UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Metoda maximální věrohodnosti a podobné techniky odhadu parametrů

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracovala: **Bc. Veronika Vykydalová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2021

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Veronika Vykydalová

Název práce: Metoda maximální věrohodnosti a podobné techniky odhadu parametrů

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2021

Abstrakt: Metoda maximální věrohodnosti je velmi rozšířenou metodou pro stanovení odhadu neznámých parametrů různých rozdělení, která díky velmi dobrým vlastnostem získaného odhadu poskytuje také nástroje pro testování hypotéz o neznámých parametrech. Cílem práce je seznámit se s vybranými modely využívajících tuto metodu odhadu, prostudovat využití modifikací metody (partial likelihood, restringovaná maximální věrohodnost) a především aplikovat získané znalosti při analýze reálných dat. První vybranou metodou byla analýza přežití využitá ke zpracování dat o onkologických pacientech mj. Coxovým modelem proporcionálních rizik. Druhou vybranou metodou byly lineární smíšené modely, kde jsme pro longitudinální data o pacientech trpících aneurysmatem aorty břišní získali populační průměrný vývoj odhadnutím marginálního modelu.

Klíčová slova: maximální věrohodnost, restringovaná maximální věrohodnost, analýza přežití, Kaplanův-Meierův odhad funkce přežití, Coxův model proporcionálních rizik, lineární smíšené modely, marginální model, longitudinální data, neúplná pozorování

Počet stran: 82

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Veronika Vykydalová

Title: Maximum likelihood and similar estimation techniques

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2021

Abstract: The maximum likelihood method is a broadly used estimation method for unknown parameters of different types of model with very good estimators properties and tools for hypothesis testing. The aim of the thesis is to study the selected models using maximum likelihood estimation, study its modifications (partial likelihood, restricted maximum likelihood) and mainly to use the acquired knowledge for real data analysis. The first selected method is survival analysis which was applied on cancer patients data including, but not limited to, the Cox proportional hazards model. The second selected method is linear mixed-effects models which was applied on longitudinal data about patients with abdominal aortic aneurysm, population-averaged evolution was obtained by the marginal model.

Key words: maximum likelihood, partial likelihood, restricted maximum likelihood, survival analysis, Kaplan-Meier estimator of the survival function, Cox proportional hazards model, linear mixed-effects models, marginal model, longitudinal data, missing data

Number of pages: 82

Number of appendices: 1

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala svému vedoucímu diplomové práce Mgr. Ondřeji Vencálkovi, Ph.D. za ochotu, věnovaný čas při konzultacích, odborné vedení a všechny udělené cenné rady. Dále bych ráda poděkovala své rodině a přátelům za podporu při studiu a tvorbě této práce.

Úvod

Metoda maximální věrohodnosti je velmi rozšířenou metodou pro získání odhadů parametrů různorodých modelů, která nám také poskytuje nástroje pro testování hypotéz o parametrech a srovnání získaných modelů.

Cílem práce je seznámit se s vybranými modely, které využívají teorie maximální věrohodnosti, resp. její modifikaci, prostudovat využití modifikací metody a především veškeré nabyté znalosti uplatnit při analýze reálných dat z oblasti medicínského výzkumu. Modely, na které se v této práci zaměříme, jsou analýza přežití a lineární smíšené modely. Obě metody slouží pro zpracování dat složitějších struktur, se kterými se často můžeme setkat v medicínském výzkumu, ale jejich využití je mnohem širší.

Analýza přežití se zabývá modelací sledovaného času do dané události, často však dochází k tzv. cenzorování, které vede k získaní neúplné informace a je třeba zohlednit při modelaci. Principy metody maximální věrohodnosti jsou využity mj. při odhadu parametrů Coxova modelu proporcionálních rizik, kde se však využívá tzv. partial likelihood modifikace. Tuto metodu využijeme pro zpracování dat o pacientech trpících rakovinou v oblasti dutiny ústní.

Lineární smíšené modely se zabývají modelací dat, u kterých je porušen předpoklad nezávislosti pozorování. Průměrný populační vývoj lze získat pomocí tzv. marginálního modelu, jehož parametry se odhadují buď klasickou metodou maximální věrohodnosti, častěji však tzv. restringovanou metodou maximální věrohodnosti, která poskytuje nevychýlené odhady. Tuto metodu využijeme pro zpracování dat pacientů trpících aneurysmatem aorty břišní, kteří byli sledováni pravidelně každých šest měsíců.

Kapitola 1

Metoda maximální věrohodnosti

Metoda maximální věrohodnosti je hojně využívanou metodou pro stanovení odhadů parametrů různých rozdělení s častým využitím také pro odhad parametrů různých uvažovaných modelů. Nejprve se tedy seznámíme s tím, dle jakých principů tato metoda stanovuje odhad, jaké vlastnosti takový odhad má a zda a jak se případně dají testovat různé hypotézy o parametrech.

Předpokládejme, že máme náhodný výběr[1] $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení pravděpodobnosti s pravděpodobnostní funkcí $p(x; \boldsymbol{\theta})$, resp. hustotou $f(x; \boldsymbol{\theta})$, jejíž obecně vektorový, neznámý parametr $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ je prvkem parametrického prostoru $\boldsymbol{\Theta}$, tj. $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Sdružená pravděpodobnostní funkce $p(\mathbf{x}; \boldsymbol{\theta})$ náhodného výběru $\mathbf{X}$, resp. sdružená hustota $f(\mathbf{x}; \boldsymbol{\theta})$, jakožto funkce proměnné $\boldsymbol{\theta}$ při pevně zvoleném $\mathbf{X} = \mathbf{x}$, se nazývá *věrohodnostní funkce*, ozn. $L(\mathbf{x}; \boldsymbol{\theta})$. Vzhledem k nezávislosti veličin $X_1, \ldots, X_n$ můžeme psát:

$$L(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}), \ \textit{resp.} \qquad L(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

Bod $\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$, ve kterém věrohodnostní funkce nabývá svého maxima, se nazývá *maximálně věrohodným odhadem parametru* $\boldsymbol{\theta}$, tj. je to bod $\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ takový, že pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí:

$$L(\mathbf{x}; \boldsymbol{\theta}) \leq L(\mathbf{x}; \widehat{\boldsymbol{\theta}}).$$

[1] Nechť $X_1, \ldots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin s rozdělením Q. Pak říkáme, že $X_1, \ldots, X_n$ je náhodný výběr z rozdělení Q. [1]

Často se můžeme také setkat s tzv. *logaritmickou věrohodnostní funkcí*, ozn. $l(\mathbf{x};\boldsymbol{\theta})$, kde $l(\mathbf{x};\boldsymbol{\theta}) = \ln L(\mathbf{x};\boldsymbol{\theta})$, které se využívá pro možné zjednodušení výpočetního vztahu odhadu, neboť platí: $\arg\max_{\theta\in\boldsymbol{\Theta}} l(\mathbf{x};\boldsymbol{\theta}) = \arg\max_{\theta\in\boldsymbol{\Theta}} L(\mathbf{x};\boldsymbol{\theta})$. [1], [2]

1.1 Obecný princip hledání maximálně věrohodného odhadu

Matematický aparát pro následující úvahy lze nalézt v odborné literatuře, např. [1, str. 146-162]. Při konstrukci maximálně věrohodného odhadu postupujeme obecně v následujících krocích:

1. Vyjádříme si předpis věrohodnostní funkce pro dané $\mathbf{X} = \mathbf{x}$.

2. Je-li to pro další výpočty výhodné, funkci zlogaritmujeme a získáme tak předpis logaritmické věrohodnostní funkce.

3. Analyticky, resp. numericky (nejčastěji s využitím Newtonovy-Raphsonovy metody) hledáme maximum (logaritmické) věrohodnostní funkce řešením tzv. systému věrohodnostních rovnic

$$\frac{\partial L(\mathbf{X},\boldsymbol{\theta})}{\partial\theta_j} = 0, \qquad \text{resp.} \quad \frac{\partial l(\mathbf{X},\boldsymbol{\theta})}{\partial\theta_j} = 0 \quad \text{pro } j = 1,\ldots,p.$$

4. Splňuje-li nalezené řešení systému věrohodnostních rovnic podmínky pro maximum funkce, je bod, ve kterém nabývá tohoto maxima, hledaným maximálně věrohodným odhadem $\widehat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$.

1.2 Vlastnosti

Vycházíme nadále z teorie uvedené především v [1].

Věta 1.1 (Princip invariance) *Je-li $\widehat{\boldsymbol{\theta}}$ maximálně věrohodný odhad parametru $\boldsymbol{\theta}$, pak je $u(\widehat{\boldsymbol{\theta}})$ maximálně věrohodný odhad parametrické funkce $u(\boldsymbol{\theta})$.*

Rozšířené využití odhadů metodou maximální věrohodnosti je také způsobeno jejich velmi dobrými asymptotickými vlastnostmi.

Věta 1.2 (Konzistence, Asymptotická normalita) *Za splnění daných, relativně jednoduchých předpokladů, uvedených v [1, str. 159, Věta 7.100], platí:*

(i) Jestliže $n \to \infty$, pak ke každému $\varepsilon > 0$ existuje s pravděpodobností blížící se jedné takové řešení $\widehat{\boldsymbol{\theta}}$ systému věrohodnostních rovnic, že $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| < \varepsilon$.

(ii) Existuje-li pro každé dostatečně velké n a pro každou hodnotu $\mathbf{X}$ takový kořen $\widehat{\boldsymbol{\theta}}$ systému věrohodnostních rovnic, že $\widehat{\boldsymbol{\theta}}$ je konzistentním odhadem parametru $\boldsymbol{\theta}$, pak

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, [\mathbf{J}(\boldsymbol{\theta})]^{-1}),$$

kde $\mathbf{J}(\boldsymbol{\theta})$ je Fisherova informační matice.

Z věty 1.2 vyplývají také vlastnosti **asymptotické nevychýlenosti** a **asymptotické eficience**, úvahy vedoucí k těmto závěrům lze nalézt např. [28, str. 27-28].

1.3 Testy založené na věrohodnostní funkci

Odhad pomocí metody maximální věrohodnosti umožňuje také získat nástroje pro testování hypotéz o neznámých parametrech:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ proti alternativě } H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Tyto testy se opírají především o znalost rozdělení maximálně věrohodného odhadu – asymptotickou normalitu. Označme pro $\boldsymbol{\theta}$ obecně p-rozměrný parametr:

$$\mathbf{U}(\boldsymbol{\theta}) = \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_p} \right)'.$$

1.3.1 Skórový test

Dá se ukázat, že testová statistika

$$LM(\boldsymbol{\theta}_0) = \frac{1}{n}[\mathbf{U}(\boldsymbol{\theta}_0)]'[\mathbf{J}(\boldsymbol{\theta}_0)]^{-1}\mathbf{U}(\boldsymbol{\theta}_0)$$

má za platnosti nulové hypotézy, předpokladů věty 1.2 a předpokladu spojitosti matice $\mathbf{J}(\boldsymbol{\theta})$ v bodě $\boldsymbol{\theta}_0$ asymptoticky χ^2-rozdělení o p stupních volnosti. Skórový test (označovaný také jako *Raův test* nebo *test založený na Lagrangeových multiplikátorech*) zamítá platnost nulové hypotézy na hladině významnosti α v případě, že $LM(\boldsymbol{\theta}_0) \geq \chi^2_p(1-\alpha)$. Výhodou tohoto testu je, že testová statistika neobsahuje vůbec maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}$, takže jej není třeba stanovovat.[1]

1.3.2 Waldův test

Dá se ukázat, že testová statistika

$$W(\boldsymbol{\theta}_0) = n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'[\mathbf{J}(\widehat{\boldsymbol{\theta}})]^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

má za platnosti nulové hypotézy a za předpokladů věty 1.2 asymptoticky χ^2-rozdělení o p stupních volnosti. Waldův test zamítá nulovou hypotézu na hladině významnosti α, pokud $W(\boldsymbol{\theta}_0) \geq \chi^2_p(1-\alpha)$. Využívá se nejčastěji pro testy hypotéz o významnosti jednotlivých parametrů v modelu. [1]

1.3.3 Likelihood-ratio test

Dá se ukázat, že testová statistika

$$LR(\boldsymbol{\theta}_0) = -2\ln\left(\frac{L(\boldsymbol{\theta}_0)}{L(\widehat{\boldsymbol{\theta}})}\right) = 2[l(\widehat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}_0)]$$

má za platnosti nulové hypotézy a za předpokladů věty 1.2 asymptoticky χ^2-rozdělení o p stupních volnosti. Likelihood-ratio test, neboli *test založený na věrohodnostním poměru*, zamítá nulovou hypotézu na hladině významnosti α, pokud $LR(\boldsymbol{\theta}_0) \geq \chi^2_p(1-\alpha)$. Výhodou oproti předcházejícím testům je, že nevyžaduje určení Fisherovy informační matice.

Likelihood-ratio test se využívá nejčastěji pro testování významnosti rozdílu mezi modelem (ozn. m_1) a jeho podmodelem[2] (ozn. m_2). Testujeme tedy, zda

[2]např. model, který vznikl odebráním některých parametrů původního modelu nebo lineární kombinací některých parametrů

pozorovaný rozdíl mezi věrohodností původního modelu a věrohodností jeho podmodelu je signifikantní. Pokud by tomu tak bylo, není možné přejít od původního modelu k jeho podmodelu bez zhoršení kvality modelu. Původní testová statistika lze převést do tvaru

$$LR = -2 \ln \left(\frac{L(m_2)}{L(m_1)} \right) = 2[l(m_1) - l(m_2)],$$

přičemž se dá ukázat, že má χ^2-rozdělení o r stupních volnosti, kde r je rozdíl mezi počtem parametrů m_1 a m_2. Nulovou hypotézu zamítáme opět pro velké hodnoty testové statistiky, tj. pro $LR \geq \chi^2_r(1 - \alpha)$ zamítneme možnost přejít k jednoduššímu modelu. [1], [19]

Kapitola 2

Analýza přežití

Analýza přežití je nástrojem pro zpracování dat zachycující dobu do definované události. V této části se seznámíme s nutnou teorií pro pochopení následné praktické aplikace metody na reálná data. Nejčastěji se využívá v oblasti medicíny ve vztahu k úmrtí, odkud také získala svůj název a terminologii. Sledovaná událost může být ale také cokoliv jiného, např. v oblasti průmyslu porucha systému s časem sledování od jeho spuštění do této poruchy, je však nutné na začátku analýzy tuto událost jasně definovat a dle toho také přistupovat k interpretaci získaných výsledků. Jelikož se i v praktické části budeme zabývat daty z medicínského výzkumu, uvedeme si dva nejčastější příklady uvažovaných událostí v tomto prostředí:

- **Celkové přežití** (ozn. OS, z angl. overall survival) – zkoumanou událostí je smrt z jakékoliv příčiny.
- **Přežití specifické pro sledované onemocnění** (ozn. DSS, z angl. disease specific survival) – zkoumanou událostí je smrt v důsledku sledovaného onemocnění, úmrtí z jiného důvodu je v tomto případě tzv. cenzorovaným pozorováním, jak vysvětlíme dále.

Cenzorovaná pozorování jsou problémem, se kterým se při této analýze můžeme často setkat. Nejčastěji je pozorování označeno jako cenzorované, pokud nedošlo během sledované doby k námi definované události. Tento typ cenzorování je označován za *cenzorování zprava* – víme, že skutečná hodnota je větší

než naše napozorovaná. Tento typ cenzorování budeme také dále uvažovat. Údaj může být však také *cenzorován zleva* – skutečná hodnota je menší než pozorovaná, např. měřená hodnota pod detekčním limitem přístroje, nebo také existuje *intervalové cenzorování* – skutečná hodnota je mezi známými hodnotami, např. úmrtí v intervalu mezi dvěma kontrolními vyšetřeními bez přesného data úmrtí.[5]

Uvažujme nyní sledovanou dobu přežití jako realizaci náhodné veličiny ozn. T, pro kterou si zde uvedeme další základními pojmy, vztahující se zejména k popsání jejího chování, definované dle [6], nebude-li uvedeno jinak.

- **Funkce přežití**

  $$S(t) = P(T > t), \quad 0 < t < \infty$$

  Funkce přežití určuje pravděpodobnost, s jakou bude čas přežití větší než t. Je vlastně doplňkem k distribuční funkci[1] náhodné veličiny T sledovaného času. Je to nerostoucí, zprava spojitá funkce, nabývajících hodnot od 1 (v čase 0) do 0, pod kterou nikdy neklesne.

- **Hazardní funkce**

  Obecně dle [4] definujeme hazardní funkci jako

  $$h(t) = \lim_{\delta \to 0^+} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta}.$$

  Hazardní funkce určuje okamžitou míru selhání, tj. pravděpodobnost, že za předpokladu, že se osoba dožila času t, dojde k události v následujícím velmi blízkém okamžiku. Pro spojité rozdělení náhodné veličiny T lze hazardní funkce také vyjádřit pomocí její hustoty $f(t)$ a funkce přežití jako

  $$h(t) = \frac{f(t)}{S(t)}.$$

  Často také využijeme ještě tzv. **kumulativní hazardní funkci**, která vyjadřuje plochu pod hazardní funkcí do času t, tj.

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

[1] $F(t) = P(T \leq t) = 1 - S(t), \quad 0 < t < \infty$

pro kterou při spojitém rozdělení náhodné veličiny T platí vztah

$$S(t) = \exp(-H(t)).$$

Nejčastěji se setkáváme s předpokladem spojitého rozdělení, je-li však uvažováno diskrétní rozdělení náhodné veličiny T, lze nalézt definice výše uváděných vztahů v literatuře, např. [4, kap. 1.2.2], podobně pro smíšené rozdělení náhodné veličiny T [4, kap. 1.2.3].

- **Medián přežití**

  Medián přežití je definován jako čas t, pro který platí, že $S(t) = \frac{1}{2}$. Není-li funkce $S(t)$ spojitá v bodě $\frac{1}{2}$, je medián určen jako nejmenší hodnota t, pro kterou platí: $S(t) \leq \frac{1}{2}$. Pokud funkce přežití neklesá pod hodnotu $\frac{1}{2}$, pak není medián přežití definován.

2.1 Kaplanův-Meierův odhad funkce přežití

Kaplanův-Meierův odhad je neparametrickým odhadem funkce přežití, neklade tedy žádné předpoklady na distribuci náhodné veličiny T popisující sledovaný čas přežití. Je to jeden ze základních nástrojů popisné statistiky využívané pro analýzu dat s cenzorovanými pozorováními.

Při stanovování odhadu využívá informaci necenzorovaných, ale také cenzorovaných pozorování – v libovolném časovém okamžiku je počet pozorování, u kterých ještě nedošlo do daného času k události či cenzorování, označován jako pozorovaný počet v riziku, ozn. n_i. Označme dále počet událostí, ke kterým došlo v daný časový okamžik, jako d_i a seřazený pozorovaný čas přežití $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$, přičemž m udává počet časových okamžiků, u kterých došlo k události z celkových n pozorování. Kaplanův-Meierův odhad funkce přežití v čase t je poté dán vztahem

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i},$$

kde dle konvence $\hat{S}(t) = 1$, pokud $t < t_{(1)}$. Při takto stanoveném odhadu cenzorovaná pozorování přispívají svojí informací v počtu pozorování v riziku, dokud nejsou ztraceni ze sledování. Je-li pozorování s nejdelším sledovaným časem cenzorované, poté Kaplanův-Meierův odhad funkce přežití neklesá až k 0, nejmenší odhadnutá hodnota náleží poslední pozorované události.

Nejčastěji se využívá k vykreslení odhadu funkce přežití a grafické analýze této křivky. Její tvar závisí na pozorovaném čase přežití a podílu cenzorovaných dat. [3]

Kromě bodového odhadu lze konstruovat také intervalový odhad funkce přežití pro daný časový okamžik t. V teorii je několik možných přístupů pro odvození, dle [3] a [4] uvedeme přístup vycházející z teorie čítacích procesů, který je doporučen a bude také využit v praktické části. Dle této teorie je dokázáno, že Kaplanův-Meierův odhad a jeho funkce mají asymptoticky normální rozdělení, což nám umožní sestrojit intervalový odhad pomocí známých principů. Otázkou pouze zůstává, jak stanovit rozptyl odhadu funkce přežití. Prvním zkonstruovaným odhadem variability je tzv. *Greenwoodův vzorec* 2.1, který pro odvození využívá funkce přirozeného logaritmu odhadu funkce přežití a platných vztahů (více např. viz [3, str. 28-29], [4, str. 16-17]). Dle tohoto přístupu získáme

$$\widehat{var}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{2.1}$$

Využití vztahu 2.1 však nezaručí, aby hodnoty intervalového odhadu respektovaly definiční obor funkce přežití. Proto se doporučuje využití tzv. log-log transformace funkce přežití $\ln(-\ln[\hat{S}(t)])$, pro kterou platí

$$\widehat{var}(\ln(-\ln[\hat{S}(t)])) = \frac{1}{(\ln(\hat{S}(t)))^2} \sum_{t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{2.2}$$

Pro *log-log funkci přežití* můžeme získat $100(1-\alpha)\%$ intervalový odhad, ozn. $(\hat{c}_l, \hat{c}_u)$, pro daný časový okamžik t využitím normality a odhadu rozptylu dle vztahu 2.2, kde

$$\hat{c}_l = \ln(-\ln[\hat{S}(t)]) - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\ln(-\ln[\hat{S}(t)]))}$$
$$\hat{c}_u = \ln(-\ln[\hat{S}(t)]) + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\ln(-\ln[\hat{S}(t)]))},$$

přičemž $100(1-\alpha)\%$ intervalový odhad pro původní odhadnutou funkci přežití je získán zpětnou transformací jako

$$(\exp[-\exp(\hat{c}_u)]; \exp[-\exp(\hat{c}_l)])\,.$$

2.1.1 Odhad mediánu přežití

Z definice mediánu přežití vyplývá, že jeho odhad můžeme získat z Kaplanova-Meierova odhadu funkce přežití jako čas t, pro který platí $\hat{S}(t) = ½$, resp. $\widehat{t}_{med} = \min\{t : \hat{S}(t) \leq 0,5\}$. Často se interpretuje tato číselná charakteristika jako souhrn celého Kaplanova-Meierova odhadu. Vyjadřuje časový okamžik, kterého se dožije polovina sledovaných. Pro získání $100(1-\alpha)\%$ intervalového odhadu $\widehat{t}_{med}$ využíváme znalosti konstrukce intervalových odhadů funkce přežití. Intervalový odhad $\widehat{t}_{med}$ totiž musí zahrnovat všechny časové okamžiky t, v nichž intervalový odhad funkce přežití obsahuje hodnotu $\widehat{t}_{med}$. Graficky jej získáme jako průsečíky horizontální osy konstruované v ½ s nejkrajnějšími body intervalových odhadů vykresleného Kaplanova-Meierova odhadu funkce přežití. Vyčíslit jej pak můžeme pomocí uvažované testové statistiky využívající *log-log transformace*:

$$\frac{\ln\{-\ln[\hat{S}(t)]\} - \ln\{-\ln[(0,5)]\}}{\sqrt{\widehat{var}(\ln(-\ln[\hat{S}(t)]))}}$$

pro $H_0 : S(t) = ½$, za jejíž platnosti má tato testová statistika normované normální rozdělení a konfidenční interval pro $\widehat{t}_{med}$ je definovaný jako všechny hodnoty t, pro které bychom nezamítali nulovou hypotézu. Blíže jsou tyto úvahy vysvětleny v [3, str. 36-40] nebo v [6, kap. 3.2].

2.1.2 Porovnání odhadnutých funkcí přežití

Obvykle je v centru našeho zájmu také porovnání chování přežití dle definovaných kategorií různých proměnných. Pro tyto skupiny můžeme zvlášť získat Kaplanovy-Meierovy odhady funkcí přežití a rozdíl mezi nimi porovnat nejen graficky, ale také dle tzv. log-rank testu.

Log-rank test

Předpokládejme, že obecně nás zajímá porovnání p funkcí přežití, přičemž nulová hypotéza je, že funkce přežití si jsou rovny. Testová statistika je založena na rozdílu mezi pozorovanými počty událostí v daný časový okamžik t a očekávanými počty událostí v tomto okamžiku t za předpokladu platnosti nulové hypotézy.

Předpokládejme nyní, že máme zhruba dodrženu proporcionalitu rizik, tj. poměr rizik mezi p skupinami je v čase konstantní. Označme dále $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ časy všech událostí, ke kterým došlo. Pro každý časový okamžik t_j, kde $j = 1, \ldots, m$, lze pozorování v jednotlivých skupinách shrnout do kontingenční tabulky 2.1.

	Skupina 1	...	Skupina i	...	Skupina p	Celkem
Počet událostí	d_{1j}	...	d_{ij}	...	d_{pj}	d_j
Počet přeživších	$n_{1j} - d_{1j}$	...	$n_{ij} - d_{ij}$	...	$n_{pj} - d_{pj}$	$n_j - d_j$
Počet v riziku	n_{1j}	...	n_{ij}	...	n_{pj}	n_j

Tabulka 2.1: Kontingenční tabulka shrnující počty nastalých událostí a přeživších v čase události t_j, pro $j = 1, \ldots, m$, dle rozdělení do p skupin.

Dá se ukázat, že podmíněné rozdělení $d_{1j}, \ldots, d_{pj}$ při daném d_j je mnohorozměrné hypergeometrické rozdělení a podmíněná střední hodnota pro d_{ij} je rovna

$$e_{ij} = n_{ij} d_j n_j^{-1}.$$

Proto statistika $w_j = (d_{1j} - e_{1j}, \ldots, d_{pj} - e_{pj})'$ bude mít nulovou střední hodnotu a kovarianční matici ozn. W_j rozměrů $p \times p$. Dále označme $w = \sum_{j=1}^{m} w_j$. Za předpokladu nezávislosti m časových okamžiků, ve kterých došlo k události, jejichž pozorování dle jednotlivých skupin lze shrnout do celkem m kontingenčních

tabulek tvaru 2.1, platí, že kovarianční matice statistiky w, ozn. W, se získá jako: $W = W_1 + \cdots + W_m$. Potom testová statistika

$$w'W^{-1}w$$

bude mít asymptoticky χ^2-rozdělení o $p-1$ stupních volnosti, nulovou hypotézu zamítáme pro velké hodnoty testové statistiky. Podrobné úvahy a odvození všech výše nastíněných potřebných vztahů lze nalézt např. v knize [4, str. 20-23], podrobněji potom v článku [12].

Pokud je proporcionalita rizik porušena, je možné využití tzv. *váženého log-rank testu*, který přiřazuje váhy rozdílům mezi pozorovanými počty událostí v daný časový okamžik t a očekávanými počty událostí v tomto okamžiku t za předpokladu platnosti nulové hypotézy dle toho, jak se podíl rizika mezi skupinami liší. Dle stanovení podoby těchto vah rozlišujeme např. *Gehanův-Breslův test [13], Taronův-Warův test [14], Petovu-Petovu modifikaci [15], Flemingův-Harringtonův test [16].* Více o podobě jednotlivých testových statistik se můžeme dočíst v uvedené literatuře. Dá se opět ukázat, že mají asymptoticky χ^2-rozdělení o $p-1$ stupních volnosti, přičemž nulovou hypotézu o rovnosti funkcí přežití zamítáme pro velké hodnoty testové statistiky. [4]

2.2 Coxův model proporcionálních rizik

Coxův model proporcionálních rizik je semi-parametrickým modelem, který nám umožní zahrnout a interpretovat především vlivy různých rizikových faktorů a jejich kombinací pro riziko nastání dané události. To nejlépe vystihuje hazardní funkce, která je v případě Coxova modelu proporcionálních rizik definována jako

$$h(t, \mathbf{x}, \boldsymbol{\beta}) = h_0(t) \cdot e^{\mathbf{x}'\boldsymbol{\beta}}, \tag{2.3}$$

tedy jako součin základního hazardu $h_0(t)$ (tzv. *baseline hazard function*), který je společný pro všechna pozorování, závisí pouze na čase t, a parametrizované složky modelující vliv p faktorů $\mathbf{x}$ pomocí parametrů $\boldsymbol{\beta}$ nezávisející na čase, který se projevuje násobkem základního rizika. Baseline hazard function je ponechána

bez předpokladů na typ rozdělení doby přežití, je odhadována neparametricky, což se projeví především později při postupu odhadu parametrů Coxova modelu. [5], [3]

Poměrem rizik rozumíme

$$HR = \frac{h(t, \mathbf{x_1}, \boldsymbol{\beta})}{h(t, \mathbf{x_0}, \boldsymbol{\beta})},$$

přičemž dle definice hazardní funkce 2.3 platí

$$HR = \frac{h(t, \mathbf{x_1}, \boldsymbol{\beta})}{h(t, \mathbf{x_0}, \boldsymbol{\beta})} = \frac{h_0(t) \cdot e^{\mathbf{x_1}'\boldsymbol{\beta}}}{h_0(t) \cdot e^{\mathbf{x_0}'\boldsymbol{\beta}}} = e^{(\mathbf{x_1}-\mathbf{x_0})'\boldsymbol{\beta}}. \tag{2.4}$$

Získáváme tak jasnou interpretaci parametrů Coxova modelu bez nutnosti odhadu funkce základního hazardu. Uvažujeme-li nyní binární faktor ovlivňující riziko (např. 1 = placebo, 0 = léčba), dle vztahu 2.4 platí

$$HR = \frac{h_0(t) \cdot e^{1 \cdot \beta}}{h_0(t) \cdot e^{0 \cdot \beta}} = e^{\beta}, \tag{2.5}$$

tj. míra rizika okamžitého selhání (tj. nastání sledované události) bez ohledu na časový okamžik t, za předpokladu, že do tohoto okamžiku k události nedošlo, je pro kategorii 1 (placebo skupina) e^{β} násobně vyšší oproti kategorii 0 (léčba). Má-li uvažovaný faktor více kategorií, poté interpretujeme míru rizika vzhledem ke zvolené referenční kategorii. V případě, že je některá z vysvětlujících proměnných numerická spojitá, pak je možno e^{β} vnímat jako navýšení rizika, ke kterému dojde, pokud se uvažovaný faktor zvýší o jednotku (např. věk pacienta při vstupu do studie bude o rok vyšší). [3], [5]

Dá se také ukázat, že dle vztahů mezi funkcí přežití a hazardní funkcí, definovanou dle vztahu 2.3, platí

$$S(t, \mathbf{x}, \boldsymbol{\beta}) = [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}, \tag{2.6}$$

kde $S_0(t)$ je základní funkcí přežití označovaná jako *baseline survival function*. [3]

Podoba multiplikativně definovaného hazardu 2.3 však vede také k zásadnímu předpokladu tohoto modelu – proporcionalitě rizik, tj. poměr rizik je v čase

konstantní. Než se však budeme věnovat možnému ověření tohoto a dalších předpokladů semi-parametrického Coxova modelu, zaměřme se ještě na to, jak jeho parametry odhadnout.

2.2.1 Odhad parametrů

Již několikrát jsme pojmenovali Coxův model proporcionálních rizik jako semi-parametrický model, avšak ještě ani jednou jsme se nezamysleli nad tím, co to vlastně znamená. Ve vztahu 2.3 jsou neznámými základní hazardní funkce $h_0(t)$ a parametry $\boldsymbol{\beta}$. Jak jsme si ukázali (vztah 2.4), pro interpretaci a porovnávání vlivu jednotlivých rizikových faktorů nepotřebujeme znát odhad $h_0(t)$. Je to také jedním z důvodů, proč se baseline hazard function může ponechat bez předpokladů na rozdělení doby přežití, a proto není nijak parametrizována, což vede k označení Coxova modelu semi-parametrickým.

Odhad parametrů $\boldsymbol{\beta}$ je založen na metodě maximální věrohodnosti, která je zde však modifikována na tzv. *partial likelihood*, přičemž je dokázáno, že asymptotické vlastnosti metody maximální věrohodnosti jsou zachovány také pro partial likelihood. Teorie partial likelihood vychází z myšlenky, že sdružené rozdělení pravděpodobnosti pozorování závisí na neznámých parametrech primárního zájmu (v našem kontextu se jedná o $\boldsymbol{\beta}$) a tzv. rušivých parametrech (v našem kontextu $h_0(t)$), přičemž věrohodnostní funkci lze faktorizovat na funkci neznámých parametrů primárního zájmu, a druhou část, která je funkcí rušivých parametrů a parametrů primárního zájmu, které zde však již nepřispívají velkou měrou informace. Následující úvahy čerpají z literatury [4, kap. 4.2] a [3, kap. 3.3], kde lze také nalézt rozšiřující informace o této metodě.

Uvažujeme, že pro i-té pozorování máme k dispozici trojici $(t_i, \mathbf{x}_i, c_i)$, kde t_i je pozorovaný čas, $\mathbf{x}_i$ jsou získané hodnoty vysvětlujících proměnných a c_i je indikátor cenzorování zprava (1 = nastala událost, 0 = cenzor). Předpokládejme, že celkem došlo k m událostem v různé časové okamžiky a lze je seřadit tak, že platí: $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. Označme $\mathcal{R}(t_j)$ množinu indexů pozorování, které jsou v daný časový okamžik t_j v riziku, tj. platí $\mathcal{R}(t_j) = \{i : t_i \geq t_j\}$.

Pro i-té zprava cenzorované pozorování v čase t_i je obecně příspěvek k věrohodnosti $L_i(\boldsymbol{\beta}) = S(t_i, \mathbf{x}_i, \boldsymbol{\beta})$. Pokud by u i-tého pozorování došlo k pozorované události, potom je jeho příspěvek k věrohodnosti $L_i(\boldsymbol{\beta}) = h(t_i, \mathbf{x}_i, \boldsymbol{\beta}) \cdot S(t_i, \mathbf{x}_i, \boldsymbol{\beta})$. Věrohodnostní funkce lze pak vyjádřit jako

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} h(t_i, \mathbf{x}_i, \boldsymbol{\beta})^{c_i} \cdot S(t_i, \mathbf{x}_i, \boldsymbol{\beta}),$$

přičemž po rozšíření členem $\left[\frac{\sum_{\ell \in \mathcal{R}(t_i)} h(t_i, \mathbf{x}_\ell, \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_i)} h(t_i, \mathbf{x}_\ell, \boldsymbol{\beta})}\right]^{c_i}$ získáme

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{h(t_i, \mathbf{x}_i, \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_i)} h(t_i, \mathbf{x}_\ell, \boldsymbol{\beta})}\right]^{c_i} \cdot \left[\sum_{\ell \in \mathcal{R}(t_i)} h(t_i, \mathbf{x}_\ell, \boldsymbol{\beta})\right]^{c_i} S(t_i, \mathbf{x}_i, \boldsymbol{\beta}). \tag{2.7}$$

Dle úvah uvedených v článku [17] první člen obsahuje téměř veškerou informaci o parametru $\boldsymbol{\beta}$, zatímco další dva členy obsahují informaci především o baseline hazard function. První člen z věrohodnostní funkce 2.7 je tedy naše hledaná partial likelihood pro parametry primárního zájmu $\boldsymbol{\beta}$. Pro definici hazardní funkce 2.3 můžeme dále partial likelihood upravit

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left[\frac{h(t_i, \mathbf{x}_i, \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_i)} h(t_i, \mathbf{x}_\ell, \boldsymbol{\beta})}\right]^{c_i} = \prod_{i=1}^{n} \left[\frac{h_0(t_i) \exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_i)} h_0(t_i) \exp(\mathbf{x}_\ell' \boldsymbol{\beta})}\right]^{c_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_i)} \exp(\mathbf{x}_\ell' \boldsymbol{\beta})}\right]^{c_i} = \prod_{i=1}^{m} \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_i)} \exp(\mathbf{x}_\ell' \boldsymbol{\beta})}, \end{aligned}$$

pomocí které můžeme získat klasickými postupy známými z metody maximální věrohodnosti odhad $\widehat{\boldsymbol{\beta}}$ a testovat hypotézy o parametrech $\boldsymbol{\beta}$.

Dle vztahu 2.6 jsme také schopni získat z odhadů parametrů rizikových faktorů odhad funkce přežití pro dané hodnoty rizikových faktorů. Potřebujeme k tomu však ještě stanovit odhad baseline survival function $S_0(t)$. Nejčastěji se můžeme setkat s tzv. *Breslowovým odhadem*, který je založen na odhadu tzv. cumulative baseline hazard

$$\widehat{H_0}(t) = \sum_{i: t_i < t} \left(\frac{d_i}{\sum_{\ell \in \mathcal{R}(t_i)} \exp\left(\mathbf{x}_\ell' \widehat{\boldsymbol{\beta}}\right)}\right),$$

přičemž odhad baseline survival function získáme jako $\widehat{S_0}(t) = \exp\left(-\widehat{H_0}(t)\right)$. [7]

2.2.2 Ověření předpokladů

Prvním předpokladem, který model 2.3 klade, je **lineární vztah** numerických spojitých vysvětlujících proměnných s vysvětlovaným logaritmem hazardu, tj. předpokládáme, že platí

$$\ln\left[h(t, \mathbf{x}, \boldsymbol{\beta})\right] = \ln\left[h_0(t)\right] + x_1\beta_1 + \cdots + x_p\beta_p. \tag{2.8}$$

Zda daný rizikový faktor vyhovuje tomuto lineárnímu vztahu, či je třeba nějaké jeho transformace, můžeme ověřit nejčastěji graficky pomocí tzv. *martingal reziduí*. Martingal rezidua jsou definovaná pro i-té pozorování, které je určeno opět trojicí $(t_i, \mathbf{x}_i, c_i)$, jako

$$m_i = c_i - \widehat{H}(t_i, \mathbf{x}, \widehat{\boldsymbol{\beta}}) = c_i + \ln\left[\widehat{S_0}(t_i)\right] \exp \mathbf{x}'\widehat{\boldsymbol{\beta}}, \tag{2.9}$$

kde $\widehat{H}(t_i, \mathbf{x}, \widehat{\boldsymbol{\beta}})$ je odhad kumulativní hazardní funkce, která lze také vyjádřit dle vztahu mezi kumulativní hazardní funkcí a funkcí přežití, s využitím definice Coxova modelu 2.3. Vykreslíme-li tato rezidua získaná z „nulového“ Coxova modelu (tj. bez jakéhokoliv prediktoru) oproti hodnotám daných numerických spojitých vysvětlujících proměnných, s využitím tzv. LOESS vyhlazení získáme křivku, jejíž tvar odpovídá danému vztahu prediktoru k modelu. Je-li tedy její tvar lineární, je předpoklad splněn, jinak hledáme vhodnou transformaci, kterou této linearity dosáhneme. [6], [3]

Nejvýznamnějším předpokladem je však **proporcionalita rizik**, tj. že poměr rizik je v čase konstantní, nezávisí na časovém okamžiku t, neboť nedodržení tohoto předpokladu je hrubým porušením vlastnosti využívané pro odhad i interpretaci získaných parametrů. Nejčastěji ověřujeme dodržení tohoto předpokladu pomocí tzv. *Schoenfeldových reziduí*, resp. jejich vážené formy. Jsou založena na individuálním příspěvku daného pozorování k derivaci logaritmu partial likelihood, tj. odhad Schoenfeldova rezidua i-tého pozorování vzhledem ke k-té vysvětlující

proměnné je dán vztahem

$$\widehat{r}_{ik} = c_i \left(x_{ik} - \hat{\bar{x}}_{w_i k} \right), \tag{2.10}$$

kde

$$\hat{\bar{x}}_{w_i k} = \frac{\sum_{j \in R(t_i)} x_{jk} e^{\mathbf{x_j}' \widehat{\boldsymbol{\beta}}}}{\sum_{j \in R(t_i)} e^{\mathbf{x_j}' \widehat{\boldsymbol{\beta}}}}.$$

Z definice vyplývá, že Schoenfeldova rezidua jsou nulová pro všechna cenzorovaná pozorování, nevypovídají tedy nijak o vhodnosti zvoleného modelu. [3]

K ověřování předpokladů se využívá spíše jejich škálovaná forma ve tvaru

$$\hat{\mathbf{r}}_i^* = m \cdot \widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) \hat{\mathbf{r}}_i,$$

s využitím aproximace varianční matice samotných reziduí pomocí odhadu variační matice $\widehat{\boldsymbol{\beta}}$. Vyjádříme-li si časově závislý prediktor jako

$$\beta_j(t) = \beta_j + \gamma_j g_j(t),$$

kde $g_j(t)$ je libovolná zvolená funkce času, dá se ukázat, že přibližně platí

$$\mathsf{E}\left[\mathbf{r}_j^*(t)\right] \doteq \gamma_j g_j(t). \tag{2.11}$$

Dle vztahu 2.11 by tedy pro škálovaná Schoenfeldova rezidua vykreslená oproti sledovanému času mělo platit, že jsou při dodržení předpokladu proporcionality rizik průměrně okolo nuly. Pokud tomu tak není, dá se z tohoto grafu s využitím vyhlazení vyčíst, jakou funkční závislost na čase lze u dané vysvětlující proměnné pozorovat. Tuto možnou lineární závislost na čase lze také otestovat – nejčastěji dle zvolené modelace funkce času[2] testovými statistikami vedoucími ke skórovému testu významnosti přidání tohoto časově závislého prediktoru do modelu. [3], [18]

Není-li předpoklad proporcionality rizik dodržen, lze využít metody stratifikovaného Coxova modelu či modely zahrnující časově závislé prediktory, které jsou ovšem již nad rámec obsahu této práce a více se jim věnovat nebudeme.

[2] nejčastěji $g(t) = t, \quad g(t) = \widehat{S}_{KM}(t), \quad g(t) = rank(t)$

Kapitola 3

Analýza přežití – aplikace

V této části se budeme věnovat analýze přežití z praktického hlediska – budeme aplikovat tuto metodu na reálná data, se kterými se nejprve seznámíme nástroji popisné statistiky, následně zkonstruujeme Kaplanovy-Meierovy odhady funkcí přežití a nalezneme nejvhodnější Coxův model proporcionálních rizik, jehož předpoklady také ověříme.

3.1 Popis datové sady

Data, kterými se budeme v této části práce zabývat, pocházejí z oblasti onkologie a byly poskytnuty ke zpracování vedoucím práce. Originální datová sada obsahuje celkem 49 proměnných. Se všemi se však seznamovat nebudeme, ať už pro jejich nerelevantnost vzhledem k naší analýze (např. různé výsledky testů pro celkové určení HPV pozitivity), nebo pro jejich nedostatečné zastoupení ve vzorku pro projevení efektu dané proměnné při analýze (např. informace o onemocnění diabetem - celkem pouze 15 pacientů trpí diabetem I., nebo II. stupně).

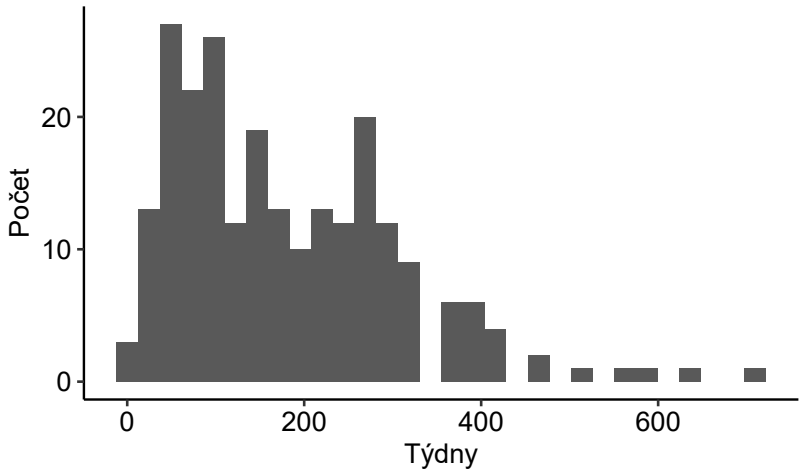
Naše datová sada tedy obsahuje anonymizované informace o celkem 234 pacientech trpících rakovinou bez metastází v oblasti orofarynx, tj. ústní části hltanu, podstupujících léčbu na pracovištích v Brně, nebo v Praze.

Celkový čas sledování vyjádřený v týdnech jsme získali jako čas přežití, resp. čas ve studii, rozdílem mezi datem diagnózy a datem úmrtí, resp. datem posledního kontaktu pro cenzorovaná pozorování. Jeho typicky zešikmené rozdělení

je zobrazeno na obrázku 3.1. Základní popisné charakteristiky jsou pak shrnuty v tabulce 3.1.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6	80	152	180	264	716

Tabulka 3.1: Základní popisné charakteristiky času sledování v týdnech.



Obrázek 3.1: Histogram času sledování v týdnech.

Při analýze budeme uvažovat sledovanou událost úmrtí z jakéholiv důvodu, zajímá nás tedy tzv. overall survival, tj. celkové přežití. Z celkového počtu 234 pacientů v průběhu sledování nezemřelo 122 z nich, celkový podíl cenzorovaných pozorování je tedy 52 %.

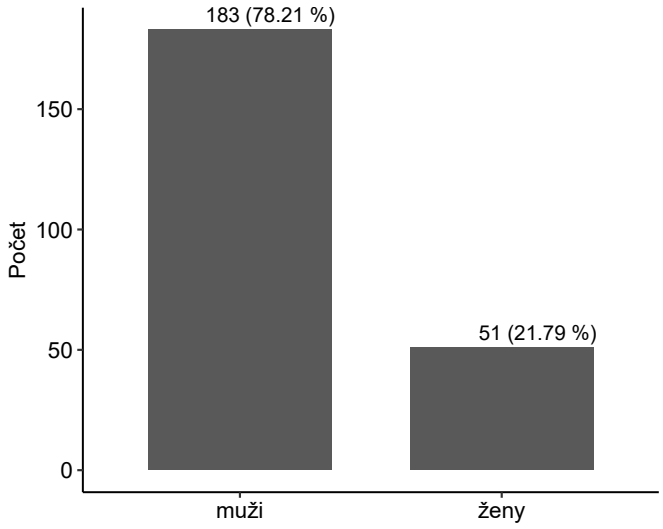
Uvažované proměnné, které mohou mít vliv na dobu přežití, můžeme pro přehlednost rozdělit do dvou kategorií:

1. faktory zachycující stav a životní styl pacienta, socioekonomické ukazatele
2. faktory vztahující se k samotné nemoci, resp. nádoru a léčbě

Nejprve se zaměřme tedy na proměnné, které by mohly mít vliv na dobu přežití, ale nejsou přímo spojeny s onemocněním:

- **pohlaví**

  Převažuje zastoupení mužů v datovém souboru, jak lze vidět z obrázku 3.2.



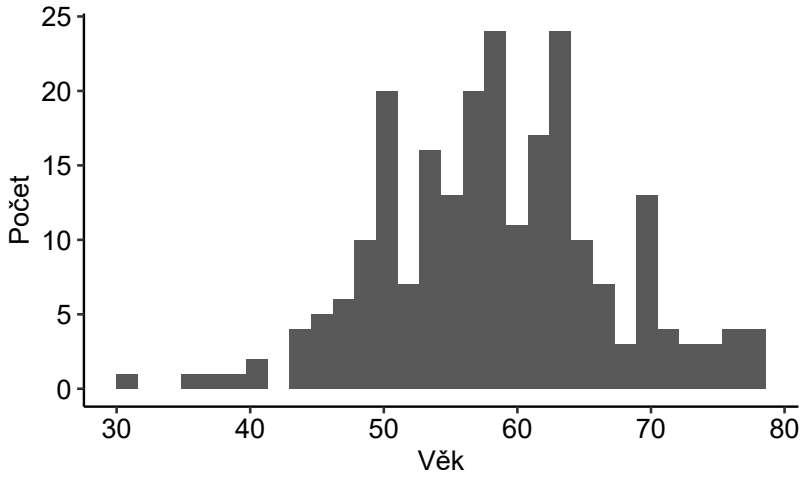
Obrázek 3.2: Četnost a procentuální rozdělení pacientů dle pohlaví v datové sadě.

- **věk v čase diagnózy**

  Základní popisné charakteristiky věkové struktury pacientů jsou zachyceny v tabulce 5.1, zobrazení tohoto rozdělení je zachyceno na obrázku 3.3.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
31	53	59	58,4	64	78

Tabulka 3.2: Základní popisné charakteristiky věku v čase diagnózy pacientů.



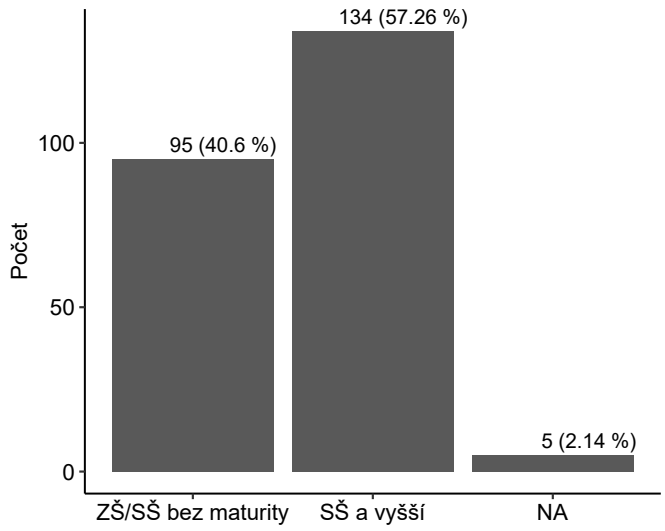
Obrázek 3.3: Rozdělení pacientů dle věku v čase diagnózy.

- **vzdělání**

  V původních datech je vyjádřeno jako počet odstudovaných let. Pro lepší interpretovatelnost jsme si rozdělili pacienty do dvou skupin dle vzdělání:

  - s dobou vzdělávání < 12 let odpovídající dosaženému základnímu, nebo středoškolskému vzdělání pravděpodobně bez maturity
  - s dobou vzdělávání ≥ 12 let zachycující osoby s ukončeným SŠ vzděláním s maturitou, nebo vyšším stupněm vzdělání

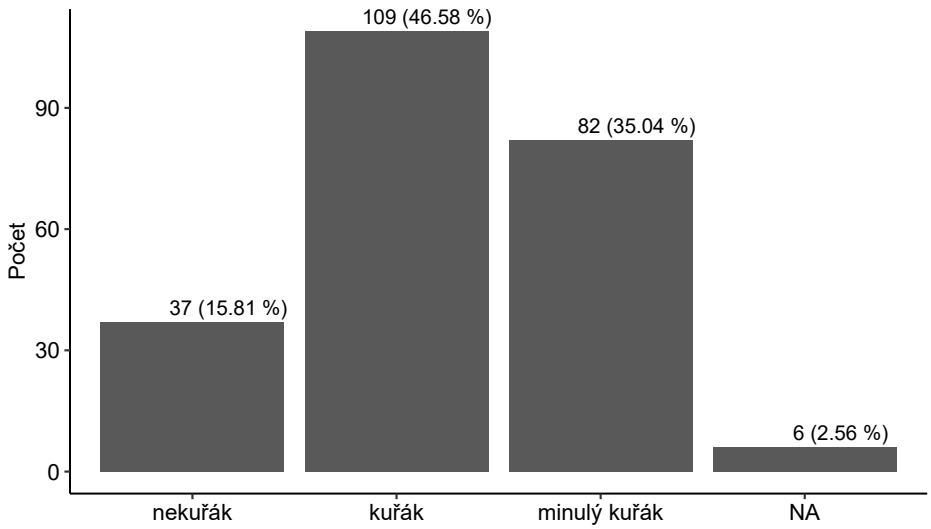
  Rozdělení vzdělání do těchto skupin nám pomůže lépe vystihnout rozdělení dle různého sociálního postavení pacienta, stejně tak jako možnou spojitost úrovně vzdělání s úrovní ústní hygieny. Právě nedostatečná ústní hygiena může také hrát roli při vzniku nádorů HPV negativního původu [11]. Zastoupení jednotlivých skupin vzdělání je zachyceno na obrázku 3.4.



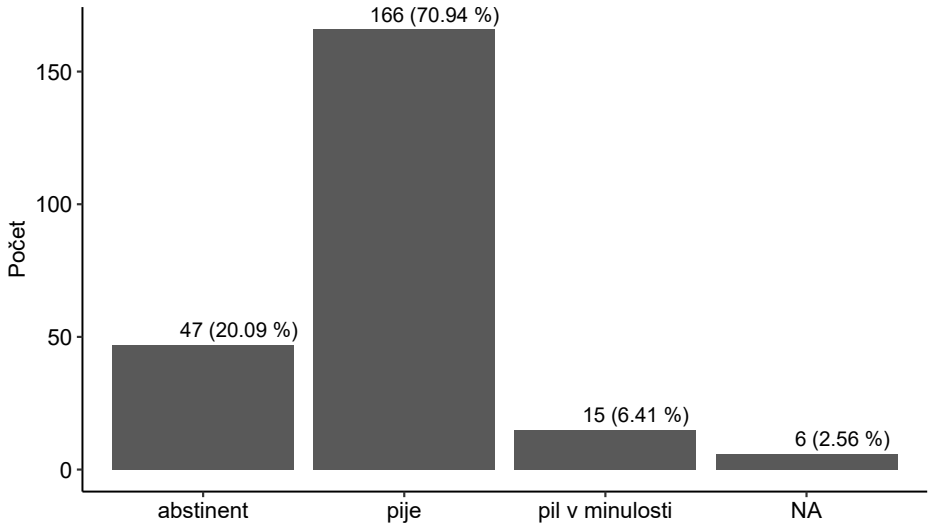
Obrázek 3.4: Četnost a procentuální rozdělení pacientů dle úrovně dosaženého vzdělání.

- **kouření**, **alkohol**

  Poslední uvažované, avšak velmi zásadní, rizikové faktory, nevztahující se přímo k nádoru a jeho léčbě, jsou kouření a alkohol. Jejich asociace s výskytem nádorů v oblasti hlavy a krku jsou také různými studiemi prokazatelné [11]. Zastoupení jednotlivých skupin kuřáků a konzumentů alkoholu je zobrazeno na obrázku 3.5, resp. 3.6.



Obrázek 3.5: Četnost a procentuální rozdělení pacientů dle vztahu ke kouření.

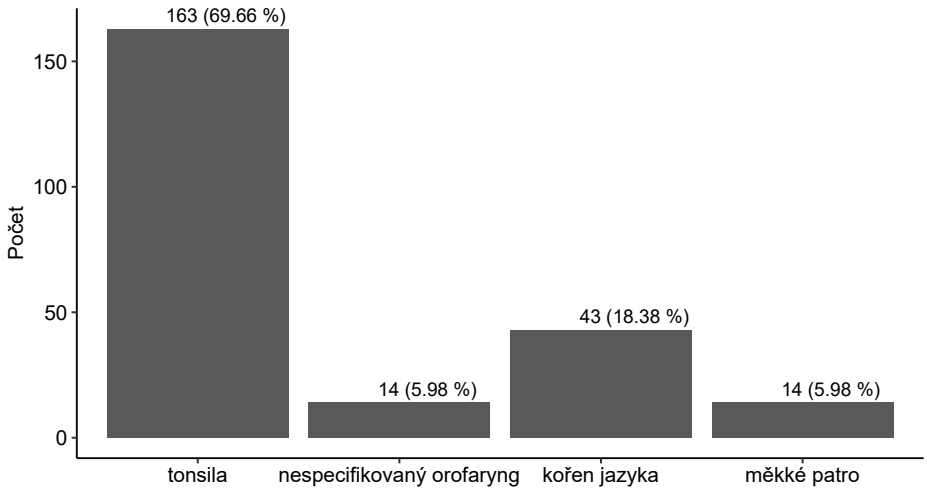


Obrázek 3.6: Četnost a procentuální rozdělení pacientů dle míry konzumování alkoholu.

Dále si přiblížíme rizikové faktory spojené se samotnou nemocí, přičemž všechny klasifikace nádoru se řídí standardy systému TNM klasifikace [8].

- **přesnější lokace**

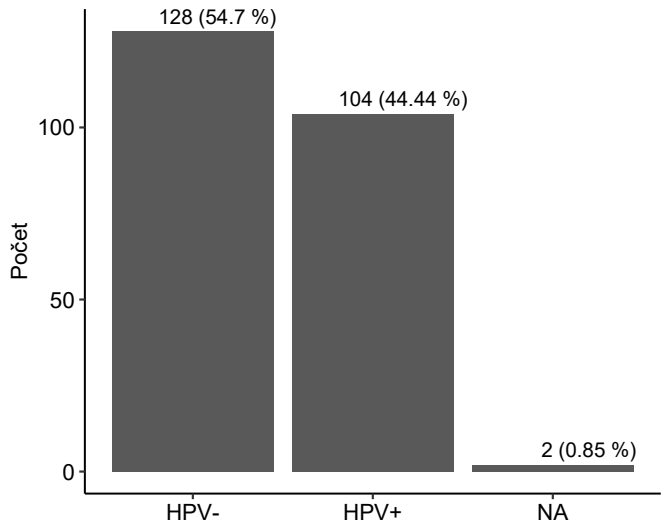
  Určuje přesnější lokaci nádoru v oblasti orofaryngu - tonsila, kořen jazyka, měkké patro. Jejich rozdělení je uvedeno na obrázku 3.7.



Obrázek 3.7: Četnost a procentuální zastoupení jednotlivých kategorií přesnější lokace v datovém souboru.

- **HPV (Human papilloma virus) pozitivita**

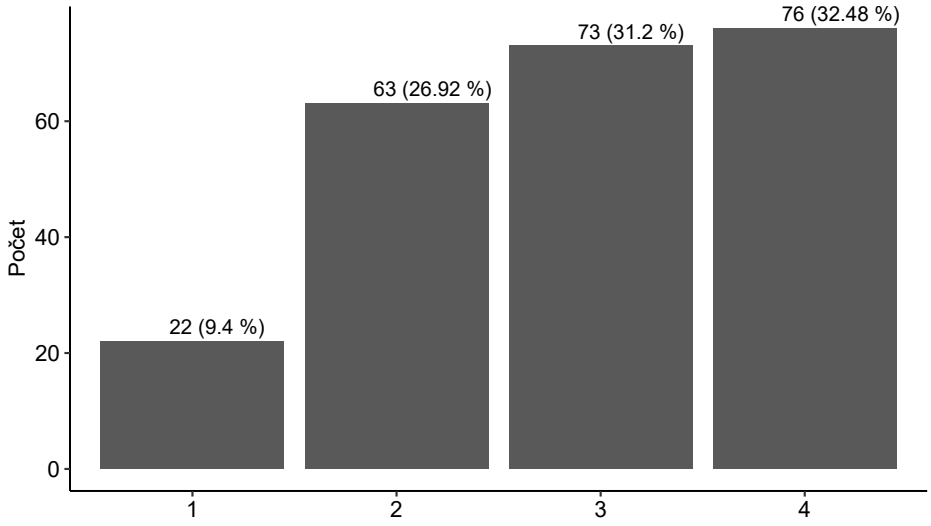
  Původ tumoru je označen jako HPV pozitivní, pokud má nenulový výsledek PCR metody (proměnná ozn. *SPF*) a zároveň je u pacienta prokázána přítomnost proteinu p16 (proměnná ozn. *p16*). Zastoupení těchto nádorů je zachyceno na obrázku 3.8. Efekt HPV pozitivity je dlouho zkoumaným rizikovým faktorem pro vznik nádorů v oblasti hlavy a krku stejně jako vliv této skutečnosti při odpovědi na léčbu. Několik studií již prokázalo, že u pacientů s HPV pozitivním původem nádoru bez dalších rizikových faktorů je lepší odpověď na léčbu i delší celkové přežití. [11] Z tohoto důvodu také očekáváme, že tento faktor může hrát významnou roli při našich dalších analýzách.



Obrázek 3.8: Četnost a procentuální rozdělení HPV pozitivního původu nádoru v datové sadě.

- **velikost nádoru**

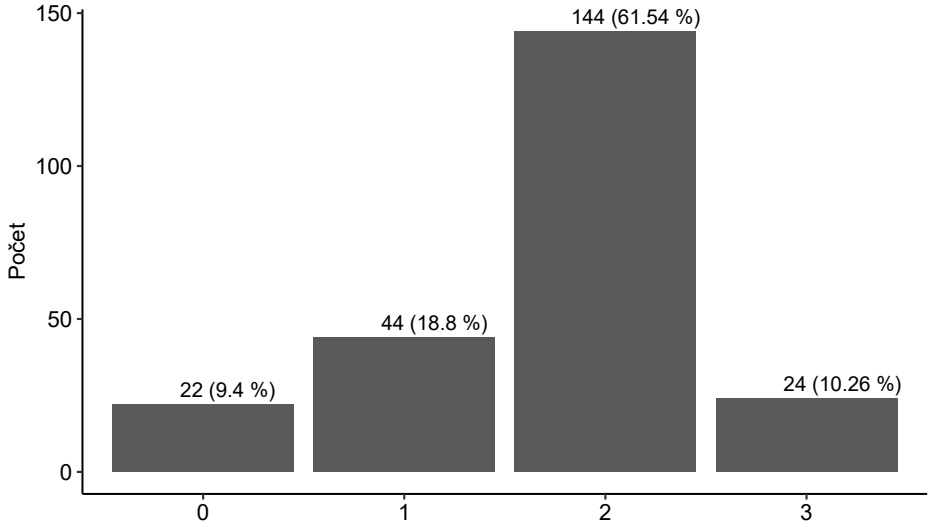
  Klasifikovaná do jednotlivých skupin 1 až 4 dle rostoucí velikosti, jejichž zastoupení je zachyceno na obrázku 3.9.



Obrázek 3.9: Četnost a procentuální rozdělení velikosti nádoru v datové sadě.

- **uzliny**

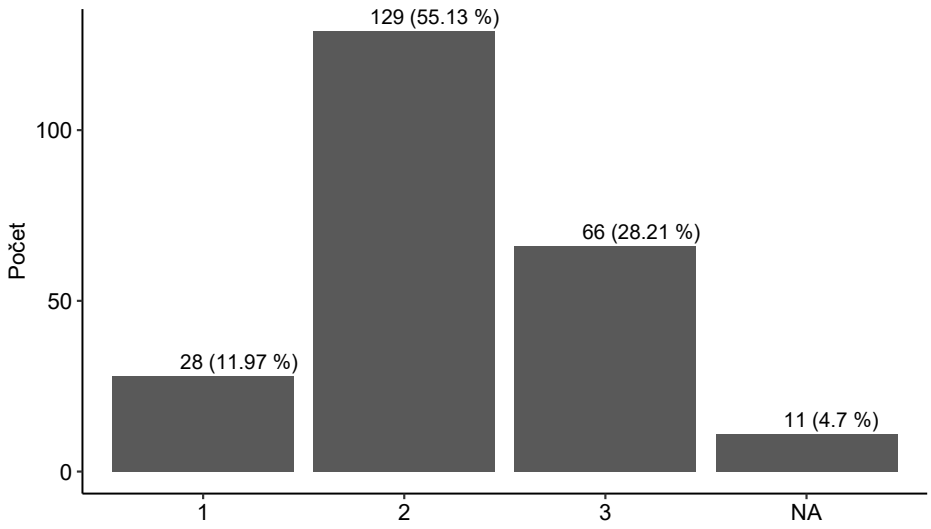
  Klasifikované do jednotlivých skupin 0 až 3 dle přítomnosti a rozsahu metastáz v regionálních mízních uzlinách. Zastoupení těchto skupin je zachyceno na obrázku 3.10.



Obrázek 3.10: Četnost a procentuální rozdělení zasažení uzlin metastázemi nádoru v datové sadě.

- **grading nádoru**

  Je definovaný pomocí tzv. stupně diferencovanosti buněk. Obvykle platí, že čím vyšší grading (tj. histopatologický stupeň), tím jsou nádory agresivnější, ale také citlivější k léčbě.[9] Zastoupení jednotlivých stupňů v datové sadě je zachyceno na obrázku 3.11.

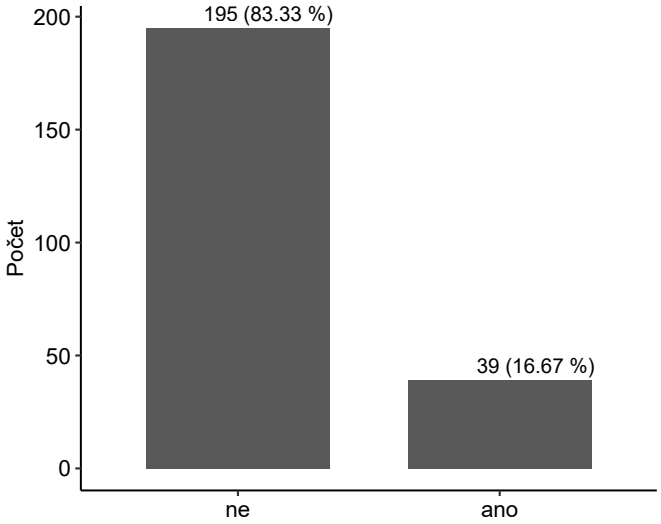


Obrázek 3.11: Četnost a procentuální rozdělení gradingu nádoru v datové sadě.

- **recidiva**

  Vypovídá o návratu onemocnění, které po předchozí léčbě nebylo prokazatelné.[9]

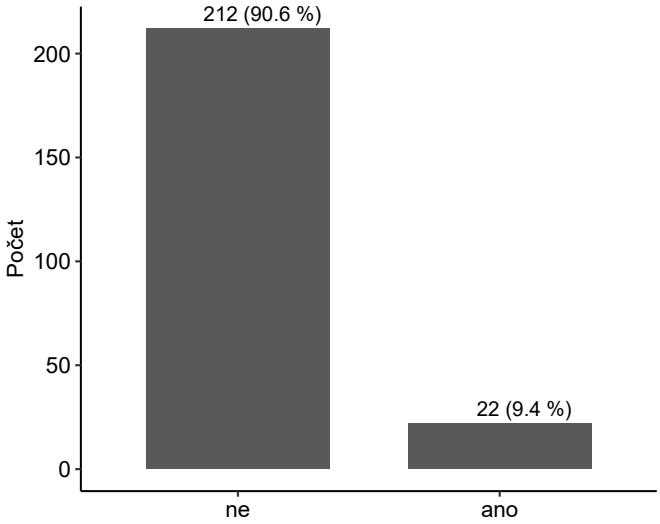
  Zastoupení pacientů s recidivou je zachyceno na obrázku 3.12.



Obrázek 3.12: Četnost a procentuální rozdělení pacientů s recidivou nádoru v datové sadě.

- **duplicita**

  Značí výskyt dalšího, nezávislého nádoru u pacienta, nejedná se tedy o metastáze primárního nádoru. [9] Zastoupení pacientů s duplicitním nádorem je zachyceno na obrázku 3.13.



Obrázek 3.13: Četnost a procentuální rozdělení pacientů s duplicitním nádorem v datové sadě.

- **léčebná terapie**

  Posledním uvažovaným faktorem je typ zvolené léčebné terapie. Celkem se vyskytují čtyři její varianty, které budeme označovat:

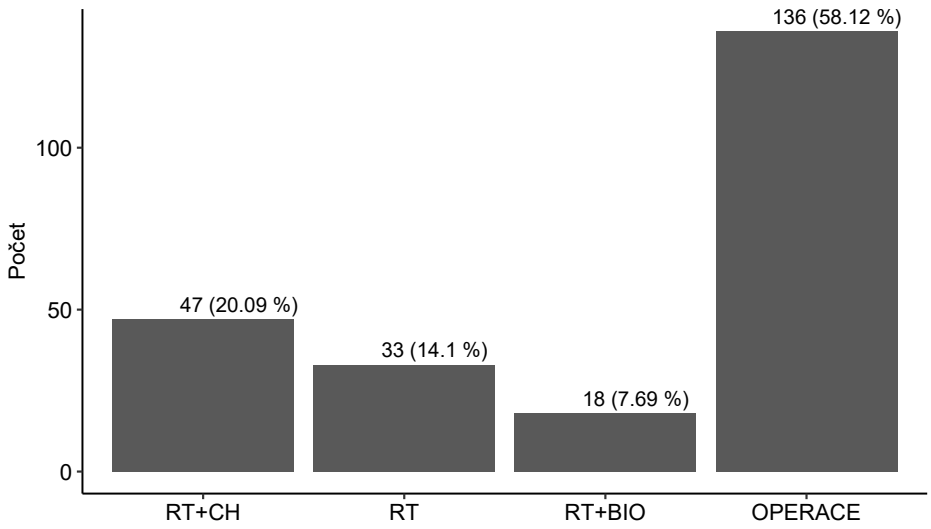
  RT+CH... chemoterapie v kombinaci s ozařováním

  RT... pouze ozařování

  RT+BIO... ozařování v kombinaci s biologickou léčbou

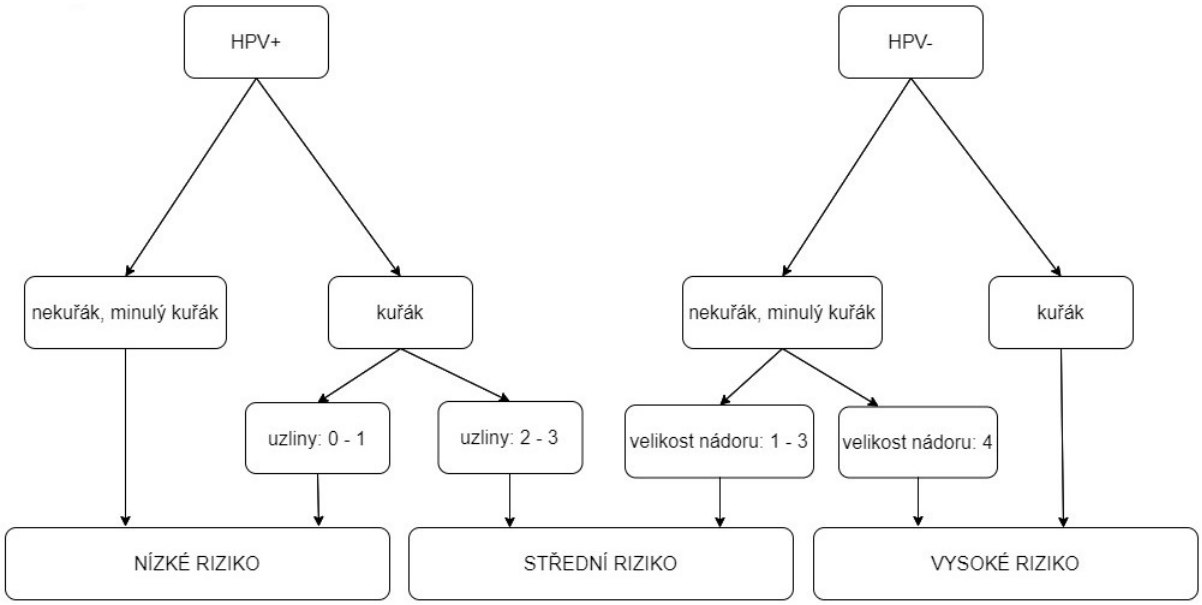
  OPERACE... operace

  Jejich zastoupení v datové sadě je zachyceno na obrázku 3.14. Chirurgický zákrok je nejčastější primární léčebnou strategií, není-li zvolen, je tomu často z důvodů špatného celkového zdravotního stavu pacienta či pokročilého stádia onemocnění, které jsou také spojeny se špatnou prognózou přežití. [10]

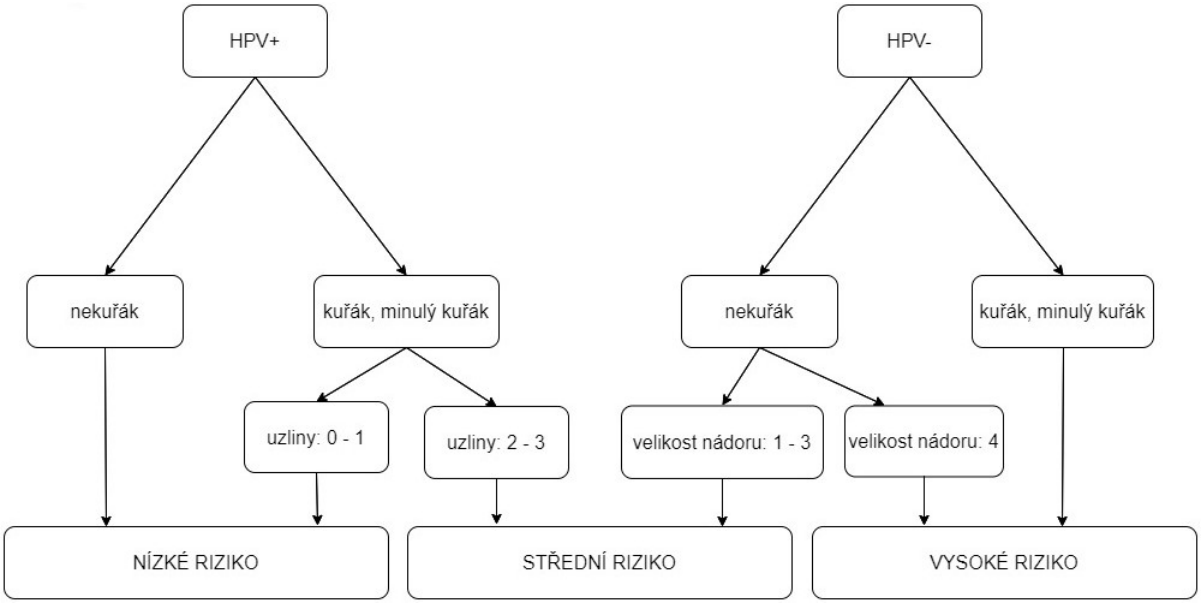


Obrázek 3.14: Četnost a procentuální rozdělení pacientů dle léčebné terapie v datové sadě.

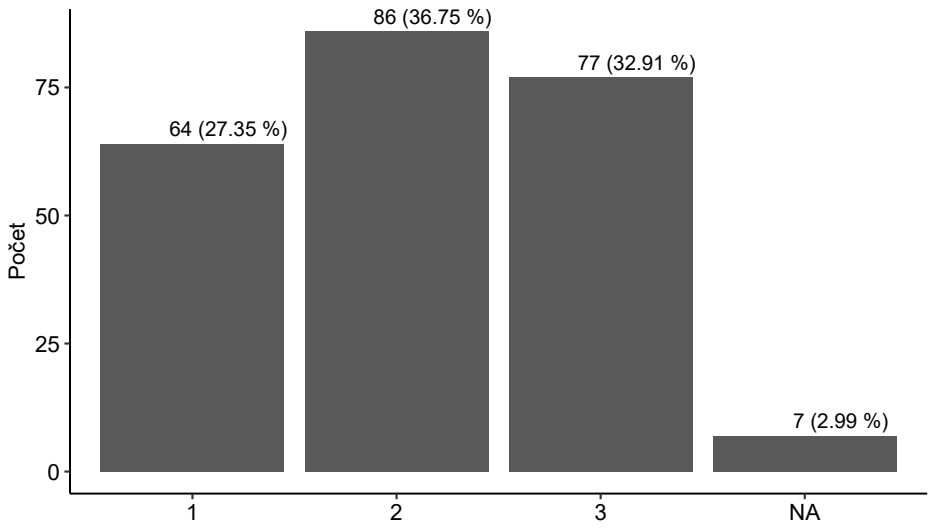
V naší analýze máme možnost jednak pracovat s těmito faktory přímo či dle navrhovaných agregovaných proměnných rizikových faktorů HPV pozitivity, kouření, velikosti nádoru a uzlin, které jsou v datech označeny *risk.group1, risk.group2*. Tyto proměnné stanovují pro každého pacienta míru rizika na základě hodnot daných rizikových faktorů řadících je do kategorií nízkého (ozn. 1), středního (ozn. 2) a vysokého rizika (ozn. 3) dle schémat zobrazených na obrázcích 3.15, 3.16, jejichž návrh byl součástí dodané datové sady. Četnost zastoupení v těchto rizikových skupinách je zobrazena na obrázku 3.17, resp. 3.18. Ze schématů 3.15, 3.16 vyplývá, že rozdíl mezi přiřazením dané míry rizika zachycených v těchto dvou proměnných spočívá v odlišném chápání rizika bývalých kuřáků. První klasifikace pomocí *risk.group1* jim přiřazuje stejné riziko jako pro nekuřáky, naopak v druhé klasifikaci *risk.group2* je jejich riziko přiřazeno stejné jako pro kuřáky.



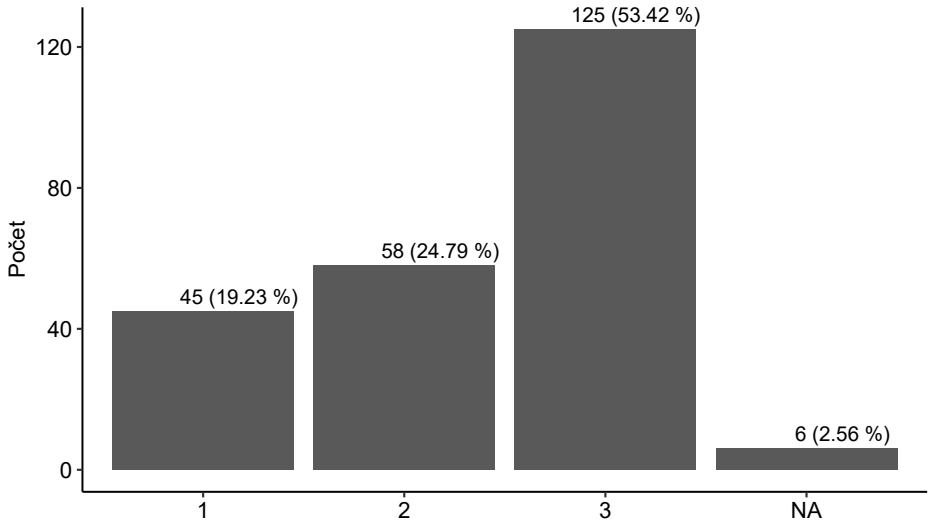
Obrázek 3.15: Schéma přiřazení míry rizika dle *risk.group1*.



Obrázek 3.16: Schéma přiřazení míry rizika dle *risk.group2*.



Obrázek 3.17: Četnost a procentuální rozdělení pacientů dle přiřazené míry rizika *risk.group1*.



Obrázek 3.18: Četnost a procentuální rozdělení pacientů dle přiřazené míry rizika *risk.group2*.

3.2 Analýza

V této části se již budeme přímo věnovat analýze doby přežití v závislosti na jednotlivých rizikových faktorech pomocí neparametrických Kaplanových-Meierových odhadů funkcí přežití a semi-parametrického Coxova modelu proporcionálních rizik. Pro účely využití testů založených na porovnávání věrohodnosti budeme v této části uvažovat soubor pouze 215 pacientů bez chybějících údajů. Budeme tak pracovat se stále stejným datovým souborem nehledě na aktuálně uvažované rizikové faktory a věrohodnosti jednotlivých modelů tak budou porovnatelné.

3.2.1 Kaplanův-Meierův odhad funkce přežití, medián přežití

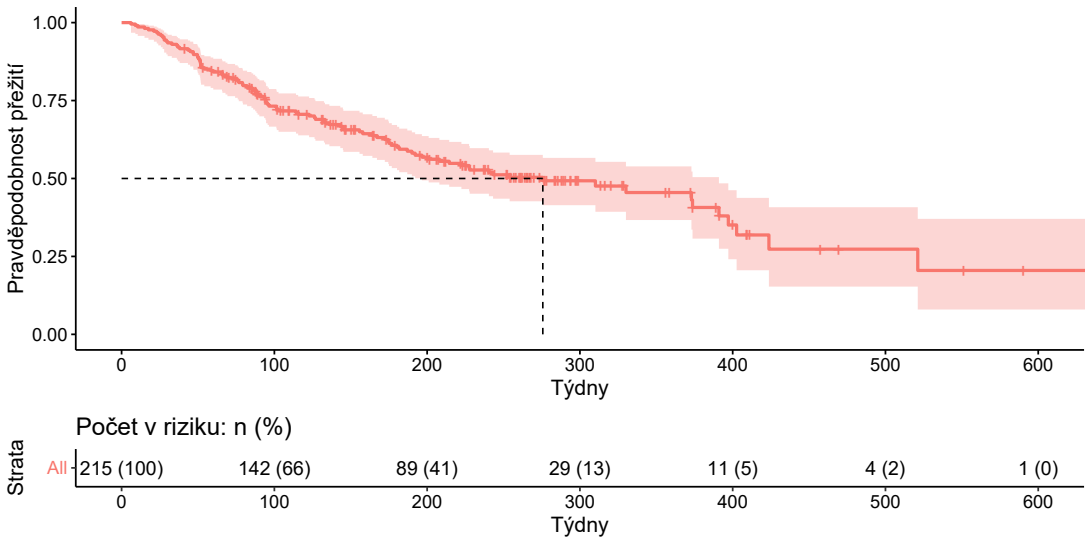
Zde využijeme Kaplanův-Meierův odhad funkce přežití zkonstruovaný nejprve pro všechna pozorování, ze kterého také odhadneme medián celkového přežití na celém souboru pozorování bez ohledu na jakékoliv rizikové faktory. Následně zkonstruujeme odhady funkcí přežití pro skupiny dat definovaných dle kategorií jednotlivých rizikových faktorů – poslouží nám především jako grafický nástroj porovnání rozdílů jejich křivek přežití. Statistickou významnost rozdílů takto definovaných křivek přežití také otestujeme pomocí tzv. log-rank testu. Tyto získané poznatky využijeme i později při tvorbě Coxova modelu proporcionálních rizik.

Celkový Kaplanův-Meierův odhad funkce přežití

Celkové přežití pacientů v této studii je zachyneno na obrázku 3.19 společně s vyznačeným mediánem přežití: 276 týdnů (95% CI: 197 − 391). Můžeme si všimnout, že odhad funkce přežití neklesá až k nule, což odráží fakt, že pacienti s nejdelší dobou sledování jsou cenzorovanými pozorováními, tj. u nich k úmrtí nedošlo.

Kaplanovy-Meierovy odhady funkcí přežití pro jednotlivé rizikové faktory

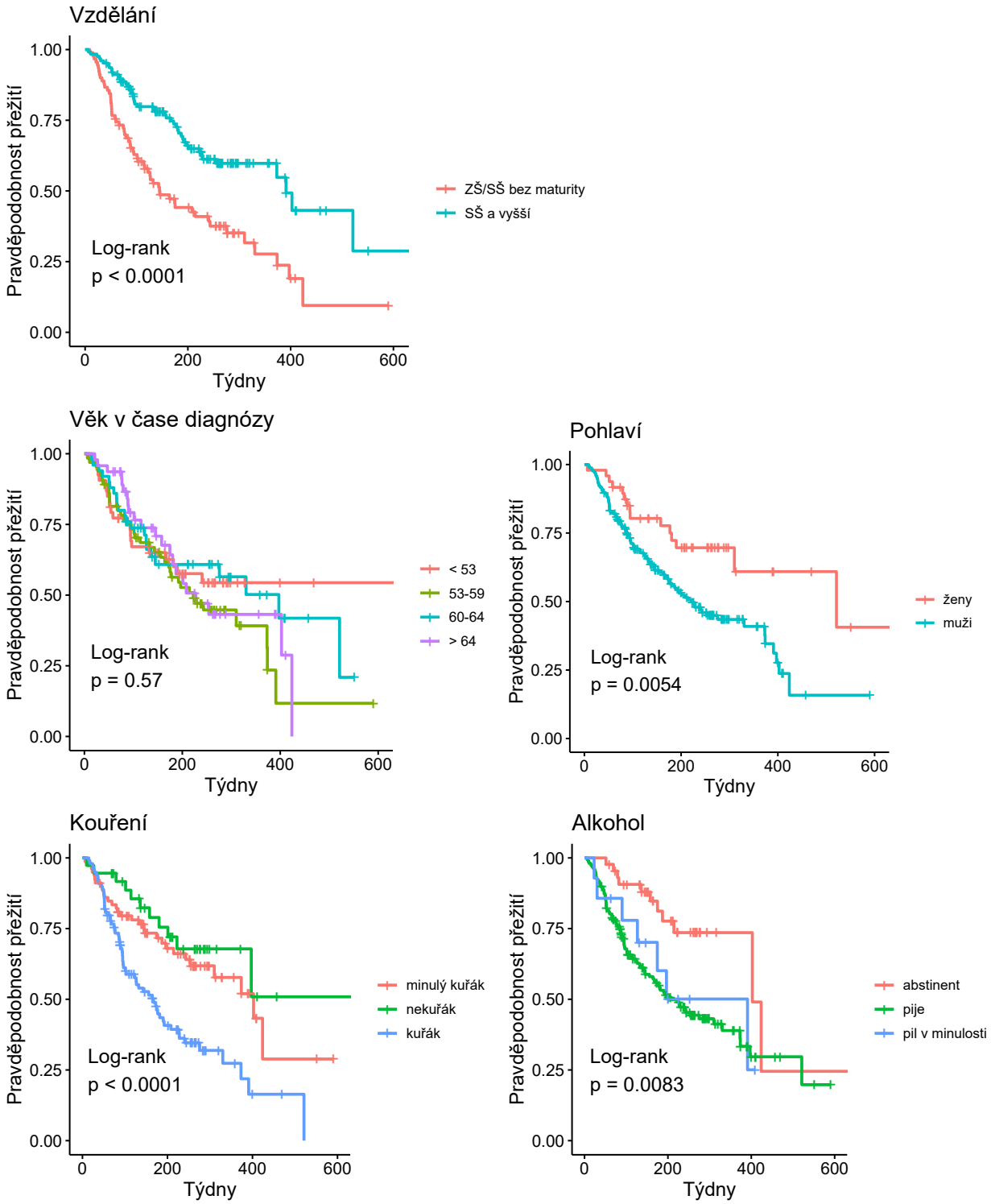
Vliv uvažovaných rizikových faktorů na funkci přežití můžeme pomocí Kaplanova-Meierova neparametrického odhadu zjistit tak, že rozdělíme data do skupin dle jednotlivých kategorií daného rizikového faktoru, pro které poté zvlášť provedeme



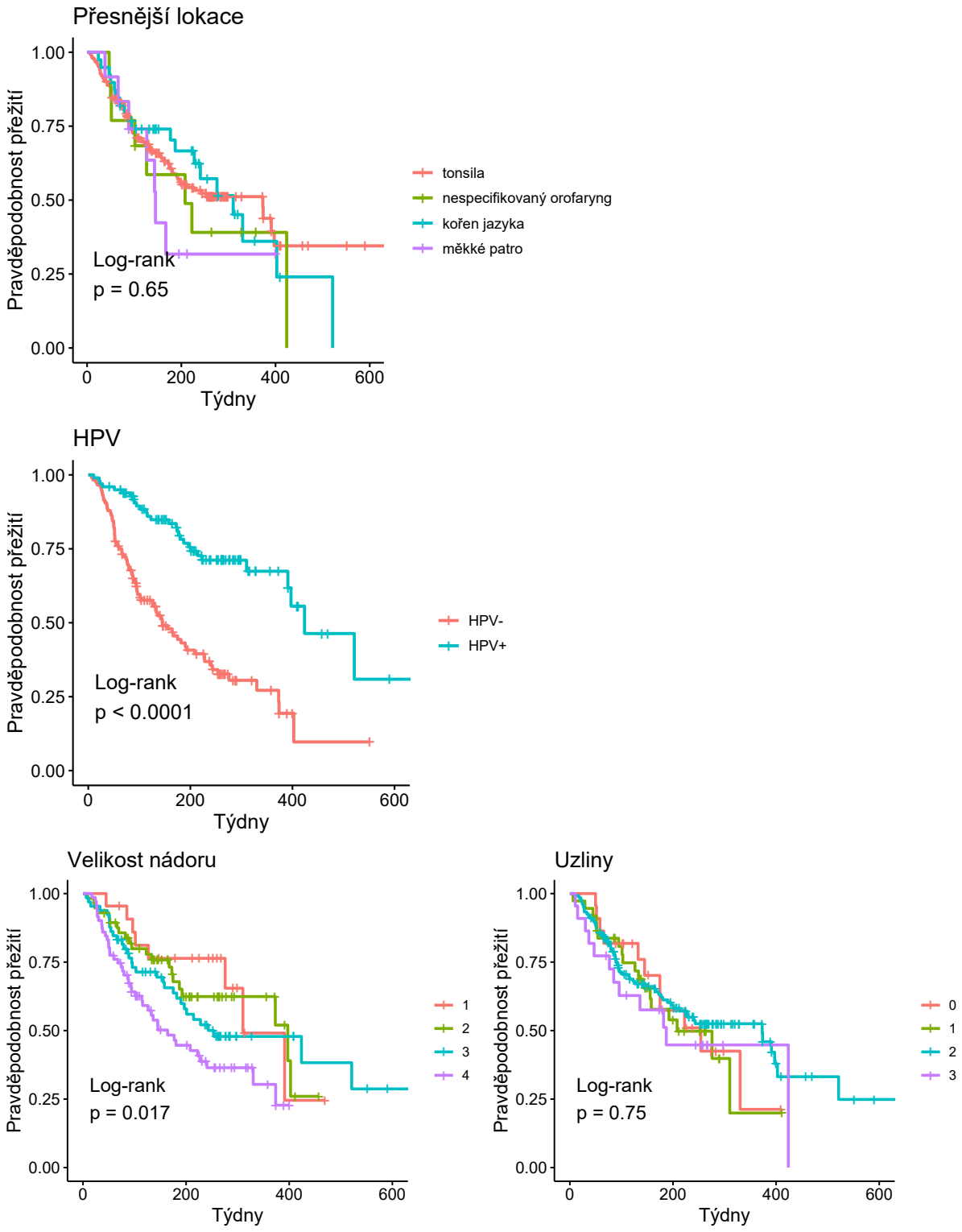
Obrázek 3.19: Kaplanův-Meierův odhad funkce přežití s 95% bodovým konfidenčním intervalem (získaný pomocí *log-log transformace*), vyznačeným mediánem přežití a tabulkou vyjadřující počet, resp. procento, pacientů v riziku v daných časových okamžicích.

Kaplanovy-Meierovy odhady funkcí přežití a vykreslíme si je společně do grafu. Například pro faktor léčba tak získáme čtyři různé odhady funkcí přežití zvlášť pro pacienty, kteří podstoupili ozařování, ozařování v kombinaci s biologickou léčbou nebo ozařování v kombinaci s chemoterapií a pro pouze operované pacienty. Pro spojité rizikové faktory si nejprve kategorie musíme vytvořit – v našem případě se jedná pouze o rizikový faktor věku v čase diagnózy, který jsme rozdělili do skupin dle kvantilů: dolního kvartilu, mediánu a horního kvartilu.

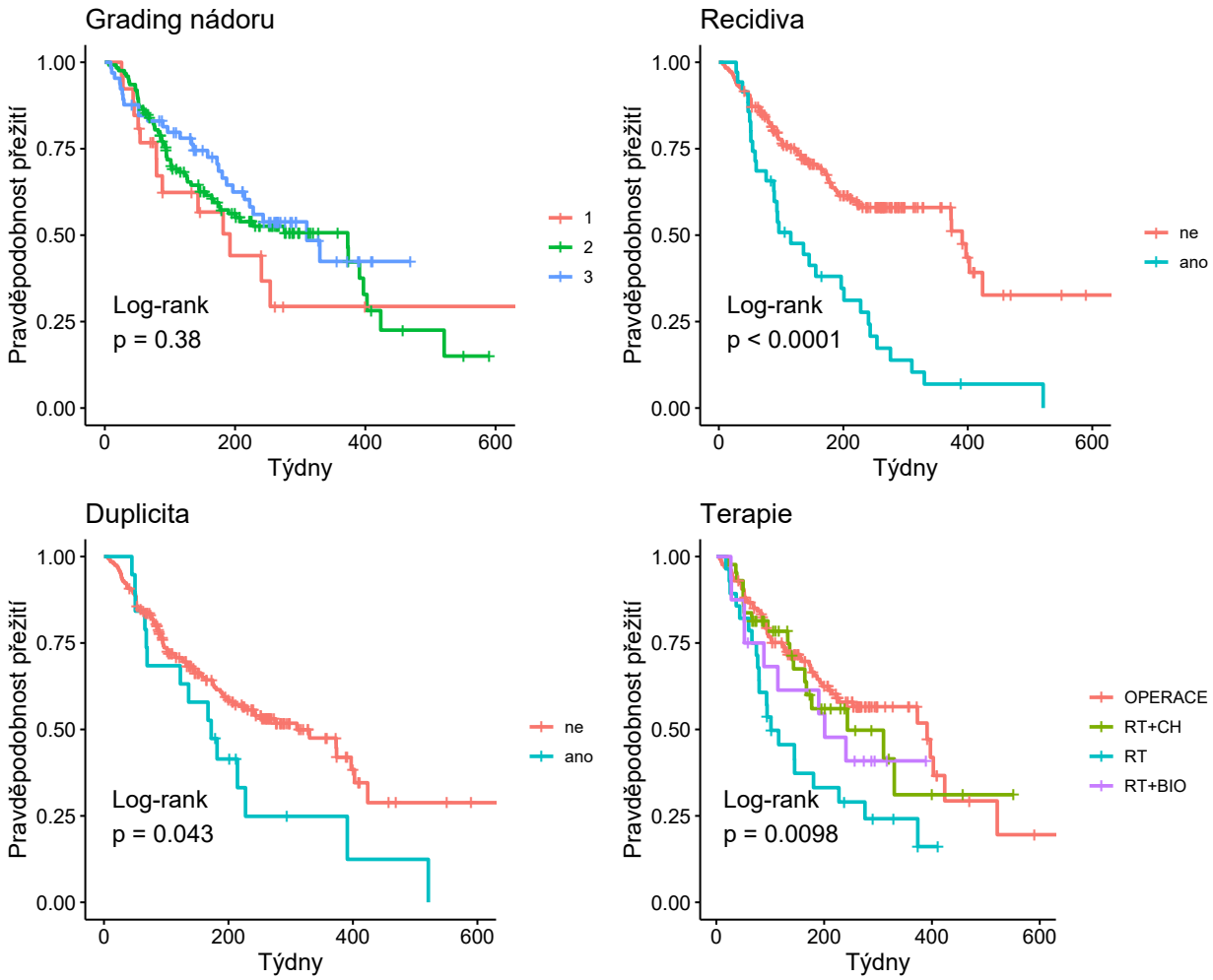
Vykreslené Kaplanovy-Meierovy odhady funkcí přežití jsme opět seskupili dle členění rizikových faktorů v popisné části – nejprve jsou tedy zobrazeny rizikové faktory nevztahující se přímo k onemocnění na obrázku 3.20, na obrázcích 3.21, 3.22 poté faktory týkající se přímo nádoru a léčby. Pro každý rizikový faktor jsme také provedli tzv. log-rank test významnosti rozdílů funkcí přežití, jehož výsledná p-hodnota je také zobrazena v příslušných grafech.



Obrázek 3.20: Kaplanovy-Meierovy odhady funkcí přežití pro rizikové faktory nevztahující se přímo k nádoru a léčbě spolu s p-hodnotou log-rank testu.



Obrázek 3.21: Kaplanovy-Meierovy odhady funkcí přežití pro jednotlivé rizikové faktory vztahující se k samotnému nádoru a léčbě spolu s p-hodnotou log-rank testu. *(1. část)*

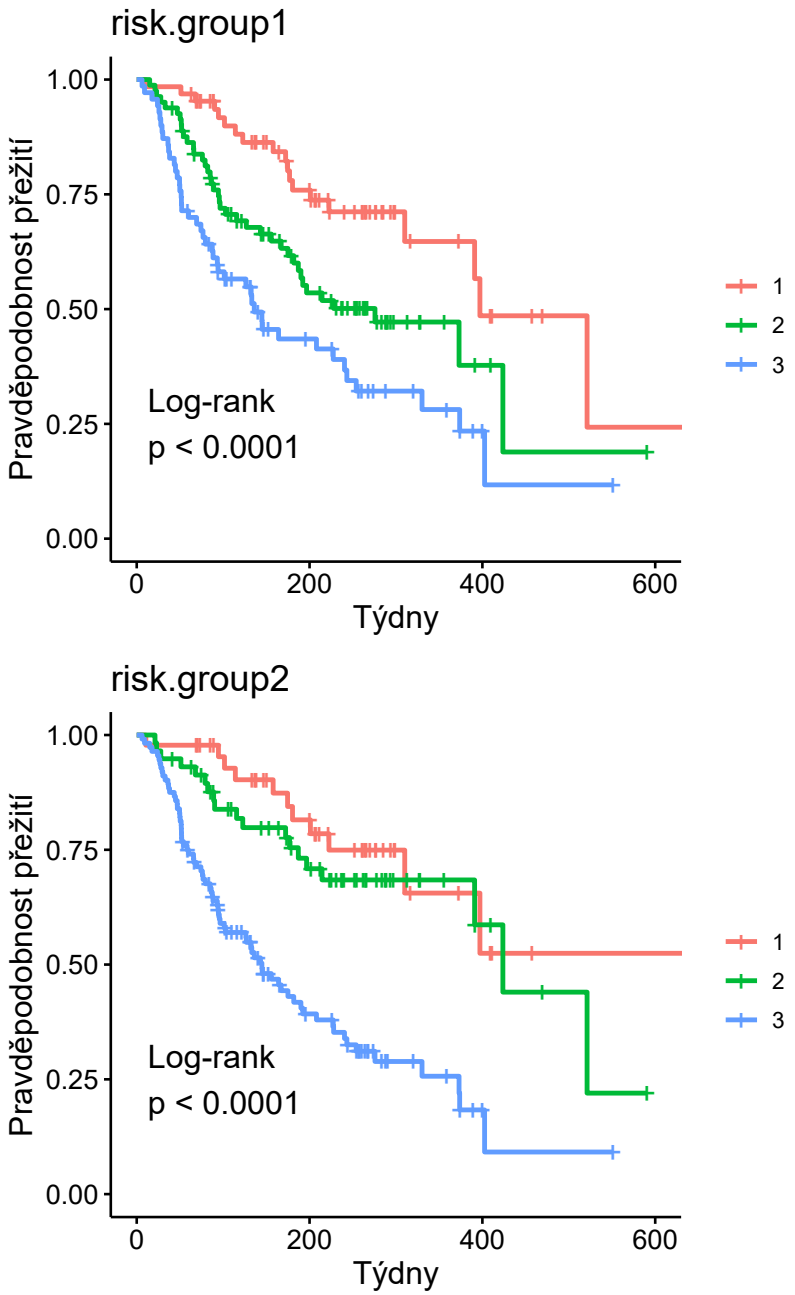


Obrázek 3.22: Kaplanovy-Meierovy odhady funkcí přežití pro jednotlivé rizikové faktory vztahující se k samotnému nádoru a léčbě spolu s p-hodnotou log-rank testu. *(2. část)*

Omezení Kaplanova-Meierova odhadu funkce přežití

Využití tohoto neparametrického přístupu ke stanovení odhadu funkce přežití je limitováno v situaci, kdy chceme vzít v úvahu hned vícero rizikových faktorů ovlivňujících přežití. Kombinace jejich jednotlivých kategorií nám zvyšuje počet skupin, pro které chceme jednotlivé křivky odhadnout, a snižují se nám tak počty pozorování, ze kterých odhad stanovujeme, což zvyšuje i míru nejistoty těchto odhadů. Zároveň také roste komplikovanost pro interpretaci získaných odhadů funkcí přežití.

V našem případě však můžeme ještě využít klasifikace v *risk.group1*, *risk.group2*, které vznikly agregací některých rizikových faktorů dle algoritmu popsaného ve schématech 3.15, 3.16 a vyjadřují míru rizikovosti pro daného pacienta. Výsledné Kaplanovy-Meierovy odhady funkcí přežití pro jednotlivé kategorie rizika opět spolu s p-hodnotou log-rank testu významnosti jejich rozdílů jsou vykresleny na obrázku 3.23. Dle tohoto výsledku můžeme říci, že je významný rozdíl v přežití mezi jednotlivými skupinami a dle očekávání nejdéle přežívají pacienti s nejnižší mírou rizika. Jak můžeme vidět, při zahrnutí bývalých kuřáku mezi nekuřáky v *risk.group1* klasifikaci je rozdíl mezi přežitím v daných kategoriích jasně vymezen, odhadnuté funkce přežití se ani nekříží. Naopak však při pohlížení na bývalé kuřáky stejně jako na kuřáky již jasný rozdíl ve funkci přežití mezi rizikovou skupinou 1 a 2 není, avšak rozdíl v přežití oproti nejrizikovější skupině 3 zůstává.



Obrázek 3.23: Kaplanovy-Meierovy odhady funkcí přežití pro stanovené kategorie rizika *risk.group1, risk.group2* s p-hodnotou log-rank testu.

3.2.2 Coxův model proporcionálních rizik

Tento model nám umožní zahrnout vliv různých rizikových faktorů na dobu přežití, klade však také předpoklady, které je nutné ověřit před interpretací výsledků modelu - především dodržení proporcionality rizik.

Sestavení modelu

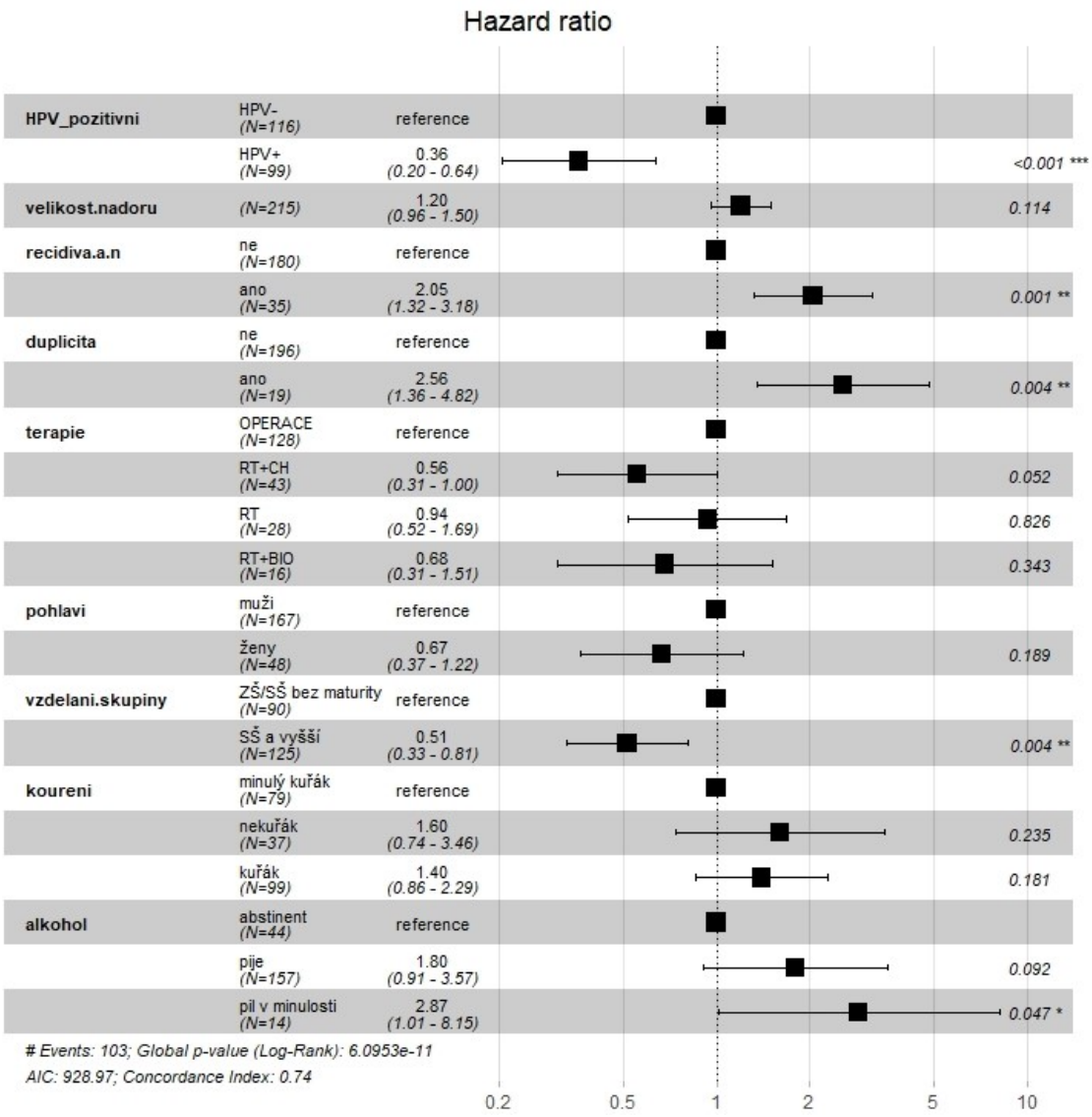
Zaměříme se na tvorbu modelu, který bude obsahovat všechny rizikové faktory, jež budou mít statisticky významný vliv na riziko úmrtí v libovolný časový okamžik a to za předpokladu, že se daného časového okamžiku pacient dožije.

Při tvorbě modelu budeme postupovat v souladu s kroky popsanými v knize Hosmer, Lemeshow, May [3]. Pro základní model, který budeme dále upravovat, jsme tedy zvolili všechny rizikové faktory, které dle log-rank testu z Kaplanových-Meierových odhadů funkcí přežití byly signifikantní na úrovni 25 %. Odhady parametrů tohoto modelu ve formě poměrů rizika, tj. e^{β_j}, jsou znázorněné spolu s 95% konfidenčními intervaly a p-hodnotou Waldova testu významnosti jednotlivých parametrů na obrázku 3.24.

Můžeme pozorovat, že ne všechny rizikové faktory jsou v tomto modelu statisticky významné, proto jsme jej dále redukovali. Vhodnost odebrání nesignifikantních proměnných byla ověřena pomocí likelihood-ratio testu a také porovnáním změny v odhadech parametrů – pokud by změna byla velká (udává se více než 20 %, dle [3]), odebraná proměnná je potencionální confounder a musí být v modelu zachována. Tento případ však v našem postupu ani jednou nenastal.

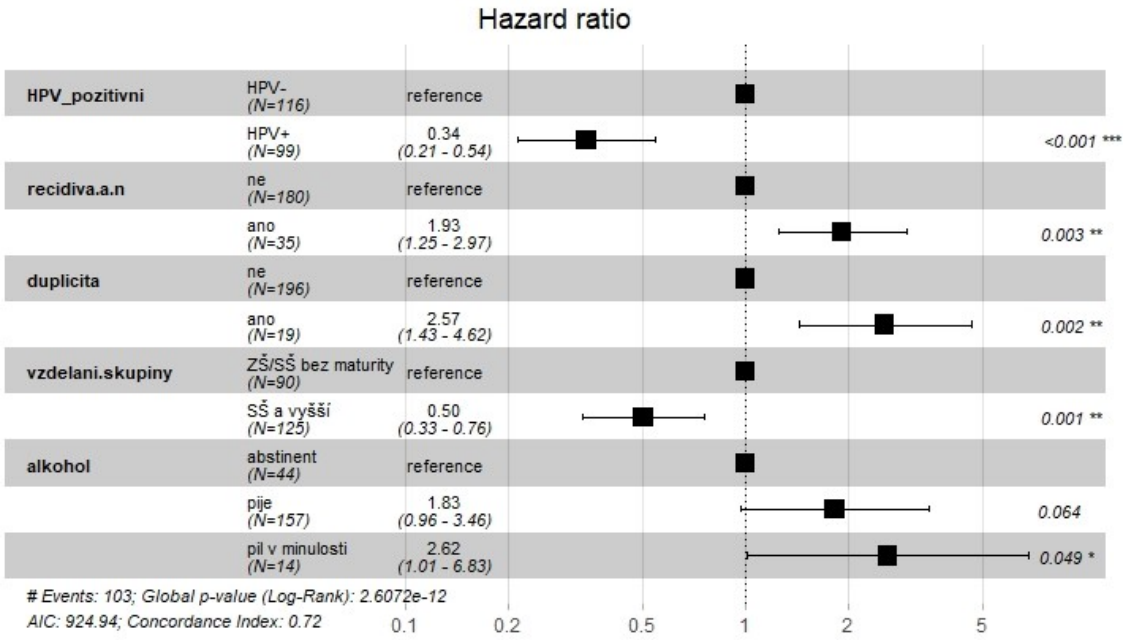
Odhady parametrů výsledného redukovaného modelu, opět ve formě poměrů rizik, jsou znázorněné spolu s 95% konfidenčními intervaly a p-hodnotou Waldova testu významnosti jednotlivých parametrů na obrázku 3.25.

Abychom tento redukovaný model mohli považovat za finální při splnění předpokladů modelu, je třeba ještě ověřit signifikantnost a možný confounding proměnných, které nebyly zahrnuty do základního modelu, a to postupným přidáním jednotlivých proměnných do redukovaného modelu - ani jedna proměnná se však neukázala být jako signifikantní či confounder. Na závěr jsme ještě ověřili možné



Obrázek 3.24: Souhrnný výstup pro základní Coxův model proporcionálních rizik.

interakce v modelu, ani jedna však také nebyla statisticky významná. Můžeme proto model, jehož výsledky jsou zobrazeny na obrázku 3.25, považovat za finální, pokud bude splňovat předpoklady.

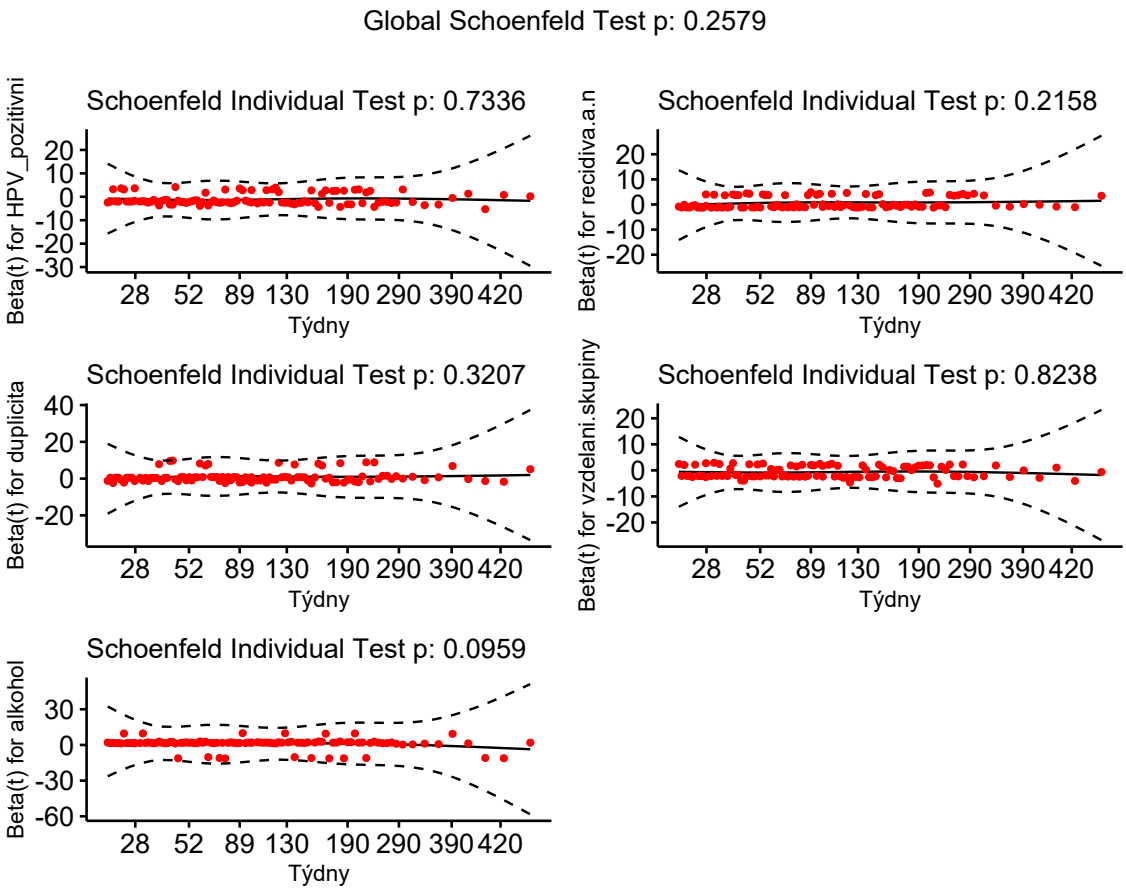


Obrázek 3.25: Souhrnný výstup pro redukovaný Coxův model proporcionálních rizik.

Ověření předpokladů modelu

Ověření předpokladů Coxova modelu proporcionálních rizik je nutným krokem před interpretováním dosažených výsledků, protože při porušení předpokladů nelze již tyto výsledky považovat za správné.

Protože náš finální model neobsahuje žádné numerické vysvětlující proměnné, nemusíme ověřovat dodržení předpokladu jejich lineárního vztahu vzhledem k vysvětlovanému logaritmu hazardu. Musíme tedy pouze ověřit nejvýznamnější předpoklad – proporcionalitu rizik. Jak můžeme vidět na obrázku 3.26, celkový test porušení tohoto předpokladu za pomoci Schoenfeldových reziduí hypotézu proporcionality rizik nezamítá, stejně tak jako ani individuální testy pro jednotlivé proměnné. Tento předpoklad je tedy splněn a můžeme získaný model 3.25 považovat za finální a jeho výsledky dále interpretovat.

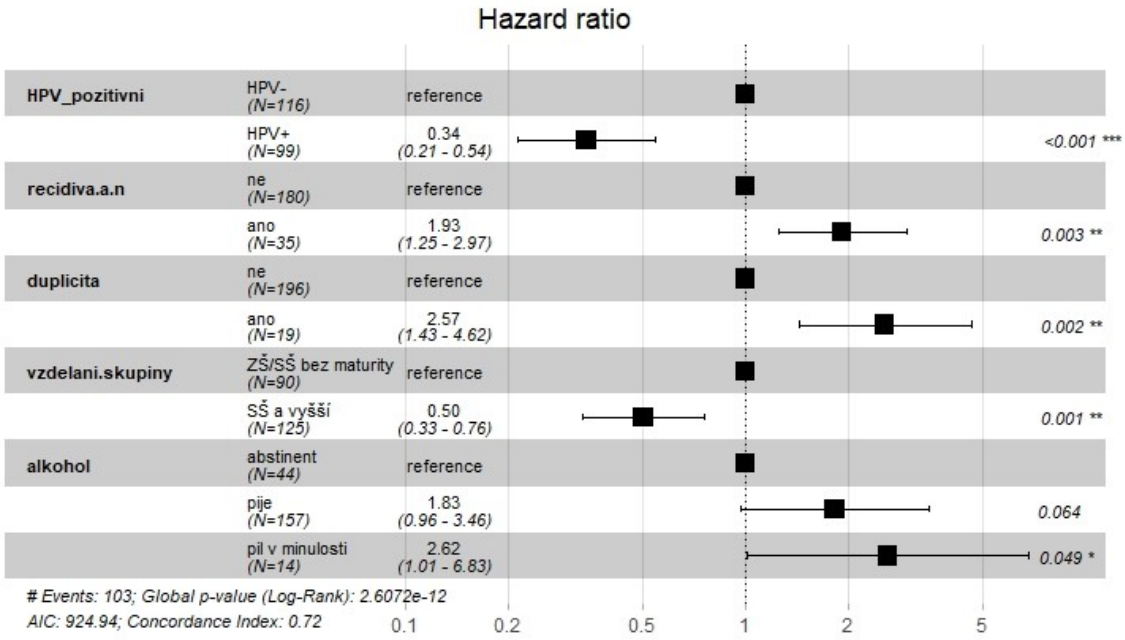


Obrázek 3.26: Schoenfeldova rezidua spolu s p-hodnotami testů proporcionality rizik pro jednotlivé rizikové faktory a p-hodnotou celkového testu dodržení proporcionality rizik s využitím těchto reziduí.

Interpretace výsledků

Podívejme se tedy ještě jednou na to, co nám obrázek 3.27 říká o vlivu statisticky významných rizikových faktorů na riziko úmrtí v libovolném časovém okamžiku za předpokladu, že se pacient tohoto času dožil.

Můžeme říci, že takové riziko je pro HPV pozitivní pacienty v průměru okolo 65 % nižší než pro pacienty s nevirovým původem nádoru. Jestliže jsme dosáhli vyššího stupně vzdělání, bude riziko úmrtí v daném okamžiku nižší asi o 50 %, přičemž faktor vzdělání je také možno chápat jako indikátor lepšího socioekonomického postavení. Přítomnost recidivy zvýší toto riziko skoro 2krát, duplicitního nádoru více než 2,5krát. Pro aktuální konzumenty alkoholu je riziko úmrtí v li-



Obrázek 3.27: Souhrnný výstup pro finální Coxův model proporcionálních rizik.

bovolném časovém okamžiku, za předpokladu dožití se tohoto času, skoro 2krát vyšší než pro abstinenty a pro bývalé konzumenty alkoholu je v porovnání s abstinenty více než 2,5krát vyšší. Je však nutno vzít v potaz menší počet pacientů zastoupených v této skupině, který také vede k velké variabilitě tohoto odhadu.

Kapitola 4

Lineární smíšené modely

V praxi se často setkáváme se situacemi, kdy dochází k porušení předpokladů nezávislosti jednotlivých pozorování. Příkladem mohou být opakovaná měření daného ukazatele na stejných subjektech při různých podmínkách (např. vliv podání léku při různé koncentraci účinné látky), skupinově uspořádaná data (např. testování znalostí žáků, kteří jsou náhodně vybráni z různých škol) nebo longitudinální datová struktura, tj. sledování daného ukazatele na stejných subjektech opakovaně v čase při stejných podmínkách (např. pozorování velikosti aneurysma v čase). Pro takovouto datovou strukturu nelze využít klasických metod lineárních modelů, která ignorují vzájemnou korelovanost některých pozorování. Vhodnou metodou pro zpracování mohou být *lineární smíšené modely*, které jsou kombinací tzv. fixních a náhodných efektů. Fixní efekty zachycují populační (ozn. *population-specific*) regresní parametry, zatímco náhodné efekty jsou parametry individuální (ozn. *subject-specific*). Dále budeme pohlížet na teorii lineárních smíšených modelů z pohledu problematiky longitudinální datové struktury. Uváděná teorie, vycházející převážně z [23], částečně z [24], je však také obecně platná pro širší využití jiných datových struktur s korelovanými daty.

Obecně pro i-tý subjekt, $i = 1, \ldots, N$, s celkem n_i pozorováními lze lineární smíšený model zapsat ve tvaru

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{4.1}$$

kde $\mathbf{X}_i$ je matice fixních vysvětlujících proměnných velikosti $n_i \times p$, $\mathbf{Z}_i$ je matice

vysvětlujících proměnných individuálního vývoje o velikosti $n_i \times q$, parametr $\boldsymbol{\beta}$ je p-rozměrný vektor fixních efektů, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ je q-rozměrný vektor náhodných efektů a $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$ n_i-rozměrný vektor reziduální složky, přičemž $\mathbf{b}_i$ a $\boldsymbol{\varepsilon}_i$ jsou nezávislé. Takto formulovaný lineární smíšený model je označován jako hierarchický. Velmi častým předpokladem kovarianční struktury reziduální složky je homoskedasticita, tj. $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$. Z tohoto předpokladu vyplývá, že při daném $\mathbf{b_i}$ a $\boldsymbol{\beta}$ jsou pozorování na i-tém subjektu nezávislé veličiny, což může být často nerealistický předpoklad zvláště pro jednodušší struktury náhodných efektů. Jestliže prvky $\boldsymbol{\varepsilon}_i$ mají konstantní rozptyl, můžeme předpokládat, že reziduální složka $\boldsymbol{\varepsilon}_i$ lze rozložit na část chyby měření $\boldsymbol{\varepsilon}_{(1)i} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}\right)$ a část vyjadřující sériovou korelaci $\boldsymbol{\varepsilon}_{(2)i} \sim N\left(\mathbf{0}, \tau^2 \mathbf{H}_i\right)$, kde $\mathbf{H}_i$ je předpokládaná korelační matice rozměrů $n_i \times n_i$ určité struktury sériové korelace, tj. $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}$. Složka $\boldsymbol{\varepsilon}_{(2)i}$ odpovídá předpokladu, že alespoň část pozorovaného individuálního profilu lze popsat časově proměnlivým stochastickým procesem v rámci daného subjektu. Často jsou to procesy známé z problematiky časových řad (AR, ARMA). Výsledný lineární smíšený model lze zapsat ve tvaru

$$\begin{cases} \mathbf{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i} \\ \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \\ \boldsymbol{\varepsilon}_{(1)i} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}\right), \\ \boldsymbol{\varepsilon}_{(2)i} \sim N\left(\mathbf{0}, \tau^2 \mathbf{H}_i\right), \\ \mathbf{b}_1, \ldots, \mathbf{b}_N, \boldsymbol{\varepsilon}_{(1)1}, \ldots, \boldsymbol{\varepsilon}_{(1)N}, \boldsymbol{\varepsilon}_{(2)1}, \ldots, \boldsymbol{\varepsilon}_{(2)N} \text{ nezávislé.} \end{cases} \tag{4.2}$$

Pokud není hierarchický model ve tvaru 4.1 nebo 4.2 odhadován bayesovsky, vychází veškerá teorie odhadu neznámých parametrů a inference z uvažovaného marginálního rozdělení závisle proměnné $\mathbf{Y}_i$. Z formulace modelu 4.1 vyplývá, že

$$\mathbf{Y}_i | \mathbf{b}_i \sim N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \boldsymbol{\Sigma}_i)$$
$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$$

s hustotou ozn. $f_i(\mathbf{y}_i|\mathbf{b}_i)$, resp.$f(\mathbf{b}_i)$. Pro hustotu marginálního rozdělení závisle proměnné $\mathbf{Y}_i$ platí

$$f_i\left(\boldsymbol{y}_i\right) = \int f_i\left(\boldsymbol{y}_i \mid \boldsymbol{b}_i\right) f\left(\boldsymbol{b}_i\right) d\boldsymbol{b}_i,$$

která lze ukázat, že je hustotou n_i-rozměrného normálního rozdělení se střední hodnotou $\mathbf{X}_i\boldsymbol{\beta}$ a kovarianční strukturou $\boldsymbol{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i$, tj.

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i). \tag{4.3}$$

Uvažujeme-li rozdělení 4.3 závisle proměnné $\mathbf{Y}_i$, hovoříme o tzv. marginálním modelu. Informace získané pomocí odhadu neznámých parametrů marginálního modelu 4.3 se týkají tedy fixních efektů, tj. průměrného populačního vývoje, náhodné efekty se podílejí pouze na kovarianční struktuře, jejich samotné odhady z marginálního modelu nezískáme.

Z hierarchické formulace modelu přirozeně vyplývá také jeho marginální struktura, ovšem z marginálního modelu není hierarchický jednoznačně určen. Různé formulace hierarchického modelu mohou totiž vést ke stejnému marginálnímu modelu. Ukažme si to na následujícím příkladu dvou různých hierarchických modelů, kdy pro jednoduchost uvažujeme pouze dvě měření subjektu, tj. $n_i = 2$. První hierarchický model bude s náhodným efektem interceptu a reziduální složkou $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} = \mathrm{diag}\,(\sigma_1^2, \sigma_2^2)$, výsledná kovarianční struktura z něj vyplývajícího marginálního modelu je ve tvaru

$$V = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (d)(1 \quad 1) + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} d + \sigma_1^2 & d \\ d & d + \sigma_2^2 \end{pmatrix}.$$

Druhý hierarchický model uvažujme vzájemně nekorelovaný náhodný efekt interceptu a lineárního trendu (směrnice) s reziduální složkou $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} = \sigma^2\mathbf{I}_2$, výsledná kovarianční struktura odpovídajícího marginálního modelu je ve tvaru

$$V = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \begin{pmatrix} d_1 + \sigma^2 & d_1 \\ d_1 & d_1 + d_2 + \sigma^2 \end{pmatrix}.$$

Pokud označíme $d_1 = d, d_2 = \sigma_2^2 - \sigma_1^2$, a $\sigma^2 = \sigma_1^2$, je jasné, že kovarianční struktura marginálního modelu vyplývající z první formulace hierarchického modelu a z druhého modelu si jsou rovny.

4.1 Odhady parametrů marginálního modelu

V této části se budeme zabývat možnostmi odhadu parametrů pro marginální model, vycházíme primárně z literatury [23], ale také z [25], ve které lze nalézt podrobnější popis a odvození uváděných tvrzeních.

Pro marginální model ve tvaru 4.3 uvažujeme normální rozdělení závisle proměnné $\mathbf{Y}_i$ s neznámými parametry fixních efektů $\boldsymbol{\beta}$ a neznámými parametry kovarianční struktury $\boldsymbol{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i$, které označme $\boldsymbol{\alpha}$, obsahující parametry kovarianční matice náhodných efektů $\mathbf{D}$ a kovarianční matice reziduální složky $\boldsymbol{\Sigma}_i$. Označme dále vektor neznámých parametrů marginálního modelu $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$. Věrohodnostní funkce marginálního modelu je ve tvaru

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} \left|\boldsymbol{V}_i(\boldsymbol{\alpha})\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)' \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha}) \left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) \right) \right\}, \tag{4.4}$$

přičemž, kdyby byl parametr $\boldsymbol{\alpha}$ známý, maximálně věrohodný odhad parametru $\boldsymbol{\beta}$ je roven

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{y}_i, \tag{4.5}$$

kde $\boldsymbol{W}_i = \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha})$. Odhad parametrů fixních efektů $\widehat{\boldsymbol{\beta}}$ lze tedy za předpokladu známého parametru $\boldsymbol{\alpha}$ získat metodou nejmenších vážených čtverců.

V praxi však parametr $\boldsymbol{\alpha}$ často není známý a je třeba jej nahradit ve vztahu 4.5 jeho odhadem $\widehat{\boldsymbol{\alpha}}$. Nejčastěji bývá nahrazen odhadem získaným metodou restringované maximální věrohodnosti, nebo klasické maximální věrohodnosti.

- **Metoda maximální věrohodnosti**, ozn. ML

  Odhad $\widehat{\boldsymbol{\alpha}}$ je získán po dosazení vztahu 4.5 do věrohodnostní funkce 4.4 a její maximalizací vzhledem k jediné neznámé $\boldsymbol{\alpha}$. Označme získaný odhad $\widehat{\boldsymbol{\alpha}}_{ML}$, přičemž odhady parametrů fixních efektů získáme zpětným dosazením $\widehat{\boldsymbol{\alpha}}_{ML}$ do vztahu 4.5, ozn. $\widehat{\boldsymbol{\beta}}_{ML}$.

  Odhady $\widehat{\boldsymbol{\alpha}}_{ML}, \widehat{\boldsymbol{\beta}}_{ML}$ získáme také maximalizací věrohodnosti 4.4 vzhledem k $\boldsymbol{\theta}$, tj. vzhledem k parametrům $\boldsymbol{\alpha}, \boldsymbol{\beta}$ současně.

- **Metoda restringované maximální věrohodnosti**, ozn. REML[1]

  Metoda klasické maximální věrohodnosti má nevýhodu v nezahrnutí ztráty stupňů volnosti vlivem odhadu parametru $\boldsymbol{\beta}$, což vede k tomu, že odhad $\widehat{\boldsymbol{\alpha}}_{ML}$ není nestranný. Tento problém je odstraněn tzv. restringovanou metodou maximální věrohodnosti. Vycházejme z modelu formulovaného dle vztahu 4.3 pro $i = 1, \ldots, N$, který zkombinujeme do jednoho modelu

  $$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\alpha})), \tag{4.6}$$

  kde $\mathbf{Y}$ je n-rozměrný vektor, $\mathbf{X}$ matice o n řádcích. Vzniknou naskládáním $\mathbf{Y}_i, \mathbf{X}_i$ za sebe. Matice $\mathbf{V}(\boldsymbol{\alpha})$ je blokovou maticí s bloky $\boldsymbol{V}_i$ na hlavní diagonále a nulami jinde. Pomocí ortogonální transformace matic[2] $\mathbf{A}$ vzhledem k $\mathbf{X}$ získáme

  $$\mathbf{U} = \mathbf{A}'\mathbf{Y} \sim N(\mathbf{0}, \mathbf{A}'\mathbf{V}(\boldsymbol{\alpha})\mathbf{A}), \tag{4.7}$$

  přičemž rozdělení $\mathbf{U}$ není již závislé na parametrech $\boldsymbol{\beta}$. Odhad parametru $\boldsymbol{\alpha}$ získáme metodou maximální věrohodnosti založené na $\mathbf{U}$ pro model 4.7. Je ukázáno [26], že tato věrohodnostní funkce lze zapsat ve tvaru

  $$\begin{aligned} L(\boldsymbol{\alpha}) =& (2\pi)^{-(n-p)/2} \left| \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{X}_i \right|^{1/2} \\ & \times \left| \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right|^{-1/2} \prod_{i=1}^{N} |\boldsymbol{V}_i|^{-1/2} \\ & \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}} \right)' \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}} \right) \right\}, \end{aligned} \tag{4.8}$$

  kde $\widehat{\boldsymbol{\beta}}$ je dán vztahem 4.5. Věrohodnost nezávisí na volbě matice $\mathbf{A}$. Maximálně věrohodný odhad získaný na základě 4.8 označme $\widehat{\boldsymbol{\alpha}}_{REML}$, dosazením do vztahu 4.5 získáme $\widehat{\boldsymbol{\beta}}_{REML}$.

[1] z angl. Restricted Maximum Likelihood

[2] $\mathbf{A}$ je libovolná matice rozměrů $n \times (n-p)$ plné (sloupcové) hodnosti, sloupcově ortogonální ke sloupcům matice $\mathbf{X}$

Simultánní odhad parametrů $\boldsymbol{\alpha}, \boldsymbol{\beta}$ lze získat metodou restringované maximální věrohodnosti díky platnosti vztahu

$$L_{REML}(\boldsymbol{\theta}) = \left| \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{X}_i \right|^{-\frac{1}{2}} L_{ML}(\boldsymbol{\theta}), \tag{4.9}$$

kde $L_{\mathrm{ML}}(\boldsymbol{\theta})$ je věrohodnostní funkce dle vztahu 4.4, pomocí maximalizace této funkce 4.9 vzhledem k parametru $\boldsymbol{\theta}$.

Oba tyto přístupy vedou k odhadům s vlastnostmi maximálně věrohodných odhadů jako je konzistence, asymptotická normalita a eficience, viz kap. 1.2. Jelikož odhad metodou maximální věrohodnosti je vychýlený, častěji se využívá metoda restringované maximální věrohodnosti. Obecně se od sebe odhady liší tím více, čím více roste počet odhadovaných fixních efektů. Zvláštní pozornost musí být kladena výběru metody odhadu pro inferenci o parametrech modelu, které se budeme věnovat v následující kapitole. Pro samotné získání odhadů se využívá nejčastěji numerických metod, např. Newtonova-Raphsonova metoda nebo EM algoritmus[3].

4.2 Inference pro marginální model

Cílem marginálního modelu 4.3 je často kromě samotného získání odhadu parametrů fixních odhadů a jejich interpretace také posuzování jejich významu na základě testů o parametrech. Pro správnou inferenci založenou na daném modelu je však potřeba mít správně formulovánu i hierarchickou strukturu modelu, ze kterého daný marginální model vychází, neboť má vliv na kovarianční matici marginálního modelu. Vzhledem k tomu, že jsme si v předcházející části odvodili možnosti odhadu parametrů pomocí metody maximální věrohodnosti i restringované maximální věrohodnosti, také testy budou založeny na věrohodnosti a testových statistikách uvedených v části 1.3 s nutnými úpravami pro náš konkrétní problém.

[3] z angl. Expectation-Maximization Algorithm, více o metodě je uvedeno např. v [23, kap. 22], [27, kap. 8]

4.2.1 Fixní efekty

Odhad fixních efektů je získán vztahem 4.5, za předpokladu známého parametru $\boldsymbol{\alpha}$ se řídí mnohorozměrným normálním rozdělením se střední hodnotou $\boldsymbol{\beta}$ a kovarianční strukturou

$$\begin{aligned}&\mathrm{var}(\widehat{\boldsymbol{\beta}})\\&= \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \,\mathrm{var}\left(\boldsymbol{Y}_i\right) \boldsymbol{W}_i \boldsymbol{X}_i\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1}\\&= \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1},\end{aligned} \tag{4.10}$$

kde $\boldsymbol{W}_i = \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha})$. V praxi však opět parametr $\boldsymbol{\alpha}$ není známý a je nahrazen svým odhadem $\widehat{\boldsymbol{\alpha}}_{ML}$ nebo $\widehat{\boldsymbol{\alpha}}_{REML}$. Touto náhradou jsou rozdělení získaných testových statistik aproximovány rozděleními vycházejících ze známých testů.

Aproximativní Waldův test

Rozdělení testové statistiky

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{var}(\widehat{\beta}_j)}} \tag{4.11}$$

aproximujeme normálním normovaným rozdělením. Obecně poté pro hypotézu

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0} \quad \text{proti alternativě} \quad H_1 : \mathbf{L}\boldsymbol{\beta} \neq \mathbf{0} \tag{4.12}$$

platí, že rozdělení testové statistiky

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{L}' \left[\mathbf{L} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\alpha}}) \boldsymbol{X}_i\right)^{-1} \mathbf{L}'\right]^{-1} \mathbf{L}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

má asymptoticky χ^2-rozdělení se stupni volnosti rovny hodnosti matice $\mathbf{L}$. Při výpočtu je třeba dosadit hypotetické hodnoty $\mathbf{L}\boldsymbol{\beta}$ za platnosti nulové hypotézy. Rozhodnutí o platnosti hypotézy je stejné jako pro klasický Waldův test uvedený v části 1.3, tj. zamítáme H_0 pro velké hodnoty testové statistiky W.

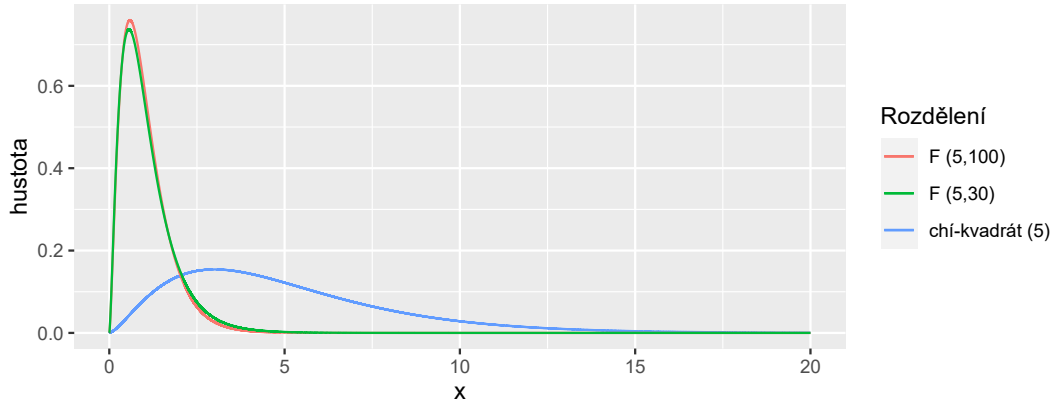
Aproximativní t-test, F-test

Waldův test nezohledňuje přidanou variabilitu nahrazením parametru $\boldsymbol{\alpha}$ jeho odhadem $\widehat{\boldsymbol{\alpha}}$, jeho použití povede k platné inferenci pouze při dostatečně velkých výběrech. V praxi se proto využívá spíše aproximativního t-testu, resp. F-testu, které vzniknou aproximací rozdělení testové statistiky 4.11 pomocí studentova t-rozdělení, z které pro obecnou hypotézu ve tvaru 4.12 vyplývá aproximativní F-test, kdy testová statistika

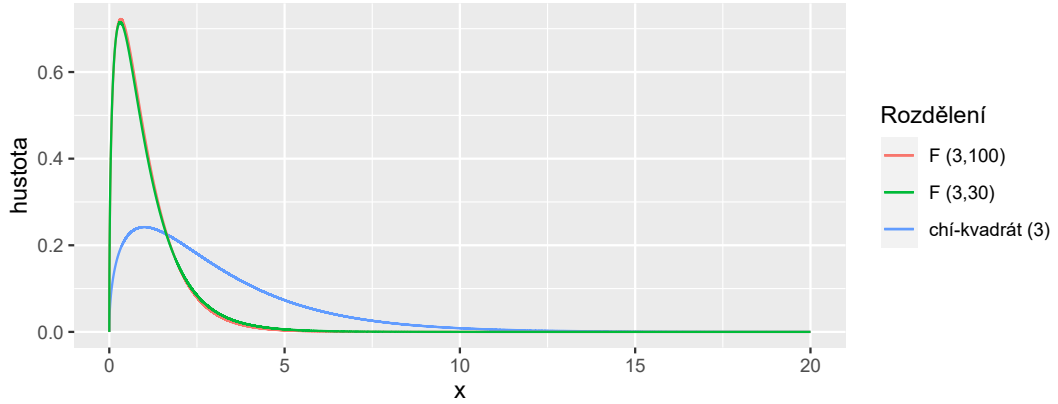
$$F = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{L}' \left[\mathbf{L} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1}(\widehat{\alpha}) \boldsymbol{X}_i \right)^{-1} \mathbf{L}' \right]^{-1} \mathbf{L}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\text{rank}(\mathbf{L})}$$

má aproximativně F-rozdělení o (p, q) stupních volnosti, kde $p = rank(\mathbf{L})$ je rovno hodnosti matice $\mathbf{L}$ a q je získáno aproximací z dat, nejčastěji využitím Satterthwaite nebo Kenward-Roger aproximace. V kontextu longitudinálních dat všechny aproximativní metody vedou k velkému počtu stupňů volnosti a dávají tak velmi podobné výsledné p-hodnoty. Při výpočtu je třeba opět v testové statistice dosadit hodnoty $\mathbf{L}\boldsymbol{\beta}$ za platnosti nulové hypotézy, kterou také zamítáme pro velké hodnoty testové statistiky F.

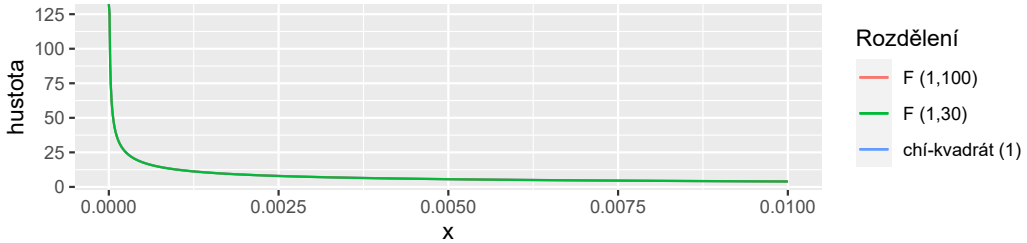
Rozdílnost hustot uvažovaných rozdělení aproximativního Waldova testu a aproximativního F-testu je zachycena na obrázku 4.1 a 4.2 pro zvolené hodnoty hodnosti matice $\mathbf{L}$, konkrétně $rank(\mathbf{L}) = 5$ a $rank(\mathbf{L}) = 3$. Pro $rank(\mathbf{L}) = 1$ je χ^2-rozdělení a F-rozdělení shodné, jak lze pozorovat i na obrázku 4.3, kde se vykreslené hustoty zcela dokonale překrývají.



Obrázek 4.1: Srovnání hustot χ^2-rozdělení o 5 stupních volnosti s F-rozdělením o (5, 100), resp. (5, 30) stupních volnosti.



Obrázek 4.2: Srovnání hustot χ^2-rozdělení o 3 stupních volnosti s F-rozdělením o (3, 100), resp. (3, 30) stupních volnosti.



Obrázek 4.3: Srovnání hustot χ^2-rozdělení o 1 stupni volnosti s F-rozdělením o (1, 100), resp. (1, 30) stupních volnosti.

Likelihood-ratio test

Klasickým nástrojem pro porovnání modelu a jeho podmodelu, jehož odhady byly získány metodou maximální věrohodnosti, je likelihood-ratio test, viz kapitola 1.3. Stejná teorie je platná také pro inferenci fixních efektů marginálního modelu, ovšem pouze za podmínky získání odhadu metodou klasické maximální věrohodnosti, tj. nahrazením neznámého parametru $\boldsymbol{\alpha}$ jeho odhadem $\widehat{\boldsymbol{\alpha}}_{ML}$. Při odhadu modelu a jeho podmodelu pomocí REML metody bychom totiž porovnávali věrohodnosti získané na různých datech, neboť jak bylo uvedeno v části 4.1, REML odhad je získán pomocí transformace $\mathbf{U}$, která má pro různou strukturu fixních efektů jinou podobu.

4.2.2 Náhodné efekty

V praxi využíváme marginálního modelu především s primárním zájmem o fixní efekty, avšak volba uvažované struktury náhodných efektů hierarchického modelu je důležitým faktorem pro platnou inferenci založenou na odhadnutém modelu, získání eficience a také pro interpretaci zvoleného náhodného chování v datech. Testy uvažovaného hierarchického modelu, ze kterého vyplývá náš odhadnutý marginální model, jsou možné pomocí likelihood-ratio testu za předpokladu homoskedasticity reziduální složky, tj. $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$, a s nutnou aproximací rozdělení testové statistiky. Pohlížíme-li totiž na parametry matice $\mathbf{D}$ z hierarchického pohledu, dostáváme se s hypotézou o jejich nulovosti na hranici parametrického prostoru, neboť diagonální prvky vyjadřují rozptyly, které jsou nezáporné hodnoty. Bylo ukázáno, že testová statistika likelihood-ratio testu pro testování takovéto nulové hypotézy nemá χ^2-rozdělení, ale jeho smíšenou podobu, tj. její rozdělení je kombinací dvou χ^2-rozdělení o rozdílných stupních volnosti. Nejčastějšími případy nulových hypotéz spolu s rozdělením jejich testové statistiky jsou:

- H_0: Žádný náhodných efekt vs. H_1: Jeden náhodný efekt

  Pro tuto hypotézu má testová statistika $\chi^2_{0:1}$ rozdělení, tj. p-hodnota bude rovna $p = P(\chi^2_{0:1} > LR) = \frac{1}{2}P(\chi^2_0 > LR) + \frac{1}{2}P(\chi^2_1 > LR)$. Rozdělení χ^2_0 je

rozdělením, které s mírou pravděpodobnosti 1 nabývá hodnoty 0.

- H_0: Jeden náhodný efekt vs. H_1: Dva náhodné efekty

  Pro tuto hypotézu má testová statistika $\chi^2_{1:2}$ rozdělení, tj. p-hodnota bude rovna $p = P(\chi^2_{1:2} > LR) = \frac{1}{2}P(\chi^2_1 > LR) + \frac{1}{2}P(\chi^2_2 > LR)$. Nejčastěji testujeme tuto hypotézu pro zjištění nutnosti zařazení náhodného lineárního efektu směrnice oproti pouze náhodnému efektu interceptu.

Na rozdíl od využití likelihood-testu pro testy o fixních efektech se nyní doporučuje využívat REML odhadu, neboť při stejné struktuře fixních efektů je transformace $\mathbf{U}$ shodná, věrohodnosti jsou v tomto případě získány ze stejných dat, proto jsou také porovnatelné.

4.3 Problém neúplných pozorováních

Ve studiích, ze kterých získáváme korelovaná pozorování, je častým problémem také nepozorování všech naplánovaných měření. Nejčastějším problémem longitudinálních studií bývá tzv. drop-out, tj. situace, kdy do určitého časového okamžiku máme kompletní pozorování, ale později k dalším naplánovaným měřením již nedojde, pozorovaný subjekt „vypadl“[4] ze studie. Problém neúplných pozorování se projeví především ve ztrátě eficience s rostoucí nejistotou v čase.

Uvedená teorie vychází stále z [23], ale čerpáme i z [27], kde čtenář nalezne také potřebné podrobnější informace pro práci s neúplnými daty jiných struktur než jsou longitudinální data, na něž se nyní zaměříme.

Uvažujme, že pro i-tý subjekt bylo naplánováno celkem n_i měření Y_{ij} v časových okamžicích $j = 1, \ldots, n_i$. Výsledek měření je uskupen do vektoru $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$. Navíc pro každý okamžik j definujme proměnnou R_{ij} jako indikátor chybějícího pozorování, tj.

$$R_{ij} = \begin{cases} 1 & \text{jestliže } Y_{ij} \text{ je pozorováno,} \\ 0 & \text{jinak.} \end{cases} \tag{4.13}$$

[4] příkladem může být změna místa bydliště, úmrtí, nutné změny léčby mimo studii aj.

Indikátor chybějících pozorování můžeme uskupit do vektoru $\mathbf{R}_i$ stejné délky jako $\mathbf{Y}_i$. Dle něj lze psát $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$, kde $\mathbf{Y}_i^o$ obsahuje všechny napozorované prvky Y_{ij} (tj. pro $R_{ij} = 1$), $\mathbf{Y}_i^m$ obsahuje všechny nenapozorované prvky Y_{ij} (tj. pro $R_{ij} = 0$). Je-li vektor $\mathbf{R}_i$ ve tvaru $(1, \ldots, 1, 0, \ldots, 0)'$, lze nahradit tzv. *drop-out indikátorem* D_i označující buď měření, při kterém došlo k opuštění studie, tj. $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$, nebo počet dosažených úplných pozorování, tj. $D_i = \sum_{j=1}^{n_i} R_{ij}$. Jestliže lze struktura chybějících dat zjednodušit pomocí drop-out indikátoru, označujeme ji za *monotónní*, v opačném případě se jedná o *nemonotónní* vzor chybějících hodnot.

Dále je potřeba rozlišit mechanismus, který vede k získání daného vzoru chybějících hodnot. Předpoklady, které učiníme o tomto mechanismu, jsou zásadní pro přístup k analýze dat s neúplnými pozorováními. Mechanismus je charakterizován na základě podmíněného rozdělení $\mathbf{R}_i$ za podmínky $\mathbf{Y}_i$. Podle $f(\mathbf{R}_i|\mathbf{Y}_i, \phi)$, kde ϕ označuje neznámé parametry rozdělení, rozlišujeme:

- **MCAR mechanismus**, *z angl. missing completely at random*

  Jestliže ztráta nezávisí na hodnotách $\mathbf{Y}_i$, označujeme mechanismus za zcela náhodnou ztrátu (MCAR), tj. pokud platí

  $$f(\mathbf{R}_i|\mathbf{Y}_i, \phi) = f(\mathbf{R}_i|\phi) \quad \text{pro všechny } \mathbf{Y}_i, \phi.$$

- **MAR mechanismus**, *z angl. missing at random*

  Jestliže ztráta nezávisí na nenapozorovaných hodnotách $\mathbf{Y}_i^m$, ale závisí na napozorovaných hodnotách $\mathbf{Y}_i^o$, označujeme mechanismus jako náhodnou ztrátu (MAR), tj. pokud platí

  $$f(\mathbf{R}_i|\mathbf{Y}_i, \phi) = f(\mathbf{R}_i|\mathbf{Y}_i^o, \phi) \qquad \text{pro všechny } \mathbf{Y}_i^m, \phi.$$

- **MNAR mechanismus**, *z angl. missing not at random*

  Jestliže ztráta závisí i na nenapozorovaných hodnotách $\mathbf{Y}_i^m$, označujeme mechanismus jako nenáhodnou ztrátu (MNAR).

Pro inferenci založenou na věrohodnosti je vzor mechanismus chybějících dat ignorovatelný,[5] jestliže chybějící data jsou náhodnou ztrátou (MAR) a parametrický prostor $\mathbf{\Theta}_\theta$ a $\mathbf{\Theta}_\phi$ odlišný. Není-li splněna podmínka odlišnosti parametrických prostorů, je inference založená na věrohodnosti při neúplných pozorováních platná, ale ne plně eficientní. Úvahy vedoucí k odvození platnosti tohoto tvrzení pro případ lineárních smíšených modelů lze nalézt např. v [23, kap. 15.5-15.9], obecněji [27, kap. 6.2].

[5] z angl. likelihood ignorability

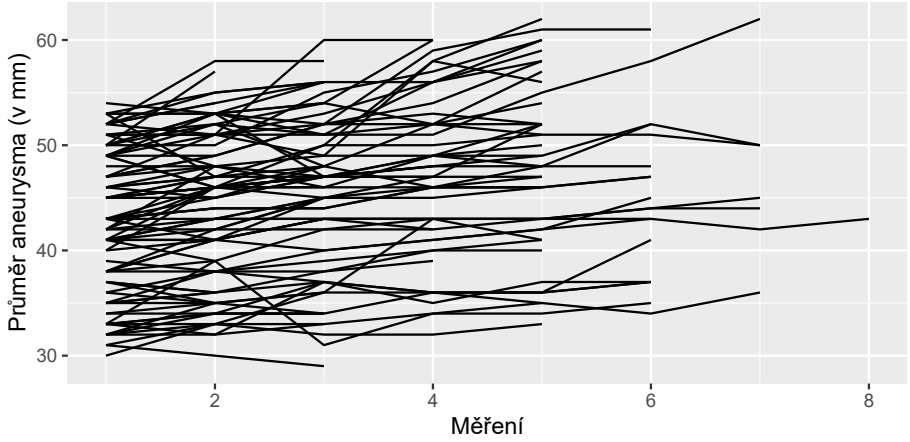
Kapitola 5

Lineární smíšené modely – aplikace

V této části práce aplikujeme teoretické poznatky o lineárních smíšených modelech na longitudinální datové sadě, která byla získána při studiu této problematiky v rámci programu Erasmus+ na KU Leuven v Belgii. Celkem máme k dispozici záznamy 101 pacientů trpících aneurysmatem břišní aorty, kteří byli sledování pravidelně každých šest měsíců, nejdéle po osm návštěv.

Aneurysma aorty břišní je vyboulení či rozšíření aorty v oblasti dutiny břišní způsobené oslabením aortální stěny. Vzhledem k vysoké mortalitě při ruptuře aneurisma je základním cílem pozorování určit faktory, které se podílejí na rychlosti růstu průměru sledovaného aneurysmatu. [21]

Vývoj tohoto parametru pro jednotlivé pacienty je zachycen na obrázku 5.1.



Obrázek 5.1: Vývoj průměru aneurysma pro jednotlivé pacienty zachycený při pravidelných měřeních opakujících se každých šest měsíců.

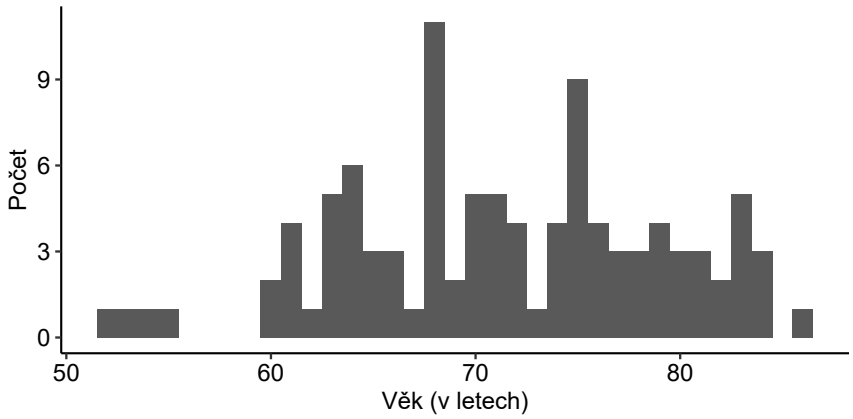
Faktory zahrnuté ve studii (hodnoty byly získány při prvním měření), které mohou mít vliv na vznik onemocnění a mohou také souviset s jeho vývojem, jsou:

- **věk**

  Základní popisné charakteristiky faktoru věku pacientů jsou uvedeny v tabulce 5.1, histogram četnosti zastoupení poté na obrázku 5.2. Riziko rozvoje tohoto onemocnění se udává vyšší pro pacienty nad 50 let [21], což odpovídá i zastoupení účastníků ve studii, kdy nejmladšímu bylo 52 let.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
52	66	71	71,36	77	86

Tabulka 5.1: Základní popisné charakteristiky vysvětlující proměnné věku při vstupu do studie vyjádřeného v letech.



Obrázek 5.2: Histogram věku pacientů při vstupu do studie.

- **BMI** ... body mass index (kg/m^2)

  BMI je jedním z možných ukazatelů zdravého životního stylu pacientů, umožňující porovnávání hmotnosti pacientů s různou výškou. Klasifikace do jednotlivých kategorií, dle výše hodnoty BMI, jsou uvedeny v tabulce 5.2 [22]. Orientační souhrn váhového rozpětí jednotlivých kategorií pro zvolené výšky dospělého jedince je uveden v tabulce 5.3. Základní popisné

charakteristiky faktoru BMI u sledovaných pacientů jsou uvedeny v tabulce 5.4, četnost zastoupení zachycuje histogram 5.3.

Podváha	méně než 18,5
Optimální váha	18,5 – 24,9
Nadváha	25 – 29,9
Obezita I. stupně	30 – 34,9
Obezita II. stupně	35 – 39,9
Obezita III. stupně	více než 40

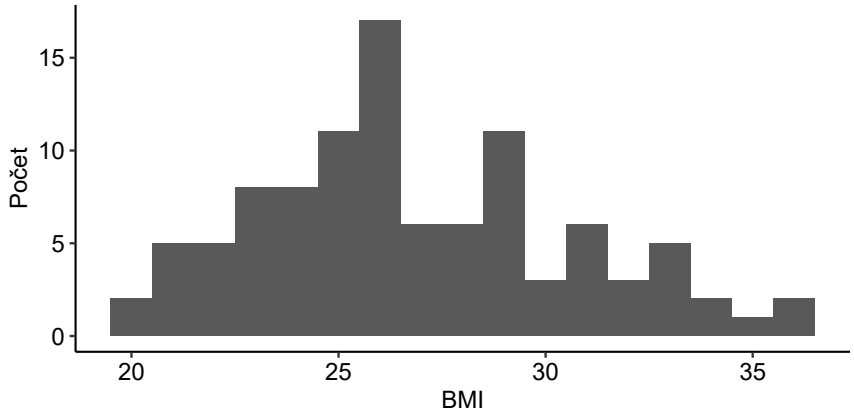
Tabulka 5.2: Klasifikace dle hodnoty BMI.

	160 cm	170 cm	180 cm	190 cm
Podváha	$<$ 47 kg	$<$ 53 kg	$<$ 60 kg	$<$ 67 kg
Optimální váha	47 – 64 kg	53 – 72 kg	60 – 81 kg	67 – 90 kg
Nadváha	64 – 77 kg	72 – 86 kg	81 – 97 kg	90 – 108 kg
Obezita I. stupně	77 – 89 kg	87 – 101 kg	97 – 113 kg	108 – 126 kg
Obezita II. stupně	90 – 102 kg	101 – 115 kg	113 – 129 kg	126 – 144 kg
Obezita III. stupně	$>$ 102 kg	$>$ 115	$>$ 129 kg	$>$ 144 kg

Tabulka 5.3: Orientační váhová klasifikace dle kategorií BMI pro zvolené výšky.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19,8	24,3	26,2	26,85	29,1	36

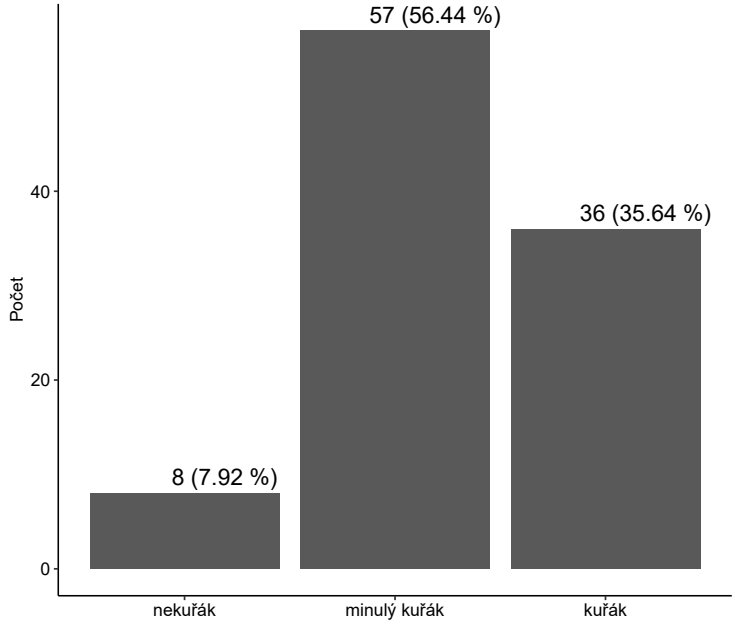
Tabulka 5.4: Základní popisné charakteristiky vysvětlující proměnné BMI při vstupu do studie v jednotkách kg/m^2.



Obrázek 5.3: Histogram BMI pacientů při vstupu do studie v jednotkách kg/m^2.

- **Kouření**

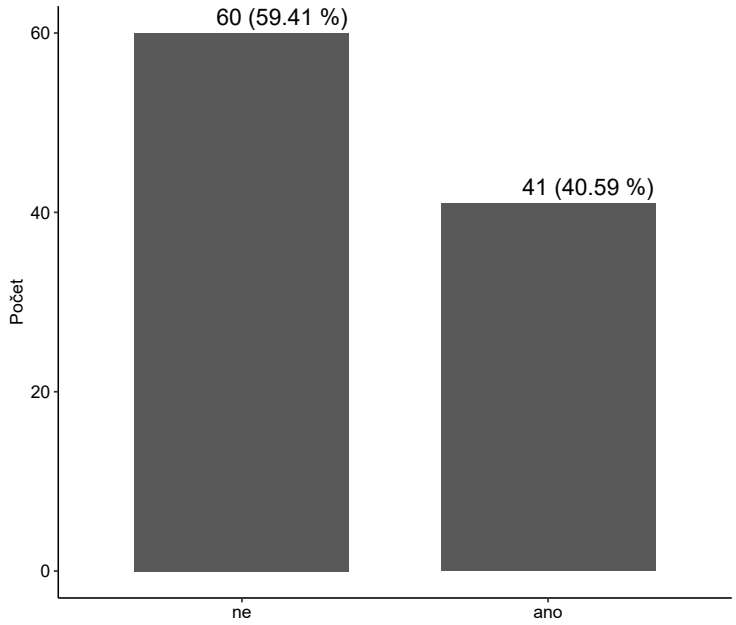
  Kuřácká historie či aktuální kouření jsou také významným rizikovým faktorem pro vznik a vývoj tohoto onemocnění [21], zastoupení jednotlivých kategorií ve studii je zobrazeno na obrázku 5.4.



Obrázek 5.4: Četnost a procentuální rozdělení pacientů dle vztahu ke kouření při vstupu do studie.

- **Ischemická choroba srdeční**, ozn. ICHS

  Udává informaci o tom, zda pacient při vstupu do studie již trpí tímto onemocněním koronárních tepen, jehož příčinou je ateroskleróza (kornatění) věnčitých tepen, které má vliv také na rozvoj aneurysmatu. [20], [21] Zastoupení pacientů s tímto onemocněním je zobrazeno na obrázku 5.5.



Obrázek 5.5: Četnost a procentuální rozdělení pacientů trpících při vstupu do studie ischemickou chorobou srdeční.

5.1 Výstavba marginálního modelu

Jelikož cílem naší analýzy je zjistit průměrný vývoj tohoto onemocnění v populaci a možné faktory, které na něj mohou mít vliv, chceme sestavit tzv. marginální model. Pro jeho určení budeme postupovat v souladu s obecnými doporučeními, viz [23, kap. 9]. Potřebujeme nejprve specifikovat jak předběžnou strukturu fixních efektů, tak také předběžnou strukturu náhodných efektů, které mají vliv na kovarianční strukturu marginálního modelu. Pro lepší interpretaci odhadu interceptu budeme dále pracovat s lehce upravenými proměnnými:

- **věk nad 50 let**, který vznikl odečtením 50 let od původní proměnné věk, díky čemuž se hodnota odhadu interceptu modelu, který bude obsahovat fixní efekt věku, vztahuje k 50letému pacientovi a nikoliv k novorozenci ve věku 0.

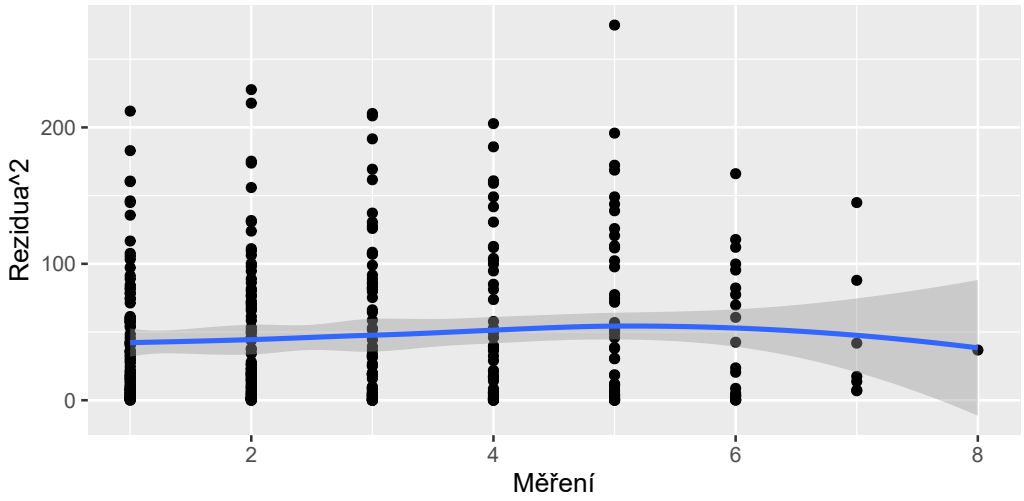
- **BMI nad 20**, které vzniklo jako rozdíl mezi skutečnou hodnotou BMI pacienta a hodnotou 20. Tato změna nám umožní interpretovat nyní hodnotu interceptu, bude-li informace o BMI obsažena v modelu, jako hodnotu průměru aneurysmatu pro člověka normální váhy.

Nejprve tedy zvolíme předběžnou strukturu fixních efektů, kterou uvažujeme co možná nejkomplexnější, avšak stále reálnou. Komplexita fixních efektů je důležitým předpokladem pro další práci s odhadovaným modelem, neboť opomenutím některé složky mající vliv na střední hodnotu závislé proměnné bychom nemuseli dospět ke konzistentní kovarianční struktuře v dalších krocích modelace. V našem případě jsme zvolili všechny výše zmiňované faktory, také jejich interakce s časovým faktorem a interakční členy mezi věkem a hodnotou BMI a také mezi kouřením a ischemickou chorobou srdeční.

Pro volbu vhodné kovarianční struktury marginálního modelu, na které se podílí také kovarianční struktura náhodných efektů, potřebujeme naše data očistit o veškerý systematický trend. Proto nejprve pro naši komplexní strukturu fixních efektů získáme odhady klasickou metodou nejmenších čtverců ignorující korelace

mezi jednotlivými pozorováními. Z tohoto modelu získáme rezidua, která představují veškerou zbylou variabilitu, jež není vysvětlena pomocí fixních efektů. Pokud si je umocněná vykreslíme a proložíme vyhlazenou křivkou, dle jejího tvaru můžeme identifikovat, které náhodné efekty bude vhodné uvažovat pro jejich předběžnou strukturu.

Jak můžeme vidět na obrázku 5.6, proložená křivka[1] má mírný lineární růst. Budeme tedy uvažovat jako předběžnou strukturu náhodných efektů intercept a také časový faktor měření pro lineární trend (směrnici). Pro tuto předběžnou



Obrázek 5.6: Umocněná rezidua získaná z klasického lineárního modelu s odhadem parametrů pomocí MNČ proložená křivkou získanou metodou *LOESS*.

strukturu náhodných efektů a předběžnou strukturu fixních efektů je třeba ještě rozhodnout, zda jeho reziduální složka je složena z části zachycující chybu měření a části sériové korelace či nikoliv, viz model 4.2 v kapitole 4. Odhadneme proto lineární smíšený model s předběžnou strukturou fixních efektů a náhodných efektů pomocí REML metody pro různou volbu funkce modelující sériovou korelaci chyb. Na základě AIC kritéria jsme provedli srovnání výsledných modelů, přičemž ani jeden z nich nebyl lepší než model s nejjednodušší varianční strukturou – za předpokladu homoskedasticity chyb, který tedy také zvolíme.

Nyní můžeme přistoupit k možné redukci předběžné struktury náhodných

[1] získaná metodou *LOESS*

efektů – k rozhodnutí o potřebě náhodného efektu času můžeme využít likelihood-ratio testu. Jelikož se však s touto hypotézou pohybujeme na hranici parametrického prostoru, viz kap. 4.2.2, rozdělením testové statistiky LR je smíšené rozdělení $\chi^2_{1:2}$. Na základě výsledku tohoto testu uvedeného v tabulce 5.5 zamítáme možnost přechodu k jednodušší struktuře náhodných efektů, naše předběžně zvolená struktura náhodného interceptu a náhodného lineárního efektu času měření je finální.

Model	$\ln L(\boldsymbol{\theta})_{REML}$	LR	df	p-hodnota
Lineární efekt času měření	−959,723			
Pouze intercept	−1006,060	92,674	1:2	$< 0{,}0001$

Tabulka 5.5: Likelihood-ratio test významnosti náhodného lineárního časového efektu založený na REML odhadu s odpovídajícím smíšeným rozdělením $\chi^2_{1:2}$.

Pro získanou vhodnou kovarianční strukturu marginálního modelu volbou náhodného interceptu a náhodného lineárního efektu času měření budeme postupně redukovat předběžnou volbu fixních efektů, neboť tato komplexní forma není nutná, jak lze vidět dle četnosti vysoce nesignifikantních výsledků aproximativního t-testu nulovosti jednotlivých parametrů uvedených v tabulce 5.6. Volba parametrů při postupné redukci byla založena na základě aproximativního t-testu o jednotlivých parametrech a také pomocí aproximovaného F-testu významnosti kombinace zvolených nejméně jednotlivě signifikantních proměnných (při zachování hierarchie efektů). Ověření možnosti přechodu k jednodušším podmodelu bylo provedeno také pomocí likelihood-ratio testu, proto všechny odhady v průběhu procesu redukce struktury fixních efektů byly získány metodou klasické maximální věrohodnosti, jinak bychom neporovnávali věrohodnosti získané na stejné datové struktuře, viz kap. 4.2.1.

Tímto postupem jsme získali finální marginální model s redukovanou strukturou fixních efektů obsahující časový faktor měření a věk nad 50 let, vycházející z uvažovaného hierarchického modelu s náhodným efektem interceptu a náhodným lineárním efektem času měření. Odhady parametrů fixních efektů finálního

	Odhad	s.e.	p-hodnota
Intercept	32,639	5,067	$< 0{,}0001$
Měření	1,576	0,637	0,014
Věk nad 50 let	0,309	0,178	0,086
BMI nad 20	0,462	0,554	0,407
Minulý kuřák	3,427	2,885	0,238
Kuřák	1,119	2,919	0,702
$ICHS_{ano}$	0,246	5,212	0,963
Měření:Věk nad 50 let	0,004	0,019	0,831
Měření:BMI nad 20	$-0{,}004$	0,037	0,925
Měření:Minulý kuřák	$-0{,}740$	0,457	0,106
Měření:Kuřák	$-0{,}397$	0,467	0,396
$Měření{:}ICHS_{ano}$	0,151	0,251	0,547
Věk nad 50 let:BMI nad 20	$-0{,}024$	0,024	0,335
$Minulý\ kuřák{:}ICHS_{ano}$	$-0{,}567$	5,511	0,918
$Kuřák{:}ICHS_{ano}$	2,633	5,638	0,642

Tabulka 5.6: (ML) Odhady fixních parametrů marginálního modelu se směrodatnou odchylkou a p-hodnotou aproximativního t-testu.

marginálního modelu, získané pomocí metody restringované maximální věrohodnosti, jsou shrnuty v tabulce 5.7. V populaci můžeme očekávat, že průměrně bude velikost aneurysmatu při prvním vyšetření u 50letého pacienta dosahovat skoro 38 mm. V našich datech není pozorováno, že by na průměrný vývoj velikosti aneurysmatu aorty břišní v populaci měl mít vliv faktor BMI, ischemická choroba srdeční či kouření. Je však nutno podotknout, že zastoupení jednotlivých skupin kuřáků není vyvážené, což může také souviset s neprokázáním jeho vlivu na průměrný růst aneurysmatu. Naopak pozorujeme, že s každým měřením, tj. každých šest měsíců, aneurysma roste v populaci v průměru o 1,15 mm, přičemž s každým rokem nad 50 let při prvním měření je aneurysma průměrně větší o 0,18 mm.

	Odhad	s.e.	p-hodnota
Intercept	37,92	1,899	$< 0{,}0001$
Měření	1,15	0,124	$< 0{,}0001$
Věk nad 50 let	0,181	0,084	0,033

Tabulka 5.7: (REML) Odhady fixních parametrů marginálního modelu se směrodatnou odchylkou a p-hodnotou aproximativního t-testu.

5.2 Problém neúplných pozorování

Již v grafu individuálních profilů 5.1 lze pozorovat, že zdaleka ne u všech pacientů došlo k osmi měřením průměru aneurysmatu. Vývoj počtu pacientů ve studii dle počtu absolvovaných měření je znázorněn společně se vzorem chybějících hodnot v tabulce 5.8. Odtud také lze pozorovat, že neúplná pozorování jsou způsobena trvalým opuštěním studie (tzv. drop-out), jedná se tedy o monotonní vzor chybějících pozorování. Z kapitoly 4.3 víme, že inference platí za předpokladu náhodného mechanismu chybějících pozorování (MAR). Dle zkoumaného problému můžeme usuzovat, že pacient opustil studii pouze v závislosti na napozorovaných hodnotách průměru aneurysmatu, které vedly buď k chirurgickému zákroku, příp. úmrtí, nebo z důvodů zcela nesouvisejících s onemocněním, proto předpoklad MAR mechanismu není nereálný a modely v předcházející části 5.1 byly získány dle platné inference i za přítomnosti neúplných pozorováních.

		Měření								Počet	Procentuální podíl	Kumulativní % podíl
		1	2	3	4	5	6	7	8			
Vzor	I.	O	O	O	O	O	O	O	O	1	0,99	0,99
	II.	O	O	O	O	O	O	O	M	6	5,94	6,93
	III.	O	O	O	O	O	O	M	M	12	11,88	18,81
	IV.	O	O	O	O	O	M	M	M	27	26,73	45,54
	V.	O	O	O	O	M	M	M	M	7	6,93	52,47
	VI.	O	O	O	M	M	M	M	M	21	20,79	73,26
	VII.	O	O	M	M	M	M	M	M	16	15,84	89,10
	VIII.	O	M	M	M	M	M	M	M	11	10,90	100

Tabulka 5.8: Vzor chybějících pozorování (ozn. O: pozorované, M: chybějící) spolu s absolutním, relativním a kumulativně relativním zastoupením vzhledem k celkovému počtu pacientů.

Závěr

Cílem práce bylo seznámit se s modely využívajícími metodu maximální věrohodnosti, resp. její modifikaci, pro odhad neznámých parametrů a využít získané znalosti při analýze reálných dat. Proto jsme se v práci nejprve seznámili s teoretickými aspekty vybraných metod, které jsme následně využili při zpracování reálných problémů pocházejících z medicínského výzkumu.

První vybranou metodou byla analýza přežití s využitím ke zpracování dat zachycujících dobu přežití pacientů trpících rakovinou v oblasti dutiny ústní. Nejprve jsme se seznámili s datovou sadou pomocí nástrojů popisné statistiky a využili jsme Kaplanových-Meierových neparametrických odhadů funkcí přežití k získání představy o rozdílnosti v přežití dle jednotlivých rizikových faktorů zvlášť. Pro zohlednění současného vlivu různých rizikových faktorů jsme zvolili Coxův model proporcionálních rizik. Dle finálního Coxova modelu, s ověřenou platností předpokladů proporcionality rizik, se jako statisticky významné rizikové faktory prokázaly: virový původ nádoru, recidiva nádoru, duplicita nádoru, konzumace alkoholu a dosažené vzdělání.

Druhou vybranou metodou byly lineární smíšené modely s využitím ke zpracování longitudinálních dat o pacientech trpících aneurysmatem aorty břišní, kteří byli pravidelně sledováni každých šest měsíců. Spolu s tímto údajem jsme měli k dispozici také možné faktory, které mohou kromě vzniku aneurysmatu ovlivňovat také tempo jeho růstu. Výsledný marginální model, vycházející z hierarchické struktury uvažující náhodné efekty interceptu a náhodného lineárního trendu (směrnice) času, neprokázal vliv jiných fixních faktorů než času a věku při vstupu do studie na populační průměrný vývoj velikosti aneurysmatu.

Literatura

[1] ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 4. Praha: Matfyzpress, 2013. ISBN 978-80-7378-162-0.

[2] ANDĚL, Jiří. *Matematická statistika*. 2. vyd. Praha: SNTL - Nakladatelství technické literatury, 1985, 352 s. (Váz.).

[3] HOSMER, David W., Stanley LEMESHOW a Susanne MAY. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd Edition. Hoboken, New Jersey: John Wiley & Sons, 2008. ISBN 978-0-471-75499-2.

[4] KALBFLEISCH, John D. a Ross L. PRENTICE. *The Statistical Analysis of Failure Time Data*. 2nd Edition. Hoboken, New Jersey: John Wiley & Sons, 2002. ISBN 0-471-36357-X.

[5] PROCHÁZKA, Bohumír. *Biostatistika pro lékaře: principy základních metod a jejich interpretace s využitím statistického systému R*. V Praze: Univerzita Karlova v Praze, nakladatelství Karolinum, 2015. ISBN 978-80-246-2782-3.

[6] MOORE, Dirk F. *Applied Survival Analysis Using R*. Switzerland: Springer International Publishing, 2016. Use R! ISBN 978-3-319-31243-9.

[7] HARRELL, JR., Frank E. *Regression Modeling Strategies*. 2. vyd. Cham: Springer International Publishing, 2015. Springer Series in Statistics. ISBN 978-3-319-19424-0.

[8] BRIERLEY, James, M. K. GOSPODAROWICZ, Christian WITTEKIND, et al., ed. *TNM: klasifikace zhoubných novotvarů*. Česká verze 2018. Přeložil Kristýna SALAČOVÁ, přeložil Miroslav ZVOLSKÝ. Praha: [Ústav zdravotnických informací a statistiky České republiky], 2018. ISBN 978-80-7472-173-1.

[9] *Slovníček* [online]. Linkos - Česká onkologická společnost České lékařské společnosti J.E. Purkyně, ©2020. [cit. 20. 10. 2020]. Dostupné z: `https://www.linkos.cz/slovnicek/`

[10] *O nádorech hlavy a krku* [online]. Česká onkologická společnost ČLS JEP pacientům a jejich blízkým, ©2020. Poslední změna 18. 10. 2017 [cit. 20. 10. 2020]. ISSN 1801-9951. Dostupné z: https://www.linkos.cz/pacient-a-rodina/onkologicke-diagnozy/nadory-hlavy-a-krku-c00-14-c30-32/o-nadorech-hlavy-a-krku/

[11] PEŘINA, Vojtěch, Jiří BLAHÁK a Oliver BULIK. Karcinomy hlavy a krku — vlivy HPV infekce. *LKS: časopis České stomatologické komory* [online]. Praha: Česká stomatologická komora, 19. 10. 2015, 25(10), 198-202 [cit. 20. 10. 2020]. ISSN 2571-2411. Dostupné z: http://www.lks-casopis.cz/clanek/karcinomy-hlavy-a-krku-vlivy-hpv-infekce/

[12] HARRINGTON, David P. a Thomas R. FLEMING. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika* [online]. 1982, **69**(3), 553–566 [cit. 19. 2. 2021]. ISSN 00063444. Dostupné z: https://www.jstor.org/stable/2335991 doi:10.2307/2335991

[13] GEHAN, Edmund A. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika* [online]. 1965, **52**(1/2), 203–223 [cit. 19. 2. 2021]. ISSN 00063444. Dostupné z: http://www.jstor.org/stable/2333825 doi:10.2307/2333825

[14] TARONE, Robert E. a James WARE. On Distribution-Free Tests for Equality of Survival Distributions. *Biometrika* [online]. 1977, **64**(1), 156–160 [cit. 19. 2. 2021]. ISSN 00063444. Dostupné z: http://www.jstor.org/stable/2335790 doi:10.2307/2335790

[15] PETO, Richard a Julian PETO. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General)* [online]. 1972, **135**(2), 185–207 [cit. 19. 2. 2021]. ISSN 00359238. Dostupné z: http://www.jstor.org/stable/2344317 doi:10.2307/2344317

[16] FLEMING, Thomas R., David P. HARRINGTON a Margaret O'SULLIVAN. Supremum Versions of the Log-Rank and Generalized Wilcoxon Statistics. *Journal of the American Statistical Association* [online]. 1987, **82**(397), 312–320 [cit. 19. 2. 2021]. ISSN 01621459. Dostupné z: http://www.jstor.org/stable/2289169 doi:10.2307/2289169

[17] COX, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* [online]. 1972, **34**(2), 187–220 [cit. 19. 2. 2021]. ISSN 00359246. Dostupné z: http://www.jstor.org/stable/2985181

[18] GRAMBSCH, Patricia M. a Terry M. THERNEAU. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* [online].

1994, **81**(3), 515–526 [cit. 24. 2. 2021]. ISSN 00063444. Dostupné z: www.jstor.org/stable/2337123 doi:10.2307/2337123

[19] FAQ: How are the likelihood ratio, Wald, and Lagrange multiplier (Score) tests different and/or similar? *UCLA: Statistical Consulting Group* [online]. ©2021 [cit. 24. 2. 2021]. Dostupné z: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-andor-similar/

[20] Ischemická choroba srdeční. *Nemocnice na Homolce* [online]. © Nemocnice Na Homolce, 2017 [cit. 1. 3. 2021]. Dostupné z: https://www.homolka.cz/nase-oddeleni/11635-kardiovaskularni-program/11635-kardiochirurgie-kch/informace-pro-pacienty/11723-nase-sluzby/kch-ischem-choroba-srdecni/

[21] Abdominal Aortic Aneurysm. *Medtronic* [online]. © Medtronic, 2021 [cit. 1. 3. 2021]. Dostupné z: https://www.medtronic.com/us-en/patients/conditions/abdominal-aortic-aneurysm.html

[22] Body mass index - BMI. *World Health Organization: Regional office for Europe* [online]. WHO, © 2021 [cit. 3. 3. 2021]. Dostupné z: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

[23] VERBEKE, Geert a Geert MOLENBERGHS. *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag, 2000. Springer Series in Statistics. ISBN 0-387-95027-3.

[24] VERBEKE, Geert a Geert MOLENBERGHS. *Models for Discrete Longitudinal Data*. New York: Springer-Verlag, 2005. Springer Series in Statistics. ISBN 0-387-25144-8.

[25] RAO, C. Radhakrishna, Helge TOUTENBURG, SHALABH a Christian HEUMANN. *Linear Models and Generalizations: Least Squares and Alternatives*. 3. vyd. Berlin Heidelberg: Springer-Verlag, 2008. Springer Series in Statistics. ISBN 978-3-540-74226-5.

[26] HARVILLE, DAVID A. Bayesian inference for variance components using only error contrasts. *Biometrika* [online]. 1974, **61**(2), 383-385 [cit. 8. 3. 2021]. ISSN 0006-3444. Dostupné z: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/61.2.383 doi:10.1093/biomet/61.2.383

[27] LITTLE, Roderick J. A. a Donald B. RUBIN. *Statistical Analysis with Missing Data*. 2. vyd. Hoboken, New Jersey: John Wiley Sons, 2002. ISBN 0-471-18386-5.

[28] VYKYDALOVÁ, Veronika. *Kolik sázenek je vsazeno ve Sportce?*. Olomouc, 2018. Bakalářská práce. Univerzita Palackého v Olomouci, Přírodovědecká fakulta. Vedoucí práce Mgr. Ondřej Vencálek, Ph.D.

Příloha: kódy v R

Datové sady byly zpracovány pomocí softwaru R (verze 4.0.3). Vybrané užité balíčky a konkrétní příkazy jsou pro čtenáře uvedeny v této příloze, mohou sloužit jako pomůcka při aplikaci metod na vlastní datovou sadu.

```
####### Analýza přežití ########
library(survminer)
library(survival)

### Kaplanův-Meierův odhad funkce přežití ###
km0 = survfit(Surv(doba.preziti, umrti)~1, data=data_vyber, conf.type='log-log')
#Odhad mediánu přežití
surv_median(km0)
#Grafické znázornění - s tabulkou počtu v riziku,
#znázorněným mediánem a 95% CI s využitím log-log transformace
ggsurvplot(km0, data=data_vyber, risk.table='abs_pct', surv.median.line = 'hv',
          break.time.by=100, legend='none', ylab='Pravděpodobnost přežití',
          xlab='Týdny', risk.table.title='Počet v riziku: n (%)')
#Grafické znázornění - s p-hodnotou log-rank testu
ggsurvplot(km, data=data_vyber, pval=T, pval.method =T)

### Coxův model ###
fit1 = coxph(Surv(doba.preziti, umrti) ~ HPV_pozitivni+ velikost.nadoru
             +recidiva.a.n+duplicita+terapie
             +pohlavi + vzdelani.skupiny + koureni + alkohol, data = data_vyber)
summary(fit1)
#Grafické znázornění
ggforest(fit1)
#Testování modelu a podmodelu likelihood-ratio testem
anova(fit1, fit2)
#Ověření předpokladu proporcionality rizik
test.ph = cox.zph(main.effect.model)
test.ph
ggcoxzph(test.ph, font.main = 12, font.x = 10, font.y=10) #grafické znázornění
```

```
####### Lineární smíšené modely ########
library(nlme)

#Smíšený model s předběžnou strukturou fixních efektů
#a předběžnou strukturou náhodných efektů
pre.lme.model.slope = lme(fixed = DMA~Measurement + years.up.50 + BMI.up.20 +
                                  smoking.status + cds + Measurement:years.up.50 +
                                  Measurement:BMI.up.20 + Measurement:cds +
                                  Measurement:smoking.status +
                                  years.up.50:BMI.up.20 + smoking.status:cds,
                          data = data,
                          random = ~Measurement|id)
#Model s reziduální složkou specifikovanou i částí sériové korelace
pre.lme.model.slope.ARMA11 = update(pre.lme.model.slope, corr=corARMA(p=1,q=1))

#Redukce předběžné struktury fixních efektů
pre.lme.model.slope = lme(fixed = DMA~Measurement + years.up.50 + BMI.up.20 +
                                  smoking.status + cds + Measurement:years.up.50 +
                                  Measurement:BMI.up.20 + Measurement:cds +
                                  Measurement:smoking.status +
                                  years.up.50:BMI.up.20 + smoking.status:cds,
                          data = data,
                          method = 'ML',
                          random = ~Measurement|id)
#aproximativní t-test nulovosti jednotlivých efektů
summary(pre.lme.model.slope)
#aproximativní F-test nulovosti kategorických vysvětlujících proměnných
anova(pre.lme.model.slope, type = 'marginal') #type-III ANOVA
#aproximativní F-test zadaných parametrů
anova(pre.lme.model.slope, Terms = c('cds', 'smoking.status:cds'))
#likelihood-ratio test
anova(pre.lme.model.slope, pre.lme.model.slope1)
```

```
####### Popisné grafy ########
library(ggplot2)

#histogram s absolutními počty a procentualním zastoupením
ggplot(data_vyber2, aes(x=lok2)) + geom_bar(width=0.9) + labs(y='Pocet',x=NULL) +
  geom_text(
    aes(label=..count..),
    stat='count') +
  geom_text(
    aes(label=paste0(' (',round((..count../sum(..count..))*100, digit=2),' %)')),
    stat='count')


####### Úprava dat ########
library(tibble)
library (dplyr)
library(tidyr)
library(skimr)
library(ggpmisc)
library(ggpubr)
```