UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Modely analýzy přežití s neproporcionálním rizikem a jejich aplikace

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Eva Fišerová, Ph.D**
Vypracovala: **Bc. Sylva Šmoldasová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2020

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Sylva Šmoldasová

Název práce: Modely analýzy přežití s neproporcionálním rizikem a jejich aplikace

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D

Rok obhajoby práce: 2020

Abstrakt: Práce je zaměřena na modely analýzy přežití v případě porušení předpokladu proporcionality rizik. V teoretické části jsou představeny základní pojmy a charakteristiky analýzy přežití včetně proporcionality a možností jejího ověření. Jsou zde vysvětleny principy výstavby Coxova modelu proporcionálních rizik, stratifikovaného Coxova modelu a modelu konkurenčních rizik. V práci jsou popsány metody pro odhad parametrů v modelech a metody pro výběr a hodnocení modelu. Teoretické znalosti jsou následně aplikovány na praktický příklad, kdy je sledována doba přežití pacientů s rakovinovým onemocněním gastrointestinálního traktu. Analýza je provedena za pomoci statistického softwaru R, verze 3.6.3.

Klíčová slova: Analýza přežití, funkce přežití, riziková funkce, Kaplanův-Meierův odhad, proporcionalita rizik, Coxův model, stratifikované modely, modely konkurenčních rizik, log-rank test, Waldův test, test poměrem věrohodností, skórový test, informační kritéria, konkordance, software R

Počet stran: 120

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Sylva Šmoldasová

Title: Survival models with nonproportional hazards and their applications

Type of thesis: Masters's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D

The year of presentation: 2020

Abstract: The thesis is focused on survival analysis models in case of violation of proportionality hazards assumption. The theoretical part presents the basic concepts and characteristics of survival analysis, including the concept of proportionality hazards and the possibility of testing it. The basic principles of construction of the Cox proportional hazards model, the stratified Cox model and the competing risks model are defined. At the same time, methods of parameter estimation and methods of model evaluation and selection are described. Theoretical knowledge is applied to a practical problem where the survival time of patients with gastrointestinal cancer is followed up. The analysis is performed using the statistical software R, version 3.6.3.

Key words: Survival analysis, survival function, risk function, Kaplan-Meier estimator, proportional hazards, Cox model, stratified models, competing risks models, log-rank test, Wald test, likelihood-ratio test, score test, Information criteria, concordance, software R

Number of pages: 120

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych na tomto místě poděkovala zejména vedoucí mé diplomové práce paní doc. RNDr. Evě Fišerové, Ph.D za odborné vedení, cenné rady, ochotu a věnovaný čas. Dále bych ráda poděkovala své rodině a přátelům za podporu při studiu a tvorbě této práce.

Úvod

Analýza přežití je odvětví statistiky, které zkoumá čas do výskytu nějaké předem určené (definované) události či jevu. Touto událostí můžeme chápat například smrt pacienta jako následek konkrétního onemocnění (odtud název analýza přežití). Uvažovaná událost však nemusí být jen negativního charakteru, jak bylo právě uvedeno, ale může znamenat například uzdravení pacienta nebo úspěšné otěhotnění při procesu umělého oplodnění. Analýza přežití nachází uplatnění nejen v medicíně, ale také v ekonomii, sociologii nebo například i v technických oborech. Předmětem analýzy tedy může být například sledování doby životnosti různých přístrojů do času jejich selhání, sledování doby do úspěšného nalezení zaměstnátní pro člověka vedeného na úřadu práce či naopak sledování doby, než zaměstanec podá výpověď. Jako další příklady lze uvést sledování délky vztahu dvou lidí do svatby či naopak délky manželství do rozvodu.

Nejpoužívanějším modelem v analýze přežití je Coxův model proporcionálních rizik [2], [12], který se používá pro modelování vlivů prediktorů na čas přežití nebo jinak řečeno na čas, než dojde k události. Mezi výhody využití tohoto modelu patří především jeho snadné použití a následná jednoduchá interpretace výsledků, nicméně i tento model má své nevýhody. Hlavním problémem tohoto modelu jsou silné předpoklady, které jsou kladeny na data a při jejichž porušení můžeme dostat nepřesné nebo zkreslené výsledky. Předpoklady nutné pro aplikaci tohoto modelu jsou následující:

- proporcionalita rizik,
- hodnoty vysvětlujících proměnných jsou konstantní v čase,
- sledovaná událost může nastat pouze jednou pro každý subjekt,
- nekorelovanost mezi časů událostí různých subjektů.

Hlavním předpokladem pro použití tohoto modelu je proporcionalita rizik. Jinými slovy, požadujeme, aby byl poměr rizik různých skupin subjektů konstantní v čase. Představme si, že sledujeme dobu přežití pacientů u nichž bylo diagnostikováno nějaké konkrétní onemocnění. V takovém případě budeme požadovat, aby poměr rizik skupiny pacientů, kteří podstupují léčbu a skupiny pacientů, kteří léčbu nepodstupují, zůstával v čase neměnný. V této práci se budeme zabývat tím, jak modelovat data, která nesplňují právě podmínku proporcionality.

V první kapitole si představíme základní pojmy a charakteristiky používané v analýze přežití. Naučíme se, jak tyto charakteristiky odhadnout z dat a jaké vzájemné vztahy mezi nimi platí. Druhá kapitola bude věnována Coxově modelu proporcionálních rizik a metodě pro odhad parametrů tohoto modelu. Na konci druhé kapitoly se také seznámíme s metodami pro ověření předpokladu proporcionality. Ve třetí kapitole se budeme věnovat modelům s neproporcionálními riziky. Představíme si zde stratifikovaný Coxův model a krátce také modely konkurenčních rizik. Čtvrtá kapitola pak bude věnována testování hypotéz a metodám pro výběr a hodnocení modelů. Konečně v páté kapitole aplikujeme získané teoretické znalosti na reálný příklad. Data, která použijeme v praktické části se budou týkat rakovinného onemocnění gastrointestinálního traktu.

Kapitola 1

Úvod do analýzy přežití

V této kapitole si zavedeme a vysvětlíme základní pojmy a charakteristiky používané v analýze přežití.

1.1. Cenzorování a krácení

Jak již bylo řečeno v úvodu této práce, analýza přežití spočívá ve sledování doby (času) než dojde k předem dané události. Často se však může stát, že nebudeme mít u některých sledovaných subjektů kompletní informace o celkovém čase od počátku studie až do dané události. V takových případech mluvíme o *cenzorovaném času přežití (cenzorování)*[9]. Cenzorování dělíme celkem do tří skupin:

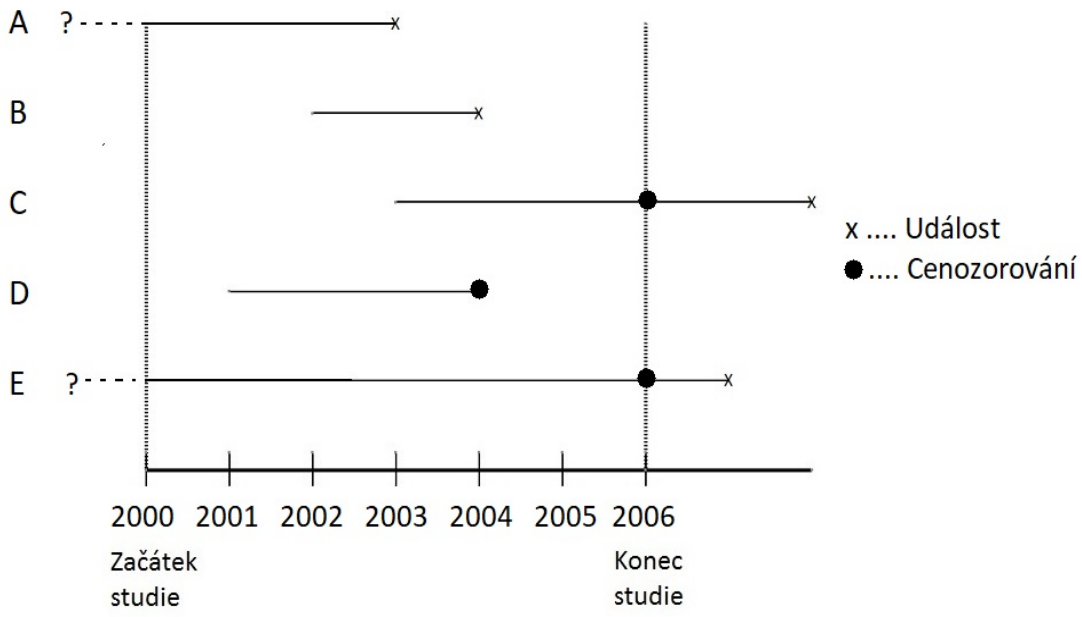
1. **Cenzorování zprava**: Nastává v situaci, kdy o subjektu ztratíme informaci během studie, například pokud: subjekt ze studie dobrovolně odejde, odstěhuje se a není možné jej nadále sledovat, je ze studie vyřazen, dojde k jiné události, jež znemožní nastání události, kterou sledujeme nebo k události během studie nedojde vůbec. Poslední případ bývá zapříčiněn faktem, že sledovat všechny subjekty od počátku až do konce studie by mohlo být nejen časově, ale i finančně náročné, a proto se předem stanovuje maximální délka studie, a tedy její konec.

2. **Cenzorování zleva**: V tomto případě nám chybí informace o přesném počátku, odkdy má být subjekt sledován. Příkladem může třeba být sledování času přežití od nakažení se určitou nemocí až do smrti pacienta. Jako začátek studie budeme

uvažovat čas prvního pozitivního testu na danou nemoc, je ale velmi pravděpodobné, že se pacient nakazil již dříve a my tedy nevíme kolik času uplynulo, než test nemoc odhalil. Stejně tak si cenzorování zleva můžeme ilustrovat na příkladu zubního kazu. Zubní kaz bývá odhalen zubařem při prohlídce, nicméně nevznikl až v době prohlídky, ale již dříve, my však nevíme kdy.

3. **Intervalové cenzorování**: Intervalové cenzorování je kombinací cenzorování zprava a zleva. Příkladem může být právě nakažení se konkrétní nemocí. Pacient může být nakažen dlouho před tím, než je nemoc odhalena testem a zároveň pacient během studie na tuto nemoc nezemře (pokud je smrt na tuto nemoc sledovanou událostí).

Pro lepší představu o těchto pojmech se nyní podívejme na obrázek 1.1, kde je znázorněna fiktivní studie probíhající od roku 2000 do roku 2006. Předpokládejme, že cílem této



Obrázek 1.1: Cenzorování: zleva (A), zprava (C,D) a intervalové (E).

studie bylo sledovat pacienty od nakažení se určitou nemocí do jejich smrti. Do studie nám vstoupilo pět pacientů, označených písmeny A-E. Pacient A představuje cenzorování zleva,

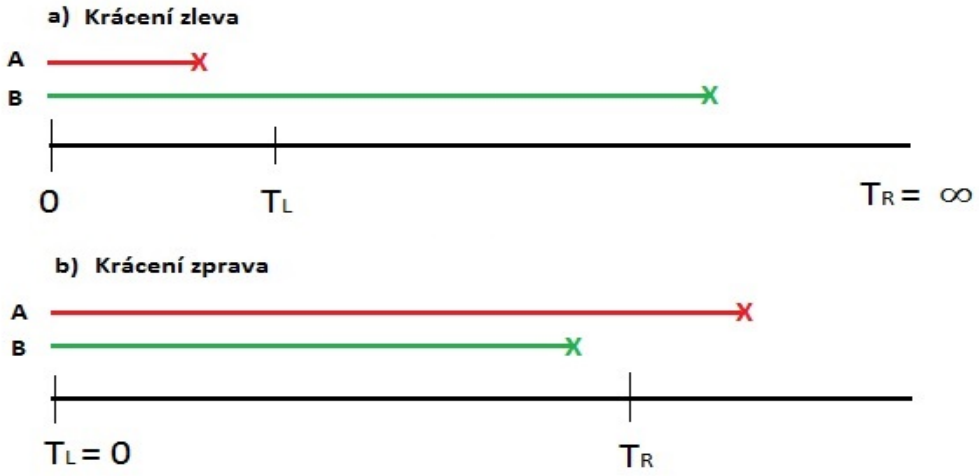
neboť nevíme, kdy došlo k nakažení – víme pouze, že v roce 2000 již nemocný byl a víme také, že k události došlo v roce 2003. Pacient B představuje pozorování s úplnou informací, neboť víme, že k nakažení došlo v roce 2002 a zemřel již v roce 2004. Pacient C představuje cenzorování zprava. Víme, že k nakažení došlo v roce 2003, nicméně dále víme pouze to, že do konce studie k události nedošlo. Pacient D opět představuje cenzorování zprava – víme, že k nakažení došlo v roce 2001, ale v roce 2004 ze studie odešel a my nemáme informace o jeho dalším vývoji. Nakonec pacient E představuje intervalové cenzorování, neboť nemáme informaci ani o čase nakažení ani o čase události – víme pouze, že pacient byl nakažený, a že k události během studie nedošlo.

V analýze přežití se kromě cenzorování používá pojem *krácení* (anglicky truncation) [3], [9]. Tento pojem bývá často nesprávně zaměňován právě s pojmem cenzorování neboť také značí určitou ztrátu informace. V tomto případě jsou do studie zahrnuty pouze subjekty, u nichž v uvažovaném časovém intervalu k události došlo. Tento interval si označme jako (T_L, T_R). Všechny ostatní subjekty jsou ze studie vyloučeny – což je významný rozdíl oproti cenzorování, kdy máme o všech subjektech alespoň nějakou informaci. V praxi můžeme například chtít do studie zahrnout jen takové pacienty, kteří byli do času T_L naživu a ke smrti následkem dané nemoci došlo do časového okamžiku T_R. Všichni ostatní pacienti nesplňující tuto podmínku jsou ze studie vyloučeni a nemáme o nich žádné informace. Obdobně jako rozlišujeme cenzorování na cenzorování zprava, zleva a intervalové cenzorování, rozlišujeme také krácení na:

1. **Krácení zleva**: V tomto případě stanovujeme pouze počátek studie, tedy časový okamžik T_L. Ze studie jsou pak vyloučeny všechny subjekty, u nichž došlo k události před tímto okamžikem. Obecně pak můžeme uvažovat, že platí $T_R = \infty$. Celý interval lze zapsat jako: (T_L, ∞). Příkladem může být sledování doby, po kterou pobírá jedinec starobní důchod. Je zřejmé, že se člověk musí nejdříve dožít určitého věku, aby měl na starobní důchod nárok. Ze studie jsou tedy vyloučeni všechni jedinci, kteří se tohoto věku nedožili. Naopak není dán věk, po jehož dosažení by jedinec na starobní důchod nárok ztratil. Krácení zleva se používá například v souvislosti se životním a důchodovým pojištěním.

2. **Krácení zprava**: Při krácení zprava naopak stanovujeme konec studie, tedy časový okmažik T_R. Ze studie poté vyloučíme všechny subjekty, u nichž nedošlo do tohoto času k události. Obecně pak můžeme počátek studie označit jako $T_L = 0$. Celý interval pak lze zapsat jako: $(0, T_R)$. Je zřejmé, že v tomto případě může dojít k opravdu zásadní ztrátě informace. Představme si, že bychom uvažovali sledování doby od nakažení se určitou nemocí až do smrti pacienta. V případě krácení zprava bychom tedy z další analýzy vyřadili pacienty, kteří žili i po ukončení studie. Tato skutečnost je však v daném případě velmi zásadní. Vyloučením přeživších pacientů ze studie totiž uměle krátíme délku přežití u dané nemoci. Je tedy velmi důležité zamyslet se, zda je použití krácení vhodným krokem. Krácení zprava však také nalézá své využití, příkladem může být často uváděná studie týkající se inkubační doby AIDS [11]. V této studii byl měřen čas od infikování jedince virem AIDS skrze kontaminovanou krevní transfúzi do času vyvinutí samotné nemoci. Dospělým jedincům byla kontaminovaná krevní transfúze podána 1.dubna 1978, konec studie byl následně stanoven na datum 30.června 1986. Jedinci, u nichž se nemoc nerozvinula ani do tohoto data, byli ze studie vyloučeni.

Pro snažší pochopení krácení zprava a zleva se podívejme na obrázek 1.2. Zde máme



Obrázek 1.2: Krácení zleva (a) a zprava (b)

postupně vyobrazeny situace pro krácení zleva a krácení zprava. Při krácení zleva uvažujeme obecně $T_R = \infty$ a jako začátek studie je považován čas T_L. U subjektu A, jehož čas přežití je vyznačen červenou barvou, došlo k události ještě před časem T_L a proto nebude do studie zahrnut. Naopak subjekt B do studie zahrnut bude, neboť k události došlo až po tomto čase. V případě krácení zprava je pevně určen začátek studie jako $T_L = 0$ a konec studie představuje libovolný, avšak předem daný čas T_R. Znamená to tedy, že ze studie bude vyloučen subjekt A a ponechán pouze subjekt B.

1.2. Základní charakteristiky analýzy přežití

Nechť T je náhodná veličina, která označuje čas přežití. Časem přežití chápeme dobu od začátku sledování subjektu do nastání předem určené události. Dále t bude představovat konkrétní hodnotu pro T, tedy její realizaci. Pravděpodobnostní rozdělení náhodné veličiny T lze popsat pomocí tří základních funkcí, a to pomocí *funkce přežití*, *hustoty pravděpodobnosti* (v případě spojitého rozdělení náhodné veličiny T) a *rizikové funkce* [12]. Někdy se k těmto třem funkcím přiřazují i následující číselné charakteristiky: *průměrná doba přežití*, *průměrná doba dožití* a *medián přežití* [9]. Přičemž platí, že známe-li alespoň jednu z výše uvedených funkcí, jsme schopni zbylé funkce dopočítat. Ačkoli se však jedná o matematicky ekvivalentní funkce, každá z nich nám data vykresluje z jiného pohledu. Představme si nyní tyto funkce podrobněji.

1.2.1. Funkce přežití

Funkce přežití (anglicky survival function) nám vyjadřuje pravděpodobnost, že subjekt přežije déle než do času t. Jinými slovy, jedná se o pravděpodobnost, že čas přežití daného subjektu bude delší než zvolený čas t. Tuto funkci budeme dále značit jako $S(t)$. Funkce přežití tedy vyjadřuje pravděpodobnost, že se náhodná veličina T realizuje hodnotou větší než zvolená hodnota t, což lze zapsat jako:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t), \tag{1.1}$$

kde $F(t)$ označuje distribuční fukci náhodné veličiny T. Protože funkce přežití $S(t)$ vyjadřuje pravděpodobnost, může nabývat hodnot pouze z intervalu $\langle 0; 1\rangle$, přičemž platí, že $S(t) = 1$ v čase $t = 0$ a $S(t) = 0$ pro $t = \infty$.

Uvažujme nejprve diskrétní čas, tedy situaci, kdy T je diskrétní náhodná veličina, která se realizuje obecně hodnotami t_i, kde $i = 1, 2, 3, \ldots$, a to s pravděpodobnostmi $p(t_i) = P(T = t_i), \quad i = 1, 2, 3, \ldots$. Funkci přežití pro diskrétní náhodnou veličinu T lze zapsat ve tvaru:

$$S(t) = P(T > t) = \sum_{t_i > t} p(t_i). \tag{1.2}$$

Nyní předpokládejme spojitý čas. Náhodná veličina T je v tomto případě popsána hustotou pravděpodobnosti $f(x)$. Funkci přežití lze ve spojitém případě zapsat jako:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx. \tag{1.3}$$

Protože však nikdy neznáme skutečnou hodnotu této pravděpodobnosti, je nutné funkci přežití odhadnout. Odhad funkce přežití budeme značit jako $\hat{S}(t)$. V případě, že nemáme v datovém souboru cenzorovaná data, můžeme odhad provést následujím způsobem:

$$\hat{S}(t) = \frac{\text{počet subjektů, jejichž čas přežití je větší než } t}{\text{celkový počet subjektů}}. \tag{1.4}$$

Odhad funkce přežití ve tvaru (1.4) nelze použít v případě, že máme v datovém souboru cenzorovaná pozorování. Přítomnost cenzorovaných pozorování je však pro data týkající se analýzy přežití zcela typická, a proto se zde budeme věnovat přístupu, který lze využít právě i v případě cenzorovaných časů přežití. Jedná se o Kaplanův-Meierův odhad funkce přežití. [9], [13] Označme si jako n_i počet pozorování, které jsou v čase t_i v riziku (tzn. jejich pozorovaný čas přežití je delší nebo roven času t_i) a d_i počet událostí jež v čase t_i nastaly. Nechť dále $t_{(1)}$ označuje čas výskytu první události (jedná se o minimální čas přežití), pak lze Kaplanův-Meierův odhad funkce přežití v čase t psát jako:

$$\hat{S}(t) = \begin{cases} 1 & \text{pro } t < t_{(1)}, \\ \prod_{t_i \le t} \left[1 - \frac{d_i}{n_i}\right] & \text{pro } t_{(1)} \le t. \end{cases} \tag{1.5}$$

Poznamenejme, že Kaplanův-Meierův odhad funkce přežití lze počítat pouze pro časové okamžiky t, které nejsou větší než nejdelší napozorovaný čas (označme $t_{(n)}$). V případě, že je tento nejdelší pozorovaný čas přežití necenzorovaný (jedná se o čas nastání události), pak $\hat{S}(t_{(n)}) = 0$, neboť v posledním členu součinu bude $d_i = n_i$. Bude-li naopak nejdelší pozorovaný čas přežití cenzorovaný, pak Kaplanův-Meierův odhad funkce přežití sice nedosáhne nulové hodnoty, nicméně za tímto nejdelším časem přežití již není tento odhad definován. [12] Existují dva různé přístupy, jak se s tímto problémem vypořádat. Prvním z nich je uvažovat odhad funkce přežití za nulový okamžitě za tímto nejdelším (cenzorovaným) časem přežití, tj. $\hat{S}(t) = 0$ pro $t > t_{(n)}$. Tento přístup pracuje s předpokladem, že u subjektu, jehož čas je cenzorován, došlo ke sledované události ihned po ukončení studie. Druhou možností je pak pracovat s předpokladem, že u daného subjektu dojde ke sledované události naopak v čase $t = \infty$, tj. $\hat{S}(t) = \hat{S}(t_{(n)})$ pro $t > t_{(n)}$. Oba tyto přístupy mají pro velký počet pozorování stejné vlastnosti (jsou asymptoticky ekvivalentní) a konvergují ke skutečné funkci přežití. Ukázalo se však, že pro menší počty pozorování je vhodnější použít druhý přístup. [8], [9]

V praxi nás kromě bodových odhadů funkce přežití zajímají také odpovídající intervalové odhady. Pro sestrojení intervalu spolehlivosti je třeba znalost odhadu směrodatné chyby, resp. odhadu rozptylu. Odhad rozptylu $\hat{S}(t)$ přitom získáme následovně:

$$\widehat{var}\left[\hat{S}(t)\right] \approx \left[\hat{S}(t)\right]^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}, \tag{1.6}$$

kde n_i je opět počet pozorování v riziku v čase t_i a d_i počet událostí, jež v tomto čase nastaly. Vzorec (1.6) se nazývá Greenwoodův vzorec. [9], [12] Pro konstrukci $100(1-\alpha)\%$ intervalu spolehlivosti však použití tohoto roztptylu, resp. odpovídající směrodatné odchylky není příliš vhodné. Pokud bychom $100(1-\alpha)\%$ interval spolehlivosti konstruovali obvyklým způsobem, tedy za pomoci aproximace normálním rozdělením (s využitím platnosti centrální litmitní věty), dostali bychom následující:

$$I_{1-\alpha}\left(S(t)\right) = \left\langle \hat{S}(t) - u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}\left[\hat{S}(t)\right]}; \hat{S}(t) + u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}\left[\hat{S}(t)\right]} \right\rangle, \tag{1.7}$$

kde $u_{1-\frac{\alpha}{2}}$ označuje $1 - \frac{\alpha}{2}$ kvantil normovaného normálního rozdělení. Ze vztahu (1.7) je

zřejmé, že v každém časovém okamžiku t se jedná o symetrický interval, což způsbuje jisté problémy. Pokud bychom se zaměřili na takto sestavený interval spolehlivosti pro časové okamžiky t, ve kterých je odhad funkce přežití blízký jedné nebo naopak nule, zjistili bychom, že pro některé časové okamžiky t tento interval spolehlivosti připouští i hodnoty funkce přežití vyšší než jedna nebo naopak hodnoty nižší než nula. Jistým řešením může být nastavení pravidel, kdy je každý horní odhad, který je vyšší než jedna, automaticky nahrazen hodnotou jedna a analogicky je i každý dolní odhad, který je menší než nula, nahrazen hodnotou nula.

Existuje však i jiný způsob, jak vytvořit $100(1-\alpha)\%$ interval spolehlivosti, který nebude překračovat hranice intervalu $\langle 0; 1\rangle$. Výchozí myšlenkou je přitom transformace odhadu funkce přežití tak, aby jeho hodnoty mohly ležet v intervalu $(-\infty; +\infty)$. Interval spolehlivosti je následně vytvořen pro transformovaný odhad funkce přežití a na závěr jsou výsledné intervaly spolehlivosti transformovány zpět tak, aby odpovídaly intervalům spolehlivosti pro odhad funkce přežití. Často používanou volbou je komplementární logaritmická transformace, která je ve tvaru [2], [9]:

$$\ln\left(-\ln\left(\hat{S}(t)\right)\right). \tag{1.8}$$

Nyní tedy chceme sestrojit $100(1-\alpha)\%$ interval spolehlivosti pro uvedenou transformaci. Nejprve si musíme vyjádřit rozptyl transformované funkce přežití: $\widehat{var}\left[\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right]$. Využijeme přitom platnosti následujícího vztahu:

$$var\left[g(X)\right] \approx \left[\frac{\partial g(x)}{\partial x}\right]^2 var(X). \tag{1.9}$$

Vztah (1.9) je znám pod pojmem delta metoda nebo také jako aproximace rozptylu náhodné veličiny pomocí rozvoje Taylorovy řady. [2] Aplikací delta metody na (1.8) dostáváme následující:

$$\widehat{var}\left[\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right] \approx \left[\frac{1}{\ln \hat{S}(t)}\frac{1}{\hat{S}(t)}\right]^2 \left[\left[\hat{S}(t)\right]^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}\right] \tag{1.10}$$

$$\widehat{var}\left[\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right] \approx \left[\frac{1}{\ln \hat{S}(t)}\right]^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}. \tag{1.11}$$

Odhad rozptylu ve tvaru (1.11) lze následně využít pro sestavení $100(1-\alpha)\%$ intervalu spolehlivosti pro odhad funkce přežití. Při konstrukci tohoto intervalu vyjdeme z odpovídajícího intervalu spolehlivosti pro $\ln\left[-\ln\left(S(t)\right)\right]$, který má tvar:

$$I_{1-\alpha}\left[\ln\left[-\ln\left(S(t)\right)\right]\right] = \left\langle \ln\left[-\ln\left(\hat{S}(t)\right)\right] \pm u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}\left[\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right]}\right\rangle. \tag{1.12}$$

Označme si nyní dolní, resp. horní krajní hodnotu tohoto intervalu jako $\hat{c}_l$, resp. $\hat{c}_u$. Interval spolehlivosti pro funkci přežití $S(t)$ pak získáme tranformací těchto hodnot jako [7]:

$$I_{1-\alpha}\left(S(t)\right) = \langle exp\left\{-exp\left\{\hat{c}_u\right\}\right\}; exp\left\{-exp\left\{\hat{c}_l\right\}\right\}\rangle \tag{1.13}$$

Povšimnout si můžeme faktu, že ve výpočtu dolní meze intervalu vystupuje hodnota $\hat{c}_u$, která v původním intervalu spolehlivosti (1.12) představovala mez horní. Analogická situace platí i pro hodnotu $\hat{c}_l$. Tato záměna je způsobena násobením mínus jedničkou pro první tranformaci exponenciální funkcí.

Výše popsaným způsobem dostáváme odhadované hodnoty funkce přežití pro různá t (včetně odpovídajících intervalů spolehlivosti). Vykreslením těchto hodnot do grafu získáme tzv. *křivku přežití* (anglicky survival curve). Ačkoli výsledný tvar této křivky záleží na konkrétních datech, je křivka přežití vždy nerostoucí.

Ilustrujme si nyní odhad Kaplanovy-Meierovy křivky přežití na příkladě. V tabulce 1.1 jsou pro tyto účely uvedena fiktivní data. V prvním sloupečku je uvedeno číslo pozorování,

Pozorování	Čas	Událost
1	5	1
2	10	1
3	6	0
4	2	0
5	9	1
6	13	0
7	8	1
8	3	1
9	10	1
10	4	0

Tabulka 1.1: Ilustrativní data přežití

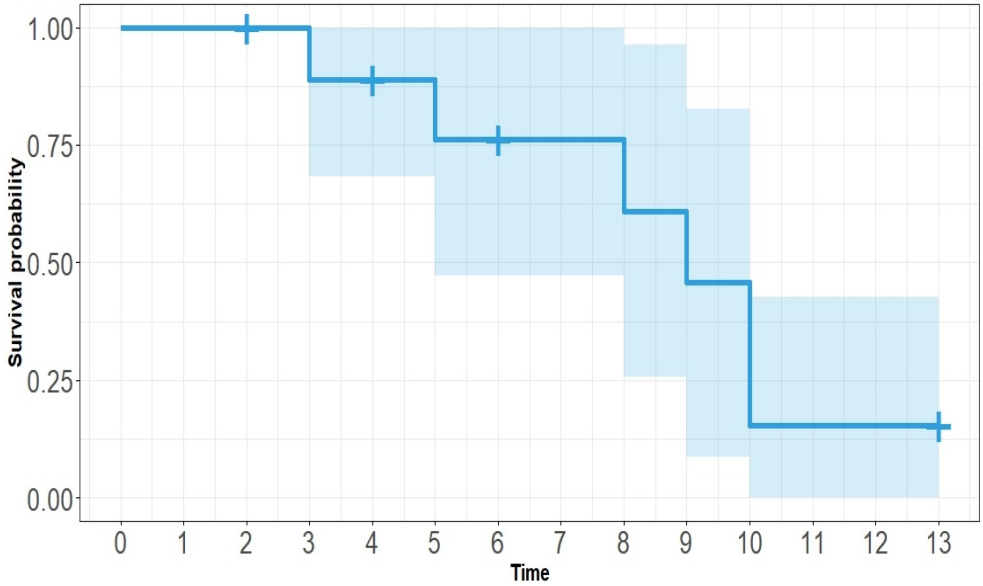
ve druhém sloupečku je pozorovaná doba přežití a ve třetím sloupečku je informace o tom,

zda došlo ke sledované události (1) nebo jde o cenzorované pozorování (0). Z tabulky 1.1 vidíme, že máme deset pozorování, přičemž čtyři z nich jsou cenzorovaná. V tabulce 1.2 pak máme údaje potřebné pro Kaplanův-Meierův odhad funkce přežití. V prvním sloupci jsou vzestupně uspořádané časy t_i, ve kterých došlo k události. Ve druhém sloupci jsou počty subjektů n_i, které jsou v daném čase t_i v riziku. Třetí sloupeček pak nese informaci o počtu událostí d_i, jež v daném čase nastaly. Ve čtvrtém sloupci jsou hodnoty $1 - d_i/n_i$. V předposledním sloupci jsou výsledné hodnoty Kaplanova-Meierova odhadu fuknce přežití v čase t_i a ve sloupci posledním jsou pak odpovídající hodnoty rozptylu (počítané dle vztahu (1.6)).

t_i	n_i	d_i	$1 - d_i/n_i$	$\hat{S}(t_i)$	$\widehat{var}\left[\hat{S}(t_i)\right]$
3	9	1	0.889	0.889	0.011
5	7	1	0.857	0.762	0.022
8	5	1	0.800	0.610	0.033
9	4	1	0.750	0.457	0.036
10	3	2	0.333	0.152	0.019

Tabulka 1.2: Výpočet Kaplanova-Meierova odhadu funkce přežití pro ilustrativní data

Výsledná křivka přežití je pak znázorněna na obrázku 1.3. Vidíme, že odhad funkce pře-



Obrázek 1.3: Kaplanova-Meierova křivka přežití pro ilustativní data

žití je roven jedné až do času $t = 3$, což je čas první události. Postupně pak klesá až do času $t = 10$, kdy došlo k události naposledy. Protože je nejdelší napozorovaný čas cenzorovaný, nedosahuje křivka přežití nuly. Cenzorovaná pozorování jsou na křivce označena symboly "+". V grafu je pro každý časový okamžik vykreslen také 95% interval spolehlivosti. Protože jsme rozptyl počítali pomocí Greewoodova vzorce, je tento interval spolehlivosti počítán pomocí odpovídajícího vztahu (1.7).

Pomocí křivky přežití můžeme též graficky porovnat časy přežití dvou nebo více skupin subjektů, a to vykreslením křivek přežití pro každou skupinu. Porovnat můžeme chtít například skupinu kuřáků a nekuřáků, mužů a žen či skupinu léčenou lékem A, lékem B a lékem C. Pokud bychom chtěli testovat hypotézu o rozdílnosti odhadovaných pravděpodobností přežití pro jednotlivé skupiny, pak lze použít log-rank test (viz podkapitola 4.1.4) [12], [13]

1.2.2. Hustota pravděpodobnosti

Hustota pravděpodobnosti (anglicky probability density function) vyjadřuje pravděpodobnost výskytu sledované události v určitém časovém intervalu na časové ose. V dalším textu ji budeme značit jako $f(t)$. Hustota pravděpodobnosti je definována jako:

$$f(t) = \lim_{\Delta t \to 0} \frac{P\left[t < T \leq t + \Delta t\right]}{\Delta t} \ . \tag{1.14}$$

Hustotu pravděpodobnosti lze získat derivací distribuční funkce $F(t)$ podle proměnné t (za předpokladu, že ve všech uvažovaných časech t tato derivace existuje). Platí tedy:

$$f(t) = \frac{dF(t)}{dt}, \ t \geq 0. \tag{1.15}$$

V praxi však nejsme schopni získat skutečné hodnoty těchto pravděpodobností, proto provádíme pouze odhad hustoty pravděpodobnosti. V případě, že nejsou přítomna cenzorovaná pozorování můžeme odhad provést následovně [12]:

$$\hat{f}(t) = \frac{\text{počet subjektů, u nichž došlo k události v časovém intervalu } (t + \Delta t)}{\text{celký počet subjektů } \cdot (\Delta t)}. \tag{1.16}$$

Protože vztah (1.16) nelze použít v přítomnosti cenzorovaných pozorování, představíme si zde jiný způsob odhadu hustoty. Využijeme k tomu Kaplanova-Meierova odhadu funkce přežití a uvažovat přitom budeme (vzestupně) uspořádané časy výskytů událostí, tj.: $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$. Nechť Δ_i označuje šířku i-tého časového intervalu, tj. $\Delta_i = t_{(i)} - t_{(i-1)}$, $i = 1, \ldots, n$. Časový okamžik odpovídající středu i-tého intervalu pak označme jako $t_{m(i)}$, tj. $t_{m(i)} = (t_{(i)} + t_{(i-1)})/2$. Odhad hustoty pravděpodobnosti je pak pravděpodobnost, že sledovaná událost nastane v i-tém časovém intervalu za jednotku času, tj. [9]:

$$\hat{f}(t_{m(i)}) = \frac{\hat{S}(t_{(i-1)}) - \hat{S}(t_{(i)})}{\Delta_i}. \tag{1.17}$$

Hustota pravděpodobnosti bývá označována také jako *nepodmíněná míra selhání* (anglicky unconditional failure rate).[12] Míra selhání vyjadřuje frekvenci (intenzitu), s jakou dochází k selhání (události) za časovou jednotku. Z grafu hustoty lze vyčíst, ve kterém časovém okamžiku je tato frekvenkce nejvyšší. Pokud by byla například sledovanou událostí smrt následkem autonehody, pak bychom očekávali, že nejvyšší intenzita nastání události bude pro časové okamžiky blízké času nehody. Naopak, pokud bychom sledovali smrt následkem rakovinného onemocnění, lze očekávat, že bude frekvence nastání události nejvyšší až po určitém čase, nikoli bezprostředně po stanovení diagnózy. Proporci nastání události v nějakém časovém intervalu pak získáme jako plochu pod křivkou v daném intervalu.

Na tomto místě ještě poznamenejme, že v případě diskrétní náhodné veličiny T neuvažujeme hustotu pravděpodobnosti, ale pravděpodobnostní funkci. Tato pravděpodobnostní funkce každému uvažovanému času t_i, kde $i = 1, 2, 3, \ldots$ přiřadí pravděpodobnost výskytu události, tedy: $p(t_i) = P(T = t_i)$, $i = 1, 2, 3, \ldots$.

1.2.3. Riziková funkce

Riziková funkce (anglicky hazard function), kterou budeme dále značit jako $h(t)$, je též známá pod pojmem podmíněná míra selhání (anglicky conditional failure rate). [9] Riziková funkce nám vyjadřuje intenzitu s níž dochází k výskytu události v čase t, ovšem za podmínky, že k události do času t nedošlo, což lze matematicky zapsat následujícím

způsobem:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left[t < T \leq t + \Delta t | T \geq t\right]}{\Delta t}. \tag{1.18}$$

V případě, že T je diskrétní náhodná veličina (uvažujeme diskrétní čas), můžeme rizikovou funkci pro časové okamžiky t_i, kde $i = 1, 2, 3, \ldots$ získat s využitím znalosti pravděpodobnostní funkce a funkce přežití jako [9]:

$$h(t_i) = P(T = t_i | T \geq t_i) = \frac{p(t_{(i)})}{S(t_{(i-1)})}, \qquad i = 1, 2, 3, \ldots, \tag{1.19}$$

přičemž platí, že $t_{(0)} = 0$ a $S(0) = 1$. Vztah (1.19) lze s využitím platnosti: $p(t_{(i)}) = S(t_{(i-1)}) - S(t_{(i)})$ přepsat do tvaru:

$$h(t_i) = 1 - \frac{S(t_{(i)})}{S(t_{(i-1)})}, \qquad i = 1, 2, 3, \ldots. \tag{1.20}$$

Nyní se podívejme na rizikovou funkci ve spojitém případě (T je spojitá náhodná veličina). S využitím hustoty pravděpodobnosti $f(t)$ a distribuční funkce $F(t)$, lze rizikovou funkci psát následovně:

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{1.21}$$

Za předpokladu znalosti funkce přežití lze s využitím platnosti $S(t) = 1 - F(t)$ zapsat vztah (1.21) ve tvaru:

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.22}$$

V praxi však nejsme nikdy schopni zjistit přesnou (skutečnou) hodnotu rizikové funkce a používáme pouze její odhad. V případě, že nejsou v datovém souboru přítomna cenzorovaná pozorování, lze odhad rizikové funkce provést následovně:

$$\hat{h}(t) = \frac{\text{počet subjektů, u nichž došlo k události v časovém intervalu } (t + \Delta t)}{\text{počet subjektů v riziku v čase } t \cdot (\Delta t)} \tag{1.23}$$

nebo také:

$$\hat{h}(t) = \frac{\text{počet subjektů, u nichž došlo k události v čase } t}{\text{počet subjektů v riziku v čase } t}. \tag{1.24}$$

Počtem subjektů v riziku je přitom myšlen počet subjektů, u nichž do času t nedošlo k události.

V případě přítomnosti cenzorovaných pozorování je nutné provést odhad rizikové funkce jiným způsobem. Za předpokladu konstantní rizikové funkce mezi časy nastání událostí, tj. obecně mezi časy $t_{(i)}$ a $t_{(i+1)}$, lze odhad rizikové funkce provést následovně [2]:

$$\hat{h}(t) = \frac{d_i}{n_i(t_{(i+1)} - t_{(i)})}, \qquad t_{(i)} \leq t < t_{(i+1)}, \tag{1.25}$$

kde d_i je počet událostí, jež nastaly v čase $t_{(i)}$ a n_i je počet pozorování v riziku v čase $t_{(i)}$. Poznamenejme, že podle vztahu (1.25) nelze provést odhad rizikové funkce pro časový interval, který začíná okamžikem nastání poslední události, neboť tento interval je zprava otevřený. Zároveň do okamžiku nastání první události je odhad rizikové funkce roven nule.

Odhad rozptylu pro $\hat{h}(t)$ získáme jako:

$$\widehat{var}\left[\hat{h}(t)\right] \approx \left[\hat{h}(t)\right]^2 \left(\frac{n_i - d_i}{n_i d_i}\right), \qquad t_{(i)} \leq t < t_{(i+1)}. \tag{1.26}$$

Pomocí rozptylu pak můžeme sestrojit $100(1-\alpha)\%$ intervaly spolehlivosti pro $h(t)$:

$$I_{1-\alpha}\left(h(t)\right) = \left\langle \hat{h}(t) - u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}\left[\hat{h}(t)\right]}; \hat{h}(t) + u_{1-\frac{\alpha}{2}}\sqrt{\widehat{var}\left[\hat{h}(t)\right]} \right\rangle, \tag{1.27}$$

kde $u_{1-\frac{\alpha}{2}}$ je $1-\frac{\alpha}{2}$ kvantil normovaného normálního rozdělení.

Odhad rizikové funkce si můžeme ilustrovat na příkladě. Uvažujme opět data z tabulky 1.1. V tabulce 1.3 jsou pak hodnoty pro odhad rizikové funkce a také odpovídajících směrodatných chyb. V prvním sloupci jsou časové intervaly pro které budeme odhady počítat.

Časový interval	Δ_i	n_i	d_i	$\hat{h}(t)$	$\widehat{var}\left[\hat{h}(t)\right]$
$\langle 0; 3)$	3	10	0	0	–
$\langle 3; 5)$	2	9	1	0.056	0.003
$\langle 5; 8)$	3	7	1	0.048	0.002
$\langle 8; 9)$	1	5	1	0.200	0.032
$\langle 9; 10)$	1	4	1	0.250	0.047

Tabulka 1.3: Výpočet odhadu rizikové funkce pro ilustrativní data

Ve druhém sloupci jsou délky odpovídajících časových intervalů (Δ_i). Následuje sloupec

s počtem pozorování v riziku (n_i) a sloupec s počtem událostí, jež v daném intervalu nastaly (d_i). V předposledním sloupci je příslušný odhad rizikové funkce a v posledním sloupci je odpovídající rozptyl.

Z tabulky 1.3 vidíme, že odhadované riziko v našem příkladě postupem času roste. Pouze ve třetím časovém intervalu ($5 \leq t < 8$) došlo oproti předchozímu intervalu k mírnému poklesu.

V praxi se kromě rizikové funkce využívá tzv. *kumulativní riziková funkce* (anglicky cumulative hazard function), která vyjadřuje celkové riziko výskytu sledované události po celou uvažovanou dobu od počátku až do času t. Kumulativní rizikovou funkci budeme značit $H(t)$. V diskrétním případě platí:

$$H(t) = \sum_{t_i \leq t} h(t_i), \tag{1.28}$$

ve spojitém případě pak platí následující:

$$H(t) = \int_0^t h(x)dx. \tag{1.29}$$

Pro praktický odhad kumulované rizikové funkce využijeme Kaplanova-Meierova odhadu funkce přežití. Nejdříve si vyjádříme rizikovou funkci pomocí funkce přežití. S využitím platnosti $f(t) = \frac{d}{dt}F(t)$ a $F(t) = 1 - S(t)$ lze psát:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\frac{dS(t)}{dt}}{S(t)} = -\frac{d}{dt}\ln S(t). \tag{1.30}$$

Odtud pak snadno získáme vztah mezi kumulativní rizikovou funkcí a funkcí přežití:

$$H(t) = \int_0^t h(t) = -\ln S(t). \tag{1.31}$$

Nakonec s využitím vztahu (1.5) získáme odhad kumulativní rizikové funkce ve tvaru:

$$\hat{H}(t) = -\ln \hat{S}(t) = -\ln \sum_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right). \tag{1.32}$$

Příslušný odhad rozptylu pak vypočítáme jako:

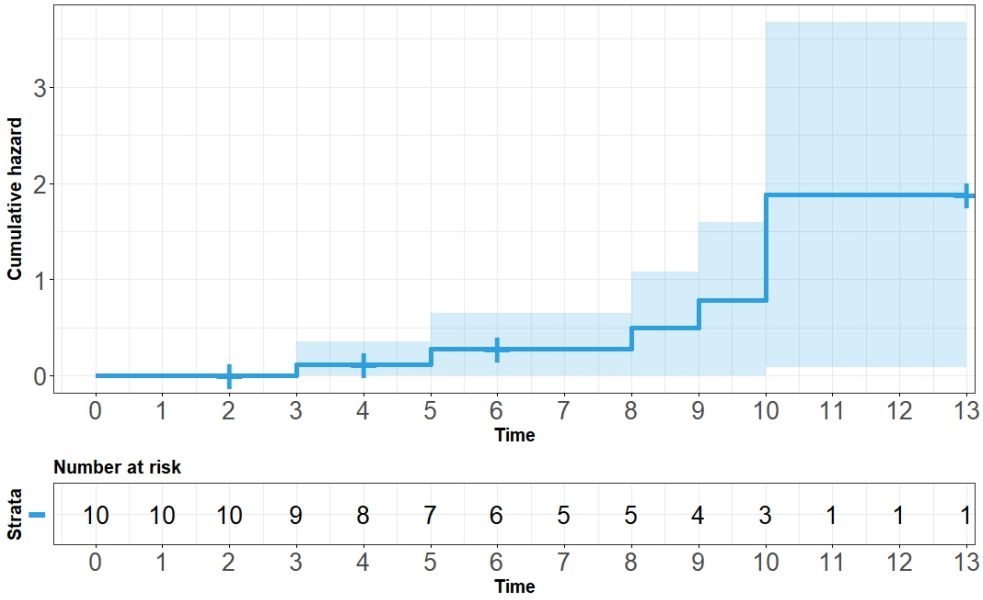
$$\widehat{var}\left[\hat{H}(t)\right] = \frac{\widehat{var}\left[\hat{S}(t)\right]}{\left[\hat{S}(t)\right]^2}. \tag{1.33}$$

Výpočet odhadu kumulované rizikové funkce si opět ilustrujme na příkladu dat z tabulky 1.1. Pro výpočet využijeme data z tabulky 1.2, kde máme vypočítané odhady funkce přežití a příslušné odhady rozptylu. Ve výsledné tabulce 1.4 tedy máme časy t_i, dále odhady funkce přežití $\hat{S}(t_i)$ a odpovídající rozptyly $\widehat{var}\left[\hat{S}(t_i)\right]$. Následuje sloupec s hodnotou odhadu kumulované rizikové funkce $\hat{H}(t_i)$. V poslední sloupci jsou pak odhady rozptylu $\widehat{var}\left[\hat{H}(t_i)\right]$

t_i	$\hat{S}(t_i)$	$\widehat{var}\left[\hat{S}(t_i)\right]$	$\hat{H}(t_i)$	$\widehat{var}\left[\hat{H}(t_i)\right]$
3	0.889	0.011	0.118	0.014
5	0.762	0.022	0.272	0.038
8	0.610	0.033	0.494	0.088
9	0.457	0.036	0.783	0.171
10	0.152	0.019	1.884	0.838

Tabulka 1.4: Výpočet kumulativní rizikové funkce pro ilustrativní data

Na obrázku 1.4 je pak pro náš příklad vykreslena křivka kumulativní rizikové funkce spolu s intervaly spolehlivosti v jednotlivých bodech. Opět vidíme, že nenulových hodnot



Obrázek 1.4: Křivka kumulativní rizikové funkce pro ilustrativní data

křivka nabývá až od třetího časového okamžiku, kdy nastala první událost. Nejvyšší nárůst

je pak sledován v desátém časovém okamžiku, což koresponduje s faktem, že v tomto jediném okamžiku nastaly hned dvě sledované události.

1.2.4. Výpočetní vztahy mezi základními funkcemi analýzy přežití

Jak již víme, základními funkcemi používanými v analýze přežití jsou funkce přežití, riziková funkce a hustota (ve spojitém případě) nebo pravděpodobnostní funkce (v diskrétním případě). Řekli jsme si také, že mezi jednotlivými funkcemi existují výpočetní vztahy a některé z nich jsme již dokonce představili. Nicméně pro přehlednost si v této části práce všechny tyto vztahy uvedeme a také odvodíme. [9], [12]

Začněme diskrétním případem. Jako výchozí výpočetní vztahy pro jednotlivé funkce uvažujeme následující:

- Pravděpodobnostní funkce: $p(t_i) = P(T = t_i)$
- Funkce přežití: $S(t_i) = P(T > t_i) = \sum_{t_i > t} p(t_i)$
- Riziková funkce: $h(t_i) = P(T = t_i | T \geq t_i) = \frac{p(t_{(i)})}{S(t_{(i-1)})}$

Nejprve předpokládejme, že známe pouze pravděpodobnostní funkci. Funkci přežití získáme přímo dle již výše uvedeného vztahu $S(t_i) = \sum_{t_i > t} p(t_i)$. Pro výpočet rizikové funkce pak lze využít funkci přežití a vztah $h(t_i) = \frac{p(t_{(i)})}{S(t_{(i-1)})}$, který jsme již také uvedli výše.

Dále předpokládejme, že známe pouze funkci přežití. Pravděpodobnostní funkci získáme následovně:

$$p(t_i) = S(t_{(i-1)}) - S(t_{(i)}). \tag{1.34}$$

Rizikovou funkci pak vypočítáme s využitím funkce přežití opět pomocí vztahu $h(t_i) = \frac{p(t_{(i)})}{S(t_{(i-1)})}$. Pokud bychom využili vztahu (1.34), lze psát:

$$h(t_i) = \frac{p(t_{(i)})}{S(t_{(i-1)})} = \frac{S(t_{(i-1)}) - S(t_{(i)})}{S(t_{(i-1)})} = 1 - \frac{S(t_{(i)})}{S(t_{(i-1)})}. \tag{1.35}$$

Nakonec předpokládejme znalost pouze rizikové funkce. Pro funkci přežití platí vztah $S(t) = \prod_{t_i \leq t} S(t_i)/S(t_{i-1})$. [9] S využitím tohoto vztahu a vztahu (1.35) lze psát:

$$S(t_i) = \prod_{t_i \leq t} \left[1 - h(t_i)\right].$$

Pravděpodobnostní funkci lze získat opět dle vztahu (1.34) nebo úpravou vztahu (1.35), tj.:

$$p(t_i) = h(t_i)S(t_{i-1})$$

Ve spojitém případě pak vycházíme z následujícího:

- Hustota pravděpodobnosti: $f(t) = \frac{dF(t)}{dt}$
- Funkce přežití: $S(t) = 1 - F(t)$
- Riziková funkce - spojitý případ: $h(t) = \frac{f(t)}{1-F(t)}$

Nejdříve předpokládejme, že známe pouze hustotu pravděpodobnosti. Pro získání funkce přežití nám stačí jednoduchá úprava původního vztahu pro výpočet této funkce:

$$S(t) = 1 - F(t) = 1 - P(T \leq t) = P(T > t) = \int_t^\infty f(x)dx.$$

Odtud pak už snadno získáváme rizikovou funkci jako:

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

Ve druhém případě předpokládejme, že známe pouze funkci přežití. Vztah pro výpočet hustoty pravděpodobnosti získáme tak, že si nejprve ze vztahu pro výpočet funkce přežití vyjádříme distribuční funkci. Hustotu pravděpodobnosti poté získáme jako derivaci tohoto vztahu, tedy:

$$S(t) = 1 - F(t) \rightarrow F(t) = 1 - S(t),$$

$$f(t) = \frac{d}{dt}\left[1 - S(t)\right] = -\frac{d}{dt}S(t).$$

Rizikovou funkci poté můžeme opět určit s využitím znalosti hustoty pravděpodobnosti a funkce přežití jako:

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

Rizikovou funkci však lze z funkce přežití vypočítat i bez znalosti hustoty. Platí totiž:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\frac{dS(t)}{dt}}{S(t)} = -\frac{d}{dt}\ln S(t).$$

Poslední případ, který lze uvažovat je ten, kdy známe pouze rizikovou funkci. Nejprve si zde odvodíme vztah pro výpočet funkce přežití. Využijeme přitom předchozího odvozeného vztahu mezi funkcí přežití a rizikovou funkcí. Tento vztah budeme integrovat od nuly do t, čímž dostaneme následující:

$$-\int_0^t h(x)dx = \ln S(t).$$

Pro získání výsledného vztahu už nyní stačí využít exponenciální funkce:

$$S(t) = exp\left[-\int_0^t h(x)dx\right].$$

Vztah pro výpočet hustoty pravděpodobnosti získáme tak, že si ji vyjádříme z rovnosti $h(t) = \frac{f(t)}{S(t)}$ a dosadíme právě odvozený vztah pro výpočet funkce přežití. Tedy:

$$f(t) = h(t)S(t) = h(t)exp\left[-\int_0^t h(x)dx\right].$$

1.2.5. Průměrná doba přežití, průměrná doba dožití a medián doby přežití

Dalšími charakteristikami v analýze přežití jsou průměrná doba přežití, průměrná doba dožití a medián. [9] Začněme nejprve průměrnou dobou přežití (anglicky mean survival time), která nepředstavuje nic jiného než střední hodnotu náhodné veličiny T a budeme ji značit jako μ. V diskrétním případě (T je diskrétní náhodná veličina, která se realizuje hodnotami $t_i, i = 1, 2, 3, \ldots, n$ s pravděpodobnostmi $p(t_i)$) je výpočetní vztah pro průměrnou dobu přežití následující:

$$\mu = E(T) = \sum_i^n t_i p(t_i). \tag{1.36}$$

Ve spojitém případě, je výpočetní vztah pro střední dobu přežití v následujícím tvaru:

$$\mu = E(T) = \int_0^\infty t f(t)dt. \tag{1.37}$$

Důležité je poznamenat, že výpočet průměrného času přežití je ovlivněn pozorováním s nejdelším časem. Je-li nejdelší pozorovaný čas necenzorovaný, pak platí výše uvedené

vztahy a navíc lze průměrný čas přežití vypočítat také s využitím funkce přežití. Nechť $t_{(i)}$ $(i = 1, 2, \ldots, s)$ jsou vzestupně uspořádané časy výskytu události (uvažujeme celkem s necenzorovaných pozorování), pak lze průměrnou dobu přežití odhadnout jako:

$$\hat{\mu} = \sum_{i=1}^{s} \hat{S}(t_{(i-1)})(t_{(i)} - t_{(i-1)}), \tag{1.38}$$

kde $t_{(0)} = 0$. Jestliže ale bude nejdelší pozorovaný čas cenzorovaný, pak lze počítat pouze tzv. restringovaný (omezený) průměrný čas přežití (budeme značit $\hat{\mu}_{RES}$). [2] Tato charakteristika představuje průměrný čas přežití od začátku studie do nějakého (námi zvoleného) času. Tímto časem může být například nejdelší pozorovaný čas výskytu události, tj. nejdelší necenzorovaný čas, nicméně zvolit můžeme i jakýkoli jiný čas. Označíme-li si tento mezní čas jako τ, pak je vztah pro výpočet restringovaného průměrného času přežití ve spojitém případě ve tvaru:

$$\hat{\mu}_{RES}(\tau) = \int_0^{\tau} \hat{S}(t)dt, \tag{1.39}$$

a v diskrétním případě ve tvaru:

$$\hat{\mu}_{RES}(\tau) = \sum_{i=1}^{s^*} \hat{S}(t_{(i-1)})(t_{(i)} - t_{(i-1)}) + \hat{S}(t_{(s^*)})(\tau - t_{(s^*)}), \tag{1.40}$$

kde s^* je počet pozorovaných časů výskytu sledované události do času τ.

Nyní se podívejme na průměrnou dobu dožití (anglicky mean residual life –MRL), kterou budeme značit μ_t. Tato charakteristika nám říká, kolik času zbývá, než dojde k události, jestliže víme, že k ní nedošlo alespoň do času t. V diskrétním případě je možné průměrnou dobu dožití pro každý časový okamžik vypočítat dle následujícího vzorce:

$$\mu_t = \frac{(t_{(i+1)} - t)S(t_{(i)}) + \sum_{j \geq i+1}(t_{(j+1)} - t_{(j)})S(t_{(j)})}{S(t)}, \quad t_{(i)} \leq t < t_{(i+1)}. \tag{1.41}$$

Ze vztahu (1.41) vidíme, že průměrný čas dožití lze počítat pouze pro časy, které předcházejí nejdelšímu (poslednímu) pozorovanému času výskytu události. Ve spojitém případě pak dostáváme:

$$\mu_t = E(T - t|T > t) = \frac{\int_t^{\infty} (x - t)\, f\,(x)\, dx}{S\,(t)} = \frac{\int_t^{\infty} S\,(x)\, dx}{S(t)}. \tag{1.42}$$

Poznamenejme ještě, že mezi průměrným časem přežití a průměrným časem dožití platí vztah $\mu = \mu_0$, čili průměrný čas přežití je shodný s průměrným časem dožití na začátku studie. V případě restringovaného průměrného času přežití, je této rovnosti dosaženo pouze zvolíme-li interval pro výpočet μ_{RES} od začátku studie do času posledního výskytu události (tj. do času $t_{(s^*)}$).

Třetí a zároveň poslední charakteristikou, kterou si představíme v této podkapitole, je medián doby přežití (anglicky median survival time). [2] Medián doby přežití (budeme značit $t_{0,5}$) představuje časový okamžik, ve kterém u 50 % subjektů ze studie nedošlo ke sledované události. Pokud bychom tedy považovali za sledovanou událost smrt následkem konkrétní nemoci, mohli bychom říct, že v čase $t_{0,5}$ je stále 50 % sledovaných pacientů naživu. Medián doby přežití zjišťujeme z funkce přežití a vyjádřit jej můžeme následovně:

$$S(t_{0.5}) = 0.5. \tag{1.43}$$

Uvažujeme-li diskrétní případ (funkce přežití je schodovitá), pak medián doby přežití odpovídá nejmenšímu času t, pro který platí, že $S(t) \leq 0.5$. Problém nastává v případě, kdy je více jak polovina pozorovaných časů přežití cenzorovaných a je zároveň cenzorovaný i nejdelší čas přežití. V takovém případě totiž není možné odhadovat medián času přežití. [12]

Nyní si tyto tři charakteristiky vypočítejme pro fiktivní data, která máme v tabulce 1.1. Začněme průměrnou dobou přežití. Protože je nejdelší pozorovaný čas cenzorovaný, využijeme k výpočtu vztah (1.40) a namísto průměrného času přežití budeme počítat restringovaný průměrný čas přežití. Toto provedeme pro dobu od začátku studie až do času $t = 10$ (čas výskytu poslední události). Hodnoty funkce přežití v časech výskytů událostí máme v tabulce 1.2 a výsledek tak získáme jednoduše jako následující součet:

$$\hat{\mu}_{RES}(10) = 1(3-0) + 0.889(5-3) + 0.762(8-5) + 0.610(9-8) + 0.457(10-9) = 8.131.$$

Dále si spočítejme průměrný čas dožití. Využijeme přitom vztah (1.41) a také hodnoty funkce přežití z tabulky 1.2. Je zřejmé, že různým časovým okamžikům t odpovídají různé hodnoty μ_t (jedná se o spojitou klesající funkci). My si zde ale vypočítáme průměrný čas dožití pouze pro (celé) časové okamžiky $t = 1, 2, \ldots, 9$. Pro ověření platnosti vztahu mezi

(restringovanou) průměrnou dobou přežití a průměrnou dobou dožití si tuto charakteristiku vypočítáme také v čase $t = 0$. Tímto časem můžeme ostatně začít:

$$\hat{\mu}(0) = (3-0)1 + [(5-3)0.889 + (8-5)0.762 + (9-8)0.610 + (10-9)0.457] = 8.131.$$

Z výsledku vidíme, že opravdu platí vztah $\hat{\mu}_{RES}(t_{(s^*)}) = \hat{\mu}_0$. Analogicky lze vypočítat průměrný čas dožití i pro další časové okamžiky t. Výsledné hodnoty jsou uvedeny v tabulce 1.5.

t	0	1	2	3	4	5	6	7	8	9
$\hat{\mu}(t)$	8.13	7.13	6.13	5.77	4.77	4.40	3.40	2.40	1.75	1.00

Tabulka 1.5: Průměrné doby dožití pro ilustrativní data.

Nakonec zbývá určit medián času přežití. Ten lze jednoduše určit jako časový okamžik, ve kterém je odhad funkce přežití roven hodnotě 0.5, resp. jako nejmenší čas t, pro který $\hat{S}(t) \leq 0.5$. V našem případě je (dle tabulky 1.2) mediánem doby přežití čas $t = 9$.

Kapitola 2

Coxův model proporcionálních rizik

Jak již bylo řečeno v úvodu této práce, v praxi nejpoužívanějším modelem v analýze přežití je jednoznačně Coxův model proporcionálních rizik. Kromě výhody jednoduchého použití a snadné interpretace výsledků však s sebou přináší i nevýhody ve formě striktních předpokladů, které jsou kladeny na analyzovaná data. Coxův model proporcionálních rizik v analýze přežití výborně poslouží, jsou-li splněny tyto základní předpoklady: je zachována proporcionalita rizik, proměnné jsou nezávislé na čase, sledovaná událost může pro každý subjekt nastat nejvýše jedenkrát a časy do nastání událostí různých subjektů jsou nekorelované. [12] Ačkoli se tato práce zabývá případy, kdy je předpoklad proporcionality porušen a použití Coxova modelu proporcionálních rizik není možné, tento model a s ním spjatou teorii si zde představíme. Učiníme tak zejména z toho důvodu, že modely pro neproporcionální rizika vycházejí právě z Coxova modelu proporcionálních rizik. V této kapitole se také naučíme ověřit předpoklad proporcionality rizik.

2.1. Sestavení Coxova modelu

Coxův model proporcionálních rizik lze definovat pomocí rizikové funkce. Pro i-tý subjekt ($i = 1, 2, \ldots, n$) je tato riziková funkce ve tvaru:

$$h(t, \mathbf{x}_i) = h_i(t) = h_0(t) exp\left\{x_{i1}\beta_1 + \ldots + x_{ik}\beta_k\right\} = h_0(t) exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\}, \tag{2.1}$$

kde $h_0(t)$ je hodnota základní rizikové funkce v čase t, x_{ij} je hodnota j-té vysvětlující proměnné ($j = 1, 2, \ldots, k$) pro i-tý subjekt a β_j jsou regresní parametry. [2], [9], [12]

Pro interpretaci regresních parametrů β_j si nejprve upravme vztah (2.1). Použitím logaritmické transformace dostáváme následující:

$$\ln h_i(t) = \ln h_0(t) + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k. \tag{2.2}$$

Ze vztahu (2.2) vidíme, že regresní parametr β_j $(j = 1, 2, \ldots, k)$ lze interpretovat jako hodnotu, o kterou se změní přirozený logaritmus rizikové funkce funkce při jednotkové změně j-té proměnné, a to za předpokladu, že hodnoty zbylých $(k-1)$ proměnných zůstanou nezměněné. Je-li hodnota regresního parametru β_j kladná, pak j-tou vysvětlující proměnnou označíme jako špatný (nepříznivý) prognostický faktor, neboť vyšší hodnoty této proměnné znamenají vyšší riziko výskytu sledované události. Je-li naopak hodnota regresního parametru β_j záporná, pak mluvíme o dobrém (příznivém) prognostickém faktoru (vyšší hodnoty dané proměnné znamenají nižší riziko). [12]

Alternativně lze pro interpretaci regresních parametrů β_j využít poměr rizikových funkcí pro dva různé subjekty [2]:

$$HR = \frac{h(t, \mathbf{x}_s)}{h(t, \mathbf{x}_r)} = \frac{h_0(t) exp\left\{x_{s1}\beta_1 + \ldots + x_{sk}\beta_k\right\}}{h_0(t) exp\left\{x_{r1}\beta_1 + \ldots + x_{rk}\beta_k\right\}} = \frac{exp\left\{\mathbf{x}_s'\boldsymbol{\beta}\right\}}{exp\left\{\mathbf{x}_r'\boldsymbol{\beta}\right\}}. \tag{2.3}$$

Vztah (2.3) nazýváme poměr rizik (anglicky hazard ratio). Úpravou získáváme vztah:

$$HR = exp\left\{\beta_1\left(x_{s1} - x_{r1}\right) + \ldots + \beta_k\left(x_{sk} - x_{rk}\right)\right\}. \tag{2.4}$$

Logaritmováním vztahu (2.4) pak dostaneme:

$$\ln\left[HR\right] = \beta_1\left(x_{s1} - x_{r1}\right) + \ldots + \beta_k\left(x_{sk} - x_{rk}\right). \tag{2.5}$$

Pokud bychom uvažovali dva subjekty, u nichž se hodnota j-té vysvětlující proměnné liší o jednotku a hodnoty zbývajících $(k-1)$ vysvětlujících proměnných jsou shodné, potom bychom mohli psát:

$$\ln\left[HR\right] = \beta_j. \tag{2.6}$$

Odtud ihned vidíme další možnou interpretaci regresních parametrů. Lze říci, že parametr β_j, $j = 1, \ldots, k$, představuje logaritmus podílu rizik pro dva subjekty, jejichž hodnoty všech vysvětlujících proměnných jsou stejné, až na hodnoty j-té proměnné, které se liší

o jednotku. V případě, že sledovaná j-tá proměnná bude kvantitativní, pak β_j udává hodnotu logaritmu poměru rizik při jednotkové změně dané proměnné. Bude-li j-tá proměnná kvalitativní, pak β_j vyjadřuje logaritmus poměru rizik uvažované kategorie dané proměnné vůči kategorii referenční.

Spíše než hodnoty logaritmovaných poměrů rizik nás v praxi zajímají samotné hodnoty poměrů rizik. Tyto hodnoty získáme aplikací exponenciální funkce na vztah (2.6):

$$HR = exp\left\{\beta_j\right\}. \tag{2.7}$$

Je-li hodnota HR vyšší než jedna, potom vyšší hodnoty dané proměnné znamenají vyšší riziko výskytu sledované události (ve spojitém případě), resp. vyšší riziko výskytu události oproti referenční katerogii (v diskrétním případě). Naopak hodnota HR nižší než jedna znamená, že s vyššími hodnotami dané proměnné je spojeno nižší riziko výskytu události (spojitý případ), resp. že má uvažovaná kategorie nižší riziko výskytu události než kategorie referenční (diskrétní případ). Je-li $HR = 1$, potom nemá danná proměnná na riziko výskytu události vliv.

Na tomto místě si ještě uveďme, jak lze z rizikové funkce, kterou máme ve tvaru (2.1), vyjádřit kumulovanou rizikovou funkci a funkci přežití. [2] Začněme kumulativní rizikovou funkcí $H(t, \mathbf{x}_i)$, která pro i-tý subjekt udává celkové riziko výskytu sledované události od začátku sledování až do času t. Využijeme-li vztahu (1.29), dostaneme:

$$H(t, \mathbf{x}_i) = H_i(t) = exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} \int_0^t h_0(s)ds = exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} H_0(t). \tag{2.8}$$

Ze vztahu (1.31), který platí mezi funkcí přežití a kumulativní rizikovou funkcí dostáváme:

$$S(t) = exp\left[-H(t)\right]. \tag{2.9}$$

Odtud pak snadno vyjádříme funkci přežití pro i-tý subjekt:

$$S(t, \mathbf{x}_i) = exp\left[-H_i(t)\right] = exp\left[-exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} H_0(t)\right] = exp\left[-H_0(t)\right]^{exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\}}. \tag{2.10}$$

Označíme-li si $exp\left[-H_0(t)\right] = S_0(t)$, pak lze vztah (2.10) přepsat do tvaru:

$$S(t, \mathbf{x}_i) = S_i(t) = S_0(t)^{exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\}}. \tag{2.11}$$

$S_0(t)$ přitom představuje základní funkci přežití, která je totožná pro všechny subjekty. Poznamenejme ještě, že hodnota funkce přežití je pro všechny subjekty v čase $t = 0$ rovna jedné, tj. $S_i(0) = 1$, $i = 1, 2, \ldots, n$.

2.2. Odhady regresních parametrů metodou maximální věrohodnosti

V této části práce se naučíme, jak odhadnout parametry β_j $(j = 1, 2, \ldots, k)$ Coxova modelu proporcionálních rizik, který je popsán rizikovou funkcí ve tvaru (2.1). Pro odhad těchto parametrů je využívána metoda maximální věrohodnosti (anglicky maximum-likelihood estimation, zkráceně MLE). [2], [9], [12]

Uvažujme celkem n nezávislých pozorování, přičemž u s pozorování $(s \leq n)$ došlo ke sledované události. Zbývajících $(n - s)$ pozorování je cenzorovaných zprava. Dále předpokládejme, že v jednom časovém okamžiku mohlo dojít ke sledované události jen u jednoho pozorování. Z toho plyne, že časy výskytů událostí lze uspořádat: $t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(s)}$. Metoda maximální věrohodnosti hledá takové odhady parametrů, které maximalizují parciální funkci věrohodnosti. V tomto případě je parciální funkce věrohodnosti ve tvaru:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{s} \frac{exp\left\{\sum_{j=1}^{k} \beta_j x_{(i)j}\right\}}{\sum_{l \in R(t_{(i)})} exp\left\{\sum_{j=1}^{k} \beta_j x_{lj}\right\}}, \tag{2.12}$$

kde $R(t_{(i)})$ je riziková skupina, která obsahuje všechna pozorování, u nichž do času $t_{(i)}$ nedošlo k události a jsou stále ve studii (nejsou do tohoto času cenzorovaná). V čitateli je tedy obsaženo pozorování u něhož došlo v čase $t_{(i)}$ ke sledované události a ve jmenovateli sčítáme hodnoty pro pozorování, která jsou v témže čase v riziku.

Zavedením indikátorové proměnné δ_i, $i = 1, \ldots, n$, která bude nabývat hodnot nula nebo jedna, přičemž hodnoty nula bude nabývat v případě, že jde o zprava cenzorované pozorování a jinak bude nabývat hodnoty jedna, lze vztah (2.12) přepsat do tvaru:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{exp\left\{\sum_{j=1}^{k} \beta_j x_{ij}\right\}}{\sum_{l \in R(t_i)} exp\left\{\sum_{j=1}^{k} \beta_j x_{lj}\right\}} \right]^{\delta_i}. \tag{2.13}$$

Z (2.13) vidíme, že zavedením indikátorové proměnné jsme se zbavili nutnosti uspořádání pozorování dle jejich časů přežití od nejnižšího po nejvyšší.

Protože maximalizace parciální funkce věrohodnosti v tomto tvaru by byla příliš složitá,

používá se přechod k *logaritmické parciální věrohodnostní funkci*, která je ve tvaru:

$$l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\sum_{j=1}^{k} \beta_j x_{ij} - \ln \sum_{l \in R(t_i)} exp \left\{ \sum_{j=1}^{k} \beta_j x_{lj} \right\} \right]. \tag{2.14}$$

Odhady parametrů $\hat{\beta}_j$ $(j = 1, 2, \ldots, k)$ pak získáme jako řešení následujícího systému simultánních rovnic:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = 0, \qquad j = 1, 2, \ldots, k. \tag{2.15}$$

Jednotlivé rovnice z (2.15) získáme jako:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{s} \left[x_{(i)j} - \frac{\sum_{l \in R(t_i)} x_{lj} exp \left\{ \sum_{j=1}^{k} \beta_j x_{lj} \right\}}{\sum_{l \in R(t_i)} exp \left\{ \sum_{j=1}^{k} \beta_j x_{lj} \right\}} \right]. \tag{2.16}$$

Řešení systému rovnic (2.15) se následně hledá pomocí iteračních metod, nejčastěji pomocí *Newtonovy-Raphsonovy metody* [2], [12].

Předpokládejme nyní, že jsme již odhady $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)'$ nalezli. Kromě těchto bodových odhadů nás ale v praxi zajímají také odhady intervalové. Pro sestrojení $100(1-\alpha)\%$ intervalů spolehlivosti pro odhady $\hat{\beta}_j$ je třeba znalosti odpovídajích rozptylů, resp. směrodatných chyb. Varianční matici příslušející odhadnutému vektoru parametrů $\hat{\boldsymbol{\beta}}$ získáme jako [2]:

$$\widehat{var}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \left(\left[-\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \beta_u \partial \beta_w} \right]^{-1} \right)_{u,w=1}^{k}. \tag{2.17}$$

Prvky varianční matice $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ budeme dále značit jako $\hat{v}_{uw}$ $(u, w = 1, 2, \ldots, k)$. Označíme-li si $z = \sum_{j=1}^{k} \beta_j x_{lj}$, pak hodnoty druhých parciálních derivací ze vztahu (2.17) získáme jako:

$$\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \beta_u \partial \beta_w} = -\sum_{i=1}^{s} \frac{\sum_{l \in R(t_i)} x_{lu} x_{lw} exp\{z\}}{\sum_{l \in R(t_i)} exp\{z\}} - \frac{\sum_{l \in R(t_i)} x_{lu} exp\{z\}}{\sum_{l \in R(t_i)} exp\{z\}} \frac{\sum_{l \in R(t_i)} x_{lw} exp\{z\}}{\sum_{l \in R(t_i)} exp\{z\}}. \tag{2.18}$$

Nyní jsme již schopni získat $100(1-\alpha)\%$ intervaly spolehlivosti pro $\hat{\beta}_j$ $(j = 1, 2, \ldots, k)$, a to ve tvaru:

$$I_{1-\alpha}(\beta_j) = \left\langle \hat{\beta}_j - u_{1-\frac{\alpha}{2}} \sqrt{\hat{v}_{jj}}; \hat{\beta}_j + u_{1-\frac{\alpha}{2}} \sqrt{\hat{v}_{jj}} \right\rangle. \tag{2.19}$$

Obsahuje-li takto sestrojený interval spolehlivosti nulu, znamená to, že opovídající proměnná nebude pro daný model statisticky významná. Testům hypotéz o významnosti odhadnutých parametrů, resp. proměnných v modelu, se budeme dále podrobněji věnovat v kapitole 4.

2.3. Odhad základní rizikové funkce

Kromě znalosti odhadu regresních parametrů β_j $(j = 1, 2, \ldots, k)$, je pro Coxův model proporcionálních rizik ve tvaru (2.1) nutná také znalost odhadu základní rizikové funkce $h_0(t)$. Jak tento odhad získat si nyní ukážeme. [2], [12]

Stejně jako při odhadu regresních parametrů předpokládejme, že máme k dispozici n nezávislých pozorování, přičemž u s, $s \leq n$, z nich nastala sledovaná událost. Časy výskytu sledovaných událostí lze uspořádat: $t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(s)}$. Nyní navíc předpokládejme také znalost odhadu regresních parametrů $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)'$. Odhad základní rizikové funkce v čase nastání události $t_{(j)}$ je dán jako [2]:

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\xi}_j, \qquad j = 1, 2, \ldots, s, \tag{2.20}$$

kde $\hat{\xi}_j$ je řešením rovnice:

$$\sum_{i \in D(t_{(j)})} \frac{exp\left\{\mathbf{x}_i' \hat{\boldsymbol{\beta}}\right\}}{1 - \hat{\xi}_j^{exp\left\{\mathbf{x}_i' \hat{\boldsymbol{\beta}}\right\}}} = \sum_{i \in R(t_{(j)})} exp\left\{\mathbf{x}_i' \hat{\boldsymbol{\beta}}\right\}, \tag{2.21}$$

kde $j = 1, 2, \ldots, s$, $D(t_{(j)})$ označuje skupinu všech pozorování, u nichž došlo k události v čase $t_{(j)}$ a $R(t_{(j)})$ je skupina všech pozorování, která jsou v čase $t_{(j)}$ v riziku, jinými slovy u kterých do tohoto času včetně k události nedošlo. Hodnotu $\hat{\xi}_j$ můžeme přitom interpretovat jako odhad pravděpodobnosti, že u subjektu během časového intervalu od $t_{(j)}$ do $t_{(j+1)}$ nedojde ke sledované události. Řešení rovnice (2.21) je složité, nicméně za předpokladu, že v každém časovém okamžiku může sledovaná událost nastat pouze u jednoho subjektu (tj. v levé části rovnice budeme vždy uvažovat pouze jedno pozorování, a to takové, u něhož v j-tém čase došlo ke sledované události), můžeme řešení vyjádřit celkem snadno. Za uvedené

podmínky uvažujeme tedy rovnici ve tvaru:

$$\frac{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}{1-\hat{\xi}_j^{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}} = \sum_{i\in R(t_{(j)})} exp\left\{\mathbf{x}'_i\hat{\boldsymbol{\beta}}\right\}, \tag{2.22}$$

kde $\mathbf{x}_{(j)}$ označuje vektor hodnot vysvětlující proměnné pro subjekt, u něhož došlo k události v čase $t_{(j)}$. Rovnici (2.22) nejprve vynásobíme jmenovatelem z levé části rovnosti tj.:

$$exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\} = \sum_{i\in R(t_{(j)})} exp\left\{\mathbf{x}'_i\hat{\boldsymbol{\beta}}\right\} - \hat{\xi}_j^{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}\left[\sum_{i\in R(t_{(j)})} exp\left\{\mathbf{x}'_i\hat{\boldsymbol{\beta}}\right\}\right]. \tag{2.23}$$

Dále odečteme od rovnice člen $\sum_{i\in R(t_{(j)})} exp\left\{\mathbf{x}'_i\hat{\boldsymbol{\beta}}\right\}$, stejným členem celou rovnici vydělíme a následně ještě vynásobíme mínus jedničkou, čímž dostáváme následující:

$$1-\frac{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}{\sum_{i\in R(t_{(j)})} exp\left\{\mathbf{x}'_i\hat{\boldsymbol{\beta}}\right\}} = \hat{\xi}_j^{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}. \tag{2.24}$$

Nakonec celou rovnost umocníme na $\frac{1}{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}$ a tím získáme výsledný odhad parametru $\hat{\xi}_j$ ve tvaru:

$$\hat{\xi}_j = \left[1-\frac{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}{\sum_{i\in R(t_{(j)})} exp\left\{\mathbf{x}'_i\hat{\boldsymbol{\beta}}\right\}}\right]^{exp\left\{-\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}}, \tag{2.25}$$

přičemž jsme zde využili platnosti: $\frac{1}{exp\left\{\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}} = exp\left\{-\mathbf{x}'_{(j)}\hat{\boldsymbol{\beta}}\right\}$. Zatím jsme ale získali odhad základní rizikové funkce pouze v časech, ve kterých došlo ke sledované události. Hodnota rizikové funkce mezi dvěma nejbližšími časy událostí je pak (za předpokladu konstantního rizika v daném časovém intervalu) dána vztahem [2]:

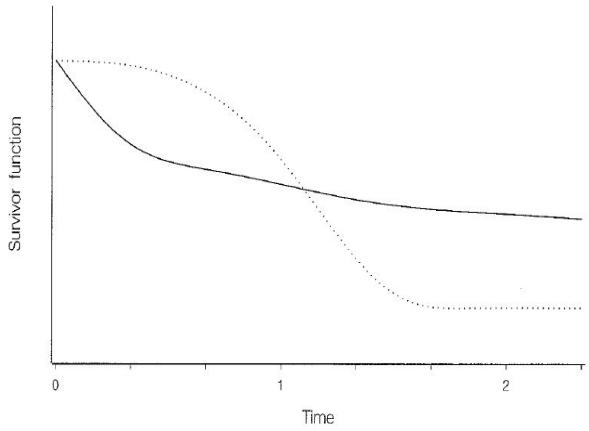
$$\hat{h}_0(t) = \frac{1-\hat{\xi}_j}{t_{(j+1)}-t_{(j)}}, \quad t_{(j)} \le t < t_{(j+1)}, \tag{2.26}$$

kde $j = 1, 2, \ldots, s-1$, přičemž platí: $\hat{h}_0(t) = 0$ pro všechna $t < t_{(1)}$.

V případě nesplnění podmínky, která říká, že v jednom časovém okamžiku může nastat sledovaná událost nejvýše u jednoho subjektu, nelze rovnici (2.20) řešit explicitně a je nutné použít iterační postup. [2]

2.4. Proporcionalita rizik

Dříve, než se začneme zabývat modely neproporcionálních rizik, podívejme se na to, co vlastně proporcionalita rizik znamená a jak lze ověřit její (ne)splnění. Proporcionalita rizik je hlavním předpokladem pro použití základního Coxova modelu – v případě jejího nesplnění je třeba použít některý z modelů neproporcionálních rizik. Tento předpoklad říká, že poměr rizik dvou skupin subjektů je v čase konstantní. Nejjednodušším způsobem ověření proporcionality je vykreslení křivek přežití. Pokud se křivky po celou dobu sledování rovnoměrně rozcházejí, je to znak proporcionality. V případě, že tomu tak není nebo se křivky v průběhu doby dokonce kříží, potom jsou rizika neproporcionální. [2] Na obrázku 2.1 je příklad toho, jak mohou vypadat vykreslené křivky přežití v případě porušení předpokladu proporcionality.



Obrázek 2.1: Ukázka neproporcionálních rizik na křivkách přežití [2]

Posouzení splnění předpokladu proporcionality dle právě uvedené metody však není vždy zcela přesné a dostačující. Představme si proto některé další metody. První metoda je opět založena na grafickém ověřování, a sice pomocí grafu logaritmované kumulativní rizikové funkce. Riziková funkce v jakémkoli čase t je dle Coxova modelu proporcionálních rizik pro i-tý subjekt dána jako [2]:

$$h_i(t) = exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} h_0(t), \tag{2.27}$$

kde $\mathbf{x}_i$ je vektor hodnot vysvětlujících proměnných pro i-tý subjekt, $\boldsymbol{\beta}$ je vektor regresních koeficientů a $h_0(t)$ je základní riziková funkce, která je stejná pro všechny subjekty. Základní

riziková funkce $h_0(t)$ představuje riziko nastání události, pokud bychom nebrali v úvahu žádné vysvětlující proměnné. [12] Nyní provedeme několik úprav tohoto vztahu. Začneme tím, že budeme obě strany rovnosti integrovat přes t:

$$\int_0^t h_i(u)du = exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} \int_0^t h_0(u)du. \tag{2.28}$$

S využitím platnosti vztahu (1.29) můžeme vztah (2.28) přepsat do tvaru:

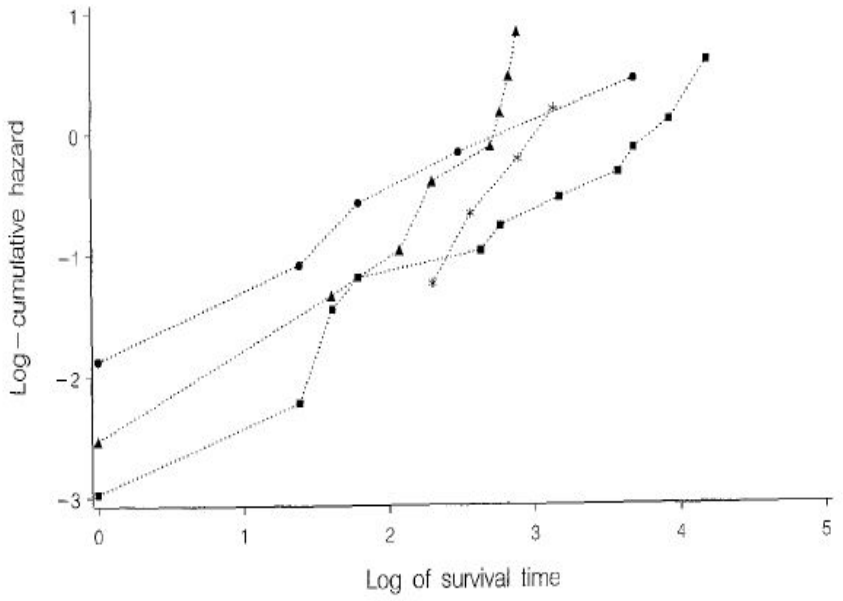
$$H_i(t) = exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} H_0(t), \tag{2.29}$$

kde $H_i(t)$ a $H_0(t)$ jsou kumulativní rizikové funkce. Nyní zbývá vztah (2.29) logaritmovat, čímž dostáváme:

$$\ln H_i(t) = \mathbf{x}_i'\boldsymbol{\beta} + \ln H_0(t). \tag{2.30}$$

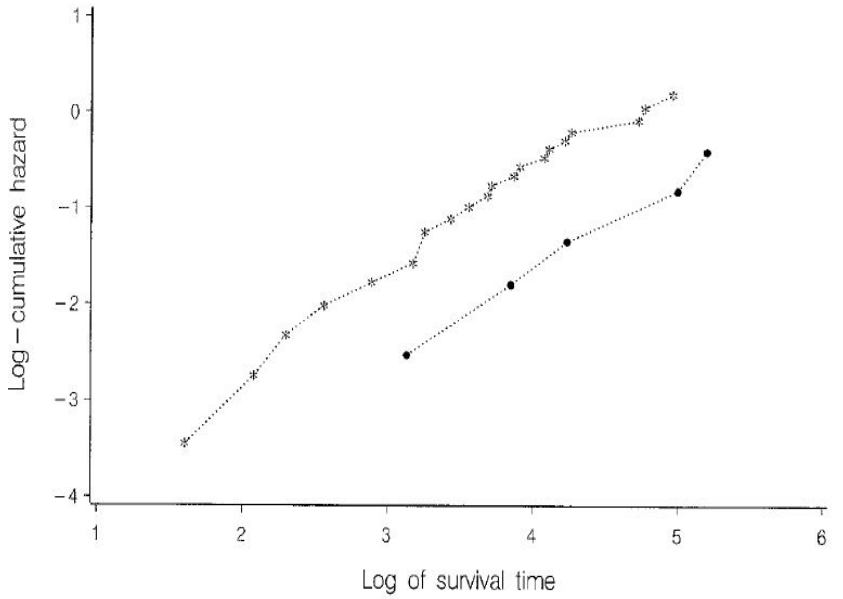
Z výsledného vztahu (2.30) vidíme, že rozdíly mezi logaritmy kumulovaných rizikových funkcí jednotlivých subjektů v čase t závisí pouze na hodnotách vysvětlujících proměnných a nikoli na čase. Za splnění předpokladu proporcionality a tedy platnosti vztahu (2.27), budou vykreslené křivky logaritmovaných kumulovaných funkcí pro jednotlivé subjekty souběžné (paralelní) v každém čase t. Zároveň se ale ukázalo, že při testování je lepší tyto křivky spíše vykreslovat oproti logaritmovanému času t než jen času t.[2]

Abychom však mohli tuto metodu použít, je třeba data seskupit dle hodnot kategoriální proměnné. Pokud máme k dispozici pouze spojité proměnné, je třeba vytvořit skupiny dle nějakých rozumných intervalů, zároveň je ale třeba, aby byl v každé skupině rozumný počet pozorování. V každé skupině poté provedeme odhad kumulativní rizikové funkce a pokud je předpoklad proporcionality splněn, pak budou vykreslené křivky logaritmované kumulativní rizikové funkce jednotlivých skupin navzájem paralelní. Příklad, jak může výsledný graf vypadat je na obrázku 2.2. Zde vidíme, že pozorování byla rozdělena to čtyř skupin, přičemž křivky odpovídající skupinám označeným čtverečky a kolečky jsou přibližně paralelní. Zbylé dvě skupiny už však v souladu nejsou. Je tedy namístě pochybovat o splnění předpokladu proporcionality.



Obrázek 2.2: Logaritmované kumulativní funkce - neproporcionální rizika [2]

Naopak křivky logaritmované kumulované funkce vykreslené na obrázku 2.3 jsou vzájemně paralelní a dle tohoto hlediska je proporcionalita rizik splněna.



Obrázek 2.3: Logaritmované kumulativní funkce - proporcionální rizika [2]

Další možností, jak lze ověřit (ne)proporcionalitu rizik je pomocí Schoenfeldových reziduí, a to jak graficky, tak i testem. [2], [12] Schoenfeldova rezidua se od klasických reziudí (používaných například v lineární regresi) liší zejména tím, že se nejedná o jednu hodnotu pro každé pozorování, ale o celý vektor hodnot. Složky tohoto vektoru přitom odpovídají jednotlivým proměnným v modelu. Schoenfeldova rezidua jsou pro j-tou proměnnou ($j = 1, 2, \ldots, k$) u i-tého pozorování ($i = 1, 2, \ldots, n$) definována jako:

$$r_{ij} = \delta_i \left[x_{ij} - \frac{\sum_{l \in R(t_{(i)})} x_{lj} exp\left\{ \mathbf{x}_l'(t)\hat{\boldsymbol{\beta}} \right\}}{\sum_{l \in R(t_{(i)})} exp\left\{ \mathbf{x}_l'(t)\hat{\boldsymbol{\beta}} \right\}} \right], \tag{2.31}$$

kde x_{ij} je hodnota j-té vysvětlující proměnné pro i-té pozorování, $R(t_{(i)})$ označuje skupinu pozorování, která jsou v čase $t_{(i)}$ v riziku, $\hat{\boldsymbol{\beta}}$ je vektor odhadnutých regresních koeficientů a δ_i je indikátorová funkce, která nabývá hodnoty jedna, je-li i-té pozorování necenzorované (v opačném případě nabývá hodnoty nula). Znamená to tedy, že Schoenfeldova rezidua počítáme pouze pro necenzorovaná pozorování.

Pro posouzení proporcionality rizik se doporučuje použití škálovaných (vážených) Schoenfeldových reziduí. Nechť $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{ik})'$ je vektor Schoenfeldových reziduí pro i-té pozorování. Odpovídající vektor škálovaných Schoenfeldových reziduí potom získáme jako:

$$\mathbf{r}_i^* = [\widehat{var}(\mathbf{r}_i)]^{-1} \mathbf{r}_i, \tag{2.32}$$

kde $\widehat{var}(\mathbf{r}_i)$ je odhad varianční matice vektoru Schoenfeldových reziduí i-tého pozorování. Pro zjednodušení výpočtů lze využít následující aproximace:

$$[\widehat{var}(\mathbf{r}_i)]^{-1} \approx s \left[\widehat{var}(\hat{\boldsymbol{\beta}}) \right], \tag{2.33}$$

kde s je počet dostupných necenzorovaných pozorování a $\widehat{var}(\hat{\boldsymbol{\beta}})$ je odhad varianční matice pro odhadované regresní koeficienty. Dohromady dostáváme škálovaná Schoenfeldova rezidua ve tvaru:

$$\mathbf{r}_i^* = s \left[\widehat{var}(\hat{\boldsymbol{\beta}}) \right] \mathbf{r}_i. \tag{2.34}$$

Grafické ověření předpokladu proporcionality potom spočívá ve vykreslení škálovaných reziduí pro jednotlivé proměnné do grafu oproti času. Je-li proporcionalita rizik splněna, pak jsou škálovaná Schoenfeldova rezidua náhodně rozmístěna kolem nuly, a to po celou

uvažovanou dobu. V případě, že je v grafech pozorovatelný nějaký trend, systematické chování či schodovité změny, pak je proporcionalita rizik s největší pravděpodobností porušena.

V souvislosti s grafickým ověřením proporcionality rizik pomocí Schoenfeldových reziduí si ještě uveďme, že byla zjištěna platnost přibližného vztahu [6]:

$$E(r_{ij}^*) \approx \beta_j(t_i) - \hat{\beta}_j, \tag{2.35}$$

kde $\hat{\beta}_j$ je odhad parametru v Coxově modelu proporcionálních rizik pro j-tou proměnnou a $\beta_j(t_i)$ je neznámý parametr pro j-tou proměnnou v čase t_i (tento neznámý parametr je brán jako závislý na čase). Vztah (2.35) nám tedy říká, že očekávaná hodnota škálovaných Schoenfeldových reziduí pro j-tou proměnnou v čase t_i je přibližně rovna rozdílu hodnoty neznámého (na čase závislého) parametru pro j-tou proměnnou a příslušného odhadnutého parametru, který na čase nezávisí. Pomocí metody Monte Carlo bylo dále zjištěno, že pokud proložíme body $r_{ij}^* + \hat{\beta}_j$ $(i = 1, \ldots, s,\ j = 1, \ldots, k)$ křivku, pak tato křivka bude odpovídat odhadu $\hat{\beta}_j(t)$. [6], [18] Předpoklad proporcionality přitom říká, že poměr rizik je v čase konstantní, tzn. konstantní jsou také parametry v modelu, které vyjadřují vliv jednotlivých proměnných. Z právě uvedeného vyplývá, že za předpokladu proporcionality bude takto vytvořená křivka v čase konstantní.

Pro jednotlivé proměnné v modelu lze testovat nulovou hypotézu, která tvrdí, že škálovaná Schoenfeldova uvažované proměnné jsou nezávislá na čase, resp. na jakékoli funkci času. [6] [7] Toto odpovídá grafickému testu, kdy chceme, aby byla odpovídající rezidua v čase náhodně rozmístěna kolem nuly a nevykazovala žádný trend. Testová statistika má pro j-tou proměnnou $(j = 1, 2, \ldots, k)$ asymptoticky chí kvadrát rozdělení o jednom stupni volnosti a je ve tvaru:

$$T_j = \frac{\left[\sum_{i=1}^{n} \left(\delta_i g(t_i) - \bar{g}(t)\right) r_{ij}^*\right]^2}{s\left[\widehat{var}(\hat{\beta}_j)\right] \sum_{i=1}^{n} \left(\delta_i g(t_i) - \bar{g}(t)\right)^2} \quad \dot{\sim} \quad \chi^2(1), \tag{2.36}$$

kde s je opět počet necenzorovaných pozorování, δ_i indikátorová funkce (nabývá hodnoty jedna, je-li i-té pozorování necenzorované), $g(t)$ je předem definovaná funkce času a $\bar{g}(t)$ je průměr hodnot $g(t_i)$, $i = 1, 2, \ldots, s$ (i odpovídá pouze necenzorovaným pozorováním). Vidíme, že stejně jako u grafického přístupu, i zde jsou pro test použita pouze necenzorovaná

pozorování. Nulovou hypotézu zamítáme pro hodnoty $T_j \geq \chi^2_{1-\alpha}(1)$. Jinými slovy, předpoklad proporcionality pro j-tou proměnnou zamítáme na zvolené hladině významnosti α, pokud bude hodnota vypočítané testové statistiky T_j větší než příslušný $1-\alpha$ kvantil chí kvadrát rozdělení o jednom stupni volnosti.

Testovat splnění předpokladu proporcionality lze ale také pro všechny uvažované proměnné dohromady. Testová statistika pro celkový test je ve tvaru:

$$T = \left[\sum_{i=1}^{n} \left(\delta_i g(t_i) - \bar{g}(t)\right) \mathbf{r}_i\right]' \left[\frac{s\left[\widehat{var}(\hat{\boldsymbol{\beta}})\right]}{\sum_{i=1}^{n} \left(\delta_i g(t_i) - \bar{g}(t)\right)^2}\right] \left[\sum_{i=1}^{n} \left(\delta_i g(t_i) - \bar{g}(t)\right) \mathbf{r}_i\right] \;\dot{\sim}\; \chi^2(k). \tag{2.37}$$

Na rozdíl od testů proporcionality pro jednotlivé proměnné zvlášť, u celkového testu používáme neškálovaná Schoenfeldova rezidua $\mathbf{r}_i$. Nulovou hypotézu pak zamítáme pro hodnoty $T \geq \chi^2_{1-\alpha}(k)$.

Kapitola 3

Modely neproporcionálních rizik

V předchozí kapitole jsme se naučili sestavit Coxův model proporcionálních rizik a odhadnout jeho parametry. Víme také, že výsledky tohoto modelu lze považovat za platné (správné) pouze při splnění určitých předpokladů. My zde budeme dále uvažovat, že kromě podmínky proporcionality rizik, jsou všechny tyto předpoklady splněny. Možným řešením, jak se vypořádat s neproporcionalitou rizik, je využití stratifikovaného Coxova modelu. Jak už název napovídá, jedná se o rozšíření původního Coxova modelu proporcionálních rizik. Znalosti získané v předchozí kapitole nyní využijeme pro stratifikované Coxovy modely. [2], [9], [12]

Závěrem této kapitoly si ještě krátce představíme modely konkurenčních rizik. O konkurenčním riziku mluvíme v případě, že může nastat jiná než sledovaná (konkurenční) událost, která způsobí, že námi sledovaná událost již nastat nemůže. Typickým příkladem konkurenční události je smrt z jiné než sledované příčiny. [9], [12]

3.1. Stratifikované modely proporcionálních rizik

V případě, kdy je porušen předpoklad proporcionality rizik, by použití Coxova modelu proporcionálních rizik vedlo k zavádějícím a nesprávným výsledkům. Často je proporcionalita rizik porušena vlivem jedné vysvětlující proměnné. Jistou možností by samozřejmě bylo tuto proměnnou vyloučit z analýzy, nabízí se však lepší varianta, a to stratifikace, resp. využití *stratifikovaného Coxova modelu*, který je rozšířením původního Coxova modelu proporcionálních rizik. [9], [12]

Stratifikace v podstatě znamená, že celý datový soubor rozdělíme do několika strat (skupin), a to právě podle hodnot (kategorií či vhodných intervalů) vysvětlující proměnné, která stojí za porušením proporcionality rizik. Podmínkou pro použití stratifikovaného Coxova modelu je proporcionalita rizik v rámci jednotlivých strat. Obecně tedy datový soubor rozdělíme do m skupin a v každé takto vytvořené skupině aplikujeme Coxův model proporcionálních rizik, tímto způsobem dostáváme celkem m různých rizikových funkcí. Riziková funkce je pro i-tý subjekt patřící do r-tého strata ve tvaru:

$$h_i^r(t|\mathbf{x}_i) = h_0^r(t) exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\}, \tag{3.1}$$

kde $r = 1, \ldots, m$ označuje stratum, $i = 1, \ldots, n_r$ označuje index pozorování v rámci r-tého strata, $\boldsymbol{\beta}$ je vektor regresních koeficientů, který je stejný pro všechna strata, $\mathbf{x}_i$ je vektor hodnot vysvětlujících proměnných pro i-tý subjekt v daném stratu a $h_0^r(t)$ je základní riziková funkce pro r-té stratum.

Jak už bylo řečeno, vektor regresních koeficientů je stejný pro všechna strata, znamená to tedy, že vliv vysvětlujících proměnných je stejný pro všechny skupiny. Znamená to ale také to, že odhad vektoru regresních koeficientů musí probíhat komplexně pro všechna strata najednou. Pro odhad regresních koeficicentů opět využíváme metodu maximální věrohodnosti, přičemž parciální funkce věrohodnosti je dána jako:

$$L(\boldsymbol{\beta}) = \prod_{r=1}^{m} L_r(\boldsymbol{\beta}), \tag{3.2}$$

kde $L_r(\boldsymbol{\beta})$ je věrohodnostní funkce r-tého strata. Funkce $L_r(\boldsymbol{\beta})$ $(r = 1, \ldots, m)$ jsou analogií k věrohodnostní funkci ve tvaru (2.12), jen s tím rozdílem, že pro výpočet $L_r(\boldsymbol{\beta})$ využíváme pouze pozorování, resp. hodnoty x_{ij}, příslušející k danému stratu.

Pro odhad vektoru regresních parametrů $\boldsymbol{\beta}$ je opět výhodnější využít logaritmickou parciální věrohodnostní funkci, která je ve tvaru:

$$\ln L(\boldsymbol{\beta}) = \sum_{r=1}^{m} \ln L_r(\boldsymbol{\beta}). \tag{3.3}$$

Výsledný odhad vektoru regresních parametrů $\boldsymbol{\beta}$ lze opět získat Newtonovou-Raphsonovou metodou. [2]

3.2. Modely konkurenčních rizik

Do této chvíle jsme se zabývali metodami analýzy přežití, které uvažují pouze jednu sledovanou událost. V praxi se ovšem stává, že během studie nastane jiná (než naše sledovaná) událost, která způsobí, že námi sledovaná událost již nastat nemůže. Typickým příkladem je smrt následkem jiné než námi sledované nemoci. Pokud budeme například sledovat pacienty s rakovinou, konkurečním rizikem pro nás může být smrt následkem dopravní nehody, smrt následkem selhání srdce apod.. Základní předpoklad pro tyto modely je tedy ten, že výskyt jednoho typu události u subjektu znamená vyloučení daného subjektu z rizikové skupiny pro všechny další typy událostí. Nyní se podívejme, jakým způsobem lze konkurenční rizika zahrnout do modelu. [9], [12]

Nechť T opět značí čas přežití a $\mathbf{x}$ je vektor hodnot vysvětlujících proměnných. V případě konkurenčních rizik přidáváme navíc ještě proměnnou R, která značí událost, která nastala (např. příčina smrti). Riziková funkce je poté definovaná jako:

$$ {}^{r}h_i(t;\mathbf{x}_i) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, R = r | T \geq t, \mathbf{x}_i)}{\Delta t}, \tag{3.4} $$

kde $r = 1, \ldots, m$. Uvažujeme tedy, že může nastat celkem m různých událostí. Vztah (3.4) nám vyjadřuje míru selhání důsledkem příčiny r (resp. intenzitu, s níž dochází k události r) v čase t, při daném $\mathbf{x}_i$ a to za přítomnosti zbylých $(m-1)$ konkurenčních událostí. Můžeme si povšimnout, že právě uvedený vztah se od vztahu (1.18) vyjadřujícího rizikovou funkci liší pouze tím, že uvažujeme i další (konkurenční) události. Celkovou rizikovou funkci, která bude vyjadřovat intenzitu, s níž dojde k jakékoli z uvažovaných událostí v čase t při daném $\mathbf{x}_i$ pak můžeme získat jako součet rizikových funkcí pro jednotlivé události, tedy:

$$ h_i(t;\mathbf{x}_i) = \sum_{r=1}^{m} {}^{r}h_i(t;\mathbf{x}_i). \tag{3.5} $$

Pro použití vztahu (3.5) však musí být splněn předpoklad, že se konkureční rizika vzájemně vylučují.

Za předpokladu proporcionálních rizik bychom následně mohli vztah (3.4) psát ve tvaru:

$$ {}^{r}h_i(t;\mathbf{x}_i) = {}^{r}h_0(t) exp\left\{\mathbf{x}_i'\boldsymbol{\beta}_j\right\}, \tag{3.6} $$

kde $r = 1, \ldots, m$ značí typ události a ${}^{r}h_0(t)$ je základní riziková funkce pro r-tou sledovanou událost v čase t.

Parciální funkce věrohodnosti je v případě konkurenčních rizik ve tvaru:

$$L(\boldsymbol{\beta}_r) = \prod_{r=1}^{m} \prod_{i=1}^{s_r} \frac{exp\left\{\mathbf{x}'_{(i)r}\boldsymbol{\beta}_r\right\}}{\sum_{l \in R(t_{(i)r})} exp\left\{\mathbf{x}'_l\boldsymbol{\beta}_r\right\}}, \tag{3.7}$$

kde r značí typ události, s_r je počet necenzorovaných událostí typu r, $\mathbf{x}_{(i)r}$ je vektor hodnot vysvětlujících proměnných pro pozorování, u něhož došlo k události typu r jako u i-tého v pořadí, $\boldsymbol{\beta}_r$ je vektor regresních parametrů vyjadřujících vliv proměnných na výskyt události typu r a $R(t_{(i)r})$ je riziková skupina v čase $t_{(i)r}$ (i-té nastání události typu r v pořadí). Výslední odhady parametrů pak opět získáme například využitím Newtonovy-Raphsonovy metody. [2], [12]

V případě porušení předpokladu proporcionality lze využít stratifikaci. Jednotlivé typy událostí je pak nutné modelovat zvlášť v rámci každého strata. Namísto vztahu (3.6), který vyjadřuje rizikovou funkci pro r-tý typ události pro i-tý subjekt, dostaneme vztah vyjadřující rizikovou funkci pro r-tý typ události pro i-tý subjekt v q-tém stratu, tj.:

$${}^{r}h_i^q(t; \mathbf{x}_i) = {}^{r}h_0^q(t) exp\left\{\mathbf{x}'_i\boldsymbol{\beta}_j\right\}, \tag{3.8}$$

kde $r = 1, \ldots, m$ označuje typ události a $q = 1, \ldots, Q$ značí stratum. Využitím analogie ke vztahu (3.5) lze získat celkové rizikové funkce v jednotlivých stratech. Tyto funkce budou vyjadřovat intenzitu, s níž dojde v daném stratu u i-tého subjektu k jakékoli z m uvažovaných událostí a lze je získat následovně:

$$h_i^q(t; \mathbf{x}_i) = \sum_{r=1}^{m} {}^{r}h_i^q(t; \mathbf{x}_i). \tag{3.9}$$

Odhad parametrů bychom opět provedli metodou maximální věrohodnosti.

Kapitola 4

Výběr a hodnocení modelu

V této kapitole si ukážeme metody, které nám pomohou s výběrem modelu, resp. s výběrem proměnných, které do modelu zahrneme. Nejdříve se budeme věnovat testování hypotéz, a to jak o významnosti jednotlivých proměnných, tak i o podmodelech. Dále se seznámíme s metodami pro porovnání modelů mezi sebou. Nakonec probereme možnosti, jakými lze ohodnotit kvalitu takto nalezených modelů.

4.1. Testování hypotéz

V předchozí kapitole jsme se naučili, jak sestavit model a odhadnout jeho parametry. Nyní je třeba rozhodnout, které proměnné do modelu zahrnout, případně, zda má smysl do modelu zahrnout také jejich interakce. S tímto úkolem nám pomůžou testy hypotéz o významnosti parametrů, resp. testy o podmodelech. My si zde uvedeme tři nejpoužívanější, a to Waldův test, test poměrem věrohodností a skórový test. Nakonec si zde představíme ještě log-rank test, který testuje hypotézu o shodě odhadovaných pravděpodobností přežití pro různé skupiny subjektů.

4.1.1. Waldův test

Základní statistikou využívanou při výběru proměnných do modelu je Waldova statistika [13], resp. Waldův test, který testuje následující hypotézu:

$$\mathrm{H}_0 : \beta_j = 0 \;\text{ vs. }\; \mathrm{H}_1 : \beta_j \neq 0.$$

Nulová hypotéza říká, že j-tá proměnná není statisticky významná. Pokud tuto hypotézu nezamítneme, měli bychom zvážit vyjmutí dané proměnné z modelu. Za platnosti nulové hypotézy má Waldova statistika asymptoticky normované normální rozdělení a je ve tvaru:

$$\mathrm{W}_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{var}(\hat{\beta}_j)}} \quad \dot{\sim} \quad N(0,1), \tag{4.1}$$

kde $\sqrt{\widehat{var}(\hat{\beta}_j)}$ je odhad směrodatné chyby pro příslušný odhad j-tého regresního parametru, který získáme ze vztahu (2.17) jako odmocninu j-tého diagonálního prvku. Nulovou hypotézu zamítáme pro hodnoty: $|\mathrm{W}_j| > u_{1-\frac{\alpha}{2}}$, kde $u_{1-\frac{\alpha}{2}}$ je $1-\frac{\alpha}{2}$ kvantil normovaného normálního rozdělení. V literatuře je Waldova statistika často uváděna v transformaci $\mathrm{W}_j^2 = \frac{\hat{\beta}_j^2}{\widehat{var}(\hat{\beta}_j)}$, která má za platnosti nulové hypotézy asymptoticky chí kvadrát rozdělení o jednom stupni volnosti.

Waldův test lze využít také pro ověření významnosti více proměnných zároveň. Nejčastěji takto testujeme významnost všech proměnných v modelu. Označíme-li si $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, pak lze testovat hypotézu ve tvaru:

$$\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0} \quad \text{vs.} \quad \mathrm{H}_1 : \exists\, j, \;\; \beta_j \neq 0.$$

Nezamítneme-li nulovou hypotézu, znamená to, že náš model je srovnatelný s nulovým modelem (modelem uvažujícím pouze základní rizikovou funkci). Testová statistika je ve tvaru:

$$\mathrm{W} = \hat{\boldsymbol{\beta}}' \left[\widehat{var}(\hat{\boldsymbol{\beta}})\right]^{-1} \hat{\boldsymbol{\beta}} \quad \dot{\sim} \quad \chi^2(k), \tag{4.2}$$

kde $\hat{\boldsymbol{\beta}}$ je vektor odhadnutých regresních koeficientů, $\widehat{var}(\hat{\boldsymbol{\beta}})$ je odhadnutá varianční matice pro regresní parametry (získaná ze vztahu (2.17)) a k je počet parametrů (proměnných), jejichž významnost testujeme.

4.1.2. Test poměrem věrohodností

Test poměrem věrohodností lze využít (stejně jako Waldův test) pro testování hypotézy o nulovosti j-tého regresního koeficientu, testování hypotézy o nulovosti více parametrů zároveň, včetně testování hypotézy o nulovosti všech parametrů dohromady a také

pro testování hypotéz o podmodelech. [12], [13] Uvažujme obecně dva modely: model M_1 s počtem parametrů p_1 a model M_2 s p_2 parametry, přičemž $p_1 > p_2$ (model M_2 uvažuje méně parametrů než model M_1). Naším cílem je mít co nejlepší model s co nejmenším počtem parametrů, a proto budeme testovat hypotézu ve tvaru:

$$H_0 : \text{Model } M_2 \text{ je podmodelem modelu } M_1.$$

V případě, že tuto hypotézu zamítneme, znamená to, že nelze přejít od složitějšího modelu k modelu jednoduššímu, neboť bychom ztratili významnou část informace. Naopak, pokud danou hypotézu nezamítneme, můžeme uvažovat jednodušší model s menším počtem parametrů. Zároveň tím říkáme, že parametry, které jsme uvažovali navíc oproti jednoduššímu modelu, jsou statisticky nevýznamné.

Testová statistika je, jak už název napovídá, založena na věrohodnostních funkcích, resp. na logaritmech věrohodnostních funkcí testovaných modelů. Označíme-li si postupně hodnoty logaritmů věrohodnostních fukncí pro modely M_1 a M_2 jako l_{M_1} a l_{M_2}, pak můžeme testovou statistiku zapsat ve tvaru:

$$\mathrm{LRT}(M_1, M_2) = 2(l_{M_1} - l_{M_2}) \quad \dot{\sim} \quad \chi^2(p_1 - p_2). \tag{4.3}$$

Nulovou hypotézu zamítáme pro hodnoty $\mathrm{LRT}(M_1, M_2) \geq \chi^2_{1-\alpha}(p_1 - p_2)$. Rozdíl $p_1 - p_2$ je přitom roven počtu parametrů, jejichž významnost tímto způsobem testujeme. Poznamenejme ještě, že tento test lze použít pouze pro vnořené modely, tj. množina proměnných, které uvažujeme v modelu M_2, musí být podmnožinou množiny proměnných uvažovaných v modelu M_1. Jestliže chceme testovat významnost všech parametrů zároveň, stačí nám tímto způsobem srovnat model uvažující pouze základní rizikovou funkci a model se všemi parametry. Zamítnutí nulové hypotézy by v takovém případě znamenalo, že některé parametry statisticky významné jsou.

4.1.3. Skórový test

Skórový test opět slouží k ověření významnosti proměnných v modelu. Testovat lze jak nulovost jednotlivých parametrů zvlášť, tak i nulovost celého vektoru parametrů. [2], [12], [13]

Než si představíme testovou statistiku, označme si parciální derivaci logaritmické parciální věrohodnostní funkce $l(\boldsymbol{\beta})$ podle proměnné β_j jako $u_j(\beta)$, tj:

$$u_j(\beta) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} \tag{4.4}$$

a záporně vzatou druhou parciální derivaci téže funkce podle téže proměnné β_j jako $i_j(\beta)$, tj.:

$$i_j(\beta) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial^2 \beta_j}. \tag{4.5}$$

S právě zavedeným značením lze testovou statistiku pro ověření nulové hypotézy o významnosti j-tého parametru ($\mathrm{H}_0 : \beta_j = 0$) psát ve tvaru:

$$Z_j = \frac{u_j(0)}{\sqrt{i_j(0)}} \quad \dot{\sim} \quad N(0,1), \tag{4.6}$$

kde $u_j(0)$, resp. $i_j(0)$, je hodnota $u_j(\beta)$, resp. $i_j(\beta)$ pokud za parametr β_j dosadíme nulu. Jinými slovy, díváme se na hodnoty první a druhé parciální derivace logaritmické parciální funkce věrohodnosti podle proměnné β_j v bodě $\beta_j = 0$. Vztahy pro výpočet $u_j(\beta)$ a $i_j(\beta)$ jsme si uvedli již dříve (vztahy (2.16) a (2.18)). Nulovou hypotézu zamítáme pro hodnoty $|Z_j| > u_{1-\frac{\alpha}{2}}$, kde $u_{1-\frac{\alpha}{2}}$ je $1 - \frac{\alpha}{2}$ kvantil normovaného normálního rozdělení. Stejně jako Waldův test, i skórový test lze tranformovat umocněním na druhou, přičemž výsledná testová statistika pak má asymptoticky chí kvadrát rozdělení o jednom stupni volnosti:

$$Z_j^2 = \frac{[u_j(0)]^2}{i_j(0)} \quad \dot{\sim} \quad \chi^2(1). \tag{4.7}$$

V případě, kdy chceme chceme testovat nulovost všech parametrů v modelu ($\mathrm{H}_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' = \mathbf{0}$), je testová statistika ve tvaru:

$$Z = \mathbf{u}'(\mathbf{0})\mathbf{I}^{-1}(\mathbf{0})\mathbf{u}(\mathbf{0}) \quad \dot{\sim} \quad \chi^2(k), \tag{4.8}$$

kde $\mathbf{u}(\boldsymbol{\beta}) = (u_1(\beta), u_2(\beta), \ldots, u_k(\beta))'$ a $\mathbf{I}^{-1}(\boldsymbol{\beta}) \approx var(\boldsymbol{\beta})$. Vektor hodnot $\mathbf{u}(\mathbf{0})$ a matici $\mathbf{I}^{-1}(\mathbf{0})$ pak získáme dosazením nuly za příslušné parametry β_j, $j = 1, 2, \ldots, k$. Nulovou hypotézu zamítáme pro hodnoty $Z \geq \chi^2_{1-\alpha}(k)$.

4.1.4. Log-rank test

Log-rank test je jednou z možností, jak otestovat hypotézu o shodě odhadovaných pravděpodobností přežití pro různé skupiny. [2], [12], [13] V praxi takto můžeme například otestovat, zda má podávání léku A namísto léku B signifikantní vliv na pravděpodobnost přežití nebo zda se významně liší pravděpodobnost přežití pro ženy a pro muže. V případě, že zjistíme významné rozdíly mezi skupinami, je na místě zařadit danou proměnnou do modelu.

Uvažujme nejprve dvě skupiny. Formálně můžeme testovanou hypotézu zapsat následovně:

$$\mathrm{H}_0 : S_1(t) = S_2(t) \quad \text{vs.} \quad \mathrm{H}_1 : S_1(t) \neq S_2(t).$$

Sledujeme-li například skupinu pacientů léčenou lékem A a skupinu pacientů léčenou lékem B, pak výše uvedená nulová hypotéza říká, že mezi léky A a B není statisticky významný rozdíl co se týká přežití pacientů. Alternativou je pak tvrzení, že mezi danými léky je signifikantní rozdíl. V případě, kdy očekáváme, že lék B, resp. lék A, bude v léčbě efektivnější, pak je možné testovat výše uvedenou nulovou hypotézu proti jednostranné alternativě $S_1(t) < S_2(t)$, resp. $S_1(t) > S_2(t)$.

Nechť O_1 a O_2 jsou počty pozorovaných a E_1 a E_2 počty očekávaných událostí v první a druhé skupině. Počet očekávaných událostí E_k (obecně pro k-tou skupinu, kde $k = 1, 2$) vypočítáme jako:

$$E_k = \sum_i \frac{n_{ki}}{n_i} d_i, \tag{4.9}$$

kde n_{ki} je počet pozorování v riziku ve skupině k v čase t_i, n_i je počet pozorování v riziku v čase t_i v obou skupinách dohromady a d_i je počet událostí, jež nastaly v čase t_i. Testová statistika pro log-rank test je pak ve tvaru [2], [13]:

$$X^2_{LR} = \frac{(O_1 - E_1)^2}{var(O_1 - E_1)} \quad \dot{\sim} \quad \chi^2(1), \tag{4.10}$$

kde $var(O_1 - E_1)$ vypočítáme jako:

$$var(O_1 - E_1) = \sum_i \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}.$$

Jedná o součet rozptylů rozdílů pozorovaných a očekávaných počtů událostí v první skupině přes všechny uvažované časy událostí t_i. Nulovou hypotézu zamítáme pro hodnoty $X^2_{LR} \geq \chi^2_{1-\alpha}(1)$.

Poznamenejme, že kromě log-rank testu ve tvaru (4.10) se pro testování téže nulové hypotézy používá také chí kvadrát test. Odpovídající testová statistika je přitom rovněž založena na počtech pozorovaných a očekávaných událostí [12]:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad \dot{\sim} \quad \chi^2(1). \tag{4.11}$$

Nulovou hypotézu zamítáme pro hodnoty $X^2 \geq \chi^2_{1-\alpha}(1)$.

Ilustrujme si nyní log-rank test na příkladu fiktivních dat z tabulky 1.1. Předpokládejme, že prvních pět pozorování patří do skupiny A a zbylých pět pozorování do skupiny B. V tabulce 4.1 jsou uvedeny údaje potřebné pro výpočet testové statistiky log-rank testu. V prvním sloupci jsou časové okamžiky t_i, ve kterých došlo k události. Ve druhém sloupci jsou počty pozorování n_i, která jsou v čase t_i v riziku. Následují sloupce n_{1i} a d_{1i}, resp. n_{2i} a d_{2i} s počty pozorování v riziku a s počty událostí pro první a druhou skupinu v čase t_i. Další dva sloupce pak obsahují očekávané počty událostí v čase t_i v první (e_{1i}) a v druhé (e_{2i}) skupině. Poslední sloupec pak odpovídá rozptylu rozdílu pozorovaného a očekávaného počtu událostí v čase t_i v první skupině (v_{1i}). V posledním řádku tabulky jsou provedeny součty, které budeme dále potřebovat k výpočtu testové statistiky, tedy $O_1 = \sum_i d_{1i}$, $O_2 = \sum_i d_{2i}$, $E_1 = \sum_i e_{1i}$, $E_2 = \sum_i e_{2i}$ a $var(O_1 - E_1) = \sum_i v_{1i}$.

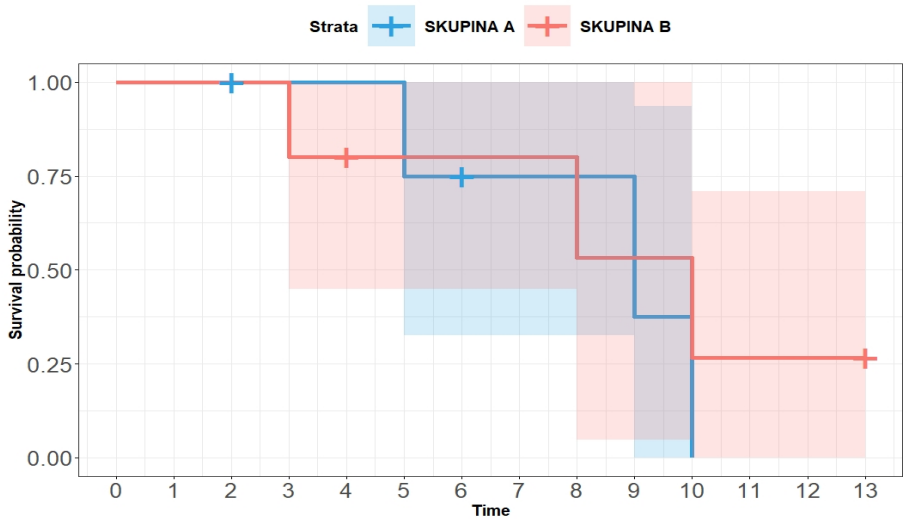
t_i	n_i	n_{1i}	d_{1i}	n_{2i}	d_{2i}	e_{1i}	e_{2i}	v_{1i}
3	9	4	0	5	1	0.444	0.556	0.247
5	7	4	1	3	0	0.571	0.429	0.245
8	5	2	0	3	1	0.400	0.600	0.240
9	4	2	1	2	0	0.500	0.500	0.250
10	3	1	1	2	1	0.667	1.333	0.222
			$O_1 = 3$		$O_2 = 3$	$E_1 = 2.583$	$E_2 = 3.417$	$var(O_1 - E_1) = 1.204$

Tabulka 4.1: Hodnoty pro výpočet log-rank testu pro ilustativní data

Dle vztahu (4.10) dostáváme:

$$X^2_{LR} = \frac{(3 - 2.583)^2}{1.204} = 0.1444 < 3.841 = \chi^2_{0.95}(1).$$

Z výsledků vyplývá, že na hladině významnosti 0.05 nelze nulovou hypotézu zamítnout. A dále tedy můžeme předpokládat, že se pravděpodobnosti přežití pro dané skupiny významně neliší. Toto lze ostatně zhodnotit i graficky pomocí křivek přežití na obrázku 4.1. Vidíme, že křivky se po celou uvažovanou dobu vzájemně překřižují a odpovídající 95% intervaly spolehlivosti se překrývají.



Obrázek 4.1: Ilustrativní Kaplanovy-Meierovy křivky přežití pro skupiny A a B

Pokud bychom pro ověření uvedené nulové hypotézy využili chí kvadrát test, kterému odpovídá vztah (4.11), došli bychom ke stejnému závěru (křivky přežití se pro uvažované skupiny významně neliší).

$$X^2 = \frac{(3-2.583)^2}{2.583} + \frac{(3-3.417)^2}{3.417} = 0.118 < 3.841 = \chi^2_{0.95}(1).$$

Nyní předpokládejme, že chceme porovnat celkem K ($K > 2$) skupin. Nulovou hypotézu v takovém případě můžeme zapsat jako:

$$\mathrm{H}_0 : S_1(t) = S_2(t) = \ldots = S_K(t) \quad \text{vs.} \quad \mathrm{H}_1 : \exists k \neq l, \;\; S_k(t) \neq S_l(t).$$

Pro výpočet log-rank testu využijeme modifikovanou verzi vztahu (4.10). [2] Podobně jako v případě dvou skupin, kdy jsme pro výpočet využívali pouze hodnoty odpovídající první skupině (O_1, E_1 a $var(O_1 - E_1)$), budeme nyní využívat hodnoty pouze pro $K-1$ skupin.

Nechť O_k jsou pozorované a E_k očekávané počty událostí v k-té skupině ($k = 1, \ldots, K$). Očekávané počty událostí E_k lze vypočítat dle vztahu (4.9). Namísto rozptylu $var(O_1 - E_1)$ budeme v tomto případě ale potřebovat varianční matici $\mathbf{V}$ o rozměrech $(K-1) \times (K-1)$. Diagonála matice $\mathbf{V}$ bude tvořena prvky $v_{kk} = var(O_k - E_k)$:

$$v_{kk} = var(O_k - E_k) = \sum_i \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)}.$$

Mimodiagonální prvky (v_{kl}) pak budou představovat odpovídající kovariance mezi k-tou a l-tou skupinou.

$$v_{kl} = \sum_i \frac{n_{ki}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)}.$$

Výsledná matice $\mathbf{V}$ je ve tvaru:

$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1(K-1)} \\ v_{21} & v_{22} & \cdots & v_{2(K-1)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{(K-1)1} & \cdots & \cdots & v_{(K-1)(K-1)} \end{pmatrix}.$$

Označíme-li $\mathbf{O} = \left(O_1, O_2, \ldots, O_{(K-1)}\right)'$ a $\mathbf{E} = \left(E_1, E_2, \ldots, E_{(K-1)}\right)'$, pak je testová statistika pro log-rank test ve tvaru [2]:

$$X^2 = (\mathbf{O} - \mathbf{E})'\mathbf{V}^{-1}(\mathbf{O} - \mathbf{E}) \quad \dot{\sim} \quad \chi^2(K-1). \tag{4.12}$$

Nulovou hypotézu budeme zamítat na hladině α pro hodnoty $X^2 \geq \chi^2_{1-\alpha}(K-1)$

4.2. Výběr modelu

Již tedy víme, jak vytvořit model, odhadnout parametry a otestovat jejich významnost. Nyní si ukážeme metody, jakými lze takto vytvořené modely vzájemně porovat. Cílem těchto metod je vybrat takový model, který bude co nejlépe vysvětlovat porozovaná data, a který bude zároveň obsahovat co nejmenší počet proměnných.

4.2.1. Informační kritéria

Informační kritéria nabízí možnost, jak porovnat vytvořené modely, a to s ohledem na počet uvažovaných proměnných. [12], [13] Hlavní myšlenkou informačních kritérií je skutečnost, že při tvorbě každého modelu dochází ke ztrátě informace obsažené v původních datech a my chceme tuto ztrátu minimalizovat. Pro každý vytvořený model vypočítáme hodnotu informačního kritéria a následně vybíráme hodnotu nejnižší. Mezi nejpoužívanější informační kritéria patří Akaikeho informační kritérium a Bayesovo informační kritérium.

Akaikeho informační kritérium je ve tvaru:

$$AIC = -2l_M + 2p, \tag{4.13}$$

kde l_M je hodnota logaritmu věrohodnostní funkce uvažovaného modelu a p je odpovídající počet parametrů. Protože se ztráta informace snižuje s rostoucím počtem parametrů, je ve výpočtu zahrnut člen $+2p$, který má za úkol počet parametrů penalizovat.

Bayesovo informační kritérium je pak ve tvaru:

$$BIC = -2l_M + p\ln(n), \tag{4.14}$$

kde l_M je opět hodnota logaritmu věrohodnostní funkce uvažovaného modelu, p počet parametrů a n počet pozorování. Vidíme, že penalizační člen má v tomto případě podobu $+p\ln(n)$, což způsobuje mnohem větší znevýhodnění (zvláště pro modely s vyšším počtem parametrů) než tomu je u Akaikeho informačního kritéria.

4.2.2. Koeficienty determinace

Při výběru modelu se můžeme podívat také na hodnoty (upravených) koeficientů determinace. [1], [15] V lineární regresi nám hodnota koeficientu determinace udává, jakou část celkové variability se nám podařilo vysvětlit pomocí uvažovaného modelu. Ideálně bychom tedy chtěli hodnotu koeficientu determinace rovnu jedné. Pro potřeby analýzy přežití bylo vytvořeno několik zobecnění (modifikací) koeficientu determinace. Tyto statistiky by přitom měly být pokud možno nezávislé na počtu cenzorovaných pozorování

v datovém souboru. Dále by mělo platit, že pokud uvažujeme model a jeho podmodel, bude hodnota koeficientu determinace vždy nižší pro podmodel.

Existuje mnoho různých modifikací koeficientů determinace, nicméně my se zde omezíme pouze na tři z nich a začneme od nejstaršího. [1], [15] V roce 1991 byl definován Nagelkerkův koeficient determinace, a to následovně :

$$R_N^2 = 1 - \exp\left\{-\frac{2}{n}\left(l_{\hat{\boldsymbol{\beta}}} - l_0\right)\right\}, \tag{4.15}$$

kde $l_{\hat{\boldsymbol{\beta}}}$ je hodnota logaritmu funkce věrohodnosti uvažovaného modelu, l_0 hodnota logaritmu funkce věrohodnosti pro nulový model (model uvažující pouze základní rizikovou funkci) a n je počet pozorování. Hodnota tohoto indexu determinace byla dříve součástí výstupu Coxova modelu proporcionálních rizik v softwaru R, nicméně následně byla odstraněna, a to z důvodu špatných vlastností. Bylo zjištěno, že takto definovaný koeficient determinace je záporně korelován s podílem cenzorovaných pozorování v datovém souboru. V případě, že se podíl cenzorovaných pozorování v datovém souboru blíží k jedné, pak se hodnota Nagelkerkova koeficientu determinace blíží k nule. [15]

Následně byl roku 1999 definován nový koeficient determinace, jehož autory jsou R. Xu, a J. O'Quigley. Použití tohoto koeficientu determinace je striktně určeno pro modely s proporcionálním rizikem. Výpočet vychází ze vzorce pro upravený index determinace v lineární regresi, přičemž zde pracujeme s (váženými) čtverci Schoenfeldových reziduí, které si označíme jako $J(\beta)$. [1] Výhodou této statistiky je skutečnost, že nevykazuje korelovanost s podílem cenzorovaných pozorování v datech. Koeficient determinace získáme jako:

$$R_{XO}^2 = 1 - \frac{J(\hat{\boldsymbol{\beta}})}{J(0)} = 1 - \frac{\sum_{\delta_{ij}=1}\left(r_{ij}(\hat{\boldsymbol{\beta}})\right)^2}{\sum_{\delta_{ij}=1}\left(r_{ij}(0)\right)^2}, \tag{4.16}$$

kde δ_{ij} značí indikátorovou funkci pro cenzorování (jeli i-té pozorování necenzorované, pak je funkce rovna jedné), r_{ij} pak značí Schoenfeldova rezidua, přičemž $r_{ij}(\hat{\boldsymbol{\beta}})$ náleží k modelu s odhadnutými parametry $\hat{\boldsymbol{\beta}}$ a $r_{ij}(0)$ k modelu nulovému, tj. k modelu uvažujícímu pouze základní rizikovou funkci. Schoenfeldova rezidua počítáme podle vztahu (2.31).

Poslední koeficient determinace, o kterém se tu zmíníme, byl představen v roce 2005. Autory tohoto koeficientu determinace jsou J. O'Quigley, R. Xu a J. Stare. Jedná se o modifikaci původního Nagelkerkova koeficientu, přičemž počet pozorování n je ve výpočetním vzorci nahrazen počtem událostí (budeme značit e). Dostáváme koeficient determinace ve tvaru:

$$R^2_{OXS} = 1 - \exp\left\{-\frac{2}{e}\left(l_{\hat{\boldsymbol{\beta}}} - l_0\right)\right\}. \tag{4.17}$$

Takto definovaný koeficient determinace již není negativně korelován s podílem cenzorovaných dat v souboru, nicméně byla naopak pozorována korelace pozitivní.

4.3. Hodnocení modelu

V předchozí podkapitole jsme se zabývali metodami pro výběr modelu, které spočívaly v porovnávání modelů mezi sebou. Nyní nás bude zajímat, jak dobrý takto nalezený model je, a to bez ohledu na modely ostatní. Snadno by se nám totiž mohlo stát, že vybereme model, který bude v porovnání s ostatními modely velmi dobrý, nicméně jeho prediktivní vlastnosti budou naopak velmi špatné.

4.3.1. Konkordance

Konkordance se řadí mezi testy dobré shody a je také jednou z nejpoužívanějších metod pro hodnocení kvality modelu přežití. [19], [20], [21] Používá se také v lineární či logistické regresi. Základem je porovnávání párů pozorovaných a očekávaných hodnot. Nechť y_i a y_j jsou pozorované a x_i a x_j očekávané hodnoty pro i-té a j-té pozorování. Konkordance, kterou budeme dále značit jako C, je pak definována jako:

$$C = P\left(x_i > x_j | y_i > y_j\right). \tag{4.18}$$

Jedná se tedy o podmíněnou pravděpodobnost toho, že modelem odhadovaná (vyrovnaná) hodnota bude pro i-tý subjekt vyšší než pro j-tý subjekt, a to za podmínky, že je pozorovaná hodnota pro i-tý subjekt vyšší než pro subjekt j-tý.

Než se seznámíme se vzorcem pro praktický výpočet konkordance, je třeba zavést několik pojmů. Jak už bylo řečeno, konkordance je založena na porovnávání párů pozorování. Pár

pro který platí: $(y_i > y_j, x_i > x_j)$ nebo $(y_i < y_j, x_i < x_j)$ nazveme konkordantním. V případě, že toto není splněno, nazveme takový pár diskordantním. Samozřejmě se může stát, že pozorované hodnoty obou subjektů budou shodné $(y_i = y_j)$ nebo budou shodné hodnoty vyrovnané $(x_i = x_j)$ či budou dokonce shodné jak hodnoty pozorované tak i hodnoty očekávané. V takových případech mluvíme o vázaných (vyrovnaných) párech.

V analýze přežití rozumíme pozorovanými hodnotami délky dob přežití, resp. délky doby subjektů ve studii (pokud je pozorování cenzorované). Pod čekávanými hodnotami pak kromě modelem predikované délky doby přežití rozumíme i tzv. predikované rizikové skóry. Rizikovým skórem pro i-té pozorování budeme dále rozumět hodnotu $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$. U konkordantního páru by tedy mělo platit, že subjektu s delším pozorovaným časem odpovídá menší rizikový skór. Z výše uvedeného vyplývá, že v analýze přežití bychom měli prohodit definice pro konkordantní a diskordantní pár. Nechť y_i a y_j představují pozorované časy přežití pro i-tý a j-tý subjekt a $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ a $\mathbf{x}_j'\hat{\boldsymbol{\beta}}$ odpovídající modelem odhadované rizikové skóry. Konkordantním párem v analýze přežití nazveme pár splňující: $\left(y_i < y_j, \mathbf{x}_i'\hat{\boldsymbol{\beta}} > \mathbf{x}_j'\hat{\boldsymbol{\beta}}\right)$ nebo $\left(y_i > y_j, \mathbf{x}_i'\hat{\boldsymbol{\beta}} < \mathbf{x}_j'\hat{\boldsymbol{\beta}}\right)$. Diskordantní pár je potom pár splňující: $\left(y_i > y_j, \mathbf{x}_i'\hat{\boldsymbol{\beta}} > \mathbf{x}_j'\hat{\boldsymbol{\beta}}\right)$ nebo $\left(y_i < y_j, \mathbf{x}_i'\hat{\boldsymbol{\beta}} < \mathbf{x}_j'\hat{\boldsymbol{\beta}}\right)$

V analýze přežití musíme navíc řešit i problém cenzorovaných pozorování. V případě, že je i-té pozorování cenzorováno v čase t_i a j-té pozorování je cenzorováno (nebo u něj došlo k události) v čase t_j, kde $t_i < t_j$, potom takový pár též označujeme jako vázaný. Nelze totiž rozhodnout, zda u i-tého pozorování došlo ke sledované události dříve než v čase t_j nebo až později. Výjimkou je případ, kdy u i-tého subjektu dojde k cenzorování v čase t_i a v témže času dojde k události u subjektu j, potom platí: $y_i > y_j$, neboť v tomto případě s jistotou víme, že ke sledované události u i-tého subjektu dojde (došlo by) až později.

Označme si K počet konkordantních párů v souboru, D počet diskordantních párů, T_x počet vázaných párů z důvodu rovnosti predikovaných rizikových skórů, T_y počet vázaných párů z důvodu rovnosti pozorovaných časů přežití a T_{xy} počet vázaných párů z důvodu rovnosti jak predikovaných rizikových skórů, tak i pozorovaných časů přežití. Celkový počet párů, které takto zkoumáme je při rozsahu datového souboru n roven $\frac{n(n-1)}{2}$. K výpočtu

konkordance využijeme Somersovo d, které získáme jako:

$$d = \frac{K - D}{K + D + T_x}. \tag{4.19}$$

Platí, že $d \in \langle -1; 1 \rangle$. Pokud by byly všechny páry v datovém souboru konkordantní, pak $D = T_x = 0$ a tím pádem $d = 1$. Naopak, pokud by byly všechny páry v souboru diskordantní, pak $K = T_x = 0$ a dostaneme hodnotu Somerdova d rovnu mínus jedné. Konkordanci získáme upravou vzorečku Somersova d:

$$C = \frac{d + 1}{2} = \frac{K + \frac{T_x}{2}}{K + D + T_x}. \tag{4.20}$$

Konkordance může na rozdíl od Somersova d nabývat hodnot z intervalu od nuly do jedné, nicméně obvykle se pohybuje v hodnotách od jedné poloviny do jedné. Hodnota $C = \frac{1}{2}$ přitom představuje situaci, kdy predikce pro každý subjekt tvoříme zcela náhodně (bez systematického pravidla). Naproti tomu hodnota $C = 1$ pro nás představuje ideál, znamená to totiž, že jsou všechna pozorování v souladu s očekáváními. V analýze přežití se však nejčastěji setkáme s hodnotami $C = 0.6 - 0.7$.

V případě stratifikovaných modelů (uvažujeme strata $s = 1, 2, \ldots, S$) se výpočet korkondance provede následujícím způsobem:

$$C = \frac{\sum_{s=1}^{S} \left(K_s + \frac{T_{xs}}{2} \right)}{\sum_{s=1}^{S} \left(K_s + D_s + T_{xs} \right)}, \tag{4.21}$$

kde K_s, D_s a T_{xs} jsou postupně počty konkordantních párů, diskordantních párů a vázaných párů (v důsledku shody predikovaného skóre) v jednotlivých stratech.

Konkordance vypočítaná podle výše uvedených vzorců (4.20) a (4.21) se označuje jako Harrellova konkordance (Harrellovo C) a je součástí základního výstupu Coxova modelu proporcionálních rizik v softwaru R. Nevýhodou Harrellova C je fakt, že v případě přítomnosti cenzorovaných pozorování je tento odhad vychýlený. To je způsobeno tím, že je-li kratší pozorovaný čas přežití ve dvojici cenzorovaný, pak je tato dvojice z výpočtu vyloučena.

Stejně jako je tomu u koeficientů determinace, tak i pro konkordanci existuje několik různých modifikací. Jednotlivé vzorce pro výpočet konkordance se přitom liší zejména tím,

jak a jestli uvažují vázané páry a jak pracují s cenzorovanými pozorováními. Možnou alternativou k Harrellovu C je pak konkordance, která byla představena autory Gönenem a Hellerem. [19] Tato statistika je v přítomnosti cenzorovaných pozorování asymptoticky nevychýlená, neboť není přímo závislá na pozorovaných časech. Gönenova a Hellerova konkordance je funkcí regresních parametrů a je ve tvaru:

$$C_{GH} = C_{GH}\left(\hat{\boldsymbol{\beta}}\right) = \frac{2}{n(n-1)} \sum_{i<j} \sum \left\{ \frac{I(\Delta \mathbf{x}_{ji}\hat{\boldsymbol{\beta}} \leq 0)}{1+\exp\left(\Delta \mathbf{x}_{ji}\hat{\boldsymbol{\beta}}\right)} + \frac{I(\Delta \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} < 0)}{1+\exp\left(\Delta \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}\right)} \right\}, \tag{4.22}$$

kde $I(\cdot)$ je indikátorová funkce (nabývá hodnoty jedna, je-li výraz v závorce pravdivý, v opačném případě je rovna nule), $\Delta \mathbf{x}_{ij}$ představuje rozdíl hodnot vysvětlujících proměnných pro i-té a j-té pozorování (tj. $\mathbf{x}_i - \mathbf{x}_j$) a n je počet pozorování. Gönenův a Hellerův index konkordance má stejně jako Harrelovo C vazbu na Somersovo d (platí: $d = 2C_{GH} - 1$).

4.3.2. Brierovo skóre

Brierovo skóre je vedle konkordance další možností, jak lze posoudit kvalitu nalezeného modelu. Jedná se o odhad čtvercové odchylky očekávaných hodnot od hodnot pozorovaných [4], [10] tj.:

$$BS(t) = E(Y_i(t) - \hat{S}(t, x_i))^2, \tag{4.23}$$

kde $Y_i(t)$ představuje pozorovanou a $\hat{S}(t, x_i)$ očekávanou hodnotu pro i-tý subjekt v čase t. Očekávanou hodnotu lze přitom vypočítat podle vztahu (2.11), ve kterém pouze dosadíme za vektor neznámých parametrů $\boldsymbol{\beta}$ příslušný odhad $\hat{\boldsymbol{\beta}}$ získaný metodou maximální věrohodnosti. Brierovo skóre hodnotí přesnost odhadu funkce přežití v čase. V případě, že bychom neměli přítomna žádná cenzorovaná pozorování, pak bychom hodnotu Brierova skóre v čase t vypočítali následovně:

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left(I(t_i > t) - \hat{S}(t, x_i) \right)^2, \tag{4.24}$$

kde $I(\cdot)$ je indikátorová funkce (bude rovna jedné, pokud je pozorovaný čas i-tého subjetu vyšší než je uvažovaný čas t, v opačném případě bude rovna nule) a n je počet pozorování. Pro praxi je však přítomnost cenzorovaných pozorování běžnou situací a pro výpočet

Brierova skóre je třeba použít vztah, který tato cenzorování zohledňuje. [5] Označme si cenzorovaný čas jako t_C. Nechť dále $G(t) = P(t_C > t)$, resp. $G(t_i) = P(t_C > t_i)$ značí pravděpodobnost toho, že k cenzorování dojde až po čase t, resp. až po čase t_i. Jedná se o analogii k funkci přežití, kdy je předmětem zájmu namísto času cenzorování čas výskytu sledované události. Podobně jako lze vypočítat Kaplanův-Meierův odhad funkce přežití, lze vypočítat také odhad pravděpodobnosti cenzorování. Označíme-li si tyto odhady jako $\hat{G}(t)$ a $\hat{G}(t_i)$, lze vzorec pro výpočet Brierova skóre v čase t psát ve tvaru:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(t_i \leq t, \delta_i = 1)(0 - \hat{S}(t, x_i))^2}{\hat{G}(t_i)} + \frac{I(t_i > t)(1 - \hat{S}(t, x_i))^2}{\hat{G}(t)}\right), \tag{4.25}$$

kde δ_i je indikátorová funkce, přičemž $\delta_i = 1$ nastane v případě, kdy u i-tého pozorování došlo v čase t_i ke sledované události. Vzorec pro výpočet Brierova skóre ve tvaru (4.25) je v podstatě váženou verzí vzorce (4.24), přičemž váhy jsou rovny $1/\hat{G}(t_i)$ pokud k události došlo před časem t a $1/\hat{G}(t)$, pokud k události došlo až po čase t. Pokud je pozorování cenzorováno již před časem t, potom není do výpočtu hodnoty Brierova skóre pro čas t vůbec zařazeno.

Abychom získali jednu hodnotu, která bude mít vypovídací hodnotu o kvalitě (přesnosti) predikce sestaveného modelu, můžeme se podívat na integrované Brierovo skóre, tj.:

$$IBS = \frac{1}{max(t_i)}\int_0^{max(t_i)} BS(t)dt. \tag{4.26}$$

V ideálním případě chceme, aby byla hodnota Brierova skóre nižší než 0.25, neboť tato hodnota odpovídá situaci, kdybychom každému subjektu přiřadili rizikový skór 0.50. Hodnota Brierova skóre 0.33 by pak znamenala, že rizikové skóry jednotlivým subjektů přiřazujeme naprosto náhodně (z intervalu $\langle 0, 1\rangle$). Dobré je taky porovnat takto získanou hodnotu Brierova skóre s integrovaným Brierovým skórem pro nulový model (model pouze s absolutním členem). Jak už bylo uvedeno výše, Brierovo skóre měří čtvercovou odchylku pozorovaných a predikovaných hodnot. Alternativou je pak přístup, kdy se odhadují absolutní odchylky pozorovaných a predikovaných hodnot [16].

Kapitola 5

Praktická část

V této části práce provedeme praktickou analýzu dat s využitím znalostí z předchozích kapitol. K analýze bude využit statistický software R, přičemž zde budou uvedeny použité příkazy a funkce. V případě použití jiných než základních balíčků softwaru R bude uveden odpovídající název knihovny.

5.1. Popisná statistika

Datový soubor, který budeme analyzovat, se týká přežívání u rakovinného onemocnění gastrointestinálního traktu (dále budeme značit GIT). Sledovanou událostí uvažované studie je smrt následkem tohoto onemocnění, přičemž čas je uvažován od operace pacienta. K dispozici máme celkem osm proměnných a 119 pozorování. Pro názornost je v tabulce 5.1 ukázka prvních šesti řádků tohoto datového souboru. V prvním sloupci je informace o věku pacienta. Druhý sloupec nese informaci o tom, zda se jedná o ženu (1) či muže (0). Třetí sloupec s názvem OSevent (OS – overall survival) nám dává informaci o tom, zda pacient zemřel (bez ohledu na příčinu), přičemž OSevent = 0 značí cenzorování a OSevent = 1 smrt. Ve čtvrtém sloupci je uváděn čas v měsících (počítáno od operace pacienta), po který byl pacient ve studii. Pátý sloupec s názvem CSSevent (CSS – cancer-specific survival) obsahuje informaci o tom, zda u pacienta došlo ke sledované události (CSSevent = 1) nebo se jedná o cenzorované pozorování (CSSevent = 0). Následuje sloupec s informací, zda pacient podstoupil (1) nebo nepodstoupil (0) léčbu chemoterapií. Předposlední sloupec s názvem Grade označuje úroveň nemoci. V datovém souboru jsou rozlišovány celkem tři

různé úrovně: G1, G2 a G3. Obecně lze říci, že čím vyšší tato úroveň je, tím jsou nádorové buňky agresivnější a také rychleji rostou. Poslední sloupec s názvem Stage pak označuje stádium rakovinného onemocnění – zde rozlišujeme stádia: IA, IB, IIA, IIB a IIIA. Stádium onemocnění je určeno velikostí nádoru a také tím, jaké tkáně jsou nádorovými buňkami zasaženy - vyšší označení odpovídá závažnějšímu stádiu onemocnění. [14]

age	sex	OSevent	OSm	CSSevent	CHT	Grade	Stage
66	1	1	36.34	0	0	G2	IA
64	1	0	48.99	0	0	G2	IB
62	1	0	38.47	0	1	G2	IA
54	0	1	7.62	1	1	G3	IB
77	0	0	41.92	0	0	G2	IA
77	1	1	41.56	0	0	G2	IA

Tabulka 5.1: Ukázka datového souboru pro analýzu přežití u onemocnění GIT

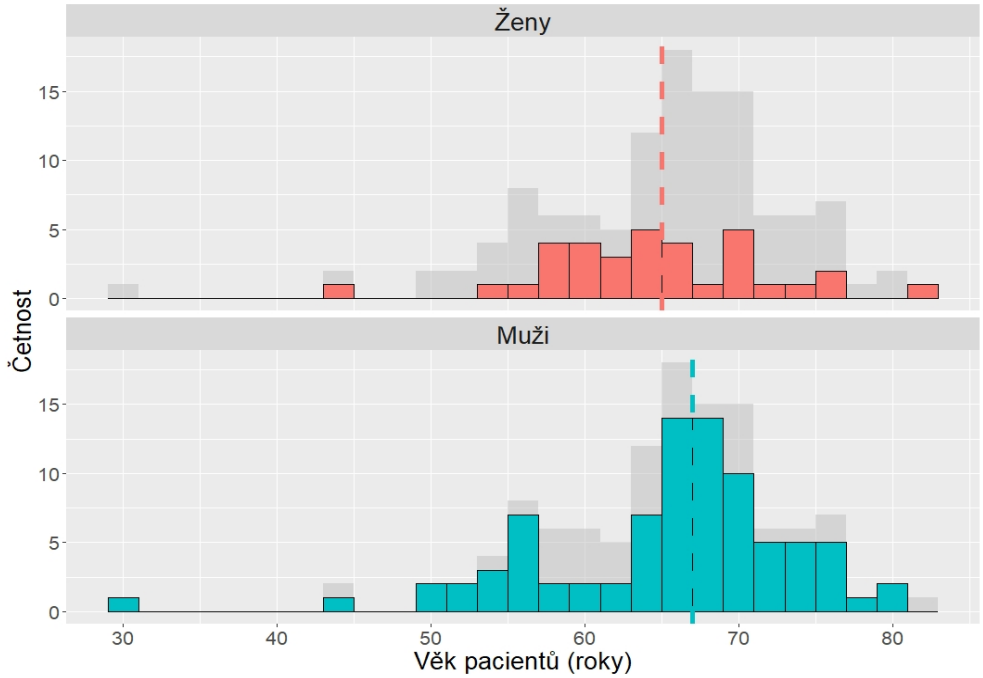
Použitím příkazu `summary` získáme pro proměnné věk a čas základní číselné charakteristiky (minimum, maximum, dolní a horní kvartil, průměr a medián) a pro zbylé proměnné počty pozorování u jednotlivých úrovní dané faktorové proměnné.

```
> summary(GIT)
      age        sex    OSevent       OSm            CHT    CSSevent
 Min.   :29.00   1:34   0:79     Min.   : 0.5257    0:51   0:93
 1st Qu.:61.00   0:85   1:40     1st Qu.: 9.0351    1:68   1:26
 Median :67.00                   Median :15.9018
 Mean   :65.67                   Mean   :18.9537
 3rd Qu.:70.00                   3rd Qu.:29.0930
 Max.   :82.00                   Max.   :48.9866

Grade   stage
G1:10   IA  :37
G2:39   IB  :27
G3:70   IIA :19
        IIB :20
        IIIA:16
```

Vidíme, že pacienti, kteří se zúčastnili studie, byli ve věku mezi 29 a 82 roky, přičemž zde bylo 34 žen (29 %) a 85 mužů (71 %). Doba sledování se pohybovala mezi půl měsícem a přibližně čtyřmi lety. Smrt následkem onemocnění GIT nastala celkem u 26 pacientů (22

%). Z toho vyplývá, že celkem 14 pacientů (12 %) zemřelo následkem jiné příčiny, neboť během studie zemřelo (z jakékoli příčiny) celkem 40 pacientů (34 %). Léčbu chemoterapií pak postoupilo 68 pacientů (57 %). Pokud se podíváme na proměnnou Grade, vidíme, že nejvíce pacientů mělo onemocnění na úrovni G3, což odpovídá rychle se rozrůstající nádorové tkáni a naopak nejméně pacientů mělo onemocnění ve stádiu G1. Co se týče stádia onemocnění, nejvíce pacientů bylo zařazeno do skupin IA a IB, které značí malý nádor s nejmenším rozsahem zasažení tkání.

Nyní se pro lepší představu o věku pacientů se podívejme na obrázek 5.1, kde jsou vykresleny histogramy věku zvlášť pro ženy (červeně) a zvlášť pro muže (modře). V his-



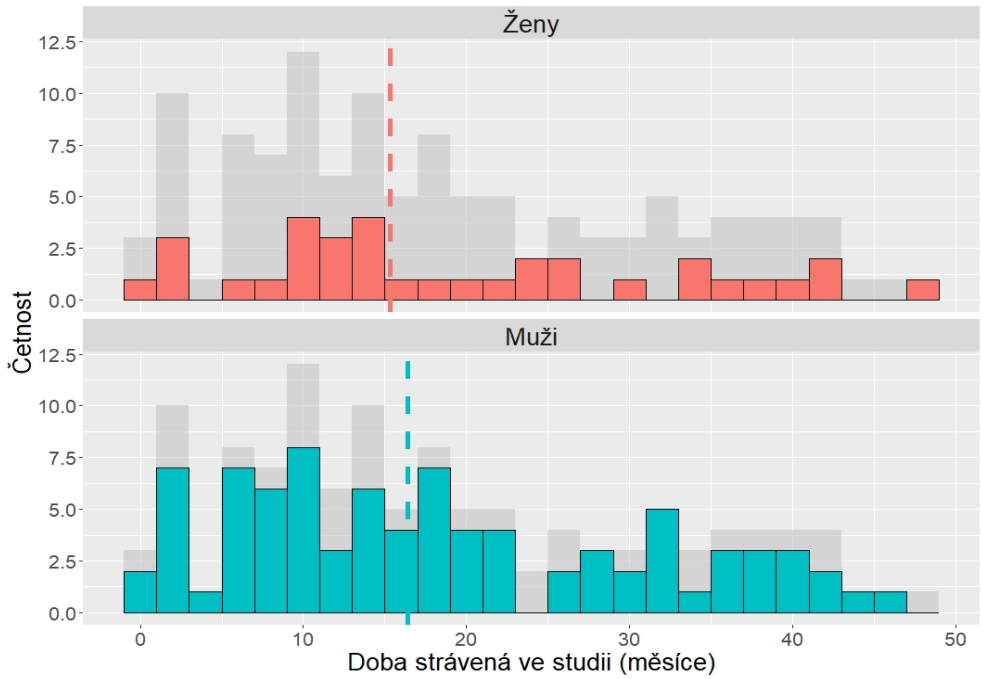
Obrázek 5.1: Histogramy věku pro ženy a pro muže s vyznačenými mediány

togramech jsou vyznačeny mediány – pro ženy (65 let) a pro muže (67 let). V pozadí obou těchto histogramů jsou šedou barvou vykresleny histogramy věku všech pacientů dohromady (bez rozlišení pohlaví). Graf byl vytvořen s využitím knihoven `ggplot2` a `plyr` následujícími příkazy:

```
> GIT_bg = GIT[, -2]
> mu = ddply(GIT, "sex", summarise, grp.median=median(age))
> ggplot(GIT, aes(x = age, fill = sex)) +
  geom_histogram(data = GIT_bg, alpha = 0.5, binwidth = 2,
                          colour = "black", fill = "grey") +
  labs(x = "Věk pacientů (roky)", y = "Četnost") +
  geom_vline(data=mu, aes(xintercept = grp.median,color = sex),
                  linetype = "dashed",size = 2, show.legend = F)+
  facet_wrap(~ sex, ncol = 1) + guides(fill = F) +
  theme(text = element_text(size = 20), strip.text.x = element_text(size = 20))
```

Z histogramů výše vidíme, že jedno pozorování se od ostatních výrazně liší. Toto pozorování odpovídá muži ve věku 29 let. Pomocí příkazu `boxplot.stats("GIT$age")` si můžeme ověřit, že se opravdu jedná o odlehlé pozorování. Zároveň tímto příkazem můžeme zjistit, že i další dvě pozorování jsou považována za odhlehlá, jedná se přitom o muže ve věku 44 let a ženu ve věku 45 let.

Histogramy si můžeme vykreslit také pro dobu (v měsících) strávenou ve studii (obrázek 5.2), a to analogickými příkazy jakým byly vytvořeny předchozí histogramy. Opět jsme



Obrázek 5.2: Histogramy délky doby ve studii pro ženy a muže s vyznačenými mediány

přitom barevně rozlišili histogram pro ženy (červeně) a histogram pro muže (modře) a vyznačili odpovídající mediány. Medián délky doby ve studii je přibližně 15 měsíců pro ženy a asi 16 měsíců pro muže. Z obrázku také vidíme, že mezi pohlavími nejspíše není signifikantní rozdíl co se týče délky doby ve studii, tj. čas přežití při této nemoci nejspíše nebude příliš ovlivněn pohlavím.

Dále se můžeme podívat na hodnoty korelačních koeficientů mezi jednotlivými proměnnými. Tím si uděláme představu o tom, které proměnné se navzájem ovlivňují, a které proměnné by mohly mít významný vliv na čas přežití. Společnou vlastností všech koeficientů, které zde využijeme, je totiž fakt, že nabývají hodnot od mínus jedné do jedné. Hodnoty blízké mínus jedné přitom značí záporný vztah (velké hodnoty jedné proměnné jsou asociovány s malými hodnotami druhé proměnné), hodnoty blízké jedničce značí kladný vztah (velké hodnoty jedné proměnné jsou asociovány s velkými hodnotami druhé proměnné) a hodnoty blízké nule pak znamenají, že mezi danými veličinami není žádný lineární vztah (závilost mezi hodnotami existovat může, ale nelze ji popsat lineárně). Uvažovat zde přitom nebudeme proměnnou OSevent. Než si koeficienty korelace vypočítáme, upravíme si ještě proměnnou Stage. Namísto dělení do skupin: IA, IB, IIA, IIB, IIIA, budeme uvažovat pouze jednodušší dělení, a to: I, II a III.

Pro popsání míry asociace mezi dvěmi kategoriálními proměnnými využijeme Cramerovo V, které je v případě dvou dichotomických proměnných rovno koeficientu Phi. V softwaru R jeho hodnotu získáme pomocí příkazu `CramerV`, který nalezneme v knihovně `DescTools`. Výsledné hodnoty pro naše proměnné jsou uvedeny v tabulce 5.2. Nejvyšší hodnoty je dosaženo u proměnných chemoterapie a stádium onemocnění, druhé u proměnných stádium onemocnění a CSS event a třetí u proměnných chemoterapie a úroveň onemocnění. Ve všech třech případech se přitom jedná o kladný vztah a vyšší hodnoty jedné proměnné jsou asociovány s vyššími hodnami druhé proměnné. Uvažujme například nejsilnější zjištěný vztah (chemoterapie a stádium onemocnění). Zde nám hodnota korelačního koeficientu říká, že podstoupení léčby chemoterapií je spojeno s vyšším stádiem onemocnění. Současně vidíme, že pro pohlaví a CSS event je tato hodnota velmi nízká (blízká nule), což je v souladu s tím, co jsme si uvedli již dříve.

Proměnné	Cramerovo V
pohlaví - chemoterapie	0.129
pohlaví - CSS event	0.064
pohlaví - grade	0.078
pohlaví - stage	0.091
chemoterapie - CSS event	0.211
chemoterapie - grade	0.284
chemoterapie - stage	0.517
CSS event - grade	0.217
CSS event - stage	0.326
grade - stage	0.096

Tabulka 5.2: Hodnoty Cramerova V pro kategoriální proměnné

Pro popsání síly a směru vztahu mezi spojitou a dichotomickou proměnnou využijeme bodového biseriálního koeficientu. K výpočtu lze využít příkaz `biserial.cor` z knihovny `ltm`. Protože je bodový biseriální koeficient roven Pearsonovu korelačnímu koeficientu, lze využít také příkaz `cor.test`. Vypočítané hodnoty pro naše proměnné jsou v tabulce 5.3. Lze říci, že kromě poslední kombinace proměnných (doba strávená ve studii a CSS event) jsou všechny uvažované korelace nevýznamné. Mezi dobou strávenou ve studii a proměnnou, která říká, zda nastala sledovaná událost, je přitom vztah záporný. Nastání sledované události je tedy asociováno s kratší dobou strávenou ve studii.

Proměnné	Bodový biseriální koeficient
věk - pohlaví	0.060
věk -chemoterapie	-0.124
věk -CSS event	0.123
doba ve studii - pohlaví	-0.045
doba ve studii - chemoterapie	-0.088
doba ve studii - CSS event	-0.282

Tabulka 5.3: Hodnoty bodového biseriálního koeficientu korelace pro spojité a dichotomické proměnné.

V případě, že kategoriální proměnná nabývá více než dvou hodnot, nelze použít pro posání vztahu mezi spojitou proměnnou a touto kategoriální proměnnou bodový biseriální koeficient. Využít ale můžeme jeho zobecnění, a to polyseriální koeficient korelace. Pro jeho výpočet lze v softwaru R využít příkaz `polyserial` z knihovny `polycor`. V tabulce 5.4 jsou

uvedeny vypočítané hodnoty. Vidíme, že nejvyšší hodnota korelace náleží době strávené ve studii a stádiu onemocnění. Zároveň je tato hodnota záporná, což znamená, že s nižší stupněm onemocnění souvisí delší doba strávená ve studii a zároveň také delší doba přežití.

Proměnné	Polyseriální koeficient korelace
věk - stage	-0.208
věk - grade	0.173
doba ve studii -stage	-0.289
doba ve studii - grade	0.006

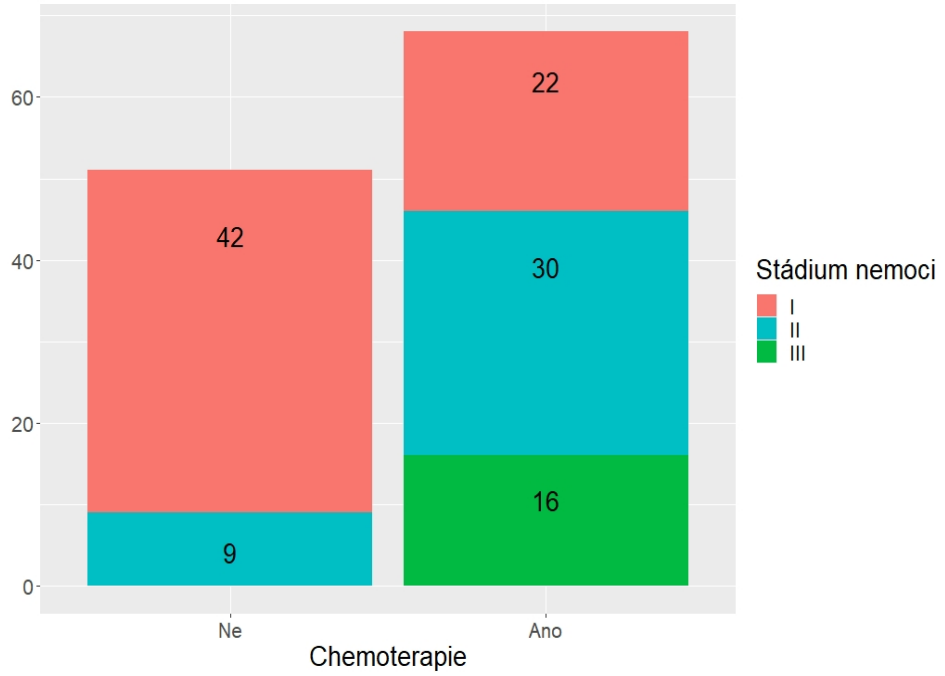
Tabulka 5.4: Hodnoty polyseriálního koeficientu korelace pro spojité a kategoriální proměnné.

Nakonec pro kombinaci dvou spojitých proměnných můžeme využít Spearmanova korelačního koeficientu. Výpočet lze provést pomocí příkazu `cor.test`. Hodnota Spearmanova korelačního koeficientu je pro věk a dobu strávenou ve studii přibližně 0.027. Jedná se o hodnotu velmi blízkou nule, z čehož vyplývá, že mezi věkem a dobou strávenou ve studii není žádný lineární vztah. Je proto na místě domnívat se, že v našem případě nebude věk významným prediktorem.

Nejsilnější vztah jsme identifikovali mezi proměnnými chemoterapie a stádiem onemocnění. Druhý nejsilnější vztah je mezi proměnnou CSS event a stádiem onemocnění a třetí mezi dobou strávenou ve studii a proměnnou CSS event. Nyní se podívejme na grafické vizualizace těchto tří závilostí.

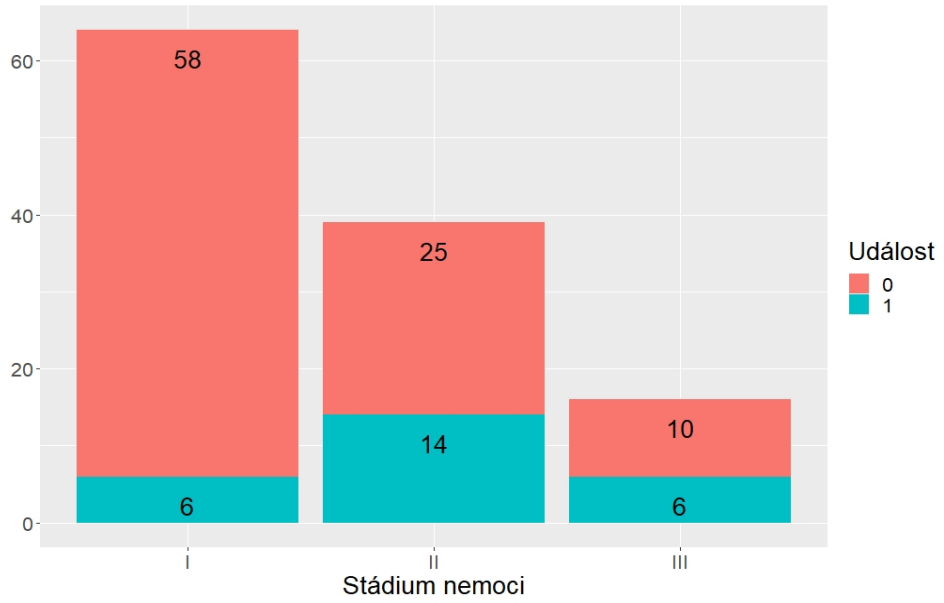
Začněme závislostí nejsilnější. Na obrázku 5.3 je graf znázorňující počty pacientů kteří (ne)podstoupili chemoterapii, který je navíc barevně rozlišen podle stádia onemocnění. Vidíme, že všichni pacienti, kteří byli ve stádiu onemocnění III, léčbu chemoterapií podstoupili (100 %). Tuto léčbu podstoupila také většina pacientů ve stádiu onemocnění II (77 %). Naopak tomu je u skupiny pacientů ve stádiu onemocnění I, kde chemoterapii podstoupila menšina (34 %). Graf je vytvořen s využitím knihovny `ggplot2` příkazem:

```
> ggplot(GIT,aes(x=factor(CHT),fill=factor(Stage_upr)))+
  geom_bar()+ labs(x='Chemoterapie', y = "", fill='Stádium nemoci')+
  geom_text(aes(label=..count..),stat="count",
            position=position_stack(0.9),vjust=2, size = 7)+
  theme(text = element_text(size = 20))+
  scale_fill_manual("Stádium nemoci",
            values = c("I" = "#F8766D", "II" = "#00BFC4", "III" = "#00BA42"))
```



Obrázek 5.3: Počty pacientů, kteří (ne)podstoupili chemoterapii – barevně rozlišené dle stádia nemoci

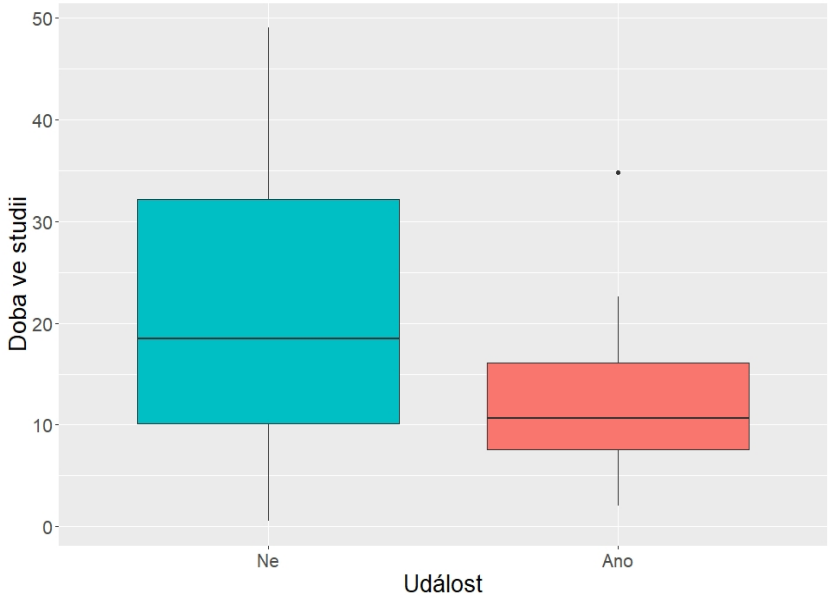
Druhá nejsilnější závilost byla zjištěna mezi proměnnými CSS event a Stage. Na obrázku 5.4 je graf znázorňující počty pacientů v jednotlivých stádiích nemoci, které jsou navíc barevně rozlišeny podle toho, zda u pacienta došlo ke sledované události. Vidíme, že ve stádiu I bylo 64 pacientů a z toho pouze u 6 z nich ke sledované události došlo, tj. ke sledované události došlo ve stádiu I u 10 % pacientů. Ve stádiu II došlo ke sledované události u 14 z 39 pacientů, to znamená u přibližně 36 % pacientů. Ve stádiu III pak bylo celkem 16 pacientů a ke sledované události došlo u šesti z nich, tj. u přibližně 38 %.



Obrázek 5.4: Počty pacientů v jednotlivých stádiích nemoci – barevně rozlišené dle proměnné CSS event

Třetí nejsilnější závislost je mezi dobou strávenou ve studii a proměnnou, která říká, zda došlo ke sledované události. Na obrázku 5.5 jsou postupně vykresleny boxploty délky doby strávené ve studii pro skupinu pacientů u nichž nedošlo ke sledované události a pro skupinu pacientů, kteří na onemocnění GIT zemřeli. Vidíme, že medián doby strávené ve studii, je vyšší pro pacienty, kteří sledovanému odemocnění GIT nepodlehli. Dále je vidět, že pro tuto skupinu pacientů se časy strávené ve studii pohybují prakticky po celé uvažované časové ose, tj. až do přibližně 50. měsíce od operace, zatímco u druhé skupiny pacientů se časy pohybují pouze do 25. měsíce (s jedinnou výjimkou, kdy u pacienta došlo ke sledované události až v 35. měsíci). Graf byl vytvořen v knihovně `ggplot2` příkazem:

```
> ggplot(GIT, aes(x=CSSevent, y=OSm, fill=CSSevent)) +
  geom_boxplot()+
  labs(x='', y = "Doba ve studii",fill="Událost")+
  theme(legend.position = "none", text = element_text(size = 20))
```

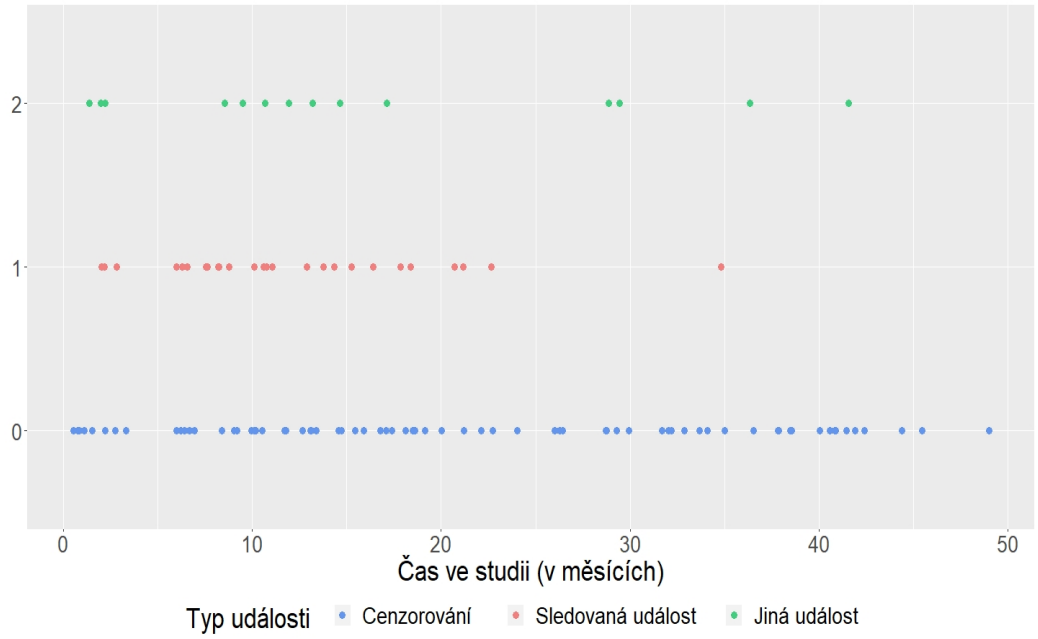


Obrázek 5.5: Boxploty délky doby ve studii dle proměnné CSS event

5.2. Analýza přežití

V této podkapitole se už budeme věnovat samotné analýze přežití. Připomeňme, že máme k dispozici celkem 119 pozorování, přičemž u 26 z nich došlo ke sledované události – tedy ke smrti následkem onemocnění GIT. U zbylých 93 pozorování k této události nedošlo (jde tedy o cenzorovaná pozorování), nicméně víme, že u 14 pozorování došlo ke smrti z jiné příčiny – což lze označit za konkurenční riziko, neboť po této události již námi sledovaná událost nastat nemůže. Na obrázku 5.6 jsou vykresleny časy ve studii pro jednotlivá pozorování, a to s ohledem na skutečnost, zda jde o pozorování u něhož ke sledované události došlo, nebo došlo k jiné události, anebo nedošlo k žádné události. Pro účely vytvoření tohoto grafu byla vytvořena nová proměnná "typ", kde typ = 0 označuje cenzorované pozorování, typ = 1 sledovanou událost a typ = 2 jinou než sledovanou událost. Můžeme si povšimnout, že čas strávený ve studii pro cenzorovaná pozorování se pohybuje rovnoměrně po celé časové ose, zatímco doba přežití subjektů u nichž došlo ke sledované události se pohybuje (až na jednu výjimku) maximálně do 23. měsíce. Z toho bychom mohli usuzovat, že pokud má u pacienta dojít ke sledované události, nejspíše se tak stane v tomto období. Graf byl opět vytvořen s využitím knihovny `ggplot2`, a to příkazem:

```
> ggplot(GIT, aes(x = OSm, y = typ))+
  geom_point(aes(colour = typ), size = 3, show.legend = T)+
  labs(x="Čas ve studii (v měsících)", y = "", colour = "Typ události")+
  theme(text = element_text(size = 20), legend.position = "bottom")+
  scale_color_manual(
          labels = c("Cenzorování", "Sledovaná událost","Jiná událost"),
          values = c("cornflowerblue", "lightcoral", "seagreen3"))
```

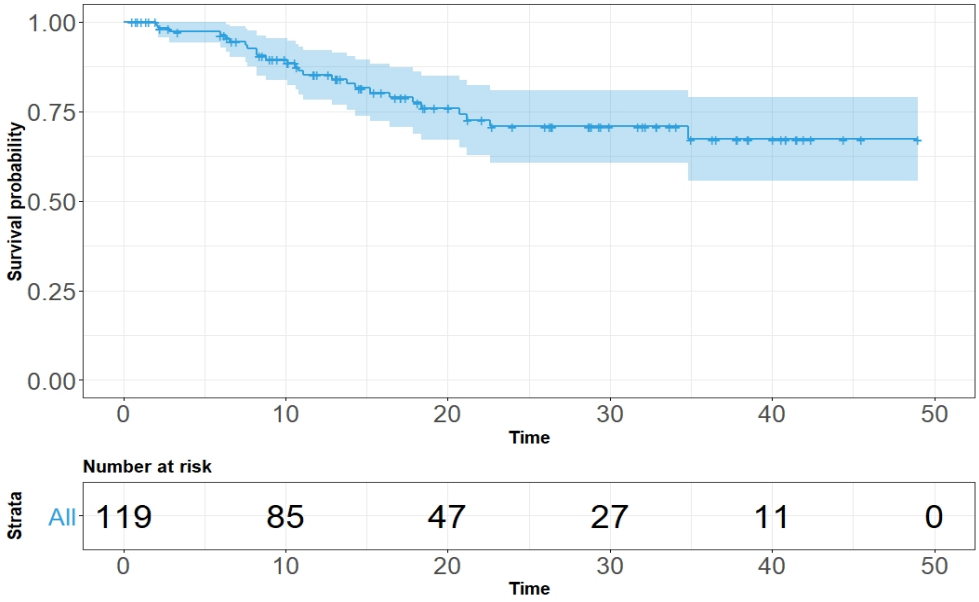


Obrázek 5.6: Doba přežití dle typu události

5.2.1. Kaplanovy-Meierovy křivky přežití

Jedním ze základních nástrojů využívaných v analýze přežití jsou Kaplanovy-Meierovy křivky přežití, které nám poskytují odhad funkce přežití. Začněme tedy jejich vykreslením. Nejříve uvažujme pouze proměnnou CSSevent, která říká, zda u pozorování došlo ke sledované události. Kaplanův-Meierův odhad funkce přežití a odpovídající křivku přežití (obrázek 5.7) získáme pomocí následujících příkazů (s využitím knihoven `survival`, `survminer` a `ggplot2`):

```
> surv_object = Surv(time=GIT$OSm, event= GIT$CSSevent)
> kmfit1 <- survfit(surv_object ~ 1, conf.type = "plain")
> ggsurv = ggsurvplot(kmfit1, data=GIT, risk.table = TRUE,
           palette = "#2E9FDF",ggtheme = theme_bw(),
           legend='none', risk.table.fontsize = 10)
> ggpar(ggsurv, font.tickslab = c(20,"plain"), font.title = c(15,"bold"),
      font.x = c(15,"bold"), font.y = c(15,"bold"))
```



Obrázek 5.7: Základní křivka přežití pro proměnnou CSSevent.

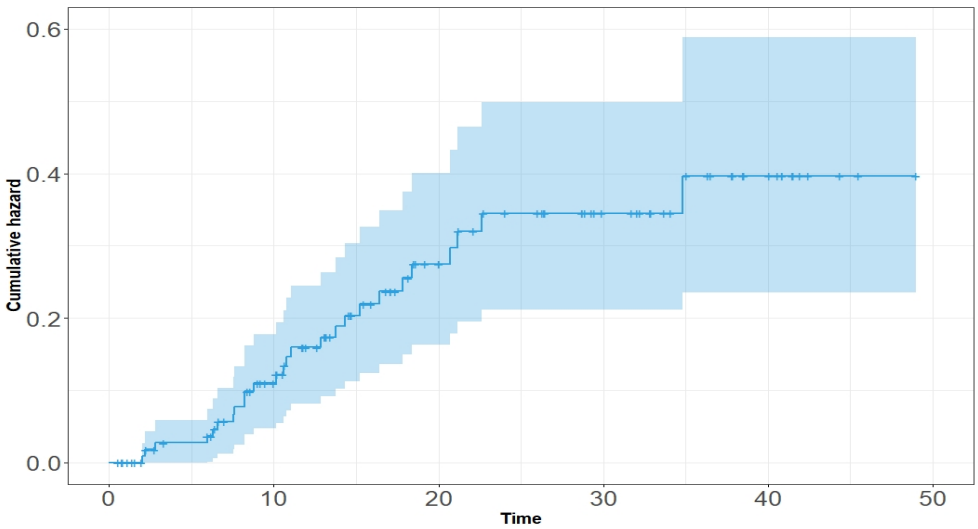
V hlavní části grafu na obrázku 5.7 je vykreslena křivka přežití, vyznačené křížky na této křivce odpovídají cenzorovaným pozorováním. Z grafu vidíme, že hodnota odhadu funkce přežití s postupem času klesá, nicméně přibližně po 23. měsíci se stabilizuje na hodnotě asi 0.70. Jinými slovy, pravděpodobnost přežití se ode dne operace snižuje, a to po dobu 23 měsíců, po kterých se ustálí na hodnotě 70 %. V grafu je také vyznačen pás spolehlivosti kolem křivky přežití, který s pravděpodobností 95 % pokrývá skutečnou hodnotu pravěpodobnosti přežití v jednotlivých časech. V tabulce pod grafem je pak pro každý desátý měsíc uvedený počet pozorování, která jsou v riziku, tzn. počet subjektů u nichž do daného času nedošlo k události a zároveň neopustili studii.

Zajímá-li nás průměrná doba přežití, lze využít následujícího příkazu:

```
> print(kmfit1, print.rmean = TRUE)

Call: survfit(formula = surv_object ~ 1)
     n     events     *rmean *se(rmean)     median    0.95LCL    0.95UCL
119.00      26.00      37.85       1.86         NA         NA         NA
    * restricted mean with upper limit =  49
```

Ve výstupu se zobrazuje hned několik údajů, přičemž první dva informují o počtu pozorování a počtu sledovaných událostí jež nastaly. Následuje průměrná doba přežití, která je téměř 38 měsíců. Protože nejdelší čas přežití ve studii je cenzorovaný, jedná se o tzv. restringovanou (omezenou) průměrnou dobu přežití. Jako horní limit pro výpočet tohoto průměru je brán právě onen nejdelší čas přežití ve studii (49 měsíců). Následuje směrodatná chyba průměrné doby přežití. Poslední tři údaje se týkají mediánu přežití a jeho horního a dolního intervalového odhadu. Připomeňme, že medián přežití nalezneme jako časový okamžik, ve kterém křivka přežití dosahuje hodnoty 0.50 (pravděpodobnost přežití je 50 %). V našem případě je ale pravděpodobnost přežití po celou uvažovanou dobu vyšší než 50 %, proto nejsou tyto tři hodnoty definované.

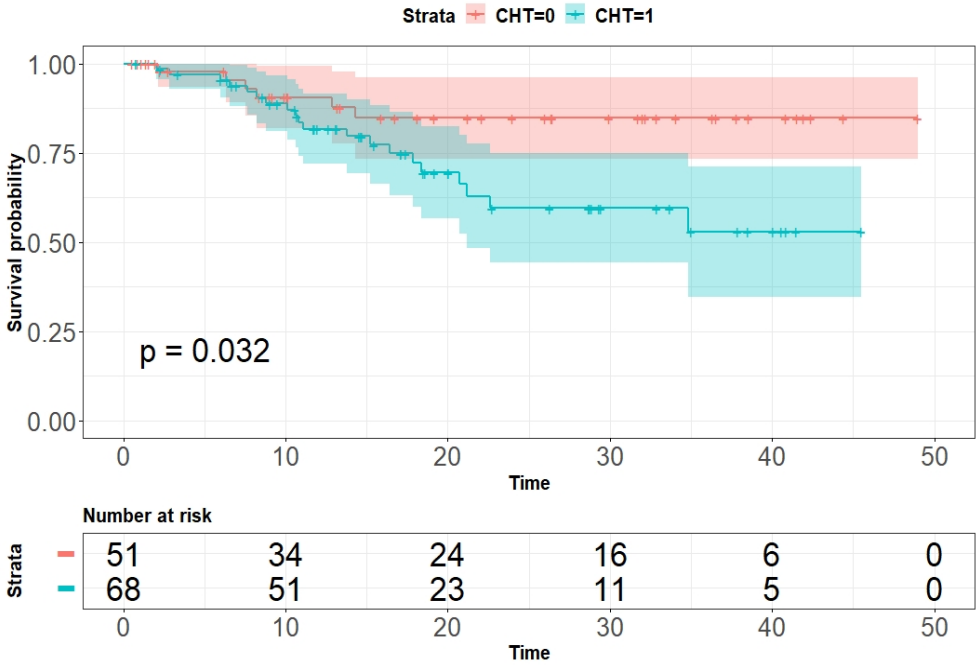
Nakonec si můžeme vykreslit ještě kumulativní rizikovou funkci (obrázek 5.8), která vyjadřuje celkové riziko toho, že od začátku sledování až do nějakého zvoleného času t dojde ke sledované události.



Obrázek 5.8: Kumulativní riziková funkce pro základní model

Jinými slovy, kumulativní riziková funkce udává očekávaný počet sledovaných událostí, které by u jedince nastaly od začátku sledování až do času t (za předpokladu, že by se daná událost mohla opakovat).Z grafu na obrázku 5.8 vidíme, že v našem případě dosahuje kumulativní riziková funkce nejvýše hodnoty 0.4, což je pozitivní číslo – znamená to totiž, že neočekáváme výskyt sledované události u každého pacienta. Toto samozřejmě odpovídá již výše uvedenému. Kolem křivky kumulativní rizikové funkce je vyznačen také 95% pás spolehlivosti, který v jednotlivých časových bodech pokrývá skutečnou hodnotu kumulativního rizika s pravděpodobností 95 %. Graf byl vytvořen analogickým způsobem jako graf funkce přežití (obrázek 5.7) s tím rozdílem, že mezi parametry funkce `ggsurvplot` jsme přidali argument `fun = "cumhaz"`.

Nyní si přidejme informaci o tom, zda pacient podstoupil nebo nepodstoupil léčbu chemoterapií. V takovémto případě budeme mít dva různé odhady funkce přežití, a tedy dvě křivky přežití. Z grafu na obrázku 5.9 vidíme, že lépe jsou na tom pacienti, kteří léčbu chemoterapií nepodstupují. Odpovídající křivka přežití se totiž přibližně po 14. měsíci stabilizuje na hodnotě 0.85, zatímco křivka přežití pro pacienty, kteří léčbu chemoterapií podstupují se stabilizuje až přibližně po 22. měsíci na hodnotě 0.60.

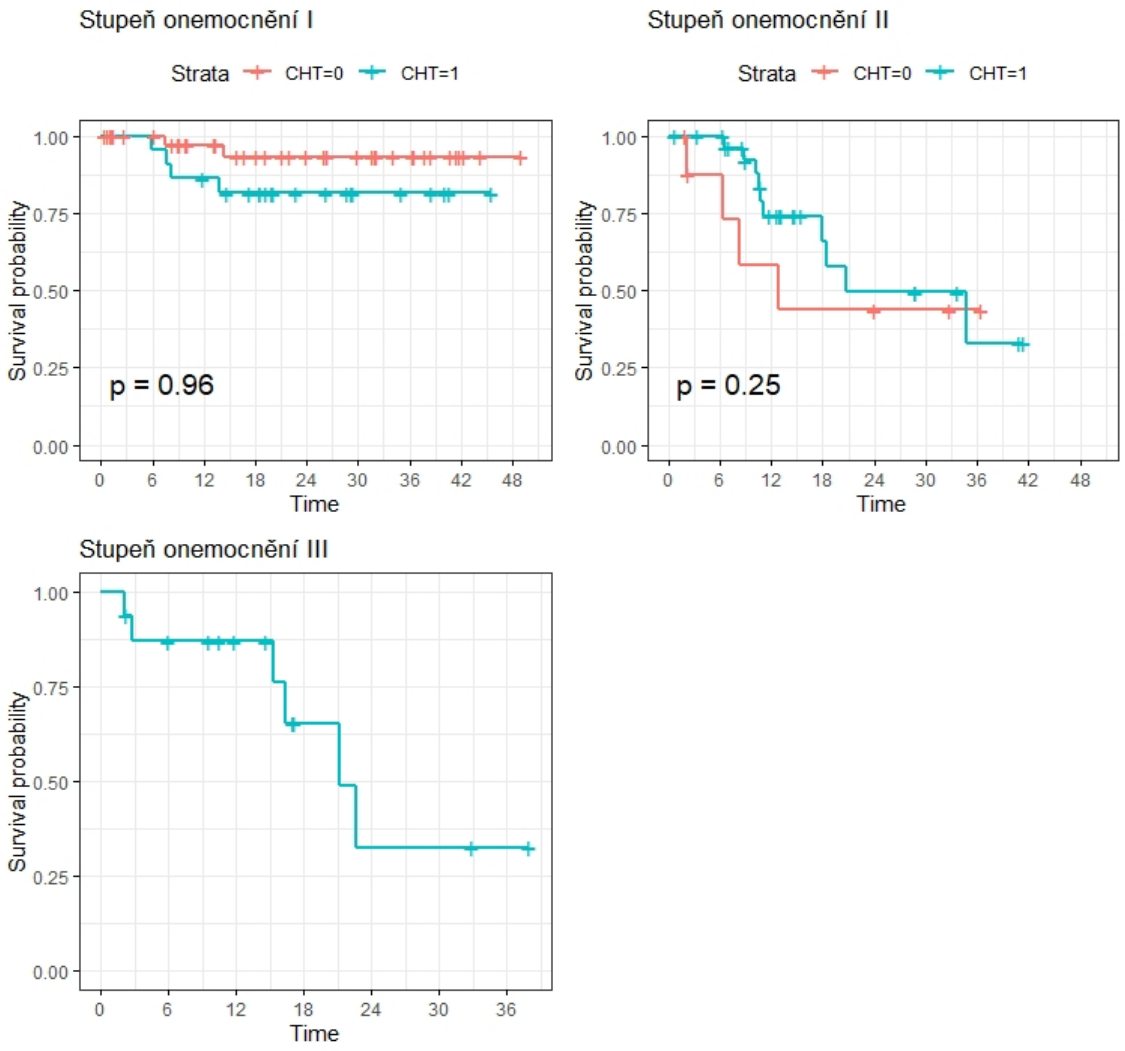


Obrázek 5.9: Křivky přežití pro skupiny dle proměnné CHT

Tento výsledek je sice překvapivý, nicméně může souviset právě se skutečností, že léčbu chemoterapií postupují zejména pacienti ve vážnějším stádiu onemocnění – viz obrázek 5.3. Připomeňme, že hodnota korelačního koeficientu mezi proměnnými CHT – (ne)podstoupení chemoterapie a Stage – stupeň onemocnění je 0.517. V tabulce pod grafem jsou opět pro každý desátý měsíc zobrazeny počty pozorování, které jsou v riziku, tentokrát jsou však rozděleny podle proměnné CHT. Oproti předchozímu grafu je zde navíc vypočítána p-hodnota, která odpovídá hodnotě log-rank testu, který slouží k rozhodnutí, zda se od sebe křivky pro dané skupiny významně liší. P-hodnota 0.032 přitom svědčí o tom, že mezi danými skupina skutečně signifikantní rozdíl je. Samostatně lze log-rank test v softwaru R provést s využitím knihovny `survival` pomocí příkazu `survdiff`. Poznamenejme ještě, že restringovaný průměrný čas přežití je 33 měsíců ($\pm$2.20) pro pacienty, kteří podstoupili léčbu chemoterapií a 41 měsíců ($\pm$2.52) pro pacienty, kteří tuto léčbu nepostoupili.

Vzhledem ke zjištěné závilosti mezi proměnnými CHT a Stage bude vhodné si vykreslit analogické křivky přežití pro jednotlivé stupně onemocnění (obrázek 5.10). V prvním grafu jsou křivky přežití pro stádium I. Vidíme, že obě křivky se drží nad hodnotou 0.75, což odpovídá tomu, že se jedná o nejlehčí stádium onemocnění. V tomto stádiu ale stále platí, že vyšší pravděpodobnost přežití mají pacienti, kteří léčbu chemoterapií nepodstupují. Nicméně dle p-hodnoty vidíme, že se od sebe tyto křivky významně neliší. Druhý graf je pro stádium II. Zde vidíme, že obě křivky dosahují hodnot menších než 0.50. Oproti stádiu I tedy pozorujeme velký pokles pravděpodobnosti přežití. Zároveň si můžeme povšimnout, že vyšší pravděpodobnost přežití má skupina, která léčbu chemoterapií podstupuje. Pravděpodobnosti přežití 50 % tato skupina dosahuje přibližně 21 měsíců od operace, zatímco pro skupinu, která léčbu nepodstoupila je medián času přežití jen 13 měsíců od operace. P-hodnota je v tomto případě nižší než u stádia I, nicméně stále platí, že mezi křivkami není signifikantní rozdíl. Poslední graf pak odpovídá stádiu III. V grafu je pouze jedna křivka přežití, neboť všichni pacienti z této skupiny léčbu chemoterapií podstoupili. Pravděpodobnost přežití je u těchto pacientů až do 15. měsíce vyšší než 85 % a až poté začne výrazně klesat. Medián času přežití je pak 21. měsíc od operace – tedy stejný čas jako pro pacienty ve stádiu II, kteří nepodstoupili léčbu chemoterapií. Oproti předchozím grafům jsme nyní záměrně vynechali pásy spolehlivosti kolem křivek přežití, a to zejména pro lepší přehled-

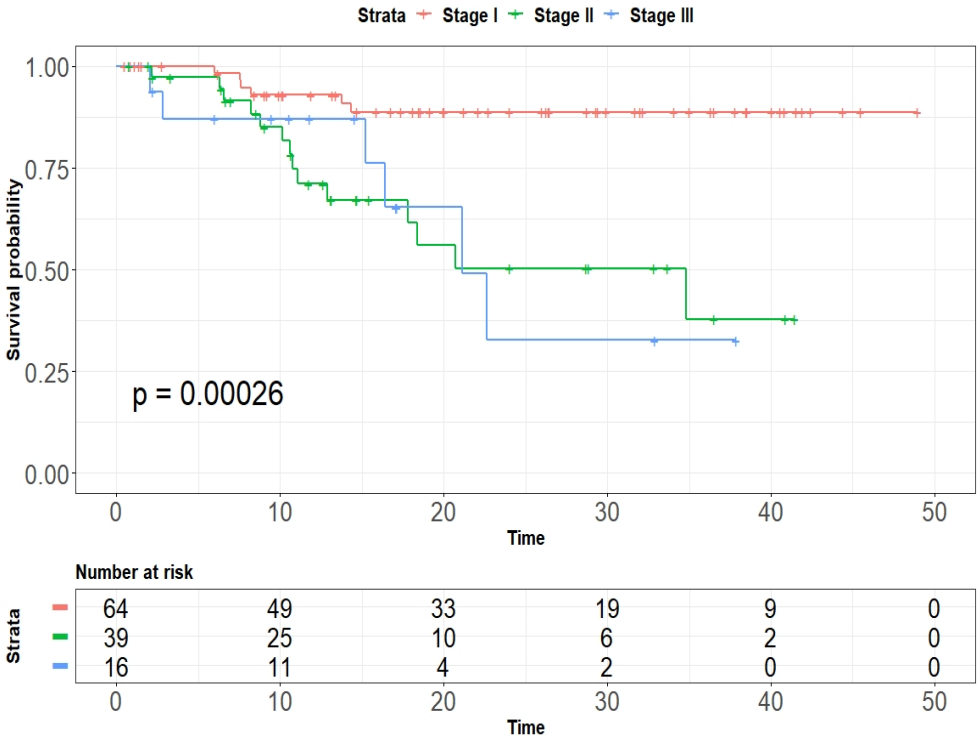
nost. Grafy na obrázku 5.10 byly vytvořeny opět v knihovně `ggplot2` a následně spojeny pomocí funkce `arrange_ggsurvplots`.



Obrázek 5.10: Křivky přežití pro jednotlivá stádia onemocnění - rozlišené dle proměnné CHT

Dále se můžeme podívat na to, jak je pravděpodobnost přežití ovlivněna jen samotnými stádii nemoci (opět budeme pro přehlednost uvažovat pouze základní dělení I, II a III). Na obrázku 5.11 jsou vykresleny příslušné křivky přežití pro daná stádia. Tyto křivky přežití samozřejmě korespondují s křivkami přežití z grafů na obrázku 5.10. Nyní ale můžeme křivky lépe porovnat, a to zejména křivky přežití pro stádium nemoci II a III. Z tohoto

grafu totiž vyplývá, že nelze obecně říci, která ze skupin pacientů (II nebo III) má vyšší pravděpodobnost přežití. Nízká p-hodnota je tedy nejspíše způsobena odlišností křivky přežití pro pacienty ve stádiu nemoci I, která se po celou dobu drží nad hodnotou 85 %. Pod grafem jsou opět uvedeny počty pacientů v riziku v daném čase rozdělené podle stádia nemoci.

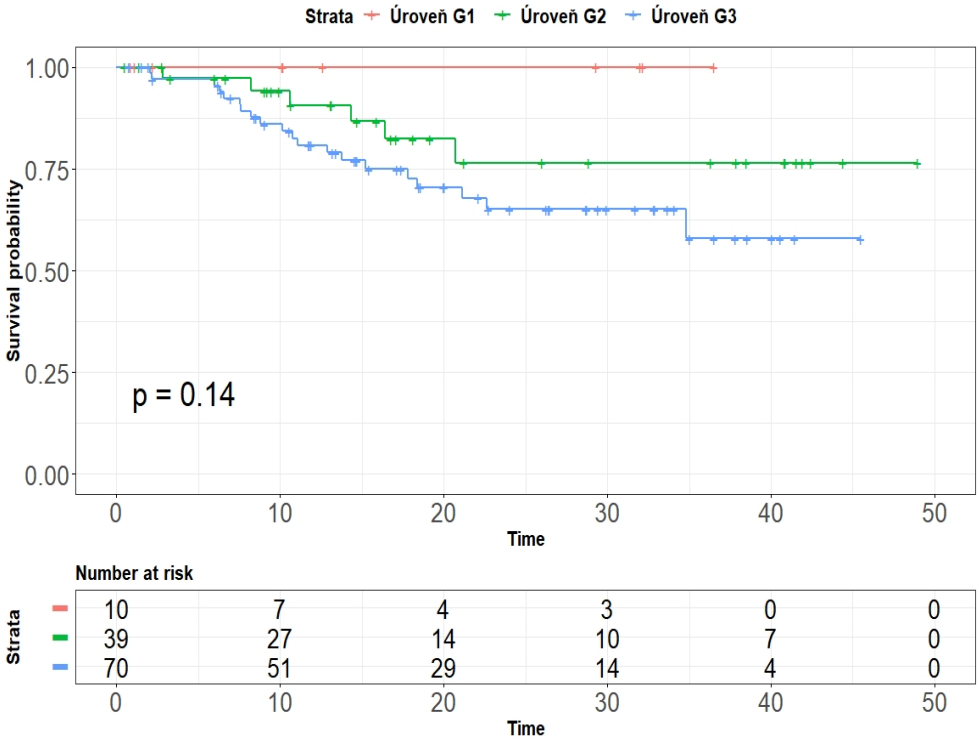


Obrázek 5.11: Křivky přežití pro jednotlivá stádia onemocnění

Podívat se můžeme ještě na průměrné časy přežití pro jednotlivá stádia nemoci. Z výstupu níže vidíme, že nejvyšší průměrný čas 37.9 měsíců (±1.37) odpovídá stádiu I, medián přežití pro tuto skupinu není definován (všichni pacienti v této skupině mají pravděpodobnost přežití vyšší než 50 %). Průměrný čas přežití pro pacienty ve stádiu II je 26.2 měsíců (±2.90) a medián přežití pak 34.8 měsíců od operace. Nicméně z grafu 5.11 vidíme, že pravděpodobnosti přežití 50 % je dosaženo již přibližně ve 21. měsíci od operace. Vypočítaný medián je zde ovlivněn faktem, že tato pravděpodobnost přežití zůstává konstantní téměř až do 35. měsíce od operace. Ve stádiu III této nemoci je průměrný čas přežití 24.4 měsíců (±4.40) od operace a medián přežití 21.2 měsíců od operace.

```
> print(survfit(surv_object ~ GIT$Stage_upr), print.rmean = TRUE)

Call: survfit(formula = surv_object ~ GIT$Stage_upr)

          n events *rmean *se(rmean) median 0.95LCL 0.95UCL
StageI   64      6   37.9       1.37     NA      NA      NA
StageII  39     14   26.2       2.90   34.8    17.8      NA
StageIII 16      6   24.4       4.40   21.2    16.4      NA
    * restricted mean with upper limit =  41.5
```

Analogickým způsobem si můžeme nechat vykreslit také křivky přežití pro proměnnou Grade (obrázek 5.12). Na první pohled vidíme, že pravděpodobnost přežití pro pacienty

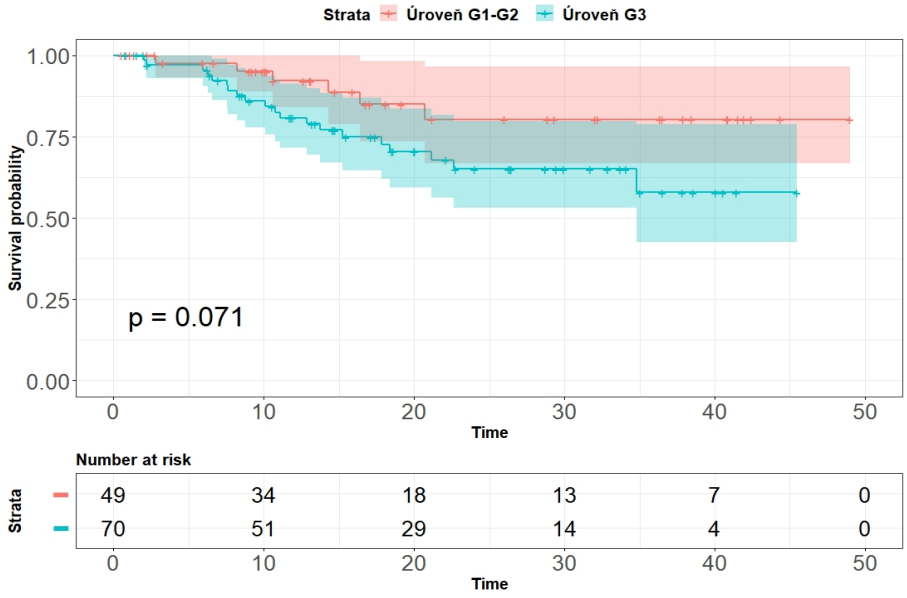


Obrázek 5.12: Křivky přežití pro jednotlivé úrovně onemocnění

s úrovní nemoci G1 je po celou uvažovanou dobu 100 %. Důvodem je fakt, že žádný z pacientů, který byl do této skupiny zařazen, na sledovanou nemoc nezemřel. Nicméně je třeba myslet na to, že do této skupiny patřilo pouze 10 pacientů z celkových 119. Je tedy možné, že je tento výsledek zapříčiněn malým počtem pozorování. Pokud se zaměříme na zbylé dvě úrovně nemoci (G2 a G3), uvidíme, že vyšší odhadovanou pravděpodobnost

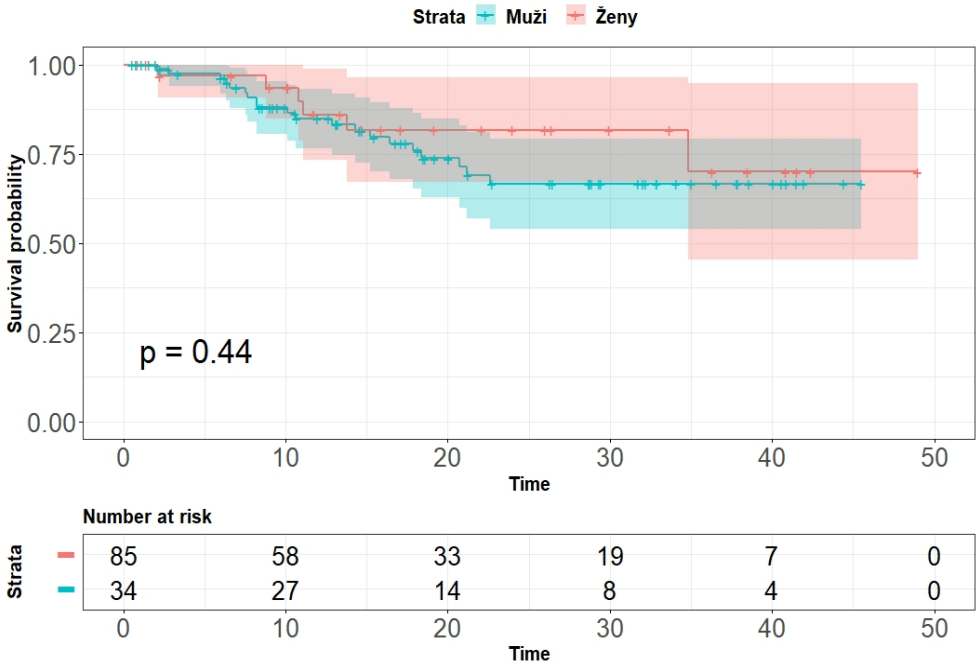
přežití má téměř po celou dobu skupina pacientů s úrovní nemoci G2. Podle p-hodnoty, kterou navíc ovlivňuje i skupina G1, bychom však mohli říct, že tento rozdíl významný není. Ani u jedné ze skupin odhadovaná pravděpodobnost přežití neklesne pod 50 %. Restringovaný průměrný čas přežití je 38 měsíců (±2.73) pro pacienty s úrovní nemoci G2 a 33.1 měsíců (±2.25) pro pacienty s úrovní nemoci G3. Uvažovaná hranice pro tyto výpočty je pak brána jako 45.5 měsíce.

Jak už bylo řečeno, žádný z celkových deseti pacientů ze skupiny G1 nezemřel v důsledku onemocnění GIT. Fakt, že žádný z těchto pacientů nezemřel na sledované onemocnění je velmi dobrá zpráva, nicméně z hlediska další analýzy a výpočtů přináší skupina G1 hned dva potenciální problémy. Prvním z nich je malý počet pozorování a druhým je fakt, že pokud v dané skupině nedošlo ani k jedné sledované události, pak nebude možné odhadnout odpovídající regresní parametr Coxova modelu. Nabízí se proto dvě možná řešení. Buď můžeme pozorování zařazená do skupiny G1 z analýzy vyřadit nebo můžeme sloučit skupiny G1 a G2. Přistoupením k první možnosti by došlo k poměrně významné ztrátě informace, neboť se jedná o 10 pacientů z celkových 119, tj. vyřadili bychom 8.4% pozorování. Z tohoto důvodu se druhá možnost jeví jako výhodnější. Spojíme-li tedy skupiny G1 a G2, budou křivky přežití pro jednotlivé úrovně vypadat následovně:



Obrázek 5.13: Křivky přežití pro úrovně onemocnění G1-G2 a G3.

Nakonec si takto vykreslíme křivky přežití pro ženy a pro muže (obrázek 5.14). Křivky přežití se od sebe dle očekávání nijak výrazně neodlišují. Nutné je ale myslet také na to, že počet žen ve studii bylo více než dvakrát více jak mužů. Restringovaný průměrný čas přežití je pak 38.9 měsíců (±2.98) pro ženy a 35.8 měsíců (±2.13) pro muže. I z těchto údajů tedy cítíme, že pohlaví nebude hrát u této nemoci roli významného faktoru.

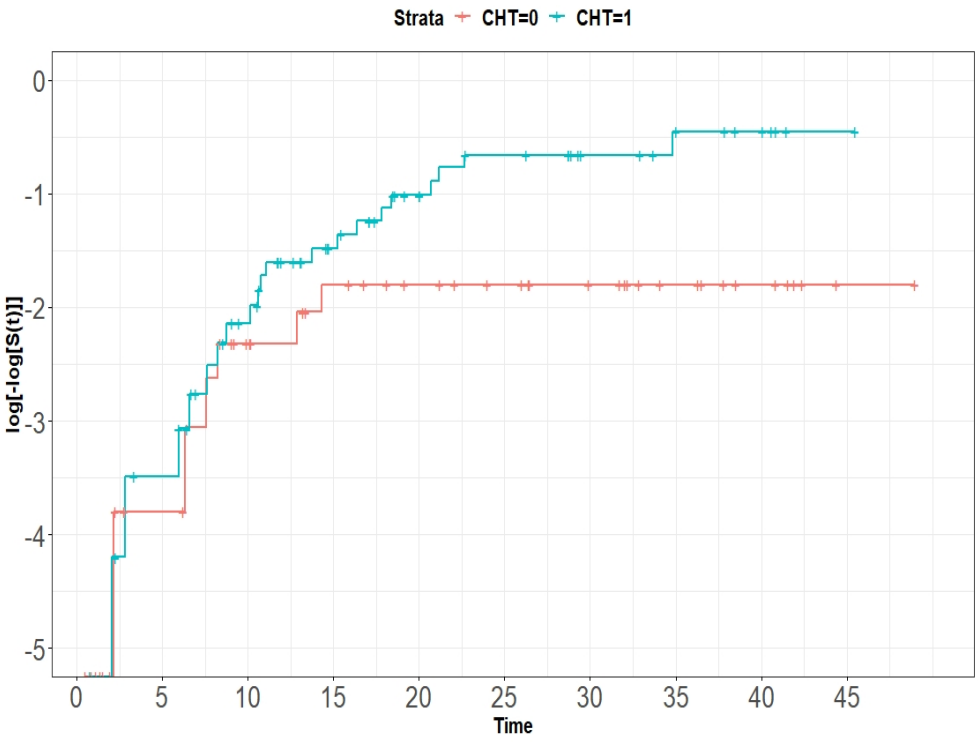


Obrázek 5.14: Křivky přežití pro ženy a muže

5.2.2. Ověření předpokladu proporcionality

Dříve, než přijde na řadu tvorba modelu přežití, s jehož pomocí budeme schopni co nejlepší (nejpřesnější) predikce pro nová pozorování, je třeba otestovat splnění předpokladu proporcionality. V případě jejího splnění lze použít Coxův model proporcionálních rizik, v opačném případě musíme zvolit jiný postup. Základním nástrojem pro ověření proporcionality rizik je vykreslení log-log Kaplanova-Meierova odhadu funkce přežití. V případě, že křivky nebudou paralelní nebo se dokonce překříží, pak je předpoklad porušen. Z grafu na obrázku 5.15 je vidět, že v tomto případě je skutečně proporcionalita rizik porušena, neboť se křivky v několika místech kříží. Graf byl vytvořen pomocí knihovny `ggplot2`, a to analogickým příkazem jakým byly tvořeny křivky přežití výše. Do předpisu funkce

`ggsurvplot` jsme pouze přidali pouze argument `fun = "cloglog"`.



Obrázek 5.15: Log-log křivky přežití

Další možností, jak otestovat předpoklad proporcionality rizik, je využití Schoenfeldových reziduí. Pokud je tento předpoklad splněn, pak jsou Schoenfeldova rezidua nezávislá na čase. Abychom je získali, je nutné nejprve vytvořit Coxův model proporcionálních rizik. Toto provedeme s využitím knihovny `survival` následujícím příkazem:

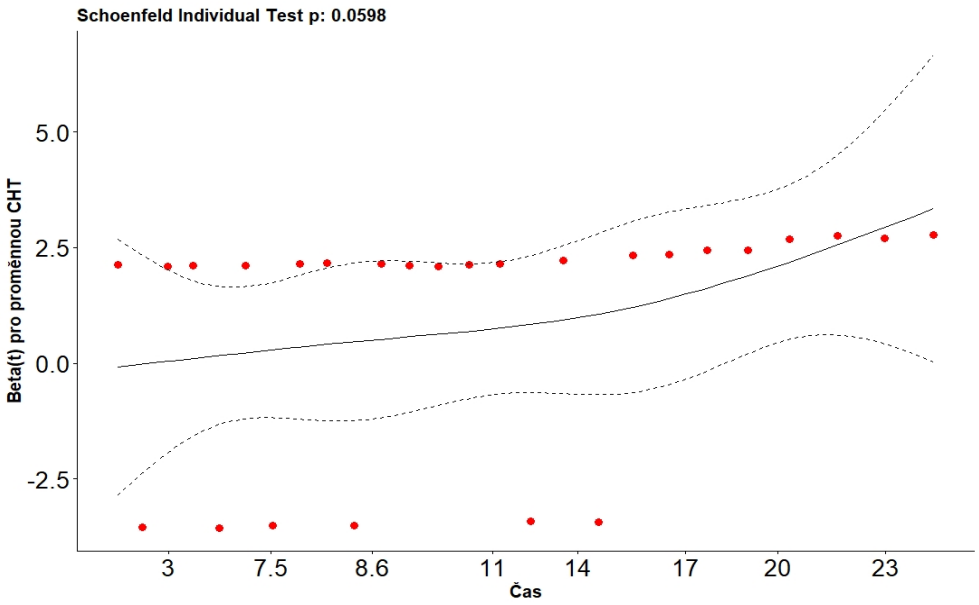
```
> Coxfit1 <- coxph(surv_object ~ CHT, data = GIT)
```

Detaily výstupu se nyní zabývat nebudeme, neboť jsou pro nás relevantní pouze při splnění předpokladu proporcionality rizik. Předpoklad proporcionality rizik lze v softwaru R otestovat (opět s využím knihovny `suvival`) pomocí příkazu:

```
> Cox.zph.fit1 <- cox.zph(Coxfit1,transform="km")
> print(Cox.zph.fit1)
          rho chisq      p
GIT$CHT 0.359  3.54 0.0598
```

Hodnota `rho=0.359` z výstupu je hodnota korelačního koeficientu mezi Schoenfeldovými rezidui a časem. Pokud bychom se striktně řídili p-hodnotou testu, potom bychom jen těsně nezamítli předpoklad proporcionality rizik. Nicméně p-hodnota je v tomto případě velmi nízká a použitím Coxova modelu proporcionálních rizik bychom mohli dostat nepřesné či zavádějící výsledky. Konečně, nemusíme se dívat pouze na tento výstup, ale Schoenfeldova rezidua si můžeme vykreslit (obrázek 5.16). Z teorie víme, že Schoenfeldova rezidua jsou definována pouze pro necenzorovaná pozorování. V našem případě máme celkem 26 necenzorovaných pozorování, čemuž odpovídá 26 červeně vykreslených bodů v grafu. Hodnoty Schoenfeldových reziduí jsou v grafu vykresleny vždy v čase odpovídajícím výskytu události. Jestliže je předpoklad proporcionality splněn, neměli bychom v grafu nalézt žádný trend či systematické skoky a body by měly být náhodně rozmístěny kolem nuly. S posouzením by nám měla pomoci také plná černá křivka v grafu, která představuje odhad příslušného na čase závislého regresního parametru (2.35). Přerušované linky pak představují hranice pro intervaly spolehlivosti dané jako: $\hat{\beta}_j(t) \pm 2\sqrt{\widehat{var}(\hat{\beta}_j(t))}$. Z teorie již víme, že vykreslená křivka v grafu by měla být za předpokladu proporcionality rizik konstantní s hodnotou nula ve všech časech. Případný sklon či zvlnění pak znamená porušení předpokladu proporcionality. Z grafu na obrázku 5.16 vidíme, že většina bodů (20 z 26) leží v kladných hodnotách. Tomu také odpovídá rostoucí trend křivky v grafu. Z výše uvedeného tedy vyplývá, že v tomto případě je podmínka proporcionality skutečně porušena. Graf byl vytvořen s využitím knihovny `survminer` pomocí příkazu:

```
> schoefRes = ggcoxzph(Cox.zph.fit1, point.size = 3,
              xlab = "Čas", ylab="Beta(t) pro proměnnou CHT")
> ggpar( schoefRes,
       font.x        = c(15, "bold"),
       font.y        = c(15, "bold"),
       font.main     = c(15, "bold"),
       font.tickslab = c(20, "plain"))
```



Obrázek 5.16: Schoenfeldova rezidua pro model s proměnnou CHT

Jestliže jsme dospěli k závěru, že předpoklad proporcionality splněn není, pak je nutné najít jiný způsob (než je Coxův model proporcionálních rizik), jak data modelovat.

5.2.3. Stratifikované modely

Základní možností, jak se vypořádat s porušením předpokladu proporcionality rizik je, jak již víme z teoretické části, použití stratifikovaných modelů. Protože byl předpoklad proporcionality porušen u proměnné chemoterapie, budeme modely stratifikovat právě podle této proměnné. Výběr nejvhodnějšího modelu provedeme tak, že budeme postupně přidávat nebo odebírat proměnné, přičemž nás bude zajímat hned několik faktorů, a to: významnost regresního parametru (Waldův test), hodnota indexů determinace a samozřejmě hodnota informačního kritéria (Akaikeho informační kritérium).

Dříve, než začneme s tvorbou a výběrem modelů, připomeňme si, že v části popisné statistiky jsme u proměnné věk odhalili tři odlehlá pozorování. Abychom dostali vypovídající výsledky, tato tři pozorování nebudeme dále uvažovat. Pokud bychom si totiž vytvořili dva modely obsahující pouze proměnnou věk, kde první model uvažuje všechna pozorování včetně těch odlehlých a druhý model odlehlá pozorování neuvažuje, potom bychom zjistili, že první model vykazuje (na rozdíl od druhého modelu) srovnatelné či dokonce méně kva-

litnější výsledky než nulový model (např. Brierovo skóre pro nulový model je BS $= 0,2272$, pro první model pak BS $= 0,2413$). Dále připomeňme, že u proměnné s názvem Grade (úroveň onemocnění) budeme uvažovat kategorie G1 a G2 sloučené dohromady, neboť kategorii G1 odpovídá pouze 10 pozorování a zároveň u žádného z nich nedošlo ke sledované události, což by dále působilo problémy při výpočtech. U proměnné Stage (stádium onemocnění) pak opět uvažujeme pouze základní dělení na skupiny I, II a III.

Nejprve se podívejme na stratifikovaný model, který obsahuje všechny vysvětující proměnné (věk, pohlaví, stádium a úroveň onemocnění). Na příkladě tohoto modelu si zároveň představíme informace, které nám nabízí výstup v softwaru R. Model vytvoříme s využitím knihovny `survival` pomocí funkce `coxph` (funkce pro tvorbu Coxova modelu proporcionálních rizik), přičemž stratifikace dle proměnné chemoterapie docílíme přidáním členu `strata(CHT)` k ostatním proměnným.

```
> model.full = coxph(Surv(OSm, CSSevent) ~ age + sex + Stage_upr +
                     Grade + strata(CHT) , data = GIT_upr)

> summary(model.full)
Call:
coxph(Surv(OSm, CSSevent) ~ age + sex + Stage_upr +
      Grade + strata(CHT) , data = GIT_upr)

  n= 116, number of events= 26

                coef exp(coef) se(coef)     z Pr(>|z|)
age          0.02977   1.03022  0.02819 1.056  0.29089
sex0         0.02756   1.02795  0.49960 0.055  0.95600
Stage_uprII  1.67214   5.32354  0.51124 3.271  0.00107 **
Stage_uprIII 1.74879   5.74762  0.63969 2.734  0.00626 **
GradeG3      0.65356   1.92236  0.48344 1.352  0.17641
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
age              1.030     0.9707    0.9748     1.089
sex0             1.028     0.9728    0.3861     2.737
Stage_uprII      5.324     0.1878    1.9545    14.500
Stage_uprIII     5.748     0.1740    1.6405    20.137
GradeG3          1.922     0.5202    0.7453     4.958
```

```
Concordance= 0.68  (se = 0.065 )
Likelihood ratio test= 16.54  on 5 df,   p=0.005
Wald test            = 15.21  on 5 df,   p=0.01
Score (logrank) test = 16.79  on 5 df,   p=0.005
```

Z výstupu vidíme, že jsme uvažovali cekem 116 pozorování, z nichž u 26 došlo ke sledované události. V modelu jsou zahrnuty celkem tři kategoriální proměnné, a to pohlaví (referenční kategorie ženy), stádium onemocnění (referenční kategorie I) a úroveň onemocnění (referenční kategorie G1-G2). Sloupec s názvem `coef` obsahuje hodnoty odhadnutých regresních parametrů, následuje sloupec `exp(coef)`, který odpovídá poměru rizik (HR). Je-li poměr rizik nižší než jedna, potom hovoříme o tzv. dobrém prognostickém faktoru, neboť vyšší hodnoty proměnné (ve spojitém případě) přinášejí nižší riziko nastání sledované události. V diskrétním případě daná kategorie znamená nižší riziko oproti referenční kategorii. Naopak je-li poměr rizik vyšší než jedna, pak hovoříme o tzv. špatném prognostickém faktoru. Je-li poměr rizik roven jedné, pak daná proměnná nemá na velikost rizika žádný efekt. Ve sloupci označeném `se(coef)` jsou směrodatné chyby pro odhady regresních parametrů. Sloupec označený jako `z` pak odpovídá hodnotě Waldovy statistiky (`coef`/`se(coef)`). V posledním sloupci jsou p-hodnoty příslušné pro Waldův test. Z výstupu vidíme, že statisticky významná je pouze proměnná Stage. V další části výstupu je znovu uveden sloupec s názvem `exp(coef)` udávající poměr rizik a po něm následuje sloupec `exp(-coef)`, který udává převrácený poměr rizik. Máme-li tedy v druhém řádku poměr rizik mužů oproti ženám, pak převrácený poměr rizik vyjadřuje poměr rizik žen oproti mužům. Poslední dva sloupce pak postupně udávají dolní a horní mez pro 95% interval spolehlivosti pro poměr rizik. V dolní části výstupu je uvedena hodnota konkordance (odpovídá Harrellovu C) včetně směrodatné chyby. Poslední tři řádky výstupu se týkají testů hypotézy o nulovosti všech regresních parametrů. Máme zde test poměrem věrohodností (vytvořený model srovnáváme s nulovým modelem), Waldův test a skórový (logrank) test.

Dále můžeme snadno pomocí funkce `cox.zph` ověřit předpoklad proporcionality:

```
> cox.zph(model.full)
            chisq df    p
age         0.176  1 0.67
sex         0.503  1 0.48
Stage_upr   1.463  2 0.48
Grade       0.435  1 0.51
GLOBAL      3.515  5 0.62
```

Vidíme, že tentokrát je předpoklad proporcionality splněn jak pro jednolitvé proměnné zvlášť, tak i celkově (poslední řádek). Můžeme se proto přesunout k další analýze vytvořeného modelu.

Vypočítat si můžeme Akaikeho informační kritérium, které získáme pomocí příkazu `extractAIC`. Pro náš vytvořený model dostáváme hodnotu $AIC = 180.9711$. Tato hodnota sama o sobě vypovídací hodnotu nemá, ale pomůže nám dále v procesu výběru modelu. Stejně tak tomu je i u indexů determinace. Indexy determinace vypočítáme s využitím knihovny `survAUC`. Před výpočtem si datový soubor rozdělíme na trénovací (80 % dat) a testovací sadu (20 % dat). Na základě trénovací sady vytvoříme požadovaný model, ale také model nulový (uvažující pouze základní rizikovou funkci). Tyto modely poté využijeme pro výpočet koeficientů determinace. Ve výstupu je zachováno pořadí koefecientů determinace podle teoretické části této práce (R^2_N, R^2_{XO} a R^2_{OXS}). Vypočítané hodnoty opět využijeme až později při porovnávání modelů.

```
> GIT_upr$id = 1:nrow(GIT_upr)
> set.seed(123)
> ind.split = sample(1:nrow(GIT_upr), round(nrow(GIT_upr)*4/5),
              replace = FALSE)
> GIT.train.upr = GIT_upr[ ind.split,]
> GIT.test.upr  = GIT_upr[-ind.split,]

> m0 = coxph(Surv(OSm, CSSevent) ~ 1, data=GIT.train.upr)
> m1 = coxph(Surv(OSm, CSSevent) ~ age + sex + Stage_upr + Grade +
             strata(CHT), data = GIT.train.upr)
> f0 = rep(0,nrow(GIT.test.upr))
> f1 = predict(m1, newdata=GIT.test.upr)
> Surv.res = Surv(GIT.test.upr$OSm, GIT.test.upr$CSSevent)
```

```
> Nagelk(Surv.res, f1, f0)
[1] 0.3871872

> XO(Surv.res, f1, f0)
[1] 0.2018043

> OXS(Surv.res, f1, f0)
[1] 0.7622221
```

Kombinováním proměnných lze vytvořit další modely a analogickým způsobem pak vypočítat hodnoty statistik pro následné porovnání modelů. Takto podrobně už zde ale další modely rozebírat nebudeme, neboť by to neúměrně zvětšilo rozsah práce. Vypočítané hodnoty uvažovaných kritérií jsou pro všechny modely uvedeny v příloze v tabulce A1. Dle každého z kritérií (AIC, R^2_N, R^2_{XO} a R^2_{OXS}) byly vybrány tři nejlepší modely. Tyto modely jsou uvedeny spolu s hodnotami kritérií v tabulce 5.5. Prozatím nebyly uvažovány interakce mezi proměnnými.

Model	AIC	R^2_N	R^2_{XO}	R^2_{OXS}
STAGE + strata(CHT)	179.196	0.2426	0.1006	0.5857
STAGE + GRADE + strata(CHT)	178.2775	0.3430	0.1525	0.7175
STAGE + VĚK + strata(CHT)	178.9906	0.3225	0.1586	0.6943
STAGE + GRADE + POHLAVÍ + strata(CHT)	180.1242	0.3567	0.1686	0.7322
model.full	180.9711	0.3872	0.2018	0.7622

Tabulka 5.5: Hodnoty AIC, R^2_N, R^2_{XO} a R^2_{OXS} pro vybrané modely bez interakcí.

Dle Akaikeho informačního kritéria je nejlepší model obsahující stádium a úroveň nemoci, tento model má zároveň třetí nejlepší hodnoty kritérií R^2_N a R^2_{OXS}. Nejlepší model dle kritéria R^2_N je model obsahující všechny proměnné, stejně tak je tento model nejlepší i dle kritérií R^2_{XO} a R^2_{OXS}. Druhé nejvyšší hodnoty upravených indexů determinace má model obsahující proměnné pohlaví, stádium a úroveň onemocnění. Třetí nejvyšší hodnotu kritéria R^2_{XO} má model obsahující proměnné věk a stádium onemocnění. Tento model má zároveň čtvrté nejvyšší hodnoty kritérií R^2_N a R^2_{OXS}. Protože poslední dva modely z tabulky 5.5 obsahují velký počet proměnných a zároveň je v obou těchto modelech dle Waldova testu významná pouze proměnná stádium onemocnění, vybrali bychom si dle těchto krité-

rií buď model s proměnnými stádium a úroveň onemocnění nebo model s proměnnou věk a stádium onemocnění.

Nyní přichází na řadu zjistit, jak dobré jsou jednotlivé modely. K tomu využijeme konkordance (Harrellovo C a Gönnenovo a Hellerovo C) a vypočítáme si také Brierovo skóre. Způsob výpočtu si opět ukážeme pouze na modelu, který obrahuje všechny proměnné. Harrellovo C můžeme primárně zjistit z výstupu funkce `coxph`. Protože jsme se již na tento výstup dívali, víme, že hodnota Harrellova C je 0.68 (± 0.065).

Hodnotu Harellova C lze alternativně získat i pomocí funkce `concordance` z knihovny `survival`:

```
> concordance(model.full)
Call:
concordance.coxph(object = model.full)

n= 116
Concordance= 0.6798 se= 0.06546
     concordant discordant tied.x tied.y tied.xy
[1,]        178         37      1      0       0
[2,]        510        286      3      0       0
```

Vidíme, že výsledek je totožný s výsledkem z výstupu funkce `coxph`, zde máme ale navíc rozepsáno, kolik párů je konkordantních, diskordantních a kolik jich je vázaných. Tyto počty jsou navíc rozdělené dle jednotlivých strat.

Gönenovo a Hellerovo C vypočítáme pomocí příkazu `GHCI` z knihovny `survAUC`. K výpočtu využijeme trénovací a testovací sadu, kterou jsme si vytvořili již při výpočtu indexů determinace:

```
> m1 = coxph(Surv(OSm, CSSevent) ~ age + sex + Stage_upr + Grade +
             strata(CHT), data = GIT.train.upr)
> lp.new = predict(m1, newdata = GIT.test.upr)
> GHCI(lp.new)
[1] 0.6902
```

Pro tento model jsme tedy získali hodnotu $C_{GH} = 0.6902$, což je poměrně uspokojivý výsledek.

Brierovo skóre získáme opět s využitím knihovny `survAUC` a trénovací a testovací sady. Protože se jedná o funkci času, definujeme si předem časové okamžiky, ve kterých budeme

toto skóre počítat. Jako vypovídající hodnotu pak budeme brát integrované Brierovo skóre, které je součástí výstupu:

```
> m1 = coxph(Surv(OSm, CSSevent) ~ age + sex + Stage_upr + Grade +
             strata(CHT), data = GIT.train.upr)
> lp = predict(m1)
> lp.new = predict(m1, newdata = GIT.test.upr)
> Surv.rsp = Surv(GIT.train.upr$OSm, GIT.train.upr$CSSevent)
> Surv.rsp.new = Surv(GIT.test.upr$OSm, GIT.test.upr$CSSevent)

> times = 1:40
> predErr(Surv.rsp, Surv.rsp.new, lp, lp.new, times,
          type = "brier", int.type = "weighted")
$error
 [1] 0.000000000 0.000000000 0.001975685 0.001928821 0.001928821
 [6] 0.003334070 0.006110923 0.010488597 0.018339743 0.018202850
[11] 0.041016457 0.042773391 0.031918085 0.028221322 0.026872143
[16] 0.021692241 0.019843697 0.017779707 0.016135367 0.016135367
[21] 0.014532038 0.013084039 0.011719640 0.011719640 0.011719640
[26] 0.011719640 0.011719640 0.011719640 0.011719640 0.011719640
[31] 0.011719640 0.011719640 0.011719640 0.011719640 0.008787658
[36] 0.008787658 0.008787658 0.008787658 0.008787658 0.008787658

$times
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
[21] 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

$ierror
[1] 0.03697274
```

Ve výstupu máme uvedeny hodnoty Brierova skóre v jednotlivých časech, dále uvažované časové body a nakonec také hodnotu integrovaného Brierova skóre ($IBS = 0.0370$). Tato hodnota je výrazně nižší než 0.25, což je hodnota odpovídající přiřazení stejného (50%) rizika všem subjektům. Získali jsme tedy velmi pozitivní výsledek.

Analogickým způsobem lze určit hodnoty konkordance i Brierova skóre pro ostatní modely. Stejně jako u výběru modelu dle AIC a upravených indexů determinace, i zde jsou v tabulce 5.6 uvedeny pouze tři nejlepší modely dle každého kritéria. Hodnoty pro všechny modely jsou uvedeny v sekci přílohy v tabulce A1.

Model	C	C_{GH}	BS
STAGE + VĚK + strata(CHT)	0.6793	0.6843	0.0393
STAGE + GRADE + strata(CHT)	0.6606	0.6016	0.0379
STAGE + POHLAVÍ + VĚK + strata(CHT)	0.6803	0.6859	0.0393
STAGE + POHLAVÍ + GRADE + strata(CHT)	0.6665	0.6465	0.0385
model.full	0.6798	0.6902	0.0370

Tabulka 5.6: Hodnoty Harrellova C, Gönenova a Hellerova C a Brierova skóre pro vybrané modely bez interakcí

Povšimnout si můžeme toho, že kromě třetího modelu (model uvažující proměnné věk, pohlaví a stádium onemocnění) se všechny ostatní modely objevily už v tabulce 5.5. Nejvyšší hodnodu Harrellova C má právě třetí model, který má zároveň druhou nejvyššího hodnotu Gönenova a Hellerova C. Nejvyšší hodnotu Gönenova a Hellerova C má pak model obsahující všechny promněnné. Tento model má také nejnižší hodnotu Brierova skóre. Druhou nejnižší hodnotu Brierova skóre má model uvažující stádium a úroveň onemocnění. Dále tento model ale nijak nevyniká. Třetí nejvyšší hodnoty konkordace pak náleží modelu s proměnnými věk a stádium onemocnění. Nakonec třetí nejnižší hodnotu Brierova skóre má model předposlední. Velmi dobře je na tom tedy model se všemi proměnnými, nicméně jak jsme si již uvedli dříve, v tomto modelu je významná pouze proměnná stádium onemocnění. Protože se jedná o model s poměrně velkým počtem proměnných a jen jedna z nich je významná, nemá smysl tento model dále uvažovat. Stejný problém se týká také modelu třetího a předposledního. Pokud se zaměříme na první dva modely v tabulce, zjistíme, že (zejména kvůli rozdílu v hodnotách C_{GH}) si vede lépe model první. Nicméně celkově jsou hodnoty daných kritérií pro všechny modely poměrně srovnatelně dobré.

Podle kritérií pro porovnání modelů mezi sebou i podle statistik hodnotících predikční schopnosti jednotlivých modelů jsme shodně označili jako nejlepší model uvažující proměnnou věk a stádium onemocnění (ozn. model 1) a model uvažující stádium a úroveň onemocnění (ozn. model 2). V tabulce 5.7 jsou pro názorné porovnání znovu uvedeny tyto dva modely s hodnotami všech uvažovaných statistik. Vidíme, oba uvažované modely jsou velmi dobré a výrazněji se liší pouze podle Gönenova a Hellerova C, kde si lépe vede model první. Protože jsou ale jinak modely srovnatelně dobré, lze je pro závěrečnou analýzu využít oba.

Model	AIC	R^2_N	R^2_{XO}	R^2_{OXS}	C	C_{GH}	BS
Model 1	178.9906	0.3225	0.1586	0.6943	0.6793	0.6843	0.0393
Model 2	178.2775	0.3430	0.1525	0.7175	0.6606	0.6016	0.0379

Tabulka 5.7: Hodnoty AIC, R^2_N, R^2_{XO}, R^2_{OXS}, Harrellova C, Gönenova a Hellerova C a Brierova skóre pro nejlepší modely bez interakcí

Do této chvíle jsme uvažovali pouze modely bez interakcí. Nyní je třeba podívat se, zda by nám přidání interakce do modelu nepřineslo jeho zlepšení. V sekci přílohy jsou v tabulce A2 vypočítané hodnoty Akaikeho informačního kritéria, upravených indexů determinace (R^2_N, R^2_{XO}, R^2_{OXS}), konkordance (Harrellova C i Gönenovo a Hellerovo C) a též Brierova skóre pro modely s interakcemi. Pro názornost jsou tabulce 5.8 uvedeny modely, které byly alespoň podle jednoho kritéria nejlepší.

Model	AIC	R^2_N	R^2_{XO}	R^2_{OXS}	C	C_{GH}	BS
STAGE:VĚK + strata(CHT)	178.4273	0.3240	0.1658	0.6960	0.6783	0.6362	0.0394
STAGE:VĚK + GRADE + strata(CHT)	178.4552	0.3828	0.2068	0.7581	0.6773	0.6545	0.0373
STAGE + VĚK:strata(CHT)	180.8228	0.2923	0.1180	0.6568	0.6793	0.6332	0.0391
VĚK + STAGE:strata(CHT)	179.3735	0.3206	0.3112	0.6920	0.6783	0.6552	0.0435

Tabulka 5.8: Hodnoty AIC, R^2_N, R^2_{XO}, R^2_{OXS}, Harrellova C, Gönenova a Hellerova C a Brierova skóre pro vybrané modely s interakcemi.

Z tabulky vidíme, že stejně jako u modelů bez iterakcí, tak i v modelech s interakcemi jsou dle uvažovaných kritérií nejlepší ty, které obsahují proměnné věk, stádium nebo úroveň onemocnění. Oproti modelům bez interakcí zde však nedošlo k výraznějšímu zlepšení u žádného z kritérií. V prvních dvou modelech není dle Waldova testu žádná proměnná významná. Ve třetím modelu je významná pouze proměnná stádium onemocnění, stejně jako v modelu posledním. Protože nedošlo přidáním interakcí do modelů k výraznému zlepšení a proměnné v modelech nejsou významné, nemá smysl se dále těmito modely zabývat.

Z provedené analýzy vyplývá, že nejlepším modelem pro popis našich dat a následnou predikci je model s proměnnými věk a stádium onemocnění a model s proměnnými stádium a úroveň onemocnění. Nyní se na tyto modely podíváme podrobněji. Postupně se podíváme na odhady parametrů v obou modelech, které si následně interpretujeme. Provedeme také základní diagnostiku obou modelů, abychom se ujistili, že získané výsledky nejsou zavádějící. Začněme modelem uvažujím proměnné věk a stádium onemocnění.

```
> model1 = coxph(Surv(OSm, CSSevent) ~ age + Stage_upr + strata(CHT),
                                        data = GIT_Out)
> summary(model1)
Call:
coxph(formula = Surv(OSm, CSSevent) ~ age + Stage_upr + strata(CHT),
    data = GIT_Out)

  n= 116, number of events= 26
                coef exp(coef) se(coef)     z Pr(>|z|)
age          0.03947   1.04026  0.02701 1.462  0.14386
Stage_uprII  1.65644   5.24064  0.52119 3.178  0.00148 **
Stage_uprIII 1.76392   5.83524  0.64623 2.730  0.00634 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
age              1.040     0.9613    0.9866     1.097
Stage_uprII      5.241     0.1908    1.8869    14.555
Stage_uprIII     5.835     0.1714    1.6443    20.707

Concordance= 0.679  (se = 0.064 )
Likelihood ratio test= 14.52  on 3 df,   p=0.002
Wald test            = 12.84  on 3 df,   p=0.005
Score (logrank) test = 14.29  on 3 df,   p=0.003
```

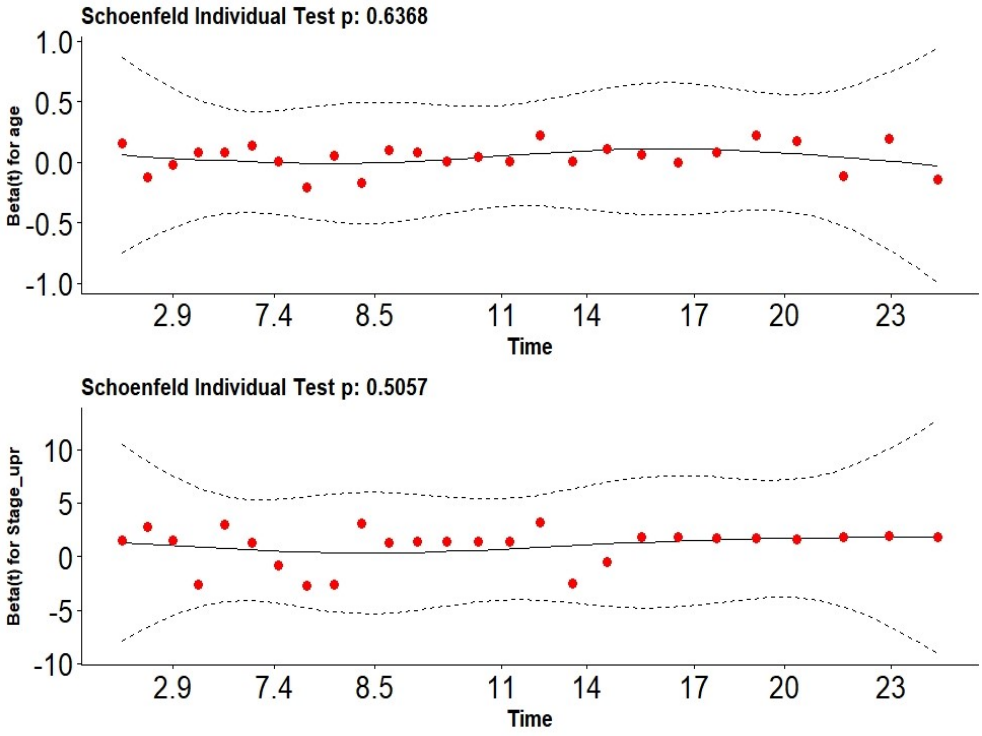
Z výstupu vidíme, že proměnná věk není statisticky významná, a to jak podle p-hodnoty u Waldova testu, tak i podle poměru rizik, který je roven přibližně jedné. Odpovídající 95% interval spolehlivosti pro poměr rizik navíc obsahuje jedničku, což nám jen potvzuje, že věk nemá na riziko výskytu sledované události vliv. Nicméně proměnná Stage významná je. Pacienti ve II. stádiu onemocnění mají asi 5.2-krát vyšší riziko, že zemřou na onemocnění GIT než pacienti ve stádiu I. Vidíme také, že odpovídající 95% interval spolehlivosti je pro tento poměr rizik poněkud široký $\langle 1.89; 14.56\rangle$. Podobná situace je pak u pacientů ve stádiu onemocnění III. Zde je odhadovaný poměr rizik roven hodnotě asi 5.8, nicméně 95% interval spolehlivosti je ještě širší než pro pacienty ve stádiu II ($\langle 1.64; 20.71\rangle$). Jak už bylo uvedeno v tabulce 5.6, hodnota Harrellova C je 0.679 (± 0.064). Na konci výstupu jsou pak údaje pro testy významnosti všech regresních parametrů dohromady. Vidíme, že podle testu poměrem věrohodností, Waldova testu i skórového testu nulovou hypotézu

o nevýznamnosti regresních parametrů zamítáme.

Abychom mohli považovat právě uvedené výsledky za vypovídající (správné), je třeba splnění předpokladu proporcionality. Nejprve si proporcionalitu otestujme. Z výstupu funkce `cox.zph` níže vidíme, že výsledné p-hodnoty jsou velmi vysoké, a to jak pro jednotlivé proměnné, tak i pro celkový test. Z toho vyplývá, že předpoklad proporcionality je v případě tohoto modelu jistě splněn.

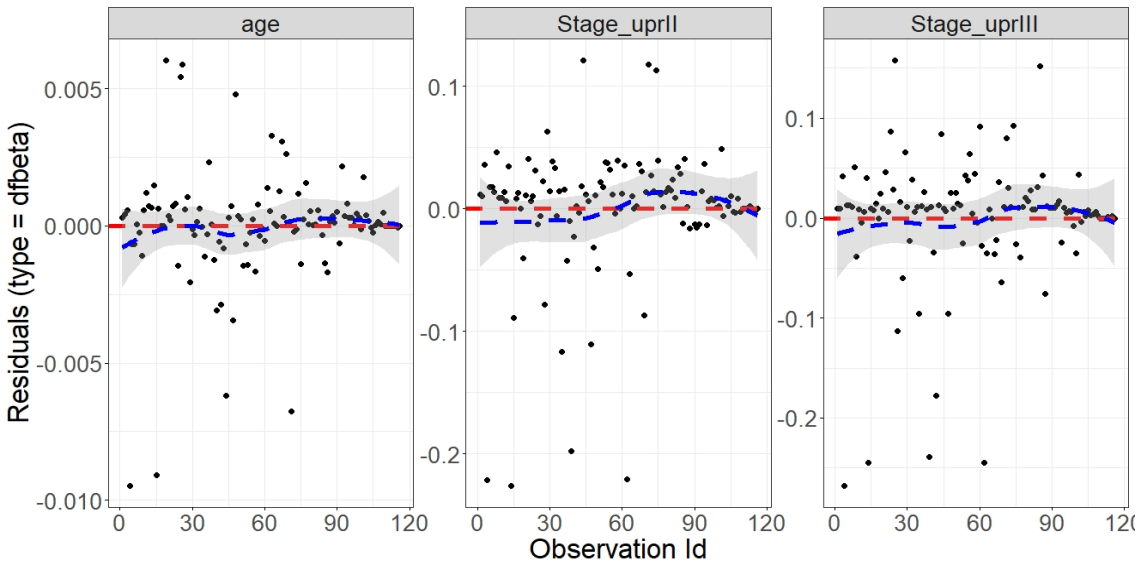
```
> cox.zph(model1)
           chisq df    p
age        0.223  1 0.64
Stage_upr 1.364  2 0.51
GLOBAL     1.678  3 0.64
```

Stejný výsledek vidíme i z obrázku 5.17. Schoenfeldova rizidua jsou v případě obou proměnných symetricky rozmístěna kolem nuly, a to po celou uvažovanou dobu. Stejně tak křivky znázorňující na čase závislé odhady parametrů pro obě proměnné jsou konstantní s hodnotou nula.



Obrázek 5.17: Schoenfeldova rezidua pro model 1

Z hlediska diagnostiky bychom měli ještě zkontrolovat přítomnost vlivných pozorování. Vlivnými pozorováními v tomto případě myslíme ta, jenž svoji přítomností významně ovlivňují hodnoty odhadovaných regresních parametrů. Využijeme k tomu knihovnu `survminer` a příkaz `ggcoxdiagnostics`, přičemž zde specifikujeme `type = "dfbeta"` (obrázek 5.18). Výstupem jsou v tomto případě tři grafy, které odpovídají regresním parametrům modelu. Osy x představují indexy jednotlivých pozorování a na osách y jsou odhadované změny v hodnotách regresních koeficientů za předpokladu, že by dané pozorování bylo při tvorbě modelu vynecháno. V ideálním případě bychom chtěli, aby byly body v grafech rovnoměrně rozmístěny kolem nuly. Z obrázku 5.18 vidíme, že tomu tak ve většině pozorování je, nicméně najdou se i pozorování, jejichž odchylka od nuly je výraznější. Tyto odchylky je dobré porovnat se směrodatnou chybou odhadu regresního parametru, v případě, že odchylka pro dané pozorování není větší než směrodatná chyba odhadu, pak lze předpokládat, že se o vlivné pozorování nejedná. Pro proměnnou věk je odhad regresního parametru 0.03947 ($\pm$0.02701), pro II. stádium onemocnění pak 1.65644 ($\pm$0.52119) a pro III. stádium 1.76392 ($\pm$0.64623). Hodnoty směrodatných chyb tedy překročeny nebyly ani u jednoho pozorování.



Obrázek 5.18: Diagnostický graf pro posouzení přítomnosti vlivných pozorování u prvního modelu

Protože jsme v uvažovaném modelu nenašli žádná vlivná pozorování, jenž by ovlivňovala výsledky a zároveň jsme potrvdili splnění předpokladu proporcionality, můžeme získané výsledky považovat za správné.

Nyní se podívejme na druhý model, který uvažuje proměnné stádium a úroveň onemocnění.

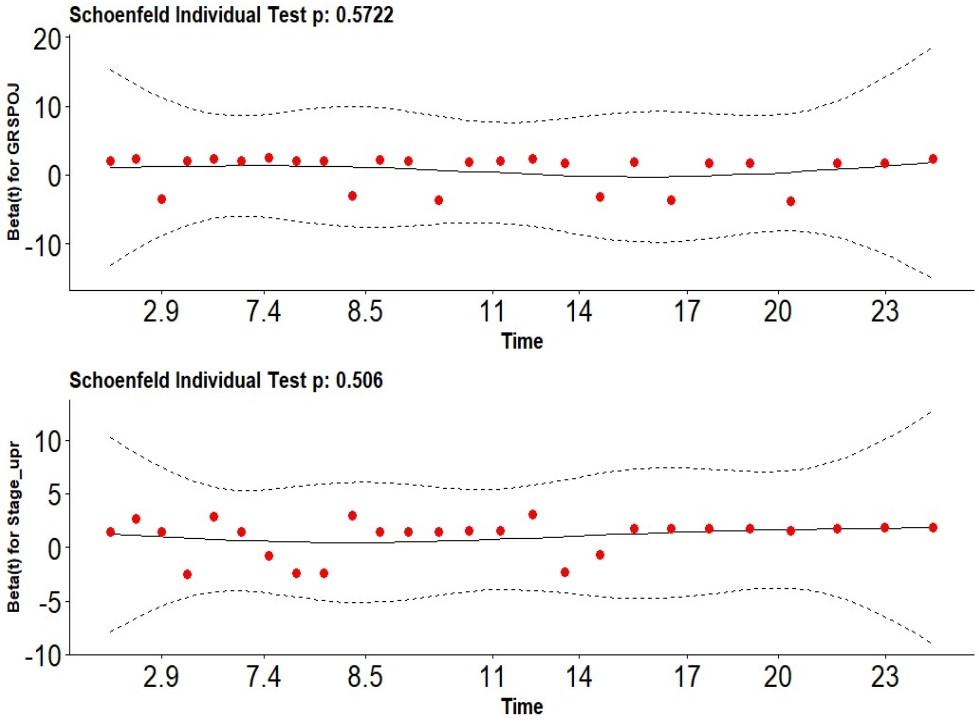
```
> model2 = coxph(Surv(OSm, CSSevent) ~ Stage_upr + GRSPOJ + strata(CHT),
                                        data = GIT_Out)
> summary(model2)
Call:
coxph(formula = Surv(OSm, CSSevent) ~ Stage_upr + GRSPOJ + strata(CHT),
    data = GIT_Out)

  n= 116, number of events= 26
               coef exp(coef) se(coef)     z Pr(>|z|)
Stage_uprII  1.6615    5.2672   0.5137 3.234  0.00122 **
Stage_uprIII 1.7007    5.4780   0.6371 2.670  0.00759 **
GRSPOJG3     0.7548    2.1272   0.4684 1.611  0.10711
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
Stage_uprII      5.267     0.1899    1.9244    14.417
Stage_uprIII     5.478     0.1825    1.5716    19.094
GRSPOJG3         2.127     0.4701    0.8493     5.328

Concordance= 0.661  (se = 0.06 )
Likelihood ratio test= 15.23  on 3 df,   p=0.002
Wald test            = 13.5  on 3 df,   p=0.004
Score (logrank) test = 15.15  on 3 df,   p=0.002
```

Z výstupu vidíme, že významná je opět pouze proměnná stádium onemocnění, přičemž interpretačně dostáváme velmi podobné výsledky jak u modelu předchozího. Pacienti ve stádiu nemoci II mají oproti pacientům ve stádiu I asi 5.3-krát vyšší riziko, že na onemocnění GIT zemřou. Odpovídající 95% interval spolehlivost je pak $\langle 1.9244; 14.417\rangle$. Jedná se tedy opět o poměrně široké rozpětí. Pacienti ve stádiu onemocnění III pak mají asi 5.5-krát vyšší riziko úmrtí na GIT než pacienti ve stádiu I, přičemž 95% interval spolehlivosti je $\langle 1.5716; 19.094\rangle$. Proměnná označující úroveň onemocnění není dle Waldova testu

významná. Odhadovaný poměr rizik skupiny G3 oproti skupině, kde jsou dohromady pacienti G1 a G2, je roven 2.1272. Znamená to, že pacienti ve skupině G3 mají přibližně dvakrát vyšší riziko, že zemřou na onemocnění GIT než mají pacienti ve skupinách G1 a G2. Nicméně odpovídající 95% interval spolehlivosti $\langle 0.08493; 5.328\rangle$ obsahuje jedničku. Z toho vyplývá, že úroveň onemocnění ve skutečnosti vliv na výskyt sledované události mít nemusí. Dle testu poměrem věrohodností, Waldova i skórového testu hypotézu o nevýznamnosti všech proměnných v modelu zamítáme.

Stejně jako u prvního vybraného modelu, i zde je nutné provést základní diagnostiku modelu. Začněme opět ověřením proporcionality rizik. Z výstupu níže vidíme, že dle testu proporcionalitu nezamítáme.

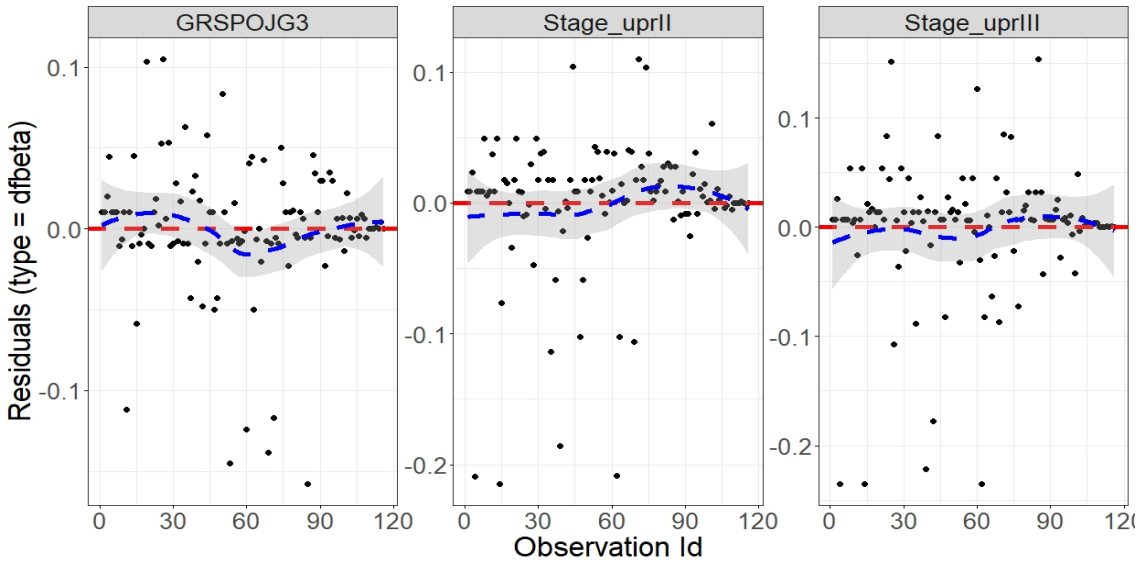
```
> cox.zph(model2)
          chisq df    p
GRSPOJ    0.319  1 0.57
Stage_upr 1.362  2 0.51
GLOBAL    1.600  3 0.66
```



Obrázek 5.19: Schoenfeldova rezidua pro model 2

Na obrázku 5.19 jsou grafy pro vizuální posouzení proporcionality. Vidíme, že škálovaná Schoenfeldova rezidua jsou náhodně rozmístěna kolem nuly a odhady na čase závislých parametrů jsou po celou dobu pro obě proměnné konstantně nulové.

Zbývá ověřit, zda nejsou přítomna vlivná pozorování, která by významně ovlivňovala odhadované parametry v modelu. Odpovídající grafy jsou na obrázku 5.20. Opět vidíme, že některá pozorování mají na odhad parametrů větší vliv než jiná, nicméně v souvislosti s žádným z nich nedochází k větší změně hodnoty parametru než je jeho odpovídající směrodatná ochylka. Pro úroveň onemocnění G3 je odhad parametru 0.7548 (±0.4684), pro stádium II 1.6615 (±0.5137) a pro stádium III pak 1.7007 (±0.6371). Ani v tomto modelu jsme tedy nezaznamenali nic, co by mělo znehodnocovat získané výsledky a můžeme je považovat za správné.



Obrázek 5.20: Diagnostický graf pro posouzení přítomnosti vlivných pozorování u druhého modelu

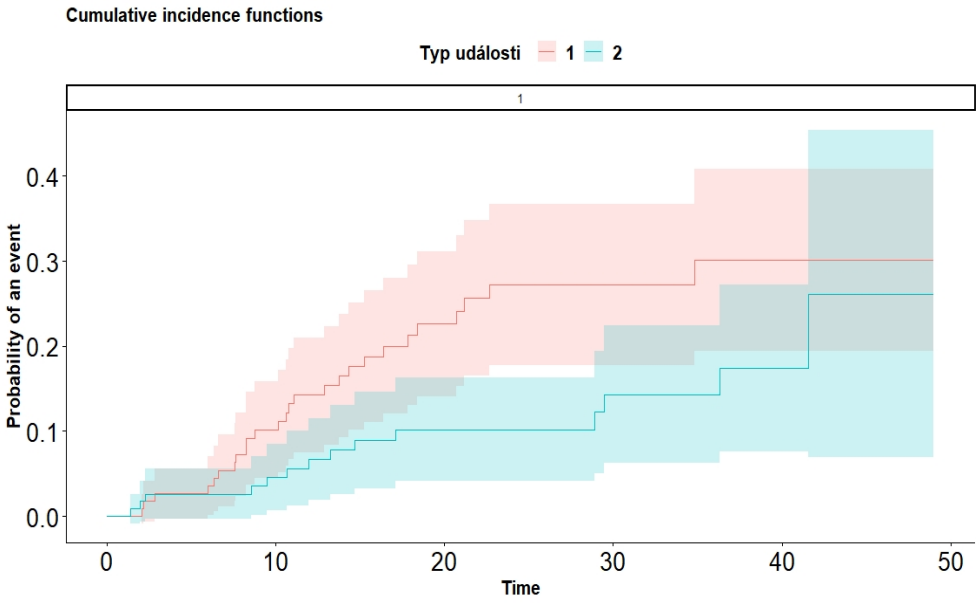
Na tomto místě si shrňme poznatky z provedené analýzy. Na riziko výskytu sledované události má největší vliv stádium onemocnění a fakt, zda pacient podstoupil léčbu chemoterapií. Naopak žádný vliv nemá pohlaví pacienta. Velmi malý vliv pak mají proměnné věk a úroveň onemocnění. Protože proměnná označující, zda pacient prodělal léčbu chemoterapií způsobovala porušení proporcionality v modelech, přistoupili jsme ke stratifikaci

modelů právě dle této proměnné. Podle hodnot Akaikeho informačního kritéria, upravených indexů determinace, konkordance a Brierova skóre jsme nakonec vybrali dva nejlepší modely, a to model uvažující proměnné věk a strádium onemocnění (model 1) a model uvažující proměnné úroveň a stádium onemocnění (model 2). Pro oba tyto modely byla provedena základní diagnostika a interpretace výsledků. Zároveň jsme zjistili, že přidání interakcí do modelu nevede k jeho zlepšení.

5.2.4. Modely konkurenčních rizik

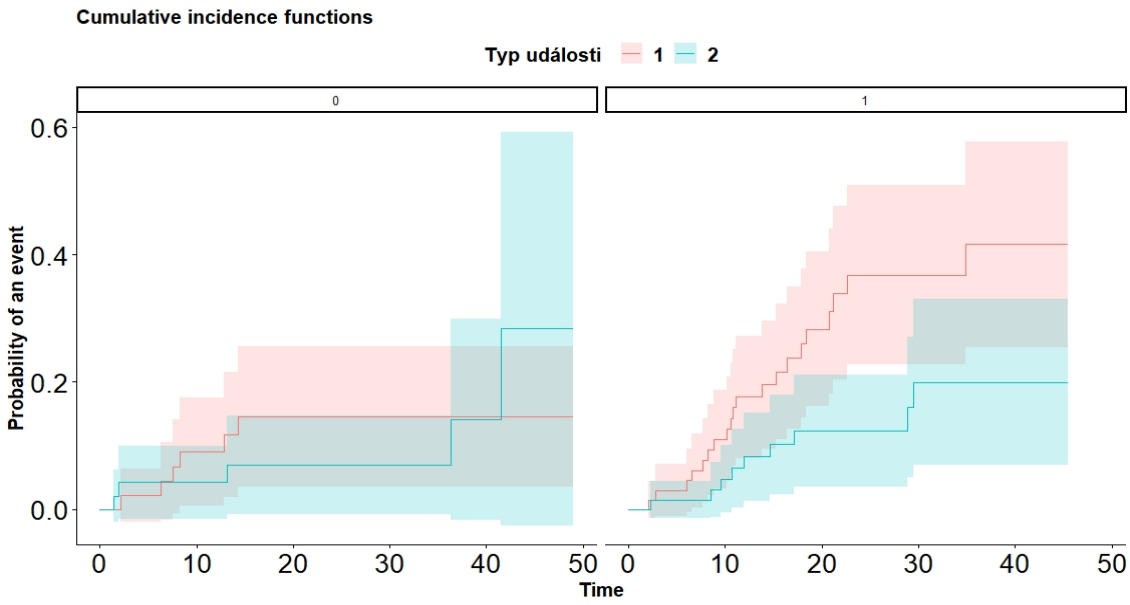
V této části využijeme kromě informace o tom, zda pacient podlehl sledovanému onemocnění (proměnná CSSevent), také informaci o tom, zda pacient zemřel následkem jakékoli příčiny (proměnná OSevent). Mluvíme přitom o tzv. konkurenčním riziku, neboť nastane-li nějaká jiná (konkurenční) událost, není možné, aby následně nastala událost sledovaná. Pro modelování konkurenčních rizik v softwaru R využijeme knihovny `cmprsk`. První co můžeme udělat, je vykreslit si křivky kumulativní incidience pro obě možné události. Získáme tím kumulativní pravděpodobnosti, s jakými dojde ke sledované či konkurenční události v různých časech. Uvažovat budeme nejdříve opět nejjednoduší případ, tedy pouze časy přežití – obrázek 5.21. Typ události 1 přitom označuje smrt následkem námi sledovaného onemocnění a typ události 2 pak označuje smrt následkem jiné příčiny. Z grafu vidíme, že po většinu času je kumulativní pravděpodobnost nastání sledované události vyšší než nastání události konkurenční. Graf byl vytvořen s využitím knihovny `ggplot2` pomocí následujících příkazů:

```
> CInc = cuminc(ftime = GIT$OSm, fstatus = GIT$typ, cencode = 0)
> GGCOMP1=ggcompetingrisks(CInc, conf.int = T,
          legend.title = "Typ události")
> ggpar( GGCOMP1,
       font.title    = c(15, "bold"),
       font.x        = c(15, "bold"),
       font.y        = c(15, "bold"),
       font.legend   = c(15, "bold"),
       font.tickslab = c(20, "plain"))
```



Obrázek 5.21: Křivky kumulativní incidience sledované a konkurenční události

Tyto křivky si můžeme vykreslit také pro jednotlivé úrovně kvalitativních proměnných. Uvažujme nejprve proměnnou CHT. Křivky kumulativní incidience si vykreslíme zvlášť pro pacienty, kteří nepostoupili léčbu chemoterapií a zvlášť pro pacienty, kteří tuto léčbu podstoupili (obrázek 5.22). Levý graf z obrázku odpovídá skupině pacientů, kteří léčbu chemoterapií nepodstoupili. Vidíme že kumulativní pravděpodobnost nastání sledované události roste do 15. měsíce od operace a poté se ustálí přibližně na 15%. Kumulativní pravděpodobnost, že takovýto pacient zemře následkem jiné příčiny je po většinu sledované doby nižší. Zvýší se až na konci této doby, kdy je dokonce vyšší než kumulativní pravděpodobnost nastání sledované události, nicméně zvětší se zde také pás spolehlivosti a tento odhad je tedy velmi nepřesný. Co se týče skupiny pacientů, kteří léčbu chemoterapií podstoupili, je kumulativní pravděpodobnost nastání sledované události opět po celou dobu vyšší než nastání konkurenční události. Oproti předchozí skupině pacientů se tato pravděpodobnost zvyšuje přibližně až do 23. měsíce od operace, kdy dosáhne hodnoty asi 40%. Kumulativní pravděpodobnost nastání konkurenční události je pak nejvýše 20%. Oproti první skupině pacientů se zde také zúžily pásy spolehlivosti kolem obou křivek a tyto odhady jsou tedy přesnější. Graf byl vytvořen analogickým způsobem jako graf předchozí, jen s tím rozdílem, že jsme do příkazu `cuminc` přidali argument `group = GIT$CHT`.



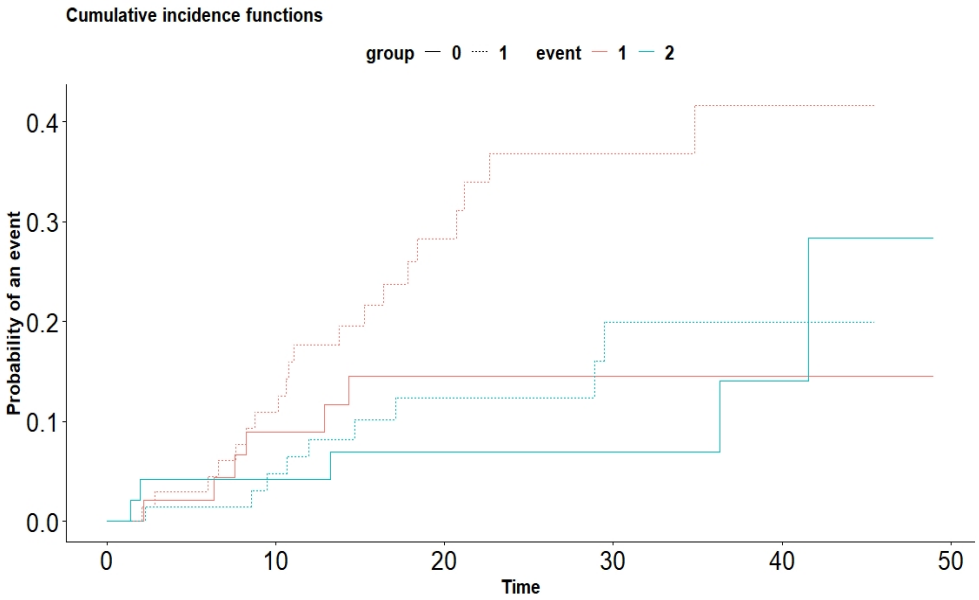
Obrázek 5.22: Křivky kumulativní incidience sledované a konkurenční události – rozlišené dle proměnné CHT

Kromě grafického porovnání lze rozdílnost odhadovaných kumulativních pravděpodobností mezi skupinami také testovat. Využijeme přitom Grayův test. V softwaru R lze výsledky Grayova testu získat následovně (první z příkazů jsme využili již k tvorbě samotného grafu):

```
> CincCHT = cuminc(ftime = GIT$OSm, fstatus = GIT$typ,
            group = GIT$CHT, cencode = 0)
> CincCHT[["Tests"]]

       stat         pv df
1 4.2930160 0.03826922  1
2 0.4540695 0.50040871  1
```

Řádky z výstupu postupně odpovídají oběma typům událostí (1 – sledovaná událost, 2 – konkurenční událost). První sloupec udává hodnotu testové statistiky, druhý sloupec příslušnou p-hodnotu a poslední sloupec počet stupňů volnosti. Výsledky testu nám říkají, že rozdíl mezi kumulativními pravděpodobnostmi nastání sledované události v daných skupinách je statisticky významný, naproti tomu rozdíl mezi kumulativními pravděpodobnostmi nastání konkurenční události statisticky významný není. Pro lepší grafické porovnání odpo-

vídajících křivek kumulativních pravděpodobností existuje možnost vykreslit všechny tyto křivky pro obě skupiny do jednoho grafu – obrázek 5.23.

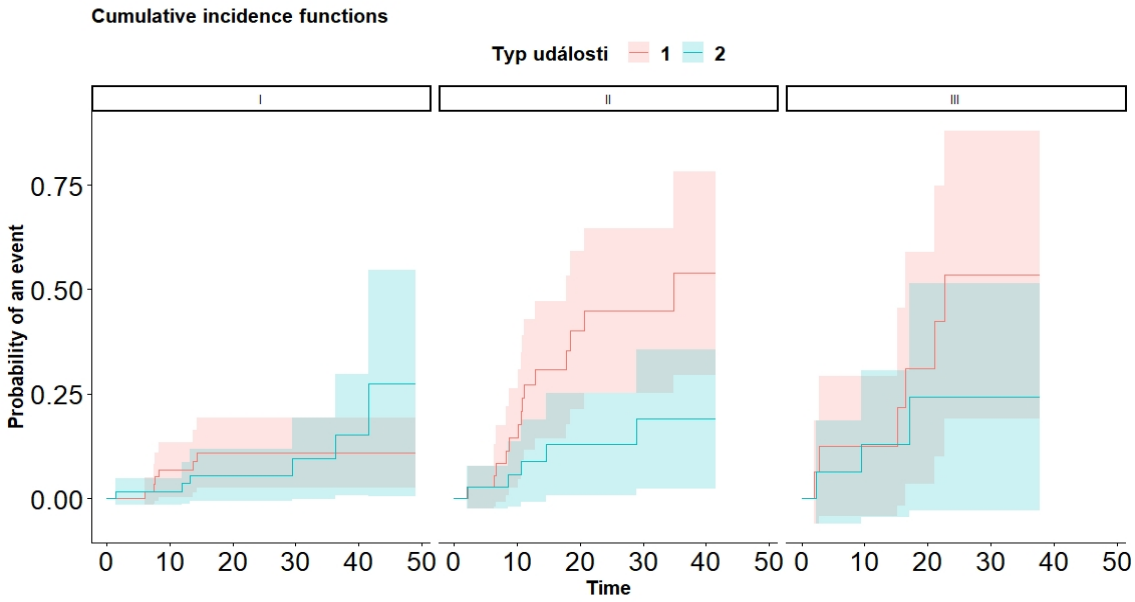


Obrázek 5.23: Křivky kumulativní incidience sledované a konkurenční události – rozlišené dle proměnné CHT

Červené křivky odpovídají kumulativní pravděpodobnosti nastání sledované události, modré pak nastání konkurenční události. Plná čára přitom reprezentuje skupinu pacientů, kteří léčbu chemoterapií nepodstoupili a přerušová čára skupinu pacientů, kteří tuto léčbu podstoupili. Vidíme, že přibližně od 8. měsíce po operaci se od sebe začínají červené křivky vzdalovat a toto vzdalování pokračuje po celou sledovanou dobu, zatímco modré křivky zůstávají blíž u sebe a kolem 42. měsíce od operace se překříží. Graf byl vytvořen opět pomocí příkazu `ggcompetingrisks`, přičemž bylo třeba přidat agrument `multiple_panels=FALSE`.

Analogicky si můžeme pacienty rozdělit do skupin podle proměnné Stage (uvažujeme opět pouze rozdělení I, II a III). Na obrázku 5.24 jsou odpovídající grafy. První graf patří skupině se stádiem nemoci I. Kumulativní pravděpodobnost nastání sledované události je v této skupině v porovnání s ostatními skupinami nejnižší. Prostřední graf patří skupině pacientů se stádiem nemoci II, v této skupině se od sebe obě křivky výrazně odlišují a na rozdíl od první skupiny zde kumulativní pravděpodobnost nastání sledované události překročí hranici 50%. Kumulativní pravděpodobnost nastání konkurenční události

má víceméně podobný průběh jako ve skupině první. Poslední graf pak odpovídá skupině se stupněm onemocnění III. Zde vidíme jistou podobnost se skupinou II, nicméně pásy spolehlivosti jsou zde širší a jedná se tedy o méně přesný odhad.



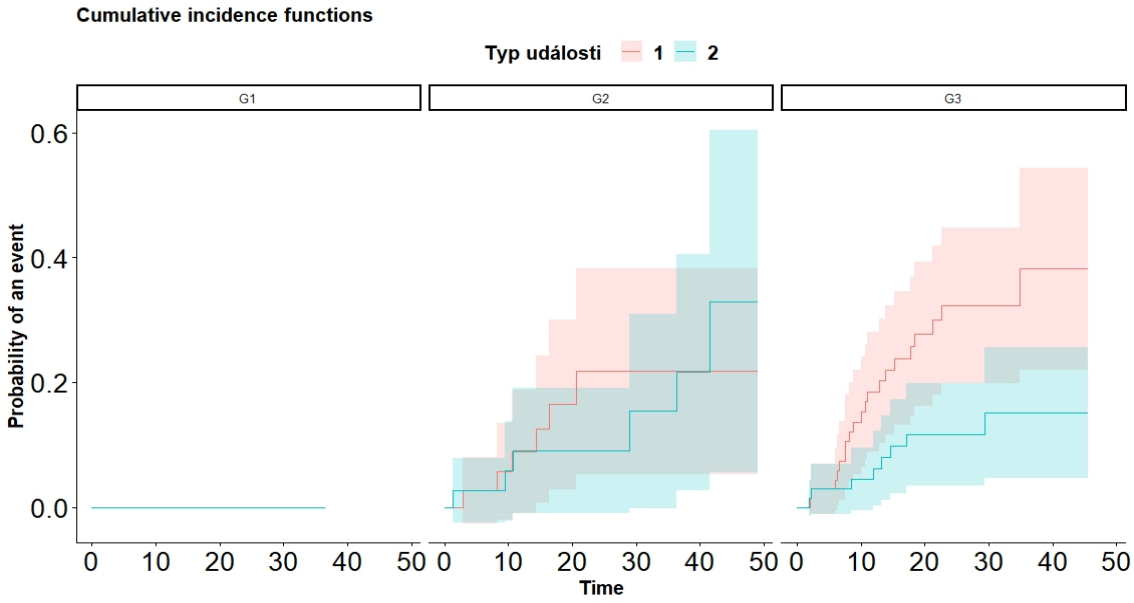
Obrázek 5.24: Křivky kumulativní incidience sledované a konkurenční události – rozlišené dle proměnné Stage

Podívat se také můžeme na výsledky Grayova testu (níže). Vidíme, že rozdíl mezi kumulativními pravděpodobnostmi nastání sledované události v jednotlivých skupinách je signifikantně významný, zatímco rozdíl mezi kumulativními pravděpodobnostmi nastání konkurenční události významný není. Získali jsme tedy analogický výsledek jako v případě rozdělení pacientů podle proměnné CHT.

```
       stat           pv df
1 13.827589 0.0009939792  2
2  1.849808 0.3965694551  2
```

Dále si pacienty rozdělme podle proměnné Grade, tedy podle úrovně nemoci. Na obrázku 5.25 jsou postupně odpovídají tři grafy znázorňující kumulativní pravděpodobnosti nastání sledované, resp. konkurenční události pro úrovně G1, G2 a G3. V první grafu je křivna pouze pro konkurenční událost, neboť jak už jsme dříve poznamenali, v této skupině nedošlo ke sledované události ani u jednoho pacienta. Zajímat nás ale budou následující

dva grafy. Pokud tyto dva grafy budeme porovnávat zjistíme, že vyšší kumulativní pravděpodobnost nástání sledované události je ve skupině pacientů s úrovní nemoci G3 – zde dosahuje hodnoty až 40%, zatímco ve skupině pacientů s úrovní nemoci G2 je to přibližně 20%. Co se týče kumulativní pravděpodobnosti nastání konkurenční události, ta dosahuje vyšších hodnot ve skupině pacientů s úrovní nemoci G2. Zároveň jsou však v této skupině pacientů pásy spolehlivosti kolem obou křivek mnohem širší než ve skupině pacientů s úrovní nemoci G3.

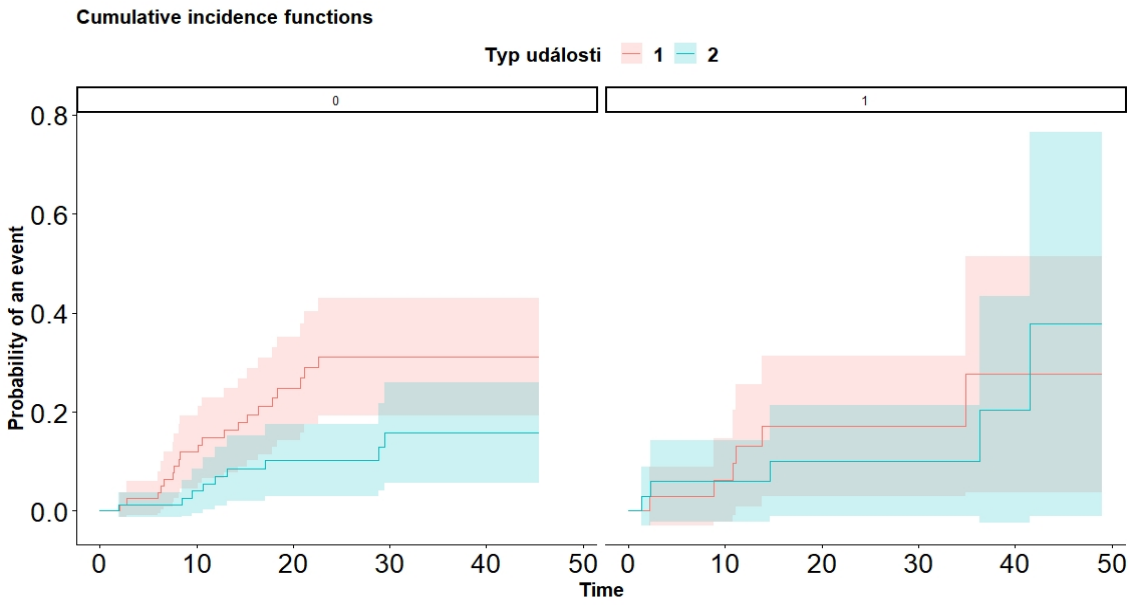


Obrázek 5.25: Křivky kumulativní incidience sledované a konkurenční události – rozlišené dle proměnné Grade

Již podle grafů lze předpokládat, že rozdílnost kumulativních pravděpodobností nastání obou jevů nebude statisticky významná. Toto také potvrzuje výstup Grayova testu:

```
      stat        pv df
1 3.941146 0.1393770  2
2 1.047196 0.5923852  2
```

Nakonec si můžeme pacienty rozdělit také podle pohlaví – obrázek 5.26. Levý graf přitom odpovídá mužům a pravý graf ženám. V případě mužů se od sebe křivky vzdalují více, než v případě žen. Tento fakt může být zapříčiněn pouze malým počtem pozorování u žen (34 ze 119), který se také odráží v šířce pásů. spolehlivosti.



Obrázek 5.26: Křivky kumulativní incidience sledované a konkurenční události – rozlišené dle pohlaví

Stejně jako tomu bylo u proměnné Grade, ani v tomto případě není rozdílnost kumulativních pravděpodobností signifikantní.

```
       stat        pv df
1 0.6332626 0.4261611  1
2 0.2389293 0.6249805  1
```

Kromě vizuálního posouzení vlivu jednotlivých proměnných na výskyt daných událostí můžeme pro konkurenční rizika také sestavit modely. Stejně jako v případě stratifikovaných modelů, i nyní budeme pracovat s datovým souborem, ze kterého budou vyloučena odlehlá pozorování, jenž byla detekována u proměnné věk. U proměnné stádium onemocnění budeme opět pracovat pouze s jednodušším dělením na skupiny I, II a III a u proměnné úroveň onemocnění pak opět spojíme skupiny G1 a G2. Pro sestavené modelu konkurenčních rizik existuje v softwaru R mnoho různých funkcí. My si zde představíme funkci `CSC` z knihovny `riskRegression`. Tato funkce nám vytvoří Coxův model proporcionálních rizik pro každý typ rizika. Protože my už víme, že bychom díky proměnné chemoterapie dostali modely, které nesplňují podmínku proporcionality, budeme rovnou uvažovat odpovídající stratifikované modely. Tvorbu modelu si pro ukažme na případu, kdy uvažujeme všechny proměnné.

```
> model.comp.full = CSC(Hist(OSm, typ) ~ age + sex + Stage_upr + GRSPOJ

> print(model.comp.full)

CSC(formula = Hist(OSm, typ) ~ age + sex + Stage_upr + GRSPOJ +
    strata(CHT), data = GIT_Out)

Right-censored response of a competing.risks model
No.Observations: 116
Pattern:
Cause     event right.censored
  1          26              0
  2          13              0
  unknown     0             77

----------> Cause:  1
Call:
survival::coxph(formula = survival::Surv(time, status) ~ age +
    sex + Stage_upr + GRSPOJ + strata(CHT), x = TRUE, y = TRUE)

  n= 116, number of events= 26
              coef exp(coef) se(coef)     z Pr(>|z|)
age         0.02977   1.03022  0.02819 1.056  0.29089
sex0        0.02756   1.02795  0.49960 0.055  0.95600
Stage_upr2 1.67214   5.32354  0.51124 3.271  0.00107 **
Stage_upr3 1.74879   5.74762  0.63969 2.734  0.00626 **
GRSPOJG3   0.65356   1.92236  0.48344 1.352  0.17641
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
age            1.030     0.9707    0.9748     1.089
sex0           1.028     0.9728    0.3861     2.737
Stage_upr2     5.324     0.1878    1.9545    14.500
Stage_upr3     5.748     0.1740    1.6405    20.137
GRSPOJG3       1.922     0.5202    0.7453     4.958

Concordance= 0.68  (se = 0.065 )
Likelihood ratio test= 16.54  on 5 df,   p=0.005
Wald test            = 15.21  on 5 df,   p=0.01
Score (logrank) test = 16.79  on 5 df,   p=0.005
```

```
----------> Cause:  2
Call:
survival::coxph(formula = survival::Surv(time, status) ~ age +
    sex + Stage_upr + GRSPOJ + strata(CHT), x = TRUE, y = TRUE)

  n= 116, number of events= 13
               coef exp(coef) se(coef)      z Pr(>|z|)
age          0.04411   1.04510  0.04183  1.055    0.292
sex0        -0.46967   0.62521  0.60453 -0.777    0.437
Stage_upr2   0.82374   2.27900  0.70596  1.167    0.243
Stage_upr3   0.97068   2.63975  0.96264  1.008    0.313
GRSPOJG3    -0.12357   0.88376  0.62847 -0.197    0.844

            exp(coef) exp(-coef) lower .95 upper .95
age            1.0451     0.9568    0.9628     1.134
sex0           0.6252     1.5995    0.1912     2.045
Stage_upr2     2.2790     0.4388    0.5713     9.092
Stage_upr3     2.6398     0.3788    0.4001    17.417
GRSPOJG3       0.8838     1.1315    0.2579     3.029

Concordance= 0.622  (se = 0.095 )
Likelihood ratio test= 2.98  on 5 df,   p=0.7
Wald test            = 2.79  on 5 df,   p=0.7
Score (logrank) test = 2.89  on 5 df,   p=0.7
```

V první části výstupu je tabulka s počty událostí jednotlivých typů, které nastaly a je zde uveden také celkový počet pozorování, o nichž nemáme žádnou informaci. Následuje část výstupu označená jako `Cause:1`, která odpovídá stratifikovanému Coxově modelu pro první typ události. V našem případě je prvním typem události smrt následkem rakovinného onemocnění GIT. Pokud srovnáme tuto část výstupu s odpovídajícím výstupem funkce `coxph`, zjistíme, že jsme dostali totožné výsledky. Nově je zde ale část označená jako `Cause:2`, která modeluje událost druhého typu, což je v našem případě smrt následkem jiné příčiny. Z této části výsptupu vidíme, že žádná z proměnných nemá na úmrtí z jiné příčiny než je onemocnění GIT významný vliv. Toto je ostatně výsledek, který jsme očekávali.

V praktické části věnované stratifikovaným modelům jsme vybrali jako nejlepší model uvažující proměnnou věk a stádium onemocnění a model uvažující úroveň a stádium onemocnění, proto se na ně podívejme i v této části práce. Nejprve uvažujme model s pro-

měnnými věk a stádium onemocnění.

```
> model.comp.1 = CSC(Hist(OSm, typ) ~ age + Stage_upr
                          + strata(CHT), data = GIT_Out)
> print(model.comp.1)
CSC(formula = Hist(OSm, typ) ~ age + Stage_upr + strata(CHT),
    data = GIT_Out)

Right-censored response of a competing.risks model
No.Observations: 116
Pattern:
Cause     event right.censored
  1          26              0
  2          13              0
  unknown     0             77

----------> Cause:  1

Call:
survival::coxph(formula = survival::Surv(time, status) ~ age +
    Stage_upr + strata(CHT), x = TRUE, y = TRUE)

  n= 116, number of events= 26

              coef exp(coef) se(coef)     z Pr(>|z|)
age         0.03947   1.04026  0.02701 1.462  0.14386
Stage_upr2 1.65644   5.24064  0.52119 3.178  0.00148 **
Stage_upr3 1.76392   5.83524  0.64623 2.730  0.00634 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
age            1.040     0.9613    0.9866     1.097
Stage_upr2     5.241     0.1908    1.8869    14.555
Stage_upr3     5.835     0.1714    1.6443    20.707

Concordance= 0.679  (se = 0.064 )
Likelihood ratio test= 14.52  on 3 df,   p=0.002
Wald test            = 12.84  on 3 df,   p=0.005
Score (logrank) test = 14.29  on 3 df,   p=0.003

----------> Cause:  2
```

```
Call:
survival::coxph(formula = survival::Surv(time, status) ~ age +
    Stage_upr + strata(CHT), x = TRUE, y = TRUE)

  n= 116, number of events= 13

              coef exp(coef) se(coef)     z Pr(>|z|)
age         0.03791   1.03864  0.03971 0.955    0.340
Stage_upr2 0.80985   2.24756  0.68867 1.176    0.240
Stage_upr3 0.89932   2.45793  0.94143 0.955    0.339

           exp(coef) exp(-coef) lower .95 upper .95
age            1.039     0.9628    0.9609     1.123
Stage_upr2     2.248     0.4449    0.5828     8.668
Stage_upr3     2.458     0.4068    0.3883    15.557

Concordance= 0.642  (se = 0.093 )
Likelihood ratio test= 2.36  on 3 df,   p=0.5
Wald test            = 2.33  on 3 df,   p=0.5
Score (logrank) test = 2.39  on 3 df,   p=0.5
```

Stejně jako u modelu se všemi proměnnými je i v tomto případě první část výstupu totožná s výstupem odpovídajícího stratifikovaného Coxova modelu. V druhé části opět vidíme, že ani jedna z proměnných nemá na nastání smrti následkem jiné příčiny významný vliv. Pro úplnost se ještě podívejme na model uvažující úroveň a stádium onemocnění. Z výstupu níže vidíme, že i v případě tohoto modelu je situace zcela analogická.

```
> model.comp.2 = CSC(Hist(OSm, typ) ~ GRSPOJ + Stage_upr
                                + strata(CHT), data = GIT_Out)
> print(model.comp.2)
CSC(formula = Hist(OSm, typ) ~ GRSPOJ + Stage_upr + strata(CHT),
    data = GIT_Out)

Right-censored response of a competing.risks model
No.Observations: 116
Pattern:
Cause     event right.censored
  1          26              0
  2          13              0
```

```
  unknown     0             77

----------> Cause:  1
Call:
survival::coxph(formula = survival::Surv(time, status) ~ GRSPOJ +
    Stage_upr + strata(CHT), x = TRUE, y = TRUE)

  n= 116, number of events= 26

             coef exp(coef) se(coef)     z Pr(>|z|)
GRSPOJG3   0.7548    2.1272   0.4684 1.611  0.10711
Stage_upr2 1.6615    5.2672   0.5137 3.234  0.00122 **
Stage_upr3 1.7007    5.4780   0.6371 2.670  0.00759 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
GRSPOJG3       2.127     0.4701    0.8493     5.328
Stage_upr2     5.267     0.1899    1.9244    14.417
Stage_upr3     5.478     0.1825    1.5716    19.094

Concordance= 0.661  (se = 0.06 )
Likelihood ratio test= 15.23  on 3 df,   p=0.002
Wald test            = 13.5  on 3 df,   p=0.004
Score (logrank) test = 15.15  on 3 df,   p=0.002

----------> Cause:  2
Call:
survival::coxph(formula = survival::Surv(time, status) ~ GRSPOJ +
    Stage_upr + strata(CHT), x = TRUE, y = TRUE)

  n= 116, number of events= 13

               coef exp(coef) se(coef)     z Pr(>|z|)
GRSPOJG3   0.004135  1.004143 0.607818 0.007    0.995
Stage_upr2 0.775927  2.172606 0.693905 1.118    0.263
Stage_upr3 0.782057  2.185965 0.940672 0.831    0.406

           exp(coef) exp(-coef) lower .95 upper .95
GRSPOJG3       1.004     0.9959    0.3051     3.305
Stage_upr2     2.173     0.4603    0.5576     8.465
Stage_upr3     2.186     0.4575    0.3459    13.815
```

```
Concordance= 0.634  (se = 0.086 )
Likelihood ratio test= 1.41  on 3 df,   p=0.7
Wald test            = 1.37  on 3 df,   p=0.7
Score (logrank) test = 1.4  on 3 df,   p=0.7
```

Výsledky, které jsme získali z vytvořených modelů pro konkurenční rizika, nejsou nijak překvapivé, neboť druhým typem události je smrt následkem jiné než sledované příčiny. Tato jiná příčina není blíže specifikovaná a pro každé pozorování může být zcela odlišná. Pokud bychom konkurenčním rizikem rozumněli úplně vyléčení (za předpokladu, že bychom neuvažovali možnost navrácení onemocnění), pak by pravděpodobně některé z proměnných významné byly, neboť by se konkurenční událost vztahovala přímo k danému onemocnění. V našem případě tomu tak ale není. Ostatně tento fakt byl znatelný již z křivek kumulativní incidience, kdy jsme mezi různými úrovněmi kategoriálních proměnných neidentifikovali signifikantní rozdíly v pravděpodobnosti výskytu události druhého typu.

Závěr

Tato práce měla dva hlavní cíle. Prvním cílem bylo seznámení se s metodami analýzy přežití v případě, kdy není splněn předpoklad proporcionality. Druhým cílem pak byla praktická analýza dat týkajících se přežívání u rakovinného onemocnění gastrointestinálního traktu.

V první kapitole jsme si představili základní pojmy a charakteristiky užívané v analýze přežití. Těmito základními charakteristikami v analýze přežití přitom rozumíme funkci přežití, rizikovou funkci a hustotu, resp. pravděpodobnostní funkci. Naučili jsme se, jak tyto charakteristiky odhadnout z dat, a to i v případě přítomnosti cenzorovaných pozorování. Na závěr první kapitoly jsme se seznámili s postupy pro ověření splnění předpokladu proporcionality rizik.

Druhá kapitola byla věnována tvorbě Coxova modelu proporcionálních rizik a odhadům parametrů. Seznámili jsme se také s metodami pro ověřování předpokladu proporcionality.

Ve třetí kapitole jsme rozšířili znalosti o Coxově modelu proporcionálních rizik na případ stratifikovaného Coxova modelu, který lze využít v případě porušení předpokladu proporcionality. Závěrem této kapitoly jsme se navíc seznámili s modely konkurenčních rizik.

Čtvrtá kapitola byla věnována testování hypotéz v modelech, metodám pro výběr nejlepšího modelu a metodám pro hodnocení modelů. Naučili jsme se, jak otestovat významnost proměnných v modelech pomocí Waldova testu, skórového testu nebo testu poměrem věrohodností. Pomocí log-rank testu jsme se pak naučili testovat hypotézu o shodě odhadovaných pravděpodobností přežití pro dvě nebo více skupin pozorování. Za účelem výběru nejlepšího modelu jsme si představili informační kritéria a upravené indexy determinace. Dále jsme se seznámili s konkordancí a Brierovým skóre, které slouží k hodnocení modelů.

V poslední kapitole jsme se pak věnovali praktické analýze dat. K dispozici jsme měli reálná data týkající se rakovinného onemocnění gastrointestinálního traktu. Nejprve jsme se s daty seznámili a provedli základní popisnou statistiku. Dále jsme se věnovali Kaplanovým-Meierovým křivkám přežití, z nichž jsme získali prvnotní představu o tom, které proměnné ovlivňují pravděpodobnost přežití. Zjistili jsme, že nejsilněji je pravděpodobnost přežití ovlivněna stádiem onemocnění a také tím, zda pacient podstoupil či nepodstoupil léčbu chemoterapií. U proměnné chemoterapie bylo poněkud překvapivým zjištěním, že vyšší pravděpodobnost přežití odpovídala pacientům, kteří tuto léčbu nedpodstoupili. Následně jsme ale zjistili, že podstoupení této léčby silně souvisí právě se stádiem onemocnění. V nejvážnějším stádiu onemocnění podstoupili chemoterapii všichni pacienti. Naopak většina pacientů v nejlehčím stádiu tuto léčbu nepostoupila. Dále jsme si ukázali, jak ověřit předpoklad proporcionality. Z provedených testů jsme zjistili porušení proporcionality u proměnné chemoterapie. Z tohoto důvodu jsme se dále věnovali tvorbě stratifikovaných modelů přežití, přičemž jsme stratifikaci prováděli právě přes tuto proměnnou. Postupně jsme se dívali jak na kritéria pro výběr modelů, tak na kritéria pro hodnocení modelů. Ze všech uvažovaných proměnných (pohlaví, věk, podstoupení chemoterapie, stádium onemocnění a úroveň onemocnění) se ukázala být významná pouze proměnná označující stádium onemocnění. Jako nejlepší modely jsme pak vybrali model s proměnnými věk a stádium onemocnění a model s proměnnými úroveň a stádium onemocnění. U obou těchto modelů však byla dle Waldova testu významná pouze proměnná stádium onemocnění. Stádium onemocnění je přitom kategoriální proměnná, kde jako referenční kategorie byla uvažována skupina I. Skupiny II a III pak měly v průměru asi pětkrát vyšší riziko výskytu sledované události kategorii referenční. Mezi skupinami II a III nebyl detekován výrazný rozdíl. Na závěr praktické části jsme se věnovali modelům konkurenčních rizik. Nejprve jsme se dívali na grafy kumulativní incidience, která vyjadřuje kumulativní pravděpodobnost nastání události daného typu. V našem případě jsme uvažovali dva typy událostí, a to úmrtí důsledkem onemocnění GIT a úmrtí z jiné příčiny. Dle očekávání neměla žádná z uvažovaných proměnných vliv na výskyt druhého typu události (konkurenční události). Toto jsme si nakonec ověřili i sestavením modelů konkurenčních rizik.

Literatura

[1] Bertrand, F.; Bastien P.; Maumy-Bertrand, M.: *Cross validating extensions of kernel, sparse or regular partial least squares regression models to censored data.* 2018

[2] Collet, D.: *Modelling Survival Data in Medical Research.* 2. vydání, Chapman & Hall/CRC, London. 2003.

[3] Dörre, A.; Emura, T.: *Analysis of Doubly Truncated Data, An Introduction.* Springer, 2019.

[4] Gerds, T. A.; Schumacher M.: *Consistent estimation of the expected Brier score in general survival models with right-censored event times.* Biometrical Journal 48, Issue 6. 2006.

[5] Graf, E.; Schmoor, C.; Sauerbrei, W.; Schumacher, M.: *Assessment and comparison of prognostic classification schemes for survival data.* Stat Med. 1999

[6] Grambsch, P.; Patricia, M.; Therneau, T.: *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.* Biometrika Vol. 81, No. 3. 1994.

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

[8] Klein, J.P.: *Small Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators* Scandinavian Journal of Statistics Vol. 18, No. 4. 1991.

[9] Klein, J.P.; Moeschberger, M.L.: *Survival Analysis: Techniques for Censored and Truncated Data.* 2. vydání, Springer New York. 2006.

[10] Kronek, L. P.; Reddy, A.: *Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data.* Bioinformatics, Volume 24, Issue 16. 2008.

[11] Lagakos, S.W.; Barraj, L.M.; De Gruttola, V.: *Nonparametric analysis of truncated survival data, with application to AIDS.* 3. vydání, Biometrika, 1988.

[12] Lee, E.T.; Wang, W.J.: *Statistical Methods for Survival Data Analysis.* 3. ilustrované vydání, John Wiley & Sons. 2003.

[13] Moore, D.F.: *Applied Survival Analysis Using R.* Springer New York. 2016

[14] National Cancer Institute. *PDQ Gastric Cancer Treatment.* [online]. [cit. 2020-01-20]. Dostupné z: https://www.cancer.gov/types/stomach/patient/stomach-treatment-pdq.

[15] Royston, P.: *Explained variation for survival models.* The Stata Journal 6, Number 1, pp. 83-96. 2006

[16] Schmid, M.; Hielscher, T.; Augustin, T.; Gefeller, O.: *A robust alter - native to the Schemper-Henderson estimator of prediction error.* Biometrics 67. 2011.

[17] Sharaf, T.; Tsokos, C.P.: *Predicting Survival Time of Localized Melanoma Patients Using Discrete Survival Time Method.* 2014.

[18] Tableman, M.; Kim, J.S.: *Survival Analysis Using S: Analysis of Time-to-Event Data* CRC Press. 2003.

[19] Stata. *Postestimation tools for stcox.* [online]. [cit. 2020-03-16]. Dostupné z: https://www.stata.com/manuals/ststcoxpostestimation.pdf

[20] Therneau, T. M.; Atkinson E.: *Concordance.* R package version 3.1-11 [online]. [cit. 2020-03-17]. Dostupné z: https://cran.r-project.org/web/packages/survival/

[21] Therneau, T. M.; Grambsch P. M.: *A Package for Survival Analysis in R.* R package version 3.1-8 [online]. [cit. 2020-03-15]. Dostupné z: https://CRAN.R-project.org/package=survival

Přílohy

Seznam příloh

Příloha I. - Tabulka pro výběr a hodnocení modelů bez interakcí

Příloha II. - Tabulka pro výběr a hodnocení modelů s interakcemi

Tabulka pro výběr a hodnocení modelů bez interakcí

Model	AIC	R^2_N	R^2_{XO}	R^2_{OXS}	C_H	C_{GH}	BS
VĚK + strata(CHT)	188.1279	0.0895	0.0055	0.2725	0.5808	0.5474	0.0461
POHLAVÍ + strata(CHT)	189.4728	0.0017	0.0000	0.0061	0.5158	0.3553	0.0482
STAGE + strata(CHT)	179.196	0.2426	0.1006	0.5857	0.6140	0.5393	0.0411
GRADE + strata(CHT)	187.6109	0.0926	0.0195	0.2806	0.5651	0.4590	0.0463
VĚK + POHLAVÍ + strata(CHT)	190.1278	0.0859	0.0055	0.2630	0.5837	0.5618	0.0462
VĚK + STAGE + strata(CHT)	178.9906	0.3225	0.1586	0.6943	0.6793	0.6843	0.0393
VĚK + GRADE + strata(CHT)	188.8279	0.1403	0.0177	0.3948	0.6020	0.5961	0.0449
POHLAVÍ + STAGE + strata(CHT)	181.1610	0.2469	0.1046	0.5924	0.6193	0.5984	0.0410
POHLAVÍ + GRADE + strata(CHT)	189.6009	0.0904	0.0193	0.2749	0.5632	0.5124	0.0464
STAGE + GRADE + strata(CHT)	178.2775	0.3430	0.1525	0.7175	0.6513	0.6016	0.0379
VĚK + POHLAVÍ + STAGE + strata(CHT)	180.9722	0.3209	0.1562	0.6924	0.6803	0.6859	0.0393
VĚK + POHLAVÍ + GRADE + strata(CHT	190.8242	0.1362	0.0170	0.3858	0.6059	0.6005	0.0451
POHLAVÍ + STAGE + GRADE + strata(CHT)	180.1242	0.3567	0.1686	0.7322	0.6513	0.6465	0.0385
VĚK + POHLAVÍ + STAGE + GRADE + strata(CHT)	180.9711	0.3872	0.2018	0.7622	0.6798	0.6902	0.0370

Tabulka A1: Hodnoty AIC, R^2_N, R^2_{XO}, R^2_{OXS}, Harrellova C, Gönenova a Hellerova C a Brierova skóre pro modely bez interakcí

Tabulka pro výběr a hodnocení modelů s interakcemi

Model	AIC	R^2_N	R^2_{XO}	R^2_{OXS}	C_H	C_{GH}	BS
VĚK:strata(CHT)	190.1226	0.0895	0.0055	0.2725	0.5808	0.5185	0.0461
VĚK:POHLAVÍ + strata(CHT)	190.1215	0.0880	0.0055	0.2687	0.5759	0.5216	0.0461
VĚK:STAGE + strata(CHT)	178.4273	0.3240	0.1658	0.6960	0.6783	0.6362	0.0394
VĚK:GRADE + strata(CHT)	188.9451	0.1458	0.0156	0.4070	0.6030	0.5560	0.0448
VĚK:STAGE + POHLAVÍ + strata(CHT)	180.3972	0.3217	0.1627	0.6933	0.6782	0.6382	0.0395
VĚK:STAGE + GRADE + strata(CHT)	178.4552	0.3828	0.2068	0.7581	0.6773	0.6545	0.0373
VĚK:STAGE + STAGE + strata(CHT)	181.9808	0.3113	0.1520	0.6809	0.6645	0.6307	0.0396
POHLAVÍ + VĚK:strata(CHT)	192.1224	0.0177	0.0048	0.0604	0.5837	0.5187	0.0475
VĚK + VĚK:strata(CHT)	190.1226	0.0288	0.0038	0.0967	0.5808	0.5114	0.0473
STAGE + VĚK:strata(CHT)	180.8228	0.2923	0.1180	0.6568	0.6793	0.6332	0.0391
GRADE + VĚK:strata(CHT)	190.8275	0.0821	0.0160	0.2531	0.6030	0.5597	0.0462
STAGE:strata(CHT)	179.2774	0.2886	0.2167	0.6519	0.6320	0.4890	0.0422
VĚK + STAGE:strata(CHT)	179.3735	0.3206	0.3112	0.6920	0.6783	0.6552	0.0435
POHLAVÍ + STAGE:strata(CHT)	181.2773	0.2886	0.2167	0.6519	0.6369	0.5505	0.0422
GRADE + STAGE:strata(CHT)	179.0307	0.3486	0.2574	0.7236	0.6606	0.5957	0.0407
POHLAVÍ:STAGE + strata(CHT)	182.3774	0.1515	0.1684	0.4192	0.6754	0.6024	0.0432
GRADE:STAGE + strata(CHT)	181.6481	0.3463	0.1843	0.7211	0.6685	0.6085	0.0390

Tabulka A2: Hodnoty AIC, R^2_N, R^2_{XO}, R^2_{OXS}, Harrellova C, Gönenova a Hellerova C a Brierova skóre pro modely s interakcemi