FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Zadání bakalářské práce

Název:	Odhady parametrů v modelu zrychleného času
Student:	Martin Benedikt
Vedoucí:	Mgr. Petr Novák, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Umělá inteligence 2021
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2024/2025

Elektronicky schválil/a Ing. Magda Friedjungová, Ph.D. dne 1. listopadu 2023 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Pokyny pro vypracování

Regresní model zrychleného času (Accelerated failure time, AFT) je nástrojem pro statistickou analýzu a interpretaci cenzorovaných dat s využitím zejména v oblasti přežití a spolehlivosti. Pokud není specifikovaná základní riziková funkce modelu, pro odhady jeho parametrů neexistuje explicitní vyjádření a je potřeba je získat numericky. Cílem práce je prozkoumat možnosti odhadů parametrů, postupy implementovat a vyzkoušet na reálných i simulovaných datech.

Cíle:
1. Nastudujte a popište problematiku statistické analýzy přežití s cenzorovanými daty.
2. Představte nejčastěji používané regresní modely v této oblasti. Zaměřte se na model zrychleného času a způsoby odhadu jeho parametrů.
3. Implementujte postupy odhadů formou balíčku pro jazyk R.
4. Prozkoumejte chování odhadů pomocí simulačních experimentů s různými velikostmi vzorků, mírou cenzorování apod.
5. Předveďte použití metod na reálných datech, včetně interpretace výsledků.

Literatura:
1. Buckley J., James I.: Linear regression with censored data. Biometrika, Volume 66, Issue 3, December 1979, Pages 429–436, https://doi.org/10.1093/biomet/66.3.429
2. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. 2nd Edition. John Wiley and Sons, New York. 2002, ISBN 978-0-471-36357-6.
3. Martinussen T., Scheike T. H.: Dynamic Regression Models for Survival Data. Springer, New York, 2006, ISBN 978-0387-20274-7.

Elektronicky schválil/a Ing. Magda Friedjungová, Ph.D. dne 1. listopadu 2023 v Praze.

Bakalářská práce

ODHADY PARAMETRŮ V MODELU ZRYCHLENÉHO ČASU

Martin Benedikt

Fakulta informačních technologií
Katedra aplikované matematiky
Vedoucí: Mgr. Petr Novák, Ph.D.
13. května 2024

Obsah

Seznam obrázků

Seznam tabulek

Seznam výpisů kódu

Chtěl bych poděkovat především Mgr. Petru Novákovi, Phd., za trpělivost, konzultace a výborné vedení po celou dobu psaní této práce. Dále bych chtěl poděkovat mé rodině a blízkým za neutuchající podporu při studiu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 citovaného zákona.

V Praze dne 13. května 2024

Abstrakt

Bakalářská práce se zaměřuje na implementaci odhadu regresních parametrů u semiparametrického modelu zrychleného času. V práci je postupně představena analýza přežití, přístup k analýze přežití pomocí čítacích procesů a regresní modelování v analýze přežití. Poté následují kapitoly, kde se detailně popisují zvolené metody odhadu regresních parametrů a sestavený R balíček. Vybrané metody jsou v závěru otestovány jak na reálných, tak na simulovaných datech. Na umělých datech je provedena simulační studie, kde jsou vybrané postupy testovány v několika vybraných scénářích s různou procentuální mírou cenzorování, velikostí vzorků a různými rozděleními odchylek modelu. Praktický výsledek práce je naimplementovaný R balíček, který realizuje vybrané metody odhadu regresních koeficientů a obsahuje všechny potřebné nástroje pro semiparametrickou analýzu cenzorovaných dat.

Klíčová slova model zrychleného času, odhad regresních parametrů, analýza přežití, cenzorováná data, semiparametrická analýza, pořadové metody, metoda nejmenších čtverců

Abstract

The bachelor's thesis focuses on the implementation of regression parameter estimation in a semiparametric accelerated failure time model. The thesis first introduces survival analysis, the approach to survival analysis using counting processes, and then regression modeling in survival analysis. Subsequent chapters provide detailed descriptions of the chosen methods for estimating regression parameters of semiparametric accelerated failure time model and the constructed R package. The selected methods are tested at the end using both real and simulated data. An extensive simulation study was conducted on simulated data, where were selected methods tested with different scenarios including different procentual number of censoring, number of observations and different distributions of model deviations. The practical result of the thesis is an implemented R package that executes selected methods for estimating regression coefficients and includes all necessary tools for the semiparametric analysis of censored data.

Keywords accelerated failure time model, regression parameter estimation, survival analysis, censored data, semiparametric analysis, rank based methods, least squares

Seznam zkratek

AFT	Accelerated failure time model - model zrychleného času
i.i.d.	independent and identically distributed - nezávislé stejně rozdělené

Úvod

Analýza cenzorovaných dat je důležitá disciplína zejména v odvětvích biologie, medicíny, ekonomie či elektrotechniky. Zajímavé je především regresní modelování, kde zkoumáme vliv jednotlivých kovariát na čas přežití, tedy čas ve kterém subjekt zažil nějakou předem specifikovanou sledovanou událost, kde pod pojmem sledované události rozumíme například smrt pacienta či porouchání nějaké mechanické součástky.

V analýze přežití je Coxův model proporcionálního rizika bezesporu nejpoužívanějším regresním modelem, pokud ovšem nejsou splněny předpoklady jeho použití, tedy předpoklad proporcionality rizik, tak dává nepřesné a zkreslené výsledky. V tomto případě se nabízí použít jiný alternativní model jako je například model zrychleného času, který zároveň poskytuje mnohem více intuitivní interpretaci regresních koeficientů, jelikož na rozdíl od Coxova modelu nemodeluje intezitu nýbrž samotné časy přežití.

Tato práce si klade za cíl představit možnosti odhadů regresních parametrů a jejich rozptylu pro model zrychleného času. Konkrétně se zaměříme na jeho semiparametrickou variantu, která nevyžaduje informaci o statistickém rozdělení časů přežití.

V práci je nejdříve představena statistická analýza přežití od úplného základu. Je také představena její moderní formulace pomocí čítacích procesů a martingalů, které jsou popsány i s příslušnou vybudovanou teorií ohledně náhodných procesů. Dále jsou představeny základní semiparametrické regresní modely používané v analýze přežití a to konkrétně Coxův model proporcionálního rizika, Aalenův aditivní model a samotný model zrychleného času. Poté jsou detailně popsány možnosti odhadů regresních parametrů u semiparametrického modelu zrychleného času. Je kladen důraz jak na odhady založené na pořadových metodách, které pro odhad používají lineární pořadové testy, tak na metody založené na vhodné úpravě metody nejmenších čtverců, které staví na Buckleyho algoritmu ze 70.let minulého století.

Vybrané způsoby odhadů regresních parametrů a jejich rozptylu jsou naprogramovány v rámci R balíčku, který obsahuje všechny základní nástroje pro statistickou analýzu cenzorovaných dat. Balíček je v samotném závěru použit na analýzu reálných dat ze Stanfordského transplantačního programu a také je provedena simulační studie na umělých datech. V simulační studii jsou zkoumány vlastnosti jednotlivých odhadů v závislosti na různých parametrech dat jako je procentuální míra cenzorování, velikost datasetu či pravděpodobnostní rozdělení odchylek. Také je experimentováno s různými optimalizačními algoritmy a počátečními odhady. Dosažené experimentální výsledky jsou poté zhodnoceny a jsou vybrány nejlepší odhady pro jednotlivé scénáře.

Kapitola 1

Analýza přežití

1.1 Úvod do analýzy přežití

Analýza přežití, jinde označována jako analýza spolehlivosti, je odvětvím statistiky zaměřené na analýzu času do výskytu nějaké sledované události. Uplatnění nachází zejména v medicíně, ekonomii, sociologii, biologii či v elektrotechnice. Hlavní specifikum dat přežití je výskyt nezáporné časové složky T^*, označované jako čas přežití, která reflektuje čas do výskytu sledované události. Sledovaná událost musí být jasně definována a ideálně snadno zjistitelná [1]. Včasné pozorování totiž umožňuje co nejpřesněji naměřit čas přežití a tedy vede ke kvalitním výsledkům. Jako příklad sledované události se nejčastěji uvádí smrt pacienta či porouchání součástky, obecně ale může jít o jakoukoliv událost. Dále je potřeba dobře definovat počáteční a koncový bod pozorování pro každý subjekt studie, jelikož zvolení špatného počátku může zásadně změnit interpretaci výsledků. Základním cílem analýzy přežití je poskytnutí kvantitativních odpovědí na otázky týkající se časů přežití a identifikovat faktory, které mohou tyto časy ovlivnit.

1.2 Cenzorování

Během pozorování nemusí u všech subjektů nastat očekávaná událost. Například při zkoumání životnosti součástek nemusí všechny selhat. Mohou také nastat kompetitivní události, které zabraňují výskytu hlavní události (např. selhání součástky z důvodu pádu). Ačkoliv pozorování těchto subjektů je neúplné, je důležité je zahrnout do výsledné analýzy, aby se předešlo zkreslení interpretace výsledků.

Čas přežití u subjektů, u nichž hlavní událost nenastala, se označuje jako cenzorovaný. Předpokládáme, že cenzorování je nezávislé, tedy že pravděpodobnost cenzorování nesouvisí s pravděpodobností výskytu sledované události. V této práci se zaměřujeme výhradně na cenzorování zprava.

▶ **Definice 1.1.** *Definujme $T_1^*, T_2^*, ..., T_n^*$ jako skutečné časy přežití a $C_1, C_2, ..., C_n$ jako časy cenzorování. Pokud platí $T_i^* > C_i$ pro $i = 1, ..., n.$,tak označujeme i-tý čas přežití jako zprava cenzorovaný.*

Pro jednotlivé subjekty tedy zavádíme časy přežití jako $T_i = min(T_i^*, C_i)$. K rozlišení dat zavádíme indikátorovou proměnnou:

$$\delta_i = \begin{cases} 1 & \text{pokud je čas u i-tého subjektu necenzorovaný} \\ 0 & \text{jinak} \end{cases}$$

Data jsou tedy reprezentována jako náhodný vektor (T_i, δ_i). V této práci se budeme zabývat regresními modely v analýze přežití, tudíž pro jednotlivé subjekty ještě přidáváme vektor kovariát $\mathbf{Z_i}$ a výsledná reprezentace je ve tvaru $(T_i, \delta_i, \mathbf{Z_i})$.

▶ Poznámka 1.2. V analýze přežití existují i další formy cenzorování, např. cenzorování zleva nebo intervalové cenzorování.

1.3 Funkce v analýze přežití

Mějme $T_1^*, T_2^*, ..., T_n^*$ nezáporné i.i.d. hodnoty s distribuční funkcí $F(t) = P(T^* \leq t)$. Zároveň budeme předpokládat, že hodnoty pochází ze spojitého rozdělení s hustotou $f(t)$. Pravděpodobnostní chování náhodné veličiny T^* potom popisujeme pomocí následujících funkcí[1],[2],[3]

1.3.1 Funkce přežití

V analýze přežití se často využívá funkce přežití (survival function), která vyjadřuje pravděpodobnost dožití subjektu do času t. Definujeme jí jako

$$S(t) = P(T^* > t). \tag{1.1}$$

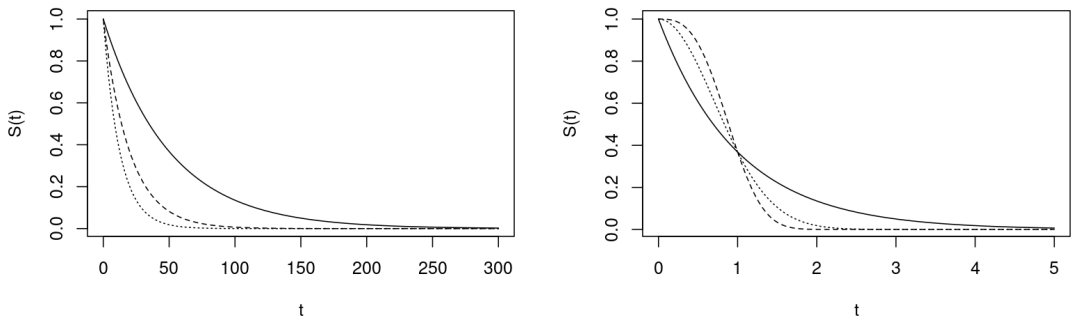
Pokud je T spojitá náhodná veličina, tak S(t) je spojitá, ostře klesající funkce [3], jinak je vždy nerostoucí. Snadno zjistíme, že platí

$$S(t) = P(T^* > t) = 1 - P(T^* \leq t) = 1 - F(t)$$

a také

$$S(t) = P(T^* > t) = \int_t^\infty f(u)du.$$

Ukázky tvarů různých funkcí přežití jsou zobrazeny na obrázku 4.1



■ **Obrázek 1.1** Vlevo jsou znázorněny funkce přežití pro data z exp. rozdělení s $\lambda = 0.02$ (plně), $\lambda = 0.05$ (čárkovaně), $\lambda = 0.08$ (tečkovaně). Vpravo jsou funkce přežití pro data z Welbuillova rozdělení s $\lambda = 1, k = 1$ (plně), $\lambda = 1, k = 3$ (čárkovaně), $\lambda = 1, k = 2$ (tečkovaně)

▶ Poznámka 1.3. Funkce přežití nabývá vždy v čase nula hodnoty jedna a s přirůstajícím časem se blíží limitně k 0. To vede k hezké interpretaci, jelikož subjekt je jistě naživu na začátku studie a s přirůstajícím časem se jeho šance na přežití stále zmenšují. Rychlost poklesu se liší v závislosti na rozdělení dat.

1.3.2 Riziková funkce

Další důležitou funkcí v analýze přežití je riziková funkce (hazard function), často také označovaná jako intenzita selhání. Riziková funkce popisuje intenzitu výskytu sledované události v čase t, za předpokladu že dosud nenastala [3]. Matematicky jí definujeme následovně

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < T^* \leq t + \Delta t | T^* > t). \tag{1.2}$$

Uveďme také její kumulativní variantu

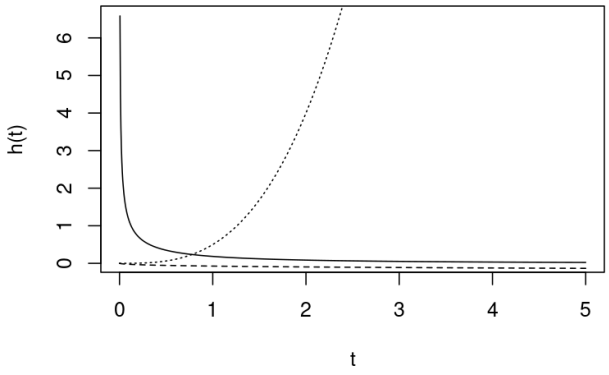
$$H(t) = \int_0^t h(u)du.$$

Rizikovou funkci můžeme vyjádřit i jinak. Pokud do rovnice 1.2 dosadíme vzoreček pro podmíněnou pravděpodobnost, dostaneme

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < T^* \leq t + \Delta t | T^* > t) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t < T^* \leq t + \Delta t \cap T^* > t)}{P(T^* > t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t < T^* \leq t + \Delta t)}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{1.3}$$

A funkci přežití můžeme vyjádřit dále jako

$$\begin{aligned} h(t) &= -\frac{d}{dt} \ln(S(t)) \\ S(t) &= \exp(-H(t)). \end{aligned} \tag{1.4}$$



Obrázek 1.2 Ukázka různých tvarů rizikové funkce pro data pocházející z Wellbuillova rozdělení

▶ Poznámka 1.4. Riziková funkce může nabývat různých tvarů, její funkční hodnoty jsou ale vždy nezáporné.

1.4 Teorie čítacích procesů

Zabýváme se zkoumáním času do nějaké určité události. Ukáže se, že je výhodné taková data reprezentovat pomocí čítacích procesů. Tento přístup poprvé představil Odd O. Aalen v kontextu charakteristik neparametrických odhadů [4]. Později vyšla další literatura prohlubující tento přístup k analýze přežití, uveďme například [5],[6]. Než přijdeme k definici čítacích procesů a jejich využití v analýze přežití, tak budeme muset vybudovat příslušnou teorii. V této sekci používáme definice z knihy [2].

▶ **Definice 1.5.** *Soubor náhodných veličin nad stejným pravděpodobnostním prostorem indexovaných časem* $\{X(t) : t \geq 0\}$*, nazýváme náhodným procesem*

▶ Poznámka 1.6 (značení). Zkrácený zápis procesu $\{X(t) : t \geq 0\}$ budeme značit pro zjednodušení zápisu jako $\mathbf{X}$, tedy stejně jako matici. Aby se předešlo nedorozumění, tak v textu bude vždy jasně specifikováno zda mluvíme o náhodném procesu či například o matici kovariát.

▶ **Definice 1.7.** *Mějme* $\mathbf{X}$ *náhodný proces. Zobrazení* $t \to \mathbf{X}(t, \omega)$*, kde* $\omega \in \Omega$*, nazýváme realizací (trajektorií) náhodného procesu*

▶ **Definice 1.8.** *Náhodný proces* $\mathbf{X}$ *je zprava/zleva spojitý pokud je jeho realizace zprava/zleva spojitá*

Náhodné procesy, které jsou zprava spojité a mají definovanou levou limitu nazýváme **cadlagovými** (continue à droite, limite à gauche) procesy.

▶ **Definice 1.9.** *Systém zvětšujících se* σ*-algeber idukovaný nějakým náhodným procesem* $\mathbf{X}$ *nazýváme fitrací a značíme jako* $\mathcal{F}_t^X = \sigma\{X(s) : 0 \leq s \leq t\}$

Filtraci chápeme jako vnitří historii procesu $\mathbf{X}$. Jinými slovy si jí můžeme představit jako dostupnou informaci do nějakého určitého času. Zobecněnou filtraci pro více náhodných procesů, tedy informaci generovanou těmito procesy, značíme jako $(\mathcal{F}_t : t \geq 0)$.

▶ **Definice 1.10.** *Nezápornou náhodnou veličinu* S *nazýváme zastavujícím časem s ohledem na fitraci* $\mathcal{F}_t$ *pokud* $(S \leq t) \in \mathcal{F}_t$ *pro všechny* $t \geq 0$

Zastavující čas nám definuje okamžik kdy můžeme říci, že máme dostatek informací k tomu abychom potvrdili výskyt určité události (v kontextu analýzy přežití se jedná o sledovanou událost).

▶ **Definice 1.11.** *Mějme náhodný proces* $\mathbf{X}$ *a zastavující čas* S*, proces* $\mathbf{X}^S$ *definovaný jako* $X(t) = X\left[min(t, S)\right]$ *nazýváme zastavený proces*

▶ **Definice 1.12.** *Jako lokalizační posloupnost označujeme posloupnost neklesajících zastavujících časů* S_n *kde* $S_n \to \infty$ *pokud* $n \to \infty$

▶ **Definice 1.13.** *Mějme filtraci* $\mathcal{F}_t$ *pokud jsou* X_s*, kde* $s \in [0, t]$*,* $\mathcal{F}_t$*-měřitelné reálné náhodné veličiny pak označujeme proces* $\{X(t) : t \geq 0\}$ *jako adaptovaný k fitraci* $\mathcal{F}_t$

1.4.1 Martingaly

Důležitým nástrojem v teorii čítacích procesů jsou martingaly, koncept původně označovaný jako sázecí strategie. Zjistíme, že pomocí martingalů a centrální limitní věty pro martingaly můžeme snadno popisovat asymptotické vlastnosti modelů, přesněji vlastnosti jejich odhadujících funkcí.

▶ **Definice 1.14.** *Jako martingal označujeme cadlagový proces* $\mathbf{X}$ *adaptovaný na filtraci* $\mathcal{F}_t$ *který splňuje podmínky*

(i) $E(X(t)) < \infty$ *pro všechna* t

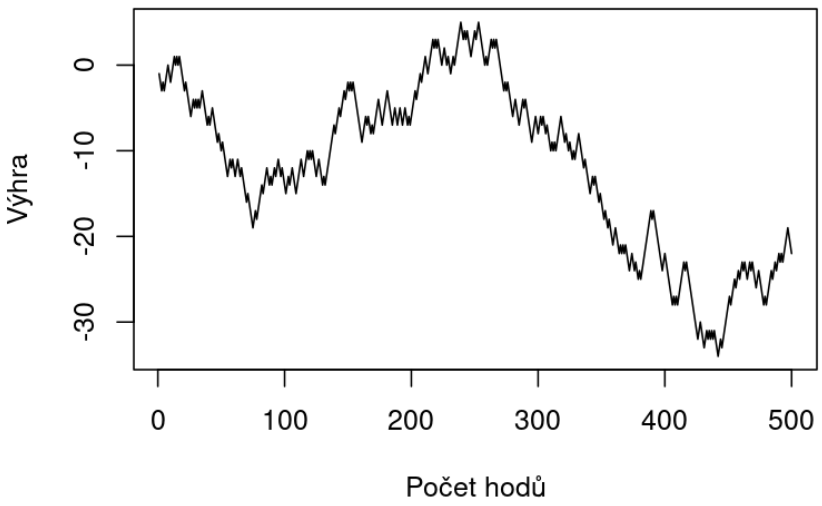
(ii) $E(X(t)|\mathcal{F}_s) = X(s)$ *pro všechna* $s \leq t$

Martingal má tedy nulové přírůstky nezávisle na historii. Pokud navíc platí, že martingal je rovný nule v čase nula, pak hovoříme martingalu s nulovou střední hodnotou $\{E(X(t)) = E(X(0))\}$.

S martingaly se díky jejich vlastnostem často pracuje jako s chybovými procesy vyjadřující rozdíl mezi skutečnou hodnotou parametru a vypočteným odhadem. Na těchto martingalových reziduech se často diagnostikují vhodnosti modelů.

Pokud místo druhé podmínky v definici 1.14 platí, že $E(X(t)|\mathcal{F}_s) \geq X(s)$ pro všechny $s \leq t$ nazýváme proces $\mathbf{X}$ submartingalem.

Uveďme nyní jednoduchou ukázku martingalu na příkladu hodu mincí. Definujme X_n jako náhodnou veličinu reprezentující výhru po n hodech. Hod panny znamená zisk jedné koruny, hod orla naopak ztrátu. Jelikož uvažujeme, že je mince férová, tak po každém hodu je očekávaná hodnota výhry stejná jako dosavadní výdělek. Navíc pro všechny n jistě platí $E(X_n) < \infty$. Proces $\{X_n : n \geq 0\}$ je tedy martingal.



Obrázek 1.3 Ukázka martingalu na simulaci hodu mincí.

▶ **Definice 1.15.** *Za lokální (sub)martingal budeme označovat proces* $\mathbf{X}$ *adaptovaný na filtraci* $\mathcal{F}_t$*, pro který existuje lokalizační posloupnost zastavujících časů* S_n *taková, že pro každé* n, $\mathbf{X}^{S_n}$ *je (sub)martingal.*

Často budeme potřebovat upravit náhodný proces na martingal. K tomu využijeme kompenzátor.

▶ **Definice 1.16.** *Mějme cadlagový proces* $\mathbf{X}$ *adaptovaný na filtraci* $\mathcal{F}_t$*. Řekneme že náhodný proces* $\mathbf{A}$ *je kompenzátor procesu* $\mathbf{X}$ *pokud* $\mathbf{X} - \mathbf{A}$ *je lokální martingal adaptovaný k filtraci* $\mathcal{F}_t$ *s nulovou střední hodnotou. Tento kompenzátor je unikátní*

Nyní můžeme definovat klíčovou větu v teorii martingalů.

▶ **Věta 1.17** (Doob-Meyerova dekompozice)**.** *Cadlagový proces* $\mathbf{X}$ *adaptovaný k filtraci* $\mathcal{F}_t$ *má kompezátor pouze tehdy pokud je* $\mathbf{X}$ *rozdíl dvou lokálních submartingalů*

Jelikož je z definice konstatní proces 0 lokální submartingal, tak můžeme díky Doob-Meyerovy dekompozici říci, že každý submartingal má kompenzátor. Toto tvrzení je velmi důležité, neboť později zjistíme, že čítací proces je také submartingal.

1.4.2 Čítací procesy v analýze přežití

Zabývejme se nyní čítacími procesy. Čítací proces definujeme následovně

▶ **Definice 1.18.** *Cadlagový proces* $\mathbf{N}$ *adaptovaný na fitraci* $\mathcal{F}_t$ *nazýváme čítacím procesem pokud platí*

(i) $N(0) = 0$, $N(t) < \infty$ *pro všechna* t

(ii) jeho realizace je po částech konstantní se skoky o velikosti 1

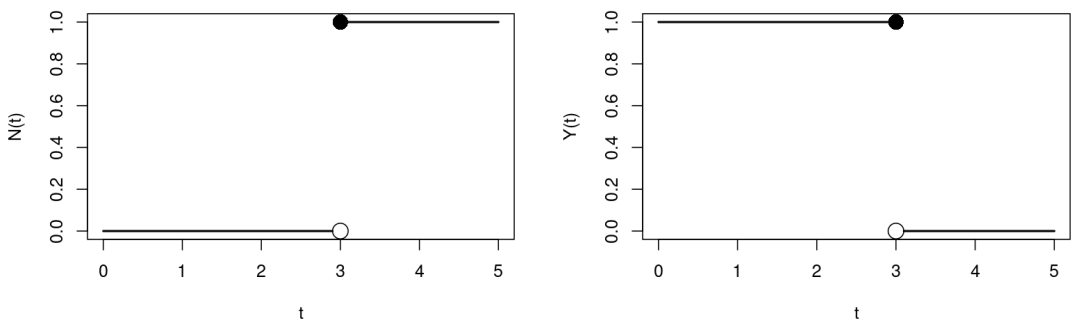
Z kapitoly o cenzorování víme, že máme data reprezentovaná jako náhodný vektor $(T, \delta, \mathbf{Z})$. Pomocí čítacích procesů můžeme tuto reprezantaci nahradit funkcemi času $(N(t), Y(t), \mathbf{Z})$, kde

$$N(t) = I(T \leq t, \delta = 1).$$

Tedy čítací proces, který je nulový a při výskytu necenzorované události skočí na jedničku (v případě dalších událostí pokračuje na dvojku atd.) a

$$Y(t) = I(t \leq T).$$

Proces, který nazýváme indikátorem rizika. Tento proces nabývá hodnoty jedna pokud u subjektu nedošlo ke sledované události či cenzorování, v opačném případě je rovný nule.



■ **Obrázek 1.4** Vlevo je ukázka čítacího procesu se sledovanou událostí v čase t=3. Vpravo je přidružený indikátor rizika

Jak už bylo zmíněno v podsekci o martingalech, čítací proces je submartingal. Toto tvrzení plyne přímo z definice čítacího procesu, jelikož jistě platí že $E\left(N(t)\right) < \infty$ a s přírůstajícím časem se hodnota čítacího procesu vždy zvedá či zůstává stejná (v novém času může vzrůst o 1). Podle Doob-Meyerovi dekompozice může být tedy pomocí kompenzátoru převeden na martingalový proces

$$\mathbf{M} = \mathbf{N} - \mathbf{A}$$

Pokud rovnici přepíšeme na

$$\mathbf{N} = \mathbf{A} + \mathbf{M}$$

chápeme výslednou reprezentaci jako

$$pozorování = model + chyba$$

Užitečnost tohoto zápisu tkví v existenci centrální limitní věty pro martingaly. Ta tvrdí, že martingalový proces za nějakých určitých podmínek, konverguje v limitě ke Gaussovskému procesu s nulovou střední hodnotou a s rozptylem rovným jedné [2]. Pomocí vlastností, která nám tato věta dává, můžeme snadno popisovat asymptotické vlastnosti odhadů pro neparametrické a semiparametrické modely. Samotnou přesnou formulaci centrální limitní věty pro martinglaly v této práci nebudeme ukazovat, lze jí nalézt například zde [7].

Ukáže se, že kompenzátor **A** je vyjádřen jako

$$\mathcal{A}(t) = \int_0^t \lambda(u)du.$$

kde $\lambda(u)$ je procesem intenzity pro $N(t)$ a platí $\lambda(u) = Y(u)h(u)$ [2]. Proces intenzity můžeme interpretovat jako očekávaný počet výskytů sledované události v čase t a kompenzátor jako kumulovaný součet sledovaných událostí od počátku sledování do času t.

1.5 Tvorba věrohodnosti v analýze přežití

Pro odhad regresních parametrů, rizikové funkce či intenzity v různých modelech přežití, je klíčová metoda maximální věrohodnosti. Opět předpokládáme, že máme data reprezentovaná jako náhodný vektor $(T_i, \delta_i)_{i=1}^n$ (prozatím vynechme vektor kovariát), kde cenzorování zprava je nezávislé. Data mějme naměřená na nějakém časovém intervalu $[0, \tau]$ kde τ značí konec studie. Pro pozorování platí

$$P(T,\delta) = \begin{cases} P(T,(\delta=1)) = P(T=T^*|\delta=1)P(\delta=1) = f(T) \\ P(T,(\delta=0)) = P(T=C|\delta=0)P(\delta=0) = P(\delta=0) = P(T^*>C) = S(C) \end{cases}$$

Tyto dvě vyjádření můžeme spojit do jednoho, kde první část odpovídá necenzorovaným pozorováním a druhá cenzorovaným

$$P(T_i,\delta_i) = f(T_i)^{\delta_i} S(T_i)^{(1-\delta_i)}.$$

Věrohodnostní funkce poté nabývá tvaru

$$L = \prod_{i=1}^n P(T_i,\delta_i) = \prod f(T_i)^{\delta_i} S(T_i)^{(1-\delta_i)} \tag{1.5}$$

a její logaritmická verze

$$l = \sum_{i=1}^n \delta_i \log(f(T_i)) + \sum_{i=1}^n (1-\delta_i) S(T_i). \tag{1.6}$$

V literatuře se často uvádí jiný tvar věrohodnostní funkce [3]. Pokud použijeme vyjádření z rovnice 1.3 a vyjádříme hustotu pomocí rizikové funkce a funkce přežití, tak můžeme věrohodnost 1.6 přepsat následovně

$$l = \sum_{i=1}^n \left[\delta_i \log(h(T_i)) - \int_0^{T_i} h(u)du \right]. \tag{1.7}$$

▶ Poznámka 1.19. V případě parametrických odhadů uvažujeme, že riziková funkce je tvaru $h(\theta, t)$. V textu tuto skutečnost pro zjednodušení zápisu zatím vynecháváme

Přejděme nyní k vyjádření věrohodnosti pomocí čítacích procesů. Mějme data ve tvaru $(N_i(t), Y_i(t))_i^n$ a označme $dN_i(t)$ jako skok čítacího procesu v čase t. Tedy proces, který je jedna pokud došlo k necenzorované události v čase t, jinak je nulový. Ukáže se, že věrohodnostní funkce vyjádřená pomocí čítacích procesů adaptovaných k fitraci $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u), 0 \leq u \leq t\}$ má tvar

$$l = \sum_{i=1}^{n} \left[\int_0^\tau \log(\lambda_i(u)) dN_i(u) - \int_0^\tau \lambda_i(u) du \right]. \tag{1.8}$$

Pro odhady parametrů potom zavádíme skórovou funkci $U(\theta)$, která vznikne derivováním 1.8 podle parametrů θ. Samotné odhady získáme položením $U(\theta) = 0$ a vyřešením rovnic

$$U(\theta) = \sum_{i=1}^{n} \left[\int_0^\tau \frac{\partial}{\partial\theta} \log(\lambda_i(u)) dN_i(u) - \int_0^\tau \frac{\partial}{\partial\theta} \lambda_i(u) du \right]. \tag{1.9}$$

1.6 Základní neparametrické odhady

Při studování různých modelů v analýze přežití nás zajímá pravděpodobnostní rozdělení dat přežití. Je potřeba tedy odhadnout funkci přežití a kumulativní rizikovou funkci. V této sekci představíme dva základní neparametrické odhady těchto funkcí, se kterými poté budeme dále pracovat v následujících kapitolách.

Mějme $\mathbf{N}(t) = \sum_{i=1}^n N_i(t)$, $\mathbf{Y}(t) = \sum_{i=1}^n Y_i(t)$, kde $N_i(t)$ a $Y_i(t)$ je čítací proces respektive indikátor rizika adaptovaný k fitraci $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u), 0 \leq u \leq t\}$. Nelsonův-Aalenův odhad [4] je neparametrický odhad kumulativní rizikové funkce definovaný jako

$$\hat{H}(t) = \int_0^t \frac{J(u)}{\mathbf{Y}(t)} \mathbf{dN}(u), \tag{1.10}$$

kde $J(u) = I(\mathbf{Y}(u) > 0)$ a zároveň zavádíme konvenci $\frac{0}{0} = 0$. Nelsonův-Aalenův odhad vychází z Doob-Meierovy dekompozice. Označme $\mathbf{M}(t) = N(t) - \mathbf{A}(t)$, kde $\mathbf{M}(t) = \sum_{i=1}^n M_i(t)$ a $\mathbf{A}(t) = \sum_{i=1}^{A_i(t)}$. Poté platí

$$\mathbf{dN}(t) = \mathbf{dA}(t) + \mathbf{dM}(t) \tag{1.11}$$

Nelsonův-Aalenův odhad poté dostaneme postupným vyjádřením rizikové funkce z kompenzátoru A a zintegrováním rovnice 1.11 [2].

Jelikož nyní známe odhad kumulativní rizikové funkce, tak můžeme vyjádřit odhad funkce přežití pomocí definice 1.4

$$\hat{S}(t) = \exp(-\hat{H}(t)).$$

Tento odhad se nazývá Fleming-Harringtonův odhad funkce přežití. Jeho nevýhodou je, že předpokládá že data pocházejí z absolutně spojitého rozdělení. Odhad funkce přežití, který můžeme aplikovat jak na spojité tak na diskrétní data je Kaplan-Meierův [8]. Dejinujeme jej následovně

$$\hat{S}(t) = \prod_{u \leq t} \left[1 - \mathbf{d}\hat{H}(u) \right] = \prod_{u \leq t} \left[1 - \frac{\mathbf{dN}(u)}{\mathbf{Y}(u)} \right]. \tag{1.12}$$

▶ Poznámka 1.20. Kaplan-Meierův odhad se dá také odvodit pomocí metody maximální věrohodnosti. Odvození můžeme najít například v [3]

Kapitola 2

Regresní modely v analýze přežití

Základní úlohou u regresního modelování je zkoumání vlivu kovariát na rizikovou funkci, intenzitu či dokonce na čas přežití.

V této kapitole si zběžně představíme Coxův model proporcionálního rizika, který je v dnešní době standardní nástroj pro regresní modelování v analýze přežití. Pro Coxův model si ukážeme jak odhadovat regresní parametry a jejich rozptyl. Dále představíme Aalenův aditivní model a na konci kapitoly se zaměříme na samotný model zrychleného času. Možnostem odhadů regresních koeficientů u modelu zrychleného času se poté budeme věnovat v další kapitole.

2.1 Coxův model proporcionálního rizika

Jednoznačně nejpoužívanějším regresím modelem v analýze přežití je Coxův model [9], který modeluje vliv jednotlivých kovariát multiplikativně na rizikovou funkci. Model předpokládá, že riziková funkce pro i-tý subjekt nabývá tvaru

$$h_i(t) = h_0(t) \exp\left(\mathbf{Z}_i^T(t)\beta\right), \tag{2.1}$$

kde β je $p \times 1$ dimenzionální vektor regresních parametrů, $\mathbf{Z}_i^T(t)$ je $1 \times p$ dimenzionální vektor kovariát pro i-tý subjekt a h_0 je předem nespecifikovaná funkce, představující základní rizikovou funkci pro všechny subjekty. Definice 2.1 představuje tzv. zobecněný Coxův model, který předpokládá že kovariáty $\mathbf{Z}_i$ jsou funkcemi času, tedy mohou nabývat v různých časech jiných hodnot. V této práci ovšem budeme uvažovat pouze vysvětlující proměnné konstatní v čase.

Vliv kovariát je vyjádřen pomocí vektoru $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$, kde jednotlivé složky β_i $i \in (1, 2, \ldots, p)$ představují změnu, o kterou se změní přirozený logaritmus rizikové funkce při hodnotě kovariátu $\mathbf{Z}_i^T$. Tedy kladná hodnota β_i znamená, že i-tý kovariát působí negativně na jedince a zvyšuje hodnotu rizika. Záporná hodnota β_i naopak riziko snižuje a má tedy kladný účinek na pacienta.

V literatuře, využívající přístup pomocí čítacích procesů, se Coxův model častěji uvádí pomocí intenzity [2],[10]

$$\lambda_i(t) = Y_i(t)\lambda_0(t) \exp\left(\mathbf{Z}_i^T(t)\beta\right), \tag{2.2}$$

kde $Y(t)$ označuje indikátor rizika a $\lambda_0(t)$ základní intenzitu společnou pro všechny subjekty. Základní intenzita není, opět jako v případě základní rizikové funkce z definice 2.1, specifikována a odhaduje se neparametricky z dat. Musí pouze splňovat podmínku $\int_0^\tau \lambda_0(t)dt < \infty$, kde τ označuje čas ukončení studie. V praxi to znamená, že neklademe žádné předpoklady na distribuci času přežití pro pacienty s nulovými kovariáty, Coxův model patří tedy do rodiny semiparametrických modelů.

Důležitý předpoklad Coxova modelu je kladen na relativní riziko, které musí být konstatní v čase a působí na intezitu multiplikativně. Relativní riziko pro model 2.2 formulujeme následovně

$$\frac{\lambda\left(t, Z_1, Z_2, \ldots Z_k + 1, \ldots Z_n\right)}{\lambda\left(t, Z_1, Z_2, \ldots Z_n\right)} = \exp\left(\beta_k\right) \tag{2.3}$$

Jedná se o vyjádření vztahu β_k na intenzitu. Pokud není splněn předpoklad pro relativní riziko, tak může Coxův model poskytovat zavádějící výsledky [2],[11]. Je potřeba tedy testovat, zda model vystihuje daná data správně, pomocí testů dobré shody. V praxi se ovšem často na testování zapomíná [11]. Jednou z možností jak tento problém vyřešit je rozšířit model pro koeficienty měnící se v čase [12] či zkusit modelovat závislost pomocí jiné než exponencionální funkce [2].

2.1.1 Odhady parametrů a základní intenzity v Coxově modelu proporcionálního rizika

Mějme data ve tvaru $(N_i(t), Y_i(t), \mathbf{Z}_i(t))_i^n$ naměřená na nějakém časovém intervalu $[0, \tau]$, kde τ označuje konec pozorování. Čítací procesy jsou adaptované k fitraci $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u), \mathbf{Z}_i(t), 0 \leq u \leq t\}$. Dále předpokládejme, že pro každé $N_i(t)$ nabývá intenzita tvaru 2.2 a definujme $\mathcal{A}(t) = \int_0^t \lambda_0(u)d(u)$, $\mathbf{N}(t) = \sum_i N_i(t)$. Kumulovanou intenzitu můžeme, při známém odhadu β, odhadnout neparametricky pomocí Breslowova odhadu jako

$$\hat{\mathcal{A}}(t) = \int_0^t \frac{1}{S_0(u, \beta)} \mathbf{dN}(u), \tag{2.4}$$

kde

$$S_0(t, \beta) = \sum_{i=1}^{n} Y_i(t) \exp\left(\mathbf{Z}_i^T(t)\beta\right).$$

Zároveň ještě zavádíme první a druhou derivaci $S_0(t, \beta)$ podle β jako $S_1(t, \beta)$ respektive $S_2(t, \beta)$

$$S_1(t, \beta) = \sum_{i=1}^{n} Y_i(t) \exp\left(\mathbf{Z}_i^T(t)\beta\right) \mathbf{Z}_i,$$

$$S_2(t, \beta) = \sum_{i=1}^{n} Y_i(t) \exp\left(\mathbf{Z}_i^T(t)\beta\right) \mathbf{Z}_i^T \mathbf{Z}_i$$

a vážený průměr $\mathbf{Z}$ přes pozorování, která jsou stále v riziku v čase t jako

$$E(t, \beta) = \frac{S_1(t, \beta)}{S_0(t, \beta)}.$$

Samotný odhad regresního vektoru β získáme maximalizováním tzv. Coxovo parciální věrohodnosti [13].

$$L\left(\beta\right) = \prod_t \prod_i \frac{\exp\left(\mathbf{Z}_i^T(t)\beta\right)}{S_0(t, \beta)}^{dN_i(t)}. \tag{2.5}$$

Skórový vektor $U(\beta)$ získáme postupně logaritmováním 2.5 a derivováním podle β

$$l(\beta) = \sum_i^n \int_0^\tau \left[\mathbf{Z}_i^T(t)\beta - \log\left(S_0(t, \beta)\right)\right] dN_i(t),$$

$$U(\beta) = \frac{\partial}{\partial \beta} l(\beta) = \sum_i^n \int_0^\tau \left[\mathbf{Z}_i(t) - E(t, \beta)\right] dN_i(t). \tag{2.6}$$

Odhad $\hat{\beta}$ poté získáme řešením $U(\beta) = 0$, v praxi se často používá Newton-Raphsonův algoritmus či evoluční algoritmy.

Matici $\frac{-\partial}{\partial\beta}U(\beta)$ o rozměrech p×p nazýváme informační maticí. Můžeme jí vyjádřit jako

$$I(t,\beta) = \sum_{i=1}^{n} \int_0^t \left[\frac{S_2(u,\beta)}{S_0(u,\beta)} - E(u,\beta)^{\otimes 2} \right] dNi(u), \tag{2.7}$$

kde pro jakýkoliv sloupcový vektor a značíme $a^{\otimes 2} = aa^T$. Řešení $\hat{\beta}$ je poté konzistentní, asymptoticky normálně rozdělené se střední hodnotou β a rozptylem $\{I(\tau,\beta)\}^{-1}$ [2],[14]

2.2 Aalenův aditivní model

Aalenův aditivní model [15] se na rozdíl od Coxova modelu snaží modelovat intenzitu neparametricky. Model předpokládá že intenzita i-tého subjektu, pro čítací proces $[N_i(t), t \in [0,\tau], \tau < \inf]$, nabývá tvaru

$$\lambda_i(t) = Y_i(t)\mathbf{Z}_i^T(t)\beta(t), \tag{2.8}$$

kde $\mathbf{Z}_i(t)$ je $p \times 1$ rozměrný vektor kovariát pro i-tého jedince, $Y_i(t)$ je indikátor rizika a $\beta(t)$ je $p \times 1$ dimenzionální vektor regresních koeficientů. Model předpokládá, že regresní koeficienty jsou funkcemi času a můžou na časovém intervalu $[0,\tau]$ nabývat různých hodnot.

Neparametrická část modelu se skrývá právě v regresních koeficientech β. Místo odhadu jednotlivých parametrů se snažíme odhadnout kumulovaný odhad

$$B(t) = \int_0^t \beta(u)du. \tag{2.9}$$

Označme $\mathbf{W}(t) = (Y_1(t)\mathbf{Z}_1(t), Y_2(t)\mathbf{Z}_2(t), \ldots, Y_n(t)\mathbf{Z}_n(t))$. Díky Doob-Meyerově dekompozici můžeme vyjádřit martingalové procesy jako $M_i(t) = N_i(t) - \int_0^t \lambda_i(s)ds$. Kumulované koeficienty $B(t)$ můžeme poté odhadnout pomocí Aalenova odhadu jako

$$\hat{B}(t) = \int_0^t \left[W^-(s)dN(s) \right],$$

kde $W^-(t) = \left(W(t)^T W(t)\right)^{-1} W(t)^T$ [15],[2].

Neparametrické vyjádření a poměrně jednoduchý způsob odhadu regresních parametrů dělá z Aalenova modelu zajímavou alternativu pro Coxův model proporcionálního rizika. Díky tomu, že je model lineární v kovariátech, můžeme jednoduše detekovat změny v koeficientech v každém jednotlivém času přežití.

V praxi je tento model ovšem poměrně přehlížen. Jelikož nepracujeme s jednotlivými koeficienty $\beta_i(t)$, ale jenom s kumulovanou variantou $B(t)$, tak je celková interpretace vlivu kovariát na intenzitu méně intuitivní než u Coxova modelu. Model zároveň není plně vyvinut pro inferenční účely [2].

Pokud data nevystihuje multiplikativní a ani aditivní model, tak se nabízí možnost zkombinovat oba modely do jednoho. Mluvíme poté o tzv. multiplikativně-aditivních modelech [2].

2.3 Model zrychleného času

Model zrychleného času (Accelerated failure time model – Aft) [16],[17] se snaží modelovat data přežití podobně jako klasický model lineární regrese a vychází z myšlenky, že kovariáty zrychlují či zpomalují čas přežití jednotlivých subjektů. Model specifikujeme následovně

$$\log T_i = \mathbf{Z}_i^T \beta + \epsilon_i, \tag{2.10}$$

kde T_i je zprava cenzorovaný čas přežití, β je $p \times 1$ rozměrný vektor regresních parametrů (intercept se většinou neuvádí jelikož bylo ukázáno, že jej nelze spolehlivě odhadovat [18]) , $\mathbf{Z}_i$ je přidružený $p \times 1$ rozměrný vektor kovariát pro i-tý subjekt a ϵ jsou nezávislé odchylky se stejným, ale kompletně nespecifikovaným rozdělením. Jedná se tedy stejně jako v případě Coxova modelu o semiparametrický model (v praxi se ovšem používají i parametrické varianty). Toto vyjádření umožňuje více přímou interpretaci vlivu regresorů na čas přežití a zároveň neklade žádné požadavky na rozdělení odchylek, jedná se tedy o užitečnou alternativu ke Coxova modelu pokud není splněn předpoklad pro relativní riziko [18].

Vyjádření 2.10 vede k opačné interpretaci koeficientů β než u Coxova modelu. Pokud je například β_i kladné, tak i-tý kovariát prodlužuje čas přežití subjektu a má na něj pozitivní účinek. Naopak záporné β_i čas přežití zkracuje a zrychluje u pacienta výskyt sledované události. Pokud bychom chtěli mít interpretaci regresních koeficientů stejnou jako u Coxova modelu, tak můžeme vyjádření 2.10 přepsat na

$$\log T_i = -\mathbf{Z}_i^T \beta + \epsilon_i.$$

V této práci budeme ovšem používat původně zadefinovaný tvar 2.10, jelikož je pro autora matematicky více přívětivý.

Přítomnost cenzorování v časech přežití zabraňuje použití standardních regresních metod a přináší řadu problémů pro semiparametrickou analýzu [18],[19],[16],[17],[2]. V průběhu let byly vytvořeny procedury pro odhadování regresních parametrů a jejich rozdělení, tyto postupy budeme probírat v následující kapitole.

Pokud bychom chtěli vyjádřit model pomocí rizikové funkce, tak vyjdeme z vyjádření 1.3. Distribuční funkce $F(t)$ pro model 2.10 vypadá následovně

$$\begin{aligned} F(t) &= P(T \leq t) = P(\exp\left(\mathbf{Z}^T\beta\right)\exp\left(\epsilon\right) \leq t) \\ &= P(\exp\left(\epsilon\right) \leq t\exp\left(-\mathbf{Z}^T\beta\right)) \\ &= F_0(t\exp\left(-\mathbf{Z}^T\beta\right)), \end{aligned} \tag{2.11}$$

kde F_0 značí distribuční funkci pro odchylky $\exp\left(\epsilon\right)$.

$$\begin{aligned} h_i(t) &= \frac{f(t)}{S(t)} = \frac{F(t)'}{1-F(t)} \\ &= \frac{f_0(t\exp\left(-\mathbf{Z}_i^T\beta\right))\exp\left(-\mathbf{Z}_i^T\beta\right)}{1-F_0(t\exp\left(-\mathbf{Z}_i^T\beta\right))} \\ &= h_0(t\exp\left(-\mathbf{Z}_i^T\beta\right))\exp\left(-\mathbf{Z}_i^T\beta\right), \end{aligned} \tag{2.12}$$

kde h_0 značí riziko asociované s neznámými odchylkami $\exp\left(\epsilon\right)$. Z vyjádření rizikové funkce je poté dobře vidět jak kovariáty působí multiplikativně na čas t a tedy s ním manipulují o nějakou konstantu. Také je opět potřeba zmínit, že model může být pro nějaká data nedostačující a je potřeba ho testovat pomocí testů dobré shody. Model může být rozšířen pro kovariáty měnící se v čase, nebo můžeme dokonce modelovat pomocí jiné než logaritmické závislosti, odhad parametrů je ovšem mnohem složitější než u obyčejného modelu zrychleného času [2].

▶ Poznámka 2.1. Pokud je základní rozdělení Wellbuilovo, tak se dá ukázat že je model zrychleného času totožný s Coxovým modelem proporcionálního rizika [2].

Kapitola 3

Odhady parametrů v modelu zrychleného času

V této kapitole si představíme možnosti odhadů regresních parametrů v semiparametrickém modelu zrychleného času. V posledních 40 letech bylo semiparametrické odhadování intenzivě zkoumáno a bylo navrženo několik statistických metod pro odhad regresních koeficientů. Zaměříme jak na staré, tak i na novější přístupy. Vybrané přístupy poté budou shrnuty a krátce bude představena i možnost odhadu kovarianční matice pro regresní parametry.

3.1 Pořadové metody

Pořadové metody pro odhad regresních koeficientů byly poprvé představeny Prenticem v roce 1978 [20]. Asymptotické vlastnosti těchto odhadů byly poté odvozeny o dvanáct let později Tsiatisem [21]. Základní myšlenka spočívá v použití lineárních pořadových testů, pro testování hypotézy $\beta = 0$, jako odhadujících funkcí[21]. My se ke stejnému výsledku dostaneme pomocí vhodné úpravy časové škály pro čítací procesy a použitím věrohodnostní funkce.

Mějme data uspořádána jako náhodný vektor $(N_i(t), Y_i(t), \mathbf{Z}_i)_i^n$ s čítacími procesy adaptovanými na filtraci $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u), \mathbf{Z}_i(t), 0 \leq u \leq t\}$. Ukáže se, že pro vytvoření skórového vektoru je výhodnější pracovat s čítacími procesy na transformované časové škále. Označme tedy transformované procesy jako

$$N_i^*(t) = N_i(t \exp\left(\mathbf{Z}_i^T \beta\right)), \;\; Y_i^*(t, \beta) = Y_i(t \exp\left(\mathbf{Z}_i^T \beta\right)).$$

S procesy $N_i^*(t)$ a můžeme poté vyjádřit intenzitu následovně

$$\lambda_i^*(t, \beta) = Y_i^*(t, \beta) h_0(t). \tag{3.1}$$

Označme dále a $S_0^*(t, \beta) = \sum_{i=1}^n Y_i^*(t, \beta), \;\; S_1^*(t, \beta) = \sum_{i=1}^n Y_i^*(t, \beta)\mathbf{Z}_i$.

Pro odhad kumulované základní rizikové funkce můžeme opět použít Nelsonův-Aalenův odhad z definice 1.10.

$$\hat{H}_0(t) = \int_0^t \frac{J(u)}{S_0^*(u, \beta)} \mathbf{dN}^*(u), \tag{3.2}$$

kde $\mathbf{dN}^*(u) = \sum_{i=1}^n dN_i^*(u)$ a $J(u) = I(S_0^*(u, \beta) > 0)$. Samotný odhad neznámé základní rizikové funkce $h_0(t)$ můžeme poté nahradit $d\hat{H}_0(t)$. Pro odhad regresních parametrů β vyjdeme ze skórové funkce z definice 1.9.

$$U(\beta) = \sum_{i=1}^{n} \left[\int_0^\tau \frac{\partial}{\partial \beta} \log(\lambda_i(u)) dN_i(u) - \int_0^\tau \frac{\partial}{\partial \beta} \lambda_i(u) du \right]$$
$$= \sum_{i=1}^{n} \left[\int_0^\tau \frac{\frac{\partial}{\partial \beta} \lambda_i(t)}{\lambda_i(t)} \left(dN_i(t) - Y_i(t) \lambda_i(t) \right) \right].$$

Po zavedení substituce $u = t \exp\left(-\mathbf{Z}_i^T \beta\right)$ a dosazením transformovaných procesů do naší skórové funkce, dostaneme

$$U_W(\beta) = \sum_{i=1}^{n} \left[\int_0^\tau \left(\frac{h_0'(u)u}{h_0(u)} + 1 \right) \mathbf{Z}_i \left(dN_i^*(u) - Y_i^*(u, \beta) d\hat{H}_0(u) \right) \right]$$
$$= \sum_{i=1}^{n} \left[\int_0^\tau W(u) \left(\mathbf{Z}_i - E^*(u, \beta) \right) dN_i^*(u) \right], \tag{3.3}$$

kde

$$W(u) = \left(\frac{h_0'(u)u}{h_0(u)} + 1 \right), \; E^*(u, \beta) = \frac{S_1^*(u, \beta)}{S_0^*(u, \beta)}.$$

Funkce 3.3 nemůže být přímo použita k odhadujícím účelům jelikož váhová funkce $W(u)$ stále obsahuje neznámou derivaci funkce $h_0(u)$. Derivaci bychom se mohli pokusit odhadnout numericky, tento postup se ovšem nedoporučuje jelikož je těžké najít nějaký spolehlivý odhad [2][19]. Řešení spočívá v nahrazení $W(u)$ jinou, snadno vypočítatelnou, váhovou funkcí, která splňuje podmínky sepsané Yingem v [22].

Obecně není $U_W(\beta)$ spojitá funkce, ani po částech monotónní. Řešení $U_W(\beta) = 0$ je tedy složité najít. Dokonce může existovat několik možných kořenů, z nichž některé jsou nekonzistentní [21]. Nabízí se tedy možnost minimalizovat $||U_W(\beta)||$, ovšem vzhledem k nespojitosti a nemonotónnosti funkce $U_W(\beta)$ může být opět takové řešení $\hat{\beta}$ stále obtížné najít.

Označme $\hat{\beta}_W$ jako přípustné řešení $U_w(\beta) = 0$. Náhodný vektor $\left[n^{\frac{1}{2}} \left(\hat{\beta}_W - \beta \right) \right]$, kde β představuje skutečný vektor koeficientů, je poté asymptoticky normálně rozdělen, s nulovou střední hodnotou a kovarianční maticí $A_W^{-1} B_W A_W^{-1}$, kde

$$A_W = n^{-1} \sum_{i=1}^{n} \int_0^\tau \left[W(u) \left[\mathbf{Z}_i - E^*(u, \beta) \right]^{\otimes 2} \left(\frac{h_0'(u)u}{h_0(u)} + 1 \right) dN_i^*(u) \right]$$
$$B_W = n^{-1} \sum_{i=1}^{n} \int_0^\tau \left[W(u)^2 \left[\mathbf{Z}_i - E^*(u, \beta) \right]^{\otimes 2} dN_i^*(u) \right] \tag{3.4}$$

Matice A_W opět obsahuje neznámou derivaci základní rizikové funkce a její vypočítání opět představuje problém. Na možnosti jak odhadovat kovarianční matici bez nutnosti numerického vyjádření derivace h_0 se zaměříme na konci kapitoly.

3.1.1 Gehanův odhad

Výhodnou volbou pro náhradu funkce $W(u)$ je tzv. Gehanova váha [23], která nabývá tvaru

$$G(u) = n^{-1} S_0^*(u, \beta).$$

Po dosazení do skórové funkce 3.3 dostaneme

$$
\begin{aligned}
U_G(\beta) &= \sum_{i=1}^{n} \left[\int_0^{\tau} n^{-1} S_0^*(u,\beta) \left(\mathbf{Z}_i - E^*(u,\beta) \right) dN_i^*(u) \right] \\
&= n^{-1} \sum_{i=1}^{n} \left[\int_0^{\tau} \left(S_0^*(u,\beta)\mathbf{Z}_i - S_1^*(u,\beta) \right) dN_i^*(u) \right] \\
&= n^{-1} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \left[\delta_i \left(\mathbf{Z}_i - \mathbf{Z}_j \right) I(e_i(\beta) \leq e_j(\beta)) \right] \right],
\end{aligned} \tag{3.5}
$$

kde jsme použili značení $e_i(\beta) = \log(T_i) - \mathbf{Z}_i^T \beta$. Výsledná funkce $U_G(\beta)$ je monotonní v β [24] a zároveň se jedná o p-dimenzionální gradient ztrátové funkce

$$
L_G(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{ e_i(\beta) - e_j(\beta) \}^-, \tag{3.6}
$$

kde $a^- = I(a < 0)|a|$. Jelikož víme, že funkce $U_G(\beta)$ je monotónní a jedná se o gradient $L_G(\beta)$, tak díky vlastnostem konvexních funkcí můžeme s jistotou říct, že $L_G(\beta)$ je po částech (kvůli nespojitosti $U_G(\beta)$) konvexní funkce. Jako Gehanův odhad poté označujeme takové řešení $\hat{\beta}_G$, ve kterém ztrátová funkce $L_G(\beta)$ nabývá globálního minima. Jelikož je $L_G(\beta)$ pouze po částech konvexní funkce, tak nemusí být $\hat{\beta}_G$ unikátní, všechny možné řešení jsou ovšem asymptoticky ekvivalentní [19]. Zároveň, kvůli nediferencovatelnosti $L_G(\beta)$, nemůžeme pro hledání $\hat{\beta}_G$ přímo použít gradientní metody.

Spolehlivé řešení jak najít $\hat{\beta}_G$ navrhuje Jin v [19]. Problém hledání globálního minima se dá přeformulovat na úlohu lineárního programovaní, konkrétně

$$
\begin{aligned}
&\underset{u,\beta}{\text{minimalizuj}} && \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i u_{ij} \\
&\text{za podmínek:} && u_{ij} = (e_j(\beta) - e_i(\beta)), \\
& && u_{ij} \leq 0, \\
& && \text{pro } i, j = 1, \ldots, n.
\end{aligned} \tag{3.7}
$$

Výhodou tohoto přístupu je jednoduchá implementace a záruka nalezení přesného řešení po konečně mnoho iterací (při použití simplexové metody). Ovšem s rostoucí dimenzí dat roste dramaticky i výpočetní náročnost. Problém obsahuje $\mathcal{O}(n^2 + p)$ parametrů a n^2 omezujících podmínek. Pro vysoko-dimenzionální data je tento přístup nevhodný.

Minimalizace $L_G(\beta)$ je také ekvivalentní minimalizaci

$$
\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i |e_i(\beta) - e_j(\beta)| + |M - \beta^T \sum_{k=1}^{n} \sum_{l=1}^{n} \delta_k (\mathbf{Z}_k - \mathbf{Z}_l)|, \tag{3.8}
$$

kde M je předem specifikované velmi velké číslo (M musí být minimálně větší než $\sum_{k=1}^{n} \sum_{l=1}^{n} \delta_k (\mathbf{Z}_k - \mathbf{Z}_l)$) [19]. Minimalizace funkce 3.8 je tudíž problém nejmenších absolutních odchylek a dá se řešit pomocí mediánové regrese. Můžeme například použít algoritmus, představený Koenkerem & D'Oreym v [25], který vhodně modifikuje simplexový algoritmus a dosahuje mnohem rychlejších výsledků než metoda lineárního programování pro malé a středně velké datasety [26] (v R je implementován v balíčku **quantreg**). Pro vysoko-dimenzionální data je ale minimalizace pomocí mediánové regrese opět výpočetně velmi náročná.

3.1.2 Polynomiálně vyhlazený Gehanův odhad

Další možností jak můžeme nalézt regresní koeficienty $\hat{\beta}_G$ je pomocí aproximace $L_G(\beta)$ nějakou spojitou konvexní funkcí a následného použití výpočetně efektivních gradientních metod. V této sekci se zaměříme na aproximaci $L_G(\beta)$ přímo pomocí polynomiálně vyhlazující funkce [27],[28] [29]. Definujme vyhlazenou ztrátovou funkci $L_G(\beta)$ jako

$$L_{G\epsilon} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i c_\epsilon \left[e_i(\beta) - e_j(\beta) \right], \tag{3.9}$$

kde

$$c_\epsilon(x) = \begin{cases} -x, & \text{pokud } x \leq -\epsilon, \\ -\frac{1}{16\epsilon^3}(\epsilon - x)^4 + \frac{1}{4\epsilon^2}(\epsilon - x)^3, & \text{pokud } x \in (-\epsilon, \epsilon], \\ 0, & \text{pokud } x > \epsilon, \end{cases} \tag{3.10}$$

Pro nějaké velmi malé ϵ, v této práci volíme $\epsilon = 10^{-6}$. Obecně může ovšem být definice c_ϵ i jiná [27]. Z definice 3.10 je zřejmé, že $\lim_{\epsilon \to 0} L_{G\epsilon}(\beta) = L_G(\beta)$ a tedy pro $\epsilon = 0$ jsou obě ztrátové funkce totožné. Vyhlazování probíhá pouze na intervalu $(-\epsilon, \epsilon]$. Jako řešení poté označujeme takové $\hat{\beta}_{G\epsilon}$ ve kterém nabývá funkce $L_{G\epsilon}(\beta)$ globálního minima.

Gradient $L_{G\epsilon}(\beta)$ podle β můžeme vyjádřit jako

$$\begin{aligned} \frac{\partial}{\partial \beta} L_{G\epsilon} = U_{G\epsilon} &= \frac{\partial}{\partial \beta} \left[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i c_\epsilon \left[e_i(\beta) - e_j(\beta) \right] \right] \\ &= \left[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \frac{\partial}{\partial \beta} c_\epsilon \left[\log T_i - \mathbf{Z}_i^T \beta - \log T_j + \mathbf{Z}_j^T \beta \right] \right] \\ &= \left[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \left(\mathbf{Z}_i - \mathbf{Z}_j \right) k_\epsilon \left[e_i(\beta) - e_j(\beta) \right] \right], \end{aligned} \tag{3.11}$$

kde $k_\epsilon = -c'_\epsilon$

$$k_\epsilon(x) = \begin{cases} 1, & \text{pokud } x \leq -\epsilon, \\ -\frac{1}{4\epsilon^3}(\epsilon - x)^3 + \frac{3}{4\epsilon^2}(\epsilon - x)^2, & \text{pokud } x \in (-\epsilon, \epsilon], \\ 0, & \text{pokud } x > \epsilon, \end{cases} \tag{3.12}$$

Z vyjádření 3.12 je opět vidět, že $\lim_{\epsilon \to 0} U_{G\epsilon}(\beta) = U_G(\beta)$.

Jak už bylo zmíněno, tak velkou výhodou tohoto přístupu je možnost využití gradientních metod. Jak funkce $L_{G\epsilon}(\beta)$ tak její derivace podle β je spojitá a konvexní pro jakékoliv $\epsilon > 0$ [27]. Pro efektivní minimalizaci funkce $L_{G\epsilon}(\beta)$ je výhodné využít Broyden–Fletcher–Goldfarb–Shanno (BFGS) algoritmus, který patří do rodiny tzv. kvazi-newtonovských metod. Přesnou formulaci BFGS algoritmu v této práci nebudeme ukazovat a odkážeme se na literaturu [28],[30]. Důležitý je pro nás fakt, že BFGS algoritmus na rozdíl od Newtonovy metody nevyžaduje ke hledání kroku sestupu Hessovu matici, jejíž numerický výpočet může být pro funkci $LG\epsilon(\beta)$ velmi náročný.

Při aproximaci nějaké funkce pomocí jiné funkce, s požadovanými vlastnostmi, nás obecně zajímá chyba aproximace. Tu můžeme spočítat následovně

$$
\begin{aligned}
|L_{G\epsilon} - L_G| &= |L_{G\epsilon} - L_{G0}| \\
&= n^{-1} \left| \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i c_\epsilon \left[e_i(\beta) - e_j(\beta) \right] \right] - \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i c_0 \left[e_i(\beta) - e_j(\beta) \right] \right] \right| \\
&\leq n^{-1} \left| \left[\sum_{i=1}^{n} \sum_{j=1}^{n} c_\epsilon \left[e_i(\beta) - e_j(\beta) \right] - \ c_0 \left[e_i(\beta) - e_j(\beta) \right] \right] \right| \\
&\leq n^{-1} \left| \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{max}(c_\epsilon \left[e_i(\beta) - e_j(\beta) \right] - c_0 \left[e_i(\beta) - e_j(\beta) \right]) \right] \right| \\
&= n^{-1} \left| \left[\sum_{i=1}^{n} \sum_{j=1}^{n} (c_\epsilon \left[0 \right] - c_0 \left[0 \right]) \right] \right| \\
&= -\frac{1\epsilon}{16} + \frac{1\epsilon}{4} \\
&= \frac{3\epsilon}{16}
\end{aligned} \tag{3.13}
$$

Průměrná odchylka od $L_G(\beta)$ je tedy $\frac{3\epsilon}{16}$ pro jakékoliv možné β.

Jako počáteční β je možné zvolit např. $\hat{\beta}_{0p} = (0_0, 0_1, \ldots, 0_p)$ či námi použitý odhad $\hat{\beta}_{OLS}$ pomocí metody nejmenších čtverců na celém datasetu.

3.1.3 Hellerův odhad

Další možnost jak vyhladit Gehanovu odhadující funkci $U_G(\beta)$ navrhuje Heller v [31]. Základní myšlenka Hellerova odhadu spočívá v nahrazení indikátorové funkce $I(e_i(\beta) \leq e_j(\beta))$ funkcí $D\left(\frac{e_i(\beta)-e_j(\beta)}{a}\right) = 1 - \phi\left(\frac{e_i(\beta)-e_j(\beta)}{a}\right)$, kde a je bandwidth (škálovací parametr) a ϕ nějaká lokální distribuční funkce. Nespojitost původní funkce $U_G(\beta)$ v β pramení právě v použité indikátorové funkci [31], její aproximací bychom potom měli dostat hladkou funkci na kterou můžeme opět aplikovat gradientní metody. Výsledná funkce nabývá tvaru

$$
U_H(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\delta_i (\mathbf{Z}_i - \mathbf{Z}_j) D\left(\frac{e_i - e_j}{a} \right) \right] \tag{3.14}
$$

Funkce $U_H(\beta)$ je poté spojitá, monotónní a pokud zároveň platí, že ϕ je standardní normální distribuční funkce, tak se jedná o gradient ztrátové konvexní funkce

$$
L_H(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\delta_i (e_j(\beta) - e_i(\beta)) \phi\left(\frac{e_i(\beta) - e_j(\beta)}{a} \right) + a\phi\left(\frac{e_i(\beta) - e_j(\beta)}{a} \right) \right] .[27] \tag{3.15}
$$

Obtíž v použití Hellerova odhadu tkví v nalezení vhodné funkce ϕ a škálovacího parametru a.

Parametr a by měl konvergovat k 0 při zvětšujícím se vzorku pozorováních [31]. Při dostatečně velkém n platí, že pokud $e_i(\beta) \leq e_j(\beta)$ tak $D\left(\frac{e_i(\beta)-e_j(\beta)}{a}\right) \to 1$. Naopak pokud $e_i(\beta) > e_j(\beta)$ tak $D\left(\frac{e_i(\beta)-e_j(\beta)}{a}\right) \to 0$. Tedy čím více máme pozorování, tím můžeme s větší jistotou říci, že můžeme nahradit indikátorovou funkci funkcí D bez ztráty asymptotických vlastností U_G. Škálovací parametr by měl být volen tak aby splňoval podmínky sepsané Hellerem v [31]. Jako praktická volba se jeví použití $a = \sigma n^{-0.26}$, kde σ je standardní odchylka necenzorovaných

reziduí $e_i(\beta)$, $i \in 1, 2, \ldots, n$ pro počáteční hodnotu β. Jsou ovšem možné i jiné volby jako například $\sigma n^{-\frac{1}{5}}$ či přímo najít vhodný bandwidth pomocí metod strojového učení, konkrétně křížové validace která se používá například v [32].

Lokální distribuční funkce ϕ musí být spojitá a splňovat podmínky sepsané Hellerem v [31]. Jako nejvýhodnější volba se jeví použití standardní normální distribuční funkce, jelikož můžeme jednoduše díky integraci per partes vyjádřit funkci $L_H(\beta)$ z $U_H(\beta)$. Jako řešení poté označujeme $\hat{\beta}_H$ ve kterém funkce L_H nabývá globálního minima.

Pro hledání $\hat{\beta}_H$ opět doporučujeme použít BFGS algoritmus. Pro větší datasety je také možné vyzkoušet jeho variantu L-BFGS, která používá pouze limitovanou část počítačové paměti, více se lze dozvědět například v [28].

Jako počáteční odhad β navrhuje Heller použít $\hat{\beta}_G$. V této práci ale ovšem používáme jako počáteční odhad $\hat{\beta}_{OLS}$, jelikož dosahoval v testování mnohem rychlejších výsledků s minimální odchylkou od použití $\hat{\beta}_G$ jako počátečního odhadu.

Nutnost zvolení škálovacího parametru a a distribuční funkce ϕ dělá z Hellerova odhadu složitější alternativu pro poměrně přímočarý polynomiálně vyhlazený Gehanův odhad. Jeho hlavní síla ovšem spočívá v odhadu kovarianční matice $A_W^{-1} B_W A_W^{-1}$. U polynomiálního odhadu musíme aktualizovat $A_W^{-1} B_W A_W^{-1}$ při každé iteraci. Heller poskytuje pomocí metod U-statistiky způsob jak kovarianční matici vypočítat přímo s již nalezeným $\hat{\beta}_H$. Tento způsob si představíme na konci kapitoly.

3.1.4 Další možnosti odhadů pomocí pořadových metod

V této části práce pracujeme pouze s vybranými funkcemi založenými na pořadových metodách. V práci je z časového důvodu mimo jiné vynechán efektivní Zengův a Linův jádrový odhad představený v [33] a Brownova a Wangova verze vyhlazeného Gehanova odhadu z [34].

Zároveň zde není uveden žádný odhad založený na nějaké jiné než Gehanově váze. To je zapříčiněno hned z několika důvodů. Gehanova váha na rozdíl od jiných transformuje funkci $U_W(\beta)$ na monotonní a poskytuje tedy dobré podmínky pro optimalizační algoritmy. Funkce $U_W(\beta)$ založené na jiné váze monotonnost negarantují a hledání nějakého lokálního minima je tedy velmi složitý problém (vzhledem k nespojitosti v β) [35]. Může se ovšem stát, že použití jiné než Gehanovy váhy bude vystihovat data lépe a bylo by jí tedy vhodnější použít. Jin navrhuje v [19] iterativní algoritmus, který využívá vhodné modifikace $U_W(\beta)$ a počátečního řešení $\hat{\beta}_G$ pro použití pro jakoukoliv obecnou váhu splňující Yingovy podmínky z [22]. Výsledný odhad poté můžeme opět získat pomocí metod lineárního programování. Další způsob odhadu regresních parametrů pro nějakou obecnou váhu je navržen v článku [35].

3.1.5 Odhad rozptylu u pořadových metod

Výpočet kovarianční matice $A_W^{-1} B_W A_W^{-1}$ je numericky náročný úkol, jelikož matice A_W^{-1} obsahuje neznámou derivaci základní rizikové funkce.

Možným způsob jak aproximovat rozdělení nalezeného odhadu $\hat{\beta}$ navrhuje Jin v [19]. Pracujeme nyní s řešením $\hat{\beta}_G$, které minimalizuje ztrátovou funkci $L_G(\beta)$. Pro odhad kovarianční matice náhodného vektoru $n^{0.5}\left(\hat{\beta}_G - \beta\right)$ kde β označuje skutečné regresní koeficienty můžeme využít resamplingové metody.

Definujme novou Gehanovu ztrátovou funkci

$$L_G^*(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} R_i R_j \delta_i |e_i(\beta) - e_j(\beta)| + |M - \beta^T \sum_{k=1}^{n}\sum_{l=1}^{n} R_k R_l \delta_k (\mathbf{Z}_k - \mathbf{Z}_l)|, \tag{3.16}$$

kde R_i, $i \in 1, 2, \ldots, n$ jsou i.i.d. pozitivní náhodné veličiny splňující $\mathrm{E}(R_i) = \mathrm{var}(R_i) = 1$. Náhodné veličiny R_i můžeme generovat ze standardního exponencionálního rozdělení [19] [36].

Funkce $L_G^*(\beta)$ je tedy vlastně jenom perturbovaná funkce $L_G(\beta)$ o náhodné veličiny R_i. Označme $\hat{\beta}^*$ jako řešení, které minimalizuje funkci $L_G^*(\beta)$. Pokud budeme opětovně generovat N realizací náhodných veličin R, dostaneme N hodnot $\hat{\beta}^*$. Kovarianční matici poté můžeme odhadnout standardně pomocí výběrového rozptylu

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\beta}_i^* - \overline{\beta}\right)^{\otimes 2}, \tag{3.17}$$

kde $\overline{\beta} = N^{-1}\sum_{i=1}^{N}\hat{\beta}_i^*$.

Odhad rozdělení pomocí resamplingové metody je používaná statistická technika [37],[38] a dá se rozšířit i pro Hellerův a polynomiálně vyhlazený odhad. Jeho nevýhoda je vysoká výpočetní náročnost, která může být pro větší datasety až nerozumná.

Jednou z motivací Hellerova odhadu bylo nalezení lepšího způsobu odhadu kovarianční matice $A_W^{-1}B_W A_W^{-1}$. Označme $\hat{\beta}_H$ jako řešení, které minimalizuje ztrátovou funkci $L_H(\beta)$. Kovarianční matici poté můžeme odhadnout následovně

$$\hat{A} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i a^{-1}\left(\mathbf{Z}_i - \mathbf{Z}_j\right)^{\otimes 2}\phi\left(\frac{e_i(\hat{\beta}_H) - e_j(\hat{\beta}_H)}{a}\right)$$

$$\hat{B} = n^{-2}\sum_{i}\sum_{j}\sum_{k\neq j}(\mathbf{Z}_i - \mathbf{Z}_j)(\mathbf{Z}_i - \mathbf{Z}_k)(e_{ij}(\hat{\beta}_H) - e_{ji}(\hat{\beta}_H))(e_{ik}(\hat{\beta}_H) - e_{ki}(\hat{\beta}_H))$$

Kde $e_{ij}(\hat{\beta}_H) = \delta_i\left[1 - \phi\left(\frac{e_i(\hat{\beta}_H) - e_j(\hat{\beta}_H)}{a}\right)\right]$ a $\hat{A}$ je matice druhých derivací původní skórové funkce L_H.

Hellerův odhad tedy značně usnadňuje výpočet distribuce nalezeného řešení a nevyžaduje takovou výpočetní náročnost jako u resamplingové metody.

Odhad kovarianční matice pro Gehanův polynomiálně vyhlazený odhad v této práci neřešíme. Jak už bylo zmíněno v předchozích odstavcích, tak bychom opět mohli perturbovat ztrátovou funkci a získat odhad pomocí resamplingu.

3.2 Odhady založené na metodě nejmenších čtverců

Druhým způsobem jak přistupovat k hledání regresních parametů β je pomocí rozšíření metody nejmenších čtverců pro cenzorovaná data. Metoda nejmenších čtverců je standardní nástroj pro datovou a regresní analýzu a její použití či rozšíření pro cenzorovaná data se jeví jako přirozený nápad. Poprvé tento způsob zkoumal Miller v [17] a následně ho rozšířil a zpopularizoval Buckley v [16]. Miller následně v [39] ukázal, že Buckleyho způsob je spolehlivější v praxi a z Buckleyho metody se poté stal klasický způsob pro odhad regresních parametrů pomocí metody nejmenších čtverců pro cenzorovaná data [36],[40]. Ritov poté v [41] poskytl asymptotickou teorii pro Buckleyho odhad a dokázal, že odhady založené na pořadových metodách a Buckleyho odhad jsou asymptoticky ekvivalentní.

3.2.1 Buckleyho odhad

U klasické metody nejmenších čtverců se snažíme minimalizovat funkci

$$n^{-1}\sum_{i=0}^{n}\left[T_i - \alpha - \mathbf{Z}_i^T\beta\right]^2, \tag{3.18}$$

kde α představuje intercept. Připomeňme, že v této práci intercept neuvažujeme, jelikož ho neumíme spolehlivě odhadovat [18]. Řešení funkce 3.18 poté pro β poté nabývá tvaru

$$\sum_{i=0}^{n} \left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\left(T_i - \mathbf{Z}_i^T\beta\right) = 0, \tag{3.19}$$

kde $\overline{\mathbf{Z}} = n^{-1}\sum_{i=0}^{n}\mathbf{Z}_i$. Rovnici 3.19 ovšem nemůžeme zatím přímo použít k odhadu regresních koeficientů, jelikož cenzorované hodnoty T_i (tedy hodnoty kde $\delta_i = 0$) jsou neúplné a odhad by tak dával zkreslené výsledky. Buckley a James navrhují nahradit pozorování T jinou náhodnou veličinou, se stejnou střední hodnotou a s možností pojmutí cenzorovaných dat [16],[2]. Konkrétně můžeme nahradit každé T_i hodnotami $\hat{T}_i$

$$\hat{T}_i(\beta) = \delta_i T_i + (\delta_i - 1)\,\mathrm{E}(T_i|T_i > t_i). \tag{3.20}$$

Hodnota podmíněné pravděpodobnosti $\mathrm{E}(T_i|T_i > t_i)$ je ovšem stále neznámá a můžeme jí aproximovat pomocí

$$\frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}_\beta(u)}{1 - \hat{F}_\beta(e_i(\beta))} + \mathbf{Z}_i\beta, \tag{3.21}$$

kde $\hat{F}$ reprezentuje Kaplan-Meierův odhad (viz definice 1.12) distribuční funkce reziduí. Jako výsledný odhad regresních koeficientů β poté označujeme $\hat{\beta}_{BJ}$ které je kořenem skórové funkce $U_{BJ}(\beta, b)$

$$U_{BJ}(\beta, b) = \sum_{i=1}^{n}(\mathbf{Z}_i - \bar{\mathbf{Z}})\{\hat{T}(b) - \mathbf{Z}_i^T\beta\} \tag{3.22}$$

Funkce $U_{BJ}(\beta, \beta)$ je nespojitá a nemonotónní v β [36], nalezení řešení $\hat{\beta}_{BJ}$ je tedy opět poměrně komplikované a iterace nemusí v mnoho případech zkonvergovat či může oscilovat mezi několika možnými hodnotami [16].

Pro použití Buckleyho metody musí data splňovat dva základní předpoklady. Model musí být lineární v koeficientech a data musí být homoskedatická. V praxi je ovšem téměř nemožné tyto předpoklady nějak ověřit [42].

Jako počáteční odhad β navrhuje Buckley v [16] zvolit $\hat{\beta}_{OLS}$.

▶ Poznámka 3.1. Pro numerickou efektivitu doporučuje Buckley předefinovat největší reziduál $e_i(\beta)$ jako necenzorovaný, více viz literatura [43]

3.2.2 Jinovo rozšíření Buckleyho odhadu

Linearizace skórové funkce $U_{BJ}(\beta, b)$ pro nějaký počáteční odhad $\hat{\beta}$ a následné vyřešení $U_{BJ}(\beta, \hat{\beta})$ vede k iterativnímu algoritmu, kde postupně aktualizujeme hodnotu β dokud nepřekročíme maximální počet iterací, či není splněna nějaká podmínka konvergence. Tedy

$$\hat{\beta}_k = L_{BJ}(\hat{\beta}_{k-1})\ k \in 1, 2, \ldots, m$$

Pro

$$L_{BJ}(\beta) = \left\{\left[\sum_{i=1}^{n}\mathbf{Z}_i - \overline{\mathbf{Z}}\right]^{\otimes 2}\right\}^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\left(\hat{T}_i - \overline{T}\right)\right], \tag{3.23}$$

kde $\overline{T} = n^{-1}\sum_{i=1}^{n}\hat{T}_i$.

Jin v [36] dokázal, že pokud se jako počáteční $\hat{\beta}$ zvolí konzistentní odhad β (tedy konzistentní odhad pravých regresních koeficientů), tak poté pro jakékoliv fixní k bude $\hat{\beta}_k$ také konzistentní

a asymptoticky normální. Jako počáteční odhad tedy můžeme zvolit $\hat{\beta}_G$. Výsledné $\hat{\beta}_k$ je poté lineární kombinací odhadu $\hat{\beta}_G$ a $\hat{\beta}_{BJ}$ [36]. Konkrétně platí

$$\hat{\beta}_k = (I - D^{-1}A)^k \hat{\beta}_G + \{I - (I - D^{-1}A)^k\}\hat{\beta}_{BJ} + o_p\left(n^{-0.5}\right), \tag{3.24}$$

kde $D = n^{-1}\sum_{i=0}^{n}\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)^{\otimes 2}$, I je jednotková matice a A představuje matici koeficientů z Buckleyho odhadu a její definici můžeme nalézt v literatuře v [36].

Pokud pracujeme s necenzorovanými daty, tak řešení $\hat{\beta}_k$ konverguje ke klasickému odhadu pomocí obyčejné lineární regrese. Dále pokud iterativní algoritmus z vyjádření 3.23 zkonverguje v nějakých k krocích, tak je výsledné řešení $\hat{\beta}_k$ přímo rovné odhadu $\hat{\beta}_{BJ}$. Pokud algoritmus nezkonverguje nebo osciluje mezi několika možnými hodnotami, tak je výsledné řešení stále konzistentní a asymptoticky normální [36].

Volba konvergenční podmínky může být různá. V této práci používáme (stejná podmínka je použita i pro Buckleyho odhad)

$$\frac{||\hat{\beta}_k - \hat{\beta}_{k-1}||}{||\hat{\beta}_k|| + 10^{-6}} < \epsilon \tag{3.25}$$

Pro nějaké zvolené $\epsilon > 0$ (v práci používáme $\epsilon = 10^{-5}$). K normě vektoru $\hat{\beta}_k$ ve jmenovateli zlomu 3.25 přičítáme číslo 10^{-6} abychom předešli dělení nulou.

Jinovo rozšíření tedy primárně spočívá ve zvolení vhodného počátečního odhadu regresních koeficientů. Ačkoliv Jin v článku [36] nijak neřeší problém předpokladu homoskedaticity dat, tak je jeho odhad konzistentní i v případě heteroskedatických dat. Důkaz tohoto tvrzení lze najít v apendixu v článku [32].

3.2.3 Další možnosti odhadů pomocí metod založených na metodě nejmenších čtverců

Zásádním nedostatkem Buckleyho odhadu je omezení pouze na homoskedatická data a nekonzistentní odhad rozptylu vektoru $n^{-1}(\hat{\beta}_{BJ} - \beta)$ [44]. Tyto problémy řeší již zmíněný Jinův odhad, který kovarianční matici odhaduje opět pomocí resamplingové metody.

Další způsob odhadu ukazuje například Yu v [32]. Jeho modifikace Buckleyho odhadu využívá jádrové odhady a postupně aktualizuje odhad rozptylu pomocí lokální polynomiální regrese. Zároveň dokazuje, že je jeho odhad vhodný jak pro homoskedatická tak pro heteroskedatická data.

Za zmínku také stojí lokální Buckleyho odhad představený v [45]. Tato modifikace Buckleyho odhadu se opět zaměřuje na scénář, kde pracujeme s heteroskedatickými daty a autoři dokazují, že jejich odhad pro tato data funguje mnohem efektivněji než ostatní odhady založené na metodě nejmenších čtverců.

Nakonec zmiňme ještě fakt, že nemusíme modifikovat pouze metodu nejmenších čtverců a místo toho můžeme použít tzv. M-odhady. L2 ztrátovou funkci v 3.18 můžeme nahradit nějakou jinou obecnou ztrátovou funkcí [36],[41]. Poté můžeme vytvořit podobný iterativní algoritmus jako Buckleyho a odhadnout kovarianční matici pomocí resamplingových metod. Všechna asymptotická tvrzení o Buckleyho odhadu poté platí i pro nově vytvořenou modifikaci [36].

3.2.4 Odhad rozptylu u Buckleyho odhadu

Možné způsoby odhadu kovarianční matice vektoru $n^{0.5}\left(\hat{\beta}_k - \beta\right)$ jsou popsány v [44]. V této sekci se zaměříme na odhad pomocí resamplingové metody představený Jinem v [36].

Mějme opět i.i.d. náhodné veličiny R_i, pro které platí $\mathrm{E}(R_i) = \mathrm{var}(R_i) = 1$. Definujme nové perturbované funkce

$$\hat{T}_i^*(\beta) = \delta_i T_i + (\delta_i - 1)\left[\frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}_\beta^*(u)}{1 - \hat{F}_\beta^*(e_i(\beta))} + \mathbf{Z}_i\beta\right],$$

$$\hat{F}_\beta^*(e_i(\beta)) = 1 - \prod_{e_i(\beta)\leq t}\left[1 - \frac{R_i \mathbf{dN}(e_i(\beta))}{\sum_{j=1}^n R_j Y(e_i(\beta))}\right],$$

$$L_{BJ}^*(\beta) = \left\{\left[\sum_{i=1}^n R_i\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\right]^{\otimes 2}\right\}^{-1}\left[\sum_{i=1}^n R_i\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\left(\hat{T}_i^* - \overline{T}^*\right)\right].$$

Poté definujeme iterativní proces

$$\hat{\beta}_k^* = L_{BJ}(\hat{\beta}_{k-1}^*)\ k > 0, \tag{3.26}$$

kde jako počáteční odhad $(\hat{\beta}_0^*)$ bereme řešení minimalizující funkci 3.16. Je tedy potřeba nejdříve vygenerovat N realizací odhadnutých regresních koeficientů pro Gehanovu perturbovanou funkci a poté každý odhadnutý regresní koeficient použít jako počáteční řešení pro našich N iterativních procesů 3.26. Poté opět dostaneme N odhadnutých řešení $\hat{\beta}_k^*$ a kovariančí matici můžeme následovně odhadnout stejně jako v předchozím vyjádření zde 3.17 pomocí výběrového rozptylu.

Kapitola 4
Sestavený balíček

V této kapitole stručně popíšeme praktický výstup práce, tedy balíček pro programovací jazyk R, který je v dnešní době standardní nástroj pro statistické výpočty a analýzu dat.

Balíček, dále označovaný jménem **aftsem**, poskytuje implementaci vybraných možností odhadů regresních parametrů (a jejich rozptylu) z předchozí kapitoly. Celková implementace je kompletně zdokumentována pomocí nástroje roxygen [46] a umožňuje jednoduchou integraci dalších odhadů. Balíček **aftsem** se tedy může stát vhodným nástrojem pro analýzu cenzorovaných dat pomocí semiparametrického modelu zrychleného času.

4.1 Aftsem

Rozhraní balíčku je koncipováno stejně jako u balíčku pro výpočet standardní lineární regrese (součást základní verze R) či u balíčku **survival** [47], který implementuje například Coxův model proporcionálního rizika. Volání programu probíhá pomocí hlavní funkce `aftsem`, která vypadá následovně

```
aftsem(
  formula,
  data,
  control = aftsem.control(),
  method = "buckley",
  binit = "auto",
  ties = NULL,
  na.action = na.omit,
  subset = NULL,
  resample = 0,
  ...
)
```

Výpis kódu 4.1 Funkce aftsem

Povinné argumenty jsou `formula` a `data`. Argument `formula` specifikuje model se kterým chceme pracovat a má klasický tvar typu `response ~ variables`, který je dobře známý uživatelům R. Vysvětlovaná proměnná, tedy `response` v argumentu `formula`, musí být typu `Surv`. Datový objekt `Surv` se sestává ze dvou hodnot, kde první reprezentuje logaritmované časy přežití a druhá vektor cenzorování, argument `formula` tedy vypadá například takto: `formula = Surv(log(data$time),data$censor) ~ data$age`. Proměnné `variables` nemusí být čistě numerické, balíček je v případě faktorové proměnné sám převede na číselnou reprezentaci pomocí binárních indikátorů.

Argument `data` specifikuje dataframe, nad kterým funkci voláme. Argument může být vynechán pokud používáme ve volání `formula` operátor `$`.

Argument `method` specifikuje jaká metoda bude použita pro odhad regresních parametrů. Jako výchozí je zvolena klasická neupravená Buckleyho metoda. Na výběr jsou dále k dispozici metody (`"jin"`,`"gehan"`,`"gehan-poly"`,`"gehan-heller"`). Všechny metody jsou poté popsány v předchozí kapitole. Metody `"buckley"` a `"jin"` jsou kompletně napsány v jazyku c++ pomocí knihovny **armadillo** [48]. Knihovna **armadillo** je určená pro rychlé vědecké výpočty a lineární algebru a disponuje spousty připravenými statistickými funkcemi. Knihovna byla zvolena za účelem zrychlení samotného R kódu. Do jazyku R jsou poté c++ kódy integrovány přes prostředí Rcpp [49]. Metoda `"gehan"` je implementována pomocí medianové regrese (viz 3.8) s balíčkem **quantreg**. Přístup pomocí mediánové regrese byl již dříve implementován Jinem v [50] a Johnsonem v [51], implementace v balíčku **aftsem** je založená na stejném přístupu a příliš se neodlišuje od Jinovy implementace. Metody `"gehan-poly"` a `"gehan-heller"` jsou implementovány pomocí baličku **optimx** [52], který poskytuje jednoduché a efektivní prostředí pro optimalizaci bez omezujících podmínek. Definice samotných funkcí (Hellerova funkce a polynomiálně vyhlazená Gehanova funkce) je opět napsána pomocí knihovny **armadillo** v jazyce c++ a samotné volání knihovny optimx poté probíhá v jazyce R. Toto propojení c++ a R poskytuje několika násobné zrychlení celé minimalizace oproti kódu psaném v čistém R.

Argument `binit` představuje počáteční odhad regresních koeficientů β. Jeho výchozí hodnota je nastavena na `"auto"`, která inicializuje počáteční odhad pro jednotlivé metody stejně, jako je popsáno v předchozí kapitole. Uživatel dále může zvolit alternativní hodnoty `binit` a to konkrétně

- `"lm"`: Počáteční hodnota je vypočítána pomocí klasické metody nejmenších čtverců na celém datasetu
- `"gehan"`: Počáteční hodnota je vypočítána pomocí minimalizace Gehanovy ztrátové funkce
- `Vlastní numerický vektor`: Uživatel může přímo vložit numerický vektor dimenze p $\times$ 1

Argument `times` reprezentuje způsob řešení shody v časech přežití. Výchozí hodnota je nastavena na `NULL` a v případě nalezené shody v nějakých časech přežití je vypsáno varování na terminál. Další možností je nastavení argumentu `ties` na `"jitter"`, který perturbuje časy přežití podle následujícího schématu

```
perturbed_times <- survival_times + seq_along(survival_times) * .Machine
$double.eps^0.5
```

Výpis kódu 4.2 Perturbace času přežití

Tedy ke každému času přežití přičte index vynásobený nejmenší hodnotou reprezentovanou datovým typem `double`. Obecně shody v časech přežití nejsou velký problém a doporučujeme nechat argument `times` na výchozí hodnotě. Pokud ovšem budeme pracovat s datasetem s malým množstvím pozorování (např. 50 a méně) a budeme chtít odhadovat regresní parametry pomocí metody `"gehan"`, tak shoda časů přežití může vyvolat nečekanou chybu při vytváření augumentovaných dat pro mediánovou regresi, konkrétně může být designová matice singulární. V tomto případě je vhodné data modifikovat, či použít zmíněnou volbu `"jitter"`.

Argument `na.action` představuje akci, která bude zacházet s chybějícími hodnotami v datasetu. Defaultní hodnota `na.omit` jednoduše chybějící hodnoty vynechá a bude pracovat pouze s úplnými daty. Dále můžeme použít volbu `na.fail`, která vyvolá chybu v případě nalezení chybějících hodnot a shodí výpočet.

Argument `subset` slouží jako přepínač, kterým můžeme vybrat konkrétní data se kterými chceme pracovat. Například nás může zajímat jak vypadají odhady pro data se subjekty starší 50 let. Poté můžeme ve volání funkce `aftsem` použít přepínač `subset = age>50`.

Argument `resample` představuje počet vygenerovaných náhodných veličin R_i. Jinými slovy se jedná o číslo N ve vzorci 3.17. Výchozí hodnota je nastavena na 0, což znamená že odhad

rozptylu pomocí resamplingu nebude probíhat. Odhad pomocí resamplingu je naprogramován pouze pro metody `"gehan"` a `"jin"`.

Poslední argument je list `control`. Výchozí hodnota tohoto argumentu je nastavena na `aftsem.control()`, která vytvoří list ve tvaru

```
1    aftsem.control <- function(eps = 0.00001,
2                               maxiter = 15,
3                               gehan_eps = 10^-6,
4                               optimx.alg = "BFGS",
5                               variance.estimation = FALSE,
6                               quantile.method = "br",
7                               use.grad = FALSE)
```

Výpis kódu 4.3 aftsem.control()

kde

- `eps`: Epsilon pro kritérium konvergence u metod `"jin"` a `"buckley"`
- `maxiter`: Maximální počet iterací u metod `"jin"` a `"buckley"`
- `gehan_eps`: Epsilon pro polynomiálně vyhlazenou Gehanovu funkci
- `optimx.alg`: Algoritmus minimalizace pro metody `"gehan-poly"` a `"gehan-heller"`
- `variance.estimation`: Indikátor, zda chce uživatel odhadnout rozptyl při použití metody `"gehan-heller"`
- `quantile.method`: Metoda optimalizace mediánové regrese pro Gehanovu funkci.
- `use.grad`: Indikátor, zda chce uživatel použít manuálně napsané výpočty gradientů u metod `"gehan-poly"` a `"gehan-heller"`

Jako `optimx.alg` si uživatel může zvolit libovolný algoritmus z nabídky knihovny **optimx**, jako příklad uveďme algoritmy `"BFGS"`, `"L-BFGS"` či `"Nelder-Mead"`. Přepínač `use.grad` indikuje zda k minimalizaci bude použit manuálně napsaný gradient či jeho numerická aproximace. Jako výchozí hodnota je zvolena numerická aproximace, v průběhu testování se totiž zjistilo, že minimalizace pomocí knihovny **optimx** je citlivá na počáteční odhad β. Ovšem použití manuálně napsaného gradientu dosahuje o něco rychlejších výsledků, volba je tedy ponechána uživateli. V případě argumentu `quantile.method` je výchozí hodnota nastavena na algoritmus `"br"`, který koresponduje s algoritmem Koenkera D'Oreyho. Uživatel si alternativně může zvolit minimalizující algoritmus z nabídky knihovny **quantreg**.

Návratovou hodnotou funkce `aftsem` je list obsahující informace o proběhnutém odhadu. K funkci jsou dále napsány podporující funkce `print` a `summary`. Více informací bude ukázáno v použití balíčku na reálných datech.

4.2 Stanford heart transplant data

Jako jednoduchou ukázku uveďme použití balíčku **aftsem** na reálných datech, konkrétně na známem datasetu *Stanford heart transplant data* [53] (všechny následující kódy jsou k dispozici v příloze práce).

Stanfordský transplantační program začal v říjnu roku 1967 a přerušení studie pro tuto analýzu se datuje na únor roku 1980. V průběhu tohoto časového úseku byla poskytnuta transplantace srdce 184 pacientům. Jako sledovanou událost označujeme smrt pacienta po provedené transplantaci a jako čas přežití označujeme dobu přežití pacienta ve dnech od provedené transplantace. Někteří pacienti podstoupili transplantaci srdce vícekrát než jednou, čas přežití u těchto

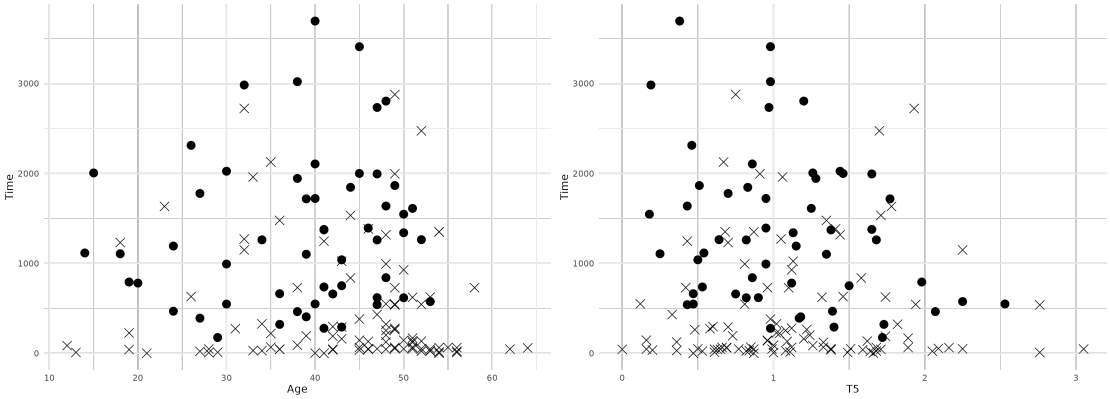
jedinců je poté měřen jako součet času přežití po první transplantaci sečtený s časem přežití po druhé transplantaci. Z datasetu je vynecháno 27 pacientů s chybějícími hodnotami, celkem tedy pracujeme se 157 pozorováními. Ke každému subjektu zaznamenáváme dva příznaky, které byly *apriorně* zvoleny za možné proměnné mající vliv na přežití pacienta. Konkrétně pracujeme s příznaky **age** a **t5**.

Příznak **age** reprezentuje věk pacienta v době provedené transplantace. Jako počáteční odhad můžeme uvažovat, že mladší pacienti mají větší šanci na přežití po provedené transplantaci než staří pacienti [17].

Příznak **t5** označuje T5 skóre vytvořené doktorem Charlesem Bieberem ze Stanfordské univerzity. T5 skóre měří neshodu tkáně mezi dárcem orgánu a pacientem transplantace s ohledem na HL-A antigeny [17],[39]. Hodnota T5 skóre menší než 1 reprezentuje dobrou shodu, naopak hodnota větší než 1 označuje špatnou shodu a může tedy vést k nepřijmutí orgánu imunitním systémem pacienta.

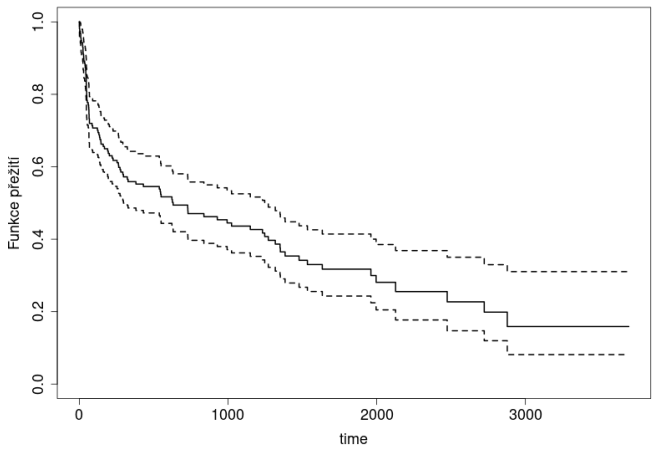
Zajímá nás tedy jak tyto dva kovariáty dobře vystihují čas přežití u jednotlivých pacientů. Nejprve se ovšem zaměřme krátce na samotná data.

- **Průměrný věk**: 41.7 let
- **Nejstarší pacient**: 64 let, zemřel po 60 dnech od provedení transplantace s naměřeným T5 skóre rovným 0.69
- **Nejmladší pacient**: 12 let, zemřel po 86 dnech od operace s naměřeným T5 skórem rovným 1.26
- **Nejmenší naměřené T5 skóre**: Naměřené u 36 letého pacienta s hodnotou 0, pacient zemřel po 44 dnech od provedení transplantace
- **Nejvyšší naměřené T5 skóre**: Naměřené u 53 letého pacienta s hodnotou 3.05, pacient také zemřel po 44 dnech od provedení transplantace
- **Nejmenší naměřený čas přežití**: Jeden den od provedení transplantace, tuto hodnotu pozorujeme u třech pacientů, všichni měli naměřené dobré T5 skóre (≤ 1)
- **Největší naměřený čas přežití**: 3695 dní, pacient má status označený jako cenzorovaný, tedy nevíme přesný čas přežití
- **Celkový počet cenzorování**: 55 pacientů ze 157, tedy cca 35% cenzorování



Obrázek 4.1 Vlevo je vykreslen scatter plot věku na přidružený čas přežití, vpravo je vykreslen plot T5 skóre na čas přežití. Cenzorovaná pozorování jsou označena kroužkem, necenzorovaná naopak křížkem

Neparametrický odhad funkce přežití pomocí Kaplan-Meierova odhadu (viz definice 1.12) z obrázku 4.2 nám říká, že medián doby přežití se nachází lehce pod 1000 dny, tedy něco málo pod tři roky.



Obrázek 4.2 Kaplan-Meierovo odhad funkce přežití s 95% konfidečním intervalem

Zaměřme se nyní na regresní modelování. Definujme tedy náš model jako

$$\log T = \beta_1 \mathbf{age} + \beta_2 \mathbf{t5} + \epsilon. \tag{4.1}$$

Model použijeme následovně

```
fit <- aftsem(Surv(log(stan$time),stan$status) ~ stan$age + stan$t5,
method = "jin", resample = 500)
summary(fit)
```

a dostaneme výsledek

```
    AFT Semiparametric Model Summary
================================

Model Call: aftsem(formula = Surv(log(stan$time), stan$status) ~ stan$age +
 Model Call:     stan$t5, method = "jin", resample = 500)

Used parameter estimation method:
[1] "jin"

Convergence Status: Converged

          |Estimate| |Std. Error| |Z value| |Pr(>|Z|)|
stan$age    -0.03417      0.02237     -1.53       0.13
stan$t5     -0.00641      0.34732     -0.02       0.99
Number of Iterations: 12
Number of Observations: 157
Percent of Censored Observations: 35.030000
```

Modelování pomocí metody `"jin"` zkonvergovalo po 12 iteracích a dostali jsme bodový odhad regresních parametrů $\beta = (-0.03417, -0.00641)$. Jelikož výpočet zkonvergoval, tak víme že Buckleyho odhad bude stejný. Podle hodnot standardní chyby, Z hodnoty a p-hodnoty naměřené pomocí resamplingové metody s $N = 500$ vidíme, že jak proměnná **age** tak proměnná **t5** jsou

statisticky nesignifikantní na standardní pětiprocentní hladině významnosti. Balíček využívá pro výpočet Z-hodnoty Waldův test [54], počítáme

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})},$$

kde SE označuje směrodatnou odchylku.

Pokud bychom se chtěli podívat na počáteční odhad (tedy v tomto případě Gehanův), tak můžeme využít návratovou hodnotu `betafirst`.

```
    print(fit$betafirst)
                [,1]
    [1,] -0.04861196
    [2,] -0.06112750
```

Pozorujeme menší odchylku ve druhém odhadnutém koeficientu od původního Gehanova odhadu.

Další zajímavý objekt je vektor reziduí pro modifikované časy přežití z definice 3.20. Uživatel může tento vektor získat pomocí volání `fit$resid`. Obecně je testování dobré shody modelu s danými daty složitý problém [55],[56]. Menší představu můžeme ovšem zjistit i pomocí standardních regresních metod pro lineární model, například můžeme porovnávat průměr reziduí u jednotlivých skupin rozdělených podle kvartilů kovariát [56].

Pokud bychom chtěli použít metodu `"gehan-heller"` společně s odhadnutím kovarianční matice, tak musíme použít volání s upraveným listem `control`, zároveň také musíme uvést všechny důležité informace pro metodu `"gehan-heller"`, tedy parametry `use.grad` a `optimx.alg` (v případě použití Gehanova odhadu jako počátečního řešení je potřeba specifikovat i argument `quantile.method`)

```
    fit2 <- aftsem(Surv(log(stan$time),stan$status) ~ stan$age + stan$t5,
    method = "gehan-heller", control = list(variance.estimation = TRUE, use.
    grad = FALSE, optimx.alg = "BFGS"))
    summary(fit2)
```

```
AFT Semiparametric Model Summary
================================

Model Call: aftsem(formula = Surv(log(stan$time), stan$status) ~ stan$age +
 Model Call:     stan$t5, control = list(variance.estimation = TRUE, use.
    grad = FALSE,
 Model Call:     optimx.alg = "BFGS"), method = "gehan-heller")

Used parameter estimation method:
[1] "gehan-heller"

Convergence Status: Converged

          |Estimate| |Std. Error| |Z value| |Pr(>|Z|)|
stan$age     -0.0536       0.0255     -2.10      0.035 *
stan$t5      -0.0325       0.3526     -0.09      0.927
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘ ’ 1
Number of Observations: 157
Percent of Censored Observations: 35.030000
```

Výsledek nám vyšel lehce odlišný, vidíme že proměnná **age** už není statisticky nesignifikatní a působí na logaritmovaný čas přežití negativním způsobem s odhadnutým koeficientem -0.0536.

Způsob odhadu rozptylu je odlišný než u resamplingové metody a může se do něj propsat i naše definice ztrátových funkcí, škálovacího parametru a lokální distribuční funkce. Můžeme si také povšimnout lehké odchylky v odhadnutých koeficientech. Porovnání s ostatními metodami můžeme nalézt v tabulce 4.1.

Tabulka 4.1 Odhadnuté koeficienty u jednotlivých metod pro model 4.1

	gehan.poly	gehan.heller	gehan	jin	buckley
age	-0.04865840	-0.05359555	-0.04861196	-0.034165820	-0.034165755
t5	-0.06106086	-0.03247692	-0.06112750	-0.006408728	-0.006408925

Vidíme, že metody `"gehan-poly"` a `"gehan"` dávájí téměř totožné výsledky, to samé platí u metod `"jin"` a `"buckley"`. Metoda `"gehan-heller"` se od ostatních metod více liší v odhadnutém druhém koeficientu, nepřesnost je pravděpodobně zaviněna zvoleným škálovacím parametrem a lokální distribuční funkcí (viz předchozí kapitola).

Podle celkových výsledků vidíme, že T5 skóre nemá na vysvětlovanou proměnnou takový vliv jako věk pacienta. Obě proměnné snižují čas přežití velmi minimalisticky a obě působí na pacienta negativním účinkem. Čas přežití závisí pravděpodobně na jiných proměnných. Výsledek pomocí metody `"gehan-heller"` nám ovšem říká, že proměnná **age** může nést zajímavou informaci. Zkusme tedy ještě postavit model

$$\log T = \beta_1 \mathbf{age} + \beta_2 \mathbf{age}^2 + \epsilon \tag{4.2}$$

Tento model je podobný Millerovu modelu z [39]. Výsledek pomocí metody `"gehan-heller"` nám říká následující

```
fit3 <- aftsem(Surv(log(time),status) ~ age + I(age^2), data = stan,
method = "gehan-heller", binit="gehan", control = list(variance.
estimation = TRUE, use.grad = FALSE, optimx.alg = "BFGS", quantile.
method = "br"))
summary(fit3)
```

```
    AFT Semiparametric Model Summary
================================

Model Call: aftsem(formula = Surv(log(time), status) ~ age + I(age^2), data
    = stan,
 Model Call:     control = list(variance.estimation = TRUE, use.grad = FALSE
    ,
 Model Call:         optimx.alg = "BFGS", quantile.method = "br"), method =
    "gehan-heller", binit = "gehan",
 Model Call:     resample = 200)

Used parameter estimation method:
[1] "gehan-heller"

Convergence Status: Converged

          |Estimate| |Std. Error| |Z value| |Pr(>|Z|)|
age          0.28209      0.12453      2.27     0.0235 *
I(age^2)    -0.00434      0.00156     -2.79     0.0053 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘ ’ 1
Number of Observations: 157
Percent of Censored Observations: 35.030000
```

Obě dvě proměnné jsou na hladině významnosti 5% statisticky významné. Čas přežití se tedy dá dobře modelovat kvadraticky pomocí proměnné **age**. Koeficienty modelu nám nyní říkají že logaritmus času přežití se s jednotkou přirůstajícího věku změní o 0.28209 - 0.00868**age**. Porovnání odhadnutých koeficientů pro jednotlivé metody jsou k dispozici v tabulce 4.2

Tabulka 4.2 Odhadnuté koeficienty u jednotlivých metod pro model 4.2

	gehan.poly	gehan.heller	gehan	jin	buckley
age	0.278937101	0.282086913	0.259299796	0.259093080	0.259059593
age^2	-0.004277525	-0.004336355	-0.003995583	-0.003864943	-0.003864744

V odhadnutých hodnotách tentokrát nepozorujeme žádné odchylky a jedná se téměř o totožná čísla.

4.3 Podobné balíčky

Do dne 4.5.2024 byl autorem této práce nalezen pouze jeden aktivní R balíček implementující semiparametrický model zrychleného času a to konkrétně balíček **aftgee** [57]. Náš balíček **aftsem** může sloužit jako alternativa pro balíček **aftgee**, jelikož poskytuje implementaci jiných metod odhadu regresních parametrů. Zároveň je náš balíček velmi uživatelsky škálovatelný a na rozdíl od balíčku **aftgee** poskytuje jednoduchou kostru programu a umožňuje optimalizaci využívající balíčku **optimx**. Tento krok přináší perfektní prostor pro inovaci a jednoduché rozšíření balíčku o další dosud neimplementované metody a podpůrné funkce. Pokud tedy dojde v následujících letech k dalšímu vývoji odhadu regresních parametrů u semiparametrického modelu zrychleného času, tak může být balíček jednoduše rozšířen a dále použit k porovnání s dalšími metodami. Případný uživatel může navíc využít i programování pomocí jazyku c++ přes rozhraní Rcpp.

Kapitola 5
Simulační studie

V této kapitole provedeme několik experimentů na simulovaných datech. Zajímá nás hlavně přesnost a rychlost jednotlivých odhadů a v jakých situacích je můžeme použít.

Všechny experimenty byly provedeny na jednom zařízení s procesorem AMD Ryzen 3 4300U s frekvencí 2,7 GHz, 4 jádry a operačním systémem Ubuntu 20.04.3 LTS.

Veškeré kódy používané k provádění experimentů jsou poté k dispozici v přílohách práce.

5.1 Simulovaná data

Pro generování časů přežití byl zvolen model zrychleného času inspirovaný literaturou [36],[35],[19]

$$\log T^* = 2 + \mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3 + \epsilon, \tag{5.1}$$

kde $\mathbf{Z}_1$ je Bernouliho náhodná veličina s parametrem 0.5 a $\mathbf{Z}_2$,$\mathbf{Z}_3$ jsou standardně normálně rozdělené náhodné veličiny. Odchylky ϵ jsou generovány ze standardního normálního rozdělení, extrémního (extremálního) rozdělení s parametry $[0, 1, 0]$ (konkrétně se jedná o Gumbelovo rozdělení, definice hustoty je k dispozici k vidění v literatuře [58]) a logistického rozdělení s parametry $[0, 1]$ (hustota je k dispozici k vidění na stránce [59]). Cenzorované časy C jsou generovány z uniformních rozdělení $[0, c]$, kde c bylo nastaveno tak aby reflektovalo požadovanou procentuální míru cenzorování. Jako T (časy přežití) poté označujeme hodnoty $\min(T^*, C)$. Pro experimenty byly zvoleny čtyři procentuální míry cenzorování $\{0, 25, 50, 90\}$. Zvolené míry cenzorování se snaží co nejvíce napodobit možná data přežití se kterými se uživatel může v praxi setkat. Vysoko cenzorovaná data například asociujeme s nějakými vzácnými nemoci, málo cenzorovaná data naopak můžou být příkladem sledování účinku nějakého léku (například sledujeme za jak dlouho začne působit lék na bolest). Dále uvažujeme tři velikosti vzorků a to konkrétně $\{100, 200, 400\}$ pozorování. Jako skutečný vektor regresních koeficientů označujeme vektor $\beta = (1, 1, 1)$.

5.2 Experimenty

Nejprve se zaměřme na celkovou přesnost odhadů a jejich výpočetní náročnost. Pro každou kombinaci parametrů dat, tedy rozdělení odchylek, procentuální míry cenzorování, velikosti vzorku a zvolené metodě odhadu bylo vytvořeno 1000 simulovaných datasetů. Celkem máme tedy 144 různých scénářů. Zaměřujeme na tři statistiky, konkrétně

- `BIAS`: Průměrná odchylka od skutečných regresních parametrů
- `MSE`: Střední kvadratická chyba od skutečných parametrů

- `AVG TIME`: Průměrný čas výpočtu

V experimentu je vynecháno měření standardní odchylky, jelikož není implementována u všech metod (konkrétně u `"gehan-poly"`) a zároveň je výpočetní náročnost pomocí resamplingové metody velmi náročná. Místo standardní odchylky byla zvolena střední kvadratická chyba, která nám říká podobnou charakteristiku a můžeme jí snadno spočítat. Také vynecháváme ve studii neupravený Buckleyho odhad, jelikož v mnoha případech nekonvergoval a dostávali bychom tedy chybné výsledky, zároveň víme že v případě konvergence je totožný s Jinovým upraveným odhadem. U všech metod ponecháváme výchozí nastavení parametrů, které je popsáno v předchozí kapitole.

Výsledky experimentů jsou k dispozici k vidění v tabulkách A.1, A.2, A.3, čísla jsou z důvodu přehlednosti tabulky zaokrouhlena na 4 desetinná místa. Všechny bodové odhady se zdají být nestranné. U metod (`"gehan"`, `"gehan-poly"`, `"jin"`) pozorujeme skutečně minimální Bias ve všech odhadnutých koeficientech pro každý scénář, s tím že největší odchylku (ovšem stále minimální) pozorujeme u scénářů s vysokou mírou cenzorování, tedy 90%.

Zajímavé pozorování zaznamenáváme u metody `"gehan-heller"`. Z hodnot sloupce `MSE` a `Bias` pozorujeme velké nepřesnosti pro data s 90% cenzorováním. Zaměřme se například na hodnoty pro scénáře se 100 pozorováními a s odchylkami z extrémního rozdělení, z hodnot `MSE` $= [8.67, 4.28, 4.06]$ usuzujeme vysoký rozptyl odhadů. Korespondující hodnoty u sloupce `Bias`, tedy $[1.47, 1.21, 1.16]$ (viz tabulka 5.1), nám říkají že Hellerův odhad má tendenci nadhodnocovat skutečný odhad regresních parametrů β. S rostoucím počtem dat se ovšem nepřesnosti zmenšují. Ukázky nadhodnocení odhadu prvního regresního koeficientu jsou vidět na obrázcích 5.2, 5.3, 5.4, kde na vykreslujeme na osu y klasický Gehanův odhad a na osu x zbylé tři odhady. Z obrázků je poté evidentně vidět odlišné chování metody `"gehan-heller"` oproti zbylým metodám. U druhého a třetího regresního koeficientu jsou výsledky podobné. Zároveň pro $n = 100$ pozorujeme menší nepřesnosti i u metody `"gehan-poly"`, s rostoucím počtem dat tyto špatné odhady ovšem nezaznamenáváme.

Tabulka 5.1 Porovnání Bias a MSE u jednotlivých metod pro data se 100 pozorováními a odchylkami z extrémního rozdělení při 90% cenzorování.

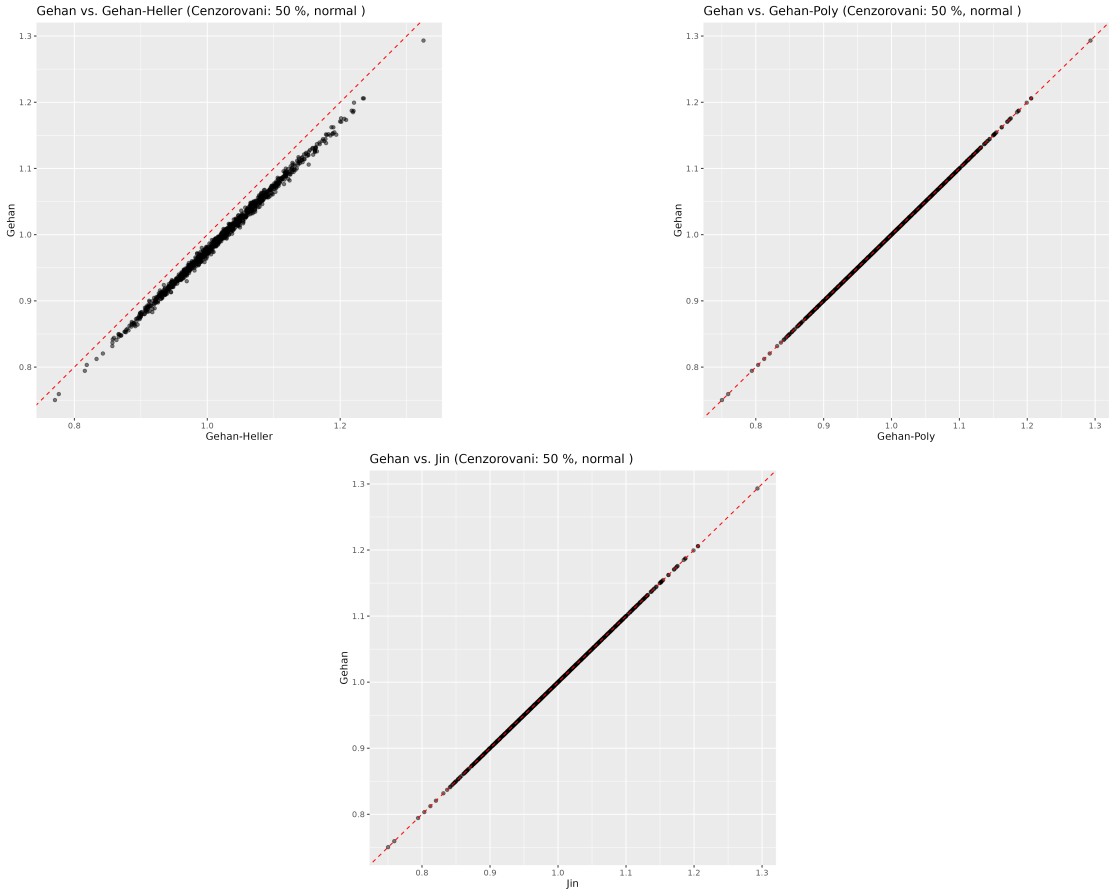
Metoda	b1		b2		b3	
	Bias	MSE	Bias	MSE	Bias	MSE
Gehan	0.1417	0.5942	0.0582	0.2104	0.0400	0.1923
Jin	0.1335	0.5926	0.0558	0.2097	0.0374	0.1936
Gehan-Poly	0.2839	1.2585	0.0574	0.2116	0.0388	0.1930
Gehan-Heller	1.4735	8.6726	1.2138	4.2837	1.1657	4.0629

Nepřesnosti, ačkoliv menší, zaznamenáváme u metody `"gehan-heller"` i pro scénáře s 90% cenzorováním a odchylkami ze standardního normálního a logistického rozdělení. Opět vidíme, že s rostoucím počtem dat tyto nepřesnosti klesají. Pravděpodobná příčina nadhodnocování tkví v naší definici funkce $D(u)$ z rovnice 3.14. Při zvolení ϕ jako standardní normální distribuční funkce dostaneme vyjádření ztrátové funkce $L_H(\beta)$ z 3.15. Při velkém cenzorování se první člen v sumách vynuluje a zbyde pouze člen $a\phi\left(\frac{e_i(\beta)-e_j(\beta)}{a}\right)$. Hustota standardního normálního rozdělení tedy špatně zachycuje data z extrémního rozdělení a to vede k nadhodnocování bodového odhadu. Možným řešením by mohlo být zvolení jiné distribuční funkce ϕ či škálovacího parametru a, tento krok by ovšem zároveň nemusel opět fungovat pro jiné rozdělení odchylek a vyjádření ztrátové funkce by už nebylo tak jednoduché jako v případě standarní normální distribuční funkce. Dalším způsobem by mohlo být zavedení nějaké regularizace a použití robustních statistických technik, tento krok by ovšem vyžadoval extenzivnější experimentování a delší studii robustní statistiky a proto je v práci vynechán. Jako nejjednoduší řešení by poté mohlo být malinko jiné vyjádření funkce $U_H(\beta)$ z definice 3.14, kde bychom místo námi zvoleného dělitele n^{-1} mohli

použít malinko větší číslo, například $n^{-0.5}$. Tyto nápady ovšem zatím ponecháváme k dalšímu zkoumání.

Pomineme-li scénáře s vysokým procentem cenzorování, tak si metoda `"gehan-heller"` vede velmi dobře a nepozorujeme žádné podezřelé nepřesnosti odhadů regresních parametrů.

S rostoucím počtem pozorování dávají metody téměř stejné výsledky. Ukázku odhadu druhého regresního parametru pro data se 400 pozorováními, standardně rozdělenými normálními odchylkami a 50% mírou cenzorování můžeme vidět na obrázku 5.1.

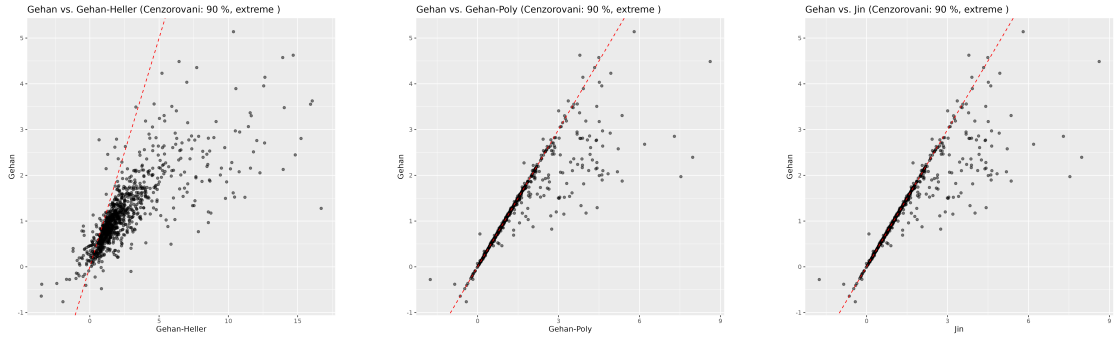


Obrázek 5.1 Porovnání odhadů s metodou Gehan při 50% cenzorování, s odchylkami ze standardního normálního rozdělení a 400 pozorováními

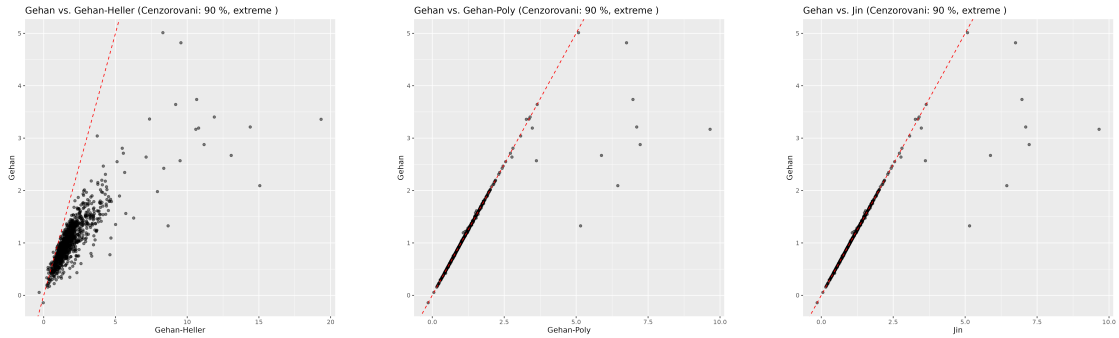
Z obrázku 5.1 opět pozorujeme menší nadhodnocení odhadu a vyšší rozptyl u metody `"gehan-heller"` oproti odhadům pomocí metody `"gehan"`. Jedná se ovšem o minimální chybu, kterou v praxi můžeme ignorovat.

Pokud porovnáme metody založené na pořadových metodách, tedy `"gehan"`,`"gehan-poly"` a `"gehan-heller"` oproti metodě `"jin"`, která modifikuje metodu nejmenších čtverců, tak nepozorujeme žádné signifikantní rozdíly v přesnostech odhadů (kromě již výše zmíněného scénáře s vysokým procentem cenzorování). Pokud budeme chtít data prozkoumat velice podrobně, tak můžeme říci že pořadové metody fungují lépe pro odchylky z extrémního rozdělení a naopak metody založené na nejmenších čtvercích fungují malinko lépe pro data s odchylkami z normálního a logistického rozdělení.

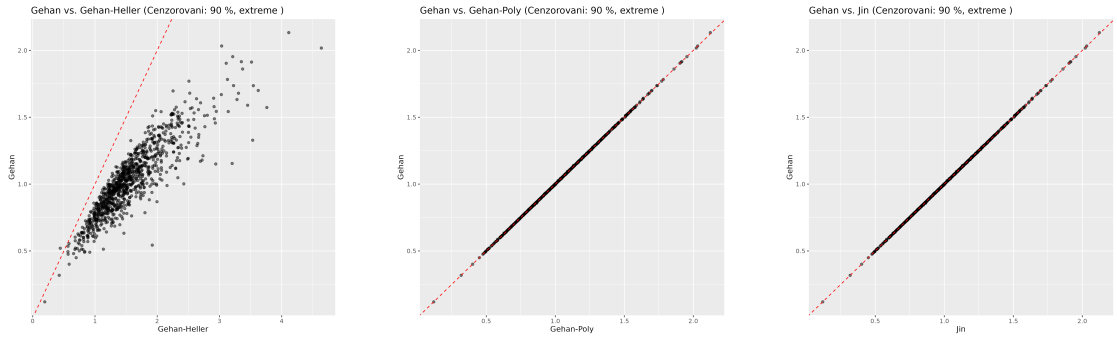
Zaměřme se nyní na výpočetní náročnost, tedy na sloupec `Avg Time`. Porovnání průměrného času výpočtu pro jednotlivé metody v závislosti na jednotlivých scénářích můžeme vidět na obrázcích 5.5, 5.6, 5.7.



Obrázek 5.2 Porovnání odhadů s metodou Gehan pro data s 90% mírou cenzorování, odchylkami z extrémního rozdělení a 100 pozorováními



Obrázek 5.3 Porovnání odhadů s metodou Gehan pro data s 90% mírou cenzorování, odchylkami z extrémního rozdělení a 200 pozorováními



Obrázek 5.4 Porovnání odhadů s metodou Gehan pro data s 90% mírou cenzorování, odchylkami z extrémního rozdělení a 400 pozorováními

Z časových výsledků odhadu jednotlivých metod vidíme několik pozorování. Metoda `"gehan"` dokáže na menších datech získat regresní koeficienty velmi rychle, to je zapříčiněno hlavně díky způsobu minimalizace pomocí mediánové regrese, která využívá efektivně algoritmus Koenkera & D'Oreyho. Metoda `"jin"` je na tom s výpočetní složitostí velmi podobně jako metoda `"gehan"`. S rostoucím počtem dat ovšem délka výpočtu pro tyto dvě metody lineárně roste. Pro data s $n = 100$ dokážou metody `"gehan"` a `"jin"` najít regresní parametry v každém scénáři do 0.4 sekund. Ovšem pro data s $n = 400$ výpočet trvá již v řádech nízkých sekund. Jako nejhorší scénář se jeví situace s nulovým procentem cenzorování, kde výpočet dosahuje až 5.7 sekund. Nejrychleji naopak tyto metody pracují s daty s vysokým procentem cenzorování, kde i pro data

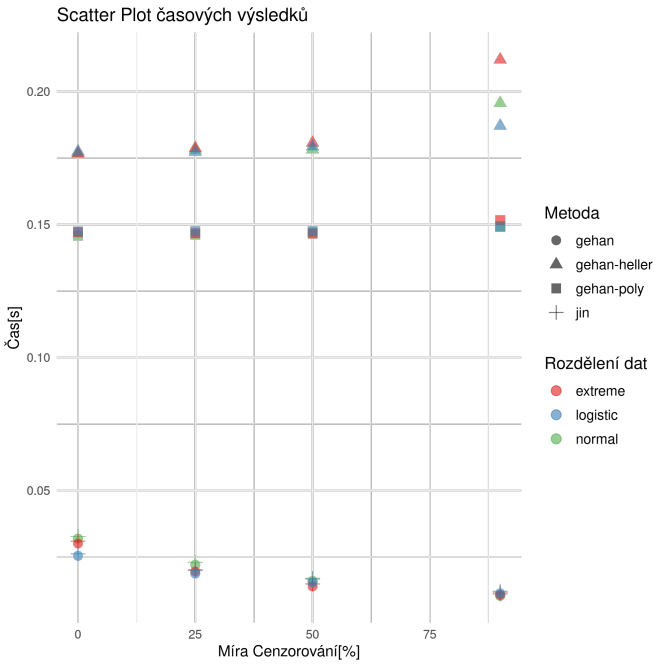
s $n = 400$ trvá výpočet pro scénáře s 90% cenzorováním pouze 0.2 sekund. Při bližším zkoumání, také zjistíme že pracují lehce rychleji s daty s odchylkami z extrémního a logistického rozdělení (viz porovnávající tabulka 5.2).

Tabulka 5.2 Průměrná doba výpočtů pro různé metody na datech se 400 pozorováními

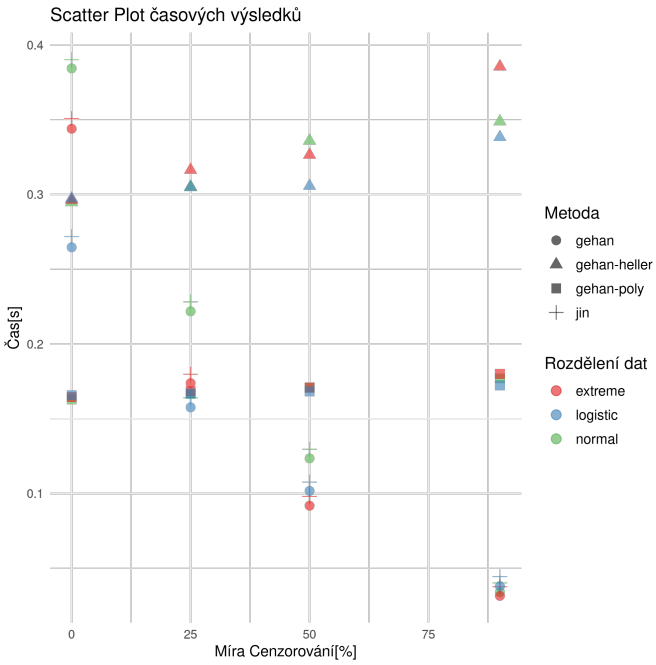
Cenzorování	Gehan	Jin	Gehan-Poly	Gehan-Heller
		Normální rozdělení		
0%	5.6964	5.7153	0.2295	0.7828
25%	3.2124	3.2492	0.2360	0.8104
50%	1.6211	1.6424	0.2408	0.8074
90%	0.2092	0.2339	0.2731	0.9805
		Extrémní rozdělení		
0%	5.1487	5.1751	0.2319	0.7776
25%	2.4321	2.4580	0.2354	0.8011
50%	1.1164	1.1416	0.2480	0.8357
90%	0.1694	0.1970	0.2919	1.1930
		Logistické rozdělení		
0%	3.9262	3.9136	0.2496	0.9390
25%	2.1702	2.1911	0.2290	0.7889
50%	1.2723	1.2943	0.2352	0.8047
90%	0.2746	0.2956	0.2518	0.8433

Metoda, která se ukázala jako velmi rychlá pro všechny scénáře a velikosti datasetů je `"gehan-poly"`. Rychlost této metody a její celková přesnost dělá z `"gehan-poly"` ideálního kandidáta pro odhad regresních parametrů. Je potřeba ovšem zmínit, že pracujeme pouze s časy odhadu regresních parametrů a nezapočítáváme časy odhadu rozptylu. Pokud bychom měřili i čas odhadu rozptylu, tak je jednoznačně nejrychlejší metoda `"gehan-heller"`, která zvládne vypočítat kovarianční matici pouhým dosazením odhadnutých koeficientů $\hat{\beta}_H$. U ostatních metod by odhad kovarianční matice trval v řádech sekund pro menší data a pro větší v řádech minut či dokonce hodin (například pro metodu `"gehan"` trvá celkový odhad pro data s $n = 400$ a s argumentem `resampling` nastaveným na 1000 realizací 25 minut).

Čas odhadu pomocí metody `"jin"` je přímo závislý na době trvání odhadu pomocí metody `"gehan"`, jelikož se toto řešení bere jako počáteční odhad. Možné řešení, které by zrychlilo výpočet metody `"jin"` by mohlo být použití odhadu z metody `"gehan-poly"` jako počátečního řešení. Víme, že pokud je jako počáteční řešení zvoleno $\hat{\beta}_G$, které minimalizuje Gehanovu funkci, tak je výsledné $\hat{\beta}_k$ asymptoticky normální. Tedy jinými slovy záleží na počátečním odhadu $\hat{\beta}_0$, pokud se jedná o konzistentní odhad skutečných regresních koeficientů, tak je výsledné $\hat{\beta}_k$ také konzistetní. Z výsledků experimentů pozorujeme shodu metod `"gehan"` a `"gehan-poly"`, zároveň víme, že můžeme udělat průměrnou chybu odhadu velmi malou zvolením vhodného ϵ. Mohli bychom tedy vzít řešení z metody `"gehan-poly"` aniž bychom přišli o asymptotické vlastnosti $\hat{\beta}_k$. Stále by se ovšem nevyřešil problém s odhadem kovarianční matice, který by musel být opět proveden pomocí resamplingu.

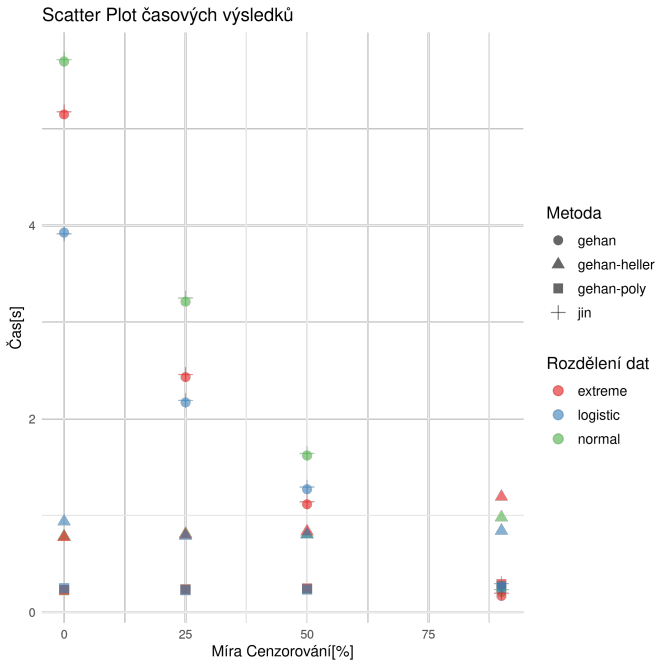


Obrázek 5.5 Průměrné časy výpočtů jednotlivých metod pro data se 100 pozorováními



Obrázek 5.6 Průměrné časy výpočtů jednotlivých metod pro data s 200 pozorováními

Další možné zrychlení se týká přímo metody `"gehan"`. V balíčku je minimalizace prováděna pomocí mediánové regrese s již zmíněným algoritmem Koenkera & D'Oreyho. Tento algoritmus je velmi rychlý pro menší datasety, s větším vzorkem pozorování si ovšem nevede tak dobře. Můžeme ovšem zkusit minimalizovat i pomocí jiného algoritmu. Po prostudování dokumentace balíčku



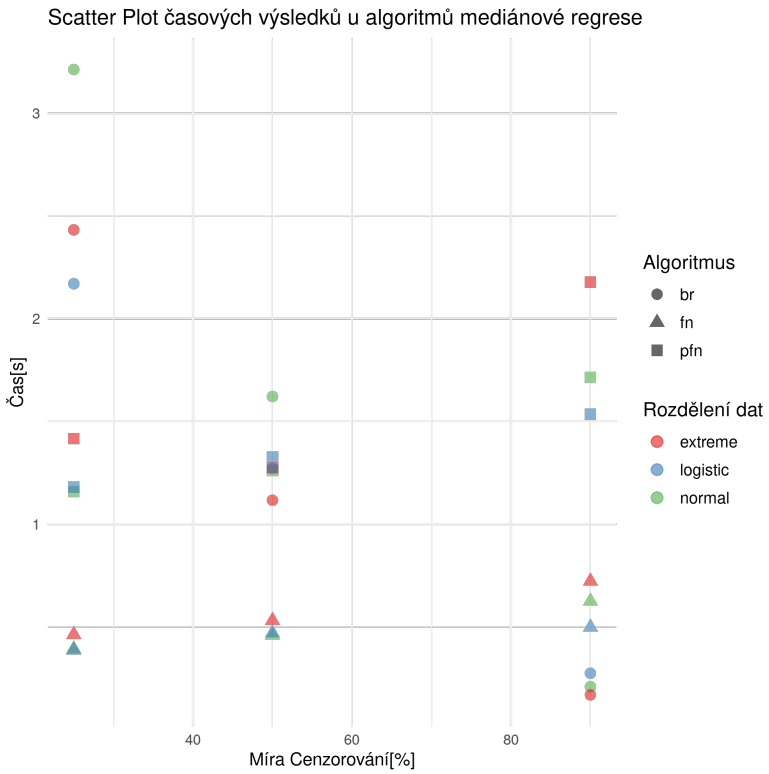
Obrázek 5.7 Průměrné časy výpočtů jednotlivých metod pro data se 400 pozorováními

quantreg byly zvoleny dvě další metody minimalizace, konkrétně Frisch–Newtonova metoda vnitřního bodu (označovaná jako `"fn"`) a její upravená varianta Frisch–Newtonova metoda po předzpracování (označovaná jako `"pnf"`). Předzpracování v našem kontextu znamená redukce dimenzionality problému vyhozením některých vybraných pozorování. Popis těchto algoritmů je nad rámec této bakalářské práce, jejich detailní formulaci lze najít v původním článku [60]. Dle dokumentace balíčku **quantreg** a článku [60] zmíněné algoritmy pracují rychleji pro větší datasety. Byl proveden tedy experiment na 300 simulovaných datasetech se 400 pozorováními, kde jsme otestovali rychlost těchto metod na datech s $\{25, 50, 90\}$ procentním cenzorováním a odhylkami ze všech rozdělení. Výsledek experimentu je k dispozici k vidění v tabulce 5.3. Časové porovnání je poté k dispozici k vidění na obrázku 5.8.

Z časových výsledků vidíme velké zrychlení při použití algoritmu `"fn"`. Z původních 3,21 sekund (nejhorší naměřené číslo pro data s odchylkami z normálního rozdělení a 25% cenzorováním, viz tabulka 5.2) se celý proces odhadu zrychlil na 0.39 sekund. Pro větší datasety je tedy velmi výhodné použít Frisch–Newtonovu metodu, která klasický `"br"` algoritmus poráží ve všech scénářích kromě těch s vysokým procentem cenzorování. Při 90% cenzorování je algoritmus Koenkera & D'Oreyho nejrychlejší a změna algoritmu minimalizace se tedy nevyplatí. U druhé testované metody (`"pfn"`) výrazné zrychlení nezaznamenáváme, ovšem její síla by se dle dokumentace balíčku **quantreg** a článku [60] mohla projevit při ještě větších datasetech. Z časových důvodů ovšem nejsou tyto experimenty v práci uvažovány. Z tabulky 5.3 zároveň nepozorujeme žádné podezřelé hodnoty sloupců `Bias` a `MSE` pro obě metody.

Tabulka 5.3 Různé optimalizační algoritmy mediánové regrese pro Gehanovu metodu na 400 pozorováních, zaokrouhleno na 4 desetiná místa

		FN				PNF		
	Cenzorování	Bias	MSE	Avg Time		Bias	MSE	Avg Time
					Normální rozdělení			
	25%	-0.0060	0.0123	0.3920		-0.0060	0.0123	1.1584
b1	50%	-0.0065	0.0152	0.4586		-0.0065	0.0152	1.2613
	90%	0.0009	0.0638	0.6248		0.0009	0.0638	1.7141
	25%	-0.0051	0.0032	0.3920		-0.0051	0.0032	1.1584
b2	50%	-0.0047	0.0050	0.4586		-0.0047	0.0050	1.2613
	90%	0.0006	0.0233	0.6248		0.0006	0.0233	1.7141
	25%	0.0014	0.0035	0.3920		0.0014	0.0035	1.1584
b3	50%	0.0035	0.0051	0.4586		0.0035	0.0051	1.2613
	90%	0.0028	0.0211	0.6248		0.0028	0.0211	1.7141
					Extrémní rozdělení			
	25%	-0.0092	0.0135	0.4628		-0.0092	0.0135	1.4164
b1	50%	0.0013	0.0170	0.5315		0.0013	0.0170	1.2744
	90%	0.0289	0.0653	0.7226		0.0290	0.0653	2.1783
	25%	0.0004	0.0036	0.4628		0.0004	0.0036	1.4164
b2	50%	0.0025	0.0048	0.5315		0.0025	0.0048	1.2744
	90%	0.0161	0.0221	0.7226		0.0161	0.0221	2.1783
	25%	-0.0033	0.0043	0.4628		-0.0033	0.0043	1.4164
b3	50%	0.0009	0.0052	0.5315		0.0009	0.0052	1.2744
	90%	0.0165	0.0236	0.7226		0.0165	0.0236	2.1783
					Logistické rozdělení			
	25%	0.0083	0.0316	0.3873		0.0083	0.0316	1.1803
b1	50%	0.0163	0.0408	0.4701		0.0163	0.0408	1.3272
	90%	0.0300	0.1230	0.4985		0.0300	0.1230	1.5345
	25%	0.0004	0.0099	0.3873		0.0004	0.0099	1.1803
b2	50%	-0.0002	0.0127	0.4701		-0.0002	0.0127	1.3272
	90%	0.0048	0.0363	0.4985		0.0048	0.0363	1.5345
	25%	-0.0058	0.0097	0.3873		-0.0058	0.0097	1.1803
b3	50%	-0.0044	0.0121	0.4701		-0.0044	0.0121	1.3272
	90%	0.0019	0.0388	0.4985		0.0019	0.0388	1.5345



Obrázek 5.8 Průměrné časy výpočtů Gehanova odhadu v závislosti na různých minimalizujících algoritmech pro data se 400 pozorováními

Dále byly provedeny experimenty na 100 simulovaných datasetech se 400 pozorováními pro metody `"gehan-heller"` a `"gehan-poly"`. Byl zkoumán dopad zvolení různých optimalizačních algoritmů na výsledek minimalizace. Jako zvolené algoritmy byly vybrány metody

- `"BFGS"`
- `"L-BFGS"`
- `"Nelder-Mead"`

Abychom předešli zkreslení výsledků použitím manuálně napsaného gradientu, tak je pro metody `"BFGS"` a `"L-BFGS"` použita numerická aproximace z balíčku **optimx**. Metoda `"Nelder-Mead"` k optimalizaci nevyžaduje derivování ztrátové funkce a k nalezení minima vytváří iterativně, pomocí heuristických funkcí, body simplexu a postupně konverguje k lokálnímu minimu, více viz literatura [28]. Intuitivně předpokládáme, že by mohl algoritmus `"L-BFGS"` fungovat rychleji pro větší datasety, jelikož pracuje pouze s omezenou pamětí počítače. Výsledky experimentu jsou k dispozici v tabulce 5.4. V tabulce jsou opět hodnoty zaokrouhleny na 4 desetinná místa, zároveň jsou z ní vynechány údaje s odchylkami z extrémního a logistického rozdělení, jelikož jsme u nich nezpozorovali výrazné změny. Také vynecháváme scénář s nulovou mírou cenzorování a soustředíme se pouze na cenzorovaná data.

Celkově nepozorujeme žádné podezřelé hodnoty, všechny optimalizační algoritmy poskytují podobné výsledky ve všech sloupcích. Pro zajímavost byl proveden i experiment pro dataset s 800 pozorováními (viz tabulka 5.5), kde metoda `"L-BFGS"` dosáhne minima až o půl sekundy rychleji než u klasického algoritmu `"BFGS"`. Naopak metoda `"Nelder-Mead"` si pro větší datasety vede nejhůře s průměrným časem o více než sekundu delším než u ostatních metod. Pro středně

Tabulka 5.4 Různé optimalizační algoritmy na 400 vzorcích, zaokrouhleno na 4 desetinná místa

		Gehan-Poly			Gehan-Heller		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time
				BFGS			
b1	25%	0.0194	0.0136	0.2846	0.0285	0.0143	1.0232
	50%	0.0193	0.0160	0.2800	0.0451	0.0186	1.0396
	90%	0.0314	0.0556	0.3094	0.2889	0.1957	1.2301
b2	25%	-0.0137	0.0037	0.2846	-0.0045	0.0036	1.0232
	50%	0.0025	0.0041	0.2800	0.0108	0.0061	1.0396
	90%	-0.0125	0.0245	0.3094	0.2217	0.0965	1.2301
b3	25%	-0.0008	0.0029	0.2846	0.0193	0.0160	1.0232
	50%	0.0025	0.0041	0.2800	0.0275	0.0051	1.0396
	90%	-0.0079	0.0181	0.3094	0.2274	0.0882	1.2301
				L-BFGS			
b1	25%	0.0194	0.0136	0.3530	0.0285	0.0143	1.0058
	50%	0.0193	0.0160	0.3145	0.0451	0.0186	1.0447
	90%	0.0314	0.0556	0.3522	0.2889	0.1957	1.0783
b2	25%	-0.0137	0.0037	0.3530	-0.0045	0.0036	1.0058
	50%	-0.0139	0.0059	0.3145	0.0451	0.0186	1.0447
	90%	-0.0125	0.0245	0.3522	0.2217	0.0965	1.0783
b3	25%	-0.0008	0.0029	0.3530	0.0086	0.0031	1.0058
	50%	0.0025	0.0041	0.3145	0.0108	0.0061	1.0447
	90%	-0.0079	0.0181	0.3522	0.2274	0.0882	1.0783
				Nelder-Mead			
b1	25%	0.0194	0.0136	0.3073	0.0285	0.0143	1.1921
	50%	0.0192	0.0160	0.3010	0.0451	0.0186	1.1001
	90%	0.0313	0.0555	0.3107	0.2889	0.1958	1.1943
b2	25%	-0.0137	0.0037	0.3073	-0.0045	0.0036	1.1921
	50%	-0.0139	0.0059	0.3010	0.0451	0.0186	1.1001
	90%	-0.0125	0.0245	0.3107	0.2217	0.0965	1.1943
b3	25%	-0.0009	0.0029	0.3073	0.0086	0.0031	1.1921
	50%	0.0025	0.0042	0.3010	0.0275	0.0051	1.1001
	90%	-0.0079	0.0181	0.3107	0.2274	0.0883	1.1943

velké a malé datasety ovšem usuzujeme, že na optimalizačním algoritmu nezáleží a doporučujeme pracovat s výchozí metodou `"BFGS"`. Pokud ovšem uživatel bude pracovat s opravdu velkými daty, tak se může vyplatit optimalizační algoritmus změnit na `"L-BFGS"`.

Tabulka 5.5 Různé optimalizační algoritmy na 800 vzorcích, zaokrouhleno na 4 desetinná místa

		Gehan-Poly			Gehan-Heller		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time
				BFGS			
b1	25%	-0.0063	0.0077	0.5576	-0.0001	0.0075	2.9130
	50%	0.0157	0.0182	0.5966	0.0366	0.0204	3.0410
	90%	0.0360	0.0247	0.7298	0.2417	0.1014	3.5370
b2	25%	0.0066	0.0013	0.5576	0.0139	0.0014	2.9130
	50%	0.0022	0.0014	0.5966	0.0219	0.0019	3.0410
	90%	0.0046	0.0021	0.7298	0.1897	0.0465	3.5370
b3	25%	-0.0005	0.0005	0.5576	0.0068	0.0006	2.9130
	50%	0.0111	0.0015	0.5966	0.0307	0.0025	3.0410
	90%	0.0046	0.0021	0.7298	0.1798	0.0353	3.5370
				L-BFGS			
b1	25%	-0.0063	0.0077	0.7284	-0.0001	0.0075	2.8480
	50%	0.0156	0.0182	0.7235	0.0366	0.0204	2.8220
	90%	0.0359	0.0247	0.8267	0.2417	0.1014	2.9230
b2	25%	0.0066	0.0013	0.7284	0.0139	0.0014	2.8480
	50%	0.0022	0.0014	0.7235	0.0219	0.0019	2.8220
	90%	0.0082	0.0070	0.8267	0.1897	0.0465	2.9230
b3	25%	-0.0005	0.0005	0.7284	0.0068	0.0006	2.8480
	50%	0.0111	0.0015	0.7235	0.0306	0.0025	2.8220
	90%	0.0046	0.0021	0.8267	0.1798	0.0353	2.9230
				Nelder-Mead			
b1	25%	-0.0062	0.0077	0.6167	-0.0003	0.0075	3.5110
	50%	0.0157	0.0182	0.6699	0.0366	0.0204	3.7630
	90%	0.0360	0.0247	0.7614	0.2418	0.1015	4.2460
b2	25%	0.0066	0.0013	0.6167	0.0138	0.0014	3.5110
	50%	0.0022	0.0014	0.6699	0.0219	0.0019	3.7630
	90%	0.0082	0.0070	0.7614	0.1898	0.0465	4.2460
b3	25%	-0.0005	0.0005	0.6167	0.0068	0.0006	3.5110
	50%	0.0111	0.0015	0.6699	0.0306	0.0025	3.7630
	90%	0.0047	0.0021	0.7614	0.1798	0.0353	4.2460

Nakonec byl uskutečněn experiment zaměřený na zvolení počátečního odhadu regresních koeficientů β. Experiment byl proveden opět na 100 simulovaných datasetech ale pouze se 100 pozorováními. Jako vybrané počáteční odhady byly zvoleny

- `"gehan"`: Řešení minimalizující Gehanovu funkci
- `"lm"`: Řešení pomocí obyčejné metody nejmenších čtverců na celém datasetu.
- `zero`: Nulový vektor o dimenzi 3

Výsledek experimentu je k dispozici k vidění v tabulce 5.6. Z tabulky jsou vynechány údaje o datech z extrémního a logistického rozdělení, jelikož jsme u nich nezaznamenali žádné významné změny. Celkově nepozorujeme žádné různorodé výsledky a pro všechny počáteční odhady je tedy optimalizační cesta velmi podobná. Z výpočetních důvodů doporučujeme ponechat argument `binit` na výchozí hodnotě, tedy na `"lm"`. Výpočet pomocí obyčejné metody nejmenších čtverců je totiž velmi rychlý i pro velké datasety a nemusí se tak zbytečně dlouho čekat na výpočet Gehanova odhadu. Zároveň počáteční odhad `lm` dosahuje lepších výsledků než odhad `zero`, nutno ovšem podotknout že vskutku minimálních.

Tabulka 5.6 Různé počáteční odhady vektoru β na 100 vzorcích, zaokrouhleno na 4 desetinná místa

		Gehan-Poly			Gehan-Heller		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time
				Gehan			
b1	25%	-0.0371	0.0474	0.1891	-0.0234	0.0485	0.2220
	50%	-0.0190	0.0642	0.1864	0.0240	0.0711	0.2199
	90%	-0.0560	0.2141	0.1804	0.3495	0.8976	0.2287
b2	25%	-0.0057	0.0119	0.1891	0.0098	0.0123	0.2220
	50%	0.0045	0.0189	0.1864	0.0484	0.0233	0.2199
	90%	-0.0359	0.0789	0.1804	0.4069	0.5024	0.2287
b3	25%	-0.0018	0.0125	0.1891	0.0135	0.0133	0.2220
	50%	0.0004	0.0171	0.1864	0.0431	0.0216	0.2199
	90%	0.0017	0.0786	0.1804	0.4125	0.4799	0.2287
				Lm			
b1	25%	-0.0371	0.0474	0.1696	-0.0236	0.0485	0.2064
	50%	-0.0189	0.0643	0.1679	0.0232	0.0714	0.2042
	90%	-0.0516	0.2170	0.1697	0.4013	1.1087	0.2137
b2	25%	-0.0056	0.0119	0.1696	0.0095	0.0123	0.2064
	50%	0.0046	0.0189	0.1679	0.0471	0.0233	0.2042
	90%	-0.0355	0.0791	0.1697	0.4238	0.5803	0.2137
b3	25%	-0.0017	0.0125	0.1696	0.0133	0.0133	0.2064
	50%	0.0005	0.0171	0.1679	0.0419	0.0214	0.2042
	90%	0.0022	0.0788	0.1697	0.4360	0.4865	0.2137
				Zero			
b1	25%	-0.0370	0.0474	0.1750	-0.0055	0.0503	0.1991
	50%	-0.0189	0.0642	0.1690	0.0584	0.0795	0.2013
	90%	-0.0547	0.2125	0.1711	0.4888	1.3857	0.2202
b2	25%	-0.0057	0.0119	0.1750	0.0294	0.0140	0.1991
	50%	0.0046	0.0189	0.1690	0.0835	0.0307	0.2013
	90%	0.0016	0.0786	0.1711	0.5129	0.7994	0.2202
b3	25%	-0.0017	0.0125	0.1750	0.0328	0.0151	0.1991
	50%	0.0004	0.0171	0.1690	0.0756	0.0276	0.2013
	90%	-0.0361	0.0790	0.1711	0.5178	0.6521	0.2202

Pro zajímavost byl experiment proveden i na datech s 800 pozorováními, tentokrát pouze s počátečními odhady `"lm"` a `zero` (viz tabulka 5.7). Z výsledků experimentu se potvrdilo že počáteční odhad `"lm"` je pro minimalizaci nejvýhodnější, jelikož je celkový proces odhadu regresních parametrů o dost rychlejší než pro počáteční hodnotu `zero`. Největší rozdíl je vidět u metody `"Gehan-Heller"`, kde se s použitím počátečního odhadu pomocí metody nejmenších čtverců dostaneme k výsledku o více než sekundu rychleji pro data s 25% cenzorováním. Usuzujeme tedy, že počáteční řešení `"lm"` je blíže ke globálnímu minimu ztrátové funkce.

Tabulka 5.7 Různé počáteční hodnoty vektoru β na 800 vzorcích bez odhadu založeným na Gehanově metodě, zaokrouhleno na 4 desetinná místa

		Gehan-Poly			**Gehan-Heller**		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time
				Lm			
	25%	-0.0011	0.0065	0.6094	0.0057	0.0066	2.8853
b1	50%	-0.0024	0.0089	0.6435	0.0171	0.0096	2.7438
	90%	0.0036	0.0308	0.7307	0.1945	0.0851	3.3546
	25%	0.0012	0.0019	0.6094	0.0083	0.0020	2.8853
b2	50%	0.0016	0.0028	0.6435	0.0209	0.0033	2.7438
	90%	0.0052	0.0103	0.7307	0.1850	0.0514	3.3546
	25%	-0.0028	0.0017	0.6094	0.0043	0.0018	2.8853
b3	50%	-0.0022	0.0023	0.6435	0.0168	0.0027	2.7438
	90%	-0.0012	0.0098	0.7307	0.1784	0.0490	3.3546
				Zero			
	25%	-0.0011	0.0065	0.7324	0.0144	0.0069	4.3428
b1	50%	-0.0024	0.0089	0.6981	0.0328	0.0108	2.9665
	90%	0.0036	0.0308	0.7601	0.2264	0.1017	3.4247
	25%	0.0012	0.0019	0.7324	0.0174	0.0023	4.3428
b2	50%	0.0016	0.0028	0.6981	0.0364	0.0044	2.9665
	90%	0.0052	0.0103	0.7601	0.2151	0.0649	3.4247
	25%	-0.0028	0.0017	0.7324	0.0132	0.0020	4.3428
b3	50%	-0.0022	0.0023	0.6981	0.0322	0.0035	2.9665
	90%	-0.0012	0.0098	0.7601	0.2082	0.0620	3.4247

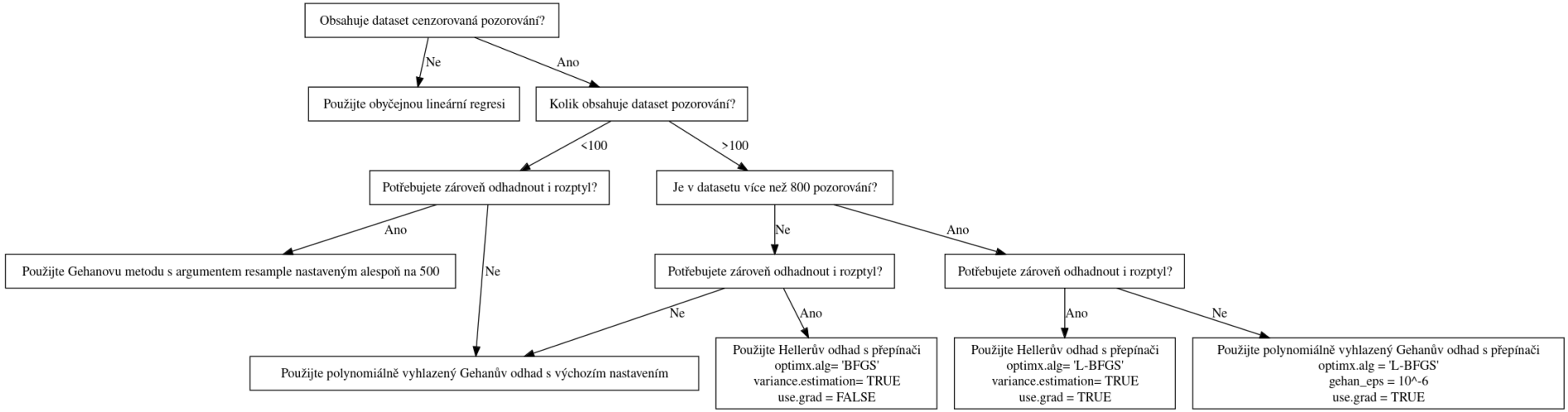
5.3 Zajímavé pozorování

V průběhu testování bylo zjištěno pár pozorování, které se mohou hodit vědět případnému uživateli balíčku **aftsem**. Jedná se o rady jak efektivně s balíčkem pracovat a případně rozšiřovat.

- Optimalizace pomocí balíčku **optimx** je poměrně citlivá na počáteční odhad v případě minimalizace s použitím manuálně napsaného gradientu. V případě nekonvergence či chyby minimalizace tedy doporučujeme zkusit více počátečních odhadů. Použití manuálně napsaného gradientu může být velmi výhodné pro velké datasety, jelikož balíček **optimx** nemusí sám numericky aproximovat gradient a stačí pouhé dosazení do předem napsané funkce.
- Při případném přidávání nových metod se vyplatí psát nové odhady v jazyku c++ pomocí knihovny **armadillo**. Výsledný kód poté běží několika násobně rychleji než v čistém R a také vyžaduje mnohem méně počítačové paměti. Doporučujeme psát pouze definice ztrátových funkcí (v případě pořadových metod) a minimalizaci nechat na efektivních algoritmech z knihovny **optimx**.
- Při odhadu kovarianční matice pomocí resamplingu může být doba výpočtu až nerozumně dlouhá. Je potřeba tedy zvolit vhodné číslo realizací náhodných veličin R_i, neboli číslo N. Obecně platí, že čím větší bude N, tím přesnější odhad dostaneme. Wilcox ve své knize [61] doporučuje použít alespoň 1000 realizací. Při testování balíčku se ovšem jeví i použití pouze 500 realizací jako dostatečné a zároveň se ušetří velké množství času.
- Z výsledků experimentů nepozorujeme žádné významné rozdíly mezi metodami založenými na pořadových metodách a Jinovo metodou (tedy upravená metoda nejmenších čtverců). Hlavní výhoda pořadových metod spočívá v tom, že nemusí odhadovat pravděpodobnostní rozdělení reziduí modelu a jsou tedy jak teoreticky tak výpočetně méně náročné. Výsledné hodnoty odhadů jsou pro oba dva způsoby ovšem velmi podobné a případný uživatel si tak může vybrat, kterou použije podle vlastního uvážení.

5.4 Zvolení vhodné metody

Z výsledků experimentů byl sestaven rozhodovací strom (viz obrázek 5.9), který může pomoci uživateli zvolit vhodnou metodu odhadu regresních parametrů při provádění semiparametrické analýzy cenzorovaných dat pomocí balíčku **aftsem**. Strom samozřejmě nedokáže pokrýt všechny možné scénáře a je inspirován hlavně povahou provedených experimentů a vlastnostmi simulovaných dat, stále ovšem může sloužit jako odrazový můstek pro základní analýzu.



Obrázek 5.9 Rozhodovací strom pro výběr vhodné metody odhadu regresních parametrů

Kapitola 6
Závěr

Tato bakalářská práce se zaměřuje na představení možných odhadů regresních koeficientů u semiparametrického modelu zrychleného času, který je jedním ze základních regresních modelů používaných ve statistické analýze přežití. V práci je kladen důraz jak na představení samotné analýzy přežití a regresního modelování, tak na popis jednotlivých metod odhadů. V práci jsou představeny vybrané metody založené na lineárních pořadových testech a také i metody které modifikují algoritmus nejmenších čtverců a staví na známém Buckleyho algoritmu. Vybrané metody jsou detailně popsány v samostatné kapitole a uvádějí tak čtenáře do problému semiparametrického regresního modelování pomocí modelu zrychleného času.

Praktickým výstupem práce je poté R balíček **aftsem**, který implementuje popsané postupy a u některých přidává i odhad rozptylu. Balíček je zdokumentován a otestován jak na simulovaných tak na reálných datech. Na umělých datech byla navíc provedena simulační studie, kde bylo zjištěno vhodné použití jednotlivých metod v různých scénářích zahrnující odlišné procentuální míry cenzorování, velikosti vzorků a statistické rozdělení šumu. Zároveň byly nalezeny vhodné parametry balíčku pro tyto scénáře (optimalizační algoritmy, počáteční odhady vektoru koeficientů β). Implementace balíčku využívá jak vlastního kódu, tak i funkcí z ostatních balíčků a poznatků z přečtené literatury. Balíček může být přidán po drobných úpravách do systému CRAN a poskytnut k použití širší veřejnosti zajímající se o analýzu přežití.

Balíček by bylo možné časem rozšířit o odhad kovarianční matice pro Gehanův polynomiální odhad pomocí resamplingu, testy dobré shody s daty a poté další dosud neimplementované metody jako je například Linova jádrová metoda [33] či modifikované Buckleyho metody [45],[32]. Dále je možné prozkoumat numerickou stabilitu již naimplementovaných odhadů či existující odhady modifikovat přidáním nějakých robustních metod jako je například regularizace. Celkově slouží práce jako základní úvod do regresního modelování v analýze přežití a může se na ní dále stavět či ji upravovat.

Příloha A

Výsledky Experimentů

Tabulka A.1 Výsledky na datech se 100 pozorováními, zaokrouhleno na 4 desetinná místa

		Gehan			Jin			Gehan-Poly			Gehan-Heller		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time	Bias	MSE	Avg Time	Bias	MSE	Avg Time
							Normální rozdělení						
b1	0%	-0.0138	0.0448	0.0319	-0.0133	0.0423	0.0327	-0.0138	0.0448	0.1458	-0.0138	0.0442	0.1764
	25%	-0.0123	0.0526	0.0221	-0.0122	0.0508	0.0230	-0.0123	0.0526	0.1462	0.0028	0.0542	0.1780
	50%	0.0008	0.0733	0.0160	-0.0026	0.0698	0.0170	0.0007	0.0734	0.1469	0.0441	0.0836	0.1782
	90%	0.0841	0.4516	0.0104	0.0823	0.4372	0.0111	0.1142	0.6432	0.1494	0.7466	3.1344	0.1956
b2	0%	-0.0037	0.0103	0.0319	-0.0035	0.0097	0.0327	-0.0037	0.0103	0.1458	-0.0036	0.0102	0.1764
	25%	-0.0012	0.0127	0.0221	-0.0019	0.0119	0.0230	-0.0012	0.0127	0.1462	0.0148	0.0134	0.1780
	50%	0.0025	0.0187	0.0160	0.0020	0.0179	0.0170	0.0025	0.0187	0.1469	0.0473	0.0233	0.1782
	90%	0.0231	0.1128	0.0104	0.0203	0.1090	0.0111	0.0234	0.1140	0.1494	0.5523	0.9450	0.1956
b3	0%	0.0016	0.0115	0.0319	0.0015	0.0109	0.0327	0.0016	0.0115	0.1458	0.0015	0.0114	0.1764
	25%	0.0016	0.0153	0.0221	0.0012	0.0145	0.0230	0.0016	0.0153	0.1462	0.0177	0.0162	0.1780
	50%	0.0061	0.0216	0.0160	0.0041	0.0203	0.0170	0.0061	0.0216	0.1469	0.0504	0.0269	0.1782
	90%	0.0323	0.1052	0.0104	0.0330	0.1012	0.0111	0.0326	0.1058	0.1494	0.5579	0.8636	0.1956
							Extrémní rozdělení						
b1	0%	-0.0002	0.0542	0.0301	-0.0037	0.0636	0.0310	-0.0002	0.0541	0.1471	-0.0005	0.0543	0.1768
	25%	0.0034	0.0568	0.0196	-0.0019	0.0665	0.0201	0.0034	0.0568	0.1467	0.0239	0.0600	0.1787
	50%	0.0109	0.0705	0.0140	0.0047	0.0794	0.0149	0.0108	0.0705	0.1466	0.0719	0.0873	0.1807
	90%	0.1417	0.5942	0.0108	0.1335	0.5926	0.0115	0.2839	1.2585	0.1516	1.4735	8.6726	0.2120
b2	0%	0.0002	0.0138	0.0301	-0.0025	0.0165	0.0310	0.0002	0.0138	0.1471	-0.0002	0.0139	0.1768
	25%	-0.0011	0.0153	0.0196	-0.0035	0.0167	0.0201	-0.0011	0.0153	0.1467	0.0200	0.0165	0.1787
	50%	0.0036	0.0211	0.0140	0.0022	0.0224	0.0149	0.0036	0.0211	0.1466	0.0615	0.0288	0.1807
	90%	0.0582	0.2104	0.0108	0.0558	0.2097	0.0115	0.0574	0.2116	0.1516	1.2138	4.2837	0.2120
b3	0%	-0.0027	0.0147	0.0301	-0.0047	0.0175	0.0310	-0.0027	0.0147	0.1471	-0.0031	0.0147	0.1768
	25%	-0.0047	0.0180	0.0196	-0.0066	0.0200	0.0201	-0.0047	0.0180	0.1467	0.0161	0.0192	0.1787
	50%	-0.0043	0.0238	0.0140	-0.0057	0.0265	0.0149	-0.0043	0.0238	0.1466	0.0522	0.0307	0.1807
	90%	0.0400	0.1923	0.0108	0.0374	0.1936	0.0115	0.0388	0.1930	0.1516	1.1657	4.0629	0.2120
							Logistické rozdělení						
b1	0%	0.0158	0.1199	0.0255	0.0186	0.1278	0.0262	0.0157	0.1198	0.1474	0.0162	0.1185	0.1774
	25%	0.0173	0.1433	0.0188	0.0177	0.1476	0.0200	0.0173	0.1433	0.1474	0.0479	0.1538	0.1772
	50%	0.0289	0.1910	0.0154	0.0273	0.1944	0.0168	0.0289	0.1911	0.1475	0.1064	0.2364	0.1793
	90%	0.0676	0.6341	0.0114	0.0729	0.5816	0.0121	0.0679	0.6354	0.1490	0.4405	1.5657	0.1871
b2	0%	0.0074	0.0309	0.0255	0.0081	0.0329	0.0262	0.0074	0.0309	0.1474	0.0074	0.0305	0.1774
	25%	0.0173	0.1433	0.0188	0.0125	0.0417	0.0200	0.0105	0.0394	0.1474	0.0433	0.0442	0.1772
	50%	0.0157	0.0496	0.0154	0.0170	0.0509	0.0168	0.0158	0.0496	0.1475	0.0917	0.0670	0.1793
	90%	0.0714	0.2008	0.0114	0.0622	0.1800	0.0121	0.0713	0.2009	0.1490	0.4293	0.6175	0.1871
b3	0%	0.0008	0.0351	0.0255	0.0016	0.0369	0.0262	0.0008	0.0351	0.1474	0.0012	0.0347	0.1774
	25%	0.0054	0.0428	0.0188	0.0050	0.0444	0.0200	0.0054	0.0428	0.1474	0.0373	0.0476	0.1772
	50%	0.0120	0.0609	0.0154	0.0091	0.0608	0.0168	0.0119	0.0609	0.1475	0.0884	0.0792	0.1793
	90%	0.0409	0.1992	0.0114	0.0342	0.1836	0.0121	0.0409	0.1991	0.1490	0.3859	0.5475	0.1871

Tabulka A.2 Výsledky na datech s 200 pozorováními, zaokrouhleno na 4 desetinná místa

		Gehan			Jin			Gehan-Poly			Gehan-Heller		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time	Bias	MSE	Avg Time	Bias	MSE	Avg Time
							Normální rozdělení						
b1	0%	-0.0016	0.0215	0.3843	-0.0029	0.0206	0.3902	-0.0016	0.0215	0.1630	-0.0017	0.0214	0.2947
	25%	-0.0001	0.0263	0.2218	-0.0009	0.0254	0.2281	-0.0001	0.0263	0.1664	0.0114	0.0272	0.3051
	50%	-0.0002	0.0334	0.1235	-0.0006	0.0322	0.1296	-0.0002	0.0334	0.1708	0.0335	0.0374	0.3359
	90%	0.0248	0.1449	0.0340	0.0274	0.1420	0.0402	0.0248	0.1451	0.1769	0.4009	0.5844	0.3488
b2	0%	-0.0016	0.0053	0.3843	-0.0015	0.0051	0.3902	-0.0016	0.0053	0.1630	-0.0015	0.0052	0.2947
	25%	-0.0026	0.0075	0.2218	-0.0023	0.0071	0.2281	-0.0026	0.0075	0.1664	0.0099	0.0078	0.3051
	50%	-0.0023	0.0106	0.1235	-0.0021	0.0102	0.1296	-0.0023	0.0106	0.1708	0.0315	0.0126	0.3359
	90%	0.0017	0.0501	0.0340	0.0007	0.0467	0.0402	0.0017	0.0500	0.1769	0.3404	0.2467	0.3488
b3	0%	-0.0006	0.0052	0.3843	-0.0010	0.0050	0.3902	-0.0006	0.0052	0.1630	-0.0006	0.0052	0.2947
	25%	0.0020	0.0072	0.2218	0.0013	0.0069	0.2281	0.0020	0.0072	0.1664	0.0142	0.0077	0.3051
	50%	0.0037	0.0100	0.1235	0.0024	0.0097	0.1296	0.0037	0.0100	0.1708	0.0369	0.0124	0.3359
	90%	0.0087	0.0464	0.0340	0.0069	0.0442	0.0402	0.0087	0.0465	0.1769	0.3528	0.2549	0.3488
							Extrémní rozdělení						
b1	0%	0.0078	0.0279	0.3439	0.0048	0.0332	0.3508	0.0079	0.0279	0.1646	0.0074	0.0281	0.2963
	25%	0.0100	0.0297	0.1738	0.0073	0.0340	0.1798	0.0100	0.0297	0.1683	0.0251	0.0313	0.3163
	50%	0.0126	0.0358	0.0918	0.0082	0.0404	0.0981	0.0126	0.0358	0.1708	0.0566	0.0432	0.3264
	90%	0.0679	0.2437	0.0315	0.0624	0.2460	0.0377	0.1005	0.5011	0.1800	0.9686	3.3429	0.3855
b2	0%	0.0003	0.0068	0.3439	0.0006	0.0076	0.3508	0.0003	0.0068	0.1646	0.0002	0.0068	0.2963
	25%	0.0031	0.0076	0.1738	0.0024	0.0083	0.1798	0.0031	0.0076	0.1683	0.0190	0.0083	0.3163
	50%	0.0037	0.0109	0.0918	0.0035	0.0118	0.0981	0.0037	0.0109	0.1708	0.0473	0.0143	0.3264
	90%	0.0115	0.0491	0.0315	0.0130	0.0496	0.0377	0.0115	0.0490	0.1800	0.7062	0.9047	0.3855
b3	0%	-0.0012	0.0067	0.3439	-0.0030	0.0079	0.3508	-0.0012	0.0067	0.1646	-0.0014	0.0067	0.2963
	25%	0.0000	0.0081	0.1738	-0.0008	0.0091	0.1798	0.0000	0.0081	0.1683	0.0158	0.0087	0.3163
	50%	0.0001	0.0102	0.0918	-0.0014	0.0114	0.0981	0.0001	0.0102	0.1708	0.0428	0.0133	0.3264
	90%	0.0033	0.0517	0.0315	0.0005	0.0525	0.0377	0.0033	0.0517	0.1800	0.7062	0.9047	0.3855
							Logistické rozdělení						
b1	0%	-0.0011	0.0617	0.2646	-0.0024	0.0658	0.2718	-0.0011	0.0617	0.1658	-0.0014	0.0614	0.2972
	25%	-0.0033	0.0699	0.1576	-0.0040	0.0729	0.1640	-0.0033	0.0699	0.1670	0.0200	0.0740	0.3048
	50%	-0.0061	0.0873	0.1017	-0.0072	0.0911	0.1076	-0.0061	0.0873	0.1684	0.0517	0.1025	0.3056
	90%	0.0385	0.2961	0.0381	0.0354	0.2782	0.0444	0.0385	0.2960	0.1724	0.3214	0.6389	0.3382
b2	0%	0.0029	0.0162	0.2646	0.0012	0.0172	0.2718	0.0029	0.0162	0.1658	0.0026	0.0161	0.2972
	25%	0.0052	0.0209	0.1576	0.0033	0.0216	0.1640	0.0052	0.0209	0.1670	0.0289	0.0231	0.3048
	50%	0.0055	0.0274	0.1017	0.0043	0.0277	0.1076	0.0055	0.0274	0.1684	0.0630	0.0352	0.3056
	90%	0.0291	0.0885	0.0381	0.0253	0.0803	0.0444	0.0291	0.0885	0.1724	0.3006	0.2473	0.3382
b3	0%	0.0004	0.0138	0.2646	-0.0007	0.0151	0.2718	0.0004	0.0138	0.1658	0.0003	0.0138	0.2972
	25%	0.0033	0.0179	0.1576	0.0008	0.0191	0.1640	0.0033	0.0179	0.1670	0.0268	0.0197	0.3048
	50%	0.0047	0.0237	0.1017	0.0017	0.0245	0.1076	0.0047	0.0236	0.1684	0.0622	0.0303	0.3056
	90%	0.0269	0.0887	0.0381	0.0210	0.0810	0.0444	0.0269	0.0887	0.1724	0.2929	0.2386	0.3382

Tabulka A.3 Výsledky na datech se 400 pozorováními, zaokrouhleno na 4 desetinná místa

		Gehan			Jin			Gehan-Poly			Gehan-Heller		
	Cenzorování	Bias	MSE	Avg Time	Bias	MSE	Avg Time	Bias	MSE	Avg Time	Bias	MSE	Avg Time
							Normální rozdělení						
b1	0%	-0.0023	0.0110	5.6964	-0.0008	0.0103	5.7153	-0.0023	0.0110	0.2295	-0.0021	0.0109	0.7828
	25%	-0.0037	0.0127	3.2124	-0.0027	0.0121	3.2492	-0.0037	0.0127	0.2360	0.0055	0.0128	0.8104
	50%	-0.0038	0.0163	1.6211	-0.0040	0.0157	1.6424	-0.0038	0.0163	0.2408	0.0219	0.0178	0.8074
	90%	0.0102	0.0606	0.2092	0.0067	0.0548	0.2339	0.0102	0.0606	0.2731	0.2717	0.1938	0.9805
b2	0%	-0.0033	0.0026	5.6964	-0.0031	0.0025	5.7153	-0.0033	0.0026	0.2295	-0.0033	0.0026	0.7828
	25%	-0.0040	0.0034	3.2124	-0.0039	0.0033	3.2492	-0.0039	0.0034	0.2360	0.0054	0.0035	0.8104
	50%	-0.0043	0.0052	1.6211	-0.0046	0.0049	1.6424	-0.0043	0.0052	0.2408	0.0209	0.0059	0.8074
	90%	-0.0027	0.0222	0.2092	-0.0039	0.0210	0.2339	-0.0027	0.0222	0.2731	0.2399	0.1035	0.9805
b3	0%	-0.0038	0.0028	5.6964	-0.0038	0.0027	5.7153	-0.0038	0.0028	0.2295	-0.0038	0.0028	0.7828
	25%	-0.0022	0.0037	3.2124	-0.0021	0.0035	3.2492	-0.0022	0.0037	0.2360	0.0070	0.0038	0.8104
	50%	-0.0021	0.0053	1.6211	-0.0023	0.0051	1.6424	-0.0021	0.0053	0.2408	0.0230	0.0061	0.8074
	90%	-0.0029	0.0205	0.2092	-0.0030	0.0194	0.2339	-0.0029	0.0205	0.2731	0.2392	0.0988	0.9805
							Extrémní rozdělení						
b1	0%	-0.0033	0.0134	5.1487	-0.0017	0.0163	5.1751	-0.0033	0.0134	0.2319	-0.0033	0.0134	0.7776
	25%	-0.0045	0.0140	2.4321	-0.0040	0.0164	2.4580	-0.0045	0.0140	0.2354	0.0071	0.0144	0.8011
	50%	-0.0002	0.0172	1.1164	-0.0025	0.0198	1.1416	-0.0002	0.0172	0.2480	0.0325	0.0197	0.8357
	90%	0.0263	0.0686	0.1694	0.0243	0.0726	0.1970	0.0263	0.0685	0.2919	0.5632	0.5891	1.1930
b2	0%	0.0022	0.0033	5.1487	0.0023	0.0041	5.1751	0.0022	0.0033	0.2319	0.0022	0.0033	0.7776
	25%	0.0012	0.0038	2.4321	0.0007	0.0043	2.4580	0.0011	0.0038	0.2354	0.0128	0.0040	0.8011
	50%	0.0024	0.0049	1.1164	0.0014	0.0057	1.1416	0.0024	0.0049	0.2480	0.0345	0.0065	0.8357
	90%	0.0094	0.0232	0.1694	0.0078	0.0239	0.1970	0.0095	0.0232	0.2919	0.4934	0.3458	1.1930
b3	0%	0.0008	0.0037	5.1487	0.0013	0.0045	5.1751	0.0008	0.0037	0.2319	-0.0009	0.0311	0.9390
	25%	-0.0007	0.0040	2.4321	-0.0025	0.0047	2.4580	-0.0007	0.0040	0.2354	0.0186	0.0379	0.7889
	50%	0.0017	0.0048	1.1164	-0.0003	0.0055	1.1416	0.0017	0.0048	0.2480	0.0503	0.0523	0.8047
	90%	0.0086	0.0229	0.1694	0.0080	0.0243	0.1970	0.0086	0.0229	0.2919	0.2369	0.2629	0.8433
							Logistické rozdělení						
b1	0%	-0.0007	0.0311	3.9262	-0.0005	0.0338	3.9136	-0.0007	0.0311	0.2496	-0.0021	0.0075	0.9390
	25%	0.0010	0.0360	2.1702	-0.0002	0.0380	2.1911	0.0010	0.0360	0.2290	0.0160	0.0100	0.7889
	50%	0.0050	0.0447	1.2723	0.0058	0.0458	1.2943	0.0050	0.0447	0.2352	0.0422	0.0151	0.8047
	90%	0.0178	0.1312	0.2746	0.0178	0.1202	0.2956	0.0178	0.1312	0.2518	0.2169	0.1063	0.8433
b2	0%	-0.0019	0.0076	3.9262	-0.0031	0.0080	3.9136	-0.0019	0.0076	0.2496	-0.0021	0.0075	0.9390
	25%	-0.0021	0.0093	2.1702	-0.0032	0.0096	2.1911	-0.0021	0.0093	0.2290	0.0160	0.0100	0.7889
	50%	-0.0019	0.0121	1.2723	-0.0034	0.0122	1.2943	-0.0019	0.0121	0.2352	0.0422	0.0151	0.8047
	90%	0.0110	0.0386	0.2746	0.0085	0.0360	0.2956	0.0110	0.0386	0.2518	0.2169	0.1063	0.8433
b3	0%	0.0013	0.0073	3.9262	0.0005	0.0081	3.9136	0.0013	0.0073	0.2496	0.0012	0.0073	0.9390
	25%	0.0015	0.0088	2.1702	0.0014	0.0096	2.1911	0.0015	0.0088	0.2290	0.0196	0.0096	0.7889
	50%	0.0024	0.0115	1.2723	0.0021	0.0121	1.2943	0.0024	0.0115	0.2352	0.0469	0.0149	0.8047
	90%	0.0185	0.0432	0.2746	0.0132	0.0391	0.2956	0.0185	0.0432	0.2518	0.2281	0.1195	0.8433

Bibliografie

1. COX, D.R.; OAKES, D. *Analysis of Survival Data*. New York: Chapman a Hall/CRC., 1984. ISBN 9781315137438.
2. MARTINUSSEN, T.; SCHEIKE, T.H. *Dynamic Regression Models for Survival Data*. Switzerland: Springer, 2006. ISBN 978-0-387-20274-7.
3. KLEIN, John P.; MOESCHBERGER, Melvin L. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media, 2005. ISBN 038795399X, ISBN 9780387953991.
4. AALEN, Odd. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* [online]. 1978, roč. 6, č. 4, s. 701–726 [cit. 2023-10-30]. ISSN 00905364. Dostupné z: http://www.jstor.org/stable/2958850.
5. FLEMING, Thomas R.; HARRINGTON, David P. *Counting processes and survival analysis*. New York [u.a.]: Wiley, 1991. Wiley series in probability and mathematical statistics. ISBN 047152218X. Dostupné také z: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+019456689&sourceid=fbw_bibsonomy.
6. PER KRAGH ANDERSEN Ørnulf Borgan, Richard D. Gill; KEIDING, Niels. *Statistical Models Based on Counting Processes*. New York: Springer New York, NY, 1993. ISBN 978-1-4612-4348-9. Dostupné také z: https://link.springer.com/book/10.1007/978-1-4612-4348-9.
7. MATTHES, K. Brémaud, P.: Point Processes and Queues. Martingale Dynamics. Springer-Verlag, Berlin – Heidelberg – New York 1981, 373 S., 31 Abb., DM 88,–. *Biometrical Journal*. 1988, roč. 30, č. 2, s. 248–249. Dostupné z DOI: https://doi.org/10.1002/bimj.4710300220.
8. KAPLAN, E. L.; MEIER, Paul. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* [online]. 1958, roč. 53, č. 282, s. 457–481 [cit. 2023-11-21]. ISSN 01621459. Dostupné z: http://www.jstor.org/stable/2281868.
9. COX, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* [online]. 1972, roč. 34, č. 2, s. 187–220 [cit. 2024-01-12]. ISSN 00359246. Dostupné z: http://www.jstor.org/stable/2985181.
10. ANDERSEN, P. K.; GILL, R. D. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* [online]. 1982, roč. 10, č. 4, s. 1100–1120 [cit. 2024-01-13]. ISSN 00905364. Dostupné z: http://www.jstor.org/stable/2240714.
11. MOOLGAVKAR, Suresh H.; CHANG, Ellen T.; WATSON, Heather N.; LAU, Edmund C. An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies. *Risk Analysis*. 2018, roč. 38, č. 4, s. 777–794. Dostupné z DOI: https://doi.org/10.1111/risa.12865.

12. WEBB, Annabel; MA, Jun. Cox models with time-varying covariates and partly-interval censoring–A maximum penalised likelihood approach. *Statistics in Medicine*. 2023, roč. 42, č. 6, s. 815–833. Dostupné z DOI: https://doi.org/10.1002/sim.9645.

13. COX, D. R. Partial Likelihood. *Biometrika* [online]. 1975, roč. 62, č. 2, s. 269–276 [cit. 2024-02-05]. ISSN 00063444. Dostupné z: http://www.jstor.org/stable/2335362.

14. THERNEAU, T.M.; GRAMBSCH, P.M. *Modeling Survival Data: Extending the Cox Model*. Springer New York, 2013. Statistics for Biology and Health. ISBN 9781475732948. Dostupné také z: https://books.google.cz/books?id=oj0mBQAAQBAJ.

15. AALEN, Odd. A model for nonparametric regression analysis of counting processes. In: *Mathematical Statistics and Probability Theory: Proceedings, Sixth International Conference, Wisła (Poland), 1978*. Springer, 1980, s. 1–25.

16. BUCKLEY, Jonathan; JAMES, Ian. Linear Regression with Censored Data. *Biometrika* [online]. 1979, roč. 66, č. 3, s. 429–436 [cit. 2024-02-13]. ISSN 00063444. Dostupné z: http://www.jstor.org/stable/2335161.

17. MILLER, Rupert G. Least Squares Regression with Censored Data. *Biometrika* [online]. 1976, roč. 63, č. 3, s. 449–464 [cit. 2024-02-13]. ISSN 00063444. Dostupné z: http://www.jstor.org/stable/2335722.

18. WEI, L. J. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*. 1992, roč. 11, č. 14-15, s. 1871–1879. Dostupné z DOI: https://doi.org/10.1002/sim.4780111409.

19. JIN, Zhezhen; LIN, D. Y.; WEI, L. J.; YING, Zhiliang. Rank-Based Inference for the Accelerated Failure Time Model. *Biometrika* [online]. 2003, roč. 90, č. 2, s. 341–353 [cit. 2024-02-13]. ISSN 00063444. Dostupné z: http://www.jstor.org/stable/30042044.

20. PRENTICE, Ross L. Linear rank tests with right censored data. *Biometrika*. 1978, roč. 65, č. 1, s. 167–179.

21. TSIATIS, Anastasios A. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*. 1990, s. 354–372.

22. YING, Zhiliang. A large sample study of rank estimation for censored regression data. *The Annals of Statistics*. 1993, s. 76–99.

23. GEHAN, Edmund A. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. 1965, roč. 52, č. 1-2, s. 203–224.

24. FYGENSON, Mendel; RITOV, Ya'acov. Monotone estimating equations for censored data. *The Annals of Statistics*. 1994, s. 732–746.

25. KOENKER, Roger W; D'OREY, Vasco. Algorithm AS 229: Computing regression quantiles. *Applied statistics*. 1987, s. 383–393.

26. CHERNOZHUKOV, Victor; FERNÁNDEZ-VAL, Iván; MELLY, Blaise. Fast algorithms for the quantile regression process. *Empirical economics*. 2022, s. 1–27.

27. CHUNG, Matthias; LONG, Qi; JOHNSON, Brent A. A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems. *Statistics and computing*. 2013, roč. 23, s. 601–614.

28. NOCEDAL, Jorge; WRIGHT, Stephen J. *Numerical optimization*. Springer, 1999.

29. HUBER, Peter J. Robust estimation of a location parameter. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, s. 492–518.

30. FLETCHER, R. Newton-Like Methods. In: *Practical Methods of Optimization*. John Wiley Sons, Ltd, 2000, kap. 3, s. 44–79. ISBN 9781118723203. Dostupné z DOI: https://doi.org/10.1002/9781118723203.ch3.

31. HELLER, Glenn. Smoothed rank regression with censored data. *Journal of the American Statistical Association*. 2007, roč. 102, č. 478, s. 552–559.
32. YU, Lili; LIU, Liang; CHEN, Ding-Geng(Din). Weighted Least-Squares Method for Right-Censored Data in Accelerated Failure Time Model. *Biometrics* [online]. 2013, roč. 69, č. 2, s. 358–365 [cit. 2024-04-03]. ISSN 0006341X, ISSN 15410420. Dostupné z: `http://www.jstor.org/stable/44695252`.
33. ZENG, Donglin; LIN, D. Y. Efficient Estimation for the Accelerated Failure Time Model. *Journal of the American Statistical Association* [online]. 2007, roč. 102, č. 480, s. 1387–1396 [cit. 2024-04-03]. ISSN 01621459. Dostupné z: `http://www.jstor.org/stable/27639988`.
34. BROWN, Bruce M; WANG, You-Gan. Induced smoothing for rank regression with censored survival times. *Statistics in medicine*. 2007, roč. 26, č. 4, s. 828–836.
35. CHIOU, S; KANG, Sangwook; YAN, Jun. Rank-based estimating equations with general weight for accelerated failure time models: an induced smoothing approach. *Statistics in Medicine*. 2015, roč. 34, č. 9, s. 1495–1510.
36. JIN, Zhezhen; LIN, DY; YING, Zhiliang. On least-squares regression with censored data. *Biometrika*. 2006, roč. 93, č. 1, s. 147–161.
37. JIN, Zhezhen; YING, Zhiliang; WEI, LJ. A simple resampling method by perturbing the minimand. *Biometrika*. 2001, roč. 88, č. 2, s. 381–390.
38. JIN, Zhezhen; SHAO, Yongzhao; YING, Zhiliang. A Monte Carlo method for variance estimation for estimators based on induced smoothing. *Biostatistics*. 2015, roč. 16, č. 1, s. 179–188.
39. MILLER, Rupert; HALPERN, Jerry. Regression with censored data. *Biometrika*. 1982, roč. 69, č. 3, s. 521–531.
40. PÖTTER, U. A MULTIVARIATE BUCKLEY-JAMES ESTIMATOR. In: *Proceedings of the 6th Tartu Conference, Tartu, Estonia, 19-22 August 1999*. Ed. KOLLO, T.; TIIT, E.-M.; SRIVASTAVA, M. Berlin, Boston: De Gruyter, 2000, s. 117–132. ISBN 9783110944655. Dostupné z DOI: `doi:10.1515/9783110944655.117`.
41. RITOV, Y. Estimation in a Linear Regression Model with Censored Data. *The Annals of Statistics*. 1990, roč. 18, č. 1, s. 303–328. Dostupné z DOI: `10.1214/aos/1176347502`.
42. STARE, Janez; HARRELL JR, Frank E; HEINZL, Harald. BJ: An S-plus program to fit linear regression models to censored data using the Buckley–James method. *Computer methods and programs in biomedicine*. 2001, roč. 64, č. 1, s. 45–52.
43. KOLLO, T et al. A MULTIVARIATE BUCKLEY-JAMES ESTIMATOR U. PÖTTER. In: *Multivariate Statistics: Proceedings of the 6th Tartu Conference, Tartu, Estonia, 19–22 August 1999*. Walter de Gruyter, 2011, s. 117.
44. HILLIS, Stephen L. A comparison of three Buckley-James variance estimators. *Communications in Statistics-Simulation and Computation*. 1993, roč. 22, č. 4, s. 955–973.
45. PANG, Lei; LU, Wenbin; WANG, Huixia Judy. Local Buckley-James estimation for heteroscedastic accelerated failure time model. *Statistica Sinica*. 2015, roč. 25, s. 863.
46. WICKHAM, Hadley; DANENBERG, Peter; CSÁRDI, Gábor; EUGSTER, Manuel. *roxygen2: In-Line Documentation for R*. 2022. Dostupné také z: `https://CRAN.R-project.org/package=roxygen2`. R package version 7.2.3.
47. TERRY M. THERNEAU; PATRICIA M. GRAMBSCH. *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000. ISBN 0-387-98784-3.
48. SANDERSON, Conrad; CURTIN, Ryan. Armadillo: a template-based C++ library for linear algebra. *Journal of Open Source Software*. 2016, roč. 1, č. 2, s. 26.
49. EDDELBUETTEL, Dirk. Seamless R and C++ integration with Rcpp. 2013.

50. HUANG, Lin; JIN, Zhezhen. LSS: An S-Plus/R program for the accelerated failure time model to right censored data based on least-squares principle. *Computer Methods and Programs in Biomedicine*. 2007, roč. 86, č. 1, s. 45–50. ISSN 0169-2607. Dostupné z DOI: https://doi.org/10.1016/j.cmpb.2006.12.005.

51. JOHNSON, Brent A. Rank-based estimation in the 1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics*. 2009, roč. 10, č. 4, s. 659–666.

52. JOHN C. NASH. On Best Practice Optimization Methods in R. *Journal of Statistical Software*. 2014, roč. 60, č. 2, s. 1–14. Dostupné z DOI: 10.18637/jss.v060.i02.

53. ANDREWS, David. F.; HERZBERG, AM. *DATA: A Collection of Problems from Many fields for the Student and Research Worker*. Springer-Verlag, 1985.

54. WILCOX, Rand R. Chapter 10 - Robust Regression. In: WILCOX, Rand R. (ed.). *Introduction to Robust Estimation and Hypothesis Testing (Fifth Edition)*. Fifth Edition. Academic Press, 2022, s. 577–651. ISBN 978-0-12-820098-8. Dostupné z DOI: https://doi.org/10.1016/B978-0-12-820098-8.00016-6.

55. CHOI, Dongrak; BAE, Woojung; YAN, Jun; KANG, Sangwook. *A general model-checking procedure for semiparametric accelerated failure time models*. 2023. Dostupné z arXiv: 2305.11445 [stat.ME].

56. NOVÁK, Petr. Goodness-of-fit test for the accelerated failure time model based on martingale residuals. *Kybernetika*. 2013, roč. 49, č. 1, s. 40–59.

57. CHIOU, Sy Han; KANG, Sangwook; YAN, Jun. Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee. *Journal of Statistical Software*. 2014, roč. 61, č. 11, s. 1–23. Dostupné také z: https://doi.org/10.18637/jss.v061.i11.

58. WRIGHT, Linda; MURALEEDHARAN, G.; GUEDES SOARES, Carlos; LUCAS, Cláudia. CHARACTERISTIC AND MOMENT GENERATING FUNCTIONS OF GENERALISED EXTREME VALUE DISTRIBUTION (GEV). In: 2010, s. 269–276. ISBN 978-1-61728-655-1.

59. RESEARCH, Wolfram. *LogisticDistribution* [https://reference.wolfram.com/language/ref/LogisticDistribution.html]. 2016. [Accessed: 12-May-2024].

60. PORTNOY, Stephen; KOENKER, Roger. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*. 1997, roč. 12, č. 4, s. 279–300.

61. WILCOX, Rand R. *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Sv. 249. Springer, 2010.

Obsah příloh

- `readme.txt` ..... stručný popis obsahu média
- `aftsem_1.0.tar.gz` ..... build soubor balíčku, připravený na instalaci
- `src`
  - `aftsem` ..... zdrojové kódy implementace
  - `notebooks` ..... notebooky se simulační studií a analýzou reálných dat
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF