Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Martin Dungl

R-knihovna Robustbase jako obraz současné robustní statistiky

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Jan Dienstbier

Studijní program: obecná matematika

2007

Na tomto místě bych chtěl poděkovat vedoucímu práce, Mgr. Janu Dienstbierovi, za konzultace, cenné rady, velmi přátelský přístup a za to, že mi ukazoval cestu.

Dále děkuji Honzovi Olšinovi za vydatnou pomoc s LaTeXem.

Můj dík patří také rodičům a Lucce, že to se mnou během psaní práce vydrželi.

V neposlední řadě bych chtěl poděkovat ing. Boudovi za konzultace a morální podporu.

Prohlašuji, že jsem svou bakalářskou práci napsal(a) samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne Martin Dungl

Obsah

Název práce: R-knihovna Robustbase jako obraz současné robustní statistiky
Autor: Martin Dungl
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky MFF UK
Vedoucí bakalářské práce: Mgr. Jan Dienstbier
e-mail vedoucího: dienstbi@karlin.mff.cuni.cz

Abstrakt: Předložená práce popisuje užití robustních statistických metod v R-knihovně robustbase. Nejprve jsou představeny robustní metody a různými způsoby definován pojem robustnosti. Podrobněji jsou popsány M-odhady s grafy používaných ψ-funkcí. Dále je představena R-knihovna robustbase, z hlediska historie jejího vývoje i z hlediska jejího složení. Detailně je rozebrán M-odhad polohy *huberM* a odvozen jeho bod selhání. Ve třetí části je na několika příkladech v programu R ukázáno použití tohoto odhadu a jsou diskutovány výsledky, zejména ve srovnání s průměrem a mediánem.

Klíčová slova: robustní metody, R, robustbase, M-odhady

Title: The R-library Robustbase documenting the contemporary state of robust statistics
Author: Martin Dungl
Department: Department of Probability and Mathematical Statistics, MFF UK
Supervisor: Mgr. Jan Dienstbier
Supervisor's e-mail adress: dienstbi@karlin.mff.cuni.cz

Abstract: The contribution documents use of robust statistical methods in R-library robustbase. First we present basic ideas of robust statistical methods and define the concept of their robustness in different ways. We focus on M-estimators, describing basic theory and present some of mostly used ψ-functions. Next the project of R-robustbase is introduced with a view to its history and architecture. We detail the M-estimator *huberM* and count its breakdown point. In the third part of the work the practical use of this estimator is illustrated through several examples in R. The results of these examples are analyzed and confronted with results obtained by sample mean and median.

Keywords: robust methods, R, robustbase, M-estimators

Úvod

Při používání klasických parametrických statistických metod vždy vycházíme z jistého modelu. Na správné těchto metod pak požadujeme jisté předpoklady týkající se pozorovaných dat. Často to je nezávislost a stejné rozdělení jednotlivých pozorování, při regresi pak zpravidla normální rozdělení chyb. Mnohdy však tyto předpoklady splněny nejsou a použití klasických metod vede k chybným závěrům. Proč tomu tak je? Splnění těchto předpokladů je podmíněno našimi předběžnými úvahami o vzniku a způsobu získání dat. Takové úvahy však nemusejí být přesné. Příčina, proč data nesplňují naše předpoklady, může být systematická, vycházející z toho, že jsme použili špatný model, například takový, který příliš zjednodušuje realitu. Dále může být příčina svým způsobem náhodná, například pokud zpracovávaná data obsahují několik nepřirozeně odlehlých pozorování, která lze považovat za chybná (jsou způsobena například špatně provedeným měřením).

Bohužel je nesprávná úvaha, že pokud se data "takřka" shodují s modelem, budou závěry získané klasickou metodou, která funguje za platnosti modelu, správné. Mnohdy je vypovídací hodnota takto získaných výsledků naopak velmi malá. Toto bylo ověřeno zejména v poslední době, kdy není problém počítačově zpracovat velké množství dat. Nastává tedy otázka, jaké použít v těchto případech, kdy se realita mírně liší od modelu, statistické postupy. Zajímá nás, jaký lze zvolit obecnější model, který by zahrnoval možnost odchylek dat, a za jakých okolností lze stále použít klasické postupy, které bychom použili při úplné shodě dat s konkrétnějším modelem. V případě, že by tyto klasické postupy vedly k mylným závěrům, zajímá nás, jaké jiné postupy, odolnější na nesplnění základních předpokladů, lze použít.

Tato základní úvaha nás vede k *robustním statistickým metodám.* Volně řečeno, je *robustnost* statistické metody odolnost výsledků získaných touto metodou vzhledem odchylkám jednotlivých pozorování od předpokládaného rozdělení. Robustní jsou tedy ty statistické metody, které si v okolí nějakého základního rozdělení zachovávají svoji optimalitu. Robustní metody mají základ již v 19. století, doopravdy se však začaly rozvíjet až v 60.letech minulého století, velkou zásluhu na jejich rozvoji má John Tukey. V této době byly vypracovány matematicky dobře založené M-odhady, L-odhady a R-odhady.

Jelikož jsou robustní statistické metody zpravidla výpočetně mnohem náročnější nežli tradiční parametrické metody, došlo k jejich bouřlivému rozvoji a rutinnímu používání až s nástupem počítačů na konci minulého století. Začaly vznikat také nové postupy, které tolik nevycházejí ze striktní teorie o třídách robustních odhadů, často jsou založeny na využití hrubé výpočetní síly.

Jedním z nejoblíbenějších současných statistických programů je volně šiřitelný program *R*. Není proto překvapivé, že robustní metody byly implementovány i zde. K pokusu o sjednocení jednotlivých robustních funkcí došlo v roce 2006, kdy byla vydána R-knihovna *robustbase*. Většina metod obsažených v této knihovně je založena jednak na použití M-odhadů, jednak na použití výše zmíněných postupů, které nemají tak pevný matematický základ.

V první části této práce se budu zabývat robustními statistickými metodami - jak obecně, tak na příkladech. Ve druhé části se zaměřím na knihovnu robustbase. Budu se zabývat jak procesem jejího vzniku (je dílem statistiků z celého světa, kteří do ní přidávali vlastní naprogramované funkce), tak jejím složením. Budu hodnotit robustbase z různých úhlů pohledu. Pokusím se též ilustrovat, jakým způsobem jsou do ní implementovány metody popsané v první kapitole, na příkladu Huberova M-odhadu polohy. Jako poslední kapitolu této práce uvádím krátkou simulaci v programu R. Na tři sady dat, získané různými způsoby, aplikuji různé odhady polohy a diskutuji výsledky. Ilustruji zde, že použití robustních metod je v praxi skutečně opodstatněné. Uvádím zdrojový kód v R, demonstrující snadnost použití obsahu robustbase v praxi.

Na konci úvodu si dovolím drobnou jazykovou poznámku. Ačkoliv v názvu práce je uvedeno Robustbase s velkým R, používám v práci malé první písmeno. Důvodem je fakt, že se jedná o knihovnu programu R, který je citlivý na velká a malá písmena. A zde vystupuje robustbase s malým r, což budu respektovat.

Kapitola 1

Úvod do robustních metod

1.1 Ilustrační příklad

Maronna a kol. uvádí v [1] následující příklad, který vhodně ilustruje, co si lze představit pod pojmem robustnost.

Příklad: *Uvažujme následujících 24 naměřených hodnot obsahu mědi v celozrnné mouce uspořádaných podle velikosti (údaje jsou v počtu částic na milion okolních).*

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90	3.03	3.03	3.10	3.37
3.40	3.40	3.40	3.50	3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

Hodnota 28.95 je výrazně vyšší než ostatní hodnoty, proto lze předpokládat, že se jedná o chybu měření. Tato odchylka má nezanedbatelný vliv na odhad střední hodnoty a směrodatné odchylky obsahu mědi. Buď $x = (x_1, \ldots, x_{24}) = (2.20, \ldots, 28.95)$ uvedený soubor pozorování. Pak výběrový průměr a výběrová směrodatná odchylka

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

nabývají hodnot $\overline{x} = 4.28$ a $s = 5.30$. Zjevně není tímto způsobem dobře odhadnuta poloha ani měřítko. Pokud bychom odstranili podezřelou hodnotu 28.95, dostaneme $\overline{x} = 3.21$ a $s = 0.69$.

Dostatečně odlehlou volbou jediného pozorování mohou $\overline{x}$ a s nabývat neomezených hodnot. Lze tedy říci, že jedna odchylka má *neomezený vliv* na tyto dvě statistiky. Proto jsou průměr a výběrová směrodatná odchylka velmi *nerobustními* statistikami. Protože po vymazání x_{24} se $\overline{x}$ a s chovají mnohem rozumněji, nabízí se možnost, metodicky vyhledávat a odstraňovat odlehlá pozorování. Obecně lze říci, že je to lepší než nic, přesto ovšem vznikají problémy:

- Jak lze ospravedlnit vymazání konkrétního pozorování? Kdy už je dostatečně odlehlé? Nehrozí, že by se vymazalo dobré pozorování?

- Jelikož by podobný postup byl ryze subjektivní, nebylo by možno předvídat jeho chování jako celku a porovnávat získané výsledky.

Zvolím-li za odhad polohy výběrový medián $\widetilde{x}$, dostanu $\widetilde{x} = 3.38$ pro výběr s hodnotou 28.95 a $\widetilde{x} = 3.37$ pro výběr bez této hodnoty. Podobně pomocí mediánu lze definovat pro odhad měřítka analogii s jako $\widetilde{s} = \frac{\mathrm{Med}_{\{|x_i - \widetilde{x}|,\ i=1,\dots,n\}}}{t}$, kde $t \approx 0.6745$ je konstanta zaručující, že pro výběr n prvků z normálního rozdělení je $\widetilde{s} \to s$ pro $n \to \infty$ (veličina $\widetilde{s}$ bývá někdy označována jako $\mathrm{mad}(x)$). V našem případě dostáváme $\widetilde{s} = 0.53$ resp. $\widetilde{s} = 0.50$.

Při libovolné, jakkoliv odlehlé, volbě x_{24} může $\widetilde{x}$ nabývat v našem případě pouze hodnot z intervalu $[\frac{x_{11}+x_{12}}{2}, \frac{x_{12}+x_{13}}{2}]$. Délka tohoto intervalu je v našem případě 0.15. Vidíme tedy, že vliv jednoho pozorování na $\widetilde{x}$ je velmi omezený. Analogicky je velmi omezený vliv jednoho pozorování na $\widetilde{s}$. Proto jsou $\widetilde{x}$ a $\widetilde{s}$ velmi *robustními* statistikami.

V následujících dvou podkapitolách dám pojmu robustnost matematické základy a pokusím se robustnost měřit. Za tímto účelem je výhodné pojmout odhad jako statistický funkcionál, tím se budu zabývat v podkapitole 1.2. Opírám se přitom zejména o publikaci [3]. Poté, prostřednictvím obecnějšího funkcionálního počtu, bude možno robustnost odhadu měřit, viz. podkapitola 1.3.

1.2 Odhad jako statistický funkcionál

Pokud odhad ztotožníme s funkcionálem na množině rozdělení, získáváme celou řadu nástrojů k teoretickému popisu robustnosti. Robustnost lze chápat jako stabilitu odhadu na větší množinu rozdělení. Bude nás proto zajímat hodnota funkcionálu nejen v rozdělení P, ale i v jeho okolí. Důležitými nástroji, které umožňují stabilitu odhadu coby statistického funkcionálu přirozeným způsobem porovnávat, jsou jeho Gâteauxova derivace a charakteristiky na ní založené. V této podkapitole se pokusím tyto základní ideje systematizovat, přesto podrobné teoretické úvahy v této věci zdaleka přesahují náplň této práce. Pro podrobnosti odkazuji na [3] a [4].

Předpokládejme, že X je náhodná veličina s rozdělením $P_\theta \in \mathcal{P}$, kde $\mathcal{P}$ je nějaká rodina rozdělení a θ nějaký reálný parametr. Mějme nějaký funkcionál $T : \mathcal{P} \to \mathbb{R}$.

Buď $\mathbf{X} = (X_1, \dots, X_n)$ náhodný výběr z rozdělení P_θ s empirickým rozdělením pravděpodobnosti P_n definovaným jako

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \in A} \qquad \text{pro } A \in \mathcal{B}.$$

Předpokládejme, že výraz $T(P_n)$ má smysl.

Nakonec vznesme požadavek $\theta = T(P_\theta)$, kterému se říká *fisherovská konzistence* funkcionálu T. $T(P_n)$ lze pak použít jako odhad parametru θ.

Příkladem T je $T(P_\theta) = \int_{\mathbb{R}} x dP = \mathbb{E}_P X$ s analogií $T(P_n) = \int_{\mathbb{R}} x dP_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$. Druhá rovnost platí, protože P_n je vlastně čítací míra. T se často vyskytuje ve tvaru $T(\mu) = \int_{\mathbb{R}} f(x, \mu) d\mu$ pro pravděpodobnostní míru μ a nějakou funkci f. Tato třída funkcionálů je dostatečně bohatá a umožňuje značnou obecnost tak, aby $T(P_n)$ byl přirozeným

odhadem $\theta = T(P)$.

Poznámka: Ne každý přirozený odhad je fisherovsky konzistentní - např. rozptyl rozdělení P lze odhadovat dvěmi statistikami aplikovanými na náhodný výběr $\{X_i, i = 1, \ldots, n\}$ z tohoto rozdělení:

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{a} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

viz [2], str. 35. Fisherovsky konzistentním odhadem rozptylu je M_2, na rozdíl od s^2 však není nestranný. V robustních metodách je zpravidla fisherovská konzistence silnějším požadavkem, nežli nestrannost.

Pozorování: *Označme $\mathcal{P}$ systém všech rozdělení na měřitelném výběrovém prostoru $(\mathcal{X}, \mathcal{B})$, kde $\mathcal{B}$ je borelovská σ-algebra. Pak $\mathcal{P}$ je konvexní množina.*

Chceme-li vyjádřit robustnost funkcionálu T, je třeba znát jeho chování v okolí nějakého rozdělení P. Proto přistupujeme k zavedení *Gâteauxovy derivace* odhadu.

Definice: *Předpokládejme, že T je definován na $\mathcal{P}$. Gâteauxovu derivaci funkcionálu T podle P ve směru Q definujeme jako*

$$T'_Q(P) = \lim_{t \to 0+} \frac{T(P + t(Q - P)) - T(P)}{t},$$

má-li pravá strana smysl.

Poznámka k definici: Díky hlubším úvahám, opírajícím se o vybudování metriky na $\mathcal{P}$ a práci s $\mathcal{P}$ jako s metrickým prostorem a jiné matematické nástroje, je ve [3] dokázáno, že Gâteauxova derivace je definována pro velmi širokou škálu statistických funkcionálů. Dá se říci, že za "rozumných" okolností lze předpokládat, že je definována. Zároveň lze jednoduše dokázat, že

$$\sup_{Q \in \mathcal{P}} T'_Q(P) < \infty \quad \Leftrightarrow \quad \exists \epsilon > 0 : \quad T \text{ je omezené na otevřeném okolí } U_d(P, \epsilon),$$

vzhledem k libovolné metrice d na $\mathcal{P}$ (mají-li všechny výrazy v této ekvivalenci smysl). To tedy znamená, že Gâteauxova derivace T podle P skutečně popisuje chování funkcionálu T na okolí rozdělení P. Pro podrobnosti odkazuji na [3].

1.3 Jak měřit robustnost?

Obraťme pozornost od teoretických základů ke konkrétnímu měření robustnosti odhadů. Budeme přitom vycházet především z [3]. Základními nástroji používanými pro popis míry

robustnosti funkcionálu jsou *citlivost k přidání dalšího pozorování*, vlastnosti založené na *influenční funkci* a *bod selhání*.

1.3.1 Citlivost k přidání dalšího pozorování

Mějme dán vektor pozorování $\mathbf{X} = (X_1, \ldots, X_n)$ generující empirické rozdělení P_n. Přidejme další pozorování $Y \in \mathbb{R}$. Empirické rozdělení generované $(X_1, \ldots, X_n, Y)$ označme P_Y. Důležitý pak je rozdíl

$$T(P_Y) - T(P_n) = I(T, \mathbf{X}, Y).$$

Definice: *Citlivostí funkcionálu T k přidání dalšího pozorování při daných pozorováních $\mathbf{X} = (X_1, \ldots, X_n)$ rozumíme*

$$S(T, \mathbf{X}) = \sup_Y |I(T, \mathbf{X}, Y)|.$$

Lze lehce ověřit, že zvolíme-li za T např. střední hodnotu, je $S(T, \mathbf{X}) = \infty$, pro libovolný vektor $\mathbf{X}$. Průměr má tudíž nekonečnou citlivost. Naopak pro medián z výběru o lichém počtu prvků dostaneme citlivost omezenou maximem z rozdílů tří "prostředních" hodnot. Viz. oddíl 1.1.

1.3.2 Influenční funkce

Připomeňme, že pro $x \in \mathcal{X}$ je δ_x míra na $(\mathcal{X}, \mathcal{B})$ definovaná vztahem $\delta_x(A) = I_{x \in A}, A \in \mathcal{B}$.

Definice: *Nechť x je prvkem výběrového prostoru. Pak definujeme*

$$IF(x; T, P) = T'_{\delta_x}(P) (= \lim_{t \to 0+} \frac{T((1-t)P + t\delta_x) - T(P)}{t}).$$

$IF(x; T, P)$ nazveme influenční funkcí funkcionálu T v rozdělení P ve směru x.

V následujících vztazích se ukáže, že influenční funkce funkcionálu T za jistých předpokladů úzce souvisí s jeho citlivostí k přidání dalšího pozorování do výběru. Předně, platí $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, a tedy

$$\begin{aligned} I(T, \mathbf{X}, Y) = &\quad T\Big(\frac{1}{n+1}(\delta_Y + \sum_{i=1}^n \delta_{X_i})\Big) - T(P_n) = \\ = &\quad T\Big((1 - \frac{1}{n+1})P_n + \frac{1}{n+1}\delta_Y\Big) - T(P_n). \end{aligned}$$

Nechť P_n je empirické rozdělení dané n pozorováními z rozdělení P. Předpokládáme-li konvergenci $T(P_n)$ k $T(P)$ (ta lze předpokládat za poměrně obecných okolností, důkaz je opřen o použití centrální limitní věty, podrobnosti viz. [3]), lze psát

$$\begin{aligned} IF(Y;T,P) = & \lim_{t\to 0+} \frac{T((1-t)P + t\delta_Y) - T(P)}{t} = \\ = & \lim_{n\to\infty} \frac{T\left((1-\frac{1}{n+1})P_n + \frac{1}{n+1}\delta_Y\right) - T(P_n)}{\frac{1}{n+1}} = \lim_{n\to\infty}(n+1)I(T,\mathbf{X},Y). \end{aligned}$$

Definice: *Globální citlivostí funkcionálu T pro rozdělení P nazveme*

$$\gamma^* = \sup_{x\in\mathcal{X}} |IF(x;T,P)|.$$

Je-li $\mathcal{X} = \mathbb{R}$, pak lokální citlivostí T pro rozdělení P nazveme

$$\lambda^* = \sup_{x,y\in\mathbb{R}, x\neq y} \left| \frac{IF(y;T,P) - IF(x;T,P)}{y-x} \right|$$

1.3.3 Bod selhání

Bod selhání je pravděpodobně nejoblíbenější charakteristikou robustnosti odhadu. Intuitivně lze bod selhání popsat jako podíl odlehlých pozorování v souboru, který je třeba k neomezenému znehodnocení odhadu. Mějme dán vektor pozorování $\mathbf{X} = (X_1, \ldots, X_n)$. Označme obecně rozdělení generované vektorem pozorování $\mathbf{A}$ jako $P_{\mathbf{A}}$ a označme $\mathbf{X}^a, a \in \{0, \ldots, n\}$ množinu vektorů, které dostaneme z $\mathbf{X}$ po nahrazení a složek jinými čísly. Nechť m je nejmenší takové přirozené číslo, že

$$\sup_{\mathbf{Y}\in\mathbf{X}^m} |T(P_X) - T(P_Y)| = \infty.$$

Předpokládáme, že m existuje (což platí vždy, když se T opravdu chová jako odhad).

Definice: *Výběrový bod selhání odhadu T ve výběru $\mathbf{X} = (X_1, \ldots, X_n)$ se definuje jako $\epsilon_n^*(T, \mathbf{X}) = \frac{m}{n}$.*

Jestliže (alespoň pro dost velká n) nezávisí $\epsilon_n^*(T, \mathbf{X})$ na $\mathbf{X}$, lze dokázat, že existuje $\lim_{n\to\infty} \epsilon_n^*(T, \mathbf{X})$.

Definice: *Za situace z předchozího odstavce definujeme bod selhání odhadu T jako*

$$\epsilon^*(T) = \lim_{n\to\infty} \epsilon_n^*(T, \mathbf{X}).$$

Samozřejmě platí, že čím je větší hodnota bodu selhání odhadu T, tím je T robustnější.

Příklady:

Medián: Mějme vektor pozorování $\mathbf{X} = (X_1, \ldots, X_{2k-1})$, $X_1 \leq X_2 \leq \cdots \leq X_{2k-1}$, funkcionál T nechť přiřazuje medián výběru. Pak $T(P_{\mathbf{X}}) = X_k$. Protože

$$\sup_{\mathbf{Y} \in \mathbf{X}^{k-1}} |T(P_X) - T(P_Y)| = \max\{|X_k - X_1|, |X_k - X_{2k-1}|\}$$

a

$$\sup_{\mathbf{Y} \in \mathbf{X}^{k}} |T(P_X) - T(P_Y)| = \infty,$$

platí $\epsilon^*_{2k-1}(T, \mathbf{X}) = \frac{k}{2k-1}$. Stejnými argumenty lze dokázat, že $\epsilon^*_{2k}(T, \mathbf{X}) = \frac{k}{2k}$. Jelikož

$$\lim_{k \to \infty} \frac{k}{2k-1} = \lim_{k \to \infty} \frac{k}{2k} = \frac{1}{2},$$

je

$$\epsilon^*(T) = \frac{1}{2}.$$

Průměr: Buď T průměr. Pak $\epsilon^*_n(T, \mathbf{X}) = \frac{1}{n}$, tedy $\epsilon^*(T) = 0$.

Useknutý průměr: T buď takový funkcionál, který z vektoru pozorování setříděného podle velikosti odebere t% ($t \in [0, 50)$) nejnižších a nejvyšších pozorování a zbylé hodnoty zprůměruje. Dostaneme $\epsilon^*(T) = \frac{t}{100}$.

1.4 M-odhady

Opět uvažujeme situaci s vektorem pozorování $\mathbf{X} = (X_1, \ldots, X_n)$ generujícím empirické rozdělení P_n. Nechť $\mathbf{X}$ je realizace náhodného výběru z rozdělení P. Pak existuje několik, poměrně obecně stanovených, tříd fisherovsky konzistentních odhadů neznámého parametru rozdělení P. V poslední době se zejména při odhadu polohy pro své výhodné vlastnosti, jako je obecnost, účinnost a vysoký bod selhání, nejvíce prosazují tzv. *M-odhady*. M-odhady jsou, velmi zjednodušeně řečeno, zobecněním odhadu metodou maximální věrohodnosti. Jak ukážeme, jsou M-odhady v praxi nejpoužívanější třídou robustních odhadů. Uvidíme, že i struktura knihovny robustbase se opírá o tuto třídu odhadů. Ve druhé kapitole podrobně rozeberu Huberův M-odhad polohy.

Definice: *M-odhad T parametru θ rozdělení P je dán vztahem*

$$T(P) = \arg\min_{\theta \in \Theta} \mathbb{E}_P \rho(X, \theta),$$

tedy speciálně

$$T(P_n) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho(X_i, \theta),$$

kde $\rho : \mathbb{R} \times \Theta \to \mathbb{R}$ *je vhodně zvolená funkce.*

Pro srovnání a ilustraci, při odhadu metodou maximální věrohodnosti pro výběr z rozdělení s hustotou f_θ hledáme

$$\arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f_\theta(X_i) = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} (-\log f_\theta(X_i)),$$

jde tedy o M-odhad s funkcí $\rho(X_i, \theta) = -\log f_\theta(X_i)$.

Je-li funkce ρ spojitě diferencovatelná vzhledem k θ, je výhodné se zabývat funkcí $\psi(x, \theta) = \frac{\partial\rho(x,\theta)}{\partial\theta}$. V tomto případě totiž $\hat{\theta} = T(P_n)$ řeší rovnici

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0.$$

Předpokládáme, že tato rovnice má jen jedno řešení. V praxi se pro názornost a z výpočetních důvodů používají zejména odhady založené na ψ, tzv. *M-odhady typu* ψ. Spojitost ψ vzhledem k x není vyžadována.

Za robustní se považují ty M-odhady typu ψ, pro které je $\psi(x, \theta)$ stejně omezená vzhledem k x pro všechna θ.

V dalším textu se budu zabývat výhradně M-odhady polohy.

Příklad: Definujme

$$\rho(x, \theta) = \frac{(x-\theta)^2}{2}.$$

Dostáváme $\psi(x, \theta) = \theta - x$. Máme-li realizaci výběru $(X_1, \ldots, X_n)$, hledáme pro něj řešení rovnice $\sum_{i=1}^{n}(\theta - X_i) = 0$. Vidíme tedy, že $\hat{\theta} = \overline{X}$. Odhad střední hodnoty pomocí výběrového průměru je tedy také speciálním případem M-odhadu, definovaným funkcí $\psi(x, \theta) = \theta - x$. Podobně lze odlišnou volbou ψ dostat useknutý průměr či medián.

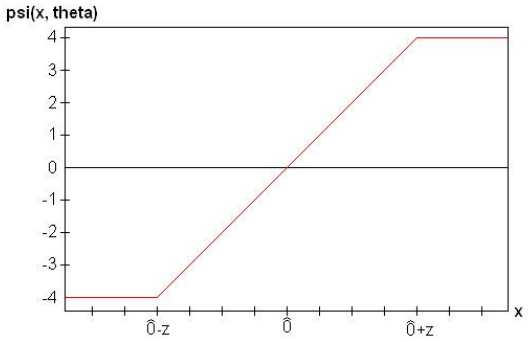
1.4.1 Grafy používaných ψ-funkcí

V praxi se používá zejména Huberova ψ-funkce - viz obrázek 1.1. Hodnota z lze volit následujícími nejběžnějšími způsoby:

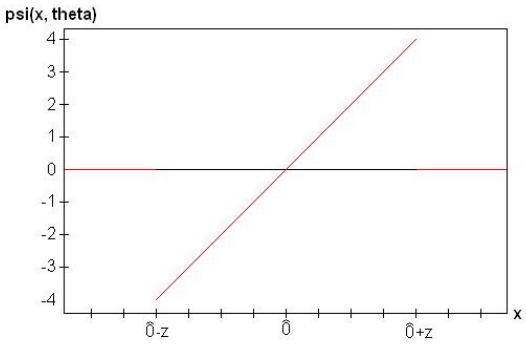
- Pevně - volíme zejména v případech, kdy známe očekávaný rozptyl dat.
- Přímo úměrně naměřenému měřítku - volíme nejčastěji. Ve druhé kapitole se budu zabývat funkcí *huberM*, která takto pracuje. Poznamenávám, že v této práci uvažuji měření měřítka rozdělení zejména pomocí směrodatné odchylky s a pomocí její robustní analogie $\widetilde{s}$.

- Pomocí kvantilů dat - jedná se o modifikovaný useknutý průměr, ne tolik radikální k odchylkám. V praxi se často volí 10%-ní kvantil. V tomto případě započítáváme místo 5% nejmenších a 5% největších hodnot hodnotu $\hat{\theta} - z$ resp. $\hat{\theta} + z$. Proces určení z je iterační.

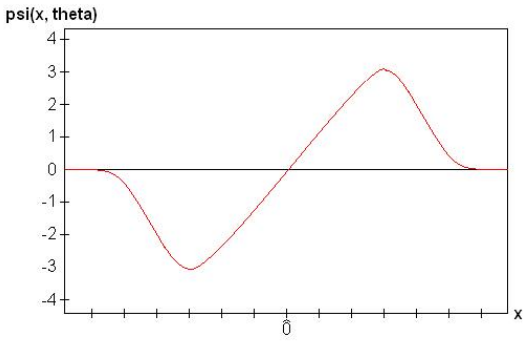
Obrázek 1.2 ukazuje ψ-funkci useknutého průměru. Kompromisem mezi uvedenými dvěma funkcemi je tzv. *Tukeyho biweight funkce*, viz. obrázek 1.3. Pro ilustraci ukazuji na obrázku 1.4 ψ-funkci mediánu a na obrázku 1.5 ψ-funkci průměru. Grafy dalších používaných ψ-funkcí mohou zájemci nalézt v [4].



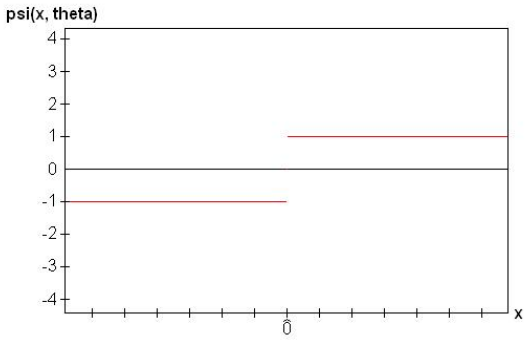
Obrázek 1.1: Huberova ψ-funkce



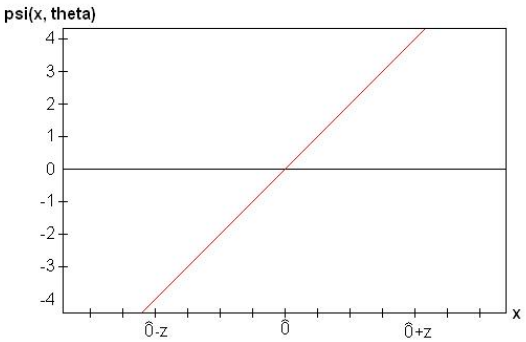
Obrázek 1.2: ψ-funkce useknutého průměru



Obrázek 1.3: Tukeyho biweight funkce



Obrázek 1.4: ψ-funkce mediánu



Obrázek 1.5: ψ-funkce průměru

Kapitola 2

R-knihovna robustbase

2.1 Prostředí R - vývoj a struktura

R je název jak pro programovací jazyk, tak pro softwarové prostředí používané ke statistickým výpočtům a jejich grafickým výstupům. R je volně šiřitelnou implementací programovacího jazyka S (existuje i komerční implementace S zvaná S-plus). Základní verze R 0.16 byla vytvořena v dubnu roku 1997 Rossem Ihakou a Robertem Gentlemanem na University of Auckland, Nový Zéland. R se postupně stalo velmi oblíbeným. Pro jednoduchost ovládání, velkou adaptabilitu, kvalitní grafické výstupy, široké programovací možnosti a jiné příznivé vlastnosti je nyní standardem zejména v akademickém prostředí. V softwarovém prostředí R se pracuje na příkazovém řádku.

R se šíří a obohacuje o nové funkce zejména zásluhou aktivity vědců a univerzitních studentů. R je volně ke stažení na síti zvané *CRAN*, což je zkratka pro *Comprehensive R Archive Network*. Volně přeloženo jde o síť, která obsahuje kompletní data projektu R. Jedná se o strukturu více než 60 serverů, tzv. zrcadel. Centrální bod této sítě je ve Vídni. Projekt R není tvořen jen jedním, pevně daným, souborem. Zatímco základní balíček obsahující samotný program vybavený nejdůležitějšími funkcemi je poměrně malý, existuje více než 1000 dalších volně dostupných knihoven (nebo též balíčků - z anglického a běžně používaného *package*) zaměřených na některé speciální funkce a odvětví statistiky. Je tomu tak díky tomu, že R je i programovací jazyk a umožňuje uživatelům vytvořit si statistické nástroje vhodné k řešení i velmi specifických problémů. Rozvoj statistiky v prostředí R tedy probíhá právě pomocí tvorby a zdokonalování uvedených balíčků. Jedním z nich je právě knihovna *robustbase*.

2.2 Historie a cíle projektu robustbase

Vzhledem k obecnosti prostředí R v něm byly programovány robustní metody již od jeho vzniku. Začlenění metod do R však nebylo systematické. Panoval značný zmatek - ostatně typický pro rozvoj volně šiřitelných programů - související s rychlým a nepřehledným nárůstem počtu balíčků. Bylo známo, že robustní metody jsou "někde" v R implemen-

továny, nebyly však soustředěny pohromadě a nebyla zaručena jejich kvalita. V důsledku toho pak nedostávaly metody robustní statistiky v praxi takový prostor, jaký by si zasloužily (tato dvě tvrzení jsou samozřejmě subjektivní - zde jde o mínění "robustních" statistiků). Tyto problémy byly náplní konference *Robust Statistics and R*, která se uskutečnila v italském Trevisu 26.-28.10.2005. Zúčastnilo se jí 77 vědeckých pracovníků a studentů, až na několik výjimek z Evropy, většinu tvořili Italové a Rakušané. Cílem této konference bylo sjednotit a zastřešit úsilí o začlenění robustních metod do R. Účastníci byli rozděleni do šesti následujících skupin podle zaměření: regresní modely, analýza více pozorování, časové řady, popisná statistika, velké soubory dat a ekonometrické modely. Každá skupina se zaměřila na problémy ve své oblasti zkoumání, rozdělila si práci a později (v horizontu měsíců) naprogramovali její členové příslušné funkce v R. Jelikož neexistovala v R žádná knihovna, která by se na robustní metody specializovala, byl schválen návrh ji vytvořit.

Na konferenci bylo dohodnuto založení internetové konference - mailing listu, původně určeného pro delegáty z Trevisa, později otevřeného pro veřejnost. Založil jej 5.11.2005 Švýcar Dr. Martin Mälcher z ETH Zürich. Na konferenci tak prostřednictvím tohoto mailing-listu v podstatě navázal projekt se stejným názvem Robust Statistics and R, jehož cílem je zejména umožnit uživatelům R používat kvalitně naprogramované robustní metody, soustředěné v jednom balíčku. V jeho čele neoficiálně stanul právě Dr.Mälcher. Z dalších zmíním zejména velké zásluhy Dr. Valentina Todorova z vídeňského řízení letového provozu, který dodal většinu sad dat a naprogramoval některé funkce.

Z mailing listu lze vyčíst některé zajímavé informace o vývoji projektu. Například otázkou bylo, jaké má mít nový balíček jméno. Pro jasné vymezení obsahu se většina shodla na tom, že název má obsahovat řetězec "robust". Proto nakonec došlo k hlasování mezi těmito čtyřmi návrhy - robusta, robustat, robustats a robustbase. Hlasovat mohli členové mailing-listu, každý měl 3 hlasy, které mohl rozdělit mezi návrhy. Hlasování se nakonec zúčastnilo 20 voličů a skončilo v poměru 5:1:9:45. Zvítězil tedy název *robustbase* - snad i proto, že se přišlo na to, že název robusta, který se dlouho zdál být nejvhodnějším, je již používán v botanice.

Knihovna robustbase byla vydána, tedy nahrána na CRAN, 9.2.2006 jako *robustbase 0.1-2* (nyní, v červenci 2007, lze stáhnout verzi 0.2-8). Původní verze obsahovala několik souborů dat, části spíše experimentálního kódu určeného k dalšímu dopracování a zejména robustní verze některých funkcí obsažených ve standardním balíku R. Tyto robustní funkce byly dílem nově naprogramované, dílem převzaté z jiných databází. V e-mailu z 10.2. Martin Mälcher uvažuje nad dalšími směry rozšíření robustbase. Zejména určuje funkce, které by měly být přidány, a to v souvislosti s nedokončenými úkoly pracovních skupin z Trevisa a s potřebami robustní statistiky. Dále vytyčuje ideální cíl implementovat metody použité v publikaci [1], která měla brzy vyjít.

Asi o měsíc později, v březnu 2006, skutečně vychází monografie [1], která ukazuje obraz současné robustní statistiky. Zabývá se i praktickým použitím robustních metod. Nacházejí se v ní odkazy na soubory dat a již naprogramované softwarové verze popisovaných metod, implementované však pouze do komerčního prostředí S-plus. Důvodem může být fakt, že zatímco v R byl vývoj nových komponent živelný a robustní knihovna dlouho neexistovala, S-plus mělo přehledně implementovány robustní metody v jednom balíčku již od roku 2000.

I díky vydání [1] došlo k rozvoji tohoto balíčku (příznačně nazvaného *Robust library*).

V roce 2006 se vyvíjely oba projekty paralelně vedle sebe. Nejnověji, na začátku března 2007, se pak na síti CRAN objevila R-knihovna *robust*, jejíž delší název zní "Robust: Insightful Robust Library". Insightful je firma vydávající softwarové prostředí S-plus, jako autoři knihovny robust jsou mimo jiné uvedeni i všichni autoři [1] - Maronna, Martin a Yohai. Robust je, stejně jako všechny komponenty R, vydána volně, pod všeobecnou veřejnou licencí GNU. Lze spekulovat o tom, že se Insightful rozhodl uvolnit tuto (jistě kvalitní) knihovnu právě proto, aby zde naprogramované metody zůstaly nejpoužívanějšími. Bohužel pokus o zhodnocení knihovny robust a podrobné porovnání s knihovnou robustbase je nad možnosti této práce, bylo by totiž potřeba jít až do hloubky obou knihoven. Přesto je na první pohled zřejmý fakt, že obsahová podobnost mezi robust a robustbase je značná. Je to patrné i z názvů funkcí, viz. dokumenace [5] a [6]. Jako příklad uvedu funkce *lmrob* a *lmRob*, které dělají velmi podobné věci (při simulaci v R jsem nepostřehl rozdíl), přesto nejsou kompatibilní (mají různé názvy). Přirozeně se tak vynořuje otázka, zda je šťastná současná situace, kdy dochází k paralelnímu vývoji hned dvou knihoven, které se obě tváří jako soubor základních robustních metod. Čas ukáže, co vydání knihovny robust s projektem Robust Statistics and R a knihovnou robustbase udělá.

2.3 Složení knihovny robustbase

V listopadu 2006 měla robustbase 88 elementů, rozdělitelných do následujících tří skupin:

1. Sady dat.
2. Robustní metody samotné.
3. Pomocné funkce robustních metod, technické funkce.

Podrobněji se budu zabývat jen některými elementy, zatímco většinu proberu pouze zběžně nebo zcela vynechám. V důsledku tohoto postupu nebude robustbase popsána v celé šíři. Je třeba si však uvědomit, že toto není cílem této bakalářské práce, nechci a ani nemohu nahradit manuály a oficiální dokumentaci. Mým cílem je přinést určité shrnutí a zhodnocení, čemuž tento postup není snad překážkou. Konkrétní položky nejsou zmíněny či podrobněji probrány kvůli jednomu ze tří následujících důvodů:

- Jsou zcela technické, nesouvisející s ideou robusbase, ani robustní statistiky. Do této kategorie spadá např. celá třetí skupina elementů, viz. výše.
- Svou strukturou nebo použitím jsou velmi podobné některému elementu, který popisuji. Z tohoto důvodu vynechávám většinu sad dat.
- Pro svou složitost překračují rámec bakalářské práce. Sem patří některé robustní metody, například robustní verze zobecněného lineárního modelu.

Zájemce o podrobnější rozbor knihovny odkazuji na oficiální dokumentaci [5].

2.3.1 Sady dat

V původní verzi robustbase tvořily sady dat 24 elementů, nyní jich je 30. Tyto kolekce mají kromě toho, že se samy o sobě mohou jevit zajímavé, za úkol umožnit uživatelům vyzkoušet použití klasických i robustních metod na příkladech z praxe. Některé sady dat obsahují odlehlá pozorování. Uvedu 9 příkladů:

- *alcohol* - rozpustnost různých alkoholů ve vodě (2001, 44 pozorování, 7 měřených veličin)
- *Animals2* - hmotnosti mozku a těla u 62 savců a 3 pravěkých ještěrů (1987, 65 pozorování, 2 měřené veličiny). Použití viz. podkapitola 3.3.
- *bushfire* - satelitně pořízené statistiky lesních požárů (1984, 38 pozorování, 5 měřených veličin)
- *milk* - chemická analýza mléka (1988, 85 pozorování, 8 měřených veličin)
- *NOxEmissions* - časové záznamy koncentrace oxidů dusíku u frekventované silnice (8088 pozorování, 4 měřené veličiny)
- *pension* - finanční stav penzijních fondů holandských firem (1981, 18 pozorování, 2 měřené veličiny)
- *starsCYG* - data o hvězdách ve směru souhvězdí labutě (1987, 47 pozorování, 2 měřené veličiny)
- *wood* - vliv "anatomických" vlastností druhu dřeviny na hustotu dřeva (1966, 20 pozorování, 5 měřených veličin)
- *hbk* - umělá sada dat složená ze 75 čtyřrozměrných pozorování. Je oblíbená pro testování regresních modelů, data jsou totiž zvolena tak, že různé regresní metody různě vyhodnotí odchylky (1984, 75 pozorování, 4 měřené veličiny),

více informací viz. [5].

2.3.2 Odhady

Odhady obsažené v knihovně robustbase lze dělit podle dvou kritérií. Jedno z nich je použití, druhým je použitá metoda. Budu se zde zabývat pouze prvním kritériem, použité metody zmíním v oddíle 2.3.3.

Po stránce použití lze najít v knihovně robustní regresní modely, odhady polohy, rozptylu a výpočty výběrových korelačních koeficientů. Regresní modely tvoří největší podíl na funkcích robustbase. Uvedu rámcový přehled nejdůležitějších funkcí:

- *lmrob* - Lineární model. Základní nástroj lineární regrese.

- *glmrob* - Zobecněný lineární model.
- *nlrob* - Nelineární regresní model.
- *ltsReg* - Regrese pomocí metody nejmenších oříznutých čtverců.
- *HuberM* - Huberův M-odhad polohy.
- *psiFunc* - Funkce sloužící ke konstrukci vhodné psi-funkce pro M-odhad.
- *wgt.himedian* - Vážený medián.
- *Qn, Sn* a *scaleTau2* - Různé metody robustních odhadů rozptylu.
- *covGk, covOGK* - Odhady výběrového korelačního koeficientu, založené na funkci *cov.rob* z balíčku MASS.

Pro podrobnosti ohledně syntaxe a použití jednotlivých metod v R odkazuji na [5], pro teoretické základy většiny z nich na [1].

2.3.3 Robustbase v kontextu robustní statistiky

V teoretických publikacích, jako např. [3], se popisují různé třídy robustních odhadů, jako jsou kromě M-odhadů ještě L-odhady a R-odhady. Pokud bychom chtěli ale odpovědět na otázku, co jsou to robustní metody, na základě odhadů prakticky implementovaných do knihovny robustbase, ukazuje se, že pojem M-odhadů by byl centrální. Po prostudování mailing listu skupiny *Robust statistics and R* si dovoluji tvrdit, že M-odhady jsou základním stavebním kamenem knihovny robustbase, více viz. [7]. O velké oblíbenosti M-odhadů svědčí i prostor jim věnovaný v prakticky zaměřené monografii [1].

Z tohoto důvodu se ve zbytku této kapitoly zaměřím na Huberův M-odhad polohy obsažený ve funkci *huberM*, kterou podrobně zanalyzuji. Velké množství dalších funkcí, jako např. regresní modely v čele s funkcí *lmrob*, je pak založeno právě na M-odhadech a mediánu. Z hlediska struktury práce bych měl o funkci *huberM* pojednávat v pododdílu oddílu 2.3.2, pro přehlednost a z důvodu adekvátního rozsahu však tento odhad zařadím jako samostatnou podkapitolu.

2.4 M-odhad a funkce *huberM*:

V praxi se pro odhad polohy často používá Huberova verze M-odhadu. Proto je zařazena do robustbase jako funkce *huberM*. Jde o M-odhad s tzv. *Huberovou ψ-funkcí*, viz obrázek 1.1. Jedná se o funkci

$$\psi(x,\theta) = z \cdot \operatorname{sgn}(x-\theta) \cdot \mathbb{I}_{x\in(-\infty,\theta-z)\bigcup(\theta+z,\infty)} + (x-\theta)\cdot \mathbb{I}_{x\in[\theta-z,\theta+z]},$$

kde z je předem zadaná konstanta. Připomínám, že pro vektor pozorování $\mathbf{X} = (X_1, \ldots, X_n)$ hledáme $\hat{\theta}$, aby

$$\sum_{i=1}^{n} \psi(X_i, \hat{\theta}) = \sum_{i=1}^{n} \left(z \cdot \mathrm{sgn}(X_i - \hat{\theta}) \cdot \mathbb{I}_{X_i \in (-\infty, \hat{\theta}-z) \bigcup (\hat{\theta}+z, \infty)} + (X_i - \hat{\theta}) \cdot \mathbb{I}_{X_i \in [\hat{\theta}-z, \hat{\theta}+z]} \right) = 0.$$

Funkce *huberM* počítá M-odhad daný uvedenou ψ-funkcí. Přesná podoba této funkce závisí na hodnotě konstanty z. Tato hodnota může být zadána uživatelem nebo ji může určit podle struktury dat pomocná funkce, zavolaná funkcí *huberM*. Podrobnosti zmíním dále. Pro představu, zatímco pro dostatečně velká z bude výstupem funkce průměr ze zadaných hodnot, pro malá z bude výsledek naopak blízký mediánu. Funkce *huberM* pracuje iteračně. Algoritmus lze popsat následovně po krocích:

1. V prvním kroku l-té iterace zvolíme za střed určitý bod $\hat{\theta}_l$.
2. Ve druhém kroku zkonstruujeme soubor hodnot $\mathbf{T} = (T_1, \ldots, T_n)$ následovně:

$$T_i = (\hat{\theta}_l + z \cdot \mathrm{sgn}(X_i - \hat{\theta}_l)) \cdot \mathbb{I}_{|X_i - \hat{\theta}_l| > z} + X_i \cdot \mathbb{I}_{|X_i - \hat{\theta}_l| \leq z}$$

3. Ve třetím kroku položíme $\hat{\theta}_{l+1} = \overline{\mathbf{T}}$ a přejdeme k 1.kroku $l+1$. iterace.

Posloupnost $\hat{\theta}_l$ konverguje nezávisle na vstupních hodnotách, to však nebudu dokazovat. Jestliže

$$|\hat{\theta}_{l+1} - \hat{\theta}_l| < \texttt{tol}$$

pro danou konstantu `tol`, algoritmus skončí s výsledkem $\hat{\theta} = \hat{\theta}_{l+1}$. V následujících odstavcích se funkci *huberM* věnuji z pohledu programu R a z praktičtějšího hlediska.

Povinným vstupem funkce *huberM* je vektor hodnot $\mathbf{X}$. Dále existují následující volitelné vstupy, hodnoty různých parametrů funkce *huberM*:

- `mu` - neboli $\hat{\theta}_1$, viz. výše
- `tol` - kritérium konvergence $(\hat{\theta}_i)_{i=1}^{\infty}$, viz. výše
- `k` - prostředek pro zadání z, ukazatel robustnosti funkcionálu
- `s` - prostředek pro výpočet z, ukazatel předpokládaného měřítka
- `weights` - vektor délky n udávající váhy jednotlivých pozorování
- `warn0scale` - logická hodnota, pokud je zapnuta, $s = 0$ a $n > 1$, výpočet nebude proveden.

Zmínil jsem, že hodnoty `k` a `s` slouží k výpočtu z. Jednoduše platí, že $z = s \cdot k$. Každý volitelný vstup má výchozí hodnotu (tj. takovou, se kterou se počítá, pokud uživatel nestanoví jinak). Rozdíl mezi `s` a `k` je ten, že zatímco `k` má udávat robustnost odhadu a pro výchozí hodnotu `k` platí `k` = 1.5, `s` má udávat, volně řečeno, jakési předpokládané měřítko. Ze vstupních dat se spočte jednoduše: `s` = $\widetilde{s}$ (připomínám, že $\widetilde{s} = \frac{\mathrm{Med}_{\{|X_i - \widetilde{\mathbf{X}}|,\ i=1,\ldots,n\}}}{t}$, kde $t \approx 0.6745$, viz kap. 1.1; v prostředí R se $\widetilde{s}$ spočte pomocí příkazu `mad`). V praxi se osvědčilo, že tato volba `s` vede k ψ-funkci zaručující dostatečnou robustnost i efektivitu odhadu. Další výchozí hodnoty:

- `mu` = $\widetilde{\mathbf{X}}$
- `tol` = 10^{-6}
- `warn0scale` = "FALSE"
- `weights` = $(1, \ldots, 1)$.

2.4.1 Robustnost odhadu při použití funkce *huberM*

Charakteristiky robustnosti odhadu pomocí funkce *huberM* úzce korespondují se zvolenými parametry této funkce. Pokud bychom za `mu` zvolili průměr a za `s` směrodatnou odchylku, nebyl by zkonstruovaný M-odhad robustní. Původní nastavení funkce *huberM* však napovídá, jaké použití se považuje v robustní statistice za vhodnější. V tomto případě, kdy volíme za `mu` výběrový medián a za `s` veličinu $\widetilde{s}$, připomíná naopak tento odhad svou robustností tyto statistiky. Přitom je, na rozdíl od nich, výrazně efektivnější. Velkou robustnost takto nastaveného Huberova odhadu budu lustrovat na jeho bodu selhání, který vypočtu.

Bod selhání

Mějme vektor pozorování $\mathbf{X} = (X_1, \ldots, X_n)$ generující empirické rozdělení $P_{\mathbf{X}}$. Mějme M-odhad $\hat{\theta}$ středu $P_{\mathbf{X}}$ získaný pomocí výše popsané funkce *huberM* bez volitelných parametrů upravených uživatelem (tedy zejména platí $z = 1.5 \cdot \widetilde{s}$ a $\hat{\theta}_1 = \widetilde{\mathbf{X}}$) a položme $T(P_{\mathbf{X}}) = \hat{\theta}$.

Lemma: *Nechť* $\hat{\theta}_1 = \widetilde{\mathbf{X}}$. *Pak* $\hat{\theta} \in (\hat{\theta}_1 - 2z, \hat{\theta}_1 + 2z)$.

Důkaz: Bez újmy na obecnosti položme $\hat{\theta}_1 = 0$ a $z = 1$. Díky symetrii stačí dokázat pouze, že $\hat{\theta} < 2$.

Vzhledem k definici mediánu platí pro množinu $A \subset \{X_1, \ldots, X_n\}$ definovanou vztahem $X_j \in A \Leftrightarrow X_j \leq 0$, že $|A| \geq \lceil \frac{n}{2} \rceil$. Předpokládejme, že $\hat{\theta}_i \geq 1$. Pokud by tomu tak bylo, nahradil by se každý prvek A při výpočtu $\hat{\theta}_{i+1}$ prvkem

$$T_j = (\hat{\theta}_i + z \cdot \mathrm{sgn}(X_j - \hat{\theta}_i)) \cdot \mathbb{I}_{|X_j - \hat{\theta}_i| > z} = \hat{\theta}_i - 1,$$

viz. popis 2.kroku algoritmu *huberM*. Buď $B \subset \{X_1, \ldots, X_n\}$ množina definovaná vztahem $X_j \in B \Leftrightarrow X_j \geq \hat{\theta}_i$. Prvky množiny B budou při výpočtu $\hat{\theta}_{i+1}$ nahrazeny prvky $T_j \in$

$[\hat{\theta}_i, \hat{\theta}_i + 1]$. Protože A a B jsou disjunktní a $|A| \geq \lceil \frac{n}{2} \rceil$, je $|B| \leq |A|$. Odtud vidíme, že $\hat{\theta}_{i+1} = \overline{\mathbf{T}} \leq \hat{\theta}_i$.

Dokázali jsme tedy, že jestliže $\hat{\theta}_i \geq 1$, tak $\hat{\theta}_{i+1} \leq \hat{\theta}_i$. Pokud $\hat{\theta}_i < 1$, tak vzhledem k hodnotě $z = 1$ je $\hat{\theta}_{i+1} < 2$. Z uvedených úvah již lehce vyplývá, že $\hat{\theta} < 2$.

Věta: *$\epsilon^*(T)$ existuje a platí $\epsilon^*(T) = \frac{1}{2}$.*

Důkaz: Nechť $n = 2k + 1$, $k \in \mathbb{N}$ a $X_1 \leq X_2 \leq \cdots \leq X_{2k+1}$. Pak $\hat{\theta}_1 = \widetilde{\mathbf{X}} = X_{k+1}$. Nejprve dokážeme, že při libovolné změně k pozorování zůstane jak $\widetilde{\mathbf{X}}$, tak $\widetilde{s}$ omezené bez ohledu na velikost této změny. Poté ukážeme, že změna $k+1$ pozorování již znehodnocuje odhad a limitním přechodem přejdeme k $\epsilon^*(T)$.

Jak jsem ukázal v části 1.3.3, platí pro výběrový medián L $\epsilon^*_{2k+1}(L, \mathbf{X}) = \frac{k+1}{2k+1}$. Nechť $\mathbf{Y}$ je libovolný vektor, získaný z vektoru $\mathbf{X}$ pozměněním právě k hodnot. Z předchozího platí, že rozdíl výběrových mediánů $a = |\widetilde{\mathbf{X}} - \widetilde{\mathbf{Y}}|$ je omezený nezávisle na této změně. Označme J funkcionál, přiřazující empirickému rozdělení příslušné $\widetilde{s}$ a položme

$$\mathbf{U} = (U_1, \ldots, U_{2k+1}): \quad U_i = |X_i - \widetilde{\mathbf{X}}|, \quad i = 1, \ldots, 2k+1,$$

$$\mathbf{V} = (V_1, \ldots, V_{2k+1}): \quad V_i = |Y_i - \widetilde{\mathbf{Y}}|, \quad i = 1, \ldots, 2k+1.$$

Platí

$$|J(P_{\mathbf{X}}) - J(P_{\mathbf{Y}})| = |\frac{\widetilde{\mathbf{U}}}{t} - \frac{\widetilde{\mathbf{V}}}{t}| = \frac{|\widetilde{\mathbf{U}} - \widetilde{\mathbf{V}}|}{t}.$$

Jestliže $X_i = Y_i$, je $|U_i - V_i| \leq a$. Buď $\mathbf{W}$ takový $(2k+1)$-složkový vektor, že $W_i = V_i$, jestliže $X_i = Y_i$ a $W_i = U_i$ jinak. Vzhledem k tomu, že $\mathbf{V}$ a $\mathbf{W}$ se liší v nejvýše k hodnotách a bod selhání mediánu je $k+1$, je $b = |\widetilde{\mathbf{V}} - \widetilde{\mathbf{W}}|$ nezávislý na velikosti odchylky těchto k hodnot. Dále platí $|\widetilde{\mathbf{U}} - \widetilde{\mathbf{W}}| \leq a$. Dostáváme odhad

$$|\widetilde{\mathbf{U}} - \widetilde{\mathbf{V}}| \leq |\widetilde{\mathbf{U}} - \widetilde{\mathbf{W}}| + |\widetilde{\mathbf{W}} - \widetilde{\mathbf{V}}| \leq a + b,$$

tedy po změně k pozorování je i tento rozdíl omezený nezávisle na velikosti této změny. V posledním kroku jsme vlastně dokázali, že $\epsilon^*_{2k+1}(J, \mathbf{X}) > \frac{k}{2k+1}$.

Dokázali jsme, že pokud změníme libovolně k pozorování, $\hat{\theta}_1$ i $\widetilde{s}$ se změní pouze omezeně bez závislosti na velikosti této změny. Pokud si uvědomíme, co to znamená pro konstrukci odhadu a užijeme výše uvedené lemma, dostaneme, že ani $\hat{\theta}$ se nemůže neomezeně odchýlit. Proto $\epsilon^*_{2k+1}(T, \mathbf{X}) > \frac{k}{2k+1}$.

Dále je třeba dokázat, že změna $k+1$ pozorování již znehodnocuje odhad. To je však jednoduché. Stačí zvolit libovolnou konstantu G a

$$Y_1 = \cdots = Y_{k+1} = G, \qquad Y_i = X_i, \quad i = k+2, \ldots, 2k+1.$$

Potom $\widetilde{\mathbf{Y}} = G$, $\widetilde{\mathbf{V}} = 0$, tedy $J(P_{\mathbf{Y}}) = 0$, tedy, přejdeme-li k funkci *huberM* aplikované na $\mathbf{Y}$, dostaneme $z = 0$, neprovádí se tedy žádné iterace a $\hat{\theta} = \hat{\theta}_1 = G$.

Dokázali jsme tedy, že $\epsilon^*_{2k+1}(T, \mathbf{X}) = \frac{k+1}{2k+1}$. Analogicky lze dokázat, že $\epsilon^*_{2k}(T, \mathbf{X}) = \frac{k}{2k}$. Platí tedy

$$\lim_{k\to\infty} \epsilon^*_{2k+1}(T, \mathbf{X}) = \lim_{k\to\infty} \epsilon^*_{2k}(T, \mathbf{X}) = \frac{1}{2},$$

tedy $\epsilon^*(T)$ existuje a $\epsilon^*(T) = \frac{1}{2}$.

Kapitola 3

Použití robustbase v praxi

Tato kapitola je zasvěcena malé simulaci v prostředí R. Ilustruji zde na několika příkladech použití funkce *huberM* pro odhad polohy, dále pak pro srovnání uvádím odhad polohy pomocí průměru a mediánu.

Odhady ilustruji na třech příkladech.

1. První z nich je aplikace Huberova odhadu na uměle zvolený číselný vektor (1, 4, 2, 2, 11). Budu ilustrovat, že podle různě zvolených parametrů funkce *huberM* může být poslední složka započítána jako korektní pozorování i jako odchylka.

2. Ve druhém příkladu generuji dvě sady dat - jednu z normálního rozdělení a druhou z rozdělení s těžkými chvosty. Diskutuji výsledky.

3. Sadu dat pro třetí příklad beru z praxe, vytvářím ji ze sady *Animals2*, obsažené v robustbase.

3.1 Příklad ukázkového vektoru

```
> a <- c(1, 4, 2, 2, 11)

> a

[1]  1  4  2  2 11
```

Nyní přistupme k odhadům:

```
> huberM(a)

$mu [1] 2.805975    $s [1] 1.4826    $it [1] 10

> mean(a)

[1] 4
```

```
> median(a)

[1] 2
```

Výsledek funkce *huberM* pro vektor a leží asi uprostřed mezi mediánem a průměrem. Budu ilustrovat použití dalších parametrů funkce *huberM*, viz. 2.4. Nejprve ověřím, že spočtená hodnota \$s $= \widetilde{s}$.

```
> mad(a)

[1] 1.4826
```

Vidíme, že tomu tak je. Nyní ověřme, že skutečně $z = 1.5 \cdot \widetilde{s}$. Jelikož zřejmě $\{1, 4, 2, 2\} \in U(2.805975, 1.5 \cdot 1.4826)$, budu nahrazovat (ve smyslu vysvětleném v 2.4) pouze 5.složku vektoru a.

```
>  mean (c(1, 4, 2, 2, 2.805975 + 1.5 * mad(a)))

[1] 2.805975
```

Volbou dostatečně velkého k nebo manuální volbou s by bylo možno docílit, aby se v tomto případě choval M-odhad jako průměr. Naopak malou volbou k bychom docílili chování podobné, jako u mediánu:

```
> huberM(a, k = 10)

$mu [1] 4    $s [1] 1.4826    $it [1] 2

> huberM(a, s = 10)

$mu [1] 4    $s [1] 10    $it [1] 2

> huberM(a, k = 0.5)

$mu [1] 2.370647    $s [1] 1.4826    $it [1] 24

> huberM(a, k = 0.1)

$mu [1] 2.074127    $s [1] 1.4826    $it [1] 21
```

Nakonec budu ilustrovat bod selhání, a to tak, že přidám 4 pozorování, silně kontaminující tento vektor.

```
> a <- c(a,1378,1423,1491,1602)

> huberM(a)

$mu [1] 21.79117    $s [1] 14.826    $it [1] 17
```

Tento příklad ukázal použití a význam volitelných parametrů funkce *huberM*. Dále ukázal, že Huberův M-odhad je odolný, pokud se odchýlí méně než polovina pozorování.

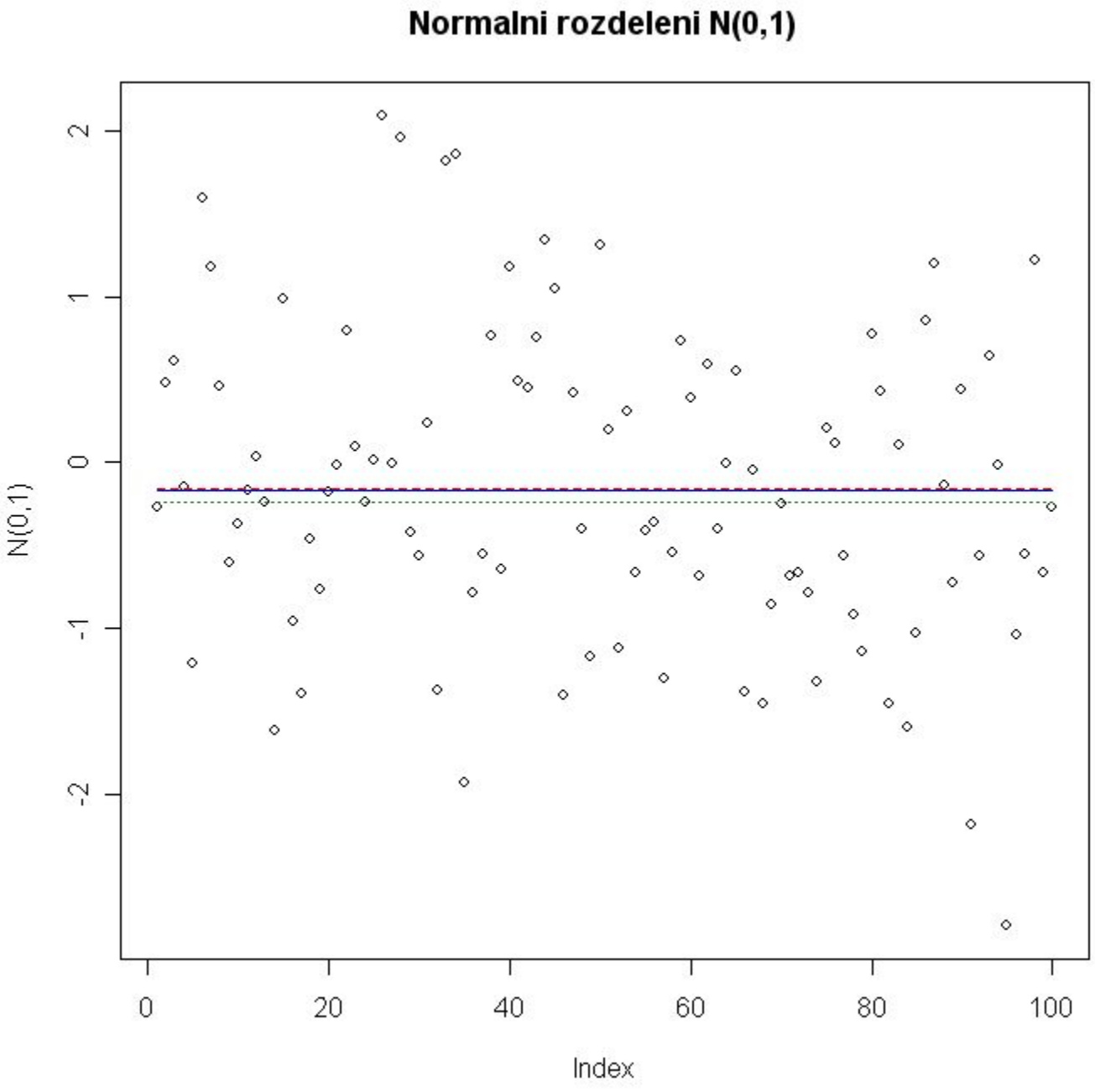
3.2 Výběry z normálního rozdělení a Studentova t-rozdělení

V tomto příkladě ilustruji rozdílé chování Huberova M-odhadu pro rozdělení bez odlehlých hodnot a pro rozdělení s těžkými chvosty. Jako zástupce těchto rozdělení, ze kterých jsem vygeneroval realizace náhodných výběrů, jsem zvolil normální rozdělení $N(0,1)$ a Studentovo t-rozdělení s jedním stupněm volnosti $t(1)$. Realizace obou výběrů čítají 100 pozorování, označuji je b a t. Uvádím syntaxi v R, nevypisuji však získané datové vektory. Výsledky budu pro větší názornost ilustrovat graficky. Používám Huberův M-odhad bez změny nastavení jeho počátečních parametrů.

```
 > b <- rnorm(100)

 > mean(b)

[1] -0.1582808

 > median(b)

[1] -0.238989

 > huberM(b)

 $mu [1] -0.1667562    $s [1] 0.9799045    $it [1] 6

 > plot(b, ylab="N(0,1)", main="Normalni rozdeleni N(0,1)")

 > lines(1:100, rep(mean(b),100), col='red', lty=2)

 > lines(1:100, rep(median(b), 100), col='forestgreen', lty=3)

 > lines(1:100, rep(huberM(b)[1], 100), col='blue', lty=1)
```

Z číselných hodnot i z obrázku 3.1 lze odečíst, že pro normální rozdělení se Huberův M-odhad chová velmi podobně jako průměr. Je to způsobeno faktem, že data neobsahují žádná odlehlá pozorování. To je doloženo také tím, že $\widetilde{s} \approx var(N(0,1)) = 1$. Jiné výsledky dostaneme, budeme-li zkoumat polohu výběru z $t(1)$.

```
 > t <- rt(100,1)

 > mean(t)

[1] -0.929628

 > median(t)
```

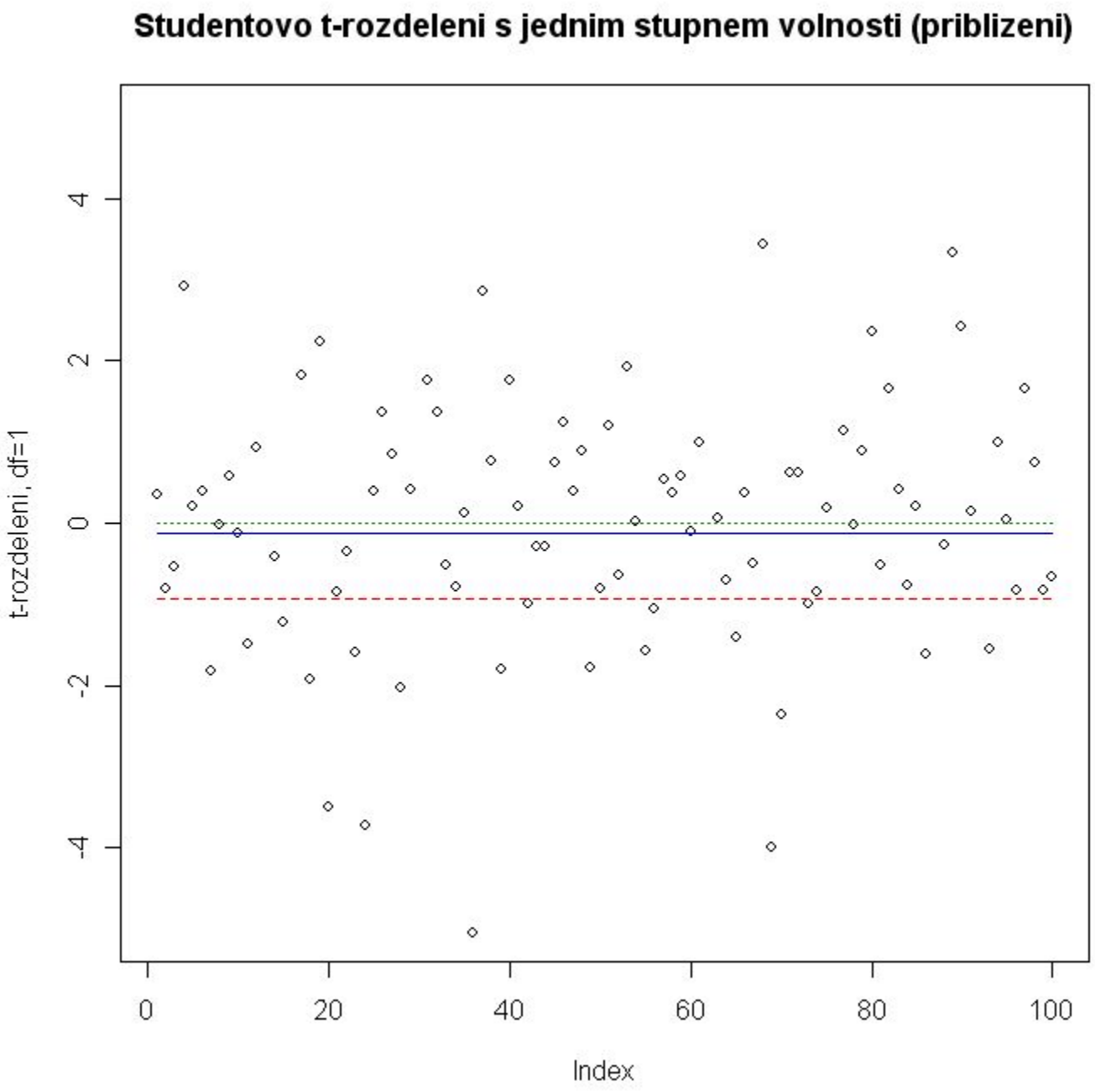


Obrázek 3.1: Rozdělení $N(0,1)$ a odhady polohy. Průměr (červený čárkovaný), medián (zelený čárkovaný) a Huberův odhad (modrý plný).

```
[1] 0.0002848629

 > huberM(t)

 $mu [1]  -0.1153378    $s [1] 1.284288    $it [1] 9

 > plot(t, ylim=c(-5,5), ylab="t-rozdeleni, df=1", main="Studentovo

 t-rozdeleni s jednim stupnem volnosti (priblizeni)")

 > lines(1:100, rep(mean(t),100), col='red', lty=2)

 > lines(1:100, rep(median(t), 100), col='forestgreen', lty=3)

 > lines(1:100, rep(huberM(t)[1], 100), col='blue', lty=1)
```

Obrázek 3.2 zobrazuje pouze data v intervalu $[-5, 5]$, proto zde nejsou vidět opravdu odlehlá data, která pozorování vychylují. Toto řešení jsem zvolil, aby vynikl rozdíl mezi Huberovým odhadem, mediánem a průměrem. Z číselných hodnot i přiblížného obrázku 3.2 je zřejmé, že pro data s velkým množstvím odlehlých pozorování se Huberův M-odhad chová podobně jako medián.

Dle mého názoru v obou případech splňuje Huberův M-odhad to, co bychom od robustního odhadu polohy očekávali.



Obrázek 3.2: Rozdělení $t(1)$ a odhady polohy. Barvy stejné jako v obr.3.1

3.3 Ilustrace odhadu na sadě *Animals2*

Jak jsem již uvedl, sada dat použitá v této podkapitole bude je připravena z hodnot obsažených v sadě *Animals2*, obsažené v robustbase. Jedná se výsledky měření hmotnosti mozku a těla u 65 živočišných druhů. Sada tedy obsahuje dva sloupce dat. V prvním sloupci je uvedena hmotnost živočicha v kilogramech, ve druhém hmotnost jeho mozku v gramech. Data jsou seřazena vzestupně podle prvního sloupce. Zatímco prvních 62 druhů jsou savci, žijící v současnosti, 63.-65. druh jsou pravěcí ještěři - triceratops, diplodocus a brachiosaurus. Vektor dat získám tak, že vydělím hodnoty v prvním sloupci hodnotami ve druhém sloupci - bude se tedy jednat o tisícinu poměru hmotnosti těla a mozku.

```
> c <- round((Animals2[,1]/Animals2[,2]),3)

> c

 [1]   0.036   0.040   0.077   0.057   0.145   0.060   0.062   0.025   0.042

[10]   0.120   0.041   0.040   0.147   0.066   0.031   0.229   0.061   0.224

[19]   0.346   0.161   0.152   0.189   0.167   0.112   0.081   0.142   0.270

[28]   0.163   0.207   0.120   0.129   0.076   0.897   0.324   0.171   0.238

[37]   0.072   0.084   0.109   0.038   0.087   0.059   0.151   0.241   0.625

[46]   0.304   0.119   0.317   0.741   0.047   0.262   0.637   0.947   0.447

[55]   1.067   0.510   0.510   1.099   0.795   0.778   0.553   1.165 134.286

[64] 234.000 563.107

> c1<-c[1:62]
```

Vektor, který jsme získali, je zajímavý tím, že zatímco u převážné většiny současných savců nepřekročí příslušná složka hodnotu 1, u pravěkých ještěrů dosahuje hodnoty více než 100. Ještěři byli velcí a měli malý mozek. Dále stojí v tomto vektoru za povšimnutí mírný nárůst hodnot zleva doprava (tedy malí savci mají relativně spíše těžší mozek, než velcí savci) a dále významně nízké hodnoty u paviána (č. 42) a člověka (č. 50).

Data jsou silně kontaminovaná, ovšem kontaminuje je pouze malý počet odlehlých pozorování. Příklad ilustruje, že Huberův M-odhad skutečně odlehlé hodnoty nebere příliš v potaz. Abych prokázal toto tvrzení, vytvořil jsem z prvních 62 hodnot vektoru c vektor $c1$. Huberův M-odhad budu aplikovat na oba tyto vektory. Dále počítám průměr jak vektoru c, tak vektoru $c1$ a medián vektoru c.

```
> huberM(c)
```

```
 $mu [1] 0.2145258    $s [1]  0.1512252    $it [1] 11

 > huberM(c1)

 $mu [1] 0.1938621    $s [1]  0.1363992    $it [1] 10

 > mean(c)

[1] 14.59392

 > mean(c1)

[1] 0.2776129

 > median(c)

[1] 0.161
```

Rozdíl mezi odhady polohy ve výběrech c a $c1$ je zanedbatelný, což je potěšující výsledek.

Kapitola 4

Závěr

Cílem této práce bylo popsat a zhodnotit R-knihovnu robustbase, jak po stránce složení, tj. v kontextu robustních statistických metod, tak po stránce vývoje, v kontextu prostředí R.

Za účelem popisu složení robustbase jsem se v přípravné 1. kapitole seznámil s pojmem robustnost a s robustními statistickými metodami. Ztotožnění odhadu se statistickým funkcionálem na množině rozdělení mi umožnilo zabývat se měřením robustnosti. V poslední podkapitole 1. kapitoly jsem se zaměřil na třídu M-odhadů.

2. kapitola je klíčovou součástí této práce. Seznamuji v ní čtenáře s knihovnou robustbase, zaměřuji se jak na její vývoj, tak na její složení. Historie robustbase je zatím krátká, proto popis vývoje nevede k přesvědčivým závěrům. V práci upozorňuji na některé problematické skutečnosti, jako je především existence konkurenčního projektu firmy Insightful. Je tedy otázkou, jak se do bude do budoucna robustbase vyvíjet. K dalším úvahám na toto téma odkazuji na konec podkapitoly 2.2.

Vzhledem ke komplexnosti knihovny robustbase, nebylo možno se věnovat jejím funkcím jednotlivě a do detailů. Faktem však je, že základními stavebními kameny robustbase jsou M-odhady. Rozhodl jsem se proto demonstrovat možnosti knihovny na jednoduchém příkladu Huberova odhadu parametru polohy, *huberM*. V práci se pokouším podat popis tohoto odhadu jak z teoretického, tak z praktického pohledu. Podávám důkaz, že bod selhání odhadu funkcí *huberM* je $\frac{1}{2}$ a ilustruji vlastnosti odhadu na jednoduchých příkladech, které jsou obsahem třetí kapitoly. Huberův odhad je zde srovnáván s průměrem a výběrovým mediánem.

Doufám, že v podobě, v jaké práci předkládám, bude pro případného čtenáře zajímavým uvedením k projektu knihovny robustbase. V celé práci jsem se pokoušel na problematiku nahlížet jak z úhlu praxe, dané obsahem robustbase, tak z úhlu teorie robustní statistiky, jak je prezentována v učebnicích, jako je třeba [3]. Snad se mi rozdíl mezi oběma pohledy podařilo uchopit.

Seznam použité literatury

[1] Maronna R., Martin R., Yohai V. (2006): Robust Statistics: Theory and Methods. John Wiley & Sons Ltd, Chichester, England.

[2] Anděl J. (2005): Základy matematické statistiky. Matfyzpress, Praha.

[3] Jurečková J. (2001): Robustní statistické metody. Karolinum, Praha.

[4] Jurečková J., Picek J. (2006): Robust Statistical Methods with R. Chapman & Hall/CRC, Boca Raton, Florida.

[5] Oficiální dokumentace R-knihovny robustbase:
http://bg9.imslab.co.jp/Rhelp/R-2.4.0/src/library/robustbase.html.

[6] Oficiální dokumentace R-knihovny robust:
http://rss.acs.unt.edu/Rdoc/library/robust/html/00Index.html.

[7] Mailing list skupiny Robust statistics and R:
https://stat.ethz.ch/pipermail/r-sig-robust/2005/thread.html.