# ROBUSTNÍ

# STATISTICKÉ METODY

Jana Jurečková

PRAHA 2001

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fyzikální fakulty University Karlovy v Praze

Vedoucí katedry: Prof. RNDr Josef Štěpán, DrSc


ISBN

# Předmluva

Tento učební text je určen pro posluchače magisterského studia Matematicko-fyzikální fakulty UK, kteří se zaměřují na matematickou statistiku a ekonometrii, ale také pro doktorandy oboru pravděpodobnost a matematická statistika a pro další zájemce. Materiál nejen pokrývá robustní statistické metody, které jsou částí přednášky *Robustní a neparametrické metody*, ale je i širší, aby poskytl zájemci ucelený obraz o současném stavu problematiky. Četba předpokládá základní znalosti pravděpodobnosti a matematické statistiky. Pokud některá tvrzení nejsou doplněna důkazy, neboť ty by požadovaly hlubší matematický výklad, jsou doplněna odkazy na literaturu, aby se zájemce mohl s nimi seznámit. Bibliografie je doplněna dalšími tituly, zejména knižními, z oblasti robustních statistických metod, která se bouřlivě rozvíjela zejména od šedesátých let 20. století.

Učební text samozřejmě nepokrývá celou rozsáhlou oblast robustních statistických metod, ke které je pouze úvodem. Zaměřili jsme se pouze na robustní statistické odhady, založené na nezávislých pozorováních, která lze popsat lineárním modelem nebo modelem s parametrem posunutí. Nedotkli jsme se robustních statistických testů ani robustních metod v časových řadách, kde našly





četné zajímavé aplikace. Ale věřím, že čtenář získá představu, co jsou robustní metody, a zapojí je do své práce.

*Praha, leden 2001,* *Jana Jurečková*

# Obsah

# Úvod

Jestliže zpracováváme data klasickými statistickými postupy, založenými na parametrických modelech, obvykle předpokládáme linearitu regrese, nezávislost pozorování, homoskedasticitu, a normální rozdělení chyb. Jak se snadno můžeme přesvědčit dnes, kdy pomocí počítačů můžeme snadno simulovat data z kteréhokoli rozdělení pravděpodobností a modelu, tyto předpoklady často nejsou splněny. Pak nás samozřejmě zajímají hlavně dvě otázky:

a) Do jaké míry jsou klasické statistické postupy použitelné, a za jakých podmínek zachovávají svou optimalitu?

b) Existují jiné statistické postupy, které nejsou tak vázány na splnění určitých podmínek?

Klasické statistické postupy mají typicky parametrický charakter: model je plně určen až na hodnoty několika parametrů, které nabývají reálných nebo vektorových hodnot. Často jsou to parametry rozdělení pravděpodobností náhodných chyb měření. Jakmile se nám podaří tyto parametry odhadnout nebo otestovat jejich obor, můžeme učinit víceméně jednoznačný závěr, plynoucí z našich dat, ovšem za platnosti modelu.





Často se setkáme s *neparametrickými statistickými postupy*, které jsou protipólem parametrických: jsou to takové postupy, které jsou nezávislé nebo málo závislé na tvaru základního rozdělení pravděpodobností a zachovávají si některé dobré vlastnosti pro co nejširší třídu distribučních funkcí, většinou těch, které mají hustotu, případně symetrickou. Diskrétní rozdělení pravděpodobností nás v tomto směru ani tolik netrápí; tvar takového rozdělení většinou poznáme už z povahy experimentu. Typickým představitelem neparametrických statistických postupů jsou *pořadové testy statistických hypotéz*, u kterých je rozdělení pravděpodobností testové statistiky za hypotézy (nulové rozdělení, tj. za $\mathbf{H}_0$) shodné za všech spojitých distribučních funkcí pozorování. U neparametrických postupů chápeme celou hustotu, případně celou regresní funkci jako neznámý parametr (nekonečné dimenze); tento parametr je buď *rušivý*, tj. naše závěry se ho přímo netýkají a pokud možno se vyhýbáme jeho odhadování, nebo naopak je středem našeho zájmu a hledáme postupy, jak tuto funkci odhadnout (odhady hustoty, odhady regresní funkce), nebo otestovat, do jaké třídy patří (testy dobré shody o tvaru rozdělení).

Naproti tomu *robustní statistické postupy* jsou takové, které si zachovávají určitou optimalitu v okolí nějakého základního rozdělení pravděpodobností, např. normálního. K robustním postupům vedlo zjištění, že i malé odchylky od normálního rozdělení mají značný vliv na kvalitu klasického odhadu metodou nejmenších čtverců, klasického $F$-testu a dalších klasických postupů. Robustní postupy lze pak chápat jako určitá vylepšení, modifikace klasických postupů, které neselžou při malých odchylkách od základních předpokladů. Robustní postupy jsou optimální v okolí daného rozdělení, vzhledem k určité vzdálenosti a k určitému kriteriu optimality. Jako takové jsou vydatnější než neparametrické postupy, které svou funkčnost pro široký model platí určitou ztrátou vydatnosti. Mluvíme-li o robustních statistických postupech, většinou máme na mysli robustní statistické odhady; pokud používáme robustní testy, jsou to testy Waldova typu, založené na robustních odhadech, a tyto testy doporučujeme použít v situaci, kdy nemáme vhodný pořadový test pro danou hypotézu.

Během posledních dvaceti let se značně rozvinuly i *semiparametrické statistické postupy,* které chápou hustotu rozdělení pravděpodobností, influenční funkci statistického odhadu nebo další funkci jako rušivý parametr, který obvykle nejprve odhadují, a pak hledají postup, vhodný pro tuto funkci.

V tomto výčtu nelze opominout *adaptivní statistické postupy,* které konvergují (skoro jistě nebo v pravděpodobnosti) k optimálnímu parametrickému odhadu nebo testu tak, že se s rostoucím počtem pozorování adaptují na příslušný parametrický model; jakkoli by tato situace byla ideální, konvergence je natolik pomalá, že optimality bychom dosáhli při nerealisticky velkém počtu pozorování. Existují také *částečně adaptivní postupy*, které se postupně blíží k rozhodnutí, nejlepšímu z předepsané konečné množiny možností.

Protože se adaptivní, neparametrické, robustní a semiparametrické metody rozvíjely postupně, hlavně od čtyřicátých let 20. století, není mezi nimi ostrá hranice, a jednotlivé pojmy, hlediska a cíle se vzájemně prolínají. I v této učebnici, zaměřené hlavně na robustní statistické postupy, se často dotkneme i ostatních postupů. Naším hlavním cílem je ukázat, jaké možné alternativy klasických statistických postupů můžeme použít, pokud si nejsme jisti naším modelem. Matematicky chápeme robustní postupy jako statistické funkcionály, definované na prostoru distribučních funkcí. Zajímá nás jejich chování v okolí určitého rozdělení pravděpodobností,



případně modelu, a toto okolí je definováno vzhledem k nějaké vzdálenosti. Proto musíme nejprve uvažovat možné vzdálenosti na prostoru distribučních funkcí a příslušné základní vlastnosti a charakteristiky statistických funkcionálů, jako je jejich spojitost a derivace. To je teoretickým základem robustních statistických postupů.

# Kapitola 1

# Matematické nástroje robustnosti

## 1.1 Statistický model

Předpokládejme, že pokus vede k pozorováním $X_1, \ldots, X_n$. Klasický statistický model předpokládá, že vektor pozorování $(X_1, \ldots, X_n)$ může nabývat hodnot z *výběrového prostoru* $\mathcal{X}$ se $\sigma$-algebrou podmnožin $\mathcal{B}$, a pravděpodobnostní chování studovaných jevů popisuje rozdělení pravděpodobností $P$, definované na $\mathcal{B}$. Rozdělení $P$ patří do třídy $\mathbb{P} = \{P_\theta, \theta \in \mathbf{\Theta}\}$, indexované parametrem $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^p$, kde $p$ je přirozené číslo.

Trojice $\{\mathcal{X}, \mathcal{B}, P_\theta : \theta \in \mathbf{\Theta}\}$ je (parametrický) statistický model. Ve většině případů je $\mathcal{X}$ podmnožinou $\mathbb{R}^{p \times n}$, tedy náhodný pokus vede k $n$ nezávislým $p$-rozměrným pozorováním.

V některých případech je charakter parametrického statistického modelu plně určen povahou experimentu: např. snadno poznáme binomické, multinomické, Poissonovo či hypergeometrické





rozdělení. Podobně, pravděpodobnostní chování doby čekání (na obsluhu apod.) obvykle charakterizujeme gama rozdělením.

Většina statistických postupů však byla odvozena za předpokladu, že pozorování pocházejí z normálního rozdělení. Tyto postupy jsou většinou algebraicky jednoduché, proto se automaticky používají ve všech situacích, kdy nosičem hustoty pozorování je celá přímka, a na předpoklad normality se jaksi "zapomíná". Např. odhad metodou nejmenších čtverců, jakkoli se zdá univerzální, je úzce spjat s normálním rozdělením chyb a selhává, pokud i jen část pozorování pochází z jiného rozdělení, jehož hustota má těžší chvosty než normální, nebo vyskytují-li se mezi daty odlehlá pozorování, která data kontaminují. O tom se můžeme přesvědčit nejen numericky, ale byly též dokázány přesvědčivé teoretické argumenty, založené na charakterizaci normálního rozdělení: např. Kagan, Linnik a Rao [49] dokázali, že odhad metodou nejmenších čtverců v lineárním regresním modelu je přípustný vzhledem ke kvadratické ztrátové funkci (tj. neexistuje jiný odhad se stejnoměrně menším kvadratickým rizikem) tehdy a jen tehdy, je-li rozdělení chyb normální.

Studentův $t$-test a Snedecorův $F$-test, podobně jako $F$-test lineární hypotézy, byly odvozeny za předpokladu normality; zatímco $t$-test je poměrně robustní k odchylkám od normálního rozdělení, $F$-test je k nim velice citlivý; nejsme-li si jisti normálním rozdělením, použijeme příslušných pořadových testů.

Jestliže si nejsme jisti parametrickou formou modelu, máme dvě možnosti:

a) Vzdáme se parametrizace $P_\theta$ reálným nebo vektorovým parametrem $\theta$ a nahradíme rodinu $\{P_\theta : \theta \in \mathbf{\Theta}\}$ rozsáhlejší rodinou rozdělení pravděpodobností; tj. přijmeme *neparametrický přístup.*

b) Na prostoru $\{\mathcal{X}, \mathcal{B}\}$ zavedeme vhodnou topologii, která nám umožní studovat stabilitu klasických postupů, optimálních za $P_\theta$, při malých odchylkách od $P_\theta$, tj. přijmeme *robustní přístup*.

## 1.2 Ilustrace na statistickém odhadu

Nechť $X_1, \ldots, X_n$ jsou nezávislá pozorování se stejným rozdělením pravděpodobností $P_\theta$, kde $\theta$ je nepozorovatelný parametr, $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^p$; nechť $F(x, \theta)$ je distribuční funkce, příslušná $P_\theta$.

Chceme-li odhadnout parametr $\theta$, máme řadu možností, např.

(1) Metoda maximální věrohodnosti.

(2) Metoda momentů.

(3) Metoda $\chi^2$-minima nebo metoda minimalizující jiný typ vzdálenosti.

(4) Metoda založená na postačujících statistikách (Rao-Blackwellova věta) a na úplných postačujících statistikách (Lehmann-Schefféo věta). Připomeňme si, že vektor uspořádaných pozorování (vektor pořádkových statistik) $X_{n:1} \leq X_{n:2} \leq \ldots \leq X_{n:n}$ je úplnou postačující statistikou pro systém rozdělení s hustotami $\prod_{i=1}^n f(x_i)$, kde $f$ je libovolná spojitá jednorozměrná hustota; v případě, že parametr $\theta$ je reálný, to přirozeně vede ke třídě *L-odhadů* typu

$$T_n = \sum_{i=1}^n c_{ni} h(X_{n:i})$$

založených na pořádkových statistikách.



(5) Minimalizace určité (kriteriální) funkce pozorování a $\theta$ : např. minimalizace

$$\sum_{i=1}^n \rho(X_i, \theta) := \min, \quad \theta \in \mathbf{\Theta},$$

kde $\rho(\cdot, \cdot)$ je vhodná nekonstantní funkce, např. $\rho(x, \boldsymbol{\theta}) = -\log f(x, \boldsymbol{\theta})$ vedoucí k maximálně věrohodnému odhadu. Tím se dostáváme ke třídě *M-odhadů*, tj. odhadů maximálně věrohodného typu.

(6) Inverzí pořadových testů o posunutí v poloze, o významnosti regrese aj. dostáváme třídu *R-odhadů,* založených na pořadích pozorování nebo jejich residuí.

V dalších kapitolách této knížky se seznámíme s $M$-, $L$- a $R$- odhady a s některými dalšími metodami.

## 1.3 Statistický funkcionál

Nechť $X$ je náhodná veličina s rozdělením pravděpodobností $P_\theta$, kde $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \mathbf{\Theta} \subseteq \mathbb{R}^p\}$. Pak v mnoha případech lze $\theta$ chápat jako funkcionál $\theta = T(P)$ definovaný na $\mathcal{P}$; můžeme též psát $\theta = T(F)$, kde $F$ je distribuční funkce příslušná $P$. Přirozeným odhadem $\theta$, založeným na pozorováních $X_1, \ldots, X_n$ pak je $T(P_n)$, kde $P_n$ je *empirické rozdělení pravděpodobností* vektoru $(X_1, \ldots, X_n)$, tj.

$$P_n(A) = \tfrac{1}{n} \sum_{i=1}^n I[X_i \in A], \quad A \in \mathcal{B}, \tag{1.1}$$

tedy $P_n$ je rovnoměrné rozdělení na množině $\{X_1, \ldots, X_n\}$, neboť $P_n(\{X_i\}) = \frac{1}{n}$, $i = 1, \ldots, n$. Distribuční funkce příslušná $P_n$ je *empirická distribuční funkce*

$$F_n(x) = P_n((-\infty, x]) = \tfrac{1}{n} \sum_{i=1}^{n} I[X_i \leq x], \;\; x \in \mathbb{R}. \tag{1.2}$$

**Příklad 1.1** (1) *Střední hodnota:*

$$\begin{aligned} T(P) &= \textstyle\int_{\mathbb{R}} x dP &&= \mathbb{E}X, \\ T(P_n) &= \textstyle\int_{\mathbb{R}} x dP_n &&= \bar{X}_n = \tfrac{1}{n} \sum_{i=1}^{n} X_i. \end{aligned}$$

(2) *Rozptyl:*

$$\begin{aligned} T(P) &= \operatorname{var} X = \int_{\mathbb{R}} x^2 dP - (\mathbb{E}X)^2 \\ T(P_n) &= \tfrac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2. \end{aligned}$$

(3) Jestliže $T(P) = \int_{\mathbb{R}} h(x) dP$, kde $h$ je libovolná $P$-integrabilní funkce, pak empirickým protějškem $T(P)$ je

$$T(P_n) = \tfrac{1}{n} \sum_{i=1}^{n} h(X_i).$$



(4) Obráceně, k danému statistickému odhadu můžeme nalézt příslušný statistický funkcionál: např. *geometrický průměr* pozorovaní $X_1, \ldots, X_n$ je definován jako

$$T(P_n) = G_n = \Big( \prod_{i=1}^{n} X_i \Big)^{1/n},$$

$$\log G_n = \tfrac{1}{n} \sum_{i=1}^{n} \log X_i = \int_{\mathbb{R}} \log x dP_n,$$

a tedy příslušný statistický funkcionál má tvar

$$T(P) = \exp\left\{ \int_{\mathbb{R}} \log x dP \right\}.$$

Podobně *harmonický průměr* $T(P_n) = H_n$ pozorování $X_1, \ldots, X_n$ je definován vztahem

$$\frac{1}{H_n} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{X_i}$$

a jemu příslušný statistický funkcionál má tvar

$$T(P) = H = \left( \int_{\mathbb{R}} \frac{1}{x} dP \right)^{-1}.$$

Statistické funkcionály poprvé uvažoval von Mises [57].

Je žádoucí, aby $T(P_n)$ konvergovalo k $T(P)$ při $n \to \infty$ vzhledem k nějaké konvergenci na prostoru pravděpodobnostních měr: většinou uvažujeme konvergenci v pravděpodobnosti, v distribuci, skoro jistě, ale často také limitu vychýlení odhadu $T(P_n)$ od $T(P)$, tj. $\lim_{n\to\infty} |\mathbb{E}[T(P_n) - T(P)]|$. Abychom mohli studovat chování odhadu $T(P_n)$ v okolí $P$, uvažujeme rozvoj funkcionálu $(T(P_n) - T(P))$ Taylorova typu; k tomu potřebujeme některé další pojmy z funkcionální analýzy, jako různé vzdálenosti mezi $P_n$ a $P$, vzájemné vztahy těchto vzdáleností, a spojitost a diferencovatelnost funkcionálu $T$ vzhledem k příslušné vzdálenosti.

## 1.4 Fisherovská konsistence odhadu

Přirozeným požadavkem, který by měl splňovat statistický odhad, je *fisherovská konsistence*, zavedená v r. 1921 R. A. Fisherem: Odhad $\hat{\theta}_n$ založený na pozorováních $X_1 \ldots, X_n$ s rozdělením pravděpodobností $P$ je fisherovsky konsistentním odhadem parametru $\theta$, jestliže, píšeme-li jej jako funkcionál $\hat{\theta}_n = T(P_n)$ empirického rozdělení pravděpodobností vektoru $(X_1, \ldots, X_n)$, $n = 1, \ldots,$ pak platí $T(P) = \theta$. Tato podmínka není vždy automaticky splněna, jak je vidět na následujícím příkladě:

**Příklad 1.2** Nechť $\theta = \operatorname{var} X = T(P) = \int_{\mathbb{R}} x^2 dP - \left(\int_{\mathbb{R}} x dP\right)^2$ je rozptyl $P$. Pak výběrový rozptyl $\hat{\theta}_n = T(P_n) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ je fisherovsky konsistentním, ale vychýleným odhadem $\theta$. Naproti tomu nevychýlený (nestranný) odhad rozptylu $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ není fisherovsky konsistentním odhadem $\theta$, neboť

$$S_n^2 = \frac{n}{n-1} T(P_n) \quad \text{a} \quad \frac{n}{n-1} T(P) \neq T(P).$$

Fisherovská konsistence je přirozená vlastnost odhadu a z hlediska robustnosti je důležitější než jeho nevychýlenost (nestrannost); proto u každého statistického funkcionálu nejprve ověřujeme jeho fisherovskou konsistenci.



## 1.5 Některé vzdálenosti pravděpodobnostních měr

Nechť $\mathcal{X}$ je metrický prostor, úplný a separabilní s metrikou $d$, a nechť $\mathcal{B}$ je $\sigma$-algebra borelovských podmnožin $\mathcal{X}$. Nechť $\mathcal{P}$ je systém všech pravděpodobnostních měr na $(\mathcal{X}, \mathcal{B})$; pak $\mathcal{P}$ je konvexní množina, na které můžeme zavést různé typy vzdáleností dvou prvků $P, Q \in \mathcal{P}$. Popíšeme stručně některé z těchto vzdáleností, které se v matematické statistice nejčastěji užívají; čtenáře, který se chce podrobněji seznámit s dalšími vzdálenostmi a vůbec s touto problematikou, odkazujeme na literaturu z funkcionální analýzy a teorie pravděpodobnosti, např. [9].

(1) *Prochorovova vzdálenost:*

$$\begin{aligned} d_P(P,Q) &= \inf\{\varepsilon > 0 : \; P(A) \leq Q(A^\varepsilon) + \varepsilon \\ &\qquad \forall A \in \mathcal{B}, A \neq \emptyset\}, \end{aligned}$$

kde $A^\varepsilon = \{x \in \mathcal{X} : \; \inf_{y \in A} d(x,y) \leq \varepsilon\}$ je uzavřené $\varepsilon$-okolí neprázdné množiny $A$.

(2) *Lévyho vzdálenost:* Nechť $\mathcal{X} = \mathbb{R}$ je reálná přímka a nechť $F, G$ jsou distribuční funkce pravděpodobnostních měr $P, Q$. Pak

$$\begin{aligned} d_L(F,G) &= \inf\{\varepsilon > 0 : \; F(x-\varepsilon) - \varepsilon \\ &\qquad \leq G(x) \leq F(x+\varepsilon) + \varepsilon \forall x \in \mathbb{R}\}. \end{aligned}$$

(3) *Úplná variace:*

$$d_V(P,Q) = \sup_{A \in \mathcal{B}} |P(A) - Q(A)|.$$

Jak snadno ověříme, platí $d_V(P,Q) = \int_{\mathcal{X}} |dP - dQ|$.

(4) *Kolmogorovova vzdálenost:* Nechť $\mathcal{X} = \mathbb{R}$ je reálná přímka a nechť $F, G$ jsou distribuční funkce pravděpodobnostních měr $P, Q$. Pak
$$d_K(F,G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|.$$

(5) *Hellingerova vzdálenost:*
$$H(P,Q) = \left\{ \int_{\mathcal{X}} \left( \sqrt{dP} - \sqrt{dQ} \right)^2 \right\}^{1/2}.$$

Jestliže $f = \frac{dP}{d\mu}$ a $g = \frac{dQ}{d\mu}$ jsou hustoty $P, Q$ vzhledem k nějaké míře $\mu$, pak lze Hellingerovu vzdálenost psát ve tvaru
$$H^2(P,Q) = \int_{\mathcal{X}} \left( \sqrt{f} - \sqrt{g} \right)^2 d\mu = 2 \left( 1 - \int_{\mathcal{X}} \sqrt{fg} d\mu \right).$$

(6) *Lipschitzovská vzdálenost:* Předpokládejme, že $d(x,y) \leq 1 \; \forall x,y \in \mathcal{X}$ (jinak vezměme metriku $d' = \frac{d}{1+d}$). Pak
$$d_{Li}(P,Q) = \sup_{\psi\in\mathcal{L}} \left| \int_{\mathcal{X}} \psi dP - \int_{\mathcal{X}} \psi dQ \right|,$$
kde $\mathcal{L} = \{\Psi : \mathcal{X} \mapsto \mathbb{R} : \; |\psi(x) - \psi(y)| \leq d(x,y)\}$ je množina lipschitzovských funkcí.

### Vztahy mezi jednotlivými vzdálenostmi

Množina $\mathcal{P}$ všech pravděpodobnostních měr na $(\mathcal{X}, \mathcal{B})$ je metrickým prostorem vzhledem ke každé z výše popsaných vzdáleností, na kterém pak můžeme studovat spojitost a další vlastnosti statistického funkcionálu $T(P)$. Protože nás zajímá chování funkcionálu



v okolí nějakého rozdělení $P$, zajímá nás také, která vzdálenost jemněji reaguje na malé odchylky od $P$.

Následující nerovnosti mezi jednotlivými vzdálenostmi pravděpodobnostních měr ukazují nejen případnou dominanci jedné vzdálenosti nad druhou, ale i jejich vzájemné vztahy. Ověření těchto nerovností ponecháváme jako cvičení:
$$\begin{aligned} H^2(P,Q) &\leq 2d_V(P,Q) &&\leq 2H(P,Q), \\ d_P^2(P,Q) &\leq d_{Li}^2(P,Q) &&\leq 2d_P(P,Q) \quad \forall P,Q \in \mathcal{P}; \end{aligned}$$
jestliže $\mathcal{X} = \mathbb{R}$, pak dále platí:
$$\begin{aligned} d_L(P,Q) &\leq d_P(P,Q) &&\leq d_V(P,Q), \\ d_L(P,Q) &\leq d_K(P,Q) &&\leq d_V(P,Q) \quad \forall P,Q \in \mathcal{P}. \end{aligned}$$

**Příklad 1.3** Nechť $P$ je exponenciální rozdělení s hustotou
$$f(x) = \begin{cases} \mathrm{e}^{-x} & \ldots \quad x \geq 0 \\ 0 & \ldots \quad x < 0 \end{cases}$$
a $Q$ je rovnoměrné rozdělení $R(0,1)$ s hustotou
$$g(x) = \begin{cases} 1 & \ldots \quad 0 \leq x \leq 1 \\ 0 & \ldots \quad \text{jinak.} \end{cases}$$
Pak
$$2d_V(P,Q) = \int_0^1 \left(1 - \mathrm{e}^{-x}\right) dx + \int_1^\infty \mathrm{e}^{-x} dx = 1 + \tfrac{1}{\mathrm{e}} - 1 + \tfrac{1}{\mathrm{e}} = \tfrac{2}{\mathrm{e}},$$
a tedy $d_V(\exp, R(0,1)) \approx 0.3679$. Dále platí
$$\begin{aligned} d_K(P,Q) &= \sup_{x\geq 0} \left|1 - \mathrm{e}^{-x} - xI[0 \leq x \leq 1] - I[x > 1]\right| \\ &= e^{-1} \approx 0.1839 \end{aligned}$$

a $\quad H^2(\exp, R(0,1)) = 2\left(1 - \int_0^1 \sqrt{e^{-x}}dx\right) = 2\left(\frac{2}{\sqrt{e}} - 1\right),$

tedy $H(\exp, R(0,1)) \approx 0.6528$.

## 1.6 Diferencovatelné statistické funkcionály

Nechť $\mathcal{P}$ je množina všech rozdělení pravděpodobností na prostoru s mírou $(\mathcal{X}, \mathcal{B}, \mu)$, kde $\mathcal{X}$ je úplný separabilní metrický prostor s metrikou $d$ a $\mathcal{B}$ je systém borelovských podmnožin $\mathcal{X}$. Zvolme pevně vzdálenost $\delta$ na $\mathcal{P}$. Nechť $T(\cdot)$ je statistický funkcionál na $\mathcal{P}$. Abychom mohli uvažovat rozvoj statistického funkcionálu $T(\cdot)$ kolem rozdělení $P$, podobný Taylorovu rozvoji, musíme zavést pojem derivace funkcionálu. Zavedeme hned tři různé verze derivace statistického funkcionálu: Gâteauxovu, Fréchetovu a Hadamardovu a porovnáme jejich výhodné i nevýhodné vlastnosti ze statistického hlediska.

**Definice 1.1** *Nechť $P, Q \in \mathcal{P}$ a nechť $t \in [0,1]$. Rozdělení pravděpodobností*

$$P_t(Q) = (1-t)P + tQ \tag{1.3}$$

*nazýváme* **kontaminací** *$P$ rozdělením $Q$ v poměru $t$.*

**Poznámka 1.1** *Protože $\mathcal{P}$ je konvexní, je $P_t(Q)$ skutečně rozdělením pravděpodobností; $P_0(Q) = P$ znamená nepřítomnost kontaminace a $P_1(Q) = Q$ úplnou kontaminaci.*

### Gâteauxova derivace

Zvolme pevně $P, Q \in \mathcal{P}$ a označme $\varphi(t) = T((1-t)P + tQ),\ 0 \leq t \leq 1$. Předpokládejme, že funkce $\varphi(t)$ má konečnou $n$-tou derivaci



$\varphi^{(n)}$ a že pro $k = 1, \ldots, n-1$ jsou derivace $\varphi^{(k)}$ spojité v intervalu $(0,1)$ a derivace zprava $\varphi_+^{(k)}$ jsou zprava spojité v bodě $t = 0$. Pak pro $0 < u < t < 1$ můžeme uvažovat Taylorův rozvoj

$$\varphi(t) = \varphi(u) + \sum_{k=1}^{n-1} \frac{\varphi^{(k)}(u)}{k!}(t-u)^k + \frac{\varphi^{(n)}(v)}{n!}(t-u)^n,\ v \in [u,t]. \tag{1.4}$$

Nás však nejvíce zajímá rozvoj v pravostranném okolí bodu $u = 0$, který odpovídá malé kontaminaci rozdělení $P$. V tom případě nahradíme derivace $\varphi^{(k)}(0)$ pravostrannými derivacemi $\varphi_+^{(k)}(0)$. Derivace $\varphi'_+(0)$ se nazývá *Gâteauxovou derivací* funkcionálu $T$ podle $P$ ve směru $Q$.

**Definice 1.2** *Řekneme, že funkcionál $T$ je diferencovatelný v Gâteauxově smyslu podle $P$ ve směru $Q$, jestliže existuje limita*

$$T'_Q(P) = \lim_{t \to 0_+} \frac{T(P + t(Q-P)) - T(P)}{t}; \tag{1.5}$$

*$T'_Q(P)$ se nazývá Gâteauxovou derivací $T$ podle $P$ ve směru $Q$.*

**Poznámka 1.2**

*a) Gâteauxova derivace $T'_Q(P)$ funkcionálu $T$ je rovna obyčejné derivaci zprava funkce $\varphi$ v bodě 0, tj.*

$$T'_Q(P) = \varphi'(0_+).$$

*b) Podobně je definována Gâteauxova derivace řádu $k$ :*

$$T_Q^{(k)}(P) = \left[\frac{d^k}{dt^k} T(P + t(Q-p))\right]_{t=0_+} = \varphi^{(k)}(0_+).$$

*c) Ve speciálním případě $Q = \delta_x$ (Diracova pravděpodobnost v bodě $x$, rozdělení degenerované v bodě $x$) budeme používat jednoduššího značení $T'_{\delta_x}(P) = T'_x(P)$.*

Taylorův rozvoj (1.4) ve speciálním případě $t = 1,\ u = 0$ dává

$$T(Q)-T(P) = \sum_{k=1}^{n-1} \frac{T_Q^{(k)}(P)}{k!} + \frac{1}{n!}\left[\frac{d^n}{dt^n}T(P+t(Q-p))\right]_{t=t^*} \tag{1.6}$$

kde $0 \le t^* \le 1$.

**Příklad 1.4** (a) *Střední hodnota*

$$T(P) = \int_{\mathcal{X}} x dP = \mathbb{E}_P X$$

$$\varphi(t) = \int_{\mathcal{X}} x d((1-t)P + tQ) = (1-t)\mathbb{E}_P X + t\mathbb{E}_Q X$$

$$\Longrightarrow \varphi'(t) = \mathbb{E}_Q X - \mathbb{E}_P X$$

$$T'_Q(P) = \varphi'(0_+) = \mathbb{E}_Q X - \mathbb{E}_P X.$$

Pro $Q = \delta_x$ je $T'_x = x - \mathbb{E}_P X$.

(b) *Rozptyl*

$$T(P) = \mathrm{var}_P X = \mathbb{E}_P(X^2) - (\mathbb{E}_P X)^2$$

$$T((1-t)P + tQ) = \int_{\mathcal{X}} x^2 d((1-t)P + tQ)$$

$$-\left[\int_{\mathcal{X}} x d((1-t)P + tQ)\right]^2$$



$$\Longrightarrow \varphi(t) = (1-t)\mathbb{E}_P X^2 + t\mathbb{E}_Q X^2 - (1-t)^2(\mathbb{E}_P X)^2$$

$$-t^2(\mathbb{E}_Q X)^2 - 2t(1-t)\mathbb{E}_P X \cdot \mathbb{E}_Q X$$

$$\varphi'(t) = -\mathbb{E}_P X^2 + \mathbb{E}_Q X^2$$

$$+2(1-t)(\mathbb{E}_P X)^2 - 2t(\mathbb{E}_Q X)^2$$

$$-2(1-2t)\mathbb{E}_P X \cdot \mathbb{E}_Q X.$$

Odtud plyne

$$\lim_{t\to 0_+} \varphi'(t) = T'_Q(P)$$

$$= \mathbb{E}_Q X^2 - \mathbb{E}_P X^2 - 2\mathbb{E}_P X \cdot \mathbb{E}_Q X + 2(\mathbb{E}_P X)^2$$

a pro $Q = \delta_x$ nakonec dostáváme

$$T'_x(P) = x^2 - \mathbb{E}_P X^2 - 2x\mathbb{E}_P X + 2(\mathbb{E}_P X)^2$$

$$= (x - \mathbb{E}_P X)^2 - \mathrm{var}_P X.$$

### Fréchetova derivace

**Definice 1.3** *Řekneme, že funkcionál $T$ je diferencovatelný podle $P$ ve Fréchetově smyslu, jestliže existuje lineární funkcionál $L_P(Q-P)$ tak, že stejnoměrně pro $Q \in \mathcal{P}$, $\delta(P,Q) \le C$ pro libovolné pevné $C \in (0,\infty)$*

$$\lim_{t\to 0} \frac{T(P+t(Q-P)) - T(P)}{t} = L_P((Q-P)). \tag{1.7}$$

*Lineární funkcionál $L_P(Q-P)$ nazýváme Fréchetovou derivací funkcionálu $T$ podle $P$ ve směru $Q$.*

**Poznámka 1.3**

*a) Protože $L_P$ je lineární funkcionál, existuje funkce $g : \mathcal{X} \mapsto \mathbb{R}$ taková, že*

$$L_P(Q-P) = \int_{\mathcal{X}} g d(Q-P). \tag{1.8}$$

*b) Jestliže je $T$ diferencovatelné ve Fréchetově smyslu, je diferencovatelné i v Gâteauxově smyslu, tj. existuje $T'_Q(P)$ $\forall Q \in \mathcal{P}$, a platí*

$$T'_Q(P) = L_P(Q-P) \quad \forall Q \in \mathcal{P}. \tag{1.9}$$

*Speciálně,*

$$T'_x(P) = L_P(\delta_x - P) = g(x) - \int_{\mathcal{X}} g dP \tag{1.10}$$

*a odtud dále plyne*

$$\mathbb{E}_P(T'_x(P)) = \int_{\mathcal{X}} T'_x(P) dP = 0. \tag{1.11}$$

*c) Nechť $P_n$ je empirické rozdělení pravděpodobností vektoru $(X_1 \dots, X_n)$. Pak $P_n - P = \frac{1}{n}\sum_{i=1}^n (\delta_{X_i} - P)$, a tedy, protože $L_P$ je lineární funkcionál,*

$$L_P(P_n - P) = \frac{1}{n}\sum_{i=1}^n L_P(\delta_{X_i} - P)$$

$$= \frac{1}{n}\sum_{i=1}^n T'_{X_i}(P) = T'_{P_n}(P). \tag{1.12}$$



**Důkaz** (1.9):
Skutečně, podle (1.7), protože $L_P(\cdot)$ je lineární funkcionál,

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

$$= \lim_{t\to 0_+} \frac{T(P+t(Q-P)) - T(P)}{t} - L_P((Q-P))$$

$$+ L_P(Q-P) = 0 + L_P(Q-P) = L_P(Q-P). \quad \blacksquare$$

## Hadamardova (kompaktní) derivace

Jestliže existuje lineární funkcionál $L(Q-P)$ takový, že konvergence (1.7) je stejnoměrná nikoli nutně pro ohraničené množiny metrického prostoru $(\mathcal{P}, \delta)$, pokrývající $P$, tj. pro všechna $Q$ taková, že $\delta(P,Q) \leq C$, $0 < C < \infty$, ale pouze pro $Q$ patřící do libovolné pevné kompaktní množiny $\mathbf{K} \subset \mathcal{P}$ pokrývající $P$, pak říkáme, že funkcionál $T$ je diferencovatelný *v Hadamardově smyslu* a funkcionál $L(Q-P)$ nazýváme *Hadamardovou (kompaktní) derivací* $T$. Funkcionál, diferencovatelný ve Fréchetově smyslu, je zřejmě diferencovatelný i v Hadamardově smyslu, a z diferencovatelnosti v Hadamardově smyslu dále plyne diferencovatelnost v Gâteauxově smyslu podobným způsobem jako v Poznámce 1.3. Čtenáři, kterého zajímají vlastnosti diferencovatelnosti různých statistických funkcionálů, doporučujeme knížku [23].

Fréchetova diferencovatelnost klade dost omezující podmínky na funkcionál a ne každý robustní funkcionál je splňuje. Na druhé straně, je-li funkcionál fréchetovsky diferencovatelný, pak snadno odvodíme asymptotické (normální) rozdělení pravděpodobností jeho empirického protějšku, pro počet pozorování rostoucí $n$ nade všechny meze. Asymptotickou normalitu často odvodíme i pomocí Hadamardovy derivace, není-li funkcionál dostatečně "hladký". Pokud chceme pouze dokázat, že $T(P_n)$ je konsistentním odhadem $T(P)$, vystačíme jen se spojitostí funkcionálu. Gâteauxova derivace $T'_x(P)$, zvaná *influenční funkcí funkcionálu* $T$, je jednou z nejdůležitějších charakteristik robustnosti funkcionálu. Influenční funkcí se budeme zabývat ve 2. kapitole.

## 1.7 Asymptotické rozdělení empirického funkcionálu

Uvažujme opět metrický prostor $(\mathcal{P}, \delta)$ všech rozdělení pravděpodobností na $(\mathcal{X}, \mathcal{B})$ s metrikou $\delta$ takovou, že

$$\sqrt{n}\delta(P_n, P) = O_p(1) \quad \text{při} \quad n \to \infty, \tag{1.13}$$

kde $P_n$ je empirické rozdělení pravděpodobností, příslušné náhodnému výběru $(X_1, \ldots, X_n)$, $n = 1, 2, \ldots$. Poznamenejme, že (1.13) je splněno např. pro Kolmogorovovu vzdálenost empirické distribuční funkce od skutečné, což má pro statistické aplikace největší význam, ale platí to i pro další vzdálenosti

Ukážeme, že fréchetovská diferencovatelnost spolu s klasickou formou centrální limitní věty dávají asymptotické rozdělení pravděpodobností empirického funkcionálu $T(P_n)$.

**Věta 1.1** *Nechť $T$ je statistický funkcionál, fréchetovsky diferencovatelný podle $P$ a předpokládejme, že empirické rozdělení $P_n$ náhodného výběru $(X_1, \ldots, X_n)$ splňuje podmínku (1.13) při $n \to \infty$. Jestliže Gâteauxova derivace $T'_{X_1}(P)$ má kladný rozptyl,*



$\text{var}_P T'_{X_1}(P) > 0$, *pak posloupnost $\sqrt{n}(T(P_n) - T(P))$ má asymptoticky normální rozdělení při $n \to \infty$, neboli*

$$\mathcal{L}\Big(T(P_n) - T(P)\Big) \longrightarrow \mathcal{N}\Big(0, \text{var}_P T'_{X_1}(P)\Big). \tag{1.14}$$

**Důkaz.** Podle (1.12) je $T'_{P_n}(P) = \frac{1}{n}\sum_{i=1}^n T'_{X_i}(P)$. Dále podle (1.6) a podmínky (1.13) dostáváme

$$\begin{aligned}
\sqrt{n}(T(P_n) - T(P)) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n T'_{X_i}(P) + R_n \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n L_P(P_n - P) + \sqrt{n}\, o(\delta(P_n, P)) \qquad (1.15) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n T'_{X_i}(P) + o_p(1).
\end{aligned}$$

Jestliže společný rozptyl $\text{var}_P T'_{X_i}(P) = \text{var}_P T'_{X_1}(P)$, $i = 1, \ldots, n$, je konečný, pak (1.14) plyne z (1.15) a z klasické centrální limitní věty. ■

**Příklad 1.5** Nechť $T(P) = \text{var}_P X = \sigma^2$. Pak

$$T(P_n) = S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

a podle příkladu 1.4 b)

$$T'_x(P) = (x - \mathbb{E}_P X)^2 - \text{var}_P X,$$

tedy

$$\text{var}_P T'_X(P) = \mathbb{E}_P(X - \mathbb{E}_P X)^4 - \mathbb{E}_P^2(X - \mathbb{E}_P X)^2 = \mu_4 - \mu_2^2$$

a podle Věty 1.1 dostáváme asymptotické rozdělení výběrového rozptylu

$$\mathcal{L}\Big(\sqrt{n}(S_n^2 - \sigma^2)\Big) \longrightarrow \mathcal{N}\Big(0, \mu_4 - \mu_2^2\Big).$$

# Kapitola 2

# Základní charakteristiky robustnosti

## 2.1 Influenční funkce

Vraťme se k rozvoji (1.15) rozdílu $T(P_n) - T(P)$, podle kterého

$$T(P_n) - T(P) = \tfrac{1}{n} \sum_{i=1}^{n} T'_{X_i}(P) + n^{-1/2} R_n, \tag{2.1}$$

kde $n^{-1/2} R_n = o_p(n^{-1/2})$. Pak $\frac{1}{n}\sum_{i=1}^{n} T'_{X_i}(P)$ můžeme chápat jako chybu odhadu $T(P)$ pomocí $T(P_n)$ a $T'_{X_i}(P)$ můžeme chápat jako příspěvek $X_i$ k této chybě, neboli jako *vliv* $X_i$ na tuto chybu. To nás intuitivně vede k výkladu Gâteauxovy derivace $T'_x(P),\ x \in \mathcal{X}$ jako *influenční funkce* funkcionálu $T(P)$.

**Definice 2.1** *Influenční funkcí funkcionálu $T$ v rozdělení pravděpodobností $P$ nazveme Gâteauxovu derivaci $T$ podle $P$ ve směru*





*$\delta_x,\ x \in \mathcal{X},$ tj.*

$$IF(x; T, P) = T'_x(P) = lim_{t \to 0_+} \frac{T(P_t(\delta_x)) - T(P)}{t} \tag{2.2}$$

*kde $P_t(\delta_x) = (1-t)P + t\delta_x$.*

**Vlastnosti $IF$:**

a) $\mathbb{E}_P(IF(x;T,P)) = \int_{\mathcal{X}} T'_x(P) dP = 0,$
tedy průměrný vliv na chybu odhadování přes všechny body $x$ je roven nule.

b) Jestliže $T$ je fréchetovsky diferencovatelný, je splněna podmínka (1.13) a

$$\mathrm{var}_P(IF(x;T,P)) = \mathbb{E}_P(IF(x;T,P))^2 > 0,$$

pak $\Big(\sqrt{n}(T(P_n) - T(P)\Big) \longrightarrow \mathcal{N}\Big(0, \mathrm{var}_P(IF(x;T,P))\Big).$

**Příklad 2.1** (a) *Střední hodnota:* $T(P) = \mathbb{E}_P(X) = m_P$. Pak

$$\begin{aligned}
&T(P_n) = \bar{X}_n, \\
&IF(x;T,P) = T'_x(P) = x - m_p, \\
&\mathbb{E}_P(IF(x;T,P)) = 0, \\
&\mathrm{var}_P(IF(x;T,P)) = \mathrm{var}_P X = \sigma_P^2, \\
&\mathbb{E}_Q(IF(x;T,P)) = m_Q - m_P \quad \text{pro} \quad Q \neq P, \\
&\mathcal{L}\Big(\sqrt{n}(\bar{X}_n - m_p)\Big) \longrightarrow \mathcal{N}(0, \sigma_P^2)
\end{aligned}$$

pokud $P$ je skutečné rozdělení pravděpodobností náhodného výběru $(X_1,\ldots,X_n)$.
(b) *Rozptyl:* $T(P)=\mathrm{var}_P X=\sigma_P^2$. Pak

$$IF(x;T,P)=(x-m_P)^2-\sigma_P^2,$$

$$I\!E_P(IF(x;T,P))=0,$$

$$\mathrm{var}_P(IF(x;T,P))=\mu_4-\mu_2^2=\mu_4-\sigma_P^4$$

$$I\!E_Q(IF(x;T,P))=I\!E_Q(X-m_p)^2-\sigma_P^2$$

$$=\sigma_Q^2+(m_Q-m_P)^2+2I\!E_Q(X-m_Q)(m_Q-m_P)$$

$$-\sigma_P^2=\sigma_Q^2-\sigma_P^2+(m_Q-m_P)^2.$$

### 2.1.1 Diskretizovaná forma influenční funkce

Označme $T_n=T(P_n)=T_n(X_1,\ldots,X_n)$ empirický funkcionál odpovídající vektoru pozorování $(X_1,\ldots,X_n)$. Přidejme k pozorováním $X_1,\ldots,X_n$ další pozorování $Y$. Pak vliv $Y$ na $T_n$ charakterizujeme rozdílem

$$T_{n+1}(X_1,\ldots,X_n,Y)-T_n(X_1,\ldots,X_n):=I(T_n,Y). \tag{2.3}$$

Protože

$$\begin{aligned}P_n &= \tfrac{1}{n}\sum_{i=1}^n \delta_{X_i},\\ P_{n+1} &= \frac{1}{n+1}\left(\sum_{i=1}^n \delta_{X_i}+\delta_Y\right)\\ &= \frac{n}{n+1}P_n+\frac{1}{n+1}\delta_Y=\end{aligned}$$



$$\left(1-\frac{1}{n+1}\right)P_n+\frac{1}{n+1}\delta_Y,$$

můžeme říci, že $P_{n+1}$ vzniklo z $P_n$ kontaminací degenerovaným rozdělením $\delta_Y$ v poměru $\frac{1}{n+1}$, a tedy

$$I(T_n,Y)=T\left[\left(1-\frac{1}{n+1}\right)P_n+\frac{1}{n+1}\delta_Y\right]-T(P_n).$$

Protože

$$\begin{aligned}&\lim_{n\to\infty}(n+1)I(T_n,Y) \qquad (2.4)\\ &=\lim_{n\to\infty}\frac{T\left[\left(1-\frac{1}{n+1}\right)P_n+\frac{1}{n+1}\delta_Y\right]-T(P_n)}{\frac{1}{n+1}}\\ &=IF(Y;T,P),\end{aligned}$$

můžeme chápat $(n+1)I(T_n,Y)$ jako diskretizovanou verzi influenční funkce. Supremum $|I(T_n,Y)|$ přes $Y$ představuje míru citlivosti empirického funkcionálu $T_n$, při pevných $X_1,\ldots,X_n$, ke přidání dalšího pozorování.

**Definice 2.2** *Citlivostí funkcionálu $T_n(X_1,\ldots,X_n)$ k přidání dalšího pozorování při daných $X_1,\ldots,X_n$ nazýváme číslo*

$$S(T_n)=\sup_Y |I(T_n(X_1,\ldots,X_n),Y)|. \tag{2.5}$$

**Příklad 2.2** (a) *Střední hodnota:*

$$T(P)=I\!E_P X,\quad T_n=\bar X_n,\ T_{n+1}=\bar X_{n+1}$$

$$\Longrightarrow T_{n+1} = \frac{1}{n+1}(n\bar{X}_n + Y)$$

$$I(T_n, Y) = \left(\frac{n}{n+1} - 1\right)\bar{X}_n + \frac{1}{n+1}Y$$

$$= \frac{1}{n+1}(Y - \bar{X}_n)$$

$$\Longrightarrow (n+1)I(T_n, Y) = Y - \bar{X}_n \stackrel{P}{\Longrightarrow} Y - I\!E_P X$$

$$\Longrightarrow S(\bar{X}_n) = \frac{1}{n+1}\sup_Y |Y - \bar{X}_n| = \infty,$$

tedy výběrový průměr má nekonečnou citlivost k přidání dalšího pozorování.
(b) *Medián:*
Nechť $n = 2m+1$ a nechť $X_{(1)} \leq \ldots \leq X_{(n)}$ jsou pozorování uspořádaná podle velikosti. Pak $T_n = T_n(X_1, \ldots, X_n)$ $= X_{(m+1)}$ a $T_{n+1} = T_{n+1}(X_1 \ldots, X_n, Y)$ nabývá následujících hodnot v závislosti na poloze $Y$ vzhledem k ostatním pozorováním:

$$T_{n+1} = \begin{cases} \frac{X_{(m)}+X_{(m+1)}}{2} & \ldots \quad Y \leq X_{(m)} \\ \frac{X_{(m+1)}+X_{(m+2)}}{2} & \ldots \quad Y \geq X_{(m+2)} \\ \frac{Y+X_{(m+1)}}{2} & \ldots \quad X_{(m)} \leq Y \leq X_{(m+2)} \end{cases}$$

a odtud odvodíme míru vlivu přidání $Y$ k pozorováním $X_1, \ldots, X_n$ :

$$I(T_n, Y) = \begin{cases} \frac{X_{(m)}-X_{(m+1)}}{2} & Y \leq X_{(m)} \\ \frac{X_{(m+2)}-X_{(m+1)}}{2} & Y \geq X_{(m+2)} \\ \frac{Y-X_{(m+1)}}{2} & X_{(m)} \leq Y \leq X_{(m+2)}. \end{cases}$$



Protože $|\frac{1}{2}(Y - X_{(m+1)})|$ je nejmenší ze tří možných hodnot $|I(T_n, Y)|$, dostáváme hodnotu citlivosti mediánu ke přidání dalšího pozorování

$$S(T_n) = \max\left\{\tfrac{1}{2}(X_{(m+1)} - X_{(m)}), \tfrac{1}{2}(X_{(m+2)} - X_{(m+1)})\right\};$$

tato hodnota je konečná při libovolných pevných hodnotách $X_1, \ldots, X_n$.

## 2.2 Kvalitativní robustnost

Na příkladu 2.1 jsme viděli, že influenční funkce průměru a rozptylu jsou neohraničené a mohou nabýt libovolně velkých hodnot. Rovněž příklad 2.2 ukazuje, že přidání dalšího pozorování může způsobit selhání výběrového průměru. Podobné chování pozorujeme na odhadu metodou nejmenších čtverců (a vlastně průměr je speciálním případem odhadu metodou nejmenších čtverců); připomeňme si větu Kagana, Linnika a Rao, citovanou v paragrafu 1.1, podle které je odhad metodou nejmenších čtverců velice citlivý k odchylkám od normálního rozdělení chyb. Odtud intuitivně usuzujeme, že odhad metodou nejmenších čtverců (a průměr) je velmi nerobustní. Jak však matematicky definovat robustnost? Definice robustnosti není zcela jednoznačná, protože historicky se tento pojem vyvíjel po mnoho let a problémy citlivosti statistických postupů k odchylkám od daných podmínek uvažovalo mnoho statistiků v průběhu dlouhého období a z různých hledisek.

Je zajímavé, že prvně si uvědomili citlivost průměru a rozptylu k odlehlým pozorováním astronomové a fyzikové, kteří se snažili určit hodnoty různých fyzikálních, geofyzikálních a astronomických konstant jakožto průměru několika měření. Tato část historie je velmi zajímavá a poučná a je poutavě popsána ve Stiglerově knize [70]. R. J. Boskovič [12] již v roce 1757 navrhl alternativní metodu k nejmenším čtvercům při vyhodnocování svých pokusů spějících k charakterizování tvaru zeměkoule. E. S. Pearson [58] již v r. 1931 pozoroval citlivost klasických metod analýzy rozptylu k odchylkám od normálního rozdělení. J. W. Tukey a jeho princetonská skupina začali se systematickým studiem různých alternativ k metodě nejmenších čtverců od 40. let 20. století. Označení "robustní" poprvé použil Box [13] v r. 1953. Box a Anderson [14] v r. 1955 argumentovali tím, že dobrý statistický postup má být málo citlivý ke změnám parametrů, které jsou pro něj rušivé nebo se ho netýkají, ale má být *vydatný*, tj. citlivý ke změnám parametrů, které jsou středem jeho zájmu.

Ve většině případů uvažujeme robustnost statistického postupu vzhledem k odchylkám od předpokládaného rozdělení chyb. Jsou však i jiné důležité typy robustnosti, např. k odchylkám od předpokladu nezávislosti pozorování. Hampel [29], [30] uvažoval pojem robustnosti statistického funkcionálu, založený na jeho spojitosti v okolí daného rozdělení pravděpodobností $P_0 \in \mathcal{P}$ vzhledem k Prochorovově metrice na prostoru $\mathcal{P}$.

Nechť náhodná veličina [vektor] $X$ nabývá hodnot ve výběrovém prostoru $(\mathcal{X}, \mathcal{B})$ a $(X_1, \ldots, X_n)$ je vektor nezávislých realizací $X$ nabývající hodnot v součinovém prostoru $(\mathcal{X}, \mathcal{B})^{\otimes n}$. Nechť $T_n = T_n(X_1, \ldots, X_n)$ je posloupnost statistik (empirických funkcionálů), $T_n : (\mathcal{X}, \mathcal{B})^{\otimes n} \mapsto (\mathcal{T}_n, \mathcal{A}_n)$. Nechť $\mathcal{P}$ je systém všech rozdělení pravděpodobností na $\mathcal{B}$ s Prochorovovou metrikou $d_P$.

**Definice 2.3** *Řekneme, že posloupnost statistik $\{T_n\}$ je (kvalitativně) robustní pro rozdělení pravděpodobností $P$, jestliže k libovolnému $\varepsilon > 0$ existuje $\delta > 0$ a přirozené číslo $n_0$ tak, pro všechna $Q \in \mathcal{P}$ a $n \geq n_0$,*

$$d_P(P, Q) < \delta \implies d_P\left(\mathcal{L}_P(T_n), \mathcal{L}_Q(T_n)\right) < \varepsilon, \tag{2.6}$$

*kde $\mathcal{L}_P(T_n)$ je rozdělení $T_n$ za $P$ a $\mathcal{L}_Q(T_n)$ je rozdělení $T_n$ za $Q$.*



Takto chápanou robustnost nazýváme *kvalitativní* , protože pouze říká, jestli funkcionál je nebo není robustní, a tuto charakteristiku nijak neměří. Je to také robustnost *infinitesimální*, protože uvažuje chování funkcionálu v okolí $P_0$. Samozřejmě podobně můžeme uvažovat spojitost i vzhledem k jiné metrice na $\mathcal{P}$, např. k Lévyho metrice.

Protože chceme srovnávat funkcionály mezi sebou z hlediska robustnosti, snažíme se robustnost nějakým způsobem kvantifikovat, tj. charakterizovat ji nějakým číslem. Jak ukážeme, takových možných kvantifikací je celá řada; náhrada složitějšího pojmu jedním číslem je však většinou jednostranná a zjednodušující.

## 2.3 Kvantitativní charakteristiky robustnosti

### 2.3.1 Charakteristiky založené na influenční funkci

Influenční funkce je jednou z nejdůležitějších charakteristik statistického funkcionálu/odhadu. Hodnota $IF(x; T, P)$ měří vliv kontaminace funkcionálu $T$ hodnotou $x$, a tedy má-li být $T$ robustní, měl by mít ohraničenou influenční funkci. Ohraničenost influenční funkce však neplyne ze spojitosti funkcionálu, tj. z jeho kvalitativní robustnosti ; např. odhad parametru polohy nebo posunutí, vzniklý inverzí van der Waerdenova pořadového testu, má neohraničenou influenční funkci, zatímco je globálně robustní.

Nejužívanějšími číselnými charakteristikami funkcionálu $T$, založenými na influenční funkci, jsou jeho *globální a lokální citlivost:*

a) *Globální citlivostí* funkcionálu $T$ pro rozdělení pravděpodobností $P$ nazýváme maximální hodnotu influenční funkce, příslušnou argumentu $P$, tj.

$$\gamma^* = \sup_{x \in \mathcal{X}} |IF(x; T, P)|. \tag{2.7}$$

b) *Lokální citlivostí* funkcionálu $T$ pro rozdělení pravděpodobností $P$ nazýváme hodnotu

$$\lambda^* = \sup_{x,y;\ x \neq y} \left| \frac{IF(y; T, P) - IF(x; T, P)}{y - x} \right|, \tag{2.8}$$

která zobrazuje vliv nahrazení hodnoty $x$ hodnotou $y$ na funkcionál $T$.

Rozdíl mezi globální a lokální citlivostí je dobře vidět na následujícím příkladě.

**Příklad 2.3** (a) *Průměr*

$T(P) = \mathbb{E}_P(X), \quad IF(x; T, P) = x - \mathbb{E}_P X \implies \gamma^* = \infty, \quad \lambda^* = 1$; tedy průměr není robustní, ale není citlivý k lokálnímu nahrazování hodnot.

(b) *Rozptyl*

$$T(P) = \mathrm{var}_P X = \sigma_P^2,$$

$$IF(x; T, P) = (x - \mathbb{E}_P(X))^2 - \sigma_P^2, \quad \gamma^* = \infty,$$



$$\lambda^* = \sup_{y \neq x} \left| \frac{(x - \mathbb{E}_P(X))^2 - (y - \mathbb{E}_P(X))^2}{x - y} \right|$$

$$= \sup_{y \neq x} \left| \frac{x^2 - y^2 - 2(x - y)\mathbb{E}_P X}{x - y} \right|$$

$$= \sup_{y \neq x} |x + y - 2\mathbb{E}_P X| = \infty,$$

a tedy rozptyl není robustní ani k velkým, ani k lokálním odchylkám.

### 2.3.2 Bod selhání

Velmi často používanou charakteristikou robustnosti odhadu je jeho bod selhání, navržený Donoho a Huberem [20] v r. 1983. Uvažujme náhodný výběr $\mathbf{x}^0 = (x_1, \ldots, x_n)$ a jemu příslušnou hodnotu $T_n(\mathbf{x}^0)$ odhadu funkcionálu $T$. V tomto "počátečním" výběru nahradíme $m$ jakýchkoli složek libovolnými hodnotami; představme si co nejnepříznivější nahrazení co nejnepříznivějšími hodnotami, případně nekonečnými. Označme $\mathbf{x}^{(m)}$ nový výběr vzniklý po takovém nahrazení a $T_n(\mathbf{x}^m)$ příslušnou hodnotu odhadu.

Pak *bodem selhání odhadu* $T_n$ *ve výběru* $\mathbf{x}^{(0)}$ nazýváme číslo

$$\varepsilon_n^*(T_n, \mathbf{x}^{(0)}) = \frac{m^*(\mathbf{x}^{(0)})}{n},$$

kde $m^*(\mathbf{x}^{(0)})$ je nejmenší celé číslo $m$, pro které

$$\sup_{\mathbf{x}^{(m)}} \|T_n(\mathbf{x}^{(m)}) - T_n(\mathbf{x}^{(0)})\| = \infty,$$

tj. nejmenší podíl pozorování, který po nahrazení libovolnými hodnotami může přivést $T_n$ k nekonečným hodnotám. Bod selhání některých odhadů je univerzální v tom smyslu, že $m^*$ nezávisí na počátečním výběru $\mathbf{x}^{(0)}$. V takovém případě můžeme stanovit limitu $\varepsilon^* = \lim_{n\to\infty} \varepsilon_n^*$, která se také někdy nazývá bodem selhání.

Modifikaci bodu selhání dostaneme, jestliže místo nahrazení $m$ složek přidáme k původnímu výběru $m$ nepříznivých hodnot.

**Příklad 2.4** (a) *Průměr* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ :

$\varepsilon_n^*(\bar{X}_n, \mathbf{x}^{(0)}) = \frac{1}{n}$ pro libovolný počáteční výběr $\mathbf{x}^{(0)}$, a tedy

$\lim_{n\to} \varepsilon_n^*(\bar{X}_n, \mathbf{x}^{(0)}) = 0$ pro libovolný počáteční výběr $\mathbf{x}^{(0)}$.

(b) *Medián* $\tilde{X}_n = X_{(\frac{n+1}{2})}$ (pro jednoduchost uvažujeme liché $n$):

$\varepsilon_n^*(\tilde{X}_n, \mathbf{x}^{(0)}) = \frac{n+1}{2n}$ pro libovolný počáteční výběr $\mathbf{x}^{(0)}$, a tedy

$\lim_{n\to} \varepsilon_n^*(\tilde{X}_n, \mathbf{x}^{(0)}) = \frac{1}{2}$ pro libovolný počáteční výběr $\mathbf{x}^{(0)}$.

### 2.3.3 Míra chvostů statistického odhadu

Tato míra se uplatňuje zejména při posuzování odhadů parametrů posunutí a regrese, kde je v překvapivé shodě s intuicí; zde ji budeme ilustrovat na parametru posunutí a později se vrátíme k regresi. Uvažujme model, ve kterém $(X_1, \ldots, X_n)$ je náhodný výběr z rozdělení pravděpodobností se spojitou distribuční funkcí $F(x-\theta)$, $\theta \in \mathbb{R}$, a chceme odhadnout parametr $\theta$. V takovém modelu je přirozené omezit se na odhady $T_n$ *ekvivariantní vzhledem k posunutí,* tj. splňující

$$T_n(X_1 + c, \ldots, X_n + c) = T_n(X_1, \ldots, X_n) + c$$
$$\forall\ c \in \mathbb{R} \text{ a } \ \forall X_1 \ldots, X_n.$$

Chování odhadu $T_n$ parametru $\theta$ můžeme charakterizovat pomocí průběhu pravděpodobnosti $P_\theta(|T_n - \theta| > a)$, buď při pevném



$a > 0$ a $n \to \infty$, nebo při pevném $n$ a $a \to \infty$. Skutečně, jestliže $\{T_n\}$ je konsistentním odhadem $\theta$, pak pro libovolné pevné $a > 0$ platí $\lim_{n\to 0} P_\theta(|T_n - \theta| > a) = 0$. Někteří autoři, např. Bahadur [4], Fu [25] a Sievers [67] uvažovali jako míru vydatnosti odhadu $T_n$ limitu

$$\lim_{n\to\infty} \left\{ -\frac{1}{n} \log P_\theta(|T_n - \theta| > a) \right\} \quad \text{ při pevném } a > 0$$

(pokud tato limita existuje), a porovnávali odhady z hlediska této vydatnosti.

Od dobrého odhadu $T_n = T_n(X_1, \ldots, X_n)$ také očekáváme, že

$$\lim_{a\to\infty} P_\theta(|T_n - \theta| > a) = 0$$

při pevném $n$, a že tato konvergence je co nejrychlejší vzhledem k $a \to \infty$. Pravděpodobnosti $P_\theta(T_n - \theta > a)$ nebo $P_\theta(T_n - \theta < -a)$ při velkých $a > 0$ nazýváme pravým, resp. levým *chvostem rozdělení pravděpodobností* $T_n$. V případě symetrického rozdělení charakterizujeme chvosty pravděpodobností $P_\theta(|T_n - \theta| > a) = P_0(|T_n| > a)$. Lze tedy říci, že zajímavé jsou odhady s co nejrychlejšími chvosty; existuje však horní hranice rychlosti chvostů ekvivariantního odhadu $T_n$, a ta je dána hodnotami $1 - F(a)$ a $F(-a)$, při velkých $a > 0$.

Pro jednoduchost uvažujme symetrickou distribuční funkci, tj. předpokládejme, že $F(-x) = 1 - F(x) \quad \forall x \in \mathbb{R}$. Jurečková [43] navrhla následující míru chování chvostů ekvivariantního odhadu $T_n$ (viz [43]):

$$\begin{aligned} B(T_n; a) &= \frac{-\log\ P_\theta(|T_n - \theta| > a)}{-\log\ (1 - F(a))} \\ &= \frac{-\log\ P_0(|T_n| > a)}{-\log\ (1 - F(a))},\ a > 0. \end{aligned} \tag{2.9}$$

Hodnota $B(T_n;a)$ udává, kolikrát rychleji konverguje pravděpodobnost
$P_0(|T_n| > a)$ k 0 při $a \to \infty$ než $1 - F(a)$, a tedy zajímavý je odhad $T_n$ s co největšími hodnotami $B(T_n;a)$ při $a \gg 0$. Snadno ověříme následující lemma:

**Lemma 2.1** *Nechť $X_1, \ldots, X_n$ je náhodný výběr z populace s distribuční funkcí $F(x-\theta)$, $0 < F(x) < 1, F(-x) = 1-F(x)$, $x, \theta \in \mathbb{R}$. Nechť $T_n$ je ekvivariantní odhad $\theta$ takový, že pro libovolné pevné $n$ platí*

$$\begin{aligned} \min_{1\le i\le n} X_i > 0 \quad &\Longrightarrow \quad T_n(X_1,\ldots,X_n) > 0 \\ \max_{1\le i\le n} X_i < 0 \quad &\Longrightarrow \quad T_n(X_1,\ldots,X_n) < 0. \end{aligned} \tag{2.10}$$

*Pak, pro libovolné pevné $n$,*

$$1 \le \underline{\lim}_{a\to\infty} B(T_n;a) \le \overline{\lim}_{a\to\infty} B(T_n;a) \le n. \tag{2.11}$$

**Důkaz.** Skutečně, pro ekvivariantní odhad $T_n$ platí

$$\begin{aligned} &P_0(|T_n(X_1,\ldots,X_n)| > a) \\ &= P_0(T_n(X_1,\ldots,X_n) > a) \\ &+ P_0(T_n(X_1,\ldots,X_n) < -a) \\ &= P_0(T_n(X_1-a,\ldots,X_n-a) > 0) \\ &+ P_0(T_n(X_1+a,\ldots,X_n+a) < 0) \\ &\ge P_0\left(\min_{1\le i\le n} X_i > a\right) + P_0\left(\max_{1\le i\le n} X_i < -a\right) \end{aligned}$$



$$= 2^{-n+1}\left[P_0(|X_1| > a)\right]^n,$$

a tedy

$$\begin{aligned} &-\log P_0(|T_n(X_1,\ldots,X_n)| > a) \\ &\le -\log 2 - n\log(1-F(a)) \\ &\Longrightarrow \overline{\lim}_{a\to\infty} \frac{-\log P_0(|T_n| > a)}{-\log(1-F(a))} \le n. \end{aligned}$$

Podobně,

$$\begin{aligned} &P_0(|T_n(X_1,\ldots,X_n)| > a) \le P_0\left(\min_{1\le i\le n} X_i \le -a\right) \\ &+ P_0\left(\max_{1\le i\le n} X_i \ge a\right) = 2\left\{1 - \left[1 - \tfrac{1}{2}P_0(|X_1| > a)\right]^n\right\} \\ &= 2\left\{1 - (F(a))^n\right\} \\ &= 2(1-F(a))\left[1 + F(a) + \ldots + (F(a))^{n-1}\right] \\ &\le 2n(1-F(a)), \end{aligned}$$

a tedy

$$\begin{aligned} &-\log P_0(|T_n(X_1,\ldots,X_n)| > a) \\ &\ge -\log 2 - \log n - \log(1-F(a)) \\ &\Longrightarrow \underline{\lim}_{a\to\infty} \frac{-\log P_0(|T_n| > a)}{-\log(1-F(a))} \ge 1. \end{aligned}$$

■

Pokud odhad $T_n$ dosahuje horní hranice ve (2.11), pak je zřejmě nejlepší pro distribuční funkci $F$, protože jeho chvosty konvergují k nule $n$-násobně rychleji než $(1-F(a))$, a rychleji nelze. Vznikají ovšem otázky,

- zda je tato horní hranice dosažitelná a pro která $T_n$ a $F$,
- zda nějaký odhad $T_n$ dosahuje vysokých hodnot $B(T_n;a)$ robustně pro velkou třídu distribučních funkcí.

Ukazuje se, že dolní i horní hranice ve (2.11) jsou dosažitelné výběrovým průměrem $\bar{X}_n$, a to horní hranice pro normální rozdělení a pro rozdělení s exponenciálními chvosty a dolní hranice pro Cauchyho rozdělení a pro rozdělení s těžkými chvosty. To znamená, že $\bar{X}_n$ je opět velmi nerobustní. Naproti tomu, chování výběrového mediánu $\tilde{X}_n$ je robustní i z hlediska chvostů: $\tilde{X}_n$ však nedosahuje horní hranice ve (2.11), naopak, $\lim_{a\to\infty} B(\tilde{X}_n;a)$ se drží uprostřed mezi 1 a $n$ pro širokou třídu distribučních funkcí $F$.

Protože tyto závěry dobře charakterizují pojem robustnosti, upřesníme je v následující větě:

**Věta 2.1** *Nechť $X_1,\dots,X_n$ je náhodný výběr z populace s distribuční funkcí $F(x-\theta)$, $0<F(x)<1$, $F(-x)=1-F(x)$, $x,\theta\in\mathbb{R}$.*

*(i) Nechť $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ značí výběrový průměr. Má-li distribuční funkce $F$ exponenciální chvosty, tj.*

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{ba^r}=1 \quad \text{pro nějaká } b>0,\ r\geq 1, \tag{2.12}$$

*pak*

$$\lim_{a\to\infty} B(\bar{X}_n;a)=n. \tag{2.13}$$

*(ii) Má-li distribuční funkce $F$ těžké chvosty, tj.*

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{m\log a}=1 \quad \text{pro nějaké } m>0, \tag{2.14}$$

*pak*

$$\lim_{a\to\infty} B(\bar{X}_n;a)=1. \tag{2.15}$$

*(iii) Nechť $\tilde{X}_n$ je výběrový medián. Jestliže $F$ splňuje buď (2.12) nebo (2.14), pak*

$$\frac{n}{2}\leq \overline{\lim}_{a\to\infty} B(\tilde{X}_n;a)\leq \frac{n}{2}+1 \quad \text{pro sudé } n, \tag{2.16}$$

$$\lim_{a\to\infty} B(\tilde{X}_n,a)=\frac{n+1}{2} \quad \text{pro liché } n. \tag{2.17}$$

**Poznámka 2.1** *Distribuční funkci s exponenciálními chvosty, splňující (2.12), označíme krátce jako typ I: mezi tato rozdělení patří např. normální ($r=2$), logistické a Laplaceovo ($r=1$) rozdělení. Distribuční funkci s těžkými chvosty, splňující (2.14), označíme krátce jako typ II: mezi tato rozdělení patří např. Cauchyho ($m=1$) nebo t-rozdělení o $m$ stupních volnosti $m>1$.*

**Důkaz věty 2.1.** (i) Stačí dokázat, že v případě $F$ s exponenciálními chvosty existuje střední hodnota

$$E_\varepsilon=\mathbb{E}_0\left[\exp\left\{n(1-\varepsilon)b|\bar{X}_n|^r\right\}\right]<\infty, \tag{2.18}$$

pro libovolné $\varepsilon \in (0,1)$. Skutečně, pak plyne z Markovovy nerovnosti

$$P_0(|\bar{X}_n| > a) \le E_\varepsilon \cdot \exp\{-n(1-\varepsilon)ba^r\}$$

$$\Longrightarrow \underline{\lim}_{a\to\infty} \frac{-\log P_0(|\bar{X}_n| > a)}{ba^r}$$

$$\ge \lim_{a\to\infty} \frac{n(1-\varepsilon)ba^r - \log E_\varepsilon}{ba^r} = n(1-\varepsilon),$$

a tedy tvrzení (2.13).

Konečnost střední hodnoty (2.18) dokážeme pomocí Hölderovy nerovnosti:

$$I\!E_0\left[\exp\left\{n(1-\varepsilon)b|\bar{X}_n|^r\right\}\right]$$

$$\le I\!E_0\left[\exp\left\{(1-\varepsilon)b\sum_{i=1}^n |X_i|^r\right\}\right] \tag{2.19}$$

$$\le \left(I\!E_0\left[\exp\{(1-\varepsilon)b|X_1|^r\}\right]\right)^n$$

$$= 2^n\left(\int_0^\infty \left[\exp\{(1-\varepsilon)bx^r\}\right] dF(x)\right)^n.$$

Z podmínky (2.12) vyplývá, že ke každé volbě $\varepsilon$ existuje $A_\varepsilon > 0$ tak, že pro $a \ge A_\varepsilon$ platí

$$1 - F(a) < \exp\left\{-(1-\tfrac{\varepsilon}{2})ba^r\right\}.$$

Poslední integrál v (2.19) můžeme postupně upravit následujícím způsobem:

$$\int_0^\infty \exp\{(1-\varepsilon)bx^r\}\, dF(x)$$



$$= \int_0^{A_\varepsilon} \exp\{(1-\varepsilon)bx^r\}\, dF(x)$$

$$- \int_{A_\varepsilon}^\infty \exp\{(1-\varepsilon)bx^r\}\, d(1-F(x))$$

$$= \int_0^{A_\varepsilon} \exp\{(1-\varepsilon)bx^r\}\, dF(x)$$

$$+(1-F(A_\varepsilon))\cdot \exp\{(1-\varepsilon)bA_\varepsilon^r\}$$

$$+\int_{A_\varepsilon}^\infty (1-F(x))(1-\varepsilon)brx^{r-1}\cdot \exp\{(1-\varepsilon)bx^r\}\, dx$$

$$\le \int_0^{A_\varepsilon} \exp\{(1-\varepsilon)bx^r\}\, dF(x) + \exp\left\{-\frac{\varepsilon}{2}bA_\varepsilon^r\right\}$$

$$+\int_{A_\varepsilon}^\infty (1-\varepsilon)brx^{r-1}\cdot \exp\left\{-\frac{\varepsilon}{2}bx^r\right\} dx < \infty$$

a odtud plyne tvrzení (i).

(ii) Nechť $F$ má těžké chvosty. Pak

$$P_0(|\bar{X}_n| > a) = P_0(\bar{X}_n > a) + P_0(\bar{X}_n < -a)$$

$$\ge P_0\Big(X_1 > -a, \ldots, X_{n-1} > -a, X_n > (2n-1)a\Big)$$

$$+P_0\Big(X_1 < a, \ldots, X_{n-1} < a, X_n < -(2n-1)a\Big)$$

$$= 2(F(a))^{n-1}[1 - F((2n-1)a)],$$

a tedy

$$\overline{\lim}_{a\to\infty} B(\bar{X}_n, a) \le \overline{\lim}_{a\to\infty} \frac{-\log[1 - F(2n-1)a]}{m\log a}$$

$$= \lim_{a\to\infty} \frac{-\log[1-F(2n-1)a]}{m\log((2n-1)a)} = 1.$$

(iii) Nechť $\tilde{X}_n$ je výběrový medián a $n$ je liché. Pak $\tilde{X}_n$ je prostřední pořádková statistika, $\tilde{X}_n = X_{(m)}$, $m = \frac{n+1}{2}$ a $F(\tilde{X}_n) = U_{(m)}$ má beta-rozdělení pravděpodobností, a platí

$$P_0(|\tilde{X}_n| > a) = P_0(\tilde{X}_n > a) + P_0(\tilde{X}_n < -a)$$
$$= 2n \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \int_{F(a)}^{1} u^{m-1}(1-u)^{m-1} du$$
$$\leq 2n \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} (1-F(a))^m,$$

a podobně

$$P_0(|\tilde{X}_n| > a) \geq 2n \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} (F(a))^{m-1}(1-F(a))^m,$$

což po zlogaritmování dává (2.17). Důkaz pro sudé $n$ je analogický.
■

### 2.3.4 Rozptyl asymptoticky normálního rozdělení

Jestliže odhad $T_n$ funkcionálu $T(\cdot)$ má asymptoticky normální rozdělení při $n \to \infty$,

$$\mathcal{L}_P\left(\sqrt{n}(T_n - T(P))\right) \to \mathcal{N}(0, V^2(P,T)),$$

pak vhodnou mírou robustnosti $T_n$ je supremum rozptylu $V^2(P,T)$ přes okolí $\mathcal{P}_0 \subset \mathcal{P}$ předpokládaného modelu,

$$\sigma^2(T) = \sup_{P\in\mathcal{P}_0} V^2(P,T).$$



Odhad, který minimalizuje $\sup_{P\in\mathcal{P}_0} V^2(P,T)$ přes určitou třídu $\mathcal{T}$ odhadů parametru $\theta$, se nazývá *minimaximálně robustní* ve třídě $\mathcal{T}$. Později ukážeme, že třídy M-odhadů, L-odhadů i R-odhadů obsahují minimálně robustní odhad parametru posunutí i regrese v množině kontaminovaných normálních rozdělení.

# Kapitola 3

# Robustní odhady reálného parametru

Mějme náhodný výběr $X_1, \ldots, X_n$ z populace s rozdělením pravděpodobností $P$; rozdělení je obecně neznámé, pouze předpokládáme, že jeho distribuční funkce $F$ patří do nějaké třídy $\mathcal{F}$ distribučních funkcí. Hledáme vhodný odhad parametru $\theta$, který lze vyjádřit jako funkcionál $T(P)$ rozdělení $P$. Tentýž parametr $\theta$ může být vyjádřen i více funkcionály: např. střed symetrie může být zároveň střední hodnotou, mediánem, modem rozdělení, a může být vyjádřen i jinými způsoby. Funkcionál $T(P)$ může být vyjádřen i implicitně jako řešení rovnice (soustavy rovnic) nebo minimalizační (maximalizační) úlohy: připomeňme si maximálně věrohodný odhad, odhad momentovou metodu aj. Odhad parametru $\theta$ získáme tak, že nahradíme $P$ v příslušném funkcionálu $T(\cdot)$ empirickým rozdělením příslušným vektoru pozorování $X_1, \ldots, X_n$.

Budeme se zabývat hlavně třemi nejrozšířenějšími třídami robustních odhadů reálného parametru: $M$-odhady, $L$-odhady a $R$-





odhady, které později rozšíříme na jiné modely, zejména na lineární regresní model.

## 3.1 $M$-odhady

Třídu $M$-odhadů zavedl P. J. Huber v práci [37] a vlastnosti $M$-odhadů jsou podrobně studovány v jeho knize [39]; viz také [3], [15], [19], [32], [46], [52], aj.

$M$-odhad $T_n$ je definován jako řešení minimalizační úlohy

$$\sum_{i=1}^{n} \rho(X_i, \theta) := \min \quad \text{vzhledem k} \quad \theta \in \mathbf{\Theta},$$

neboli (3.1)

$$\mathbb{E}_{P_n}\left[\rho(X, \theta)\right] = \min, \quad \theta \in \mathbf{\Theta},$$

kde $\rho(\cdot, \cdot)$ je vhodně zvolená funkce. V parametrickém modelu, kde rozdělení $P_\theta$ má hustotu $f(x, \theta)$, je speciálním případem $M$-odhadu i *maximálně věrohodný odhad,* který je řešením minimalizace

$$\sum_{i=1}^{n} (-\log\ f(X_i, \theta)) = \min, \quad \theta \in \mathbf{\Theta}.$$

Jestliže $\rho$ je diferencovatelná vzhledem k $\theta$ se spojitou derivací $\psi(\cdot, \theta) = \frac{\partial}{\partial\theta}\rho(\cdot, \theta)$, pak $T_n$ je řešením (případně jedním z řešení) rovnice

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0, \quad \theta \in \mathbf{\Theta}, \tag{3.2}$$

a tedy

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, T_n) = \mathbb{E}_{P_n}\left[\psi(X, T_n)\right] = 0, \quad T_n \in \mathbf{\Theta}. \tag{3.3}$$

Z (3.1) a (3.3) vyplývá, že statistický funkcionál, příslušný $T_n$, neboli $M$-funkcionál, je definován jako řešení minimalizace

$$\int_{\mathcal{X}} \rho(x, T(P))\ dP(x) = I\!E_P\left[\rho(X, T(P))\right] := \min, \quad T(P) \in \mathbf{\Theta} \tag{3.4}$$

nebo jako řešení rovnice

$$\int_{\mathcal{X}} \psi(x, T(P))\ dP(x) = I\!E_P\left[\psi(X, T(P))\right] = 0, \quad T(P) \in \mathbf{\Theta}. \tag{3.5}$$

Aby funkcionál $T(P)$ byl fisherovsky konsistentní, je třeba předpokládat, že úlohy (3.4) a (3.5) mají jediné řešení.

### 3.1.1 Influenční funkce $M$-odhadu

Předpokládejme, že $\rho(\cdot, \theta)$ je diferencovatelná, derivace $\psi(\cdot, \theta)$ je absolutně spojitá vzhledem k $\theta$ a rovnice (3.5) má jediné řešení $T(P)$. Nechť $P_t = (1-t)P + t\delta_x$; pak $T(P_t)$ je řešením rovnice

$$\int_{\mathcal{X}} \psi(y, T(P_t))d((1-t)P + t\delta_x) = 0,$$

tedy

$$(1-t)\int_{\mathcal{X}} \psi(y, T(P_t))\ dP(y) + t\psi(x, T(P_t)) = 0. \tag{3.6}$$

Derivujme (3.6) vzhledem k $t$ :

$$-\int_{\mathcal{X}} \psi(y, T(P_t))dP(y) + \psi(x, T(P_t))$$
$$+(1-t)\frac{dT(P_t)}{dt} \int_{\mathcal{X}} \left[\frac{\partial}{\partial\theta}\psi(y,\theta)\right]_{\theta=T(P_t)} dP(y)$$



$$+t\ \frac{dT(P_t)}{dt}\left[\frac{\partial}{\partial\theta}\psi(x,\theta)\right]_{\theta=T(P_t)} = 0.$$

Dosadíme-li $t = 0$, dostaneme influenční funkci

$$IC(x; T, P) = \frac{\psi(x, T(P))}{-\int_{\mathcal{X}} \dot{\psi}(y, T(P)dP(y)} \tag{3.7}$$

kde $\dot{\psi}(y, T(P) = \left[\frac{\partial}{\partial\theta}\psi(y,\theta)\right]_{\theta=T(P)}$.

#### $M$-odhad parametru posunutí

Důležitý speciální případ je model s parametrem posunutí $\theta$, ve kterém $X_1, \ldots, X_n$ jsou nezávislá pozorování se stejnou distribuční funkcí
$F(x-\theta)$, $\theta \in \mathbb{R}$; distribuční funkce $F$ je obecně neznámá. $M$-odhad $T_n$ je definován jako řešení minimalizace

$$\sum_{i=1}^{n} \rho(X_i - \theta) := \min, \tag{3.8}$$

a pokud $\rho(\cdot)$ je diferencovatelná s absolutně spojitou derivací $\psi(\cdot)$, je $T_n$ řešením rovnice

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0. \tag{3.9}$$

Aby byl příslušný $M$-funkcionál $T(F)$ fisherovsky konsistentní, je třeba předpokládat, že úloha $\int_{\mathcal{X}} \rho(x-\theta)dP(x) := \min$ má jediné řešení $\theta = 0$. Influenční funkce $T(F)$ pak je

$$IC(x; T, P) = \frac{\psi(x - T(P))}{\int_{\mathcal{X}} \psi'(y)dP(y)}. \tag{3.10}$$

Z (3.8) a (3.9) okamžitě vyplývá, že $T_n$ je *ekvivariantní vzhledem k posunutí*, tj. že splňuje

$$T_n(X_1 + c, \ldots, X_n + c) = T_n(X_1, \ldots, X_n) + c \ \forall c \in \mathbb{R}. \quad (3.11)$$

Na druhé straně, $T_n$ obecně *není ekvivariantní vzhledem k měřítku*, tj. obecně neplatí

$$T_n(cX_1, \ldots, cX_n) = cT_n(X_1 \ldots, X_n) \quad \text{pro } c > 0.$$

V symetrickém modelu volíme $\rho$ symetrickou kolem 0 ($\psi$ je pak lichá funkce). Jestliže je $\rho(x)$ ryze konvexní (a tedy $\psi(x)$ rostoucí), je i $\sum_{i=1}^n \rho(X_i - \theta)$ ryze konvexní v $\theta$ a $M$-odhad je určen jednoznačně. Jestliže je $\rho(\cdot)$ v některém úseku lineární, je $\psi(\cdot)$ v tomto úseku konstantní: pak rovnice $\sum_{i=1}^n \psi(X_i - \theta) = 0$ může mít vice kořenů a obvykle volíme jeden z těchto kořenů podle pravidla

$$\begin{aligned} T_n &= \tfrac{1}{2}(T_n^+ + T_n^-), \\ T_n^- &= \sup\{t : \ \sum_{i=1}^n \psi(X_i - t) > 0\}, \qquad (3.12) \\ T_n^+ &= \inf\{t : \ \sum_{i=1}^n \psi(X_i - t) < 0\}. \end{aligned}$$

Stejným způsobem určíme $M$-odhad v situaci, že $\psi$ je neklesající nespojitá funkce se skoky. Pokud je $\psi(\cdot)$ neklesající, ať už spojitá nebo se skoky, pak zřejmě platí pro libovolné $a \in \mathbb{R}$ :

$$P_\theta\Big(\sum_{i=1}^n \psi(X_i - a) > 0\Big) \le P_\theta(T_n > a) \le P_\theta(T_n \ge a)$$



$$\le P_\theta\Big(\sum_{i=1}^n \psi(X_i - a) \ge 0\Big) \qquad (3.13)$$

$$= P_\theta\Big(\sum_{i=1}^n \psi(X_i - a) > 0\Big) + P_\theta\Big(\sum_{i=1}^n \psi(X_i - a) = 0\Big);$$

pokud $P_\theta\Big(\sum_{i=1}^n \psi(X_i - a) = 0\Big) = 0$, přecházejí nerovnosti v (3.13) v rovnosti. Odtud dále dostáváme

$$\begin{aligned} &P_0\left\{n^{-\frac{1}{2}} \sum_{i=1}^n \psi\left(X_i - \frac{x}{\sqrt{n}}\right) < 0\right\} \\ &\le P_\theta(\sqrt{n}(T_n - \theta) < x) \le P_\theta(\sqrt{n}(T_n - \theta) \le x) \\ &\le P_0\left\{n^{-\frac{1}{2}} \sum_{i=1}^n \psi\left(X_i - \frac{x}{\sqrt{n}}\right) \le 0\right\}. \end{aligned}$$

Protože $n^{-\frac{1}{2}} \sum_{i=1}^n \psi\left(X_i - \frac{x}{\sqrt{n}}\right)$ je normovaný součet nezávislých stejně rozdělených náhodných veličin, můžeme nalézt asymptotické rozdělení pravděpodobností $\sqrt{n}(T_n - \theta)$ při $n \to \infty$, pro $\psi$ neklesající, podle centrální limitní věty.

**Bod selhání $M$-odhadu parametru posunutí** určíme podle paragrafu 2.3.2: Jestliže je funkce $\psi(\cdot)$ neohraničená, je $\varepsilon^* = \lim_{n\to\infty} \varepsilon_n^* = 0$. Naopak, je-li $\theta$ středem symetrie rozdělení pravděpodobností a funkce $\psi$ je ohraničená a lichá, je $\varepsilon^* = \lim_{n\to\infty} \varepsilon_n^* = \frac{1}{2}$. Třída $M$-odhadů tedy obsahuje robustní i nerobustní elementy.

**Příklad 3.1** (a) *Střední hodnota:*
Střední hodnotu $\theta = \mathbb{E}_P X$ lze chápat jako $M$-funkcionál s kriteriální funkcí $\rho(x) = x^2$, $\psi(x) = 2x$, $\psi'(x) = 2$, a podle (3.10)

dostaneme

$$IC(x;T,P)=\frac{2(x-\mathbb{E}_P(X))}{\int_{\mathbb{R}}2dP}=x-\mathbb{E}_P(X),$$

což je ve shodě s předcházejícími výsledky.

Příslušným $M$-odhadem střední hodnoty je aritmetický průměr $\bar{X}_n$ s bodem selhání $\varepsilon^*=\lim_{n\to\infty}\varepsilon_n^*=0$ a s globální citlivostí $\gamma^*=+\infty$.

(b) *Medián:*
Medián $\tilde{X}=F^{-1}(\frac{1}{2})$ lze chápat jako $M$-funkcionál s kriteriální funkcí $\rho(x)=|x|$ a výběrový medián $T_n=\tilde{X}_n$ je pak řešením minimalizace

$$\sum_{i=1}^{n}|X_i-\theta|:=\min,\quad \theta\in\mathbb{R}.$$

Předpokládejme, že rozdělení pravděpodobností $P$ má spojitou distribuční funkci $F$, ryze rostoucí v intervalu $(a,b)$, $-\infty\le a<b\le\infty$ a diferencovatelnou v okolí $\tilde{X}$. Nechť $F_t$ je distribuční funkce kontaminovaného rozdělení $P_t=(1-t)P+t\delta_x$. Medián $T(P_t)$ je řešením rovnice $F_t(u)=\frac{1}{2}$, tj.

$$(1-t)F(T(P_t))+tI[x<T(P_t)<\infty]=\tfrac{1}{2}.$$

Řešením této rovnice dostaneme

$$T(P_t)=\begin{cases}F^{-1}\left(\frac{1}{2(1-t)}\right) & \ldots\quad x>T(P_t)\\ F^{-1}\left(\frac{1-2t}{2(1-t)}\right) & \ldots\quad x\le T(P_t).\end{cases}$$

Funkce $T(P_t)$ je spojitá v bodě $t=0$, neboť $T(P_t)\to\tilde{X}=T(P)$ při $t\to 0$; s použitím rozvoje

$$\frac{1}{2(1-t)}=\frac{1}{2}+\frac{t}{2}+\mathcal{O}(t^2)\quad\text{a}\quad\frac{1-2t}{2(1-t)}=\frac{1}{2}-\frac{t}{2}+\mathcal{O}(t^2)$$



při $t\to 0$ dostaneme

$$\lim_{t\to 0}\tfrac{1}{t}[T(P_t)-F^{-1}(\tfrac{1}{2})]=\tfrac{1}{2}\ \text{sign}\ (x-F^{-1}(\tfrac{1}{2}))\left[\frac{dF^{-1}(u)}{du}\right]_{u=\frac{1}{2}},$$

a odtud dostaneme influenční funkci mediánu

$$IC(x;\tilde{X},F)=\frac{\text{sign}\ (x-\tilde{X})}{2f(\tilde{X})}.\tag{3.14}$$

Medián je robustní, neboť jeho influenční funkce je ohraničená, na rozdíl od střední hodnoty. Bod selhání mediánu je $\varepsilon^*=\frac{1}{2}$ a globální citlivost $\gamma^*=\frac{1}{2f(\tilde{X})}$, (pro standardní normální rozdělení $N(0,1)$ je $\gamma^*=1.253$).

Podle (3.14) je $(IF(x;\tilde{X},P))^2=\frac{1}{4f^2(\tilde{X})}=$ konst a lze dokázat, při $n\to\infty$ má $\sqrt{n}(\tilde{X}_n-\tilde{X})$ asymptoticky normální rozdělení,

$$\mathcal{L}\{\sqrt{n}(\tilde{X}_n-\tilde{X})\}\to\mathcal{N}\left(0,\frac{1}{4f^2(\tilde{X})}\right).$$

Speciálně, je-li $F$ distribuční funkce normálního rozdělení $\mathcal{N}(\mu,\sigma^2)$, je $f^2(\tilde{X})=f^2(\mu)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2$ a

$$\mathcal{L}\{\sqrt{n}(\tilde{X}_n-\tilde{X})\}\to\mathcal{N}\left(0,\frac{\pi}{2}\sigma^2\right).$$

(c) *Maximálně věrohodný odhad* parametru $\theta$ rozdělení pravděpodobností s hustotou $f(x,\theta)$ :

$$\rho(x,T(P))=-\log\ f(x,T(P)),$$

$$\psi(x,T(P))=-\frac{\partial}{\partial\theta}\log\ f(x,\theta)\Big|_{\theta=T(P)},$$

$$IF(x;T,P) = \frac{1}{\mathcal{I}_f(T(P))} \cdot \frac{\dot{f}(x,T(P))}{f(x,T(P))},$$

$$\text{kde } \dot{f}(x,T(P)) = \frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=T(P)};$$

$$\mathcal{I}_f(T(P)) = \int_{\mathcal{X}} \left[\frac{\partial}{\partial\theta} \log\ f(x,\theta)\Big|_{\theta=T(P)}\right]^2 f(x,T(P))dx$$

je Fisherova informace rozdělení $f$ v bodě $\theta = T(P)$.

### 3.1.2 Volba funkce $\psi$ u $M$-odhadu parametru posunutí

$M$-odhad je určen volbou kriteriální funkce $\rho$ nebo její derivace $\psi$. Jestliže parametr polohy je zároveň středem symetrie rozdělení pravděpodobností, volíme $\rho$ symetrickou podle nuly a tudíž $\psi$ lichou.

Podle (3.10) je influenční funkce $M$-odhadu úměrná $\psi(x - T(P))$; tedy, má-li být odhad robustní, musí být $\psi$ ohraničená. Uveďme příklady nejčastější volby funkce $\psi$ (a tedy i $\rho$), které se vyskytují v literatuře.

Střední hodnota je $M$-funkcionál s lineární, a tedy neohraničenou funkcí $\psi$. Příslušný $M$-odhad, $\bar{X}_n$, je maximálně věrohodným odhadem parametru polohy normálního rozdělení. Tento funkcionál je však úzce vázán na normální rozdělení a je velmi nerobustní. Hledáme-li $M$-odhad parametru polohy rozdělení pravděpodobností, vhodný pro okolí normálního rozdělení, použijeme funkci $\psi$, kterou navrhl a zdůvodnil P. J. Huber [37]. Tato funkce je lineární v ohraničeném intervalu $[-k,k]$, a konstantní vně tohoto intervalu. Kdybychom hledali rozdělení pravděpodobností s takovou věrohodnostní funkcí, zjistili bychom, že jeho hustota je normální v intervalu $[-k,k]$ a exponenciální vně:



$$\psi_H(x) = \begin{cases} x & \dots\ |x| \le k \\ k \text{ sign } x & \dots\ |x| > k, \end{cases} \tag{3.15}$$

kde $k > 0$ je pevně zvolená konstanta. Příslušný $M$-odhad, který se často vyskytuje v literatuře jako *Huberův odhad*, má ohraničenou influenční funkci, bod selhání $\varepsilon^* = \frac{1}{2}$, globální citlivost $\gamma^* = \frac{k}{2F(k)-1}$ a míru chvostů $\lim_{a\to\infty} B(a,T_n,F) = \frac{1}{2}$ pro rozdělení jak s exponenciálními, tak s těžkými chvosty. Je to tedy robustní odhad středu symetrie, necitlivý k extrémním a odlehlým pozorováním. Jak dokázal Huber [37], odhad generovaný funkcí (3.15) je minimaximálně robustní pro kontaminované normální rozdělení , přičemž hodnota $k$ závisí na podílu kontaminace.

Někteří autoři doporučují ještě více omezit vliv odlehlých pozorování volbou funkce $\psi(x)$, která konverguje k 0 při $x \to \pm\infty$, případně která je rovna 0 vně ohraničeného intervalu pokrývajícího 0. Takovou je např. *věrohodnostní funkce Cauchyho rozdělení*,

$$\psi_C(x) = -\frac{f'(x)}{f(x)} = \frac{2x}{1+x^2} \tag{3.16}$$

kde $f(x) = \frac{1}{\pi(1+x^2)}$ je hustota Cauchyho rozdělení; dále *Tukeyho biweight* funkce,

$$\psi_T(x) = \begin{cases} x\left[1 - \left(\frac{x}{k}\right)^2\right] & \dots\ |x| \le k \\ 0 & \dots\ |x| > k \end{cases} \tag{3.17}$$

nebo *Andrewsova sinusová funkce,*

$$\psi_A(x) = \begin{cases} \sin \frac{x}{k} & \ldots \quad |x| \leq k\pi \\ 0 & \ldots \quad |x| > k\pi. \end{cases} \tag{3.18}$$

Hampel [31] navrhl spojitou, po částech lineární funkci $\psi$, nulovou vně ohraničeného intervalu:

$$\psi_{HA}(x) = \begin{cases} |x| \text{ sign } x & \ldots \quad |x| < a \\ a \text{ sign } x & \ldots \quad a \leq |x| < b \\ \frac{c-|x|}{c-b}\, a \text{ sign } x & \ldots \quad b \leq |x| < c \\ 0 & \ldots \quad |x| > c. \end{cases} \tag{3.19}$$

V literatuře se také vyskytuje *skipped mean,* generovaný funkcí

$$\psi^*(x) = \begin{cases} x & \ldots \quad |x| \leq k \\ 0 & \ldots \quad |x| > k \end{cases} \tag{3.20}$$

nebo *skipped median,* generovaný funkcí

$$\tilde{\psi}(x) = \begin{cases} -1 & \ldots \quad -k \leq x < 0 \\ 0 & \ldots \quad |x| > k \\ 1 & \ldots \quad 0 \leq x \leq k. \end{cases} \tag{3.21}$$

Je však třeba si uvědomit, že tyto funkce nejsou monotonní a jim příslušné primitivní funkce $\rho$ nejsou konvexní. Vedle globálního minima může mít funkce $\sum_{i=1}^n \rho(X_i - \theta)$ lokální extrémy, které jsou dalšími kořeny rovnice $\sum_{i=1}^n \psi(X_i - \theta) = 0$. Poslední dvě funkce $\psi$ navíc nejsou spojité, tedy rovnice $\sum_{i=1}^n \psi(X_i - \theta) = 0$ obecně nemá řešení a $M$-odhad musí být hledán jako globální minimum funkce $\sum_{i=1}^n \rho(X_i - \theta)$.



### 3.1.3 Studentizované $M$-odhady

$M$-odhad parametru posunutí je ekvivariantní vzhledem k posunutí, ale obecně není ekvivariantní vzhledem k měřítku (viz (3.11)). K překonání tohoto nedostatku můžeme použít jedné z následujících metod:

- Zároveň s parametrem posunutí odhadujeme i měřítko: např. Huber [39] navrhuje zároveň s $\theta$ odhadnout parametr měřítka $\sigma$ řešením následující soustavy rovnic:

$$\sum_{i=1}^n \psi_H\left(\frac{X_i - \theta}{\sigma}\right) = 0 \tag{3.22}$$

$$\sum_{i=1}^n \chi\left(\frac{X_i - \theta}{\sigma}\right) = 0 \tag{3.23}$$

  kde $\chi(x) = \psi_H^2(x) - \int_{\mathbb{R}} \psi_H^2(y) d\Phi(y)$, $\psi_H$ je Huberova funkce (3.15) a $\Phi$ je distribuční funkce standardního normálního rozdělení.

- Odhad, ekvivariantní vzhledem k posunutí i měřítku získáme *studentizací* $M$-odhadu vhodnou škálovou (měřítkovou) statistikou $S_n = S_n(X_1, \ldots, X_n)$, splňující následující podmínky:

  (a) $S_n(\mathbf{x}) > 0$ *s.v.* pro $\mathbf{x} \in \mathbb{R}$

  (b) $S_n(x_1 + c, \ldots, x_n + c) = S_n(x_1, \ldots, x_n)$, $c \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$ *(invariance vzhledem k posunutí)*

  (c) $S_n(cx_1, \ldots, cx_n) = cS_n(x_1, \ldots, x_n)$, $c > 0$, $\mathbf{x} \in \mathbb{R}^n$ *(ekvivariance vzhledem k měřítku)*

Dále předpokládáme,

$$n^{\frac{1}{2}}(S_n - S(F)) = \mathcal{O}_p(1) \quad \text{při } n \to \infty \tag{3.24}$$

kde $S(F)$ je statistický funkcionál, příslušný $S_n$. Studentizovaný $M$-odhad je řešením minimalizace

$$\sum_{i=1}^{n} \rho\left(\frac{X_i - \theta}{S_n}\right) := \min, \quad \theta \in \mathbb{R}. \tag{3.25}$$

Takto definovaný odhad je skutečně ekvivariantní vzhledem k posunutí i k měřítku. Příslušný statistický funkcionál je definován implicitně jako řešení minimalizace

$$\int_{\mathcal{X}} \rho\left(\frac{x - t}{S(F)}\right) dF(x) := \min, \quad t \in \mathbb{R} \tag{3.26}$$

a funkcionál je fisherovsky konsistentní, pokud má minimalizace (3.26) jediné řešení. Pokud $\rho$ má spojitou derivaci $\psi$, je odhad též řešením rovnice

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta}{S_n}\right) = 0. \tag{3.27}$$

Pokud je $\rho$ konvexní a tedy $\psi$ je neklesající, ale nespojitá v některých bodech nebo konstantní na některých intervalech, uvažujeme studentizovaný odhad ve tvaru analogickém (3.12), tedy

$$\begin{aligned} T_n &= \tfrac{1}{2}(T_n^+ + T_n^-), \\ T_n^- &= \sup\{t : \sum_{i=1}^{n} \psi\left(\frac{X_i - t}{S_n}\right) > 0\} \\ T_n^+ &= \inf\{t : \sum_{i=1}^{n} \psi\left(\frac{X_i - t}{S_n}\right) < 0\}. \end{aligned} \tag{3.28}$$



Pozastavme se u volby škálové statistiky $S_n$. Na rozdíl od středu symetrie rozdělení pravděpodobností, který je zároveň průměrem, mediánem, modem atd., neexistuje univerzální měřítko, přesněji funkcionál měřítka, a volba určitého funkcionálu závisí na nás. Uveďme některé příklady:

- *Výběrová směrodatná odchylka:*

  $$S_n = \left(\tfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^{\frac{1}{2}},$$

  $$S(F) = (\mathrm{var}_F(X))^{\frac{1}{2}}.$$

  Protože tento funkcionál je nerobustní, používá se ke studentizaci jen ve speciálních případech.

- *Mezikvartilová odchylka:*

  $$S_n = X_{n:[\frac{3}{4}n]} - X_{n:[\frac{1}{4}n]},$$

  kde $X_{n:[np]}$, $0 < p < 1$ je empirický $p$-kvantil stanovený z uspořádaného výběru $X_{n:1} \leq \ldots \leq X_{n:n}$. Příslušný funkcionál má tvar

  $$S(F) = F^{-1}(\tfrac{3}{4}) - F^{-1}(\tfrac{1}{4}).$$

- *Mediánová absolutní odchylka (MAD):*

  $$S_n = \mathrm{med}_{1 \leq i \leq n}|X_i - \tilde{X}_n|.$$

  Příslušný statistický funkcionál $S(F)$ je řešením rovnice

  $$F\left(S(F) + F^{-1}(\tfrac{1}{2})\right) - F\left(-S(F) + F^{-1}(\tfrac{1}{2})\right) = \tfrac{1}{2}$$

  a pokud distribuční funkce $F$ je symetrická podle 0, a tedy $F^{-1}(\frac{1}{2}) = 0$, je $S(F) = F^{-1}(\frac{3}{4})$.

Lze ukázat, že v symetrickém modelu odpovídajícím $F(-x) = 1 - F(x)$, $\rho(-x) = \rho(x)$ a $\psi(-x) = -\psi(x)$, $x \in \mathbb{R}$, s absolutně spojitou $\psi$, má influenční funkce, studentizovaného $M$-funkcionálu tvar

$$IF(x, T, F) = \frac{S(F)}{\gamma(F)}\ \psi\left(\frac{x - T(F)}{S(F)}\right),$$

kde $\gamma(F) = \int_{\mathbb{R}} \psi'\left(\frac{y}{S(F)}\right) dF(y)$. To znamená, že v symetrickém modelu influenční funkce, $T(F)$ sice závisí na hodnotě $S(F)$, ale nezávisí na influenční funkci funkcionálu $S(F)$.

## 3.2 $L$-odhady

$L$-odhady jsou odhady, založené na uspořádaných pozorováních (pořádkových statistikách) $X_{n:1} \leq \ldots \leq X_{n:n}$, příslušné k náhodnému výběru $X_1, \ldots, X_n$. Obecný $L$-odhad píšeme ve tvaru

$$T_n = \sum_{i=1}^{n} c_{ni} h(X_{n:i}) + \sum_{j=1}^{k} a_j h^*(X_{n:[np_j]+1}), \tag{3.29}$$

kde $c_{n1}, \ldots, c_{nn}$ a $a_1, \ldots, a_k$ jsou dané koeficienty, $0 < p_1 < \ldots < p_k < 1$ a $h(\cdot)$ a $h^*(\cdot)$ dané funkce. Koeficienty $c_{ni}$, $1 \leq i \leq n$ jsou určeny ohraničenou váhovou funkcí $J : [0,1] \mapsto \mathbb{R}$ následujícím způsobem:

$$c_{ni} = \int_{\frac{i-1}{n}}^{\frac{i}{n}} J(s)ds, \quad i = 1, \ldots, n, \tag{3.30}$$

nebo přibližným způsobem

$$c_{ni} = \tfrac{1}{n} J\left(\tfrac{i}{n+1}\right), \quad i = 1, \ldots, n. \tag{3.31}$$



První složka $L$-odhadu (3.29) obecně zahrnuje všechny pořádkové statistiky, zatímco druhá složka je lineární kombinací konečně mnoha výběrových kvantilů. Řada $L$-odhadů má tvar pouze jedné ze složek ve (3.29) ($L$-odhad typu I a II).

Jednoduchými příklady $L$-odhadů jsou výběrový medián a střed rozpětí

$$T_n = \tfrac{1}{2}(X_{n:1} + X_{n:n}),$$

které odhadují parametr polohy, a dále např. výběrové rozpětí

$$R_n = X_{n:n} - X_{n:1}$$

a Giniho průměrná diference

$$G_n = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} |X_i - X_j| = \frac{2}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1) X_{n:i},$$

což jsou škálové statistiky.

Uvažujme $L$-odhad typu I s váhovou funkcí $J$ takovou, že $\int_0^1 J(u)du = 1$. Abychom nalezli příslušný statistický funkcionál, zavedeme empirickou kvantilovou funkci $Q_n(t) = F_n^{-1}(t)$, $0 < t < 1$ jako $Q_n(t) = \inf\{x : F_n(x) \geq t\}$, $0 < t < 1$. Tato funkce je empirickým protějškem kvantilové funkce $Q(t) = F^{-1}(t) = \inf\{x : F(x) \geq t\}$, $0 < t < 1$ a je rovna

$$Q_n(t) = \begin{cases} X_{n:i} & \ldots \quad \frac{i-1}{n} < t \leq \frac{i}{n}, \qquad i = 1, \ldots, n-1 \\ X_{n:n} & \ldots \quad \frac{n-1}{n} < t \leq 1. \end{cases} \tag{3.32}$$

Pomocí ní můžeme $L$-odhad vyjádřit alternativním způsobem

$$T_n = \int_0^1 J(s) h\left(Q_n(s)\right) ds \tag{3.33}$$

a jemu příslušný funkcionál má tvar

$$T(F) = \int_0^1 J(s)h\left(Q(s)\right)ds. \tag{3.34}$$

**Influenční funkce** $T(F)$ :
Předpokládejme, že $F$ je rostoucí a absolutně spojitá a funkce $h$ je absolutně spojitá. Označme

$$F_t(y) = (1-t)F(y) + t\delta_x = \begin{cases} (1-t)F(y) & y < x \\ (1-t)F(y) + t & y \geq x \end{cases}$$

Pak

$$F_t^{-1}(u) = \begin{cases} F^{-1}\left(\frac{u}{1-t}\right) & u \leq (1-t)F(x) \\ x & (1-t)F(x) < u \leq (1-t)F(x) + t \\ F^{-1}\left(\frac{u-t}{1-t}\right) & u>(1-t)F(x)+t, \end{cases}$$

a tedy

$$\frac{dF_t^{-1}(u)}{dt} = \begin{cases} \frac{u}{(1-t)^2} \cdot \frac{1}{f\left(F^{-1}\left(\frac{u}{1-t}\right)\right)} & u < (1-t)F(x) \\ \frac{u-1}{(1-t)^2} \cdot \frac{1}{f\left(F^{-1}\left(\frac{u-t}{1-t}\right)\right)} & u>(1-t)F(x)+t. \end{cases}$$

Odtud vyplývá

$$\frac{dT(F_t)}{dt} = \int_0^1 J(u)h'\left(F_t^{-1}(u)\right) \cdot \frac{dF_t^{-1}(u)}{dt}du$$

$$= \int_0^{F_t(x)} \frac{u}{(1-t)^2} \cdot \frac{h'\left(F^{-1}\left(\frac{u}{1-t}\right)\right)}{f\left(F^{-1}\left(\frac{u}{1-t}\right)\right)}J(u)du$$



$$+ \int_{F_t(x)}^1 \frac{u-1}{(1-t)^2} \cdot \frac{h'\left(F^{-1}\left(\frac{u-t}{1-t}\right)\right)}{f\left(F^{-1}\left(\frac{u-t}{1-t}\right)\right)}J(u)du$$

a influenční funkci funkcionálu (3.34) dostaneme při $t \to 0_+$ :

$$\begin{aligned} IF(x,T,F) &= \int_0^{F(x)} s \cdot \frac{h'(F^{-1}(u))}{f(F^{-1}(u))}J(u)du \\ &+ \int_{F(x)}^1 (u-1) \cdot \frac{h'(F^{-1}(u))}{f(F^{-1}(u))}J(u)du \\ &= \int_0^1 u \cdot \frac{h'(F^{-1}(u))}{f(F^{-1}(u))}J(u)du \\ &- \int_{F(x)}^1 \frac{h'(F^{-1}(u))}{f(F^{-1}(u))}J(u)du \\ &= \int_{-\infty}^{\infty} F(y)h'(y)J(F(y))dy - \int_x^{\infty} J(F(y))dy \end{aligned} \tag{3.35}$$

a tedy

$$\frac{d}{dx}IF(x,T,F) = h'(x)J(F(x)).$$

Ve speciálním případě $h(x) \equiv x,\ F(-x) = 1 - F(x),\ x \in \mathbb{R}$ a $J(u) = J(1-u),\ 0 < u < 1$, se influenční funkce zjednoduší:

$$IF(x,T,F) = \int_{-\infty}^{\infty} F(y)J(F(y))dy - \int_x^{\infty} J(F(y))dy$$

$$= \int_0^\infty F(y)J(F(y))dy$$

$$+ \int_{-\infty}^0 (1 - F(-y))J(1 - F(-y))dy$$

$$- \int_x^\infty J(F(y))dy = \int_0^\infty F(y)J(F(y))dy$$

$$+ \int_0^\infty (1 - F(y))J(F(y))dy - \int_x^\infty J(F(y))dy$$

$$= \int_0^\infty J(F(y))dy - \int_x^\infty J(F(y))dy$$

a tedy

$$IF(x,T,F) = \textstyle\int_0^x J(F(y))dF(y) \quad \ldots\ x \geq 0 \tag{3.36}$$

$$IF(-x,T,F) = -IF(x,T,F) \quad \ldots\ x \in \mathbb{R}$$

**Poznámka 3.1** *Nechť $M_n$ je $M$-odhad středu symetrie, vytvořený absolutně spojitou funkcí $\psi$ a nechť $L_n$ je $L$-odhad s váhovou funkcí $J(u) = c\,\psi'(F^{-1}(u))$. Pak $M_n$ a $L_n$ mají stejnou influenční funkci.*

**Bod selhání $L$-odhadu:** Jestliže $J(u) = 0$ pro $0 < u < \alpha$ a $1 - \alpha < u < 1$ a $\varepsilon_n^* = \frac{m_n}{n}$ je bod selhání $L$-odhadu (3.29), pak $\lim_{n\to\infty} \varepsilon_n^* = \alpha$.

**Příklad 3.2** (a) *$\alpha$-useknutý průměr* $(0 < \alpha < \frac{1}{2})$

$$\bar{X}_{n\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n:i}$$



$$c_{ni} = \begin{cases} \frac{1}{n-[n\alpha]} & \ldots\ [n\alpha] + 1 \leq i \leq n - [n\alpha] \\ 0 & \ldots\ \text{jinak} \end{cases}$$

$$J(u) = \tfrac{1}{1-2\alpha} I[\alpha \leq u \leq 1 - \alpha]$$

$$T_n = T(F_n) = \tfrac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F_n^{-1}(u)du$$

$$T(F) = \tfrac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(u)du.$$

*Influenční funkci* useknutého průměru vyjádříme pomocí (3.35):

$$IF(x,T,F) = \int_{\mathbb{R}} F(y)J(F(y))dy - \int_x^\infty J(F(y))dy$$

$$= \frac{1}{1-2\alpha}\left\{ \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F(y)dy - \int_x^\infty I[\alpha < F(y) < 1 - \alpha]dy \right\}$$

a tedy

$$IF(x,T,F) + \mu_\alpha =$$

$$= \begin{cases} -\frac{1}{1-2\alpha}\left[\alpha F^{-1}(1-\alpha) - (1-\alpha)F^{-1}(\alpha)\right] I\left[x < F^{-1}(\alpha)\right] \\ \frac{1}{1-2\alpha}\left[x - \alpha F^{-1}(\alpha) - \alpha F^{-1}(1-\alpha)\right] I\left[F^{-1}(\alpha) \leq x \leq F^{-1}(1-\alpha)\right] \\ \frac{1}{1-2\alpha}\left[-\alpha F^{-1}(\alpha) + (1-\alpha)F^{-1}(1-\alpha)\right] I\left[x > F^{-1}(1-\alpha)\right] \end{cases}$$

kde

$$\mu_\alpha = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(u)du = \frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y dF(y).$$

Speciálně, pro $F$ symetrickou splňující $F(x)+F(-x)=1\ \forall x$ a $F^{-1}(u)=-F^{-1}(1-u),\ 0<u<1$, je $\mu_\alpha=0$ a

$$IF(x,T,F)=\begin{cases} -\frac{F^{-1}(1-\alpha)}{1-2\alpha} & \ldots\ x<-F^{-1}(1-\alpha) \\ \frac{x}{1-2\alpha} & \ldots\ {\scriptstyle -F^{-1}(1-\alpha)\leq x\leq F^{-1}(1-\alpha)} \\ \frac{F^{-1}(1-\alpha)}{1-2\alpha} & \ldots\ x>F^{-1}(1-\alpha). \end{cases}$$

Globální citlivost useknutého průměru je

$$\gamma^*=\frac{F^{-1}(1-\alpha)}{1-2\alpha}.$$

**Poznámka 3.2** *Nechť $M_n$ je Huberův odhad středu symetrie $\theta$ rozdělení $F(x-\theta)$, vytvořený Huberovou funkcí $\psi_H$ s $k=F^{-1}(1-\alpha)$ (viz (3.15). Pak $M_n$ a $\bar{X}_{n\alpha}$ mají stejnou influenční funkci.*

**Poznámka 3.3** *(i) Bod selhání $\alpha$-useknutého průměru $\bar{X}_{n\alpha}$ je $\lim_{n\to\infty}\varepsilon_n^*=\alpha$.*

*(ii) Nechť $\alpha=[k/n]$, $n\geq 3$ a nechť $B(\bar{X}_{n\alpha};a)$ je míra chvostů $\bar{X}_{n\alpha}$, definovaná v (2.9). Pak*

$$n-2k\leq \underline{\lim}_{a\to\infty}B(\bar{X}_{n\alpha};a)\leq \overline{\lim}_{a\to\infty}B(\bar{X}_{n\alpha};a)\leq n-k \quad (3.37)$$

*pokud $F$ má exponenciální chvosty (2.12), zatímco pro $F$ s těžkými chvosty (2.14) platí*

$$\lim_{a\to\infty}B(\bar{X}_{n\alpha};a)=k+1 \quad (3.38)$$

*pokud $k<\frac{n-1}{2}$.*



**Příklad 3.3** *$\alpha$-winsorizovaný průměr:*

$$\begin{aligned}\bar{W}_{n\alpha}=T(F_n)&=\frac{1}{n}\Big\{[n\alpha]X_{n:[n\alpha]+1}+\sum_{i=[n\alpha]+1}^{n-[n\alpha]}X_{n:i} \\ &+[n\alpha]X_{n:n-[n\alpha]}\Big\} \qquad (3.39)\\ &=\alpha F_n^{-1}(\alpha)+\int_\alpha^{1-\alpha}F_n^{-1}(u)du+\alpha F_n^{-1}(1-\alpha)\\ &=\sum_{i=1}^n c_{ni}X_{n:i}+\frac{[n\alpha]+1}{n}X_{n:[n\alpha]+1}+\frac{[n\alpha]+1}{n}X_{n:n-[n\alpha]}\end{aligned}$$

kde

$$c_{ni}=\begin{cases}\frac{1}{n} & \ldots\ 1+[n\alpha]<i<n-[n\alpha]\\ 0 & \ldots\ \text{jinak.}\end{cases}$$

To znamená, že extrémní kvantily nejsou useknuty, ale jsou nahrazeny kvantilem $X_{n:[n\alpha]+1}$ nebo $X_{n:n-[n\alpha]}$. Pro jednoduchost uvažujme model se symetrickou distribuční funkcí $F$. Statistický funkcionál $T(F)$ má tvar

$$T(F)=T_1(F)+T_2(F)=\int_\alpha^{1-\alpha}F^{-1}(u)du+\alpha F^{-1}(\alpha)$$
$$+\alpha F^{-1}(1-\alpha).$$

Influenční funkce $T_1(F)$ plyne z (3.35), zatímco influenční funkce $T_2(F)$ je modifikací influenční funkce mediánu (3.14), který je kvantilem s $\alpha=\frac{1}{2}$; tedy průměru z (3.36):

$$IF(x,\bar{W}_{n\alpha},F)=$$

$$= \begin{cases} F^{-1}(\alpha) - \frac{\alpha}{f(F^{-1}(\alpha))}\, I[x < F^{-1}(\alpha)] \\ x\, I[F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha)] \\ F^{-1}(1-\alpha) + \frac{\alpha}{f(F^{-1}(1-\alpha))} I[x > F^{-1}(1-\alpha)]. \end{cases}$$

Globální citlivost winsorizovaného průměru je

$$\gamma^* = F^{-1}(\alpha) + \frac{\alpha}{f(F^{-1}(1-\alpha))}$$

a bod selhání $\varepsilon^* = \alpha$. Zatímco influenční funkce $\alpha$-useknutého průměru je spojitá, influenční funkce winsorizovaného průměru má body nespojitosti $F^{-1}(\alpha)$ a $F^{-1}(1-\alpha)$.

Jako další příklady uveďme

*Senův vážený průměr* (Sen [64]):

$$T_{n,k} = \begin{pmatrix} n \\ 2k+1 \end{pmatrix}^{-1} \sum_{i=k+1}^{n-k} \begin{pmatrix} i \\ k \end{pmatrix} \begin{pmatrix} n-i+1 \\ k \end{pmatrix} X_{n:i+1},$$

kde $0 < k < \frac{n}{2}$. Všimněme si, že $T_{n,0} = \bar{X}_n$ a $T_{n,k} = \tilde{X}_n$ pro $k = [(n+1)/2]$;

*Harrell-Davisův odhad p-kvantilu* [33]:

$$T_n = \sum_{i=1}^{n} c_{ni} X_{n:i},$$

$$c_{ni} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \int_{(i-1)/n}^{i/n} u^{k-1}(1-u)^{n-k} du,$$

$i = 1, \dots, n$, kde $k = [np]$, $0 < p < 1$.



*BLUE (asymptotically best linear unbiased estimator)* odhad parametru polohy ([41], [42], [10]). Nechť $X_1, X_2, \dots$ jsou nezávislá pozorovaní s distribuční funkcí $F(x-\theta)$, kde $F$ má absolutně spojitou hustotu $f$
s derivací $f'$. Pak *BLUE* je $L$-odhad s váhovou funkcí

$$T_n = \sum_{i=1}^{n} c_{ni} X_{n:i},\ c_{ni} = \frac{1}{n} J\left(\frac{i}{n+1}\right),\ i = 1, \dots, n$$

$$J(F(x)) = \psi_f'(x),\ \psi_f(x) = -\frac{f'(x)}{f(x)},\ x \in \mathbb{R}.$$

## 3.3 $R$-odhady

Uvažujme náhodný výběr $X_1, \dots, X_n$ z populace se spojitou distribuční funkcí. Nechť $R_i$ je pořadí $X_i$ mezi
$X_1, \dots, X_n$, $i = 1, \dots, n$. Formálně lze pořadí vyjádřit ve tvaru

$$R_i = \sum_{j=1}^{n} I[X_j \le X_i],\ i = 1, \dots, n, \tag{3.40}$$

a tedy $R_i = nF_n(X_i)$, $i = 1, \dots, n$, kde $F_n$ je empirická distribuční funkce $X_1, \dots, X_n$. Pořadí jsou invariantní ke třídě ryze monotonních transformací pozorovaní a pořadové testy mají mnoho výhodných vlastností, z nichž nejdůležitější je, že rozdělení testového kriteria za platnosti hypotézy nezávisí na distribuční pozorování.

Hodges a Lehmann [36] navrhli třídu odhadů, tzv. *R-odhadů*, které jsou inverzí pořadových testů.

Omezme se na situaci, kdy
$X_1, \dots, X_n$ mají spojitou distribuční funkci $F(x-\theta)$ se středem

symetrie $\theta$. Hypotézu

$$\mathbf{H}_0: \ \theta = \theta_0$$

o středu symetrie testujeme *znaménkovým pořadovým testem* (jinak jednovýběrovým pořadovým testem), založeným na statistice

$$S_n(\theta_0) = \text{sign } (X_i - \theta_0) a_n(R^+_{ni}(\theta_0)) \tag{3.41}$$

kde $R^+_{ni}(\theta_0)$ je pořadí $|X_i - \theta_0|$ mezi $|X_1 - \theta_0|, \ldots, |X_n - \theta_0|$ a $a_n(1) \leq \ldots \leq a_n(n)$ jsou dané *skóry*, obvykle generované neklesající skórovou funkcí $\varphi^+ : \ [0,1) \mapsto \mathbb{R}^+, \ \varphi^+(0) = 0$ jako $a_n(i) = \varphi^+\left(\frac{i}{n+1}\right), \ i = 1, \ldots, n$. Jestliže např. volíme $a_n(i) = \frac{i}{n+1}, \ i = 1, \ldots, n$, dostáváme *Wilcoxonův jednovýběrový test*. Jestliže platí $\theta = \theta_0, \ F(x) + F(-x) = 1, \ x \in \mathbb{R}$, jsou sign $(X_i - \theta_0)$ a $R^+_{ni}(\theta_0)$ stochasticky nezávislé a $S_n(t)$ je neklesající a schodovitá funkce $t$. Odtud plyne, že $I\!E_{\theta_0} S_n(\theta_0) = 0$ a rozdělení $S_n(\theta_0)$ je za platnosti $\mathbf{H}_0$ symetrické kolem 0. Jako odhad $\theta_0$ navrhujeme hodnotu $t$, která je řešením rovnice $S_n(t) = 0$. Taková rovnice ovšem nemusí mít řešení, protože $S_n(t)$ je nespojitá; podobně jako u $M$-odhadů tedy definujeme $R$-odhad ve tvaru

$$T_n = \tfrac{1}{2}(T_n^- + T_n^+), \tag{3.42}$$

$$T_n^- = \sup\{t : S_n(t) > 0\}, \qquad T_n^+ = \inf\{t : S_n(t) < 0\}.$$

Jestliže $a_n(i) = 1, \ i = 1, \ldots, n$, je $T_n$ rovno výběrovému mediánu. Odhad, odpovídající jednovýběrovému Wilcoxonovu testu se skóry $a_n(i) = \frac{i}{n+1}, \ i = 1, \ldots, n$, se nazývá *Hodges-Lehmannův odhad*. Dá se ukázat, že Hodges-Lehmannův odhad lze vyjádřit explicitně; je roven

$$T_{nH} = \text{med}\left\{\frac{X_i + X_j}{2} : \ 1 \leq i \leq j \leq n\right\}. \tag{3.43}$$



Ostatní $R$-odhady, s výjimkou mediánu a Hodges-Lehmannova odhadu, se nedají vyjádřit explicitně a musí být počítány iteračně.

Na rozdíl of $M$-odhadů jsou $R$-odhady ekvivariantní nejen vzhledem k posunutí v poloze, ale také ke změně měřítka, tj. platí

$$T_n(X_1 + c, \ldots, X_n + c) = T_n(X_1, \ldots, X_n) + c, \ c \in \mathbb{R} \tag{3.44}$$

$$T_n(cX_1, \ldots, cX_n) = cT_n(X_1, \ldots, X_n), \ c > 0.$$

Distribuční funkce statistiky $S_n(\theta)$ není spojitá, i když $X_1, \ldots, X_n$ mají spojitou distribuční funkci $F(x-\theta)$. Jestliže však $\theta$ je skutečný střed symetrie, pak distribuční funkce statistiky $S_n(\theta)$ nezávisí na $F$. Označíme-li

$$p_n = P_\theta\left(S_n(\theta) = 0\right) = P_0\left(S_n(0) = 0\right),$$

pak $0 \leq p_n < 1$ a $\lim_{n\to\infty} p_n = 0$ a

$$\tfrac{1}{2}(1 - p_n) \leq P_\theta(T_n < \theta) \leq P_\theta(T_n \leq \theta) \leq \tfrac{1}{2}(1 + p_n). \tag{3.45}$$

To znamená, že v případě symetrické $F$ je $T_n$ *mediánově nestranným* odhadem $\theta$, tj. $\theta = \text{med}_\theta \ T_n$.

Jestliže vyjádříme pořadí $R_i^+$ ve (3.41) podle (3.40), vidíme, že Hodges-Lehmannův odhad $T_n$ lze alternativně vyjádřit jako řešení rovnice

$$\int_{-\infty}^{\infty} [F_n(y) - F_n(2T_n - y)] dF_n(y) = 0 \tag{3.46}$$

a obecně, $R$-odhad vytvořený skórovou funkcí $\varphi^+$ lze vyjádřit jako řešení rovnice

$$\int_{-\infty}^{\infty} \varphi\left(F_n(y) - F_n(2T_n - y)\right) dF_n(y) = 0, \tag{3.47}$$

kde $\varphi(u) = \operatorname{sign}(u-\frac{1}{2})\varphi^+(2u-1),\ 0 < u < 1$. Příslušný statistický funkcionál je tedy řešením rovnice

$$\int_{-\infty}^{\infty} \varphi\left(F(y) - F(2T(F) - y)\right) dF(y) \tag{3.48}$$
$$= \int_0^1 \varphi\left(u - F(2T(F) - F^{-1}(u))\right) du = 0.$$

Influenční funkci $T(F)$ odvodíme z (3.48) analogickým způsobem, jako jsme odvodili influenční funkci $L$-odhadu, a pro $F$ symetrickou s absolutně spojitou hustotou $f$ dostaneme

$$IF(x,T,F) = \frac{\varphi(F(x))}{\int_{\mathbb{R}} \varphi(F(y))(-f'(y))dy}. \tag{3.49}$$

**Poznámka 3.4** *Jestliže $\psi(x) = c\varphi(F(x)),\ x \in \mathbb{R}$, pak $M$-odhad vytvořený funkcí $\psi$ a $R$-odhad vytvořený funkcí $\varphi$ mají stejné influenční funkce.*

Na závěr porovnáme některé numerické charakteristiky výběrového průměru $\bar{X}_n$, výběrového mediánu $\tilde{X}_n$, 5%-useknutého průměru $\bar{X}_{.05}$, 10%-useknutého průměru $\bar{X}_{.10}$, 5%-winsorizovaného průměru $W_{.05}$ a Hodges-Lehmannova odhadu HL:



| Odhad | $\gamma^*$ | $\lambda^*$ | $\varepsilon^*$ | $\mathrm{var}_{\mathcal{N}}$ | $\mathrm{var}_{c\mathcal{N}}$ |
|---|---|---|---|---|---|
| $\bar{X}_n$ | $\infty$ | 1 | 0 | 1 | $\infty$ |
| $\tilde{X}_n$ | $\sqrt{\frac{\pi}{2}}$ | $\infty$ | $\frac{1}{2}$ | $\frac{\pi}{2}$ | 1.74 |
| $\bar{X}_{.05}$ | 1.83 | 1.11 | 0.05 | 1.03 | 1.30 |
| $\bar{X}_{.10}$ | 1.60 | 1.25 | 0.10 | 1.26 | 1.26 |
| $W_{.05}$ | 2.13 | $\infty$ | 0.05 | 1.01 | 1.46 |
| HL | 1.77 | 1.41 | 0.29 | 1.05 | 1.29 |

Zde značíme

- $\gamma^*$ ... globální citlivost,
- $\lambda^*$ ... lokální citlivost,
- $\varepsilon^*$ ... bod selhání,
- $\mathrm{var}_{\mathcal{N}}$ - asymptotický rozptyl za normálního rozdělení $\mathcal{N}(0,1)$,
- $\mathrm{var}_{c\mathcal{N}}$ ... asymptotický rozptyl za kontaminovaného normálního rozdělení $0.95\ \mathcal{N}(0,1) + 0.05\ \mathcal{N}(0,\sigma^2)$, $\sigma^2 \to \infty$.

## 3.4 Asymptotické vlastnosti $M$-, $L$- a $R$-odhadů

Robustní odhady jsou nelineárními funkcemi pozorování, často definované implicitně. Odvodit jejich distribuční funkci při konečném počtu pozorování je velmi obtížné; proto ji aproximujeme limitní distribuční funkcí odpovídající neomezeně rostoucímu počtu pozorování $n \to \infty$. Limitní rozdělení je většinou normální a rozptyl asymptoticky normálního rozdělení je důležitou charakteristikou odhadu.

Asymptoticky normální rozdělení robustních odhadů nemůžeme odvodit přímo použitím centrální limitní věty, protože nejsou lineárními kombinacemi nezávislých náhodných veličin. Nejprve musíme $\sqrt{n}(T_n - T(F))$ lineární kombinací nezávislých náhodných veličin aproximovat.

Připomeňme si rozvoj (1.15), který platí pro fréchetovsky diferencovatelné funkcionály $T(P)$. Tento rozvoj můžeme přepsat pomocí influenční funkce $IF(x, T, P)$ ve tvaru

$$\sqrt{n}(T_n - T(F)) = \tfrac{1}{n}\sum_{i=1}^{n} IF(X_i, T, F) + R_n, \tag{3.50}$$

kde $R_n = o_p(1)$. Podobný rozvoj, který nazýváme *asymptotickou reprezentací* odhadu $T_n$, lze odvodit i pro funkcionály, které nejsou fréchetovsky diferencovatelné, různými metodami a za nejrůznějších podmínek na hladkost distribuční funkce $F$ a skórové funkce odhadu $(\psi, J, \varphi)$. Různé formy asymptotických reprezentací robustních odhadů jsou odvozeny v knize [46].

Jestliže pro odhad $T_n$ platí reprezentace (3.50), pak $T_n$ má asymptotické rozdělení pravděpodobností při $n \to \infty$ v tom smyslu, že

$$\mathcal{L}\left\{\sqrt{n}(T_n - T(F))\right\} \to \mathcal{N}(0, \sigma_F^2), \tag{3.51}$$

kde $\sigma_F^2 = I\!E_F(IF(X, T, F))^2$. Aplikujme tento výsledek na $M$-, $L$- a $R$-odhady, jejichž influenční funkce jsme odvodili. Podrobné podmínky, za nichž tyto asymptotické výsledky platí, lze nalézt v [46].



### 3.4.1 $M$-odhady

#### $M$-odhad obecného skalárního parametru

Nechť $\{X_i,\ i = 1, 2, \ldots,$ je posloupnost nezávislých pozorování se stejnou distribuční funkcí $F(x, \theta),\ \theta \in \mathbf{\Theta}$, kde $\mathbf{\Theta}$ je otevřený interval $\mathbb{R}^1$. $M$-odhad parametru $\theta$ je řešením minimalizace

$$\sum_{i=1}^{n} \rho(X_i, \theta) = \min, \ \theta \in \mathbf{\Theta}.$$

Předpokládejme, že $\rho(x, \theta)$ je absolutně spojitá v $\theta$ s derivací $\psi(x, \theta) = \frac{\partial}{\partial\theta}\rho(x, \theta)$. Jestliže $\psi(x, \theta)$ je spojitá v $\theta$, pak hledáme $M$-odhad $T_n$ mezi kořeny rovnice

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0. \tag{3.52}$$

Jestliže funkce $I\!E_\theta \rho(X, t)$ má jediné minimum v bodě $t = \theta$ (fisherovská konsistence) a jsou splněny další podmínky buď na hladkost $\psi(x, \theta)$ nebo $F(x, \theta)$, pak existuje posloupnost $\{T_n\}$ kořenů rovnice (3.52) taková, že při $n \to \infty$

$$\sqrt{n}(T_n - \theta) = 0_p(1), \tag{3.53}$$

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}\gamma(\theta)} \sum_{i=1}^{n} \psi(X_i, \theta) + \mathcal{O}_p(n^{-1/2}),$$

$$\text{kde } \gamma(\theta) = I\!E_\theta \dot\psi(X, \theta), \ \dot\psi(x, \theta) = \frac{\partial}{\partial\theta}\psi(x, \theta).$$

Odtud dále vyplývá, že $\sqrt{n}(T_n - \theta)$ má asymptotické normální rozdělení

$$\mathcal{N}(0, \sigma^2(\psi, F)), \quad \text{kde} \ \ \sigma^2(\psi, F)) = \frac{I\!E_\theta(\psi^2(X, \theta)}{\gamma^2(\theta)}. \tag{3.54}$$

**$M$-odhady parametru posunutí**

Nechť $X_1, X_2, \dots$ jsou nezávislá pozorování s distribuční funkcí $F(x-\theta)$. $M$-odhad $\theta$ je řešením minimalizace

$$\sum_{i=1}^{n} \rho(X_i - \theta) = \min, \ \theta \in \mathbb{R}^1.$$

Předpokládejme, že $\rho(x)$ je absolutně spojitá s derivací $\psi(x)$ a že funkce $h(t) = \int_{\mathbb{R}} \rho(x-t)dF(x)$ má jediné minimum v bodě $t=0$. Jestliže $\psi$ je absolutně spojitá s derivací $\psi'$ a $\gamma = \int \psi'(x)dF(x) > 0$, pak existuje posloupnost $\{T_n\}$ kořenů rovnice $\sum_{i=1}^{n} \psi(X_i - t) = 0$ taková, že při $n \to \infty$

$$\sqrt{n}(T_n - \theta) = \mathcal{O}_p(1), \tag{3.55}$$

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}\gamma} \sum_{i=1}^{n} \psi(X_i - \theta) + \mathcal{O}_p(n^{-1/2})$$

$$P_\theta\left(\sqrt{n}(T_n - \theta) \le x\right) \to \Phi\left(\frac{x}{\sigma^2(\psi, F)}\right),$$

kde $\sigma^2(\psi, F) = \gamma^{-2} \int_{\mathbb{R}} \psi^2(x)dF(x)$ a $\Psi$ je distribuční funkce normálního rozdělení $\mathcal{N}(0,1)$. Pokud $F$ má absolutně spojitou hustotu $f$ s derivací $f'$ a konečnou Fisherovu informaci $\mathcal{I}(F) = \int [f'(x)/f(x)]^2 dF(x)$, pak při speciální volbě $\rho(x) = -\log\ f(x)$ je $M$-odhad roven maximálně věrohodnému odhadu $\theta$, jehož asymptotický rozptyl je roven Rao-Cramérově dolní hranici $1/\mathcal{I}(F)$.

Jestliže $\psi(x)$ má body nespojitosti, je třeba, aby distribuční funkce $F$ měla dvě derivace $f, f'$ v jejich okolí. $M$-odhad je určen jednoznačně, pokud $\psi$ je neklesající, a to vztahy (3.12). Pak řešení



$T_n$ úlohy $\sum_{i=1}^{n} \rho(X_i - \theta) :=$ min není obecně kořenem rovnice $\sum_{i=1}^{n} \psi(X_i - \theta) = 0$, ale platí

$$n^{-1/2} \sum_{i=1}^{n} \psi(X_i - T_n) = \mathcal{O}_p(n^{-1/2}) \ \text{ při } \ n \to \infty \tag{3.56}$$

a

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}\gamma^*} \sum_{i=1}^{n} \psi(X_i - \theta) + \mathcal{O}_p(n^{-1/4}),$$

$$\gamma^* = \int_{\mathbb{R}} f(x)d\psi(x), \tag{3.57}$$

$$P_\theta\left(\sqrt{n}(T_n - \theta) \le x\right) \to \Phi\left(\frac{x}{\sigma^2(\psi, F)}\right),$$

kde $\sigma^2(\psi, F) = (\gamma^*)^{-2} \int_{\mathbb{R}} \psi^2(x)dF(x)$ a $\Psi$ je distribuční funkce normálního rozdělení $\mathcal{N}(0,1)$.

Více o asymptotických reprezentacích $M$-odhadů, jakož i asymptotické reprezentace studentizovaných $M$-odhadů lze nalézt v [46].

### 3.4.2 $L$-odhady

Nechť $X_1, X_2, \dots,$ jsou nezávislá pozorování s distribuční funkcí $F$. Nejprve uvažujme lineární kombinaci pořádkových statistik $T_n = \sum_{i=1}^{n} c_{ni} X_{n:i}$ s koeficienty generovanými váhovou funkcí $J$ buď podle (3.30) nebo podle (3.31) ($L$-odhad typu I). Omezíme se na useknuté $L$-odhady splňující $J(u) = 0$ pro $0 \le u < \alpha$ a $1-\alpha < u \le 1$, $0 < \alpha < \frac{1}{2}$. Předpokládejme, že distribuční funkce $F$ je skoro všude spojitá a $F^{-1}(u)$ je lipschitzovská v okolí bodů nespojitosti funkce $J$, kterých je nejvýše konečně mnoho. Pak při $n \to \infty$

$$\sqrt{n}(T_n - T(F)) = n^{-1/2} \sum_{i=1}^{n} \psi_1(X_i) + \mathcal{O}_p(n^{-1/2}),$$

$$T(F) = \int_0^1 J(u) F^{-1}(u) du, \tag{3.58}$$

$$\psi_1(x) = -\int_{\mathbb{R}} \{I[y \geq x] - F(y)\} J(F(y)) dy, \ \ x \in \mathbb{R}$$

a $\sqrt{n}(T_n - T(F))$ má asymptoticky normální rozdělení $\mathcal{N}\left(0, \sigma^2(J,F)\right)$, kde

$$\sigma^2(J,F) = \int_{\mathbb{R}} \psi_1^2(x) dF(x)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} J(F(x)) J(F(y)) [F(x \wedge y) - F(x)F(y)] dx dy.$$

Jestliže distribuční funkce $F$ má absolutně spojitou hustotu $f$ s derivací $f'$ a konečnou Fisherovu informaci $\mathcal{I}(F) = \int [f'(x)/f(x)]^2 dF(x)$, pak volba váhové funkce

$$J(u) = J_F(u) = \frac{\psi'(F^{-1}(u))}{\mathcal{I}(F)},$$

$$0 < u < 1, \quad \psi(x) = \frac{f'(x)}{f(x)}, \ \ x \in \mathbb{R} \tag{3.59}$$

vede k asymptoticky vydatnému $L$-odhadu s asymptotickým rozptylem

$$\sigma^2(J,F) = \frac{1}{\mathcal{I}(F)}.$$



Všimněme si, že pokud je $J_F(u) = 0$ pro $0 < u < \alpha$ a $1-\alpha < u < 1$, je $\frac{f'(x)}{f(x)} = \frac{d \log f(x)}{dx} =$ konst pro $x < F^{-1}(\alpha)$ a $x > F^{-1}(1-\alpha)$, a tedy chvosty hustoty $f$ klesají exponenciálně k 0.

Uvažujme odhad typu II, tj. lineární kombinaci konečně mnoha kvantilů $T_n = \sum_{i=1}^{k} a_j X_{n:[np_j]+1}$, $0 < p_1 < \ldots < p_k < 1$. Předpokládejme, že $F$ je dvakrát diferencovatelná v $F^{-1}(p_j)$ a $F'\left(F^{-1}(p_j)\right) > 0$, $j = 1, \ldots, k$. Pak při $n \to \infty$

$$\sqrt{n}\Big(T_n - \sum_{j=1}^{k} a_j F^{-1}(p_j)\Big) = n^{-1/2} \sum_{i=1}^{n} \psi_2(X_i) + R_n,$$

$$R_n = \mathcal{O}\left(n^{-1/4} (\log\ n)^{1/2} (\log\log\ n)^{1/4}\right) \text{ skoro jistě}, \tag{3.60}$$

$$\psi_2(x) = \sum_{j=1}^{k} \frac{a_j}{F'\left(F^{-1}(p_j)\right)} \left\{p_j - I\left[x \leq F^{-1}(p_j)\right]\right\},$$

$x \in \mathbb{R}$, a $\sqrt{n}\Big(T_n - \sum_{j=1}^{k} a_j F^{-1}(p_j)\Big)$ má asymptoticky normální rozdělení $\mathcal{N}\left(0, \int_{\mathbb{R}} \psi_2^2(x) dF(x)\right)$.

### 3.4.3 $R$-odhady

Uvažujme $R$-odhad $T_n$ středu symetrie $\theta$ distribuční funkce $F(x-\theta)$, vytvořený pořadovou statistikou $S_n(t)$ (3.41) pomocí vztahů (3.42), se skórovou funkcí $\varphi(u)$, neklesající a integrabilní se čtvercem, $0 < u < 1$. Předpokládejme, že $F$ má absolutně spojitou hustotu $f$ a konečnou Fisherovu informaci $\mathcal{I}(F)$. Pak při $n \to \infty$

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}\gamma} \sum_{i=1}^{n} \varphi(F(X_i - \theta)) + o_p(1), \tag{3.61}$$

kde $\gamma = \int_{\mathbb{R}} \varphi(F(x))(-f'(x))dx$, a tedy $\sqrt{n}(T_n - \theta)$ má asymptoticky normální rozdělení $\mathcal{N}\left(0, \gamma^{-2} \int_0^1 \varphi^2(u)du\right)$. Speciálně, jestliže volíme

$$\varphi(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \quad 0 < u < 1,$$

dostaneme asymptoticky vydatný $R$-odhad s asymptotickým rozptylem $1/\mathcal{I}(F)$.

### 3.4.4 Asymptotické vztahy $M$-, $L$- a $R$-odhadů

Nechť $X_1, X_2, \ldots$ je posloupnost nezávislých pozorování s distribuční funkcí $F(x-\theta)$, $F(x) + F(-x) = 1$ $\forall x$, a nechť $\{T_{1n}\}$ a $\{T_{2n}\}$ jsou dvě posloupnosti odhadů $\theta$. Jestliže $\sqrt{n}(T_{jn} - \theta)$ má při $n \to \infty$ asymptoticky normální rozdělení $\mathcal{N}(0, \sigma_j^2)$, $j = 1, 2$, pak podíl rozptylů $e_{1,2} = \sigma_1^2/\sigma_2^2$ nazýváme *asymptotickou relativní vydatností* $\{T_{2n}\}$ vzhledem k $\{T_{1n}\}$. Alternativně, jestliže $\{T_{2n'}\}$ je založeno na $n'$ pozorováních, pak $\sqrt{n}(T_{2n'} - \theta)$ má asymptoticky normální rozdělení $\mathcal{N}(0, \sigma_1^2)$, stejně jako $\sqrt{n}(T_{1n} - \theta)$, jestliže posloupnost $n' = n'(n)$ je volena tak, že existuje limita

$$\lim_{n\to\infty} \frac{n}{n'(n)} = \frac{\sigma_1^2}{\sigma_2^2} = e_{1,2}.$$

Jestliže $e_{1,2} = 1$, znamená to, že $\{T_{1n}\}$ a $\{T_{2n}\}$ jsou stejně asymptoticky vydatné. V takovém případě dále srovnáváme $\{T_{1n}\}$ a $\{T_{2n}\}$ pomocí tzv. *deficience* $\{T_{2n}\}$ vzhledem k $\{T_{1n}\}$ : jestliže platí

$$\mathbb{E}_\theta\left[n(T_{nj} - \theta)^2\right] = \tau^2 + \frac{a_j}{n} + o(n^{-1}), \; j = 1, 2,$$

pak deficiencí $\{T_{2n}\}$ vzhledem k $\{T_{1n}\}$ nazýváme

$$d_{1,2} = \frac{a_2 - a_1}{\tau^2}.$$



Jestliže $n'(n)$ zvolíme tak, že

$$\mathbb{E}_\theta[n(T_{2n'} - \theta)^2] = \mathbb{E}_\theta[n(T_{1n} - \theta)^2] + \mathcal{O}(n^{-1}),$$

pak

$$d_{1,2} = \lim_{n\to\infty} [n'(n) - n].$$

V předcházejících paragrafech jsme viděli, že $M$- a $L$-odhady založené na pozorováních s distribuční funkcí $F$ mají stejné influenční funkce $IF(x, T_1, F)$
$\equiv IF(x, T_2, F)$, pokud $J(u) = \psi'(F^{-1}(u))$, $0 < u < 1$. Podobné vztahy platí i mezi $M$- a $R$-odhady a $L$- a $R$-odhady. V těchto úvahách můžeme pokračovat dále: z asymptotických reprezentací paragrafů 3.4.1–3.4.3 plyne, že tyto odhady nejen mají stejné influenční funkce, ale pokud $\{T_{n1}\}$ a $\{T_{n2}\}$ mají stejné asymptotické reprezentace, (až na tvar zbytku), pak jsou asymptoticky blízké ve snyslu

$$\sqrt{n}(T_{2n} - T_{1n}) = R_n = o_p(1) \quad \text{při} \quad n \to \infty, \tag{3.62}$$

V tom případě říkáme, že posloupnosti odhadů $\{T_{n1}\}$ a $\{T_{n2}\}$ jsou *asymptoticky ekvivalentní*. Další informaci o vztahu $\{T_{n1}\}$ a $\{T_{n2}\}$ získáme, podaří-li se nám odvodit přesný řád zbytku $R_n$ ve (3.62), případně jeho asymptotické rozdělení, po vynásobení vhodnou mocninou $n$. Toto rozdělení už ovšem není normální.

Pro úplnost shrňme nejzajímavější z těchto asymptotických vztahů.

#### $M$- a $L$-odhady

Nechť $X_1, X_2, \ldots$ jsou nezávislé náhodné veličiny se stejnou distribuční funkcí $F(x-\theta)$ takovou, že $F(x) + F(-x) = 1$, $x \in \mathbb{R}$;

nechť $X_{n:1} \leq X_{n:2} \leq \ldots \leq X_{n:n}$ jsou pořádkové statistiky příslušné $X_1, \ldots, X_n$.

**I.** Nechť $M_n$ je $M$-odhad $\theta$ generovaný neklesající schodovitou funkcí $\psi$

$$\psi(x) = \alpha_j \ \ldots \ s_j < x < s_{j+1}, \ j = 1, \ldots, k, \tag{3.63}$$

kde

$$-\infty = s_0 < s_1 < \ldots < s_k < s_{k+1} = \infty,$$

$$-\infty < \alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_k < \infty,$$

$$\alpha_j = -\alpha_{k-j+1}, \ s_j = -s_{k-j+1}, \ j = 1, \ldots, k,$$

a alespoň dvě z čísel $\alpha_j$ jsou různá. To znamená, že $M_n$ je řešením minimalizace $\sum_{i=1}^n \rho(X_i - t) = \min$, kde $\rho$ je spojitá, konvexní, symetrická a po částech lineární funkce s derivací $\rho' = \psi$ *s.v.* Předpokládejme, že $F$ má dvě ohraničené derivace $f, f'$, $f$ kladnou, v okolí $s_1, \ldots, s_k$.

Pak $L$-odhad $L_n$, asymptoticky ekvivalentní $M_n$, je lineární kombinace konečně mnoha kvantilů,
$L_n = \sum_{j=1}^k a_j X_{n:[np_j]}$, kde

$$p_j = F(s_j), \quad a_j = \frac{1}{\gamma}(\alpha_j - \alpha_{j-1}) f(s_j),$$

$$\gamma = \sum_{j=1}^k (\alpha_j - \alpha_{j-1}) f(s_j) \ (> 0); \tag{3.64}$$

a platí $M_n - L_n = \mathcal{O}_p\left(n^{-\frac{3}{4}}\right)$ při $n \to \infty$.



**II.** Předpokládejme, že $F$ má absolutně spojitou symetrickou hustotu $f$ a konečnou Fisherovu informaci $\mathcal{I}(F)$. Nechť $M_n$ je Huberův $M$-odhad $\theta$, generovaný funkcí $\psi$

$$\psi(x) = \begin{cases} x & \ldots \ |x| \leq c \\ c \cdot \operatorname{sign} x & \ldots \ |x| > c, \end{cases}$$

kde $c > 0$, a nechť $L_n$ je $\alpha$-useknutý průměr,

$$L_n = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n:i},$$

kde $\alpha = 1 - F(c)$. Jestliže $F$ dále splňuje $f(x) > a > 0$ a $f'(x)$ existuje pro

$$F^{-1}(\alpha - \varepsilon) < x < F^{-1}(1 - \alpha + \varepsilon), \ \varepsilon > 0,$$

pak při $n \to \infty$

$$M_n - L_n = \mathcal{O}_p\left(n^{-1}\right). \tag{3.65}$$

**III.** Nechť $L_n$ je $\alpha$-winsorizovaný průměr

$$L_n = \frac{1}{n}\left\{[n\alpha] X_{n:[n\alpha]+1} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n:i} + [n\alpha] X_{n:n-[n\alpha]}\right\}.$$

Pak za stejných podmínek jako ve **II** platí

$$M_n - L_n = \mathcal{O}_p\left(n^{-\frac{3}{4}}\right), \ n \to \infty, \tag{3.66}$$

kde $M_n$ je $M$-odhad vytvořený funkcí

$$\psi(x) = \begin{cases} F^{-1}(\alpha) - \frac{\alpha}{f(F^{-1}(\alpha))} & x < F^{-1}((\alpha) \\ x & F^{-1}(\alpha) \leq x \leq F^{-1}(1-\alpha) \\ F^{-1}(1-\alpha) + \frac{\alpha}{f(F^{-1}(\alpha))} & x > F^{-1}(1-\alpha). \end{cases}$$

**IV.** Nechť $L_n = \sum_{i=1}^{n} c_{ni} X_{n:i}$, kde koeficienty $c_{ni}$ jsou generovány funkcí $J : (0,1) \mapsto \mathbb{R}$ takovou, že

$$J(1-u) = J(u),\ 0 < u < 1,\ \int_0^1 J(u)du = 1,$$

$$J(u) = 0 \ \text{ pro } \ u \in (0,\alpha) \cup (1-\alpha, 1),\ 0 < \alpha < \tfrac{1}{2},$$

$J$ je spojitá v $(0,1)$ až na konečně mnoho bodů$s_1, \ldots, s_m$, kde $\alpha < s_1 \ldots < s_m < 1-\alpha$, a $J$ je lipschitzovská v intervalech $(\alpha, s_1), (s_1, s_2), \ldots,$ $(s_m, 1-\alpha)$.
O distribuční funkci $F$ předpokládáme, že má symetrickou hustotu a že $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ je lipschitzovská v okolí $s_1, \ldots, s_m$, a

$$\int_{-A}^{A} f^2(x)dx < \infty, \ \text{ kde } \ A = F^{-1}(1-\alpha+\varepsilon),\ \varepsilon > 0.$$

Pak asymptoticky ekvivalentní $M$-odhad $M_n$ je vytvořený funkcí

$$\psi(x) = -\int_{\mathbb{R}} \left(I[y \geq x] - F(y)\right) J(F(y))dy,\ x \in \mathbb{R}$$

a platí

$$M_n - L_n = \mathcal{O}_p\left(n^{-1}\right),\ n \to \infty. \tag{3.67}$$

### $M$- a $R$-odhady

Nechť $X_1, X_2, \ldots$ jsou nezávislé náhodné veličiny se stejnou distribuční funkcí $F(x-\theta)$ takovou, že $F(x) + F(-x) = 1,\ x \in \mathbb{R}$. Předpokládejme, že $F$ má absolutně spojitou hustotu $f$ a konečnou Fisherovu informaci $\mathcal{I}(F)$. Nechť $\varphi : (0,1) \mapsto \mathbb{R}$ je neklesající skórová funkce, $\varphi(1-u) = -\varphi(u),\ 0 < u < 1$ a $\int_0^1 \varphi^2(u)du < \infty$.
Nechť

$$\gamma = -\int_0^1 \varphi(F(x))f'(x)dx \neq 0.$$

Nechť $R_n$ je $R$-odhad , definovaný v (3.41) a (3.42) se skóry $a_n(i) = \varphi^+\left(\frac{i}{n+1}\right),\ i = 1, \ldots, n$, kde $\varphi^+(u) = \varphi\left(\frac{u+1}{2}\right)$,
$0 \leq u < 1$. Nechť $M_n$ je $M$-odhad vytvořený funkcí $\psi(x) = c\varphi(F(x)),\ x \in \mathbb{R},\ c > 0$. Pak

$$M_n - R_n = o_p\left(n^{-\frac{1}{2}}\right),\ n \to \infty. \tag{3.68}$$

Speciálně, Hodges-Lehmannův $R$-odhad je vytvořen skórovou funkcí $\varphi(u) = u - \frac{1}{2},\ 0 < u < 1$, a tedy asymptoticky ekvivalentní $M$-odhad je vytvořen $\psi$-funkcí $\psi(x) = F(x) - \frac{1}{2},\ x \in \mathbb{R}$.



### $R$- a $L$-odhady

Kombinací předcházejících výsledků dostaneme asymptotické vztahy mezi $R$- a $L$-odhady; nemusíme je tedy podrobně rozepisovat. Jako zajímavý příklad uveďme $R$-odhad, asymptoticky ekvivalentní $\alpha$-useknutému průměru, který je generovaný skórovou funkcí

$$\varphi(u) = \begin{cases} F^{-1}(\alpha) & \ldots \ 0 < u < \alpha \\ F^{-1}(u) & \ldots \ \alpha \leq u \leq 1-\alpha \\ F^{-1}(1-\alpha) & \ldots \ 1-\alpha < u < 1. \end{cases}$$

## 3.4.5 Minimaximálně robustní odhady

Většina odhadů $T_n = T(F_n)$ má asymptoticky normální rozdělení, tj. při $n \to \infty$ rozdělení $\sqrt{n}(T_n - T(F))$ konverguje k normálnímu

rozdělení
$\mathcal{N}(0, V_{as}(F,T))$, kde $V_{as}(F,T) = \int_{\mathbb{R}} IF^2(x,T,F)dF(x)$.

Jakožto míru robustnosti funkcionálu $T$ (a odhadu $T_n$) můžeme uvažovat maximum asymptotického rozptylu
$V_{as}(F,T)$ přes určitou třídu $\mathcal{F}$ distribučních funkcí:

$$\sigma^2(T) = \sup_{F\in\mathcal{F}} V_{as}(F,T).$$

Na druhé straně, uvažujme určitou třídu funkcionálů $\mathcal{T}$, např. $M$-funkcionálů, a hledejme funkcionál $T_0$ takový, že $\sigma^2(T_0) \leq \sigma^2(T)\ \forall T \in \mathcal{T}$. Jestliže takový funkcionál existuje, nazývá se *minimaximálně robustní,* neboť splňuje

$$\sigma^2(T_0) = \inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} V_{as}(F,T). \tag{3.69}$$

Uvažujme speciální případ odhadu parametru polohy. Nechť $X_1,\ldots,X_n$ je náhodný výběr z rozdělení s distribuční funkcí $F(x-\theta)$, kde $\theta$ je neznámý parametr a $F$ je neznámý prvek systému distribučních funkcí $\mathcal{F}$. Nejčastěji se uvažují následující třídy $\mathcal{F}$ :

(i) *Kontaminační model:*

$$\mathcal{F}_G = \{F:\ F = (1-\varepsilon)G + \varepsilon H,\ H \in \mathcal{P}\}, \tag{3.70}$$

kde $G$ je pevná distribuční funkce, $\varepsilon \in [0,1)$ je pevné číslo a $H$ probíhá pevnou třídu $\mathcal{P}$ distribučních funkcí.

(ii) *Kolmogorův model:*

$$\mathcal{F}_G = \left\{F:\ \sup_{x\in\mathbb{R}} |F(x) - G(x)| \leq \varepsilon\right\},\ \varepsilon \in [0,1) \text{ pevné.} \tag{3.71}$$



Nechť $F_0 \in \mathcal{F}$ je distribuční funkce, která minimalizuje Fischerovu informaci na $\mathcal{F}$ (nejméně příznivé rozdělení systému $\mathcal{F}$), tj.

$$\mathcal{I}(F_0) = \int_{\mathbb{R}} \left(\frac{f_0'(x)}{f_0(x)}\right)^2 dF_0 = \min_{F\in\mathcal{F}} \mathcal{I}(F).$$

Nechť $T_0$ je prvek třídy odhadů $\mathcal{T}$, který je asymptotickým odhadem $\theta$ pro distribuční funkci $F_0$, tj. $V_{as}(F_0,T_0) = \frac{1}{\mathcal{I}(F_0)}$. Jestliže dále platí

$$\frac{1}{\mathcal{I}(F_0)} = V_{as}(F_0,T_0) \geq \sup_{F\in\mathcal{F}} V_{as}(F,T_0),$$

pak

$$\inf_{T\in\mathcal{T}} \sup_{F\in\mathcal{F}} V_{as}(F,T) = \frac{1}{\mathcal{I}(F_0)}$$
$$\text{tj.} \tag{3.72}$$
$$V_{as}(F_0,T) \geq V_{as}(F_0,T_0) \geq V_{as}(F,T_0)$$

$\forall T \in \mathcal{T}$ a $\forall F \in \mathcal{F}$. Minimaximálně robustní odhad existuje mezi $M$-, $L$- i $R$-odhady v symetrickém kontaminačním modelu (Huber [37], Jaeckel [40]).

### Minimaximálně robustní $M$-, $L$- a $R$-odhady

Uvažujme kontaminační model (3.70), kde $G$ je symetrická *jednovrcholová* distribuční funkce s dvakrát diferencovatelnou hustotou $g$ takovou, že
$(-\log\ g(x))$ je konvexní v $x$; nechť $H$ probíhá symetrické distribuční funkce; označme tento systém $\mathcal{F}_1$. Nechť $T(F)$ je $M$-funkcionál, definovaný jako kořen rovnice $\int_{\mathbb{R}} \psi(x - T(F)) = 0$.

Pak

$$V_{as}(F,T) = \frac{\int_{\mathbb{R}} \psi^2(x - T(F))dF(x)}{\left(\int_{\mathbb{R}} \psi'(x - T(F))dF(x)\right)^2} \geq \frac{1}{\mathcal{I}(F)}.$$

Huber [37] dokázal, že nejméně příznivé rozdělení třídy $\mathcal{F}_1$ má hustotu

$$f_0(x) = \begin{cases} (1-\varepsilon)g(x_0)\mathrm{e}^{k(x-x_0)} & \dots \quad x \leq x_0 \\ (1-\varepsilon)g(x) & \dots \quad x_0 \leq x \leq x_1 \\ (1-\varepsilon)g(x_1)\mathrm{e}^{-k(x-x_1)} & \dots \quad x \geq x_1 \end{cases} \tag{3.73}$$

kde

$$x_0 = -x_1 = \inf\left\{x: \ -\frac{g'(x)}{g(x)} \geq -k\right\}$$

a $k > 0$ je určeno vztahem

$$\frac{2}{k}g(x_1) + \int_{x_0}^{x_1} g(x)dx = \frac{1}{1-\varepsilon}$$

a $T_n$ je maximálně věrohodný odhad pro rozdělení $f_0$, tedy $M$-odhad generovaný funkcí

$$\psi_0(x) = -\frac{f_0'(x)}{f_0(x)} = \begin{cases} -k & \dots \quad x \leq x_0 \\ -\frac{g'(x)}{g(x)} & \dots \quad x_0 < x < x_1 \\ k & \dots \quad x \geq x_1. \end{cases}$$

Z asymptotických vztahů v \$ 3.4.4 hned plyne, že existují
i minimaximálně robustní L- a R- odhady; speciálně, minimaximálně robustní $L$-odhad je vytvořen váhovou funkcí

$$J_0(u) = \frac{1}{\mathcal{I}(F_0)}\psi_0'(F^{-1}(u)), \ 0 < u < 1$$



a minimaximálně robustní $R$-odhad je vytvořen skórovou funkcí

$$\varphi_0(u) = \psi_0(F_0^{-1}(u)), \ 0 < u < 1.$$

Důležitý speciální případ je minimaximálně robustní odhad v modelu *kontaminovaného normálního rozdělení:* v modelu (3.70) položme $G \equiv \Phi$, kde $\Phi$ je distribuční funkce $\mathcal{N}(0,1)$. Pak nejméně příznivé rozdělení má hustotu

$$f_0(x) = \begin{cases} \frac{1-\varepsilon}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2} & \dots \quad |x| \leq k \\ \frac{1-\varepsilon}{\sqrt{2\pi}}\mathrm{e}^{-k^2/2-k|x|} & \dots \quad |x| > k, \end{cases} \tag{3.74}$$

a tedy je normální v centrální části $[-k,k]$ a exponenciální vně tohoto intervalu. Věrohodnostní funkce, příslušná $f_0$, je

$$\psi_0(x) = -\frac{f_0'(x)}{f_0(x)} = \begin{cases} x & \dots \quad |x| \leq k \\ k \ \mathrm{sign}\ x & \dots \quad |x| > k \end{cases}$$

což je známá Huberova funkce. Konstanta $k > 0$ je určena vztahem

$$2\Phi(k) - 1 + 2\frac{\Phi'(k)}{k} = \frac{1}{1-\varepsilon}.$$

Minimaximálně robustní $M$-odhad pro kontaminované normální rozdělení je generovaný funkcí $\psi_0$ a je shodný s maximálně věrohodným odhadem příslušným hustotě $f_0$. Minimaximálně robustní $L$-odhad je vytvořen váhovou funkcí $J_0$, která musí splňovat

$$J_0\left(F_0(x)\right) = \frac{1}{\mathcal{I}(F_0)}I[-k \leq x \leq k], \ x \in \mathbb{R},$$

a tedy je rovna

$$J_0(u) = \frac{1}{\mathcal{I}(F_0)}I\left[F_0^{-1}(-k) \leq u \leq F_0^{-1}(k)\right], \ 0 < u < 1.$$

Příslušný $L$-odhad je $\alpha$-useknutý průměr, kde $\alpha = F_0^{-1}(-k)$. Podobně minimaximálně robustní $R$-odhad pro kontaminované normální rozdělení je vytvořen skórovou funkcí
$\varphi_0(u) = \psi_0\left(F_0^{-1}(u)\right),\ 0 < u < 1.$

# Kapitola 4

# Robustní odhady v lineárním modelu

## Úvod

Uvažujme lineární regresní model

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + U_i, \ i = 1, \ldots, n, \tag{4.1}$$

kde $Y_1, \ldots, Y_n$ jsou pozorování, $\boldsymbol{\beta} \in \mathbb{R}^p$ je neznámý parametr, $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ jsou pevně dané vektory nebo náhodné pozorovatelné vektory (*regresory*) a $U_1, \ldots, U_n$ jsou vzájemně nezávislé náhodné chyby se stejnou distribuční funkcí $F$. Distribuční funkce $F$ je obecně neznámá; jen předpokládáme, že patří do určitého systému $\mathcal{F}$ distribučních funkcí.

Označíme-li

$$\mathbf{Y} = (Y_1, \ldots, Y_n)',$$





$$\mathbf{X} = \mathbf{X}_n = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix},$$

$$\mathbf{U} = (U_1, \ldots, U_n)',$$

můžeme (4.1) přepsat v maticovém tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}. \tag{4.2}$$

Nejznámějším odhadem $\boldsymbol{\beta}$ je klasický odhad *metodou nejmenších čtverců* $\widehat{\boldsymbol{\beta}}$. Pokud $\mathbf{X}$ je nenáhodná a má hodnost $p$, je $\widehat{\boldsymbol{\beta}}$ roven

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{4.3}$$

Je-li $F$ normální, je $\widehat{\boldsymbol{\beta}}$ maximálně věrohodným odhadem $\boldsymbol{\beta}$. Pro obecnou distribuční funkci $F$, která má konečný druhý moment, je podle známé Gauss-Markovovy věty $\widehat{\boldsymbol{\beta}}$ nejlepším nestranným lineárním odhadem $\boldsymbol{\beta}$. Protože $\widehat{\boldsymbol{\beta}}$ je rozšířením výběrového průměru na lineární regresní model, má i podobné vlastnosti, zejména je velmi nerobustní a citlivý k odlehlým pozorováním $Y_i$, k odchylkám od normálního rozdělení chyb $U_i$ a selhává, pokud toto rozdělení má těžké chvosty. Avšak navíc je odhad $\widehat{\boldsymbol{\beta}}$ v lineárním regresním modelu silně ovlivněn regresní maticí $\mathbf{X}$ a je velmi citlivý k odlehlým hodnotám jejích elementů.

Chyby, způsobené odchylkami od předpokládaného modelu a od předpokládaného rozdělení pravděpodobností v lineárních modelech, zejména ekonometrických, mohou mít dalekosáhlejší důsledky než v modelu s parametrem posunutí. Proto právě zde musíme hledat robustní alternativy ke klasickým odhadům, jejichž hlavním představitelem je odhad metodou nejmenších ctverců.

Než zavedeme robustní alternativy metody nejmenších čtverců, ukážeme, v čem spočívá vliv odlehlých prvků regresní matice $\mathbf{X}$ na chování odhadu $\widehat{\boldsymbol{\beta}}$.

## 4.1 Metoda nejmenších čtverců

Jestliže odhadneme $\boldsymbol{\beta}$ metodou nejmenších čtverců, pak regresní nadrovina prochází body $(\mathbf{x}_i, \hat{Y}_i),\ i = 1, \ldots, n$, kde

$$\hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \mathbf{h}_i'\mathbf{Y},\ i = 1, \ldots, n,$$

a $\mathbf{h}_i'$ je $i$-tý řádek projekční matice $\widehat{\mathbf{H}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Tedy $\hat{\mathbf{Y}} = \widehat{\mathbf{H}}\mathbf{Y}$ je projekcí vektoru $\mathbf{Y}$ do prostoru nad sloupci matice $\mathbf{X}$. Protože $\widehat{\mathbf{H}}$ je projekční matice, platí $\mathbf{h}_i'\mathbf{h}_j = h_{ij},\ i, j = 1, \ldots, n$, a tedy

$$0 \le \sum_{k \ne i} h_{ik}^2 = h_{ii}(1 - h_{ii}) \Longrightarrow 0 \le h_{ii} \le 1,\ i = 1, \ldots, n,$$

$$\Longrightarrow |h_{ij}| \le \|\mathbf{h_i}\|\|\mathbf{h}_j\| = (h_{ii}h_{jj})^{\frac{1}{2}} \le 1,\ i, j = 1, \ldots, n. \tag{4.4}$$

Matice $\widehat{\mathbf{H}}$ je řádu $n \times n$ a hodnosti $p$; její diagonální prvky leží v mezích $0 \le h_{ii} \le 1,\ i = 1, \ldots, n$ a stopa $trace(\widehat{\mathbf{H}}) = \sum_{i=1}^n h_{ii} = p$. Jestliže se stane, že $h_{ii} = 1$ pro nějaké $i$, pak

$$1 = h_{ii} = \|\mathbf{h}_i\|^2 = \sum_{k=1}^n h_{ik}^2 = 1 + \sum_{k \ne i} h_{ij}^2$$

$$\Longrightarrow h_{ij} = 0 \quad \text{pro} \quad j \ne i,$$

což znamená, že

$$\hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \mathbf{h}_i'\mathbf{Y} = h_{ii}Y_i = Y_i,$$



a regresní nadrovina prochází bodem $(\mathbf{x}_i, Y_i)$, bez ohledu na hodnoty ostatních pozorování. Hodnota $h_{ii} = 1$ je extrémní případ, který však ukazuje, že vysoká hodnota diagonálního prvku $h_{ii}$ matice $\widehat{\mathbf{H}}$ způsobuje, že regresní nadrovina prochází v blízkosti bodu $(\mathbf{x}_i, Y_i)$. Takový bod proto nazýváme *vlivným* (*leverage*) bodem množiny pozorování. V literatuře není shoda v názoru, kterou hodnotu $h_{ii}$ je třeba považovat za vysokou. Je však známo, (viz např. [39]), že pokud $\mathbb{E}U_i = 0$ a $0 < \sigma^2 = \mathbb{E}U_i^2 < \infty,\ i = 1, \ldots, n$, pak

$$\lim_{n\to\infty} \max_{1 \le i \le n} h_{ii} = 0$$

je nutnou a postačující podmínkou k tomu, aby platilo

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 \to 0,$$

$$\mathcal{L}\left\{(\mathbf{X}'\mathbf{X})^{-1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\right\} \to \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_p\right)$$

při $n \to \infty$, kde $\mathbf{I}_p$ je jednotková matice řádu $p$.

Uvažujme, jaký vliv může mít maximální diagonální prvek matice $\widehat{\mathbf{H}}$ na pravděpodobnost velkých hodnot residuí odhadu $\widehat{\boldsymbol{\beta}}$; zdá se nám, že právě zde je vliv diagonály $\mathbf{X}$ na $\widehat{\boldsymbol{\beta}}$ nejnázornější.

Předpokládejme, že distribuční funkce $F$ je symetrická podle nuly, tj. $F(x) + F(-x) = 1,\ x \in \mathbb{R}$, a má nedegenerované chvosty, tj. $0 < F(x) < 1,\ x \in \mathbb{R}$. Uvažujme následující míru chvostů odhadu $\widehat{\boldsymbol{\beta}}$

$$B(a, \widehat{\boldsymbol{\beta}}) = \frac{-\log P_{\boldsymbol{\beta}}\left(\max_i |\mathbf{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})| > a\right)}{-\log(1 - F(a))}. \tag{4.5}$$

Přirozeně očekáváme, že

$$\lim_{a\to\infty} P_{\boldsymbol{\beta}}\left(\max_i |\mathbf{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})| > a\right) = 0$$

a zajímá nás, kdy je tato konvergence nejrychlejší, a kdy naopak je velmi pomalá. Označme

$$\tilde{h} = \max_{1\le i\le n} h_{ii}, \ \ h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \ \ i = 1,\dots,n. \tag{4.6}$$

Následující věta popisuje vliv $\tilde{h}$ na limitní chování $B(a,\widehat{\boldsymbol\beta})$ :

**Věta 4.1** *Nechť $\widehat{\boldsymbol\beta}$ je odhad $\boldsymbol\beta$ metodou nejmenších čtverců v modelu (4.2).*

*(i) Jestliže $F$ má exponenciální chvosty, tj.*

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{ba} = 1, \ \ b>0, \ \ pak$$

$$\tilde{h}^{-1/2} \le \underline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) \le \overline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) \le \tilde{h}^{-1}.$$

*(ii) Jestliže $F$ má exponenciální chvosty s exponentem $r$, tj.*

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{ba^r} = 1, \ \ b>0 \ \ a \ \ r\in(1,2],$$

*pak*

$$\tilde{h}^{1-r} \le \underline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) \le \overline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) \le \tilde{h}^{-r}.$$

*(iii) Jestliže $F$ je normální, pak*

$$\lim_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) = \tilde{h}^{-1}.$$



*(iv) Jestliže $F$ má těžké chvosty, tj.*

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{m\log a} = 1, \ \ m>0,$$

*pak*

$$\lim_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) = 1.$$

Věta 4.1 ukazuje, že velká hodnota maximálního diagonálního prvku $\tilde{h}$ matice $\widehat{\mathbf{H}}$ způsobuje, že pravděpodobnost $P_{\boldsymbol\beta}(\max_i|\mathbf{x}_i'(\widehat{\boldsymbol\beta}-\boldsymbol\beta|>a)$ klesá k 0 s rostoucím $a$ pomalu, i při normálním rozdělení chyb a při velkém počtu pozorování $n$. Zároveň vidíme, že při normálním rozdělení chyb vždy platí

$$\overline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) \le \frac{n}{p}, \tag{4.7}$$

přičemž rovnost nastává při vyrovnaném designu odpovídajícím $h_{ii} = \frac{p}{n}, \ i = 1,\dots,n.$

**Důkaz věty 4.1.** Bez újmy obecnosti předpokládejme, že $\tilde{h} = h_{11}$. Protože $0 < \tilde{h} \le 1$ a $\hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol\beta} = \mathbf{h}_i'\mathbf{Y}$, můžeme psát

$$\begin{aligned}
&P_{\boldsymbol\beta}(\max_i\left|\mathbf{x}_i'(\widehat{\boldsymbol\beta}-\boldsymbol\beta)\right|>a)\\
&= P_{\mathbf{0}}(\max_i|\mathbf{h}_i'\mathbf{Y}|>a) \ge P_{\mathbf{0}}(\mathbf{h}_1'\mathbf{Y}>a)\\
&\ge P_{\mathbf{0}}(\tilde{h}Y_1>a, h_{12}Y_2\ge 0,\dots,h_{1n}Y_n\ge 0)\\
&\ge P_{\mathbf{0}}(Y_1>a/\tilde{h})\left(\tfrac{1}{2}\right)^{n-1} = (1-F(a/\tilde{h}))\left(\tfrac{1}{2}\right)^{n-1}.
\end{aligned}$$

Odtud vyplývá

$$\overline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol\beta}) \le \overline{\lim}_{a\to\infty}\frac{-\log(1-F(a/\tilde{h}))}{-\log(1-F(a))}. \tag{4.8}$$

Jestliže $F$ má exponenciální chvosty s indexem $r$, pak ze (4.8) dále plyne

$$\overline{\lim}_{a\to\infty} B(a, \widehat{\boldsymbol{\beta}}) \le \overline{\lim}_{a\to\infty} \frac{b(a/\tilde{h})^r}{ba^r} = \tilde{h}^{-r}, \tag{4.9}$$

což dává horní hranici v (i) a (ii). Pro $F$ s těžkými chvosty ze (4.8) plyne

$$\overline{\lim}_{a\to\infty} B(a, \widehat{\boldsymbol{\beta}}) \le \overline{\lim}_{a\to\infty} \frac{m \log(a/\tilde{h})}{m \log a} = 1 \tag{4.10}$$

a odtud plyne (iv), protože $\widehat{\boldsymbol{\beta}}$ má alespoň jeden kladný a alespoň jeden záporný residuál, a tedy $\underline{\lim}_{a\to\infty} B(a, \widehat{\boldsymbol{\beta}}) \ge 1$.

Na druhé straně, jestliže $F$ má exponenciální chvosty s exponentem $r$, $1 < r \le 2$, pak s užitím Markovovy nerovnosti můžeme psát pro libovolné $\varepsilon \in (0,1)$

$$P_{\boldsymbol{\beta}}(\max_i |\mathbf{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})| > a) \tag{4.11}$$

$$\le \frac{\mathbb{E}_{\mathbf{0}}[\exp\{(1-\varepsilon)b\tilde{h}^{1-r}(\max_i |\hat{Y}_i|^r)\}]}{\exp\{(1-\varepsilon)b\tilde{h}^{1-r}a^r\}},$$

a tedy pokud můžeme ověřit, že

$$\mathbb{E}_{\mathbf{0}}[\exp\{(1-\varepsilon)b\tilde{h}^{1-r}(\max_i |\hat{Y}_i|)^r\}] \le C_r < \infty, \tag{4.12}$$

pak bude platit

$$-\log P_{\mathbf{0}}(\max_i |\hat{Y}_i| > a) \ge -\log C_r + (1-\varepsilon)b\tilde{h}^{1-r}a^r,$$

a odtud dostaneme dolní hranici ve (ii) a vlastně také dolní hranici pro normální rozdělení ve (iii). Musíme tedy dokázat konečnost střední hodnoty ve (4.12). Označme $\|\mathbf{x}\|_s = \left(\sum_{i=1}^n |x_i|^s\right)^{1/s}$, $s > 0$



a položme $s = \frac{r}{r-1}$ ($> 2$). Pak (s přihlédnutím ke vztahu $\sum_{k=1}^n h_{ik}^2 = h_{ii}$)

$$(\max_i |\hat{Y}_i|)^r = \max_i |\mathbf{h}_i'\mathbf{Y}|^r \le \max_i (\|\mathbf{h}_i\|_s \|\mathbf{Y}\|_r)^r$$

$$\le \max_i (\sum_{k=1}^n h_{ik}^2)^{r/s} \sum_{k=1}^n |Y_k|^r \le \tilde{h}^{r-1} \sum_{k=1}^n |Y_k|^r,$$

a tedy

$$\mathbb{E}_{\mathbf{0}} \exp\{(1-\varepsilon)b\tilde{h}^{1-r}(\max_i |\hat{Y}_i|^r\}$$

$$\le \mathbb{E}_{\mathbf{0}} \exp\{(1-\varepsilon)b \sum_{k=1}^n |Y_k|^r\}$$

$$\le (\mathbb{E}_{\mathbf{0}} \exp\{(1-\varepsilon)b|Y_1|^r\})^n.$$

Má-li $F$ exponenciální chvosty s exponentem $r$, pak existuje $K > 0$ takové, že pro $x > K$ platí $1 - F(x) \le \exp\{-(1-\frac{\varepsilon}{2}bx^r\} = C_K$ a integrací per partes dostaneme

$$0 < \mathbb{E}_{\mathbf{0}}[\exp\{(1-\varepsilon)b|Y_1|^r\}]$$

$$= -2\int_0^\infty \exp\{(1-\varepsilon)by^r\}d(1-F(y)) \tag{4.13}$$

$$\le 2\int_0^K \exp\{(1-\varepsilon)by^r\}dF(y)$$

$$+2\exp\{(1-\varepsilon)bK^r\}(1-F(K))$$

$$+2\int_K^\infty r(1-\varepsilon)by^{r-1}(1-F(y))\exp\{(1-\varepsilon)by^r\}dy$$

$$\leq 2\int_0^K \exp\{(1-\varepsilon)by^r\}dF(y)$$
$$+2(1-F(K))\exp\{(1-\varepsilon)bK^2\}$$
$$+2\int_K^\infty r(1-\varepsilon)by^{r-1}\exp\{-\tfrac{\varepsilon}{2}by^r\}dy \leq C_\varepsilon < \infty$$

a tím jsme dokázali (4.12) pro $1 < r \leq 2$. Pro $r = 1$ postupujeme takto: nejprve si uvědomíme, že ze (4.4) vyplývá $|h_{ij}| \leq \sqrt{h_{ii}},\ i,j = 1,\ldots,n$, a tedy

$$\max_i |\hat{Y}_i| = \max_i |\mathbf{h}_i'\mathbf{Y}| = \max_i |\sum_{j=1}^n h_{ij}Y_j|$$
$$\leq \max_{ij} |h_{ij}| \sum_{j=1}^n |Y_j| \leq \tilde{h}^{1/2} \sum_{j=1}^n |Y_j|.$$

Z Markovovy nerovnosti vyplývá

$$P_{\mathbf{0}}(\max_i |\hat{Y}_i| > a) \leq \frac{\mathbb{E}_{\mathbf{0}} \exp\{(1-\varepsilon)b\tilde{h}^{-1/2}\max_i |\hat{Y}_i|\}}{\exp\{(1-\varepsilon)b\tilde{h}^{-1/2}a\}}$$
$$\leq \frac{(\mathbb{E}_{\mathbf{0}} \exp\{(1-\varepsilon)b|Y_1|\})^n}{\exp\{(1-\varepsilon)b\tilde{h}^{-1/2}a\}}$$

a ze (4.13) vyplývá, že $\mathbb{E}_{\mathbf{0}} \exp\{(1-\varepsilon)b|Y_1|\} < \infty$; odtud dostaneme dolní hranici v (i).

Jestliže $F$ je distribuční funkce normálního rozdělení $\mathcal{N}(0,\sigma^2)$, pak $\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}$ má $n$-rozměrné normální rozdělení $\mathcal{N}_n\left(\mathbf{0},\sigma^2\widehat{\mathbf{H}}\right)$, a tedy

$$P_{\mathbf{0}}(\max_i |\hat{Y}_i| > a) \geq P_{\mathbf{0}}(\mathbf{h}_1'\mathbf{Y} > a) = 1 - \Phi(a\sigma^{-1}\tilde{h}^{-1/2})$$



a $\overline{\lim}_{a\to\infty} B(a,\widehat{\boldsymbol{\beta}}) \leq \tilde{h}^{-1}$. ■

## 4.2 $M$-odhady

$M$-odhad parametru $\boldsymbol{\beta}$ v modelu (4.1) je definován jako řešení $\mathbf{M}_n$ minimalizace

$$\sum_{i=1}^n \rho(Y_i - \mathbf{x}_i'\mathbf{t}) := \min \tag{4.14}$$

vzhledem k $\mathbf{t} \in \mathbb{R}_p$, kde $\rho : \mathbb{R}_1 \mapsto \mathbb{R}_1$ je absolutně spojitá, obvykle konvexní funkce s derivací $\psi$. Zřejmě $\mathbf{M}_n$ je *ekvivariantní vzhledem k regresi*, tj.

$$\mathbf{M}_n(\mathbf{Y} + \mathbf{X}\mathbf{b}) = \mathbf{M}_n(\mathbf{Y}) + \mathbf{b}\ \forall \mathbf{b} \in \mathbb{R}_p, \tag{4.15}$$

ale $\mathbf{M}_n$ obecně není *ekvivariantní vzhledem k měřítku*: obecně neplatí

$$\mathbf{M}_n(c\mathbf{Y}) = c\mathbf{M}_n(\mathbf{Y}) \quad \text{pro} \quad c > 0. \tag{4.16}$$

$M$-odhad, ekvivariantní vzhledem k měřítku, získáme buď *studentizací* nebo tak, že zároveň s regresním parametrem odhadujeme měřítko. Studentizovaný $M$-odhad je řešením minimalizace

$$\sum_{i=1}^n \rho\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{S_n}\right) := \min, \tag{4.17}$$

kde $S_n = S_n(\mathbf{Y}) \geq 0$ je vhodná škálová statistika. Aby $\mathbf{M}_n$ bylo ekvivariantní vzhledem k regresi i k měřítku, je třeba, aby škálová statistika $S_n$ byla invariantní vzhledem k regresi a ekvivariantní vzhledem k měřítku, tj.

$$S_n(c(\mathbf{Y} + \mathbf{X}\mathbf{b})) = cS_n(\mathbf{Y})\ \forall \mathbf{b} \in \mathbb{R}_p \ \text{ a } \ c > 0. \tag{4.18}$$

Takovou statistikou je např. odmocnina z residuálního součtu čtverců,

$$S_n(\mathbf{Y}) = [(\hat{\mathbf{Y}} - \mathbf{Y})'(\hat{\mathbf{Y}} - \mathbf{Y})]^{1/2} = [\mathbf{Y}'(\mathbf{I}_n - \widehat{\mathbf{H}})\mathbf{Y}]^{1/2},$$

ale ta je úzce spojena s odhadem metodou nejmenších čtverců a tedy nerobustní. Robustní škálové statistiky mohou být založeny na regresních kvantilech nebo regresních pořadových skórech, o kterých se zmíníme později.

Minimalizace (4.17) musí být doplněna pravidlem, jak definovat $\mathbf{M}_n$ v případě, že $S_n(\mathbf{Y}) = 0$; ve většině případů však toto nastane s pravděpodobností 0 a speciální tvar pravidla nemá vliv na asymptotické chování $\mathbf{M}_n$.

Jestliže $\psi(x) = \frac{d\rho(x)}{dx}$ je spojitá funkce, pak $\mathbf{M}_n$ je kořenem soustavy rovnic

$$\sum_{i=1}^{n} \mathbf{x}_i \psi\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{S_n}\right) = \mathbf{0}. \tag{4.19}$$

Tato soustava rovnic však může mít více kořenů a pouze jeden z nich vede ke globálnímu minimu úlohy (4.17). V knize [46] je dokázáno, že za obecných podmínek vždy existuje alespoň jeden kořen (4.19), který je $\sqrt{n}$-konsistentním odhadem $\boldsymbol{\beta}$. Jestliže $\psi$ je neklesající schodovitá funkce, a tedy $\rho$ je konvexní, po částech lineární funkce, pak $\mathbf{M}_n$ je bodem minima konvexní funkce $\sum_{i=1}^{n} \rho\left((Y_i - \mathbf{x}_i'\mathbf{t})/S_n\right)$ přes $\mathbf{t} \in \mathbb{R}_p$, a i v tomto případě můžeme dokázat jeho konsistenci a asymptotickou normalitu.

Měřítko zároveň s regresním parametrem můžeme odhadovat různými způsoby: např. $(\mathbf{M}_n, \hat{\sigma})$ je řešením minimalizace

$$\sum_{i=1}^{n} \sigma\rho\left(\sigma^{-1}(Y_i - \mathbf{x}_i'\mathbf{t}\right) + a\sigma := \min, \ \mathbf{t} \in \mathbb{R}_p, \ \sigma > 0, \tag{4.20}$$



kde $a > 0$ je vhodná konstanta. Tato minimalizace vede k soustavě $p + 1$ rovnic

$$\sum_{i=1}^{n} \mathbf{x}_i \psi\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{\sigma}\right) = \mathbf{0}$$

$$\sum_{i=1}^{n} \chi\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{\sigma}\right) = a, \tag{4.21}$$

$$\text{kde} \qquad \chi(x) = x\psi(x) - \rho(x) \ \text{ a } \ a = \int_{\mathbb{R}} \chi(x) d\Phi(x)$$

a $\Phi$ je distribuční funkce $\mathcal{N}(0,1)$. Za funkci $\psi$ se obvykle volí Huberova funkce (3.15).

Matice $\mathbf{X}$ může být náhodná, nenáhodná i smíšená, tj. některé prvky $\mathbf{X}$ jsou pevné a jiné náhodné. Při náhodné matici $\mathbf{X}$ je třeba vzít v úvahu i možné rozdělení pravděpodobností řádků $\mathbf{X}$ a influenční funkce závisí na dvou argumentech, $\mathbf{x}$ a $y$. Podobně i bod selhání odhadu je třeba uvažovat nejen vzhledem k možným změnám pozorování $y$, ale i pozorování $\mathbf{x}$.

Asymptotické vlastnosti $M$-odhadů s pevnou maticí $\mathbf{X}$ jsou podrobně studovány v knize [46]. Pro ilustraci uvedeme asymptotické rozdělení pravděpodobností $M$-odhadu v nejjednodušším případě, tj. nestudentizovaného $M$-odhadu s nenáhodnou maticí $\mathbf{X}$.

### 4.2.1 Asymptotické rozdělení $M$-odhadu s nenáhodnou maticí

Předpokládejme, že distribuční funkce $F$ chyb $U_i$ v modelu (4.1) je symetrická podle nuly. Uvažujme $M$-odhad $\mathbf{M}_n$ jakožto řešení minimalizace (4.14), kde $\psi = \rho'$ je lichá, absolutně spojitá a předpokládejme, že $\mathbb{E}_F \psi^2(U_1) < \infty$. O matici $\mathbf{X} = \mathbf{X}_n$

předpokládejme, že má hodnost $p$ a že $\max_{1\leq i\leq n} h_{ii}^{(n)} \to 0$ při $n\to\infty$, kde $h_{ii}^{(n)}$ je maximální diagonální element projekční matice $\widehat{\mathbf{H}}_n = \mathbf{X}_n(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'$. Pak při $n\to\infty$ platí

$$\mathbf{M}_n \xrightarrow{p} \boldsymbol{\beta} \tag{4.22}$$

$$\mathcal{L}\left\{(\mathbf{X}_n'\mathbf{X}_n)^{1/2}(\mathbf{M}_n-\boldsymbol{\beta})\right\} \to \mathcal{N}_p\left(\mathbf{0}, \sigma^2(\psi,F)\mathbf{I}_p)\right),$$

$$\text{kde}\quad \sigma^2(\psi,F) = \frac{\mathbb{E}_F\psi^2(U_1)}{(\mathbb{E}_F\psi'(U_1))^2}.$$

Jestliže za stejných předpokladů $\frac{1}{n}\mathbf{X}_n'\mathbf{X}_n \to \mathbf{Q}$, kde $\mathbf{Q}$ je pozitivně definitní matice řádu $p\times p$, pak

$$\mathcal{L}\left\{\sqrt{n}(\mathbf{M}_n-\boldsymbol{\beta})\right\} \to \mathcal{N}_p\left(\mathbf{0}, \sigma^2(\psi,F)\mathbf{Q}^{-1}\right).$$

Jestliže $\psi$ může mít skoky, ale je neklesající, a $F$ je absolutně spojitá s hustotou $f$, pak (4.22) zůstává v platnosti s tím rozdílem, že

$$\sigma^2(\psi,F) = \frac{\mathbb{E}_F\psi^2(U_1)}{(\int_{\mathbb{R}} f(x)d\psi(x))^2}.$$

Všimněme si, že $\sigma^2(\psi,F)$ je totéž jako ve (3.54) u asymptotického rozdělení $M$-odhadu parametru polohy. Asymptotické rozdělení studentizovaného $M$-odhadu závisí na vlastnostech studentizující statistiky $S_n$.

### 4.2.2 Influenční funkce $M$-odhadu s náhodnou maticí

Uvažujme model (4.1) s náhodnou maticí $\mathbf{X}$, ve kterém $(\mathbf{x}_i', Y_i)'$, $i=1,\ldots,n$ jsou nezávislé náhodné vektory s hodnotami v $\mathbb{R}_p\times\mathbb{R}_1$,



stejně rozdělené s distribucí $P(\mathbf{x},y)$. Jestliže $\rho$ má absolutně spojitou derivaci $\psi$, pak statistický funkcionál $\mathbf{T}(P)$, příslušný odhadu (4.14), je řešením soustavy $p$ rovnic

$$\int_{\mathbb{R}_{p+1}} \mathbf{x}\psi(y-\mathbf{x}'\mathbf{T}(P)dP(\mathbf{x},y) = \mathbf{0}. \tag{4.23}$$

Uvažujme kontaminované rozdělení

$$P_t = (1-t)P + t\delta(\mathbf{x}_0,y_0),\ 0\leq t\leq 1,\ (\mathbf{x}_0,y_0)\in\mathbb{R}_p\times\mathbb{R},$$

kde $\delta(\mathbf{x}_0,y_0)$ je rozdělení pravděpodobností degenerované v bodě $(\mathbf{x}_0,y_0)$. Pak funkcionál $\mathbf{T}(P_t)$ je řešením soustavy rovnic

$$(1-t)\int_{\mathbb{R}_{p+1}} \mathbf{x}\psi(y-\mathbf{x}'\mathbf{T}(P_t))dP(\mathbf{x},y)$$
$$+t\mathbf{x}_0\psi(y_0-\mathbf{x}_0'\mathbf{T}(P_t)) = \mathbf{0}.$$

Derivováním podle $t$ dostaneme

$$-\int_{\mathbb{R}_{p+1}} \mathbf{x}\psi(y-\mathbf{x}'\mathbf{T}(P_t))dP(\mathbf{x},y) + \mathbf{x}_0\psi(y_0-\mathbf{x}_0'\mathbf{T}(P_t))$$
$$-(1-t)\int_{\mathbb{R}_{p+1}} \mathbf{x}'\mathbf{x}\frac{d\mathbf{T}(P_t)}{dt}\psi'(y-\mathbf{x}'\mathbf{T}(P_t))dP(\mathbf{x},y)$$
$$-t\mathbf{x}_0'\mathbf{x}_0\frac{d\mathbf{T}(P_t)}{dt}\psi'(y_0-\mathbf{x}_0'\mathbf{T}(P_t)) = \mathbf{0}.$$

Influenční funkci $\mathbf{IF}(\mathbf{x}_0,y_0;\mathbf{T},P) = \left.\frac{d\mathbf{T}(P_t)}{dt}\right|_{t=0}$ dostaneme, položíme-li $t=0$ a uvědomíme si, že vzhledem ke (4.23) je $\int_{\mathbb{R}_{p+1}} \mathbf{x}\psi(y-$

$\mathbf{x}'\mathbf{T}(P_t))dP(\mathbf{x}, y) = \mathbf{0}$ :

$$\mathbf{IF}(\mathbf{x}_0, y_0; \mathbf{T}, P) \int_{\mathbb{R}_{p+1}} \mathbf{x}'\mathbf{x}\psi'(y - \mathbf{x}'\mathbf{T}(P))dP(\mathbf{x}, y)$$
$$= \mathbf{x}_0\psi(y_0 - \mathbf{x}_0'\mathbf{T}(P)),$$

a tedy influenční funkce $M$-odhadu má tvar

$$\mathbf{IF}(\mathbf{x}_0, y_0; \mathbf{T}, P) = \mathbf{B}^{-1}\mathbf{x}_0\psi(y_0 - \mathbf{x}_0'\mathbf{T}(P)), \tag{4.24}$$

kde

$$\mathbf{B} = \int_{\mathbb{R}_{p+1}} \mathbf{x}'\mathbf{x}\psi'(y - \mathbf{x}'\mathbf{T}(P))dP(\mathbf{x}, y). \tag{4.25}$$

Vidíme, že volbou $\psi$ lze dosáhnout toho, aby influenční funkce (4.24) byla ohraničená vzhledem k $y_0$; influenční funkce $M$-odhadu je však neohraničená vzhledem k $\mathbf{x}_0$, a tedy $M$-odhad je nerobustní vzhledem k $\mathbf{X}$. To vedlo řadu autorů k zavedení zobecněných $M$-odhadů, tzv. $GM$-odhadů, které vhodnými vahami vyrovnávají vliv odlehlých hodnot $\mathbf{x}$.

#### Asymptotické vlastnosti $M$-odhadu s náhodnou maticí

Jestliže soustava rovnic

$$E_P[\mathbf{x}\psi(y - \mathbf{x}'\mathbf{t}) = \mathbf{0}$$

má jediné řešení $\mathbf{T}(P) = \boldsymbol{\beta}$, pak

$$\mathbf{T}(P_n) \to \mathbf{T}(P)$$



při $n \to \infty$, kde $P_n$ je empirické rozdělení příslušné pozorováním $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$. Za určitých podmínek na rozdělení pravděpodobností $P$ platí asymptotická reprezentace

$$\mathbf{T}(P_n) = \mathbf{T}(P) + \tfrac{1}{n}\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P) + \mathbf{o}_p(n^{-1/2}).$$

Jestliže $E_P\|\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P)\|^2 < \infty$, dostaneme odtud asymptotické rozdělení pravděpodobností $\mathbf{T}(P_n)$ :

$$\mathcal{L}\left\{\sqrt{n}(\mathbf{T}(P_n) - \mathbf{T}(P))\right\} \to \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \tag{4.26}$$

kde

$$\boldsymbol{\Sigma} = E_P[\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P)]'[\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P)] = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1},$$

$\mathbf{B}$ je matice definovaná ve (4.25) a

$$\mathbf{A} = \int_{\mathbb{R}_{p+1}} \mathbf{x}'\mathbf{x}\psi^2(y - \mathbf{x}'\mathbf{T}(P))dP(\mathbf{x}, y).$$

### 4.2.3 $GM$-odhady

Influenční funkce (4.24) $M$-odhadu je neohraničená vzhledem k $\mathbf{x}$, a tedy $M$-odhad je citlivý k případným vlivným bodům v matici $\mathbf{X}$. Tuto skutečnost nemůžeme ovlivnit volbou funkce $\psi$. Řada autorů navrhla doplnit definici $M$-odhadu vhodnými vahami $w$, které redukují vliv velkých hodnot $x_{ij}$.

Mallows [54], [55] navrhl zobecněný $M$-odhad jako řešení minimalizace

$$\sum_{i=1}^{n} \sigma w(\mathbf{x}_i)\rho\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{\sigma}\right) := \min, \ \mathbf{t} \in \mathbb{R}_p, \sigma > 0. \tag{4.27}$$

Jestliže $\psi = \rho'$ je spojitá, je zobecněný $M$-odhad kořenem rovnice

$$\sum_{i=1}^{n} \mathbf{x}_i w(\mathbf{x}_i) \psi \left( \frac{Y_i - \mathbf{x}_i' \mathbf{t}}{\sigma} \right) = \mathbf{0} \tag{4.28}$$

a influenční funkce příslušného funkcionálu $\mathbf{T}(P)$ je rovna

$$\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P) = \mathbf{B}^{-1} \mathbf{x} w(\mathbf{x}) \psi \left( \frac{y - \mathbf{x}' \mathbf{T}(P)}{S(P)} \right), \tag{4.29}$$

kde $S(P)$ je funkcionál, příslušný řešení $\sigma$ v minimalizaci (4.27). Ohraničené influenční funkce dosáhneme volbou $w$, při které je $\mathbf{x} w(\mathbf{x})$ ohraničené.

Takto definovaný odhad je speciálním případem následujícího $GM$-odhadu, který je řešením soustavy rovnic

$$\begin{aligned} \sum_{i=1}^{n} \eta \left( \mathbf{x}_i, \frac{Y_i - \mathbf{x}_i' \mathbf{t}}{\sigma} \right) &= \mathbf{0} \\ \sum_{i=1}^{n} \chi \left( \frac{Y_i - \mathbf{x}_i' \mathbf{t}}{\sigma} \right) &= \mathbf{0}, \end{aligned} \tag{4.30}$$

kde $\eta, \chi$ jsou funkce, $\eta : \mathbb{R}_p \times \mathbb{R} \mapsto \mathbb{R}$ a $\chi : \mathbb{R} \mapsto \mathbb{R}$.

Odhadu metodou nejmenších čtverců odpovídá volba

$$\eta(\mathbf{x}, u) = u \quad \text{a} \quad \chi(u) = u^2 - 1,$$

$M$-odhadu volba $\eta(\mathbf{x}, u) = \psi(u)$ a Mallowsově $GM$-odhadu odpovídá volba

$$\eta(\mathbf{x}, u) = w(\mathbf{x}) \psi(u).$$



Obvyklá volba funkce $\eta$ je $\eta(\mathbf{x}, u) = \frac{\psi_1(\mathbf{x})}{\|\mathbf{x}\|} \psi(u)$, kde $\psi$ je např. Huberova funkce. Funkce $\chi$ se obvykle volí stejně jako ve (4.21).

Statistické funkcionály $\mathbf{T}(P)$ a $S(P)$ odpovídající $\mathbf{M}_n$ a $\sigma_n$ jsou definovány implicitně jako řešení soustavy rovnic:

$$\begin{aligned} \int_{\mathbb{R}_{p+1}} \mathbf{x} \, \eta \left( \mathbf{x}, \frac{y - \mathbf{x}' \mathbf{T}(P)}{S(P)} \right) dP(\mathbf{x}, y) &= \mathbf{0} \\ \int_{\mathbb{R}_{p+1}} \chi \left( \mathbf{x}, \frac{y - \mathbf{x}' \mathbf{T}(P)}{S(P)} \right) dP(\mathbf{x}, y) &= \mathbf{0}. \end{aligned} \tag{4.31}$$

Influenční funkce funkcionálu $\mathbf{T}(P)$ ve speciálním případě $\sigma = 1$ má tvar

$$\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P) = \mathbf{B}^{-1} \mathbf{x} \eta(\mathbf{x}, y - \mathbf{x}' \mathbf{T}(P)),$$

kde

$$\mathbf{B} = \int_{\mathbb{R}_{p+1}} \mathbf{x}' \mathbf{x} \left[ \frac{\partial}{\partial u} \eta(\mathbf{x}, u) \right]_{u = y - \mathbf{x}' \mathbf{T}(P)} dP(\mathbf{x}, y).$$

Asymptotické vlastnosti $GM$-odhadů studovali Maronna a Yohai [56]. Za určitých podmínek jsou $GM$-odhady silně konsistentní a $\sqrt{n}(\mathbf{T}(P_n) - \mathbf{T}(P))$ má asymptoticky $p$-rozměrné normální rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ s kovarianční maticí $\mathbf{\Sigma} = \mathbf{B}^{-1} \mathbf{A} \mathbf{B}^{-1}$, kde

$$\mathbf{A} = \int_{\mathbb{R}_{p+1}} \mathbf{x}' \mathbf{x} \eta^2(\mathbf{x}, y - \mathbf{x}' \mathbf{T}(P)) dP(\mathbf{x}, y).$$

Krasker a Welsch [51] navrhli $GM$-odhad jako řešení soustavy rovnic

$$\sum_{i=1}^{n} \mathbf{x}_i w_i \frac{Y_i - {}' \mathbf{t}}{\sigma} = \mathbf{0}$$

s vahami $w_i = w(\mathbf{x}_i, Y_i, \mathbf{t}) > 0$, které jsou určeny tak, aby maximalizovaly asymptotickou vydatnost odhadu (vzhledem k asymptotické kovarianční matici $\mathbf{\Sigma}$) za omezení $\gamma^* \leq a < \infty$, kde $\gamma^*$ je globální citlivost funkcionálu $\mathbf{T}$ vzhledem k rozdělení $P$, tj.

$$\gamma^* = \sup_{\mathbf{x},y} \left[ (\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P))' \, \mathbf{\Sigma}^{-1} \, (\mathbf{IF}(\mathbf{x}, y; \mathbf{T}, P)) \right]^{1/2}.$$

Řešením jsou váhy ve tvaru

$$w(\mathbf{x}, y, \mathbf{t}) = \min \left\{ 1, \frac{a}{\left| \frac{y - \mathbf{x}'\mathbf{t}}{\sigma} \right| (\mathbf{x}'\mathbf{A}\mathbf{x})^{1/2}} \right\},$$

kde

$$\mathbf{A} = \int_{\mathbb{R}_{p+1}} \mathbf{x}'\mathbf{x} \left( \frac{y - \mathbf{x}'\mathbf{t}}{\sigma} \right) w^2(\mathbf{x}, y, \mathbf{t}) dP(\mathbf{x}, y).$$

Krasker-Welschův odhad má ohraničenou influenční funkci, ale je třeba ho počítat iteračně, protože matice $\mathbf{A}$ závisí na $w$.

## 4.3 $L$-odhady

$L$-odhady parametru polohy ve tvaru lineárních kombinací pořádkových statistik nebo funkcí pořádkových statistik jsou velmi atraktivní, protože jsou definovány explicitně a snadno se vypočítají. Proto se přirozeně statistikové snažili rozšířit $L$-odhady na lineární regresní model. Toto rozšíření však není snadné, protože neexistovalo žádné přirozené rozšíření empirického (výběrového) kvantilu na regresní model. To se podařilo až Koenkerovi a Bassettovi [50], kteří v r. 1978 definovali regresní $\alpha$-kvantil $\widehat{\boldsymbol{\beta}}(\alpha)$ pro



model (4.1) za předpokladu, že $\beta_1$ je absolutní člen, tj. že matice $\mathbf{X}$ vyhovuje podmínce

$$x_{i1} = 1, \ i = 1, \ldots, n. \tag{4.32}$$

Regresní $\alpha$-kvantil $\widehat{\boldsymbol{\beta}}(\alpha)$, $0 < \alpha < 1$, je definován jako řešení minimalizace

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i'\mathbf{t}) := \min, \ \mathbf{t} \in \mathbb{R}_p, \tag{4.33}$$

kde

$$\rho_\alpha(x) = |x| \{ \alpha I[x > 0] + (1 - \alpha) I[x < 0] \}, \ x \in \mathbb{R}. \tag{4.34}$$

Protože $\rho_\alpha(x)$ je konvexní, po částech lineární funkce $x$, je nasnadě myšlenka řešit minimalizaci (4.33) upravenou simplexovou metodou. Skutečně, Koenker a Bassett navrhli počítat $\widehat{\boldsymbol{\beta}}(\alpha)$ jako složku $\boldsymbol{\beta}$ optimálního řešení $(\boldsymbol{\beta}, \mathbf{r}^+, \mathbf{r}^-)$ úlohy parametrického lineárního programování

$$\begin{aligned} & \alpha \sum_{i=1}^{n} r_i^+ + (1 - \alpha) \sum_{i=1}^{n} r_i^- : \min \\ & \text{za podmínky} \\ & \sum_{j=1}^{p} x_{ij}\beta_j + r_i^+ - r_i^- = Y_i, \ i = 1, \ldots, n; \\ & \beta_j \in \mathbb{R}_1, \ j = 1, \ldots, p, \ r_i^+, r_i^- \geq 0, \ i = 1, \ldots, n, \\ & 0 < \alpha < 1. \end{aligned} \tag{4.35}$$

Proměnné $r_i^+$ a $r_i^-$ v (4.35) jsou rovny kladné a záporné části residuí $Y_i - \mathbf{x}_i'\boldsymbol{\beta}, \ i = 1, \ldots, n$.

Úloha (4.35) nám nejen umožňuje vypočítat regresní kvantily simplexovou metodou, ale zároveň vypovídá o struktuře regresních kvantilů. Z teorie lineárního programování víme, že množina $B(\alpha)$ řešení (4.35) (a tedy i (4.33)) je neprázdná, kompaktní a polyedrální. Pokud není dáno jiné omezení, lze volit $\widehat{\boldsymbol{\beta}}(\alpha)$ jako lexikograficky maximální element $B(\alpha)$. Jakožto funkce argumentu $\alpha \in (0,1)$ je $\widehat{\boldsymbol{\beta}}(\alpha)$ schodovitou funkcí $\alpha$.

Asymptotické vlastnosti $\widehat{\boldsymbol{\beta}}(\alpha)$ jsou analogické vlastnostem výběrových kvantilů v modelu s parametrem posunutí. Populačním partnerem (statistickým funkcionálem) příslušným $\widehat{\boldsymbol{\beta}}(\alpha)$ je *populační regresní kvantil*

$$\boldsymbol{\beta}(\alpha) = (\beta_1 + F^{-1}(\alpha), \beta_2, \ldots, \beta_p)' \tag{4.36}$$

a jestliže distribuční funkce $F$ chyb v modelu (4.1) je symetrická a ryze rostoucí v okolí $F^{-1}(\alpha)$ s derivací $f$ a matice $\mathbf{X}_n$ je buď pevná a $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}_n'\mathbf{X}_n = \mathbf{Q}$ nebo je náhodná (až na první sloupec) a $\lim_{n\to\infty} \mathbb{E}\mathbf{x}_1'\mathbf{x}_1 = \mathbf{Q}$, kde $\mathbf{Q}$ je pozitivně definitní matice řádu $p \times p$, pak $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n(\alpha) - \boldsymbol{\beta}(\alpha))$ má asymptoticky $p$-rozměrné normální rozdělení

$$\mathcal{N}_p\left(\mathbf{0}, \frac{\alpha(1-\alpha)}{(f(F^{-1}(\alpha))^2}\mathbf{Q}^{-1}\right), \tag{4.37}$$

což je ve shodě s asymptotickým rozdělením výběrového $\alpha$-kvantilu odpovídajícího matici $\mathbf{X} = \mathbf{1}_n = (1,\ldots,1)' \in \mathbb{R}_n$.

Máme-li k dispozici regresní kvantily, můžeme definovat řadu $L$-odhadů parametru $\boldsymbol{\beta}$ v lineárním regresním modelu. Nejznámější je $L_1$-odhad, neboli regresní medián, což je regresní $\alpha$-kvantil s $\alpha = 1/2$. Dále můžeme uvažovat $L$-odhady, které jsou rovny lineární kombinaci konečně mnoha regresních kvantilů. Z hlediska praktického použití je nejzajímavější *useknutý odhad metodou nejmenších čtverců,* který navrhli Koenker a Bassett [50], a který je



rozšířením useknutého průměru na lineární regresní model: Zvolme $\alpha_1, \alpha_2,\ 0 < \alpha_1 < \alpha_2 < 1$ a položme

$$a_i = I\left[\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_n(\alpha_1) < Y_i < \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_n(\alpha_2)\right], \tag{4.38}$$

a vypočtěme vážený odhad metodou nejmenších čtverců s vahami $a_i,\ i = 1,\ldots,n$. Tento odhad $\mathbf{T}_n(\alpha_1,\alpha_2)$, který nazveme $(\alpha_1,\alpha_2)$-useknutým odhadem metodou nejmenších čtverců, můžeme psát v explicitním tvaru

$$\mathbf{T}_n(\alpha_1,\alpha_2) = (\mathbf{X}_n'\mathbf{A}_n\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{A}_n\mathbf{Y}_n, \tag{4.39}$$

kde $\mathbf{A}_n = \mathrm{diag}(a_i)$ je diagonální matice s diagonálou $(a_1,\ldots,a_n)$.

Za určitých podmínek regularity kladených na matici $\mathbf{X}_n$ a distribuční funkci $F$ (která má být rostoucí a diferencovatelná v intervalu $(F^{-1}(\alpha_1)-\varepsilon, F^{-1}(\alpha_2)+\varepsilon)$) lze ukázat, že $\mathbf{T}_n(\alpha_1,\alpha_2)$ má asymptoticky normální rozdělení; přesněji řečeno,

$$\mathcal{L}\left\{\sqrt{n}(\mathbf{T}_n - \boldsymbol{\beta} - \delta\mathbf{e}_1)\right\} \to \mathcal{N}_p\left(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}\right), \tag{4.40}$$

kde $\mathbf{e}_1 = (1,0,\ldots,0)' \in \mathbb{R}_p$ a

$$\begin{aligned}
\delta &= (\alpha_2-\alpha_1)^{-1}\int_{\alpha_1}^{\alpha_2} F^{-1}(u)du,\\
\sigma^2 &= \sigma^2(\alpha_1,\alpha_2,F) \qquad\qquad (4.41)\\
&= (\alpha_2-\alpha_1)^{-1}\Big\{\int_{\alpha_1}^{\alpha_2} \alpha_2(F^{-1}(u)-\delta)^2du\\
&+\alpha_1(F^{-1}(\alpha_1)-\delta)^2 + (1-\alpha_2)(F^{-1}(\alpha_2)-\delta)^2\\
&-\left[\alpha_1(F^{-1}(\alpha_1)-\delta) + (1-\alpha_2)(F^{-1}(\alpha_2)-\delta)\right]^2\Big\}.
\end{aligned}$$

V symetrické situaci, kdy $F(x) + F(-x) = 1$, $x \in \mathbb{R}$ a $\alpha_1 = \alpha$, $\alpha_2 = 1-\alpha$, $0 < \alpha < \frac{1}{2}$, je $\delta = 0$ a $\sqrt{n}(\mathbf{T}_n(\alpha) - \boldsymbol{\beta})$ má asymptoticky normální rozdělení $\mathcal{N}_p(\mathbf{0}, \sigma^2(\alpha, F)\mathbf{Q}^{-1})$, kde

$$\sigma^2(\alpha, F) = \frac{\int_\alpha^{1-\alpha}(F^{-1}(u))^2 du + 2\alpha(F^{-1}(\alpha))^2}{1-2\alpha}. \tag{4.42}$$

Všimněme si, že $\sigma^2(\alpha, F)$ se shoduje s asymptotickým rozptylem $\alpha$-useknutého průměru v modelu s parametrem posunutí.

Vedle useknutého odhadu metodou nejmenších čtverců můžeme uvažovat obecnou třídu $L$-odhadů tvaru

$$\mathbf{T}_n^\nu = \int_0^1 \widehat{\boldsymbol{\beta}}_n(\alpha) d\nu(\alpha), \tag{4.43}$$

kde $\nu$ je vhodná znaménková míra na $(0,1)$ (konečná a s kompaktním nosičem, který je podmnožinou $(0,1)$). Atomická míra $\nu$ vede ke kombinaci konečně mnoha regresních kvantilů. Jiné rozšíření $(\alpha_1, \alpha_2)$-useknutého průměru dostaneme, jestliže $\nu$ je absolutně spojitá vzhledem k Lebesgueově míře s hustotou

$$J(u) = \frac{I[\alpha_1 \le u \le \alpha_2]}{\alpha_2 - \alpha_1}, \; 0 < \alpha_1 < \alpha_2 < 1.$$

Na rozdíl od $M$-odhadů jsou $L$-odhady regresního parametru ekvivariantní nejen vzhledem k regresi, ale též vzhledem k měřítku. $L$-odhady různých typů a jejich vlastnosti lze nalézt např. v [26], [27] a [46].

### 4.3.1 Regresní pořadové skóry

K úloze lineárního programování (4.35) přirozeně existuje duální úloha. Řešení této duální úlohy má velmi zajímavou interpretaci:



zatímco řešení úlohy (4.35) jsou regresní kvantily, řešení duální úlohy, zvaná *regresní pořadové skóry*, mají řadu vlastností podobných vlastnostem pořadí pozorování.

Napišme úlohu duální ke (4.35) ve tvaru

$$\begin{aligned}
&\sum_{i=1}^n Y_i \hat{a}_i := \max \quad \text{za podmínky} \\
&\sum_{i=1}^n \hat{a}_i = n(1-\alpha), \\
&\sum_{i=1}^n x_{ij}\hat{a}_i = (1-\alpha)\sum_{i=1}^n x_{ij}, \; j = 2, \dots, p, \\
&0 \le \hat{a}_i \le 1, \; i = 1, \dots, n, \; 0 < \alpha < 1.
\end{aligned} \tag{4.44}$$

Optimální řešení úlohy (4.44)

$$\hat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \dots, \hat{a}_{nn}(\alpha))', \; 0 \le \alpha \le 1$$

nazveme *regresní pořadové skóry.* Přepišme úlohu (4.44) v maticovém tvaru (a připomeňme si, že podle předpokladu (4.32) je $x_{i1} = 1$, $i = 1, \dots, n$):

$$\begin{aligned}
&\mathbf{Y}_n'\hat{\mathbf{a}} := \max \quad \text{za podmínky} \\
&\mathbf{X}_n'\hat{\mathbf{a}} = (1-\alpha)\mathbf{X}_n'\mathbf{1}_n, \\
&\hat{\mathbf{a}} \in [0,1]^n, \; 0 \le \alpha \le 1.
\end{aligned} \tag{4.45}$$

Z této formy je vidět, že regresní pořadové skóry jsou invariantní vzhledem k parametru $\boldsymbol{\beta}$, tj.

$$\hat{\mathbf{a}}_n(\alpha, \mathbf{Y} + \mathbf{X}\mathbf{b}) = \hat{\mathbf{a}}_n(\alpha, \mathbf{Y}) \; \forall \mathbf{b} \in \mathbb{R}_p. \tag{4.46}$$

Z duality $\widehat{\boldsymbol\beta}(\alpha)$ a $\hat{\mathbf{a}}_n(\alpha)$ vyplývají vztahy

$$\hat{a}_{ni}(\alpha) = \begin{cases} 1 & \dots \quad Y_i > \mathbf{x}_i'\widehat{\boldsymbol\beta}_n(\alpha), \\ 0 & \dots \quad Y_i < \mathbf{x}_i'\widehat{\boldsymbol\beta}_n(\alpha), \quad i = 1, \dots, n \end{cases} \tag{4.47}$$

a pokud $Y_i = \mathbf{x}_i'\widehat{\boldsymbol\beta}_n(\alpha)$, je $0 < \hat{a}_{ni}(\alpha) < 1$; těchto složek je přesně $p$. Jakožto funkce $\alpha$ je $\hat{a}_{ni}(\alpha)$ spojitá, po částech lineární funkce, $\hat{a}_{ni}(0) = 1,\ \hat{a}_{ni}(1) = 0$.

Regresní pořadové skóry mají řadu aplikací. Invariance (4.46) zaručuje, že pokud v nějaké statistické úloze je $\boldsymbol\beta$ rušivým parametrem, zatímco chceme testovat hypotézu o jiném parametru nebo o tvaru rozdělení chyb, testy založené na regresních pořadových skórech jsou vzhledem k $\boldsymbol\beta$ invariantní a tedy $\boldsymbol\beta$ není třeba odhadovat. To nejen usnadňuje výpočet, ale tím se též vyhneme riziku, že bychom $\boldsymbol\beta$ odhadli nevhodným odhadem. Testy lineárních hypotéz s rušivým parametrem $\boldsymbol\beta$, založené na regresních pořadových skórech, jsou zkonstruovány ve [28]. Jako jinou aplikaci zmiňme škálové statistiky, založené na regresních pořadovych skórech, podrobněji popsané v následujícím odstavci; tyto statistiky jsou invariantní vzhledem k regresi a ekvivariantní vzhledem ke změně měřítka, což je žádoucí např. při studentizaci $M$-odhadů. Více o regresních pořadových skórech se dozvíme např. v [19] nebo [46].

## 4.4 Robustní škálové statistiky

Pro studentizaci $M$-odhadů i v mnoha jiných souvislostech potřebujeme škálovou statistiku $S_n(\mathbf{Y})$, která je invariantní vzhledem k regresi a ekvivariantní vzhledem ke změně měřítka, tj. vyhovuje identitě

$$S_n(c(\mathbf{Y} + \mathbf{X}\mathbf{b})) = cS_n(\mathbf{Y})\ \forall \mathbf{b} \in \mathbb{R}_p,\ c > 0,\ \mathbf{Y} \in \mathbb{R}_n \tag{4.48}$$



(viz (4.18)). Takových statistik není v literatuře mnoho a někteří autoři užívají statistiky, které jsou invariantní jen vzhledem k posunutí, ale nikoli k regresi, aniž by si uvědomili, že studentizovaný odhad tím ztrácí svou regresní ekvivarianci. Proto v této části popíšeme některé statistiky tohoto typu; pojmy k tomu potřebné již známe.

(i) *Mediánová absolutní odchylka od regresního mediánu* (MAD).
Statistika MAD je hojně používána v modelu s parametrem posunutí. Na lineární regresní model ji rozšířil Welsh [73]. Nechť $\boldsymbol\beta^0$ je počáteční odhad $\boldsymbol\beta$, který je $\sqrt{n}$-konsistentní a ekvivariantní vzhledem k regresi i k měřítku (tedy nikoli např. obyčejný $M$-odhad). Pak Welshova škálová statistika má tvar

$$S_n = \mathrm{med}_{1\le i\le n}\left|Y_i(\boldsymbol\beta^0) - \xi_{\frac{1}{2}}(\boldsymbol\beta^0)\right|, \tag{4.49}$$

kde

$$Y_i(\boldsymbol\beta^0) = Y_i - \mathbf{x}_i'\boldsymbol\beta^0,\ i = 1, \dots, n$$

$$\xi_{\frac{1}{2}}(\boldsymbol\beta^0) = \mathrm{med}_{1\le i\le n} Y_i(\boldsymbol\beta^0).$$

Tato statistika zřejmě splňuje (4.48). Její asymptotické vlastnosti lze nalézt v [73].

(ii) *L-statistiky založené na regresních kvantilech.*
Eukleidovská vzdálenost dvou regresních kvantilů

$$S_n = \left\|\widehat{\boldsymbol\beta}_n(\alpha_2) - \widehat{\boldsymbol\beta}_n(\alpha_1)\right\|, \tag{4.50}$$

$0 < \alpha_1 < \alpha_2 < 1$, zřejmě vyhovuje (4.48) a $S_n \xrightarrow{p} S(F) = F^{-1}(\alpha_2) - F^{-1}(\alpha_1)$. Další asymptotické vlastnosti $S_n$ plynou např. z vět ve [63]. Eukleidovská norma může být nahrazena $L_p$-normou nebo jinou vhodnou normou. Jiná možnost je uvažovat pouze rozdíl prvních složek regresních kvantilů. tj.

$$S_n = \left|\hat{\beta}_{n1}(\alpha_2) - \hat{\beta}_{n1}(\alpha_1)\right|.$$

Obecněji, Bickel a Lehmann [8] navrhli několik měr rozpětí rozdělení $F$, např.

$$S(F) = \left\{\int_{\frac{1}{2}}^{1} \left[F^{-1}(u) - F^{-1}(1-u)\right]^2 d\Lambda(u)\right\}^{1/2},$$

kde $\Lambda$ je rovnoměrné rozdělení na $(\frac{1}{2}, 1-\delta)$, $0 < \delta < \frac{1}{2}$; to nás vede k zavedení třídy škálových statistik založených na regresních kvantilech typu

$$S_n = \left\{\int_{\frac{1}{2}}^{1} \left\|\widehat{\boldsymbol{\beta}}_n(u) - \widehat{\boldsymbol{\beta}}_n(1-u)\right\|^2 d\Lambda(u)\right\}^{1/2}.$$

(iii) *Odhady* $1/f(F^{-1}(\alpha))$ *založené na regresních kvantilech.* Tyto odhady zobecňují odhady, navržené Falkem [22], na lineární regresní model; jejich asymptotické a další vlastnosti jsou studovány ve [19]. Můžeme uvažovat odhad typu histogram

$$H_n^{(\alpha)} = \frac{\hat{\beta}_{n1}(\alpha+\nu_n) - \hat{\beta}_{n1}(\alpha-\nu_n)}{2\nu_n}, \tag{4.51}$$

kde

$$\nu_n = o\left(n^{-1/2}\right) \text{ a } \lim_{n\to\infty} n\nu_n = \infty.$$



Jiným odhadem $1/f(F^{-1}(\alpha))$ je jádrový odhad s jádrem $k : \mathbb{R}_1 \mapsto \mathbb{R}_1$, které má kompaktní nosič a vyhovuje vztahům

$$\int k(x)dx = 0 \text{ a } \int xk(x)dx = -1.$$

Jádrový odhad má tvar

$$\chi_n^{(\alpha)} = \frac{1}{\nu_n^2}\int_0^1 \hat{\beta}_{n1}(u)k\left(\frac{\alpha-u}{\nu_n}\right)du, \tag{4.52}$$

kde

$$\nu_n \to 0,\ n\nu_n^2 \to \infty,\ n\nu_n^3 \to 0$$

při $n \to \infty$. Oba odhady jsou $\sqrt{n\nu_n}$-konsistentními odhady $1/f(F^{-1}(\alpha))$, vyhovujícími (4.48). Vzhledem k jejich nižšímu řádu konsistence (který plyne z povahy problému a nelze jej za daných podmínek výrazně zlepšit) se nepoužívají ke studentizaci, ale jsou nutné např. při statistické inferenci o kvantilech rozdělení $F$.

(iv) *Škálové statistiky založené na regresních pořadových skórech.* Nechť $(\hat{a}_{n1}(\alpha), \ldots, \hat{a}_{nn}(\alpha)), 0 < \alpha < 1$ jsou regresní pořadové skóry pro model (4.1). Zvolme neklesající skórovou funkci $\varphi : (0,1) \mapsto \mathbb{R}_1$ standardizovanou tak, že $\int_{\alpha_0}^{1-\alpha_0} \varphi^2(\alpha)d\alpha = 1$ pro pevně zvolené $\alpha_0$, $0 < \alpha_0 < \frac{1}{2}$. Vypočtěme skóry

$$\hat{b}_{ni} = -\int_{\alpha_0}^{1-\alpha_0} \varphi(\alpha)d\hat{a}_{ni}(\alpha),\ i = 1, \ldots, n.$$

Škálová statistika

$$S_n = \frac{1}{n}\sum_{i=1}^{n} Y_i\hat{b}_{ni} \tag{4.53}$$

vyhovuje (4.48) a je $\sqrt{n}$-konsistentním odhadem funkcionálu $S(F) = \int_{\alpha_0}^{1-\alpha_0} \varphi(\alpha)F^{-1}(\alpha)d\alpha$.

## 4.5 Jednokrokové verze odhadů

Mnoho odhadů, jako $M$-odhady, regresní kvantily a maximálně věrohodné odhady jsou definovány implicitně jako řešení minimalizace nebo soustavy rovnic. Někdy může být obtížné vyřešit tento problém algebraicky, jindy může existovat více řešení a pouze jedno z nich je vydatné apod. V souvislosti s $M$-odhady jsme se již zmínili, že mezi kořeny soustavy rovnic (4.19) existuje alespoň jeden $\sqrt{n}$-konsistentní odhad, ale nevíme jak rozhodnout, který z kořenů to je.

V mnoha případech lze eficientní kořen soustavy rovnic aproximovat tzv. jednokrokovou verzí, což je vlastně první krok Newton-Raphsonova iteračního algoritmu řešení algebraických rovnic. Ilustrujme tento přístup na jednokrokové verzi $M$-odhadu v lineárním regresním modelu (4.1), vytvořeného funkcí $\rho$, která derivaci $\psi$, a studentizovaného škálovou statistikou $S_n = S_n(\mathbf{Y})$.

Procedura začíná počátečním odhadem $\mathbf{M}_n^{(0)}$ parametru $\boldsymbol{\beta}$,j konsistentním s řádem $\sqrt{n}$. Jednokroková verze $M$-odhadu je dána vztahem

$$\mathbf{M}_n^{(1)} = \begin{cases} \mathbf{M}_n^{(0)} + \frac{1}{\hat{\gamma}_n}\mathbf{W}_n & \ldots \quad \hat{\gamma}_n \neq 0, \\ \mathbf{M}_n^{(0)} & \ldots \quad \hat{\gamma}_n = 0, \end{cases} \tag{4.54}$$

kde

$$\mathbf{W}_n = \mathbf{Q}_n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \psi\left(\frac{Y_i - \mathbf{x}_i \mathbf{M}_n^{(0)}}{S_n}\right),$$



$$\mathbf{Q}_n = \tfrac{1}{n}\mathbf{X}_n'\mathbf{X}_n$$

a $\hat{\gamma}_n$ je odhad funkcionálu $\gamma = 1/S(F)\int_{\mathbb{R}} \psi'(x/S(F))dF(x)$ nebo $\gamma = \int_{\mathbb{R}} f(xS(F))d\psi(x)$ podle toho, zda volená funkce $\psi$ generující $M$-odhad je spojitá nebo nespojitá. Např. u absolutně spojité $\psi$ můžeme použít odhad

$$\hat{\gamma}_n = \frac{1}{nS_n}\sum_{i=1}^{n} \psi'\left(\frac{Y_i - \mathbf{x}_i'\mathbf{M}_n^{(0)}}{S_n}\right).$$

$\mathbf{M}_n^{(1)}$ je dobrou aproximací konsistentního $M$-odhadu $\mathbf{M}_n$: pokud $\psi$ je dostatečně hladká, lze dokázat

$$\|\mathbf{M}_n - \mathbf{M}_n^{(1)}\| = \mathcal{O}_p\left(n^{-1}\right),$$

zatímco za přítomnosti skoků ve funkci $\psi$ platí

$$\|\mathbf{M}_n - \mathbf{M}_n^{(1)}\| = \mathcal{O}_p\left(n^{-3/4}\right).$$

Více o jednokrokových verzích lze nalézt v [7], [45], [48], [68] a pro $k$-krokové verze v modelu posunutí též [44]. Obecně lze říci, že jednokrokové verze dávají dobré aproximace pro $M$-odhady s hladkými funkcemi $\psi$.

## 4.6 Odhady s vysokým bodem selhání

Bod selhání odhadu v lineárním modelu bere v úvahu nejen možná nahrazení pozorování $Y_1, \ldots, Y_n$ libovolnými hodnotami, ale též možná nahrazení vektorů $\mathbf{x}_1, \dots, \mathbf{x}_n$. Přesněji řečeno, naše pozorování tvoří matici

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \dots \\ \mathbf{z}_n' \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1', & y_1 \\ \mathbf{x}_2', & y_2 \\ \dots & \\ \mathbf{x}_n', & y_n \end{bmatrix}$$

a bod selhání odhadu $\mathbf{T}$ parametru $\boldsymbol{\beta}$ je nejmenší celé číslo $m_n(\mathbf{Z})$ takové, že nahradíme-li libovolných $m$ řádků v matici $\mathbf{Z}$ libovolnými jinými řádky a označíme vzniklý odhad $\mathbf{T}_m^*$, pak $\sup \|\mathbf{T} - \mathbf{T}_m^*\| = \infty$, kde supremum bereme přes všechny možné náhrady $m$ řádků. Často měříme bod selhání také limitou $\varepsilon^* = \lim_{n\to\infty} \frac{m_n}{n}$, pokud existuje. Je zřejmé, že i odhady, které dosahovaly bodu selhání 1/2 v modelu s parametrem posunutí, těžko mohou dosahovat 1/2 v regresním modelu, ovlivněny maticí $\mathbf{X}$. V této souvislosti vzniká několik otázek, hlavně zdali vůbec existují odhady s maximálním možným bodem selhání, jaké jsou jejich další vlastnosti a kdy má smysl je použít.

První otázku odpověděl kladně Siegel [66] již v roce 1982, kdy sestrojil tzv. opakovaný medián s 50% bodem selhání, který však není vhodný pro praktické aplikace. Krátce nato Rousseeuw [60] v r. 1984 publikoval *odhad metodou nejmenšího mediánu čtverců* (LMS), který minimalizuje

$$\mathrm{med}_{1\le i\le n}\{[Y_i - \mathbf{x}_i'\mathbf{t}]^2\}, \ \mathbf{t} \in \mathbb{R}_p. \tag{4.55}$$

Tento odhad je sice konsistentním odhadem $\boldsymbol{\beta}$, ale s řádem konsistence $n^{1/3}$ je velmi málo vydatný. Rousseeuw též navrhl odhad metodou useknutých čtverců, který je řešením minimalizace

$$\sum_{i=1}^{h_n} (Y_i - \mathbf{x}_i'\mathbf{t}) := \min, \ \mathbf{t} \in \mathbb{R}_p,$$



kde $h_n = [n/2] + [(p+1)/2]$ a $[a]$ značí celou část $a$. Tento odhad má bod selhání 1/2 a jeho řád konsistence je již $\sqrt{n}$.

$S$-odhad, navržený Rousseeuwem a Yohaiem [62] v r. 1984, je řešením minimalizace

$$\mathcal{S}(Y_1 - \mathbf{x}_1'\mathbf{t}, \dots, Y_n - \mathbf{x}_n'\mathbf{t}) := \min, \ \mathbf{t} \in \mathbb{R}_p,$$

kde $\mathcal{S}(z_1, \dots, z_n)$ je řešení rovnice

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{z_i}{s}\right) = k$$

vzhledem k $s > 0$ při pevných $z_1, \dots, z_n$; volbou funkce $\rho$ (obvykle ohraničené) a konstanty $k$ se určuje poměr mezi vydatností a bodem selhání odhadu. Tyto odhady a jejich výpočetní aspekty jsou podrobně popsány v knize [61].

Jinou možností vyváženějšího poměru mezi vysokým bodem selhání a vysokou vydatností je vhodně upravená jednokroková verze $M$-odhadu nebo $GM$-odhadu, začínající odhadem s vysokým bodem selhání (viz [45] a [68]).

Vysoký bod selhání těchto a podobných odhadů je však na druhé straně zaplacen některými nevýhodami, vzhledem k nimž nejsou odhady příliš využívány v praxi. Přes obtížný výpočet těchto odhadů již existují účinné algoritmy, zabudované do standardních balíků, jako S-PLUS. Nedostatkem těchto odhadů však může být, že zatímco jsou resistentní vzhledem k vysoce odlehlým hodnotám pozorování, mohou být velmi citlivé i k malým odchylkám v centru dat. Tento aspekt zatím nebyl zevrubně teoreticky analyzován, ale existuje k němu dostatečná numerická evidence, viz [35].

## 4.7 Výpočetní algoritmy

Výpočetní aspekty robustních odhadů v lineárním modelu i odhadu metodou nejmenších čtverců jsou podrobně analyzovány v knize [19], kde je obsažen i výpočetní program ADAPTIVE v systému S-PLUS, vypracovaný J. Pickem s užitím podprogramů pro regresní kvantily vypracovaných R. Koenkerem.
Program ADAPTIVE je také ke stažení na adresách
http://www.karlin.mff.cuni.cz/˜jurecko/adaptive.s
a http://www.fp.vslib.cz/picek/adaptive.htm.

# Literatura

[1] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (1972). *Robust Estimates of Location. Survey and Advances.* Princeton University Press, Princeton.

[2] J. Antoch and J. Á. Víšek (editoři) (1992). *Computational Aspects of Model Choice.* Physica-Verlag, Heidelberg.

[3] J. Antoch, H. Ekblom and J. Á. Víšek (1998). *Robust Estimation in Linear Model.* XploRe Macros: http://www.quantlet.de/codes/rob/ROB.html

[4] R. R. Bahadur (1967). Rates of convergence of estimators and test statistics. *Ann. Math. Statist.* 38, 303–324.

[5] V. Barnett and T. Lewis (1994). *Outliers in Statistical Data* (3. vydání). J. Wiley, Chichester.

[6] D. A. Belsley, E. Kuh and R. E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* J. Wiley, New York.





[7] P. J. Bickel (1975). One-step Huber estimates in the linear model. *Ann. Statist.* 1, 597–616.

[8] P. J. Bickel and E. L. Lehmann (1979). Descriptive statistics for nonparametric model. IV. Spread. *Contributions to Statistics: Jaroslav Hájek Memorial Volume* (ed. J. Jurečková), str. 33–40. Academia, Prague and Reidel, Dordrecht.

[9] P. Billingsley (1998). *Convergence of Probability Measures*, 2nd Edition. J. Wiley, New York.

[10] G. Blom (1956). On linear estimates with nearly minimum variance. *Arkiv für Mathematik* 3, 365–369.

[11] P. Bloomfield and W. L. Steiger (1983). *Least Absolute Deviations. Theory, Applications and Algorithms.* Birkhäuser, Boston.

[12] R. J. Boskovič (1757). De literaria expeditione per pontificiam ditionem et synopsis amplioris operis... *Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii* 4, 353–396.

[13] G. E. P. Box (1953). Non-normality and tests of variance. *Biometrika* 40, 318–335.

[14] G. E. P. Box and S. L. Anderson (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. Royal Statist. Soc., Ser. B* 17, 1–34.

[15] H. Bunke and O. Bunke (editoři) (1986). *Statistical Inference in Linear Models.* J. Wiley, Chichester.

[16] R. J. Carroll and D. Ruppert (1988). *Transformations and Weighting in Regression*. Chapman & Hall, London.

[17] S. Chaterjee and A. S. Hadi. *Sensitivity Analysis in Linear Regression*. J. Wiley, New York.

[18] R. D. Cook and S. Weisberg (1982). *Resials and Influence in Regression*. Chapman & Hall, London.

[19] Y. Dodge and J. Jurečková (2000). *Adaptive Regression*. Springer, New York.

[20] D. L. Donoho and P. J. Huber (1983). The notion of breakdown point. *A Festschrift for Erich Lehmann* (editoři P. J. Bickel, K. A. Doksum a J. L. Hodges). *Wadsworth, California*.

[21] N. R. Draper and H. Smith (1988). *Applied Regression Analysis,* 3. vydání. J. Wiley, New York.

[22] M. Falk (1986). On the estimation of the quantile density function. *Statist. & Probab. Letters* 4, 69–73.

[23] L. T. Fernholz (1983). *von Mises Calculus for Statistical Functionals*. *Lecture Notes in Statistics* 19, Springer-Verlag, New York.

[24] C. A. Field and E. M. Ronchetti (1990). *Small Sample Asymptotics*. IMS Lecture Notes 13, IMS, Hayward, California.

[25] J. C. Fu (1975). The rate of convergence of consistent point estimators. *Ann. Statist.* 3, 234–240.



[26] C. Gutenbrunner (1986). Zur Asymptotik von Regression Quantil Prozessen und daraus abgeleiten Statistiken. *Disertace, Universität Freiburg*.

[27] C. Gutenbrunner and J. Jurečková (1992). Regression rank scores and regression quantiles. *Ann. Statist.* 20, 305–330.

[28] C. Gutenbrunner, J. Jurečková, R. Koenker and S. Portnoy (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonpar. Statist.* 2, 307–331.

[29] *Contribution to the Theory of Robust Estimators*. PhD Thesis. University of California, Berkeley.

[30] A general qualitative definition of robustness. *Ann. Math. Statist.* 42, 1887–1896.

[31] F. R. Hampel (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* 69, 383–393.

[32] F. R. Hampel, P. J. Rousseeuw, E. Ronchetti, and W. Stahel (1986). *Robust Statistics - The Approach Based on Influence Functions*. J. Wiley, New York.

[33] F. Harrell and C. Davis (1982). A new distribution-free quantile estimator. *Biometrika* 69, 636–640.

[34] T. P. Hettmansperger (1985). *Statistical Inference Based on Ranks*. J. Wiley, New York.

[35] T. P. Hettmansperger and S. Sheather (1992). A cautionary note on the method of least median squares. *Amer. Statist.* 46, 79–83.

[36] J. L. Hodges and E. L. Lehmann (1963). Estimation of location based on rank tests. *Ann. Math. Statist.* 34, 598–611.

[37] P. J. Huber (1964). Robust estimation od a location parameter. *Ann. Math. Statist.* 36, 73–101.

[38] P. J. Huber (1969). *Théorie de l'inférence de statistique robuste.* Presses de l'Université de Montréal.

[39] P. Huber (1981). *Robust Statistics.* J. Wiley, New York.

[40] L. A. Jaeckel (1971). Robust estimation of location: Symmetry and asymmetric contamination. *Ann. Math. Statist.* 42, 1020–1034.

[41] J. Jung (1955). On linear estimates defined by a continuous weight function. *Arkiv für Mathematik* 3, 199–209.

[42] J. Jung (1962). Approximation to the best linear estimates. *Contribution to Order Statistics* (editoři A. E. Sarhan a B. G. Greenberg), str. 28–33. J. Wiley, New York.

[43] J. Jurečková (1981). Tail-behavior of location estimators. *Ann. Statist.* 9, 578–585.

[44] J. Jurečková and M. Malý (1995). The asymptotics for studentized $k$-step $M$-estimators of location. *Sequen. Anal.* 14, 229–245.

[45] J. Jurečková and S. Portnoy (1987). Asymptotics for one-step $M$-estimators in regression with application to combining efficiency and high breakdown point. *Commun. Statist. Theory and Methods A* 16, 2187–2199.



[46] J. Jurečková and P.K. Sen (1996). *Robust Statistical Procedures: Asymptotics and Interrelations.* J. Wiley, New York.

[47] J. Jurečková and P.K. Sen (1994). Regression rank scores statistics and studentization in the linear model. *Proc. 5th Prague Conf. on Asymptotic Statistics* (editoři M. Hušková and P. Mandl), str. 111–121. *Physica-Verlag, Vienna.*

[48] J. Jurečková and A. H. Welsh (1990). Asymptotic relations between $L$- and $M$-estimators in the linear model. *Ann. Inst. Statist. Math.* 42, 671–698.

[49] A. M. Kagan, J. V. Linnik and C. R. Rao (1973). *Characterization Problems in Mathematical Statistics.* J. Wiley, New York.

[50] R. Koenker and G. Bassett (1978). Regression quantiles. *Econometrica* 46, 33–50.

[51] W. Krasker and R. Welsch (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* 77, 595–604.

[52] J.-P. Lecoutre et P. Tassi (1987). *Statistique non parametrique et robustesse.* Economica, Paris.

[53] E. L. Lehmann (1983). *Theory of Point Estimators.* J. Wiley, New York.

[54] C. Mallows (1973). Influence functions. *National Bureau of Economic Research, Conference on Robust Regression*, Cambridge, Massachusetts.

[55] C. Mallows (1975). On some topics in robustness. *Memorandum, Bell Tel. Laboratories,* Murray Hill, New Jersey.

[56] R. Maronna and V. Yohai (1981). Asymptotic behavior of general $M$-estimates for regression and scale with random carriers. *Z. Wahrscheinlichkeitstheorie und verw. Gebiete* 58, 7–20.

[57] R. von Mises (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* 35, 73–101.

[58] E. S. Pearson (1931). The analysis of variance in cases of nonnormal variation. *Biometrika* 23, 114–133.

[59] H. Rieder (1994). *Robust Asymptotic Statistics.* Springer, New York.

[60] P. J. Rousseeuw (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* 79, 871–880.

[61] P. J. Rousseeuw and A. M. Leroy (1987). *Robust Regression and Outlier Detection.* J. Wiley, New York.

[62] P. J. Rousseeuw and V. Yohai (1984). Robust regression by means of $S$-estimators. *Robust and Nonlinear Time Series Analysis* (editoři J. Franke, W. Härdle a R. D. Martin), str. 256–272. *Springer, New York.*

[63] D. Ruppert and R. J. Carroll (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* 75, 828–838.

[64] P. K. Sen (1964). On some properties of the rank-weighted means. *J. Indian Soc. Agricul. Statist.* 16, 51–61.

[65] R. J. Serfling (1980). *Approximation Theorems of Mathematical Statistics.* J. Wiley, New York.



[66] A. F. Siegel (1982). Robust regression using repeated medians. *Biometrika* 69, 242–244.

[67] G. L. Sievers (1978). Estimation of location: A large deviation comparison. *Ann. Statist.* 6, 610–618.

[68] D. G. Simpson, D. Ruppert and R.J. Carroll (1992). On one-step $GM$-estimates and stability of inference in linear regression. *J. Amer. Statist. Assoc.* 87, 439–450.

[69] R. J. Staudte and S. J. Sheather (1990). *Robust Estimation and Testing.* J. Wiley, New York.

[70] S. M. Stigler (1986). *The History of Statistics. The measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press, London.

[71] J. W. Tukey (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachussets.

[72] I. Vajda (1988). *Theory of Statistical Inference and Information.* Reidel, Dordrecht.

[73] A. H. Welsh (1986). Bahadur representation for robust scale estimators based on regression residuals. *Ann. Statist.* 14, 1246–1251.

# Rejstřík

# ROBUSTNÍ STATISTICKÉ METODY

Prof. RNDr Jana Jurečková, DrSc

Lektorovali: Doc. RNDr Jaromír Antoch, CSc
RNDr Jan Picek, CSc

Vydala Univerzita Karlova v Praze
Nakladatelství Karolinum Praha 1, Ovocný trh 3
jako učební text pro posluchače Matematicko-fyzikální fakulty UK
Praha 2001
Dáno do tisku:
Vytiskla tiskárna Nakladatelství Karolinum
AA - VA -1. vydání - Náklad výtisků

Cena Kč
Publikace neprošla jazykovou ani redakční úpravou