NĚKTERÉ POSTUPY PRO VÝPOČET ROBUSTNÍCH ODHADŮ V LINEÁRNÍM REGRESNÍM MODELU

Jaromír ANTOCH

1. ÚVOD

Uvažujme klasický lineární regresní model

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}\,, \tag{1.1}$$

kde $\underset{\sim}{X} = (X_{ij})_{i=1,\dots,n}^{j=1,\dots,p}$ je matice známých regresních konstant, $\underset{\sim}{Y} = (Y_1,\dots,Y_n)'$ vektor pozorování a $\underset{\sim}{\varepsilon} = (\varepsilon_1,\dots,\varepsilon_n)'$ vektor chyb. O vektoru chyb budeme předpokládat, že jeho složky jsou nezávislé, stejně rozdělené dle distribuční funkce F, jež není obecně specifikována.

Je-li F normální, je všeobecně známo, že optimálním způsobem odhadu vektoru parametrů $\underset{\sim}{\beta}$ je klasická metoda nejmenších čtverců. Tzn., že odhad $\hat{\underset{\sim}{\beta}}$ dostaneme jako

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{p} X_{ij}\,\beta_j\Big)^2. \tag{1.2}$$

V praxi se ale občas setkáváme s pozorováními, která určitým způsobem "vybočují z řady", čímž způsobují, že klasické metody ztrácejí svoji účinnost. Proto byly v nedávné minulosti navrženy nové metody odhadu, méně citlivé na výskyt takovýchto pozorování. Jednu z nejdůležitějších tříd mezi nimi zaujímají tzv. robustní metody, jimž je v poslední době věnována velká pozornost ve všech oborech matematické statistiky.

Cílem tohoto příspěvku je zrekapitulovat některé algoritmy a ukázat postupy pro výpočet robustních odhadů v lineárním regresním modelu. Zaměřím se přitom na robustnost vzhledem k rozdělení chyb, tj. proti výskytu odlehlých pozorování v datech.

2. Submodel parametru polohy

Pro submodel parametru polohy, tj. $p=1 \;\&\; X_{ij}=1$, $i=1,\dots,n$, byly navrženy a podrobně studovány v řadě prací v zásadě 3 alternativy vzhledem k (1.2), totiž: M-odhady (odhady maximálně věrohodného typu), L-odhady (odhady založené na lineárních kombinacích pořádkových statistik) a R-odhady (odhady založené na pořadí pozorování). Zájemci o praktické použití patrně najdou nejlepší poučení v knize Andrewse all. [2], která shrnuje nejenom četné teoretické vlastnosti odhadů vhodných pro tento model, ale především výsledky rozsáhlých simulací a z nich vyplývající hodnocení a porovnání 62 různých odhadů výše uvedených tříd. Pro praktika, jakož i osoby nezabývající se programováním, je více než příjemným faktem to, že v knize jsou uvedeny programy pro výpočet všech 62 uvažovaných odhadů, napsané ve Fortranu IV., jež byly optimalizovány jak z hlediska spotřeby strojového času, tak nároků na paměť.

Pro ty, jež nechtějí ztrácet mnoho času hledáním v literatuře, připomeňme, že jsou to 10% useknutý průměr, tj. průměr počítaný z pozorování, jež zbudou po odstranění cca 10% minimálních a 10% maximálních pozorování výběru, spolu s M-odhady značenými v [2] H12, resp. H15, jež lze doporučit pro převážnou většinu praktických situací.

Vedle těchto tříd tzv. robustních odhadů je třeba se ještě v krátkosti zmínit i o existenci dalších typů odhadů, které za jistých podmínek mohou dát neméně dobré výsledky jako odhady předchozí. Jedná se především o tzv. SA-odhady, zavedené Martinem v [36] a adaptivní odhady, viz např. [18], [22] či [23].

Vypracováno během studijního pobytu v Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse.

3. Postupy robustní vzhledem k rozdělení chyb

Společným rysem všech metod robustních vzhledem k rozdělení chyb (distributional robustness) je omezit možný vliv velkých reziduí

$$\delta_i(\underset{\sim}{\beta}) = Y_i - \sum_{j=1}^{p} X_{ij}\beta_j , \qquad i=1,\dots,n, \tag{3.1}$$

odpovídajících většinou odlehlým pozorováním či chybám v datech. Zpravidla se tak děje volbou funkce méně citlivé na extrémní hodnoty reziduí než je tomu v (1.2)

V následujících paragrafech se zaměříme na některé způsoby výpočtu robustních odhadů v lineárním modelu, a to především na R,L a M odhady, včetně krátké zmínky o SA-odhadech a adaptivních odhadech.

3.1. R - odhady

R-odhad vektoru parametrů $\underset{\sim}{\beta}$ dostaneme, zhruba řečeno, jako řešení úlohy minimalizovat

$$\sum_{i=1}^{n} a_n(R_i)\,\delta_i(\underset{\sim}{\beta}) \tag{3.2}$$

vzhledem k $\underset{\sim}{\beta} = (\beta_1,\dots,\beta_p)'$, kde R_i je pořadí i-tého rezidua $\delta_i(\underset{\sim}{\beta})$ v $(\delta_1(\underset{\sim}{\beta}), \dots, \delta_n(\underset{\sim}{\beta}))$ a $a_n(.)$ je některá monotónní skórová funkce [většinou pro jednoduchost normovaná tak, aby $\sum_{i=1}^{n} a_n(i) = 0$]. Diferencujeme-li v (3.2) podle β_j, $j=1,\dots,p$, lze převést řešení minimalizační úlohy (3.2) na řešení soustavy přibližných rovnic

$$\sum_{i=1}^{n} a_n(R_i)\,X_{ij} \approx 0, \qquad j=1,\dots,p. \tag{3.3}$$

Tato varianta byla zkoumána např. Jaeckelem v [25].

Jurečková v [26] naopak místo řešení (3.3) uvažovala minimalizaci výrazu

$$\sum_{j=1}^{p} \left| \sum_{i=1}^{n} a_n(R_i)\,X_{ij} \right| . \tag{3.4}$$

Lze ukázat, že oba výše uvedené přístupy jsou (za jistých podmínek regularity) asymptoticky ekvivalentní a mají mnoho dobrých asymptotických vlastností. Bohužel, ani pro jeden z nich se mi nepodařilo nalézt v literatuře podrobně rozebraný fungující algoritmus pro jejich výpočet v obecném lineárním modelu. Metody numerické matematiky jsou však stále účinnější, což může v dohledné době přivodit zvrat i v této oblasti.

Jednokrokovým R-odhadem se v [33] zabývali Kraft a van Eeden. Jejich idea spočívá ve snaze vylepšit některý dobrý počáteční odhad (reasonable preliminary estimator) dodatečným členem počítaným za použití pořadí reziduí. Přesněji řečeno:

(a) Nechť skórová funkce $a_n(.)$ je generována nekonstantní a neklesající funkcí $\varphi(t) \in L_2(0,1)$ tak, že $a_n(i) = \varphi(\frac{i}{n+1})$. Označme $K_\varphi = \int_0^1 \varphi^2(u)\,du$.

(b) Nechť $\hat{\underset{\sim}{\beta}}^1$ je některý počáteční odhad $\underset{\sim}{\beta}$ a $\hat{\sigma}$ je některý konzistentní odhad parametru měřítka chyb σ. Autoři doporučují vzít za $\hat{\underset{\sim}{\beta}}^1$ odhad získaný metodou nejmenších čtverců či LAE odhad (tj. least absolute error estimator).

Potom odhad $\hat{\underset{\sim}{\beta}}$ vektoru parametrů $\underset{\sim}{\beta}$ je definován vztahem

$$\hat{\underset{\sim}{\beta}} = \hat{\underset{\sim}{\beta}}^1 + \frac{\hat{\sigma}}{K_\varphi} (\underset{\sim}{X}'\underset{\sim}{X})^{-1} \cdot \underset{\sim}{X}' \cdot \mathcal{R}(\hat{\underset{\sim}{\beta}}^1), \tag{3.5}$$

kde $\mathcal{R}(\hat{\underset{\sim}{\beta}}^1) = \left\{ \varphi\left(\frac{R_{(\underset{\sim}{Y}-\underset{\sim}{X}\hat{\underset{\sim}{\beta}}^1)i}}{n+1} \right) \right\}_{i=1}^{n}$ značí $(n\times 1)$ rozměrný vektor a $R_{(\underset{\sim}{Y}-\underset{\sim}{X}\hat{\underset{\sim}{\beta}}^1)i}$ je pořadí i-té komponenty reziduí počítaných pro počáteční odhad $\hat{\underset{\sim}{\beta}}^1$.

Poznámka 3.1. Základní nevýhodou tohoto přístupu je požadavek některého dostatečně dobrého počátečního odhadu. Autoři článku uvažují především klasický odhad metodou nejmenších čtverců. Jsou-li počáteční data "slušná", je i počáteční odhad

metodou nejmenších čtverců sluŠný a je sporné, zda "případné nevelké vylepšení" tohoto odhadu odpovídá námaze s jeho výpočtem. Obsahují-li však počáteční data hrubé chyby, a nemusí jich být tak mnoho, je odhad získaný metodou nejmenších čtverců zpravidla špatný a lze jen těžko věřit, že jej lze výrazně vylepšit jednou opravou. Na základě osobních zkušeností s M-odhady tomu věřit nelze. Autoři sice v poznámce pod čarou hovoří o možnosti opakovaného použití vztahu (3.5), ale bez důkazu o konvergenci takovéto procedury, nemluvě o stanovení stop pravidla.

Poznámka 3.2. Kraft a van Eeden v [33] pracují se známým rozdělením F vektoru chyb a vzhledem k němu volí optimální funkci φ generující skóry. Je-li však F neznámá, je třeba doplnit proceduru o odhad funkce φ [resp. F], jak je tomu u tzv. adaptivních procedur, viz např. [18], [22] či [23]. Adaptivní postupy mají sice výborné asymptotické vlastnosti, ale vzhledem k jejich pomalé konvergenci je třeba mít k dispozici značně rozsáhlá data pro dosažení dostatečné přesnosti, což poněkud omezuje možnosti jejich nasazení. Blíže se lze o těchto odhadech dočíst v příspěvku M.Huškové v tomto sborníku.

3.2. L - odhady

Lineární kombinace pořádkových statistik byly dlouhou dobu středem zájmu jako odhady parametru polohy či měřítka. Důvody jsou nasnadě - snadno se s nimi počítá, jsou dostatečně asymptoticky vydatné a mnohé z nich, např. α- useknutý průměr, se ukázaly být robustní vzhledem k mnoha různým navrženým kritériím. Jejich zobecnění pro lineární model však nebylo zdaleko tak přímočaré jako u M a R odhadů. Naopak, po dlouhou dobu to byl pouze návrh Bickela [8], který se touto problematikou zabýval. Teprve pojem tzv. regresních kvantilů, zavedených Koenkerem a Bassetem v [31], zlomil předchozí bariéru a dovolil definovat L-odhady parametrů v lineárním modelu, které díky snadnosti výpočtu a velmi dobrým teoretickým vlastnostem směle konkurují ostatním robustním odhadům a často je předčí.

Zmiňme se nejprve krátce o Bickelově návrhu, který je částečně analogický návrhu Krafta a van Eednové. Vychází se opět z některého počátečního odhadu $\hat{\beta}^*$ a nový odhad $\hat{\beta}$ je počítán ze vztahu

$$\hat{\beta} = \hat{\beta}^* + (X'X)^{-1} \cdot \mathcal{L}(\hat{\beta}^*), \tag{3.6}$$

kde $\mathcal{L}(\hat{\beta}^*)$ je oprava počítaná z uspořádaných reziduí pro $\hat{\beta}^*$. Vzhledem k tomu, že přesný tvar této opravy je dosti složitý a potřebuje dlouhé zavedení, bylo od něj zde upuštěno a zájemci se odkazují na [8]. Pro praktické použití Bickelova postupu nelze než zopakovat poznámku 3.1.

O mnoho zajímavější a schůdnější pro praxi se jeví definice L-odhadů v regresním modelu založená na pojmu tzv. regresních kvantilů, zavedených Koenkerem a Bassettem v [31]. Vzhledem k modelu (1.1) lze α-tý regresní kvantil, $0<\alpha<1$, definovat jako řešení úlohy minimalizovat, vzhledem k $\beta = (\beta_1, \ldots, \beta_p)'$

$$\sum_{i=1}^{n} \varrho_\alpha (Y_i - \sum_{j=1}^{p} x_{ij} \beta_j), \tag{3.7}$$

kde

$$\varrho_\alpha(x) = x \cdot (\alpha - I_{[x<0]}), \quad x \in R_1. \tag{3.8}$$

Nevychází se zde přitom z obvyklé definice výběrového kvantilu založeného na pořádkových statistikách, nýbrž z definice následující, viz [16]. Je-li $X_1, \ldots, X_n$ náhodný výběr rozdělený podle téže distribuční funkce F, potom výběrový α-kvantil, $0<\alpha<1$, budiž definován jako řešení minimalizačního problému

$$\min_{\beta \in R_1} \left\{ \sum_{\{i \mid X_i > \beta\}} \alpha \cdot |X_i - \beta| + \sum_{\{i \mid X_i < \beta\}} (1-\alpha) |X_i - \beta| \right\}. \tag{3.9}$$

Prohlédneme-li si vzorce (3.7)-(3.9) pečlivěji, vidíme, že se vlastně nejedná o nic jiného, než o jistý druh minimalizace v l_1 normě. Přesněji řečeno, řešením

je vektor minimalizující vážený součet absolutních hodnot reziduí, což je problém, o jehož řešení je známo poměrně mnoho. Tato metoda, tj. hledání řešení minimalizujícího l_1 normu, byla již v "dávné statistické minulosti" doporučována v některých učebnicích na místo metody nejmenších čtverců pro případ, obáváme-li se výskytu divokých pozorování. Podívejme se nyní na 2 základní typy L-odhadů v našem modelu, totiž α - useknutý a α-winsorizovaný odhad metodou nejmenších čtverců. O regresním medián, tj. 1/2-regresním kvantilu, se není třeba blíže zmiňovat, neboť se shoduje s řešením v l_1 normě, tj. jedná se o LAE (least absolute error) odhad.

_$L_n(\alpha)$: α - useknutý odhad metodou nejmenších čtverců_

Nechť $0 < \alpha < \frac{1}{2}$ je předem zvolená konstanta odpovídající procentu useknutí.

1⁺) Nechť $\hat{\beta}(\alpha) = (\hat{\beta}_1(\alpha), \ldots, \hat{\beta}_p(\alpha))'$ je řešení (3.7) pro α, tj. $\hat{\beta}(\alpha)$ je α-tý regresní kvantil.

2⁺) Nechť $\hat{\beta}(1-\alpha) = (\hat{\beta}_1(1-\alpha), \ldots, \hat{\beta}_p(1-\alpha))'$ je řešení (3.7) pro $(1-\alpha)$, tj. $\hat{\beta}(1-\alpha)$ je $(1-\alpha)$-tý regresní kvantil.

3⁺) Odstraňme z původního výběru všechna pozorování taková, pro něž

$$Y_i - \sum_{j=1}^{p} X_{ij}\hat{\beta}_j(\alpha) < 0 \quad \text{nebo} \quad Y_i - \sum_{j=1}^{p} X_{ij}\hat{\beta}_j(1-\alpha) > 0, \quad i=1,\ldots,n.$$

4⁺) α-useknutý odhad metodou nejmenších čtverců $L_n(\alpha)$ pak obdržíme jako odhad metodou nejmenších čtverců ze zbývajících pozorování.

_$W_n(\alpha)$: α - winsorizovaný odhad metodou nejmenších čtverců_

Nechť $0 < \alpha < \frac{1}{2}$ je předem zvolená konstanta odpovídající procentu winsorizace. Nechť $\hat{\beta}(\alpha), \hat{\beta}(1-\alpha)$ a $L_n(\alpha)$ jsou postupně α-tý regresní kvantil, $(1-\alpha)$-tý regresní kvantil a α-useknutý odhad metodou nejmenších čtverců počítané podle 1⁺) - 4⁺).

5⁺) α-winsorizovaný odhad metodou nejmenších čtverců $W_n(\alpha)$ vypočteme ze vztahu

$$(3.10) \qquad W_n(\alpha) = n^{-1}.\left\{[n\alpha].(\hat{\beta}(\alpha) + \hat{\beta}(1-\alpha)) + (n-2.[n\alpha]).L_n(\alpha)\right\}.$$

Poznámka 3.3. : O některých zkušenostech s těmito odhady se lze dočíst např. v [3] či [5]. Naše zkušenosti s nimi jsou více než dobré a lze je, zvláště α-useknutý odhad metodou nejmenších čtverců, jenom doporučit.

Domníváme-li se, že naše data jsou "celkem čistá", jak tomu zpravidla v praxi bývá, pak volbou $\alpha \sim 0.05$-0.1 jsme chráněni mnohem lépe než při přímém použití metody nejmenších čtverců před případným výskytem několika nahodilých chyb. Obáváme-li se výskytu mnoha hrubých chyb, a máme-li k dispozici dost pozorování, pak lze doporučit zvýšení α až do hranice cca 0.2. Obsahují-li data i potom hrubé chyby, je nejlepší je zahodit a provést nová měření, neb co jiného dělat s daty, v nichž je více než 40% "podivných pozorování". Pokud si takový "luxus" nemůžeme dovolit, použijme regresní medián, tj. LAE odhad.

Programy ve Fortranu IV. spolu s návodem na jejich použití lze obdržet na KPMS MFF UK od autora tohoto příspěvku.

Poznámka 3.4. O odhadech parametrů a jejich vlastnostech je známo poměrně mnoho. Mnohem méně prací však bylo věnováno problémům testování hypotéz o odhadovaných parametrech, tolik potřebných při tvorbě modelu etc. Ve speciálním případě pro model polynomické regrese je k dispozici práce J.Jurečkové [28]. Pro obecný lineární model pak lze vyjít např. z [32] či [21]. Nicméně lze konstatovat, že na tomto poli stále zůstává dostatek místa pro další iniciativu.

3.3. M - odhady

Podíváme-li se do literatury, vidíme, že jsou to M-odhady, jimž byla věnována největší pozornost. Pro některá shrnutí viz např. [21] . Podívejme se nyní krátce, co nám nabízejí.

Pro omezení vlivu odlehlých pozorování bylo navrženo místo (1.2) hledat spíše

$$\min \sum_{i=1}^{n} \varrho\left(Y_i - \sum_{j=1}^{p} X_{ij} \beta_j \right) \tag{3.11}$$

vzhledem k $\underset{\sim}{\beta} = (\beta_1, \ldots, \beta_p)'$. Funkce ϱ by přitom měla omezit vliv velkých reziduí a měla by tudíž být funkcí, jež konverguje do nekonečna pomaleji než kvadratická funkce. O volbě ϱ optimální vzhledem k apriorní informaci o rozdělení chyb se lze dočíst např. v [42] či [21] . Naneštěstí však takováto procedura není obecně invariantní vzhledem ke změně měřítka a tento parametr je třeba odhadnout současně s $\underset{\sim}{\beta}$, viz [21] . Tedy, místo (3.11) musíme hledat

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} \varrho\left(\frac{Y_i - \sum_{j=1}^{p} X_{ij} \beta_j}{\sigma} \right), \tag{3.12}$$

např. za podmínky

$$\sum_{i=1}^{n} \chi\left(\frac{Y_i - \sum_{j=1}^{p} X_{ij} \beta_j}{\sigma} \right) = (n-p)\, \frac{\gamma}{2} \tag{3.13}$$

kde $\chi(t) = t\psi(t) - \varrho(t)$, $\psi = \varrho'$, $\gamma = 2\int x\, \chi(x)\, d\Phi(x)$, $\Phi(x)$ je distribuční funkce $N(0,1)$. Důvod k zavedení podmínky (3.13) je snaha získat asymptoticky nestranný odhad σ v případě, jsou-li chyby normálně rozděleny, viz např. [21] .

Předchozí úloha se dá převést na hledání minima výrazu

$$\sum_{i=1}^{n} \varrho\left(\frac{Y_i - \sum_{j=1}^{p} X_{ij} \beta_j}{\sigma} \right) \cdot \sigma + a\sigma \tag{3.14}$$

vzhledem k σ a $\underset{\sim}{\beta}$. Má-li ϱ spojitou derivaci ψ a zvolíme-li $a = (n-p) \cdot \frac{\gamma}{2}$, lze se snadno přesvědčit, že minimalizace (3.14) vzhledem k σ a $\underset{\sim}{\beta}$ je ekvivalentní minimalizací (3.12) vzhledem k $\underset{\sim}{\beta}$, přičemž (3.13) je splněna.

Uveďme nyní tzv. jednoduchý algoritmus pro výpočet robustních odhadů, tak jej nazývají jeho autoři Huber a Dutter v [20]. Pro úplnost dodejme, že jeho konvergence byla dokázána za následujících podmínek na funkci $\varrho(x)$: $\varrho(x)$ je 2x diferencovatelná a konvexní symetrická funkce, $\varrho(t) \geq 0$, $\varrho(0) = 0$, $\psi(t)/t$ je konvexní pro $t < 0$ a konkávní pro $t > 0$, $\lim_{t \to 0} \psi(t)/t < \infty$, $0 \leq \varrho'' \leq 1$.

Jednoduchý algoritmus pro M-odhady

Nechť $\hat{\underset{\sim}{\beta}}^{(0)}$ a $\hat{\sigma}^{(0)}$ jsou některé počáteční odhady $\underset{\sim}{\beta}$ a σ.

1°) Nechť $m=0$, kde m je konstanta označující počet iterací.

2°) Spočteme rezidua $\delta_i^{(m)} = Y_i - \sum_{j=1}^{p} X_{ij} \hat{\beta}_j^{(m)}$, $i=1,\ldots,n$.

3°) Spočteme nový odhad σ^2 z výrazu

$$(\hat{\sigma}^{(m+1)})^2 = \frac{1}{a} \sum_{i=1}^{n} \chi\left(\frac{\delta_i^{(m)}}{\hat{\sigma}^{(m)}} \right) \cdot (\hat{\sigma}^{(m)})^2 .$$

4°) Spočtěme hodnoty $\Delta_i^{(m)}$ tzv. winsorizovaných reziduí, tj.

$$\Delta_i^{(m)} = \psi\left(\frac{\delta_i^{(m)}}{\hat{\sigma}^{(m+1)}} \right) \cdot \hat{\sigma}^{(m+1)}, \qquad i=1,\ldots,n.$$

5°) Spočtěme minimum, vzhledem k $\underset{\sim}{\tau}$, výrazu

$$\sum_{i=1}^{n} \left(\Delta_i^{(m)} - \sum_{j=1}^{p} X_{ij} \tau_j \right)^2 .$$

a označme jej $\hat{\underset{\sim}{\tau}}^{(m)}$.

6°) Nový odhad $\hat{\beta}^{(m+1)}$ dostaneme jako

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \hat{\tau}^{(m)} \cdot K,$$

kde $0 < K < 2$ je některá předem zvolená konstanta pomáhající nám zrychlit (zpomalit) proceduru, blíže viz [10].

7°) Jestliže $\| \hat{\tau}^{(m)} \| < \varepsilon_0$, kde ε_0 je některá předem předepsaná tolerance, ukončeme proceduru a položme $\hat{\beta} = \hat{\beta}^{(m+1)}$ a $\hat{\sigma} = \hat{\sigma}^{(m+1)}$. V opačném případě položme $m = m+1$ a vraťme se do bodu 2°.

Podívejme se ještě na 2 modifikace předchozího algoritmu.

Modifikace I. : spočívá v přehození bodů 3° a 4° a poněkud odlišném způsobu odhadu σ., tj.

$3^{o`}$) $\Delta_i^{(m+1)} = \psi\left(\frac{\delta_i^{(m)}}{\hat{\sigma}^{(m)}}\right) \cdot \hat{\sigma}^{(m)}$, $i = 1, \ldots, n$,

$4^{o`}$) $(\hat{\sigma}^{(m+1)})^2 = \frac{1}{2a} \sum_{i=1}^{n} (\Delta_i^{(m+1)})^2$.

Ostatní kroky jsou shodné s předchozí procedurou. Soudě podle výsledků v [11], nedává nám tato modifikace odhady zásadně odlišné od metody předchozí. Naopak lze říci, že odhady počítané z týchž dat jsou si zcela rovnocenné.

Modifikace II. : vychází také z předchozího algoritmu, ale využívá více vlastností lineárního modelu.

Předpokládejme, že matice X má plnou hodnost, potom řešení bodu 5° lze psát ve tvaru

$$\hat{\tau}^{(m)} = (X`X)^{-1} X` \Delta^{(m)}, \tag{3.15}$$

kde $\Delta^{(m)}$ je vektor winsorizovaných reziduí z bodu 4°. Pro další iterační krok potřebujeme vektor reziduí $\delta^{(m+1)}$ vzhledem $\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + K\hat{\tau}^{(m)}$, který lze ale počítat přímo z $\delta^{(m)}$ a $\Delta^{(m)}$, neboť

$$\delta^{(m+1)} = Y - X\hat{\beta}^{(m+1)} = Y - X\hat{\beta}^{(m)} - X\hat{\tau}^{(m)} \cdot K = \delta^{(m)} - X(X`X)^{-1} X` \Delta^{(m)}. \tag{3.16}$$

Spočteme-li si tedy na počátku projekční matici $X(X`X)^{-1}.X`$ a schováme její horní trojúhelníkovou část, lze využít předchozí algoritmus, aniž bychom explicitně v každém kroku počítali opravný vektor $\hat{\tau}^{(m)}$. Proceduru zastavíme tehdy, je-li $\| X(X`X)^{-1} X` \Delta^{(m)} \| < \varepsilon_0^*$, kde ε_0^* je předem zadaná tolerance (různá od ε_0). Výsledný odhad vektoru parametrů $\hat{\beta}$ pak má tvar

$$\hat{\beta} = (X`X)^{-1} X`(Y - \delta^{(m+1)}), \tag{3.17}$$

tj. $\hat{\beta} = \hat{\beta}^{(m+1)}$.

Poznámka 3.5.: Předložený algoritmus (a jeho varianty) není jediný způsob pro hledání M-odhadů, viz např. [10] . Prostudujeme-li si však většinu ostatních algoritmů pečlivěji, zjistíme, že se jedná více méně o variace na předchozí téma, zpravidla využívají apriorní informaci ať již o funkci φ, či tvaru chyb ap. Výjimkou jsou metody využívající technik vážených nejmenších čtverců, viz [41] , [15] či [24].

Poznámka 3.6.: Připomeňme, že v MS BÚ ČSAV Praha je k dispozici rozsáhlá knihovna podprogramů pro výpočet M-odhadů, viz [17] . Od vydání sborníku ROBUST 80 byla mezitím p.Malinou napsána řada programů umožňujících zájemcům o praktické vyzkoušení pouze přijít, naděrovat data v předepsaném tvaru a odeslat k výpočtu. Více snad již ani nabídnout nelze - jen to děrování, ale na to zase nejsou lidi.

3.4. SA - odhady

Odstavec 3 by nebyl úplný, kdybychom se vedle R, L a M odhadů nezmínili o odhadech založených na algoritmu stochastických aproximací (SA) Robins-Monroova (RM) typu, jež jsou dány jednoduchým rekurzivním vztahem a jsou vhodné pro aplikace v reálném čase.

Nechť $\hat{\beta}^{(0)}$ je některý počáteční odhad a $\gamma>0$ konstanta. Potom odhad $\hat{\beta}_{SA}$ vektoru β v modelu (1.1), definovaný rekurzivním vztahem

$$\hat{\beta}^{(n)} = \hat{\beta}^{(n-1)} + \frac{\gamma}{n} (X'X)^{-1} X' \Psi(Y - X\hat{\beta}^{(n-1)}), \tag{3.18}$$

$n=1,2,\ldots$ nazveme SA-odhadem příslušným k funkci Ψ. Funkce Ψ se volí zpravidla táž jako u M-odhadů. Bližší informaci o těchto odhadech lze nalézt např. v [36], [37] či [38].

Odhady tohoto typu se v ČSSR zabývá např. dr.J.Novovičová z ÚTIA ČSAV Praha se svými spolupracovníky. Tento tým nejenže odvodil řadu zajímavých teoretických výsledků o odhadech typu SA-RM, ale připravil řadu programů pro jejich výpočet. Zájemce o případné praktické využití proto odkazujeme na výše uvedené pracovníky.

4. Závěr

Jak už jsem se zmínil v úvodu a snažil se dokumentovat v textu, četnost a kvalita prací věnovaných výpočetní stránce robustních metod není zdaleka ve shodě s ohromným množství prací teoretických. A to jak v oblasti algoritmů, tak ještě ve větší míře v oblasti skutečných realizací.

Nicméně nelze říci, že by na tomto poli nic neexistovalo. Naopak, možnosti existují nejenom ve světě, ale i u nás. K dispozici jsou poměrně velmi dobře fungující knihovny programů na výpočet jak L, tak M odhadů. Tyto programy umožňují experimentovat nejenom na poli statistiky - při zkoumání chování odhadů na datech umělých i z praxe, ale i na poli numeriky a programování - vylepšování programů, jejich optimalizace, hledání nových algoritmů etc.

Zbývá však ještě krok třetí, nejtěžší. Je jím dovedení robustních metod mezi běžné uživatele do každodenní praxe, neboť teprve potom bude všechno to nezměrné úsilí náležitě zhodnoceno. Na tomto poli jsme zatím postoupili nejméně, a to nejenom u nás, ale i ve světě. I zde se však vlak pomalu rozjíždí a akce typu ROBUST se snaží jej co nejvíce zrychlit.

Literatura

[1] Adelmalek N.N., On the discrete linear l_1-approximation and l_1 solutions od overdetermined linear equations. Journal of Approximation theory 11(1974), 38-53.

[2] Andrews D.F., Bickel P.J., Hampel F.R., Huber P.J., Tukey J.W., Robust estimates of location. Survey and advances. Princeton University Press, 1972.

[3] Antoch J., Collomb G., Hassani S., Robustness in parametric and non-parametric regression estimation: An investigation by computer simulations. Proceedings of COMPSTAT 1984, Physica Verlag, Viena, 1984.

[4] Antoch J., Soubor programů pro robustní regresní analýzu. Publ.de Laboratoire de Statistique et Probabilités, Toulouse, 1984.

[5] Antoch J., L-estimates in linear regression model. Publ.de Laboratoire de Statistique et Probabilités, Toulouse, 1984.

[6] Barrodale I., Roberts F.D.K., An improved algorithm for discrete l_1-linear approximation.SIAM J.Numer.Anal. 10,5 (1973), 839-848.

[7] Barrodale I., Roberts F.D.K., Solution of an overdetermined system of equations in the l_1-norm. CACM 17 (1974), 319-320.

[8] Bickel P.J., On some analogues to linear combinations of order statistics in the linear model. AS 1 (1973), 597-616.

[9] Dione L., Efficient nonparametric estimators of parameters in the general linear hypothesis. AS 4 (1976).

[10] Dutter R., Robust regression: Different approaches to numerical solutions and algorithms. Research report N° 6, ETH Zürich 1975.

[11] Dutter R., Numerical solution of robust regression problems: Computational aspects and comparison. Research report N° 7, ETH Zürich, 1975.

[12] Dutter R., Blinwdr: Robust and bounded influence regression. Proc.of COMPSTAT 1982, Caussinus et al.(eds.), Physica-Verlag, Viena, 1982

[13] Dutter R; Computer program BLINWDR for robust and bounded influence regression. Res.Rep.No 6 (Version 2). Inst.für Statistik, Technische Univeristät Graz, 1983.

[14] Dutter R., Covinter : A computer program for computing robust covariances and for plotting tolerance elipses. Res.Rep. No.10, idem.

[15] Dutter R., Statistical computing: robust regression.Handout at 6th International Summer School on Problems of Model Choise and Parameter Estimation in Regression Analysis, Sellin 1983, NDR.

[16] Ferguson T.S., Mathematical statistics. A decision theoretic approach. New York, Academic Press, 1967.

[17] Havránek T. - Antoch,J., ROBETH-knihovna programů pro robustní statistické metody. Sborník ROBUST 80, 25-33, JČSMF 1980.

[18] Hogg R.V., Adaptive robust procedures. JASA 69(1974), 909-923.

[19] Huber P.J., Robust estimation of regression parameter. AMS 35(1964), 73-101.

[20] Huber P.J., Dutter R., Numerical solution of robust regression problems. Proc.of COMPSTAT 1974, Physica Verlag, Viena ,1974.

[21] Huber P.J., Robust Statistics. J.Wiley & Sons, Inc., 1981.

[22] Hušková M., Adaptive procedures. Acta Univ.Carolinae-Math. et Ph.24(1983), 41-48.

[23] Hušková M., Adaptivní odhady. Sborník ROBUST 84, JČSMF 1984.

[24] Chambers J.M., Computational methods of data analysis, J.Wiley, 1977.

[25] Jaeckel L.A., Estimating regression coefficients by minimizing the dispersion of the residuals. AMS 43 (1972), 1449-1458.

[26] Jurečková J., Nonparametric estimate of regression coefficients. AMS 42 (1971), 1328-1338.

[27] Jurečková J., Winsorized least-squares estimator and its M-counterpart. Contributions to Statistics. Essays in Honour of N.L.J.Johnson,P.K.Sen(ed.), 237-245, North Holland, 1983.

[28] Jurečková J., Trimmed polynomial regression.CMUC 24,4(1983), 597-607.

[29] Jurečková J., Robust estimators of location and regression parameters and their second order asymptotic relations. Trans. 9th Prague Conf.on Inf. Theory etc.,

[30] Jurečková J., Regression quantiles and trimmed least-squares estimator under a general design. Subm.to Kybernetika,1984.

[31] Koenker R., Basset G., Regression quantiles.Econometrica 46(1978),33-51.

[32] Koenker R., Bassett G., Tests of linear hypotheses and l_1-estimation. Econometrica 50 (1982), 1577-1583.

[33] Kraft C., van Eeden C., Linearized rank estimates and sign-rank estimates based on ranks in regression. AMS 42 (1972), 42-57.

[34] Lawson C.L., Hanson R.J., Solving least squares problems. Prentice Hall,Inc., Englewood Cliffs, New Jersey, 1974.

[35] Marazzi A., Robeth : A subroutine library for robust statistical procedures. Proc.of COMPSTAT 1980, Barrit et al.(eds.), Physica-Verlag, Viena, 1980.

[36] Martin R.D., Robust estimation of signal amplitude. IEEE Trans.Inform.Th. II-18, 596-606.

[37] Novovičová J., M-odhady a SA-odhady v lineárním regresním modelu. Sborník ROBUST 80, 119-133, JČSMF, 1980.

[38] Novovičová,J., SA-odhady regrese. Kandidátská disertační práce,ÚTIA ČSAV Praha, 1983.

39 Ruppert D., Carroll R.J., Robust regression by trimmed least-squares estimation. Inst.of Statist. Mimeo Series No 1186, Univ.of North Carolina at Chapell Hill, 1978.

40 Ruppert D., Carroll R.J., Trimmed least squares estimation in the linear model. JASA 75 (1980), 828-838.

41 Samarov A.,Welsch R.E., Computational procedures for bounded-influence regression. Proc.of COMPSTAT 1982, Caussinuss et all.(eds.), Physica Verlag, Viena, 1982.

42 Цыпкин Я.З.,Поляк Б.Т.,Огрубленный метод максимального правдоподобия. Сб."Динамика систем",12,22-46.