ROBETH - knihovna podprogramů pro robustní statistické metody

T. Havránek, J. Antoch

0. Úvod: V posledních 15-20.letech můžeme ve statistické literatuře vysledovat široký proud tzv. "robustní statistiky". Její vliv neustále vzrůstá a již daleko přesáhl rámec teorie odhadu, odkud základní myšlenkové postupy vzešly. Při tom nejde pouze o vliv teoretický, ale i praktický.

Bohužel,na rozdíl od klasické metody nejmenších čtverců, není s robustními metodami vesměs spojena snadná vypočitatelnost, spíše naopak. Tato překážka však odpadá s nasazením moderní výpočetní techniky. Nestačí ale jen výkonná technika. Jeden z hlavních cílů, jež si robustní metody kladou, je bránit se proti některým chybám tak častým v praxi (odlehlá pozorování, závislost...). Chceme-li tyto metody uvést do každodenní praxe - především nespecialistů - je třeba mít po ruce vhodné balíky programů s jasnou a srozumitelnou motivací, jednoduchou použitelností a snadnou interpretací výsledků. Soudíme, že toto je hlavní cesta, jak dané pěkné matematické výsledky přiblížit nejširší veřejnosti .

Předkládaný soubor ROBETH je jedna z možných nabídek.

I. Základní počítačová struktura

Soubor ROBETH není systém pro"přímé" počítání robustních metod,nýbrž jde o knihovnu podprogramů pro tyto metody. Hlavní zájem je přitom soustředěn na odhady v lineárním modelu.

Celý soubor, který byl autorem napsán ve FORTRANU IV. a odladěn na počítači CDC 3300, byl zimplementován na počítači IBM 370/135 Matematickým střediskem MBÚ ČSAV. Implementaci provedl B.Zouvar, jemuž patří náš dík. Přenos (včetně testování) proběhl bez problémů. Nicméně ve formě, v níž je nyní k dispozici,se nehodí pro běžné "rutinní" používání a je třeba napsat příslušné hlavní programy.

Základem systému jsou tři moduly MAIN 1-3,obsahující zákl. subroutiny.Tyto jsou pak doprovázeny moduly AUX 1-3, obsahujícími pomocné podprogramy a modulem UTILITY, ve kterém jsou uloženy podprogramy pro maticové operace, třídění, generování apod.

Celý systém je bohatě strukturován. Jako příklad si ukažme uživatelem napsaný hlavní program USER pro vypočtení odhadů koeficientů v lineárním modelu Schweppeho metodou s Hampel-Kraskerovými vahami.

Část definující potřebná pole, načítání dat a tisk výsledků vynecháme.

```
      Program  USER
```

zde následuje deklarace, základní informace o modelu,polích COMMON, čtení dat apod.

```
                CALL WHKO (...)
                CALL BETC (...)
                CALL MTRF (...)
                CALL UCOV (...)
                CALL CLLS (...)
                CALL RREG (...)
                CALL SCOV (...)
```

následuje tisk výsledků

```
                END
```

Vysvětlíme si nyní význam jednotlivých podprogramů a zároveň uvedeme všechny další podprogramy z AUX 1-3 a UTILITY, které jsou v jednotlivých podprogramech volány.

WHKO - výpočet vah pro jednotlivá pozorování
Volá : ICOV,ACOV .

BETC - výpočet korekčního faktoru pro nestrannost odhadu rozptylu.
Volá : NORM - normální rozdělení;
ERROR- chybová hlášení.

MTRF - úprava matice pro výpočet (provede se horní triangularizace), výsledek slouží jako vstup pro UCOV a RREG.
Volá : M12 - jednoduchá Hausholderova transformace;
DIFF- výpočet differencí;
ERROR.

UCOV - Výpočet kovarianční matice parametrů odhadu (s horní trojúhelníkovou maticí plánu)
Volá : ERROR.

CLLS - klasická metoda odhadu parametrů metodou nejmenších čtverců. Dává počáteční hodnoty pro iterativní získávání robustních odhadů.
Volá : M12, ERROR;
SWAP - výměna obsahu dvou jednorozměrných polí;
SOLV - řešení trojúhelníkového systému lineárních rovnic;
RES - výpočet residuálů pomocí speciální Hausholderovy transformace (R3V);
PERM - permutace složek vektoru;
NRM2 - výpočet euklidovské normy vektoru.

RREG - provádí výpočet robustních odhadů regresních parametrů.
Volá : M12;
MALG - modifikovaný Huberův algoritmus řešení

Volá : RES, SOLV, M12, SWAP;
NEWSIG - nové hodnoty odhadů v iteraci;
HUB - počítá winsorisované residuály;
PSI,CHI- základní ztrátová M-funkce (typu external)definující model.

SCOV - výpočet kovarianční matice parametrů odhadu v robustním modelu.
Volá : PSI, CHI;
PSIPRM(X) - výpočet(numerický) první derivace funkce PSI.

Snažili jsem se zde ukázat mnohotvárnost celého systému a použití prvků z AUX 1-3 a UTILITY jako jednotlivých "stavebních kamenů". Zdůrazňujeme ještě jednou, že uživatel se stará pouze o volání "hlavních podprogramů z MAIN 1-3.

II. Hlavní obsah systému

V tomto odstavci se podrobněji zmíníme o obsahu hlavních modulů MAIN 1-3. Hlavní pozornost při tom je soustředěna na MAIN 2, který tvoří jádro celého souboru.

2.1.: MAIN 1 - tento modul obsahuje základní typy tzv. R-odhadů (tj. robustních odhadů založených na pořadí) pro případ odhadu parametru polohy symetrické hustoty $f(x-\theta)$; $\theta \in R_1$. Protože netvoří podstatnou část souboru a je zařazen především pro možnost srovnání, odkazujeme zájemce spíše na [1] , kde je uvedeno mnoho dalších typů (včetně simulace a hodnocení). Některé z těchto odhadů jsou též zařazeny do systému statistických programů BMDP (viz [2]).

2.2.: MAIN 2 - umožňuje nalezení odhadů parametrů v lineárním modelu a tvoří jádro celého systému. Obsahuje podprogramy pro :

1) Klasické odhady parametrů metodou nejmenších čtverců (ale netradičním numerickým postupem).
2) M-odhady parametrů lineárního modelu.
3) Určování vah pro redukci vlivu odlehlých hodnot.
4) Výpočet kovarianční matice odhadů parametrů.

2.2.1.: Model: Budeme uvažovat klasický lineární model

$$Y = X\theta + e ,$$

kde $Y(nx1)$ je vektor náhodných veličin, $X(nxp)$ matice experimentu (uspořádání), $\theta(px1)$ vektor neznámých parametrů a $\mathcal{L}(e) \sim N(0, \sigma^2 I_n)$, kde I_n je jednotková matice a σ^2 označuje neznámý rozptyl.

M - odhadem parametrů θ a σ rozumíme řešení systému rovnic (v θ a σ)

$$\sum \psi(r_i/\sigma w_i) w_i x_{ij} = 0 \qquad j = 1,\ldots,p$$

$$\sum \chi(r_i/\sigma w_i) w_i^2 = \text{CONST} \tag{2.1}$$

(odpovídající

$$\min_{\theta} \sum_i \varphi(r_i/\sigma w_i)\sigma w_i^2 + \sigma.\text{CONST})$$

(tzv. obecný SCHWEPPEHO model), kde $r_i = y_i - x_i\theta$ (x_i je i-tý řádek matice X a y_i napozorovaná hodnota). Funkce ψ, φ a χ mohou být buď uživatelem definované "vhodné funkce" [viz [1],[5]] (typu external), jinak se používá tzv. Huberovy funkce

$$\psi(x) = \min(c, \max(-c, x)) \quad \text{a} \quad \chi(x) = \psi^2(x)/2 .$$

Váhy pro redukci vlivu odlehlých pozorování mohou být opět definovány uživatelem dle "určitých pravidel" (blíže viz např. [1] , [5a]), nebo lze použít standardní volby W0-W4 obsažené v MAIN 2. Motivací k zavedení vah byl fakt, že dané postupy nás mají především chránit před vlivem odlehlých (nehomogenních) pozorování, jež se dostaly do výběru "omylem" a vážně narušují klasické postupy.

Kromě obecného Schweppeho modelu (2.1) ROBETH uvažuje specielně:

1) Huberův model - získáme z (2.1) volbou $w_i = 1$ $\forall i$.

2) Mallowsův model - získáme z (2.1) transformacemi

$$w_i^* = \sqrt{w_i}, \qquad x_i^* = w_i^* . x_i \quad \text{a} \quad y_i^* = w_i^* . y_i .$$

Při určení vah se postupuje následovně. Pro danou konstantu β se nalezne regulární matice A(pxp) taková, aby

$$\frac{1}{n} \sum_{i=1}^{n} u(|z_i|) z_i z_i' = \beta I_p ,$$

kde $z_i = Ax_i'$ a $u(t)$ je některá "vhodná" funkce. Váhy se nyní definují vztahem $w_i = f(|z_i|)$, kde f se liší podle jednotlivých verzí. Systém nabízí (kromě W0 ≡ Huberův model - $w_i = 1$ $\forall i$) následující možnosti.

W1) Mallowsův model - váhy Hampel (WHMP) - získáme volbou $u(t) = \min[1, \text{BOUND}/t^2]$, $f(z) = u(|z|)$ a $\beta = 1$. Při výpočtu se používají podprogramy ICOV a ACOV z MAIN 3.

W2) Mallowsův model - váhy Mallows a Maronna (WMMP) - získáme volbou

$$u(t) = \min[1, \text{BOUND}/t^2], \quad f(|z|) = \sqrt{u(z)}$$

a $$\beta = (1/p)\, E\{u(|Z|)|Z|^2\},$$

kde střední hodnota je počítána za předpokladu $\mathcal{L}(Z) \sim N(0, I_p)$.

W3) Schweppeho (původní) návrh (WHAT) - získáme volbou $w_i = \sqrt{1-h_{ii}}$, kde h_{ii} je i-tý diagonální prvek matice $H = X(X'X)^{-1}X'$.

Volba CONST v (2.1) je podřízena snaze získat asymptoticky nestranný odhad σ^2 (za předpokladu normálního rozložení reziduí). Nepoužije-li uživatel své konstanty, umožňuje systém stanovit CONST = (n-k) BETA podprogramem BETC následovně:

1) Huberův návrh - $\text{BETA} = \int \chi(u)\varphi(u)du$.

2) Schweppeho návrh -

$$\text{BETA} = \frac{1}{n}\sum_{i=1}^{n}\left\{w_i^2 \int \chi(u/w_i)\varphi(u)du\right\}.$$

3) Mallowsův návrh -

$$\text{BETA} = \frac{1}{n}\sum_{i=1}^{n}\left\{w_i \int \chi(u)\varphi(u)du\right\}.$$

2.3. Konvergence algoritmu: Jak je všeobecně známo, řešením odhadu parametrů v lineárním modelu je vektor θ minimalizující Euklidovskou normou. Daný problém se při tom převádí obvykle na řešení příslušných normálních rovnic. Hledáme-li nyní intuitivní smysl rovnic (2.1), vidíme, že zde jde v podstatě o odhad minimalizující jinou než kvadratickou ztrátovou funkci a klasický odhad metodou nejmenších čtverců je speciálním případem M-odhadů. Přechod k adekvátním "normálním" rovnicím je zřejmý. Vzniká při tom přirozená otázka, za jakých podmínek daný algoritmus konverguje. Odpověď dává (Huber 1977):

Tvrzení: Nechť ztrátová funkce $\varphi(x)$ splňuje : $\varphi(0) = 0$, $\varphi(x)$ je konvexní a nezáporná, nechť

$$0 < \lim_{|x|\to\infty} \frac{\varphi(x)}{|x|} = L \leq +\infty .$$

Nechť $\psi(x) = \varphi'(x)$, $\chi(x) = x\psi(x) - \varphi(x)$. Potom algoritmus pro řešení odhadu podprogramem RREG konverguje.

2.4: MAIN 3 - slouží především pro výpočet afinní kovariační matice pomocí následující metody.

Nechť $X_1,\ldots,X_n$ jsou iid rv's dané hustotou $f(x,b,A) = |\det A|.g(|A(x-b)|)$ (g-sféricky symetrická v R^p) a $x \to A(x-b)$ je nedegenerovaná afinní transformace. Odhady $\hat{A}$ a $\hat{b}$ získáme řešením rovnic

$$\frac{1}{n}\sum_{i=1}^{n} w(|z_i|)z_i' = 0 \quad \text{a}$$

$$\frac{1}{n}\sum_{i=1}^{n} (u(|z_i|))z_i z_i' - v(|z_i|)I_p = 0_p.$$

kde $z_i = A(x_i - b)$, $i=1,\ldots,n$; x_i značí i-tý řádek matice X a u,v,w jsou váhové funkce. Váhy mohou být opět buď zvoleny ze systému nebo individuelně. ROBETH navrhuje:

$$u(t)=\begin{cases} a2/t^2 & t^2 < a2 \\ 1 & a2 \le t^2 \le b2 \\ b2/t^2 & b2 < t^2 \end{cases} \qquad w(t)=\begin{cases} 1 & t < c \\ \frac{c}{t} & t \ge c \end{cases}$$

$$v(t) = \beta \quad \text{(konst.)}.$$

Koeficienty a2,b2 a c mohou být (poměrně složitě) spočteny pomocí (a2b2), (LOC) z ořezaných pozorování podobně jako β.

Výpočet odhadů $\hat{A},\hat{b}$ je iterativní a vychází z následujících počátečních hodnot (blíže viz [1])

$$b_j^0 = \underset{e}{\text{med}}\{x_{ej}\}, \quad j=1,\ldots,p$$

$$a_{jj}^0 = 0.6745/\underset{e}{\text{med}}\,(x_{ej} - b_j^0), \quad j=1,\ldots,p$$

$$a_{ij}^0 = 0 \quad i \neq j.$$

Robustní kovarianční matice $(A'A)^{-1}$ (nalezená podprogramem (RCOV)) je pak dána vztahem $\tau^2 \hat{A}^{-1}(\hat{A}^{-1})'$, kde τ je korekční faktor - není-li stanoven jinak, je vypočten pomocí (TAU2) jako řešení rovnice

$$E(u(\tau|Z|)\,|\tau Z|^2 = p.$$

2.5. Výpočty kovarianční matice odhadů - mohou být provedeny jedním z následujících podprogramů:

UCOV - výpočet kovarianční matice COV v řešení metodou nejmenších čtverců. Používá algoritmu COV z [3].

SCOV - výpočet kovarianční matice při řešení Huberova modelu; hledanou matici získáme z předchozí ve tvaru

$$(PV).(COV).V'P',$$

kterou je třeba násobit vhodným měřítkovým faktorem (o maticích P,V viz odstavec III.).

SCOW - výpočet kovarianční matice pro Schweppeho model.

III. Vlastní algoritmus odhadu parametrů

3.1. : ROBETH může provádět nejprve odhad parametrů klasickou metodou nejmenších čtverců (CLLS). Výsledky buď slouží jako počáteční hodnoty pro iterativní hledání robustních odhadů (RREG),

nebo mohou být použity jako takové. Postup je volen dle [3], o Hausholderově rozkladu viz [6] str.105-6, 116.

Nejprve je provedena horní triangularizace matice X a určena pseudohodnost k (MTRL). Zároveň jsou hledány matice P,Q a R tak,že X.P = Q.R, kde P je (pxp) permutační matice, R je horní trojúhelníková matice (nxp) a Q je ortogonální matice (nxn). Zároveň $r_{11} \geq r_{22} \geq \ldots$. Pseudohodnost k je největší j takové, že $r_{jj} > \tau$, kde τ je vhodně zvolená konstanta (obrana proti "skoro" singulárním maticím). Zároveň je nalezena ortogonální matice V tak,že

$$R = \begin{bmatrix} R_1 & R_2 \\ \hline 0 & R_3 \end{bmatrix} . V = [U, 0] ,$$

kde R_1 je řádu (kxk), U je horní trojúhelníková a 0 nulová matice. Matice P,Q,R,V a U, jakož i k, jsou dále používány při stanovování robustních odhadů.

3.2. : Iterativní hledání robustních odhadů

Jde o upravený Huberův, resp. Huber-Dutterův algoritmus, viz [5b], [4]. Počáteční hodnoty mohou být získány pomocí CLLS nebo i řešením vázané lineární regrese (viz odst.II.).

Vlastní program pro iterativní hledání robustních odhadů (RREG) potřebuje jako vstupní hodnoty matice Q,R,P,V a U, počáteční hodnoty odhadů θ_o, čísla TOL a MAXIT (dána tolerance a maximální počet iterací), matici COV (jde o výsledek podprogramu UCOV hledající kovariančí matici odhadů) a číslo BETA. Podprogram při tom pracuje s hodnotami CONST= (n-k)BETA, $T=(PV)^{-1}\theta_o$ a Z = QY (tj. s transformovanými parametry). Počet provedených iterací (limitováno hodnotou MAXIT) lze nalézt v konstantě NIT. Každá iterace má pět kroků:

1) polož $q = Z - RVT$, $r1 = Q^{-1}q$

2) vypočti novou hodnotu pro $\sigma = \sqrt{s^2}$, kde

$$s^2 = \frac{1}{CONST} \sum_{i=1}^{n} \chi\left\{\frac{r1_i}{\sigma w_i}\right\} . (\sigma w_i)^2$$

3) winsorizuj reziduály

$$r1_i = \psi\left\{\frac{r1_i}{\sigma w_i}\right\} . (\sigma w_i) \qquad i=1,\ldots,n$$

4) vypočti $q = Qr1$ a d takové, že

$$\begin{bmatrix} d_1 \\ \vdots \\ d_k \end{bmatrix} = U^{-1} \begin{bmatrix} q_1 \\ \vdots \\ q_k \end{bmatrix}$$

$$d_i = 0 \qquad i=k+1,\ldots,p$$

5) oprava odhadů : polož T : = T + d.

Iterace se zastavuje, jestliže jsme vyčerpali maximální počet iterací MAXIT nebo jsme pod předepsanou toleranci, tj.

$$|d_j| < TOL.\sqrt{(COV)_{jj}} \qquad j=1,\ldots,p.$$

Jako výstup dostáváme odhad parametrů $\hat{\theta}$ = RVT, residuály a poslední změnu v parametrech Δ = PVd.
Residuály jsou přitom dvojího druhu; jednak

r1 – týkající se "uřiznuté"matice uspořádání a dále pro "celou" matici X = Q' RP'.

IV. Numerický příklad. Ilustrujme si dané metody na případu lineární regrese. K dispozici je 13 dvojic dat, odhady značíme klasicky $\hat{a},\hat{b}$. Data pro kontrolu udáváme:

① = (17.6, 20.9, 21.6, 26.0, 27.1, 27.6, 27.8, 32.6, 33.4, 35.1, 37.0, 38.7, 77.6),

② = (15.7, 18.0, 19.9, 23.4, 19.7, 23.1, 23.8, 24.9, 26.1, 27.6, 26.1, 31.3, 44.9).

Hodnoty ① (nezávislé proměnné) tvoří s jednotkovým vektorem matici X, hodnoty 2 sloupcový vektor Y. V následující tabulce jsou shrnuty výsledky jednotlivých odhadů (pro zajímavost jsou připojeny i hodnoty NIT a BETA).

Typ odhadu	$\hat{a}$	$\hat{b}$	c	NIT	BETA
CLLS	9.514	0.475	1.785	-	-
A:HUBER	9.512	0.473	1.952	6	0.356280
B:SCHWEPPE (váhy Schweppe)	8.840	0.498	1.929	37	0.321857
C:SCHWEPPE(váhy Hampel-Krasker)	8.707	0.504	1.740	36	0.438303
D:MALLOWS (váhy Marona)	6.897	0.568	1.700	15	0.386877
E:MALLOWS (váhy Hampel)	6.141	0.596	1.652	6	0.332456

Uveďme si ještě váhy pro jednotlivá data a odhady:

pořadové číslo dat	B	C	D	E
1	0.917	0.636	1.000	0.470
2	0.934	0.734	1.000	0.747
3	0.937	0.757	1.000	0.831
4	0.952	0.896	1.000	1.000
5	0.955	0.924	1.000	1.000
6	0.956	0.936	1.000	1.000
7	0.956	0.939	1.000	1.000
8	0.961	0.960	1.000	1.000
9	0.961	0.965	1.000	1.000
10	0.960	0.910	1.000	1.000
11	0.957	0.855	1.000	0.849
12	0.953	0.800	1.000	0.659
13	0.411	0.231	0.326	0.036

Tento příklad z [7] byl úspěšně testován v MS MBÚ na IBM 370/135 s numericky shodnými výsledky. Hodnoty vah, NIT a BETA byly získány navíc v MS MBÚ pro zajímavosti a možnost lepšího porovnání.

Již při letmé kontrole dat vidíme, že poslední hodnota nám vypadá z rámce ostatních dat. Soud o tom, jak se s touto "těžkostí" jednotlivé metody vypořádaly, ponecháme čtenáři. Nicméně se domníváme, že robustní metody získaly plus. Uvědomujeme si, že šlo o umělý příklad. Proto soudíme, že je nutné dané metody (a možnosti ROBETHu) podrobit intenzivnímu zkoumání jak na simulovaných, tak na reálných datech. Pro tento účel je nyní hlavní vyrobit flexibilní hlavní programy.

Kdokoliv by se chtěl této práce zúčastnit, je srdečně vítán a stále platí pozvání ke spolupráci tak, jak bylo vysloveno v Načetíně.

LITERATURA:

1 Andrews D.F. & all: Robust Estimation of Location. Princeton University Press. Princeton 1972.

2 Dixon W.J.(ed.): Biomedical Computer Programs-BMDP, University of California Press, Los Angeles 1975.

3 Hanson R.J., Lawson C.L.: Solving least squares problems, Prentice Hall, Englewood Clif, 1974.

4 Huber P.J., Dutter R.: Numerical solutions of robust regression problems, COMPSTAT 1974, Physica-Verlag, Wien 1974.

5 a) Huber P.J.: Robust statistics: A review. AMS 43, 1041-67.

b) Huber P.J.: Robust confidence limits. Z.f.Wahr..., 1968.

6 J.M.Chambers: Comptutational Methods for data analysis, J.Wiley, 1977

7 a) Marazzi A.: ROBETH-a subroutine library for robust statistical procedures, COMPSTAT 80, Physica-Verlag

b) Marazzi A.: ROBETH-document No 2, Robust liner regression programs, Res.Rep.23, Fachgruppe für Statistik, ETH, Zürich, 80.

c) Marazzi A: ROBETH-document No3, Robust affine invariant covariances, Res.Rep.24, Fachgruppe für Statistik, ETH, Zürich, 80.