STEINERŮV PŘÍSTUP K ROBUSTNÍ REGRESI

Ivan Křivý

1. Úvod

Statistické metody jsou jen částečně založeny na pozorováních. Druhou neméně významnou složku představují apriorní předpoklady o skutečné situaci (předpoklady o náhodnosti, nezávislosti pozorování, typu rozdělení sledované náhodné veličiny apod.). Tyto předpoklady nemohou být striktně splněny, jsou v nejlepším případě pouhou aproximací (idealizací) skutečnosti. Odchylky od idealizovaného modelu jsou způsobeny např. výskytem hrubých chyb, jež se projevují přítomností odlehlých údajů v datovém souboru, zaokrouhlováním jednotlivých pozorování, kontaminací modelového rozdělení (zpravidla normálního) jinými rozděleními nebo porušením předpokladu o nezávislosti pozorování. V závislosti na velikosti těchto odchylek mohou statistické metody vytvořené pro idealizovaný model v reálné situaci podstatně ztratit na své účinnosti nebo dokonce zcela selhat. Proto se pozornost statistiků, ale i uživatelů statistických metod, soustřeďuje na hledání nových přístupů, které jsou málo citlivé na odchylky od idealizovaného modelu.

2. Steinerovy charakteristiky

Steiner [1] zavádí pro spojitou náhodnou veličinu s hustotou $f(x)$ dvě nové číselné charakteristiky: nejčastější hodnotu (most frequent value) M a dihezi c. Tyto veličiny se definují vztahy:

$$\mathrm{M}: \; H_c(\mathrm{M}) = \int_{-\infty}^{\infty} \frac{x - \mathrm{M}}{c^2 + (x - \mathrm{M})^2} f(x)\, dx = 0, \tag{1}$$

$$c: \; F_{\mathrm{M}}(c) = \int_{-\infty}^{\infty} \frac{c^{3/2}}{c^2 + (x - \mathrm{M})^2} f(x)\, dx = \operatorname*{Max}_{c}. \tag{2}$$

Hodnota M je tedy určena řešením rovnice $H_c(\mathrm{M}) = 0$, zatímco hodnota c polohou maxima funkce $F_{\mathrm{M}}(c)$ za podmínky (1). Funkce $F_{\mathrm{M}}(c)$ může mít několik maxim; v takovém případě se uvažuje

maximum s nejvyšší hodnotou c. Je-li funkce $F_M(c)$ diferencovatelná podle c ($c > 0$), používá se namísto podmínky (2) vztahu

$$F'_M(c) = \frac{c^{1/2}}{2} \int_{-\infty}^{\infty} \frac{3(x - M)^2 - c^2}{[c^2 + (x - M)^2]^2} f(x)\, dx = 0. \qquad (3)$$

Veličina M je zřejmě charakteristikou polohy, dheze c charakteristikou variability uvažované náhodné veličiny. V teoretických úvahách se někdy dává přednost veličině $\kappa = 1/c$, kterou Steiner označuje termínem koheze (míra spolehlivosti).

Otázkami existence a vlastností navržených charakteristik se podrobněji zabýval Csernyák [2]. Jeho výsledky lze shrnout do následujících vět.

a) Pro libovolné (pevně zvolené) konečné $c > 0$ existuje takové konečné M, jež splňuje rovnici (1).

b) Je-li hustota $f(x)$ spojitá v bodě $x = M$ a platí-li navíc $f(M) \neq 0$, pak existuje konečné $c > 0$ takové, že pro něj funkce $F_M(c)$ nabývá svého maxima.

c) Pro libovolné konečné M platí, že všechna řešení c rovnice (3) splňují nerovnost

$$c^2 \leq 3 \int_{-\infty}^{\infty} (x - M)^2 f(x)\, dx,$$

tj. jsou ohraničená.

Tvrzení a) je možno doplnit takto: Je-li hustota $f(x)$ navíc symetrická a unimodální (jednovrcholová), existuje právě jediné konečné M takové, že vyhovuje rovnici (1).

Zřejmou předností charakteristik M a c je skutečnost, že existují (nabývají konečných hodnot) i v případě takových abs. spojitých rozdělení, pro něž neexistuje rozptyl D nebo dokonce ani střední hodnota E (Cauchyho rozdělení nebo celá třída rozdělení s hustotou typu

$$f_\alpha(x) = \frac{\Gamma(\alpha/2)}{\sqrt{\pi}\,\Gamma\left(\frac{\alpha-1}{2}\right)} (1 + x^2)^{-\alpha/2} \quad \text{pro} \quad 1 < \alpha \leq 3.$$

Poznámka. Charakteristiky M a c lze pomocí Stieltjesova integrálu definovat i pro náh. veličiny diskrétního typu.

3. Steinerovy odhady a jejich robustnost

Odhady Steinerových charakteristik M a c budeme dále značit symboly $\hat{M}$, resp. $\hat{c}$. Pro tyto odhady (statistiky) platí (viz vztahy (1) a (3)):

$$\hat{M} = \frac{\sum_{j=1}^{N} \frac{x_i}{\hat{c}^2 + (x_j - \hat{M})^2}}{\sum_{j=1}^{N} \frac{1}{\hat{c}^2 + (x_j - \hat{M})^2}}, \tag{4}$$

$$\hat{c}^2 = \frac{3 \sum_{j=1}^{N} \frac{(x_j - \hat{M})^2}{[\hat{c}^2 + (x_j - \hat{M})^2]^2}}{\sum_{j=1}^{N} \frac{1}{[\hat{c}^2 + (x_j - \hat{M})^2]^2}}, \tag{5}$$

kde $x_1, x_2, \ldots, x_N$ představují realizaci nějakého náh. výběru o rozsahu N. Statistika $\hat{M}$ je evidentně váženým aritmetickým průměrem se statistickými vahami typu

$$w_j = \frac{c}{\hat{c}^2 + (x_j - \hat{M})^2}, \quad c > 0, \; j = 1, 2, \ldots, N.$$

Nejprve ukážeme, že statistiky $\hat{M}$, $\hat{c}$ jsou vlastně *M*-odhady polohy (location), resp. škálového parametru (scale). *M*-odhady (zobecněné maximálně věrohodné odhady) polohy T_N a škály S_N jsou podle [3] definovány soustavou rovnic

$$\sum_{j=1}^{N} \psi\left(\frac{x_j - T_N}{S_N}\right) = 0, \quad \sum_{j=1}^{N} \chi\left(\frac{x_j - T_N}{S_N}\right) = 0. \tag{6}$$

Je-li $F_N(x)$ empirická distribuční funkce odpovídající realizaci uvažovaného výběru, může být řešení T_N, S_N soustavy (6) zapsáno jako $T(F_N)$, $S(F_N)$, kde T, S jsou funkcionály dané soustavou

$$\int_{-\infty}^{\infty} \psi\left(\frac{x - T(F)}{S(F)}\right) dF(x) = 0, \quad \int_{-\infty}^{\infty} \chi\left(\frac{x - T(F)}{S(F)}\right) dF(x) = 0. \tag{7}$$

Předpokládáme-li, že rozdělení F přísluší hustota f, pak prosté srovnání vztahů (7) se vztahy (1) a (3) [nebo vztahů (6) se vztahy (4) a (5)] vede k závěru, že statistiky $\hat{M}$, $\hat{c}$ jsou skutečně M-odhady, přičemž

$$\psi(x) = \frac{x}{x^2 + 1}, \qquad \chi(x) = \frac{3x^2 - 1}{(x^2 + 1)^2}. \tag{8}$$

Obě funkce jsou spojité a ohraničené na množině R, přičemž $\lim_{x \to \pm\infty} \psi(x) = \lim_{x \to \pm\infty} \chi(x) = 0$. Funkce $\psi(x)$ je lichá, nabývá maximální (minimální) hodnoty 1/2 (-1/2) pro $x = 1$ ($x = -1$). Funkce $\chi(x)$ je sudá, maximum v bodě $x = \pm\sqrt{5/3}$ má hodnotu 9/16 a minimum pro $x = 0$ hodnotu -1.

Při posuzování robustnosti odhadů $\hat{M}$ a $\hat{c}$ vycházíme z koncepce založené na pojmu funkce vlivu IF (influence function). IF odhadu T pro rozdělení F je obecně definována vztahem

$$IF(x;\ T,\ F) = \lim_{t \to 0} \frac{T((1-t)F + t\delta(x)) - T(F)}{t},$$

a to pro všechna taková x, pro něž uvedená limita existuje. ($\delta(x)$ značí Heavisideovu funkci.) Tato funkce zřejmě charakterizuje vliv infinitezimální kontaminace rozdělení F poruchou lokalizovanou v bodě x.

Výrazy pro IF odhadů polohy polohy T a škály S určíme tak, že ve vztazích (7) nahradíme distribuci F rozdělením $(1-t)F + t\delta(x)$ a takto upravené vztahy derivujeme podle t v bodě $t = 0$. Za zjednodušujícího předpokladu, že rozdělení F má hustotu f symetrickou vzhledem k počátku, dostaneme pro IF odhadů $\hat{M}$ a $\hat{c}$ vztahy

$$IF(x;\ \hat{M},\ F) = \frac{x}{\hat{c}^2 + x^2} \frac{1}{\int_{-\infty}^{\infty} \frac{\hat{c}^2 - y^2}{(\hat{c}^2 + y^2)^2} f(y)\, dy}, \tag{9}$$

$$IF(x;\ \hat{c},\ F) = \frac{3x^2 - \hat{c}^2}{(x^2 + \hat{c}^2)^2} \frac{\hat{c}}{\int_{-\infty}^{\infty} \frac{10\hat{c}^2 - 6y^2}{(\hat{c}^2 + y^2)^2} y^2 f(y)\, dy}.$$

Pomocí IF lze zavést některé kvantitativní míry robustnosti příslušného odhadu T pro dané rozdělení F, a to zejména citlivost vůči hrubým chybám (gross-error sensitivity) $\gamma^*(T, F)$ a asymptotický rozptyl odhadu (asymptotic variance) $V(T, F)$. Tyto veličiny jsou obecně definovány takto:

$$\gamma^*(T, F) = \sup_x |IF(x; T, F)| ,$$

$$V(T, F) = \int_{-\infty}^{\infty} IF^2(x; T, F)\, dF(x) .$$

Lze snadno dokázat, že veličiny $\gamma^*(\hat{M}, F)$ i $\gamma^*(\hat{c}, F)$ mohou pro symetrická rozdělení nabývat pouze konečných kladných hodnot. (Integrály ve vztazích (9) jsou evidentně nenulové a funkce $x/(\hat{c}^2 + x^2)$ a $(3x^2 - \hat{c}^2)/(\hat{c}^2 + x^2)^2$ mají pro dané konečné $\hat{c} > 0$ kvalitativně stejný průběh jako funkce $\psi(x)$, resp. $\chi(x)$.). To znamená, že malá kontaminace rozdělení F nemůže mít (ani v nejhorším případě) katastrofální vliv na hodnoty odhadů $\hat{M}$ a $\hat{c}$. Odhady $\hat{M}$ i $\hat{c}$ je tedy možno považovat za robustní vzhledem ke zkreslení.

Pro asymptotický rozptyl $V(\hat{M}, F)$ odvozuje Czernyák [5] vztah

$$V(\hat{M}, F) = \frac{1}{\int_{-\infty}^{\infty} \frac{1}{\hat{c}^2 + x^2} f(x)\, dx} ,$$

z něhož vyplývá, že veličina $V(\hat{M}, F)$ je pro symetrická rozdělení a dané $\hat{c} > 0$ vždy konečná. Konečnost asymptotického rozptylu $V(\hat{M}, F)$ je dalším dokladem robustnosti odhadu $\hat{M}$.

4. Použití Steinerových odhadů v regresní analýze

Odhady $\hat{M}$, $\hat{c}$ se pro nějakou realizaci $x_1, x_2, \ldots, x_N$ jednorozměrného výběru určují simultánním řešením soustavy rovnic (4) a (5). Steiner [6] navrhuje algoritmus "dvojnásobné" iterace. Ve vnějším cyklu se počítají postupně iterace $\hat{M}$, ve vnitřním se upřesňují hodnoty diheze. Výchozí hodnoty pro iterace se volí takto:

$$\hat{M}: \quad \frac{1}{N}\sum_{j=1}^{N} x_j, \qquad \hat{c}: \quad \frac{\sqrt{3}}{2}(x_{max} - x_{min});$$

přitom x_{max}, x_{min} představují největší, resp. nejmenší, hodnotu realizace uvažovaného výběru.

Pro účely regresní analýzy je nutno uvedený postup modifikovat. Uvažujme např. exp. data ve formě posloupnosti uspořádaných r-tic $[y_j, x_{1j}, x_{2j}, \ldots, x_{nj}]$, kde $r = n + 1$, y_j jsou hodnoty náh. veličiny η a $x_{1j}, x_{2j}, \ldots, x_{nj}$ hodnoty fixních veličin $x_1, x_2, \ldots, x_n$ $(j = 1, 2, \ldots, N)$. Předpokládejme, že závislost η na skupině nezávisle proměnných $x_1, x_2, \ldots, x_n$ je vystižena regresní funkcí

$$\eta = \varphi(p; x_1, x_2, \ldots, x_n) ,$$

v níž $p = (p_1, p_2, \ldots, p_m)$ je vektor regresních koeficientů. Pak se ve vnějším cyklu provádí iteračně upřesnění odhadů regresních koeficientů $(\hat{p}_i \longrightarrow \hat{p}_{i+1})$ řešením úlohy

$$\sum_{j=1}^{N} w_{ji}\left[y_j - \varphi(\hat{p}_{i+1}; x_{1j}, x_{2j}, \ldots, x_{nj})\right]^2 = \text{Min} ,$$

v níž

$$w_{ji} = \frac{\hat{c}_i^2}{\hat{c}_i^2 + d_{ji}^2}, \qquad d_{ji} = y_j - \varphi(\hat{p}_i; x_{1j}, x_{2j}, \ldots, x_{nj}).$$

Přitom lze s výhodou použít standardního podprogramu pro metodu nejmenších čtverců se stat. vahami w_{ji}. Iterace ve vnitřním cyklu se provádějí podle vzorce

$$\hat{c}_{i,k+1}^2 = \frac{3\sum_{j=1}^{N} \frac{d_{ji}^2 \hat{c}_{ik}^4}{[\hat{c}_{ik}^2 + d_{ji}^2]^2}}{\sum_{j=1}^{N} \frac{\hat{c}_{ik}^4}{[\hat{c}_{ik}^2 + d_{ji}^2]^2}} .$$

kde $\hat{c}_{ik}$ představuje hodnotu diheze v i-tém vnějším cyklu a v k-tém cyklu vnitřním. Složky $\hat{p}_{11}, \hat{p}_{21}, \ldots, \hat{p}_{m1}$ vektoru výchozích regresních koeficientů se určí metodou nejmenších čtverců bez stat. vah; pro výchozí hodnotu diheze se doporučuje volba

$$\frac{\sqrt{3}}{2}\,(d_{max,1} - d_{min,1}) ,$$

kde $d_{max,1}$, $d_{min,1}$ značí největší, resp. nejmenší, z hodnot d_{j1}.

Na závěr uvádíme ukázku rezistence navrženého postupu vůči odlehlým údajům. Uvažujeme jednoduchou lineární regresi s následujícími daty, jež obsahují 20 % odlehlých údajů (hodnot náh. veličiny η).

x_j	10	20	30	40	50	60	70	80	90	100
y_j	21	29	45	45	62	68	81	89	1000	1000

Výpočty realizujeme s pevným počtem šesti iterací ve vnitřním cyklu. Ukazuje se, že odhady směrnice regresní přímky nejsou již po sedmém iteračním kroku prakticky zkresleny odlehlými hodnotami nezávisle proměnné. Steiner se svými spolupracovníky uvádí celou řadu příkladů úspěšného použití M-fitting při analýze exp. dat z geofyzikálního výzkumu.

Odhady regresních koeficientů s využitím Steinerova přístupu je možno plným právem pokládat za robustní z hlediska koncepce založené na IF (viz [4]). Jejich robustnost je výrazné v těch případech, kdy se provádí regrese náh. veličiny na veličinách nenáhodné povahy.

5. Závěr

Tento příspěvek je věnován dvěma novým (Steinerem zavedeným) číselným charakteristikám náh. veličiny: nejčastější hodnotě M a dihezi c. Ukazuje se, že odpovídající statistiky $\hat{M}$ a $\hat{c}$ (odhady M a c) je možno považovat za tzv. *M*-odhady s výraznými robustními vlastnostmi. Obě statistiky jsou především robustní vzhledem ke zkreslení a veličina $\hat{M}$ vykazuje vždy konečný asymptotický rozptyl. Steinerova přístupu se pak využívá k návrhu algoritmu pro robustní odhad regresních koeficientů. Popsaný algoritmus lze doporučit pro zpracování exp. dat zejména tam, kde se sleduje závislost náh. veličiny na jedné či několika fixních veličinách.

Literatura

[1] Steiner, F.: Most frequent value and cohesion of probability distributions. The most frequent value. Introduction to a modern conception of statistics. Ed. F. Steiner. Budapest, Akadémiai Kiadó 1991 (dále MFV), pp. 17-35.

[2] Czernyák, L.: Investigations concerning the existence and determination of the most frequent value and c. MFV, pp. 209-217.

[3] Huber, P.J.: Robust Statistics. New York, John Wiley & Sons, 1981.

[4] Hampel, F.R. - Ronchetti,E.M. - Rousseeuw,P.J. - Stahel,W.A.: Robust Statistics. The Approach Based on Influence Function. New York, John Wiley & Sons 1986.

[5] Czernyák, L.: The most frequent value as a robust estimate. MFV, App. VI, pp. 253-257.

[6] Steiner, F.: Introductory instructions for the computation of the most frequent value of a series of data. MFV, App. VIII, pp. 263-270.