Vlivné body v lineární regresi

Jiří Militký, VÚTZ, Dvůr Králové n.L.

1. Úvod

Je známo, že klasická metoda nejmenších čtverců je silně nerobustní. Je tedy velmi důležité identifikovat vlivné body a omezit jejich působení na odhady parametrů resp. jejich statistické charakteristiky.

V této práci jsou na jednoduchém simulačním experimentu demonstrovány nerobustní vlastnosti metody nejmenších čtverců a porovnány různé postupy pro identifikaci resp. eliminaci vlivných měření (při odhadu regresních parametrů).

2. Základní pojmy

Vyjděme ze standardní situace, kdy máme k dispozici n-tici bodů, $\{y_i, \underline{x}_i\}_{i=1,\dots,n}$ kde vektory deterministických vysvětlujících proměnných mají složky $\underline{x}_i^T = (1, x_{i1}, \dots, x_{im})$. Nechť dále platí aditivní model měření

$$y_i = \underline{x}_i^T \underline{a} + \varepsilon_i \qquad (1)$$

kde $\underline{a}^T = (a_1, \dots, a_m)$ jsou regresní parametry. Pak za předpokladu, že náhodné chyby $\underline{\varepsilon}^T = (\varepsilon_1, \dots, \varepsilon_n)$ mají

a) nulovou střední hodnotu $E(\underline{\varepsilon}) = \underline{0}$

b) diagonální kovariační matici $E(\underline{\varepsilon} \cdot \underline{\varepsilon}^T) = \sigma^2 \mathbf{E}_n \quad (\sigma^2 < \infty)$

c) stejné rozdělení, takže $f(\underline{\varepsilon}) = \prod_{i=1}^{n} f(\varepsilon_i)$

a navíc, že

d) hodnost matice $\mathbf{X}$ je právě m

e) na parametry $\underline{a}$ nejsou kladena omezení

f) platí model (2)

kde nejlepší nestranné lineární odhady $\hat{\underline{a}}$ parametrů $\underline{a}$ získat minimalizací kritéria nejmenších čtverců odchylek

$$S(\underline{a}) = \| \underline{y} - \mathbf{X}\underline{a} \|_2^2 \qquad (2)$$

kde $\mathbf{X}^T = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$ je (m x n) matice. Z rov. (2) je zřejmé, že vektor

$$\hat{\underline{y}} = \mathbf{X}\hat{\underline{a}} \qquad (3)$$

kde $\hat{\underline{a}}$ minimalizuje rov. (2) je kolmá projekce vektoru $\underline{y}$ do roviny vymezené sloupci matice $\mathbf{X}$. Z toho také plyne, že vektor reziduí

$$\hat{\underline{e}} = \underline{y} - \hat{\underline{y}} \qquad (4)$$

musí být kolmý na sloupce matice $\mathbf{X}$ a tedy

$$\mathbf{X}^T \hat{\underline{e}} = \underline{0} \qquad \text{resp.} \qquad \mathbf{X}^T (\underline{y} - \mathbf{X}\hat{\underline{a}}) = \underline{0} \qquad (5)$$

Z rov. (5) již přímo plyne, že $\hat{\underline{a}}$ minimalizující kritérium $S(\underline{a})$ lze vyjádřit ve tvaru

$$\hat{\underline{a}} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \underline{y} \qquad (5')$$

Dosadíme-li z rov. (5') do rov. (3)dostaneme

$$\hat{\underline{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \underline{y} = \mathbf{H}\underline{y} \qquad (6)$$

kde $\mathbf{H}$ (n x n) je zřejmě projekční matice. Pokud je $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \mathbf{E}_n)$, jsou odhady $\hat{\underline{a}}$ z rov. (5) maximálně věrohodné.

Snadno lze dokázat, že kovariační matice odhadů $\hat{\underline{a}}$ má tvar

$$C(\hat{\underline{a}}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} \qquad (7)$$

kde reziduální rozptyl σ^2 lze nahradit odhadem $\hat{\sigma}^2 = \| \hat{\underline{e}} \|_2^2 / (n-m)$.

Z rov. (6) plyne, že kovarianční matice predikce je rovna

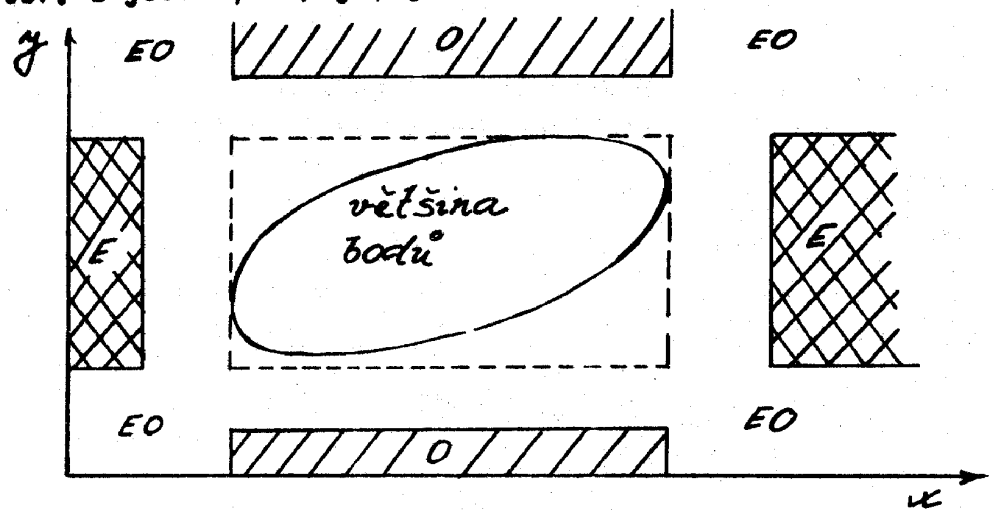
$$C(\hat{\underline{y}}) = \sigma^2 H \qquad (8)$$

a z rov. (4) při dosazení z rov. (6) lze určit kovarianční matici reziduí

$$C(\hat{\underline{e}}) = \sigma^2 (E_n - H) \qquad (9)$$

Platí, že odhady $\hat{a}_1, \hat{\underline{y}}$ a jejich statistické charakteristiky jsou citlivé na
- vybočující body (*outliers*) - označení O
- extrémní body (remote points) - označení E

Na obr. 1 jsou tyto pojmy graficky znázorněny



Obr. 1. Rozdíl mezi oblastmi kde leží vybočující pozorování ///// a extrémní body XXXX. EO označující vlivné body, které jsou jak vybočující, tak i extrémní.

3. Simulační experiment

Pro účely této práce je použito simulovaných dat (y_i, x_i) $i = 1, \ldots, 50$. Vysvětlující proměnná x_i byla generována z rovnoměrného rozdělení R (0, 1) a transformována do intervalu $\langle 10, 20 \rangle$. Odpovídající hodnoty y_i byly určeny ze vztahu

$$y_i = 1 \cdot x_i + 1 + N(0,1)$$

kde $N(0,1)$ je realizace náhodné veličiny s normalizovaným normálním rozdělením.

Do takto získané n-tice bodů se zaváděly vybočující (O) resp. extrémní (E) hodnoty o různých velikostech A pro bod s indexem j = 37 resp. body s indexy j = 12, 25, 37.

Pro ilustraci se dále v tabulkách většinou uvádějí odhady směrnice $\hat{a}_1$ regresní přímky určené různými postupy z takto generovaných dat.

V tab. 1 jsou uvedeny odhady směrnice $\hat{a}_1$ určené metodou nejmenších čtverců pro případ, že se měnily souřadnice pouze bodu j = 37. (Uveďme, že původní hodnoty tohoto bodu jsou $x_{37} = 19{,}2$ a $y_{37} = 20{,}8$.

Tab. 1 : Odhady $\hat{a}_1$ získané MNČ pro data s jedním vlivným měřením (O resp. E o velikostech A)

A	E	O
20	0,968	0,967
30	0,71	1,07
40	0,46	1,18
50	0,31	1,28
60	0,22	1,39
70	0,17	1,49
80	0,13	1,602

Z tab. 1 je patrné, že přítomnost extrémů E vede k výrazně horším výsledkům než přítomnost vybočujících pozorování. (Klasická analýza reziduí obyčejně E nezjistí).

4. Identifikace vlivných bodů

Jako vlivné se obecně označují takové body, které výrazně ovlivňují výsledky regrese.

Poznámka :
Ve většině postupů identifikace vlivných měření se využívá diagonálních prvků h_{ii} projekční matice H. Tyto prvky mají řadu zajímavých vlastností plynoucích z faktu, že H je idempotentní symetrická matice. Platí, že /1/

a) $0 \leq h_{ii} \leq 1$

b) $\sum_{(i)} h_{ii} = m$ (pokud je hodnost matice rovna m)

c) průměrná hodnota $h_{ii} \sim m/n$

d) lze psát, že $h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$. Z toho plyne, že pro $h_{ii} = 1$ resp. $h_{ii} = 0$ jsou vždy všechna $h_{ij} (j \neq i)$ také nulová.

e) pokud mají vysvětlující proměnné vícerozměrné normální rozdělení, má veličina
$$F = (n-m)[h_{ii} - (1/n)] \,/\, [(1 - h_{ii})(m-1)]$$
Fisherovo rozdělení s (m - 1) a (n - m) stupni volnosti.

f) Čím je h_{ii} větší, tím více ovlivňuje bod $[x_i, y_i]$ predikci $\hat{y}_i$. To plyne přímo z rov. (6), kterou lze pro $\hat{y}_i$ vyjádřit ve tvaru

$$\hat{y}_i = y_i h_{ii} + \sum_{j \neq i}^{n} h_{ij} y_j \qquad (10)$$

Je zřejmé, že pro $h_{ii} \to 1$ jsou $h_{ij} \to 0$ a $\hat{y}_i \sim y_i$ (tedy veškerá variabilita v místě x_i^T je objasněna regresním modelem).

g) V práci /2/ je dokázáno, že pro centrované proměnné x_{ci}^T je h_{cii} velké, pokud je $x_{ci} x_{ci}$ velké (to je případ extrémních bodů) resp. pokud je x_{ci} ve směru vlastního vektoru odpovídajícího malému vlastnímu číslu kovariační matice vysvětlujících proměnných. Malé h_{cii} znamená vždy, že $x_{ci} x_{ci}$ je malé (x_{ci} leží blízko "těžiště" vysvětlujících proměnných).

h) Z rov. (8) a rov. (9) přímo plyne, že $\mathrm{var}(\hat{y}_i) = h_{ii} \sigma^2$ a $\mathrm{var}(\hat{e}_i) = \sigma^2 (1 - h_{ii})$. Pro extrémní body (s velkým h_{ii}) bude tedy rozptyl reziduí malý, takže bude obtížné na základě analýzy reziduí identifikovat vybočující měření.

i) Pokud $h_{ii} \to 1$, nejsou klasické metody robustní regrese efektivní.

Z těchto vlastností je patrné, že h_{ii} mohou přímo identifikovat extrémní body.

Speciálně pro případy, kdy je účelem identifikovat vybočující body, využívá se různých typů reziduí.

Klasická nenormalizovaná rezidua $\hat{a}_i$ jsou definována vztahy

$$\hat{a}_i = y_i - x_i^T \hat{a} \qquad i = 1, \ldots, n \qquad (11)$$

Z rov. (4) a (6) přímo plyne, že

$$\hat{e} = (E_n - H) y = (E_n - H) e \qquad (12)$$

Platí tedy, že každé reziduum je lineární kombinací náhodných chyb $e_1, \ldots, e_n$. (Proto se u malých výběrů projevuje efekt supernormality, t.j. rozdělení $\hat{e}$ je přibližně normální, i když rozdělení e normální není).

Z rov. (9) plyne, že rezidua $\hat{e}$ jsou vzájemně korelované a robustním rozptylem. Také tzv. normalizovaná rezidua

$$\hat{e}_N = (n-m)\hat{e} \,/\, \|\hat{e}\|_2^2 \qquad (13)$$

zůstávají nadále korelované a mají nekonstantní rozptyl $\mathrm{var}(\hat{e}_N) = (E_n - H)$. Znestejnění rozptylu lze docílit podělením $\hat{e}$ odpovídajícími rozptyly z rov. (8). Vyjdou tzv. Studentizované rezidua

$$\hat{r}_i = \frac{\hat{a}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \qquad i = 1, \ldots, n \qquad (14)$$

Veličiny $\hat{r}_i$ již mají nulovou střední hodnotu a jednotkový rozptyl. Navíc platí, že $\hat{r}_i^2/(n-m)$ má beta rozdělení.

Mnohonásobnými transformacemi reziduí $\hat{r}_i$ jsou tzv. Jackknife rezidua

$$\hat{t}_i = \hat{r}_i \sqrt{\frac{n-m-1}{n-m-\hat{r}_i^2}} \qquad i = 1,..,n \tag{15}$$

Platí, že $\hat{t}_i$ jsou testační statistiky pro test hypotézy $H_0 : \alpha = 0$ v modelu jednoduchého modelu vybočujícího pozorování vychýleného co do polohy

$$\underline{y} = X\underline{a} + \underline{z}_i \alpha + \underline{\varepsilon} \tag{16}$$

kde vektor $\underline{z}_i$ má i-tou souřadnici rovnou jedné a ostatní nulové. Pokud platí H_0, má $\hat{t}_i$ Studentovo rozdělení s n - m - 1 stupni volnosti.

Poznámka :
Pokud se zavede reziduální rozptyl $\hat{\sigma}^2_{(-i)}$ počítaný vždy bez i-tého bodu, pro který platí

$$\hat{\sigma}^2_{(-i)} = \frac{(n-m)\hat{\sigma}^2 - \hat{a}_i^2/(1-h_{ii})}{n-m-1} = \hat{\sigma}^2\left(\frac{n-m-\hat{r}_i^2}{n-m-1}\right) \tag{17}$$

můžeme vyjádřit Jackknife rezidua analogicky jako Studentizované rezidua. Platí

$$\hat{t}_i = \frac{\hat{a}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \tag{17a}$$

V mnoha aplikacích je výhodné použít predikované rezidua

$$\hat{a}_{(-i)} = y_i - \underline{x}_i^T \hat{\underline{a}}_{(-i)} \qquad i = 1,..,n \tag{18}$$

kde $\hat{\underline{a}}_{(-i)}$ jsou odhady získané metodou nejmenších čtverců při vynechání i-tého bodu (y_i, $\underline{x}_i^T$). Lze snadno ukázat, že $\hat{a}_{(i)}$ jsou odhady $\hat{\alpha}$ parametru α v modelu (16) získané metodou nejmenších čtverců. Z výpočetního hlediska je výhodné použít pro jejich určení vztahu

$$\hat{a}_{(-i)} = \hat{a}_i / (1-h_{ii}) = \hat{\alpha} \tag{19}$$

Platí, že pokud ε_i mají normální rozdělení $N(0,\sigma^2)$, mají $\hat{a}_{(-i)}$ také normální rozdělení $N(0,\sigma^2/(1-h_{ii}))$ a jsou stejně korelované jako ε_i.
V tabulce 2 jsou pro data ze simulačního experimentu (kap.3) s jedním vlivným měřením (O resp. E)o velikosti A v bodě i = 37 určeny h_{ii}, $\hat{a}_{Ni}$, $\hat{r}_i$, $\hat{t}_i$, $\hat{a}_{(-i)}$.
Tabulka 2 : Různé typy reziduí a h_{ii} pro data s jedním vlivným měřením (O,E) o velikostech A.

	O					E				
A	$\hat{a}_{Ni}$	$\hat{r}_i$	$\hat{t}_i$	$\hat{a}_{(-i)}$	h_{ii}	$\hat{a}_{Ni}$	$\hat{r}_i$	$\hat{t}_i$	$\hat{a}_{(-i)}$	h_{ii}
20	-0,41	-0,43	-0,42	-0,41	0,06	-0,384	-0,4	-0,396	-0,39	0,078
40	6,36	6,56	20,1	19,6	0,06	-3,64	-6,08	-12,6	-19,8	0,641
60	6,62	6,83	40,6	39,6	0,06	-2,44	-6,36	-15,9	-39,3	0,853
80	6,67	6,88	61,1	59,6	0,06	-1,77	-6,43	-17,1	-58,7	0,924

Z tab. 2 plyne, že :
- a_{Ni} vyhovují (alespoň částečně) pro případ vybočujících pozorování. Pro případ extrémních pozorování nejsou vhodné pro identifikaci vlivných bodů (s růstem A jejich hodnota klesá)
- ostatní typy reziduí $\hat{r}_i$, $\hat{t}_i$, $\hat{a}_{(-i)}$ jsou vhodné pro identifikaci obou typů vlivných bodů. Vyjdeme-li ze skutečnosti, že původní hodnota $Y_{37} = 20,8$ jsou $\hat{a}_{(i)}$ poměrně přesné odhady ($A - Y_{37}$).

- veličina h_{ii} dobře identifikuje extrémní body. (Za extrémní se orientačně považuje bod, pro který je $h_{ii} > \frac{2m}{n}$, což je pro tento případ 0,08).

Poznámka :
Všechny výše uvedené druhy reziduí jsou stejně korelované. Nekorelovaných reziduí je pouze (n - m). Patří mezi ně např. BLUS rezidua resp. rekurzivní rezidua /2,3/.

Pro identifikaci vlivných bodů je možno použít různých modifikací vlivové křivky, která je obecně definována např. v práci /4/.

Empirická vlivová křivka (vektor) $\underline{EIC}_i$ je definována vztahem /2/

$$\underline{EIC}_i = n(X^TX)^{-1}\underline{x}_i\hat{e}_i \qquad (20)$$

Pro Jackknife vlivovou křivku (počítanou bez i-tého bodu) platí /2/

$$\underline{EIF}_i = (n-1)(X^TX)^{-1}\underline{x}_i\hat{a}_i / (1-h_{ii})^2 \qquad (20a)$$

Pro konečné výběry je možno definovat také výběrovou vlivovou křivku /5/

$$\underline{SIC}_i = (n-1)(X^TX)^{-1}\underline{x}_i\hat{a}_i / (1-h_{ii}) \qquad (20b)$$

Je zřejmé, že rozdíl mezi těmito vlivovými křivkami je pouze v mocnině členu $(1-h_{ii})$ ve jmenovateli (která určuje citlivost dané vlivové křivky na přítomnost extrémních bodů).

S využitím známého vztahu

$$\hat{\underline{a}}_{(-i)} = \hat{\underline{a}} - (X^TX)^{-1}\underline{x}_i\hat{e}_i / (1-h_{ii}) \qquad (21)$$

lze snadno ukázat, že výběrová vlivová křivka $\underline{SIC}_i$ je (až na konstantu n - 1) rovna rozdílu $\hat{\underline{a}} - \hat{\underline{a}}_{(-i)}$.

Poznámka :
Také řada dalších diagnostických veličin pro identifikaci vlivných bodů porovnává statistické charakteristiky pro všechna data a pro data bez i-tého bodu.

Pro charakterizaci vzdálenosti mezi $\hat{\underline{a}}$ a $\hat{\underline{a}}_{(-i)}$ je vhodné použít Cookovu statistiku

$$D_i = \frac{(\hat{\underline{a}}_{(-i)} - \hat{\underline{a}})(X^TX)(\hat{\underline{a}}_{(-i)} - \hat{\underline{a}})}{m\hat{\sigma}^2} = \frac{1}{m}\cdot\hat{r}_i^2\cdot\frac{h_{ii}}{1-h_{ii}} \qquad (22)$$

která má F - rozdělení s m a n-m stupni volnosti. (Z rov. (22) plyne, že jde o analogii známého F-testu pro lineární regresní modely).

Atkinson /6/ navrhuje místo D_i použít modifikovanou funkci

$$T_i = \sqrt{\frac{(n-m)\,h_{ii}}{m\,(1-h_{ii})}}\,|\hat{t}_i| \qquad (23)$$

která je citlivější na vybočující měření i extrémní body (Velká T_i identifikují vlivné body).

Pro vyjádření vlivu i-tého bodu na predikci lze použít rozdílu

$$F_i = \hat{y}_i - \hat{y}_{(-i)} = h_{ii}\hat{a}_i / (1-h_{ii}) \qquad (24a)$$

resp. jako standardizované verze

$$DF_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{\hat{\sigma}^2_{(-i)}h_{ii}} = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\cdot\hat{t}_i \qquad (24)$$

Orientačně platí, že vlivný bod má DF_i vyšší než $2\sqrt{m/n}$.

Pro vyjádření vlivu i-tého bodu na změnu kovariančních matice odhadů lze využít determinantů (objemů konfidenčních oblastí).

Tedy

$$COV_i = det\left[\hat{\sigma}^2_{(-i)}\left(X_{(i)}^T X_{(i)}\right)^{-1}\right] / det\left[\hat{\sigma}^2 (X^T X)^{-1}\right] = \left[\frac{\hat{\sigma}^2_{(-i)}}{\hat{\sigma}^2}\right]^m \cdot \frac{1}{1-h_{ii}} \qquad (25)$$

V rov. (25) je $X_{(i)}$ matice, ve které chybí i-tý řádek $\underline{x}_i^T$. Pokud vyjde $|COV_i - 1| > 3m/n$, považuje se i-tý bod za významný.

Jednoduchou možností indikace vlivných bodů je uvažovat rozšířenou matici $Z = (X \mid \underline{y})$, jejíž řádky lze (pro stochastické vysvětlující proměnné) chápat jako m+1 rozměrné pozorování. Vybočující pozorování $\underline{z}_i^T$ se indikuje na základě rozdílu mezi tímto pozorováním a aritmetickým průměrem ostatních pomocí Wilksovy statistiky

$$\lambda_i = [n/(n-1)] \cdot (1-h_{ii}) \left[1 + \frac{\hat{t}_i}{n-m-1}\right]^{-1} \qquad (26)$$

Pokud je matice X výběrem z vícerozměrného Gaussova rozdělení, má veličina

$$w_i = (1-\lambda_i)(n-m-1)[m \cdot \lambda_i]^{-1} \qquad (27)$$

Fisherovo rozdělení s m a n - m stupni volnosti. Další druhy charakteristik vlivných bodů lze nalézt v /1 - 3/.

V tabulce 3 jsou pro stejná data jako v tab. 2 určeny charakteristiky D_i, T_i, DF_i, COV_i, a W_i.

Tabulka 3 : Různé charakteristiky vlivných bodů pro data s jedním vlivným měřením (O, E) o velikostech A

	O					E				
A	D_i	T_i	DF_i	COV_i	W_i	D_i	T_i	DF_i	COV_i	W_i
20	6.10^{-3}	0,52	-0,11	1,1	0,78	$6,7.10^{-3}$	0,562	-0,12	1,12	1,25
40	1,38	24,9	5,08	0,012	11,5	33	82,3	-16,8	0,15	23,5
60	1,5	50,3	10,3	$8,5.10^{-4}$	22,2	118	188	-38,4	0,17	80,2
80	1,52	75,8	15,5	$1,7.10^{-4}$	32,9	251	292	-59,7	0,26	169

Z tabulky 3 plyne, že :

- D_i nejsou vhodné pro identifikaci vybočujících měření (pro n = 50 m = 2 je $F_{95}(50,48) = 1,61$). Na extrémní body jsou D_i dostatečně citlivé.
- statistiky T_i a W_i mohou identifikovat jak vybočující, tak i extrémní body
- kritická hodnota pro DF_i je $\pm$ 0,4. Tedy i tato statistika indikuje oba typy vlivných bodů
- kritická hodnota pro $|COV_i - 1|$ je $\sim 0,12$.

  Tedy pouze pro A = 20 (kde nedošlo k prakticky žádnému zkreslení) nejsou nalezeny vlivné body (obou typů).

Pro rychlé identifikační účely je tedy výhodné použít místo běžně doporučovaných D_i modifikované statistiky T_i.

5. Odhady parametrů v přítomnosti vlivných bodů

Pro odhady parametrů, které jsou méně citlivé na přítomnost vlivných bodů se používá různých variant robustní regrese.

Klasické metody robustní regrese vedoucí na M - odhady parametrů nahrazují čtverce odchylek méně rychle rostoucí funkcí. Odpovídající kritérium regrese má tvar

$$S_M(\underline{a}) = \sum_{i=1}^{n} \rho(d_i) = \sum_{i=1}^{n} \rho\left[(y_i - \underline{x}_i^T \underline{a})^2 / \sigma^2\right] \qquad (28)$$

Minimalizace $S_M(\underline{a})$ je ekvivalentní řešení soustavy rovnic

$$\sum_{i=1}^{n} x_{ij}\, \psi(d_i) = 0 \qquad (29)$$

kde $\psi(d_i) = \rho'(d_i)$.

Poznámka :
Pro odhady parametrů $\hat{\underline{a}}_M$ je možno využít faktu, že rov. (29) lze při zavedení vah $w_i = \psi(d_i)/d_i$ převést na soustavu normálních rovnic vážené metody nejmenších čtverců. Postačuje tedy iterativně počítat odhady parametrů $\hat{\underline{a}}_M$ pomocí vážené MNČ s vahami určenými vždy z předchozí iterace.

V této práci je zvolena robustní Krasherova funkce

$$\rho(d) = \frac{w^2}{2}\left[1 - \exp\left(-\left(\frac{d}{w}\right)^2\right)\right] \qquad (30)$$

Při volbě W = 2,985 má v případě, že data mají normální rozdělení, robustní regrese s funkcí (30) 95 %ní efektivnost.

Ve třetím a čtvrtém sloupci tab. 4 jsou pro simulační model s jedním vlivným bodem o velikosti A (pro i = 37) uvedeny odhady směrnice regresní přímky $\hat{a}_{M1}$ minimalizující rov. (28) při $\rho(d)$ definovaném rov. (30).

Tabulka 4 : Směrnice regresní přímky $\hat{a}_1$ určené různými robustními metodami (Krasher pro $\rho(d)$ z rov. (30) resp. Krasher - Welsch se speciálními vahami).

Metoda	Krasher		Krasher - Welsch	
A	E	O	E	O
40	0,926	0,973	0,958	0,962
70	0,2033	0,969	0,965	0,959
80	0,1401	0,969	0,965	0,959

Z tab. 4 je patrné, že klasické postupy robustní regrese jsou málo robustní vůči extrémním bodům (viz 2. sl. tab. 4)

Robustnost vůči vybočujícím měřením je velmi dobrá.

Eliminace extrémních bodů se dá provést pomocí speciálních vah w_{Si}, které souvisejí s velkými hustotami vlivných bodů.

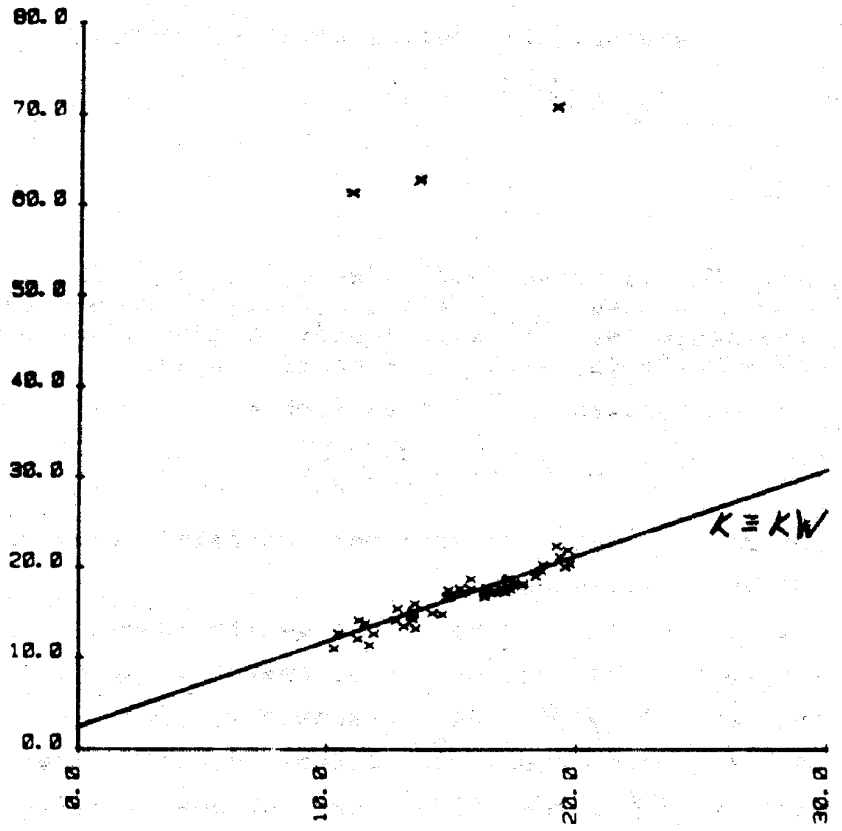
V této práci se využívé jednoduchých vah doporučených Krasherem a Welschem /8/.

$$w_{Si} = (1 - h_{ii}) / \sqrt{h_{ii}} \qquad (31)$$

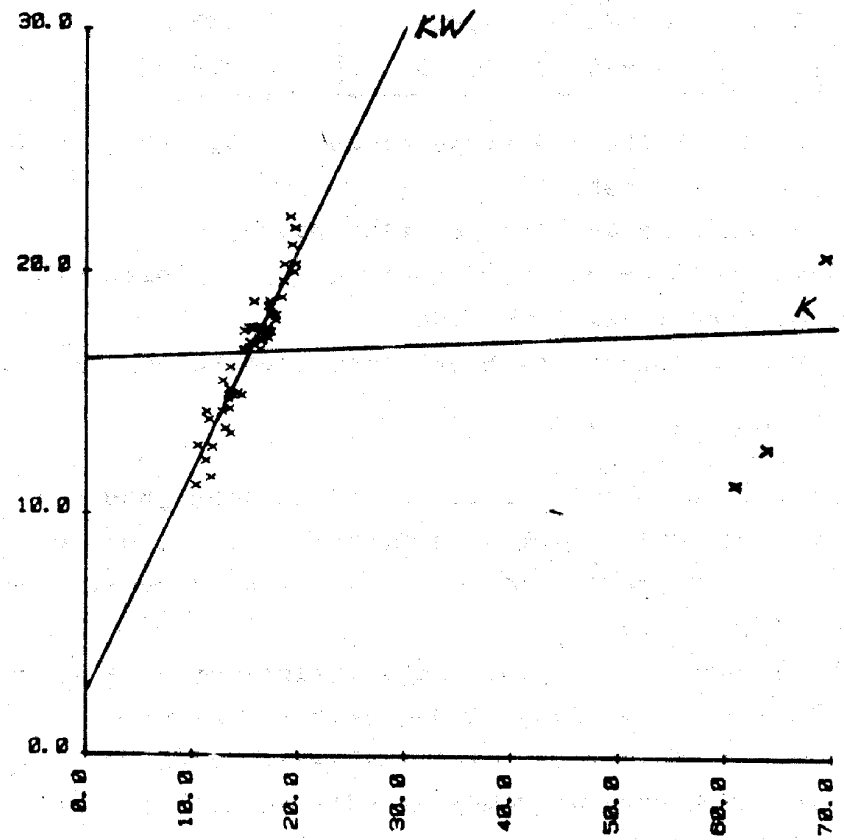
Ostatní způsoby volby vah eliminujících extrémní body jsou uvedeny např. v /7 - 9/. Obecné zavedení vah w_{Si} nijak nekomplikuje výpočty, protože stačí zavést globální váhy $w_{Ci} = w_{Si}\psi(d_i)/d_i$ a lze opět použít iterativně váženou metodu nejmenších čtverců /9/.

V pátém a šestém sloupci tab. 4 jsou pro simulační model s jedním vlivným bodem o velikosti A (pro i = 37) uvedeny odhady směrnice regresní přímky $\hat{a}_{M1}$ určené při použití vah dle (31) pro korekci extrémních bodů. Je zřejmé, že teprve pro tento případ je výsledek regrese robustní vůči oběma typům vlivných bodů.

Pro ilustraci rozdílů mezi klasickou robustní regresí a regresí robustní i vůči extrémním bodům jsou na obr. 2 a obr. 3 odhady regresní přímky určené oběma způsoby pro případ, kdy jsou v simulačním modelu tři vlivné body o velikosti A = 50 pro i = 18, 25 a 37.



Obr. 2 Regresní přímky (pro data se třemi vybočujícími body) počítané dle Krashera (K) resp. Krashera - Welsche (KW).



Obr. 3 Regresní přímky (pro data se třemi extrémními body) počítané dle Krashera (K) resp. Krashera - Welsche (KW)

Opět je jasně patrné, že klasické robustní techniky (Krasherova) jsou robustní vůči vybočujícím měřením a nikoliv extrémním bodům.

Pro automatickou eliminaci vybočujících měření byla navržena také technika "ořezané" regrese, kde se ořezání provádí s využitím regresních kvantilů. To znamená, že se nejdříve vyloučí body ležící mimo pás vymezený regresními kvantily pro $\alpha = 0{,}1$ a $\alpha = 0{,}9$ a pak se pro zbylé body použije standardní metoda nejmenších čtverců.

Poznámka :

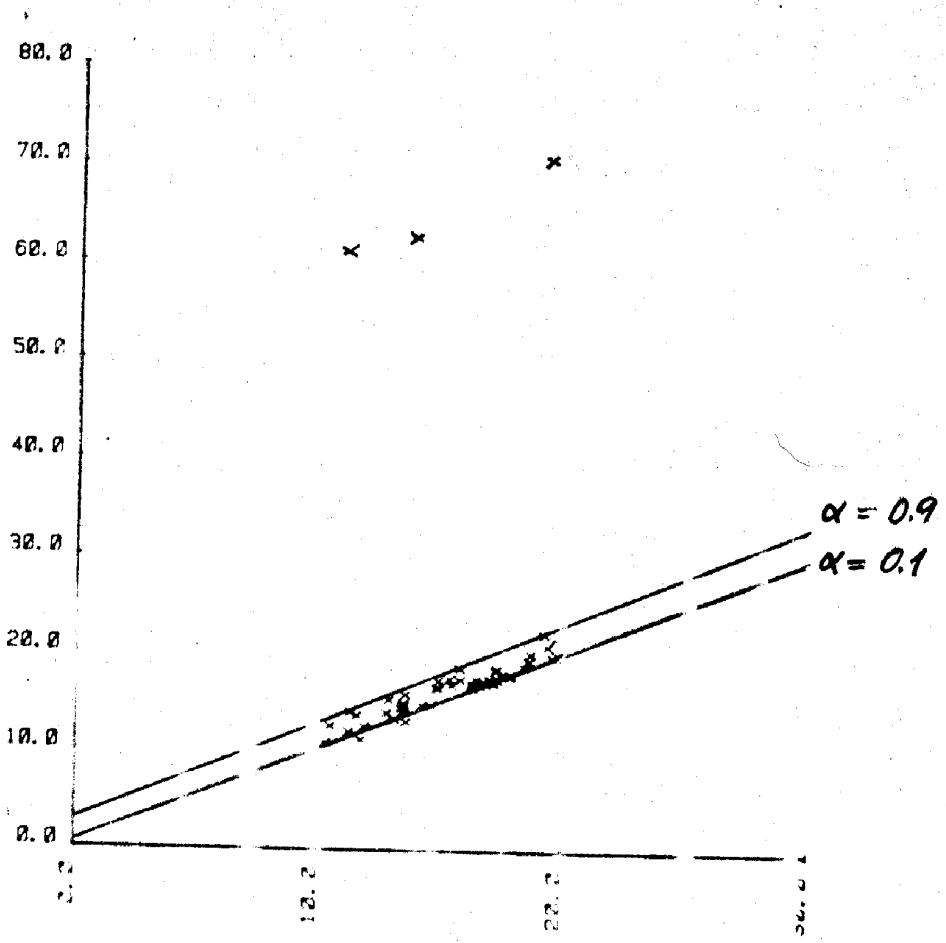
Uvědomme, že α - regresní kvantil ($0 < \alpha < 1$) je regresní model $\hat{y} = x\hat{a}_\alpha$, kde pro odhad $\hat{a}_\alpha$ se minimalizuje vztah

$$\hat{a}_\alpha = \min_{a} \left[\alpha \sum_{i \in (y_i \geq x_i^T a)} |a_i| + (1-\alpha) \sum_{i \in (y_i < x_i^T a)} |a_i| \right] \qquad (32)$$

kde $a_i = y_i - x_i^T a$.

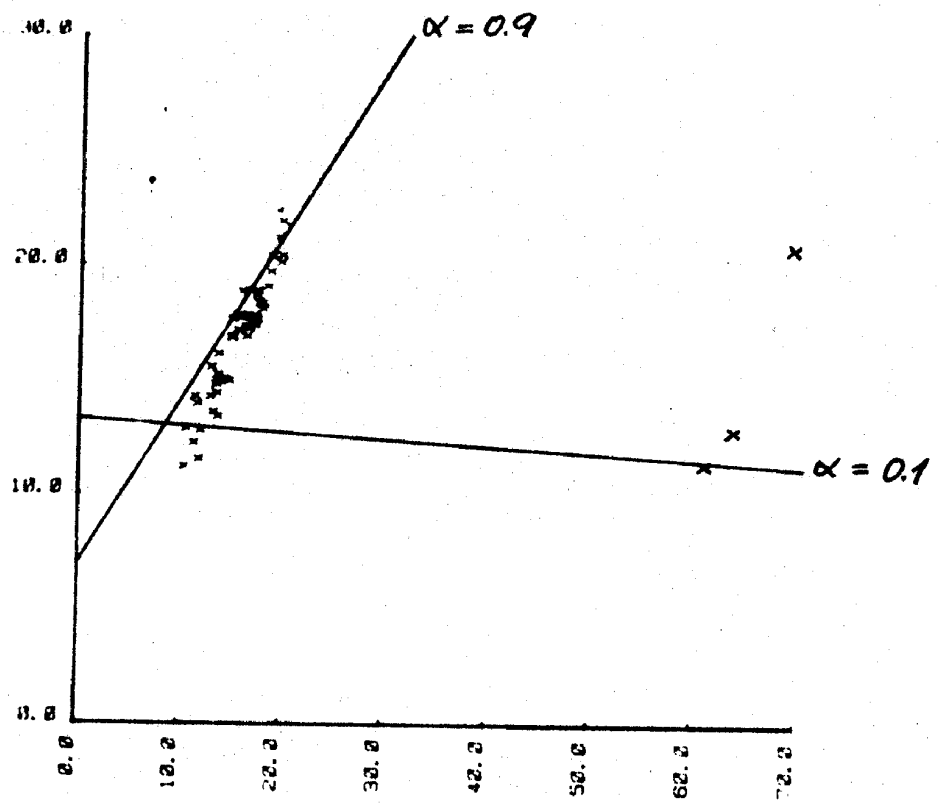
Pro α - regresní kvantil platí, že
- nejvíce $n-\alpha$ bodů je pod a nejméně $(n-m)\cdot\alpha$ bodů je nad modelem $x\hat{a}_\alpha$
- pro homoskedastická normální rozdělená data jsou kvantilové čáry (pro α nepříliš blízké 0 resp. 1) přibližně lineární.

Na obr. 4 jsou pro data z obr. 2 vybočující měření uvedeny α- kvantilové regresní přímky.



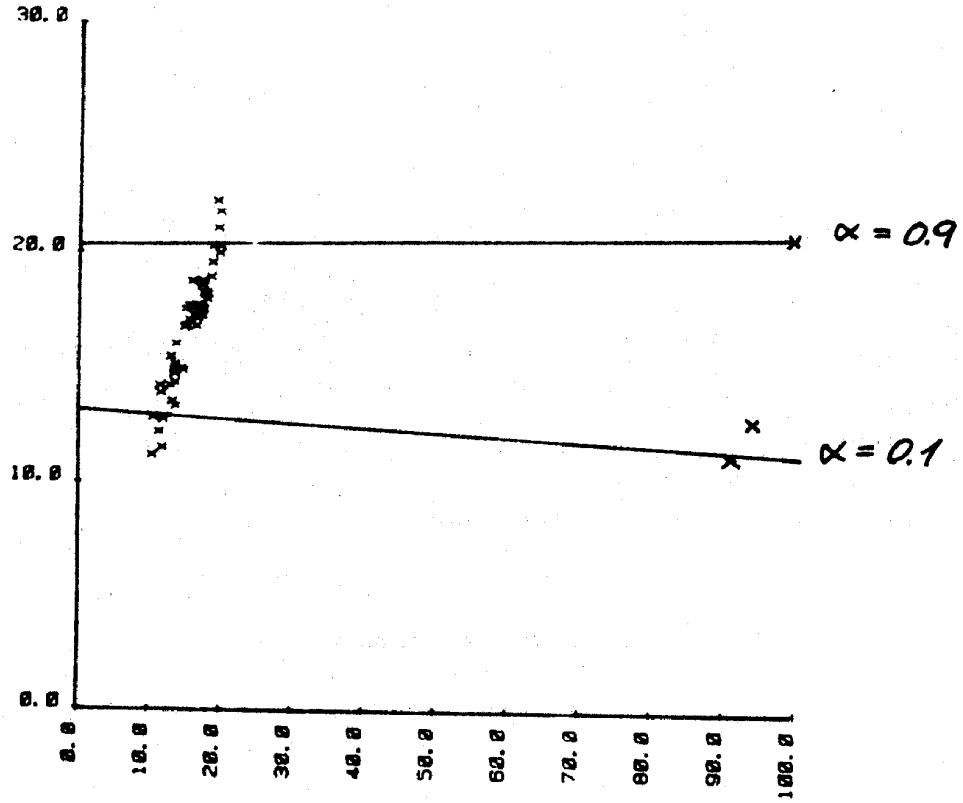
Obr. 4 Regresní α- kvantilové přímky pro data z obr. 2.

Na obr. 5 jsou pro data z obr. 3 (extrémní body) uvedeny α - kvantilové regresní přímky.



Obr. 5 Regresní α - kvantilové přímky pro data z obr. 3.

Z těchto obrázků je zřejmé, že pro případ, kdy se v datech vyskytují extrémní body, nejsou regresní α - kvantily robustní a nebude možno použít postup z práce /10/. To, že pro dostatečně "extrémní" body jsou oba regresní α- kvantily nerobustní, je patrné z obr. 6 (simulované data se třemi extrémními body při A = 80 a i = 12, 25, 37)



Obr. 6 Regresní α- kvantilové přímky pro simulovaná data se třemi extrémními body (A = 80, i = 12, 25, 37).

Z uvedeného je zřejmé, že pro vyloučení vlivných bodů bude nutné nejdříve provést eliminaci extrémních bodů (např. přes Cookovou statistiku D_i z rov. (22)). Pak již bude možné využít postup eliminace vybočujících bodů a využitím regresních kvantilů.

6. Závěr

V příspěvku byly zmíněny některé postupy identifikace vlivných bodů a jejich eliminace resp. omezení jejich vlivu při odhadu parametrů lineárních modelů. Nebyly zde zmíněny postupy využívající různých typů grafů (ty jsou sumárně diskutovány např. v práci /11/)ani metody pro testaci většího počtu vlivných bodů /2/. Při konstrukci regresních α- kvantilů bylo využito programu, který sestavil J.Antoch z MFF UK.

Literatura

/1/ Belsey D.A., Kuk E., Welsch R.E.: Regression Diagnostics, J.Wiley, New York 1980

/2/ Cook R.D., Weisberg S.: Residuals and Influence in Regression, Champmann and Hall, New York 1982

/3/ Militký J.: Tvorba matematických modelů - I, skripta pro kurs DT ČSVTS Ostrava 1983

/4/ Huber P.J.: Robust Statistics, J.Wiley, New York 1981 (překlad v ruštině 1984)

/5/ Cook R.D., Weisberg S.: Technometrics 22, 495 (1980)

/6/ Atkinson A.C. : Biometrika 68, 13 (1981)

/7/ Huber P.J.: J.A.S.A. 78, 66 (1983)

/8/ Kraeher W.S., Welsch R.E.: J.A.S.A. 77, 595 (1982)

/9/ Hill R.W., Commun. Statist. A 11, 849 (1982)

/10/ Antoch J. a kol., Proc. Compstat'84, Praha, září 1984

/11/ Militký J.: Sborník přednášek z konference, Numerické metody ve fyzikální metalurgii, Blansko, říjen 1984