GRAFICKÉ POSTUPY IDENTIFIKACE VLIVNÝCH BODŮ

Jiří Militký, VÚTZ Dvůr Králové n.L.

1. Úvod

Vlivné body (VB) představují jeden ze základních problémů, které se vyskytují při praktické aplikaci regresních metod. Obecně lze říci, že jako VB označujeme takové body, které výrazně ovlivňují výsledky regrese. Lze je rozdělit do tří základních skupin:

a) hrubé chyby způsobené měřením (outliers), nehomogenitou ve vysvětlujících proměnných (extrémy) resp. chyby vznikající při manipulaci s daty /1/
b) speciálně vybrané "přesně" měřené body s vysokým vlivem "golden points" /2/
c) důsledky nesprávně specifikovaného regresního modelu /3/.

V reálných situacích, kdy se neprovádí speciální plánování experimentů, se vyskytují zpravidla VB ze skupiny ad a) a ad b).
Při zpracování regresních úloh, kde se dá očekávat přítomnost VB (v praxi prakticky vždy), se využívá dvou základních přístupů.
První spočívá v použití robustních metod odhadu parametrů. Problematice robustních metod je neustále věnována pozornost zejména ze strany teoretiků. Od původních postupů, vhodných zejména pro eliminaci outliers (ležících co do hodnoty vysvětlované veličiny y mimo ostatní body), se postupem často dospělo k realističtějším postupům s ohraničeným vlivem, které eliminují také působení extrémů (ležícím co hodnoty minimálně jedné vysvětlující proměnné x_j mimo ostatní body).
Přehled novějších výsledků v této oblasti byl podán např. Jurečkovou /4/.
Základní výhodou robustních metod regrese je, že při "automatické analýze" na počítači (kdy nedochází k interakci mezi zadavatelem a počítačem) poskytují oproti MNČ adekvátnější výsledky. Mají však řadu nevýhod:

- vycházejí z předpokladu, že jde o VB ze skupiny ad a) a pouze omezují jejich působení
- neumožňují analýzu vzniku VB (a z výsledků nemusí být ani patrné, zda jsou v datech nějaké chyby)
- pro chybně specifikovaný regresní model vedou ke kuriózním výsledkům
- následná statistická analýza (intervalové odhady, testy) je komplikovaná

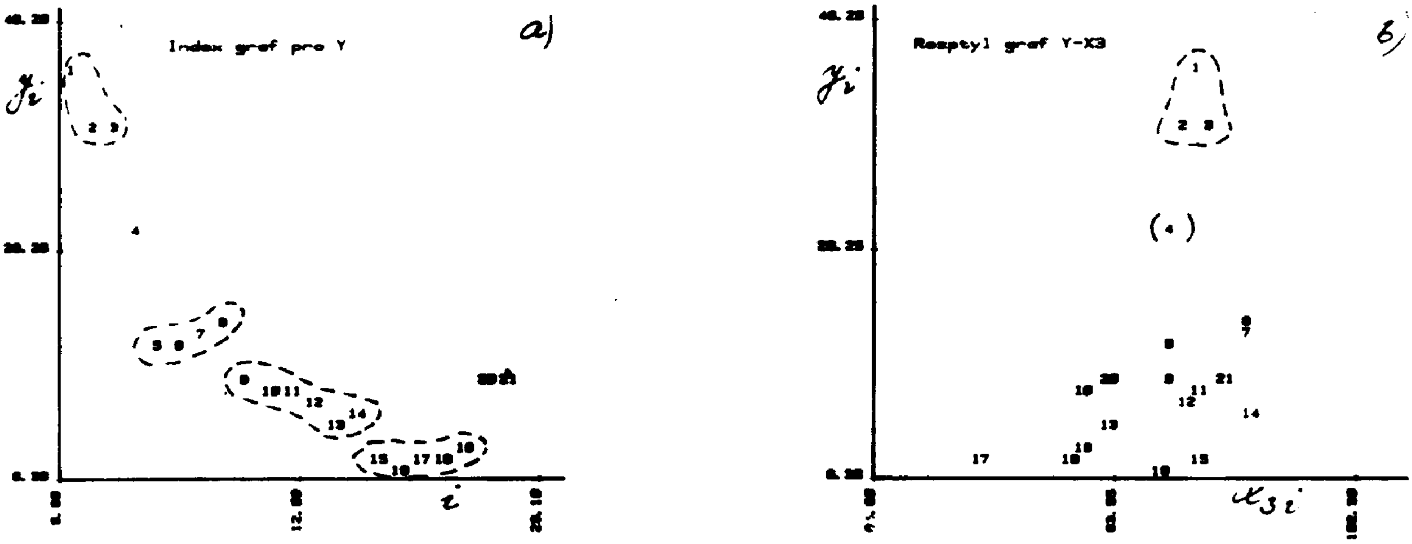
Druhý přístup spočívá ve využití postupů regresní diagnostiky, které umožňují identifikaci VB. Je vhodný pro interaktivní tvorbu regresních modelů a vyžaduje obecně znalost informací o vzniku dat. Kromě nesporných výhod, které poskytuje možnost identifikace typu VB příp. určení příčin jejich vzniku, má tento přístup také některé nevýhody. Mezi základní patří:

- časová náročnost (nikoliv z hlediska strojového času počítače) při budování regresního modelu
- problémy vznikající s možností "maskování" u skupin VB
- nutnost subjektivního rozhodnutí co se zjištěnými VB (jejich vylučování není vždy nejlepším postupem).

Mezi oběma přístupy existuje mnoho společného. Např. funkční váhy jednotlivých bodů vznikající při použití iterativní vážené metody nejmenších čtverců pro konstrukci robustních odhadů (IRWLS) jsou dobrou diagnostickou mírou vlivných bodů. Na druhé straně lze jako váhy při IRWLS použít diagnostické charakteristiky, vyjadřující míru vlivu každého bodu. Pro praktické účely je vždy vhodné využít syntézu obou přístupů. To obyčejně znamená začít od regresní diagnostiky a použít robustní regrese až ve fázi, kdy již známe správně specifikovaný model a rozhodujeme, co dále s VB (pokud je nelze jednoznačně vyloučit jako hrubé chyby, vzniklé manipulací s daty).

<u>Příklad 1:</u>

V řadě studií o robustní regresi se jako ilustrativní příklad volí tzv. Brownleeho data /6/, obsahující údaje o 21denním provozu oxidační aparatury pro výrobu kyseliny dusičné z amoniaku. Uvažují se tři vysvětlující proměnné:
x_1 - průtok vzduchu, x_2 - teplota chladicí vody, x_3 - koncentrace kyseliny dusičné v absorpční kapalině. Vysvětlovaná proměnná y je procentní ztráta amoniaku. Robustní odhady parametrů odpovídajícího lineárního modelu indikují jako VB body č. 1, 3, 4 a 21. Také při aplikaci metod pro identifikaci vlivných bodů lze určit jako nejvlivnější podmnožinu body č. 1, 2, 3, 4 a 21.
Jednoduchý indexový graf (vynesení y_i vs. i, kde i odpovídá dnu měření) na obr. 1a však indikuje vznik několika skupin bodů vzhledem k času. Také rozptylové grafy (obr. 1b) indikují odlišnost bodů č. 1, 2, 3 od ostatních.



Obr. 1 a) Indexový graf pro vysvětlovanou proměnnou (data /6/)
b) Rozptylový graf pro závislost y vs. x_3 (data /6/).

Detailnějším rozborem lze zjistit, že dochází k "náběhu" zařízení v prvních třech dnech a navíc se v dalších dnech projevuje efekt času /7/.
V práci /8/ je ukázáno, že data neumožňují konstrukci požadovaného regresního modelu. Tento příklad jasně dokumentuje, že pouze komplexnější přístup využívající řady diagnostických postupů spolu s věcnou interpretací zjištěných závěrů umožňuje dosažení prakticky použitelných výsledků. Robustní regrese zde vlastně pouze indikuje, že nejsou splněny všechny předpoklady klasické MNČ (neumožňuje však "nápravu" v žádném směru).

Pro identifikaci VB jsou prakticky výhodné grafické metody, protože /10/:

- poskytují informace o všech datech současně
- umožňují často identifikaci i skupin VB
- usnadňují posouzení i jiných zvláštností v datech (např. trendy, nelinearity, atp.), které ovlivňují výsledky regrese
- indikují často nejen velikost, ale i "směr" vlivu jednotlivých bodů vzhledem k ostatním

Na druhé straně je třeba mít na paměti, že grafické znázornění může vést ke vzniku artefaktů a nemusí poskytovat vždy jednoznačné výsledky.

V této práci jsou diskutovány základní typy grafů, vhodné pro identifikaci vlivných bodů. Označení navazuje na práci /7/.

2. Základní pojmy

Při konstrukci regresních modelů se vychází z dat $\{y_i, \underline{x}_i\}$ i=1,...,n, které tvoří n-tici bodů v prostoru E^{m+1}. O vysvětlovaných proměnných x_j j=1,...,m se běžně předpokládá, že jsou deterministické (bez újmy na obecnosti je $x_1 = 1$ pro všechny body, tj. uvažuje se absolutní člen). Pro náhodnou vysvětlovanou proměnnou y se (v případě, že byla data získána měřením) uvažuje tzv. aditivní model měření

$$y_i = \underline{x}_i^T \underline{a} + \varepsilon_i \tag{1}$$

kde $f(\underline{x}, \underline{a}) = \underline{x}^T \underline{a}$ je regresní model, o kterém se (předběžně) předpokládá, že vyhovuje pro vystižení variability vysvětlované proměnné. Na regresní parametry $\underline{a} = (a_1, \ldots, a_m)^T$ nejsou kladena žádná omezení. Pokud jsou chyby ε_i stejně rozdělené nezávislé náhodné veličiny s nulovou střední hodnotou $E(\varepsilon_i) = 0$ a konstantním rozptylem $D(\varepsilon_i) = \sigma^2 (i = 1, \ldots, n)$, lze nejlepší nestranné odhady $\hat{\underline{a}}$ získat minimalizací kritéria nejmenších čtverců (MNČ).

$$S(\underline{a}) = (\underline{y} - \mathbb{X}\underline{a})^T (\underline{y} - \mathbb{X}\underline{a}) \tag{2}$$

kde $\mathbb{X}^T = (\underline{x}_1, \ldots, \underline{x}_n)$ je $(m \times n)$ matice "plánu" a $\underline{y}$ je vektor měřených hodnot. Pokud navíc platí, že $\underline{\varepsilon}$ má normální rozdělení $N(\underline{0}, \sigma^2 \mathbb{E})$, jsou odhady $\hat{\underline{a}} = (\mathbb{X}^T\mathbb{X})^{-1} \mathbb{X}^T \underline{y}$, získané minimalizací rov. (2) navíc maximálně věrohodné a mají normální rozdělení se střední hodnotou $E(\hat{\underline{a}}) = \underline{a}$ resp. rozptylem $D(\hat{\underline{a}}) = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}$.

Na základě odhadů $\hat{\underline{a}}$ lze určit vektor predikce $\hat{\underline{y}}$ ve tvaru

$$\hat{\underline{y}} = \mathbb{X}\hat{\underline{a}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \underline{y} = \mathbb{H}\underline{y} \tag{3}$$

resp. vektor reziduí

$$\hat{\underline{e}} = \underline{y} - \hat{\underline{y}} = (\mathbb{E} - \mathbb{H})\underline{y} \tag{4}$$

Na základě rov. (1) a rov. (3) lze snadno určit, že rozptyl predikce je roven

$$D(\hat{\underline{y}}) = \sigma^2 \mathbb{H} \tag{5}$$

a rozptyl reziduí je roven

$$D(\hat{\underline{e}}) = \sigma^2 (\mathbb{E} - \mathbb{H}) \tag{6}$$

Nevychýleným odhadem rozptylu chyb je tzv. reziduální rozptyl $\hat{\sigma}^2 = \hat{\underline{e}}^T \hat{\underline{e}} / (n - m)$.

Je užitečné znázornit si geometrický význam MNČ ve výběrovém prostoru R^n. Z rov. (2) a (3) plyne, že vektor projekce $\hat{\underline{y}}$ je projekcí vektoru $\underline{y}$ do nadroviny R^m, definované sloupci matice $\mathbb{X}$. Vektor reziduí $\hat{\underline{e}}$ je projekcí vektoru $\underline{y}$ do ortogonálního doplňku $R^{\perp}_{n-m}$ k nadrovině R^m.
Z toho přímo plyne, že rezidua $\hat{\underline{e}}$ nemohou být nezávislé.
Projekční matice $\mathbb{H}$ a $\mathbb{E} - \mathbb{H}$ umožňují projekci libovolného n-rozměrného vektoru $\underline{z}$ do zvoleného podprostoru.

2.1 Projekční matice

V oblasti regresní diagnostiky hraje centrální roli projekční matice $\mathbb{H}$, která je jako všechny projekční matice idempotentní ($\mathbb{H}^2 = \mathbb{H}$) a symetrická ($\mathbb{H}^T = \mathbb{H}$). Její diagonální prvky $h_{ii} = x_i^T(\mathbb{X}^T\mathbb{X})^{-1}x_i$ se označují jako "leverage" a mají řadu zajímavých vlastností (viz /8, 10/):

1. Z vlastnosti projekční matice přímo plyne, že $0 \le h_{ii} \le 1$ a $-1 \le h_{ij} \le -1$ $(i \neq j)$. Navíc, pokud model obsahuje absolutní člen a hodnost matice $\mathbb{X}$ je rovna m, je $1/n \le h_{ii} \le 1/c$, kde n je počet měření a c je počet opakování měření (opakování i-tého řádku matice $\mathbb{X}$).

2. Pro modely s absolutním členem a plnou hodností matice $\mathbb{X}$ dále platí, že $\sum_{(i)} h_{ii} = m$ (průměrná hodnota $h_{ii} \approx m/n$) a $\sum_{(j)} h_{ij} = \sum_{(i)} h_{ij} = 1$

3. Z vlastnosti idempotence plyne, že $h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$. Z toho plynou dvě důležité vlastnosti diagonálních prvků projekční matice:
   a) pokud je $h_{ii} \to 0$, jsou všechna $h_{ij} \to 0$ $(j = 1, n)$
   b) pokud je $h_{ii} \to 1$, jsou všechna $h_{ij} \to 0$ $(j = 1, \neq i)$

4. Malé hodnoty h_{ii} vždy znamenají, že daný bod má x-ové souřadnice blízké těžišti x_T vysvětlujících proměnných. Velké hodnoty h_{ii} (uvádí se hraniční hodnoty mezi 0.2 až 0.5) indikují obyčejně extrémní bod (ne však vždy viz /8/)

5. Z rov. (3) plyne, že
$$\hat{y}_i = y_i \cdot h_{ii} + \sum_{j \neq i}^{n} h_{ij} \cdot y_j$$
To znamená, že pro $h_{ii} \to 1$ (kdy jsou $h_{ij} \to 0$) je přibližně $\hat{y}_i = y_i$ a veškerá variabilita v i-tém bodě je objasněna regresním modelem. Také pro dostatečně vysoké y_i (odlehlé měření) bude možné zbylé členy zanedbat a vyjde $\hat{y}_i \approx h_{ii} y_i$. Z rov. (4) pak plyne, že pro $h_{ii} \to 1$ je $\hat{e}_i \to 0$ bez ohledu na velikost y_i. Pro $h_{ii} \ll 1$ a dostatečně veliké y_i je $\hat{e}_i \approx y_i(1 - h_{ii})$.

6. Z rov. (5) plyne, že rozptyl $var(\hat{y}_i) = \sigma^2 h_{ii}$ a rozptyl $var\,\hat{e}_i = (1 - h_{ii})\sigma^2$. Pro extrémní body $(h_{ii} \to 1)$ bude rozptyl reziduí malý, což opět zkomplikuje idetifikaci VB

7. Pokud jsou vysvětlující proměnné x_j také náhodné veličiny s normálním rozdělením, platí pro velká n přibližně, že $n \cdot h_{ii} - 1$ má χ^2_m rozdělení.

V práci /11/ je ukázáno, jak využít této aproximace pro určení střední hodnoty maximálního diagonálního prvku matice $\mathbb{H}$.

Některé další vlastnosti veličin h_{ii} jsou uvedeny v /7, 8, 10/. Je zřejmé, (viz ad 5) že h_{ii} ukazují míru vlivu hodnoty y_i na predikci $\hat{y}_i$ v tomtéž bodě. Charakterizují dobře extrémní body (viz ad 4) a ovlivňují variabilitu reziduí i predikce. Používají se samostatně pro indikce extrémních VB .

2.2 Rezidua

Je přirozené používat pro identifikaci outliers funkcí reziduí $\hat{e}_i$. Samotná rezidua $\hat{e}_i$ mají řadu negativních vlastností, které omezují jejich přímé použití jako diagnostických charakteristik. Platí pro ně, že jsou:

a) lineární kombinací chyb ε_i, protože $\hat{e}_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij} \varepsilon_j$. Rozdělení reziduí závisí pak na rozdělení chyb, velikosti výběru n a prvcích matice $\mathbb{H}$. Pro malé výběry se pak projevuje supernormalita (rezidua se jeví normálněji rozdělená než chyby ε_i).

b) korelované s nekonstantním rozptylem (viz rov. (6)

c) špatnými odhady chyb ε_i, pokud jsou h_{ii} dostatečně vysoké ($h_{ii} \to 1$). To plyne přímo z ad a).

Pro velké počty bodů ($n \to \infty$) jsou všechna h_{ii}, h_{ij} malá, takže $\hat{e}_i$ pak dobře odhadují chování chyb ε_i. Z geometrie MNČ dále plyne, že $\hat{\underline{e}} \perp \hat{\underline{y}}$. Platí však, že $\hat{\underline{e}}$ jsou korelované s vektorem $\underline{y}$, kde $cov(\hat{\underline{e}}, \underline{y}) = \hat{\underline{e}}^T \hat{\underline{e}}$. Podobně jako rezidua, můžeme promítnout do prostoru $R^{\perp}_{n-m}$ také jednotkové vektory $\underline{i}$ (sloupce matice $\mathbb{E}$). Platí, že $\underline{i}^{\perp} = (\mathbb{E} - \mathbb{H})\underline{i}$ a také $\hat{e}_i = (\underline{i}^{\perp})^T \hat{\underline{e}}$.

Zkonstantnění rozptylu lze docílit podělením rezidui odpovídajícím rozptylem.
__Studentizovaná rezidua__ $\hat{r}_i$ (korektně standardizovaná) mají tedy tvar

$$\hat{r}_i = \frac{\hat{e}_i}{\sqrt{var(\hat{e}_i)}} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{7}$$

Studentizovaná rezidua již mají konstantní rozptyl. Zůstávají však korelovaná. Navíc platí, že náhodná veličina $\hat{r}_i^2/(n-m)$ má beta rozdělení $Be\left[0{,}5, \frac{n-m-1}{2}\right]$. Lze snadno dokázat, že $\hat{r}_i = \sqrt{n-m} \cdot \cos \Theta_i$, kde Θ_i je úhel, který svírá vektor $\hat{\underline{e}}$ s vektorem $\underline{i}^{\perp}$ v prostoru $R^{\perp}_{n-m}$. Pro $\hat{r}_i$ platí:

- pokud je $\hat{\underline{e}} \parallel \underline{i}^{\perp}$, vyjde maximální hodnota $\hat{r}_{max} = \sqrt{n-m}$
- pokud je $h_{ii} \to 1$ ($\hat{e}_i \to 0$), vyjde po detailnější analýze, že $\hat{r}_i \to 0$
- pro velká y_i a $h_{ii} \ll 1$ ($\hat{e}_i \gg \hat{e}_j$, $i \neq j$) vyjde přibližně $\hat{r}_i \approx \sqrt{n-m}/\sqrt{1-h_{ii}}$

Je zřejmé, že ani studentizovaná rezidua nemusí (pro případ, kdy $h_{ii} \to 1$) dobře indikovat přítomnost VB. Místo rozptylu $\hat{\sigma}^2$ lze použít nezávislý odhad $\hat{\sigma}^2_{(-i)}$, při jehož výpočetu byl vynechán i-tý bod.

$$\hat{\sigma}^2_{(-i)} = \left(\frac{n - m - \hat{r}_i^2}{n - m - 1}\right) \cdot \hat{\sigma}^2 \tag{8}$$

Dosazením $\hat{\sigma}^2_{(-i)}$ z rov. (8) za $\hat{\sigma}^2$ v rov. (7) vyjdou tzv. __plně studentizovaná "Jackknife" rezidua__ ve tvaru

$$\hat{t}_i = \hat{r}_i \sqrt{\frac{n-m-1}{n-m-\hat{r}_i^2}} \tag{9}$$

Je zřejmé, že Jackknife reziua $\hat{t}_i$ jsou monotónní transformací studentizovaných reziduí $\hat{r}_i$. V případě, že i-tý bod není "outlier", mají $\hat{t}_i$ Studentovo rozdělení s n-m-1 stupni volnosti. Platí také, že $\hat{t}_i$ jsou testační statistiky hypotézy H_0: $d = 0$ v modelu jednoduchého vybočujícího pozorování $\underline{y} = \mathbb{X}\underline{a} + \underline{i}\, d + \underline{\varepsilon}$ ($\underline{i}$ je i-tý sloupec matice $\mathbb{E}$).

Pro $\hat{t}_i$ lze říci, že:

- pokud je $h_{ii} \to 1$ ($\hat{e}_i \to 0$), vyjde také $\hat{t}_i \to 0$
- pokud je y_i veliké a $h_{ii} \approx 0$ ($\hat{e}_i \gg \hat{e}_j$; $j \neq i$), vyjde $\hat{t}_i \to \infty$

Také Jackknife rezidua zůstávají pochopitelně korelovaná.

Odhadem velikosti d v modelu jednoduchého vybočujícího pozorování jsou tzv. __predikovaná rezidua__ $\hat{e}_{(-i)}$, pro která platí $\hat{e}_{(-i)} = y_i - \underline{x}_i^T \hat{\underline{a}}_{(-i)}$. Odhady $\hat{a}_{(-i)}$ jsou získány ze všech bodů kromě $(y_i, \underline{x}_i^T)$. Pro predikovaná rezidua lze psát

$$\hat{e}_{(-i)} = \hat{e}_i / (1 - h_{ii}) \tag{10}$$

Pro velké výběry, kdy $h_{ii} \approx h$, budou predikované rezidua odpovídat reziduím klasickým. Dále pro $\hat{e}_{(-i)}$ platí:

- pokud je $h_{ii} \to 1$ ($\hat{e}_i \to 0$), vyjde $\hat{e}_{(-i)} = y_i$
- pokud je $h_{ii} << 1$ a y_i dostatečně veliké, je také $\hat{e}_{(-i)} = y_i$

V limitních případech jsou tedy predikované rezidua přímo rovna odpovídajícím hodnotám y_i (které mohou být v případě, že $h_{ii} \to 1$, v rozmezí ostatních hodnot vysvětlované proměnné).

Z uvedených typů reziduí se pro účely regresní diagnostiky používá buď přímo Jackknife reziduí $\hat{t}_i$ nebo různých funkcí studentizovaných reziduí $\hat{r}_i$.

Jako příklad nekorelovaných reziduí (musí jich být pouze n-m) uveďme tzv. <u>rekurzivní rezidua</u> w_i , pro která platí

$$w_i = \left(y_i - \underline{x}_i^T \hat{\underline{a}}_{R\,i-1}\right) \cdot \left[1 + \underline{x}_i^T \left(X_{i-1}^T X_{i-1}\right)^{-1} \underline{x}_i\right]^{-1/2} \tag{11}$$

$$i = m+1, \ldots n$$

V rov. (11) jsou $\hat{\underline{a}}_{Ri-1}$ odhady získané z (i-1) bodů a matice X_{i-1} obsahuje prvních (i-1) řádků matice X . Platí, že $w_i = 0$ pro $i = 1, \ldots m$. Vychází se z m-tice bodů tzv. <u>báze,</u> která se volí tak, aby neobsahovala VB . Na rekurzivní rezidua se dá také nahlížet jako na "jednokrokové" standardizované predikce reziduí. Za předpokladu, že $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 E)$, jsou rezidua w_i nezávislá, normálně rozdělená s konstantním rozptylem.

<u>2.3 Vlivné body</u>

Při analýze vlivných bodů se používají dva základní přístupy /12/. Přístup, založený na <u>vypouštění jednotlivých bodů</u> vychází z určení změn ve výsledcích regrese, ke kterým dochází vlivem vynechání i-tého bodu. Je použit zejména v knize /13/. Základním vztahem je zde rozdíl mezi odhadem $\hat{\underline{a}}$ a odhadem $\hat{\underline{a}}_{(-i)}$, pro který platí

$$\hat{\underline{a}} - \hat{\underline{a}}_{(-i)} = \left(X^T X\right)^{-1} \underline{x}_i \, \hat{e}_i \,/\, (1 - h_{ii}) \tag{12}$$

Pro vyjádření relativní významnosti složek rozdílu mezi odhady $\hat{a}$ a $\hat{a}_{(-i)}$, tj. <u>vlivu i-tého bodu na odhad parametrů</u>, se používají standardizované složky rov. (12) ve tvaru

$$DFS_{ij} = \left[\hat{a}_j - \hat{a}_{j(-i)}\right] / \left(\hat{\sigma}_{(-i)} \cdot \sqrt{c_{ii}}\right) \tag{13}$$

kde c_{ii} jsou diagonální prvky matice $(X^T X)^{-1}$. Vliv i-tého bodu na odhad j-tého parametru se považuje za významný, pokud $DFS_{ij} > 2/\sqrt{n}$.
Podobně lze vyjádřit vliv i-tého bodu na predikci ve standardizovaném tvaru

$$DF_i = \frac{|\hat{y}_i - \hat{y}_{i(-i)}|}{\hat{\sigma}_{(-i)} \sqrt{h_{ii}}} = |\hat{t}_i| \cdot \sqrt{h_{ii}/(1 - h_{ii})} \tag{14}$$

Přitom platí, že pro $DF_i > 2\sqrt{m/n}$ se vliv i-tého bodu na predikci považuje za významný.
Řada dalších charakteristik tohoto typu je uvedena v /10, 13/.

Přístup, založený na <u>infinitezimálních změnách</u>, využívá sledování vlivu "perturbací" v rozptylu chyb ε na výsledky regrese. Je použit v knize /8/.

Vychází se z předpokladu, že i-tá chyba má rozdělení $\varepsilon_i \sim N(\emptyset, \sigma^2/w)$, zatímco ostatní chyby mají rozdělení $N(\emptyset, \sigma^2)$. Pro odhad parametrů $\hat{a}(w)$ je v tomto případě třeba použít vážené MNČ. Pro charakterizaci lokálních změn odhadů $\hat{a}(w)$ se používá vlivových funkcí, což jsou derivace $\partial \hat{a}(w)/\partial w$ v oblasti zvoleného w.

Tak pro případ, že $w \to \emptyset$, rezultuje tzv. Jackknife vlivová funkce

$$\frac{\partial \hat{a}(w)}{\partial w}\Big|_{w \to \emptyset} = (X^T X)^{-1} x_i \, \hat{e}_i / (1 - h_{ii})^2 \tag{15}$$

Pro konečné výběry se definuje výběrová vlivová funkce, která je (až na konstantu) rovna rozdílu $\hat{a} - \hat{a}_{(-i)}$ (viz rov. (12)).

Další varianty vlivových funkcí jsou obsaženy např. v /8, 14/.

Welsch /14/ doporučuje pro globální vyjádření vlivu i-tého bodu statistikou

$$G_i = \sqrt{n-1} \cdot |\hat{t}_i| \cdot \frac{\sqrt{h_{ii}}}{1 - h_{ii}} \tag{16}$$

které je až na faktor $\sqrt{1 - h_{ii}}$ rovna charakteristice DF_i (viz rov. (14)). Pro vyjádření vzdálenosti mezi odhady $\hat{a} - \hat{a}_{(i)}$ se používá Cookova statistika

$$D_i = \frac{(\hat{a}_{(-i)} - \hat{a})^T (X^T X)(\hat{a}_{(-i)} - \hat{a})}{m \cdot \hat{\sigma}^2} = \frac{1}{m} \cdot \hat{r}_i^2 \cdot \frac{h_{ii}}{1 - h_{ii}} \tag{17}$$

Z rov. (17) je patrné, že jde o analogii s klasickým konfidenčním elipsoidem, což umožňuje porovnat D_i s α-kvantilem $F_\alpha(m, n-m)$ Fisherova rozdělení.

Atkinson /15/ nahrazuje pro zvýšení citlivosti statistiky D_i studentizovaná rezidua Jackknife rezidui. Dále doporučuje použít odmocniny z D_i (pro možnost znázornění jako speciálních reziduí) a provést vhodnou standardizaci pro zvýraznění rozdílu proti D-optimálnímu plánu (kde $h_{ii} = m/n$). Výsledná modifikovaná Cookova statistika má tvar

$$T_i = |\hat{t}_i| \cdot \sqrt{\frac{n-m}{m} \cdot \frac{h_{ii}}{1 - h_{ii}}} \tag{18}$$

Lze snadno ukázat, že až na konstantu, je T_i stejné jako DF_i v rov. (14) a tedy blízké G_i v rov. (16).

Přehled ostatních charakteristik pro vyjádření relativního vlivu jednotlivých charakteristik na výsledek regrese je uveden v /8, 10/. Důležité je, že naprostá většina z nich je funkcí $\hat{r}_i$ a h_{ii}.

3. Grafické metody identifikace VB

Pro identifikaci VB se používá řada různých grafických postupů, které se vzájemně doplňují a částečně i překrývají. Zde se omezíme na tyto základní:

1. Parciální regresní grafy
2. Indexové a rankitové grafy
3. Grafické analogie charakteristik VB
4. Grafy využívající projekce do hlavních komponent

Pro ilustraci jednotlivých grafů je použit testační příklad z práce /16/. Tam byla pro m=4 a n=26 připravena simulovaná data s předem známými VB. Proměnná y byla generována podle vztahu

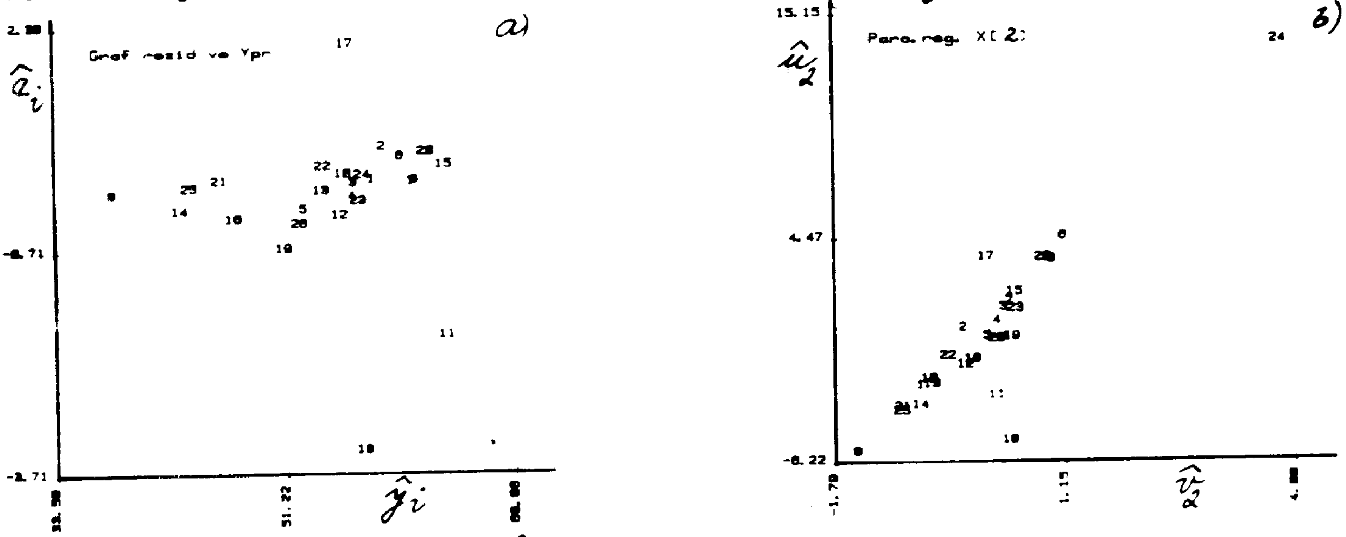
$$y = 10 + 3x_2 - 2x_3 + N(0;\ 0.025) \tag{19}$$

Proměnná x_4 byla generována jako přibližně lineární kombinace proměnných x_2 a x_3 podle vztahu

$$x_{4i} = 60 - 3x_2 - 1.5x_3 + N(0;\ 0.16) \tag{20}$$

(v těchto vztazích je $N(0, \sigma^2)$ pseudonáhodné číslo s normálním rozdělením a rozptylem σ^2). Dále byly upraveny y-nové hodnoty tak, aby body č. 11, 17 a 18 byly výrazné "outliers". Bod č. 24 byl modifikován tak, aby ležel mimo rovinu ostatních vzhledem k proměnné x_4 (skrytý extrém). Pro tato data byl použit lineární regresní model $y = a_1 + \sum_{j=2}^{4} a_j x_j$

Na obr. 2 je uveden klasický reziduový graf $\hat{e}_i$ vs. $\hat{y}_i$ pro tento příklad.



Obr. 2 a) Graf rezidua vs. $\hat{y}$ pro testační data

b) Parciální regresní graf pro x_2 .

Je patrné, že již tento graf indikuje všechny tři outliers. Extrémní bod č. 24 však indikován není (i když ovlivňuje výsledek regrese v rozhodující míře).

Dále uváděné grafy jsou výsledkem programu "REG-DIA", vytvořeného v jazyce HPL pro stolní počítač HP 9825. Odhady parametrů se počítají rekurzivní MNČ z libovolně volených bázových bodů. Kromě dále uvedených grafů umožňuje program tvorbu grafů typu CUSUM, Komponenta+reziduum, Konstrukční proměnné a Přidané proměnné (viz /10/). Tento program je součástí souboru programů pro regresní diagnostiku.

3.1 Parciální regresní grafy

Tyto grafy se považují za základní nástroj interaktivní tvorby regresních modelů /15/, protože kromě indikace VB diagnostikují další porušení předpokladů o chybách. Umožňují hodnocení závislosti mezi y a x_k při zkonstantnění vlivu ostatních proměnných.

Označme $\hat{u}_k$ rezidua regrese y na $X[k]$ (kde $X[k]$ je $(n \times m-1)$ matice, která vznikne vynecháním k-tého sloupce matice X) a $\hat{v}_k$ rezidua regrese x_k^S na $X[k]$ (x_k^S je k-tý sloupec matice X). Je zřejmé, že platí

$$\hat{u}_k = (E - H_k)\, y \qquad \hat{v}_k = (E - H_k)\, x_k^S \tag{21}$$

kde $\mathbb{H}_k = \mathbb{X}[k](\mathbb{X}[k]^T \mathbb{X}[k])^{-1} \mathbb{X}^T[k]$ je projekční matice. Parciální regresní graf pro k-tou vysvětlující proměnnou je pak závislost $\hat{\underline{u}}_k$ (na ose y) vs. $\hat{\underline{v}}_k$ (na ose x). Tento graf má řadu zajímavých vlastností:

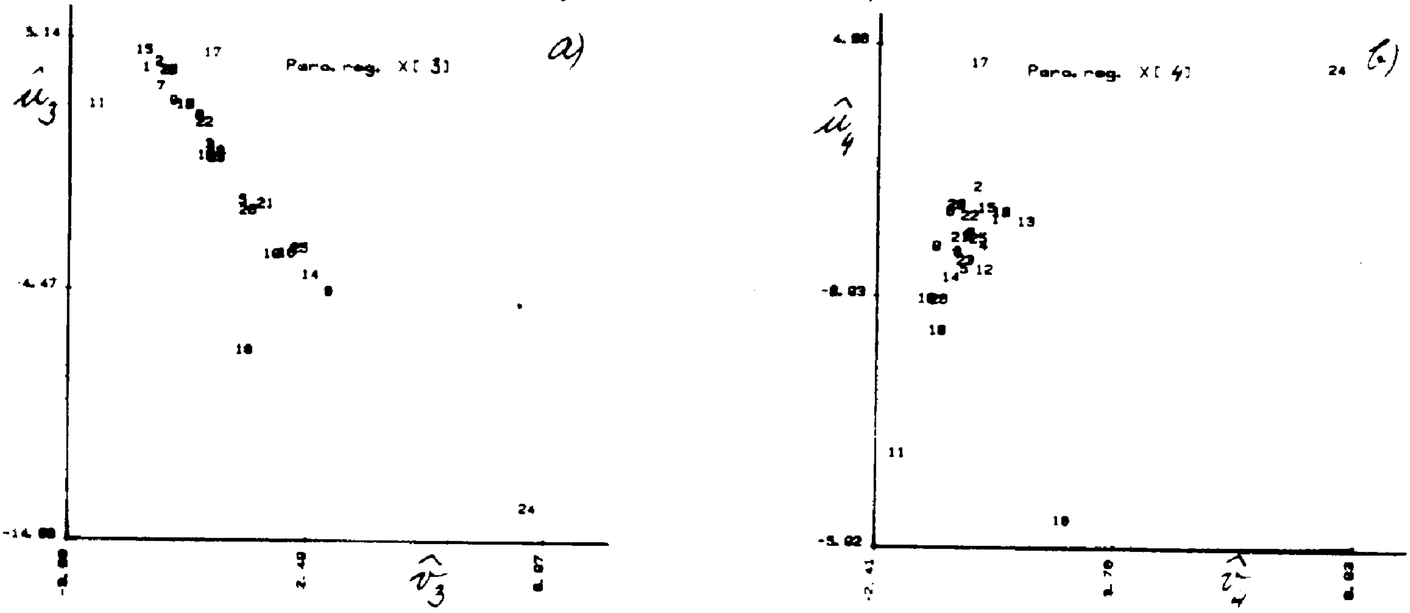
- směrnice je rovna $\hat{a}_k$ (pro model $\mathbb{X}\underline{a}$) a úsek je roven nule (pokud je regresní model dobře specifikován),
- korelační koeficient je přímo parciální regresní koeficient $R_{yx_k(x_1,\ldots x_m)}$
- odchylky od regresní přímky jsou přímo rezidua $\hat{\underline{e}} = (\mathbb{E}-\mathbb{H})\underline{y}$, protože platí

$$\hat{\underline{u}}_k = \hat{\underline{v}}_k^T \hat{a}_k + \hat{\underline{e}} \tag{22}$$

- jsou zvýrazněny vlivné body a případná heteroskedasticita.

Samotná $\hat{\underline{v}}_k$ indikuje účinek odstranění k-té vysvětlující proměnné na změnu významnosti jednotlivých bodů (parciální leverage).

Na obr. 2b je parciální regresní graf pro x_2 a na obr. 3a,b pro x_3 a x_4 z testačního příkladu. Je zřejmé, že všechny grafy dobře indikují všechny VB . Navíc je patrné, že proměnná x_4 prakticky nepřispívá k výsledku regrese (body bez VB tvoří shluk s prakticky nulovou směrnicí).

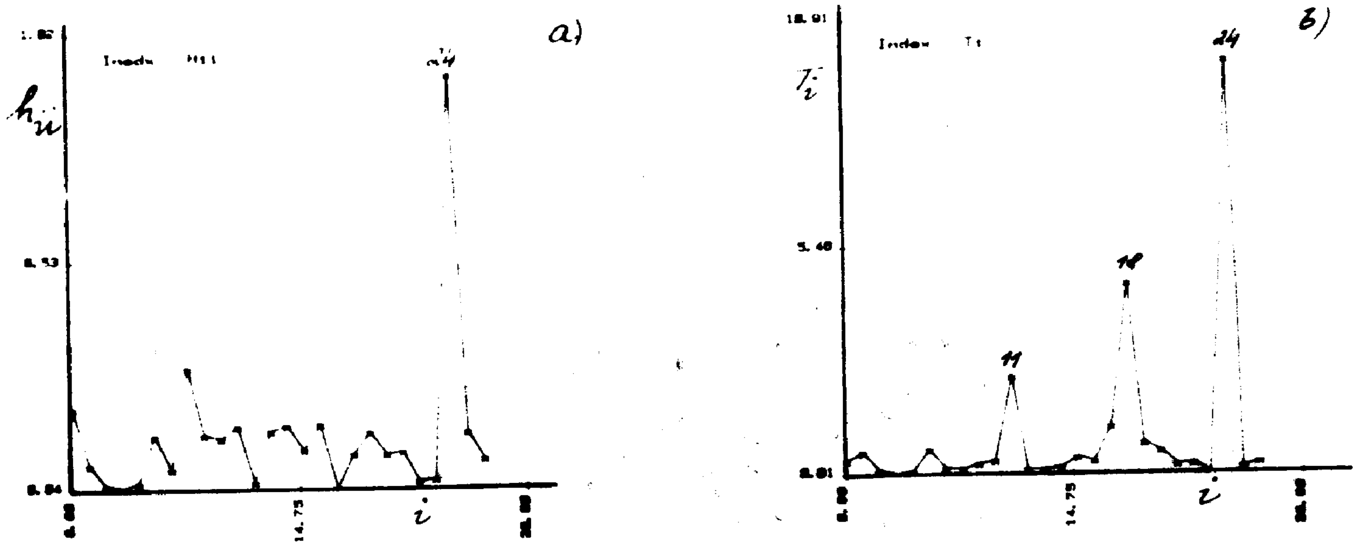


Obr. 3 a) Parciální regresní graf pro x_3
b) Parciální regresní graf pro x_4 .

3.2 Indexové a rankitové grafy

Indexové grafy jsou velmi jednoduché. Vynáší se zvolená charakteristika VB v závislosti na jejím indexu (je vhodné, aby data byla přirozeně uspořádána v pořadí, jak byla měřena). V programu "REG-DIA" jsou indexové grafy pro různé typy **reziduí** ($\hat{r}_i$, $\hat{t}_i$, w_i, $\hat{e}_{(-i)}$), **charakteristiky vlivných bodů** (DFS_i, DF_i, T_i, Q_i), **diagonální prvky projekční matice** (h_{ii}) a **rekurzivní odhady** ($\hat{a}_{Ri}$). Některé z těchto grafů jsou pro testační příklad uvedeny dále.

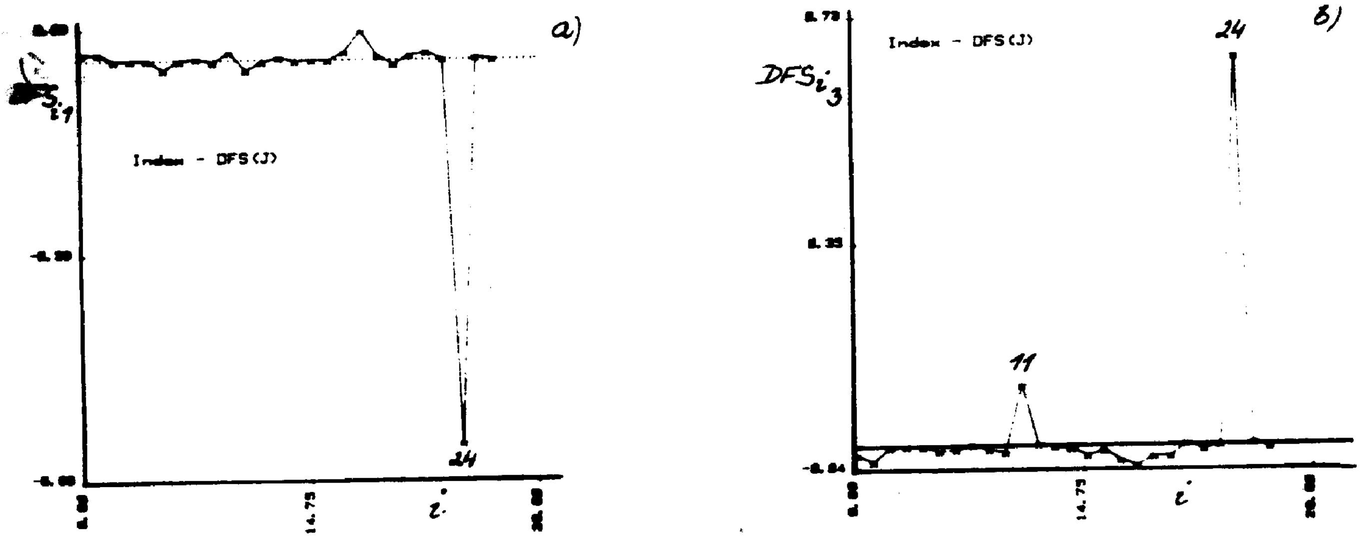
Na obr. 4a je indexový graf pro diagonální prvky matice $\mathbb{H}$. Je dobře indikován extrémní bod č. 24. Na obr. 4b je indexový graf pro modifikovanou Cookovu statistiku T_i (viz rov. (18)). Opět vychází jako nejvlivnější bod č. 24. Následuje bod č. 18 a 11.



Obr. 4 a) Indexový graf pro h_{ii}

b) Indexový graf pro T_i .

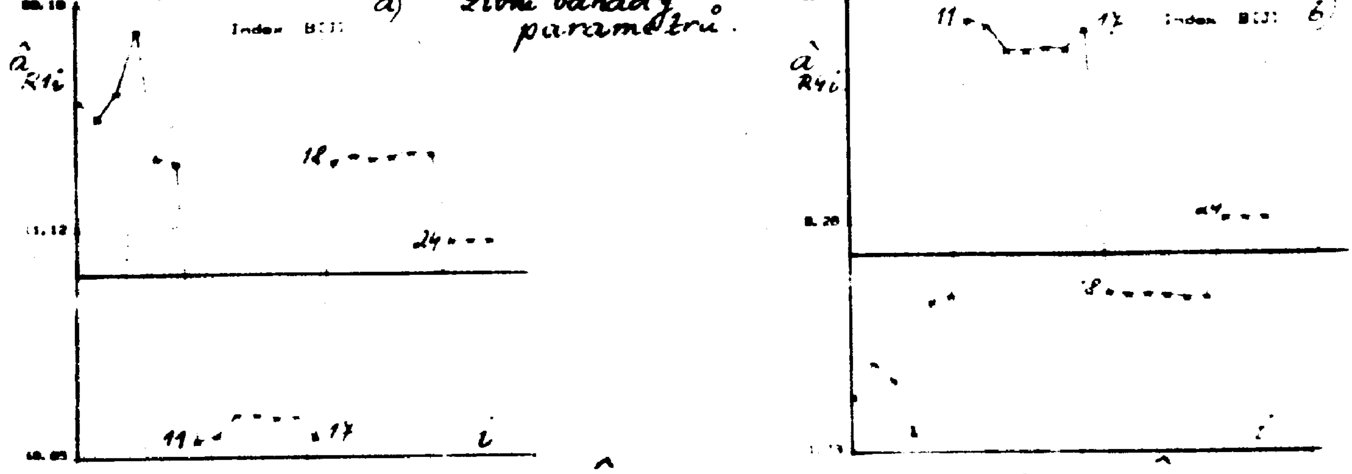
Na obr. 5a je indexový graf pro veličinu DFS_{i1} a na obr. 5b pro DFS_{i3} . Je patrné, že nejvlivnější je opět bod č. 24 a výrazně méně vlivný je bod č. 11.



Obr. 5 a) Indexový graf pro DFS_{i1}

b) Indexový graf pro DFS_{i3} .

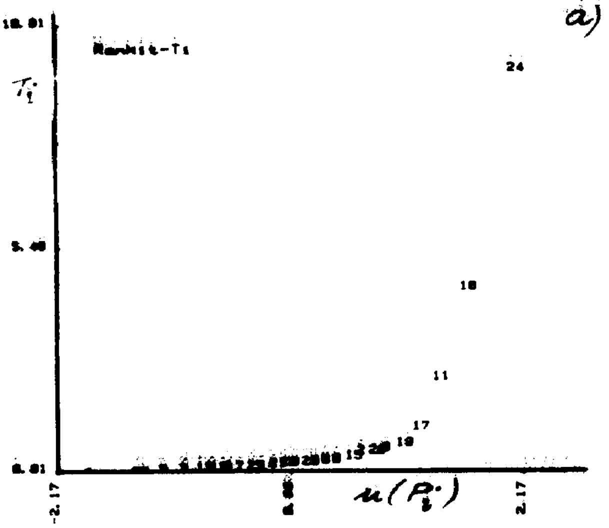
Konečně na obr. 6a je indexový graf pro rekurzivní odhad úseku $\hat{a}_{R1j}$ a na obr. 6b indexový graf pro rekurzivní odhad koeficientu $\hat{a}_{R4j}$ (odpovídající proměnné x_4). Je dobře viditelné, jak přítomnost všech VB skokově mění rekurzivní odhady parametrů.

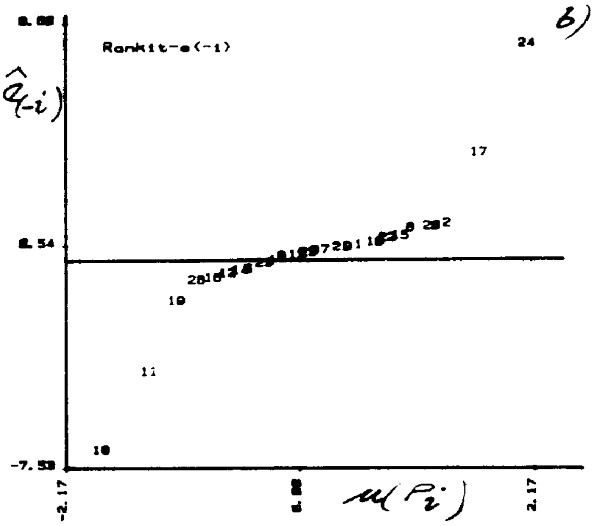


Obr. 6 a) Indexový graf pro $\hat{a}_{R1i}$ b) Indexový graf pro $\hat{a}_{R4i}$

Rankitové grafy jsou vlastně Q-Q grafy pro testaci normality. Na y-novou osu se vynášejí pořádkové statistiky $R_{(i)}$ (vzestupně setříděné hodnoty rezidui) a na osu x kvantily normovaného normálního rozdělení $u(P_i)$ pro $P_i = i/(n+1)$. Lineární průběh bez výrazně vybočujících "konců" je zde příznakem "normality" reziduí (pozor na supernormalitu) a nepřítomnosti VB.

V programu "REG-DIA" jsou rankitové grafy pro rezidua $R_i \sim \hat{e}_{(-i)}, \hat{e}_i, w_i$ a statistiku T_i. Speciálně v případě T_i způsobuje přítomnost VB konvexní průběh.
Na obr. 7a je rankitový graf pro modifikovanou Cookovu statistiku T_i a na obr. 7b rankitový graf pro predikované reziduum $\hat{e}_{(-i)}$. V obou případech jsou dobře indikovány všechny vybočující body (nejvlivnější vychází opět č. 24).



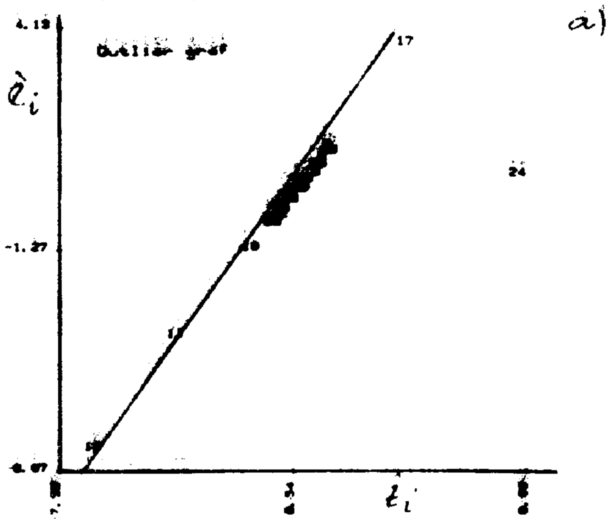


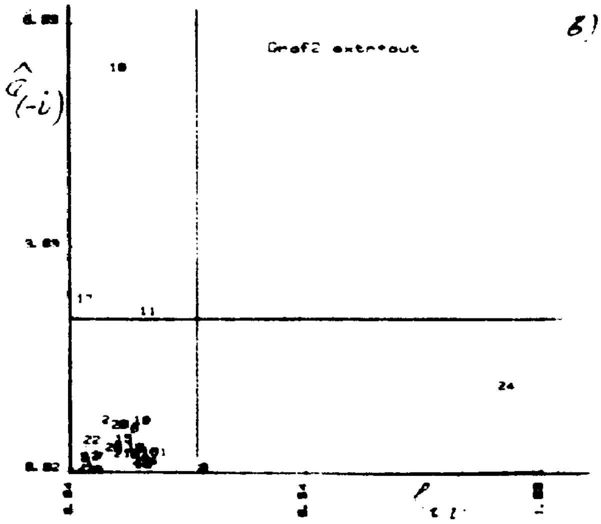
Obr. 7 a) Rankitový graf pro T_i
b) Rankitový graf pro $\hat{e}_{(-i)}$

Z porovnání obr. 4b a 7a je patrné, že pro modifikovanou Cookovu statistiku T_i jsou rankitové grafy zřejmě ilustrativnější.

3.3 Grafická analýza charakteristik VB

Do této skupiny lze zařadit řadu grafů, které vycházejí z charakteristik VB. Mezi nejjednodušší patří tzv. graf predikovaných reziduí (GPR), ve kterém se na osu y vynáší rezidua $\hat{e}_i$ a na osu x rezidua predikovaná $\hat{e}_{(-i)}$.
V tomto grafu leží extrémy mimo přímku $y = x$. "Outliers" leží v blízkosti této přímky, ale mimo ostatní data. To je dobře patrné z obr. 8a.





Obr. 8 a) průběh GPR
b) ...

Jednoduchý je také **Williamsův graf** (WG), kdy se na osu y vynáší Jackknife rezidua a na osu x prvky projekční matice h_{ii}. Do tohoto grafu se zakreslují mezní linie pro outliere $y = t_{0.95}(n-m-1)$ a mezní linie pro extrémy $x = 2m/n$. Tento graf je znázorněn na obr. 8b.

V programu "REG-DIA" je ještě grafická analogie Cookovy statistiky D_i a tzv. Pregibonův garf (viz /10/).
Mezi nejvhodnější "univerzální" znázornění různých charakteristik vlivných bodů patří **L-R grafy** /17/. Zde se na osu y vynáší normovaná rezidua $\hat{e}_i^2/RSC$ a na osu x prvky h_{ii} projekční matice. Lze snadno ukázat, že všechny možné body v tomto grafu padnou do tzv. L-R trojúhelníku, definovaného vztahy

$$0 \le h_{ii}\ ; \quad 0 \le \hat{e}_i^2/RSC\ ; \qquad h_{ii} + \hat{e}_i^2/RSC \le 1$$

Dále platí (viz kap. 2.3), že řada charakteristik vlivných bodů je ve tvaru $K(n,m) \cdot f(h_{ii},\ \hat{e}_i^2/RSC)$, kde $K(n,m)$ je konstanta závislá pouze na parametrech n, m. Pak lze snadno do L-R trojúhelníku zakreslit **izočáry** stejného vlivu a tak porovnávat polohy jednotlivých bodů. Tak např. charakteristiku DF_i z rov. (14) je možno přepsat na tvar

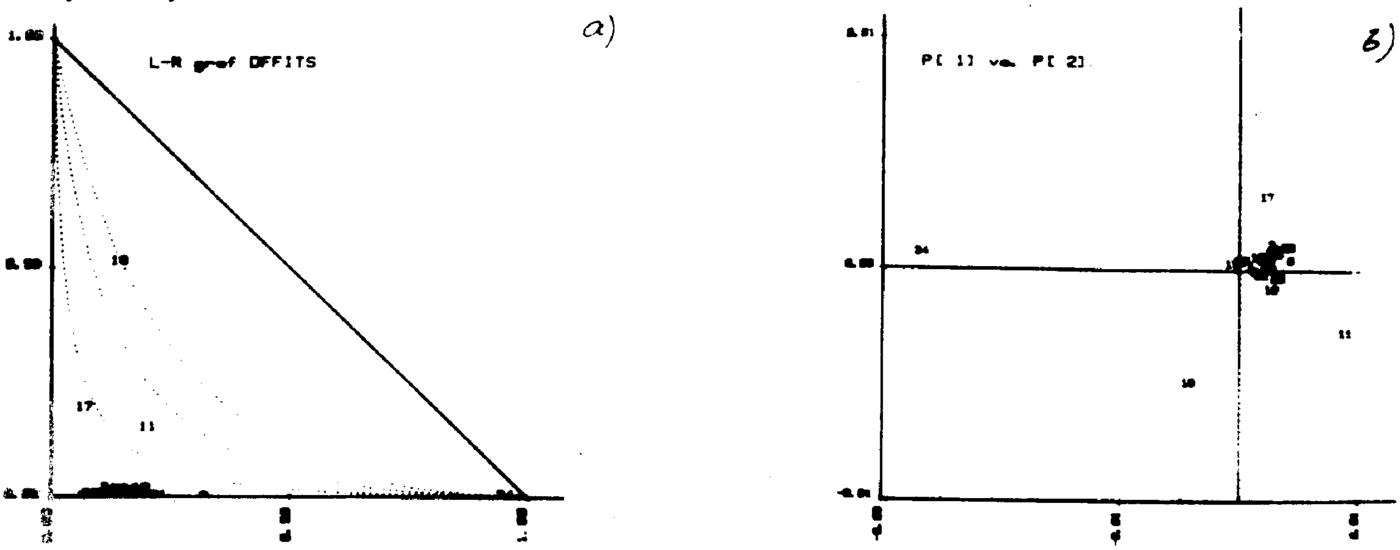
$$DF_i = \sqrt{n-m-1} \cdot \sqrt{\frac{h_{ii}(\hat{e}_i^2/RSC)}{(1-h_{ii})[1-h_{ii}-\hat{e}_i^2/RSC]}} \tag{23}$$

Po jednoduchých úpravách lze zjistit, že izočáry stejného vlivu jsou v tomto případě hyperboly typu

$$y = \frac{2x - x^2 - 1}{x(K-1)-1} \qquad \text{kde} \qquad K = \frac{n(n-m-1)}{c^2 \cdot m}$$

Zde c je volitelná konstanta ($c = 2$ odpovídá hranici $2\sqrt{m/n}$).

Na obr. 9a je zakreslen L-R graf pro DF_i. Čárkovaně jsou znázorněny izolinie stejného vlivu pro c=2, 4, 8.
Opět je patrné, že jsou indikovány všechny VB a lze posoudit i jejich relativní vliv. L-R grafy lze analogickým způsobem sestrojit také pro jiné charakteristiky z kap. 2.3.



Obr. 9 a) L-R graf pro DF_i b) Rozšířené hlavní komponenty.

3.4 Grafy, využívající projekce do hlavních komponent

Dnes již existuje celá řada grafů, které umožňují vhodnou projekci vícerozměrných dat do vhodně vybrané roviny. V oblasti exploratorní analýzy se tato rovina často definuje pomocí vlastních vektorů, odpovídajících největším vlastním číslům matice $\mathcal{X}^T\mathcal{X}$. Pro případ regrese doporučuje Hocking (viz /2/) jako doplněk graf rozšířených hlavních komponent (GRHK). Vychází se z kovarianční matice $\mathcal{C}^*=(\mathcal{X}^*\mathcal{Y}^*)^T(\mathcal{X}^*\mathcal{Y}^*)$, kde $\mathcal{X}^*, \mathcal{Y}^*$ jsou standardizované (mají nulovou střední hodnotu a jednotkový součet čtverců). Určí se vlastní vektory $\mathcal{V}$, odpovídající matici $\mathcal{C}^*$ a následně matice $\mathcal{S}=(\mathcal{X}^*\mathcal{Y}^*)\mathcal{V}$. Pokud je provedeno přeuspořádání sloupců matice $\mathcal{V}$ takové, že první patří vlastnímu vektoru s největším vlastním číslem a druhý vlastnímu vektoru s druhým největším vlastním číslem, je GRHK závislost $\mathcal{S}_1$ vs. $\mathcal{S}_2$ (kde $\mathcal{S}_i$ je i-tý sloupec matice $\mathcal{S}$). Na těchto grafech je možné jednoduše definovat všechny typy VB (i když to není obecně zajištěno - viz /2/).
Na obr. 9b je GRHK pro testační příklad. Je na první pohled patrné, které body jsou výrazně vybočující.

4. Závěr

V této práci bylo provedeno omezení především na identifikaci jednotlivých VB . Při identifikaci podskupin VB je obecně třeba použít komplikovanějších postupů (viz /2/), i když dle dosavadních omezených experimentů lze ve většině případů úspěšně použít kombinací uváděných grafů. Je zřejmé, že grafické metody identifikace VB jsou užitečné jako jeden ze základních "nástrojů" regresní diagnostiky resp. jako jeden z prostředků interaktivní tvorby regresních modelů.

Literatura

/1/ Beckman R.J., Cook R.D.: Technimetrics 25, 119 (1983)

/2/ Gray I.B., Ling R.F.: Technimetrics 26, 326 (1984)

/3/ Atkinson A.C.: Technometrics 25, 23 (1983)

/4/ Jurečková J.: Přednáška na zimní škole ROBUST ´86, Teplice n.M., 1986

/5/ Brownlee K.A.: Statistical Theory and Methodology in Science and Engineering, New York, J. Wiley 1965

/6/ Barnard G.A.: J.R.Stat.Soc. B44, 27 (1982)

/7/ Militký J.: Vlivné body v lineární regresi, letní škola ROBUST ´84, Slavonice 1984

/8/ Cook R.D., Weisberg S.: Residuals and Influence in Regression, Champmann and Hall, New York 1982

/9/ Daniel C., Wood F.: Fitting Equations to Data, 2nd Edition, J. Wiley, New York 1980

/10/ Militký J.: Vybrané matematicko-statistické metody v textilním průmyslu, díl III, Modelování a regrese, DT Pardubice, říjen 1984

/11/ Ascombe F.J.: Computing in Statistical Science through APL, Springer, New York 1981

/12/ Pregibon D.: Annals of Statist. 9, 705 (1981)

/13/ Belsey D.A., Kuh E., Welsch R.E.: Regression Diagnostics, J. Wiley, New York 1980

/14/ Welsch R.E. in Lanner R.L., Siegel A.F. Eds.: Modern Data Analysis, Academic Press, New York 1982, str. 149

/15/ Henderson H.V., Vellemann P.F.: Biometrics 37, 391 (1981)

/16/ Hockink R.R., Pendleton O.J.: Commun. Statist. A12, 497 (1983)

/17/ Gray J.B.: Technometrics (v tisku)