Regresní diagnostika v materiálovém výzkumu

JIŘÍ MILITKÝ ,
Katedra textilních materiálů, Technická universita v Liberci, Hálkova 6
461 17 Liberec, e- mail: jiri.miliky@vslib.cz

MILAN MELOUN,
Katedra analytické chemie, Universita Pardubice, Pardubice

Abstrakt: *Jsou vedeny základní problémy při použití regresních metod pro popis vztahu mezi složením a vlastnostmi materiálů. Je přehledně pojednáno o logických principech metody nejmenších čtverců pro lineární modely. Jsou uvedeny základní úlohy regresní diagnostiky. Detailněji jsou probrány metoly hodnocení vlivných bodů a jejich skupin. Je popsán program REGDIA v jazyce MATLAB pro identifikaci vlivných bodů v lineárních regresních modelech.*

1.Úvod

Jednou z frekventovaných úloh řešených v rámci hutní analytiky je analýza složení rud, resp. obsahu prvků v kovech a slitinách. Účelem je kromě hodnocení kvality hledat vztahy mezi složením a vlastnostmi materiálů. Představou je, že vlastnost materiálu P se dá vyjádřit funkcí

$$P = f(s_1 \ldots s_m),$$

kde s_i jsou obsahy jednotlivých prvků, resp. sloučenin v materiálu.Modely pro vyjádření vlastností materiálu v závislosti na jeho složení se vyskytují frekventovaně také v dalších oborech souvisejících s materiálovým výzkumem. Význam těchto modelů tkví zejména v představě, že umožní předvídat vlastnosti a optimalizovat složení. Vyžaduje se tedy prognostická schopnost modelu související s možností rozšíření mimo oblast sledovaného složení.

Formálně se funkce $f(s_1 \ldots s_m)$ hledá s využitím metod matematického modelování. Vzhledem k tomu, že neexistuje fyzikální teorie, která by byla východiskem pro nalezení typu modelové funkce se využívá aparátu regrese. Vychází se z lineárního regresního modelu typu

$$P = b_0 + \sum_{j=1}^{m} b_j * s_j \qquad (1)$$

který se dále rozšiřuje a upravuje tak, aby měl postačující predikční schopnosti.

Vzhledem k tomu, že je snahou postihnout nejvýznamnější složky *s* ovlivňující vlastnost *P* je třeba řešit zejména tyto úlohy:

- stanovení vazeb mezi proměnnými s_i $= 1 \ldots m$) za účelem odstranění multikolinearit a parazitních proměnných
- nalezení vazeb mezi vysvětlovanou proměnnou *P* a vysvětlujícími proměnnými s_i za účelem zpřesnění modelu (1), resp. jeho rozšíření o interakce a nelinearity
- posouzení kvality dat s ohledem na omezený rozsah (obsah prvků je omezen jak shora, tak i ze zdola), přítomnost vlivných bodů (vybočující body, extrémy) a případně nenormální rozdělení.

Řada vhodných technik pro řešení těchto úloh je uvedena v knize [1]. V tomto příspěvku jsou popsány pouze vybrané problémy týkající se posuzování kvality dat, které jsou pro konstrukci kvalitního modelu jednou z rozhodujících součástí. Je přehledně pojednáno také o obecném postupu tvorby regresních modelů.

2. Základy regrese

Regresní analýza umožňuje nalezení závislosti výstupní veličiny (odezvy) *y* na nastavované kombinaci hodnot m-tice vstupních proměnných $\boldsymbol{x} = (x_1, x_2..x_m)$.

Vychází se z naměřených hodnot *y* při různých kombinacích nastavovaných proměnných $x_1, x_2, \ldots x_m$. Jde vlastně o n-tici bodů $\{y_i, x_{ij}\}$, $j = 1, ..., m$, $i = 1, ..., n$, vyjádřených ve zkráceném maticovém zápisu $\{\boldsymbol{y}, X\}$. Vektor $\boldsymbol{y}$ má rozměr (n×1) a matice $\boldsymbol{X}$ (n×m). Cílem statistické analýzy je objasnění variability měřené, výstupní ***závisle proměnné*** (vysvětlované) veličiny *y* s využitím regresní funkce $y = f(\boldsymbol{x}, \boldsymbol{\beta})$ obsahující nastavované, vstupní, ***nezávisle proměnné*** (vysvětlující) veličiny $\boldsymbol{x}$. Běžně se předpokládá, že veličina *y* je náhodná a veličiny $\boldsymbol{x}$ jsou nenáhodné a libovolně nastavovatelné. Tento předpoklad je možné akceptovat i pro hutnická data pouze s tím, rozdílem, že obsah jednotlivých složek v materiálu není libovolně nastavitelný. Je neovlivnitelný experimentátorem a jeho velikost je omezena. To může činit problémy zejména při posuzování významnosti přes korelační koeficient, kdy omezení v datech působí výrazně na jeho velikost. Dalším předpokladem je aditivní model měření který lze vyjádřit ve tvaru

$$y_i = f(\mathbf{x_i}, \mathbf{b}) + \varepsilon_i \qquad (2)$$

kde ε_i jsou náhodné veličiny.

Omezme se na lineární regresní modely, kde je regresní model lineární v parametrech a obyčejně je přímo lineární kombinací vysvětlujících proměnných. Podmíněná střední hodnota proměnné *y* pro dané $\boldsymbol{x}$ (regrese) je pak ve tvaru

$$E(y/x) = \sum_{i=1}^{m} \beta_j x_j \qquad (1a)$$

Je patrné, že tomuto modelu vyhovuje také rov. (1) výchozí pro hledání vztahu mezi složením a vlastnostmi materiálů. Odhady $\boldsymbol{b}$ parametrů $\boldsymbol{\beta}$ je pak možné určit metodou nejmenších čtverců, která bývá v praxi nejpoužívanější. Ukažme si geometrický význam této metody.

V případě platnosti aditivního modelu měření pro lineární regresní model je možné zapsat výsledky experimentů jednoduše s pomocí lineární kombinace sloupcových vektorů.

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdot & x_{1m} \\ x_{21} & x_{22} & \cdot & x_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \qquad (3)$$

(nx1) ***(nxm)*** ***(mx1)*** ***(nx1)***

Sloupce x_j matice ***X*** definují z geometrického hlediska m-rozměrný souřadnicový systém resp. nadrovinu L v *n*-rozměrném eukleidovském prostoru E^n. Vektor ***y*** obecně neleží v nadrovině L, (viz. obr. 1 pro případ dvou nezávisle proměnných m = 2). V nadrovině L však leží všechny lineární kombinace sloupců matice ***X*** tj. vektory $\mathbf{X}\,\boldsymbol{\beta}$. Parametry $\boldsymbol{\beta}$ lze tedy chápat jako koeficienty úměrnosti u jednotlivých složek x_j souřadnicového systému (vysvětlujících proměnných) jejichž lineární kombinace tvoří regresní model. Bez ohledu na užité kritérium regrese bude tedy u lineárních regresních modelů ležet modelová funkce $\mathbf{X}\,\mathbf{b}$ stejně jako teoretický model $\mathbf{X}\,\boldsymbol{\beta}$ v m-rozměrné nadrovině L.

Metoda nejmenších čtverců (MNČ) hledá odhady parametrů ***b*** tak, aby byla minimalizována vzdálenost mezi vektorem ***y*** a nadrovinou L. To je ekvivalentní požadavku minimální délky vektoru reziduí

$$\mathbf{e} = \mathbf{y} - \mathbf{y_P} \tag{4}$$

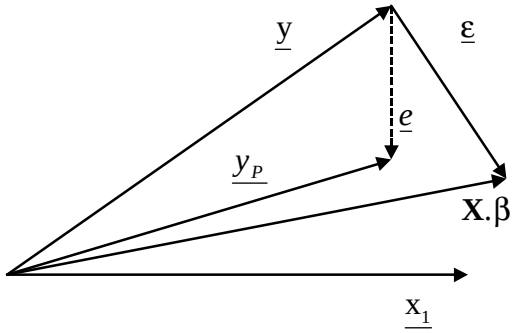
kde $\mathbf{y_p} = \mathbf{X}\,\mathbf{b}$ je vektor predikce. V eukleidovském prostoru lze délku vektoru e vyjádřit vztahem

$$d = \sqrt{\sum_{i=1}^{n} e_i^2} \tag{5}$$

Čtverec délky vektoru ***e*** je tedy číselně shodný s hodnotou kriteriální podmínky ***S(b)*** metody nejmenších čtverců. Odhady modelových parametrů ***b*** pak minimalizují výraz

$$S(\mathbf{b}) = \sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{m} x_{ij} b_j \right]^2 \tag{6}$$

Vektory $\mathbf{e}$ a $\mathbf{y}_P$ jsou znázorněny na obr.1. Vektor $\mathbf{y}_P$ nazývaný ***vektor predikce*** představuje ***kolmou projekci*** vektoru ***y*** do nadroviny L. Vektor $\mathbf{e}$ nazývaný ***vektor reziduí*** leží v (n-m) rozměrné nadrovině L*, ***kolmé*** na nadrovinu L.



Obr. 1 Geometrie lineárního regresního modelu

Na základě tohoto geometrického znázornění lze hledat odhady parametrů ***b*** tak, aby byla minimalizována vzdálenost mezi vektorem ***y*** a nadrovinou L. Je patrné, že vektor reziduí **e** je kolmý na všechny sloupce matice **X**, a proto jsou odpovídající skalární součiny nulové.Tuto

soustavu podmínek lze zapsat maticově jako

$$\mathbf{X}^{\mathbf{T}}\mathbf{e} = 0 \qquad (7)$$

Po dosazení za **e** = **y - X b** a úpravě vyjde odhad **b**, minimalizující vzdálenost *d* ve tvaru

$$\mathbf{b} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \qquad (8)$$

kde symbol $\mathbf{A}^{-1}$ označuje inverzi matice **A**. Z rovnice (8) lze určit tvar projekční matice **H** pomocí které se promítá vektor ***y*** do nadroviny L**.** Tedy

$$\mathbf{y_P} = \mathbf{H}\,\mathbf{y} \qquad (9)$$

Pomocí vektoru **b** lze vyjádřit rovnici (9) ve tvaru

$$\mathbf{y_P} = \mathbf{X}\,\mathbf{b} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \qquad (10)$$

Projekční matice $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}$ má tu vlastnost, že promítne libovolný vektor **V** do roviny L. Projekční matice **P** pro kolmou projekci do nadroviny L*, kolmé na nadrovinu L má tvar

$$\mathbf{P} = \mathbf{E} - \mathbf{H} \qquad (11)$$

kde **E** je jednotková matice. S využitím těchto projekčních matic lze provést rozklad vektoru **y** do dvou složek

$$\mathbf{y} = \mathbf{H}\mathbf{y} + \mathbf{P}\mathbf{y} = \mathbf{y_P} + \mathbf{e}$$

Geometricky to znamená, že vektor ***y*** byl rozložen na dva vzájemně kolmé vektory. Jeden souvisí s částí variability objasněné regresním modelem a druhý se zbytkovou (reziduální variabilitou). Ke stejným vztahům lze dospět analytickou minimalizací kritéria MNČ, tzn. derivováním rovnice (6) a dalšími úpravami.

Pro určení statistických vlastností náhodných vektorů $\mathbf{y_P}$, **e** resp. ***b*** se užívá ***předpokladů***, za kterých má metoda nejmenších čtverců (MNČ) optimální vlastnosti [1]:

I. Regresní parametry $\boldsymbol{\beta}$ mohou nabývat libovolných hodnot. V praxi však často existují omezení parametrů, která vycházejí z jejich fyzikálního smyslu.

II. Regresní model je lineární v parametrech a platí aditivní model měření (2).

III. Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných ***X*** má hodnost rovnou právě m. To znamená, že žádné její dva sloupce $\boldsymbol{x_j}$, $\boldsymbol{x_k}$ nejsou ***kolineární***, tj. rovnoběžné vektory. Tomu odpovídá i formulace, že matice $\boldsymbol{X^TX}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula. Z geometrického hlediska to znamená, že rovina L je m-rozměrná a vektory ***X b*** jsou jednoznačně určeny. Jednoznačné jsou i odhady ***b*** parametrů $\boldsymbol{\beta}$, stanovené metodou nejmenších čtverců.

IV. Náhodné chyby ε_i mají nulovou střední hodnotu $E(\varepsilon_i) = 0$. To musí u korelačních modelů platit vždy. U regresních modelů se může stát, že $E(\varepsilon_i) = K$, $i = 1, ..., n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení však bude $E(\varepsilon'_i) = 0$, kde $\varepsilon'_i = y_i - y_{P,i} - K$. Modely typu (1a) obsahují absolutní člen, pokud je poslední proměnná $x_{im} = 1$ pro všechna $i = 1, ..., n$. Poslední sloupec matice **X** obsahuje tedy samé jedničky a b_m představuje absolutní člen.

V. Náhodné chyby ε_i mají konstantní a konečný rozptyl $E(\varepsilon_i^2) = \sigma^2$. Také podmíněný rozptyl $D(y/\boldsymbol{x}) = \sigma^2$ je konstantní a jde o ***homoskedastický*** případ.

VI. Náhodné chyby ε_i jsou vzájemně nekorelované a platí $cov(\varepsilon_i\ \varepsilon_j) = E(\varepsilon_i\ \varepsilon_j) = 0$. Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin y.

VII. Chyby ε_i mají normální rozdělení $N(0, \sigma^2)$. Vektor $\boldsymbol{y}$ má pak vícerozměrné normální rozdělení se střední hodnotou $\boldsymbol{X\beta}$ a kovarianční maticí $\sigma^2\boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice.

Pokud platí prvních šest předpokladů, jsou odhady $\boldsymbol{b}$, získané minimalizací kritéria nejmenších čtverců, ***nejlepší nevychýlené lineární odhady*** regresních parametrů:

Nejlepší odhady **b** jsou proto, že jejich libovolná lineární kombinace má ***nejmenší*** rozptyl ze všech lineárních nevychýlených odhadů. Znamená to, že i jednotlivé rozptyly odhadů $D(b_j)$ jsou minimální ze všech možných lineárních nevychýlených odhadů (Gaussova-Markova věta). Je třeba poznamenat, že existují vychýlené odhady, jejichž rozptyly jsou menší než rozptyly odhadů $D(b_j)$.

Nevychýlené odhady b jsou proto, že platí $E(\boldsymbol{\beta} - \boldsymbol{b}) = 0$, což znamená, že střední hodnota vektoru odhadů $E(\boldsymbol{b})$ je rovna vektoru regresních parametrů $\boldsymbol{\beta}$.

Lineární odhady **b** jsou proto, že je lze zapsat jako lineární kombinaci měření $\boldsymbol{y}$ s váhami Q_{ij}, které závisí pouze na polohách proměnných x_j, $j = 1, ..., m$. Za jistých předpokladů o matici $\boldsymbol{X}$ navíc platí, že odhady $\boldsymbol{b}$ mají asymptoticky vícerozměrné normální rozdělení s kovarianční maticí

$$D(\mathbf{b}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} \qquad (12)$$

V případě, že platí také předpoklad VII., mají odhady $\boldsymbol{b}$ normální rozdělení už pro konečné rozsahy výběru n.

Protože je matice **H** nenáhodná, platí pro ***kovarianční matici predikce*** vztah

$$D(\mathbf{y_P}) = \sigma^2\mathbf{H} \qquad (13)$$

a analogicky platí pro ***kovarianční matici reziduí*** vztah

$$D(\mathbf{e}) = \sigma^2\mathbf{P} = \sigma^2(\mathbf{E} - \mathbf{H}) \qquad (14)$$

Oba vztahy vyplývají z důležitých vlastností projekčních matic, tj. ***idempotentnosti***, kdy **H** = **H H** a ***symetrie***, kdy $\mathbf{H} = \mathbf{H}^T$. Součet čtverců reziduí RSC lze napsat ve tvaru

$$\mathbf{RSC} = S(\mathbf{b}) = \mathbf{e}^{\mathbf{T}}\mathbf{e} = \mathbf{y}^{\mathbf{T}}(\mathbf{E} - \mathbf{H})\mathbf{y} = \mathbf{y}^{\mathbf{T}}\mathbf{P}\mathbf{y}$$

a pro jeho střední hodnotu platí, že

$$E(\mathbf{RSC}) = \sigma^2 \; tr(\mathbf{P}) = \sigma^2 \; (n - m) \qquad (15)$$

kde tr(**P**) je stopa matice **P**. Ta je vzhledem k idempotentnosti a symetrii matice **P** rovna její hodnosti. Pro nestranný odhad s^2 rozptylu chyb σ^2 lze tedy využít reziduální rozptyl

$$s^2 = \frac{S(\mathbf{b})}{n-m} = \frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{n-m}$$

Při použití odhadů parametrů ***b*** je třeba mít na paměti, že jde o bodové odhady parametrů $\boldsymbol{\beta}$. Tyto bodové odhady jsou náhodné veličiny, a mají proto pro praxi menší význam. Důležitější jsou ***konfidenční oblasti***, nazývané také oblasti nebo intervaly spolehlivosti, ve kterých leží teoretická hodnota $\boldsymbol{\beta}$ se zvolenou pravděpodobností (1-α). Stejně jako u jednorozměrných výběrů, se volí hladina významnosti α = 0.05 nebo 0.01. Této volbě odpovídají 95 %ní nebo 99 %ní intervaly (oblasti) spolehlivosti.

Při konstrukci intervalů spolehlivosti se vychází ze skutečnosti, že náhodná veličina (n - m) s^2 / σ^2 má χ^2 rozdělení s (n - m) stupni volnosti a náhodná veličina $(\mathbf{b} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\mathbf{b} - \boldsymbol{\beta}) / \sigma^2$ má χ^2-rozdělení s m stupni volnosti. Podíl těchto veličin korigovaný stupni volnosti má F-rozdělení s m a (n - m) stupni volnosti. Pro hranice 100×(1-α) %ního intervalu spolehlivosti pak vyjde

$$(\mathbf{b} - \beta)^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} \mathbf{X} (\mathbf{b} - \beta) = m s^2 F_{1-\alpha}(m, n-m) \qquad (17)$$

kde $F_{1-\alpha}$(m, n - m) je (1-α) kvantil F-rozdělení s m a (n - m) stupni volnosti. Vzhledem k tomu, že matice $\boldsymbol{X}^T\boldsymbol{X}$ je regulární, definuje rov. (17) hyperelipsoid, jehož osy jsou orientovány do směrů vlastních vektorů $\boldsymbol{V_j}$ matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Délky jednotlivých poloos jsou rovny p $\sqrt{\lambda_j}$, kde λ_j jsou vlastní čísla matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ a

$$p^2 = m \, s^2 F_{1-\alpha}(m, n-m) \qquad (18)$$

Jak je patrné, jsou jak odhady parametrů, tak i další statistické charakteristiky regrese závislé jak na hodnotách *y* tak i ***X***.

Metoda nejmenších čtverců poskytuje správné výsledky jenom při současném splnění předpokladů o datech a o regresním modelu. K ověřování těchto předpokladů se používá ***regresní diagnostika***, která zahrnuje :

1. Metody pro průzkumovou analýzu jednotlivých proměnných
2. Metody pro analýzu vlivných bodů.
3. Metody pro odhalení porušení předpokladu metody nejmenších čtverců

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu a jsou přitom odhaleny typy odchylek od ideálního regresního tripletu „ data - model - metoda odhadu“.

3 Průzkumová analýza dat

Účelem průzkumové analýzy je zkoumání statistických zvláštností v datech, Problémem použití těchto metod v regresi je to, že jde o strukturovaná data s vazbami vyjádřenými regresní funkcí. O metodách průzkumové analýzy jednorozměrných dat je detailně pojednáno v knize [1]. V regresní analýze se vybraných postupů průzkumové analýzy používá pro:

a) určení statistických zvláštností jednotlivých proměnných nebo reziduí,
b) posouzení "párových" vztahů mezi všemi sledovanými proměnnými,
c) ověření předpokladu o rozdělení proměnných nebo reziduí.

V řadě případù již pouhé vynesení naměřené veličiny y_i proti indexu i může odhalit ***skrytou proměnnou***, často související s časem nebo pořadím měření.

K orientačnímu posouzení vztahů mezi jednotlivými proměnnými se užívá rozptylových grafů, kde se na osy vynášejí přímo hodnoty sledovaných proměnných. Informace o multikolinearitě lze získat vynesením dvojic vysvětlujících proměnných x_j proti x_k,. Přibližně lineární závislost zde indikuje silnou multikolinearitu. Na druhé straně však může vést vynášení y proti x_j, j = 1, ..., m, i k mylným závěrům o nelinearitě modelu, který je ve skutečnosti lineární.

K ověření normality dat se často používá Q-Q grafů [1]. Mezi základní techniky průzkumové analýzy patří i stanovení rozsahu a rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. K tomu lze využít grafù rozptýlení s kvantily a řady dalších postupů [1]. Přes svoji jednoduchost umožňuje průzkumová analýza identifikovat ještě před vlastní regresní analýzou:

1. ***nevhodnost dat*** jako důsledek malého rozmezí nebo přítomnosti vybočujících bodů,
2. ***nesprávnost navrženého modelu*** (skryté proměnné),
3. ***multikolinearitu*** (přibližně lineární vztah mezi sloupci matice ***X***)
4. ***nenormalitu*** v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

4. Posouzení kvality dat

Kvalita dat úzce souvisí s použitým regresním modelem. Při posuzování se sleduje především výskyt ***vlivných bodů*** (VB), které jsou hlavním zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních odhadů parametrů. Ve zvláštních případech však vlivné body zlepšují predikční schopnosti modelů.

Vlivné body silně ovlivňují většinu výsledků regrese. Lze je rozdělit do tří základních skupin:

a) ***Hrubé chyby***, které jsou způsobeny měřenou veličinou (***vybočující pozorování***) nebo nevhodným nastavením vysvětlujících proměnných (***extrémy***). Jsou obyčejně důsledkem chyb při manipulaci s daty.

b) ***Body s vysokým vlivem*** (tzv. golden points) jsou speciálně vybrané body, které byly přesně změřeny, a které obvykle rozšiřují predikční schopnosti modelu.

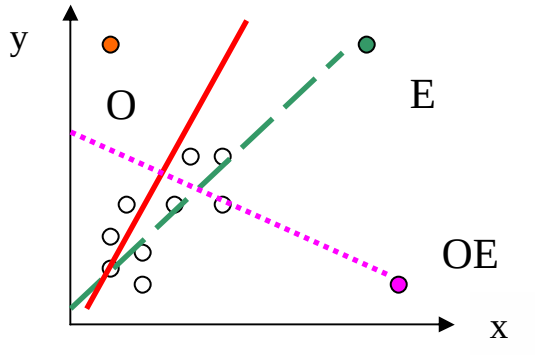
c) ***Zdánlivě vlivné body*** vznikají jako důsledek nesprávně navrženého regresního modelu.

Podle složky dat, ve které se vlivné body vyskytují, lze provést dělení na:

1. ***vybočující pozorování*** (outliers O), které se na ose y výrazně liší od ostatních,
2. ***extrémy*** (high leverage points E), které se liší v hodnotách na ose x, nebo v jejich

kombinaci (v případě multikolinearity) od ostatních bodů.

Vyskytují se však i body, které jsou jak vybočující tak i extrémní (OE). O jejich výsledném vlivu však především rozhoduje to, že jsou extrémy.



Obr. 2 Vliv vybočujícího bodu (O plná čára), extrému (E čárkovaná čára) a kombinace (OE tečkovaná čára) na průběh regresní přímky určené MNČ

K identifikaci vlivných bodů typu vybočujícího pozorování se využívá zejména reziduí a k identifikaci ***extrémů*** pak diagonálních prvků H_{ii} projekční matice **H**.

Obecnější charakteristiky vlivných bodů jsou funkcí **reziduí** e_i a **diagonálních prvků projekční matice H_{ii}** s faktorem souvisejícím s počtem bodů *n* a počtem proměnných *m*.

5. Statistická analýza reziduí

Rezidua jsou základem pro identifikaci podezřelých bodů a nekorektnosti navrženého regresního modelu. Při jejich interpretaci se však vyskytuje řada chyb a nepřesností.
Statistická analýza reziduí $e_i = y_i - \boldsymbol{x_i b}$, kde $\boldsymbol{x_i}$ je i-tý řádek v matici **X**, vychází z předpokladu, že jde o odhady chyb ε_i. Nesprávné představy o klasických reziduích jsou, že:

1. rozdělení reziduí je stejné jako rozdělení chyb a statistické vlastnosti reziduí jsou shodné s vlastnosti chyb
2. čím je reziduum e_i vetší, tím je daný bod vlivnější, a tím spíše by se měl z dat vyloučit.

Z geometrie na obr.1 plyne, že rezidua e_i ***nejsou nezávislá***, i když chyby jsou nezávislé. Jde totiž o projekci vektoru **y** do podprostoru rozměru *(n - m)*. S využitím projekční matice **P** lze psát, že

$$e = Py = P(X\beta + \varepsilon) = P\varepsilon \qquad (19)$$

Při úpravě rovnice (19) bylo využito faktu, že vektor $\boldsymbol{X\,b}$ leží v rovině kolmé na rovinu, do které se provádí projekce, takže výsledkem je nulový vektor.Pro i-té reziduum vyjde

$$e_i = (1 - H_{ii})\, y_i - \sum_{j\_i}^{n} H_{ij}\, y_j = (1 - H_{ii})\, \varepsilon_i - \sum_{j\_i}^{n} H_{ij}\, \varepsilon_j \qquad (20)$$

Každé reziduum e_i je tedy lineární kombinací všech chyb ε_i. Rozdělení reziduí je obecně závislé na rozdělení chyb, na prvcích projekční matice ***H*** a na velikosti výběru *n*.

Protože je reziduum e_i součtem náhodných veličin s ohraničeným rozptylem, projevuje se zejména u menších výběrů tzv. ***efekt supernormality***. To znamená, že i když chyby ε nemají normální rozdělení, vychází rozdělení reziduí blízké normálnímu. U menších výběrů jsou prvky projekční matice ***H*** veliké a převažující roli hraje součet členů $H_{ij}\varepsilon_j$. Rozdělení tohoto součtu se více blíží normalitě než rozdělení původních chyb ε. U dostatečně velikých výběrů, kdy 1/n je blízké 0 je $e_i \approx \varepsilon_i$ a analýza rozdělení reziduí podává informace o rozdělení chyb.

Pro rozptyl reziduí platí

$$D(e_i) = (1 - H_{ii})s^2 \qquad (21)$$

Rozptyl reziduí $D(e_i)$ je tedy ***nekonstantní***, i když rozptyl chyb je konstantní. Pro párový korelační koeficient r_{ij} mezi dvěma rezidui e_i a e_j platí

$$r_{ij} = \frac{-H_{ij}}{\sqrt{(1-H_{ii})(1-H_{jj})}}$$

Je tedy patrné,že rezidua jsou ***korelovaná***, i když chyby ε_i a ε_j jsou nezávislé.

Pro silně extrémní body platí, že diagonální prvky $H_{ii} \to 1$, zatímco všechny nediagonální prvky $H_{ij} \to 0$. Z rovnice (20) pak ovšem plyne, že $e_i = 0$ je ***bez ohledu*** na velikost y_i. Rezidua proto ***neindikují*** vždy správně silně odchýlené hodnoty.

Klasická rezidua jsou tedy korelovaná, s nekonstantním rozptylem, jeví se normálnější a nemusí indikovat silně odchýlené body.

V odborné literatuře se často doporučuje užívání normovaných reziduí $e_{Ni} = e_i / s$, o kterých se soudí, že to jsou normálně rozdělené veličiny s nulovou střední hodnotou a jednotkovým rozptylem $e_{Ni} \sim N(0, 1)$. K ***vyjádření*** jejich vlivu se používá pravidla 3s, tj. hodnoty větší než ± 3s jsou považovány za vybočující. Pro případ normálního rozdělení leží za hranicí $x_A \pm 3s$ pouze 0.3 % hodnot.

Rozptyl $D(e_{Ni}) = (1 - H_{ii})$ není ani konstantní, ani jednotkový. Navíc bylo ukázáno, že pro silně vlivné extrémní body je $e_i \to 0$, takže užití pravidla ± 3s může vést i k vylučování správných dat při zachování chybných hodnot.

Konstantní rozptyl mají teprve ***standardizovaná rezidua*** e_{Si}, která vzniknou dělením reziduí jejich směrodatnou odchylkou s, tedy

$$e_{Si} = \frac{e_i}{s\sqrt{1-H_{ii}}} \qquad (22)$$

Vlastnosti standardizovaných reziduí e_{Si} jsou téměř stejné jako klasických reziduí e_i. Maximální hodnota e_{Si} je $\sqrt{n-m}$.

Veličina $e^2_{Si}/(n - m)$ má beta-rozdělení $Be[0.5; (n - m - 1) / 2]$.

Pokud se v rov. (22) pro výpočet standardizovaného rezidua e_{Si} použije místo odhadu *s* odhadu směrodatné odchylky $s_{(-i)}$, získané při vynecháním i-tého bodu, resultují ***plně Studentizované***, resp. ***Jackknife rezidua*** e_{Ji}

$$e_{Ji} = e_{Si}\sqrt{\frac{n - m - 1}{n - m - e_{Si}^2}} = \sqrt{n - m}\,\mathrm{cotg}\,\Theta_i \qquad (23)$$

Tato rezidua mají za předpokladu normality chyb Studentovo rozdělení s (n - m - 1) stupni volnosti. Odpovídají testovací statistice Studentova t-testu nulové hypotézy H_0: C = 0 v modelu jednoduchého posunutí

$$y = X\beta + i * C + \varepsilon \qquad (24)$$

kde ***i*** je jednotkový vektor, obsahující jako i-tý prvek jedničku a ostatní prvky jsou nulové. Model (24) vystihuje nejen případ vybočujícího měření, kde C je přímo velikost vychýlení, ale i případ extrému, kdy je C = a $\mathbf{d}_i$ je vektor vychýlení jednotlivých x-ových složek i-tého bodu. Jackknife rezidua jsou běžně využívána místo klasických reziduí e_i k identifikaci vybočujících bodů.. Ani tato rezidua však nemusí být spolehlivá v případě extrémů. Další skupiny reziduí jsou popsány v práci [1].

6. Analýza prvků projekční matice

Analýza prvků projekční matice hraje v regresní diagnostice důležitou roli. Diagonální prvky této matice $H_{ii} = x^T_i\,(\boldsymbol{X}^T\,\boldsymbol{X})^{-1}x_i$ indikují přítomnost extrémních bodů, které nejsou zachyceny analýzou reziduí. Diagonální prvky H_{ii} mají řadu vlastností, plynoucích ze symetrie a idempotentnosti matice ***H***:

1. Z vlastností projekční matice ***H*** přímo plyne podmínka pro diagonální prvky $0 < H_{ii} < 1$ a prvky mimo diagonálu $-1 \le H_{ij} \le 1$. Pokud model obsahuje absolutní člen a hodnost matice ***X*** je *m*, platí pro diagonální prvky podmínka $1/n \le H_{ii} \le 1/C$, kde *C* je počet opakování měření t.j. opakování i-tého řádku matice ***X***.

2. Pro model s absolutním členem a plnou hodností matice **X** platí, že

$$\sum_{i=1}^{n} H_{ii} = 1\,, \qquad \sum_{i=1}^{n} H_{ij} = 1$$

a průměrná hodnota diagonálního prvku je $H_{ii} = m/n$.

3. Z idempotentnosti matice **H** plyne, že $H_{ii} = H_{ii}^2 + \sum_{j\neq i}^{n} H_{ij}^2 = \sum_{j=1}^{n} H_{ij}^2$

Z těchto rovností vyplývají dvě důležité vlastnosti diagonálních prvků H_{ii}:

a) pokud jsou diagonální prvky blízké nule, $H_{ii} \to 0$, jsou i všechny mimodiagonální prvky blízké nule $H_{ij} \to 0$, pro j = 1, ..., n;

b) pokud jsou diagonální prvky blízké jedné, $H_{ii} \to 1$, jsou všechny mimodiagonální prvky blízké nule, $H_{ij} \to 0$, pro j = 1, ..., n.

4. Jestliže matice **X** pochází z vícerozměrného normálního rozdělení, má veličina

$$F = (n - m)\,[H_{ii} - 1/n]\,[(1 - H_{ii})\,(m - 1)]$$

F-rozdělení F (m - 1, n - m).

5. Čím jsou diagonální prvky H_{ii} vyšší, tím více ovlivňuje i-tý bod predikci y_{Pi}. Jsou-li hodnotou H_{ii} blízké jedné $H_{ii} \to 1$, je $y_{Pi} = y_i$ a veškerá variabilita v místě $\boldsymbol{x}_i$ je objasněna regresním modelem.(viz tečkovaná a čárkovaná čára na obr. 2)

6. Diagonální prvky $H_{ii} = d\,y_{Pi} / dy_i$ vyjadřují citlivost predikce y_{Pi}. na změnu hodnoty y_i. Jejich nulová hodnota $H_{ii} = 0$ potom indikuje bod, který nemá žádný vliv na predikci.

7. Diagonální prvky H_{ii} jsou neklesající funkcí počtu vysvětlujících proměnných *m*

a nerostoucí funkcí počtu bodů *n*.

8. Čím je bod $\boldsymbol{x}_i$ vzdálenější od těžiště ostatních bodů, tím více se bude jevit extrémní, a tím více poroste i hodnota diagonálních prvků H_{ii}.

9. Pokud mají vysvětlující proměnné $\boldsymbol{x}$ normální rozdělení, platí pro velké poèty bodù n, že n H_{ii} - 1 má přibližně χ^2_m (2) rozdělení.

Pro komplexnější analýzu je vhodné provést rozšíření matice $\boldsymbol{X}$ o vektor $\boldsymbol{y}$, takže vznikne matice $\boldsymbol{X}^* = (\boldsymbol{X}\, \boldsymbol{y})$. Této matici odpovídá projekční matice

$$\mathrm{H}^* = \mathrm{H} + \frac{\mathrm{e}\,\mathrm{e}^{\mathrm{T}}}{\mathrm{e}^{\mathrm{T}}\mathrm{e}}$$

Protože matice $\mathbf{H}^*$ obsahuje informace o všech datech, je vhodná jako celková míra vlivných bodù. Pro diagonální prvky této matice platí vztah

$$\mathrm{H}^*_{\mathrm{ii}} = \mathrm{H}_{\mathrm{ii}} + \frac{\mathrm{e}_{\mathrm{i}}^2}{(\mathrm{n} - \mathrm{m})\, s^2}$$

Pro grafické znázornění se používá ***indexový graf*** prvkù H_{ii} proti indexu *i*.

7. Charakteristiky vlivných bodù

Při posuzování vlivných bodů je třeba mít na paměti, že mohou nestejně výrazně ovlivňovat různé charakteristiky regrese. Například, body ovlivňující výrazně predikci y_{Pi} nemusí být z hlediska rozptylu parametrů vůbec vlivné. Stupeň vlivu jednotlivých bodů je třeba posuzovat vždy s ohledem na to, které charakteristiky regrese ovlivňují. K identifikaci vlivných bodů existuje řada dalších diagnostik, které lze rozdělit dle dvou základních skupin

.

Zvětšený rozptyl

Tento přístup vychází z platnosti lineárního regresního modelu (1a) se speciální strukturou rozptylů chyb. Pro i-tou chybu ε_i platí, že má normální rozdělení N(0, s^2 / w_i), zatímco ostatní chyby ε_j, $j \neq i$, mají normální rozdělení N(0, s^2) s konstantním rozptylem. Váhový parametr w_i leží v intervalu $0 < w_i < 1$. Takový model působení vlivných bodù se označuje jako ***model zvětšeného rozptylu*** (inflated variance).

Pro $w_i = 1$ se jedná o klasickou metodu nejmenších čtverců. Označme $\mathbf{b}(w_i)$ odhad parametrù **b**, určený MNČ pro případ, že rozptyl i-té chyby je roven právě s^2/w_i. Pak platí

$$\mathrm{b}(1) - \mathrm{b}(w_i) = \frac{(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{x}_i(1-w_i)e_i}{1-(1-w_i)\mathrm{H}_{ii}} \qquad (25)$$

kde $\mathbf{x}_i$ je i-tý řádek matice $\boldsymbol{X}$, který obsahuje x-ové složky i-tého bodu. Pro w = 0 vyjde z rovnice (25), že $\mathbf{b}(1) - \mathbf{b}(0) = \mathbf{b} - \mathbf{b}_{(i)}$, kde $\mathbf{b}_{(i)}$ je odhad získaný metodou nejmenších čtverců ze všech bodù kromě i-tého. Vynechání i-tého bodu je tedy stejné, jako když má tento bod neohraničený, tj. nekonečný rozptyl.

Vypouštění bodů

Tento přístup je založen na sledování změn charakteristik regrese, ke kterým dojde při vypuštění jednotlivých bodů nebo jejich skupin. Je snahou používat vhodné skalární míry regresních charakteristik, které se snadno interpretují a graficky znázorňují. Nejznámější

skalární míra je ***Cookova vzdálenost*** D_i související s konfidenčním elipsoidem odhadů.

$$D_i = \frac{(b - b_{(i)})^T X^T X (b - b_{(i)})}{m s^2} = \frac{e_{Si}}{m} \frac{H_{ii}}{1 - H_{ii}} \tag{26}$$

To umožňuje její porovnání s kvantily F-rozdělení. Jde zde však o posun odhadů, který vznikl vynecháním i-tého bodu. Orientačně platí, že pro $D_i > 1$ posun přesahuje 50%ní konfidenční oblast a daný bod je proto vlivný. Další možné vysvětlení Cookovy vzdálenosti D_i vychází z toho, že jde o Eukleidovskou vzdálenost mezi vektorem predikce y_P metody nejmenších čtverců a vektorem predikce $y_{P(i)}$, který odpovídá odhadům metodou nejmenších čtverců při vynechání i-tého bodu. Cookova vzdálenost D_i vyjadřuje vliv i-tého bodu pouze na odhady parametrů ***b***. Pokud tedy i-tý bod neovlivní odhady regresních parametrů ***b*** výrazně, bude hodnota Cookovy vzdálenosti D_i malá. Takový bod však může silně ovlivnit odhad reziduálního rozptylu s^2.

K vyjádření relativní změny odhadů parametrů, způsobené vynecháním i-tého bodu je možné užít standardizovaných odchylek j-tého odhadu b_j od téhož odhadu $b_{(i)j}$, získaného při vynechání i-tého bodu. Odpovídající diagnostika má tvar

$$DS_{ij} = \frac{b_j - b_{(i)j}}{s_{(i)} \sqrt{V_{ii}}} \tag{27}$$

kde V_{ii} je diagonální prvek matice $\boldsymbol{X}^T\boldsymbol{X}$. Vliv i-tého bodu na odhad j-tého regresního parametru je významný, pokud je $DS > 2 / \sqrt{2}$.

Andrewsova-Pregibonova diagnostika AP_i vyjadřuje vliv i-tého bodu na změnu objemu konfidenčního elipsoidu

$$AP_i = \frac{\det(X^{*T}_{(i)} X^*_{(i)})}{\det(X^{*T} X^*)} \tag{28}$$

kde $\boldsymbol{X}^* = (\boldsymbol{X}\ \boldsymbol{y})$ je matice $\boldsymbol{X}$ rozšířená o vektor $\boldsymbol{y}$. Diagnostika AP_i souvisí s prvky rozšířené projekční matice $\mathbf{H}^*$ vztahem

$$AP_i = 1 - H_{ii} - e^2_{Ni} = 1 - H^*_{ii} \tag{29}$$

Za výrazně vlivné se považují body, pro které je $H^*_{ii} = (1 - AP_i) > 2\,(m + 1) / n$.

K unifikovanému vyjádření vlivných bodů se používá ***věrohodnostní vzdálenost*** LD_i definovanou výrazem

$$LD_i = 2(L(\Theta) - L(\Theta_{(i)}))$$

kde $L(\Theta)$ je maximum logaritmu věrohodnostní funkce při použití všech bodů a $L(\Theta_{(i)})$ je totéž s vynecháním i-tého bodu. Vektor parametrů Θ obsahuje jak odhady regresních parametrů ***b*** tak i rozptylu s^2. Za silně vlivné se považují body, pro které je $LD_i > \chi^2_{1-\alpha}(m + 1)$, kde $\chi^2_{1-\alpha}(m + 1)$ je kvantil χ^2 rozdělení s $(m + 1)$ stupni volnosti.

Pomocí různých variant LD_i lze vyšetřovat vliv i-tého bodu na odhady parametrů, rozptyl

chyb nebo kombinaci obojích.

Pro sledování vlivu jednotlivých bodů pouze na odhady regresních parametrů ***b*** vyjde věrohodnostní vzdálenost ve tvaru

$$LD_I = n \ln\left[\frac{d_I H_{ii}}{1 - H_{ii}} + 1\right]$$

Pro sledování citlivosti odhadu reziduálního rozptylu s^2 na přítomnost vlivných bodů má věrohodnostní vzdálenost tvar

$$\mathrm{LD_i}(s^2) = \mathrm{nln}\frac{\mathrm{n}}{\mathrm{n-1}} + \mathrm{nln}(1 - \mathrm{d_i}) + \frac{\mathrm{d_i}(\mathrm{n-1})}{1 - \mathrm{d_i}} - 1$$

Pro sledování vlivu i-tého bodu na odhady parametrů i rozptylu má věrohodnostní vzdálenost tvar

$$LD_i(b, s^2) = n \ln\left(\frac{n}{n-1}\right) + n \ln(1 - d_i) + \frac{(n-1)d_i}{(1 - d_i)(1 - H_{ii})} - 1$$

V těchto vztazích je

$$\mathrm{d_i} = \frac{\mathrm{e_{si}^2}}{\mathrm{n - m}} \qquad (30)$$

Z rozboru těchto tří variant věrohodnostní vzdálenosti plyne:

a) Diagnostika $LD_i(\boldsymbol{b})$ je monotónní funkcí Cookovy vzdálenosti D_i a v porovnání s ní nepřináší žádné nové poznatky.
b) Diagnostika $LD_i(s^2)$ nezávisí na H_{ii} a nebude tedy ovlivněna extrémními body.
c) Diagnostika $LD_i(\boldsymbol{b}, s^2)$ vystihuje vliv jednotlivých bodů na ***b*** a s^2. Je výhodná zejména pro modely bez absolutního členu. Diagnostika $LD_i(\boldsymbol{b},s^2)$ ohraničuje shora veličiny $LD_i(\boldsymbol{b})$ a $LD_i(s^2)$ a postačuje proto v prvním přiblížení sledovat pouze ji.

Ani veličiny LD_i nejsou zcela univerzální a k vyšetření vlivných bodů se proto užívá kombinace řady různých diagnostik. Z jejich hodnot se usuzuje, zda je nutné dané body z další analýzy vypustit či nikoliv.

K testování vlivu i-tého bodu na součet středních kvadratických chyb odhadů, středních kvadratických chyb predikce a integrální střední kvadratické chyby predikce se doporučuje jako testovací statistika Jackknife reziduum e_{Ji}, které je vhodné jak pro modely jednoduchého posunutí tak i pro modely zvětšeného rozptylu $\mathrm{D}(\varepsilon_i) = s^2 / w_i$.

Pokud se sleduje současně *n* bodů, platí pro model jednoduchého posunutí podmínka

$$\mathrm{e_{Ji}^2} \le \mathrm{F_{1-\alpha/n}}(1, \mathrm{n-m-1}, 0.5)$$

Její splnění pro všechna *i* znamená nepřítomnost vlivných bodů v datech. Veličina $\mathrm{F_{1-\alpha/n}}(1, \mathrm{n - m - 1}, 0.5)$ je $100´(1 - \alpha / \mathrm{n})$ %ní kvantil necentrálního F-rozdělení s parametrem necentrality 0.5 a (1, n - m - 1) stupni volnosti. Pro model zvětšeného rozptylu platí analogicky, že splnění

nerovnosti

$$e_{Ji}^2 \leq 2\, F_{1-\alpha/n}(1, n\ m\ 1)$$

pro všechna *i* znamená nepřítomnost vlivných bodù. Zde $F_{1-}\alpha_{/n}(1, n - m - 1)$ je 100´(1 - α / n) %ní kvantil centrálního F-rozdělení s 1 a (n - m - 1) stupni volnosti. Na základě těchto dvou testů lze definovat orientační pravidlo: silně vlivné body mají čtverce Jackknife reziduí e_{Ji}^2 větší než 10.

K analýze vlivných bodů je vhodné užít také diagnostických grafů:

a) ***Indexové grafy*** (IG) obsahují charakteristiky vlivných bodů v závislosti na indexu *i* daného bodu, stejně jako indexové grafy pro prvky projekční matice H_{ii}, atd. Výhodnější jsou však speciální grafy, které využívají faktu, že všechny charakteristiky vlivných bodù jsou jednoduchými funkcemi reziduí e_i a prvků H_{ii} projekční matice

b) V ***L-R grafech*** se vynáší na osu y čtverce normovaných reziduí $e_{Ni}^2 = e_i^2 / RSC$ a na osu x prvky H_{ii}. Všechny body pak leží pod přeponou v pravoúhlém trojúhelníku s pravým úhlem v počátku souřadnic a přeponou, definovanou limitní rovností $H_{ii} + e_{Ni}^2 = 1$.

Většinu charakteristik vlivných bodů lze vyjádřit ve tvaru $K(m, n)\ f(H_{ii},\ e_{Ni}^2)$, kde $K(m, n)$ je konstanta, závisející jen na m a n. [1].

V praktických aplikacích je problémem, že přítomnost více vlivných bodů se může projevit maskováním nebo překrytím [2]. Diagnostiky simultánního posuzování skupin vlivných bodů lze snadno definovat na základě diagnostik založených na vypouštění bodů.

Nechť $\boldsymbol{I} = (i_1, i_2, \ldots i_k)$ pro $k < (n-m)$ je množina *k* indexů jejichž vliv se má posoudit. S výhodou se využije přeuspořádání tak, že podezřelých k bodů jsou poslední řádky matice ***X*** a vektoru ***y.*** Zaveďme označení

$$X = \begin{pmatrix} X_{(I)} \\ X_I \end{pmatrix} \begin{matrix} (n-k)\ x\ m \\ k\ x\ m \end{matrix} \qquad y = \begin{pmatrix} y_{(I)} \\ y_I \end{pmatrix} \begin{matrix} (n-k)\ x\ 1 \\ k\ x\ 1 \end{matrix} \qquad e = \begin{pmatrix} e_{(I)} \\ e_I \end{pmatrix} \begin{matrix} (n-k)\ x\ 1 \\ k\ x\ 1 \end{matrix}$$

Projekční matice odpovídající podezřelým bodům je pak definována vztahem.

$$H_I = X_I^T (X^T X)^{-1} X_I \qquad (31)$$

Veličina $S_I = e_I^T (E - H_I)^{-1} e_I$ odpovídá snížení reziduálního součtu čtverců vlivem odstranění k tice indexovaných bodů ***I***. Analogií klasických standardizovaných reziduí pro více bodů je veličina

$$e_{SI}^2 = \frac{S_I}{s^2} \qquad (32)$$

Pro skupinu vlivných bodů má Cookova vzdálenost tvar

$$D_I^2 = \frac{e_I^T (e - H_I)^{-1} H_I e_I}{ms^2} \qquad (33)$$

a pro Andrews Pregibonovu statistiku platí

$$\mathrm{AP_I} = 1 - (1 - \frac{\mathrm{e_{SI}^2}}{n-m}) \det(\mathrm{E} - \mathrm{H_I}) \qquad (34)$$

Věrohodnostní vzdálenost $L(b, s^2)$ má pro případ *k* vyloučených bodů tvar

$$\mathrm{LD_I}(b, s^2) = n \ln\left[\frac{n(n-m-\mathrm{e_{SI}^2})}{n-m}\right] + \frac{(n-1)(n-m+m\mathrm{D_I^2})}{n-m-\mathrm{e_{SI}^2}} - n$$

Je patrné, že že při vhodném přeuspořádání indexů lze poměrně snadno nahradit skalár *i* vektorem indexů ***I.***

Dosavadní míry byly vhodné pro vybrané charakteristiky regrese a nepostihovali komplexně vliv bodů na výsledek regrese. Hadi [3] navrhl jednu míru vycházející z předpokladu, že vlivné body mohou vybočovat vzhledem k prostoru proměnných ***x*** a vzhledem k vektoru ***y.*** Kombinací charakteristik vyjadřujících vliv v prostoru ***x*** (vzdálenost podezřelých hodnot od ostatních) a v prostoru ***y*** (chyba predikce) resultuje vztah

$$\mathrm{HA_I^2} = \frac{m}{k} \frac{\mathrm{e_I^T}(\mathrm{E} - \mathrm{H_I})\mathrm{e_I}}{\mathrm{e^T e} - \mathrm{e_I^T e_I}}$$

Pro případ, kdy *k = 1* a *I = (i)* pak vyjde

$$HA_i^2 = \frac{m}{(1-H_{ii})} * \frac{d_i^2}{(1-d_i^2)} + \frac{H_{ii}}{1-H_{ii}} \qquad (31)$$

kde d_I^2 je definováno rov. (30). První člen v rov (31) je funkce *i* tého rezidua a diagonály projekční matice (chyba predikce). Druhý člen se nazývá potenciál. Potenciál reziduový graf (PRG) má na ose x první člen a na ose y druhý člen matice (31). Tedy pro *k = 1* a *I = i* se vynášejí $\frac{H_{ii}}{1-H_{ii}}$ proti $\frac{m}{1-H_{ii}} \frac{d_i^2}{(1-d_i^2)}$.

V tomto grafu jsou extrémy v levém horním rohu a vybočující hodnoty jsou v pravém dolním rohu.

Další diagnostiky vlivných bodů jsou popsány v práci [4]. Zajímavou možností je také kombinace robustních metod s identifikací vlivných bodů [2].

8. Program REGDIA

Na základě výše popsaných charakteristik vlivných bodů byl sestaven program REGDIA v jazyce MATLAB. Tento program počítá základní charakteristiky regrese a diagnostiky založené na vypouštění jednotlivých bodů. Kromě zde uvedených charakteristik jsou v programu obsaženy i další charakteristiky, jejichž popis lze nalézt např. v článku [2]. Je použit také PR graf pro posouzení obecného vlivu jednotlivých bodů na výsledek regrese. Pro řešení odhadu parametrů se užívá interní zabudované funkce. Invertace se provádí pomocí zabudované funkce **inv.** Uživatel může volit model bez nebo s absolutním členem. Jsou k dispozici jak rozsáhlé tabelární výstupy tak i řada grafů.

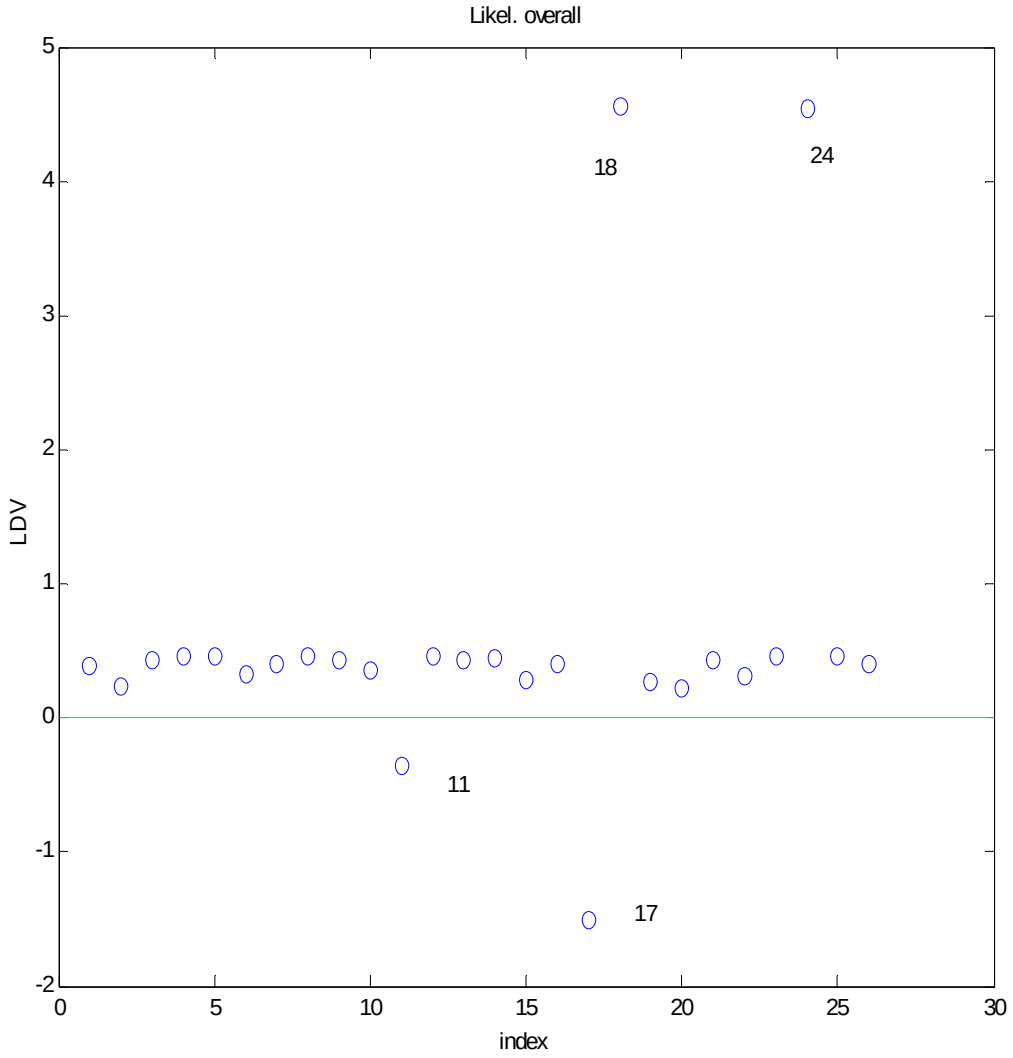
Pro ilustraci činnosti tohoto programu byla použita Hockingova syntetická data [6] určená pro regresní diagnostiku .Počet bodů *n=26 a počet proměnných , m = 4*

Model: Y = a0 + a1*x1 + a2*x2 + a3*x3

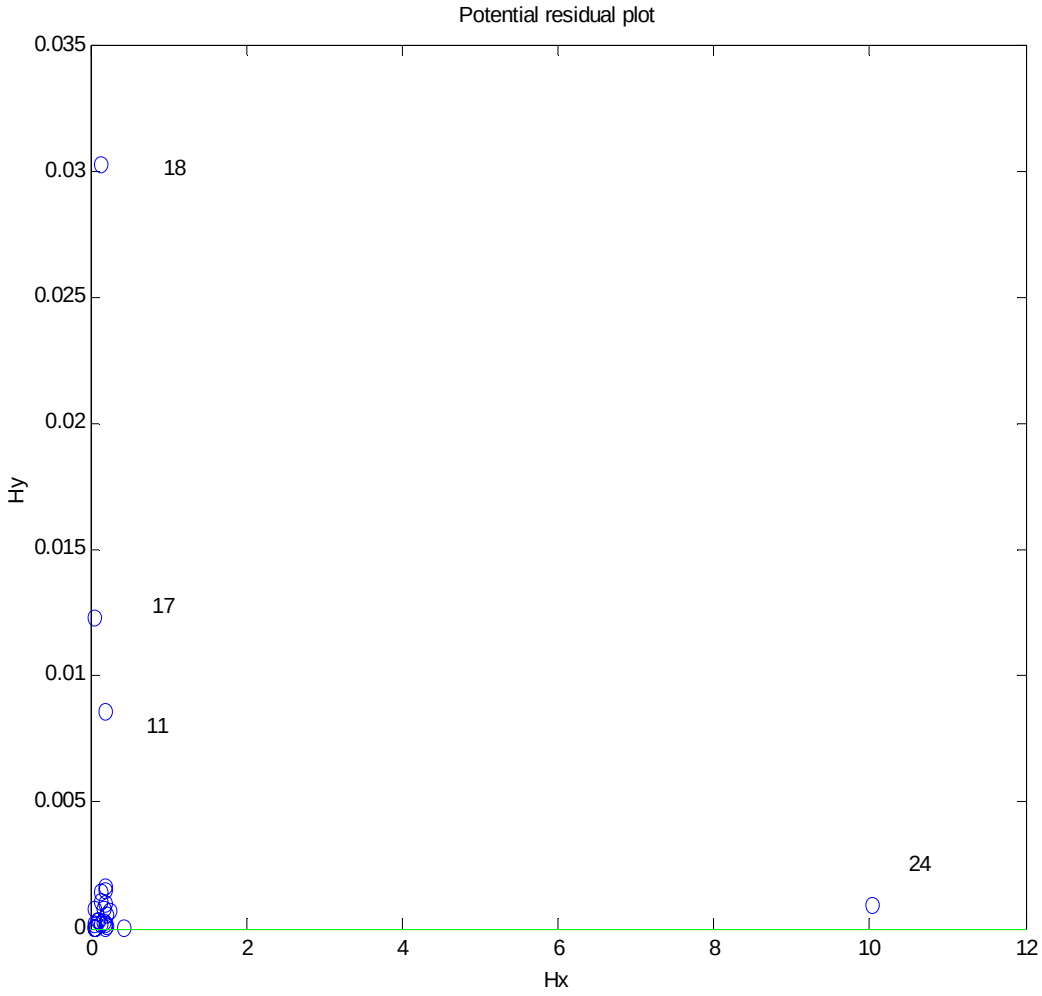
Generace dat : y= 20 + 3*x1 - 2*x2 + eps1 eps1...náhodná čísla z N(0,.25)

multikolinearita: 2*x3 = 60-3*x1 -1.5*x2 + eps2,
eps2... náhodná čísla z N(0,.16)

Vybočující body : č.11,17,18. Extrém : č.24 (leží mimo rovinu multikolinearity). Data byla zpracována programem REGDIA. S ohledem na zaměření této práce byly vybrány dva typické grafické výstupy. Indexový graf pro celkovou věrohodnostní vzdálenost $LD_i(\boldsymbol{b}, s^2)$ a potenciál reziduový graf (PR graf) jsou zobrazeny na obr.3 a 4. Je patrné, že v obou případech byly identifikovány všechny narušující body.



Obr3. Indexový graf pro celkovou věrohodnostní vzdálenost $LD_i(\boldsymbol{b}, s^2)$



Obr. 4 Potenciál reziduový graf

9. Závěr

Byly uvedeny základní myšlenky a souvislosti pro metodu nejmenších čtverců. Byly popsány vybrané metody regresní diagnostiky. Pozornost byla zaměřena především na postupy identifikace vlivných bodů. Byla zmíněno také použití technik průzkumové analýzy dat. Byl uveden program v jazyce MATLAB.

Poděkování:
Tato práce vznikla s podporou výzkumného centra Textil LN00B090

10. Literatura

[1] Meloun M., Militký J.*: Zpracování experimentálních dat*, East Publishing Praha 1998
[2] Militký J., Meloun M*: Vybočující body ve vícerozměrných datech*, Sborník z konference „Zajištění kvality analytických výsledků“, Komorní Lhotka , březen 2002
[3] Hadi A. A.: Comput. Statist. Data Anal. **14,** 1 (1992)
[4] Brown G.P., Lawrance A.J.: Commun. Statist. A29, 2079 (2000)
[5]. Meloun M., Militký J.::Anal. Chim. Acta **439**, 16 (2001)
[6] Hocking R.R.,Pendleton O.J.: Commun.Statist. A12,497 (1983)