IDENTIFIKACE ODLEHLÝCH POZOROVÁNÍ V LINEÁRNÍ REGRESI

Karel Zvára, MFF UK, Praha

V praxi se občas setkáváme s pozorováními, která nějak vybočují z řady. V tomto textu si připomeneme řešení problému odlehlých pozorování v náhodném výběru. Pro regresní model ukážeme různé modifikace reziduí, sloužící k vyhledání či prokázání odlehlých pozorování a zmíníme se o grafických metodách, určených spíše k naší orientaci v datech. Zmíníme se též o jiném přístupu (vlivná pozorování), kdy nás zajímá nejen neočekávaná hodnota závisle proměnné, ale také vliv hodnot nezávisle proměnných daného pozorování.

1. Odlehlá pozorování ve statistice

S pozorováními výrazně se odlišujícími od ostatních se museli vědci vypořádávat již dávno. Například americký astronom Chauvenet (1863) navrhl následující postup. Označme jako G(x) pravděpodobnost, že chyba měření překračuje hodnotu x. Pracujeme-li s n pozorováními, pak kritickou hodnotu x_0 pro vylučování odlehlých pozorování určíme řešením rovnice $n\,G(x_0)=0.5$. Pro větší hodnoty n však tento postup vylučuje správné pozorování přibližně s 40% pravděpodobností.

Také dnes známé robustní metody používali experimentující vědci již před mnoha roky. Tak Mendělejev pracoval s useknutým průměrem, přičemž usekával 1/3 největších a 1/3 nejmenších pozorování.

Uvažujme náhodný výběr $X_1,\ldots,X_n$, kde veličiny X mají všechny stejné rozdělení s distribuční funkcí F(x). V souvislosti s odlehlými pozorováními může jít o **nesouhlas s modelem** při alternativách:

(i) všechny veličiny mají jiné rozdělení, s "těžšími chvosty";
(ii) jde o směs veličin s rozdělením podle hypotézy s veličinami s jiným rozdělením;
(iii) několik veličin je odlehlých v užším smyslu.

Zpravidla se uvažuje poslední alternativa, a to tak, že některých m veličin z oněch n veličin má buď

A: jinou střední hodnotu nebo

B: jiný (větší) rozptyl.

Nejčastěji se přitom předpokládá, že veličiny mají normální rozdělení. Z dlouhého přehledu testů proti odlehlosti v užším smyslu uvedeného v monografii Barnett and Lewis (1977) jich připomeneme několik:

N1: $\dfrac{x_{(n)} - \bar{x}}{s}$ — maximálně věrohodný test při alternativě s jedním pozorováním $N(\mu+a,\sigma^2)$, $a>0$

N2: $\max\left(\dfrac{x_{(n)} - \bar{x}}{s}, \dfrac{\bar{x} - x_{(1)}}{s}\right)$ — " — $N(\mu+a,\sigma^2)$, $a\neq 0$

N3: $\dfrac{x_{(n-m+1)} + \ldots + x_{(n)} - m\,\bar{x}}{s}$ — " — s m pozorováními $N(\mu+a_i,\sigma^2)$, $a_i>0$, $i=1,\ldots,m$

N14: $\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\,s^3}}$ — lokálně nejlepší test proti alternativě s m pozorováními $N(\mu+a_i,\sigma^2)$, $a_i>0$, $m/n<0,5$ (při $m=1$ téměř tak silný jako N1)

N15: $\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\,s^4}$ — lok. nejlepší nestranný invariantní test proti alternativě s m pozorováními $N(\mu+a_i,\sigma^2)$, $a_i\neq 0$, $m/n<0,21$ nebo lokálně nejlepší invariantní test proti alternativě s m pozorováními $N(\mu, b_i\sigma^2)$, $b_i>1$

2. Jedno pozorování a lineární model

Uvažujme klasický normální lineární regresní model

(1) $\quad y = X\beta + e\,, \quad e \sim N(0, \sigma^2 I)\,,$

u kterého předpokládáme úplnou hodnost regresní matice $X_{n\times p}$. Označme

$$b = (X'X)^{-1}X'y, \qquad u = y - Xb\,,$$
$$RSS = \|y - Xb\|\,, \qquad s^2 = RSS/(n-p)\,.$$

Vektor **u** se nazývá **vektor reziduí**.

Nyní je třeba zavést vhodnou alternativu odlehlého pozorování také do modelu (1): i-té pozorování bude odlehlé, když střední hodnota i-tého chybového členu e_i bude nenulová. Ovšem na rozdíl od 1. kapitoly vektor **e**

(tam vektor **x**) nyní nepozorujeme přímo, ale pouze prostřednictvím vektoru **y**. Nabízí se možnost použít vektor reziduí **u**, který lze považovat za odhad náhodného vektoru **e**. Všimněme si však rozdělení vektoru **u**. Platí

$$(3) \qquad u \sim N(0,\ \sigma^2(I - H)),$$

kde **H** je hojně používaná projekční (tzv. hat) matice

$$(4) \qquad H = X\,(X'X)^{-1}X' ,$$

která je symetrická, idempotentní a má hodnost p. Odtud plyne, že složky vektoru **u** nemusí mít stejný rozptyl a nemohou být vesměs nezávislé.

Především v souvislosti s hledáním odlehlých pozorování bylo proto navrženo několik úprav vektorů reziduí. Pokusme se nejprve náhodnou veličinu u_i normovat. Neznámý parametr σ^2 při tom nahradíme jeho nestranným odhadem s^2. Takto dostaneme **normovaná** (standardizovaná) **rezidua**

$$(5) \qquad v_i = \frac{u_i}{s\sqrt{1-h_{ii}}}, \qquad i=1,\ldots,n.$$

Není třeba obávat se o regulérnost výrazu ve jmenovateli, pokud budeme předpokládat $1-h_{ii}>0$. Protože jsou náhodné veličiny v_i a u (pro každé i) nezávislé, snadno se spočte, že platí $E\,v_i = 0$ a $var\,v_i = 1$, $i=1,\ldots,n$. Ovšem rozdělení veličiny v_i není právě z nejpoužívanějších. Jak uvidíme dále, náhodná veličina $(n-p)\,v_i^2$ má beta rozdělení s parametry 1/2, (n-p-1)/2.

Pravděpodobně častěji se v této souvislosti používají **studentizovaná** (jackknife) **rezidua**

$$(6) \qquad v_i^* = \frac{u_i}{s_{(i)}\sqrt{1-h_{ii}}}, \qquad i=1,\ldots,n.$$

Jediný rozdíl proti definici normovaných reziduí je ve jmenovateli. Místo běžného nestranného odhadu rozptylu je zde uveden odhad založený na všech pozorováních **kromě** i-tého.

Zaveďme označení $X_{(i)}$, $y_{(i)}$ pro matici (vektor) vzniklou z matice **X** (vektoru **y**) vynecháním i-tého řádku x_i' (složky y_i). Dříve uvedený předpoklad $1-h_{ii}>0$ je ekvivalentní předpokladu, že matice **X** a $X_{(i)}$ mají stejnou hodnost. V takovém případě dostáváme v modelu $y_{(i)} \sim N(X_{(i)}\beta, \sigma^2 I_{n-1})$ odhady

$$b_{(i)} = (X'_{(i)}\,X_{(i)})^{-1}X'_{(i)}\,y_{(i)} ,$$

$$RSS_{(i)} = \|y_{(i)} - X_{(i)}\,b_{(i)}\|^2 ,$$

$$s^2_{(i.)} = RSS_{(i.)}/(n-1-p) .$$

Protože je

$$\text{var}\,(y_i - \mathbf{x}_i'\mathbf{b}_{(i.)}) = \sigma^2 + \sigma^2\,\mathbf{x}_{i.}'(\mathbf{X}_{[i.]}'\mathbf{X}_{[i.]})^{-1}\mathbf{x}_{i.} ,$$

má podíl

$$(7) \qquad \frac{y_i - \mathbf{x}_{i.}'\mathbf{b}_{(i.)}}{s_{(i.)}\sqrt{1+\mathbf{x}_{i.}'(\mathbf{X}_{[i.]}'\mathbf{X}_{[i.]})^{-1}\mathbf{x}_{i.}}}$$

Studentovo t rozdělení s n-p-1 stupni volnosti. V čitateli a jmenovateli jsou totiž nezávislé náhodné veličiny, protože $\mathbf{b}_{(i.)}$ a $s^2_{(i.)}$ jsou nezávislé. Použijeme-li známý vzorec (např. Anděl (1978), str. 73, cvič. 11), dostaneme

$$(\mathbf{X}_{[i.]}'\mathbf{X}_{[i.]})^{-1} = (\mathbf{X}'\mathbf{X} - \mathbf{x}_{i.}\mathbf{x}_{i.}')^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_{i.}\,\frac{1}{1-\mathbf{x}_{i.}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{i.}}\,\mathbf{x}_{i.}'(\mathbf{X}'\mathbf{X})^{-1}.$$

Po vynásobení vektorem $\mathbf{X}_{[i.]}'\mathbf{y}_{[i]} = \mathbf{X}'\mathbf{y} - \mathbf{x}_{i.}y_i$ dostaneme důležitý vztah, svazující odhad parametru β ze všech pozorování s odhadem stejného parametru, založeným na všech pozorováních s výjimkou i-tého

$$(8) \qquad \mathbf{b}_{(i.)} = \mathbf{b} - \frac{u_i}{1-h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i .$$

Odtud dostaneme po úpravě také

$$RSS_{(i.)} = \mathbf{y}_{[i]}'\mathbf{y}_{[i]} - \mathbf{y}_{[i]}'\mathbf{X}_{[i.]}\,\mathbf{b}_{(i.)}$$

$$= RSS - \frac{u_i^2}{1-h_{ii}},$$

takže nestranné odhady rozptylu spolu souvisejí vztahem

$$(9) \qquad \frac{s^2_{(i)}}{s^2} = \frac{n-p-v_i^2}{n-p-1} .$$

Zřejmě tedy se po vyloučení i-tého pozorování odhad rozptylu zmenší právě když je $v_i^2>1$. Vraťme se ale k výrazu (7). Čitatele lze upravit na tvar

$$y_i - x'_{i.} b_{(i.)} = y_i - x'_{i.} b + \frac{u_i h_{ii}}{1-h_{ii}} = \frac{u_i}{1-h_{ii}} ,$$

podobně jmenovatele

$$1 + x'_{i.} (X'_{(i.)} X_{(i.)})^{-1} x_{i.} = 1 + h_{ii} + \frac{h_{ii}^2}{1-h_{ii}} = \frac{1}{1-h_{ii}} .$$

Uvážíme-li tyto úpravy, vidíme, že v (7) je jen jinak zapsáno studentizované reziduum v_i, které má tedy Studentovo t rozdělení s n-p-1 stupni volnosti.

Vztah (9) můžeme použít také k explicitnímu vyjádření vztahu normovaných a studentizovaných reziduí. Po nepříliš složité úpravě dostaneme

$$v_i^* = v_i \left(\frac{n-p-1}{n-p-v_i^2} \right)^{1/2}$$

K jiné důležité představě nás přivede článek Andrewse (1971). Uvidíme, že na vektoru reziduí není zajímavá jen jeho délka (reziduální součet čtverců). Normujme vektor reziduí na jednotkovou délku. Zaveďme vektor $r = u/\|u\|$. Stejně jako vektor u leží tento vektor v ortogonálním doplňku $M(X)^1$ lineárního obalu sloupců matice X, tedy v lineárním prostoru dimenze n-p. Vzhledem k jeho jednotkové délce si můžeme představovat, že bod r leží na povrchu jednotkové koule. Pokusme se ukázat, že tento náhodný bod má na jednotkové kouli rovnoměrné rozdělení.

Vezměme libovolný vektor $a \in M(X)^1$, $\|a\|=1$. Protožepředpokládáme, že je $e \sim N(0,\sigma^2 I)$, platí nutně $a'e \sim N(0,\sigma^2)$. Ovšem (I-H) je projekční maticí na $M(X)^1$, takže platí (I-H)a=a. Proto je také $a'u=a'(I-H)e=a'e \sim N(0,\sigma^2)$, takže kvadratická forma $u'aa'u/\sigma^2$ má má chí kvadrát rozdělení o jednom stupni volnosti χ_1^2. Náhodná veličina $e'aa'a/\sigma^2$ má rozdělení χ_1^2. Je známo, že reziduální součet čtverců RSS=u'u vydělený σ^2 má rozdělení χ_{n-p}^2. Protože je matice (I-aa') idempotentní, jak se lze snadno přesvědčit, mají kvadratické formy $u'aa'u/\sigma^2$ a $u'(I-aa')u/\sigma^2$ po řadě rozdělení χ_1^2 a χ_{n-p-1}^2 a jsou nezávislé. Označíme-li symbolem φ náhodný úhel, který svírají vektory a a u (resp. r), pak lze psát

$$(\cos\varphi)^2 = (r'a)^2 = \frac{(u'a)^2}{u'u} = \frac{u'aa'u}{u'aa'u+u'(I-aa')u} ,$$

takže náhodná veličina $(\cos\varphi)^2$ má rozdělení beta(1/2,(n-p-1)/2). Protože toto rozdělení je pro všechny vektory a jednotkové délky stejné, pro vektor r zbývá pouze rovnoměrné rozdělení.

3. Odlehlé pozorování v lineární regresi

Vedle modelu (1) zaveďme alternativní model s jedním odlehlým pozorováním. Má-li být tímto odlehlým pozorováním pozorování t-té, pak lze model vyjádřit ve tvaru

$$(9) \qquad y \sim N(X\beta + \gamma j_t,\ \sigma^2 I)$$

a o odlehlé pozorování půjde, bude-li $\gamma \neq 0$. V takovém případě se však poněkud změní vlastnosti vektoru u, neboť bude

$$E\,u = (I-H)(X\beta + \gamma j_t) = \gamma\,(I-H)j_t = \gamma\, m_{.t},$$

kde jsme v posledním výrazu použili zápis $m_{.t}$ pro t-tý sloupec matice M=I-H.

Je-li tedy t-té pozorování odlehlé, pak očekáváme, že bude blízké (podobné) t-tému sloupci matice M. O této vlastnosti vypovídá úhel mezi uvedenými vektory. Protože všechny sloupce matice M leží v prostoru $\mathcal{M}(X)^{\perp}$ (jde o projekční matici do tohoto podprostoru), má čtverec kosinu tohoto úhlu rozdělení beta(1/2,(n-p-1)/2). Tento čtverec je ale roven čtverci skalárního součinu

$$\left(\frac{u}{\|u\|} \cdot \frac{m_{.t}}{\|m_{.t}\|}\right)^2 = \frac{(u'Mj_t)^2}{u'u\ j_t'Mj_t} = \frac{u_t^2}{(n-p)s^2 m_{tt}} = \frac{1}{(n-p)}\, v_t^2 ,$$

což dává novou interpretaci pro normované reziduum v_t a také důkaz dříve zmíněného tvrzení o rozdělení normovaného rezidua.

K testování odlehlosti t-tého pozorování můžeme použít také studentizované reziduum v_t^*. Test na předem zvolené hladině α dostaneme jen když index t byl určen nikoliv v závislosti na vektoru y. Stačí porovnat hodnotu $|v_t^*|$ s kritickou hodnotou t rozdělení $t_{n-p-1}(\alpha)$.

Častěji ovšem podezříváme nějaké pozorování právě proto, že jemu odpovídající reziduum je mimořádně veliké. Pokud porovnáváme rezidua pomocí absolutních hodnot studentizovaných (nebo normovaných) reziduí, můžeme použít Bonferroniho nerovnost

$$P(\max_{1\le j\le n} |v_j^*| \ge c) = P(\bigcup_{j=1}^{n} \{|v_j^*| \ge c\}) \le \sum_{j=1}^{n} P(|v_j^*| \ge c) .$$

Abychom zachovali hladinu α, stačí zvolit

$$c = t_{n-p-1}(\alpha/n) ,$$

což zaručí, že tuto hladinu nepřekročíme.

Kdybychom podobně jako u jednoho odlehlého pozorování podezřívali celkem m pozorování, avšak znovu bez ohledu na hodnotu vektoru y (a tedy

bez ohledu na u), můžeme použít zobecnění výše uvedeného postupu. Bez újmy na obecnosti předpokládejme, že jde o posledních m pozorování. Místo modelu (9) budeme uvažovat model

$$(10) \qquad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = N\left(\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \beta + \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_m \end{pmatrix} \gamma,\ \sigma^2 \mathbf{I} \right),$$

což vede k soustavě normálních rovnic pro odhady b a c parametrů β a γ

$$(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)\mathbf{b} + \mathbf{X}_2'\mathbf{c} = \mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2 ,$$

$$\mathbf{X}_2\mathbf{b} + \mathbf{c} = \mathbf{y}_2 .$$

Řešením je (předpokládáme, že matice $\mathbf{X}_1$ má hodnost p)

$$\mathbf{b} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 , \quad \mathbf{c} = \mathbf{y}_2 - \mathbf{X}_2\mathbf{b}$$

s reziduálním součtem čtverců $RSS_m = \|\mathbf{y}_1 - \mathbf{X}_1\mathbf{b}\|^2$. Testování hypotézy $\gamma=0$ vede k běžné F statistice

$$F = \frac{RSS - RSS_m}{RSS_m} \cdot \frac{n-(p+m)}{m} ,$$

která má za platnosti hypotézy $\gamma=0$ F rozdělení s m a n-p-m stupni volnosti. Hypotézu tedy zamítáme při překročení kritické hodnoty $F_{m,n-p-m}(\alpha)$. Pokud bychom neměli pozorování, která mohou být odlehlá, určena předem, museli bychom podobně jako u jediného pozorování použít Bonferroniho nerovnost.

4. Vlivná pozorování

Pokusme se vyjádřit co možná jednoduše vliv i-tého pozorování na chování odhadů. Velikost tohoto vlivu můžeme charakterizovat pomocí vektoru $\mathbf{b}-\mathbf{b}_{(i.)}$ (viz (9)), případně pomocí délky tohoto vektoru. Vezmeme-li v úvahu nestejnou přesnost odhadu jednotlivých složek vektoru β a jejich korelovanost, dospějeme ke Cookově míře vlivu

$$D_i = (\mathbf{b}_{(i.)}-\mathbf{b})' \mathbf{X}'\mathbf{X} (\mathbf{b}_{(i.)}-\mathbf{b})/(ps^2)$$

$$= \|\mathbf{X}\mathbf{b}_{(i.)}-\mathbf{X}\mathbf{b}\|^2/(ps^2)$$

Když dosadíme podle (9), dostaneme po malé úpravě

$$(11) \qquad D_i = v_i^2 \frac{1}{p} \frac{h_{ii}}{1-h_{ii}} = v_i^2 \frac{1}{p} \frac{\operatorname{var} \hat{y}_i}{\operatorname{var} u_i} .$$

Jak je vidět, záleží na počtu parametrů, na "odlehlosti" i-tého pozorování

měřené velikostí v_i^2, a také na velikosti i-tého diagonálního prvku matice H.

Jak lze tuto poslední hodnotu interpretovat? Předpokládejme, že matice X obsahuje sloupec ze samých jedniček, že tedy regresní vztah obsahuje absolutní člen. Potom je h_{ii} rovno čtverci jisté vzdálenosti i-tého řádku matice X od "těžiště" všech těchto řádků zvětšenému o konstantu 1/n. Pozorování s velkou hodnotou h_{ii}, jak je patrné ze vztahu (9), mají mimořádně velký vliv na hodnotu vektoru b. Zpravidla se za **vzdálená** označují pozorování, pro něž hodnota h_{ii} překračuje hodnotu 2p/n.

5. Grafické postupy

Zatím jsme se zmiňovali o různých charakteristikách, které umožňují určování odlehlých či mimořádně vlivných (vzdálených) pozorování. V praxi však zpravidla vystačíme s grafickými diagnostickými postupy. O chování jednotlivých reziduí se rychle přesvědčíme, když si je znázorníme jako funkci pořadového indexu i, některého regresoru nebo vyhlazených hodnot $\hat{y}_i$. Takovýto rozptylový diagram nás upozorní na nestejný rozptyl, nevhodně zvolenou funkční závislost i na odlehlá pozorování. Na odlehlá pozorování nás upozorní také různé normální diagramy, byť byly původně určeny k ověřování normality. V normálních diagramech se zpravidla zobrazují uspořádané hodnoty reziduí proti hodnotám kvantilové funkce normálního rozdělení. Na normální rozdělení ukazuje diagram, ve kterém leží jednotlivé body přibližně na přímce. Odlehlé pozorování se projevuje jako krajní bod, který se odchyluje směrem k velkým či malým hodnotám.

6. Literatura

J. Anděl (1978). Matematická statistika. SNTL Praha.

D.F. Andrews (1971). Significance tests based on residuals. Biometrika 58, 139-148.

D.F. Andrews and D. Pregibon (1978). Finding outliers that matter. JRRS B 40, 85-93.

A.C. Atkinson (1981). Two graphical displays for outlying and influential observations in regression. Biometrika 68, 13-20.

V. Barnett and T. Lewis (1977). Outliers in statistical data. Wiley.

R.D. Cook (1977). Detection of influential observations in linear regression. Technometrics 19, 15-18.

R.D. Cook (1979). Influential observations in linear regression. JASA 74, 169-174.

N.R. Draper and J.A. John (1981). Influential observations and outliers in regression. Technometrics 23, 21-26.

R.E. Lund (1975). Tables for an approximate test on residuals. Biometrika 58, 139-148.

K.S. Srikantan (1961). Testing for a single outlier in a regression model. Sankhya A 23, 251-260.

K. Zvára (1989). Regresní analýza. Academia Praha.