KŘIVOST V NELINÁRNÍ REGRESI

KAREL ZVÁRA

ABSTRACT. Když si něco začneme s regresí, brzy dojde i na tu nelineární. Ne vždy však doporučené postupy fungují, možná je to právě tou nelinearitou. Co to znamená nelineární? Lze nelinearitu nějak měřit nebo dokonce odstranit? Co všechno ovlivňuje? Je vůbec skutečným problémem, vyskytuje se v praxi?

1. ÚVOD

Nejprve zavedeme označení. Mějme náhodný vektor y, jehož střední hodnotu označíme $\eta(\theta)$, kde θ je vektor neznámých parametrů. Čím se liší regresní úloha od zpracování n-tice nezávislých náhodných veličin se stejnou střední hodnotou a stejným rozptylem? V regresi vyšetřované veličiny obecně nemají stejné střední hodnoty, ale před úplnou volností nás tu chrání požadavek, že vyhovují nějaké vedlejší podmínce, která jim nechá jen několik stupňů volnosti. V případě lineární regrese leží všechny n-tice přípustných středních hodnot ve známém lineárním podprostoru, afinní prostor nestačí.

Příklad 1. Mějme $\eta_1(\theta) = \beta$, $\eta_2(\theta) = 1$. Je to lineární regrese nebo není? Střední hodnoty θ jsou pro několik ekvidistantních hodnot β znázorněny v horní části obrázku 1, přičemž z určitého důvodu jsou druhé souřadnice zvětšeny o jedničku. Všechny střední hodnoty jsou lineárními kombinacemi dvou jednotkových vektorů, ovšem nulový vektor vytvořit nelze. Taylorův rozvoj vektoru středních hodnot lze zapsat bez problémů. Pro $\beta = 1$ je to

$$\eta(\beta) = \begin{pmatrix} \beta \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} (\beta - 1). \tag{1}$$

Vektor středních hodnot můžeme psát také jako

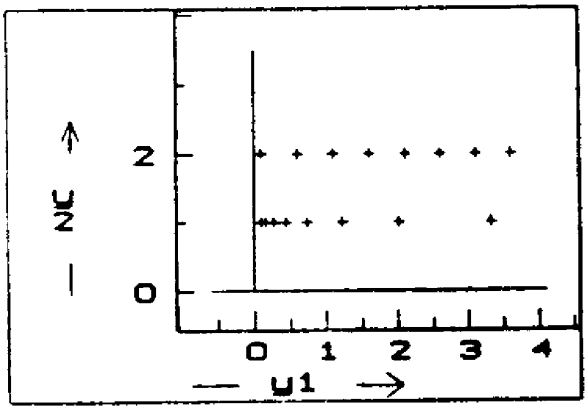
$$\eta(\theta) = \begin{pmatrix} e^{\theta} \\ 1 \end{pmatrix},$$

kde $\theta = \ln(\beta)$. Rozvoj provedeme kolem $\theta = \ln(1) = 0$

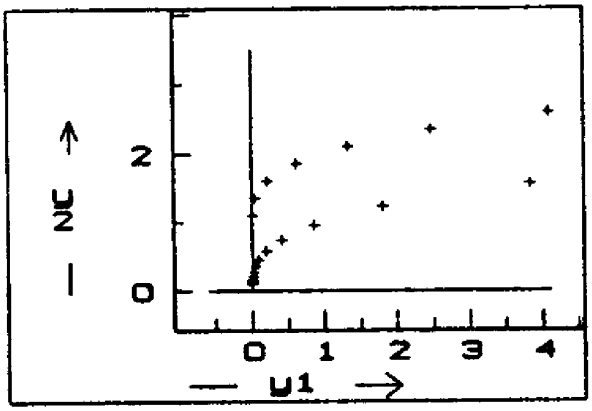
$$\eta(\theta) = \begin{pmatrix} e^{\theta} \\ 1 \end{pmatrix} \doteq \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} (\theta - 0). \tag{2}$$

Tentokrát platí rozvoj opravdu jen přibližně, což je vidět už z toho, že na pravé straně vztahů (1) a (2) jsou identické funkce. V čem je rozdíl? Podívejme se znovu na obrázek 1. Dolní řada bodů odpovídá vektorům středních hodnot pro ekvidistantní hodnoty θ. Tyto body nejsou rozmístěny rovnoměrně, takže při rovnoměrné změně hodnoty

Typeset by AMS-TEX



Obr. 1: Ekvidistantní střední hodnoty pro příklad 1



Obr. 2: Ekvidistantní střední hodnoty pro příklad 2

parametru probíháme množinu řešení nerovnoměrně. Uvidíme později, že právě tato vlastnost je jedním ze zdrojů potíží v nelineární regresi.

Zkusme vektor středních hodnot přepsat ještě jednou. Obě vyjádření můžeme vyjádřit pomocí $x_1 = 1$ a $x_2 = 0$ ve tvaru

$$\eta(\beta) = \begin{pmatrix} \beta^{x_1} \\ \beta^{x_2} \end{pmatrix} \tag{3}$$

resp.

$$\eta(\theta) = \begin{pmatrix} e^{\theta x_1} \\ e^{\theta x_2} \end{pmatrix} \tag{4}$$

Zvolíme-li hodnoty x_1, x_2 jinak, bude nelinearita nepochybná, jak je vidět z dalšího příkladu.

Příklad 2. Předpokládejme vektor středních hodnot (3) resp. (4), kde zvolíme $x_1 = 3, x_2 = 1$. Tyto vektory jsou pro ekvidistantní hodnoty parametru a pro obě parametrizace (pro β je opět druhá souřadnice zvětšena o 1) znázorněny na obrázku 2. Množina řešení, kterou v obou případech probíháme nerovnoměrně, je tentokrát zakřivená, což naznačuje další možný problém.

Jako úlohu nelineární regrese lze zformulovat i tak běžnou úlohu, jakou je porovnání středních hodnot dvou nezávislých náhodných výběrů.

Příklad 3. Mějme nezávislé náhodné veličiny

$$y_1, \ldots, y_m \sim \mathrm{N}(\theta_1, \sigma^2), \qquad y_{m+1}, \ldots, y_{2m} \sim \mathrm{N}(\theta_1\theta_2, \sigma^2).$$

Pomocí regresní funkce

$$f(x_i, \theta) = x_i\theta_1 + (1 - x_i)\theta_1\theta_2, \qquad x_i = \begin{cases} 1 & 1 \le i \le m, \\ 0 & m+1 \le i \le 2m \end{cases}$$

lze zapsat tuto úlohu jako úlohu nelineární regrese. Je však zřejmé, že při jiné parametrizaci ($\beta_1 = \theta_1$, $\beta_2 = \theta_1\theta_2$) jde o úlohu lineární.

2. Lineární aproximace

Budeme předpokládat

$$y \sim \mathrm{N}(\eta(\theta), \sigma^2 I),$$

kde $\theta \in \Theta$ a $\sigma^2 > 0$ jsou neznámé parametry, Θ je parametrický prostor. Matici prvních parciálních derivací rozměru $n \times k$ označíme

$$\dot{F}(\theta) = \frac{\partial \eta(\theta)}{\partial \theta'}.$$

Množina všech možných středních hodnot $\{\eta(\theta), \theta \in \Theta\}$ se zpravidla nazývá *množina řešení*. Aniž bychom se jimi podrobně zabývali, budeme předpokládat platnost splnění vhodných předpokladů regularity, mezi nimiž je zejména lineární nezávislost sloupců matice $\dot{F}(\theta)$ v okolí bodu $\theta = \theta^*$.

Odhad parametru θ hledáme minimalizací funkce $S(\theta) = \|y - \eta(\theta)\|^2$. Protože má náhodný vektor y normální rozdělení, dostaneme tak maximálně věrohodný odhad t parametru θ. Vlastní minimalizace se provádí iteračně. Výsledný odhad t je řešením *normální rovnice*

$$\dot{F}(\theta)'(y - \eta(\theta)) = 0$$

Označme jako θ^* skutečnou hodnotu parametru θ. Podobně označme $\eta^* = \eta(\theta^*)$, $\dot{F}^* = \dot{F}(\theta^*)$. Dosadíme-li do normální rovnice přibližná vyjádření

$$\eta(t) \doteq \eta^* + \dot{F}^*(t - \theta^*), \quad \dot{F}(t) \doteq \dot{F}^*, \tag{3}$$

dostaneme po úpravě

$$t \doteq \eta^* + (\dot{F}^{*\prime}\dot{F}^*)^{-1}\dot{F}^{*\prime}e$$

s náhodným vektorem $e \sim \mathrm{N}(0, \sigma^2 I)$ na pravé straně. Odtud plyne přibližné tvrzení

$$t \sim \mathrm{N}(\theta^*, \sigma^2(\dot{F}^{*\prime}\dot{F}^*)^{-1}).$$

Když matici $\dot{F}^*$ aproximujeme pomocí $\dot{F}(t)$, dostaneme důležité přibližné tvrzení klasické statistiky v nelineární regresi

$$t \sim \mathrm{N}(\theta^*, \sigma^2(\dot{F}'(t)\dot{F}(t))^{-1}).$$

3. Kvadratická aproximace a vychýlení odhadu parametrů

Zaveďme nejprve další potřebné označení. Matici druhých parciálních derivací rozměru $n \times k \times k$ označíme

$$\ddot{F}(\theta) = \frac{\partial^2 \eta(\theta)}{\partial \theta \partial \theta'}.$$

V případě potřeby označíme (jr)-tý sloupec třírozměrného číselného pole $\ddot{F}(\theta)$ jako $\ddot{f}_{.jr}(\theta)$, i-tou vrstvu jako $\ddot{F}_{i..}(\theta)$.

Abychom lépe vystihli nelineární charakter úlohy, použijeme kvadratickou aproximaci regresní funkce

$$\eta(t) \doteq \eta(\theta^*) + \dot{F}(\theta^*)(t - \theta^*) + \tfrac{1}{2}(t - \theta^*)'\ddot{F}(\theta^*)(t - \theta^*) \tag{9}$$

a porovnáme ji s aproximací lineární z (5). Jiné odvození kvadratické aproximace, podle Boxe (1971), je uvedeno v knize Zvára (1989). Vynásobme poslední rovnici maticí H^*, což je projekční matice na tečnou nadrovinu k množině řešení v bodě $\eta(\theta^*)$. Předpokládáme při tom, že v okolí tohoto bodu je množina řešení natolik symetrická, že je $\mathsf{E}\, H^*\eta(t) = H^*\eta(\theta^*)$. Vzhledem k předpokládané lineární nezávislosti sloupců matice $\dot{F}^*$ dostaneme po úpravě aproximaci

$$\mathsf{E}\, t \doteq \theta^* - \tfrac{1}{2}(\dot{F}^{*\prime}\dot{F}^*)^{-1}\dot{F}^{*\prime}\mathsf{E}(t-\theta^*)'\ddot{F}^*(t-\theta^*).$$

Střední hodnotu uvedenou na pravé straně posledního vzorce vyjádříme pomocí vektoru m, jehož i-tá složka je rovna

$$\begin{aligned} m_i &= \sigma^{-2}\mathsf{E}(t-\theta^*)'\ddot{F}^*_{i..}(t-\theta^*) \\ &= \sigma^{-2}\operatorname{tr}\ddot{F}^*_{i..}\mathsf{E}(t-\theta^*)(t-\theta^*)' \\ &= \sigma^{-2}\operatorname{tr}\ddot{F}^*_{i..}\operatorname{var} t \\ &= \operatorname{tr}(\dot{F}^{*\prime}\dot{F}^*)^{-1}\ddot{F}^*_{i..} \end{aligned}$$

(V symbolu tr čtenář jistě poznal součet diagonálních prvků čtvercové matice, tedy její stopu.) Vychýlení $\operatorname{bias} t$ můžeme vyjádřit jako

$$\operatorname{bias} t = -\tfrac{\sigma^2}{2}(\dot{F}^{*\prime}\dot{F}^*)^{-1}\dot{F}^{*\prime}m. \tag{5}$$

V lineární regresi je pole druhých derivací $\ddot{F}$ nulové, je tedy nulový vektor m i vychýlení $\operatorname{bias} t$. Vraťme se ještě na okamžik k interpretaci vektoru m. Když porovnáme jeho definici se vztahem (6), zjistíme že výraz $\frac{\sigma^2}{2}m$ udává rozdíl mezi lineární a kvadratickou aproximací regresní funkce.

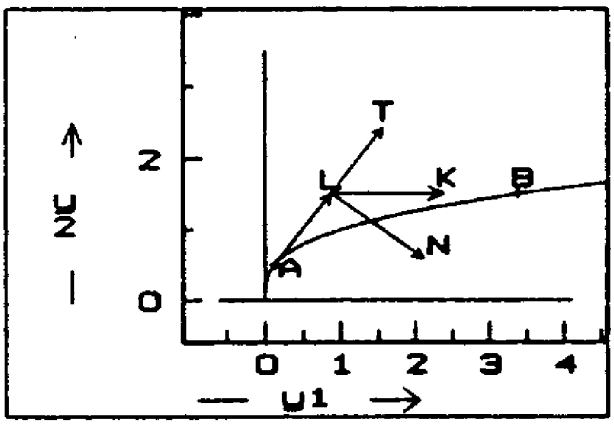
4. Míry křivosti

Vraťme se ke kvadratické aproximaci (7). Vyšetřujme ji v okolí θ^* jako funkci jediné proměnné. Dostaneme tak křivku v n-rozměrném prostoru. K tomu zvolme pevně h. Pro reálné τ v okolí nuly máme pak kvadratickou aproximaci regresní funkce $\eta(\theta)$ v bodě θ^* ve směru h

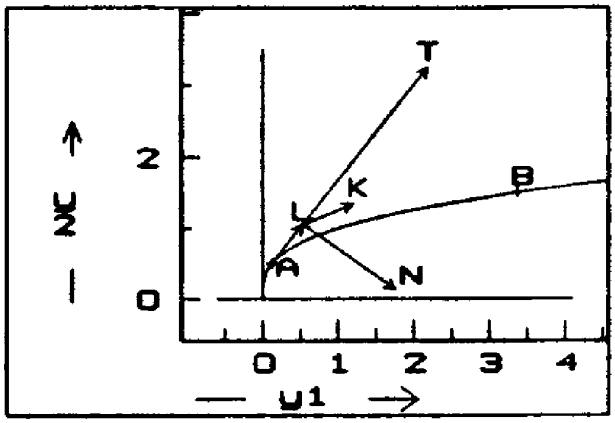
$$\eta(\theta^* + \tau h) = \eta^* + \tau\dot{f}^*_h + \tfrac{1}{2}\tau^2\ddot{f}^*_h,$$

kde $\dot{f}^*_h = \dot{F}^*h$ a $\ddot{f}^*_h = \sum\sum h_j h_r \ddot{f}^*_{jr}$. Rozložme vektor $\ddot{f}^*_h$ druhých derivací na součet dvou ortogonálních vektorů $\ddot{f}^*_h = \ddot{f}^{*T}_h + \ddot{f}^{*N}_h$, kde $\ddot{f}^{*T}_h$ leží v nadrovině tečné k množině řešení v bodě $\eta(\theta^*)$. Lineární a kvadratickou aproximaci porovnáme pro takovou hodnotu τ, aby lineární aproximace $\eta^* + \tau\dot{f}^*_h$ byla od η^* v jednotkové vzdálenosti, takže zvolíme $\tau = \|\dot{F}^*h\|^{-1}$. Rozdíl aproximací podobně rozdělíme do dvou ortogonálních složek a v poměru délek těchto složek zavedeme dvě míry křivosti (viz Bates, Watts (1980)):

$$K^N_h = \frac{\|\ddot{f}^{*N}_h\|}{\|\dot{f}^*_h\|^2}\sigma\sqrt{k} = \frac{\|h'\ddot{F}^{*N}h\|}{\|\dot{F}^*h\|^2}\sigma\sqrt{k} \qquad \text{(vnitřní křivost)}$$



Obr. 3: Rozklad vektoru druhých derivací do dvou složek. Parametrizace pomocí β.



Obr. 4: Rozklad vektoru druhých derivací do dvou složek. Parametrizace pomocí θ.

$$K_h^T = \frac{\|\ddot{f}_h^{*T}\|}{\|\dot{f}_h^*\|^2}\sigma\sqrt{k} = \frac{\|h'\ddot{F}^{*T}h\|}{\|\dot{F}^*h\|^2}\sigma\sqrt{k} \qquad \text{(parametrická křivost)}$$

Symbol k použitý v posledních vzorcích značí počet složek vektoru parametrů. Poznamenejme, že převrácená hodnota zlomku v definici vnitřní křivosti je rovna poloměru oskulační kružnice ke křivce v bodě dotyku k tečné nadrovině. Tato hodnota je *nezávislá na parametrizaci*, lze ji ovlivnit pouze tvarem množiny řešení.

Na obrázcích 3 a 4 je tato situace znázorněna pro úlohu z příkladu 2. Jako A je označen bod dotyku tečné nadroviny $\eta(\theta^*)$, jako B bod $\eta(\theta)$. Jeho lineární aproximace je označena jako L, kvadratická jako K. Rozklad vektoru $\tau^2\ddot{f}_h^*$ je umístěn do bodu L (lineární aproximace). Orientovaná úsečka $\overrightarrow{LT}$ značí složku $\tau^2\ddot{f}_h^{*T}$ ležící v tečné nadrovině, orientovaná úsečka $\overrightarrow{LN}$ pak složku na tuto nadrovinu kolmou. Porovnání obrázků 3 a 4 potvrzuje, že velikost složky tohoto vektoru kolmé na tečnou nadrovinu nezávisí na zvolené parametrizaci. Na druhé straně délka průmětu stejného vektoru do tečné nadroviny je u obou parametrizací podstatně odlišná.

Příklad 2a. Pokračujme v příkladu 2 a spočítejme postupně jednotlivé veličiny, které jsme právě zavedli. Protože parametr β resp. θ je pouze jednorozměrný, není třeba zavádět vektor h. Nejprve budeme pracovat s parametrizací (3). Běžnými úpravami dostaneme postupně (b značí odhad pro β)

$$\dot{F}(\beta) = \begin{pmatrix} 3\beta^2 \\ 1 \end{pmatrix}, \qquad \ddot{F}(\beta) = \begin{pmatrix} 6\beta \\ 0 \end{pmatrix}$$

$$m = \frac{1}{1+9\beta^4}\begin{pmatrix} 6\beta \\ 0 \end{pmatrix}, \qquad \text{bias } b = -\frac{\sigma^2}{2}\frac{18\beta^2}{(1+9\beta^4)^2},$$

takže vnitřní a parametrická křivost jsou rovny

$$K_h^N = \frac{6\beta}{(1+9\beta^4)^{3/2}}\sigma, \qquad K_h^T = \frac{18\beta^3}{(1+9\beta^4)^{3/2}}\sigma.$$

Pro parametrické vyjádření z (4) dostaneme podobně

$$\dot{F}(\theta) = \begin{pmatrix} 3e^{3\theta} \\ e^{\theta} \end{pmatrix}, \qquad \ddot{F}(\theta) = \begin{pmatrix} 9e^{3\theta} \\ e^{\theta} \end{pmatrix},$$

$$m = \frac{1}{e^{2\theta}+9e^{6\theta}}\begin{pmatrix} 9e^{3\theta} \\ e^{\theta} \end{pmatrix}, \qquad \text{bias } b = -\frac{\sigma^2}{2}\frac{e^{2\theta}+27e^{6\theta}}{(e^{2\theta}+9e^{6\theta})^2},$$

takže vnitřní a parametrická křivost jsou rovny

$$K_h^N = \frac{6e^{\theta}}{(1+9e^{4\theta})^{3/2}}\sigma, \qquad K_h^T = \frac{1+27e^{4\theta}}{e^{\theta}(1+9e^{4\theta})^{3/2}}\sigma.$$

Uvážíme-li, že je $\beta = e^{\theta}$, snadno nahlédneme, že vnitřní křivosti jsou opravdu totožné, kdežto parametrická křivost je pro β vždy menší než pro odpovídající $\theta = \ln(\beta)$.

Příklad 3a. Vraťme se k příkladu 3, který je jen jiným parametrickým vyjádřením známého dvouvýběrového t-testu. Snadno dostaneme

$$\dot{F}(\theta) = \begin{pmatrix} 1 & 0 \\ \theta_2 1 & \theta_1 1 \end{pmatrix}$$

$$f_{.12} = f_{.21} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad f_{.11} = f_{.22} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

takže vychýlení odhadu je po řadě úprav dáno vztahem

$$\text{bias } t = \frac{\sigma^2}{m}\frac{\theta_2}{\theta_1^2}\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Při výpočtu křivostí musíme tentokrát zvolit směr h. Vnitřní křivost je ovšem identicky nulová, protože sloupce matice $\ddot{F}$ jsou lineárními kombinacemi sloupců matice $\dot{F}$, takže složka vektoru druhých derivací ortogonální k tečné nadrovině je vždy nulová. Jak lze ukázat, parametrická křivost splňuje nerovnost

$$K_h^T = \frac{\sqrt{8mh_1^2h_2^2}\sqrt{2}\sigma}{mh_1^2 + m(h_1\theta_2 + h_2\theta_1)^2} \leq \frac{(\sqrt{1+\theta_2^2}+|\theta_2|)\sigma}{\sqrt{m}|\theta_1|},$$

přičemž rovnost může nastat. Všimněme si zajímavé vlastnosti právě uvedené maximální parametrické křivosti. Představme si, že porovnáváme dva výběry se středními hodnotami $\theta_1 = 20°C$ a $\theta_1\theta_2 = 30°C$. Maximální parametrická křivost je rovna hodnotě $0,234\sigma/\sqrt{m}$. Když však stejné teploty vyjádříme v absolutní stupnici, bude maximální křivost rovna pouze hodnotě $0,012\sigma/\sqrt{m}$. Samozřejmě, pokud použijeme parametry $\beta_1 = \theta_1$ a $\beta_2 = \theta_1\theta_2$, budou obě křivosti identicky nulové.

Zatím jsme zavedli vnitřní a parametrickou křivost tak, že záleží na volbě vektoru h. V posledním příkladu jsme ukázali nejběžnější řešení tohoto problému — nalézt křivosti maximální. Jinou možností je počítat průměr (integrál) přes všechny směry, viz Bates, Watts (1980), kde je také uvedena souvislost s mírami křivosti, jak je navrhl Beale (1960).

5. Rezidua

Užitečným nástrojem regresní analýzy jsou rezidua definovaná tradičně vztahem $u = y - \eta(t)$. V lineární regresi mají tato rezidua nulovou střední hodnotu a používají se především k ověřování platnosti předpokladů, na kterých jsou v regresi založeny

statistické závěry. Když do uvedené definice dosadíme známou kvadratickou aproximaci (6) a odhad vychýlení (7), dostaneme postupně pro střední hodnotu reziduí

$$\begin{aligned}\mathsf{E}\,u &\doteq y - \eta^* - \dot{F}^*(t-\theta^*) - \tfrac{1}{2}(t-\theta^*)'\ddot{F}^*(t-\theta^*)\\ &\doteq -\dot{F}^*\Big(-\frac{\sigma^2}{2}(\dot{F}^{*\prime}\dot{F}^*)^{-1}\dot{F}^{*\prime}m - \tfrac{1}{2}\sigma^2 m\\ &\doteq -\frac{\sigma^2}{2}(I - H^*)m.\end{aligned}$$

Zjišťujeme tedy, že rezidua už nemají nulovou střední hodnotu. Označme symbolem $\mathcal{M}(A)$ lineární obal sloupců matice A. Právě nalezená aproximace střední hodnoty reziduí je průmětem jisté lineární kombinace sloupců matice $\ddot{F}^*$, totiž vektoru m, do ortogonálního doplňku podprostoru $\mathcal{M}(\dot{F}^*)$. Ve shodě s ostatním naším značením bychom mohli střední hodnotu $\mathsf{E}\,u$ psát jako $-\frac{\sigma^2}{2}m^{*N}$. Nenulovost této střední hodnoty způsobila ta část lineárního obalu sloupců matice $\ddot{F}^*$, která se nevešla do $\mathcal{M}(\dot{F}^*)$. Označme tedy jako $\tilde{H}$ projekční matici do podprostoru $\mathcal{M}(\dot{F}^*, \ddot{F}^{*N}) = \mathcal{M}(\dot{F}^*, \ddot{F}^*)$ a zaveďme *promítaná* (projected) *rezidua* vztahem (Cook, Tsai (1985))

$$\tilde{u} = (I - \tilde{H})u.$$

Aspoň v okolí $\eta(\theta^*)$ budou tato rezidua mít běžné vlastnosti reziduí z lineární regrese, jako jsou

$$\mathsf{E}\,\tilde{u} = 0,$$

$$\mathrm{var}\,\tilde{u} = \sigma^2(I - \tilde{H}),$$

$$\mathsf{E}\,e'(I - \tilde{H})e = \sigma^2 \mathrm{tr}(I - \tilde{H}).$$

Příklad 1b. Navážeme na příklad 1. Snadno se zjistí, že platí

$$\dot{F}(\beta) = \begin{pmatrix}1\\0\end{pmatrix}, \quad \ddot{F}(\beta) = \begin{pmatrix}0\\0\end{pmatrix},$$

a

$$\dot{F}(\theta) = \begin{pmatrix}e^\theta\\0\end{pmatrix}, \quad \ddot{F}(\theta) = \begin{pmatrix}e^\theta\\0\end{pmatrix}.$$

Matice $\ddot{F}$ je tedy násobkem matice $\dot{F}$, takže platí nutně $\tilde{u} = u$ a dvojí vektory reziduí se neliší.

Příklad 2b. V tomto příkladě nastala zcela netypická situace, neboť jak je zřejmé z příkladu 2a, lineární obal matic $\dot{F}$ a $\ddot{F}$ obsahuje všechny dvourozměrné vektory, takže na promítaná rezidua nezbývají žádné stupně volnosti. Je tedy identicky $\tilde{u} = 0$.

Příklad 3b. Každý sloupec matice $\ddot{F}$ je lineární kombinací sloupců matice $\dot{F}$ (viz příklad 3a), takže je $\mathcal{M}(\dot{F}, \ddot{F}) = \mathcal{M}(\dot{F})$ a tudíž $\tilde{u} = u$.

6. Pázmanova klasifikace

Užitečné třídění úloh nelineární regrese navrhl Pázman (1992). *Lineární úlohu* lze charakterizovat podmínkou $\ddot{F}(\theta) = O$ pro všechna $\theta \in \Theta$. Pro *vnitřně lineární* úlohu je charakteristické, že platí $\ddot{F}^N(\theta) \equiv O$. To mimo jiné znamená, že klasická rezidua mají nulovou střední hodnotu. Úlohy s *konstantní informační maticí* splňují podmínku $\ddot{F}^T(\theta) \equiv O$. Nulová parametrická křivost znamená, že vychýlení odhadu regresních koeficientů podle (7) je nulové. Posledním speciálním případem této klasifikace jsou úlohy s *nulovou Riemannovou křivostí*. Jejich regresní funkci lze parametrizovat tak, aby parametrická křivost byla identicky nulová. Existencí takové parametrizace se zabývá zvláště P. Hougaard, viz např. Hougaard (1982).

Důležitá je souvislost tohoto třídění s kvalitou běžných konfidenčních množin pro vektor parametrů θ. Chceme-li dělat závěry pro konečný rozsah výběru, musíme předpokládat normální rozdělení náhodného vektoru y nebo aspoň pro velký rozsah výběru použít asymptotická tvrzení. Klasická eliptická konfidenční množina

$$\left\{\theta \in \Theta : (\theta - t)' \left(\dot{F}'(t)\dot{F}(t)\right)^{-1} (\theta - t) \leqq \sigma^2 \chi_k^2(1-\alpha)\right\} \tag{8}$$

je přesná (má zaručenu nominální spolehlivost) pouze v lineárním modelu. Ovšem v úloze s konstantní informační maticí (s nulovou parametrickou křivostí) je tato množina velmi blízká množině

$$\left\{\theta \in \Theta : \|H(\theta)(y - \eta(\theta)\|^2 \leqq \sigma^2 \chi_k^2(1-\alpha)\right\},$$

která je přesná v každé úloze.

Ve vnitřně lineární úloze (bez ohledu na parametrickou křivost) je přesnou konfidenční množina

$$\left\{\theta \in \Theta : \|y - \eta(\theta\|^2 - \|y - \eta(t)\|^2 \leqq \sigma^2 \chi_k^2(1-\alpha)\right\}. \tag{9}$$

Pozorný čtenář si jistě uvědomil, že všechny tři uvedené konfidenční množiny předpokládají známou hodnotu rozptylu σ. Když použijeme jako odhad rozptylu statistiku

$$s^2 = \|y - \eta(t)\|^2/(n-k),$$

a kvantil rozdělení χ^2 nahradíme kvantilem rozdělení $F_{k,n-k}$, můžeme konfidenční množinu (9) upravit na tzv. věrohodnostní konfidenční množinu

$$\left\{\theta \in \Theta : \|y - \eta(\theta)\|^2 \leqq s^2(n - k + kF_{k,n-k}(1-\alpha))\right\}. \tag{10}$$

Podobnou náhradou dostaneme z konfidenční množiny (8) často používanou množinu

$$\left\{\theta \in \Theta : (\theta - t)' \left(\dot{F}'(t)\dot{F}(t)\right)^{-1} (\theta - t) \leqq s^2 k\, F_{k,n-k}(1-\alpha)\right\}. \tag{11}$$

7. Jaké úlohy se v praxi vyskytují

Brzy po tom, co vyšel článek Donaldson, Schnabel (1987), upozornil na něj Tomáš Havránek na pravidelném středečním semináři. V článku je kromě jiného popsán simulační experiment se reálnými daty. Autoři zvolili 20 úloh publikovaných v literatuře. Pro každou úlohu použili známý odhad parametrů jako skutečnou hodnotu parametru θ^*. Potom opakovaně (500 krát) simulovali nová pozorování, našli odhad t a zjistili, zda θ^* leží v konfidenční množině (11). Relativní četnost tohoto jevu odhaduje její skutečnou spolehlivost. U 95% konfidenční množiny (11) se ukázalo, že skutečná spolehlivost elipsoidická množiny silně koreluje se zjištěnou *parametrickou* křivostí. U tří úloh, kde byla parametrická křivost menší než křivost mezní, za kterou je považována hodnota $(F_{k,n-k}(1-\alpha))^{-1/2}$, se empirická spolehlivost lišila od nominální nejvýše o 3,8%. Naproti tomu u úloh, kde je parametrická křivost větší než desetinásobek mezní křivosti, byla empirická spolehlivost v rozmezí od 82,6% do 12,4%! Pro data s parametrickou křivostí v rozmezí od jednonásobku do desetinásobku mezní křivosti byla empirická spolehlivost v rozmezí od 83,2% do 91,6%.

Empirická spolehlivost 95% věrohodnostní konfidenční množiny (10) vyšla u všech úloh velmi blízká své nominální hodnotě.

Jistě by bylo zajímavé zjistit závislost na hodnotě vnitřní křivosti. Bohužel (raději bohudík) se v praxi setkáváme zpravidla s daty s malou vnitřní křivostí.

8. Hledáme lepší parametrizaci

Co dělat, když zjistíme velkou hodnotu parametrické křivosti, případně velký odhad vychýlení odhadu některého parametru? Zkusme hledat jiné parametrické vyjádření. Místo parametru θ zvolme parametr $\psi \equiv g(\theta)$, kde $g(\theta) = (g_1(\theta), \ldots, g_k(\theta))'$ je dostatečně hladká invertovatelná vektorová funkce. Opět musíme použít kvadratickou aproximaci

$$g(t) \doteq g(\theta) + \frac{\partial g}{\partial \theta'}(t-\theta) + \frac{1}{2}(t-\theta)' \frac{\partial^2 g}{\partial \theta \partial \theta'}(t-\theta).$$

Když najdeme střední hodnoty obou stran, dostaneme aproximaci pro vychýlení odhadu parametru $\psi = g(t)$

$$\text{bias}\, g(t) \doteq \frac{\partial g}{\partial \theta'} \text{bias}\, t + \frac{1}{2} \begin{pmatrix} \text{tr}\, \dfrac{\partial^2 g_1}{\partial \theta \partial \theta'} \\ \vdots \\ \text{tr}\, \dfrac{\partial^2 g_k}{\partial \theta \partial \theta'} \end{pmatrix} \tag{12}$$

Ukažme si použití uvedeného vztahu na příkladu.

Příklad 4. Použijme regresní funkci (Janků, Zvára (1991)) tvaru

$$f(x; \beta_1, \beta_2) = \beta_1 \left(ux + (q-ux)(1 - \beta e^{-\frac{\beta_2}{1-x}}) \right),$$

kde q a u jsou pevné konstanty. Experimentální data vedla k bodovému odhadu $t = (0,285; 1,965)'$ s odhadem vychýlení $\text{bias}\, t = (0,050; 0,012)'$, což je 17,5% resp. 0,6% hodnoty příslušného bodového odhadu. Směrodatné odchylky odhadů jsou po řadě rovny hodnotám 0,106 a 1,381. Pro zajímavost, maximální vnitřní křivost (neznámý rozptyl

σ^2 jsme nahradili výběrovým rozptylem $s^2 = 3,183$) je rovna 0,14, kdežto maximální parametrická křivost je 10,82. Přitom nahoře zmíněná mezní křivost je rovna hodnotě $(F_{2,30}(0,95))^{-1/2} = 0,55$. Parametrická křivost je zřejmě příliš velká, což se také projevilo na odhadech vychýlení. Protože má velké vychýlení první parametr, pokusili jsme se o jiné vyjádření právě tohoto parametru. Zvolme $g_1(\theta) = 1/\theta_1$ a $g_2(\theta) = \theta_2$. Podle (12) dostaneme odhad vychýlení

$$\begin{aligned}\operatorname{bias} g(t) &\doteq \begin{pmatrix} -1/t_1^2 & 0 \\ 0 & 1 \end{pmatrix} \operatorname{bias} t + \frac{1}{2} \begin{pmatrix} \operatorname{tr}\begin{pmatrix} 2/t_1^3 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \widehat{\operatorname{var} t_1} & \widehat{\operatorname{cov}(t_1,t_2)} \\ \widehat{\operatorname{cov}(t_1,t_2)} & \widehat{\operatorname{var} t_2} \end{pmatrix} \\ \operatorname{tr}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \widehat{\operatorname{var} t_1} & \widehat{\operatorname{cov}(t_1,t_2)} \\ \widehat{\operatorname{cov}(t_1,t_2)} & \widehat{\operatorname{var} t_2} \end{pmatrix} \end{pmatrix} \\ &\doteq \begin{pmatrix} -0,050/0,285^2 \\ 0,0120 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 2 \cdot 0,106^2/0,285^3 \\ 0 \end{pmatrix} \\ &\doteq \begin{pmatrix} -0,129 \\ 0,012 \end{pmatrix}.\end{aligned}$$

Protože je $g_1(t) = 3,50$, je odhad velikosti vychýlení roven pouze -3,7% hodnoty tohoto parametru a transformace je tudíž nadějná. Skutečně, tento úsudek se v nové parametrizaci potvrdil. Maximální parametrická křivost klesla na stále ještě poměrně velkou hodnotu 2,65, kdežto maximální vnitřní křivost zůstala beze změny.

9. Závěrem

Rozlišujeme dvojí křivost — parametrickou a vnitřní. Velikost vnitřní křivosti ovlivňuje vychýlení klasických reziduí a tudíž jejich rozdíl proti reziduím promítaným. U reálných dat nebývá tato křivost příliš veliká. Parametrická křivost, na rozdíl od křivosti vnitřní, závisí na parametrizaci regresní funkce. Velikost parametrické křivosti ovlivňuje vychýlení odhadů regresních koeficientů a skutečný koeficient spolehlivosti konfidenčních množin. Na velkou hodnotu parametrické křivosti nás upozorní velké vychýlení některého parametru, což také může usnadnit hledání vhodnějšího parametrického vyjádření. Navíc výpočetní zkušenost ukazuje, že při srovnatelných výchozích odhadech je v případě parametrizace s menší křivostí potřeba méně iterací k nalezení odhadu parametrů.

References

D. M. Bates, D. G. Watts (1980), *Relative curvature measures of nonlinearity*, J. Roy. Statist. Soc. B 42, 1-25.

E. M. L. Beale (1960), *Confidence regions in non-linear estimation*, J. Roy. Statist. Soc. B 22, 41-88.

M. J. Box (1971), *Bias in nonlinear estimation*, J. Roy. Statist. Soc B 33, 171-190.

D. R. Cook, C. L. Tsai (1985), *Residuals in nonlinear regression*, Biometrika vol 72, 23-29.

R. D. Cook, J. A. Witmer (1985), *A note on parameter-effects curvature*, J. Amer. Statist. Assoc. 80, 872-878.

J. R. Donaldson, R. B. Schnabel (1987), *Computational experience with confidence regions and confidence intervals for nonlinear least squares*, Technometrics 29 no. 1, 67-82.

P. Hougaard (1982), *Parametrizations in non-linear models*, J. Roy. Statist. Soc B 44, 244-252.

I. Janků, K. Zvára (1991), *Quantitative analysis of drug handling by the kidney using a physiological model of renal drug clearence*, Grant č. 70801, ČSAV.

A. Pázman (1992), *A classification of nonlinear regression models and parameter confidence regions*, Kybernetika 28 no. 6.
K. Zvára (1989), *Regresní analýza*, Academia, Praha.

KPMS MFF UK, Sokolovská 83, 186 00 Praha 8-Karlín