ODHADY PRO NEPARAMETRICKÝ REGRESNÍ MODEL COXOVA TYPU

Petr VOLF

Model s proporcionálním rizikem slouží jako alternativní model pro popis regrese a má nesporné výhody v oblasti analýzy přežívání. Umí zacházet s náhodně cenzorovanými i nenáhodně useknutými daty. To lineárnímu modelu regrese působí značné potíže. A přitom si tyto dva modely nemusí být tak vzdáleny. V mnoha případech se nám podaří modelovat dobu přežití pomocí Weibulova rozložení pravděpodobnosti. Pokud zde použijeme pro regresi doby T na regresoru x modelu s proporcionálním rizikem, tj. doba přežití bude mít distribuční funkci $1 - \exp(-B(x).t^{\mu})$, tak regrese logaritmu doby přežití splňuje lineární model $\ln T = S = \beta(x) + e$, náhodné odchylky e mají dvojné exponenciální rozložení s distibuční funkcí $1 - \exp(-\exp \frac{s-a}{b})$, kde $B(x) = \exp[-\beta(x)+a)\mu]$, $b = 1/\mu$. Vidíme na tomto příkladě už i vztah mezi lineárním parametrickým modelem regrese s $\beta(x) = \alpha . x$ a Coxovým modelem regrese s $B(x) = C . e^{\gamma x}$. Takže zkoumání průběhu funkce B(x) neparametrickými metodami, což je téma tohoto příspěvku, je asi stejně motivováno snahou udělat si jasno o tvaru závislosti, jako v případě lineárního modelu neparametrické odhadování průběhu funkce $\beta(x)$. Jen pro pořádek musíme říci, že i po této transformaci je regrese T na x vyjádřitelná ve formě modelu s proporcionálním rizikem, s touž funkcí B(x). Je to způsobeno tím, že ln je ryze rostoucí funkce:

$$\Pr\{S > y|x\} = \Pr\{T > e^{y|x}\} = \exp(-e^{\mu y}.B(x)).$$

Mějme tedy N nezávislých n. v. $Y(x_i)$, $i = 1,\ldots,N$, tj. v N hodnotách nějaké veličiny - regresoru $x \in \mathcal{X} \subset R_1$. Předpokládáme regresní model s konstantním poměrem intenzit poruch (s proporcionálním rizikem). To znamená, že n. v. Y(x) má kumulativní intenzitu poruch (KIP) vyjádřitelnou jako

$$L(t,x) = A(t) . B(x).$$

Hned vidíme, že funkce A, B jsou určeny až na násobek (podíl) kladnou konstantou. Budeme proto spíš mluvit o identifikaci nějakých vhodných funkcí A, B než o odhadování.

Jenže realizace n. v. $Y(x_i)$ jsou cenzorovány zprava prostřednictvím n.v. $V(x_i)$ nezávislých vzájemně i na $\{Y(x_j)\}$. Takže pozorujeme $T(x_i) = \min(Y(x_i), V(x_i))$ a $\delta(x_i) = I[Y(x_i) \le V(x_i)]$. Distribuční funkci n. v. V(x) budeme značit G(t,x), $Q(t,x) = 1 - G(t,x)$. Pro distribuční funkci n.v. Y(x) použijeme označení F(t,x), a $P(t,x) = 1 - F(t,x) = \exp(-L(t,x))$. Zajímá nás chování funkcí pro $t \in [0,T]$, kde T je takové, že stále ještě $P(T,x) . Q(T,x) > 0$ pro všechna $x \in \mathcal{X}$. Samozřejmě, aby odhadování mělo smysl, mělo by být $P(T,x) < 1$.

Předpoklad 1. Funkce B(x) je spojitá na $\mathcal{X}$, nabývá pouze kladných hodnot. Funkce A(t) je spojitá, nezáporná, neklesající na $[0,T]$.

Předpoklad 2. G(t,x) je funkce spojitá v obou argumentech t i x.

Postup identifikace funkcí A a B by mohl být následující: Zvolíme v $\mathcal{I}$ pevně několik bodů $z_1,\dots,z_K$ a v nich odhadneme KIP $L(t,z_j)$. Odhad provedeme běžným "product limit" způsobem [1], ale z hodnot $T(x_i)$, $\delta(x_i)$naměřených v okolí bodu z_j, tj. v bodech $x_i \in O_{d_N}(z_j) = \{x_i : |x_i - z_j| \le d_N\}$. Půjde vlastně o jednoduchý jádrový odhad. Při $N \to \infty$ volíme $d_N \to 0$ tak, aby $N \cdot d_N \to \infty$. Po rozložení hodnot regresoru x budeme požadovat určitou rovnoměrnost v $\mathcal{I}$. To už pak stačí k tomu, aby počet měření v okolí z_j rostl bez omezení. Označíme tento počet $M_N(z_j)$ a náš požadavek můžeme zformulovat třeba jako:

Předpoklad 3. Nechť při $N \to \infty$; $M_N(z_j) \to \infty$, a to tak, že

$$0 < \underline{\lim}(M_N(z_j)/Nd_N) \le \overline{\lim}(M_N(z_j)/Nd_N) < \infty \text{ (pro každé } j = 1,\dots,K).$$

Teď tedy budeme mít odhadnuté kumulativní funkce rizika (KIP) vždy z části dat, pro určitou úroveň hodnot regresoru. Takže nyní je ta pravá chvíle pro testování proporcionality rizika, a to metodami (ať už grafickými či numerickými), které prověří, zda skutečně $\ln L(t,z_j) - \ln L(t,z_k)$ jsou (jako funkce t na $[0,T]$) vůči sobě posunuty. O metodách takovýchto testů viz i příspěvek V. Lánské v tomto sborníku. To posunutí by v našem modelu mělo odpovídat $\ln B(z_j) - \ln B(z_k)$. Zde je vidět i jiná cesta k identifikaci funkce B(x). Jenom bychom fixovali v nějakém z_0 $B(z_0) = 1$ a tím i $A(t) = L(t,z_0)$. Tato cesta je vhodná zejména v případě, kdy zvyšujeme počet vybraných bodů z_j ($j = 1,\dots,K$, $K = K_N \to \infty$), nevýhodou je, že k identifikaci funkce A využijeme jen malou část dat.

Pokud se spokojíme s představou, že K a body $z_1,\dots,z_K$ zvolíme pevně, tak pomyslnou násobící konstantu můžeme vybrat tak, aby

$$\frac{1}{K}\sum_{j=1}^{K} B(z_j) = 1.$$

Potom ale máme $A(t) = \frac{1}{K}\sum_{j=1}^{K} L(t,z_j)$ a z odhadů KIP $L_N(t,z_j)$ tak také přímo získáme odhad $A_N(t)$. Pak už zbývá jen odhad K "parametrů" $B(z_j)$, třeba metodou nejmenších čtverců $B_N(z_j) = \arg\min_{B_j} \sum_{i=1}^{N} (L_N(T_i,z_j) - A_N(T_i).B_j)^2 . I[T_i \le T]$. Funkce B bude tedy charakterizována odhady funkčních hodnot ve vybraných bodech z_j.

Asymptotické vlastnosti odhadů KIP

Ať už je další postup jakýkoli, pro kvalitu identifikace funkcí A a B jsou rozhodující vlastnosti odhadů KIP. Představme si, že jsme v okolí bodu z (jeden z vybraných $z_1,\dots,z_K$), okolí se zužuje při celkovém počtu měření $N \to \infty$, ale v souladu s předpokladem 3 počet bodů v tomto okolí $M = M_N(z) \to \infty$. V každém okamžiku přečíslujeme měření tak, aby právě $x_1,\dots,x_M$ ležely v $O_{d_N}(z)$ a $T(x_1) \le T(x_2) \le \dots \le T(x_M)$. Odhad KIP v bodě $\underline{z}$ z realizací v

$O_{d_N}(z)$ pak je

$$L_N(t,z) = \sum_{i=1}^{M} \frac{\delta(x_i)}{M-i+1} \cdot I[T(x_i) \le t].$$

Věta 1. Jsou-li v okolí bodu z splněny předpoklady 1 - 3, pak

$$\limsup_{N\to\infty} \sup_{t\in[0,T]} (L_N(t,z)-L(t,z)) = 0 \text{ s. j.}$$

Důkaz může krok za krokem sledovat postup v [3]. Dík předpokladu 2 nám komplikace nepůsobí fakt, že empirické odhady jsou založeny na realizacích v bodech blízkých z, kdežto [3] se týká případu bez regrese. V [3] se vyšetřuje konvergence odhadu distribuční funkce, resp. jejího doplňku, což v našem modelu pro okolí bodu z je

$$P_N(t,z) = \prod_{i=1}^{N} \left(\frac{M-i}{M-i+1}\right)^{\delta(x_i).I[T(x_i)\le t]}$$

(s konvencí $0^0 = 1$ a $\delta(x_M) = 1$).
To je tzv. "product limit" odhad - viz i [2]. Ale k vlastnostem odhadu KIP se dostaneme lehce pomocí tvrzení, které je opakováno v mnoha článcích, i v [1]. My je potřebujeme ve tvaru:

Lemma: Za předpokladů 1-3 je $\limsup_{N\to\infty} \sup_{t\in[0,T]} \sqrt{M}(\ln P_N(t,z)+L_N(t,z)) = 0$ s. j.

Jak je vidět, maximálně využíváme předpokladu o spojitosti všech funkcí rozložení v proměnné x. Je možné, že při důkladnějším rozboru by se tento předpoklad dal výrazněji zeslabit (požadavek existence limit $\sum_{i=1}^{M} Q(t,x_i)/M$ při $x_i \in O_{d_N}(z)$ není výrazné zeslabení).

Věta 2. Nechť platí předpoklady 1 - 3, navíc spojitost všech distribučních funkcí v proměnné x je Lipschitzovská a šířka okna je volena tak, aby $\sqrt{M}.d_N \to 0$. Potom náhodná funkce $Z_N(t) = \sqrt{M}(L_N(t,z)-L(t,z))$ konverguje na $[0,T]$ slabě ke gaussovské náhodné funkci $Z(t)$, která je centrovaná a má následující kovariančí strukturu pro $0 \le s \le t \le T$:

$$\mathrm{cov}(Z(s),Z(t)) = C(s) = \int_0^s \frac{dF}{P^2Q}.$$

Důkaz tentokrát sleduje postup důkazu podobného tvrzení (opět bez regrese) v [1], část 7. Základ důkazu spočívá v tom, že $Z_N(t)$ vyjádříme jako lineární transformaci několika empirických procesů. Hlavní roli tu hrají distribuční funkce

$$\tilde{F}(t,z) = \Pr\{T(z) < t \ \&\ \delta(z) = 1\} = \int_0^t Q(s,z)dF(s,z),$$

$$H(t,z) = \Pr\{T(z) < t\} = 1 - P(t,z)Q(t,z),$$

které lehce odhadneme přímo z našich dat jako relativní četnosti. Tuď si tyto odhady představíme jako normované součty nezávislých nulajedničkových veličin:

$H_N(t,z) = \frac{1}{M}\sum \varepsilon(t,x_i)$, kde $Pr\{\varepsilon(t,x_i) = 1\} = H(t,x_i)$,

$\tilde{F}_N(t,z) = \frac{1}{M}\sum \tilde{\varepsilon}(t,x_i)$, $Pr\{\tilde{\varepsilon}(t,x_i) = 1\} = \tilde{F}(t,x_i)$.

Pak 2-rozměrný náhodný proces $\nu_N^*(t) = \sqrt{M}(H_N(t,z) - \frac{1}{M}\sum H(t,x_i))$, $\rho_N^*(t) = \sqrt{M}(\tilde{F}_N(t,z) - \frac{1}{M}\sum \tilde{F}(t,x_i))$ na $[0,T]$ je centrovaný a má kovariační strukturu danou kovariancemi n. v. ε a $\tilde{\varepsilon}$, pro $s \le t$ je například

$$cov(\nu_N^*(s),\rho_N^*(t)) = \frac{1}{M}\sum cov(\varepsilon(s,x_i),\tilde{\varepsilon}(t,x_i)) = \frac{1}{M}\sum (\tilde{F}(s,x_i) - H(s,x_i).\tilde{F}(t,x_i))$$

Předpoklady 1-3 by zřejmě stačily ke konvergenci kovariancí a tedy dík C.L.V. k asymptotické normalitě $\nu_N^*(t)$, $\rho_N^*(t)$, a jistě i k silné konzistenci $H_N(t)$ či $\tilde{F}_N(t,z)$ na $[0,T]$. Pro platnost naší věty potřebujeme navíc, aby $\sqrt{M}(\frac{1}{M}\sum H(t,x_i) - H(t,z)) \to 0$, totéž pro $\tilde{F}$. Proto ty dodatečné předpoklady ve Větě 2. Pokud volíme d_N tvaru $C.N^{-\alpha}$, tak požadavek $\sqrt{M}.d_N \to 0$ spolu s předpokladem ($M \sim N.d_N$) znamenají, že $\alpha \in (\frac{1}{3},1)$.

Máme teď k dispozici v pevně zvolených bodech $z_1,\dots,z_K$ dobré (aspoň asymptoticky) odhady kumulativních intenzit poruch, a chceme odhadnout funkci $A(t)$ a hodnoty $B(z_j)$, které jsme svázali podmínkou $\frac{1}{K}\sum B(z_j) = 1$.

Důsledek: a) Za předpokladů 1-3 je $A_N(t) = \frac{1}{K}\sum L_N(t,z_j)$ na $[0,T]$ silně konzistentním odhadem funkce $A(t)$, stejnoměrně v t.
b) Nechť jsou splněny předpoklady Věty 2. Označme $M_N^* = \sum_j M_N(z_j)$ a požadujme existenci limit $I_j = M_N(z_j)/M_N^*$ (které jsou kladné dík předpokladu 3). Pak $W_N = \sqrt{M_N^*}\,(A_N(t) - A(t))$ je na $[0,T]$ asymptoticky centrovaná gaussovská náhodná funkce, s kovariancí pro $0 \le s \le t \le T$:

$$ascov(W_N(s).W_N(t)) = \frac{1}{K^2}\sum_{j=1}^{K}\frac{C_j(s)}{I_j}, \text{ kde}$$

$C_j(s) = \int_0^s \frac{dF(u,z_j)}{P^2(u,z_j)Q(u,z_j)}$ (srovnej s asymptotickou kovariancí ve Větě 2).

Odhady metodou nejmenších čtverců pro $B(z_j)$ jsou (při označení $T_i = T(x_i)$) $B_N(z_j) = \sum L_N(T_i z_j)A_N(T_i)/\sum A_N(T_i)^2$, kde součet je přes $T_i \le T$.
Podíváme-li se na popsanou metodu z praktického hlediska, je jistě rozumné odhadnout $A(t)$ jako průměr kumulativních funkcí rizika na různých úrovních hodnot regresoru. Pak $B(z_j)$ charakterizují vliv regresoru na přežívání pro úroveň hodnot x okolo z_j.

Testování shody modelu a dat. Samozřejmě, předpokládáme-li určitý typ modelu, pak procedurami analýzy takovéhoto modelu dojdeme k jeho identifikaci, k odhadům parametrů, regresní funkce apod. Je dobré mít k ruce procedury, které otestují, zda zvolený typ je vůbec vhodný. Třeba v rámci Coxova modelu se pro test hypotézy $\gamma = \gamma_0$ nabízí statistika vytvořená na základě asymptotického rozložení odhadu γ. Ale tento test nám v případě zamítnutí hypotézy nepoví nic o případném lepším modelu. Jednoduchý postup pro test hypotézy H_0 : $B(x) = B_0(x)$, který je schopen ukázat i odchylky od této hypotézy, může vypadat následovně: Máme naše pozorování $\{T(x_i), \delta(x_i)\}$, $i = 1,\dots,N$ a hypotézu H_0. Pokud v hypotéze není

zahrnuta i funkce A(t), můžeme ji maximálně věrohodně odhadnout jako

$$A_0(t) = \sum_i \left\{ \delta(x_i) . I[T(x_i) \le t] / \sum_j B_0(x_j) I[T(x_j) \ge T(x_i)] \right\}$$

Pak se tato funkce pro nás stane součástí hypotézy. Transformujeme pozorování na $S(x_i) = 1 - [\exp -A_0(T(x_i)) B_0(x_i)] = F_0(T(x_i), x_i)$, kde F_0 představuje hypotetické distribuční funkce n. v. $Y(x_i)$. Dostaneme zase schema náhodného cenzorování s $S(x_i) = \min(U_i, W(x_i))$, $\delta(x_i)$ zůstávají tatéž, $W(x_i) = F_0(V(x_i), x_i)$ a $U_i = F_0(Y(x_i), x_i)$ jsou n. v. s rovnoměrným rozložením pravděpodobnosti na (0,1). Takže když z takto přetransformovaných pozorování uděláme PL-odhad distribuční funkce, při H_0 by měla projít testem dobré shody s distribuční funkcí $H_+(u) = u$ na (0,1). A tuto shodu lehce otestujeme třeba testy Kolmogorova-Smirnova, modifikovanými pro náhodně cenzorovaná data (viz Robust 84).

Ještě bychom potřebovali nějakou proceduru na testování nevýznamnosti regrese, tj. pro test hypotézy $B(x) \equiv 1$. K tomu se jistě dá využít předchozí test. Nebo můžeme tuto úlohu pojmout jako test rovnosti rozložení n. v. Y(x) pro různé úrovně veličiny x. K tomu by mohly posloužit již zmíněné testy na proporcionalitu. Nabízejí se i testy homogenity, zobecněné pro cenzorovaná data. O takových testech pro 2 výběry (zde 2 úrovně) se psalo už v Robustu 84. O testech pro více výběrů naráz je možné se dočíst například v Breslow N.: A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorhip, Biometrika 57(1970),No 3, 579-594.

Literatura

[1] *Breslow N., Crowley J. (1974)*, A large sample study of the life table and product limit estimates under random censorhip, AS 2 No 3, 437-453.

[2] *Lánská V. (1988)* - ROBUST 88.

[3] *Winter B. B., Földes A., Rejtö L. (1978)*, Glivenko-Cantelli theorems for the product limit estimate, PCIT 7, No 3, 213-225.