ODHADY REGRESNÍ FUNKCE V MODELU S CENZOROVANÝMI DATY

Petr V o l f, ÚTIA

Pokusíme se řešit úlohu odhadu regresní funkce v případě, že data měříme s náhodným censorováním. S takovými daty se můžeme setkat při zkoumání životnosti či doby bezporuchového provozu. Zmíníme se i o nenáhodném censorování, i o Coxově modelu popisujícím regresi jiným způsobem.

1. Model. Představme si náhodné veličiny $Y_1, Y_2, \ldots, Y_N$ odpovídající modelu

$$Y_i = r(x_i) + \varepsilon_i\ , \quad i=1,\ldots,N,$$

kde x_i jsou známé hodnoty,(pro jednoduchost zůstaneme v R_1), ε_i jsou nezávislé, stejně rozložené náhodné veličiny, centrované, se spojitou distribuční funkcí F, $r(.)$ je reálná funkce, buď neznámá, nebo známého typu s neznámým parametrem $\theta \in R_m$. V praxi takový model může být vhodný pro logaritmus doby bezporucového provozu.

Náhodné censorování zprava spočívá v tom, že nepozorujeme hodnoty Y_i, ale

$$T_i = \min(Y_i, V_i) \qquad a \qquad \delta_i = I[Y_i \leq V_i]\ ,$$

kde $I[A]$ označuje indikátor množiny A a $V_1, \ldots, V_N$ jsou nějaké náhodné veličiny, které mohou také záviset na hodnotách regresoru x_i. Předpokládáme, že tato závislost je "rozumná", třeba že je ji možné také vyjádřit ve formě modelu

$$V_i = v(x_i) + e_i\ ,$$

s náhodnými veličinami e_i nezávislými navzájem i na $\{\varepsilon_i\}$.

2. Odhady parametru. Můžeme být v situaci, kdy známe tvar regresní funkce r_θ a potřebujeme odhadnout hodnotu parametru θ. Zcela uspokojivý postup jsem zatím nenašel. Miller (1976) a později Buckley a James (1979) se snažili dospět k odhadu s nejmenšími čtverci iteračním postupem. Je navržen v několika modifikacích a výsledky v předvedených simulovaných příkladech nejsou špatné. Ale postupy jsou početně složité, navíc iterace mohou někdy začít oscilovat. Ani konsistence není zaručena.

Šikovně je odvozena metoda odhadu v práci Koul et al. (1981), pro případ veličin V_i nezávislých na x_i a stejně rozložených se spojitou distribuční funkcí G. Tu pak můžeme konsistentně odhadnout, při označení p_i=pořadí hodnoty T_i,

$$\hat{G}(t) = 1 - \prod_{i=1}^{N} \left(\frac{N-p_i+1}{N-p_i+2}\right)^{(1-\delta_i)\cdot I[T_i \leq t]}.$$

Je to tedy variace na Kaplanův a Meierův (1958) "product limit estimator"-PLE.

Všimneme-li si, že pro pevné x je

$$E_x(\delta_i T_i) = \int_{-\infty}^{\infty} t\,[1-G(t)]\; dF[t - r_\theta(x)]\ ,$$

tj.
$$E_x\left\{\frac{\delta_i T_i}{1-G(T_i)}\right\} = \int_{-\infty}^{\infty} t\; dF[t - r_\theta(x)] = r_\theta(x)$$

a vidíme, že náhodné veličiny $\delta_i T_i / [1-G(T_i)]$ splňují regresní model s toutéž regresní funkcí. Šumy v něm jsou stále ještě nezávislé, ale již různě rozložené a s daleko většími rozptyly. To činí metodu velice citlivou a nespolehlivou, jak ukazují výsledky. V praxi samozřejmě použijeme hodnot odhadu $\hat{G}(T_i)$, parametr pak odhadneme metodou nejmenších čtverců s hodnot $W_i = \delta_i T_i / [1-\hat{G}(T_i)]$. Autoři dokázali středně kvadratickou konsistenci takto získaných odhadů parametrů lineární funkce $r_\theta(x) = \alpha + \beta x$. Z důležitých předpokladů kromě spojitosti distribučních funkcí upozorňujeme na základní, odlišující náhodné censorování od nenáhodného useknutí, že

$$\sup\{t:\ G(t) < 1\} \geq \sup\{t:\ F(t) < 1\}\ .$$

Můžeme jej obejít, použijeme-li vhodně "useknutého" výběru.

3. Neparametrické odhady. K-odhad hodnoty regresní funkce v daném bodě x je v podstatě vážený průměr z hodnot naměřených v okolních bodech, váhy potlačují vliv měření ve vzdálených bodech x_i. Zvolíme nejjednodušší formu s "oknem" šířky $2\cdot d_N$:

$$O_{d_N}(x) = \{z \in R_1 : |x-z| \leq d_N\} .$$

Musíme si ještě všimnout rozmístění bodů x_i. Potřebujeme, aby ve zvoleném okolí bodu x byl dostatečný počet měření. Předpokládejme, že x_i jsou realizace náhodné veličiny $\mathcal{X}$ definované v $(R_1, \mathcal{B}_1)$, rozložené s hustotou $\underline{h}$ spojitou v bodě x, $h(x)>0$. Označíme M_N počet bodů v $O_{d_N}(x)$. Pokud $d_N \to 0$ a $N\cdot d_N \to \infty$,už máme zajištěno, že aspoň ve středním kvadrátě

$$M_N/(2d_N\cdot N) \to h(x), \text{ čili } M_N \to \infty \text{ v pravděpodobnosti.}$$

Volí se běžně tvar $d_N = C\cdot N^{-\delta}$, $\delta \in (0,1)$. Pohledem z druhé strany je volba d_N tak, aby byl zajištěn určitý počet bodů M_N -metoda N nejbližších bodů, např. Stone (1977). To spíš odpovídá přístupu v konkrétní praktické situaci.

Označíme-li $T_{k1} \leq \ldots \leq T_{kM}$ $(M=M_N)$ seřazené všechny výsledky v bodech $x_i \in O_{d_N}(x)$, uděláme odhad (PLE) distribuční funkce veličiny $\{Y/x_i \in O_{d_N}(x)\}$:

$$F^Y_{N,x}(t) = 1 - \prod_{j=1}^{M} \left(\frac{M-j}{M-j+1}\right)^{\delta_{kj}\cdot I[T_{kj} \leq t]} .$$

Tato funkce má skoky $\Delta F^Y_{N,x}$ v bodech T_{kj} s $\delta_{kj}=1$. Pokládáme $\delta_{kM}=1$, aby $\sum \Delta F_{N,x} = 1$.

V práci Volf (1985) je dokázána pro $N \to \infty$ P konzistence odhadu

$$r_N(x) = \sum_{i=1}^{M} T_i \cdot \Delta F^Y_{N,x}(T_i)\cdot I\left[|T_i| \leq A_N\right] ,$$

a to za následujících předpokladů:

1. Volíme $d_N = C\cdot N^{-\delta}$, $C>0$, $\delta \in \left(\frac{2}{3}, 1\right)$!?
2. Body x_i jsou rozmístěny tak, že $M_N \to \infty$.
3. Funkce r , v jsou spojité v bodě x.
4. Veličiny e_i jsou stejně rozděleny ,s distribuční funkcí G (to není podstatné) a distribuční funkce F a G jsou stejnoměrně spojité.
5. Označíme-li $\vartheta_F = \sup\{t : F(t)<1\}$,tak existují a,b $\in (0,1\rangle$ tak, že
$$a\cdot\left[1-F(t)\right]^b \leq 1-G(t) \text{ pro každé t z nějakého intervalu } \langle\vartheta, \vartheta_F\rangle.$$
6. A_N je posloupnost čísel, $A_N \to \infty$,$A_N\cdot\left[\frac{\log(N\cdot d_N)}{(N\cdot d_N)^a}\right]^{\frac{1}{2+b}} \to 0$.

Příkré omezení hodnot δ nevypadá nejlépe, pokusím se ukázat, k čemu bylo potřeba. Označím-li $\tau_i = \min(\varepsilon_i, e_i + v(x) - r(x))$, dostanu se v důkazu k bodu, kdy vím, že

(1) $$\left|r_N(x) - r(x)\cdot\sum \Delta F^Y_{N,x}(T_i) - \sum \tau_i\cdot\Delta F^Y_{N,x}(T_i)\right| = O(d_N)$$ /součet je přes i: $|T_i| \leq A_N$/

Přitom podmínka 5. zaručuje podle Rejtő (1982) s.j. stejnoměrnou v t konzistenci odhadu $F_{N,x}$ pro F, konstruovaného jako PLE z hodnot $\{\tau_{kj}\}_{j=1}^{M_N}$. Podmínka 6. pak zajišťuje silnou konzistenci odhadu střední hodnoty ε_1:

$$\sum \tau_i\cdot\Delta F_{N,x}(\tau_i) \to E\,\varepsilon_1 = 0 ,$$

to je vlastně vedlejší výsledek pro odhad parametru polohy. Pokud následující pravděpodobnost

(2) $$\Pr\left\{\tau_{k1} < \tau_{k2} \leq \ldots \leq \tau_{kM_N} \text{ a zároveň } \delta_{kj} = I\left[\varepsilon_{kj} \leq e_{kj} + v(x) - r(x)\right], j=1,\ldots,M_N\right\}$$

poroste k jedné s $N \to \infty$, bude tvrzení dokázáno, jak vidíme z (1), protože pak i

$$\Pr\left\{\Delta F^Y_{N,x}(T_{kj}) = \Delta F_{N,x}(\tau_{kj}), j=1,\ldots,M_N\right\} \to 1 .$$

Přitom T_{kj} se od τ_{kj} liší o $r(x) + O(d_N)$ takže k (2) stačí, aby $\Pr\left\{\min_{i \neq j} \frac{|\tau_i - \tau_j|}{d_N} > K\right\} \to 1$ pro libovolné K. Není těžké ukázat, že při 4. podmínce je pro $\alpha > 2$ a libovolné K

$$\Pr\left\{\min_{i\neq j}\frac{|\tau_{ki}-\tau_{kj}|}{M_N^{-\alpha}} > K\right\} \to 1 .$$

Pak už stačí vzít $d_N = M_N^{-\alpha}$ při $\alpha > 2$, tj. $N^{-1} = C_1(N^{1-\beta})^{-\alpha}$, což vymezí β na interval $(\frac{1}{3},1)$.

Výsledky příkladů se simulovanými daty ukazují, že toto omezení asi není nutné, zároveň je však pravda, že asymptotické vlastnosti se zde projevují velice pomalu. Takže i volbou C se dá vhodně určit rozměr okna pro značně velké rozpětí N. Příklady, v nichž bylo $\mathcal{X}$ rozloženo rovnoměrně, ukazují, že asi 10 až několik desítek bodů pro rozsahy N=50 - 1 000 jsou nejvhodnější.

4. Znovu parametrický případ - a robustnost. Protože zatím v části 2 uvedené metody odhadu parametru regresní funkce r_θ nejsou uspokojivé, tento případ jsme zkusili řešit následujícím způsobem: V 1. kroku odhadneme neparametricky hodnoty regresní funkce, buď ve všech x_i, nebo jen ve vybraných bodech. V 2. kroku pak odhadneme parametr θ z takto získaných hodnot $r_N(x_j)$, třeba metodou nejm. čtverců.

Metody robustního odhadování sem můžeme vnést dvěma způsoby. Buď do 1. kroku, kdy počítáme "průměr"; můžeme použít některé z technik robustního odhadování, třeba trimování. Druhý krok můžeme řešit s trimováním pomocí regresních kvantilů místo prostou metodou nejm. čtverců. Tak bychom se mohli bránit proti vlivu neočekávaných odchylek v šumech ε_i. Jenže se ukazuje, že to není vždy nutné. Samotný K-odhad sice není robustní, ale vyrovnává data a odchylky tedy zmenšuje. Přidání robustní metody již nepřinese výrazný pokrok, nejsou-li data příliš nečistá.

Pokud jde o pozorování v odlehlých bodech x_i, zvláště na krajích oboru $\mathcal{X}$, mají ... nadále větší vliv, což je o to horší, že v nich jsou K-odhady počítány z méně bodů (a jen z jedné strany). Takže v 2. kroku by se měly použít běžné metody na oslabení vlivu pozorování v odlehlých bodech na odhad parametru. Stejně tak se musíme zmínit o dalších typických vlastnostech jednoduchých K-odhadů, že totiž podceňují konkávní regresní funkci a přeceňují konvexní.

O neparametrických odhadech regresní funkce je možné se dočíst v tomto sborníku v práci J. Antocha, dále např. v sborníku Gasser a Rosenblatt (1979).

5. Cenzorování nenáhodné. Zvláště v technických aplikacích se setkáváme s daty cenzorovanými v okamžiku jejich sběru, při současných počátcích pozorování. Pak vysoké hodnoty jsou useknuty, říká se tomu cenzorování časem. V jiném případě mohou být useknuty i malé hodnoty, jsou třeba mimo rozsah přístroje. Veličiny Y_i pak pozorujeme přímo jen v nějakém intervalu $\langle\vartheta_1, \vartheta_2\rangle$. To nám jistě připomene data usečená záměrně při trimování. Pokud předpokládáme symetricky rozložené šumy ε_i, nic nám nebrání v parametrizovaném případě výběr symetricky trimovat pomocí regresních α-kvantilů, α zvolíme tak velké, abychom se zbavili všech cenzorovaných hodnot. Pokud se nemůžeme spolehnout na symetrii odchylek nebo je-li cenzorování příliš veliké, je lépe se uchýlit k regresnímu mediánu. O těchto postupech viz Antoch, Robust 84.

Při neparametrickém odhadování hodnoty regresní funkce můžeme opět odhad v každém bodě počítat jako trimovaný průměr, až jako medián v krajním případě. Stejně se dá postupovat při kombinovaném cenzorování (usečené kraje a ještě náhodné cenzorování mezi nimi).

6. Coxův regresní model. Duchovním otcem je D.R.Cox (1972), i když vlastně rozvíjí model závislosti parametru exponenciálního rozložení. Popisuje vývoj intenzity poruch (hazard rate), což je veličina definovaná jako $\lambda(t) = -d \ln[1-F(t)]/dt$, když $F(\cdot)$ je distribuční funkce. Předpokládá a modeluje tedy vývoj rozložení veličiny v závislosti na x, nikoli vývoj střední hodnoty. V nejjednodušším případě má tvar

$$\lambda^{Y}(t,x) = \lambda_o^{Y}(t)\cdot \exp(\beta x) .$$

Je vidět, β vesměs záporná znamenají pokles intenzity poruch s růstem x, tedy ... životnost. Dobře se dá tento model začlenit do věrohodnostní funkce, takže je možné dospět k maximálně věrohodným odhadům parametru β a pak i k odhadu kumulativní intenzity poruch definované jako $L(\tau)=\int_{-\infty}^{\tau}\lambda(t)dt$. Maximalizací parciální věrohodnostní funkce

$$\mathcal{L}(\beta)=\prod_{i=1}^{N}\left\{\exp(\beta\cdot x_i)\Big/\sum_{j=1}^{N}\exp(\beta x_j)\cdot I[T_j\geq T_i]\right\}^{\delta_1}$$

dostaneme odhad $\hat{\beta}$. Nabízí se iterační postup, konečné řešení je silně konzistentní. Potom můžeme také odhadnout

$$\hat{L}_0^{\gamma}(t)=\sum_{i=1}^{N}\frac{\delta_i\cdot I\,[T_i\leq t]}{\sum_{j=1}^{N}\exp(\hat{\beta}x_j)\cdot I[T_j\geq T_i]}\ .$$

Navíc asymptotická normalita odhadů $\hat{\beta}$ umožňuje testovat významnost regrese. Z literatury na toto téma uveďme článek Tsiatise(1981), variantu pro tříděná data včetně ukázky praktického použití v Prentice, Gloeckler(1978) i knihu Kalbfleisch, Prentice(1980), kde je i ukázka programu . Novější verze BMDP také již obsahují tuto proceduru. Někdy jsou obě pojetí regrese ekvivalentní, odpovídá-li log Y lineárnímu modelu s dvojně exponenciálními šumy, odpovídá Y Coxově modelu a je rozloženo Weibullovsky. Výhodou Coxova modelu je, že nevyžaduje žádné specifické vlastnosti od censorující náhodné veličiny V; může také záviset na x, může i veličinu Y uřezávat tj. $\sup\{V\}\leq\sup\{Y\}$, i tím spíš odpovídá praktické situaci s kombinovaným cenzorováním.

<u>Literatura:</u>

Antoch J., sborníky Robust 1984 a Robust 1986.

Buckley J., James I.(1979),Linear regression with censored data, Biometrika 66, 429.

Cox D.R.(1972),Regression models and life tables -with discussion, J.R.Statist.Soc. B 34, 187-220.

Gasser Th.,Rosenblatt M. -editors(1979),Smoothing Techniques for Curve Estimation, Lecture Notes on Math. No 757, Springer-Verlag.

Kalbfleisch J.D.,Prentice R.L.(1980),The Statistical Analysis of Failure Time Data, N.Y.,Wiley.

Kaplan E.L.,Meier P.(1958),Nonparametric estimation from incomplete observations, J.Amer.Stat.Assoc. 53,457-81.

Koul H.,Susarla V.,Van Ryzin J.(1981),Regression analysis with randomly right-censored data, Ann.Stat.9,1276-88.

Miller R.G.(1976),Least squares regression with censored data,Biometrika 63,449-64.

Prentice R.L.,Gloeckler L.A.(1978),Regression analysis of grouped data with application to breast cancer data, Biometrics 34, 57-67.

Rejtö L.(1982),On the fixed censoring model and consequences for the stochastic case,9th Prague Conf. on Inf. Theory ...,Vol B, Academia.

Stone C.J.(1977),Consistent nonparametric regression -with discussion, Ann.Stat. 5, No 4.

Tsiatis A.A.(1981),Large sample study of Cox's regression model, Ann.Stat.9, 93-108.

Volf P.(1985),Estimation in linear model with censored data, posláno do Proc. of 5th Pannonian Symposium on Math. Stat., Budapest.