ROBUST 2006, Lhota nad Rohanovem, 23. – 27. ledna 2006

Zobecněné lineární modely pro značkované bodové procesy

David Kraus

ÚTIA AV ČR Praha & KPMS MFF UK Praha, kraus@karlin.mff.cuni.cz, http://www.karlin.mff.cuni.cz/~kraus/robust2006/

Motto: *Mladá nadějná generace hodně pochytila od svých učitelů a daří se jí slibně komplikovat jednoduché úlohy.* Jiří Žváček v hodnocení ROBUSTu 2004
http://www.stahroun.me.cz/eseje/robust2004/

1 Úvod

Uvažujme longitudinální pozorování

$$T_{ij},\ Z_i(T_{ij}),\ (X_{i1}(T_{ij}),\dots,X_{ip}(T_{ij}))^\mathsf{T},\quad j=1,2,\dots,$$

tedy pro i-tého jedince pozorujeme v náhodných časech $T_{i1} < T_{i2} < \dots$ značky $Z_i(T_{ij})$ a kovariáty $(X_{i1}(T_{ij}),\dots,X_{ip}(T_{ij}))^\mathsf{T}$. Cílem je modelovat závislost rozdělení značek na kovariátách pomocí *zobecněných lineárních modelů* (GLM) s časově závislými koeficienty.

2 Značkované bodové procesy

Dvojice $T_{ij}, Z_i(T_{ij})$ tvoří *značkovaný bodový proces* (MPP) v časovém intervalu $[0,\tau]$ se značkami v nějakém prostoru E (Brémaud, 1981). Proces označme $p_i(dt\times dz_i)$. Jedná se o náhodnou čítací míru na $[0,\tau]\times E$. Proces $p_i(dt\times A) = \int_A p_i(dt\times dz_i)$ (A je nějaká borelovská množina) je čítací proces adaptovaný na nějakou filtraci, tzn. $p_i(dt\times A)$ počítá události se značkami v A.

Chování procesu je popsáno jádrem intensity

$$\lambda_i(t,dz_i) = \lambda_i(t)\Phi_i(t,dz_i).$$

Zde $\lambda_i(t)$ je *intensita* procesu $p_i(dt\times E)$ (proces všech událostí bez ohledu na jejich značky). $\Phi_i(t,dz_i)$ je *podmíněné rozdělení značek* v čase t za podmínky historie do t a toho, že t je časem pozorování.

3 GLM pro MPP

Modelujme rozdělení $\Phi_i(t,dz_i)$ pomocí GLM. Předpokládejme, že jeho střední hodnota $\mu_i(t)$ závisí na lineárním prediktoru $\eta_i(t) = X_i(t)^\mathsf{T}\beta(t)$ pomocí linkové funkce g, čili

$$g(\mu_i(t)) = \eta_i(t) = X_i(t)^\mathsf{T}\beta(t).$$

Rozptyl v rozdělení $\Phi_i(t,dz_i)$ předpokládejme ve tvaru $\psi(t)V(\mu_i(t))$, kde $\psi(t)$ je dispersní parametr a $V(\mu)$ je varianční funkce. Jak koeficienty $\beta(t)$, tak disperse $\psi(t)$ mohou záviset na čase (uvažujeme *neparametrický model*). Speciálním případem je lineární model, který uvažovali Martinussen & Scheike (2001).

4 Odhad

Odhadneme kumulativní koeficienty $B(t) = \int_0^t \beta(s)ds$. *Proč kumulativní?* Abychom mohli o nich dělat inferenci. Předpokládejme, že intensita $\lambda_i(t)$ splňuje Aalenův aditivní model, tj. $\lambda_i(t) = Y_i(t)U_i(t)^\mathsf{T}\alpha(t)$, kde $U_i(t)$ jsou nějaké kovariáty a $\alpha(t)$ koeficienty. *Proč předpokládáme nějaký model intensity časového procesu?* Protože chceme odhadovat kumulativní efekty $B(t)$, nikoli $\beta(t)$.

Odhadovací rovnicí je

$$\sum_{i=1}^n \frac{X_i(t)}{g'(\mu_i(t))\psi(t)V(\mu_i(t))}\left[\int_E z_i p_i(dt\times dz_i) - \hat\lambda_i(t)\mu_i(t)dt\right] = 0$$

($\hat\lambda_i(t)$ se získá standardně vyhlazením odhadu z Aalenova modelu). *Kde se rovnice vzala?* Lze ji motivovat jako *kvazivěrohodnostní rovnici*, poněvadž

$$\frac{X_i(t)\lambda_i(t)}{g'(\mu_i(t))} = \frac{\partial}{\partial\beta(t)}\mathsf{E}\left[\int_E z_i p_i(dt\times dz_i)\middle|\mathcal{F}_{t-}\right] = \frac{\partial}{\partial\beta(t)}\lambda_i(t)\mu_i(t)dt$$

a

$$\psi(t)V(\mu_i(t))\lambda_i(t)dt = \mathrm{var}\left[\int_E z_i p_i(dt\times dz_i)\middle|\mathcal{F}_{t-}\right].$$

K rovnici lze dojít také z věrohodnosti spočtením skorových operátorů.

5 Algoritmus: IRLS s vyhlazováním

Rovnici řešíme pomocí *iterativně vážených nejmenších čtverců* (IRLS), přičemž mezi jednotlivými kroky spočtené přírůstky *vyhladíme* (jádrově). Odhad je inspirován metodou popsanou v Martinussen, Scheike & Skovgaard (2002). *Proč vyhlazujeme?* Iterovat v každém bodě pozorování odděleně nemůžeme, protože v každém bodě je nejvýše jedno pozorování.

Označme $\tilde B(t)$ předchozí odhad. Průběh iterace:

(1) Vyhlazením $\tilde B(t)$ získej $\tilde\beta(t)$.

(2) Spočti 1 krok IRLS pro všechny časy pozorování $t\in[0,\tau]$:

(2i) Spočti pracovní odezvu (working response)

$$\tilde r_i(t)dt = \tilde\eta_i(t)dt + \frac{g'(\tilde\mu_i(t))}{\hat\lambda_i(t)}\left[\int_E z_i p_i(dt\times dz_i) - \hat\lambda_i(t)\tilde\mu_i(t)dt\right].$$

(2ii) Spočti váhovou matici $W(t) = \mathrm{diag}[W_i(t)]$, kde

$$W_i(t) = \frac{\hat\lambda_i(t)}{g'(\tilde\mu_i(t))^2 V(\tilde\mu_i(t))}.$$

(2iii) Lineární regresí $\tilde r(t)dt$ na $X(t)$ pomocí vážených nejmenších čtverců získej přírůstky nové iterace $d\tilde B(t)$

$$\begin{aligned} d\tilde B(t) &= \left[X(t)^\mathsf{T}W(t)X(t)\right]^{-1}X(t)^\mathsf{T}W(t)\tilde r(t)dt \\ &= \tilde\beta(t)dt + \left[X(t)^\mathsf{T}W(t)X(t)\right]^{-1}X(t)^\mathsf{T}W(t) \\ &\times \mathrm{diag}\left[\frac{g'(\tilde\mu_i(t))}{\hat\lambda_i(t)}\right]\left[\int_E zp(dt\times dz) - \hat\lambda(t)\tilde\mu(t)dt\right]. \end{aligned}$$

(3) Jdi na (1).

Algoritmus je pochopitelně nutné někde nastartovat: například z lokálně polynomiálního odhadu $\beta(t)$ (zde je použit lokálně konstantní odhad).

6 Testování konstantnosti

Testujme *konstantnost* funkcí $\beta(t)$ proti alternativě závislosti efektů na čase. Konstantnost $\beta(t)$ je ekvivalentní linearitě $B(t)$. Test (typu KS, například) můžeme založit na procesu

$$\hat B(t) - \frac{t}{\tau}\hat B(\tau).$$

K aproximaci jeho rozdělení použijeme modifikaci *simulační metody* Lin, Wei & Ying (1993) původně navržené pro čítací procesy.

Pro proces $R(t) = \hat B(t) - B(t)$ za určitých předpokladů platí

$$R(\cdot) = \int_0^\cdot\int_E H(s,z)^\mathsf{T}q(ds\times dz) + o_P(n^{-1/2}),$$

kde $q_i(dt\times dz_i) = p_i(dt\times dz_i) - \lambda_i(t)\Phi_i(t,dz_i)$ jsou martingaly a $H(s,z)$ je $(n\times p)$-matice nějakých predikovatelných procesů.

V této *martingalové representaci* můžeme $H(s,z)$ odhadnout. Jediné, co nemůžeme pozorovat ani odhadnout, jsou martingaly $q_i(dt\times dz_i)$. Nahradíme je v každém okamžiku pozorování T_{ij} simulovanými hodnotami $q_i^*(dt\times dz_i) = G_{ij}p_i(dt\times dz_i)$, kde G_{ij} jsou iid $\mathrm{N}(0,1)$ nezávislé na datech. *Proč to funguje?* Protože limita podmíněného rozdělení (při datech) takto vyrobeného procesu $R^*(t)$ je stejná jako limitní rozdělení procesu $R(t)$.

Generováním vhodného počtu realisací $R^*(t)$ můžeme procesem $R^*(t) - \frac{t}{\tau}R^*(\tau)$ aproximovat rozdělení testového procesu $R(t) - \frac{t}{\tau}R(\tau)$.

Poděkování a prosba

Poděkování. Děkuji za podporu, jíž se mi dostalo prostřednictvím grantů 201/05/H007 a 402/04/1294.
Prosba. Pokud by někdo měl reálná data zde popsaného typu a byl mi je ochoten poskytnout, byl bych mu velice vděčný.

Odkazy

Brémaud, P. (1981). *Point Processes and Queues. Martingale Dynamics*. Springer, New York.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557–572.

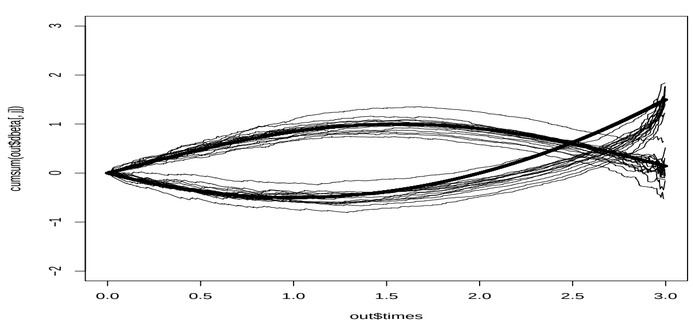
Martinussen, T. & Scheike, T. H. (2001). Sampling adjusted analysis of dynamic additive regression models for longitudinal data. *Scand. J. Statist.*, 28, 303–323.

Martinussen, T., Scheike, T. H. & Skovgaard, I. M. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scand. J. Statist.*, 29, 57–74.

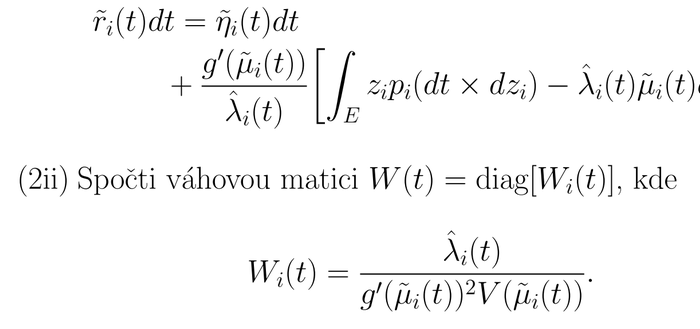
7 Ilustrace: odhady a testy

Obrázky 1–3 ukazují fungování *odhadů* na simulovaných datech. Silnou čárou jsou znázorněny skutečné funkce $B_1(t)$ a $B_2(t)$, tenkou čárou 15 odhadů z 15 vygenerovaných výběrů (každý o rozsahu 100). Ve všech případech je kovariáta $X_1 = 1$ a $X_2 \sim \mathrm{U}(-2,2)$. Intensita je konstantní $\lambda(t) = 2$.

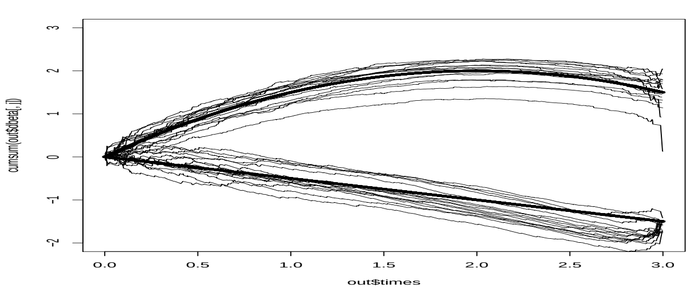
Obr. 1: Poissonovská regrese s (kanonickým) linkem $g(\mu) = \log\mu$; $\beta_1(t) = t-1$, $B_1(t) = t^2/2-t$, $\beta_2(t) = \cos t$, $B_2(t) = \sin t$.



Obr. 2: Normální regrese s linkem $g(\mu) = \mu$; $\beta_1(t) = 2\cos 5t$, $B_1(t) = 2\sin(5t)/5$, $\beta_2(t) = 2-t$, $B_2(t) = 2-0.5(t-2)^2$.



Obr. 3: Poissonovská regrese s linkem $g(\mu) = \log\mu$; $\beta_1(t) = -0.5$, $B_1(t) = -0.5t$, $\beta_2(t) = 2-t$, $B_2(t) = 2-0.5(t-2)^2$.



Obrázek 4 ukazuje *testy konstantnosti* regresních funkcí $\beta_1(t)$ a $\beta_2(t)$ pro výběr o rozsahu 100 vygenerovaný z modelu z obr. 3. Silná čára zachycuje testový proces, slabé čáry 50 simulovaných trajektorií z rozdělení testového procesu za platnosti hypotesy konstantnosti.

Obr. 4: Test konstantnosti $\beta_1(t)$ (nahoře) a $\beta_2(t)$.