REGRESNÍ MODELY V ANALÝZE PŘEŽÍVÁNÍ

Petr VOLF, ÚTIA ČSAV, Praha

Každá oblast využití matematické statistiky má své zvláštnosti, nejinak tomu je i s analýzou přežívání. Obecněji by se dalo mluvit o analýze výskytu událostí v čase (event history analysis), analýze proudu událostí. Odlišnosti se projevují ve výběru modelů, ve výběru charakteristik pro analýzu, i ve formě dat. Oblast analýzy přežívání se vyznačuje tím, že se často modeluje a zkoumá intenzita (riziková funkce, hazard rate) distribuce doby čekání na nějakou událost. Budeme se proto zabývat modely pro intenzitu, odhady, analýzou regrese v modelech s intenzitou. Pro seznámení s tématem a základními výsledky stačí sborníky "ROBUSTů", viz V. Lánská (1988) i J. Hurt (1992). Dokonalý přehled metodologie poskytne kniha Prentice a Kalbfleisch (1980). Moderní trendy modelování intenzit pro counting procesy (načítající v čase pozorované události) je možné sledovat v článcích P. K. Andersena a dalších (také E. Arjas, 1989). Současný stav oblasti je popsán v právě vycházející knize Andersen, Borgan, Gill, Keiding (1992). V tomto pojetí už je intenzita hlavní charakteristikou modelu, od ní se pak dají odvodit distribuce dílčích veličin, jako např. doba do k-té události, počet událostí do určité doby apod. Tento přístup se pokusíme vyložit v souvislosti s modelováním regrese.

Začneme s klasickou definicí intenzity jako charakteristiky distribuce nějaké náhodné veličiny T – např. doby do poruchy. Soustředíme se na následující okruhy otázek:

1. Definice, typické průběhy intenzity rozdělení pravděpodobnosti.
2. Neparametrické metody odhadu intenzity, vlastnosti odhadů.
3. Regresní modely pro counting procesy, multiplikativní model, Coxův model.
4. Aditivní regresní funkce, její neparametrické odhadování metodou lokální věrohodnosti.
5. Jiné užívané modely pro intenzitu, možnosti testování typu regresního modelu.

1 INTENZITA ROZDĚLENÍ PRAVDĚPODOBNOSTI

Mějme n. v. T se spojitým rozdělením pravděpodobnosti na $[0,\infty)$. Hustotu její distribuce označme $f(t)$, distribuční funkci $F(t)$. $P(T) = 1 - F(t)$ se nazývá *funkce přežití* a *intenzita* je definována jako

$$h(t) = \lim_{\Delta \to 0+} \frac{F(t+\Delta) - F(t)}{\Delta \cdot (1 - F(t))} = \frac{f(t)}{P(t)} = -d \log P(t)/dt.$$

Zavádí se ještě *kumulativní intenzita* $H(t) = \int_0^t h(s)\,ds = -\log P(t)$. Nyní bychom mohli uvést příklady typů distribucí, které se běžně používají pro popis doby přežití a v dílčích úsecích časové osy často popisují přežívání dosti přesně. Je to např. Weibullovo rozdělení s $P(t) = \exp -(t/\alpha)^\mu$, čili s $h(t) = \mu/\alpha \cdot (t/\alpha)^{\mu-1}$. Takže intenzita (poruch, řekněme) roste s časem, je-li $\mu > 1$, naopak klesá při $\mu < 1$. Mezní distribucí pak je pro $\mu = 1$ exponenciální rozdělení, s konstantní intenzitou $h = 1/\alpha$. Další používané typy distribucí jsou gamma rozdělení, s $f(t) = \nu^\lambda t^{\lambda-1} e^{-\nu t}/\Gamma(\lambda)$, $\lambda, \nu > 0$, nebo i lognormální rozdělení. Pro $\Gamma(\nu, \lambda)$ distribuci je limitní hodnota $\lim_{t\to\infty} h(t) = \nu$, přičemž je $h(t)$ klesající pro $\lambda < 1$, rostoucí pro $\lambda > 1$. Lognormální distribuce nemá intenzitu monotónní, limitní hodnoty v 0 i v ∞ jsou 0.

My se ovšem nechceme omezovat na určitou parametrizovanou rodinu distribucí, naším cílem je neparametrický způsob popisu a odhadu průběhu intenzity. Velice často má intenzita (doby přežití)

$h(t)$ zhruba *"vanovitý"* průběh na definičním oboru hodnot n. v. T, tj. zpočátku klesá, pak je téměř konstantní, ke konci roste (projeví se "stárnutí"). To se týká nejen například elektronických součástek, ale i živých organismů. I rizikové funkce po náročné operaci, i třeba rizika úmrtí během celého lidského života. Přesto máme tendenci chápat onu charakteristiku pro neživé a živé objekty trochu jinak. Soubor součástek považujeme za více méně homogenní, značná variabilita živých organismů nás daleko více nutí vzít v úvahu tuto nesourodost – *heterogenitu*. Může být způsobena vlivy genetickými, prostředím, životosprávou apod. Někdy se do modelu zavádí další n. v. ξ, která má za úloh popsat tu rozmazanost (variabilitu) určení rizikové funkce. Pak je intenzita náhodná $h(t) = \xi \cdot h_0(t)$. Aby model byl identifikovatelný, normuje se veličina "dispozice" ("frailty", = také "křehkost") tak, že $E\xi = 1$. Často se pro ni používá gamma distribuce s $\lambda = \nu$. Jiná možnost je pochopitelně popsat část heterogenity pomocí vhodných vysvětlujících proměnných a regresního modelu. Dostáváme se tak k multiplikativnímu modelu (také zvanému "s proporcionální intenzitou"), ale o tom později. Toto pojetí heterogenity předpokládá, že intenzita má zhruba týž tvar pro různé skupiny sledovaného souboru, ale riziko je zesílené či zeslabené nějakou dispozicí, či působením nějaké kovariáty.

Jindy bývají tvary intenzit pozorované z nesourodého souboru výsledkem agregace veličin několika typů. Takový tvar pak už nepopisuje "šanci" jednotlivého objektu, protože vznikl kompozicí ze skupin mající zcela odlišný tvar intenzity.

Mnoho o kompozici intenzit poruch pro technické objekty je v klasické knize Barlow, Proschan (1967). Nejznámější výsledek je následující: Označme $P(t/v)$, $h(t/v)$ funkci přežití a intenzitu pro rozdělení n. v. T při hodnotě n. v. (příznaku) $V = v$. Nechť $\varphi(v)$ je hustota rozdělení n. v. V. Pak agregovanou funkcí přežití myslíme $P(t) = \int P(t/v)\varphi(v)\,dv$, příslušná intenzita poruch je $h(t) = -d\log P(t)/dt$.

TVRZENÍ. Pokud intenzity poruch $h(t/v)$ jsou konstantní, pak $h(t)$ je nerostoucí.

Viděli jsme zatím dvě oblasti užití modelů pro rizikovou funkci, a to oblasti analýzy spolehlivosti a biostatistiky. Další oblastí, kde se zkoumají doby trvání jevů, je pochopitelně demografie, i ekonomie a sociální vědy vůbec. Použití modelů pro intenzity je nutně spjato se sledováním vývoje systému v čase, a tedy zavádí do statistické analýzy dynamický prvek. To vede k chápání dat jako realizace náhodného procesu, což už je dost důležitý posun od i. i. d. schématu, ať už v metodologii, či v možnostech teoretické podpory pro výsledky analýzy.

2 METODY ODHADU INTENZITY ROZDĚLENÍ PRAVDĚPODOBNOSTI

Distribuci (náhodné veličiny T – doby od poruchy apod.) jsme charakterizovali čtyřmi vzájemně svázanými funkcemi. Dvě z nich mají kumulativní charakter, F (resp. P) a H – kumulativní intenzita, další dvě jsou lokální, f a h. Mějme zatím k dispozici i. i. d. náhodný výběr $T_1, \ldots, T_n$. Když se konstruuje empirická distribuční funkce, každému realizovanému bodu T_i se jakoby přiřadí váha $\Delta F_n(T_i) = 1/n$. Podobně váha přiřazená každému realizovanému datu T_i pro odhad intenzity je $\Delta H_n(T_i) = 1/R_i$, kde R_i je počet rizikových objektů v okamžiku T_i^-. Kumulativní charakteristiku v t pak standardně odhadneme součtem přes $T_i \le t$. Empirickou funkci přežití si ale můžeme také představit jako součin: Nechť $T_1 < T_2 < \cdots < T_n$. Pak $P_n(T_i) = P_n(T_{i-1}) \cdot P_{R_i}(T_i) = \prod_{j=1}^{n} 1\{T_j \le$

$T_i](R_j-1)/R_j = \prod 1[T_j \le T_i](1-1/R_j)$. V tomto rozkladu je také každému datu T_j přiřazena váha $1/R_j$.

Nyní si představme složitější situaci s cenzorováním, které je u životnostních dat častým jevem. Modelujme cenzorování pomocí n. v. V se spojitou distribuční funkcí G na $[0,\infty)$, s $Q = 1-G$. Pak pozorujeme veličiny $Y_i = \min(T_i, V_i)$, $\delta_i = 1[T_i \le V_i]$. Toto je schéma náhodného cenzorování zprava, s tím se nejčastěji setkáme. Předpokládáme přitom vzájemnou nezávislosti n. v. $\{V_i\}$ mezi sebou i s $\{T_i\}$. Když nyní každému momentu, ve kterém nastal sledovaný jev (tj. bodu $(Y_i, \delta_i = 1)$) přiřadíme znovu váhu $1/R_i$, pak pro kumulativní intenzitu poruch dostaneme odhad

$$H_n(t) = \sum_{i=1}^{n} 1[Y_i \le t]\frac{\delta_i}{R_i}, \qquad H_n(t) = 0 \quad \text{pro } t < \min\{Y_i\}.$$

To je tzv. Nelson-Aalenův odhad. Z předchozího rozkladu empirické funkce přežití dostaneme odhad

$$P_n(t) = \prod_{i=1}^{n} 1[Y_i \le t]\cdot\left(1-\frac{\delta_i}{R_i}\right), \qquad P_n(t) = 1 \quad \text{pro } t < \min\{Y_i\},$$

což je známý Product Limit Estimate Kaplana a Meiera (viz znovu i Lánská 1988, Hurt 1992). Vlastnosti těchto odhadů jsou i nadále velice dobré, byly popsány v mnoha článcích (např. Breslow, Crowley 1974), a také v Robustu 84, kde se mluvilo o modifikaci testu Kolmogorova-Smirnova pro cenzorovaná data.

Označme $S_T = \sup\{t : F(t) < 1\}$, $S_V = \sup\{t : G(t) < 1\}$, $S = \min(S_V, S_T)$.

Platí:

1. $P_n(t)$ je silně konzistentním odhadem, stejnoměrně v $t \in [0, S)$, s řádem

$$\sup_{t<S}|P_n(t) - P(t)| \sim \mathcal{O}\left[\left(\frac{\log n}{n}\right)^{\frac{1}{2+b}}\right], \quad \text{s. j.},$$

pro nějaké $b > 0$ (za podmínek, které jsou formulovány např. v příspěvku L. Rejtö na 9. Pražské konferenci, 1982). Pokud $S < S_V$, pak $b = 0$.

$H_n(t)$ není ohraničená, proto pro ni je dokázáno totéž jen na každém ohraničeném intervalu $[0, T]$, s T takovým, že $P(T)\cdot Q(T) > 0$.

2. Na $[0, T]$ je dokázána asymptotická normalita: $\sqrt{n}\,(P_n(t)/P(t) - 1) \sim \sqrt{n}\,(H_n(t) - H(t)) \sim Z(t)$, což je gaussovský náhodný proces s nulovou střední hodnotou a kovariancí pro $0 \le s \le t \le T$

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

Hned je vidět možnost transformace na Brownův proces ($Z(t)$ je proces s nezávislými přírůstky) a tedy možnost zkonstruovat pro $P(t)$ (resp. $H(t)$) pásy spolehlivosti typu Kolmogorova-Smirnova (znovu, viz i Robust 84).

Obraťme nyní pozornost k odhadům hustoty a intenzity. Jak jsme řekli, při realizaci náhodného výběru přiřazujeme realizovaným bodům empirickou váhu ΔF_n (resp. ΔH_n), coby "primitivní" odhad okamžité hustoty či intenzity. A rozumný odhad dostaneme standardním způsobem – jádrovým

vyrovnáním těchto hodnot. Mějme jádro $w(u)$ - hustotu na $(-\infty, \infty)$, symetrickou, ≥ 0 (což není obecně nutné), navíc, řekněme, Parzen-Rosenblattova typu, tj. takové, že $\lim u \cdot w(u) = 0$ při $u \to \infty$. Pak definujeme jádrový odhad hustoty f resp. intenzity h jako

$$f_n(t) = \frac{1}{d} \sum_{i=1}^{n} w\left(\frac{t - Y_i}{d}\right) \Delta F_n(Y_i),$$

$$h_n(t) = \frac{1}{d} \sum_{i=1}^{n} w\left(\frac{t - Y_i}{d}\right) \Delta H_n(Y_i).$$

Zde d je zvolená "šíře okna", parametr vyrovnání.

Jsou i jiné možnosti odhadů. Samozřejmě, $P_n(t)$ i $H_n(t)$ jsou "schodovité" funkce, můžeme z nich vhodnou modifikací získat spojité odhady. Intenzitu můžeme odhadovat přímo z definice, jako $\hat{h}_n(t) = f_n(t)/P_n(t)$. Nebo, nechť $f^1(t)$ je hustota rozdělení pravděpodobnosti pro veličinu $(Y, \delta = 1)$, to jest pro dobu pozorovaných událostí. Označme $P^* = P \cdot Q$ funkci přežití pro distribuci n. v. $Y = \min(T, V)$. Oboje je přímo pozorovatelné, už bez cenzorování. A protože $h(t) = f(t) \cdot Q(t) / (P(t) \cdot Q(t)) = f^1(t) / P^*(t)$ na $[0, T]$, můžeme odhadnout $h(t)$ jako $\tilde{h}_n(t) = f_n^1(t) / P_n^*(t)$, kde $f_n^1(t)$ a $P_n^*(t)$ jsou standardní odhady hustoty f^1 respektive funkce přežití P^*.

Vlastnosti jádrových odhadů hustoty byly dost podrobně zkoumány už v Robustu 82 (J. Antoch). Vlastnosti odhadů intenzity jsou obdobné a ani cenzorování jim příliš neublíží. Citujme několik výsledků z článku Tanner, Wong (1983):

Mějme symetrické nezáporné jádro w s omezeným nosičem. Pak při $n \to \infty$, $d_n \to 0$, $n \cdot d_n \to \infty$, na $t \in [0, T]$ je

1. $\mathrm{E}\, h_n(t) \longrightarrow h(t)$,
2. $\mathrm{var}\, h_n(t) = \frac{1}{n d_n} \int_{-\infty}^{\infty} w^2(u)\, du \cdot \frac{h(t)}{P(t) Q(t)} + o(1/n d_n)$.
3. $\sqrt{n d_n}\,(h_n(t) - h(t))$ má asymptoticky normální rozdělení, centrované, s variancí z 2.

   V navazujícím článku uvažuje M. Tanner (AS 1983, No 3) jádrový odhad s proměnnou šířkou okna danou k nejbližšími sousedy (resp. k nejbližšími sousedy s $\delta_i = 1$, tj. k nejbližšími momenty pozorovaných událostí). Takový odhad má v podstatě tytéž vlastnosti jako běžný odhad jádrový, pokud hustota $f(t)$ je kladná. Pak je asymptoticky $k_n \sim 2n \cdot d_n f(t)$ v okolí bodu t, pokud d_n označuje poloměr intervalu kolem t, v němž je právě k_n dat.

4. Nechť $k_n = n^\alpha$ pro $\alpha \in (\frac{1}{2}, 1)$. Označme d_n příslušný parametr vyrovnání (neboli $d_n \sim n^{-\beta}$, $\beta \in (0, \frac{1}{2})$). Nechť jádro w má ohraničený nosič a je symetrické, nezáporné. Potom skoro jistě

$$h_n(t) \longrightarrow h(t).$$

Velká pozornost byla (a stále ještě je) věnována dílčím otázkám, jako je optimalizace parametru d (pomocí krosvalidace a různých penalizačních funkcí) a otázkám volby jádrových funkcí.

Všimněme si, že zatímco u odhadů kumulativních charakteristik asymptotické výsledky popisovaly chování náhodných procesů při t probíhajícím $[0, T]$, výsledky pro lokální charakteristiky se týkají jen odhadů v určitém bodě $t \in [0, T]$.

3 REGRESNÍ MODEL S PROPORCIONÁLNÍ INTENZITOU

Jak už jsme řekli, heterogenita zkoumaného souboru je v modelu pro intenzitu často vyjádřena součinem více členů. Pokud je ona nesourodost způsobena vlivem další veličiny (kovariáty) X, pak intenzitu rozdělení pravděpodobnosti n. v. T při $X = x$ modelujeme jako $h(t,x) = h_0(t) \cdot B(x)$. Tomuto multiplikativnímu modelu se také říká "proportional hazard regression model", $h_0(t)$ je tzv. "baseline" intenzita. Jako data budeme uvažovat realizace dvojic n. v. (T_i, X_i), $i = 1, \ldots, n$. Předpokládáme, že při daných hodnotách $X_i = x_i$ už jsou n. v. T_i nezávislé (tedy podmíněně). Pokud budeme mít co dělat s cenzorovanými daty (náhodně, zprava), cenzorující veličiny znovu modelujeme jako V_i závislé na X_i, ale při daných $X_i = x_i$ už nezávislé vzájemně i na T_i. Pak pozorujeme (Y_i, X_i, δ_i), $i = 1, \ldots, n$, kde opět $Y_i = \min(T_i, V_i)$ a $\delta_i = 1[T_i \leq V_i]$. Zopakujeme některé vlastnosti modelu, o kterém na Robustu 88 mluvila V. Lánská:

1. Pro dvě různé hodnoty kovariáty x_1, x_2 je $h(t,x_1) / h(t,x_2) = B(x_1) / B(x_2)$ pro každé t v definičním oboru. Podle této vlastnosti je model pojmenován. Pokud je tato vlastnost splněna, umožňuje metodu analýzy přežívání zvanou "accelerated testing". Například při testování spolehlivosti, pokud kovariáta X charakterizuje zátěž, můžeme v experimentu použít zátěž větší než je zátěž běžná ve skutečném provozu, a tím zkrátit čas potřebný pro doběhnutí většiny pokusů, či podstatně zmenšit míru cenzorování.
2. Mějme n. v. T_x, která při $X = x$ má distribuci s intenzitou $h_0(t) \cdot B(x)$. Nechť $\varphi(t)$ je ostře rostoucí transformace, $Z_x = \varphi(T_x)$. Potom rozdělení veličiny Z_x má intenzitu tvaru $h_1(t) \cdot B(x)$, tedy model s tímtéž "poměrem intenzit". To je i důvod, proč je odhadování funkce $B(x)$ možné oddělit od zkoumání funkce $h_0(t)$ a je možné vyjít z částečné věrohodnostní funkce (partial likelihood)

$$L_p = \prod_{i=1}^{n} \left[\frac{B(X_i)}{\sum_{j \in R_i} B(X_j)} \right]^{\delta_i},$$

kde opět R_i je "risk set", množina indexů objektů, které v okamžiku Y_i^- jsou rizikové. Částečná věrohodnost nezávisí přímo na hodnotách Y_i, ale jen na jejich pořadí, na rizikových množinách v momentech Y_i^-, takže se nezmění při rostoucí transformaci φ.

Vlastnostmi částečné věrohodnostní funkce a odhadů z ní pocházejících se zabýval Cox (1982). Byly připomenuty i v článku V. Lánské (1988), v souvislosti s analýzou nejznámější semiparametrické verze modelu, Coxovým modelem, v němž $B(X) = \exp \beta \cdot X$.

Nyní je chvíle, kdy bychom měli přejít k obecnějšímu pohledu na modely výskytu událostí v čase. Je to model tzv. counting procesů, který se velice často vyskytuje v posledním desetiletí v pracech zejména "skandinávské školy", viz Andersen, Borgan, Gill, Keiding, i E. Arjas, O. O. Aalen. Další zobecnění je v tom, že se připouští změna hodnot kovariát během sledované doby. Okamžitá intenzita pak závisí na okamžitém stavu kovariáty. Příkladů je mnoho – třeba doba trvání prosperity podniku. Riziková funkce se mění v závislosti na momentálním stavu nějakých hospodářských ukazatelů (pro nás kovariát), nestačí uvažovat stav jen v okamžiku vzniku podniku. Často tak uvažujeme dva paralelně běžící časy. Například jeden je sledovaná doba přežití, druhý je kalendářní čas, ve které se mění prostředí, které ovlivňuje dobu přežití.

Označíme $N_i(t)$, $i = 1, \ldots, n$ counting procesy, náhodné procesy nabývající hodnot $0, 1, \ldots$ a mající trajektorie po částech konstantní, se skoky $+1$. Dále uvažujme $X_i(t)$ – náhodné procesy kovariát,

$I_i(t)$ – nula-jedničkové procesy indikující zda i-tý counting proces je v čase t^- v rizikové množině, tj. vlastně je-li pozorován v momentě t^-. Procesy $I_i(t)$ pozorujeme plně v nějakém intervalu $t \in [0, T]$, procesy $X_i(t)$ jen když $I_i(t) = 1$. Counting proces $N_i(t)$ pak prostě v čase načítá *pozorované* události i-tého druhu (či události týkající se i-tého objektu). Připouštíme tedy i opakované události.

Aby bylo teorii učiněno zadost, je nutné předpokládat existenci neklesající posloupnosti σ-algeber $\sigma^n(t)$ ve výběrovém prostoru Ω^n, a to takové, že všechny uvažované náhodné procesy jsou vždy v $[0, t)$ "adaptovány" na tuto posloupnost. Neboli $\sigma^n(t)$ musí obsahovat všechny jevy, které mohou nastat v $[0, t)$, pro n-rozměrný model $N_1(t), \ldots, N_n(t)$. Předpokládejme, že vývoj procesu $N_i(t)$ začíná v $N_i(0) = 0$ a je popsán a určen spojitou intenzitou $h_i(t)$. Znamená to vlastně, že $\lim_{\Delta \to 0+} \frac{1}{\Delta} \Pr\{N_i(t+\Delta) - N_i(t^-) = 1 \mid \sigma^n(t)\} = h_i(t)$, či $h_i(t)\,dt = \Pr\{dN_i(t) = 1 \mid \sigma^n(t)\}$. Přitom podmíněnost $\sigma^n(t)$ interpretujeme tak, že intenzita v t může záviset na celé minulosti sledovaného procesu v $[0, t)$. To by asi bylo příliš široké pojetí, matematicky těžko vyjádřitelné. Proto většina užívaných modelů uvažuje intenzitu pro $N_i(t)$ závisící jen na "okamžitém" stavu, $h_i(t) = h(t, X_i(t)) \cdot I_i(t)$. Ale vhodně vybranými kovariátami můžeme do modelu dostat i důležité události z minulosti za tu cenu, že zvýšíme dimenzi $X_i(t)$.

Studium intenzitou popsaných counting procesů je úzce spjato s užitím teorie martingalů pro odvození vlastností odhadů a testů. Protože $\mathrm{E}\{dN_i(t) \mid \sigma^n(t)\} = h_i(t)\,dt$, tak náhodný proces definovaný jako $M_i(t) = N_i(t) - \int_0^t h_i(s)\,ds$ je martingal. Skoro všechny jeho trajektorie jsou po částech spojité a nerostoucí funkce, se skoky $+1$ přesně v bodech, kde jsou skoky counting procesu $N_i(t)$. Přitom $\mathrm{E}\, M_i(t) = 0$, $\mathrm{var}\, M_i(t) = \int_0^t h_i(s)\,ds$. Celý matematický rámec je přehledně podán v článku Andersen a Borgan (1985).

Pozorujeme naráz n probíhajících procesů, jedním důležitým předpokladem teorie je, že nenastanou současně skoky u dvou či více sledovaných procesů. Teoreticky z toho plyne, že $\mathrm{cov}\,(dM_i(t), dM_j(t)) = \mathrm{E}\,\{dN_i(t) \cdot dN_j(t) \mid \sigma_t^n\} = 0$ pro $i \neq j$. V praxi ovšem tento předpoklad splněn nebývá, ať už kvůli zaokrouhlování či vůbec z toho důvodu, že většina v čase probíhajících procesů se měří ve vybraných diskrétních okamžicích. Je třeba to vzít v úvahu a posoudit, jestli ještě naše data mají spojitý charakter, nejsou-li to již vlastně data tříděná (grouped) – pak i asymptotické výsledky jsou jen přibližné (viz Prentice, Gloeckler 1978).

Multiplikativní regresní model předpokládá, že counting proces $N_i(t)$ má intenzitu rozložení "skoků" (tj. pozorovaných událostí) $h_i(t) = h_0(t) \cdot B(X_i(t)) \cdot I_i(t)$, kde h_0 je opět společná spojitá "baseline" intenzita. Můžeme k ní definovat příslušnou kumulativní intenzitu $H_0(t) = \int_0^t h_0(s)\,ds$. Proces se zkoumá na nějakém intervalu $[0, T]$, pro jehož horní mez platí $H_0(T) < \infty$. Předpokládáme dále $B(\cdot)$ spojitou, nezápornou, procesy $X_i(t)$ skoro jistě ohraničené v $[0, T]$. Andersen et al. pracují výhradně s Coxovým modelem, kde $B(\mathbf{x}) = \exp \boldsymbol{\beta}'\mathbf{x}$. Základem pro odhad $\boldsymbol{\beta}$ je částečná věrohodnostní funkce, její logaritmus má nyní tvar

$$\ell = \sum_{i=1}^{n} \int_0^T \log \frac{\exp(\boldsymbol{\beta}\, X_i(t))}{\sum_{j=1}^{n} \exp(\boldsymbol{\beta}\, X_j(t)) \cdot I_j(t)}\, dN_i(t).$$

Je to jen obecnější výraz než jsme zvyklí z klasické analýzy přežívání. K modelu s daty (Y_i, X_i, δ_i), popisovanému např. v práci V. Lánské (1988), se dostaneme, položíme-li $I_i(t) = 1$ v $[0, Y_i]$, $dN_i(t) = 1$ právě v $t = Y_i$ při $\delta_i = 1$.

Odhad parametru $\boldsymbol{\beta}$ dostaneme řešením rovnic $\partial \ell / \partial \beta_\ell = 0$, $\ell = 1, \ldots, K$ (pro K-rozměrné $\boldsymbol{\beta}$ a $X(t)$). Při praktickém řešení si musíme pomoci iterací, Newtonovým–Raphsonovým algoritmem, který používá i matici druhých derivací $S(\boldsymbol{\beta}) = \{\partial^2 \ell / \partial \beta_k\, \partial \beta_\ell\}$.

Hned vidíme, že při proměnných kovariátách $X'(t) = (X_1(t), \ldots, X_K(t))$ potřebujeme k řešení věrohodnostních rovnic znát hodnoty $X_j(T_i)$, $j = 1, \ldots, K$, $i = 1, \ldots, n$, pokud $I_j(T_i) = 1$, kde T_i jsou okamžiky pozorovaných událostí, tj. body, v nichž $dN_i(T_i) = 1$. Takový je tedy nárůst potřebné informace. Ještě daleko větší nároky na paměť bude mít procedura pro neparametrické odhadování funkce $B(x)$ - tam je pak vhodné si pomoci interpolací, což zase prodlužuje čas potřebný k výpočtům.

Máme-li již odhad funkce B (resp. odhad parametru β v Coxově modelu), můžeme odhadnout i kumulativní základní intenzitu:

$$\hat{H}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} I_j(s)\, \hat{B}(X_j(s))}.$$

Vyrovnáním jejich přírůstků pomocí jádra w pak dostaneme

$$\hat{h}_0(t) = \frac{1}{d} \sum_{i=1}^{n} w\left(\frac{T_i - t}{d}\right) \Delta \hat{H}_0(T_i).$$

Celou teorií odhadů pro Coxův model se zabývají Andersen a Gill (1982). Uvádějí podmínky, které zajišťují asymptotickou stabilitu částečné věrohodnosti i derivací jejího logaritmu, dále podmínky pro asymptotickou regularitu (konkavitu logaritmu věrohodnostní funkce), i variantu Lindebergovy podmínky pro zajištění asymptotické normality odhadů. Implicitně je v těchto podmínkách skryt i požadavek, aby proces byl vůbec dostatečně pozorován, tj. aby $\sum_{i=1}^{n} I_i(t) \to \infty$ v skoro každém $t \in [0, T)$, kde $h_0(t) > 0$. Shrneme teď výsledky z práce Andersen a Gill:

1. Odhad $\hat{\beta}$ (řešení rovnic $\partial \ell / \partial \beta_k = 0$) je konzistentní v pravděpodobnosti. Příslušná pravděpodobnostní míra je limitou $P^{(n)}$, kde $P^{(n)}$ je pravděpodobnost na Ω^n, $\sigma^n(\mathcal{T})$.
2. $\sqrt{n}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \Sigma)$
3. P-konzistentní odhad matice Σ se získá z matice druhých derivací:

$$\hat{\Sigma}_n = n \cdot \mathcal{S}^{-1}(\hat{\beta}).$$

4. $\sqrt{n}(\hat{\beta} - \beta)$ a náhodný proces $W_n(t) = \sqrt{n}(\hat{H}_0(t) - H_0(t)) + \sqrt{n}(\hat{\beta} - \beta)' \cdot \int_0^t e(\beta, s)\, h_0(s)\, ds$ (kde e je jistá K-rozměrná funkce z podmínek stability a regularity) jsou asymptoticky nezávislé. Proces $W_n(t)$ je asymptoticky rozdělen jako Gaussovský martingal, s variancí $\int_0^t \frac{h_0(u)}{s^0(u)}\, du$ (zde $s^0(t)$ je opět funkce z předpokladů, P limita výrazu $\frac{1}{n} \sum_{i=1}^{n} B(X_i(t))\, I_i(t)$).

Jako důsledek 4. dostaneme následující tvrzení: Známe-li funkci B (resp. parametr β), je $\sqrt{n}(\hat{H}_0(t) - H_0(t))$ asymptoticky rozloženo jako Gaussovský martingal, s variancí P-konzistentně odhadnutelnou:

$$\sup_{t \le T} \left| \text{as var } W_n(t) - n \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\left[\sum_{j=1}^{n} B(X_j(s))\, I_j(s)\right]^2} \right| \xrightarrow{P} 0.$$

4 NEPARAMETRICKÝ MODEL, ADITIVNÍ REGRESNÍ FUNKCE

Coxův model je velice často užívaný v mnohých oblastech uplatnění analýzy přežívání. Jeho parametrizovaná část nemusí být nutně lineární ($\beta' X$), pro jinou parametrizovanou funkci v exponentu bychom došli k podobným výsledkům odhadování (viz také Prentice, Self, AS 1983, No 1). Ale

každá parametrizovaná funkce je nutně omezující. Proto je dobré se snažit odhadnout celý model jako neparametrický, pak teprve na základě odhadnutého tvaru funkcí zkusit pro ně zvolit vhodnou parametrizovanou formu.

Vraťme se nejprve ke standardnímu modelu pro intenzitu $h(t,x)$ doby do poruchy, závisící na kovariátě, která se nemění v čase. V Robustu 88 byla popsána metoda, jak odhadnout kumulativní intenzitu $H(t,x) = \int_0^t h(s,x)\,ds$, jako funkci t pro pevné x. Odhad je prostě počítán jako Nelsonův–Aalenův odhad z výsledků majících hodnotu kovariáty v nějakém zvoleném okolí x. Opět, pokud šířku okolí d zmenšujeme k 0, počet dat $n \to \infty$ a $d \cdot n \to \infty$, pak lze ukázat, že získaný odhad je konzistentní a asymptoticky normální. Je možné pokračovat dál a pro model s proporcionální intenzitou, tj. když $H(t,x) = H_0(t) \cdot B(x)$, konzistentně odhadnout vhodně normované funkce H_0, B – viz stále Volf, 1988 a,b. McKeague a Utikal (1991) vyvinuli na základě tohoto vlastně jádrového odhadu sérii testů pro rozlišení mezi modelem s proporcionální intenzitou a jinými formami modelu intenzity.

Jenže jakmile je počet kovariát (tj. dimenze x) větší, stačí už $K > 2$, málokdy máme dat tolik a tak "hustá", aby se dal efektivně použít globální (v R_K) jádrový odhad. To je důvod, proč se zavádí aditivní regresní funkce. Předpokládáme, že vliv kovariát $x_1, \ldots, x_K$ na regresní model je zprostředkován nějakou regresní funkcí $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$ – třeba až po nějaké vhodné transformaci modelu či kovariát. Například v lineárním regresním modelu popisujícím závislost nějaké veličiny Y na kovariátě x, regresní funkce je $\mathrm{E}[Y/\mathbf{X} = \mathbf{x}]$. V rámci modelů pro analýzu přežívání budeme předpokládat, že intenzita je proporcionální, tvaru

$$h(t,\mathbf{x}) = h_0(t) \cdot \exp(b(\mathbf{x}))$$

a budeme $b(\mathbf{x})$ odhadovat jako aditivní funkci kovariát. Ale nejprve se podíváme na problém neparametrického odhadování regresních funkcí obecně.

Metoda maxima lokální věrohodnosti. Metoda byla popsána a modifikována na "local scoring" v článku Hastie, Tibshirani (1986). Spočívá v následujícím přístupu k odhadování komponent $b_k(x_k)$ aditivní regresní funkce $b(\mathbf{x}) = \sum_{k=1}^{K} b_k(x_k)$:

Chceme-li odhadnout hodnotu funkce b_k v bodě $x_k = z$, považujeme funkci b_k za konstantní (b_z) v nějakém zvoleném okolí $\mathcal{O}_z$. Předpokládejme, že můžeme zkonstruovat log-likelihood ℓ_n na základě náhodného výběru $\{Y_i, X_i,\ i = 1, \ldots, n\}$. S b_z nyní zacházíme jako s parametrem a řešíme rovnici $\partial \ell_n / \partial b_z = 0$. Je to tedy opět jakýsi "jádrový" přístup, tentokrát ke každé složce zvlášť. V mnoha běžných modelech má logaritmus věrohodnostní funkce tvar

$$\ell_n = \sum_{i=1}^{n} \ell_1(Y_i, b(\mathbf{X}_i)), \tag{1}$$

takže

$$\frac{\partial \ell_n}{\partial b_z} = \sum_{i=1}^{n} \mathbf{1}[X_{ki} \in \mathcal{O}_z] \cdot \ell_1' \left(Y_i, b_z + \sum_{j \neq k} b_j(X_{ji}) \right).$$

Je vidět, že lokální věrohodnostní rovnici můžeme řešit (pro b_z), pokud máme k dispozici odhady ostatních komponent b_j, $j \neq k$, z nějakého předchozího kroku odhadování. Takže tento přístup vede k iterativní proceduře, startující z nějakého počátečního odhadu všech funkcí b_j, $j = 1, \ldots, k$, třeba $b_j^{(0)} \equiv 0$, nebo z $b_j^{(0)}(x_j) = \hat{\alpha}_j + \hat{\beta}_j x_j$, kde $\hat{\alpha}_j$, $\hat{\beta}_j$ jsou MVO pro lineární regresní funkce. Věrohodnostní

rovnice bývají zpravidla řešeny také iterativně, Newtonovou–Raphsonovou procedurou. Kombinací těchto dvou iterací dospějeme k následujícímu kroku od s-tého k $(s+1)$-tému přiblížení

$$b_z^{(s+1)} = b_z^{(s)} - \left[\frac{\partial \ell_n}{\partial b_z} \Big/ \frac{\partial^2 \ell_n}{\partial b_z^2} \right] \Bigg|_{b=b^{(s)}} . \tag{2}$$

Tak výpočet projde hodnoty $z = x_k$, odhadne tím funkci $b_k^{(s+1)}$, a začne odhadovat funkci $b_{k+1}^{(s+1)}$. Hastie a Tibshirani (1986) ve své modifikaci, kterou nazvali "local scoring", jádrově vyrovnávali v každém kroku i derivace ℓ_n.

Stone (1986) se zabýval odhadováním regresních funkcí, ale nikoli zcela neparametricky. Uvažoval pro funkce b_k aproximace pomocí polynomiálních splinů. Vlastně je tedy reparametrizoval, a parametry těchto splinů odhadoval běžnou, globální metodou maximální věrohodnosti. Došel však k zajímavým závěrům, pokud jde o aditivní regresní funkce.

Předpokládal, že data odpovídají modelu exponenciálního typu. To znamená, že likelihood je typu (1), s $\ell_1(Y,\theta) = Y \cdot c(\theta) + d(\theta)$, kde c, d jsou známé funkce a $\theta = \theta(\mathbf{x})$ je regresní funkce popisující závislost Y na $\mathbf{X}$. Stone formuloval vhodné podmínky a dokázal, že platí:

1. Mezi aditivními regresními funkcemi $\sum_{j=1}^{K} b_j(x_j) + b_0$ existuje taková, která je nejblíž k $\theta(\mathbf{x})$ vzhledem k příslušné Kullbackově–Leiblerově vzdálenosti (dané věrohodností ℓ_1).
2. Tato aditivní funkce z 1. je konzistentně odhadnutelná polynomiálními spliny.

Aby složky b_j byly jednoznačné (s. j.), předpokládal Stone, že $\mathrm{E}\, b_j(X_j) = 0$, $j = 1, \ldots, K$.

Konzistence může být zajištěna jenom tak, že s růstem rozsahu dat (n) roste i počet parametrů. V případě splinů to znamená, že zůstává zvolený stupeň polynomů, ale zvětšuje se počet uzlových bodů, ve kterých jeden polynom přechází v druhý.

Metoda lokální věrohodnosti vychází vlastně z téže aproximace jako Stone. Místo polynomů v pevných oknech (triviálním splinem je histogram) uvažuje okna posouvající se. Podstatný rozdíl může být ve způsobu výpočtu. Standardní věrohodnostní odhad řeší naráz soustavu rovnic, zatímco lokální přístup řeší postupně rovnice pro jednu neznámou (a dokonce provede vždy jen jeden krok iterativního řešení, a přejde k rovnici pro další "okno" -okolí, viz (2)). Pokud je kovariáta jednorozměrná a je-li likelihood typu (1), ukáže se, že matice druhých derivací (při aproximaci funkce b histogramem) je diagonální. Fakticky tak řešíme rovnice "globální" věrohodnosti stejně jako při použití metody lokální věrohodnosti. Jakmile ale pracujeme s aditivní regresní funkcí vícerozměrné kovariáty, je konzistence dokázána jen pro klasický Gaussovský regresní model, kde $b(\mathbf{x}) = \mathrm{E}[Y/\mathbf{x}]$. Pro tento speciální případ je iterativní procedura známa pod jménem "backfitting" algoritmus (Friedman, Stuetzle, 1981), konzistence odhadu a existence nejlepší aditivní aproximace pro $\mathrm{E}(Y/\mathbf{x})$ je také dokázána v rámci obecnějšího ACE algoritmu (Alternating Conditional Expectations, Breiman, Friedman 1985), který řeší otázku nejlepších aditivních aproximací i pro vícerozměrnou závislou proměnnou $\mathbf{Y}$. V Gaussovském modelu se totiž shoduje MVO regresní funkce s pravoúhlou projekcí.

V obecnějších případech si však konzistencí výsledků metody maxima lokální věrohodnosti nemůžeme být jisti (i když zkušenosti jsou povzbuzující). Koneckonců, neparametrické odhadování můžeme považovat za první krok analýzy, kterou pak prohloubíme uvažováním vhodného modelu parametrizovaného.

Věnujme se nyní opět regresnímu modelu s proporcionální intenzitou v analýze přežívání. Protože částečná věrohodnostní funkce není typu (1), nemůžeme použít pro tento model závěry ze Stonea (1986). Popíšeme řešení úlohy odhadu aditivní regresní funkce metodou lokální věrohodnosti. Pro

standardní model s konstantními hodnotami kovariát bylo řešení navrženo i předvedeno na příkladě v Volf (1990). Zde budeme pracovat již s modelem pro counting procesy a s regresí na procesech kovariát. Model byl popsán v předešlé části článku, připomeňme, že vývoj counting procesu $N_i(t)$ je řízen intenzitou $h_i(t) = h_0(t) \cdot \exp b(\mathbf{X}_i(t)) \cdot I_i(t)$. Odhadování vychází z logaritmu částečné věrohodnostní funkce

$$\ell_n = \sum_{i=1}^{n} \int_0^T \frac{\exp b(\mathbf{X}_i(t))}{\sum_{j=1}^{n} \exp b(\mathbf{X}_j(t)) \cdot I_j(t)} \, dN_i(t). \tag{3}$$

Budeme se snažit odhadnout složky aditivní regresní funkce $b(\mathbf{x}) = \sum_{k=1}^{K} b_k(x_k)$. Zvolme si nějaké z z oboru hodnot, řekněme, $X_\ell(t)$, $t \in [0, T]$, a považujme $b_\ell(x_\ell)$ za konstantu $b_\ell(z)$ v nějakém okolí $\mathcal{O}_z$. Potom

$$\frac{\partial \ell_n}{\partial b_\ell(z)} = \sum_i \int_0^T \left\{ 1[X_{\ell i}(t) \in \mathcal{O}_z] - \exp b_\ell(z) \cdot \frac{R_\ell(z, b, t)}{S(b, t)} \right\} dN_i(t),$$

kde $R_\ell(z, b, t) = \sum_{j=1}^{n} 1[X_{\ell j}(t) \in \mathcal{O}_z] \exp\left\{ \sum_{k \neq \ell} b_k(X_{kj}(t)) \right\} \cdot I_j(t)$ a $S(b, t) = \sum_{j=1}^{n} \exp\{b(\mathbf{X}_j(t)) \cdot I_j(t)\}$. Přímo z požadavku na řešení rovnice $\partial \ell_n / \partial b_\ell(z) = 0$ se nabízí následující iterační krok:

$$b_\ell^{(s+1)}(z) = -\log \left[\sum_{i=1}^{n} \int_0^T \frac{R_\ell(z, b^{(s)}, t)}{S(b^{(s)}, t)} \, dN_i(t) \Big/ \sum_{i=1}^{n} \int_0^T 1[X_{\ell i}(t) \in \mathcal{O}_z] \, dN_i(t) \right]. \tag{4}$$

Potřebujeme tedy v každém takovém kroku už mít s-té odhady ($b^{(s)}$) všech komponent regresní funkce, a to alespoň v bodech $X_{kj}(T_i)$, pokud $I_j(T_i) = 1$ – tj. ve všech pozorovaných hodnotách kovariát. Přitom $i, j = 1, \ldots, n$, $k = 1, \ldots, K$, čili množství uložených dat je nejméně $n \times n \times K$, pokud si nepomůžeme nějakou interpolací (a pokud skutečně všechny kovariáty jsou proměnné).

Metoda odhadu je zformulována (alternativní postup by mohl být založen přímo na (2), s pomocí ještě druhých derivací ℓ_n). Její možnosti nejlépe předvedeme na příkladě.

PŘÍKLAD. Řekli jsme, že se sledováním dob (do nějaké událostí) se setkáváme v mnoha oblastech, perspektivní pro statistickou analýzu jsou nyní zejména demografie, ekologie, sociální vědy. Následující příklad je jednoduchý, vymyšlený, ale snad naznačí možnosti použití popisované analýzy výskytu událostí v čase.

Představme si, že v nějakém podniku byl po nějakou dobu sledován vývoj zaměstnanosti, mimo jiné i odchody zaměstnanců. Získala se data

$$\{T_i, \delta_i, X_{1i}, \ldots, X_{4i}, \; i = 1, \ldots, n\}.$$

Doba T probíhá posledních 10 let, je udávána v měsících od 0 do 120. Moment T_i je buď doba skončení pracovního poměru i-tého pracovníka, nebo moment cenzorování, při $\delta_i = 0$ – to se většinou týká pracovníků zůstávajících v podniku po době ukončení sledování, pak je $T_i = 120$. Uvažujeme dva typy odchodů, a to dobrovolný ($\delta_i = 2$), zde je zahrnut i odchod do penze, a propuštění ($\delta_i = 1$). Naměřené hodnoty kovariát mají následující význam: X_{1i} je věk pracovníka v momentě T_i, X_{2i} je délka zaměstnání v podniku do okamžiku T_i. Obě jsou udány v rocích. X_3 charakterizuje kategorii profese: 1 – vědecký pracovník, 2 – odborný pracovník, 3 – administrativa, 4 – technici a kvalifikovaní dělníci, 5 – ostatní, $X_4 = 1$ pro muže, 2 pro ženy. Veličina δ tedy kromě cenzorování indikuje i konkurující si rizika. X_1 a X_2 musíme uvažovat jako veličiny měnící se s časem, například $X_{1i}(t) = \max\{0, X_{1i} - (T_i - t)/12\}$. (tj. $X_{1i}(T_i) = X_{1i}$), stejně pro $X_{2i}(t)$. Indikátor rizikové množiny $I_i(t) = 1$ pro $t \in [0, 120]$

takové, že i-tý pracovník byl zaměstnán v podniku v době t. Neboli pro $t \in [\max\{0, T_i - 12\,X_{2i}\}, T_i]$, jinak $I_i(t) = 0$.

Data jsou připravena pro neparametrický odhad složek aditivní regresní funkce $b(\mathbf{x}) = \sum_{k=1}^{4} b_k(x_k)$ pomocí iterativního schématu (4). Pak odhadneme kumulativní intenzitu $H_0(t)$ jako

$$\hat{H}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{S(\hat{b}, s)},$$

případně jádrovým vyrovnáním z ní dostaneme odhad pro $h_0(t)$. Proceduru odstartujeme z $b_1^{(0)}, \ldots, b_4^{(0)} \equiv 0$. Po každém dokončeném s-tém kroku proložíme odhadnutými body $\{x_{ki}, b_k^{(s)}(x_{ki}),\ i = 1, \ldots, n\}$ optimální přímky $\alpha_k^{(s)} + \beta_k^{(s)} x_k$ (metodou nejmenších čtverců). Vývoj parametrů těchto přímek bude pro nás indikací postupu iterování a jeho konvergence.

V našem případě už pro $s = 6$ byly rozdíly mezi $\beta_k^{(6)}$ a $\beta_k^{(5)}$ menší než 10^{-3}, výpočet jsme ukončili. Výsledky této konečné lineární aproximace funkcí b_k jsou shrnuty v Tabulce 1. Jsou tam i odhady korelace x_{ki} a $\hat{b}_k(x_{ki})$ a variance reziduí $\hat{b}_k(x_{ki})$ od přímky $\hat{\alpha}_k + \hat{\beta}_k x_k$. Celou analýzu jsme provedli zvlášť pro $\delta = 1$ a pro $\delta = 2$.

Obrázek 1 ukazuje výsledné odhady funkcí b_1, b_2, b_3. Protože čtvrtá kovariáta měla jen 2 hodnoty, je její vliv zcela popsatelný přímkou, jejíž odhad je v Tabulce 1. Na Obrázku 2 je zobrazen odhad "baseline" intenzity $h_0(t)$.

5 MOŽNOSTI TESTOVÁNÍ TVARU MODELU, ALTERNATIVNÍ MODEL AALENŮV

Když už provedeme analýzu v rámci určitého modelu, měli bychom se také přesvědčit, že jsme model nevybrali špatně. Pokud jde o model pro intenzitu, už jsme se zde zmínili o článku McKeague, Utikal (1991). V něm je navržena testovací procedura vycházející z asymptotické normality odhadu kumulativní intenzity $H(t, x)$ pro pevné x, tj. jen z dat, která mají hodnoty kovariát ve zvoleném okolí x. Důkaz asymptotických vlastností je naznačen také v Robustu'88. Metoda se dá použít i pro případ s časově proměnnými kovariátami. Pak sledujeme vždy jen counting procesy probíhající za podmínky $X(t) \in \mathcal{O}_x$. Pokud je model n counting procesů charakterizován jejich intenzitami $h_i(t) = h(t, X_i(t)) \cdot I_i(t)$, pak ony podmíněné counting procesy jsou popsány přibližně intenzitami $h(t, x) \cdot I_i(t) \cdot 1[X_i(t) \in \mathcal{O}_x]$. Nelsonovým–Aalenovým odhadem jsme schopni odhadnout $H(t, x) = \int_0^t h(s, x)\, ds$.

Pokud jde konkrétně o model s proporcionálními intenzitami, pro jeho testování existuje řada procedur, numerických i grafických (viz také V. Lánská, 1988). Většinou jsou založeny na té vlastnosti, která dala modelu jméno. Konkrétně, jsou-li z_1 a z_2 dvě úrovně hodnot kovariáty, pak

$$\log h(t, z_1) - \log h(t, z_2) = b(z_1) - b(z_2)$$

pro každé $t \in (0, \mathcal{T}]$ takové, že $h_0(t) > 0$. Jako hladinu kovariáty z uvažujeme vždy nějakou strátu (okolí) kolem hodnoty z.

Další otázkou, která nás zajímá, je to, zda závislost na kovariátách je významná, zda není zanedbatelná.

Pro doplnění i kontrolu předchozího příkladu jsme provedli i analýzu za předpokladu, že intenzita splňuje Coxův model, neboli $b_k(x) = \beta_k \cdot x$. Procedura výpočtu odhadů je popsána v části 3 tohoto

článku, vychází z výsledků Andersena a Gilla (1982). Využijeme-li asymptotické normality odhadů β_k a konzistence odhadu asymptotické kovarianční funkce, můžeme zkonstruovat test hypotézy $\beta_k = 0$ na základě statistiky (označme ji G_k), která má asymptoticky standardní normální rozdělení. Tabulka 2 přináší výsledky této analýzy. Vidíme, že už na (asymptotické) hladině 0.1 zamítneme hypotézu "nevýznamnosti" regrese jen pro složky 2, 4 (pro $\delta = 1$) a pro složky 2, 3 (pro $\delta = 2$), neboť gaussovský kvantil $q(0.05) \doteq -1.645$. Pro test významnosti regrese na více komponentách naráz se lehce odvodí chi-kvadrát kriterium.

S určitou tolerancí můžeme přijmout závěry předvedeného testu i v případě, že Coxův model je vzdálen od skutečnosti.

Dostáváme se k třetí, "nejjemnější" úrovni testování, které by mělo rozhodnout, zda regresní funkce je (například) lineární. Jedna možnost je ta, že odhadneme regresní funkcí jako polynom vyššího řádu, a pak testujeme hypotézu o tom, že koeficienty nelineárních členů jsou nulové.

Druhá cesta může vést přes pokus lineárně aproximovat neparametrický odhad regresní funkce. Tabulka 1 obsahuje výsledky takové aproximace. Jenže výsledky (tj. odhad korelace i variance) příliš závisí na použité vyrovnávací proceduře (na šířce okna) při neparametrickém odhadování.

Aalenův aditivní model pro intenzitu

Jde o alternativní model pro intenzitu counting procesu $N_i(t)$, uvedl ho Aalen (1980). Intenzita je modelována jako

$$h_i(t) = \left\{ \beta_0(t) + \sum_{j=1}^{K} \beta_j(t)\, X_{ji}(t) \right\} \cdot I_i(t).$$

Aditivní je přímo tvar intenzity, zatímco v předešlé části jsme uvažovali aditivní regresní funkci (v multiplikativním modelu pro intenzitu). Vyskytují se zde "parametry" $\beta_j(t)$ coby funkce času na nějakém intervalu $[0, T]$, X_{ji} jsou procesy kovariát, $I_i(t)$ indikátory rizika, $i = 1, \ldots, n$ jsou indexy sledovaných objektů či sledovaných typů událostí. Model musí být volen tak, aby intenzita byla nezáporná.

Přirozený vznik multiplikativního modelu pro určité pojetí vlivu kovariát jsme se snažili vysvětlit hned v první části práce. Aalenův aditivní model odpovídá zase trochu jiné situaci. Představme si například, že kovariáty jsou jen nula-jedničkové veličiny, tj. indikace, že nějaký vliv působí či nepůsobí. Aalenův model pak představuje situaci, kdy se "zapínají" další zdroje rizika, kromě nějakého základního popsaného intenzitou $\beta_0(t)$, kdežto v multiplikativním modelu $h_0(t) \cdot \exp \sum \beta_j X_{ji}(t)$ přepnutí kovariáty z 0 na 1 způsobí "zesílení" rizika popsaného základní intenzitou $h_0(t)$.

Stojíme nyní před úkolem odhadnout neznámé funkce $\beta_j(t)$, $j = 0, 1, \ldots, K$. Odhadují se jejich kumulativní verze, tj. $B_j(t) = \int_0^t \beta_j(s)\, ds$. Označme $Z_{ji}(t) = X_{ji}(t) \cdot I_i(t)$, $Z_{0i}(t) = I_i(t)$, matice $\mathbf{Z}(t)$ složená z těchto prvků má rozměr $(K+1) \times n$. Dále označíme $J(t) = \lim_{\Delta \to 0+} 1[\text{hodn}\, \mathbf{Z}(t-\Delta) = K+1]$, což je jakýsi indikátor regularity úlohy odhadu. Pokud si nyní vzpomeneme na vztah mezi přírůstky counting procesu a příslušného martingalu, $d\,M_i(t) = d\,N_i(t) - h_i(t)\, dt$, tak nyní je tento vztah možné napsat

$$d\,\mathbf{M}(t) = d\,\mathbf{N}(t) - \mathbf{Z}'(t) \cdot \beta(t)\, dt$$

kde $\mathbf{M}$, $\mathbf{N}$, β jsou vektory, $\mathbf{Z}$ je matice. Z tohoto vztahu vlastně metodou nejmenších čtverců dosta-

neme výraz pro odhad

$$\begin{aligned}\hat{\mathbf{B}}(t) &= \int_0^t J(s)\,[\mathbf{Z}(s)\,\mathbf{Z}'(s)]^{-1}\,\mathbf{Z}(s)\,d\mathbf{N}(s) \\ &= \sum_{T_i \le t} J(T_i)\cdot \overline{\mathbf{Z}}(T_i)\,d\mathbf{N}(T_i),\end{aligned}$$

kde T_i jsou pozorované momenty událostí, $\overline{\mathbf{Z}} = (\mathbf{Z}\,\mathbf{Z}')^{-1}\,\mathbf{Z}$ je zobecněná inverzní matice k $\mathbf{Z}'$, $d\mathbf{N}(T_i)'$ je $(0,0,\ldots,1,\ldots,0)$, přičemž "1" je na i-tém místě.

I pro tyto odhady je vyvinuta asymptotická teorie (Aalen 1980, monografie Andersen et al. 1992). Pokud označíme $\mathbf{B}^*(t) = \int_0^t J(s)\,\beta(s)\,ds$, tak zjistíme, že $\sqrt{n}\,\left(\hat{\mathbf{B}}(t) - \mathbf{B}^*(t)\right) = \sqrt{n}\,\int_0^t J(s)\,\overline{\mathbf{Z}}(s)\,d\mathbf{M}(s)$ je lokální martingal s kovariančním procesem $n\,\int_0^t J(s)\,\overline{\mathbf{Z}}(s)\,Q(s)\,\overline{\mathbf{Z}}'(s)\,ds$, kde $Q(s)$ je diagonální matice $\mathrm{diag}\,(h_1(s),\ldots,h_n(s))$, protože, jak jsme viděli, $\mathrm{var}\,dM_i(t) = h_i(t)\,dt$. Z toho již se dají odvodit asymptotické vlastnosti. Tento model je už svou podstatou neparametrický, ale kromě výpočtu inverzních matic není jeho numerická analýza nijak obtížná.

Literatura

Aalen O. O. (1980). A model for nonparametric regression analysis of counting processes. Springer Lect. Notes in Statistics 2, 1–25.

Andersen P. K., Borgan O. (1985). Counting process model for life history data: a review. Scand. J. Statist. 12, 97–158.

Andersen P. J., Borgan O., Gill R., Keiding N. (1992). Statistical Models Based on Counting Processes. New York, Springer.

Andersen P. K., Gill R. (1982). Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100–1120.

Arjas E. (1989). Survival models and martingale dynamics. Scand. J. Statist. 16, 177–225.

Antoch J. (1982). Odhady hustoty. Robust 82.

Barlow R. E., Proschan F. (1967). Mathematical Theory of Reliability. New York, Wiley.

Breiman L., Friedman J. H. (1985). Estimating optimal transformations for multiple regression and correlation. J.A.S.A. 80, 580–597.

Breslow N., Crowley J. (1974). A large sample study of the life table and product limit estimates under random censorship. Ann. Statist. 2, 437–453.

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

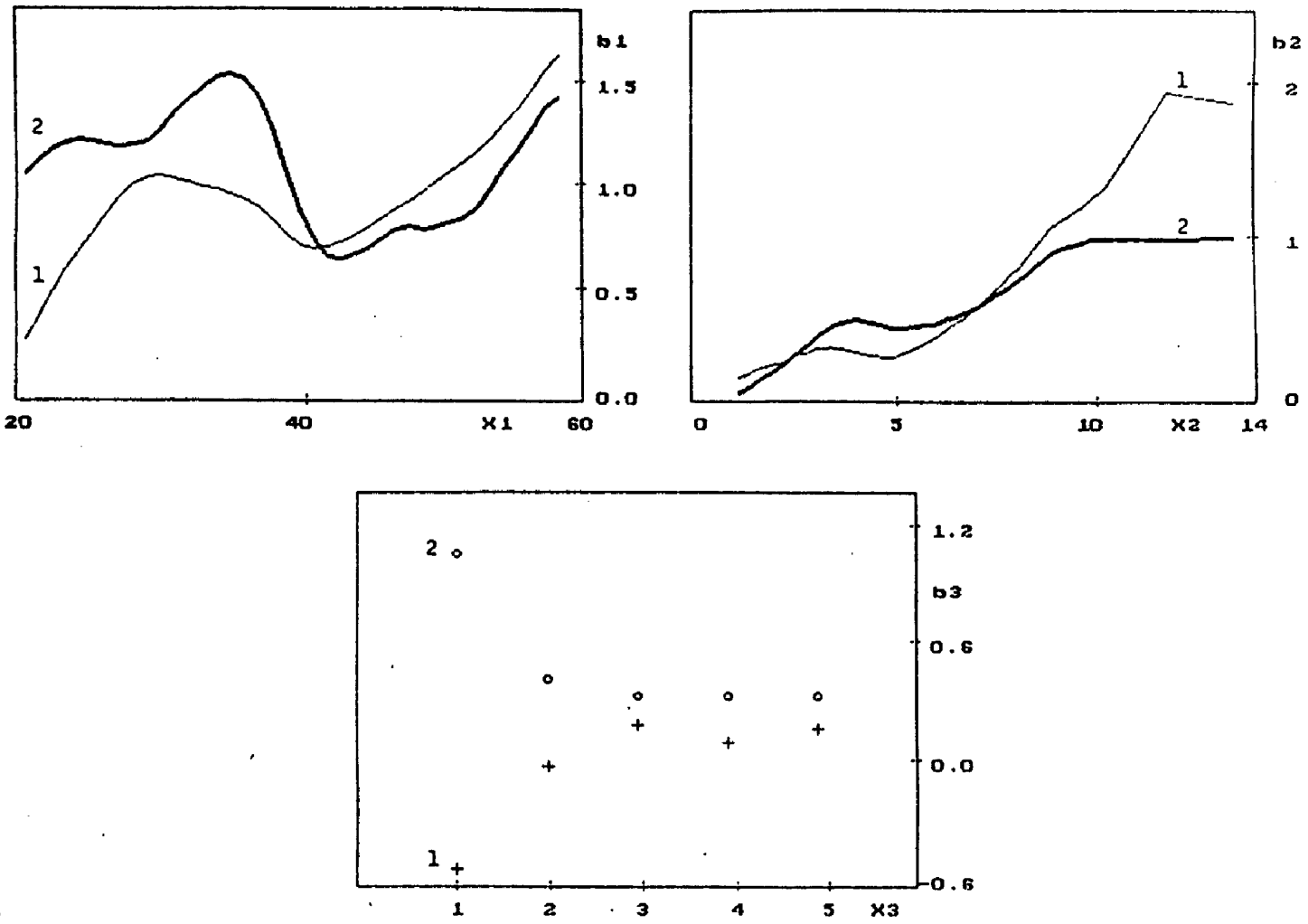
Cox R. D. (1975). Partial likelihood. Biometrika 62, 269–276.

Dabrowska D. A., Doksum K. A. (1987). Estimates and confidence intervals for median and mean life in the proportional hazard model. Biometrika 74, 799–807.

Fleming T. R., Harrington D. P. (1991). Counting Processes and Survival Analysis. New York, Wiley.

Friedman J., Stuetzle W. (1981). Projection pursuit regression. J.A.S.A. 76, 817–823.

Hastie T. J., Tibshirani R. J. (1986). Generalized additive models (with discussion). Statist. Sci. 1, 297–310.


b1
1.5
1.0
0.5
0.0
2
1
20
40
X1
60
b2
2
1
2
1
0
0
5
10
X2
14
b3
1.2
0.6
0.0
-0.6
2
1
1
2
3
4
5
X3

Obr. 1.

	k	$\hat{\alpha}$	$\hat{\beta}$	corr	var
$\delta = 1$:	1	0.3771	0.0137	0.5001	0.0549
	2	-0.1285	0.1218	0.8412	0.0435
	3	-0.5346	0.2004	0.7753	0.0251
	4	-1.0673	0.6258	1.0	0.0
$\delta = 2$:	1	1.6455	-0.0144	-0.3897	0.1129
	2	-0.2972	0.1131	0.8343	0.0396
	3	0.9571	-0.1991	-0.6960	0.0398
	4	0.6551	-0.4478	-1.0	0.0

Tab. 1.

Obr. 2.

	$\delta = 1$		$\delta = 2$	
k	$\tilde{\beta}_k$	G_k	$\tilde{\beta}_k$	G_k
1	0.0075	0.5726	-0.0259	-1.2259
2	0.0712	1.6528	0.1428	2.0139
3	0.1124	0.8628	-0.4773	-1.8863
4	0.7474	2.8004	-0.2057	-0.4932

Tab. 2.

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

Lánská V. (1988). Analýza přežívání, Coxův model. Robust 88.

McKeague I. W., Utikal K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards model. Scand. J. Statist. 18, 177–195.

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

Rejtõ L. (1982). On the fixed censoring model and consequences for the stochastic case. In: Transactions of 9th Prague Conf. on Information Theory,.... Academia, Prague.

Sleeper J. A., Harrington D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. J. A. S. A. 85, 941–949.

Stone C. J. (1986). The dimensionality reduction principle for generalized additive models. Ann. Statist. 14, 590–606.

Tanner M., Wong W. (1983). The estimation of the hazard function from randomly censored data by the kernel method. Ann. Statist. 11, 989–993.

Tsiatis A. A. (1981). A large sample study of Cox's regression model. Ann. Statist. 9, 93–108.

Volf P.: Robust 84, Robust 88, 4th Prague Symposium on Asymptotic Statistics 1988, 11th Prague Conference on Information Theory 1990.