ODHADOVÁNÍ A TESTOVÁNÍ V REGRESNÍCH MODELECH PRO ŽIVOTNOST

Petr Volf

Tento příspěvek je věnován metodám diagnostiky a testování shody modelů a dat pro modely intenzity (poruch, proudu událostí). Přitom intenzita může záviset na některých vysvětlujících veličinách, půjde tedy o modely regresní. Nejprve připomeneme counting (čítací) procesy jako nástroj pro popis proudu události (či pro popis rozdělení doby čekání na určitou událost) ve spojitém čase. Dále zopakujeme metody odhadů pro regresní modely intenzit čítacích procesů, hlavně metody neparametrické, i když toto téma bylo už podrobně popsáno v příspěvku pro Robust'92.

Základem pro posuzování vhodnosti regresního modelu je analýza reziduí (či reziduálů) – odchylek naměřených dat od modelových (predikovaných) hodnot. V našem případě můžeme reziduály definovat několikerým způsobem, v podstatě jako odchylku individuální (porovnáváme-li doby čekání), či odchylku celé skupiny (to pak porovnáváme spíš frekvence výskytu událostí). Předvedeme především jednoduché a účinné grafické metody pro regresní diagnostiku (tj. zejména pro porovnání chování různých podsouborů dat vzhledem k modelu). Použité veličiny mají ale také vhodné asymptotické vlastnosti, které poslouží při numerickém vyhodnocování testů.

Klíčová slova: čítací proces, intenzita, zobecněný reziduál, Coxův regresní model, Poissonův proces, neparametrický (jádrový) odhad regresní funkce.

1. SCHÉMA SLEDOVÁNÍ OKAMŽIKŮ PORUCH, ČÍTACÍ PROCES

Připomeňme si nejdříve nejčastější situaci ve statistické analýze doby přežití. Máme náhodnou veličinu T popisující dobu do poruchy. Tu sledujeme na n objektech. Často ale je pozorování cenzorováno (je ukončeno dřív než dojde k poruše – to je cenzorování zprava). Předpokládáme, že dobu do cenzorování můžeme také popsat jako náhodnou veličinu V. Výsledkem pozorování je n-tice (Y_i, δ_i), $i = 1, \ldots, n$, kde $Y_i = \min(T_i, V_i)$, $\delta_i = \mathbf{1}[Y_i \leq V_i]$, T_i, V_i jsou vzájemně nezávislé kopie n. v. T, resp. V. Předpokládáme, že n. v. T má spojité rozdělení na $(0, \infty)$, s distribuční funkcí $F(t)$, hustotou $f(t)$ a intenzitou (rizikovou funkcí) $h(t) = f(t)/(1-F(t))$. Čas zde zpravidla běží pro každý objekt zvlášť (nemusejí začínat ve stejný "kalendářní" okamžik). Například, sledujeme-li životnost nějakých přístrojů, "čas" pro každý z nich běží od spuštění přístroje do jeho poruchy (či do konce jeho sledování).

Intenzita je tedy jednou z možných charakteristik rozdělení náhodné veličiny. Přitom ale tato charakteristika má určité výsadní postavení. Jak její jméno i definice říkají, intenzita charakterizuje okamžitou míru rizika. Protože ale už i z fyzikální podstaty věci je vznik poruchy dynamický proces (probíhající v čase), musí intenzita $h(t)$ v okamžiku t v sobě skrývat vliv celé minulosti provozu sledovaného přístroje v $[0,t)$ na možný vznik poruchy v t. Tento aspekt je lépe vidět v následujícím schématu, který popíše celou situaci (a i situace obecnější) pomocí aparátu náhodných čítacích procesů (counting processes).

Counting proces $N_1(t)$ je bodový náhodný proces, probíhající v čase $t \geq 0$, který, řekněme, má na začátku hodnotu $N_1(0) = 0$ a jehož trajektorie jsou po částech konstantní, spojité zprava, se skoky $+1$. Takový proces je tedy vhodný pro popis proudu náhodných událostí (nepřipouštíme-li 2 události současně). Podstatné je (to předpokládáme), že to, zda v $[t, t+dt]$ nastane skok, závisí pouze na historii spjaté s procesem v $[0,t)$. Označme $\sigma_1(t)$ σ-algebru (v prostoru elementárních jevů Ω_1), která obsahuje právě jevy z oné historie v $[0,t)$ relevantní pro další chování procesu. Jde tedy o neklesající posloupnost σ-algeber. nechť existuje taková nezáporná, ohraničená (a Leb. měřitelná) funkce h_1, že

$$\Pr\left(N_1(t+\Delta) - N_1(t^-) = 1 | \sigma_1(t)\right) = h_1(t)\,\Delta + o(\Delta). \tag{1}$$

Pak $h_1(t)$ je právě *intenzita* counting procesu N_1. Tady vidíme, že v "okamžité" intenzitě je skryt vliv minulosti. $N_1(t)$ je nyní vlastně obecný Poissonův proces. Z takového popisu vztahu procesu s historií (v té "historii" může být zahrnuto chování celého prostředí kolem sledovaného objektu) plynou hned další důsledky (podrobněji viz P. K. Andersen a spolupracovníci, Arjas, i Robust'92). Například existence rozkladu

$$N_1(t) = M_1(t) + H_1(t),$$

kde $M_1(t)$ je martingál a $H_1(t) = \int_0^t h_1(s)\,ds$ je kumulativní intenzita.

Při statistické analýze sledujeme zpravidla více objektů, tedy i vícerozměrný counting proces $N(t) = N_1(t), \ldots, N_n(t)$. $N_i(t)$ nechť načítá pozorované události týkající se i-tého objektu (událostí může být samozřejmě více za sebou – je spousta událostí, které se v životě přístroje i člověka několikrát opakují – například rozhodnutí přestat kouřit, že). Obecně nyní každý counting proces má svou intenzitu $h_i(t)$, ale předpokládáme, že mají společné historie $\sigma(t) = \sigma_1(t) \otimes \cdots \otimes \sigma_n(t)$. Neboli, prostřednictvím této společné historie jsou vývoje jednotlivých objektů závislé. Pak je tedy i nutné, aby čas t byl "společný" čas, nejlépe tedy čas kalendářní, s $t = 0$ v momentě, kdy se začalo s pozorováním prvního objektu.

Představme si třeba případ nějakého zařízení, v němž pracuje n prvků. Je jasné, že stav každého prvku má vliv na intenzitu poruchy ostatních. Nemůžeme tedy pominout vliv společné historie. Zároveň však pro každý prvek běží jeho individuální čas s_i, který začal v momentě jeho uvedení do provozu. Ten se v modelu pro intenzitu musí objevit také, jako další vysvětlující proměnná. Intenzita poruchy pro i-tý prvek by se pak nejspíš

modelovala jako

$$h_i(t) = I_i(t) \cdot \lambda_i(\mathbf{X}_i(t), s_i(t), t), \tag{2}$$

kde λ_i by byla zvolená funkce (asi $\lambda_i \equiv \lambda$, kdyby všechny prvky byly téhož druhu), $\mathbf{X}_i(t)$ = vysvětlující proměnná (kovariáta) popisující historii celého systému v $[0, t)$, $s_i(t) = t - S_i$ = individuální čas (S_i je moment instalace prvku) a konečně $I_i(t)$ je indikátor = 1 když je prvek v provozu, $I_i(t) = 0$ jinak. Takto jsme se dostali vlastně k *regresnímu modelu.*

2. REGRESNÍ MODEL PRO INTENZITU, ODHADOVÁNÍ REGRESNÍ FUNKCE

Většinou vystačíme s předpokladem (stejně jako v předchozím příkladě), že intenzitu i-tého counting procesu ovlivňují další dva procesy, a to nějaká vysvětlující proměnná $\mathbf{X}_i(t)$ (její hodnota se může – ale nemusí – měnit v čase, může to samozřejmě být vektor kovariát) a indikátor toho, zda je proces $N_i(t)$ vůbec pozorován, $I_i(t)$.

Model pro intenzitu je pak $h_i(t) = I_i(t) \cdot \lambda(t, \mathbf{X}_i(t))$ (individuální čas $s_i(t)$ z předchozího případu je vlastně jen jednou z komponent v $\mathbf{X}_i(t)$. Samozřejmě předpokládáme, že konkrétní realizaci "dat" $I_i(t)$, $N_i(t)$ a $\mathbf{X}_i(t)$ (to potřebujeme jen v t, kde $I_i(t) = 1$) už plně známe. Funkci $\lambda(t, \mathbf{x})$ volíme, a to je právě kámen úrazu a důvod pro vývoj metod testování dobré shody. Nejznámějším regresním modelem pro intenzitu je zřejmě Coxův, s $\lambda(t, \mathbf{x}) = \lambda_0(t) \cdot \exp \boldsymbol{\beta}'\mathbf{x}$, kde $\lambda_0(t)$ je jakási základní (baseline) intenzita a vliv kovariát je "loglineární".

Ještě k onomu problému se dvěma paralelně běžícími časy: Samozřejmě, pokud vývoj každého objektu nezávisí na historii objektů ostatních, nemusí být referenčním časem čas kalendářní, ale je lepší mít jako referenční čas individuální, neboť intenzita pak závisí hlavně na něm. Například v oné základní situaci sledování doby přežití (viz začátek článku) i každý counting proces $N_i(t)$ běží od (individuálního) $t = 0$, $N_i(0) = 0$, $h_i(t) = h(t) \cdot I_i(t)$. $N_i(t)$ zaznamená jeden jediný skok právě v okamžiku poruchy Y_i při $\delta_i = 1$ (při cenzorování zůstane $N_i(t) \equiv 0$) a objekt přestane být dále sledován. neboli $I_i(t) = 1$ v $[0, Y_i]$, $= 0$ v $t > Y_i$.

Tento příspěvek je sice věnován testování modelů pro intenzitu, ale k testování většinou potřebujeme odhadnout neznámé části zvoleného modelu. Jen někdy je možné testovat typ modelu na základě nějakých typických vlastností, dobře graficky (například) vyhodnotitelných. To se týká linearity, to se týká i proporcionality rizika (tj. Coxova modelu v případě, že kovariáty nezávisí na čase).

Odhadování regresních funkcí ve "věrohodnostních modelech", tj. takových, že lze sestavit věrohodnostní funkci a regresní funkce je parametrem tohoto likelihoodu, byla věnována část článku v Robustu'92, viz i Kybernetika (1993), Hastie, Tibshirani (1986), Stone (1986 – 1991). V modelech pro counting procesy máme k dispozici hned dvě věrohodností funkce. Předpokládejme, že pozorujeme procesy $N_i(t)$, $I_i(t)$, $\mathbf{X}_i(t)$, $i = 1, \ldots, n$

v $t \in [0,T]$. Úplný (podmíněný pozorovanými procesy $\mathbf{X}$, I) likelihood (vzpomeň na likelihood pro Poissonův proces) je

$$\mathcal{L}_n = \prod_{i=1}^{n} \left\{ \prod_{t \le T} h_i(t)^{dN_i(t)} \cdot \exp - \int_0^T h_i(s)\, ds \right\}, \tag{3}$$

kde tedy $h_i(s) = I_i(s) \cdot \lambda(s, \mathbf{X}_i(s))$. Vidíme, že likelihood je součinem individuálních likelihoodů, jako kdyby osudy jednotlivých objektů byly nezávislé. Ony sice jsou závislé na společné historii, ale jen prostřednictvím procesů $\mathbf{X}_i$, případně I_i (tak byl v (2) model vybudován). Čili součin pro $\mathcal{L}_n$ vyjadřuje *podmíněnou nezávislost při daných* $\mathbf{X}_i(t)$, $I_i(t)$, $t \in [0,T]$.

Představme si nyní, že pro funkci λ předpokládáme nějaký specifický tvar, například $\lambda(t,\mathbf{x}) = \alpha(a(t), b(\mathbf{x}))$, funkci α známe, funkce $a(t)$, $b(\mathbf{x})$ je třeba odhadnout. To už můžeme zkusit s pomocí logaritmu věrohodnostní funkce (3), a to buď metodou postupné maximalizace lokální věrohodnosti (local scoring – Hastie, Tibshirani, Robust'92) – což je ekvivalent jádrových odhadů, nebo po reparametrizaci funkcí a, b, např. kombinací splinů (Stone). Pro nejčastěji užívaný model, multiplikativní (nepaparametrický Coxův) s $\lambda(t,\mathbf{x}) = \lambda_0(t) \cdot \exp b(\mathbf{x})$ máme pro odhadování regresní funkce $b(\mathbf{x})$ k dispozici tzv. částečnou věrohodnostní funkci, jejíž logaritmus je

$$\ell_n^* = \sum_{i=1}^{n} \int_0^T \log \left\{ \frac{\exp b(\mathbf{X}_i(t))}{S(b,t)} \right\} dN_i(t), \quad \text{kde } S(b,t) = \sum_{j=1}^{n} I_j(t) \exp b(\mathbf{X}_j(t)). \tag{4}$$

Víme, že odhad kumulativní verze $L_0(t) = \int_0^t \lambda_0(s)\, ds$ základní intenzity pak dostaneme odhadem (Breslow, Crowley)

$$\hat{L}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{S(b,s)}. \tag{5}$$

Připomeňme nyní metodu maxima lokální věrohodnosti pro odhad jednotlivých složek aditivní regresní funkce $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$. (Tj. uvažujeme K-rozměrný vektor kovariát a předpokládáme, že vliv každé z nich na logaritmus intenzity je aditivní) – viz Robust'92.

Odhadujme tedy funkci b_ℓ v bodě z. Pro tuto chvíli považujme $b_\ell(x_\ell)$ za konstantu $b_\ell(z)$ v nějakém okolí $O(z)$, a řešíme rovnici

$$\frac{\partial \ell_n^*}{\partial b_\ell(z)} = \sum_{i=1}^{n} \int_0^T \left\{ 1[X_{\ell_i}(t) \in O(z)] - \exp b_\ell(z) \cdot \frac{R_\ell(z,b,t)}{S(b,t)} \right\} dN_i(t) = 0, \tag{6}$$

kde $R_\ell(z,b,t) = \sum_{j=1}^{n} 1[X_{\ell_j}(t) \in O(z)] \exp\left\{ \sum_{k \neq \ell} b_k(X_{k_j}(t)) \right\} \cdot I_j(t)$. Přímo z rovnice (6) se nabízí iterační krok pro získání nové hodnoty $b_\ell(z)$, máme-li už nějaký "starý" odhad všech komponent b_k, a to alespoň v bodech $X_{k_j}(T_i)$, pokud $I_j(T_i) = 1$ – tj. ve všech pozorovaných hodnotách kovariát. Přitom $i, j = 1, \ldots, n$. $k = 1, \ldots, K$, neboli musíme

mít uloženo $n \times n \times K$ dat, pokud nepoužijeme nějakou interpolaci (a pokud jsou všechny kovariáty proměnné v čase). Alternativní iterační krok by mohl být založen na běžném kroku Newtonova algoritmu s druhou derivací, tak to doporučují Hastie a Tibshirani (1986).

Způsob odhadu používal pohybujícího se obdélníkového okna, není tedy problém použít podobně nějakého okna obecnějšího (v praxi dávám přednost přístupu s m-nejbližšími sousedy, výsledné odhady funkcí $b_j(x_j)$ pak ještě jednou vyhladím).

3. REZIDUA A TESTY SHODY DAT S MODELEM

Jak jsme řekli v úvodu, pro hodnocení toho, jak zvolený model odpovídá skutečnosti (reprezentované daty) je nutné mít vhodně definovaná rezidua. Jaká veličina by mohla dobře odrážet odchylky modelu a dat v naší situaci, kdy model je zadán intenzitou counting procesu?

Vybereme dvě různá pojetí reziduí, která ale samozřejmě budou mít leccos společného. To společné je dáno příbuzností counting procesu s Poissonovým procesem. Vlastně, couting proces s intenzitou je obecnější jen o tu podmíněnost (společnou) historií. Takže při retrospektivním pohledu (když už realizaci oné historie známe), není rozdílu mezi nimi.

3.1. Rezidua "individuální", transformace na standardní Poissonův proces

Uvažujme i nadále n-složkový counting proces $N_i(t)$, $i = 1, \ldots, n$, na $[0, T]$. Model nechť je zadán kumulativními intenzitami $H_i(t) = \int_0^t h_i(s)\, ds$. Představme si na chvíli, že každý proces má nanejvýš 1 skok. Označme $S_i = \sup\{t \in [0.T],\ I_i(t) = 1\}$, $\delta_i = 1$ je-li S_i momentem skoku procesu $N_i(t)$, $\delta_i = 0$ jinak.

Neboli opět sledujeme nějakou n. v. T_i – dobu do události (skoku procesu $N_i(t)$), ta událost je pozorována s cenzorováním, tj. pozorujeme (S_i, δ_i), $i = 1, \ldots, n$. Připomeňme si ještě onu podmíněnou nezávislost jednotlivých procesů, jakmile známe celou "historii", tj. jakmile plně známe intenzity v $[0, T]$. Nechť $H_i(t)$ jsou "správné" kumulativní intenzity. Pak platí:

VĚTA. Náhodné veličiny $H_i(T_i)$, $i = 1, \ldots, n$ mají standardní exponenciální rozdělení a jsou vzájemně (podmíněně) nezávislé. $(H_i(S_i), \delta_i)$ jsou výsledky náhodného cenzorování zprava veličin $H_i(T_i)$.

Podobně, pokud budeme uvažovat opakující se události, $H_i(t)$ představuje jakousi transformaci času na čas standardního Poissonova procesu. Pokud T_{ij}, $j = 1, \ldots, N_i(T)$, jsou okamžiky skoků counting procesu $N_i(t)$, který má kumulativní intenzitu $H_i(t)$, tak $H_i(T_{ij})$ jsou okamžiky skoků pro Poissonův proces s jednotkovou intenzitou.

Rezidua můžeme tedy definovat jako $r_i = H_i(T)/N_i(T)$, je-li $N_i(T) > 0$, nebo přímo jako $H_i(T) = H_i(S_i)$ pro první případ jediného skoku.

Poznámka: Veličina $z_i = 1/r_i$ charakterizuje jakousi latenci (náchylnost k poruše) pro i-tý objekt. Také se jí říká "frailty" (křehkost) v analýze doby přežití. Je spíš zajímavé tuto veličinu zjišťovat pro nějakou skupinu S objektů, protože se často očekává, že úroveň latence je specifická pro objekty (pacienty) určité kategorie. Pak, označíme-li $\overline{H}_S(T) = \sum_{i \in S} H_i(T)$, $\overline{N}_S(T) = \sum_{i \in S} N_i(T)$, je $z_S = \overline{N}_S(T)/\overline{H}_S(T)$ ona skupinová charakteristika latence.

Mimochodem, z definice je vidět, že pokud $z_S \neq 1$, je vhodné ji přidat multiplikativně k modelu intenzity. Vylepšený model intenzity pak je $h_i^*(t) = h_i(t) \cdot z_S \cdot 1[i \in S]$.

PŘÍKLAD 1. Pro nedostatek prostoru budou příklady k tomuto příspěvku jednoduché a simulované. Nasimulovali jsme $n = 20$ nezávislých "dob do poruch" T_i splňujících Coxův model s intenzitou $h_i(t) = \alpha \exp(x_i^\beta + y_i^\gamma)$, $\alpha = 2$, $\beta = -0.5$, $\gamma = 0.3$. Hodnoty kovariát x, y byly simulovány nezávisle a rovnoměrně v $(0, 10)$, resp. v $(0, 20)$. Hodnoty T_i byly řádu 10^{-2} až 10^1. Přidali jsme dva outliery $T_{21} \sim 10^{-6}$, $T_{22} \sim 10^3$. Data nebyla cenzorována. Nyní jsme v souladu s předchozím textem pro každé i-té datum spočetli jackknifeovaná rezidua $r_i^* = H_i^*(T_i)$, kde v $H_i^*(t)$ jsou použity odhady parametrů α, β, γ z celého výběru bez i-tého pozorování. Podobně jako se pro rezidua při standardní analýze regrese dělá normální graf, můžeme nyní něco podobného udělat pro rezidua r_i^* – či pro jejich další transformace. Tak na obr. 1 jsou hodnoty $u_i^* = 1 - \exp(-r_i^*)$, což by v ideálním případě měla být realizace nezávislých veličin, rovnoměrně rozdělených na $[0, 1]$. Pásy kolem jsou sjednocením nasimulovaných 90 % intervalů spolehlivosti pro pořádkové statistiky z 22 členného i.i.d. výběru s $U(0, 1)$ rozdělením. Na obrázku 2 je tentýž případ, tentokrát po transformaci na standardní dvojitě-exponenciální rozdělení, tj. s veličinami $d_i^* = \log r_i^*$.

Numerické metody testování pro tento typ reziduí mohou být založeny na jakékoli proceduře ověřující určitý typ distribuce např. standardní exponenciální distribuci, nebo gamma distribuci, zkoumáme-li transformovaný čas $\overline{H}_S(t)$. Je-li totiž T_{Sk} čas do k-té události ve skupině S, má veličina $\overline{H}_S(T_{Sk})$ gamma $(1, k)$ rozdělení (je-li model "správný").

3.2. Využití rozkladu "kompenzátor + martingal"

Předchozí metoda byla spíš vhodná pro menší soubory dat, a pro zkoumání případné odlehlosti jednotlivých dat. Nyní popíšeme grafickou metodu vhodnou pro testování shody dat a modelu pro intenzitu pro větší soubory dat. Pro testování Coxova modelu ji navrhl Arjas (1988) a přímo v ní počítá s rozdělením celého souboru na 2 či více podsouborů a s testováním pro každý soubor zvlášť. Metoda vychází z už zmíněného rozkladu (stále na $[0, T]$)

$$H_i(t) = N_i(t) - M_i(t).$$

Nechť nyní $S \subset \{1,\ldots,n\}$ je opět skupina (podsoubor) objektů, pak po sečtení přes $i \in S$ dostaneme $\overline{H}_S(t) = \overline{N}_S(t) - \overline{M}_S(t)$. Grafické testování bude založeno na porovnání modelů (zde reprezentovaného kumulativními intenzitami $H_i(t)$) a dat ($N_i(t)$), rozdíl je martingál (s nulovou střední hodnotou). Označme nyní opět T_{Sk} k-tý pozorovaný skok ve skupině S. Pak $\overline{N}_S(T_{Sk}) = k$ a grafická metoda přímo se nabízející je porovnávat hodnoty $\overline{H}_S(T_{Sk})$ s k.

V případě, že předpokládáme platnost Coxova modelu, máme

$$H_i(t) = \int_0^t I_i(s)\,\lambda_0(s)\exp b(X_i(s))\,ds.$$

Pokud $\lambda_0(s)$ neznáme, můžeme ji nahradit odhadem (viz (5))

$$\hat{\lambda}_0(s)\,ds = \sum_{i=1}^{n} dN_i(s)/S(b,s).$$

Pak dostáváme dosazením do $\overline{H}_S(t)$

$$(7) \qquad H_S^*(t) = \int_0^t \left[\sum_{i\in S} I_i(s)\exp b(X_i(s))\right] \Big/ \left[\sum_{i=1}^{n} I_i(s)\exp b(X_i(s))\right] \sum_{i=1}^{n} dN_i(s).$$

To je ona původní Arjasova statistika. Pokud neznáme ani funkci b, ale nahradíme ji "dobrým" odhadem, bude grafický test mít stále svou diagnostickou cenu.

PŘÍKLAD 2. Zůstali jsme u téhož modelu jako v Př. 1, tentokrát jsme nasimulovali 200 hodnot. Rozdělili jsme data na 2 přibližně stejně velké skupiny, podle toho, zda bylo $y_i < 10$ či $y_i \geq 10$. Pro Obr. 3 jsme odhadli parametry β a γ a zobrazili jsme pro obě skupiny statistiku (7) v $t = T_{Sk}$ (s $b(x,y) = x^{\hat{\beta}} + y^{\hat{\gamma}}$). Grafy hodnot $H_S^*(T_{Sk})$ pro obě skupiny jsou blízko diagonály, není tedy důvod model zamítnout. Pak jsme "zapomněli" na závislost intenzity na y a zkoušeli, zda nestačí model s $b(x) = x^{\beta}$. Výsledek, opět pro statistiku (7) a skupiny $\{y_i < 10\}$ a $\{y_i \geq 10\}$, je na obr. 4. Grafy pro obě skupiny se vzdalují od diagonály $H_S^*(T_{Sk}) = k$, obrázek napovídá, že model není dobrý a že intenzita ve skutečnosti závisí (kladně) na kovariátě y (doby T_{Sk} ve druhé skupině s $y \geq 10$ jsou značně menší – čili intenzita je tam ve skutečnosti značně vyšší – než předpokládá model).

3.3. Asymptotické vlastnosti reziduálů

V předchozím přístupu jsme vlastně sledovali chování skupinových "reziduálů" (které jsme dostali jako součet individuálních odchylek) $\overline{N}_S(t) - \overline{H}_S(t)$. Pokud by model byl "správný", je $\overline{N}_S(t) - \overline{H}_S(t) = \overline{M}_S(t) = \sum_{i\in S} M_i(t)$, tj. martingal. Označme $|S| = \sum_{i\in S} 1$ počet prvků v S. Motivací pro zkoumání asymptotiky v tomto případě je očekávání, že $|S|^{-\frac{1}{2}}\overline{M}_S(t)$ splňuje nějakou verzi centrální limitní věty.

Zkoumejme nejprve případ Coxova modelu (stejně jako Marzec a Marzec, 1993), $\lambda(t,x) = \lambda_0(t) \cdot \exp(\beta' x)$. Řekněme, že jsme v situaci, kdy chceme otestovat, že Coxův model je vhodný pro naše data, ale nevíme, který (s kterými parametry). Budeme tedy testovat hypotézu H_0, že data odpovídají Coxovu modelu s nějakým nám neznámým β_0, použijeme testovou statistiku $H_S^*(t)$ (pro podsoubor S i pro podsoubor doplňkový $\overline{S}$ – tuto statistiku nelze použít přímo pro celý soubor, jak snadno zjistíme, její hodnota by byla vždy přímo k v T_{Sk}), do které dosadíme $\hat{\beta}$ = odhad Coxova parametru. Ten získáme standardním způsobem, z částečné věrohodnostní funkce. $L_0(t)$ odhadneme dodatečně, k testování ji nepotřebujeme. Označme $D_S^*(\beta,t) = H_S^*(t) - \overline{N}_S(t)$, když v $H_S^*(t)$ je dosazena hodnota β. Označme ještě $R_S(\beta,t) = \frac{1}{|S|}\sum_{i\in S} I_i(t)\exp(\beta' \mathbf{X}_i(t))$, $R(\beta,t) = \frac{1}{n}\sum_{i=1}^{n} I_i(t)\exp(\beta' \mathbf{X}_i(t))$, R' derivace R podle β (budou to p-rozměrné vektory, je-li p dimenze β). Stále zůstáváme na $t \in [0,T]$, $\lambda_0(t)$ tam předpokládáme ohraničenou.

PŘEDPOKLADY: 1. Nechť $n \to \infty$, $|S| \to \infty$ tak, že $|S|/n \to q \in (0,1)$.

2. Jsou splněny předpoklady B, D z Andersen a Gill (1982), potřebné k silné konzistenci a as. normalitě odhadu $\hat{b}$. Konkrétně jde o to, že existují funkce $r(\beta,t)$, $r'(\beta,t)$ které jsou stejnoměrné (v t a v β v nějakém okolí β_0) P-limity R a R' (a také R_S a R'_S), přitom $r' = dr/d\beta$ a $r \geq \varepsilon > 0$.

Za těchto předpokladů, při hypotéze H_0 pak platí:

VĚTA. (Marzec a Marzec, 1993). $n^{-\frac{1}{2}} D_S^*(\hat{\beta},t)$ konverguje na $[0,T]$ slabě (tj. v distribuci) ke (gaussovskému) Wienerovu procesu $W(t)$, který má varianci

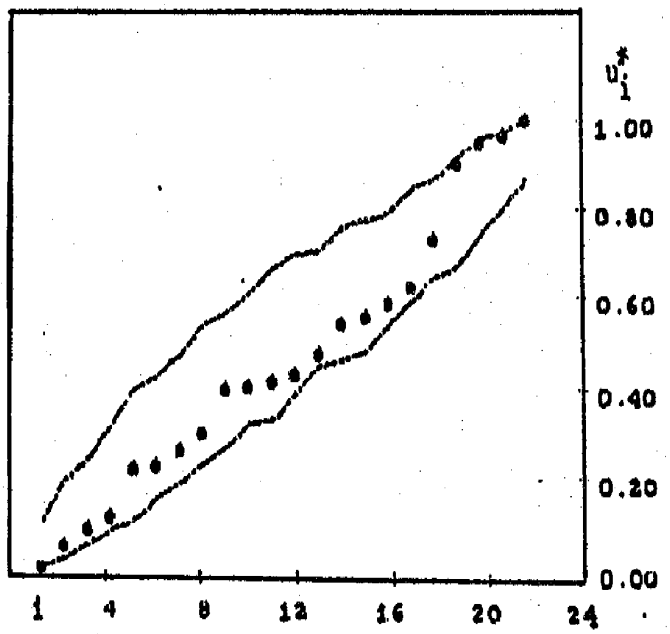
$$\operatorname{var} W(t) = q(1-q)\int_0^t r(\beta_0,s)\,\lambda_0(s)\,ds.$$

Na základě tohoto výsledku můžeme tedy pro statistiku $n^{-\frac{1}{2}} D_S^*(\hat{\beta}t)$ sestrojit přibližné pásy spolehlivosti v $[0,T]$ (asymptotickou varianci odhadneme).

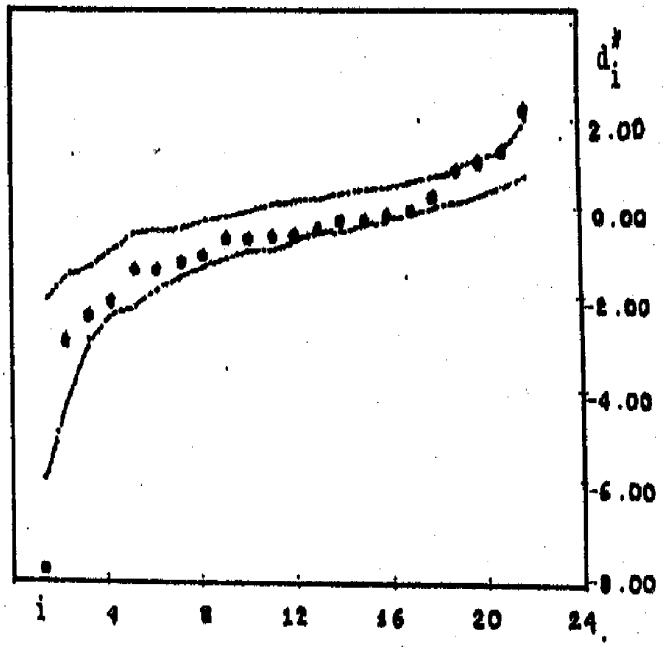
Jiná situace by nastala, jakmile by v Předpokladu 2 byly limitní funkce různé pro různé podsoubory (tj. $\lim R_S \neq \lim R$ například). Marzec a Marzec ukazují, že stále ještě má $n^{-\frac{1}{2}} D_S^*(\hat{\beta},t)$ limitní rozdělení gaussovského procesu, ale už nejde o Wienerův proces.

Představme si nyní, že testovaný model není multiplikativního typu, testujeme hypotézu, že intenzita je $\lambda_0(t,x)$, k testování použijeme statistiku $D_S(\lambda_0,t) = \overline{N}_S(t) - \overline{H}_S(\lambda_0,t)$, kde $\overline{H}_S(\lambda,t) = \int_0^t \sum_{i\in S} I_i(s)\,\lambda(s,X_i(s))\,ds$. Nyní je přímo $n^{-\frac{1}{2}} D_S(\lambda_0,t) = n^{-\frac{1}{2}}\overline{M}_S(t)$. Protože (viz 1. část či Robust'92) je $\operatorname{var} dM_i(t) = I_i(t)\,\lambda_0(t,X_i(t))\,dt$ a $\operatorname{cov}(dM_i(t), dM_j(t)) = 0$ pro $i \neq j$, je

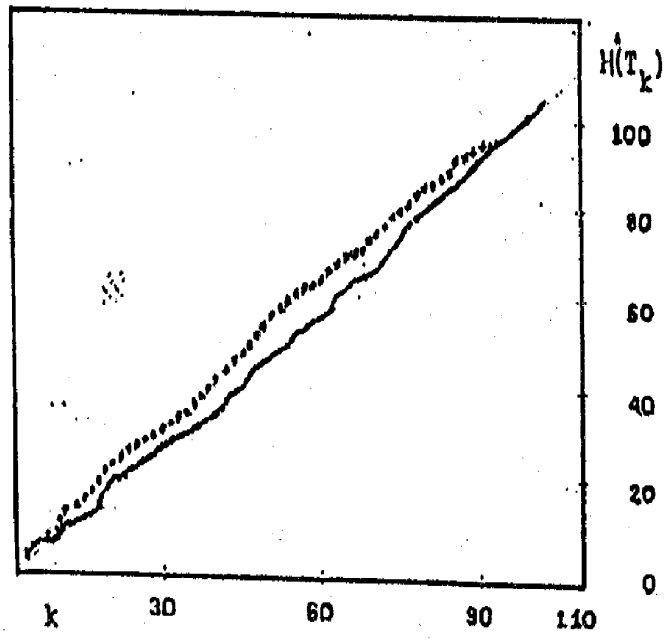
$$\operatorname{var}\left(|S|^{-\frac{1}{2}}\overline{M}_S(t)\right) = \int_0^t \frac{1}{|S|}\sum_{i\in S} I_i(s)\,\lambda_0(s,X_i(s))\,ds.$$



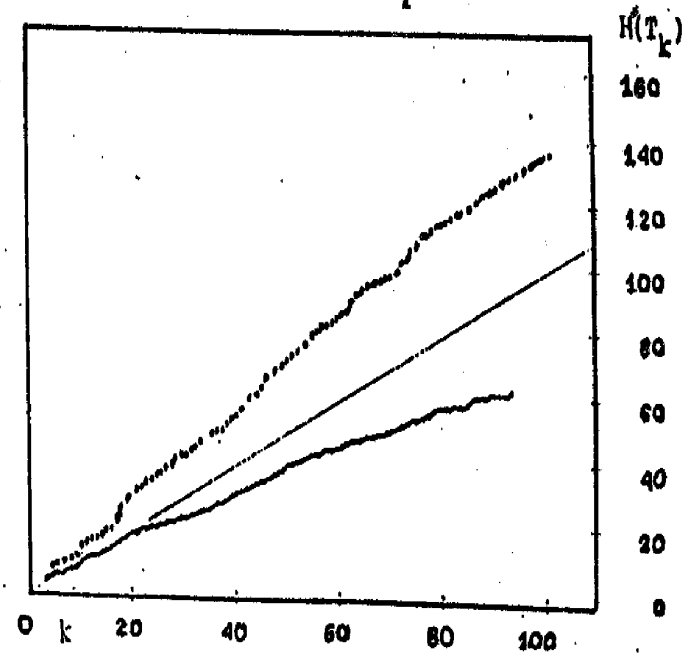
Obr.1. Seřazená rezidua u_i^* a 90% konf. intervaly



Obr.2. Totéž pro rezidua d_i^*



Obr.3. Tečkovaně pro y>10, plně pro y<10



Obr.4. Pro změnu tečkovaně pro y<10, plně pro y>10

Předpoklad 3. Existuje $\mathcal{L}(\lambda, s) = P - \lim_{n\to\infty} \frac{1}{|S|} \sum_{i\in S} I_i(s)\,\lambda(s, X_i(s))$, a to stejnoměrně v $s \in [0,T]$ a pro funkce $\lambda(s,x)$ v nějakém okolí funkce $\lambda_0(s,x)$.

Samozřejmě předpokládáme, že funkce $\lambda(s,x)$ jsou stejnoměrně ohraničené na $[0,T] \times \mathcal{X}$ (což je oblast hodnot x). Za předpokladů 1 a 3, je-li λ_0 "skutečnou" intenzitou pro námi pozorované counting procesy, konverguje $n^{-\frac{1}{2}}\overline{M}_S(t)$ slabě ke gaussovskému Wienerovu procesu s variancí

$$q \int_0^t \mathcal{L}(\lambda_0, s)\, ds.$$

Jak to vypadá v případě, kdy místo λ_0 musíme použít její odhad $\hat{\lambda}$? K testu nyní používáme statistiku

$$\begin{aligned} n^{-\frac{1}{2}} D_S(\hat{\lambda}, t) &= n^{-\frac{1}{2}}(\overline{N}_S(t) - \overline{H}_S(\hat{\lambda}, t)) = \\ &= n^{-\frac{1}{2}}\overline{M}_S(t) + n^{-\frac{1}{2}} \int_0^t \sum_{i\in S} I_i(s)[\lambda_0(s, X_i(s)) - \hat{\lambda}(s, X_i(s))]\, ds. \end{aligned}$$

Tuhle situaci je asi lepší rozebrat zvlášť pro konkrétní tvar intenzity.

Ještě malou poznámku k tomu, proč jsme zdůrazňovali testy shody modelů s podsoubory dat. Prostě proto, že kromě globální shody (pro všechna data) nás většinou velmi zajímá i to, jak zvolený model vyhoví určitým skupinám objektů, kde jsou případné nedostatky našeho modelu či odchylky dat – viz poznámka o latencích a "křehkostech". Mimochodem, jakmile porovnáváme podsoubory, děláme vlastně už testy homogenity.

Literatura

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

Arjas E. (1988): A graphical method for assesing goodness-of-fit in Cox's proportional hazard model. JASA 83, 204–212.

Atkinson A. C. (1981): Two graphical displays for outlying and influential observations in regression. Biometrika 68, 13–20.

Barlow W. E., Prentice R. L. (1988): Residuals for relative risk regression. Biometrika 75, 65–74.

Hastie T., Tibshirani R. (1986): Generalized additive models (with discussion). Statist. Science 1, 297–318.

Marzec L., Marzec P. (1993): Goodness of fit inference based on stratification in Cox's regression model. Scand. J. Statist. 20, 227–238.

Stone C. J. (1991): Asymptotics for doubly flexible logspline response models. Ann. Statist. 19, 1832–1854.

Thernau T. M., Grambsch P. M., Fleming T. R. (1990): Martingale-based residuals for survival models. Biometrika 77, 147–160.

Volf P.: In Robust'92, Kybernetika 1993, č. 4.