ZÁKLADY STATISTICKÉ ANALÝZY PŘEŽITÍ, S APLIKACÍ V ANALÝZE SPOLEHLIVOSTI

P. Volf
ÚTIA AV ČR, *volf@utia.cas.cz*

1 Úvod, analýza přežití

Každá oblast využití matematické statistiky má své zvláštnosti, nejinak tomu je i s analýzou přežití (survival analysis). 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 přežití 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, statistickými odhady, i analýzou regrese v modelech s intenzitou. Pro seznámení s modely i metodami je k dispozici mnoho literatury, ze základní jmenujme třeba Kalhfteisch a Prentice (2002), Andersen, Borgan, Gill, Keiding (1993), Fleming a Harrington (1991). Základní informace a odkazy lze najít i na Wikipedii.

Oblastmi užití modelů pro rizikovou funkci je především statistická analýza spolehlivosti a biostatistika, ale i další oblasti, kde se zkoumají doby trvání jevů, jako demografie i ekonomie a sociální vědy. 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 Základní charakteristiky spolehlivosti

Zde máme na mysli kvantitativní charakteristiky odvozené z pravděpodobnostního pohledu na výskyt poruch. Nechť doba fungování nějakého zařízení až do poruchy je náhodná veličina T ($T > 0$). Ta nechť je dána hustotou rozdělení pravděpodobnosti $f(t)$, příslušnou distribuční funkcí $F(t) = \int_0^t f(s)\,ds = P(T \leq t)$, dále se užívá pojem **"funkce spolehlivosti"** či "funkce přežití" $S(t) = 1 - F(t)$. Ta má význam pravděpodobnosti, že zařízení přežije bez poruchy dobu t, tj. $S(t) = P(T > t)$. Ještě další zajímavou charakteristikou je **intenzita poruch** (či riziková funkce), která je definována jako

$$h(t) = \lim_{d \to 0_+} \frac{P(T \in (t, t+d])}{d \cdot P(T > t)} = \frac{f(t)}{S(t)} = -\frac{d \ln S(t)}{dt},$$

neboli vlastně jako hustota pravděpodobnosti pro jev, že porucha nastane právě "teď" – v čase t, za podmínky, že zařízení přežilo do času t. Další charakteristikou je **kumulovaná intenzita poruch** $H(t) = \int_0^t h(s)ds$. Z předchozích vztahů plyne, že $H(t) = -ln(S(t))$.

Podívejme se, jaké typy rozdělení pravděpodobnosti se nejčastěji používají pro popis náhodné doby do poruchy.

2.1 Exponenciální rozdělení

Exponenciální rozdělení má hustotu rozdělení pravděpodobnosti

$$f(t) = \frac{1}{a}\, e^{-\frac{t}{a}}, \quad \text{pro } t \geq 0,$$

$f(t) = 0$ pro $t < 0$, $a > 0$ je parametr tohoto rozdělení. Má význam jak střední hodnoty $\mathrm{E}T$, tak směrodatné odchylky. Distribuční funkce je $F(t) = 1 - e^{-\frac{t}{a}}$, funkce přežití $S(t) = e^{-\frac{t}{a}}$, a

intenzita poruch je konstantní, $h(t) = \lambda = \frac{1}{a}$. To znamená, že očekávání poruchy se nemění v závislosti na čase, exponenciální rozdělení popisuje dobu do poruchy pro zařízení, které (ve sledovaném období) nestárne. Zkusme spočítat pravděpodobnost pro dobu do poruchy v případě, že zařízení již přežilo dobu D, jako podmíněnou pravděpodobnost:

$$P(T > D + s|T > D) = \frac{P(T > D + s, T > D)}{P(T > D)} = \frac{P(T > D + s)}{P(T > D)} = \frac{e^{-(D+s)/a}}{e^{-D/a}} = e^{-s/a}.$$

Vidíme, že rozdělení dalšího přežití doby s vůbec nezávisí na již uběhnuté době D. To znovu ukazuje, že v takto popsaném období zařízení "nestárne". Často se můžeme setkat i se značením $f(t) = \lambda e^{-\lambda t}$, $S(t) = e^{-\lambda t}$, tj. používá se parametr λ místo $\frac{1}{a}$. Tento tvar se užívá například v Excelu, kdežto Matlab užívá tvar s $\frac{1}{a}$.

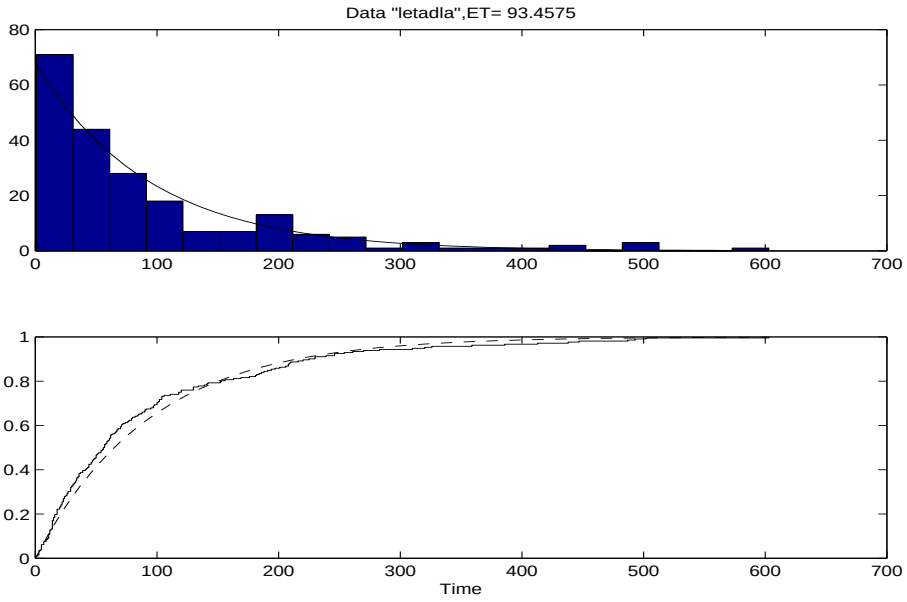

Figure 1: Histogram dat (nahoře) a porovnání empirické distribuční funkce s exponenciální s $\lambda = 0.0107$ (dole)

.

Příklad 1

Příklad analyzuje doby mezi poruchami klimatizací v letadlech Boening, data jsou uvedena v Barlow, Proschan (1967). Jde o opakované poruchy po opravách u 13 letadel, celkem 212 poruch, čas je měřen v hodinách provozu. Hypotéza je, že jde o Poissonův proces obnovy se stálým exponenciálním rozdělením dob mezi poruchami, tj. se stálou intenzitou poruch λ. Obrázky 1. a 2. to potvrzují, přesný test dobré shody také, nutno použít variantu testu Kolmogorova Smirnova pracující s odhadem parametru, tzv. test Lillieforse, či nějaký jiný test na exponencialitu rozdělení. Podstatné je, jak ukazuje Obr. 2, že kumulovaná intenzita poruch je skutečně zhruba lineární, tj. intenzita je konstantní. Znamená to, že v sledovaném období jsou zařízení po opravě vždy "jako nová". Z dat jsme spočetli následující hodnoty jejich charakteristik: průměr a směrodatná odchylka: 93.4575, 106.9036, šikmost a špičatost: 2.1062, 7.8949, kvartily: 22, 56.5, 119, minimum a maximum: 1, 603. Odhad intenzity je tedy $\hat{\lambda} = 1/93.4575 = 0.0107$.

Jenže typický průběh intenzity poruch pro reálná zařízení v celém jejich životním cyklu má velice často "vanovitý" tvar, viz Obrázek 3a.

První část, kde intenzita klesá, popisuje interval, kdy jsou odstraňovány výrobní závady, např. během záběhu, zkušebního provozu resp. jsou vyřazovány vadné kusy. Představme si

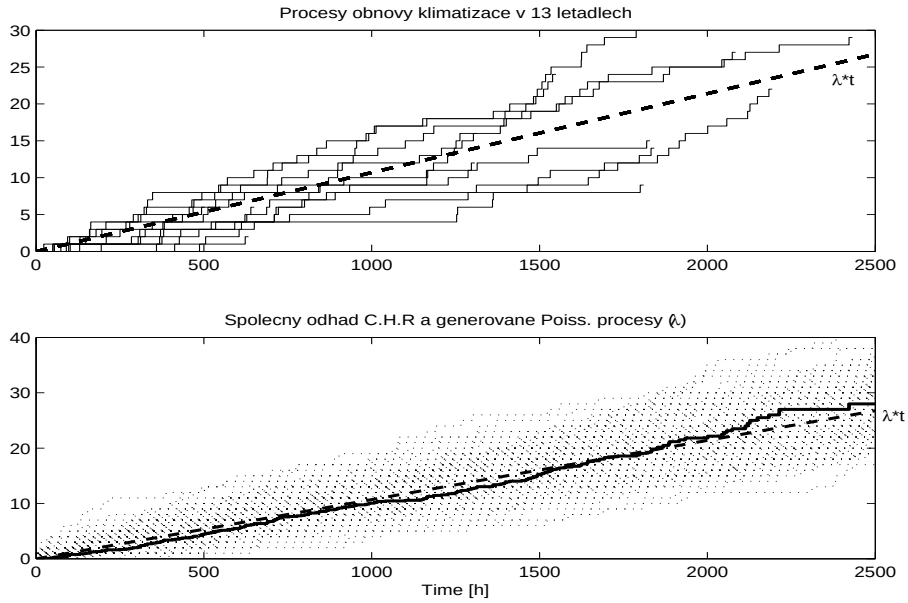

Figure 2: Procesy poruch na jednotlivých objektech (2a) i společný odhad kumulované intenzity poruch (2b) -průměr procesů z Obr.2a, v 2b) je i (tečkovaně) ukázka generovaných Poissonových procesů s $\lambda = 0.0107$

.

i třeba nově napsaný program, než se odstraní chyby (které tam vždy jsou). Pak následuje část s konstantní intenzitou poruch, kdy zařízení spolehlivě pracuje. Toto období by mělo být co nejdelší (vzhledem k účelu zařízení) a samozřejmě s co nejnižší intenzitou poruch. A v posledním období se již intenzita poruch zvyšuje, roste v důsledku stárnutí zařízení, tj. projeví se opotřebení i únava materiálu. Mimochodem, i živé organizmy (včetně lidí) mají takovýto tvar "intezity poruch" během svého života.

Jindy bývají tvary intenzit pozorované z nesourodého souboru výsledkem smíchání veličin několika typů. Takový tvar pak už nepopisuje "šanci" jednotlivého objektu, protože vznikl kompozicí ze skupin mající i zcela odlišný tvar intenzity. Mnoho o kompozici intenzit poruch pro technické objekty je v klasické knize Barlow, Proschan (1967). Takový případ je znázorněn na Obr. 3b. Například, mezi nezaměstnanými lidmi jsou dvě rozlišitelné skupiny. Jedni jsou jednak schopní i dost kvalifikovaní i mají zájem novou práci najít (tzv. "movers"), druhou skupinu tvoří tzv. "stayers" kteří zůstavají bez zaměstnání delší dobu.

Z příkladů je tedy vidět, že je potřeba i dalších typů rozdělení pravděpodobnosti, z nichž je pak možné takový reálný model intenzity složit.

2.2 Weibullovo rozdělení

Weibullovo rozdělení pravděpodobnosti má funkci přežití a hustotu

$$S(t) = \exp\left(-\left(\frac{t}{a}\right)^b\right), \quad f(t) = -\frac{dS(t)}{dt} = \frac{b\,t^{b-1}}{a^b} \cdot S(t),$$

čili intenzita poruch je

$$h(t) = \frac{b}{a^b} \cdot t^{b-1}.$$

Přitom parametry jsou a, $b > 0$, čas $t > 0$. Vidíme, že pro $b \in (0, 1)$ dostáváme klesající intenzitu $h(t)$, pro $b > 1$ rostoucí $h(t)$, a pro $b = 1$ dostáváme exponenciální rozdělení s konstantní $h(t) = 1/a$. Takže vanovitý průběh celoživotní intenzity poruch lze sestavit z takovýchto 3 částí, jako jejich směs, tak tomu je i na Obrázku 3a.

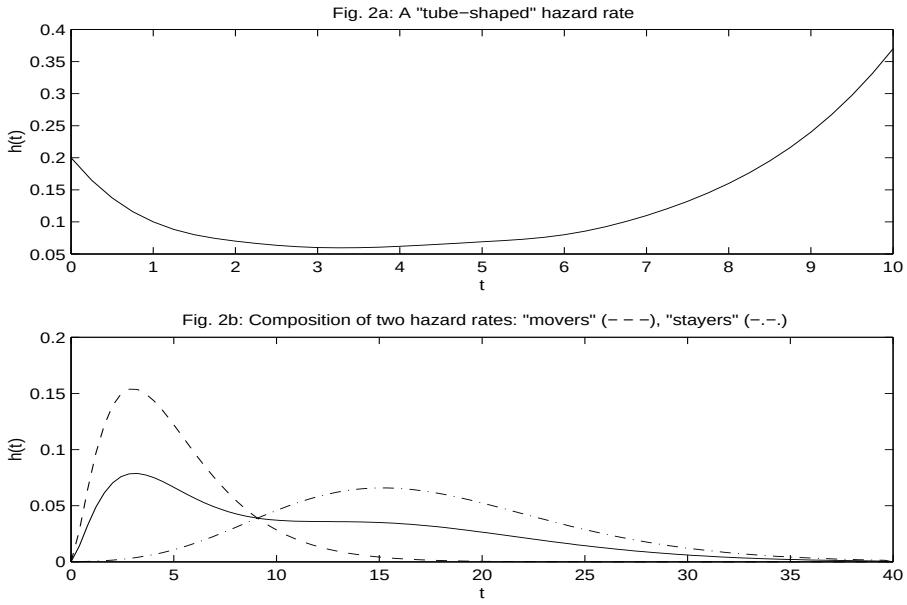

Figure 3: Příklady intenzit

Pro Weibullovo rozdělení je výpočet střední hodnoty, rozptylu (i dalších momentů) trochu komplikovaný:

$$EX = \int_0^\infty x \cdot f(x)\,dx = \int_0^\infty x\frac{bx^{b-1}}{a^b}exp(-(\frac{x}{a})^b)dx = \int_0^\infty az^{\frac{1}{b}}e^{-z}dz = a\cdot\Gamma(\frac{1}{b}+1).$$

Použili jsme substituci $z = (x/a)^b$, symbol $\Gamma(c+1) = \int_0^\infty z^c e^{-z}dz$ označuje t.zv. (úplnou) gamma funkci. Obdobně dostaneme

$$E(X^2) = \int_0^\infty x^2 \cdot f(x)\,dx = \int_0^\infty a^2 z^{\frac{2}{b}}e^{-z}dz = a^2\cdot\Gamma(\frac{2}{b}+1),$$

Takže rozptyl je $\mathrm{var}(X) = E(X^2) - (EX)^2 = a^2\Gamma(\frac{2}{b}+1) - (a\Gamma(\frac{1}{b}+1))^2$.

Někdy se používá rozdělení pravděpodobnosti, které má "počátek" posunutý z bodu 0 do nějakého bodu T_0, tj. příslušná náhodná veličina se popisuje od času T_0. Pak je např. funkce spolehlivosti takto posunutého Weibullova rozdělení $S(t) = exp(-(\frac{t-T_0}{a})^b)$ pro $t > T_0$, $S(t) = 1$ pro $t \le T_0$.

Dalšími typy rozdělení pravděpodobnosti, které se dají úspěšně použít pro modelování náhodné veličiny T = doba do poruchy, jsou například: Gamma rozdělení, log-normální rozdělení (to znamená, že $\ln T$ má normální rozdělení), nebo i normální (Gaussovo) rozdělení (s tím, že to připouští i záporné hodnoty, ale se zanedbatelnou pravděpodobností, pokud $\mu \gg \sigma$), i řada dalších.

Gamma rozdělení **Gamma**(λ, a) má hustotu $f(t) = \lambda^a t^{a-1}e^{-\lambda t}/\Gamma(a)$, s parametry $\lambda, a > 0$. Limitní hodnota $lim_{t\to\infty}h(t) = \lambda$, přičemž je $h(t)$ klesající pro $a < 1$, rostoucí pro $a > 1$. Lognormální distribuce nemá intenzitu monotónní, limitní hodnoty v 0 i v ∞ jsou 0.

2.3 Příklad 2

Další příklad ukazuje výdrž 99 nylonových vláken při postupném napínání (síla, resp. odpor vlákna, tj. vlastně "stres", je měřena v Newtonech). Jako model pro tato data se hodí víceméně kterékoli ze zmiňovaných rozdělení (Weibullovo s počátkem posunutým do minima dat), Obr. 4 zachycuje shodu s normálním rozdělením. Odhadnuté charakteristiky byly: průměr a směrodatná odchylka: 7.2061, 0.1530, šikmost a špičatost: -0.1780, 3.0072, kvartily: 7.1120, 7.2120, 7.3080, minimum a maximum: 6.7940, 7.5780.

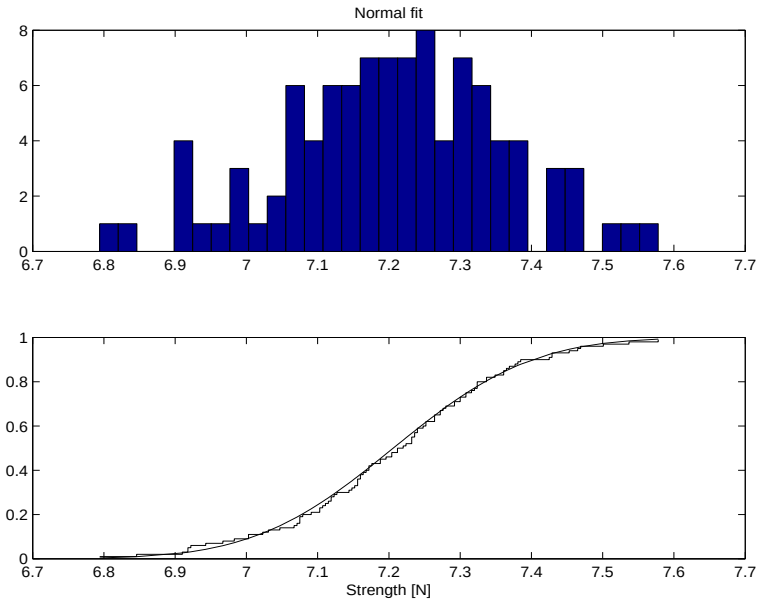

Figure 4: Histogram dat výdrže nylonových vláken (nahoře) a porovnání empirické distribuční funkce s distribuční funkcí normálního rozdělení s $\mu = 7.2061, \sigma = 0.1530$ (dole).

Poznámka: Při výpočtu špičatosti se opět použité formulky mohou lišit, zde uvedené hodnoty jsou spočteny v Matlabu, Excel odečítá -3 (3 je špičatost normálního rozdělení). Neboli ideální normální rozdělení má v Matlabu špičatost 3, v Excelu 0.

3 Číselné charakteristiky pro spolehlivost

Často se při kvantitativním popisu spolehlivosti zařízení používají jen průměrné hodnoty, které odpovídají některým důležitým charakteristikám (parametrům) pravděpodobnosti poruch:

1. **Střední doba do poruchy (MTTF = Mean Time to Failure)** není nic jiného než ET (střední hodnota náhodné veličiny T, doby do poruchy). Například při exponenciálním rozdělení je $ET = a = 1/\lambda$. Obecně tedy

$$ET = \int_0^\infty t\, f(t)\, \mathrm{d}t = \int_0^\infty S(t)\, \mathrm{d}t.$$

2. Stejným způsobem se popisuje

   Střední doba mezi poruchami (MTBF = Mean Time Between Failures), tj. jako střední hodnota náhodné veličiny T^* = doba do další poruchy. V případě exponenciálního rozdělení (a tedy Poissonova procesu poruch s konstantní intenzitou λ) je tedy i MTBF $= 1/\lambda$, jinak obecně MTBF závisí na stáří zařízení a může záviset i na počtu předchozích poruch, pokud je s poruchou spojena nějaká degradace (zhoršení stavu) zařízení, a závisí samozřejmě i na údržbě (zlepšení stavu).

   V těchto úvahách, kdy se pracuje s takto jednoduše definovanými (průměrnými) dobami, se vlastně předpokládá, že spolehlivost zařízení je popsatelná konstantní intenzitou poruch a že i po případné opravě je ve stejném stavu jako před poruchou (a vlastně "jako nové"). Zavádí se také:

3. **Střední doba opravy (MTTR = Mean Time to Repair)**. Dobu opravy (označme ji třeba Y) lze také chápat jako (nezápornou) náhodnou veličinu s nějakým rozdělením (ať už stejných typů jako pro dobu do poruchy, nebo třeba i s rovnoměrným rozdělením), a tedy s nějakou střední hodnotou MTTR $= EY$. Pak v takto "ustáleném" (či zprůměrovaném)

procesu poruch a oprav má smysl uvažovat i další průměrné charakteristiky, například **dostupnost** (availability): $A = MTTF/(MTTF + MTTR)$.

4. $1-\alpha$-procentní **zaručená doba životnosti**. To je taková doba T_α, že $P(T \geq T_\alpha) = 1-\alpha$ neboli T_α je α-kvantil rozdělení náhodné veličiny T - doby do poruchy.

Poznámka: V některých oblastech statistické analýzy, např. při modelování finančních řad, se používají rozdělení pravděpodobnosti "s těžkými konci", pro které je integrál udávající střední hodnotu či rozptyl nekonečný, jde např. o Paretovo rozdělení s $S(t) = (A/t)^a$ definovaném na $[A, \infty)$, má-li parametr $a \leq 1$.

4 Cenzorovaná data

U dat, která nějak souvisí s událostmi v čase, se často stane, že k dispozici není dost dlouhý časový interval a nestačí dojít k události, na kterou čekáme. Tak vzniknou neúplná data, kterým se říká data cenzorovaná. Představme si ten případ, kterým se zde převážně zabýváme, a to pozorování doby do poruchy. Nechť v čase 0 začneme pozorovat výdrž N součástek a pozorování všech ukončíme v čase C. U některých součástek došlo k poruše v čase $T_i \leq C$, u dalších pak do času C k poruše nedošlo, naše informace je, že příslušné T_i by bylo větší než C. Toto je schema **cenzorování 1. typu, cenzorování pevným časem**.

Cenzorování 2. typu je takové, kdy si řekneme, že pozorování ukončíme po R-té poruše, $R \leq N$. Tím si řídíme, kolik úplných pozorování (R z N) budeme mít k dispozici, ale doba, než dojde k R-té poruše, je náhodná.

Při obou typech cenzorování vlastně usekáváme některá (ta dlouhá) měření zprava. Dost časté, zejména v podobných sledováních v oblasti medicíny, je **náhodné cenzorování zprava**. Představme si třeba situaci, kdy pacienti s určitou diagnózou přicházejí v náhodných časech, podstoupí lékařský zákrok a sledujeme náhodnou veličinu – dobu do nějaké reakce, třeba do návratu k normálu. A pak v určitý okamžik data chceme shromáždit k analýze. Jenže dík náhodnému začátku byl každý pacient sledován po jinou dobu, C_i, takže dobu do projevu reakce, T_i, pozorujeme jen pokud $T_i \leq C_i$.

Statistická analýza takových dat tedy pracuje s informací, že některá data známe přesně (ta $T_i \leq C_i$), označme si tyto případy indikátorem $\delta_i = 1$, a o dalších datech víme, že jsou $T_i > C_i$. Tyto případy označme indikátorem $\delta_i = 0$. Dále označme $X_i = min(T_i, C_i)$, to jsou vlastně hodnoty, které "vidíme". Pak například, pokud předpokládáme, že náhodné veličiny (doby do poruchy) T_i mají rozdělení s hustotou $f(x;\theta)$ a funkcí spolehlivosti $S(x;\theta)$, s neznámým parametrem θ, tak věrohodnostní funkce má tvar

$$L(\theta; data) = \prod_{i=1}^{N} f(X_i;\theta)^{\delta_i} \cdot S(X_i;\theta)^{1-\delta_i}.$$

Její maximalizací, resp. jak je zvykem, maximalizací jejího logaritmu $\ln(L(\theta; data))$ přes θ, lze najít maximálně věrohodný odhad parametru θ.

Pro schema cenzorování se předpokládá, že veličiny C_i jsou nezávislé vzájemně i na T_i. Pokud tomu tak není, dostaneme složitější situaci tzv. **konkurenčních rizik** (competing risks). Ta je v statistické analýze přežití také řešena, zde se jí ale věnovat nebudeme.

Intervalové cenzorování vznikne, když se o stavu sledovaného objektu přesvědčujeme jen během občasných inspekcí. Pak neznáme přesný čas události, víme jen, v jakém intervalu k ní došlo.

Obecneji, zejména v oblasti mediciny, demografie, pojišťovnictví, podobné **životnostní tabulky** (Life tables) mohou obsahovat i počty jedinců "ztracených" z pozorování během daného intervalu (tedy cenzorovaných v původním smyslu), u nichž přesně nevíme, v které části intervalu bylo pozorování ukončeno. Z toho pak plyne nejednoznačnost následných odhadů intenzity.

Krajní případy jsou jestliže předpokládáme, že k jejich vyřazení došlo na počátku intervalu, nebo na konci. Skutečnost může být někde mezi tím, i proto se často výsledky pro tyto dva krajní případy průměrují.

O statistické analýze cenzorovaných dat existuje rozsáhlá literatura, samozřejmě se tímto tématem zabývají publikace o analýze přežití.

5 Neparametrické metody odhadu intenzity

Pokud se nechceme omezovat na určitou parametrizovanou rodinu distribucí, naším cílem je neparametrický způsob popisu i odhadu průběhu intenzity, resp. kumulovaných charakteristik.

Distribuci (náhodné veličiny T - doby do poruchy apod.) jsme charakterizovali čtyřmi vzájemně svázanými funkcemi. Dvě z nich mají kumulativní charakter, F (resp. S) a H - kumulativní intenzita, další dvě jsou lokální, f a h. Mějme k dispozici náhodný výběr $T_1, ..., T_n$. Když se konstruuje empirická distribučni funkce $F_n(t)$, 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 objektů ještě pozorovaných (tj. ještě v riziku poruchy) v okamžiku T_i. Nechť jsou data seřazena, $T_0 = 0 < T_1 \leq T_2 \leq ... \leq T_n$, pak $R_i = n - i + 1$. Kumulativni charakteristiku v t pak standardně odhadneme součtem těchto vah přes $T_i \leq t$. Empirickou funkci přežití si ale můžeme také představit jako součin: Pak $S_n(T_0) = 1$ a $S_n(T_i) = S_n(T_{i-1})(1 - 1/R_i) = \prod_{j=1}^{n} I[T_j \leq T_i](1 - 1/R_j)$. 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. C se spojitou distribuční funkcí G na $[0, \infty)$, s $Q = 1 - G$. Pak pozorujeme veličiny $X_i = \min(T_i, C_i)$, $\delta_i = I[(T_i \leq C_i]$. Toto je schéma náhodného cenzorování zprava. Když nyní každému momentu. ve kterém nastal sledovaný jev (tj. bodu $X_i, \delta_i = 1$) přiřadíme váhu $1/R_i$, pak pro kumulativní intenzitu poruch dostaneme odhad

$$H_n(t) = \sum_{i=1}^{n} I[X_i \leq t] \frac{\delta_i}{R_i}, \quad H_n(t) = O \; pro \; t < min\{X_i\}.$$

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

$$S_n(t) = \prod_{i=1}^{n} (1 - \frac{\delta_i}{R_i})^{I[X_i \leq t]}, \quad S_n(t) = 1 \; pro \; t < min\{X_i\},$$

což je známý "Product Limit Estimator" Kaplana a Meiera. Vlastnosti těchto odhadů jsou i nadále velice dobré, jsou popsány v mnoha článcích i monografiích. Označme $T_F = \sup\{t : F(t) < 1\}$, $T_C = \sup\{t : G(t) < l\}$, $T_M = \min\{T_F, T_C\}$.

1. $S_n(t)$ je silně konzistentním odhadem, stejnoměrně v $t \in (0, T_M)$.

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

2. Na $(0, T_*]$ je dokázána asymptotická normalita:

$\sqrt{n}(S_n(t)/S(t) - 1) \sim \sqrt{n}(H_n(t) - H(t)) \sim Z(t)$, kde $Z(t)$ je gaussovský náhodný proces s nulovou střední hodnotou a kovariancí pro $0 \leq s \leq t \leq T_*$

$$cov(Z(s), Z(t)) = var(Z(s)) = \int_0^s \frac{dF}{S^2 Q}.$$

Hned je vidět možnost transformace na Brownův proces, protože $Z(t)$ je proces s nezávislými přírůstky, a tedy je možné zkonstruovat pro $S(t)$, resp. pro $H(t)$, pásy spolehlivosti typu Kolmosorova-Smimova.

6 Náhodné bodové a čítací procesy

Náhodný bodový proces (v R_1) je obecně posloupnost náhodných "hodnot" $T_0 = 0 < T_1 < T_2 < \ldots$. Jak tyto hodnoty vznikají, to je dáno typem procesu. To, čemu se říká čítací proces, je pak proces s bodovým jednoznačně svázaný (lze je vlastně ztotožnit, bodový proces popsat jako čítací a naopak): $N(t) = \sum_{i=1}^{\infty} \mathbf{1}[T_i \leq t]$, tj. tento proces si připočte +1 v každém bodě odpovídajícího bodového procesu (a má trajektorie spojité zprava), při t rostoucím od 0 (nejčastěji jde skutečně o čas, ale v jiných aplikacích může jít o vzdálenost, rostoucí sílu působící na objekt apod.).

Z bodových procesů je nejznámější homogenní **Poissonův proces**, kdy $T_i = T_{i-1} + X_i$, X_i, $i = 1, 2, \ldots$, jsou i.i.d. náhodné veličiny s exponenciálním rozdělením (a $T_0 = 0$). Parametr λ exponenciálního rozdělení ($\lambda = 1/EX_i$) je intenzita procesu (zároveň je to intenzita rozdělení n.v. X_i dle definice $\lambda(x) = f(x)/(1 - F(x))$, když $f(x)$ je hustota a $F(x)$ je distribuční funkce). Počet "bodů" do času t má pak Poissonovo rozdělení s parametrem $\lambda \cdot t$.

Obecnější procesy jsou **nehomogenní Poissonův** (intenzita již není konstantní, ale je to stále deterministická funkce) a pak případy, kdy intenzita je náhodná: Tak **smíšený (mixed) Poissonův proces** je Poissonův proces s intenzitou - náhodnou veličinou. Rozdělením této náhodné veličiny (s nějakou distribuční funkcí G_0) vlastně mícháme (mixujeme) různé homogenní Poissonovy procesy. Pro každé $t > 0$ a celé $k \geq 0$ j e

$$P(N(t) = k) = \int_0^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} dG_0(\lambda).$$

Všimněme si, že to odpovídá i Bayesovskému pohledu na situaci s neznámým λ a jeho priorem $G_0(\lambda)$. Dále, střední hodnota počtu událostí do t je $E\ N(t) = t \cdot E\lambda$ a rozptyl tohoto procesu v t je $var(N(t)) = t^2 \cdot var(\lambda) + t \cdot E\lambda$.

Dvojitě stochastický (double stochastic) Poissonův proces, také zvaný Coxův proces, je nehomogenní variantou mixed procesu: Existuje náhodný proces $\lambda(t)$ tak, že pro každá $t > 0$ a celé $k \geq 0$ je

$$P(N(t) = k) = E\{\frac{(\int_0^t \lambda(s)ds)^k}{k!} exp(-\int_0^t \lambda(s)ds)\}.$$

Ta "dvojitá stochastičnost" znamená, že tu působí dva náhodné mechanizmy $\lambda(t)$ a $N(t)$, "nezávislá" realizace $\lambda(t)$ je intenzitou pro $N(t)$. Opět můžeme z každé realizace procesu $N_j(t)$, $j = 1, ..., n$, odhadnout zhruba , jak vypadala jeho kumulovaná intenzita, $\Lambda_j(t) = \int_0^t \lambda_j(s)ds \sim N_j(t)$ pro každé t v $[0, T]$. Tím získáme představu o rozdělení náhodné intenzity $\lambda(t)$.

Čítací proces (counting process) je pak náhodný proces, jehož body přírůstků +1 jsou řízeny náhodnou intenzitou, která je "predikovatelná a adaptovaná" – tj. měřitelná vzhledem k (a tedy závislá na) neklesající posloupnosti σ-algeber, filtraci $\mathcal{F}_{t-}$, tj. chápanou jako spojitou zleva. Jde tedy o zobecnění dvojitě stochastického Poissonova procesu. Predikovatelnost zde znamená, že proces je pozorovatelný a hodnota v t je odvoditelná z hodnot těsně před t (takže například spojitost trajektorií zleva dostačuje). Základní vztah mezi pravděpodobností výskytu bodu v malém časovém intervalu Δt a momentální intenzitou $\lambda(t)$ je nyní podmíněný:

$$P\left(N(t + \Delta t) - N(t) = 1 | \mathcal{F}_{t-}\right) = \lambda(t)\, \Delta t + o(\Delta t).$$

Interpretace je většinou taková, že $\mathcal{F}_{t-}$ obsahuje informaci o událostech, které se staly před t, tj. **historii** procesu, která má vliv na pravděpodobnost skoku $N(t)$ v t a v nejbližší budoucnosti, tj. na "inovaci" procesu. Význam tohoto matematického aparátu je, že za těchto okolností je kumulovaná intenzita $A(t) = \int_0^t \lambda(s)\, \mathrm{d}s$ kompenzátorem čítacího procesu, tj. proces $M(t) = N(t) - A(t)$ je martingal (adaptovaný na σ algebry $\mathcal{F}_{t+}$ obsahující historii i inovaci). Navíc, dík

tomu, že v každém $[t, t+\Delta t)$ je $\Delta N(t)$ "nulajedničková" náhodná veličina (přesněji, s $P \to 1$ při $\Delta t \to 0$), tak varianční proces $\langle M \rangle(t) = A(t)$. Podrobněji je toto vše probráno v mnoha článcích i knihách (Andersen et al, 1993, Fleming a Harrington, 1991). Často se ještě přidává proces – indikátor, $I(t) = 1$ když je proces (objekt) pozorován v čase t, $I(t) = 0$ jinak. Tento proces může tedy indikovat i různé typy cenzorování.

Nyní je vhodné i rozlišovat mezi pojmy: intenzita $\lambda(t)$ je náhodná funkce, závislá na historii procesu do času t, riziková funkce $h(t)$ je nenáhodná modelová funkce. V nejjednodušším případě je prostě $\lambda(t) = h(t) \cdot I(t)$, tedy $\lambda(t) = 0$ pokud již proces není pozorován (je cenzorován či ukončenj poruchou).

Ukažme si, jak v tomto značení vypadá Nelson-Aalenův odhad kumulované rizikové funkce $H(t) = \int_0^t h(s)ds$. Uvažujme, že pozorujeme m procesů, pro $t \in [0, T]$, s událostmi v časech $\{T_{ij}, j = 1, ..., m, i = 1, ..., n_j$, tj. n_j událostí bylo pozorováno pro j-tý proces). Věrohodnostní funkce (která je nyní také náhodná) je

$$L = \prod_{j=1}^{m} \{ \prod_{i=1}^{n_j} \lambda_j(T_{ij}) \cdot exp(- \int_0^T \lambda_j(s)ds) \}.$$

Logaritmus je pak (už použijeme nové značení)

$$\mathcal{L} = \ln(L) = \sum_{j=1}^{m} \{ \int_0^T \ln\, h(t) dN_j(t) - \int_0^T h(s) I_j(s) ds \}.$$

Z toho je odhad odvozen maximalizací,

$$\hat{H}(t) = \int_0^t \sum_{j=1}^{m} \frac{\mathrm{d}N_j(t)}{\sum_{k=1}^{m} I_k(t)},$$

kde znovu $I_k(t)$ jsou indikátory, zda je k-tý objekt (proces) v t pozorován. Pokud chceme z odhadu $H(t)$ dostat odhad pro její derivaci $h(t)$, můžeme to provést tradičním způsobem – jádrovým vyrovnáním přírůstků $dH(t)$:

$$\hat{h(t)} = \frac{1}{d} \int_0^T W(\frac{t-s}{d}) d\hat{H}(s),$$

kde $W(.)$ je jádrová funkce.

7 Martingalové reziduály

V teorii pro čítací procesy je mnoho výsledků založeno na již výše zmíněném rozdkladu procesu na martingal a kompenzátor, tj. $N_i(t) = A_i(t) + M_i(t)$ pro individuální procesy, $i = 1, 2, ..., n$, $N(t) = A(t) + M(t)$ pro jejich součty přes i. Zde $A_i(t)$, $A(t)$ značí kumulované intenzity (chápané jako náhodné procesy), $M_i(t)$, $M(t)$ jsou martingaly s nulovou střední hodnotou, s nekorelovanými přírůstky, s rozptylovým procesem $\langle M_i \rangle(t) = A_i(t)$, $\langle M \rangle(t) = A(t)$, $M_i(t)$ jsou také nekorelované vzájemně pro různá i. Tento rozklad ilustruje Obr. 5.

Pak je zcela přirozené jako proces charakterizující kvalitu (dobrý fit) odhadu uvažovat proces tzv. martingalových reziduálů

$$R(t) = N(t) - \hat{A}(t) = M(t) + A(t) - \hat{A}(t),$$

kde $\hat{A}(t)$ je odhadnutá kumulovaná intenzita. V nejjednodušším případě je $\hat{A}(t) = \int_0^t I(s) d\hat{H}(s)$, kde $\hat{H}(s)$ je Nelsonův-Aalenův odhad kumulované rizikové funkce. Vlastnosti reziduálů tedy závisejí na vlastnostech odhadu. Testy dobré shody modelu s daty jsou prováděny buď graficky nebo i numericky, kriteria pro zamítnutí - nezamítnutí dobré shody jsou odvozena právě od (asymptotických) vlastností odhadu.

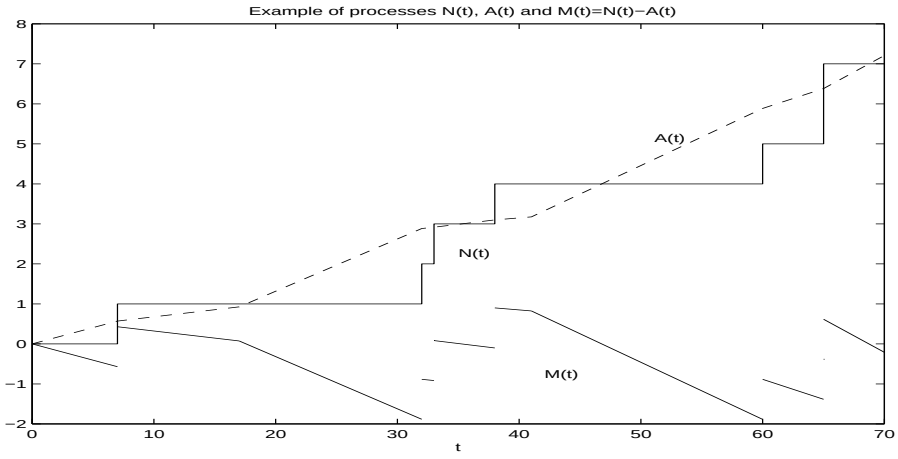

Figure 5: Příklad realizace čítacího procesu $N(t)$, intenzity $A(t)$ a $M(t) = N(t) - A(t)$, dle Andersen et al (1993).

Vlastnosti Nelsonova-Aalenova odhadu již byly popsány výše. Protože po dosazení odhadu dostaneme, že přímo $\hat{A}(t) = N(t)$, konstruují se reziduály zvlášť pro různé podskupiny dat (objektů), $S \subset \{1, .., n\}$. Označme tedy

$$R_S(t) = N_S(t) - \hat{A}_S(t) = M_S(t) + A_S(t) - \hat{A}_S(t),$$

kde je $N(t) = \sum_{i=1}^{n} N_i(t)$, $N_S(t) = \sum_{i \in S} N_i(t)$, obdobně pro $I(t)$, $M(t)$, $A(t)$, $\hat{A}(t)$. Protože

$$\hat{A}_S(t) = \int_0^t \sum_{i \in S} d\hat{H}(r) I_i(r) = \int_0^t \frac{dN(r)}{I(r)} \cdot I_S(r) =$$

$$= \int_0^t \frac{dH(r)I(r) + dM(r)}{I(r)} \cdot I_S(r) = A_S(t) + \int_0^t \frac{dM(r)}{I(r)} \cdot I_S(r),$$

dostaneme (při označení $\bar{S}$ = doplněk S)

$$R_S(t) = M_S(t) - \int_0^t \frac{dM(r)}{I(r)} \cdot I_S(r) = \int_0^t \frac{dM_S(r) I_{\bar{S}}(r) - dM_{\bar{S}}(r) I_S(r)}{I(r)}.$$

Z toho je vidět, že proces $R_S(t)$ je opět martingal, je možné i spočítat jeho rozptylový proces a odvodit kritické oblasti, ve kterých by se měl pohybovat s vysokou pravděpodobností, pokud skutečně zvolený tvar $H(t)$ odpovídá datům. Dokonce můžeme i pro různé podskupiny zjišťovat, kde případně zvolený model není dobrý.

V složitějších modelech, např. při předpokladu Coxova regresního modelu (viz dále) již proces reziduálů není martingal. Pak se k určení kritických mezí používá náhodné generování dat z hypotetického modelu. To je popsáno třeba v článku Volf, Timková (2014).

A vždy lze použít alespoň grafické testy. Ty jsou dělány buď tak, že na osu X vynášíme počty událostí v čítacím procesu $N_S(t)$, tj. 1,2,3,..., na osu Y pak odhad $\hat{A}_S(t)$. Ten by se měl pohybovat kolem diagonály (jako bychom přetransformovali čítací proces na Poissonův proces s intenzitou 1). Nebo na osu Y vynášíme přímo reziduály $R_S(t) = N_S(t) - \hat{A}_S(t)$, ty by se tedy měly pohybovat kolem nuly. Použití bude také později ukázáno.

8 Regresní modely v analýze přežití

Regresní model popisuje závislost sledované veličiny na nějakých dalších veličinách, kovariátách, regresorech. V analýze přežití je opět tato závislost často zadána formou závislosti rizikové funkce nejen tedy na čase t, ale i na oněch kovariátách, označme je z. Samozřejmě se používají i

standardní regresní modely, například lineární, ale bylo navrženo několik typů modelů speciálně pro tuto oblast. Na začátku jsme zmínili některá typická rozdělení pravděpodobnosti užívaná v analýze přežití, jako exponenciální, Weibullovo, lognormální, gamma. Například riziková funkce Weibullova rozdělení je $h_W(t) = \alpha\,\beta\,t^{\beta-1}$, s parametry α, β. Pokud některý z nich ještě závisí na kovariátách, model popisuje závislost doby přežití či rizikové funkce. Pokud je závislost ještě specifikována pomocí dalších parametrů, jde o "parametrický" model. Pokud model obsahuje jak parametrizovanou tak neparametrickou část, jde o model "semiparametrický", jedním příkladem je následující Coxův regresní model pro rizikovou funkci.

8.1 'Proportional hazard' či Coxův model

Heterogenita (nesourodost) zkoumaného souboru je v modelu pro intenzitu často vyjádřena součinem více členů. Pokud je ona nesourodost způsobena vlivem kovariáty Z, pak intenzitu rozdělení pravděpodobnosti n.v. T při $Z = z$ modelujeme jako $h(t, z) = ho(t) \cdot B(z)$, samozřejmě Z může být vícerozměrná. Tomuto multiplikativnímu modelu se také říká "proportional hazard regression model", $ho(t)$ je základní, tzv. "baseline" intenzita. Jako data budeme uvažovat realizace dvojic $(T_i, Z_i), i = 1, ..., n$, případně trojic (X_i, δ_i, Z_i) při cenzorování. Předpokládáme, že při daných hodnotách $Z_i = z_i$ už jsou n. v. T_i nezávislé vzájemně i s C_i cenzorujícími. Zároveň se předpokládá, že cenzorující veličina nezávisí na regresní funkci $B(z)$.

Základní vlastností, která dala modelu jméno, je ta, že pro dvě různé hodnoty kovariáty $z_1 \neq z_2$ poměr $h(t, z_1)/h(t, z_2) = B(z_1)/B(z_2)$ pro každé t. Pokud je tato vlastnost splněna, model umožňuje metodu analýzy přežívání zvanou "accelerated testing", neboť je i v této situaci dobře odhadnutelný. Například při testování spolehlivosti, pokud kovariáta charakterizuje zátěž, můžeme v experimentu použít zátěž větší než je zátěž běžná ve skutečném provozu, a tím zkrátit čas potřebný pro doběhnutí většiny testů, čili podstatně zmenšit míru cenzorování.

Původní a nejčastěji užívaný tvar **Coxova modelu** je jeho standardní "semiparametrická" verze, ve které je $B(z) = \exp(\beta \cdot z)$ log-lineární funkce s parametrem β stejného rozměru jako Z, zatímco 'baseline' riziková funkce $h_0(t)$ je neparametrizovaná. Označíme $H_0(t) = \int_0^t h_0(s)\,\mathrm{d}s$ kumulovanou základní intenzitu, její odhad (srovnej s odhadem Nelsona-Aalena), tzv. **Breslow - Crowley odhad** je

$$\hat{H}_0(t) = \int_0^t \sum_{j=1}^{n} \frac{\mathrm{d}N_j(t)}{\sum_{k=1}^{m} I_k(t)\, e^{\beta Z_k}}$$

Dále, β se odhaduje z tzv. **částečné (parciální) věrohodnostní funkce**, resp. maximalizací jejího logaritmu

$$\mathcal{L}^P = \sum_{j=1}^{n} \int_0^T \ln \left\{ \frac{e^{\beta Z_j}}{\sum_{k=1}^{n} e^{\beta Z_k} I_k(t)} \right\} \mathrm{d}N_j(t),$$

což vše jsou vlastně jen součty hodnot v bodech skoků $N_j(t)$. V případě standardního Coxova modelu to vede na úlohu odhadu parametrů β metodou maxima věrohodnosti. Platí tu i další vlastnost MVO, a to sice že odhady jsou asymptoticky normální a jejich limitní rozptyl lze odhadnout z druhých derivací $\mathcal{L}^P$ podle β, viz například Andersen et al (1993), Kalbfleish and Prentice (2002). Tento model může obsahovat i kovariáty měnící se v čase, podstata odhadu i jeho vlastnosti se zásadně nezmění.

Příklad 3

Na simulovaných datech ukážeme onu proporcionalitu rizik. Měla by se projevit rovnoběžností (zhruba) logaritmů odhadů kumulované intenzity pro různé skupiny hodnot kovariát. Tuto vlastnost lze použít jako grafický test vhodnosti Coxova modelu.

Vygenerovali jsme 100 hodnot z Coxova modelu s parametrem $\beta = 1$ a s základní intenzitou odpovídající lognormálnímu rozdělení s parametry $\mu = 1$, $\sigma = 0.5$. Hodnoty kovaríáty byly rozděleny rovnoměrně v (0,2), cenzorující hodnoty byly generovány z rovnoměrného rozdělení

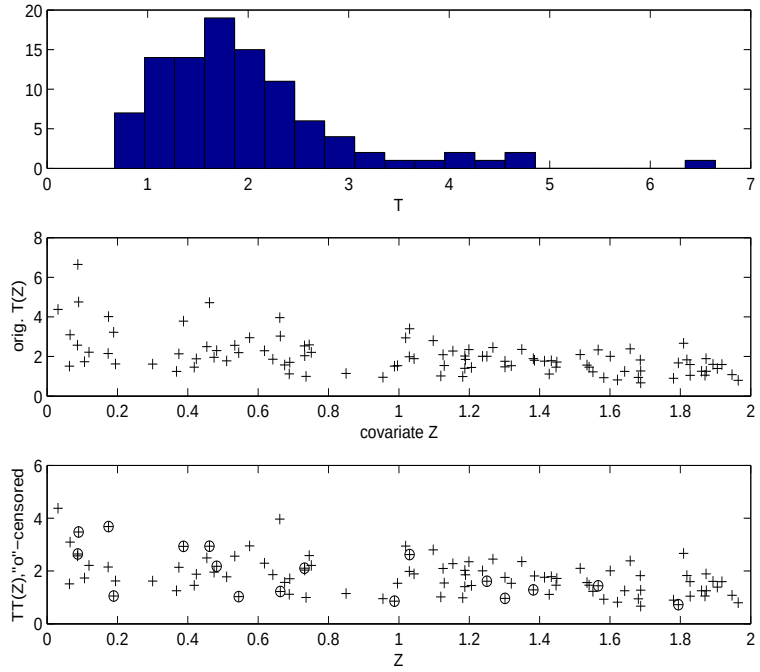

Figure 6: Data generovaná dle Coxova modelu, o=cenzorovaná data.

na intervalu (1,7), což vedlo k asi 20% cenzorování. Obr. 6 ukazuje data, nejprve vygenerovaná data bez cenzorování (Obr. 6 a,b), dolní obrázek pak data po cenzorování. Data jsme rozdělili do dvou skupin, s kovariátou $Z < 1$ a s $Z > 1$, v obou skupinách jsme odhadli kumulovanou rizikovou funkci Nalson-Aalenovým odhadem. Logaritmy (přirozené) těchto odhadů jsou na Obr. 7. Při vlastnosti proporcionálních rizikových funkcí by měly být zhruba rovnoběžné, což je zřejmě splněno.

8.2 Aditivní regresní model

V aditivním (také Aalenově) modelu je riziková funkce zadána jako $h(t, z) = z' \cdot \beta(t)$, kde z je kovariáta, $\beta(t)$ jsou funkce času, oboje, z i β mohou být mnohorozměrné (označme jejich dimenzi K). Samozřejmě musí i být $h(t, z) \geq 0$. Zpravidla $\beta_1(t)$ má význam jakési základní rizikové funkce, v tom případě první komponenta kovariáty je rovna jedné. Nechť data opět tvoří pozorování n objektů, indexujme je i, $i = 1, .., n$. Kovariáty $Z_i(t)$ se mohou měnit v čase. Individuální intenzita čítacího procesu $N_i(t)$ je pak

$$\lambda_i(t) = Z_i(t)' \cdot \beta(t) \cdot I_i(t), \;\; i = 1, \ldots, n.$$

Kumulované funkce $B(t)$, s komponentami $B_k(t) = \int_0^t \beta_k(s)\, ds$, pro $k = 1, ..., K$, se odhadují metodou vážených nejmenších čtverců, a to následovně: Jelikož opět platí rozklad čítacího procesu na intenzitu a martingal, $dN_i(t) = X_i(t)' dB(t) + dM_i(t)$, kde $X_i(t) = Z_i(t) \cdot I_i(t)$, pak

$$\hat{B}(t) = \int_0^t (X(r)'W(r)X(r))^{-1} X(r)'W(r)dN(r),$$

kde $W(r)$ je matice vah; její nejjednodušší verze je identická matice, $W(r) = I_n$, optimální váhy jsou $W(r) = diag\{1/\lambda_i(r)\}$. V praxi se pak použijí odhady $\hat{\lambda}_i(r)$, výpočet se iteruje.

Lze dokázat jak konzistenci tak asymptotickou normalitu $\hat{B}(t)$, platí, že

$$\sqrt{n}(\hat{B}(t) - B(t)) = \sqrt{n} \int_0^t \overline{X}(r) dM(r),$$

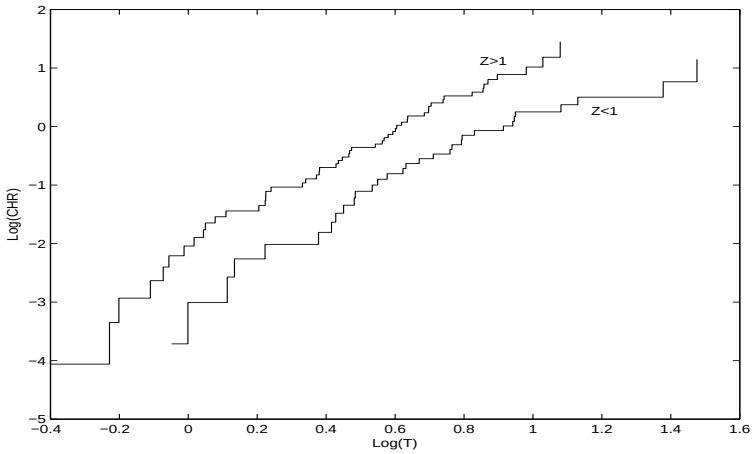

Figure 7: Grafický test rovnoběžnosti logaritmu kumulovaných rizikových funkcí.

je asymptoticky rozděleno jako Gaussův proces s nezávislými přírůstky.
Zde $\overline{X}(r) = (X(r)'W(r)X(r))^{-1}X(r)'W(r)$. Varianční funkce tohoto procesu je odhadnutelná empirickou verzí výrazu

$$n\int_0^t \overline{X}(s)\, D(s, B(s))\, \overline{X}'\, ds,$$

kde $D(s, B(s))$ je diagonální matice s komponentami $\lambda_i(s)$.

8.3 Model se zrychleným časem

Tento model ("accelerated failure time", AFT model) je často brán jako alternativa k Coxovu modelu, když zjevně proporcionalita rizik neplatí. Model předpokládá, že subjektivní čas (např. stárnutí) běží pro každý objekt v závislosti právě na jeho kovariátách. Obecně, předpokládá se, že objekt při hodnotě kovariát z má čas přežití T zadán distribuční funkcí $F(t) = F_0(t{\cdot}\exp(g(z)))$, kde F_0 je opět nějaká základní distribuční funkce, odpovídající náhodné veličině T_0 s kovariátami z_0 takovými, že $g(z_0) = 0$. Funkci $g(z)$ zde tedy můžeme nazvat regresní funkcí, nejčastější verze předpokládá prostou log-lineární závislost $g(z;\beta) = \beta \cdot z$. Logaritmováním dostaneme

$$\ln T = -g(z) + \ln T_0,$$

což odpovídá transformaci času $t \to t \cdot exp(g(z))$. Jenže obecně rozdělení T_0 neznáme, v tom případě jde znovu o semiparametrický model. Další problém způsobí přítomnost cenzorovaných dat, k cíli (odhadu jak regresní funkce tak rozdělení $\log T_0$) vede iterační metoda, v podstatě verze tzv. EM algoritmu (James, Smith, 1984). Proto je i zde často preferován přístup založený na rizikové funkci: Označme opět $\{X_i, z_i, \delta_i, i = 1, .., n\}$ pozorované časy (poruch či cenzorování), hodnoty kovariát, indikátor cenzorování pro i-tý objekt, pak věrohodnostní funkce je

$$L = \prod_{i=1}^{n} h_i(X_i)^{\delta_i} \cdot \exp(-\int_0^{X_i} h_i(t)dt),$$

kde $h_i(t) = h_0(t \cdot \exp(g(z_i))) \cdot \exp(g(z_i))$ je hodnota rizikové funkce pro i-tý objekt v čase t a h_0 je znovu základní riziková funkce, odpovídající veličině T_0. A opět je ukázáno, že jak regresní funkci (případně její parametry, pokud je parametrizovaná), tak základní rizikovou funkci lze odhadnout konzistentně, a že odhady mají vlastnost asymptotické normality (což je důležité pro testování vhodnosti modelu). Viz třeba Bagdonavicius and Nikulin (2002, Ch. 6). Pro semiparametrický model je dokonce dokázáno (Lin et al, 1998), že je možné, vhodnou aproximací skórové funkce (tj derivace log-věrohodnosti podle odhadovaného parametru) konzistentně odhadnout parametr regresní funkce bez znalosti (odhadu) základní intenzity, tj. odhadování

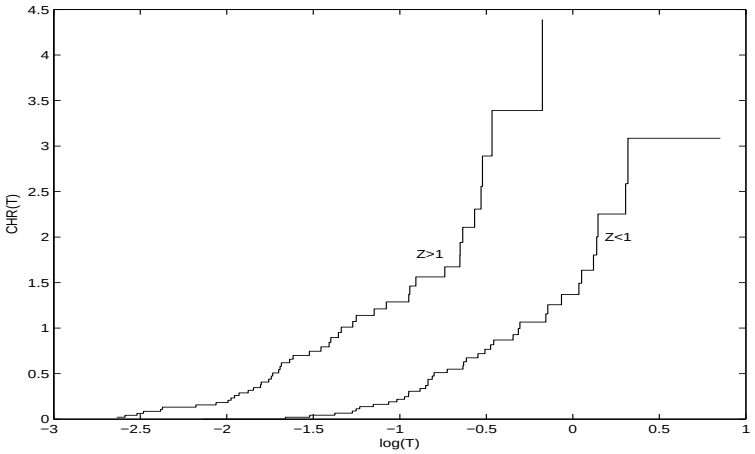

Figure 8: Grafický test posunutí kumulovaných rizikových funkcí.

rozdělit do dvou kroků podobně jako v případě Coxova modelu. I AFT model může být rozšířen na případ, kdy se kovariáty mění v čase.

Kromě testů s využitím reziduálů (jejichž grafická verze je snadná, ale protože reziduály v tomto případě také už nejsou martingaly, numericky se zkoumají pomocí Monte Carlo metod), je možné použít i další jednoduchý grafický test. Protože kumulovaná riziková funkce je $H(t,z) = H_0(\exp(log(t) + g(z)))$, tak je jako funkce logaritmu času pro různé hodnoty kovariát prostě jen posunutá. Takže můžeme rozdělit data do několika skupin podle hodnot kovariát a graficky porovnat Nelsonovy-Aalenovy odhady kumulované rizikové funkce v těchto skupinách, jako v následujícím Příkladě 4. Přehled a základní vlastnosti zde uvedených regresních modelů a testů lze nalézt i v Novák (2008).

Příklad 4

Opět jsme vygenerovali 100 hodnot, tentokrát odpovídající semiparametrickému AFT modelu s parametrem $\beta = 1$, základní rozdělení veličiny T_0 odpovídalo log-normálnímu rozdělení s parametry $\mu = 0$, $\sigma = 0.5$. Kovariáta byla rovnoměrně rozdělena v (0,2), hodnoty $\ln T$ byly cenzorovány náhodně zprava veličinou $\ln C$ s rovnoměrným rozdělením na (-3,1), tj. tak aby opět bylo zhruba 20% dat cenzorovaných. Obr. 8 ukazuje, že skutečně kumulované rizikové funkce odhadnuté Nelson-Aalenovým odhadem zvlášť z dat s $Z < 1$ a $Z > 1$ jsou při logaritmickém měřítku na ose času vzájemně posunuté.

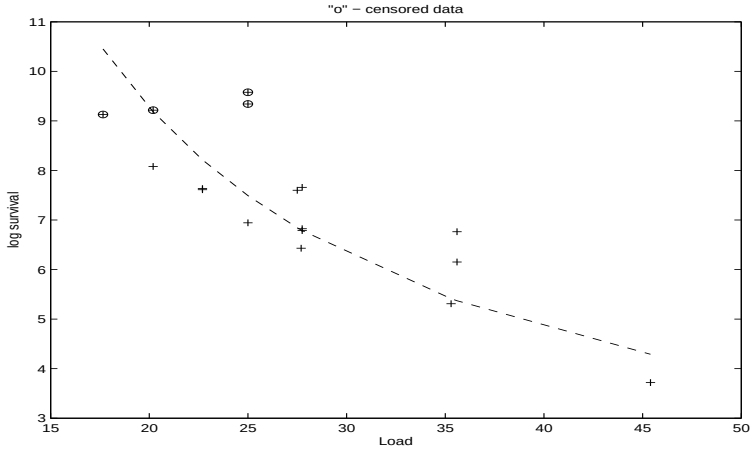

Figure 9: Data a model Příkladu 5, censorovaná data jsou označena 'o'.

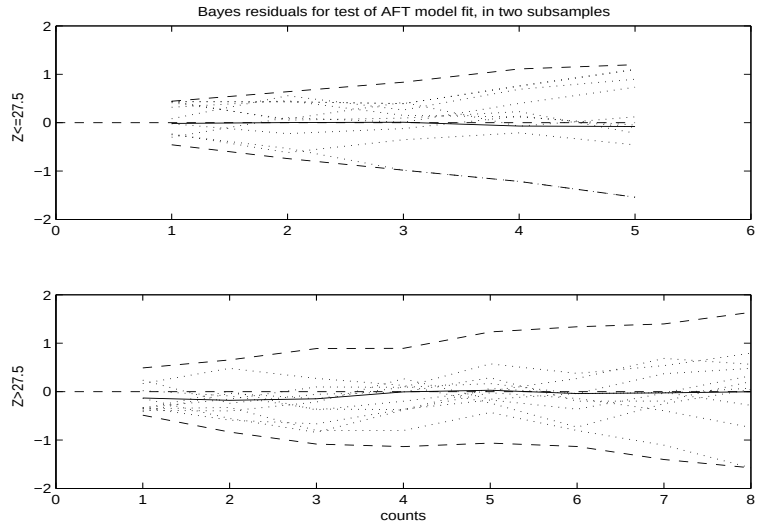

Figure 10: Test pomocí martingalových reziduálů ve 2 skupinách s $Z \leq 27.5$ and $Z > 27.5$.

8.4 Příklad 5

Data v Tabulce 1 pocházejí z před mnoha lety prováděného testování výdrže ocelových pružin z podvozku nákladních vozů Tatra. Jde o pouhých 17 měření, neb zkoušky byly náročné časově i finančně. Součástky byly v laboratoři podrobeny cyklům vysokofrekvenčních vibrací s různou zátěží, která zůstala stálá během každého experimentu. Zátěž je dána v Kp/cm^2, výdrž v cyklech vibrací do defektu, indikátor cenzorování je 0 když byl experiment ukončen před zaznamenáním defektu.

Zátěž	Výdrž	Cenz.	Zátěž	Výdrž	Cenz.	Zátěž	Výdrž	Cenz.
45.4	41.2	1	35.6	470.1	1	35.6	865.8	1
35.3	202.4	1	27.7	620.0	1	27.75	884.9	1
27.75	919.3	1	27.75	2119.9	1	27.5	1998.9	1
25.0	1036.2	1	25	11390.0	0	25	14443.9	0
22.7	2020.0	1	22.7	2065.9	1	20.2	3231.4	1
20.2	10064.8	0	17.66	9219.0	0			

Table 1. Data: Zátěž v Kp/cm^2, výdrž v cyklech vibrací, indikátor cenzorování.

Data jsou také zobrazena na Obr. 9, v logaritmické stupnici pro výdrž. Uvažovali jsme AFT model, před log-lineárním trendem jsme na základě dat dali přednost log-hyperbolickému, variantě tzv. Arrheniova modelu s tvarem

$$\ln T = -\beta \cdot (C - 1/Z) + \ln T_0.$$

Zde Z je zátěž, $\beta > 0$ neznámý parametr, konstantu jsme zvolili $C = 0.1$. To znamená, že T_0 odpovídá zátěži $Z = 10\, Kp/cm^2$. To je zde samozřejmě jen jakási referenční hodnota. Takovýto model umožňuje také metodu zrychleného testování (tj. se zvětšenou zátěží), ale samozřejmě je vhodný jen v určité oblasti zátěže. A přirozeně, při skutečném provozu se zátěž i vibrace mění.

Řešení bylo založeno na Monte Carlo metodě v rámci Bayesova přístupu, tj. na generování možných řešení a jejich optimalizaci v rámci tzv. metody MCMC (viz Volf a Timková, 2014). Cílem bylo získat reprezentaci aposteriorního rozdělení pro parametr β: mělo průměr 178.34, medián 178.12 a směrodatnou odchylku 0.70. Obr. 9 ukazuje i výslednou mediánovou křivku $\text{med}\,\{-\beta \cdot (C - 1/Z_i) + \ln T_0\}$, s dosazenou odhadnutou (vygenerovanou v průběhu řešení) reprezentací β a T_0.

Grafický test dobré shody s AFT modelem je na Obr. 10. Je vidět, že skutečně distribuce reziduálů vygenerovaných z modelu se drží kolem nuly, pro obě skupiny dat rozdělených podle hodnoty kovariáty.

Literatura

Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical Models Based on Counting Processes. New York, Springer, 1993.

Arjas E. (1988): A graphical method for assesing goodnes of fit in Cox's proportional hazard model. Journal of the Amer. Statist. Association 83, 204–212.

Aven T. and Jensen U. (1999). Stochastic Models in Reliability. New York, Springer.

Bagdonavicius, V., Nikulin, M.: Accelerated Life Models, Modeling and Statistical Analysis. Boca Raton, Chapman&Hall/CRC, 2002.

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

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

Hoyland, A., Rausand, M.: System Reliability Theory: Models and Statistical Methods. New York, Wiley, 1994.

James, I.R., Smith, P.J.: Consistency results for linear regression with censored data. Annals Statist. 12, 1984, 590-600.

Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. New York, Wiley, 2002.

Lin, D.Y., WeiL, J., Ying, Z. Accelerated failure time models for counting processes. Biometrika 1998;85:60518.

Novák, P. Regresní modely pro intenzity poruch v analýze spolehlivosti. Dipl. práce KPMS MFF UK v Praze, 2009.

Singpurvala N. D. (1995). Survival in dynamic environments. Statist. Science 10, 86–103.

Snyder D. L. (1975). Random Point Processes. Wiley, N.Y.

Volf P. (2004). An application of nonparametric Cox regression model in reliability analysis: A case study, Kybernetika vol.40, 5 (2004), 639-648.

Volf P., Timková J. (2014). On selection of optimal stochastic model for accelerated life testing , Reliability Engineering & System Safety vol.131, 1 (2014), 291-297.

Wikipedia, The Free Encyclopedia.
Retrieved on: en.wikipedia.org , cs.wikipedia.org