Vera Lanská, IKEM, Praha


1. ÚVOD


2. PARAMETRIČKÉ MODELY

Uvažujeme nezápornou náhodnou veličinu T, která popisuje dobu do sahání. Funkci přežití rozumíme S(t)=PT(t). Budeme předpokládat, že existuje hustota rozložení f(t) = -S'(t). Intenzita sahání (risiková funkce) je definována vztahem

\[ \lambda(t) = \lim_{h \to 0} \frac{P(T \leq t+h | T \leq t)}{h} = \frac{f(t)}{S(t)}. \]

Neži funkci přežití a intenzitu sahání platí

\[ S(t) = \exp(-\int_0^t \lambda(s) \, ds). \]

Intenzita sahání popisuje okamžitě riziko sahání a je velmi často používána v analýze přežívání.

Nyní uveďeme některé pravděpodobnostní rozložení vhodné pro doby do sahání:
(i) exponenciální rozložení, \( \lambda(t) = \lambda \)
(ii) Gamma rozložení, intenzita má kombinované vyjádření s 2 parametry
(iii) Weibullovo rozložení, \( \lambda(t) = \lambda a^k t^{k-1} \)
(iv) log normální rozložení, intenzita není monotónní a má dva parametry
(v) Inversní Gaussovo rozložení, dvouparametrické, není vhodné pro cenzory
Při odhadování neznámých parametrů v parametrických modelech se najízdelší užívá metoda maximální věrnostnosti. Nechť f(x; θ) je hustota rozložení doby do selhání. Vektor neznámých parametrů uvažujeme ve tvaru \( \theta = (\omega, \lambda) \), kde vektor parametrů u nás zajišťuje a vektor \( \lambda \) obsahuje rušivé parametry. Jediněc, u kterého dojde k selhání v okamžiku \( t \) přispívá k funkcii věrnosti členům \( f(t; \theta) \) a jedinec, který je cenzorován v okamžiku s přispívá členům \( S_i(t; \theta) \). Funkce věrnosti pro n nezávislých jedinců s indexem i má tvar

\[
L(\theta) = \prod_{i=1}^{n} f(t_i; \theta) \prod_{i=1}^{n} S_i(t_i; \theta).
\]

Tedy pro logaritmus věrnostní fungce platí vztah

\[
\ln L(\theta) = \sum_{\text{nečenzor}} \ln f(t_i; \theta) + \sum_{\text{čenzor}} \ln S_i(t_i; \theta).
\]

Oznáme-li \( X_i = \min(t_i, q_i, \Lambda_i) \), pak dostaneme

\[
\ln L(\theta) = \sum_{\text{nečenzor}} \ln f(x_i; \theta) + \sum_{i=1}^{n} \ln S_i(t_i; \theta) = \sum_{\text{nečenzor}} \ln f(x_i; \theta) - \sum_{i=1}^{n} \Lambda_i(t_i; \theta).
\]

Tento vztah pro logaritmus věrnostní funkce ukazuje jednu z výhod zavedení integrační zatemnění. Odnadně neznámých parametrů maximizují \( L(\theta) \). Testování hypotéza o neznámých parametrech je založeno na asymptotických vlastnostech maximálně věrohodných odhadů. Uvedeme tři základní statistiky pro testování hypotéz \( H: \omega = \omega_0 \).

1) Věrohodnostní poměr \( W(\psi) = 2 \{ \ln \psi - \ln \psi(\psi) \} \)

2) Waldova statistika \( W(\psi) = (\hat{\psi} - \psi_0)(\hat{\psi} - \psi_0) \psi_0^{-1} \psi_0^{-1} \)

3) skórová funkce \( W(\psi) = V(\psi) - V(\psi_0) \psi_0^{-1} \psi_0^{-1} \)

Všechny tato statistiky mají za platnost hypotézy \( \chi^2 \) rozložení pravděpodobnosti se stupněmi volnosti, které se rovnají \( \nu \). Podrobná odvození lze najít např. v (11), (21), (25), (77).

3. NEPARAMETRICKÉ METODY

V předchozí části jsme předpokládali, že známe tři řady pravděpodobnostních rozložení, kterou se řídal doba do selhání. Nyní uvedeme dva základní neparametrické odhady pravděpodobnosti předčástí. Klasické metody, užívané hlavně v epidemiologii a v pojišťovací matematici, neznačí "life-table". Sledovaný časový úsek se rozdelí na pevné podintervaly \( t_1 = (0, \tau_1), ..., t_k = (\tau_k, \tau_{k+1}) \) (např. 0-6, 6-12, 12-24, 24 a více). Oznáme si

- \( n_i \) = počet živých jedinců na počátku \( t_i \),
- \( d_i \) = počet selhání během \( t_i \),
- \( l_i = d_i \) ztracených během \( t_i \),
- \( m_i = d_i \) ukončení během \( t_i \),
- \( p_i = P(\text{vzrostl} \text{ na počátku } t_{i+1} \text{ je naživu}) \),
- \( w_i = l_i \) cenzorován pozorování v intervalu \( t_i \).

Pro pravděpodobnost předčástí platí

\[
S(t) = P(T > t) = \prod_{i=1}^{k} p_i.
\]
Metodou ‘life-table’ odhadneme \( p_i \) pomocí \( \hat{\beta}_j = 1 - d_j/n_j' \), kde \( n_j' = n_j - 1/2 (w_j + 1) \) se nazývá efektivní rozsaň a odhad pravděpodobnosti přežití je pak
\[
\hat{S}(\tau)_k = \prod_{i=1}^{k} \hat{\beta}_i.
\]
Rozpůlňad je dáno Greenwoodovou formulí
\[
\text{var}(\hat{S}(\tau)_k) = (\hat{S}(\tau)_k)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j'-d_j)}.
\]
Tento odhad se nehodí pro malé výběry, kdy značnou volblo intervallů můžeme dostat velmi odlišné výsledky.

Pro přesnější odhad funkce přežití se užívá Kaplan-Meierov odhad (nazývaný též product-limit). Odpovídá předchozímu, kde za \( \tau_j \) bereme i-té uspořádané pozorování. Tedy předpokládáme, že pozorujeme dvojice \((Y_i, \delta_i)\), \((Y_i, \delta_i)\), kde \( \delta_i \) označuje indikátor selhání a platí \( Y_i (\delta_i = 1) < \ldots < Y_i (\delta_i = 0) \) (nenastávaly shody). Pak odhad funkce přežití má tvar
\[
\hat{S}(t) = \prod_{Y_i \leq t} (1 - \frac{1}{n_i+1} \delta_i).
\]
Pokud \( \delta_i = 0 \) je třeba dedefinovat \( \hat{S}(t) = 0 \) pro \( t < \min(\tau_i) \).

Pro srovnání dvou a více výběrů existuje mnoho statistik. Pro jednoduchost budeme uvažovat pouze dva výběry a uvedeme dva základní testy. Nechť pro první výběr \( T_1, \ldots, T_n \) dovedně jsou i.i.d. s distribuční funkcí \( F_1 \); \( C_1, \ldots, C_n \) jsou i.i.d. s distribuční funkcí \( C_1 \) (ním příslušné okamžiky cenzorování). Pozorujeme dvojice \((X_i, \delta_i)\), \((X_i, \delta_i)\), kde \( X_i = T_i \) a \( \delta_i = 1 \) je \( T_i \leq C_i \).

Pro druhý výběr máme \( U_1, \ldots, U_n \) s \( F_2 \) a \( D_1, \ldots, D_n \) s \( G_2 \) a pozorujeme \((Y_i, \delta_i)\), \((Y_i, \delta_i)\). Obyčejné hypotéza je \( H_0 : F_1 = F_2 \).

První test se nazývá Gehanův a je zobecněným Wilcoxonova (resp. Mann-Whitneyho) nesparametrického testu na cenzorované pozorování. Testovací statistika má tvar
\[
U = \sum_{i=1}^{n} \sum_{j=1}^{n} U_{ij}, \quad \text{kde} \quad U_{ij} = \begin{cases} 1 & \text{pro } t_j \leq t_i, \\ 0 & \text{jinak}, \\ -1 & \text{pro } t_j > t_i. \end{cases}
\]

Druhý, Mantal-Haenschelův, je založen na posouzení čtyřplošních kontingenčních tabulek. Výběry sloučíme, uspořádáme a pro každé cenzorované pozorování vytvoříme tabulku pro počty v kategoriích skupin 1, žalně selhal, žalně selhal. Testovací statistika se rovná
\[
M-H = \frac{\sum (a_i - E(a_i))^2}{\sqrt{\sum \text{var}(a_i)}},
\]

kde \( a_i \) je počet jedinců v levém horním polišku kontingenční tabulky u i-tého cenzorovaného pozorování.

Oba výše uvedené testy lze zapsat pomocí
\[
\sum w_i(a_i - E(a_i)),
\]

kde \( w_i \) se pro Mantal-Haenschelův a \( w_i \) pro Gehanův test, kde \( n_i \) je počet jedinců v riziku v okamžiku i-tého uspořádaného selhání. Odhad je vidět, že Gehanův test klade větší váhu na počet větší selhání, zatímco M-H test dává větší vahy většině. Oba testy mají zobecnění pro více skupin. (Podrobnější literatura viz např. [23, 661, 91].)
V následující části uvedeme několik modellů, které jsou vhodné pro vyšetřování vlivu vysvětlujících faktorů (např. proměnných) na dobu do selhání. Pro každého jedince může dán různý vektor z vysvětlujících faktorů. Složky z reprezentují různé veličiny, jako např. způsob oběti, vnitřní vlastnosti jedinců (pohlaví, věk) a vnější podmínky. Další složky vektoru z mohou být dodatečně vytvořeny: např. složky k vyšetřování interakcí mezi faktory. Vysvětlující faktory mohou být konstantní a nebo proměnné v čase. Často je výhodné definovat vektor z tak, aby z = (0, ..., 0) odpovídal nějaké "základní" množině podmínek. Modelování pak má dvě úrovni:

(a) model pro dobu do selhání při z = (0, ..., 0);
(b) reprezentace změn při nenulovém z.

4.1 Parametrické regresní modely

Intenzita selhání \( \lambda(t; z) \) je určena až na vektor neznámých parametrů \( \theta = (\phi, \beta) \), kde \( \phi \) ovlivňuje intenzitu a \( \beta \) se vztahuje k vysvětlujícím faktorům. Přesná přesně, kdy aplikujír plné parametrický model neexistuje, ale je vhodné se na následující doporučení:

(a) pro velké složité situace je vhodná alespoň částečná parametrizace
(b) pro pouze zrovnačnění několika skupin s relativně velkým počtem pozorování je většinou nejlépe užít test na porovnání němocí funkcí příslušnosti
(c) můžeme studovat relativně složité závislosti a data jsou cenzurována, pak proporcionální Coxův model je výhodný
(d) klademe hlavní důraz na vysvětlující faktory, ochlad funkce příslušnosti není důležitý
(e) jest třeba provést alespoň nějakou kontrolu správnosti parametrického vyjadření
(f) statistické závěry jsou citlivé na externí hodnoty jak v prostoru \( z \) tak v prostoru do do selhání


(i) exponenciální rozložení s intenzitou \( \lambda(t; z) = \exp(\beta'z) \)
(ii) Weibullovo rozložení s intenzitou \( \lambda(t; z) = k t^{k-1} \exp(\beta'z) \)
(iii) obecný proporcionální model s intenzitou \( \lambda(t; z) = \lambda_0(t; \phi) h(z) \), kde \( \lambda_0 \) a \( h \) jsou známá a nezáporné funkce
(iv) aditivní model s intenzitou \( \lambda(t; z) = \lambda_0(t; \phi) + h(z) \)

(Podrobnější literaturu lze najít např. v [21], [6], [9])

4.2 COXŮV REGRESNÍ MODEL

Tato část bude věnována proporcionálnímu semiparametrickému modelu s intenzitou

\[ \lambda(t; z) = \lambda_0(t) \exp(\beta'z), \]

kde \( \lambda_0(t) \) je neznámá nezáporná funkce a \( \beta \) je p-rozumný vektor neznámých regresních koeficíntů. Napříč se budeme zabývat ochodem regresních koeficíntů. Odvodíme funkci, kterou Cox nazval "podmíněná" věrohodnost. S jeho pomocí získáme ochody regresních koeficíntů, které mají stejné asymptotické vlastnosti jako ochody maximální věrohodné a jeho vládnou eficínciu je poměrně velká. Řáděte se jeho nástrojem vědeckého kroku věděka (např. shodně). Pro uspořádání pozorování platí \( w_{ij} (w_{ij} = 0 \) je indikátor selhání a \( z_{ij} \) je vektor vysvětlujících faktorů příslušného \( z_{ij} \)). Přidáme-možnou podmínku, kde jsou v riziku lékaři před přijímáním \( z_{ij} \) tedy možno \( z_{ij} \) obsahuje jedinco. Když přišlo k časům \( w_{ij}, \ldots, w_{ij} \). Pro každé neocenzurované
pouoování platí

\( P( d o j e k s e l h n í v \{ w_{i0}, w_{i1}, \ldots, w_{i} \} i s \{ r_{i} \} = \sum_{j \in S_{i0}} \beta_{j}^{2} \lambda_{0} w_{i0} \lambda_{w} \)

\( P( s e l j e w_{i} v č a s e w_{i0} i j e n e j e n e c z s \{ r_{i} \} s e l j e v č a s e w_{i} ) = \frac{\exp(\beta^{2} z_{i})}{\sum_{j \in S_{i0}} \exp(\beta^{2} z_{j})} \)

Vezmeme-li součin těchto podmíněných pravděpodobností, dostaneme tzv. 'podmíněnou' věrohodnost

\[ L_{o}(\beta) = \prod_{\text{necenzor}.} \frac{\exp(\beta z_{w})}{\sum_{j \in S_{i0}} \exp(\beta z_{j})} \]

Při odhadování regresních koeficientů zacházejeme s \( L_{o}(\beta) \) jako s obvyklou věrohodnostní funkcí. Tedy k náležení odhadů užijeme skorový vektor a výběrovou informační matici

\[ U(\beta) = \frac{3}{3^{2}} \log L_{o}(\beta) a \quad i(\beta) = - \frac{3}{3^{2}} \log L_{o}(\beta) \]

Vektor \( \hat{\beta} \) je řešením soustavy rovnic \( U(\beta) = 0 \), která většinou vyžaduje iterativní metodu. Stejně jako pro maximálně věrohodné odhady platí, že \( \hat{\beta} \) je asymptoticky normální se střední hodnotou \( \beta \) a rozptylem \( i^{2}(\beta) \). K testování hypotézy \( H_{0}: \beta = 0 \) lze užít skorovou statistiku \( U^{2}(\beta) \) \( i^{2}(\beta) \), která má za platnostnou hypotézu \( x^{2} \) s \( p \) stupněmi volnosti. Jednotlivé složky skorového vektoru a informační matice mají pro \( \beta = 0 \) jednodušší tvar

\[ \frac{3}{3^{2}} \log L_{o}(\beta) = \sum_{\text{necenzor}.} \left( z_{w} - z_{w0} \right) \]

\[ i(\beta) = \sum_{\text{necenzor}.} \left\{ \frac{1}{n_{i}} \sum_{j \in S_{i0}} (z_{w} - z_{w0})(z_{w} - z_{w0}) \right\} \]

kde \( z_{w} = r_{i}^{-1} \sum_{j \in S_{i0}} z_{j} \) a \( n_{i} \) je počet prvků množiny \( S_{i0} \).

Speciálně pro \( p = 1 \) a nevstříkatučí faktor, který označuje příslušnost do skupiny, tedy \( z_{w} = 0 \) pro jedince z 1.skupiny a \( z_{w} = 1 \) pro jedince z 2.skupiny. Výše uvedená statistika odpovídá Mantel-Haesaurově za předpokladu, že nedochází ke shodám.

Dále uvedeme odhad funkce přeřížit \( Sl(x) \). V literatuře existuje mnoho přístupů, zde uvedeme pouze jeden. Funkci přeřížit vyjadříme ve tvaru

\[ S(t; x) = \exp( - t^{2} \int_{0}^{t} \lambda_{0}(u) du ) = S_{0}(t)^{e^{Bx}} \]

a \( S_{0}(t) \) odhademe podle Breslowa pomocí

\[ \hat{S}_{0}(t) = \prod_{w_{0} \leq t} \left( 1 - \sum_{j \in S_{i0}} \exp(\beta z_{j}) \right) \]

Literatura týkající se Coxova modelu je např. [22], [61], [81], [91].

Nyní se budeme zabývat vztahem mezi 'podmíněnou' věrohodností a některými jinými typy věrohodností. Nejprve uveďeme základní věrohodnost pro pořadí a předpokládejme, že nedochází ke shodám. Pozorujeme dvě věci \( \{ X_{0}, \delta_{1} \}, \ldots, \{ X_{w}, \delta_{n} \} \), kde k nim příslušná \( T \) májí distribuční funkci \( F_{T} \) s hustotou \( f_{T} \). Vektor pořadí \( R = (R_{0}, \ldots, R_{w}) \) je vytvořen následujícím způsobem:

- \( R_{j} = 1 \) pokud je \( X_{j} \) mezi necenzorovanými počátečními \( \delta_{j} = 1 \), \( R_{j} = 0 \) pokud je předchozího necenzorovaného \( \delta_{j} = 0 \).
Marginální věrohodnost pro pořadí má pak vyzkoušení

\[ p(r, s) = \prod_{u_k \in u_k} \prod_{u \in u_{vec}} (f_{u_0(u)} \prod_{i \in u_{vec}} [1 - F_i(u_i)]) \, du_1 \cdots du_{u_{vec}}. \]

Dosadíme-li

\[ F_j(t) = 1 - \exp(-e^{\beta Z_j} \int_0^t \lambda_0(s) \, ds), \]

dostaneme

\[ L_0(s) = p(r, s). \]

Dále předpokládáme, že máme posloupnost dvojic náhodných veličin \(X_i, S_i\), \(X_i, S_i\). Nechť \(u_{vec}\) označuje i-té uspořádané necenzurované pozorování, \(X_i\) obsahuje informaci o cenozorování v intervalu \((u_{vec-1}, u_{vec})\) a informaci o tom, že k selhání došlo v \(u_{vec}\) a \(S_i\) obsahuje informaci o tom, že právě jedinec se \(X_i\) selže v \(u_{vec}\).

Potom platí, že marginální věrohodnost \(S_i, \ldots, S_n\) má tvar

\[ p(S_i, \ldots, S_n \mid X_i, \ldots, X_n) = \prod_{i=1}^m p(S_i \mid S_i, \ldots, S_{i-1}). \]

Postupně věrohodnost \(S_i, \ldots, S_n\) dán \(X_i, \ldots, X_n\) má tvar

\[ p(S_i, \ldots, S_n \mid X_i, \ldots, X_n) = \prod_{i=1}^m p(S_i \mid S_i, \ldots, S_{i-1}, X_i, \ldots, X_{i-1}). \]

Plné věrohodnost se rovná

\[ p(S_i, \ldots, S_n, X_i, \ldots, X_n) = \prod_{i=1}^m p(X_i \mid S_i, \ldots, S_{i-1}, X_i, \ldots, X_{i-1}, S_i, \ldots, S_{i-1}) = \]

\[ = \prod_{i=1}^m p(X_i \mid S_i, \ldots, S_{i-1}, X_i, \ldots, X_{i-1}, S_i, \ldots, S_{i-1}) \]

Cox v (15) dokázal, že platí

\[ L_0(s) = \prod_{i=1}^m \prod_{S_i, X_i, X_i, S_i, \ldots, S_{i-1}}. \]

Proto se \(L_0\) také nazývá čiastočná věrohodnost.

Před použitím Coxova modelu je vhodné ověřit, zda jsou splněny předpoklady proporcionality. Obvykle se používá test grafický, který je vhodný ve sporných případech doplnit testem numerickým. Nyní uvedeme hlavní myšlenku grafického testu. Vektor vyššetřujících faktorů napišeme ve tvaru \(z = (z_1, \ldots, z_k, z_{k+1})\), kde prvních \(k\) složek splňuje předpoklad proporcionality. Intenzita pro i-tého jedince má tvar

\[ \lambda_i(t, z_{i, 1}, \ldots, z_{i, k+1}) = \mu_0(t, z_{i, k+1}) \exp(\beta Z_i). \]

Za platnosti předpokladu proporcionality platí

\[ \mu_0(t, z_{i, k+1}) = \lambda_0(t) \exp(\beta z_{i, k+1}). \]

Vyššetřující faktor \(z_{k+1}\) se kategorizuje do p úrovní. Potom platí

\[ \lambda_i(t, z_{i, 1}, \ldots, z_{i, k+1}) = \lambda_{0s}(t) \exp(\beta z_{i, 1}), s=1, \ldots, p. \]

Za platnosti hypotézu dostaneme

\[ \lambda_{0s}(t) = \lambda_{0s}(t) e^{\beta z_{i, 1}}, s=1, \ldots, p. \]
Označime-li

\[ \Lambda(t) = \int_0^t \Lambda(s) \, ds \, \text{d}w, \quad s = 1, \ldots, p. \]

pak grafy \( \log \Lambda(t) = 1, \ldots, p \) versus t mají za platnosti předpokladu proporcionality konstantní úsečky \( \varepsilon_1, \ldots, \varepsilon_p \).

Poměrně jednoduchý numerický test lze nalézt v [12] a nebo v [24].

5. LITERATURA

Knihu


Články