

Odhady aktuárských předpokladů pomocí analýzy přežití

Seminář z aktuárských věd – 11.5.2012

- **analýza přežití, teorie spolehlivosti**
- **survival analysis, failure time analysis, event history analysis**

→ zkoumání doby do události

- **využití**
 - demografie
 - medicína
 - technika
 - sociální vědy
 - pojišťovnictví
 - storna
 - další (např. doba do zaplacení dlužného pojistného)
 - úmrtnost

- **Kaplanův – Meierův odhad**
- **Coxův regresní model**
- **Parametrické modely**

Analýza přežití

Základní model

■ doba do události

- náhodná veličina $T \geq 0$

■ charakteristiky statistického rozložení T

- distribuční funkce

$$F(t) = P(T \leq t)$$

- funkce přežití

$$S(t) = P(T > t) = 1 - F(t)$$

- spojitý případ

- hustota

$$f(t)$$

- riziková funkce

$$\lambda(t) = \lim_{h \rightarrow 0+} \frac{P(t < T \leq t+h | T > t)}{h} = \frac{f(t)}{1-F(t)} = -\frac{d}{dt} \log(S(t))$$

- kumulativní riziková funkce

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

- diskrétní případ

- pravděpodobnostní funkce

$$f(t_j) = f_j = P(T = t_j)$$

- riziková funkce

$$\lambda(t_j) = \lambda_j = P(T = t_j | T \geq t_j)$$

- kumulativní riziková funkce

$$\Lambda(t) = \sum_{\{j | 0 \leq t_j \leq t\}} \lambda_j = \sum_{s \leq t} \lambda(s)$$

Analýza přežití

Základní model

■ podoba pozorování

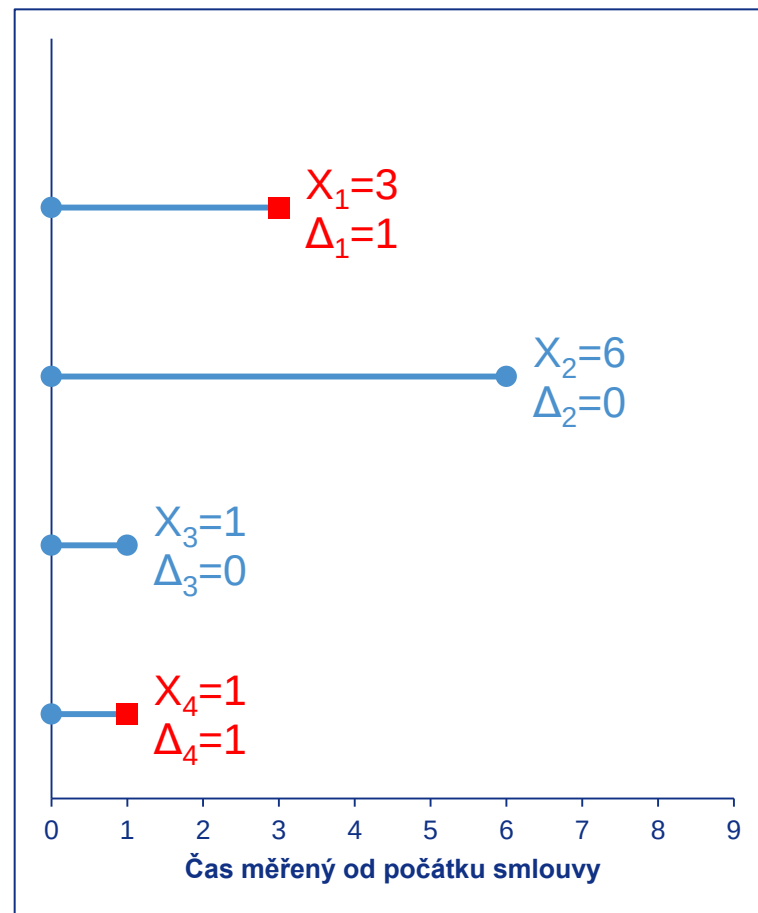
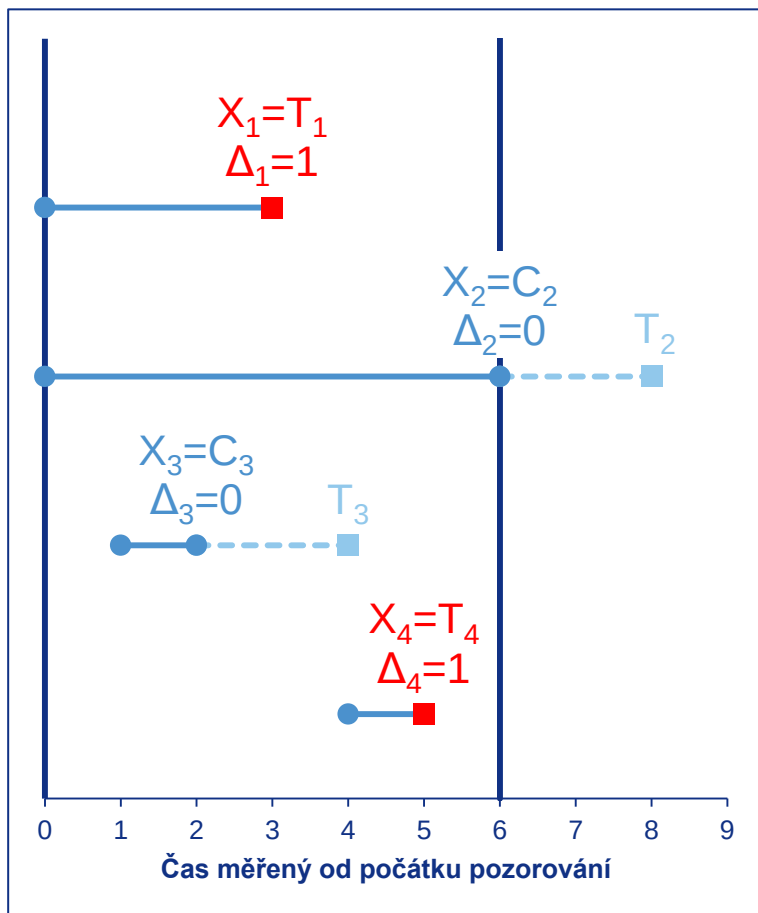
- pozorováno n objektů s realizacemi $T_1, T_2, \dots, T_n$ veličiny T

■ cenzorování

- přímo časy T_i nejsou známy pro všechna pozorování
- pozorování ukončená cenzorováním v čase C_i nesou pouze informaci, že $T_i \geq C_i$
- k dispozici dvojice (X_i, Δ_i)
 - $X_i = \min(T_i, C_i)$
 - $\Delta_i = I(T_i < C_i)$
- typy cenzorování
 - typ I (událostí)
 - typ II (časem)
 - náhodné – nezávislé na T
 - náhodné – informativní (nese informaci o riziku)

Analýza přežití

Obrázek



Kaplanův – Meierův odhad

Konstrukce odhadu

■ čítací proces

- $N_i(t) = I(X_i \leq t, \Delta_i = 1)$
- $N(t) = \sum_i N_i(t)$

■ pozorovací proces

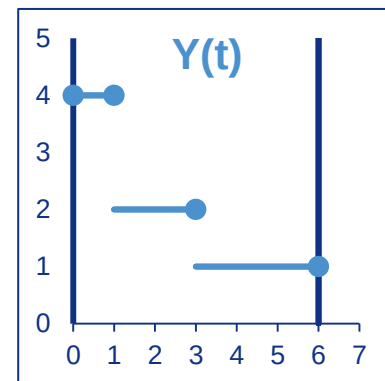
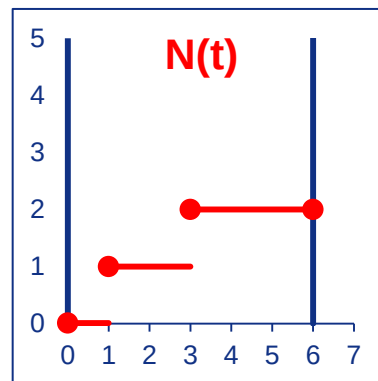
- $Y_i(t) = I(X_i \geq t)$
- $Y(t) = \sum_i Y_i(t)$

■ předpoklady

- T_i i. i. d., nezávislé na C_i

■ výsledky pozorování

- $t_1 < t_2 < \dots < t_k$ časy pozorovaných událostí
- $dN(t_1), dN(t_2), \dots, dN(t_k)$ počty pozorovaných událostí v časech $t_1, t_2, \dots, t_k$
- $t_0 = 0, t_{k+1} = \infty$
- v intervalu (t_j, t_{j+1}) , $j = 0, 1, \dots, k$ nastalo $Y(t_j) - Y(t_{j+1}) - dN(t_j)$ cenzorování v časech t_{il} , $l = 1, 2, \dots, Y(t_j) - Y(t_{j+1}) - dN(t_j)$



Kaplanův – Meierův odhad

Konstrukce odhadu

■ odhad funkce přežití $S(t)$

■ věrohodnostní funkce

- $$L = \prod_{j=0}^k \left(\left(S(t_j^-) - S(t_j) \right)^{dN(t_j)} \prod_{l=1}^{Y(t_j) - Y(t_{j+1}) - dN(t_j)} S(t_{jl}) \right)$$

- $\hat{S}(t_j) = \prod_{m=1}^j (1 - \widehat{\lambda}_m)$ pro $j = 0, 1, \dots, k$

- $\hat{S}(t_{jl}) = \prod_{m=1}^j (1 - \widehat{\lambda}_m)$ pro $t \in \langle t_j, t_{j+1} \rangle$

- $L =$

$$= \prod_{j=0}^k \left(\lambda_j^{dN(t_j)} \prod_{m=1}^{j-1} (1 - \lambda_m)^{dN(t_j)} \prod_{m=1}^j (1 - \lambda_m)^{Y(t_j) - Y(t_{j+1}) - dN(t_j)} \right)$$

- $L = \prod_{j=1}^k \lambda_j^{dN(t_j)} (1 - \lambda_j)^{Y(t_j) - dN(t_j)}$

$$\rightarrow \hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{dN(s)}{Y(s)} \right)$$

Kaplanův – Meierův odhad

Vlastnosti odhadu

■ vychýlení odhadu ($E[\hat{S}(t)] \geq S(t)$)

■ odhad rozptylu

- $\log(\hat{S}(t)) = \sum_{\{j|t_j \leq t\}} \log\left(1 - \frac{dN(t_j)}{Y(t_j)}\right)$

- $\mathcal{P}(dN(t_j)|Y(t_j) = n_j) \sim \mathcal{Bi}(p_j, n_j)$

- $f(X) \approx f(c) + f'(c)(X - c), \text{var}[f(X)] \approx f'(c)^2 \text{var}[X]$

- $\log(\hat{S}(t)) \approx \sum_{\{j|t_j \leq t\}} \left(\log(1 - p_j) - \frac{1}{1-p_j} \left(\frac{dN(t_j)}{Y(t_j)} - p_j \right) \right)$

- $\text{var}[\log(\hat{S}(t))] \approx \sum_{\{j|t_j \leq t\}} \frac{1}{(1-p_j)^2} \frac{p_j(1-p_j)}{n_j}$

$$\rightarrow \widehat{\text{var}}[\hat{S}(t)] = (\hat{S}(t))^2 \sum_{s \leq t} \frac{dN(s)}{Y(s)(Y(s) - dN(s))}$$

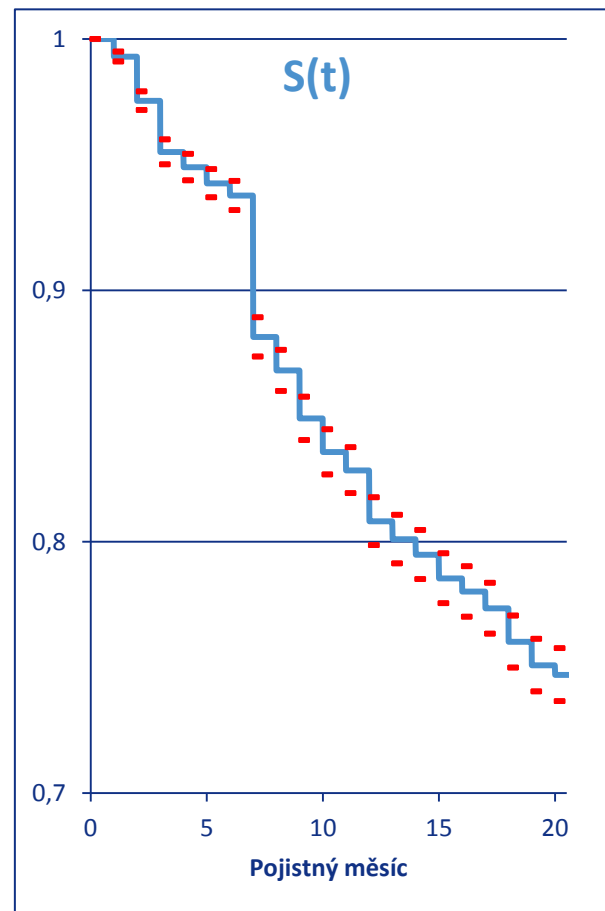
■ asymptotické rozložení

$$\rightarrow \hat{S}(t) \sim \mathcal{N}(S(t), \widehat{\text{var}}[\hat{S}(t)])$$

Kaplanův – Meierův odhad

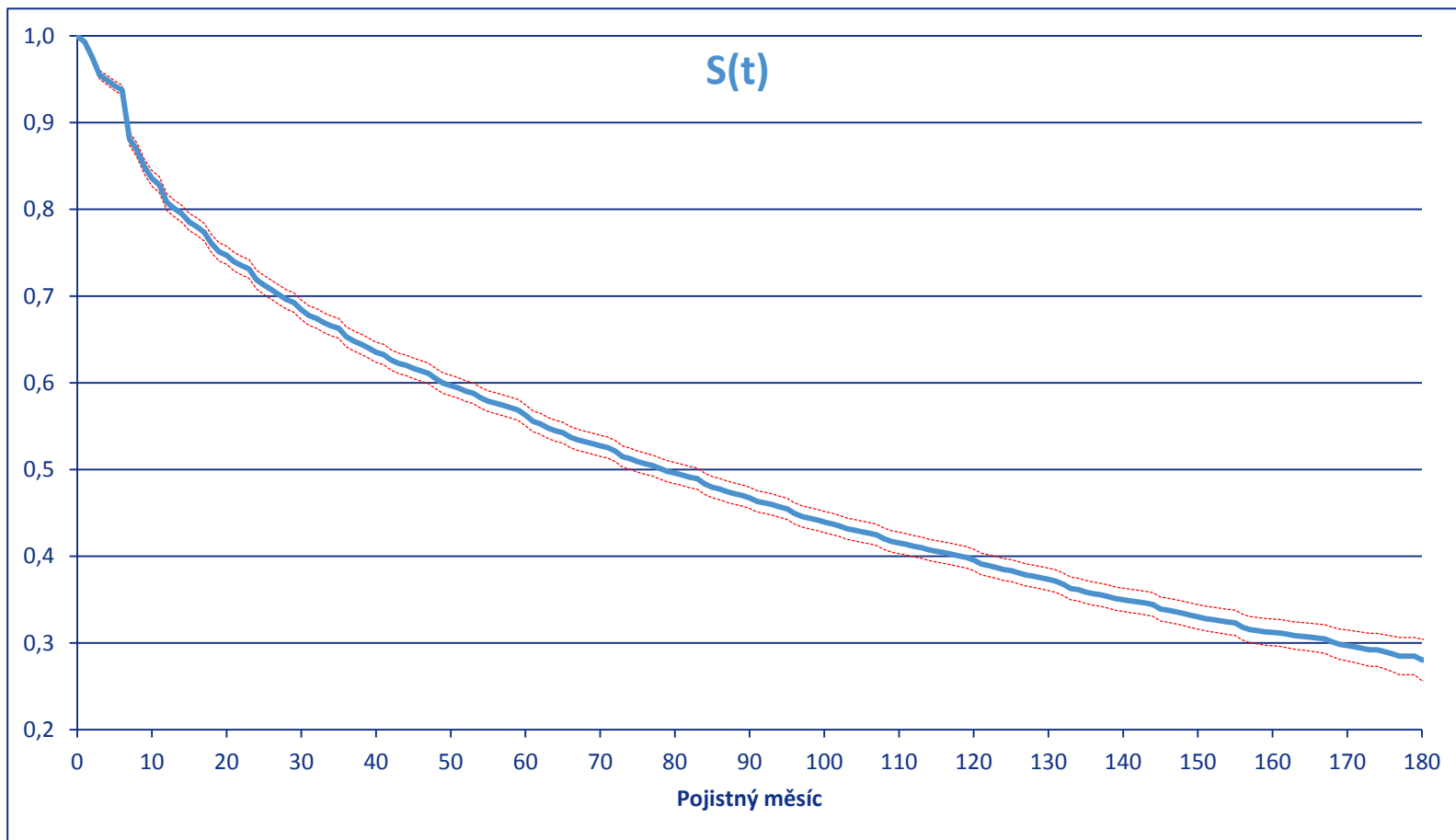
Příklad

t	dN(t)	Y(t)	S(t)	$\sigma[S(t)]$	$Q_{0,5\%}[S(t)]$	$Q_{99,5\%}[S(t)]$
1	82	11653	0,9930	0,0008	0,9910	0,9950
2	204	11523	0,9754	0,0014	0,9717	0,9791
3	234	11203	0,9550	0,0019	0,9500	0,9600
4	68	10815	0,9490	0,0020	0,9437	0,9543
5	73	10743	0,9426	0,0022	0,9370	0,9481
6	55	10663	0,9377	0,0023	0,9319	0,9435
7	636	10602	0,8814	0,0030	0,8736	0,8892
8	150	9931	0,8681	0,0032	0,8600	0,8763
9	215	9770	0,8490	0,0034	0,8404	0,8577
10	150	9549	0,8357	0,0035	0,8267	0,8446
11	82	9396	0,8284	0,0035	0,8193	0,8375
12	228	9313	0,8081	0,0037	0,7986	0,8176
13	80	9082	0,8010	0,0038	0,7913	0,8107
14	69	9000	0,7949	0,0038	0,7851	0,8046
15	106	8930	0,7854	0,0039	0,7755	0,7954
16	59	8823	0,7802	0,0039	0,7701	0,7902
17	75	8763	0,7735	0,0039	0,7634	0,7836
18	149	8685	0,7602	0,0040	0,7499	0,7706
19	105	8535	0,7509	0,0041	0,7404	0,7613
20	43	8429	0,7470	0,0041	0,7365	0,7576
...	...	...	...	...	...	...
175	1	136	0,2899	0,0076	0,2703	0,3096
176	1	122	0,2876	0,0079	0,2671	0,3080
177	1	105	0,2848	0,0083	0,2634	0,3062
178	0	91	0,2848	0,0083	0,2634	0,3062
179	0	79	0,2848	0,0083	0,2634	0,3062
180	1	64	0,2804	0,0093	0,2564	0,3043



Kaplanův – Meierův odhad

Příklad



Kaplanův – Meierův odhad

Pravděpodobnost události v intervalu

- pravděpodobnost události v intervalu (s, t) za podmínky, že k události nedošlo do času s

- $R(s, t) = P(s < T_i \leq t | T_i > s) = 1 - \frac{S(t)}{S(s)}$

- odhad pravděpodobnosti

$$\rightarrow \hat{R}(s, t) = 1 - \frac{\hat{S}(t)}{\hat{S}(s)}$$

- odhad rozptylu

- $\log(1 - R(s, t)) = \sum_{\{j | s < t_j \leq t\}} \log\left(1 - \frac{dN(t_j)}{Y(t_j)}\right)$
 $\rightarrow \widehat{\text{var}}[\hat{R}(s, t)] = \left(1 - \hat{R}(s, t)\right)^2 \sum_{s < u \leq t} \frac{dN(u)}{Y(u)(Y(u) - dN(u))}$

- asymptotické rozložení

$$\rightarrow \hat{R}(s, t) \sim \mathcal{N}(R(s, t), \widehat{\text{var}}[\hat{R}(s, t)])$$

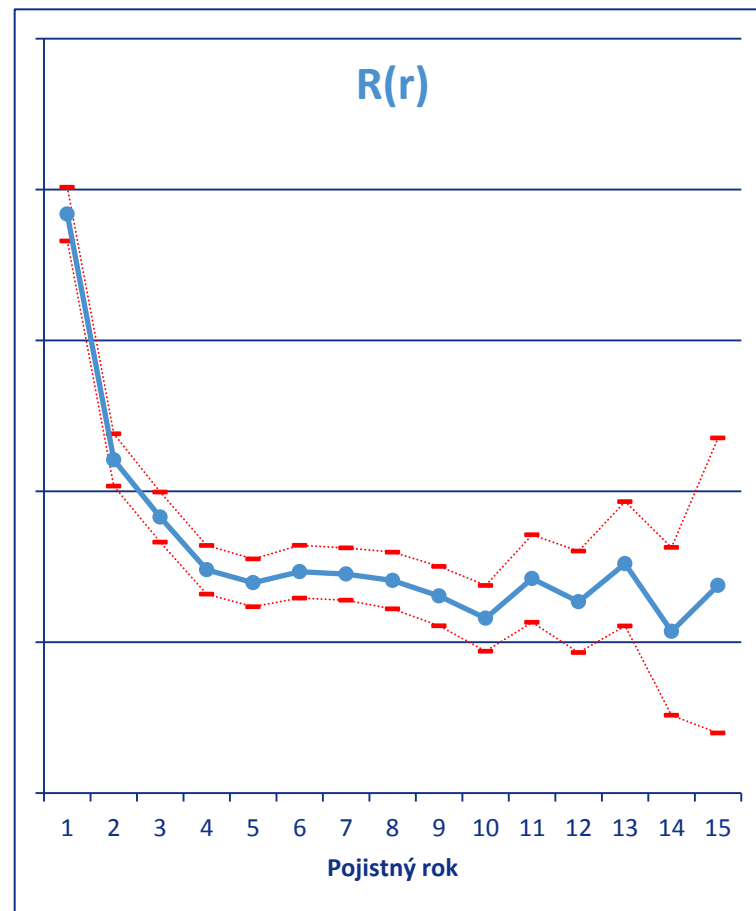
- speciálně při měsíčním pozorování ($t = 1, 2, \dots$) pro pravděpodobnost události v roce r

- $R(r) = R(12r, 12(r - 1))$

Kaplanův – Meierův odhad

Příklad

r	R(r)	$\sigma[R(r)]$	$Q_{0,5\%}[R(r)]$	$Q_{99,5\%}[R(r)]$
1	15,35%	0,002763	14,64%	16,06%
2	8,83%	0,002689	8,14%	9,53%
3	7,32%	0,002579	6,65%	7,98%
4	5,92%	0,002511	5,27%	6,57%
5	5,58%	0,002458	4,94%	6,21%
6	5,87%	0,002716	5,17%	6,57%
7	5,81%	0,002694	5,11%	6,50%
8	5,64%	0,002918	4,89%	6,39%
9	5,22%	0,003060	4,44%	6,01%
10	4,63%	0,003383	3,76%	5,51%
11	5,69%	0,004506	4,53%	6,85%
12	5,07%	0,005220	3,73%	6,42%
13	6,08%	0,006395	4,43%	7,73%
14	4,29%	0,008638	2,06%	6,51%
15	5,50%	0,015185	1,59%	9,41%



Kaplanův – Meierův odhad

Shrnutí

- velmi časté použití
- výpočetně nenáročný přístup
- předpoklad nezávislosti jednotlivých pozorování
- předpoklad stejného rozložení doby do události
 - pro všechna pozorování
 - v průběhu (kalendářního) času
- zajištění splnění předpokladů
 - vytvoření dostatečně homogenních skupin pozorování
 - test shody výběrů
 - pozorování za dostatečně krátké období
 - useknutí pozorování zleva
- rozšíření modelu
 - závislost na dalších proměnných
 - Coxův regresní model

Analýza přežití

Test shody výběrů

■ test hypotézy shody funkce přežití

- $S_0(t) = S_1(t) = \dots = S_p(t)$ pro všechna $t \geq 0$

→ (vážený) logrankový test

■ test hypotézy shody pravděpodobnosti přežití v určitém čase

- $S_0(t|\mathcal{T}) = S_1(t|\mathcal{T}) = \dots = S_p(t|\mathcal{T})$

- pro konkrétní čas t
- pro $t \in \langle t_1, t_2 \rangle$

- možnost využití vlastností Kaplanova – Meierova odhadu

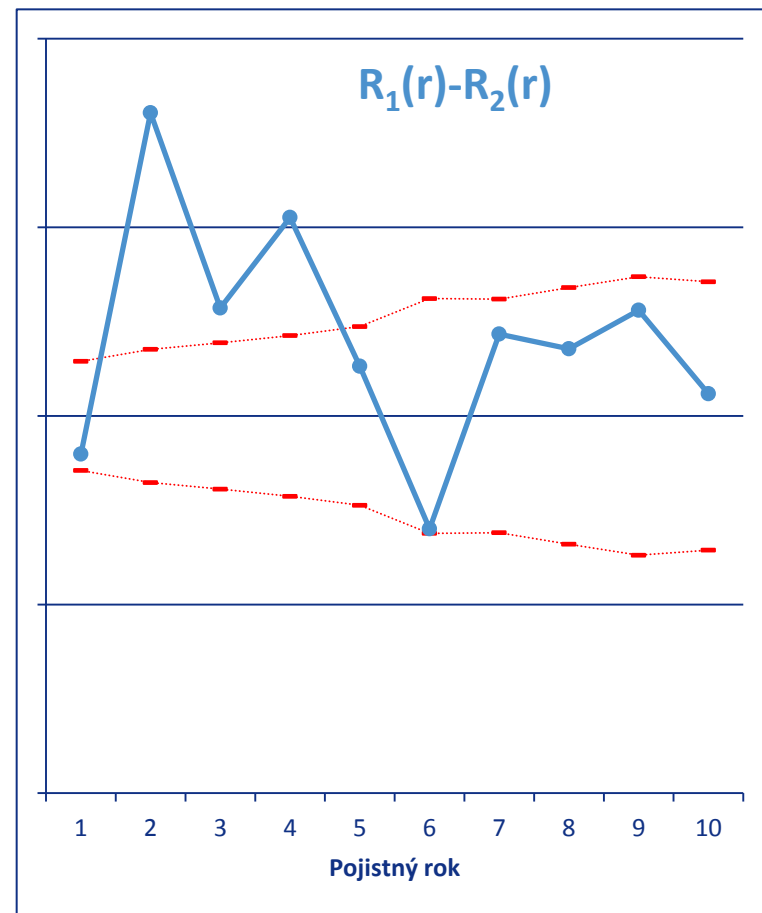
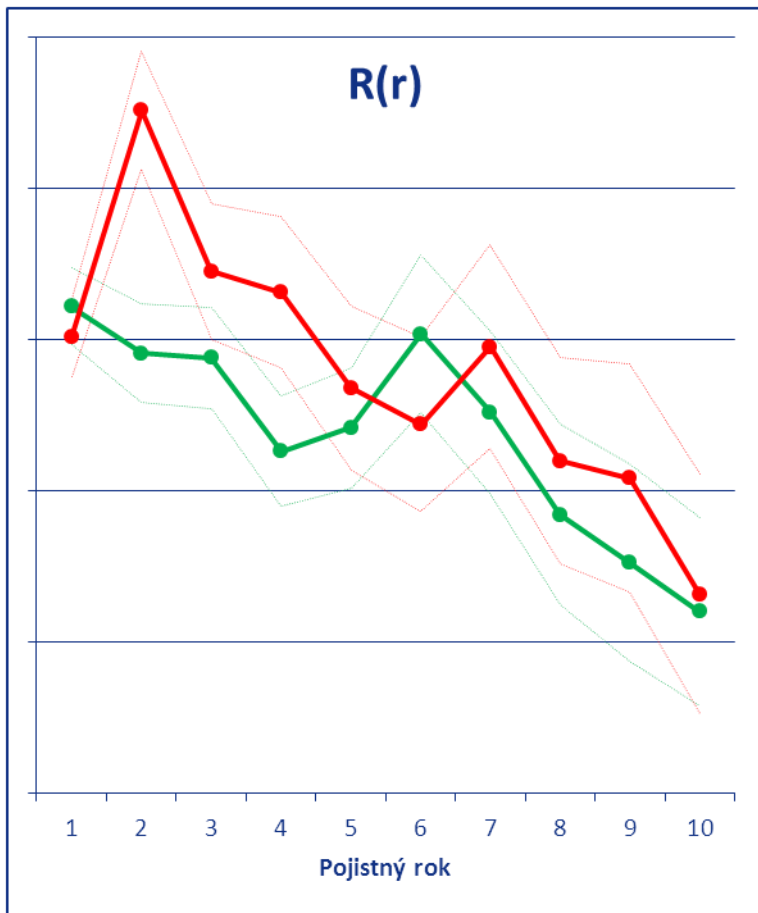
- například na základě $R(r)$

- $\widehat{\text{var}}[\hat{R}_i(r)] = \left(1 - \hat{R}_i(r)\right)^2 \sum_{[12(r-1) < u \leq 12r]} \frac{dN(u)}{Y_i(u)(Y(u) - dN(u))}$

→ $\hat{R}_1(r) - \hat{R}_2(r) \sim \mathcal{N}\left(0, \widehat{\text{var}}[\hat{R}_1(r)] + \widehat{\text{var}}[\hat{R}_2(r)]\right)$

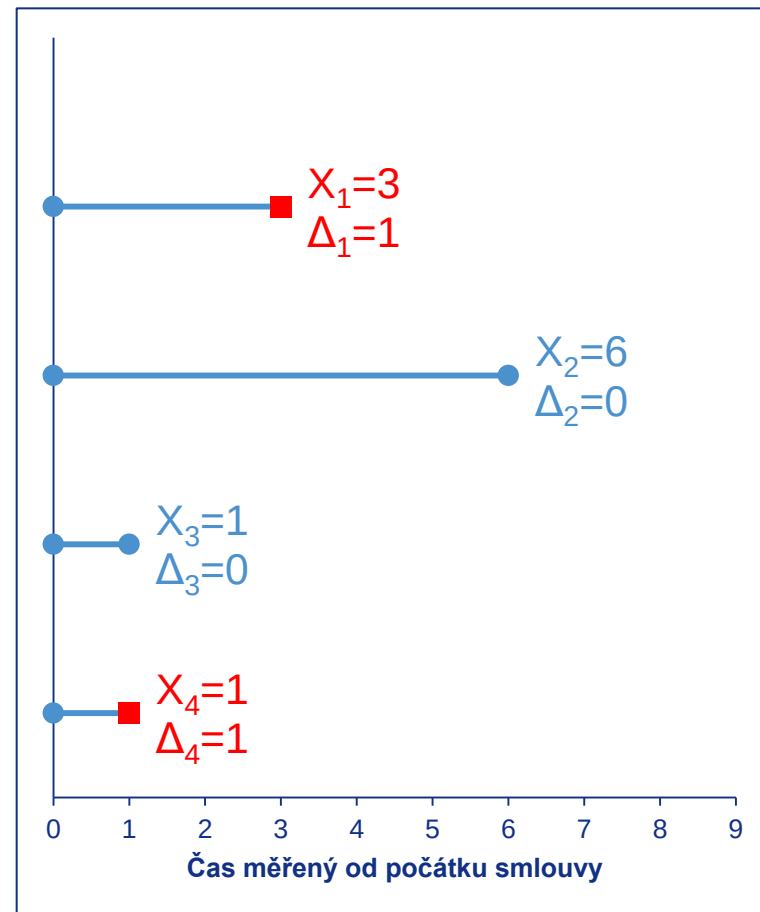
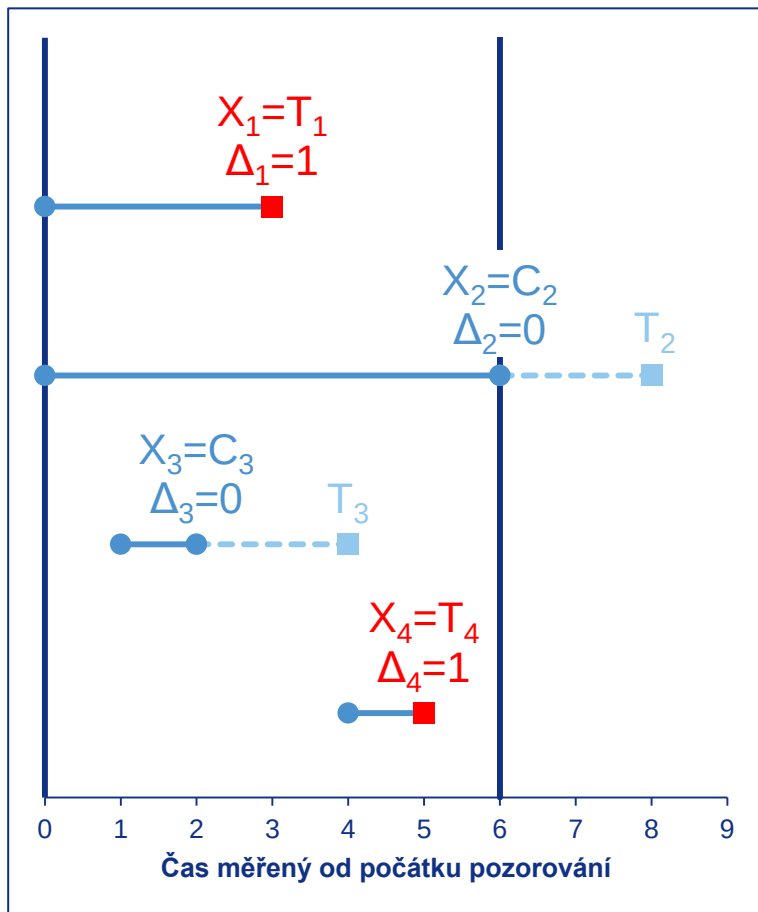
Analýza přežití

Test shody výběrů: Příklad



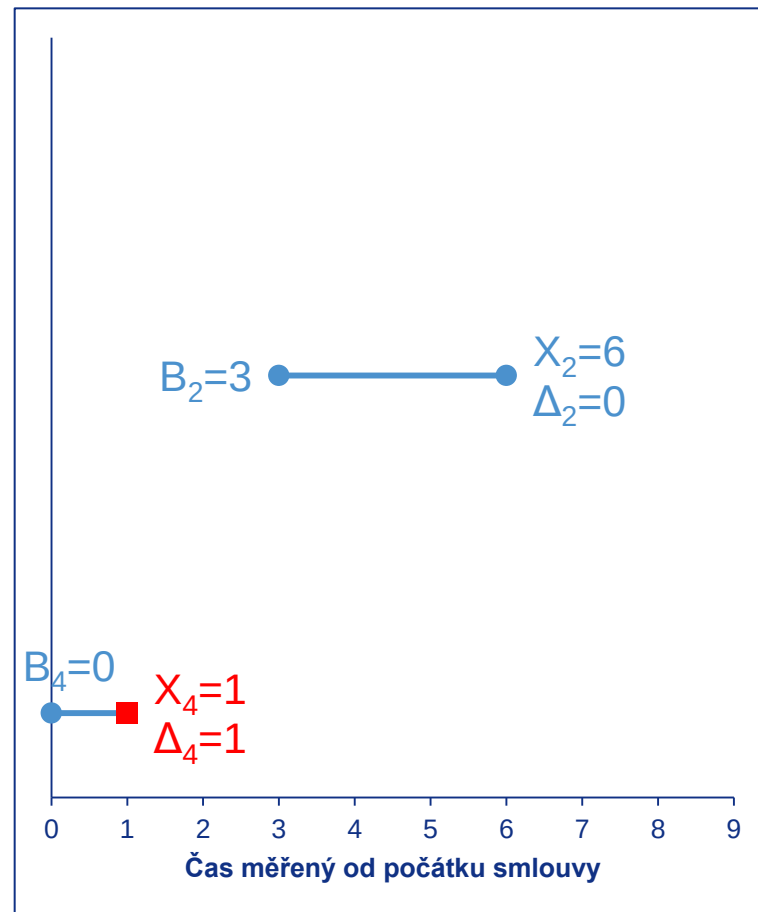
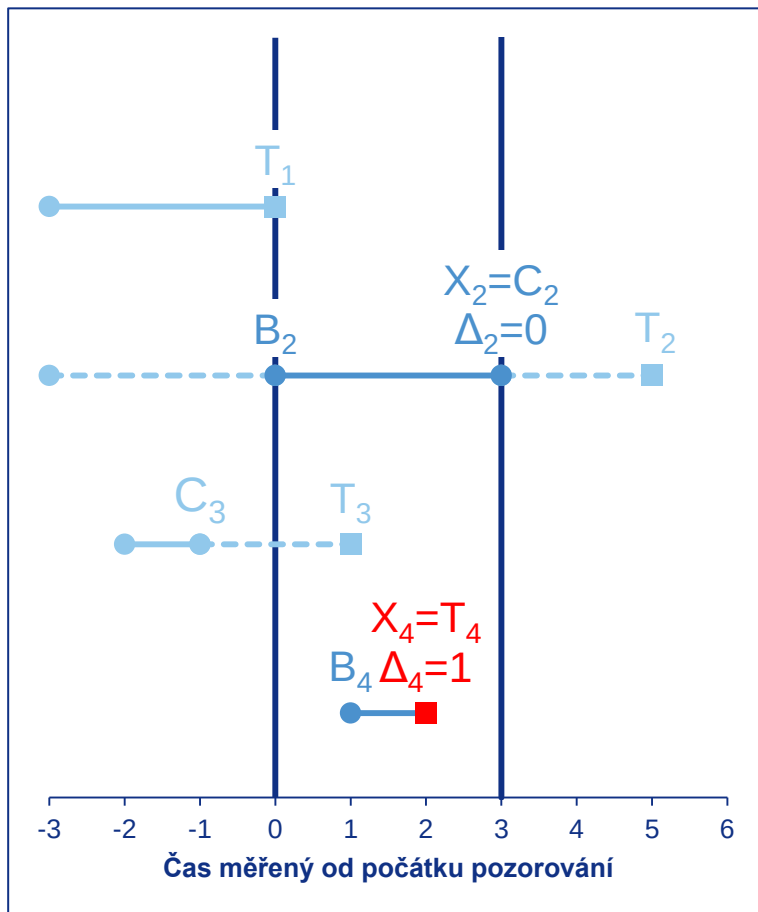
Analýza přežití

Obrázek



Analýza přežití

Useknutí pozorování zleva: Obrázek



Analýza přežití

Useknutí pozorování zleva

■ základní model – cenzorování zprava

- některá pozorování nekončí časem události T_i , ale časem cenzorování C_i
- informace, že $T_i \geq C_i$

■ useknutí zleva

- některé objekty nevstupují do pozorování v čase 0, ale v čase B_i
- žádná informace, pokud by událost nastala v čase $T_i < B_i$

■ další možnosti cenzorování nebo zkrácení

- cenzorování zleva
- intervalové cenzorování
- useknutí zprava

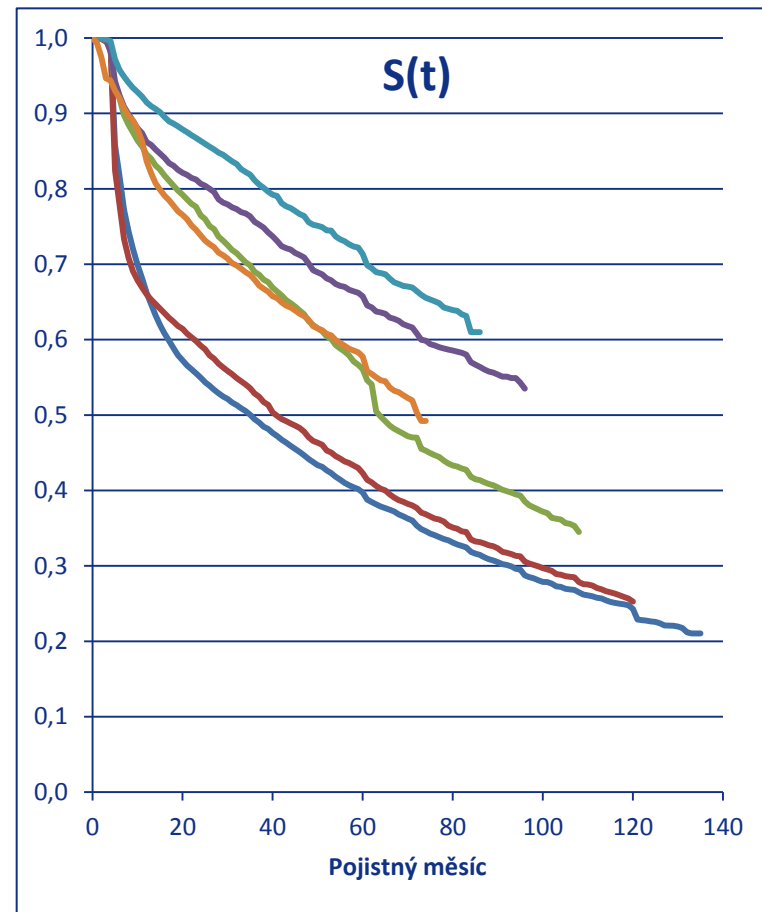
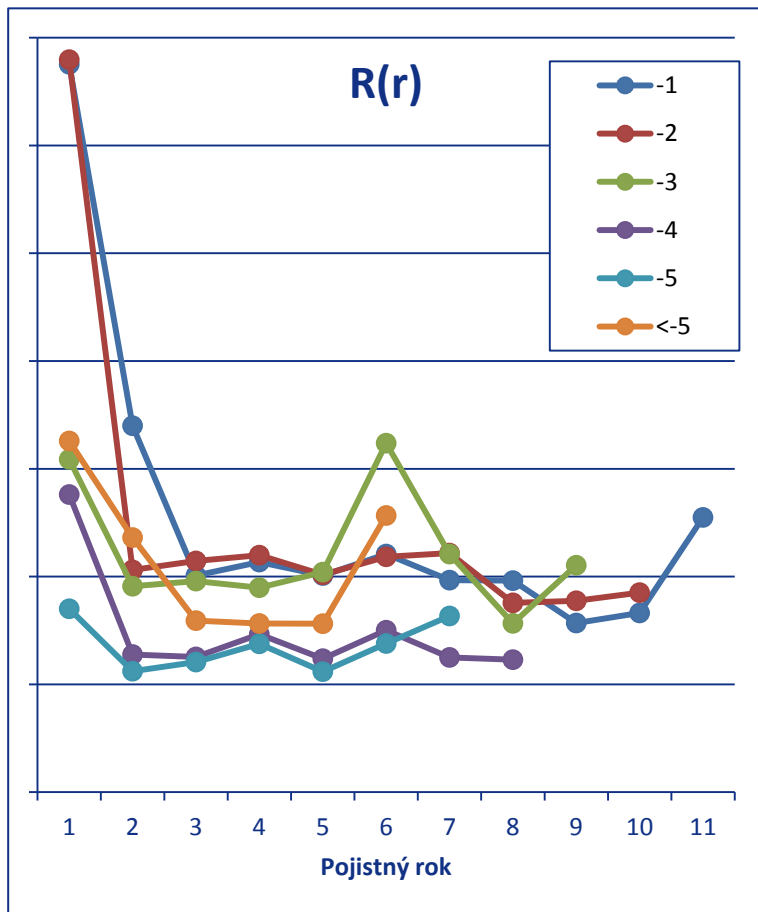
■ úprava základního modelu pro useknutí zleva

- pozorovací proces $Y_i(t) = I(B_i < t \leq X_i)$

■ možnost využití testu shody k určení délky období pro pozorování

Analýza přežití

Useknutí pozorování zleva: Příklad



Coxův regresní model

Formulace modelu

■ podoba pozorování

- $(X_i, \Delta_i, \mathbf{Z}_i(t)), i = 1, 2, \dots, n$
- $\mathbf{Z}_i(t)$ známé veličiny, které mohou mít vliv na dobu do události

■ tvar rizikové funkce

- $\lambda(t|\mathbf{Z}(t)) = \lambda_0(t)e^{\boldsymbol{\beta}^T \mathbf{Z}(t)}$
- značení pro i -tý objekt $\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}^T \mathbf{Z}_i(t)\}$

■ případ vektoru $\mathbf{Z}$ nezávislého na čase

- $\lambda(t|\mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}^T \mathbf{Z}}$
- model proporcionálních rizik
 - $\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{\boldsymbol{\beta}^T \mathbf{Z}_i\}}{\lambda_0(t)\exp\{\boldsymbol{\beta}^T \mathbf{Z}_j\}} = \frac{\exp\{\boldsymbol{\beta}^T \mathbf{Z}_i\}}{\exp\{\boldsymbol{\beta}^T \mathbf{Z}_j\}}$ nezávisí na čase
- tvar funkce přežití
 - $S(t|\mathbf{Z}) = \exp\left\{-\int_0^t \lambda_0(s)e^{\boldsymbol{\beta}^T \mathbf{Z}} ds\right\} = S_0(t)\exp\{\boldsymbol{\beta}^T \mathbf{Z}\}$
 - $S_0(t) = \exp\left\{-\int_0^t \lambda_0(s) ds\right\}$

Coxův regresní model

Konstrukce odhadů

■ věrohodnostní funkce pro odhad základní funkce přežití $S_0(t)$

- $L = \prod_{j=0}^k \left(\prod_{\{i|X_i=t_j, \Delta_i=1\}} \left(S(t_j^- | \mathbf{Z}) - S(t_j | \mathbf{Z}) \right) \prod_{\{i|t_j \leq X_i < t_{j+1}, \Delta_i=0\}} S(t_j | \mathbf{Z}) \right)$
 $\rightarrow \hat{S}_0(t) = \prod_{s \leq t} \left(1 - \frac{dN(s)}{\sum_{i=1}^n Y_i(s) \exp\{\beta^T \mathbf{Z}_i(s)\}} \right)$

■ parciální věrohodnostní funkce pro odhad vektoru parametrů β

- příspěvek události i -tého objektu $L_i(\beta) = \frac{\lambda_i(t_i) dt}{\sum_{j=1}^n Y_j(t_i) \lambda_j(t_i) dt}$
- $L(\beta) = \prod_{i=1}^n \prod_{s \geq 0} \left(\frac{\exp\{\beta^T \mathbf{Z}_i(s)\}}{\sum_{j=1}^n Y_j(s) \exp\{\beta^T \mathbf{Z}_j(s)\}} \right)^{dN_i(s)}$
 $\rightarrow \sum_{s \geq 0} \sum_{i=1}^n \left(\mathbf{Z}_i(s) - \frac{\sum_{j=1}^n \mathbf{Z}_j(s) Y_j(s) \exp\{\beta^T \mathbf{Z}_j(s)\}}{\sum_{j=1}^n Y_j(s) \exp\{\beta^T \mathbf{Z}_j(s)\}} \right) dN_i(s) = 0$

Coxův regresní model

Vlastnosti odhadů

■ asymptotické vlastnosti $\hat{\beta}$

- $S^{(0)}(\beta, t) = \sum_{j=1}^n Y_j(t) \exp\{\beta^T Z_j(t)\}$
- $S^{(1)}(\beta, t) = \sum_{j=1}^n Z_j(t) Y_j(t) \exp\{\beta^T Z_j(t)\}$
- $S^{(2)}(\beta, t) = \sum_{j=1}^n Z_j(t) Z_j(t)^T Y_j(t) \exp\{\beta^T Z_j(t)\}$
- $E(\beta, t) = Z - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}$
- $\Sigma(\beta) = \sum_{s \geq 0} \left(\frac{S^{(0)}(\beta, s) S^{(2)}(\beta, s) - (S^{(1)}(\beta, s))(S^{(1)}(\beta, s))^T}{(S^{(0)}(\beta, s))^2} \right) dN(s)$
 $\rightarrow \hat{\beta} \sim \mathcal{N}(\beta, \Sigma(\hat{\beta})^{-1})$

■ test významnosti parametru β

$$\rightarrow (\hat{\beta} - \beta_0)^T \Sigma(\hat{\beta})^{-1} (\hat{\beta} - \beta_0) \sim \chi^2_p$$

Coxův regresní model

Vlastnosti odhadů

- asymptotické vlastnosti $\widehat{S}(t) = \widehat{S}_0(t) \exp\{\widehat{\beta}^T Z\}$

$$\rightarrow \widehat{S}(t) \sim \mathcal{N}(S(t), \widehat{\text{var}}[\widehat{S}(t)])$$

$$\rightarrow \widehat{\text{var}}[\widehat{S}(t)] = \left(\widehat{S}(t)\right)^2 e^{2\widehat{\beta}^T Z} \cdot$$

$$\cdot \left(\sum_{s \leq t} \frac{dN(s)}{(S^{(0)}(\widehat{\beta}, s))^2} + \left(\sum_{s \leq t} \frac{E(\widehat{\beta}, s) dN(s)}{S^{(0)}(\widehat{\beta}, s)} \right)^T \Sigma(\widehat{\beta})^{-1} \left(\sum_{s \leq t} \frac{E(\widehat{\beta}, s) dN(s)}{S^{(0)}(\widehat{\beta}, s)} \right) \right)$$

- asymptotické vlastnosti $\widehat{R}(s, t) = 1 - \left(\frac{\widehat{S}_0(t)}{\widehat{S}_0(s)} \right) \exp\{\widehat{\beta}^T Z\}$

$$\rightarrow \widehat{\text{var}}[\widehat{R}(s, t)] = \left(1 - \widehat{R}(s, t)\right)^2 e^{2\widehat{\beta}^T Z} \cdot$$

$$\cdot \left(\sum_{s < w \leq t} \frac{dN(w)}{(S^{(0)}(\widehat{\beta}, w))^2} + \left(\sum_{s < w \leq t} \frac{E(\widehat{\beta}, w) dN(w)}{S^{(0)}(\widehat{\beta}, w)} \right)^T \Sigma(\widehat{\beta})^{-1} \left(\sum_{s < w \leq t} \frac{E(\widehat{\beta}, w) dN(w)}{S^{(0)}(\widehat{\beta}, w)} \right) \right)$$

$$\rightarrow \widehat{\text{cov}}[\widehat{R}(s, t), \widehat{R}(u, v)] = \left(1 - \widehat{R}(s, t)\right) \left(1 - \widehat{R}(u, v)\right) e^{2\widehat{\beta}^T Z} \cdot$$

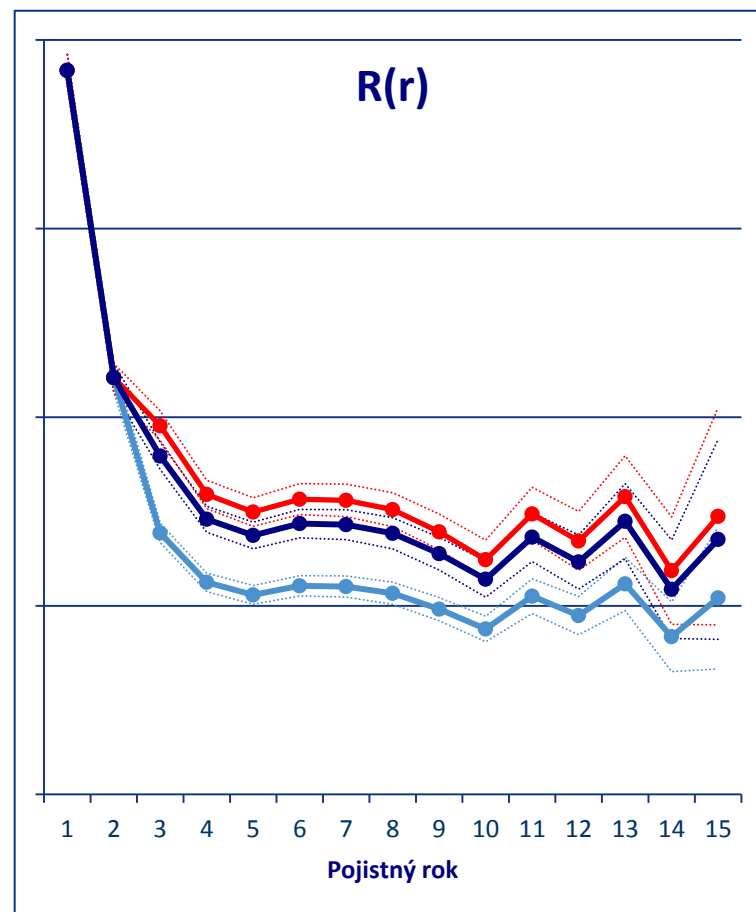
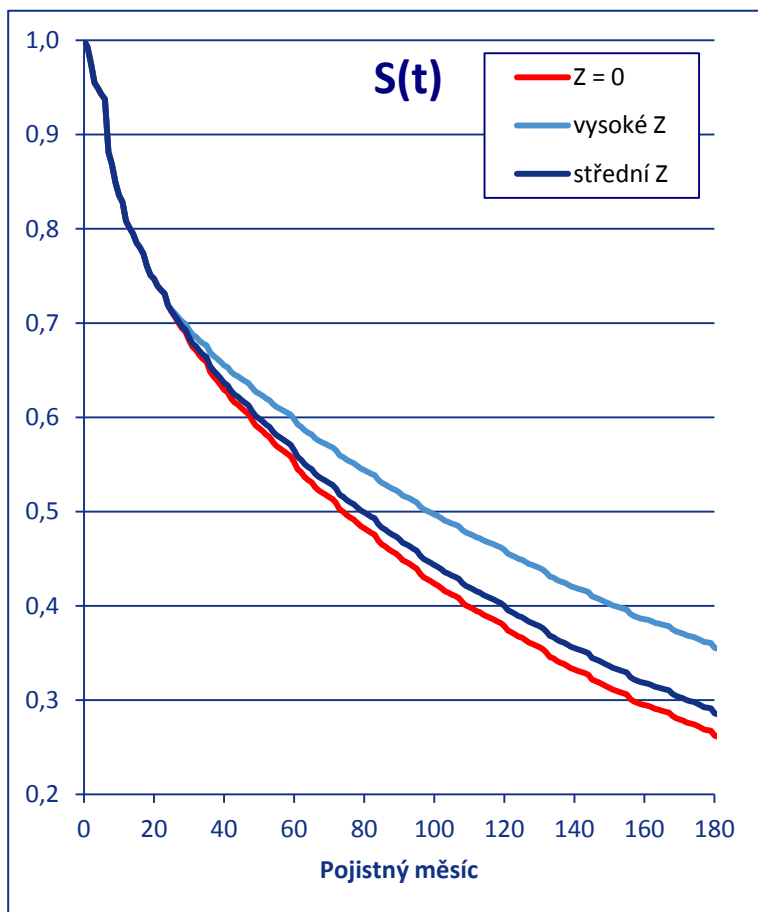
$$\cdot \left(\left(\sum_{s < w \leq t} \frac{E(\widehat{\beta}, w) dN(w)}{S^{(0)}(\widehat{\beta}, w)} \right)^T \Sigma(\widehat{\beta})^{-1} \left(\sum_{u < w \leq v} \frac{E(\widehat{\beta}, w) dN(w)}{S^{(0)}(\widehat{\beta}, w)} \right) \right) \text{ pro } s < t \leq u < v$$

Coxův regresní model

Příklad

■ $Z(t) = Z$

$\beta \cdot 10^6$	$Q_{0,5\%}[\beta \cdot 10^6]$	$Q_{99,5\%}[\beta \cdot 10^6]$
-7,4820	-9,4533	-5,5107

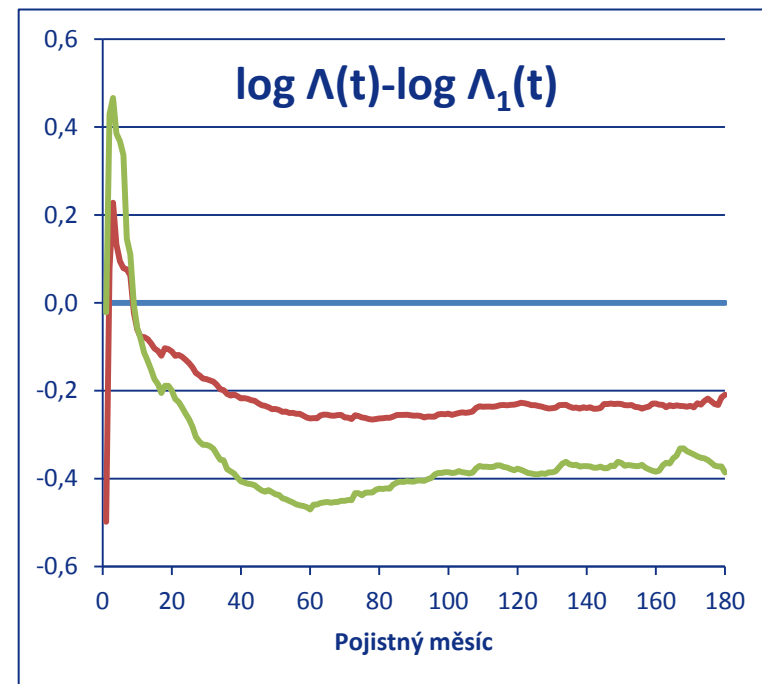
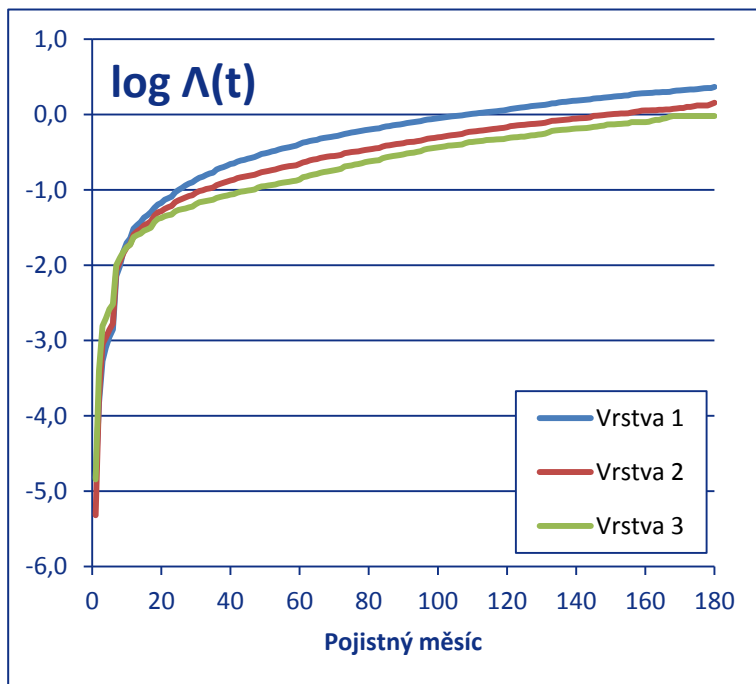


Coxův regresní model

Identifikace a verifikace modelu

■ tvar kumulativní rizikové funkce

- $\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}^T \mathbf{Z}_i\}$ $\Lambda_i(t) = \Lambda_0(t)\exp\{\boldsymbol{\beta}^T \mathbf{Z}_i\}$
- $\log(\Lambda_i(t)) - \log(\Lambda_j(t)) = \boldsymbol{\beta}^T (\mathbf{Z}_i - \mathbf{Z}_j)$ nezávisí na čase
→ identifikace pomocí paralelních křivek $\log(\Lambda_i(t))$



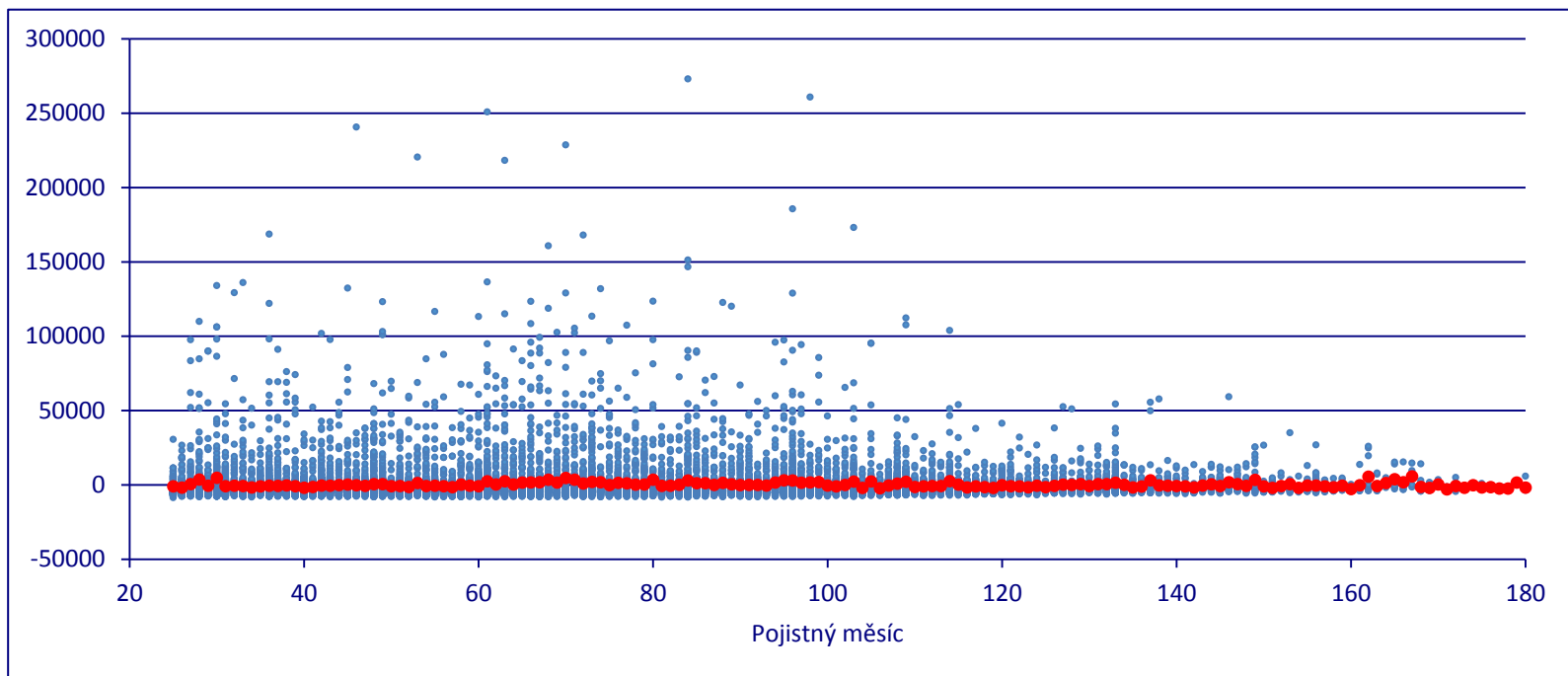
Coxův regresní model

Identifikace a verifikace modelu

■ Schoenfeldova rezidua

- $E_i(\hat{\beta}, t) = Z_i - \frac{s^{(1)}(\hat{\beta}, t)}{s^{(0)}(\hat{\beta}, t)}$ pouze pro necenzorovaná pozorování

→ neměla by vykazovat systematickou odchylku v závislosti na čase



Coxův regresní model

Shrnutí

■ široká třída modelů

- flexibilní volba $\mathbf{Z}(t)$
- obecný multiplikativní model $\lambda(t|\mathbf{Z}(t)) = \lambda_0(t)r(\boldsymbol{\beta}^T \mathbf{Z}(t))$
- aditivní model $\lambda(t|\mathbf{Z}(t)) = \beta_0(t) + \boldsymbol{\beta}(t)^T \mathbf{Z}(t)$

■ větší výpočtová náročnost

- numerické řešení rovnice pro $\hat{\boldsymbol{\beta}}$
- asymptotické kvantily a intervaly spolehlivosti
- kontrola vhodnosti modelu

■ nevýhoda neparametrických modelů (i semiparametrických typu Coxova regresního modelu)

- neposkytují odhad pro časy přesahující maximální pozorovanou dobu
→ použití parametrických modelů

Parametrické modely

Exponenciální model

■ formulace modelu

- $T \sim \text{Exp}(\lambda)$

- charakteristiky rozložení

- $f(t) = \lambda e^{-\lambda t}$ $S(t) = e^{-\lambda t}$ $R(s, t) = 1 - e^{-\lambda(t-s)}$
 - $\lambda(t) = \lambda$ $\log(S(t)) = -\lambda t$

→ identifikace pomocí lineární závislosti $\log(\hat{S}(t))$ na t

■ konstrukce odhadu

- $L(\lambda) = \prod_{i=1}^n (\lambda e^{-\lambda X_i})^{\Delta_i} (e^{-\lambda X_i})^{1-\Delta_i} = \prod_{i=1}^n \lambda^{\Delta_i} e^{-\lambda X_i}$

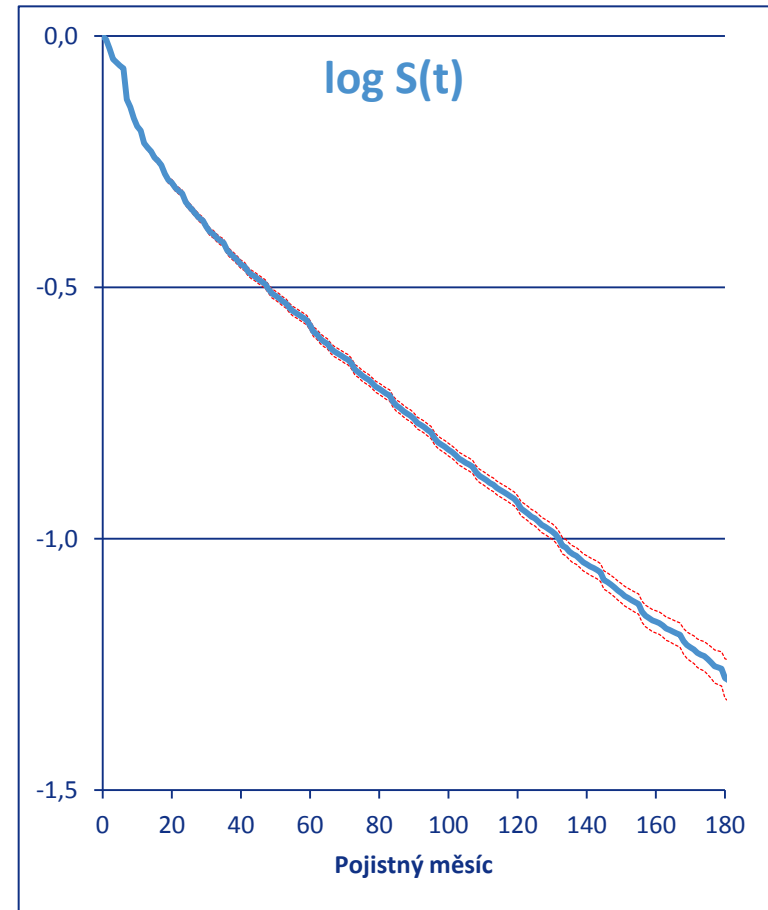
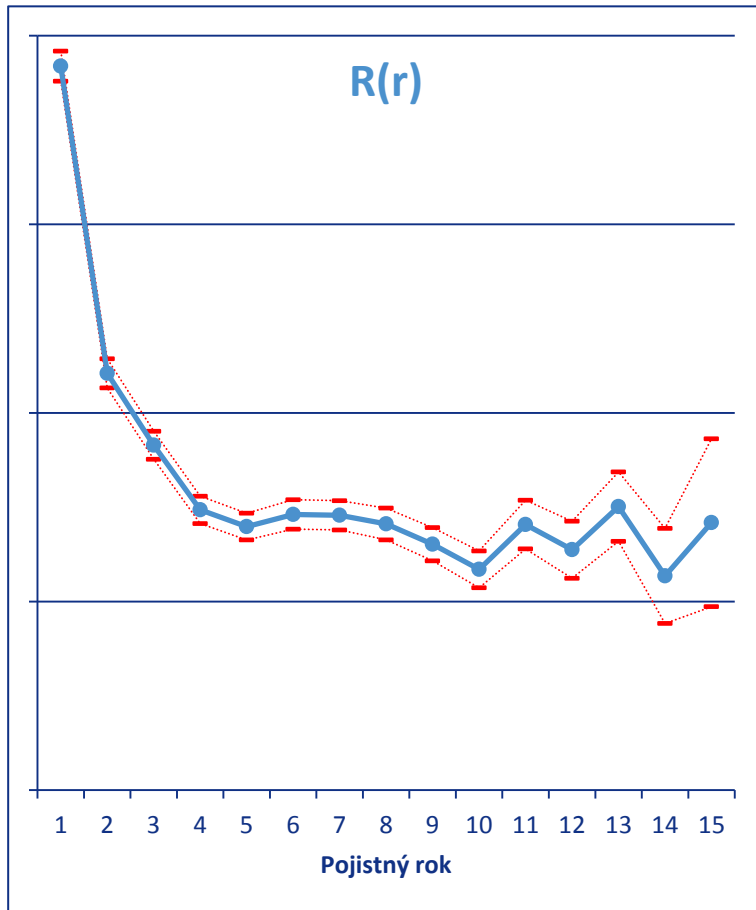
$$\rightarrow \hat{\lambda} = \frac{\max_{s \geq 0} N(s)}{\sum_{i=1}^n X_i}$$

■ asymptotické vlastnosti odhadu

$$\rightarrow \hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{\max_{s \geq 0} N(s)}{(\sum_{i=1}^n X_i)^2}\right)$$

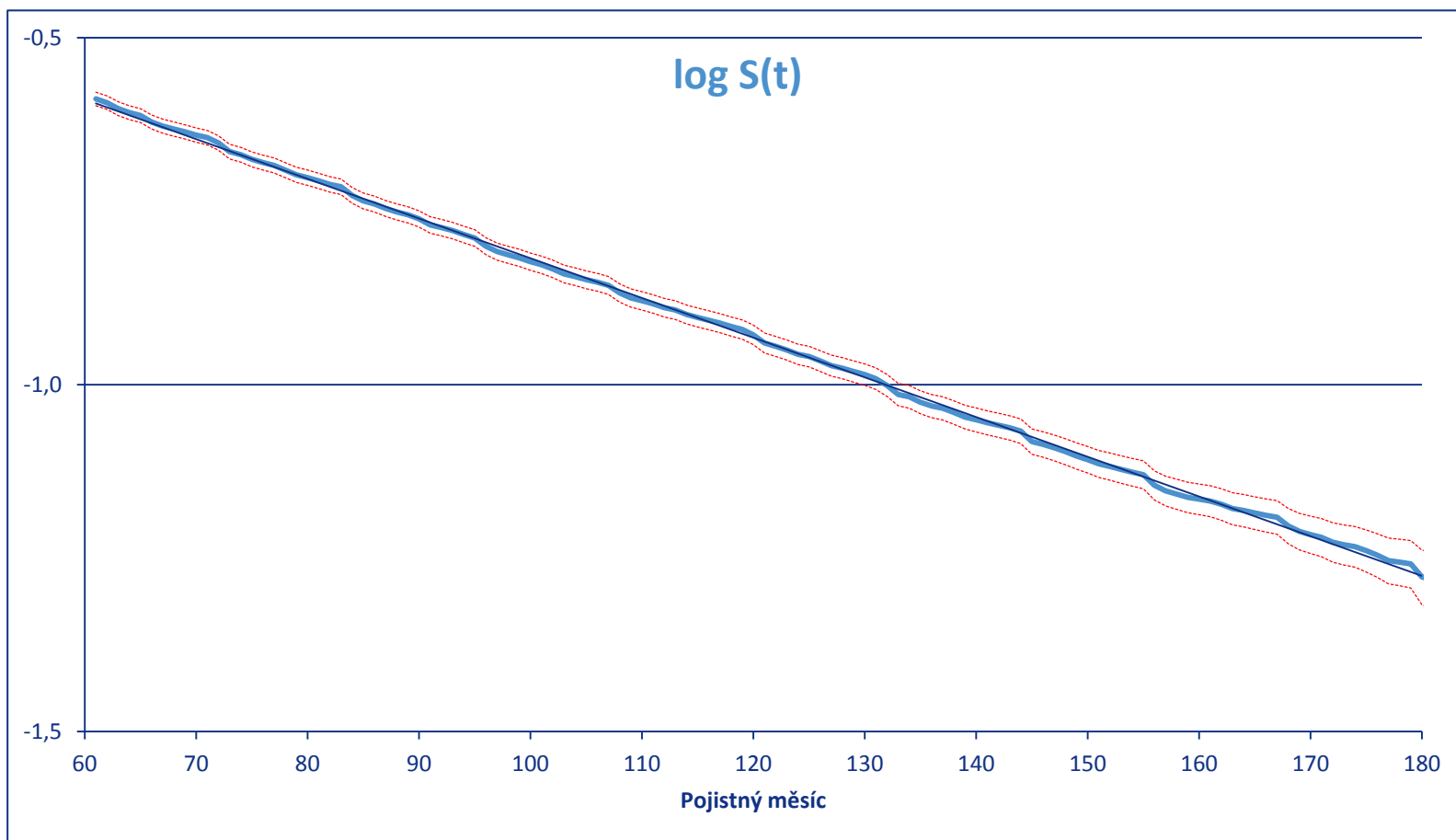
Parametrické modely

Exponenciální model: Příklad



Parametrické modely

Exponenciální model: Příklad

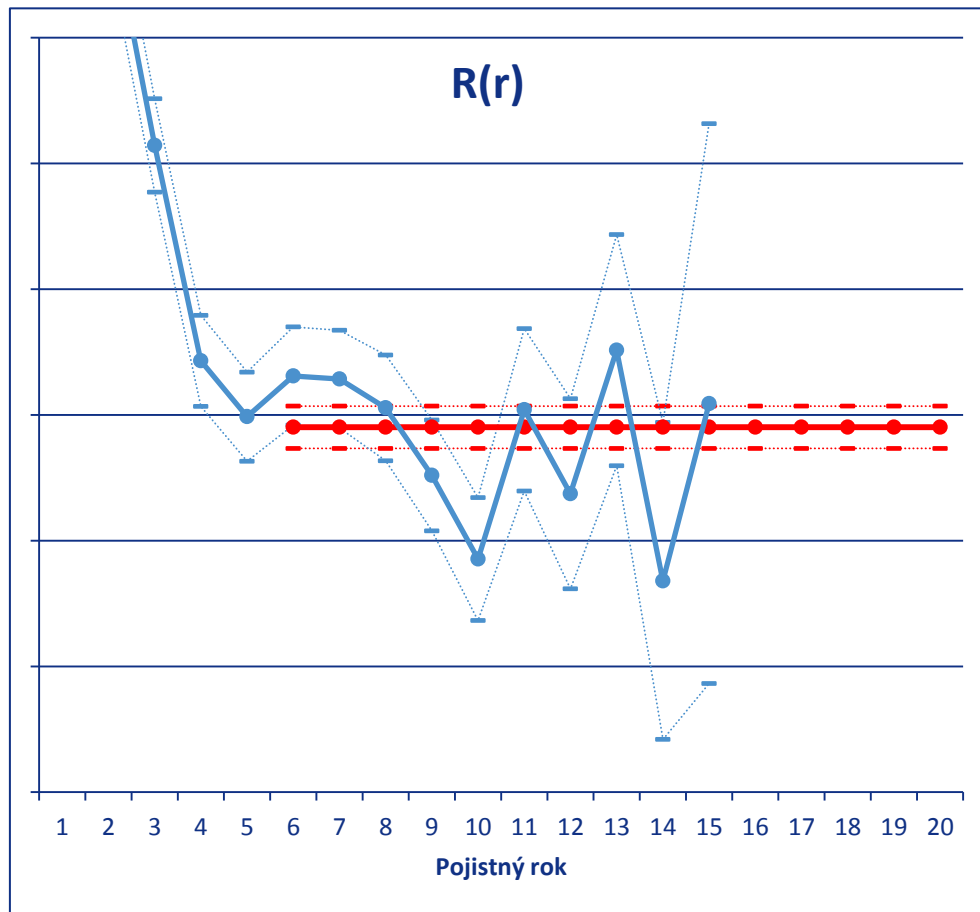


Parametrické modely

Exponenciální model: Příklad

■ Odhad parametru

λ	$Q_{0,5\%}[\lambda]$	$Q_{99,5\%}[\lambda]$
0,005960	0,005809	0,006111



Parametrické modely

Weibullův model

■ formulace modelu

- $T \sim \mathcal{W}(\lambda, \gamma)$
- charakteristiky rozložení

- $f(t) = \lambda \gamma t^{\gamma-1} e^{-\lambda t^\gamma}$
- $\lambda(t) = \lambda \gamma t^{\gamma-1}$

$$S(t) = e^{-\lambda t^\gamma} \quad R(s, t) = 1 - e^{-(\lambda t^\gamma - \lambda s^\gamma)}$$
$$\log(-\log(S(t))) = \gamma \log t + \log \lambda$$

→ identifikace pomocí lineární závislosti $\log(-\log(\hat{S}(t)))$ na $\log(t)$

■ konstrukce odhadu

- $L(\lambda, \gamma) = \prod_{i=1}^n (\lambda \gamma (X_i)^{\gamma-1} e^{-\lambda (X_i)^\gamma})^{\Delta_i} (e^{-\lambda (X_i)^\gamma})^{1-\Delta_i}$

$$\rightarrow \lambda - \frac{\max_{s \geq 0} N(s)}{\sum_{i=1}^n (X_i)^\gamma} = 0$$

$$\rightarrow \frac{\sum_{i=1}^n (X_i)^\gamma \log(X_i)}{\sum_{i=1}^n (X_i)^\gamma} - \frac{1}{\gamma} - \frac{\sum_{s \geq 0} \log(s) dN(s)}{\max_{s \geq 0} N(s)} = 0$$

Parametrické modely

Regresní model proporcionálních rizik

■ tvar rizikové funkce

- $\lambda(t|\mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}^T \mathbf{Z}}$

■ exponenciální regresní model

- $\lambda(t|\mathbf{Z}) = \lambda e^{\boldsymbol{\beta}^T \mathbf{Z}}$

- lze psát ve tvaru $\lambda(t|\mathbf{Z}) = e^{\boldsymbol{\beta}^T \mathbf{Z}}$, při $Z_0 \equiv 1$ je $\lambda = e^{\beta_0}$

- $S(t|\mathbf{Z}) = \exp\{-te^{\boldsymbol{\beta}^T \mathbf{Z}}\}$

- konstrukce odhadu

$$\rightarrow \sum_{i=1}^n \mathbf{Z}_i \left(\Delta_i - X_i e^{\boldsymbol{\beta}^T \mathbf{Z}_i} \right) = 0$$

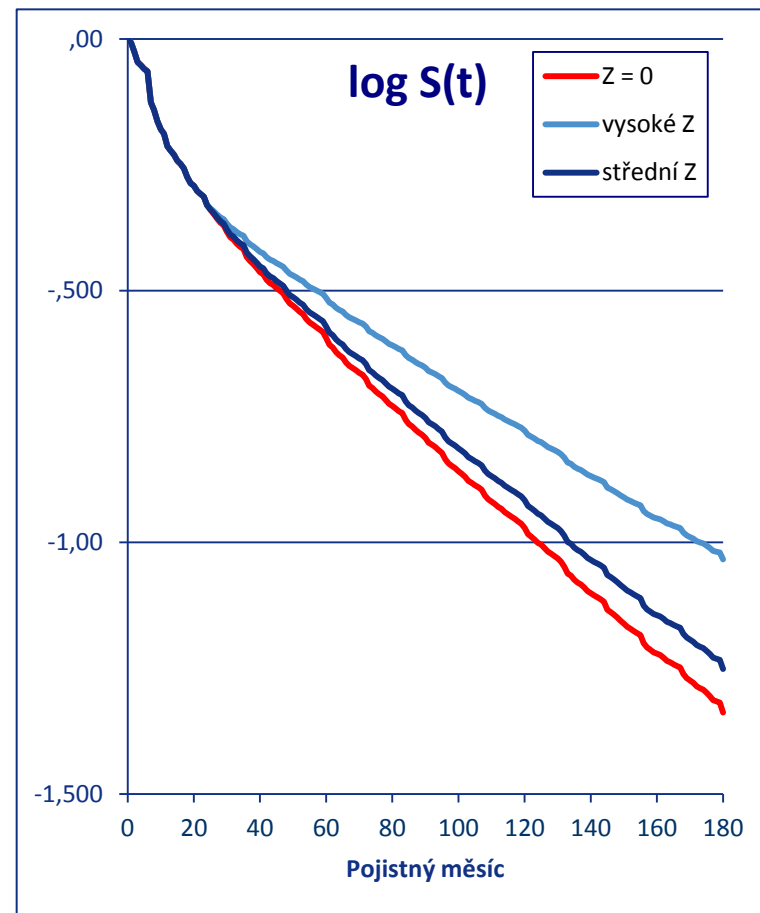
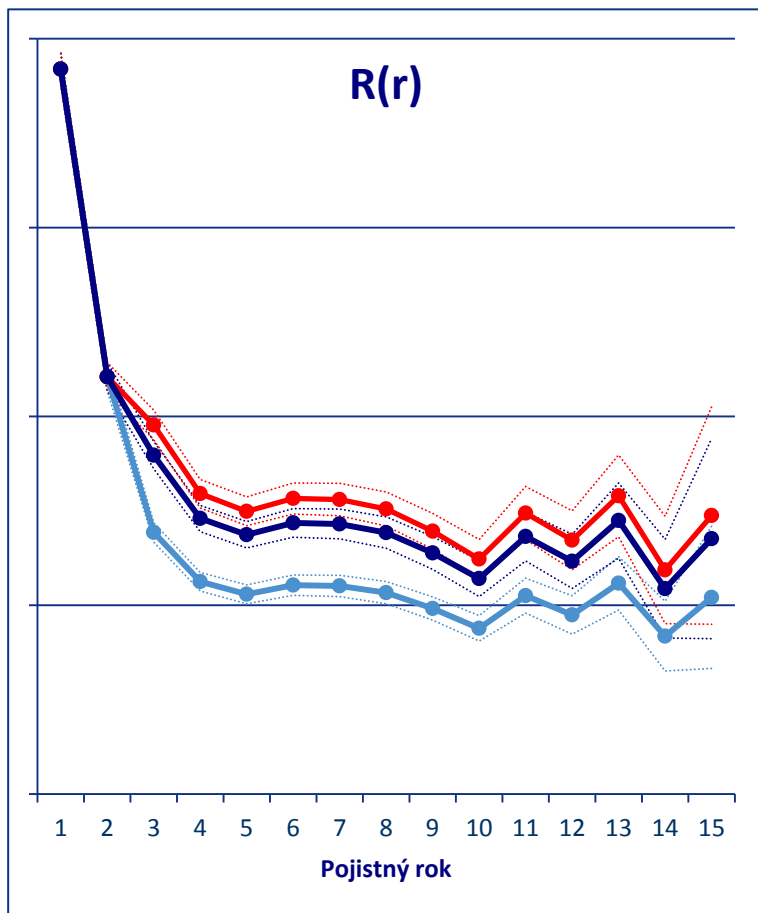
- asymptotické vlastnosti odhadu

- $\Sigma(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T X_i e^{\boldsymbol{\beta}^T \mathbf{Z}_i}$

$$\rightarrow \hat{\boldsymbol{\beta}} \sim \mathcal{N} \left(\boldsymbol{\beta}, \Sigma(\hat{\boldsymbol{\beta}})^{-1} \right)$$

Parametrické modely

Regresní model proporcionálních rizik: Příklad



Parametrické modely

Regresní model proporcionálních rizik: Příklad

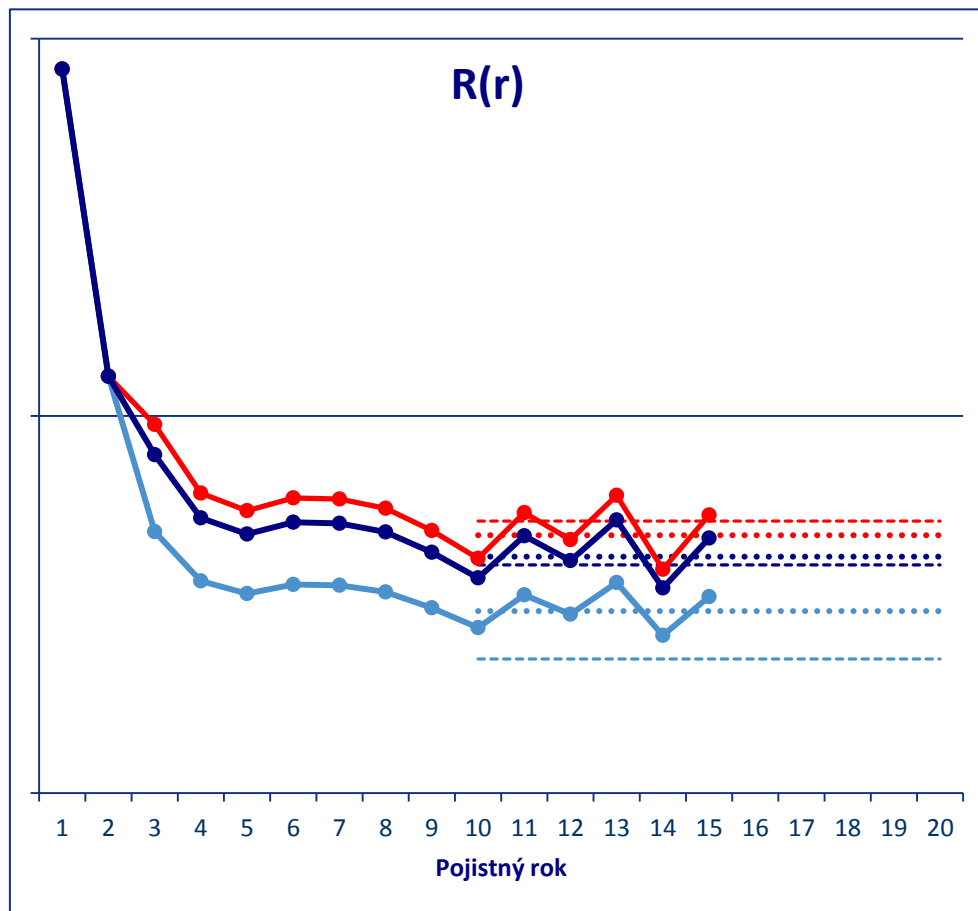
■ Odhady parametrů

- odhad β z
původního modelu

λ	$Q_{0,5\%}[\lambda]$	$Q_{99,5\%}[\lambda]$
0,005901	0,005483	0,006351
$\beta \cdot 10^6$	$Q_{0,5\%}[\beta \cdot 10^6]$	$Q_{99,5\%}[\beta \cdot 10^6]$
-7,4820	-9,4533	-5,5107

- odhad obou
parametrů

λ	$Q_{0,5\%}[\lambda]$	$Q_{99,5\%}[\lambda]$
0,006238	0,005778	0,006734
$\beta \cdot 10^6$	$Q_{0,5\%}[\beta \cdot 10^6]$	$Q_{99,5\%}[\beta \cdot 10^6]$
-15,1365	-23,1984	-7,0746



Parametrické modely

Regresní model zrychlených časů

■ tvar rizikové funkce a funkce přežití

- $\lambda(t|\mathbf{Z}) = e^{\beta^T \mathbf{Z}} \lambda_0 \left(t e^{\beta^T \mathbf{Z}} \right)$
- $S(t|\mathbf{Z}) = \exp \left\{ -\Lambda_0 \left(t e^{\beta^T \mathbf{Z}} \right) \right\} = S_0 \left(t e^{\beta^T \mathbf{Z}} \right)$

■ speciálně v případě Wiebullova (a exponenciálního) rozložení je současně modelem proporcionálních rizik

- $\lambda(t|\mathbf{Z}) = \lambda_\gamma \left(t e^{\beta^T \mathbf{Z}} \right)^{\gamma-1} e^{\beta^T \mathbf{Z}}$
- lze přepsat do tvaru $\lambda(t|\mathbf{Z}) = \lambda_\gamma t^{\gamma-1} e^{\beta^T \mathbf{Z}}$

■ konstrukce odhadů Weibullova regresního modelu

$$\begin{aligned} \rightarrow \lambda - \frac{\max_{s \geq 0} N(s)}{\sum_{i=1}^n (X_i)^\gamma \exp\{\beta^T \mathbf{Z}_i\}} &= 0 \\ \rightarrow \frac{\sum_{i=1}^n (X_i)^\gamma \log(X_i)}{\sum_{i=1}^n (X_i)^\gamma} - \frac{1}{\gamma} - \frac{\sum_{s \geq 0} \log(s) dN(s)}{\max_{s \geq 0} N(s)} &= 0 \\ \rightarrow \sum_{i=1}^n \mathbf{Z}_i \left(\Delta_i - \lambda X_i^\gamma e^{\beta^T \mathbf{Z}_i} \right) &= 0 \end{aligned}$$

Parametrické modely

Shrnutí

■ další modely

- logaritmicko-normální rozložení $f(t) = \frac{1}{\sigma t \sqrt{2\pi}} e^{-\frac{(\log t - \mu)^2}{2\sigma^2}}$
- gama rozložení $f(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-\frac{t}{\theta}}$
- za rizikovou funkcí λ je možno zvolit i jiné funkce

■ výhody

- možnost extrapolace na delší časové období
- odhad jen několika parametrů namísto celé funkce přežití

■ nevýhody

- riziko špatné identifikace modelu
- některá data nemusejí odpovídat žádnému z obvyklých statistických rozložení
- možná výpočtová náročnost (pro některé modely je nutné hledat odhady parametrů numericky)

- **řada metod v závislosti na charakteru dat**
 - základem analýzy neparametrické modely
 - zahrnutí regrese
 - parametrické modely
 - další rozšíření (např. analýza opakovaných událostí)
- **možnost využití softwaru**
 - statistické programy
 - pro menší objemy dat MS Excel

Literatura

- Andersen, Borgan, Gill, Keiding (1993): Statistical Models Based on Counting Processes
- Fleming, Harrington (1991): Counting Processes and Survival Analysis
- Kalbfleisch, Prentice (2002): The Statistical Analysis of Failure Time Data
- Cox (1972): Regression Models and Life Tables
- Kaplan, Meier (1958): Nonparametric Estimation from Incomplete Observations

Kontakt

■ martin.hrba@axa.cz