Vybrané kapitoly z analýzy prežívania

Regresné modely

Stanislav Katina[1]

[1]Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita v Brne

ZS 2013

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Poďakovanie

Tento učební text vznikl za přispění Evropského sociálního fondu a státního rozpočtu ČR prostřednictvím Operačního programu Vzdělávání pro konkurenceschopnost v rámci projektu Univerzitní výuka matematiky v měnícím se světě

(CZ.1.07/2.2.00/15.0203)

Parametrické regresné modely

Prehľad rozdelení pravdepodobnosti

1. **Exponenciálne rozdelenie**
2. **Weibullovo rozdelenie**
3. **Rozdelenie extrémnych (minimálnych) hodnôt**
4. **Log-normálne rozdelenie**
5. **Log-logistické rozdelenie**
6. **Gama rozdelenie**

Parametrické regresné modely

Exponenciálne rozdelenie

Náhodná premenná T má ***exponenciálne rozdelenie*** $Exp(\lambda)$ práve vtedy, ak jej **hustota** má tvar

$$f(t,\lambda) = \lambda e^{-\lambda t}, \text{ kde } \lambda > 0, t \geq 0$$

a naviac

- **funkcia rizika** $\lambda(t,\lambda) = \lambda$
- **distribučná funkcia** $F(t,\lambda) = \int_0^t \lambda e^{-\lambda t} = 1 - e^{-\lambda t}$
- **funkcia prežívania** $S(t,\lambda) = e^{-\lambda t}$
- **stredná hodnota** $E[T] = \frac{1}{\lambda}$
- **rozptyl** $Var[T] = \frac{1}{\lambda^2}$
- p**-ty kvantil** $t_p = -\frac{1}{\lambda}\log(1-p)$

Parametrické regresné modely
Exponenciálne rozdelenie

- jednoduchosť modelu – **riziko je konštantné**, teda
  - riziko nezávisí na t a
  - pravdepodobnosť zlyhania v intervale $[y, y+\delta y]$ je nezávislá na tom, aký dlhý čas jedinec doteraz prežil
- **riziko bez pamäte** – obmedzenie v praxi, nakoľko riziko zvyčajne s časom narastá
- **podmienka konštantnosti rizika** – odhad kumulatívneho rizika zobraziť voči t – graf bude potom predstavovať priamku
  - **kumulatívne riziko** $\Lambda(t) = -\log S(t)$
  - pre exponenciálne rozdelenie $\log(\Lambda(t)) = \log(-\log S(t)) = \log(\lambda) + \log(t)$, kde $\log(t) = -\log(\lambda) + \log(-\log S(t))$
  - graf $\log(t)$ voči $\log(-\log S(t))$ je priamka so **sklonom** rovným jednej a **interceptom** $-\log(\lambda)$
  - graf času t voči riziku λ je **horizontálna** priamka
- **medián času prežívania** je daný riešením rovnice $F(t, \lambda) = \frac{1}{2}$, a teda $t_{0.5} = \frac{1}{\lambda}\log 2$
- exponenciálny model je **špeciálnym prípadom Weibullovho a Gama rozdelenia s parametrom tvaru rovným jednej**

Parametrické regresné modely
Exponenciálne rozdelenie

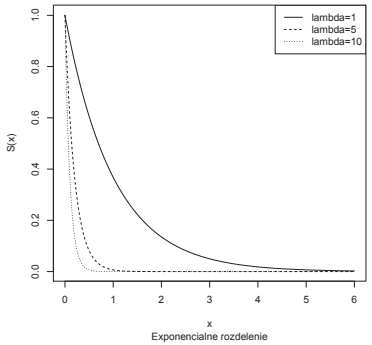

Figure: Funkcie prežívania (Exponenciálne rozdelenie)

Parametrické regresné modely
Weibullovo rozdelenie

Náhodná premenná T má ***Weibullovo rozdelenie*** $W(\lambda, \theta)$ práve vtedy, ak jej **hustota** má tvar

$$f(t, \lambda, \theta) = \lambda\theta(\lambda t)^{\theta-1} e^{-(\lambda t)^{\theta}}, \text{kde } \lambda > 0, \theta > 0, t \geq 0$$

a naviac

- **funkcia rizika** $\lambda(t, \lambda, \theta) = \lambda\theta(\lambda t)^{\theta-1}$
- **distribučná funkcia** $F(t, \lambda, \theta) = 1 - e^{-(\lambda t)^{\theta}}$
- **funkcia prežívania** $S(t, \lambda, \theta) = e^{-(\lambda t)^{\theta}}$
- **stredná hodnota** $E[T] = \int_0^{\infty} \theta(\lambda t)^{\theta} e^{-(\lambda t)^{\theta}} = \lambda^{-1}\Gamma\left(1 + \frac{1}{p}\right)$
- **rozptyl** $Var[T] = \lambda^{-2}\left[\Gamma\left(1 + \frac{2}{\theta}\right) - \Gamma^2\left(1 + \frac{1}{\theta}\right)\right]$
- p**-ty kvantil** $t_p = \frac{1}{\lambda}(-\log(1-p))^{\frac{1}{\theta}}$

Parametrické regresné modely
Weibullovo rozdelenie

- $\Gamma(x)$ predstavuje **Gama funkciu** a je definovaná ako $\int_0^{\infty} u^{x-1}e^{-u}du, x > 0$
- parameter θ sa nazýva **parameter tvaru**
- **funkcia rizika** je
  - **rastúca**, keď je $\theta > 1$
  - **klesajúca**, keď $\theta < 1$
  - **konštantná**, keď $\theta = 1$
- parameter λ sa nazýva **škála**, nakoľko jeho hodnoty menia iba škálu na horizontálnej časovej t osi a nie tvar funkcie
- Weibullov model je veľmi flexibilný a ukázalo sa, že je v praxi často aplikovateľný
- často očakávame **rastúce riziko** s časom, ako napr. modelovanie času prežívania pacientov s leukémiou, ktorí neodpovedajú na liečbu, kde je udalosťou smrť
- avšak môžeme mať aj opačnú situáciu, teda **klesajúce riziko**, ak sa pacienti napr. zotavujú po chirurgickom zákroku

- pre Weibullovo rozdelenie
  $\log(\Lambda(t)) = \log(-\log S(t)) = \theta(\log(\lambda) + \log(t))$, kde
  $\log(t) = -\log(\lambda) + \sigma \log(-\log S(t))$, kde $\sigma = \frac{1}{\theta}$
- graf $\log(t)$ voči $\log(-\log S(t))$ je priamka so **sklonom** rovným $\sigma = \frac{1}{\theta}$ a **interceptom** $-\log(\lambda)$
- **medián času prežívania** je daný riešením rovnice $F(t, \lambda, \theta) = \frac{1}{2}$, a teda $t_{0.5} = \frac{1}{\lambda}(-\log 2)^{\frac{1}{\theta}}$
- Weibullovo rozdelenie je úzko previazané s **rozdelením extrémnych hodnôt** [extreme value distribution] – **logaritmická transformácia** Weibullovej náhodnej premennej nám dáva náhodnú premennú majúcu rozdelenie extrémnych hodnôt

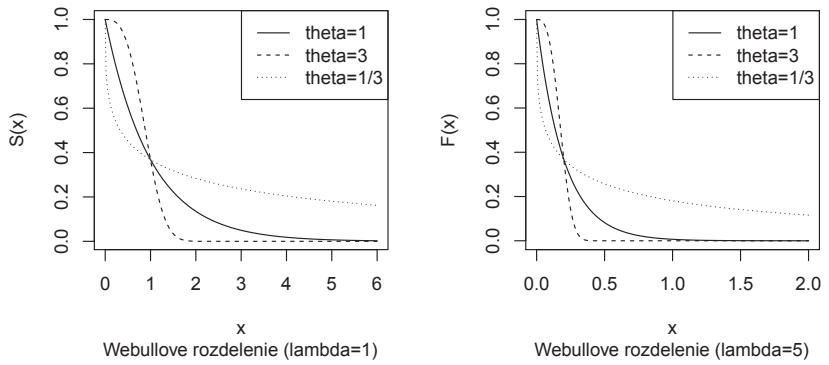

Figure: Funkcie prežívania (Weibullove rozdelenie)

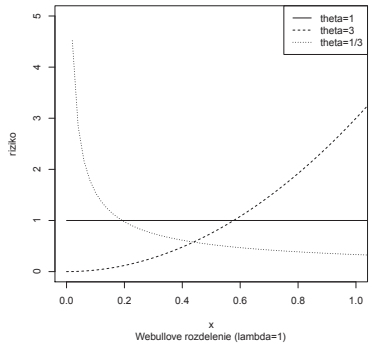

Figure: Funkcie rizika (Weibullove rozdelenie)

Nech $\mu \in \mathbb{R}$ a $\sigma > 0$ sú parametre polohy a škály – ***štandardizované rozdelenie extrémnych hodnôt*** má $\mu = 0$ a $\sigma = 1$; potom

- **hustota** $f(t, \mu, \sigma) = \frac{1}{\sigma} \exp(\frac{t-\mu}{\sigma} - \exp\frac{t-\mu}{\sigma})$
- **distribučná funkcia** $F(t, \mu, \sigma) = 1 - \exp(-\exp\frac{t-\mu}{\sigma})$
- **funkcia prežívania** $S(t, \mu, \sigma) = \exp(-\exp\frac{t-\mu}{\sigma})$
- **stredná hodnota** $E[T] = \mu - \gamma\sigma$
- **rozptyl** $Var[T] = \frac{\pi^2}{6}\sigma^2$
- ***p*-ty kvantil** $t_p = \mu + \sigma \log(-\log(1-p))$

kde $\gamma \doteq 0.5772$ je Eulerovo číslo, parameter polohy μ je 0.632-tý kvantil a $t \in \mathbb{R}$

- potom ak T je Weibullova náhodná premenná s paramatrami θ a λ, $Y = \log(T)$ má rozdelenie extrémnych hodnôt s parametrami $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$
- naviac $Y = \mu + \sigma Z$, kde Z má štandardizované rozdelenie extrémnych hodnôt
- parametre μ a σ nemenia tvar rozdelenia, iba polohu a škálu

Parametrické regresné modely
Log-normálne rozdelenie

Ak je čas prežívania T ***log-normálne*** *rozdelený*, potom $Y = \log(T)$ je normálne rozdelené s parametrami μ a σ, teda $Y \sim N(\mu, \sigma^2)$. Nech $Y = \mu + \sigma Z$, potom $Z \sim N(0,1)$. Nech $\theta, \lambda > 0$, potom

- **hustota** $f(t,\theta,\lambda) = \frac{1}{\sqrt{2\pi}}\theta t^{-1}\exp(\frac{-\theta^2(\log(\lambda t))^2}{2})$
- **distribučná funkcia** $F(t,\theta,\lambda) = \Phi(\theta\log(\lambda t))$
- **funkcia prežívania** $S(t,\theta,\lambda) = 1 - \Phi(\theta\log(\lambda t))$
- **funkcia rizika** $\lambda(t,\theta,\lambda) = \frac{f(t,\theta,\lambda)}{S(t,\theta,\lambda)}$
- **stredná hodnota** $E[T] = \exp(\mu + \frac{\sigma^2}{2})$
- **rozptyl** $Var[T] = (\exp(\sigma^2) - 1)(\exp(2\mu + \sigma^2))$

kde $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$

- **funkcia rizika** sa rovná nule v čase $t = 0$, **rastie** do maxima, potom **klesá** smerom k nule, keď t dosahuje veľké hodnoty – riziko klesá pre veľké hodnoty t, čo sa môže zdať ako nepoužiteľná vlastnosť tohoto rozdelenia, napr. pri modelovaní prežívania tuberkulóznych pacientov, kde ich zlyhávanie najprv stúpa a potom neskôr klesá, tento model vhodný je
- dobrou aproximáciou log-normálneho rozdelenia je **log-logistické rozdelenie**

Parametrické regresné modely
Log-logistické rozdelenie

Ak je čas prežívania T ***log-logisticky*** *rozdelený*, potom $Y = \log(T)$ je logisticky rozdelené s parametrom **polohy** μ a **škály** σ. Nech $Y = \mu + \sigma Z$, potom Z má štandardizované logistické rozdelenie s hustotou

$$f(z) = \frac{\exp z}{(1+\exp z)^2}, z \in \mathbb{R},$$

ktorá je symetrická, jej chvosty sú o niečo ťažšie ako chvosty normálneho rozdelenia, šikmosť a špicatosť sú rovné 1.2, $E[Z] = 0$ a $Var[Z] = \frac{\pi^2}{3}$. Potom pre T a $\lambda, \theta > 0$ platí

- **hustota** $f(t,\lambda,\theta) = \lambda\theta(\lambda t)^{\theta-1}(1+(\lambda t)^\theta)^{-2}$
- **distribučná funkcia** $F(t,\lambda,\theta) = 1 - \frac{1}{1+(\lambda t)^\theta}$
- **funkcia prežívania** $S(t,\lambda,\theta) = \frac{1}{1+(\lambda t)^\theta}$
- **funkcia rizika** $\lambda(t,\lambda,\theta) = \frac{\lambda\theta(\lambda t)^{\theta-1}}{1+(\lambda t)^\theta}$
- ***p*-ty kvantil** $t_p = \lambda^{-1}(\frac{p}{1-p})^{\frac{1}{\theta}}$

kde $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$

Parametrické regresné modely
Log-logistické rozdelenie

- tento model sa stal populárnym, podobne ako Weibullov model, aj pre jednoduché vyjadrenie funkcie prežívania a funkcie rizika
- aproximácia cenzurovaných dát pomocou tohoto modelu sa v praxi ukázala lepšia ako aproximácia log-normálny rozdelením (okrem extrémov v chvostoch rozdelenia)
- **riziková funkcia je podobná Weibullovej**, líšia sa len menovateľom $1 + (\lambda t)^\theta$.
- ak $\theta < 1 (\sigma > 1)$ je riziková funkcia monotónne **klesajúca** z ∞ a
- ak $\theta = 1 (\sigma = 1)$, je riziková funkcia monotónne **klesajúca** z λ
- Ak $\theta > 1 (\sigma < 1)$ je riziková funkcia monotónne **rastúca** z nuly do maxima v $t = \frac{(\theta-1)^{\frac{1}{\theta}}}{\lambda}$ a **klesajúca** potom k nule
- dá sa ľahko ukázať, že **šanca prežitia za časom** t sa dá zapísať ako

$$\frac{S(t)}{1-S(t)} = (\lambda t)^{-\theta}$$

a potom $\log t$ je lineárnou funkciou logaritmu šance prežívania za časom t, teda $\log t = \mu + \sigma(-\log\frac{S(t)}{1-S(t)})$, kde $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$

- graf $\log t$ oproti $-\log\frac{S(t)}{1-S(t)}$ je priamka so **sklonom** σ a **interceptom** μ

Parametrické regresné modely
Log-logistické rozdelenie

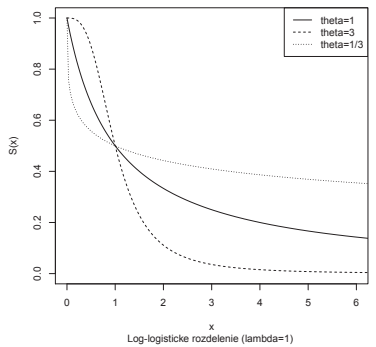

Figure: Funkcie prežívania (Log-logistické rozdelenie)

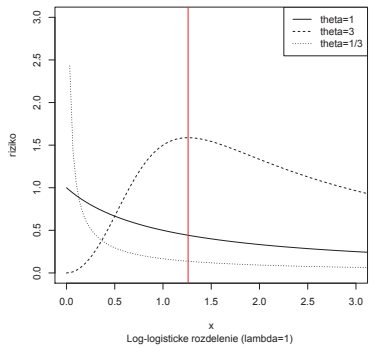

Figure: Funkcie rizika (Log-logistické rozdelenie)

Náhodná premenná T má **Gama rozdelenie** $Gama(\lambda, k)$ práve vtedy, ak jej **hustota** má tvar

$$f(t) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t}, \text{ kde } t \geq 0, \lambda > 0, k > 0.$$

Potom ak neúplná Gama funkcia $I_k(s) = \left(\int_0^s t^{k-1} e^{-t} dt \right) / \Gamma(k)$, tak

- **funkcia rizika** $\lambda(t, \lambda, k) = \frac{\lambda(\lambda t)^{k-1} e^{-\lambda t} \Gamma^{-1}(k)}{1 - I_k(\lambda t)}$
- **distribučná funkcia** $F(t, \lambda, k) = I_k(\lambda t)$
- **funkcia prežívania** $S(t, \lambda, k) = 1 - I_k(\lambda t)$
- **stredná hodnota** $E[T] = \frac{k}{\lambda}$
- **rozptyl** $Var[T] = \frac{k}{\lambda^2}$

- ak $k > 1$ funkciu rizika monotónne **rastie** z 0
- ak $k < 1$ funkciu rizika monotónne **klesá** z ∞
- ak $k = 1$, potom sa Gama rozdelenie redukuje na **exponenciálne rozdelenie** a riziková funkcia je **konštantná**, t.j. rovná λ
- **model** pre $Y = \log T$ môžeme písať v tvare $Y = \mu + Z$, kde Z má hustotu $\frac{\exp(kz - \exp(z))}{\Gamma(k)}$
- náhodná premenná Y sa nazýva **log-Gama náhodná premenná** s parametrami k a $\mu = -\log \lambda$
- Z má negatívne zošikmené rozdelenie s klesajúcou šikmosťou ak k rastie
- ak $k = 1$ potom ide o **exponenciálny model** a Z má štandardizované rozdelenie extrémnej hodnoty
- s výnimkou $k = 1$ **nie je Gama rozdelenie členom rodiny polohovo-škálových rozdelení, ale iba členom rodiny polohových rozdelení**

Okrem Gama rozdelenia majú všetky spomenuté rozdelenia času prežívania T vlastnosť, že **rozdelenie** $\log T$ **je členom rodiny polohovo-škálových rozdelení**. Spoločné črty spomenutých rozdelení sú nasledovné

1. rozdelenie T má dva parametre škálu λ a tvar θ
2. rozdelenie $\log T$ má dva parametre škálu $\mu = -\log \lambda$ a tvar $\sigma = 1/\theta$
3. pre všetky rozdelenia platí, že $Y = \log T = \mu + \sigma Z$, kde Z má štandardizované rozdelenie s $\mu = 0 (\lambda = 1)$ a $\sigma = 1 (\theta = 1)$
4. tieto modely sú **log-lineárnymi modelmi**

T	$Y = \log T$
Weibullove rozdelenie	rozdelenie extrémnych hodnôt
log-normálne rozdelenie	normálne rozdelenie
log-logistické rozdelenie	logistické rozdelenie

Parametrické regresné modely
Konštrukcia qq-diagramu

- nech $\widehat{S}_{KM}(t)$ je **Kaplan-Meierov odhad funkcie prežívania** v čase t
- nech $t_i, i = 1, 2, \ldots r \leq n$ sú **zoradené necenzurované časy do zlyhania**
- pre každý necenzurovaný výberový kvantil $y_i = \log t_i$ je odhadnutá **pravdepodobnosť zlyhania** $\widehat{p}_i^{(KM)} = 1 - \widehat{S}_{KM}(t_i)$
- **parametrický štandardizovaný kvantil** získame použitím $\widehat{p}_i = F_{0,1}(z_i) = \Pr(Z \leq z_i)$, kde $F_{0,1}(z_i)$ je **distribučná funkcia štandardizovaného parametrického modelu** ($\mu = 0, \sigma = 1$)
- tieto štandardizované kvantily porovnávame s kvantilmi neparametrického odhadu $\widehat{S}_{KM}(t)$ – ak je navrhovaný model adekvátny, potom graf (z_i, y_i) leží na priamke so **sklonom** σ a **interceptom** μ

t_p kvantil	$y_p = \log t_p$ kvantil	štandardizovaný kvantil z_p
Weibull r.	roz.extr.h.	$\log(-\log \widehat{S}(t_p)) = \log \Lambda(t_p) = \log(-\log(1-p))$
log-norm.r.	norm.r.	$\Phi^{-1}(p)$
log-logis.roz.	logis.r.	$= -\log \frac{\widehat{S}(t_p)}{1-\widehat{S}(t_p)} = -\log(\frac{1-p}{p})$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre necenzurované dáta

- **funkcia vierohodnosti**

$$L(\lambda|\mathbf{t}) = \prod_{i=1}^{n} \lambda \exp(-\lambda t_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} t_i)$$

- **logaritmus funkcie vierohodnosti**

$$l(\lambda|\mathbf{t}) = n \log \lambda - \lambda \sum_{i=1}^{n} t_i$$

- **maximálne vierohodné odhady (MLE)**

$$\frac{\partial l(\lambda|\mathbf{t})}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0, \text{ potom}$$

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\overline{T}}, \widehat{E[T]} = \frac{1}{\widehat{\lambda}} = \overline{T}, \widehat{Var[T]} = \frac{1}{\widehat{\lambda}^2} = \overline{T}^2$$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre necenzurované dáta

- **exaktné rozdelenie** – ak $T_i \sim Exp(\lambda)$, potom $\sum_{i=1}^{n} T_i \sim Gama(\lambda, k = n)$, potom

$$2\lambda \sum_{i=1}^{n} T_i = 2n \frac{\lambda}{\widehat{\lambda}} \sim \chi^2_{2n}$$

potom $(1-\alpha) \times 100\%$ **interval spoľahlivosti (IS) pre** λ

$$\left\{ \lambda : \lambda \in (\chi^2_{2n}(1-\alpha/2)\frac{\widehat{\lambda}}{2n}, \chi^2_{2n}(\alpha/2)\frac{\widehat{\lambda}}{2n}) \right\},$$

$(1-\alpha) \times 100\%$ **IS pre** $1/\lambda$ (IS pre strednú hodnotu $E[T]$)

$$\left\{ 1/\lambda : 1/\lambda \in (\frac{2n\overline{T}}{\chi^2_{2n}(\alpha/2)}, \frac{2n\overline{T}}{\chi^2_{2n}(1-\alpha/2)}) \right\}$$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre necenzurované dáta

- **p-ty kvantil** t_p, kde $p = F(t_p|\lambda) = 1 - \exp(-\lambda t_p)$, potom $t_p = -\log(1-p)/\lambda$; nech $\widehat{t}_p$ je MLE t_p, potom $\widehat{t}_p = -\log(1-p)/\widehat{\lambda} = -\overline{T}\log(1-p)$ a MLE mediánu je $\widehat{t}_{0.5} = -\overline{T}\log(\frac{1}{2}) = \overline{T}\log 2$
- **test pomerom vierohodnosti**

$$LR(\lambda_0|\mathbf{t}) = -2 \log \frac{L(\lambda_0|\mathbf{t})}{L(\lambda|\mathbf{t})} \sim \chi^2_1$$

kde

$$l(\lambda_0|\mathbf{t}) = n \log \lambda_0 - \lambda_0 n \overline{T}$$

a

$$l(\lambda|\mathbf{t}) = n \log \frac{1}{\overline{T}} - \frac{1}{\overline{T}} n \overline{T}$$

Example (exponenciálny model)

(pokrač.) Majme AML dáta. Ak predpokladáme, že cenzúry sú zlyhania (problém B, kap.1), potom vypočítajte MLE pre strednú hodnotu $1/\lambda$, medián $\widehat{t}_{0.5}$ ako aj 95%IS pre λ a $1/\lambda$. Otestujte $H_0 : E[T] = 30$ oproti $H_1 : E[T] \neq 30$ na $\alpha = 0.05$

Skupina	Čas po kompletný relaps (v týždňoch)	n	udalostí	cenzúr
skupina A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4
skupina B	$5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45$	12	11	1

- $u \in \mathbb{U}$ – **necenzurované pozorovania**
- $c \in \mathbb{C}$ – **cenzurované pozorovania**
- n_u **počet necenzurovaných pozorovaní**
- **vektor pozorovaní** $\mathbf{x} = (x_1, \ldots, x_n)^T$ a
- **vektor indikácií zlyhania a cenzúry** $\delta = (\delta_1, \ldots, \delta_n)^T$ (zodpovedajúci $\mathbf{x}$; 1 je zlyhanie a 0 je cenzúra)
- **Fisherova informačná matica** $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} l(\theta|\mathbf{x})\right] = nI_1(\theta)$
- **Fisherova informačná matica nejakého jedného pozorovania** x_1 $I_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} f(x_1|\theta)\right]$

- Potom platí $\widehat{\theta} \sim N_d(\theta, I^{-1}(\theta))$, kde
  - d je dimenzia $\mathbf{x}$
  - $I(\theta)$ je matica $d \times d$
  - i-ty diagonálny element $I^{-1}(\theta)$ je asymptotickým rozptylom i-teho elementu θ
  - mimodiagonálne elementy sú asymptotickými kovarianciami korešpondujúcich elementov θ
- ak je θ skalár, potom $Var(\theta) = \frac{1}{I(\theta)}$, kde $I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})]$
- **pre cenzurované dáta** ide o funkciu cenzurovaného rozdelenia G ako aj rozdelenia času prežívania F
- preto je potrebné aproximovať $I(\theta)$ **pozorovanou informačnou maticou** $I^*(\theta)$ v bode $\widehat{\theta}$, kde $I^*(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} l(\theta|\mathbf{x})\right]$
- v jednorozmernom prípade $I^*(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})\right]$, kde rozptyl $Var(\theta) = (I^*(\theta))^{-1}$

- **funkcia vierohodnosti** – vo všeobecnosti pre náhodne cenzurované dáta platí

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)^{\delta_i} S_f(x_i|\theta)^{1-\delta_i}$$

potom

$$\begin{aligned} L(\lambda|\mathbf{x}) &= \prod_{i:u\in\mathbb{U}} \lambda \exp(-\lambda x_i) \prod_{i:c\in\mathbb{C}} \exp(-\lambda x_i) \\ &= \lambda^{n_u} \exp(-\lambda \sum_{i:u\in\mathbb{U}} x_i) \exp(-\lambda \sum_{i:c\in\mathbb{C}} x_i) \\ &= \lambda^{n_u} \exp(-\lambda \sum_{i=1}^{n} x_i) \end{aligned}$$

- **logaritmus funkcie vierohodnosti**

$$l(\theta|\mathbf{x}) = \sum_{i:u\in\mathbb{U}} \log f(x_i|\theta) + \sum_{i:c\in\mathbb{C}} \log S_f(x_i|\theta),$$

$$l(\lambda|\mathbf{x}) = n_u \log\lambda - \lambda\sum_{i=1}^{n} x_i$$

- **maximálne vierohodné odhady (MLE)** – prvá a druhá derivácia $l(\lambda|\mathbf{x})$

$$\frac{\partial l(\lambda|\mathbf{x})}{\partial\lambda} = \frac{n_u}{\lambda} - \sum_{i=1}^{n} x_i, \frac{\partial^2 l(\lambda|\mathbf{x})}{\partial\lambda^2} = -\frac{n_u}{\lambda^2} = -I^*(\lambda),$$

potom

$$\widehat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} x_i}, \widehat{Var\left[\widehat{\lambda}\right]} = \frac{\widehat{\lambda}^2}{E[n_u]}, \text{ kde } E[n_u] = n\Pr(T \le C)$$

- **asymptotické rozdelenie** λ

$$\frac{\widehat{\lambda}-\lambda}{\sqrt{\widehat{Var}(\widehat{\lambda})}} \sim N(0,1), \text{ kde } \widehat{Var}(\widehat{\lambda}) \approx \frac{\widehat{\lambda}^2}{n_u} = \frac{1}{I^*(\lambda)}$$

kde $E[n_u]$ substituujeme za n_u, lebo rozdelenie $G(\cdot)$ nepoznáme. Treba si uvedomiť, že rozptyl je závislý na λ; potom $(1-\alpha)\times 100\%$ **IS pre** λ

$$\left\{\lambda : \lambda \in (\widehat{\lambda} - u_{\alpha/2}SE(\widehat{\lambda}), \widehat{\lambda} + u_{\alpha/2}SE(\widehat{\lambda}))\right\},$$

$(1-\alpha)\times 100\%$ **IS pre** $1/\lambda$

$$\left\{1/\lambda : 1/\lambda \in (1/\widehat{\lambda} - u_{\alpha/2}SE(1/\widehat{\lambda}), 1/\widehat{\lambda} + u_{\alpha/2}SE(1/\widehat{\lambda}))\right\},$$

kde $Var(1/\lambda)$ dostaneme pomocou *delta metódy*, kde $g(\lambda) = 1/\lambda$, $g'(\lambda) = -1/\lambda^2$ a $Var(1/\lambda) = \frac{1}{\lambda^2 E[n_u]} \approx \frac{1}{\lambda^2 n_u}$

- p**-ty kvantil** t_p, $\widehat{t}_p = -\log(1-p)/\widehat{\lambda} = -\frac{\sum_{i=1}^{n} x_i}{n_u}\log(1-p)$ a MLE mediánu je $\widehat{t}_{0.5} = -\frac{\sum_{i=1}^{n} x_i}{n_u}\log(\frac{1}{2})$; **rozptyl** p**-teho kvantilu**

$$\widehat{Var(\widehat{t}_p)} = \log^2(1-p)Var(1/\widehat{\lambda}) \approx \log^2(1-p)\frac{1}{\widehat{\lambda}^2 n_u}$$

$(1-\alpha)\times 100\%$ **IS pre** $t_{0.5}$

$$\left\{t_{0.5} : t_{0.5} \in (\widehat{t}_{0.5} - u_{\alpha/2}SE(\widehat{t}_{0.5}), \widehat{t}_{0.5} + u_{\alpha/2}SE(\widehat{t}_{0.5}))\right\}$$

- pomocou **delta metódy** môžeme vypočítať IS, ktorého hranice sú menej vychýlené nájdením **transformácií**, ktoré eliminujú závislosť rozptylu nejakého parametra na parametri samotnom. Napr. $g(\lambda) = \log\lambda$, $g'(\lambda) = 1/\lambda$ a potom $Var(\log\lambda) = \lambda^{-2}\frac{\lambda^2}{E[n_u]} \approx \frac{1}{n_u}$

Potom $\log(\widehat{\lambda}) \sim N(\log\lambda, 1/n_u)$, $(1-\alpha)\times 100\%$ **IS pre** $\log\lambda$

$$\left\{\log\lambda : \log\lambda \in (\log\widehat{\lambda} - u_{\alpha/2}\frac{1}{\sqrt{n_u}}, \log\widehat{\lambda} + u_{\alpha/2}\frac{1}{\sqrt{n_u}})\right\}$$

Analogicky dostaneme

$$\log\frac{1}{\widehat{\lambda}} \sim N(\log\frac{1}{\lambda}, \frac{1}{n_u}), \log\widehat{t}_{0.5} \sim N(\log t_{0.5}, \frac{1}{n_u}),$$

potom $(1-\alpha)\times 100\%$ **IS pre** $\log(1/\lambda)$

$$\left\{\log(1/\lambda) : \log(1/\lambda) \in (\log(1/\widehat{\lambda}) - u_{\alpha/2}\frac{1}{\sqrt{n_u}}, \log(1/\widehat{\lambda}) + u_{\alpha/2}\frac{1}{\sqrt{n_u}})\right\}$$

a $(1-\alpha)\times 100\%$ **IS pre** $\log t_{0.5}$

$$\left\{\log t_{0.5} : \log t_{0.5} \in (\log\widehat{t}_{0.5} - u_{\alpha/2}\frac{1}{\sqrt{n_u}}, \log\widehat{t}_{0.5} + u_{\alpha/2}\frac{1}{\sqrt{n_u}})\right\}$$

Spätnou transformáciou hraníc IS pre log(paramater) dostaneme hranice pre IS parametra ako exp(koncové body)

- **ML odhad funkcie prežívania** $\widehat{S}(t) = \exp(-\widehat{\lambda}t)$, ktorej rozdelenie môžeme dostať pomocou delta metódy

Alternatívne zoberieme **log-log transformáciu**, ktorá zvyčajne urýchli konvergenciu k normalite (kvôli nezávislosti rozptylu na neznámom parametri λ). Vieme, že $\log(\widehat{\lambda}) \sim N(\log\lambda, 1/n_u)$, potom

$$\log(-\log\widehat{S}(t)) = \log(\widehat{\lambda}) + \log t$$

a pre rozptyl platí

$$Var(\widehat{\log(-\log\widehat{S}(t))}) = Var(\widehat{\log(\widehat{\lambda})}) \approx \frac{1}{n_u}$$

Z delta metódy pre každé fixované t platí

$$\log(-\log\widehat{S}(t)) \sim N\left(\log(-\log S(t)), \frac{1}{n_u}\right), \text{ kde } \log(-\log S(t)) = \log(\lambda t)$$

$(1-\alpha)\times 100\%$ **IS pre** $S(t)$

$$\left\{S(t) : S(t) \in \left(\exp\left(\log\widehat{S}(t)\exp\left(\frac{u_{\alpha/2}}{\sqrt{n_u}}\right)\right), \exp\left(\log\widehat{S}(t)\exp\left(\frac{-u_{\alpha/2}}{\sqrt{n_u}}\right)\right)\right)\right\}$$

- **test pomerom vierohodnosti**

$$LR(\lambda_0) = -2\log\frac{L(\lambda_0)}{L(\lambda)} \sim \chi^2_1,$$

kde

$$l(\lambda_0|\mathbf{x}) = n_u\log\lambda_0 - \lambda_0\sum_{i=1}^{n} x_i$$

a

$$l(\lambda|\mathbf{x}) = n_u\log\frac{n_u}{\sum_{i=1}^{n} x_i} - n_u$$

Example (exponenciálne rozdelenie)

(pokrač.) Majme AML dáta. Vypočítajte MLE pre strednú hodnotu $1/\lambda$, medián $\widehat{t}_{0.5}$ ako aj 95%IS pre λ a $1/\lambda$. Otestujte $H_0 : E[T] = 30$ oproti $H_1 : E[T] \neq 30$ na $\alpha = 0.05$.

Skupina	Čas po kompletný relaps (v týždňoch)	n	udalostí	cenzúr
skupina A	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+	11	7	4
skupina B	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45	12	11	1

Example (exponenciálne rozdelenie)

(pokrač.) Použitím funkcie `survreg` pre cenzurované dáta zopakujte výpočty urobené ručne, teda vypočítajte bodové odhady strednej hodnoty a mediánu ako aj ich 95%IS a zostrojte qq-diagram. Nakreslite funkcie prežívania pre exponenciálne, Weibullovo a log-logistické rozdelenie a porovnajte ich s $\widehat{S}_{KM}(t)$.

Pozn.: **Exponenciálne rozdelenie** času t je Weibullovo rozdelenie s parametrom tvaru $\theta = 1$ alebo $\log t$ má rozdelenie extrémnej hodnoty so škálou $\sigma = 1$. Funkcia `survreg` fituje $\log t$ a výstupom je $\widehat{\mu} = -\log\widehat{\lambda}$ (parameter polohy rozdelenia extrémnej hodnoty). Pre **Weibullove rozdelenie** $\widehat{\lambda} = \exp(-\widehat{\mu})$ a odhad $1/\widehat{\lambda} = \exp(\widehat{\mu})$ a naviac $\widehat{\sigma} = 1/\widehat{\theta}$. Funkcia `summary(fit)`, kde `fit` je vypočítaný pomocou funkcie `survreg`, nám poskytuje práve $\widehat{\mu}$ a $\widehat{\sigma}$. Vo funkcii `survreg` treba špecifikovať rozdelenie `dist="weibull"`. Funkcia `predict` je doplnkom ku funkcii `survreg` a jej výstupom sú odhady kvantilov a ich štandardných chýb. Jeden z argumentov funkcie `predict` je `type`. Nastavte `type="uquantile"`, kedy ide o log-transformáciu. Default počíta odhady rozptylu a IS pomocou delta metódy. Ak chceme odhadnúť lineárny prediktor (`lp`), potom použijeme funkciu `predict(fit, type="lp", newdata = list(skupina = 1))`.

Pri výpočte $\widehat{S}(t)$ pre **log-logistické rozdelenie** treba mať na zreteli, že parametre z log-logistického modelu ($\widehat{\mu}$ a $\widehat{\sigma}$, funkcia `survreg`) sa aplikujú na $Y = \log T$. Vo funkcii `survreg` treba špecifikovať rozdelenie `dist="loglogistic"`. Pre **lognormálne rozdelenie** `dist="lognormal"`. Označenia distribučných funkcii v `R` sú nasledovné

	Weibullove r.	logistické r.($Y = \log T$)	normálne r.($Y = \log T$)
$F(t)$	`pweibull`$(q, \theta, \lambda^{-1})$	`plogis`(q, μ, σ)	`pnorm`(q, μ, σ)
t_p	`qweibull`$(p, \theta, \lambda^{-1})$	`qlogis`(p, μ, σ)	`qnorm`(p, μ, σ)

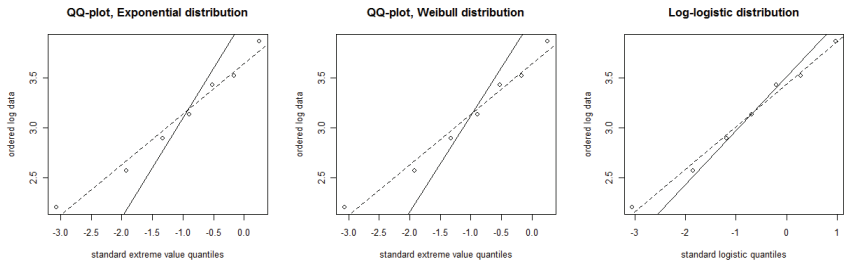

Figure: Porovnanie qq-diagramov

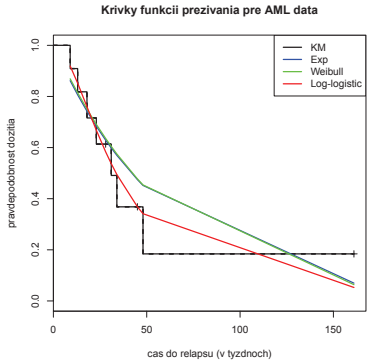

Figure: Porovnanie funkcií prežívania

- zamerajme sa na **porovnanie dvoch škálových parametrov** λ – v **log-transformácii** pôjde o porovnanie dvoch parametrov polohy $\mu = -\log \lambda$
- použijeme `AML dáta`, kde **naparametrický log-rank test** zamietol H_0 o rovnosti dvoh kriviek prežívania (p-hodnota=0.03265)
- najprv si položíme otázku, **či nejaký model** (log-rozdelenie, ktoré patrí do rodiny rozdelení polohy a škály) **fituje dáta adekvátne**
- **saturovaný (plný) log-lineárny model** môžeme písať ako

$$\begin{aligned} Y = \log T &= \widetilde{\mu} + \epsilon \\ &= \beta_0^* + \beta^* \text{skupina} + \epsilon \\ &= \begin{cases} \beta_0^* + \beta^* \text{skupina} + \epsilon, & \text{kde skupina=1} \\ \beta_0^* + \epsilon, & \text{kde skupina=0} \end{cases} \end{aligned}$$

- parameter $\widetilde{\mu} = \beta_0^* + \beta^* \text{skupina}$ – **lineárny prediktor**, ktorý nadobúda dve hodnoty
  - $\mu_1 = \beta_0^* + \beta^*$
  - $\mu_2 = \beta_0^*$
- vieme, že $\widetilde{\mu} = -\log \widetilde{\lambda}$, kde $\widetilde{\lambda}$ znamená **parameter škály** rozdelenia premennej T
- potom $\widetilde{\lambda} = \exp(-\beta_0^* - \beta^* \text{skupina})$ nadobúda dve hodnoty
  - $\lambda_1 = \exp(-\beta_0^* - \beta^*)$
  - $\lambda_2 = \exp(-\beta_0^*)$
- **nulová hypotéza**

  $H_0 : \lambda_1 = \lambda_2$ vtedy a len vtedy, keď $\mu_1 = \mu_2$ vtedy a len vtedy, keď $\beta^* = 0$
- treba si uvedomiť, že **parameter škály** σ v log-transformovanom modeli je rovný (**parameter polohy**)$^{-1}$, kde ide o parameter polohy θ originálneho (netransformovaného) modelu, teda $\sigma = 1/\theta$

Regresný model pre dvojvýberový prípad
Úvod

H_0 testujeme za nasledovných dvoch predpokladov

P1 Predpokladáme **rovnaký parameter tvaru** θ, teda predpokladáme, že $\sigma_1 = \sigma_2$ (**parametre škály**) sa rovnajú. Potom sa chyba $\epsilon = \sigma Z$, kde náhodná premenná Z pochádza z rozdelenia extrémnych hodnôt, štandardizovaného logistického rozdelenia alebo štandardizovaného normálneho rozdelenia.

P2 Predpokladáme **rôzne parametre tvaru** θ, teda $\sigma_1 \neq \sigma_2$

Na testovanie H_0 použijeme tri nasledovné modely

M1 Dáta pochádzajú z **rovnakého rozdelenia**, kde **nulový model** bude mať tvar $Y = \log T = \beta_0^* + \sigma Z$, kde Z pochádza z rozdelenia extrémnych hodnôt.

M2=P1 **Model s rôznymi parametrami polohy** μ **a rovnakými parametrami škály** σ bude mať tvar $Y = \log T = \beta_0^* + \beta^* \text{skupina} + \sigma Z$

M3=P2 **Model s rôznymi parametrami polohy** μ **a škály** σ bude mať tvar $Y = \log T = \beta_0^* + \beta^* \text{skupina} + \epsilon$

Regresný model pre dvojvýberový prípad
Úvod

model	parametre	slovné vyjadrenie
M1	β_0^*, σ	rovnaká poloha a škála
M2	$\beta_0^*, \beta^*, \sigma \equiv \mu_1, \mu_2, \sigma$	rôzna poloha a rovnaká škála
M3	$\beta_0^*, \beta^*, \sigma_1, \sigma_2 \equiv \mu_1, \mu_2, \sigma_1, \sigma_2$	rôzna poloha a škála

Regresný model pre dvojvýberový prípad
Exponenciálny a Weibullov regresný model pre dve skupiny

Nech $\widetilde{\mu} = -\log \widetilde{\lambda}$ a $\sigma = 1/\theta$. Potom

$$\lambda(t|\text{skupina}) = \theta \widetilde{\lambda}^{\theta} t^{\theta-1} = \theta \lambda^{\theta} t^{\theta-1} \exp(\beta \text{skupina}) = \lambda_0(t) \exp(\beta \text{skupina}),$$

kde $\lambda = \exp(-\beta_0^*)$ a $\beta = -\beta^*/\sigma$, $\lambda_0(t)$ je **baseline riziko** (teda ak skupina=0 alebo $\beta = 0$). Teda $\lambda_0(t)$ je **riziková funkcia pre Weibullovo rozdelenie so škálou** λ **nezávislá na ďalších premenných**.
Pomer rizík skupina=1 a skupiny=0

$$\text{HR} = \frac{\lambda(t|1)}{\lambda(t|0)} = \frac{\exp \beta}{\exp 0} = \exp \beta,$$

teda môžeme povedať, že Weibullov model spĺňa **podmienku proporcionality rizík**. Ak $\theta = 1$ ide o exponenciálny regresný model.

Všeobecné zápisy regresných modelov
Exponenciálny regresný model

Nech **funkcia rizika** $\lambda(t, \lambda) = \lambda(t) = \lambda$ je konštantná pri rôznych hodnotách t a $E[T] = 1/\lambda$. Modelujme riziko λ ako funkciu vektora premenných $\mathbf{x}$, potom **funkcia rizika** bude mať tvar

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}^T \beta) = \lambda \exp(\sum_{i=1}^{k} \beta_i x_i),$$

odkiaľ je zreteľné, že premenné ovplyvňujú riziko **multiplikatívne**. V log-lineárnom modeli

$$\log(\lambda(t|\mathbf{x})) = \log(\lambda) + \mathbf{x}^T \beta = \log(\lambda) + \sum_{i=1}^{k} \beta_i x_i$$

premenné ovplyvňujú logaritmus rizika **aditívne** a $\sum_{i=1}^{k} \beta_i x_i$ sa nazýva *lineárny prediktor log-rizika*.

Funkcia prežívania

$$S(t|\mathbf{x}) = \exp(-\lambda(t|\mathbf{x})t) = \exp(-\lambda t \exp(\mathbf{x}^T\beta))$$

Hustota

$$f(t|\mathbf{x}) = \lambda(t|\mathbf{x})S(t|\mathbf{x}) = \lambda \exp(\mathbf{x}^T\beta)\exp(-\lambda t \exp(\mathbf{x}^T\beta))$$

Ak T je **rozdelené exponenciálne**, potom $Y = \log T$ má **rozdelenie extrémnych hodnôt s parametrom škály** $\sigma = 1$. Potom

$$\widetilde{\mu} = -\log(\lambda(t|\mathbf{x})) = -\lambda \exp(\mathbf{x}^T\beta) = -\log\lambda - \mathbf{x}^T\beta, \sigma = 1$$

a platí

$$Y = \log T = \widetilde{\mu} + \sigma Z = \beta_0^* + \mathbf{x}^T\beta^* + Z,$$

kde $\beta_0^* = -\log\lambda$, $\beta^* = -\beta$ a $Z \sim f(z) = \exp(z - e^z), z \in \mathbb{R}$ je **rozdelenie extrémnych hodnôt**.
Zhrnutie: $\lambda(t|\mathbf{x}) = \lambda\exp(\mathbf{x}^T\beta)$ je log-lineárny model zlyhania a je transformovaný na lineárny model $Y = \log T$.

Zovšeobecnime teraz funkciu rizika $\lambda(t) = \theta\lambda^\theta t^{\theta-1}$ tak, že budeme modelovať $\lambda(t)$ ako funkciu vektora premenných $\mathbf{x}$. Potom **funkcia rizika** bude mať tvar

$$\begin{aligned}\lambda(t|\mathbf{x}) &= \lambda_0(t)\exp(\mathbf{x}^T\beta)\\ &= \theta\lambda^\theta t^{\theta-1}\exp(\mathbf{x}^T\beta)\\ &= \theta\left(\lambda(\exp(\mathbf{x}^T\beta))^{\frac{1}{\theta}}\right)^\theta t^{\theta-1} = \theta\widetilde{\lambda}^\theta t^{\theta-1},\end{aligned}$$

kde $\widetilde{\lambda} = \lambda(\exp(\mathbf{x}^T\beta))^{\frac{1}{\theta}}$. Potom **log-lineárny model** bude mať tvar

$$\begin{aligned}\log(\lambda(t|\mathbf{x})) &= \log(\theta) + \theta\log(\widetilde{\lambda}) + (\theta-1)\log(t)\\ &= \log(\theta) + \theta\log(\lambda) + \sum_{i=1}^k \beta_i x_i + (\theta-1)\log(t)\end{aligned}$$

Ak $T \sim W(\lambda, \theta)$, potom $Y = \log T = \widetilde{\mu} + \sigma Z$, podmienené dátami $\mathbf{x}$,

$$\widetilde{\mu} = -\log(\widetilde{\lambda}) = \log(\lambda(\exp(\mathbf{x}^T\beta))^{\frac{1}{\theta}}) = -\log(\lambda) - \frac{1}{\theta}\sum_{i=1}^k \beta_i x_i, \sigma = \frac{1}{\theta}$$

a Z má **štandardizované rozdelenie extrémnych hodnôt**. Preto

$$Y = \widetilde{\mu} + \sigma Z = \beta_0^* + \mathbf{x}^T\beta^* + \sigma Z$$

kde $\beta_0^* = -\log\lambda$, $\beta^* = -\sigma\beta$.

Funkcia prežívania

$$S(t|\mathbf{x}) = \exp\left(-(\widetilde{\lambda}t)^\theta\right).$$

Vieme, že $\Lambda(t|\mathbf{x}) = -\log(S(t|\mathbf{x}))$. Potom **logaritmus funkcie kumulatívneho rizika**

$$\begin{aligned}\log(\Lambda(t|\mathbf{x})) &= \theta\log(\widetilde{\lambda}) + \theta\log(t)\\ &= \theta\log(\lambda) + \theta\log(t) + \sum_{i=1}^k \beta_i x_i\\ &= \log(\Lambda_0(t)) + \sum_{i=1}^k \beta_i x_i,\end{aligned}$$

kde $\Lambda_0(t) = -\log(S_0(t)) = (\lambda t)^\theta$ je **baseline funkcie kumulatívneho rizika**.

Všeobecné zápisy regresných modelov
Weibullov regresný model

Logaritmus funkcie kumulatívneho rizika je lineárny v $\log t$ a v regresných koeficientoch β. Potom vo fixovaných x graf $\Lambda(t|x)$ oproti t v log-log škále je priamka so **sklonom** θ a **interceptom** $\mathbf{x}^T\beta + \theta \log \lambda$. Dá sa ľahko ukázať, že

$$\Lambda(t|\mathbf{x}) = \Lambda_0(t)\exp(\mathbf{x}^T\beta) = (\lambda t)^\theta \exp(\mathbf{x}^T\beta).$$

Weibullov regresný model je jediný log-lineárny model, ktorý spĺňa podmienku **proporcionality rizika**. Vieme, že $\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}^T\beta)$, kde $\lambda_0(t) = \theta\lambda^\theta t^{\theta-1}$ ako aj to, že $\Lambda_0(t) = (\lambda t)^\theta$. Ďalej vieme, že $\log(\Lambda(t|\mathbf{x})) = \theta\log(\lambda) + \theta\log(t) + \mathbf{x}^T\beta$. **Vzťah medzi koeficientami log-lineárnom modeli a koeficientami v rizikovej funkcii** je nasledovný

$$\beta = -\sigma^{-1}\beta^* \text{ a } \lambda = \exp(-\beta_0^*).$$

Pomer rizík sa dá napísať nasledovne

$$\mathrm{HR}(t|\mathbf{x}_1, \mathbf{x}_2) = \frac{\lambda(t|\mathbf{x}_2)}{\lambda(t|\mathbf{x}_1)} = \left(\exp((\mathbf{x}_2^T - \mathbf{x}_1^T)\beta^*)\right)^{1/\sigma}.$$

Všeobecné zápisy regresných modelov
Log-logistický regresný model

Majme model

$$T = \exp Y = \exp(\beta_0^* + \mathbf{x}^T\beta^* + \sigma Z) = \exp(\mathbf{x}^T\beta^*)T^*,$$

kde $T^* = \exp Z^*$, $Z^* = \beta_0^* + \sigma Z$. Teda $\mathbf{x}$ má **multiplikatívny efekt** na čas T. **Riziková funkcia** času T (pre dané $\mathbf{x}$) sa dá zapísať ako funkcia $\lambda_0^*(\cdot)$ v podobe

$$\lambda(t|\mathbf{x}) = \lambda_0^*\left(\exp(-\mathbf{x}^T\beta^*)t\right)\exp(-\mathbf{x}^T\beta^*),$$

čo vyplýva z transformačnej vety.
Pozn.: $f(t) = f^*(g^{-1}(t))\left|\frac{\partial g^{-1}(t)}{\partial t}\right|$, kde $t = g(t^*)$.
Nech $Y = \log T$, potom má **log-lineárny model** tvar

$$Y = \beta_0^* + \mathbf{x}^T\beta^* + \sigma Z,$$

kde $Z \sim$ **štandardizované logistické rozdelenie**.

Všeobecné zápisy regresných modelov
Log-logistický regresný model

Nech $Z^* = \sigma Z$. Potom môžeme písať

$$Y = \beta_0^* + \mathbf{x}^T\beta^* + Z^*,$$

kde $Z^* \sim$ **logistické rozdelenie so strednou hodnotou** β_0^* **a škálou** σ. Potom **baseline riziková funkcia** log-logistického času $T^* = \exp Z^*$ je

$$\lambda_0^*(t^*) = \frac{\lambda\theta(\lambda t^*)^{\theta-1}}{1 + (\lambda t^*)^\theta}$$

kde $\beta_0^* = -\log\lambda$ a $\sigma = 1/\theta$. Treba si uvedomiť, že $\lambda_0^*(t^*)$ je nezávislý na β^*. Potom pre **rizikovú funkciu** T píšeme

$$\lambda(t|\mathbf{x}) = \frac{\theta\widetilde{\lambda}^\theta t^{\theta-1}}{1 + \widetilde{\lambda}^\theta t^\theta},$$

kde $\widetilde{\lambda} = \lambda\exp(-\mathbf{x}^T\beta^*)$. A naviac pre dané $\mathbf{x}$, $T \sim$ **log-logisticky rozdelené s parametrami** $\widetilde{\lambda}$ **a** θ.

Všeobecné zápisy regresných modelov
Log-logistický regresný model

Teda

$$Y = \log T = \widetilde{\mu} + \sigma Z,$$

kde $\widetilde{\mu} = -\log(\widetilde{\lambda}) = -\log\lambda + \mathbf{x}^T\beta^* = \beta_0^* + \mathbf{x}^T\beta^*$ a $Z \sim$ **štandardizované logistické rozdelenie**. Log-logistický model je log-lineárnym modelom, ale nie je modelom proporcionálneho rizika ako napr. Weibullov regresný model.
Log-logistický regresný model má jednu výhodu oproti ostatným modelom, pretože ide o **model propocionálnych šancí** a **proporcionálneho času**. Log-logistická funkcia prežívania má tvar $S(t|\mathbf{x}) = S_0^*(\exp(-\mathbf{x}^T\beta^*)t) = S_0^*(t^*)$, kde $= S_0^*(\cdot)$ je **baseline funkcia prežívania**.
Pozor ! $\mathbf{x}$ mení škálu horizontálnej (t) osi. Ak $\mathbf{x}^T\beta^*$ rastie, potom čas do zlyhania klesá a naopak. Preto sa takýto log-lineárny model nazýva **spomaľovací (zrýchlovací) model času zlyhania (accelerated (decelerated) failure time model**). Do tejto skupiny modelov patrí aj exponenciálny, Weibullov a aj log-normálny regresný model.

Pre **funkciu prežívania** platí

$$S(t|\mathbf{x}) = S_0^*(\exp(-\mathbf{x}^T\beta^*)t) = \frac{1}{1+\left(\exp(y-\beta_0^*-\mathbf{x}^T\beta^*)\right)^\theta},$$

kde $y = \log t$, $\beta_0^* = -\log\lambda$ a $\theta = 1/\sigma$.
Šanca prežitia za časom t je daná ako

$$\frac{S(t|\mathbf{x})}{1-S(t|\mathbf{x})} = \left(\exp(y-\beta_0^*-\mathbf{x}^T\beta^*)\right)^{-\theta},$$

kde treba zdôrazniť, že $-\log$(šanca) je lineárnou funkciou $\log t$ ako aj premenných $x_j, j = 1, 2, \ldots k$. **Pomer šancí za časom** t počítaný medzi x_1 a x_2 sa dá vyjadriť nasledovne

$$\mathrm{OR}(t|\mathbf{x}_2, \mathbf{x}_1) = \left(\exp(\mathbf{x}_2-\mathbf{x}_1)^T\beta^*\right)^\theta$$

naviac **časový pomer** v t počítaný medzi x_1 a x_2

$$\mathrm{TR}(t|\mathbf{x}_2, \mathbf{x}_1) = (\mathrm{OR}(\mathbf{x}_2, \mathbf{x}_1))^{1/\theta}$$

- pomer šancí je často používaný na meranie efektu premenných x_j a je nezávislý na čase t, čo nazývame **vlastnosť proporcionality šancí**.

Example

Ak OR $= 2$, potom šanca prežitia za časom t jedincov s x_2 je $2\times$ tak veľká ako u jedincov s x_1, čo platí pre všetky t.

- rovnako aj TR je nezávislý na čase, čo nazývame **vlastnosť proporcionality časov**.

Example

Podobne ak TR $= 2$, tak čas prežívania jedincov s x_2 je $2\times$ väčší ako u jedincov s x_1.

- OR $= \mathrm{TR}^\theta$, kde pomer OR je kontrolovaný θ, tvarovým parametrom log-logistického rozdelenia

Example

Ak $\theta = 1$, tak OR $=$ TR. Ak $\theta = 2$, tak TR $= 2$ a OR $= 2^2 = 4$.

- na jednotku nárastu jednej premennej, fixovaním ostatných premenných v x, OR $\to +\infty \vee 0$ ak $\theta \to \infty$ (v závislosti od znamienka korešpondujúcej premennej pri β^*)
- na zistenie $\widehat{\mathrm{HR}}$ a $\widehat{\mathrm{OR}}$ – odhady vypočítané funkciou `survreg`
- ľubovoľný $p \times 100\%$ **percentil** log-logistického modelu času t

$$t_p(\mathbf{x}) = \left(\frac{p}{1-p}\right)^\sigma \exp(\beta_0^* + \mathbf{x}^T\beta^*)$$

- log-logistický regresný model je jediným akcelerovaným modelom zlyhania, ktorý spĺňa podmienky **propocionality šancí** a **proporcionality času**

Example

(pokrač.) Použitím funkcie `survreg` pre cenzurované dáta odhadnite parametre Weibullovho, log-logistického a log-normálneho modelu a samotné modely porovnajte.

- Na základe výsledkov LR testu medzi modelmi M1 a M2 ako aj M2 a M3 môžeme konštatovať, že model M2, ktorý predpokladá rovnakú škálu je adek-vátny.
- Výsledky z modelov M1, M2 a M3 zhrnieme v nasledovnej tabuľke

rozdelenie	$L(\widehat{\beta}_0^*, \widehat{\beta}^*)$	p-hodnota	$\widehat{\beta}_0^*$	$\widehat{\beta}^*$	p-hodnota (skupina)
Weibullovo	−80.5	0.021	3.180	0.929	0.0151
log-logistické	−79.4	0.120	2.899	0.604	0.1240
log-normálne	−78.9	0.062	2.854	0.724	0.0568

- Vo Weibullovom modeli je efekt skupiny signifikantný (prvá skupina zostáva v remisii dlhšie, t.j. prežívanie je v druhej skupine kratšie) s odhadmi $\widehat{\mu}_1 = 4.109$ a $\widehat{\mu}_2 = 4.109 - 0.929 = 3.18$ rozdelenia extrémnych hodnôt.
- Log-normálny model má maximálnu vierohodnosť najväčšiu a Weibullov model najmenšiu. Avšak LR test fitu modelu je signifikantný len pre Weibullov model, kde p-hodnota=0.0151.

- Mediány s ich 95%IS zhŕňa pre všetky modely nasledovná tabuľka

model	med 1	95%IS pre med 1	med 2	95%IS pre med 2
Weibullov	45.566	(24.901,83.381)	17.990	(10.966,29.514)
Log-logistický	33.215	(18.902,58.366)	18.147	(10.747,30.641)
Log-normálny	35.833	(20.510,62.605)	17.364	(10.556,28.561)

- Odhad lineárneho prediktora $\widehat{\widetilde{\mu}} = \widehat{\beta}_0^* + \widehat{\beta}^*$skupina, $\widehat{\widetilde{\mu}} = \log(\widehat{\widetilde{\lambda}})$, $\widehat{\widetilde{\lambda}} = \exp(-\widehat{\widetilde{\mu}})$ = $\exp(-\widehat{\beta}_0^* - \widehat{\beta}^*$skupina), $\widehat{\theta} = 1/\widehat{\sigma}$ pre všetky tri modely (Weibullov, log-logistický a log-normálny; zľava doprava) zhŕňa nasledovná tabuľka

skupina	$\widehat{\widetilde{\lambda}}$	$\widehat{\sigma}$	$\widehat{\widetilde{\lambda}}$	$\widehat{\sigma}$	$\widehat{\widetilde{\lambda}}$	$\widehat{\sigma}$
0	0.042	1.264	0.055	1.950	0.058	1.160
1	0.016	1.264	0.030	1.950	0.028	1.160

- Z qq-diagramu je zrejmé, že log-logistický a log-normálny model fitujú skupinu `Maintained` lepšie ako Weibullov regresný model. Avšak fit skupiny `Nonmain- tained` sa použitím týchto modelov nezlepšil.

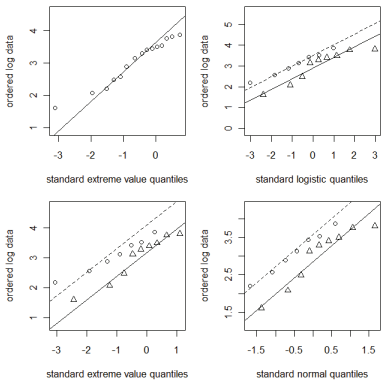

Figure: Porovnanie kvantilových diagramov (Weibullove modely M1 a M2, log-logistický model, log-normálny model)

- Neparametrický prístup (KM odhad) dáva lepší náhľad na AML dáta

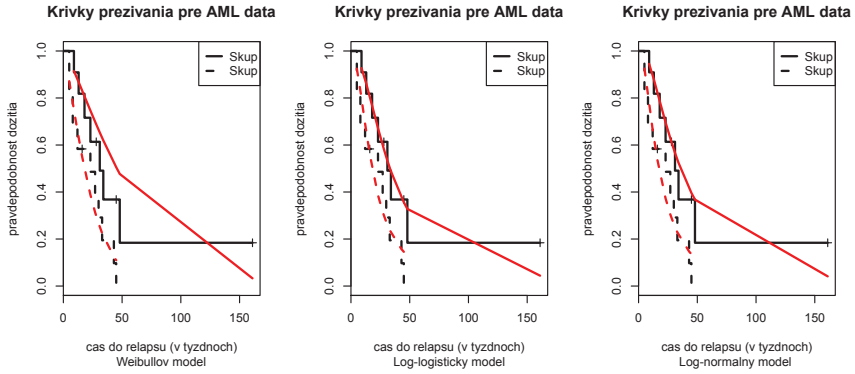

Figure: Porovnanie funkcií prežívania

- Majme Weibullov model. Nech $\widehat{\beta} = -\widehat{\beta}^*/\widehat{\sigma} = -0.929/0.791 = -1.1745$, potom odhad pomeru rizík $\widehat{\text{HR}} = \widehat{\lambda}(t|1)/\widehat{\lambda}(t|0) = \exp(\widehat{\beta})/\exp(0) = \exp(\widehat{\beta}) = 0.31$. Skupina `Maitained` má 31% riziko z rizika kontrolnej skupiny `Nonmaitained`. Alebo kontrolná skupina má $1/0.31 = 3.23\times$ väčšie riziko ako skupina `Maitained`. HR je teda miera efektu vplyvu skupiny na čas prežívania (čas do relapsu alebo recidívy).
- Majme opäť Weibullov model. Ak vypočítame pomer odhadnutých pravdepodobností prežívania, napr. v čase $t = 31$ týždňov ($\widehat{\widetilde{\lambda}} = \exp(-\widehat{\widetilde{\mu}})$), dostaneme $\widehat{S}(31|1)/\widehat{S}(31|0) = 0.6531634/0.2517891 = 2.594$, čo znamená, že skupina `Maitained` má $2.594\times$ väčšiu pravdepodobnosť zostať v remisii (dočasné zlepšenie stavu) najmenej 31 týždňov ako skupina `Nonmaitained`.

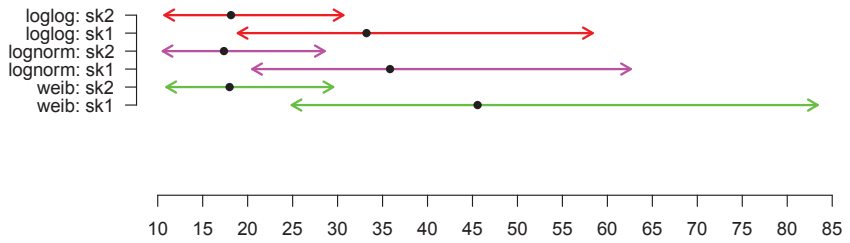

Figure: Porovnanie intervalov spoľahlivosti pre medián (Weibullov, log-logistický a log-normálny model)