Regresní modelování v analýze přežití

Tomáš Pavlík

pavlik@iba.muni.cz

Bi7491 Regresní modelování

9. května 2012

Motivace

- V klasické regresi nás zajímá vliv vysvětlujících proměnných (sledovaných faktorů) na hodnoty sledované náhodné veličiny.
- V logistické regresi nás zajímá vliv vysvětlujících proměnných na výskyt nebo nevýskyt nějakého jevu.
- Tyto metody jsou však nevhodné ve chvíli, kdy
  1. studujeme čas do výskytu nějakého jevu
  2. může se stát, že u všech subjektů nemáme o výskytu tohoto jevu kompletní informace.

- **Analýza přežití (*survival analysis*) – soubor metod pro popis a modelování doby do výskytu sledovaného jevu v čase.**

Motivace - příklad

Jméno pacienta	Zemřel	Délka sledování
Pan Silný	Ne	12,0 měsíců
Pan Slabý	Ano	6,8 měsíců
Pan Moucha	Ano	4,8 měsíců
Pan Komár	Ano	9,8 měsíců
Pan Skála	Ne	10,8 měsíců

Příklady

- Čas od diagnózy do úmrtí onkologického pacienta
- Čas od zahájení léčby do progrese onemocnění
- Čas od propuštění z nemocnice do opakované hospitalizace
- Čas od infekce HIV do propuknutí AIDS
- Čas od narození do prvního požití alkoholu/drog

Modelová data - projekt Alert

Projekt **Alert** - Klinický registr pro standardizovaný sběr diagnostických, prognostických a léčebných dat u pacientů s akutní leukémií.

- charakteristika pacientů s AML, ALL a APL v ČR a SR
- sběr a vyhodnocování základních epidemiologických a klinických dat
- sběr a vyhodnocování molekulárně-biologických a genetických dat
- modelování přežití

Definice souboru pacientů:

- Dospělí pacienti s de novo AML i sekundární AML, diagnostikovaní v ČR (5 hematologických center) v období 1999–2009
- Kurativně léčení (léčba se snahou o odstranění nemoci)
- N = 1091

Literatura

- Marubini & Valsecchi: *Analysing Survival Data from Clinical Trials and Observational Studies*
- Collett: *Modelling Survival Data in Medical Research*
- Hosmer & Lemeshow: *Applied Survival Analysis*
- Klein & Moeschberger: *Survival Analysis: Techniques for Censored and Truncated Data*
- Therneau & Grambsch: *Modeling Survival Data: Extending the Cox Model*
- Andersen, Borgan, Gill & Keiding: *Statistical Models Based on Counting Processes*

Úvod a definice pojmů

Definice pojmů

Čas do výskytu sledované události (*time to event*), jinak také čas přežití (*survival time*, *failure time*, *event time*), je nezáporná náhodná veličina. Budeme ji značit T. Lze ji jednoznačně popsat následujícími charakteristikami:

- Distribuční funkcí: $F(t) = \mathrm{P}(T \leq t), t \geq 0$
- Hustotou pravděpodobnosti: $f(t) = F'(t), t \geq 0$
- Funkcí přežití: $S(t) = \mathrm{P}(T > t) = 1 - F(t), t \geq 0$

Náhodná veličina T může být jak **spojitá**, definovaná na intervalu $(0, \infty)$, tak **diskrétní**, nabývající nejvýše spočetně mnoha hodnot, např. $a_1, a_2, \ldots, a_n$.
My se zde budeme zabývat pouze spojitým případem.

Klíčové prvky v analýze přežití

Pro definici času přežití jako náhodné veličiny potřebujeme stanovit následující:

- **Jednoznačný počátek sledování**
  (např. datum diagnózy, narození, zahájení léčby)
- **Časovou škálu**
  (např. reálný čas v měsících nebo letech)
- **Koncovou událost**
  (např. úmrtí, progrese onemocnění, rehospitalizace, otěhotnění)

Sledovaná událost by měla být snadno pozorovatelná či měřitelná (úmrtí $\times$ progrese onemocnění).

Úlohy v analýze přežití

1. Popis – bodové a intervalové odhady pravděpodobnosti přežití v čase t
2. Srovnání – hodnocení rozdílů v přežívání dvou a více skupin (pacientů, osob, zvířat, věcí, bakterií, apod.)
3. **Modelování – hodnocení vlivu vysvětlujících proměnných na pozorovaný čas přežití**

Specifikum dat přežití - Cenzorování

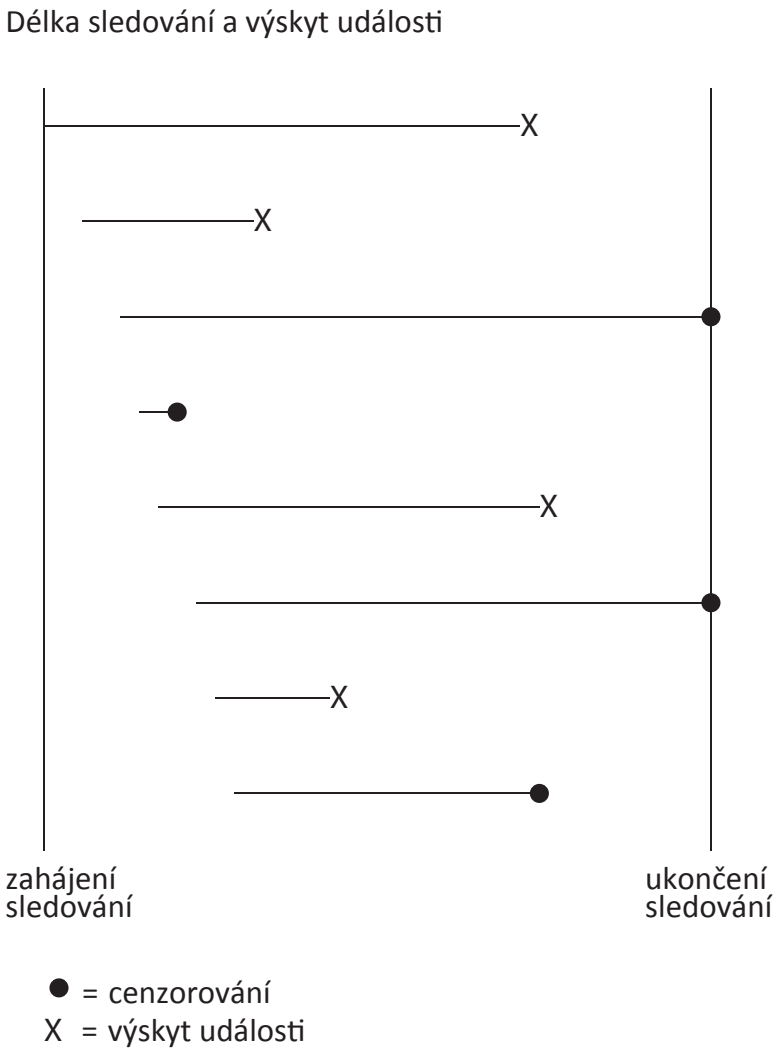
- Definovaná událost se nemusí v průběhu sledování vyskytnout u všech subjektů (pacientů) $\rightarrow$ Nevíme tedy, kdy a jestli vůbec daná událost nastala. Víme pouze, že nenastala před ukončením sledování.
- **Cenzorování je ztrátou určité informace, cenzorované časy přežití však nelze z analýzy vyřadit.**
- Typy cenzorování:
  1. Cenzorování zprava
  2. Cenzorování zleva
  3. Intervalové cenzorování
- Bez ohledu na typ cenzorování, hlavním předpokladem analýzy přežití je, že **cenzorování je** tzv. **neinformativní** (*non-informative*). To znamená, že výskyt cenzorování nijak nesouvisí s výskytem sledovaných událostí.

Typy cenzorování

1. **Cenzorování zprava** – pozorujeme minimum z hodnot skutečného a cenzorovaného času přežití.
   - Klinické studie
   - Observační studie
2. **Cenzorování zleva** – pozorujeme maximum z hodnot skutečného a cenzorovaného času přežití.
   - Věk prvního požití drog, kouření
3. **Intervalové cenzorování** – víme pouze, že skutečný čas přežití leží v intervalu (D, H).
   - Sledování pacientů v dlouhodobějších intervalech (screening karcinomu prostaty, pacienti s CML)

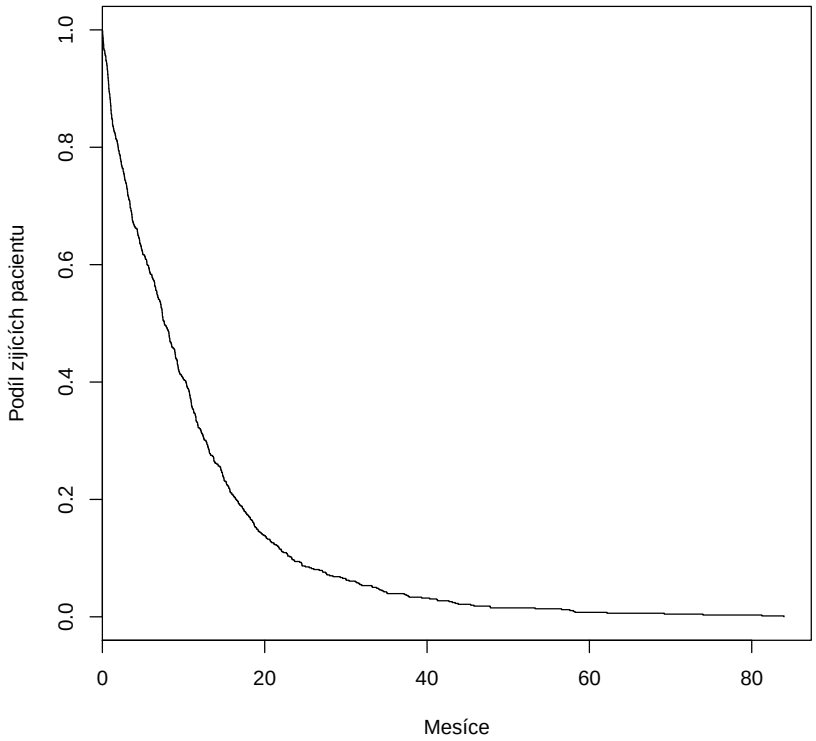
Nejčastějším typem cenzorování je v biolgii a medicíně cenzorování zprava.

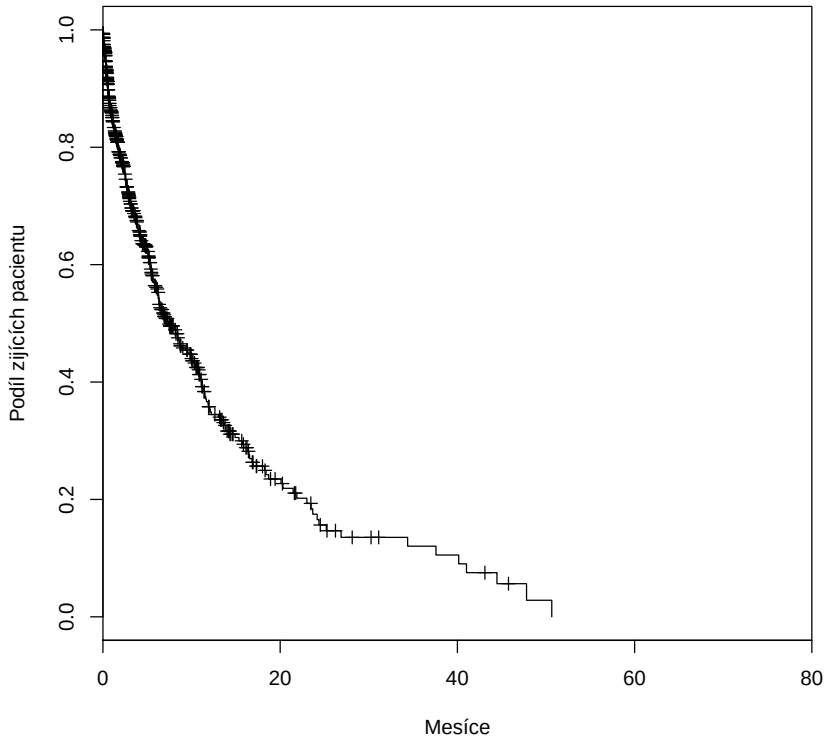
Cenzorování zprava - příklad



Obrázek 1. Sledování času přežití v přítomnosti cenzorování

Cenzorování - vliv procenta cenzorovaných hodnot na odhad funkce přežití





Obrázek 2. Křivka přežití bez cenzorovaných hodnot a s 50% cenzorovaných hodnot.

Délka sledování

- Design (plán) studie má vždy vliv na její analýzu.
- Informace o výskytu sledované události v čase je primárně zachycena v kompletních (skutečných) časech přežití.
- **Délka sledování** (studie) **proto musí být dostatečná**, abychom pozorovali dostatek událostí a tedy i informace o sledovaném procesu.
- Souvisí s **četností výskytu události** (např. mortalitou u rakoviny).

Čas přežití v přítomnosti cenzorování

Jak jsou data o přežití reprezentována v přítomnosti cenzorování?

- T_i je zaznamenaný čas přežití pro pacienta i
- Označme δ_i indikátor pozorované události. Pak

$$\delta_i = \begin{cases} 1 \text{ když pozorujeme skutečný čas do události} \\ 0 \text{ když je čas } T_i \text{ cenzorován} \end{cases}$$

- Data přežití n subjektů pak budeme reprezentovat dvojicemi $(t_i, \delta_i), i = 1, \ldots, n$.

Hustota pravděpodobnosti (*density function*)

Hustota udává pravděpodobnost výskytu sledované události v čase t.

- T jako spojitá náhodná veličina:

$$f(t) = \lim_{u \to 0} \frac{1}{u} \mathrm{P}(t \leq T \leq t + u)$$

- T jako diskrétní náhodná veličina:

$$f(t) = \mathrm{P}(T = t) = \begin{cases} f_j & \text{když } t = a_j, j = 1, \ldots, n \\ 0 & \text{když } t \neq a_j, j = 1, \ldots, n \end{cases}$$

Funkce přežití (*survival function*)

Hlavní charakteristikou je tzv. **funkce přežití**. Ta udává pravděpodobnost, že jedinec přežije (vzhledem k výskytu sledované události) déle než do času t.

- T jako spojitá náhodná veličina:

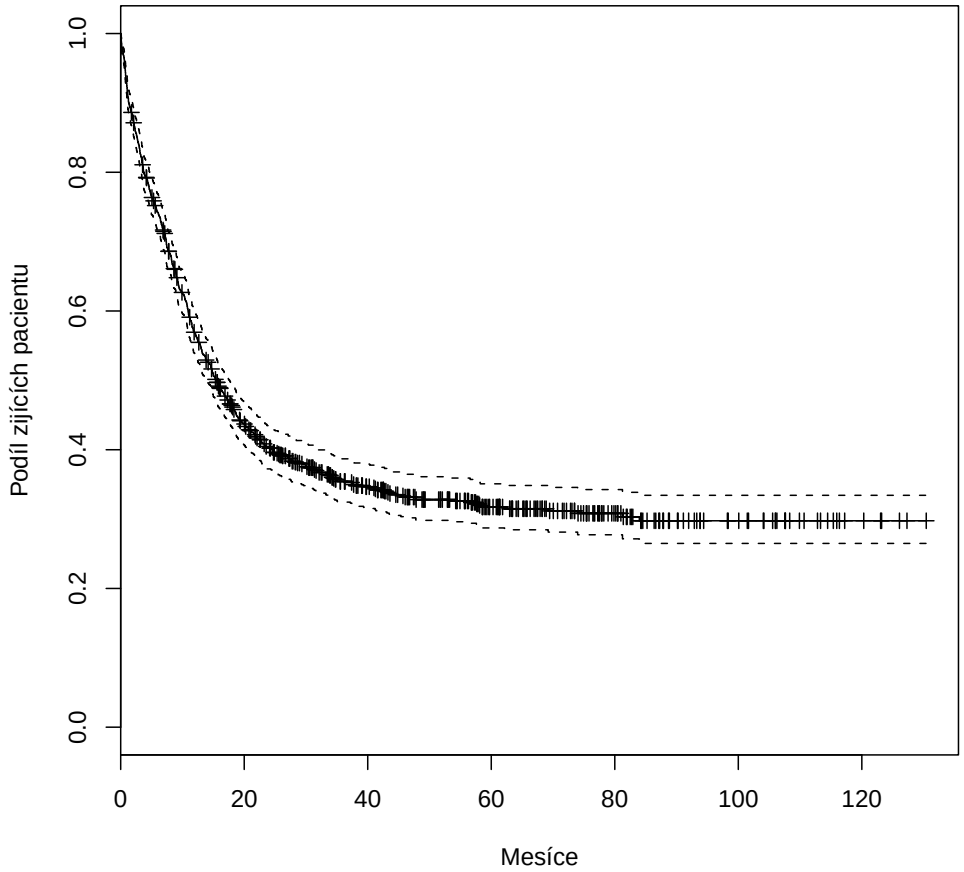
$$S(t) = \mathrm{P}(T > t) = \int_t^{\infty} f(x)dx = 1 - F(t)$$

- T jako diskrétní náhodná veličina:

$$S(t) = \sum_{a_j \geq t} f(a_j) = \sum_{a_j \geq t} f_j$$

- Nabývá hodnot mezi 1 a 0 (respektive 100% a 0%), kdy hodnotu 1 má v počátečním čase a hodnotou 0 při výskytu poslední události
- Je to nerostoucí funkce

Funkce přežití - příklad



Obrázek 3. Odhad funkce přežití pacientů s AML léčených kurativně (N = 1091).

Riziková funkce (*hazard function*)

Druhou důležitou charakteristikou je tzv. **riziková funkce**. Ta vyjadřuje intenzitu výskytu sledované události v čase t za podmínky, že subjekt přežil do času t.

- Riziková funkce je klíčová pro modelování přežití - řada modelů je definována a interpretována pomocí rizikové funkce.

- T jako spojitá náhodná veličina:

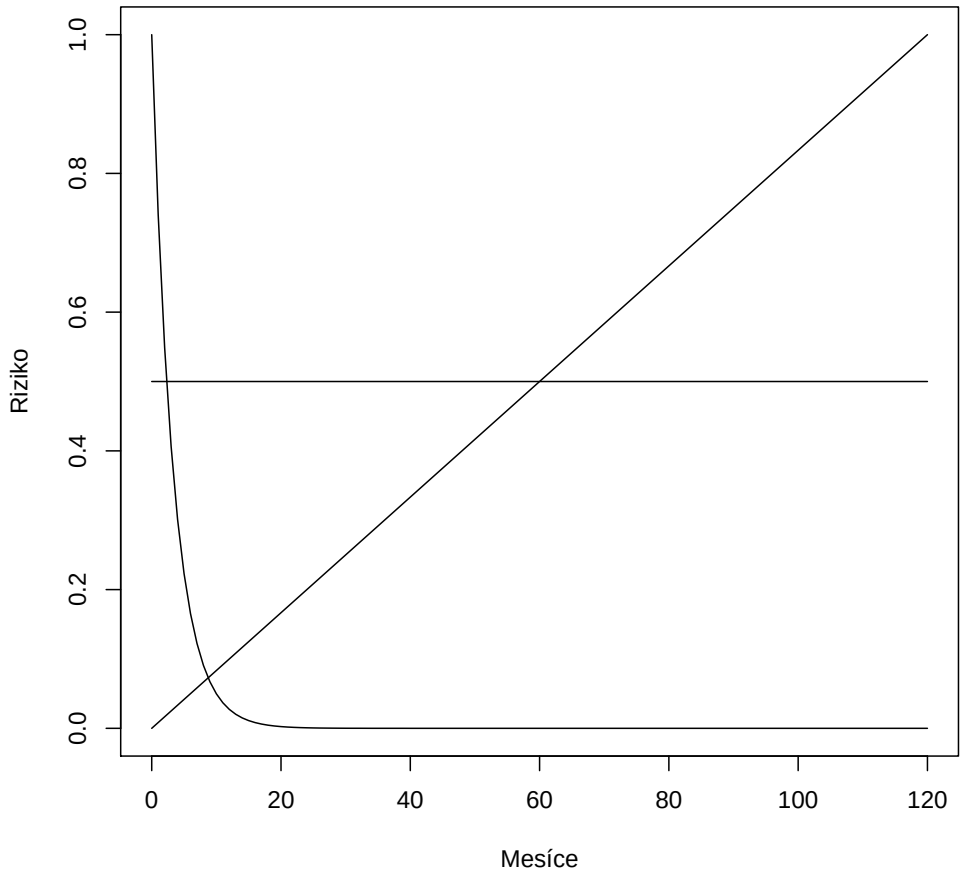
$$h(t) = \lim_{u \to 0} \frac{\mathrm{P}\{t < T \leq t + u | T > t\}}{u} = \lim_{u \to 0} \frac{\mathrm{P}\{t < T \leq t + u\}/\mathrm{P}\{T > t\}}{u}$$
$$= \lim_{u \to 0} \frac{[F(t+u) - F(t)]/u}{S(t)} = \frac{dF(t)/dt}{S(t)}$$
$$= \frac{f(t)}{S(t)} = -\frac{d \ln S(t)}{dt}$$

Riziková funkce (*hazard function*)

- T jako diskrétní náhodná veličina:

$$\begin{aligned} h(a_j) = h_j &= \mathrm{P}(T = a_j \mid T \geq a_j) \\ &= \frac{\mathrm{P}(T = a_j)}{\mathrm{P}(T \geq a_j)} \\ &= \frac{f(a_j)}{S(a_j)} \\ &= \frac{f(a_j)}{\sum_{k:a_k \geq a_j} f(a_k)} \end{aligned}$$

Riziková funkce - příklad



Obrázek 4. Modelové příklady rizikové funkce, respektive vývoje rizika výskytu sledované události v čase.

Kumulativní riziková funkce (*cumulative hazard function*)

Kumulativní riziková funkce odpovídá celkovému riziku výskytu sledované události od začátku sledování až do času t.

- T jako spojitá náhodná veličina:

$$H(t) = \int_0^t h(x)dx$$

- T jako diskrétní náhodná veličina:

$$H(t) = \sum_{j:a_j<t} h_j$$

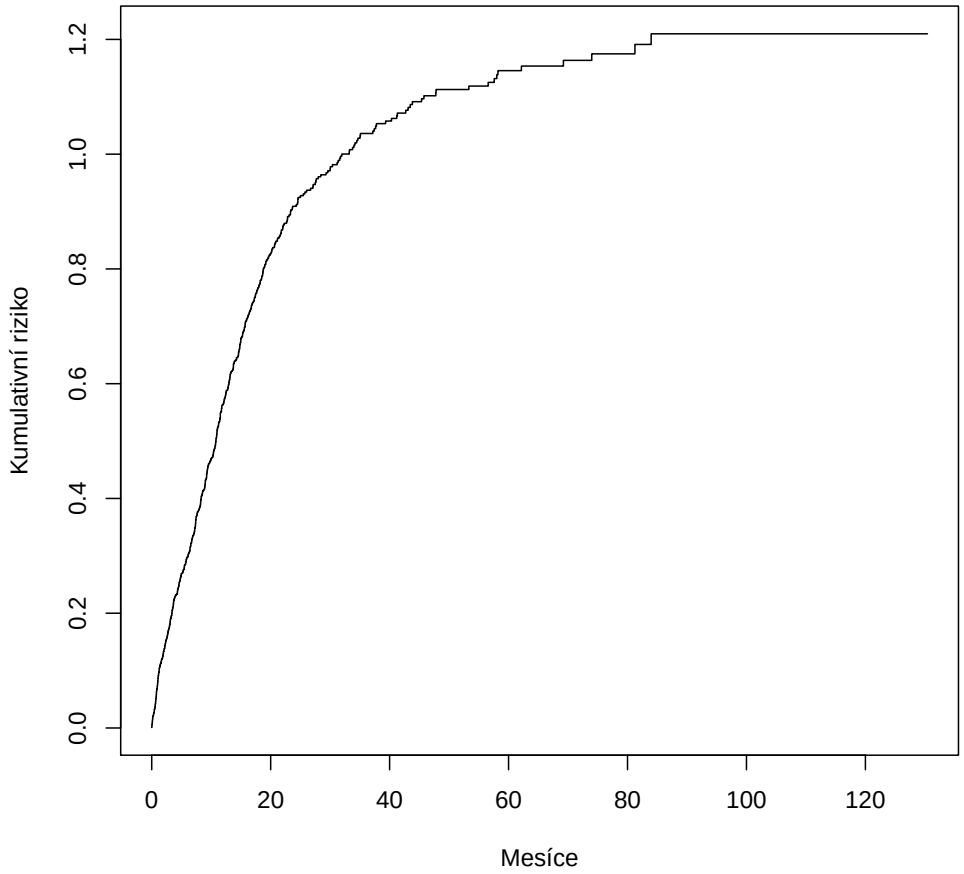
Kumulativní riziková funkce (*cumulative hazard function*)

- Kumulativní rizikovou funkci, $H(t)$, lze použít k odhadu funkce přežití:

$$h(t) = -\frac{d \ln S(t)}{dt} \longrightarrow \int_0^t h(x)dx = -\ln S(t)$$

$$S(t) = \exp(-H(t))$$

Kumulativní riziková funkce - příklad



Obrázek 4. Odhad kumulativní rizikové funkce vzhledem k výskytu úmrtí u pacientů s AML (N = 1091).

Medián přežití a průměrné přežití

Medián přežití, τ, je definován jako čas, ve kterém má funkce přežití hodnotu 0,5:

$$S(\tau) = 0,5$$

Prakticky ho najdeme jako nejmenší čas t, pro který platí, že $\hat{S}(t) \leq 0,5$.

Průměrnou dobu přežití, μ, představuje odhad střední doby přežití náhodné veličiny T. Podle definice střední hodnoty platí

$$\mu = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt.$$

Výpočet je pak jednoduchý:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt,$$

kde $\hat{S}(t)$ je libovolný odhad funkce přežití. Problém představují neparametrické odhady funkce přežití s cenzorovaným posledním časem přežití.

Neparametrické odhady

Neparametrické odhady v analýze přežití

1. Odhady **funkce přežití**
   - Kaplanův-Meierův odhad (Kaplan-Meier estimator) – Kaplan & Meier (1958)
   - Odhad metodou úmrtnostních tabulek (Life-table estimator, actuarial estimator)
   - Breslowův odhad (Breslow estimator) – Breslow (1972)
2. Nelsonův-Aalenův odhad **kumulativní rizikové funkce** (Nelson-Aalen estimator) – Nelson (1972); Aalen (1978)

Kaplanův-Meierův odhad – motivace

- V nepřítomnosti cenzorování je odhad funkce přežití, $\hat{S}(t)$, podíl pacientů s časem přežití větším než t.

$$\hat{S}(t) = \frac{\text{počet subjektů s } T \geq t}{\text{celkový počet subjektů}}$$

- V přítomnosti cenzorování můžeme využít podmíněnou pravděpodobnost

$$\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B)$$

$$\begin{aligned}\mathrm{P}(A_1 \cap A_2 \ldots \cap A_k) =& \mathrm{P}(A_k|A_1 \cap A_2 \ldots \cap A_{k-1}) \times \\ & \times \mathrm{P}(A_{k-1}|A_1 \cap A_2 \ldots \cap A_{k-2}) \\ & \ldots \\ & \times \mathrm{P}(A_2|A_1) \\ & \times \mathrm{P}(A_1)\end{aligned}$$

Kaplanův-Meierův odhad $S(t)$

Předpokládejme k různých časů přežití takových, že $t_1 < t_2 < \ldots < t_k$. Dále předpokládejme $t \geq t_k$ Pak

$$\begin{aligned} S(t) &= \mathrm{P}(T \geq t_k) \\ &= \mathrm{P}(T \geq t_1, T \geq t_2, \ldots, T \geq t_k) \\ &= \mathrm{P}(T \geq t_1) \times \prod_{i=1}^{k-1} \mathrm{P}(T \geq t_{i+1} | T \geq t_i) \\ &= \prod_{i=1}^{k} [1 - \mathrm{P}(T = t_i | T \geq t_i)] \\ &= \prod_{i=1}^{k} [1 - h(t_i))] \end{aligned}$$

Kaplanův-Meierův odhad $S(t)$

1. Jak odhadneme $h(t_i)$?
2. A jak tento odhad rozšíříme i na cenzorované časy přežití?

Kaplanův-Meierův odhad $S(t)$

- Pozorovaná data máme ve formě dvojic $(t_i, \delta_i), i = 1, \ldots, n$. Označme R_i počet subjektů v riziku sledované události v čase t_i.
- **Kaplanův-Meierův odhad** $S(t)$:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{\delta_i}{R_i}\right).$$

- Součin je přes všechny pozorované časy přežití, cenzorované časy ale k odhadu přispívají pouze prostřednictvím R_i, neboť pro cenzorované $\delta_i = 0$.
- Pro konstrukci $100(1-\alpha)\%$ intervalu spolehlivosti potřebujeme odhad rozptylu $\hat{S}(t)$. **Greenwoodův odhad**:

$$\widehat{\text{var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{\delta_i}{R_i(R_i - \delta_i)}.$$

Odhad $S(t)$ metodou úmrtnostních tabulek

- Pozorovaná data máme v agregované formě pro J časových intervalů.
- Označme d_j počet sledovaných událostí (úmrtí) v $j$tém intervalu, $j = 1, \ldots, J$.
- Dále R_j bude počet subjektů (pacientů) v riziku výskytu sledované události na začátku intervalu j.
- Nakonec c_j je počet subjektů s časem cenzorovaným v průběhu $j$tého intervalu.
- Odhad pravděpodobnosti přežití do konce Jtého intervalu **metodou úmrtnostních tabulek** je pak následující:

$$\hat{S}(J) = \prod_{j=1}^{J} p(j) = \prod_{j=1}^{J} \left(1 - \frac{d_j}{R_j - c_j/2}\right).$$

Nelsonův-Aalenův odhad rizikové funkce

- Značení δ_i a R_i je stejné jako u K-M odhadu.
- **Nelsonův-Aalenův odhad** kumulativní rizikové funkce má tvar:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{\delta_i}{R_i}.$$

- Opět platí, že suma je přes všechny pozorované časy přežití, cenzorované časy ale k odhadu přispívají pouze prostřednictvím R_i, neboť pro cenzorované $\delta_i = 0$.
- Aalen následně doplnil bodový N-A odhad odhadem jeho rozptylu:

$$\widehat{\text{var}}(\hat{H}(t)) = \sum_{t_i \leq t} \frac{\delta_i}{R_i^2}.$$

Breslowův odhad $S(t)$

Breslow v roce 1972 navrhl neparametrický odhad funkce přežití pomocí N-A odhadu kumulativní rizikové funkce.

- Platí:

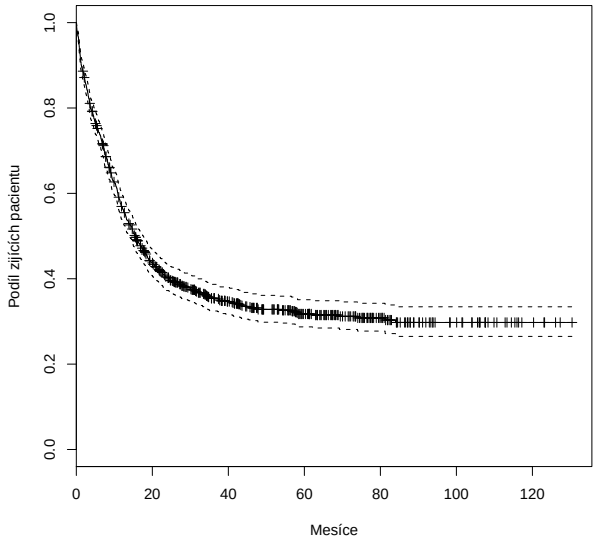
$$h(t) = -\frac{d \ln S(t)}{dt} \longrightarrow \int_0^t h(x)dx = H(t) = -\ln S(t)$$

- Pak můžeme odhad $S(t)$ vyjádřit jako:

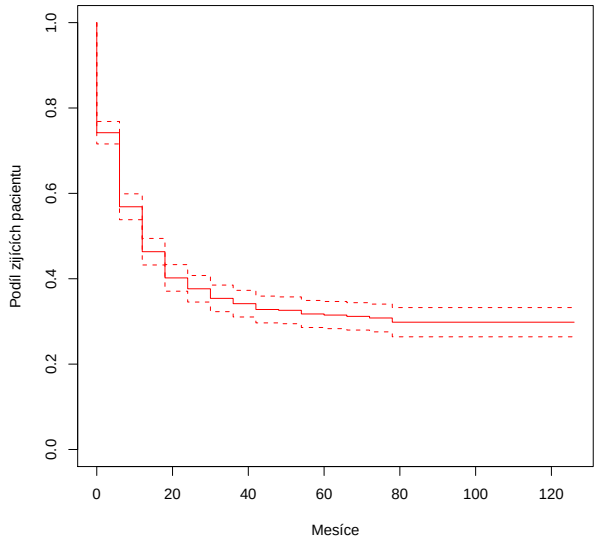
$$\hat{S}(t) = \exp(-\hat{H}(t)) = \exp\left(-\sum_{t_i \leq t} \frac{\delta_i}{R_i}\right)$$

Srovnání naparametrických odhadů $S(t)$

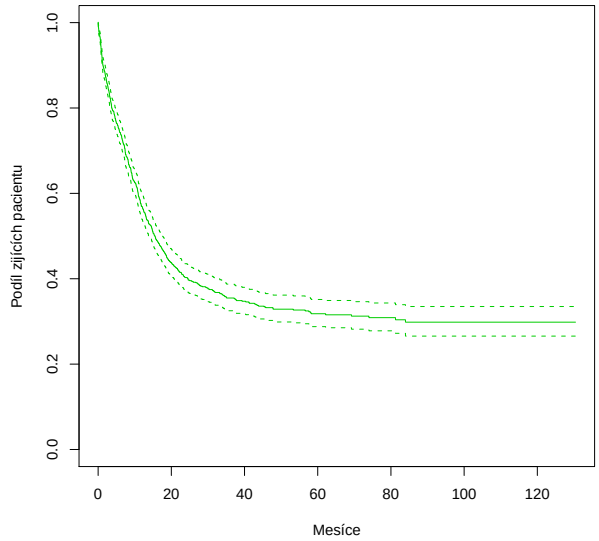
K-M odhad



L-T odhad

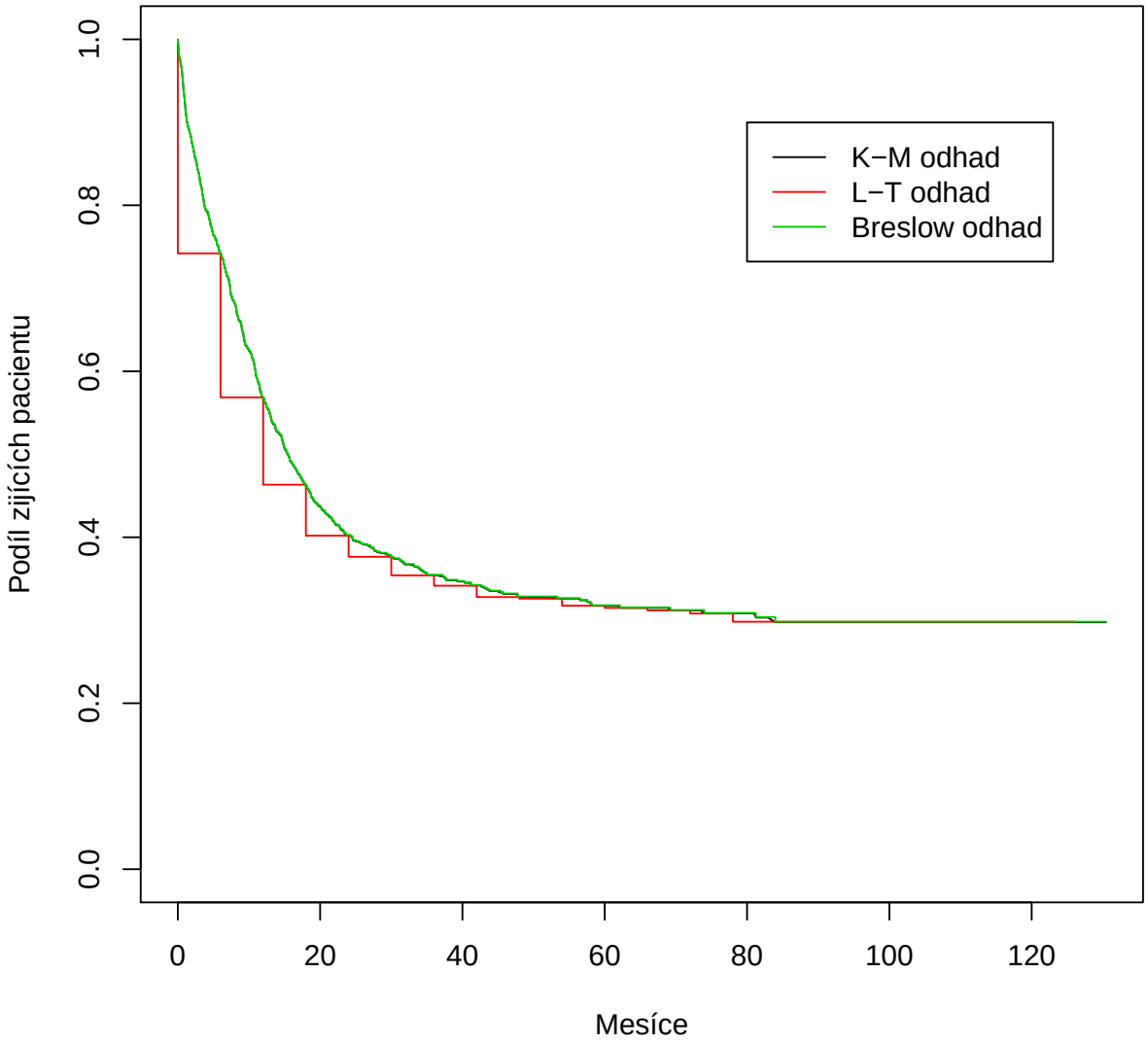


Breslowův odhad



- L-T odhad respektuje časové intervaly.
- Při dostatečné velikosti souboru jsou odhady téměř shodné.

Srovnání naparametrických odhadů $S(t)$



Parametrické odhady

Parametrické odhady funkce přežití

K-M odhad je velmi dobrým nástrojem pro odhad $S(t)$ - je to neparametrický maximálně věrohodný odhad $S(t)$, nicméně funkci přežití můžeme odhadnout i pomocí předpokladu o parametrické formě rozdělení pravděpodobnosti veličiny T.

Výhody parametrického odhadu $S(t)$ jsou následující:

- Můžeme jednoduše odhadnout kvantily $S(t)$ a střední dobu dožití.
- Můžeme najít vyjádření $S(t), H(t)$ a $h(t)$ pomocí spojité funkce.
- Můžeme odhadnout $S(t)$ přesněji než pomocí K-M odhadu.

Předpoklad určitého rozdělení pravděpodobnosti pro T je velmi silný! Pokud není správný, odhady mohou být úplně mimo realitu.

Rozdělení pravděpodobnosti v analýze přežití

Časy přežití mají většinou kladně sešikmené rozdělení $\rightarrow$ normální rozdělení není vhodné.

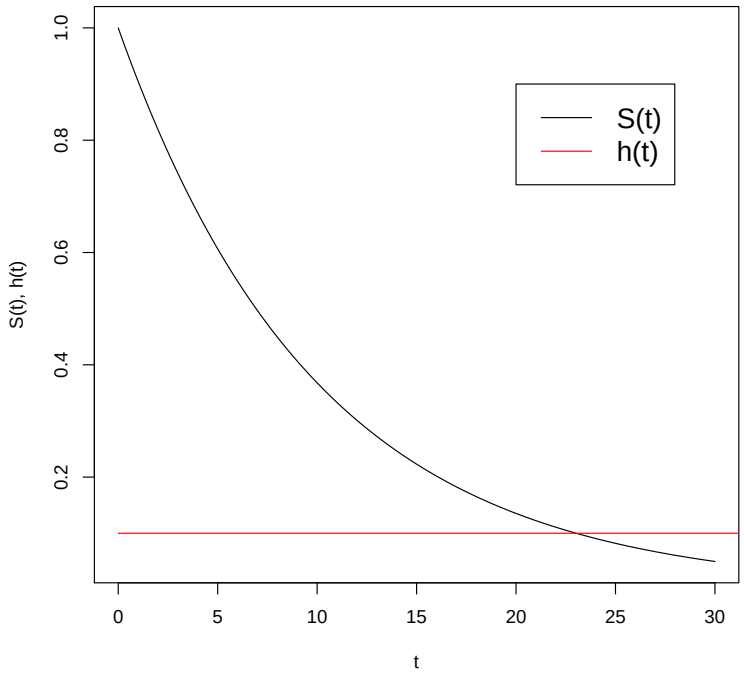
Nejčastěji používaná rozdělení pravděpodobnosti v analýze přežití jsou:

- Exponenciální rozdělení
- Weibullovo rozdělení
- Lognormální rozdělení ($\log(T)$ má normální rozdělení)
- Gamma rozdělení

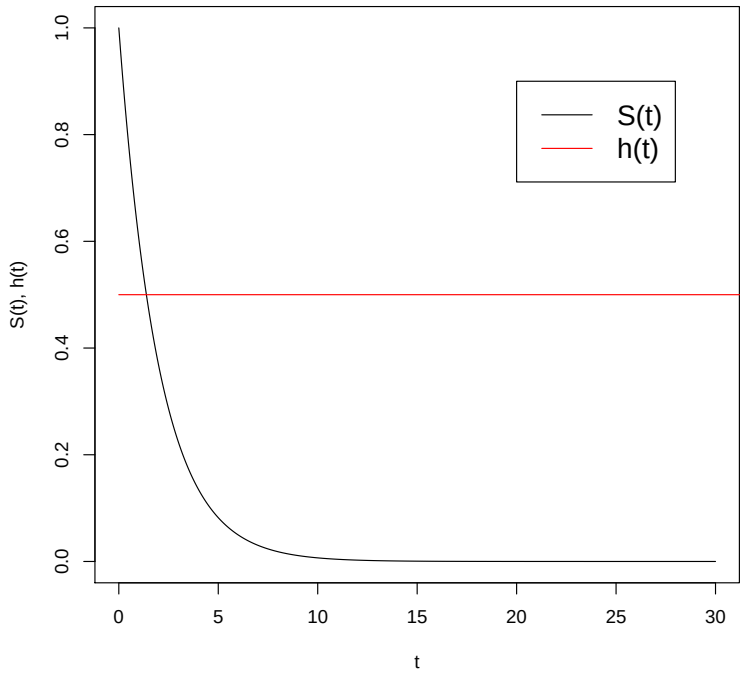
Příklad - Exponenciální rozdělení

$$h(t) = \lambda, S(t) = \exp(-\lambda t)$$

$\lambda = 0.1$



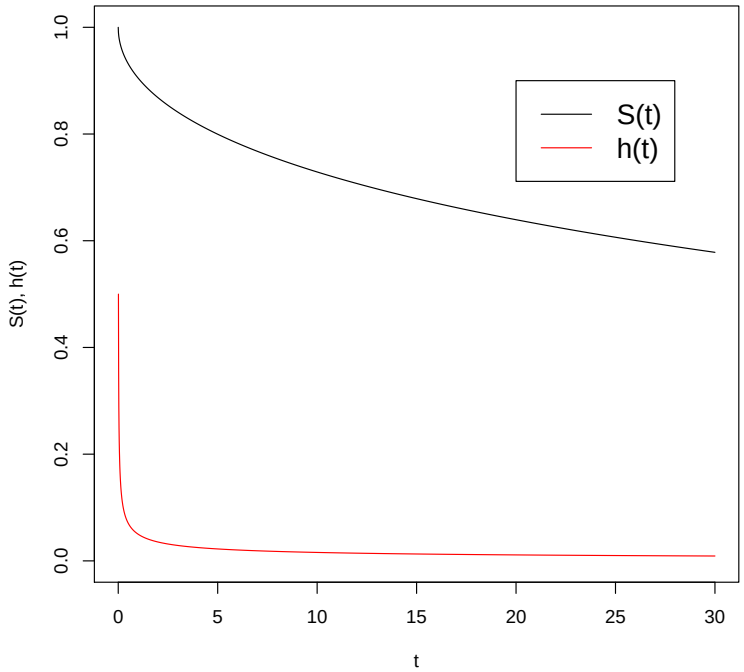
$\lambda = 0.5$



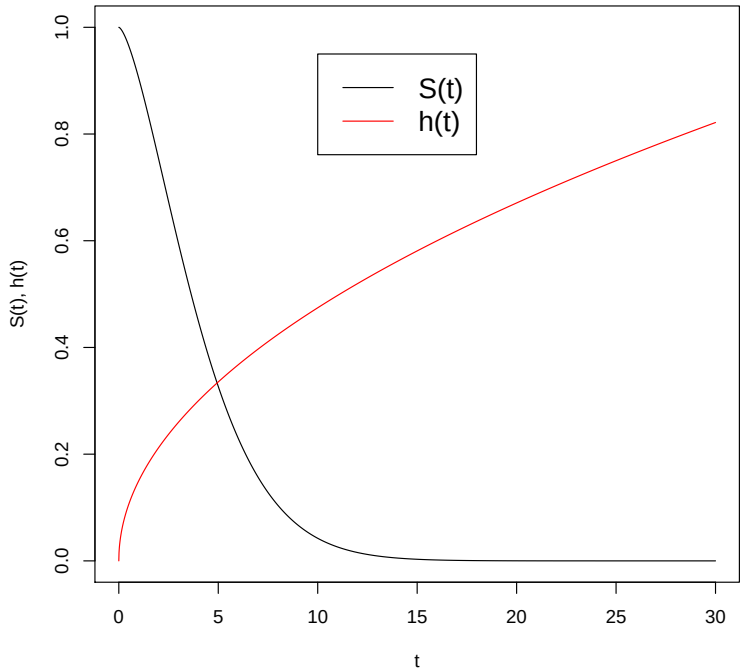
Příklad - Weibullovo rozdělení

$$h(t) = \lambda\gamma t^{\gamma-1}, S(t) = \exp(-\lambda t^{\gamma})$$

$\lambda = 0.1, \gamma = 0.5$



$\lambda = 0.1, \gamma = 1.5$



Věrohodnostní funkce přežití

Parametrické odhady $S(t)$ jsou v analýze přežití založeny na **metodě maximální věrohodnosti**. Věrohodnostní funkce je použita pro odhad neznámých parametrů vybraného rozdělení pravděpodobnosti.

- Příspěvek $i$tého pacienta k věrohodnostní funkci je $f(t_i) = h(t_i)S(t_i)$, když je čas t_i úplným pozorováním.
- Příspěvek $i$tého pacienta k věrohodnostní funkci je $S(t_i)$, když je čas t_i cenzorovaný.

Věrohodnostní funkce v přítomnosti cenzorování pak má tvar

$$L(\beta, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \prod_{i=1}^{n} \mathrm{P}\{t_i, \delta_i\} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i).$$

Věrohodnostní funkci lze s pomocí výrazu $f(t) = h(t)S(t)$ přepsat na

$$L(\beta, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \prod_{i=1}^{n} \mathrm{P}\{t_i, \delta_i\} = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.$$

Logaritmus věrohodnostní funkce

K odhadu parametrů modelu se však nepoužívá věrohodnostní funkce přímo, používá se zejména její logaritmus.

$$L(\beta, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i).$$

$$\begin{aligned} \Rightarrow l(\beta, (t_1, \delta_1), \ldots, (t_n, \delta_n)) &= \log \left(\prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \right) \\ &= \sum_{i=1}^{n} \left(\log h(t_i)^{\delta_i} + \log S(t_i) \right) \\ &= \sum_{i=1}^{n} \left(\delta_i \log h(t_i) - H(t_i) \right) \end{aligned}$$

Příklad - Exponenciální rozdělení

Věrohodnostní funkce pro exponenciální rozdělení (bez vysvětlujících proměnných) má tvar:

$$L(\lambda, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda t_i).$$

Věrohodnostní funkci zlogaritmujeme:

$$l(\lambda, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \sum_{i=1}^{n} (\delta_i \log \lambda - \lambda t_i) = d \log \lambda - \lambda \sum_{i=1}^{n} t_i,$$

kde $d = \sum_{i=1}^{n} \delta_i$ je celkový počet sledovaných událostí.

Maximálně věrohodný odhad parametru λ pak má tvar:

$$\frac{d}{d\lambda} l(\lambda, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \frac{d}{\lambda} - \sum_{i=1}^{n} t_i \rightarrow \hat{\lambda} = \frac{d}{\sum_{i=1}^{n} t_i}.$$

Ověření předpokladu exponenciálního rozdělení

V případě exponenciálního rozdělení je riziková funkce v čase konstantní:

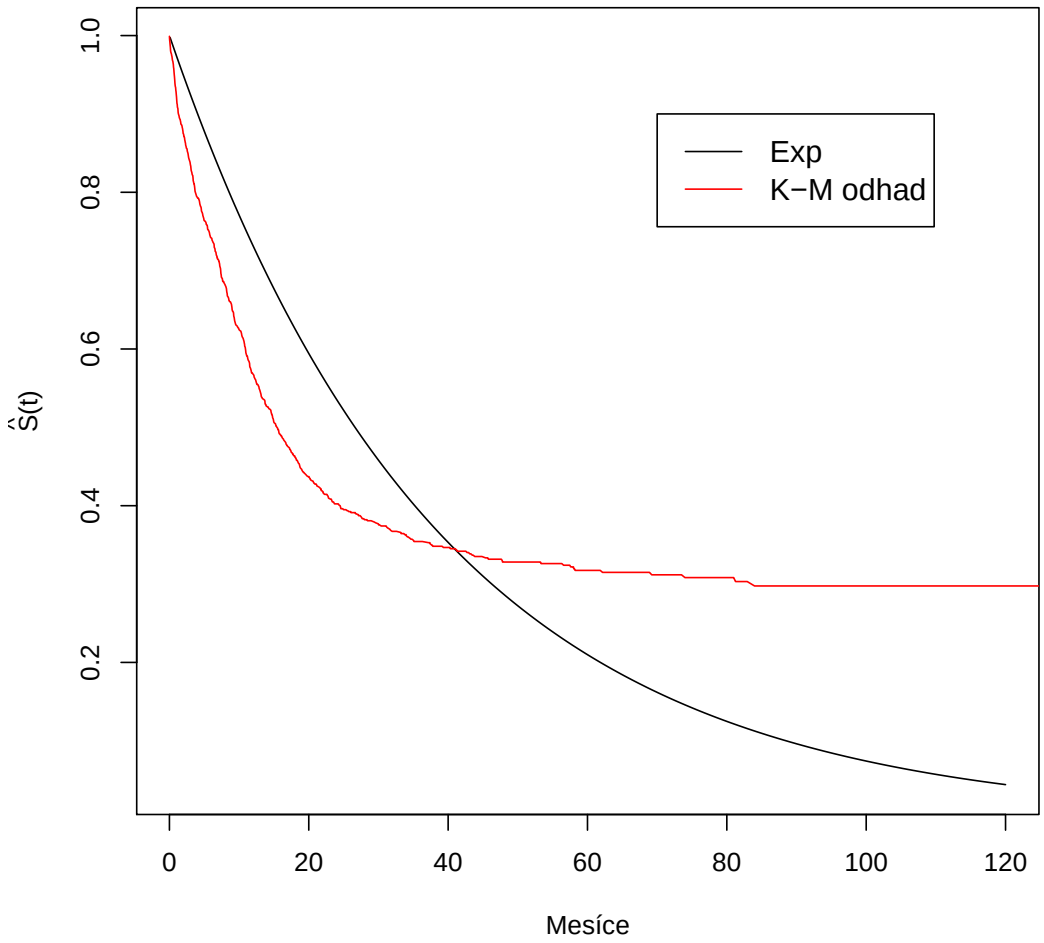
$$h(t) = \lambda$$

- Z toho plyne, že kumulativní riziková funkce je lineární funkcí času.

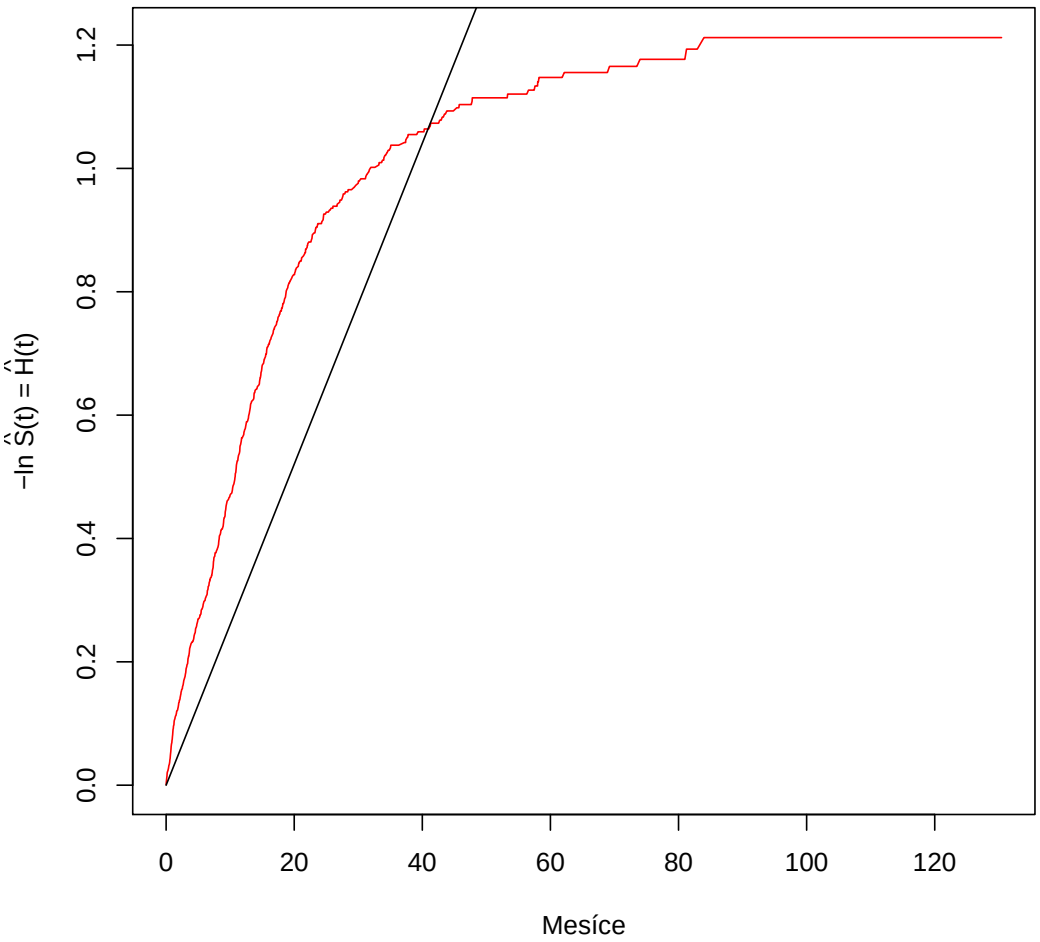
$$h(t) = \lambda \Rightarrow H(t) = \lambda t$$

- Při splnění předpokladu exponenciálního rozdělení, by tedy N-A odhad kumulativní rizikové funkce měl být přímkou. V praxi se častěji vizualizuje $-\log \hat{S}_{K-M}(t)$.

Příklad - data pacientů s AML (N = 1091)



Ověření předpokladu exponenciálního rozdělení



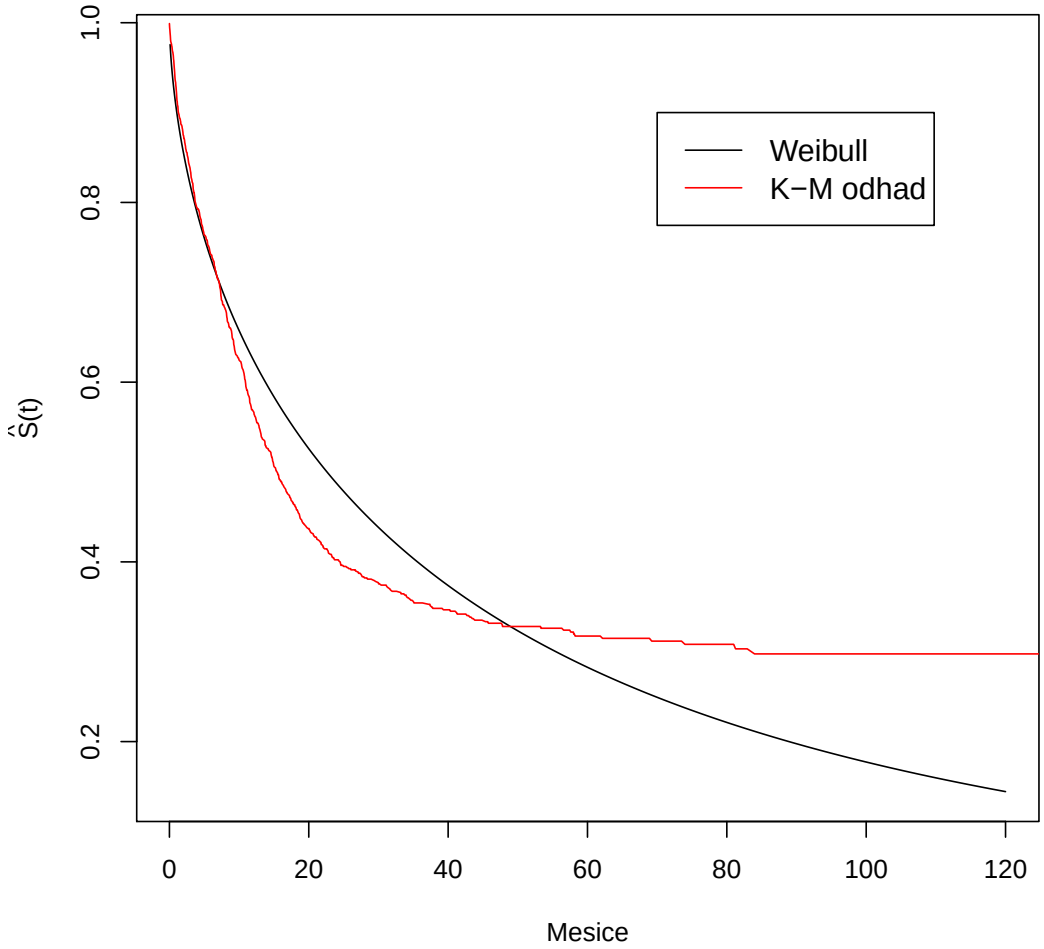
Příklad - Weibullovo rozdělení

Funkce přežití má v případě Weibullova rozdělení tvar $S(t) = \exp(-\lambda t^{\gamma})$. Platí ted následující

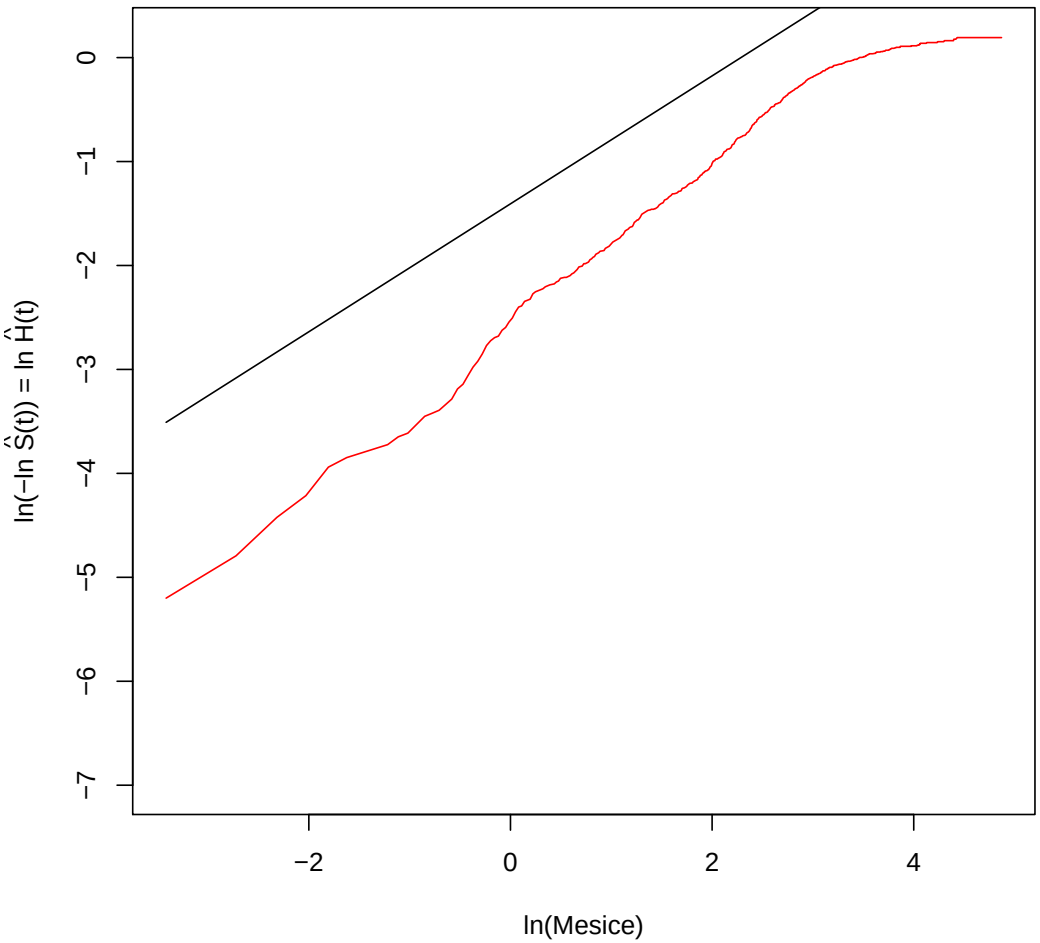
$$\log\left(-\log S(t)\right) = \log H(t) = \gamma\ log\lambda + \gamma \log t.$$

- Logaritmus kumulativní rizikové funkce je tedy v případě vhodnosti Weibullova modelu lineárně závislý na logaritmu času.
- Předpoklad můžeme ověřit pomocí K-M odhadu $S(t)$. Vizualizujeme $\log\left(-\log \hat{S}_{K-M}(t)\right)$ versus $\log t$.

Příklad - data pacientů s AML (N = 1091)



Ověření předpokladu Weibullova rozdělení



Regresní modely v analýze přežití

Pointa regresního modelování

Co dělat, když chceme analyzovat vliv vysvětlující proměnné na přežití (dobu do sledované události)? Možnosti:

- **Vizualizace K-M křivek pro jednotlivé skupiny**
- **Logrank test**
- **Regresní model**

Jaký je mezi nimi rozdíl?

Pointa regresního modelování

Co dělat, když chceme analyzovat vliv vysvětlující proměnné na přežití (dobu do sledované události)? Možnosti:

- **Vizualizace K-M křivek pro jednotlivé skupiny** - umožňuje pouze vizualizaci a optické zhodnocení rozdílu mezi skupinami danými jednou proměnnou. Nebere v úvahu vliv dalších proměnných.
- **Logrank test** - umožňuje statistické zhodnocení rozdílu v přežití (testová statistika, p-hodnota), ale neposkytuje kvantifikaci tohoto rozdílu neboli efektu ("effect size"). Nebere v úvahu vliv dalších proměnných.
- **Regresní model** - umožňuje současně uvažovat vliv více proměnných a vzájemně tak adjustovat jejich vlivy. Zároveň umožňuje kvantifikaci statistické významnosti i velikosti rozdílu/efektu.

Význam regresního modelování

Regresní model umožňuje současně uvažovat vliv více proměnných a vzájemně tak adjustovat jejich vlivy:

Table I Hazard ratios from the Cox PH model for the ovarian dataset

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.809	2.24	(2.03–2.48)	<0.001	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.727	0.48	(0.38–0.61)		−0.422	0.66	(0.50–0.85)	
Endometroid	−1.162	0.31	(0.22–0.45)		0.198	1.22	(0.80–1.85)	
Clear cell	−0.343	0.71	(0.52–0.97)		0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.119	1.13	(0.74–1.72)		0.501	1.65	(0.91–2.99)	
Undifferentiated	0.390	1.48	(0.81–2.70)		0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.614	1.85	(1.28–2.66)		0.789	2.20	(1.45–3.35)	
Grade				<0.001				<0.001
1	(0.000)	(1.00)			(0.000)	(1.00)		
2	1.116	3.05	(1.90–4.91)		0.885	2.42	(1.40–4.19)	
3	1.650	5.20	(3.31–8.18)		0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.798	0.45	(0.37–0.55)	<0.001	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.153	1.17	(1.12–1.21)	<0.001	0.133	1.14	(1.09–1.19)	<0.001

HR = hazard ratio, CI = confidence interval.

Regresní modely - značení

- Index i označuje pacienty.
- Vektor vysvětlujících proměnných: $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$
- Vektor regresních koeficientů příslušných jednotlivým proměnným: $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$
- Základní riziková funkce (baseline hazard function): $h_0(t)$. Může být konstantní nebo závislá na čase.

Regresní modely obecně

- V lineárním regresním modelu jsou vysvětlující proměnné a závisle proměnná spojeny vztahem

$$\boldsymbol{E}(Y_i, \boldsymbol{x}_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p.$$

  Jak lze interpretovat koeficienty $\beta_k, k = 1, \ldots, p$?

- V logistické regresi jsou vysvětlující proměnné a závisle proměnná spojeny vztahem

$$\text{logit}(\pi_i, \boldsymbol{x}_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p.$$

  Jak lze interpretovat koeficienty $\beta_k, k = 1, \ldots, p$?

Lze použít standardní modely v analýze přežití?

- Standardní regresní modely můžeme pro hodnocení dat přežití použít pouze tehdy, nemáme-li cenzorované hodnoty.

- Specifika dat přežití:
  1. **Čas přežití** může nabývat pouze kladných hodnot a **má kladně sešikmené rozdělení pravděpodobnosti**.
  2. Místo odhadu střední hodnoty nás většinou zajímá **odhad pravděpodobnosti přežití v daném časovém bodě**.
  3. Regresní modely v analýze přežití jsou často založeny na rizikové funkci, která nám může dávat lepší informaci o chování přežití sledované skupiny subjektů.

Regresní modely v analýze přežití

V analýze přežití jsou dva hlavní přístupy, jak vyjádřit vztah vysvětlující proměnné a závisle proměnné:

- **Modely proporcionálních rizik** – vysvětlující proměnné modifikují rizikovou funkci:

$$h(t, \boldsymbol{x}_i) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p) = h_0(t) \exp(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

- **AFT modely** (Accelerated Failure Time) – vysvětlující proměnné modifikují funkci (pravděpodobnost) přežití:

$$S(t) = S_0(t \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p)) = S_0(t \exp(\boldsymbol{x}_i'\boldsymbol{\beta})).$$

Modely proporcionálních rizik

Modely proporcionálních rizik

- Nejpoužívanější modely v analýze přežití
- Regresní model je vyjádřen pomocí rizikové funkce takto:

$$h(t, \boldsymbol{x}_i) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p).$$

kde $h_0(t)$ je **základní** (bazální) **riziková funkce** (společná pro všechny subjekty ve studii). Výraz $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ vyjadřuje **relativní riziko** daného subjektu vzhledem k subjektu se základním rizikem ($\boldsymbol{x}_i = \boldsymbol{0}$).

- Funkci $h_0(t)$ můžeme specifikovat parametricky.
- Model můžeme linearizovat jako

$$\log h(t, \boldsymbol{x}_i) = \log h_0(t) + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p.$$

Jak lze interpretovat koeficienty $\beta_1, \ldots, \beta_p$?

- Regresní koeficient β_k pro proměnnou x_k představuje hodnotu, o kterou se navýší logaritmus rizikové funkce při zvýšení x_k o jednu jednotku s tím, že ostatní vysvětlující proměnné se nemění.

$$\begin{aligned}\beta_k &= \log h(t, x_1, x_2, \ldots, x_k + 1, \ldots, x_p) - \log h(t, x_1, x_2, \ldots, x_k, \ldots, x_p) \\ &= \log \frac{h(t, x_1, x_2, \ldots, x_k + 1, \ldots, x_p)}{h(t, x_1, x_2, \ldots, x_k, \ldots, x_p)}\end{aligned}$$

- To znamená:

$$\exp(\beta_k) = \frac{h(t, x_1, x_2, \ldots, x_k + 1, \ldots, x_p)}{h(t, x_1, x_2, \ldots, x_k, \ldots, x_p)}$$

Předpoklady modelů proporcionálních rizik

1. Vztah mezi vysvětlujícími proměnnými a $\log h(t)$ je **lineární**.
2. Vysvětlující proměnné mají **aditivní** vliv na škále $\log h(t)$ (nebereme-li v úvahu interakce).
3. Vliv vysvětlujících proměnných na rizikovou funkci je stejný $\forall t$.

Parametrické modely proporcionálních rizik

Parametrické regresní modely proporcionálních rizik

- **Exponenciální regresní model**

$$h(t, \boldsymbol{x}_i) = h_0 \exp(\boldsymbol{x}_i' \beta) = h_0 \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p),$$

- **Weibullův regresní model**

$$h(t, \boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{x}_i' \beta) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p),$$

- V čem se oba modely liší?

Odhad regresních koeficientů v parametrickém modelu

- Založen na **věrohodnostní funkci** pro cenzorovaná data
- Maximalizujeme logaritmus věrohodnostní funkce

$$\Rightarrow l(\boldsymbol{\beta}, (t_1, \delta_1), \ldots, (t_n, \delta_n)) = \sum_{i=1}^{n} \left(\log h(t_i, \boldsymbol{x}_i)^{\delta_i} + \log S(t_i, \boldsymbol{x}_i) \right)$$

$$= \sum_{i=1}^{n} \left(\delta_i \log h(t_i, \boldsymbol{x}_i) - H(t_i, \boldsymbol{x}_i) \right)$$

- Můžeme použít libovolné parametrické vyjádření $h(t)$ a $S(t)$

Exponenciální regresní model

$$h(t, \boldsymbol{x}_i) = h_0 \exp(\boldsymbol{x}_i' \beta) = \exp(\log h_0 + \boldsymbol{x}_i' \beta)$$

- Důležitým předpokladem je konstantní základní riziko v čase - je třeba ho ověřit.
- Není moc flexibilní (předpoklad konstantního základního rizika v čase je svazující).

Exponenciální regresní model - příklad

Uvažujme data přežití dvou skupin pacientů, tedy jednu vysvětlující proměnnou x_1 nabývající hodnot 0 a 1 (např. podání standardní a experimentální léčby).

$$h(t, x_1 = 0) = h_0 \exp(0 * \beta_1) = h_0$$

$$h(t, x_1 = 1) = h_0 \exp(1 * \beta_1) = h_0 \exp(\beta_1)$$

Model lze přepsat takto (místo x_{i1} budeme psát pouze x_i):

$$h(t, x_i) = h_0 \exp(\beta_1 x_i) = \exp(\log h_0 + \beta_1 x_i) = \exp(\beta_0 + \beta_1 x_i)$$

Logaritmus věrohodnostní funkce má pak tvar:

$$l(\beta_0, \beta_1, (t_i, \delta_i)) = \sum_{i=1}^{n} (\delta_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i) t_i) .$$

Derivace věrohodnostní funkce podle β_0

$$\begin{aligned}\frac{\partial}{\partial \beta_0} l(\beta_0, \beta_1, (t_i, \delta_i)) &= \sum_{i=1}^{n} (\delta_i - \exp(\beta_0 + \beta_1 x_i) t_i) \\ &= \sum_{i=1}^{n_1} (\delta_i - \exp(\beta_0) t_i) + \sum_{i=n_1+1}^{n} (\delta_i - \exp(\beta_0 + \beta_1) t_i) \\ &= d_1 + d_2 - \exp(\beta_0) \sum_{i=1}^{n_1} t_i - \exp(\beta_0 + \beta_1) \sum_{i=n_1+1}^{n} t_i\end{aligned}$$

kde d_1 a d_2 jsou počty pozorovaných událostí a n_1 a $n - n_1$ jsou počty subjektů ve sledovaných skupinách.

Derivace věrohodnostní funkce podle β_1

$$\begin{aligned}\frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1, (t_i, \delta_i)) &= \sum_{i=1}^{n} (\delta_i x_i - \exp(\beta_0 + \beta_1 x_i) t_i x_i) \\ &= \sum_{i=n_1+1}^{n} (\delta_i - \exp(\beta_0 + \beta_1) t_i) \\ &= d_2 - \exp(\beta_0 + \beta_1) \sum_{i=n_1+1}^{n} t_i\end{aligned}$$

kde d_1 a d_2 jsou počty pozorovaných událostí a n_1 a $n - n_1$ jsou počty subjektů ve sledovaných skupinách.

Weibullův regresní model

$$h(t, \boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{x}_i' \beta) = \lambda \gamma t^{\gamma - 1} \exp(\boldsymbol{x}_i' \beta)$$

- Vektor proměnných $\boldsymbol{x} = \boldsymbol{0}$ odpovídá základnímu riziku.
- Tvar rizikové funkce závisí zejména na parametru γ. Je-li např. $\gamma > 1$, riziková funkce je monotónně rostoucí ve všech podskupinách definovaných vysvětlujícími proměnnými.
- Vliv proměnných je multiplikativní. S nárůstem x_k o 1 se zvýší riziko $\exp(\beta_k)$-krát.
- Pro dva subjekty s vektory $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ platí

$$\mathrm{HR} = \frac{h(t, \boldsymbol{x}_2)}{h(t, \boldsymbol{x}_1)} = \exp\left((\boldsymbol{x}_2 - \boldsymbol{x}_1)' \beta\right)$$

Coxův model proporcionálních rizik

Coxův model proporcionálních rizik

Coxův model proporcionálních rizik je dán vztahem

$$h(t, \boldsymbol{x}_i) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p),$$

Předpoklady

- vysvětlující proměnné mají aditivní vliv na škále $\log h(t, \boldsymbol{x}_i)$.
- vliv vysvětlujících proměnných je stejný pro $\forall t$.

Na rozdíl od parametrických modelů nespecifikujeme $h_0(t)$.

Poměr rizik (Hazard ratio, *HR*)

Poměr rizik pro subjekty s vektory vysvětlujících proměnných $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ lze vyjádřit takto:

$$HR = \frac{h(t, \boldsymbol{x}_i)}{h(t, \boldsymbol{x}_j)} = \frac{h_0(t)\exp(\boldsymbol{x}_i'\beta)}{h_0(t)\exp(\boldsymbol{x}_j'\beta)} = \exp((\boldsymbol{x}_i - \boldsymbol{x}_j)'\beta),$$

Bodový odhad pro poměr rizik je

$$\hat{HR} = \frac{h(t, \mathbf{x}_i)}{h(t, \mathbf{x}_j)} = \frac{h_0(t)\exp(\mathbf{x}_i'\hat{\beta})}{h_0(t)\exp(\mathbf{x}_j'\hat{\beta})} = \exp((\mathbf{x}_i - \mathbf{x}_j)'\hat{\beta}),$$

kde $\hat{\beta}$ je maximálně věrohodný odhad β. $100(1-\alpha)\%$ interval spolehlivosti lze získat takto

$$\exp((\mathbf{x}_1 - \mathbf{x}_2)'\hat{\beta} \pm z_{1-\alpha/2}\hat{SE}((\mathbf{x}_1 - \mathbf{x}_2)'\hat{\beta})).$$

Základní riziková funkce

Základní riziková funkce, $h_0(t)$, vyjadřuje riziko pro subjekty, pro které platí $x_1 = 0, \ldots, x_p = 0$. Jedná se o riziko odpovídající referenční skupině subjektů:

$$h(t, \mathbf{x}_i = \mathbf{0}) = h_0(t) \exp(0 * \beta_1 + 0 * \beta_2 + \ldots + 0 * \beta_p) = h_0(t).$$

Pro odhad vlivu vysvětlujících proměnných, HR, základní rizikovou funkci nepotřebujeme. Potřebujeme ji ale pro odhad funkce přežití a rizikové funkce pro každé $\boldsymbol{x}$. Můžeme použít odhad kumulativní rizikové funkce, $H_0(t)$, dle Breslowa:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \hat{h}_0(t_i) = \sum_{t_i \leq t} \frac{d_i}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}})},$$

kde d_i je počet úmrtí v čase t_i a R_i je příslušný počet pacientů v riziku sledované události.

Odhad regresních koeficientů

- Pro odhad regresních koeficientů navrhl Cox metodu parciální věrohodnosti (partial likelihood), kdy je místo věrohodnostní funkce maximalizována tzv. parciální věrohodnostní funkce.
- Ta má pro vektor regresních koeficientů β tvar:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta)}.$$

Logaritmus parciální věrohodnostní funkce pak má tvar:

$$\log L(\beta) = \sum_{i=1}^{k} \left\{ \mathbf{x}_i'\beta - \log \left[\sum_{j \in R_i} \exp(\mathbf{x}_j'\beta) \right] \right\} = \sum_{i=1}^{k} l_i.$$

Sestavení modelu

Cíl studie ovlivňuje výběr vysvětlujících proměnných

Před modelováním je třeba si ujasnit, proč vlastně modelujeme a co od modelu požadujeme.
Modely umí zároveň zohlednit více proměnných – **jejich výběr ale vždy leží na analytikovi**!

Scénáře mohou být následující:

1. Cílem hodnocení je **jediná proměnná**
2. Cílem je **prediktivní model**
3. Cílem je **identifikace potenciálních prediktorů**

Cílem hodnocení je jediná proměnná

- Jde nám o **kvantifikaci a statistickou významnost vlivu jedné vysvětlující proměnné**, např. podané léčby (předpokládáme totiž její vliv na přežití). Typické je modelování výsledků klinických studií.
- Ostatní vysvětlující proměnné vystupují v modelu v roli **adjustačních faktorů** - chceme odstranit jejich vliv.
- V modelu by měly kvůli adjustaci zůstat i proměnné s nevýznamným vlivem - i ty mohou hrát roli v identifikaci vlivu sledované proměnné.

Význam regresního modelování

Regresní model umožňuje současně uvažovat vliv více proměnných a vzájemně tak adjustovat jejich vlivy:

Table I Hazard ratios from the Cox PH model for the ovarian dataset

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.809	2.24	(2.03–2.48)	<0.001	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.727	0.48	(0.38–0.61)		−0.422	0.66	(0.50–0.85)	
Endometroid	−1.162	0.31	(0.22–0.45)		0.198	1.22	(0.80–1.85)	
Clear cell	−0.343	0.71	(0.52–0.97)		0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.119	1.13	(0.74–1.72)		0.501	1.65	(0.91–2.99)	
Undifferentiated	0.390	1.48	(0.81–2.70)		0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.614	1.85	(1.28–2.66)		0.789	2.20	(1.45–3.35)	
Grade				<0.001				<0.001
1	(0.000)	(1.00)			(0.000)	(1.00)		
2	1.116	3.05	(1.90–4.91)		0.885	2.42	(1.40–4.19)	
3	1.650	5.20	(3.31–8.18)		0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.798	0.45	(0.37–0.55)	<0.001	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.153	1.17	(1.12–1.21)	<0.001	0.133	1.14	(1.09–1.19)	<0.001

HR = hazard ratio, CI = confidence interval.

Cílem je prediktivní model

- Z množiny zaznamenávaných proměnných chceme **vybrat sadu s významným vlivem na přežití a schopností jeho predikce**.
- U všech předpokládáme relevanci vzhledem k přežití, ale ve výsledném modelu se vyskytují pouze ty **"nejsilnější"**.
- Klíčové téma je **složitost modelu** (počet vysvětlujících proměnných a jejich interakcí). Statistická významnost proměnné nemusí zaručovat její přínos pro predikci.
- Je třeba rozlišovat regresní model a prediktivní model.

Cílem je identifikace potenciálních prediktorů

- Částečně **exploratorní** práce – z množiny zaznamenávaných proměnných chceme vybrat ty významně asociované s přežitím, např. expresní profily zkoumaných genů (mohou jich být až tisíce).
- Jde nám o redukci počtu vysvětlujících proměnných a identifikaci těch nejvýznamnějších.
- Je třeba dávat pozor na **falešně pozitivní** výsledky.
- Problém $n \ll p$.

Výběr vysvětlujících proměnných I

"The data analyst knows more than computer"

- Klíčové téma, roli zde hraje i úplnost dat - které proměnné si vůbec můžeme dovolit do modelování zahrnout.
- Je-li vysvětlujících proměnných mnoho, je třeba zapojit vícerozměrné metody - **identifikovat shluky proměnných**, které jsou korelované. Následně vybrat pouze reprezentativní zástupce.
- Cílem použití vícerozměrných metod je odstranit redundantní informaci.
- Tento postup nelze automatizovat, závisí na znalosti konkrétního problému.

Výběr vysvětlujících proměnných II

"The data analyst knows more than computer"

- Nejoblíbenější jsou tzv. **stepwise procedury**:
  - **Backward elimination** - vysvětlující proměnné z modelu postupně ubíráme.
  - **Forward selection** - vysvětlující proměnné do modelu postupně přidáváme.
- Nelze je používat slepě, analytik musí vždy nad modelováním přemýšlet a pracovat s literaturou/odborníkem.
- Zvlášť obtížné je **modelování interakcí**.

Rozvaha nad velikostí vzorku

Velikost vzorku je zvlášť v analýze přežití extrémně důležitá $\Rightarrow$ jde nám hlavně o dostatečný **počet sledovaných událostí** (problém cenzorování).

- Právě čas do sledované události představuje v analýze přežití informaci!
- Velikost vzorku by měla být vždy plánována před zahájením experimentu (pomocí software).
- Peduzzi a kol. (1995) navrhli na základě simulací **alespoň 10 událostí na 1 vysvětlující proměnnou** zahrnutou do modelu.

Co když nemáme alespoň 10 událostí na 1 vysvětlující proměnnou?

- Bodové i intervalové odhady regresních koeficientů budou nevěrohodné.
- Model může selhat ve výběru významných proměnných.

Kategorizace spojitých proměnných

Kategorizace spojitých proměnných by měla být prováděna co nejméně, protože se jedná o **ztrátu informace**, která může zvýšit variabilitu a vést ke zkreslení výsledků. Někdy je však důležitá z hlediska interpretace, např. věkové kategorie (0 – 50 let, 51 – 60 let, 61 – 70 let, nad 70 let).

Kategorizace může být použita pro řešení nelinearity vlivu vysvětlující proměnné. Doporučení pro kategorizaci spojitých proměnných:

1. Kategorie by měly mít logiku, případně odrážet biologickou podstatu.
2. Je třeba dávat pozor na počty subjektů v kategoriích, nejlépe je mít kategorie vyvážené.

Regresní diagnostika

Co je regresní diagnostika?

Nástroje regresní diagnostiky slouží pro **hodnocení vhodnosti modelu** (*goodness of fit*) a **splnění předpokladů modelu** (*model assumptions*). Cílem je zjistit, zda data **nejsou v rozporu** s předpoklady modelu a zda náš model **adekvátně** vystihuje přežití sledovaného souboru.

Hodnocení vhodnosti modelu je těsně spjato s výběrem modelu a výběrem vysvětlujících proměnných.

Nástroje regresní diagnostiky:

1. **Vizualizace neparametrických odhadů charakteristik přežití**
2. **Výpočet reziduí modelu a jejich vizualizace (grafy reziduí)**
3. **Akaikeho informační kritérium** (AIC)
4. **Statistické testy**

Hodnocení vhodnosti modelu

"Každý model je špatný, ale některý může být i užitečný."

- Velmi náročný úkol, protože objektivně není dáno, co je vhodný model a co už ne.
- Pro **srovnání dvou modelů** lze použít tzv. Akaikeho informační kritérium.
- Pro **hodnocení celkového fitu** lze použít test dle Parzena & Lipsitze (1999) založený na podobném principu jako Pearsonův chí-kvadrát test pro kontingenční tabulky.

Akaikeho informační kritérium (AIC)

AIC (Akaike, 1974) je statistika **zahrnující věrohodnost modelu a jeho složitost**. Slouží k posouzení schoposti různých modelů fitovat pozorovaná data.

$$AIC = -2\ell(\beta, (t_1, \delta_1), \ldots, (t_n, \delta_n)) + 2(c + a)$$

kde $\ell(\beta, (t_1, \delta_1), \ldots, (t_n, \delta_n))$ je logaritmus věrohodnostní funkce modelu, c je počet vysvětlujících proměnných v modelu a a je počet parametrů uvažovaného rozdělení pravděpodobnosti.

- Nižší hodnoty AIC indikují lepší model.
- Nelze srovnávat AIC parametrických modelů a Coxova modelu. Proč?

Rezidua modelu

Obecně lze rezidua definovat jako **rozdíl mezi pozorovanou a predikovanou hodnotou** sledované veličiny. Velké hodnoty reziduí, případně jejich "systematické chování", jsou indikátorem špatného modelu.

Kvůli cenzorování není jednoduché hodnoty reziduí interpretovat – přínosem je **grafická vizualizace** a **vyhlazení trendu** (např. jádrovým vyhlazováním). Obecně lze říci, že by **grafy reziduí neměly vykazovat žádný trend** (rezidua by měla tovřit rovnoměrný horizontální pás).

Vybrané typy reziduí v analýze přežití:

- **Martingale** rezidua
- **Deviance** rezidua
- **Skórová** rezidua
- **Schoenfeldova** rezidua

Martingale rezidua

Martingale reziduum představuje **rozdíl mezi pozorovaným a předpokládaným počtem událostí** (dle daného modelu) u subjektu i. Předpokládaný počet událostí ($\hat{E}_i$) je reprezentován kumulativním rizikem do času t_i.

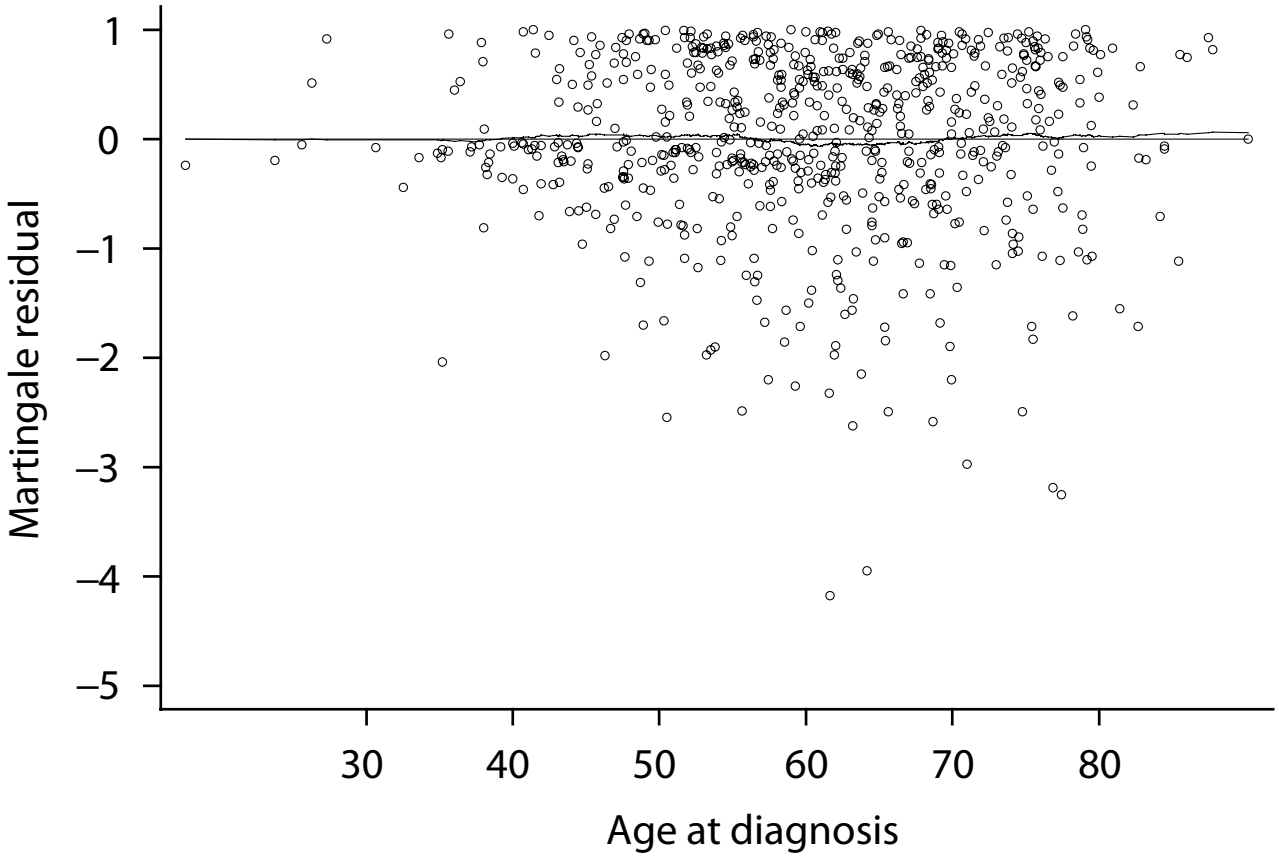
$$\hat{r}_i^M = \delta_i - \hat{E}_i = \delta_i - \hat{H}_0(\hat{\beta}, t_i) \exp(\mathbf{x}_i' \hat{\beta})$$

- Martingale reziduum je jedno číslo charakterizující shodu pozorování s předpokládaným rizikem. Reziduum deviance je také jedno číslo.
- Skórové reziduum i Schoenfeldovo reziduum je vektor hodnot - každý subjekt má jednu hodnotu pro každou z vysvětlujících proměnných.

Význam jednotlivých reziduí

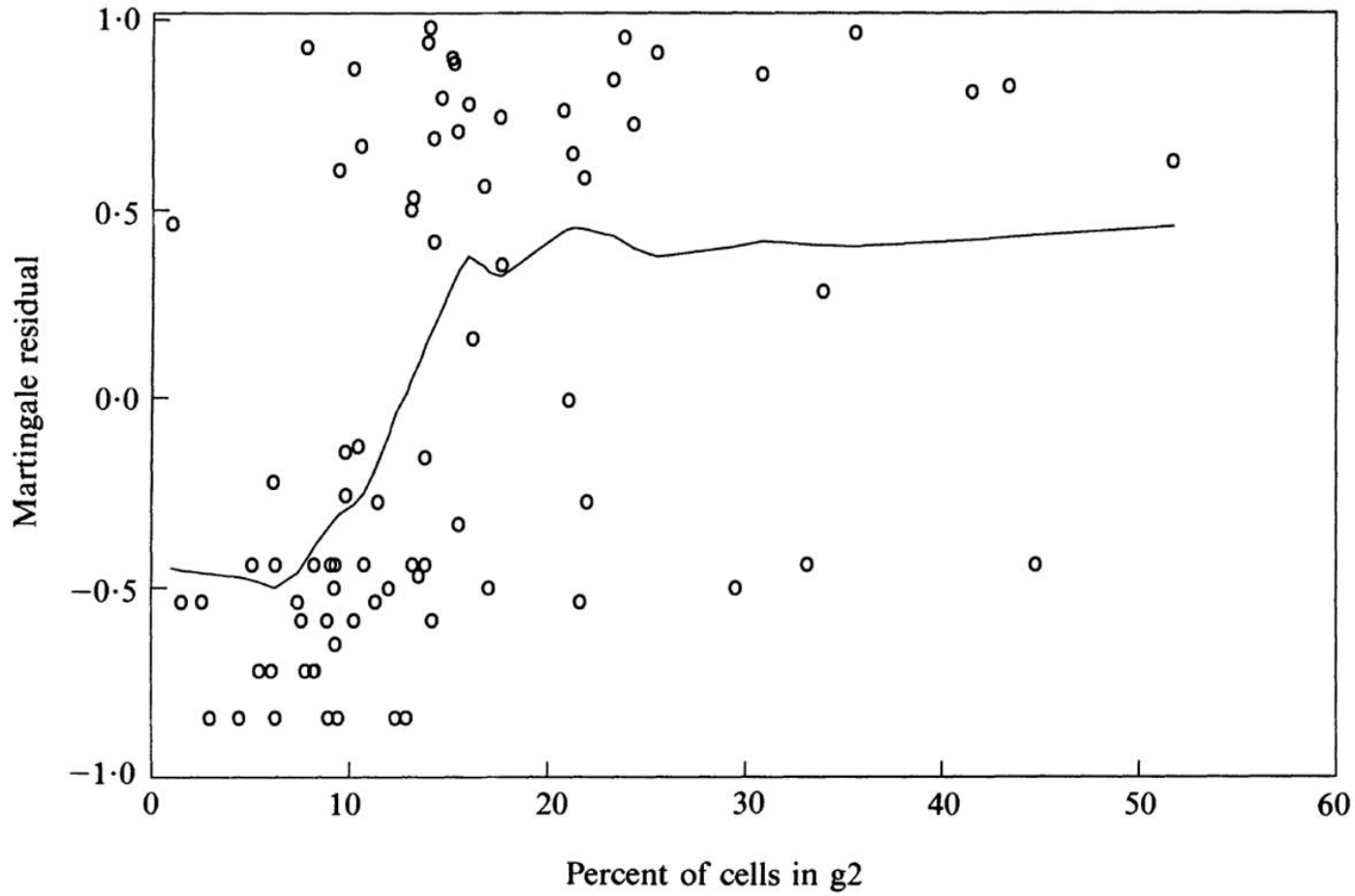
Rezidua	Použití v regresní diagnostice
Martingale	Identifikace nelineárního vlivu proměnné
Martingale	Identifikace nesprávně vyloučené proměnné
Deviance	Identifikace odlehlých pozorování/jedinců
Skórová	Identifikace vlivných pozorování/jedinců
Schoenfeldova	Testování předpokladu proporcionality rizik

Martingale rezidua - příklad 1



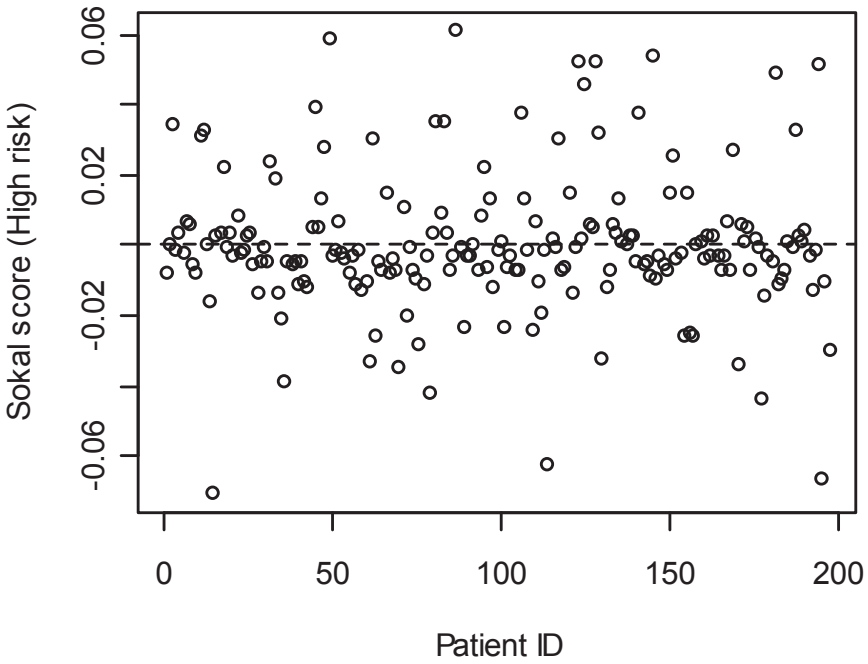
Zdroj: Bradburn et al. (2003) Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit.

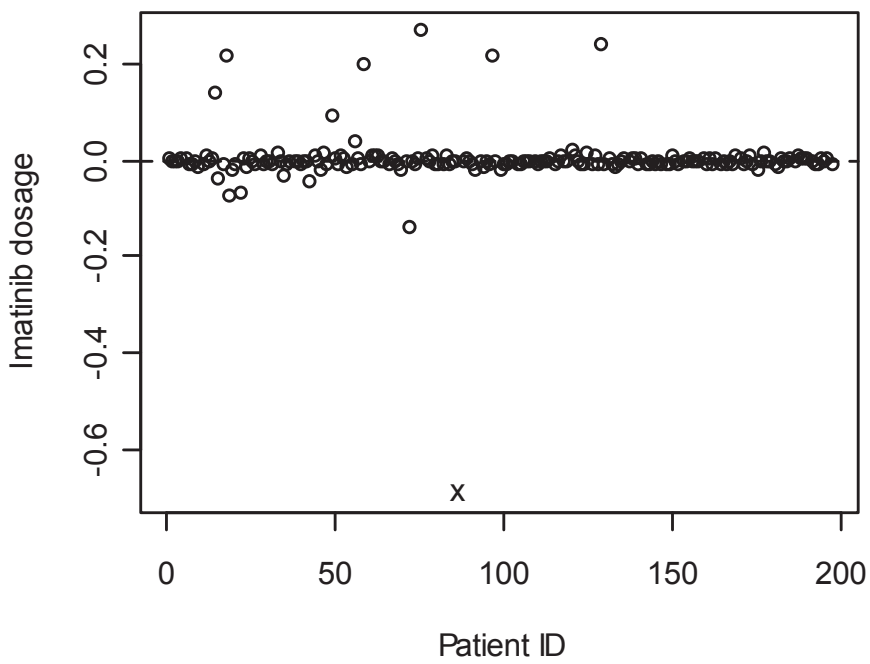
Martingale rezidua - příklad 2



Zdroj: Therneau et al. (1990) Martingale-Based Residuals for Survival Models.

Skórová rezidua - příklad





Zdroj: Analýza dat registru CAMELIA.

Ověření předpokladu proporcionality rizik

1. **Grafické ověření** - nejjednodušší forma ověření, K-M křivky by se měly "rovnoměrně vzdalovat", v žádném případě se nesmí křížit. Ještě vhodnější je použití logaritmu kumulativní rizikové funkce.
2. **Test pomocí časově závislé proměnné**
3. **Test založený na škálovaných Schoenfeldových reziduích**

Grafické ověření proporcionality rizik

Máme k dispozici jednoduchou pomůcku pro ověření *jednorozměrné* proporcionality, neoť funkce přežití vzhledem k proměnné k musí v případě Coxova modelu splňovat

$$S_k(t) = \exp(-H_0(t)\exp(x_k\beta_k)),\ k = 1, \ldots, p,$$

a tedy platí

$$\log[-\log(S_k(t))] = \log H_0(t) + x_k\beta_k,$$

- Pokud je předpoklad proporcionality splněn, křivky $\log[-\log(S_k(t))]$, pro jednotlivé hodnoty proměnné k budou přibližně paralelní.
- Grafické ověření však nezohledňuje více faktorů.

Accelerated Failure Time model

Accelerated Failure Time (AFT) model

1. Čas přežití $i$tého subjektu, T_i, je nezáporný $\rightarrow$ můžeme modelovat jeho logaritmus.
2. AFT model je pak definován jako:

$$\log T_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$$

nebo

$$\log T_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \mu + \sigma\epsilon_i,$$

kde ϵ_i je reziduální člen s daným rozdělením pravděpodobnosti.

$$T_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\exp(\epsilon_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\,T_{0i}$$

AFT model - alternativní zápis

Jinou formou zápisu AFT modelu je zápis pomocí funkce přežití:

$$S_i(t) = S_0(t/\exp(\boldsymbol{x}_i'\beta)),$$

kde $S_0(t)$ představuje základní funkci přežití odpovídající referenční skupině s vektorem vysvětlujících proměnných $\boldsymbol{x}_i = \mathbf{0}$.

- Regresní AFT model je vhodnou alternativou pro model proporcionálního rizika tehdy, když je předpoklad proporcionality rizik porušen.
- Odhad regresních koeficientů je opět založen na metodě maximální věrohodnosti.

Relativní přežití

Vyjádření pravděpodobnosti přežití pacientů

- Nejčastěji jako tzv. **celkové přežití** (*overall survival*), někdy označováno také jako pozorované přežití (*observed survival*).
- Celkové přežití odráží celkovou mortalitu pacientů bez ohledu na přesnou příčinu úmrtí.
- Chceme-li kvantifikovat mortalitu spojenou pouze se sledovanou diagnózou, musíme zohlednit pouze *vybrané* příčiny úmrtí. Jak?

1. Výpočet **přežití specifického dle diagnózy** (*cause-specific survival*)
2. Výpočet **relativního přežití** (*relative survival*)

Přežití specifické pro sledovanou diagnózu

- **Výpočet je jednoduchý**, jediný rozdíl proti výpočtu celkového přežití je v tom, že časy pacientů, kteří zemřeli z jiné příčiny než z příčiny základního onemocnění, jsou cenzorovány.
- Problém s výpočtem přežití specifického pro sledovanou diagnózu však nastává při hodnocení populačních dat, která svojí kvalitou záznamů nemohou monitorovaným klinickým studiím konkurovat.
- **Problém s kódováním přesné příčiny úmrtí** u onkologických pacientů nemusí být v administrativě záznamu, ale spíše v jeho nejednoznačnosti, protože ne vždy je klinicky zřejmé, jestli pacient zemřel v souvislosti se sledovaným onemocněním nebo ne.

Relativní přežití

Relativní přežití představuje **poměr celkového** a tzv. **očekávaného přežití** (*expected survival*). Očekávané přežití odráží mortalitu v obecné populaci, která odpovídá sledované skupině pacientů věkem, pohlavím a obdobím diagnózy.

- Relativní přežití $\approx$ vážený ekvivalent celkového přežití, váhou je přežití obecné populace.
- Relativní přežití $\approx$ celkové přežití korigované na mortalitu spojenou s dalšími chorobami, na něž může pacient zemřít.
- Relativní přežití $\approx$ odhad pravděpodobnosti přežití, který odpovídá pouze zátěži představované sledovanou diagnózou.

Relativní přežití představuje míru tzv. *excess mortality*, která je asociována s danou chorobou bez ohledu na to, zda přímo či nepřímo.

Relativní přežití

Označme $h^*(t)$ a $S^*(t)$ očekávanou rizikovou funkci a očekávanou funkci přežití obecné populace. Obdobně označme $h(t)$ a $S(t)$ pozorovanou rizikovou funkci a funkci celkového přežití.

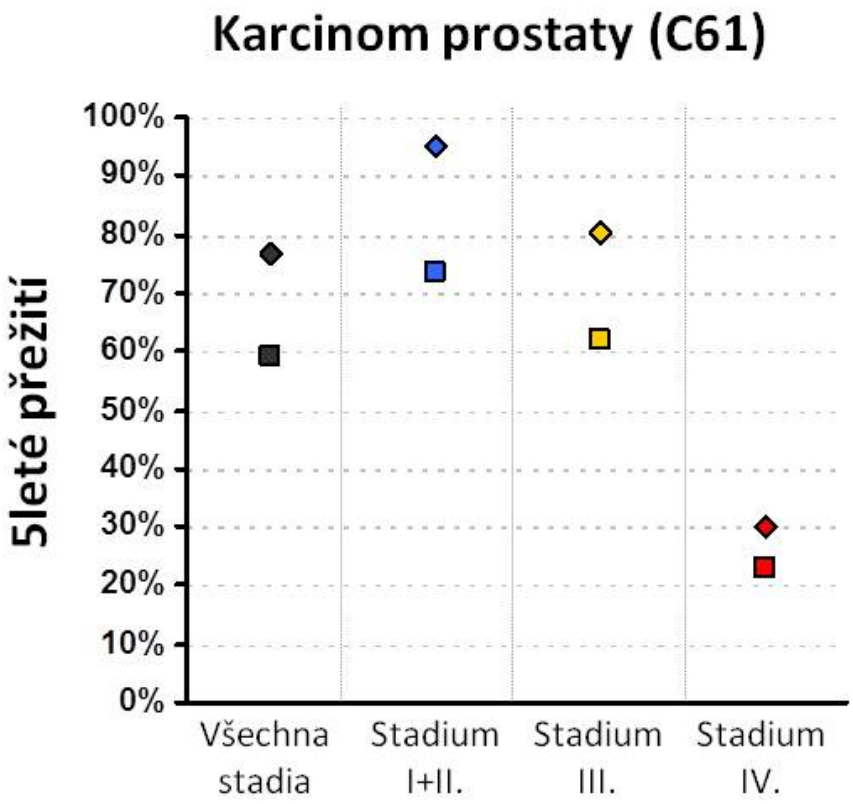
Relativní rizikovou funkci, označme ji $h^R(t)$, pak vypočteme jako

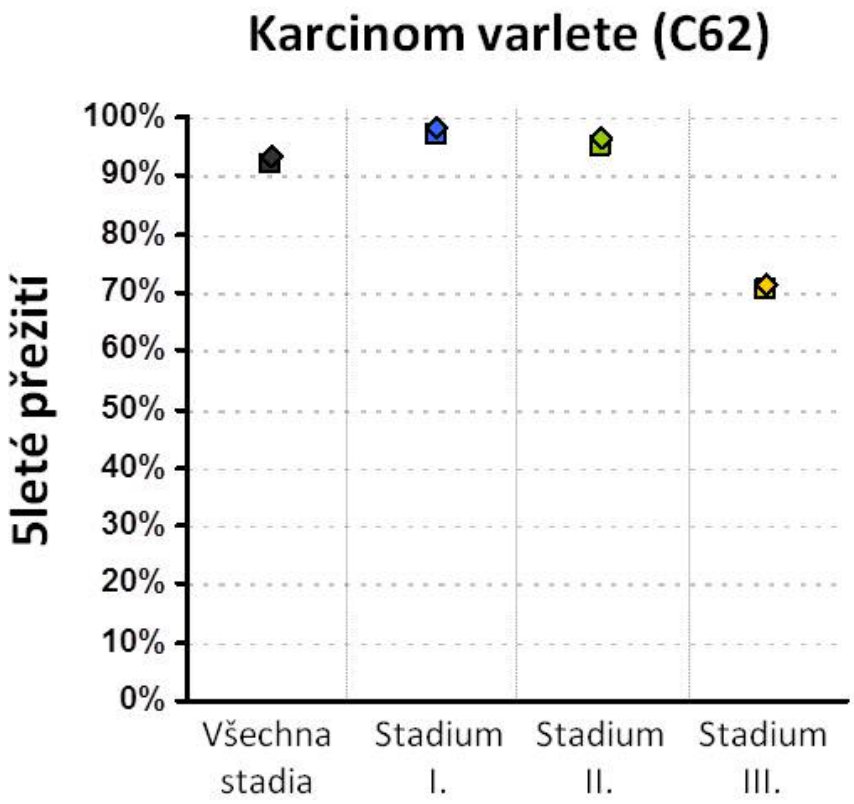
$$h^R(t) = h(t) - h^*(t),$$

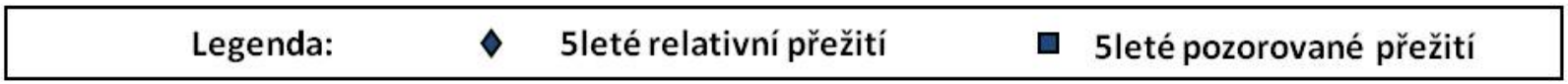
a **relativní funkci přežití**, $S^R(t)$, jako

$$S^R(t) = \frac{S(t)}{S^*(t)}.$$

Relativní přežití - příklad







Zdroj: Data NOR.

Metody odhadu očekávaného přežití

Principem výpočtu **očekávaného přežití** je výpočet přežití z populačních mortalitních tabulek **odpovídajícího srovnatelné skupině** (vzhledem k věku, pohlaví a roku diagnózy) z obecné populace, u které předpokládáme, že prakticky není zasažena sledovaným onemocněním. Metody výpočtu očekávaného přežití:

- Edererova metoda I
- Edererova metoda II
- Hakulinenova metoda

Všechny metody vycházejí ze stejných datových podkladů, tedy úmrtnostních tabulek pro danou populaci (stát). Mezinárodní zdroj: www.mortality.org.

Rozdíl v metodách odhadu očekávaného přežití

- **Edererova metoda I**: srovnatelná skupina z obecné populace je uvažována *v riziku* bez omezení, nebereme tedy ohled na mortalitu a cenzorování v hodnocené kohortě.
- **Edererova metoda II**: srovnatelná skupina z obecné populace je uvažována *v riziku* pouze do úmrtí nebo cenzorování odpovídajícího jedince v hodnocené kohortě.
- **Hakulinenova metoda**: zařazení do skupiny osob *v riziku* bere ohled na cenzorování, ale v případě úmrtí je srovnatelná skupina z obecné populace *v riziku* až do ukončení sledování.

Metody jsou sice výpočetně odlišné, nicméně pokud odhadujeme pouze krátkodobé přežití (např. 5leté relativní přežití), jejich výsledky jsou velmi podobné.

Interval spolehlivosti pro relativní přežití

Bodový odhad relativního přežití v čase t je nutné doplnit intervalovým odhadem. Pointou výpočtu je, že očekávané přežití bereme za konstantní (jeho variabilitu zanedbáváme).

$$\text{var}\left(S^R(t)\right) = \text{var}\left(\frac{S(t)}{S^*(t)}\right) = \frac{\text{var}(S(t))}{S^*(t)^2} = \frac{SE(S(t))^2}{S^*(t)^2}.$$

- Můžeme si to dovolit?

Intervalově specifické relativní přežití

- Relativní přežití nejčastěji počítáme pomocí metody úmrtnostních tabulek.
- Rozdělíme-li délku sledování do l intervalů, lze kumulativní formu relativního přežití vyjádřit jako

$$S^R(t) = \prod_{i=1}^{l} p_i^R,$$

kde p_i^R je tzv. **intervalově specifické relativní přežití**.

- Hodnoty intervalově specifického relativního přežití souvisí s pojmem statistické vyléčení. Jak?

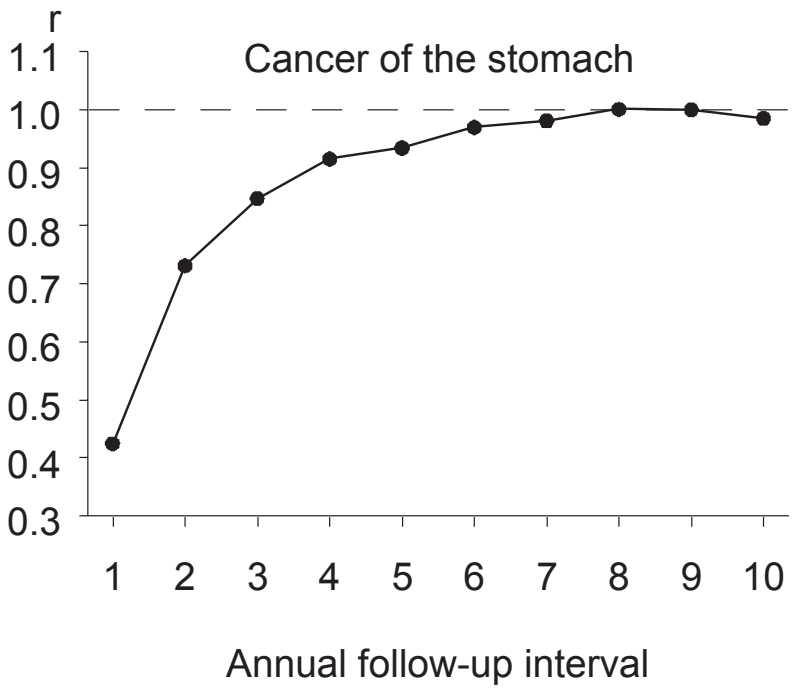
Statistické vyléčení

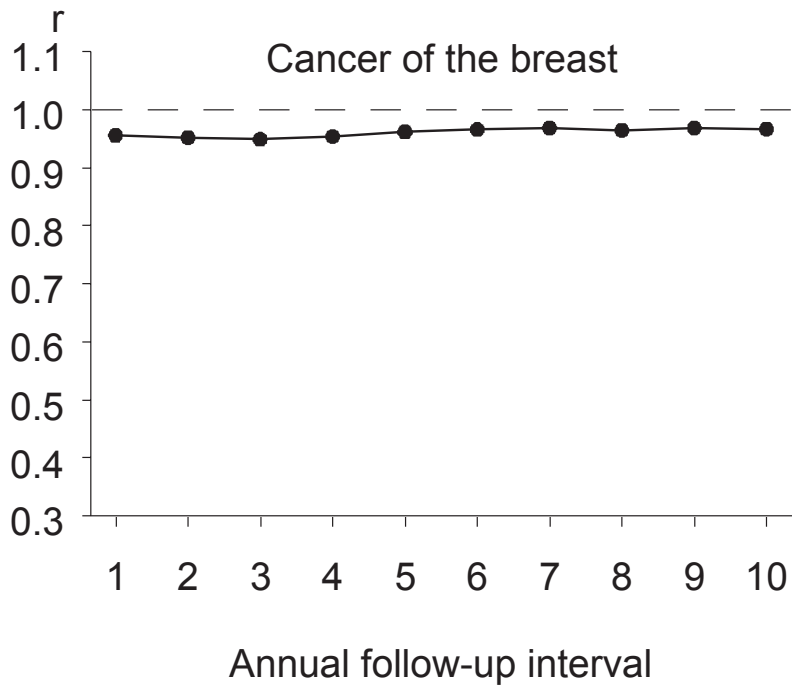
Ve chvíli, kdy se odhady intervalově specifického relativního přežití dosáhnou hodnoty 1, lze říci, že se mortalita sledovaných pacientů v daném intervalu dostala na úroveň mortality populační.

V tomto případě pak mluvíme o tzv. **statistickém vyléčení**.

- Nelze zaměňovat pojmy klinické a statistické vyléčení.
- Pojem **klinické vyléčení** chápeme na úrovní jedince jako vymizení všech klinických projevů nemoci.
- Pojem **statistické vyléčení** chápeme na úrovni skupiny pacientů jako srovnání mortality s populační úrovní.

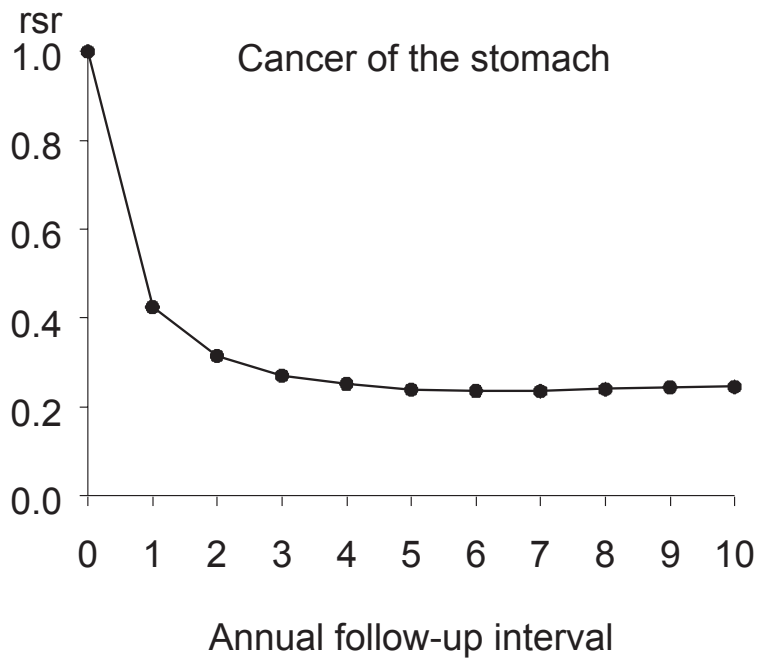
Statistické vyléčení - příklad

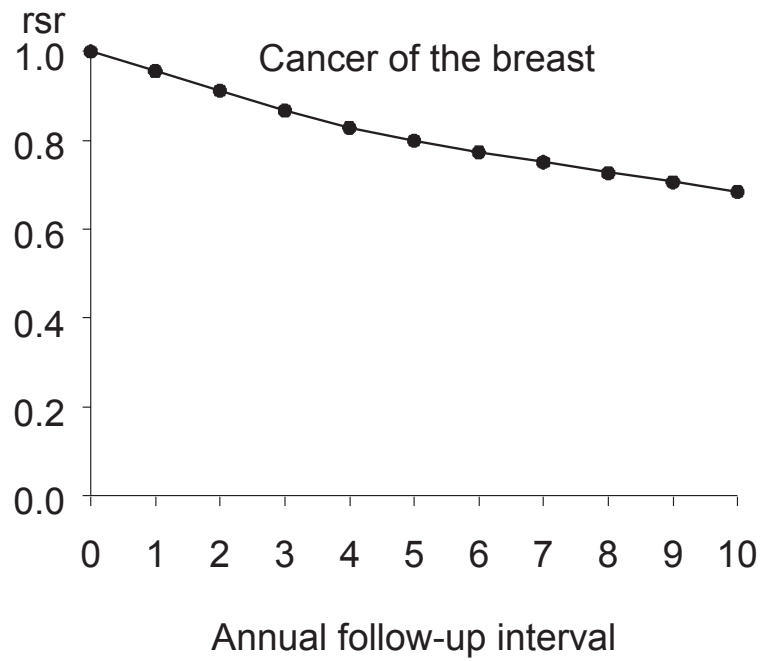




Zdroj: Dickman PW. (2004) Population-based cancer survival analysis.

Statistické vyléčení - příklad





Zdroj: Dickman PW. (2004) Population-based cancer survival analysis.

Modely s podílem vyléčených pacientů

Modely s podílem vyléčených pacientů

Pointou **modelů s podílem vyléčených pacientů** (*cure fraction models*) je rozdělení sledovaných pacientů (na základě rizika úmrtí) do dvou skupin:

- **Nevyléčení pacienti**, jejichž úmrtí lze přičítat sledovanému onemocnění.
- **Vyléčení pacienti**, u nichž k úmrtí ve sledovaném období nedošlo (nebo případně zemřeli z jiných příčin, než bylo sledované onemocnění).

Jsou-li v dané kohortě pacientů přítomni tzv. *statisticky vyléčení pacienti*, křivka funkce přežití, která jde zpravidla k nule, se *narovná* k pomyslné nenulové asymptotě.

Modelování celkového a relativního přežití

Modely s podílem vyléčených pacientů lze použít jak pro modelování **celkového přežití** (využití v klinických analýzách), tak i **relativního přežití** (využití v populačních analýzách). Nejpoužívanějšími modely jsou tzv. **smíšené modely**, které jsou definovány jako

- Model pro **celkové přežití**:

$$S(t) = c + (1 - c)S_U(t),$$

kde c je tzv. *podíl statisticky vyléčených pacientů* a $S_U(t)$ je funkce přežití pro tzv. *nevyléčené pacienty* (*uncured, bound to die*).

- Model pro **relativní přežití**:

$$S(t) = S^*(t)S^R(t) = S^*(t)\left(c + (1 - c)S_U(t)\right),$$

kde $S^*(t)$ je očekávané přežití odpovídající populace.

Smíšené a nesmíšené modely

Kromě smíšeného modelu s podílem vyléčených pacientů byly definovány i tzv. **smíšené modely** s podílem vyléčených pacientů, které jsou definovány jako

- Model pro **celkové přežití**:

$$S(t) = c^{F(t)},$$

kde c je tzv. *podíl statisticky vyléčených pacientů* a $F(t)$ je vhodně zvolená distribuční funkce.

- Model pro **relativní přežití**:

$$S(t) = S^*(t)c^{F(t)},$$

kde $S^*(t)$ je očekávané přežití odpovídající populace.

Používaná rozdělení pravděpodobnosti

Lze definovat i **semi-parametrické modely** s podílem vyléčených pacientů, ale většina těchto modelů je parametrických s některým z následujících rozdělení pravděpodobnosti:

- Weibullovo rozdělení - 2 parametry
- Log-normální rozdělení - 2 parametry
- Gamma rozdělení - 3 parametry
- Směs dvou Weibullových rozdělení

Co vlastně u těchto modelů modelujeme?

Naším cílem je modelování charakteristik týkajících se

1. **vyléčených pacientů** - modelujeme c, tedy *podíl statisticky vyléčených pacientů*.
2. **nevyléčených pacientů** - modelujeme parametry rozdělení pravděpodobnosti (např. λ a γ v případě Weibullova rozdělení).

- Odhad regresních koeficientů je opět založen na metodě maximální věrohodnosti.
- Podle toho, co všechno může záviset na vysvětlujících proměnných, roste náročnost na počet pozorování (sledovaných událostí).