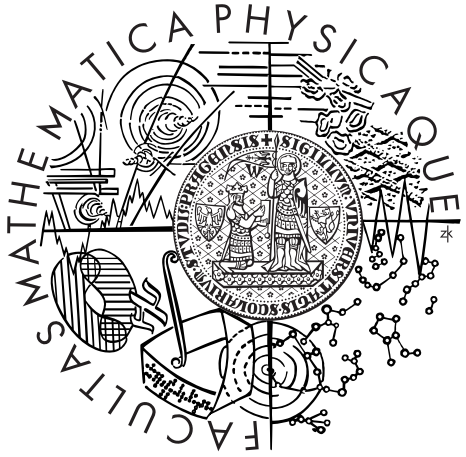
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Adéla Drabinová

Modely pro přežití s možností vyléčení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	doc. Mgr. Michal Kulich, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie

Praha 2016

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 13. května 2016 Adéla Drabinová

Název práce: Modely pro přežití s možností vyléčení

Autor: Adéla Drabinová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Mgr. Michal Kulich, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zabýváme modely pro přežití, kdy uvažujeme, že s kladnou pravděpodobností k relapsu nikdy nedojde, protože pacient se vyléčí. Zaměřujeme se především na dvousložkový směsový model a model s biologickou motivací. Pro každý z nich je odvozen odhad pravděpodobnosti vyléčení a pro nevyléčené pacienty také odhad funkce přežití pro čas do relapsu metodou maximální věrohodnosti. Dále předpokládáme, že jak pravděpodobnost vyléčení, tak doba do relapsu mohou být ovlivněny vysvětlujícími veličinami. Modely jsou následně porovnány v simulační studii.

Klíčová slova: analýza přežití, pravděpodobnost vyléčení, EM algoritmus

Title: Cure-rate models

Author: Adéla Drabinová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Michal Kulich, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this work we deal with survival models, when we consider that with positive probability some patients never relapse because they are cured. We focus on two-component mixture model and model with biological motivation. For each model, we derive estimate of probability of cure and estimate of survival function of time to relaps of uncured patients by maximum likelihood method. Further we consider, that both probability of cure and survival time can depend on regressors. Models are then compared through simulation study.

Keywords: survival analysis, probability of cure, EM algorithm

Chtěla bych poděkovat svému vedoucímu doc. Mgr. Michalu Kulichovi, Ph.D. za trpělivost, ochotu a cenné rady. Děkuji také Mgr. Lence Drahotuské za gramatickou korekturu práce.

Obsah

Úvod

V klinických studiích se stále častěji setkáváme s daty, u kterých lze pozorovat, že křivky přežití mají tendenci se ustálit s rostoucím časem na nenulové hodnotě. To naznačuje, že u části pacientů se již zkoumané onemocnění neobjeví a tyto pacienty lze považovat za vyléčené. Názorným příkladem může být studie karcinomu prsu analyzovaná v Farewell (1986), kde se Kaplan-Meierovy křivky přežití pacientů pro tři různé léčby ustalují nad hodnotou 0,4. Otázkou této studie pak je, jaký podíl pacientů je vyléčen a které faktory, včetně léčebné metody, mají na tento podíl vliv. Přirozeně nás také zajímá doba přežití nebo doba do relapsu u nevyléčených pacientů.

Událost vyléčení pacienta ovšem nelze přímo nikdy pozorovat. Navíc pacient je na konci studie censorován bez ohledu na to, zda se uzdravil, či nikoliv. V takovém případě klasické modely pro přežití nemusí být postačující.

V této práci se zabýváme modely pro přežití s možností vyléčení. Jedná se o obecnější třídu modelů, která nám může pomoci lépe popsat data. Toto téma bylo v literatuře hojně diskutováno mnohými autory, kteří uvádí různé přístupy k dané problematice, včetně dvousměsových modelů (Berkson a Gage, 1952; Kuk a Chen, 1992; Sy a Taylor, 2000; Yin a Ibrahim, 2005), Bayesovských modelů (Chen a kol., 1999; Ibrahim a kol., 2001), modelu procesu vzniku a zániku (Hanin, 2001), modelu založeném na Box-Coxově transformaci (Yin a Ibrahim, 2005) či třídě modelů založené na nelineární transformaci (Tsodikov, 2002).

Hlavním úkolem této práce je popsat a porovnat různé přístupy k formulaci, analýze a interpretaci těchto modelů. Zaměříme se především na dvousměsový model představený poprvé v Boag (1949) , později také v Berkson a Gage (1952), a model s biologickou motivací uvedený v Chen a kol. (1999) jako na dva nejvýraznější představitele této třídy modelů. Pro ně pak odvodíme odhady parametrů a jejich vlastnosti metodou maximální věrohodnosti a následně je porovnáme pomocí simulační studie.

Kapitola 1

Modely

V této kapitole se budeme věnovat konkrétním modelům pro přežití, které počítají s možností vyléčení. Na začátku si uvedeme dva základní pojmy.

Nechť X je náhodná veličiny představující čas do relapsu s funkcí přežití $\bar{F}(t) = 1 - F(t)$, kde $F(t)$ je distribuční funkce. Dále předpokládejme, že k relapsu nemusí v konečném čase dojít. To nás vede k pojmu nevlastní funkce přežití.

Definice 1.1. Řekneme, že funkce $\bar{F}(t) = \mathsf{P}(X < t)$ je nevlastní funkce přežití nezáporné náhodné veličiny X, jestliže je zprava spojitá, nerostoucí a platí, že $\lim_{t\to 0} \bar{F}(t) = 1$ a $\lim_{t\to\infty} \bar{F}(t) \geq 0$.

V práci dále předpokládáme, že čas do relapsu je spojitá náhodná veličina. Dalším důležitým pojmem je pravděpodobnost vyléčení.

Definice 1.2. Nechť $\bar{F}(t)$ je nevlastní funkce přežití nezáporné náhodné veličiny X. Řekneme, že p je pravděpodobnost vyléčení, jestliže platí

$$p = \lim_{t\to\infty} \bar{F}(t) = \mathsf{P}(X = \infty).$$

Pro spojité modely můžeme pravděpodobnost vyléčení p vyjádřit pomocí rizikové funkce $\lambda(t)$:

$$p = \lim_{t\to\infty} \bar{F}(t) = \lim_{t\to\infty} \exp\left(-\int_0^t \lambda(s)\mathrm{d}s\right) = \exp\left(-\int_0^\infty \lambda(s)\mathrm{d}s\right).$$

Asi nejznámější a nejpoužívanější model byl představen v Berkson a Gage (1952). Tento dvousložkový směsový model pro přežití předpokládá, že pacienty lze rozdělit do dvou skupin; první skupina je vyléčená, druhá nikoliv. Neboli každou nevlastní funkci přežití lze zapsat v následujícím tvaru:

$$\bar{F}(t) = p + (1 - p)\bar{F}_0(t), \tag{1.1}$$

kde p je pravděpodobnost vyléčení a $\bar{F}_0(t)$ je funkce přežití pro relaps pro nevyléčené pacienty. Jinými slovy, je-li X čas do relapsu, pak $\bar{F}_0(t) = \mathsf{P}(X \geq t | X < \infty)$. Model (1.1) budeme dále nazývat jako standardní model.

Následující tvrzení uvádí alternativní reprezentaci nevlastní funkce přežití, která byla uvedena v Chen a kol. (1999).

Tvrzení 1.1. Nechť $G(t)$ je distribuční funkce nějaké nezáporné náhodné veličiny. Potom funkce $\bar{F}(t)$ daná vztahem

$$\bar{F}(t) = \exp\left(-\theta G(t)\right), \ \theta > 0$$

je nevlastní funkce přežití.

Důkaz. Jelikož $G(t)$ je distribuční funkce, je funkce $\bar{F}(t)$ zřejmě zprava spojitá. Navíc platí:

$$\lim_{t\to 0} \bar{F}(t) = \lim_{t\to 0} \exp\left(-\theta G(t)\right) = \exp(0) = 1,$$
$$\lim_{t\to \infty} \bar{F}(t) = \lim_{t\to \infty} \exp\left(-\theta G(t)\right) = \exp(-\theta) \geq 0.$$

Stačí již pouze ověřit, že funkce $\bar{F}(t)$ je nerostoucí.

$$\frac{\partial \bar{F}}{\partial t} = -\theta g(t) \exp\left(-\theta G(t)\right),$$

kde $g(t) = \frac{\partial G(t)}{\partial t}$. Výraz $\frac{\partial \bar{F}(t)}{\partial t}$ je záporný právě tehdy, když $\theta > 0$. A tedy funkce $\bar{F}(t)$ je klesající.

❑

Chen a kol. (1999) předkládají pro reprezentaci uvedenou v tvrzení 1.1 následující biologickou motivaci. Předpokládejme nyní, že u pacientů pozorujeme počty aktivních karcinogenních buněk. Označme N_i počet aktivních karcinogenních buněk i-tého pacienta a předpokládejme, že N_i jsou nezávislé, stejně rozdělené náhodné veličiny s Poissonovým rozdělení se střední hodnotou θ. Nechť Z_{ji} je inkubační doba pro j-tou karcinogenní buňku i-tého pacienta. Tedy Z_{ji} je období mezi vznikem j-té karcinogenní buňky u i-tého pacienta a vypuknutím nemoci. Dále předpokládejme, že $Z_{ji},\ j = 1, \ldots, N_i$ jsou nezávislé, stejně rozdělené veličiny s distribuční funkcí $G(t)$, které jsou nezávislé na N_i. Čas do recidivy i-tého pacienta definujme jako náhodnou veličinu $T_i^r = \min\{Z_{ji},\ 0 \leq j \leq N_i\}$, kde $\mathsf{P}(Z_{0i} = \infty) = 1$. Náhodné veličiny $T_i^r,\ i = 1, \ldots, n$ jsou nezávislé. Předpokládáme, že k recidivě nemusí v konečném čase dojít, a tedy nevlastní funkce přežití náhodné veličiny T_i^r je $\bar{F}(t) = \mathsf{P}(\text{bez recidivy do času } t)$. Pokud do času t nedošlo k recidivě, pak pacient buď nemá žádné karcinogenní buňky, nebo inkubační doby karcinogenních buněk pacienta jsou větší než t, a tedy platí

$$\begin{aligned}
\bar{F}(t) &= \mathsf{P}(\text{bez recidivy do času } t) \\
&= \mathsf{P}(N_i = 0) + \mathsf{P}(Z_{i1} > t, \ldots, Z_{iN_i} > t, N_i \geq 1) \\
&= \exp(-\theta) + \sum_{k=1}^{\infty} \bar{G}(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\
&= \exp\left(-\theta + \theta \bar{G}(t)\right) = \exp\left(-\theta G(t)\right). \qquad (1.2)
\end{aligned}$$

Vztah (1.2) budeme nadále nazývat jako bio model, jelikož má biologickou motivaci. Počty karcinogenních buněk nejsou v praxi pozorovány a v modelu (1.2) mohou vystupovat jako latentní veličiny.

Mezi uvedenými reprezentacemi zřejmě existuje matematický vztah. Každou nevlastní funkci přežití ve tvaru (1.2) lze zapsat ve tvaru (1.1).

$$\begin{aligned}
\bar{F}(t) &= \exp\left(-\theta G(t)\right) \\
&= \exp(-\theta) - \exp(-\theta) + \exp\left(-\theta G(t)\right) \\
&= \exp(-\theta) + \frac{1 - \exp(-\theta)}{1 - \exp(-\theta)} \Big[\exp\left(-\theta G(t)\right) - \exp(-\theta)\Big] \\
&= \exp(-\theta) + \Big(1 - \exp(-\theta)\Big) \bar{F}_0(t).
\end{aligned}$$

Navíc pro funkci do relapsu nevyléčených pacientů platí:

$$\begin{aligned}\mathsf{P}(X \geq t | X < \infty) &= \frac{\mathsf{P}(t \leq X < \infty)}{\mathsf{P}(X < \infty)} = \frac{1 - \exp(-\theta) - 1 + \exp\left(-\theta G(t)\right)}{1 - \exp(-\theta)} \\ &= \frac{\exp\left(-\theta G(t)\right) - \exp(-\theta)}{1 - \exp(-\theta)}, \end{aligned} \tag{1.3}$$

a tedy se jedná o standardní model (1.1) s pravděpodobností vyléčení $p = \exp(-\theta)$ a funkcí přežití pro relaps pro nevyléčené pacienty $\bar{F}_0(t) = \frac{\exp(-\theta G(t)) - \exp(-\theta)}{1 - \exp(-\theta)}$. Naopak každá reprezentace (1.1) odpovídá (1.2) pro nějaký parametr θ a nějakou distribuční funkci $G(t)$ nezáporné náhodné veličiny.

$$\begin{aligned}\bar{F}(t) &= p + (1-p)\bar{F}_0(t) \\ &= \exp\left(\log\left(p + (1-p)\bar{F}_0(t)\right)\right) \\ &= \exp\left(\frac{\log p}{\log p}\ \log\left(p + (1-p)\bar{F}_0(t)\right)\right) \\ &= \exp\left(\log p\ \ \frac{\log\left(p + (1-p)\bar{F}_0(t)\right)}{\log p}\right) \\ &= \exp\left(-\theta G(t)\right). \end{aligned}$$

Jedná se o reprezentaci (1.2) s parametrem $\theta = -\log p$ a $G(t) = \frac{\log\left(p + (1-p)\bar{F}_0(t)\right)}{\log p}$. Stačí ověřit, že $G(t)$ je opravdu distribuční funkce nějaké nezáporné náhodné veličiny. Funkce $G(t)$ je zřejmě zprava spojitá a neklesající a navíc platí

$$\begin{aligned}\lim_{t \to \infty} G(t) &= \frac{\log p}{\log p} = 1, \\ \lim_{t \to 0} G(t) &= \frac{\log 1}{\log p} = 0. \end{aligned}$$

Riziková funkce příslušná nevlastní funkci přežití pro bio model (1.2) je dána vztahem

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)} = \frac{\theta g(t) \exp(-\theta G(t))}{\exp(-\theta G(t)} = \theta g(t),$$

kde $f(t) = \frac{\partial F(t)}{\partial t}$ a $g(t) = \frac{\partial G(t)}{\partial t}$. Hlavní rozdíl v uvedených modelech je, že bio model, na rozdíl od standardního modelu, má strukturu proporcionálního rizika, pokud uvažujeme, že parametr θ závisí na regresorech (Chen a kol., 1999).

Další reprezentace nevlastní funkce přežití byla uvedena v Yin a Ibrahim (2005). Pomocí Box-Coxovy transformace

$$x^{(a)} = \begin{cases} \frac{x^a - 1}{a} & \text{jestliže } a \neq 0, \\ \log x & \text{jestliže } a = 0 \end{cases}$$

je nevlastní funkce přežití $\bar{F}(t)$ transformována a dává tak vzniknout nové třídě parametrizací nevlastní funkce přežití.

Tvrzení 1.2. Nechť $G(t)$ je distribuční funkce nějaké nezáporné náhodné veličiny. Nechť $0 \leq a \leq 1$, $\theta > 0$ a $0 \leq a\theta \leq 1$. Potom funkce $\bar{F}(t)$ daná vztahem

$$\bar{F}(t)^{(a)} = -\theta G(t), \tag{1.4}$$

kde $\bar{F}(t)^{(a)}$ značí Box-Coxovu transformaci funkční hodnoty $\bar{F}(t)$, je nevlastní funkce přežití.

Důkaz. Pro $a = 0$ je $\bar{F}(t)^{(0)} = \log \bar{F}(t) = -\theta G(t)$, tedy $\bar{F}(t) = \exp(-\theta G(t))$, a tedy se jedná o tvrzení 1.1.

Pro $a > 0$ je $\bar{F}(t)^{(a)} = \frac{\bar{F}(t)^a - 1}{a}$, tedy $\bar{F}(t) = (1 - a\theta G(t))^{1/a}$. Tato funkce je zprava spojitá, jelikož $G(t)$ je distribuční funkce, a tedy také zprava spojitá. Dále platí:

$$\begin{aligned}
\lim_{t\to 0} \bar{F}(t) &= \lim_{t\to 0} (1 - a\theta G(t))^{1/a} = 1^{1/a} = 1, \\
\lim_{t\to \infty} \bar{F}(t) &= \lim_{t\to \infty} (1 - a\theta G(t))^{1/a} = (1 - a\theta)^{1/a} \geq 0.
\end{aligned}$$

Navíc

$$\frac{\partial \bar{F}(t)}{\partial t} = \frac{1}{a}(1 - a\theta G(t))^{1/a-1}(-a\theta g(t)) = -\theta g(t)(1 - a\theta G(t))^{1/a-1},$$

kde $g(t) = \frac{\partial G(t)}{\partial t}$. Výraz $\frac{\partial \bar{F}(t)}{\partial t}$ je nekladný právě tehdy, když $\theta > 0$ a $0 \leq a\theta \leq 1$. A tedy funkce $\bar{F}(t)$ je klesající.

❑

Jak bylo zmíněno v předchozím důkaze, reprezentace (1.2) je speciálním případem (1.4) pro $a = 0$. Zvolíme-li $a = 1$ v reprezentaci (1.4), pak pro nevlastní funkci přežití platí $\bar{F}(t)^{(1)} = \bar{F}(t) - 1 = -\theta G(t)$. Tedy

$$\begin{aligned}
\bar{F}(t) &= 1 - \theta G(t) = 1 - \theta + \theta - \theta G(t) \\
&= (1 - \theta) + \theta(1 - G(t)) = (1 - \theta) + \theta \bar{G}(t),
\end{aligned}$$

což je standardní model (1.1) s pravděpodobností vyléčení $p = 1 - \theta$ a s funkcí přežití pro relaps pro nevyléčené pacienty $\bar{F}_0(t) = \bar{G}(t)$.

Alternativní třída parametrizací nevlastních funkcí přežití byla uvedena v Tsodikov (2002). Autor navrhuje transformovat nevlastní funkci přežití pomocí nelineární transformační funkce, která je specifikována v následující definici.

Definice 1.3. Řekneme, že funkce $\gamma(x)$ je transformační funkce, jestliže je absolutně spojitá s nosičem na $[0, 1]$, neklesající a platí $\gamma(1) = 1$ a $\gamma(0) \geq 0$.

Tvrzení 1.3. Nechť $\bar{G}(t)$ je funkce přežití nějaké nezáporné náhodné veličiny. Nechť γ je transformační funkce. Potom funkce $\bar{F}(t)$ daná vztahem

$$\bar{F}(t) = \gamma\left(\bar{G}(t)\right) \tag{1.5}$$

je nevlastní funkce přežití.

Důkaz. Jelikož $\bar{G}(t)$ je funkce přežití, je funkce $\bar{F}(t)$ zřejmě zprava spojitá. Navíc platí:

$$\begin{aligned}
\lim_{t\to 0} \bar{F}(t) &= \lim_{t\to 0} \gamma\left(\bar{G}(t)\right) = \gamma(1) = 1, \\
\lim_{t\to \infty} \bar{F}(t) &= \lim_{t\to \infty} \gamma\left(\bar{G}(t)\right) = \gamma(0) \geq 0.
\end{aligned}$$

Stačí již pouze ověřit, že funkce $\bar{F}(t)$ je nerostoucí. Nechť $t_1 < t_2$, potom platí $\bar{G}(t_1) \geq \bar{G}(t_2)$, protože $\bar{G}(t)$ je funkce přežití, a tedy nerostoucí funkce. A protože γ je neklesající, platí $\gamma\left(\bar{G}(t_1)\right) \geq \gamma\left(\bar{G}(t_2)\right)$, a tedy funkce $\bar{F}(t)$ je nerostoucí.

❑

Třídu parametrizací (1.5) nazýváme NTM (z anglického Nonlinear Tranformation Models). Přechozí reprezentace jsou speciálními případy NTM třídy. Pro tranformační funkci $\gamma(x) = p + (1-p)x$ dostáváme $\bar{F}(t) = p + (1-p)\bar{G}(t)$, tedy standardní model (1.1). Pro transformační funkci $\gamma(x) = \exp(-\theta(1-x))$ je $\bar{F}(t) = \exp(-\theta G(t))$, tedy reprezentace (1.2). Pro volbu $\gamma(x) = [1 - a\theta(1-x)]^{1/a}$, $a > 0$ je nevlastní funkce přežití $\bar{F}(t) = (1 - a\theta G(t))^{1/a}$, tedy se jedná o reprezentaci (1.4).

Kapitola 2

Odhady metodou maximální věrohodnosti

V této kapitole se budeme věnovat odhadování parametrů metodou maximální věrohodnosti pro nezávislá stejně rozdělená data. V sekci 2.1 se zaměříme na standardní model (1.1) a v sekci 2.2 na bio model (1.2).

2.1 Standardní model

Předpokládejme, že náš datový soubor obsahuje informace o n pacientech a naše pozorování jsou nezávislá a stejně rozdělená. Jako T_i^u označíme čas relapsu a T_i^c čas censorování i-tého pacienta. Nechť $Y_i = \min(T_i^u, T_i^c)$ a $\delta_i = \mathsf{I}_{\left(T_i^u \leq T_i^c\right)}$ je indikátor relapsu i-tého pacienta. Tedy $\delta_i = 1$, pokud u i-tého pacienta došlo k recidivě onemocnění, nebo $\delta_i = 0$, pokud k recidivě nedošlo. Předpokládejme, že funkce přežití pro relaps pro nevyléčené pacienty $\bar{F}_0(t)$ je parametricky specifikována a závisí na vektoru parametrů $\boldsymbol{\alpha}$, tedy $\bar{F}_0(t) = \bar{F}_0(t|\boldsymbol{\alpha})$.

K odhadnutí pravděpodobnosti vyléčení p a funkci přežití pro relaps nevyléčených pacientů je možno použít metodu maximální věrohodnosti. Věrohodnostní funkci pro model (1.1) lze zapsat následovně:

$$L(p, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \left[(1-p) f_0(y_i|\boldsymbol{\alpha})\right]^{\delta_i} \left[p + (1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^{1-\delta_i}. \tag{2.1}$$

Dále budeme pracovat s logaritmem věrohodnostní funkce (2.1)

$$\begin{aligned} l(p, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \Big\{ & \delta_i \left[\log(1-p) + \log f_0(y_i|\boldsymbol{\alpha})\right] \\ & + (1-\delta_i) \log\Big(p + (1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\Big) \Big\}. \end{aligned} \tag{2.2}$$

Pro nalezení maximálně věrohodných odhadů derivujeme logaritmickou věrohodnost (2.2) podle jednotlivých parametrů. Složky skórové funkce jsou

$$\begin{aligned} \frac{\partial l(p, \boldsymbol{\alpha})}{\partial p} &= \sum_{i=1}^{n} \left[-\delta_i \frac{1}{1-p} + (1-\delta_i) \frac{1 - \bar{F}_0(y_i|\boldsymbol{\alpha})}{p + (1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} \right], \\ \frac{\partial l(p, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n} \left[\frac{\delta_i}{f_0(y_i|\boldsymbol{\alpha})} \frac{\partial f_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} + \frac{(1-\delta_i)(1-p)}{p + (1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} \frac{\partial \bar{F}_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right]. \end{aligned}$$

Maximálně věrohodný odhad pravděpodobnosti vyléčení p při známých parametrech $\boldsymbol{\alpha}$ je řešením věrohodnostní rovnice $\frac{\partial l(p,\boldsymbol{\alpha})}{\partial p} = 0$. Postupně dostáváme:

$$\begin{aligned}
\frac{\partial l(p,\boldsymbol{\alpha})}{\partial p} &= 0,\\
\sum_{i=1}^{n}(1-\delta_i)\frac{1-\bar{F}_0(y_i|\boldsymbol{\alpha})}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} &= \sum_{i=1}^{n}\frac{\delta_i}{1-p},\\
\sum_{i=1}^{n}(1-\delta_i)\frac{(1-p)\left(1-\bar{F}_0(y_i|\boldsymbol{\alpha})\right)}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} &= \sum_{i=1}^{n}\delta_i,\\
\sum_{i=1}^{n}(1-\delta_i)\frac{1-p-(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} &= \sum_{i=1}^{n}\delta_i,\\
\sum_{i=1}^{n}\frac{1-\delta_i}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} - \sum_{i=1}^{n}(1-\delta_i) &= \sum_{i=1}^{n}\delta_i,\\
\sum_{i=1}^{n}\frac{1-\delta_i}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} &= n. \qquad (2.3)
\end{aligned}$$

Maximálně věrohodný odhad parametrů $\boldsymbol{\alpha}$ při známé pravděpodobnosti vyléčení p je řešením rovnice $\frac{\partial l(p,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$, tedy rovnice

$$\sum_{i=1}^{n}\left[\frac{\delta_i}{f_0(y_i|\boldsymbol{\alpha})}\frac{\partial f_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} + \frac{(1-\delta_i)(1-p)}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})}\frac{\partial \bar{F}_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right] = 0. \qquad (2.4)$$

Ani jedna z rovnic (2.3) a (2.4) nemá obecně explicitní řešení. K nalezení odhadu lze použít numerické metody, například Newton-Raphsonovu metodu.

Uvažujme nyní, že čas do relapsu nevyléčených pacientů odpovídá exponenciálnímu rozdělení s neznámým parametrem λ, tedy předpokládáme, že pro funkci přežití nevyléčených pacientů platí $\bar{F}_0(t|\lambda) = \exp(-\lambda t)$. Na tomto jednoduchém příkladě ukážeme odvození maximálně věrohodného odhadu parametru λ. Věrohodnostní rovnice pro parametr λ má tvar:

$$\sum_{i=1}^{n}\left[\delta_i\frac{1-\lambda y_i}{\lambda} - (1-\delta_i)\frac{y_i(1-p)\exp(-\lambda y_i)}{p+(1-p)\exp(-\lambda y_i)}\right] = 0. \qquad (2.5)$$

Rovnice (2.5) ani v tomto jednoduchém případě nemá explicitní řešení. Maximálně věrohodný odhad pravděpodobnosti vyléčení při známém parametru λ je v tomto jednoduchém příkladě řešením rovnice

$$\sum_{i=1}^{n}\frac{1-\delta_i}{p+(1-p)\exp(-\lambda y_i)} = n.$$

Za předpokladu $\bar{F}_0(t|\lambda) = \exp(-\lambda t)$ má Fisherova informační matice tvar

$$I(p,\lambda) = -\,\mathsf{E}\begin{pmatrix}\frac{\partial^2 l_i(p,\lambda)}{\partial p^2} & \frac{\partial^2 l_i(p,\lambda)}{\partial p\partial\lambda}\\ \frac{\partial^2 l_i(p,\lambda)}{\partial p\partial\lambda} & \frac{\partial^2 l_i(p,\lambda)}{\partial\lambda^2}\end{pmatrix},$$

kde $l_i(p,\lambda)$ je příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (2.2). Pozorovaná informační matice má tvar

$$I_n(p,\lambda) = -\frac{1}{n}\begin{pmatrix}\frac{\partial^2 l(p,\lambda)}{\partial p^2} & \frac{\partial^2 l(p,\lambda)}{\partial p\partial\lambda}\\ \frac{\partial^2 l(p,\lambda)}{\partial p\partial\lambda} & \frac{\partial^2 l(p,\lambda)}{\partial\lambda^2}\end{pmatrix} \qquad (2.6)$$

s následujícími složkami

$$\begin{aligned}
\frac{\partial^2 l(p,\lambda)}{\partial p^2} &= \sum_{i=1}^{n}\left\{-\frac{\delta_i}{(1-p)^2} - \frac{(1-\delta_i)\left(1-\exp(-\lambda y_i)\right)^2}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\right\},\\
\frac{\partial^2 l(p,\lambda)}{\partial p \partial \lambda} &= \sum_{i=1}^{n}\Bigg\{(1-\delta_i)\frac{y_i \exp(-\lambda y_i)\left[p+(1-p)\exp(-\lambda y_i)\right]}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\\
&\qquad + (1-\delta_i)\frac{y_i\left(1-\exp(-\lambda y_i)\right)(1-p)\exp(-\lambda y_i)}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\Bigg\}\\
&= \sum_{i=1}^{n}\frac{(1-\delta_i) y_i \exp(-\lambda y_i)}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2},\\
\frac{\partial^2 l(p,\lambda)}{\partial \lambda^2} &= \sum_{i=1}^{n}\Bigg\{\delta_i\frac{-\lambda y_i - (1-\lambda y_i)}{\lambda^2}\\
&\qquad - (1-\delta_i)(1-p) y_i \frac{-y_i \exp(-\lambda y_i)\left[p+(1-p)\exp(-\lambda y_i)\right]}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\\
&\qquad + (1-\delta_i)(1-p) y_i \frac{\exp(-\lambda y_i)\left[-y_i(1-p)\exp(-\lambda y_i)\right]}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\Bigg\}\\
&= \sum_{i=1}^{n}\left\{-\frac{\delta_i}{\lambda^2} + \frac{(1-\delta_i)(1-p) p y_i^2 \exp(-\lambda y_i)}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\right\}.
\end{aligned}$$

Označíme-li pro matici (2.6) $D_1 = \det\left(I_n(p,\lambda)_{11}\right)$ a $D_2 = \det\left(I_n(p,\lambda)\right)$, pak

$$\begin{aligned}
D_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{(1-p)^2} + \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\delta_i)\left(1-\exp(-\lambda y_i)\right)^2}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2},\\
D_2 &= \frac{1}{n} D_1 \times \sum_{i=1}^{n}\left\{\frac{\delta_i}{\lambda^2} - \frac{(1-\delta_i)(1-p) p y_i^2 \exp(-\lambda y_i)}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\right\}\\
&\qquad - \frac{1}{n^2}\left\{\sum_{i=1}^{n}\frac{(1-\delta_i) y_i \exp(-\lambda y_i)}{\left[p+(1-p)\exp(-\lambda y_i)\right]^2}\right\}^2.
\end{aligned}$$

Výraz D_1 je součtem dvou nezáporných sum. První suma je nulová právě tehdy, když nedojde ani k jedné události, tj. $\sum_{i=1}^{n}\delta_i = 0$. Druhá suma je nulová právě tehdy, když dojde u každého pacienta k události, tj. pro každé i platí $\delta_i = 1$, nebo v případě, kdy čas události, respektive čas censorování, je nulový pro všechna pozorování. Pokud ani jeden z popsaných případů nenastal, pak je výraz D_1 kladný. Z tvaru výrazu D_2 usuzujeme, že může nabývat kladných i záporných hodnot, a tedy pozorovaná informační matice není obecně pozitivně definitní dle Sylvestrova kritéria.

Jak již bylo zmíněno, soustava věrohodnostních rovnic nemá analytické řešení a při hledání maxima logaritmické věrohodnostní funkce (2.2) je potřeba použít numerické metody. Řešení nemusí ovšem nutně existovat nebo může existovat více řešení, přičemž ne všechna jsou maximálně věrohodným odhadem. Jelikož pozorovaná informační matice není obecně pozitivně definitní, pak logaritmická věrohodnost (2.2) není konkávní funkce a řešení soustavy věrohodnostních rovnic není nutně globálním maximem.

Peng a Dear (2000) navrhují v modelu (1.1) zavést latentní veličinu. Předpokládejme nyní, že u pacientů pozorujeme, zda se vyléčili či nikoliv. Označme C_i indikátor vyléčení i-tého pacienta, tedy $C_i = 1$, pokud se pacient vyléčil a $C_i = 0$, pokud k vyléčení nedošlo. Došlo-li k relapsu pacienta (tj. $\delta_i = 1$), pak pacient nemohl být vyléčen a $C_i = 0$. Naopak, pokud k relapsu nedošlo, pacient mohl, ale nemusel být vyléčen, a tedy C_i může nabývat obou hodnot. Věrohodnostní rovnici (2.1) lze přepsat takto:

$$\begin{aligned} L^C(p, \boldsymbol{\alpha}) &= \prod_{i=1}^{n} \left\{ \left[(1-p) f_0(y_i|\boldsymbol{\alpha})\right]^{1-c_i} \right\}^{\delta_i} \left\{ p^{c_i} \left[(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^{1-c_i} \right\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[f_0(y_i|\boldsymbol{\alpha})^{\delta_i} \bar{F}_0(y_i|\boldsymbol{\alpha})^{1-\delta_i} \right]^{1-c_i} \times \prod_{i=1}^{n} p^{c_i}(1-p)^{1-c_i}. \end{aligned} \tag{2.7}$$

Věrohodnostní příspěvek nevyléčeného pacienta ($c_i = 0$), u kterého došlo k recidivě onemocnění ($\delta_i = 1$), je roven výrazu

$$f_0(y_i|\boldsymbol{\alpha})(1-p),$$

pro nevyléčeného pacienta, u kterého doposud nedošlo k recidivě ($\delta_i = 0$), pak

$$\bar{F}_0(y_i|\boldsymbol{\alpha})(1-p).$$

Příspěvek vyléčeného pacienta ($c_i = 1$) do věrohodnostní funkce je p, což odpovídá pravděpodobnosti vyléčení.

Zlogaritmováním věrohodnostní funkce (2.7) získáváme logaritmickou věrohodnost:

$$\begin{aligned} l^C(p, \boldsymbol{\alpha}) &= \sum_{i=1}^{n} (1-c_i) \left[\delta_i \log f_0(y_i|\boldsymbol{\alpha}) + (1-\delta_i) \log \bar{F}_0(y_i|\boldsymbol{\alpha}) \right] \\ &\quad + \sum_{i=1}^{n} \left[c_i \log p + (1-c_i) \log(1-p) \right]. \end{aligned} \tag{2.8}$$

Z předchozích úvah vyplývá, že součin $c_i\delta_i$ je vždy nulový a logaritmickou věrohodnostní funkci (2.8) lze zjednodušit:

$$\begin{aligned} l^C(p, \boldsymbol{\alpha}) &= \sum_{i=1}^{n} \left[\delta_i \log f_0(y_i|\boldsymbol{\alpha}) + (1-\delta_i-c_i) \log \bar{F}_0(y_i|\boldsymbol{\alpha}) \right] \\ &\quad + \sum_{i=1}^{n} \left[c_i \log p + (1-c_i) \log(1-p) \right] \\ &= \sum_{i=1}^{n} \left[\delta_i \log \lambda_0(y_i|\boldsymbol{\alpha}) + (1-c_i) \log \bar{F}_0(y_i|\boldsymbol{\alpha}) \right] \\ &\quad + \sum_{i=1}^{n} \left[c_i \log p + (1-c_i) \log(1-p) \right], \end{aligned} \tag{2.9}$$

kde $\lambda_0(t|\boldsymbol{\alpha}) = \frac{f_0(t|\boldsymbol{\alpha})}{\bar{F}_0(y_i|\boldsymbol{\alpha})}$ je riziková funkce pro relaps.

Motivací pro zavedení latentní veličiny C_i do věrohodnosti je použití EM algoritmu (z anglického Expectation–Maximization algorithm), který popíšeme na konci této sekce. Navíc logaritmickou věrohodnost (2.9) lze nyní zapsat jako

součet dvou sum, přičemž první nezávisí na parametru p a druhá na parametrech $\boldsymbol{\alpha}$, což zjednodušuje další odvození. Dále si podrobně odvodíme maximálně věrohodné odhady pro pravděpodobnost vyléčení p a pro parametry $\boldsymbol{\alpha}$ funkce přežití pro relaps, které v dostupné literatuře nejsou uvedeny. Derivací (2.9) podle jednotlivých parametrů získáváme složky skórové funkce:

$$\frac{\partial l^C(p,\boldsymbol{\alpha})}{\partial p} = \sum_{i=1}^{n}\left(\frac{c_i}{p} - \frac{1-c_i}{1-p}\right),$$
$$\frac{\partial l^C(p,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n}\left(\frac{\delta_i}{\lambda_0(y_i|\boldsymbol{\alpha})}\frac{\partial \lambda_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} + \frac{1-c_i}{\bar{F}_0(y_i|\boldsymbol{\alpha})}\frac{\partial \bar{F}_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right).$$

Maximálně věrohodný odhad pravděpodobnosti vyléčení $\hat{p}$ a odhad parametrů $\hat{\boldsymbol{\alpha}}$ funkce přežití pro relaps pro nevyléčené pacienty získáme řešením věrohodnostních rovnic $\frac{\partial l^C(p,\boldsymbol{\alpha})}{\partial p} = 0$ a $\frac{\partial l^C(p,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$.

$$\begin{aligned}
\frac{\partial l^C(p,\boldsymbol{\alpha})}{\partial p} &= 0,\\
\sum_{i=1}^{n}\left(\frac{c_i}{p} - \frac{1-c_i}{1-p}\right) &= 0,\\
(1-p)\sum_{i=1}^{n} c_i &= p\sum_{i=1}^{n}(1-c_i),\\
\sum_{i=1}^{n} c_i &= p\sum_{i=1}^{n}(1-c_i) + p\sum_{i=1}^{n} c_i,\\
\hat{p} &= \frac{\sum_{i=1}^{n} c_i}{n}.
\end{aligned}$$

Pokud bychom znali indikátory vyléčení pacientů, pak bychom odhad $\hat{p}$ mohli interpretovat jako poměr počtu vyléčených pacientů ku všem pacientům.

Věrohodnostní rovnice $\frac{\partial l^C(p,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$ nemá obecně explicitní řešení. Nyní, stejně jako v předchozí části, předpokládejme, že čas do relapsu nevyléčených pacientů se řídí exponenciálním rozdělením s parametrem λ. Riziková funkce pro relaps má potom tvar $\lambda_0(t|\lambda) = \frac{\lambda\exp(-t\lambda)}{\exp(-t\lambda)} = \lambda$. Složka skórové funkce příslušná parametru λ je

$$\begin{aligned}
\frac{\partial l^C(p,\lambda)}{\partial \lambda} &= \sum_{i=1}^{n}\left[\frac{\delta_i}{\lambda} - (1-c_i)\frac{y_i\exp(-y_i\lambda)}{\exp(-y_i\lambda)}\right]\\
&= \sum_{i=1}^{n}\left[\frac{\delta_i}{\lambda} - (1-c_i)y_i\right].
\end{aligned}$$

Řešením věrohodnostní rovnice získáváme odhad $\hat{\lambda}$.

$$\begin{aligned}
\frac{\partial l^C(p,\lambda)}{\partial \lambda} &= 0,\\
\sum_{i=1}^{n}\left[\frac{\delta_i}{\lambda} - (1-c_i)y_i\right] &= 0,\\
\sum_{i=1}^{n}\delta_i &= \lambda\sum_{i=1}^{n} y_i(1-c_i),
\end{aligned}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i(1-c_i)}.$$

Pokud bychom znali indikátory vyléčení pacientů, pak bychom odhad $\hat{\lambda}$ mohli interpretovat jako podíl počtu relapsů a součtu časů relapsů nevyléčených pacientů.

Za předpokladu $\bar{F}_0(t|\lambda) = \exp(-\lambda t)$ má Fisherova informační matice tvar

$$I(p,\lambda) = -\,\mathsf{E}\begin{pmatrix} \frac{\partial^2 l_i^C(p,\lambda)}{\partial p^2} & \frac{\partial^2 l_i^C(p,\lambda)}{\partial p \partial \lambda} \\ \frac{\partial^2 l_i^C(p,\lambda)}{\partial p \partial \lambda} & \frac{\partial^2 l_i^C(p,\lambda)}{\partial \lambda^2} \end{pmatrix},$$

kde $l_i^C(p,\lambda)$ je příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (2.8). Navíc platí

$$\begin{aligned}
-\,\mathsf{E}\left(\frac{\partial^2 l_i^C(p,\lambda)}{\partial p^2}\right) &= \mathsf{E}\left(\frac{c_i}{p^2} + \frac{1-c_i}{(1-p)^2}\right) = \mathsf{E}\left(\frac{c_i}{p^2}\right) + \mathsf{E}\left(\frac{1-c_i}{(1-p)^2}\right) \\
&= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}, \\
-\,\mathsf{E}\left(\frac{\partial^2 l_i^C(p,\lambda)}{\partial p \partial \lambda}\right) &= \mathsf{E}\,0 = 0, \\
-\,\mathsf{E}\left(\frac{\partial^2 l_i^C(p,\lambda)}{\partial \lambda^2}\right) &= \mathsf{E}\left[\frac{\delta_i}{\lambda^2}\right] = \frac{\mathsf{P}\left(T^u \le T^c\right)}{\lambda^2},
\end{aligned}$$

a tedy Fisherova informační matice má tvar

$$I(p,\lambda) = \begin{pmatrix} \frac{1}{p(1-p)} & 0 \\ 0 & \frac{\mathsf{P}(T^u \le T^c)}{\lambda^2} \end{pmatrix}.$$

Pozorovaná informační matice má tvar

$$I_n(p,\lambda) = -\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l^C(p,\lambda)}{\partial p^2} & \frac{\partial^2 l^C(p,\lambda)}{\partial p \partial \lambda} \\ \frac{\partial^2 l^C(p,\lambda)}{\partial p \partial \lambda} & \frac{\partial^2 l^C(p,\lambda)}{\partial \lambda^2} \end{pmatrix} \tag{2.10}$$

s následujícími složkami

$$\begin{aligned}
\frac{\partial^2 l^C(p,\lambda)}{\partial p^2} &= \sum_{i=1}^{n}\left[-\frac{c_i}{p^2} - \frac{1-c_i}{(1-p)^2}\right], \\
\frac{\partial^2 l^C(p,\lambda)}{\partial p \partial \lambda} &= 0, \\
\frac{\partial^2 l^C(p,\lambda)}{\partial \lambda^2} &= \sum_{i=1}^{n} \delta_i \frac{-y_i\lambda - (1-\lambda y_i)}{\lambda^2} \\
&= \sum_{i=1}^{n}\left(-\frac{\delta_i}{\lambda^2}\right).
\end{aligned}$$

Označíme-li pro matici (2.10) $D_1 = \det\left(I_n(p,\lambda)_{11}\right)$ a $D_2 = \det\left(I_n(p,\lambda)\right)$, pak

$$\begin{aligned}
D_1 &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{c_i}{p^2} + \frac{1-c_i}{(1-p)^2}\right], \\
D_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{c_i}{p^2} + \frac{1-c_i}{(1-p)^2}\right] \times \sum_{i=1}^{n}\frac{\delta_i}{\lambda^2}.
\end{aligned}$$

Výraz D_1 je součtem dvou nezáporných sum, přičemž první je rovna nule právě tehdy, když se žádný z pacientů neuzdraví, to jest $\sum_{i=1}^{n} c_i = 0$. Druhá suma je rovna nule pouze v případě, že se uzdraví všichni pacienti, tedy $\sum_{i=1}^{n}(1 - c_i) = 0$. Tyto dva případy nemohou nastat zároveň, a tedy výraz D_1 je vždy kladný. Výraz D_2 je zřejmě nezáporný. Rovnost $D_2 = 0$ může nastat pouze, pokud u žádného z pacientů nedošlo k recidivě onemocnění, tedy $\sum_{i=1}^{n} \delta_i = 0$. Pozorovaná informační matice (2.10) je ve všech ostatních případech dle Sylvestrova kritéria pozitivně definitní, a tedy $\hat{p}$ a $\hat{\lambda}$ jsou maximálně věrohodné odhady parametrů p a λ.

Zavedením latentních veličin C_i se odvození odhadů značně zjednodušují. V praxi ovšem hodnoty, kterých veličina nabývá, neznáme. Na vektor indikátorů vyléčení $\mathbf{c} = (c_1, \ldots, c_n)$ můžeme nahlížet jako na chybějící data a pro odhadnutí parametrů p a $\boldsymbol{\alpha}$ (respektive λ) lze použít EM algoritmus. Dále budeme postupovat stejně jako Peng a Dear (2000) a navíc předvedeme použití EM algoritmu pro speciální případ, kdy se čas do relapsu řídí exponenciálním rozdělením.

Označme $\mathbf{D} = (\boldsymbol{\delta}, \mathbf{y})$ vektor dat, kde $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ je vektor indikátorů relapsu a $\mathbf{y} = (y_1, \ldots, y_n)$ je vektor pozorovaných časů. Dále nechť $p^{(k)}$ a $\boldsymbol{\alpha}^{(k)}$ jsou odhady parametrů p a $\boldsymbol{\alpha}$ v k-tém kroku. Označme $\boldsymbol{\zeta}^{(k)} = (p^{(k)}, \boldsymbol{\alpha}^{(k)})$, přičemž počáteční odhady $p^{(0)}$ a $\boldsymbol{\alpha}^{(0)}$ lze volit na základě dat. EM algoritmus je iterační metoda, která se sestává ze dvou kroků. V E-kroku je vypočtena podmíněná střední hodnota logaritmické věrohodnosti (2.8) z kompletního modelu při daných hodnotách $\boldsymbol{\zeta}^{(k)}$ a při daných datech $\mathbf{D}$.

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}\left(l^C(p, \boldsymbol{\alpha})\right) = &\sum_{i=1}^{n}\left[\delta_i \log \lambda_0(y_i|\boldsymbol{\alpha}) + \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(1 - c_i)\log \bar{F}_0(y_i|\boldsymbol{\alpha})\right] \\ &+ \sum_{i=1}^{n}\left[\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(c_i)\log p + \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(1 - c_i)\log(1 - p)\right]. \qquad (2.11)\end{aligned}$$

Označíme-li $c_i^{(k)} = \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(c_i)$, pak platí

$$c_i^{(k)} = \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(c_i) = \frac{(1 - \delta_i)p^{(k)}}{p^{(k)} + (1 - p^{(k)})\bar{F}(y_i|\boldsymbol{\alpha}^{(k)})}. \qquad (2.12)$$

Pro speciální případ, kdy se čas do relapsu řídí exponenciálním rozdělením, má podmíněná střední hodnota (2.12) následující tvar

$$c_i^{(k)} = \frac{(1 - \delta_i)p^{(k)}}{p^{(k)} + (1 - p^{(k)})\exp(-\lambda^{(k)} y_i)}.$$

V M-kroku vypočtené hodnoty (2.12) fixujeme a metodou maximální věrohodnosti aplikovanou na (2.11) odhadujeme parametry $p^{(k+1)}$ a $\boldsymbol{\alpha}^{(k+1)}$ tak, jak bylo popsáno výše. Uvažujeme-li, že čas do relapsu se řídí exponenciálním rozdělením, potom odhady parametrů p a λ v $k + 1$ kroku jsou

$$\begin{aligned}p^{(k+1)} &= \frac{\sum_{i=1}^{n} c_i^{(k)}}{n}, \\ \lambda^{(k+1)} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i\left(1 - c_i^{(k)}\right)}.\end{aligned}$$

2.2 Bio model

V této sekci se budeme zabývat odhadem parametrů v modelu (1.2). Předpokládejme, že distribuční funkce $G(t)$ je parametricky specifikována skrze vektor parametrů $\boldsymbol{\alpha}$, tedy $G(t) = G(t|\boldsymbol{\alpha})$. Nejprve uvedeme věrohodnost bez latentních proměnných a odhady založené na ní. Tento přístup v dostupné literatuře zatím nebyl popsán. Obdobně jako v předchozí sekci, i zde zavedeme latentní proměnné a ukážeme možnou implementaci EM algoritmu.

Věrohodnostní funkci pro model (1.2) lze dle Kalbfleisch a Prentice (2011) zapsat následovně:

$$\begin{aligned} L(\theta, \boldsymbol{\alpha}) &= \prod_{i=1}^{n} \left(\theta g(y_i|\boldsymbol{\alpha}) \exp\left(-\theta G(y_i|\boldsymbol{\alpha})\right)\right)^{\delta_i} \exp\left(-\theta G(y_i|\boldsymbol{\alpha})\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(\theta g(y_i|\boldsymbol{\alpha})\right)^{\delta_i} \exp\left(-\theta G(y_i|\boldsymbol{\alpha})\right). \end{aligned} \tag{2.13}$$

Zlogaritmováním funkce (2.13) dostáváme logaritmickou věrohodnost:

$$l(\theta, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \left[\delta_i \left(\log\theta + \log g(y_i|\boldsymbol{\alpha})\right) - \theta G(y_i|\boldsymbol{\alpha})\right]. \tag{2.14}$$

Maximálně věrohodný odhad parametru θ a odhad parametrů distribuční funkce $G(t|\boldsymbol{\alpha})$ získáme pomocí skórové funkce a řešením věrohodnostních rovnic $\frac{\partial l(\theta,\boldsymbol{\alpha})}{\partial\theta} = 0$ a $\frac{\partial l(\theta,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} = 0$. Složky skórové funkce jsou

$$\begin{aligned} \frac{\partial l(\theta, \boldsymbol{\alpha})}{\partial\theta} &= \sum_{i=1}^{n} \left(\frac{\delta_i}{\theta} - G(y_i|\boldsymbol{\alpha})\right), \\ \frac{\partial l(\theta, \boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} &= \sum_{i=1}^{n} \left(\frac{\delta_i}{g(y_i|\boldsymbol{\alpha})} \frac{\partial g(y_i|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} - \theta \frac{\partial G(y_i|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right). \end{aligned}$$

Řešíme věrohodnostní rovnice:

$$\begin{aligned} \frac{\partial l(\theta, \boldsymbol{\alpha})}{\partial\theta} &= 0, \\ \sum_{i=1}^{n} \left(\frac{\delta_i}{\theta} - G(y_i|\boldsymbol{\alpha})\right) &= 0, \\ \frac{1}{\theta}\sum_{i=1}^{n} \delta_i &= \sum_{i=1}^{n} G(y_i|\boldsymbol{\alpha}), \\ \hat{\theta} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} G(y_i|\boldsymbol{\alpha})}. \end{aligned}$$

Řešením věrohodnostní rovnice při daných parametrech $\boldsymbol{\alpha}$ je odhad $\hat{\theta}$.

Věrohodnostní rovnice $\frac{\partial l(\theta,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} = 0$ nemá obecně explicitní řešení. Opět si na jednoduchém příkladě ukážeme odvození odhadů parametrů distribuční funkce $G(t|\boldsymbol{\alpha})$. Předpokládejme nyní, že G je distribuční funkce náhodné veličiny z exponenciálního rozdělení s parametrem λ, tj. $G(t|\lambda) = 1 - \exp(-t\lambda)$. Postupujeme obdobně jako v případě odhadu parametru θ. Skóre příslušné parametru λ je

$$\begin{aligned} \frac{\partial l(\theta, \lambda)}{\partial\lambda} &= \sum_{i=1}^{n} \left(\delta_i \frac{\exp(-\lambda y_i) - \lambda y_i \exp(-\lambda y_i)}{\lambda \exp(-\lambda y_i)} + \theta y_i \exp(-\lambda y_i)\right) \\ &= \sum_{i=1}^{n} \left(\delta_i \frac{1 - \lambda y_i}{\lambda} + \theta y_i \exp(-\lambda y_i)\right). \end{aligned}$$

Příslušná věrohodnostní rovnice je tedy

$$\sum_{i=1}^{n}\left(\delta_i\frac{1-\lambda y_i}{\lambda}+\theta y_i\exp(-\lambda y_i)\right)=0$$

a ani v tomto jednoduchém případě nemá explicitní řešení a je třeba jej hledat numericky.

Za předpokladu $G(t|\lambda)=1-\exp(-t\lambda)$ má Fisherova informační matice tvar

$$I(p,\lambda)=-\,\mathsf{E}\begin{pmatrix}\frac{\partial^2 l_i(\theta,\lambda)}{\partial\theta^2} & \frac{\partial^2 l_i(\theta,\lambda)}{\partial\theta\partial\lambda}\\ \frac{\partial^2 l_i(\theta,\lambda)}{\partial\theta\partial\lambda} & \frac{\partial^2 l_i(\theta,\lambda)}{\partial\lambda^2}\end{pmatrix},$$

kde $l_i(\theta,\lambda)$ je příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (2.14). Pozorovaná informační matice je

$$I_n(\theta,\lambda)=-\frac{1}{n}\begin{pmatrix}\frac{\partial^2 l(\theta,\lambda)}{\partial\theta^2} & \frac{\partial^2 l(\theta,\lambda)}{\partial\theta\partial\lambda}\\ \frac{\partial^2 l(\theta,\lambda)}{\partial\theta\partial\lambda} & \frac{\partial^2 l(\theta,\lambda)}{\partial\lambda^2}\end{pmatrix} \tag{2.15}$$

se složkami

$$\begin{aligned}\frac{\partial^2 l(\theta,\lambda)}{\partial\theta^2}&=\sum_{i=1}^{n}\left(-\frac{\delta_i}{\theta^2}\right),\\ \frac{\partial^2 l(\theta,\lambda)}{\partial\theta\partial\lambda}&=\sum_{i=1}^{n}\left(-y_i\exp(-\lambda y_i)\right),\\ \frac{\partial^2 l(\theta,\lambda)}{\partial\lambda^2}&=\sum_{i=1}^{n}\left[\delta_i\frac{\lambda y_i-(1-\lambda y_i)}{\lambda^2}-\theta y_i\exp(-\lambda y_i)(-y_i)\right]\\ &=\sum_{i=1}^{n}\left(-\frac{\delta_i}{\lambda^2}-\theta y_i^2\exp(-\lambda y_i)\right).\end{aligned}$$

Označme pro matici (2.15) $D_1=\det\left(I_n(\theta,\lambda)_{11}\right)$ a $D_2=\det\left(I_n(\theta,\lambda)\right)$, potom platí

$$D_1=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\theta^2}\geq 0,$$

přičemž rovnost $D_1=0$ nastává právě tehdy, když $\sum_{i=1}^{n}\delta_i=0$, tedy u žádného pacienta nedošlo k recidivě onemocnění. Situace při určování nezápornosti, respektive kladnosti, výrazu D_2 je složitější. Platí:

$$D_2=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\delta_i}{\theta^2}\times\sum_{i=1}^{n}\left(\frac{\delta_i}{\lambda^2}+\theta y_i^2\exp(-\lambda y_i)\right)-\frac{1}{n^2}\left(\sum_{i=1}^{n}y_i\exp(-\lambda y_i)\right)^2.$$

Usuzujeme, že výraz D_2 může nabývat kladných i záporných hodnot, a tedy matice (2.15) není obecně pozitivně definitní. Tedy logaritmická věrohodnost (2.14) není ryze konkávní funkce a řešení soustavy věrohodnostních rovnic není nutně globálním maximem.

Jak bylo zmíněno v kapitole 1, model (1.2) má biologickou motivaci, avšak

v praxi počty karcinogenních buněk nejsou pozorovány. Obdobně jako v sekci 2.1, i zde lze na počty karcinogenních buněk nahlížet jako na latentní proměnné.

Věrohodnostní funkci pro model (1.2), pokud bychom pozorovali počty karcinogenních buněk, má dle Chen a kol. (1999) následující tvar:

$$L^N(\theta, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \bar{G}(y_i|\boldsymbol{\alpha})^{N_i-\delta_i} \left(N_i g(y_i|\boldsymbol{\alpha})\right)^{\delta_i} \times \prod_{i=1}^{n} \frac{\theta^{N_i}}{N_i!} \exp(-\theta). \tag{2.16}$$

Věrohodnostní příspěvek pacienta s aktivními karcinogenními buňkami ($N_i > 0$), u kterého došlo k recidivě onemocnění ($\delta_i = 1$), je roven výrazu

$$\bar{G}(y_i|\boldsymbol{\alpha})^{N_i-1} N_i g(y_i|\boldsymbol{\alpha}) \times \frac{\theta^{N_i}}{N_i!} \exp(-\theta),$$

pro pacienta s aktivními karcinogenními buňkami, u kterého k relapsu nedošlo ($\delta_i = 0$) pak

$$\bar{G}(y_i|\boldsymbol{\alpha})^{N_i} \times \frac{\theta^{N_i}}{N_i!} \exp(-\theta).$$

Definujeme-li $0^0 = 1$, pak věrohodnostní příspěvek pacienta bez aktivních karcinogenních buněk ($\delta_i = 0$ a $N_i = 0$) je

$$\exp(-\theta),$$

což odpovídá pravděpodobnosti vyléčení.

Zlogaritmováním funkce (2.16) dostáváme logaritmickou věrohodnost:

$$\begin{aligned} l^N(\theta, \boldsymbol{\alpha}) =& \sum_{i=1}^{n} \left[(N_i - \delta_i) \log \bar{G}(y_i|\boldsymbol{\alpha}) + \delta_i \left(\log N_i + \log g(y_i|\boldsymbol{\alpha})\right)\right] \\ &+ \sum_{i=1}^{n} \left(N_i \log \theta - \log(N_i!) - \theta\right) \\ =& \sum_{i=1}^{n} \left[N_i \log \bar{G}(y_i|\boldsymbol{\alpha}) + \delta_i \left(\log N_i + \log \lambda(y_i|\boldsymbol{\alpha})\right)\right] \\ &+ \sum_{i=1}^{n} \left(N_i \log \theta - \log(N_i!) - \theta\right), \end{aligned} \tag{2.17}$$

kde $\lambda(y_i|\boldsymbol{\alpha}) = \frac{g(y_i|\boldsymbol{\alpha})}{\bar{G}(y_i|\boldsymbol{\alpha})}$ je riziková funkce.

Stejně jako v předchozí sekci i zde si odvodíme maximálně věrohodné odhady parametrů θ a $\boldsymbol{\alpha}$. Tato odvození a konkrétní tvary odhadů nejsou v literatuře podrobně popsány. Na konci této sekce uvedeme také implementaci EM algoritmu.

Logaritmickou věrohodnost (2.17) lze zapsat jako součet dvou sum, přičemž první nezávisí na parametru θ a druhá suma nezávisí na parametrech $\boldsymbol{\alpha}$. Následující odvození se tedy značně zjednodušuje. Složky skórové funkce mají tvar

$$\begin{aligned} \frac{\partial l^N(\theta, \boldsymbol{\alpha})}{\partial \theta} &= \sum_{i=1}^{n} \left(\frac{N_i}{\theta} - 1\right), \\ \frac{\partial l^N(\theta, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n} \left(\frac{N_i}{\bar{G}(y_i|\boldsymbol{\alpha})} \frac{\partial \bar{G}(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} + \frac{\delta_i}{\lambda(y_i|\boldsymbol{\alpha})} \frac{\partial \lambda(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right). \end{aligned}$$

Věrohodnostní rovnice pro parametr θ je

$$\begin{aligned}\frac{\partial l^N(\theta,\boldsymbol{\alpha})}{\partial\theta} &= 0,\\ \sum_{i=1}^{n}\left(\frac{N_i}{\theta}-1\right) &= 0,\\ \frac{1}{\theta}\sum_{i=1}^{n}N_i &= n,\\ \hat{\theta} &= \frac{\sum_{i=1}^{n}N_i}{n}.\end{aligned}$$

Pokud bychom znali počty karcinogenních buněk pacientů, pak bychom odhad $\hat{\theta}$ spočítali jako průměrný počet karcinogenních buněk. Věrohodnostní rovnice pro parametry $\boldsymbol{\alpha}$ opět nemá obecně explicitní řešení.

Předpokládejme nyní, že inkubační doba karcinogenních buněk se řídí exponenciálním rozdělením s parametrem λ, tedy pro distribuční funkci $G(t)$ platí $G(t|\lambda) = 1-\exp(-t\lambda)$. Skórová funkce příslušná parametru λ je

$$\begin{aligned}\frac{\partial l^N(\theta,\lambda)}{\partial\lambda} &= \sum_{i=1}^{n}\left[\frac{N_i}{\exp(-\lambda y_i)}\left(-y_i\exp(-\lambda y_i)\right)+\frac{\delta_i}{\lambda}\right]\\ &= \sum_{i=1}^{n}\left(-y_iN_i+\frac{\delta_i}{\lambda}\right).\end{aligned}$$

Řešíme následující věrohodnostní rovnici:

$$\begin{aligned}\frac{\partial l^N(\theta,\lambda)}{\partial\lambda} &= 0,\\ \sum_{i=1}^{n}\left(-y_iN_i+\frac{\delta_i}{\lambda}\right) &= 0,\\ \lambda\sum_{i=1}^{n}y_iN_i &= \sum_{i=1}^{n}\delta_i,\\ \hat{\lambda} &= \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n}y_iN_i}.\end{aligned}$$

Za předpokladu $G(t|\lambda) = 1-\exp(-t\lambda)$ má Fisherova informační matice tvar

$$I(p,\lambda) = -\,\mathsf{E}\begin{pmatrix}\frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\theta^2} & \frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\theta\partial\lambda}\\ \frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\theta\partial\lambda} & \frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\lambda^2}\end{pmatrix},$$

kde $l_i^N(\theta,\lambda)$ je příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (2.14). Navíc platí

$$\begin{aligned}-\,\mathsf{E}\left(\frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\theta^2}\right) &= -\,\mathsf{E}\left(-\frac{N_i}{\theta^2}\right) = \frac{\theta}{\theta^2} = \frac{1}{\theta},\\ -\,\mathsf{E}\left(\frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\theta\partial\lambda}\right) &= \mathsf{E}\,0 = 0,\\ -\,\mathsf{E}\left(\frac{\partial^2 l_i^N(\theta,\lambda)}{\partial\lambda^2}\right) &= -\,\mathsf{E}\left(\delta_i\frac{-\lambda y_i-(1-\lambda y_i)}{\lambda^2}\right)\\ &= -\,\mathsf{E}\left(-\frac{\delta_i}{\lambda^2}\right) = \frac{\mathsf{P}\left(T^u\le T^c\right)}{\lambda^2}.\end{aligned}$$

Fisherova informační matice má tedy tvar

$$I(p,\lambda) = \begin{pmatrix} \frac{1}{\theta} & 0 \\ 0 & \frac{\mathsf{P}(T^u \leq T^c)}{\lambda^2} \end{pmatrix}.$$

Pozorovaná informační matice má tvar

$$I_n(\theta,\lambda) = -\frac{1}{n} \begin{pmatrix} \frac{\partial^2 l^N(\theta,\lambda)}{\partial\theta^2} & \frac{\partial^2 l^N(\theta,\lambda)}{\partial\theta\partial\lambda} \\ \frac{\partial^2 l^N(\theta,\lambda)}{\partial\theta\partial\lambda} & \frac{\partial^2 l^N(\theta,\lambda)}{\partial\lambda^2} \end{pmatrix} \tag{2.18}$$

se složkami

$$\begin{aligned} \frac{\partial^2 l^N(\theta,\lambda)}{\partial\theta^2} &= \sum_{i=1}^{n} \Big(- \frac{N_i}{\theta^2} \Big), \\ \frac{\partial^2 l^N(\theta,\lambda)}{\partial\theta\partial\lambda} &= 0, \\ \frac{\partial^2 l^N(\theta,\lambda)}{\partial\lambda^2} &= \sum_{i=1}^{n} \Big(- \frac{\delta_i}{\lambda^2} \Big). \end{aligned}$$

Analogicky jako v sekci 2.1 označme pro pozorovanou informační matici (2.18) $D_1 = \det\left(I_n(\theta,\lambda)_{11}\right)$ a $D_2 = \det\left(I_n(\theta,\lambda)\right)$, potom platí

$$\begin{aligned} D_1 &= \frac{1}{n} \sum_{i=1}^{n} \frac{N_i}{\theta^2} \geq 0, \\ D_2 &= \frac{1}{n^2} \sum_{i=1}^{n} \frac{N_i}{\theta^2} \times \sum_{i=1}^{n} \frac{\delta_i}{\lambda^2} \geq 0, \end{aligned}$$

přičemž rovnost $D_1 = 0$ nastává právě tehdy, když $\sum_{i=1}^{n} N_i = 0$, tedy žádný pacient nemá karcinogenní buňky. Rovnost $D_2 = 0$ nastává buď v předchozím případě, nebo pokud u žádného pacienta nedošlo k relapsu, to jest $\sum_{i=1}^{n} \delta_i = 0$. Ve všech dalších případech je pozorovaná informační matice (2.18) dle Sylvestrova kritéria pozitivně definitní, a tedy $\hat{\theta}$ a $\hat{\lambda}$ jsou maximálně věrohodné odhady parametrů θ a λ.

Obdobně jako v sekci 2.1 lze vektor $\mathbf{N} = (N_1, \ldots, N_n)$ chápat jako chybějící data a k odhadu parametrů θ a $\boldsymbol{\alpha}$ pak použít EM-algoritmus. Ve zbývající části budeme postupovat podle Chen a Ibrahim (2001), kde autoři navrhují použít po částech exponenciální distribuční funkci $G(t)$. Ukážeme implementaci EM algoritmu pro parametricky specifikovanou distribuční funkci $G(t|\boldsymbol{\alpha})$ a také uvedeme tvary odhadů pro případ, kdy uvažujeme, že čas do relapsu se řídí exponenciálním rozdělením, což je speciální případ po částech exponenciální distribuční funkce.

Označme $\mathbf{D} = (\boldsymbol{\delta}, \mathbf{y})$ vektor pozorovaných dat, kde $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ je vektor indikátorů relapsu a $\mathbf{y} = (y_1, \ldots, y_n)$ je vektor pozorovaných časů. Nechť $\theta^{(k)}$ a $\boldsymbol{\alpha}^{(k)}$ jsou odhady parametrů θ a $\boldsymbol{\alpha}$ v k-tém kroku. Označme $\boldsymbol{\zeta}^{(k)} = (\theta^{(k)}, \boldsymbol{\alpha}^{(k)})$. V E-kroku spočítáme podmíněnou střední hodnotu logaritmické věrohodnosti (2.17) při daných hodnotách $\boldsymbol{\zeta}^{(k)}$ a při daných datech $\mathbf{D}$, tedy

$$\begin{aligned} \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}\left(l^N(\theta,\boldsymbol{\alpha})\right) = &\sum_{i=1}^{n} \Big[\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(N_i) \log \bar{G}(y_i|\boldsymbol{\alpha}) + \delta_i \Big(\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(\log N_i) + \log \lambda(y_i|\boldsymbol{\alpha}) \Big) \Big] \\ &+ \sum_{i=1}^{n} \Big(\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(N_i) \log\theta - \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(\log(N_i!)) - \theta \Big). \end{aligned} \tag{2.19}$$

Členy $\log N_i$ a $\log(N_i!)$ lze považovat za konstanty vzhledem k odhadování parametrů θ a $\boldsymbol{\alpha}$, a tudíž tvary podmíněných středních hodnot $\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(\log N_i)$ a $\mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(\log(N_i!))$ není nutné odvozovat. Označíme-li $p(y_i, \delta_i, N_i|\theta, \boldsymbol{\alpha})$ podmíněnou hustotu úplných dat při daných hodnotách parametrů θ a $\boldsymbol{\alpha}$, pak platí

$$\begin{aligned}
p(y_i, \delta_i, N_i|\theta, \boldsymbol{\alpha}) &= p(N_i|y_i, \delta_i, \theta, \boldsymbol{\alpha}) p(y_i, \delta_i|\theta, \boldsymbol{\alpha}), \\
p(N_i|y_i, \delta_i, \theta, \boldsymbol{\alpha}) &= \frac{p(y_i, \delta_i, N_i|\theta, \boldsymbol{\alpha})}{p(y_i, \delta_i|\theta, \boldsymbol{\alpha})}, \\
p(N_i|y_i, \delta_i, \theta, \boldsymbol{\alpha}) &= \frac{\bar{G}(y_i|\boldsymbol{\alpha})^{N_i - \delta_i} \left(N_i g(y_i|\boldsymbol{\alpha})\right)^{\delta_i} \frac{\theta^{N_i}}{N_i!} \exp(-\theta)}{\left(\theta g(y_i|\boldsymbol{\alpha})\right)^{\delta_i} \exp\left(-\theta G(y_i|\boldsymbol{\alpha})\right)}, \\
p(N_i|y_i, \delta_i, \theta, \boldsymbol{\alpha}) &= \frac{\exp\left(-\theta \bar{G}(y_i|\boldsymbol{\alpha})\right) \left(\theta \bar{G}(y_i|\boldsymbol{\alpha})\right)^{N_i}}{N_i!} \left(\frac{N_i}{\theta \bar{G}(y_i|\boldsymbol{\alpha})}\right)^{\delta_i}.
\end{aligned}$$

Pro $\delta_i = 0$ platí $p(N_i|y_i, \delta_i, \theta, \boldsymbol{\alpha}) = \frac{\exp\left(-\theta \bar{G}(y_i|\boldsymbol{\alpha})\right)\left(\theta \bar{G}(y_i|\boldsymbol{\alpha})\right)^{N_i}}{N_i!}$ a pro $\delta_i = 1$ pak $p(N_i|y_i, \delta_i, \theta, \boldsymbol{\alpha}) = \frac{\exp\left(-\theta \bar{G}(y_i|\boldsymbol{\alpha})\right)\left(\theta \bar{G}(y_i|\boldsymbol{\alpha})\right)^{N_i - 1}}{(N_i - 1)!}$, tedy $N_i - \delta_i$ má Poissonovo rozdělení se střední hodnotou $\theta \bar{G}(y_i|\boldsymbol{\alpha})$.

Označíme-li $N_i^{(k)} = \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(N_i)$ podmíněnou střední hodnotu N_i při daných hodnotách $\boldsymbol{\zeta}^{(k)}$ a daných datech $\mathbf{D}$, pak platí

$$N_i^{(k)} = \mathsf{E}_{\boldsymbol{\zeta}^{(k)}}(N_i) = \theta^{(k)} \bar{G}(y_i|\boldsymbol{\alpha}^{(k)}) + \delta_i.$$

Pro speciální případ, kdy se čas do relapsu řídí exponenciálním rozdělením s parametrem λ, potom platí

$$N_i^{(k)} = \theta^{(k)} \exp(-\lambda^{(k)} y_i) + \delta_i.$$

V M-kroku fixujeme hodnoty $N_i^{(k)}$ a metodou maximální věrohodnosti aplikovanou na (2.19) odhadujeme parametry $\theta^{(k+1)}$ a $\boldsymbol{\alpha}^{(k+1)}$ tak, jak již bylo popsáno výše. Pro speciální případ, kdy se čas do relapsu řídí exponenciálním rozdělením, mají odhady parametrů θ a λ v $k+1$ kroku následující tvar

$$\begin{aligned}
\theta^{(k+1)} &= \frac{\sum_{i=1}^n N_i^{(k)}}{n}, \\
\lambda^{(k+1)} &= \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n y_i N_i^{(k)}}.
\end{aligned}$$

Kapitola 3

Vliv regresorů

Některé veličiny, jako je věk či pohlaví, mohou mít vliv na vyléčení, nebo na dobu do relapsu, nebo na obě tyto veličiny. Avšak pravděpodobnost vyléčení a funkce přežití pro relaps nutně nezávisí na stejných vysvětlujících proměnných a jednotlivé regresory mohou mít na ně různý efekt.

V této kapitole budeme uvažovat, že vysvětlující veličiny mají vliv buď na pravděpodobnost vyléčení, nebo na funkci přežití pro relaps ve standardním modelu (1.1) a v bio modelu (1.2). Označme $\mathbf{X}_i$ a $\mathbf{Z}_i$ vektory vysvětlujících proměnných i-tého pacienta, přičemž $\mathbf{X}_i$ má vliv na pravděpodobnost vyléčení a $\mathbf{Z}_i$ na dobu do relapsu. Některé prvky těchto vektorů mohou být společné. Uvedeme různé způsoby modelování závislosti parametrů pravděpodobnosti vyléčení a funkce přežití pro relaps nevyléčených pacientů na vektorech $\mathbf{X}_i$ a $\mathbf{Z}_i$ a pro některé z nich odvodíme odhady metodou maximální věrohodnosti.

3.1 Standardní model

V této sekci se budeme věnovat zavádění vlivu regresorů do standardního modelu (1.1). Pokud $\mathbf{X}_i$ je vektor vysvětlujících proměnných i-tého pacienta, které ovlivňují pravděpodobnost vyléčení, a $\mathbf{Z}_i$ je vektor vysvětlujících proměnných i-tého pacienta, které mají vliv na funkci přežití pro relaps, potom pro model (1.1) můžeme psát

$$\bar{F}(t|\mathbf{X}_i, \mathbf{Z}_i) = p(\mathbf{X}_i) + (1 - p(\mathbf{X}_i))\, \bar{F}_0(t|\mathbf{Z}_i).$$

Farewell (1982) navrhuje parametr pravděpodobnosti vyléčení p nechat záviset na regresorech $\mathbf{X}_i = (1, \mathrm{X}_{i1}, \ldots, \mathrm{X}_{ip})$ skrze logistickou regresi:

$$p_i = p(\mathbf{X}_i) = g^{-1}(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i) = \frac{\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)}{1 + \exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)}, \tag{3.1}$$

kde $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ je vektor regresních parametrů, $g(\cdot)$ je logitová linková funkce a $g^{-1}(\cdot)$ její inverze. Potom pro každé i platí $0 < p_i < 1$ a p_i můžeme interpretovat jako pravděpodobnost vyléčení pacienta, jehož vysvětlující veličiny odpovídají vektoru $\mathbf{X}_i$.

V části 3.1.1 budeme uvažovat, že pouze pravděpodobnost vyléčení závisí na vysvětlujících proměnných, kdežto funkce přežití pro relaps na vysvětlujících proměnných nezávisí a je parametricky specifikována srkze parametry α. Podrobně

odvodíme odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ metodou maximální věrohodnosti a stejně jako v sekci 2.1 budeme uvažovat i zavedení latentních veličin.

Jedním ze způsobů, jak modelovat čas do relapsu pomocí vysvětlujících proměnných je použít Weibullův model (Farewell, 1977). Tedy pro rizikovou funkci $\lambda_0(t|\mathbf{Z}_i)$ a pro funkci přežití $\bar{F}_0(t|\mathbf{Z}_i)$ pro relaps nevyléčených pacientů platí

$$\begin{aligned} \lambda_0(t|\mathbf{Z}_i) &= \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\kappa t^{\kappa-1}, \\ \bar{F}_0(t|\mathbf{Z}_i) &= \exp\left(-\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)t^\kappa\right), \end{aligned} \tag{3.2}$$

kde $\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)$ je parametr měřítka a κ je parametr tvaru Weibullova rozdělení.

Alternativní přístup představují Kuk a Chen (1992), kteří navrhují použít Coxův model proporcionálního rizika. Pro rizikovou funkci $\lambda_0(t)$ pro relaps nevyléčených pacientů platí

$$\lambda_0(t|\mathbf{Z}_i) = \lambda_B(t)\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i), \tag{3.3}$$

kde $\lambda_B(t)$ je základní riziko. Funkci přežití pro relaps potom můžeme zapsat následovně:

$$\bar{F}_0(t|\mathbf{Z}_i) = \exp\left(-\int_0^t \lambda_0(s|\mathbf{Z}_i)\mathrm{d}s\right) = \exp\left(-\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\int_0^t \lambda_B(s)\mathrm{d}s\right).$$

Klasickou metodu maximální věrohodnosti pro odhadování parametrů v takto specifikovaném modelu není obecně možné použít. Jiné přístupy k odhadování navrhují například Kuk a Chen (1992) a Sy a Taylor (2000). Model daný rizikovou funkcí Weibullova rozdělení (3.2) je speciálním případem Coxova modelu pro základní riziko $\lambda_B(t) = \kappa t^{\kappa-1}$.

Speciálním případem modelů (3.2) a (3.3) je exponenciální model, kde riziková funkce pro relaps nevyléčených pacientů je dána vztahem

$$\lambda_0(t|\mathbf{Z}_i) = \lambda_B\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i),$$

kde uvažujeme konstantní základní riziko v čase, tj. $\lambda_B(t) = \lambda_B,\ \forall t$ v modelu (3.3), respektive $\kappa = 1$ v modelu (3.2). Příslušná funkce přežití je pak

$$\bar{F}_0(t|\mathbf{Z}_i) = \exp\left(-t\lambda_B\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right). \tag{3.4}$$

V části 3.1.2 budeme předpokládat, že pravděpodobnost vyléčení na regresorech nezávisí, kdežto funkce přežití pro relaps ano skrze vztah (3.4). Podrobně odvodíme odhady parametrů p, $\boldsymbol{\gamma}$ a λ_B metodou maximální věrohodnosti a stejně jako v předchozích sekcích zavedeme i latentní veličiny.

3.1.1 Regresní model pro pravděpodobnost vyléčení

V této části se budeme věnovat odvození odhadů parametrů v modelu (1.1), kdy uvažujeme, že pravděpodobnost vyléčení p závisí na regresorech $\mathbf{X}_i$ skrze logistickou regresi (3.1), tedy $p_i = \frac{\exp(\beta^\mathsf{T}\mathbf{X}_i)}{1+\exp(\beta^\mathsf{T}\mathbf{X}_i)}$ (Farewell, 1982). Nechť funkce přežití pro relaps je parametricky specifikována skrze parametry $\boldsymbol{\alpha}$, ale nechť na vysvětlujících proměnných nezávisí. Nyní odvodíme odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ metodou maximální věrohodnosti a budeme uvažovat i zavedení latentních veličin. Tato

odvození a konkrétní tvary odhadů nejsou v dostupné literatuře takto podrobně uvedeny.

Uvažujeme-li model bez latentních proměnných, potom logaritmickou věrohodnostní funkci (2.2) lze zapsat následovně:

$$l(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n}\Big\{\delta_i\left[\log(1-p_i) + \log f_0(y_i|\boldsymbol{\alpha})\right] + (1-\delta_i)\log\Big(p_i + (1-p_i)\bar{F}_0(y_i|\boldsymbol{\alpha})\Big)\Big\}. \tag{3.5}$$

Nyní postupujeme stejně jako v kapitole 2. Pro odvození skórové funkce je vhodné použít řetízkové pravidlo, tedy $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial p_i}\frac{\partial p_i}{\partial \boldsymbol{\beta}}$. Označíme-li

$$p_i' = \frac{\partial p_i}{\partial \boldsymbol{\beta}} = \frac{\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\mathbf{X}_i}{\Big(1+\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\Big)^2} = p_i(1-p_i)\mathbf{X}_i,$$

potom složky skórové funkce jsou

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\left[-\frac{\delta_i}{1-p_i} + \frac{(1-\delta_i)\left(1-\bar{F}_0(y_i|\boldsymbol{\alpha})\right)}{p_i + (1-p_i)\bar{F}_0(y_i|\boldsymbol{\alpha})}\right]p_i' \\
&= \sum_{i=1}^{n}\left[-\delta_i + (1-\delta_i)\frac{(1-p_i)\left(1-\bar{F}_0(y_i|\boldsymbol{\alpha})\right)}{p_i + (1-p_i)\bar{F}_0(y_i|\boldsymbol{\alpha})}\right]p_i\mathbf{X}_i \\
&= \sum_{i=1}^{n}\left\{-\delta_i + (1-\delta_i)\left[\frac{1}{p_i + (1-p_i)\bar{F}_0(y_i|\boldsymbol{\alpha})} - 1\right]\right\}p_i\mathbf{X}_i \\
&= \sum_{i=1}^{n}\left[\frac{1-\delta_i}{p_i + (1-p_i)\bar{F}_0(y_i|\boldsymbol{\alpha})} - 1\right]p_i\mathbf{X}_i, \\
\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n}\left[\frac{\delta_i}{f_0(y_i|\boldsymbol{\alpha})}\frac{\partial f_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right. \\
&\qquad \left. + \frac{(1-\delta_i)(1-p_i)}{p_i + (1-p_i)\bar{F}_0(y_i|\boldsymbol{\alpha})}\frac{\partial \bar{F}_0(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right].
\end{aligned}$$

Soustava příslušných věrohodnostních rovnic $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = 0$ a $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$ nemá explicitní řešení a odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ se musí hledat numericky.

Nechť čas do relapsu nevyléčených pacientů se řídí exponenciálním rozdělením, tedy $\bar{F}_0(t|\lambda) = \exp(-\lambda t)$. Stejně jako v kapitole 2, i zde ukážeme věrohodnostní rovnice a tvary informačních matic pro tento jednoduchý speciální případ.

Věrohodnostní rovnice

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\left[\frac{1-\delta_i}{p_i + (1-p_i)\exp(-\lambda y_i)} - 1\right]p_i\mathbf{X}_i = \ 0, \\
\frac{\partial l(\boldsymbol{\beta},\lambda)}{\partial \lambda} &= \sum_{i=1}^{n}\left[\delta_i\frac{1-\lambda y_i}{\lambda} - \frac{(1-\delta_i)(1-p_i)y_i\exp(-\lambda y_i)}{p_i + (1-p_i)\exp(-\lambda y_i)}\right] \\
&= \sum_{i=1}^{n}\left[\frac{\delta_i - \lambda y_i}{\lambda} + \frac{(1-\delta_i)p_i y_i}{p_i + (1-p_i)\exp(-\lambda y_i)}\right] = \ 0
\end{aligned}$$

nemají explicitní řešení a odhady parametrů $\boldsymbol{\beta}$ a λ je třeba hledat pomocí numerických metod. Fisherova informační matice je

$$I(p,\lambda) = -\,\mathsf{E}\begin{pmatrix} \frac{\partial^2 l_i(p,\lambda)}{\partial p^2} & \frac{\partial^2 l_i(p,\lambda)}{\partial p \partial \lambda} \\ \frac{\partial^2 l_i(p,\lambda)}{\partial p \partial \lambda} & \frac{\partial^2 l_i(p,\lambda)}{\partial \lambda^2} \end{pmatrix},$$

kde $l_i(p,\lambda)$ je příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (3.5). Pozorovaná informační matice má tvar

$$I_n(p,\lambda) = -\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l(p,\lambda)}{\partial p^2} & \frac{\partial^2 l(p,\lambda)}{\partial p \partial \lambda} \\ \frac{\partial^2 l(p,\lambda)}{\partial p \partial \lambda} & \frac{\partial^2 l(p,\lambda)}{\partial \lambda^2} \end{pmatrix} \tag{3.6}$$

s následujícími složkami

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}^2} =& \sum_{i=1}^{n}\Bigg\{ -\frac{(1-\delta_i)p_i\left(1-\exp(-\lambda y_i)\right)}{\left[p_i+(1-p_i)\exp(-\lambda y_i)\right]^2} \\
&\qquad + \frac{1-\delta_i}{p_i+(1-p_i)\exp(-\lambda y_i)} - 1\Bigg\} p_i(1-p_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}} \\
=& \sum_{i=1}^{n}\left\{ \frac{(1-\delta_i)\exp(-\lambda y_i)}{\left[p_i+(1-p_i)\exp(-\lambda y_i)\right]^2} - 1\right\} p_i(1-p_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}} \\
\frac{\partial^2 l(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial \lambda} =& \sum_{i=1}^{n}\frac{(1-\delta_i)y_i\exp(-\lambda y_i)}{\left[p_i+(1-p_i)\exp(-\lambda y_i)\right]^2}\; p_i(1-p_i)\mathbf{X}_i, \\
\frac{\partial^2 l(\boldsymbol{\beta},\lambda)}{\partial \lambda^2} =& \sum_{i=1}^{n}\left\{ -\frac{\delta_i}{\lambda^2} + \frac{(1-\delta_i)p_i(1-p_i)y_i^2\exp(-\lambda y_i)}{\left[p_i+(1-p_i)\exp(-\lambda y_i)\right]^2}\right\}.
\end{aligned}$$

Z tvaru druhých parciálních derivací usuzujeme, že pozorovaná informační matice není obecně pozitivně definitní, a tedy řešení věrohodnostních matic nemusí být nutně globálním maximem.

Nyní, stejně jako v kapitole 2, zaveďme do standardního modelu (1.1) latentní indikátor vyléčení C_i, jak navrhují například Peng a Dear (2000) nebo Sy a Taylor (2000). V následující části odvodíme maximálně věrohodné odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ za předpokladu, že známe indikátory vyléčení. Toto podrobné odvození není v literatuře uvedeno.

Logaritmická věrohodnostní funkce, pokud bychom znali indikátory vyléčení, má tvar

$$\begin{aligned}
l^C(p,\boldsymbol{\alpha}) =& \sum_{i=1}^{n}\left[\delta_i \log f_0(y_i|\boldsymbol{\alpha}) + (1-\delta_i-c_i)\log \bar{F}_0(y_i|\boldsymbol{\alpha})\right] \\
&+ \sum_{i=1}^{n}\left[c_i \log p_i + (1-c_i)\log(1-p_i)\right].
\end{aligned} \tag{3.7}$$

Díky zavedení latentní proměnné lze logaritmickou věrohodnost zapsat jako součet dvou sum, přičemž první nezávisí na parametrech $\boldsymbol{\beta}$ a má stejný tvar jako v rovnici (2.8). Druhá suma nezávisí na parametrech $\boldsymbol{\alpha}$, a tedy skórová funkce

a příslušná věrohodnostní rovnice pro parametr $\boldsymbol{\alpha}$ mají stejný tvar jako pro logaritmickou věrohodnostní funkci (2.8). Tyto rovnice nebudeme znovu uvádět.

Pro odvození skórové funkce pro parametry $\boldsymbol{\beta}$ použijeme řetízkové pravidlo. Složka příslušná parametrům $\boldsymbol{\beta}$ je

$$\begin{aligned}\frac{\partial l^C(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\left(\frac{c_i}{p_i}-\frac{1-c_i}{1-p_i}\right)p_i' \\ &= \sum_{i=1}^{n}\left(\frac{c_i}{p_i}-\frac{1-c_i}{1-p_i}\right)p_i(1-p_i)\mathbf{X}_i \\ &= \sum_{i=1}^{n}\left[c_i(1-p_i)-(1-c_i)p_i\right]\mathbf{X}_i \\ &= \sum_{i=1}^{n}\left(c_i-p_i\right)\mathbf{X}_i,\end{aligned}$$

což odpovídá skórové funkci logistické regrese. Příslušná věrohodnostní rovnice $\frac{\partial l^C(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}} = 0$ nemá explicitní řešení a je nutno jej hledat numericky.

Dále uvažujme, že čas do relapsu se řídí exponenciálním rozdělením s neznámým parametrem λ. Fisherova informační matice má tvar

$$I(\boldsymbol{\beta},\lambda) = \begin{pmatrix} \mathsf{E}\left(p_i(1-p_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}\right) & 0 \\ 0 & \frac{\mathsf{P}(T^u \leq T^c)}{\lambda^2} \end{pmatrix}.$$

Pozorovaná informační matice je

$$I_n(\boldsymbol{\beta},\lambda) = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} p_i(1-p_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}} & 0 \\ 0 & \sum_{i=1}^{n}\frac{\delta_i}{\lambda^2} \end{pmatrix},$$

tedy se jedná o blokovou matici se dvěma nenulovými submaticemi:

$$\begin{aligned} I_{1n}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} p_i(1-p_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}, \\ I_{2n}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\lambda^2}. \end{aligned}$$

Jelikož $0 < p_i < 1,\ \forall i$, pak matice $I_{1n}(\boldsymbol{\beta})$ je pozitivně definitní právě tehdy, když vektory $\mathbf{X}_i$ jsou lineárně nezávislé (Agresti a Kateri, 2011). Pokud dojde aspoň u jednoho pacienta k recidivě onemocnění, tj. $\sum_{i=1}^{n}\delta_i \neq 0$, pak i matice $I_{2n}(\lambda)$ je pozitivně definitní. Pokud jsou obě matice $I_{1n}(\boldsymbol{\beta})$ a $I_{2n}(\lambda)$ pozitivně definitní, pak i pozorovaná informační matice je pozitivně definitní, a tedy odhady parametrů $\boldsymbol{\beta}$ a λ jsou maximálně věrohodné.

Obdobně jako v sekci 2.1, i zde je možné k odhadu parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ (respektive λ) použít EM algoritmus. Implementace algoritmu je analogická.

3.1.2 Regresní model pro funkci přežití pro relaps

Nyní předpokládejme, že pravděpodobnost vyléčení p ve standardním modelu (1.1) na vysvětlujících proměnných nezávisí. Nechť $\mathbf{Z}_i$ je vektor regresorů a nechť

závislost funkce přežití pro relaps na tomto vektoru je dána exponenciálním modelem (3.4), tedy platí

$$\begin{aligned}\lambda_0(t|\mathbf{Z}_i) &= \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i),\\ \bar{F}_0(t|\mathbf{Z}_i) &= \exp\left(-t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right).\end{aligned}$$

Potom pro příslušnou hustotu platí:

$$f_0(t|\mathbf{Z}_i) = \lambda_0(t|\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i) = \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\exp\left(-t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right).$$

V této části odvodíme odhady parametrů p, $\boldsymbol{\gamma}$ a λ_B metodou maximální věrohodnosti. Budeme uvažovat i zavedení latentních proměnných. Postupujeme obdobně jako v sekci 2.1. Tato odvození nejsou uvedena v literatuře.

Logaritmickou věrohodnost standardního modelu (1.1) bez latentních proměnných lze zapsat ve tvaru

$$\begin{aligned}l(p,\boldsymbol{\gamma},\lambda_B) = \sum_{i=1}^{n}\Big[&\delta_i\left(\log(1-p) + \log f_0(y_i|\mathbf{Z}_i)\right)\\ &+ (1-\delta_i)\log\left(p + (1-p)\bar{F}_0(y_i|\mathbf{Z}_i)\right)\Big].\end{aligned} \tag{3.8}$$

Složky skórové funkce příslušné parametrům $\boldsymbol{\gamma}$ a λ_B získáme pomocí řetízkového pravidla. Platí

$$\begin{aligned}\frac{\partial \bar{F}_0(t|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}} &= -t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\exp\left(-t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right)\mathbf{Z}_i\\ &= -t\lambda_0(t|\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i)\mathbf{Z}_i,\\ \frac{\partial \bar{F}_0(t|\mathbf{Z}_i)}{\partial \lambda_B} &= -t\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\exp\left(-t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right)\\ &= -t\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i),\\ \frac{\partial f_0(t|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}} &= \frac{\partial \lambda_0(t|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}}\bar{F}_0(t|\mathbf{Z}_i) + \lambda_0(t|\mathbf{Z}_i)\frac{\partial \bar{F}_0(t|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}}\\ &= \lambda_0(t|\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i)\mathbf{Z}_i - t\lambda_0^2(t|\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i)\mathbf{Z}_i\\ &= \lambda_0(t|\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i)\left(1 - t\lambda_0(t|\mathbf{Z}_i)\right)\mathbf{Z}_i,\\ \frac{\partial f_0(t|\mathbf{Z}_i)}{\partial \lambda_B} &= \frac{\partial \lambda_0(t|\mathbf{Z}_i)}{\partial \lambda_B}\bar{F}_0(t|\mathbf{Z}_i) + \lambda_0(t|\mathbf{Z}_i)\frac{\partial \bar{F}_0(t|\mathbf{Z}_i)}{\partial \lambda_B}\\ &= \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i) - t\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\lambda_0(t|\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i)\\ &= \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(t|\mathbf{Z}_i)\left(1 - t\lambda_0(t|\mathbf{Z}_i)\right).\end{aligned}$$

Složky skórové funkce jsou

$$\begin{aligned}\frac{\partial l(p,\boldsymbol{\gamma},\lambda_B)}{\partial p} &= \sum_{i=1}^{n}\left[-\frac{\delta_i}{1-p} + (1-\delta_i)\frac{1-\bar{F}_0(y_i|\mathbf{Z}_i)}{p + (1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\right]\\ \frac{\partial l(p,\boldsymbol{\gamma},\lambda_B)}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n}\left[\delta_i\frac{\frac{\partial f_0(y_i|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}}}{f_0(y_i|\mathbf{Z}_i)} + (1-\delta_i)\frac{(1-p)\frac{\partial \bar{F}_0(y_i|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}}}{p + (1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\right]\\ &= \sum_{i=1}^{n}\left[\delta_i\frac{\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)\left(1 - y_i\lambda_0(y_i|\mathbf{Z}_i)\right)\mathbf{Z}_i}{\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)}\right.\end{aligned}$$

$$
\begin{aligned}
&\quad - (1-\delta_i)\frac{(1-p)y_i\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)\mathbf{Z}_i}{p+(1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\Bigg] \\
&= \sum_{i=1}^{n}\Bigg[\delta_i\left(1-y_i\lambda_0(y_i|\mathbf{Z}_i)\right) \\
&\quad - (1-\delta_i)\frac{(1-p)y_i\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)}{p+(1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\Bigg]\mathbf{Z}_i, \\
\frac{\partial l(p,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B} &= \sum_{i=1}^{n}\left[\delta_i\frac{\frac{\partial f_0(y_i|\mathbf{Z}_i)}{\partial\lambda_B}}{f_0(y_i|\mathbf{Z}_i)} + (1-\delta_i)\frac{(1-p)\frac{\partial\bar{F}_0(y_i|\mathbf{Z}_i)}{\partial\lambda_B}}{p+(1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\right] \\
&= \sum_{i=1}^{n}\Bigg[\delta_i\frac{\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)\left(1-y_i\lambda_0(y_i|\mathbf{Z}_i)\right)}{\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)} \\
&\quad - (1-\delta_i)\frac{(1-p)y_i\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)}{p+(1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\Bigg] \\
&= \sum_{i=1}^{n}\Bigg[\frac{\delta_i}{\lambda_B}\left(1-y_i\lambda_0(y_i|\mathbf{Z}_i)\right) \\
&\quad - (1-\delta_i)\frac{(1-p)y_i\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(y_i|\mathbf{Z}_i)}{p+(1-p)\bar{F}_0(y_i|\mathbf{Z}_i)}\Bigg].
\end{aligned}
$$

Soustava věrohodnostních rovnic $\frac{\partial l(p,\boldsymbol{\gamma},\lambda_B)}{\partial p}=0$, $\frac{\partial l(p,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}}=0$ a $\frac{\partial l(p,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B}=0$ nemá explicitní řešení, a tedy je jej třeba hledat pomocí numerických metod.

Označíme-li $l_i(p,\boldsymbol{\gamma},\lambda_B)$ příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (3.8), pak pro složky Fisherovy a pozorované informační matice platí

$$
\begin{aligned}
\frac{\partial l_i^2(p,\boldsymbol{\gamma},\lambda_B)}{\partial p^2} &= -\frac{\delta_i}{(1-p)^2} - (1-\delta_i)\left[\frac{1-\bar{F}_0(y_i|\boldsymbol{\alpha})}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})}\right]^2, \\
\frac{\partial l_i^2(p,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}^2} &= -\delta_i y_i\lambda_0(y_i|\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T} - (1-\delta_i)(1-p)y_i\Bigg\{\frac{\frac{\partial\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})}{\partial\boldsymbol{\gamma}}}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} \\
&\quad - \frac{\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})(1-p)\frac{\partial\bar{F}_0(y_i|\boldsymbol{\alpha})}{\partial\boldsymbol{\gamma}}}{\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^2}\Bigg\} \\
&= -\delta_i y_i\lambda_0(y_i|\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T} - (1-\delta_i)(1-p)y_i \\
&\quad \times \frac{\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})-py_i\lambda_0(y_i|\mathbf{Z}_i)\right]\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T}}{\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^2}, \\
\frac{\partial l_i^2(p,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B^2} &= -\frac{\delta_i}{\lambda_B^2} + \frac{(1-\delta_i)(1-p)py_i^2\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)^2\bar{F}_0(y_i|\boldsymbol{\alpha})}{\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^2}, \\
\frac{\partial^2 l_i(p,\boldsymbol{\gamma},\lambda_B)}{\partial p\partial\boldsymbol{\gamma}} &= (1-\delta_i)\frac{y_i\lambda_0(y_i|\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})\mathbf{Z}_i}{\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^2},
\end{aligned}
$$

$$\frac{\partial^2 l_i(p,\boldsymbol{\gamma},\lambda_B)}{\partial p \partial \lambda_B} = (1-\delta_i)\frac{y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})}{\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^2},$$

$$\begin{aligned}\frac{\partial^2 l_i(p,\boldsymbol{\gamma},\lambda_B)}{\partial \boldsymbol{\gamma} \partial \lambda_B} =& -\delta_i y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i - (1-\delta_i)(1-p)y_i \\ &\times \left\{ \frac{\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})(1-\lambda_0(y_i|\mathbf{Z}_i))\mathbf{Z}_i}{p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})} \right. \\ &\left. + \frac{\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\bar{F}_0(y_i|\boldsymbol{\alpha})^2(1-p)y_i\lambda_0(y_i|\mathbf{Z}_i)\mathbf{Z}_i}{\left[p+(1-p)\bar{F}_0(y_i|\boldsymbol{\alpha})\right]^2} \right\}.\end{aligned}$$

Ze tvaru těchto vzorců usuzujeme, že pozorovaná informační matice nemusí být vždy pozitivně definitní.

Obdobně jako v předchozí sekci, i zde zavedeme do modelu (1.1) latentní indikátor uzdravení pacientů. Logaritmická věrohodnostní funkce pak odpovídá tvaru

$$\begin{aligned}l^C(p,\boldsymbol{\gamma},\lambda_B) =& \sum_{i=1}^n \left[\delta_i \log \lambda_0(y_i|\mathbf{Z}_i) + (1-c_i)\log \bar{F}_0(y_i|\boldsymbol{\alpha})\right] \\ &+ \sum_{i=1}^n \left[c_i \log p + (1-c_i)\log(1-p)\right] \\ =& \sum_{i=1}^n \left[\delta_i(\gamma^\mathsf{T}\mathbf{Z}_i + \log\lambda_B) - (1-c_i)\lambda_B y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right] \\ &+ \sum_{i=1}^n \left[c_i \log p + (1-c_i)\log(1-p)\right]. \qquad (3.9)\end{aligned}$$

Logaritmickou věrohodnost (3.9) lze vyjádřit jako součet dvou sum, přičemž první nezávisí na parametru p a druhá na parametrech $\boldsymbol{\gamma}$ a λ_B. Druhá suma navíc odpovídá výrazu v rovnici (2.9), a tedy skórová funkce, věrohodnostní rovnice a části Fisherovy a pozorované informační matice budou stejné jako v sekci 2.1 při latentních veličinách. Tyto rovnice tedy nebudeme znovu odvozovat.

Složky skórové funkce příslušné parametrům $\boldsymbol{\gamma}$ a λ_B jsou

$$\begin{aligned}\frac{\partial l^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \gamma} =& \sum_{i=1}^n \left[\delta_i - (1-c_i)y_i\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right]\mathbf{Z}_i, \\ \frac{\partial l^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \lambda_B} =& \sum_{i=1}^n \left[\frac{\delta_i}{\lambda_B} - (1-c_i)y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right].\end{aligned}$$

Věrohodnostní rovnice $\frac{\partial l_C(p,\gamma,\lambda_B)}{\partial \gamma} = 0$ nemá explicitní řešení. Řešením rovnice $\frac{\partial l_C(p,\gamma,\lambda_B)}{\partial \lambda_B}$ při známých parametrech $\boldsymbol{\gamma}$ je

$$\hat{\lambda}_b = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n (1-c_i)y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)}.$$

Označme $l_i^C(p,\boldsymbol{\gamma},\lambda_B)$ příspěvek i-tého pacienta do věrohodnostní funkce (3.9).

Potom pro jednotlivé složky Fisherovy a pozorované informační matice platí

$$\begin{aligned}
\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial p^2} &= -\frac{c_i}{p^2} - \frac{1-c_i}{(1-p)^2},\\
\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \gamma^2} &= -(1-c_i) y_i \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T},\\
\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \lambda_B^2} &= -\frac{\delta_i}{\lambda_B^2},\\
\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \gamma \partial \lambda_B} &= -(1-c_i) y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i.
\end{aligned}$$

Navíc platí $\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial p \partial \lambda_B} = \frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial p \partial \gamma} = 0$. A tedy pozorovaná informační matice $I_n(p,\gamma,\lambda_B)$ je bloková matice s dvěma nenulovými submaticemi

$$\begin{aligned}
I_{1n}(p) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial p^2} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{c_i}{p^2} + \frac{1-c_i}{(1-p)^2}\right],\\
I_{2n}(\boldsymbol{\gamma},\lambda_B) &= -\frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n}\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \gamma^2} & \sum_{i=1}^{n}\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \gamma \partial \lambda_B} \\ \sum_{i=1}^{n}\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \gamma \partial \lambda_B} & \sum_{i=1}^{n}\frac{\partial^2 l_i^C(p,\boldsymbol{\gamma},\lambda_B)}{\partial \lambda_B^2} \end{pmatrix}. \qquad (3.10)
\end{aligned}$$

Pokud jsou obě submatice $I_{1n}(p)$ a $I_{2n}(\boldsymbol{\gamma},\lambda_B)$ pozitivně definitní, pak i pozorovaná informační matice $I_n(p,\gamma,\lambda_B)$ je pozitivně definitní. V sekci 2.1 bylo ukázáno, že platí $\sum_{i=1}^{n}\left[\frac{c_i}{p^2} + \frac{1-c_i}{(1-p)^2}\right] > 0$, a tedy matice $I_{1n}(p)$ je pozitivně definitní. Z tvaru druhých parciálních derivací usuzujeme, že submatice $I_{2n}(\boldsymbol{\gamma},\lambda_B)$ není vždy pozitivně definitní, tudíž ani pozorovaná informační matice $I_n(\theta,\gamma,\lambda_B)$ nemusí být vždy pozitivně definitní.

3.2 Bio model

V této sekci zavedeme do bio modelu (1.2) vysvětlující veličiny. Pokud $\mathbf{X}_i$ je vektor regresorů i-tého pacienta, které ovlivňují parametr pravděpodobnosti vyléčení, a $\mathbf{Z}_i$ je vektor regresorů i-tého pacienta, které mají vliv na distribuční funkci $G(t)$, potom pro model (1.2) můžeme psát

$$\bar{F}(t|\mathbf{X}_i,\mathbf{Z}_i) = \exp\left(\theta(\mathbf{X}_i)G(t|\mathbf{Z}_i)\right).$$

Chen a kol. (1999) navrhují nechat parametr θ záviset na regresorech skrze vztah

$$\theta_i = \theta(\mathbf{X}_i) = g^{-1}(\beta^\mathsf{T}\mathbf{X}_i) = \exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i), \qquad (3.11)$$

kde $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$ je vektor regresních parametrů a $g^{-1}(\cdot)$ je inverze logaritmické linkové funkce. Tento vztah mezi parametry θ a $\boldsymbol{\beta}$ odpovídá kanonické linkové funkci Poissonova modelu pro parametr θ (Chen a kol., 1999). Pravděpodobnost vyléčení i-tého pacienta s regresory $\mathbf{X}_i$ je tedy dána výrazem $\exp\left(-\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\right)$.

V části 3.2.1 budeme předpokládat, že pouze parametr θ závisí na regresorech a distribuční funkce $G(t)$ je parametricky specifikována skrze parametry $\boldsymbol{\alpha}$, avšak na vysvětlujících proměnných nezávisí. Podrobně odvodíme odhady parametrů $\boldsymbol{\beta}$

a $\boldsymbol{\alpha}$ metodou maximální věrohodnosti a stejně jako v sekci 2.2 budeme uvažovat i zavedení latentních proměnných.

Distribuční funkci $G(t)$ lze nechat záviset na regresorech obdobně, jak bylo popsáno na začátku sekce 3.1. Na rozdíl od modelu (1.2), kde modelujeme nevlastní funkci přežití pomocí funkce přežití pro relaps, zde se používá distribuční funkce.

Za platnosti Weibullova modelu pro rizikovou funkci platí

$$\lambda(t|\mathbf{Z}_i) = \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\kappa t^{\kappa-1}$$

a distribuční funkce $G(t)$ má tvar

$$G(t|\mathbf{Z}_i) = 1 - \exp\left(-\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)t^\kappa\right), \tag{3.12}$$

kde $\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)$ je parametr měřítka a κ je parametr tvaru Weibullova rozdělení.

Yin a Ibrahim (2005) navrhují použít Coxův model proporcionálního rizika, tedy pro rizikovou funkci $\lambda(t)$ a distribuční funkci $G(t)$ platí

$$\begin{aligned}\lambda(t|\mathbf{Z}_i) &= \lambda_B(t)\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i),\\ G(t|\mathbf{Z}_i) &= 1 - \exp\left(-\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\int_0^t \lambda_B(s)\mathrm{d}s\right),\end{aligned} \tag{3.13}$$

kde $\lambda_B(t)$ je základní riziko. Pro tento model ovšem nelze použít pro odhadování parametrů metodu maximální věrohodnosti.

Další alternativou je použití po částech exponenciálního modelu (Chen a Ibrahim, 2001). Označme $0 = s_0 < s_1 < \cdots < s_J$ konečné dělení časové osy, kde $s_J > \max_{1\le i\le n} y_i$. Pak riziková funkce je po částech konstantní, tedy

$$\lambda(t|\mathbf{Z}_i) = \lambda_j,\ t \in (s_{j-1}, s_j]$$

a distribuční funkci $G(t|\mathbf{Z}_i)$ lze zapsat ve tvaru

$$G(t|\mathbf{Z}_i) = 1 - \exp\left(-\lambda_j(t - s_{j-1}) - \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right),\ t \in (s_{j-1}, s_j]. \tag{3.14}$$

Speciálním případem výše uvedených modelů je exponenciální model. Tedy uvažujeme-li $\kappa = 1$ ve Weibullově modelu (3.12), respektive konstantní základní riziko v čase, tj. $\lambda_B(t) = \lambda_B,\ \forall t$ v Coxově modelu proporcionálního rizika (3.13), respektive $J = 1$ v po částech exponenciálního modelu (3.14), pro distribuční funkci $G(t)$ platí

$$G(t|\mathbf{Z}_i) = 1 - \exp\left(-t\lambda_B\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right). \tag{3.15}$$

V části 3.2.2 budeme předpokládat, že pravděpodobnost vyléčení na vysvětlujících proměnných nezávisí, kdežto distribuční funkce $G(t)$ ano skrze exponenciální model (3.15). Opět podrobně odvodíme odhady parametrů p, $\boldsymbol{\gamma}$ a λ_B metodou maximální věrohodnosti a do modelu zavedeme i latentní veličiny počtu karcinogenních buněk.

3.2.1 Regresní model pro pravděpodobnost vyléčení

Předpokládejme, že parametr θ v modelu (1.2) závisí na regresorech $\mathbf{X}_i$ skrze vztah (3.11), tedy $\theta_i = \exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)$ (Chen a kol., 1999). Nechť distribuční funkce $G(t)$ je parametricky specifikována skrze vektor parametrů $\boldsymbol{\alpha}$ a na regresorech nezávisí. V této části odvodíme odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ metodou maximální věrohodnosti. Tato odvození v literatuře zatím nebyla uvedena.

Logaritmická věrohodnostní funkce, neuvažujeme-li latentní veličiny, má tvar

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \Big[\delta_i \left(\log \theta_i + \log g(y_i|\boldsymbol{\alpha})\right) - \theta_i G(y_i|\boldsymbol{\alpha})\Big]. \tag{3.16}$$

Nyní budeme postupovat analogicky jako v sekci 2.2. Složky skórové funkce příslušné logaritmické věrohodnosti (3.16) odvodíme pomocí řetízkového pravidla, tedy $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}}$. Označme

$$\theta_i' = \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\mathbf{X}_i = \theta_i \mathbf{X}_i,$$

potom složky skórové funkce jsou

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left(\frac{\delta_i}{\theta_i} - G(y_i|\boldsymbol{\alpha})\right) \theta_i' = \sum_{i=1}^{n} \left(\frac{\delta_i}{\theta_i} - G(y_i|\boldsymbol{\alpha})\right) \theta_i \mathbf{X}_i \\
&= \sum_{i=1}^{n} \left(\delta_i - \theta_i G(y_i|\boldsymbol{\alpha})\right) \mathbf{X}_i, \\
\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n} \left(\frac{\delta_i}{g(y_i|\boldsymbol{\alpha})}\frac{\partial g(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} - \theta_i \frac{\partial G(y_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right).
\end{aligned}$$

Soustava věrohodnostních rovnic $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = 0$ a $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$ nemá explicitní řešení a pro hledání odhadů parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ je nutno použít numerické metody.

Pro speciální případ, kdy $G(t|\lambda) = 1 - \exp(-\lambda y_i)$, uvedeme tvar Fisherovy a pozorované informační matice. Složky skórové funkce jsou

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left[\delta_i - \theta_i \left(1 - \exp(-\lambda y_i)\right)\right] \mathbf{X}_i, \\
\frac{\partial l(\boldsymbol{\beta}, \lambda)}{\partial \lambda} &= \sum_{i=1}^{n} \left[\frac{\delta_i(1 - \lambda y_i)}{\lambda} + y_i \theta_i \exp(-\lambda y_i)\right].
\end{aligned}$$

Označme $l_i(\boldsymbol{\beta}, \lambda)$ příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (3.16). Potom platí

$$\begin{aligned}
\frac{\partial^2 l_i(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}^2} &= -\theta_i \left(1 - \exp(-\lambda y_i)\right) \mathbf{X}_i \mathbf{X}_i^\mathsf{T}, \\
\frac{\partial^2 l_i(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta} \partial \lambda} &= -y_i \theta_i \exp(-\lambda y_i) \mathbf{X}_i, \\
\frac{\partial^2 l_i(\boldsymbol{\beta}, \lambda)}{\partial \lambda^2} &= -\frac{\delta_i}{\lambda^2} - \theta_i y_i^2 \exp(-\lambda y_i).
\end{aligned}$$

Fisherova informační matice má tvar

$$I(\boldsymbol{\beta},\lambda) = -\,\mathsf{E}\begin{pmatrix} \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}^2} & \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial\lambda} \\ \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial\lambda} & \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \lambda^2} \end{pmatrix}.$$

Pozorovaná informační matice je

$$I_n(\boldsymbol{\beta},\lambda) = -\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}^2} & \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial\lambda} \\ \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial\lambda} & \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\beta},\lambda)}{\partial \lambda^2} \end{pmatrix},$$

přičemž z jejího tvaru se zdá, že nemusí být obecně pozitivně definitní.

Zaveďme nyní do modelu (1.2) latentní počty karcinogenních buněk stejně jako v sekci 2.2. Pak logaritmickou věrohodnostní funkci lze dle Chen a kol. (1999) zapsat následovně

$$\begin{aligned} l^N(\boldsymbol{\beta},\boldsymbol{\alpha}) =& \sum_{i=1}^n \Big[(N_i-\delta_i)\log \bar{G}(y_i|\boldsymbol{\alpha}) + \delta_i\left(\log N_i + \log g(y_i|\boldsymbol{\alpha})\right]\Big) \\ &+ \sum_{i=1}^n \Big(N_i \log\theta_i - \log(N_i!) - \theta_i\Big) \\ =& \sum_{i=1}^n \Big[N_i \log \bar{G}(y_i|\boldsymbol{\alpha}) + \delta_i\left(\log N_i + \log \lambda(y_i|\boldsymbol{\alpha})\right)\Big] \\ &+ \sum_{i=1}^n \Big(N_i \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i - \log(N_i!) - \exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\Big), \end{aligned} \tag{3.17}$$

kde $\lambda(y_i|\boldsymbol{\alpha}) = \frac{g(y_i|\boldsymbol{\alpha})}{\bar{G}(y_i|\boldsymbol{\alpha})}$ je riziková funkce. Po zavedení latentních proměnných N_i lze logaritmickou věrohodnost (3.17) zapsat jako součet dvou sum, kde první nezávisí na parametrech $\boldsymbol{\beta}$ a druhá na parametrech $\boldsymbol{\alpha}$. Navíc první ze sum je shodná s první sumou v logaritmické věrohodnostní funkci (2.17), a tedy skórová funkce, věrohodnostní rovnice a příslušné členy ve Fisherově a pozorované informační matici budou stejné. Odvození těchto výrazů zde nebude znovu uvedeno. Druhá suma odpovídá logaritmické věrohodnosti pro Poissonův model.

Nyní odvodíme odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ metodou maximální věrohodnosti. Tato odvození nejsou v literatuře podrobně ukázána. Pokud bychom znali počty karcinogenních buněk N_i, potom skórová funkce příslušná parametru $\boldsymbol{\beta}$ je

$$\frac{\partial l^N(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \Big(N_i - \exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\Big)\,\mathbf{X}_i.$$

Soustava věrohodnostních rovnic $\frac{\partial l^N(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = 0$ a $\frac{\partial l^N(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$ nemá explicitní řešení a je jej nutné hledat numericky.

Pro volbu $G(t|\lambda) = 1 - \exp(-\lambda y_i)$ odvodíme tvar Fisherovy a pozorované informační matice. Je-li $l_i^N(\boldsymbol{\beta},\lambda)$ příspěvek i-tého pacienta do logaritmické věrohodnostní funkce (3.17), potom pro Fisherovu informační matici platí:

$$I(\boldsymbol{\beta},\lambda) = -\,\mathsf{E}\begin{pmatrix} \frac{\partial^2 l_i^N(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}^2} & \frac{\partial^2 l_i^N(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial\lambda} \\ \frac{\partial^2 l_i^N(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}\partial\lambda} & \frac{\partial^2 l_i^N(\boldsymbol{\beta},\lambda)}{\partial \lambda^2} \end{pmatrix} = \begin{pmatrix} \mathsf{E}\left(\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathsf{T}}\right) & 0 \\ 0 & \frac{\mathsf{P}(T^u \le T^c)}{\lambda^2} \end{pmatrix}.$$

Pozorovaná informační matice je

$$I_n(\boldsymbol{\beta},\lambda) = -\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \frac{\partial^2 l_i^N(\beta,\lambda)}{\partial\beta^2} & \sum_{i=1}^n \frac{\partial^2 l_i^N(\beta,\lambda)}{\partial\beta\partial\lambda} \\ \sum_{i=1}^n \frac{\partial^2 l_i^N(\beta,\lambda)}{\partial\beta\partial\lambda} & \sum_{i=1}^n \frac{\partial^2 l_i^N(\beta,\lambda)}{\partial\lambda^2} \end{pmatrix}$$
$$= \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^\mathsf{T} & 0 \\ 0 & \sum_{i=1}^n \frac{\delta_i}{\lambda^2} \end{pmatrix}. \tag{3.18}$$

Pozorovaná informační matice (3.18) je tedy bloková matice se dvěma nenulovými submaticemi

$$I_{1n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^\mathsf{T},$$
$$I_{2n}(\lambda) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\lambda^2}.$$

Jelikož $\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i) > 0$, pak submatice $I_{1n}(\boldsymbol{\beta})$ je pozitivně definitní právě tehdy, když vektory $\mathbf{X}_i$ jsou lineárně nezávislé (Agresti a Kateri, 2011). Pokud aspoň u jednoho pacienta došlo k relapsu, pak $\sum_{i=1}^n \frac{\delta_i}{\lambda^2} > 0$, a tedy submatice $I_{2n}(\lambda)$ je pozitivně definitní. Jsou-li obě submatice pozitivně definitní, pak i pozorovaná informační matice (3.18) je pozitivně definitní, a tedy odhady parametrů $\boldsymbol{\beta}$ a λ jsou maximálně věrohodné.

Parametry $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ lze v tomto případě odhadnout pomocí EM algoritmu, jehož implementace je analogická postupu uvedeného v sekci (2.2).

3.2.2 Regresní model pro distribuční funkci

Nyní předpokládejme, že parametr θ na vysvětlujících veličinách nezávisí. Nechť $\mathbf{Z}_i$ je vektor regresorů a nechť distribuční funkce $G(t)$ na těchto regresorech závisí skrze vztah (3.15), tedy se řídí exponenciálním modelem. Riziková funkce $\lambda(t)$ a distribuční funkce $G(t)$ mají následující tvar

$$\lambda(t|\mathbf{Z}_i) = \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i),$$
$$G(t|\mathbf{Z}_i) = 1 - \exp\left(-t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right),$$

kde λ_B je základní riziko. Hustota příslušná distribuční funkci $G(t|\mathbf{Z}_i)$ je

$$g(t|\mathbf{Z}_i) = \lambda(t|\mathbf{Z}_i)\bar{G}(t|\mathbf{Z}_i) = \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\exp\left(-t\lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right),$$

kde $\bar{G}(t|\mathbf{Z}_i) = 1 - G(t|\mathbf{Z}_i)$.

V této části odvodíme odhady parametrů θ, $\boldsymbol{\gamma}$ a λ_B metodou maximální věrohodnosti. Stejně jako v předchozích sekcích budeme zde uvažovat i zavedení latentních veličin počtu karcinogenních buněk. Podrobná odvození odhadů nejsou uvedena v dostupné literatuře.

Logaritmická věrohodnostní funkce pro model (1.2), neuvažujeme-li latentní veličiny, je

$$l(\theta,\boldsymbol{\gamma},\lambda_B) = \sum_{i=1}^n \left[\delta_i\left(\log\theta + \log g(y_i|\mathbf{Z}_i)\right) - \theta G(y_i|\mathbf{Z}_i)\right]. \tag{3.19}$$

Pro odvození složek skórové funkce použijeme následující vztahy:

$$\frac{\partial G(t|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}} = t\lambda(t|\mathbf{Z}_i)(1 - G(t|\mathbf{Z}_i))\mathbf{Z}_i,$$
$$\frac{\partial G(t|\mathbf{Z}_i)}{\partial \lambda_B} = t\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)(1 - G(t|\mathbf{Z}_i)).$$

V sekci 3.1.2 jsme odvodili parciální derivace hustoty, tedy platí

$$\frac{\partial g(t|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}} = \lambda(t|\mathbf{Z}_i)\left(1 - G(t|\mathbf{Z}_i)\right)\left(1 - t\lambda(t|\mathbf{Z}_i)\right)\mathbf{Z}_i,$$
$$\frac{\partial g(t|\mathbf{Z}_i)}{\partial \lambda_B} = \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\left(1 - G(t|\mathbf{Z}_i)\right)\left(1 - t\lambda(t|\mathbf{Z}_i)\right).$$

Složky skórové funkce pro logaritmickou věrohodnost (3.19) jsou

$$\begin{aligned}
\frac{\partial l(\theta, \boldsymbol{\gamma}, \lambda_B)}{\partial \theta} &= \sum_{i=1}^{n}\left(\frac{\delta_i}{\theta} - G(y_i|\mathbf{Z}_i)\right), \\
\frac{\partial l(\theta, \boldsymbol{\gamma}, \lambda_B)}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n}\left(\delta_i \frac{\frac{\partial g(y_i|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}}}{g(y_i|\mathbf{Z}_i)} - \theta\frac{\partial G(y_i|\mathbf{Z}_i)}{\partial \boldsymbol{\gamma}}\right) \\
&= \sum_{i=1}^{n}\left[\delta_i \frac{\lambda(y_i|\mathbf{Z}_i)\left(1 - G(y_i|\mathbf{Z}_i)\right)\left(1 - y_i\lambda(y_i|\mathbf{Z}_i)\right)\mathbf{Z}_i}{\lambda(y_i|\mathbf{Z}_i)\left(1 - G(y_i|\mathbf{Z}_i)\right)}\right. \\
&\qquad \left. - \theta y_i \lambda(t|\mathbf{Z}_i)(1 - G(y_i|\mathbf{Z}_i))\mathbf{Z}_i\right] \\
&= \sum_{i=1}^{n}\left[\delta_i\left(1 - y_i\lambda(y_i|\mathbf{Z}_i)\right) - \theta y_i \lambda(y_i|\mathbf{Z}_i)(1 - G(y_i|\mathbf{Z}_i))\right]\mathbf{Z}_i, \\
\frac{\partial l(\theta, \boldsymbol{\gamma}, \lambda_B)}{\partial \lambda_B} &= \sum_{i=1}^{n}\left(\delta_i \frac{\frac{\partial g(y_i|\mathbf{Z}_i)}{\partial \lambda_B}}{g(y_i|\mathbf{Z}_i)} - \theta\frac{\partial G(y_i|\mathbf{Z}_i)}{\partial \lambda_B}\right) \\
&= \sum_{i=1}^{n}\left[\delta_i \frac{\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\left(1 - G(y_i|\mathbf{Z}_i)\right)\left(1 - y_i\lambda(y_i|\mathbf{Z}_i)\right)}{\lambda(y_i|\mathbf{Z}_i)\left(1 - G(y_i|\mathbf{Z}_i)\right)}\right. \\
&\qquad \left. - \theta y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)(1 - G(y_i|\mathbf{Z}_i))\right] \\
&= \sum_{i=1}^{n}\left[\frac{\delta_i}{\lambda_B}\left(1 - y_i\lambda(y_i|\mathbf{Z}_i)\right) - \theta y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)(1 - G(y_i|\mathbf{Z}_i))\right].
\end{aligned}$$

Označíme-li $l_i(\theta, \boldsymbol{\gamma}, \lambda_B)$ příspěvek i-tého pacienta do logaritmické věrohodnosti, potom složky Fisherovy a pozorované informační matice jsou dány násle-

dujícími výrazy:

$$
\begin{aligned}
\frac{\partial^2 l_i(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta^2} &= -\frac{\delta_i}{\theta^2},\\
\frac{\partial^2 l_i(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta\partial\boldsymbol{\gamma}} &= -\frac{\partial G(y_i|\mathbf{Z}_i)}{\partial\boldsymbol{\gamma}} = -y_i\lambda(y_i|\mathbf{Z}_i)(1-G(y_i|\mathbf{Z}_i))\mathbf{Z}_i,\\
\frac{\partial^2 l_i(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta\partial\lambda_B} &= -\frac{\partial G(y_i|\mathbf{Z}_i)}{\partial\lambda_B} = -y_i\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)(1-G(y_i|\mathbf{Z}_i)),\\
\frac{\partial^2 l_i(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\gamma^2} &= \Big\{ -\delta_i y_i\lambda(y_i|\mathbf{Z}_i) - \theta y_i\Big[\lambda(y_i|\mathbf{Z}_i)\,(1-G(y_i|\mathbf{Z}_i))\\
&\quad - \lambda(y_i|\mathbf{Z}_i)^2 y_i\,(1-G(y_i|\mathbf{Z}_i))\,\Big]\Big\}\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T}\\
&= \left[-\delta_i y_i\lambda(y_i|\mathbf{Z}_i) - \theta y_i\lambda(y_i|\mathbf{Z}_i)\,(1-G(y_i|\mathbf{Z}_i))\,(1-y_i\lambda(y_i|\mathbf{Z}_i))\right]\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T}\\
&= -\,y_i\lambda(y_i|\mathbf{Z}_i)\left[\delta_i + \theta\,(1-G(y_i|\mathbf{Z}_i))\,(1-y_i\lambda(y_i|\mathbf{Z}_i))\right]\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T},\\
\frac{\partial^2 l_i(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B^2} &= -\frac{\delta_i}{\lambda_B} + \theta y_i^2\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)^2\,(1-G(y_i|\mathbf{Z}_i))\,,\\
\frac{\partial^2 l_i(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}\partial\lambda_B} &= -\,\delta_i y_i\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i - \theta y_i\Big[\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\,(1-G(y_i|\mathbf{Z}_i))\,\mathbf{Z}_i\\
&\quad - \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)y_i\lambda(y_i|\mathbf{Z}_i)\,(1-G(y_i|\mathbf{Z}_i))\,\mathbf{Z}_i\Big]\\
&= -\,y_i\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\Big[\delta_i + \theta\,(1-G(y_i|\mathbf{Z}_i))\,(1-y_i\lambda(y_i|\mathbf{Z}_i))\,\Big]\mathbf{Z}_i.
\end{aligned}
$$

Ze tvaru druhých parciálních derivací usuzujeme, že pozorovaná informační matice nemusí být obecně pozitivně definitní.

Nyní do modelu (1.2) zavedeme latentní počty karcinogenních buněk, jak navrhují Chen a kol. (1999). Potom logaritmickou věrohodnostní funkci lze zapsat ve tvaru

$$
\begin{aligned}
l^N(\theta,\boldsymbol{\gamma},\lambda_B) &= \sum_{i=1}^{n}\Big[N_i\log\bar{G}(y_i|\mathbf{Z}_i) + \delta_i\,(\log N_i + \log\lambda(y_i|\mathbf{Z}_i))\Big]\\
&\quad + \sum_{i=1}^{n}\Big(N_i\log\theta - \log(N_i!) - \theta\Big)\\
&= \sum_{i=1}^{n}\Big[-N_i y_i\lambda_B\exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i) + \delta_i\Big(\log N_i + \log\lambda_B + \boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i\Big)\Big]\\
&\quad + \sum_{i=1}^{n}\Big(N_i\log\theta - \log(N_i!) - \theta\Big). \qquad (3.20)
\end{aligned}
$$

Díky zavedení latentních proměnných N_i lze logaritmickou věrohodnostní funkci (3.20) zapsat jako součet dvou sum, kde první nezávisí na parametrech $\boldsymbol{\beta}$ a druhá na parametrech $\boldsymbol{\gamma}$ a λ_B. Druhá ze sum je shodná s druhou sumou v logaritmické věrohodnosti (2.17), a tedy skórová funkce, věrohodnostní rovnice a příslušné členy ve Fisherově a pozorované informační matici budou stejné a nebudeme je znovu uvádět.

Nyní odvodíme odhady parametrů $\boldsymbol{\gamma}$ a λ_B metodou maximální věrohodnosti. Tato podrobná odvození nejsou v literatuře obsažena. Složky skórové funkce pro

parametry $\boldsymbol{\gamma}$ a λ_B jsou

$$\begin{aligned}\frac{\partial l^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}} &= \sum_{i=1}^{n}\left(-N_i y_i \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i + \delta_i \mathbf{Z}_i\right)\\ &= \sum_{i=1}^{n}\left(\delta_i - N_i y_i \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right)\mathbf{Z}_i,\\ \frac{\partial l^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B} &= \sum_{i=1}^{n}\left(\frac{\delta_i}{\lambda_B} - N_i y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\right).\end{aligned}$$

Pří známých parametrech $\boldsymbol{\gamma}$ je odhad parametru základního rizika dán vztahem

$$\hat{\lambda}_B = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} N_i y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)}.$$

Označíme-li $l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)$ příspěvek i-tého pacienta do logaritmické věrohodnosti (3.20), potom jednotlivé prvky Fisherovy a pozorované informační matice dostaneme pomocí druhých parciálních derivací, tedy

$$\begin{aligned}\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta^2} &= -\frac{N_i}{\theta^2},\\ \frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}^2} &= -N_i y_i \lambda_B \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^\mathsf{T},\\ \frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B^2} &= -\frac{\delta_i}{\lambda_B^2},\\ \frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}\partial\lambda_B} &= -N_i y_i \exp(\boldsymbol{\gamma}^\mathsf{T}\mathbf{Z}_i)\mathbf{Z}_i.\end{aligned}$$

Navíc platí $\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta\partial\boldsymbol{\gamma}} = \frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta\partial\lambda_B} = 0$, a tedy pozorovaná informační matice $I_n(\theta,\gamma,\lambda_B)$ je bloková matice s dvěma nenulovými submaticemi:

$$\begin{aligned}I_{1n}(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\theta^2} = \sum_{i=1}^{n}\frac{N_i}{\theta^2},\\ I_{2n}(\boldsymbol{\gamma},\lambda_B) &= -\frac{1}{n}\begin{pmatrix}\sum_{i=1}^{n}\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}^2} & \sum_{i=1}^{n}\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}\partial\lambda_B}\\ \sum_{i=1}^{n}\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\boldsymbol{\gamma}\partial\lambda_B} & \sum_{i=1}^{n}\frac{\partial^2 l_i^N(\theta,\boldsymbol{\gamma},\lambda_B)}{\partial\lambda_B^2}\end{pmatrix}.\end{aligned}$$

Jsou-li obě submatice $I_{1n}(\theta)$ a $I_{2n}(\boldsymbol{\gamma},\lambda_B)$ pozitivně definitní, potom i pozorovaná informační matice $I_n(\theta,\gamma,\lambda_B)$ je pozitivně definitní. V sekci 2.2 jsme ukázali, že $\sum_{i=1}^{n}\frac{N_i}{\theta^2} > 0$, pokud aspoň jeden z pacientů má aspoň jednu karcinogenní buňku, a v takovém případě je submatice $I_{1n}(\theta)$ pozitivně definitní. Z tvaru druhých parciálních derivací usuzujeme, že submatice $I_{2n}(\boldsymbol{\gamma},\lambda_B)$ není vždy pozitivně definitní, tudíž ani pozorovaná informační matice $I_n(\theta,\gamma,\lambda_B)$ nemusí být vždy pozitivně definitní.

Kapitola 4

Simulační studie

V této kapitole se zaměříme na porovnání standardního modelu (1.1) a bio modelu (1.2) pomocí simulací. Nejdříve vygenerujeme data na základě předem určených parametrů a následně tyto parametry odhadneme pomocí obou modelů a EM algoritmu. Všechny výpočty jsou provedeny ve statistickém programu R , verze 3.2.3 (R Development Core Team, 2015).

Uvažujeme-li data bez vlivu regresorů, pak pro pevně daný počet pozorování n vygenerujeme latentní indikátor uzdravení c_i z alternativního rozdělení s pevně danou pravděpodobností p. Pro standardní model generujeme čas události (relapsu) nevyléčených pacientů z exponenciálního rozdělení se zvoleným parametrem λ. Pro bio model pak předpokládáme, že inkubační doba karcinogenních buněk se řídí exponenciálním rozdělením s parametrem λ a funkce přežití nevyléčených pacientů pro čas události je pak dána vztahem (1.3), tedy pro distribuční funkci platí

$$F_0(t) = 1 - \bar{F}_0(t) = 1 - \frac{\exp\left(-\log(p)\left(1-\exp(-\lambda t)\right)\right) - p}{1-p}.$$

Čas události pak generujeme pomocí metody inverzní transformace, kde inverze distribuční funkce je

$$F_0^{-1}(t) = -\frac{\log\left(1 + \frac{\log((1-t)(1-p)+p))}{\log p}\right)}{\lambda}.$$

Čas události vyléčených pacientů položíme roven nekonečnu. Čas censorování generujeme z rovnoměrného rozdělení tak, aby podíl zcensorovaných pozorování nebyl příliš velký. Pozorovaný čas je dán jako minimum času události a času censorování. Nastala-li událost před censorováním, pak identifikátor události nabývá hodnoty 1, jinak je roven 0.

Uvažujeme-li data s vlivem vysvětlujících proměnných na pravděpodobnost vyléčení, pak postupujeme následovně. U prvního regresoru předpokládáme, že pochází z normálního rozdělení s nulovou střední hodnotou a jednotkovým rozptylem. Druhý regresor je diskrétní a nabývá hodnot 0 a 1. Je generován z alternativního rozdělení s pravděpodobností 0,2, pokud hodnota normálního regresoru je vyšší než hodnota 70. kvantilu, a s pravděpodobností 0,8 v opačném případě. Pro dané regresní parametry $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ a regresory vypočteme pravděpodobnost vyléčení i-tého pacienta p_i pomocí vztahu (3.1), tedy $p_i = \frac{\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)}{1+\exp(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)}$.

Indikátor vyléčení i-tého pacienta generujeme z alternativního rozdělení s pravděpodobností p_i. Při generování časů událostí a censorování a identifikátorů událostí postupujeme stejně jako pro data bez vlivu regresorů.

Tímto způsobem získáme sady dat, na nichž odhadneme regresní koeficienty pomocí obou modelů a EM algoritmu, kde počáteční hodnoty parametrů volíme následovně: pro parametr λ převrácenou hodnotu průměru časů pacientů, u kterých došlo k události; pro pravděpodobnost vyléčení p průměrný počet událostí vydělený dvěma; pro parametr θ pak mínus logaritmus této hodnoty; pro regresní koeficienty regresorů volíme logit průměrného počtu událostí pro standardní model a logaritmus mínus logaritmu průměrného počtu událostí pro bio model. Pro všechny regresní koeficienty uvažujeme stejnou počáteční hodnotu.

Tento proces budeme opakovat 1 000 krát. Potom spočítáme průměr odhadů regresních koeficientů a empirický rozptyl odhadů, respektive jeho odmocninu - empirickou směrodatnou odchylku. Také vypočítáme vychýlení průměrů od skutečných hodnot.

V simulacích budeme uvažovat několik scénářů, které se budou lišit v počtu uvažovaných regresorů. Pro každý scénář budeme uvažovat sadu pevně daných regresních koeficientů a budeme porovnávat, který z modelů je schopen přesněji tyto koeficienty odhadnout v závislosti na měnícím se počtu pozorování. Tabulka s výsledky pro oba modely bude uvedena pro každý příklad a model zvlášť a bude obsahovat průměr odhadů parametrů, empirickou směrodatnou odchylku, vychýlení a průměrný počet iterací EM algoritmu. Pokud uvažujeme vliv regresorů na pravděpodobnost vyléčení, pak pro samotný odhad regresních koeficientů je v každé iteraci EM algoritmu použita funkce `glm` z knihovny `stats`. Pokud došlo k překročení maximálního počtu iterací EM algoritmu (10 000) nebo k nezkonvergování algoritmu pro odhadování parametrů v rámci jedné iterace EM algoritmu, pak odhad je označen za chybný a není zahrnut ve výsledcích.

4.1 Data bez vlivu regresorů

Nejdříve budeme uvažovat data bez regresorů. Parametry jsou určeny následovně: pravděpodobnost vyléčení je $p = 0{,}4$, což pro bio model odpovídá hodnotě parametru $\theta = -\log(0{,}4) \doteq 0{,}916$. Pro standardní model je čas událostí generován z exponenciálního rozdělení s parametrem $\lambda = 2$. Pro bio model předpokládáme, že inkubační doba karcinogenních buněk se řídí exponenciálním rozdělením s parametrem $\lambda = 2$. Čas censorování je pro oba modely generován z rovnoměrného rozdělení; pro standardní model na intervalu $(0, 1{,}5)$ a pro bio model na intervalu $(0, 1)$, abychom dosáhli podobného podílu zcensorovaných pozorování pro obě sady dat.

U obou modelů došlo k několika případům překročení maximálního počtu iterací EM algoritmu. S rostoucím počtem pozorování se podíl těchto událostí zmenšuje. Průměrné odhady se s rostoucím počtem pozorování zpřesňují pro oba modely (Tabulky 4.1 a 4.2). Empirické směrodatné odchylky parametru λ jsou nižší pro standardní model pro všechny uvažované počty pozorování. Odhady parametrů p a θ nelze porovnávat přímo. Je nutné nejdříve vypočítat odhady pravděpodobnosti vyléčení pro bio model pomocí transformace $p = \exp(-\theta)$. Empirické směrodatné odchylky pravděpodobnosti vyléčení pro standardní model jsou v tomto případě opět menší než pro bio model pro všechny uvažované počty

pozorování (Tabulky 4.1 a 4.2). U bio modelu můžeme navíc pozorovat výrazně vyšší průměrný počet iterací EM algoritmu (Tabulka 4.2).

Počet pozorování	Parametr	Průměr odhadů	Empirická směrodatná odchylka	Vychýlení	Průměrný počet iterací
100	p	0,385	0,106	−0,015	170
	λ	2,048	0,571	0,048	
200	p	0,392	0,068	−0,008	113
	λ	2,009	0,397	0,009	
300	p	0,397	0,055	−0,003	105
	λ	2,011	0,329	0,011	
500	p	0,398	0,042	−0,002	97
	λ	2,016	0,251	0,016	

Tabulka 4.1: Standardní model. Výsledky simulací pro data bez vlivu regresorů. Čas událostí $\sim$ Exp($\lambda = 2$), čas cenzorování $\sim$ U(0, 1,5). Skutečná hodnota parametru $p = 0{,}4$. Průměrný podíl zcenzorovaných pozorování 0,591. Podíl nezkonvergovaných simulací je postupně 0,003, 0, 0, 0.

Počet pozorování	Parametr	Průměr odhadů	Empirická směrodatná odchylka	Vychýlení	Průměrný počet iterací
100	θ	1,012	0,381	0,096	540
	p	0,385	0,113	−0,015	
	λ	2,119	0,874	0,119	
200	θ	0,965	0,251	0,048	375
	p	0,392	0,085	−0,008	
	λ	2,067	0,648	0,067	
300	θ	0,949	0,179	0,033	303
	p	0,393	0,065	−0,007	
	λ	2,040	0,507	0,040	
500	θ	0,941	0,143	0,025	286
	p	0,394	0,052	−0,006	
	λ	1,998	0,391	−0,002	

Tabulka 4.2: Bio model. Výsledky simulací pro data bez vlivu regresorů. Čas událostí $\sim$ Exp($\lambda = 2$), čas cenzorování $\sim$ U(0, 1). Skutečná hodnota parametru $\theta = 0{,}916$. Parametr p představuje pravděpodobnost vyléčení, skutečná hodnota $p = 0{,}4$. Průměrný podíl zcenzorovaných pozorování 0,610. Výsledky pouze ze zkonvergovaných simulací. Podíl nezkonvergovaných simulací je postupně 0,026, 0,001, 0, 0.

4.2 Data s vlivem regresorů na pravděpodobnost vyléčení

4.2.1 Jeden regresor

Nyní budeme předpokládat, že na pravděpodobnost vyléčení má vliv jeden regresor, který je generován z normálního rozdělení s nulovou střední hodnotou a jednotkovým rozptylem. Pro standardní model generujeme čas událostí z exponenciálního rozdělením s parametrem $\lambda = 2$. Pro bio model předpokládáme, že inkubační doba karcinogenních buněk pochází také z exponenciálního rozdělení s parametrem $\lambda = 2$. Čas censorování pro obě sady volíme z rovnoměrného rozdělení, pro standardní model na intervalu $(0, 1{,}5)$ a pro bio model na intervalu $(0,\ 1)$, abychom dosáhli srovnatelného podílu zcensorovaných pozorování.

Necháme-li záviset pravděpodobnost vyléčení na vysvětlující proměnné, pak odhady regresních koeficientů získané pomocí standardního a bio modelu nemohou být přímo porovnávány. Můžeme ovšem porovnat odhadnuté pravděpodobnosti vyléčení. Označme $\boldsymbol{\beta} = (\beta_0, \beta_1)$ regresní koeficienty standardního modelu a $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)$ regresní koeficienty v bio modelu. Potom pravděpodobnost vyléčení pacienta s nulovým normálním regresorem je ve standardním modelu dána vztahem $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ a v bio modelu pak $\exp(-\exp(\beta_0^*))$. Pro pacienta s hodnotou normálního regresoru rovnou jedné pak ve standardním modelu $\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$ a v bio modelu $\exp(-\exp(\beta_0^* + \beta_1^*))$. Chceme-li, aby tyto pravděpodobnosti byly stejné, pak pro regresní koeficienty bio modelu platí

$$\beta_0^* = \log\left(-\log\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)\right), \tag{4.1}$$

$$\beta_1^* = \log\left(-\log\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)\right) - \beta_0^*. \tag{4.2}$$

Pro standardní model volíme hodnoty regresních koeficientů pro pravděpodobnost přežití $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-1, 0{,}5)$. Pak výše popsané pravděpodobnosti vyléčení jsou přibližně 0,269 a 0,378, což pro bio model odpovídá regresním koeficientům $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*) \doteq (0{,}273, -0{,}299)$.

V případě standardního modelu došlo v některých simulacích k nezkonvergování algoritmu pro odhadování parametrů v rámci jednotlivých iterací EM algoritmu. U bio modelu došlo k několika překročením maximálního počtu iterací EM algoritmu. Výsledky simulací jsou uvedeny pouze pro zkonvergované simulace (Tabulky 4.3 a 4.4). Odhady regresních koeficientů $\boldsymbol{\beta}$ pro standardní model a pro menší počty pozorování jsou poměrně vychýlené. Empirická směrodatná odchylka odhadu pravděpodobnosti vyléčení je pro bio model menší v případě, kdy v sadě dat máme 100 pozorování (Tabulky 4.3 a 4.4). Pro vyšší počty pozorování jsou empirické směrodatné odchylky pravděpodobností vyléčení menší u standardního modelu, avšak rozdíly mezi jednotlivými modely nejsou příliš velké. Odhady parametrů λ jsou srovnatelné s výsledky pro data bez vlivu regresorů, kde, stejně jako v tomto případě, jsou odhady přesnější pro standardní model. Rozdíl v empirických směrodatných odchylkách mezi modely je opět výrazný. Průměrný počet iterací EM algoritmu je u standardního modelu srovnatelný s daty bez vlivu regresorů, avšak u bio modelu můžeme sledovat nárůst, který se značně promítá na výpočetním čase.

Počet pozorování	Parametr	Průměr odhadů	Empirická směrodatná odchylka	Vychýlení	Průměrný počet iterací
100	β_0	−1,355	1,600	−0,355	157
	β_1	0,691	0,831	0,191	
	p_0	0,251	0,105	−0,018	
	p_1	0,370	0,126	−0,008	
	λ	2,069	0,521	0,069	
200	β_0	−1,248	2,707	−0,248	140
	β_1	0,658	1,641	0,158	
	p_0	0,258	0,076	−0,011	
	p_1	0,372	0,087	−0,006	
	λ	2,021	0,359	0,021	
300	β_0	−1,090	0,459	−0,090	131
	β_1	0,545	0,312	0,045	
	p_0	0,260	0,062	−0,009	
	p_1	0,371	0,071	−0,007	
	λ	2,016	0,296	0,016	
500	β_0	−1,042	0,333	−0,042	125
	β_1	0,527	0,231	0,027	
	p_0	0,265	0,047	−0,004	
	p_1	0,376	0,053	−0,002	
	λ	2,006	0,230	0,006	

Tabulka 4.3: Výsledky simulací pro standardní model a data s 1 regresorem $\sim \mathcal{N}(0,\ 1)$ ovlivňující pravděpodobnost vyléčení. Čas událostí $\sim \text{Exp}(\lambda = 2)$, čas censorování $\sim \text{U}(0,\ 1{,}5)$. Skutečné hodnoty regresních parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-1,\ 0{,}5)$. Parametr p_0, respektive p_1, značí pravděpodobnost vyléčení pacienta s nulovou, respektive jednotkovou, hodnotou normálního regresoru. Skutečné hodnoty pravděpodobností $p_0 = 0{,}269$ a $p_1 = 0{,}378$. Výsledky pouze ze zkonvergovaných simulací. Podíl nezkonvergovaných simulací je postupně 0,013, 0,001, 0,001, 0. Průměrný podíl zcensorovaných pozorování 0,508.

Počet pozorování	Parametr	Průměr odhadů	Empirická směrodatná odchylka	Vychýlení	Průměrný počet iterací
100	β_0	0,296	0,307	0,023	656
	β_1	−0,284	0,151	0,015	
	p_0	0,267	0,098	−0,002	
	p_1	0,365	0,118	−0,012	
	λ	2,068	0,808	0,068	
200	β_0	0,287	0,240	0,014	493
	β_1	−0,273	0,104	0,026	
	p_0	0,268	0,079	−0,001	
	p_1	0,364	0,092	−0,014	
	λ	2,023	0,610	0,023	
300	β_0	0,285	0,194	0,012	445
	β_1	−0,279	0,085	0,020	
	p_0	0,267	0,065	−0,002	
	p_1	0,366	0,076	−0,012	
	λ	1,973	0,482	−0,027	
500	β_0	0,270	0,144	−0,003	377
	β_1	−0,276	0,064	0,023	
	p_0	0,271	0,049	0,002	
	p_1	0,370	0,056	−0,008	
	λ	2,006	0,381	0,006	

Tabulka 4.4: Výsledky simulací pro bio model a data s 1 regresorem $\sim \mathcal{N}(0,\, 1)$ ovlivňující pravděpodobnost vyléčení. Čas událostí $\sim \text{Exp}(\lambda = 2)$, čas censorování $\sim$ U(0, 1). Skutečné hodnoty regresních parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1) = (0{,}273, -0{,}299)$. Parametr p_0, respektive p_1, značí pravděpodobnost vyléčení pacienta s nulovou, respektive jednotkovou, hodnotou normálního regresoru. Skutečné hodnoty pravděpodobností $p_0 = 0{,}269$ a $p_1 = 0{,}378$. Podíl nezkonvergovaných simulací je postupně 0,033, 0,005, 0, 0. Průměrný podíl zcensorovaných pozorování 0,505.

4.2.2 Dva regresory

V této části budeme předpokládat, že na pravděpodobnost vyléčení má vliv nejen normální regresor, ale také binární vysvětlující veličina. Při generování dat postupujeme analogicky jako v předchozí části. Čas událostí pro standardní model a inkubační doba karcinogenních buněk pro bio model pocházejí z exponenciálního rozdělení s parametrem $\lambda = 2$. Čas censorování pak z rovnoměrného rozdělení, pro standardní model na intervalu (0, 2,5) a pro bio model na intervalu (0, 2), abychom dosáhli srovnatelného podílu zcensorovaných pozorování.

I zde se zaměříme na odhady pravděpodobnosti vyléčení a použijeme stejnou myšlenku ohledně porovnávání regresních koeficientů jednotlivých modelů na základě pevně zvolené pravděpodobnosti vyléčení. Označme $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ regresní koeficienty ve standardním modelu a $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ v bio modelu. Pak platí vztahy (4.1) a (4.2). Pro koeficient β_2^* navíc platí

$$\beta_2^* = \log\left(-\log\left(\frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}\right)\right) - \beta_0^*.$$

Pro generování pravděpodobnosti vyléčení pomocí standardního modelu volíme hodnoty regresních koeficientů $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (-1, 0{,}5, -0{,}7)$, což pro bio model odpovídá přibližně hodnotám $\boldsymbol{\beta}^* = (\beta_0, \beta_1, \beta_2) = (0{,}273, -0{,}299, 0{,}352)$. Skutečná pravděpodobnost vyléčení pro pacienta s nulovými regresory je tedy $p_0 = 0{,}269$, pro pacienta s hodnotou normálního regresoru rovnou jedné a nulovou hodnotou diskrétního regresoru pak $p_1 = 0{,}378$ a pro pacienta s nulovou hodnotou normálního regresoru a jednotkovou hodnotou diskrétního regresoru pak $p_2 = 0{,}154$.

V tomto příkladě jsou téměř všechny odhady pravděpodobností vyléčení u bio modelu přesnější. Pouze u pravděpodobnosti p_2 můžeme sledovat nepatrně vyšší empirické směrodatné odchylky pro vyšší počty pozorování (Tabulky 4.5 a 4.6). Obdobně jako v předchozích příkladech, i zde jsou odhady parametru λ přesnější pro standardní model. V porovnání s předchozími scénáři jsou empirické směrodatné odchylky nižší, což může být způsobeno zmenšením podílu zcensorovaných pozorování v obou sadách dat. U obou modelů také došlo k prudkému snížení průměrného počtu iterací.

Počet pozorování	Parametr	Průměr odhadů	Empirická směrodatná odchylka	Vychýlení	Průměrný počet iterací
100	β_0	−1,205	1,220	−0,205	52
	β_1	0,581	0,448	0,081	
	β_2	0,875	1,240	0,175	
	p_0	0,259	0,101	−0,010	
	p_1	0,369	0,112	−0,009	
	p_2	0,422	0,089	−0,004	
	λ	2,028	0,397	0,028	
200	β_0	−1,056	0,416	−0,056	43
	β_1	0,532	0,256	0,032	
	β_2	0,746	0,495	0,046	
	p_0	0,266	0,074	−0,003	
	p_1	0,375	0,076	−0,002	
	p_2	0,424	0,059	−0,001	
	λ	2,021	0,281	0,021	
300	β_0	−1,020	0,313	−0,020	41
	β_1	0,520	0,197	0,020	
	β_2	0,709	0,368	0,009	
	p_0	0,269	0,059	0,000	
	p_1	0,380	0,060	0,002	
	p_2	0,424	0,048	−0,002	
	λ	2,019	0,219	0,019	
500	β_0	−1,000	0,225	−0,000	40
	β_1	0,504	0,141	0,004	
	β_2	0,699	0,273	−0,001	
	p_0	0,271	0,044	0,002	
	p_1	0,380	0,046	0,002	
	p_2	0,426	0,037	0,000	
	λ	2,007	0,166	0,007	

Tabulka 4.5: Výsledky simulací pro standardní model a data se 2 regresory - spojitý $\sim \mathcal{N}(0,\, 1)$ a diskrétní, ovlivňující pravděpodobnost vyléčení. Čas událostí $\sim$ Exp($\lambda = 2$), čas censorování $\sim$ U(0, 2,5). Skutečné hodnoty regresních parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (-1,\ 0{,}5,\ -0{,}7)$. Parametr p_0, respektive p_1 a p_2, značí pravděpodobnost vyléčení pacienta s nulovou hodnotou obou regresorů, respektive jednotkovou hodnotou normálního regresoru a nulovou hodnotou diskrétního, respektive nulovou hodnotou normálního a jednotkovou diskrétního regresoru. Skutečné hodnoty pravděpodobností $p_0 = 0{,}269$, $p_1 = 0{,}378$, $p_2 = 0{,}154$. Podíl nezkonvergovaných simulací je roven 0 pro všechny počty pozorování. Průměrný podíl zcensorovaných pozorování 0,495.

Počet pozorování	Parametr	Průměr odhadů	Empirická směrodatná odchylka	Vychýlení	Průměrný počet iterací
100	β_0	0,276	0,275	0,003	125
	β_1	−0,311	0,163	−0,012	
	β_2	−0,427	0,330	0,003	
	p_0	0,272	0,092	0,003	
	p_1	0,381	0,097	0,003	
	p_2	0,422	0,083	−0,004	
	λ	2,016	0,530	0,016	
200	β_0	0,272	0,194	−0,000	109
	β_1	−0,303	0,109	−0,004	
	β_2	−0,417	0,234	0,013	
	p_0	0,271	0,067	0,002	
	p_1	0,379	0,069	0,001	
	p_2	0,420	0,059	−0,005	
	λ	2,002	0,361	0,002	
300	β_0	0,260	0,151	−0,013	102
	β_1	−0,300	0,090	−0,001	
	β_2	−0,415	0,190	0,015	
	p_0	0,275	0,053	0,006	
	p_1	0,382	0,056	0,005	
	p_2	0,424	0,049	−0,001	
	λ	2,007	0,296	0,007	
500	β_0	0,254	0,116	−0,019	98
	β_1	−0,298	0,068	0,001	
	β_2	−0,411	0,143	0,019	
	p_0	0,276	0,041	0,007	
	p_1	0,384	0,045	0,007	
	p_2	0,425	0,038	−0,000	
	λ	2,017	0,233	0,017	

Tabulka 4.6: Výsledky simulací pro bio model a data se 2 regresory - spojitý $\sim \mathcal{N}(0,\, 1)$ a diskrétní, ovlivňující pravděpodobnost vyléčení. Čas událostí $\sim$ Exp($\lambda = 2$), čas censorování $\sim$ U(0, 2). Skutečné hodnoty regresních parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (0{,}273,\ -0{,}299,\ 0{,}352)$. Parametr p_0, respektive p_1 a p_2, značí pravděpodobnost vyléčení pacienta s nulovou hodnotou obou regresorů, respektive jednotkovou hodnotou normálního regresoru a nulovou hodnotou diskrétního, respektive nulovou hodnotou normálního a jednotkovou diskrétního regresoru. Skutečné hodnoty pravděpodobností $p_0 = 0{,}269$, $p_1 = 0{,}378$, $p_2 = 0{,}154$. Podíl nezkonvergovaných simulací je roven 0 pro všechny počty pozorování. Průměrný podíl zcensorovaných pozorování 0,485.

Závěr

V práci jsme se zabývali modely pro přežití s možností vyléčení, kde předpokládáme, že u části pacientů se v konečném čase již neobjeví symptomy zkoumané nemoci, a lze je tedy považovat za vyléčené. U nevyléčených pacientů nás pak zajímá rozdělení doby do relapsu. Obě tyto veličiny mohou záviset na regresorech, přičemž jejich vliv na vyléčení a dobu do relapsu může být rozdílný.

Hlavním cílem této práce bylo porovnat různé přístupy k formulaci, analýze a interpretaci těchto modelů. Věnovali jsme se především dvěma modelům - standardnímu, který je v současnosti nejznámější a také nejpoužívanější, a bio modelu, který byl v literatuře popsán relativně nedávno. V první kapitole jsme uvedli definice těchto modelů a jejich zobecnění. Také jsme podrobně ukázali jejich vzájemné vztahy.

V kapitole 2 jsme předpokládali, že vysvětlující veličiny nemají vliv ani na přežití, ani na vyléčení. Pro oba modely jsme odvodili odhady parametrů a jejich vlastnosti metodou maximální věrohodnosti. Do modelů jsme následně zavedli latentní veličiny a demonstrovali jsme použití EM algoritmu.

V třetí kapitole jsme předpokládali, že pravděpodobnost vyléčení a doba do relapsu mohou být ovlivněny vysvětlujícími veličinami. I zde jsme podrobně odvodili odhady parametrů a naznačili jsme implementaci EM algoritmu s latentními proměnnými.

V kapitole 4 jsme tyto dva modely porovnali pomocí simulací. Výsledky naznačují, že pokud vysvětlující veličiny mají vliv na pravděpodobnost vyléčení, pak bio model dává přesnější odhady této pravděpodobnosti. Ve všech scénářích však standardní model dává přesnější odhady parametru λ. Je důležité mít na paměti, že interpretace tohoto parametru v jednotlivých modelech je odlišná. Navíc předpokládaný tvar funkce přežití času do relapsu nevyléčených pacientů je u bio modelu složitější. Pomalejší konvergence EM algoritmu u bio modelu byla zjevná ve všech scénářích a značně se projevovala na časové výpočetní náročnosti.

V simulační studii jsme se omezili pouze na případy, kdy regresory buď nemají vliv na pravděpodobnost vyléčení ani na dobu relapsu, nebo kdy mají vliv pouze na pravděpodobnost vyléčení. Dále jsme předpokládali, že doba do relapsu u standardního modelu, respektive inkubační doba karcinogenních buněk v bio modelu, se řídí exponenciálním modelem.

Možným rozšířením této práce by mohlo být srovnání modelů, kdy vysvětlující veličiny mají vliv na dobu do relapsu, nebo na pravděpodobnost vyléčení a dobu do relapsu zároveň. Také by bylo možné porovnat modely za předpokladu složitější struktury funkce přežití, respektive inkubační doby karcinogenních buněk.

Literatura

Agresti, A. a Kateri, M. (2011). *Categorical data analysis*. Springer.

Berkson, J. a Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 15–53.

Chen, M.-H. a Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, **57**(1), 43–52.

Chen, M.-H., Ibrahim, J. G., a Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**(447), 909–919.

Farewell, V. T. (1977). The combined effect of breast cancer risk factors. *Cancer*, **40**(2), 931–936.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**(3), 257–262.

Hanin, L. G. (2001). Iterated birth and death process as a model of radiation cell survival. *Mathematical biosciences*, **169**(1), 89–107.

Ibrahim, J. G., Chen, M.-H., a Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, **57**(2), 383–388.

Kalbfleisch, J. D. a Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kuk, A. Y. a Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**(3), 531–541.

Peng, Y. a Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**(1), 237–243.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Sy, J. P. a Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, **56**(1), 227–236.

Tsodikov, A. (2002). Semi-parametric models of long-and short-term survival: an application to the analysis of breast cancer survival in utah by age and stage. *Statistics in medicine*, **21**(6), 895–920.

Yin, G. a Ibrahim, J. G. (2005). Cure rate models: a unified approach. *Canadian Journal of Statistics*, **33**(4), 559–570.

Seznam tabulek