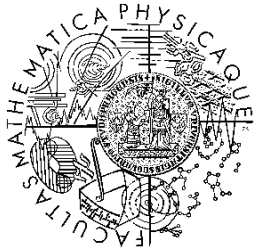


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCA

Bc. Kristína Mečiarová

Dynamická predikcia v analýze prežitia

Katedra pravděpodobnosti a matematické statistiky

Vedúci diplomovej práce:	doc. RNDr. Arnošt Komárek, Ph.D.
Študijný program:	Pravděpodobnost, matematická statistika a ekonometrie
Študijný odbor:	Pravděpodobnost, matematická statistika a ekonometrie

Praha 2023

Prehlasujem, že som túto diplomovú prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov. Táto práca nebola využitá k získaniu iného alebo rovnakého titulu.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova má právo na uzavretie licenčnej zmluvy o použití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V ..... dňa ....

Podpis autora

Týmto by som chcela poďakovať môjmu vedúcemu diplomovej práce, doc. RNDr. Arnoštovi Komárkovi, Ph.D. za trpezlivosť, ochotu, cenné rady a čas, ktorý mi venoval počas písania diplomovej práce. Veľká vďaka patrí taktiež mojej rodine, priateľom a snúbencovi, ktorí mi boli oporou počas celého štúdia.

Názov práce: Dynamická predikcia v analýze prežitia

Autor: Bc. Kristína Mečiarová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci diplomovej práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Častou motiváciou na budovanie štatistického modelu je predikcia výsledkov. V kontexte analýzy prežitia je dôležité rozlišovať dva druhy časovo premenlivých prediktorov a starostlivo zvážiť voľbu prežívacieho modelu. Združený model pre longitudinálne a cenzorované dáta umožňuje, oproti štandardnému Coxovmu modelu, zohľadniť spojitý vývoj longitudinálnej premennej v čase v modeli prežitia. V práci sú uvedené dva typy združených modelov, združený model so spoločnými náhodnými efektami a združený model s latentnými kategóriami. Pre prvý spomínaný typ modelu je podrobne popísané bayesovské odhadovanie parametrov a zhrnutá metodika dynamickej predikcie individuálnej pravdepodobnosti prežitia. Teoretické poznatky sú aplikované v ilustračnej analýze dát k primárnej biliárnej cirhóze. Následne je v simulačnej štúdii skúmaný vplyv počtu pacientov, počtu longitudinálnych meraní a percenta cenzorovania na kvalitu predikcií a odhady parametrov modelu.

Kľúčové slová: dynamická predikcia, pravdepodobnosť prežitia, združený model, Coxov model, lineárny zmiešaný model, bayesovské metódy

Title: Dynamic prediction in survival analysis

Author: Bc. Kristína Mečiarová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Often the motivation behind building a statistical model is to provide prediction for an outcome of interest. In the context of survival analysis it is important to distinguish between two types of time-varying covariates and take into careful consideration the appropriate type of analysis. Joint model for longitudinal and time-to-event data, in contrast to standard Cox model, enables to account for continuous change of the covariate over time in the survival model. In this thesis two examples of joint models are presented, the shared random-effect model and the joint latent class model. Bayesian estimation of the model parameters and summary of methodology for dynamic prediction of individual survival probability is provided for the first one of the aforementioned types of models. Application of the theoretical knowledge is illustrated in the analysis of the data on primary biliary cirrhosis. The impact of number of patients, number of longitudinal measurements and per-cent of censoring on the quality of prediction and estimates of the model parameters is examined in the simulation study.

Keywords: dynamic prediction, survival probability, joint model, Cox model, linear mixed effects model, Bayesian methods

Obsah

Značenie

$\mathbf{Y}$	vektor longitudinálnych meraní (marker)
$\mathbb{X}$	regresná matica fixných efektov
$\mathbb{Z}$	regresná matica náhodných efektov
C	čas cenzorovania
T^*	skutočný čas udalosti
$T = min(T^*,C)$	cenzorovaný čas do udalosti
$\delta = \mathbb{1}(T^* \leq C)$	identifikátor udalosti
$N(t) = \mathbb{1}(T^* \leq t, \delta = 1)$	čítací proces
$R(t) = \mathbb{1}(T \geq t)$	proces v riziku
$\lambda(t)$	riziková funkcia náhodnej veličiny T^*
$\Lambda(t)$	kumulatívna riziková funkcia náhodnej veličiny T^*
$F(t)$	distribučná funkcia náhodnej veličiny T^*
$S(t)$	funkcia prežitia náhodnej veličiny T^*
$\hat{\Lambda}(t)$	Nelsonov-Aalenov odhad kumulatívneho rizika
$\hat{S}(t)$	Kaplan-Meierov odhad funkcie prežitia
$\boldsymbol{\epsilon}$	chybový vektor
$\mathbf{b}$	vektor náhodných efektov
$\lvert\mathbb{D}\rvert$	determinant matice $\mathbb{D}$
$\mathrm{tr}(\mathbb{D})$	stopa matice $\mathbb{D}$
$p(\mathbf{b})$	hustota náhodného vektoru pri bayesovskom prístupe
$f(\mathbf{b})$	hustota náhodného vektoru pri frekventistickom prístupe

Úvod

Kľúčovou otázkou mnohých klinických štúdií je správna predikcia prognózy pacientov. V dnešnej dobe majú lekári k dispózícii širokú škálu pacientskych údajov a meraní. Tieto merania sú často vykonávané v pravidelných časových intervaloch, aby bolo možné lepšie sledovať vývoj stavu pacienta. Cieľom je využiť v maximálnej možnej miere zaznamenané informácie a poskytnúť medicínsky relevatné informácie, napríklad, ako je biomarker, meniaci sa v čase, spojený s rizikom konkrétnej udalosti. Možným prostriedkom na zodpovedanie tejto otázky sú združené modely pre longitudinálne dáta a dáta pre čas do udalosti. V tejto práci sa zameriame na určitý typ združených modelov a popíšeme bayesovské odhady parametrov modelu, predikcie individuálnych pravdepodobností prežitia na základe združeného modelu a miery presnosti predikcií.

V prvej kapitole zavedieme definíciu lineárneho zmiešaného modelu a stručne popíšeme možné prístupy a odhady parametrov modelu pre longitudinálne dáta. Samostatná časť práce je venovaná dátam pre čas do udalosti a cenzorovaniu, kde zavádzame základné pojmy z analýzy prežitia a Coxov model proporčných rizík. Posledná časť prvej kapitoly je venovaná úvodu do bayesovskej štatistiky, keďže tento prístup je využívaný v celej práci.

Ďalej pokročíme k odvodeniu združených modelov. Hlavná myšlienka združených modelov spočíva v použití modelu pre longitudinálne dáta v modeli pre riziko udalosti. Pre longitudinálne dáta volíme lineárny zmiešaný model a pre dáta pre čas do udalosti Coxov model proporčných rizík. Spomenieme dva typy združených modelov - združené modely so spoločnými náhodnými efektami a združené modely s latentnými kategóriami. Zameriame sa na združené modely so spoločnými náhodnými efektami, v ktorých je kľúčový predpoklad, že podmienene na náhodnom efekte pacienta, sú longitudinálna premenná a čas do udalosti nezávislé. Pre tento typ združených modelov podrobne popíšeme odhad parametrov bayesovskými metódami. V združených modeloch s latentými kategóriami je hlavnou myšlienkou, že populáciu možno rozdeliť na homogénne latentné kategórie subjektov, ktoré zdieľajú rovnaký model pre longitudinálnu premennú a rovnaké riziko udalosti. Podmienene na kategórii sú longitudinálna premená a čas do udalosti nezávislé. Tento typ modelov v práci nie je podrobnejšie rozoberaný.

Zavŕšením teoretickej časti práce je kapitola o dynamickej predikcii. Zavádzame tu pojem individuálna pravdepodobnosť prežitia a spôsob jej odhadovania pomocou bayesovskej štatistiky. Ďalej sa venujeme mieram presnosti predikcie a pojmom ako špecificita, senzitivita, ROC krivka, plocha pod ROC krivkou a dynamický index predikcie. Pre všetky spomínané kvantity sú odvodené odhady. Záver kapitoly je venovaný kalibrácii modelu, konkrétne chybe predikcie, integrovanej chybe predikcie a ich odhadom.

Nadobudnuté teoretické poznatky aplikujeme v ilustračnej analýze reálnych dát. V analýze skúmame, ako sa mení riziko transplantácie pečene alebo úmrtia na primárnu biliárnu cirhózu v závislosti na hladine bilirubínu v čase, veku, pohlaví a liečbe pacienta. Rolu longitudinálnej premennej v združenom modeli hrá hladina bilirubínu, na ktorú používame logaritmickú transformáciu. Výsledky z vybudovaného modelu, ktorými sa zaoberáme, sú odhady parametrov modelu, porovnanie predikcií podmienených pravdepodobností prežitia na základe klasic-

kého Coxovho modelu a združeného modelu a odhadnuté miery presnosti predikcie.

Nakoniec v simulačnej štúdii skúmame vplyv zvyšujúceho sa počtu pacientov, počtu návštev a percenta cenzorovania na predikcie, ich presnosť a presnosť odhadov parametrov združeného modelu.

1. Základné pojmy

Kľúčovým nástrojom pre dynamickú predikciu je použitie modelu pre longitudinálne dáta v modeli prežitia. K tomu, aby sme mohli podrobne popísať metodiku dynamickej predikcie, najskôr v tejto kapitole zavedieme potrebné základné pojmy, s ktorými budeme pracovať. Zadefinujeme lineárne zmiešaný model a zavedieme základné pojmy z analýzy prežitia. V závere kapitoly uvedieme základy bayesovskej štatistiky, ako možného prostriedku na odhadovanie parametrov.

1.1 Lineárne zmiešané modely

Predpokladajme, že máme k dispozícii K nezávislých náhodných vektorov $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$, kde $\mathbf{Y}_i = (Y_{i_1}, \ldots, Y_{i_{n_i}})^\top$, $i = 1, \ldots, K$. Dáta pozostávajú z K nezávislých skupín (subjektov), ktoré zahŕňajú rôzny počet korelovaných pozorovaní. V rámci každej skupiny sú pozorovania závislé, ale medzi skupinami sú nezávislé. Medzi takéto skupinovo závislé dáta patria aj longitudinálne dáta. Jedná sa o opakované merania so zaznamenaným časom merania. Z toho plynie, že štandardné štatistické nástroje, ako t-test a obyčajná lineárna regresia, ktoré predpokladajú nezávislosť pozorovaní, nie sú vhodné pre analýzu longitudinálnych dát. Priamočiarym prístupom, ako modelovať korelované dáta, je mnohorozmerná regresia. Jednou z metód, na odhadovanie regresných parametrov pre skupinovo závislé dáta, sú zovšeobecnené odhadovacie rovnice, ktoré vyžadujú voľbu pracovnej korelačnej štruktúry. Alternatívny, a možno intuitívnejší prístup, je uvažovať, že každý subjekt v populácii má svoj vlastný špecifický očakávaný vývoj odozvy v čase. Subjekty sú volené náhodne z populácie, teda regresné koeficienty špecifické pre jednotlivé subjekty, sú tiež náhodne vybraté z „populácie regresných koeficientov“. Jedným z modelov, založených na tomto princípe, je lineárny zmiešaný model, podrobne zadefnovaný v nasledujúcej definícii podľa [Laird and Ware, 1982].

Definícia 1. *Vektory $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$ spĺňajú lineárny zmiešaný model, ak sú nezávislé a môžu byť rozpísané ako*

$$\mathbf{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, K,$$

kde $\boldsymbol{\beta}$ je p-rozmerný vektor fixných efektov, $\mathbb{X}_i = (\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,p})$ je $n_i \times p$ regresná matica pre fixné efekty, $\mathbb{Z}_i$ je $n_i \times q$ regresná matica pre náhodné efekty, $\mathbf{b}_i$ sú nezávislé vektory náhodných efektov, ktoré spĺňajú

$$\mathbf{b}_i \sim \mathsf{N}_q(\mathbf{0}, \mathbb{D}),$$

$\boldsymbol{\epsilon}_i$ sú nezávislé vektory chybových členov, spĺňajúce

$$\boldsymbol{\epsilon}_i \sim \mathsf{N}_{n_i}(\mathbf{0}, \sigma^2\mathbb{I}_{n_i}),$$

a chybové členy $(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_K)^\top$ sú nezávislé od náhodných efektov $(\mathbf{b}_1, \ldots, \mathbf{b}_K)^\top$.

Interpretácia koeficientov z lineárne zmiešaného modelu je následne jednoznačná:

- β_j, $j \in \{1, \ldots, p\}$, značí zmenu $\mathbf{Y}_i$ priemerne v populácii, keď sa $\mathbf{X}_{i,j}$ zväčší o jednu jednotku,
- $\mathbf{b}_i$ interpretujeme v zmysle, ako sa podmnožina regresných parametrov pre i-ty subjekt líši od populačných parametrov.

Výhodou lineárne zmiešaného modelu sú populačné a pre subjekt špecifické predikcie:

- $\boldsymbol{\beta}$ popisuje zmeny odozvy priemerne v populácii,
- $\boldsymbol{\beta} + \mathbf{b}_i$ popisuje individuálne trajektórie odozvy.

Model z definície 1 je špeciálnym prípadom marginálnej formy lineárne zmiešaného modelu

$$\mathbf{Y}_i \sim \mathsf{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta},\sigma^2\boldsymbol{\Sigma}_i),\ i = 1,\ldots,K, \tag{1.1}$$

kde $\boldsymbol{\Sigma}_i = \frac{1}{\sigma^2}\mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\top + \mathbb{I}_{n_i}$ je pozitívne definitná matica. Nie každý model tvaru (1.1) nutne spĺňa definíciu 1. Môže sa stať, že matica $\boldsymbol{\Sigma}_i$ je pozitívne definitná, ale matica $\mathbb{D}$ nie je. Marginálnu formu možno zapísať združene ako

$$\mathbf{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta},\sigma^2\boldsymbol{\Sigma}),$$

kde $n = \sum_{i=1}^K n_i$, $\mathbf{Y}^\top = (\mathbf{Y}_1^\top,\ldots,\mathbf{Y}_K^\top)^\top$, $\mathbb{X}^\top = (\mathbb{X}_1^\top,\ldots,\mathbb{X}_K^\top)^\top$ a $\boldsymbol{\Sigma}$ je blokovo diagonálna matica s diagonálnymi blokmi $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K$.
Model z defincície 1 sa nazýva aj jednoúrovňový lineárny zmiešaný model. Pridaním ďalších skupinových faktorov možno jednoúrovňový lineárny zmiešaný model zovšeobecniť na viacúrovňový model. V práci sa budeme zaoberať iba jednoúrovňovými lineárnymi zmiešanými modelmi.
Neznáme parametre v modeli podľa definície 1 sú $\boldsymbol{\beta}$, σ^2 a variančná matica $\mathbb{D}$. Jedným z možných odhadov neznámych parametrov je pomocou maximálnej vierohodnosti, resp. profilovej vierohodnosti. Tvar tejto profilovej vierohodnosti je však príliš komplikovaný a analytické výpočty skórovej štatistiky a informačnej matice sú možné iba v špeciálnych jednoduchých prípadoch. V praxi často používaná metóda, ktorá umožnuje odhadovať súčasne regresné parametre $\boldsymbol{\beta}$ a predikovať latentné premenné $\mathbf{b}_i$, sa nazýva Hendersonove rovnice pre zmiešaný model [Henderson, 1984]. Je založená na počítaní vážených najmenších štvorcov pre model z definície 1, zapísaný v tvare súčasne pre všetky pozorovania. Za predpokladu, že regresné matice $\mathbb{X}$ a $\mathbb{Z}$ majú plnú stĺpcovú hodnosť a parametre σ^2 a $\mathbb{D}$ sú známe, dostaneme najlepší lineárny nestranný odhad parametra $\boldsymbol{\beta}$ tvaru

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top\boldsymbol{\Sigma}^{-1}\mathbb{X})^{-1}(\mathbb{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{Y}),$$

a najlepšiu lineárnu nestrannú predikciu premennej $\mathbf{b}^\top = (\mathbf{b}_1^\top,\ldots,\mathbf{b}_K^\top)^\top$ tvaru

$$\hat{\mathbf{b}} = \mathbb{D}_*\mathbb{Z}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}),$$

kde $\mathbb{D}_*$ je blokovo diagonálna matica s K blokmi $\mathbb{D}$. V prípade, že parametre σ^2 a variančná matica $\mathbb{D}$ nie sú známe, použije sa na ich odhad metóda maximálnej

vierohodnosti alebo metóda obmedzenej maximálnej vierohodnosti (angl. *restricted maximum likelihood – REML*) [Patterson and Thompson, 1971], [Harville, 1974]. Rozdiel medzi REML a klasickou vierohodnosťou je asymptoticky zanedbateľný (pri $K \to \infty$). Odhady variančných parametrov sa následne dosadia do odhadov $\hat{\boldsymbol{\beta}}$ a $\hat{\mathbf{b}}$. Viac detailov možno nájsť v [Jennrich and Schluchter, 1986] a [Lindstrom and Bates, 1988].

1.2 Analýza cenzorovaných dát

Štandardné štatistické metódy na odhadovanie a testovanie by mohli byť použité na analýzu dát pre čas do udalosti, ak by bol čas udalosti pozorovaný u každého subjektu. Zvyčajne však čas udalosti nie je pozorovaný z rôznych dôvodov- dĺžka štúdie, finančné náklady, choroba sa u pacienta neprejaví/nedôjde k úmrtiu pacienta. Dáta pre čas do udalosti preto vyžadujú špecializované metódy, ktoré popíšeme v tejto časti.

1.2.1 Cenzorované dáta

Označme $T^* \geq 0$ nezápornú náhodnú veličinu, ktorá predstavuje čas do udalosti (alebo inak aj čas zlyhania, doba prežitia) a $C \geq 0$ čas cenzorovania, ktorý predstavuje dobu pozorovania subjektu. Podľa polohy skutočnej udalosti na časovej osi vzhľadom k cenzorovaniu, rozlišujeme medzi troma druhmi cenzorovania: cenzorovanie sprava ($T^* > C$), zľava ($T^* < C$) a intervalové cenzorovanie, ktoré je dané dvoma cenzorovacími premennými $0 < C_L < C_U < \infty$, $T^* \in (C_L; C_U)$. Z hľadiska pravdepodobnostných vzťahov medzi skutočným časom udalosti a časom cenzorovania, rozlišujeme medzi informatívnym a neinformatívnym cenzorovaním. Ak pravdepodobnosť cenzorovania závisí na procese zlyhania, jedná sa o informatívne cenzorovanie. Tento mechanizmus je podobný nenáhodnému mechanizmu chýbania dát (angl. *missing not at random – MNAR*). V prípade, že ukončenie sledovania subjektu nezávisí na procese zlyhania, jedná sa o neinformatívne cenzorovanie, ktoré odpovedá náhodnému mechanizmu chýbania dát (angl. *missing at random – MAR*). V tejto práci budeme pracovať s neinformatívnym cenzorovaním sprava. Ak dôjde k ukončeniu pozorovania skôr, než nastane udalosť, čas udalosti nemáme k dispozícii. Preto zavádzame nasledujúce značenie. Náhodnú veličinu $T = \min(T^*,C)$ nazveme cenzorovaný čas do udalosti a $\delta = \mathbb{1}(T^* \leq C)$ indikátor zlyhania. Majme teda latentné premenné pre čas do udalosti a čas cenzorovania $(T_1^*,C_1), \ldots, (T_n^*,C_n)$ generované z n nezávislých subjektov. Pre daný náhodný výber sme schopní pozorovať iba $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ a na základe týchto pozorovaní robiť inferenciu o T_i^*, pre každé $i \in \{1, \ldots, n\}$.

Cenzorovacie premenné $C_1, \ldots, C_n$ sú náhodné veličiny z nejakého rozdelenia (každá C_i môže mať iné rozdelenie). Tento model sa nazýva model náhodného cenzorovania, ktorý má dva špeciálne prípady:

- **Cenzorovanie I. typu** (cenzorovanie časom) Všetky cenzorovacie premenné sú rovné predom špecifikovanej konštante τ, ktorá vyjadruje spoločnú maximálnu dobu pozorovania, t.j. $C_i = \tau$ pre každé $i = 1, \ldots, n$ skoro iste.

- **Cenzorovanie II. typu** (cenzorovanie poruchou) Všetky zostávajúce pozorovania sú cenzorované, keď nastane k-te zlyhanie ($k \in \{1, \ldots, n\}$ je predom zadaná), t.j., $C_i = T^*_{(k)}$ pre každé $i = 1, \ldots, n$, $T^*_{(k)}$ je k-ta poriadková štatistika z náhodného výberu $T^*_1, \ldots, T^*_n$.

Tieto dve schémy cenzorovania sú používané predovšetkým v technických aplikáciách. Pre väčšinu aplikácií v skutočnom živote sú nereálne. Zvyčajne pacienti vstupujú do štúdie priebežne a k cenzorovaniu dôjde ukončením štúdie v istom časovom okamihu. Cenzorované časy jednotlivých pacientov sú teda individuálne. Ďalej zadefinujeme funkcie, ktoré sú kľúčovým nástrojom na prácu s cenzorovanými dátami.

Značenie. *Pre sprava spojitú funkciu f zavedieme nasledujúce značenie: $f(t-) = \lim_{h \searrow 0} f(t-h)$ (ak limita na pravej strane existuje). Táto funkcia je zľava spojitá.*

Definícia 2. *Funkcia $S(t) = 1 - F(t) = \mathsf{P}(T^* > t)$ sa nazýva funkcia prežitia náhodnej veličiny T^* s distribučnou funkciou $F(t)$.*

Funkcia $S(t)$ je nerastúca sprava spojitá, $S(0) = 1 - \mathsf{P}(T^* = 0)$, $\lim_{t\to\infty} S(t) = 0$. Ak je T^* spojitá s hustotou $f(t)$ vzhľadom k Lebesgueovej miere, potom $S(t) = \int_t^\infty f(s)\, ds$ a ak je T^* diskrétna s hodnotami $t_1, t_2, \ldots$ a $p_i = P(T^* = t_i)$, potom $p_i = S(t-) - S(t)$ a $S(t) = \sum_{\{i:t_i>t\}} p_i$. Zvyčajne predpokladáme $\mathsf{P}(T = 0) = 0$ (zlyhanie nemôže nastať v čase 0). Potom platí $S(0) = 1$. Funkcia prežitia jednoznačne určuje rozdelenie náhodnej veličiny T^*. Ďalšia funkcia, ktorá tiež určuje rozdelenie náhodnej veličiny jednoznačne, je riziková funkcia.

Definícia 3. *Nech T^* je spojitá nezáporná náhodná veličina. Potom riziková funkcia $\lambda(t)$ veličiny T^* je definovaná ako*

$$\lambda(t) = \lim_{h \searrow 0} \frac{1}{h} \mathsf{P}(t \le T^* < t + h | T^* \ge t).$$

Nech T^* je diskrétna náhodná veličina s hodnotami $0 \le t_1 < t_2 < \ldots$. Potom riziková funkcia $\lambda(t)$ veličiny T^* je definovaná v t_1,$t_2, \ldots$ ako

$$\lambda(t_i) \equiv \lambda_i = \mathsf{P}(T^* = t_i | T^* \ge t_i).$$

Inak povedané, riziková funkcia meria pravdepodobnosť, že nastane udalosť v čase t za podmienky, že nenastala skôr. Vyjadruje teda riziko, že v čase t dôjde k udalosti. Ďalej definujeme kumulatívnu rizikovú funkciu, ktorá vyjadruje kumulované riziko do času t.

Definícia 4. *Funkcia $\Lambda(t)$ definovaná ako*

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

pre spojité T^* a

$$\Lambda(t) = \sum_{\{i:t_i \le t\}} \lambda(t_i)$$

pre disktrétne T^*, sa nazýva kumulatívna riziková funkcia.

Predpokladajme, že náhodné veličiny T^* a C sú nezávislé a pozorujeme iba náhodnú veličinu $T = \min(T^*,C)$. Pre funkciu prežitia náhodnej veličiny T platí:

$$S_T(t) = \mathsf{P}(T > t) = \mathsf{P}(T^* > t,C > t) = S(t)\,\mathsf{P}(C > t) \leq S(t),$$

kde $S(t)$ je funkcia prežitia náhodnej veličiny T^*. Pokiaľ nepoznáme rozdelenie náhodnej veličiny C, tak $S(t)$ nevieme odvodiť. Predpokladajme ďalej, že T^* má spojité rozdelenie s rizikovou funkciou $\lambda(t)$

$$\begin{aligned}\lambda(t) &= \lim_{h\searrow 0}\frac{1}{h}\,\mathsf{P}(t \leq T^* < t+h|T^* \geq t) = \lim_{h\searrow 0}\frac{1}{h}\,\mathsf{P}(t \leq T^* < t+h|T^* \geq t, C > t)\\ &= \lim_{h\searrow 0}\frac{1}{h}\,\mathsf{P}(t \leq T^* < t+h|T \geq t),\end{aligned}$$

kde druhá rovnosť platí z nezávislosti T^* a C. Za určitých podmienok, teda vieme z cenzorovaných dát získať rizikovú funkciu T^*, a preto je riziková funkcia vhodným nástrojom na analýzu cenzorovaných dát. Stochastická nezávislosť medzi T^* a C je postačujúca, ale nie nutná podmienka pre vyššie uvedenú rovnosť. Zadefinujeme preto podmienku nezávislého cenzorovania.

Definícia 5. *Cenzorovacia premenná C spĺňa podmienku nezávislého cenzorovania pre čas zlyhania T^* s kumulatívnou rizikovou funkciou Λ práve vtedy, keď*

$$\Lambda(t) = -\int_0^t \frac{d\,P[T^* \geq s,C \geq T^*]}{P[T^* \geq s,C \geq s]},\ \forall t \text{ spĺňajúce } P[T^* \geq t,C \geq t] > 0.$$

Poznámka. Pre spojitú náhodnú veličinu T^* je podmienka z definície 5 ekvivalentná rovnosti

$$\begin{aligned}\lambda(t) &= \frac{\frac{-\partial}{\partial s}\,\mathsf{P}[T^* \geq s,C \geq t]|_{s=t}}{\mathsf{P}[T^* \geq t,C \geq t]}\\ &= \lim_{h\searrow 0}\frac{1}{h}\,\mathsf{P}(t \leq T^* < t+h|T^* \geq t, C \geq t)\ \forall t \geq 0.\end{aligned}$$

Ak sú T^* a C nezávislé, podmienka nezávislého cenzorovania je automaticky splnená. V zvyšku práce budeme predpokladať splnenie podmienky nezávislého cenzorovania.

1.2.2 Neparametrické odhady rozdelenia času zlyhania

Nech $(T_1^*,C_1),\ldots,(T_n^*,C_n)$ sú nezávislé a $T_1^*,\ldots,T_n^*$ rovnako rozdelené náhodné veličiny s funkciou prežitia S a kumulatívnou rizikovou funkciou Λ. Buďte $T_i = \min(T_i^*,C_i)$ cenzorované časy zlyhania a $\delta_i = \mathbb{1}(T_i^* \leq C_i)$ indikátory zlyhania. Chceme odhadovať funkciu prežitia S a kumulatívnu rizikovú funkciu Λ z nezávislých pozorovaní $(T_1,\delta_1),\ldots,(T_n,\delta_n)$, bez akýchkoľvek predpokladov o rozdelení T_i^*.
Označme čítacie procesy $N_i(t) = \mathbb{1}(T_i^* \leq t,\delta_i = 1)$ a procesy v riziku $R_i(t) = \mathbb{1}(T_i \geq t)$. Ďalej označme $\overline{N}(t) = \sum_{i=1}^n N_i(t)$, $\overline{R}(t) = \sum_{i=1}^n R_i(t)$ a $\tau^* = \inf\{s : \overline{R}(s) = 0\}$ čas, v ktorom dôjdu dáta. Pre ďalšiu prácu potrebujeme zadefinovať pojem filtrácia a prediktabilita. Prediktabilita procesu je dôležitým predpokladom pre budovanie modelu v nasledujúcej časti.

Definícia 6. *Nech* $(\Omega,\mathcal{F},P)$ *je pravdepodobnostný priestor a* $\{\mathcal{F}_t, t \geq 0\}$ *trieda* σ*-algebier na* Ω*. Ak platí, že* $\forall\ 0 \leq s < t : \mathcal{F}_s \subseteq \mathcal{F}_t \subset \mathcal{F}$*, potom* $\{\mathcal{F}_t\}$ *nazveme filtráciou.*

Definícia 7. *Nech* $(\Omega,\mathcal{F},P)$ *je pravdepodobnostný priestor s filtráciou* $\{\mathcal{F}_t\}$*. Prediktabilná* σ*-algebra* $\mathcal{P}(\mathcal{F}_t)$ *je najmenšia* σ*-algebra obsahujúca množiny typu:* $\{0\} \times A$, $A \in \mathcal{F}_0$ *a* $(s,t] \times A$, $A \in \mathcal{F}_s$, $0 \leq s < t$*. Náhodný proces* $X = \{X_t, t \geq 0\}$ *je* $\mathcal{F}_t$*-prediktabilný, ak platí* $\{(t,\omega) : X_t(\omega) \leq a\} \in \mathcal{P}(\mathcal{F}_t)\ \forall a \in \mathbb{R}$.

Majme filtráciu $\mathcal{F}_t = \sigma\{N_i(u), R_i(u+), 0 \leq u \leq t, i = 1,\ldots,n\}$. Prediktabilitou procesu, budeme myslieť $\mathcal{F}_t$-predikatibilitu. So zavedeným potrebným značením môžeme zadefinovať odhad kumulatívneho rizika.

Definícia 8. *Funkcia*

$$\hat{\Lambda}(t) = \int_0^t \frac{d\overline{N}(u)}{\overline{R}(u)}$$

sa nazýva Nelsonov-Aalenov odhad kumulatívneho rizika.

Odhad navrhol [Nelson, 1969] a jeho konzistenciu a slabú konvergenciu pomocou martingalovej teórie dokázal [Aalen, 1978]. Pre $t > \tau^*$ je Nelsonov-Aalenov odhad konštantný. V dátach nemáme informáciu o riziku potom, čo posledné pozorovanie zlyhá alebo je cenzorované. Označme $t_1 < t_2 \cdots < t_d$ usporiadané rôzne časy zlyhania vypozorované z dát. Potom

$$\hat{\Lambda}(t) = \sum_{\{j:t_j \leq t\}} \frac{\Delta\overline{N}(t_j)}{\overline{R}(t_j)} = \sum_{\{j:t_j \leq t\}} \hat{\lambda}_j$$

je vzorec, podľa ktorého sa Nelsonov-Aalenov odhad počíta a $\Delta\overline{N}(t_j) = \overline{N}(t_j) - \overline{N}(t_j-)$. Člen $\hat{\lambda}_j$ je empirický odhad diskrétneho rizika v čase t_j. Odhad je pomer počtu subjektov, ktorí zlyhali v čase t_j a počtu subjektov v riziku v čase t_j. Keď máme k dispozícii odhad kumulatívneho rizika, môžeme ho využiť k odhadu funkcie prežitia pre spojité T^*, pre ktorú platí $S(t) = e^{-\Lambda(t)}$.

Definícia 9. *Funkciu*

$$\hat{S}(t) = e^{-\hat{\Lambda}(t)}$$

nazveme Flemingov-Harringtonov odhad funkcie prežitia pre spojitý čas do udalosti.

Univerzálnejší odhad funkcie prežitia, ktorý možno použiť aj ak sú zhody v časoch zlyhania, je Kaplan-Meierov odhad.

Definícia 10. *Funkcia*

$$\hat{S}(t) = \prod_{u \leq t} \Big[1 - \frac{\Delta\overline{N}(u)}{\overline{R}(u)}\Big]$$

sa nazýva Kaplan-Meierov odhad funkcie prežitia.

Odhad ako prvý skonštruovali [Kaplan and Meier, 1958]. Pre $t_1 < t_2 \cdots < t_d$ usporiadané rôzne časy zlyhania platí

$$\hat{S}(t) = \prod_{\{j:t_j \leq t\}} \Big[1 - \frac{\Delta\overline{N}(t_j)}{\overline{R}(t_j)}\Big] = \prod_{\{j:t_j \leq t\}} \Big[1 - \hat{\lambda}_j\Big].$$

Kaplan-Meierov odhad je sprava spojitá po častiach konštantná funkcia. Keď v dátach nie je cenzorovanie, $1 - \hat{S}$ je rovné empirickej distribučnej funkcii. Pre $t \geq \tau^*$ je Kaplan-Meierov odhad konštantný. Nespadne na nulu v poslednom pozorovanom čase zlyhania t_d, pokiaľ v tom čase nezlyhajú všetky zostávajúce subjekty.

1.2.3 Model relatívneho rizika

Majme n nezávislých náhodných vektorov $(T_i, \delta_i, \tilde{\mathbf{X}}_i)$, $i = 1, \ldots, n$, kde $T_i = \min(T_i^*, C_i)$ je cenzorovaný čas zlyhania, $\delta_i = \mathbb{1}(T_i^* \leq C_i)$ je indikátor zlyhania a $\tilde{\mathbf{X}}_i = (\tilde{X}_{i_1}, \ldots, \tilde{X}_{i_r})^\top$ je r-rozmerný vektor regresorov. Naším cieľom je vyjadriť potencionálny vplyv vektoru regresorov $\tilde{\mathbf{X}}_i$ na rozdelenie T_i^* pomocou nejakého regresného modelu. Budeme predpokladať, že rozdelenie T_i^* je spojité. Ďalším cieľom bude odhadovanie efektu $\tilde{\mathbf{X}}_i$ na čas zlyhania a testovanie významnosti vplyvu zložiek $\tilde{\mathbf{X}}_i$ na T_i^*. Pre dáta pozorujeme cenzorovaný čas udalosti ako dvojice čítacích procesov $N_i(t) = \mathbb{1}(T_i^* \leq t, \delta_i = 1)$ a procesov v riziku $R_i(t) = \mathbb{1}(T_i \geq t)$. Budeme brať do úvahy, že vektor regresorov $\tilde{\mathbf{X}}_i$ môže závisieť na čase. Teda $\tilde{\mathbf{X}}_i(t)$ budú vektory r sprava spojitých náhodných procesov (tento scenár pripúšťa aj konštantné zložky $\tilde{\mathbf{X}}_i(t)$).
Podmienka nezávislého cenzorovania musí vziať do úvahy, že riziko môže byť ovplyvnené regresormi. Vyjadríme ju teda v tvare podmieneného rizika:

$$\begin{aligned}\lambda(t|\tilde{\mathbf{X}}) &\equiv \lim_{h \searrow 0} \frac{1}{h} \mathsf{P}(t \leq T^* < t + h | T^* \geq t, \tilde{\mathbf{X}}(t)) \\ &= \lim_{h \searrow 0} \frac{1}{h} \mathsf{P}(t \leq T^* < t + h | T^* \geq t, C \geq t, \tilde{\mathbf{X}}(t)). \end{aligned} \tag{1.2}$$

Postačujúcou podmienkou pre nezávislé cenzorovanie je, aby T^* a C boli podmienene na regresoroch nezávislé. To umožňuje, aby časy cenzorovania záviseli na regresoroch (napr. muži môžu mať rozdielne rozdelenie času cenzorovania ako ženy, pokiaľ je pohlavie ako regresor zahrnuté v modeli). Namiesto modelovania strednej hodnoty regresným modelom, budeme budovať model pre podmienené riziko. Coxov model proporčných rizík [Cox, 1972] predpokladá špeciálny predpis efektu regresorov na rizikovú funkciu.

Definícia 11. *Pozorovania $(T_i, \delta_i, \tilde{\mathbf{X}}_i)$, $i = 1, \ldots, n$, spĺňajú Coxov model proporčných rizík, ak platia nasledujúce 2 podmienky:*

1. *trojice $(T_i, \delta_i, \tilde{\mathbf{X}}_i)$ sú navzájom nezávislé,*
2. *podmienená riziková funkcia má, podmienene pri danom regresnom vektore náhodných procesov, tvar*

$$\lambda(t|\tilde{\mathbf{X}}) = \lambda_0(t) e^{\boldsymbol{\beta}_0^\top \tilde{\mathbf{X}}(t)},$$

kde $\lambda_0(t)$ je neznáma bližšie nešpecifikovaná riziková funkcia (tzv. základné riziko) a $\boldsymbol{\beta}_0 \in \mathbb{R}^r$ je neznámy vektor regresných koeficientov.

Základné riziko odpovedá riziku subjektu, ktorého všetky zložky vektoru regresorov sú nulové. Z tohto dôvodu v modeli nie je zahrnutý absolútny člen, úlohu

absolútneho členu zohráva základné riziko. Coxov model patrí medzi tzv. semiparametrické modely. Predpokladá konkrétny tvar asociácie medzi regresormi a rizikovou funkciou, ale pre tvar rizikovej funkcie nie sú vyslovené žiadne predpoklady. Odhad parametrov v Coxovom modeli proporčných rizík teda nemôže byť založený na maximálnej vierohodnosti, keďže sa nejedná o parametrický model. [Cox, 1972] navrhol tzv. parciálnu vierohodnosť, ktorá nezávisí na $\lambda_0(t)$.
Ak hodnoty regresorov závisia na čase, riziko v čase t môže závisieť iba na hodnote regresoru v rovnakom čase. Daný regresor môže byť transformovaný tak, že hodnota v čase t zahŕňa v istom zmysle históriu regresoru. Každopádne regresory nesmú závisieť na ničom zmeranom po čase t, keďže to by porušilo prediktabilitu. Predpokladajme, že regresory sú v čase konštantné, t.j. $\tilde{\mathbf{X}}(t) \equiv \tilde{\mathbf{X}}$. Potom z definície 11 pre ľubovoľné 2 vektory regresorov $\tilde{\mathbf{X}}$ a $\tilde{\mathbf{X}}^*$ plynie:

$$\frac{\lambda(t|\tilde{\mathbf{X}}^*)}{\lambda(t|\tilde{\mathbf{X}})} = \exp\{\boldsymbol{\beta}_0^\top(\tilde{\mathbf{X}}^* - \tilde{\mathbf{X}})\},$$

teda, pomer rizík (relatívne riziko) sa pre akékoľvek dva subjekty v čase nemení. Tento predpoklad sa nazýva predpoklad proporčných rizík. Pre $\tilde{\mathbf{X}}^* = \tilde{\mathbf{X}} + \mathbf{e}_j$, kde $\mathbf{e}_j$ je r-rozmerný kanonický vektor s j-tou zložkou rovnou 1 a ostatnými 0 dostaneme:

$$\exp\{\beta_j\} = \frac{\lambda(t|\tilde{\mathbf{X}} + \mathbf{e}_j)}{\lambda(t|\tilde{\mathbf{X}})},$$

pre akékoľvek $\tilde{\mathbf{X}}$ a t. Na základe tohto výrazu možno jednoducho interpretovať regresné koeficienty: umocnený regresný koeficient odpovedá relatívnemu riziku udalosti pri jednotkovom náraste daného regresoru. Pre konštantné regresory možno Coxov model vyjadriť aj pomocou funkcie prežitia ako

$$S(t|\tilde{\mathbf{X}}) = \exp\{-\int_0^t \lambda(s|\tilde{\mathbf{X}})ds\} = [S_0(t)]^{\exp\{\beta_0^T\tilde{\mathbf{X}}\}}, \tag{1.3}$$

kde $S_0(t) = \exp\{-\Lambda_0(t)\}$ je základná funkcia prežitia a $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ je základné kumulatívne riziko. Keďže parciálna vierohodnosť eliminuje základné riziko λ_0, odhad kumulatívneho základného rizika je odvodený momentovou metódou [Breslow, 1972].

Definícia 12. *Funkcia*

$$\hat{\Lambda}_0 = \int_0^t \frac{d\overline{N}(s)}{\sum_{i=1}^n R_i(s)\exp\{\hat{\boldsymbol{\beta}}^\top\tilde{\mathbf{X}}_i(s)\}}$$

sa nazýva Breslowov odhad kumulatívneho základného rizika.

Ak nie je splnený predpoklad proporčných rizík, sú dva spôsoby, ako do modelu zakomponovať premenné, ktoré tento predpoklad nespĺňajú: stratifikácia a časovo závislé efekty. Stratifikovaný Coxov model je založený na modelovaní rôznych rizikových funkcií v rámci úrovní daných kategóriami kategorickej premennej, ktorá nespĺňa predpoklad proporčných rizík. Časovo závislé efekty môžu byť modelované pomocou interakcie časovo nezávislej premennej a nejakej funkcie času $g(t)$ (napr. $g(t) = t$, $g(t) = \log(t+1)$, $g(t) = \mathbb{1}(s_1 \leq t < s_2)$, kde $s_1, s_2 > 0$).

Umocnený regresný koeficient vyjadruje relatívne riziko v danom čase t pri jednotkovom nárastne príslušnej premennej a nezmenených ostatných regresoroch.

Pri časovo premenlivých regresoroch podmienená riziková funkcia nemôže byť jednoducho vyintegrovaná a funkcia prežitia (1.3) sa nedá vyjadriť týmto spôsobom. Rozlišujeme 2 typy časovo premenlivých regresorov [Kalbfleisch and Prentice, 2002]:

- externé (exogénne): hodnota regresoru v čase t nezávisí na výskyte udalosti v čase u pre $t > u$,
- interné (endogénne): hodnota regresoru v čase t je asociovaná s výskytom udalosti v čase u pre $t > u$.

Je dôležité rozlišovať medzi týmito dvoma typmi časovo premenlivých regresorov, keďže typ regresoru určuje vhodný typ analýzy. Práca s endogénnymi regresormi ako s exogénnymi môže vyústiť v nesprávne výsledky. Coxov model možno rozšíriť na prácu s exogénnymi časovo premenlivými regresormi. Interpretácia príslušných regresných koeficientov je opäť viazaná ku konkrétnemu času t. Pre prácu s endogénnymi časovo premenlivými regresormi navrhol [Rizopoulos, 2012] použitie združeného modelu pre longitudinálne dáta a dáta pre čas do udalosti. Tento model bližšie popíšeme v ďalšej kapitole, keďže je základom pre dynamickú predikciu.

Zovšeobecnenie Coxovho modelu možno získať použitím všeobecnej linkovej funkcie g, ktorá vyjadruje vzťah medzi lineárnym pediktorom $\boldsymbol{\beta}_{\mathbf{0}}^{\top}\tilde{\mathbf{X}}$ a rizikom $\lambda(t|\tilde{\mathbf{X}})$. Model môžeme zapísať v tvare $\lambda(t|\tilde{\mathbf{X}}(t)) = \lambda_0(t)g(\boldsymbol{\beta}_{\mathbf{0}}^{\top}\tilde{\mathbf{X}})$, kde $g(.)$ je rastúca, dvakrát diferencovateľná a spĺňa $g(0) = 1$. Predpoklad proporčných rizík je aj v takomto prípade splnený. Napríklad funkcia $g(y) = 1 + y$ generuje takzvaný aditívny model relatívneho rizika, ktorý sa používa napríklad v radiačnej epidemiológii na modelovanie efektu vystavenia radiácii na výskyt rakoviny.

1.3 Bayesovské metódy

Základným princípom bayesovských metód je, že neznámy parameter môžeme považovať za náhodnú veličinu a informácia o hodnote tohto parametra môže byť vyjadrená pomocou pravdepodobnostného rozdelenia. K záverom o hodnote neznámeho parametra využívame apriórnu informáciu o hodnote parametra a experimetálne výsledky, čiže dáta (nezávislé na tejto apriórnej informácii). V nasledujúcej časti práce stručne zhrnieme úvod do bayesovskej teórie a popíšeme niektoré metódy odhadovania parametrov.

1.3.1 Základy bayesovského prístupu

Ako prvé zavedieme potrebné značenie. Symbolom $p(\cdot|\cdot)$ budeme označovať podmienenú hustotu s argumentami podľa kontextu a podobne $p(\cdot)$ marginálnu hustotu. Hustotu diskrétnej a spojitej náhodnej veličiny budeme značiť rovnako. Nech $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\top} \in \Theta \subset \mathbb{R}^p$ je náhodný vektor, predstavujúci neznámy parameter, s hustotou $p(\boldsymbol{\theta})$ vzhľadom k nejakej σ-konečnej miere λ na $(\Theta, \mathcal{B}(\Theta))$, kde $\mathcal{B}(\Theta)$ označuje borelovské podmnožiny Θ. Ďalej nech $\mathbf{X} = (X_1, \ldots, X_n)^{\top}$ je

náhodný vektor s podmienenou hustotou $p(\mathbf{x}|\boldsymbol{\theta})$, pri danom $\boldsymbol{\theta}$, vzhľadom k σ-konečnej miere ν_n na $(\mathbb{R}^n, \mathcal{B}^n)$, kde $\mathcal{B}^n$ označuje borelovské množiny $\mathbb{R}^n$. Platí teda:

$$\mathsf{P}(\boldsymbol{\theta} \in B, \mathbf{X} \in C) = \int_B \Big(\int_C p(\mathbf{x}|\boldsymbol{\theta}) d\nu_n(\mathbf{x}) \Big) p(\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta}),$$

kde B, C sú ľubovoľné merateľné množiny.

Kľúčové postavenie v bayesovskej teórii má nasledujúca veta.

Veta 1 (Bayesova)**.** *Pre podmienenú hustotu $p(\boldsymbol{\theta}|\mathbf{X})$ náhodného vektoru $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, ktorá je absolútne spojitá vzhľadom k Lebesgueovej miere, pri danom $\mathbf{X}$ platí*

$$p(\boldsymbol{\theta}|\mathbf{X}) = \begin{cases} \frac{p(\boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta})}{\int_\Theta p(\boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta})}, & ak \int_\Theta p(\boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta}) \neq 0, \\ 0, & inak. \end{cases}$$

Dôkaz. [Hušková, 1985][1. Kapitola]

□

Pre parameter $\boldsymbol{\theta}$ nazveme hustotu $p(\boldsymbol{\theta})$ *apriórnou hustotou*, keďže vyjadruje informáciu o $\boldsymbol{\theta}$ ešte pred realizáciou $\mathbf{X}$. Pre $\mathbf{x} = (x_1, \ldots, x_n)^\top$, pozorované hodnoty vektoru $\mathbf{X}$, potom podmienenú hustotu $p(\boldsymbol{\theta}|\mathbf{x})$ nazývame *aposteriórna hustota*, keďže sa jedná o hustotu po realizácii $\mathbf{X}$. Pomocou Bayesovej vety teda aktualizujeme informáciu o $\boldsymbol{\theta}$ získaním informácie o $\boldsymbol{\theta}$ obsiahnutej v pozorovaniach $\mathbf{x}$. Bayesovský model môžeme formálne zadefinovať podľa definície [Robert, 2007].

Definícia 13. *Bayesovský štatistický model pozostáva z parametrického modelu $p(\mathbf{X}|\boldsymbol{\theta})$ a apriórneho rozdelenia parametra $p(\boldsymbol{\theta})$.*

Vierohodnostnú funkciu $\boldsymbol{\theta}$ možno značiť ako $L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta})$. Z Bayesovej vety máme, že aposteriórne rozdelenie $p(\boldsymbol{\theta}|\mathbf{X})$ je proporčné vierohodnosti prenásobenej apriórnym rozdelením

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto L(\boldsymbol{\theta}; \mathbf{X}) p(\boldsymbol{\theta}). \tag{1.4}$$

Vierohodnostná funkcia $\boldsymbol{\theta}$ a apriórne rozdelenie $\boldsymbol{\theta}$ teda určujú jednoznačne združené rozdelenie $p(\mathbf{X},\boldsymbol{\theta})$ vzťahom $p(\mathbf{X}; \boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{X}) p(\boldsymbol{\theta})$. Menovateľ v predpise z Bayesovej vety, $\int_\Theta p(\boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta})$, je normujúca konštanta $p(\boldsymbol{\theta}|\mathbf{X})$ a marginálne rozdelenie $\mathbf{X}$ pri $\boldsymbol{\theta}$. Spočítať hodnotu tejto funkcie presne je zvyčajne problematické, no v niektorých situáciách to ani nebude potrebné, ako bude vysvetlené v ďalšej časti. V bayesovskom modeli je spravidla užitočné rozložiť vierohodnosť a apriórne rozdelenie na niekoľko podmienených rozdelení, čím vznikne tzv. hierarchický bayesovský model. Hierarchická špecifikácia je často prirodzeným spôsobom na konštrukciu realistických pravdepodobnostných modelov na popis reálnej situácie. Dôvodom na použitie hierarchického modelu je, že celková informácia je rozložená na viacerých úrovniach experimentálnych jednotiek. Napríklad v medicíne, biológii, ekonomike, atď., možno skúmanú populáciu vnímať ako subpopuláciu globálnej populácie. Hierarchické modelovanie hrá tiež dôležitú

úlohu vo výpočtových metódach odhadov parametrov, ako bude vidieť v ďalšej časti.

Nech $t : \Theta \to \mathbb{R}^q$ je merateľná funkcia a hlavným parametrom záujmu je $t(\boldsymbol{\theta})$. Za bodový odhad $t(\boldsymbol{\theta})$ budeme brať aposteriórnu strednú hodnotu $\overline{t(\boldsymbol{\theta})} = \mathsf{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[t(\boldsymbol{\theta})] = \int_\Theta t(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\lambda(\boldsymbol{\theta})$, ak existuje. Na spočítanie tohto odhadu je potrebné poznať aposteriórne rozdelenie parametra $\boldsymbol{\theta}$ alebo vedieť spočítať daný integrál. Takmer vždy je tak potrebné spočítať marginálne aposteriórne rozdelenia všetkých zložiek $\boldsymbol{\theta}$ a následne spočítať daný integrál, čo je v zložitejších modeloch náročné. Ak by sme mali k dispozícii náhodný výber z rozdelenia $p(\boldsymbol{\theta}|\mathbf{X})$, ktorý označíme $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}, M \in \mathbb{N}$, mohli by sme použiť Silný zákon veľkých čísel a použiť odhad $\widehat{t(\boldsymbol{\theta})} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}^{(m)} \xrightarrow[M\to\infty]{sj.} \overline{t(\boldsymbol{\theta})}$. Presný tvar rozdelenia $p(\boldsymbol{\theta}|\mathbf{X})$ však typicky nie je známy, preto je potrebné zvoliť inú metódu. Odhadovacie metódy, popísané v ďalšej časti, umožňujú generovať markovské reťazce, ktorých limitným rozdelením je požadované aposteriórne rozdelenie. Na odhadovanie aposteriórnej strednej hodnoty možno následne použiť členy týchto reťazcov.

1.3.2 Metódy na výpočty odhadov

Pri odhadovaní integrálov a iných funkcionálov, ktoré závisia na aposteriórnom rozdelení parametra $\boldsymbol{\theta}$, býva vo všeobecnosti komplikované generovať $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ priamo z $p(\boldsymbol{\theta}|\mathbf{X})$. Čoraz častejšie používanou bayesovskou metódou na odhadovanie parametrov je metóda Monte Carlo pre markovské reťazce (anglicky *Markov Chain Monte Carlo – MCMC*). Ako napovedá názov, hlavná myšlienka tejto metódy je generovať markovské reťazce $\{\boldsymbol{\theta}^{(m)}\}_m$, ktorých limitným rozdelením je požadované aposteriórne rozdelenie $p(\boldsymbol{\theta}|\mathbf{X})$. V tejto časti uvedieme dva základné algoritmy na výpočet MCMC, Gibbsov algoritmus a Metropolisov–Hastingsov algoritmus [Robert, 2007].

Gibsov algoritmus je jedným z najznámejších MCMC algoritmov. Formálne ho zaviedli [Geman and Geman, 1984]. K implementácii algoritmu sú potrebné nasledujúce predpoklady:

- vieme generovať z každého plne podmieneného rozdelenia $\{p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{j\neq i},\mathbf{x}), i = 1, \ldots, p\}$,
- cieľové (stacionárne) rozdelenie má hustotu $p(\boldsymbol{\theta}|\mathbf{X})$ vzhľadom k súčinovej miere $\lambda_1 \otimes \cdots \otimes \lambda_p$, kde λ_i je nejaká σ-konečná miera spĺňajúca $\lambda_i(\Theta_i) > 0,\ i = 1, \ldots, p$,
- podmienka pozitivity: $\Theta = \prod_{i=1}^{p} \Theta_i$, $\Theta = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}|\mathbf{X}) > 0\}$.

Podmienené rozdelenia, s ktorými budeme pracovať v ďalších kapitolách, budú absolútne spojité vzhľadom k Lebesgueovej miere. Druhá podmienka tak bude automaticky splnená. Za platnosti podmienky pozitivity je výsledný reťazec ergodický a rozdelenie markovského reťazca konverguje k cieľovému rozdeleniu[Robert, 2007][6. Kapitola]. Schéma algoritmu je nasledujúca:

Gibbsov algoritmus: Zvoľ počiatočnú hodnotu $\boldsymbol{\theta}_0 = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^\top$, $m = 0$ a simuluj:

1. $\theta_1^{(m+1)} \sim p(\theta_1 | \theta_2^{(m)}, \ldots, \theta_p^{(m)}, \boldsymbol{x})$,
2. $\theta_2^{(m+1)} \sim p(\theta_2 | \theta_1^{(m+1)}, \theta_3^{(m)}, \ldots, \theta_p^{(m)}, \boldsymbol{x})$,

$\vdots$

p. $\theta_p^{(m+1)} \sim p(\theta_p | \theta_1^{(m+1)}, \ldots, \theta_{p-1}^{(m+1)}, \boldsymbol{x})$,

Výstupom po m-tej iterácii je p-tica $(\theta_1^{(m)}, \ldots, \theta_p^{(m)})^\top$, ktorá konverguje v distribúcii k aposteriórnemu rozdeleniu $p(\boldsymbol{\theta}|\boldsymbol{x})$. Teda pre m dostatočne veľké (väčšie ako nejaké stanovené $m_0 \in \mathbb{N}$) možno $\{\boldsymbol{\theta}^{(m)}, m = m_0, \ldots, M\}$ považovať za náhodnú postupnosť zo skutočného aposteriórneho rozdelenia, na základe ktorej možno robiť ďalšiu inferenciu. Napríklad, ako bolo naznačené v predchádzajúcej časti, na odhadovanie aposteriórnej strednej hodnoty možno využiť výberový priemer

$$\widehat{t(\boldsymbol{\theta})} = \frac{1}{M - m_0} \sum_{m=m_0+1}^{M} \boldsymbol{\theta}^{(m)}. \tag{1.5}$$

Úsek medzi $m = 0$ a $m = m_0$ sa nazýva „rozohrievacia“ fáza (angl. *burn-in period*) a postupnosti získané z tejto fázy algoritmu nie sú používané na inferenciu o aposteriórnom rozdelení.

Metropolisov–Hastingsov algoritmus bol skonštruovaný pôvodne pre mechanickú fyziku [Metropolis et al., 1953] a neskôr zovšeobecnený na štatistické výpočty [Hastings, 1970]. Predpoklad na použitie algoritmu je znalosť cieľového rozdelenia $p(\boldsymbol{\theta}|\mathbf{X})$, až na normujúcu konštantu, vzhľadom k nejakej σ-konečnej miere λ, ktorá spĺňa $\lambda(\Theta) > 0$ pre parametrický priestor $\Theta = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}|\mathbf{X}) > 0\}$. Toto bude opäť automaticky splnené pre všetky podmienené rozdelenia, s ktorými budeme pracovať vo zvyšku práce. Na začiatok treba špecifikovať tzv. návrhovú hustotu $q(\cdot|\boldsymbol{\theta})$ vzhľadom k σ-konečnej miere λ. Algoritmus potom generuje reťazec $\{\boldsymbol{\theta}^{(m)}\}_{m=1,\ldots,M}$ nasledovne:

Metropolisov–Hastingsovs algoritmus:

1. začni s ľubovoľnou počiatočnou hodnotou $\boldsymbol{\theta}^{(0)}$
2. aktualizuj $\boldsymbol{\theta}^{(m)}$ na $\boldsymbol{\theta}^{(m+1)}$, $(m = 0,1,2,\ldots,M)$ postupom
   (a) generuj $\boldsymbol{\xi} \sim q(\boldsymbol{\xi}|\boldsymbol{\theta}^{(m)})$
   (b) $\rho = min\{\frac{p(\boldsymbol{\xi})q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\xi})}{p(\boldsymbol{\theta}^{(m)})q(\boldsymbol{\xi}|\boldsymbol{\theta}^{(m)})}, 1\}$
   (c) vezmi $\boldsymbol{\theta}^{(m+1)} = \begin{cases} \boldsymbol{\xi} & \text{s pravdepodobnosťou } \rho, \\ \boldsymbol{\theta}^{(m)}, & \text{inak.} \end{cases}$

Na odhadovanie parametrov sa opäť používajú členy markovského reťazca bez zahrievacej fázy, tj. $\{\boldsymbol{\theta}^{(m)}\}_{m=m_0}^{M}$. Výhodami tohto algoritmu je, že je nie je potrebné poznať normujúcu konštantu a návrhovú hustotu je možné voliť ľubovoľne. Voľba návrhovej hustoty však môže mať veľký dopad na rýchlosť konvergencie algoritmu. Aby bola zaručená existencia limitného rozdelenia, musí návrhová hustota spĺňať isté podmienky. Tie sú splnené napríklad pre symetrickú náhodnú prechádzku. Formálne odôvodnenie funkčnosti MCMC algoritmov možno nájsť v [Robert, 2007][6. Kapitola].

1.3.3 Bayesovská predikcia

Dôležitým aspektom bayesovskej teórie je predikcia. Majme nezávislé náhodné veličiny a model

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top,\ \mathbf{X}_i \sim p_i(\mathbf{x}_i|\boldsymbol{\theta}), i = 1, \ldots, n,$$

kde p_i má rovnakú funkcionálnu formu pre všetky i a závisí na i iba prostredníctvom známych faktorov (napríklad vysvetľujúce premenné v regresnom modeli). Majme apriórne rozdelenie $p(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Potom aposteriórne rozdelenie parametra $\boldsymbol{\theta}$ pri napozorovaných dátach $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ je dané predpisom

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \prod_{i=1}^{n} p_i(\mathbf{x}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Označme $\mathbf{Z}$ nový d_z-rozmerný náhodný vektor, ktorý je nezávislý od $\mathbf{X}$, pri danom modeli a generovaný rovnakým pravdepodobnostným mechanizmom ako $\mathbf{X}$, teda $\mathbf{Z} \sim p_z(\mathbf{z}; \boldsymbol{\theta})$, v ktorom sú dodatočné faktory známe (napr. dodatočné vysvetľujúce premenné).

Definícia 14. *Aposteriórne prediktívne rozdelenie budúceho pozorovania $\boldsymbol{Z}$, pri napozorovaných dátach $\mathbf{x}$, je rozdelenie s hustotou v $\mathbf{z} \in \mathbb{R}^{d_z}$, danou predpisom*

$$p_{pred}(\mathbf{z}|\mathbf{x}) = \int_\Theta p_z(\mathbf{z}; \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\lambda(\boldsymbol{\theta}).$$

Z predpisu vidíme, že sa jedná o aposteriórnu strednú hodnotu $p(\mathbf{z}|\boldsymbol{\theta})$ a náhodný výber z $p_{pred}(\mathbf{z}|\mathbf{x})$ možno ľahko získať pomocou Gibbsovho algoritmu. Hustota $p_{pred}(\mathbf{z}|\mathbf{x})$ sa nazýva aposteriórna prediktívna hustota. Aposteriórne prediktívne rozdelenie nového pozorovania $\mathbf{Z}$ je teda marginálne rozdelenie $\mathbf{Z}$ pri zohľadnení informácie o $\boldsymbol{\theta}$ obsiahnutej v dátach $\mathbf{X}$. Ako bodovú predikciu na základe aposteriórneho prediktívneho rozdelenia budeme brať

$$\widehat{\mathbf{Z}} = \mathsf{E}_{p_{pred}(\mathbf{z}|\mathbf{x})}\, \mathbf{Z} = \int_{\mathbb{R}^{d_z}} \mathbf{z}^* p_{pred}(\mathbf{z}^*|\mathbf{x}) d\lambda_z(z^*) \text{ (ak existuje)},$$

kde λ_z je σ-konečná miera príslušná hustote $\mathbf{Z}$. Okrem bodovej predikcie možno v jednorozmernom prípade ($d_z = 1$) robiť aj intervalovú predikciu, pomocou vierohodnostných intervalov, založenú na aposteriórnom prediktívnom rozdelení. Využitím bayesovskej teórie teda možno, pri znalosti $p(\boldsymbol{\theta}|\boldsymbol{x})$, predikcie spočítať veľmi ľahko, preto ju uprednostníme pred frekventistickým prístupom.

2. Združené modely

Motivácia pre budovanie združených modelov je rozšíriť model relatívneho rizika, v ktorom máme iba diskrétne časové okamihy tak, aby zahŕňal aj endogénne časové premenné. Keď je cieľom inferencia o čase do udalosti, s ohľadom na longitudinálne merania (napr. tlak, teplota pacienta alebo hladina určitej látky v krvi), postupy popísané v predchádzajúcej kapitole nemôžeme použiť. Intuitívna myšlienka za týmito modelmi je nasledujúca: použiť vhodný model pre popísanie endogénnej časovo závislej premennej pre každého pacienta a následne odhadnutý vývoj použiť v modeli pre čas do udalosti. Hlavným predpokladom pre použitie združeného modelu je, že existuje latentná premenná, ktorá zachytáva závislosť medzi longitudinálnou premennou a časom udalosti a podmienene na tejto premennej sú čas do udalosti a longitudinálna premenná nezávislé [Hogan and Laird, 1997]. Modely pre longitudinálne dáta a pre dáta pre čas do udalosti môžeme chápať tak, že združene závisia na zdieľaných náhodných efektoch [Wulfsohn and Tsiatis, 1997]. Ďalším typom združených modelov sú združené modely s latentnými kategóriami, v ktorých sa predpokladá, že longitudinálna premenná a čas do udalosti sú nezávislé podmienene na nejakej diskrétnej latentnej premennej [Proust-Lima et al., 2014]. V tejto práci budeme pracovať najmä so združenými modelmi so spoločnými náhodnými efektami, preto im venujeme väčšiu pozornosť. Združeným modelom s latentnými kategóriami sa podrobne venuje napr. práca [Vorlíčková, 2020].

2.1 Združené modely so spoločnými náhodnými efektami

Rozšíreným typom združených modelov sú združené modely so spoločnými náhodnými efektami (anglicky *the Shared Random-Effect Model- SREM*). Ako model pre longitudinálnu premennú možno voliť lineárny zmiešaný model [Wulfsohn and Tsiatis, 1997] alebo zovšeobecnený lineárny zmiešaný model [Xu and Zeger, 2001]. Pre dáta pre čas do udalosti je zvyčajne volený Coxov model proporčných rizík. Latentnú premennú, ktorá popisuje závislosť medzi longitudinálnou premennou a časom do udalosti, predstavujú náhodné efekty linárneho zmiešaného modelu [Tsiatis and Davidian, 2004], ako ďalej ukážeme. K zavedeniu podrobnej definície združeného modelu najskôr zopakujeme značenie z prvej kapitoly a zavedieme potrebné nové značenie.

2.1.1 Zavedenie modelu

Rovnako ako v kapitole 1.1, bude K značiť počet subjektov. Pre každé $i = 1, \ldots, K$ označme náhodný proces $Y_i(t), 0 \leq t \leq L$, kde L predstavuje nejakú konštantu ohraničujúcu dobu sledovania subjektov. Proces $Y_i(t)$ bude (istým spôsobom) predstavovať endogénnu časovo závislú premennú v Coxovom modeli, ktorý uvedieme o chvíľu. Longitudinálne údaje pre jednotlivé subjekty nameriame v časoch $0 \leq t_{i1} < \cdots < t_{in_i}$, $i = 1, \ldots, K$. Hodnotu $Y_i(t)$ nepozorujeme v ktorom-

koľvek časovom okamihu t, ale iba v časových okamihoch, kedy boli vykonané merania. Budeme teda značiť $Y_{ij} = Y_i(t_{ij}), i = 1,\ldots,K, j = 1,\ldots,n_i$ a $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{in_i})^\top$, vektor meraní pre i-teho pacienta. Symbolom $\mathbf{Y} = (\mathbf{Y}_1^\top,\ldots,\mathbf{Y}_K^\top)^\top$ budeme značiť vektor pozorovaných longitudinálnych premenných pre všetky subjekty dohromady.

Ďalej využijeme pojmy a značenie z kapitoly 1.2 a označíme pre každý subjekt nezáporné (nepozorované) náhodné veličiny pre skutočný čas udalosti a čas cenzorovania $(T_1^*,C_1),\ldots,(T_K^*,C_K)$. Budeme značiť $T_i = \min(T_i^*,C_i), i = 1,\ldots,K$ a $\mathbf{T} = (T_1,\ldots,T_K)^\top$ vektor cenzorovaných časov udalostí pre všetky subjekty, $\delta_i = \mathbb{1}(T_i^* \leq C_i), i = 1,\ldots,K$, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_K)^\top$ príslušné indikátory udalostí. Budeme predpokladať splnenie podmienky nezávislého cenzorovania z definície 5. Združenú hustotu veličín $\mathbf{Y}$,$\mathbf{T}$ a $\boldsymbol{\delta}$ budeme značiť $f_{(\mathbf{Y},\mathbf{T},\boldsymbol{\delta})}$.

Združený model je štruktúrovaný hierarchicky. Prvý krok pri definícii tohto modelu je definovať model pre longitudinálne dáta (endogénnu časovú premennú v Coxovom modeli), ďalej model prežitia a nakoniec asociáciu medzi týmito dvoma modelmi – model pre združené rozdelenie ($\mathbf{Y}$,$\mathbf{T}$,$\boldsymbol{\delta}$). V modeli prežitia bude jediný modelovaný regresor – odozva z longitudinálneho modelu, zvyšné regresory sa nemodelujú (sú dané). Symbolom $m_i(t)$ budeme značiť takzvanú systematickú zložku premennej $Y_i(t)$ a $\epsilon_i(t)$ bude značiť náhodnú zložku merania v čase t, teda $Y_i(t) = m_i(t) + \epsilon_i(t)$. Pre náhodnú zložku budeme predpokladať $\epsilon_i(t) \sim \mathsf{N}(0,\sigma^2)$. Históriu systematickej zložky $m_i(t)$ longitudinálnej premennej označíme $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$. Zadefinujeme model pre longitudinálne dáta:

1. Lineárny zmiešaný model z definície 1 rozpísaný po zložkách:

$$\begin{aligned} Y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= \mathbf{X}_i(t)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t)^\top\mathbf{b}_i + \epsilon_i(t),\ i = 1,\ldots K. \end{aligned} \tag{2.1}$$

Vektor náhodných efektov bude spĺňať $\mathbf{b}_i \sim \mathsf{N}_q(\mathbf{0},\mathbb{D})$. Rovnako ako v definícii 1, bude $\boldsymbol{\beta}$ p-rozmerný vektor fixných efektov, $\mathbf{X}_i(t)$ regresný vektor pre fixné efekty v čase t a $\mathbf{Z}_i(t)$ regresný vektor pre náhodné efekty v čase t.

Ďalej zadefinujeme model prežitia:

2. Coxov model proporčných rizík:

$$\begin{aligned} \lambda_i(t|\mathbf{b}_i) &= \lambda_0(t,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(t) + \alpha m_i(t)\}, \\ &= \lambda_0(t,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(t) + \alpha[\mathbf{X}_i(t)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t)^\top\mathbf{b}_i]\},\ i = 1,\ldots K, \end{aligned} \tag{2.2}$$

kde $\lambda_0(t,\boldsymbol{\xi})$je nešpecifikovaná základná riziková funkcia, ktorá závisí na vektore neznámych parametrov $\boldsymbol{\xi}$, $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_r)^\top$ je vektor neznámych parametrov a α je neznámy parameter, ktorý vyjadruje vzťah medzi $m_i(t)$ (endogénnou časovou premennou) a rizikom udalosti v čase t. Vektor $\tilde{\mathbf{X}}_i(t)$ je r-rozmerný vektor (časovo závislých) regresorov v čase t. Tieto regresory chápeme, na rozdiel od $m_i(t)$,

ako po častiach konštatné premenné, u ktorých dochádza k zmene hodnôt len v časoch cenzorovania alebo ak dôjde k udalosti.

Posledný krok je definovať asociáciu medzi týmito dvoma modelmi:

3. Združené rozdelenie $(\mathbf{Y}_i, T_i, \delta_i)$ $i = 1, \dots K$:

$$f_{(\mathbf{Y}_i,T_i,\delta_i)}(\mathbf{y}_i, t_i, d_i) = \int_{-\infty}^{\infty} f_{\mathbf{Y}_i|\mathbf{b}_i}(\mathbf{y_i}|\mathbf{b}_i)[\lambda_i(t_i|\mathbf{b}_i)]^{\delta_i} S(t_i|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i. \quad (2.3)$$

Kľúčovými predpokladmi pre budovanie modelu sú [Rizopoulos, 2012, Kapitola 4]:

1. nezávislosť subjektov: pre $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_K)^\top$, $\mathbf{y}_i \in \mathbb{R}^{n_i}$, $\mathbf{t} = (t_1, \dots, t_K)^\top$, $t_i > 0$, $\boldsymbol{d} = (d_1, \dots, d_K)^\top$, $d_i \in \{0,1\}$ $\forall i = 1, \dots, K$ platí

$$f_{(\mathbf{Y},\mathbf{T},\boldsymbol{\delta})}(\mathbf{y}, \mathbf{t}, \boldsymbol{d}) = \prod_{i=1}^{K} f_{(\mathbf{Y}_i,T_i,\boldsymbol{\delta}_i)}(\mathbf{y_i}, t_i, d_i), \quad (2.4)$$

2. podmiene na náhodných efektoch, sú longitudinálna premenná a čas do udalosti nezávislé: pre každé $i = 1, \dots, K$ a $f(\mathbf{b}_i)$ hustotu $\mathbf{b}_i$ platí

$$f_{(\mathbf{Y}_i,T_i,\delta_i)}(\mathbf{y_i}, t_i, d_i) = \int_{-\infty}^{\infty} f_{\mathbf{Y}_i|\mathbf{b}_i}(\mathbf{y_i}|\mathbf{b}_i) f_{(T_i,\delta_i)|\mathbf{b}_i}(t_i, d_i|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i. \quad (2.5)$$

Značenie. *Ďalej budeme hustotu náhodnej veličiny/náhodného vektoru používať bez spodného indexovania, pre prehľadnosť značenia. Teda $f(\mathbf{x})$ bude značiť hodnotu hustoty pre napozorovanú hodnotu $\mathbf{X} = \mathbf{x}$ a $f(\mathbf{X})$ bude všeobecné značenie hustoty náhodného vektoru $\mathbf{X}$.*

Model relatívneho rizika (2.2) predpokladá, že riziko udalosti v čase t závisí iba na aktuálnej skutočnej hodnote $m_i(t)$. Pre funkciu prežitia však dostávame $S_i(t|\mathbf{b}_i) = \mathsf{P}(T_i^* > t|\mathbf{b}_i) = \exp\{-\int_0^t \lambda_0(t,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(s) + \alpha m_i(s)\}ds\}$, čo implikuje, že funkcia prežitia závisí na celej histórii longitudinálnej premennej $(\mathcal{M}_i(t))$. Táto vlastnosť je pre odhadovanie parametrov v združených modeloch dôležitá, ako si ukážeme ďalej.

Predpis (2.3) platí z predpokladu (2.5) a rozpísaním tvaru združenej hustoty $f(T_i,\delta_i|\mathbf{b}_i) \propto [\lambda_i(T_i|\mathbf{b}_i)]^{\delta_i} S(T_i|\mathbf{b}_i)$ [Kulich, 2021]. Treba zdôrazniť, že nevyhnutným predpokladom pre daný rozpis združenej hustoty je, že pracujeme s neinformatívnym cenzorovaním. Keď označíme vektor neznámych parametrov z modelov (2.1) a (2.2) ako $\boldsymbol{\theta} = (\boldsymbol{\theta}_Y^\top, \boldsymbol{\theta}_T^\top, \boldsymbol{\theta}_b^\top)^\top$, kde $\boldsymbol{\theta}_Y = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ predstavuje vektor parametrov modelu (2.1), $\boldsymbol{\theta}_T = (\boldsymbol{\xi}^\top, \boldsymbol{\gamma}^\top, \alpha)^\top$ predstavuje vektor parametrov modelu (2.2) a $\boldsymbol{\theta}_b = (\text{vec}(\mathbb{D}))^\top = (\mathbb{D}_{1,1}, \dots, \mathbb{D}_{q,1}, \mathbb{D}_{1,2}, \dots, \mathbb{D}_{q,2}, \dots, \mathbb{D}_{1,q}, \dots, \mathbb{D}_{q,q})^\top$ parametre variančnej matice náhodných efektov, môžeme pre združený model (2.3) rozpísať vierohodnostnú funkciu:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{K} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{K} f(\mathbf{Y_i}, T_i, \delta_i, \boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{K} \int_{-\infty}^{\infty} f(\mathbf{Y_i}|\mathbf{b}_i, \boldsymbol{\theta}_Y)[\lambda_i(T_i|\mathbf{b}_i, \boldsymbol{\theta}_T, \boldsymbol{\beta})]^{\delta_i} S(T_i|\mathbf{b}_i, \boldsymbol{\theta}_T, \boldsymbol{\beta}) f(\mathbf{b}_i|\boldsymbol{\theta}_b) d\mathbf{b}_i. \end{aligned} \quad (2.6)$$

Pre prehľadnejší zápis rozpíšeme jednotlivé členy:

$$
\begin{aligned}
L_i(\boldsymbol{\theta}) \propto \int_{-\infty}^{\infty} & \frac{1}{(2\pi\sigma^2)^{n_i/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y_i}-\mathbf{X}_i(T_i)^\top\boldsymbol{\beta}-\mathbf{Z}_i(T_i)^\top\mathbf{b}_i)^\top(\mathbf{Y_i}-\mathbf{X}_i(T_i)^\top\boldsymbol{\beta}-\mathbf{Z}_i(T_i)^\top\mathbf{b}_i)} \\
& \times \left[\lambda_0(T_i,\boldsymbol{\xi}) e^{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(T_i)+\alpha[\mathbf{X}_i(T_i)^\top\boldsymbol{\beta}+\mathbf{Z}_i(T_i)^\top\mathbf{b}_i]}\right]^{\delta_i} \\
& \times e^{-\int_0^{T_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\
& \times \frac{1}{(2\pi)^{q/2}|\mathbb{D}|^{1/2}} e^{-\frac{1}{2}\mathbf{b_i}^\top\mathbb{D}^{-1}\mathbf{b_i}} d\mathbf{b_i}. \qquad (2.7)
\end{aligned}
$$

Na odhadovanie parametrov sa používa ako metóda maximálnej vierohodnosti, tak aj bayesovský prístup. Na spočítanie maximálnej vierohodnosti je potrebná numerická aproximácia integrálov (2.7), ktorá je najmä vzhľadom k $\mathbf{b}_i$ výpočtovo náročná. Na maximalizáciu vierohodnostnej funkcie je potrebné využiť optimalizačné algoritmy, ako napr. EM-algoritmus [Wulfsohn and Tsiatis, 1997]. Z hľadiska výpočtovej náročnosti je preto často výhodnejšie zvoliť bayesovský prístup, v ktorom je inferencia založená na aposteriórnom rozdelení [Xu and Zeger, 2001, Hanson et al., 2011]. Vo väčšej miere sa budeme venovať bayesovskému prístupu aj v tejto práci.

Závislosť medzi longitudinálnou premennou a rizikom udalosti nemusí byť daná priamo aktuálnou hodnotou premennej v čase t. Premenná $m_i(t)$ môže byť parametrizovaná použitím rôznych funkcionálnych foriem, napr.:

1. oneskorené efekty: $m_i(t_+^c)$, $t_+^c = \max(t-c, 0)$, vyžaduje sa, aby riziko v čase t záviselo na systematickej zložke longitudinálnej premennej v čase $t-c$, kde c predstavue požadované oneskorenie,

2. časovo závislá smernica: $\alpha_1 m_i(t) + \alpha_2 m_i'(t)$, kde $m_i'(t) = \frac{d}{dt}\{\mathbf{X}_i(t)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t)^T\mathbf{b}_i\}$, interpretácia α_1 je rovnaká ako štandardne, parameter α_2 meria, ako je asociovaná smernica systematickej longitudinálnej trajektórie v čase t s rizikom udalosti v tomto čase, za predpokladu, že $m_i(t)$ zostane nezmenené,

3. časovo závislá smernica inak: $\Delta m_i(t) = m_i(t) - m_i(t-1)$, riziko udalosti v čase t je asociované so zmenou trajektórie za jednu jednotku času,

4. kumulatívny efekt: $\int_0^t m_i(s)ds$, umožňuje, aby bola celá história longitudinálnej premennej, až do času t, asociovaná s rizikom udalosti v čase t.

Podrobnosti k použiu rôznych funkcionálnych foriem možno nájsť v [Rizopoulos, 2012, Kapitola 5]. Model (2.2) možno rozšíriť pridaním ďalších longitudinálnych premenných alebo uvažovaním viacerých časov zlyhania. V tejto práci sa budeme venovať iba modelom s jednou longitudinálnou premennou a jedným časom zlyhania. Rozšírenie modelov je opäť podrobne rozpísané napr. v [Rizopoulos, 2012, Kapitola 5].

2.1.2 Bayesovské odhadovanie

Pre uvedenie do problematiky bayesovských odhadov v združenom modeli, je potrebné rozpísať tvar hierarchického bayesovského modelu. Pre jednotlivé členy

modelu sa určia podmienené rozdelenia, vďaka čomu bude možné odvodiť tvar združenej hustoty (2.6). Pre združenú hustotu platí:

$$p(\mathbf{Y}_i,T_i,\delta_i,\mathbf{b}_i,\boldsymbol{\theta}) = p(\mathbf{Y}_i|\mathbf{b}_i,\boldsymbol{\theta})p(T_i,\delta_i|\mathbf{b}_i,\boldsymbol{\theta})p(\mathbf{b}_i|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.8}$$

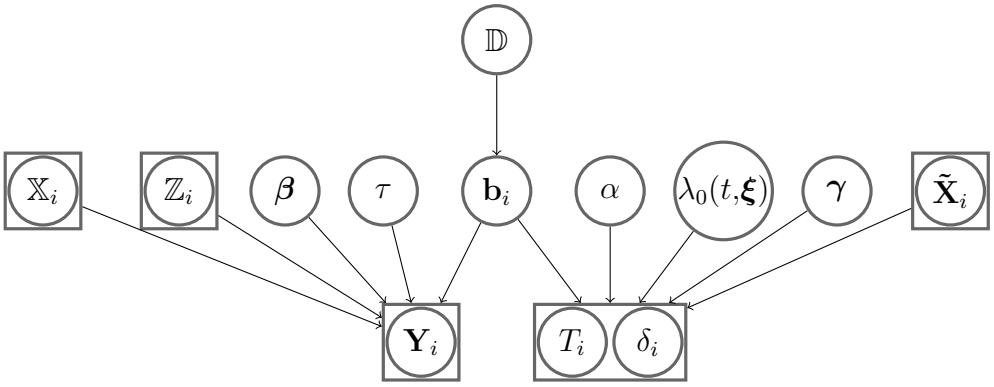
V bayesovskej štatistike sa často používa namiesto rozptylu σ^2 inverzný rozptyl $\tau = \sigma^{-2}$, tzv. presnosť. Ďalej budeme pracovať s parametrom presnosti namiesto rozptylu σ^2. Pre zložky vektoru $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top,\tau,(\text{vec}(\mathbb{D}))^\top,\boldsymbol{\xi}^\top,\boldsymbol{\gamma}^\top,\alpha)^\top$ budeme predpokladať vzájomnú nezávislosť apriórnych rozdelení, teda:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\tau)p(\mathbb{D})p(\boldsymbol{\xi})p(\boldsymbol{\gamma})p(\alpha). \tag{2.9}$$

Na odhadovanie parametrov modelu je potrebné určiť, čo sú pozorované a latentné premenné, čo sú parametre modelu a určiť ich apriórne rozdelenia a definovať tak hierarchický bayesovský model, ako na obrázku 2.1, ktorý graficky znázorňuje štruktúru modelu danú vzťahmi (2.8) a (2.9). Pre apriórne rozdelenia budeme uvažovať obvyklé voľby [Leiva-Yamaguchi and Alvares, 2020, Andrinopoulou et al., 2016] a voľby hyperparametrov odpovedajúce slabo informatívnym apriórnym rozdeleniam. Rozpísaním dostaneme:

1. Pozorované premenné:
   - $\mathbf{Y_i}$, pre ktoré platí $\mathbf{Y_i}|\mathbf{b}_i,\mathbb{X}_i,\mathbb{Z}_i,\boldsymbol{\beta},\tau \sim \mathsf{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta},\mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\top + \tau^{-1}\mathbb{I}_{n_i})$,
   - (T_i,δ_i), ktoré spĺňajú Coxov model (2.2),
2. Latentná premenná:
   - $\mathbf{b}_i$, spĺňajúca $\mathbf{b}_i|\mathbb{D} \sim \mathsf{N}_q(\mathbf{0},\mathbb{D})$,
3. Parametre modelu:
   - $\boldsymbol{\beta} \sim \mathsf{N}_p(\mathbf{0},\boldsymbol{\Sigma}_\beta)$, kde $\boldsymbol{\Sigma}_\beta = \text{diag}(\sigma^2_{\beta_1},\dots,\sigma^2_{\beta_p})$ je pozitívne definitná matica a $\sigma^2_{\beta_j} > 0$, $j = 1,\dots,p$ je volené veľké[1],
   - $\tau \sim \Gamma(a_\tau,b_\tau); a_\tau,b_\tau > 0$, $a_\tau \in (0,1]$, b_τ blízko 0,
   - $\mathbb{D}$, $\mathbb{D}^{-1} \sim \mathsf{W}_q(\nu_D,\Upsilon)$, $\nu_D \in (q-1,q]$, $\Upsilon = \text{diag}(\upsilon_1^{-1},\dots,\upsilon_q^{-1})$, $\upsilon_j^{-1} > 0$, $j = 1,\dots,q$ je volené veľké
   - $\alpha \sim \mathsf{N}(0,\sigma^2_\alpha)$, $\sigma^2_\alpha > 0$ je veľké,
   - $\boldsymbol{\gamma} \sim \mathsf{N}_r(\mathbf{0},\boldsymbol{\Sigma}_\gamma)$, $\boldsymbol{\Sigma}_\gamma = \text{diag}(\sigma^2_{\gamma_1},\dots,\sigma^2_{\gamma_r})$ je pozitívne definitná matica a $\sigma^2_{\gamma_j} > 0$, $j = 1,\dots,r$ je volené veľké,
   - $\lambda_0(t,\boldsymbol{\xi})$, pre ktoré možno uvažovať napr. Weibullovo rozdelenie, po častiach konštantný predpis, alebo model s použitím B-splajnov, ktorý zadefinujeme nižšie.

[1]Vo všetkých prípadoch pod pojmom „veľké“ požadujeme, aby aposteriórny rozptyl parametra bol rádovo menší ako apriórny rozptyl.



Obr. 2.1: Orientovaný acyklický graf pre hierarchický združený model so spoločnými náhodnými efektami.

Ďalej je potrebné odvodiť plne podmienené rozdelenia zo združeného rozdelenia (2.8). Najskôr odvodíme plne podmienené rozdelenia pre model pre longitudinálne dáta a následne pre model prežitia. Združené aposteriórne rozdelenie $(\boldsymbol{\theta},\mathbf{b})^\top$, kde $\mathbf{b} = (\mathbf{b_1},\ldots,\mathbf{b_K})^\top$ dostaneme rozpisom:

$$\begin{aligned}
p(\boldsymbol{\theta},\mathbf{b}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}) &\propto p(\mathbf{y},\mathbf{t},\boldsymbol{\delta}|\boldsymbol{\theta},\mathbf{b}) = p(\mathbf{y}|\boldsymbol{\theta}_Y,\mathbf{b})p(\mathbf{t},\boldsymbol{\delta}|\boldsymbol{\theta}_T,\boldsymbol{\beta},\mathbf{b})p(\mathbf{b}|\boldsymbol{\theta}_b)p(\boldsymbol{\theta}) \\
&= \prod_{i=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}_Y,\mathbf{b}_i)p(t_i,\delta_i|\boldsymbol{\theta}_T,\boldsymbol{\beta},\mathbf{b}_i)p(\mathbf{b}_i|\boldsymbol{\theta}_b)p(\boldsymbol{\theta}_Y)p(\boldsymbol{\theta}_T)p(\boldsymbol{\theta}_b) \\
&\propto \prod_{i=1}^{K} \left(\frac{\tau}{2\pi}\right)^{\frac{n_i}{2}} \exp^{-\frac{\tau}{2}(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)} \\
&\times \left[\lambda_0(t_i,\boldsymbol{\xi})e^{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(t_i)+\alpha[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta}+\mathbf{Z}_i(t_i)^\top\mathbf{b}_i]}\right]^{\delta_i} \\
&\times e^{-\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\
&\times \frac{1}{(2\pi)^{q/2}|\mathbb{D}|^{1/2}} e^{-\frac{1}{2}\mathbf{b_i}^\top\mathbb{D}^{-1}\mathbf{b_i}} p(\boldsymbol{\beta})p(\tau)p(\mathbb{D}^{-1})p(\boldsymbol{\xi})p(\boldsymbol{\gamma})p(\alpha). \qquad (2.10)
\end{aligned}$$

Ako prvé odvodíme plne podmienené rozdelenie vektoru parametrov $\boldsymbol{\beta}$, pričom rozdelenie stačí poznať až na normujúcu konštantu. Pre apriórne rozdelenie $\boldsymbol{\beta}$ predpokladáme $\boldsymbol{\beta} \sim \mathsf{N}_p(\mathbf{0},\boldsymbol{\Sigma}_\beta)$. Využitím apriórnej nezávislosti, nezávislosti $\boldsymbol{\beta}$ a $\mathbf{b}$ a vynechaním členov, ktoré nezávisia na $\boldsymbol{\beta}$ dostaneme:

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} \exp^{-\frac{\tau}{2}(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)} \left[e^{\alpha[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta}+\mathbf{Z}_i(t_i)^\top\mathbf{b}_i]}\right]^{\delta_i} \\
&\times e^{-\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} e^{-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}} \\
&\propto \prod_{i=1}^{K} \exp^{-\frac{\tau}{2}(-\mathbf{y}_i^\top\mathbb{X}_i\boldsymbol{\beta}-(\mathbb{X}_i\boldsymbol{\beta})^\top\mathbf{y}_i+(\mathbb{X}_i\boldsymbol{\beta})^\top(\mathbb{X}_i\boldsymbol{\beta})+(\mathbb{X}_i\boldsymbol{\beta})^\top\mathbb{Z}_i\mathbf{b}_i+(\mathbb{Z}_i\mathbf{b}_i)^\top\mathbb{X}_i\boldsymbol{\beta})} \\
&\times e^{\alpha\delta_i\mathbf{X}_i(t_i)^\top\boldsymbol{\beta}} e^{-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}} e^{-\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\
&\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{K}-2\tau(\mathbb{X}_i\boldsymbol{\beta})^\top(\mathbf{y}_i-\mathbb{Z}_i\mathbf{b}_i)+\tau\sum_{i=1}^{K}(\mathbb{X}_i\boldsymbol{\beta})^\top(\mathbb{X}_i\boldsymbol{\beta})+\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}\right]} \\
&\times e^{\sum_{i=1}^{K}\alpha\delta_i\mathbf{X}_i(t_i)^\top\boldsymbol{\beta}-\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\
&\propto e^{-\frac{1}{2}\left[\boldsymbol{\beta}^\top\left(\Sigma_\beta^{-1}+\tau\sum_{i=1}^{K}\mathbb{X}_i^\top\mathbb{X}_i\right)\boldsymbol{\beta}-2\boldsymbol{\beta}^\top\sum_{i=1}^{K}\left(\tau\mathbb{X}_i^\top(\mathbf{y}_i-\mathbb{Z}_i\mathbf{b}_i)-\alpha\delta_i\mathbf{X}_i(t_i)\right)\right]} \\
&\times e^{-\sum_{i=1}^{K}\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du}. \qquad (2.11)
\end{aligned}
$$

Matica $\boldsymbol{\Sigma}_\beta^{-1}$ je symetrická pozitívne definitná, keďže sa jedná o variančnú maticu. Za predpokladu plnej stĺpcove hodnosti matice $\mathbb{X}_i$ je aj matica $\mathbb{X}_i^\top\mathbb{X}_i$ symetrická pozitívne definitná (SPD). Keďže súčet SPD matíc je opäť SPD matica, dostávame, že $\mathbb{V}_\beta = \Sigma_\beta^{-1} + \tau\sum_{i=1}^{K}\mathbb{X}_i^\top\mathbb{X}_i$ je SPD matica. Môžeme teda písať $\mathbb{V}_\beta = \mathbb{V}_\beta^{1/2}\mathbb{V}_\beta^{1/2}$. Ďalej označíme $\mathbf{U}_\beta = \sum_{i=1}^{K}\tau\mathbb{X}_i^\top(\mathbf{y}_i - \mathbb{Z}_i\mathbf{b}_i) - \alpha\delta_i\mathbf{X}_i(t_i)$ a $S_\Lambda = -\sum_{i=1}^{K}\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du$. Potom platí

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto e^{-\frac{1}{2}[(\mathbb{V}_\beta^{1/2}\boldsymbol{\beta}-\mathbb{V}_\beta^{-1/2}\mathbf{U}_\beta)^\top(\mathbb{V}_\beta^{1/2}\boldsymbol{\beta}-\mathbb{V}_\beta^{-1/2}\mathbf{U}_\beta)-\mathbf{U}_\beta^\top\mathbb{V}_\beta^{-1}\mathbf{U}_\beta]+S_\Lambda} \\
&\propto e^{-\frac{1}{2}[(\boldsymbol{\beta}-\mathbb{V}_\beta^{-1/2}\mathbb{V}_\beta^{-1/2}\mathbf{U}_\beta)^\top\mathbb{V}_\beta^{1/2}\mathbb{V}_\beta^{1/2}(\boldsymbol{\beta}-\mathbb{V}_\beta^{-1/2}\mathbb{V}_\beta^{-1/2}\mathbf{U}_\beta)]+S_\Lambda} \\
&\propto e^{-\frac{1}{2}[(\boldsymbol{\beta}-\mathbb{V}_\beta^{-1}\mathbf{U}_\beta)^\top\mathbb{V}_\beta(\boldsymbol{\beta}-\mathbb{V}_\beta^{-1}\mathbf{U}_\beta)]+S_\Lambda}. \qquad (2.12)
\end{aligned}
$$

Ak by sa v predpise (2.12) nenachádzal člen S_Λ, tak by platilo aposteriórne rozdelenie $\boldsymbol{\beta}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\cdots \propto \mathsf{N}_p(\mathbb{V}_\beta^{-1}\mathbf{U}_\beta,\mathbb{V}_\beta)$. V našom prípade sa však nejedná o žiadne známe rozdelenie, čo motivuje použitie Metropolisovho-Hastingsovho algoritmu na odhadovanie parametrov $\boldsymbol{\beta}$.

Ako ďalšie odvodíme plne podmienené rozdelenie parametra τ, pre ktorý predpokladáme apriórne rozdelenie $\tau \sim \Gamma(a_\tau,b_\tau)$. Keďže sa tento parameter nenachádza v predpise pre model prežitia, výpočet sa výrazne zjednoduší.

$$
\begin{aligned}
p(\tau|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}_Y,\mathbf{b}_i)p(\tau) \\
&\propto \prod_{i=1}^{K}\left(\frac{\tau}{2\pi}\right)^{\frac{n_i}{2}} \exp^{-\frac{\tau}{2}(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)} \tau^{a_\tau-1}e^{-b_\tau\tau} \\
&\propto \tau^{\sum_{i=1}^{K}\frac{n_i}{2}+a_\tau-1} e^{-\tau\left[\sum_{i=1}^{K}(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)+b_\tau\right]} \qquad (2.13)
\end{aligned}
$$

Zo vzťahu $\sum_{i=1}^{K}n_i = n$ a označením $b^* = \sum_{i=1}^{K}(\mathbf{y}_i - \mathbb{X}_i\boldsymbol{\beta} - \mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i - \mathbb{X}_i\boldsymbol{\beta} - \mathbb{Z}_i\mathbf{b}_i) + b_\tau$ dostávame

$$p(\tau|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto e^{\frac{n}{2}+a_\tau-1}e^{-\tau b_\tau^*}, \tag{2.14}$$

teda $\tau|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\cdots \propto \Gamma(\frac{n}{2}+a_\tau,b_\tau^*)$. Oba parametre sú kladné, keďže a_τ je parameter gamma rozdelenia, $\frac{n}{2}>0$ a b_τ^* je súčet štvorcov. Na generovanie markovských reťazcov môžeme použiť Gibbsov algoritmus.

Vektor náhodných efektov $\mathbf{b}$ vystupuje ako v modeli pre longitudinálne dáta, tak aj v modeli prežitia. Výpočty tak budú analogické ako pre parameter $\boldsymbol{\beta}$.

$$\begin{aligned}
p(\mathbf{b}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} p(\mathbf{y}_i|\mathbf{b}_i,\boldsymbol{\theta})p(t_i,\delta_i|\mathbf{b}_i,\boldsymbol{\theta})p(\mathbf{b}_i|\boldsymbol{\theta}) \\
&\propto \prod_{i=1}^{K} \exp^{-\frac{\tau}{2}(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta}-\mathbb{Z}_i\mathbf{b}_i)} e^{\alpha\delta_i[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta}+\mathbf{Z}_i(t_i)^\top\mathbf{b}_i]} \\
&\times e^{-\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du}|\mathbb{D}|^{-1/2}e^{-\frac{1}{2}\mathbf{b_i}^\top\mathbb{D}^{-1}\mathbf{b_i}} \\
&\propto \prod_{i=1}^{K} e^{\frac{\tau}{2}[\mathbf{b_i}^\top\mathbb{Z}_i^\top\mathbb{Z}_i\mathbf{b_i}-2\mathbf{b_i}^\top(\mathbb{Z}_i^\top\mathbf{y}_i-\mathbb{Z}_i^\top\mathbb{X}_i\boldsymbol{\beta}-\alpha\delta_i\mathbf{Z}_i(t_i))]}e^{-\frac{1}{2}\mathbf{b_i}^\top\mathbb{D}^{-1}\mathbf{b_i}} \\
&\times e^{-\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\
&\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{K}-2\mathbf{b}_i^\top\{\tau\mathbb{Z}_i^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta})-\alpha\delta_i\mathbf{Z}_i(t_i)\}+\mathbf{b}_i^\top(\tau\mathbb{Z}_i^\top\mathbb{Z}_i+\mathbb{D}^{-1})\mathbf{b}_i\right]+S_\Lambda}.
\end{aligned} \tag{2.15}$$

Označíme $\mathbf{U}_{b_i}=\tau\mathbb{Z}_i^\top(\mathbf{y}_i-\mathbb{X}_i\boldsymbol{\beta})-\alpha\delta_i\mathbf{Z}_i(t_i)$ a $\mathbb{V}_{b_i}=\tau\mathbb{Z}_i^\top\mathbb{Z}_i+\mathbb{D}^{-1}$. Ak má matica regresorov pre náhodné efekty $\mathbb{Z}_i$ plnú stĺpcovú hodnosť, matica $\tau\mathbb{Z}_i^\top\mathbb{Z}_i$ je symetrická pozitívne definitná (SPD), matica $\mathbb{D}^{-1}$ je variančná matica, teda SPD. Ich súčet je tak tiež SPD matica a môžeme písať $\mathbb{V}_{b_i}=\mathbb{V}_{b_i}^{1/2}\mathbb{V}_{b_i}^{1/2}$. Dostávame tak

$$\begin{aligned}
p(\mathbf{b}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{K}\mathbf{b}_i^\top\mathbb{V}_{b_i}\mathbf{b}_i-2\mathbf{b}_i^\top\mathbf{U}_{b_i}\right]+S_\Lambda} \\
&\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{K}(\mathbb{V}_{b_i}^{1/2}\mathbf{b}_i-\mathbb{V}_{b_i}^{-1/2}\mathbf{U}_{b_i})^\top(\mathbb{V}_{b_i}^{1/2}\mathbf{b}_i-\mathbb{V}_{b_i}^{-1/2}\mathbf{U}_{b_i})\right]+S_\Lambda} \\
&\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{K}(\mathbf{b}_i-\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i})^\top\mathbb{V}_{b_i}(\mathbf{b}_i-\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i})\right]+S_\Lambda} \\
&\propto e^{-\frac{1}{2}(\mathbf{b}-\mathbb{V}_b^{-1}\mathbf{U}_b)^\top\mathbb{V}_b(\mathbf{b}-\mathbb{V}_b^{-1}\mathbf{U}_b)+S_\Lambda},
\end{aligned} \tag{2.16}$$

kde $\mathbf{U}_b=(\mathbf{U}_{b_1}^\top,\dots,\mathbf{U}_{b_K}^\top)^\top$ a $\mathbb{V}_b$ je blokovo diagonálna matica s blokmi $\mathbb{V}_{b_i}$, $i=1,\dots,K$. Pre jednotlivé $\mathbf{b}_i$ a $\Lambda_i=\int_0^{t_i}\lambda_0(u,\boldsymbol{\xi})e^{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]}du$ tak dostávame

$$p(\mathbf{b}_i|\mathbf{y}_i,t_i,\delta_i,\dots) \propto e^{-\frac{1}{2}(\mathbf{b}_i-\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i})^\top\mathbb{V}_{b_i}(\mathbf{b}_i-\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i})-\Lambda_i}.$$

Opäť by sa jednalo o normálne rozdelenie $\mathbf{b}_i|\mathbf{y}_i,t_i,\delta_i,\cdots\propto\mathsf{N}_q(\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i},\mathbb{V}_{b_i})$, ak by sa vo výraze nenachádzal člen Λ_i.

Variančná matica vektoru náhodných efektov $\mathbb{D}$ vystupuje iba v predpise pre podmienené rozdelenie náhodných efektov a v apriórnom rozdelení $\mathbb{D}^{-1} \sim \mathsf{Wish}_q(\nu_D, \Upsilon)$. Predpis hustoty tohto rozdelenia je uvedený v prílohe A.1. Plne podmienené rozdelenie jednoducho odvodíme ako:

$$\begin{aligned} p(\mathbb{D}^{-1}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} p(\mathbf{b}_i|\mathbb{D}^{-1})p(\mathbb{D}^{-1}) \\ &\propto \prod_{i=1}^{K} |\mathbb{D}|^{-1/2} e^{-\frac{1}{2}\mathbf{b}_i^\top \mathbb{D}^{-1}\mathbf{b}_i} |\mathbb{D}^{-1}|^{\frac{\nu_d-q-1}{2}} e^{-\frac{\mathrm{tr}(\Upsilon^{-1}\mathbb{D}^{-1})}{2}} \\ &\propto e^{-\frac{1}{2}\left(\sum_{i=1}^{K} \mathbf{b}_i^\top \mathbb{D}^{-1}\mathbf{b}_i + \mathrm{tr}(\Upsilon^{-1}\mathbb{D}^{-1})\right)} |\mathbb{D}^{-1}|^{\frac{\nu_d-q+K-1}{2}}. \end{aligned} \tag{2.17}$$

Prepísaním

$$\begin{aligned} &\sum_{i=1}^{K} \mathbf{b}_i^\top \mathbb{D}^{-1}\mathbf{b}_i + \mathrm{tr}(\Upsilon^{-1}\mathbb{D}^{-1}) = \sum_{i=1}^{K} \mathrm{tr}(\mathbf{b}_i^\top \mathbb{D}^{-1}\mathbf{b}_i) + \mathrm{tr}(\Upsilon^{-1}\mathbb{D}^{-1}) \\ &= \sum_{i=1}^{K} \mathrm{tr}(\mathbf{b}_i\mathbf{b}_i^\top \mathbb{D}^{-1}) + \mathrm{tr}(\Upsilon^{-1}\mathbb{D}^{-1}) = \mathrm{tr}\Big(\Big[\sum_{i=1}^{K} \mathbf{b}_i\mathbf{b}_i^\top + \Upsilon^{-1}\Big]\mathbb{D}^{-1}\Big) \end{aligned}$$

a označením $\mathbb{V}_D^{-1} = \sum_{i=1}^{K} \mathbf{b}_i\mathbf{b}_i^\top + \Upsilon^{-1}$ získame

$$p(\mathbb{D}^{-1}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto |\mathbb{D}^{-1}|^{\frac{\nu_d-q+K-1}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(\mathbb{V}_D^{-1}\mathbb{D}^{-1}\right)}, \tag{2.18}$$

teda plne podmienené rozdelenie parametra $\mathbb{D}^{-1}$ je Wishartovo rozdelenie s parametrami $\mathbb{V}_D$ a $\nu_d - q + K$, $\mathbb{D}^{-1}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\cdots \propto \mathsf{Wish}_q(\nu_d - q + K, \mathbb{V}_D)$.

Ďalej odvodíme plne podmienené rozdelenia parametrov z modelu prežitia. Parameter α vyjadruje závislosť rizika udalosti na endogénnej časovo závislej premennej. Použitím predpokladu normálneho apriórneho rozdelenia $\alpha \sim \mathsf{N}(0, \sigma_\alpha^2)$ rozpíšeme jeho aposteriórne rozdelenie ako:

$$\begin{aligned} p(\alpha|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} \left[e^{\alpha[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t_i)^\top\mathbf{b}_i]}\right]^{\delta_i} (2\pi\sigma_\alpha^2)^{-1/2} e^{-\frac{\alpha^2}{2\sigma_\alpha^2}} \\ &\times e^{-\int_0^{t_i} \lambda_0(u,\boldsymbol{\xi}) \exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\ &\propto e^{\sum_{i=1}^{K} \alpha\delta_i[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t_i)^\top\mathbf{b}_i] - \frac{\alpha^2}{2\sigma_\alpha^2} + S_\Lambda} \\ &\propto e^{-\frac{1}{2}\left[\frac{\alpha^2}{\sigma_\alpha^2} - 2\alpha\sum_{i=1}^{K} \delta_i[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t_i)^\top\mathbf{b}_i\right] + S_\Lambda}, \end{aligned} \tag{2.19}$$

kde $S_\Lambda = -\sum_{i=1}^{K} \int_0^{t_i} \lambda_0(u,\boldsymbol{\xi}) \exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du$ je značenie z predchádzajúcich výpočtov. Keď označíme $U_\alpha = \sum_{i=1}^{K} \delta_i[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t_i)^\top\mathbf{b}_i\Big]$, dostaneme zjednodušený predpis

$$p(\alpha|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto e^{-\frac{1}{2}\left(\frac{\alpha-\sigma_\alpha U_\alpha}{\sigma_\alpha}\right)^2 + S_\Lambda}. \tag{2.20}$$

Opäť vidíme, že ak by sa v predpise nenachádzal člen S_Λ, jednalo by sa o normálne rozdelenie $\mathsf{N}(\sigma_\alpha U_\alpha,\sigma_\alpha^2)$. Aj v prípade tohto parametra teda bude potrebný Metropolisov-Hastingsov algoritmus.

Pri odvodení aposteriórneho rozdelenia parametra $\boldsymbol{\gamma}$ budeme postupovať analogicky ako pre parameter $\boldsymbol{\beta}, \alpha$ a vektor náhodných efektov $\mathbf{b}$:

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} \left[\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(t_i)\right]^{\delta_i} |\boldsymbol{\Sigma}_\gamma|^{-1/2} e^{-\frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}} \\ &\quad \times e^{-\int_0^{t_i} \lambda_0(u,\boldsymbol{\xi}) \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\ &\quad \propto e^{\sum_{i=1}^{K} \delta_i \boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(t_i) - \frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}+S_\Lambda} \\ &\quad \propto e^{-\frac{1}{2}\left[\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma} - 2\sum_{i=1}^{K} \delta_i \boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(t_i)\right]+S_\Lambda}. \end{aligned} \tag{2.21}$$

Označením $\mathbf{U}_\gamma = \sum_{i=1}^{K} \delta_i \tilde{\mathbf{X}}_i(t_i)$ dostaneme

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto e^{-\frac{1}{2}(\boldsymbol{\Sigma}_\gamma^{-1/2}\boldsymbol{\gamma}-\boldsymbol{\Sigma}_\gamma^{1/2}\mathbf{U}_\gamma)^\top(\boldsymbol{\Sigma}_\gamma^{-1/2}\boldsymbol{\gamma}-\boldsymbol{\Sigma}_\gamma^{1/2}\mathbf{U}_\gamma)+S_\Lambda} \\ &\propto e^{-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\Sigma}_\gamma\mathbf{U}_\gamma)^\top\boldsymbol{\Sigma}_\gamma^{-1}(\boldsymbol{\gamma}-\boldsymbol{\Sigma}_\gamma\mathbf{U}_\gamma)+S_\Lambda}, \end{aligned} \tag{2.22}$$

čo vedie opäť na použitie Metropolisovho-Hastingsovho algoritmu.

Ako posledné zostáva určiť plne podmienené rozdelenie základného rizika $\lambda_0(t,\boldsymbol{\xi})$, resp. jeho parametrov $\boldsymbol{\xi}$. Ako sme načrtli vyššie, najčastejšie používané základné riziká sú po častiach konštantné základné riziko, riziko z Weibullovho rozdelenia $\mathsf{Weibull}(\xi_1,\xi_2)$, $\xi_1,\xi_2 > 0$ a riziko modelované regresnými splajnami. Aposteriórne rozdelenie odvodíme pre všetky 3 tvary.

Riziková funkcia Weibullovho rozdelenia $\mathsf{Weibull}(\xi_1,\xi_2)$, $\xi_1,\xi_2 > 0$ (predpis hustoty možno nájsť v prílohe A.1) má predpis $\lambda_0(t,\xi_1,\xi_2) = \xi_1\xi_2 t^{\xi_1-1}$. Pre oba parametre ξ_1 a ξ_2 budeme predpokladať apriórne rozdelenie $\xi_1 \sim \Gamma(a_{\xi_1},b_{\xi_1})$, $a_{\xi_1},b_{\xi_1} > 0$ a $\xi_2 \sim \Gamma(a_{\xi_2},b_{\xi_2})$, $a_{\xi_2},b_{\xi_2} > 0$, analogicky ako pre parameter τ. Aposteriórne rozdelenia parametra ξ_1 dostaneme využitím apriórnej nezávislosti ako:

$$\begin{aligned} p(\xi_1|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} [\xi_1\xi_2 t_i^{\xi_1-1}]^{\delta_i} e^{-\int_0^{t_i} \xi_1\xi_2 u^{\xi_1-1} \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du} \\ &\quad \times \xi_1^{a_{\xi_1}-1} e^{-b_{\xi_1}\xi_1} \\ &\quad \propto \prod_{i=1}^{K} \xi_1^{\delta_i+a_{\xi_1}-1} e^{\delta_i(\xi_1-1)\log t_i - \xi_1 I_{i\xi_1} - \xi_1 b_{\xi_1}}, \end{aligned} \tag{2.23}$$

kde sme označili $I_{i\xi_1} = \int_0^{t_i} \xi_2 u^{\xi_1-1} \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du$. Ďalším upravovaním vyjde:

$$p(\xi_1|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto \xi_1^{\sum_{i=1}^{K} \delta_i+a_{\xi_1}-1} e^{\xi_1(\sum_{i=1}^{K} \delta_i \log t_i - I_{i\xi_1} - b_{\xi_1})}. \tag{2.24}$$

Plne podmienené rozdelenie tohto parametra opäť nepatrí do žiadnej známej rodiny rozdelení, čo vedie na použitie Metropolisovho-Hastingsovho algoritmu. Pre odvodenie aposteriórneho rozdelenia parametra ξ_2 budeme postupovať analogicky a označíme $I_{i\xi_2} = \int_0^{t_i} \xi_1 u^{\xi_1 - 1} \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top \boldsymbol{\beta} + \mathbf{Z}_i(u)^\top \mathbf{b}_i]\} du$. Potom platí

$$\begin{aligned} p(\xi_2|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} [\xi_1 \xi_2 t_i^{\xi_1 - 1}]^{\delta_i} e^{-\int_0^{t_i} \xi_1 \xi_2 u^{\xi_1 - 1} \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top \boldsymbol{\beta} + \mathbf{Z}_i(u)^\top \mathbf{b}_i]\} du} \\ &\times \xi_2^{a_{\xi_2} - 1} e^{-b_{\xi_2} \xi_2} \\ &\propto \prod_{i=1}^{K} \xi_2^{\delta_i} e^{-\xi_2 I_{\xi_2}} \xi_2^{a_{\xi_2} - 1} e^{-b_{\xi_2} \xi_2} \propto \xi_2^{\sum_{i=1}^{K} \delta_i + a_{\xi_2} - 1} e^{-\xi_2 (\sum_{i=1}^{K} I_{i\xi_2} + b_{\xi_2})}. \end{aligned} \tag{2.25}$$

Z predpisu vidíme, že plne podmienené rozdelenie parametra ξ_2 je gama rozdelenie $\Gamma(\sum_{i=1}^{K} \delta_i + a_{\xi_2}, \sum_{i=1}^{K} I_{i\xi_2} + b_{\xi_2})$, ak sú oba parametre kladné. Člen $\sum_{i=1}^{K} \delta_i$ je nezáporný a parameter a_{ξ_2} je parameter gama rozdelenia, teda kladný. Súčet $\sum_{i=1}^{K} \delta_i + a_{\xi_2}$ je tak kladný. Člen $I_{i\xi_2}$ predstavuje kumulatívnu rizikovú funkciu $\frac{\Lambda_i(T_i^*|\boldsymbol{\theta})}{\xi_2}$, ktorá je nezáporná a parameter ξ_2 je kladný. Parameter b_{ξ_2} je kladný parameter gama rozdelenia. Celkovo je teda súčet $\sum_{i=1}^{K} I_{i\xi_2} + b_{\xi_2}$ kladný.

Po častiach konštantná základná riziková funkcia je daná predpisom $\lambda_0(t,\boldsymbol{\lambda}) = \sum_{r=1}^{R} \lambda_r \mathbb{1}\{t \in (D_{r-1},D_r]\}$, kde $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_R)^\top$ je vektor konštantných rizík a $0 < D_1 < \cdots < D_R$ je konečné delenie časovej osi tak, že $D_R > T_i$ pre všetky $i = 1,2,\dots,K$. Pre zložky vektoru $\boldsymbol{\lambda}$ budeme predpokladať nezávislé apriórne rozdelenie gama rozdelenie $\Gamma(a_{\lambda_r},b_{\lambda_r})$, $a_{\lambda_r},b_{\lambda_r} > 0$ [Ibrahim and Chen, 2001], teda

$$p(\boldsymbol{\lambda}) = \left(\frac{b_{\lambda_r}^{a_{\lambda_r}}}{\Gamma(a_{\lambda_r})} \right)^R \prod_{r=1}^{R} \lambda_r^{a_{\lambda_r} - 1} e^{-b_{\lambda_r} \lambda_r}.$$

Pre každé λ_r, $r = 1,\dots R$ tak dostávame plne podmienené rozdelenie:

$$\begin{aligned} p(\lambda_r|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} \Big[\sum_{j=1}^{R} \lambda_j \mathbb{1}\{t_i \in (D_{j-1},D_j]\} \Big]^{\delta_i} \lambda_r^{a_{\lambda_r} - 1} e^{-b_{\lambda_r} \lambda_r} \\ &\times e^{-\int_0^{t_i} \sum_{j=1}^{R} \lambda_j \mathbb{1}\{u \in (D_{j-1},D_j]\} \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top \boldsymbol{\beta} + \mathbf{Z}_i(u)^\top \mathbf{b}_i]\} du} \\ &\propto \lambda_r^{\sum_{i=1}^{K} \delta_i + a_{\lambda_r} - 1} e^{-\lambda_r b_{\lambda_r} + I_\lambda}, \end{aligned} \tag{2.26}$$

kde $I_\lambda = \sum_{i=1}^{K} \int_0^{t_i} \sum_{j=1}^{R} \lambda_j \mathbb{1}\{u \in (D_{j-1},D_j]\} e^{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top \boldsymbol{\beta} + \mathbf{Z}_i(u)^\top \mathbf{b}_i]} du$. Vidíme, že ak by bol člen I_λ nulový, jednalo by sa o gamma rozdelenie $\Gamma(\sum_{i=1}^{K} \delta_i + a_{\lambda_r},b_{\lambda_r})$. Opäť nemáme uzavretú formu rozdelenia a využijeme Metropolisov-Hastingsov algoritmus.

V knihe [Rizopoulos, 2012] je základné riziko modelované regresnými splajnami a tento model je následne implementovaný aj v balíku JMbayes [Rizopoulos, 2016], ktorý budeme využívať pri aplikovaní teórie. V modeli je logaritmus základného rizika vyjadrený ako

$$\log(\lambda_0(t,\boldsymbol{\xi})) = \xi_0 + \sum_{j=1}^{J} \xi_j B_j(t,\mathbf{v}), \tag{2.27}$$

kde $\boldsymbol{\xi} = (\xi_0,\dots,\xi_J)^\top$ je vektor splajnových koeficientov a $B_1(t,\mathbf{v}),\dots,B_J(t,\mathbf{v})$ je splajnová báza stupňa d s uzlami $\mathbf{v} = (v_1,\dots,v_{J-d+1})$. Podrobnú a pomerne rozsiahlu definíciu bázického splajnu a splajnovej bázy možno nájsť v [Komárek, 2021] a v práci ju nebudeme uvádzať. Zvyšovanie počtu uzlov $J-d+1$ umožňuje väčšiu flexibilitu v aproximácii $\log(\lambda_0(t,\boldsymbol{\xi}))$, no na druhej strane sa chceme vyhnúť problému presného prekladania regresnej krivky dátami (anglicky *overfitting*). Používané pravidlo palca je zachovať celkový počet parametrov v modeli (t.j. dimenziu vektoru $\boldsymbol{\theta}$) medzi 1/10 a 1/20 celkového počtu udalostí v dátach [Harrell, 2001, Kapitola 4]. Po zvolení počtu uzlov treba zvoliť ich polohu. Tá môže byť volená na základe percentilov cenzorovaných časov udalosti alebo pozorovaných skutočných časov udalosti. Alternatívna možnosť, pri ktorej sa vyhneme výberu vhodného počtu a umiestnenia uzlov, je zahrnúť pomerne veľký počet uzlov (napr. 15 až 20) a vhodne penalizovať regresné koeficienty B-splajnu [Eilers and Marx, 1996]. Pre zložky vektoru $\boldsymbol{\xi}$ je za predpokladu apriórnej nezávislosti štandardnou voľbou apriórneho rozdelenia jednorozmerné slabo informatívne normálne rozdelenie, t.j. $\xi_j \sim \mathsf{N}(0,\sigma^2_{\xi_j})$ pre $j = 0,\dots,J$. Pre $l \in \{1,\dots,J\}$ tak dostaneme plne podmienené rozdelenie:

$$\begin{aligned}
p(\xi_l|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) &\propto \prod_{i=1}^{K} [e^{\xi_0+\sum_{j=1}^{J}\xi_j B_j(t_i,\mathbf{v})}]^{\delta_i} e^{-\frac{1}{2}\xi_l^2/\sigma^2_{\xi_l}} \\
&\quad \times e^{-\int_0^{t_i} e^{\xi_0+\sum_{j=1}^{J}\xi_j B_j(t_i,\mathbf{v})+\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]}du} \\
&\propto e^{\xi_l\sum_{i=1}^{K}\delta_i B_l(t_i,\mathbf{v})-\frac{1}{2}\xi_l^2/\sigma^2_{\xi_l}+I_\xi} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{\xi_l-\sigma_{\xi_l}\sum_{i=1}^{K}\delta_i B_l(t_i,\mathbf{v})}{\sigma_{\xi_l}}\right]^2 + I_\xi\right\},
\end{aligned} \tag{2.28}$$

kde sme označili $I_\xi = -\sum_{i=1}^{K} e^{-\int_0^{t_i} e^{\xi_0+\sum_{j=1}^{J}\xi_j B_j(t_i,\mathbf{v})+\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]}du}$. Aj v prípade tohto aposteriórneho rozdelenia bude treba využiť Matropolisov-Hastingsov algoritmus, keďže sa nejedná o žiadne známe rozdelenie. Pre ξ_0 dostaneme analogickým výpočtom

$$p(\xi_0|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto \exp\left\{-\frac{1}{2}\left[\frac{\xi_0-\sigma_{\xi_0}\sum_{i=1}^{K}\delta_i}{\sigma_{\xi_0}}\right]^2 + I_\xi\right\}. \tag{2.29}$$

K voľbe návrhovej hustoty v Metropolisovom-Hastingsovom algoritme možno pristupovať viacerými spôsobmi. Jednou z možných volieb návrhovej hustoty je náhodná prechádzka s vhodnou voľbou rozptylu, resp. variančnej matice. V značení z kapitoly 1.3.2 v kroku $(m+1)$ je pre dané $\boldsymbol{\theta}^{(m)}$ návrh $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)}+\mathbf{Z}$, kde častou voľbou rozdelenia $\mathbf{Z}$ býva (viacrozmerné) t-rozdelenie, resp. normálne rozdelenie s nulovou strednou hodnotou a zvyčajne diagonálnou variančnou maticou.

Na odhad variančnej matice sa využíva Laplaceova transformácia [Penny et al., 2007]. Označme $\tilde{\boldsymbol{\kappa}}_0 = argmax_\Theta \log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}))$ a použijeme Taylorov rozvoj 2. rádu na logaritmus aposteriórnej hustoty:

$$\log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta})) \approx \log(p(\tilde{\boldsymbol{\kappa}}_0|\mathbf{y},\mathbf{t},\boldsymbol{\delta})) + (\boldsymbol{\kappa}-\tilde{\boldsymbol{\kappa}}_0)^\top [\nabla \log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}))]|_{\boldsymbol{\kappa}=\tilde{\boldsymbol{\kappa}}_0} + \frac{1}{2}(\boldsymbol{\kappa}-\tilde{\boldsymbol{\kappa}}_0)^\top [\nabla^2 \log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}))]_{|\boldsymbol{\kappa}=\tilde{\boldsymbol{\kappa}}_0}(\boldsymbol{\kappa}-\tilde{\boldsymbol{\kappa}}_0). \tag{2.30}$$

Platí $\nabla \log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}))]|_{\boldsymbol{\kappa}=\tilde{\boldsymbol{\kappa}}_0} = 0$ a použitím exponenciály dostaneme:

$$p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}) \approx p(\tilde{\boldsymbol{\kappa}}_0|\mathbf{y},\mathbf{t},\boldsymbol{\delta}) \cdot \exp\big(-\frac{1}{2}(\boldsymbol{\kappa}-\tilde{\boldsymbol{\kappa}}_0)^\top [-\nabla^2 \log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}))]_{|\boldsymbol{\kappa}=\tilde{\boldsymbol{\kappa}}_0}(\boldsymbol{\kappa}-\tilde{\boldsymbol{\kappa}}_0))\big). \tag{2.31}$$

Matica $[-\nabla^2 \log(p(\boldsymbol{\kappa}|\mathbf{y},\mathbf{t},\boldsymbol{\delta}))]_{|\boldsymbol{\kappa}=\tilde{\boldsymbol{\kappa}}_0}$ sa následne použije ako odhad variančnej matice náhodnej prechádzky [van der Vaart, 1998, Veta Bernstein–von Mises].

Predchádzajúce odvodenia zhrnieme do nasledujúceho tvrdenia.

Tvrdenie 2. *Nech platí združený model 2.1.2 a označme súčet kumulatívnych rizík* $S_\Lambda = -\sum_{i=1}^K \int_0^{t_i} \lambda_0(u,\boldsymbol{\xi}) \exp\{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top \boldsymbol{\beta} + \mathbf{Z}_i(u)^\top \mathbf{b}_i]\} du$. *Potom sú plne podmienené rozdelenia parametrov modelu* $\boldsymbol{\beta},\tau,\mathbb{D}^{-1},\boldsymbol{\xi},\gamma,\alpha$ *a náhodných efektov* $\mathbf{b}_1,\dots,\mathbf{b}_k$ *tvaru:*

1. $$p(\boldsymbol{\beta}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto e^{-\frac{1}{2}[(\boldsymbol{\beta}-\mathbb{V}_\beta^{-1}\mathbf{U}_\beta)^\top \mathbb{V}_\beta (\boldsymbol{\beta}-\mathbb{V}_\beta^{-1}\mathbf{U}_\beta)]+S_\Lambda},$$
   kde $\mathbf{U}_\beta = \sum_{i=1}^K \tau \mathbb{X}_i^\top(\mathbf{y}_i - \mathbb{Z}_i\mathbf{b}_i) - \alpha\delta_i \mathbf{X}_i(t_i)$ *a* $\mathbb{V}_\beta = \Sigma_\beta^{-1} + \tau\sum_{i=1}^K \mathbb{X}_i^\top \mathbb{X}_i$,

2. $$\tau|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\cdots \propto \Gamma(\frac{n}{2}+a_\tau, b_\tau^*),$$
   kde $b^* = \sum_{i=1}^K (\mathbf{y}_i - \mathbb{X}_i\boldsymbol{\beta} - \mathbb{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i - \mathbb{X}_i\boldsymbol{\beta} - \mathbb{Z}_i\mathbf{b}_i) + b_\tau$,

3. *pre jednotlivé* $\mathbf{b}_i$ *je*
   $$p(\mathbf{b}_i|\mathbf{y}_i,t_i,\delta_i,\dots) \propto e^{-\frac{1}{2}(\mathbf{b}_i-\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i})^\top \mathbb{V}_{b_i}(\mathbf{b}_i-\mathbb{V}_{b_i}^{-1}\mathbf{U}_{b_i})-\Lambda_i},$$
   kde $\Lambda_i = \int_0^{t_i} \lambda_0(u,\boldsymbol{\xi}) e^{\boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top \mathbf{b}_i]} du$, $\mathbf{U}_{b_i} = \tau\mathbb{Z}_i^\top(\mathbf{y}_i - \mathbb{X}_i\boldsymbol{\beta}) - \alpha\delta_i\mathbf{Z}_i(t_i)$ *a* $\mathbb{V}_{b_i} = \tau\mathbb{Z}_i^\top\mathbb{Z}_i + \mathbb{D}^{-1}$ *a*
   $$p(\mathbf{b}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto e^{-\frac{1}{2}(\mathbf{b}-\mathbb{V}_b^{-1}\mathbf{U}_b)^\top \mathbb{V}_b(\mathbf{b}-\mathbb{V}_b^{-1}\mathbf{U}_b)+S_\Lambda},$$
   kde $\mathbf{U}_b = (\mathbf{U}_{b_1}^\top,\dots,\mathbf{U}_{b_K}^\top)^\top$ *a* $\mathbb{V}_b$ *je blokovo diagonálna matica s blokmi* $\mathbb{V}_{b_i}$, $i = 1,\dots,K$,

4. $$\mathbb{D}^{-1}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\cdots \propto \mathit{Wish}_q(\nu_d - q + K, \mathbb{V}_D),$$
   kde $\mathbb{V}_D^{-1} = \sum_{i=1}^K \mathbf{b}_i\mathbf{b}_i^\top + \Upsilon^{-1}$

5.

$$p(\alpha|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto e^{-\frac{1}{2}\left(\frac{\alpha-\sigma_\alpha U_\alpha}{\sigma_\alpha}\right)^2+S_\Lambda},$$

pre $U_\alpha = \sum_{i=1}^K \delta_i[\mathbf{X}_i(t_i)^\top\boldsymbol{\beta} + \mathbf{Z}_i(t_i)^\top\mathbf{b}_i\Big]$,

6.

$$p(\boldsymbol{\gamma}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto e^{-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\Sigma}_\gamma\mathbf{U}_\gamma)^\top\boldsymbol{\Sigma}_\gamma^{-1}(\boldsymbol{\gamma}-\boldsymbol{\Sigma}_\gamma\mathbf{U}_\gamma)+S_\Lambda},$$

pre $\mathbf{U}_\gamma = \sum_{i=1}^K \delta_i\tilde{\mathbf{X}}_i(t_i)$,

7. pre základné riziko z Weibulloovho rizika je $\boldsymbol{\xi} = (\xi_1,\xi_2)^\top$,

$$p(\xi_1|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto \xi_1^{\sum_{i=1}^K \delta_i+a_{\xi_1}-1}e^{\xi_1(\sum_{i=1}^K \delta_i\log t_i - I_{i\xi_1}-b_{\xi_1})},$$

pre $I_{i\xi_1} = \int_0^{t_i}\xi_2 u^{\xi_1-1}\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du$,

8.

$$\xi_2|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\cdots \propto \Gamma(\sum_{i=1}^K \delta_i + a_{\xi_2}, \sum_{i=1}^K I_{i\xi_2} + b_{\xi_2}),$$

kde $I_{i\xi_2} = \int_0^{t_i}\xi_1 u^{\xi_1-1}\exp\{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u) + \alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta} + \mathbf{Z}_i(u)^\top\mathbf{b}_i]\}du$,

9. pre po častiach konštantné základné riziko dostávame $\boldsymbol{\xi} = (\lambda_1,\dots,\lambda_R)^\top$,

$$p(\lambda_r|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto \lambda_r^{\sum_{i=1}^K \delta_i+a_{\lambda_r}-1}e^{-\lambda_r b_{\lambda_r}+I_\lambda},\ r=1,\dots,R,$$

kde $I_\lambda = \sum_{i=1}^K\int_0^{t_i}\sum_{j=1}^R \lambda_j\mathbb{1}\{u\in(D_{j-1},D_j]\}e^{\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]}du$,

10. pre základné riziko splňajúce $\lambda_0(t,\boldsymbol{\xi})) = e^{\xi_0+\sum_{j=1}^J\xi_j B_j(t,\mathbf{v})}$ *je* $\boldsymbol{\xi} = (\xi_0,\dots,\xi_J)^\top$ *a pre* $l = 1,\dots,J$ *platí*

$$p(\xi_l|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto \exp\left\{-\frac{1}{2}\left[\frac{\xi_l - \sigma_{\xi_l}\sum_{i=1}^K\delta_i B_l(t_i,\mathbf{v})}{\sigma_{\xi_l}}\right]^2 + I_\xi\right\},$$

a

$$p(\xi_0|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\dots) \propto \exp\left\{-\frac{1}{2}\left[\frac{\xi_0 - \sigma_{\xi_0}\sum_{i=1}^K\delta_i}{\sigma_{\xi_0}}\right]^2 + I_\xi\right\},$$

kde $I_\xi = -\sum_{i=1}^K e^{-\int_0^{t_i} e^{\xi_0+\sum_{j=1}^J\xi_j B_j(t_i,\mathbf{v})+\boldsymbol{\gamma}^\top\tilde{\mathbf{X}}_i(u)+\alpha[\mathbf{X}_i(u)^\top\boldsymbol{\beta}+\mathbf{Z}_i(u)^\top\mathbf{b}_i]}du}$.

Dôkaz. Viď predchádzajúce výpočty.

□

2.2 Združené modely s latentnými kategóriami

Alternatívny prístup k združenému modelovaniu je pomocou združených modelov s latentnými kategóriami (anglicky *Joint Latent Class Model-JLCM*), v ktorých sa berie do úvahy heterogenita populácie. Hlavnou myšlienkou za týmito modelmi je, že populácia je zložená z homogénnych latentných podskupín/kategórií subjektov, ktoré zdieľajú rovnaký model pre longitudinálnu premennú a rovnaké riziko udalosti. Kľúčovým predpokladom pre použitie JLCM je, že podmienene na kategórii, sú longitudinálna premenná a čas do udalosti nezávislé. V porovnaní s modelmi so spoločnými náhodnými efektami je JLCM venovaná menšia pozornosť a aplikácia týchto modelov je zameraná hlavne na popis vývoja ochorení, analýzu citlivosti na chýbajúce dáta a v poslednej dobe na dynamickú predikciu [Proust-Lima et al., 2014].

Buď K počet nezávislých subjektov, ktoré možno rozdeliť do G latentných homogénnych podskupín. Priradenie do latentnej triedy pre každý subjekt $i = 1,\ldots,K$ je definované pomocou kategorickej latentnej premennej V_i, ktorá nadobúda hodnoty $g = 1,\ldots,G$. Pre budovanie tohto typu modelov budeme opäť potrebovať predpoklad (2.4) a predpoklad (2.5), upravený pre kategorickú latentnú premennú [Rizopoulos, 2012, Kapitola 5]:

$$f(\mathbf{Y}_i, T_i,\delta_i) = \sum_{g=1}^{G} f(\mathbf{Y}_i|V_i = g) f(T_i,\delta_i|V_i = g)\, \mathsf{P}(V_i = g). \tag{2.32}$$

Pravdepodobnosť, že i-ty subjekt patrí do skupiny g označíme π_{ig}. Združený model s latentnými kategóriami budujeme v 3 krokoch:

1. multinomická logistická regresia pre π_{ig}:

$$\pi_{ig} = \mathsf{P}[V_i = g|\tilde{\mathbf{Z}}_i] = \frac{\exp\{\boldsymbol{\zeta}_g^T \tilde{\mathbf{Z}}_i\}}{\sum_{g=1}^{G} \boldsymbol{\zeta}_g^\top \tilde{\mathbf{Z}}_i}, \tag{2.33}$$

2. zmiešaný lineárny model pre $\mathbf{Y}_i(t)|_{V_i=g} = (Y_{i,1},\ldots,Y_{i,n_i})^\top|_{V_i=g}$

$$\mathbf{Y}_i(t)|_{V_i=g} = \mathbf{X}_i(t)^\top \boldsymbol{\beta}_g + \mathbf{Z}_i(t)^\top \mathbf{b}_{ig} + \boldsymbol{\epsilon}_i(t),\ i = 1,\ldots K, \tag{2.34}$$

3. model proporčných rizík:

$$\lambda_i(t|V_i = g, \boldsymbol{\xi}_g, \boldsymbol{\gamma}_g) = \lambda_{0g}(t, \boldsymbol{\xi}_g)\exp\{\tilde{\mathbf{X}}_i(t)^\top \boldsymbol{\gamma}_g\}, \tag{2.35}$$

kde

- $\boldsymbol{\zeta}_g = (\zeta_{0,g},\ldots,\zeta_{m-1,g})^T$, $g = 1,\ldots,G$, sú vektory neznámych parametrov, ktoré z dôvodu identifikovateľnosti musia spĺňať podmienku $\boldsymbol{\zeta}_G = \mathbf{0}$,
- $\tilde{\mathbf{Z}}_i$ je m-rozmerný vektor časovo nezávislých regresorov pre i-ty subjekt,
- $\mathbf{X}_i(t)$ p-rozmerný vektor (aj časovo závislých) regresorov pre fixné efekty,
- $\boldsymbol{\beta}_g$ je p-rozmerný vektor fixných efektov,

- $\mathbf{Z}_i(t)$ je q-rozmerný vektor časovo závislých regresorov pre náhodné efekty,
- $\mathbf{b}_{ig}$ je vektor náhodných efektov spĺňajúci $\mathbf{b}_{ig} = \mathbf{b}_i|_{V_i=g} \sim \mathsf{N}_q(\mu_g,\mathbb{D}_g)$, $\mathbf{b}_i \sim \sum_{g=1}^{G} \pi_{ig}\mathsf{N}_q(\mu_g,\mathbb{D}_g)$,
- $\boldsymbol{\epsilon}_i = (\epsilon_{i,1},\ldots,\epsilon_{i,n_i})^\top$ je chybový vektor spĺňajúci $\boldsymbol{\epsilon}_i \sim \mathsf{N}_{n_i}(\mathbf{0},\boldsymbol{\Sigma}_i)$,
- $\tilde{\mathbf{X}}_i(t)$ je r-rozmerný vektor regresorov v čase t,
- $\boldsymbol{\gamma}_g = (\gamma_{0,g},\ldots,\gamma_{r-1,g})^\top$ je vektor neznámych parametrov,
- $\lambda_{0g}(t,\boldsymbol{\xi}_g)$ je základné riziko, špecifické pre každú skupinu, ktoré závisí na vektore neznámych paramerov $\boldsymbol{\xi}_g,\ g = 1,\ldots,G$.

Z dôvodu identifikovateľnosti predpokladáme, že zložky $\mathbf{X}_i(t)$ a $\mathbf{Z}_i(t)$ sú rôzne. Variančná matica $\mathbb{D}_g$ môže byť pre kategórie $g = 1,\ldots,G$ spoločná alebo špecifická pre každú kategóriu. V prípade, že sa jedná o špecifické matice, volí sa zvyčajne $\mathbb{D}_g = \omega_g^2\mathbb{D}$, $\omega_G^2 = 1$, z dôvodu identifikovateľnosti. Variančná matica Σ_i je zvyčajne volená ako diagonálna matica $\sigma^2\mathbb{I}_{n_i}$ pre homoskedastické nezávislé chyby, ale zložky $\boldsymbol{\epsilon}_i$ môžu byť aj korelované.

Označme pre pevný počet latentných kategórií G vektor neznámych parametrov združeného modelu (2.33)–(2.35) ako $\boldsymbol{\theta}_G$. Analogicky, ako pre združené modely so spoločnými náhodnými efektami, môžeme, vďaka predpokladom (2.4), (2.32) a vzťahu $f(T_i,\delta_i|V_i = g) \propto [\lambda_i(T_i|V_i = g)]^{\delta_i}S(T_i|V_i = g)$, rozpísať vierohodnostnú funkciu:

$$\begin{aligned}
L(\boldsymbol{\theta}_G) &= \prod_{i=1}^{K} L_i(\boldsymbol{\theta}_G) = \prod_{i=1}^{K} f(\mathbf{Y}_i, T_i,\delta_i,\boldsymbol{\theta}_G) \\
&= \prod_{i=1}^{K}\{\sum_{g=1}^{G} f(\mathbf{Y}_i|V_i = g,\boldsymbol{\theta}_G)f(T_i,\delta_i|V_i = g,\boldsymbol{\theta}_G)\,\mathsf{P}(Vi = g|\boldsymbol{\theta}_G)\} \\
&\propto \prod_{i=1}^{K}\{\sum_{g=1}^{G} f(\mathbf{Y}_i|V_i = g,\boldsymbol{\theta}_G)[\lambda_i(T_i|V_i = g,\boldsymbol{\theta}_G)]^{\delta_i}S(T_i|V_i = g,\boldsymbol{\theta}_G) \\
&\quad \mathsf{P}(V_i = g,\boldsymbol{\theta}_G)\}. \qquad (2.36)
\end{aligned}$$

Jednotlivé členy vierohodnostnej funkcie (2.36) ďalej rozpíšeme:

$$\begin{aligned}
L_i(\boldsymbol{\theta}_G) = &\sum_{g=1}^{G}\left(\frac{1}{2\pi}\right)^{n_i/2}|\mathbb{Z}_i\mathbb{D}_g\mathbb{Z}_i^\top + \boldsymbol{\Sigma}_i|^{-1/2} \\
&\times e^{-1/2[(\mathbf{Y}_i-\mathbb{Z}_i\boldsymbol{\mu}_g-\mathbb{X}_i\boldsymbol{\beta}_g)^\top(\mathbb{Z}_i\mathbb{D}_g\mathbb{Z}_i^\top+\boldsymbol{\Sigma}_i)^{-1}(\mathbf{Y}_i-\mathbb{Z}_i\boldsymbol{\mu}_g-\mathbb{X}_i\boldsymbol{\beta}_g)]} \\
&\times [\lambda_{0g}(T_i,\boldsymbol{\xi}_g)\exp\{\tilde{\mathbf{X}}_i(T_i)^\top\boldsymbol{\gamma}_g\}]^{\delta_i}e^{-\int_0^{T_i}\lambda_{0g}(u,\boldsymbol{\xi}_g)\exp\{\tilde{\mathbf{X}}_i(u)^\top\boldsymbol{\gamma}_g\}du} \\
&\times \frac{\exp\{\boldsymbol{\zeta}_g^\top\tilde{\mathbf{Z}}_i\}}{\sum_{g=1}^{G}\boldsymbol{\zeta}_g^\top\tilde{\mathbf{Z}}_i}. \qquad (2.37)
\end{aligned}$$

Oproti modelom so spoločnými náhodnými efektami je vierohodnosť jednoduchšia, a teda výpočetne menej náročná. Spočítať maximum vierohodnostnej funkcie analyticky je však stále náročné. Na maximalizovanie vierohodnosti sa

preto opäť používajú numerické metódy, ako napríklad Newton-Raphsonov algoritmus, EM-algoritmus alebo modifikovaný Marquardtov algoritmus [Proust-Lima et al., 2009]. Keďže vierohodnosť môže mať viaceré lokálne maximá, [Hipp and Bauer, 2006] odporúčajú spustiť algoritmus na spočítanie maximálne vierohodných odhadov viackrát, s rôznymi počiatočnými hodnotami, aby bolo zaručené dosiahnutie globálneho maxima. Tento problém tak v zásade vyruší výhodu nepotrebnej numerickej integrácie, keďže opakované nasadzovanie modelu je tiež výpočetne náročné. Pre použitie metódy maximálnej vierohodnosti na odhadovanie parametrov je potrebný parametrický predpoklad pre základné riziko. [Proust-Lima et al., 2009] uvádzajú napr. Weibullovo rozdelenie, po častiach konštatný predpis alebo parametrizácia pomocou splajnov.

Nevýhoda použitia JLCM je, že interpretácia koeficientov nie je priamočiara, preto sa použitie tohto typu modelov odporúča najmä na predikcie a pri potrebe odhaliť latentnú heterogenitu. Ďalšou nevýhodou je, že odhady sa počítajú pre daný počet latentných kategórií. V praxi to znamená, že musia byť použité viaceré modely s rôznym počtom latentných kategórií. Modely sa následne porovnávajú na základe rôznych kritérií [Hawkins et al., 2001]. Výsledky štúdií ukazujú, že spoľahlivým nástrojom pre porovnanie združených modelov je Bayesovo informačné kritérium (anglicky *Bayes Information Criterion-BIC*) [Rizopoulos, 2012]. Kritérium možno spočítať pomocou predpisu $\mathrm{BIC}(G) = -2L(\boldsymbol{\theta}_G) + n_\theta \log(K)$, kde n_θ je počet odhadnutých parametrov v modeli.

Tieto modely opäť možno rozšíriť uvažovaním viacerých longitudinálnych premenných [Rizopoulos and Ghosh, 2011] alebo viacerých časov zlyhania [Huang et al., 2011]. V tejto práci sa ani jednému z týchto rozšírení nebudeme venovať.

3. Dynamická predikcia

Cieľom tejto kapitoly je popísať postup predikcie individuálnych funkcií prežitia, na základe ktorých možno robiť inferenciu o individuálnych pravdepodobnostiach prežitia. Združené modely z predchádzajúcej kapitoly využívame ako prostriedok na modelovanie pravdepodobností prežitia. Princíp dynamickosti predikcie spočíva v tom, že tieto pravdepodobnosti prežitia sú aktualizované vždy, keď sú zaznamenané nové longitudinálne informácie. Okrem pravdepodobností prežitia možno dynamickú predikciu využiť aj na predikciu longitudinálnej odozvy. Touto problematikou sa podrobne zaoberal autor [Rizopoulos, 2012] a v práci ju nebudeme popisovať.

3.1 Uvedenie do problematiky

V značení z 2. kapitoly máme pre každého z K pacientov vektor longitudinálnych meraní $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, $i = 1, \ldots, K$. Ďalej značíme pre každý subjekt nezáporné (nepozorované) náhodné veličiny pre skutočný čas udalosti a čas cenzorovania $(T_1^*, C_1), \ldots, (T_K^*, C_K)$, $T_i = \min(T_i^*, C_i)$, $i = 1, \ldots, K$ cenzorovaný čas udalosti pre každého pacienta a $\delta_i = \mathbb{1}(T_i^* \leq C_i)$, $i = 1, \ldots, K$, príslušné indikátory udalostí. Predpokladáme splnenie podmienky nezávislého cenzorovania z definície 5. Označme $\mathcal{D}_K = \{T_i, \delta_i, \mathbf{Y}_i, i = 1, \ldots, K\}$ dáta o K subjektoch, na základe ktorých bol vybudovaný združený model 2.1.2. Pre (nového) pacienta $i+1$, bude $\mathcal{Y}_{i+1}(t) = \{\mathbf{Y}_{i+1}(s); 0 \leq s < t\}$ predstavovať množinu longitudinálnych meraní zaznamenaných do času t, $\tilde{\mathbf{X}}_{i+1}$ všetky zvyšné regresory v združenom modeli, T_{i+1}^* skutočný (neznámy) čas udalosti, $S_{i+1}(\cdot)$ budeme značiť funkciu prežitia náhodnej veličiny T_{i+1}^* a $\boldsymbol{\theta}^*$ skutočnú hodnotu vektoru parametrov združeného modelu. Predpoklad, že máme o pacientovi merania do času t implikuje, že do času t pacient prežil, resp. u pacienta nedošlo k udalosti. Zaujíma nás predikcia pravdepodobnosti, že u pacienta $i + 1$ nastane udalosť v časovom intervale $[t, u]$, $0 \leq t < u$. Budeme sa teda sústrediť na individuálne pravdepodobnosti prežitia do času $u > t$ za podmienky prežitia do času $t \geq 0$:

$$\pi_{i+1}(u|t) = \mathsf{P}(T_{i+1}^* \geq u | T_{i+1}^* > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \mathcal{D}_K, \boldsymbol{\theta}^*), \quad t > 0. \tag{3.1}$$

Dynamický charakter (3.1) spočíva v tom, že pri nameraní nových informácií o pacientovi v čase $t' > t$, môžeme aktualizovať predikciu a dostaneme $\pi_{i+1}(u|t'), u > t'$.

Za platnosti modelu 2.1.2 môžeme spočítať $\pi_{i+1}(u|t)$, $u > t$ ako aposteriórnu strednú hodnotu:

$$\pi_{i+1}(u|t) = \int_\Theta \mathsf{P}(T_{i+1}^* \geq u | T_{i+1}^* > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_K) d\boldsymbol{\theta}. \tag{3.2}$$

Výpočet prvého člena integrandu využíva predpoklad podmienenej nezávislosti (2.5) a môžeme ho rozpísať ako

$$
\begin{aligned}
&\mathsf{P}(T^*_{i+1} \geq u | T^*_l > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \boldsymbol{\theta}) \\
&= \int_{\mathbb{R}^q} \mathsf{P}(T^*_{i+1} \geq u | T^*_{i+1} > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \mathbf{b}_{i+1}, \boldsymbol{\theta}) p(\mathbf{b}_{i+1} | T^*_{i+1} > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \boldsymbol{\theta}) d\mathbf{b}_{i+1} \\
&= \int_{\mathbb{R}^q} \mathsf{P}(T^*_{i+1} \geq u | T^*_{i+1} > t, \tilde{\mathbf{X}}_{i+1}, \mathbf{b}_{i+1}, \boldsymbol{\theta}) p(\mathbf{b}_{i+1} | T^*_{i+1} > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \boldsymbol{\theta}) d\mathbf{b}_{i+1} \\
&= \int_{\mathbb{R}^q} \frac{\mathsf{P}(T^*_{i+1} \geq u, T^*_{i+1} > t, \tilde{\mathbf{X}}_{i+1}, \mathbf{b}_{i+1}, \boldsymbol{\theta})}{\mathsf{P}(T^*_{i+1} > t, \tilde{\mathbf{X}}_{i+1}, \mathbf{b}_{i+1}, \boldsymbol{\theta})} p(\mathbf{b}_{i+1} | T^*_{i+1} > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \boldsymbol{\theta}) d\mathbf{b}_{i+1} \\
&= \int_{\mathbb{R}^q} \frac{S_{i+1}(u | \tilde{\mathbf{X}}_{i+1}, \mathbf{b}_{i+1}, \boldsymbol{\theta})}{S_{i+1}(t | \tilde{\mathbf{X}}_{i+1}, \mathbf{b}_{i+1}, \boldsymbol{\theta})} p(\mathbf{b}_{i+1} | T^*_{i+1} > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \boldsymbol{\theta}) d\mathbf{b}_{i+1}, \qquad (3.3)
\end{aligned}
$$

kde druhá rovnosť platí z nezávisloti longitudinálnych dát a času do udalosti, podmienene na náhodných efektoch. Na základe prvého člena integrandu (3.3) môžeme ďalej pomocou bayesovskej teórie odvodiť odhad pravdepodobnosti prežitia $\pi_{i+1}(u|t)$.

3.2 Odhady pravdepodobností prežitia

Kľúčovým krokom k predikcii a odhadovaniu pravdepodobností prežitia je získať výber K vektorov náhodných efektov $\mathbf{b}_i$, $i = 1, \ldots K$, z aposteriórneho rozdelenia $p(\mathbf{b}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\boldsymbol{\theta})$ z tvrdenia 2. Podmienene na m-tej postupnosti $\boldsymbol{\theta}^{(m)}$, $m = 1, \ldots, M$, generujeme m-tú vzorku náhodných efektov z aposteriórneho rozdelenia $p(\mathbf{b}|\mathbf{y},\mathbf{t},\boldsymbol{\delta},\boldsymbol{\theta}^{(m)})$. Odhad pravdepodobnosti prežitia následne spočítame dosadením postupností vektorov parametrov a vektorov náhodných efektov $\{\boldsymbol{\theta}^{(m)}, \mathbf{b}^{(m)}, m = 1, \ldots, M\}$ do modelu:

$$
\begin{aligned}
\hat{\pi}_{i+1}(u|t) &= \frac{1}{M} \sum_{m=1}^{M} \mathsf{P}(T^*_{i+1} \geq u | T^*_{i+1} > t, \mathcal{Y}_{i+1}(t), \tilde{\mathbf{X}}_{i+1}, \mathbf{b}^{(m)}_{i+1}, \boldsymbol{\theta}^{(m)}) \\
&= \frac{1}{M} \sum_{m=1}^{M} \frac{\mathsf{P}(T^*_{i+1} \geq u, T^*_{i+1} > t, \tilde{\mathbf{X}}_{i+1}, \mathbf{b}^{(m)}_{i+1}, \boldsymbol{\theta}^{(m)})}{\mathsf{P}(T^*_{i+1} > t, \tilde{\mathbf{X}}_{i+1}, \mathbf{b}^{(m)}_{i+1}, \boldsymbol{\theta}^{(m)})} \\
&= \frac{1}{M} \sum_{m=1}^{M} \frac{S_{i+1}(u | \tilde{\mathbf{X}}_{i+1}, \mathbf{b}^{(m)}_{i+1}, \boldsymbol{\theta}^{(m)})}{S_{i+1}(t | \tilde{\mathbf{X}}_{i+1}, \mathbf{b}^{(m)}_{i+1}, \boldsymbol{\theta}^{(m)})} \\
&= \frac{1}{M} \sum_{m=1}^{M} \pi^{(m)}_{i+1}(u|t). \qquad (3.4)
\end{aligned}
$$

Pri odhadovaní môžeme postupovať rovnako ako v prvej kapitole, zvoliť nejaké $m_0 > 0$ dostatočne veľké a prvých m_0 členov postupnosti ignorovať, t.j. brať ako členy z rozohrievacej fázy, ktoré nebudú použité na odhadovanie. Označme ďalej

$$\rho_k = \mathrm{cov}[\pi^{(m)}_{i+1}(u|t), \pi^{(m+k)}_{i+1}(u|t)],\ k = 0,1,2,\ldots.$$

Za platnosti istých podmienoky regularity [Meyn and Tweedie, 1993, Kapitola 17], platí Centrálna limitná veta, ktorá umožňuje štatistickú inferenciu výstupov z MCMC. Z Centrálnej limitnej vety platí:

$$\sqrt{M}\Big(\hat{\pi}_{i+1}(u|t) - \pi_{i+1}(u|t)\Big) \xrightarrow[M\to\infty]{D} \mathsf{N}(0,\sigma^2_h),$$

kde $\sigma_h^2 = \rho_0 + 2\sum_{k=1}^{\infty} \rho_k$ určuje veľkosť takzvanej MCMC chyby [Brooks et al., 2011]. Prirodzená miera presnosti $\hat{\pi}_{i+1}(u|t)$ je teda daná MCMC štandardnou chybou (MCMCSE) definovanou ako

$$\text{MCMCSE}(\hat{\pi}_{i+1}(u|t)) = \sqrt{\frac{\rho_0 + 2\sum_{k=1}^{\infty} \rho_k}{M}}. \tag{3.5}$$

Odhad MCMC chyby $\hat{\pi}_{i+1}(u|t)$ môžeme spočítať pomocou empirickej autokovariančnej funkcie

$$\hat{\rho}_k = \frac{1}{M} \sum_{m=1}^{M-k} [\pi_{i+1}^{(m)}(u|t) - \hat{\pi}_{i+1}(u|t)][\pi_{i+1}^{(m+k)}(u|t) - \hat{\pi}_{i+1}(u|t)], \tag{3.6}$$

pre $k = 0,1,2,\ldots$.

Interval spoľahlivisti s pokrytím 95% je z Centrálnej limitnej vety pre Markovské reťazce tvaru

$$\Big(\hat{\pi}_{i+1}(u|t) \pm u_{0.975}\text{MCMCSE}(\hat{\pi}_{i+1}(u|t))\Big),$$

kde $u_{0.975}$ je 97.5% kvantil štandardného normálneho rozdelenia.

3.3 Miery presnosti predikcie

Okrem samotnej predikcie je dôležité určiť, ako dobre longitudinálny marker predikuje pravdepodobnosť prežitia. Na zodpovedanie tejto otázky sa používajú miery ako kalibrácia a diskriminácia. Kalibrácia meria, ako dobre model predikuje pozorované dáta. Pomocou diskriminácie hodnotíme, ako dobre model rozlišuje pacientov, u ktorých dôjde k udalosti v krátkodobom časovom horizonte od pacientov, u ktorých dôjde k udalosti neskôr. V štandardnej analýze prežitia sa používajú miery, ktoré kombinujú tieto dva koncepty, ako napríklad chyba predikcie [Graf et al., 1999]. V tejto časti sa budeme zaoberať jednotlivými pojmami podrobnejšie.

3.3.1 Diskriminácia

Na vyhodnotenie rozlišovacej schopnosti modelu budeme predpokladať nasledujúcu situáciu:

- použitie dostupných longitudinálnych meraní do času $t > 0$,
- zaujíma nás výskyt udalosti v medicínsky relevantnom časovom intervale $(t,t + \Delta t],\ \Delta > 1$.

V závislosti na použitom modeli a pre konkrétnu prahovú hodnotu $c \in [0,1]$ označíme pacienta i za prípad, ak $\pi_i(t + \Delta t|t) \leq c$. Následne môžeme zadefinovať pojmy ako senzitivita a špecificita.

Definícia 15. *Pre* $c \in [0,1]$ *a* $\pi_i(t+\Delta t|t)$ *pravdepodobnosť prežitia i-teho pacienta v časovom intervale* $(t,t+\Delta t]$, $t > 0,\Delta > 1$, *definujeme senzitivitu (tiež aj pomer správne pozitívnych) ako*

$$SN_t^{\Delta t}(c) = \mathsf{P}(\pi_i(t+\Delta t|t) \leq c|,T_i^* > t, T_i^* \in (t,t+\Delta t]).$$

Definícia 16. *Pre* $c \in [0,1]$ *a* $\pi_i(t+\Delta t|t)$ *pravdepodobnosť prežitia i-teho pacienta v časovom intervale* $(t,t+\Delta t]$, $t > 0,\Delta > 1$, *definujeme špecificitu ako pravdepodobnosť, že pacient je správne zaradený do kategórie pacientov, u ktorých nenastala udalosť:*

$$SP_t^{\Delta t}(c) = \mathsf{P}(\pi_i(t+\Delta t|t) > c|T_i^* > t, T_i^* > t+\Delta t).$$

Ďalej pomocou špecificity a senzitivity definujeme operačnú charakteristiku prijímača (anglicky *Receiver Operating Characteristic – ROC*) a plochu pod krivkou (anglicky *Area Under the Curve – AUC*) [Hanley and McNeil, 1982].

Definícia 17. *Pre* $p \in [0,1]$ *a* $(1-SP_t^{\Delta t}(p))^{-1} = inf\{c : 1-SP_t^{\Delta t}(c) \leq p\}$ *definujeme operačnú charakteristiku prijímača ako*

$$ROC_t^{\Delta t}(p) = SN_t^{\Delta t}\{(1-SP_t^{\Delta t}(p))^{-1}\}.$$

Plochu pod ROC krivkou potom definujeme ako

$$AUC_t^{\Delta t} = \int_0^1 ROC(p)dp.$$

Intuitívnejšiu interpetáciu AUC poskytuje formulácia podľa[Hanley and McNeil, 1982]:

$$\mathrm{AUC}_t^{\Delta t} = \mathsf{P}[\pi_i(t+\Delta t|t) < \pi_j(t+\Delta t|t)|\{T_i^* \in (t,t+\Delta t]\} \cap \{T_j^* > t+\Delta t\}].$$

Pre každý náhodný pár $\{i,j\}$ pacientov, ktorých stav vieme kategorizovať podľa toho, či došlo k udalosti alebo nie, AUC reprezentuje pravdepodobnosť, že klasifikácia podľa úrovne markerov je zhodná s klasifikáciou podľa výskytu udalosti. AUC rovné jednej značí maximálnu diskrimináciu, zatiaľ čo $\mathrm{AUC} = 0.5$ značí náhodnú diskrimináciu (t.j. marker nerozlišuje medzi pacientami lepšie ako hod mincou). Čím je teda hodnota AUC bližšie k 1, tým má model lepšiu diskriminačnú schopnosť. Výhoda používania metodiky ROC je, že môže byť použitá na porovnávanie rôznych markerov.

Vyššie zadefinované miery presnosti rozlišujú pacientov v špecifickom čase t, a teda v rôznych časových okamihoch môže mať model rôzne úrovne diskriminácie. Na zhrnutie diskriminačnej schopnosti markeru počas celej doby sledovania navrhol [Rizopoulos, 2011] použitie dynamického indexu diskriminácie založeného na vážených priemerných AUC:

$$C_{\mathrm{dyn}}^{\Delta t} = \int_0^\infty \mathrm{AUC}_t^{\Delta t} u(t)dt,$$

kde

$$u(t) = \frac{\mathsf{P}(T_t^* > t)}{\int \mathsf{P}(T_j^* > t)dt}.$$

Voľba váhovej funkcie $u(t)$ je založená na myšlienke, že časové okamihy neprispievajú do porovnania rovnako, pretože v neskorších okamihoch očakávame menej subjektov k dispozícii. V praxi sledujeme pacientov na nejakom ohraničenom časovom intervale $(0,L)$, $L > 0$. V takom prípade je dynamický index diskriminácie upravený do tvaru:

$$\left[C_{\text{dyn}}^{\Delta t}\right]^L = \int_0^L \text{AUC}_t^{\Delta t} u^L(t)dt,$$

kde

$$u^L(t) = \frac{u(t)}{\int_0^L u(t)dt}.$$

Odhadovanie senzitivity, špecificity a AUC je v jednoduchom prípade založené na počítaní príslušných početností v napozorovanom výbere. Napríklad, odhad senzitivity možno spočítať ako pomer pacientov spĺňajúcich $\pi_i(t + \Delta t|t) \leq c$ a pacientov, u ktorých nastala udalosť v intervale $(t,t + \Delta t]$. V súvislosti s cenzorovaním však tieto odhady nemožno založiť čisto na početnostiach pacientov, u ktorých nastala udalosť. [Li et al., 2018] odvodili konzistentný odhad senzitivity a špecificity použitím váh založených na modeli. Označme $W_i \in \{0,1\}$, $i = 1, \ldots, K$ váhu $i-$teho pacienta a $\mathbb{1}(T_i^* \leq t+\Delta t)$ je identifikátor, že u $i-$teho pacienta došlo k udalosti do času $t + \Delta t$, $\Delta > 1$. Odvodenie tvaru W_i je založené na princípe nasledujúcich 4 scenárov, ktoré môžu nastať v prítomnosti cenzorovania.

1. $T_i > t + \Delta t \implies \mathbb{1}(T_i^* \leq t + \Delta t) = 0 \implies W_i = 0,$
2. $T_i \leq t + \Delta t, \delta_i = 1 \implies \mathbb{1}(T_i^* \leq t + \Delta t) = 1 \implies W_i = 1,$
3. $T_i = t + \Delta t, \delta_i = 0 \implies \mathbb{1}(T_i^* \leq t + \Delta t) = 0 \implies W_i = 0,$
4. $T_i < t + \Delta t, \delta_i = 0 \implies T_i^* > T_i \implies W_i = \mathsf{P}(T_i^* \leq t + \Delta t|T_i < t + \Delta t, \delta_i = 0) = 1 - \frac{S_i(t+\Delta t|\tilde{\mathbf{X}}_i, \boldsymbol{\theta}^*)}{S_i(T_i|\tilde{\mathbf{X}}_i, \boldsymbol{\theta}^*)}.$

kde $\boldsymbol{\theta}^*$ je skutočná hodnota vektoru parametrov. Všimneme si, že pre spojitý čas do udalosti, je pravdepodobnosť tretej možnosti nulová. Celkovo tak dostávame

$$\begin{aligned} W_i = \mathsf{P}(T_i^* \leq t + \Delta t|T_i, \delta_i) &= \mathsf{E}\left[\mathbb{1}(T_i^* \leq t + \Delta t|T_i, \delta_i)\right] \\ &= \left[1 - (1 - \delta_i)\frac{S_i(t + \Delta t|\tilde{\mathbf{X}}_i, \boldsymbol{\theta}^*)}{S_i(T_i|\tilde{\mathbf{X}}_i, \boldsymbol{\theta}^*)}\right]\mathbb{1}(T_i \leq t + \Delta t), \end{aligned} \tag{3.7}$$

Odhad W_i potom dostaneme jednoducho ako

$$\hat{W}_i = \left[1 - (1 - \delta_i)\hat{\pi}_i(t + \Delta t|T_i)\right]\mathbb{1}(T_i \leq t + \Delta t)$$

a odtiaľ odhad senzitivity ako

$$\widehat{SN}_t^{\Delta t}(c) = \frac{\sum_{i=1}^{K} \mathbb{1}(\pi_i(t+\Delta t|t) \leq c)\hat{W}_i}{\sum_{i=1}^{K} \hat{W}_i}. \tag{3.8}$$

Všimneme si, že v prípade necenzorovaných dát, odhad odpovedá empirickému odhadu pomerom početností. Odhad špecificity spočítame analogicky ako

$$\widehat{SP}_t^{\Delta t}(c) = \frac{\sum_{i=1}^{K} \mathbb{1}(\pi_i(t+\Delta t|t) > c)(1-\hat{W}_i)}{\sum_{i=1}^{K} (1-\hat{W}_i)}. \tag{3.9}$$

Odhad ROC následne spočítame dosadením odhadov pre senzitivitu a špecificitu do definície. Voľba vyššie spočítaných váh umožňuje aj alternatívny spôsob odhadovania AUC pomocou priamočiareho výpočtu, ktorý nevyžaduje numerickú integráciu. Definícia AUC podľa [Hanley and McNeil, 1982] evokuje podobnosť s Mann-Whitneyho testovou štatistikou. Ak by bol čas udalosti známy pre každého pacienta, odhad AUC by bol tvaru

$$\begin{aligned}\widehat{\mathrm{AUC}}_t^{\Delta t} = & \frac{1}{\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{1}(T_i^* \leq \Delta t + t)\mathbb{1}(T_j^* > \Delta t + t)} \\ & \times \sum_{i=1}^{K}\sum_{j=1}^{K} \{\mathbb{1}(T_i^* \leq \Delta t + t)\mathbb{1}(T_j^* > \Delta t + t) \\ & \times [\mathbb{1}(\hat{\pi}_i(t+\Delta t|t) < \hat{\pi}_j(t+\Delta t|t)) + 0.5 \times \mathbb{1}(\hat{\pi}_i(t+\Delta t|t) = \hat{\pi}_j(t+\Delta t|t))]\},\end{aligned}$$

Člen $0.5 \times \mathbb{1}(\pi_i(t+\Delta t|t) = \pi_j(t+\Delta t|t))$ je vo vzorci zavedený kvôli zhodám. V prípade cenzorovania je člen $\mathbb{1}(T_i^* \leq \Delta t + t)$ nahradený váhami, resp. ich odhadmi, teda plat:

$$\begin{aligned}\widehat{\mathrm{AUC}}_t^{\Delta t} = & \frac{1}{\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{W}_i(1-\hat{W}_j)} \sum_{i=1}^{K}\sum_{j=1}^{K} \{\hat{W}_i(1-\hat{W}_j) \\ & \times [\mathbb{1}(\hat{\pi}_i(t+\Delta t|t) < \hat{\pi}_j(t+\Delta t|t)) + 0.5 \times \mathbb{1}(\hat{\pi}_i(t+\Delta t|t) = \hat{\pi}_j(t+\Delta t|t))]\}.\end{aligned}$$

Dôkaz konzistencie odhadu AUC možno nájsť v [Li et al., 2018]. Ďalšími spôsobmi odhadovania senzitivity, špecificity a AUC, ako napríklad použitím váženia pomocou inverznej pravdepodobnosti cenzorovania (angl. *Inverse Probability of Censoring Weighting – IPCW*), sa zaoberali napr. [Blanche et al., 2013].

3.3.2 Kalibrácia a validácia

Ďalšou dôležitou mierou na vyhodnotenie predikčnej schpnosti modelu je kalibrácia, t.j. kvantifikovanie zhody medzi predikovanými a pozorovanými pravdepodobnosťami prežitia. V analýze prežitia sa štandardne používa miera, ktorá kombinuje koncept diskriminácie a kalibrácie – Brierovo skóre (angl. *Brier score – BS*). K zadefinovaniu a odvodeniu tvaru odhadu Brierovho skóre najskôr zopakujeme značenie z kapitoly 1.2. Značíme $R_i(t) = \mathbb{1}(T_i > t)$ $i = 1, \ldots, K$ identifikátor, či je subjekt v riziku v čase t a $\overline{N}(t) = \sum_{i=1}^{K} N_i(t)$ počet subjektov v riziku v čase t. Ďalej budeme značiť $D_i(t) = \mathbb{1}(T_i^* > t)$ status skutočnej udalosti.

Definícia 18. *Očakávaná kvadratická chyba predikcie (Brierovo skóre) je tvaru*

$$PE(t+\Delta t|t) = \mathsf{E}[\{D_i(t+\Delta t) - \pi_i(t+\Delta t|t)\}^2].$$

Pri odhade $PE(t+\Delta t|t)$ treba brať opäť do úvahy cenzorovanie. Analogicky ako pri odvodení tvaru (3.7), situáciu s cenzorovaním rozdelíme na 3 scenáre:

1.

$$\begin{aligned} T_i > t+\Delta t &\implies D_i(t+\Delta t) = 1 \\ &\implies PE(t+\Delta t|t) = \mathsf{E}[\{1 - \pi_i(t+\Delta t|t)\}^2], \end{aligned} \tag{3.10}$$

2.

$$\begin{aligned} T_i \le t+\Delta t, \delta_i = 1 &\implies D_i(t+\Delta t) = 0 \\ &\implies PE(t+\Delta t|t) = \mathsf{E}[\{0 - \pi_i(t+\Delta t|t)\}^2], \end{aligned} \tag{3.11}$$

3.

$$\begin{aligned} &T_i \le t+\Delta t, \delta_i = 0 \\ &\implies PE(t+\Delta t|t) = \mathsf{P}(T_i^* \le t+\Delta t)\,\mathsf{E}[\{0 - \pi_i(t+\Delta t|t)\}^2] \\ &\qquad\qquad + \mathsf{P}(T_i^* > t+\Delta t)\,\mathsf{E}[\{1 - \pi_i(t+\Delta t|t)\}^2]. \end{aligned} \tag{3.12}$$

Odhad Brierovho skóre, zohľadňujúci cenzorovanie [Henderson et al., 2002], tak dostaneme v tvare

$$\begin{aligned} \widehat{PE}(t+\Delta t|t) = \frac{1}{\overline{N}(t)} \sum_{i=1}^{\overline{N}(t)} & \mathbb{1}(T_i > t+\Delta t)\{1 - \pi_i(t+\Delta t|t)\}^2 \qquad (3.10) \\ & + \mathbb{1}(T_i \le t+\Delta t)\delta_i\{0 - \pi_i(t+\Delta t|t)\}^2 \qquad (3.11) \\ & + \mathbb{1}(T_i \le t+\Delta t)(1-\delta_i)\Big[(1 - \hat{\pi}_i(t+\Delta t|T_i))\{0 - \pi_i(t+\Delta t|t)\}^2 \\ & + \hat{\pi}_i(t+\Delta t|T_i)\{1 - \pi_i(t+\Delta t|t)\}^2\Big]. \qquad (3.12) \end{aligned}$$

V odhade sú použité váhy založené na modeli, ktorých výhodou je, že cenzorovanie môže závisieť na longitudinálnej histórii. Nevýhodou ich použitia je, že model musí byť dobre špecifikovaný.

Použitím longitudinálnych informácií do času t, $PE(t+\Delta t|t)$ meria presnosť predikcie v konkrétnom čase $t+\Delta t$. Alternatívne môžeme zhrnúť chybu predikcie na konkrétnom časovom intervale $[t, t+\Delta t]$, spočítaním váženého priemeru $\{PE(s|t), t < s < t+\Delta t\}$, ktorý zohľadňuje cenzorovanie, podobne ako $C_{\text{dyn}}^{\Delta t}$. [Schemper and Henderson, 2000] navrhli odhad integrovanej chyby predikcie, ktorý je po zohľadnení časovej dynamickosti tvaru:

$$\widehat{IPE}(t+\Delta t|t) = \frac{\sum_{i:t\le T_i\le t+\Delta t} \delta_i \frac{\hat{S}_C(t)}{\hat{S}_C(T_i)} \widehat{PE}(T_i|t)}{\sum_{i:t\le T_i\le t+\Delta t} \delta_i \frac{\hat{S}_C(t)}{\hat{S}_C(T_i)}}.$$

Na získanie objektívneho vyhodnotenia schopnosti modelu predikovať pravdepodobnosti prežitia je potrebné vyhodnotiť miery presnosti predikcie. Interná validácia miery presnosti predikcie sa vykonáva pomocou štandardných prevzorkovacích techník, ako napríklad krosvalidácia (pomocu vynechaného jedného pozorovania) alebo Bootstrapom. Vo všeobecnosti je tento proces časovo náročný, keďže vyžaduje opakované budovanie modelu, preto sa odporúča využiť paralelné počítanie. Externá validácia je založená na počítaní mier presnosti predikcie z dátového súboru z inej analýzy.

4. Ilustračná analýza reálnych dát

Aplikáciu združeného modelu ilustrujeme na dátach k primárnej biliárnej cirhóze (PBC). Jedná sa o chronické a smrteľné, ale zriedkavé ochorenie pečene, ktoré je charakteristické zápalovým rozpadom žlčových vývodov v pečeni, ktoré neskôr vedie k cirhóze pečene. Dáta pochádzajú z klinickej štúdie, ktorú uskutočnila Klinika Mayo v rokoch 1974 – 1984 [Murtaugh et al., 1994]. Cieľom štúdie bolo vyhodnotiť použitie lieku D-penicillamine na liečbu PBC. Pacienti s PBC majú anomálie vo viacerých krvných testoch, ako napríklad zvýšenú hladinu bilirubínu. Počas sledovania boli zaznamenávané viaceré biomarkery spájané s PBC. V analýze sa zameriame na hladinu bilirubínu, ktorý je považovaný za jeden z najdôležitejších markerov spojený s priebehom ochorenia. Na analýzu a spracovanie dát bol využitý software R [R Core Team, 2023] a balík JMbayes [Rizopoulos, 2016]. Táto analýza slúži iba na ilustráciu aplikácie teórie z predchádzajúcich kapitol a získanie realistických scenárov pre simulačnú štúdiu, ktorá bude obsahom nasledujúcej kapitoly. Cieľom preto nie je zaoberať sa budovaním kompletného modelu. Uvádzané modely sú preto jednoduché a detailnú analýzu možno nájsť napríklad v [Dickson et al., 1989].

4.1 Popis dátového súboru

V tejto ilustračnej analýze budeme skúmať 312 pacientov, ktorí boli náhodne rozdelení na podanie lieku D-penicillamine alebo placeba. Celkovo máme k dispozícii 1945 pozorovaní. V knižnici JMbayes sú PBC dáta k dispozícií v dátových súboroch pbc2 a pbc2.id, ktoré obsahujú po rade longitudinálne a cenzorované údaje (tj. prvý súbor je v dlhom formáte a druhý obsahuje prvé meranie od každého pacienta). V tabuľke 4.1 vidieť v dátach výrazný nepomer medzi mužmi a ženami. Ako udalosť budeme chápať transplantáciu alebo úmrtie pacienta. Cenzorovanie, vyvolané zvyčajne odchodom pacienta zo štúdie, bolo približne v 45 % prípadov. V tabuľke 4.2 sú základné popisné charakteristiky numerických premenných. Merania boli robené na dospelých pacientoch vo veku od 26 do 78 rokov. Doba sledovania, resp. čas prežitia, rovnako ako aj počet meraní, sa medzi jednotlivými pacientami výrazne líšila, od jedného, až po 16 meraní, v priebehiu 40 dní, až 14 rokov. Nulová hodnota posledného času longitudiálnych meraní odpovedá situácii, keď pacient absolvoval iba jedno meranie a následne došlo k cenzorovaniu alebo udalosti.

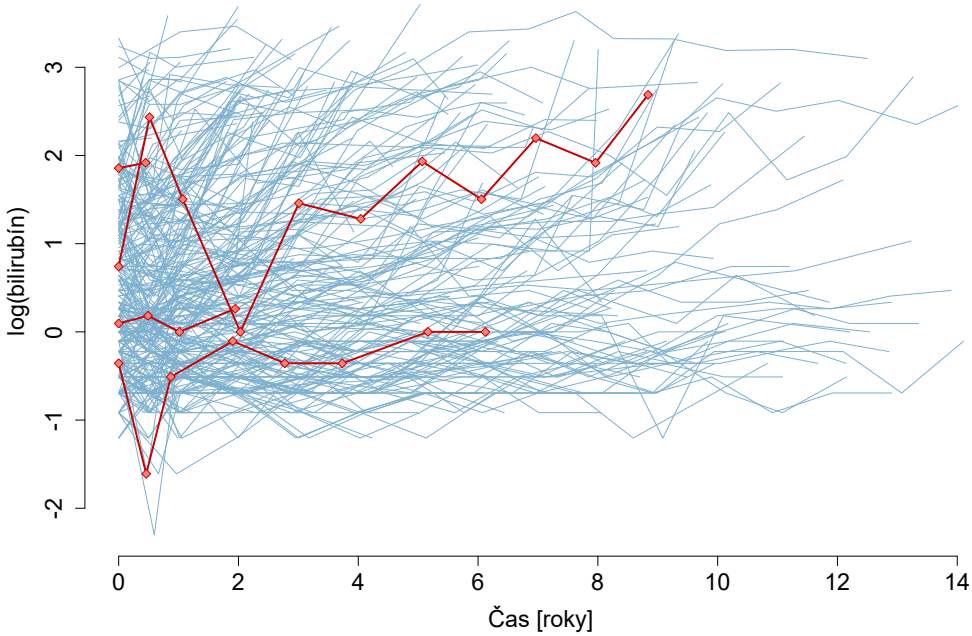
Faktor	Skupina	Zastúpenie	Počet pozorovaní	Relatívne početnosti
Pohlavie	žena	276	1708	88,46%
	muž	36	237	11,54%
Liečba	Liek	158	978	50,64%
	placebo	154	967	49,36%
Indikátor udalosti	Cenzorovanie	143	1073	45,83%
	Udalosť	169	872	54,17%

Tabuľka 4.1: Údaje o pohlaví, liečbe a počte cenzorovaných pacientov.

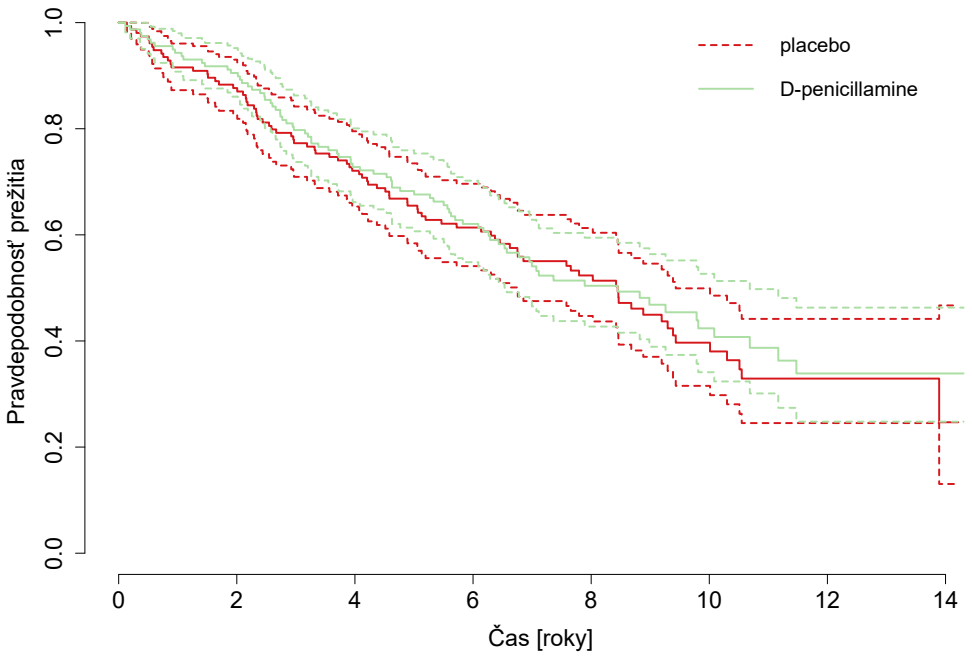
	Min	Max	Priemer	1. kvartil	Medián	3. kvartil	Smerodajná odchýlka
Vek (roky)	26,28	78,44	50,02	42,24	49,80	56,72	10,58
Bilirubín (mg/dl)	0,10	41,00	3,67	0,80	1,40	3,90	4,49
Čas prežitia (roky)	0,11	14,31	6,41	3,71	6,30	8,88	3,56
Posledný čas merania (roky)	0,00	14,11	6,84	3,72	6,82	9,62	3,70
Počet návštev	1,00	16,00	4,76	2,00	4,00	7,00	3,25

Tabuľka 4.2: Výberové údaje o veku a množstve bilirubínu na počiatku štúdie, dĺžke sledovania subjektov a počte návštev.

Z dát sme, s nastavením *set.seed(0610)*, náhodne vybrali 4 pacientov, postupne s počtom meraní 2,4,8,11, pre ktorých budeme ilustrovať predikcie pravdepodobností prežitia. Týchto pacientov sme vyradili z dátového súboru, na základe ktorého bol vybudovaný združený a klasický Coxov model. Longitudinálne trajektórie týchto pacientov sú vyznačené červenou na obrázku 4.1. Na základe grafickej analýzy bola zvolená logaritmická transformácia bilirubínu ako odozvy v LME, aj ako regresoru v klasickom Coxovom modeli. Na obrázku 4.2 sú vykreslené odhady funkcií prežitia bez transplantácie pre pacientov s placebom a liečených pacientov, z dátového súboru na budovanie modelu. Na základe obrázku sa zdá, že efekt liečby nebude signifikantný.



Obr. 4.1: Subjekt-špecifické longitudinálne trajektórie pre *log*(*bilirubín*). Červenou sú pacienti, na ktorých nebol budovaný model a pre ktorých budú ilustrované predikcie.



Obr. 4.2: Kaplan-Meierov odhad pravdepodobnosti prežitia bez transplantácie a bodové konfidenčné intervaly s pokrytím 95% pre pacientov s placebom a s liekom D-penicillamine.

4.2 Model a odhady parametrov

Cieľom ilustračnej analýzy je aplikovať poznatky z teórie na reálne dáta a porovnať odhady koeficientov a predikcie pravdepodobností prežitia získané zo združeného modelu a klasického Coxovho modelu. Na ilustráciu volíme združený model (4.1), v ktorom uvažujeme prirodzený logaritmus bilirubínu ako longitudinálnu premennú a Coxov model (4.2), v ktorom sú hodnoty logaritmu bilirubínu použité konštatne v čase. V modeloch pre riziko úmrtia budeme ako regresory uvažovať ďalej vek pacientov na začiatku štúdie, pohlavie a liečbu pacientov. Logaritmus bilirubínu budeme modelovať pre jednoduchosť len v závislosti na čase, ako fixnom, tak aj náhodnom efekte. Pre všetky 3 modely (LME, Coxov model a SREM) bola vykonaná grafická diagnostika na posúdenie platnosti predpokladov pre použitie modelov.

$$\begin{aligned}\mathsf{SREM}:\ \log(\text{bili})(t_{ij}) &= \beta_1 + \beta_2 * t_{ij} + \mathrm{b}_i^0 + \mathrm{b}_i^1 * t_{ij},\\ \lambda_i(t|\mathrm{b}_i^0,\mathrm{b}_i^1) &= \exp\{\xi_0 + \sum_{j=1}^{8} \xi_j B_j(t,\mathbf{v}) + \gamma_1 * vek_i + \gamma_2 * \mathbb{1}\{\check{z}ena\}_i\\ &\quad + \gamma_3 * \{liek\}_i + \alpha[\beta_1 + \beta_2 * t_{ij} + \mathrm{b}_i^0 + \mathrm{b}_i^1 * t_{ij}]\},\\ &\quad j = 1,\ldots,n_i,\ i = 1,\ldots 308. \end{aligned} \tag{4.1}$$

$$\begin{aligned}\mathsf{Cox}: \lambda(t|vek,pohlavie,liek,log(bili)(t)) &= \lambda_0(t) + \tilde{\gamma}_1 * vek + \tilde{\gamma}_2 * \mathbb{1}\{\check{z}ena\}\\ &\quad + \tilde{\gamma}_3 * \mathbb{1}\{liek\} + \tilde{\gamma}_4 * log(bili)(t). \end{aligned} \tag{4.2}$$

Združený model bol použitý so 100 000 iteráciami MCMC, s predvoľbou 3000 iterácii v zahrievacej fáze a 3000 iterácií na adaptovanie modelu. Na zníženie autokorelácie v rámci reťazcov bol použitý stenčovací parameter (anglicky *thinning parameter*) $\mathsf{n.thin} = 2$. Uzly a rád splajnov boli ponechané v predvoľbe, t.j. uzly boli spočítané pomocou percentilov cenzorovaných časov udalosti, zvolený počet uzlov bol 5 a použitá bola splajnovová funkcia rádu 4 (počet koeficientov na každom po častiach polynomiálnom úseku, teda rád 4 odpovedá kubickému splajnu). Apriórne rozdelenia boli, pri značení z modelu 2.1.2, volené nasledovne:

- $\boldsymbol{\beta} \sim \mathsf{N}_2(\mathbf{0},\text{diag}(0{,}001; 0{,}001))$,
- $\tau \sim \Gamma(1; 0{,}005)$,
- $\mathbb{D}$, $\mathbb{D}^{-1} \sim \mathsf{W}_q(2,\text{diag}(0{,}001; 0{,}001))$,
- $\alpha \sim \mathsf{N}(0; 0{,}01)$,
- $\boldsymbol{\gamma} \sim \mathsf{N}_3(\mathbf{0},\text{diag}(0{,}005; 0{,}001; 0{,}001))$,
- $\xi_0,\ldots,\xi_8 \sim \mathsf{N}(0; 0{,}001)$.

Tabuľka 4.3 obsahuje odhady parametrov zo združeného a Coxovho modelu. Pripomenieme, že na odhadovanie parametrov združeného modelu je využívaná

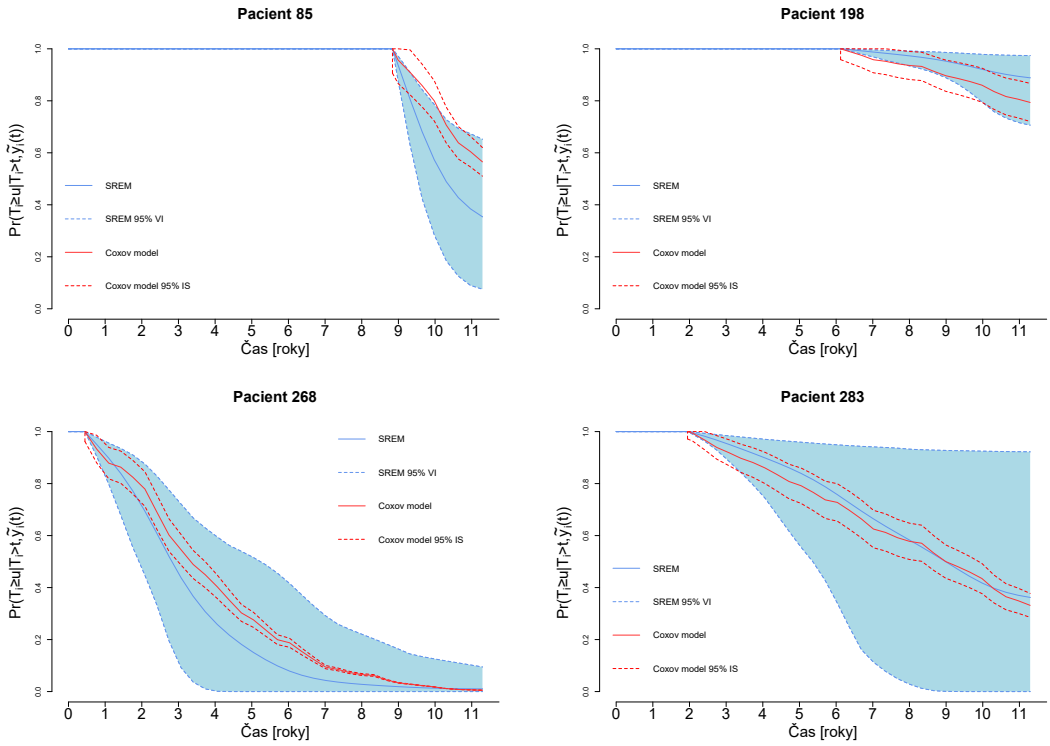
Parameter	Odhad parametra	Smerodatná chyba	P-hodnota	Vierohodnostný/ Konfidečný interval
γ_1	0,043	0,001	$< 0{,}001$	(0,026; 0,060)
$\tilde{\gamma}_1$	0,027	0,007	$< 0{,}001$	(0,013; 0,042)
γ_2	0,071	0,021	0,751	(−0,372; 0,526)
$\tilde{\gamma}_2$	−0,077	0,216	0,723	(−0,500; 0,347)
γ_3	−0,091	0,017	0,564	(−0,418; 0,243)
$\tilde{\gamma}_3$	−0,122	0,160	0,447	(−0,436; 0,192)
α	1,310	0,003	$< 0{,}001$	(1,132; 1,491)
$\tilde{\gamma}_4$	1,048	0,084	$< 0{,}001$	(0,882; 1,213)
β_1	0,494	0,0003	$< 0{,}001$	(0,380; 0,610)
β_2	0,185	0,0001	$< 0{,}001$	(0,162; 0,209)
ξ_0	−6,426	0,086	$< 0{,}001$	(−8,188; −4,854)
ξ_1	−7,287	0,074	$< 0{,}001$	(−8,950; −5,820)
ξ_2	−5,938	0,085	$< 0{,}001$	(−7,754; −4,403)
ξ_3	−6,822	0,060	$< 0{,}001$	(−8,130; −5,466)
ξ_4	−5,858	0,076	$< 0{,}001$	(−7,359; −4,326)
ξ_5	−6,630	0,048	$< 0{,}001$	(−8,006; −5,276)
ξ_6	−4,430	0,080	$< 0{,}001$	(−6,890; −1,879)
ξ_7	−10,350	0,073	$< 0{,}001$	(−15,434; −6,168)
ξ_8	−6,493	0,209	$< 0{,}001$	(−13,493; −2,448)

Tabuľka 4.3: Odhadnuté fixné efekty, príslušné smerodajné chyby, p-hodnoty a vierohodnostné intervaly, resp. konfidenčné intervaly združeného modelu (4.1) a Coxovho modelu (4.2) (šedou).

bayesovská teória a uvedené smerodatné chyby, p-hodnoty a vierohodnostné intervaly majú mierne odlišnú interpretáciu ako údaje získané frekventistickým odhadovaním, v prípade Coxovho modelu. Vidíme, že oba modely sa zhodujú v signifikantnosti efektov regresorov, no vierohodnostné intervaly sú oproti konfidečným intervalom mierne posunuté. Vierohodnostný interval pre parameter α interpretujeme ako množinu, ktorá za podmienky napozorovaných dát pokryje skutočnú hodnotu parametra s pravdepodobnosťou 95%. V prípade združeného modelu je najmenšia pravdepodobnosť, že vierohodnostný interval neobsahuje skutočnú hodnotu parametra α menšia ako 0,001. V prípade Coxovho modelu, pri opakovanom použití modelu na rôzne dáta, interval spoľahlivosti pokryje skutočnú hodnotu parametra v 95% prípadov. P-hodnotu interpretujeme ako najmenšiu možnú hladinu testu, na ktorej by sme zamietali nulovú hypotézu, že parameter α je rovný nule. Bodové odhady parametrov majú podobné hodnoty, až na odhad efektu pohlavia, ktorý má v Coxovom modeli opačné znamienko ako v združenom modeli. Tento efekt vyšiel v oboch modeloch nesignifikantný, preto z tohto výsledku nebudeme vyvodzovať ďalšie závery.

4.3 Predikcie

Pri odlišnosti hodnôt odhadov efektov musíme brať ohľad aj na rôzne prístupy použité pri odhadovaní, preto sa na základe výsledkov v tabuľke nemusí zdať, že v použití rôznych modelov je nejaký výrazný rozdiel. Väčšie rozdiely vo výstupoch modelov sa však ukazujú na predikcii pravdepodobností prežitia, ako vidieť na nasledujúcich obrázkoch. Na obrázku 4.3 sú vykreslené predikcie pravdepodobností prežitia pre pacientov 85, 198, 268 a 283 z PBC datasetu. Odhadované funkcie prežitia zo združeného modelu (modrou) zohľadňujú informáciu, že o pacientovi máme informácie aj medzi začiatkom štúdie a časom udalosti/cenzorovania. Odhady funkcií prežitia sú tak rovné 1 až do posledného času longitudinálneho merania. Pri použití štandardného Coxovho modelu tieto informácie nie sú k dispozícii a odhad podmienenej pravdepodobnosti prežitia $\hat{\pi}_i(u|t)$ možno spočítať len ako podiel odhadov funkcií prežitia $\frac{\hat{S}(u)}{\hat{S}(t)}$ z modelu, v ktorom nevieme zohľadniť, že do času t bol pacient nažive. Vidíme, že bodové odhady z Coxovho modelu sú u pacientov 85 a 268 nadhodnotené, zatiaľ čo u pacienta 198 je odhad podhodnotený. U pacienta 283 sa zdajú byť bodové odhady funkcií prežitia z dvoch modelov pomerne v súlade, no vidíme, že vierohodnostný interval je oveľa širší, ako konfidenčný interval. Podobný trend medzi vierohodnostným a konfidenčným intervalom možno pozorovať aj u pacienta 268. U pacientov 85 a 198 sú posunuté ako bodové, tak aj intervalové odhady. Podrobnejšie budeme skúmať správanie sa odhadov pri rôznych počtoch návštev pacientov v simulačnej štúdii v kapitole 5.



Obr. 4.3: Predikované pravdepodobnosti prežitia, vierohodnostné intervaly (VI) a bodové intervaly spoľahlivosti (IS) s pokrytím 95% pre pacientov 85,198, 268 a 283 z PBC datasetu na základe združeného modelu (modrou) a Coxovho modelu (červenou).

4.4 Diskriminácia a kalibrácia

Tabuľka 4.4 obsahuje vyhodnotenie diskriminácie a kalibrácie združeného modelu v časoch $t = 3{,}5{,}7$, resp. na intervaloch $[t,t + \Delta t]$ pre $\Delta t = 2{,}4$. Vidíme, že pri použití longitudinálnych meraní z prvých 5 rokov, bilirubín vykazuje najlepšiu diskriminačnú schopnosť pre pacientov, ktorí by mohli zomrieť v nasledujúcich dvoch rokoch $\left(\widehat{\mathrm{AUC}}_t^{\Delta t} = 0{,}850\right)$. Aby sme overili, že toto platí počas celej doby sledovania, spočítali sme dynamický index predikcie, v tabuľke 4.5, pre rovnaké časové úseky. Odhad $\hat{C}_{\mathrm{dyn}}^{\Delta t=2}$ má podobnú hodnotu ako $\widehat{\mathrm{AUC}}_{t=5}^{\Delta t=2}$, čo naznačuje, že podľa hodnôt bilirubínu model dobre rozlišuje pacientov. Najmenšiu chybu predikcie dostávame v čase $t + \Delta t = 5$ pri využití longitudinálnych informácií do času $t = 3$. Na časovom intervale $[3; 5]$ je tak isto najmenšia aj integrovaná chyba predikcie $\widehat{IPE}(t + \Delta t|t) = 0{,}056$.

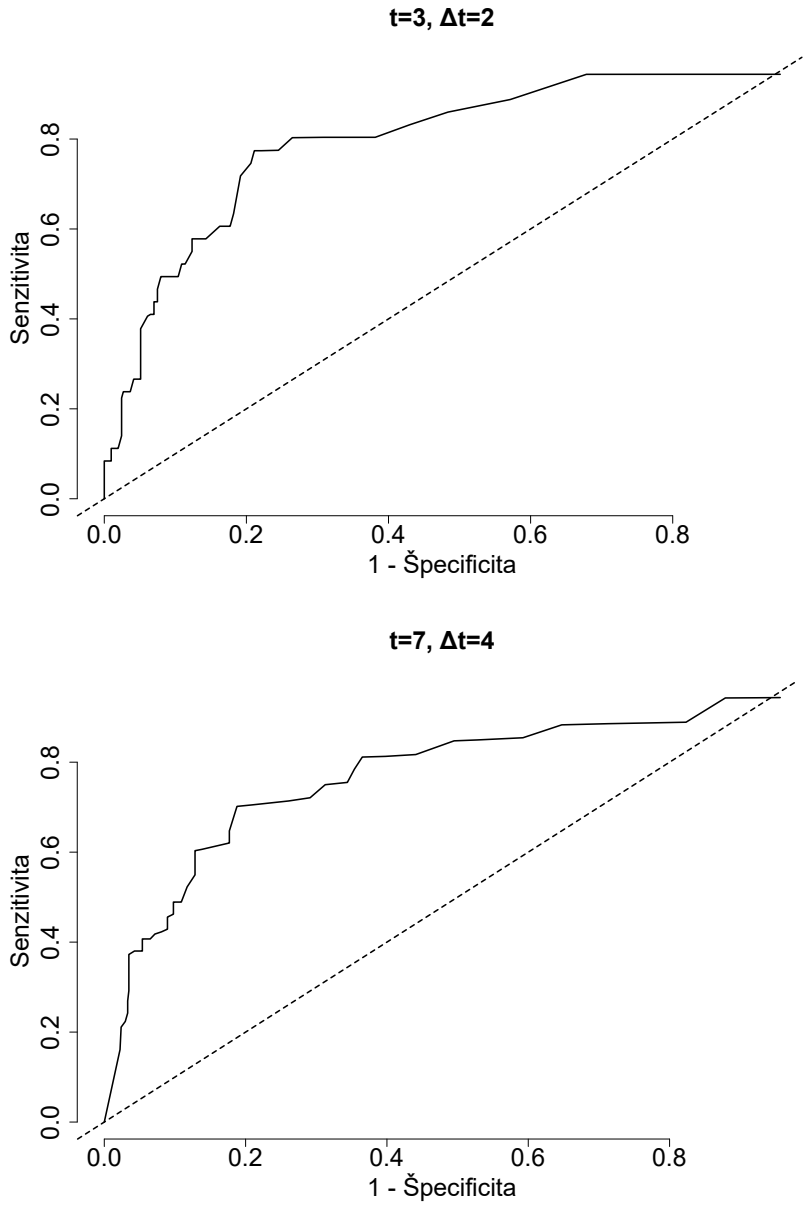
Na obrázku 4.4 sú vykreslené ROC krivky v časoch 5 a 11 s využitím longitudinálnych meraní do časov 3 a 7. Na základe oboch kriviek by sme zhodnotili, že model má pomerne dobrú diskriminačnú schopnosť a drobný rozdiel je viditeľný viac na vyčíslenonom príslušnom AUC v tabuľke 4.4, než na krivkách.

	$\widehat{\mathrm{AUC}}_t^{\Delta t}$		$\widehat{PE}(t + \Delta t\|t)$		$\widehat{IPE}(t + \Delta t\|t)$	
	$\Delta t = 2$	$\Delta t = 4$	$\Delta t = 2$	$\Delta t = 4$	$\Delta t = 2$	$\Delta t = 4$
$t = 3$	0,803	0,834	0,098	0,139	0,056	0,092
$t = 5$	0,850	0,842	0,111	0,145	0,069	0,093
$t = 7$	0,767	0,784	0,117	0,163	0,069	0,113

Tabuľka 4.4: Miery presnosti predikcie pre model (4.1)

	$\Delta t = 2$	$\Delta t = 4$
$\hat{C}_{\mathrm{dyn}}^{\Delta t}$	0,794	0,780

Tabuľka 4.5: Dynamický index diskriminácie pre model (4.1).



Obr. 4.4: ROC krivky pre model (4.1) zhora v čase 3 pre $\Delta t = 2$ a v čase 7 pre $\Delta t = 4$.

5. Simulačná štúdia

V tejto záverečnej kapitole práce sa budeme zaoberať simulačnou štúdiou, ktorej cieľom je porovnať výstupy z Coxovho modelu a združeného modelu a vyhodnotiť presnosť odhadov združeného modelu. Pre modely budeme simulovať dáta za rôznych scenárov, ktoré zahŕňajú zvyšujúci sa počet pacientov, zvyšujúci sa počet longitudinálnych meraní a rastúce percento cenzorovania. Na výpočty a spracovanie dát bol opäť využitý software R [R Core Team, 2023] a balík JMbayes [Rizopoulos, 2016]. Ako prvé popíšeme simulovanie dát a ciele, ktoré sledujeme a následne popíšeme a prediskutujeme získané výsledky.

5.1 Popis a ciele simulácií

V simulačnej štúdii uvažujeme nasledujúcu situáciu: pacientov sledujeme po dobu jedného roku, počas ktorého zaznamenávame longitudinálne merania. Všetci pacienti majú merania robené približne v rovnakom čase, ekvidistančne rozloženom v priebehu roku. Niektorí pacienti po tretej návšteve (alebo neskôr) prestanú chodiť na kontroly a strácame o nich akúkoľvek informáciu. Čas poslednej kontroly sa teda stáva časom cenzorovania. U pacientov, ktorí chodili na vyšetrenia počas celej doby sledovania, máme informáciu o čase úmrtia. Zaujíma nás správanie sa odhadov a predikcií na základe združeného modelu, ak budeme zvyšovať počet pacientov, počet návštev a percento cenzorovania. Pre rôzne scenáre volíme počet pacientov $K = 50,\ 100,\ 150$, počet návštev $n = 4,\ 7,\ 9$ a percento cenzorovania $p = 10\%,\ 30\%,\ 50\%$. Dáta sú generované vždy pre $K + 4$ pacientov, s nastavením *set.seed(123456)*. Na základe údajov o K pacientoch sú budované modely a pre zvyšných 4 pacientov sú počítané predikcie podmienených pravdepodobností prežitia. Podrobnejší popis výberu pacientov na predikovanie uvedieme neskôr. V simuláciách pracujeme s takmer rovnakým modelom ako v ilustračnej analýze. Jediný rozdiel v modeloch je ten, že v simuláciách nebudeme uvažovať nesignifikantný efekt liečby. Pri simulovaní dát volíme hodnoty koeficientov na základe ilustračnej analýzy, teda $\boldsymbol{\beta} = (\beta_1,\beta_2)^\top = (0{,}5; 0{,}2)^\top$, $\boldsymbol{\gamma} = (\gamma_1,\gamma_2)^\top = (0{,}05; 0{,}1)^\top$, $\alpha = 1{,}3$, $\sigma^2 = 0{,}3$ a $\text{vec}(\mathbb{D}) = (d_{11},d_{12},d_{12},d_{22})^\top = (1; 0{,}08; 0{,}08; 0{,}03)^\top$. Časy longitudinálnych meraní generujeme ekvidistančne na intervale $[0; 1]$, v závislosti na počte návštev, s pridaním šumu z rovnomerného rozdelenia na $[-0{,}05; 0{,}05]$ (aby simulácie viac kopírovali reálnu situáciu, v ktorej nechodia všetci pacienti presne v rovnakom čase). Ďalej generujeme náhodné efekty (absolútny člen a smernicu) a chybové členy z normálneho rozdelenia s nulovou stredou hodnotou, po rade rozptylmi d_{11}, d_{22} a σ^2, všetky pre jednoduchosť nezávislé. Odozva z lineárneho zmiešaného modelu je následne počítaná predpisom $\mathbf{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,\ i = 1,\ldots,K+4$, kde $\mathbb{X}_i = \mathbb{Z}_i$ sú po rade matice fixných a náhodných efektov obsahujúce v stĺpcoch vektor jednotiek a časy longitudinálnych meraní i-teho pacienta. Pohlavie pacientov generujeme z alternatívneho rozdelenia s pravdepodobnosťou ženského pohlavia 0,7 (znížená pravdepodobnosť z ilustračnej analýzy) a vek pacientov generujeme z rovnomerného rozdelenia na intervale $[20; 80]$. Pre čas prežitia uvažujeme exponenciálne rozdelenie, teda základné riziko je tvaru $\lambda_0(t,\boldsymbol{\xi}) = \lambda$ a volíme $\lambda = 0{,}001$. Čas prežitia následne generujeme na základe združeného modelu (5.2) podľa [Austin, 2012]. Simulované dáta teda spĺňajú model:

$$
\begin{aligned}
\textsf{SREM}: \ \log(\text{bili})(t_{ij}) &= 0{,}5 + 0{,}2 * t_{ij} + \mathrm{b}_i^0 + \mathrm{b}_i^1 * t_{ij} + \epsilon_{ij}, \\
\lambda_i(t|\mathrm{b}_i^0,\mathrm{b}_i^1) &= \exp\{0{,}001 + 0{,}05 * vek_i + 0{,}1 * \mathbb{1}\{\check{z}ena\}_i \\
&\quad 1{,}3 * [0{,}5 + 0{,}2 * t_{ij} + \mathrm{b}_i^0 + \mathrm{b}_i^1 * t_{ij}]\}, \\
&\quad j = 1, \ldots, n, \ i = 1, \ldots K + 4. \qquad (5.1)
\end{aligned}
$$

Z dôvodu dosiahnutia požadovaného percenta cenzorovania, u všetkých pacientov predpokladáme, že na intervale $[0; 1]$, s pravdepodobnosťou rovnou jednej, nikto nezomrie. Aby sme zachovali nezávislé cenzorovanie a zároveň kontrolovali percento cenzorovania, generujeme čas cenzorovania nasledovne:

1. Náhodne volíme pacientov, ktorí budú cenzorovaní, z dôvodu odchodu zo štúdie, podľa daného percenta cenzorovania.

2. Náhodne volíme poslednú návštevu daného pacienta (všetci pacienti však musia absolvovať aspoň prvé tri longitudinálne merania).

3. Ako čas cenzorovania berieme čas poslednej návštevy u vyššie uvedených náhodne vybraných pacientov.

Cenzorovaný čas udalosti následne spočítame štandardne ako minimum času cenzorovania a času udalosti, pre každého pacienta. U pacientov, u korých došlo k cenzorovaniu, je ešte potrebné upraviť longitudinálne dáta tak, aby neobsahovali údaje po cenzorovaní. Výsledky predikcií chceme ilustrovať pre pacientov s rôznym časom úmrtia, resp. cenzorovania. Pacientov, pre ktorých budú počítané predikcie volíme tak, aby mali postupne zvyšujúci sa (cenzorovaný) čas udalosti, teda $T_i \approx 0{,}66;\ 1{,}3,\ 3,\ 6$. Pri výbere pacientov, ktorí budú cenzorovaní tak navyše vyberáme 1 pacienta, aby sme mali zachované percento cenzorovaných pacientov pri budovaní modelov. Ku každému modelu spočítame miery diskriminácie a kalibrácie. Z dôvodu časovo náročných výpočtov počítame iba $\widehat{\mathrm{AUC}}_t^{\Delta t}$, $\widehat{PE}(t+\Delta t|t)$ a $\hat{C}_{\mathrm{dyn}}^{\Delta t}$ pre $t = 3$, $\Delta t = 2$ a $t = 7$, $\Delta t = 4$. Cieľom simulácií je ukázať, že s rastúcim počtom pacientov, rastúcim počtom návštev a znižujúcim sa percentom cenzorovania, budú odhady zo združeného modelu, oproti klasickému Coxovmu modelu, presnejšie. Taktiež očakávame väčší rozdiel medzi odhadnutými funkciami prežitia z klasického Coxovho modelu a združeného modelu. Na kvantifikovanie rozdielu medzi funkciami prežitia využijeme suprémovú metriku, ktorú spočítame ako testovú štatistiku z dvojvýberového Kolmogorovho-Smirnovho testu. Každý scenár zreplikujeme v $d = 100$ vygenerovaných datasetoch a výsledky budeme prezentovať ako priemer výstupov z každého datasetu, pre jednotlivé voľby K, n, p.

5.2 Výsledky

V tejto časti zhrnieme výsledky simulácií. Ako bolo spomenuté na začiatku kapitoly, zaujímajú nás miery diskriminácie a kalibrácie, vzdialenosť odhadov funkcií prežitia a presnosť bodových odhadov parametrov združeného modelu pre rôzne scenáre. V simuláciach sa zaoberáme odhadmi parametrov a predikciami na základe združeného modelu (5.2) a klasického Coxovho modelu (5.3).

$$\begin{aligned}\mathsf{SREM}: \ \log(\text{bili})(t_{ij}) &= \beta_1 + \beta_2 * t_{ij} + \mathrm{b}_i^0 + \mathrm{b}_i^1 * t_{ij}, \\ \lambda_i(t|\mathrm{b}_i^0,\mathrm{b}_i^1) &= \exp\{\xi_0 + \sum_{j=1}^{8} \xi_j B_j(t,\mathbf{v}) + \gamma_1 * vek_i + \gamma_2 * \mathbb{1}\{\check{z}ena\}_i \\ &\quad + \alpha[\beta_1 + \beta_2 * t_{ij} + \mathrm{b}_i^0 + \mathrm{b}_i^1 * t_{ij}]\}, \\ &\quad j = 1,\ldots,n_i, \ i = 1,\ldots K. \end{aligned} \tag{5.2}$$

$$\begin{aligned}\mathsf{Cox}: \lambda(t|vek,pohlavie,log(bili)(t)) &= \lambda_0(t) + \tilde{\gamma}_1 * vek + \tilde{\gamma}_2 * \mathbb{1}\{\check{z}ena\} \\ &\quad + \tilde{\gamma}_3 * log(bili)(t). \end{aligned} \tag{5.3}$$

Združený model bol v každom zo scenárov použitý z dôvodu výpočetnej náročnosti len s 20 000 iteráciami MCMC, s predvoľbou 3000 iterácii v zahrievacej fáze a 3000 iterácií na adaptovanie modelu. Na zníženie autokorelácie v rámci reťazcov bol použitý stenčovací parameter $\mathsf{n.thin} = 2$. Uzly a rád splajnov boli ponechané opäť v predvoľbe, t.j. zvolený počet uzlov bol 5 a použitá bola splajnovová funkcia rádu 4. Apriorné rozdelenia boli, v značení z modelu 2.1.2, volené nasledovne:

- $\boldsymbol{\beta} \sim \mathsf{N}_2(\mathbf{0},\text{diag}(0{,}001; 0{,}001))$,
- $\tau \sim \Gamma(1; 0{,}005)$,
- $\mathbb{D}$, $\mathbb{D}^{-1} \sim \mathsf{W}_q(2,\text{diag}(0{,}001; 0{,}001))$,
- $\alpha \sim \mathsf{N}(0; 0{,}01)$,
- $\boldsymbol{\gamma} \sim \mathsf{N}_(\mathbf{0},\text{diag}(0{,}005; 0{,}001))$,
- $\xi_0,\ldots,\xi_8 \sim \mathsf{N}(0; 0{,}001)$.

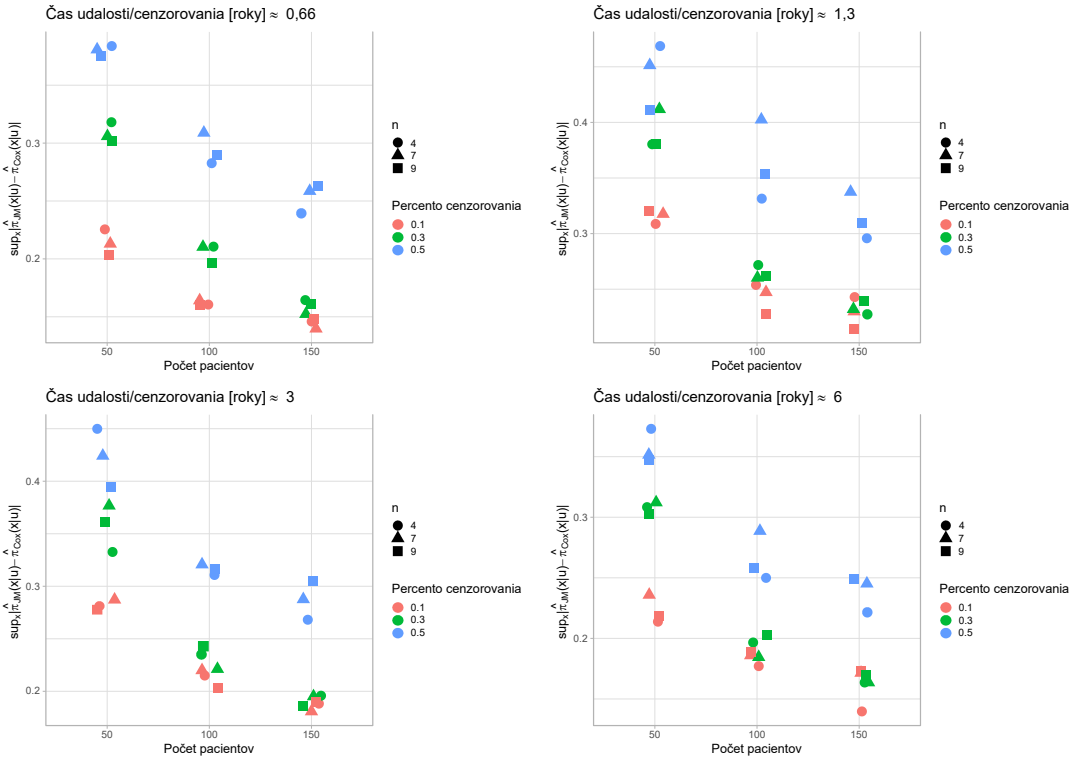
V tabuľke 5.1 sú uvedené priemerné hodnoty odhadov AUC, dynamického indexu predikcie a chyby predikcie, počítané zo 100 aplikácií modelu 5.2 pre každú kombináciu počtu pacientov, počtu návštev a percenta cenzorovania. Na základe hodnôt pre každý z 27 scenárov vidíme, že model má pomerne dobrú diskriminčnú schopnosť ($\widehat{\text{AUC}}$, $\hat{C}_{\text{dyn}}^{\Delta t} > 0{,}7$). Vidíme, že hodnoty AUC a dynamického indexu predikcie sa pri zvyšujúcom sa počte pacientov zlepšujú, aj keď sa nejedná o výrazný trend. S rastúcim percentom cenzorovania pozorujeme zreteľnejší pokles AUC a dynamického indexu. U chyby predikcie nevidieť žiadny trend. Zdá sa, že zvyšujúci sa počet návštev nemá vplyv na žiadnu z diskriminačných a kalibračných mier.

Na obrázku 5.1 je vykreslená priemerná suprémová vzdialenosť odhadnutej podmienenej funkcie prežitia zo združeného modelu a Coxovho modelu. Ľavý horný obrázok zobrazuje situáciu pre pacientov, u ktorých došlo k cenzorovaniu približne v 8 mesiacoch (0,66 roku). Zvyšné 3 obrázky zobrazujú situáciu pre pacientov, u ktorých došlo k udalosti. Časový bod u, na osi y, je vo všeobecnosti čas posledného longitudinálneho merania, ktorý bol u každého pacienta v každom

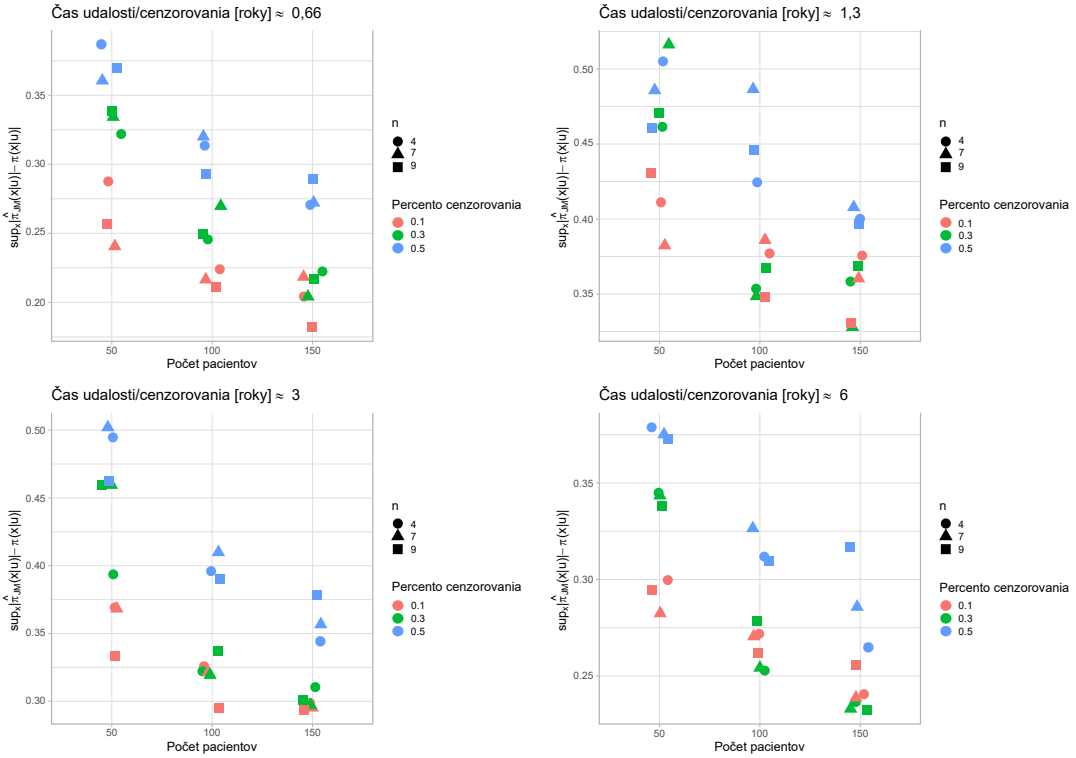
Počet pacientov	Počet návštev	Percento cenzorovania	$\widehat{\mathrm{AUC}}_3^2$	$\widehat{\mathrm{AUC}}_7^4$	$\hat{C}_{\mathrm{dyn}}^2$	$\hat{C}_{\mathrm{dyn}}^4$	$\widehat{PE}(5\|3)$	$\widehat{PE}(11\|7)$
K=50	n=4	p=0,1	0,85	0,81	0,80	0,81	0,09	0,17
		p=0,3	0,78	0,74	0,73	0,74	0,10	0,19
		p=0,5	0,73	0,75	0,72	0,73	0,09	0,18
	n=7	p=0,1	0,84	0,83	0,80	0,81	0,09	0,16
		p=0,3	0,75	0,76	0,73	0,74	0,11	0,18
		p=0,5	0,70	0,75	0,73	0,74	0,10	0,18
	n=9	p=0,1	0,83	0,83	0,81	0,82	0,10	0,15
		p=0,3	0,77	0,75	0,74	0,75	0,10	0,18
		p=0,5	0,73	0,73	0,74	0,75	0,09	0,19
K=100	n=4	p=0,1	0,84	0,82	0,79	0,80	0,09	0,16
		p=0,3	0,81	0,82	0,78	0,79	0,10	0,16
		p=0,5	0,76	0,78	0,75	0,76	0,10	0,18
	n=7	p=0,1	0,84	0,82	0,80	0,81	0,09	0,16
		p=0,3	0,82	0,80	0,78	0,79	0,09	0,17
		p=0,5	0,78	0,77	0,75	0,76	0,11	0,18
	n=9	p=0,1	0,83	0,81	0,81	0,82	0,09	0,17
		p=0,3	0,84	0,82	0,80	0,81	0,09	0,16
		p=0,5	0,78	0,77	0,76	0,77	0,11	0,19
K=150	n=4	p=0,1	0,83	0,82	0,79	0,80	0,10	0,17
		p=0,3	0,83	0,82	0,79	0,80	0,10	0,16
		p=0,5	0,80	0,80	0,76	0,77	0,11	0,17
	n=7	p=0,1	0,82	0,83	0,80	0,81	0,10	0,16
		p=0,3	0,84	0,83	0,80	0,81	0,09	0,16
		p=0,5	0,78	0,79	0,75	0,76	0,12	0,18
	n=9	p=0,1	0,84	0,83	0,81	0,82	0,09	0,16
		p=0,3	0,83	0,82	0,80	0,82	0,10	0,17
		p=0,5	0,79	0,78	0,76	0,78	0,12	0,18

Tabuľka 5.1: Priemerné odhadnuté miery diskriminácie a kalibrácie spomedzi 100 modelov pre každý scenár.

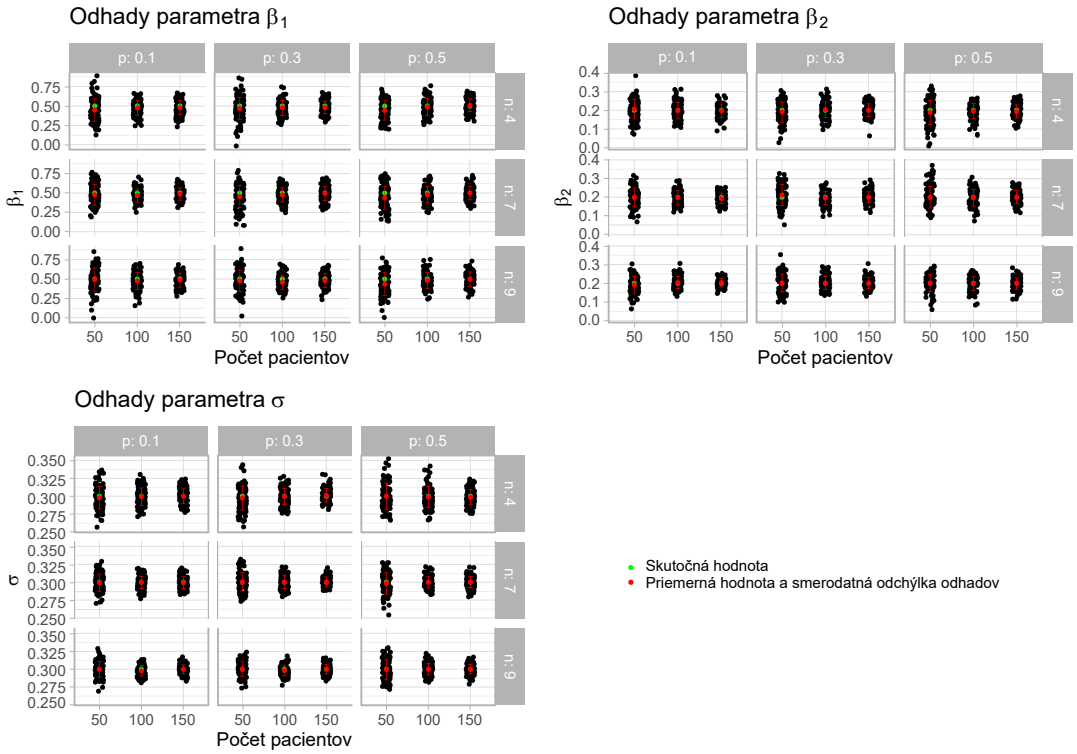
datasete iný. Máme iba informáciu, že sa jednalo približne o čas 0,66 roku, v prípade ľavého horného obrázku, a približne o 1 rok, v prípade zvyšných obrázkov. Čas u preto nechávame vo všeobecnom zápise. Na obrázku vidíme, že s rastúcim počtom pacientov, na ktorých boli budované modely, sú si odhady bližšie. Tak isto, ako by sme aj čakali, sa zdá, že s klesajúcim percentom cenzorovania klesá aj vzdialenosť odhadov. Pri zvyšujúcom sa počte návštev nie je na obrázkoch viditeľný monotónny trend. Na základe obrázkov sa zdá, že s neskorším skutočným časom udalosti, je vzdialenosť medzi odhadmi funkcie prežitia menšia. Na obrázku 5.2 je znázornená priemerná suprémová vzdialenosť odhadnutej podmienenej funkcie prežitia zo združeného modelu a skutočnej podmienenej funkcie prežitia, pre každý scenár. Opäť vidíme, že najväčší vplyv na vzdialenosť funkcií má počet pacientov zahrnutých v modeli (klesajúci trend vzdialenosti s rastúcim počtom pacientov). Vplyv rastúceho percenta cenzorovania už nie je taký jednoznačný, ako v prípade vzdialeností medzi odhadnutými funkciami. Aj v tomto prípade je však na obrázkoch stále viditeľný klesajúci trend vzdialeností s rastúcim percentom cenzorovania. Zdá sa, že zvyšujúci sa počet návštev výsledky opäť nezlepšuje. Vidíme, že vzdialenosť medzi odhadom a skutočnou funkciou prežitia, s rastúcim skutočným časom udalosti, klesá.



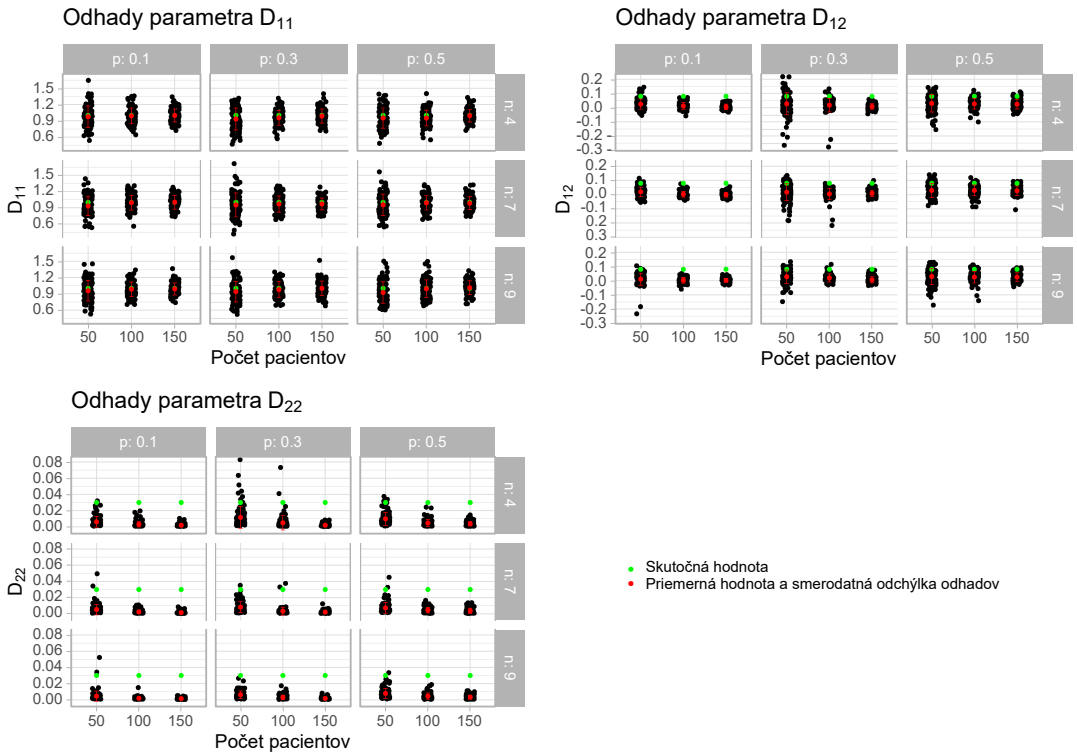
Obr. 5.1: Suprémová vzdialenosť odhadnutej podmienej funkcie prežitia zo združeného modelu (5.2) a odhadnutej podmienenej funkcie prežitia z Coxovho modelu (5.3). Meniaci sa počet návštev je odlíšený tvarom bodov a percento cenzorovania ja odlíšené farebne.



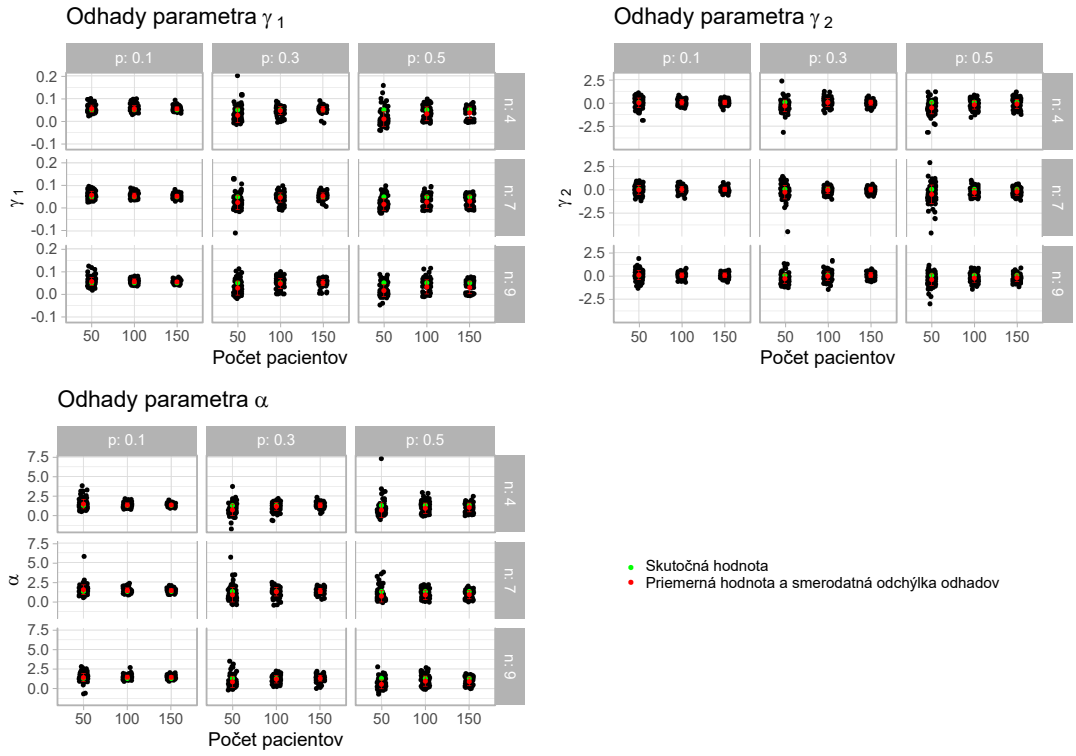
Obr. 5.2: Suprémová vzdialenosť odhadnutej podmienej funkcie prežitia zo združeného modelu (5.2) a skutočnej podmienenej funkcie prežitia. Meniaci sa počet návštev je odlíšený tvarom bodov a percento cenzorovania ja odlíšené farebne.



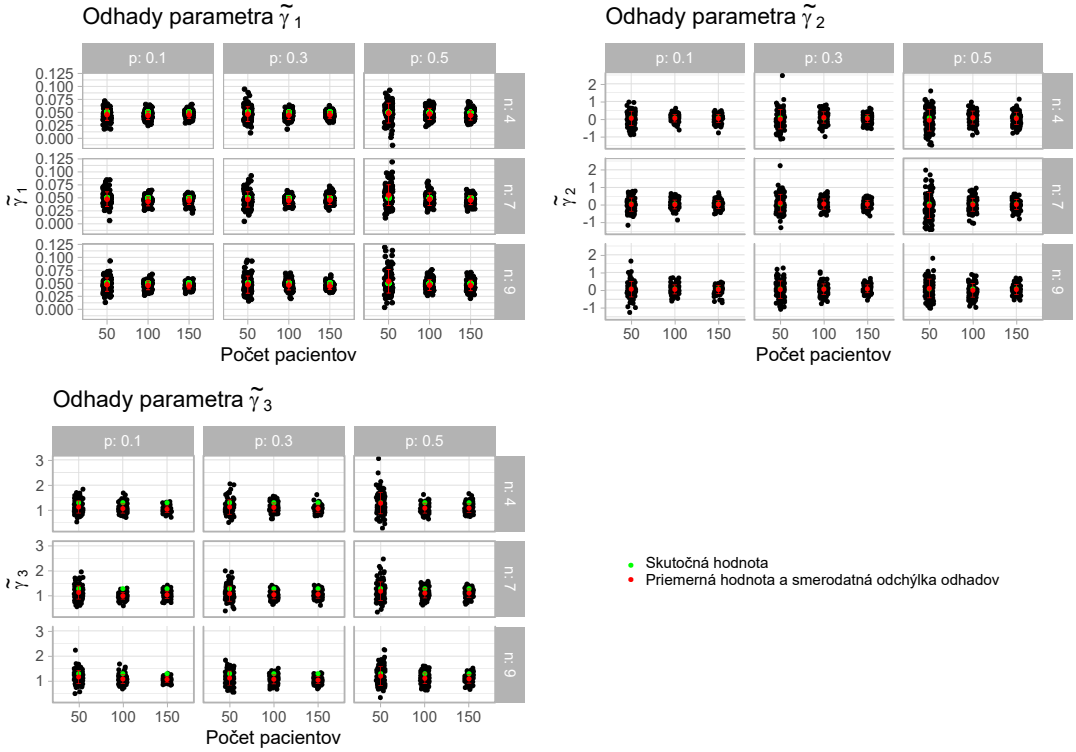
Obr. 5.3: Porovnanie bodových odhadov jednotlivých parametrov zo združeného modelu (5.2), spočítaných z každého datasetu (čiernou), so skutočnou hodnotou parametra (zelenou) a priemerným odhadom spomedzi všetkých modelov (červenou).



Obr. 5.4: Porovnanie bodových odhadov jednotlivých parametrov zo združeného modelu (5.2), spočítaných z každého datasetu (čiernou), so skutočnou hodnotou parametra (zelenou) a priemerným odhadom spomedzi všetkých modelov (červenou).



Obr. 5.5: Porovnanie bodových odhadov jednotlivých parametrov zo združeného (5.2), spočítaných z každého datasetu (čiernou), so skutočnou hodnotou parametra (zelenou) a priemerným odhadom spomedzi všetkých modelov (červenou).



Obr. 5.6: Porovnanie bodových odhadov jednotlivých parametrov z klasického Coxovho modelu (5.3), spočítaných z každého datasetu (čiernou), so skutočnou hodnotou parametra (zelenou) a priemerným odhadom spomedzi všetkých modelov (červenou).

Na obrázkoch 5.3, 5.4, 5.5 a 5.6 sú vykreslené bodové odhady parametrov zo združeného a Coxovho modelu pre každý scenár a každý spočítaný model. Zelenou je vyznačená skutočná hodnota parametra, s ktorou boli dáta generované a červenou je priemerná hodnota a smerodatná odchýlka medzi všetkými spočítanými modelmi pre každý scenár. Odhady parametra $\mathbb{D}_{22}$ (druhý diagonálny prvok variančnej matice náhodných efektov) sú úplne mimo skutočných hodnôt parametra. Príčinou je zrejme vysoká autokorelácia odhadov v rámci MCMC a malý počet iterácií. Pripomenieme, že v ilustračnej analýze sme volili desaťnásobne väčšie množstvo iterácií. Podobne vidíme, že aj reťazec pre odhad parametra $\mathbb{D}_{12}$ zrejme ešte neskonvergoval k stacionárnemu rozdeleniu. Pre zvyšné parametre združeného a Coxovho modelu pokrýva priemer odhadov pomerne dobre skutočnú hodnotu parametrov. U všetkých parametrov sa pri zvýšujúcom sa počte pacientov zlepšila presnosť odhadov (menšia smerodatná odchýlka odhadov). Vplyv počtu návštev na odhady parametrov nie je viditeľný. V prípade parametrov γ_1, γ_2 a α vyzerá, že vyššie percento cenzorovania priemerne zhoršilo hodnoty odhadov. Opäť treba zdôrazniť, že odhady zo združeného modelu a štandardného Coxovho modelu boli spočítané použitím dvoch rôznych princípov. Keď porovnáme odhady parametra α a $\tilde{\gamma}_3$, ktoré vyjadrujú efekt endogénnej časovej premennej, vidíme, že združený model dáva spoľahlivejšie výsledky. Tieto výsledky potvrdzujú hlavnú myšlienku použitia zduženého modelu.

5.3 Zhrnutie výsledkov simulácií

Výsledky simulácii čiastočne potvrdili naše očakávania. Ukázali sme, že rastúci počet pacientov a klesajúce percento cenzorovania zlepšuje predikcie funkcií prežitia. Výsledky taktiež ukázali, že rastúce percento cenzorovania mierne zhoršilo diskriminačné miery modelu a presnosť odhadov parametrov z modelu prežitia v rámci združeného modelu. Zvyšujúci sa počet pacientov zlepšil presnosť odhadov parametrov ako združeného, tak aj klasického Coxovho modelu. Pri meniacom sa počte návštev sme nenašli výrazné zmeny vo výsledkoch.

Situácia, v ktorej k cenzorovaniu dochádza iba počas doby longitudinálneho sledovania, neodpovedá úplne skutočnosti. Tento scenár však bol zvolený, aby sme boli schopní jednoducho kontrolovať percento cenzorovania. Zrejme aj z dôvodu použitia takéhoto cenzorovania vykazovali niektoré výpočty pre združený model výpočtové problémy. Dôvod, prečo toto nastávalo, zrejme súvisel s tým, že s rastúcim percentom cenzorovania sme strácali väčšie množstvo longitudinálnych dát. Z tohto dôvodu bola volená aj požiadavka, aby všetci pacienti mali aspoň 3 longitudinálne merania a aby s pravdepodobnosťou rovnou jednej, počas doby longitudinálneho sledovania, nedošlo k udalosti.

Záver

V tejto práci sme sa zaoberali dynamickou predikciou pravdepodobnosti prežitia na základe združeného modelu. Zaviedli sme definíciu a použitie lineárneho zmiešaného modelu, stručne sme zopakovali frekventistické metódy odhadovania parametrov zmiešaného modelu. Čitateľa sme oboznámili so základnými pojmami z analýzy cenzorovaných dát, uviedli sme neparametrické odhady rozdelenia času zlyhania a zadefinovali sme Coxov model proporčných rizík. V závere kapitoly sme popísali základné princípy bayesovskej štatistiky a metódu MCMC na výpočty odhadov, konkrétne Gibbsov a Metropolisov-Hastingsov algoritmus.

Zaoberali sme sa dvoma špecifickými typmi združených modelov. Zadefinovali sme združený model s náhodnými efektami, predpoklady pre platnosť modelu a uviedli sme rôzne možné funkcionálne formy na vyjadrenie závislosti medzi longitudinálnou premennou a rizikom udalosti. Vlastný prínos v tejto časti práce bol v bayesovskom odhadovaní, konkrétne v podrobnom odvodení plne podmienených rozdelení parametrov pre Gibbsov algoritmus. Ďalší typ združených modelov, ktorý sme stručne popísali, boli združené modely s latentnými kategóriami.

V kapitole o dynamickej predikcii sme zadefinovali individuálnu pravdepodobnosť prežitia a vysvetlili princíp dynamickosti predikcií. Pomocou bayesovskej teórie sme odvodili odhad podmienenej pravdepodobnosti prežitia. Na vyhodnotenie kvality predikcií sme uviedli miery presnosti predikcie a ich konzistentné odhady. V závere kapitoly sme zhrnuli kalibračné miery spolu s ich odhadmi.

Teoretické poznatky sme aplikovali v ilustračnej analýze dát k primárnej biliárnej cirhóze. Porovnali sme odhady parametrov a predikcií podmienených pravdepodobností prežitia zo združeného a klasického Coxovho modelu. Viditeľné rozdiely vo výstupoch z modelov sa neprejavili na samotných odhadoch parametrov, ale až na predikciách podmienených funkcií prežitia. Zvolený združený model vykazoval uspokojivú diskrimináciu aj kalibráciu.

Vlastným prínosom bola aj celá záverečná časť práce, venovaná simulačnej štúdii. V štúdii sme ukázali, že pri zvyšujúcom sa počte pacientov sa zlepšujú odhady parametrov, diskriminačná schopnosť modelu a predikcie podmienenej pravdepodobnosti prežitia. Taktiež sa zmenšuje suprémová vzdialenosť odhadov podmienenej pravdepodobnosti prežitia zo združného modelu a Coxovho modelu. Zvyšujúce sa percento cenzorovania, ako sme predpokladali, spôsobilo mierne zhoršenie vyššie zmienených výstupov modelu. Meniaci sa počet návštev, prekvapivo, viditeľne neovplyvnil výsledky.

Zoznam použitej literatúry

O. Aalen. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701 – 726, 1978.

E.-R. Andrinopoulou, D. Rizopoulos, and J. Eilers, P.and Takkenberg. Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using p-splines. *Biometrics*, 74, 09 2016.

P. C. Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958, 2012.

P. Blanche, J.-F. Dartigues, and H. Jacqmin-Gadda. Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704, 2013.

N. E. Breslow. Contribution to the discussion on the paper by d. r. cox, regression and life tables. *Journal of the Royal Statistical Society*, 34:216–217, 1972.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

D. R. Cox. Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10(1):1–7, 1989.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89 – 121, 1996.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

T. Hanson, A. Branscum, and W. Johnson. Predictive comparison of joint longitudinal-survival modeling: A case study illustrating competing approaches. *Lifetime data analysis*, 17:3–28, 01 2011.

F. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag, New York, 2001.

D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385, 08 1974. ISSN 0006-3444.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444.

D. S. Hawkins, D. M. Allen, and A. J. Stromberg. Determining the number of components in mixtures of linear models. *Computational Statistics & Data Analysis*, 38(1):15–48, 2001. ISSN 0167-9473.

Ch. R. Henderson. Applications of linear models in animal breeding. 1984.

R. Henderson, P. Diggle, and A. Dobson. Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3(1):33–50, 03 2002.

J. R. Hipp and D. J. Bauer. Local solutions in the estimation of growth mixture models. *Psychological methods*, 11(1):36, 2006.

J. W. Hogan and N. M. Laird. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16(3):259–272, 1997.

X. Huang, G. Li, R. M. Elashoff, and J. Pan. A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis*, 17(1):80–100, 2011.

M. Hušková. *Bayesovské metody: skripta pro posl. matematicko-fyz. fakulty Univ. Karlovy*. Univerzita Karlova, 1985.

J. G. Ibrahim and D. Chen, M.-H.and Sinha. *Bayesian Survival Analysis*. Springer New York, New York, NY, 2001.

R. I. Jennrich and M. D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986. ISSN 0006341X, 15410420.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. Second Edition. John Wiley & Sons, Hoboken, 2002. ISBN 0-471-36357-X.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

A. Komárek. Linear regression. `http://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2021_22/nmsa407/2021-NMSA407-notes.pdf`, September 2021. Posledný prístup: 2-9-2021.

M. Kulich. Censored data analysis. `https://www2.karlin.mff.cuni.cz/~kulich/vyuka/cens/doc/cens_notes_ext_220102.pdf`, October 2021. Posledný prístup: 18-10-2021.

M. Laird, N. and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. ISSN 0006341X, 15410420.

V. Leiva-Yamaguchi and D. Alvares. A two-stage approach for bayesian joint models of longitudinal and survival data: Correcting bias with informative prior. *Entropy*, 23:1–10, 12 2020.

L. Li, T. Greene, and B. Hu. A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical methods in medical research*, 27(8):2264–2278, 2018.

M. J. Lindstrom and D. M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988. ISSN 01621459.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1993.

P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips. Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1):126–134, 1994.

W. Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1:27–52, 1969.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. ISSN 00063444.

W. Penny, K. Friston, J. Ashburner, S. Kiebel, and T. Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. 01 2007.

C. Proust-Lima, P. Joly, J.-F. Dartigues, and H. Jacqmin-Gadda. Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational statistics & data analysis*, 53(4):1142–1154, 2009.

C. Proust-Lima, M. Séne, J. Taylor, and H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1):74–90, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

D. Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press, Boca Raton, 2012.

D. Rizopoulos. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–45, 2016. doi: 10.18637/jss.v072.i07.

D. Rizopoulos and P. Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12):1366–1380, 2011.

Ch. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, New York, 2007.

M. Schemper and R. Henderson. Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1):249–255, 2000.

A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14(3):809–834, 2004. ISSN 10170405, 19968507.

A. W. van der Vaart. *Asymptotic Statistics*. Prvé vydanie. Cambridge University Press, Cambgridge, 1998. ISBN 0-521-78450-6.

J. Vorlíčková. Joint models for longitudinal and time-to-event data. Diploma Thesis, Faculty of Mathematics and Physics, Charles University, 2020.

M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

J. Xu and S. L. Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50(3):375–387, 2001. ISSN 00359254, 14679876.

Zoznam obrázkov

Zoznam tabuliek

A. Prílohy

A.1 Užitočné rozdelenia

Zhrnutie základných vlastností rozdelení použitých v práci. Rozdelenia sú definované pre náhodnú veličinu X, náhodný vektor $\mathbf{X} \in \mathbb{R}^k$ alebo náhodnú maticu $\mathbb{M} \in \mathbb{R}^{k \times k}$.

A.1.1 Weibullovo rozdelenie

$X \sim \mathsf{Weibull}(\gamma, \lambda), \lambda > 0, \gamma > 0$

- Hustota: $f(x;\gamma,\lambda) = \begin{cases} \gamma\lambda x^{\gamma-1} e^{-x^\gamma \lambda}, & x \geq 0, \\ 0, & x < 0, \end{cases}$
- Stredná hodnota: $\mathsf{E}\ X = \lambda^{-\frac{1}{\gamma}} \Gamma(1 + \gamma^{-1})$
- Rozptyl: $\operatorname{var} X = \lambda^{-\frac{2}{\gamma}} \left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \left(\Gamma\left(1 + \frac{1}{\gamma}\right)\right)^2 \right]$

A.1.2 Gamma rozdelenie

$X \sim \Gamma(a, p), a > 0, p > 0$

- Hustota: $f(x;a,p) = \begin{cases} \frac{p^a}{\Gamma(a)} x^{a-1} e^{-px}, & x \geq 0, \\ 0, & x < 0, \end{cases}$
- Stredná hodnota: $\mathsf{E}\ X = \frac{a}{p}$
- Rozptyl: $\operatorname{var} X = \frac{a}{p^2}$

A.1.3 Mnohorozmerné normálne rozdelenie

$\mathbf{X} \sim \mathsf{N}_k(\boldsymbol{\mu},\boldsymbol{\Sigma}), \boldsymbol{\mu} \in \mathbb{R}^k, \boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$

- Hustota: $f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-k/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$, existuje iba pre $\boldsymbol{\Sigma}$ pozitívne definitnú.
- Stredná hodnota: $\mathsf{E}\ \mathbf{X} = \boldsymbol{\mu}$
- Variančná matica: $\operatorname{Var} \mathbf{X} = \boldsymbol{\Sigma}$

A.1.4 Wishartovo rozdelenie

Označme maticu $\mathbb{X}^{k \times n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$, $\mathbf{X}_i \sim \mathsf{N}_k(\mathbf{0},\boldsymbol{\Sigma}), \boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$, potom $\mathbb{M}^{k \times k} = \mathbb{X}\mathbb{X}^T = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \sim \mathsf{Wish}_k(n,\boldsymbol{\Sigma})$, kde $n > k - 1$ sú stupne voľnosti rozdelenia a $\boldsymbol{\Sigma}$ je škálovacia matica. Jedná sa o mnohorozmerné zovšeobecnenie chí-kvadrát rozdelenia.

- Hustota: $f(\mathbb{M};\boldsymbol{\Sigma},n) = \frac{|\mathbb{M}|^{(n-k-1)/2} e^{-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbb{M})/2}}{2^{\frac{nk}{2}} |\boldsymbol{\Sigma}|^{n/2} \Gamma_k(\frac{n}{2})}$, a Γ_k je mnohorozmerná gamma funkcia $\Gamma_k\left(\frac{n}{2}\right) = \pi^{\frac{k(k-1)}{4}} \prod_{j=1}^{k} \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right)$ podmienky pozitivity, tj. $\Theta = \prod_{i=1}^{p} \Theta_i$, $\Theta = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}) > 0\}$,
- Stredná hodnota:$\mathsf{E}\ \mathbb{M} = n\boldsymbol{\Sigma}$
- Variančná matica: $\mathrm{var}\, \mathbb{M} = n(\boldsymbol{\Sigma}_{ij}^2 + \boldsymbol{\Sigma}_{ii}\Sigma_{jj})_{i,j=1}^{k,k}$