Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Lukáš Bandas

Klasifikace na základě longitudinálních pozorování

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Arnošt Komárek, Ph.D.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2012

Na tomto místě bych rád poděkoval především vedoucímu práce RNDr. Arnoštu Komárkovi, Ph.D. za poskytnutá data, mnoho cenných rad a připomínek při konzultacích a při psaní tohoto textu. Dále bych rád poděkoval za podporu, jazykovou korekturu a četné připomínky svojí partnerce Michaele Tiché.

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

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

V Praze dne 11.04.2012 Lukáš Bandas

Název práce: Klasifikace na základě longitudinálních pozorování

Autor: Lukáš Bandas

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá klasifikací obecně různých objektů na základě longitudinálních pozorování. Čtenáře seznámí s lineárním smíšeným modelem a jeho základními vlastnostmi, který je vhodný pro modelování dat longitudinálního typu. Hlavní část práce se zaměřuje na popis metod diskriminační analýzy, které jsou vhodné pro klasifikaci na základě longitudinálních dat. Jednotlivé metody jsou nejprve se sjednoceným značením představeny z teoretického hlediska. Metoda s rozdělením náhodných efektů je zobecněna na spojitý čas. Poté jsou jednotlivé metody a vlastnosti lineárního smíšeného modelu aplikovány na reálná data. V poslední části jsou zkoumány vlastnosti uvedených metod v navržených simulačních studiích.

Klíčová slova: lineární smíšený model, longitudinální data, diskriminační analýza, Bayesova věta

Title: Classification based on longitudinal observations

Author: Lukáš Bandas

Department: Department of Probability and Mathematical Statistics

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

Abstract: The concern of this thesis is to discuss classification of different objects based on longitudinal observations. In the first instance the reader is introduced to a linear mixed-effects model which is useful for longitudinal data modeling. Description of discriminant analysis methods follows. These methods ares usually used for classification based on longitudinal observations. Individual methods are introduced in the theoretic aspect. Random effects approach is generalized to continuous time. Subsequently the methods and features of the linear mixed-effects model are applied to real data. Finally features of the methods are studied with help of simulations.

Keywords: linear mixed-effects model, longitudinal data, discriminant analysis, Bayes' theorem

Obsah

Seznam použitého značení

- $\boldsymbol{\beta}$ ... pevné efekty v lineárním smíšeném modelu
- $\mathbf{b}_i$ ... náhodné efekty i-tého subjektu v lineárním smíšeném modelu
- $\mathbb{D}$ ... varianční matice náhodných efektů
- $\boldsymbol{\Sigma}_i$ ... varianční matice chybového vektoru i-tého subjektu v lineárním smíšeném modelu
- $\boldsymbol{\alpha}$ ... všechny parametry varianční struktury lineárního smíšeného modelu
- $\boldsymbol{\theta}$ ... všechny parametry lineárního smíšeného modelu
- $\mathbf{L}_{ML}(\boldsymbol{\theta})$ ... věrohodnostní funkce
- $\mathbf{L}_{REML}(\boldsymbol{\alpha})$ ... věrohodnostní funkce při odhadu metodou REML

Kapitola 1

Úvod

Hlavním cílem úvodní kapitoly je seznámit čtenáře s obsahem a směrem této práce. První část nás nejprve motivuje a seznámí se zkoumanými problémy. V další části je pak stručně popsána náplň následujících kapitol.

1.1 Motivace

S pozorováními longitudinálního typu se setkáváme v mnoha oblastech lidského zkoumání. Za příklad může sloužit třeba sociální výzkum, kde nás můžou zajímat změny v mentální oblasti po aplikování různých terapií či spotřební chování populace po shlédnutí nějaké reklamy. Nejčastěji se však setkáváme s longitudinálními pozorováními v lékařské oblasti. Můžeme říci, že pozorování longitudinálního typu se vyskytují všude tam, kde u subjektů, které vstupují do výzkumu opakovaně (nejčastěji v čase), zjišťujeme konkrétní údaje, které jsou předmětem našeho zájmu. V závislosti na druhu výzkumu se může např. jednat o hladinu nějakého enzymu či proteinu v těle, metrické údaje při zátěžové studii výrobků či v ekonomické oblasti nějaké vlastnosti ovlivňující poskytnutí bankovního úvěru klientovi.

Takto napozorované hodnoty lze následně použít ke klasifikaci daných objektů do různých skupin podobně jako v klasické diskriminační nebo shlukové analýze. Tyto metody jsou v klasické interpretaci založeny na výběru z vícerozměrného normálního rozdělení a předpokládají, že v rámci jednotlivých skupin jsou pozorování nezávislá a stejně rozdělená. Typická longitudinální pozorování se vyznačují dvěma charakteristickými rysy, kterými jsou:

1) Počet pozorování u jednotlivých subjektů může být různý.

2) Jednotlivá pozorování nejsou pro všechny jednotky prováděna ve stejných časech.

V důsledku právě uvedených vlastností nemůžeme předpokládat, že náhodné vektory reprezentující napozorovaná data u jednotlivých subjektů tvoří náhodný výběr z vícerozměrného (normálního) rozdělení. V průběhu posledních let však bylo v literatuře popsáno několik různých přístupů pro klasifikaci na základě longitudinálních pozorování, které nejčastěji v nějaké podobě kombinují lineární smíšený model, který je velmi vhodný pro popis longitudinálních pozorování a klasické přístupy ke klasifikaci.

Klasifikaci na základě longitudinálních pozorování lze považovat za moderní diagnostický nástroj. Základní možnosti využití této metody si představíme na následujících motivačních příkladech.

Příklad 1.1. *Uvažujme studii, ve které máme výběr pacientů z nějaké mužské populace. U těchto pacientů se snažíme zjistit výskyt rakoviny prostaty, případně zjistit riziko této rakoviny. Studie probíhá tak, že se u jednotlivých pacientů provádí v průběhu času měření specifického antigenu - **glykoproteinu**, který je produkován prostatickým epitelem a jehož hladina je úzce spjata s objemem rakoviny v prostatě. Jednotlivá měření probíhají u různých pacientů v různých časových intervalech a může jich být i libovolný počet.*

Zřejmě se tedy jedná o longitudinální data. Snažíme se o sestavení modelu, který by dokázal modelovat nějakým způsobem hladinu glykoproteinu. Na základě tohoto modelu se pak snažíme pacienty klasifikovat pomocí diskriminační analýzy do několika skupin (např. riziko rakoviny, není riziko rakoviny).

Bližší informace k této problematice můžeme nalézt v článku Larry J. Branta a kol. [Brant et al., 2003].

△

Příklad 1.2. *Podobně jako v předešlém Příkladu 1.1 uvažujme tentokrát longitudinální studii těhotných žen. Je známo, že na dramatické změny během těhotenství ukazuje látka **beta-globin**. Naším cílem je stanovit, zda těhotenství u dané ženy bude nebo nebude rizikové na základě změn ve výši hladiny beta-globinu.*

Všimněme si, že jak v tomto, tak v předešlém příkladu jsme nikde neuvažovali podmínku, že okamžiky měření jednotlivých subjektů jsou stejné, případně, že intervaly mezi jednotlivými měřeními jsou stejně dlouhé, či počet měření u i-tého subjektu je konstantní. Podobně jako v Příkladu 1.1 vytvoříme model pro hladinu beta-globinu a na základě tohoto modelu se pomocí diskriminační analýzy snažíme klasifikovat jednotlivé ženy do jedné ze dvou (nebo případně více) skupin.

Více se o tomto tématu můžeme dočíst v článku autorů Marshalla G. a Barón E. A. [Marshall and Barón, 2000].

△

Příklad 1.3. *V následujícím příkladu se naše pozornost přesune ke studii zabývající se onemocněním jater. K dispozici máme několik set pacientů. U každého pacienta různý počet měření. Ze zkoumaných faktorů nás zajímá hladina bilirubinu, albuminu a počet krevních destiček v jistém objemu krve. Na základě těchto informací chceme pacienty klasifikovat do jedné ze dvou (případně více) skupin. K tomuto příkladu se budeme v dalších částech vracet, neboť budou použita k ukázce reálné aplikace popsaných metod*

△

Příklad 1.4. *V případě, že nás zajímá příklad longitudinálních dat z oblasti ekonomie či ekonomiky, můžeme na základě ukazatelů, kterými mohou být např. předchozí úvěrové chování žadatele, jeho aktuální finanční situace atd., klasifikovat potencionálního klienta banky, který žádá o úvěr, půjčku atd.*

V oblasti výroby můžeme naopak podrobovat opakovaně výrobky zátěžovým testům a po každém testu zjišťovat strukturální či metrické změny. Výstupem těchto testů pak může být klasifikace výrobků podle míry poškození či změn.

△

Vidíme tedy, že klasifikace na základě longitudinálních pozorování má široké uplatnění v mnoha odvětvích či oborech. Přitom všechny příklady spojuje klíčová myšlenka - vytvoření odpovídajícího modelu pro odezvu, která nějakým způsobem dostatečně vysvětluje klasifikované skupiny (riziko rakoviny, délka přežití, rizikovost úvěru atd.). Jako vhodný nástroj pro modelování tohoto typu dat se jeví lineární smíšený model, který bude představen v následující kapitole.

1.2 Základní informace

Práce si klade za hlavní cíl seznámit čtenáře s metodami diskriminační analýzy, které jsou vhodné ke klasifikaci longitudinálních dat, a to jak v teoretické, tak i praktické oblasti.

Text je rozdělen do čtyř hlavních kapitol. V první kapitole je představen lineární smíšený model a jeho základní vlastnosti včetně základních metod, které slouží k odhadu jeho parametrů. Lineární smíšený model je vhodný nástroj k popisu longitudinálních dat, u kterých se vyskytují opakovaná pozorování.

Ve druhé kapitole jsou nejprve připomenuty základy klasické diskriminační analýzy včetně části, která se zabývá teorií ztrátových funkcí. Výsledky týkající se optimálního rozhodovacího pravidla jsou pak použity v další kapitole k samotné klasifikaci. Dále tato kapitola představuje modifikované metody, které jsou založeny na použití odhadnutého normálního lineárního smíšeného modelu a Bayesovy věty. Poslední část se věnuje zobecnění jedné z popsaných metod na spojitý případ (uvažujeme spojitý čas).

Ve třetí kapitole je čtenáři představena reálná aplikace teoretických postupů uvedených v předešlé kapitole. V práci jsou použity data, jejichž originální zdroj je na stránkách *http://lib.stat.cmu.edu/datasets/* a jejich oficiální název je *Mayo Clinic Primary Biliary Cirrhosis, sequential data*. Data též můžeme nalézt v tištěné podobě v knize *Counting processes and survival analysis* autorů Thomase R. Fleminga a Davida P. Harrigtona [Fleming and Harrigton, 2005] v apendixu. Třetí kapitola je zaměřena na modelování hladiny různých látek v závislosti na době od prvního měření s cílem použít nalezené modely k následné klasifikaci subjektů. Jsou zde uvedeny konečné podoby modelů, grafické znázornění dat, hodnoty odhadů a intervaly spolehlivosti pro jednotlivé neznámé parametry. Dále se tato kapitola věnuje dosaženým výsledkům klasifikace, která využívá metody popsané v předešlé kapitole.

V závěrečné, tedy páté kapitole, je navrhnuta a zpracována simulační studie zabývající se základními vlastnostmi navrhnutých metod.

Jako základní literatura pro popis lineárního smíšeného modelu byla použita kniha *Linear Mixed Models for Longitudinal Data* od autorů G. Verbeka a G. Molenberghse. K popisu modifikovaných metod diskriminační analýzy byly použity především články *Comparing approaches for predicting prostate cancer from longitudinal data* autorů C.H. Morella a kol. a také článek *Functional modelling and classification of longitudinal data* autora G. H. Müllera, který uvádí zobecnění metody s rozdělením náhodných efektů na spojitý čas.

Kapitola 2

Lineární smíšený model

Tématem lineárního smíšeného modelu a studiem jeho základních vlastností, se zabývá první část této kapitoly. Ve druhé části se pak zaměříme na odhady parametrů tohoto modelu. Právě lineární smíšený model pak použijeme k popisu longitudinálních pozorování a odhadnutý model následně k diskriminaci. Struktura této kapitoly a značení je převážně převzato z knihy G.Verbeka a G. Molenberghse [Verbeke and Molenberghs, 2000].

2.1 Základní vlastnosti

Uvažujme následující strukturu.
Nechť

$$\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})^{\mathbf{T}}, \ldots, \mathbf{Y}_k = (Y_{k,1}, \ldots, Y_{k,n_k})^{\mathbf{T}}$$

jsou nezávislé náhodné vektory.
Nechť

$$\mathbb{X}_1, \ldots, \mathbb{X}_k$$

jsou matice známých čísel typu $n_i \times p$, kde $p \geq 0$, $i = 1, \ldots, k$ o plné sloupcové hodnosti.
Nechť

$$\mathbb{Z}_1, \ldots, \mathbb{Z}_k$$

jsou matice známých čísel o plné sloupcové hodnosti typu $n_i \times q$, $q \geq 0$, $i = 1, \ldots, k$.
Nechť

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathbf{T}}$$

je vektor neznámých parametrů a nechť

$$\mathbf{b}_1, \ldots, \mathbf{b}_k$$

jsou nezávislé q-složkové náhodné vektory se střední hodnotou $\mathbb{E}(\mathbf{b}_i) = \mathbf{0}$ a (obecně neznámou) varianční maticí $\mathrm{var}(\mathbf{b}_i) = \mathbb{D}$, $i = 1, \ldots, k$, kde $\mathbb{D}$ je symetrická,

pozitivně semidefinitní matice. Na pozadí této struktury vyslovíme následující definici lineárního smíšeného modelu.

Definice 2.1. Řekneme, že $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ se řídí *lineárním smíšeným modelem* se strukturou uvedenou výše, jestliže lze psát

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

kde

$$\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,n_i})$$

jsou pro $i = 1, \ldots, k$ nezávislé náhodné vektory se střední hodnotou $\mathbb{E}(\boldsymbol{\epsilon}_i) = \mathbf{0}$ a (obecně neznámou) varianční maticí $\mathrm{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i$, kde $\boldsymbol{\Sigma}_i$ je symetrická pozitivně semidefinitní matice, nezávislé také s vektory $\mathbf{b}_1, \ldots, \mathbf{b}_k$.

Lineární smíšený model budeme od této chvíle označovat jako **LMM** z anglického *Linear mixed-effects model*. Dále poznamenejme, že matice $\mathbb{Z}_i$ je typicky podmaticí matice $\mathbb{X}_i$.

Dále označme

$$n = \sum_{i=1}^{k} n_i$$

celkový rozsah výběru a

$$\mathbf{Y} = (\mathbf{Y}_1^{\mathbf{T}}, \ldots, \mathbf{Y}_k^{\mathbf{T}})^{\mathbf{T}}$$

vektor všech pozorování. Pak můžeme náš model z Definice 2.1 přepsat maticově pro všechna pozorování najednou následovně:

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{Z}\mathbf{B} + \boldsymbol{\epsilon}, \tag{2.1}$$

kde

$$\mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_k \end{pmatrix}$$

je matice typu $n \times p$,

$$\mathbb{Z} = \begin{pmatrix} \mathbb{Z}_1 & \ldots & \mathbb{O} \\ \ldots & \ddots & \ldots \\ \mathbb{O} & \ldots & \mathbb{Z}_k \end{pmatrix}$$

je matice typu $n \times k * q$,

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_k \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_k \end{pmatrix}.$$

Nyní uvedeme základní vlastnosti **LMM**.

Lemma 2.2. *Pro každé $i = 1, \ldots, k$ má lineární smíšený model popsaný v Definici 2.1 následující vlastnosti:*

1) $\mathbb{E}\mathbf{Y}_i = \mathbb{X}_i\boldsymbol{\beta}$,

2) $\mathrm{var}\mathbf{Y}_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^{\mathbf{T}} + \boldsymbol{\Sigma}_i =: \mathbb{V}_i$,

3) $\mathbb{E}[\mathbf{Y}_i|\mathbf{b}_i] = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i$,

4) $\mathrm{var}[\mathbf{Y}_i|\mathbf{b}_i] = \boldsymbol{\Sigma}_i$.

Důkaz. Všechna tvrzení dostaneme triviálně dosazením a využitím vlastností modelu. □

Poznámka 2.3. *V mnoha případech se za matici $\boldsymbol{\Sigma}_i$ volí matice $\sigma^2\mathbf{I}_{n_i}$, kde $\mathbf{I}_{n_i}$ značí jednotkovou matici dimenze n_i, $i = 1, \ldots, k$, a $\sigma^2 > 0$ je neznámý parametr. Tuto volbu interpretujeme tak, že v podmíněném rozdělení $\mathbf{Y}_i$ za podmínky $\mathbf{b}_i$ jsou složky vektoru $\mathbf{Y}_i$ nekorelované a mají rozptyl σ^2.*

Nyní vyslovme následující definici:

Definice 2.4. Lineární smíšený model, kde $\mathbf{b}_i$ jsou nezávislé stejně rozdělené s rozdělením $\mathcal{N}_q(\mathbf{0}, \mathbb{D})$ a $\boldsymbol{\epsilon}_i$ jsou nezávislé s rozdělením $\mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, k$, nazveme *normálním lineárním smíšeným modelem.*

Poznámka 2.5. *Zvolíme-li matici $\boldsymbol{\Sigma}_i$ stejně jako v Poznámce 2.3, jsou v důsledku normálního rozdělení v podmíněném rozdělení $\mathbf{Y}_i$ za podmínky $\mathbf{b}_i$ složky vektoru $\mathbf{Y}_i$ nezávislé. Mluvíme o tzv.* ***modelu s podmíněnou nezávislostí****.*
Na tomto místě poznamenejme, že nebude-li řečeno jinak, budeme vždy uvažovat normální lineární smíšený model daný Definicí 2.4.

Věta 2.6. *Marginální rozdělení odezvy v modelu z Definice 2.4 je*

$$\mathbf{Y}_i \sim \mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}, \mathbb{V}_i), \tag{2.2}$$

kde $\mathbb{V}_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^{\mathbf{T}} + \boldsymbol{\Sigma}_i$, $i = 1, \ldots, k$.

Důkaz. Důkaz je opět zřejmý. Střední hodnota a rozptyl odezvy $\mathbf{Y}_i$ jsou uvedeny v Lemmatu 2.2.
Další tvrzení uvedené v této větě plynou z předpokladu normálních rozdělení $\mathbf{b}_i$ a $\boldsymbol{\epsilon}_i$ a vlastností normálního rozdělení. □

Popsali jsme tedy (normální) lineární smíšený model, který nám později poslouží jako vhodný nástroj k popisu longitudinálních pozorování. Avšak abychom mohli daný model plně využít, potřebujeme znát nejen samotný tvar modelu, ale i odhad neznámých parametrů. Této problematice se věnuje další část.

2.2 Odhady parametrů v LMM

Jak již bylo řečeno výše, v této části se zaměříme na metody odhadu neznámých parametrů vyskytujících se v normálním lineárním smíšeném modelu. Uvažujme model daný rovnicí (2.1) a Definicí 2.4.

Parametry tohoto modelu jsou pevné efekty $\boldsymbol{\beta}$ a složky matic $\mathbb{D}$ a $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, k$. Pro přehlednost zaveďme následující značení:

- $\boldsymbol{\alpha}$: všechny parametry varianční struktury (všechny parametry matic $\mathbb{D}$ a $\boldsymbol{\Sigma}_i$),
- $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathbf{T}}, \boldsymbol{\alpha}^{\mathbf{T}})^{\mathbf{T}}$: s-dimenzionální vektor všech parametrů modelu,
- $\boldsymbol{\Theta}_{\boldsymbol{\beta}} = \mathbb{R}^p$: parametrický prostor pro pevné efekty,
- $\boldsymbol{\Theta}_{\boldsymbol{\alpha}}$: parametrický prostor pro parametry varianční struktury,
- $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\boldsymbol{\beta}} \times \boldsymbol{\Theta}_{\boldsymbol{\alpha}}$: celkový parametrický prostor.

2.2.1 Odhad metodou maximální věrohodnosti

Odhad neznámých parametrů metodou maximální věrohodnosti patří mezi klasické a často používané metody. Postup je takový, že se snažíme maximalizovat věrohodnostní funkci $\mathbf{L}_{ML}(\boldsymbol{\theta})$ v proměnné $\boldsymbol{\theta}$ na parametrickém prostoru $\boldsymbol{\Theta}$. Věrohodnostní funkce (plynoucí z marginálního rozdělení odezvy, viz 2.2) je v tomto případě ve tvaru

$$\mathbf{L}_{ML}(\boldsymbol{\theta}) = \prod_{i=1}^{k} \left\{ (2\pi)^{-\frac{n_i}{2}} |\mathbb{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \right. \\ \left. \times \exp\left(-\frac{1}{2} (\mathbf{Y}_i - \mathbb{X}_i\boldsymbol{\beta})^{\mathbf{T}} \mathbb{V}_i^{-1}(\boldsymbol{\alpha}) (\mathbf{Y}_i - \mathbb{X}_i\boldsymbol{\beta}) \right) \right\}. \tag{2.3}$$

Pokud budeme předpokládat, že je parametr $\boldsymbol{\alpha}$ znám, maximalizujeme funkci danou rovnicí (2.3) přes $\boldsymbol{\beta}$ a získáme analytické vyjádření pro výsledný odhad ve tvaru

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}} \mathbb{W}_i \mathbb{X}_i \right)^{-1} \sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}} \mathbb{W}_i \mathbf{Y}_i, \tag{2.4}$$

kde $\mathbb{W}_i = \mathbb{V}_i^{-1}$. Odhad dostaneme velice snadno z (2.3), kterou zlogaritmujeme, zderivujeme a položíme pravou stranu rovnu nule.

Ve většině případů jsou však parametry vyskytující se v $\boldsymbol{\alpha}$ neznámé. V takovém případě můžeme postupovat následujícím způsobem. Zvolíme si nějakou inicializační hodnotu $\boldsymbol{\beta}^0$. K tomu dopočteme příslušné $\boldsymbol{\alpha}^0$ maximalizací $L_{\mathrm{ML}}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0)$, které použijeme na odhad matice $\mathbb{W}_i(\boldsymbol{\alpha})$ a odhadneme $\boldsymbol{\beta}^1$ pomocí výsledku z rovnice (2.4) atd. Maximalizace vzhledem k $\boldsymbol{\alpha}$ při známém $\boldsymbol{\beta}$ se většinou řeší numericky. V současné době se využívá tzv. **Newtonova-Raphsonova algoritmu** či **Fisherova skórování**.

2.2.2 Odhad pomocí REML

Maximálně věrohodný odhad pomocí kontrastů (*Restricted Maximum Likelihood Estimation*-**REML**) je alternativou ke standardnímu maximálně věrohodnému odhadu.

Uvažujme opět model daný rovnicí (2.1), tj.

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{Z}\mathbf{B} + \boldsymbol{\epsilon}.$$

Dále nechť $\mathbb{A}$ je matice typu $n \times (n-p)$, která má lineárně nezávislé sloupce a jejíž sloupce jsou ortogonální na sloupce matice $\mathbb{X}$. Jinými slovy, $\mathbb{A}^{\mathbf{T}}\mathbb{X} = \mathbb{O}_{(n-p)\times p}$. Dále označme

$$\mathbf{M} = \mathbb{A}^{\mathbf{T}}\mathbf{Y}.$$

Marginální model pro vektor $\mathbf{M}$ je tedy roven

$$\begin{aligned}\mathbf{M} &\sim \mathcal{N}(\mathbb{A}^{\mathbf{T}}\mathbb{X}\boldsymbol{\beta}, \mathbb{A}^{\mathbf{T}}\mathbb{V}(\boldsymbol{\alpha})\mathbb{A}) \\ &\sim \mathcal{N}(\mathbb{O}, \mathbb{A}^{\mathbf{T}}\mathbb{V}(\boldsymbol{\alpha})\mathbb{A}).\end{aligned}$$

Vidíme, že marginální model pro $\mathbf{M}$ nezávisí na $\boldsymbol{\beta}$. Stejně jako v případě klasické maximální věrohodnosti nás bude nyní zajímat věrohodnost pro náhodný vektor $\mathbf{M}$. Harville [Harville, 1974] ukázal, že věrohodnost pro vektor $\mathbf{M}$ je rovna

$$\begin{aligned}\mathbf{L}^{\star}(\boldsymbol{\alpha}) = (2\pi)^{-\frac{n-p}{2}} &\left|\sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}}\mathbb{X}_i\right|^{\frac{1}{2}} \\ &\times \left|\sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}}\mathbb{V}_i^{-1}\mathbb{X}_i\right|^{-\frac{1}{2}} \prod_{i=1}^{k} |\mathbb{V}_i|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{Y}_i - \mathbb{X}_i\hat{\boldsymbol{\beta}})^{\mathbf{T}}\mathbb{V}_i^{-1}(\alpha)(\mathbf{Y}_i - \mathbb{X}_i\hat{\boldsymbol{\beta}})\right\}, \qquad (2.5)\end{aligned}$$

kde tvar $\hat{\boldsymbol{\beta}}$ je uveden v (2.4). Poznamenejme, že právě uvedená věrohodnost nezávisí na volbě matice kontrastů $\mathbb{A}$. Věrohodnost $\mathbf{L}^{\star}(\boldsymbol{\alpha})$ lze pak s využitím (2.3) přepsat na tvar

$$\begin{aligned}\mathbf{L}^{\star}(\boldsymbol{\alpha}) &= (2\pi)^{\frac{p}{2}} \left|\sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}}\mathbb{X}_i\right|^{\frac{1}{2}} \left|\sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}}\mathbb{V}_i^{-1}(\boldsymbol{\alpha})\mathbb{X}_i\right|^{-\frac{1}{2}} \mathbf{L}_{LM}(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) \\ &= \text{konst.} \cdot \mathbf{L}_{REML}(\boldsymbol{\alpha}). \qquad (2.6)\end{aligned}$$

Z výrazu (2.6) tedy dostáváme:

$$\mathbf{L}_{REML}(\boldsymbol{\alpha}) = \left|\sum_{i=1}^{k} \mathbb{X}_i^{\mathbf{T}}\mathbb{V}_i^{-1}(\boldsymbol{\alpha})\mathbb{X}_i\right|^{-\frac{1}{2}} \mathbf{L}_{LM}(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}), \boldsymbol{\alpha}), \qquad (2.7)$$

kde $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ je uvedeno v (2.4). Poznamenejme, že maximalizace se obvykle opět řeší numericky pomocí algoritmů uvedených u klasického **ML** odhadu.

Oba výše popsané přístupy jsou založeny na věrohodnostním principu a získané odhady mají za určitých podmínek vlastnosti maximálně věrohodných odhadů (konzistence, asymptotická normalita). V souvislosti s asymptotikou poznamenejme, že většinou uvažujeme, že počet subjektů jde do nekonečna, tedy $k \to \infty$.

Z předešlého textu není zcela jasné, zda můžeme nějak využít několika různých měření na jednom subjektu (pacientovi). Pro názornou představu si připomeňme Příklad 1.3, kde sledujeme u i-tého pacienta hladinu bilirubinu, albuminu a počet krevních destiček v určitém objemu krve. Pokud máme ke každému subjektu k dispozici více různých měření, nemusíme se omezovat pouze na separátní modely pro hladinu bilirubinu, albuminu a počet krevních destiček, ale můžeme sestavit společný model pro všechny sledované veličiny najednou a v důsledku toho zlepšit výsledky následné klasifikace. Společný model pro sledované veličiny má pak konkrétně tvar, který můžeme opět pro i-tý subjekt maticově vyjádřit ve tvaru

$$\mathbf{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, k, \tag{2.8}$$

kde

$$\mathbf{Y}_i = \left(\mathbf{Y}_i^{1.\text{měř. vel.}}, \ldots, \mathbf{Y}_i^{s.\text{měř. vel.}}\right),$$

je vektor dimenze $\sum_{h=1}^{s} n_i^h$, kde s značí počet měřených veličin (v našem příkladě by bylo $s = 3$, neboť máme tři sledované veličiny - bilirubin, albumin a počet krevních destiček). Analogicky dále platí, že $\mathbb{X}_i$ je matice známých čísel tvaru

$$\mathbb{X}_i = \begin{pmatrix} \mathbb{X}_i^{1.\text{měř. vel.}} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbb{X}_i^{2.\text{měř. vel.}} & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbb{X}_i^{s.\text{měř. vel.}} \end{pmatrix},$$

kde $\mathbb{X}_i^{h.\text{měř. vel.}}$ je matice známých čísel dimenze $(n_i^h \times p^h)$. Nulové matice $\mathbf{0}$ jsou obecně různé dimenze v závislosti na dimenzích matic $\mathbb{X}_i^{h-t\text{měř. vel.}}$. Tedy matice $\mathbb{X}_i$ je blokově diagonální matice, kde na diagonále jsou matice známých čísel, které přísluší h-té měřené veličině, $h = 1, \ldots, s$. Matice $\mathbb{Z}_i$ má podobnou strukturu jako matice $\mathbb{X}_i$. Opět se jedná o diagonálně blokovou matici dimenze $(n_i^h \times q^h)$. A analogicky sestavíme i vektory pevných a náhodných efektů pro společný model.

V případě společného modelu, který popisuje více sledovaných veličin najednou, je ale zcela nepřirozené předpokládat, že varianční matice $\boldsymbol{\Sigma}_i$ má tvar $\sigma^2\mathbf{I}_{(\sum_{h=1}^{s} n_i^h)}$. Můžeme však předpokládat, že varianční struktura je u h-té sledované veličiny a i-tého subjektu rovna $(\sigma^h)^2\mathbf{I}_{n_i^h}$. Tedy výsledná varianční matice $\boldsymbol{\Sigma}_i$ je blokově diagonální s maticemi $(\sigma^h)^2\mathbf{I}_{n_i^h}$ na diagonále. Problém nastává u varianční matice náhodných efektů, kterou jsme označili $\mathbb{D}$. Ta totiž obecně není blokově diagonální (tato informace se musí vzít v potaz hlavně u praktické aplikace společného

modelu).

Vlastnosti modelu, který najednou popisuje několik sledovaných veličin, případně metody odhadu, jsou pak totožné s těmi, které jsme uvedli výše. Praktická ukázka společného modelu bude uvedena v Kapitole 4, kde se i dozvíme, zda takto složený model přináší zlepšení klasifikace.

Představili jsme tedy (normální) lineární smíšený model a jeho základní vlastnosti, včetně dvou metod na odhadování neznámých parametrů. Normální lineární smíšený model v dalších kapitolách použijeme jako nástroj pro modelování longitudinálního typu dat a jeho odhad poslouží jako základ modifikovaných metod diskriminační analýzy.

Kapitola 3

Diskriminační analýza pro longitudinální data

V následující kapitole si představíme několik modifikací klasické diskriminační analýzy, které můžeme využít ke klasifikaci pozorování longitudinálního typu. Modifikované metody diskriminační analýzy, které budou uvedeny v této kapitole jsou částečně převzaty z výsledků Morrella a kol. [Morrell and Sheng, 2007] a Branta a kol. [Brant et al., 2003]. V poslední části této kapitoly je představeno zobecnění metody, která k diskriminaci využívá rozdělení náhodných efektů získaných z odhadnutého lineárního smíšeného modelu, na spojitý čas. Zobecnění je převzato z článku H. G. Müllera [Müller, 2005]. Nejprve si ale připomeňme klasickou diskriminační analýzu.

3.1 Klasická diskriminační analýza

Za zakladatele klasické diskriminační analýzy je považován sir **Ronald Aylmer Fisher**, který ve své práci z roku 1936 použil metodu dnes označovanou jako *Fisherova diskriminační analýza*, ke klasifikaci druhového jména rostlin kosatců na základě čtyř pozorovaných znaků. Metodu demostroval na datech známých jako *Iris flower data set*.

Z počátku byla diskriminační analýza využívána v biologii k třídění rostlin a zemědělských plodin, dále také v medicíně a v antropologii při klasifikaci koster. S nástupem a rozmachem výpočetní techniky, která umožnila rychle zpracovávat velké datové soubory, se diskriminační analýza dočkala širšího uplatnění například v sociologii, politice nebo bankovnictví.

Zajímavé aplikace klasické diskriminační analýzy lze najít v nejrůznějších oblastech, v biologii, medicíně, archeologii či technických oborech. Uveďme jako ukázku několik příkladů.

Příklad 3.1. *Při kontrole jakosti či spolehlivosti můžeme u výrobků zkoumat zátěžové vlastnosti. Nejprve změříme nějaké veličiny (hmotnost, rozměry, hustotu, chemické složení atd.). Poté je podrobíme nějaké zátěži a budeme sledovat, zda se nějak poškodí, či nikoliv. K předpovědi chování dalších výrobků při zátěži nám*

pak stačí provést potřebná kvantitativní měření a dle výsledků výrobek zařadit do jedné ze skupin.

△

Příklad 3.2. *Banka sleduje u svých stávajících klientů způsob splácení poskytnutého úvěru a další ukazatele (věk, pohlaví, výše příjmů,...). Na základě tohoto zjišťování pak může vyhodnocovat potencionální nové žadatele o úvěr jako důvěryhodné nebo méně důvěryhodné atd.*

△

Příklad 3.3. *Lékař ve výběrovém souboru pacientů provádí různé testy. Na základě těchto testů pak může u nově příchozích pacientů diagnostikovat onemocnění, případně fázi průběhu onemocnění.*

△

Klasická diskriminační analýza slouží k zařazení určitého subjektu do jedné z g disjunktních skupin, $g = 1, \ldots, G$, dle určitých charakteristik (věk, pohlaví,...). Tradiční lineární diskriminační analýza předpokládá:

1) Náhodné vektory $\mathbf{Y}_i^g$ reprezentující měření na jednotlivých subjektech patřící do stejné skupiny g, $g = 1, \ldots, G$, jsou nezávislé, stejně rozdělené, $i = 1, \ldots, k^g$.

2) Obvykle navíc předpokládáme, že $\mathbf{Y}_i^g$ mají vícerozměrné normální rozdělení

$$\mathbf{Y}_i^g \sim \mathcal{N}_p(\boldsymbol{\mu}^g, \boldsymbol{\Sigma}^g),$$

kde n_i je dimenze vektoru $\mathbf{Y}_i^g$, $i = 1, \ldots, k^g$.

Je tedy zřejmé, že jednotlivé pozorované vektory $\mathbf{Y}_i^g$ musí být v rámci jednotlivých skupin stejné dimenze a že jednotlivé náhodné vektory reprezentující subjekty g-té skupiny musí mít stejnou kovarianční strukturu.

K zařazování určitých subjektů (v našem případě pacientů) můžeme přistoupit následovně:

1) Vytvoříme pravidlo pro zařazení i-tého subjektu do tříd na základě pozorování $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p}) \in \mathbb{R}^p$.

2) Nové subjekty zařazujeme do tříd dle pravidla vytvořeného v bodě 1).

Poznámka 3.4. *K vytvoření rozhodovacích pravidel máme k dispozici tzv. „trénovací množinu“ subjektů, u nichž známe jak příslušnost do jedné z G tříd, tak i vektor pozorování $\mathbf{y}_i$. Můžeme rozlišit dva typy rozhodovacích pravidel.*

Deterministická pravidla fungují tak, že pokud na objektu i pozorujeme znaky $\mathbf{y}_i$ a $\mathbf{y}_i \in W^g$, kde $\mathbb{R}^p = \bigcup_g W^g$, $g = 1, \ldots, G$, je disjunktní rozklad na G tříd, pak subjekt i zařadíme do skupiny g.

Jinou možností je znáhodněné (randomizované) rozhodovací pravidlo. Subjekt $\mathbf{y}_i$ zařadíme do skupiny g s pravděpodobností p^g, kde $\Gamma(\mathbf{y}_i) = (p^1(\mathbf{y}_i), \ldots, p^G(\mathbf{y}_i)) : \mathbb{R}^p \to [0,1]^G$, je vektorová funkce splňující pro každé $\mathbf{y}_i$:

1) $p^g(\mathbf{y}_i) \geq 0$,

2) $\sum_{g=1}^{G} p^g(\mathbf{y}_i) = 1$.

Jednoduše nahlédneme, že deterministické rozhodovací pravidlo je speciální případ znáhodněného rozhodovacího pravidla.

Naším dalším cílem je nalézt takové rozhodovací pravidlo, které je za určitých předpokladů optimální. K tomu bude zapotřebí zavést základy teorie ztrátových funkcí.

Označme hustotu náhodného vektoru, který reprezentuje měření na i-tém subjektu, který náleží do skupiny g symbolem f^g (hustota je myšlena vůči nějaké σ-konečné míře ν). V závislosti na rozhodovacím pravidlu zařadíme i-tý subjekt do skupiny g. Toto rozhodnutí je spojeno s rizikem, že subjekt, který patří do skupiny g, chybně zařadíme do skupiny l, kde $l, g = 1, \ldots, G$. Toto chybné rozhodnutí ohodnotíme ztrátou r^{gl}.

Jestliže i-tý subjekt patří do skupiny g, je střední ztráta rovna výrazu

$$\begin{aligned} L^g &= \mathrm{E} r^{gl} = r^{g1}\mathrm{P}(\text{zařazení do skupiny 1}) + \\ &+ \ldots + r^{gG}\mathrm{P}(\text{zařazení do skupiny G}) = \\ &= r^{g1} \int_{W^1} f^g(\mathbf{y}_i)\mathrm{d}\nu(\mathbf{y}) + \\ &+ \ldots + r^{gG} \int_{W^G} f^g(\mathbf{y}_i)\mathrm{d}\nu(\mathbf{y}) = \\ &= \sum_{l=1}^{G} r^{gl} \int_{W^l} f^g(\mathbf{y}_i)\mathrm{d}\nu(\mathbf{y}). \end{aligned} \tag{3.1}$$

Dále předpokládejme, že $\pi^1, \ldots, \pi^G$ jsou dané apriorní pravděpodobnosti, kde π^g značí apriorní pravděpodobnost příslušnosti do skupiny g. Můžeme tedy říci, že objekt i je vybrán ze směsi populací s vahami $\pi^1, \ldots, \pi^G$ s celkovou střední ztrátou

$$L = \sum_{g=1}^{G} \pi^g L^g = \sum_{g=1}^{G} \pi^g \sum_{l=1}^{G} r^{gl} \cdot \int_{W^l} f^g(\mathbf{y}_i)\mathrm{d}\nu(\mathbf{y}). \tag{3.2}$$

Nyní můžeme definovat g-tý diskriminační skór následujícím způsobem.

Definice 3.5. Nechť $\mathbb{R}^p = \bigcup_g W^g$, $g = 1, \ldots, G$, je disjunktní rozklad na G tříd. Nechť $\pi^1, \ldots, \pi^G$ jsou dané apriorní pravděpodobnosti a nechť u i-tého subjektu pozorujeme charakteristiky $\mathbf{y}_i \in \mathbb{R}^p$, $i = 1, \ldots, k$. Pak *g-tým diskriminačním skórem* rozumíme výraz $S^g(\mathbf{y}_i)$, kde

$$S^g(\mathbf{y}_i) = -\sum_{l=1}^{G} \pi^l r^{lg} f^l(\mathbf{y}_i), \quad g = 1, \ldots, G.$$

Právě definovaný g-tý diskriminační skór můžeme dosadit do výrazu (3.2) a dostaneme tak vyjádření pro celkovou střední ztrátu

$$L = -\sum_{l=1}^{G} \int_{W^l} S^l(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}).$$

Nyní vyslovme důležitou větu, která nám později poskytne návod na požadované optimální rozhodovací pravidlo.

Věta 3.6. *Nechť $\mathbb{R}^p = \bigcup_g W_\star^g$ je disjunktní rozklad $\mathbb{R}^p$ takový, že platí implikace*

$$\mathbf{y}_i \in W_\star^g \Rightarrow S^g(\mathbf{y}_i) \geq S^l(\mathbf{y}_i), \quad \textit{pro každé } l = 1, \ldots, G.$$

Potom platí

$$L_\star = -\sum_{l=1}^{G} \int_{W_\star^l} S^l(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}) \leq L = -\sum_{l=1}^{G} \int_{W^l} S^l(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}) \tag{3.3}$$

pro každý disjunktní rozklad $W^1, \ldots, W^G$ prostoru $\mathbb{R}^p$.

Důkaz. Protože platí

$$W^l = W^l \cap \mathbb{R}^p = \cup_{g=1}^{G} (W^l \cap W^g),$$

dostáváme postupně s použitím předpokladu

$$\begin{aligned} L &= -\sum_{l=1}^{G} \int_{W^l} S^l(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}) \\ &= -\sum_{l=1}^{G} \sum_{g=1}^{G} \int_{W_\star^g \cap W^l} S^l(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}) \geq \\ &\geq -\sum_{l=1}^{G} \sum_{g=1}^{G} \int_{W_\star^g \cap W^l} S^g(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}) = \\ &= -\sum_{g=1}^{G} \int_{W_\star^g} S^g(\mathbf{y}_i) \mathrm{d}\nu(\mathbf{y}) = L_\star. \end{aligned}$$

□

Tedy dle Věty 3.6 je rozhodovací pravidlo založené na $W_\star^1, \ldots, W_\star^G$ optimální.

Poznámka 3.7. *Poznamenejme, že optimální rozklad ve Větě 3.6 není dán jednoznačně. Nejčastěji se v praxi za ztrátovou funkci volí $r^{gl} = 1$, pokud $g \neq l$ a $r^{gg} = 0$. Tedy L^g pak můžeme interpretovat jako střední hodnotu podílu špatně zařazených subjektů třídy g a L jako střední hodnotu všech špatně zařazených subjektů. Dále*

$$\begin{aligned} S^l(\mathbf{y}_i) &= -\sum_{g \neq l} \pi^g f^g(\mathbf{y}_i) = \pi^l f^l(\mathbf{y}_i) - \sum_{g=1}^{G} \pi^g f^g(\mathbf{y}_i) = \\ &= c + \pi^l f^g(\mathbf{y}_i), \qquad l = 1, \ldots, G. \end{aligned} \tag{3.4}$$

Tedy subjekt i zařadíme na základě napozorovaných charakteristik $\mathbf{y}_i$ do třídy g, pro kterou je

$$g = \arg\max_l \{\pi^l f^l(\mathbf{y}_i)\}.$$

Pro názornou ilustraci, kterou využijeme i v dalším pokračování práce předpokládejme, že f^g je hustota p-rozměrného normálního rozdělení se střední hodnotou $\boldsymbol{\mu}^g$ a varianční maticí $\boldsymbol{\Sigma}^g$ vzhledem k nějaké σ-konečné míře ν. Výše jsme uvedli, že subjekt i zařadíme do takové skupiny g, pro kterou platí

$$\pi^g f^g(\mathbf{y}_i) \geq \pi^l f^l(\mathbf{y}_i), \text{ pro každé } l = 1, \ldots, G. \tag{3.5}$$

Pravidlo (3.5) je zřejmě ekvivalentní zápisu

$$\log \pi^g + \log f^g(\mathbf{y}_i) \geq \log \pi^l + \log f^l(\mathbf{y}_i), \text{ pro každé } l = 1, \ldots, G. \tag{3.6}$$

Pak za předpokladu hustoty p-rozměrného normálního rozdělení, která je v závislosti na skupině g rovna

$$f^g(\mathbf{y}_i) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}^g|}} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}^g)^{\mathbf{T}} (\boldsymbol{\Sigma}^g)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}^g)\right],$$

můžeme g-tý diskriminační skór po zlogaritmování zapsat ve tvaru

$$\begin{aligned} S^g(\mathbf{y}_i) &= \log \pi^g + \log f^g(\mathbf{y}_i) = \\ &= \log \pi^g - \frac{1}{2}\log |\boldsymbol{\Sigma}^g| - \frac{p}{2}\log 2\pi - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}^g)^{\mathbf{T}} (\boldsymbol{\Sigma}^g)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}^g) \end{aligned} \tag{3.7}$$

a rozhodovací pravidlo k zařazení i-tého subjektu do g-té skupiny můžeme uvést ve tvaru

$$g = \arg\max_l S^l(\mathbf{y}_i).$$

Ve speciálním případě, kdy budeme předpokládat, že $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^1, \ldots, \boldsymbol{\Sigma}^G$, tedy že kovarianční struktura je ve všech skupinách stejná, je zřejmé, že člen $-\frac{1}{2}\log \boldsymbol{\Sigma} - \frac{1}{2}\mathbf{y}_i^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{y}_i$ je konstantní pro všechna $g = 1, \ldots, G$. Tedy pro rozhodování o zařazení i-tého jedince použijeme (a tedy i definujeme) g-tý diskriminační skór předpisem

$$S^g(\mathbf{y}_i) = \mathbf{y}_i^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^g - \frac{1}{2}\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^g + \log \pi^g.$$

Speciálně, pro případ $G = 2$, který nás v této práci nejvíce zajímá, se budeme rozhodovat o zařazení i-tého jedince podle pravidla

$$S^1(\mathbf{y}_i) - S^2(\mathbf{y}_i) =: L(\mathbf{y}_i) - c, \tag{3.8}$$

kde

$$L(\mathbf{y}_i) = (\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{y}_i$$

a

$$c = \frac{1}{2}\left((\boldsymbol{\mu}^1)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^1 - (\boldsymbol{\mu}^2)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^2\right) + \log\pi^2 - \log\pi^1.$$

Tedy optimální rozhodovací pravidlo můžeme interpretovat následujícím způsobem:

- Pokud $L(\mathbf{y}_i) \geq c$, pak zařadíme subjekt i na základě $\mathbf{y}_i$ do skupiny 1.
- Pokud $L(\mathbf{y}_i) < c$, pak zařadíme subjekt i na základě $\mathbf{y}_i$ do skupiny 2.

Jelikož jsme ukázali, že toto pravidlo je za uvedených předpokladů optimální, bude použito ke klasifikaci v následujících kapitolách. Pro zajímavost můžeme poznamenat, že funkce $L(\mathbf{y}_i)$ uvedená v (3.8) se většinou označuje jako *Fisherova diskriminační funkce*.

Pro přehlednost uveďme další možnosti kritérií ke klasifikaci subjektů do jedné z G disjunktních skupin. Poznamenejme, že u těchto kritérií se již neuvažuje ztrátová funkce.

- **Mahalanobisova vzdálenost**: na základě $\mathbf{y}_i$ zařadíme i-tý subjekt do skupiny $g = 1, \ldots, G$, kde

$$g = \arg\min_l \left\{(\mathbf{y}_i - \boldsymbol{\mu}^l)^{\mathbf{T}}(\boldsymbol{\Sigma}^l)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^l)\right\}.$$

- **Princip maximální věrohodnosti**: i-tý subjekt zařadíme do g-té skupiny, kde

$$g = \arg\max_l f^l(\mathbf{y}_i).$$

- **Bayesovský přístup**: i-tý subjekt zařadíme do jedné z G skupin podle klíče

$$g = \arg\max_l \pi^l f^l(\mathbf{y}_i).$$

V případě zájmu čtenáře o další informace se o klasické diskriminační analýze můžeme více dozvědět například v knize *Applied discriminant analysis*, jejímž autorem je Carl J. Huberty [Huberty, 1994], nebo v českém jazyce v knize Petra Hebáka a kol. *Vícerozměrné statistické metody (1)* [Hebák et al., 2004].

V další části této kapitoly uvedeme problémy, se kterými se potýkáme u klasifikace pozorování longitudinálního typu. Následně se zaměříme na modifikaci metod klasické diskriminační analýzy na longitudinální data.

3.2 Modifikace lineární diskriminační analýzy na longitudinální data

Jak již bylo naznačeno v předešlém textu, směřujeme postupně k modifikacím klasické lineární diskriminační analýzy, které se dají aplikovat na pozorování longitudinálního typu. Hlavním důvodem, proč nelze použít klasický přístup k lineární diskriminační analýze, je ten, že naše pozorování $\mathbf{Y}_i^g$ nejsou nutně v rámci skupiny stejně rozdělena, a to ze dvou důvodů:

- Dimenze vektoru pozorování $\mathbf{Y}_i^g$ závisí na i-tém subjektu (tj. různí subjekti mohou přinášet různý počet měření).
- Jednotlivá měření různých subjektů probíhají obecně v různých časech a doby mezi jednotlivými měřeními jsou také různé.

Pro popis vektoru pozorování $\mathbf{Y}_i^g$ v g-té skupině, $g = 1, \ldots, G$, uvažujme normální lineární smíšený model popsaný v Definici 2.4 ve tvaru

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

kde $g = 1, \ldots, G$ značí příslušnost ke g-té klasifikační skupině. A v této souvislosti předpokládejme, že $\boldsymbol{\epsilon}_i^g$ mají rozdělení $\mathcal{N}_{n_i}(\mathbf{0}, (\sigma^g)^2 \mathbf{I}_{n_i})$ a jsou nezávislé s $\mathbf{b}_i^g$, které mají rozdělení $\mathcal{N}_q(\mathbf{0}, \mathbb{D}^g)$.

Poznámka 3.8. *Předpoklad, že varianční matice $\boldsymbol{\epsilon}_i^g$ je rovna $(\sigma^g)^2 \mathbf{I}_{n_i}$, není nutný. Dokonce je v některých případech velice nevhodný a nepřirozený, jak si ukážeme v poslední části následující kapitoly a jak již bylo naznačeno na konci minulé kapitoly.*

Subjekt (pacient) $\mathbf{Y}_i^g$ lze nyní v rámci normálního lineárního smíšeného modelu reprezentovat několika náhodnými vektory a jimi souvisejícími rozděleními. Zároveň hledáme i nějakou míru, která by dostatečně určovala, do které skupiny i-tý subjekt náleží. Mezi možné kandidáty patří

1) marginální rozdělení odezvy $\mathbf{Y}_i^g$,
2) podmíněné rozdělení odezvy $\mathbf{Y}_i^g$ za podmínky náhodných efektů $\mathbf{b}_i^g$,
3) rozdělení náhodných efektů $\mathbf{b}_i^g$.

Dále uvažujme, že $\pi^1, \ldots, \pi^G$ jsou dané apriorní pravděpodobnosti, $\sum_{g=1}^G \pi^g = 1$ a $0 < \pi^g < 1$, kde π^g značí apriorní pravděpodobnost příslušnosti ke g-té skupině, $g = 1, \ldots, G$.

Aposteriorní pravděpodobnost p_i^g, tj. aposteriorní pravděpodobnost i-tého jedince, že náleží do g-té skupiny, vypočteme pomocí klasické Bayesovy věty. Tedy

$$p_i^g = \frac{\pi^g f_i^g(\mathbf{t}_i)}{\sum_{g=1}^G \pi^g f_i^g(\mathbf{t}_i)}, \tag{3.9}$$

kde $f_i^g(\mathbf{t}_i)$ je hustota (vzhledem k nějaké σ-konečné míře ν) daná jedním ze tří přístupů, které jsme uvedli výše a které budeme nadále zkoumat a porovnávat.

Poznamenejme, že všechny neznámé parametry vyskytující se v (3.9), se při praktickém výpočtu nahrazují odhady, které získáme jednou z metod popsaných v Kapitole 2.

Při výpočtu aposteriorních pravděpodobností můžeme postupovat dvěma způsoby. Pokud máme k dispozici již všechna pozorování daného subjektu, který chceme zařadit do jedné z G skupin, použijeme všechna tato pozorování najednou. Pokud ale např. pacient dochází na jednotlivá měření a my chceme po každém tomto

měření klasifikovat pacienta dle určitého pravidla, můžeme postupovat sekvenčně.

V dalších částech této kapitoly si trochu více přiblížíme možné přístupy ke klasifikaci na základě longitudinálních pozorování. Naším cílem je vlastně nalézt nějakou míru podobnosti v rámci jednotlivých skupin.

3.2.1 Přístup s marginálním rozdělením odezvy

V této části se podíváme na první z možností volby hustoty v (3.9). Věta 2.6 uvádí, že marginální rozdělení odezvy $\mathbf{Y}_i^g$ je v závislosti na skupině $g = 1, \ldots, G$, vícerozměrné normální se střední hodnotou $\mathbb{X}_i^g \boldsymbol{\beta}^g$ a varianční maticí $\mathbb{V}_i^g$, kde $\mathbb{V}_i^g = \mathbb{Z}_i^g \mathbb{D}^g (\mathbb{Z}_i^g)^{\mathbf{T}} + (\sigma^g)^2 \mathbf{I}_{n_i}$. Tedy

$$\mathbf{Y}_i^g \sim \mathcal{N}_{n_i}(\mathbb{X}_i^g \boldsymbol{\beta}^g, \mathbb{V}_i^g), \qquad i = 1, \ldots, k^g.$$

Podívejme se na tuto možnost trochu blíže. Hustota $\mathbf{Y}_i^g$ má konkrétně pro i-tý subjekt tvar

$$f^g(\mathbf{y}_i) = (2\pi))^{-\frac{n_i}{2}} |\mathbb{V}_i^g|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbb{X}_i^g \boldsymbol{\beta}^g)^{\mathbf{T}} (\mathbb{V}_i^g)^{-1} (\mathbf{y}_i - \mathbb{X}_i^g \boldsymbol{\beta}^g)\right]. \tag{3.10}$$

V předchozím textu výše (Část 3.1) jsme ukázali, že za určitých podmínek je optimální rozhodovací pravidlo pro zařazení i-tého subjektu do jedné z G skupin maximum z G diskriminačních skórů. Tedy v našem případě zařadíme i-tý subjekt do skupiny g, kde je

$$g = \arg\max_l \left\{ \frac{\pi^l f_i^l(\mathbf{y}_i)}{\sum_{j=1}^G \pi^j f_i^j(\mathbf{y}_i)} \right\}. \tag{3.11}$$

Uvědomíme-li si, že pro i-tý subjekt, který chceme klasifikovat do jedné z G skupin, je jmenovatel v (3.11) konstantní pro všechna $g = 1, \ldots, G$, dostaneme (po zlogaritmování) diskriminační pravidlo ve tvaru

$$\log \pi^g - \frac{1}{2} \log |\mathbb{V}_i^g| - \frac{1}{2} (\mathbf{y}_i - \mathbb{X}_i^g \boldsymbol{\beta}^g)^{\mathbf{T}} (\mathbb{V}_i^g)^{-1} (\mathbf{y}_i - \mathbb{X}_i^g \boldsymbol{\beta}^g), \tag{3.12}$$

což je analogický výsledek, který jsme uvedli v (3.7). Výše popsaný přístup vlastně zkoumá, jak moc je vývoj odezvy i-tého subjektu, který chceme klasifikovat, podobný střednímu vývoji odezvy v g-té skupině, $g = 1, \ldots, G$.

3.2.2 Přístup s podmíněným rozdělením odezvy

Další možností volby hustoty v (3.9) je podmíněná hustota odezvy za podmínky, že známe hodnoty náhodných efektů. V Lemmatu (2.2) je uvedeno, že podmíněné rozdělení

$$\mathbf{Y}_i^g | \mathbf{b}_i^g$$

je vícerozměrné normální dimenze n_i se střední hodnotou

$$\mathbb{E}\left[\mathbf{Y}_i^g | \mathbf{b}_i^g\right] = \mathbb{X}_i^g \boldsymbol{\beta}_i^g + \mathbb{Z}_i^g \mathbf{b}_i^g$$

a varianční maticí $(\sigma^g)^2\mathbf{I}_{n_i}$. Tedy

$$\mathbf{Y}_i^g|\mathbf{b}_i^g \sim \mathcal{N}_{n_i}(\mathbb{X}_i^g\boldsymbol{\beta}_i^g + \mathbb{Z}_i^g\mathbf{b}_i^g, (\sigma^g)^2\mathbf{I}_{n_i}).$$

Konkrétní tvar hustoty (vzhledem k Lebesgueově míře ν) je pro i-tý subjekt roven

$$f^g(\mathbf{y}_i|\mathbf{b}_i^g) = (2\pi))^{-\frac{n_i}{2}} \left|(\sigma^g)^2\mathbb{I}_{n_i}\right|^{-\frac{1}{2}} \times \\ \times \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbb{X}_i^g\boldsymbol{\beta}^g + \mathbb{Z}_i^g\mathbf{b}_i^g)^{\mathbf{T}}((\sigma^g)^2\mathbb{I}_{n_i})^{-1}(\mathbf{y}_i - \mathbb{X}_i^g\boldsymbol{\beta}^g + \mathbb{Z}_i^g\mathbf{b}_i^g)\right]. \tag{3.13}$$

Vidíme, že ve tvaru hustoty daném (3.13), který vstupuje podobně jako v případě metody s marginálním rozdělením odezvy do klasifikace, se vyskytují hodnoty náhodných efektů i-tého subjektu. Tedy v případě použití této metody je potřeba náhodné efekty nějakým způsobem odhadnout. Přirozeně požadujeme, aby uvažovaný odhad splňoval nějaké vlastnosti. Jednou z přirozených vlastností odhadu je jeho nestrannost. D. Harville ve svém článku *Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects* z roku 1976 ([Harville, 1976]) ukázal, že tzv. nejlepší nestranný lineární odhad (Best linear unbiased estimation - BLUE) náhodných efektů je roven

$$\hat{\mathbf{b}}_i^g = \mathbb{E}[\mathbf{b}_i^g|\mathbf{y}_i^g], \tag{3.14}$$

což je konkrétně rovno

$$\hat{\mathbf{b}}_i^g = \mathbb{D}^g(\mathbb{Z}_i^g)^{\mathbf{T}}(\mathbb{V}_i^g)^{-1}(\mathbf{y}_i - \mathbb{X}_i^g\hat{\boldsymbol{\beta}}_i^g), \tag{3.15}$$

kde $\mathbb{V}_i^g = \mathbb{Z}_i^g\mathbb{D}^g(\mathbb{Z}_i^g)\mathbf{T} +^{(} \sigma^g)^2\mathbf{I}_{n_i}$.

Do samotné klasifikace vstupují hodnoty odhadnutých náhodných efektů. Můžeme tedy intuitivně tvrdit, že pokud se náhodné efekty v rámci i-tého subjektu nebudou mezi skupinami příliš lišit, měla by klasifikace touto metodou poskytovat horší výsledky než metoda s marginálním přístupem.

3.2.3 Přístup s rozdělením náhodných efektů

Poslední možností volby hustoty v (3.9), kterou zde představíme, je přímo hustota náhodných efektů. Dle předpokladů normálního lineárního smíšeného modelu víme, že rozdělení $\mathbf{b}_i^g$ je vícerozměrné normální se střední hodnotou $\mathbf{0}$ a varianční maticí $\mathbb{D}^g$. Tedy

$$\mathbf{b}_i^g \sim \mathcal{N}_q(\mathbf{0}, \mathbb{D}^g).$$

Konkrétně je tvar hustoty roven

$$f^g(\mathbf{b}_i) = (2\pi))^{-\frac{q}{2}} |\mathbb{D}^g|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{b}_i)^{\mathbf{T}}(\mathbb{D}_i^g)^{-1}(\mathbf{b}_i)\right]. \tag{3.16}$$

Tento tvar hustoty opět dosadíme do výrazu (3.9) a uvědomíme-li si opět, že jmenovatel ve (3.9) je pro i-tý subjekt konstantní pro každé $g = 1, \ldots, G$, dostaneme po zlogaritmování tvar g-tého diskriminačního skóru v podobě

$$\log \pi^g - \frac{1}{2} \log |\mathbb{D}_i^g| - \frac{1}{2} \left(\mathbf{b}_i\right)^{\mathbf{T}} \left(\mathbb{D}_i^g\right)^{-1} \left(\mathbf{b}_i\right). \tag{3.17}$$

Z výrazu (3.17) je patrné, že v případě použití této metody ke klasifikaci budeme opět potřebovat odhadnout hodnoty náhodných efektů. Jako odhad náhodných efektů přirozeně použijeme ten, který byl představen v Sekci 3.2.2.

Motivací pro přístup s náhodnými efekty je to, že pokud je hodnota $\hat{\mathbf{b}}_i^g$ v závislosti na skupině více vzdálená od nuly, pak takovéto skupiny mají menší hodnotu věrohodnosti, a tedy i nízkou aposteriorní pravděpodobnost. Naopak, pokud je hodnota $\hat{\mathbf{b}}_i^g$ pro nějakou skupinu blízko nuly, pak to indikuje, že tento subjekt patří do takovéto skupiny.

Až dosud jsme uvažovali, že jednotlivá pozorování přicházejí v čase, který je diskrétní. Jednou z možností, jak zobecnit konkrétně metodu využívající rozdělení náhodých efektů, je vnímat čas spojitě. Konkrétněji se tomuto tématu věnuje poslední část této kapitoly.

3.2.4 Vývoj odezvy jako náhodný proces v čase

Jak již bylo naznačeno na konci minulé části, budeme v následujících řádcích považovat vývoj odezvy i-tého subjektu za náhodný proces se spojitým časem. Naším cílem je opět nějakým způsobem reprezentovat odezvu i-tého subjektu konečným vektorem nějakých náhodných veličin. Postupně se tedy budeme snažit zobecnit metodu hlavních komponent na náhodný proces se spojitým časem.

Předpokládejme, že $\mathbf{Y}_i = \{Y_{ij}, j \in \mathbb{R}\}$ je náhodný proces se spojitým časem, který popisuje trajektorii i-tého subjektu, $i = 1, \ldots, k$. Varianční struktura náhodného procesu je popsána jeho autokovarianční funkcí. Připomeňme si její definici.

Definice 3.9. Nechť $\mathbf{Y}_i = \{Y_{ij}, j \in \mathbf{T}\}$, $T \subset \mathbb{R}$, je náhodný proces takový, že pro každé $j \in \mathbf{T}$ existuje střední hodnota $\mathrm{E}Y_{ij}$. Potom funkci $\mu_{ij} = \mathrm{E}Y_{ij}$ definovanou na T nazveme *střední hodnotou procesu* $\mathbf{Y}_i = \{Y_{ij}, j \in \mathbf{T}\}$.

Jestliže $\mathbf{Y}_i = \{Y_{ij}, j \in \mathbf{T}\}$ je reálný proces s konečnými druhými momenty, tedy $\mathrm{E}|Y_{ij}|^2 < \infty$ pro všechna $j \in \mathbf{T}$, pak funkci dvou proměnných definovanou na $T \times T$ předpisem

$$G_i(h, j) = \mathrm{cov}(Y_{ih}, Y_{ij}) = \mathrm{E}(Y_{ih} - \mu_{ih})(Y_{ij} - \mu_{ij})$$

nazveme *autokovarianční funkcí procesu* $\mathbf{Y}_i = \{Y_{ij}, j \in \mathbf{T}\}$.

Dále definujme autokovarianční operátor předpisem

$$(A_{G_i} f)(j) = \int f(s) G_i(s, j) \mathrm{d}s. \tag{3.18}$$

Označme symbolem a_i vlastní funkci operátoru A_{G_i}, kterou získáme jako řešení rovnice $(A_{G_i} a_i)(j) = \lambda a_i(j)$, kde λ je vlastní hodnota operátoru A_{G_i}. Uspořádejme vlastní hodnoty do klesající posloupnosti, tedy $\lambda_{i1} \geq \lambda_{i2} \geq \ldots$, a nechť

$\sum_l \lambda_{il} < \infty$. Předpokládejme, že k operátoru A_{G_i} máme posloupnost ortonormálních vlastních funkcí ϕ_{il}, která splňuje

$$\int \phi_{ij}(s)\phi_{il}(s)\mathrm{d}s = \delta_{jl},$$

kde δ_{jk} je Kroneckerův symbol. Jinými slovy toto znamená, že $(A_{G_i}\phi_{il})(j) = \lambda_l\phi_{il}(j)$.

Autokovarianční funkci G_i (která reprezentuje jádro lineárního integrálního operátoru A) můžeme pomocí vlastních hodnot λ_{il} a vlastních funkcí ϕ_{il} rozvinout do nekonečné řady [Karhunen, 1946]. Tedy

$$G_i(j,h) = \sum_{l=1}^{\infty} \lambda_{il}\phi_{il}(j)\phi_{il}(h), \tag{3.19}$$

a námi uvažovaný proces $\mathbf{Y}_i = \{Y_{ij}, j \in \mathbf{T}\}$ rozvineme do řady

$$Y_{ij} = \mu_{ij} + \sum_{l=1}^{\infty} \xi_{il}\phi_{il}(j), \tag{3.20}$$

kde koeficienty ξ_{il} jsou nekorelované náhodné veličiny se střední hodnotou 0 a rozptylem λ_{il} takové, že

$$\xi_{il} = \langle Y_i - \mu_i, \phi_{il}\rangle = \int (Y_{ij} - \mu_{ij})\phi_{il}(j)\mathrm{d}j. \tag{3.21}$$

Poznámka 3.10. *Podařilo se nám tedy zobecnit metodu hlavních komponent. Funkce ϕ_{il} jsou vlastně zobecněné hlavní komponenty a koeficienty ξ_{il} hrají roli skórů hlavních komponent.*

Další problém, který nyní vyřešíme, je ten, že jak zobecněných hlavních komponent, tak skórů je obecně nekonečně mnoho. Uvažujme nějaké konečné L. Pak vyjádření (3.20) můžeme aproximovat výrazem

$$Y_{ij} = \mu_{ij} + \sum_{l=1}^{L} \xi_{il}\phi_{il}(j),$$

kde L lze hiearchicky určit z výrazu

$$F(L) = \frac{\sum_{l=1}^{L} \lambda_{il}}{\sum_{l=1}^{\infty} \lambda_{il}}. \tag{3.22}$$

Jmenovatel funkce $F(L)$ je vlastně celková variabilita i-tého subjektu. L se většinou volí tak, že podíl v (3.22) překročí určitou hranici (např. 0.95). Nyní máme každý subjekt reprezentován L-rozměrným náhodným vektorem $(\xi_{i1}, \ldots, \xi_{iL})$. Pokud nás zajímá analogie s lineárním smíšeným modelem, tak koeficienty ξ_{il} hrají roli náhodných efektů $\mathbf{b}_i$ a vlastní funkce ϕ_{il} hrají roli matice $\mathbb{Z}_i$.

Dále si uvědomíme, že trajektorie každého subjektu je náhodná, tedy pro i-tý subjekt máme v čase T_j, který je také náhodný, model

$$Y_{iT_j} = \mu_{iT_j} + \sum_{l=1}^{\infty} \xi_{il}\phi_{il}(T_j) + \epsilon_{iT_j}, \tag{3.23}$$

kde náhodné chyby měření ϵ_{iT_j} jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotu 0 a rozptylem σ^2, nezávislé s ostatními veličinami vyskytujícími se v (3.23). Potom platí, že $\mathrm{var}(Y_{iT_j}) = G_i(T_j, T_j) + \sigma^2$. Dále dostáváme

$$\mathrm{E}[\xi_{il}|\mathrm{Y}_i] = \mathrm{E}(\xi_{il}) + \mathrm{cov}(\xi_{il}, \mathrm{Y}_i)(\mathrm{cov}(\mathrm{Y}_i, \mathrm{Y}_i))^{-1}(\mathrm{Y}_i - \boldsymbol{\mu}_i) = \lambda_{il}\phi_{il}^{\mathbf{T}}\boldsymbol{\Sigma}_i^{-1}(\mathrm{Y}_i - \boldsymbol{\mu}_i), \tag{3.24}$$

kde

$$\boldsymbol{\mu}_i = (\mu_{iT_1}, \ldots, \mu_{iT_{n_i}}),$$

$$(\boldsymbol{\Sigma}_i)_{T_j,T_l} = \mathrm{cov}(Y_{iT_j}, Y_{iT_l}) + \sigma^2\delta_{T_jT_l} = G_i(T_j, T_l) + \sigma^2\delta_{T_jT_l},$$

což je analogie k výsledku získanému v (3.15). K praktickému výpočtu je potřeba neznámé parametry, kterými jsou $\boldsymbol{\mu}_i$, autokovarianční funkce G_i, vlastní funkce ϕ_{il} a koeficienty ξ_{il}, odhadnout. V tomto textu pouze naznačíme jeden z možných neparametrických postupů. V případě dalších podrobností se o odhadech neznámých parametrů můžeme dozvědět například v článku H. G. Müllera [Müller, 2005].

Nejprve je potřeba odhadnout funkci střední hodnoty $\boldsymbol{\mu}_i$. Jeden z možných přístupů, jak odhadnout funkci střední hodnoty pro i-tý subjekt, je vyhladit ji například pomocí hladkých splinů, kde vstupními údaji jsou časy pozorování t_{ij} a k nim příslušné napozorované hodnoty y_{ij}. Pomocí získaného odhadu funkce střední hodnoty můžeme dále jednoduše odhadnout autokovarianční funkci i-tého subjektu výrazem $\hat{G}_i(j,l) = (y_{ij} - \hat{\mu}_{ij})(y_{il} - \hat{\mu}_{il})$. Dále se spočte odhad vlastních funkcí ϕ_{il} a vlastních hodnot λ_{il}, a to pomocí spektrálního rozkladu příslušné konečné kovarianční matice, kterou jsme získali z odhadů autokovarianční funkce G_i a funkce střední hodnoty $\boldsymbol{\mu}_i$. Zbývá odhadnout koeficienty ξ_{il}. Víme, že pro koeficienty ξ_{il} platí $\xi_{il} = \int (Y_{ij} - \mu_{ij})\phi_{il}(j)\mathrm{d}j$. Toto vyjádření můžeme aproximovat pomocí Riemannovy sumy následovně

$$\hat{\xi}_{il} = \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_{ij})\hat{\phi}_{il}(j)(t_{i,j} - t_{i,j-1}),$$

kde zvolíme $t_{i,0} = 0$.

Celkově tedy dostáváme odhad (opět připomínáme analogii s náhodnými efekty)

$$\hat{\xi}_{il} = \mathrm{E}[\xi_{il}|\mathbf{y}_i] = \hat{\lambda}_l\hat{\phi}_{il}^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i). \tag{3.25}$$

Uvedený výsledek v (3.25) můžeme použít nyní ke klasifikaci. Za předpokladu, že $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iL})$ má vícerozměrné normální rozdělení, dostaneme výsledek analogický výsledku (3.17).

Uvedené přístupy ke klasifikaci na základě longitudinálních pozorování určitě plně

nevyčerpávají naše možnosti. Jednou z dalších možností, jak zvolit míru příslušnosti ke g-té skupině, může být například sdružené rozdělení $(\mathbf{Y}_i, \mathbf{b}_i)$. Pojítkem mezi popsanými, ale i dalšími metodami, je využití lineárního smíšeného modelu, který dostatečným způsobem popisuje chování i-tého subjektu a dokáže identifikovat rozdíly mezi skupinami subjektů v rámci disjunktních skupin. Reálnou aplikaci uvažovaných metod si předvedeme v následující kapitole. Vlastnostmi zkoumaných metod se pak zabývá poslední kapitola.

Dále poznamenejme, že obecně nemůžeme říci nebo určit, která z popsaných metod funguje lépe a která hůře. Vždy to záleží na konkrétní úloze. Na variabilitě subjektů patřících do g-té skupiny, vzdálenosti průběhu odezvy středních subjektů v jednotlivých skupinách atd.

K uzavření této kapitoly poznamenejme, že jsme v jejím obsahu všude předpokládali spojitost odezvy $\mathbf{Y}_i$. V případě, že máme k dispozici taková pozorování, která nemůžeme pokládat za výběr ze spojitého rozdělení, nestojí před námi neřešitelný problém. Jednou z cest, po které se můžeme vydat, se dostaneme k zobecněným lineárním smíšeným modelům (GLMM), což je vlastně zobecnění logistické nebo poissonovské regrese. Další postup je pak koncepčně shodný s postupem a úvahami uvedených v této kapitole.

Kapitola 4

Aplikace lineárního smíšeného modelu na longitudinální data

V této části se zaměříme na modely dat, na jejichž základě bude dle Kapitoly 3 provedena klasifikace pacientů. Poznamenejme, že tato práce se primárně nezabývá modelováním longitudinálních dat, ale jejich klasifikací. Budeme tedy uvádět pouze finální podobu modelů. Pro případné zájemce jsou skripty určené k hledání konkrétních modelů součástí přílohy na CD. Dále poznamenejme, že modely se odhadují zvlášť pro každou ze sledovaných skupin. Ke každému modelu budou uvedeny odhady koeficientů jednou z metod uvedených v Podkapitole 2.2 a příslušné 95 % intervaly spolehlivosti, které jsou založeny na Waldově přístupu a asymptotické normalitě **(RE)ML** odhadů.

Postupně se v této kapitole podíváme na modely a výsledky pro vývoj hladiny bilirubinu, albuminu a velice zajímavý bude pohled na vývoj počtu krevních destiček. V další části této kapitoly bude představen společný model pro všechny sledované odezvy a výsledky popsaných metod aplikované na model se všemi sledovanými veličinami.

Nejprve uveďme další podrobnosti ke zkoumaným datům. Jak již bylo řečeno dříve, máme k dispozici data obsahující údaje o pacientech, kteří trpí onemocněním jater. Jednotlivé pacienty chceme klasifikovat do jedné ze dvou skupin. Zajímá nás, zda pacient přežije dané období osmi let (*skupina 1*) - v dalším textu označované jako skupina živých - či nikoliv (*skupina 2*) - v dalším textu označované jako skupina mrtvých - bez transplantace jater. Přitom se zaměříme pouze na pacienty, kteří přežili bez transplantace 910 dnů. Za účelem diskriminační analýzy je potřeba z tréninkových dat vyloučit cenzorované pacienty. Přesněji, z dat vyloučíme ty pacienty, u kterých došlo ve sledovaném období osmi let buď k transplantaci jater, nebo se výzkumu přestali účastnit nějakým jiným způsobem a my nevíme, do které skupiny je zařadit. Celkově máme k dispozici 195 pacientů, z toho 104 pacientů patří do skupiny živých (*skupina 1*) a 91 do skupiny mrtvých (*skupina 2*). Celkový počet pozorování je pro albumin a bilirubin 706 pozorování. U krevních destiček jsou nějaká chybějící pozorování navíc (to může být způsobeno například vysokými náklady na zjištění počtu krevních destiček), což nám ovlivní pouze celkový počet pozorování, který je pro počet krevních destiček roven 693. Počet pacientů je v obou skupinách stejný jako u albuminu a bilirubinu.

Poslední oddíly jednotlivých částí obsahují výsledky popsaných metod. Apriorní pravděpodobnosti jsou dány četnostmi ve skupinách. Tedy apriorní pravděpodobnost pro skupinu živých (*skupina 1*) je rovna 0.53 a pro skupinu mrtvých (*skupina 2*) je rovna 0.47.

Kvalita klasifikace byla ohodnocena metodou *cross-validation*.

4.1 Aplikace metod na model s bilirubinem

Bilirubin je odpadovým produktem metabolismu červeného krevního barviva hemu. Vzniká v játrech při filtraci krve ze zaniklých červených krvinek. Jeho obsah v krvi se může zvyšovat jako příznak určitých onemocnění.

Běžná hladina bilirubinu je v rozmezí 3.4-17.1 μmol/l. Zvýšení hladiny svědčí o poruše metabolismu bilirubinu. Hlavním příznakem poruch metabolismu bilirubinu je *hyperbilirubinemie*. Při zvýšení hladiny nad 30 μmol/l pozorujeme žluté zabarvení kůže a sliznic, které je způsobeno ukládáním bilirubinu ve tkáních.

4.1.1 Lineární smíšený model pro hladinu bilirubinu

Nyní se již zaměříme na konkrétní modely pro hladinu bilirubinu. Na Obrázku 4.1, resp. Obrázku 4.2 vidíme vývoj hladiny bilirubinu, resp. logaritmu bilirubinu jednotlivých pacientů v čase. Z obrázků je patrné, že pacienti, kteří žili bez transplantace pouze krátce, mají výrazně vyšší hladinu bilirubinu.

Pro zajímavost se můžeme podívat i na Obrázek 4.3, kde je znázorněn vývoj několika konkrétních pacientů, kteří přežili bez transplantace, a na Obrázek 4.4, kde je znázorněn vývoj hladiny bilirubinu konkrétních pacientů, kteří nepřežili bez transplantace dané období osmi let. Opět vidíme výrazně nižší hladinu bilirubinu u pacientů, kteří přežili dané období. Na druhou stranu nemůžeme jednoznačně říci, že by v té či oné skupině hladina bilirubinu v čase rostla nebo klesala.

Výsledný model pro hladinu bilirubinu pro i-tého pacienta a j-té měření pro jednotlivé skupiny má tvar

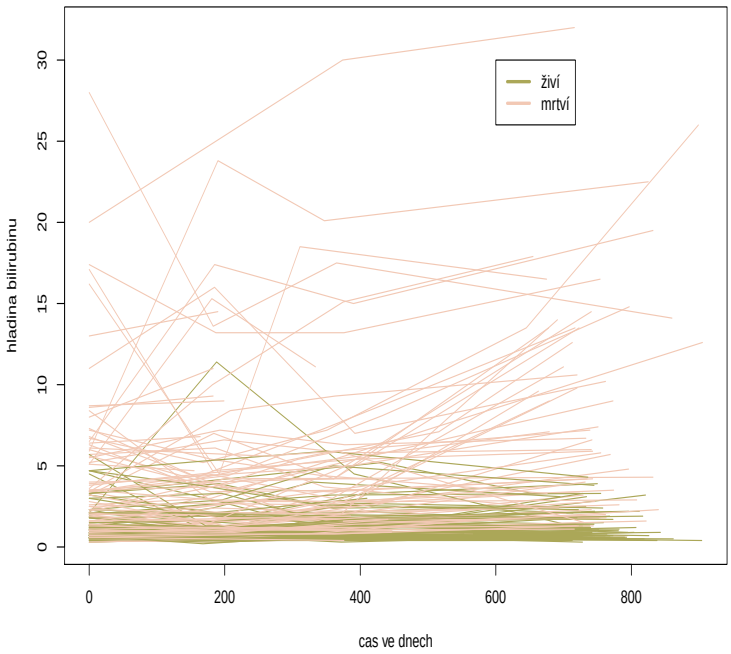
$$Y_{i,j}^g = \beta_0^g + \beta_1^g x_{i,j} + \beta_2^g x_{i,j}^2 + b_{i,0}^g + b_{i,1}^g x_{i,j} + \epsilon_{i,j}^g, \tag{4.1}$$

kde $Y_{i,j}^g = \log(\mathrm{bili_{i,j}})$, tj. j-tá hodnota měření logaritmu bilirubinu u i-tého pacienta v g-té skupině, $x_{i,j} = \mathrm{day}_{ij}$, tj. počet dní od začátku prvního měření, $g = 1, 2$. Pomocí značení z Kapitoly 2 můžeme přepsat tento model do maticové podoby. Pro i-tého pacienta má pak model tvar

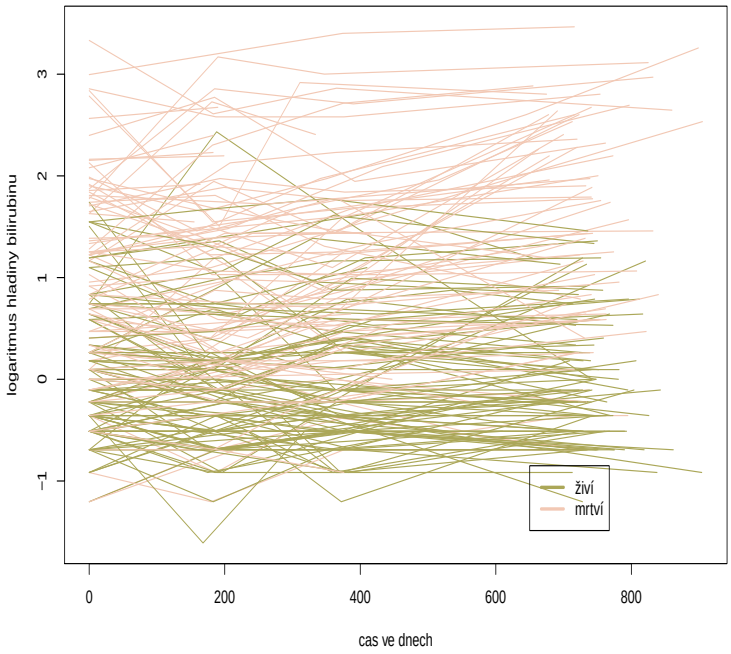
$$\mathbf{Y}_i^g = \mathbb{X}_i \boldsymbol{\beta}^g + \mathbb{Z}_i^g \mathbf{b}_i^g + \boldsymbol{\epsilon}_i^g, \tag{4.2}$$

kde

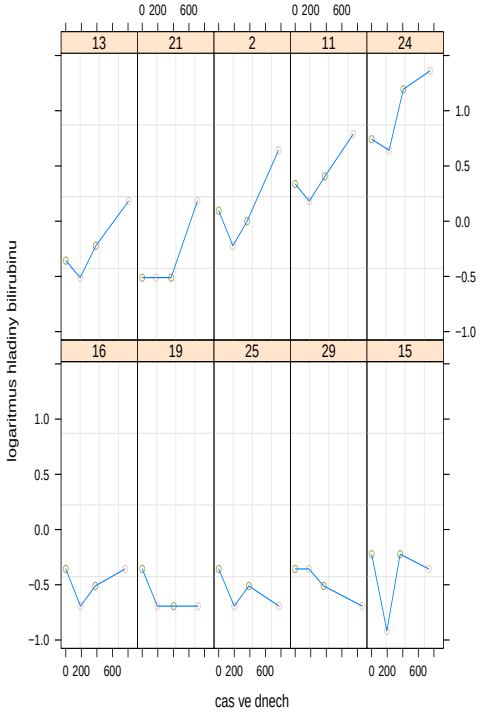
$$\boldsymbol{\beta}^g = (\beta_0^g, \beta_1^g, \beta_2^g),$$



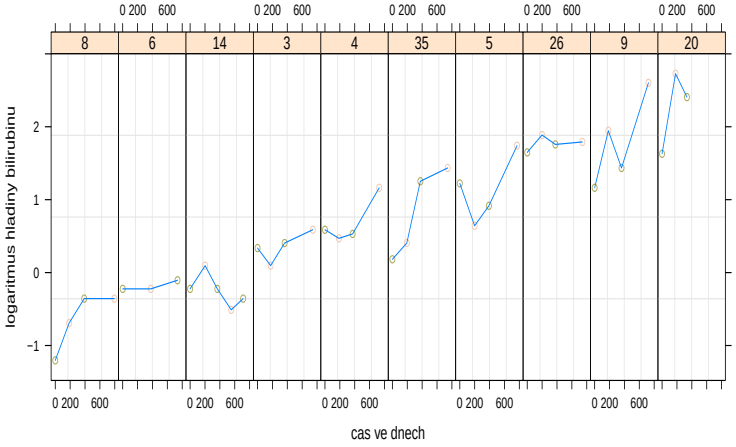
Obrázek 4.1: Vývoj hladiny bilirubinu pro jednotlivé pacienty



Obrázek 4.2: Vývoj hladiny logaritmu bilirubinu pro jednotlivé pacienty



Obrázek 4.3: Vývoj hladiny logaritmu bilirubinu pro jednotlivé přežité pacienty



Obrázek 4.4: Vývoj hladiny logaritmu bilirubinu pro jednotlivé nedožité pacienty

$$\mathbf{b}_i^g = (b_{i,0}^g, b_{i,1}^g),$$

$$\mathbb{X}_i = \begin{pmatrix} 1 & x_{i,1} & x_{i,1}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{i,n_i} & x_{i,n_i}^2 \end{pmatrix},$$

a

$$\mathbb{Z}_i = \begin{pmatrix} 1 & x_{i,1} \\ \vdots & \vdots \\ 1 & x_{i,n_i} \end{pmatrix}.$$

Maticový zápis celého modelu je

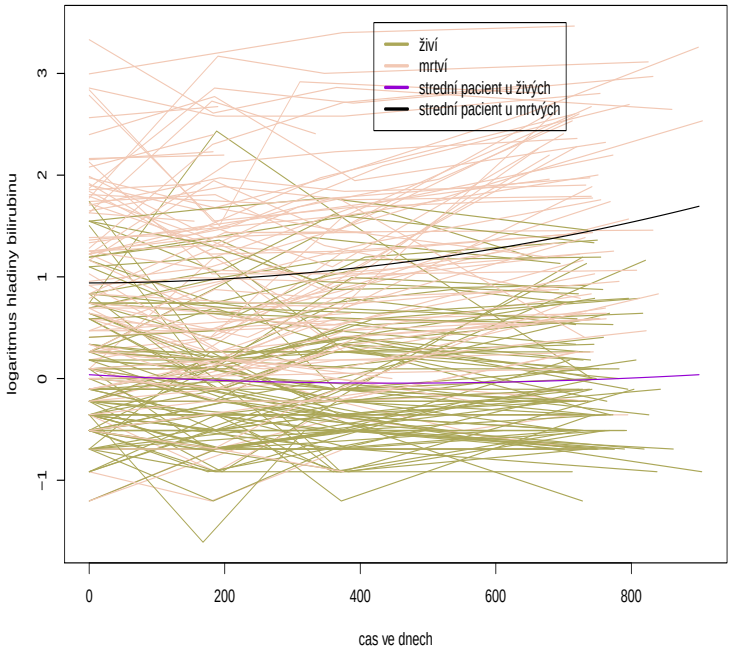
$$\mathbf{Y}^g = \mathbb{X}\boldsymbol{\beta}^g + \mathbb{Z}^g\mathbf{B}^g + \boldsymbol{\epsilon}^g,$$

kde jednotlivé složky jsou uvedeny v Podkapitole 2.1. Pro názornou představu se můžeme podívat na Obrázek 4.5, kde můžeme pozorovat odhadnuté křivky pro obě skupiny. Odhadnuté křivky můžeme interpretovat jako vývoj odezvy středního (průměrného) pacienta v g-té skupině, $g = 1, 2$. Na Obrázku 4.5 si také můžeme všimnout, že odhadnuté křivky jsou od sebe poměrně daleko. To interpretujeme tak, že skupina mrtvých a živých se liší hlavně v počáteční hodnotě hladiny bilirubinu. Tento poznatek bude mít i vliv na výsledky klasifikace, které uvedeme v následující části.

Pro úplnost nám zbývá doplnit výsledné odhady parametrů (pevné efekty a residuální směrodatná odchylka) a 95 % intervaly spolehlivosti. Tyto údaje jsou v závislosti na skupině uvedeny v Tabulce 4.1.

4.1.2 Výsledky aplikace metod s hladinou bilirubinu

V této části se podíváme na výsledky jednotlivých klasifikačních metod pro hladinu bilirubinu. V Tabulce 4.2 můžeme sledovat chybovost zařazování, senzitivitu a specificitu. Jak jsme již uvedli dříve, obě skupiny se liší hlavně v počáteční hladině bilirubinu, tedy intuitivně by měla nejlépe fungovat metoda s marginálním rozdělením. V Tabulce 4.2 konkrétně vidíme, že pro data bilirubinu je klasifikace výrazně úspěšnější a kvalitnější pro metodu s marginálním rozdělením a metodu využívající rozdělení náhodných efektů než metoda s podmíněným rozdělením. Dále si můžeme všimnout, že u všech metod vychází poměrně vysoká specificita, což nám vlastně říká, že nízké procento pacientů, kteří jsou pozitivní (tedy v našem případě nepřežijí dané období osmi let, tj. *skupina 2*), označíme jako negativní, tedy zařadíme do *skupiny 1*. Naopak senzitivita u všech metod vyšla výrazně nižší než specificita. To může být způsobeno tím, že dat pro skupinu živých je více než pro skupinu mrtvých, a tedy odhadnutý model by měl být pro skupinu živých přesnější. Tedy že pacienti, kteří nepřežijí dané období osmi let, mají v některých případech vývoj odezvy (tedy hladiny bilirubinu) podobný odhadnutému vývoji hladiny bilirubinu ve skupině pacientů, kteří přežili dané období. V Tabulce 4.3 můžeme pak sledovat přímo procenta (řádková) zařazených



Obrázek 4.5: Odhadnuté křivky hladiny logaritmu bilirubinu pro středního pacienta v obou skupinách

Skupina živých			
parametr	odhad parametru	dolní hranice	horní hranice
$\hat{\beta}_0^1$	$3.8 \cdot 10^{-2}$	$-8.9 \cdot 10^{-2}$	$1.6 \cdot 10^{-1}$
$\hat{\beta}_1^1$	$-3.7 \cdot 10^{-4}$	$-7.6 \cdot 10^{-4}$	$1.8 \cdot 10^{-5}$
$\hat{\beta}_2^1$	$4.1 \cdot 10^{-7}$	$-5.9 \cdot 10^{-8}$	$8.8 \cdot 10^{-7}$
$\hat{\sigma}^1$	0.29	0.27	0.33
Skupina mrtvých			
parametr	odhad parametru	dolní hranice	horní hranice
$\hat{\beta}_0^2$	$9.4 \cdot 10^{-1}$	$7.4 \cdot 10^{-1}$	1.14
$\hat{\beta}_1^2$	$1.3 \cdot 10^{-5}$	$-4.7 \cdot 10^{-4}$	$-4.9 \cdot 10^{-4}$
$\hat{\beta}_2^2$	$9.2 \cdot 10^{-7}$	$3.3 \cdot 10^{-7}$	$1.5 \cdot 10^{-6}$
$\hat{\sigma}^2$	0.32	0.28	0.36

Tabulka 4.1: Odhady parametrů a 95% interval spolehlivosti pro model s bilirubinem

pacientů do jednotlivých skupin u konkrétních metod.

4.2 Aplikace metod na model s hladinou albuminu

Albumin je jeden z proteinů krevní plazmy, tvoří 60% všech plazmatických bílkovin. Kromě krve se vyskytuje také v dalších tělních tekutinách, jako je tkáňový a mozkomíšní mok. Je důležitý hlavně při transportu různých látek krví (mastné kyseliny, minerály, léky) a pomáhá udržet stálé vnitřní prostředí organismu.

Albumin je, podobně jako většina proteinů krevní plazmy, syntetizovám v játrech. Ty produkují za den asi 12 g albuminu, což je 50% proteinů vylučovaných játry a čtvrtina celkových proteinů v játrech tvořených. Proto se jakákoliv porucha schopnosti jater syntetizovat proteiny projeví sníženým množstvím albuminu. Jedním z možných onemocnění je *hypoalbuminemie*, tedy snížená koncentrace albuminu, která může být způsobena poklesem syntézy v játrech, například při proteinové podvýživě nebo při onemocnění jater, jako je jaterní cirhóza. Pokles koncentrace albuminu může být příznakem zánětů, akutních stavů nebo nádorů. Vzácně se vyskytuje *analbuminemie*, defekt syntézy albuminu. Pacienti mají v plazmě jen velmi malé množství albuminu (max. 10% normálních hodnot). Kupodivu jsou pacienti většinou klinicky zdraví. Opačný stav - *hyperalbuminemie*, kdy je koncentrace albuminu naopak vyšší, je většinou způsoben dehydratací, kdy dojde ke snížení objemu vody v plazmě a albuminu je relativně více. Pro zajímavost dodejme, ze albumim slouží i k transportu bilirubinu, neboť ten je špatně rozpustný v krevní plazmě, přenáší se tedy navázaný na albumin.

4.2.1 Lineární smíšený model pro hladinu albuminu

Představení modelu pro hladinu albuminu bude probíhat podle podobného scénáře jako pro hladinu bilirubinu. Na Obrázku 4.6 a Obrázku 4.7 si opět můžeme prohlédnout hladinu albuminu pro všechny pacienty dohromady a na Obrázku 4.8 a Obrázku 4.9 se pak můžeme podívat na průběh několika jednotlivých pacientů u obou sledovaných skupin. Opět z obrázků můžeme usoudit, že skupina pacientů, kteří přežili dané období (*skupina 1*), má tentokrát vyšší hladinu albuminu (resp. logaritmu albuminu) než skupina pacientů, kteří dané období nepřežili (*skupina 2*).

Výsledný model, pro hladinu albuminu pro i-tého pacienta a j-té měření v g-té skupině má jednodušší strukturu než v případě bilirubinu. Konkrétní tvar je

$$Y_{i,j}^g = \beta_0^g + \beta_1^g x_{i,j} + b_{i,0}^g + \epsilon_{i,j}^g, \tag{4.3}$$

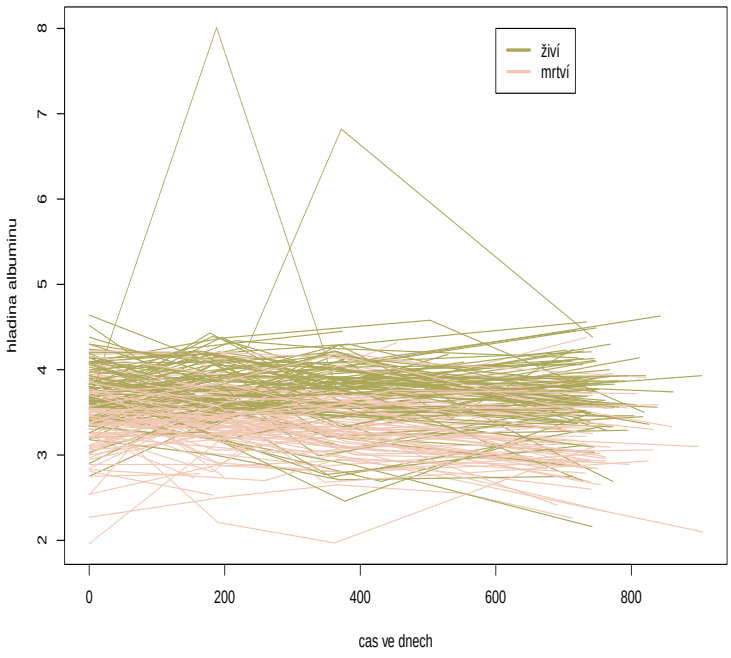
kde $Y_{i,j}^g$ je logaritmus hladiny albuminu i-tého pacienta při j-tém pozorování ve skupině g, $g = 1, 2$, a interpretace ostatních parametrů vyskytujících se v (4.3) je stejná jako u modelů v (4.1) Maticový zápis pro i-tého pacienta, případně maticový zápis modelů odvodíme analogicky jako v Části 4.1.1.

metoda	chybovost	senzitivita	specificita
Přístup s marginálním rozdělením	23.6 %	68.1 %	83.7 %
Přístup s podmíněným rozdělením	27.7 %	62.6 %	80.8 %
Přístup s náhodnými efekty	23.1 %	60.4 %	91.3 %

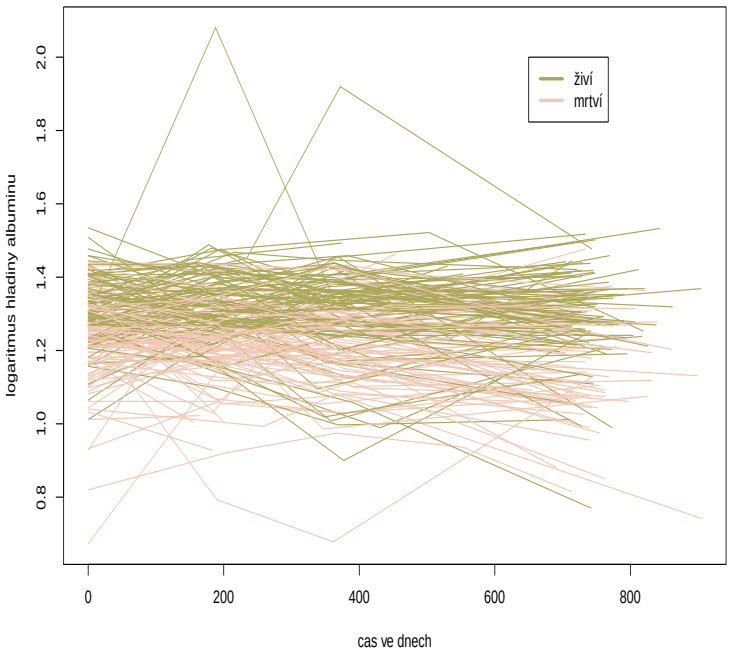
Tabulka 4.2: Shrnutí výsledků pro hladinu bilirubinu

		Výpočet		
			Živí	Mrtví
Skutečnost	1.metoda	Živí	83.7 %	16.3 %
		Mrtví	31.9 %	68.1 %
	2.metoda	Živí	80.8 %	19.2 %
		Mrtví	37.4%	62.6%
	3.metoda	Živí	91.3 %	8.7 %
		Mrtví	39.6%	60.4%

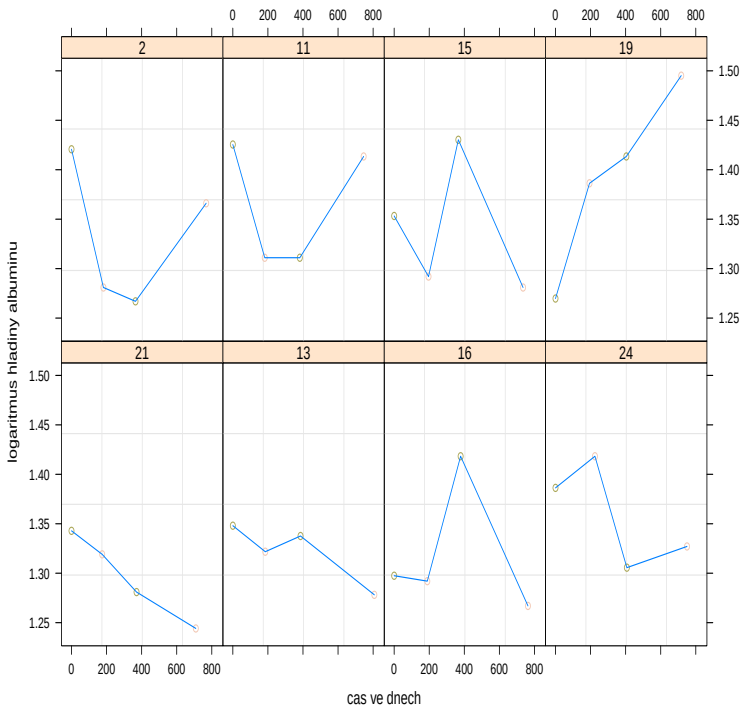
Tabulka 4.3: Shrnutí výsledků metod s hladinou bilirubinu



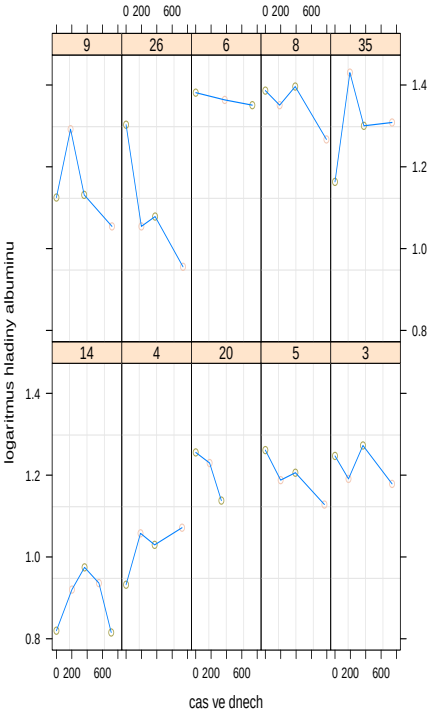
Obrázek 4.6: Vývoj hladiny albuminu pro jednotlivé pacienty



Obrázek 4.7: Vývoj hladiny logaritmu albuminu pro jednotlivé pacienty



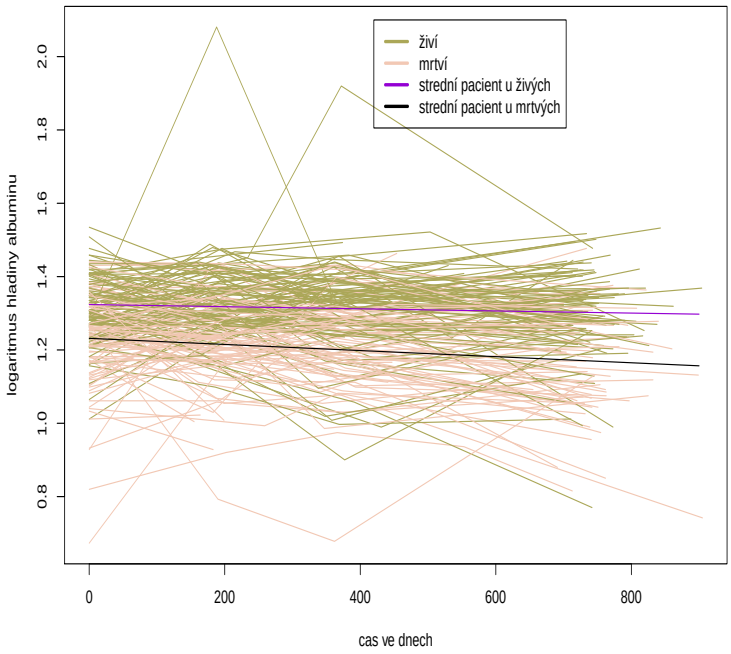
Obrázek 4.8: Vývoj hladiny logaritmu albuminu pro jednotlivé přežité pacienty



Obrázek 4.9: Vývoj hladiny logaritmu albuminu pro jednotlivé nedožité pacienty

Na Obrázku 4.10 pak můžeme opět sledovat odhadnuté křivky pro tyto modely. Vidíme, že odhadnuté průběhy odezev středních pacientů obou skupin, jsou tentokrát mnohem blíže u sebe. Tedy výsledky klasifikací by měly být minimálně v případě metody, která využívá marginální rozdělení odezvy, horší než v případě bilirubinu.

V Tabulce 4.4 jsou opět uvedeny odhady a intervaly spolehlivosti pro pevné efekty a residuální rozptyl modelu podle skupin.



Obrázek 4.10: Odhadnutá křivka hladiny logaritmu albuminu pro středního pacienta

Skupina živých			
parametr	odhad parametru	dolní hranice	horní hranice
$\hat{\beta}_0^1$	1.32	1.30	1.34
$\hat{\beta}_1^1$	$-2.9 \cdot 10^{-5}$	$-6.4 \cdot 10^{-5}$	$6.2 \cdot 10^{-6}$
$\hat{\sigma}^1$	0.10	0.09	0.11
Skupina mrtvých			
parametr	odhad parametru	dolní hranice	horní hranice
$\hat{\beta}_0^2$	1.23	1.21	1.26
$\hat{\beta}_1^2$	$-8.3 \cdot 10^{-5}$	$-1.2 \cdot 10^{-4}$	$-4.4 \cdot 10^{-5}$
$\hat{\sigma}^2$	0.09	0.08	0.10

Tabulka 4.4: Odhady parametrů a 95% interval spolehlivosti pro model s albuminem

metoda	chybovost	senzitivita	specificita
Přístup s marginálním rozdělením	23.1 %	69.2 %	83.7 %
Přístup s podmíněným rozdělením	26.2 %	60.4 %	85.6 %
Přístup s náhodnými efekty	23.6 %	65.9 %	85.6 %

Tabulka 4.5: Shrnutí výsledků pro hladinu albuminu

4.2.2 Výsledky aplikace metod s hladinou albuminu

Analogicky jako u bilirubinu můžeme v Tabulce 4.5 sledovat základní výsledky, jakými jsou celková chybovost zařazování, senzitivita a specificita. Z Tabulky 4.5 můžeme vidět podobné výsledky jako u bilirubinu. Opět se jeví metoda s podmíněným rozdělením odezvy jako nejhorší z hlediska úspěšnosti klasifikace. Naopak tato metoda má společně s třetí metodou nejvyšší specificitu. Celkově vidíme, že klasifikace na základě modelů pro albumin přináší podobné výsledky jako model s bilirubinem.

V Tabulce 4.6 pak můžeme opět sledovat podíly zařazených pacientů.

4.3 Aplikace metod na model s počtem krevních destiček

Krevní destička (*trombocyt*) savců je bezjaderné tělísko se schopností přilnavosti a shlukování, které se podílí na procesu zástavy krvácení a srážení krve. Krevní destičky jsou stálou součástí krve. Jejich množství v 1 mikrolitru krve se pohybuje mezi 200-400 tisíci. Při poklesu pod určitou úroveň začnou játra produkovat hormon *thrombopoetin*, který stimuluje tvorbu dalších destiček. Snížení počtu krevních destiček se označuje jako *trombocytopenie*. Tento stav je příčinou vážných poruch srážlivosti krve. Je tedy zřejmé, že pokud jsou játra nějakým způsobem poškozená nebo nemocná, sníží se v důsledku nedostatku hormonu *thrombopoetinu* počet krevních destiček v krvi.

		Výpočet		
Skutečnost			Živí	Mrtví
	1.metoda	Živí	83.7 %	16.3 %
		Mrtví	30.8 %	69.2 %
	2.metoda	Živí	85.6 %	14.4 %
		Mrtví	39.6%	60.4%
	3.metoda	Živí	85.6 %	14.4 %
		Mrtví	34.1%	65.9%

Tabulka 4.6: Shrnutí výsledků metod s hladinou albuminu

4.3.1 Lineární smíšený model pro počet krevních destiček

Při představení modelu opět postupujeme analogicky jako v předešlých dvou případech. Na Obrázku 4.11 a 4.12 můžeme opět sledovat vyvíjející se počet krevních destiček u všech pacientů v obou skupinách. Na těchto obrázcích si můžeme všimnout, že pacienti, kteří přežili dané období, a pacienti, kteří dané období nepřežili, netvoří v grafu žádné shluky, tak jak je tomu například u hladiny albuminu nebo bilirubinu (viz např. Obrázek 4.1 nebo Obrázek 4.6). Klasifikace na základě této veličiny bude tedy jistě velmi zajímavá. Na Obrázku 4.13 a Obrázku 4.14 je pak znázorněn vývoj počtu krevních destiček několika pacientů v obou skupinách.

Výsledný model pro počet krevních destiček se v tomto případě pro jednotlivé skupiny liší, uvedeme je tedy zvlášť. Model pro skupinu živých má tvar

$$Y^1_{i,j} = \beta^1_0 + \beta^1_1 x_{i,j} + \beta^1_2 x^2_{i,j} + b^1_{i,0} + b^1_{i,1} x_{i,j} + \epsilon^1_{i,j}, \tag{4.4}$$

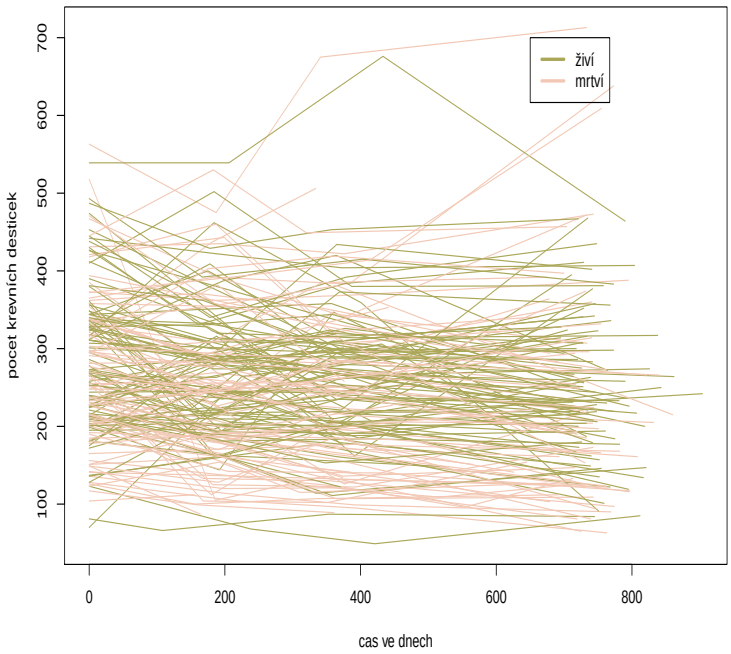
a pro skupinu mrtvých má výsledný model podobu

$$Y^2_{i,j} = \beta^2_0 + \beta^2_1 x_{i,j} + b^2_{i,0} + b^2_{i,1} x_{i,j} + \epsilon^2_{i,j}, \tag{4.5}$$

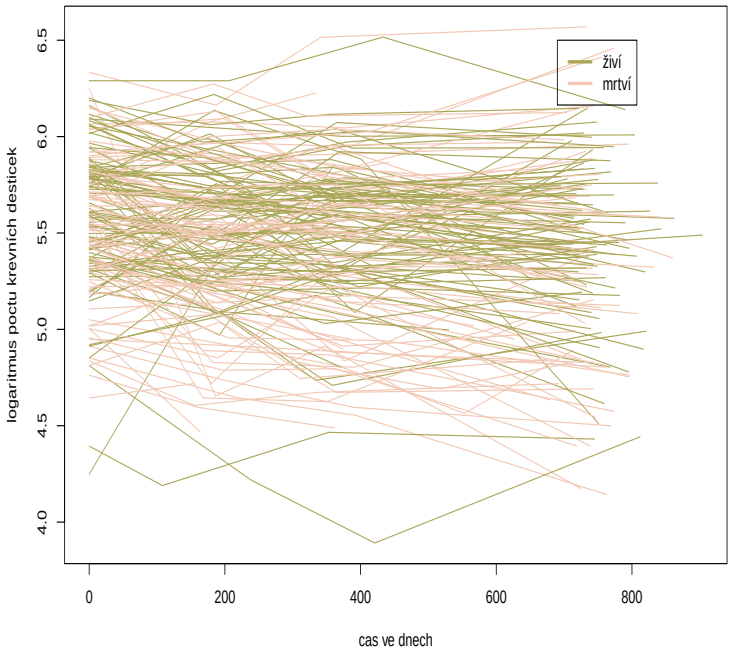
kde $Y^g_{i,j}$ je logaritmus počtu krevních destiček i-tého pacienta při j-tém měření v g-té skupině, $g = 1, 2$. Vysvětlení dalších parametrů vyskytujících se v (4.4) a v (4.5) jsou uvedeny u modelu (4.1). Maticový zápis modelu je pak obdobný maticovému zápisu v Sekci 4.1.1.

Na Obrázku 4.15 pak můžeme sledovat odhadnuté křivky reprezentující středního pacienta v jednotlivých skupinách. Jak jsme již poznamenali dříve, vidíme na Obrázku 4.15, že odhadnuté křivky obou skupin jsou velmi blízko u sebe, což bude mít jistě vliv na klasifikaci (intuitivně by se pacienti měli klasifikovat do správných skupin s menší úspěšností než u bilirubinu a albuminu).

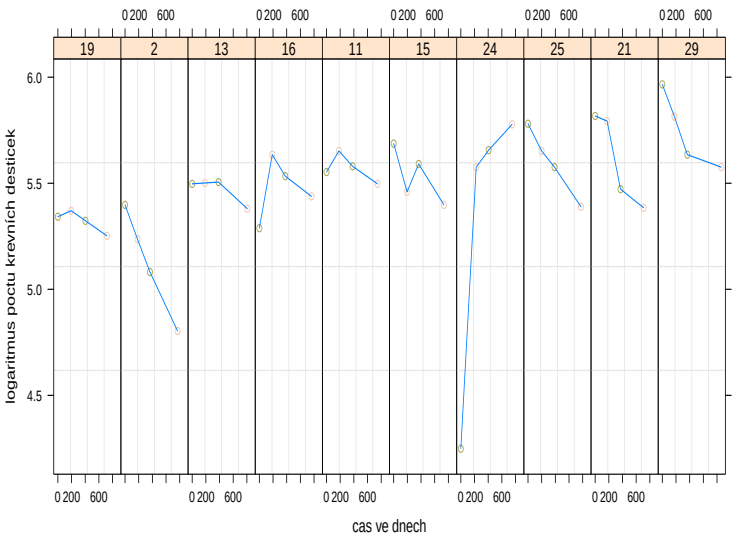
V Tabulce 4.7 jsou uvedeny odhady a intervaly spolehlivosti pro pevné efekty a residuální směrodatnou odchylku modelu podle jednotlivých skupin.



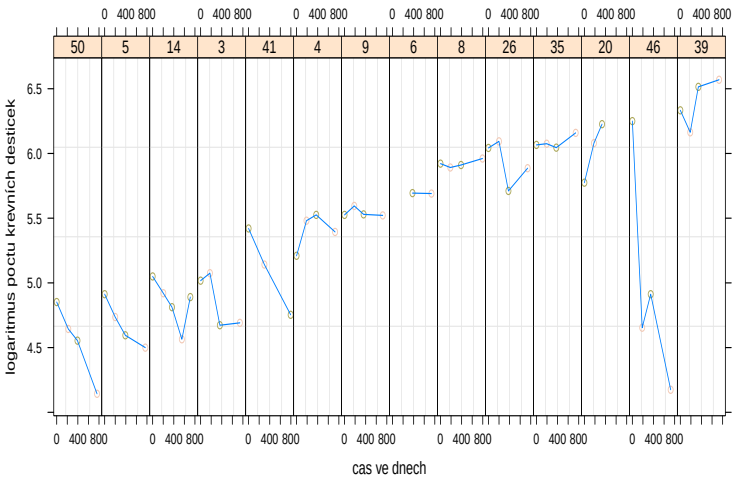
Obrázek 4.11: Vývoj počtu krevních destiček pro jednotlivé pacienty



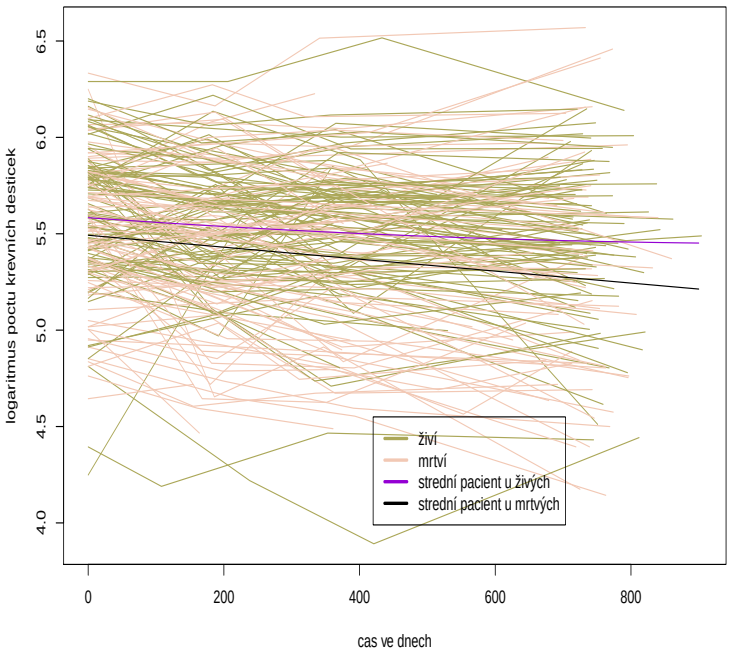
Obrázek 4.12: Vývoj logaritmu počtu krevních destiček pro jednotlivé pacienty



Obrázek 4.13: Vývoj logaritmu počtu krevních destiček pro jednotlivé přežité pacienty



Obrázek 4.14: Vývoj logaritmu počtu krevních destiček pro jednotlivé nedožité pacienty



Obrázek 4.15: Odhadnutá křivka počtu logaritmu krevních destiček pro středního pacienta

Skupina živých			
parametr	odhad parametru	dolní hranice	horní hranice
$\hat{\beta}_0^1$	5.58	5.51	5.66
$\hat{\beta}_1^1$	$-2.6 \cdot 10^{-4}$	$-5.1 \cdot 10^{-4}$	$-3.8 \cdot 10^{-7}$
$\hat{\beta}_2^1$	$1.2 \cdot 10^{-7}$	$-1.9 \cdot 10^{-9}$	$4.3 \cdot 10^{-7}$
$\hat{\sigma}^1$	0.19	0.17	0.21
Skupina mrtvých			
parametr	odhad parametru	dolní hranice	horní hranice
$\hat{\beta}_0^2$	5.49	5.42	5.57
$\hat{\beta}_1^2$	$-3.1 \cdot 10^{-4}$	$-4.2 \cdot 10^{-4}$	$-1.9 \cdot 10^{-4}$
$\hat{\sigma}^2$	0.18	0.16	0.20

Tabulka 4.7: Odhady parametrů a 95% interval spolehlivosti pro model s krevními destičkami

metoda	chybovost	senzitivita	specificita
Přístup s marginálním rozdělením	33.9 %	42.9 %	86.6 %
Přístup s podmíněným rozdělením	48.2 %	63.7 %	41.3 %
Přístup s náhodnými efekty	37.4 %	38.5 %	83.7 %

Tabulka 4.8: Shrnutí výsledků pro počet krevních destiček

			Výpočet	
			Živí	Mrtví
Skutečnost	1.metoda	Živí	86.6 %	13.4 %
		Mrtví	57.1 %	42.9 %
	2.metoda	Živí	41.3 %	58.7 %
		Mrtví	36.3 %	63.7 %
	3.metoda	Živí	83.7 %	16.3 %
		Mrtví	61.5 %	38.5 %

Tabulka 4.9: Shrnutí výsledků metod s počtem krevních destiček

4.3.2 Výsledky aplikace metod s počtem krevních destiček

V této části jsou na řadě výsledky metod aplikované na počet krevních destiček. Opět můžeme v Tabulce 4.8 sledovat celkovou chybovost klasifikace, senzitivitu a specificitu. Znázorněná Tabulka 4.8 je jistě zajímavější než v případě výsledků s bilirubinem nebo albuminem. Všimněme si hlavně vysoké chybovosti u všech tří metod. Metoda s marginálním přístupem a metoda využívající rozdělení náhodných efektů má velice nízkou senzitivitu, naopak specificita je poměrně vysoká. U metody s podmíněným rozdělením odezvy je tomu naopak.

V Tabulce 4.9 můžeme sledovat procentuální zastoupení pacientů dle výsledků klasifikace.

4.4 Společný model pro bilirubin, albumin a počet krevních destiček

K popisu chování i-tého jedince se nemusíme omezovat pouze na jednu ze sledovaných proměnných, ale můžeme vytvořit společný model pro všechny možné odezvy. V našem případě popíšeme společný model pro hladinu bilirubinu, albuminu a počet krevních destiček.

4.4.1 Společný lineární smíšený model

Nejprve sestavme společný model pro všechny sledované odezvy dohromady. Teoretický zápis modelu není téměř nic nového. Využijeme naznačené struktury

ze závěru Kapitoly 2. Společný model pro i-tého pacienta má vzhledem ke skupině tvar

$$\mathbf{Y}_i^g = \mathbb{X}_i^g \boldsymbol{\beta}^g + \mathbb{Z}_i^g \mathbf{b}_i^g + \boldsymbol{\epsilon}_i^g, \tag{4.6}$$

kde

$$\mathbf{Y}_i^g = \left(\mathbf{Y}_i^{g,\mathrm{bil}}, \mathbf{Y}_i^{g,\mathrm{alb}}, \mathbf{Y}_i^{g,\mathrm{plat}}\right)^{\mathbf{T}},$$

tj. nejprve měření pro hladinu bilirubinu, pak měření pro hladinu albuminu a nakonec měření pro počet krevních destiček. Dále $\mathbb{X}_i^g$ je blokově diagonální matice tvaru

$$\mathbb{X}_i^g = \begin{pmatrix} \mathbb{X}_i^{g,\mathrm{bil}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{X}_i^{g,\mathrm{alb}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{X}_i^{g,\mathrm{plat}} \end{pmatrix},$$

kde matice $\mathbf{X}_i^{g,bil}$, $\mathbf{X}_i^{g,alb}$, $\mathbf{X}_i^{g,plat}$ jsou shodné s maticemi v (4.2), (4.3) a (4.4) resp. (4.5). Analogicky matice $\mathbb{Z}_i^g$ je opět blokově diagonální matice, kde prvky na diagonále jsou určeny analogicky jako v případě matice $\mathbb{X}_i^g$, tj.

$$\mathbb{Z}_i^g = \begin{pmatrix} \mathbb{Z}_i^{g,\mathrm{bil}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{Z}_i^{g,\mathrm{alb}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{Z}_i^{g,\mathrm{plat}} \end{pmatrix}.$$

Vyjádření $\boldsymbol{\beta}^g = (\boldsymbol{\beta}^{g,\mathrm{bil}}, \boldsymbol{\beta}^{g,\mathrm{alb}}, \boldsymbol{\beta}^{g,\mathrm{plat}})^{\mathbf{T}}$ a $\mathbf{b}_i^g = (\mathbf{b}_i^{g,\mathrm{bil}}, \mathbf{b}_i^{g,\mathrm{alb}}, \mathbf{b}_i^{g,\mathrm{plat}})$ opět odpovídá jednotlivým samostatným modelům. Na tomto místě ještě poznamenejme, že odhady $\boldsymbol{\beta}^g$ jsou podobné odhadům v samostatných modelech. Avšak varianční matice náhodných efektů a varianční matice odhadů pevných efektů nejsou obecně blokově diagonální. Dále poznamenejme, že o $\boldsymbol{\epsilon}_i^g$ přirozeně nepředpokládáme, že je rovno $(\sigma^g)^2 \mathbb{I}_{t_i}$, kde t_i je dimenze vektoru $\mathbf{Y}_i^g$, ale že hodnota $(\sigma^g)^2$ je různá pro složky bilirubinu, albuminu a krevních destiček. Odhady residuálních rozptylů jsou opět podobné odhadům v separátních modelech.

Pro přehlednost zapišme společný model pro i-tého pacienta vzhledem ke skupině g následujícím způsobem

$$\begin{aligned} \mathbf{Y}_i^g = \begin{pmatrix} \mathbf{Y}_i^{g,\mathrm{bil}} \\ \mathbf{Y}_i^{g,\mathrm{alb}} \\ \mathbf{Y}_i^{g,\mathrm{plat}} \end{pmatrix} &= \begin{pmatrix} \mathbb{X}_i^{g,\mathrm{bil}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{X}_i^{g,\mathrm{alb}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{X}_i^{g,\mathrm{plat}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{g,\mathrm{bil}} \\ \boldsymbol{\beta}^{g,\mathrm{alb}} \\ \boldsymbol{\beta}^{g,\mathrm{plat}} \end{pmatrix} + \\ &+ \begin{pmatrix} \mathbb{Z}_i^{g,\mathrm{bil}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{Z}_i^{g,\mathrm{alb}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{Z}_i^{g,\mathrm{plat}} \end{pmatrix} \begin{pmatrix} \mathbf{b}_i^{g,\mathrm{bil}} \\ \mathbf{b}_i^{g,\mathrm{alb}} \\ \mathbf{b}_i^{g,\mathrm{plat}} \end{pmatrix} + \boldsymbol{\epsilon}_i^g \end{aligned} \tag{4.7}$$

metoda	chybovost	senzitivita	specificita
Přístup s marginálním rozdělením	17.4 %	76.9 %	87.5 %
Přístup s podmíněným rozdělením	23.6 %	64.8 %	86.5 %
Přístup s náhodnými efekty	21.0 %	80.2 %	77.9 %

Tabulka 4.10: Shrnutí výsledků pro společný model

			Výpočet	
			Živí	Mrtví
Skutečnost	1.metoda	Živí	87.5 %	12.5 %
		Mrtví	23.1 %	76.9 %
	2.metoda	Živí	86.5 %	13.5 %
		Mrtví	35.2 %	64.8 %
	3.metoda	Živí	77.9 %	22.1 %
		Mrtví	19.8 %	80.2 %

Tabulka 4.11: Shrnutí výsledků metod pro společný model

4.4.2 Aplikace metod na společný model a výsledky

V poslední části této kapitoly se podíváme na výsledky aplikovaných metod na společný model pro bilirubin, albumin a krevní destičky. V Tabulce 4.10 se můžeme podívat na celkovou chybovost klasifikace, senzitivitu specificitu. Vidíme, že dosažené výsledky jsou lepší než v případě separátních modelů. U každé z metod se celková chybovost klasifikace dle očekávání snížila. Můžeme tedy říci, že společný model, který jsme použili ke klasifikaci, je nejlepší volbou.

V Tabulce 4.11 můžeme sledovat procentuální zastoupení pacientů dle výsledků klasifikace.

Na závěr této kapitoly shrneme dosažené výsledky. Nejkvalitnější výstupy byly dosaženy pro společný model bilirubinu, albuminu a krevních destiček. Na druhou stranu je společný model celkem dost složitý. Ze separátních modelů se nejlépe jevil model, který jako odezvu využíval hladinu bilirubinu, případně albuminu. Model s hladinou bilirubinu měl tu výhodu, že odhadnuté průběhy středních pacientů v obou skupinách jsou relativně daleko od sebe, model s hladinou albuminu má pak jednodušší strukturu.

Nejlepší metodu nemůžeme obecně určit, neboť velice záleží na konkrétních datech a úloze. V našem případě se nejlépe jevila metoda využívající marginální rozdělení odezvy. Ve špatném světle se neukázala ani metoda založená na rozdělení náhodných efektů. Všechny metody se vyznačovaly vysokou specificitou a v některých případech velice nízkou senzitivitou, což mohlo být způsobeno kvalitou odhadnutých modelů, případně konkrétním zadáním úlohy.

Kapitola 5

Simulační studie

V poslední kapitole této práce se zaměříme na simulační studii, která bude zkoumat některé vlastnosti navržených metod. Podíváme se na dvě zkoumané simulace. U obou navržených simulací bylo při daném scénáři zvoleno 500 opakování pro 20, 50, 100 a 500 pacientů. V Tabulce 5.1 můžeme sledovat výsledky simulací pro základní nastavení takové, aby odpovídalo reálným datům, které jsme použily v předešlé kapitole. Konkrétně byla za scénář pro odezvu zvolena hladina bilirubinu. Naopak v Tabulce 5.2 můžeme sledovat výsledky simulací, u kterých jsme jako scénář pro odezvu použili počet krevních destiček.

Na první pohled je zřejmé, že výsledky pro bilirubin jsou výrazně lepší než pro krevní destičky. To je způsobeno především tím, že střední průběh pacienta, který přežije dané období osmi let, je u hladiny bilirubinu výrazně odlišný od pacienta, který dané období nepřežije. Naopak u počtu krevních destiček nic takového nepozorujeme. U metody, která využívá marginální rozdělení odezvy, a také u metody, která využívá ke klasifikaci rozdělení náhodných efektů, můžeme sledovat docela nízkou chybovost klasifikace. Ta je vlastně odhad střední hodnoty ztrátové funkce, která nemusí být vždy nutně nulová (proto v našem případě celková chybovost s rostoucím počtem pozorování hned neklesá). Z důvodu konzistence tohoto odhadu je ale velice důležité, aby klesal rozptyl této chybovosti, což se potvrdilo.

Na závěr této kapitoly poznamenejme, že i když dle výsledků dopadla nejhůř metoda s podmíněným rozdělením odezvy, neznamená to, že v některých případech nemůže fungovat nejlépe. Důležité jsou vždy konkrétní vlastnosti úlohy, rozdílnost uvažovaných skupin atd. Vidíme například, že v porovnání s ostatními metodami vyšla metoda s pomíněným rozdělením odezvy lépe u scénáře s počtem krevních destiček, neboť právě u těchto dat je marginální rozlišení obou skupin minimální.

	marginální přístup				podmíněný přístup				přístup s náhodnými efekty			
počet pacientů	chybovost	rozptyl chyb.	senz.	spec.	error	rozptyl chyb.	senz.	spec.	error	rozptyl chyb.	senz.	spec.
20	16.4 %	0.60 %	73.9 %	91.2 %	28.9 %	0.96 %	66.8 %	74.7 %	22.2 %	0.88 %	59.6 %	92.3 %
50	19.5 %	0.30 %	68.5 %	89.7 %	31.2 %	0.52 %	59.9 %	75.8 %	21.9 %	0.30 %	57.8 %	94.0 %
100	20.6 %	0.13 %	67.2 %	89.3 %	32.0 %	0.26 %	57.0 %	77.1 %	23.0 %	0.14 %	56.7 %	93.6 %
500	21.7 %	0.03 %	65.3 %	88.9 %	31.9 %	0.04 %	51.2 %	81.4 %	23.2 %	0.03 %	56.2 %	93.6 %

Tabulka 5.1: Výsledky simulací pro jednotlivé metody s nastavením bilirubin

	marginální přístup				podmíněný přístup				přístup s náhodnými efekty			
počet pacientů	chybovost	rozptyl chyb.	senz.	spec.	error	rozptyl chyb.	senz.	spec.	error	rozptyl chyb.	senz.	spec.
20	24.3 %	0.89 %	67.6 %	82.4 %	32.1 %	0.85%	69.0 %	67.0 %	34.0 %	1.11 %	55.5 %	74.5 %
50	29.7 %	0.37 %	57.9 %	80.1 %	36.1 %	0.43%	62.4 %	65.0 %	35.1 %	0.56 %	53.4 %	73.9 %
100	32.4 %	0.19 %	52.0 %	80.5 %	37.5 %	0.20%	60.0 %	64.5 %	35.5 %	0.26 %	48.0 %	77.9 %
500	34.3 %	0.03 %	45.2 %	82.5 %	38.6 %	0.06%	61.6 %	61.3 %	35.2 %	0.03 %	40.3 %	84.8 %

Tabulka 5.2: Výsledky simulací pro jednotlivé metody s nastavením krevních destiček

Kapitola 6

Závěr

V práci jsem se zaměřil na modifikaci metod klasické diskriminační analýzy na pozorování longitudinálního typu. V první části jsem představil základní vlastnosti (normálního) lineárního smíšeného modelu a metody jeho odhadu. Lineární smíšený model je velice vhodný nástroj na popis longitudinálních dat. Další část jsem již věnoval diskriminační analýze. Nejprve jsem připomenul klasickou diskriminační analýzu a souvislost s teorií ztrátových funkcí. Poté jsem klasické metody pomocí lineárního smíšeného modelu modifikoval na longitudinální data. Postupně jsem představil přístup s marginálním rozdělením odezvy, s podmíněným rozdělením odezvy a přístup s rozdělením náhodných efektů. Tento oddíl jsem zakončil částí, která zobecňuje přístup s rozdělením náhodných efektů na případ se spojitým časem.

Ve druhé části této práce jsem se zaměřil na reálnou aplikaci popsaných metod. V jednotlivých oddílech jsem nejprve představil konkrétní modely pro hladinu bilirubinu, albuminu a počet krevních destiček. Poté jsem pro každou odezvu uvedl výsledky aplikovaných metod. Metody jsem samostatně naprogramoval pomocí statistického programu R. V práci jsou uvedeny pouze dosažené výsledky, pro případné zájemce jsou skripty součástí přílohy na CD. V poslední části jsem se pak zaměřil na simulační studii zkoumající vlastnosti popsaných metod.

Práce se mi psala velmi dobře, neboť statistická témata s přímou aplikací v lékařství mě velice zajímají. Využil jsem především anglicky psané literatury, neboť popsané metody dosud nebyly publikovány v českém jazyce.

Za hlavní přínos práce považuji popis metod se sjednoceným značením, které se do této doby převážně objevovaly s rozdílným značením v různých článcích a také jejich přímou aplikaci na reálná data a simulační studii.

Dalšími možnostmi, jak tuto práci rozšířit, je modifikace klasických metod shlukové analýzy na longitudinální data, případně zkoumat vlastnosti popsaných metod v případě diskrétní odezvy.

Literatura

L. J. Brant, S. L. Sheng, C. H. Morrell, G. N. Verbeke, E. Lesaffre, and H. B. Carter. Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society, Series A*, 166:51–62, 2003.

R. T. Fleming and P. D. Harrigton. *Counting processes and survival analysis.* John Wiley and Sons, NY, 2005. ISBN 0-471-52218-X.

D. Harville. Extension of the gauss-markov theorem to include the estimation of random effects. *The Annals of Statistics*, 4(2):384–395, 1976.

D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 1974.

P. Hebák, J. Hustopecký, E. Jarošová, and I. Pecáková. *Vícerozměrné statistické metody (1).* INFORMATORIUM, Praha, 2004. ISBN 978-80-7333-056-9.

J. C. Huberty. *Applied discriminant analysis.* John Wiley and Sons, NY, 1994. ISBN 0-471-31145-6.

K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Feennicae*, (A I 37), 1946.

G. Marshall and E. A. Barón. Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine*, (19):1969–1981, 2000.

H. G. Müller. Functional modelling and classification of longitudinal data. *Board of the Foundation of the Scandinavian Journal of Statistics*, pages 223–240, 2005.

Brant L. J. Morrell, C. H. and S. Sheng. Comparing approaches for predicting prostate cancer from longitudinal data. *In 2007 Proceedings of the American Statistical Association, Biometrics Section*, pages 127–133, Alexandria, 2007. American Statistical Association, 2007.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data.* Springer, New York, 2000. ISBN 0-387-95027-3.

Příloha

K práci je přiloženo CD, které obsahuje následující soubory:

- Soubor **diplomova_prace_LB.pdf** obsahuje text diplomové práce.
- Soubor **model_dat_bili.R** obsahuje skript, v němž je nalezen vhodný model pro popis vývoje hladiny bilirubinu v čase pro jednotlivé skupiny. Používá data **data_bil.RData**, která jsou také součástí CD.
- Program **model_dat_albumin.R** obsahuje skript, v němž je nalezen model pro popis vývoje hladiny albuminu v čase pro jednotlivé skupiny. Využívá data **data_alb.RData**, která jsou také součástí CD.
- Soubor **model_dat_platelet.R** obsahuje skript, ve kterém je nalezen model pro popis vývoje počtu krevních destiček v čase pro jednotlivé skupiny. Využívá data **data_plat.RData**, ktera jsou součástí CD.
- Soubor **aplikace_metod_bilirubin.R** obsahuje skript s naprogramovanými modifikacemi diskriminační analýzy pro hladinu bilirubinu. Pracuje s daty **data_bil.RData**, která jsou součástí CD.
- Soubor **aplikace_metod_albumin.R** obsahuje skript s naprogramovanými modifikacemi diskriminační analýzy pro hladinu albuminu. Využívá data **data_alb.RData**, ktera jsou součástí CD.
- Soubor **aplikace_metod_platelet.R** obsahuje skript s naprogramovanými modifikacemi diskriminační analýzy pro počet krevních destiček. Používá data **data_plat.RData**, která jsou součástí CD.
- Soubor **aplikace_metod_spolecny_model.R** obsahuje skript s naprogramovanými modifikacemi diskriminační analýzy pro společný model bilirubinu, albuminu a počtu krevních destiček. Pracuje se společnými daty **data_spol.RData**, která jsou součástí CD. Soubor dále obsahuje netriviální sestavení společného modelu.
- Soubor **simulacni_studie_platelet.R** obsahuje skript s naprogramovanou simulační studií použitých metod se základním nastavením krevních destiček. Používá data **data_plat.RData**, která jsou součástí CD.
- Soubor **simulacni_studie_bilirubin.R** obsahuje skript s naprogramovanou simulační studií použitých metod se základním nastavením bilirubinu. Využívá data **data_bil.RData**, která jsou součástí CD.