UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

POROVNÁNÍ ZOBECNĚNÉHO LINEÁRNÍHO MODELU A KOMPOZIČNÍHO MODELU PŘI ANALÝZE DAT

Vedoucí bakalářské práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval: **Bc. Dan Šafařík**
Studijní program: N1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Dan Šafařík

Název práce: Porovnání zobecněného lineárního modelu a kompozičního modelu při analýze dat

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Cílem diplomové práce je porovnat dvě různé metody analýzy kategoriálních dat, v poslední době velmi populární kompoziční data a zobecněné lineární modely (konkrétně je používán ACL model). K porovnání obou přístupů je využito datové množiny o nehodách cyklistů.

Klíčová slova: GLM, zobecněný lineární model, ACL model, model logitů sousedních kategorií, kompoziční data, ilr transformace, porovnání modelů

Počet stran: 51

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Dan Šafařík

Title: Comparison of generalized linear models and compositional models within a data analysis

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2016

Abstract: The goal of this thesis is to compare two different methods of categorical data analysis, recently very popular compositional data and generalized linear models (specifically the ACL model is used). To compare both approaches the data set on accidents of cyclists is used.

Key words: GLM, generalized linear model, ACL model, adjacent categories logit model, compositional data, ilr transformation, model comparison

Number of pages: 51

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně pod vedením pana Mgr. Ondřeje Vencálka, PhD. s použitím uvedené literatury.

V Olomouci dne 21. dubna 2016

Obsah

Poděkování

Rád bych poděkoval zejména vedoucímu mé diplomové práce panu Mgr. Ondřeji Vencálkovi PhD. za cenné rady, spolupráci a veškerý čas, který mi věnoval během konzultací. Velký dík patří také mé rodině, která mě po celou dobu studia podporovala.

Úvod

Cílem této práce je čtenáři představit dva rozdílné přístupy analýzy kategoriálních dat, aplikovat oba přístupy na data a na základě výsledků je porovnat. První použitou metodou bude zobecněný lineární model, konkrétně model logitů sousedních kategorií, tou druhou je analýza kompozičních dat. Metody budou aplikovány na reálná data o nehodách cyklistů a analýzy budou prováděny výhradně s využitím softwaru *R*.

V úvodní kapitole si představíme základní poznatky pro pochopení zobecněných lineárních modelů, připomeneme si klasický lineární model a následně uvedeme, v čem spočívá zobecnění a popíšeme princip metody maximální věrohodnosti využívané k odhadu parametrů. V druhé kapitole bude představeno několik různých logitových modelů pro analýzu nominálních a ordinálních dat. V závěru této kapitoly bude také uveden jeden z alternativních modelů pro analýzu ordinálních dat a to model logitů sousedních kategorií (z anglického adjacent categories logit model), který bude v praktické části používán pro porovnání s kompozičním modelem. Bude zde také uveden názorný příklad pro aplikaci ACL modelu. Ve třetí kapitole se podíváme na teoretické základy analýzy kompozičních dat. Uvedeme si zde základní definice pro kompoziční data, definujeme si Aitchisonovu geometrii a logratio transformaci. V závěru této kapitoly si také představíme binární sekvenční dělení, které nám v praktické části umožní vhodné vyjádření souřadnic získaných z logratio transformace.

Ve čtvrté kapitole se dostáváme k praktické části, kde si v úvodu představíme data, se kterými jsme pracovali, a provedeme pár grafických znázornění těchto dat pro snazší představu toho, co data znázorňují. Postupně přejdeme k vyjádření potřebných modelů, které budou porovnávány, a vyjádříme si vztahy mezi jejich parametry. V další podkapitole provedeme s využitím softwaru *R* odhady parametrů těchto modelů, parametry interpretujeme a podíváme se i na přesnost těchto odhadů. V následující podkapitole modely obou přístupů porovnáme, dále budou obě metody aplikovány ještě na jiný datový soubor a v závěrečných dvou

podkapitolách odhadneme pro oba přístupy varianční matici metodou maximální věrohodnosti.

1. Zobecněný lineární model

V kapitole 1.1. si připomeneme podobu klasického lineárního modelu, v 1.2. si pak ukážeme, v čem spočívá zobecnění a jak vypadají jednotlivé složky zobecněného modelu a na závěr se v kapitole 1.3. podíváme na odhad parametrů metodou maximální věrohodnosti. K vypracování této kapitoly bylo čerpáno zejména z [1], [4], [8], [9].

1.1. Klasický lineární model

Připomeňme si nejdříve, jak vypadá klasický lineární model. V klasickém lineárním regresním modelu hledáme souvislost mezi náhodnými vysvětlovanými veličinami $Y_1, Y_2, ..., Y_n$ a vysvětlujícími nenáhodnými veličinami $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$, kde $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$. Předpokládejme tedy, že hodnoty závislé proměnné Y_i se skládají ze systematické a náhodné složky

$$Y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i, \quad i = 1, ..., n \qquad \varepsilon_i \sim N(0, \sigma^2).$$

Náhodná složka má v obyčejném regresním modelu normální rozdělení s nulovou střední hodnotou, konstantním rozptylem σ^2 a náhodné složky jsou nekorelované. Díky tomu můžeme střední hodnotu vyjádřit jako

$$E(Y_i) = \beta_0 + \sum_j \beta_j x_{ij}, \quad i = 1, ..., n.$$

Parametry modelu $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ odhadujeme pomocí metody nejmenších čtverců.

1.2. Složky zobecněného lineárního modelu

Zobecnění nám umožní porušit některé z předpokladů klasického modelu a uvažovat pro veličiny $Y_1, Y_2, ..., Y_n$ jiné než normální rozdělení. Náhodná složka zobecněného lineárního modelu se skládá z vysvětlované proměnné Y s nezávislými pozorováními $(y_1, ..., y_n)$ z přirozené rodiny exponenciálních rozdělení. Funkce hustoty pravděpodobnosti této rodiny má potom tvar

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)].$$

Speciálním případem je několik důležitých rozdělení, včetně Poissonova, binomického, ale také normálního rozdělení. Hodnota parametru θ_i se může pro různá $i = 1, ..., n$ lišit v závislosti na hodnotách vysvětlujících proměnných a a, b zde představují nezáporné funkce. Výraz $Q(\theta)$ se nazývá přirozený parametr.

Nechť x_{ij} vyjadřuje hodnotu j-tého prediktoru pro i-té pozorování. Označme

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., n.$$

Systematická složka zobecněného lineárního modelu vyjadřuje lineární vztah mezi vektorem $(\eta_1, ..., \eta_n)$ a vysvětlujícími proměnnými. Lineární kombinace vysvětlujících proměnných se nazývá lineární prediktor. Obvykle je jedno z $x_{ij} = 1$ pro všechna i, odpovídající koeficient v modelu představuje absolutní člen.

Třetí složkou modelu je tzv. linková funkce, která propojuje systematickou a náhodnou složku. Nechť $\mu_i = E(Y_i), i = 1, ..., n$. Model propojuje μ_i a η_i podle vztahu $\eta_i = g(\mu_i)$, kde linková funkce g je monotónní a diferencovatelná. Odtud můžeme vidět, že g spojuje $E(Y_i)$ s vysvětlujícími proměnnými pomocí vztahu

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., n.$$

V případě klasické regrese je linkovou funkcí identita.

1.3. Odhad parametrů metodou maximální věrohodnosti

Principem metody maximální věrohodnosti je najít odhad parametru θ (popřípadě vektoru parametrů), který maximalizuje pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení pravděpodobnosti.

Uvažujme náhodný výběr $Y_1, ..., Y_n$. Máme tedy n nezávislých stejně rozdělených náhodných veličin (i.i.d.) s hustotou $f(\boldsymbol{y}, \boldsymbol{\theta})$, kde $\boldsymbol{\theta}$ představuje vektor neznámých parametrů. Sdružená hustota odpovídající n realizacím $y_1, ..., y_n$ náhodné veličiny Y pak má tvar:

$$f(y_1, ..., y_n \mid \boldsymbol{\theta}) = \textstyle\prod_{i=1}^n f(y_i;\boldsymbol{\theta}).$$

Hlavní myšlenkou metody maximální věrohodnosti je dívat se na sdruženou hustotu nikoliv jako na funkci $y_1, ..., y_n$, ale jako na funkci vektoru $\boldsymbol{\theta}$ při pevně daných hodnotách $y_1, ..., y_n$ a vybrat ze všech možných hodnot $\boldsymbol{\theta}$ tak, že výše uvedený výraz nabývá svého maxima. Za tímto účelem zavádíme tzv. věrohodnostní funkci ve tvaru

$$L(\boldsymbol{\theta} \mid y_1, ..., y_n) = f(y_1, ..., y_n \mid \boldsymbol{\theta}),$$

což je vyjádření shodné se sdruženou hustotou, kde ovšem jako proměnná vystupuje vektor neznámých parametrů $\boldsymbol{\theta}$. Maximálně věrohodný odhad vektoru parametrů $\boldsymbol{\theta}$ značíme jako $\widehat{\boldsymbol{\theta}}_{MLE}$, a je to číselný vektor, který maximalizuje funkci věrohodnosti, tedy

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \mid y_1, ..., y_n),$$

kde Θ představuje parametrický prostor, tedy prostor všech možných hodnot vektoru $\boldsymbol{\theta}$.

Mnohdy je lepší maximalizovat logaritmus věrohodnostní funkce. V této práci budeme všechny logaritmy brát jako přirozené logaritmy. Tuto tzv. logaritmickou věrohodnostní funkci pak můžeme uvést ve tvaru

$$l(\boldsymbol{\theta} \mid y_1, ..., y_n) = \log L(\boldsymbol{\theta} \mid y_1, ..., y_n) = \log \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}).$$

Je-li věrohodnostní funkce diferencovatelná, lze najít maximálně věrohodný odhad jako stacionární bod funkce L nebo l. Řešíme tedy systém rovnic, kdy položíme první derivace věrohodnostní funkce (jejího logaritmu) podle parametrů rovny nule. Následně bychom měli také ověřit, zda nalezené řešení je opravdu maximum. Toho docílíme například pomocí druhých derivací.

2. Logitové modely pro multinominální data

V této kapitole budou uvedeny různé modely, které využívají tzv. logitů, kde logit pravděpodobnosti p chápeme jako

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

V kapitole 2.1. si představíme model pro nominální data, který pro každý pár kategorií používá jiný logitový model. V kapitole 2.2. se potom podíváme na model pro ordinální data, který využívá logitů kumulativních pravděpodobností. V kapitole 2.3. použijeme pro tyto pravděpodobnosti jiné linkové funkce. V kapitole 2.4. bude ještě představen jeden z alternativních modelů pro ordinální data a to model pro logity sousedních kategorií (adjacent categories logit model – ACL model). V závěrečné části druhé kapitoly si ještě uvedeme vzorový příklad pro užití ACL modelu. K vypracování této kapitoly bylo čerpáno výhradně z [1], [2].

2.1. Nominální data: logitové modely s referenční kategorií

Nechť Y je kategorická proměnná s J kategoriemi. Označme $\pi_j(\boldsymbol{x}) = P(Y = j \mid \boldsymbol{x})$. Musí platit $\sum_j \pi_j(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x}$. Počty výskytů jednotlivých variant považujeme za náhodný vektor s multinomickým rozdělením s pravděpodobnostmi $\{\pi_1(\boldsymbol{x}), ..., \pi_J(\boldsymbol{x})\}$. Pravděpodobnosti $\pi_j(\boldsymbol{x})$ můžeme po dvojicích porovnávat, například pro skupiny π_1 a π_2 můžeme vztah mezi nimi vyjádřit jejich podílem π_1/π_2, který vyjadřuje šanci kategorie 1 proti kategorii 2. Počet všech dvojic, které můžeme porovnávat, je roven $\binom{J}{2}$. S daným výběrem $J-1$ dvojic se ty zbylé stávají nadbytečnými, což vyplývá z níže uvedeného vztahu (1).

Logitové modely párují každou kategorii odpovědí s referenční kategorií, za kterou je často brána poslední kategorie nebo například ta, kterou považujeme za nejběžnější, což může být taková kategorie, jejíž pravděpodobnost je největší.

Model

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} = \alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}, \quad j = 1, ..., J-1,$$

popisuje efekt $\boldsymbol{x}$ na těchto $J-1$ logitů. Efekty se liší v závislosti na tom, která kategorie je zrovna párována s referenční. Těchto $J-1$ rovnic určuje parametry i pro logity ostatních párů kategorií, neboť platí

$$\log \frac{\pi_a(\boldsymbol{x})}{\pi_b(\boldsymbol{x})} = \log \frac{\pi_a(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} - \log \frac{\pi_b(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}. \tag{1}$$

Odvoďme teď vzorec pro pravděpodobnosti zastoupení jednotlivých skupin $\pi_j(\boldsymbol{x})$

$$\begin{aligned}
\log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} &= \alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}, \\
\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} &= \exp(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}), \\
\pi_j(\boldsymbol{x}) &= \pi_J(\boldsymbol{x}) \exp(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}).
\end{aligned}$$

Potřebujeme vyjádřit pravděpodobnost $\pi_J(\boldsymbol{x})$. Vektor pravděpodobností $(\pi_1(\boldsymbol{x}), \pi_2(\boldsymbol{x}), ..., \pi_J(\boldsymbol{x}))$ můžeme přepsat ve tvaru $(\pi_J(\boldsymbol{x}) \exp(\alpha_1 + \boldsymbol{\beta}_1'\boldsymbol{x}), \pi_J(\boldsymbol{x}) \exp(\alpha_2 + \boldsymbol{\beta}_2'\boldsymbol{x}), ..., \pi_J(\boldsymbol{x}))$. Odtud se dostáváme k následujícím úpravám

$$\begin{aligned}
\pi_J(\boldsymbol{x}) \left(1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h'\boldsymbol{x})\right) &= 1, \\
\pi_J(\boldsymbol{x}) &= \frac{1}{\left(1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h'\boldsymbol{x})\right)}.
\end{aligned}$$

S takto vyjádřeným $\pi_J(\boldsymbol{x})$ už můžeme přejít k finálnímu vzorci pro pravděpodobnosti $\pi_j(\boldsymbol{x})$

$$\pi_j(\boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h'\boldsymbol{x})}, \quad j = 1, ..., J-1.$$

2.2. Ordinální data: modely kumulativních logitů

Modely, které berou v úvahu ordinální povahu dat, mají řadu výhod a zlepšují sílu modelu. V této kapitole se podíváme na nejpopulárnější modely pro ordinální data.

2.2.1. Kumulativní logity

Jeden způsob, jak využít ordinality kategorií, je s využitím logitů kumulativních pravděpodobností

$$P(Y \leq j \mid \boldsymbol{x}) = \pi_1(\boldsymbol{x}) + ... + \pi_j(\boldsymbol{x}), \quad j = 1, ..., J-1.$$

Kumulativní logity jsou definované jako

$$\begin{aligned} \text{logit}[P(Y \leq j \mid \boldsymbol{x})] &= \log \frac{P(Y \leq j \mid \boldsymbol{x})}{1 - P(Y \leq j \mid \boldsymbol{x})} = \\ &= \log \frac{\pi_1(\boldsymbol{x}) + ... + \pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + ... + \pi_J(\boldsymbol{x})}, \quad j = 1, ..., J-1. \end{aligned}$$

Každý kumulativní logit využívá všech J kategorií.

2.2.2. Model proporcionálních šancí

Model, který využívá všech kumulativních logitů zároveň, má tvar

$$\text{logit}[P(Y \leq j \mid \boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, ..., J-1.$$

Každý kumulativní logit má svůj absolutní člen. Tyto členy $\{\alpha_j\}$ rostou s tím, jak roste j, neboť pro pevně dané $\boldsymbol{x}$ roste s j také pravděpodobnost $P(Y \leq j \mid \boldsymbol{x})$ a logit je rostoucí funkcí této pravděpodobnosti. Tento model má stejné parametry $\boldsymbol{\beta}$ pro všechny logity.

Model kumulativního logitu uvedený výše splňuje vztah

$$\begin{aligned} &\text{logit}[P(Y \leq j \mid \boldsymbol{x}_1)] - \text{logit}[P(Y \leq j \mid \boldsymbol{x}_2)] = \\ &= \log \frac{P(Y \leq j \mid \boldsymbol{x}_1)/P(Y > j \mid \boldsymbol{x}_1)}{P(Y \leq j \mid \boldsymbol{x}_2)/P(Y > j \mid \boldsymbol{x}_2)} = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2). \end{aligned}$$

Nazvěme podíl $P(Y \leq j \mid \boldsymbol{x}_1)/P(Y > j \mid \boldsymbol{x}_1)$ kumulativní šancí, jejich poměr potom nazýváme kumulativní poměr šancí. Šance, že při $\boldsymbol{x} = \boldsymbol{x}_1$ nastane situace ze skupiny $\leq j$, je $\exp[\boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)]$-krát větší než při $\boldsymbol{x} = \boldsymbol{x}_2$. Logaritmus kumulativního poměru šancí je proporcionální ku vzdálenosti mezi $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$. Stejná proporcionalita platí pro každý logit. V případě jednoho prediktoru je kumulativní poměr šancí roven e^{β}, pokud $x_1 - x_2 = 1$.

2.3. Ordinální data: modely kumulativního linku

Modely kumulativního logitu využívají logitový link. Stejně jako v případě jednorozměrného zobecněného lineárního modelu existují i další linkové funkce. Nechť G^{-1} je linková funkce, která je inverzí distribuční funkci G nějaké spojité náhodné veličiny. Model kumulativního linku

$$G^{-1}[P(Y \leq j \mid \boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}$$

vyjadřuje vztah mezi kumulativními pravděpodobnostmi a lineárním prediktorem. Logitová linková funkce $G^{-1}(u) = \log[u/(1-u)]$ je inverzí distribuční funkce logistického rozdělení, kterou můžeme vyjádřit jako $G(v) = \dfrac{1}{1+e^{-v}}$.

2.4. Logity sousedních kategorií (ACL)

Modely pro ordinální data nemusí využívat kumulativních pravděpodobností. Jedním z takových alternativních modelů je ACL. Logity sousedních kategorií mají tvar

$$\text{logit}[P(Y = j \mid Y = j \text{ nebo } Y = j+1)] = \log\frac{\pi_j}{\pi_{j+1}}, \quad j = 1, ..., J-1.$$

Logity ACL mají souvislost s logity s referenční kategorií. Logity s referenční kategorií můžeme vyjádřit pomocí ACL jako

$$\log\frac{\pi_j}{\pi_J} = \log\frac{\pi_j}{\pi_{j+1}} + \log\frac{\pi_{j+1}}{\pi_{j+2}} + ... + \log\frac{\pi_{J-1}}{\pi_J}, \tag{2}$$

a naopak ACL můžeme vyjádřit pomocí logitů s referenční kategorií:

$$\log \frac{\pi_j}{\pi_{j+1}} = \log \frac{\pi_j}{\pi_J} - \log \frac{\pi_{j+1}}{\pi_J}, \quad j = 1, ..., J-1.$$

Modely využívající logity sousedních kategorií mohou být vyjádřeny jako modely logitů s referenční kategorií. Uvažujme například ACL model

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, ..., J-1,$$

se společným parametrem $\boldsymbol{\beta}$. Přidáním $(J-j)$ výrazů se dostaneme k ekvivalentnímu modelu pro logit sousedních kategorií. Využijeme zde vztahu (2).

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} = \sum_{k=j}^{J-1} \frac{\pi_k}{\pi_{k+1}} = \sum_{k=j}^{J-1} (\alpha_k + \boldsymbol{\beta}'\boldsymbol{x}) = \left(\sum_{k=j}^{J-1} \alpha_k \right) + (J-j)\boldsymbol{\beta}'\boldsymbol{x} =$$
$$= \alpha_j^* + \boldsymbol{\beta}'\boldsymbol{u}_j, \quad j = 1, ..., J-1,$$

kde $\boldsymbol{u}_j = (J-j)\boldsymbol{x}$. ACL bere v potaz uspořádání kategorií Y. Pokud bychom například neuvažovali ordinalitu, tak dostaneme $(J-1)$ modelů, kde každý z nich bude mít svůj rozdílný parametr $\boldsymbol{\beta}_j$. ACL nám právě díky uspořádání v kategoriích umožní efekt popsat jedním parametrem namísto $(J-1)$.

2.5. Příklad: spokojenost v práci

Na závěr této kapitoly si ještě uvedeme příklad na tvorbu ACL modelu, jelikož právě tuto metodu budeme při analyzování dat používat nejvíce. Využijeme příkladu z kapitoly 7.4.2 z [1]. Tabulka 1 se zabývá vztahem mezi spokojeností afroameričanů (muž, žena) v práci (Y) a jejich příjmem. Pro zjednodušení používáme pro příjem skóry $(1, 2, 3, 4)$. Pro příjem x a pohlaví g (1 = ženy, 0 = muži) uvažujme model

$$\log(\pi_j/\pi_{j+1}) = \alpha_j + \beta_1 x + \beta_2 g, \quad j = 1, 2, 3.$$

Tabulka 1: Spokojenost v práci

Pohlaví	Příjem (dolary)	Spokojenost v práci: Velmi nespokojený	Trochu spokojený	Středně spokojený	Velmi spokojený
Žena	< 5000	1	3	11	2
	5000 - 15 000	2	3	17	3
	15 000 - 25 000	0	1	8	5
	> 25000	0	2	4	2
Muž	< 5000	1	1	2	1
	5000 - 15 000	0	3	5	1
	15 000 - 25 000	0	0	7	3
	> 25000	0	1	9	6

Tento model popisuje šanci, že člověk bude velmi nespokojen oproti trochu spokojen, trochu oproti středně spokojen a středně oproti velmi spokojen. Model je také ekvivalentní s modelem logitů s referenční kategorií

$$\log(\pi_j/\pi_4) = \alpha_j^* + \beta_1(4-j)x + \beta_2(4-j)g, \qquad j = 1, 2, 3.$$

Využijme teď softwaru R a knihovny VGAM. Nejdříve si vytvoříme datovou množinu, načteme potřebnou knihovnu a aplikujeme ACL model.

```
> gender = c(1, 1, 1, 1, 0, 0, 0, 0)
> money = c(1, 2, 3, 4, 1, 2, 3, 4)
> A = matrix(c(1,3,11,2,2,3,17,3,0,1,8,5,0,2,4,2,1,1,2,1,0
+ ,3,5,1,0,0,7,3,0,1,9,6), nrow = 8, ncol = 4,
+ byrow = TRUE)
> data = data.frame(gender, money, A)
> colnames(data) = c('gender', 'money', 'sk1', 'sk2',
+ 'sk3', 'sk4')
> attach(data)

> library(VGAM)

> acm2 = vglm(cbind(sk1, sk2, sk3, sk4) ~ money + gender,
```

```
+ family=acat(reverse=TRUE, parallel=TRUE),data=data)
> summary(acm2)
```

Na obrázku 1 můžeme vidět odhady parametrů výše popsaného ACL modelu.

Obrázek 1: Odhady parametrů

```
Coefficients:
                    Value Std. Error
(Intercept):1 -0.550668    0.67945
(Intercept):2 -0.655007    0.52527
(Intercept):3  2.025934    0.57581
money         -0.388757    0.15465
gender         0.044694    0.31444
```

Odhady parametrů tedy jsou $\widehat{\beta_1} = -0,389$ a $\widehat{\beta_2} = 0,045$. $\widehat{\beta_1} < 0$ znamená, že šance nižší pracovní spokojenosti klesá s tím, jak roste příjem, což je očekávaný výsledek. Poměr šancí $\exp(-0,389) = 0,68$ nám udává, že šance vždy té nižší z obou skupin spokojenosti klesne $0,68$ - krát, když přejdeme do vyšší kategorie příjmu. Poměr šancí pro druhý parametr $\exp(0,045) = 1,05$ vyjadřuje, že šance nižší pracovní spokojenosti je pro ženy nepatrně větší než pro muže.

3. Kompoziční data

V této kapitole se seznámíme se základy teorie kompozičních dat. V úvodní kapitole 3.1. budou uvedeny základní definice kompozičních dat. V kapitole 3.2. se podíváme na Aitchisonovu geometrii, v kapitole 3.3. pak na logratio transformaci a v závěrečné kapitole 3.4. si představíme princip binárního sekvenčního dělení. K vypracování této kapitoly bylo čerpáno zejména z [3], [5], [6], [7].

Kompoziční data jsou speciálním typem mnohorozměrných dat, kde informaci nenesou absolutní hodnoty, ale jejich podíly. To znamená, že dávají smysl jen v tom případě, kdy jsou vázána k nějakému celku. Může se tak jednat například o data s procentuálními podíly, kde součet všech složek je roven 100, nebo můžeme brát data s proporcionálními podíly na celku, kde je součet složek roven 1.

3.1. Základní definice kompozičních dat

Definice 3.1. *Kladný reálný vektor* $\boldsymbol{x} = (x_1, ..., x_D)'$ *popisující kvantitativně části nějakého celku nesoucí výhradně relativní informaci mezi složkami se nazývá D-složkový kompoziční vektor.*

Definice 3.2. *Subkompozice* $\boldsymbol{x_s}$ *kompozice* $\boldsymbol{x}$ *je vektor* $(x_{i1}, ..., x_{is})$ *určující vybrané složky. Subindexy* $\boldsymbol{s} = (i_1, ..., i_s), 1 \leq i_1 < \ldots < i_s \leq D$, *označují složky zahrnuté do subkompozice, nemusí jít nutně o s prvních složek.*

Definice 3.3. *Uzávěr kompozice* $\boldsymbol{x} = (x_1, ..., x_D)'$ *vzhledem ke konstantnímu součtu k je vektor*

$$C(\boldsymbol{x}) = \left(\frac{kx_1}{\sum_{i=1}^{D} x_i}, ..., \frac{kx_D}{\sum_{i=1}^{D} x_i} \right)'.$$

Definice 3.4. *Vektory* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D$ *jsou kompozičně ekvivalentní, jestliže existuje číslo* $\lambda \in \mathbb{R}^+$ *takové, že platí* $\boldsymbol{x} = \lambda \boldsymbol{y}$ *Ekvivalentně lze psát jako podmínku na uzávěry* $C(\boldsymbol{x}) = C(\boldsymbol{y})$.

3.2. Aitchisonova geometrie

Pro většinu reálných dat je jejich výběrovým prostorem reálný prostor s euklidovskou geometrií, ta však není pro kompoziční data vhodná. D-složkové kompozice přirozeně indukují jiný výběrový prostor, kterým je D-složkový simplex definovaný jako

$$S^D = \left\{ \boldsymbol{x} = (x_1, ..., x_D)'; x_i > 0, i = 1, ..., D, \sum_{i=1}^{D} x_i = k \right\}.$$

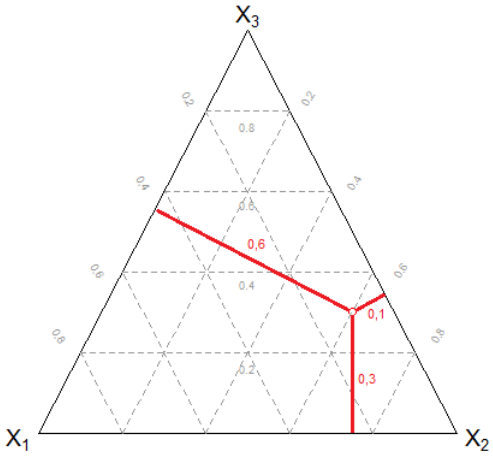
Tento simplex S^D je podmnožinou $\mathbb{R}^D$, kde $D \geq 2$. Pro $D = 2$ tak simplex tvoří úsečku a pro $D = 3$ trojúhelník. Kompozici tedy můžeme interpretovat jako bod na tomto simplexu.

Protože v praktické části pracujeme výhradně s třísložkovými kompozicemi, tak se teď podíváme na to, jak je možné je zobrazit. Při práci s daty můžeme třísložkové kompozice graficky zobrazit s využitím ternárního diagramu. Za ternární diagram považujeme rovnostranný trojúhelník s vrcholy X_1, X_2, X_3. Kompozice je v něm zobrazena jako bod a hodnoty jednotlivých složek kompozice jsou vyjádřeny jako vzdálenosti tohoto bodu od jednotlivých stran. Uvažujme například kompozici

$$\boldsymbol{a} = (0, 1, 0, 6, 0, 3).$$

Na následujícím obrázku vidíme zobrazení této kompozice v ternárním diagramu. Červené úsečky znázorňují vzdálenost bodu od jednotlivých stran, tyto vzdálenosti jsou složkami kompozice.

Obrázek 2: Ternární diagram pro třísložkové kompozice



Pro práci s kompozičními daty je nutné na simplexu zavést vhodnou geometrii, Aitchisonovu geometrii, s operacemi, které poskytují smysluplné informace o kompozicích podobně jako v euklidovském prostoru. Zavedeme teď dvě operace na simplexu, které jsou v D-dimenzionálním reálném prostoru analogické sčítání vektorů a násobení vektorů skalárem.

Definice 3.5. *Perturbace kompozice* $\boldsymbol{x} = C(x_1, ..., x_D)' \in S^D$ *kompozicí* $\boldsymbol{y} = C(y_1, ..., y_D) \in S^D$ *je kompozice* $\boldsymbol{x} \oplus \boldsymbol{y} \in S^D$ *definovaná vztahem*

$$\boldsymbol{x} \oplus \boldsymbol{y} = C(x_1 y_1, ..., x_D y_D)'.$$

Definice 3.6. *Mocninná transformace* $\boldsymbol{x} = C(x_1, ..., x_D)' \in S^D$ *číslem* $\alpha \in \mathbb{R}$ *je kompozice* $\alpha \odot \boldsymbol{x} \in S^D$ *definovaná vztahem*

$$\alpha \odot \boldsymbol{x} = C(x_1^\alpha, ..., x_D^\alpha)'.$$

Perturbace a mocninná transformace splňují následující axiomy:

I. $(S^D, \oplus)$ tvoří komutativní grupu, tj. pro libovolné kompozice $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z} \in S^D$ platí

   (a) komutativita: $\boldsymbol{x} \oplus \boldsymbol{y} = \boldsymbol{y} \oplus \boldsymbol{x}$,

(b) asociativita: $(\boldsymbol{x} \oplus \boldsymbol{y}) \oplus \boldsymbol{z} = \boldsymbol{x} \oplus (\boldsymbol{y} \oplus \boldsymbol{z})$,

(c) neutrální prvek: $\boldsymbol{n} = C(1, ..., 1)'$ a platí $\boldsymbol{x} \oplus \boldsymbol{n} = \boldsymbol{x}$,

(d) inverze: $\boldsymbol{x} \oplus \boldsymbol{x}^{-1} = \boldsymbol{n}$, kde $\boldsymbol{x}^{-1} = C(x_1^{-1}, ..., x_D^{-1})'$;

II. pro libovolné kompozice $\boldsymbol{x}, \boldsymbol{y} \in S^D$ a $\alpha, \beta \in \mathbb{R}$ platí

(a) neutrální prvek: $1 \odot \boldsymbol{x} = \boldsymbol{x}$,

(b) asociativita: $\alpha \odot (\beta \odot \boldsymbol{x}) = (\alpha \cdot \beta) \odot \boldsymbol{x}$,

(c) distributivita: $(\alpha + \beta) \odot \boldsymbol{x} = (\alpha \odot \boldsymbol{x}) \oplus (\beta \odot \boldsymbol{x})$,

(d) distributivita: $\alpha \odot (\boldsymbol{x} + \boldsymbol{y}) = (\alpha \odot \boldsymbol{x}) \oplus (\alpha \odot \boldsymbol{y})$;

$(S^D, \oplus, \odot)$ je tedy reálným vektorovým prostorem.

Nyní ještě definujeme skalární součin, normu a vzdálenost na simplexu.

Definice 3.7. *Aitchisonův skalární součin kompozic* $\boldsymbol{x}, \boldsymbol{y} \in S^D$ *definujeme vztahem*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}.$$

Definice 3.8. *Aitchisonova norma kompozice* $\boldsymbol{x} \in S^D$ *je dána jako*

$$\|\boldsymbol{x}\|_a = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_a} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j} \right)^2}.$$

Definice 3.9. *Aitchisonovu vzdálenost kompozic* $\boldsymbol{x}, \boldsymbol{y} \in S^D$ *definujeme vztahem*

$$d_a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2},$$

kde $\boldsymbol{x} \ominus \boldsymbol{y} = \boldsymbol{x} \oplus (-1 \odot \boldsymbol{y})$.

Prostor $(S^D, \oplus, \odot)$ spolu s operacemi uvedenými výše tvoří $(D-1)$-dimenzionální euklidovský vektorový prostor, který nazýváme Aitchisonova geometrie na simplexu.

3.2.1. Logratio transformace

Aitchisonova geometrie na simplexu má vlastnosti euklidovské geometrie, avšak až na výjimky není vhodná pro statistickou analýzu dat tohoto typu. Aby bylo možné využít běžné statistické metody pro analýzu kompozic, byla navržena aditivní logratio (alr) a centrovaná logratio (clr) transformace. Obě transformace však mají své nedostatky. Alr transformace není izometrická a nezachovává tak vzdálenosti, clr transformace sice je izometrická, ale vede k singulární varianční matici. Tyto nežádoucí vlastnosti tak vedly k zavedení izometrické logratio (ilr) transformace. Jejím výsledkem je reálný vektor, jehož složky jsou souřadnice vzhledem k nějaké zvolené ortonormální bázi.

Nyní se můžeme na jednotlivé transformace podívat podrobněji. Pokud využijeme alr transformace pro D-složkovou kompozici $\boldsymbol{x} = (x_1, ..., x_D)$ ze simplexu S^D, tak obdržíme $(D-1)$-rozměrný reálný vektor

$$\boldsymbol{y} = (y_1, ..., y_{D-1})' = alr(\boldsymbol{x}) = \left(\log \frac{x_1}{x_D}, ..., \log \frac{x_{D-1}}{x_D} \right)'.$$

Jak už bylo uvedeno výše, tato transformace však není izometrická a nezachovává tedy vzdálenosti, což znamená, že euklidovská vzdálenost vypočtená z alr transformovaných kompozic a Aitchisonova vzdálenost vypočtená z původních kompozic nejsou stejné. Uvažujme kompoziční vektory

$$\boldsymbol{a} = (0, 1, 0, 6, 0, 3),$$
$$\boldsymbol{b} = (0, 6, 0, 3, 0, 1),$$

a jejich alr transformace

$$alr(\boldsymbol{a}) = (-1, 099, 0, 683),$$
$$alr(\boldsymbol{b}) = (1, 792, 1, 099).$$

Teď vypočteme Aitchisonovu vzdálenost kompozic $\boldsymbol{a}$ a $\boldsymbol{b}$ a euklidovskou vzdálenost jejich alr transformací

$$d_a(\boldsymbol{a}, \boldsymbol{b}) = 2, 213,$$
$$d_e(alr(\boldsymbol{a}), alr(\boldsymbol{b})) = 2, 919.$$

Jak můžeme vidět, tak alr transformace opravdu nezachovává vzdálenost. Příčinou je, že souřadnice vypočtené alr transformací jsou vyjádřeny k bázi na simplexu, která není ortonormální. Z tohoto důvodu byla navržena clr transformace.

Zobrazením D-složkové kompozice ze simplexu S^D do reálného prostoru pomocí clr transformace dostáváme D-rozměrný vektor

$$\boldsymbol{w} = (w_1, ..., w_D)' = clr(\boldsymbol{x}) = \left(\log \frac{x_1}{g(\boldsymbol{x})}, ..., \log \frac{x_D}{g(\boldsymbol{x})} \right)',$$

kde $g(\boldsymbol{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D}$ je geometrický průměr složek kompozice. Clr transformace je sice izometrická, problémem však je, že pro její složky platí $\sum_{i=1}^{D} w_i = 0$. Uvažujme opět kompozici

$$\boldsymbol{a} = (0, 1, 0, 6, 0, 3),$$

a její clr transformaci

$$clr(\boldsymbol{a}) = (-0, 963, 0, 828, 0, 135).$$

Problém je hned vidět, neboť součet složek clr transformované kompozice je roven 0

$$-0, 963 + 0, 828 + 0, 135 = 0.$$

Dopočítejme ještě clr transformaci kompozice $\boldsymbol{b}$

$$clr(\boldsymbol{b}) = (0, 828, 0, 135, -0, 963),$$

a ověříme konzistenci

$$d_a(\boldsymbol{a}, \boldsymbol{b}) = 2, 213,$$
$$d_e(clr(\boldsymbol{a}), clr(\boldsymbol{b})) = 2, 213.$$

Můžeme vidět, že clr transformace opravdu vzdálenosti zachovává.

Všechny požadované vlastnosti pro statistické zpracování kompozic splňuje ilr transformace, která je pro vybranou ortonormální bázi definovaná jako

$$\boldsymbol{z} = (z_1, ..., z_{D-1})' = ilr(\boldsymbol{x}), \quad z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \quad i = 1, ..., D-1.$$

Definujme si vektor

$$\boldsymbol{l} = (\log x_1, \log x_2, ..., \log x_D)'.$$

Jednotlivé transformace tak můžeme vyjádřit jako

$$\boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{l}, \qquad \boldsymbol{M}_1^{(D-1)\times D} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \vdots & & & & & \vdots \\ 0 & 0 & & \cdots & 1 & -1 \end{pmatrix},$$

$$\boldsymbol{w} = \boldsymbol{M}_2 \boldsymbol{l}, \qquad \boldsymbol{M}_2^{D\times D} = \begin{pmatrix} 1 - \frac{1}{D} & -\frac{1}{D} & \cdots & -\frac{1}{D} \\ -\frac{1}{D} & 1 - \frac{1}{D} & \cdots & -\frac{1}{D} \\ \vdots & & & \vdots \\ -\frac{1}{D} & -\frac{1}{D} & \cdots & 1 - \frac{1}{D} \end{pmatrix},$$

$$\boldsymbol{z} = \boldsymbol{M}_3 \boldsymbol{l}, \qquad \boldsymbol{M}_3^{(D-1)\times D} = \begin{pmatrix} \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 & \cdots & 0 \\ \frac{\sqrt{\frac{2}{3}}}{2} & \frac{\sqrt{\frac{2}{3}}}{2} & -\sqrt{\frac{2}{3}} & \cdots & 0 \\ \vdots & & & & \vdots \\ \frac{\sqrt{\frac{D-1}{D}}}{D-1} & \frac{\sqrt{\frac{D-1}{D}}}{D-1} & \cdots & \frac{\sqrt{\frac{D-1}{D}}}{D-1} & -\sqrt{\frac{D-1}{D}} \end{pmatrix}.$$

Odtud dostáváme vztahy mezi logratio transformacemi, které můžeme vyjádřit jako

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{U}\boldsymbol{w}, \\ \boldsymbol{y} &= \boldsymbol{C}\boldsymbol{z}, \\ \boldsymbol{y} &= \boldsymbol{F}\boldsymbol{w}, \end{aligned}$$

kde

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{F}\boldsymbol{M}_2\boldsymbol{M}_2^{-1} = \boldsymbol{M}_1\boldsymbol{M}_2^{-1}, \\ \boldsymbol{U} &= \boldsymbol{U}\boldsymbol{M}_2\boldsymbol{M}_2^{-1} = \boldsymbol{M}_3\boldsymbol{M}_2^{-1}, \\ \boldsymbol{C} &= \boldsymbol{C}\boldsymbol{M}_3\boldsymbol{M}_3^{-1} = \boldsymbol{M}_1\boldsymbol{M}_3^{-1}. \end{aligned}$$

Matice $\boldsymbol{U}$, $\boldsymbol{F}$ jsou matice $(D-1) \times D$ s plnou řádkovou hodností a $\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}_{D-1}$, z čehož plyne, že matice $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{U}'$ je regulární a tak můžeme uvažovat její inverzi. Pokud se podíváme na vektory matice $\boldsymbol{U}$, tak ty tvoří vektory ortogonální báze nadroviny $w_1 + w_2 + \cdots + w_D = 0$ a $\boldsymbol{F} = [\boldsymbol{I}_{D-1}, -\boldsymbol{1}_{D-1}]$, kde $\boldsymbol{1}_{D-1}$ je vektor jedniček dimenze $D-1$. Tímto se dostáváme k dalším vztahům, které můžeme vyjádřit jako

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{U}'\boldsymbol{z}, \\ \boldsymbol{z} &= \boldsymbol{C}^{-1}\boldsymbol{y}, \\ \boldsymbol{w} &= \boldsymbol{F}^{+}\boldsymbol{y}. \end{aligned}$$

3.3. Binární sekvenční dělení

Některé speciální ortonormální báze jsou spojeny s tzv. sekvenčním binárním dělením kompozičního vektoru. Tohle je praktický způsob, jak definovat ortonormální bázi a souřadnice. Hlavní myšlenka spočívá v rozdělení souboru složek kompozice do dvou skupin těchto složek. Tyto dvě získané skupiny dále dělíme a pokračujeme tak dlouho, až každá skupina obsahuje pouze jednu složku. Počet dělení nezbytných k získání binárního sekvenčního dělení je $D-1$ a přímo souvisí s $D-1$ vektory ortogonální báze S^D.

Existuje více způsobů, jak takové binární sekvenční dělení určit. Tabulka 2 nám ukazuje jeden z nich.

Tabulka 2: Kód určující sekvenční binární dělení 6-složkové kompozice

Krok	x_1	x_2	x_3	x_4	x_5	x_6
1	+1	+1	−1	−1	+1	−1
2	+1	−1	0	0	−1	0
3	0	+1	0	0	−1	0
4	0	0	−1	+1	0	−1
5	0	0	+1	0	0	−1

Označme si hodnoty v tabulce jako k_{ij}, kde i značí krok a j skupinu. V každém kroku dělení je skupina předchozí úrovně rozdělena do dvou podskupin: v jedné

skupině jsou složky označené $+1$ a v druhé skupině složky označené jako -1. Označení 0 značí, že tato složka není v tomhle kroku součástí dělení.

Z tabulky tak můžeme vidět, že v prvním kroku dělení je skupina $\{1, 2, 5\}$ oddělena od $\{3, 4, 6\}$. V druhém kroku je potom skupina $\{1, 2, 5\}$ rozdělena na $\{1\}$ a $\{2, 5\}$ a tak dále.

i-tý prvek ortonormální báze spojený se sekvenčním binárním dělením je dán jako $\boldsymbol{e}_i = C[\exp(a_{i1}, a_{i2}, ..., a_{i,D-1})]$, kde clr koeficienty a_{ij} nabývají různých hodnot v závislosti na kódování sekvenčního binárního dělení, jejich určení si nyní ukážeme. Předpokládejme, že v kroku i je skupina o $r + s$ složkách rozdělena do dvou skupin o r složkách (kódovány pozitivně) a s složkách (kódovaných negativně). Pokud se znovu podíváme na tabulku 2, tak například pro i = 2, r = 1, s = 2. Hodnoty a_{ij} jsou potom

$$a_{ij} = a_+ = \sqrt{\frac{s}{r(r+s)}} \quad \text{pro } k_{ij} = +1,$$
$$a_{ij} = a_- = \sqrt{\frac{r}{s(r+s)}} \quad \text{pro } k_{ij} = -1,$$
$$a_{ij} = a_0 = 0 \quad \text{pro } k_{ij} = 0,$$

kde indexy odpovídají pozitivnímu, negativnímu a nulovému kódování.

Vyjádření souřadnic kompozice $\boldsymbol{x}$ s ohledem na ortonormální bázi definovanou binárním sekvenčním dělením je

$$z_i = \log \frac{\left(\prod_+ x_j\right)^{a_+}}{\left(\prod_- x_k\right)^{a_-}}, \quad i = 1, ..., D-1,$$

kde součiny $\prod_+$ a $\prod_-$ platí pro složky kódované jako $+1$ a -1 v i-tém kroku dělení.

4. Praktická část

V úvodní kapitole 4.1. si představíme data a provedeme několik grafických znázornění těchto dat; v kapitole 4.2. si vyjádříme modely pro oba přístupy a uvedeme vztahy mezi parametry těchto modelů. V kapitole 4.3. následně provedeme odhady jednotlivých parametrů a v kapitole 4.4. oba přístupy porovnáme. V kapitole 4.5. se pak ještě podíváme na výsledky po aplikaci užitých metod na jiných datech.

4.1. Představení dat

Data, na kterých budou dříve představené metody aplikovány, se týkají počtů hospitalizací cyklistů v důsledku dopravních nehod za 11 let, od roku 1999 do roku 2009, ve čtyřech věkových skupinách. Datový soubor je ještě rozdělen do dvou tabulek v závislosti na způsobu hospitalizace, máme tedy zvlášť tabulku pro ošetření hlavy a tabulku hospitalizace, kde jsou údaje všech hospitalizací bez ohledu na typ ošetření, viz tabulka 3.

Tabulka 3: Počty hospitalizovaných cyklistů za období 1999 – 2009 podle věku

Rok	0 až 14	15 až 17	18 až 26	27+
1999	2087	596	1196	3454
2000	2138	530	1280	3991
2001	1676	470	1011	3678
2002	1589	542	1127	3866
2003	1719	504	1179	4286
2004	1583	507	1072	4147
2005	1490	500	948	3871
2006	1122	380	770	3269
2007	1063	414	751	3466
2008	1046	366	745	3511
2009	946	402	669	3749

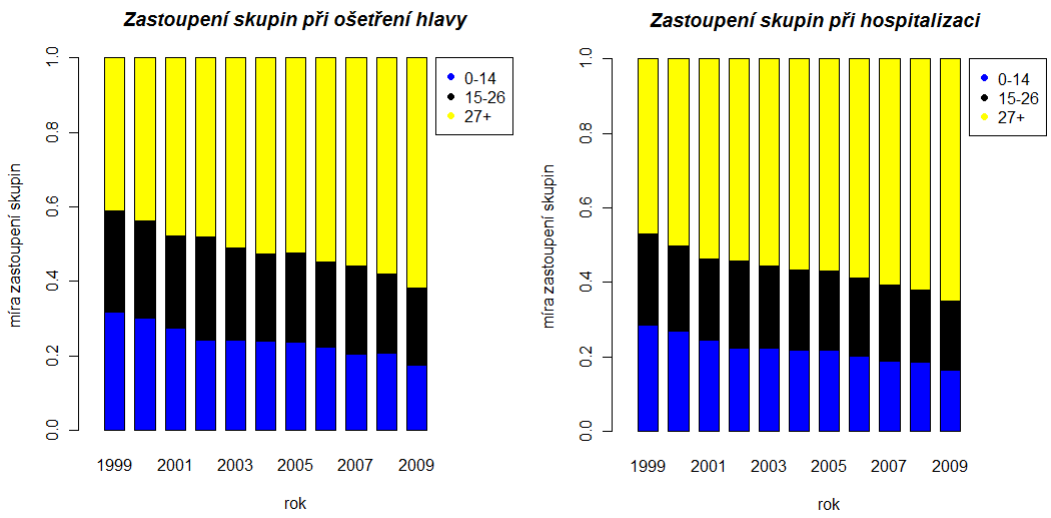
Kvůli malému věkovému rozsahu a malým četnostem druhé a třetí skupiny byly tyto skupiny spojeny do jedné. Ve všech případech tedy budeme uvažovat tři věkové skupiny cyklistů (0 až 14, 15 až 26, 27+).

Tabulka 4: Počty hospitalizovaných cyklistů za období 1999 – 2009 pro tři věkové skupiny

Rok	0 až 14	15 až 26	27+
1999	2087	1792	3454
2000	2138	1810	3991
2001	1676	1481	3678
2002	1589	1669	3866
2003	1719	1683	4286
2004	1583	1579	4147
2005	1490	1448	3871
2006	1122	1150	3269
2007	1063	1165	3466
2008	1046	1101	3511
2009	946	1071	3749

Pro lepší představu zastoupení jednotlivých skupin a toho, jak se zastoupení vyvíjela v jednotlivých letech, se teď podíváme na grafické znázornění zastoupení skupin při hospitalizaci a zastoupení skupin při ošetření hlavy.

Obrázek 3: Grafické znázornění zastoupení věkových skupin



Na první pohled můžeme vidět, že zastoupení jednotlivých věkových skupin je v obou případech velmi podobné. Z obou grafů je tak patrné, že zastoupení nejmladší skupiny v čase klesá, naopak zastoupení té nejstarší roste. Hlavním

důvodem může být, že obecně dochází ke stárnutí celé populace a tedy zastoupení nejstarší věkové skupiny v populaci je čím dál tím větší, naopak zastoupení nejmladší skupiny klesá. Odráží se to i v počtech nehod (a posléze ošetření) jednotlivých skupin.

Můžeme se podívat ještě na ternární diagram, který jsme si dříve představili jako dobrý grafický nástroj pro třísložková kompoziční data. My však zatím kompozice nemáme a tak musíme naše data upravit. Zatím budeme pracovat s datovým souborem hospitalizace. K převedení dat využijeme tohoto vztahu

$$f_i = \frac{n_i}{\sum_i n_i},$$

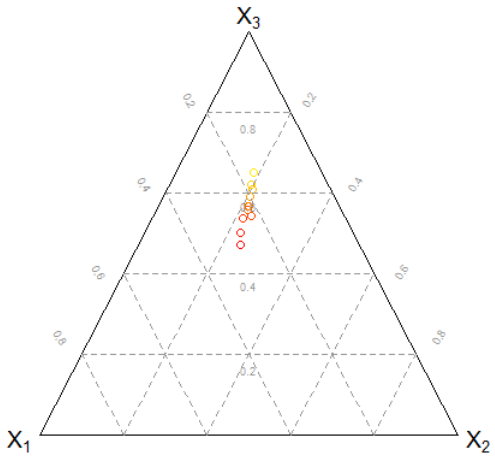
kde n_i značí počet nehod v i-té skupině a f_i relativní podíl i-té skupiny na celku. Tímto dostáváme pro každý rok třísložkovou kompozici s proporcionálními podíly na celku, kde součet všech složek kompozice je roven 1.

Tabulka 5: Třísložkové kompozice pro hospitalizace

Rok	0 až 14	15 až 26	27+
1999	0, 285	0, 244	0, 471
2000	0, 269	0, 228	0, 503
2001	0, 245	0, 217	0, 538
2002	0, 223	0, 234	0, 543
2003	0, 224	0, 219	0, 557
2004	0, 217	0, 216	0, 567
2005	0, 219	0, 213	0, 569
2006	0, 202	0, 208	0, 590
2007	0, 187	0, 205	0, 609
2008	0, 185	0, 195	0, 621
2009	0, 164	0, 186	0, 650

Těchto 11 kompozičních vektorů teď už můžeme graficky znázornit pomocí ternárního diagramu.

Obrázek 4: Ternární diagram pro hospitalizace (jednotlivé roky barevně odlišeny)



Připomeňme, že každá z kompozic je v grafu vyjádřena bodem, jehož vzdálenost od jednotlivých stran vyjadřuje velikosti jednotlivých složek. V grafu můžeme vidět, že jednotlivé body jsou odlišeny barevně od tmavě červené po světle žlutou. Tmavě červená značí rok 1999 a světle žlutá rok 2009. Z grafu tedy můžeme dobře vidět, jak právě zastoupení nejstarší skupiny v čase roste, neboť body se postupně vzdalují od strany x_3.

Ještě se podíváme na jeden způsob, jak data graficky znázornit, a to vykreslení souřadnic z_1, z_2, které získáme po ilr transformaci kompozic, čímž dojde ke snížení dimenze o jedna. Tyto souřadnice dostaneme snadno s využitím softwaru R:

```
> library('robCompositions')

> data = read.csv2('nehody.csv',header=TRUE)
> data = data.frame(data[,1:3],
+ apply(data[,4:5],1,sum), data[,6])
> names(data) = c('rok','typ','sk1','sk2','sk3')

> d1 = data[1:11,]
```

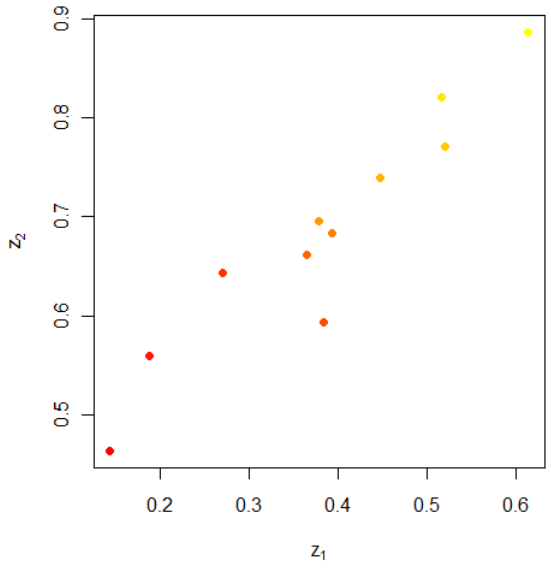
```
> prav = prop.table(as.table(as.matrix(d1[,3:5])),1)
> irld=ilr(prav)
> colnames(irld)=c('souradnice1','souradnice2')
```

Tabulka 6: Souřadnice z_1, z_2

z_1	z_2
$0,143$	$0,464$
$0,187$	$0,559$
$0,270$	$0,643$
$0,383$	$0,594$
$0,364$	$0,661$
$0,392$	$0,683$
$0,378$	$0,695$
$0,447$	$0,739$
$0,520$	$0,771$
$0,515$	$0,820$
$0,613$	$0,886$

Pokud na osu x vyneseme souřadnice z_1 a na osu y souřadnice z_2, dostaneme následný graf.

Obrázek 5: Vykreslení souřadnic z_1, z_2



Používáme zde stejného barevného kódování jako u ternárního diagramu. Můžeme tedy vidět, jak v čase obě souřadnice rostou. Na růst souřadnice z_1 má vliv především klesající zastoupení skupiny 1. Za růstem souřadnice z_2 pak stojí především rostoucí zastoupení skupiny 3.

4.2. Vyjádření modelů

Abychom mohli postupně přejít k porovnání obou přístupů, tak si nejdříve musíme určit konkrétní modely, které budeme porovnávat. Na jedné straně budeme mít ACL model a oproti němu model pro souřadnice ilr transformace definované binárním sekvenčním dělením.

Jako první si tedy uvedeme model ACL. Jak jsme si dříve ukázali, tento model porovnává vždy sousední kategorie. Uvažujeme 2 modely, kdy v prvním z nich budeme modelovat logaritmus podílu pravděpodobností 1. a 2. skupiny a v druhém modelu pak logaritmus podílu pravděpodobností 2. a 3. skupiny. Oba modely tedy můžeme vyjádřit jako

$$\log \frac{\pi_1(x)}{\pi_2(x)} = \alpha_1 + \beta_1 x,$$
$$\log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta_2 x,$$

kde x vyjadřuje jednotlivé roky. Ještě než přejdeme k druhému přístupu, tak si zde uvedeme interpretaci parametru β_1. Až totiž přejdeme k odhadu parametrů, tak tento samotný odhad nám toho bez potřebné interpretace moc neřekne. Uvažujme tedy dvě rovnice pro dva po sobě jdoucí roky.

$$\log \frac{\pi_1(x)}{\pi_2(x)} = \alpha_1 + \beta_1 x,$$
$$\log \frac{\pi_1(x+1)}{\pi_2(x+1)} = \alpha_1 + \beta_1 (x+1).$$

Obě rovnice od sebe odečteme a po úpravě se dostaneme až k interpretaci β_1

$$\log \frac{\pi_1(x+1)}{\pi_2(x+1)} - \log \frac{\pi_1(x)}{\pi_2(x)} = \beta_1(x+1) - \beta_1 x,$$
$$\log \frac{\pi_1(x+1)/\pi_2(x+1)}{\pi_1(x)/\pi_2(x)} = \beta_1,$$
$$\frac{\pi_1(x+1)/\pi_2(x+1)}{\pi_1(x)/\pi_2(x)} = e^{\beta_1}.$$

e^{β_1} tedy vyjadřuje změnu v podílu π_1/π_2, konkrétně kolikrát se tento podíl změní v následujícím roce oproti předchozímu. Interpretace parametru β_2 je podobná s tím rozdílem, že se vztahuje k podílu π_2/π_3.

Nyní můžeme přejít k druhému přístupu, kde budeme modelovat souřadnice z_1, z_2 získané z ilr transformace původních kompozic. Tyto souřadnice budeme chtít vyjádřit ve tvaru s využitím zastoupení jednotlivých skupin π_1, π_2, π_3, čehož docílíme pomocí binárního sekvenčního dělení. Postupné dělení skupin je uvedené v následující tabulce.

Tabulka 7: Kód určující sekvenční binární dělení třísložkové kompozice hospitalizací

Krok	x_1	x_2	x_3
1	$+1$	-1	-1
2	0	$+1$	-1

V prvním kroku dělení je tedy oddělena skupina $\{1\}$ od skupin $\{2, 3\}$. V druhém kroku se pak ještě osamostatní skupiny 2 a 3. Zaměřme se teď na první krok dělení, pro který platí, že

$$r = 1,$$
$$s = 2.$$

Hodnoty a_+ a a_- pak jsou

$$a_+ = \sqrt{\frac{s}{r(r+s)}} = \sqrt{\frac{2}{3}},$$
$$a_- = \sqrt{\frac{r}{s(r+s)}} = \sqrt{\frac{1}{6}}.$$

Souřadnici z_1 můžeme vyjádřit v následujícím tvaru

$$z_1 = \log \frac{(\pi_1)^{\sqrt{\frac{2}{3}}}}{(\pi_2\pi_3)^{\sqrt{\frac{1}{6}}}} = \sqrt{\frac{2}{3}} \log \pi_1 - \sqrt{\frac{2}{3}} \log(\pi_2\pi_3)^{\frac{1}{2}} = \sqrt{\frac{2}{3}} \log \frac{\pi_1}{\sqrt{\pi_2\pi_3}},$$

kde hodnoty π_1, π_2 a π_3 nám v kompozičním přístupu nepředstavují pravděpodobnosti, ale jednotlivé složky kompozic.

Přejděme teď ke druhému kroku dělení, abychom dostali i vyjádření pro druhou souřadnici. Pro druhý krok platí

$$r = 1,$$
$$s = 1.$$

A pro hodnoty a_+ a a_-

$$a_+ = \sqrt{\frac{s}{r(r+s)}} = \sqrt{\frac{1}{2}},$$
$$a_- = \sqrt{\frac{r}{s(r+s)}} = \sqrt{\frac{1}{2}}.$$

Souřadnici z_2 vyjádříme ve tvaru

$$z_2 = \log \frac{(\pi_2)^{\sqrt{\frac{1}{2}}}}{(\pi_3)^{\sqrt{\frac{1}{2}}}} = \sqrt{\frac{1}{2}} \log \frac{\pi_2}{\pi_3}.$$

Souřadnice získané tímto binárním dělením ještě přímo neodpovídají těm získaným z ilr transformace v softwaru R, liší se však pouze ve znaménku. Vynásobíme-li souřadnice získané dělením hodnotou -1, tak získáme stejné souřadnice jako po ilr transformaci. Dostáváme se tedy k vyjádření 2 modelů. Pro připomenutí a přehlednost uvádím na pravé straně také odpovídající ACL modely

$$z_1 = -\sqrt{\frac{2}{3}} \log \frac{\pi_1(x)}{\sqrt{\pi_2(x)\pi_3(x)}} = \alpha_1^* + \beta_1^* x \qquad \longleftrightarrow \qquad \log \frac{\pi_1(x)}{\pi_2(x)} = \alpha_1 + \beta_1 x,$$

$$z_2 = -\sqrt{\frac{1}{2}} \log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2^* + \beta_2^* x \qquad \longleftrightarrow \qquad \log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta_2 x.$$

Už na první pohled můžeme vidět spojitost mezi druhými modely, které se liší

pouze o konstantu $-\sqrt{\frac{1}{2}}$, kterou si označíme jako c_2. Hodnotu $-\sqrt{\frac{2}{3}}$ přítomnou v prvním kompozičním modelu si označíme jako c_1.

Než přejdeme k odhadům jednotlivých parametrů, tak si ještě vyjádříme vztahy mezi parametry. Začneme zde druhými modely, kde je vyjádření na první pohled patrné, a to

$$\alpha_2 = \frac{\alpha_2^*}{c_2},$$
$$\beta_2 = \frac{\beta_2^*}{c_2}.$$

Vyjádření pro α_1 a β_1 z prvního modelu je složitější. Dostaneme se k nim následnými úpravami

$$c_1 \log \frac{\pi_1}{\sqrt{\pi_2\pi_3}} = \alpha_1^* + \beta_1^* x =$$
$$= \log \frac{\pi_1}{\sqrt{\pi_2\pi_3}} = \frac{\alpha_1^*}{c_1} + \frac{\beta_1^*}{c_1} x =$$
$$= \frac{\pi_1}{\sqrt{\pi_2\pi_3}} = e^{\left(\frac{\alpha_1^*}{c_1} + \frac{\beta_1^*}{c_1} x\right)} =$$
$$= \frac{\pi_1^2}{\pi_2\pi_3} = e^{2\left(\frac{\alpha_1^*}{c_1} + \frac{\beta_1^*}{c_1} x\right)} =$$
$$= \left(\frac{\pi_1}{\pi_2}\right)^2 \cdot \frac{\pi_2}{\pi_3} = e^{2\left(\frac{\alpha_1^*}{c_1} + \frac{\beta_1^*}{c_1} x\right)} =$$
$$= 2 \log \frac{\pi_1}{\pi_2} + \log \frac{\pi_2}{\pi_3} = \frac{2\alpha_1^*}{c_1} + \frac{2\beta_1^*}{c_1} x =$$
$$= 2\alpha_1 + 2\beta_1 x + \alpha_2 + \beta_2 x = \frac{2\alpha_1^*}{c_1} + \frac{2\beta_1^*}{c_1} x =$$
$$= 2\alpha_1 + 2\beta_1 x + \frac{\alpha_2^*}{c_2} + \frac{\beta_2^*}{c_2} x = \frac{2\alpha_1^*}{c_1} + \frac{2\beta_1^*}{c_1} x.$$

Odtud už můžeme vidět, že

$$\alpha_1 = \frac{\alpha_1^*}{c_1} - \frac{\alpha_2^*}{2c_2},$$
$$\beta_1 = \frac{\beta_1^*}{c_1} - \frac{\beta_2^*}{2c_2}.$$

Pro přehlednost jsou dříve odvozené vztahy mezi parametry uvedeny v následující tabulce.

Tabulka 8: Vztahy mezi parametry ACL modelu a kompozičního modelu

$\alpha_1 = \frac{\alpha_1^*}{c_1} - \frac{\alpha_2^*}{2c_2}$	$\alpha_2 = \frac{\alpha_2^*}{c_2}$
$\beta_1 = \frac{\beta_1^*}{c_1} - \frac{\beta_2^*}{2c_2}$	$\beta_2 = \frac{\beta_2^*}{c_2}$

4.3. Odhady parametrů

V této části provedeme s využitím softwaru R odhady jednotlivých parametrů dříve uvedených modelů. Začněme prvním z ACL modelů

$$\log \frac{\pi_1(x)}{\pi_2(x)} = \alpha_1 + \beta_1 x.$$

Po odhadu se dostáváme k rovnici

$$\log \frac{\pi_1(x)}{\pi_2(x)} = 53,176 - 0,027x.$$

Odhady obou parametrů jsou významně nenulové a směrodatné odchylky těchto odhadů jsou

$$s(\widehat{\alpha_1}) = 7,239,$$
$$s(\widehat{\beta_1}) = 0,004.$$

Dříve jsme si uvedli interpretaci parametru β v ACL modelu, tak se podíváme, co nám zde parametr β_1 říká

$$e^{\beta_1} = e^{-0,027} = 0,974.$$

Hodnota $0,974$ znamená, že podíl π_1/π_2 v čase mírně klesá a to asi o $2,5$ % za rok, zastoupení druhé skupiny oproti první se tedy v čase významně zvyšuje.

Přejděme k druhému ACL modelu, který jsme si dříve uvedli v tomto tvaru

$$\log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta_2 x.$$

Po odhadu parametrů se pak dostáváme k rovnici

$$\log \frac{\pi_2(x)}{\pi_3(x)} = 98,502 - 0,050x.$$

Oba odhady jsou i v tomto případě významně nenulové se směrodatnými odchylkami

$$s(\widehat{\alpha_2}) = 6,053,$$
$$s(\widehat{\beta_2}) = 0,003.$$

Také zde se podíváme na interpretaci odhadnutého parametru β_2

$$e^{\beta_2} = e^{-0,050} = 0,952.$$

Podíl π_2/π_3 v čase také mírně klesá a to asi o 5% za rok. To, že oba podíly se v čase zmenšují, odpovídá tomu, co bylo vidět už na obrázku 3, kde jsme si řekli, že zastoupení nejmladší skupiny klesá na úkor skupiny nejstarší (prostřední skupina zůstává téměř beze změny). Pokud tedy dochází k zmenšování nejmladší skupiny při téměř neměnných hodnotách skupiny prostřední, tak vlastně dochází k zmenšování podílu π_1/π_2. To samé platí i pro podíly π_2/π_3, neboť dochází k růstu zastoupení nejstarší skupiny při stejných hodnotách zastoupení prostřední skupiny.

Přejdeme teď ke kompozičním modelům. Začneme prvním modelem, který je

$$z_1 = c_1 \log \frac{\pi_1(x)}{\sqrt{\pi_2(x)\pi_3(x)}} = \alpha_1^* + \beta_1^* x.$$

Po odhadnutí parametrů dostaneme rovnici

$$z_1 = c_1 \log \frac{\pi_1(x)}{\sqrt{\pi_2(x)\pi_3(x)}} = -82,514 + 0,041x.$$

Oba parametry jsou významně nenulové. Směrodatné odchylky odhadů jsou

$$s(\widehat{\alpha_1^*}) = 7,507,$$
$$s(\widehat{\beta_1^*}) = 0,004.$$

Co se týče druhého kompozičního modelu, tak ten jsme si dříve vyjádřili jako

$$z_2 = c_2 \log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2^* + \beta_2^* x.$$

Odhadnuté parametry jsou

$$z_2 = c_2 \log \frac{\pi_2(x)}{\pi_3(x)} = -69,644 + 0,035x.$$

Oba parametry jsou i v tomto případě významně nenulové a směrodatné odchylky odhadů jsou

$$s(\widehat{\alpha_2^*}) = 5,967,$$
$$s(\widehat{\beta_2^*}) = 0,003.$$

Pro přehlednost si ještě uvedeme všechny odhady i s jejich přesností pohromadě.

ACL1: $\log \frac{\pi_1(x)}{\pi_2(x)} = 53,176(7,239) - 0,027x(0,004)$

ACL2: $\log \frac{\pi_2(x)}{\pi_3(x)} = 98,502(6,053) - 0,050x(0,003)$

KOMP1: $z_1 = c_1 \log \frac{\pi_1(x)}{\sqrt{\pi_2(x)\pi_3(x)}} = -82,514(7,507) + 0,041x(0,004)$

KOMP2: $z_2 = c_2 \log \frac{\pi_2(x)}{\pi_3(x)} = -69,644(5,967) + 0,035x(0,003)$

Všechny uvedené směrodatné odchylky jsou zaokrouhleny na 3 desetinná místa, v následující kapitole však budeme počítat s nezaokrouhlenými hodnotami.

4.4. Porovnání modelů

V této kapitole budeme chtít oba modely porovnat v závislosti na přesnosti odhadů parametrů β. K tomu využijeme vztahy mezi parametry uvedené v tabulce 8. Začněme porovnáním modelů ACL1 a KOMP1

$$\log \frac{\pi_1(x)}{\pi_2(x)} = 53,176 - 0,027x,$$

$$z_1 = c_1 \log \frac{\pi_1(x)}{\sqrt{\pi_2(x)\pi_3(x)}} = -82,514 + 0,041x.$$

Připomeňme si ještě, jak můžeme parametr β_1 vyjádřit pomocí parametrů z kompozičních modelů β_1^* a β_2^*

$$\beta_1 = \frac{\beta_1^*}{c_1} - \frac{\beta_2^*}{2c_2}.$$

A v následující tabulce si ukažme bodové odhady parametrů $\widehat{\beta_1}$, $\widehat{\beta_2}$ a jejich odpovídajících vyjádření.

Tabulka 9: Bodové odhady parametrů

$\widehat{\beta_1} = -0,0265$	$\widehat{\beta_2} = -0,0496$
$\dfrac{\widehat{\beta_1^*}}{c_1} - \dfrac{\widehat{\beta_2^*}}{2c_2} = -0,0258$	$\dfrac{\widehat{\beta_2^*}}{c_2} = -0,0496$

Cílem bude vypočítat rozptyly pro $\widehat{\beta_1}$ a jeho vyjádření $\left(\dfrac{\widehat{\beta_1^*}}{c_1} - \dfrac{\widehat{\beta_2^*}}{2c_2}\right)$, ty následně porovnat a rozhodnout, který přístup podává přesnější odhady.

Rozptyl pro $\widehat{\beta_1}$ vypočítáme snadno pouhým umocněním směrodatné odchylky na druhou

$$\widehat{var}(\widehat{\beta_1}) = s^2(\widehat{\beta_1}) = 1,306 \cdot 10^{-5}.$$

Výpočet rozptylu pro vyjádření $\left(\dfrac{\widehat{\beta_1^*}}{c_1} - \dfrac{\widehat{\beta_2^*}}{2c_2}\right)$ je o něco složitější a využijeme při něm známého vztahu pro počítání s rozptyly

$$var(aX + bY) = var(aX) + var(bY) + 2cov(aX, bY) =$$
$$a^2 var(X) + b^2 var(Y) + 2ab\ cov(X, Y).$$

S využitím tohoto vztahu se tedy dostáváme k následnému výpočtu, ve kterém můžeme kovarianci zanedbat, neboť jde o parametry dvou rozdílných modelů

$$\widehat{var}\left(\frac{\widehat{\beta_1^*}}{c_1} - \frac{\widehat{\beta_2^*}}{2c_2}\right) =$$
$$= \widehat{var}\left(\frac{\widehat{\beta_1^*}}{c_1}\right) + \widehat{var}\left(\frac{\widehat{\beta_2^*}}{2c_2}\right) =$$
$$= \frac{1}{c_1^2}\widehat{var}\left(\widehat{\beta_1^*}\right) + \frac{1}{4c_2^2}\widehat{var}\left(\widehat{\beta_2^*}\right) =$$
$$= \frac{3}{2}\widehat{var}\left(\widehat{\beta_1^*}\right) + \frac{1}{2}\widehat{var}\left(\widehat{\beta_2^*}\right) =$$
$$= \widehat{var}\left(\frac{\widehat{\beta_1^*}}{c_1} - \frac{\widehat{\beta_2^*}}{2c_2}\right) = 2,548 \cdot 10^{-5}.$$

Pokud porovnáme oba vypočtené rozptyly, vidíme, že hodnota rozptylu získaná z ACL modelu je téměř dvakrát tak menší oproti hodnotě rozptylu získané kompozičním přístupem.

Přejděme k druhým modelům ACL2 a KOMP2

$$\log \frac{\pi_2(x)}{\pi_3(x)} = 98,502 - 0,050x,$$
$$z_2 = c_2 \log \frac{\pi_2(x)}{\pi_3(x)} = -69,644 + 0,035x.$$

A připomeňme si ještě vztah mezi parametry

$$\beta_2 = \frac{\beta_2^*}{c_2}.$$

Výpočet rozptylu β_2 je opět snadný

$$\widehat{var}(\widehat{\beta_2}) = s^2(\widehat{\beta_2}) = 0,913 \cdot 10^{-5}.$$

Výpočet rozptylu pro vyjádření bude v tomto případě také snadný

$$\widehat{var}\left(\frac{\widehat{\beta_2^*}}{c_2}\right) = \frac{1}{c_2^2}\widehat{var}(\widehat{\beta_2^*}) = \frac{1}{\left(-\sqrt{\frac{1}{2}}\right)^2}\widehat{var}(\widehat{\beta_2^*}) = \frac{1}{\frac{1}{2}}\widehat{var}(\widehat{\beta_2^*}) = 2\widehat{var}(\widehat{\beta_2^*}) = \\ = 1,774 \cdot 10^{-5}.$$

I pro případ druhých modelů se nám dostalo podobného výsledku, a to, že rozptyl pro ACL model je téměř dvakrát tak menší než rozptyl pro kompoziční přístup.

4.5. Data pro ošetření hlavy

Doposud jsme pracovali jen s daty týkajícími se celkového počtu hospitalizovaných. Námi analyzovaná datová množina však obsahuje i informace o zastoupení skupin pouze při ošetření hlavy. V této kapitole tedy aplikujeme oba přístupy na nová data a podíváme se, zda se výsledky nějak liší. Vzhledem k podobnému zastoupení skupin, které je vidět na obrázku 3, se však dá očekávat, že dostaneme podobné výsledky jako v předchozím případě.

Vše potřebné máme připravené a můžeme přejít rovnou k odhadu parametrů. Pro ACL modely dostáváme tyto odhady

ACL1: $\log \dfrac{\pi_1(x)}{\pi_2(x)} = 52,892 - 0,026x,$

ACL2: $\log \dfrac{\pi_2(x)}{\pi_3(x)} = 121,799 - 0,061x.$

Pro kompoziční modely se pak dostaneme k následujícím odhadům

KOMP1: $z_1 = c_1 \log \dfrac{\pi_1(x)}{\sqrt{\pi_2(x)\pi_3(x)}} = -89,842 + 0,045x,$

KOMP2: $z_2 = c_2 \log \dfrac{\pi_2(x)}{\pi_3(x)} = -86,309 + 0,043x.$

Co se týče interpretace parametru β pro modely ACL1 a ACL2, tak se dostáváme k velmi podobným výsledkům jako v předchozím případě, neboť hodnoty

$$e^{\beta_1} = 0,974,$$
$$e^{\beta_2} = 0,941,$$

jsou téměř stejné jako v případě dat pro hospitalizace. I zde oba podíly π_1/π_2 a π_2/π_3 v čase klesají. První podíl zhruba o $2,5$ % za rok a druhý podíl zhruba o 6 % za rok.

Nyní opět vypočítáme rozptyly a porovnáme je. Pro ACL1 a KOMP1 dostaneme tyto rozptyly

$$\widehat{var}(\widehat{\beta}_1) = 2,490 \cdot 10^{-5},$$

$$\widehat{var}\left(\frac{\widehat{\beta_1^*}}{c_1} - \frac{\widehat{\beta_2^*}}{2c_2}\right) = \frac{3}{2}\widehat{var}\left(\widehat{\beta_1^*}\right) + \frac{1}{2}\widehat{var}\left(\widehat{\beta_2^*}\right) = 4,828 \cdot 10^{-5}.$$

A pro druhé modely ACL2 a KOMP2 dostáváme rozptyly

$$\widehat{var}(\widehat{\beta}_2) = 1,842 \cdot 10^{-5},$$

$$\widehat{var}\left(\frac{\widehat{\beta_2^*}}{c_2}\right) = 2\widehat{var}(\widehat{\beta_2^*}) = 2,738 \cdot 10^{-5}.$$

Pro první modely dostáváme stejný výsledek jako v předchozím případě, a to, že ACL modely dávají téměř dvakrát tak přesné odhady. Pro druhé modely je rozdíl v přesnosti o něco menší, ACL2 model je v tomto případě zhruba 1,5-krát přesnější.

4.6. Odhad varianční matice ACL modelu

Vraťme se k původním datům celkové hospitalizace. Rozptyly odhadů jednotlivých parametrů pomocí vlastností maximálně věrohodného odhadu, pro který platí

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mle} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}^{-1}).$$

Tato vlastnost nám říká, že maximálně věrohodný odhad parametru $\boldsymbol{\theta}$ má asymptoticky normální rozdělení, je asymptoticky nevychýlený a varianční matice je rovna $\boldsymbol{I}^{-1}$. Matice $\boldsymbol{I}$ zde představuje Fisherovu informační matici, pro jejíž prvky platí

$$\boldsymbol{I}(\theta)_{i,j} = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\boldsymbol{Y};\boldsymbol{\theta})\right].$$

Abychom mohli určit Fisherovu informační matici, tak nejdříve musíme dostat věrohodnostní funkci pro náš případ. Uvažujme tedy již dříve představené ACL modely

$$\log\frac{\pi_1}{\pi_2} = \alpha_1 + \beta_1 x,$$
$$\log\frac{\pi_2}{\pi_3} = \alpha_2 + \beta_2 x.$$

Data pochází z multinomického rozdělení a pravděpodobnostní funkci máme tedy v tomto tvaru

$$\mathrm{C}\prod_{i=1}^{3} \pi_i^{y_i} = \mathrm{C}\pi_1^{y_1}\pi_2^{y_2}\pi_3^{y_3},$$

kde

$$\mathrm{C} = \frac{n!}{y_1!y_2!y_3!}, \quad n = \sum_{i=1}^{3} y_i.$$

Pravděpodobnostní funkci ještě upravíme

$$\mathrm{C}\pi_1^{y_1}\pi_2^{y_2}\pi_3^{y_3} = \mathrm{C}\left(\frac{\pi_1}{\pi_2}\right)^{y_1}\left(\frac{\pi_2}{\pi_3}\right)^{y_1+y_2}\pi_3^{y_1+y_2+y_3},$$

a zlogaritmujeme

$$\log\mathrm{C} + y_1\log\frac{\pi_1}{\pi_2} + (y_1+y_2)\log\frac{\pi_2}{\pi_3} + n\log\pi_3.$$

Čemu jsou rovny $\log\frac{\pi_1}{\pi_2}$ a $\log\frac{\pi_2}{\pi_3}$ víme z tvaru modelů. Nevíme však, čemu se rovná π_3. Využijeme toho, že $\pi_3 = 1-\pi_1-\pi_2$ a po několika krocích se dostaneme k vhodnému vyjádření π_3

$$\frac{\pi_2}{1-\pi_1-\pi_2} = e^{\alpha_2+\beta_2 x},$$
$$\frac{\pi_1}{1-\pi_1-\pi_2} = \frac{\pi_1}{\pi_2}\frac{\pi_2}{1-\pi_1-\pi_2} = e^{\alpha_1+\alpha_2+(\beta_1+\beta_2)x}.$$

A dosadíme do následujícího výrazu

$$\frac{1}{1-\pi_1-\pi_2} - \frac{\pi_1}{1-\pi_1-\pi_2} - \frac{\pi_2}{1-\pi_1-\pi_2} = \frac{1-\pi_1-\pi_2}{1-\pi_1-\pi_2} = 1 =$$
$$= \frac{1}{1-\pi_1-\pi_2} - e^{\alpha_1+\alpha_2+(\beta_1+\beta_2)x} - e^{\alpha_2+\beta_2 x}.$$

Odtud se už snadno dostaneme k vyjádření $\log \pi_3$

$$\frac{1}{1-\pi_1-\pi_2} = 1 + e^{\alpha_1+\alpha_2+(\beta_1+\beta_2)x} + e^{\alpha_2+\beta_2 x},$$
$$\log(1-\pi_1-\pi_2) = \log \pi_3 = -\log(1 + e^{\alpha_1+\alpha_2+(\beta_1+\beta_2)x} + e^{\alpha_2+\beta_2 x}).$$

Dosadíme do logaritmické věrohodnostní funkce pro jedno pozorování

$$\log \mathrm{C} + y_1(\alpha_1+\beta_1 x) + (y_1+y_2)(\alpha_2+\beta_2 x) - n \log\left(1 + e^{\alpha_1+\alpha_2+(\beta_1+\beta_2)x} + e^{\alpha_2+\beta_2 x}\right),$$

a vyjádříme celkovou logaritmickou věrohodnostní funkci

$$\sum_{i=1}^{11} \log \mathrm{C} + y_{1i}(\alpha_1+\beta_1 x_i) + (y_{1i}+y_{2i})(\alpha_2+\beta_2 x_i) -$$
$$n_i \log\left(1 + e^{\alpha_1+\alpha_2+(\beta_1+\beta_2)x_i} + e^{\alpha_2+\beta_2 x_i}\right).$$

S takto vyjádřenou logaritmickou věrohodnostní funkcí už můžeme přejít k výpočtu Fisherovy informační matice. Označme

$$a_i = \alpha_1 + \alpha_2 + \beta_1 x_i + \beta_2 x_i,$$
$$b_i = \alpha_2 + \beta_2 x_i,$$
$$c_i = a_i + b_i = \alpha_1 + 2\alpha_2 + \beta_1 x_i + 2\beta_2 x_i,$$

a vypočteme Fisherovu informační matici

$$\widehat{\boldsymbol{I}} = \begin{pmatrix} \sum n_i \frac{e^{a_i}+e^{c_i}}{(1+e^{a_i}+e^{b_i})^2} & \sum n_i \frac{e^{a_i}}{(1+e^{a_i}+e^{b_i})^2} & \sum n_i \frac{x_i(e^{a_i}+e^{c_i})}{(1+e^{a_i}+e^{b_i})^2} & \sum n_i \frac{x_i e^{a_i}}{(1+e^{a_i}+e^{b_i})^2} \\ & \sum n_i \frac{e^{a_i}+e^{b_i}}{(1+e^{a_i}+e^{b_i})^2} & \sum n_i \frac{x_i e^{a_i}}{(1+e^{a_i}+e^{b_i})^2} & \sum n_i \frac{x_i(e^{a_i}+e^{b_i})}{(1+e^{a_i}+e^{b_i})^2} \\ & & \sum n_i \frac{x_i^2(e^{a_i}+e^{c_i})}{(1+e^{a_i}+e^{b_i})^2} & \sum n_i \frac{x_i^2 e^{a_i}}{(1+e^{a_i}+e^{b_i})^2} \\ & & & \sum n_i \frac{x_i^2(e^{a_i}+e^{b_i})}{(1+e^{a_i}+e^{b_i})^2} \end{pmatrix}.$$

Matice je symetrická, proto stačí doplnit horní trojúhelník. Po dosazení odhadů parametrů a inverzi této matice dostáváme varianční matici ACL modelu

$$\widehat{var}\begin{pmatrix}\widehat{\alpha_1}\\ \widehat{\alpha_2}\\ \widehat{\beta_1}\\ \widehat{\beta_2}\end{pmatrix} = \begin{pmatrix} 52,573 & -26,332 & -2,624\cdot 10^{-2} & 1,314\cdot 10^{-2}\\ & 36,492 & 1,314\cdot 10^{-2} & -1,821\cdot 10^{-2}\\ & & 1,310\cdot 10^{-5} & -6,560\cdot 10^{-6}\\ & & & 9,090\cdot 10^{-6}\end{pmatrix}.$$

Můžeme vidět, že při "ručním" výpočtu varianční matice dostáváme téměř shodné rozptyly, jako ty, které jsme získali odhadem v softwaru R, který pro odhad parametrů ACL modelu také využívá metody maximální věrohodnosti

$$\widehat{var}(\widehat{\beta_1}) = 1,306\cdot 10^{-5} \doteq 1,310\cdot 10^{-5}\ ,$$
$$\widehat{var}(\widehat{\beta_2}) = 0,913\cdot 10^{-5} \doteq 0,909\cdot 10^{-5}.$$

4.7. Odhad varianční matice kompozičního modelu

Přejděme ke kompozičnímu přístupu, modely máme v nám již známém tvaru

$$c_1 \log\frac{\pi_1}{\sqrt{\pi_2\pi_3}} = \alpha_1^* + \beta_1^* x,$$
$$c_2 \log\frac{\pi_2}{\pi_3} = \alpha_2^* + \beta_2^* x,$$

kde

$$c_1 = -\sqrt{\frac{2}{3}},$$
$$c_2 = -\sqrt{\frac{1}{2}}.$$

Určíme si pravděpodobnostní funkci

$$\mathrm{C}\pi_1^{y_1}\pi_2^{y_2}\pi_3^{y_3} = \mathrm{C}\left(\frac{\pi_1}{\sqrt{\pi_2\pi_3}}\right)^{y_1}\left(\frac{\pi_2}{\pi_3}\right)^{\frac{y_1}{2}+y_2}\pi_3^{\frac{y_1}{2}+y_2+\frac{y_1}{2}+y_3},$$

a zlogaritmujeme ji

$$\log \mathrm{C} + y_1 \log \frac{\pi_1}{\sqrt{\pi_2 \pi_3}} + (\tfrac{y_1}{2} + y_2) \log \frac{\pi_2}{\pi_3} + n \log \pi_3.$$

Opět se dostáváme k problému, jak vyjádřit π_3. Víme, že

$$\frac{\pi_2}{1 - \pi_1 - \pi_2} = e^{\frac{1}{c_2}(\alpha_2^* + \beta_2^* x)},$$

ale zatím nevíme, čemu je rovno $\dfrac{\pi_1}{1 - \pi_1 - \pi_2}$. Upravíme vyjádření našich modelů

$$\log \frac{\pi_1^2}{\pi_2 \pi_3} = \tfrac{2}{c_1}(\alpha_1^* + \beta_1^* x),$$
$$\log \frac{\pi_2}{\pi_3} = \tfrac{1}{c_2}(\alpha_2^* + \beta_2^* x),$$

obě rovnice sečteme a dostaneme se k výrazu

$$\log \frac{\pi_1^2}{\pi_2 \pi_3} \cdot \frac{\pi_2}{\pi_3} = \tfrac{2}{c_1}(\alpha_1^* + \beta_1^* x) + \tfrac{1}{c_2}(\alpha_2^* + \beta_2^* x) =$$
$$= \log \frac{\pi_1}{\pi_3} = \tfrac{1}{c_1}(\alpha_1^* + \beta_1^* x) + \tfrac{1}{2c_2}(\alpha_2^* + \beta_2^* x).$$

Teď už můžeme dosadit do následujícího výrazu

$$\frac{1}{1 - \pi_1 - \pi_2} - \frac{\pi_1}{1 - \pi_1 - \pi_2} - \frac{\pi_2}{1 - \pi_1 - \pi_2} = \frac{1 - \pi_1 - \pi_2}{1 - \pi_1 - \pi_2} = 1 =$$
$$= \frac{1}{1 - \pi_1 - \pi_2} - e^{\frac{1}{c_1}(\alpha_1^* + \beta_1^* x) + \frac{1}{2c_2}(\alpha_2^* + \beta_2^* x)} - e^{\frac{1}{c_2}(\alpha_2^* + \beta_2^* x)},$$

a vyjádřit $\log \pi_3$

$$\frac{1}{1 - \pi_1 - \pi_2} = 1 + e^{\frac{1}{c_1}(\alpha_1^* + \beta_1^* x) + \frac{1}{2c_2}(\alpha_2^* + \beta_2^* x)} + e^{\frac{1}{c_2}(\alpha_2^* + \beta_2^* x)},$$
$$\log(1 - \pi_1 - \pi_2) = \log \pi_3 = -\log(1 + e^{\frac{1}{c_1}(\alpha_1^* + \beta_1^* x) + \frac{1}{2c_2}(\alpha_2^* + \beta_2^* x)} + e^{\frac{1}{c_2}(\alpha_2^* + \beta_2^* x)}.$$

Dosadíme do logaritmické věrohodnostní funkce pro jedno pozorování

$$\log \mathrm{C} + \tfrac{y_1}{c_1}(\alpha_1^* + \beta_1^* x) + \tfrac{\frac{y_1}{2} + y_2}{c_2}(\alpha_2^* + \beta_2^* x) -$$
$$n \log \left(1 + e^{\frac{1}{c_1}(\alpha_1^* + \beta_1^* x) + \frac{1}{2c_2}(\alpha_2^* + \beta_2^* x)} + e^{\frac{1}{c_2}(\alpha_2^* + \beta_2^* x)}\right),$$

a vyjádříme celkovou logaritmickou věrohodnostní funkci

$$\sum_{i=1}^{11} \log \mathrm{C} + \tfrac{y_{1i}}{c_1}(\alpha_1^* + \beta_1^* x_i) + \tfrac{\frac{y_{1i}}{2}+y_{2i}}{c_2}(\alpha_2^* + \beta_2^* x_i) - n_i \log\left(1 + e^{\frac{1}{c_1}(\alpha_1^*+\beta_1^* x_i)+\frac{1}{2c_2}(\alpha_2^*+\beta_2^* x_i)} + e^{\frac{1}{c_2}(\alpha_2^*+\beta_2^* x_i)}\right).$$

Nyní můžeme vypočítat Fisherovu informační matici. Označme

$$a_i = \tfrac{1}{c_1}(\alpha_1^* + \beta_1^* x_i) + \tfrac{1}{2c_2}(\alpha_2^* + \beta_2^* x_i),$$
$$b_i = \tfrac{1}{c_2}(\alpha_2^* + \beta_2^* x_i),$$
$$c_i = a_i + b_i.$$

Fisherovu informační matici pak máme ve tvaru

$$\widehat{\boldsymbol{I}}^* = = \begin{pmatrix} \Sigma \frac{n_i}{c_1^2}\frac{e^{a_i}+e^{c_i}}{(1+e^{a_i}+e^{b_i})^2} & \Sigma \frac{n_i}{2c_1c_2}\frac{e^{a_i}-e^{c_i}}{(1+e^{a_i}+e^{b_i})^2} & \Sigma \frac{n_i}{c_1^2}\frac{x_i(e^{a_i}+e^{c_i})}{(1+e^{a_i}+e^{b_i})^2} & \Sigma \frac{n_i}{2c_1c_2}\frac{x_i e^{a_i}-e^{c_i}}{(1+e^{a_i}+e^{b_i})^2} \\ & \Sigma \frac{n_i}{4c_2^2}\frac{e^{a_i}+e^{c_i}+4e^{b_i}}{(1+e^{a_i}+e^{b_i})^2} & \Sigma \frac{n_i}{2c_1c_2}\frac{x_i e^{a_i}-e^{c_i}}{(1+e^{a_i}+e^{b_i})^2} & \Sigma \frac{n_i}{4c_2^2}\frac{x_i(e^{a_i}+e^{c_i}+4e^{b_i})}{(1+e^{a_i}+e^{b_i})^2} \\ & & \Sigma \frac{n_i}{c_1^2}\frac{x_i^2(e^{a_i}+e^{c_i})}{(1+e^{a_i}+e^{b_i})^2} & \Sigma \frac{n_i}{2c_1c_2}\frac{x_i^2 e^{a_i}-e^{c_i}}{(1+e^{a_i}+e^{b_i})^2} \\ & & & \Sigma \frac{n_i}{4c_2^2}\frac{x_i^2(e^{a_i}+e^{c_i}+4e^{b_i})}{(1+e^{a_i}+e^{b_i})^2} \end{pmatrix}.$$

Po inverzi a dosazení odhadů parametrů dostáváme varianční matici

$$\widehat{var}\begin{pmatrix} \widehat{\alpha_1^*} \\ \widehat{\alpha_2^*} \\ \widehat{\beta_1^*} \\ \widehat{\beta_2^*} \end{pmatrix} = \begin{pmatrix} 23,568 & -4,662 & -1,176\cdot 10^{-2} & 2,327\cdot 10^{-2} \\ & 18,235 & 2,327\cdot 10^{-3} & -9,101\cdot 10^{-3} \\ & & 5,873\cdot 10^{-6} & -1,162\cdot 10^{-6} \\ & & & 4,542\cdot 10^{-6} \end{pmatrix}.$$

Při "ručním" výpočtu varianční matice dostáváme velmi rozdílné hodnoty rozptylů oproti těm získaným softwarem R. Důvodem je, že zatímco zde jsme využili metody maximální věrohodnosti využívající multinomického rozdělení, tak v softwaru R se parametry odhadovaly metodou nejmenších čtverců, která vychází z jiného předpokladu na rozdělení

$$\widehat{var}(\widehat{\beta_1^*}) = 1,403\cdot 10^{-5} \neq 5,873\cdot 10^{-6}\ ,$$
$$\widehat{var}(\widehat{\beta_2^*}) = 8,868\cdot 10^{-6} \neq 4,542\cdot 10^{-6}.$$

Vzhledem k tomu, že jsme tímto způsobem dostali rozdílné rozptyly, tak se vrátíme k tabulce 8, kde máme vyjádřeny vztahy mezi parametry a rozptyly přepočítáme a porovnáme. Rozptyly pro ACL modely zůstávají téměř stejné

$$\widehat{var}(\widehat{\beta}_1) = 1,310 \cdot 10^{-5},$$
$$\widehat{var}(\widehat{\beta}_2) = 0,909 \cdot 10^{-5}.$$

Rozptyly pro kompoziční modely přepočítáme pomocí nám známých vztahů

$$\widehat{var}\left(\frac{\widehat{\beta_1^*}}{c_1} - \frac{\widehat{\beta_2^*}}{2c_2}\right) = \frac{3}{2}\widehat{var}\left(\widehat{\beta_1^*}\right) + \frac{1}{2}\widehat{var}\left(\widehat{\beta_2^*}\right) - \sqrt{3}cov(\beta_1^*, \beta_2^*) = 1,310 \cdot 10^{-5},$$
$$\widehat{var}\left(\frac{\widehat{\beta_2^*}}{c_2}\right) = 2\widehat{var}(\widehat{\beta_2^*}) = 0,908 \cdot 10^{-5}.$$

Můžeme vidět, že pro ACL model i kompoziční přístup zde dostáváme téměř shodné rozptyly.

Závěr

Ve své práci jsem se snažil o nastudovaní a pochopení dvou různých způsobů analýzy kategoriálních dat, a to zobecněného lineárního modelu a v poslední době velmi oblíbeného kompozičního přístupu. To vše za účelem následné aplikace těchto metod na reálná data o nehodách cyklistů. Hlavním cílem práce pak bylo tyto metody na základě dosažených výsledků porovnat.

Úvodní tři kapitoly byly věnovány položení teoretických základů jednotlivých metod pro analýzu kategoriálních dat. Seznámili jsme se s klasickým a zobecněným lineárním modelem, včetně metody maximální věrohodnosti, používané pro odhad parametrů. Dále bylo uvedeno několik různých typů logitových modelů pro multinominální data. Jedním z představených modelů byl model logitů sousedních kategorií (ACL), který v praktické části sloužil jako jeden z modelů pro porovnání. Z tohoto důvodu byl uveden i vzorový příklad včetně kódu ze softwaru R, aby čtenář měl názornou ukázku, jak tento model funguje. V závěru teoretické části byly uvedeny základy pro pochopení kompozičního přístupu analýzy dat. Právě s kompozičním modelem byl ACL model porovnáván.

Stěžejní čtvrtá kapitola byla věnována praktickému příkladu. V jejím úvodu bylo potřeba určit podobu modelů, které budou porovnávány. Na jedné straně jsme měli dva ACL modely a oproti nim příslušné kompoziční modely. Po vyjádření vztahů mezi jednotlivými parametry a jejich následném odhadu s využitím softwaru R, jsme přešli až k výpočtu rozptylů odhadů parametrů. Vypočtené rozptyly ukazovaly, že lepší výsledky podává ACL model, což mě celkem překvapilo, protože jsem očekával, že obě metody budou podávat zhruba stejně dobré výsledky. Oba přístupy jsme pak aplikovali ještě na jiný datový soubor, došli jsme však velmi podobným výsledkům. V závěrečných podkapitolách praktické části jsme ještě "ručně" odhadli varianční matice obou přístupů.

Věřím, že práce povede čtenáře k zamyšlení a základnímu pochopení principů práce s kategoriálními daty. Studium jednotlivých metod a především pak jejich aplikace na reálná data pro mě byly jednoznačně přínosem.

Literatura

[1] Agresti, A.: *Categorical Data Analysis*, John Wiley & Sons, New Jersey, 2002.

[2] Agresti, A.: *Examples of using R for modeling ordinal data*, John Wiley & Sons, New Jersey, 2002.

[3] Aitchison, J.: *The statistical analysis of compositional data*, Chapman & Hall, London, 1986

[4] Anděl, J.: *Základy matematické statistiky*, MATFYZPRESS, Praha, 2005.

[5] Brodinová, Š.: *Diskriminační analýza pro kompoziční data*, bakalářská práce, Přírodovědecká fakulta Univerzity Palackého v Olomouci, 2012

[6] Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V.: *Compositional data analysis in the geosciences: From theory to practise*, The Geological Society, London, 2006

[7] Donevska, S.: *Testy parametrických hypotéz pro kompoziční data*, diplomová práce, Přírodovědecká fakulta Univerzity Palackého v Olomouci, 2010

[8] McCullagh, P., Nelder, J.A.: *Generalized linear models*, 2. vydání, Chapman & Hall, New York, 1989.

[9] Rendlová, J.: *Analýza kategoriálních dat - problém vícenásobné volby v odpovědi*, diplomová práce, Přírodovědecká fakulta Univerzity Palackého v Olomouci, 2015.