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

BAKALÁŘSKÁ PRÁCE

Regresní analýza s kompozičními vysvětlujícími proměnnými

Vedoucí bakalářské práce: **Doc. RNDr. Karel Hron, Ph.D.**
Vypracoval: **Tomáš Zdražil**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2015

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Tomáš Zdražil

Název práce: Regresní analýza s kompozičními vysvětlujícími proměnnými

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2015

Abstrakt: V praktických úlohách se často setkáváme se situací, kdy vysvětlovaná proměnná závisí na několika vysvětlujících proměnných, popisujících relativní příspěvky složek na nějakém celku (typicky pak můžeme tyto proměnné vyjádřit v procentuálních podílech). Úloha regresní analýzy se v takovém případě řeší pomocí vyjádření vysvětlujících proměnných v ortonormálních souřadnicích tak, aby se mohlo použít standardních statistických testů. Teoretické výstupy demonstrujeme na příkladech z oblasti medicíny a ekonomie.

Klíčová slova: kompoziční data, regresní analýza, ortonormální souřadnice

Počet stran: 38

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Tomáš Zdražil

Title: Regression analysis with compositional explanatory variables

Type of thesis: Bachelor's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2015

Abstract: In practical tasks we often experience the situation when the response depends on several explanatory variables, describing relative contributions of parts on a whole (we typically express these variables in percentages). In this case, the task of the regression analysis is solved via expressing the covariates in orthonormal coordinates for which standard statistical tests can be employed. We apply the methodological outputs to examples from medicine and economics.

Key words: compositional data, regression analysis, orthonormal coordinates

Number of pages: 38

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně pod vedením pana Doc. RNDr. Karla Hrona, Ph.D. s použitím uvedené literatury.

V Olomouci dne 21. dubna 2015

Poděkování

Rád bych poděkoval vedoucímu mé bakalářské práce panu Doc. RNDr. Karlu Hronovi, Ph.D. za čas, který mi věnoval při konzultacích, a za velmi cenné připomínky.

Obsah

Úvod

Tématem mé bakalářské práce je regresní analýza s kompozičními vysvětlujícími proměnnými. Důvodem jeho výběru byl fakt, že mě zaujala problematika regrese. Hledal jsem nějakou oblast statistiky, kde se tento typ úlohy vyskytuje, a našel jsem velmi zajímavou aplikaci na kompoziční data, což jsou mnohorozměrná pozorování se specifickými vlastnostmi. Ve své práci bych se rád zaměřil na metodiku práce s lineární regresí v rámci logratio metodiky kompozičních dat, a to v úlohách, kdy jsou kompozičního tvaru pouze proměnné vysvětlující.

V první kapitole se seznámíme s mnohonásobnou lineární regresí. Podíváme se na význam regresních parametrů a uvedeme si dva testy, jejichž úkolem je zjistit, zda jsou vybrané regresní parametry statisticky významné.

V kapitole druhé představíme pojem kompoziční data včetně jejich vlastností, kvůli kterým pro ně vyvíjíme speciální postupy řešení statistických úloh. Dozvíme se, co to jsou ortonormální souřadnice a jak se vytvářejí.

Následující kapitola bude věnována spojení dvou předcházejících, tedy konstrukci regresního modelu s kompozičními vysvětlujícími proměnnými. Podíváme se, jak se postupem času tyto modely vyvíjely. Dále se seznámíme s konstrukcí regresního modelu v ortonormálních souřadnicích, tedy s úlohou, která je motivací této práce. Závěrem se zmíníme o přechodu na ortogonální souřadnice, což nám umožňuje lepší interpretaci regresních parametrů.

Poslední kapitola teoretické části práce se zabývá zpracováním úlohy lineární regrese s kompozičními vysvětlujícími proměnnými ve statistickém softwaru R. Nejprve představíme, jak R pracuje, a poté se seznámíme s nejdůležitějšími funkcemi pro řešení naší úlohy.

Nakonec si uvedeme dva modelové příklady z praxe, na kterých si použití lineární regrese s kompozičními vysvětlujícími proměnnými názorně ukážeme.

1. Mnohonásobná regrese

1.1. Co to je mnohonásobná regrese

V případě kompozičních dat, kdy máme vícesložkové kompoziční vektory, si již s klasickou regresí (závisle proměnná vysvětlená jednou vysvětlující proměnnou) nevystačíme, proto se zde využívá tzv. *mnohonásobné regrese*, kde máme vysvětlujících proměnných více. S větším počtem vysvětlujících proměnných nám samozřejmě i přibývá na obtížnosti najít nejvhodnější regresní funkci. Tuto problematiku řešíme pomocí aparátu matematické statistiky, např. statistických testů či měr těsnosti.

My si uvedeme nejjednodušší případ, a sice regresní funkci závisející na vysvětlujících proměnných lineárně. Mějme tedy $x_1, \ldots, x_p$ naše vysvětlující proměnné. Regresní funkci pro vysvětlovanou proměnnou Y definujeme jako

$$\mathsf{E}(Y|(x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

kde $\beta_0, \ldots, \beta_p$ jsou neznámé parametry, které je třeba odhadnout. V praxi to znamená, že při konkrétním datovém souboru $(\mathbf{x}_1, Y_1)', \ldots, (\mathbf{x}_n, Y_n)'$ bude pro výsledek i-tého pozorování platit

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ jsou i-té hodnoty p vysvětlujících proměnných a ε_i je náhodná chyba při i-tém pozorování.

Typicky se odhad vektoru regresních parametrů provádí *metodou nejmenších čtverců (MNČ)*. Uvědomíme-li si, že obecný tvar tohoto modelu můžeme zapsat maticově ve tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

podrobněji

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

kde $\mathbf{X}$ je tzv. matice plánu a $\boldsymbol{\varepsilon}$ je vektor chyb, tak odhad vektoru regresních parametrů pomocí MNČ za předpokladu $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $r(\mathbf{X}) = p+1$ bude tvaru

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)',$$

s charakteristikami $\mathsf{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, popřípadě jejich odhadem $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = S^2(\mathbf{X}'\mathbf{X})^{-1}$ pro nestranný odhad parametru σ^2,

$$S^2 = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n-p-1}.$$

Nyní, po odhadnutí vektoru regresních parametrů $\widehat{\boldsymbol{\beta}}$, můžeme zavést tzv. *vyrovnanou hodnotu* i-tého pozorování jako

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip}, \quad i = 1, \ldots, n.$$

Odhad regresní funkce je potom tvaru

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p.$$

1.2. Význam regresních parametrů

Zmiňme se ještě o významu regresních parametrů. V případě mnohonásobné regrese se parametrům kromě absolutního členu β_0 říká *(dílčí) regresní koeficienty* a přeznačují se: $\beta_1, \ldots, \beta_p \quad \to \quad \beta_{Y x_1.x_2,\ldots,x_p}, \ldots, \beta_{Y x_p.x_1,\ldots,x_{p-1}}$. Jejich interpretace je jednoduchá: uvádějí nám, jak se změní (průměrně) vysvětlovaná proměnná Y při jednotkové změně vysvětlující proměnné před tečkou za předpokladu, že vysvětlující proměnné za tečkou zůstanou beze změny.

Máme-li naměřené hodnoty v různých jednotkách, nemůžeme dost dobře posoudit, která vysvětlující proměnná má na hodnoty proměnné vysvětlované největší vliv. Z tohoto důvodu se zavádějí tzv. *beta koeficienty*, které vzniknou normalizací regresních koeficientů. Máme-li vektor $(x_{i1}, \ldots, x_{ip}, y_i)'$ pro $i = 1, \ldots, n$ jako výsledek i-tého pozorování, vypočteme

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad s_{x_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2},$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

pro $j = 1, \ldots, p$. Nyní provedeme normalizaci, což je transformace, kdy dostáváme

$$\widehat{Y}^* = \frac{\widehat{Y} - \overline{y}}{s_y} \quad , \quad x_j^* = \frac{x_j - \overline{x}_j}{s_{x_j}},$$

pro $j = 1, \ldots, p$, a zavedeme beta koeficienty (zkrácený zápis: B-koeficienty)

$$B_{Yx_1.x_2,\ldots,x_p} = \frac{s_{x_1}}{s_y}\widehat{\beta}_{Yx_1.x_2,\ldots,x_p}, \ldots, B_{Yx_p.x_1,\ldots,x_{p-1}} = \frac{s_{x_p}}{s_y}\widehat{\beta}_{Yx_p.x_1,\ldots,x_{p-1}}.$$

Potom můžeme odhad regresní funkce vyjádřit ve tvaru

$$\widehat{Y}^* = B_{Yx_1.x_2,\ldots,x_p}x_1^* + \ldots + B_{Yx_p.x_1,\ldots,x_{p-1}}x_p^*.$$

Výsledné B-koeficienty jsou bezrozměrné, tedy nezávislé na měrných jednotkách, a součet jejich absolutních hodnot je roven jedné. Tyto vlastnosti nám již umožňují jejich vzájemné srovnávání. Čím větší hodnota B-koeficientu, tím větší má příslušná vysvětlující proměnná vliv na proměnnou vysvětlovanou.

1.3. Testy o regresních parametrech

1.3.1. Významnost jednoho regresního parametru

Často se stává, že máme o hodnotách regresních parametrů (nebo jen o hodnotách některých z nich), které odhadujeme, už nějakou předběžnou představu. V takovýchto případech nás pak zajímá, zda odhadnuté hodnoty $\widehat{\beta}_j$ odpovídají naší představě β_j^0. Tedy máme

$$H_0 : \beta_j = \beta_j^0 \quad , \quad H_A : \beta_j \neq \beta_j^0,$$

pro $j = 0, \ldots, p$. Pokud se speciálně ptáme, zda $\beta_j = 0$ pro $j = 1, \ldots, p$, chceme tím zjistit, jestli vůbec Y závisí na dané proměnné x_j (populárně hovoříme o *testu statistické významnosti*).

Jestliže jsou splněny předpoklady pro odhad regresních parametrů pomocí MNČ, tedy $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $r(\mathbf{X}) = p + 1$, můžeme přidáním předpokladu normality, tzn. $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, k testování statistické významnosti jednotlivých parametrů β_j, $j = 0, \ldots, p$, využít testovací statistiku

$$T_j = \frac{\widehat{\beta}_j - \beta_j^0}{S\sqrt{\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}}} \sim t_{n-p-1}, \tag{1}$$

která se za platnosti H_0 řídí Studentovým t-rozdělením o $n - p - 1$ stupních volnosti, kde n je počet napozorovaných hodnot, $p + 1$ je počet regresních parametrů a $\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}$ značí $(j + 1)$-ní diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$. Statistika S je odmocnina z $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-p-1}$, což je nestranný odhad parametru σ^2. Proč se tato statistika T_j řídí zmíněným rozdělením, není těžké dokázat. Z $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ a z [3] plyne nezávislost náhodných veličin

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\Big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\Big) \text{ a } S^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}.$$

Dále uvažujme pouze regresní parametr $\widehat{\beta}_j \sim \mathrm{N}\Big(\beta_j, \sigma^2\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}\Big)$, odkud normováním obdržíme

$$\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}}} \sim \mathrm{N}(0, 1).$$

Víme, že

$$S^2\frac{n-p}{\sigma^2} \sim \chi^2_{n-p}.$$

Studentovo t-rozdělení má dle definice náhodná veličina

$$\mathrm{Y} = \frac{X}{\sqrt{\frac{Z}{q}}} \sim t_{q-1},$$

kde X má normované normální rozdělení a Z se řídí χ^2 rozdělením o q stupních volnosti. Dohromady tedy máme

$$\frac{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}}}}{\sqrt{\frac{S^2\frac{n-p}{\sigma^2}}{n-p}}} = \frac{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}}}}{\frac{S}{\sigma}} = \frac{\widehat{\beta}_j - \beta_j}{S\sqrt{\{(\mathbf{X}'\mathbf{X})^{-1}\}_{jj}}} \sim t_{n-p-1},$$

což je přesně statistika T_j ze vztahu (1).

Testovanou hypotézu H_0 na základě porovnání realizace testové statistiky s kritickým oborem

$$W = (-\infty, -t_{n-p-1;1-\frac{\alpha}{2}}\rangle \cup \langle t_{n-p-1;1-\frac{\alpha}{2}}, \infty)$$

buďto zamítneme, nebo nezamítneme na námi zvolené hladině testu α (obvykle $\alpha = 0.1$, $\alpha = 0.05$ nebo $\alpha = 0.01$).

Otázkou může být, proč jsme kritický obor zadefinovali zrovna pomocí kvantilů $\pm\, t_{n-p-1;1-\frac{\alpha}{2}}$ Studentova t-rozdělení. Víme, že pro spojitou a rostoucí distribuční funkci F_X na intervalu $(0, 1)$ je α-kvantil x_α takové reálné číslo, pro které platí, že $\mathsf{P}(X \leq x_\alpha) = \alpha$ a současně $\mathsf{P}(X \geq x_\alpha) = 1 - \alpha$. V našem případě by to bylo $\mathsf{P}(|T| \geq t_{n-p-1;1-\frac{\alpha}{2}}) = \alpha$ a současně $\mathsf{P}(|T| \leq t_{n-p-1;1-\frac{\alpha}{2}}) = 1 - \alpha$. Tedy kritický obor by pokrýval za platnosti H_0 realizaci testové statistiky (a my bychom ji tedy mylně zamítali) s pravděpodobností α, což je hodnota pravděpodobnosti chyby prvního druhu, kterou před testováním nastavíme a s níž pak vstupujeme do testu.

Rozhodnutí o hypotéze můžeme také provést na základě porovnání našeho zvoleného α s tzv. *p-hodnotou (p-value)*, která vyjadřuje pravděpodobnost (počítanou za platnosti H_0), že při stávajících datech dostaneme právě naši realizaci testové statistiky nebo realizaci příznivější pro H_A (tedy příznivější pro zamítnutí H_0). Je-li tedy p-hodnota menší nebo rovna α, zamítáme H_0 na hladině testu α, v opačném případě H_0 na dané hladině nelze zamítnout. S p-hodnotami se pracuje především ve statistických softwarech, kde jako výsledek testování hypotézy dostáváme nejčastěji právě p-hodnotu.

1.3.2. Souhrnná významnost regresních koeficientů

Pomocí statistiky T_j jsme mohli testovat i $H_0 : \beta_j = 0$, $j = 1, \ldots, p$, tedy je-li regresní koeficient β_j statisticky významný. Existuje ale také jiný pohled na danou problematiku. Může nás totiž zajímat, zda jsou vlivy všech vysvětlujících proměnných souhrnně statisticky významné. Testujeme proto všechny re-

gresní koeficienty současně při

$$H_0 : \beta_1 = \ldots = \beta_p = 0, \quad H_A : \text{alespoň jedno } \beta_j \text{ se nerovná nule.}$$

Pro konstrukci tohoto testu potřebujeme určit několik charakteristik, a sice *celkový součet čtverců*,

$$S_c = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

reziduální součet čtverců,

$$S_r = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2,$$

součet čtverců vyrovnaných veličin,

$$S_v = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2,$$

a konečně *koeficient determinace*,

$$R^2 = \frac{S_v}{S_c} = 1 - \frac{S_r}{S_c},$$

jenž se realizuje v intervalu $\langle 0, 1 \rangle$ a udává nám, jak velkou část variability dat se nám podařilo vysvětlit zvoleným modelem.

Za předpokladu normality vektoru náhodných chyb $\boldsymbol{\varepsilon}$ nyní můžeme zkonstruovat testovou statistiku

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p} \sim \mathrm{F}_{p,n-p-1}, \tag{2}$$

která se za platnosti H_0 řídí Fisher-Snedecorovým F-rozdělením o p a $n - p - 1$ stupních volnosti, kde n je počet napozorovaných hodnot a p je počet regresních koeficientů odpovídajících vysvětlujícím proměnným. Hypotézu H_0 zamítneme na příslušné hladině testu α, jestliže realizace testové statistiky F překročí kvantil $\mathrm{F}_{p,n-p-1;1-\alpha}$.

Na závěr kapitoly se ještě vraťme ke koeficientu determinace. Předpis koeficientu determinace nezávisí na počtu parametrů regresní funkce, což znamená, že s rostoucím počtem parametrů nám obvykle roste i hodnota koeficientu determinace. Z tohoto důvodu byl vytvořen tzv. *upravený koeficient determinace* jako

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2),$$

kde n je počet pozorování a $p+1$ je počet parametrů v modelu. Nutno ovšem podotknout, že pro velké výběry n se oba koeficienty už prakticky neliší. R^2_{adj} lze využít i pro hledání vhodného modelu, jelikož po přidání vysvětlující proměnné do regresního modelu (která model zlepšuje) se R^2_{adj} zvětšuje, a naopak.

2. Kompoziční data

2.1. Co to jsou kompoziční data

Při sběru dat nám mnohdy nejde o absolutní hodnoty proměnných, ale pouze o relativní příspěvky daných částí na celku. Takovému typu dat se říká *kompoziční data*. Nejčastěji se vyskytují ve formě proporcí (např. proporce jednotlivých chemických složek v nápoji) nebo procent (např. procentuální zastoupení politických stran ve vládě), ovšem mohou být i jiného typu (ppm, ppb, mg/kg, ...).

Kompoziční data potom můžeme zapsat ve formě vektoru, který má obecně D složek, jenž jsou kladné a nesou pouze relativní informaci. Ve většině případů mají tyto složky konstantní součet κ. Nejčastěji je $\kappa = 1$ (případ proporcí) nebo $\kappa = 100$ (případ procent). Díky této skutečnosti můžeme dále definovat dva pojmy:

- *Uzávěr* libovolného D-složkového kompozičního vektoru $\mathbf{z}$ je vektor

$$C(\mathbf{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, \ldots, \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i} \right].$$

- *Simplex* je výběrový prostor kompozičních dat. Je to prostor všech D-složkových kompozičních vektorů, kdy součet D složek těchto vektorů je roven konstantě κ. Poznamenejme, že dimenze takového výběrového prostoru je pak zřejmě pouze $D - 1$.

Po aplikaci uzávěru dostáváme vlastně stejný vektor přeškálovaný tak, aby součet jeho složek byl roven konstantě κ (její hodnota závisí na jednotkách měření). To znamená, že z každého libovolného vektoru s kladnými složkami můžeme vytvořit kompoziční vektor v uvedeném smyslu (s předepsaným součtem složek κ). Dále se této skutečnosti využije při definování pojmu *podkompozice*, což je při zadané kompozici $\mathbf{x}$ vektor $\mathbf{x}_s$, na jehož složky $(x_{i_1}, x_{i_2}, \ldots, x_{i_s})'$ aplikujeme právě operaci uzávěru. Indexy $i_1, i_2, \ldots, i_s$ nám udávají, kterých s složek jsme do podkompozice vybrali. Je tedy zřejmé, že to nemusí být jen prvních s složek.

2.2. Vlastnosti kompozičních dat

2.2.1. Invariance na změnu měřítka

Ukažme si tuto vlastnost na jednoduchém příkladu: uvažujme, že jedna tabulka čokolády o hmotnosti 120g obsahuje 12g bílkovin, 72g sacharidů a 36g tuků. Kompozice složení je tedy vektorově $\mathbf{a} = (12, 72, 36)'$. Ovšem můžeme ji také vyjádřit i v jiných jednotkách, například v procentech, $\mathbf{b} = (10, 60, 30)'$, nebo v proporcích, $\mathbf{c} = \left(\frac{1}{10}, \frac{6}{10}, \frac{3}{10}\right)'$. Ačkoliv je patrné, že absolutní hodnoty těchto tří vektorů jsou různé, z hlediska kompozice nehrají žádnou roli. Nás zajímají pouze poměry mezi jednotlivými složkami a ty jsou ve všech případech stejné, tedy kompozice jsou invariantní na změnu měřítka.

2.2.2. Invariance vůči permutacím

Funkce je invariantní vůči permutacím složek kompozice, jestliže při změně pořadí složek dostaneme stejnou funkční hodnotu. Jedná se o další z vlastností, analogických situaci v mnohorozměrné statistice, kterou budeme pro relevantní analýzu kompozičních dat vyžadovat.

2.2.3. Podkompoziční soudržnost

V určitých případech chceme pracovat pouze s některými složkami D-složkové kompozice, čili vezmeme pouze jistou podkompozici. Podkompozice přitom hrají stejnou roli jako marginální vektory v klasické mnohorozměrné analýze a tomu by pak měly odpovídat jejich geometrické vlastnosti. Tento princip se nazývá *podkompoziční soudržnost* a má několik praktických důsledků. Za všechny si můžeme uvést tyto:

- Jestliže měříme vzdálenost mezi dvěma D-složkovými kompozicemi, tato musí být větší, když bereme v úvahu všech D složek, než když měříme pouze s využitím jakýchkoliv odpovídajících si podkompozicí. Této vlastnosti se říká *podkompoziční dominance*. Tuto podmínku například standardní euklidovská vzdálenost aplikovaná na kompozice nesplňuje.

- Pokud máme vhodný model pro D-složková kompoziční data, výsledky pro sledovanou podkompozici by se neměly změnit přidáním (odebráním) ostatních složek.

2.2.4. Ortonormální souřadnice

V předchozí podkapitole jsme si uvedli, že euklidovská geometrie není vhodná pro práci s kompozicemi. Z tohoto důvodu byla vytvořena na simplexu tzv. *Aitchisonova geometrie*, jejíž vlastnosti již plně vyhovují vlastnostem kompozičních dat. Byla pojmenována po skotském statistikovi Johnu Aitchisonovi, jenž jako první přišel s úvahami o vytvoření speciální geometrie pro kompoziční data.

Návodů na přechod z Aitchisonovy geometrie na simplexu do euklidovské geometrie v reálném prostoru (pro niž je definována většina standardních statistických metod) je hned několik, pro naše další úvahy bude ale důležitý zejména jeden, a to vyjádření kompozic v tzv. *isometrických log-ratio souřadnicích (ilr)*. Myšlenkou tohoto postupu je zkonstruovat ortonormální bázi na simplexu a vyjádřit kompozici v souřadnicích vzhledem k této bázi. Jedna její konkrétní volba vede pro D-složkový kompoziční vektor $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ na $(D-1)$-složkový reálný vektor $\mathbf{z} = (z_1, z_2, \ldots, z_{D-1})'$, jehož složky jsou dány vztahem

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1. \tag{3}$$

Z něj vyplývá, že i-tá složka vektoru $\mathbf{z}$ je definována jako (normovaný) logaritmus podílu jí pořadím odpovídající složky x_i a složek v pořadí následujících. To znamená, že pouze první souřadnice z_1 obsahuje plnou relativní informaci o složce x_1, jelikož je tvořena logaritmem podílu mezi x_1 a zbylými složkami vektoru $\mathbf{x}$, reprezentovanými ve formě jejich geometrického průměru. S každou další složkou vektoru $\mathbf{z}$ se ztrácí vliv jedné složky kompozice $\mathbf{x}$.

Fakt, že pouze první souřadnice z_1 nám podává kompletní informaci o složce x_1, není ideální v případech, kdy nám jde o analýzu k-té složky. Můžeme ovšem zkonstruovat obdobnou ortonormální bázi, která bude vycházet z původního vek-

toru $\mathbf{x}$, s permutací složek $\mathbf{x}^{(k)} = (x_k, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_D)'$, kde máme k-tou složku, $k = 1, \ldots, D$, na první pozici. Pořadí ostatních složek může být vzhledem k dalším účelům naší analýzy libovolné. Přeznačíme-li si nyní $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_k^{(k)}, x_{k+1}^{(k)}, \ldots, x_D^{(k)})'$, můžeme psát vztah (3) v obecnějším tvaru jako

$$z_i^{(k)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(k)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(k)}}}, \quad i = 1, \ldots, D-1, \tag{4}$$

pro $\mathbf{z}^{(k)} = (z_1^{(k)}, \ldots, z_{D-1}^{(k)})$, $z_i^{(1)} \equiv z_i$. Poznamenejme zde pro pozdější účely, že jednou z charakteristických vlastností takovýchto vektorů $\mathbf{z}^{(k)}$ je i jejich vzájemná ortogonalita.

Pro zpětný přechod z euklidovské geometrie do Aitchisonovy lze využít následujících vztahů:

$$x_i = \begin{cases} \exp\left(\sqrt{\frac{D-1}{D}} z_i\right) & , \quad i = 1 \\ \exp\left(-\sum_{j=1}^{i-1} \frac{z_j}{\sqrt{(D-j+1)(D-j)}} + \sqrt{\frac{D-i}{D-i+1}} z_i \right) & , \quad i = 2, \ldots, D-1 \\ \exp\left(-\sum_{j=1}^{D-1} \frac{z_j}{\sqrt{(D-j+1)(D-j)}} \right) & , \quad i = D. \end{cases}$$

3. Kompoziční regrese

3.1. Vývoj kompoziční regrese

Užití standardní regresní analýzy je zcela opodstatněné, pokud máme data (vysvětlující i vysvětlovanou proměnnou), která nesou absolutní informaci. Kompoziční data ovšem nesou informaci pouze relativní, což znamená, že součet složek zde nehraje podstatnou roli, a proto není vhodné standardního modelu vícenásobné regrese využít.

Z tohoto důvodu byly snahy o rozšíření regresní analýzy i pro kompoziční případ. Stavebním kamenem těchto myšlenek byl fakt, že kompoziční data můžeme interpretovat jako pozorování s jednotkovým součtem složek (proporce), který byl ovšem pro účely dalšího statistického zpracování poněkud nevhodně považován za pevně daný. To dalo vzniknout tzv. *experimentování se směsmi* (v anglickém originále *experiments with mixtures*), kdy dostáváme následující tvary regresní funkce:

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i, \quad \mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \beta_{ij} x_i x_j.$$

Parametry se zde odhadují standardní metodou nejmenších čtverců, nicméně takto vzniklé modely se obvykle vyznačují špatnou podmíněností.

Velkým posunem vpřed v této oblasti bylo zavedení tzv. *log-kontrastu*. Podstatou této metodiky je zakomponování přirozeného logaritmu do regresní funkce, tedy pro její část s vysvětlujícími proměnnými dostaneme tvar $\beta_1 \ln x_1 + \beta_2 \ln x_2 + \ldots + \beta_D \ln x_D$, kde musí být splněna podmínka na parametry typu $\beta_1 + \beta_2 + \ldots + \beta_D = 0$. Modely s využitím log-kontrastu potom můžeme definovat těmito vztahy (lineární a kvadratický model):

$$\mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i \ln x_i, \quad \mathsf{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i \ln x_i + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \beta_{ij} \left(\ln \frac{x_i}{x_j} \right)^2.$$

I zde ovšem těžkosti s interpretací parametrů a jejich odhady přetrvávaly.

3.2. Lineární regrese v ortonormálních souřadnicích

Máme-li kompoziční vektor $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ vyjádřený v ortonormálních souřadnicích jako vektor $\mathbf{z} = (z_1, z_2, \ldots, z_{D-1})'$, můžeme k výstavbě lineárního modelu pro Y a $\mathbf{x}$ využít vektoru $\mathbf{z}$. Lineární model tak bude tvaru

$$\mathsf{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_{D-1} z_{D-1}.$$

Vektor regresních parametrů $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{D-1})'$ odhadneme standardní metodou nejmenších čtverců. Žádné podmínky na regresní parametry zde neklademe.

Bohužel, z tohoto modelu je interpretovatelný pouze parametr absolutního členu (γ_0) a regresní koeficient γ_1 příslušný souřadnici z_1. Ostatní regresní koeficienty se totiž vztahují k souřadnicím, které nenesou kompletní kompoziční informaci o jednotlivých složkách. Z tohoto důvodu tvoříme permutací kompozičních složek pro konstrukci vektoru $\mathbf{z}$ (kdy se na první pozici vystřídá všech D složek) D regresních modelů tak, abychom postupně mohli interpretovat vliv všech složek (resp. relativní informace o těchto složkách) vektoru $\mathbf{x}$ na Y pomocí $\gamma_1^{(k)}$, $k = 1, \ldots, D$. Lineární modely budou tvaru

$$\mathsf{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1^{(k)} z_1^{(k)} + \ldots + \gamma_{D-1}^{(k)} z_{D-1}^{(k)}, \quad k = 1, \ldots, D. \tag{5}$$

Jelikož je regresní parametr γ_0 přímo spjatý s vysvětlovanou proměnnou Y, tudíž nemá žádnou spojitost s výběrem ortonormální báze na simplexu, a také díky vzájemné ortogonalitě mezi D vektory $\mathbf{z}^{(k)}$, zůstává stejný napříč všemi těmito modely.

V situaci, kdy máme k dispozici datový soubor typu $(\mathbf{x}_1, Y_1)', \ldots, (\mathbf{x}_n, Y_n)'$, dostáváme výsledek i-tého pozorování pro první z uvažovaných modelů jako

$$Y_i = \gamma_0 + \gamma_1 z_{i1} + \ldots + \gamma_{D-1} z_{i,D-1} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{6}$$

Odhad vektoru regresních parametrů $\widehat{\boldsymbol{\gamma}}$ dostaneme pomocí metody nejmenších čtverců. Za předpokladu platnosti všech podmínek pro použití MNČ popsaných v kapitole 1.1 můžeme z modelu

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

podrobněji

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & z_{11} & \dots & z_{1,D-1} \\ \vdots & \vdots & & \vdots \\ 1 & z_{n1} & \dots & z_{n,D-1} \end{pmatrix} \cdot \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_{D-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

určit vektor odhadnutých regresních parametrů

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = (\widehat{\gamma}_0, \dots, \widehat{\gamma}_{D-1})'.$$

Dále lze přidáním předpokladu $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$ provést testy o regresních parametrech. Zkonstruujeme testovou statistiku (1), v tomto případě tvaru

$$T_j = \frac{\widehat{\gamma}_j - \gamma_j^0}{S\sqrt{\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{jj}}} \sim t_{n-D}, \quad j = 0, 1, \dots, D-1, \tag{7}$$

kde

$$S = \sqrt{S^2} = \sqrt{\frac{(\mathbf{Y} - \mathbf{Z}\boldsymbol{\gamma})'(\mathbf{Y} - \mathbf{Z}\boldsymbol{\gamma})}{n - D}},$$

tedy S je odmocnina z nestranného odhadu σ^2 a $\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{jj}$ značí $(j+1)$-ní diagonální prvek matice $(\mathbf{Z}'\mathbf{Z})^{-1}$. Za platnosti H_0 se T_j řídí Studentovým t-rozdělením o $n-D$ stupních volnosti. Tento test budeme využívat zejména pro testování statistické významnosti jednotlivých regresních parametrů.

Jelikož jediným interpretovatelným regresním koeficientem je γ_1, bude mít význam konstruovat pouze testové statistiky T_0 a T_1 za účelem testování parametrů γ_0 a γ_1, obecněji γ_0 a $\gamma_1^{(k)}$, pro $k = 1, \dots, D$ (pomocí odpovídajících statistik T_0 a $T_1^{(k)}$). Naším cílem je určit pomocí testu parametru $\gamma_1^{(k)}$ (příslušícího $z_1^{(k)}$, jenž obsahuje kompletní relativní informaci o kompoziční složce x_k), zda podkompozice, vzniklá vypuštěním složky x_k, může plnohodnotně nahradit původní kompozici v regresním modelu. Zde je na místě upozornit, že test bude dávat odlišné výsledky, vezmeme-li celou kompozici a vezmeme-li jen nějakou její podkompozici.

To si lépe uvědomíme, podíváme-li se na vztah (4): tvar $z_i^{(k)}$ přímo závisí na tom, které složky jsou v podkompozici obsaženy.

Důležitým požadavkem je invariance testových statistik T_0 a $T_1^{(k)}$ vůči permutaci složek kompozice při konstrukci ortonormálních souřadnic. Statistiku T_0 jsme využívali pro test $H_0 : \gamma_0 = 0$ vs. $H_A : \gamma_0 \neq 0$, pomocí $T_1^{(k)}$ jsme zase testovali $H_0 : \gamma_1^{(k)} = 0$ vs. $H_A : \gamma_1^{(k)} \neq 0$, pro $k = 1, \ldots, D$. Ukazuje se, že tato invariance zde přítomná je. Uveďme si ji ve tvaru věty, abychom následně mohli provést i její důkaz.

Věta 1: Mějme lineární model (5).

a) Testové statistiky T_0 a $T_1^{(k)}$ jsou invariantní vůči změně pořadí $x_2^{(k)}, \ldots, x_D^{(k)}$ ve (4). Tato invariance přetrvává i pro odhadnuté hodnoty regresního modelu.

b) Testová statstika T_0 je invariantní vůči změně pořadí $x_1^{(k)}, \ldots, x_D^{(k)}$ ve (4).

Důkaz: Bez újmy na obecnosti uvažujme $k = 1$, jsme tedy v situaci modelu (6) s ortonormálními souřadnicemi definovanými jako (3).

a) Změna pořadí $x_2, \ldots, x_D$ v (3) odpovídá změně ortonormální báze na simplexu, tedy matice $\mathbf{Z}$ se vynásobí zprava ortogonální maticí řádu D,

$$\mathbf{P} = \begin{pmatrix} 1 & & \\ & 1 & \\ & & \mathbf{P}_1 \end{pmatrix},$$

s jedničkami na prvních dvou pozicích diagonály, s ortogonální maticí $\mathbf{P}_1$ řádu $D-2$ a s nulami všude jinde. Dostaneme tak $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}_D$, kde $\mathbf{I}_D$ je jednotková matice řádu D. Je dokonce i možné vybrat k vyjádření podkompozice $x_2, \ldots, x_D$ též souřadnice vzhledem k nějaké bázi, která není

ortonormální. Důsledkem bude ztráta ortogonality matice $\mathbf{P}$, která ovšem neomezuje úvahy níže. Užitím vztahu

$$[(\mathbf{ZP})'\mathbf{ZP}]^{-1}(\mathbf{ZP})'\mathbf{Y} = \mathbf{P}^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{P}')^{-1}\mathbf{P}'\mathbf{Z}'\mathbf{Y} = \mathbf{P}^{-1}\widehat{\boldsymbol{\gamma}}$$

můžeme vidět, že hodnoty odhadů $\widehat{\gamma}_0$, $\widehat{\gamma}_1$ a S^2 zůstávají, stejně jako první dva diagonální prvky matice $(\mathbf{Z}'\mathbf{Z})^{-1}$, neměnné přes zmíněnou regulární transformaci. Z toho vyplývá, že pro statistiky T_0 a $T_1^{(k)}$ je požadovaná invariance splněna. Nakonec si uvědomme, že platí vztah $\mathbf{ZPP^{-1}}\widehat{\boldsymbol{\gamma}} = \mathbf{Z}\widehat{\boldsymbol{\gamma}}$, tedy odhad vysvětlované proměnné regresním modelem je stejný bez ohledu na ortogonalitu vybrané báze.

b) Provedení důkazu je stejné jako v předchozím případě, ovšem tentokrát je matice $\mathbf{P}$ nahrazena maticí

$$\mathbf{Q} = \begin{pmatrix} 1 & \\ & \mathbf{Q}_1 \end{pmatrix},$$

tedy s jedničkou na první pozici diagonály, s ortogonální maticí $\mathbf{Q}_1$ řádu $D-1$ a s nulami všude jinde. Matice $\mathbf{P}$ je tak pouze speciálním případem matice $\mathbf{Q}$. Invariance $\widehat{\gamma}_0$, S^2, prvního diagonálního prvku matice $(\mathbf{Z}'\mathbf{Z})^{-1}$, a tedy i invariance testové statistiky T_0, je zřejmá. □

Nyní provedeme souhrnný test o regresních koeficientech. Testová statistika bude dle (2) tvaru

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-D}{D-1} \sim \mathrm{F}_{D-1,n-D}.$$

Za platnosti H_0 se tedy F řídí Fisher-Snedecorovým F-rozdělením o $D-1$ a $n-D$ stupních volnosti. Alternativní (maticový) výpočet statistiky F by byl

$$F = \frac{1}{(D-1)S^2}\widehat{\boldsymbol{\gamma}}_*'\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}\widehat{\boldsymbol{\gamma}}_*, \tag{8}$$

kde $\widehat{\boldsymbol{\gamma}}_* = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_{D-1})'$ a $\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}$ značí, že první řádek a první sloupec byly z matice $(\mathbf{Z}'\mathbf{Z})^{-1}$ vyňaty.

Také zde máme požadavek na invarianci testové statistiky vůči permutaci složek kompozice při konstrukci ortonormálních souřadnic (4). Znovu se ukazuje, že tato invariance přítomná je. Opět si tedy uveďme příslušnou větu i s důkazem.

Věta 2: Testová statistika F je invariantní vůči změně pořadí $x_1^{(k)}, \ldots, x_D^{(k)}$ ve (4).

Důkaz: Postup důkazu je analogický jako u věty 1. Matice $\mathbf{Z}$ je vynásobena zprava regulární maticí $\mathbf{Q}$, tedy z (8) dostaneme vztah

$$\widehat{\boldsymbol{\gamma}}_*'\mathbf{Q}_1\mathbf{Q}_1^{-1}\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}(\mathbf{Q}_1')^{-1}\mathbf{Q}_1'\widehat{\boldsymbol{\gamma}}_* = \widehat{\boldsymbol{\gamma}}_*'\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(-1,-1)}\widehat{\boldsymbol{\gamma}}_*,$$

čímž je invariance dokázána. □

3.3. Lineární regrese v ortogonálních souřadnicích

Přestože předchozí regresní modely v ortonormálních souřadnicích jsou teoreticky velmi dobře zdůvodnitelné, normalizační konstanta i přirozený logaritmus jako takový ústí v poměrně složitou interpretaci regresních parametrů. Řešením by mohl být přechod na ortogonální souřadnice, kde neztrácíme žádnou z výše zmíněných vlastností, přičemž zároveň dostáváme velké zjednodušení interpretace parametrů. Uvažujme ortonormální souřadnice tvaru (4). Ortogonální souřadnice potom budou tvaru

$$z_i^{(k)*} = \log_2 \frac{x_i^{(k)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(k)}}}, \quad i = 1, \ldots, D-1,$$

kde $k = 1, \ldots, D$. Jak můžeme vidět, změna se projevila v tom, že jsme vypustili normalizační konstantu a přirozený logaritmus jsme nahradili logaritmem binárním (dvojkovým).

Nyní se podívejme, jak se projeví změna ortonormálních souřadnic na ortogonální v modelech z předchozí kapitoly. Mějme tedy lineární regresi s kompozičními vysvětlujícími proměnnými. Z vlastností MNČ odhadu a ze vztahu mezi logaritmy o různých základech dostaneme, že

$$\gamma_0^* = \gamma_0 \quad \text{a} \quad \gamma_1^{(k)*} = \ln(2)\sqrt{\frac{D-1}{D}}\gamma_1^{(k)}, \tag{9}$$

obecněji

$$\gamma_i^{(k)*} = \ln(2)\sqrt{\frac{D-i}{D-i+1}}\gamma_i^{(k)}, \quad i = 1, \ldots, D-1.$$

Analogické vztahy platí i mezi odhady těchto parametrů.

Takže co můžeme říci o interpretaci regresních koeficientů teď - v této podobě? Víme, že jednotkový přírůstek v souřadnici $z_1^{(k)*}$ lze za užití binárního logaritmu vyjádřit pomocí původní kompozice $\mathbf{x}$, a sice vztahem

$$\Delta z_1^{(k)*} = \log_2 \left(\frac{x_1^{(k)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(k)}}} \cdot 2 \right) - \log_2 \frac{x_1^{(k)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(k)}}} = 1.$$

Koeficient $\gamma_1^{(k)*}$ má potom význam přírůstku u vysvětlované proměnné Y při zvyšování $z_1^{(k)*}$ o 1 (tj. zdvojnásobení podílu složky x_k vůči průměru ostatních složek v kompozici), za předpokladu zachování konstantní úrovně všech ostatních souřadnic.

Závěrem poznamenejme, že i pro lineární regresi v ortogonálních souřadnicích platí tvrzení věty 1 a věty 2. Tedy i pro tento případ regresního modelu můžeme provést jak testy o jednotlivých regresních parametrech, tak souhrnný test o regresních koeficientech.

4. Kompoziční data v R

4.1. Co je to R

Při praktické práci nejen s kompozičními daty, ale i obecně při statistické analýze, je k výpočtům příslušných charakteristik vhodné využít některého ze statistických softwarů. V dnešní době je pro analýzu dat jedním z nejpoužívanějších softwarů R. Jeho obrovskou výhodou je nejen volná šiřitelnost (status freeware), ale i fakt, že má otevřený zdrojový kód. Pracuje na bázi tzv. *knihoven*, které obsahují zdrojové kódy příslušných početních operací a funkcí, které se týkají tématu, jemuž je knihovna věnována. Konkrétní knihovnu, kterou daný uživatel zrovna potřebuje, si může stáhnout jednoduchou cestou přímo v rozhraní R. Pro dokonalou orientaci obsahují knihovny také svůj vlastní *help list*, kde jsou podrobně rozepsány všechny jejich funkce (jejich vstupy, výstupy, možné argumenty, příklady). Některé z knihoven dokonce obsahují i různé soubory dat, na kterých si můžete příslušné výpočty a funkce vyzkoušet. Tohle všechno dělá z R kvalitní a velmi uživatelsky příjemný statistický software.

4.2. Knihovna robCompositions

Jelikož se zabýváme kompozičnímy daty, bude nás zajímat, co v tomto ohledu R nabízí. V této práci je užita knihovna `robCompositions`, která poskytuje potřebné funkce k provádění statistické analýzy kompozičních dat. Jejich seznam si je možno prohlédnout, jestliže do R zadáme příkaz

```
> help(package = "robCompositions")
```

Poznamenejme, že v dalším textu bude taková tabulka značit prostředí R, znak `>` bude označovat příkaz. Knihovna `robCompositions` využívá také několik dalších knihoven, ze kterých importuje vybrané funkce. Jsou to knihovny `car`, `MASS`, `robustbase`, `rrcov` a `utils`.

4.2.1. Funkce lmCoDaX

Funkce, kterou budeme z knihovny `robCompositions` primárně využívat, se nazývá `lmCoDaX`. Abychom s ní mohli pracovat, musíme si nejprve tuto knihovnu stáhnout, a poté ji načíst příkazem

```
> library(robCompositions)
```

Nyní již něco o této funkci samotné. Všechny potřebné údaje o ní, její vstup, výstup, argumenty, aj., si můžeme přečíst v help listu, který vyvoláme zadáním

```
> help(lmCoDaX)
```

Ve zkratce řekněme, že je to funkce, která nám vrací jako výstup kompoziční lineární regresi matice $\mathbf{X}$ hodnot kompozičních vysvětlujících proměnných na vektor $\mathbf{Y}$ hodnot vysvětlované proměnné. A to dvěma způsoby: buď robustně pomocí metody nejmenších useknutých čtverců (LTS), anebo klasicky pomocí MNČ. My se vydáme cestou klasickou, tedy využijeme MNČ, jelikož budeme chtít pro odhady regresních parametrů využít všechna naměřená data.

Mějme tedy data - matici $\mathbf{X}$ (vysvětlující proměnné) a vektor $\mathbf{Y}$ (vysvětlovaná proměnná). Po aplikaci funkce `lmCoDaX` ve tvaru

```
> lmCoDaX(y, X, method = "classical")
```

dostaneme jako výstup jednak tabulku s odhady regresních parametrů, a dále dvě souhrnné tabulky regresní analýzy (jednu tabulku vycházející z původních dat, jednu tabulku vycházející z dat vyjádřených v ortonormálních souřadnicích - ta nás bude zajímat primárně). Uvedený souhrn přitom obsahuje v první řadě minimum, maximum, medián a horní i dolní kvartil odhadnutých hodnot vysvětlované proměnné, v řadě druhé potom pro nás mnohem důležitější testy významnosti jednotlivých regresních parametrů v podobě příslušných p-hodnot, souhrnou významnost regresních koeficientů, koeficient determinace i upravený koeficient determinace. Tohle všechno nám pomůže určit, jak moc bychom měli být s regresí spokojeni. Kompoziční regresi v R si názorně ukážeme na dvou příkladech.

5. Příklady

5.1. Příklad 1: Úmrtí na zhoubné nádory

V prvním příkladě máme tuto situaci: v rámci sledování dlouhodobého vývoje zhoubných nádorů u 4 vybraných orgánů (močový měchýř, slinivka břišní, tlusté střevo, žaludek) jsme získali z 25 států Evropy roční data (2004) absolutních četností úmrtí na tyto jednotlivé nádory. Naším úkolem je zjistit, jak střední délka života v našich státech závisí na relativních příspěvcích jednotlivých nádorů.

Začneme načtením dat [2], která jsou v excelovském souboru abs_čet.csv (uložili jsme je s příponou *.csv proto, abychom je mohli úspěšně načíst příkazem `read.csv2`). Ten R úspěšně vyhledá pouze v případě, jestliže pracujeme ve složce, ve které je daný soubor umístěn. To si můžeme v R lehce nastavit pomocí File → Change dir... → naše složka. Načtení provedeme zadáním

```
> abs_cet = read.csv2("abs_čet.csv", row.names = 1)
```

Argument `row.names` nám udává přítomnost sloupce s názvy řádků, jeho hodnota 1 potom ukazuje na první sloupec. Vidíme, že data, která byla načtena, jsme uložili pod názvem `abs_cet`. Od této chvíle nám bude vždy stačit zadat pouze příkaz `abs_cet` a R nám zobrazí příslušnou datovou množinu (viz tabulka 1).

V dalším kroku si vytvoříme vektor vysvětlovaných proměnných, což jsou střední délky života v jednotlivých státech [1]. Vyhledáme si příslušné hodnoty a seřadíme je tak, aby jejich pořadí odpovídalo pořadí států v datové množině `abs_cet`. Tedy

```
> vek = c(78.88, 72.56, 75.72, 77.49, 78.68, 71.91, 78.20, 79.04,
  79.87, 80.16, 72.03, 71.96, 72.65, 79.10, 79.18, 74.85, 77.67,
  71.59, 77.21, 73.96, 78.71, 80.50, 78.75, 79.84, 75.52)
```

Vektor vysvětlovaných proměnných jsme si uložili pod názvem `vek`.

```
> abs_cet
                 mechyr  slinivka  strevo  zaludek
Belgie              797      1305    2230      841
Bulharsko           464       948    1255     1500
Česká republika     786      1711    2557     1409
Dánsko              583       813    1360      365
Německo            5552     13008   19420    11473
Estonsko            107       184     255      348
Irsko               149       380     617      305
Řecko              1016      1274    1898     1325
Španělsko          4496      4549    9803     5811
Francie            4613      7861   12394     5110
Lotyšsko            146       330     410      571
Litva               222       417     451      797
Maďarsko            840      1683    3092     1938
Nizozemsko         1126      1971    3466     1546
Rakousko            527      1317    1622     1125
Polsko             2795      3923    6093     5716
Portugalsko         666       976    2337     2404
Rumunsko           1150      2321    2573     3992
Slovinsko           125       242     339      363
Slovensko           251       555     933      771
Finsko              255       908     568      573
Švédsko             597      1454    1769      833
Velká Británie     4818      7063   10327     5877
Norsko              390       622    1170      408
Chorvatsko          297       537     858      967
```

Tabulka 1: Datová množina `abs_cet`

Nyní již můžeme přistoupit k provedení samotné regresní analýzy. Tedy po vzoru kapitoly 4.2.1 využijeme funkci `lmCoDaX`, která je součástí knihovny `robCompositions`. Před každou prací s její libovolnou funkcí je třeba ji nejprve načíst, což provedeme již známým příkazem

```
> library(robCompositions)
```

Teď už nám nic nebrání provést požadovaou regresní analýzu. Zadáme

```
> lmCoDaX(vek, abs_cet, method = "classical")
```

a výstupem dostaneme tabulku odhadů regresních parametrů následovanou dvěma souhrnnými tabulkami regresní analýzy. Pro nás má význam především poslední tabulka - regresní analýza s využitím ortonormálních souřadnic (viz tabulka 2).

```
$ilr

Call:
lm(formula = y   ., data = d)

Residuals:
     Min       1Q  Median      3Q     Max
 -3.9983  -1.6390  0.4432  1.6729  3.6668

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   77.023       2.189   35.184   < 2e-16  ***
 X.mechyr       1.676       2.643    0.634     0.533
 X.slinivka     1.405       2.297    0.612     0.547
 X.strevo       2.468       2.692    0.917     0.370
 X.zaludek     -5.548       1.129   -4.914  7.34e-05  ***
 ---
 Signif. codes:  0    '***'  0.001  '**'  0.01  '*'  0.05  '.'
                 0.1  ' '   1

Residual standard error: 2.187 on 20 degrees of freedom
Multiple R-squared: 0.5531, Adjusted R-squared: 0.4892
F-statistic: 8.663 on 3 and 21 DF, p-value: 0.0006186
```

Tabulka 2: Výstup regresní analýzy prvního příkladu

Zkusme interpretovat výsledky obsažené v tabulce 2: v posledním sloupci `Pr(>|t|)` máme údaje o p-hodnotách testů o statistické významnosti jednotlivých regresních parametrů. Můžeme vidět, že funkce nám označila absolutní člen `(Intercept)` a koeficient `X.zaludek` jako statisticky významné, a to dokonce i pro nejpřísnější užívanou volbu hladiny spolehlivosti $\alpha = 0.01$. U absolutního členu to znamená, že regresní přímka neprochází počátkem. U regresního koeficientu `X.zaludek` výsledek značí, že právě tento koeficient rozhodně

není roven nule, a tedy má významný vliv na podobu regresní přímky. U ostatních regresních koeficientů jsme vzhledem k vysokým p-hodnotám jejich nulovost zamítnout nemohli. Dále se podívejme na test o souhrnné statistické významnosti regresních koeficientů. Opět budeme usuzovat z jeho p-hodnoty, tedy sledujeme `p-value` u `F-statistic`. Je to hodnota velmi malá, takže je zřejmé, že hypotézu o nulovosti všech regresních koeficientů současně budeme zamítat. To ale není nic překvapivého vzhledem k tomu, že jsme již zamítli nulovost koeficientu `X.zaludek`. Podívejme se ještě na koeficient determinace (`Multiple R-squared`) a jeho upravenou podobu (`Adjusted R-squared`). Velkoryse můžeme říci, že v obou případech se jeho hodnota pohybuje kolem 0.5, tedy naším modelem se nám podařilo vysvětlit okolo 50% celkové variability dat. To by se na první pohled mohlo zdát poněkud málo, ovšem vezmeme-li v potaz, že jsme vycházeli z reálných dat, můžeme být s touto hodnotou spokojeni.

Vytvořme si také přehlednou tabulku, kde shrneme odhady regresních koeficientů regresní analýzy vycházející z ortonormálních souřadnic, ortogonálních souřadnic (s využitím (9)) a příslušné B-koeficienty.

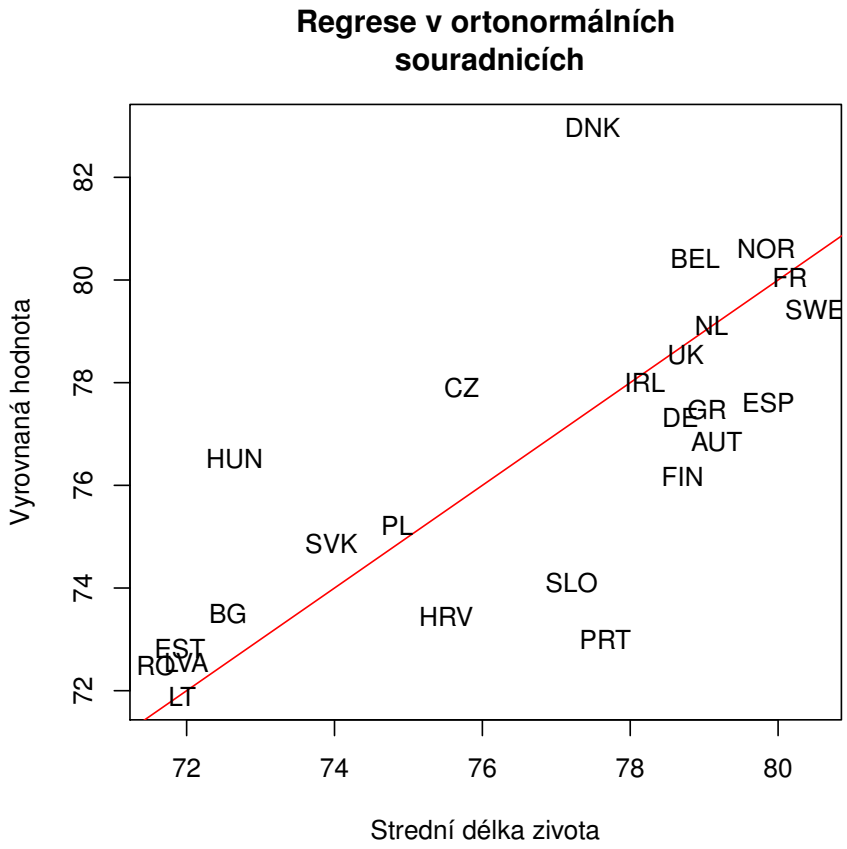
	ortonormální	ortogonální	B-koeficienty
X.mechyr	1.676	1.006	0.115
X.slinivka	1.405	0.843	0.092
X.strevo	2.468	1.481	0.174
X.zaludek	- 5.548	- 3.330	- 0.720

Tabulka 3: Přehled regresních koeficientů pro první příklad

Díky této tabulce se nám nabízí další možnosti interpretace. Z hodnot odhadů regresních koeficientů pro data vyjádřená v ortogonálních souřadnicích je patrné, že vysvětlovaná proměnná se se zvýšením hodnoty u `mechyr` o 1 (aneb zdvojnásobením podílu rakoviny měchýře vůči ostatním druhům rakoviny v průměru) zvýší jen o 1.006. Podobně nepatrné přírůstky při jednotkovém zvýšení příslušné proměnné evidujeme u souřadnic `slinivka` a `strevo`. Naproti tomu u proměnné `zaludek` se při jejím jednotkovém zvýšení vysvětlovaná proměnná sníží

o celých 3.330 a má tedy z uvažovaných vysvětlujících proměnných největší vliv na střední délku života ve zkoumaných zemích. Dodejme, že tato interpretace platí pouze v případě, kdy jsou ostatní souřadnice fixní v rámci daného ortogonálního systému souřadnic. Hodnoty B-koeficientů nás potom jenom utvrzují v tom, že největší vliv na vysvětlovanou proměnnou má souřadnice `zaludek`, jelikož 0.720 značně převyšuje všechny B-koeficienty z ostatních regresních modelů (záporné znaménko zde má pouze význam určení směru vlivu).

Na závěr prvního příkladu si ukažme graf, kde proti sobě postavíme původní střední délky života a jejich vyrovnané hodnoty tak, jak jsme je obdrželi pomocí uvažovaného regresního modelu. Z něj můžeme vidět, že regresní model kvalitně zachytil vztahy mezi vysvětlujícími veličinami a veličinou vysvětlovanou a také některá odlehlá pozorování (DNK, PRT, ...).



5.2. Příklad 2: Struktura vývozu zboží států OECD

Ve druhém příkladu věnujme pozornost společnosti OECD (Organizace pro hospodářskou spolupráci a rozvoj), jež čítá momentálně 34 zemí světa. Budeme sledovat roční data (2012) exportu států OECD v jednotlivých kategoriích zboží do celého světa [10]. Tyto kategorie dle oficiálního členění OECD jsou polotovary, domácí spotřeba, kapitálové (průmyslové) zboží, nejednoznačné zboží (zboží, které je zařaditelné do více kategorií současně) a ostatní (zboží, které nelze zařadit do žádné kategorie). Hodnoty exportu jsou uvedeny v tisících amerických dolarů. Naší vysvětlovanou proměnnou zde bude HDP na hlavu v jednotlivých státech [4]. Ptáme se tedy, jak HDP na hlavu ve státech OECD závisí na relativních příspěvcích sledovaných kategorií exportu.

Jelikož je postup výpočtu prakticky totožný jako v prvním příkladě, bez dalších komentářů provedeme již známé kroky:

```
> export = read.csv2("export.csv", row.names = 1)
```

```
> hdp_na_hlavu = c(67524.8, 48348.2, 44827.7, 52409.2, 15245.5,
  19670.4, 57636.1, 17102.2, 47243.7, 40908.3, 43931.7,
  22494.4, 12784.3, 44221.7, 48391.3, 32514.6, 35132.5,
  46679.3, 24454, 106022.8, 9817.8, 49128.1, 38678.4,
  99635.9, 12876.5, 20732.6, 17151.2, 22488.4, 28992.6,
  57134.1, 83295.3, 10660.7, 41053.7, 51495.9)
```

Z důvodu velkého rozsahu uvádíme výchozí data `export` v příloze. Dále:

```
> library(robCompositions)
```

```
> lmCoDaX(hdp_na_hlavu, export, method = "classical")
```

Obratem obdržíme tabulku regresní analýzy naší úlohy s využitím ortonormálních souřadnic (viz tabulka 4).

```
$ilr

Call:
lm(formula = y   ., data = d)

Residuals:
    Min      1Q  Median    3Q    Max
 -29514  -9085   -1340  6271  57342

Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)         38076       13386    2.844   0.00808  **
 X.polotovary        12697        7572    1.677   0.10432
 X.domaci           -11288        5516   -2.047   0.04986  *
 X.kapitalove        -6225        6938   -0.897   0.37693
 X.nejednoznacne     -2631        5006   -0.526   0.60320
 X.ostatni            7447        2487    2.995   0.00557  **
 ---
 Signif. codes:  0    '***'  0.001  '**'  0.01  '*'  0.05  '.'
                 0.1  ' '   1

Residual standard error: 20420 on 28 degrees of freedom
Multiple R-squared: 0.3489, Adjusted R-squared: 0.2591
F-statistic: 3.885 on 4 and 29 DF, p-value: 0.01204
```

Tabulka 4: Výstup regresní analýzy druhého příkladu

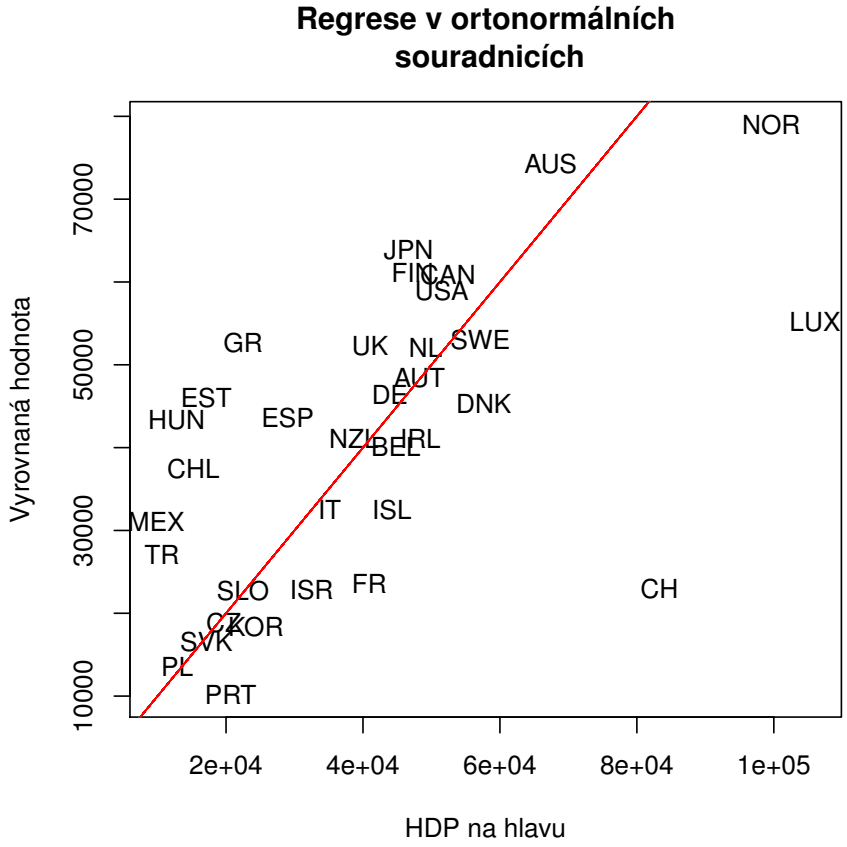
Také pro tento příklad srovnáme hodnoty regresních koeficientů při regresní analýze v ortonormálních a ortogonálních souřadnicích spolu s B-koeficienty.

	ortonormální	ortogonální	B-koeficienty
X.polotovary	12697	7871.755	0.304
X.domaci	- 11288	- 6998.218	- 0.386
X.kapitalove	- 6225	- 3859.311	- 0.161
X.nejednoznacne	- 2631	- 1631.140	- 0.106
X.ostatni	7447	4616.914	0.474

Tabulka 5: Přehled regresních koeficientů pro druhý příklad

Interpretujme výsledky druhého příkladu: v tabulce 4 vidíme, že dle testu statistické významnosti je nejvýznamější koeficient `X.ostatni`. Dále se ukázal jako významný koeficient `X.domaci`, jehož p-hodnota jen velmi těsně nepřekročila hladinu spolehlivosti $\alpha = 0.05$. Za povšimnutí také stojí p-hodnota u koeficientu `X.polotovary`, který nám sice R neoznačil jako statisticky významný pro žádnou rozumnou volbu α, nicméně hodnotu $\alpha = 0.1$ překročila příslušná p-hodnota jen velmi těsně. Koeficient determinace, resp. upravený koeficient determinace, nám vyšel okolo 0.35, resp. 0.26. Fakt že náš regresní model dokázal vysvětlit minimálně 26% celkové variability dat je znovu vzhledem ke skutečnosti, že data jsou reálného charakteru, poměrně uspokojivý. Obdržená p-hodnota testu o souhrnné statistické významnosti regresních koeficientů nám opět napovídá, že s vysokou pravděpodobností nebudou všechny regresní koeficienty rovny nule současně. Co se týče B-koeficientů, jejich hodnoty odrážejí skutečnost, kterou jsem již nastínil výše při testech statistické významnosti parametrů. A sice to, že největší význam na tvaru regrese má koeficient `X.ostatni` (potažmo příslušná souřadnice), následovaný koeficientem `X.domaci`, ovšem i koeficient `X.polotovary` má nezanedbatelný vliv. Z toho nám plyne ponaučení, abychom striktně neodsuzovali koeficienty, jejichž p-hodnota testu významnosti překročí stanovenou hladinu testu, ale abychom si také všímali, jak moc ji překročí.

Jako poslední výstup z druhého příkladu tu máme opět porovnání hodnot vysvětlované veličiny s hodnotami vyrovnanými pomocí regresního modelu v ortonormálních souřadnicích. Můžeme si přitom povšimnout několika odlehlých hodnot (CH, LUX, NOR), jejichž postavení v rámci studovaného problému by jistě stálo za hlubší reflexi příslušnými odborníky.



Na úplný závěr ještě jedna poznámka k B-koeficientům: v kapitole 1.2 jsme si řekli, že součet absolutních hodnot B-koeficientů v modelu je roven jedné. Jak ale vidíme, u druhého příkladu je tato vlastnost porušena. Vše se dá ovšem jednoduše vysvětlit, uvědomíme-li si, jak se zde tyto B-koeficienty tvoří. Z kapitoly 3.2 víme, že konstruujeme D regresních modelů tak, aby v každém z nich byla na první pozici jiná vysvětlující proměnná (kompoziční složka), jejíž regresní koeficient (resp. příslušná ortonormální souřadnice) je v tom případě jediný interpretovatelný. V každém z těchto D modelů je součet absolutních hodnot B-koeficientů skutečně roven jedné, ale napříč modely toto neplatí. Do naší tabulky pak vypisujeme jen B-koeficienty příslušné daným $\gamma_1^{(k)}$, které jsou tedy všechny z odlišných modelů. Tento fakt se neprojevil na prvním příkladu, kde absolutní hodnoty B-koeficientů byly v součtu přibližně rovny jedné, což bylo ovšem pouze dílem náhody.

Závěr

V práci jsme se zabývali tím, jak si nejlépe poradit s úlohou lineární regrese v případě, kdy máme vysvětlující proměnné kompozičního typu. Po úvodním seznámení se s teorií regresní analýzy jsme si řekli, co jsou to kompoziční data a jaké mají vlastnosti, včetně možnosti vyjádřit je v ortonormálních souřadnicích (vzhledem k Aitchisonově geometrii na simplexu). Tato vlastnost se následně ukázala jako velmi důležitá pro konstrukci regresního modelu s kompozičními vysvětlujícími proměnnými. Důraz jsem kladl na interpretovatelnost regresních koeficientů, k čemuž se kromě známých B-koeficientů jevila jako vhodná též volba ortogonálních souřadnic v regresním modelu; odhady regresních parametrů pak dostaneme jako násobky odhadů parametrů z modelu v ortonormálních souřadnicích. Další výhodou je i skutečnost, že s ortogonálními souřadnicemi neztrácíme možnost testovat významnosti regresních parametrů. Jelikož jsou veškeré výpočty časově náročné, představil jsem možnost provést uvedenou regresní analýzu pomocí statistického softwaru R, což jsem následně názorně předvedl na dvou příkladech.

Této práci vděčím za rozšíření vědomostí v oblasti regresní analýzy, zejména ale za seznámení se s pojmem kompoziční data, což bylo pro mě naprosto nové téma. Také jsem získal cenné zkušenosti ohledně práce v R a v TEXu, jež jsem předtím znal jen velmi okrajově. Nejtěžší pro mě přitom bylo osvojit si teorii kompozičních dat, zvláště pak jejich vyjádření v ortonormálních souřadnicích. Při pohledu zpět rozhodně nelituji výběru tohoto tématu, i když jsem o něm na začátku příliš vědomostí neměl.

Doufám, že se mi podařilo vysvětlit vše srozumitelně a že tato práce dokáže pomoci vnést světlo do problému regresní analýzy s kompozičními vysvětlujícími proměnnými těm, kteří se rozhodnou touto problematikou zabývat.

Literatura

[1] Countries Compared by Health > Life expectancy at birth, total > Years. International Statistics at NationMaster.com [online]. [cit. 2015-03-02]. Dostupné z: http://www.nationmaster.com/country-info/stats/Health/Life-expectancy-at-birth,-total/Years#2004.

[2] Eurostat - Data Explorer [online]. [cit. 2015-03-02]. Dostupné z: http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do.

[3] Fišerová, E.: *Lineární statistické modely (1. vydání).* Vydavatelství Univerzity Palackého, Olomouc, 2013, ISBN: 978-80-244-3402-5.

[4] GDP per capita (current US$) | Data | Table [online]. [cit. 2015-03-16]. Dostupné z: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD.

[5] Hron, K., Filzmoser, P., Thompson, K.: *Linear regression with compositional explanatory variables.* Journal of Applied Statistics, ročník 39, číslo 5, stránky 1115-1128, 2012.

[6] Hron, K., Kunderová, P.: *Základy počtu pravděpodobnosti a metod matematické statistiky (1. vydání).* Vydavatelství Univerzity Palackého, Olomouc, 2013, ISBN: 978-80-244-3396-7.

[7] Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P., Vančáková, J.: *Time budget analysis using logratio methods.* Odesláno 2015.

[8] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R.: *Lecture Notes on Compositional Data Analysis.* 2007.

[9] Pawlovsky-Glahn, V., Buccianti, A.: *Compositional data analysis: Theory and applications.* Wiley, Chichester, 2011, ISBN: 978-0470711354.

[10] STAN Bilateral Trade in Goods by Industry and End-use (BTDIxE), ISIC Rev.3 [online]. [cit. 2015-03-16]. Dostupné z: http://stats.oecd.org/Index.aspx?DataSetCode=BTDIXE_I3.

[11] van den Boogaart, K. G., Tolosana-Delgado, R.: *Analyzing Compositional Data with R.* Springer, 2013, ISBN: 978-3-642-36808-0.

Příloha

Datový soubor ke druhému příkladu (struktura exportu v zemích OECD).

```
> export
                   polotovary      domaci   kapitalove nejednoznacne     ostatni
Austrálie        216321785.86 19537887.232 6669448.704     6676966.4 7036819.968
Rakousko          91332689.92 20919015.424 28060985.344 12265110.528 6243148.288
Belgie           262169001.98   67877294.08 33929953.28 77809975.296 5068199.936
Kanada           313165283.33 35933323.264  33817907.2  55206154.24 15258232.832
Chile             61292529.28 14694136.832   926774.912   693800.32    39742.36
Česká republika  87911858.176 18906353.664 20631128.064 28802828.288 170571.024
Dánsko           50525855.744 28146331.648 14872912.896 4783556.608 7797306.88
Estonsko          10859960.32  3184025.344  3268616.96  360160.544 484693.472
Finsko           50910232.576  3756952.576 13462232.064 2958502.144 1886570.752
Francie          283915714.56 114337210.37 103008927.74 52342956.032 2970875.392
Německo          695615553.54 161782087.68 285814489.09  208429056 64543031.296
Řecko             24209025.024   7551229.44 1153876.224 1553405.696 712173.056
Maďarsko          54548152.32 16307815.424 12147893.248 16896601.088 3105550.848
Island             2697307.392 2021536.896  179421.552  133060.344 32115.572
Irsko            68170981.376 18969270.272 7010517.504  22200678.4 1418244.48
Izrael           26980145.152  5061319.168 8299058.176 22680203.264  119908
Itálie           259875913.73 120931434.5 83096723.456 30072365.056 7552423.424
Japonsko         440718065.66 23708016.64 196325605.38 102298460.16 35517431.808
Jižní Korea      349174857.73 19656665.088 118573572.1  60357365.76 91980.632
Lucembursko       9961078.784  1769420.416 1031975.424  657127.488 327943.936
Mexiko           192092258.3 55836901.376 65844318.208 53570306.048 3298767.104
Nizozemsko       300236210.18 101168726.02 73013239.808 35538919.424 44720820.224
Nový Zéland      15104881.664 18588018.688 1922848.768  319978.016 1368959.232
Norsko           137809870.85 10549133.312 6473428.48    982944 5184098.304
Polsko           97132314.624 48748077.056 20597135.36 12885202.944 240869.472
Portugalsko      33811226.624 15450963.968  5409843.2  3671097.856 35853.48
Slovensko        41256361.984 14334681.088 6959574.016 17211719.68 104659.216
Slovinsko        15855156.224  4010620.672 2581084.16 4588933.632 44227.064
Španělsko        145086119.94 66346160.128 26843746.304 34882920.448 12777491.456
Švédsko          102946578.43 19119740.928 27566374.912 14987455.488 8135880.704
Švýcarsko        108097282.05 55113191.424 26607343.616 34821079.04 1309863.168
Turecko          85476253.696 45804408.832  13700828.16 6770950.656 784209.152
Velká Británie   250263584.77 69866110.976 53313605.632 72246263.808 35536187.392
USA              880162111.49 159919833.09 234236608.51 132854775.81 138391830.53
```