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

DIPLOMOVÁ PRÁCE

Ortogonální regrese pro kompoziční data

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2012

Vypracoval:
Bc. Martin Petera
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořil tuto diplomovou práci samostatně za vedení RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 31. března 2012

Poděkování

Rád bych na tomto místě poděkoval vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále si zaslouží poděkování můj počítač, že vydržel moje pracovní tempo, a typografický systém TEX, jímž je práce vysázena.

Obsah

Úvod

Řada statistiků i odborníků z oblastí aplikací v poslední době věnuje pozornost kompozičním datům (kompozicím). Ta nezávisí na znalosti absolutních hodnot, ale pouze na relativní informaci v datech obsažené. Při práci s těmito daty nás tedy zajímají pouze podíly mezi jednotlivými složkami datového souboru.

Musíme dávat pozor na to, že při práci s kompozičními daty selhávají standardní postupy známé z euklidovské metriky. Kompozice se totiž ve skutečnosti řídí vlastní, tzv. *Aitchisonovou geometrií*.

Pracujeme-li s kompozičními daty, jeví se jako nejvhodnější regresní metoda tzv. *ortogonální regrese*. V situaci, kdy pracujeme s rovinnými daty, prokládá tato metoda daty přímku tak, že součet čtvercových vzdálenosti dat k odhadované přímce je minimální. Na rozdíl od klasické *metody nejmenších čtverců*, v níž minimalizujeme vertikální vzdálenost od přímky, pro ortogonální regresi hraje hlavní roli ortogonální vzdálenost dat od přímky.

S výhodou lze tedy v případě ortogonální regrese použít tzv. *lineární regresní modely s podmínkou II. typu*. Navíc dodatečné předpoklady na normalitu umožňují zkonstruovat konfidenční oblasti a současně testovat hypotézy.

Hlavním cílem diplomové práce je zpracovat uvedenou problematiku. V první kapitole jsou stručně nastíněny základní pojmy vztahující se ke kompozičním datům. Základní lineární modely měření s názornými ilustračními příklady osvětluje druhá část této práce. Nelineární zákon šíření chyb a linearizaci modelu popisuje následující kapitola.

Stěžejní teoretickou částí je čtvrtá kapitola, v níž budou podrobně popsány postupy při odhadu regresní přímky pomocí metody ortogonální regrese (kalibrační přímky) s využitím teorie lineárních modelů. Popsán bude samotný algoritmus ortogonální regrese. Chybět nesmí ani intervalové odhady, testování hypotéz a konfidenční oblasti. V předposlední kapitole se budeme zabývat srovnáním ortogonální regrese a metody hlavních komponent.

Důležitou součástí diplomové práce je aplikace poznatků na příklady využívající reálná data. Konkrétně budeme zkoumat složení půdy ve členských státech Evropské unie (podíl složek orné půdy, lesní plochy a půdy určené k pěstování trvalých plodin). Druhý příklad se zabývá poslechovostí rádií napříč všemi okresy v České republice.

1 Kompoziční data

Základem této práce jsou kompoziční data. Jelikož informace o nich jsou podrobně rozpracovány v bakalářské práci [15], budou v následující kapitole uvedeny zejména základní definice a věty. Při psaní kapitoly bylo dále čerpáno ze zdrojů [13] a [14].

Definice 1.1. *Sloupcový vektor* $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ *se nazývá D-složková kompozice, jestliže všechny jeho složky jsou* kladná čísla *a nesou pouze* relativní informaci.

Jak tedy z výše uvedeného vyplývá, nikoli absolutní hodnoty, ale pouze podíly mezi jednotlivými složkami jsou pro nás v tomto případě relevantní. Jinými slovy řečeno, vektory $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ a $a\mathbf{x} = (ax_1, ax_2, \ldots, ax_D)'$ nám dávají totožnou informaci ($a > 0$). Abychom mohli kompoziční data snadněji interpretovat, lze je vyjádřit jako kladné vektory se součtem složek rovným kladné konstantě κ. Při volbě $\kappa = 1$ nebo 100 je takto kompozice $\mathbf{x}$ (bez ztráty informace) reprezentovaná ve formě proporcí nebo procentuálních podílů, což právě budeme používat.

Předchozí uvedené úvahy nás oprávněně vedou k několika důležitým definicím, které budou nyní uvedeny.

Definice 1.2. *Výběrový prostor kompozičních dat je simplex, podmnožina dimenze* $D - 1$ *reálného prostoru* $\mathbb{R}^D$*, definovaná jako*

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)', x_i > 0, \ \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Definice 1.3. *Pro každou D-složkovou kompozici*

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)' \in \mathbb{R}_+^D$$

($x_i > 0, \forall i = 1, 2, \ldots, D$*) je* uzávěrem *rozuměn vektor*

$$\mathcal{C}(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \cdots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right]'.$$

Definice 1.4. *Pro danou kompozici* $\mathbf{x}$ *obdržíme* podkompozici $\mathbf{x}_s$ *(obsahující s částí,* $s < D$*) aplikací operace uzávěru na podvektor* $(x_{i_1}, x_{i_2}, \ldots, x_{i_s})'$ *vektoru* $\mathbf{x}$*. Pro indexy* $i_1, \ldots, i_s$*, které určují vybrané složky kompozice* $\mathbf{x}$*, přitom platí* $1 \leq i_1 < \ldots < i_s \leq D$*.*

V celém textu diplomové práce jsou náhodné objekty i jejich realizace značeny stejnými symboly. Není třeba rozlišovat, jelikož nám kontext napoví, zda operujeme s teoretickým, nebo výběrovými charakteristikami objektů.

Simplex je speciálním případem výběrového prostoru, vzhledem k charakteru dat ovšem musíme operace založené na standardní vektorové algebře a euklidovské metrice upravit. Zadefinujeme analogické operace plynoucí ze známých ekvivalentů na reálném prostoru. Pro kompoziční data zavedeme *Aitchisonovu geometrii* s vlastnostmi euklidovského vektorového prostoru.

Definice 1.5. *Perturbací kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *rozumíme kompozici*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left[x_1 y_1, x_2 y_2, \ldots, x_D y_D\right].$$

Definice 1.6. *Mocninná transformace kompozice* $\mathbf{x} \in \mathcal{S}^D$ *konstantou* $\alpha \in \mathbb{R}$ *je definována jako*

$$\alpha \odot \mathbf{x} = \mathcal{C}\left[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right].$$

Věta 1.1. *Simplex společně s perturbací a mocninnou transformací* $(\mathcal{S}^D, \oplus, \odot)$ *tvoří reálný vektorový prostor.*

Důkaz: Viz [15], str. 12, Věta 2.1.

□

Následující definice, které zde budou uvedeny, představují základní operace určené pro kompoziční data. Vychází opět z analogie operací na elementárním euklidovském prostoru.

Definice 1.7. *Skalární součin dvou kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *lze definovat následovně:*

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

Definice 1.8. *Normu vektoru* $\mathbf{x} \in \mathcal{S}^D$ *definujeme*

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

Definice 1.9. *Vzdálenost mezi* $\mathbf{x}$ *a* $\mathbf{y} \in \mathcal{S}^D$ *je dána*

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

1.1 Vyjádření v souřadnicích

Tzv. *logratio transformace* (neboli transformace pomocí logaritmů podílů složek původní kompozice) byly navrženy pro zobrazení kompozice z $\mathcal{S}^D$ do $(D-1)$-rozměrného, resp. D-rozměrného reálného prostoru, kde již můžeme pro jejich zpracování užít standardních statistických metod. První z uvažovaných transformací, zavedených v literatuře [1] je *aditivní logratio (alr)*, ta ovšem nezachovává metrické vlastnosti kompozic. Proto byla navržena *centrovaná logratio (clr)* transformace.

Definice 1.10. *[15] Pro kompozici $\mathbf{x} \in \mathcal{S}^D$ jsou clr koeficienty složky vektoru $\boldsymbol{\xi} = \mathrm{clr}(\mathbf{x}) = (\xi_1, \xi_2, \ldots, \xi_D)'$, jediného vektoru splňujícího*

$$\mathbf{x} = \mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}(\exp(\boldsymbol{\xi})), \quad \sum_{i=1}^{D} \xi_i = 0.$$

Potom i-tý clr koeficient je

$$\xi_i = \ln \frac{x_i}{g(\mathbf{x})}.$$

Ačkoli clr transformace zachovává metrické vlastnosti kompozic, její nevýhodou skutečnost, že vede k singulární varianční matici, což je z hlediska statistické analýzy nežádoucí. Proto se v poslední době odborníci přiklání k *izometrické logratio (ilr) transformaci*. Klíčovým požadavkem je ortonormalita báze na simplexu, vzhledem k níž jsou ilr koeficienty odpovídajícími souřadnicemi. Více podrobností je uvedeno v [15].

V této diplomové práci bude cílem najít souřadnice ortogonální báze na simplexu, které umožní dobrou interpretaci výsledků ortogonální regrese trojsložkových kompozic.

1.2 Ortonormální souřadnice

Pro jejich konstrukci je zásadním požadavkem ortonormalita báze na simplexu. Nejprve je tedy třeba uvažovat ortonomální bázi $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ na simplexu $\mathcal{S}^D$ a dále matici $\boldsymbol{\Psi}_{(D-1),D}$, jejíž řádky tvoří vektory $\mathrm{clr}(\mathbf{e}_i)$. Platí, že $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \delta_{ij}$, kde δ_{ij} je tzv. Kroneckerovo delta, $\delta_{ij} = 1$ pro $i = j$, jinak $\delta_{ij} = 0$.

V druhém kroku vyjádříme kompozici $\mathbf{x} \in \mathcal{S}^D$ jako

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} z_i \odot \mathbf{e}_i, z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a; \tag{1}$$

tak $\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = (z_1, \ldots, z_{D-1})'$, vektor souřadnic $\mathbf{z}$ vzhledem k vybrané ortonormální bázi.

Izometrická logratio transformace je označována jako předpis, kterým vyjádříme kompozici $\mathbf{x}$ v souřadnicích $\mathbf{z}$, tedy jako zobrazení z $\mathcal{S}^D$ do $\mathbb{R}^{D-1}$. Vztah mezi ilr souřadnicemi $\mathbf{z}$ a clr transformací je dán vztahem

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \boldsymbol{\Psi} \cdot \mathrm{clr}(\mathbf{x}).$$

1.3 Vlastnosti ilr transformace plynoucí z izometrie

Jak je již uvedeno v [15], existuje několik možností způsobu při definování ortonormální báze na simplexu. Hlavním kritériem je interpretace kompozice v souřadnicích.

Situaci v případě trojsložkových kompozic se budeme zabývat dále v této práci. Jednou oblíbenou volnou ortonormálních souřadnic je

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{\sqrt[D-i]{\Pi_{j=i+1}^{D} x_j}}{x_i}, \; i = 1, \ldots, D-1. \tag{2}$$

1.4 Inverzní ilr transformace

Celkem můžeme vytvořit D různých ilr transformací odvozených z (2), každou tvoří $D-1$ nových proměnných. Každou D-složkovou kompozici $\mathbf{x} = (x_1, \ldots, x_D)'$ lze dle [12] asociovat s jinou kompozicí $\mathbf{x}^{(l)} = (x_1^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$, což je výsledek permutace složek $\mathbf{x} = (x_1, \ldots, x_D)'$, kde l-tá složka je přesunuta na první pozici, tj. $\mathbf{x}^{(l)} = (x_l, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_D)'$. Ilr transformace (2) převádí kompozice $\mathbf{x}^{(l)}$ do $(D-1)$-rozměrného reálného vektoru $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})', l = 1, \ldots, D$. Výsledný vztah pro ilr transformace bude ve tvaru

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{\sqrt[D-i]{\Pi_{j=i+1}^{D} x_j^{(l)}}}{x_i^{(l)}}, \; i = 1, \ldots, D-1.$$

Inverzní transformace vektoru $\mathbf{z}^{(l)}$ do původní (permutované) kompozice $\mathbf{x}^{(l)}$ je dána vztahy

$$x_1^{(l)} = \exp\left(-\frac{\sqrt{D-1}}{\sqrt{D}} z_1^{(l)}\right), \tag{3}$$

$$x_i^{(l)} = \exp\left(\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)} - \frac{\sqrt{D-1}}{\sqrt{D-i+1}} z_i^{(l)}\right), i = 2, \ldots, D-1, \tag{4}$$

$$x_D^{(l)} = \exp\left(\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)}\right). \tag{5}$$

2 Základní lineární modely měření

Sestavujeme-li regresní modely, zajímá nás vždy určitá forma vztahu mezi vysvětlovanou a vysvětlující proměnnou. Odhady parametrů v lineárních regresních modelech se provádí metodou nejmenších čtverců.

V této práci se ovšem budeme zabývat ortogonální regresí, jež se jeví jako vhodnější pro práci s kompozičními daty. Přesto i zde lze použít speciálního typu lineárního regresního modelu. V následující kapitole si představíme jeho systematické odvození, při kterém bylo použito literatury [10].

2.1 Zpracování dat, vytvoření modelu

Při práci s lineárními regresními modely nejprve zkoumáme určitý objekt, na němž následně provádíme opakovaná měření, čímž dostáváme soubor naměřených hodnot. Cílem je tak určit odhady neznámých parametrů (obecně neznámé konstanty), odhadové funkce. Jakmile známe odhad, je nutné znát statistické charakteristiky odhadů.

Zpracování probíhá ve třech etapách. V první etapě vytvoříme tzv. *teoretický / deterministický model*. Úkolem je určit vztah mezi veličinou, kterou umíme měřit, a parametrem, tedy tím, co chceme změřit. Začínáme vždy nejjednodušším modelem. Dále je třeba určit vztah mezi teoretickými hodnotami parametrů $\beta_1, \ldots, \beta_k$,

$$Z_i = f_i(\beta_1, \ldots, \beta_k),\ i = 1, \ldots, p,$$

kde k značí počet parametrů a p je počet pozorovatelných veličin. V některých úlohách existují podmínky, jimž musí parametry vyhovovat,

$$g_j(\beta_1, \ldots, \beta_k) = 0,\ j = 1, \ldots, q,$$

kde q představují počet podmínek na parametry. Vektorově zapíšeme:

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix};\ \mathbf{f}(\boldsymbol{\beta}) = \begin{pmatrix} f_1(\boldsymbol{\beta}) \\ f_2(\boldsymbol{\beta}) \\ \vdots \\ f_p(\boldsymbol{\beta}) \end{pmatrix};\ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix};$$

$$\mathbf{g}(\boldsymbol{\beta}) = \begin{pmatrix} g_1(\boldsymbol{\beta}) \\ g_2(\boldsymbol{\beta}) \\ \vdots \\ g_q(\boldsymbol{\beta}) \end{pmatrix};\ \mathbf{0} = \left.\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right\} q.$$

V druhé etapě vytváříme *stochastický model*. Výsledky měření přímo pozorovatelných veličin jsou zatíženy chybami měření (vždy při každém rozdělení). Jsou organickou součástí procesu měření. Obvykle provádíme n_i-krát měření veličiny

$Z_i, i = 1, \ldots, p$. Výsledky opakovaného měření budeme pokládat za hodnoty náhodných veličin (zapíšeme jako vektor): $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$.

Nyní p-krát měříme výsledky a zapíšeme do vektoru, označme si

$$n = n_1 + n_2 + \ldots + n_p.$$

Výsledky všech provedených měření odpovídají hodnotám náhodného vektoru, modelují tak celý experiment.

Označme

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_p \end{pmatrix}, \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}.$$

Tento náhodný vektor je složen z podvektorů $\boldsymbol{\varepsilon}_i$, chyb vznikajících při n_i násobném měření Z_i, $i = 1, \ldots, p$. Dodejme, že jednotlivá měření se mohou vzájemně lišit.

Dále předpokládáme, že $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$. n-složkový náhodný vektor $\boldsymbol{\varepsilon}'$ obvykle interpretujeme jako vektor chyb měření.

Nyní si pro ilustraci a bližší vysvětlení uvedeme příklad stochastického modelu:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_p \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} f_1(\boldsymbol{\beta}) \\ \vdots \\ \mathbf{1}_{n_p} f_p(\boldsymbol{\beta}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_p \end{pmatrix} = \mathbf{J}\mathbf{f}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon},$$

kde

$$\mathbf{1}_{n_i} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}; \mathbf{J} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ & & \cdots & \\ \mathbf{0}_{n_p} & \mathbf{0}_{n_p} & \cdots & \mathbf{1}_{n_p} \end{pmatrix};$$

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} f_i(\boldsymbol{\beta}) \\ \vdots \\ f_i(\boldsymbol{\beta}) \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}.$$

Je třeba upozornit, že platí $\mathsf{E}(\mathbf{Y}) = \mathbf{J}\mathbf{f}(\boldsymbol{\beta})$, přičemž $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)'$ je neznámý vektorový parametr.

O varianční matici vektoru $\mathbf{Y}$ budeme dále předpokládat vždy jednu z následujících možností:

$$\mathrm{var}(\mathbf{Y}) = \begin{cases} \boldsymbol{\Sigma} & \text{známá, pozitivně definitní matice;} \\ \sigma^2\mathbf{V} & \sigma^2 \text{ je neznámý parametr, } \mathbf{V} \text{ je známá, poz.definitní;} \\ \sum_{i=1}^{r} \theta_i \mathbf{V}_i & \theta_1, \ldots, \theta_r \text{ - neznámé varianční komponentní parametry;} \\ & \mathbf{V}_1, \ldots, \mathbf{V}_r \text{ - pozitivně definitní matice, které známe;} \\ \boldsymbol{\Sigma} & \text{neznámá.} \end{cases}$$

V závěrečné, třetí etapě přistoupíme k vytvoření samotného statistického modelu. Úkolem v této fázi je stanovit odhady $\boldsymbol{\beta}$ a získat příslušné číselné charakteristiky. Užitím metod teorie odhadu (nejčastěji již zmíněnou metodu nejmenších čtverců) tedy určíme odhadové funkce neznámých parametrů

$$\widehat{\boldsymbol{\beta}}(\mathbf{Y}) = (\hat{\beta}_1(\mathbf{Y}), \ldots, \hat{\beta}_k(\mathbf{Y}))';$$

v praxi dosadíme pozorované hodnoty $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ a dostaneme realizaci náhodného vektoru $\widehat{\boldsymbol{\beta}}(\mathbf{Y})$.

Následující část druhé kapitoly nabídne přehled šesti základních lineárních modelů, z nichž stěžejní bude pro naše potřeby právě model poslední. Pro snadnější pochopení uvedené problematiky budou modely opatřeny krátkými ilustračními příklady.

Uvedení všech možných modelů není přitom samoúčelné. Mějme na paměti, že poslední, šestý model je logickým rozšířením předchozích. Na nejjednodušší model číslo 1 klademe jen minimální předpoklady, navíc dokáže zachytit pouze základní matematické problémy (například nedokáže operovat s parametry, které nedokážeme přímo změřit, resp. není schopen zachytit podmínky na regresní parametry). Postupným rozšiřováním uvedených lineárních modelů získáme větší možnosti pro využití těchto modelů, což budeme následně demonstrovat v případě ortogonální regrese.

2.2 Přehled základních lineárních modelů

V rámci teorie lineárních modelů se omezujeme na následující funkce a odhady,

- lineární funkce $f_i(\boldsymbol{\beta})$, $i = 1, \ldots, p$ - jedná se o lineární funkce zkoumaných parametrů. Jako příklad uveďme situaci kvadratické regrese, v tomto případě se jedná o lineární funkci tří parametrů, přestože vysvětlující proměnná definuje kvadratickou funkci.

- lineární funkce $g_j(\boldsymbol{\beta})$, $j = 1, \ldots, q$ - značí opět lineární funkce parametrů. Tyto funkce ovšem vymezují vzájemné vztahy mezi jednotlivými sledovanými parametry. Příkladem budiž podmínka na součet úhlů v trojúhelníku, který musí být roven 180°.

- lineární odhady $\widehat{\boldsymbol{\beta}}(\mathbf{Y})$ - lineární odhady zkoumaného parametru, jež se snažíme získat.

Restrikce na lineární modely neznamená velké zúžení okruhu problémů řešitelných v této třídě modelů, neomezuje možnosti aplikace. Mnohé z nelineárních funkcí lze totiž linearizovat tak, že funkci $\mathbf{f}(\boldsymbol{\beta})$, resp. $\mathbf{g}(\boldsymbol{\beta})$ rozvineme do Taylorovy řady a zanedbáme členy druhého a vyšších řádů. Postup si podrobně ukážeme o něco později, ve třetí kapitole této práce.

Předpoklady uváděné na hodnosti matic zaručují odhadnutelnost (to znamená, že existuje hledaný nejlepší nestranný lineární odhad) každé složky $\boldsymbol{\beta}$ a současně existenci maticových inverzí.

- **Model 1** - Přímé měření skalárního parametru
  - Při sestavování prvního modelu nejprve vytvoříme teoretický model, ten bude v jednoduchém tvaru $Z = \beta$, $k = p = 1$.

    Stochastický model představuje n-krát opakovaná měření $\mathbf{Y}$. Tato měření jsou prováděná přímo, nezahrnujeme do sestavování žádné početní operace. Pro představu použití modelu uveďme jako příklad situaci, kdy měříme výšku nějakého objektu.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta \\ \vdots \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \mathbf{1}_{(n)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

    Předpoklady na tento model jsou tyto: $\mathsf{E}(\mathbf{Y}) = \mathbf{1}_n\beta$, $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}$, kde $\boldsymbol{\Sigma}$ je pozitivně definitní matice a $h(\boldsymbol{\Sigma}) = n$.

- **Model 2** - Nepřímé měření vektorového parametru
  - Vytvořený teoretický model zapíšeme $\mathbf{Z} = \mathbf{A}\boldsymbol{\beta}$. Matice $\mathbf{A}$ určuje lineární funkce parametru $\boldsymbol{\beta}$. Zatížíme tento model chybami a vytvoříme stochastický model.

    Stochastický model zapíšeme ve tvaru $\mathbf{Y} = \mathbf{JA}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, doplněný je o následující předpoklady $\mathsf{E}(\mathbf{Y}) = \mathbf{JA}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$; $h(\mathbf{JA}) = h(\mathbf{X}) = k < n$; $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}$.

    Model využíváme v případě, kdy jsou regresní parametry svázány nějakým lineárním vztahem. Příkladem nechť je časová řada $y(t) = \beta_1 + \beta_2 t + \beta_3 t^2$, realizovaná v jednotlivých bodech $t_1, t_2, \ldots, t_n$, vždy ale jen jednou. Tento model se také často vyskytuje v případě analýzy nějaké regulární závislosti.

- **Model 3** - Přímé měření vektorového parametru s podmínkou

  – V prvním kroku zapíšeme teoretický model. Tentokrát bude ve tvaru $\mathbf{Z} = \boldsymbol{\beta}$, navíc s podmínkou $\mathbf{b} + \mathbf{B}\boldsymbol{\beta} = \mathbf{0}$, $\mathbf{B}$ je typu $q \times k$ a $\mathbf{b}$ je typu $q \times 1$, přičemž

  $$\boldsymbol{\beta} \in \tau \subset \mathbb{R}^k,\ \tau = \left\{\mathbf{u} \in \mathbb{R}^k : \mathbf{b} + \mathbf{B}\mathbf{u} = \mathbf{0}\right\}.$$

  Stochastický model má potom tvar $\mathbf{Y} = \mathbf{J}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{b} + \mathbf{B}\boldsymbol{\beta} = \mathbf{0}$. Dále předpokládáme, že $h(\mathbf{J}) = k < n$ a současně $h(\mathbf{B}) = q < k$. Dále je nutné myslet na fakt, že $\mathsf{E}(\mathbf{Y}) = \mathbf{J}\boldsymbol{\beta}$, $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}$ a je zároveň pozitivně definitní.

  Model se využívá v případě, kdy jsme sice schopni všechny sledované parametry měřit, ale jsou navíc vázány logickou podmínkou. Pro ilustraci uveďme opět stručný příklad. Mějme dány úhly v rovinném trojúhelníku. Opakovaně je měříme a získáme parametry (α, β, γ). Víme, že platí $\alpha + \beta + \gamma = 180° = \pi$ a tedy logicky také $\alpha + \beta + \gamma - \pi = 0$.

- **Model 4** - Neúplné přímé měření vektorového parametru se systémem podmínek

  – Model najde uplatnění ve chvíli, kdy jsme schopni část složek parametrů změřit přímo a část nikoli. Ovšem i tyto nepřímo měřitelné parametry nás zajímají a potřebujeme je znát. Právě proto existuje model číslo 4, mějme

  $$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \begin{matrix} \}k_1 \\ \}k_2 \end{matrix} \text{ složek parametru.}$$

  V našem případě se $\mathbf{Z} = \boldsymbol{\beta}_1$. Zatímco složky $\boldsymbol{\beta}_1$ jsou přímo pozorovatelné (tudíž měřitelné), složky $\boldsymbol{\beta}_2$ nelze přímo změřit, přesto nás zajímají a potřebujeme je pro naše další výpočty.

  Máme proto dán systém podmínek $\mathbf{b}+\mathbf{B}\boldsymbol{\beta}_1+\mathbf{C}\boldsymbol{\beta}_2 = \mathbf{0}$ (podmínka typu II, kde $\mathbf{C}\boldsymbol{\beta}_2 = \mathbf{0}$ neměříme). Dále jsou stanoveny podmínky regularity: $\mathbf{b}$ číselný vektor pro $(q \times 1)$, matici $\mathbf{B}$ (rozměry $q \times k_1$) a matici $\mathbf{C}$ typu $q \times k_2$ platí, že hodnost blokově rozdělené matice $h([\mathbf{B}, \mathbf{C}]) = q < k_1 + k_2$, přičemž matice $\mathbf{C}$ má plnou sloupcovou hodnost $h(\mathbf{C}) = k_2 < q$.

  $$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \in \tau = \left\{ \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \in \mathbb{R}^{k_1+k_2} : \mathbf{b} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{v} = \mathbf{0} \right\}.$$

  V druhém kroku vytvoříme stochastický model. Ten je tentokrát stanoven ve tvaru $\mathbf{Y} = \mathbf{J}\boldsymbol{\beta}_1+\boldsymbol{\varepsilon}$, $\mathsf{E}(\mathbf{Y}) = \mathbf{J}\boldsymbol{\beta}_1$, $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}$, $h(\mathbf{J}) = k_1 < n$.

- **Model 5** - Nepřímé měření vektorového parametru se systémem podmínek

  – Model se využívá ve stejném případě jako v předchozím modelu, tedy tehdy máme-li část parametrů měřitelných přímo a část nepřímo, ale chceme znát i ty nepřímo měřitelné. Tentokrát je ovšem model svázán logickými podmínkami. Teoretický model vyjádříme jako

  $$\mathbf{Z} = \mathbf{A}\boldsymbol{\beta}, \mathbf{b} + \mathbf{B}\boldsymbol{\beta} = \mathbf{0}, \boldsymbol{\beta} \in \tau \subset \mathbb{R}^k, \tau = \{\mathbf{u} \in \mathbb{R}^k : \mathbf{b} + \mathbf{B}\mathbf{u} = \mathbf{0}\}.$$

  Stochastický model posléze sestavíme ve tvaru: $\mathbf{Y} = (\mathbf{JA})\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, v patrnost vedeme předpoklady $\mathsf{E}(\mathbf{Y}) = \mathbf{JA}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}, \mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}, h(\mathbf{JA}) = h(\mathbf{X}) = k < n$.
  Uveďme ještě stručný ilustrační příklad pro nastínění použití předposledního lineárního modelu. V rovinném trojúhelníku je naším cílem změřit úhly α, β. Oba jsou měřitelné přímo, úhel γ však přímo změřit není možné. Hledáme tedy odhad vektoru parametrů
  $\boldsymbol{\beta} = (\alpha, \beta, \gamma)', \boldsymbol{\beta}_1 = (\alpha, \beta)', \beta_2 = \gamma$ pro $k = 3, k_1 = 2, k_2 = 1$. Podmínku pro výše uvedený model stanovíme ve tvaru:

  $$-\pi + (1, 1)(\alpha, \beta)' + \gamma = 0.$$

- **Model 6** - Neúplné nepřímé měření vektorového parametru se systémem podmínek

  – Stejně jako v modelu číslo 4, také v tomto případě je parametr $\boldsymbol{\beta}$ složen ze dvou skupin složek, první lze přímo měřit, u druhé části toto není možné. Teoretický model tedy sestavíme ve tvaru: $\mathbf{Z} = \mathbf{A}\boldsymbol{\beta}_1$, se systémem podmínek $\mathbf{b} + \mathbf{B}\boldsymbol{\beta}_1 + \mathbf{C}\boldsymbol{\beta}_2 = \mathbf{0}$. Mějme na paměti, že

  $$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \in \tau = \left\{ \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \in \mathbb{R}^{k_1+k_2} : \mathbf{b} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{v} = \mathbf{0} \right\}.$$

  Zatížíme-li model chybami, získáme stochastický model. Ten nyní uvedeme ve tvaru $\mathbf{Y} = (\mathbf{JA})\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$. Současně přitom také platí (logická) podmínka $\mathbf{b} + \mathbf{B}\boldsymbol{\beta}_1 + \mathbf{C}\boldsymbol{\beta}_2 = \mathbf{0}$. Nesmíme opomenout ani podmínky regularity, $\mathsf{E}(\mathbf{Y}) = \mathbf{JA}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_1$. Platí tedy, že varianční matice $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma}$ je pozitivně definitní. Dále musí platit, že hodnost $h(\mathbf{JA}) = h(\mathbf{X}) = k < n$ a přitom $h([\mathbf{B}, \mathbf{C}]) = q < k_1 + k_2, h(\mathbf{C}) = k_2 < q$.

  – Pro ilustraci bude také při uvádění tohoto modelu nastíněn stručný příklad. Mějme rovinný trojúhelník, u něhož známe délku strany a. Naším úkolem je určit velikost všech úhlů tohoto trojúhelníku. K dispozici přitom máme měření strany b a úhlů α, β. Cílem příkladu je určit zbývající úhel.

Nejprve sestavíme, jako vždy u podobného příkladu, teoretický model. Položíme $Z_1 = \alpha$, $Z_2 = \beta$ a $Z_3 = b$. Důležité je, že strana b je přímo měřitelná, ale zároveň je funkcí parametrů α, β. Zmíněné veličiny nám jsou známé a máme jejich hodnoty. Ze sinové věty dostáváme

$$b = a \cdot \frac{\sin\beta}{\sin\alpha}.$$

Jak je zřejmé, funkci musíme linearizovat, abychom mohli pokračovat při následných výpočtech. Vhodným nástrojem k řešení daného problému je Taylorův rozvoj, věnovat se mu budeme ovšem až v následující kapitole.

Shrňme tedy, co dosud víme o parametru $\boldsymbol{\beta}$, do maticového zápisu,

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \boldsymbol{\beta}_1 = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \beta_2 = \gamma.$$

Právě tomuto, šestému modelu byla věnována největší pozornost. Důvodem je fakt, že bude stěžejním regresním modelem v této diplomové práci.

V následující kapitole se budeme dále věnovat modelu číslo 6, v němž měříme část parametrů nepřímo, regresní parametry jsou přitom navíc svázány podmínkami II. typu.

2.3 Nepřímá měření s podmínkou II. typu na parametry 1. řádu

Jak už nyní víme, tento model bude nadefinován ve tvaru

$$\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}_1, \boldsymbol{\Sigma}),\ \mathbf{b} + \mathbf{B}\boldsymbol{\beta}_1 + \mathbf{C}\boldsymbol{\beta}_2 = 0.$$

Budeme-li v tomto modelu uvažovat rovněž následující předpoklady pro hodnosti matic,

$$h(\mathbf{X}_{n,k}) = k_1 < n,\ h([\mathbf{B}_{q,k_1}, \mathbf{C}_{q,k_2}]) = q < k_1 + k_2,\ h(\mathbf{C}) = k_2 < q,$$

a současně (známá) varianční matice $\boldsymbol{\Sigma}$ modelu bude navíc pozitivně definitní, potom můžeme uvedený model nazvat regulární. Pro další výpočty je tento předpoklad nezbytný.

Následující věta odvozuje BLUE vektoru $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$. Zkratka BLUE vychází z anglického označení *Best Linear Unbiased Estimator*, tedy *nejlepší nestranný lineární odhad*.

Věta 2.1. *[10] BLUE vektoru $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$ je dán vztahem:*

$$\begin{pmatrix} \widehat{\widehat{\boldsymbol{\beta}}}_1 \\ \widehat{\widehat{\boldsymbol{\beta}}}_2 \end{pmatrix} = -\begin{pmatrix} (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}\mathbf{Q}_{11} \\ \mathbf{Q}_{21} \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{I} - (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'\mathbf{Q}_{11}\mathbf{B} \\ -\mathbf{Q}_{21}\mathbf{B} \end{pmatrix} \widehat{\boldsymbol{\beta}}_1,$$

kde $\widehat{\boldsymbol{\beta}_1} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$ (odhad nerespektující podmínku týkající se parametrů $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$).

Varianční matice BLUE odhadu má následující tvar:

$$\operatorname{var}\left[\begin{pmatrix} \widehat{\widehat{\boldsymbol{\beta}}}_1 \\ \widehat{\widehat{\boldsymbol{\beta}}}_2 \end{pmatrix} \middle| \boldsymbol{\Sigma}\right] = \begin{pmatrix} \operatorname{var}\left(\widehat{\widehat{\boldsymbol{\beta}}}_1\right), & \operatorname{cov}\left(\widehat{\widehat{\boldsymbol{\beta}}}_1, \widehat{\widehat{\boldsymbol{\beta}}}_2\right) \\ \operatorname{cov}\left(\widehat{\widehat{\boldsymbol{\beta}}}_1, \widehat{\widehat{\boldsymbol{\beta}}}_2\right), & \operatorname{var}\left(\widehat{\widehat{\boldsymbol{\beta}}}_2\right) \end{pmatrix},$$

kde

$$\begin{aligned} \operatorname{var}\left(\widehat{\widehat{\boldsymbol{\beta}}}_1\right) &= (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} - (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'\mathbf{Q}_{11}\mathbf{B}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}, \\ \operatorname{cov}\left(\widehat{\widehat{\boldsymbol{\beta}}}_1, \widehat{\widehat{\boldsymbol{\beta}}}_2\right) &= -(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'\mathbf{Q}_{12}, \\ \operatorname{var}\left(\widehat{\widehat{\boldsymbol{\beta}}}_2\right) &= -\mathbf{Q}_{22}, \end{aligned}$$

$$\begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{B}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}' & , & \mathbf{C} \\ \mathbf{C}' & , & \mathbf{0} \end{pmatrix}^{-1}.$$

Důkaz: Viz [10], str. 129, Věta IV.4.1.

□

Jak vypadá poslední bloková matice s neznámými maticemi $\mathbf{Q}_{i,j}$? Ukažme si její tvar, získáme jej pomocí přímého odvození. Jestliže předpokládáme inverzi blokové matice uvedené v předchozí větě, platí:

$$\begin{pmatrix} \mathbf{B}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}' & , & \mathbf{C} \\ \mathbf{C}' & , & \mathbf{0} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{6}$$

Z (6) dostáváme soustavu rovnic

$$\mathbf{B}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'\mathbf{Q}_{11} + \mathbf{C}\mathbf{Q}_{21} = \mathbf{I}, \tag{7}$$

$$\mathbf{B}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'\mathbf{Q}_{12} + \mathbf{C}\mathbf{Q}_{22} = \mathbf{0}, \tag{8}$$

$$\mathbf{C}'\mathbf{Q}_{11} = \mathbf{0}, \tag{9}$$

$$\mathbf{C}'\mathbf{Q}_{12} = \mathbf{I}. \tag{10}$$

Nechť $\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'$ je regulární. Z (8) dostáváme

$$\mathbf{Q}_{12} = -[\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}']^{-1}\mathbf{C}\mathbf{Q}_{22},$$

po dosazení do (10)

$$\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{Q}_{22} = \mathbf{I}$$

získáme první výsledek z hledané blokové matice,

$$\mathbf{Q}_{22} = (-\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}.$$

Dosadíme-li zpět do vzorce pro $\mathbf{Q}_{12}$, vyjde nám další prvek diagonální matice,

$$\mathbf{Q}_{12} = -(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C}(-\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}.$$

Z rovnice (7) obdržíme

$$\mathbf{Q}_{11} = (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1} - (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C}\mathbf{Q}_{21}.$$

Tento vztah dosadíme do rovnice (9), posléze postupnými úpravami vyjádříme $\mathbf{Q}_{21}$, budeme-li postupovat správně, získáme (až na transpozici) stejný výsledek jako pro $\mathbf{Q}_{12}$,

$$\mathbf{C}'((\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1} - (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C}\mathbf{Q}_{21}) = \mathbf{0},$$

$$\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1} - \mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C}\mathbf{Q}_{21} = \mathbf{0},$$

$$\mathbf{Q}_{21} = -(-\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}.$$

Vidíme, že skutečně $\mathbf{Q}_{21} = \mathbf{Q}_{12}'$. Nyní nám už zbývá jen dosadit do výsledku pro $\mathbf{Q}_{11}$ a získáme kompletní výsledek pro inverzní blokovou matici uvedenou ve výše uvedené větě,

$$\begin{aligned}
\mathbf{Q}_{11} &= (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1} - (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C}\times \\
&\quad \times((\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}), \\
\mathbf{Q}_{21} = \mathbf{Q}_{12}' &= (\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}, \\
\mathbf{Q}_{22} &= (-\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}.
\end{aligned}$$

Uvedené rovnice porovnáme s výsledkem, který získáme z Rohdeho formule.

Věta 2.2 (Rohdeho formule)**.** *[9] Nechť* $\mathbf{T}$ *je matice o rozměrech* $m \times m$, $\mathbf{U}$ *matice o rozměrech* $m \times n$, $\mathbf{V}$ *o rozměrech* $n \times m$ *a* $\mathbf{W}$ *o rozměrech* $n \times n$. *Předpokládejme, že* $\mathbf{T}$ *není singulární. Potom matice*

$$\begin{pmatrix} \mathbf{T} & , & \mathbf{U} \\ \mathbf{V} & , & \mathbf{W} \end{pmatrix}, \text{ resp. ekvivalentně } \begin{pmatrix} \mathbf{W} & , & \mathbf{U} \\ \mathbf{V} & , & \mathbf{T} \end{pmatrix},$$

nejsou singulární tehdy a jenom tehdy, jestliže matice o rozměrech $n \times n$, $\mathbf{Q} = \mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}$, *není singulární. Z toho plyne*

$$\begin{pmatrix} \mathbf{T} & , & \mathbf{U} \\ \mathbf{V} & , & \mathbf{W} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{T}^{-1} + \mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} & , & -\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} & , & \mathbf{Q}^{-1} \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{T}^{-1} & , & \mathbf{0} \\ \mathbf{0} & , & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\mathbf{T}^{-1}\mathbf{U} \\ \mathbf{I}_n \end{pmatrix} \mathbf{Q}^{-1}(-\mathbf{V}\mathbf{T}^{-1}, \mathbf{I}_n),$$

$$\begin{pmatrix} \mathbf{W} & , & \mathbf{V} \\ \mathbf{U} & , & \mathbf{T} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{Q}^{-1} & , & -\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} \\ -\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1} & , & \mathbf{T}^{-1} + \mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{0} & , & \mathbf{0} \\ \mathbf{0} & , & \mathbf{T}^{-1} \end{pmatrix} + \begin{pmatrix} \mathbf{I}_n \\ -\mathbf{T}^{-1}\mathbf{U} \end{pmatrix} \mathbf{Q}^{-1}(\mathbf{I}_n, -\mathbf{V}\mathbf{T}^{-1}).$$

Důkaz: Viz [9], str. 99, Theorem 8.5.11.

□

Ověříme, zda získáme stejné výsledky také pomocí Rohdeho věty. Prvky budeme zapisovat postupně jako v předchozím postupu,

$$\begin{aligned} \mathbf{Q}_{11} &= (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1} + (\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C}\times \\ &\quad \times(-\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}, \\ \mathbf{Q}_{21} = \mathbf{Q}'_{12} &= -(-\mathbf{C}'\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}'\mathbf{C})^{-1}\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}, \\ \mathbf{Q}_{22} &= (-\mathbf{C}'(\mathbf{B}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{B}')^{-1}\mathbf{C})^{-1}. \end{aligned}$$

Upravíme-li znaménka u členů $\mathbf{Q}_{11}$ a $\mathbf{Q}_{21}$, resp. $\mathbf{Q}'_{12}$, získáme stejné vztahy jako pomocí přímého odvození.

3 Nelineární struktury modelů

V praxi se s lineárními modely a strukturami nesetkáváme příliš často, zpravidla jsou modely nelineární. Výhodou však je, že můžeme modely linearizovat a pracovat s méně složitými početními operacemi.

Pro jednoduchost nejprve uvažujme model číslo 2 (na něj lze ostatně všechny ostatní výše uvedené regresní modely převést) s nelineární regresní funkcí $\mathbf{f}(\boldsymbol{\beta})$ parametrů $\boldsymbol{\beta}$.

V případě, že $\boldsymbol{\beta}_0$ bude známý (číselný) vektor, můžeme $\mathbf{f}(\boldsymbol{\beta})$ rozvinout do Taylorovy řady v bodě $\boldsymbol{\beta}_0$, přitom zanedbáme členy druhého a vyšších řádů. Rozvoj vypadá dle knihy [10] (str. 228) takto:

$$\mathbf{f}(\boldsymbol{\beta}) = \mathbf{f}(\boldsymbol{\beta}_0) + \mathbf{F}(\boldsymbol{\beta}_0)\delta\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\kappa}_{\delta\beta} \cdots ,$$

kde

$$\mathbf{F}(\boldsymbol{\beta}_0) = \left.\frac{\partial \mathbf{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right|_{\beta=\beta_0}, \boldsymbol{\kappa}_{\delta\beta} = (\delta\boldsymbol{\beta}'\mathbf{F}_1\delta\boldsymbol{\beta}, \ldots, \delta\boldsymbol{\beta}'\mathbf{F}_n\delta\boldsymbol{\beta})',$$

$$\mathbf{F}_i = \left.\frac{\partial^2 f_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\beta=\beta_0}, i = 1, \ldots, n$$

a $\delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$.

Následně může být odhad samotného parametru určen lineárně. Navíc, je-li observační vektor $\mathbf{Y}$ normální, lze použít standardní postupy při určování konfidenčních oblastí nebo třeba testování hypotézy o parametrech.

Máme tedy model

$$\mathbf{Y} \sim_n [\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Sigma}], \boldsymbol{\beta} \in \mathcal{V} \subset \mathbb{R}^k, \text{ kde } \mathbf{f} : \mathcal{V} \to \mathbb{R}^n,$$

kde $\mathbf{f}$ je známou funkcí s absolutně spojitými derivacemi, $\boldsymbol{\Sigma}$ je známá pozitivně definitivní (p.d.) matice, $\mathcal{V} \subset \mathbb{R}^k$ je parametrický prostor a $\mathbb{R}^k$ je $k-$dimenzionální euklidovský prostor.

Můžeme-li v Taylorově řadě funkce $\mathbf{f}(\cdot)$ v přibližné hodnotě parametru $\boldsymbol{\beta}_0$ zanedbat členy druhého a vyšších řádů, nazveme výše uvedený model *modelem se slabou nelinearitou*. Odstranění členů vyšších řádů značně zjednoduší následující postupy.

Ačkoli jsou výše uvedené důvody faktem, proč se linearizace modelu velmi často v praxi využívá, je třeba zkoumat vliv zanedbaných členů.

Přitom již budeme rovnou uvažovat model s podmínkami typu II. Mějme tedy model $\mathbf{Y} \sim_n (\mathbf{f}(\boldsymbol{\beta}_1), \boldsymbol{\Sigma})$ s podmínkami na parametr $\boldsymbol{\beta}_1 \in \mathbb{R}^{k_1}$, rovněž na parametr $\boldsymbol{\beta}_2 \in \mathbb{R}^{k_2}$ ve tvaru $\mathbf{h}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbf{0}$. Model dle [10] vyjádříme v kvadratickém tvaru

$$\mathbf{Y} - \mathbf{f}_0 \sim_n (\mathbf{F}\delta\boldsymbol{\beta}_1 + \frac{1}{2}\boldsymbol{\kappa}_{\delta\beta_1}, \boldsymbol{\Sigma})$$

s podmínkami

$$\mathbf{h_0} + \mathbf{H_1}\delta\boldsymbol{\beta}_1 + \mathbf{H}_2\delta\boldsymbol{\beta}_2 + \frac{1}{2}\boldsymbol{\omega}_{(\delta\beta_1,\delta\beta_2)'} = \mathbf{0},$$

kde

$$\mathbf{f}_0 = \mathbf{f}(\boldsymbol{\beta}_0), \mathbf{h}_0 = \mathbf{h}(\boldsymbol{\beta}_0),\ \mathbf{H}_1 = \left.\frac{\partial \mathbf{h}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\right|_{\beta=\beta_0},\ \mathbf{H}_2 = \left.\frac{\partial \mathbf{h}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2}\right|_{\beta=\beta_0},$$

$$\mathbf{H} = (\mathbf{H}_1, \mathbf{H}_2),\ \boldsymbol{\beta}_0 = \begin{pmatrix} \boldsymbol{\beta}_{1,0} \\ \boldsymbol{\beta}_{2,0} \end{pmatrix}$$

a

$$\left\{\boldsymbol{\omega}_{(\delta\beta_1',\delta\beta_2')'}\right\}_i = (\delta\boldsymbol{\beta}_1', \delta\boldsymbol{\beta}_2') \begin{pmatrix} \mathbf{A}, & \mathbf{B} \\ \mathbf{B}', & \mathbf{D} \end{pmatrix} \begin{pmatrix} \delta\boldsymbol{\beta}_1 \\ \delta\boldsymbol{\beta}_2 \end{pmatrix},$$

kde

$$\mathbf{A} = \left.\partial^2 h_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'\right|_{\beta=\beta_0},$$

$$\mathbf{B} = \left.\partial^2 h_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_2'\right|_{\beta=\beta_0},$$

$$\mathbf{D} = \left.\partial^2 h_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2'\right|_{\beta=\beta_0},$$

$i = 1, \ldots, q$. Budeme navíc dále předpokládat, že vektor $\delta\boldsymbol{\beta}_2$ je spojitá funkce vektoru $\delta\boldsymbol{\beta}_1$ se spojitými druhými derivacemi.

Výše uvedené poznatky shrnuje následující lemma, v němž budou uvedeny BLUE parametru. Než přejdeme k samotnému lemmatu, vysvětleme si označení $(\mathbf{H}_2')^-_{\mathcal{M}(H_1C^{-1}H_1')}$.

Mějme matici $\mathbf{A} = \mathbf{H}_2'$ o rozměrech $q \times k_2$ a matici $\boldsymbol{\Sigma} = \mathbf{H}_1\mathbf{C}^{-1}\mathbf{H}_1'$, která je čtvercová o rozměrech $k_2 \times k_2$ a je navíc pozitivně semidefinitní. Dle Frobeniovy formule platí, že

$$\mathbf{Ax} = \mathbf{y} \in \mathcal{M}(\mathbf{A}) = \{\mathbf{Au} : \mathbf{u} \in \mathbb{R}^n\},$$

přičemž zároveň

$$\min\{||\mathbf{x}||_\Sigma : \mathbf{Ax} = \mathbf{y}\} = ||\mathbf{A}^-_{\mathcal{M}(\Sigma)}\mathbf{y}||.$$

Pro vysvětlení dodejme, že $\mathbf{A}^-_{\mathcal{M}(\Sigma)}$ je označení pro matici, pro níž platí

$$\mathbf{A}\mathbf{A}^-_{\mathcal{M}(\Sigma)}\mathbf{A} = \mathbf{A}\ \wedge\ \boldsymbol{\Sigma}\mathbf{A}^-_{\mathcal{M}(\Sigma)}\mathbf{A} = (\boldsymbol{\Sigma}\mathbf{A}^-_{\mathcal{M}(\Sigma)}\mathbf{A})'.$$

V tomto případě existuje celá třída takových matic, z nichž jedním reprezentantem je následující matice

$$\boldsymbol{\Sigma}^-\mathbf{A}'(\mathbf{A}\boldsymbol{\Sigma}^-\mathbf{A}')^-,$$

ovšem jen tehdy, pokud $\mathcal{M}(\mathbf{A}') \subset \mathcal{M}(\mathbf{\Sigma})$. Obecně lze uvedeného reprezentanta vyjádřit následovně:

$$(\mathbf{\Sigma} + \mathbf{A}'\mathbf{A})^{-}\mathbf{A}'[\mathbf{A}(\mathbf{\Sigma} + \mathbf{A}'\mathbf{A})^{-}\mathbf{A}']^{-}.$$

Osvětleme si další označení, $\mathbf{M}_{H_2} = \mathbf{I} - \mathbf{P}_{H_2}$, kde $\mathbf{P}_{H_2}$ je projekční matice na podprostor $\mathcal{M}(\mathbf{H}_2)$. Navíc znaménko mínus značí pseudoinverzi matice. Matice označena znaménkem plus se nazývá *Moore-Penroseova pseudoinverze.*

Lemma 3.1. *[10] V rámci modelu*

$$\mathbf{Y} - \mathbf{f}_0 \sim_n (\mathbf{F}\delta\boldsymbol{\beta}_1, \mathbf{\Sigma}), \begin{pmatrix} \delta\boldsymbol{\beta}_1 \\ \delta\boldsymbol{\beta}_2 \end{pmatrix} \in \left\{ \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} : \mathbf{h}_0 + \mathbf{H}_1\mathbf{u} + \mathbf{H}_2\mathbf{v} = \mathbf{0} \right\},$$

$h(\mathbf{F}) = k_1 < n$, $h(\mathbf{H}_1, \mathbf{H}_2) = q < k_1 + k_2$, $h(\mathbf{H}_2) = k_2 < q$ *je BLUE parametru* $\begin{pmatrix} \delta\boldsymbol{\beta}_1 \\ \delta\boldsymbol{\beta}_2 \end{pmatrix}$ *roven*

$$\delta\widehat{\widehat{\boldsymbol{\beta}}}_1 = \delta\widehat{\boldsymbol{\beta}}_1 - \mathbf{C}^{-1}\mathbf{H}_1'[\mathbf{M}_{H_2}(\mathbf{H}_1\mathbf{C}^{-1}\mathbf{H}_1')\mathbf{M}_{H_2}]^{+}(\mathbf{h}_0 + \mathbf{H}_1\delta\widehat{\boldsymbol{\beta}}_1),$$

$$\delta\widehat{\widehat{\boldsymbol{\beta}}}_2 = -[(\mathbf{H}_2')^{-}_{\mathcal{M}(H_1C^{-1}H_1')}]',$$

kde $\delta\widehat{\boldsymbol{\beta}}_1 = \mathbf{C}^{-1}\mathbf{F}'\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{f}_0)$ *je BLUE v rámci modelu* $\mathbf{Y} - \mathbf{f}_0 \sim_n (\mathbf{F}\delta\boldsymbol{\beta}_1, \mathbf{\Sigma})$ *bez podmínek; varianční matice je*

$$\operatorname{var}\begin{pmatrix} \delta\widehat{\widehat{\boldsymbol{\beta}}}_1 \\ \delta\widehat{\widehat{\boldsymbol{\beta}}}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{A}}, & \tilde{\mathbf{B}} \\ \tilde{\mathbf{B}}', & \tilde{\mathbf{D}} \end{pmatrix},$$

kde

$$\tilde{\mathbf{A}} = (\mathbf{M}_{H_1'M_{H_2}}\mathbf{C}\mathbf{M}_{H_1'M_{H_2}})^{+},$$
$$\tilde{\mathbf{B}} = -\mathbf{C}^{-1}\mathbf{H}_1'(\mathbf{H}_2')^{-}_{\mathcal{M}(H_1C^{-1}H_1')},$$
$$\tilde{\mathbf{D}} = [\mathbf{H}_2'(\mathbf{H}_1\mathbf{C}^{-1}\mathbf{H}_1' + \mathbf{H}_2\mathbf{H}_2')^{-1}\mathbf{H}_2]^{-1} - \mathbf{I}.$$

Důkaz: Viz [10], str. 246, Lemma VI.2.2.3.1

□

Jelikož už byl pojem Taylorova rozvoje osvětlen, můžeme se vrátit k příkladu z kapitoly 2. Uvedené poznatky shrneme do komplexního stochastického modelu, který je následující:

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \beta_0 \\ a \cdot \frac{\sin\beta_0}{\sin\alpha_0} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -a \cdot \frac{\sin\beta_0\cos\alpha_0}{\sin^2\alpha_0} & a \cdot \frac{\cos\beta_0}{\sin\alpha_0} \end{pmatrix} \cdot \begin{pmatrix} \Delta\alpha \\ \Delta\beta \end{pmatrix}.$$

Současně platí podmínka na součet úhlů v trojúhelníku,

$$\alpha_0 + \beta_0 + (1,1) \begin{pmatrix} \Delta\alpha \\ \Delta\beta \end{pmatrix} + \gamma - \pi = 0.$$

4 Ortogonální regrese pro kompoziční data

Dosud nabyté poznatky o kompozičních datech a regresních modelech využijeme v této kapitole zabývající se ortogonální regresí. Ta je jinak též známá pod názvem *metoda úplných čtverců* - z anglického *Total Least Squares* = TLS).

Metoda ortogonální regrese začíná vzbuzovat zájem také v oblastech mimo statistiku. Důvodem je dostupnost efektivních a numericky robustních algoritmů, v nichž hraje velkou roli *singulární rozklad* (SVD - Singular Value Decomposition). Je vhodný zejména v těch oblastech, v nichž jsou data zatížena chybami (či šumem), což je častým případem zejména v oblasti inženýrských aplikací.

Ortogonální regresi lze považovat za vhodné rozšíření klasické metody nejmenších čtverců. V nejjednodušším případě je úkolem ortogonální regrese odhadnout a určit regresní přímku pro soubor n dvourozměrných pozorování, což jsou v našem případě výsledky ilr-transformace třísložkových kompozicí.

Transformovaná kompoziční data lze dle článku [5] přepsat jako $n \times 2$ datovou matici $(\mathbf{z}_1, \mathbf{z}_2)$. Označme matici $\mathbf{X} = (\mathbf{1}_n, \mathbf{z}_1)$, kde $\mathbf{1}_n$ je n-rozměrný vektor jedniček. Dále uvažujme $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ pro neznámé parametry regresní přímky, které chceme odhadnout. Hledaný odhad $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ metodou ortogonální regrese se obvykle vyjadřuje v následujícím tvaru:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} - \lambda_3^2\mathbf{I})^{-1}\mathbf{X}'\bar{\mathbf{z}}_2,$$

kde λ_3 je nejmenší singulární hodnota z rozkladu blokové matice $(\mathbf{X}, \mathbf{z}_2)$, o singulárním rozkladu matice pohovoříme podrobně ještě dále.

Odhady parametrů regresní přímky pomocí ortogonální regrese lze ovšem vyjádřit i jiným způsobem. Uvažujme body roviny $(z_{11}, z_{21}), \ldots, (z_{1n}, z_{2n})$. Proložíme-li jimi přímku pomocí metody nejmenších čtverců, minimalizuje tato přímka součet čtverců, jejichž strany k ní z bodů (z_{1i}, z_{2i}), $i = 1, \ldots, n$ vedeny v kolmém směru k ose z_1. Metoda ortogonální regrese přitom minimalizuje součet takových čtverců, že jejich strany jsou kolmé k této přímce (samozřejmě jsou vedeny z bodů (z_{1i}, z_{2i}) jako v předchozím případě).

Zaveďme následující označení dle [2]:

$$\bar{z}_1 = \tfrac{1}{n}\textstyle\sum z_{1i}, \quad s_{z_1}^2 = \tfrac{1}{n}\textstyle\sum (z_{1i}^2 - \bar{z}_1^2), \quad s_{z_1 z_2} = \tfrac{1}{n}\textstyle\sum (z_{1i}z_{2i} - \bar{z}_1\bar{z}_2),$$

$$\bar{z}_2 = \tfrac{1}{n}\textstyle\sum z_{2i}, \quad s_{z_2}^2 = \tfrac{1}{n}\textstyle\sum (z_{2i}^2 - \bar{z}_2^2).$$

Věta 4.1. *[2] Předpokládejme, že $s_{z_1 z_2} \neq 0, s_{z_1}^2 > 0, s_{z_2}^2 > 0$. Pak parametry přímky $z_2 = \beta_1 + \beta_2 z_1$ určené metodou ortogonální regrese jsou:*

$$\hat{\beta}_1 = \bar{z}_2 - \hat{\beta}_2\bar{z}_1, \ \hat{\beta}_2 = \frac{s_{z_2}^2 - s_{z_1}^2 + \sqrt{(s_{z_2}^2 - s_{z_1}^2)^2 + 4s_{z_1 z_2}^2}}{2s_{z_1 z_2}}.$$

Důkaz: Viz [2], str. 209, Věta 12.10.

□

Z výrazu pro $\hat{\beta}_1$ je patrné, že přímka prochází těžištěm $(\bar{z}_1, \bar{z}_2)$, tedy stejně jako v případě přímky proložené za pomocí metody nejmenších čtverců. Při ortogonální regresi ovšem bereme v potaz navíc fakt, že rovněž nezávisle proměnná je zatížena chybou. Připomeňme, že v případě metody nejmenších čtverců je přitom chybou zatížena jen závisle proměnná.

Úvaha vychází z předpokladu, že dvojice (z_{1i}, z_{2i}) jsou hodnoty, které se při uskutečněném experimentu realizují doopravdy, nemáme je ale možnost určit. Mezi závisle a nezávisle proměnnou platí přesně tento lineární vztah:

$$z_{2i} = \beta_1 + \beta_2 z_{1i},\ i = 1, \ldots, n. \tag{11}$$

Místo dvojice (z_{1i}, z_{2i}) vybereme jen dvojici (ξ_i, η_i). Její složky jsou zatíženy náhodnými chybami, budeme proto předpokládat, že:

$$\xi = z_{1i} + \delta_i,\ \eta_i = z_{2i} + \epsilon_i,\ i = 1, \ldots, n, \tag{12}$$

kde $\delta_1, \ldots, \delta_n, \epsilon_1, \ldots, \epsilon_n$ jsou takové nezávislé náhodné veličiny, že

$$\delta_i \sim \mathbb{N}(0, \sigma_\delta^2),\ \epsilon_i \sim \mathbb{N}(0, \sigma_\epsilon^2),\ i = 1, \ldots, n.$$

Poznamenejme přitom, že předpoklad normality jsme uvedli kvůli následující Větě 4.2, není ovšem obecně nutný. Dosadíme-li z (12) do (11), dostaneme

$$\eta_i = \beta_1 + \beta_2 \xi_i + (\epsilon - \beta_2 \delta_i),\ i = 1, \ldots, n.$$

Zatímco v klasickém regresním modelu hodnoty nezávisle proměnné nezávisí na vektoru chyb, zde dostáváme následující vztah:

$$\operatorname{cov}(\xi_i, \epsilon_i - \beta_2 \delta_i) = \operatorname{cov}(z_i + \delta_i, \epsilon_i - \beta_2 \delta_i) = -\beta_2 \sigma_\delta^2.$$

Jelikož je hodnota obecně nenulová, komplikuje nám to odvozování příslušných statistických metod. Jak uvádí [2], pokud jsou čísla z_i daná (v experimentu se totiž nějakým způsobem nastavují), říká se modelu (11) *funkční vztah*. Jsou-li z_i nezávislé stejně rozdělené náhodné veličiny (experimentátor tudíž nemůže jejich hodnot nijak ovlivnit), pak model (11) nazveme *strukturální relace*.

Dle zjednodušeného předpokladu $\sigma_\delta^2 = \sigma_\epsilon^2 = \sigma^2$ (tedy z_i a y_i jsou měřeny se stejnou předností) nyní budeme řešit lineární funkční vztah. Pokud jsou uvažovány tyto předpoklady, platí, že:

$$\xi_i \sim \mathbb{N}(z_{1i}, \sigma^2),\ \eta_i \sim \mathbb{N}(\beta_1 + \beta_2 z_{1i}, \sigma^2).$$

Jelikož jsou veličiny $\xi_1, \ldots, \xi_n$ a $\eta_1, \ldots, \eta_n$ nezávislé, sdružená hustota je vyjádřena takto:

$$(2\pi)^{-n} \sigma^{-2n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [(\eta_i - \beta_1 - \beta_2 z_{1i})^2 + (\xi_i - z_{1i})^2] \right\}.$$

Počet neznámých parametrů je $n+3$, k dispozici máme ovšem jen n dvojic (ξ_i, η_i). Pro odhad parametrů se v následující větě volí metoda maximální věrohodnosti.

Věta 4.2. *[2] Označme*

$$\bar{\xi} = \frac{1}{n}\sum \xi_i, \quad s_\xi^2 = \frac{1}{n}\sum \left(\xi_i^2 - \bar{\xi}^2\right), \quad s_{\xi\eta} = \frac{1}{n}\sum \left(\xi_i\eta_i - \bar{\xi}\bar{\eta}\right),$$

$$\bar{\eta} = \frac{1}{n}\sum \eta_i, \quad s_\eta^2 = \frac{1}{n}\sum \left(\eta_i^2 - \bar{\eta}^2\right).$$

Pak maximálně věrohodné odhady parametrů β_1, β_2 a z_{1i} jsou

$$\hat{\beta}_2 = \frac{s_\eta^2 - s_\xi^2 + \sqrt{(s_\eta^2 - s_\xi^2)^2 + 4s_{\xi\eta}^2}}{2s_{\xi\eta}},$$

$$\hat{\beta}_1 = \bar{\eta} - \hat{\beta}_2\bar{\xi},$$

$$\hat{z}_{1i} = \frac{\xi_i + \hat{\beta}_2\eta_i - \hat{\beta}_1\hat{\beta}_2}{1+\hat{\beta}_2^2}, \ i = 1, \ldots, n.$$

Položme

$$S^* = \sum_{i=1}^{n}[(\eta_i - \hat{\beta}_1 - \hat{\beta}_2\hat{z}_{1i})^2 + (\xi_i - \hat{z}_{1i}^2)].$$

Pak

$$\hat{\sigma}^2 = \frac{S^*}{2n}$$

je maximálně věrohodným odhadem parametru σ^2.

Důkaz: Viz [2], str. 212, Věta 12.11.

□

Všimněme si, že odhady regresních parametrů jsou totožné se situací ve Větě 4.1. Navíc je zřejmě, že odhad $\boldsymbol{\beta}$ metodou ortogonální regrese minimalizuje euklidovskou normu

$$\frac{||\mathbf{X}\boldsymbol{\beta} - \mathbf{z}_2||^2}{||\boldsymbol{\beta}||^2 + 1}.$$

Ilr souřadnice pro kompoziční data jsou konstruovány vzhledem k Aitchisonově geometrii na simplexu, ony samotné se ovšem již řídí euklidovskou geometrií. Zároveň je třeba zdůraznit, že různé volby ilr souřadnic představují pouze rotaci dat, tudíž se ortogonální regrese jeví jako vhodný nástroj pro jejich regresní analýzu.

4.1 Odhad regresní přímky s využitím teorie lineárních modelů

Jak je uvedeno v článku [5], další možností, jak najít vhodnou přímku pro n rovinných dat ve smyslu ortogonální regrese je technika lineárních statistických modelů, konkrétně jeden speciální model s podmínkou II. typu. Největší výhodou v tomto přístupu je možnost konstrukce konfidenčních oblastí a testování hypotéz s použitím teorie lineárních modelů a jejich výhodné teoretické vlastnosti oproti přístupům zmíněných v předchozí kapitole.

Dále budeme uvažovat $(n \times 2)$-rozměrnou datovou matici $(\mathbf{Z}_1, \mathbf{Z}_2)$. Jestliže skutečné hodnoty první souřadnice představuje vektor $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ a u druhé souřadnice $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$, přičemž výsledná měření jsou vektory $\mathbf{z}_1$ a $\mathbf{z}_2$, potom lineární vztah mezi těmito dvěma reprezentacemi lze zapsat takto:

$$\boldsymbol{\nu} = \beta_1 \mathbf{1}_n + \beta_2 \boldsymbol{\mu}, \tag{13}$$

kde β_1 a β_2 jsou koeficienty přímky. Poznamenejme, že v kontextu předchozí kapitoly jsou dvojice (z_{1i}, z_{2i}) a (μ_i, ν_i) rovny právě (ξ_u, η_i) a (z_{1i}, z_{2i}).

Předpokládejme, že $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$ jsou nezávislé a realizovány se směrodatnou odchylkou $\sigma > 0$. Výsledný tvar odhadů $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\nu}}$ parametrů β_1, β_2, $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$, stejně jako jejich varianční a kovarianční matice jsou uvedeny v následující větě.

Věta 4.3. *[5] Uvažujme dvourozměrný náhodný vektor $(\mathbf{Z}_1', \mathbf{Z}_2')'$ se středními hodnotami $(\boldsymbol{\mu}', \boldsymbol{\nu}')'$ a varianční matici $\sigma^2 \mathbf{1}_{2n}$, kde vektory $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$ vyhovují vztahu (13). Označme $\beta_1^{(0)}, \beta_2^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\nu}^{(0)}$ přibližné hodnoty $\beta_1, \beta_2, \boldsymbol{\mu}, \boldsymbol{\nu}$ tak, aby splňovaly (13). BLUE vektorů $\boldsymbol{\mu}, \boldsymbol{\nu}$ a parametrů β_1, β_2 v linearizovaném modelu jsou dány:*

$$\hat{\boldsymbol{\mu}} = \mathbf{Z}_1 + \frac{\beta_2^{(0)}}{[\beta_2^{(0)}]^2 + 1} \mathbf{M}^{(0)} [\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2 (\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)})],$$

$$\hat{\boldsymbol{\nu}} = \mathbf{Z}_2 - \frac{\beta_2^{(0)}}{[\beta_2^{(0)}]^2 + 1} \mathbf{M}^{(0)} [\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2 (\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)})],$$

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \beta_1^{(0)} \\ \beta_2^{(0)} \end{pmatrix} + \begin{pmatrix} n & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ [\boldsymbol{\mu}^{(0)}]'\mathbf{1} & [\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \times$$

$$\times \begin{pmatrix} \mathbf{1}' \left[\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2^{(0)} (\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)}) \right] \\ [\boldsymbol{\mu}^{(0)}]' \left[\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2^{(0)} (\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)}) \right] \end{pmatrix},$$

kde

$$\mathbf{M}^{(0)} = \mathbf{I}_n - (\mathbf{1}, \boldsymbol{\mu}^{(0)}) \begin{pmatrix} n & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ [\boldsymbol{\mu}^{(0)}]'\mathbf{1} & [\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}' \\ [\boldsymbol{\mu}^{(0)}] \end{pmatrix}.$$

Varianční matice odhadů $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\nu}}$ jsou dány vztahy:

$$\mathrm{var}[\hat{\boldsymbol{\mu}}] = \sigma^2\mathbf{I}_n - \frac{\left[\beta_2^{(0)}\right]^2\sigma^2}{\left[\beta_2^{(0)}\right]^2+1}\mathbf{M}^{(0)},$$

$$\mathrm{var}[\hat{\boldsymbol{\nu}}] = \sigma^2\mathbf{I}_n - \frac{\sigma^2}{\left[\beta_2^{(0)}\right]^2+1}\mathbf{M}^{(0)}.$$

Kovarianční matice odhadů $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\nu}}$ vypadá následovně:

$$\mathrm{cov}\,[\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}] = \frac{\beta_2^{(0)}\sigma^2}{\left[\beta_2^{(0)}\right]^2+1}\mathbf{M}^{(0)}.$$

Varianční matice odhadu $(\hat{\beta}_1,\hat{\beta}_2)'$ je

$$\mathrm{var}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}\right] = \sigma^2\left(\left[\beta_2^{(0)}\right]^2+1\right)\begin{pmatrix} n & \mathbf{1}'\boldsymbol{\mu}^{(0)}\\ [\boldsymbol{\mu}^{(0)}]'\mathbf{1} & [\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)}\end{pmatrix}^{-1}$$

a kovarianční matice odhadů $(\hat{\boldsymbol{\mu}}',\hat{\boldsymbol{\nu}}')'$ a $(\hat{\beta}_1,\hat{\beta}_2)'$ je vyjádřena ve tvaru

$$\mathrm{cov}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix},\begin{pmatrix}\hat{\boldsymbol{\mu}}\\ \hat{\boldsymbol{\nu}}\end{pmatrix}\right] = -\sigma^2\left(\left[\beta_2^{(0)}\right]^2+1\right)\begin{pmatrix} n & \mathbf{1}'\boldsymbol{\mu}^{(0)}\\ [\boldsymbol{\mu}^{(0)}]'\mathbf{1} & [\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)}\end{pmatrix}^{-1}\times$$

$$\times\begin{pmatrix}\beta_2^{(0)}\mathbf{1}' & -\mathbf{1}'\\ \beta_2^{(0)}[\boldsymbol{\mu}^{(0)}]' & -[\boldsymbol{\mu}^{(0)}]'\end{pmatrix}.$$

Důkaz: Viz [5], str. 1151, Theorem 1.

□

Uveďme pro úplnost ještě, jak vypadá příslušný lineární model. V maticovém zápisu máme lineární model dle článku [5]

$$\begin{pmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{pmatrix} + \boldsymbol{\varepsilon}, \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_{2n},$$

kde neznámé vektory $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$ vyhovují regresní přímce dané vztahem $\boldsymbol{\nu} = \beta_1\mathbf{1}_n + \beta_2\boldsymbol{\mu}$ a β_1,β_2 jsou neznámé parametry. Regresní přímku můžeme linearizovat pomocí Taylorova rozvoje v bodech $\boldsymbol{\mu}^{(0)},\boldsymbol{\nu}^{(0)}\beta_1^{(0)}$ a $\beta_2^{(0)}$, přičemž členy druhého a vyšších řádů zanedbáme.

Takto získáme lineární model

$$\begin{pmatrix} \mathbf{Z}_1 - \boldsymbol{\mu}^{(0)} \\ \mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\delta\mu} \\ \boldsymbol{\delta\nu} \end{pmatrix} + \boldsymbol{\varepsilon}, \boldsymbol{\delta\mu} = \boldsymbol{\mu} - \boldsymbol{\mu}^{(0)}, \boldsymbol{\delta\nu} - \boldsymbol{\nu}^{(0)},$$

přičemž $\boldsymbol{\delta\mu}$ a $\boldsymbol{\delta\nu}$ vyhovují vztahu

$$\delta\beta_1 \mathbf{1}_n + \delta\beta_2 \boldsymbol{\mu}^{(0)} + \beta_2^{(0)} \boldsymbol{\delta\mu} = \boldsymbol{\delta\nu}.$$

Také zde platí, že $\delta\beta_1 = \beta_1 - \beta_1^{(0)}$ a $\delta\beta_2 = \beta_2 - \beta_2^{(0)}$. Předchozí vztah můžeme přepsat jako

$$(\beta_2^{(0)} \mathbf{I}_n, -\mathbf{I}_n) \begin{pmatrix} \boldsymbol{\delta\mu} \\ \boldsymbol{\delta\nu} \end{pmatrix} + (\mathbf{1}_n, \boldsymbol{\mu}_0) \begin{pmatrix} \delta\beta_1 \\ \delta\beta_2 \end{pmatrix} = \mathbf{0}.$$

Dosud jsme uvažovali, že směrodatná odchylka σ je známá. Neznáme-li ji, lze ji odhadnout pomocí následujícího vztahu:

$$\hat{\sigma}^2 = \frac{(\mathbf{Z}_1 - \hat{\boldsymbol{\mu}})'(\mathbf{Z}_1 - \hat{\boldsymbol{\mu}}) + (\mathbf{Z}_2 - \hat{\boldsymbol{\nu}})'(\mathbf{Z}_2 - \hat{\boldsymbol{\nu}})}{n - 2}. \tag{14}$$

Odtud obdržíme

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}) = \frac{\hat{\sigma}^2}{\sigma^2} \text{var} \left[\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \right].$$

Jelikož odhady vyjádřené v prvních třech rovnicích závisí na přibližných hodnotách neznámých parametrů β_1, β_2, $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$ a aproximativní hodnoty jsou neznámé, musíme rovnice vyřešit pomocí iterativní metody o čtyřech krocích.

Algoritmus 4.1. *[5]*

Krok 1. *Algoritmus zahájíme stanovením počátečního odhadu koeficientů regresní přímky ($\beta_1^{(0)}$,$\beta_2^{(0)}$) a skutečných hodnot jednotlivých souřadnic ($\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$). Volbu může ovlivnit situace, kdy je známá nějaká předchozí informace. V našem případě se jedná o odhady získané pomocí Věta 4.1. Posléze položíme $\boldsymbol{\mu}^{(0)} = \mathbf{z}_1$. Hodnotu $\boldsymbol{\nu}^{(0)}$ získáme z rovnice $\boldsymbol{\nu}^{(0)} = \beta_1^{(0)} \mathbf{1} + \beta_2^{(0)} \boldsymbol{\mu}^{(0)}$.*

Krok 2. *Pro všechny kompozice (z_{1i}, z_{2k}), $k = 1, \ldots, n$ spočítáme odhady sledovaných parametrů ($\hat{\beta}_1, \hat{\beta}_2, \hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\nu}}$) pomocí prvních tří rovnic ve Větě 4.3.*

Krok 3. *Algoritmus teď vyžaduje aktualizaci startovacích odhadů dle následujícího schématu:*

$$\boldsymbol{\nu}^{(0)} = \hat{\boldsymbol{\nu}} + \left(\hat{\beta} - \beta_2^{(0)}\right) \left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(0)}\right), \ \boldsymbol{\mu}^{(0)} = \hat{\boldsymbol{\mu}}, \ \beta_1^{(0)} = \hat{\beta}_1, \ \beta_2^{(0)} = \hat{\beta}_2.$$

Krok 4. *Kroky 2-4 provádíme do té doby, než odhady konvergují.*

KONEC

Závěrem dodejme, že zmíněný iterativní algoritmus je standardizován v tom smyslu, že volba počátečních odhadů nemá vliv na výsledky konvergence. Obvykle algoritmus probíhá velmi rychle a po několika iteracích obdržíme stabilní výsledky.

4.2 Intervalové odhady a testování hypotéz pro regresní přímku

Statistická inference má v případě ortogonální regrese standardně asymptotický charakter, kterému je možné se vyhnout užitím přístupů pomocí teorie lineárních modelů. Inference nás většinou zajímá zejména v případě parametru β_2 Definujme tedy $\beta_2 = \mathrm{tg}\theta$ a $\hat{\beta}_2 = \mathrm{tg}\hat{\theta}$.

Jak je uvedeno v [6], už v šedesátých letech minulého století byl navrhnut $(100-\alpha)\%$ výběrový konfidenční interval pro parametr β_2 s hranicemi $\mathrm{tg}(\hat{\theta}-\Phi_1)$ a $\mathrm{tg}(\hat{\theta}+\Phi_1)$, přičemž Φ_1 je definováno jako

$$\Phi_1 = \frac{1}{2}\arcsin\sqrt{\frac{4t^2_{n-2}(1-\alpha/2)(s^2_{z_1}s^2_{z_2}-s^2_{z_1z_2})}{(n-2)[(s^2_{z_1}-s^2_{z_2})^2+4s^2_{z_1z_2}]}}. \tag{15}$$

Jen pro úplnost dodejme, že $t_{n-2}(1-\alpha/2)$ značí $(1-\alpha/2)$-kvantil Studentova rozdělení o $(n-2)$ stupních volnosti. Užitím metody hlavních komponent lze odvodit hranice $\mathrm{tg}(\hat{\theta}-\Phi_2)$ a $\mathrm{tg}(\hat{\theta}+\Phi_2)$, kde Φ_2 je vyjádřeno následovně

$$\Phi_2 = \arcsin\sqrt{\frac{\chi^2_1(1-\alpha)}{(n-1)[\lambda_1/\lambda_2+\lambda_2/\lambda_1-2]}}, \tag{16}$$

přitom $\lambda_1 > \lambda_2$ jsou vlastní čísla výběrové varianční matice znaků Z_1 a Z_2 a $\chi^2_1(1-\alpha)$ označuje $(1-\alpha)$-kvantil pro χ^2 rozdělení s jedním stupněm volnosti.

Abychom mohli otestovat nulovou hypotézu $\beta_2 = \theta = 0$, lze použít statistiku odvozenou ze vzorce představeného Kendallem a Stuartem [6],

$$T = \sqrt{(n-2)\sin^2(2\hat{\theta})\frac{1/4(s^2_{z_1}s^2_{z_2})^2+s^2_{z_1z_2}}{s^2_{z_1}s^2_{z_2}}-s^2_{z_1z_2}}.$$

Tato statistika má za platnosti nulové hypotézy Studentovo rozdělení o $n-2$ stupních volnosti. V případě výběrů s malým rozsahem lze dle [6] použít statistiku danou vztahem

$$T = \frac{(n-2)r^2}{1-r^2}, \text{ kde } r^2 = \frac{(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_2)^2},$$

která má za platnosti nulové hypotézy Fisherovo rozdělení o $(1, n-2)$ stupních volnosti.

V následujícím textu budeme uvažovat náhodný vektor $(\mathbf{Z}'_1, \mathbf{Z}'_2)$ s normálním rozdělením. Ekvivalentně lze říci, že náhodná třísložková kompozice má buď normální, nebo logisticky-normální rozdělení na simplexu, uvádí to [1] a [14]. Za těchto předpokladů je BLUE (nejlepší nestranný lineární odhad) parametru $\boldsymbol{\beta}$ rovněž normálně rozdělený. Dále předpokládáme, že σ^2 je neznámé. Lze jej odhadnout (nezávisle na $\widehat{\boldsymbol{\beta}}$) ze vztahu (14), za předpokladu normality je rozdělení odhadu $\hat{\sigma}^2$ dáno vztahem

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

Pás spolehlivosti kolem regresní přímky $\nu_0 = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\mu_0$ je systémem konfidenčních intervalů pro $\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\mu_0$ v pevně zvoleném bodě μ_0, nebo ekvivalentně pro ν_0 v bodě μ_0. Pro konstrukci intervalu spolehlivosti pro $\beta_1 + \beta_2\mu_0$ použijeme statistiku

$$T = \frac{\hat{\beta}_1 + \hat{\beta}_2\mu_0 - \beta_1 - \beta_2\mu_0}{\sqrt{(1,\mu_0)\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})(1,\mu_0)'}},$$

přičemž její rozdělení je Studentovo t_{n-2}. Následně lze vyjádřit samotný $100(1-\alpha)\%$-interval pro $\beta_1+\beta_2\mu_0$ v pevně zvoleném bodě μ_0, a to přes triviální odvození ve tvaru

$$\hat{\beta}_1 + \hat{\beta}_2\mu_0 \pm \sqrt{(1,\mu_0)\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})(1,\mu_0)'}t_{n-2}(1-\alpha/2).$$

Je třeba upozornit na to, že daný interval spolehlivosti slouží vždy pouze pro jedno konkrétně zvolené μ_0. Chceme-li vyjádřit intervaly spolehlivosti pro všechny možné hodnoty μ_0 současně (a dostat tak pás spolehlivosti pro regresní přímku), je nutné zkonstruovat sdružené intervaly pro $\beta_1+\beta_2\mu_0$. Ty jsou založeny na statistice

$$F = \frac{1}{2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'[\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})]^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

která má rozdělení $F_{2,n-2}$, platí tedy zřejmý vztah

$$P\left\{\frac{1}{2}\widehat{\boldsymbol{\beta}}'[\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})]^{-1}\widehat{\boldsymbol{\beta}} \le F_{2,n-2}(n-2)\right\} = 1-\alpha.$$

Ekvivalentně dle Scheffého věty obdržíme

$$P\left\{\forall \mathbf{a} \in \mathbb{R}^2 : |\mathbf{a}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \le \sqrt{2F_{2,n-2}(1-\alpha)}\sqrt{\mathbf{a}'\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})\mathbf{a}}\right\} = 1-\alpha.$$

Nahradíme-li $\mathbf{a} = (1,\mu_0)'$ pro všechna μ_0, získáme

$$P\left\{\forall \mu_0 \in \mathbb{R}^1 : |\hat{\beta}_1 + \hat{\beta}_2\mu_0 - \beta_1 - \beta_2\mu_0| \le \sqrt{2F_{2,n-2}(1-\alpha)}\sqrt{(1,\mu_0)\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})(1,\mu_0)'}\right\}$$

$$\ge 1-\alpha,$$

tedy sdružený $100(1-\alpha)\%$ interval spolehlivosti pro $\beta_1 + \beta_2\mu_0$ (pro všechny hodnoty současně) je roven

$$\hat{\beta}_1 + \hat{\beta}_2\mu_0 \pm \sqrt{(1,\mu_0)\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})(1,\mu_0)'}\sqrt{2F_{2,n-2}(1-\alpha)}.$$

Šířka konfidenčního intervalu charakterizuje přesnost odhadu regresní přímky. Šířka konfidenčního intervalu pro $\nu_0 = \beta_1+\beta_2\mu_0$ je funkcí μ_0. Minimum se nachází v bodě $\mu_0 = \bar{\mu}$ a jeho šířka se zvětšuje dle toho, jak roste absolutní hodnota $|\mu_0 - \bar{\mu}|$.

Testování parametrů β_1 a β_2 regresní přímky v přístupu z kapitoly 4.1 lze odvodit například pomocí přístupu k obecné lineární hypotéze. Testovací statistika pro $H_0 : \beta_1 = 0$ proti $H_a : \beta_1 \neq 0$ je dána následujícím vztahem,

$$T_1 = \frac{\hat{\beta}_1 \sqrt{n[\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)} - [1'\boldsymbol{\mu}^{(0)}]^2}}{\hat{\sigma}\sqrt{(\beta_2^{(0)} + 1)[\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)}}}, \tag{17}$$

za platnosti nulové hypotézy má statistika T_1 Studentovo rozdělení o $n-2$ stupních volnosti.

Následně vyjádříme $100(1-\alpha)\%$-ní konfidenční interval pro β_1 jako

$$\hat{\beta}_1 \pm \hat{\sigma} t_{n-2}(1-\alpha/2) \frac{\sqrt{(\beta_2^{(0)} + 1)[\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)}}}{\sqrt{n[\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)} - [1'\boldsymbol{\mu}^{(0)}]}}.$$

Obdobně vyjádříme pro $H_0 : \beta_2 = 0$

$$T_2 = \frac{\hat{\beta}_2 \sqrt{n[\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)} - [1'\boldsymbol{\mu}^{(0)}]^2}}{\hat{\sigma}\sqrt{n(\beta_2^{(0)} + 1)}}, \tag{18}$$

a T_2 má tudíž za platnosti nulové hypotézy Studentovo rozdělení (t_{n-2}). Příslušný $100(1-\alpha)\%$-interval spolehlivosti pro β_2 je dán

$$\hat{\beta}_2 \pm \hat{\sigma} t_{n-2}(1-\alpha/2) \frac{\sqrt{n(\beta_2^{(0)} + 1)}}{\sqrt{n[\boldsymbol{\mu}^{(0)}]'\boldsymbol{\mu}^{(0)} - [1'\boldsymbol{\mu}^{(0)}]}}. \tag{19}$$

Na závěr lze otestovat nulovou hypotézu $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_a : \boldsymbol{\beta} \neq \mathbf{0}$, použijeme statistiku:

$$F = \frac{1}{2}\widehat{\boldsymbol{\beta}}'[\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})]^{-1}\widehat{\boldsymbol{\beta}}, \tag{20}$$

která má Fisherovo rozdělení $(F_{2,n-2})$ za platnosti nulové hypotézy. Nulovou hypotézu zamítáme, jestliže $f \geq F_{2,n-2}(1-\alpha)$, alternativně pokud p-hodnota $\leq \alpha$.

Testovací statistiky T_1, T_2, F a konfidenční intervaly pro β_1, β_2 závisí na skutečných hodnotách $\beta_2^{(0)}$ a $\boldsymbol{\mu}^{(0)}$. V praxi se skutečné hodnoty nahrazují přibližnými hodnotami odhadů.

4.3 Konfidenční oblasti pro body regresní přímky

Za předpokladu normality uvažujeme několik konfidenčních oblastí pro regresní přímku, konkrétně se jedná o intervaly spolehlivosti pro jednotlivé parametry β_1 a β_2, elipsu spolehlivosti pro $\boldsymbol{\beta}$, konfidenční oblast pro regresní přímku (každý bod přímky zvlášť) či simultánní konfidenční oblast (všechny body přímky

dohromady). Lze ovšem odvodit též konfidenční oblast pro skutečné hodnoty bodů $P_i(\mu_i, \nu_i)$.

Dle [5] uvažujme náhodný vektor $(\mathbf{X}', \mathbf{Y}')'$, který je normálně rozdělen. Konfidenční oblast pro bod P_i je dána:

$$\mathcal{E}_{1-\alpha}(P_i) = \left\{\mathbf{u} : \mathbf{u} \in \mathbb{R}^2, (u_1 - \hat{\mu}_i, u_2 - \hat{\nu}_i) \times \right.$$
$$\left. \times \begin{pmatrix} \mathrm{var}[\hat{\mu}_i] & \mathrm{cov}[\hat{\mu}_i, \hat{\nu}_i] \\ \mathrm{cov}[\hat{\mu}_i, \hat{\nu}_i] & \mathrm{var}[\hat{\nu}_i] \end{pmatrix}^{-1} \begin{pmatrix} u_1 - \hat{\mu}_i \\ u_2 - \hat{\nu}_i \end{pmatrix} \leq \chi^2_2(1-\alpha) \right\},$$

přičemž $\mathrm{var}[\hat{\mu}_i], \mathrm{cov}[\hat{\mu}_i, \hat{\nu}_i]$ a $\mathrm{var}[\hat{\nu}_i]$ jsou i-té diagonální prvky matic uvedených ve Větě 3.3.

V případě, že parametr σ^2 neznáme, je $100(1-\alpha)\%$ konfidenční oblast pro bod P_i vyjádřena dle [5] jako

$$\mathcal{E}_{1-\alpha}(P_i) = \left\{\mathbf{u} : \mathbf{u} \in \mathbb{R}^2, (u_1 - \hat{\mu}_i, u_2 - \hat{\nu}_i) \times \right.$$
$$\left. \times \begin{pmatrix} \widehat{\mathrm{var}}[\hat{\mu}_i] & \widehat{\mathrm{cov}}[\hat{\mu}_i, \hat{\nu}_i] \\ \widehat{\mathrm{cov}}[\hat{\mu}_i, \hat{\nu}_i] & \widehat{\mathrm{var}}[\hat{\nu}_i] \end{pmatrix}^{-1} \begin{pmatrix} u_1 - \hat{\mu}_i \\ u_2 - \hat{\nu}_i \end{pmatrix} \leq 2F_{2,n-2}(1-\alpha) \right\},$$

přičemž $\widehat{\mathrm{var}}[\hat{\mu}_i], \widehat{\mathrm{cov}}[\hat{\mu}_i, \hat{\nu}_i]$ a $\widehat{\mathrm{var}}[\hat{\nu}_i]$ jsou i-té diagonální prvky matic uvedených ve Větě 3.3, kde parametr σ^2 je nahrazen odhadem $\hat{\sigma}^2$.

4.4 Ortogonální regrese - alternativní přístup

Zatímco předchozí algoritmus je podrobně popsán v [5], nyní se podíváme na alternativní přístup využívající singulárního rozkladu matice.

Na začátku čtvrté kapitoly jsme zmínili singulární rozklad, který lze pro matici $\mathbf{C}$ o m řádcích a n sloupcích vyjádřit jako $\mathbf{C}_{mn} = \mathbf{U}_{mm}\mathbf{S}_{mn}\mathbf{V}'_{nn}$, vyplývá to ze článků [11] a [17]. Metoda singulárního rozkladu (SVD - Single Value Decomposition) je přitom založena na větě lineární algebry, která říká, že obdélníkovou matici $\mathbf{C}$ lze rozložit na součin tří matic - ortogonální matice $\mathbf{U}$, diagonální matice $\mathbf{S}$ a (transponované) ortogonální matice $\mathbf{V}$. Singulární rozklad je pak prezentován výše uvedeným vztahem jako

$$\mathbf{C}_{mn} = \mathbf{U}_{mm}\mathbf{S}_{mn}\mathbf{V}'_{nn},$$

kde $\mathbf{U}'\mathbf{U} = \mathbf{I}$, $\mathbf{V}'\mathbf{V} = \mathbf{I}$. Sloupce matice $\mathbf{U}$ jsou ortonormální vlastní vektory součinu $\mathbf{C}\mathbf{C}'$, sloupce $\mathbf{V}$ jsou ortonormální vlastní vektory $\mathbf{C}'\mathbf{C}$.

Matice $\mathbf{S}$ je diagonální. Její nenulové prvky jsou odmocniny nenulových vlastních čísel matice $\mathbf{C}'\mathbf{C}$ nebo $\mathbf{C}\mathbf{C}'$ v sestupném pořadí, ostatní prvky v matici $\mathbf{S}$ jsou potom rovny nule. Nenulová vlastní čísla $\mathbf{C}'\mathbf{C}$ a $\mathbf{C}\mathbf{C}'$ jsou vždy stejná, proto nezáleží, ze které matice čísla vybíráme.

Ukažme si konkrétně, jak bude singulární rozklad vypadat. Položme $[\mathbf{X}, \mathbf{z}_2] = \mathbf{USV}'$, zároveň připomeňme, že $\mathbf{X} \in \mathbb{R}^{n\times 2}$. Jak budou vypadat jednotlivé matice?

$$\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2], \mathbf{U}_1 = [\mathbf{u}_1, \mathbf{u}_2], \mathbf{U}_2 = [\mathbf{u}_3, \ldots, \mathbf{u}_n], \mathbf{u}_i \in \mathbb{R}^n, \mathbf{U}'\mathbf{U} = \mathbf{I}_n,$$

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3], \mathbf{v}_i \in \mathbb{R}^3, \mathbf{V}\mathbf{V}' = \mathbf{I}_3,$$

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix},$$

$$\mathbf{S}_1 = \mathrm{diag}(\sigma_1, \sigma_2) \in \mathbb{R}^{2\times 2}, \mathbf{S}_2 = \sigma_3 \in \mathbb{R}^1, \sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0.$$

Hodnost matice $\mathbf{S}$ je určena počtem nenulových singulárních hodnot.

Poznatků využijeme v alternativním algoritmu ortogonální regrese (ekvivalentně též úplných nejmenších čtverců). Užitečnou pro charakterizaci řešení odhadu parametru $\widehat{\boldsymbol{\beta}}$ se ukázala následující Věta 4.4.

Věta 4.4 (Explicitní vyjádření odhadu metody ortogonální regrese)**.** *[20] Nechť máme singulární rozklad* $[\mathbf{X}, \mathbf{z}_2]$*. Jestliže* $\sigma_2 > \sigma_3$*, potom*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} - \sigma_3^2\mathbf{I})^{-1}\mathbf{X}'\mathbf{z}_2$$

a

$$\sigma_3^2\left[1 + \frac{(\mathbf{u}_1'\mathbf{z}_2)^2}{\sigma_1 - \sigma_3^2} + \frac{(\mathbf{u}_2'\mathbf{z}_2)^2}{\sigma_2 - \sigma_3^2}\right] = \min_{\mathbf{x}\in\mathbb{R}^2} ||\mathbf{X}\boldsymbol{\beta} - \mathbf{z}_2||^2.$$

Důkaz: Viz [20], str. 36, Theorem 2.7.

□

Nabyté poznatky nyní shrneme do základního algoritmu alternativního přístupu k ortogonální regresi.

Algoritmus 4.2. *[20] Základní řešení* $\mathbf{X}\boldsymbol{\beta} \approx \mathbf{z}_2$ *metodou úplných čtverců.*

Máme dáno: $\mathbf{X} \in \mathbb{R}^{m\times 2}$ *a* $\mathbf{z}_2 \in \mathbb{R}^m$.

Krok 1. *Spočítáme singulární rozklad* $[\mathbf{X}, \mathbf{z}_2] = \mathbf{USV}'$.

Krok 2. *Jestliže* $v_{3,3} \neq 0$*, potom* $\widehat{\boldsymbol{\beta}} = 1/(v_{3,3})[v_{1,3}, \ldots, v_{2,3}]'$.

KONEC

Pomocí výše uvedeného algoritmu bychom dosáhli stejných odhadů jako dle kapitoly 4.1, navíc je tento numericky efektivní. Na druhou stranu nám ale neumožňuje odvození jednotlivých nástrojů statistické inference, což se z pohledu statistické analýzy jeví jako zásadní problém.

4.5 Porovnání ortogonální regrese a metody hlavních komponent

Metoda hlavních komponent je vhodná tehdy, jestliže máme mnoho proměnných, jejichž počet chceme redukovat. Dáme-li poznatky do souvislostí s touto prací, vidíme, že trojsložkové kompozice přepočítáme na dvě souřadnice. Nejprve stručně pohovoříme o metodě hlavních komponent, poté se zaměříme na situaci, kdy máme rovinná data. Následně zjistíme, jaký vztah je mezi metodou hlavních komponent a ortogonální regresí.

Metoda hlavních komponent vysvětluje původní proměnné menším počtem nových proměnných, které vznikly lineární kombinací původních proměnných. Ačkoli má tedy původní datový soubor například $D-1$ proměnných, často lze kovarianční strukturu datového souboru popsat menším počtem hlavních komponent [16].

Algoritmus 4.3. *[16]*

Vstup: *Máme náhodný výběr pozorování* $\mathbf{z}_1, \ldots, \mathbf{z}_n$ *o* $D-1$ *proměnných.*

Krok 1. *Vektory* $\mathbf{z}_1, \ldots, \mathbf{z}_n$ *tvoří matici* $\mathbf{Z}$ *o rozměrech* $n \times (D-1)$.

Krok 2. *Stanovíme výběrový průměr jako* $\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$.

Krok 3. *Spočítáme výběrovou varianční matici*

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'.$$

Krok 4. *Vypočítáme vlastní čísla* λ_k *a vlastní vektory* $\mathbf{u}_k$ *matice* $\mathbf{U}\mathbf{\Lambda}\mathbf{U}', k = 1, \ldots, D-1$. *Vlastní vektor* $\mathbf{u}_k$ *nastavíme jako sloupce matice*

$$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{D-1}).$$

Jejich pořadí v matici $\mathbf{U}$ *by mělo odpovídat uspořádání hodnot odpovídajících vlastních čísel (a to sestupně).*

Jak uvádí [4], po singulárním rozkladu $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$, kde $\mathbf{\Lambda}$ je matice vytvořená z vlastních čísel a matice $\mathbf{U}$ z vlastních vektorů matice $\mathbf{S}$, můžeme definovat transformaci metodou hlavních komponent jako

$$\mathbf{Z}^* = (\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}')\mathbf{U}.$$

Matice $\mathbf{Z}^*$ má dle článku [4] stejnou hodnost jako $\mathbf{Z}$. Sloupce matice $\mathbf{Z}^*$ nazýváme *skóry* j-té hlavní komponenty. Sloupce matice $\mathbf{U}$ se nazývají *zátěže* j-té hlavní komponenty, představují vliv původních proměnných na nové hlavní komponenty.

Při redukci dimenze původní matice již pouze několik prvních hlavních komponent pokrývá důležitou informaci o datech (ztotožněnou s celkovou variabilitou datového souboru).

Transformace představuje rotaci původního souřadnicového systému na nový s vlastnostmi ve výše uvedeném smyslu.

Mezi metodou hlavních komponent a ortogonální regresí můžeme pozorovat velmi blízký vztah. Uvažujme náhodné proměnné $\mathbf{z}_1$ a $\mathbf{z}_2$ a jejich varianční matici

$$\boldsymbol{\rho} = \begin{pmatrix} s_{z_1}^2 & s_{z_1 z_2} \\ s_{z_2 z_1} & s_{z_2}^2 \end{pmatrix}.$$

Vlastní čísla spočítáme jednoduše ze známé formule $|\mathbf{A} - \lambda\mathbf{I}| = 0$. Dodejme, že zřejmě $s_{z_1 z_2} = s_{z_2 z_1}$, proto součin členů budeme jednoduše počítat jako $s_{z_1 z_2}^2$. Konkrétní podoba je tato:

$$\begin{vmatrix} s_{z_1}^2 - \lambda & s_{z_1 z_2} \\ s_{z_2 z_1} & s_{z_2}^2 - \lambda \end{vmatrix} = (s_{z_1}^2 - \lambda)(s_{z_2}^2 - \lambda) - s_{z_1 z_2}^2 \stackrel{!}{=} 0.$$

Pro snadnější dosazení do vzorce pro výpočet kořenů kvadratické rovnice uspořádáme,

$$\lambda^2 - \lambda(s_{z_1}^2 + s_{z_2}^2) + (s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2) = 0.$$

Vlastní čísla matice $\boldsymbol{\rho}$ jsou tato:

$$\lambda_1 = \frac{s_{z_1}^2 + s_{z_2}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}}{2},$$

$$\lambda_2 = \frac{s_{z_1}^2 + s_{z_2}^2 - \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}}{2}.$$

Zřejmě platí, že $\lambda_1 > \lambda_2$. Proto největším vlastním číslem je λ_1. Z něho získáme vlastní vektor matice $\boldsymbol{\rho}$, čili první hlavní komponentu odpovídající tomuto vlastnímu číslu. Vlastní vektor vypočítáme ze vztahu $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$. Uvedenou rovnici přepíšeme,

$$\begin{pmatrix} s_{z_1}^2 & s_{z_1 z_2} \\ s_{z_2 z_1} & s_{z_2}^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \frac{s_{z_1}^2 + s_{z_2}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}}{2} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

Z uvedeného vztahu získáme soustavu dvou rovnic o dvou neznámých parametrech u_1, u_2,

$$\begin{aligned} 2s_{z_1 z_2} u_2 &= \left(s_{z_2}^2 - s_{z_1}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}\right) u_1, \\ 2s_{z_2}^2 u_2 &= \left(s_{z_1}^2 + s_{z_2}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)} - 2s_{z_1 z_2}\right) u_1. \end{aligned}$$

Dále vyjádříme u_2 z první rovnice,

$$u_2 = \frac{\left(s_{z_2}^2 - s_{z_1}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}\right)}{2s_{z_1 z_2}} u_1.$$

Dosadíme do druhé rovnice. Musíme dojít k závěru, že rovnice jsou svými násobky, po zjednodušení tedy vychází $u_1 = u_1$. Řešením jsou tak všechna reálná čísla. Jako parametr položíme $t = s_{z_1,z_2}$, důvodem je možnost zkrácení tohoto výrazu ve jmenovateli výrazu u_2. Jednoduše dosazením do vzorce pro u_2 a dostáváme výsledný tvar vlastního vektoru pro vlastní číslo λ_1.

Vlastní vektor pro λ_1 tedy vychází:

$$\mathbf{u}_1 = \left(s_{z_1 z_2}, \frac{s_{z_2}^2 - s_{z_1}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}}{2} \right),$$

pro úplnost ještě uveďme vlastní vektor pro λ_2:

$$\mathbf{u}_2 = \left(\frac{s_{z_2}^2 - s_{z_1}^2 + \sqrt{(s_{z_1}^2 + s_{z_2}^2)^2 - 4(s_{z_1}^2 s_{z_2}^2 - s_{z_1 z_2}^2)}}{2}, -s_{z_1 z_2} \right).$$

Vidíme, že u většího vlastního čísla dochází ke shodě s Větou 4.1.

5 Praktické aplikace

Uvedené teoretické poznatky aplikujeme na dva příklady s reálnými daty. Oba příklady jsou vlastní a dosud nebyly nikde publikovány. První zpracovává data týkající se typů půdy ve 27 členských státech Evropské unie. Druhý se zabývá daty poslechovosti ve všech okresech České republiky, přičemž rádia jsou rozdělena na tři skupiny, které si popíšeme níže.

Veškeré výpočty v této diplomové práci byly provedeny ve volně dostupném statistickém softwaru R [19]. Vhodný je především proto, že dokáže plně zastoupit drahé komerční softwary a je též poměrně uživatelsky přívětivý. Jak už bylo uvedeno v [15], velmi často se s jeho využitím setkáváme jak ve státní, tak i soukromé sféře. Nejdůležitější použité příkazy a algoritmy budou uvedeny a okomentovány v následujícím textu. Stěžejní částí obou příkladů je využití knihovny `compositions` [3], resp. `robCompositions` [8]. Jak už názvy napovídají, knihovny byly vytvořeny pro práci s kompozičními daty.

Před samotným počítáním příkladů musíme v softwaru `R` načíst knihovnu `robCompositions`, která slouží pro statistickou analýzu kompozičních dat. To provedeme pomocí příkazu

```
> library(robCompositions).
```

Knihovna obsahuje klasické i robustní statistické metody pro statistické zpracování kompozičních dat, umožňuje tak například provést metodu hlavních komponent, diskriminační a faktorovou analýzu a zejména *log-ratio transformace* (*alr, clr, ilr*). Podrobněji se k algoritmům dostaneme u prvního příkladu, na kterém bude ilustrována problematika diplomové práce. Poznatky následně použijeme rovněž u druhého příkladu souvisejícího s poslechovostí rádií.

V příkladech budeme pracovat s datovými soubory, v nichž se přirozeně vyskytují odlehlá pozorování. Přestože je v datech vždy zajímavý trend, což nám umožní ortogonální regresi použít, přítomnost odlehlých hodnot přece jen trochu zkomplikuje interpretaci dosažených výsledků. Ostatně tento problém je znatelný už při prvním příkladu.

5.1 Členění půdy ve státech EU

První, původní příklad zkoumá souvislost mezi třemi typy půdy ve 27 členských státech Evropské unie. Datový soubor pochází z [7], konkrétně se jedná o plochy v rámci států, které tvoří postupně *Forest area* (lesní plocha), *Arable land* (orná půda - slouží k pěstování plodin, nezapočítává se zde potenciálně kultivovatelná půda, jedná se o jeden ze tří typů zemědělské půdy) a *Permanent crops* (trvalé kultury - plodiny, rovněž je zemědělskou půdou, slouží ale pro pěstování ovocných sadů nebo vinic). Pro úplnost připomeňme, že třetím typem zemědělské půdy jsou louky a pastviny, v příkladu ale nevystupuje. Z již zmíněného datového souboru jsme vybrali právě tyto tři kategorie půdy, jelikož se

tyto vyskytují u všech členských států (s nulovými hodnotami nelze pracovat bez uvážení dalšího matematického aparátu, což již přesahuje zaměření této práce). Zároveň nespadají jednotlivé kategorie pod jiné a jsou svébytné.

Čísla v uvedeném datovém souboru značí plochu v jednotkách km^2, již zaujímají zmíněné tři typy půdy ve státech Evropské unie. Protože nás ovšem zajímají relativní příspěvky jednotlivých typů půdy, jedná se o kompoziční datový soubor.

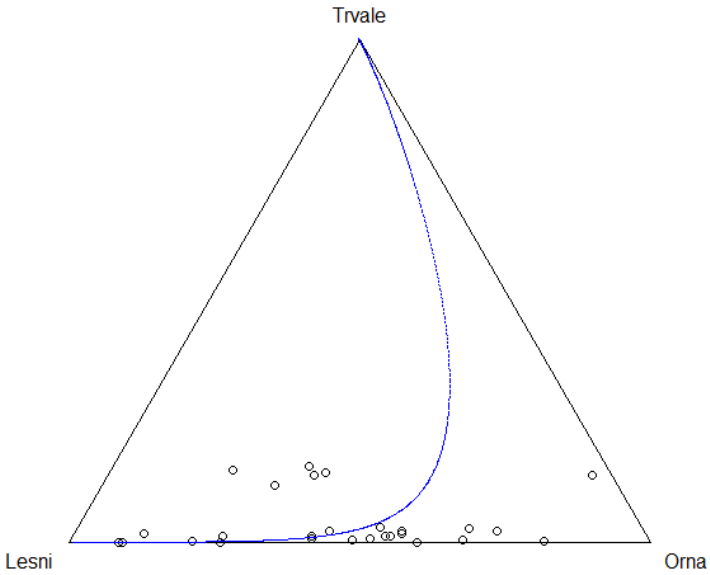
Na začátku musíme načíst data ze zdrojového souboru. Provedeme to pomocí následujícího příkazu, kde `data` značí náš datový soubor:

```
> data=read.table("data_poslechovost.txt").
```

Písmenem `n` označíme počet řádků v datové tabulce `data`, v našem případě bude příkaz následující

```
> n=nrow(data).
```

Datový soubor zobrazíme v *ternárním diagramu*, tento typ grafu je vhodný pro zobrazení kompozičních dat o třech složkách, jedná se vlastně o *simplex* jako jejich výběrový prostor. Součty složek musí být vždy rovny konstantě (zpravidla se stanovuje jako 1 nebo 100), zobrazujeme tak vlastně reprezentace kompozic. Ternární diagram pro představený datový soubor je uveden níže.



Ternární diagram vyvoláme jednoduše pomocí příkazu

```
> ternaryDiag(data),
```

kde `data` je označení našeho datového souboru, pro který daný graf aktuálně vykreslujeme. *Lesni* značí lesní plochu v jednotlivých státech, dále *Orna* ornou půdu a *Trvale* trvalé plodiny. Chceme-li též do ternárního diagramu vykreslit i regresní přímku datového souboru, musíme použít jiný postup. Nejprve vygenerujeme všechna celá čísla od 1 do 2000.

```
> RP=cbind(1:2000,1:2000)
```
Zavedeme cyklus a následně provedeme inverzní ilr transformaci na datový soubor.

```
> for(i in 1:2000)\{
> RP[i,1]=-10+i/100
> RP[i,2]=c[1]+c[2]*(-10+i/100)
\}
> RP1=invilr(-RP)
```

V posledním kroku vykreslíme kompoziční graf s regresní přímkou (vykreslena modrou barvou) na ternárním diagramu.

```
> plot.acomp(RP1,pch='.',add=TRUE,col="blue").
```

Odtud sledujeme, že se hodnoty datového souboru realizují prakticky výhradně na spodní hraně simplexu (strana Lesni, Orna), problém může být s normalitou, kterou záhy otestujeme. Z ternárního diagramu, resp. z hodnot datového souboru vidíme, že půda s trvalými plodinami zaujímá nejmenší podíl vzhledem k ostatním složkám - orné půdě a lesními plochami. Výjimkou je stát Malta, kde je vyšší plocha půdy s trvalými plodinami než v případně lesní půdy.

Abychom mohli následně aplikovat příslušnou statistickou inferenci, musíme nejprve otestovat, zda hodnoty kompozičních dat pochází z normálního rozdělení na simplexu. Bez tohoto předpokladu není možné získat například oblasti spolehlivosti či testovat hypotézy o významnosti regresních parametrů.

Pro testování normality na kompozičních datech jsme zvolili *Anderson-Darlingův test*. Jako vhodný testovací nástroj je navrhl John Aitchison v knize [1]. Funkce `adtestWrapper` knihovny `robCompositions` nejprve pomocí ilr-transformace vypočítá příslušné souřadnice, na které (zvlášť i dohromady) aplikuje zmíněný Anderson-Darlingův test. Zadáme-li v softwaru `R` následující příkaz,

```
> summary(adtestWrapper(data))
```

kde `data` je označení našeho datového souboru, získáme výsledky testování. Nezadáme-li bližší parametry, aplikace automaticky uvažuje hladinu testu $\alpha = 0{,}05$, počet simulací Monte Carlo pro získání příslušných p-hodnot bude roven jednomu tisíci. Funkce `summary` slouží k zobrazení přehledu dosažených výsledků testování. Přejděme k interpretaci výsledků z následující tabulky.

Složka	Hodnota statistiky	p-hodnota
z_1	1,1383	0,008
z_2	0,4522	0,329
z_1, z_2	0,6386	0,239

První dva řádky ukazují, kterak jsou na tom dvě složky kompozici z pohledu normality. Jelikož p-hodnota je nižší než námi stanovená hladina 0,05, musíme

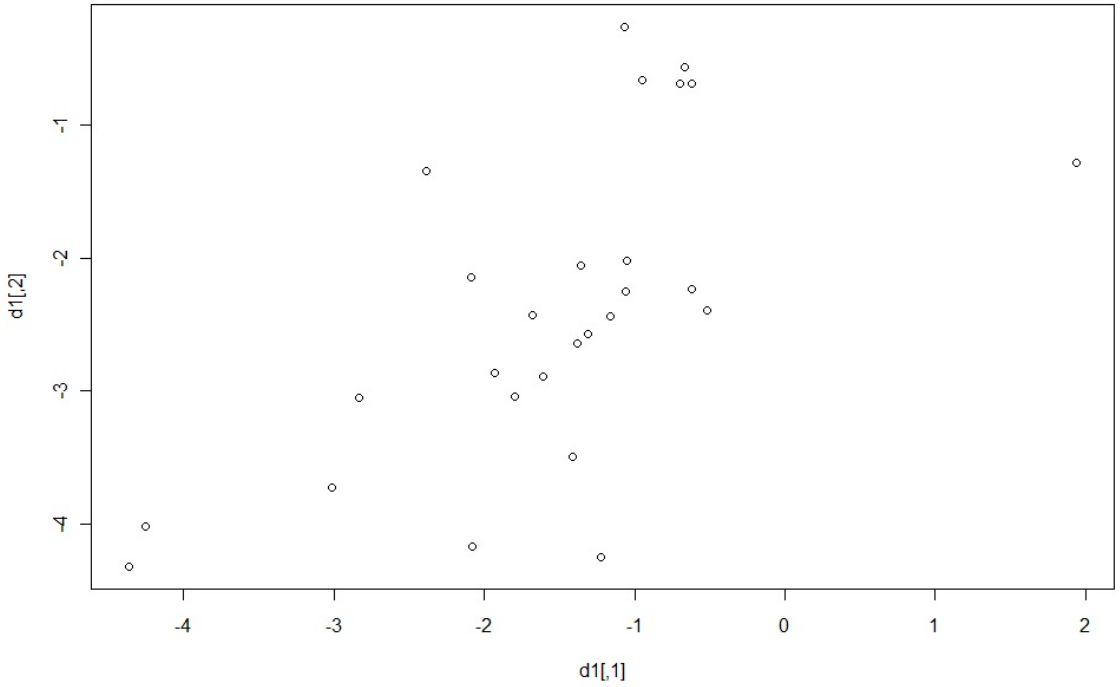
v případě první souřadnice zamítnout nulovou hypotézu o tom, že data pochází z normálního rozdělení. Ovšem u druhé souřadnice naopak na hladině 0,05 nulovou hypotézu ve prospěch alternativy zamítnout nemůžeme a data mají normální rozdělení. Ačkoli oba testy na jednotlivé souřadnice vychází rozdílně, testujeme-li obě souřadnice současně, nulovou hypotézu o tom, že data jsou z normálního rozdělení, na hladině 0,05 zamítnout nemůžeme. Celkově tak přes negativní výsledek u první souřadnice můžeme vzhledem k rozsahu datového souboru předpokládat, že data pochází z normálního rozdělení. Lze také získat potřebné charakteristiky. Při interpretaci výsledků ale budeme muset být přece jeden opatrní.

Protože jsme zvyklí pracovat v ortonormální (kartézské) soustavě souřadnic, museli jsme provést transformaci kompozičních dat. V opačném případě by totiž standardní statistické metody nebylo možné použít bez potíží a jejich výsledky by byly značně zkreslené. Jak již bylo uvedeno, jako vhodná se jeví právě ilr-transformace, například dle vztahu (1) z kapitoly 1.2. Tu aplikujeme zcela jednoduše pomocí příkazu

```
> datai=ilr(data),
```

kde jako `datai` označíme nový, transformovaný datový soubor. Graf s daty vykreslíme pomocí příkazu

```
> plot(datai).
```



Data jsme transformovali a vykreslili. Vidíme, že několik hodnot je odlehlých vzhledem k hlavnímu datovému trendu. Zejména si povšimněme pozorování v pravém horním rohu, který odpovídá složení půdy na Maltě, je zřejmě též hlavní příčinou zamítnutí normality v první souřadnici.

Ukažme si, jak pomocí softwaru `R` bude vypočítána a provedena ortogonální regrese pomocí Věty 4.1. Nejprve provedeme výchozí inicializaci vektoru regresních koeficientů,

`> b1=c(NA,NA)`, přičemž `NA` značí chybějící hodnotu.

Jako `V` označíme varianční matici transformovaného datového souboru, který jsme pojmenovali `datai`. Druhou složku vektoru `b1` vypočítáme jako

```
> b1[2]=(V[2,2]-V[1,1]+
sqrt((V[2,2]-V[1,1])^2+4*V[1,2]^2))/(2*V[1,2]).
```

Z transformovaného datové matice `datai` určíme aritmetické průměry sloupců odpovídajících jednotlivým souřadnicím. Výsledek označíme písmenem `m`.

```
> m=apply(datai,2,mean)
```

Číslo `2` uvedené v příkazu o řádek výše znamená, že zadaná funkce `mean` se počítá pouze přes sloupce, což právě požadujeme. Pro počítání s řádky by se v případě potřeby použilo analogicky číslo `1`. Složku `b1[1]` vypočítáme pomocí příkazu:

`> b1[1]=m[2]-b1[2]*m[1]`.

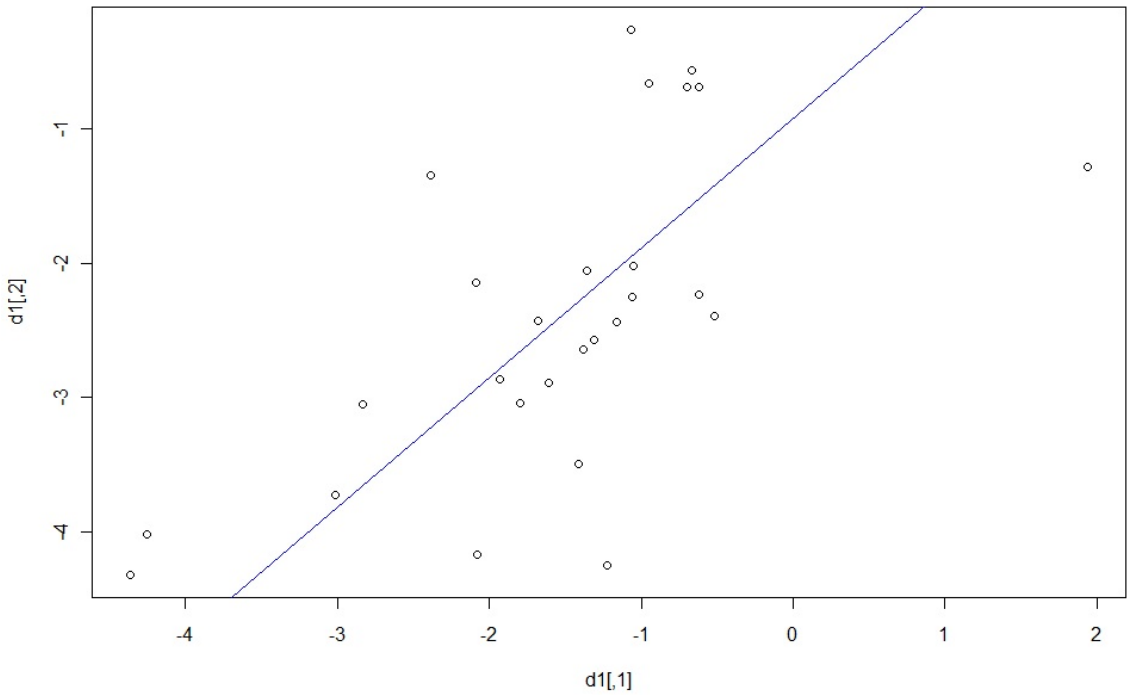
Vektor odhadu parametrů regresní přímky po výpočtu ve statistickém softwaru `R` vychází následovně:

$$\widehat{\boldsymbol{\beta}} = (-0{,}9195; 0{,}9639).$$

Regresní přímku vykreslíme pomocí příkazu `abline`.

```
> abline(b1[1],b1[2],col='blue')
```

Na grafu je modře znázorněna regresní přímka transformovaného datového souboru.



Takto bychom postupovali v případě klasické ortogonální regrese podle Věty 4.1 (resp. Věty 4.2). Ukážeme si však rovněž druhý, alternativní postup pomocí teorie lineárních regresních modelů, v němž aplikujeme poznatky z diplomové práce. Výsledek by v případě správnosti měl vyjít shodný s již získaným odhadem vektoru $\widehat{\boldsymbol{\beta}}$.

V následujícím postupu si přeznačíme písmenem `c` hledaný vektor regresních parametrů. Před zahájením samotného algoritmu pro výpočet odhadu parametrů pro regresní přímku nejprve musíme nadefinovat několik proměnných. Vektor $\boldsymbol{\mu}$ inicializujeme jako

```
> mu=datai[,1],
```

$\boldsymbol{\nu}$ potom jako

```
> nu=c[1]*rep(1,n)+c[2]*mu.
```

Příkaz `rep` přitom v uvedené podobě vytvoří vektor jedniček o n složkách.

Nesmíme zapomenout na výchozí stanovení vektoru parametrů `c1=c`, dále pak pro počáteční odhady $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$ položíme `mu0=mu`, `nu0=nu` a `c0=c`. Pro úplnost označme `datac=c(1,1)` jako výchozí hodnoty složek vektoru ukončovacího kritéria algoritmu. Průběh samotného iterativního dle Věty 4.3. je uveden v Příloze.

Celkem muselo proběhnout při použití dané-ilr transformace celkem 27 iterací. Právě tolik jich nastane do chvíle, než odhady konvergují a bude splněno ukončovací kritérium. To jsme stanovili jako $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}||^2 < 10^{-9}$, tedy s využitím druhé mocniny euklidovské normy rozdílu vektorů odhadů regresních koeficientů z aktuálního a předchozího kroku iteračního algoritmu. Získali jsme tak výsledný

odhad parametru $\boldsymbol{\beta}$. Jak vidíme, je shodný s předchozími výsledky získanými pomocí Věty 4.1,

$$(\hat{\beta}_1, \hat{\beta}_2)' = (-0{,}9195; 0{,}9639)'.$$

Nyní vypočítáme varianční matici odhadu $\widehat{\boldsymbol{\beta}}$ a to opět za pomoci statistického softwaru R. Postup je následující,

```
> var_c=s2*(c[2]^2+1)*solve(cbind(c(n,t(mu)%*%rep(1,n)),
c(t(mu)%*%rep(1,n),t(mu)%*%mu))),
```

kde s2 je odhad σ^2 a získá se jako:

```
> s2=(t(datai[,1]-mu)%*%(datai[,1]-mu)+t(datai[,2]-nu)%*%
(datai[,2]-nu))/(n-2).
```

Varianční matice vychází ve tvaru:

$$\widehat{\text{var}} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 0{,}1238 & 0{,}0541 \\ 0{,}0541 & 0{,}0355 \end{pmatrix}.$$

Pro získanou regresní přímku dále určíme pás spolehlivosti pro přímku (v následujícím grafu jej vykreslíme černou barvou), poté i pás spolehlivosti kolem přímky (v grafu je znázorněn červeně). Uveďme si algoritmy, které jsme použili v softwaru R pro získání těchto křivek.

Písmenem t označíme generování posloupnosti čísel z intervalu $\langle -10, 10\rangle$ s krokem 0,01, právě tato délka kroku je ideální pro dostatečně kvalitní zobrazení křivky. Členění lze rozdělit na menší kroky (například 0,001), počítání v softwaru R je ovšem časově a početně náročnější.

Nejprve vypočítáme hodnoty pro horní křivku pásu spolehlivosti pro přímku (upper), poté pro dolní (lower), přičemž využijeme 0,95-kvantil Fisherova rozdělení o příslušných stupních volnosti. Jednotlivé body poté vykreslíme pomocí funkce points. Následující algoritmus ukazuje, jak získat pás spolehlivosti pro přímku.

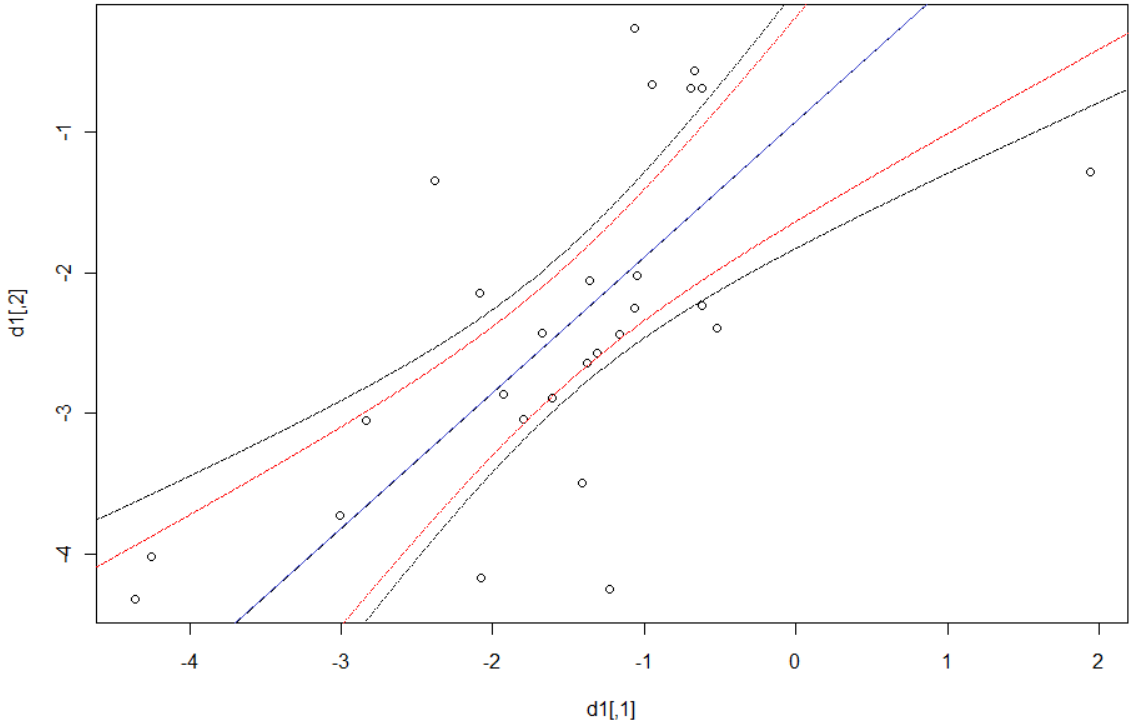
```
> t=seq(-10,10,0.01)
> upper=rep(NA,length(t))
> for(i in 1:length(t)){
> upper[i]=c[1]+c[2]*t[i]+sqrt(2*qf(0.95,2,n-2))*
  sqrt(t(c(1,t[i]))%*%var_c%*%c(1,t[i]))
}
> points(t,upper,pch='.')

> lower=rep(NA,length(t))
> for(i in 1:length(t)){
```

```
> lower[i]=c[1]+c[2]*t[i]-sqrt(2*qf(0.95,2,n-2))*
sqrt(t(c(1,t[i]))%*%var_c%*%c(1,t[i]))
}
> points(t,lower,pch='.')
```

U pásu spolehlivosti kolem přímky je postup zcela analogický, horní a dolní odhady se počítají podle upraveného vzorce. Níže je uveden případ pro horní část, dolní se liší jen znaménkem,

```
> upper[i]=c[1]+c[2]*t[i]+qt(0.975,n-2)*
sqrt(t(c(1,t[i]))%*%var_c%*%c(1,t[i])).
```



Na grafu vidíme, že jak pás spolehlivosti pro přímku, tak i pás spolehlivosti kolem přímky jsou si velmi blízké, nejspíše jsou v bodě odpovídajícím aritmetickým průměrům hodnot jednotlivých souřadnic.

Jak byly stanoveny ilr souřadnice, které odpovídají zadaným složkám kompozice? Výše užitá volba souřadnic odpovídá vztahu

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{\sqrt{x_2 x_3}}{x_1}, z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_3}{x_2}, \tag{21}$$

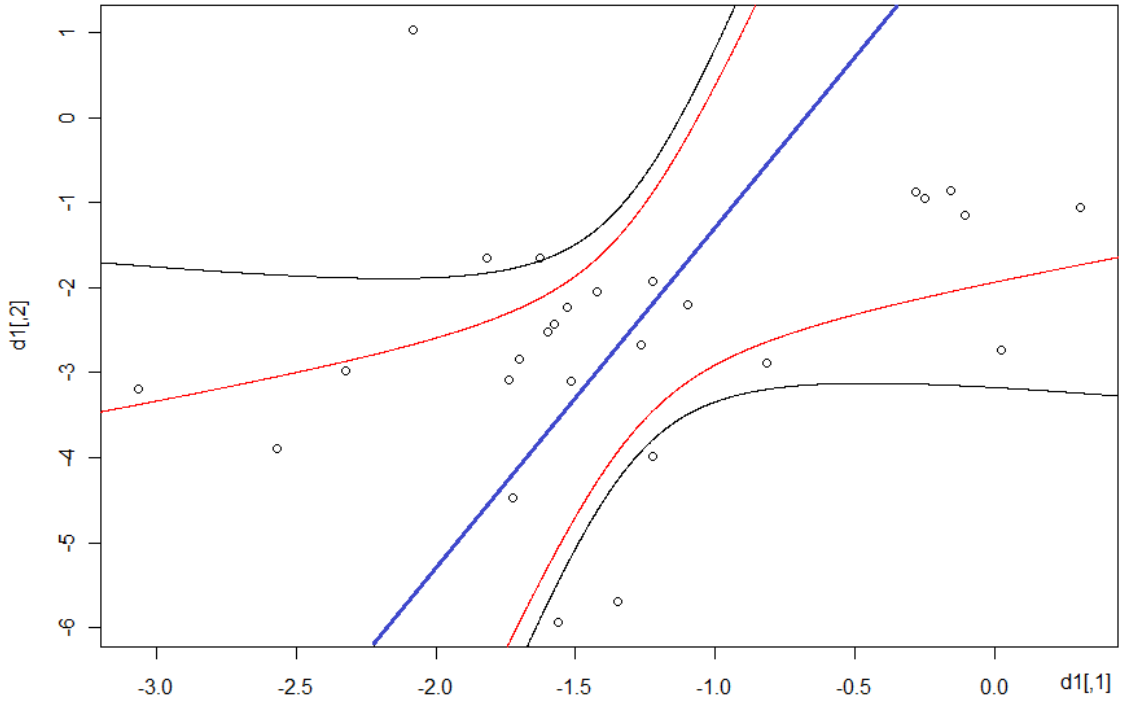
druhou volbu souřadnic vyjadřuje vztah

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{\sqrt{x_1 x_3}}{x_2}, z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_3}{x_1}, \tag{22}$$

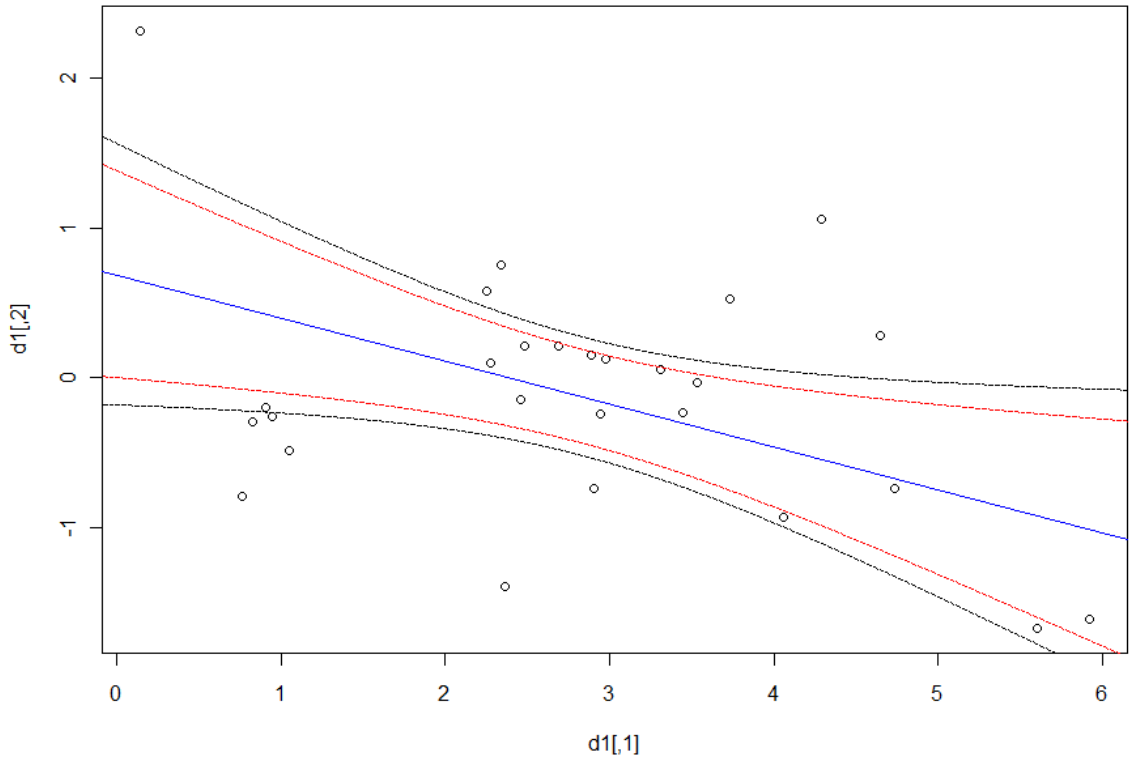
a v neposlední řadě třetí možnost, jak zvolit souřadnice pro naše kompoziční data,

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{\sqrt{x_1 x_2}}{x_3}, z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_1}. \tag{23}$$

Při volbě různých souřadnic se samozřejmě budou lišit vypočítané hodnoty v transformovaném kompozičním datovém souboru. Proto budou rovněž prezentovány grafy pro druhou a třetí volbu souřadnic.



Graf zobrazuje pásy spolehlivosti pro přímku (černě) a kolem přímky (červeně) v případě volby druhých souřadnic. Sledujeme, že pásy spolehlivosti nepokrývají všechna pozorování, zejména hodnoty odchýlené od hlavního trendu charakterizovaného regresní přímkou (vykreslena modře). Symetričnost pásu na obrázku je (zdánlivě) narušena v důsledku volby odlišných měřítek pro jednotlivé osy.



Zvolíme-li jako třetí souřadnice odpovídající vztahu (23), dostaneme poslední graf s pásy spolehlivosti kolem regresní přímky. Pozorujeme, že pásy jsou vzhledem k příznivé datové konfiguraci vykresleny jako symetrické (což je patrně zejména ve srovnání s grafem předchozím).

Tabulka níže uvedená ukazuje odhady regresních přímek a hodnoty odpovídajících statistik, týkajících se regresních parametrů, pro data s relativními příspěvky různých druhů půdy v 27 členských státech Evropské unie. Uvedeny jsou zde rovněž hodnoty statistik T_1, T_2 a F, odpovídající vztahům (17), (18), (20), stejně jako jejich p-hodnota. Každý řádek představuje právě jednu volbu souřadnic.

	regresní přímka směr. odchylky $(\hat{\beta}_1, \hat{\beta}_2)$	iterace	T_1 p-hodnota	T_2 p-hodnota	F p-hodnota
(21)	$\nu = -0{,}9195 + 0{,}9639\mu$ $(0{,}3518; 0{,}1883)$	27	$-2{,}6135$ $0{,}0150$	$5{,}1189$ $\ll 0{,}001$	$82{,}1980$ $\ll 0{,}001$
(22)	$\nu = +2{,}7467 + 4{,}0265\mu$ $(2{,}2782; 1{,}6802)$	35	$1{,}2056$ $0{,}2392$	$2{,}3965$ $0{,}0244$	$11{,}4535$ $0{,}0003$
(23)	$\nu = +0{,}6889 - 0{,}2877\mu$ $(0{,}3358; 0{,}1057)$	29	$2{,}0513$ $0{,}0509$	$-2{,}7224$ $0{,}0116$	$4{,}0477$ $0{,}0300$

V případě, kdy příslušné p-hodnoty statistik T_1, T_2, F jsou značně menší než 0,001 a zároveň bereme v potaz obvyklou hladinu významnosti $\alpha = 0{,}05$, jsou

odpovídající parametry statisticky významné (signifikantní). Toto je z p-hodnot statistik T_1, T_2 a F zřejmě případ u prvních souřadnic. Ovšem zvolíme-li druhé souřadnice, nemůžeme zamítnout nulovou hypotézu $H_0 : \beta_1 = 0$ a výsledný parametr pro nás tedy není statisticky významný. Parametr β_2 významný je, nulovou hypotézu $H_0 : \beta_2 = 0$ zamítáme ve prospěch alternativy $H_a : \beta_2 \neq 0$. Testujeme-li ovšem oba parametry zároveň, zjistíme, že nulovou hypotézu zamítáme ve prospěch alternativy a parametry jsou signifikantní.

Druhá volba souřadnic se obecně jeví o něco méně vhodná, protože počet iterací již vzrostl na 35.

Ve třetím řádku sledujeme, že i zde není parametr statisticky významný. Při souhrnném testu ale nulovou hypotézu zamítáme. Proběhnout musí ale o dvě iterace více oproti první volbě (tedy 29 versus 27) do okamžiku, kdy je splněno kritérium konvergence.

Kvůli různým strukturám statistických souborů po rotaci je jasné, že je nutné sledovat nejen statistiky T_1, T_2, ale rovněž F, jelikož i její hodnoty se při změně ortonormálních souřadnic mění.

Následující tabulky přehledně srovnávají intervaly spolehlivosti pro parametr $\hat{\beta}_2$ počítané různými přístupy pro různé volby souřadnic. Pro druhé souřadnice nelze tyto spočítat pomocí druhého a třetího přístupu, jelikož intervaly spolehlivosti vychází zcela nevhodně (konkrétně levá hranice intervalu vychází vyšší než pravá, interval by tak měl zápornou délku). Přehodíme-li je, bude jejich délka příliš velká.

1. SOUŘADNICE	intervaly spolehlivosti	délka intervalu
(19)	(0,5726; 1,3552)	0,7826
(15)	(0,5321; 1,7234)	1,1913
(16)	(0,5702; 1,6131)	1,0429

2. SOUŘADNICE	intervaly spolehlivosti	délka intervalu
(19)	(2,1566; 5,8965)	3,7399
(15)	NELZE	NELZE
(16)	NELZE	NELZE

3. SOUŘADNICE	intervaly spolehlivosti	délka intervalu
(19)	(−0,4643; −0,1112)	0,3531
(15)	(−0,6244; −0,0022)	0,6222
(16)	(−0,5846; −0,0314)	0,5532

Vidíme, že intervaly se do jisté míry kryjí. Nepřekrývají se ale zcela, jelikož jsou výchozí vzorce pro jejich počítání odvozovány jinak. Nejlépe (tedy jako nejužší) přitom vychází konfidenční interval (19).

Posledním krokem při počítání příkladu bude grafická vizualizace kompozičních složek regresní přímky, jedná se o jakousi obdobu regresní přímky pro kompoziční data na simplexu. Křivka nám ukazuje, jakým způsobem dochází k vzájemné substituci složek uvažované kompozice.

Funkce $\mathbf{y}(t) = (y_1(t), y_2(t), y_3(t))'$, $y_1(t) + y_2(t) + y_3(t) = 1$, pro hodnoty parametru $t \in (-4, 6)$, představuje transformaci obdržené regresnií přímky zpět na simplex, je zobrazena na následujícím grafu. Na vertikální ose jsou znázorněny proporce mezi jednotlivými typy půdy ve členských státech Evropské unie (lesní plocha, orná půda, trvalé plodiny).

Uveďme výchozí označení pro následující algoritmus, za jehož pomoci vykreslíme kompoziční přímku. Funkce `invilr` slouží pro *inverzní izometrickou log-ratio transformaci*, pro dané volby souřadnic je postupně rovna dle vztahů (3), (4) a (5) následujícímu,

$$x_1 = \exp\left(-\frac{\sqrt{2}}{\sqrt{3}} z_1^{(1)}\right), x_2 = \exp\left(\frac{1}{\sqrt{6}} z_1^{(1)} - z_2^{(1)}\right), x_3 = \exp\left(\frac{1}{\sqrt{6}} z_1^{(1)} + \frac{1}{\sqrt{2}} z_2^{(1)}\right)$$

platí pro první volbu souřadnic;

$$x_1 = \exp\left(\frac{1}{\sqrt{6}} z_1^{(2)} - z_2^{(2)}\right), x_2 = \exp\left(-\frac{\sqrt{2}}{\sqrt{3}} z_1^{(2)}\right), x_3 = \exp\left(\frac{1}{\sqrt{6}} z_1^{(2)} + \frac{1}{\sqrt{2}} z_2^{(2)}\right)$$

dostaneme při volbě druhých souřadnic. A konečně volíme-li třetí souřadnice, vychází vztahy následovně:

$$x_1 = \exp\left(\frac{1}{\sqrt{6}} z_1^{(2)} - z_2^{(3)}\right), x_2 = \exp\left(\frac{1}{\sqrt{6}} z_1^{(3)} + \frac{1}{\sqrt{2}} z_2^{(3)}\right), x_3 = \exp\left(-\frac{\sqrt{2}}{\sqrt{3}} z_1^{(3)}\right).$$

Funkce `constSum` uzavírá kompozice na konstantní součet (zde roven jedné) tak, že dělí každou složku kompozic řádkovým součtem složek.

```
> x0=invilr(-rbind(c(0,c[1]),c(0,c[1])))
> x=invilr(-rbind(c(1,c[2]),c(1,c[2])))

> x0a=as.matrix(x0[1,])
> xa=as.matrix(x[1,])
> constSum(t(x0a*xa^(-6)))
```

Poznamenejme, že při parametrickém vyjádření regresní přímky na simplexu jsme využili dříve uvedené operace perturbace a mocninné transformace. Jako `ti` budeme v dalším textu značit posloupnost hodnot ve stanoveném intervalu $\langle -4, 6\rangle$ s krokem 0,01 (`ti=seq(-4,6,0.01)`), který je dostatečný pro zřetelné zobrazení regresní přímky na simplexu (vzhledem k Aitschisonově geometrii).

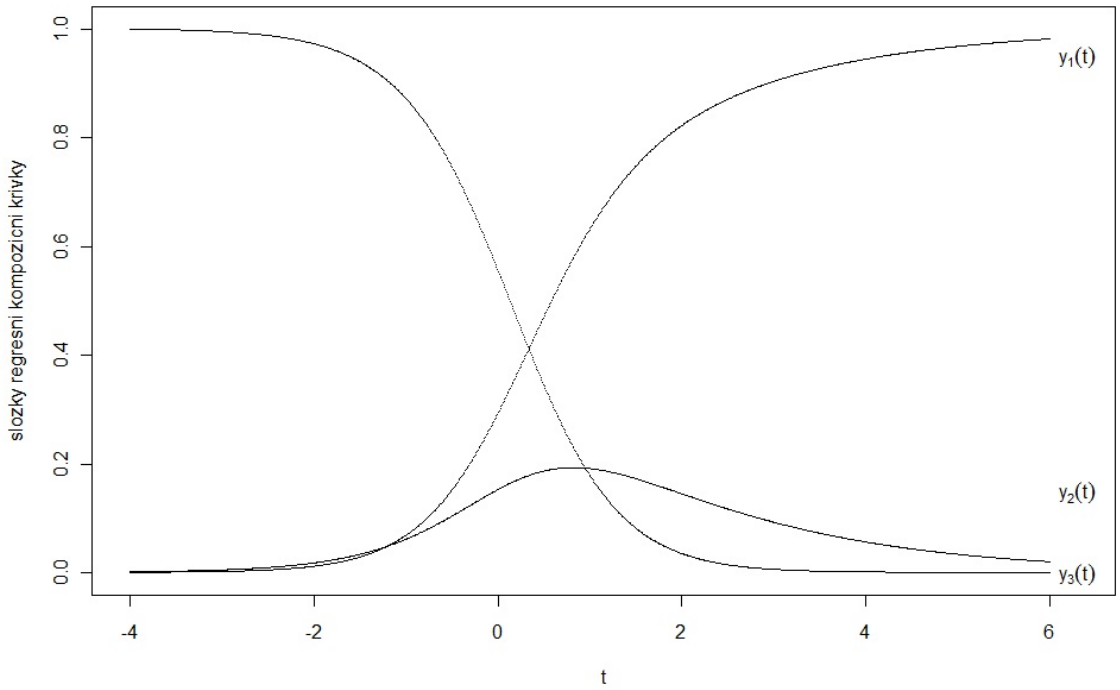
Symbolem `c1` dále označíme matici, jejíž hodnoty budeme inicializovat jako chybějící (Not Available). Počet řádků uvedené matice v zadaném algoritmu bude roven délce vektoru `ti`, tedy počtu vygenerovaných hodnot posloupnosti. Zcela logicky je počet sloupců v matici roven číslu 3.

```
> cl=matrix(NA,nrow=length(ti),ncol=3))
```

Zapíšeme tedy algoritmus, pomocí něhož budeme vizualizovat kompoziční přímku.

```
> for(i in 1:length(ti)){
> cl[i,]=constSum(t(x0a*xa^(ti[i])))
}
> plot(ti,cl[,1],xlim=c(-4,6),ylim=c(0,1),
pch='.',xlab="t",ylab="parts of the compositional line")
> points(ti,cl[,2],pch='.')
> points(ti,cl[,3],pch='.')
> text(6.3,0.6,expression(y[1](t)))
> text(6.3,0.2,expression(y[2](t)))
> text(6.3,0,expression(y[3](t)))
```

Jakmile zadáme výše uvedený algoritmus v softwaru `R`, zobrazí se nám požadovaný graf složek regresních kompozičních křivek.



Z přiloženého grafu vidíme, jak se s využitím údajů z 27 evropských států postupně mění podíly mezi jednotlivými složkami datového souboru. Jak ostatně

vyplývá z definice kompozičních dat, v tomto okamžiku nás nezajímají absolutní hodnoty, nýbrž právě relativní údaje a to vzhledem k celku (souhrnná rozloha tří půdních typů). Právě z tohoto předpokladu budeme vycházet při interpretaci grafu složek kompoziční regresní přímky vztahujících se k našemu datovému souboru.

Překvapivě nejméně výsledný průběh složek regresní přímky (na simplexu) ovlivňuje podíl orné půdy. Maximum křivky $y_2(t)$ dosahuje nejnižší hodnoty ze všech tří sledovaných. Se snižujícím se podílem půdy pro trvalé plodiny rostou současně podíly orné půdě a území, na němž rostou lesy (v případě orné půdy se po kulminaci růst mění v pokles).

Jedním z důležitých poznatků z výše uvedeného grafu je také fakt, že ve chvíli, kdy se podíl půdy, na které rostou trvalé plodiny, začne napříč státy Evropské unie postupně snižovat, výrazně stoupá podíl lesů, o něco méně tíž podíl orné půdy. Jako jedno z možných vysvětlení se nabízí možnost, že totiž využívaná zemědělská půda slouží k odlišnému účelu v zemědělství, v tomto případě právě k pěstování trvalých plodin.

Z dynamiky složek v grafu je též vidět, že v těch zemích Evropské unie, kde je vyšší podíl půdy využívané k pěstování vinné révy a pro ovocné sady (tedy k pěstování trvalých plodin), jsou vzájemné rozdíly mezi zalesněnou zemí a ornou půdou o něco nižší než v ostatních státech. Neplatí to však samozřejmě absolutně, například některé členské státy Evropské unie s nejvyšším podílem půdy určené pro trvalé plodiny vykazují větší rozdíly mezi zbývajícími dvěma složkami půdy - třeba na Maltě jsou lesy pouze ze 3% vzhledem k ostatním složkám, orná půda představuje přes 83%, zbývající podíl zaujímají pozemky s ornou půdou.

Poznamenejme, že stejný graf pro kompoziční regresní přímku jako výše uvedený by nám vyšel při použití libovolných ortonormálních souřadnic.

5.2 Poslechovost rozhlasových stanic v okresech

Druhý příklad, jenž rovněž nebyl dosud nikde publikován a je vlastní, představuje týdenní poslechovost českých rádií ve všech 77 okresech České republiky. Prezentovaná čísla udávají počet posluchačů (v tisících), kteří danou stanici poslouchali alespoň jednou v průběhu uplynulého týdne.

Data pochází z [18]. Výsledky průzkumu poslechovosti RadioProjekt vydávají agentury STEM/MARK - Median, SKMO každé tři měsíce. Datový soubor prezentuje údaje za 4. čtvrtletí roku 2010 a 1. čtvrtletí roku 2011 (sledované období je tedy od 1. října 2010 do 31. března 2011). RadioProjekt je pravidelně zkoumán na vzorku populace čítající zhruba 15 tisíc posluchačů (zpravidla je číslo nižší než tento počet, v závěru je vzorek normován), dotazovány jsou osoby ve věku od 12 do 79 let.

Soubor všech rádií v České republice nejprve rozčleníme na tři kategorie (kompoziční složky); jen pro zajímavost dodejme, že do průzkumu RadioProjekt je zapojeno téměř sto rádií. První skupinou jsou regionální rádia, podle § 2 odst. 1

písm. c) zákona 231/2001 Sb. [21] se regionálním rádiem rozumí takové, které ve vymezeném rozsahu může přijímat více než 1% a méně než 70% obyvatel České republiky. Nejúspěšnějšími regionálními stanicemi jsou například Rádio Haná, Rádio Rubi, Rádio Krokodýl a další. Trendem poslední doby je slučování rádií do tzv. *rozhlasových rodin*, resp. *sítí* (ty tvoří druhou složku kompozic). Většinou malá regionální rádia vysílají pod jednotnou značkou a to zpravidla buď s jednotným centrálním programem (například Rádio Blaník, Evropa 2, Rock Rádia), nebo s vlastním, odpojovaným programem. Tato rádia zároveň s využívají know-how provozovatele (například Hitrádio Orion, Hitrádio Dragon a jiné). Poslední skupinou jsou *celoplošné stanice*, tedy dle logiky ty, jejichž signál může přijímat více než 70% populace, jako příklad můžeme zmínit Frekvenci 1, Český rozhlas 1 - Radiožurnál nebo Rádio Impuls.

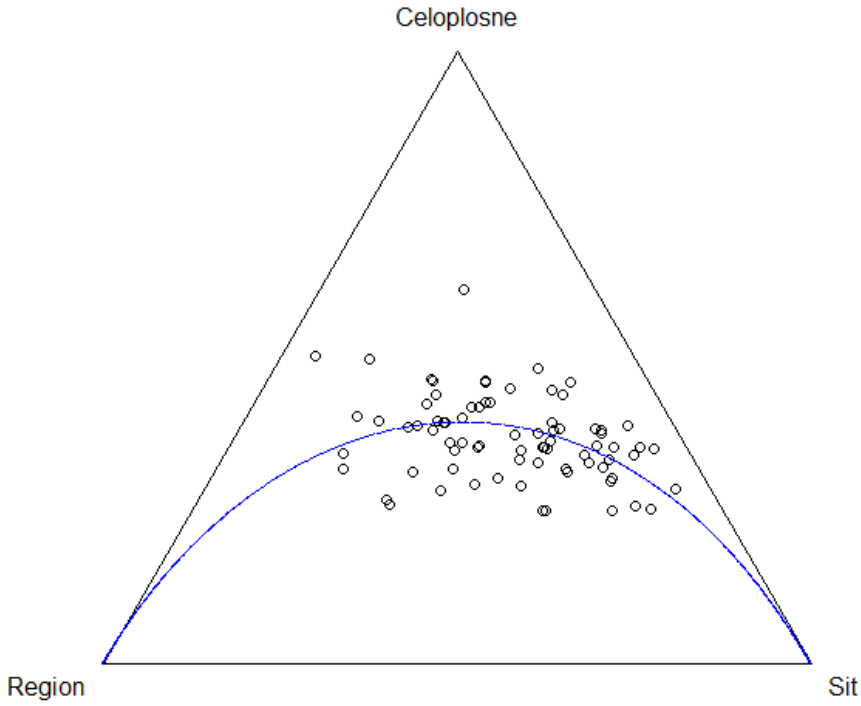
Cílem příkladu bude zjistit, jak se v různých okresech České republiky tyto stanice substituují a zda mezi jejich relativními příspěvky na celkové poslechovosti existují nějaké souvislosti. Budeme sledovat, jak ovlivní relativní nárůst jednoho typu rozhlasového subjektu zbylé dva. Sledovat budeme rovněž to, jakou váhu budou mít odlehlá pozorování.

Už z prvního pohledu na ternární diagram sledujeme, že data budou pocházet z normálního rozdělení na simplexu. Tuto hypotézu budeme ovšem muset nejprve ověřit pomocí *Anderson-Darlingova testu*.

Výsledky Anderson-Darlingova testu shrnuje následující tabulka.

Složka	Hodnota statistiky	p-hodnota
z_1	0,2744	0,708
z_2	0,7947	0,040
z_1, z_2	0,7373	0,183

U první ilr souřadnice (obdržené užitím vztahu (21)) na hladině významnosti $\alpha = 0{,}05$ nemůžeme zamítnout nulovou hypotézu o tom, že data pochází z normálního rozdělení. Zdůvodněním je p-hodnota vyšší než α, tedy hladina významnosti. U druhé souřadnice na hladině významnosti 0,05 však nulovou hypotézu zamítáme ve prospěch alternativy. Snížíme-li hladinu významnosti na 0,01, hypotézu o tom, že data pochází z normálního rozdělení, už zamítnout nemůžeme. Při testování obou souřadnic současně zjistíme, že nemůžeme zamítnout nulovou hypotézu, a data tudíž můžeme považovat za normálně rozdělená.



Vykreslíme-li si v softwaru R ternární diagram, který nám data zobrazí na simplexu, vidíme, že data se zobrazují ve střední části simplexu, blízko jeho těžiště. To signalizuje přibližně vyrovnané zastoupení všech složek kompozic v datovém souboru. Modrou barvou je vykreslena regresní přímka.

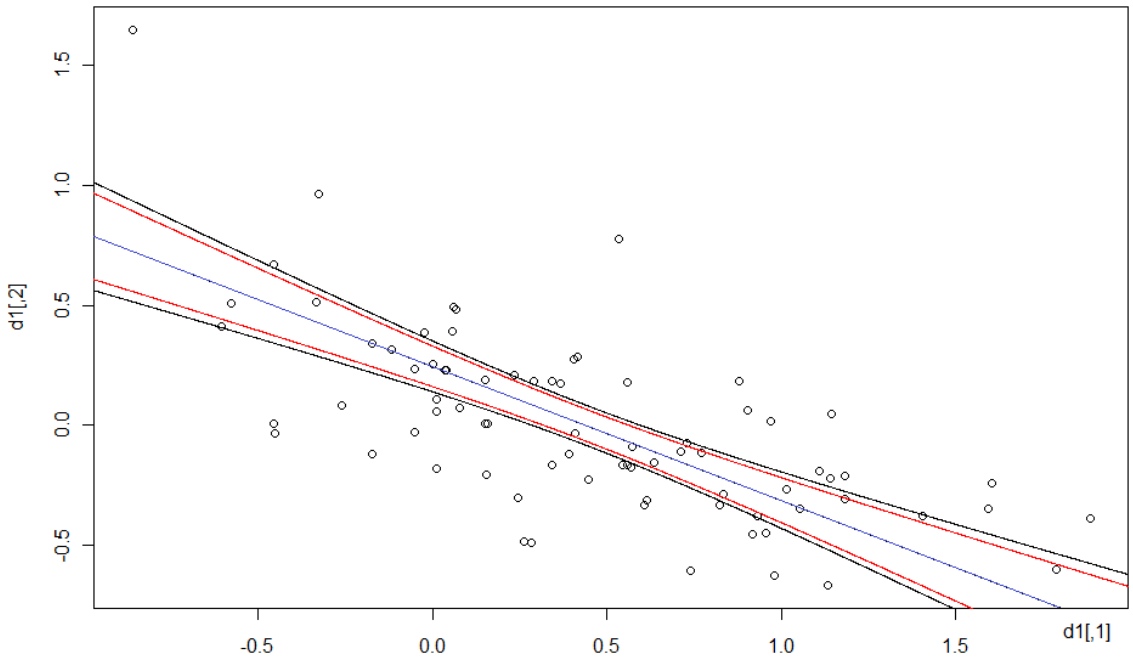
Kritérium konvergence bude stanoveno opět jako $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}||^2 < 10^{-9}$. Do okamžiku než bude kritérium splněno, proběhne celkem 21 iterace. Ve srovnání s předchozím příkladem se jedná se o velmi podobné číslo. Pomocí stejného postupu jako u Příkladu 1 vypočítáme odhad parametru $\boldsymbol{\beta}$,

$$(\hat{\beta}_1, \hat{\beta}_2)' = (0{,}2449; -0{,}5577)'.$$

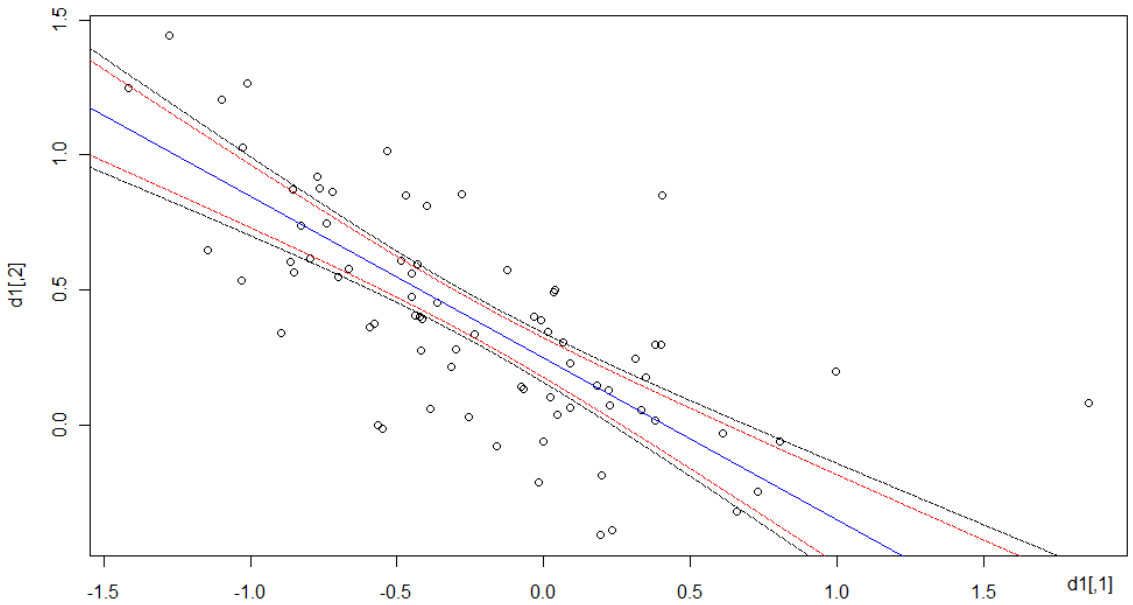
Pro úplnost je třeba vypočítat také varianční matici odhadu $\widehat{\boldsymbol{\beta}}$,

$$\widehat{\text{var}} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 0{,}0018 & -0{,}0016 \\ -0{,}0016 & 0{,}0035 \end{pmatrix}.$$

Následující graf zobrazuje transformovaná data poslechovosti, regresní přímka je znázorněna modře, pás spolehlivosti pro přímku je černý a červenou barvou jsme použili pro zdůraznění pásu spolehlivosti kolem přímky.

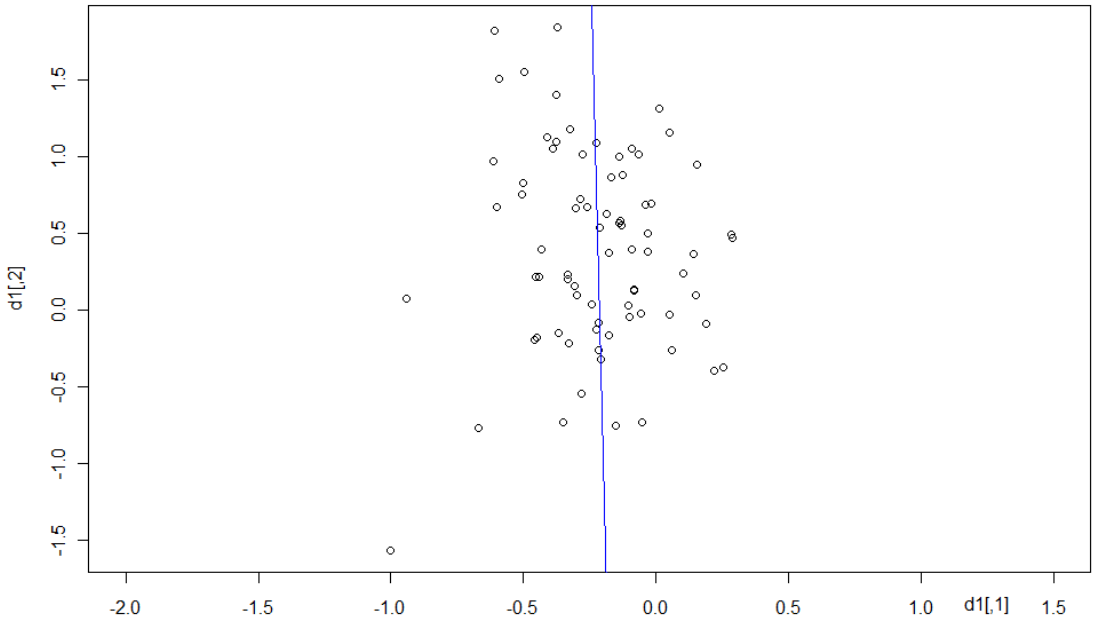


Oba pásy spolehlivosti téměř splývají (v případě první souřadnice zejména kolem hodnoty 0,5), maximální vzdálenost jejich větví je u našich pozorování jen o něco menší než 0,5, tedy pozorujeme, že se oba pásy nachází velmi těsně u regresní přímky pro transformovaná kompoziční data. I když se několik pozorování nachází ve větší vzdálenosti od regresní přímky, a lze je tudíž považovat za odlehlé hodnoty, jejich vliv zřejmě nebude tak velký jako u prvního příkladu. Podívejme se, jak budou grafy vypadat, budeme-li volit různé souřadnice.



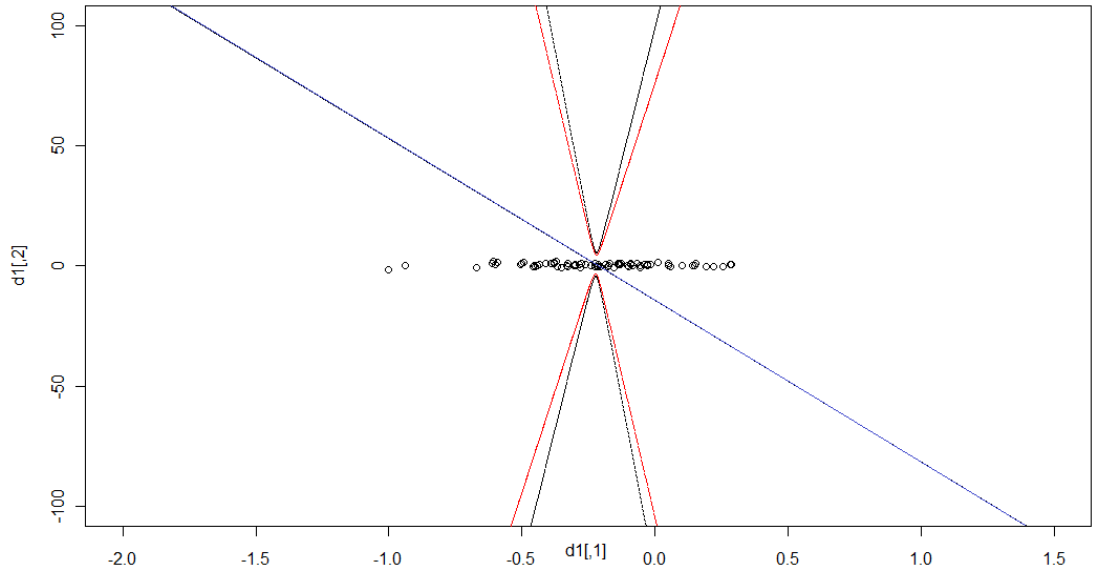
Druhá volba souřadnic (odpovídá (22) v Příkladu 1) jasně ukazuje, že pásy

spolehlivosti jsou velmi těsně okolo regresní přímky a jsou symetrické. Oba výsledky tak ukazují na vhodnost datového souboru a jeho normální rozdělení, které požadujeme. Problém ale nastane u třetí volby souřadnic (23).



Vidíme, že data (resp. zejména jim odpovídající regresní přímka) jsou prakticky rovnoběžná s vertikální osou, což bude zřejmě zdrojem velké numerické nestability iteračního algoritmu.

Pásy spolehlivosti se tak nachází ve značné vzdálenosti od regresní přímky, jak ukazuje následující graf. Taktéž sledujeme, že regresní přímka navíc ani dostatečně neodpovídá hodnotám datového souboru.



Tabulka ukazuje odhady parametrů regresních přímek a hodnoty odpovídajících statistik pro data týdenní poslechovosti českých rádií napříč všemi okresy České republiky. Souřadnice volíme obdobně jako u prvního příkladu, odpovídají vztahům (21), (22) a (23), statistiky T_1, T_2 a F jsou stejné jako v prvním příkladu.

	regresní přímka směr. odchykla $(\hat{\beta}_1, \hat{\beta}_2)$	iterace	T_1 p-hodnota	T_2 p-hodnota	F p-hodnota
(21)	$\nu = +0{,}2449 - 0{,}5577\mu$ $(0{,}0425; 0{,}0594)$	21	5,7568 $\ll 0{,}001$	$-9{,}3944$ $\ll 0{,}001$	44,1288 $\ll 0{,}001$
(22)	$\nu = +0{,}2491 - 0{,}5974\mu$ $(0{,}0367; 0{,}0614)$	21	6,7915 $\ll 0{,}001$	$-9{,}7230$ $\ll 0{,}001$	110,4527 0
(23)	$\nu = -14{,}3703 - 67{,}1847\mu$ $(44{,}9392; 204{,}4315)$	54	$-0{,}3198$ 0,7500	$-0{,}3286$ 0,7433	0,0732 0,9295

Na začátku jsme si stanovili hladinu významnosti jako obvyklou hodnotu $\alpha = 0{,}05$. Jelikož p-hodnoty jsou u první volby souřadnic značně nižší, jsou oba regresní parametry významné. Zamítáme tak nulovou hypotézu o nulovosti prvního ($H_0 : \beta_1 = 0$), resp. druhého regresního koeficientu ($H_0 : \beta_2 = 0$) ve prospěch alternativy $H_a : \beta_1 \neq 0$, resp. $H_a : \beta_2 \neq 0$. Oba parametry jsou tak statisticky významné. Situace je obdobná v případě, že zvolíme souřadnice pomocí (22). Například, testujeme-li celý model (pomocí statistiky F), opět na hladině významnosti 0,05 zamítáme nulovou hypotézu ve prospěch alternativy. Oba regresní koeficienty jsou statisticky významné.

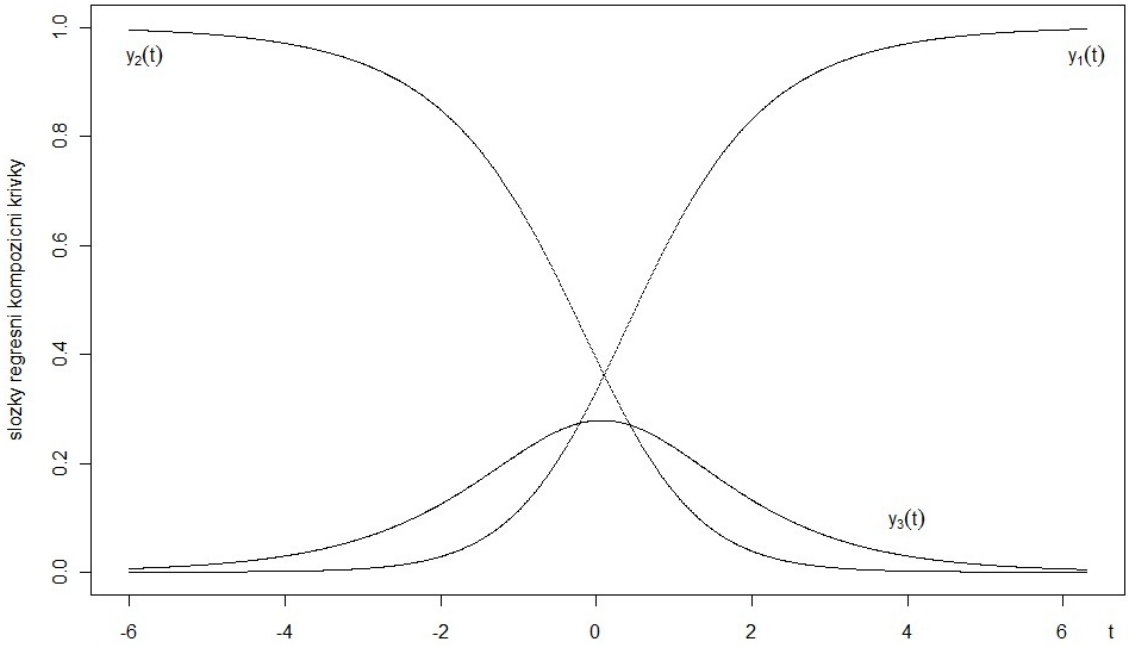
V případě volby třetích souřadnic ovšem vidíme, že ani první, ani druhý parametr nejsou statisticky významné, kvůli vysoké p-hodnotě nemůžeme zamítnout nulovou hypotézu. Tato volba souřadnic (potažmo regresního modelu) je tak evidentně pro naše data poslechovosti napříč okresy zcela nevhodná.

Následující tabulka přehledně srovnává intervaly spolehlivosti pro β_2 počítané různými přístupy. Jak u první, tak druhé volby souřadnic vychází hodnoty velmi podobně a délky intervalů se liší jen minimálně, kvalitativně nejlepší je přitom zřejmě interval (19). U třetích souřadnic ovšem intervaly spolehlivosti z důvodu numerické nestability nemůžeme spočítat, jelikož hodnota `n*(c[2]+1)` v případě prvního intervalu (19) vychází záporně. Též druhý a třetí interval spolehlivosti nelze spočítat, jelikož levá hranic intervalu spolehlivosti vychází vždy vyšší než pravá.

1. SOUŘADNICE	intervaly spolehlivosti	délka intervalu
(19)	$(-0{,}6264; -0{,}4890)$	0,1374
(15)	$(-0{,}7081; -0{,}4243)$	0,2838
(16)	$(-0{,}7035; -0{,}4279)$	0,2756

2. SOUŘADNICE	intervaly spolehlivosti	délka intervalu
(19)	$(-0{,}6640; -0{,}5307)$	0,1333
(15)	$(-0{,}7538; -0{,}4599)$	0,2939
(16)	$(-0{,}7490; -0{,}4636)$	0,2854

Provedeme vizualizaci složek kompoziční regresní přímky pro náš datový soubor. Postupujeme stejně jako u Příkladu 1.



Jelikož se jedná o kompoziční data, v každém bodě t je součet hodnot všech složek funkcí $y_1(t), y_2(t)$ a $y_3(t)$ roven právě číslu 1, což je z předchozího grafu patrné. Sledujeme měnící se podíly jednotlivých skupin rozhlasových stanic na tuzemském trhu, jelikož nás nezajímají absolutní hodnoty. Křivka funkce $y_1(t)$ odpovídá regionálním rádiím, $y_2(t)$ rozhlasovým sítím a nakonec $y_3(t)$ rádiím celoplošným.

Z grafu vyplývá, že minimálně ovlivňují dynamiku podílů poslechovosti celoplošná rádia. Jedná se z tohoto pohledu spíše o doplňkovou proměnnou. Jejich podíl je navíc napříč okresy České republiky nejstabilnější. V okamžiku, kdy začínají narůstat celoplošná rádia, rozdíl mezi regionálními a síťovými stanicemi se stírá, až do chvíle, kdy je prakticky nulový (stane se tak v případě maximálního dosahu celoplošných rádií).

Jakmile začne klesat v okresech podíl síťových rádií, kompenzuje jej nárůst jak celoplošných, tak i regionálních rádií, přičemž nárůst celoplošných rádií se děje na úkor rádií s menším pokrytím signálu.

Pro kladné hodnoty parametru t sledujeme, jak se mění podíly mezi třemi typy rádií ve prospěch regionálních rádií. Strměji se přitom snižují podíly v neprospěch síťových rádií, které někteří sice někteří posluchači vnímají jako regionální (jelikož

přináší informace právě z jejich bydliště ve větší míře), přesto dávají přednost čistě regionálním, jež k nim přeci jenom v jistém slova smyslu mají o něco blíže, a to právě proto, že sídlí v jejich blízkosti.

Poznamenejme, že uvedená regresní přímka neříká, jakým směrem se dynamika poslechovosti vyvíjí v čase (zda například od regionálních rádií k sítím či naopak). Za tímto účelem je navíc třeba sledovat časové řady poslechovosti v jednotlivých regionech.

Trendem poslední doby je slučování menších regionálních rádií do jakýchsi sítí. Tento trend již není tak silný, jako tomu bylo před rokem, nicméně nelze očekávat, že by velké společnosti přestaly usilovat o odkoupení těchto regionálních subjektů.

Jaký vývoj lze očekávat? Je velmi pravděpodobné, že posluchači budou nadále poslouchat svá rádia a to přesto, že budou vysílat v éteru pod jinou značkou. Tudíž v návaznosti na tento fakt vzroste vliv rozhlasových sítí. Z grafu sledujeme, že prudce začne klesat podíl regionálních rádií (opět logické, jelikož bude menší počet těchto rádií, bude je poslouchat i méně lidí). Na poslechovosti celoplošných rádií se nic výrazněji nemění, jen lehce posílí.

V praxi však můžeme často narazit na nečekané potíže, nastane situace, která je v rozporu s tímto modelem. Posluchači odejdou k jiné, nezávislé regionální rozhlasové stanici, která se dosud nepřiřadila k žádné rozhlasové síti. Praktickým příkladem je konec značky *Rádio Hity*, v roce 2005 se stanice transformovala pod značku *Fajn radio Hity*. Posluchači tehdy hojně odcházeli k jiným, regionálním stanicím. Formát Fajn radia Hity jim nevyhovoval. I přes větší počet rádií v rozhlasové síti se počet posluchačů snížil. Většina posluchačů nicméně neposlouchá rádio cíleně a využívá jej jako kulisu k jiným činnostem. Proto přechod pod jinou značku někteří patrně ani nepostřehnou.

Závěr

Po hlubším studiu ortogonální regrese, ke kterému mě diplomová práce motivovala, musím potvrdit prvotní úvahu, totiž že se tato regresní metoda opravdu jeví jako vhodná pro regresní analýzu složek kompozičních dat. Diplomová práce přitom demonstrovala, že lze v tomto ohledu s výhodou použít lineární regresní model s podmínkou II. typu.

Přesto, že mohou interpretaci výsledků při práci s reálnými datovými soubory poněkud pokazit vyskytující se odlehlé hodnoty, ve většině případů zůstává ortogonální regrese nejvýhodnějším nástrojem pro řešení dané problematiky.

Po prvních zkušenostech s kompozičními daty jsem se rovněž v diplomové práci rozhodl v tomto zajímavém oboru matematiky pokračovat. Volba se nakonec ukázala jako správná, i přes horké chvíle se domnívám, že téma mi může pomoci při hledání uplatnění v praxi.

Ačkoli teoretická část diplomové práce byla poměrně obtížná, díky nabytým vědomostem o lineárních modelech a zkušenostech z bakalářské práci, v níž jsem rovněž pracoval s kompozičními daty, bylo možné dosáhnout cíle a problematiku zpracovat.

Mnohem náročnější už byla praktická část, nejtěžší byla nakonec samotná interpretace grafů zobrazujících průběh složek kompoziční regresní přímky. I po bližším seznámení s problematikou a zkoumáním dat vyvstalo mnoho otázek. Přesto doufám, že se mi na ně podařilo zdárně odpovědět.

Reference

[1] Aitchison, J., *The Statistical Analysis Of Compositional Data*, London: Chapman & Hall, 1986.

[2] Anděl, J., *Statistické metody*, 2. přepracované vydání, Praha: MATFYZPRESS, 1998.

[3] van der Boogaart, K. G., *Using the R package "compositions"* [online], dostupné z http://www.stat.boogaart.de/compositions/UsingCompositions.pdf, [citováno 11.10.2011].

[4] Filzmozer P., Hron K., Reimann C., *Principal Component Analysis for Compositional Data with Outliers*, Environmetrics **20** (2009), 621-632.

[5] Fišerová, E., Hron K., *Total least squares solution for compositional data using linear models*, Journal of Applied Statistics **37** (2010), No. 7, 1137-1152.

[6] Fišerová, E., Hron K., *Statistical Inference in Orthogonal Regresion for Three-part Compositional Data Using Linear Model with Type-II Constraints*, Communications In Statistics-Theory and Methods, přijato k tisku.

[7] *Food and Agriculture Organization of the United Nations* [online], dostupné z http://faostat.fao.org/site/377/DesktopDefault.aspx?PageID=377#ancor, [citováno dne 30. 12. 2011].

[8] *Graphic Manual R* [online], dostupné z http://www.oga-lab.net/RGM2/func.php?rd_id=robCompositions:robCompositions-package, [citováno dne 18. 12. 2011].

[9] Harville, D. A., *Matrix Algebra from a Statistician's Perspective*, 1. vydání, New York: Springer, 1997.

[10] Kubáček L., Kubáčková L., *Statistika a metrologie*, 1. vydání, Olomouc: Univerzita Palackého, 2000.

[11] Markovsky I., van Huffel S., *Overview of Total Least-Squares Methods*, Signal Processing **87** (2007), No. 10, 2283-2302.

[12] Martin-Fernánder, J.A., *Model-Based Replacement of Rounded Zeros in Compositional Data: Classical and Robust Approaches*, Computational Statistics and Data Analysis (2012), přijato k tisku.

[13] Pawlowsky-Glahn, V., Egozcue J.J., *Groups of parts and their balances in compositional data analysis*, Mathematical Geology **37** (2005), No. 7, 796-801.

[14] Pawlowsky-Glahn, V., Egozcue, J.J., Tolasana-Delgado R., *Lecture Notes on Compositional Data Analysis*, The University of Girona, 2007 [online], dostupné z http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf [citováno 5. 4. 2009].

[15] Petera, M., *Bakalářská práce: Korelační analýza pro kompoziční data*, 1. vydání, Olomouc: Univerzita Palackého, 2010.

[16] *Principal Component Analysis* [online], dostupné z http://classifion.sicyon.com/References/princomp.pdf [citováno 13. 2. 2012].

[17] *Singular Value Decomposition Tutorial* [online], dostupné z http://www.cs.wits.ac.za/ michael/SVDTut.pdf [citováno 17. 11. 2011].

[18] *Tabulka poslechovosti v okresech* [online], dostupné z http://radiotv.cz/radio/poslechovost, [citováno dne 9. 9. 2011].

[19] *The R Project for Statistical Computing* [online], dostupné z http://www.r-project.org/ [citováno 18. 12. 2011].

[20] van Huffel S., Vandewalle J., *The Total Least Squares Problem: Computational Aspects and Analysis*, Flanders: Society for Industrial Mathematics, 1987.

[21] *Zákon ze dne 17. května 2001 o provozování rozhlasového a televizního vysílání a o změně dalších zákonů* [online], dostupné z http://www.rrtv.cz/cz/static/cim-se-ridime/stavajici-pravni-predpisy/pdf/231-2001.pdf [citováno 17. 11. 2011].

Příloha

A. Zápis iterativního algoritmu dle Věty 4.3

```
> iter=0

> while(sqrt(datac[1]^2+datac[2]^2)>=10^(-9)){

> P=cbind(rep(1,n),mu0)%*%
> solve(cbind(c(n,t(mu0)%*%rep(1,n)),c(t(mu0)%*%rep(1,n),
t(mu0)%*%mu0)))%*%
> t(cbind(rep(1,n),mu0))
> M=diag(rep(1,n))-P

> mu=datai[,1]+((c0[2])/(c0[2]^2+1))*M%*%(datai[,2]-nu0-c0[2]
*(datai[,1]-mu0))
> nu=datai[,2]-(1/(c[2]^2+1))*M%*%(datai[,2]-nu0-c0[2]*
(datai[,1]-mu0))

> c=c0+solve(cbind(c(n,t(mu0)%*%rep(1,n)),c(t(mu0)%*%rep(1,n),
t(mu0)%*%mu0)))%*%
> c(t(rep(1,n))%*%(d1[,2]-nu0-c0[2]*(datai[,1]-mu0)),
  t(mu0)%*%(datai[,2]-nu0-c0[2]*(datai[,1]-mu0)))

> datac=c-c0

> nu0=nu+(c[2]-c0[2])*(mu-mu0)
> mu0=mu
> c0=c

> iter=iter+1
}
> iter #počet iterací, které proběhnou do splnění kritéria
```

B. Datový soubor k Příkladu 1

Stát	Lesní plocha	Orná půda	Trvalé plodiny
Rakousko	38820	13710	660
Belgie	6768	8400	220
Bulharsko	38718	31390	1720
Kypr	1731	870	339
Česká republika	26550	31800	760
Dánsko	5420	24310	60
Estonsko	22240	5960	80
Finsko	221570	22570	50
Francie	159060	183456	10504
Německo	110760	119450	2000
Řecko	38728	25508	11484
Maďarsko	20198	45850	1940
Irsko	7302	10890	30
Itálie	90710	68800	26050
Lotyšsko	33426	11680	60
Litva	21522	20537	277
Lucembursko	868	620	20
Malta	3	80	13
Nizozemí	3650	10547	355
Polsko	93096	125390	4000
Portugalsko	34522	11250	7780
Rumunsko	65366	87890	3620
Slovensko	19328	13820	240
Slovinsko	12510	1750	260
Španělsko	179973	124970	47190
Švédsko	282030	26340	90
Spojené království	28738	60490	430

C. Datový soubor k Příkladu 2

Okres	Regionální rádia	Rozhlasové sítě	Celoplošná rádia
Praha	666	625	729
Benešov u Prahy	6	82	35
Beroun	17	59	37
Kladno	95	44	91
Kolín	19	53	45
Kutná Hora	21	51	41
Mělník	28	38	57
Mladá Boleslav	105	62	59
Nymburk	33	58	49
Praha - východ	66	76	59
Praha - západ	45	63	47
Příbram	26	99	42
Rakovník	13	18	23
České Budějovice	43	190	101
Český Krumlov	11	52	38
Jindřichův Hradec	19	55	60
Pelhřimov	7	51	30
Písek	20	43	38
Prachatice	10	41	24
Strakonice	21	56	35
Tábor	52	101	51
Domažlice	17	45	29
Cheb	48	33	37
Karlovy Vary	39	87	69
Klatovy	16	67	46
Plzeň-město	37	191	79
Plzeň - jih	13	55	29
Plzeň - sever	14	90	35
Rokycany	17	34	17
Sokolov	36	35	38
Tachov	25	24	23
Česká Lípa	44	46	51
Děčín	25	81	83
Chomutov	26	121	74
Jablonec nad Nisou	30	25	36
Liberec	88	78	108
Litoměřice	26	67	87
Louny	15	51	34

Okres	Regionální rádia	Rozhlasové sítě	Celoplošná rádia
Most	12	108	66
Teplice	9	119	69
Ústí nad Labem	11	93	66
Havlíčkův Brod	14	74	48
Hradec Králové	67	112	73
Chrudim	15	59	63
Jičín	19	49	42
Náchod	36	82	64
Pardubice	47	121	109
Rychnov nad Kněžnou	31	53	42
Semily	32	40	52
Svitavy	31	42	62
Trutnov	64	51	71
Ústí nad Orlicí	45	60	78
Blansko	53	41	81
Brno-město	283	249	210
Brno - venkov	110	76	119
Břeclav	73	26	67
Hodonín	135	48	86
Jihlava	40	81	59
Kroměříž	65	22	86
Prostějov	83	9	93
Třebíč	20	89	68
Uherské Hradiště	99	34	70
Vyškov	46	34	59
Zlín	95	72	145
Znojmo	15	74	55
Žďár nad Sázavou	42	92	73
Bruntál	24	42	54
Frýdek - Místek	75	127	121
Jeseník	9	10	30
Karviná	109	132	133
Nový Jičín	77	92	93
Olomouc	156	89	90
Opava	95	100	131
Ostrava	143	164	221
Přerov	74	47	76
Šumperk	58	47	82
Vsetín	73	65	90