evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS OSTRAVIENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

ANALÝZA VÍCEROZMĚRNÝCH DAT

URČENO PRO VZDĚLÁVÁNÍ V AKREDITOVANÝCH STUDIJNÍCH PROGRAMECH

JOSEF TVRDÍK

ČÍSLO OPERAČNÍHO PROGRAMU: CZ.1.07
NÁZEV OPERAČNÍHO PROGRAMU:
VZDĚLÁVÁNÍ PRO KONKURENCESCHOPNOST
OPATŘENÍ: 7.2
ČÍSLO OBLASTI PODPORY: 7.2.2

INOVACE VÝUKY INFORMATICKÝCH PŘEDMĚTŮ VE STUDIJNÍCH PROGRAMECH OSTRAVSKÉ UNIVERZITY

REGISTRAČNÍ ČÍSLO PROJEKTU: CZ.1.07/2.2.00/28.0245

OSTRAVA 2013

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky

Recenzent: Prof. RNDr. Ing. Ivan Křivý, CSc.

Název:	Analýza vícerozměrných dat
Autor:	Josef Tvrdík
Vydání:	třetí, 2013
Počet stran:	128

Jazyková korektura nebyla provedena, za jazykovou stránku odpovídá autor.

Obsah

1 Úvod

Tento text je určen jako studijní opora předmětu Analýza vícerozměrných dat ve všech formách studiana Přírodovědecké fakultě Ostravské university ve studijním oboru Informační systémy.

Cílem textu je poskytnout nezbytné základy pro statistickou analýzu vícerozměrných dat, zejména pro regresní analýzu a vybrané mnohorozměrné metody. Výklad je zaměřen spíše na porozumění základním pojmům, které je nutné pro správnou aplikaci metod vícerozměrné analýzy, než na matematické důkazy. Přesto předkládaný text není lehké čtení do postele před spaním. Prosím, počítejte s tím, že budete často nuceni usilovně přemýšlet, vykládanou látku si postupně vyjasňovat a k mnoha tématům se opakovaně vracet. Někdy vám může pomoci i studium citovaných učebnic a publikací nebo zdrojů z Internetu. Snad k pochopení pomohou i obsáhlé řešené příklady, které jsou připojeny k většině důležitých témat. Data pro řešené příklady jsou přiložena k učebnímu textu v samostatných souborech ve formátu tabulkového procesoru Excel, aby byla čitelná na většině běžných počítačů s různým softwarovým vybavením.

Každá kapitola začíná pokyny pro její studium. Tato část je vždy označena jako **Průvodce studiem** s ikonou na okraji stránky.

Pojmy a důležité souvislosti k zapamatování jsou vyznačeny na okraji stránky textu ikonou.

V závěru každé kapitoly je rekapitulace nejdůležitějších pojmů. Tato rekapitulace je označena textem **Shrnutí** a ikonou na okraji.

Oddíl **Kontrolní otázky** označený ikonou by vám měl pomoci zjistit, zda jste prostudovanou kapitolu pochopili a snad vyprovokuje i vaše další otázky, na které budete hledat odpověď.

U některých kapitol je připomenuta **Korespondeční úloha**. Pro kombinované studium budou korespondenční úlohy zadávány v rámci kurzu daného semestru. Úspěšné vyřešení korespondenčních úloh je součástí podmínek pro ukončení předmětu v kombinovaném studiu.

2 Vektory a matice

Průvodce studiem

Kapitola je opakováním látky o vektorech a maticích, což byste měli už znát z kurzů lineární algebry. Na tuto kapitolu počítejte se dvěma hodinami studia, pokud jste toho moc z lineární algebry nezapomněli. Jinak bude potřebný čas delší. Všechny uvedené operace s maticemi jsou pak užívány v dalším textu, tak je nutné, abyste je bezpečně zvládli.

2.1 Základní pojmy

Vektory bude označovat malými tučnými písmeny. Sloupcový vektor – příklady:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

Řádkový vektor dostaneme transpozicí sloupcového vektoru, např.

$$\mathbf{x}^T = [x_1, \cdots, x_p]$$

Skalární součin vektorů je

$$\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} = \sum_{i=1}^{p} x_i y_i$$

Specilelně $\mathbf{x}^T\mathbf{x} = \sum_{i=1}^{p} x_i^2$

Norma vektoru je

$$\| \mathbf{x} \| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^{p} x_i^2}$$

Kosinus směrového úhlu dvou vektorů je pak

$$\cos\alpha = \frac{\mathbf{x}^T\mathbf{y}}{\| \mathbf{x} \| \| \mathbf{y} \|}$$

Je-li $\cos\alpha = 0$, tj. $\mathbf{x}^T\mathbf{y} = 0$, pak říkame, že vektory jsou *ortogonální* (jsou na sebe kolmé). Vidíme, že kosinus směrového úhlu vektorů je vlastně výběrový korelační koeficient veličin $\mathbf{x}$, $\mathbf{y}$, tedy jsou-li vektory $\mathbf{x}$, $\mathbf{y}$ ortogonální, znamená to, že veličiny $\mathbf{x}$, $\mathbf{y}$ jsou nekorelované.

Matice typu $(n \times p)$ (matice budeme označovat velkými tučnými písmeny) je

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Matice $\mathbf{A}$, $\mathbf{B}$ lze sčítat (a odčítat), pokud jsou stejného typu $(n \times p)$.

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

Matice $\mathbf{C}$ je opět typu $(n \times p)$ a pro její prvky platí

$$c_{ij} = a_{ij} + b_{ij}$$

Matice $\mathbf{A}$, $\mathbf{B}$ lze násobit, pokud jsou typu $(n \times p)$ a $(p \times m)$.

$$\mathbf{AB} = \mathbf{C}$$

Matice $\mathbf{C}$ je typu $(n \times m)$ a pro její prvky platí

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

Jelikož vektor je specielní případ matice mající jen jeden sloupec nebo řádek, lze stejné pravidlo užít i pro násobení matice vektorem. Soustavu lineárních rovnic můžeme pak stručně zapsat jako

$$\mathbf{Ay} = \mathbf{b}$$

Transponovaná matice $\mathbf{X}^T$ vznikne z matice $\mathbf{X}$ tak, že zaměníme řádky a sloupce, tzn.

$$\mathbf{X}^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix}$$

Transponovaná matice $\mathbf{X}^T$ je typu $(p \times n)$. Je zřejmé, že platí $\left(\mathbf{X}^T\right)^T = \mathbf{X}$.

Pro transponování platí následující pravidla:

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

Hodnost matice typu $(m \times n)$ je přirozené číslo $h(\mathbf{C}) \leq \min(m, n)$.

Je-li matice typu $(n \times n)$, říkáme, že je to *čtvercová matice* řádu n.

Symetrická matice je čtvercová matice, pro kterou platí $\mathbf{A}^T = \mathbf{A}$, tzn. je symetrická podle hlavní diagonály.

Diagonální matice je čtvercová matice, která má všechny prvky mimo hlavní diagonálu rovny nule.

Jednotková matice je diagonální matice s jedničkami na hlavní diagonále. Označujeme ji $\mathbf{I}$ nebo $\mathbf{I}_n$, je-li nutno zmínit její rozměr.

Stopa matice je součet diagonálních prvků $Tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

Determinant matice označujeme $|\mathbf{A}|$ nebo $\det(\mathbf{A})$. Je to skalár (číselná hodnota), kterou můžeme chápat jako míru nevyváženosti matice.

Když $\mathbf{A}$ je typu 2×2, pak $|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$.

Matice $\mathbf{A}$ řádu n je regulární, když $|\mathbf{A}| \neq 0$. Pak existuje inverzní matice $\mathbf{A}^{-1}$, pro kterou platí

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Když matice $\mathbf{A}$ řádu n je regulární ($|\mathbf{A}| \neq 0$), pak hodnost matice je $h(\mathbf{A}) = n$.

Dále platí, že

$$(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Je-li matice $\mathbf{A}$ řádu 2 regulární, pak inverzní matice $\mathbf{A}^{-1}$ je

$$\mathbf{A}^{-1} = \begin{bmatrix} a_{22}/\Delta & -a_{12}/\Delta \\ -a_{21}/\Delta & a_{11}/\Delta \end{bmatrix},$$

kde $\Delta = a_{11}a_{22} - a_{12}a_{21}$, tj. determinant matice $\mathbf{A}$.

Kvadratická forma matice je skalár

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

Kvadratická forma je určena maticí $\mathbf{A}$. Matice $\mathbf{B}$, pro kterou platí $b_{ii} = a_{ii}$ a současně $b_{ij} + b_{ji} = a_{ij} + a_{ji}$ určuje tutéž kvadratickou formu. Existuje však *jediná symetrická* matice dané kvadratické formy.

2.2 Vlastní čísla a vlastní vektory matice

Nechť $\mathbf{A}$ je čtvercová matice řádu n. Pak vlastní číslo (charakteristické číslo, eigenvalue) je takový skalár λ, aby pro nenulový vektor $\mathbf{u}$ platilo:

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$$

$\mathbf{u}$ je vlastní (charakteristický) vektor. Vidíme, že výše uvedenou rovností není definován jednoznačně, neboť rovnost platí pro každý vektor $c\mathbf{u}$, $c \neq 0$. Nadále budeme tedy uvažovat jen vektory normované (s normou rovnou jedné), tzn. $\mathbf{v} = c\mathbf{u}$, kde $c = 1/ \parallel \mathbf{u} \parallel$. Rovnost pak můžeme přepsat na tvar

$$(\mathbf{A} - \mathbf{I}\lambda)\mathbf{v} = \mathbf{0}$$

Protože $\mathbf{v} \neq \mathbf{0}$, musí platit, že

$$|\mathbf{A} - \mathbf{I}\lambda| = 0$$

Tento determinant je polynom n-tého stupně, řešení je $\lambda_1, \lambda_2, \cdots, \lambda_n$ a každému vlastnímu číslu odpovídá vlastní vektor $\mathbf{v}_i$.

Když $\mathbf{A}$ je *symetrická matice*, pak vlastní čísla jsou *reálná* a vlastní vektory jsou *ortogonální*, takže platí

$$\begin{aligned} \mathbf{v}_i^T\mathbf{v}_i &= 1 \qquad & i &= 1, 2, \cdots, n \\ \mathbf{v}_i^T\mathbf{v}_i &= 0 \qquad & i &\neq j \end{aligned}$$

Vlastní vektory uspořádáme do matice $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$, pak $\mathbf{V}$ je ortogonální matice, tj. platí

$$\mathbf{V}^T = \mathbf{V}^{-1} \qquad \text{a} \qquad \mathbf{V}^T\mathbf{V} = \mathbf{I}.$$

Matici $\mathbf{A}$ můžeme diagonalizovat:

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{L} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Pak stopa matice $\mathbf{A}$ je rovna součtu jejích vlastních čísel, $Tr(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$ a determinant matice $\mathbf{A}$ je roven součinu jejích vlastních čísel, $|\mathbf{A}| = \lambda_1\lambda_2 \cdots \lambda_n$.

Spektrální rozklad matice $\mathbf{A}$ je definován jako

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

2.3 Další důležité vlastnosti matic

Jestliže $\mathbf{C}$ je ortogonální matice a $\mathbf{y} = \mathbf{C}\mathbf{x}$, pak $\mathbf{y}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}$

Symetrická matice $\mathbf{A}$ je *pozitivně definitní*, jestliže kvadratická forma $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ pro každý vektor $\mathbf{x} \neq \mathbf{0}$. Když $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$, pak $\mathbf{A}$ je *pozitivně semidefinitní*. Pozitivně definitní matice má všechna vlastní čísla kladná.

Když $\mathbf{B}$ je matice typu $(n \times m)$, s hodností m, pak $\mathbf{B}^T\mathbf{B}$ je pozitivně definitní.

Jestliže $\mathbf{A}$ je pozitivně definitní, pak existuje regulární matice $\mathbf{P}$ taková, že

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{I} \qquad \text{a} \qquad \mathbf{P}^T\mathbf{P} = \mathbf{A}^{-1}$$

Pseudoinverzní matice $\mathbf{A}^-$: Když $\mathbf{A}$ je matice typu $(m \times n)$, pak $\mathbf{A}^-$ je typu $(n \times m)$ a platí

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$$

$\mathbf{A}^-$ vždy existuje, ale není jednoznačně určena.

2.4 Derivace skalárního výrazu podle vektoru

Je-li $y = f(x_1, x_2, \ldots, x_n)$, potom

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_n \end{bmatrix}$$

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \partial \mathbf{a}^T\mathbf{x}/\partial x_1 \\ \partial \mathbf{a}^T\mathbf{x}/\partial x_2 \\ \vdots \\ \partial \mathbf{a}^T\mathbf{x}/\partial x_n \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}$$

Vidíme, že

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}.$$

Shrnutí

- *vektor, matice, transponování vektorů a matic*
- *determinant matice, hodnost matice, inverzní matice, jednotková matice, symetrická matice, diagonální matice, stopa matice,*
- *kvadratická forma, pozitivně definitní matice*
- *vlastní čísla a vlastní vektory matice*
- *derivace funkce podle vektoru*

Kontrolní otázky

1. *Nechť* $\mathbf{x}^T = [x_1, \cdots, x_n]$, $\mathbf{1}$ *je vektor* $n \times 1$, *jehož prvky jsou rovny 1. Čemu jsou rovny výrazy* $\mathbf{1}^T\mathbf{x}$, $\mathbf{1}\mathbf{x}^T$*? Jsou si tyto výrazy rovny?*
2. *Nechť* $\mathbf{x} = [x_1, x_2, , x_3]^T$, $\mathbf{a} = [1, 2, 3]^T$, $\mathbf{B} = \mathbf{a}\mathbf{x}^T$. *Spočítejte determinant matice* $\mathbf{B}$.
3. *Nechť* $\mathbf{A}$ *je čtvercová matice řádu* n, $\mathbf{y}$ *je vektor* $n \times 1$. *Čemu je rovna derivace* $\mathbf{y}^T\mathbf{A}\mathbf{y}$ *podle vektoru* $\mathbf{y}$*?*

3 Náhodný vektor a mnohorozměrné rozdělení

Průvodce studiem

Na tuto kapitolu počítejte nejméně se třemi hodinami studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí. Zejména se zaměřte na pochopení rozdílů mezi sdruženým, marginálním a podmíněným rozdělením.

Náhodný vektor $\mathbf{X} = [X_1, X_2, \cdots, X_p]^T$ je vektor, jehož složky
$X_1, X_2, \cdots, X_p$ jsou náhodné veličiny.

U náhodného vektoru musíme rozlišovat rozdělení *sdružené, marginální a podmíněné.*

3.1 Sdružené rozdělení

Pro dvousložkový náhodný vektor ($p = 2$) je sdružená distribuční funkce definována

$$F(x_1, x_2) = P(X_1 < x_1, X_2 < x_2)$$

Jsou-li X_1, X_2 *diskrétní* veličiny, pak sdružená pravděpodobnostní funkce je

$$P(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

Existuje-li nezáporná funkce $f(x_1, x_2)$ taková, že

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(u, v) \mathrm{d}u\mathrm{d}v,$$

pak náhodný vektor $[X_1, X_2]^T$ má rozdělení spojitého typu. Funkce $f(\cdot, \cdot)$ je sdružená hustota.

Pravděpodobnost, že náhodné veličiny X_1, X_2 nabývají hodnot z intervalů $[a_1, b_1)$, $[a_2, b_2)$ je určena vztahem

$$P(a_1 \leq X_1 < b_1, a_2 \leq X_2 < b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2$$

Sdružená hustota je

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$$

Pro $p > 2$ platí analogické vztahy, mimo jiné sdružená hustota je derivací distribuční funkce:

$$f(\mathbf{x}) = f(x_1, x_2, \cdots, x_p) = \frac{\partial^p F(x_1, x_2, \cdots, x_p)}{\partial x_1 \partial x_2 \cdots \partial x_p}$$

3.2 Marginální rozdělení

Když $F(x_1, x_2)$ je sdružená distribuční funkce, pak *marginální* distribuční funkce veličiny X_1, resp. X_2 jsou

$$F_1(x_1) = P(X_1 < x_1, X_2 < \infty) = F(x_1, \infty)$$

$$F_2(x_2) = P(X_1 < \infty, X_2 < x_2) = F(\infty, x_2)$$

Pro *diskrétní* rozdělení marginální pravděpodobnostní funkce jsou

$$P_1(x_1) = \sum_{M_2} P(x_1, x_2)$$

$$P_2(x_2) = \sum_{M_1} P(x_1, x_2)$$

kde M_i je množina hodnot diskrétní náhodné veličiny X_i.

Pro *spojité* rozdělení marginální hustoty jsou

$$f_1(x_1) = \int_{M_2} f(x_1, x_2)\mathrm{d}x_2$$

$$f_2(x_2) = \int_{M_1} f(x_1, x_2)\mathrm{d}x_1$$

kde M_i je obor hodnot spojité náhodné veličiny X_i.

Když $p > 2$, pak sdružené rozdělení každé neprázdné podmnožiny $\{X_1, X_2, \cdots, X_p\}$ je marginální rozdělení. Např.pro $p = 3$ jsou pak sdružená rozdělení náhodných vektorů

$$[X_1, X_2]^T, [X_1, X_3]^T, [X_2, X_3]^T, X_1, X_2, X_3$$

jsou marginální rozdělení (máme tedy 6 marginálních rozdělení).

3.3 Podmíněné rozdělení

Rozdělení, kdy jedna nebo více složek ($r < p$) náhodného vektoru je konstantní.

Pro $p = 2$ je podmíněná distribuční funkce definována jako

$$F(x_1 \mid x_2) = \lim_{\Delta x_2 \to 0} P(X_1 < x_1 \mid x_2 < X_2 \leq x_2 + \Delta x_2)$$

Pro diskrétní rozdělení je

$$F(x_1 \mid x_2) = \frac{\sum_{t<x_1} P(t, x_2)}{P_2(x_2)} \qquad P_2(x_2) \neq 0$$

a podmíněná pravděpodobnostní funkce je

$$P(x_1 \mid x_2) = \frac{P(x_1, x_2)}{P_2(x_2)} \qquad \text{pro } P_2(x_2) \neq 0$$

Pro spojitá rozdělení je podmíněná distribuční funkce

$$F(x_1 \mid x_2) = \frac{\int_{-\infty}^{x_1} f(t, x_2)\mathrm{d}t}{f_2(x_2)} \qquad \text{pro } f_2(x_2) \neq 0$$

a podmíněná hustota

$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \qquad \text{pro } f_2(x_2) \neq 0$$

3.4 Nezávislost veličin

Pro nezávislé veličiny platí

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)$$

$$P(x_1, x_2) = P_1(x_1)P_2(x_2)$$

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)$$

Jsou-li veličiny *nezávislé*, pak *podmíněná* rozdělení jsou rovna *marginálním*. Pro $p > 2$ platí podobné vztahy pro sdruženě (vzájemně) nezávislé veličiny.

3.5 Charakteristiky náhodného vektoru

Máme náhodný vektor o délce p, $\mathbf{X} = [X_1, X_2, \cdots, X_p]^T$.

Marginální charakteristiky vektoru diskrétních náhodných veličin jsou:

Střední hodnoty

$$E(X_j) = \sum_{M_j} x_j P_j(x_j), \qquad j = 1, 2, \ldots, p$$

Rozptyly

$$\mathrm{var}(X_j) = \sum_{M_j} \left[x_j - E(X_j)\right]^2 P_j(x_j), \qquad j = 1, 2, \ldots, p$$

Pro vektor spojitých náhodných veličin jsou střední hodnoty

$$E(X_j) = \int_{M_j} x_j f_j(x_j) \mathrm{d}x_j, \qquad j = 1, 2, \ldots, p$$

a rozptyly

$$\mathrm{var}(X_j) = \int_{M_j} \left[x_j - E(X_j)\right]^2 f_j(x_j) \mathrm{d}x_j, \qquad j = 1, 2, \ldots, p$$

Podmíněné charakteristiky vektoru diskrétních náhodných veličin pro $p = 2$ jsou definovány následovně.

Střední hodnoty

$$E(X_1 \mid x_2) = \sum_{M1} x_1 P_1(x_1 \mid x_2)$$

Rozptyly

$$\mathrm{var}(X_1 \mid x_2) = \sum_{M_1} \left[x_1 - E(X_1 \mid x_2)\right]^2 P(x_1 \mid x_2)$$

Podmíněné charakteristiky vektoru spojitých náhodných veličin pak jsou:

Střední hodnoty

$$E(X_1 \mid x_2) = \int_{M1} x_1 f(x_1 \mid x_2) \mathrm{d}x_1$$

Rozptyly

$$\mathrm{var}(X_1 \mid x_2) = \int_{M_1} \left[x_1 - E(X_1 \mid x_2)\right]^2 f(x_1 \mid x_2) \mathrm{d}x_1$$

Podmíněná střední hodnota $E(X_1 \mid x_2)$ se nazývá regresní funkce (závislost X1 na X2).

Podmíněný rozptyl $\mathrm{var}(X_1 \mid x_2)$ se nazývá skedastická funkce. Je-li podmíněný rozptyl $\mathrm{var}(X_1 \mid x_2)$ konstantní pro všechna x_2, pak o rozdělení říkáme, že je *homoskedastické*, není-li konstantní, mluvíme o *heteroskedastickém rozdělení* náhodného vektoru $[X_1, X_2]^T$.

3.6 Vícerozměrná rozdělení

Základními charakteristikami vícerozměrného rozdělení je vektor středních hodnot

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix}$$

a kovarianční (varianční) matice

$$\mathbf{\Sigma} = \mathrm{var}(\mathbf{X}) = \mathrm{cov}(\mathbf{X}) = E\left[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X})^T)\right] \tag{1}$$

což znamená, že

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix},$$

kde σ_{ij} je kovariance dvou náhodných veličin, tj.

$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = E\left[(X_i - EX_i)(X_j - EX_j)\right]$$

a $\sigma_{ii} = \sigma_i^2$ je rozptyl $\mathrm{var}(X_i)$. Vidíme, že kovarianční matice $\mathbf{\Sigma}$ je symetrická, nebot' $\sigma_{ij} = \sigma_{ji}$.

3.6.1 Vícerozměrné normální rozdělení

Náhodný vektor $\mathbf{X}$ má vícerozměrné normální rozdělení, jestliže jeho hustota je dána vztahem

$$f(\mathbf{x}) = (2\pi)^{p/2} \ |\mathbf{\Sigma}|^{-1/2} \ \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}\right), \tag{2}$$

kde $\boldsymbol{\mu}$ je vektor středních hodnot a $\mathbf{\Sigma}$ je kovarianční matice.

Vícerozměrné normální rozdělení má tyto vlastnosti:

- lineární kombinace prvků z $\mathbf{X}$ mají normální rozdělení
- všechny podmnožiny $\mathbf{X}$ mají normální rozdělení
- nekorelovanost veličin z $\mathbf{X}$ (složek vektoru $\mathbf{X}$) znamená i jejich nezávislost
- všechna podmíněná rozdělení jsou normální

Pro jednorozměrné normální rozdělení z rov. (2) dostaneme

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \ \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

V exponentu je čtverec vzdálenosti

$$u^2 = \left(\frac{x-\mu}{\sigma}\right)^2,$$

tedy vzdálenosti x od střední hodnoty μ kde jednotkou vzdálenosti je σ.

I pro vícerozměrné normální rozdělení je možno chápat kvadratickou formu v exponentu jako čtverec vzdálenosti vektoru $\mathbf{x}$ od vektoru $\boldsymbol{\mu}$, ve kterém je obsažena i informace z kovarinační matice

$$C^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

C je Mahalanobisova vzdálenost, pro zvolenou hodnotu $f(\mathbf{x})$ její čtverec je geometricky plocha elipsoidu se středem μ a osami $c\sqrt{\lambda_j \mathbf{v}_j}$ pro $j = 1, 2, \cdots, p$, kde λ_j jsou vlastní čísla matice $\boldsymbol{\Sigma}$ a $\mathbf{v}_j$ jsou vlastní vektory matice $\boldsymbol{\Sigma}$.

$$C^2 = (\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(p)$$

Příklad 3.1 Pro dvourozměrné normální rozdělení s parametry $EX_1 = \mu_1$, $EX_2 = \mu_2$, $\mathrm{var}X_1 = \sigma_1^2$, $\mathrm{var}X_2 = \sigma_2^2$ a kovariancí σ_{12} je korelační matice

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$$

neboť korelační koeficient $\rho = \sigma_{12}/(\sigma_1\sigma_2)$. Determinant kovarianční matice je pak

$$|\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2(1 - \rho^2).$$

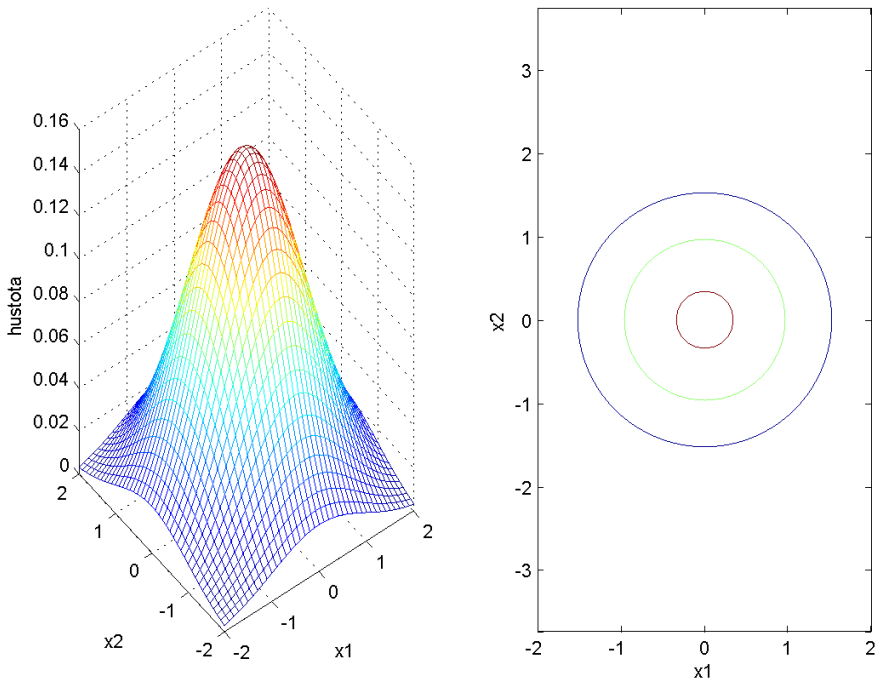
Vidíme, že tento determinant je roven nule, když $\rho^2 = 1$.

Podmíněné rozdělení $X_1|x_2$ je normální se střední hodnotou $\beta_0 + \beta_1 x_2$ a rozptylem $\sigma_1^2(1 - \rho^2)$

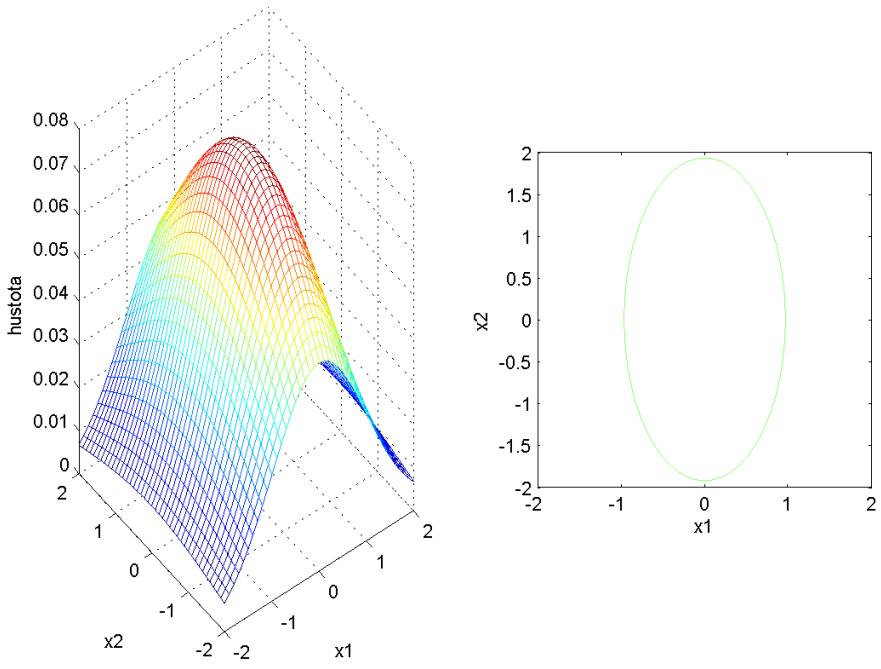
$$\beta_1 = \frac{\sigma_{12}}{\sigma_2^2} \qquad \beta_0 = \mu_1 - \beta_1\mu_2$$

Střední hodnota $X_1 \mid x_2$ závisí lineáně na x_2. Rozptyl X_1 nezávisí na x_2.

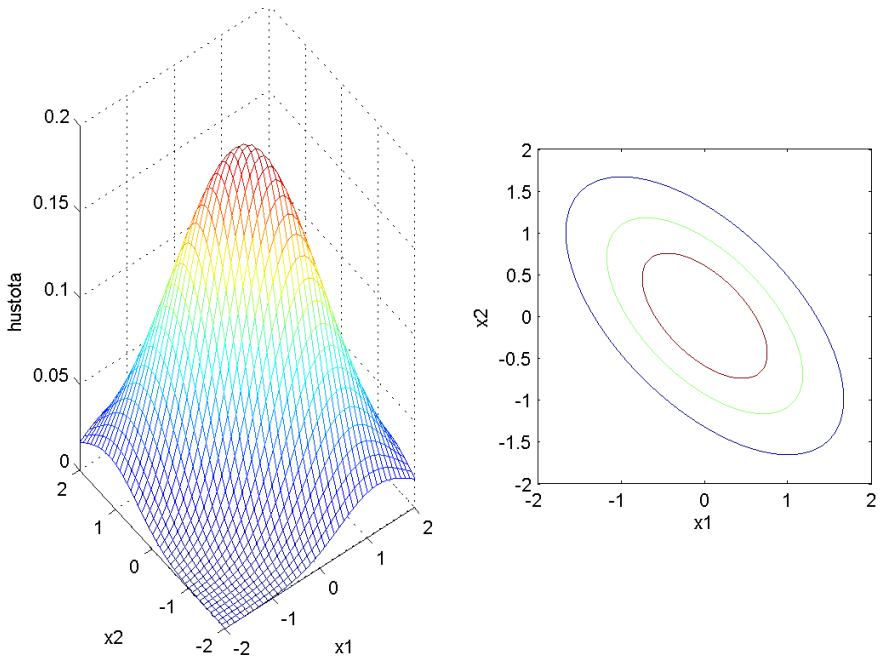
Pro dvourozměrné normální rozdělení můžeme elipsy konstantní hustoty ($f(x_1, x_2) = \mathrm{const}$) znázornit graficky. Když X_1, X_2 jsou nekorelované, tj. v případě vícerozměrného normálního rozdělení i nezávislé , osy elipsy konstantní hustoty jsou rovnoběžné s $x_1,\ x_2$, jinak jsou pootočeny. Názorně to ukazují následující obrázky.



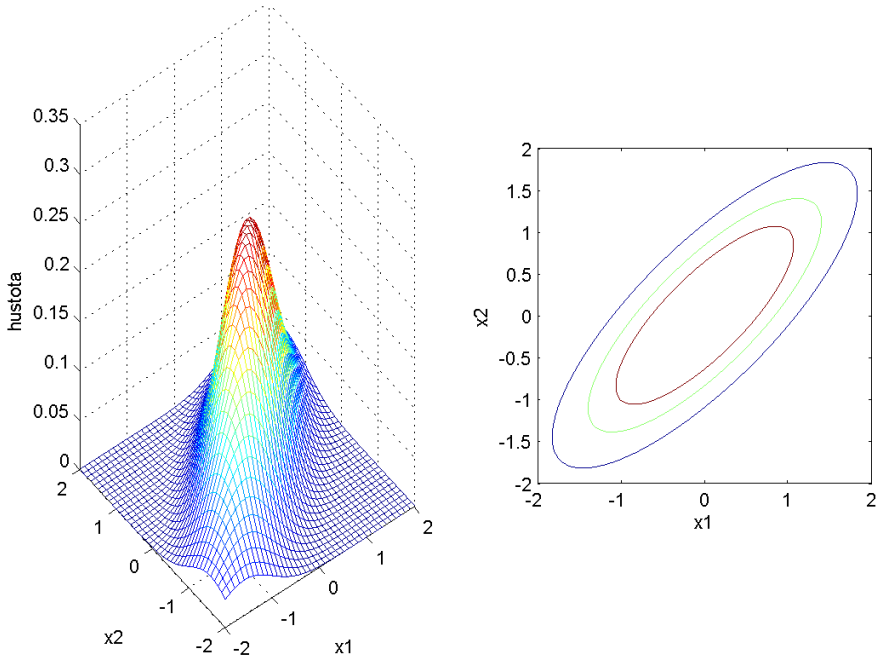
Obrázek 1: Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0$



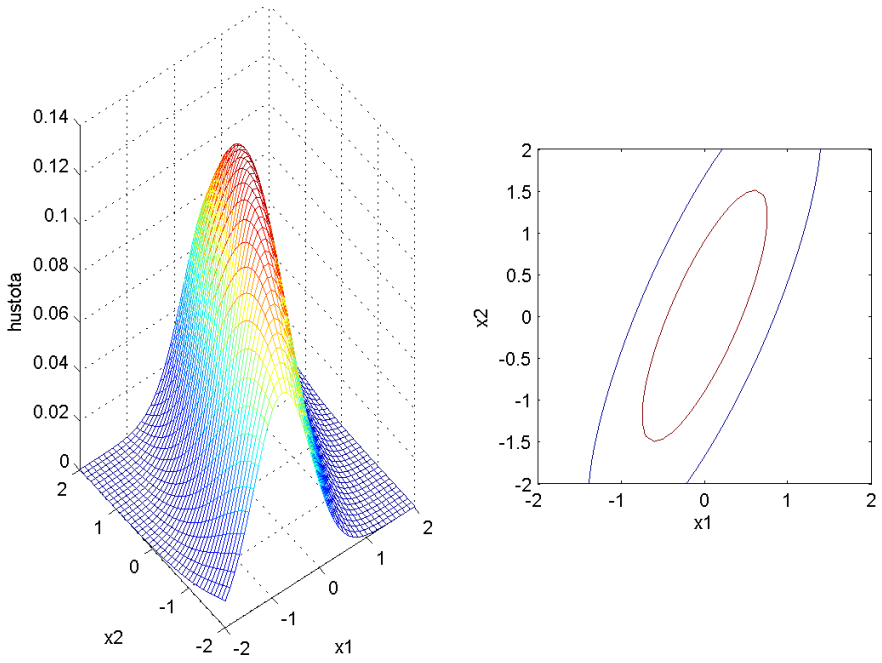
Obrázek 2: Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = 0$



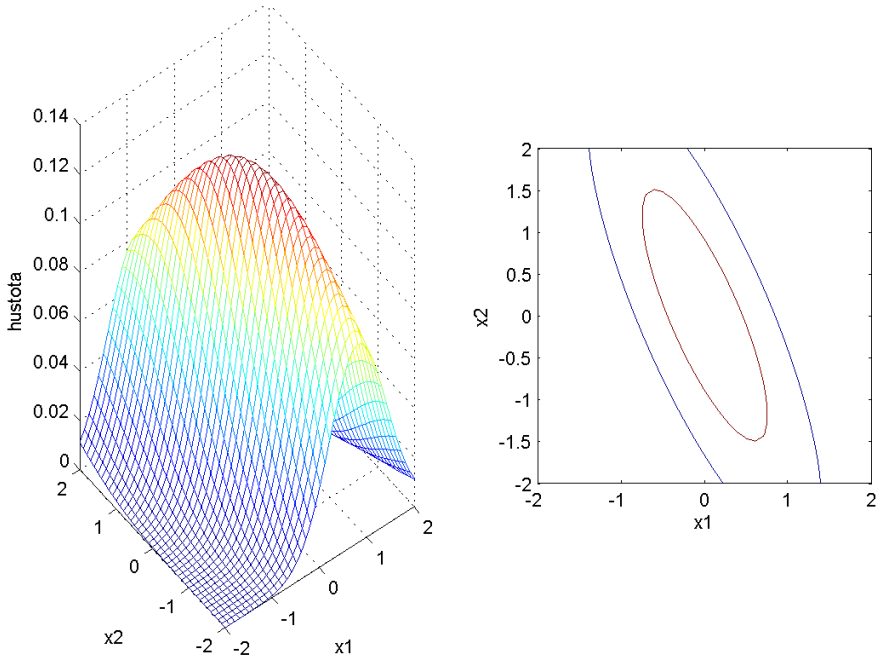
Obrázek 3: Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\rho = -0.6$



Obrázek 4: Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0.8$



Obrázek 5: Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = 0.8$



Obrázek 6: Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = -0.8$

Shrnutí

- *náhodný vektor, sdružené, marginální a podmíněné rozdělení*
- *nezávislost veličin*
- *vektor středních hodnot, kovarianční (varianční) matice*
- *vícerozměrné normální rozdělení*

Kontrolní otázky

1. *Sdružená pravděpodobnostní funkce náhodného vektoru* $[X, Y]^T$ *je zadána tabulkou*

	$X = x_1$	$X = x_2$	$X = x_3$
$Y = y_1$	p_{11}	p_{12}	p_{13}
$Y = y_2$	p_{21}	p_{22}	p_{23}

   Kolik marginálních a kolik podmíněných rozdělení má tento vektor? Vyjádřete všechny marginální pravděpodobnostní funkce a jednu zvolenou podmíněnou pravděpodobnostní funkci.

2. *Nechť* $\mathbf{A}$ *je čtvercová matice řádu* n*, jejíž prvky jsou reálná čísla (konstanty),* $\mathbf{x}$ *je náhodný vektor typu* $n \times 1$*,* $\mathbf{y} = \mathbf{Ax}$*. Vyjádřete vektor středních hodnot náhodného vektoru* $\mathbf{y}$ *a kovarianční matici náhodného vektoru* $\mathbf{y}$*.*

Korespondeční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

4 Mnohorozměrná data

Průvodce studiem

Tato kapitola má posloužit pro orientaci v problematice statistické analýzy vícerozměrných dat. Jsou zde uvedeny důležité výběrové charakteristiky, které lze z dat vyhodnotit. Také jsou zmíněny techniky ověřování předpokladů o rozdělení, zejména o normálním rozdělení a některé transformace, které lze užít v analýze dat. Počítejte se třemi hodinami studia s tím, že se k probírané látce budete ještě podle potřeby vracet.

Prozatím jsme se zabývali otázkami abstraktního popisu vztahů mezi náhodnými veličinami, především náhodným vektorem. Nyní obrátíme pozornost k praktičtějším problémům mnohorozměrných dat. Připomeňme, že veličiny, které zjišťujeme na sledovaných objektech, jsou různého typu. Jednak podle oboru jejich hodnot rozlišujeme:

- veličiny spojité - mohou nabývat nespočetně mnoho hodnot. Př: čas, délka
- veličiny nespojité (diskrétní) - nabývají jen spočetně mnoho hodnot, v praxi jen konečného počtu a často několika málo možných hodnot, např. kategoriální veličiny, vyjadřující příslušnost k nějaké skupině (kategorii) objektů
- alternativní (dichotomické,binární) veličiny, patří mezi nespojité, ale tím, že mohou nabývat jen dvou možných hodnot, často interpretovaných jako ANO/NE, TRUE/FALSE nebo 1/0, bývá někdy užitečné nahlížet na ně jako na zvláštní typ veličin.

Dále veličiny můžeme rozlišovat podle škály, ve které měříme:

- nominální (kategoriální)
- ordinální (pořadové)
- rozdílové (intervalové) kvantitativní (metrické, je definována vzdálenost dvou hodnot)
- poměrové kvantitativní (metrické, je definována vzdálenost dvou hodnot)

Kromě těchto hledisek, která klasifikují veličiny, je důležité mít i na paměti, jak vlastně data vznikla, co zobrazují a jaké jsou vztahy mezi pozorovanými objekty. Podstatné je, zda objekty můžeme považovat za nezávislé nebo zda vznikla jako řada pozorování téhož objektu v různých obdobích. Různé situace rozlišuje následující tabulka.

Úlohy řešené analýzou dat, časový prvek v datech:

počet objektů	počet veličin	počet období	typ úlohy
1	p	1	kasuistika, případová studie
n	1	1	jednorozměrná analýza
1	1	T	jednorozměrná časová řada
n	1	T	$T = 2$ párové srovnání, $T > 2$ opakovaná měření
n	p	1	vícerozměrná analýza dat
1	p	T	vícerozměrná časová řada
n	p	T	longitudinální studie

Poznámka: $n, p, T > 1$

Analýza vícerozměrných dat většinou se zabývá daty, kdy máme p veličin pozorovaných na n objektech. Rozlišit můžeme následující situace:

- všechny veličiny metrické – n bodů v $\mathbf{R}^p$
- všechny veličiny kategoriální – mnohorozměrné kontingenční tabulky
- smíšená data – datová matice se rozdělí na podsoubory – analogie s dvou/vícevýběrovými úlohami

Pro charakterizaci vícerozměrných dat potřebujeme odhadnout charakteristiky p-rozměrného vektoru náhodných veličin, tj. vektor středních hodnot a kovarianční matici (je symetrická), tedy určit následující počet výběrových charakteristik:

$$2p + \frac{p(p-1)}{2} = 2p + \frac{p^2 - p}{2} = \frac{p^2 + 3p}{2}$$

Počet odhadovaných parametrů tedy roste *kvadraticky* s počtem veličin, např. pro $p = 10$ je to 65 charakteristik, pro $p = 40$ je to už 860 charakteristik.

Pokud jsou veličiny kategoriální, potřebujeme odhadovat i sdruženou pravděpodobnostní funkci. Počet políček v tabulce roste exponenciálně s počtem veličin. Tudíž pokud mají být tyto odhady důvěryhodné, potřebujeme, aby vícerozměrná data byla měla dostatečný rozsah, tzn. n bylo velké.

4.1 Výběrové charakteristiky

Vícerozměrná dat jsou reprezentována datovou maticí $(n \times p)$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

tzn., že $i-$tý řádek datové matice je řádkový vektor pozorování p veličin na $i-$tém objektu.

Vektor průměrů jednotlivých veličin

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

tj.

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \quad j = 1, 2, \cdots, p$$

Wishartova matice typu $(p \times p)$ má prvky

$$w_{j,j'} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}), \qquad j, j' = 1, 2, \cdots, p$$

a můžeme ji zapsat také jako

$$\mathbf{W} = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

Výběrová kovarianční matice je opět typu $p \times p$

$$\mathbf{S} = \frac{1}{n-1} \mathbf{W},$$

její prvky jsou výběrové kovariance tj.

$$s_{jj'} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Korelační matice má prvky $r_{jj'} = s_{jj'}/(s_j s_{j'})$ tj. výběrové korelační koeficienty a má tvar

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

4.2 Lineární transformace proměnných

Při analýze vícerozměrných dat je často výhodné pracovat s odvozenými veličinami, které vzniknou z původních lineární transformací, např. s centrovanými proměnnými

$$v_{ij} = x_{ij} - \bar{x}_j,$$

pro které vektor průměrů je nulový, $\bar{\mathbf{v}} = \mathbf{0}$, a kovarianční a korelační matice se nezmění, tzn.

$$\mathbf{S}_v = \mathbf{S}_\mathbf{x}, \quad \mathbf{R}_v = \mathbf{R}_\mathbf{x}.$$

Další často užívanou transformací je normování. Normovaná hodnoty proměnných vzniknou vycentrováním a vydělením směrodatnou odchylkou původních proměnných, tj.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad i = 1, 2, \cdots, n \qquad j = 1, 2, \cdots, p.$$

Pak vektor průměrů je nulový $\overline{\mathbf{z}} = \mathbf{0}$, všechny rozptyly a směrodatné odchylky normovaných proměnných jsou rovny jedné a kovarianční i korelační matice normovaných proměnných jsou si rovny, tj. $\mathbf{S}_z = \mathbf{R}_z = \mathbf{R}_\mathbf{x}$ a jsou rovny korelační matici původních netrasformovaných proměnných.

Pro veličinu, která vznikne lineární kombinací původních proměnných

$$u_i = \mathbf{c}^T \mathbf{x}_i = \sum_{j=1}^{p} c_j x_{ij}$$

platí

$$\bar{u} = \mathbf{c}'^T \bar{\mathbf{x}}, \qquad s_u^2 = \mathbf{c}^T \mathbf{S}_\mathbf{x} \mathbf{c}.$$

4.3 Vzdálenost dvou objektů

Řádek datové matice, tj. vektor $\mathbf{x}^T$ můžeme považovat za souřadnice bodu v p-rozměrném prostoru. Pak je užitečné zabývat se vzdáleností dvou objektů. Jednorozměrná vzdálenost (když $p = 1$) dvou objektů i, i' je absolutní hodnota rozdílu pozorovaných hodnot,

$$d(i, i') = |x_i - x'_i|.$$

Pro vícerozměrná data můžeme definovat různé vzdálenosti. Eukleidovská vzdálenost dvou objektů i, i' je

$$D_E(i,i') = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2} = \sqrt{\sum_{j=1}^{p} d_j^2(i,i')}$$

Normovaná vzdálenost

$$D_N(i,i') = \sqrt{\sum_{j=1}^{p}(z_{ij} - z_{i'j})^2} = \sqrt{\sum_{j=1}^{p} \frac{d_j^2(i,i')}{s_j^2}}$$

Mahalanobisova vzdálenost respektuje jak rozdílnost ve variabilitě, tak korelační strukturu. Její čtverec je pak

$$D_M^2(i,i') = \mathbf{d^T S^{-1} d} = (\mathbf{x}_i - \mathbf{x}_{i'})^T \mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_{i'}),$$

kde $\mathbf{d} = \mathbf{x}_i - \mathbf{x}_{i'}$.

Pokud $\mathbf{S} = \sigma^2\mathbf{I}$ (všechny rozptyly jsou shodné, veličiny nekorelované), pak $D_N = D_M$.

Pro vyhledávání odlehlých pozorování je užitečná výběrová Mahalanobisova vzdálenost od těžiště našich pozorování (od vektoru průměrů)

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{i'})^T \mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{i'}).$$

Pro p-rozměrné normální rozdělení populace je

$$\frac{(n-p)n}{(n^2-1)p} D_i^2 \sim F(p, n-p),$$

podle toho tedy můžeme posudit, zda je pozorování odlehlé.

4.4 Chybějící hodnoty v datech

Zpracování mnohorozměrných dat v reálných úlohách je někdy komplikováno tím, že data nejsou úplná, hodnoty některých prvků datové nejsou k dispozici (slangově označovány jako missingy, missings).

Obvyklý jednoduchý postup je vypustit veličiny s mnoha missingy a vypustit případy (objekty) s mnoha missingy a úsudku o těchto veličinách a objektech se zříci.

Nejběžnější postup užívaný ve většině statistických paketů je automatické vypouštění případů (objektů) s jedním nebo více missingy, tzv. CASEWISE strategie. Tato strategie je z hlediska dalšího zpracování nejbezpečnější, ale může někdy vést k přiliš

velké ztrátě informace, kdy mnoho objektů je vyřazeno jen kvůli jedné chybějící hodnotě. Ale při této strategii výběrová kovarianční matice zůstane pozitivně definitivní (když hodnost $(\mathbf{X}) = p$), neboť hodnost $(\mathbf{X}^T\mathbf{X}) = p$.

Pro odhad kovarianční matice lze užít i strategii PAIRWISE, kdy každá kovariance se počítá ze všech možných dvojic a průměry v ní jen z hodnot užitých pro výpočet kovariance nebo strategii ALLVALUE, kdy pro průměry se užijí všechny možné (dostupné) hodnoty. Při tomto postupu se využije dat důkladněji, ale výběrová kovarianční matice nemusí být pozitivně definitivní.

Jiný přístup k práci s missingy je tak zvaná imputace, čili doplnění chybějících hodnot nějakými vhodnými, obvykle náhodnými hodnotami z rozdělení, které je shodné nebo podobné s rozdělením pozorovaných hodnot v datové matici. Cílem imputace je zabránit ztrátě informace nevyužitím všech pozorovaných hodnot v datové matici za cenu rizika, že informaci obsaženou v datech trochu zkreslíme. Běžnými metodami imputace, které jsou implementovány ve standardním statistickém software, jsou následující postupy doplnění chybějících hodnot:

- průměrem
- náhodně z předpokládaného rozdělení s využitím odhadu jeho parametrů, nejčastěji je chybějící prvek x_{ij} nahrazen hodnotami z $N(\mu, \sigma^2)$, kde hodnoty parametrů odhadneme z dostupných dat, $\hat{\mu} = \bar{x}_j,\ \widehat{\sigma}^2 = s_j^2$
- regresním modelem, jehož parametry odhadneme ze zbývajících $(n-1)$ objektů, chybějící hodnota se nahradí hodnotou predikovanou modelem, případně ještě modifikovanou náhodným kolísáním.

4.5 Ověřování normality

Mnoho metod vícerozměrné analýzy dat vychází z přepokladu, že náhodná složka v datech má normální rozdělení. Někdy je vyžadována vícerozměrná normalita, tj. p-rozměrné normální rozdělení, někdy stačí jen normalita některých veličin. Jak víme, p-rozměrné sdružené normální rozdělení má marginální rozdělení normální, avšak neplatí to naopak, tzn. marginální normalita nezaručuje sdruženou normalitu. Uvedeme stručně některé metody, kterými lze normalitu (většinou jen marginální) testovat.

Jednorozměrná normalita

Testy dobré shody, ve kterých se empirické rozdělení porovnává s normálním rozdělením. Rozpětí pozorovaných hodnot se rozdělí na r intervalů a porovnají se četnosti pozorovaných hodnot v jednotlivých intervalech s teoretickými četnostmi, které bychom očekávali při výběru stejného rozsahu z normálního rozdělení. Hranice intervalů se volí

- buď ekvidistantně (intervaly jsou stejně široké) tak, aby teoretické četnosti $\Psi_i > 1$ byly pro všechny intervaly a $\Psi_i > 5$ pro 80% hodnot (tzv. Cochranovo pravidlo).
- nebo hranice r intervalů se volí tak„ aby teoretické četnosti Ψ_i byly konstantní, nejčastěji se volí $\Psi_1 = \ldots = \Psi_r = 5$. Pokud se rozhodneme pro tento způsob volby hranic intervalů, volbou požadované hodnoty teoretické četnosti je určen při daném rozsahu výběru i počet intervalů r.

Testová statistika je pak

$$\sum_{i=1}^{r} \frac{(n_i - \Psi_i)^2}{\Psi_i^2} \sim \chi^2_{(r-1)}$$

Kolmogorovův test – v něm je výběrová distribuční funkce definována jako

$$F_n(0) = 0, \quad F_n(i) = \frac{i}{n}$$

a testovou statistikou je maximum absolutní hodnoty rozdílu porovnávaných distribučních funkcí

$$k_2 = \max(|F - F_n|).$$

Pro malé rozsahy výběru jsou kritické hodnoty této statisky tabelovány,

pro $n > 50$ se může užít asymtotická aproximace

$$k_2(0,95) \doteq \frac{1,36}{\sqrt{n}} \qquad k_2(0,99) \doteq \frac{1,63}{\sqrt{n}}$$

Testy šikmosti a špičatosti

Šikmost je definována jako

$$g_1 = \frac{M_3}{M_2^{3/2}}$$

a normální rozdělení má šikmost rovnou nule (je symetrické).

Špičatost je rovna

$$g_2 = \frac{M_4}{M_2^2} - 3$$

a i ta je pro normální rozdělení nulová, neboť u mormálního rozdělení poměr $M_4/M_2^2 = 3$.

K testování normality lze užít statistiky

$$K^{(3)} = \sqrt{\frac{g_1^2(n+1)(n+3)}{6(n-2)}}$$

a

$$K^{(4)} = \sqrt{\frac{(n+1)^2(n+3)(n+5)}{24n(n-2)(n-3)}} \left(g_2 + \frac{6}{n+1} \right),$$

které obě maji přibližně normované normální rozdělení $N(0,1)$.

4.6 Grafické metody ověřování normality

Tyto grafické metody umožňují rychlé vizuální posouzení shody empirického rozdělení s normálním (případně i jiným) rozdělením a proto jsou velmi často využívány a také jsou implementovány v běžně užívaném statistickém software. Kromě histogramů, do kterých se proloží i teoretické rozdělení a grafů porovnávajících empirickou distribuční funkci s teoretickou se užívá tzv. QQ-graf, kvantilový graf. QQ je zkratka pro kvantil (angl. quantile).

Pro sestrojení kvantilového grafu nejdříve uspořádáme výběr, tj.

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Hodnoty výběrové distribuční funkce se spočtou jako

$$VDF(x_{(i)}) = \frac{i - \frac{1}{2}}{n} \quad \text{nebo} \quad VDF(x_{(i)}) = \frac{i}{n+1}$$

a kvantily $x_{(i)}$ se vynesou do grafu proti odpovídajícím kvantilům normovaného normálního rozdělení, (tj.např. proti hodnotám $u(i/(n+1))$, kde $u(p)$ je p-kvantil normovaného normálního rozdělení). Pokud je výběrové rozdělení normální, grafem je přibližně přímka.

4.7 Transformace dat

Pokud zjistíme, že naměřená data nejsou z normálního rozdělení, je někdy užitečné použít vhodnou transformaci, aby transformací původní veličiny vznikla odvozená veličina, která normální rozdělení má. Potom lze aplikovat metody vyžadující normální rozdělení na transformované veličiny.

Transformací rozumíme takový přepočet, $y_i = f(x_i)$, aby se y_i , $i = 1, \ldots, n$ přiblížilo výběru z normálnímu rozdělení.

Ze zkušenosti lze doporučit tyto transformace:

- odmocninová transformace $y = \sqrt{x}$, když x jsou četnosti
- logitová transformace $y = \ln(\frac{x}{1-x})$, když x jsou relativní četnosti

- logaritmická transformace $y = \ln x$ pokud měřená veličina má log-normální rozdělení - např. výdaje na domácnost, náklady na výrobek atd.

Jinou možností je *transformace odvozená z dat*, kterou navrhli Box a Cox v roce 1964. Tato transformace poskytne hodnoty odvozené veličiny y , které se nejvíce přibližují normálnímu rozdělení.

$$y = \begin{cases} \frac{x^2-1}{\lambda} & \text{pro} \quad \lambda \neq 0 \\ \ln x & \text{pro} \quad \lambda = 0 \end{cases}$$

λ se odhadne jako $\lambda = \lambda_{max}$, které maximalizuje věrohodnostní funkci

$$\ln L(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \right]$$

Asymptotický interval $100(1-\alpha)\%$ – ní spolehlivosti pro λ:

$$2 \left[\ln L(\lambda_{max}) - \ln L(\lambda) \right] \leq \chi^2_{1-\alpha}(1),$$

čili v tomto intervalu jsou všechna x, pro která platí:

$$\ln L(x) \geq \ln L(\lambda_{max}) - \frac{1}{2} \chi^2_{1-\alpha}(1)$$

Charakteristiky pro data, která nejsou z normálního rozdělení
Takovými charakteristikami jsou ty, jejichž hodnoty nejsou ovlivněny odhlehlými hodnotami v datech. Jako příklad uvedeme uřezavaný průměr (trimmed mean):

$$\bar{x}(\alpha) = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} x_{(i)} \qquad m = \text{int} \left(\frac{\alpha n}{100} \right)$$

α je % uříznutých pořádkových statistik na každém konci.

Podobně lze zavést i uřezávaný odhad rozptylu atd.

Shrnutí

- *vektor výběrových průměrů, výběrová kovarianční matice, výběrová korelační matice*
- *ověřování normality, QQ graf*

Kontrolní otázky

1. *Vygenerujte si (např. v Excelu) náhodný výběr z rovnoměrného rozdělení o rozsahu 100 a zkonstruujte QQ graf*
2. *Vygenerujte náhodný výběr stejného rozsahu z normálního rozdělení, zkonstruujte QQ graf a porovnejte s QQ grafem z předchozí otázky*

5 Lineární regrese

Průvodce studiem

Tato kapitola je pro pochopení možností a technik analýzy vícerozměrných dat klíčová. Proto na tuto kapitolu počítejte nejméně se třemi hodinami usilovného studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí.

Regrese je jednou z velice často aplikovaných statistických metod, uvádí se, že dokonce naprostá většina aplikací $(70 - 90\%)$ je nějakou formou regresních metod. Počátky metody nejmenších čtverců jsou dokumentovány již na začátku 19. století (Legendre, Gauss), minimalizace součtu absolutních hodnot odchylek je připisována Galileovi ještě o pár desítek let dříve.

5.1 Klasický lineární model, metoda nejmenších čtverců

Klasický model lineární regrese lze zapsat jako

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{3}$$

kde

y_i je pozorovaná hodnota náhodné veličiny Y

$x_{i1}, \cdots, x_{ik}$ jsou hodnoty vysvětlujících proměnných (regresorů, prediktorů)

$\beta_0, \beta_1, \cdots, \beta_k$ jsou parametry modelu (fixní, leč neznámé hodnoty)

ε_i je náhodná složka

$i = 1, 2, \cdots, n$ je index pozorování (objektu)

Rovnici (3) můžeme zapsat maticově

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

Vektory jsou označeny příslušnými malými tučnými písmeny, matice velkými tučnými písmeny. Matice $\mathbf{X}$ má $k+1$ sloupců, v prvním sloupci jsou jedničky, dalších sloupcích jsou hodnoty vysvětlujících veličin.

Obvyklými předpoklady v klasickém lineární modelu jsou:

1. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, tj. vektor středních hodnot náhodné složky je roven nulovému vektoru
2. $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n, \quad \sigma^2 > 0, \quad \mathbf{I}$ je jednotková matice řádu n, tj. náhodné složky jsou nekorelované a jejich rozptyl je konstantní
3. $\mathbf{X}$ je nenáhodná matice typu $n \times (k+1)$

4. hodnost matice $\mathbf{X}$, $h(\mathbf{X}) = k + 1 \leq n$, tj. sloupce matice $\mathbf{X}$ nejsou lineárně závislé a počet pozorování je alespoň roven počtu parametrů

Z rovnice (4) dostaneme pro vektor podmíněných středních hodnot

$$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{5}$$

a pro kovarianční matici

$$\begin{gathered}\operatorname{cov}(\mathbf{y} \mid \mathbf{X}) = E\left\{[\mathbf{y} - E(\mathbf{y} \mid \mathbf{X})][\mathbf{y} - E(\mathbf{y} \mid \mathbf{X})]^T \mid \mathbf{X}\right\} = \\ = E\left\{[\mathbf{y} - \mathbf{X}\beta][\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^T \mid \mathbf{X}\right\} = E\left\{(\boldsymbol{\varepsilon} \mid \mathbf{X})(\boldsymbol{\varepsilon} \mid \mathbf{X})^T \mid \mathbf{X}\right\} = \sigma^2\mathbf{I}\end{gathered} \tag{6}$$

Neznámé parametry $\boldsymbol{\beta}$ lze odhadnout metodou nejmenších čtverců, tj. nalezením takového vektoru $\mathbf{b}$, pro který je nejmenší tzv. residuální suma čtverců (*RSS*) odchylek pozorovaných hodnot od jejich odhadů z modelu

$$\mathrm{RSS} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{7}$$

Položíme-li derivaci výrazu (7) podle vektoru $\mathbf{b}$, tj.

$$\begin{gathered}\frac{\partial\,\mathrm{RSS}}{\partial\mathbf{b}} = \frac{\partial}{\partial\mathbf{b}}(\mathbf{y}^T\,\mathbf{y} - \mathbf{b}^T\,\mathbf{X}^T\,\mathbf{y} - \mathbf{y}^T\,\mathbf{X}\,\mathbf{b} + \mathbf{b}^T\,\mathbf{X}^T\,\mathbf{X}\,\mathbf{b}) = \\ = \frac{\partial}{\partial\mathbf{b}}(-2\,\mathbf{b}^T\,\mathbf{X}^T\,\mathbf{y} + \mathbf{b}^T\,\mathbf{X}^T\,\mathbf{X}\,\mathbf{b}) = -2\,\mathbf{X}^T\,\mathbf{y} + 2\mathbf{X}^T\,\mathbf{X}\,\mathbf{b}\end{gathered}$$

rovnu nulovému vektoru, dostaneme soustavu normálních rovnic

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b} \tag{8}$$

a vzhledem k platnosti předpokladu (4) můžeme řešení této soustavy lineárních rovnic (vzhledem k $\mathbf{b}$) vyjádřit explicitně jako

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{9}$$

Odhady určené podle (9) se nazývají OLS – odhady (Ordinary Least Squares). Tyto odhady jsou nestranné, neboť

$$E(\mathbf{b}) = E(\mathbf{b}|\mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y} \mid \mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{10}$$

Lze ukázat, že tyto odhady mají další dobré vlastnosti, jsou to BLU-odhady (Best Linear Unbiased). Kovarianční matice těchto odhadů je

$$\operatorname{cov}(\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{11}$$

Na diagonále této matice jsou rozptyly odhadu parametrů $\operatorname{var}(b_i)$.

Nestranný odhad parametru σ^2 je (důkaz viz např. Anděl 1978)

$$s^2 = \frac{\text{RSS}}{n-k-1} \tag{12}$$

a nestranný odhad kovarianční matice odhadů parametru $\mathbf{b}$ je

$$\mathbf{S}_{bb} = s^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{13}$$

Na diagonále matice $\mathbf{S}_{bb}$ jsou tedy nestranné odhady rozptylů odhadu parametrů, jejich odmocninu (směrodatnou odchylku odhadu) označme $s(b_i)$

Přidáme-li k předpokladům (1) až (4) ještě předpoklad o tvaru rozdělení náhodné složky modelu (3), resp. (4), a to

$$(5) \qquad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \cdots, n$$

čili vyjádřeno vektorově s využitím předpokladu (2) $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, pak pro následující statistiku platí

$$\frac{b_i - \beta_i}{\sqrt{\text{var}(b_i)}} \sim N(0,1) \qquad i = 0, 1, \cdots, k$$

a pro

$$\frac{b_i - \beta_i}{s(b_i)} \sim t_{n-k-1} \tag{14}$$

Tuto statistiku pak můžeme užít ke stanovení intervalu spolehlivosti pro parametr β_i a testování hypotéz o tomto parametru.

5.2 Odhad parametrů metodou maximální věrohodnosti

Za předpokladu (5) můžeme odvodit i maximálně věrohodné (ML) odhady pro klasický model lineární regrese. ML-odhady odhadují hodnoty parametrů tak, aby tyto odhady maximalizovaly tzv. věrohodnostní funkci, tj. odhady jsou určeny jako nejpravděpodobnější hodnoty parametrů pro pozorovaná data. ML-odhady obecně mají řadu dobrých vlastností:

- jsou asymptoticky nestranné (s rostoucím n jejich střední hodnota konverguje k odhadovanému parametru)
- skoro vždy jsou konsistentní (s rostoucím n rozptyl odhadu konverguje k nule)

Věrohodnostní funkce (součin hustot jednotlivých pozorování) má pro klasický model lineární regrese tvar

$$L_{ML} = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{2\sigma^2}\right)$$

a její logaritmus je

$$\ln(L_{ML}) = L = (-\frac{n}{2}) \ln\left(2\pi\sigma^2\right) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/2\sigma^2 \tag{15}$$

Maximálně věrohodnými odhady jsou takové odhady $\boldsymbol{\beta}_{ML}$, pro které věrohodnostní funkce (15) nabývá maxima. Při hledání maxima funkce (15) položíme derivace podle hledaných proměnných rovny nule, tedy

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0}, \qquad \frac{\partial L}{\partial \sigma^2} = 0 \tag{16}$$

a dostaneme

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_{ML} \tag{17}$$

Vidíme, že ML-odhady regresních koeficientů jsou v klasickém lineárním modelu stejné jako odhady získané metodou nejmenších čtverců (OLS-odhady), $\boldsymbol{\beta}_{ML} = \mathbf{b}$, srovnej rov.(17) s rov.(9).

ML-odhad parametru σ^2 z rov.(16) je

$$\sigma_{ML} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\mathrm{RSS}}{n}$$

tedy se liší od OLS-odhadu v rov.(12), je pouze asymptoticky nestranný.

Shrnutí

- *lineární regresní model*
- *předpoklady v klasickém lineárním regresním modelu*
- *metoda nejmenších čtverců*
- *metoda maximální věrohodnosti*
- *kovarianční matice odhadů parametrů*

Kontrolní otázky

1. *Jaký je rozdíl mezi parametry a jejich odhady?*
2. *Jsou odhady parametrů náhodné veličiny?*
3. *Jaký je tvar kovarianční matice odhadů parametrů v klasickém modelu?*
4. *Jsou odhady získané metodou nejmeších čtverců nestranné? Dokažte to.*

Korespondeční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

6 Geometrie metody nejmenších čtverců a regresní diagnostika

Průvodce studiem

I tato kapitola je velmi důležitá pro pochopení principů statistické analýzy vícerozměrných dat. Počítejte nejméně se čtyřmi hodinami usilovného studia s tím, že se k probírané látce budete podle potřeby ještě vracet po pochopení dalších souvislostí.

6.1 Geometrie metody nejmenších čtverců

Uvažujeme klasický lineární regresní model

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

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Jak víme z předchozí kapitoly, odhad parametrů můžeme vyjádřit explicitně

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Vektor $\hat{\mathbf{y}} = \mathbf{Xb}$ je lineární kombinací vektorů regresorů, tj. leží v prostoru (přímce, rovině, nadrovině), jehož dimenze je rovna počtu regresorů. Dosadíme-li za $\mathbf{b}$, dostaneme

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{Hy}$$

Matice $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ je matice *projekce* vektoru $\mathbf{y}$ do prostoru určeného vektory regresorů. Požadavek formulovaný v metodě nejmenšich čtverců, tj. RSS $= (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$ vlastně znamená, že tato projekce je ortogonální. Pak tedy vektory $\hat{\mathbf{y}}$ a $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ jsou ortogonální vektory, tzn. $\hat{\mathbf{y}}^T\ \mathbf{e} = 0$, o čemž se velmi snadno můžeme přesvědčit:

$$(\mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb}) = \mathbf{b}^T\mathbf{X}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{Xb} = \mathbf{b}^T(\mathbf{X}^T\mathbf{y}) - \mathbf{X}^T\mathbf{Xb}) = 0,$$

neboť výraz v závorce je nulový vektor, viz normální rovnice.

Vektoru $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ se říká vektor residuí, jeho složkám $e_i = y_i - \hat{y}_i$ pak *residua*. Součet a tedy i průměr residuí je roven nule:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0,$$

neboť z první normální rovnice platí, že $\bar{y} = \mathbf{b}^T\bar{\mathbf{x}}$, kde $\bar{\mathbf{x}}^T = [1, \bar{x_1}, \bar{x_2}, \ldots, \bar{x_k}]$, tudíž

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(\hat{y}_i - \sum_{i=1}^{n} y_i = \mathbf{b}^T \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}) = 0,$$

neboť součet odchylek od průměru je nulový.

6.2 Rozklad součtu čtverců

Variabilitu vysvětlované veličiny můžeme vyjádřit jako součet čtverců odchylek pozorovaných hodnot od jejich průměru. Tuto charakteristiku nazýváme celkový součet čtverců, TSS.

$$\text{TSS} = (\mathbf{y} - \bar{\mathbf{y}})^T(\mathbf{y} - \bar{\mathbf{y}})$$

Lze ukázat, že tuto celkovou sumu čtverců můžeme rozložit na dvě složky

$$\text{MSS} = (\hat{\mathbf{y}} - \bar{\mathbf{y}})^T(\hat{\mathbf{y}} - \bar{\mathbf{y}})$$

a už dříve definovanou

$$\text{RSS} = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{e}^T\mathbf{e}$$

Platí tedy, že

$$\text{TSS} = \text{MSS} + \text{RSS},$$

MSS je ta část z celkového součtu čtverců, která je vysvětlena závislostí vysvětlované veličiny na regresorech, zbylou část (RSS) lineární závislostí vysvětlit nelze.

Nyní můžeme zavést důležitou charakteristiku toho, jak úspěšně regresní model vysvětluje variabilitu vysvětlované veličiny. Této charakteristice se říká *koeficient (index) determinace*, R^2.

$$R^2 = \frac{\text{MSS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Vidíme, že $0 \leq R^2 \leq 1$. Hodnota indexu determinace $R^2 = 1$, když RSS $= 0$, tzn. regresní model vysvětluje závislost vysvětlované veličiny na regresorech úplně (dokonalá lineární závislost). Naopak, $R^2 = 0$, když model nevysvětluje nic, tedy RSS = TSS, což nastane jen tehdy, když všechny odhady $b_1 = b_2 = \ldots = b_k = 0$ a $b_0 = \bar{y}$, např. pro $k = 1$ je regresní přímka rovnoběžná s osou x v úrovni $b_0 = \bar{y}$.

Z rozkladu celkového součtu čtverců vychází i analýza rozptylu, která je obvyklou součástí regresních programů. Tabulka analýzy rozptylu má většinou tento formát:

zdroj variability	stupně volnosti	součet čtverců	průměrný čtverec	F	$p-$value
model	k	MSS	MSS$/k$	$\frac{\text{MSS}/k}{\text{RSS}/(n-k-1)}$	$0.\ldots$
error	$n-k-1$	RSS	RSS$/(n-k-1)$		
total	$n-1$	TSS			

Za předpokladu, že průměrný čtverec $s^2 = \text{RSS}/(n-k-1)$ je opravdu nestranným odhadem rozptylu náhodné složky, σ^2, tzn. v modelu jsou zařazeny všechny relevantní regresory a tedy RSS není zvětšeno systematickou závislostí na nezařazeném regresoru (podrobněji viz např. Draper a Smith, kap. 2 a 24) a náhodné kolísání má normální rozdělení, má statistika F rozdělení $F \sim F_{k,n-k-1}$ a můžeme ji užít k testu hypotézy

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ proti

H_1 : aspoň jeden parametr $\beta_j \neq 0, \qquad j = 1, 2, \ldots, k$

Povšimněme si, že důležitou informaci o variabilitě residuí $e_i = y_i - \bar{y}_i$ a tím i o shodě modelem predikovaných hodnot $\bar{y}_i$ s pozorovanými hodnotami y_i nám poskytuje směrodatná odchylka residuí (square root mean error)

$$s = \sqrt{\frac{\text{RSS}}{n-k-1}}$$

Index determinace má tendenci nadhodnocovat podíl modelu na vysvětlení celkové variability veličiny y, mimo jiné i proto, že kvůli náhodnému kolísání jsou odhady $b_j \neq 0$ i tehdy, když $\beta_j = 0, \quad j = 1, 2, \ldots, k$. Proto se zavádí tzv. adjustovaný index determinace R^2_{adj},

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

Vidíme, že $R^2_{adj} < R^2$, rozdíl je výrazný tehdy, když počet pozorování je jen o málo větší než počet regresorů v modelu. Naopak hodnota R^2_{adj} se přibližuje R^2 pro $n \gg k$.

6.3 Regresní diagnostika

Další informace o vhodnosti modelu a o tom, zda jsou splněny předpoklady učiněné pro klasický lineární model můžeme získat z analýzy residuí. Vektor residuí můžeme vyjádřit pomocí projekční matice $\mathbf{H}$:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{I}\mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Pak kovarianční matice residuí je

$$\begin{gathered}\mathrm{cov}(\mathbf{e}) = \mathrm{cov}\left[(\mathbf{I}-\mathbf{H})\mathbf{y}\right] = (\mathbf{I}-\mathbf{H})\mathrm{cov}(\mathbf{y})(\mathbf{I}-\mathbf{H})^T = \\ (\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{H})^T = \sigma^2\mathbf{I}((\mathbf{I}-\mathbf{H})^T = \\ \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^T = \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}^T+\mathbf{H}\mathbf{H}^T) = \\ \sigma^2(\mathbf{I}-\mathbf{H})\end{gathered}$$

neboť projekční matice $\mathbf{H}$ je symetrická ($\mathbf{H}^T = \mathbf{H}$) a idempotentní ($\mathbf{H}^2 = \mathbf{H}$):

$$\mathbf{H}\mathbf{H}^T = \mathbf{H}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

Vektor reziduí $\mathbf{e}$ – náhodný (je funkcí náhodných vektorů $\mathbf{y}$ a $\mathbf{b}$)

Matice $\mathbf{H}$ s prvky $h_{ij}, \quad i,j = 1,2,\ldots,n$ je symetrická, ale nemusí být diagonální. Jak bylo v předchozím odstavci ukázáno, kovarianční matice vektoru residuí je rovna

$$\mathrm{cov}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$$

Nestranným odhadem parametru σ^2 je reziduální rozptyl (tzn. rozptyl ε_i):

$$s^2 = \frac{1}{n-k-1}\,\mathbf{e}^T\mathbf{e}$$

Dále uvedeme některé charakteristiky, které se užívají v tzv. *regresní diagnostice*, tj. při analýze vhodnosti modelu.

Klasická residua
Jsou to residua, který už jsme se zabývali,

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}.$$

Jejich rozptyly

$$\mathrm{var}(e_i) = s_e^2(1-h_{ii}),$$

nejsou konstantní, i když $\mathrm{var}(\epsilon_i) = \sigma^2$ konstantní je.

Normovaná residua
Jsou to klasická residua, vydělená reziduální směrodatnou odchylkou:

$$e_{Ni} = \frac{e_i}{s}$$

Jejich rozptyl je roven

$$\mathrm{var}(e_{Ni}) = 1 - h_{ii},$$

tedy nemusí být roven jedné.

Standardizovaná rezidua

Někdy se jim říká vnitřně studentizovaná residua (internally studentized), jsou definována takto:

$$e_{Si} = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

a jejich rozptyl je kostantní, roven jedné.

Plně studentizovaná rezidua

Podle techniky užité v jejich definici se jim říká také JACKKNIFE residua, jsou konstruována tak, že vždy pro $i-$tý bod se residuum počítá z modelu, jehož parametry byly odhadnuty ze zbývajících $n-1$ bodů, tedy vždy $i-$tý bod se vypustí.

$$e_{Ji} = \frac{e_{(-i)}}{s_{(-i)}\sqrt{1-h_{ii}}}.$$

kde $s_{(-i)}$ je residuální směrodatná odchylka při vynechání i-tého bodu (řádku datové matice). Tato residua mají $t-$rozdělení, $e_{Ji} \sim t(n-k-2)$.

Leverage

Tyto charakteristiky ohodnocují vliv i-tého bodu na hodnoty odhadů parametrů. Jsou to diagonální prvky projekční matice, tedy hodnoty h_{ii}. Platí, že

$$0 < h_{ii} < 1 \qquad \text{a} \qquad \sum_{i=1}^{n} h_{ii} = k+1,$$

kde k je počet regresorů. Hodnota h_{ii} je úměrná vzdálenosti i-tého pozorování od těžiště (v k-rozměrném prostoru regresorů), h_{ii} se považuje za velké, když $h_{ii} > 2(k+1)/n)$.

Cookova vzdálenost

Tato charakteristika slouží také k posouzení vlivu i-tého pozorování na odhady parametrů modelu, tj. hodnoty $\mathbf{b}$. Cookova vzdálenost pro i-té pozorování je definována

$$C_i = \frac{(\mathbf{b}-\mathbf{b}_{(-i)})^T(X^TX)(\mathbf{b}-\mathbf{b}_{(-\mathbf{i})})}{ps^2} = \frac{h_{ii}}{p(1-h_{ii})}e_{Si}^2$$

kde $\mathbf{b}_{(-i)}$ jsou jsou jackknife odhady (spočítané při vypuštění i-tého bodu) a p je počet odhadovaných parametrů. Cookova vzdálenost ohodnocuje vliv i-tého pozorování na odhad vektoru regresních parametrů $\mathbf{b}$. Je-li Cookova vzdálenost $C_i \geq 1$, i-pozorování velmi podstatně ovlivňuje odhady parametrů.

6.4 Autokorelace

Při posuzování předpokladu o nekorelovanosti residuí se obvykle vychází modelu autokorelačního procesu prvního řádu – AR(1):

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + u_i \qquad \text{kde} \qquad u_i \sim N(0, \sigma^2)$$

Autokorelační koeficient prvního řádu ρ_1 odhadujeme jako

$$\hat{\rho}_1 = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$$

K testování korelovanosti residuí se pak užívá *Waldův test*

$$W_a = \frac{n\hat{\rho}_1^2}{1 - \hat{\rho}_1^2} \sim \chi^2(1)$$

nebo ve statistickém software běžně implementovaná *Durbin – Watsonova* statistika

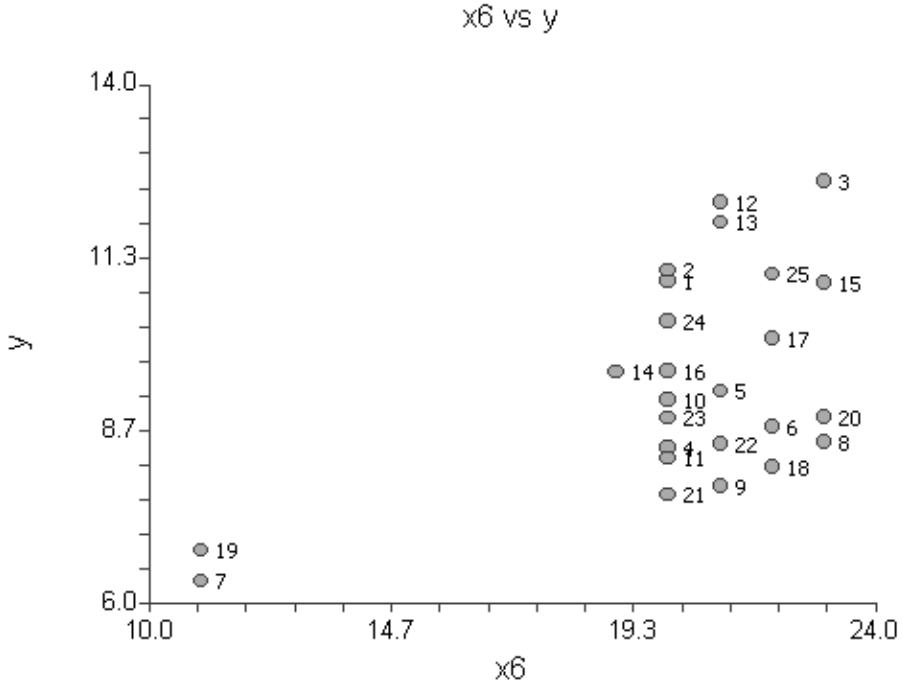
$$D_W = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \simeq 2\,(1 - \hat{\rho}_1)$$

Pro tuto statistiku platí $0 \leq D_W \leq 4$, $E(D_W) = 2$ při $\rho_1 = 0$. Kvantily této statistiky je obtížné vyjádřit explicitně, proto pro Durbin–Watsonův test statistické programy neposkytují u jiných testů obvyklý komfort, totiž i dosaženou významnost (p). Při rozhodování je pro hodnoty statistiky velmi blízké dvěma spoléhat na intuici a považovat residua za nekorelované. Pro serióznější úsudek lze využít přibližné kritické hodnoty, které jsou tabelovány, např. [16].

Příklad 6.1 Data pro tento příklad jsou převzata z knihy [8], příklad 01A, a jsou uvedena i v následující tabulce. Veličina y je měsíční spotřeba páry ve firmě, veličina $x6$ je počet pracovních dní v měsíci a veličina $x8$ je vnější teplota ve stupních Fahrenheita. Úloha, kterou máme řešit, je odhadnout parametry lineárního regresního modelu a posoudit, zda je tento model vhodný pro vysvětlení závislosti y na $x6$ a $x8$. V řešení tohoto příkladu byl užit statistický programový systém NCSS [14], zejména modul Multiple Regression, old version. Výstupy uvádíme bez větších editačních úprav v surovém stavu.

i	y	$x6$	$x8$	i	y	$x6$	$x8$
1	10.98	20	35.3	14	9.57	19	39.1
2	11.13	20	29.7	15	10.94	23	46.8
3	12.51	23	30.8	16	9.58	20	48.5
4	8.40	20	58.8	17	10.09	22	59.3
5	9.27	21	61.4	18	8.11	22	70.0
6	8.73	22	71.3	19	6.83	11	70.0
7	6.36	11	74.4	20	8.88	23	74.5
8	8.50	23	76.7	21	7.68	20	72.1
9	7.82	21	70.7	22	8.47	21	58.1
10	9.14	20	57.5	23	8.86	20	44.6
11	8.24	20	46.4	24	10.36	20	33.4
12	12.19	21	28.9	25	11.08	22	28.6
13	11.88	21	28.1				

Vyšetřujeme-li nějakou závislost, vždy je dobré data nejdříve prohlédnout jednoduchými prostředky popisné statistiky. Ty nám často pomohou odhalit zajímavé věci v datech, např. odlehlé hodnoty. To můžeme vidět i na následujícím grafu, kde pozorování na řádcích 7 a 19 jsou zcela mimo hodnoty ostatních pozorování (patrně je to počet pracovních dnů v měsících, kdy byly dovolené a ve firmě se nepracovalo). Tato pozorování jsou apriori podezřelá a při diagnostice modelu je nutné si na ně dát pozor.



Další užitečné nahlédnutí do analyzovaných dat je výběrová korelační matice, která je volitelnou součástí výstupu z modulu Multiple Regression:

```
Multiple Regression Report
Dependent   y
Correlation Matrix Section
            x6         x8         y
x6          1.000000 -0.209761  0.536122
x8         -0.209761  1.000000 -0.845244
y           0.536122 -0.845244  1.000000
```

Vidíme, že regresory $x6$ a $x8$ jsou jen slabě korelovány (korelační koeficient je -0,21), takže matice regresorů má plnou hodnost, nehrozí numerické potíže spojené se špatnou podmíněností.

Další pro nás důležitou součástí výstupu z modulu Multiple Regression je Regression Equation Section, kde jsou odhady parametrů modelu.

```
Regression Equation Section
Independent   Regression     Standard    T-Value     Prob
Variable      Coefficient    Error     (Ho: B=0)    Level
Intercept      9.12688       1.102801      8.2761   0.000000
x6             0.2028154     4.576761E-02  4.4314   0.000210
x8            -7.239294E-02  7.999381E-03  9.0498   0.000000
```

Vidíme, že u všech třech odhadovaných parametrů zamítáme nulovou hypotézu. Model tedy neobsahuje nadbytečné parametry.

Jak vidíme v následující tabulce ANOVA, model vysvětluje významnou část z celkové varibility veličiny y, podle hodnoty $R^2 = 0.8491$ asi 85% z celkové varibility. Odhad residuální směrodatné odchylky je uveden jako Root Mean Square Error a je roven přibližně 0,66. Druhá mocnina této charakteristiky je pak odhadem residuálního rozptylu σ^2.

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	2220.294	2220.294		
Model	2	54.1871	27.09355	61.9043	0.000000
Error	22	9.628704	0.4376684		

Root Mean Square Error	0.6615651
Mean of Dependent	9.424
R-Squared	0.8491
Adj R-Squared	0.8354

Pro posouzení vhodnosti modelu jsou důležité výstupy z regresní diagnostiky. Durbin-Watsonova statistika je velmi blízká hodnotě 2, tak se nemusíme znepokojovat autokorelací residuí.

Durbin-Watson Value	2.1955

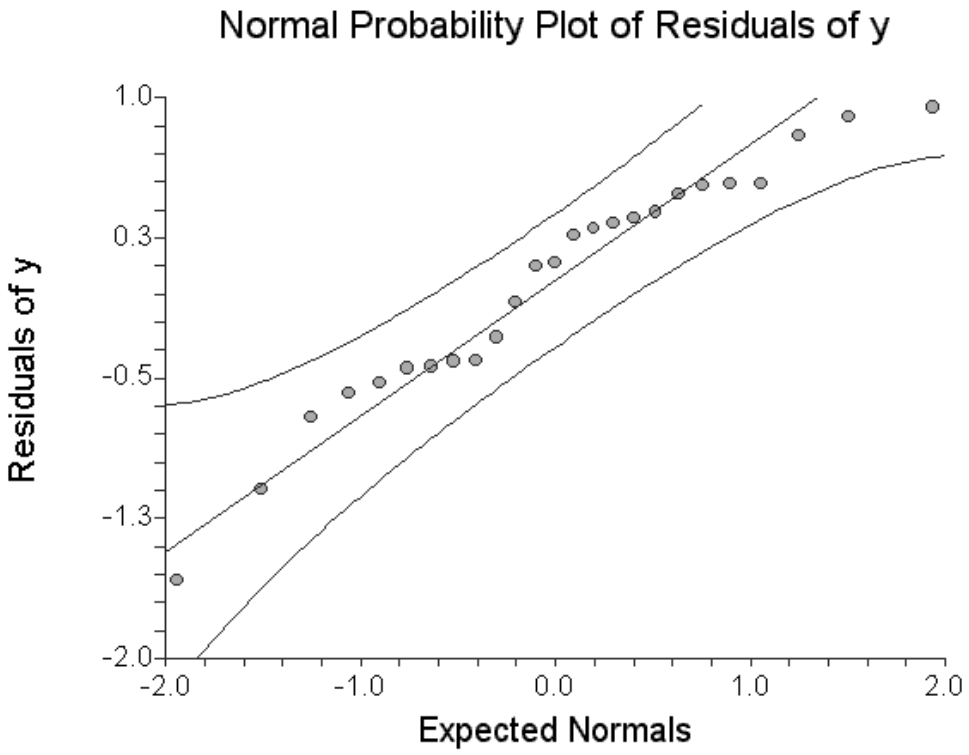
V následující tabulce jsou jackknife residua označena jako Rstudent. Pozornost věnujme především řádkům 7 a 19, které byly podezřelé už při předběžné jednoduché analýze a pak těm řádkům, kde studentizovaná residua jsou v absolutní hodnotě velká (zhruba větší než 2, což je přibližná hodnota (0.975) kvantilu $t-$rozdělení.

Regression Diagnostics Section

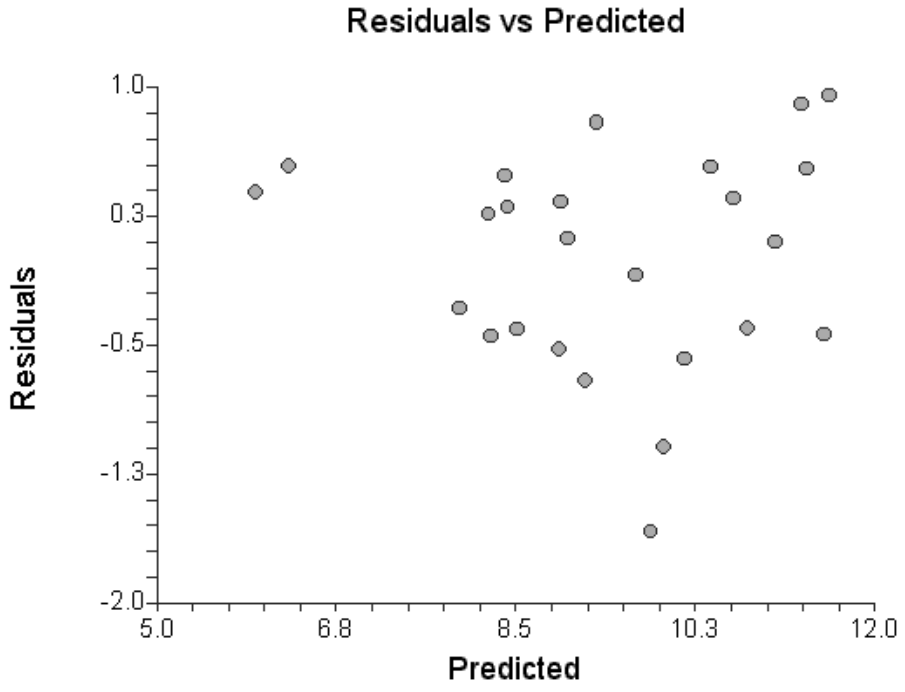
Row	Studentized Residual	Rstudent	Hat Diagonal	Cook's D
1	0.556825	0.547897	0.085491	0.009662
2	0.156002	0.152500	0.118877	0.001094
3	1.531855	1.583464	0.124826	0.111564
4	-0.814515	-0.808066	0.045374	0.010511
5	0.511834	0.503070	0.056434	0.005223
6	0.487210	0.478597	0.117502	0.010535
7	0.789332	0.782341	0.447410	0.168151
8	0.436764	0.428584	0.184719	0.014407
9	-0.711757	-0.703540	0.095491	0.017827
10	0.184529	0.180426	0.043373	0.000515
11	-2.452153	-2.810441	0.046418	0.097567
12	1.442820	1.481481	0.118566	0.093341
13	0.853098	0.847621	0.123991	0.034337
14	-0.913679	-0.910106	0.079880	0.024158
15	0.843302	0.837562	0.075758	0.019431
16	-0.142369	-0.139160	0.043079	0.000304
17	1.241672	1.258005	0.065527	0.036036
18	-0.658977	-0.650276	0.109838	0.017861
19	1.086621	1.091328	0.436460	0.304828
20	0.798049	0.791237	0.167792	0.042803
21	-0.450525	-0.442212	0.094228	0.007039
22	-1.100280	-1.105840	0.048654	0.020638
23	-1.697613	-1.779205	0.050307	0.050886
24	-0.644223	-0.635433	0.095790	0.014656
25	-0.708085	-0.699826	0.124217	0.023705

Největší residuum v absolutní hodnotě má pozorování 11, ale jak vidíme z ostatních statistik, neovlivňuje nijak významně hodnoty odhadů, je to odlehlý bod, který zvětšuje residuální rozptyl. Statistiky h_{ii} mají největší pozorování 7 a 19, jejich hodnoty jako jediné přesahují mezní hodnotu $2p/n = 6/25$, mají i největší Cookovu vzdálenost, ale zdaleka nedosahující hodnotu 1. Tyto dva body jsou tzv. vlivné, ale nevybočující. Naopak přispívají ke snížení residuálního rozptylu.

Následující obrázek ukazuje QQ graf residuí ukazuje, že rozdělení residuí můžeme považovat za normální, odchylky od přímky nejsou velké. Tedy data i model vyhovují předpokladu (5) klasického modelu a výsledky testů o parametrech modelu můžeme považovat za spolehlivé (nezavádějící).



Graf residuí proti odhadovaným hodnotám $\hat{y}$ ukazuje, že residuální rozptyl můžeme považovat za konstatní, „kazí“ to jen bod 11 s residuem menším než -2.



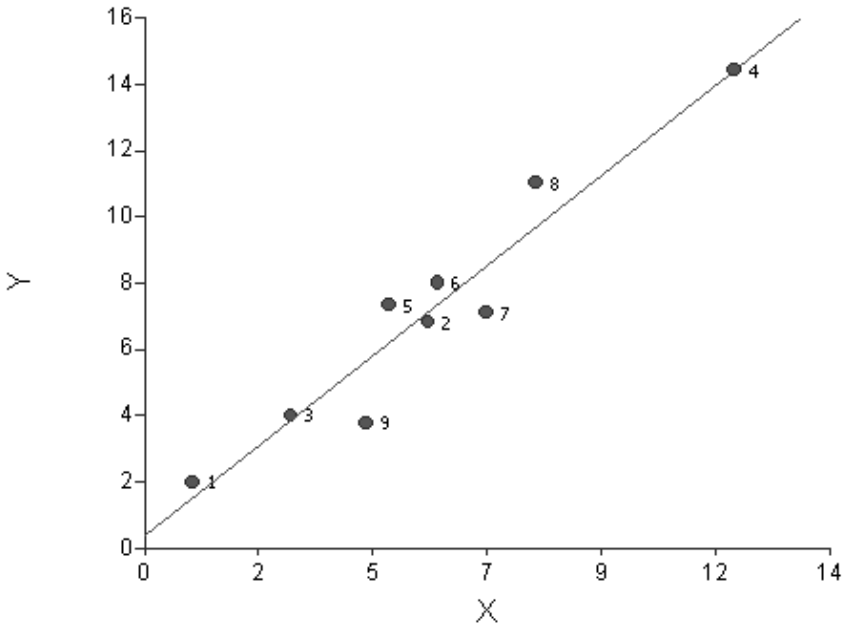
Výsledky tedy můžeme shrnout takto:

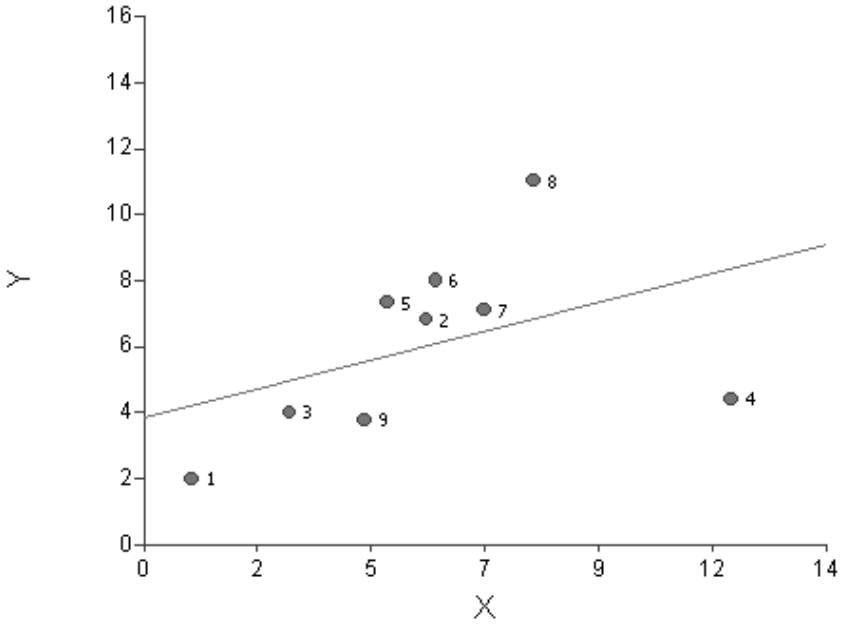
Hodnoty veličiny y (měsíční spotřeby páry ve firmě) lze uspokojivě odhadovat lineární závislostí na veličině $x6$ (počet pracovních dní v měsíci) a veličiny $x8$ (vnější teplota). Očekávanou spotřebu páry lze vyjádřit vztahem

$$\hat{y} = 9.127 + \ 0.2028\ x6 - 0.07239\ x8.$$

Směrodatná odchylka předpovědi je 0.662, model vysvětluje 85% z celkové variability spotřeby páry.

Příklad 6.2 Význam statistik pro diagnostiku ukazuje jednoduchý příklad modelu s jedním regresorem. Na obrázcích jsou data a proložené regresní přímky.





Odlišnost je jen v hodnotách vertikální souřadnice bodu 4. V obou úlohách mají body 4 velké hodnoty h_{ii}, ale liší se v hodnotách ostaních diagnostických statistik.

V případě prvním má bod 4 malé jackknife residuum i Cookovu vzdálenost, tzn. jeho vypuštěním či přidáním se hodnoty odhadů příliš nemění:

```
Row  Residual  Rstudent  Diagonal  Cook's D
4  -0.029151  -0.026991  0.605906  0.000653
```

Ve druhém případě jsou jackknife residuum i Cookova vzdálenost extrémně velké.

Row	Residual	Rstudent	Diagonal	Cook's D
4	-2.382046	-5.067308	0.605906	4.361897

Tento bod tedy má zásadní vliv na hodnoty odhadů, což je ostatně na první pohled vidět v této úloze, kdy model má je jeden regresor, i na grafech proložených regresních přímek. V případě modelů s více regresory ovšem takové vizuální posouzení není možné a regresní diagnostika je užitečným nástrojem pro ověření předpokladů a posouzení vhodnosti modelu.

Důsledky umístění bodu 4 vidíme v následující tabulce směrodatných odchylek:

	1 (vlivný)	2 (odlehlý)	bez bodu 4
intercept	0,849	1,951	1,152
směrnice	0,130	0,299	0,211
sm. odch. reziduí	1,15	2,65	1,25

Shrnutí

- *projekční matice, ortogonální projekce*
- *rozklad celkového součtu čtverců, index determinace* R^2
- *residua, jackknife residua, diagonální prvky projekční matice, Cookova vzdálenost*
- *autokorelace, Durbin–Watsonova statistika*

Kontrolní otázky

1. *Zkuste dokázat, že opravdu platí* $TSS = MSS + RSS$.
2. *Načrtněte, co znamená ortogonalita vektorů* $\hat{\mathbf{y}}$ *a* $\mathbf{e}$. *Pro graf zvolte model se dvěma regresory a výběr o rozsahu 3.*
3. *Jaká hypotéza se testuje v analýze rozptylu, uváděné ve výstupu statistických programů pro lineární regresi?*
4. *K čemu slouží regresní diagnostika?*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

7 Parciální a mnohonásobná korelace.

Průvodce studiem

Kapitola uvádí některé charakteristiky vyjadřující vztahy ve vícerozměrných datech. Jejich pochopení vám bude užitečné při studiu dalších metod analýzy dat. Na tuto kapitolu počítejte s jednou až dvěma hodinami studia.

Jak víme, korelační koeficient

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}X\,\mathrm{var}Y}}$$

vyjadřuje míru lineární závislosti dvou náhodných veličin X, Y. Platí, že

$$-1 \leq \rho(X,Y) \leq 1$$

.

Nyní uvažujme vektor náhodných veličin $[Y, X_1, X_2, \cdots, X_k]$.

Parciální korelační koeficient veličin Y, X_1 pak zapíšeme jako

$$\rho(Y, X_1 | X_2, X_3, \cdots, X_k) = \rho_{YX_1 \cdot X_2 X_3 \cdots X_k}$$

a znamená vlastně korelační koeficient mezi náhodnými veličinami, které „zbudou“ z veličin Y, X_1 po odečtení regresní funkce těchto veličin na podmiňujících veličinách, tj.

$$Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + \varepsilon^{(1)}$$

$$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon^{(2)}$$

pak parciální korelační koeficient je (obyčejný) korelační koeficient residuálních náhodných složek:

$$\rho_{YX_1 \cdot X_2 X_3 \cdots X_k} = \rho\left(\varepsilon^{(1)}, \varepsilon^{(2)}\right)$$

Parciální korelační koeficienty lze vyjádřit pomocí korelačních koeficientů dvojic veličin náhodného vektoru. Máme-li korelační matici

$$\boldsymbol{\varrho} = \begin{bmatrix} 1 & \rho(Y,X_1) & \rho(Y,X_2) & \cdots & \rho(Y,X_k) \\ \rho(X_1,Y) & 1 & \rho(X_1,X_2) & \cdots & \rho(X_1,X_k) \\ \rho(X_2,Y) & \rho(X_2,X_1) & 1 & \cdots & \rho(X_1,X_k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(X_k,Y) & \rho(X_k,X_1) & \rho(X_k,X_2) & \cdots & 1 \end{bmatrix}$$

pak

$$\rho_{YX_1\cdot X_2X_3\cdots X_k} = \frac{|\boldsymbol{\varrho}_{YX_1}|}{\sqrt{|\boldsymbol{\varrho}_{Y,Y}||\boldsymbol{\varrho}_{X_1X_1}|}}$$

kde $|\boldsymbol{\varrho}_{UV}|$ je determinant matice, která vznikne z korelační matice $\boldsymbol{\varrho}$, když je vypuštěn řádek U a sloupec V.

Např. pro $k = 2$ má korelační matice tvar

$$\begin{bmatrix} 1 & \rho(Y,X_1) & \rho(Y,X_2) \\ \rho(X_1,Y) & 1 & \rho(X_1,X_2) \\ \rho(X_2,Y) & \rho(X_2,X_1) & 1 \end{bmatrix}$$

a parciální korelační koeficient $\rho_{YX_1\cdot X_2}$ je

$$\rho_{YX_1\cdot X_2} = \frac{\rho(X_1,Y) - \rho(X_2,Y)\rho(X_1,X_2)}{[(1-\rho^2(X_1,X_2))(1-\rho^2(Y,X_2))]^{1/2}}.$$

Mnohonásobný (celkový) koeficient korelace nazývaný také *celkový* koeficient korelace vyjadřuje korelaci náhodné veličiny Y na veličinách $[X_1, X_2, \cdots, X_k]$. Vyjádříme-li náhodnou veličinu $\hat{Y}$ jako regresní funkci veličin $[X_1, X_2, \cdots, X_k]$,

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

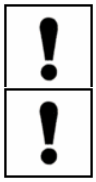
pak jednoduchý korelační koeficient $\rho\left(Y, \hat{Y}\right)$ je koeficient mnohonásobné korelace $\rho_{Y\cdot X_1X_2\cdots X_k}$,

$$\rho\left(Y, \hat{Y}\right) = \rho_{Y\cdot X_1X_2\cdots X_k}.$$

Koeficient mnohonásobné korelace lze spočítat z korelační matice,

$$\rho_{Y\cdot X_1X_2\cdots X_k} = \left(1 - \frac{|\varrho|}{|\varrho_{YY}|}\right)^{1/2}$$

Pro hodnoty mnohonásobného korelačního koeficientu platí



$$0 \leq \rho_{Y\cdot X_1X_2\cdots X_k} \leq 1,$$

při čemž nule je roven, když všechny dvojice $Y, X_i, \quad i = 1, 2, \cdots, k$ jsou nekorelované, tedy $\rho(Y, X_i) = 0$ pro $i = 1, 2, \ldots, k$. Jedničce je mnohonásobný korelační koeficient roven tehdy, když $\hat{Y} = Y$, tj. pro „čistou“ lineární závislost Y na $[X_1, X_2, \cdots, X_k]$.

Dále platí:

- $\rho_{Y\cdot X_1X_2\cdots X_k} \geq \max_{i=1,\cdots,k}[\rho(X_i, Y)]$
- $\rho(Y, X_1) \leq \rho_{Y\cdot X_1X_2} \leq \cdots \leq \rho_{Y\cdot X_1X_2\cdots X_k}$

Analogickým způsobem jsou definovány i *výběrové koeficienty parciální a mnohonásobné korelace*, jen se počítají z výběrových korelačních koeficientů (z výběrové korelační matice $\mathbf{R}$). K pochopení toho, co znamenají výběrové koeficienty parciální a mnohonásobné korelace nám pomůže jejich souvislost s lineárním regresním modelem (i když, přísně vzato, v klasickém lineárním modelu uvažujeme, že regresory jsou deterministické a korelace se týká náhodných veličin, ale chápejme tyto charakteristiky tak, že jsou počítány podle formulí zavedených pro výběrové korelační koeficienty a výběrové kovariance). Pro model s jedním regresorem $\hat{Y} = a_0 + a_1x_1$ platí, že

korelační koeficient	$r_{yx1} = \frac{s_{yx1}}{s_{x1}s_y}$
směrnice regresní přímky	$a_1 = \frac{s_{yx1}}{s^2_{x1}}$
koeficient determinace	$R^2 = r^2_{y,x1}$

Pro výběrový koeficient mnohonásobné korelace v modelu s k regresory platí, že jeho druhá mocnina je rovna koeficientu determinace, tedy

$$r^2_{y\cdot x_1x_2\cdots x_k} = R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

a pro výběrový koeficient parciální korelace platí to, že jeho druhá mocnina je rovna relativní změně residuální sumy čtverců, např.

$$r^2_{yx_2\cdot x_1} = \frac{\Delta\mathrm{RSS}(x_2|x_1)}{\mathrm{RSS}(x_1)},$$

kde jmenovatel je residuální součet čtverců modelu s jedním regresorem x_1 ($\hat{Y} = a_0 + a_1x_1$) a čitatel zlomku je snížení residuálního součtu čtverců způsobené přidáním regresoru x_2 do modelu s jedním regresorem, tj.

$$\Delta\mathrm{RSS}(x_2|x_1) = \mathrm{RSS}(x_1) - \mathrm{RSS}(x_1, x_2),$$

kde $\mathrm{RSS}(x_1, x_2)$ je residuální součet čtverců modelu se dvěma regresory $\left(\hat{Y} = b_0 + b_1x_1 + b_2x_2\right)$.

Shrnutí

- *koeficient parciální korelace, obor jeho hodnot*
- *koeficient mnohonásobné korelace, obor jeho hodnot*
- *vztah koeficientu mnohonásobné korelace a indexu determinace*

Kontrolní otázky

1. *Vyjádřete slovně, co charakterizuje koeficient parciální korelace*
2. *Vyjádřete slovně, co charakterizuje koeficient mnohonásobné korelace*

8 Výběr regresorů v mnohorozměrné regresi

Průvodce studiem

Kapitola se zabývá důležitou a často aplikovanou úlohou hledání vhodného regresního modelu. Na tuto kapitolu počítejte nejméně se třemi až čtyřmi hodinami studia. Důkladně prostudujte a promyslete i řešený příklad na konci této kapitoly.

Poměrně často se v analýze dat setkáváme s úlohami, které formálně mohou být zapsány jako klasický lineární regresní model,

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

ale v matici $\mathbf{X}$ typu $n \times (p+1)$ je počet regresorů p velký. Velký počet regresorů má často za následek, že pak pro mnoho z parametrů β_i, $i = 1, 2, \ldots, p$ nemůžeme zamítnout $H_0 : \ \beta_i = 0$, tzn. i−-tý regresor nevysvětluje změny hodnot veličiny y.

Naším cílem je najít jednodušší model s k regresory ($k < p$), obsahující jen takové regresory, které významnou měrou vysvětlují variabilitu hodnot y.

Řešit takou úlohu prozkoumáním všech lineárních modelů k regresory, $k = 1, 2, \ldots, p$, je pro větší hodnoty p časově neúnosné. Znamenalo by to prozkoumat

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p$$

modelů, tedy např. jen pro $p = 10$ výpočet a interpretaci výsledků odhadů 1024 modelů, což by představovalo práci na několik týdnů. Navíc pro velké hodnoty p bývá často matice $\mathbf{X}$ *špatně podmíněná*, tzn. determinant

$$\left|\mathbf{X}^T\mathbf{X}\right| \doteq 0$$

a odhady parametrů jsou pak numericky nestabilní a mají velké rozptyly, takže výsledky nejsou prakticky využitelné.

8.1 Kroková regrese

Jedním ze způsobů, jak nalézt pdomnožinu k regresorů do vhodného lineárního regresního modelu, je *kroková (stepwise)* regrese. Ještě dříve, než vysvětlíme tento postup, připomeneme základní pojmy a myšlenky, na kterých je založen. Víme, že celkový součet čtverců lze rozložit:

$$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS}$$

Dále, i modelovou sumu MSS můžeme rozložit. Představme si model s k regresory. Potom část MSS připadající i-tému regresoru,

$$\mathrm{MSS}(i \cdot 1, 2, \ldots, i-1, i+1, \ldots, k), \quad \text{označme ji zkratkou} \quad \mathrm{MSS}_{k(-i)}$$

$\mathrm{MSS}_{k(-i)}$ je tedy rozdíl modelové sumy čtverců $\mathrm{MSS}(k)$ při zařazení všech k regresorů a modelové sumy čtverců $\mathrm{MSS}(k-1)$ s $(k-1)$ regresory ($i-$tý regresor vynechán):

$$\mathrm{MSS}_{k(-i)} = \mathrm{MSS}(k) - \mathrm{MSS}(k-1).$$

Přidáním $i-$tého regresoru se současně změní odpovídajícím způsobem i residuální součet čtverců

$$\Delta\mathrm{RSS}_i = \mathrm{RSS}(k-1) - \mathrm{RSS}(k) = \mathrm{MSS}(k) - \mathrm{MSS}(k-1) = \mathrm{MSS}_{k(-i)},$$

neboť platí, že

$$\mathrm{TSS} = \mathrm{MSS}(k) + \mathrm{RSS}(k) = \mathrm{MSS}(k-1) + \mathrm{RSS}(k-1).$$

Současně víme, že $\Delta\mathrm{RSS}_i \geq 0$, tzn. přidáním regresoru se residuální součet čtverců nezvýší.

Myšlenka krokové regrese spočívá v tom, že v každém kroku (předpokládejme, že $k-1$ regresorů je už zařazeno v modelu) budeme z dosud nezařazených vybírat takový regresor, který nejvíce snižuje residuální sumu čtverců, tj. ten, jehož $\Delta\mathrm{RSS}_i$ je největší. Přitom zařadíme jen takový regresor, který residuální sumu čtverců snižuje významně. Kritériem (statistikou) pro posouzení významnosti je tzv. *parciální F*, což je

$$\frac{\Delta\mathrm{MSS}_i}{s^2} \sim F_{1,\nu},$$

kde $\Delta\mathrm{MSS}_i$ označuje zvýšení modelové sumy čtverců odpovídající zařazení $i-$tého regresoru z dosud v modelu nezařazených, s^2 je nestranný odhad parametru σ^2 a ν je jeho počet stupňů volnosti.

Implementace procedury krokové regrese se mohou lišit v tom, jakým způsobem je počítán s^2. Jedna z možností je počítat s^2 z residuální sumy čtverců v aktuálním kroku, tj.

$$s^2 = \frac{\mathrm{RSS}(k-1)}{n-k}.$$

Pak parciální F_i $i-$tého nezařazeného regresoru lze určit z parciálního a celkového korelačního koeficientu

$$F_i = \frac{[r^2_{iY\cdot(k-1)}]/(n-k)}{1 - r^2_{Y\cdot(k-1)}} = \frac{\Delta\mathrm{RSS}_i}{\mathrm{RSS}(k-1)/(n-k)},$$

kde $r_{iY\cdot(k-1)}$ je parciální korelační koeficient Y a x_i po „odečtení vlivu“ $(k-1)$ už zařazených regresorů a $r_{Y\cdot(k-1)}$ je mnohonásobný (celkový) koeficient korelace Y s $(k-1)$ už zařazenými regresory.

V $k-$tém kroku tedy zařadíme ten regresor, který má největší parciální F_i, a to jen tehdy, je-li F_i větší, než zadaná hodnota *F-to-entry*, kterou obvykle volíme jako takový kvantil F rozdělení, aby parciální F_i a tudíž i změna v residuálním součtu čtverců byly významné, tedy *F-to-entry* $= F_{1,(n-k-1)}(1-\alpha_1$, kde α_1 je zvolená hladina významnosti pro zařazení regresoru do modelu.

Po zařazení $i-$tého regresoru může kvůli korelaci mezi zařazenými regresory nastat situace, že parciální F některého ze zařazených regresorů přestane být významné. Jinými slovy, vypuštění tohoto regresoru z modelu pak nezvýší významně residuální sumu čtverců, tzn. regresor je v modelu nadbytečný. Proto se po zařazení regresoru spočítají parciální F_i všech dosud zařazených regresorů

$$F_i == \frac{\Delta\mathrm{RSS}(i)}{s^2},$$

kde $\Delta\mathrm{RSS}(i)$ znamená změnu (zvýšení) residuální sumy čtverců při vypuštění $i-$tého regresoru z modelu. Najde se nejmenší z těchto parciálních F_i a posuzuje se, zda bychom vypuštěním tohoto regresoru zvýšili RSS jen nepodstatně. Kriterium pro toto rozhodování je to, zda minimální F_i je menší než zadaná hodnota F-to-remove. Většinou volíme F-to-remove $= F_{1,(n-k-1)}(1-\alpha_2$, kde α_2 je zvolená hladina významnosti pro vypuštění regresoru z modelu. Abychom předešli možnosti nekonečného cyklu zařazování a vyřazování regresorů, obvykle se volí F-to-remove < F-to-entry, tj. $\alpha_2 > \alpha_1$.

Po tomto vysvětlení tedy můžeme algoritmus krokové regrese zapsat takto:

krok 0: zvol model se žádným regresorem , tj. $\hat{y} = \bar{y}$, a z p nezařazených regresorů zvol ten, který má největší absolutní hodnotu korelačního koeficientu s vysvětlovanou veličinou y (při jednom zařazeném regresoru je $R^2 = r_{xy}^2$, tedy nejvíce korelovaný regresor nejvíce snižuje residuální sumu čtverců)
if $F_i <$ F-to-entry **then** konec
else $k = 1$

krok k: mezi nezařazenými regresory vyber ten s největším F_i.
if $F_i <$ F-to-entry **then** konec
else zařaď $i-$tý regresor, $k = k + 1$
mezi zařazenými regresory najdi ten s nejmenším F_i,
if min $F_i <$ F-to-remove **then** vyřaď $i-$tý regresor, $k = k - 1$
go to krok k

Analogický krokový (stepwise) postup se, jak uvidíme dále, užívá nejen v lineární regresi, ale i v dalších metodách analýzy mnohorozměrných dat. Existují i některé další varianty, tzv. postupných procedur výběru veličin do modelu, např. zpětná *backward* procedura, která vychází z modelu, ve kterém je zařazeno všech p veličin a postupně vyřazuje nevýznamné, nebo dopředná *forward* procedura, která je podobná výše popsané krokové proceduře, avšak neumožňuje vypouštění zařazených veličin, které se stanou nevýznamné.

Obecně lze říci, že krokové procedury jsou užitečným nástrojem pro hledání vhodných modelů v mnohorozměrných datech, ale negarantují nalezení nejvhodnějšího modelu, nebot' ho mohou „minout“. Pro hledání vhodného lineárního regresního modelu je spolehlivější procedura popsaná v další kapitole, ale ta je výpočetně podstatně náročnější, takže pro velmi rozsáhlá data může být její využití problematické.

8.2 Hledání nejlepší množiny regresorů

Systematičtěji než kroková regrese pracují procedury označované jako „all possible regressions“ nebo „best subset of regressors“. Pro každé $k = 1, \ldots, p$ hledají takovou k-tici regresorů, aby $R^2 =$ bylo pro daný počet regresorů maximální. Jak bylo dříve uvedeno, počet modelů roste exponenciálně s počtem potenciálních regresorů p, procedury využívající jen hrubou sílu, tj. opravdu zkoumají všechny modely, mohou být užity jen pro poměrně malý počet potenciálních regresorů p, např. v NCSS 2000 je to $p \leq 15$. V některých statistických programech je implementována heuristika, kterou navrhli Furnival&Wilson (1974). Ta sice nezajišt'uje vyčerpávající prohledání všech modelů, ale zato dovoluje větší počet regresorů.

Jelikož s rostoucím k index determinace R^2 neklesá (obvykle roste), není vhodným kritériem pro optimalizaci modelu. Vhodnějším kriteriem je adjustovaný index determinace

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}\left(1-R^2\right) = R^2 - \left(1-R^2\right)\frac{k}{n-k-1}$$

nebo nejčastěji užívaná *Mallowsova statistika* C_p

$$C_p = [n-(k+1)]\,\frac{s_k^2}{s^2} - [n-2(k+1)] = n\left(\frac{s_k^2}{s^2}-1\right) - (k+1)\left(\frac{s_k^2}{s^2}-2\right),$$

kde s_k^2 je residuální rozptyl při k zařazených regresorech a s^2 je residuální rozptyl při všech zařazených regresorech. Střední hodnota této statistiky je

$$E(C_p) = k+1.$$

Když $(s_k^2/s^2) \approx 1$, tzn. model už nelze podstatně vylepšit, pak $C_p \approx k+1$ a zařazením zbytečného dalšího regresoru se C_p zvětší o 1. Tedy vzhledem k C_p je nejlepší ten model, který má C_p nejmenší přibližně rovné počtu zařazených regresorů zvětšených o jedničku. Obvykle je však C_p jen jedním z kriterií při hledání nejvhodnějšího modelu, musíme vzít do úvahy i residuální rozptyl a další, většinou nestatistická kriteria, jako počet regresorů (čím méně, tím obvykle lépe), cena jejich měření (levnější má přednost), interpretaci vlivu regresoru na vysvětlovanou proměnnou atd. v závislosti na konkrétní úloze.

Příklad 8.1 Užití krokové regrese a prohledání všech podmnožin regresorů při hledání vhodného regresního modelu si ukážeme na datech, která jsou v souboru STEPWISE.XLS. V datech máme 30 pozorování vysvětlované veličiny y a deseti potenciálních regresorů, $x1, x2, \ldots, x10$. Úkolem je najít vhodný lineární regresní model. Pro výpočty byly užity programy stepwise regression a all subset z [14]. Výstupy jsou opět uvedeny v surovém stavu, jen s drobným zkrácením.

```
Stepwise Regression Report

Dependent Y

Iteration Detail Section
```

Iter. No.	Action	Variab	R-Squared	Sqrt(MSE)	Max R-Sqrd Other X's
0	Unchanged		0.000000	3.010984	0.000000
1	Added	x1	0.765361	1.484322	0.000000
2	Unchanged		0.765361	1.484322	0.000000
3	Added	x5	0.826587	1.29947	0.300558
4	Unchanged		0.826587	1.29947	0.300558
5	Added	x6	0.985724	0.3799527	0.822143
6	Unchanged		0.985724	0.3799527	0.822143
7	Added	x8	0.988579	0.3465689	0.918500
8	Unchanged		0.988579	0.3465689	0.918500
9	Added	x2	0.990822	0.3170781	0.978202
10	Unchanged		0.990822	0.3170781	0.978202
11	Added	x3	0.993080	0.2812623	0.978713
12	Unchanged		0.993080	0.2812623	0.978713

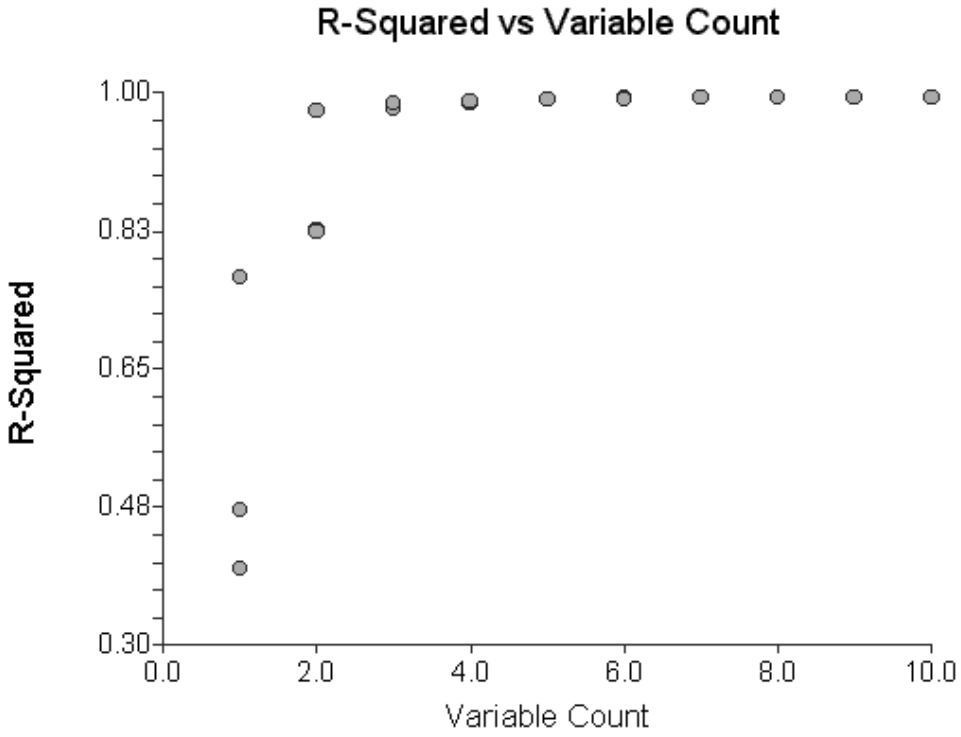
Při implicitním nastavení kritérií pro zařazování a vyřazování regresorů ($\alpha_1 = 0.05$, $\alpha_2 = 0.10$) byly postupně zařazovány regresory $x1$, $x5$, $x6$, $x8$, $x2$ a $x3$, žádný nebyl vyřazen. Při pohlednu na výsledky vidíme, podstatná změna v R^2 a residuální směrodatné odchylky nastala po zařazení regresoru $x6$, tedy pro model se třemi regresory $x1$, $x5$, $x6$. Přidávání dalších regresorů už index determinace R^2 nijak významně nezvýšilo a ani zmenšení residuální směrodatné odchylky není nikterak dramatické. Model se třemi regresory $x1$, $x5$, $x6$ je tedy nejnadějněším kandidátem na model, který vhodně vysvětluje variabilitu veličiny y. Zda je to opravdu vhodný model je nutno zkoumat podrobněji s využitím postupů uvedených v příkladu o odhadu regresních parametrů a pak i posoudit věcné souvislosti s řešeným problémem.

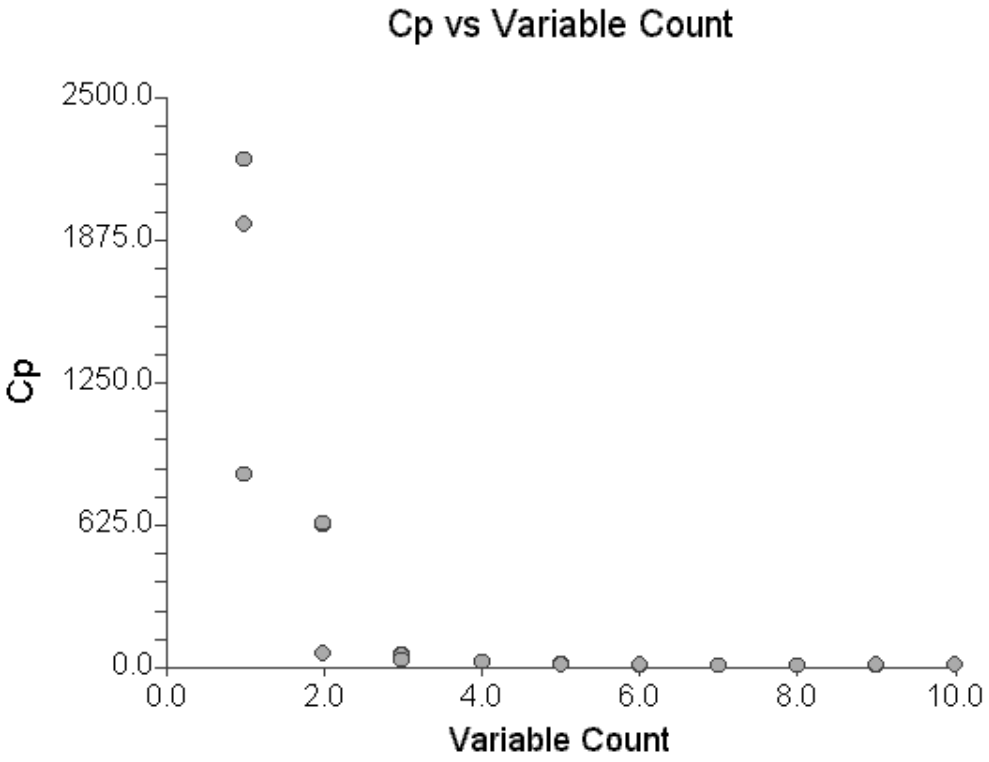
```
All Possible Results Section
```

Model Size	R-Squared	Root MSE	Cp	Model
1	0.765361	1.484322	848.389398	A (x1)
1	0.471363	2.227958	1943.984931	E (x5)
1	0.395810	2.381853	2225.534947	C (x3)
1	0.377170	2.418317	2295.000669	G (x7)
1	0.356763	2.457615	2371.046612	H (x8)
1	0.350940	2.468714	2392.745524	I (x9)
1	0.347381	2.475473	2406.008366	F (x6)
1	0.235375	2.679494	2823.404786	J (x10)
1	0.126059	2.864636	3230.773865	D (x4)
1	0.065507	2.962214	3456.422815	B (x2)
2	0.976958	0.4736817	61.867264	BE
2	0.826587	1.29947	622.230164	AE
2	0.823812	1.309826	632.570989	AC
3	0.985724	0.3799527	31.201415	AEF
3	0.980175	0.4477463	51.880208	BCE
3	0.978947	0.4614055	56.456614	BDE
4	0.988579	0.3465689	22.560822	AEFH
4	0.988347	0.3500752	23.426346	ABEF
4	0.987601	0.3611051	26.205962	ACEF
5	0.991966	0.2966695	11.939729	ACDEF
5	0.990822	0.3170781	16.200645	ABEFH
5	0.990702	0.3191541	16.649960	ACEFH
6	0.993525	0.2720509	8.127863	ACDEFH
6	0.993080	0.2812623	9.789421	ABCEFH
6	0.992527	0.29228	11.849453	ABDEFH
7	0.994075	0.2660938	8.079170	ABCDEFH
...				
8	0.994333	0.266362	9.118089	ABCDEFHJ
...				
10	0.994901	0.2656163	11.000000	ABCDEFGHIJ

Ve výstupu z programu All possible si povšimněme nejlepšího modelu se dvěma regresory $x2$ a $x5$, který má výrazně vyšší R^2 než ostatní modely se dvěma regresory a přibližuje se hodnotám R^2 s větším počtem regresorů. Přitom kroková procedura tento model „minula“. To je dosti názorná ilustrace nevýhod stepwise procedur, které jsou sice výpočetně méně náročné než úplné prohledávání, ale za cenu rizika takového minutí vhodného modelu.

Mallowsovo C_p má nejmenší hodnotu pro model se sedmi regresory, jen o málo je C_p větší pro nejlepší model se šesti regresory. Všimněme si, že tento nejlepší model se šesti regresory není shodný s tím, který byl nalezen krokovou procedurou, na místo regresoru $x2$ je zařazen regresor $x4$. Vidíme, že procedura All posible nám nabízí více kandidátů na vhodný model, než procedura Stepwise. Mezi těmito kandidáty je nutno pečlivě vybírat, rozhodně není jediný vhodný model s minimálním C_p. Pro výběr vhodných modelů jsou užitečná i grafická zobrazení statistik pro nalezené modely proti počtu regesorů. Jako příklad uvádíme grafy pro index determinace a Mallowsovo C_p, na kterých je jasně vidět výrazný skok v hodnotách statistik pro nejlepší model se dvěma regresory. Podobně je užitečný i graf závislosti residální směrodatné odchylky.





Opravdu vhodný model je však možno doporučit až po podrobnější analýze a porovnání jednotlivých kandidátů. Jak kroková procedura, tak All possible subsets nám jen generují návrhy, které je nutno podrobněji analyzovat.

Shrnutí

- *výběr regresorů do modelu*
- *kroková (stepwise) regrese*
- *hledání nejlepší množiny regresorů*
- *kriteria pro posouzení vhodnosti modelu*

Kontrolní otázky

1. *Vysvětlete principy a algoritmus krokové regrese.*
2. *Proč maximalizace R^2 není dobrou strategií při hledání vhodného modelu?*
3. *Proč minimalize Mallowsovy statistiky je přijatelnou strategií při hledání vhodného modelu?*
4. *Podle čeho se posuzuje, který model je vhodný pro danou úlohu?*
5. *Porovnejte výhody a nevýhody krokové regrese a prohledávání všech modelů.*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

9 Zobecnění klasického lineárního modelu

Průvodce studiem

V této kapitole jsou ukázány některé postupy, které rozšiřují oblast aplikace lineárního modelu, zejména za okolností, kdy předpoklady klasického modelu nejsou splněny. Prostudování kapitoly a pochopení souvislostí vyžaduje nejméně se tři až čtyři hodiny.

9.1 Transformace původních regresorů

Prozatím jsme se zabývali lineárním modelem, který obsahoval přímo hodnoty regresorů $x_{\cdot,1}, x_{\cdot,2}, \ldots, x_{\cdot,k}$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{18}$$

Jedním z mnoha možných zobecnění je model ve tvaru

$$y_i = \beta_0 + \beta_1 Z_{i1} + \cdots + \beta_{p-1} Z_{i,p-1} + \varepsilon_i, \tag{19}$$

kde každé $Z_{\cdot,j}$ je nějakou funkcí původních regresorů $x_{\cdot,1}, x_{\cdot,2}, \ldots, x_{\cdot,k}$.

Jako příklady takových modelů můžeme uvést (pro přehlednost zápisu jsou řádkové indexy vynechány):

1. $k = 1$, polynom stupně $p - 1$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1} x^{p-1} + \varepsilon$$

2. $k = 2$, $p = 6$, tzv. model 2. řádu

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

3. $k = 2$, $p = 10$, tzv. model 3. řádu

$$\begin{aligned} y = \quad & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \\ & + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \\ & + \beta_{111} x_1^3 + \beta_{112} x_1^2 x_2 + \beta_{122} x_1 x_2^2 + \beta_{222} x_2^3 + \varepsilon \end{aligned}$$

Další použitelné transformace jsou

- reciproká transformace, tj. $z_j = 1/x_j$, když $\forall x_j > 0$
- logaritmická transformace, tj. $z_j = \ln(x_j)$, když $\forall x_j > 0$
- odmocninová transformace, tj. $z_j = \sqrt{x_j}$, když $\forall x_j \geq 0$

a mnoho podobných dalších transformací a kombinace jejich užití v jednom modelu. Důležité však je, že modely 19 jsou lineární v parametrech, tj. soustava normálních rovnic je soustava p lineárních rovnic pro p odhadovaných parametrů. Splňuje-li model 19 předpoklady klasického lineárního modelu, můžeme pro analýzu a interpretaci takového modelu užít všechny techniky, které jsme dosud užívali pro klasický lineární model ve tvaru 18, tedy pro situaci, kdy jsme pracovali přímo s regresory, nikoliv s jejich funkcemi.

Na tvar lineárního modelu je možno někdy převést i modely, které na první pohled lineární nejsou, např. když závislost vysvětlované veličiny na regresorech x_1, x_2, x_3 může být proložena funkcí

$$\eta = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} \tag{20}$$

Po zlogaritmování dostaneme

$$\ln \eta = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \delta \ln x_3 \tag{21}$$

Pokud hodnoty vysvětlované náhodné veličiny y lze popsat modelem

$$\ln y = \ln \eta + \varepsilon \tag{22}$$

a platí, že $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, pak k odhadům parametrů $\alpha, \beta, \gamma, \delta$ a jejich interpretaci opět můžeme užít postupů známých z klasického lineárního modelu. Při linearizaci vztahů typu 20 však musíme být opatrní v tom, jakou roli má náhodné kolísání ε (tzv. chybový člen, error). Představa 22 znamená, že náhodné kolísání je *multiplikativní*, nikoliv aditivní, tj. hodnota náhodné veličiny y je vyjádřena modelem

$$y = \eta \exp(\varepsilon) = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} \exp(\varepsilon),$$

nikoliv modelem s aditivní chybou ve tvaru $y = \eta + \text{error}$. To, zda je oprávněné užít multiplikativní model, může vyplynout z věcné analýzy úlohy, ale často i v situacích, kdy model chyb je jiný, může být výsledek být výsledek získaný linearizací a aplikací lineárního modelu užitečným prvním přiblížením k řešení problému.

9.2 Aitkenův odhad

Řeší problém, kdy není splněn předpoklad 2 klasického model, že náhodné složky mají konstantní rozptyl a jsou nekorelované. Připust'me, že náhodné složky mohou být korelované a nemusí mít konstantní rozptyl:

$$\operatorname{cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \boldsymbol{\Omega}, \quad \sigma^2 > 0, \tag{23}$$

kde $\mathbf{\Omega}$ je pozitivně definitní matice. Pak existuje regulární matice $\mathbf{P}$, pro kterou platí

$$\mathbf{P\Omega P}^T = \mathbf{I} \quad \text{a} \quad \mathbf{P}^T\mathbf{P} = \mathbf{\Omega}^{-1} \tag{24}$$

Vynásobíme-li rov.(4), tj. klasický lineární model, maticí $\mathbf{P}$ zleva (transformujeme veličiny), dostaneme

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \tag{25}$$

Označíme-li $\mathbf{y}^* = \mathbf{Py}$, $\mathbf{X}^* = \mathbf{PX}$ a $\boldsymbol{\varepsilon}^* = \mathbf{P}\boldsymbol{\varepsilon}$, pak rov.(25) můžeme přepsat

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \tag{26}$$

Kovarianční matice náhodných složek v rov.(26) je pak

$$\mathrm{cov}(\boldsymbol{\varepsilon}^*) = E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*T}) = E(\mathbf{P}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{P}^T) = \sigma^2\mathbf{P\Omega P^T} = \sigma^2\mathbf{I}$$

tzn., že pro hvězdičkované veličiny je rov.(26) klasický lineární model. Vyjádříme rovnice pro odhady parametrů v modelu (26) pomocí původních netransformovaných veličin a dostaneme vztahy pro odhady parametrů modelu

$$\mathbf{b} = (\mathbf{X^T\Omega^{-1}X})^{-1}\mathbf{X^T\Omega^{-1}y}, \tag{27}$$

o kterých víme, že to jsou BLU-odhady s kovarianční maticí

$$\mathrm{cov}(\mathbf{b}) = \sigma^2(\mathbf{X^T\Omega^{-1}X})^{-1} \tag{28}$$

Nestranný odhad parametru σ^2 je

$$s^2 = (\mathbf{y} - \mathbf{Xb})^T\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{Xb}) \tag{29}$$

který pak můžeme užít k odhadu kovarianční matice a tedy i rozptylů odhadů b_i.

Odhady získané tímto postupem jsou nestranné, některé jsou dokonce BLU-odhady, avšak k jejich výpočtu potřebujeme znát matici $\mathbf{\Omega}$. Tu bohužel v analýze dat v naprosté většině případů neznáme. Nemůžeme tuto matici z dat ani konsistentně odhadnout, v datech máme n nezávislých pozorování a potřebujeme odhadnout $(n^2 + n)/2$ jejích prvků (diagonálu a polovinu nediagonálních prvků matice, matice $\mathbf{\Omega}$ je symetrická). Většinou nezbývá, než na místo předpokladu (23) přijmout nějaké větší omezení.

9.3 Heteroskedascita

Jedna z možností je řešit tzv. heteroskedastickou regresi, tj. připustit, že rozptyly náhodné složky nejsou konstantní, ale náhodné složky v lineárním regresním modelu (3) jsou nekorelované:

$$\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{diag}(w_1^2, w_2^2, \cdots, w_n^2) \tag{30}$$

kde $w_i^2 > 0$ je váha rozptylu i-tého pozorování. Pak matice $\boldsymbol{\Omega}$ je také diagonální,
$\boldsymbol{\Omega} = \mathrm{diag}(w_1^2, w_2^2, \cdots, w_n^2)$,
inverzní matice je $\boldsymbol{\Omega}^{-1} = \mathrm{diag}(w_1^{-2}, w_2^{-2}, \cdots, w_n^{-2})$
a $\mathbf{P} = \mathrm{diag}(1/w_1, 1/w_2, \cdots, 1/w_n)$.

Pak v modelu (3), tj. v datové matici vydělíme řádek vahou rovnou směrodatné odchylce pozorování

$$y_i/w_i = \beta_0/w_i + \beta_1 x_{i1}/w_i + \cdots + \beta_k x_{ik}/w_i + \varepsilon_i/w_i,$$

můžeme užít OLS-odhady, které budou mít dobré vlastnosti jako v klasickém modelu.

Otázkou je, jak určit váhu pozorování, w_i. Máme několik možností, záleží na řešené úloze:

(1) v modelu máme je jeden regresor x_{i1} a předpokládáme, že pozorování závislé veličiny y_i mají konstantní *relativní* chybu. Pak můžeme položit $w_i = x_{i1}$.

(2) pozorování závislé veličiny y_i mají konstantní *relativní* chybu a v modelu je více regresorů. Pak nezbývá, než vybrat jeden podle subjektivního rozhodnutí, možná ten, který nejvíce koreluje se závisle proměnnou y.

(3) nejdříve spočítat $\hat{y}_i$ jako OLS-odhad podle modelu (3) a pak ve druhém kroku položit $w_i = \hat{y}_i$

(4) postupovat jako ve variantě (9.3) a dále pokračovat v iteracích, dokud dva po sobě následující odhady nejsou dostatečně blízké. Tomuto postupu se říká metoda *iterovaných vážených čtverců*, iterated WLS (Weighted Least Squares).

9.4 Stochastické regresory

Také předpoklad v klasickém modelu, že matice $\mathbf{X}$ obsahuje pevné hodnoty, u kterých není třeba uvažovat s jejich rozptylem a korelací, je v mnoha aplikacích nerealistický. Pro tzv. *nezávislou stochastickou regresi*, kdy předpokládáme, že matice $\mathbf{X}$ je stochastická, tvoří $(k+1)$ rozměrný náhodný proces a náhodná složka $\boldsymbol{\varepsilon}$ nezávisí na $\mathbf{X}$, uvedeme stručně důležité výsledky, podrobněji viz např. [7, **?**].

OLS-odhady $\mathbf{b}$, spočítané podle rov.(9), s^2 podle rov.(12) a

$$\mathbf{S}_{bb} = s^2(\mathbf{X^T X})^{-1}$$

jsou nestranné. Ale nejsou to lineární odhady, nebot' jsou stochastickou funkcí náhodného vektoru $\mathbf{y}$ a nejsou to BLU-odhady. Za předpokladu, že v pravděpodobnosti konverguje $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/n \to \sigma^2$ a $\mathbf{X^T X}/n \to \boldsymbol{\Sigma}_{XX}$ (kde $\boldsymbol{\Sigma}_{XX}$ je kovarianční matice regresorů) však platí:

- $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ je konzistentním odhadem vektoru regresních koeficientů $\boldsymbol{\beta}$
- s^2 podle rov.(12) je konzistentním odhadem parametru σ^2
- $\mathbf{S}_{bb} = s^2(\mathbf{X^T X})^{-1}$ lze vzít za konsistentní odhad asymptotické kovarianční matice odhadovaných parametrů $\mathbf{b}$.

To znamená, že OLS-odhady lze užít k běžným testům a určení intervalů spolehlivosti pro parametry.

Pokud jsou náhodné složky modelu normálně rozděleny, jsou OLS-odhady podle rov.(9) také ML-odhady, takže mají dobré asymptotické vlastnosti, jsou konsistentní a asymptoticky eficientní.

9.5 Diskrétní regresory, umělé proměnné

Dosud jsme se zabývali úlohami, ve kterých vysvětlující veličiny byly spojité. Docela často se v analýze dat stává, že data pocházejí ze dvou nebo více populací, vzpomeňme např. na dvouvýběrové testy či analýzu rozptylu. I na taková data můžeme aplikovat lineární regresi. Uvažujme nyní nejjednodušší případ – lineární regresní model s jedním regresorem

$$EY_i = \beta_0 + \beta_1 x_i.$$

Parametr β_1 je směrnice přímky, tzn. vyjadřuje změnu střední hodnoty náhodné veličiny Y_i, změní-li se hodnota regresoru o jedničku. Uvažujme, že regresor x je diskrétní a má hodnoty $\{0, 1\}$, jinými slovy jen rozděluje data do dvou skupin (výběrů) ze dvou populací 0 a 1. Pak test hypotézy $\beta_1 = 0$ znamená totéž jako test hypotézy $\mu_0 = \mu_1$ (shoda středních hodnot obou populací), tj. dvouvýběrový t-test při shodných rozptylech.

Příklad 9.1 Máme otestovat hypotézu, že střední hodnoty veličiny Y ze dvou populací jsou shodné. Výběrové charakteristiky pro oba nezávislé výběry jsou v následující tabulce a obrázku.

výběr	n	průměr	sm. odchylka
0	30	5,06	1,04
1	20	5,96	2,09

K testu si můžeme vybrat několik metod, které nám dají shodné výsledky:

metoda	H_0	předpoklad	statistika	p
t-test(shodné rozptyly)	$\mu_0 = \mu_1$	$\sigma_0^2 = \sigma_1^2$	2,01	0,05
lineární regrese	$\beta_1 = 0$	$\sigma_0^2 = \sigma_1^2$	2,01	0,05
ANOVA	$\mu_0 = \mu_1$	$\sigma_0^2 = \sigma_1^2$	4,03	0,05

Jak vidíme, ve všech třech případech nám vyšla stejná hodnota p, pro t-test a lineární regresi i stejná hodnota statistiky, ačkoliv testujeme různé hypotézy, v analýze rozptylu je hodnota F-statistiky rovna druhé mocnině t-statistiky u ostatních dvou metod. V případě lineární regrese je odhad $b_1 = \bar{y}_1 - \bar{y}_0$ a $b_0 = \bar{y}_0$ a testujeme, zda rozdíl průměrů je dostatečně veliký k zamítnutí hypotézy $\mu_0 = \mu_1$ (připomeňme, že směrnice přímky je změna veličiny y při změně veličiny x o jedničku).

To, že uvedené statistiky vyšly stejně, není žádné překvapení, neboť se vyčíslují ze stejných formulí. Také předpoklady pro všechny uvedené testy jsou shodné, normálně rozdělená residua a shodné rozptyly v obou populacích.

Uvedený příklad ilustruje možnost podobného pohledu na analýzu rozptylu a lineární regresi, ukazuje, že diskrétní regresory mohou být docela snadno interpretovány a naznačuje směry dalšího zobecnění lineárního modelu.

Pokud diskrétní regresor nabývá více než dvou hodnot, lze k rozlišení užít tzv. *umělé proměnné, dummy variables*. Obvykle se jedna z r kategorií vybere jako referenční a $r-1$ dummy proměnných s hodnotami $\{0, 1\}$ pak kóduje kategorie. Odhad směrnice u konkrétní dummy proměnné znamená odhad změny střední hodnoty vysvětlované veličiny oproti referenční kategorii. Podrobněji viz kapitola Logistická regrese.

Při více diskrétních regresorech pomocné proměnné dovolují zkoumat regresní analýzou i velmi komplikované struktury závislosti, případně i v kombinaci s dalšími spojitými regresory tyto závislosti „očišťovat“ od vlivu jiných veličin. Podrobnější výklad takových postupů přesahuje rozsah tohoto kursu, v případě potřeby se obraťte na literaturu, např. Draper a Smith nebo Anděl atd. Tam najdete i další možnosti zavedení pomocných proměnných.

Shrnutí

- *tranformace regresorů, polynom, model druhého řádu, linearizace*
- *heteroskedascita, metoda vážených nejmenších čtverců*
- *diskrétní regresory, umělé proměnné*

Kontrolní otázky

1. *Jaké zjednodušení představuje metoda vážených nejmenších čtverců proti Aitkenovu odhadu?*
2. *Jak interpretovat směrnici regresní přímky v případě spojitých regresorů a jak v případě diskrétních regresorů?*
3. *Co jsou umělé (dummy) proměnné?*

10 Zobecněný lineární model (GLM)

Průvodce studiem

Zobecněný lineární model (GLIM) projděte spíše pro celkový přehled, snažte se však důkladně pochopit část o logistické regresi včetně řešeného příkladu. Na tuto kapitolu počítejte nejméně se čtyřmi hodinami studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí.

Významné zobecnění lineárního modelu zavedli Nelder a Wedderburn [**?**]. Tento článek patří k nejčastěji citovaným statistickým publikacím. Tento model, označovaný jako GLM nebo GLIM, je podrobněji popsán v knize McCullagh a Nelder [19] a do základních pojmů tohoto modelu nyní nahlédneme.

Zobecněný lineární modle (GLM) zahrnuje:

lineární regresi

různé modely analýzy rozptylu (ANOVA)

logistickou regresi

probitový model

log-lineární model (multinomický model pro četnosti v analýze mnohorozměrných kontingenčních tabulek)

Označení datových struktur a význam symbolů v GLIM:

Pozorování závisle proměnné (response) je sloupcový vektor náhodných veličin a je typu $(n \times 1)$, tedy $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$.

Pokud z kontextu je zřejmé, že se jedná o libovolný prvek vektoru $\mathbf{y}$, bude označován y (netučná kursiva bez indexu)

Matice $\mathbf{X}$ nezávislých proměnných (regresorů, covariates) je typu $(n \times p)$. Její j-tý sloupec označujeme $\mathbf{x}_j$.

Vektor parametrů je $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_p]^T$.

Náhodná složka modelu má vektor středních hodnot $\mathrm{E}(\mathbf{Y}) = \mu$ typu $(n \times 1)$ a kovarianční matici $\mathrm{cov}(\mathbf{Y})$

Lineární prediktor $\boldsymbol{\eta}$ je systematická složka v lineárním modelu, tedy

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$$

kde $\mathbf{x}_j$ je j – tý sloupec matice $\mathbf{X}$, tj. vektor $(n \times 1)$.

Každá složka vektoru $\mathbf{Y}$ má rozdělení z exponenciální rodiny rozdělení s hustotou

$$f_Y(y,\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\theta) + c(y,\theta)\}, \tag{31}$$

kde θ a ϕ jsou parametry rozdělení, $a(.), b(.), c(.)$ jsou funkce, jejichž tvar je dán konkrétním rozdělením z exponenciální rodiny. Pokud ϕ je známé, je rov.(31) hustota rozdělení z exponenciální rodiny a má *kanonický* parametr θ. Pokud ϕ je neznámé, pak to může, ale nemusí být dvouparametrické rozdělení z exponenciální rodiny.

Např pro normální rozdělení, $Y \sim N(\mu, \sigma^2)$

$$\begin{aligned} &f_Y(y,\theta,\phi) = \tfrac{1}{\sqrt{2\pi\sigma^2}}\exp\{-(y-\mu)^2/2\sigma^2\} = \\ &= \exp\{(y\mu - \mu^2/2)/\sigma^2 - \tfrac{1}{2}(y^2/\sigma^2 + \ln(2\pi\sigma^2))\}, \end{aligned}$$

takže v tomto případě

$$\theta = \mu, \quad \phi = \sigma^2,$$

$$a(\phi) = \phi \quad b(\theta) = \theta^2/2, \quad c(y,\phi) = -\frac{1}{2}(y^2/\sigma^2 + \ln(2\pi\sigma^2))$$

Logaritmus věrohodnostní funkce (při známém y funkce parametrů θ, ϕ) je

$$l(\theta,\phi,y) = \ln f_Y(y,\theta,\phi)$$

Střední hodnota a rozptyl může pak být určena ze vztahů známých pro věrohodnostní funkci:

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0 \qquad \mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right)^2 = 0$$

Věrohodnostní funkci pro jedno pozorování z jakéhokoli rozdělení z exponenciální rodiny lze zapsat

$$l(\theta,\phi,y) = (y\theta - b(\theta))/a(\phi) + c(y,\phi)$$

Derivace podle kanonického parametru jsou $\partial l/\partial\theta = (y - b'(\theta))/a(\phi)$ a $\partial^2 l/\partial\theta^2 = b''(\theta)/a(\phi)$.

Položíme-li je rovny nule, můžeme vyjádřit střední hodnotu a rozptyl vysvětlované náhodné veličiny:

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = (\mu - b'(\theta))/a/(\phi) = 0 \Rightarrow \mathrm{E}(Y) = \mu = b'(\theta)$$

a

$$\mathrm{var}(Y) - \frac{b''(\theta)}{a(\phi)} = 0 \Rightarrow \mathrm{var}(Y) = b''(\theta)a(\phi).$$

Střední hodnota je funkcí pouze kanonického parametru θ. Rozptyl náhodné veličiny Y je součinem dvou funkcí. Jedna, $b''(\theta)$ závisí pouze na kanonickém parametru rozdělení (a tedy na střední hodnotě μ náhodné veličiny Y). Nazývá se varianční

funkce (variance function) a můžeme ji zapsat jako funkci střední hodnoty, $V(\mu)$). Druhá funkce v součinu je nezávislá na kanonickém parametru θ a závisí jen na ϕ.

Funkce $a(\phi)$ má obvykle tvar $a(\phi) = \phi/w$. Parametr ϕ se nazývá disperzní parametr a je konstantní pro všechna pozorování, w je apriorně známá váha pozorování, může být různá pro různá pozorování.

GLIM tedy dovoluje i jiná rozdělení z exponenciální rodiny než jen normální užité v klasickém modelu. Další zobecnění je v tom, že lineární prediktor nemusí vysvětlovat jen (podmíněnou) střední hodnotu náhodné veličiny, ale i nějakou její funkci. Vztah mezi lineárním prediktorem η a střední hodnotou μ vysvětlované náhodné veličiny Y vyjadřuje spojovací funkce (link):

$$\eta_i = g(\mu_i)$$

Spojovací funkce $g(.)$ může být jakákoli monotónní diferencovatelná funkce.

V klasickém lineárním modelu je spojovací funkcí identita, tj. $\eta = \mu$. U jiných modelů se užívají zejména tyto spojovací funkce:

logit	$\eta = \ln\{\mu/(1-\mu)\}$	$0 < \mu < 1$
probit	$\eta = \Phi^{-1}(\mu)$	$0 < \mu < 1$
	$\Phi(.)$ je distribuční funkce	
	rozdělení N(0,1)	
komplementární		
log-log	$\eta = \ln\{-\ln(1-\mu)\}$	$0 < \mu < 1$
mocninové funkce	$\eta = \begin{cases} \mu^\lambda & \text{pro } \lambda \neq 0 \\ \ln \mu & \text{pro } \lambda = 0 \end{cases}$	$\mu > 0$

Jelikož střední hodnota $\mu = b'(\theta)$, je tedy jen funkcí kanonického parametru θ a spojovací funkce je monotónní, existuje inverzní funkce, kterou můžeme vyjádřit jako funkci střední hodnoty, $\theta(\mu)$. Některá rozdělení mají zvláštní spojovací funkce, kdy kanonický parametr rozdělení je roven lineárnímu prediktoru, $\theta = \eta$. Tyto spojovací funkce se nazývají kanonické (canonical link). Pro běžná rozdělení jsou kanonickými následující spojovací funkce:

normální rozdělení,$N(\mu, \sigma^2)$	$\eta = \mu$
Poissonovo,$P(\mu)$	$\eta = \ln \mu$
alternativní,$A(\pi)$	$\eta = \ln\{\pi/(1-\pi)\}$
gamma,$G(\mu, v)$	$\eta = \mu^{-1}$
inversní Gaussovo,$IG(\mu, \sigma^2)$	$\eta = \mu^{-2}$

V klasickém modelu se výstižnost modelu (těsnost proložení) vyjadřuje obvykle pomocí koeficientu determinace R^2, tj. jako podíl variability závislé veličiny vysvětlené modelem na celkové variabilitě.

$$R^2 = 1 - \frac{RSS}{\sum(y_i - \bar{y})^2}$$

Celková variabilita $\sum(y_i - \bar{y})^2$ odpovídá RSS lineárního modelu s jedním parametrem, jehož odhad $b_0 = \bar{y}$. Pro takový model je

$$R^2 = 1 - \frac{\sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 0.$$

Pro model vysvětlující variabilitu veličiny y úplně je $RSS = 0$ a tedy je $R^2 = 1$.

Ve zobecněném lineárním modelu lze model (těsnost proložení) posuzovat analogicky. Uvažujme tak zvaný úplný model s n parametry, který by vysvětloval pozorované hodnoty y přesně, tzn. $y_i = \mu_i$. Jelikož můžeme kanonický parametr vyjádřit jako funkci střední hodnoty, $\theta(\mu)$, můžeme věrohodnostní funkci zapsat jako $l(\mathbf{y}, \phi, \mathbf{y})$. To je maximálně dosažitelná hodnota věrohodnostní funkce.

Pro model jen s jedním parametrem, kdy $\mu_i = konst$ (nulový model, obsahuje jen intercept), bychom dostali věrohodnostní funkci minimální hodnoty pro daná data. Dvojnásobek rozdílu mezi těmito věrohodnostními funkcemi je jistou analogií k celkové variabilitě v klasickém modelu. Označíme-li odhad středních hodnot v modelu s p parametry jako $\hat{\boldsymbol{\mu}}$ a odhad kanonického parametru pro tento model jako $\hat{\theta} = \theta(\mu)$ a $\tilde{\theta} = \theta(\mathbf{y})$ a předpokládáme-li $a_i(\phi) = \phi/w_i$, pak dvojnásobek rozdílu věrohodnostních funkcí $l(\mathbf{y}, \phi, \mathbf{y})$ a $l(\hat{\mu}, \phi, \mathbf{y})$ je

$$\sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}/\phi = D(\mathbf{y}, \hat{\mu})/\phi.$$

$D(\mathbf{y}, \hat{\mu})/\phi$, která je funkcí pozorovaných dat, se nazývá deviance a je to analogie residuální sumy čtverců, RSS. Klasický lineární model je zvláštním případem zobecněného modelu, kdy spojovací funkce je identita a pak pro normálně rozdělenou náhodnou složku modelu je deviance rovna residuálnímu součtu čtverců, $D(\mathbf{y}, \hat{\mu})/\phi = \sum(y_i - \hat{\mu_i})^2 = RSS$

$D^*(\mathbf{y}, \hat{\mu}) = D(\mathbf{y}, \hat{\mu})/\phi$ je tzv. scaled deviance, je to deviance vyjádřená jako násobek disperzního parametru.

Ve zobecněném lineárním modelu je tedy cílem nalézt model, který zmenšuje celkovou devianci (úměrnou rozdílu logaritmů věrohodnostních funkcí mezi úplným modelem a nulovým modelem s jedním parametrem). Takový model může být vytvářen i postupně, mohou být do modelu zařazovány ty regresory, které nejvíce snižují devianci vzhledem k aktuálnímu modelu se zařazenými k parametry, tedy může být použit krokový (stepwise) postup pro vyhledávání regresního modelu. Regresory ve

zobecněném lineárním modelu mohou být i kvalitativní (faktory) a regresory mohou být i interakce (součiny) původních regresorů, takže pomocí zobecněného modelu je možné odhadovat parametry i složitých modelů analýzy rozptylu.

Logistická regrese

K logistickému regresnímu modelu dojdeme ze zobecněného lineárního modelu (GLIM)

$$g[E(Y|\mathbf{x})] = \mathbf{x}^T\boldsymbol{\beta}, \tag{32}$$

ve kterém nějaká funkce g podmíněné střední hodnoty náhodné veličiny Y je vyjádřena jako lineární funkce vektoru regresorů $\mathbf{x}^T = (1, x_1, x_2, \ldots, x_s)$ s regresními koeficienty $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_s)$. Pokud má náhodná veličina Y alternativní rozdělení, tedy $Y \sim A(p)$, které má, jak známo, střední hodnotu $E(Y) = p$, a jako spojovací (t. zv. link) funkci ve zobecněném lineárním modelu zvolíme *logit*,

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{33}$$

dojdeme k logistickému regresnímu modelu

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{x}^T\boldsymbol{\beta}, \tag{34}$$

ve kterém logit podmíněné střední hodnoty je vyjádřen jako lineární funkce regresorů.

Parametry $\beta_0, \beta_1, \ldots, \beta_s$ regresního modelu (3) lze odhadovat metodou maximální věrohodnosti. Algoritmy pro nalezení těchto odhadů $b_0, b_1, \ldots, b_s$ jsou již řadu let implementovány v dostupných statistických programech. Logistický regresní model má poměrně snadnou a přímočarou interpretaci. Poměr $p/(1-p)$, tedy poměr pravděpodobnosti „úspěchu“ ku pravděpodobnosti „neúspěchu“, je v anglosaském světě označován jako *odds* a je zcela samozřejmě používán i mimo statistiku, na př. při sázkách. Česká terminologie není ustálená, užívá se poměr šancí nebo sázkové riziko.

Nechť tedy

$$odds_0 = \frac{p_0}{1-p_0} \text{ při hodnotách regresorů } \mathbf{x} = \mathbf{x_0}$$

$$odds_1 = \frac{p_1}{1-p_1} \text{ při hodnotách regresorů } \mathbf{x} = \mathbf{x_1}$$

Poměr dvou *odds* je označován jako *odds ratio*, zkratkou *OR*.

$$OR = \frac{odds_1}{odds_0} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \tag{35}$$

Odhad regresního koeficientu b_i, $i \in [1, s]$, znamená odhad změny logitu při změně regresoru x_i o jedničku a při konstantních hodnotách regresorů ostatních, tedy

$$b_i = ln(\widehat{OR}), \text{ jestliže } x_{1,i} - x_{0,i} = 1 \text{ a } x_{1,j} = x_{0,j}, \quad j \neq i, \quad j = 1, 2, \ldots, s$$

Odhad OR při změně regresoru x_i o jedničku lze spočítat jednoduše jako

$$\widehat{OR} = e^{b_i}$$

Interpretaci výsledků logistické regrese ilustruje následující příklad nejjednoduššího logistického modelu s jedním dichotomickým regresorem. Pro větší názornost si představme, že regresor X znamená expozici (vystavení riziku), vysvětlovaná proměnná Y znamená přítomnost příznaku nemoci. Četnosti pozorovaných případů pak můžeme zapsat do čtyřpolní tabulky

Nemoc	Expozice	
	$X = 1$	$X = 0$
$Y = 1$	a	b
$Y = 0$	c	d

Pak, je-li $a, b, c, d > 0$

$$odds_1 = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}, \quad odds_0 = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

$$\widehat{OR} = \frac{ad}{bc}, \quad b_1 = \ln\left(\frac{ad}{bc}\right)$$

Pro rozptyl tohoto odhadu asymptoticky platí - viz na př. [5]

$$\text{var}(b_1) = \text{var}\left(\ln\left(\frac{ad}{bc}\right)\right) = 1/a + 1/b + 1/c + 1/d$$

Je tedy zřejmé, že logistickou regresi je možno aplikovat i v případech, kdy regresor je diskrétní dichotomická veličina. Pokud je regresor nominální, lze takovou proměnnou transformovat na dichotomické veličiny s hodnotami $\{0, 1\}$, tzv. indikátory (dummy variables). Uvažujme regresor $\mathbf{x}_i$, $i \in [1, s]$, který je nominální s k_i kategoriemi a má pozorované hodnoty $x_{li}, \quad l = 1, 2, \ldots, n$, n je počet pozorování. Hodnoty kategorií můžeme označit číselnými kódy $\{0, 1, \ldots, k_i - 1\}$. Kategorii s kódem 0 zvolíme jako referenční (t.zv. baseline category) a vytvoříme $k_i - 1$ indikátorů s ohodnocením podle následujícího pravidla

$$(d_{ij})_l = \begin{cases} 1 & \text{když} \quad x_{li} = j \\ & \qquad\qquad j = 1, 2, \ldots, k_i - 1, \quad l = 1, 2, \ldots, s \\ 0 & \text{jinak} \end{cases} \tag{36}$$

Regresní koeficienty korespondující s těmito indikátory můžeme označit β_{ij}, $j = 1, 2, \ldots, k_i - 1, \quad i = 1, 2, \ldots, s$, jejich odhady pak označíme b_{ij}. Odhady regresních koeficientů u jednotlivých indikátorů jsou vlastně logaritmem odhadovaného poměru *odds* příslušné kategorie k *odds* kategorie referenční, tedy logaritmem příslušného *odds ratio*. Pro velké výběry můžeme $100(1-\alpha)$-procentní oboustranný interval spolehlivosti pro regresní koeficient β_{ij} vyjádřit jako

$$\langle b_{ij} - u(1-\alpha/2)SE(b_{ij}), \quad b_{ij} + u(1-\alpha/2)SE(b_{ij}\rangle$$

a interval spolehlivosti pro *OR*

$$\langle \exp\left[b_{ij} - u(1-\alpha/2)SE(b_{ij})\right], \quad \exp\left[(b_{ij} + u(1-\alpha/2)SE(b_{ij})\right]\rangle, \tag{37}$$

kde $u(1-\alpha/2)$ je kvantil normovaného normálního rozdělení $N(0,1)$ a $SE(b_{ij})$ je směrodatná odchylka odhadu regresního koeficientu. Neobsahuje-li interval spolehlivosti pro *OR* jedničku, lze *odds* v této kategorii považovat za odlišný od *odds* kategorie referenční, takže interpretace výsledků regresního modelu je velice přímočará. Pokud máme regresní model s více regresory, odhad regresního parametru vyjadřuje lineární závislost predikované veličiny na daném regresoru po adjustování vlivu ostatních regresorů. Tedy v logistické regresi je odhad regresního koeficientu roven logaritmu odhadovaného *odds ratio* po adjustaci vlivu ostatních regresorů.

Příklad 10.1 Data pro tuto úlohu jsou v souboru LOGREG2.XLS. Vysvětlovaná veličina Y je dichotomická s hodnotami $\{0,\ 1\}$. Hodnota 1 znamená, že pozorovaná osoba je nemocná, hodnotu 0 má osoba zdravá. Regresory jsou veličiny *expozice* (dichotomická, hodnota 1 znamená, že osoba pracuje v rizikovém provozu, hodnota 0 znamená opak), *vek* (roky) a *koureni* (počet cigaret za den) jsou spojité, resp. i počet cigaret za spojitý můžeme považovat.

Zkrácený výstup z modulu Logistic Regression [14] následuje:

```
Logistic Regression Report

Response       Y

Parameter Estimation Section
           Regression        Standard  Chi-Square  Prob
Variable   Coefficient       Error       Beta=0    Level
Intercept  -25.35205         5.291554     22.95    0.000002
expozice     2.285141        0.4990213    20.97    0.000005
vek          0.6799906       0.1548485    19.28    0.000011
koureni      5.641818E-02    1.658742E-02 11.57    0.000671
```

```
Model in Transformation Form
-25.35205 + 2.285141*expozice + .6799906*vek +
+ 5.641818E-02*koureni
Note that this is XB. Prob(Y=1) is 1/(1+Exp(-XB)).

Odds Ratio Estimation Section
            Regression      Odds     Lower 95%   Upper 95%
Variable    Coefficient     Ratio    Conf.Limit  Conf.Limit
Intercept   -25.352054
expozice      2.285141    9.827076   3.695359    26.133163
vek           0.679991
koureni       0.056418

Model Summary Section
Model       Model      Model          Model
R-Squared    D.F.      Chi-Square     Prob
0.345596      3        77.63        0.000000
```

V první části jsou odhady parametrů logistického modelu a statistiky pro test hypotéz o nulovosti parametrů. Vidíme, že u všech čtyřech parametrů zamítáme nulovou hypotézu $\beta_j = 0$, odhadované parametry u všech tří regresorů jsou kladné, tzn. logit roste s hodnotou regresoru, je tedy vyšší u exponovaných, roste s věkem a počtem vykouřených cigaret. V části Odds Ratio Estimation jsou znovu uvedeny odhady parametrů a pro dichotomický regresor je uveden i $\widehat{OR}$ a 95%-ní interval spolehlivosti. Jelikož tento interval neobsahuje hodnotu 1 (dolní hranice intervalu je 3,7), znamená to, že $\widehat{OR}$ je významně větší než 1 i po odečtení (adjustaci) vlivu věku a kouření a že expozice významně zvyšuje riziko onemocnění. Model Summary Section je analogií sekce ANOVA v lineární regresi a slouží k testu hypotézy, že všechny parametry jsou nulové.

Shrnutí

- zobecněný lineární model (GLIM)
- spojovací funkce, kanonické spojovací funkce pro běžná rozdělení
- logistická regrese, odds ratio

Kontrolní otázky

1. Vysvětlete hlavní myšlenky zobecněného modelu.
2. Co je lineární prediktor?
3. Jaké rozdělení má vysvětlovaná veličina v logistické regresi?
4. Co je to logit? Je funkcí střední hodnoty vysvětlované veličiny?

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

11 Nelineární regresní model

Průvodce studiem

Kapitola je věnována základům nelineární regrese. Na tuto kapitolu počítejte nejméně se třemi hodinami studia. Prostudujte důkladně i řešený příklad na konci kapitoly.

Základní představa pro nelineární regresní model je, že střední hodnoty složek náhodného vektoru $\mathbf{y}$, tj. $E(y_i)$, $i = 1, 2, \ldots, n$, můžeme vyjádřit jako nějakou funkci regresorů

$$Ey_i = f(\mathbf{x}_i, \boldsymbol{\beta}), \tag{38}$$

kde $\mathbf{x}_i$ je k-členný vektor nenáhodných vysvětlujících proměnných ($\mathbf{x}_i^T$ je i-tý řádek matice regresorů $\mathbf{X}$, matice $\mathbf{X}$ je typu $n \times k$) a $\boldsymbol{\beta}$ je p-členný vektor parametrů.

Obvyklou základní úlohou nelineární regrese je pro daná data $[\mathbf{y}, \mathbf{X}]$ a daný tvar funkce $f(\mathbf{x}_i, \boldsymbol{\beta})$ odhadnout hodnoty parametrů $\boldsymbol{\beta}$ tak, aby model (38) co nejlépe vysvětloval pozorované hodnoty náhodného vektoru $[\mathbf{y}$.

Zda je model (38) opravdu nelineární v parametrech poznáme podle parciálních derivací

$$g_j = \frac{\partial f(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_j}$$

Pokud

$$g_j = \text{const} \quad \text{pro všechna } j = 1, 2, \ldots, p \tag{39}$$

tzn. parciální derivace nejsou závislé na β_j, pak model (38) je lineární v parametrech, pokud alespoň pro jeden z parametrů β_j podmínka (39) neplatí, je model nelineární. Pokud podmínka (39) není splněna pro žádný z parametrů modelu, říkáme, že model je *neseparabilní*. To je např. následující model s jedním regresorem ($k = 1$)

$$f(x_i, \boldsymbol{\beta}) = \exp(\beta_1 x_i) + \exp(\beta_2 x_i).$$

Pokud pro některé parametry je podmínka (39) splněna, je model *separabilní*, např.

$$f(x_i, \boldsymbol{\beta}) = \beta_1 + \beta_2 \exp(\beta_3 x_i).$$

To je model, který je nelineární jen vzhledem k parametru β_3. Některý tvar nelineárních modelů (38) můžeme vhodnou transformací linearizovat, např.

$$f(x_i, \boldsymbol{\beta}) = \beta_1 \exp(\frac{\beta_2}{x_i})$$

po zlogaritmování přejde na tvar

$$\ln(Ey_i) = \ln[f(x_i, \boldsymbol{\beta})] = \gamma_1 + \gamma_2 z_i,$$

což je lineární funkce proměnné $z_i = 1/x_i$ a parametry $\gamma_1 = \ln \beta_1$ a $\gamma_2 = \beta_2$.

Data $[\mathbf{y}, \mathbf{X}]$, tj. k+1 rozměrný výběr o rozsahu n jsou n-tice bodů v $(k+1)$-rozměrném prostoru. Nelineární regresní model (38) je plocha v k-rozměrném prostoru. Na rozdíl od lineárního modelu tuto plochu nelze vyjářit jako lineární kombinaci regresorů („není rovná“), má zakřivení. Při daných datech a modelové funkci je tvar této plochy závislý na hodnotách parametrů β_j. Úlohou odhadu parametrů nelineárního regresního modelu je tedy nalézt takové hodnoty parametrů, pro které tato plocha dobře aproximuje pozorované hodnoty náhodného vektoru $\mathbf{y}$. Podobně jako u lineární regrese můžeme tuto úlohu řešit *metodou nejmenších čtverců*.

Uvažujme tzv. *aditivní model* pro náhodnou složku

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i \tag{40}$$

Předpokládejme, že pro $i = 1, 2, \ldots, n$

- ε_i jsou vzájemně nezávislé náhodné veličiny (nejsou korelovány),
- $E(\varepsilon_i) = 0$, hodnoty y_i náhodně kolísají okolo prokládané plochy, střední hodnota tohoto kolísání je nulová,
- var $\varepsilon_i = \sigma^2$, tzn. mají konstantní rozptyl σ^2

Potom součet čtverců rozdílů pozorovaných a modelových hodnot vyjádřený jako funkce parametrů je

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \boldsymbol{\beta})]^2 \tag{41}$$

Odhad metodou nejmenších čtverců znamená nalézt takové odhady $\hat{\boldsymbol{\beta}}$ parametrů $\boldsymbol{\beta}$, aby $Q(\boldsymbol{\beta})$ bylo minimální. Zderivujeme-li $Q(\boldsymbol{\beta})$ podle β_j, $j = 1, 2, \ldots, p$ a položíme-li derivace rovny nule, dostaneme soustavu p rovnic

$$\sum_{i=1}^{n} \left[y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \right] \frac{\partial f(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_j}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \tag{42}$$

Na rozdíl od lineární regresní funkce, kdy je soustava normálních rovnic lineární (parciální derivace jsou konstantní) a $Q(\boldsymbol{\beta})$ eliptický paraboloid s minimem v bodě $\hat{\boldsymbol{\beta}} = \mathbf{b}$, které lze za dosti obecných podmínek jednoznačně určit, je u nelineárního modelu soustava normálních rovnic nelineární. Její řešení je náročné, neboť nemusí vždy existovat jednoznačně, navíc je nutno užívat iterativních metod, které nezaručují nalezení globálního minima funkce (41).

U nelineárních modelů lze získat informace o tvaru funkce $Q(\boldsymbol{\beta})$ v okolí bodu $\boldsymbol{\beta_k}$ z jejího Taylorova rozvoje druhého stupně:

$$Q(\boldsymbol{\beta}) \cong Q(\boldsymbol{\beta}_k) + \Delta\boldsymbol{\beta}_k \mathbf{g}_k + \frac{1}{2} \Delta\boldsymbol{\beta}_k^T \mathbf{H}_k \Delta\boldsymbol{\beta}_k, \tag{43}$$

kde $\Delta\boldsymbol{\beta}_k = \boldsymbol{\beta} - \boldsymbol{\beta}_k$, $\mathbf{g}_k$ je gradient (vektor)s prvky

$$g_j = \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_k}$$

a $\mathbf{H}_k$ je symetrická matrice řádu p (Hessián) s prvky

$$H_{ij} = \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_k}$$

Gradient lze vyjádřit

$$\mathbf{g}_k = -2\mathbf{J}^T\mathbf{d},$$

kde $\mathbf{d}$ je vektor typu $(n \times 1)$ s prvky $d_i = y_i - f(\mathbf{x_i}, \boldsymbol{\beta}_k)$, $\mathbf{J}$ je matice typu $(n \times p)$ (Jakobián) s prvky

$$J_{ij} = \frac{\partial f(\mathbf{x}_i, \boldsymbol{\beta}_k)}{\partial \beta_j}, \quad i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, p.$$

Pak pro Hessián platí:

$$\mathbf{H}_k = 2\mathbf{J}^T\mathbf{J} + \mathbf{B}_k,$$

kde $\mathbf{B}_k$ (řádu p) má prvky:

$$B_{jl} = -2\sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \beta_k)] \frac{\partial^2 f(\mathbf{x}_i, \boldsymbol{\beta}_k)}{\partial \beta_j \partial \beta_l}, \quad j, l = 1, 2, \ldots, p.$$

Regresní parametry lze odhadnout jednoznačně jen tehdy, jsou-li

$$\frac{\partial f(\mathbf{x}_i, \boldsymbol{\beta_k})}{\partial \beta_j}$$

lineárně nezávislé. To znamená, že neexistují $c_j \neq 0$, aby

$$\sum_{j=1}^{p} c_j \frac{\partial f(\mathbf{x}_i, \boldsymbol{\beta_k})}{\partial \beta_j} = 0 \tag{44}$$

Pokud platí (44), tj. $\mathbf{J}^T\mathbf{J}$ je singulární, pak je model nevhodně specifikován a nelze odhadnout jednotlivá β_j. Říkáme, že model je *přeurčen*. Jediná cesta k nápravě je změna modelu, která většinou vede přes snížení počtu parametrů, tj. zjednodušení funkce $f(\mathbf{x}_i, \boldsymbol{\beta})$. Pokud rov. (44) platí přibližně (analogie s multikolinearitou v lineárním modelu), pak jsou odhady $\hat{\beta}_j$ silně korelované a jejich odhad není spolehlivý.

Numerické metody odhadu regresních parametrů, užívané ve statistickém software, jsou většinou algoritmy, které minimalizují $Q(\boldsymbol{\beta})$ iterativně. Vycházejí z počátečního, uživatelem zadaného „nástřelu“ hodnot parametrů. Řešení probíhá po krocích

$$\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(H)}.$$

Změnu mezi jednotlivými iteračními kroky můžeme vyjádřit jako přičtení přírůstkového vektoru $\boldsymbol{\Delta}_h$

$$\boldsymbol{\beta}_{(h+1)} = \boldsymbol{\beta}_{(h)} + \boldsymbol{\Delta}_h,$$

při čemž iterační proces by měl splňovat podmínku:

$$Q(\boldsymbol{\beta}_{(h+1)}) < Q(\boldsymbol{\beta}_h)$$

Přírůstkový vektor můžeme vyjádřit jako součin tzv. směrového vektoru $\mathbf{v}_h$ a koeficientu α_h,

$$\boldsymbol{\Delta}_h = \alpha_h \mathbf{v}_h$$

Jednotlivé algoritmy se liší ve volbě směrového vektoru $\mathbf{v}_h$ a způsobu adaptace koeficientu α_h během výpočtu.

Z rovnice (43) dostaneme směrový vektor ve tvaru

$$\mathbf{v}_h = -\mathbf{H}^{-1}\mathbf{g} = (\mathbf{J}^T\mathbf{J} - \mathbf{B})^{-1}\mathbf{J}^T\hat{\mathbf{e}}, \tag{45}$$

kde $\hat{\mathbf{e}}$ je vektor residuí. Optimální je $\alpha = 1$. Tato metoda se nazývá Newtonova. Její nevýhodou je, že potřebuje výpočet matice druhých derivací kriteriální funkce $Q(\boldsymbol{\beta})$. V metodě Gauss-Newtonově se zanedbává matice $\mathbf{B}$ a směrový vektor se určuje ze vztahu

$$\mathbf{v}_h = (\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T\hat{\mathbf{e}}, \tag{46}$$

Dalšími běžně užívanými metodami jsou různé modifikace Marquardtovy metody, kdy směrový vektor se určuje podle vztahu

$$\mathbf{v}_h = (\mathbf{J}^T\mathbf{J} + \lambda\mathbf{D}_h^T\mathbf{D}_h)^{-1}\mathbf{J}^T\hat{\mathbf{e}}, \tag{47}$$

kde $\mathbf{D}_h$ je diagonální matice eliminující vliv různých velikostí složek matice $\mathbf{J}$ a λ je parametr, jehož hodnota se adaptuje během iteračního procesu. Podrobněji viz např. Meloun a Militký [20]. Jedna z variant této metody je implementována i v NCSS 2000 [14].

Pro všechny iterativní algoritmy je však velmi podstatný uživatelem zadaný „nástřel“ počátečních hodnot parametrů a jejich minima a maxima. Špatná volba počátečních hodnot může způsobit buď pomalou konvergenci, případně ukončení v nějakém lokálním minimu, nebo dokonce úplné selhání algoritmu. Vždy se vyplatí obor hod-

not parametrů věcně i numericky předem důkladně analyzovat a teprve pak spustit výpočet. Pokud standardní iterativní algoritmy selhávají, je k odhadu parametrů nelineárního regresního model užít některý ze stochastických algoritmů globální optimalizace, viz např. [**?**].

Odhadneme-li parametry nelineárního regresního modelu, můžeme (a většinou i musíme, neboť správnost odhadu, tj. nalezení globálního minima funkce $Q(\boldsymbol{\beta})$ není zaručena) posoudit vhodnost modelu a správnost nalezených odhadů. K prvnímu hrubému posouzení poslouží index determinace R^2,

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}},$$

kde $\mathrm{RSS} = Q(\hat{\boldsymbol{\beta}})$, celková suma čtverců TSS je definována stejně jako u lineárního modelu. Dalšími užitečnými jednoduchými charakteristikami je odhad residuálního rozptylu

$$\hat{\sigma^2} = s^2 = \frac{\mathrm{RSS}}{n-p},$$

případně odhad směrodatné odchylky residuí, která je odmocninou residuálního rozptylu.

Dále je vždy potřeba posoudit, zda regresní funkce s nalezenými hodnotami parametrů dobře vystihuje pozorované veličiny vysvětlované veličiny $\mathbf{y}$. Pokud máme v modelu jen jeden regresor, většinou postačí jen vizuální posouzení grafu nalezené funkce a hodnot $\mathbf{y}$ proti hodnotám regresoru. Pokud je regresorů více, můžeme užít grafy residuí podobně jako je popsáno v kapitole o lineárním regresním modelu.

Je také vhodné posoudit, jak silně jsou odhady parametrů korelovány. Kovarianční matice odhadů se obvykle nejjednodušeji aproximuje

$$\mathbf{S_b} = s^2\ (\mathbf{J}^T\mathbf{J})^{-1},$$

po případě přesnějšími aproximacemi s využitím i druhých derivací kriteriální funkce $Q(\boldsymbol{\beta})$ v bodě $\hat{\boldsymbol{\beta}}$, tj. s využitím matice $\mathbf{H}$. V matici $\mathbf{S_b}$ jsou na diagonále odhady rozptylů jednotlivých odhadů parametrů, takže lze pak snadno z kovarianční matice $\mathbf{S_b}$ i korelační matici. Pokud se absolutní hodnota některého z korelačních koeficientů blíží jedné, je model buď přeurčený nebo špatně podmíněný. Pak je potřeba celý problém znovu analyzovat a buď zjednodušit model nebo doměřit další data.

Za předpokladu, že v modelu (40) mají náhodné složky normální rozdělení, tj. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, lze pak spočítat i intervalové odhady pro parametry a testovat hypotézy H0: $\beta_j = 0,\ j = 1, 2, \ldots, p$. Jak intervaly spolehlivosti, tak testy je však nutno užívat s jistou opatrností, neboť v případě nelineární regrese odhady parametrů získané metodou nejmenších čtverců obecně nemusí být nestranné a intervaly spolehlivosti i

testy hypotéz jsou založeny jen na aproximaci tvaru funkce $Q(\boldsymbol{\beta})$ v okolí nalezených hodnot odhadů $\hat{\boldsymbol{\beta}}$.

Příklad 11.1 Data v souboru NLR1.XLS [14] obsahují 44 pozorování veličin X a Y. Regresní funkce má tvar $Y = A + (0.49 - A) * \exp(-B * X)$. Máme odhadnout parametry A, B a posoudit vhodnost modelu.

Výstup z modulu nonlinear regression [14] je následující:

```
Nonlinear Regression Report

Dependent    Y

Minimization Phase Section

Itn    Error Sum
No.    Lambda          Lambda           A           B
0      3.58328         0.00004         0.1         0.13
1      2.387233E-02    0.000016        0.3856243   7.893059E-02
Stepsize reduced to 0.9189852 by bounds.
Stepsize reduced to 0.9202212 by bounds.
Stepsize reduced to 0.9325328 by bounds.
2      1.453589E-02    0.064          0.3846167    3.805048E-02
3      9.619339E-03    0.0256         0.3635308    3.545998E-02
4      9.364404E-03    0.01024        0.3440655    2.731663E-02
5      8.930465E-03    0.004096       0.3211968    2.272922E-02
6      8.764675E-03    0.0016384      0.3023451    2.005132E-02
7      8.722906E-03    6.5536E-04     0.2922916    0.0189681
8      8.720914E-03    2.62144E-04    0.2899626    1.874566E-02
9      8.720909E-03    1.048576E-04   0.2898918    1.874063E-02
10     8.720909E-03    4.194304E-05   0.2898947    0.018741
Convergence criterion met.
```

Na výpisu vidíme, jak z počátečních hodnot parametrů ($A = 0.1$, $B = 0.13$) postupuje iterativní proces hledání minima residuální sumy čtverců.

```
Model Estimation Section

Parameter  Parameter   Asymptotic       Lower        Upper
Name       Estimate    Standard Error   95% C.L.     95% C.L.
A          0.2898947   6.939709E-02    0.1498457    0.4299437
B          0.018741    8.529059E-03    1.528661E-03 3.595333E-02
```

Z odhadů parametrů a jejich intervalů spolehlivosti vidíme, že odhady jsou významně odlišné od nuly.

```
Model    Y = A+(0.49-A)*EXP(-B*X)
R-Squared      0.779217
Iterations     10
Estimated Model
(.2898947)+(0.49-(.2898947))*EXP(-(.018741)*(X))
```

Index determinace ukazuje, že model vysvětluje asi tři čtvrtiny z celkové variability veličiny Y.

```
Analysis of Variance Table
                     Sum of         Mean
Source          DF   Squares       Square
Mean             1   7.9475        7.9475
Model            2   7.978279      3.98914
Model (Adj)      1   3.077909E-02  3.077909E-02
Error           42   8.720909E-03  2.076407E-04
Total (Adj)     43   0.0395
Total           44   7.987
```

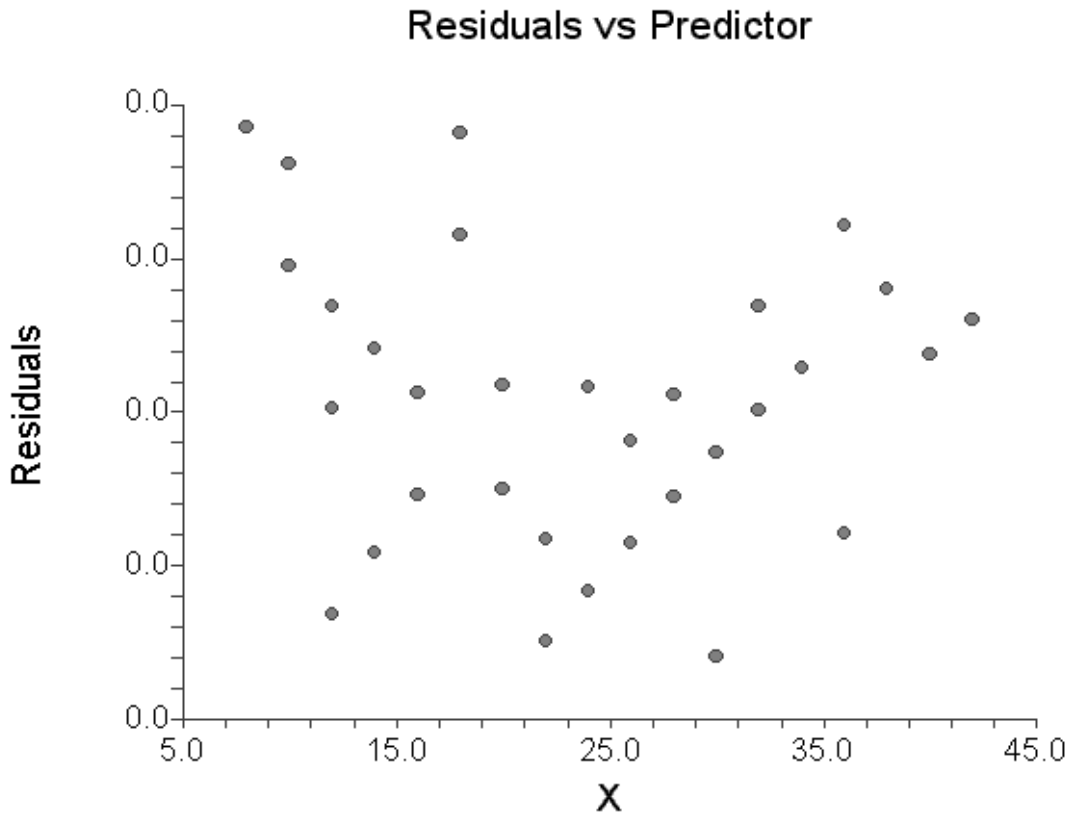
Tabulka analýzy rozptylu má podobný účel, jako u lineárního modelu, můžeme zamítnout hypotézu, že vektor parametrů je nulový.

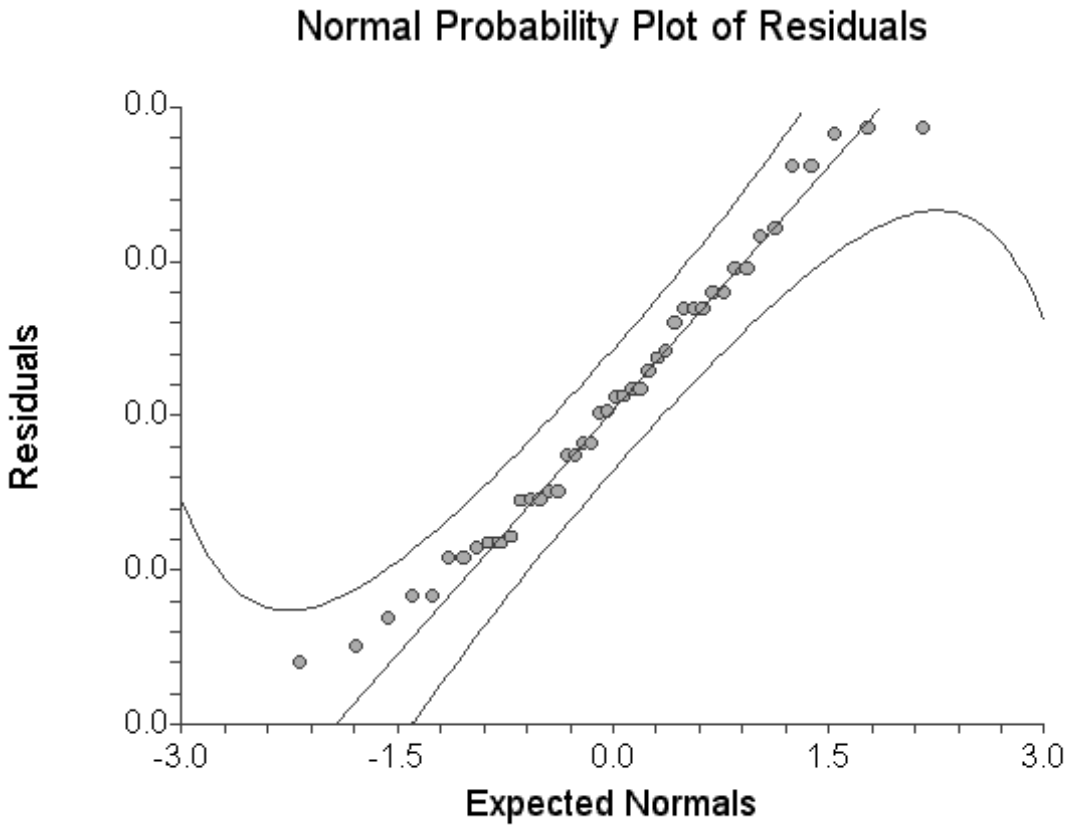
```
Asymptotic Correlation Matrix of Parameters

      A           B
A    1.000000    0.996008
B    0.996008    1.000000
```

Z korelační matice odhadů je zřejmé, že odhady jsou silně korelovány. To je u nelineárních modelů tohoto typu (jeden parametr je v součiniteli výrazu, ve kterém je druhý parametr v exponentu) častý jev. Indikuje, že minimalizovaná účelová funkce

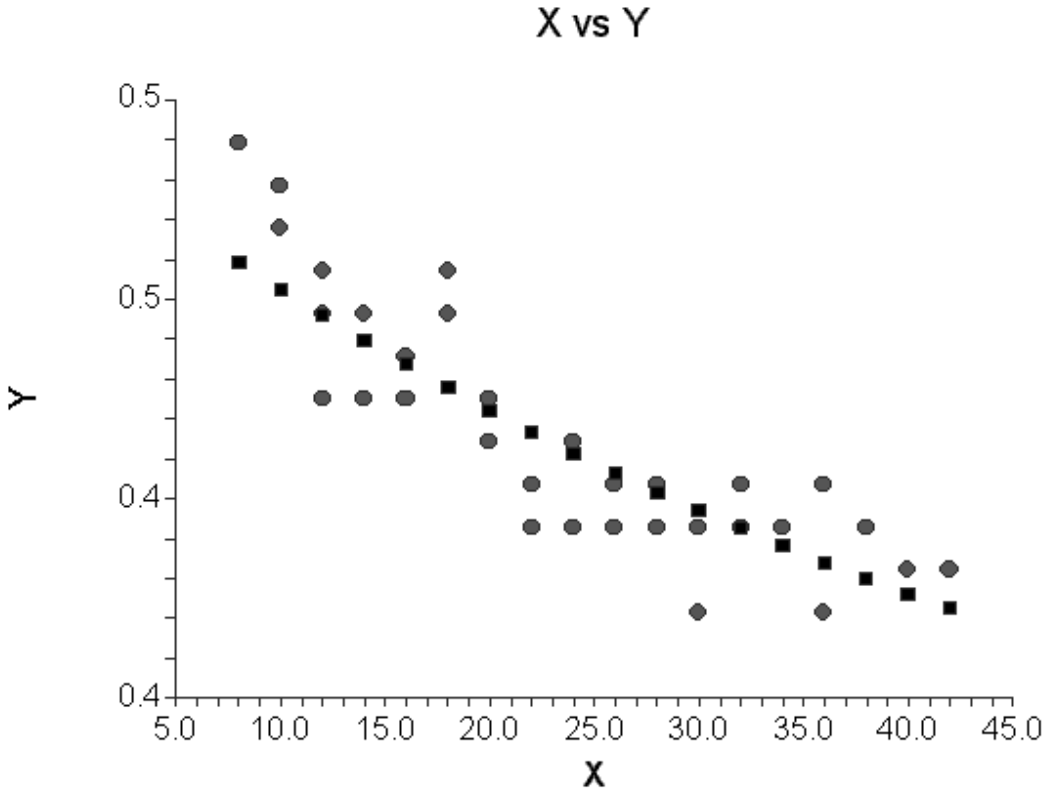
(součet residuálních čtverců) je v okolí nalezeného minima málo zakřivená a změna v hodnotách odhadů nezpůsobuje dramatickou změnu v hodnotě minimalizované funkce.

Z diagnostických grafů vidíme, že předpoklady modelu jsou zhruba splněny (rozptyl residuí můžeme považovat za konstantní, residua jsou zhruba normálně rozdělená.



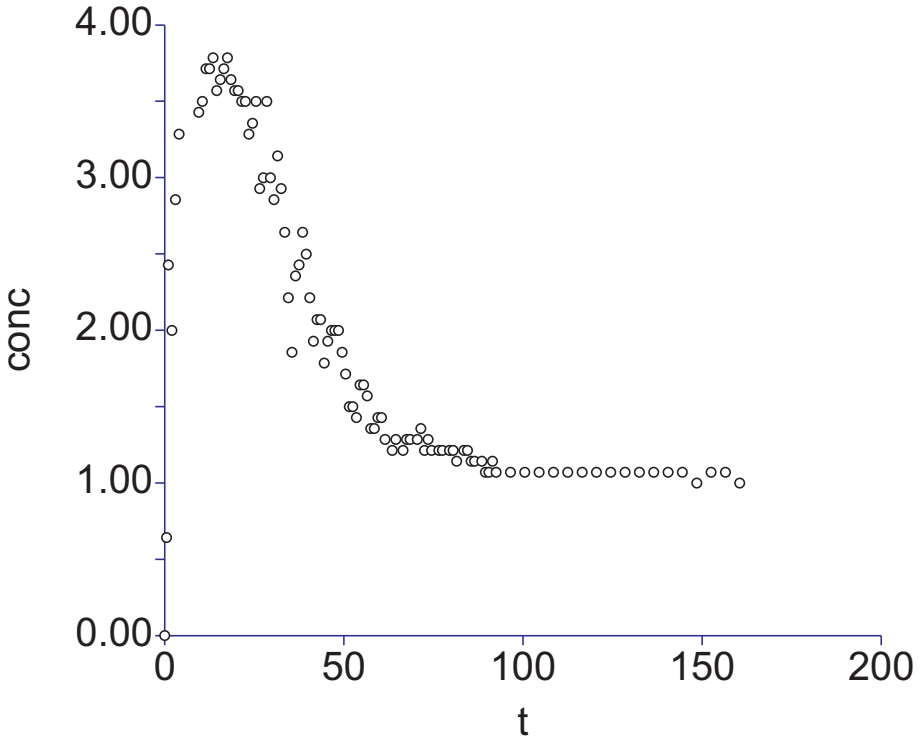


Nalezený model ukazuje následující graf. Pozorované hodnoty jsou vyznačeny kroužky, odhadované (modelové) hodnoty čtverečky.



Z grafu je zřejmé, že model dobře prokládá pozorované hodnoty. Můžeme tedy uzavřít, že navržený model je pro tuto závislost vhodný, index determinace je 0,78, odhadovaná residuální odchylka je 0,0144, odhady parametrů jsou výše.

Příklad 11.2 Na tomto příkladu si ukážeme postup při řešení jedné praktické úlohy, ve které je potřeba proložit nějakou vhodnou funkci naměřenými daty. Tato úloha je převzata od Ing. Morávky z firmy Třinecký inženýring, a.s. Data pro tento příklad jsou v souboru moravka_red_prikl.xls. Měřila se závislost koncentrace (conc) na čase (t). Zjištěná empirická závislost je nakreslena na následujícím obrázku:



Požadavek byl proložit tuto závislost funkcí, která bude mít následující vlastnosti:

- v čase $t = 0$ má mít koncentrace hodnotu rovnou nule $f(0) = 0$
- v čase $t \to \infty$ má mít koncentrace hodnotu rovnou jedné
- funkce má dobře prokládat empirickou závislost

Z grafu závislosti vidíme, že počáteční (rostoucí) část závislosti i za ní následující prudký pokles bychom mohli vystihnout součinem dvou funkcí:

- nějaké závislosti procházející počátkem, např. ve tvaru $y = A\,t$, kde A je nějaký neznámý parametr, jehož hodnotu pak odhadneme z dat, nebo funkcí s dvěma parametry $y_1 = A\,t^B$, která umožní rychlost růstu aproximovat pružněji.
- exponenciální funkce ve tvaru $y_2 = \exp(C\ t)$, která může dobře prokládat klesající část závislosti koncentrace na čase, C je opět parametr modelu.

Pak nám ještě zbývá vyřešit to, aby s rostoucím časem se funkce blížila hodnotě jedna. To znamená, že k výše uvedenému součinu potřebujeme přičíst nějakou funkci

y_3, která má malé hodnoty při malých hodnotách t a $\lim y_3(t) = 1$ pro $t \to \infty$. Tomuto požadavku vyhovuje např. funkce ve tvaru

$$y_3 = 1 - \frac{1}{\exp(D\ t)}$$

Pro lepší pochopení těchto úvah si zkuste nakreslit průběhy funkcí y_1, y_2, y_3) pro různé hodnoty jejich parametrů a $0 \leq t \leq 100$ pomocí nějakého software, můžete třeba užít možnost „function plot“ v nabídce „Graphics“ v NCSS.

Empirickou závislost tedy můžeme zkusit modelovat funkcí

$$f(t) = y_1 \times y_2 + y_3 = A\ t^B\ \exp(C\ t) + 1 - \frac{1}{\exp(D\ t)},$$

kde A, B, C, D jsou čtyři neznámé parametry, jejichž hodnoty odhadneme z dat metodou nejmenších čtverců. K tomu můžeme využít standardní statistický software, např. NCSS.

Víme už, že pro úspěšný odhad parametrů nelineárního modelu iterativními metodami je velmi důležité vhodně volit jejich počáteční „nástřely“ a dovolený obor hodnot, tj. minimum a maximum. Při tom musíme vzít v úvahu jak tvar funkcí, tak i obor hodnot nezávisle proměnné, tj. času t.

U parametru A je to celkem snadná úloha: jelikož hodnoty koncentrace jsou nezáporné, i hodnota A musí být nezáporná. Hodnoty koncentrace jsou menší než 4, takže maximální hodnota parametru A nemůže být příliš velká, interval $[0,\ 20]$ je postačující, startovací hodnota $A = 1$ může být dobrá volba.

U parametru B musíme uvážit, jak rychle má funkce růst. Pokud by $B = 1$, pak by růst podle y_2 byl lineární, pro $B < 1$ pomalejší, pro $B > 1$ rychlejší. Můžeme zkusit počáteční hodnotu $B = 1$ a dovolené hodnoty z intervalu $[0,\ 2]$.

U parametrů C a D musíme být opatrnější. Oba parametry jsou v exponentu, navíc v součinu s veličinou t, která má hodnoty z intervalu $[0,\ 160]$, tzn. hrozí numerické problémy – přetečení největší v počítači reprezentované hodnoty čísla v pohyblivé čárce. Je zřejmé, že hodnoty C a D musí být kladné, ale malé. Tedy zkusíme nástřely $C = 0.01$ a $D = 0.01$ a interval pro oba parametry $[0,\ 0.1]$ a zjistíme, že v průběhu iterací došlo k přetečení (overflow). Změníme-li nástřel na $D = 0.001$, iterační proces konverguje po 25 krocích a dostaneme následující výsledky:

```
Database D:\UZIVATELE\Moravka\moravka_prikl.S0
Dependent conc

Model Estimation Section
```

Parameter Name	Parameter Estimate	Asymptotic Standard Error	Lower 95% C.L.	Upper 95% C.L.
A	1.685585	9.694253E-02	1.493156	1.878015
B	0.5552467	3.312053E-02	0.489503	0.6209905
C	5.248904E-02	3.513403E-03	4.551499E-02	5.946309E-02
D	2.545075E-02	7.617775E-03	1.032959E-02	4.057191E-02

Model conc = (A*T^B)*EXP(-C*T)+1-1/EXP(D*T)
R-Squared 0.963863
Iterations 25
Estimated Model
((1.685585)*(T)^(.5552467))*EXP(-(5.248904E-02)*(T))
+1-1/EXP((2.545075E-02)*(T))

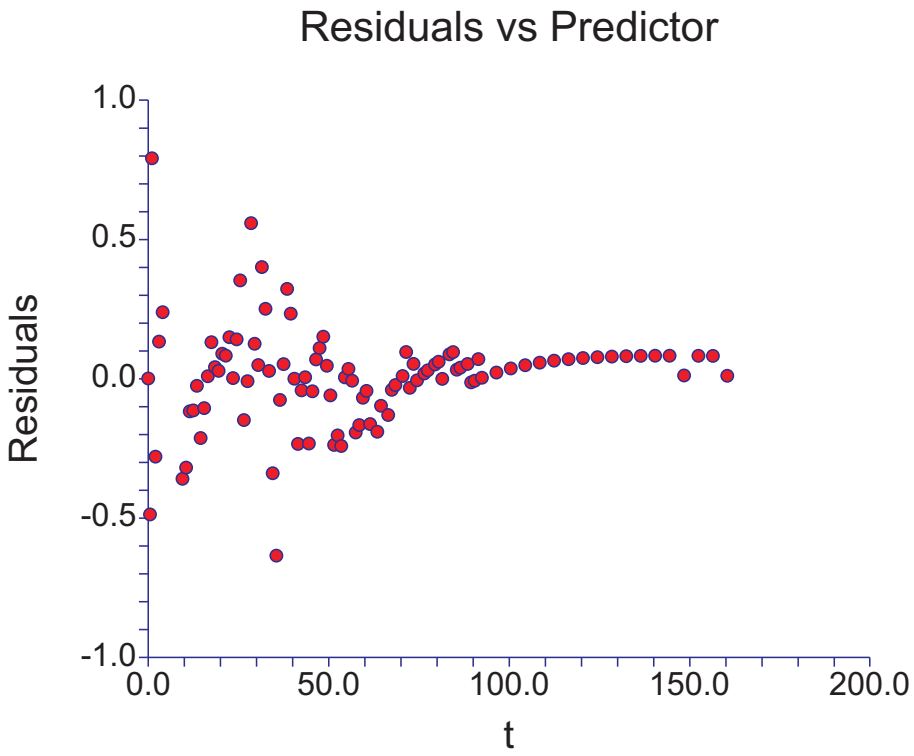
Analysis of Variance Table

Source	Sum of DF	Mean Squares	Square
Mean	1	378.8356	378.8356
Model	4	469.3829	117.3457
Model (Adjusted)	3	90.54733	30.18244
Error	96	3.394817	3.536267E-02
Total (Adjusted)	99	93.94215	
Total	100	472.7778	

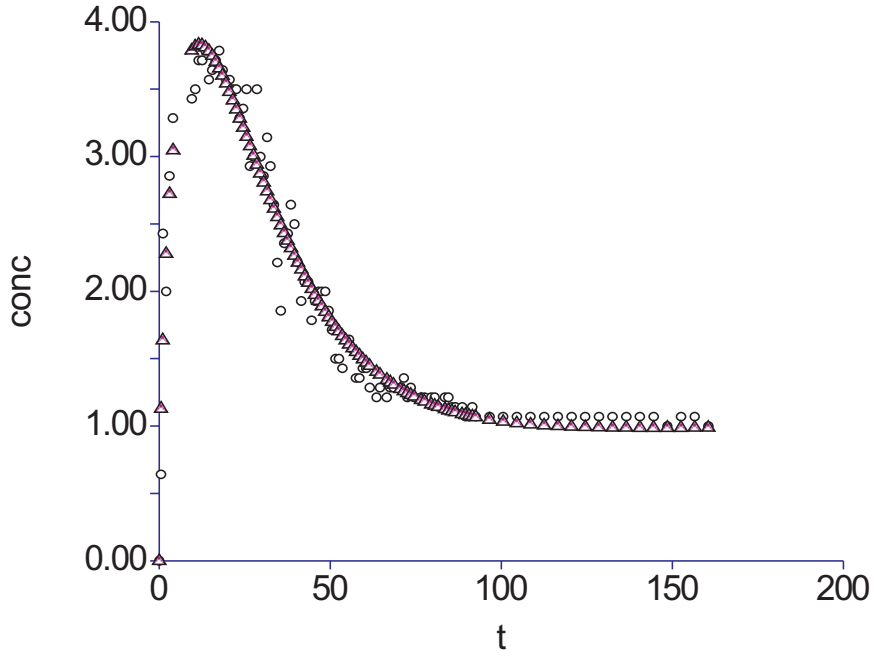
Asymptotic Correlation Matrix of Parameters

	A	B	C	D
A	1.000000	-0.913029	-0.572957	-0.339223
B	-0.913029	1.000000	0.810758	0.531945
C	-0.572957	0.810758	1.000000	0.885083
D	-0.339223	0.531945	0.885083	1.000000

Vidíme, že index determinace je přibližně 0,964, tedy model dobře prokládá empirickou závislost, graf reziduí na následujícím obrázku ukazuje, že větší odchylka modelových hodnot od pozorovaných se vyskytuje především pro malé hodnoty t, kde i z grafu empirické závislosti je patrné, že v této oblasti byla přesnost měření nejmenší.



Vhodnost zvoleného modelu ukazuje i následující obrázek, kde je kromě empirické závislosti nakresleny i modelové hodnoty (body vyznačené trojúhelníky).



Shrnutí

- *nelineární regresní model, metoda nejmenších čtverců*
- *aproximace tvaru kriteriální funkce v okolí nalezeného minima*
- *metody odhadu parametrů*

Kontrolní otázky

1. *Vysvětlete hlavní rozdíly mezi lineárním a nelineárním regresním modelem.*
2. *V čem je nalezení odhadů parametrů modelu obtížné?*
3. *Vysvětlete principy algoritmů pro odhad parametrů nelineárních regresních modelů?*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

12 Mnohorozměrné metody

Průvodce studiem

Na tuto rozsáhlou kapitolu počítejte nejméně s deseti hodinami usilovného studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí. Věnujte pozornost řešeným příkladům.

Dosud jsme se zabývali regresí, kdy jedna náhodná veličina je vysvětlována (nebo predikována) pomocí jiných veličin. Hledá se závislost podmíněné střední hodnoty náhodné veličiny na regresorech. Je to nejčastěji aplikovaná statistická metoda. Odhaduje se, že více jak 90% aplikací statistiky se opírá o regresi. Pochopení principů regrese je velmi užitečné pro pochopení ostatních metod analýzy mnohorozměrných dat. Regresní analýza bývá považována za zcela samostanou část stojící vedle mnohorozměrných metod (methods of multivariate analysis). Ve většině statistického software je regresní a korelační analýza uváděna jako samostatná položka stojící vedle mnohorozměrných metod. Také učebnice a monografie bývají věnovány samostatně regresi a samostatně zbývajícím metodám mnohorozměrné analýzy dat.

Mezi mnohorozměrné metody jsou zařazovány především:

- testy shody vektorů středních hodnot
  MANOVA (multivariate analysis of variance) - mnohorozměrná analogie analýzy rozptylu
- kanonické korelace, které můžeme považovat za jisté zobecnění lineární regrese, kdy vysvětlujeme ne jednu náhodnou veličinu, ale vektor náhodných veličin
- metody klasifikace, kdy předpokládáme, že data pocházejí z více populací a
  - hledáme pravidlo umožňující zařadit (klasifikovat) objekt charakterizovaný vektorem hodnot do jedné z populací (diskriminační analýza, logistická regrese, neuronové sítě atd.)
  - pokoušíme se najít v datech podmnožiny podobných objektů (shluková analýza – cluster analysis)
- metody redukce dimenze úlohy, kdy proměnlivost a závislosti v datech se pokoušíme vyjádřit pomocí méně veličin. Analýza hlavních komponent (principal components) vysvětluje rozptyl. Faktorová analýza vysvětluje kovarianční (korelační) strukturu.

12.1 Test shody vektoru středních hodnot

Pro test shody vektoru středních hodnot se užívá Hottelingův T^2 (čti té-kvadrát) test. Je to mnohorozměrná analogie t-testů. Ve stručnosti uvedeme základní myšlenky.

Jednovýběrový Hottelingův T^2 test:
Testuje se hypotéza, že p-rozměrný vektor středních hodnot $\boldsymbol{\mu}$ je roven nějakému danému konstantnímu vektoru. Předpokládá se, že výběr je z mnohorozměrného normálního rozdělení. Testovou statistikou je pak

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}), \tag{48}$$

Tato statistika má Hottelingovo rozdělení. Lze také užít statistiku

$$\frac{T^2}{n-1} \, \frac{n-p}{p} \sim F_{p,n-p} \tag{49}$$

Intervaly spolehlivosti pro p-rozměrný vektor středních hodnot odvodíme z (48) a (49).

$$P\left[\frac{T^2}{n-1} \, \frac{n-p}{p} < F_{1-\alpha}(p, n-p)\right] = 1 - \alpha$$

Po úpravě dostaneme

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) < \frac{n-1}{n} \frac{p}{n-p} F_{1-\alpha}(p, n-p),$$

kde $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$ znamená plochu elipsoidu se středem $\bar{\mathbf{x}}$, jehož tvar a velikost závisí na výběrové kovarianční matici $\mathbf{S}$. Volbou α určíme hodnotu c a můžeme určit intervaly spolehlivosti pro vektor středních hodnot $\boldsymbol{\mu}$.

Dvouvýběrový Hottelingův T^2 test:
Testujeme shodu dvou vektorů středních hodnot (mnohorozměrná analogie dvouvýběrového t-testu). Máme dva výběry z p-rozměrného normálního rozdělení o rozsazích $n_1,\ n_2$, $n_1 + n_2 = n$. Vektory výběrových průměrů jsou $\bar{\mathbf{x}}_1,\ \bar{\mathbf{x}}_2$. Za předpokladu shody kovariančních matic $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ můžeme z výběrových kovariančních matic $\mathbf{S}_1, \mathbf{S}_2$ odhadnout společnou výběrovou kovarianční matic

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

Označíme $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Pak statistika

$$T^2 = \frac{n_1 n_2}{n} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})$$

má Hottelingovo rozdělení a

$$\frac{n-p-1}{p}\frac{T^2}{n-2} \sim F(p, n-p-1),$$

kterou můžeme užít k testu hypotézy

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

Pokud $\boldsymbol{\Sigma_1} \neq \boldsymbol{\Sigma_2}$, jedná se o mnohorozměrnou analogii dvouvýběrového t-testu s nestejnými rozptyly. Pak je test poněkud komplikovanější, viz např. Hebák a kol.

12.2 Diskriminační analýza

Diskriminační analýza je postup, který hledá vhodné pravidlo (rozhodovací funkci) umožňující na základě zadaných hodnot vektoru $\mathbf{x}$ zařadit objekt do některé, řekněme h-té skupiny. Např. velmi jednoduchou úlohou tohoto typu je rozhodnout podle změřené teploty osoby o tom, zda je zdravá či nemocná. V tomto případě $\mathbf{x}$ je skalár (teplota) a rozhodovací pravidlo velice jednoduché: je-li teplota vyšší než 37^0 C, pak zařaď do skupiny nemocných, jinak do skupiny zdravých. Tedy klasifikujeme osobu, u níž příslušnost do skupiny neznáme a užíváme rozhodovacího pravidla získaného z dat popisujících vztah teploty příslušnosti ke skupině.

Je jasné, že naším zájmem je najít takové pravidlo, které by klasifikovalo pokud možno správně. V reálném světě většinou není možné najít pravidlo klasifikující správně vždy. Proto dobré pravidlo bude takové, které minimalizuje pravděpodobnost chybných rozhodnutí. Jak uvidíme za chvilku, za jistých předpokladů je takovým pravidlem lineární diskriminační funkce. Odvození jejího tvaru si ukážeme pro klasifikaci do dvou skupin.

Zavedeme následující označení:

$h = 1, 2$ – index skupiny

A_h – jev „příslušnost k h-té skupině“

$P(A_h) = \pi_h$ – apriorní pravděpodobnost

$f_h(\mathbf{x})$ – sdružená hustota pro h-tou skupinu

$P(A_h|\mathbf{x})$ – aposteriorní pravděpodobnost, tj. pravděpodobnost příslušnosti k h-té skupině za podmínky daných hodnot $\mathbf{x}$

Hustotu můžeme zapsat $f_h(\mathbf{x}) = f(\mathbf{x}|A_h)$ pro $h = 1, 2$, tj. sdružená hustota pro h-tou skupinu je hustota za podmínky, že nastane jev A_h.

Podle Bayesova vzorce vyjádříme aposteriorní pravděpodobnost:

$$P(A_h|\mathbf{x}) = \frac{P(A_h)f_h(\mathbf{x}|A_h)}{P(A_1)f(\mathbf{x}|A_1) + P(A_2)f(\mathbf{x}|A_2)} = \frac{\pi_h f_h(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})}, \qquad (50)$$

$$h = 1, 2.$$

Klasifikovat budeme do skupiny s největší aposteriorní pravděpodobností.

Dále označme $\mathcal{S}$ – výběrový prostor (množinu všech možných výsledků $\mathbf{x}$). Naším cílem je rozdělit tento výběrový prostor na dvě části splňující podmínky:

$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2, \quad \mathcal{S}_1 \cap \mathcal{S}_2 = \oslash.$$

Pak když $\mathbf{x} \in \mathcal{S}_h$, zařadíme do h - té skupiny.

Pravděpodobnost chybného zařazení objektu z h-té skupiny do h'-té skupiny je

$$P(\mathbf{x} \in \mathcal{S}_{h'}|A_h) = \int_{\mathcal{S}_{h'}} f_h(\mathbf{x})\mathrm{d}\mathbf{x}, \quad h = 1, 2.$$

Podle věty o úplné pravděpodobnosti je celková pravděpodobnost chybné klasifikace

$$\omega = \pi_1 \int_{\mathcal{S}_2} f_1(\mathbf{x})\mathrm{d}\mathbf{x} + \pi_2 \int_{\mathcal{S}_1} f_2(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{51}$$

Pokud obě chyby klasifikace mají stejnou váhu, je optimální rozhodovací pravidlo, které minimalizuje ω dané vztahem (51). Chceme-li chybám klasifikace dát různou váhu, užijeme ztrátovou matici:

$$\mathbf{Z} = \begin{bmatrix} 0 & z(2|1) \\ z(1|2) & 0 \end{bmatrix}$$

Pak celková ztráta z chybné klasifikace je:

$$\tau = z(2|1)\pi_1 \int_{S_2} f_1(\mathbf{x})\mathrm{d}\mathbf{x} + z(1|2)\pi_2 \int_{S_1} f_2(\mathbf{x})\mathrm{d}\mathbf{x}$$

a optimální je postup, který minimalizuje τ.

Objekt řadíme do skupiny s vyšší aposteriorní pravděpodobností, např. z rov. (50) do skupiny 1 zařadíme objekt, když $\pi_1 f_1(x) > \pi_2 f_2(x)$ (jmenovatel je shodný pro obě skupiny). Klasifikační pravidlo pro zařazení do skupiny 1 je tedy

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1} \tag{52}$$

Předpokládáme-li p-rozměrné normální rozdělení vektoru $\mathbf{x}$, tj. $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ v 1. skupině a $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ ve 2. skupině, pak hustota je:

$$f_h(\mathbf{x}) = (2\pi)^{\frac{-p}{2}} |\boldsymbol{\Sigma}_h|^{-\frac{1}{2}} \exp\left[-(\mathbf{x} - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_h^{-1} (\mathbf{x} - \boldsymbol{\mu}_h)/2\right]$$

Po dosazení do (52) a zlogaritmování dostaneme

$$\mathbf{x}^T \boldsymbol{\Gamma} \mathbf{x} + \boldsymbol{\eta}^T \mathbf{x} + \xi > 0,$$

kde

$$\boldsymbol{\Gamma} = 0,5(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}),$$

$$\boldsymbol{\eta}^T = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2 \boldsymbol{\Sigma}_2^{-1}$$

$$\xi = \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - \ln \frac{\pi_2}{\pi_1} - \frac{1}{2} \left(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2\right)$$

Jsou-li kovarianční matice v obou skupinách shodné, tj. $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, pak odpadne kvadratický člen a rozhodovací pravidlo se podstatně zjednoduší:

$$\boldsymbol{\beta}^T \mathbf{x} + \gamma > 0,$$

kde

$$\boldsymbol{\beta}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}$$

a

$$\gamma = -\frac{1}{2}\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{1}{2}\ln\frac{\pi_2}{\pi_1}$$

Funkce

$$L(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x} \tag{53}$$

se nazývá *lineární diskriminační funkce*, zkratkou LDF.

Pokud $\mathbf{x}$ má p-rozměrné normální rozdělení, pak i $L(\mathbf{x})$ má normální rozdělení. Čtverec Mahalanobisovy vzdálenosti vektorů středních hodnot je

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

střední hodnoty LDF jsou pak pro skupinu 1 a skupinu 2 jsou

$$E_1[L(\mathbf{x})] = \frac{1}{2}\triangle^2 \qquad E_2[L(\mathbf{x})] = -\frac{1}{2}\triangle^2$$

Oba rozptyly jsou shodné

$$\mathrm{var}_1[L(\mathbf{x})] = \mathrm{var}_2[L(\mathbf{x})] = \triangle^2$$

Oba podprostory $\mathcal{S}_1$ a $\mathcal{S}_2$ v p-rozměrném podprostoru $\mathcal{S}$ odděluje nadrovina určená rovnicí

$$\boldsymbol{\beta}^T \mathbf{x} + \gamma = 0 \quad \text{čili} \quad L(\mathbf{x}) = -\gamma$$

LDF (53) lze vyjádřit také jako

$$L_h(\mathbf{x}) = \boldsymbol{\mu}_h^T \mathbf{\Sigma}^{-1} \mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_h^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_h$$

a klasifikovat do té skupiny, pro kterou je $L_h(\mathbf{x})$ největší. Tak se postupuje, když se klasifikuje do více než dvou skupin.

LDF je optimální rozhodovací pravidlo pro klasifikaci do skupin, pokud náhodný vektor $\mathbf{x}$ má normální rozdělení a skupiny se liší jen vektorem středních hodnot, nikoliv kovarianční strukturou.

Procedura diskriminační analýzy z dat, u kterých je klasifikace známa, odhaduje hodnoty parametrů lineární diskriminační funkce $\boldsymbol{\beta}$. Pak LDF ve tvaru (53) s hod-

notami odhadů lze užít pro klasifikaci objektů, jejichž příslušnost do skupiny známa není.

Příklad 12.1 V souboru DISKRIM.XLS jsou na 30 objektech, které pocházejí ze dvou populací (veličina *skup*, změřeny hodnoty 10 spojitých veličin ($x1$ až $x10$). Naším úkolem je nalézt pravidlo pro klasifikaci objektů. Pravidlo má být co nejjednodušší (čím méně veličin, tím lépe). Použijeme jednak diskriminační analýzu (lineární diskriminační funkci), jednak logistickou regresi [14].

Abychom ověřili, zda vůbec můžeme lineární diskriminační funkci užít, je nutné, aby se populace lišily ve středních hodnotách. To lze zjistit pomocí dvouvýběrového Hotellinova testu:

```
Two-Sample Hotelling's T2 Report

Group     skup

Descriptive Statistics
                Means              Standard Deviations
Variable     0          1             0             1
x1       12.45333     17.25333     2.289375     2.207088
x2       14.996       13.638       2.427947     2.424848
x3       12.05333     17.23333     3.346612     2.916129
x4      183.5267     236.06       73.80299     62.96599
x5      180.28       232.2533     37.71711     47.72365
x6      187.74       233.14       43.81628     48.71579
x7      190.1267     240.2667     42.31836     47.44453
x8      188.4933     233.7267     37.21266     52.34204
x9      190.2333     238.74       31.271       54.68576
x10     188.08       231.4333     50.21964     61.33117
Count    15           15          15           15
```

Už letmý pohled na popisné statistiky výše naznačuje, že mezi průměry některých veličin ve skupinách jsou významné rozdíly. To potvrzuje jak Hotellingův test, tak dvouvýběrové t-testy pro jednotlivé veličiny.

```
Hotelling's T2 Test Section

Covariance                              Prob
Assumption      T2        DF1      DF2      Level
Equal        101.988      10       28       0.0002
Unequal      101.988      10       27       0.0002
```

```
Student's T-Test Section

                            Prob
Variable     |Student's T|  Level
All (T2)       101.988     0.0002
x1               5.846     0.0000
x2               1.533     0.1366
x3               4.520     0.0001
x4               2.097     0.0451
x5               3.309     0.0026
x6               2.684     0.0121
x7               3.055     0.0049
x8               2.728     0.0109
x9               2.982     0.0059
x10              2.118     0.0432
These individual t-test significance levels
should only be used when the overall T2 value is significant.
```

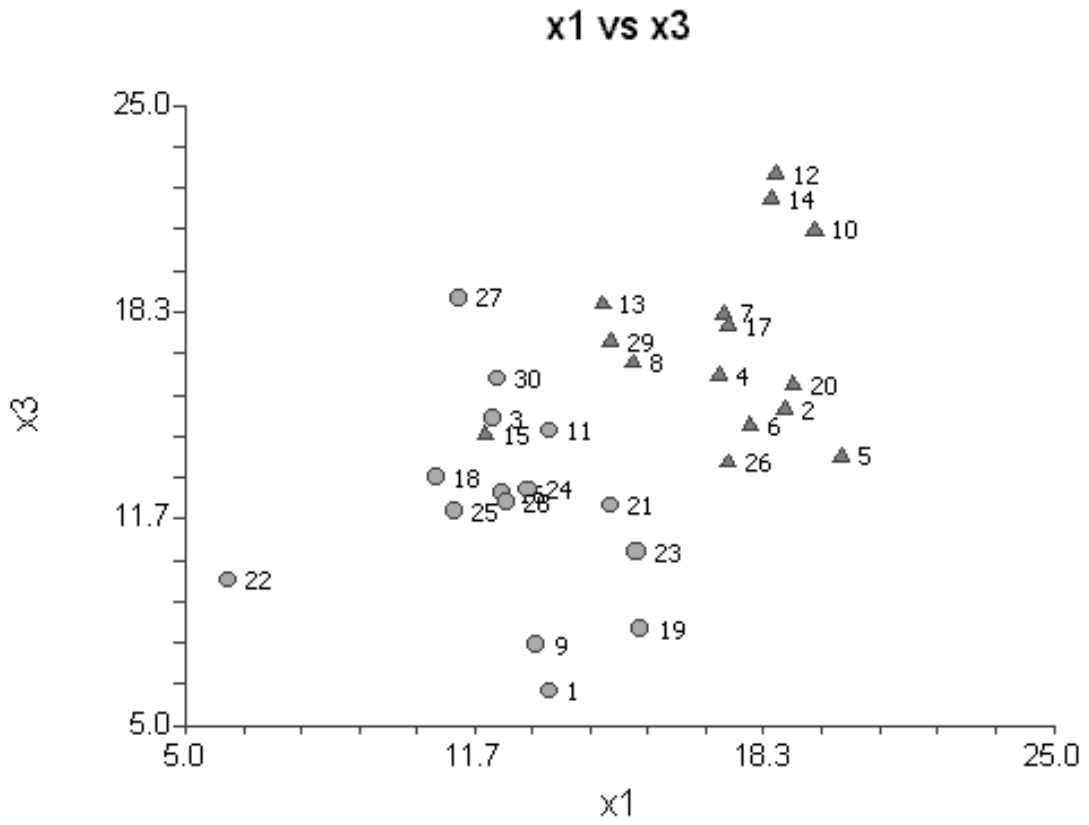
Stepwise procedura diskriminační analýzy poskytne následující výstup:

```
Discriminant Analysis Report

Dependent     skup

Variable-Selection Summary Section
            Action     Independent  Pct Chg In          Prob
Iteration   This Step  Variable     Lambda     F-Value  Level
0           None
1           Entered    x1           54.97       34.18   0.000003
2           Entered    x3           25.24        9.12   0.005481

Variable-Selection Detail Section - Step 2
       Independent  Pct Chg In           Prob       R-Squared
Status Variable     Lambda     F-Value   Level      Other X's
In      x1          41.77     19.37       0.000152   0.226687
In      x3          25.24     9.12        0.005481   0.226687
Out     x2           4.74     1.29        0.265589   0.118719
Out     x4           6.85     1.91        0.178413   0.749175
Out     x5           0.17     0.04        0.836049   0.398169
Out     x6           6.30     1.75        0.197650   0.505648
```

```
Out     x7          1.46     0.39        0.539842   0.485859
Out     x8          6.35     1.76        0.195840   0.505363
Out     x9          0.33     0.09        0.772237   0.485821
Out     x10         2.25     0.60        0.445960   0.320911
Overall Wilks' Lambda = 0.336670
Action this step:  None


Linear Discriminant Functions

                    skup
Variable        0            1
Constant    -22.93688    -44.95984
x1            2.481297     3.438486
x3            1.242258     1.775299

Classification Count Table for skup
          Predicted
Actual     0     1     Total
0         15     0     15
1          1    14     15
Total     16    14     30
```

Jako veličiny významně odlišující dvě skupiny byly do klasifikačního pravidla krokovou procedurou vybrány $x1$ a $x3$. Ze 30 objektů je pak 29 klasifikováno správně a jen jeden chybně. Toto empirické ověření spolehlivosti klasifikace je však nutno brát s opatrností, neboť spolehlivost klasifikačního pravidla je ověřována na datech, ze kterých byly koeficienty lineární diskriminační funkce spočítány, nikoliv na nezávislých pozorováních, kde musíme počítat s nižší úspěšností. Realističtější odhad očekávané úspěšnosti klasifikace lze získat buď tak, že data rozdělíme náhodně na skupinu učící a testovací, parametry lineární diskriminační funkce se spočítají jen z učící skupiny a úspěšnost klasifikace se odhadne ze zjištěné klasifikace testovací skupiny. Tento postup má ovšem nevýhodu, že parametry lineární diskriminační funkce se počítají z podstatně menšího počtu pozorování, tzn. nevyužije se informace v datech. V některých programech (v NCSS prozatím nikoliv) je proto *jackknife* procedura, která postupně spočítá parametry lineární diskriminační funkce z $n-1$ objektů a úspěšnost klasifikace se vždy ověřuje na objektu, který byl vyjmut. Tak se získá odhad spolehlivosti klasifikace na nezávislých pozorováních.

Na obrázku, který následuje, vidíme nesprávně klasifikovaný objekt 15 (ze skupiny 1), který leží uvnitř shluku objektů skupiny 0. Pro ostatní body v rovině lze skupiny od sebe oddělit lineární funkcí (přímkou).



Stejnou úlohu hledání klasifikačního pravidla pro klasifikaci do dvou skupin lze řešit i logistickou regresí. Příslušnost do skupiny je nutno označit { 0, 1 }. Klasifikace je pak založena na odhadu pravděpodobnosti, že pro dané hodnoty regresorů má veličina Y má hodnotu 1. Tvar klasifikační funkce lze snadno vyjádřit z modelu logistické regrese (34)

$$p = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T\boldsymbol{\beta})}$$

Je-li p větší než zvolená hodnota (většinou 0,5), pak objekt klasifikujeme do skupiny 1, jinak do skupiny 0. Klasifikační pravidlo na rozdíl od lineární diskriminační funkce je složitější, především není lineární funkcí regresorů. To v některých případech je výhodou, neboť lze najít vhodné klasifikační pravidlo i pro skupiny, které nejsou lineárně separabilní nebo v situacích, kdy nejsou splněny poměrně přísné předpoklady pro aplikaci lineární diskriminační funkce (mnohorozměrné normální rozdělení, shoda kovariančních matic ve skupinách). Někdy je ovšem tato výhoda problematická, jak ukazuje následující výstup z postupné logistické regrese (metoda forward, tj. postupné přidávání významných regresorů bez vylučování nadbytečných).

Příklad 12.2

```
Logistic Regression Report

Response    skup

Forward Variable-Selection
Action    Variable
Added      x1
Added      x3
Added      x5
Added      x2

Parameter Estimation Section
            Regression    Standard Chi-Square Prob
Variable    Coefficient    Error    Beta=0    Level
Intercept  -231.4874     68355.56     0.00    0.997
x1            4.703548    2773.934    0.00    0.998
x3            4.343548    1478.741    0.00    0.997
x5            2.484694     322.7473   0.00    0.993
x2           -27.89017     5084.675    0.00    0.995

Model Summary Section
Model        Model    Model        Model
R-Squared    D.F.    Chi-Square    Prob
0.624562      4       41.59        0.000000

Classification Table
                  Predicted
Actual             0     1   Total
0        Count    15     0    15
1        Count     0    15    15
Total    Count    15    15    30
Percent Correctly Classified=100
```

Logistickou regresí se podařilo nalézt pravidlo se čtyřmi regresory $x1$, $x3$, $x5$ a $x2$, které má stoprocentní úspěšnost klasifikace. Lineární diskrimininační funkce pro tato data stoprocentní úspěšnosti klasifikace nedosáhne ani při zařazení všech deseti veličin. Ale při podrobnějším pohledu na odhady parametrů logistického modelu a statistiky s nimi spojené vidíme, že směrodatné odchylky odhadů parametrů jsou velmi vysoké v porovnání s hodnotami odhadů, takže vlastně žádný z odhadů parametrů nemůžeme považovat za významně odlišný od nuly. Klasifikační pravidlo je

„ušito na míru“ datům. Pokud zpřísníme kritérium pro zařazování regresorů, dostaneme následující výstup:

```
Logistic Regression Report

Forward Variable-Selection
Action      Variable
Added        x1
Added        x3

Parameter Estimation Section
              Regression       Standard      Chi-Square  Prob
Variable      Coefficient      Error         Beta=0      Level
Intercept     -26.49609       10.43108        6.45      0.011082
x1              1.113222       0.5043868      4.87      0.027309
x3              0.7096213      0.3449301      4.23      0.039658

Model Summary Section
Model         Model      Model           Model
R-Squared      D.F.      Chi-Square      Prob
0.540694        2        31.78          0.000000

Classification Table
            Predicted
Actual              0      1      Total
0      Count       15      0      15
1      Count        1     14      15
Total  Count       16     14      30
Percent Correctly Classified=96.67
```

Toto klasifikační pravidlo obsahuje dva regesory (stejné jako LDF), všechny parametry tohoto logistického modelu můžeme považovat za nenulové a klasifikační pravidlo má stejnou úspěšnost jakou měla LDF.

12.3 Shluková analýza

Cílem shlukové analýzy je nalézt v datech podmnožiny podobných objektů. Mějme množinu m objektů, tuto množinu označme M.

Pro každé dva objekty $a, b \in M$ máme číslo $\sigma(a, b)$, kterému říkáme numerická podobnost.

$$\sigma : M \times M \to R$$

Požadavky na vlastnosti numerické podobnosti:

1. $0 \leq \sigma(a, b) \leq 1$
2. $\sigma(a, a) = 1$
3. $\sigma(a, b) = \sigma(b, a)$
4. $\min(\sigma(a, b), \sigma(b, c)) \leq \sigma(a, c)$ – slabší trojúhelníková nerovnost

Poznámka: Charakteristiku $(1-\sigma(a, b))$ můžeme chápat jako normovanou vzdálenost dvou objektů $a,\ b$.

Úlohou shlukové analýzy je najít rozklad $\{M_i\}_{i=1}^{k}$, množiny M tj.

1. $\bigcup_{i=1}^{k} M_i = M$
2. $M_i \cap M_j = 0$ pro $i \neq j$
3. vágní kritérium: objekty uvnitř M_i jsou si podobnější mezi sebou než s objekty z množiny M_j, např. když $a, b \in M_i, c \in M_j$, pak $\sigma(a, b) \leq \sigma(a, c), \sigma(a, b) \leq \sigma(b, c)$

Je mnoho možností,

- jak definovat numerickou podobnost,
- jak formulovat postup zařazování objektů do podmnožin,

tedy existuje mnoho metod shlukování.

12.3.1 Hierarchické metody

Vychází se z matice podobnosti objektů (symetrická matice s jedničkami na diagonále), nejčastější je aglomerativní procedura, začne od m shluků (každý shluk je tvořen jedním objektem a spojuje ty shluky, které jsou si nejpodobnější, až skončí jedním shlukem, obsahujícím všech m objektů. Pro takovou posloupnost rozkladů $\{M_{ij}\}_{j=1}^{k}$, pro $i_1 < i_2$ platí $M_{i_1,j} \subseteq M_{i_2,j}$, tj. rozklady jsou do sebe zasunuty, objekty

jednou spojené do shluku zůstávají spolu. Posloupnost spojování můžeme je graficky znázornit *dendrogramem*. Podobnou úlohu řeší taxonomie v biologii.

Nejčastěji užívané strategie spojování shluků jsou:

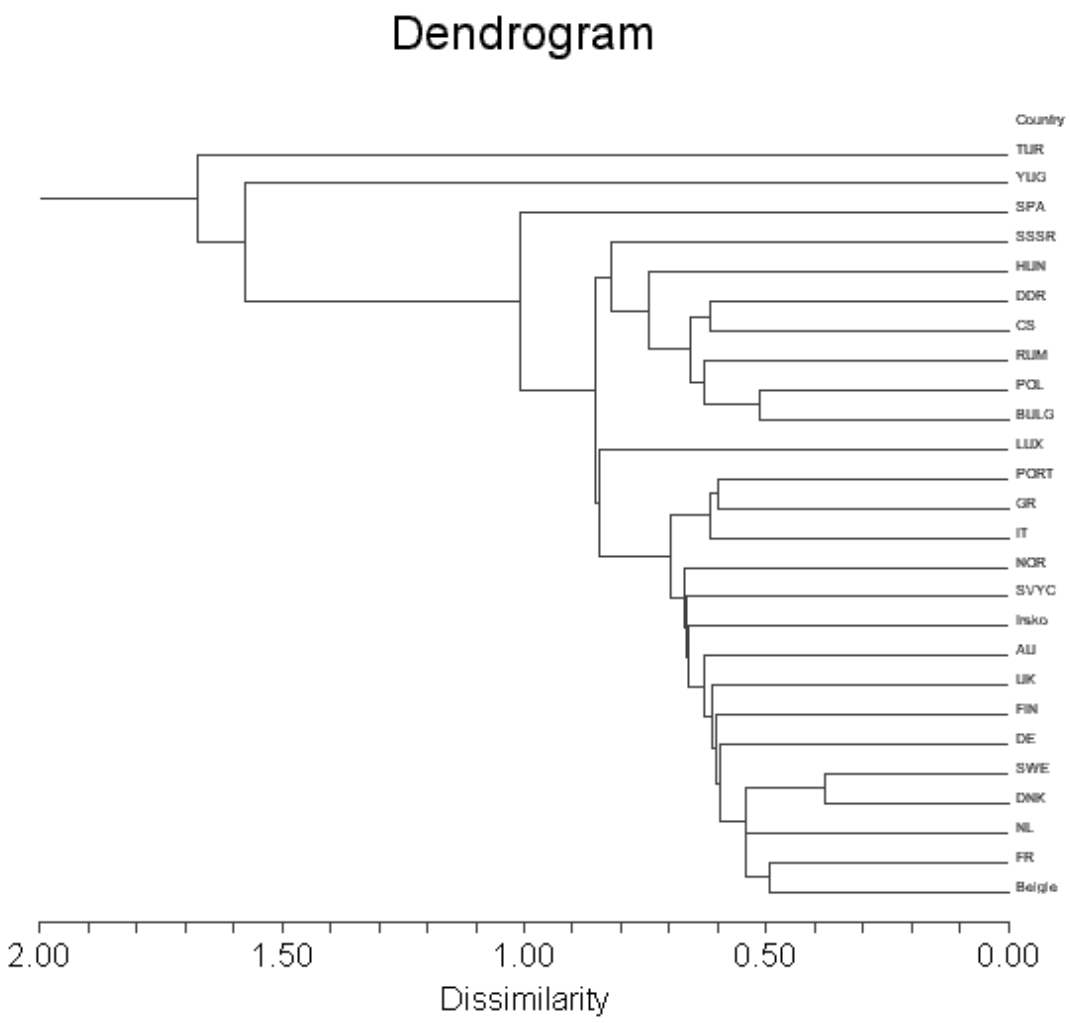
- *single linkage* (nejbližší soused, nearest neiborough) - shluk tvoří souvislý podgraf, tj. existuje aspoň jedna cesta mezi dvěma uzly podgrafu, nejméně přísná metoda na podobnost uvnitř shluků, shluky mají tvar „souhvězdí“
- *complete linkage* (nejvzdálenější soused, furthest neiborough) shluk tvoří úplný podgraf, tj. každé dva uzly podgrafu jsou spojeny hranou, nejpřísnější na podobnost uvnitř shluku
- *average linkage* - spojuje shluky podle jejich průměrné vzdálenosti
- *centroidní* - spojuje shluky podle vzdáleností jejich těžiště

Rozdíly mezi strategiemi shlukování ilustrují následující příklady.

Příklad 12.3 Data pro tento příklad jsou z knihy [18] a jsou uvedena i v souboru EMPLOY.XLS. Obsahují údaje o podílu zaměstnaných v devíti odvětvích ve 26 evropských zemí. Údaje jsou z konce 70. let 20. století, proto jsou v nich uvedeny i státy, které už nyní neexistují. Jednotlivé veličiny znamenají: AGR = agriculture (zemědělství), MIN = mining (těžba), MAN = manufacturing (těžký průmysl), PS = power supplies (energetika), CON = construction (stavebnictví), SER = service industries (lehký průmysl), FIN = finance, SPS = social and personal services (sociální služby), TC = transport and communications (doprava a spoje).

Ve všech ukázkách je zvolena Eukleidovská vzdálenost mezi objekty, všechny veličiny byly standardizovány, po standardizaci tedy mají jednotkový rozptyl. Výstupy se liší jen podle použité strategie (metody) shlukování.

```
Hierarchical Clustering Report
Variables            AGR to TC
Clustering Method    Single Linkage (Nearest Neighbor)
Distance Type        Euclidean
Scale Type           Standard Deviation
```



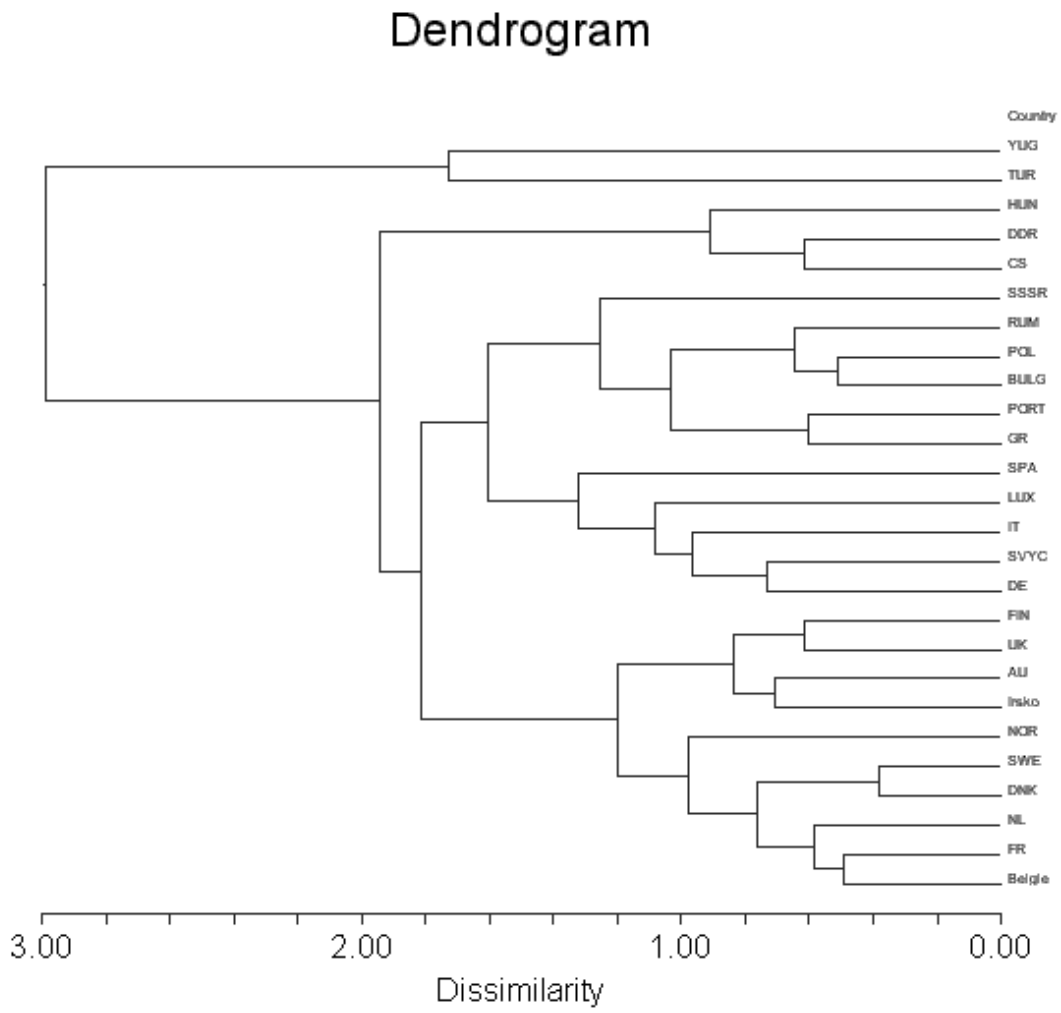
Na dendrogramu vidíme, jak postupně byly vytvářeny shluky. Jako první se spojily nejpodobnější objekty, tj. Dánsko a Švédsko, pak Belgie s Francií atd. Dendrogram ukazuje typické chování metody nejbližšího souseda, kdy k velkým shlukům jsou připojovány jednotlivé objekty nebo shluky s malým počtem objektů. Na úrovni nepodobnosti (vzdálenosti) zhruba 0,85 jsou země rozděleny do 6 shluků:

1. západoevropské země s výjimkou Španělska a Lucemburska
2. Lucembursko
3. země bývalého socialistického bloku
4. Španělsko
5. bývalá Jugoslávie
6. Turecko

Příklad 12.4

`Hierarchical Clustering Report`

```
Variables            AGR to TC
Clustering Method    Complete Linkage (Furthest Neighbor)
Distance Type        Euclidean
Scale Type           Standard Deviation
```

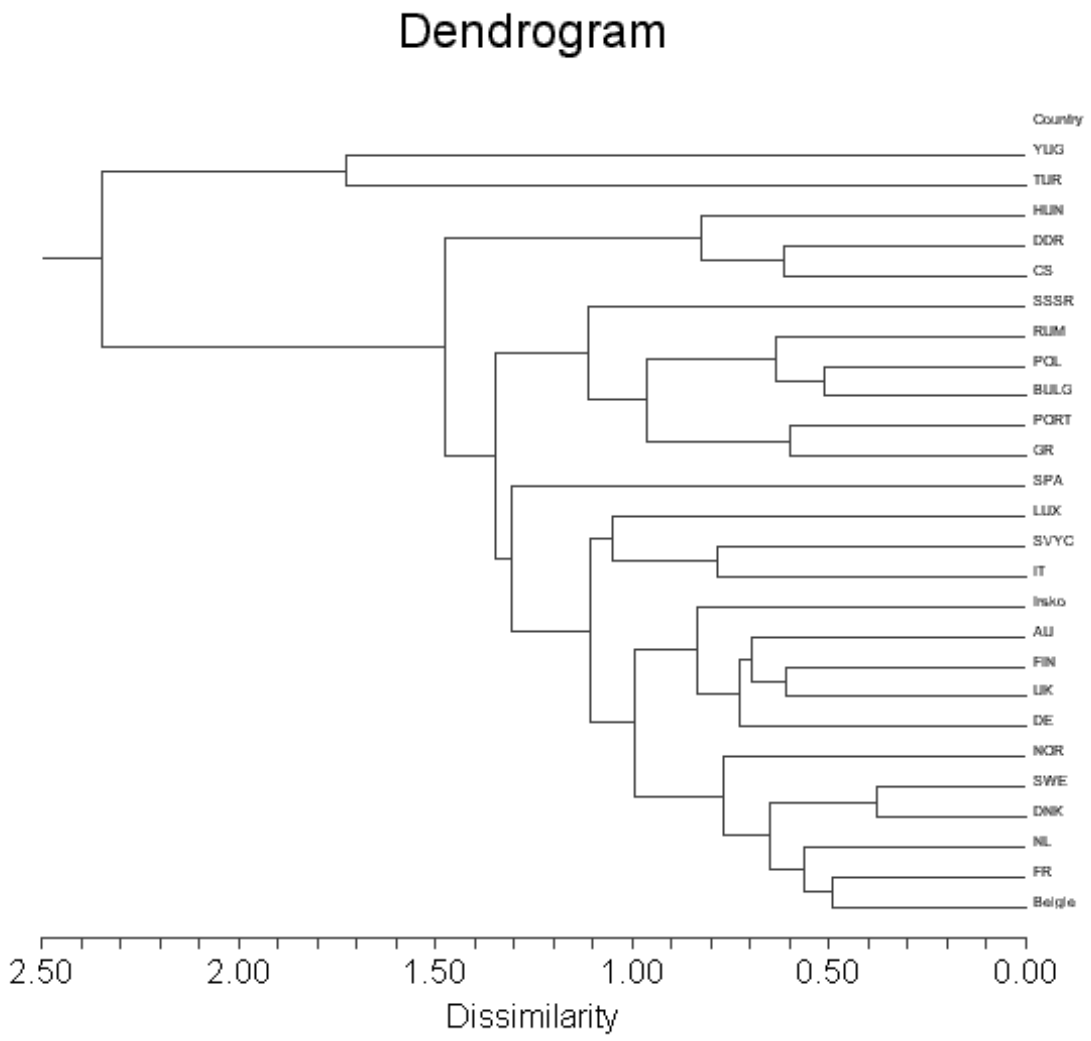


Jako první se opět spojily nejpodobnější objekty, tj. Dánsko a Švédsko, pak Belgie s Francií atd. Metoda nejvzdálenějšího souseda má tendenci vytvářet kompaktnější shluky, ve kterých je počet objektů rovnoměrnější. Na úrovni nepodobnosti (vzdálenosti) zhruba 0,95 jsou země rozděleny do 6 shluků:

1. západoevropské Belgie až Finsko, seznam viz dendrogram
2. SRN, Švýcarsko, Itálie, Lucembursko a Španělsko
3. Řecko, Portugalsko a přímořské země bývalého socialistického bloku s výjimkou NDR
4. Československo, NDR a Maďarsko
5. Turecko
6. bývalá Jugoslávie

Příklad 12.5

```
Hierarchical Clustering Report
Variables            AGR to TC
Clustering Method    Simple Average (Weighted Pair-Group)
Distance Type        Euclidean
Scale Type           Standard Deviation
```



Metoda průměrné vzdálenosti vedla k výsledkům, které jsou podobné metodě nejvzdálenějšího souseda, rozdíly jsou především uprostřed shlukovací procedury, např. Španělsko se ke shluku západoevropských zemím připojilo později než SSSR ke shluku Rumunska, Polska atd.

12.3.2 Nehierarchické metody

Mezi nejpopulárnější nehierarchické metody patří metoda k-means . Počet shluků k je předem znám, objekty se rozdělují do shluků tak, rozptyl uvnitř shluků (within sum of squares) byl co nejmenší. Jde tedy o to, abychom nalezli takové přiřazení objektů do shluků tak, aby stopa matice $\mathbf{W}$ byla minimální.

$$\mathbf{W} = \sum_{g=1}^{k} \mathbf{W}_g, \tag{54}$$

$\mathbf{W}_g$ je Wishartova matice pro shluk g, tj.

$$\mathbf{W}_g = \sum_{j=1}^{n_g} (\mathbf{x}_j^{(g)} - \bar{\mathbf{x}}^{(g)})(\mathbf{x}_j^{(g)} - \bar{\mathbf{x}}^{(g)})^T, \tag{55}$$

kde $\mathbf{x}_j^{(g)}$ je vektor hodnot veličin $j$tého objektu v g-tém shluku, $\bar{\mathbf{x}}^{(g)} = \left(\sum_{j=1}^{n_g} \mathbf{x}_j^{(g)}\right)/n_g$ vektor průměrů (centroid) g-tého shluku. Kritériem, jež má být minimalizovano, je pak

$$\mathrm{TRW} = \mathrm{tr}(\mathbf{W}). \tag{56}$$

Najít globální minimum je algoritmicky obtížný problém, který neumíme vyřešit v polynomiálním čase. Obvykle se užívá se Hartiganův algoritmus *k-means*, který umí najít přijatelné lokální minimum pro většinu jednodušších klasifikačních úloh nebo se v poslední době pro optimalizaci klasifikace využívají evoluční algoritmy.

Algoritmus *k-means* je velmi jednoduchý:

1. Nejdříve se k centroidů (těžišť shluků) zvolí náhodně, buď se vybere náhodně k objektů ze zadaných dat nebo se objekty náhodně klasifikují do k shluků a spočítají jejich těžiště (vektor průměrů).
2. Objekty se zařadí do shluku, jehož těžišti jsou nejbližší a spočítá se nové těžiště každého shluku.
3. Krok 2 se opakuje tak dlouho, dokud dochází ke změně klasifikace objektů.

Příklad 12.6 Využijeme opět data ze souboru EMPLOY.XLS. Stručné výsledky pro šest shluků následují.

```
K-Means Cluster Analysis Report

Iteration Section
Iteration No. of    Percent of    Bar Chart
No.       Clusters  Variation    of Percent
1          2        72.59        ||||||||||||||||||||
2          3        52.48        ||||||||||||||||
3          4        43.44        ||||||||||||||
4          5        36.60        ||||||||||||
5          6        33.70        ||||||||||||
```

Vidíme, jak s počtem shluků klesá podíl variability uvnitř shluků (within sum of squares) na celkové variabilitě. Nejvýraznější skok je mezi 2 a 3 shluky,pak už se rychlost snižování zmenšuje.

Průměry (tj. souřadnice těžiště shluků) jsou v následující tabulce. Podobnou tabulku volitelně obsahuje výstup z procedury k-means [14] i pro směrodatné odchylky uvnitř shluků.

```
Cluster Means
Variab Clust1    Clust2   Clust3   Clust4  Clust5  Clust6
AGR     31.7      12.9     20.13    9.76    6.78    57.75
MIN      0.95      0.97     2.45    1.20    0.4      1.1
MAN     25.17     26.75    31.68   31.9    24.04    12.35
PS       0.62      1.35     1.08    0.78    0.82     0.6
CON      9.17      7.7      8.33    8.98    8.44     3.85
SER     10.1      16.3      8.56   17.06   16.6      5.8
FIN      3.72      4.72     0.85    4.50    6.04     6.2
SPS     12.8      22.55    19.18   19.92   29.46     8.6
TC       5.72      6.77     7.71    5.88    7.46     3.6
Count    4         4        6       5        5       2
```

Země byly do shluků zařazeny takto:

1. Řecko, Portugalsko, Španělsko, Rumunsko
2. Irsko, Británie, Rakousko, Finsko
3. Bulharsko, Československo, NDR, Maďarsko, Polsko, SSSR
4. Francie, SNR, Itálie, Lucembursko, Švýcarsko

5. Belgie, Dánsko, Holandsko, Norsko, Švédsko
6. Turecko, Jugoslávie

Výsledky se s tím, co poskytly hierarchické procedury, shodují jen částečně, ovšem podstatné rysy možné klasifikace jsou společné. Právě porovnání výsledků více shlukovacích procedur a nalezení jejich společných rysů je užitečné pro úvahy o možné klasifikaci.

K této úloze se ještě vrátíme v kapitole o hlavních komponentách.

12.4 Analýza hlavních komponent

Analýza hlavních komponent (Principal components analysis, PCA) je jedna z metod redukce dimenze úlohy. Snaží se vysvětlit celkový rozptyl vektoru náhodných veličin, resp. jeho podstatnou část pomocí méně veličin.

$\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$ náhodný vektor
$\mathbf{V} = [\mathrm{cov}(X_i, X_j)] = [\sigma_{ij}]$,
$i, j = 1, 2, \ldots, p$ kovarianční (varianční) matice

Kovarianční matice $\mathbf{V}$ typu $(p \times p)$ je symetrická (vlastní čísla jsou reálná) a pozitivně semidefinitní, tj. $\mathbf{y}^T\mathbf{V}\mathbf{y} \geq 0$ pro jakýkoliv vektor $\mathbf{y} \neq 0$. Tzn., že všechna vlastní čísla jsou nezáporná, – důkaz viz např. Anděl, str. 28 [2]

Vlastní čísla můžeme uspořádat:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

Matici $\mathbf{V}$ můžeme napsat jako

$$\mathbf{V} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \qquad \mathbf{U}\mathbf{U}^T = \mathbf{I}, \tag{57}$$

kde $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ a k – tý sloupec matice $\mathbf{U}$ je vlastní vektor $\mathbf{v}_k$ matice $\mathbf{V}$, který přísluší vlastnímu číslu λ_k a pro který platí $\mathbf{v}_k^T\mathbf{v}_k = 1$. Tedy $\mathbf{v}_k$ jsou ortonormální vlastní vektory kovarianční matice $\mathbf{V}$, platí

$$\mathbf{V}\mathbf{v}_k = \lambda_k\mathbf{v}_k$$
$$\mathbf{V} = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T + \ldots + \lambda_p\mathbf{v}_p\mathbf{v}_p^T$$
$$\mathbf{I} = \mathbf{v}_1\mathbf{v}_1^T + \mathbf{v}_2\mathbf{v}_2^T + \ldots + \mathbf{v}_p\mathbf{v}_p^T$$

Vektory $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ tvoří bázi prostoru $\mathcal{R}^p$, takže libovolný vektor $\mathbf{y} \in \mathcal{R}^p$ lze vyjádřit jako

$$\mathbf{y} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_p\mathbf{v}_p$$

Platí pro každý vektor $\mathbf{y} \in \mathcal{R}^p$ jednotkové délky ($\mathbf{y}^T\mathbf{y} = 1$), že kvadratická forma $\mathbf{y}^T\mathbf{V}\mathbf{y} \leq \lambda_1$, při čemž rovnost platí pro $\mathbf{c} = \mathbf{v}_1$, tj. pro první vlastní vektor – viz Anděl, str. 297 [2].

Celková variabilita náhodného vektoru $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$ je součet diagonálních prvků kovarianční matice (rozptylů jednotlivých náhodných veličin):

$$\sigma^2 = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \ldots + \mathrm{var}(X_p)$$

Hledáme novou náhodnou veličinu, která vznikne lineární transformací vektoru $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$, tj. vlastně hledáme $\mathbf{c} \in \mathcal{R}^p$, $\mathbf{c}^T\mathbf{c} = 1$, aby náhodná veličina měla

co největší rozptyl. Tím tato nová veličina vyčerpá co největší část celkové variability, tzn. její rozptyl je roven λ_1. Tedy $\mathbf{c} = \mathbf{v}_1$. Náhodnou veličinu $Y_1 = \mathbf{v}_1^T\mathbf{x}$ nazýváme první hlavní komponentou.

Pak hledáme další náhodnou veličinu, tj. další $\mathbf{c} \in \mathcal{R}^p$, $\mathbf{c}^T\mathbf{c} = 1$, aby náhodná veličina $\mathbf{c}^T\mathbf{x}$ byla nekorelovaná s $Y_1 = \mathbf{v}_1^T\mathbf{x}$. Tomu vyhovuje $\mathbf{c} = \mathbf{v}_2$, takže druhá hlavní komponenta je $Y_2 = \mathbf{v}_2^T\mathbf{x}$ a její rozptyl je $\mathrm{var}(Y_2) = \mathrm{var}(\mathbf{v}_2^T\mathbf{x}) = \lambda_2$. Podobně i pro třetí a další hlavní komponenty.

Celková variabilita

$$\sigma^2 = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \ldots + \mathrm{var}(X_p) =$$

$$= \mathrm{var}(Y_1) + \mathrm{var}(Y_2) + \ldots + \mathrm{var}(Y_p) = \sum_{i=1}^{p} \lambda_i$$

Relativní podíl celkové variability vysvětlovaný i-tou hlavní komponentou je λ_i/σ^2. Prvních k hlavních komponent vysvětluje $\sum_{i=1}^{k} \lambda_i/\sigma^2$ z celkové variability.

V praktických aplikacích se vychází z výběrové kovarianční matice nebo častěji z výběrové korelační matice (abychom se vyhnuli vlivu volby jednotek, ve kterých měříme veličiny, na hodnoty výběrových rozptylů). Korelační matice je vlastně kovarianční matice standardizovaných veličin. Pak celková variabilita je rovna počtu veličin,

$$\sigma^2 = \sum_{i=1}^{p} \lambda_i = p$$

Příklad 12.7 Základní možnosti analýzy hlavních komponent ukážeme na příkladu o struktuře zaměstnanosti ve 26 evropských zemích, data jsou v souboru EMPLOY.XLS. Vycházíme z výběrové korelační matice.

```
Principal Components Report

Database   C:\avdat\employ.S0

Eigenvalues
                  Individual Cumulative
No.   Eigenvalue   Percent   Percent   Scree Plot
1      3.487151    38.75      38.75    ||||||||
2      2.130173    23.67      62.41    |||||
3      1.098958    12.21      74.63    |||
4      0.994483    11.05      85.68    |||
5      0.543218     6.04      91.71    ||
6      0.383428     4.26      95.97    |
7      0.225754     2.51      98.48    |
8      0.136790     1.52     100.00    |
9      0.000046     0.00     100.00    |
```

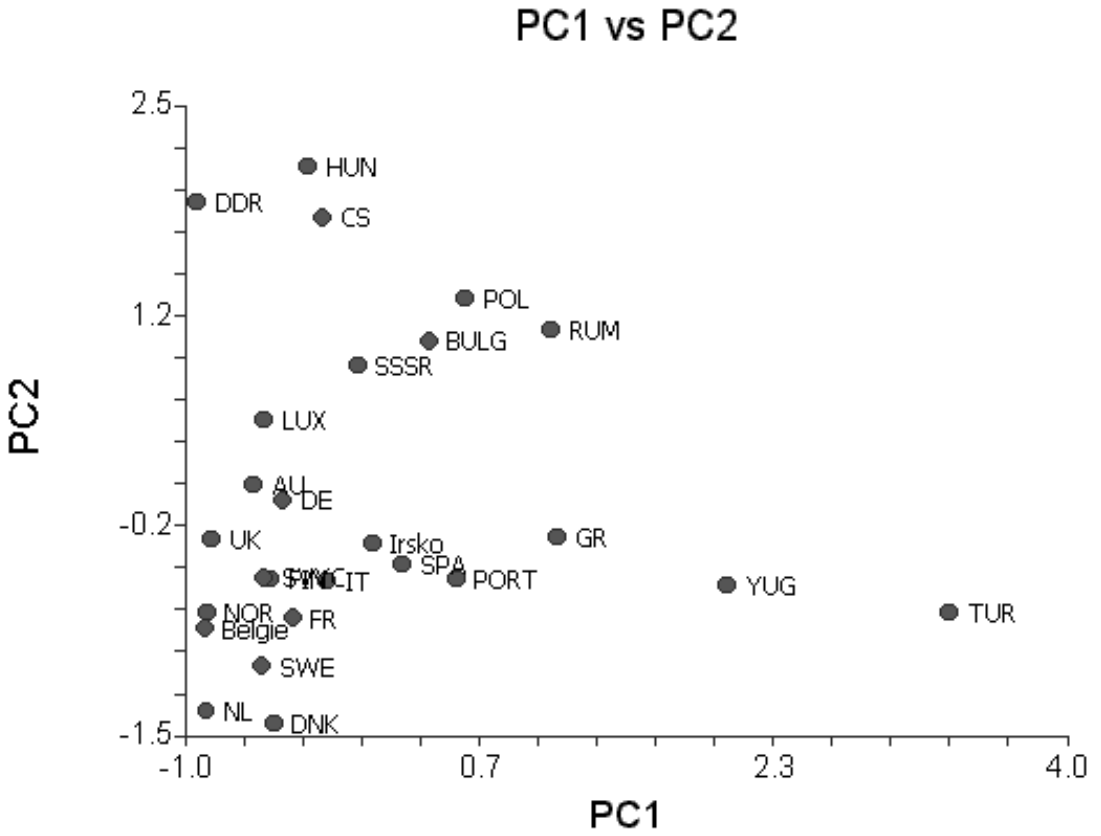
Vidíme, že tři vlastní čísla jsou větší než 1, čtvrté téměř rovno jedné. První dvě hlavní komponenty vysvětlují 62 % z celkové variability, první čtyři už 85 % celkové variability.

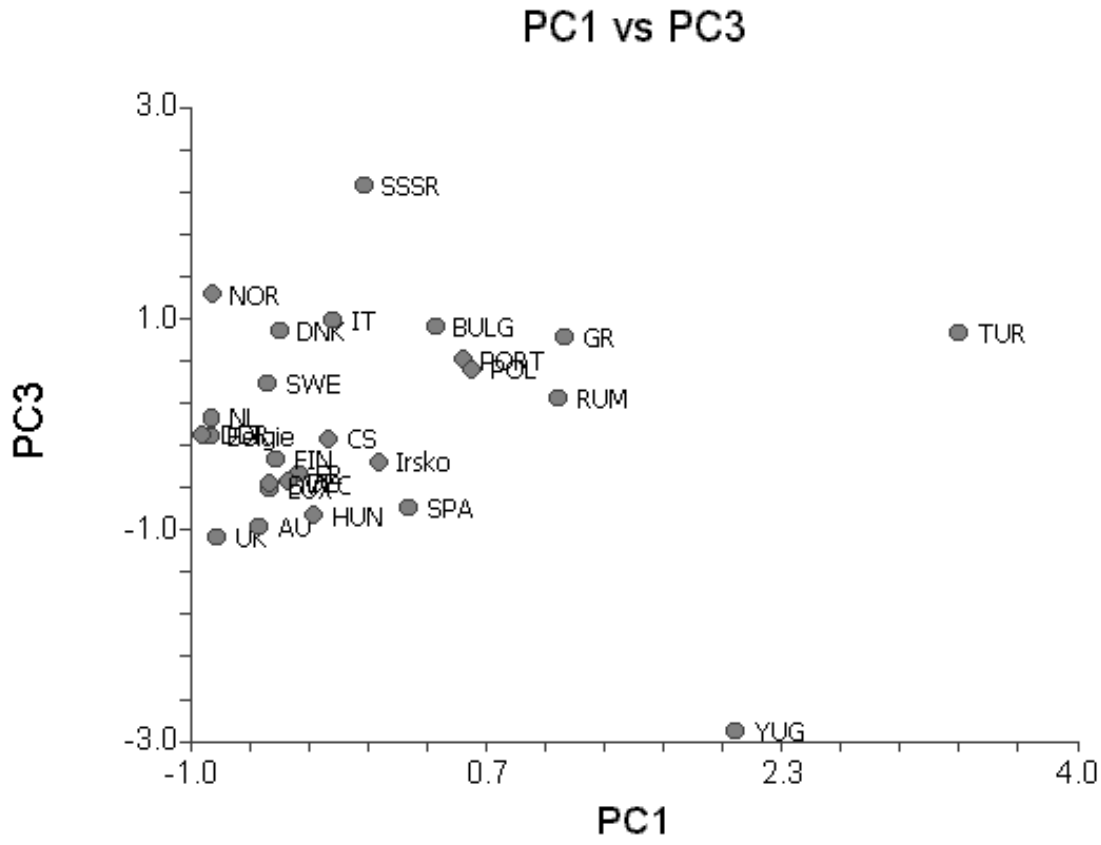
První čtyři vlastní vektory jsou v následující tabulce:

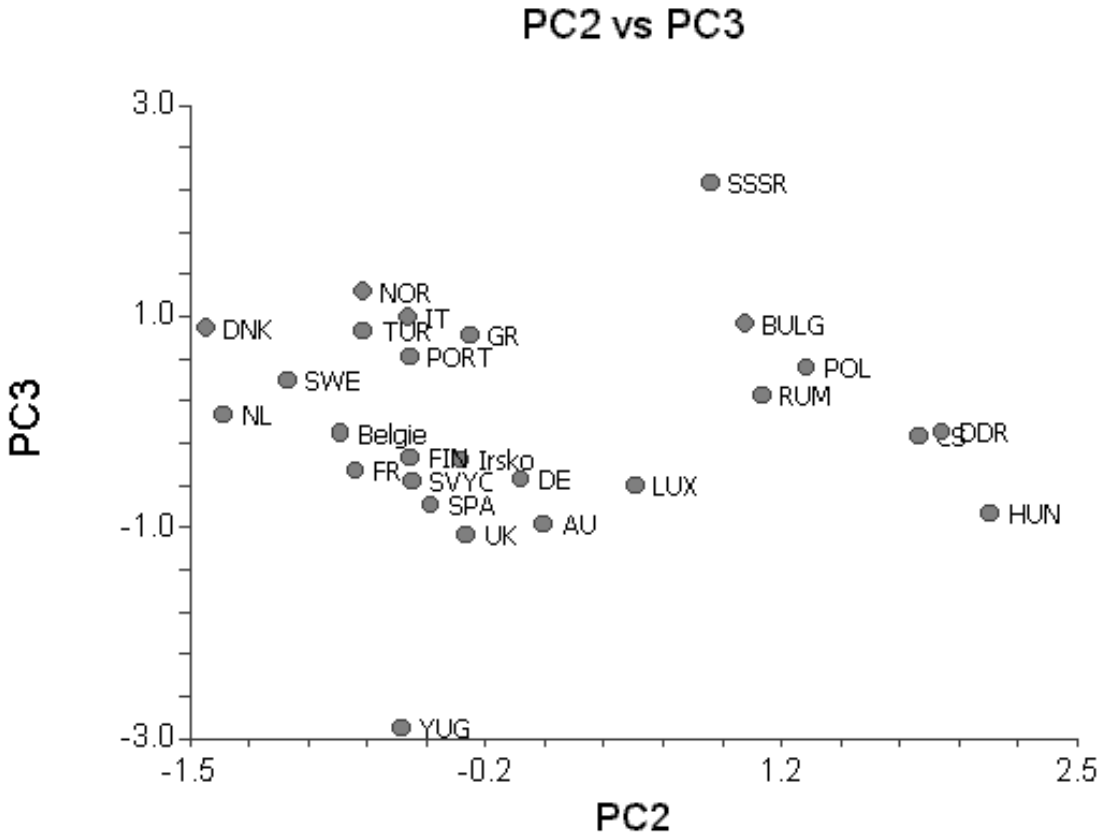
```
Eigenvectors
Variables   Factor1     Factor2     Factor3    Factor4
AGR          0.523791    0.053594    0.048674   0.028793
MIN          0.001323    0.617807   -0.201100   0.064085
MAN         -0.347495    0.355054   -0.150463  -0.346088
PS          -0.255716    0.261096   -0.561083   0.393309
CON         -0.325179    0.051288    0.153321  -0.668324
SER         -0.378920   -0.350172   -0.115096  -0.050157
FIN         -0.074374   -0.453698   -0.587361  -0.051567
SPS         -0.387409   -0.221521    0.311904   0.412230
TC          -0.366823    0.202592    0.375106   0.314372
```

Na dvourozměrných grafech v rovinách dvojic prvních tří hlavních komponent vidíme, že při takto podstatném snížení dimenze, ve kterých objekty zobrazujeme, lze sledovat podobnosti a odlišnosti zobrazených objektů. Zejména na obrázku prvních

dvou hlavních komponent jsou viditelné shluky objektů (zemí) v souladu s klasifikací nalezenou shlukovou analýzou.






PC2 vs PC3
PC3
PC2
3.0
1.0
-1.0
-3.0
-1.5
-0.2
1.2
2.5
SSSR
NOR
IT
DNK
TUR
GR
BULG
PORT
POL
SWE
RUM
NL
Belgie
DDR
FIN
Irsko
FR
SVYC
DE
LUX
SPA
HUN
AU
UK
YUG

12.5 Faktorová analýza

Faktorová analýza je jedna z metod redukce dimenze úlohy. Snaží se vysvětlit kovarianční strukturu (korelační matici) vektoru náhodných veličin, pomocí méně tzv. faktorů, tj. jakýchsi skrytých veličin, které nemůžeme nebo neumíme přímo měřit.

Uvažujme, že naměřená data jsou matice $\mathbf{X}$ typu $n \times p$ (n objektů, p veličin). Standardizovaná matice dat je matice $\mathbf{Z}$ opět typu $n \times p$, kde

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j}$$

$\bar{x_j}$, s_j jsou výběrový průměr a směrodatná odchylka j-té veličiny.

„Model“ faktorové analýzy můžeme pak zapsat

$$z_{ij} = a_{j1}f_{i1} + a_{j2}f_{i2} + \ldots + a_{jm}f_{im} + e_{ij}, \quad m < p$$

tj. naměřenou hodnotu j-té veličiny na i-tém objektu vysvětlujeme jako vážený součet m faktorů a nějaké složky, která těmito faktory vysvětlit nelze.

matice	$\mathbf{A}$	je typu	$p \times m$	faktorové zátěže (sycení, saturace, loadings)
matice	$\mathbf{F}$	je typu	$n \times m$	faktorové skóry
matice	$\mathbf{E}$	je typu	$n \times p$	

Maticově můžeme tedy model faktorové analýzy zapsat takto:

$$\mathbf{Z} = \mathbf{F}\mathbf{A}_T + \mathbf{E}$$

Předpokládejme, že faktory jsou ortonormální (nekorelované, jednotkové délky), tzn. $\mathbf{F}^T\mathbf{F} = \mathbf{I}$. Označme $\mathbf{U} = \mathbf{E}^T\mathbf{E}$. $\mathbf{U}$ je diagonální matice typu $p \times p$, tedy $\mathbf{U} = diag(u_1, u_2, \ldots, u_p)$, kde $u_j = \sum_{i=1}^{n} e_{ij}^2$, tj. variabilita j-té veličiny, kterou nelze vysvětlit faktory, tzv. specificita.

Pak

$$h_j = 1 - u_j = \sum_{k=1}^{m} a_{jk}^2$$

je tzv. komunalita j-té veličiny, tj. variabilita vysvětlitelná faktory.

Korelační matici můžeme vyjádřit jako

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{U}$$

Matice $\mathbf{A}\mathbf{A}^T$ vysvětluje korelační matici až na diagonální prvky, tam místo jedniček jsou komunality.

Faktorová analýza hledá model, aby

- příspěvek specifických faktorů (u_j) byl co nejmenší
- faktorové zátěže v absolutní hodnotě co nejbližší jedné nebo nule
- počet faktorů co nejmenší (podstatně menší než počet veličin)

Tyto požadavky jsou ve vzájemném rozporu a „umění“ faktorové analýzy spočívá v nalezení vhodného a přijatelného kompromisu:

extrakce faktorů - stanovit jejich počet, např. z vlastních čísel korelační matice

problém komunalit $R_j^2 \leq h_j \leq 1$, kde R_j^2 je koeficient mnohonásobné korelace (j-tá veličina na ostatních)

rotace faktorů tj. nalezení „jednoduché struktury“, aby faktorové zátěže v absolutní hodnotě byly co nejbližší jedné nebo nule

Ortogonální rotace znamená najít takovou transformaci matice $\mathbf{A}$ na $\mathbf{B} = \mathbf{AT}$, aby $\mathbf{B}$ splňovalo požadavek jednoduché struktury, $\mathbf{T}$ je ortogonální matice, tj. $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, takže $\mathbf{AA}^T = \mathbf{BB}^T$.

Je mnoho možných metod takové rotace faktorů, nejběžnější je tzv. VARIMAX, založená na maximalizaci výrazu

$$\frac{1}{p}\sum_{j=1}^{m}\sum_{i=1}^{p}(a_{ij}^2 - a_{\cdot j}^2)^2,$$

kde $a_{\cdot j}^2 = \frac{1}{p}\sum_{i=1}^{p} a_{ij}^2$, tj. průměr čtverců zátěží pro j-tý faktor.

Po rotaci můžeme nahlédnout, která veličina „patří“ kterému faktoru (faktorové zátěže v absolutní hodnotě blízké jedné), případně faktorové zátěže jednotlivých veličin vynést do grafů pro dvojice faktorů. Lze pak spočítat i faktorové skóry a na grafech faktorových skórů hledat, zda zobrazené objekty nevytvářejí nějaké shluky signalizující možný rozklad pozorovaných objektů do dvou či více podskupin.

Příklad 12.8 Data pro tento příklad jsou převzata z knihy [6]. Původní data byla výsledky dosažené ve výběru 220 chlapců v šesti předmětech - galštině, angličtině, dějepisu, aritmetice, algebře a geometrii. K dispozici pro analýzu máme výběrovou korelační matici (dolní trojúhelník, předměty jsou ve výše uvedeném pořadí):

```
1.00
0.44   1.00
0.41   0.35   1.00
0.29   0.35   0.16   1.00
0.33   0.32   0.19   0.59   1.00
0.25   0.33   0.18   0.47   0.46   1.00
```

Abychom si uvědomili rozdíly mezi analýzou hlavních komponent a faktorovou analýzou, uvádíme nejdříve stručné výsledky analýzy hlavních komponent, která vysvětluje celkový rozptyl, tedy hlavní diagonálu korelační matice.

```
Principal Components Report

Database     C:\avdat\subject.S0

Eigenvalues
                       Individual    Cumulative
No.     Eigenvalue     Percent       Percent    Scree Plot
1        2.728683       45.48         45.48     ||||||||||
2        1.128792       18.81         64.29     ||||
3        0.615291       10.25         74.55     |||
4        0.602809       10.05         84.59     |||
5        0.522514        8.71         93.30     ||
6        0.401910        6.70        100.00     ||


Factor Loadings
Variables     Factor1       Factor2
Gaelic       -0.660803     -0.444475
English      -0.688465     -0.289771
History      -0.516356     -0.639552
Arithmetic   -0.735620      0.417018
Algebra      -0.741868      0.372759
Geometry     -0.678168      0.354100
```

Dvě vlastní čísla jsou větší než jedna, pro faktorovou analýzu tedy zvolíme počet faktorů 2, rotaci *varimax* a dostaneme následující výsledky:

```
Factor Analysis Report

Database         C:\avdat\subject.S0

Eigenvalues after Varimax Rotation
                    Individual   Cumulative
No.    Eigenvalue   Percent      Percent     Scree Plot
1      1.596863      56.94        56.94      ||||||||||||
2      1.207981      43.08       100.02      |||||||||
3      0.050820       1.81       101.83      |
4      0.011910       0.42       102.26      |
5     -0.008657      -0.31       101.95      |
6     -0.054642      -1.95       100.00      |
```

Vlastní čísla se týkají matice $\mathbf{A}\mathbf{A}^T = \mathbf{R} - \mathbf{U}$ po rotaci, cílem faktorové analýzy je především vysvětlit korelační strukturu, tj. mimodiagonální prvky korelační matice. Celkový vysvětlený rozptyl je roven jen součtu komunalit.

Po rotaci jsou faktorové zátěže následující:

```
Factor Loadings after Varimax Rotation
Variables      Factor1       Factor2
Gaelic        -0.233132     -0.659253
English       -0.322810     -0.552071
History       -0.084713     -0.589192
Arithmetic    -0.765986     -0.170657
Algebra       -0.718105     -0.214689
Geometry      -0.573340     -0.214994
```

Absolutní hodnoty faktorových zátěží jsou znázorněny graficky:

```
Bar Chart of Absolute Factor Loadings
after Varimax Rotation

Variables      Factor1             Factor2
Gaelic         |||||               |||||||||||||
English        |||||||             ||||||||||||
History        ||                  ||||||||||||
Arithmetic     ||||||||||||||||    ||||
Algebra        ||||||||||||||||    |||||
```

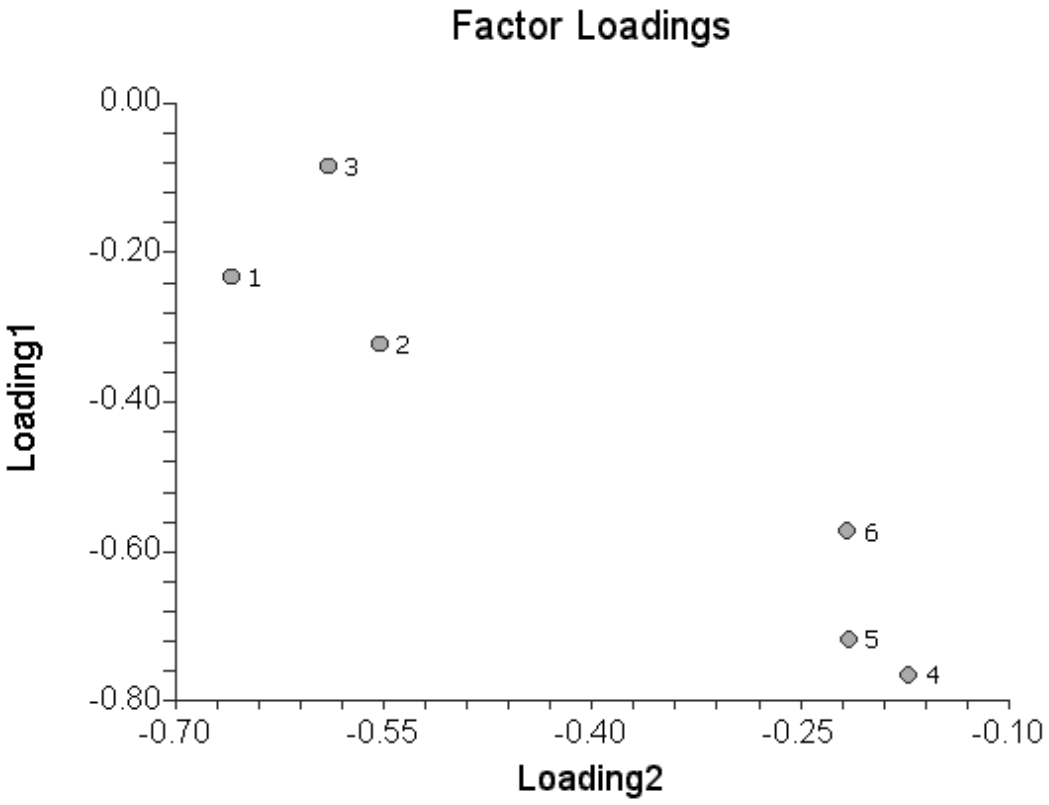
```
Geometry        ||||||||||||          |||||
```

Faktor 1 je sycen veličinami aritmetika, algebra a geometrie, faktor 2 veličinami galština, angličtina a dějepis.

Podíly faktorů na komunalitách ukazuje následující tabulka:

```
Communalities after Varimax Rotation
     Factors
Variables      Factor1       Factor2       Communality
Gaelic         0.054350      0.434614      0.488965
English        0.104206      0.304783      0.408989
History        0.007176      0.347147      0.354324
Arithmetic     0.586735      0.029124      0.615859
Algebra        0.515675      0.046091      0.561766
Geometry       0.328719      0.046222      0.374942
```

Vztah veličin k faktorům je graficky znázorněn v grafu faktorových zátěží, ve kterém vidíme shluk veličin 1, 2 a 3 s nízkými absolutními hodnotami zátěží prvního faktoru, které sytí především druhý faktor, a shluk veličin 4, 5 a 6 s nízkými absolutními hodnotami zátěží druhého faktoru, sytících především první faktor.



Korelační strukturu pozorovaných dat (studijních výsledků v šesti předmětech) lze tedy uspokojivě vysvětlit dvě faktory. První faktor vyjadřuje matematickou dispozici žáka, druhý dispozici jazykově-humanitní.

Shrnutí

- *úlohy řešené mnohorozměrnými metodami*
- *test shody vektorů středních hodnot*
- *metody klasifikace objektů, diskriminační analýza, logistická regrese*
- *metody shlukování, numerická podobnost, hierarchické a nehierarchické metody shlukování*
- *redukce dimenze, analýza hlavních komponent, faktorová analýza*

Kontrolní otázky

1. *V čem jsou shodné a v čem odlišné cíle diskriminační analýzy a shlukové analýzy?*
2. *Proč se při hledání lineární diskriminační funkce musí lišit střední hodnoty skupin?*
3. *Jaké jsou výhody a nevýhody lineární diskriminační funkce oproti jiným klasifikačním pravidlům (např. logistická regrese, neuronové sítě ap.)?*
4. *Jaké jsou rozdíly mezi hierarchickými a nehierarchickými metodami?*
5. *V čem se liší faktorová analýza od analýzy hlavních komponent?*
6. *Jak zjistíte souřadnice jednotlivých objektů (řádků datové matice) v rovině prvních dvou hlavních komponent?*
7. *Jak určit počet faktorů?*
8. *Co je to komunalita?*
9. *Co jsou faktorové zátěže? Co lze vyčíst z grafu faktorových zátěží?*
10. *Jakou část celkové variability vysvětluje první hlavní komponenta?*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.

13 Literatura

[1] Afifi A., Clark V.A., May S., Computer-Aided Multivariate Analysis, Chapman & Hall/CRC, 2004

[2] Anděl J.: Matematická statistika, SNTL, Praha, 1978

[3] Anděl J.: Statistické metody, Matfyzpress, Praha, 1993

[4] Antoch, J., Vorlíčková, D.: Vybrané metody statistické analýzy dat. Academia, Praha, 1992.

[5] Armitage P., Berry G.: Statistical Methods in Medical Research, Blackwell Sci Publ., 1994

[6] Bartholomew D.J., Steel F., Moustaki I., Galbraith J.I., The Analysis and Interpretation of Multivariate Data for Social Sciences, Chapman and Hall/CRC, 2002

[7] Cipra., T., Ekonometrie, MFF UK, SPN Praha, 1984

[8] Draper N. R., Smith H.: Applied Regression Analysis, John Wiley, 1998

[9] Hair J.F.jr, Black W.C., Babin B.J., Anderson R.E., Multivariate Data Analysis, Prentice Hall, 2010

[10] Havránek T.: Statistika pro biologické a lékařské vědy, Academia, Praha, 1993

[11] Havránek T. a kol.: Matematika pro biologické a lékařské vědy, Academia, Praha, 1981

[12] Hebák P. a kol. Vícerozměrné statistické metody (1), Informatorium, Praha, 2004

[13] Hebák P. a kol. Vícerozměrné statistické metody (2),(3), Informatorium, Praha, 2005

[14] Hintze, J., NCSS 8. NCSS, LLC. Kaysville, Utah, USA. www.ncss.com, 2012

[15] Kleinbaum D.G.: Logistic Regression: A Self-Learning Text, Springer-Verlag, New York Berlin Heidelberg, 1994

[16] Likeš J., Laga J., Základní statistické tabulky, SNTL, Praha, 1978

[17] Lukasová A., Šarmanová J., Metody shlukové analýzy, SNTL, Praha, 1981

[18] Manly, B. F. J., Multivariate Statistical Methods - A primer, Chapman and Hall/CRC, 1994

[19] McCullagh, P., Nelder, J.A.: Generalized Linear Model, 2nd Edition, Chapman and Hall, 1989

[20] Meloun M.,Militký J.: Statistické zpracování experimentálních dat, PLUS, 1994

[21] Tvrdík, J., Základy pravděpodobnosti a statistiky, OU, 2010

[22] Tvrdík, J., Analýza dat, OU, 2008

[23] Zvára, K., Regrese, Academia, Praha, 1989