Osnova přednášky Lineární diskriminační analýza

1. Motivace

Diskriminační analýza patří k vícerozměrným statistickým metodám a zabývá se klasifikací objektů do $r \geq 2$ skupin na základě znalosti vektorů pozorování těchto objektů.

Zakladatelem DA je R. A. Fisher

Řeší problém, jak získat jednu či více rovnic, které umožní klasifikovat objekty do skupin. Tyto rovnice se nazývají klasifikační neboli diskriminační funkce a kombinují jednotlivé proměnné a jejich váhy tak, aby bylo možné určit skupinu, do které klasifikovaný objekt s největší pravděpodobností patří.

2. Možnosti použití diskriminační analýzy

Technické obory:
Při kontrole jakosti či spolehlivosti lze ve výběrovém souboru výrobků změřit nějaké kvantitativní proměnné (např. rozměry, hmotnost, chemické složení apod.), pak výrobky podrobit zátěži a sledovat, zda tuto zátěž vydrží nebo ne. K predikci chování dalších výrobků při zátěži je skutečné zátěži nemusíme vystavovat, stačí, když provedeme potřebná měření kvantitativních proměnných.

Lékařství
Máme soubor pacientů, u nichž jsou diagnostikovány určité choroby. Pro každého pacienta máme k dispozici výsledky různých laboratorních testů. Pokud existuje souvislost mezi výsledky testů a diagnózou, může se lékař u nových pacientů rozhodovat pro určitou diagnózu (a tedy i způsob léčení) na základě výsledků testů.

Bankovnictví
Banka sleduje ve výběrovém souboru klientů, jak splácejí poskytnutý úvěr a kromě toho řadu dalších ukazatelů (věk, rodinný stav, výši příjmu, …). Následně na tomto základě může vyhodnocovat potenciální žadatele o úvěr jako více či méně důvěryhodné.

Archeologie
Při vykopávkách byly nalézány hroby s kostrami pravěkých lidí. Na základě nějakých charakteristických vlastností (délka určité kosti, úhly kostí na lebce,…) bylo možné další nalezené kostry zařadit k určitému historickému období, kultuře a rase.

3. LDA pro dvě skupiny objektů

3.1. Odvození bayesovského rozhodovacího pravidla

V 1. skupině je n_1 objektů, ve 2. skupině n_2 objektů. Každý objekt je charakterizován p-rozměrným vektorem pozorování $\mathbf{X} = (X_1, \ldots, X_p)'$.

Předpokládáme, že v h-té skupině má náhodný vektor $\mathbf{X}$ hustotu $\varphi_h(\mathbf{x})$, h = 1, 2.

Nechť H_h je jev „objekt patří do h-té skupiny“.

Apriorní pravděpodobnost $P(H_h)$ příslušnosti objektu k h-té skupině označíme π_h, h = 1, 2.

Známe-li u nějakého objektu vektor pozorování $\mathbf{x}$, můžeme podle Bayesova vzorce vypočítat aposteriorní pravděpodobnost příslušnosti objektu ke skupině:

$$P(H_h / \mathbf{X} = \mathbf{x}) = \frac{\pi_h \varphi_h(\mathbf{x})}{\pi_1 \varphi_1(\mathbf{x}) + \pi_2 \varphi_2(\mathbf{x})}, \quad h = 1, 2$$

Rozhodovací pravidlo: nový objekt zařadíme do té skupiny, u níž je aposteriorní pravděpodobnost větší.

Objekt s vektorem pozorování **x** zařadíme do 1. skupiny, když $\pi_1\varphi_1(\mathbf{x}) > \pi_2\varphi_2(\mathbf{x})$, jinak ho zařadíme do 2. skupiny.

Součin $\pi_h\varphi_h(\mathbf{x})$ se nazývá diskriminační skór pro h-tou skupinu.

Lze ukázat, že bayesovské rozhodovací pravidlo je optimální v tom smyslu, že minimalizuje celkovou pravděpodobnost mylné klasifikace.

3.2. Fisherova lineární diskriminační funkce pro dvě skupiny objektů

V lineární diskriminační analýze se předpokládá, že hustota v h-té skupině je normální a má parametry $\boldsymbol{\mu}_h$, $\boldsymbol{\Sigma}$, tj.

$$\varphi_h(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_h)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_h)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_h) \right), \quad h = 1, 2.$$

Lze odvodit, že **lineární diskriminační skór** pro h-tou skupinu (tzv. Andersonova diskriminační statistika) - má tvar $\lambda_h(\mathbf{x}) = \boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_h + \ln\pi_h$, h = 1, 2.

Objekt s vektorem pozorování **x** tedy zařadíme do 1. skupiny, když $\lambda_1(\mathbf{x}) > \lambda_2(\mathbf{x})$, jinak ho zařadíme do 2. skupiny.

Vzhledem k tomu, že máme jen dvě skupiny objektů, lze rozhodnutí o zařazení objektu do skupiny učinit na základě rozdílu

$$\lambda(\mathbf{x}) = \lambda_1(\mathbf{x}) - \lambda_2(\mathbf{x}) = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) + \ln \pi_1 - \ln \pi_2 .$$

Funkce $\lambda(\mathbf{x})$ se nazývá **Fisherova lineární diskriminační funkce**. Označíme-li

$$\boldsymbol{\beta}' = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1}, \gamma = -\frac{1}{2}\ \boldsymbol{\beta}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln \pi_1 - \ln \pi_2,$$

můžeme Fisherovu lineární diskriminační funkci psát ve tvaru

$\lambda(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + \gamma$.

Znamená to, že jsme našli takovou lineární kombinaci vektoru pozorování **x**, která nám umožní minimalizovat celkovou pravděpodobnost mylného zařazení objektu do skupiny. Objekt s vektorem pozorování **x** tedy zařadíme do 1. skupiny, když $\lambda(\mathbf{x}) > 0$, jinak ho zařadíme do 2. skupiny.

3.3. Modifikace pro případ neznámých parametrů

Při praktickém použití diskriminační analýzy většinou neznáme parametry $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$ ani apriorní pravděpodobnosti π_1, π_2. V takovém případě používáme odhady:

$\boldsymbol{\mu}_h \rightarrow \mathbf{M}_h$, h = 1, 2

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

$$\pi_h \rightarrow \frac{n_h}{n}, \; h = 1, 2.$$

Odhad Fisherovy lineární diskriminační funkce $\lambda(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + \gamma$:

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g$, kde

$\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}$, $g = -\frac{1}{2}\,\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2$.

3.4. Posouzení účinnosti diskriminace resubstituční metodou

Resubstituční metoda spočívá v uplatnění zkonstruovaného rozhodovacího pravidla na objekty se známou příslušností ke skupině. Uvažujeme postupně všechny tyto objekty a jejich zařazení podle rozhodovacího pravidla porovnáme se skutečnou příslušností ke skupině. Stanovíme podíl správně a mylně zařazených objektů.

skutečnost	zařazení		součet
	1. skupina	2. skupina	
1. skupina	n_{11}	n_{12}	$n_{1.} = n_1$
2. skupina	n_{21}	n_{22}	$n_{2.} = n_2$
součet	$n_{.1}$	$n_{.2}$	n

Podíl správně zařazených objektů:

$$\frac{n_{11} + n_{22}}{n}$$

Podíl mylně zařazených objektů:

$$\frac{n_{12} + n_{21}}{n}$$

3.5. Postup při lineární diskriminační analýze

1. Vzhledem k povaze úlohy určíme veličiny $X_1, ..., X_p$ a pořídíme $n_1 + n_2$ p-rozměrných pozorování tak, aby n_1 objektů pocházelo z 1. skupiny a n_2 objektů z 2. skupiny.

2. Na zvolené hladině významnosti α testujeme hypotézy o normalitě rozložení v obou skupinách a orientačně posoudíme linearitu vztahů mezi sledovanými proměnnými v obou skupinách.

3. Vypočteme odhady $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}$, p_1, p_2.

4. Na zvolené hladině významnosti α testujeme hypotézy o shodě variančních matic a vektorů středních hodnot v obou skupinách.

5. Vypočteme odhad $L(\mathbf{x})$ Fisherovy lineární diskriminační funkce. Objekt s vektorem pozorování $\mathbf{x}$ přiřadíme k 1. skupině, když $L(\mathbf{x}) > 0$, jinak ho přiřadíme ke 2. skupině.

6. Účinnost diskriminace posoudíme metodou resubstituce.

3.6. Příklad

Příklad je převzat z knihy Meloun M., Militký J., Hill, M.: Počítačová analýza vícerozměrných dat v příkladech. Academia Praha 2005.

Datový soubor lebky.sta obsahuje údaje o 32 lebkách nalezených na pohřebištích v Tibetu.

Sledují se tyto proměnné:

ID … identifikátor (1 pro lebky z okolí Sikkimu, 2 pro lebky z okolí Lhasy)

Ldelka … největší délka lebky (v mm)

Lsirka … největší horizontální šířka lebky (v mm)

Lvyska … výška lebky (v mm)

Ovyska … výška horní části obličeje (v mm)

Osirka … šířka obličeje mezi body lícních kostí (v mm)

Pro uvedená data sestrojte Fisherovu lineární diskriminační funkci, která pomocí veličin Ldelka, … Osirka umožní rozlišit lebky ze Sikkimu od lebek ze Lhasy.

Testování hypotézy o normalitě sledovaných proměnných v daných dvou skupinách pomocí Lilieforsovy varianty K-S testu a pomocí S-W testu:

Proměnná	Testy normality (lebky.sta) Zhrnout podmínku: ID=1				
	N	max D	Lilliefors p	W	p
Ldelka	13	0,149568	p > .20	0,971258	0,908822
Lsirka	13	0,188580	p > .20	0,946284	0,543198
Lvyska	13	0,196329	p < ,20	0,900168	0,134439
Ovyska	13	0,176191	p > .20	0,944919	0,523649
Osirka	13	0,148589	p > .20	0,954446	0,666891

Proměnná	Testy normality (lebky.sta) Zhrnout podmínku: ID=2				
	N	max D	Lilliefors p	W	p
Ldelka	19	0,121226	p > .20	0,946640	0,345905
Lsirka	19	0,099534	p > .20	0,973572	0,844925
Lvyska	19	0,116289	p > .20	0,969669	0,769812
Ovyska	19	0,149966	p > .20	0,965452	0,683230
Osirka	19	0,180030	p < ,10	0,873328	0,016463

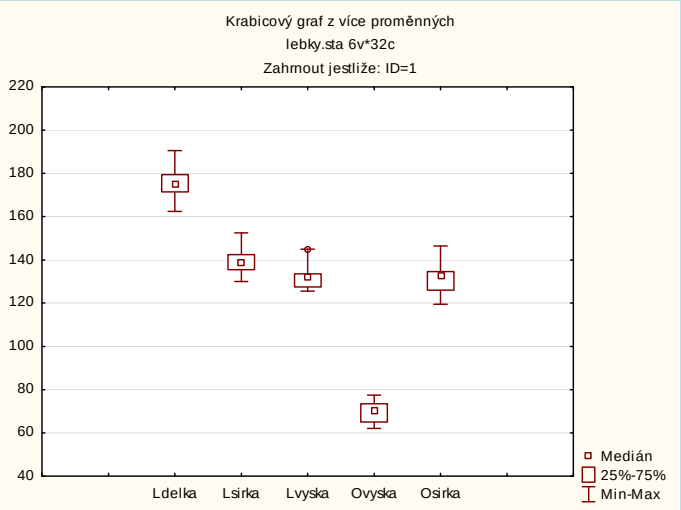
Vidíme, že ve 2. skupině zamítá S-W test hypotézu o normalitě proměnné Osirka na hladině významnosti 0,05, Lilieforsův test nikoli.

Nadále budeme data považovat za normálně rozložená.

Odhad vektorů středních hodnot v 1. skupině:

Proměnná	Popisné statistiky (lebky.sta) Zhrnout podmínku: ID=1 Průměr
Ldelka	175,1923
Lsirka	140,2692
Lvyska	132,3846
Ovyska	69,6923
Osirka	131,0000

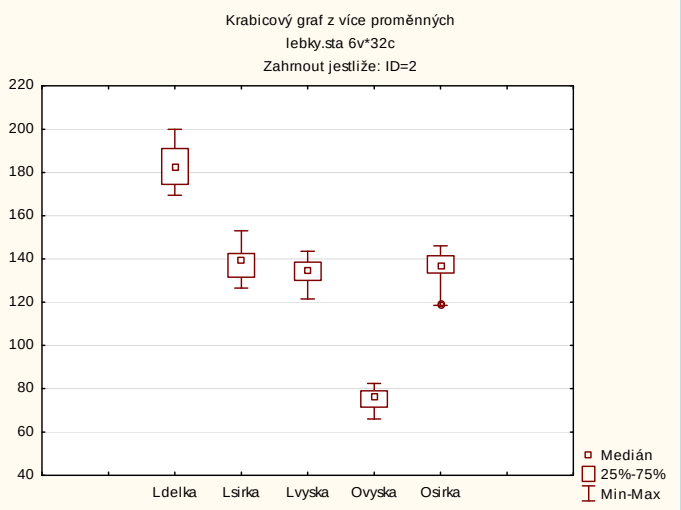
Krabicové grafy všech proměnných v 1. skupině:



Odhad vektorů středních hodnot ve 2. skupině:

Proměnná	Popisné statistiky (lebky.sta) Zhrnout podmínku: ID=2 Průměr
Ldelka	183,1842
Lsirka	138,2368
Lvyska	133,9211
Ovyska	75,1579
Osirka	135,5526

Krabicové grafy všech proměnných ve 2. skupině:



Odhad varianční matice $\mathbf{S}_1$

Proměnná	Kovariance (lebky.sta) Zhrnout podmínku: ID=1				
	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	50,02244	17,52724	25,02404	22,85577	20,04167
Lsirka	17,52724	47,31731	25,57532	-0,76442	32,35417
Lvyska	25,02404	25,57532	36,83974	5,89904	12,27083
Ovyska	22,85577	-0,76442	5,89904	22,64744	10,29167
Osirka	20,04167	32,35417	12,27083	10,29167	53,25000

Odhad varianční matice $\mathbf{S}_2$

Proměnná	Kovariance (lebky.sta) Zhrnout podmínku: ID=2				
	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	90,31140	7,3012	20,1126	31,64985	39,07310
Lsirka	7,30117	47,2880	-14,6747	10,68275	30,51462
Lvyska	20,11257	-14,6747	38,1462	8,66594	4,42105
Ovyska	31,64985	10,6827	8,6659	22,14035	25,13012
Osirka	39,07310	30,5146	4,4211	25,13012	51,05263

Odhad společné varianční matice **S**

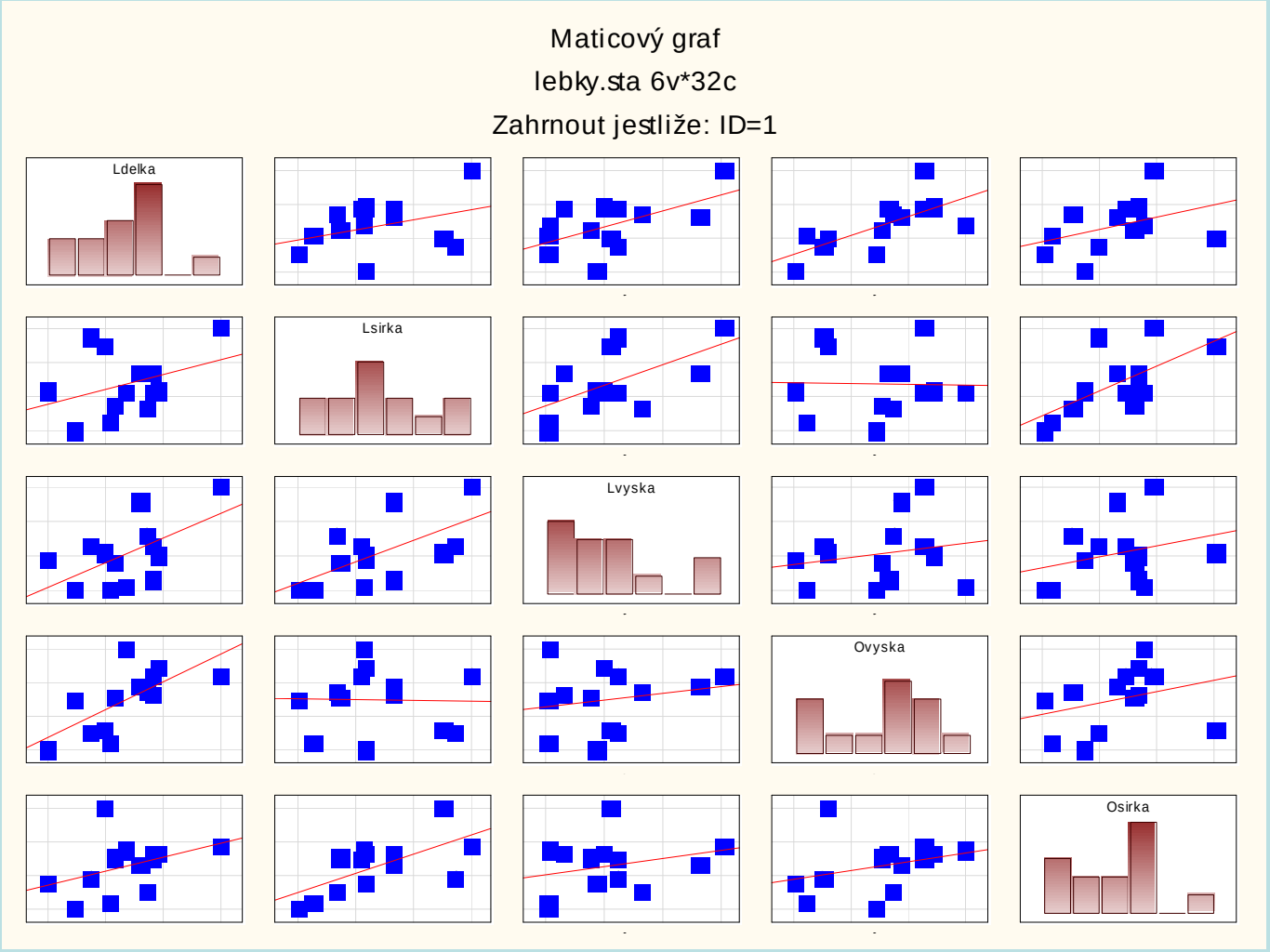
Proměnná	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	74,19582	11,3916	22,07716	28,13222	31,46053
Lsirka	11,3916	47,29973	1,425304	6,10388	31,25044
Lvyska	22,07716	1,425304	37,62362	7,559177	7,560965
Ovyska	28,13222	6,10388	7,559177	22,34318	19,19474
Osirka	31,46053	31,25044	7,560965	19,19474	51,93158

Boxův test shody variančních matic:

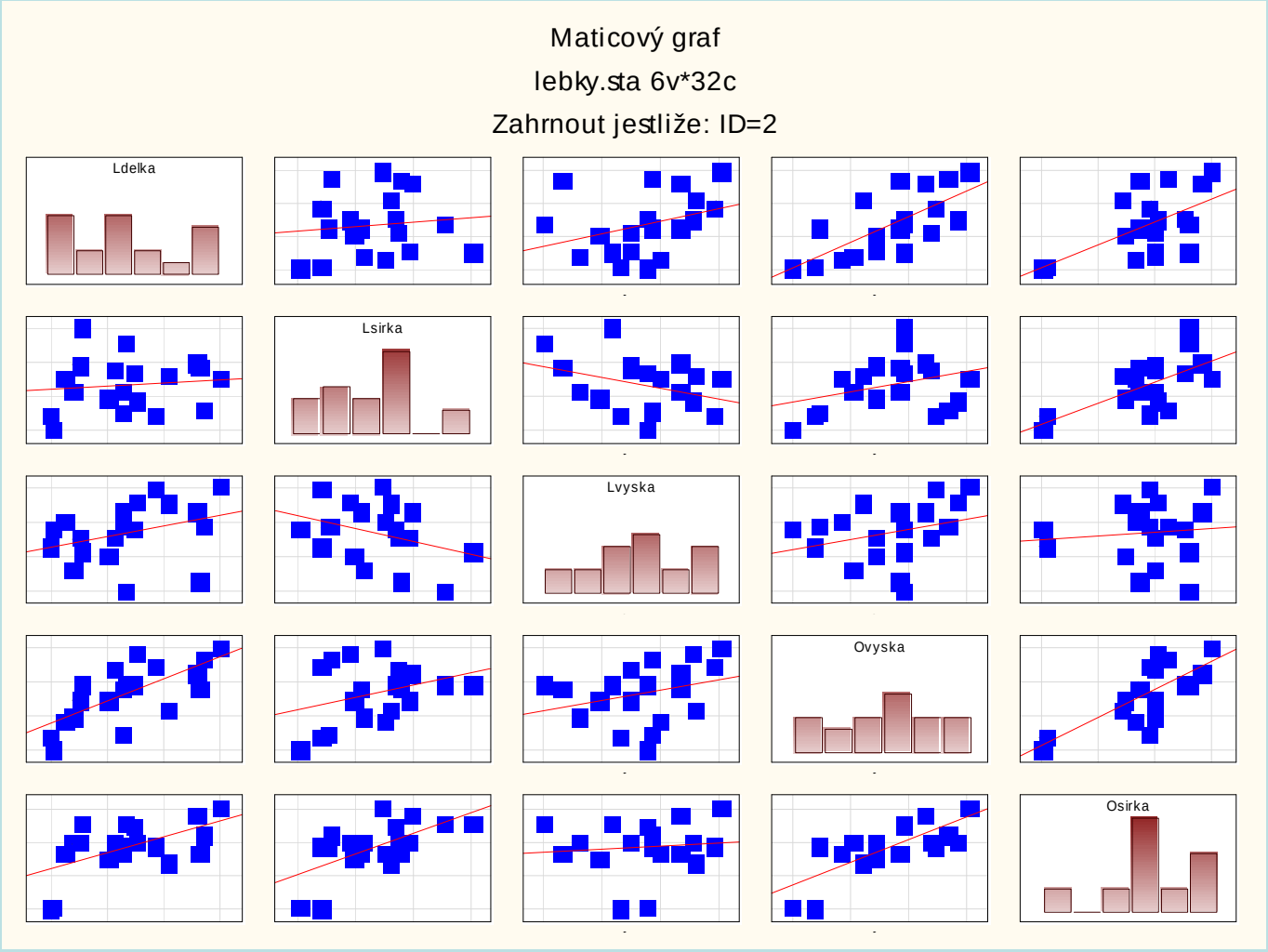
	Boxův M test (lebky.sta) Efekt: ID (Vypočteno pro všechny proměnné)			
	Boxovo M	Chí-kv.	sv	p
Boxovo M	22,65281	18,40191	15	0,242126

Hypotézu o shodě variančních matic nezamítáme na asymptotické hladině významnosti 0,05, protože p-hodnota = 0,242 je větší než 0,05.

Linearita vztahů mezi proměnnými ve skupině lebek ze Sikkimu



Linearita vztahů mezi proměnnými ve skupině lebek ze Lhasy



Test shody vektorů středních hodnot (Hotellingův T2 test):

Testová statistika $\frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2} (\mathbf{M}_1 - \mathbf{M}_2)' \mathbf{S}^{-1} (\mathbf{M}_1 - \mathbf{M}_2) = 2{,}4377$

Kvantil $F_{1-\alpha}(p, n_1+n_2-p-1) = F_{0,95}(5,26) = 2{,}5868$

Protože testová statistika se nerealizuje v kritickém oboru, nezamítáme na hladině významnosti 0,05 hypotézu o shodě vektorů středních hodnot $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$.

Proměnná	t-testy; grupováno: ID (lebky.sta) Skup. 1: 1; Skup. 2: 2 T2(celé případy 14,0638 F(5,26)=2,4377 p<,06127										
	Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr Rozptyly	p Rozptyly
Ldelka	175,1923	183,1842	-2,57771	30	0,015102	13	19	7,072654	9,503231	1,805418	0,298973
Lsirka	140,2692	138,2368	0,82101	30	0,418115	13	19	6,878758	6,876628	1,000620	0,970788
Lvyska	132,3846	133,9211	-0,69592	30	0,491837	13	19	6,069575	6,176261	1,035463	0,976491
Ovyska	69,6923	75,1579	-3,21246	30	0,003136	13	19	4,758932	4,705353	1,022903	0,938035
Osirka	131,0000	135,5526	-1,75517	30	0,089438	13	19	7,297260	7,145112	1,043041	0,909092

Individuální t-testy však prokázaly, že na hladině významnosti 0,05 se liší střední hodnoty proměnných Ldelka a Ovyska.

Význam jednotlivých proměnných v modelu

N=32	Výsledky diskriminační funkční analýzy (lebky.sta) Počet prom. v modelu: 5; grupovací: ID (2 skup) Wilk. lambda: ,68083 přibliž F (5,26)=2,4377 p< ,0613					
	Wilk. Lambda	Parc. Lambda	F na vyj (1,26)	p-hodn.	Toler.	1-toler. R^2
Ldelka	0,685248	0,993554	0,168690	0,684644	0,443129	0,556871
Lsirka	0,736910	0,923898	2,141624	0,155336	0,559228	0,440773
Lvyska	0,683292	0,996397	0,094009	0,761583	0,821817	0,178183
Ovyska	0,718740	0,947255	1,447718	0,239736	0,444337	0,555663
Osirka	0,697976	0,975435	0,654782	0,425752	0,383584	0,616416

V záhlaví této tabulky je uvedena Wilksova Lambda (na škále od 0 – nejlepší diskriminace do 1 – žádná diskriminace) a její přepočet na testovou statistiku F pro Hotellingův test shody vektorů středních hodnot (2,4377) a odpovídající p-hodnota (je blízká 0).
V 1. sloupci (Wilk. Lambda) jsou hodnoty Wilksovy Lambdy při vyřazení dané proměnné z modelu (vyšší hodnoty jsou lepší).
2. sloupec (Parc. Lambda) obsahuje unikátní příspěvky proměnných k diskriminaci.
Ve 3. sloupci jsou přepočty parciálních Lambda na testové statistiky a ve 4. sloupci pak odpovídající p-hodnoty. Podle p-hodnot u jednotlivých proměnných soudíme, že pro diskriminaci jsou důležité (nikoli však stratisticky významně na hladině významnosti 0,05) proměnné Lsirka a Ovyska.
5. sloupec (Tolerance) udává unikátní variabilitu proměnné nevysvětlenou ostatními proměnnými v modelu.
6. sloupec (1-toler., R^2) udává variabilitu proměnné vysvětlenou ostatními proměnnými.

Pro zájemce : $$\text{Parc. lambda} = \frac{\text{Wilk. lambda po vstupu prediktoru do modelu}}{\text{Wilk. lambda pred vstupem prediktoru do modelu}}$$

N=32	Wilk. Lambda	Parc. Lambda	F na vyj (1,26)	p-hodn.	Toler.	1-toler. R^2
	Výsledky diskriminační funkční analýzy (lebky.sta) Počet prom. v modelu: 5; grupovací: ID (2 skup) Wilk. lambda: ,68083 přibliž F (5,26)=2,4377 p< ,0613					
Ldelka	0,685248	0,993554	0,168690	0,684644	0,443129	0,556871
Lsirka	0,736910	0,923898	2,141624	0,155336	0,559228	0,440773
Lvyska	0,683292	0,996397	0,094009	0,761583	0,821817	0,178183
Ovyska	0,718740	0,947255	1,447718	0,239736	0,444337	0,555663
Osirka	0,697976	0,975435	0,654782	0,425752	0,383584	0,616416

Prediktor Ldelka: $$\text{Parc. lambda} = \frac{\text{Wilk. lambda po vstupu prediktoru do modelu}}{\text{Wilk. lambda pred vstupem prediktoru do modelu}} = \frac{0{,}68083}{0{,}685248} = 0{,}99355$$

Prediktor Lsirka: $$\text{Parc. lambda} = \frac{\text{Wilk. lambda po vstupu prediktoru do modelu}}{\text{Wilk. lambda pred vstupem prediktoru do modelu}} = \frac{0{,}68083}{0{,}73691} = 0{,}923898$$

Prediktor Lvyska: $$\text{Parc. lambda} = \frac{\text{Wilk. lambda po vstupu prediktoru do modelu}}{\text{Wilk. lambda pred vstupem prediktoru do modelu}} = \frac{0{,}68083}{0{,}683292} = 0{,}996397$$

Prediktor Ovyska: $$\text{Parc. lambda} = \frac{\text{Wilk. lambda po vstupu prediktoru do modelu}}{\text{Wilk. lambda pred vstupem prediktoru do modelu}} = \frac{0{,}68083}{0{,}71874} = 0{,}947255$$

Pediktor Osirka: $$\text{Parc. lambda} = \frac{\text{Wilk. lambda po vstupu prediktoru do modelu}}{\text{Wilk. lambda pred vstupem prediktoru do modelu}} = \frac{0{,}68083}{0{,}697976} = 0{,}975435$$

Mahalanobisova vzdálenost v diskriminační analýze

Používá se pro popis vzájemných vzdáleností centroidů jednotlivých skupin.

Vzdálenosti mezi skupinami:

	Mahalanobisovy vzdálenosti^2 (lebky.sta)	
ID	Sikkim	Lhasa
Sikkim	0,000000	1,822037
Lhasa	1,822037	0,000000

p-hodnoty pro testy hypotéz, že vzdálenosti jsou nulové:

	p-hodnot (lebky.sta)	
ID	Sikkim	Lhasa
Sikkim		0,061268
Lhasa	0,061268	

Lze také získat Mahalanobisovy vzdálenosti jednotlivých objektů od centroidů skupin, zde jsou uvedeny tyto vzdálenosti pro prvních 6 lebek:

	Mahalanobisovy vzdálenosti (lebky.sta) Nesprávná klasifikace je označena *		
Případ	**Pozorova Klasif.**	Sikkim p=,40625	Lhasa p=,59375
1	Sikkim	9,48563	11,81986
2	Sikkim	5,51987	9,05645
3	Sikkim	5,35887	6,73857
4	Sikkim	5,81074	13,00539
*5	Sikkim	6,71734	4,29486
6	Sikkim	3,47094	5,60249

Stanovení odhadu Fisherovy lineární diskriminační funkce:

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g$, kde $\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}$, $g = -\frac{1}{2}\,\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2$.

Odhad vektoru středních hodnot v 1. skupině:

	Popisné statistiky (lebky.sta) Zhrnout podmínku: ID=1	
Proměnná	N platných	Průměr
Ldelka	13	175,1923
Lsirka	13	140,2692
Lvyska	13	132,3846
Ovyska	13	69,6923
Osirka	13	131,0000

Odhad vektoru středních hodnot ve 2. skupině:

	Popisné statistiky (lebky.sta) Zhrnout podmínku: ID=2	
Proměnná	N platných	Průměr
Ldelka	19	183,1842
Lsirka	19	138,2368
Lvyska	19	133,9211
Ovyska	19	75,1579
Osirka	19	135,5526

Odhad společné varianční matice **S**:

Proměnná	Ldelka	Lsirka	Lwyska	Ovyska	Osirka
Ldelka	74,19582	11,3916	22,07716	28,13222	31,46053
Lsirka	11,3916	47,29973	1,425304	6,10388	31,25044
Lwyska	22,07716	1,425304	37,62362	7,559177	7,560965
Ovyska	28,13222	6,10388	7,559177	22,34318	19,19474
Osirka	31,46053	31,25044	7,560965	19,19474	51,93158

Odhady apriorních pravděpodobností:

$p_1 = \frac{n_1}{n} = \frac{13}{32} = 0{,}406,\ p_2 = \frac{n_2}{n} = \frac{19}{32} = 0{,}594$

Po dosazení dostaneme:

$\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1} = (-0{,}0335,\ 0{,}1282\ \ 0{,}0258\ \ -0{,}1742\ \ -0{,}0839)$

$g = -\frac{1}{2}\,\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2 = 8{,}1304$

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g = -0{,}033\text{Ldelka} + 0{,}1282\text{Lsirka} + 0{,}0258\text{Lvyska} - 0{,}1742\text{Ovyska} - 0{,}0839\text{Osirka} + 8{,}1304$

Získání odhadu Fisherovy lineární diskriminační funkce ve STATISTICE:

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných Ldelka až Osirka – OK – OK – na záložce Klasifikace zvolíme Klasifikační funkce. Do výstupní tabulky přidáme novou proměnnou, do jejíhož Dlouhého jména napíšeme =v1-v2

	Klasifikační funkce; grupovací : ID (lebky.sta		
Proměnná	G_1:1 p=,40625	G_2:2 p=,59375	NProm =v1-v2
Ldelka	1,168	1,202	-0,03346
Lsirka	2,820	2,692	0,128157
Lvyska	2,748	2,722	0,025791
Ovyska	0,280	0,454	-0,17415
Osirka	-0,385	-0,302	-0,0839
Konstant	-467,373	-475,503	8,130393

Posouzení účinnosti diskriminace resubstituční metodou:

Klasifikační matice:

	Klasifikační matice (lebky.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace		
Skup.	% správnýc	Sikkim p=,40625	Lhasa p=,59375
Sikkim	69,23077	9	4
Lhasa	84,21053	3	16
Celkem	78,12500	12	20

Podíl správně zařazených objektů:

$$\frac{n_{11}+n_{22}}{n}=\frac{9+16}{32}=0{,}781$$

Podíl mylně zařazených objektů:

$$\frac{n_{12}+n_{21}}{n}=\frac{4+3}{32}=0{,}219$$

Pro určení chybně zařazených případů zvolíme na záložce Klasifikace možnost Klasifikace případů. Zjistíme, že v 1. skupině došlo k mylnému zařazení u lebek č. 5, 8, 9 a 13, ve 2. skupině u lebek číslo 15, 16, 17.

Porovnání s náhodnou klasifikací:

Odhad celkové pravděpodobnosti mylné klasifikace je

$$2p_1(1-p_1)=2\cdot\frac{13}{32}\cdot\frac{19}{32}=0{,}4824.$$

Použitím diskriminační analýzy jsme tedy dosáhli značného zlepšení, pravděpodobnost mylné klasifikace klesla na 0,22.

Výpočet pomocí systému R

Načteme data:

```
data<-read.table('lebky.txt',sep=',',header=T)
```

Proměnnou ID zavedeme jako faktor:

```
ID<-factor(lebky$ID,labels=c('Sikkim','Lhasa'))
```

Zjistíme rozsahy 1. a 2. skupiny:

```
table(ID)
ID
Sikkim  Lhasa
    19     13
```

Vytvoříme datový soubor pro 1. skupinu a pro 2. skupinu:

```
data1<-lebky[1:13,2:6]
data2<-lebky[14:32,2:6]
```

Zjistíme průměry všech proměnných v 1. a 2. skupině:

```
colMeans(data1)
   Ldelka    Lsirka    Lvyska    Ovyska    Osirka
175.19231 140.26923 132.38462  69.69231 131.00000
colMeans(data2)
   Ldelka    Lsirka    Lvyska    Ovyska    Osirka
183.18421 138.23684 133.92105  75.15789 135.55263
```

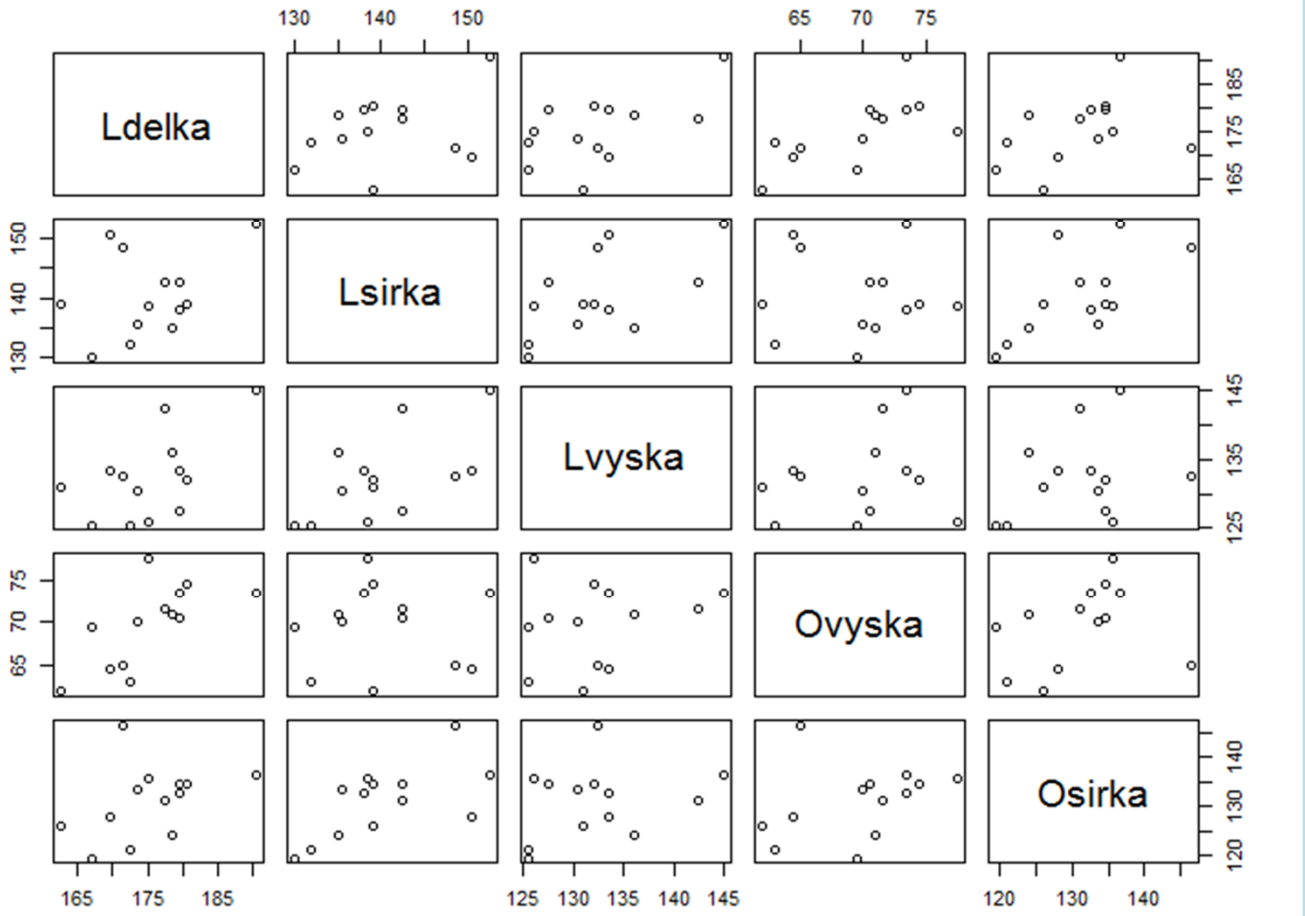
Vypočteme varianční matice v 1. a 2. skupině:

```
cov(data1)
         Ldelka     Lsirka    Lvyska     Ovyska   Osirka
Ldelka 50.02244 17.5272436 25.024038 22.8557692 20.04167
Lsirka 17.52724 47.3173077 25.575321 -0.7644231 32.35417
Lvyska 25.02404 25.5753205 36.839744  5.8990385 12.27083
Ovyska 22.85577 -0.7644231  5.899038 22.6474359 10.29167
Osirka 20.04167 32.3541667 12.270833 10.2916667 53.25000
```

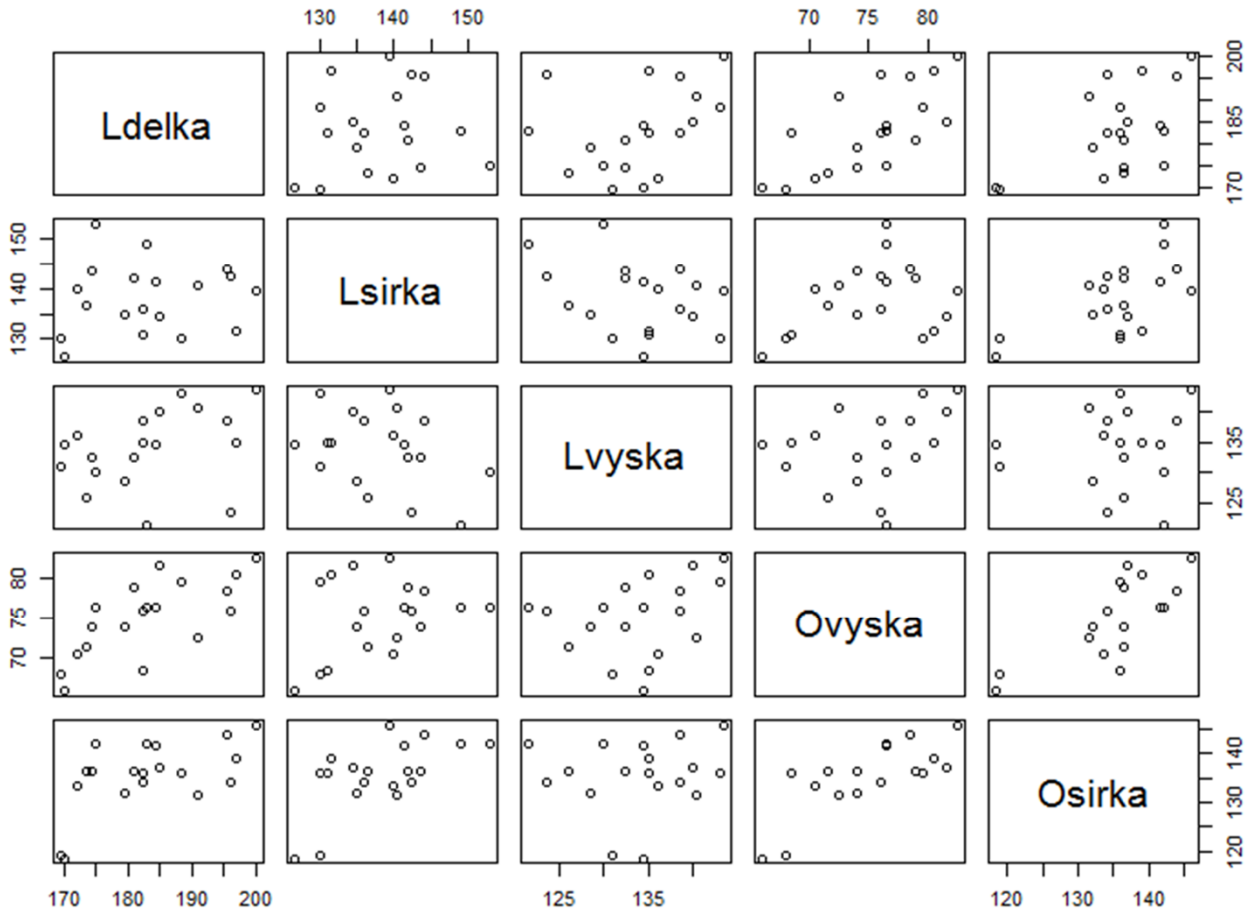
```
cov(data2)
         Ldelka    Lsirka     Lvyska    Ovyska    Osirka
Ldelka 90.31140   7.30117  20.112573 31.649854 39.073099
Lsirka  7.30117  47.28801 -14.674708 10.682749 30.514620
Lvyska 20.11257 -14.67471  38.146199  8.665936  4.421053
Ovyska 31.64985  10.68275   8.665936 22.140351 25.130117
Osirka 39.07310  30.51462   4.421053 25.130117 51.052632
```

Orientačně ověříme linearitu vztahů v 1. a 2. skupině:

```
plot(data1)
```



```
plot(data2)
```


Ldelka
Lsirka
Lvyska
Ovyska
Osirka
130 140 150
70 75 80
170 185 200
130 140 150
125 135
70 75 80
120 130 140
170 180 190 200
125 135
120 130 140

Dále ověříme vícerozměrnou normalitu dat v 1. a 2. skupině:

```
HZ.test(data1)
            Henze-Zirkler test for Multivariate Normality
  data : data1
  HZ               : 0.7604069
  p-value          : 0.386167

  Result  : Data are multivariate normal (sig.level = 0.05)
```

```
HZ.test(data2)
            Henze-Zirkler test for Multivariate Normality
  data : data2
  HZ               : 0.7842975
  p-value          : 0.4063919

  Result  : Data are multivariate normal (sig.level = 0.05)
```

Provedeme Boxův test shody variančních matic:

```
library(biotools)
boxM(lebky[,2:6],grouping=ID)

      Box's M-test for Homogeneity of Covariance Matrices

data:  lebky[, 2:6]
Chi-Sq (approx.) = 18.402, df = 15, p-value = 0.2421
```

Nyní pomocí Hotellingova T2 testu otestujeme shodu vektorů středních hodnot v obou skupinách:

```
library(ICSNP)

HotellingsT2(data1, data2)

      Hotelling's two sample T2-test

data:  data1 and data2
T.2 = 2.4377, df1 = 5, df2 = 26, p-value = 0.06127
alternative hypothesis: true location difference is not equal to c(0,0,0,0,0)
```

Sestrojíme odhad lineární diskriminační funkce:

```
library(MASS)
> lebky.lda<-lda(ID~Ldelka+Lsirka+Lvyska+Ovyska+Osirka,data=lebky)
> lebky.lda
Call:
lda(ID ~ Ldelka + Lsirka + Lvyska + Ovyska + Osirka, data = lebky)

Prior probabilities of groups:
  Lhasa  Sikkim
0.59375 0.40625

Group means:
         Ldelka   Lsirka   Lvyska   Ovyska   Osirka
Lhasa  183.1842 138.2368 133.9211 75.15789 135.5526
Sikkim 175.1923 140.2692 132.3846 69.69231 131.0000

Coefficients of linear discriminants:
                LD1
Ldelka -0.02478507
Lsirka  0.09494291
Lvyska  0.01910672
Ovyska -0.12901769
Osirka -0.06215888
```

Pomocí funkce predict získáme klasifikační tabulku. Zařazení objektu do skupiny je provedeno na základě vyšší aposteriorní pravděpodobnosti.

```
(tab<-table(fitted$class,lebky$ID))
```

```
         Lhasa Sikkim
  Lhasa     16      4
  Sikkim     3      9
```

Vypočítáme podíl správně zařazených lebek:

```
sum(diag(tab))/sum(tab)
```

```
[1] 0.78125
```

Správně bylo zařazeno 78,1 % lebek, špatně 21,9 %.

4. Výběr proměnných pro klasifikaci krokovou metodou

Kroková metoda postupně vyhledává nejvhodnější soubor proměnných pro diskriminaci. Používá se buď jako dopředná nebo jako zpětná.

Význam jednotlivých proměnných pro diskriminaci se k každém kroku zkoumá pomocí zaváděcího a odstraňovacího kritéria.
Vybírání proměnných či jejich odstraňování skončí, když žádné další proměnné nesplňují zaváděcí nebo odstraňovací kritérium.

Upozornění: Před zařazením j-té proměnné do modelu se stanoví její tolerance $1-R_j^2$ (R_j^2 je čtverec vícenásobného koeficientu korelace, tj. koeficientu, který měří těsnost lineární závislosti veličiny X_j na ostatních veličinách). Tolerance je implicitně nastavená na 0,01.

Příklad: Použijte krokovou dopřednou (a poté zpětnou) metodu pro zařazování lebek do dvou skupin.

Řešení:

Výsledky dopředné metody:

	Výsledky diskriminační funkční analýzy (lebky.sta) krok 2, poč. prom. v modelu: 2; grupovací: ID (2 skup) Wilk. lambda: ,70717 přibliž F (2,29)=6,0041 p< ,0066					
N=32	Wilk. Lambda	Parc. Lambda	F na vyj (1,29)	p-hodn.	Toler.	1-toler. R^2
Ovyska	0,978025	0,723064	11,10709	0,002359	0,964746	0,035254
Lsirka	0,744049	0,950441	1,51217	0,228692	0,964746	0,035254

Výsledky zpětné metody:

	Výsledky diskriminační funkční analýzy (lebky.sta) krok 4, poč. prom. v modelu: 1; grupovací: ID (2 skup) Wilk. lambda: ,74405 přibliž F (1,30)=10,320 p< ,0031					
N=32	Wilk. Lambda	Parc. Lambda	F na vyj (1,30)	p-hodn.	Toler.	1-toler. R^2
Ovyska	1,000000	0,744050	10,31990	0,003136	1,000000	0,00

Vidíme, že dopředná metoda skončila po dvou krocích a vybrala proměnné Ovyska, Lsirka.

Odhad Fisherovy lineární diskriminační funkce:

$L(\mathbf{x}) = -0{,}2657\text{Ovyska} + 0{,}0773\text{Lsirka} + 8{,}1071$
Úspěšnost klasifikace je 68,8 %.

Použijeme-li krokovou zpětnou metodu, je po 4 krocích vybrána pouze proměnná Ovyska:
Odhad Fisherovy lineární diskriminační funkce:

$L(\mathbf{x}) = -0{,}2446\text{Ovyska} + 17{,}371$
Účinnost diskriminace zůstala stejná jako u dopředné metody, tj. 68,8 %.

5. Lineární diskriminační analýza pro $r \geq 3$ skupin

5.1. Pravidlo pro zařazení objektu do skupiny

V h-té skupině je n_h objektů, h = 1, …, r. Každý objekt je charakterizován p-rozměrným vektorem pozorování $\mathbf{X} = (X_1, \ldots, X_p)'$.

Předpokládáme, že ve všech r skupinách se vektory pozorování řídí p-rozměrným normálním rozložením, varianční matice jednotlivých skupin jsou shodné a vztahy mezi sledovanými p proměnnými jsou přibližně lineární.

Lineární diskriminační skór pro h-tou skupinu (Andersonova diskriminační statistika) má tvar:

$$\lambda_h(\mathbf{x}) = \boldsymbol{\mu}_h' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_h' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_h + \ln \pi_h, \; h = 1, \ldots, r$$

Její odhad získáme dosazením $\mathbf{M}_h$, $\mathbf{S}$ a p_h:

$$L_h(\mathbf{x}) = \mathbf{M}_h' \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{M}_h' \mathbf{S}^{-1} \mathbf{M}_h + \ln p_h$$

Objekt neznámého původu, jehož vektor pozorování je $\mathbf{x}$, bude zařazen do skupiny s nejvyšší hodnotou $L_h(\mathbf{x})$.

5.2. Příklad

V souboru 50 rodin byly zjišťovány tyto údaje:

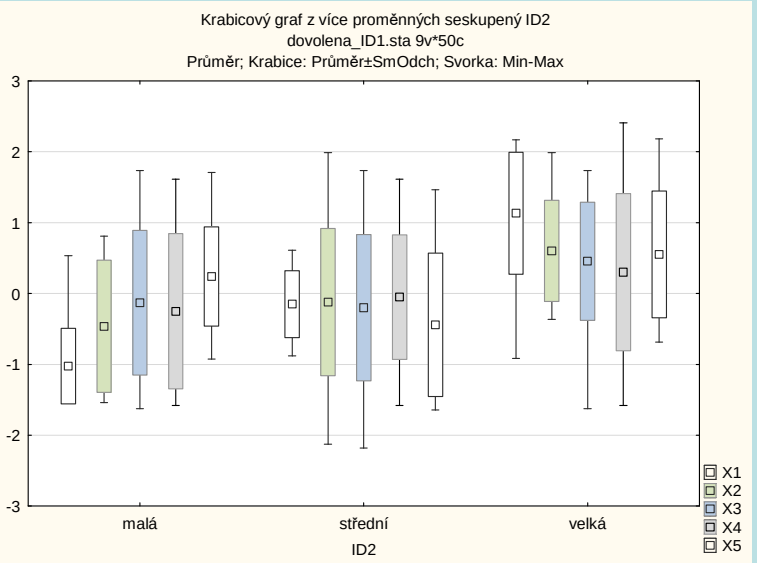
- jakou kategorizovanou částku je rodina ochotna vydat za dovolenou (veličina ID, nabývá variant „malá“ – 1, „střední“ – 2, „velká“ – 3)
- roční příjem v tisících dolarů (veličina X_1)
- postoj k cestování (veličina X_2, devítibodová škála, 1 = naprosto odmítavý, 9 = veskrze kladný)
- význam přičítaný rodinné dovolené (veličina X_3, devítibodová škála, 1 = nejnižší, 9 = nejvyšší)
- počet členů rodiny (veličina X_4)
- věk nejstaršího člena rodiny (veličina X_5).

Soubor rodin roztřiďte do tří skupin podle toho, jak velkou částku je rodina ochotna vydat za dovolenou.

Řešení:

Posouzení úrovně a variability proměnných X_1, …, X_5 v daných třech skupinách

Proměnná	ID2	N platných	Průměr	Sm.odch.
X1	malá	12	38,1	6,16
X2	malá	12	3,8	1,59
X3	malá	12	4,7	1,83
X4	malá	12	3,7	1,37
X5	malá	12	51,8	5,85
X1	střední	24	48,2	5,46
X2	střední	24	4,4	1,77
X3	střední	24	4,5	1,84
X4	střední	24	3,9	1,10
X5	střední	24	46,0	8,46
X1	velká	14	63,0	9,94
X2	velká	14	5,6	1,22
X3	velká	14	5,7	1,49
X4	velká	14	4,4	1,39
X5	velká	14	54,4	7,48



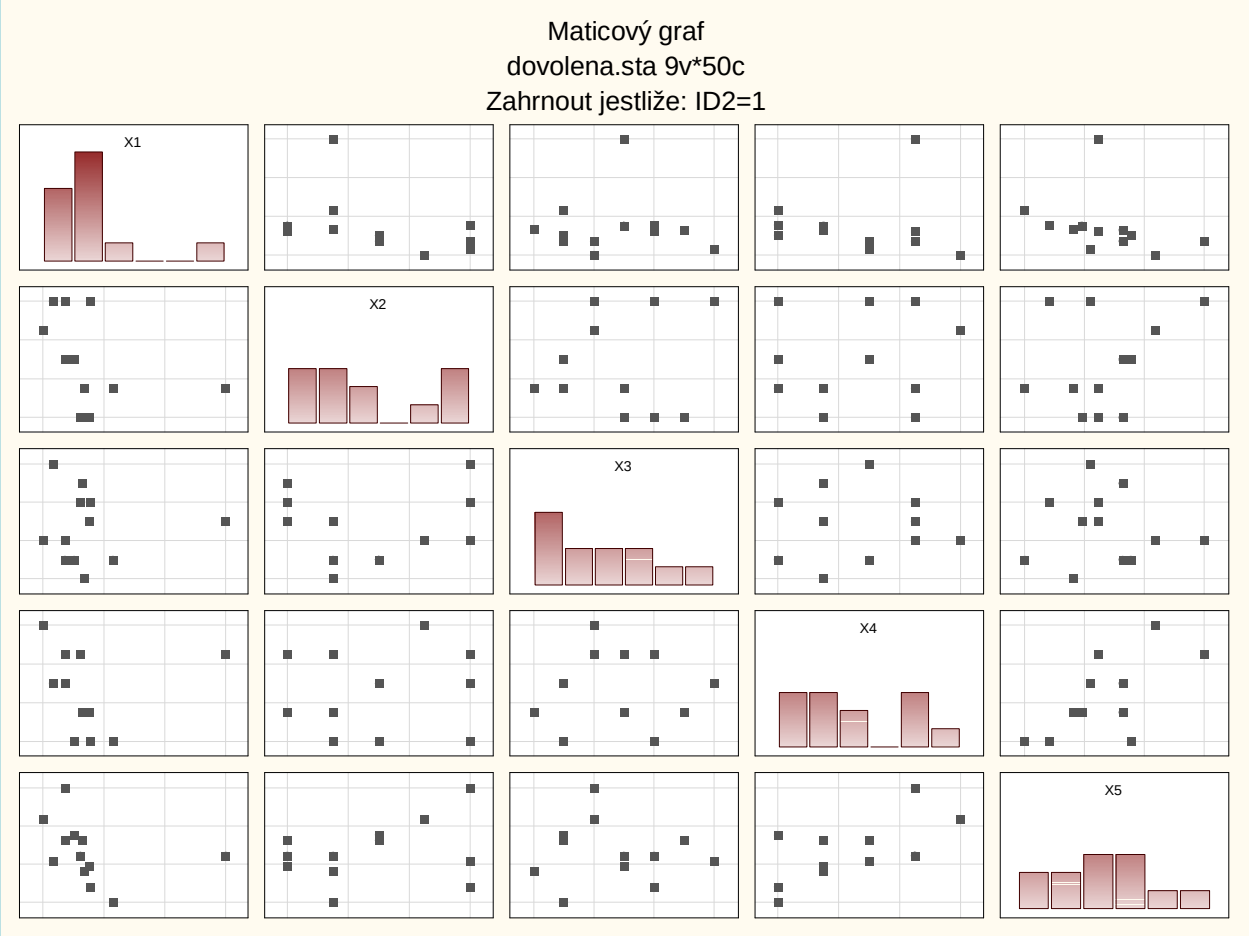
Ověření normality proměnných X_1, …, X_5 v daných třech skupinách

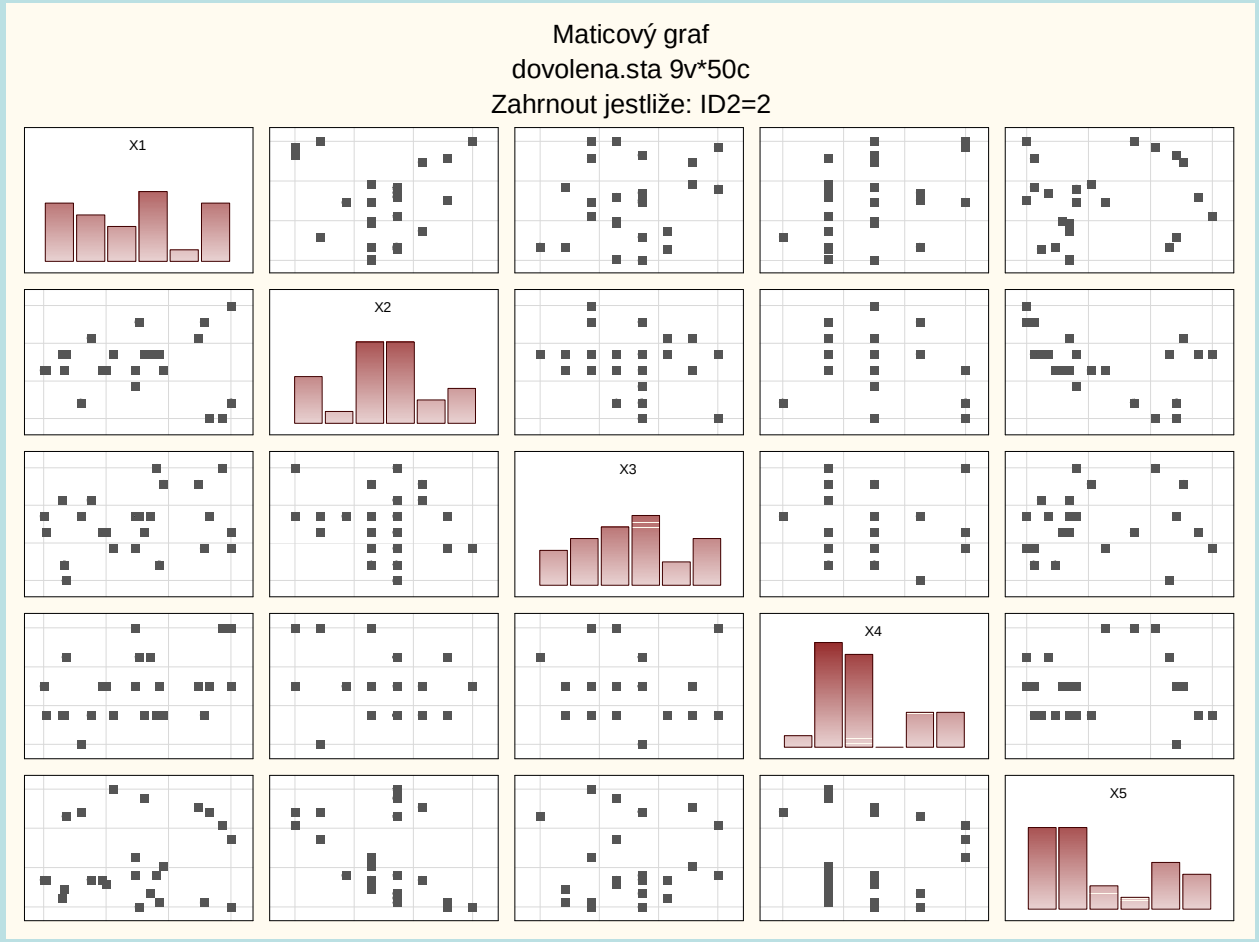
Proměnná	Souhrnné výsledky Testy normality (dovolena.sta) ID2	N	W	p
X1: roční příjem v tisících dolarů	malá	12	0,706875	0,000982
X2: postoj k cestování (škála 9 bodů)	malá	12	0,867375	0,060535
X3: význam rodinné dovolené (škála 9 bodů)	malá	12	0,955130	0,712720
X4: počet členů rodiny	malá	12	0,907871	0,200341
X5: věk nejstaršího člena	malá	12	0,976999	0,968796
X1: roční příjem v tisících dolarů	střední	24	0,947240	0,235912
X2: postoj k cestování (škála 9 bodů)	střední	24	0,943681	0,196939
X3: význam rodinné dovolené (škála 9 bodů)	střední	24	0,962008	0,480070
X4: počet členů rodiny	střední	24	0,877051	0,007252
X5: věk nejstaršího člena	střední	24	0,882154	0,009185
X1: roční příjem v tisících dolarů	velká	14	0,897737	0,104575
X2: postoj k cestování (škála 9 bodů)	velká	14	0,922488	0,238745
X3: význam rodinné dovolené (škála 9 bodů)	velká	14	0,909165	0,153244
X4: počet členů rodiny	velká	14	0,958259	0,694341
X5: věk nejstaršího člena	velká	14	0,933244	0,338619

Boxův test shody variančních matic

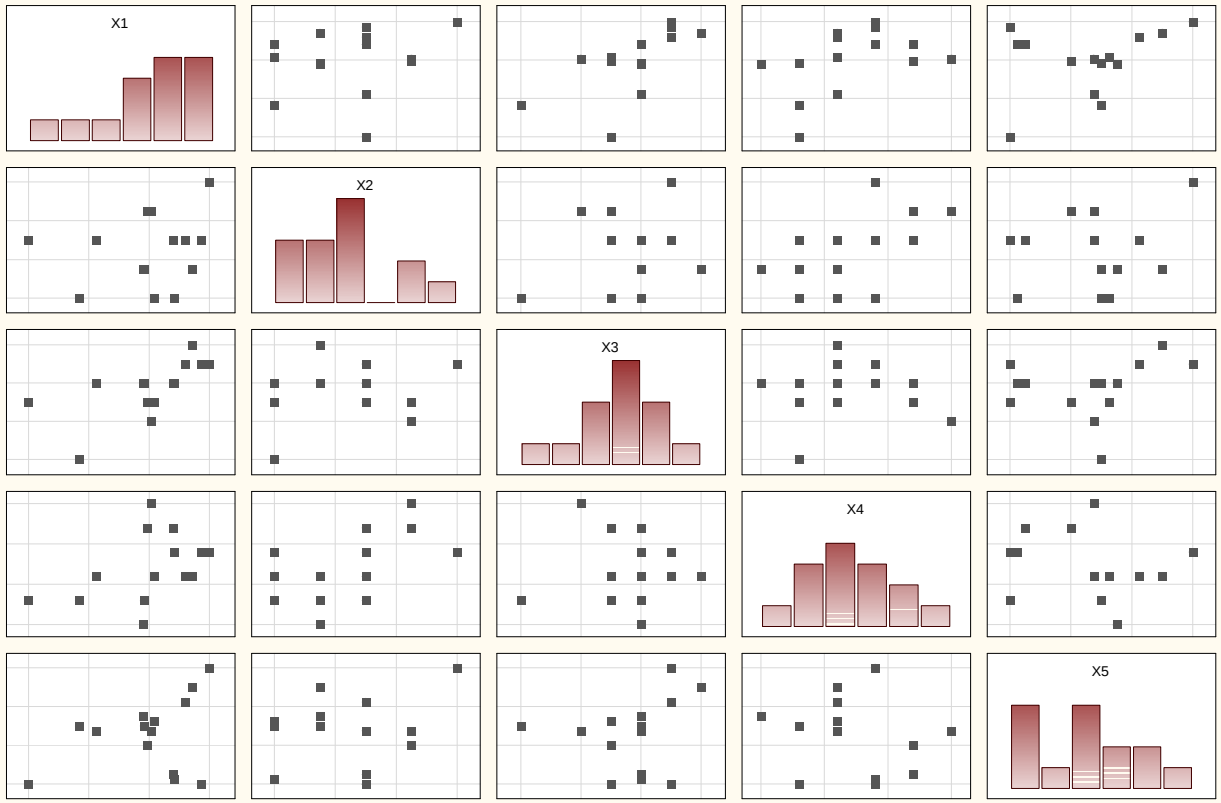
	Boxův M test (dovolena.sta) Efekt: "ID2" (Vypočteno pro všechny proměnné) Boxovo M	Chí-kv.	SV	p
Boxovo M	51,55790	42,84879	30	0,060418

Linearita vztahů proměnných X_1, …, X_5 v daných třech skupinách




Maticový graf
dovolena.sta 9v*50c
Zahrnout jestliže: ID2=2
X1
X2
X3
X4
X5

Maticový graf
dovolena.sta 9v*50c
Zahrnout jestliže: ID2=3



Testování hypotézy o shodě vektorů středních hodnot pomocí MANOVY

Efekt	Vícerozměrné testy významnosti. (dovolena.sta) Sigma-omezená parametrizace Dekompozice efektivní hypotézy Test	Hodnota	F	Efekt SV	Chyba SV	p
Abs. člen	Wilksův	0,01010	842,8765	5	43	0,000000
	Pillaiův	0,98990	842,8765	5	43	0,000000
	Hotellng	98,00890	842,8765	5	43	0,000000
	Royův	98,00890	842,8765	5	43	0,000000
"ID2"	Wilksův	0,26322	8,1626	10	86	0,000000
	Pillaiův	0,86784	6,7455	10	88	0,000000
	Hotellng	2,30122	9,6651	10	84	0,000000
	Royův	2,05945	18,1231	5	44	0,000000

Odlišnost vektorů středních hodnot ve sledovaných třech skupinách je prokázána na hladině významnosti 0,05.

Nyní provedeme simultánní testy o složkách vektorů středních hodnot.

Matice **E** reziduální variability

Efekt	Matice SSCP (Z' Z) reziduí (dovolena.sta) Sigma-omezená parametrizace Dekompozice efektivní hypotézy proměnné	X1	X2	X3	X4	X5
Chyba	X1	2386,662	-7,821	174,1762	134,0548	313,738
	X2	-7,821	118,714	-7,5119	5,9524	-103,131
	X3	174,176	-7,512	143,4821	1,1786	52,887
	X4	134,055	5,952	1,1786	73,7143	32,298
	X5	313,738	-103,131	52,8869	32,2976	2750,423

Matice **T** celkové variability

Efekt	Matice SSCP (Z' Z) odchylek (dovolena.sta) Matice SSCP (Z' Z) odchylek vektorů matice v matici schématu X Sloup.4 X1	Sloup.5 X2	Sloup.6 X3	Sloup.7 X4	Sloup.8 X5
X1	6535,025	299,6500	371,2500	250,2500	1026,550
X2	299,650	141,7800	8,1000	14,6200	-37,940
X3	371,250	8,1000	156,5000	6,9000	131,700
X4	250,250	14,6200	6,9000	76,9800	54,740
X5	1026,550	-37,9400	131,7000	54,7400	3425,620

Hodnoty testových statistik K1 až K5 a kritický obor:

	1 K1	2 K2	3 K3	4 K4	5 K5	6 kvantil
1	45,3276196	7,99016946	3,90805746	1,95069769	9,87874916	18,3070381

Na hladině významnosti 0,05 se prokázalo, že rozdíl mezi skupinami způsobuje X1.

Test shody vektorů středních hodnot a posouzení významu proměnných můžeme ve STATISTICE provést přímo v Diskriminační analýze.
Při zadávání proměnných zvolíme jako grupovací proměnnou ID2. Zvolíme-li Výpočet: proměnné v modelu, dostaneme tabulku:

	Výsledky diskriminační funkční analýzy (dovolena.sta) Počet prom. v modelu: 5; grupovací: ID2 (3 skup) Wilk. lambda: ,26322 přibliž F (10,86)=8,1626 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,43)	p-hodn.	Toler.	1-toler. R^2
X1	0,602832	0,436636	27,74006	0,000000	0,805704	0,194297
X2	0,289522	0,909148	2,14852	0,129016	0,959666	0,040334
X3	0,270302	0,973794	0,57859	0,564991	0,899531	0,100469
X4	0,269947	0,975075	0,54960	0,581183	0,883696	0,116304
X5	0,319480	0,823896	4,59552	0,015533	0,948842	0,051158

V záhlaví této tabulky je uvedena testová statistika pro Wilksův test shody vektorů středních hodnot (8,1626) a odpovídající p-hodnota (je blízká 0).
Podle p-hodnot u jednotlivých proměnných soudíme, že pro diskriminaci jsou významné proměnné X_1 a X_5.

Klasifikační funkce:

Proměnná	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
	malá p=,24000	střední p=,48000	velká p=,28000
X1	0,5525	0,8026	1,0981
X2	2,3285	2,4727	3,1155
X3	0,6466	0,3530	0,3648
X4	0,7459	0,4926	0,1242
X5	0,8874	0,7754	0,9120
Konstant	-42,2581	-45,1663	-70,7708

Zde jsou uvedeny koeficienty pro odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu:

$L_1(\mathbf{x}) = 0{,}5525*X1 + 2{,}3285*X2 + 0{,}6466*X3 + 0{,}7459*X4 + 0{,}8874*X5 - 42{,}2581$
$L_2(\mathbf{x}) = 0{,}8026*X1 + 2{,}4727*X2 + 0{,}3530*X3 + 0{,}4926*X4 + 0{,}7754*X5 - 45{,}1663$
$L_3(\mathbf{x}) = 1{,}0981*X1 + 3{,}1155*X2 + 0{,}3648*X3 + 0{,}1242*X4 + 0{,}9120*X5 - 70{,}7708$

Klasifikační matice:

Skup.	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správnýc	malá p=,24000	střední p=,48000	velká p=,28000
malá	66,66666	8	4	0
střední	91,66666	1	22	1
velká	78,57143	0	3	11
Celkem	82,00000	9	29	12

Správně zařazeno bylo $\frac{8+22+11}{50}\cdot 100\% = 82\%$ případů, chybně 18 % případů.

V 1. skupině rodin byly chybně zařazeny případy 8, 10, 19, 20 ($\frac{4}{12} = 33{,}3\%$), ve 2. skupině případy 4, 47 ($\frac{2}{24} = 8{,}3\%$) a ve 3. skupině případy 24, 34, 43 ($\frac{3}{14} = 21{,}4\%$)

Zařazení nového případu

Nyní podle těchto skórů zařadíme do jedné ze tří skupin rodinu, která

má roční příjem X_1 = 51,8 tisíc dolarů,

k cestování zaujímá postoj ohodnocený X_2 = 6 body,

rodinné dovolené přičítá význam ohodnocený X_3 = 7 body,

má X_4 = 4 členy

a nejstaršímu členovi je X_5 = 51 let.

Andersonovy diskriminační skóry:

	1 X1	2 X2	3 X3	4 X4	5 X5	6 L1	7 L2	8 L3
1	51,8	6	7	4	51	53,0996	55,23138	54,36618

Největší hodnotu má skór ve 2. skupině, tedy zkoumaná rodina vydá za dovolenou střední část-ku.

Dále v LDA použijeme pro výběr proměnných krokovou metodu.

Výsledky pro krokovou dopřednou metodu

Proměnné obsažené v modelu

N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,45)	p-hodn.	Toler.	1-toler. R^2
	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 3, poč. prom. v modelu: 3; grupovací: ID2 (3 skup) Wilk. lambda: ,27663 přibliž F (6,90)=13,519 p< ,0000					
X1	0,652311	0,424084	30,55552	0,000000	0,984948	0,015052
X5	0,338537	0,817147	5,03482	0,010635	0,953070	0,046930
X2	0,303098	0,912692	2,15236	0,128024	0,967370	0,032630

Klasifikační funkce

Proměnná	malá p=,24000	střední p=,48000	velká p=,28000
	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
X1	0,6401	0,8551	1,1311
X5	0,8991	0,7824	0,9163
X2	2,3409	2,4846	3,1046
Konstant	-41,3768	-44,8553	-70,5840

Klasifikační matice

Skup.	% správnýc	malá p=,24000	střední p=,48000	velká p=,28000
	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace			
malá	75,00000	9	3	0
střední	83,33334	3	20	1
velká	78,57143	0	3	11
Celkem	80,00000	12	26	12

Úspěšnost klasifikace poklesla z 82 % na 80 %.

Výsledky pro krokovou zpětnou metodu

Proměnné obsažené v modelu

	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 4, poč. prom. v modelu: 1; grupovací: ID2 (3 skup) Wilk. lambda: ,36521 přibliž F (2,47)=40,846 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,47)	p-hodn.	Toler.	1-toler. R^2
X1	1,000000	0,365211	40,84639	0,000000	1,000000	0,00

Klasifikační funkce

	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
Proměnná	malá p=,24000	střední p=,48000	velká p=,28000
X1	0,7506	0,9498	1,2413
Konstant	-15,7327	-23,6411	-40,3976

Klasifikační matice

	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace			
Skup.	% správnýc	malá p=,24000	střední p=,48000	velká p=,28000
malá	83,3333	10	2	0
střední	100,0000	0	24	0
velká	78,5714	1	2	11
Celkem	90,0000	11	28	11

Je-li ke klasifikaci rodin do skupin použita pouze proměnná X_1, je úspěšnost klasifikace nejvyšší, a to 90 %.
Aplikujeme-li toto klasifikační pravidlo na rodinu s vektorem pozorování (51,8 6 7 4 51)', dostaneme výsledek

	1 X1	2 X2	3 X3	4 X4	5 X5	6 L1	7 L2	8 L3
1	51,8	6	7	4	51	23,14838	25,55854	23,90174

Výpočet pomocí systému R

Načteme data:

```
dovolena3<-read.table('dovolena3.txt',sep=',',header=T)
```

Proměnnou ID zavedeme jako faktor:

```
ID<-factor(dovolena3$ID,labels=c('mala castka','stredni castka','velka castka'))
```

Zjistíme rozsahy 1., 2. a 3. skupiny:

```
ID
   mala castka stredni castka   velka castka
            12              24             14
```

Vytvoříme datový soubor pro 1., 2. a 3. skupinu:

```
data1<-dovolena3[1:12,2:6]
data2<-dovolena3[13:36,2:6]
data3<-dovolena3[37:50,2:6]
```

Zjistíme průměry všech proměnných v 1., 2. a 3. skupině:

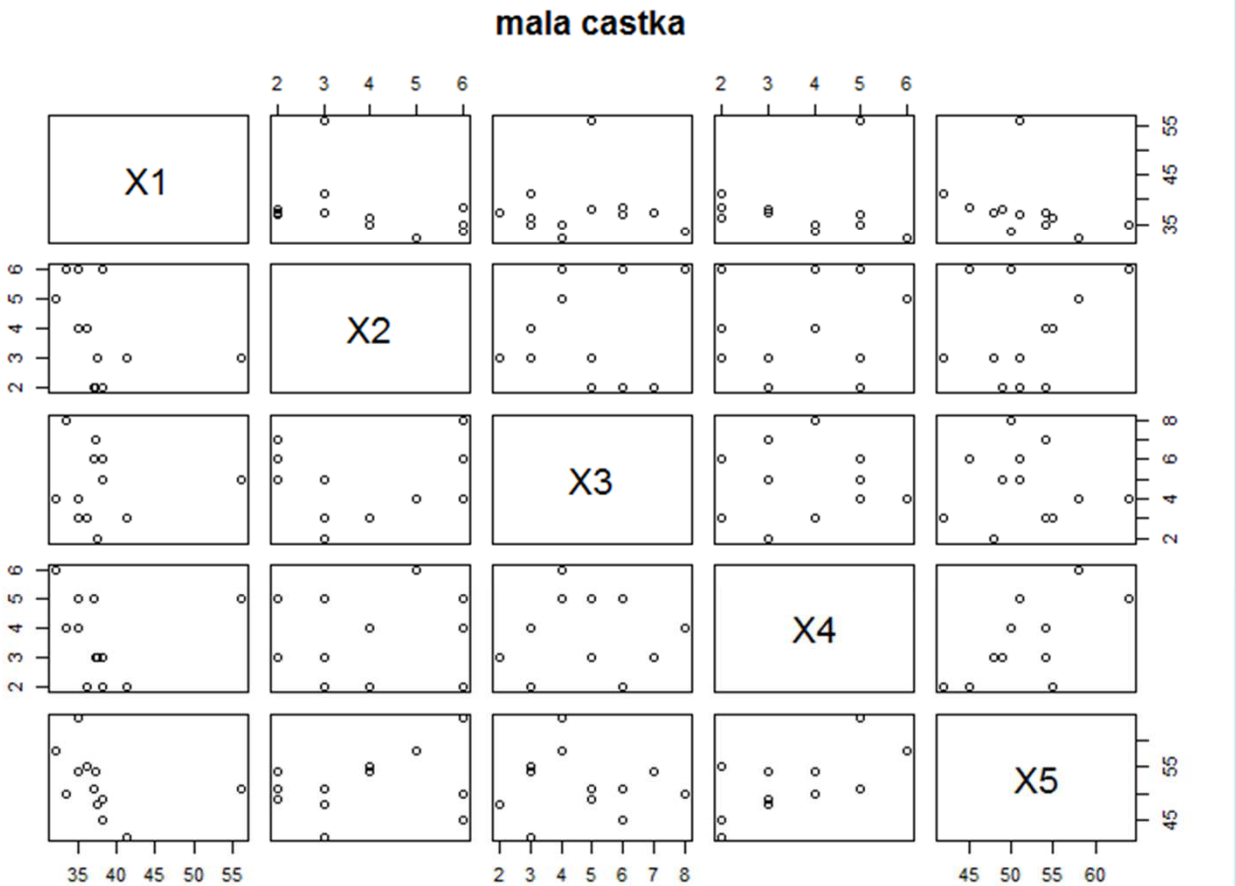
```
colMeans(data1)
       X1        X2        X3        X4        X5
38.116667  3.833333  4.666667  3.666667 51.750000
colMeans(data2)
       X1        X2        X3        X4        X5
48.233333  4.416667  4.541667  3.916667 46.041667
colMeans(data3)
       X1        X2        X3        X4        X5
63.035714  5.642857  5.714286  4.357143 54.357143
```

Vypočteme varianční matice v 1., 2. a 3. skupině:

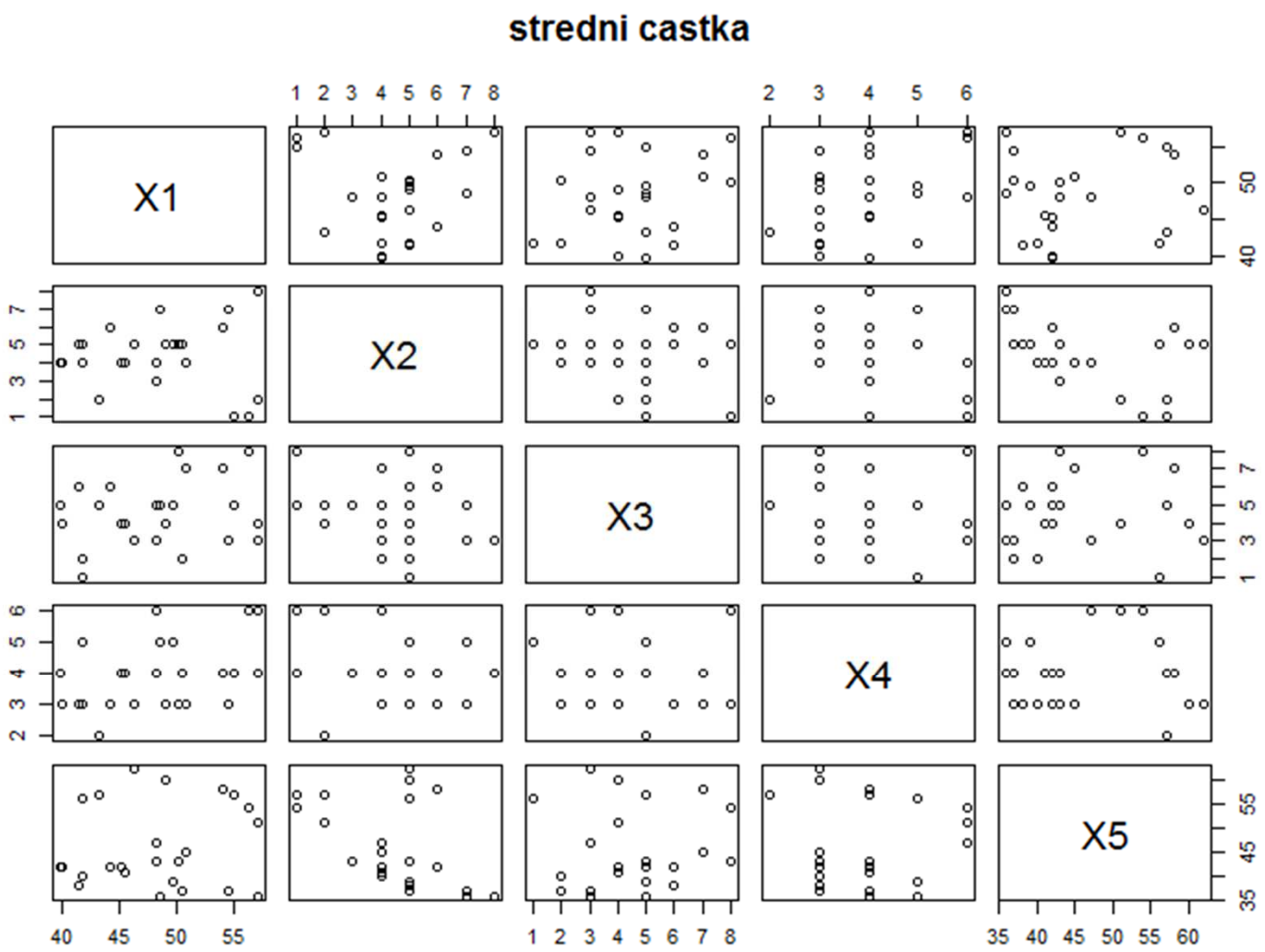
```
cov(data1)
             x1          x2         x3         x4          x5
x1  37.94696970 -3.4696970 -0.1848485 0.01515152 -11.4045455
x2  -3.46969697  2.5151515  0.3030303 0.30303030   2.7727273
x3  -0.18484848  0.3030303  3.3333333 0.33333333  -0.6363636
x4   0.01515152  0.3030303  0.3333333 1.87878788   4.9090909
x5 -11.40454545  2.7727273 -0.6363636 4.90909091  34.2045455

cov(data2)
           x1          x2         x3         x4         x5
x1 29.7701449 -0.1144928  2.3028986  2.3811594  5.5463768
x2 -0.1144928  3.1231884 -0.7137681 -0.3985507 -6.7137681
x3  2.3028986 -0.7137681  3.3894928 -0.1268116  1.2373188
x4  2.3811594 -0.3985507 -0.1268116  1.2101449  0.1775362
x5  5.5463768 -6.7137681  1.2373188  0.1775362 71.6068841

cov(data3)
           x1        x2         x3          x4        x5
x1 98.810165 2.5368132 9.48021978  6.08626374 23.970879
x2  2.536813 1.4780220 0.42857143  0.90659341  1.598901
x3  9.480220 0.4285714 2.21978022  0.03296703  2.417582
x4  6.086264 0.9065934 0.03296703  1.93956044 -1.983516
x5 23.970879 1.5989011 2.41758242 -1.98351648 55.939560
```

Orientačně ověříme linearitu vztahů v 1., 2. a 3. skupině:

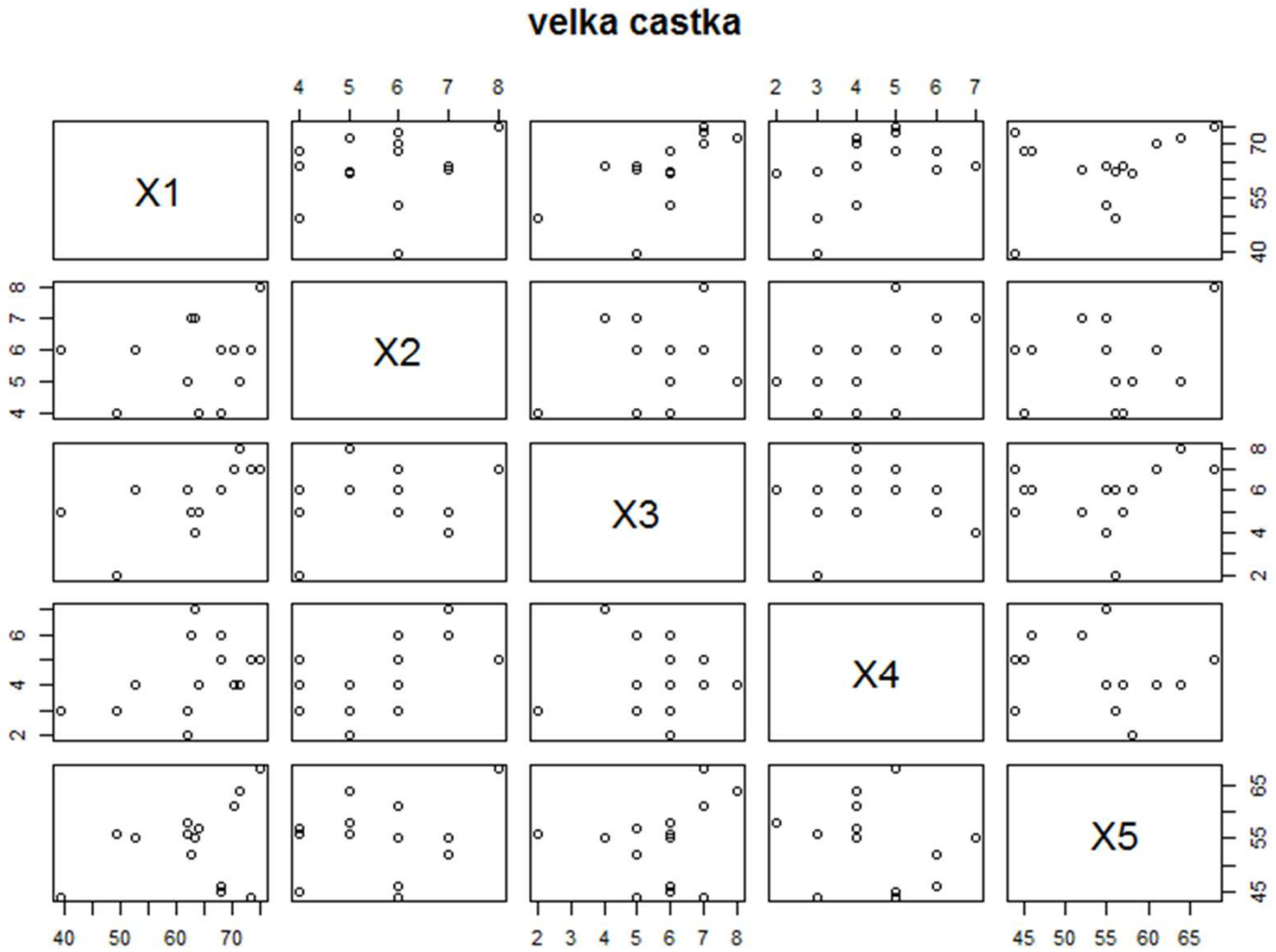
```
plot(data1, main='mala castka')
```



```
plot(data2, main='stredni castka')
```



```
plot(data3, main='velka castka')
```



Ověříme vícerozměrnou normalitu dat v 1., 2. a 3. skupině:

```
HZ.test(data1)
            Henze-Zirkler test for Multivariate Normality

  data : data1

  HZ              : 0.7754281
  p-value         : 0.2958131

  Result  : Data are multivariate normal (sig.level = 0.05)
HZ.test(data2)
            Henze-Zirkler test for Multivariate Normality

  data : data2

  HZ              : 1.043538
  p-value         : 0.002300619

  Result  : Data are not multivariate normal (sig.level = 0.05)
HZ.test(data3)
            Henze-Zirkler test for Multivariate Normality

  data : data3

  HZ              : 0.7142153
  p-value         : 0.6359006

  Result  : Data are multivariate normal (sig.level = 0.05)
```

Vícerozměrná normalita je porušena u 2. skupiny.

Provedeme Boxův test shody variančních matic:

```
library(biotools)
boxM(dovolena3[,2:6],grouping=ID)
      Box's M-test for Homogeneity of Covariance Matrices

data:  dovolena3[2:6]
Chi-Sq (approx.) = 42.849, df = 30, p-value = 0.06042
```

Pomocí MANOVY otestujeme hypotézu o shodě vektorů středních hodnot v daných třech skupinách rodin:

```
model <- manova(as.matrix(dovolena3[2:6]) ~ ID)
summary(model,test='Wilks')
           Df   Wilks approx F num Df den Df    Pr(>F)
ID          2 0.26322   8.1626     10     86 3.807e-09 ***
Residuals 47
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

Pokusíme se zjistit, které proměnné způsobují rozdíly. K tomu využijeme simultánní test založený na Wilksově statistice.

```
SSE <- summary(model,test='Wilks')$SS[[2]]
> SST<-SSB+SSE
> n<-nrow(dovolena3)
> n
[1] 50
> k<-5
> r<-3
> const <- (n- (k+r)/2 -1)
> K <- -const * log(diag(SSE)/diag(SST))
> K
       X1         X2         X3         X4         X5 
45.327617  7.990061  3.908044  1.950706  9.878755 
> ##
> (kvantil <- qchisq(0.95, df=k*(r-1)))
[1] 18.30704
```

V kritickém oboru $W = \langle 18{,}307; \infty)$ se realizuje testová statistika pro proměnnou X1 (nabývá hodnoty 45,3276), tedy rozdíl mezi rodinami je způsoben ročním příjmem.

Sestavíme funkci pro rozlišení skupin:

```
dovolena.lda <- lda(ID ~ X1+X2+X3+X4+X5, data=dovolena3)
dovolena.lda
Call:
lda(ID ~ X1 + X2 + X3 + X4 + X5, data = dovolena3)

Prior probabilities of groups:
   mala castka stredni castka   velka castka
          0.24           0.48           0.28

Group means:
                     X1       X2       X3       X4       X5
mala castka    38.11667 3.833333 4.666667 3.666667 51.75000
stredni castka 48.23333 4.416667 4.541667 3.916667 46.04167
velka castka   63.03571 5.642857 5.714286 4.357143 54.35714

Coefficients of linear discriminants:
            LD1         LD2
X1  0.14100713 -0.04449459
X2  0.22026963  0.15735553
X3 -0.06004878  0.19117537
X4 -0.16315720  0.01823570
X5  0.01594357  0.12414042

Proportion of trace:
   LD1    LD2
0.8949 0.1051
```

Resubstituční metoda provede přiřazení rodin do skupin podle aposteriorních pravděpodobností:

```
fit<-predict(dovolena.lda)
(tab.d <- table(fit$class, ID))
                ID
                 mala castka stredni castka velka castka
  mala castka                8               1            0
  stredni castka             4              22            3
  velka castka               0               1           11
```

Vypočteme podíl správně zařazených rodin:

```
sum(diag(tab.d))/sum(tab.d)
[1] 0.82
```

Nyní zařadíme do jedné ze tří skupin rodinu, která má roční příjem X1 = 51,8 tisíc dolarů, k cestování zaujímá postoj ohodnocený X2 = 6 body, rodinné dovolené přičítá význam ohodnocený X3 = 7 body, má X4 = 4 členy a nejstaršímu členovi je X5 = 51 let.

```
predict(dovolena.lda, newdata=list(X1=51.8,X2=6,X3=7,X4=4,X5=51))
$class
[1] stredni castka
Levels: mala castka stredni castka velka castka

$posterior
  mala castka stredni castka velka castka
1  0.07731109      0.6496072    0.2730817

$x
        LD1       LD2
1 0.4555586 0.6930856
```

Vidíme, že rodina byla zařazena do skupiny, která je ochotna vydat za rodinnou dovolenou středně velkou částku.