Cvičení 11.: Lineární diskriminační analýza

Třídění do dvou skupin
Úkol: V souboru 50 rodin byly zjišťovány tyto údaje:

- zda v posledních dvou letech rodina navštívila jistou rekreační oblast (veličina ID, nabývá hodnoty 0 pro odpověď „ne“, hodnoty 1 pro odpověď „ano“)
- 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).

Pro uvedená data sestrojte Fisherovu lineární diskriminační funkci, která pomocí veličin $X_1, \ldots, X_5$ umožní rozlišit rodiny navštěvující uvedenou rekreační oblast od rodin, které do této oblasti nejezdí.
Data jsou uložena v souboru dovolena.sta.

Testování normality náhodných veličin $X_1, \ldots, X_5$ v daných dvou skupinách pomocí S - W testu:
Pro skupinu rodin, které danou rekreační oblast nenavštěvují:

	Testy normality (dovolena.sta Zhrnout podmínku: ID=0		
Proměnná	N	W	p
X1: roční příjem v tisících dolarů	29	0,940188	0,101411
X2: postoj k cestování (škála 9 bodů)	29	0,964071	0,412187
X3: význam rodinné dovolené (škála 9 bodů)	29	0,964432	0,420319
X4: počet členů rodiny	29	0,917696	0,026668
X5: věk nejstaršího člena	29	0,944508	0,131598

Pro skupinu rodin, které danou rekreační oblast navštěvují:

	Testy normality (dovolena.sta Zhrnout podmínku: ID=1		
Proměnná	N	W	p
X1: roční příjem v tisících dolarů	21	0,935874	0,180430
X2: postoj k cestování (škála 9 bodů)	21	0,930271	0,139382
X3: význam rodinné dovolené (škála 9 bodů)	21	0,934717	0,171087
X4: počet členů rodiny	21	0,928224	0,126815
X5: věk nejstaršího člena	21	0,967589	0,679311

Na hladině významnosti 0,05 zamítáme hypotézu o normalitě u veličiny X_4 ve skupině rodin, které danou rekreační oblast nenavštěvují.

Odhady číselných charakteristik a krabicové grafy:

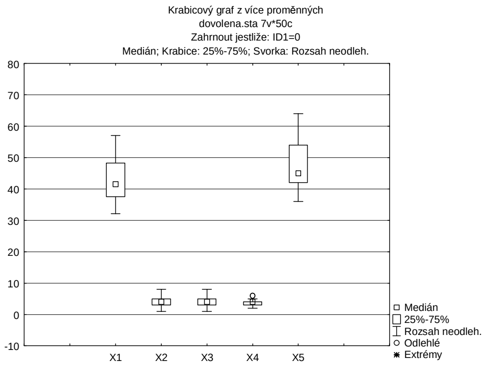
Odhad vektoru středních hodnot $\mathbf{M}_1$:

Proměnná	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=0	
	N platných	Průměr
X1	29	42,84483
X2	29	4,24138
X3	29	4,27586
X4	29	3,72414
X5	29	46,93103

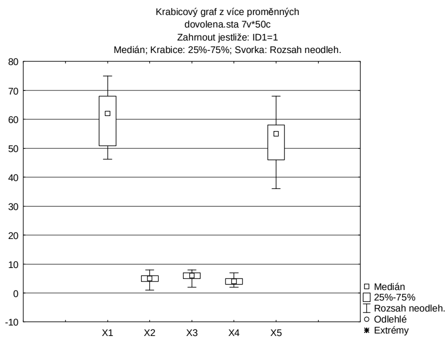
Odhad vektoru středních hodnot $\mathbf{M}_2$:

Proměnná	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=1	
	N platných	Průměr
X1	21	59,76190
X2	21	5,14286
X3	21	5,76190
X4	21	4,33333
X5	21	53,61905

Krabicové grafy:



Krabicové grafy:



Odhad varianční matice $\mathbf{S}_1$

Proměnná	Kovariance (dovolena.sta) Zhrnout podmínku: ID=0				
	X1	X2	X3	X4	X5
X1	49,1947	0,99594	-2,24138	1,094951	-24,1647
X2	0,9959	2,76108	-0,31897	0,140394	-4,7328
X3	-2,2414	-0,31897	2,63547	-0,171182	1,1268
X4	1,0950	0,14039	-0,17118	1,278325	1,9446
X5	-24,1647	-4,73276	1,12685	1,944581	57,2808

Odhad varianční matice $\mathbf{S}_2$

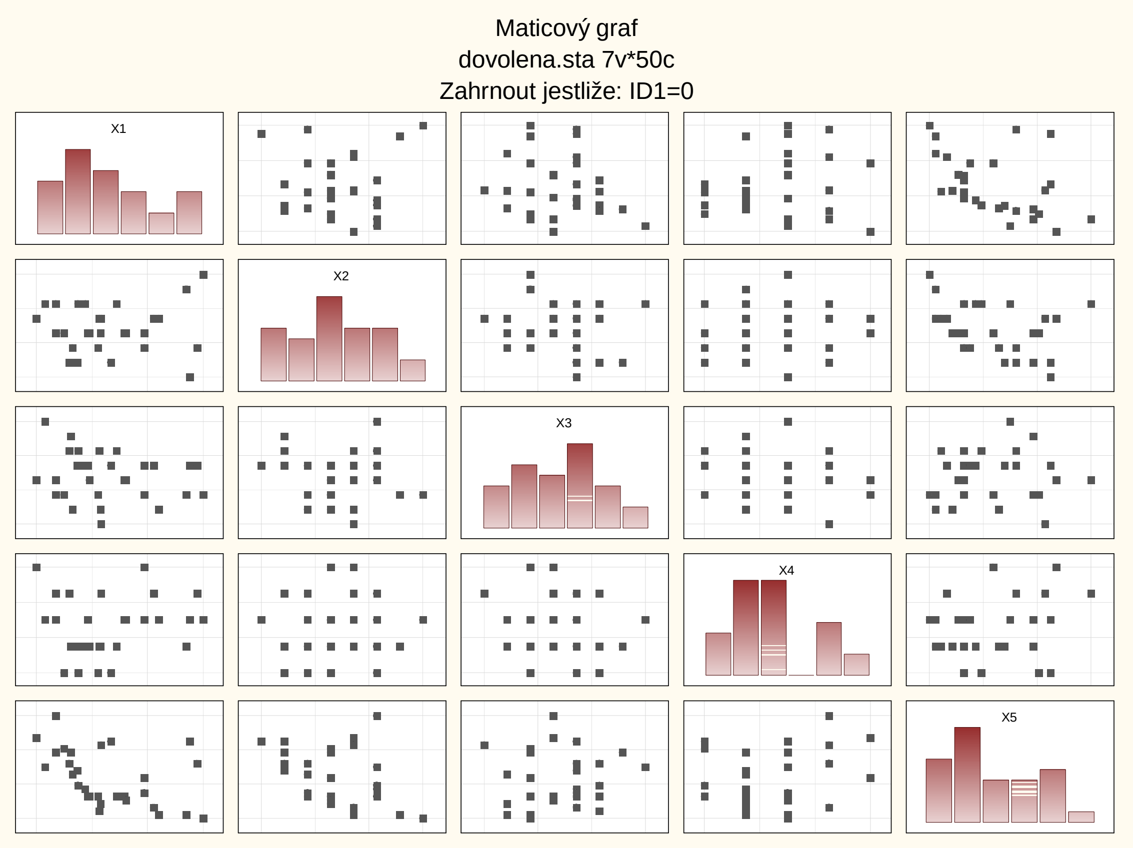
Proměnná	Kovariance (dovolena.sta) Zhrnout podmínku: ID=1				
	X1	X2	X3	X4	X5
X1	83,59048	4,300714	6,39048	4,70333	16,25476
X2	4,30071	2,728571	0,03571	0,20000	1,05714
X3	6,39048	0,035714	2,79048	0,03333	-1,04524
X4	4,70333	0,200000	0,03333	1,83333	-2,46667
X5	16,25476	1,057143	-1,04524	-2,46667	63,84762

Odhad společné varianční matice **S**
Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné – Grupovací ID, Seznam nezáv. proměnných X1-X5 – OK, zapneme Další možnosti (kroková analýza) – OK – Popisné statistiky – Zobrazit popisné statistiky – Vnitřní kovariance a korelace.

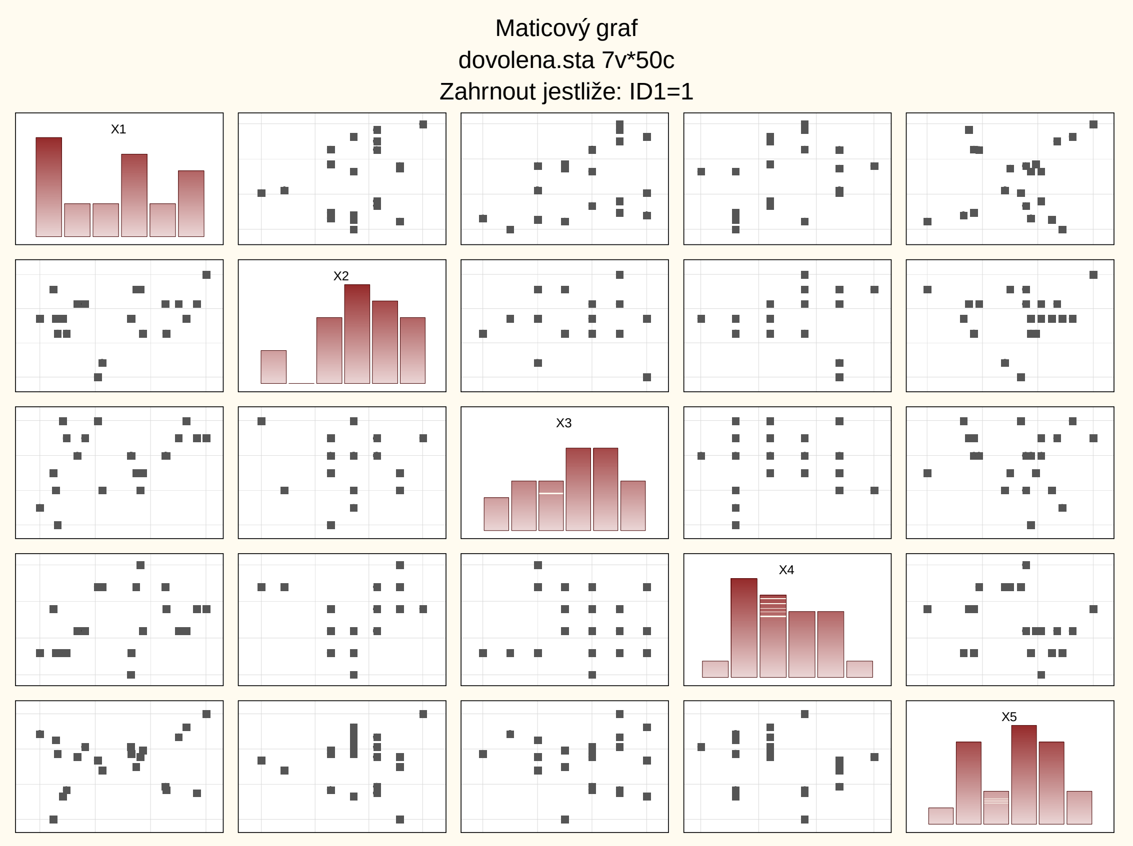
	X_1	X_2	X_3	X_4	X_5
X1	63,53	2,37	1,36	2,60	-7,32
X2	2,37	2,75	-0,17	0,17	-2,32
X3	1,36	-0,17	2,70	-0,09	0,22
X4	2,60	0,17	-0,09	1,51	0,11
X5	-7,32	-2,32	0,22	0,11	60,02

Ověření linearity vztahů mezi proměnnými:

Skupina rodin nenavštěvujících danou oblast



Skupina rodin navštěvujících danou oblast



Boxův test shody variančních matic:
Statistiky – Pokročilé lineární/nelineární modely – Obecné lineární modely – Typ analýzy: Jednofaktorová ANOVA - Metoda specifikace: Rychlé nastavení – OK – Proměnné – Seznam závislých proměnných: X1 – X5, Kategor. nezávislá proměnná (faktor): ID – OK – OK – Více výsledků – Boxův M-test.

	Boxův M test (dovolena.sta) Efekt: ID (Vypočteno pro všechny proměnné)			
	Boxovo M	Chí-kv.	sv	p
Boxovo M	26,61690	23,54681	15	0,073200

Protože p-hodnota je větší než hladina významnosti 0,05, hypotézu o shodě variančních matic nezamítáme na asymptotické hladině významnosti 0,05.

Test shody vektorů středních hodnot:
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, dle skupin – OK – Proměnné – Závisle proměnné X1 až X5, Grupovací proměnná ID – OK – na záložce Možnosti zaškrtneme Vícerozměrný test. V záhlaví výstupní tabulky se zobrazí realizace testové statistiky a příslušná p-hodnota.

	t-testy; grupováno: ID (dovolena.sta) Skup. 1: návštěva ne; Skup. 2: návštěva ano Hotellingovo 77,5606 F(5,44)=14,219 p<,00000								
Proměnná	Průměr návštěva ne	Průměr návštěva ano	t	sv	p	Poč.plat návštěva ne	Poč.plat. návštěva ano	Sm.odch. návštěva ne	Sm.odch. návštěva ano
X1	42,84483	59,76190	-7,40751	48	0,000000	29	21	7,013894	9,142783
X2	4,24138	5,14286	-1,89805	48	0,063712	29	21	1,661651	1,651839
X3	4,27586	5,76190	-3,15623	48	0,002760	29	21	1,623412	1,670472
X4	3,72414	4,33333	-1,73042	48	0,089980	29	21	1,130630	1,354006
X5	46,93103	53,61905	-3,01289	48	0,004122	29	21	7,568407	7,990471

Vidíme, že na hladině významnosti 0,05 jsou odlišné střední hodnoty proměnných X_1, X_3, X_5. U proměnných X_2 a X_4 se odlišnost neprokázala, z dalšího zpracování je však vyřazovat nebudeme.

Význam jednotlivých proměnných v modelu:
Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné - Grupovací ID1 – Seznam nezáv. proměnných X1 až X5 – OK – OK – Výpočet: proměnné v modelu.

	Výsledky diskriminační funkční analýzy (dovolena.sta) Počet prom. v modelu: 5; grupovací: ID1 (2 skup) Wilk. lambda: ,38229 přibliž F (5,44)=14,219 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (1,44)	p-hodn.	Toler.	1-toler. R^2
X1	0,627513	0,609207	28,22504	0,000003	0,879866	0,120134
X2	0,388609	0,983729	0,72778	0,398223	0,934715	0,065285
X3	0,400086	0,955507	2,04884	0,159388	0,977164	0,022836
X4	0,382565	0,999270	0,03215	0,858527	0,921303	0,078697
X5	0,439319	0,870177	6,56444	0,013904	0,956782	0,043218

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 (14,219) 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 významné proměnné X_1 a X_5.
5. sloupec (Tolerance) udává unikátní variabilitu proměnné nevysvětlenou ostatními proměnnými v modelu.
6. sloupec (1-toler., R^2) udává variabilitu proměnné vysvětlenou ostatními proměnnými.

Mahalanobisova vzdálenost v diskriminační analýze
Používá se pro popis vzájemných vzdáleností centroidů jednotlivých skupin.

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK – na záložce Detaily zvolíme Vzdálenosti mezi skupinami. Současně dostaneme i p-hodnoty pro testy hypotéz, že vzdálenosti jsou nulové:

	Mahalanobisovy vzdálenosti^2 (dovoler	
ID1	návštěva ne	návštěva ano
návštěva ne	0,000000	6,367867
návštěva ano	6,367867	0,000000

	p-hodnot (dovolena.sta)	
ID1	návštěva ne	návštěva
návštěva ne		0,0C
návštěva ano	0,000000	

Lze také získat Mahalanobisovy vzdálenosti jednotlivých objektů od centroidů skupin.

Na záložce Klasifikace zvolíme Mahalanobisovy vzdálenosti^2:

	Mahalanobisovy vzdálenosti (dovolena.sta) Nesprávná klasifikace je označena *		
Případ	**Pozorova Klasif.**	návštěva ne p=,58000	návštěva ano p=,42000
1	návštěva ne	9,18363	18,11825
2	návštěva ne	0,88533	10,53314
3	návštěva ne	3,90372	12,30937
4	návštěva ne	5,35649	8,74744
5	návštěva ne	4,41397	11,30806
6	návštěva ne	0,62136	7,62423

Stanovení odhadu Fisherovy lineární diskriminační funkce:
Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK – na záložce Klasifikace zvolíme Klasifikační funkce. Dostaneme tabulku tvaru:

	Klasifikační funkce; grupovací : ID (dovolena	
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000
X1	0,6369	0,9054
X2	1,7840	2,0395
X3	1,3391	1,7560
X4	1,1866	1,1130
X5	0,9216	1,0743
Konstant	-44,6709	-69,4375

Abychom získali odhad Fisherovy lineární diskriminační funkce, přidáme do této tabulky novou proměnnou a do jejího Dlouhého jména napíšeme =v1-v2

Proměnná	Klasifikační funkce; grupovací : ID (dovolena) návštěva ne p=,58000	návštěva ano p=,42000	NProm =v1-v2
X1	0,6369	0,9054	-0,26847
X2	1,7840	2,0395	-0,25557
X3	1,3391	1,7560	-0,41694
X4	1,1866	1,1130	0,073566
X5	0,9216	1,0743	-0,15266
Konstant	-44,6709	-69,4375	24,76658

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g = -0{,}2685X_1 - 0{,}2556X_2 - 0{,}4169X_3 + 0{,}0736X_4 - 0{,}1527X_5 + 24{,}7666$

Klasifikace nového případu:

Předpokládejme nyní, že jsme prozkoumali další 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.
Na základě těchto údajů se pokusíme pomocí Fisherovy lineární diskriminační funkce zařadit tuto rodinu do skupiny rodin, které buď navštěvují nebo nenavštěvují danou rekreační oblast:

$L(\mathbf{x}) = -0{,}2685X_1 - 0{,}2556X_2 - 0{,}4169X_3 + 0{,}0736X_4 - 0{,}1527X_5 + 24{,}7666 =$
$= -0{,}2685*51{,}8 - 0{,}2556*6 - 0{,}4169*7 + 0{,}0736*4 - 0{,}1527*51 + 24{,}7666 = -1{,}0836.$
Protože $L(\mathbf{x}) < 0$, zařadíme tuto rodinu do skupiny rodin, které navštěvují danou rekreační oblast.

Posouzení účinnosti diskriminace resubstituční metodou:

Na záložce Klasifikace zvolíme Klasifikační matice.

Skup.	Klasifikační matice (dovolena) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správných	návštěva ne p=,58000	návštěva ano p=,42000
návštěva ne	93,10345	27	2
návštěva ano	76,19048	5	16
Celkem	86,00000	32	18

Podíl správně zařazených objektů:

$$\frac{n_{11} + n_{22}}{n} = \frac{27 + 16}{50} = 0{,}86$$

Podíl mylně zařazených objektů:

$$\frac{n_{12} + n_{21}}{n} = \frac{5 + 2}{50} = 0{,}14$$

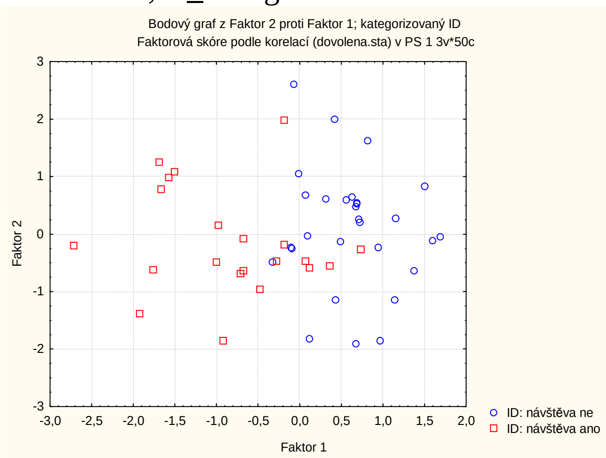
Pro určení chybně zařazených případů zvolíme na záložce Klasifikace možnost Klasifikace případů. Zjistíme, že v 1. skupině došlo k mylnému zařazení u rodin č. 9 a 10, ve 2. skupině u rodin číslo 30, 33, 36, 43, 45.

Porovnání s náhodnou klasifikací:
Kdybychom zařazovali rodiny do skupin náhodně, pouze s ohledem na apriorní pravděpodobnosti π_1, π_2, tak bychom s pravděpodobností π_1 našli rodinu patřící do 1. skupiny, avšak s pravděpodobností π_2 bychom ji mylně zařadili do 2. skupiny.
Naopak s pravděpodobností π_2 najdeme rodinu patřící do 2. skupiny, kterou s pravděpodobností π_1 mylně zařadíme do 1. skupiny. Celková pravděpodobnost mylné klasifikace je tedy: $\pi_1\pi_2 + \pi_2\pi_1 = 2\pi_1(1-\pi_1)$. Nahradíme-li apriorní pravděpodobnosti π_1, π_2 jejich odhady p_1, p_2 , dostaneme odhad celkové pravděpodobnosti mylné klasifikace:

$$2p_1(1-p_1) = 2\cdot\frac{29}{50}\cdot\frac{21}{50} = 0{,}4872.$$

Použitím diskriminační analýzy jsme tedy dosáhli výrazného zlepšení, pravděpodobnost mylné klasifikace klesla na 0,14.

Grafické znázornění případů na ploše prvních dvou hlavních komponent
Jako aktivní vstup použijeme Faktorová skóre podle korelací z analýzy hlavních komponent. Grafy – Kategorizované grafy – Bodové grafy – Rozložení Přes sebe – Proměnné X: Faktor 1, Y: Faktor 2, X_Kategorie: ID - OK



Výběr proměnných do modelu pomocí krokových metod:

Statistika – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné - Grupovací ID1 – Seznam nezáv. proměnných X1 až X5 – OK – zaškrtneme Další možnosti (kroková analýza) – OK – Metoda – zvolíme kroková dopředná. Na záložce Detaily můžeme změnit Možnosti kroku (ponecháme implicitní nastavení) a také pomocí tlačítka Výsledky můžeme zvolit, zda chceme zobrazovat výsledky po každém kroku nebo chceme pouze shrnutí (ponecháme shrnutí) – OK.
Zvolíme-li tlačítko Výpočet: proměnné v modelu, dostaneme tabulku

	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 3, poč. prom. v modelu: 3; grupovací: ID1 (2 skup) Wilk. lambda: ,38880 přibliž F (3,46)=24,104 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (1,46)	p-hodn.	Toler.	1-toler. R^2
X1	0,719493	0,540386	39,12429	0,000000	0,974791	0,025209
X5	0,441811	0,880024	6,27128	0,015879	0,985042	0,014958
X3	0,405987	0,957678	2,03285	0,160683	0,988398	0,011602

Vidíme, že algoritmus skončil po třech krocích a vybral proměnné X_1, X_5 a X_3.

Zvolíme-li tlačítko Proměnné neobsažené v modelu, zjistíme, že jde o proměnné X_2 a X_4.
Na záložce Klasifikace vybereme Klasifikační funkce. Dostaneme lineární diskriminační skóry pro 1. a 2. skupinu objektů. Do vzniklé tabulky přidáme novou proměnnou L, do jejíhož Dlouhého jména napíšeme =v1-v2 a tím získáme odhad Fisherovy lineární diskriminační funkce:

Proměnná	Klasifikační funkce; grupovací : ID1 (dovolena.sta) návštěva ne p=,58000	návštěva ano p=,42000	L =v1-v2
X1	0,7504	1,0247	-0,2742808
X5	0,8693	1,0128	-0,1434212
X3	1,1355	1,5365	-0,4009242
Konstant	-39,4479	-63,0649	23,6170025

Vidíme, že L(**x**) = -0,2743*X1 – 0,1434*X5 – 0,4009*X3 + 23,617

Klasifikační matice je stejná jako v případě diskriminace podle všech proměnných a chybně zařazené případy jsou také stejné.

Skup.	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správnýc	návštěva ne p=,58000	návštěva ano p=,42000
návštěva ne	93,10345	27	2
návštěva ano	76,19048	5	16
Celkem	86,00000	32	18

Použijeme-li krokovou zpětnou metodu, je vybrána pouze proměnná X_1 a účinnost diskriminace poklesne na 80 %.

Úkol k samostatnému řešení: Použijte datový soubor SIDS.sta, který obsahuje údaje o 65 novorozencích, z nichž někteří zemřeli na syndrom náhlého úmrtí kojence. Obsahuje tyto proměnné :
ID … má hodnotu 1, když novorozenec žije , hodnotu 2, když umřel na syndrom náhlé smrti kojence (SIDS)
X1 … počet tepů za minutu
X2 … porodní hmotnost v gramech
X3 … popisuje funkci srdce a plic
X4 … počet týdnů těhotenství (všichni se narodili aspoň v 37. týdnu, což je považováno za ukončené období zdravého vývoje plodu)

Ověřte předpoklady pro provedení LDA.
(S-W test normality prokázal porušení normality u proměnné X4 ve skupině 1, Boxův test nezamítl shodu variančních matic na hladině významnosti 0,05, linearita vztahů mezi proměnnými je v obou skupinách přibližně splněna.)

Zjistěte význam jednotlivých proměnných v modelu:

	Výsledky diskriminační funkční analýzy (SIDS.sta) Počet prom. v modelu: 4; grupovací: ID (2 skup) Wilk. lambda: ,68278 přibliž F (4,60)=6,9691 p< ,0001					
N=65	Wilk. Lambda	Parc. Lambda	F na vyj (1,60)	p-hodn.	Toler.	1-toler. R^2
X1	0,682851	0,999893	0,00641	0,936461	0,953604	0,046396
X2	0,757725	0,901089	6,58610	0,012792	0,849030	0,150970
X3	0,831482	0,821157	13,06763	0,000616	0,917977	0,082023
X4	0,686567	0,994481	0,33299	0,566064	0,835387	0,164613

Test hypotézy o shodě vektorů středních hodnot v obou skupinách je na hladině významnosti 0,05 průkazný. Největší vliv na diskriminaci mají proměnné X2 a X3.

Vypočtěte Mahalanobisovy vzdálenosti skupin a odpovídající p-hodnoty (2,4267, p = 0,000112).

Jaké jsou apriorní pravděpodobnosti příslušnosti objektů ke skupinám? (p_1 = 0,75385, p_2 = 0,24615).

Stanovte odhad Fisherovy lineární diskriminační funkce:
L(**x**) = -0,00178322553X1+0,00178850321X2-15,5253107X3+0,214815554X4-7,35265591

Posuďte účinnost diskriminace resubstituční metodou.
81,54 % objektů je správně zařazeno.

Které objekty byly zařazeny chybně?
(V 1. skupině objekty č. 14, 32, 34, ve 2. skupině 50, 52, 54, 57, 59, 60, 62, 64, 65)

Odhadněte celkovou pravděpodobnost mylné klasifikace při náhodném zařazování (0,37).

Dále proveďte LDA krokovou dopřednou metodou:
(Do modelu byly zařazeny proměnné X3 a X2, odhad Fisherovy lineární diskriminační funkce je:
L(**x**) = -16,0770541X3+0,00194756243+0,613039967, úspěšnost diskriminace je 81,54 %.)

Třídění do tří skupin
Použijte datový soubor ropa.sta.

Zjistěte význam jednotlivých proměnných v modelu:

N=45	Výsledky diskriminační funkční analýzy (ropa.sta) Počet prom. v modelu: 4; grupovací: ID (3 skup) Wilk. lambda: ,17959 přibliž F (8,78)=13,257 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,39)	p-hodn.	Toler.	1-toler. R^2
X1	0,229700	0,781858	5,44059	0,008241	0,730601	0,269399
X2	0,213007	0,843133	3,62803	0,035890	0,736104	0,263896
X3	0,219437	0,818427	4,32621	0,020096	0,648482	0,351519
X4	0,321952	0,557825	15,45717	0,000011	0,662875	0,337126

Test hypotézy o shodě vektorů středních hodnot v obou skupinách je na hladině významnosti 0,05 průkazný. Největší vliv na diskriminaci mají proměnné X1 a X4.

Vypočtěte Mahalanobisovy vzdálenosti skupin a odpovídající p-hodnoty.

ID	Mahalanobisovy vzdálenosti^2 (ropa.sta)		
	G_1:1	G_2:2	G_3:3
G_1:1	0,0000C	10,15239	17,55756
G_2:2	10,15239	0,00000	6,78236
G_3:3	17,55756	6,78236	0,0000C

ID	p-hodnot (ropa.sta)		
	G_1:1	G_2:2	G_3:3
G_1:1		0,000037	0,00000C
G_2:2	0,000037		0,000012
G_3:3	0,000000	0,000012	

Jaké jsou apriorní pravděpodobnosti příslušnosti objektů ke skupinám? (p_1 = 0,15556, p_2 = 0,17778, p_3 = 0,66667).

Najděte odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu:

Proměnná	Klasifikační funkce; grupovací : ID (ropa.sta)		
	G_1:1 p=,15556	G_2:2 p=,17778	G_3:3 p=,66667
X1	0,3645	0,4964	0,5373
X2	0,9792	0,8364	0,6613
X3	0,0499	0,0688	0,0534
X4	0,0085	-0,0064	-0,0042
Konstant	-49,0185	-50,0807	-38,9736

Posuďte účinnost diskriminace resubstituční metodou:
93,33 % objektů je správně zařazeno.

Které případy byly zařazeny chybně?
(V 1. skupině objekt č. 2, ve 2. skupině 10, ve 3. skupině 34)

Dále proveďte LDA krokovou zpětnou metodou:
(Do modelu byly zařazeny proměnné X1 a X4, odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu jsou:

Proměnná	Klasifikační funkce; grupovací : ID (ropa.sta)		
	G_1:1 p=,15556	G_2:2 p=,17778	G_3:3 p=,66667
X1	0,0773	0,18557	0,2940
X4	0,0155	0,00478	0,0044
Konstant	-11,6852	-7,66368	-12,9968

Úspěšnost diskriminace je 88,89 %.)

Nepovinný úkol: Na datovém souboru Irisdat.sta, který obsahuje údaje o délce a šířce okvětních a kališních lístků 150 rostlin tří druhů kosatců (Setosa, Virginic, Versicola) proveďte lineární diskriminační analýzu.