Vzorová úloha 4.7 *Užití lineární diskriminační funkce*

Předpokládejme, že máme data o 2 třídách objektů tibetských lebek v úloze **B4.14** *Aglomerativní hierarchické shlukování při analýze lebek Tibeťanů*: prvních 13 bylo nalezeno v hrobech v Sikkimu a okolí, zatímco druhých 15 lebek na bojištích okolo Lhasy. První třída vede ke středním hodnotám $\bar{x}_1^T$ = [174.82, 139.35, 132.00, 69.82, 130.35] a kovarianční matici

$$S_1 = \begin{vmatrix} 45.53 & & & & \\ 25.22 & 57.81 & & & \\ 12.39 & 11.88 & 36.09 & & \\ 22.15 & 7.52 & -0.31 & 20.94 & \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{vmatrix} .$$

Druhá třída vede ke středním hodnotám $\bar{x}_1^T$ = [185.73, 138.73, 134.77, 76.47, 137.50] a kovarianční matici

$$S_2 = \begin{vmatrix} 74.42 & & & & \\ -9.52 & 37.35 & & & \\ 22.74 & -11.26 & 36.32 & & \\ 17.79 & 0.70 & -10.72 & 15.30 & \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{vmatrix} .$$

Koeficienty diskriminační funkce jsou vyčísleny vztahem

$$a = S^{-1}(\bar{x}_1 - \bar{x}_2) = [-0.09, 0.16, 0.01, -0.18, -0.18]$$

a vedou k průměrům u obou tříd: $\bar{z}_1$ = -28.71 a $\bar{z}_2$ = -32.21. Hraniční bod, dle kterého se budou nezařazené objekty třídit do první nebo druhé třídy se vyčíslí jako polosuma obou průměrů $(\bar{z}_1 + \bar{z}_2)/2$ = (-28.71 + (-32.21))/2 = -30.46.

Diskriminace: vezmeme data pro lebku prvního Tibetana z dat všech lebek a pokusíme se ji diskriminovat čili zařadit do 1. nebo 2. třídy. Vyčísleme pro ni hodnotu lineární diskriminační funkce

$$z_1 = -0.09 \times 190.5 + 0.16 \times 152.5 + 0.01 \times 145.0 - 0.18 \times 73.5 - 0.18 \times 136.5 = -29.74,$$

a protože -29.74 je menší než hraniční bod -30.46, patří lebka prvního Tibeťana do první třídy.

Vzorová úloha 4.8 *Užití logistické diskriminace*

Logistickou diskriminaci budeme demonstrovat na **Úloze B4.12** *Aplikace logistické diskriminační analýzy u rakoviny prostaty*. Režim léčení je závislý na rozšíření rakoviny na lymfatické uzliny. Rozhodující metodou vyšetření je laparotomie, vyjádřená proměnnou *B412x6*: je-li výsledek laparotomického vyšetření 0 negativní výsledek a je-li roven 1 pozitivní výsledek nodálního rozšíření rakoviny. Brownův postup následujícího vyšetření pěti diskriminantů u 53 pacientů by měl do jisté míry nahradit právě toto obtížnější laparotomické vyšetření. Brown ve své studii použil databázi: *i* je index pacienta, *B412x1*

věk pacienta, *B412x2* hladina sérové kyselé fosfatázy v Kingových-Armstrongových jednotkách, *B412x3* výsledek roentgenového vyšetření (0 = negativní, 1 = pozitivní), *B412x4* velikost tumoru rektálním vyšetřením (0 = malý, 1 = velký), *B412x5* závěr pathologického bodování z biopsie (0 = méně vážný, 1 = velmi vážný).

Diskriminace: odhady parametrů (včetně svých směrodatných odchylek v závorce) k vyčíslení logistické diskriminační funkce jsou b_0 1.52 (3.56), b_1 0.10 (0.06), b_2 2.64 (1.33), b_3 1.68 (0.80), b_4 2.04 (0.83), b_5 0.35 (0.80). Tyto odhady vedou k formulaci klasifikačního pravidla, zda má pacient rakovinu lymfatických uzlin či ne. Pacient rakovinu lymfatických uzlin nemá a je diskriminován do 1. třídy, je-li splněna nerovnost

$$1.52 - 0.10\, x_1 + 2.64\, x_2 + 1.68\, x_3 + 2.04\, x_4 + 0.35\, x_5 > 0.$$

Není-li splněna tato nerovnost, je pacient diskriminován do 2. třídy s rakovinou lymfatických uzlin. Dosadíme-li do této nerovnosti hodnoty prvního pacienta z databáze, dostaneme

$$1.52 - 0.10 \times 66 + 2.64 \times 0.48 + 1.68 \times 0 + 2.04 \times 0 + 0.35 \times 0 = -3.81.$$

Protože výsledek -3.81 není větší než nula, je pacient diskriminován do 1. třídy bez rakoviny lymfatických uzlin, což potvrdilo konečně i laparotomické vyšetření.

Posouzení správnosti diskriminace: po aplikaci diskriminační funkce k zařazení objektů do tříd je třeba posoudit správnost diskriminace. Aplikaci diskriminace na data objektů vyhodnotíme jejich chybné zařazení do tříd:

(a) Křížová tabulka diskriminace. Ukážeme křížovou tabulku zařazených objektů na konkrétním příkladu, například databáze lebek Tibeťanů. Sestavíme křížovou tabulku původního (správného) umístění objektů (lebek) do tříd a nalezeného zařazení do tříd diskriminací. Výsledkem bude *tabulka správnosti klasifikace* diskriminační analýzou, kde nesprávné zařazení je zvýrazněno tučným písmem:

		Známo (správné třídy)	
		1	2
Nalezeno diskriminací	1	14	**3**
	2	**3**	12

Nesprávného umístění je 100 % . 6/32 = 19 %. Výhodou této techniky je právě její jednoduchost, nevýhodou příliš optimistické závěry, ke kterým většinou metoda dospěje.

(b) Postupné vypouštění "vždy jednoho objektu". Spolehlivější výsledky přináší modifikace předešlého způsobu. Vytvoříme primární třídy pro n - 1 objektů a vyšetřujeme zařazení jediného dosud nezařazeného objektu. Postup n-krát opakujeme tak, že postupně vyšetřujeme zařazení všech objektů testovaného souboru. Užijeme-li i zde databáze lebek Tibeťanů, obdržíme tabulku správnosti klasifikace diskriminační analýzou, kde nesprávné zařazení je zvýrazněno tučným písmem:

		Známo (správné třídy)	
		1	2
Nalezeno diskriminací	1	12	**5**
	2	**6**	9

Nesprávného umístění je 100 % . 11/32 = 34 %, což je téměř dvojnásobek než u předešlé příliš optimistické metody.

Volba proměnných: otázkou v diskriminační analýze je, zda volba proměnných je schopna provést zařazení objektů do tříd čili diskriminaci. Byla navržena řada postupů jak provést volbu těch nejúčinnějších proměnných. Principem většiny metod je zajištění dostatečné separability tříd a volba takových proměnných, které vedou k maximalizaci nějaké míry. Jindy se volí postup, který začne se všemi původními proměnnými a postupně se vypouštějí takové, které vedou k nedostatečné redukci separace.

K ilustraci užijeme databáze lebek Tibeťanů z **úlohy B4.14** *Aglomerativní hierarchické shlukování při analýze lebek Tibeťanů*. Užijeme pouze jednu proměnnou, *B414x4* výšku horní části obličeje [mm]. Dostaneme velmi jednoduché klasifikační pravidlo: lebka bude zařazena do 1. třídy tehdy, když výška horní části obličeje bude menší než 73.14 mm. Optimistický odhad chybné klasifikace je 25%.

Krokový postup u logistické diskriminace **úlohy B4.12** *Aplikace logistické diskriminační analýzy u rakoviny prostaty* vede k volbě tří nejúčinnějších proměnných: *B412x2* hladina sérové kyselé fosfatázy v Kingových-Armstrongových jednotkách, *B412x3* výsledek roentgenového vyšetření (0 = negativní, 1 = pozitivní), *B412x4* velikost tumoru rektálním vyšetřením (0 = malý, 1 = velký).

Vzorová úloha 4.9 *Užití postupu diskriminační analýzy*

V úloze **S2.18** *Fisherova úloha rozměrů okvětních lístků u 150 kosatců* analyzujte předložený výběr kosatců, obsahujících čtvero popisných rozměrů okvětních lístků (čili diskriminátorů) u 150 květů kosatců (čili objektů), pocházejících ze tří základních tříd: (1) Iris setosa, (2) Iris versicolor, (3) Iris virginica. Z botaniky je známo, že druh Iris versicolor je hybridem zbývajících dvou druhů. Iris setosa je diploidní květ s 38 chromozomy, Iris virginica je tetraploidní a Iris versicolor je hexaploidní s 108 chromozomy. Květy kosatců jsou popsány čtyřmi diskriminátory: délkou kališních lístků v mm anglicky *lsepal,* šířkou *wsepal,* dále délkou korunních plátků v mm *lpetal* a šířkou *wpetal*. Budeme proto formulovat úlohu: jsou dána data o K třídách, $K = 3$, tři druhy čili třídy kosatců: *Setosa, Versicolor* a *Virginica* s N_k, $k = 1, ..., K,$ objekty v každé třídě, pro *Setosu* $k = 1$ *je* $N_1 = 50$, pro *Versicolor* $k = 2$ *je* $N_2 = 50$ a pro *Virginica* $k = 3$ *je* $N_3 = 50$, N představuje celkový počet objektů, $N = N_1 + N_2 + N_3 = 150$. Každý objekt je popsán p diskriminátory, $p = 4$, a to *Sepal Length, Sepal Width, Petal Length, Petal Width*. Každý i-tý objekt je prezentován prvkem x_{ki}. Nechť $\bar{\boldsymbol{x}}$ představuje vektor průměrů diskriminátorů ve všech třídách dohromady a $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ je vektor průměrů objektů v k-té třídě. Cílem diskriminační analýzy je vyšetřit a ověřit botanické třídění a odpovědět na otázku, zda botanické třídění kosatců Iris do tří tříd je správné. Nelze zařadit 150 kosatců do jiného počtu tříd?

Řešení: Výstup z bloku Discriminant Analysis (NCSS2000) pro Fisherovu úlohu:

1. **Výpočet bodových odhadů parametrů polohy a rozptýlení všech diskriminátorů:**
   (a) **Aritmetický průměr** [mm] u tříd G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica) a celkově:

	G_1	G_2	G_3	
Proměnná	**Setosa**	**Versicolor**	**Virginica**	**Celkově**
SepalLength	50.06	59.36	65.88	58.43333
SepalWidth	34.28	27.7	29.74	30.57333
PetalLength	14.62	42.6	55.52	37.58
PetalWidth	2.46	13.26	20.26	11.99333
Počet	50	50	50	150

Tabulka obsahuje průměry každého diskriminátoru, a to v každé třídě kosatců. Poslední řádek obsahuje počet objektů ve třídě. Nadpisy sloupců jsou názvy dotyčné třídy kosatců. **Celkově** znamená všechny třídy dohromady.

   (b) **Směrodatné odchylky** [mm] u tříd G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica) a celkově:

	G_1	G_2	G_3	
Proměnná	**Setosa**	**Versicolor**	**Virginica**	**Celkově**
SepalLength	3.524897	5.161712	6.358796	8.280662
SepalWidth	3.790644	3.137983	3.224966	4.358663
PetalLength	1.73664	4.69911	5.518947	17.65298
PetalWidth	1.053856	1.977527	2.7465	7.622377
Počet	50	50	50	150

Tabulka obsahuje směrodatné odchylky každého diskriminátoru, a to v každé třídě kosatců. Poslední řádek obsahuje počet objektů ve třídě. Nadpisy sloupců jsou názvy dotyčné třídy kosatců. **Celkově** znamená všechny třídy dohromady. Diskriminační analýza je postavena na předpokladu, že kovarianční matice jsou stejné pro každou třídu. Tato tabulka umožňuje posoudit předpoklad, zda totiž jsou směrodatné odchylky ve třídách zhruba stejné.

   (c) **Celkové korelace a kovariance:**

		Proměnná		
Proměnná	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength	68.56935	-4.243401	127.4315	51.62707
SepalWidth	-0.117570	18.99794	-32.96564	-12.16394
PetalLength	0.871754	-0.428440	311.6278	129.5609
PetalWidth	0.817941	-0.366126	0.962865	58.10063

Tabulka obsahuje korelace a kovariance, vytvořené když jsou ignorovány smíšené proměnné *diskriminátorů.* Korelace jsou v dolní levé části, kovariance jsou v pravé horní části matice. Rozptyly jsou na diagonále matice.

   (d) **Mezitřídní korelace a kovariance:**

		Proměnná		
Proměnná	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength	3160.607	-997.6334	8262.42	3563.967
SepalWidth	-0.745075	567.2466	-2861.98	-1146.633
PetalLength	0.994135	-0.812838	21855.14	9338.7
PetalWidth	0.999768	-0.759258	0.996232	4020.667

Tabulka obsahuje korelace a kovariance, vytvořené za použití průměrů místo jednotlivých objektů. Korelace jsou v dolní levé části, mezitřídní kovariance jsou na diagonále matice a v horní pravé části matice. Všimněte si, že když by byly jenom dvě třídy kosatců, všechny korelace by byly rovny jedné, protože byly vytvořeny pouze ze dvou řádků, totiž ze dvou třídních průměrů.

   (e) **Vnitrotřídní korelace a kovariance:**

	Proměnná			
Proměnná	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength	26.50082	9.272109	16.75143	3.840136
SepalWidth	0.530236	11.53878	5.524354	3.27102
PetalLength	0.756164	0.377916	18.51878	4.266531
PetalWidth	0.364506	0.470535	0.484459	4.188163

Tabulka obsahuje korelace a kovariance, vytvořené z dat, ve kterých byly třídní průměry odečteny. Korelace jsou v dolní levé části, vnitrotřídní kovariance jsou na diagonále a v pravé horní části matice.

2. Vyšetření vlivu jednotlivých diskriminátorů:

	Při odstranění této proměnné			**Pro tuto samotnou proměnnou**			R^2
Proměnná	**Lambda**	***F*-test**	**Spočtená α**	**Lambda**	***F*-test**	**Spočtená α**	**ostatní *X***
SepalLength	0.938463	4.72	0.010329	0.381294	119.26	0.000000	0.858612
SepalWidth	0.766480	21.94	0.000000	0.599217	49.16	0.000000	0.524007
PetalLength	0.669206	35.59	0.000000	0.058628	1180.2	0.000000	0.968012
PetalWidth	0.743001	24.90	0.000000	0.071117	960.01	0.000000	0.937850

Tabulka ukazuje na vliv jednotlivých diskriminátorů proměnných na výsledky diskriminační analýzy. **Proměnná:** jméno diskriminátoru. **Lambda při odstranění této proměnné:** hodnota Wilkova lambda, vypočtená k testování důsledku odstranění této diskriminační proměnné. ***F*-test při odstranění této proměnné:** hodnota *F*-kritéria, vyčísleného k testování statistické významnosti Wilkova lambda. **Spočtená hladina významnosti při odstranění této proměnné:** vypočtená hladina významnosti výše uvedeného *F*-testu při odstranění této diskriminační proměnné. Test je totiž statisticky významný a diskriminátor je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$. **Lambda pro tuto samotnou proměnnou:** jde o hodnotu Wilkova lambda, kterou dostaneme za použití této jediné nezávisle proměnné. ***F*-test pro tuto samotnou proměnnou:** jde o testační kritérium, vyčíslené k testování statistické významnosti Wilkova lambda. **Spočtená hladina významnosti pro tuto samotnou proměnnou:** uvedený *F*-test je statisticky významný a diskriminátor je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$.

3. **Odhady neznámých parametrů** b_0 , b_1 , ..., b_p **lineární diskriminační funkce pro každou třídu** G_1 *(Setosa),* G_2 *(Versicolor),* G_3 *(Virginica)*:

	G_1	G_2	G_3
Proměnná	**Setosa**	**Versicolor**	**Virginica**
Absolutní člen -85.20985	-71.754	-103.2697	
SepalLength	2.354417	1.569821	1.244585
SepalWidth	2.358787	0.707251	0.3685279
PetalLength	-1.643064	0.5211451	1.276654
PetalWidth	-1.739841	0.6434229	2.107911

Tabulka obsahuje odhady neznámých parametrů b_0 , b_1 , ..., b_p lineární diskriminační funkce. Tyto parametry jsou také nazývány diskriminačními koeficienty. Technika předpokládá, že diskriminátory v každé třídě kosatců vykazují vícerozměrné normální rozdělení se shodnými variančně-kovariančními maticemi ve třídách. Technika je dostatečně robustní i při nesplnění těchto předpokladů. Tabulka obsahuje celkem tři klasifikační funkce, jednu pro každou třídu. Každá funkce je prezentována vertikálně hodnotami ve sloupci. Když vytvoříme vážený průměr diskriminátorů užitím těchto koeficientů jako vah (a přidáním konstanty jako absolutního členu), dostaneme diskriminační skóre.

4. **Odhady regresních parametrů b_0 , b_1 , ..., b_p lineárního regresní modelu pro každou třídu G_1 *(Setosa), G_2 (Versicolor), G_3 (Virginica)*:**

	G_1	G_2	G_3
Proměnná	**Setosa**	**Versicolor**	**Virginica**
Absolutní člen	0.1182229	1.577059	-0.6952819
SepalLength	6.602977E-03	-2.015369E-03	-4.587608E-03
SepalWidth	2.428479E-02	-4.456162E-02	2.027684E-02
PetalLength	-2.246571E-02	2.206692E-02	3.987911E-04
PetalWidth	-5.747273E-03	-4.943066E-02	5.517793E-02

Tabulka obsahuje regresní parametry b_0 , b_1 , ..., b_p lineárního regresního modelu pro každou třídu G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica), které byly vyčísleny následujícím postupem: (1) Vytvoříme tři indikátorové proměnné, jedna je pro každou ze tří druhů kosatců (Setosa, Versicolor a Virdinica). Každá indikátorová proměnná je položena rovna jedné. (2) Proložíme vícenásobnou regresí nezávisle proměnných každý ze tří kosatců. (3) Obdržíme odhady regresních parametrů, uvedené v tabulce. Těmito regresními parametry pak predikované hodnoty budou ležet mezi nulou a jedničkou. Určení, ke které třídě jedinec patří se provede tak, že se vybere třída s nejvyšším skóre.

5. **Klasifikace objektů diskriminačního funkcí** (diskriminace objektů do tříd):

(a) **Tabulka klasifikačních počtů pro kosatce u diskriminace do tříd** G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica) **a celkově:**

Predikovaná	G_1	G_2	G_3	
Známá	**Setosa**	**Versicolor**	**Virginica**	**Total**
Setosa	50	0	0	50
Versicolor	0	34	16	50
Virginica	0	7	43	50
Celkově	50	41	59	150

Redukce v klasifikační správnosti v důsledku proměnných $\boldsymbol{X}$ = 77.0 %.

Tabulka ukazuje, jak navržené diskriminační funkce klasifikují objekty v datech. Bylo-li dosaženo perfektní klasifikace, obdržíme v matici mimo diagonálu nuly. Řádky tabulky představují aktuální třídy kosatců, zatímco sloupce představují predikované třídy kosatců. **Redukce v klasifikační správnosti:** obsahuje procento redukce v klasifikační správnosti, dosažené diskriminačními funkcemi vůči očekávané hodnotě, když byly objekty klasifikovány náhodně.

(b) **Přehled chybně klasifikovaných objektů v řádcích u diskriminace do tříd** G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica):

			Procento zařazení do jednotlivé třídy		
Řádek	**Známá**	**Predikovaná**	**Třída 1**	**Třída 2**	**Třída 3**
5	Virginica	Versicolo	-1.8	58.6	43.1
9	Versicolo	Virginica	10.3	20.2	69.5
22	Versicolo	Virginica	18.8	22.6	58.6
28	Versicolo	Virginica	22.1	35.5	42.4
29	Versicolo	Virginica	22.1	27.4	50.6
38	Versicolo	Virginica	10.6	38.3	51.1
45	Virginica	Versicolo	-31.4	66.4	65.0
57	Virginica	Versicolo	-18.6	83.9	34.7
62	Versicolo	Virginica	24.4	34.0	41.6
66	Versicolo	Virginica	11.9	37.9	50.2
70	Versicolo	Virginica	12.1	41.5	46.3
78	Virginica	Versicolo	-7.3	58.4	48.9
91	Virginica	Versicolo	-16.1	83.8	32.3

95	Versicolo	Virginica	23.7	14.3	62.0
106	Versicolo	Virginica	20.7	30.7	48.7
111	Virginica	Versicolo	-21.4	63.8	57.6
112	Virginica	Versicolo	-23.9	71.8	52.1
114	Versicolo	Virginica	17.1	35.6	47.2
117	Versicolo	Virginica	22.1	38.9	39.0
130	Versicolo	Virginica	30.9	32.4	36.8
131	Versicolo	Virginica	14.0	39.6	46.4
142	Versicolo	Virginica	21.4	38.6	40.0
148	Versicolo	Virginica	6.8	36.8	56.4

V řádku se u každého chybně klasifikovaného objektu nachází vždy název známé třídy kosatců a predikované třídy kosatců. Následuje 100× zvětšená hodnota pravděpodobnosti (v procentech), že objekt se nachází v dané třídě kosatců. Procento pravděpodobnosti se jeví totiž názornější než normovaný odhad v rozmězí 0 a 1. Hodnota blízko 100 % ukazuje, že objekt patří do dotyčné třídy. ***P(i)*:** při užití lineární diskriminační techniky se vyčíslí pravděpodobnosti, že tento řádek patří do *i*-té třídy: nechť f_i , $i = 1, ..., K$, je hodnota lineární diskriminační funkce a $\max(f_k)$ je maximální skóre ze všech tříd. Označme $P(G_i)$ celkovou pravděpodobnost, klasifikující jednotlivce do třídy *i*. Hodnota *P(i)* se vypočte dle vztahu

$$P(i) = \frac{\exp[f_i - \max(f_k)]\ P(G_i)}{\sum_{j=1}^{K} \exp[f_i - \max(f_k)]\ P(G_i)} .$$

Když užijeme regresní klasifikační techniku, bude představovat predikovanou hodnotu regresní rovnice. Implicitně je Y v regresní rovnici rovno 1 nebo 0 v závislosti, zda objekt do *i*-té třídy kosatců patří či ne. Proto predikovaná hodnota blízko nuly ukazuje, že objekt nepatří do *i*-té třídy, zatímco blízko 1 ukazuje na silný důkaz, že objekt patří do *i*-té třídy. V žádném případě nemůže vyčíslena hodnota být větší než 1 a menší než 0.

(c) **Zařazení objektů predikovanou klasifikací pomocí diskriminační funkce do tříd** G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica):

			Procento zařazení do jednotlivé třídy		
Řádek	**Známá**	**Predikovaná**	**Třída 1**	**Třída 2**	**Třída 3**
1	Setosa	Setosa	92.4	21.6	-14.0
2	Virginica	Virginica	-16.4	34.9	81.5
3	Versicolo	Versicolo	10.8	47.2	42.0
..	.....	.....	.....	.....	.....
..	.....	.....	.....	.....	.....
150	Setosa	Setosa	101.8	5.4	-7.2

Tabulka obsahuje pro každý objekt kosatců vždy skutečnou, čili známou třídu kosatců, predikovanou třídu kosatců a procento pravděpodobnosti zařazení do dotyčné třídy kosatců.

6. Kanonická korelační analýza:

(a) **Analýza kanonických proměnných:**

	Inv(W)B	**Ind.**	**Total**	**Kanon.**	**Kanon.**		**Čitatel**	**Jmenov.**	**Spočtená**	**Wilkovo**
Fn	**vlast.číslo**	**Pcnt**	**Pcnt**	**korel.**	**korel2**	***F*-test**	**SV**	**SV**	**α**	**Lambda**
1	32.191929	99.1	99.1	0.9848	0.9699	199.1	8.0	288.0	0.0000	0.023439
2	0.285391	0.9	100.0	0.4712	0.2220	13.8	3.0	145.0	0.0000	0.777973

F-test testuje, zda tato funkce a další jsou statisticky významné.

Tabulka obsahuje výsledky kanonické korelační analýzy diskriminačního problému. U kanonické korelační analýzy jsou dva soubory proměnných, které jsou zde definovány následovně: první soubor obsahuje diskriminátory. Třídní proměnná definuje druhý, jiný soubor, který je generován vytvořením indikátorové proměnné pro každou třídu,

kromě poslední. **Inv(W)B vlastn. číslo:** vlastní čísla matice $\mathbf{W^{-1}B}$ ukazují, jak mnoho je celková proměnlivost vysvětlena různými diskriminačními funkcemi. První diskriminační funkce totiž odpovídá prvnímu vlastnímu číslu, atd. Počet vlastních čísel je roven minimu počtu diskriminátorů a K-1, kde K je počet tříd kosatců. **Ind. Pcnt:** procento, jež toto vlastní číslo představuje z celku vlastních čísel. **Total Pcnt:** kumulativní procento tohoto a všech předešlých vlastních čísel. **Kanon korel.:** kanonický korelační koeficient. **Kanon korel2:** čtverec kanonického korelačního koeficientu je podobný R^2 ve vícenásobné regresi. ***F*-test:** hodnota *F*-kritéria, testujícího Wilkovo lambda, které odpovídá tomuto řádku a řádkům níže. V tomto případě testuje *F*-kritérium statistickou významnost obou, první a druhé, kanonické korelace, zatímco druhá *F*-hodnota testuje významnost pouze druhé korelace. **Čitatel SV:** počet stupňů volnosti pro čitatele v tomto *F*-testu. **Jmenov. SV:** počet stupňů volnosti pro jmenovatele v tomto *F*-testu. **Spočtená α:** spočtená hladina významnosti pro *F*-test. Je-li tato hodnota α menší než uživatelem zadané 0.05, je test statisticky významný. **Wilkovo lambda:** hodnota Wilkova lambda pro tento řádek se užívá k testování statistické významnosti diskriminační funkce, odpovídající tomuto řádku a řádkům níže. Wilkovo lambda je vícerozměrným zobecněním R^2. Uvedený *F*-test je aproximativním testem Wilkova lambda.

(b) **Odhady parametrů u kanonických proměnných:**

	Kanonická proměnná	
Proměnná	**Proměnná1**	**Proměnná2**
Absolutní člen	-2.105106	6.661473
SepalLength	-0.082938	-0.002410
SepalWidth -0.153447	-0.216452	
PetalLength	0.220121	0.093192
PetalWidth	0.281046	-0.283919

Obsahuje koeficienty k výpočtu kanonického skóre. Kanonická skóre jsou vážené průměry objektů a tyto koefienty jsou pak váhy s přidaným absolutním členem.

(c) **Kanonické proměnné u třídních průměrů:**

	Kanonická funkce	
Iris	**Funkce 1**	**Funkce 2**
Setosa	-7.6076	-0.215133
Versicolor	1.82505	0.7278996
Virginica	5.78255	-0.5127666

Tabulka obsahuje výsledky kanonických koeficientů pro průměry u každé třídy.

(d) **Standardizované kanonické koeficienty:**

	Kanonická proměnná	
Proměnná	**Proměnná 1**	**Proměnná 2**
SepalLength	-0.426955	-0.012408
SepalWidth	-0.521242	-0.735261
PetalLength	0.947257	0.401038
PetalWidth	0.575161	-0.581040

Tabulka obsahuje standardizované kanonické koeficienty.

(e) **Korelace původních a kanonických proměnných:**

	Kanonická proměnná	
Proměnná	**Proměnná 1**	**Proměnná 2**
SepalLength	0.222596	-0.310812
SepalWidth	-0.119012	-0.863681
PetalLength	0.706065	-0.167701
PetalWidth	0.633178	-0.737242

Tabulka obsahuje zátěže (korelace) původních proměnných na kanonické proměnné. Každý výstup je korelací mezi kanonickou proměnnou a diskriminátorem. Tato tabulka usnadní interpretovat dotyčné kanonické proměnné.

7. Lineární diskriminační skóre všech objektů :

Řádek	Iris	Skóre1	Skóre2	Skóre3
1	Setosa	83.86837	38.65921	-6.790054
2	Virginica	1.230765	91.857	104.5692
..	.....	.....	.....	.....
150	Setosa	98.72371	46.71882	-0.3055334

Tabulka obsahuje jednotlivé hodnoty lineárních diskriminačních skóre pro všechny objekty, tj. pro všech 150 kosatců.

8. Regresní skóre všech objektů:

Řádek	Iris	Skóre1	Skóre2	Skóre3
1	Setosa	0.923755	0.215832	-0.139588
2	Virginica	-0.163732	0.348623	0.815109
3	Versicolo	0.107759	0.471953	0.420288
..	.....	.....	.....	.....
..	.....	.....	.....	.....
150	Setosa	1.018238	0.053607	-0.071844

Tabulka obsahuje jednotlivé hodnoty predikovaných skóre, založené na regresních koeficientech. I když tyto hodnoty jsou predikované indikátorové proměnné, může nastat případ, že hodnota bude menší než nula a větší než 1.

9. Kanonická skóre všech objektů:

Řádek	Iris	Skóre1	Skóre2
1	Setosa	-7.671967	0.134894
2	Virginica	6.800150	-0.580895
3	Versicolo	2.548678	0.472205
..	.....	.....	.....
..	.....	.....	.....
150	Setosa	-8.314449	-0.644953

Tabulka obsahuje skóre kanonických proměnných pro každý řádek u všech objektů, tj. 150 kosatců.

10. Automatická volba účinných diskriminátorů:

Dosavadní tabulky jsou postaveny na čtyřech diskriminátorech: Petal Length, Petal Width, Sepal Length a Sepal Width. Stěžejním úkolem v diskriminační analýze je však výběr diskriminátorů. Často máme velikou paletu možných diskriminátorů, ze kterých potřebujeme vybrat menší výběr, asi tak maximálně 8 účinných proměnných, který se bude chovat jako původní velký soubor.

Iterace	Činnost v kroku	Nezávisle proměnná	% změny v lambda	*F*-test	Spočtená hladina α	Wilkovo lambda
0	None					1.000000
1	Entered	PetalLength	94.14	1180.16	0.000000	0.058628
2	Entered	SepalWidth	37.09	43.04	0.000000	0.036884
3	Entered	PetalWidth	32.29	34.57	0.000000	0.024976
4	Entered	SepalLength	6.15	4.72	0.010329	0.023439
..	...	.....	...	...	.....	.....

Detail ve 4. kroku automatického výběru proměnné:

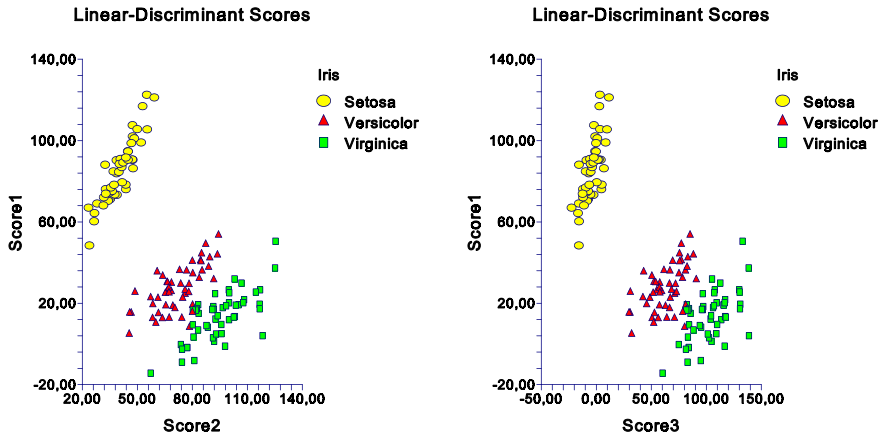
Status	Nezávisle proměnná	% změny v lambda	*F*-test	Spočtená hladina α	R^2 ostatních *X*
In	SepalLength	6.15	4.72	0.010329	0.858612
In	SepalWidth	23.35	21.94	0.000000	0.524007
In	PetalLength	33.08	35.59	0.000000	0.968012
In	PetalWidth	25.70	24.90	0.000000	0.937850

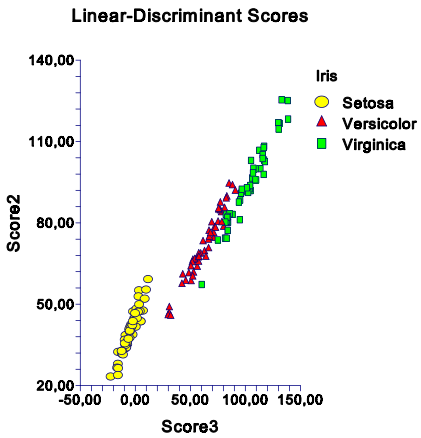
Celkové Wilkovo lambda = 0.023439

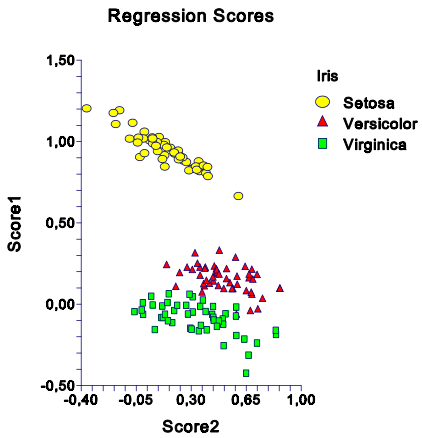
Tabulka *Automatický výběr diskriminátorů* se provádí krokově: nejprve se nalezne nejlepší diskriminátor a potom druhý nejlepší. Když byly nalezeny první dva, prověří se, zda diskriminace bude tak dokonalá, jako když byl jeden diskriminátor odebrán. Postupný (či krokový) proces přidávání nejlepšího zbývajícího diskriminátoru s následným ověřením, zda by jeden aktivní diskriminátor mohl být odebrán, pokračuje, dokud není žádný nový diskriminátor k dispozici. U nového diskriminátoru se ověřuje, zda jeho *F*-hodnota má pravděpodobnost menší než uživatelem zadaná vstupní hodnota hladiny významnosti α = 0.05. **Přehled výběru proměnných:** obsahuje protokol o činnosti v každém kroku. **Iterace:** uvádí pořadové číslo (index) kroku. **Činnost v tomto kroku:** uvádí zda diskriminátor byl zaveden do souboru aktivních diskriminátorů nebo odstraněn z tohoto souboru. **% změny v lambda:** procento snížení v hodnotě lambda, jež je výsledkem tohoto kroku. Všimněte si, že Wilkovo lambda je analogické (1-R^2) ve vícenásobné regresi. Abychom zlepšili model, budeme žádat snížit Wilkovo lambda. Např. od iterace 2 k iteraci 3 se lambda sníží z hodnoty 0.036884 na 0.024976. To je 32.29% snížení hodnoty lambda. ***F*-test:** jde o *F*-kritérium k testování statistické významnosti tohoto diskriminátoru. Je-li diskriminátor zaveden, testuje se hypotéza, že diskriminátor je třeba přidat. Je-li diskriminátor odstraněn, testuje se hypotéza, že diskriminátor je třeba odstranit. **Spočtená hladina významnosti α:** od uvedeného *F*-testu. **Wilkovo lambda:** víceparametrické rozšíření R^2 redukuje (1-R^2) ve dvojtřídě. Může být vysvětleno právě opačně než R^2. Mění se v intervalu od 1 do 0. Hodnoty blízko 1 vedou k nízké prediktibilitě, zatímco hodnoty blízko 0 k vysoké. Wilkovo lambda odpovídá právě aktivním diskriminátorům.

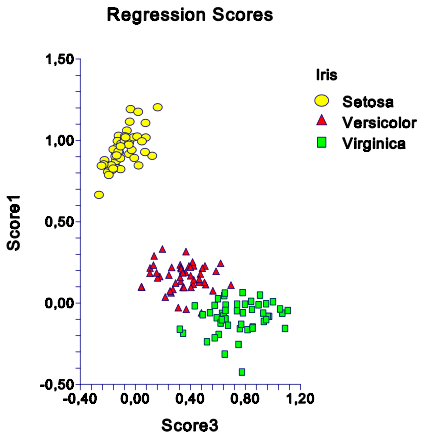
11. Výklad grafů diskriminace všech objektů do tříd:

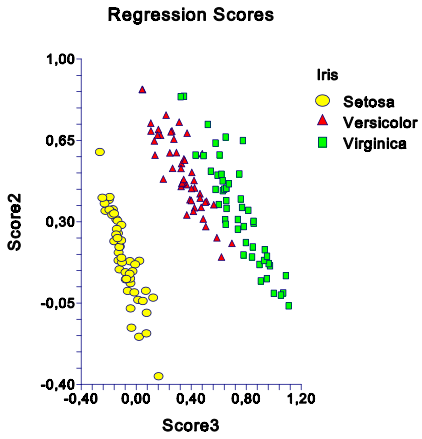
Nabízí se několik zobrazení (a) lineárních diskriminačních skóre, (b) regresních skóre nebo (c) kanonických skóre: Na základě diagramů těchto tří druhů skóre pak snáze vytvoří svou interpretaci. Diagramy totiž poskytnou vizuální vysvětlení, jak diskriminační funkce klasifikují objekty v datech. Předložený diagram ukazuje hodnoty prvního a druhého kanonického skóre. Z grafu je patrné klasifikační pravidlo: první kanonická funkce postačuje k diskriminování mezi kosatci, protože třídy kosatců mohou být snadno odděleny vertikální osou. Existuje software (S-Plus), který umožňuje 3D zobrazení s rotací podél os v prostoru. Potom by bylo vytvoření a rozlišení tříd kosatců ještě názornější.

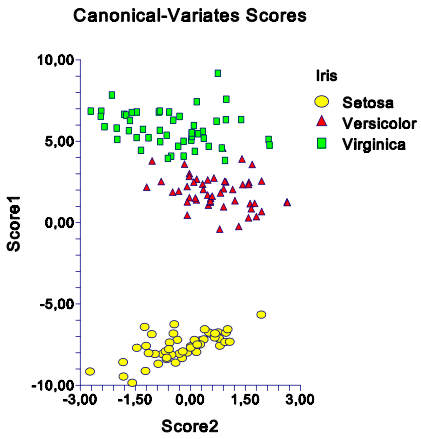












Obr. 4.15 Graf lineárního diskriminačního skóre (Linear Dis□riminant S□ores - 1 vs. 2, 1 vs. 3, 2 vs. 3).

Obr. 4.16 Graf regresního skóre (Regression S□ores - 1 vs. 2, 1 vs. 3, 2 vs. 3)

Obr. 4.17 Graf kanoni□ký□h proměnný□h (Canoni□al S□ores - 1 vs. 2).