Diskriminačná analýza (CVA, DFA)

Danka Haruštiaková

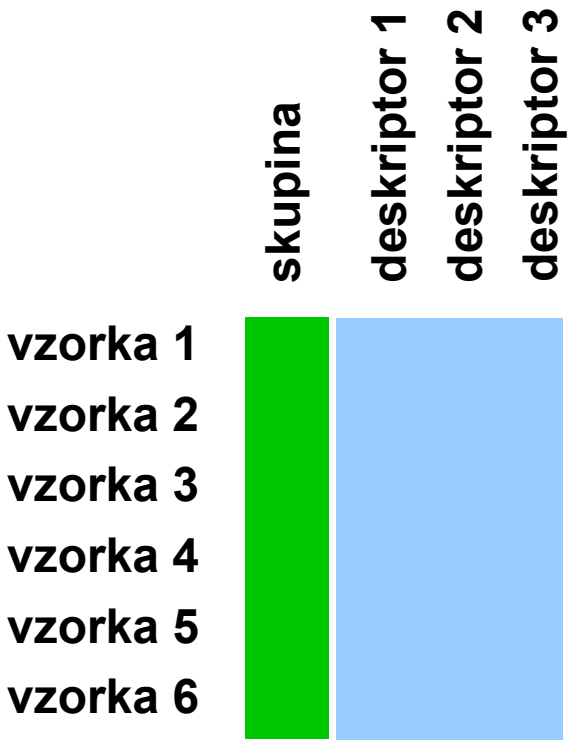
Podzim 2009

Inštitút bioštatistiky a analýz, Masarykova univerzita

Diskriminačná analýza

Diskriminačná analýza

- sa snaží zjednodušiť viacrozmernú štruktúru dát výpočtom súhrnných osí (diskriminačnej funcie).
- Je jednou z metód **ordinácie**.
- Logika osí v diskriminačnej analýze je, že maximálne diskriminuje skupiny.



- Zaoberá sa závislosťou jednej **kvalitatívnej premennej** na niekoľkých kvantitatívnych premenných.
- Objekty charakterizované sériou deskriptorov (parametrov) – kvantitatívne parametre. Známa príslušnosť objektov do skupín.

Diskriminačná analýza

Diskriminačná analýza testuje hypotézy

Ciele diskriminačnej analýzy:

- **Interpretácia rozdielov** – kanonická diskriminačná analýza

  a) či a do akej miery je možné odlíšiť stanovené skupiny objektov na základe znakov, ktoré máme k dispozícii

  b) ktoré znaky k tomuto odlíšeniu prispievajú najväčšou mierou

- **Identifikácia objektov** – klasifikačná diskriminačná analýza

  Odvodenie jednej alebo viacerých rovníc za účelom idetifikácie nových objektov

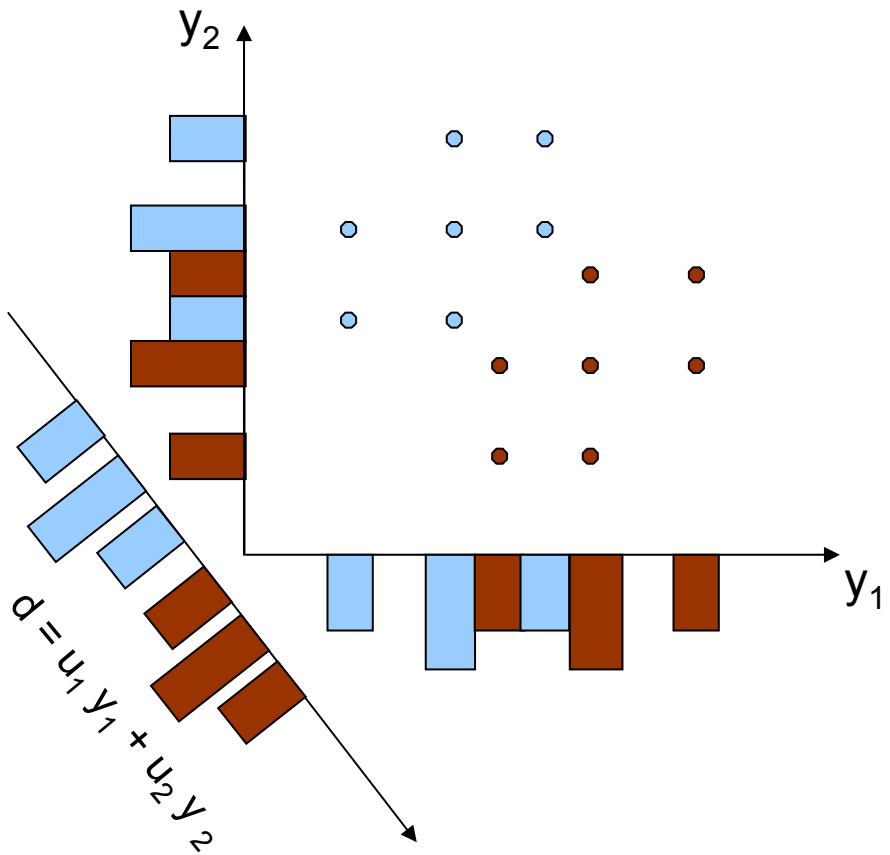
Diskriminačná analýza

- Analýza nachádza takú kombináciu vstupných parametrov, ktorá oddeľuje od seba skupiny objektov

Skupina	y_1	y_2
A	3	5
A	3	7
A	5	5
A	5	7
A	5	9
A	7	7
A	7	9
B	6	2
B	6	4
B	8	2
B	8	4
B	8	6
B	10	4
B	10	6

Kvalitatívna premenná

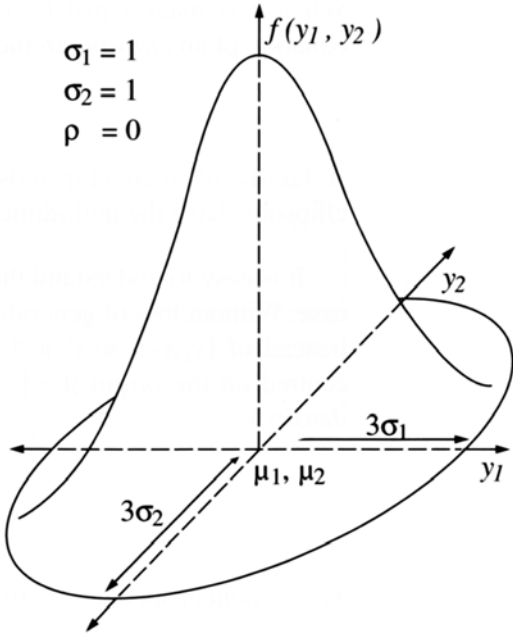
Kvantitatívne premenné (deskriptory)



Diskriminačná analýza

Požiadavky na dáta:

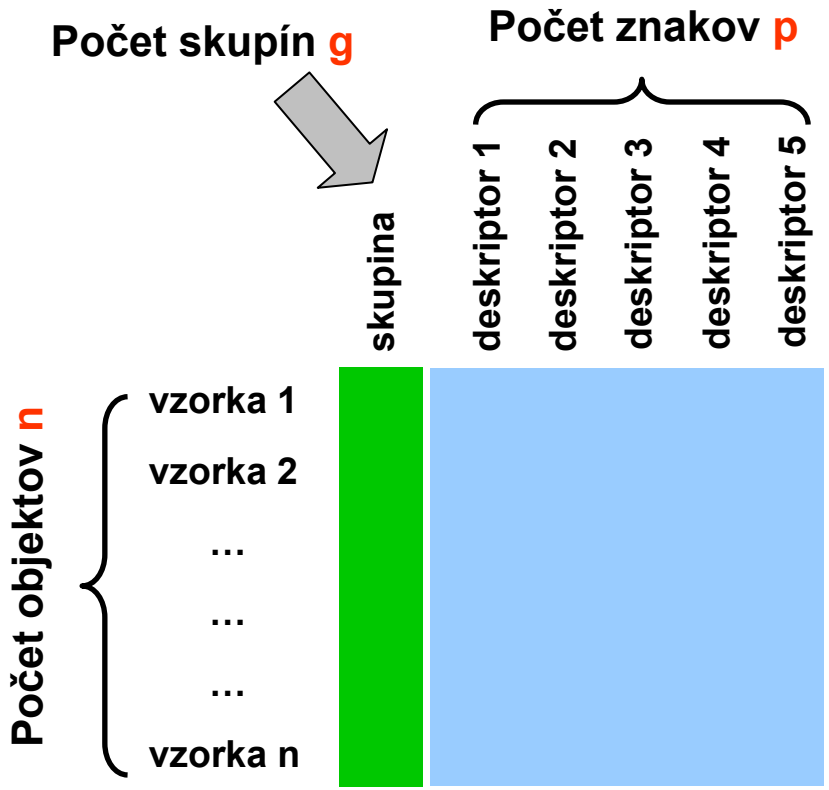
1. Kvantitatívne alebo binárne znaky
2. Žiadny zo znakov nesmie byť lineárnou kombináciou iného znaku alebo iných znakov
3. Nedá sa súčasne používať dva alebo viac silne korelovaných znakov
4. Kovariančné matice pre jednotlivé skupiny musia byť približne zhodné
5. Znaky charakterizujúce každú skupinu by mali spĺňať požiadavku mnohorozmerného normálneho rozdelenia



Diskriminačná analýza

Pre počty skupín (g), znakov (p) a objektov (n) musí platiť:

1. Musia byť aspoň dve skupiny objektov: g ≥ 2
2. V každej skupine musia byť najmenej 2 objekty
3. Počet znakov použitých v analýze musí byť menší než počet objektov znížený o počet skupín: 0 < p < (n-g)
4. Žiadny znak by nemal byť v niektorej skupine konštantný



Diskriminačná analýza

diskriminačná funkcia (kanonická)

$f_{km} = a_0 + a_1 x_{1km} + a_2 x_{2km} + ... + a_p x_{pkm}$,

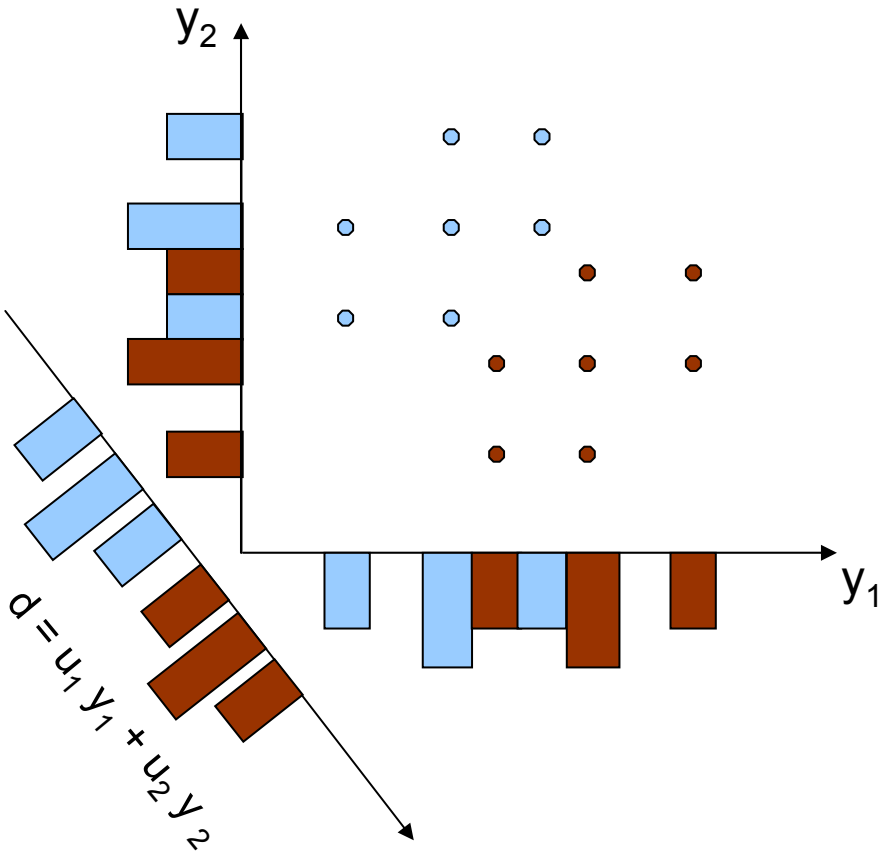
f_{km}	hodnota (skóre) kanonickej diskriminačnej funkcie pre prípad m v skupine k;
x_{ikm}	hodnota diskriminačného znaku x_i pre prípad m v skupine k
a_i	koeficienty diskriminačnej funkcie (i = 0, 1 ..., p);

Koeficienty (a) pre prvú funkciu sa odvodia tak, aby skupinové ťažiská (centroidy, priemery) boli maximálne vzdialené (v zmysle Mahalanobisovej vzdialenosti). Koeficienty vypočítané pre druhú funkciu musia ďalej maximalizovať rozdiely medzi skupinovými centroidmi a súčasne hodnoty oboch funkcií nesmú byť korelované.

Diskriminačná analýza

- Výsledkom diskriminačnej analýzy je diskriminačná funkcia (koeficienty deskriptorov).
- Premenné s najväčšími (štandardizovanými) koeficientami najviac prispievajú k predikcii príslušnosti do skupín.

Skupina	Raw coefficients	Standardized coefficients
y_1	-0.6124	-1.0
y_2	0.6124	1.0
konštanta	0.6124	
vlastná hodnota	3.9375	3.9375



- Počet diskriminačných funkcií je rovný počtu skupín znížený o jednu (prípadne počtu premenných, ak je tento nižší jako g-1)

Diskriminačná analýza

Koeficienty diskriminačnej funkcie

neštandardizované koeficienty *raw coefficients*
štandardizované koeficienty *standardized coefficients*

Klasifikačná diskriminačná analýza

1. Hľadanie identifikačného (klasifikačného) kritéria
   skupiny objektov známeho zaradenia
   skupina objektov neurčitého postavenia
2. Zistenie účinnosti klasifikačného kritéria
   resubstitúcia (resubstitution)
   krížové overenie (cross-validation)

Účinnosť klasifikačného kritéria: testovanie *cross validation*, *resubstitution*.

Diskriminačná analýza

Kroková diskriminačná analýza (*stepwise discriminant analysis; forward stepwise*)

Kroková diskriminančná analýza vyhľadáva takú kombináciu prediktorov, ktoré spoločne umožňujú čo najlepšie oddelenie stanovených skupín.

Súbor najvhodnejších prediktorov je vyberaný postupne, v jednotlivých krokoch.

Metóda začíná selekciou prediktoru, ktorý je najlepší na oddelenie vopred stanovených skupín, v ďalšom kroku posudzuje všetky zostávajúce prediktory a hľadá taký, ktorý skupiny najlepšie oddeľuje v kombinácii s už vybraným prediktorom.

V každom kroku sa počíta štatistická významnosť vybraných prediktorov (hodnota „*F-to-remove*“, *statistics for removal*) a štatistická významnosť ostatných prediktorov (hodnota „*F-to-enter*“, *statistics for entry*).

Diskriminačná analýza

Vstup diskriminačnej analýzy:

- Tabuľka objektov charakterizovaných niekoľkými kvantitívnymi parametrami a jednou kvalitatívnou premennou (ktorá priraďuje objektom príslušnosť ku skupine)

Výstup diskriminačnej analýzy:

- Ordinačný diagram (osami sú korene, čiže diskriminačné funkcie)
- Korene diskriminačnej analýzy (koeficienty diskriminačných funkcií)

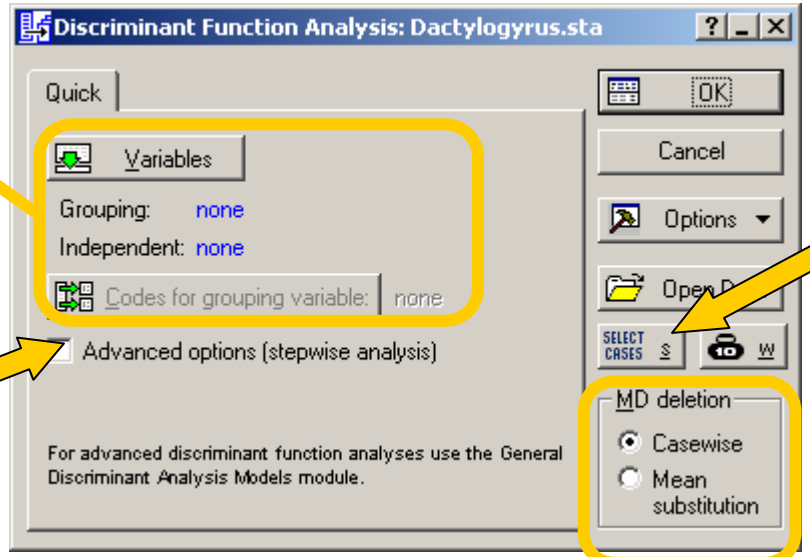
Pri použití diskriminačnej analýzy je potrebné pamätať na obmedzenia:

- parametrická metóda; vyžaduje normálne rozdelenie deskriptorov
- problém odľahlých hodnôt
- závislé na rozložení premenných
- výsledky udáva v pravdepodobnostiach
- nie je schopná zachytiť nelineárne vzťahy medzi prediktormi
- nedá sa použiť na silne korelované prediktory

Diskriminačná analýza v Statistica

Diskriminační analýza na základě námi daného rozdělení objektů do skupin vytváří model pro jejich rozdělení podle parametrů

Nastavení proměnných s hodnotami a se skupinami + definice rozlišovaných skupin



Výběr z dat

Rozšířené možnosti specifikování modelu

Smazání chybějících dat nebo jejich nahrazení průměrem

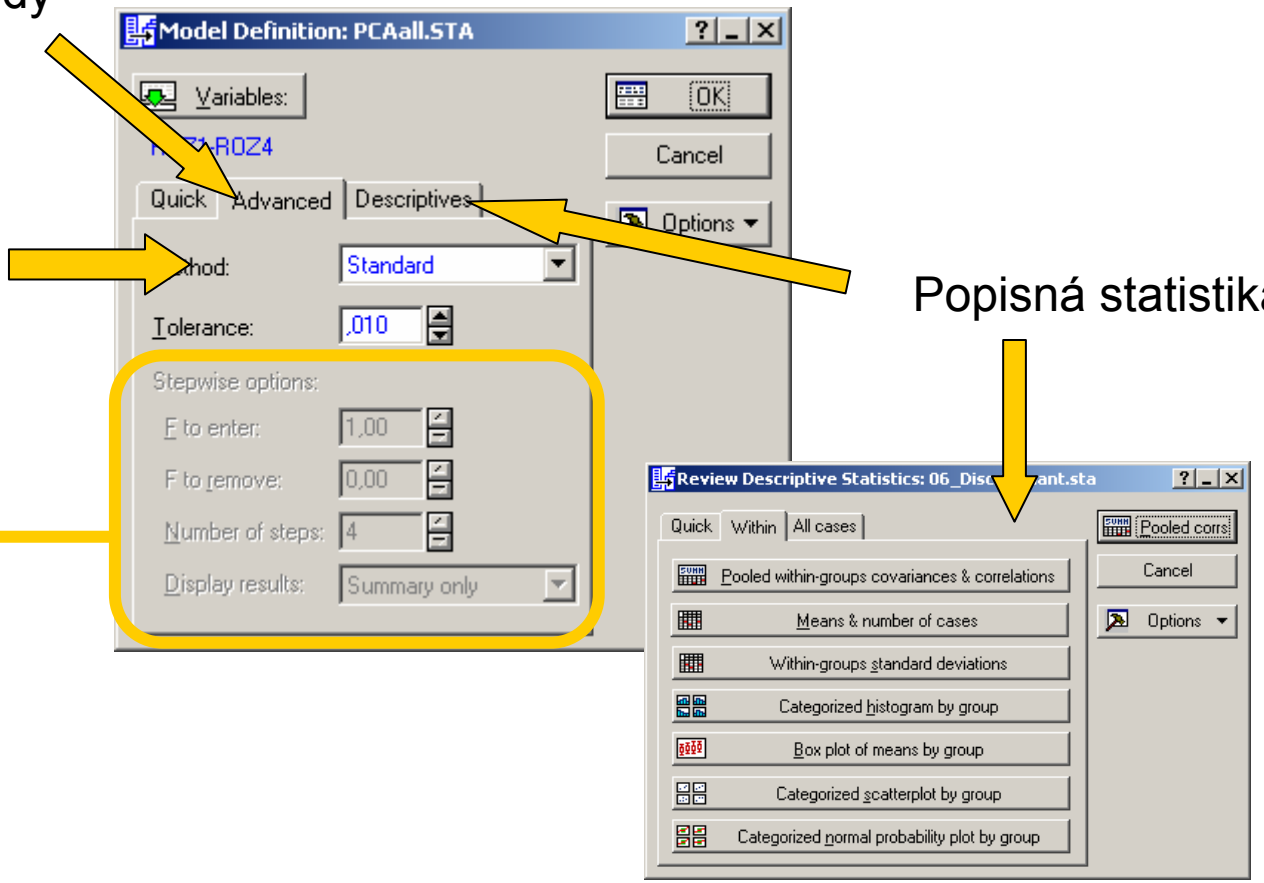
Diskriminačná analýza v Statistica

Definice modelu

Rychlé nastavení metody

Typ metody:
- Standartní
- Forward stepwise
- Backward stepwise

Nastavení stepwise metod



Popisná statistika

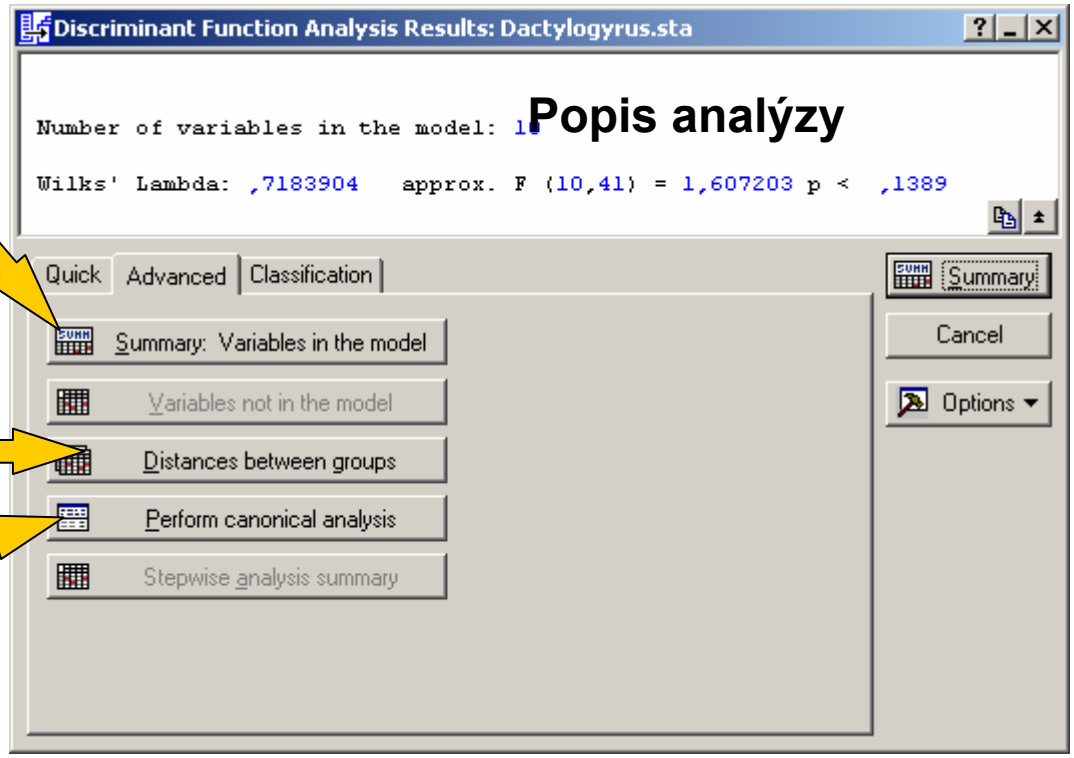
Diskriminačná analýza v Statistica

Výsledky

Popis výsledků – příspěvek jednotlivých proměnných k diskriminaci objektů

Vzdálenosti diskriminovaných skupin

Kanonická analýza



Diskriminačná analýza v Statistica

Výsledky tabulky

F spojené s danou WL

Wilk`s Lambda po vyjmutí parametru (0=perfektní diskriminace, 1=žádná diskriminace)

p spojené s daným F to remove

Discriminant Function Analysis Summary (06_Discriminant.sta)
No. of vars in model: 7; Grouping: D_UH (2 grps)
Wilks' Lambda: ,01612 approx. F (7,478)=4167,7 p<0,0000

N=486	Wilks' Lambda	Partial Lambda	F-remove (1,478)	p-level	Toler.	1-Toler. (R-Sqr.)
ROZ1	0,016500	0,976988	11,2589	0,000856	0,432261	0,567739
ROZ2	0,026691	0,6[illegible]971	313,4290	0,000000	0,58[illegible]312	0,415688
ROZ3	0,017607	0,9[illegible]584	44,0711	0,000000	0,7[illegible]944	0,276056
ROZ4	0,017084	0,9[illegible]588	28,5772	0,000000	0,4[illegible]628	0,535372
ROZ5	0,016169	0,9[illegible]7022	1,4279	0,232698	0,6[illegible]677	0,332323
ROZ6	0,016212	0,9[illegible]4356	2,7133	0,100175	0,90[illegible]271	0,097729
ROZ8	0,016610	0,9[illegible]0503	14,5281	0,000156	0,78[illegible]792	0,237208

R^2 (spjato s tolerance)

parametry

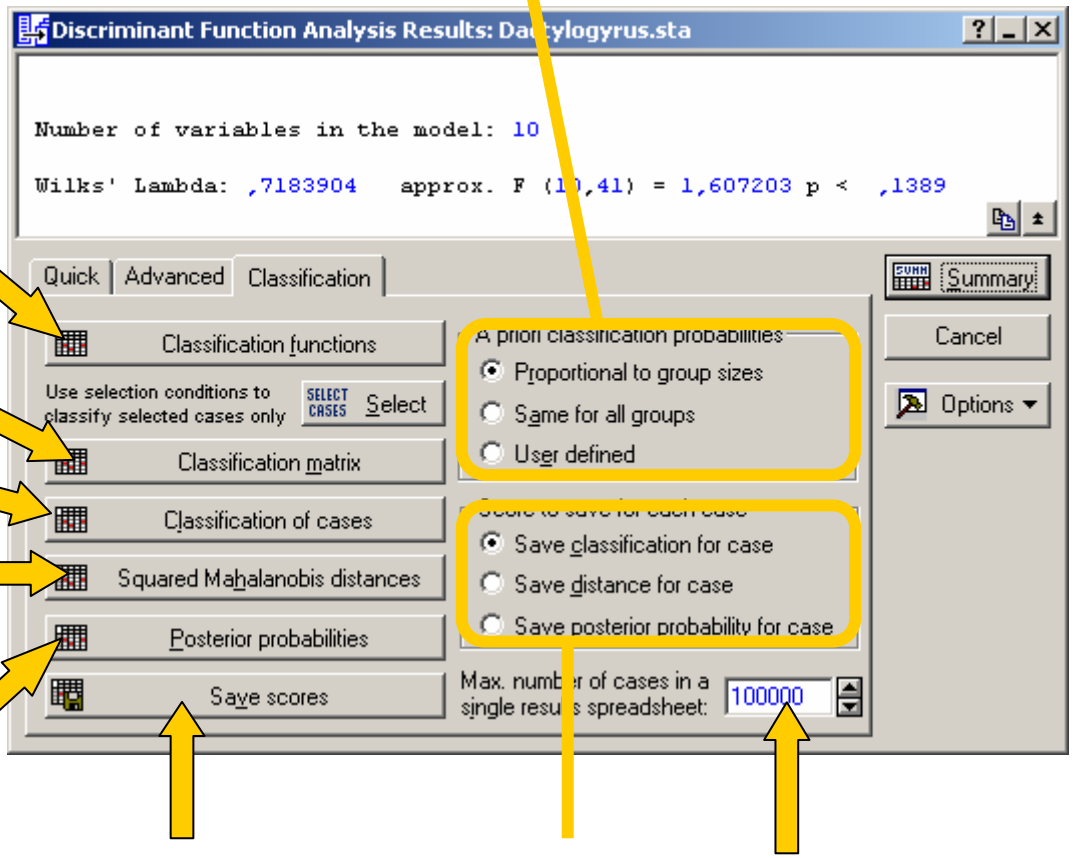
Wilk`s Lambda spojená s unikátním příspěvkem parametru k diskriminační síle modelu

Tolerance = měřítko redundance

Diskriminačná analýza v Statistica

Výsledky klasifikace

Předem nastavená pravděpodobnost zařazení do skupiny



Klasifikační funkce

Pozorované a vypočítané příslušnosti do skupin

Klasifikace objektů

Mahalanobisova vzdálenost^2 objektů od centroidů skupin

Pravděpodobnost zařazení

Uložení klasifikace (jaký parametr a kolik objektů uložit)

Diskriminačná analýza v Statistica

Výsledky klasifikace

Koeficienty klasifikační funkce

Variable	Classification Functior	
	PBIN p=,63374	PAN p=,36626
ROZ1	956,913	1923,03
ROZ2	6960,975	11766,81
ROZ3	5447,041	7612,83
ROZ4	1054,730	2527,01
ROZ5	28,245	509,99
ROZ6	2333,167	1509,32
ROZ8	2047,701	1062,15
Constant	-359,064	-861,43

Objekt patří do skupiny pro kterou mu vyšla vyšší hodnota funkce

Vzdálenost do centroidů

Vzdálenost od centroidů

Objekt

Jeho klasifikace

Case	Observed Classif.	PBIN p=,63374	PAN p=,36626
408	PBIN	4,4234	328,2919
101	PAN	140,4836	25,3236
374	PBIN	7,4163	295,4637
375	PBIN	3,3083	262,1007
376	PBIN	4,5284	298,2785
289	PAN	264,9879	8,8166
290	PAN	240,0623	6,6247
291	PAN	265,5203	2,6734
301	PAN	248,9952	8,3603
605	PBIN	5,9409	289,9818
606	PBIN	5,5818	292,0057

Diskriminačná analýza v Canoco

Canonical Variates Analysis (CVA), discriminant analysis (DFA)

Možnosť spočítať CVA v Canoco:

1. zvoliť kanonickú korešpondenčnú analýzu (CCA)
2. rozdelenie vzoriek do skupín vo forme druhových dát, ktoré sú binárne a charakterizujú príslušnosť vzorky ku skupine
3. charakteristiky prostredia ako environmentálne dáta
4. zvoliť Hillovo škálovanie so zameraním na inter-species distances

V súbore **.sol**:

- species scores sú stredmi zhlukov v CVA ordinačnom diagrame
- sample scores, ktoré sú lineárnou kombináciou charakteristík prostredia sú škálované tak, že rozptyl v rámci skupín sa rovná 1

Permutačný test môže byť použitý na hodnotenie rozdielov medzi skupinami.

Je možné špecifikovať aj kovariáty => parciálna CVA = one-way Multivariate Analysis of Covariance (MANOCO).