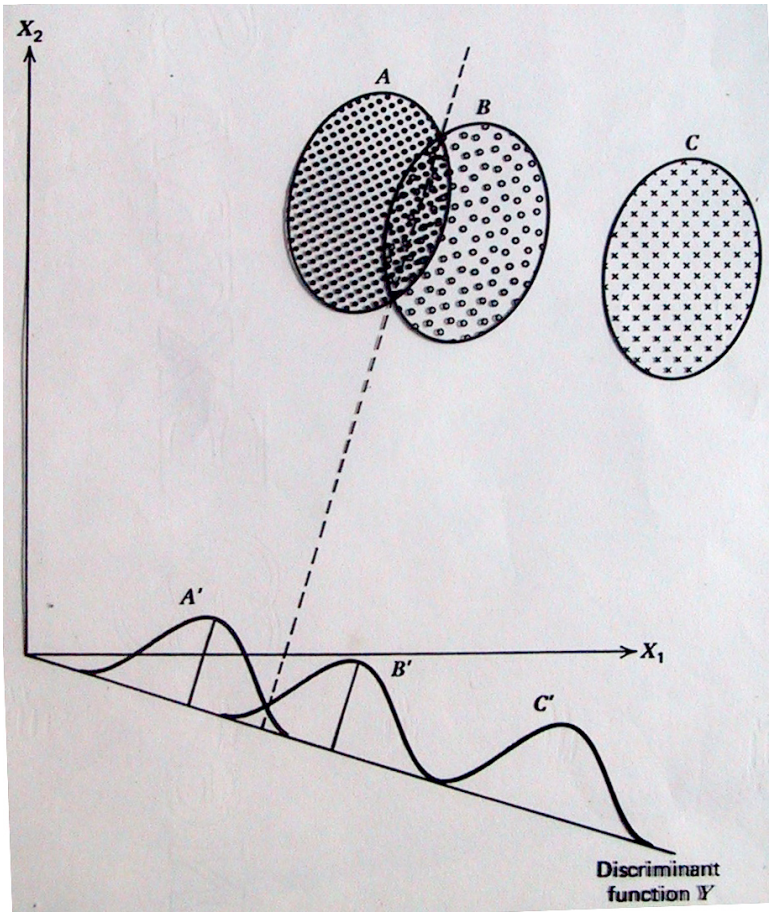
DISKRIMINANTNA ANALIZA



Z diskriminantno analizo poiščemo tako linearno kombinacijo merjenih spremenljivk, da bo maksimalno ločila **vnaprej** določene skupine in da bo napaka pri uvrščanju enot v skupine najmanjša.

Pri diskriminantni analizi torej gre za iskanje tistih razsežnosti, ki kar najbolj pojasnjujejo razlike med skupinami (**pojasnjevanje**) in za kar se da dobro prirejanje enot vnaprej danim skupinam (**napovedovanje**).

Predpostavke

1. $k \geq 2$.
2. Vsaj 2 enoti v vsaki skupini.
3. $p < n - 2$; p je število spremenljivk in n število vseh enot v vzorcu.
4. Nobena spremenljivka ne sme biti linearna kombinacija preostalih spremenljivk (multikolinearnost).
5. Pri statističnemu ocenjevanju se predpostavlja, da so v vsaki skupini enot (vzorcu) enote slučajno izbrane iz populacije, kjer se spremenljivke porazdeljujejo večrazsežno normalno.
6. Variančno-kovariančna matrika $p \times p$ je v vsaki populacijski skupini enaka.

Diskriminantna analiza v primeru dveh skupin (kanonična diskriminantna analiza, Fisher 1936)

skupini	vektorja ar. sredin	variančno-kovariančna m.
G_1	μ_1	Σ_1
G_2	μ_2	Σ_2

Predpostavka: $\Sigma_1 = \Sigma_2 = \Sigma$

Fisher je definiral diskriminantno spremenljivko (funkcijo) Y kot linearno kombinacijo p merjenih spremenljivk X_i

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p = Xb$$

tako da je kvocient razlik aritmetičnih sredin diskriminantne spremenljivke v obeh skupinah (G_1 in G_2) glede na varianco diskriminantne spremenljivke znotraj skupine maksimalen.

Aritmetični sredini diskriminantne spremenljivke v skupinama G_1 in G_2 sta:

$$\bar{y}_1 = b'\mu_1$$

$$\bar{y}_2 = b'\mu_2$$

Varianci diskriminantne spremenljivke v vsaki skupini pa:

$$\operatorname{var} Y_1 = \operatorname{var} Y_2 = b'\Sigma b$$

Kvocient, ki naj bi bil maksimalen, pa je:

$$\frac{b'\mu_1 - b'\mu_2}{b'\Sigma b} = max$$

To je pogoj, na osnovi katerega izračunamo uteži b_i najboljše diskriminantne spremenljivke. Reševanje tega optimizacijskega problema privede do rešitve za b, ki je sorazmerna

$$\Sigma^{-1}(\mu_1 - \mu_2)$$

(Rešitev je lahko pomnožena s poljubno konstanto.)

Vzorčne ocene

Ponavadi imamo vzorčne podatke za vsako populacijo G_i, iz katerih ocenimo μ_i in Σ.

Vzorčne ocene za μ_i so:

$$\bar{x}_i' = (\bar{x}_{i1}, \bar{x}_{i2}, ..., \bar{x}_{ip})$$

in za Σ (angl. pooled sample variance-covariance matrix)

$$S = \frac{1}{n_1 + n_2 - 2}(X_1'X_1 + X_2'X_2)$$

kjer je n_1 število enot v vzorcu iz G_1 in n_2 število enot v vzorcu iz G_2.

Ocena uteži je tedaj

$$\hat{b} = S^{-1}(\bar{x}_1 - \bar{x}_2)$$

Centroid skupine

Aritmetično sredino diskriminantne spremenljivke v določeni skupini imenujemo **centroid skupine**.

Centroid skupine i je

$$\bar{y}_i = b'\bar{x}_i$$

Pravila uvrščanja enot v skupine

Denimo, da smo izračunali diskriminantno spremenljivko $Y = Xb$. Pravilo uvrščanja enote v optimalno skupino je tedaj: i-to enoto (glede na p izmerjenih spremenljivk) uvrstimo v skupino G_1, če je njena vrednost na diskriminantni spremenljivki y_i

$$y_i - \bar{y}_1 \leq y_i - \bar{y}_2$$

ali v skupino G_2, če je

$$y_i - \bar{y}_1 > y_i - \bar{y}_2$$

Ekvivalenten pogoj uvrščanja je **metoda srednje točke** kot točke ločevanja skupin. Če imata skupini enako število enot ($n_1 = n_2$), je točka ločevanja

$$y_c = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

Če imamo skupini neenakih velikosti ($n_1 \neq n_2$), točko ločevanja izračunamo takole

$$y_c = \frac{n_2\bar{y}_1 + n_1\bar{y}_2}{n_1 + n_2}$$

Klasifikacijska tabela

Obnašanje diskriminatne spremenljivke lahko ocenimo tudi z deležem (ne)pravilno uvrščenih enot. Uporabimo dobljeno diskriminantno spremenljivko na podatkih, iz katerih je bila izračunana. Vsako enoto enoto uvrstimo v eno od obeh skupin glede na pravilo uvrščanja. Rezultate uvrščanja lahko predstavimo z naslednjo tabelo:

dejanski skupini	število enot	dobljeni skupini G_1	G_2
G_1	n_1	a	b
G_2	n_2	c	d

Delež pravilno uvrščenih enot je

$$\frac{a+d}{n_1+n_2}$$

Ob predpostavki, da je v obeh skupinah enako število enot, je spodnja meja tega deleža 0.50.

Ocenjeni delež pravilno uvrščenih enot je optimistično pristranski, ker pri tem uporabljamo iste podatke za pravilo uvrščanja in oceno obnašanja.

Primer: majhna podjetja

Populacijo sestavljajo majhna podjetja, v katerih je zaposlenih med 1 in 50 v vseh sektorjih razen v kmetijstvu v Sloveniji. Slučajni vzorec je bil dobljen na osnivi spiskov podjetij v Gospodarski zbornici R Slovenije in Obrtni zbornici R Slovenije. Od 200 slučajno izbranih podjetij, jih je v anketi sodelovalo 151. Podatki so bili zbrani z osebnim intervjujem. Zbiranje podatkov je potekalo v letu 1993. Skupini sta bili določeni takole: G_1 – storitvena podjetja ($n_1 = 70$) in G_2 – obrtna podjetja ($n_2 = 75$). Merjene spremenljivke so 12 faktorjev, ki pomagajo k uspešnosti podjetja.

Dobljene uteži diskriminantne spremenljivke so:

	uteži
PROD-MET	**-.54**
MARK-MET	**.40**
PRODUKT	-.00
ODNOSI	.01
USP-ZAP	.22
USP-MAN	**.51**
DRUŽINA	-.33
GOSP-ZDR	-.18
POL-ZVEZE	**.48**
LOK-OBL	-.28
DRŽAVA	.16
PODJETJA	.06

Centroidi so:

servis	.54
obrt	-.50

Klasifikacijska tabela je:

dejanski skupini	število enot	dobljeni skupini servis	dobljeni skupini obrt
servis	70	70%	30%
obrt	75	30.7%	69.3%

Odstotek pravilno uvrščenih enot je 70%.

Lastniki servisnih podjetij bolj verjamejo kot latniki obrtnih podjetij, da je izboljšanje produktov manj pomembno in bolj pomembna usposobljenost managerjev, dobre politične zveze in izboljšanje marketinških pristopov.

Diskriminantna analiza na več skupinah

V primeru več skupin razlike med skupinami popišemo z več diskriminantnimi spremenljivkami. Največ jih je $\min(p, k-1)$.

Postopek za izračun diskriminantnih spremenljivk

Označimo vsoto kvadratov in produktov odklonov od skupnega povprečja $\bar{x}$ (vse skupine skupaj)

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})(x_{ij} - \bar{x})'$$

in isto za posamezno skupino

$$W_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})(x_{ij} - \bar{x_i})'$$

Variabilnost znotraj skupin je:

$$W = W_1 + W_2 + ... + W_k$$

ker velja (podobno kot v univariatni analizi):

$$T = W + B$$

je

$$B = T - W$$

kjer je B vsota kvadratov in produktov odklonov med skupinami.

Diskriminantni kriterij, ki ga je potrebno maksimizirati, je podoben kot v primeru dveh skupin:

$$\frac{\text{variabilnost med skupinami}}{\text{variabilnost znotraj skupin}} = max$$

Varianca diskriminantne spremenljivke $Y = Xb$ je

$$\operatorname{var} Y = b'\Sigma b$$

Varianca med skupinami je

$$\operatorname{var} Y = b'Bb$$

Varianca znotraj skupin pa

$$\operatorname{var} Y = b'Wb$$

Tedaj je diskriminantni kriterij, ki ga je potrebno maksimizirati:

$$\frac{b'Bb}{b'Wb} = \max$$

Rešitev dobimo tako, da izraz odvajamo po b in dobimo

$$(B - \lambda W)b = 0$$

oziroma

$$(W^{-1}B - \lambda I)b = 0$$

Rešitev enačbe so lastne vrednosti in lastni vektorji matrike $W^{-1}B$. Lastna vrednost λ_i je sorazmerna s skupno varianco, ki je pojasnjena z i-to diskriminantno spremenljivko, ustrezni lastni vektor pa predstavlja uteži i-te diskriminantne spremenljivke. Delež skupne variance, pojasnjene z i-to diskriminantno spremenljivko, je

$$\frac{\lambda_i}{\Sigma \lambda_i}$$

Zveza med diskriminantno analizo in kanonično korelacijsko analizo

V primeru diskriminantne analize imamo k skupin in p merjenih spremenljivk. Za oceno diskriminatnih spremenljivk računamo lastne vrednosti in vektorje matrike $W^{-1}B$. Označimo lastne vrednosti z λ_j^{da}.

Naredimo iz nominalne spremenljivke, ki določa k skupin v diskriminantni analizi, $(k-1)$ 'dummy' spremenljivk. Po drugi strani imamo p merjenih spremenljivk. Za ti dve skupini spremenljivk lahko izračunamo $\min((k-1), p)$ kanoničnih rešitev. Te dobimo z izračunom lastnih vrednosti in lastnih vektorjev matrike $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$, ki jih označimo z λ_j^{kka}.

Velja zveza:

$$\lambda_j^{da} = \frac{\lambda_j^{kka}}{1 - \lambda_j^{kka}}$$