

Statistika Industri II
TIP - FTP – UB

Analisis Diskriminan

Analisis Diskriminan

- Mirip regresi linier berganda
- Metode dependen

$$\text{Non-Metrik } Y_1 = \text{Metrik } X_1 + X_2 + X_3 + \dots + X_n$$

■ Dimana :

- Variabel Independen (X_1 dan seterusnya) adalah data metrik, yaitu data berskala interval atau rasio.
- Variabel Dependen (Y_1) adalah data kategorikal atau nominal. Jika data kategorikal tersebut hanya terdiri dari 2 kode saja disebut "Two-Groups Discriminant Analysis". Namun apabila lebih dari 2 kategori disebut "Multiple Discriminant Analysis".

Tujuan analisis diskriminan

- Mengetahui perbedaan yang jelas antar grup pada variabel dependen.
- Jika ada perbedaan, variabel independen manakah pada fungsi diskriminan yang membuat perbedaan tersebut.
- Membuat fungsi atau model diskriminan (yang mirip dengan persamaan regresi).
- Melakukan klasifikasi terhadap obyek ke dalam kelompok (grup).

$$D = b_0 + b_1X + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

- Keterangan:
 - D = skor diskriminan
 - b = koefisien diskriminasi atau bobot
 - X = prediktor atau variabel independent

Jumlah sampel

- Pedoman umum untuk variabel independen sebaiknya sampel sebanyak **5 – 20 kali variabel**.
Misal: ada 6 variabel independen maka jumlah sampel minimal sebanyak $6 \times 5 = 30$ sampel.
- Pada analisis diskriminan, sebaiknya menggunakan dua jenis sampel, yaitu **sampel analisis (*analysis sample*)** yang digunakan untuk membuat fungsi diskriminan dan ***holdout sample (split sample)*** yang digunakan untuk menguji hasil analisis diskriminan.
Misal: jika ada 80 sampel, maka dapat dibagi menjadi dua, 40 sampel untuk *analysis sample* dan 40 untuk *holdout sample*.

Asumsi yang dipakai

- *Multivariate normality*. Variabel independen berdistribusi normal. Jika data dari variabel independen tidak berdistribusi normal akan menyebabkan ketidaktepatan fungsi (model) diskriminan.
- Matrik kovarians dari semua variabel independen sama
- Tidak terdapat korelasi antara variabel independen. Jika terdapat dua atau lebih variabel independen yang berkorelasi maka disebut multikolinearitas. Multikolinearitas menyebabkan terjadinya bias interpretasi model.
- Tidak terdapat data pencilan (*outlier*) pada variabel independen. Jika terdapat data pencilan dan tetap dilakukan analisis diskriminan maka dapat menyebabkan ketidaktepatan klarifikasi dari fungsi diskriminan.

Langkah

- Memilah variabel-variabel menjadi variabel terikat (*Dependent*) dan variabel bebas (*Independent*).
- Menentukan metode untuk membuat fungsi diskriminan, yaitu
 - Menguji signifikansi Fungsi Diskriminan yang terbentuk, dengan menggunakan Wilk's Lambda, Nilai F test, dan lainnya.
 - Menguji ketepatan klasifikasi dari fungsi diskriminan (secara individual dengan *Casewise Diagnostics*).
- Melakukan interpretasi Fungsi Diskriminan.
- Melakukan uji validasi fungsi diskriminan

Contoh

- Seorang produsen kerajinan tangan sedang melakukan penelitian untuk mengetahui penyebab banyaknya jumlah konsumen yang melakukan pembelian secara online (2) daripada offline (1). Penelitian dilakukan terhadap 50 responden, dengan menggunakan variabel: pendapatan konsumen (X_1), daya beli konsumen (X_2), sikap pelayanan penjaga toko (X_3), dan daya tarik produk (X_4). Sampel dibagi menjadi dua, sampel yang digunakan untuk analisis dan sampel yang digunakan untuk sampel holdout.

- Sampel

Group Statistics

d		Mean	Std. Deviation	Valid N (listwise)	
				Unweight ed	Weight ed
1.00	X1	3.8857	.68259	14	14.000
	X2	5.7857	.69929	14	14.000
	X3	6.0714	.82874	14	14.000
	X4	5.9286	.82874	14	14.000
2.00	X1	4.9813	1.04640	16	16.000
	X2	4.0625	.92871	16	16.000
	X3	4.2500	.77460	16	16.000
	X4	4.5000	.51640	16	16.000
Total	X1	4.4700	1.04127	30	30.000
	X2	4.8667	1.19578	30	30.000
	X3	5.1000	1.21343	30	30.000
	X4	5.1667	.98553	30	30.000

- Perbedaan nilai rata-rata antargrup dapat digunakan untuk mengetahui peran variabel-variabel dalam pengelompokan responden.
- Standar deviasi digunakan sebagai indikator variabel berperan baik sebagai diskriminator. Jika nilai standar deviasi grup lebih rendah daripada standar deviasi total maka variabel-variabel dalam grup berperan baik.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
X1	.715	11.161	1	28	.002
X2	.465	32.176	1	28	.000
X3	.420	38.687	1	28	.000
X4	.459	33.002	1	28	.000

Pooled Within-Groups Matrices^a

		X1	X2	X3	X4
Covariance	X1	.803	-.137	-.129	-.177
	X2	-.137	.689	.142	.046
	X3	-.129	.142	.640	.038
	X4	-.177	.046	.038	.462
Correlation	X1	1.000	-.184	-.180	-.291
	X2	-.184	1.000	.213	.081
	X3	-.180	.213	1.000	.070
	X4	-.291	.081	.070	1.000

a. The covariance matrix has 28 degrees of freedom. Pooled within-groups correlation matrices mengindikasikan adanya korelasi antarvariabel predictor yang rendah.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
dimension 1 0	2.993^a	100.0	100.0	.866

a. First 1 canonical discriminant functions were used in the analysis.

Canonical correlation menunjukkan keeratan hubungan antara discriminant score dengan grup. Nilai conanical correlation sebesar 0,866 menunjukkan keeratan yang cukup tinggi, dengan ukuran skala asosiasi antara 0 sampai 1 → koefisien determinasi.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
dimension 1 0	.250	36.001	4	.000

Wilks' Lambda adalah varians total dalam nilai diskriminan yang tidak dapat dijelaskan oleh perbedaan antara grup-grup yang dianalisis. Angka chi-square sebesar 36,001 memiliki tingkat signifikansi tinggi menunjukkan adanya perbedaan yang jelas antara kedua grup responden.

Structure Matrix	
	Function
	1
X3	.679
X4	.627
X2	.620
X1	-.365

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
 Variables ordered by absolute size of correlation within function.

Table structure matrix menjelaskan korelasi antara variabel independen dengan fungsi diskriminan yang terbentuk. Berdasarkan table di atas, variabel X_3 (sikap pelayanan penjaga toko) memiliki hubungan yang paling erat terhadap fungsi diskriminan dibanding variabel lainnya.

Canonical Discriminant Function Coefficients	
	Function
	1
X1	-.029
X2	.549
X3	.674
X4	.802
(Constant)	-10.125

Unstandardized coefficients

Berdasarkan nilai canonical discriminant function coefficients dapat disusun fungsi diskriminannya. Fungsi diskriminan untuk kasus di atas adalah:

$$D = -10,125 - 0,029X_1 + 0,549X_2 + 0,674X_3 + 0,802X_4$$

Classification Results^{b,c}

d		Predicted Group Membership		Total	
		1.00	2.00		
Original	Count	1.00	13	1	14
		2.00	0	16	16
	%	1.00	92.9	7.1	100.0
		2.00	.0	100.0	100.0
Cross-validated ^a	Count	1.00	13	1	14
		2.00	1	15	16
	%	1.00	92.9	7.1	100.0
		2.00	6.3	93.8	100.0

a. Cross validation is done only for those cases in the analysis.

In cross validation, each case is classified by the functions derived from all cases other than that case.

b. **96.7% of original grouped cases correctly classified.**

c. 93.3% of cross-validated grouped cases correctly classified.

Perhitungan sample holdout

Canonical Discriminant Function Coefficients

	Function
	1
x1	-.054
x2	1.026
x3	.929
x4	.170
(Constant)	-10.470

Unstandardized coefficients

Berdasarkan nilai table conanical function coefficient di atas, fungsi diskriminan untuk sampel holdout adalah:

$$D = -10,470 - 0,054X_1 + 1,026X_2 + 0,929X_3 + 0,170X_4$$

Classification Results^{b,c}

d		Predicted Group Membership		Total
		1.00	2.00	
Original	Count	1.00	2.00	
		10	0	10
		0	10	10
	%	1.00	2.00	
		100.0	.0	100.0
		.0	100.0	100.0
Cross-validated ^a	Count	1.00	2.00	
		10	2	10
		2	8	10
	%	1.00	2.00	
		100.0	.0	100.0
		20.0	80.0	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. **100.0% of original grouped cases correctly classified.**

c. 90.0% of cross-validated grouped cases correctly classified.

- Holdout

$$D = -10,470 - 0,054X_1 + 1,026X_2 + 0,929X_3 + 0,170X_4$$

- Analysis Sample

$$D = -10,125 - 0,029X_1 + 0,549X_2 + 0,674X_3 + 0,802X_4$$

- Tabel *classification result* menunjukkan tingkat ketepatan dari model (fungsi) diskriminan. Model diskriminan sampel holdout menunjukkan nilai ketepatan yang sangat tinggi yaitu sebesar 100% (original grup) dan 90% (cross-validated grup). Nilai model di atas 50% memiliki ketepatan yang tinggi, sehingga model bisa digunakan untuk mengklasifikasi pada kasus sejenis.