L'analyse discriminante

À Propos de ce document

Ce document a été créé dans le but d'aider toute personne qui débute dans SPSS, logiciel très puissant mais très peu sympathique.

Ce document se base sur la version 11.0 Base de SPSS, en version anglaise. La plupart des exemples sont issus des dictatiels du programme SPSS en lui-même.

Toutes les remarques, tant sur le fond que sur la forme, sont les bienvenues. N'hésitez pas à me contacter à l'adresse suivante : <lemoal@lemoal.org> ou à venir visiter mon site internet : http://www.lemoal.org/spss/

Merci.

Introduction

Le but de l'analyse discriminante est d'étudier les relations entre une variable qualitative et un ensemble de variables explicatives quantitatives. C'est une méthode utilisée notamment par les banques pour le scoring

Trois objectifs principaux peuvent être assignés à l'analyse discriminante :

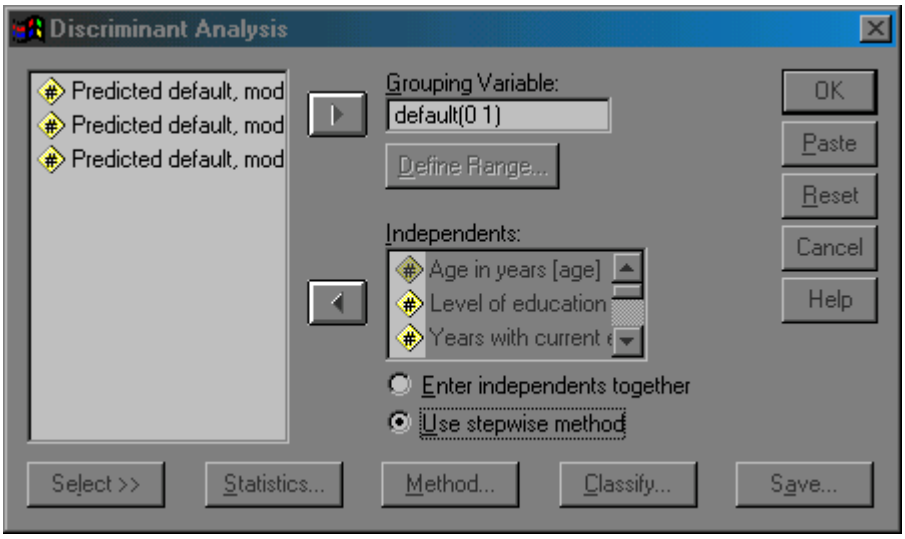
- Déterminer les variables explicatives les plus discriminantes vis à vis des classes déterminées
- Déterminer à quel groupe appartient un individu à partir de ses caractéristiques
- Mais surtout à **valider une classification** ou à **faire un choix entre plusieurs classifications pour savoir laquelle est la plus pertinente**. L'analyse discriminante intervient donc a posteriori d'une classification.

Deux conditions sont à remplir :

- Les variables explicatives doivent être métriques
- Elles ne doivent pas être trop corrélées entre elles. Cela se vérifie par l'observation des corrélations entre les variables. Si c'est le cas, on peut passer par une analyse factorielle qui permet de réduire les données à quelques axes. Ces axes sont, par propriété, non corrélés entre eux.

La démarche à suivre sous SPSS

Aller dans Analyse > Classify > Discriminant… La boîte de dialogue suivante apparaît alors :



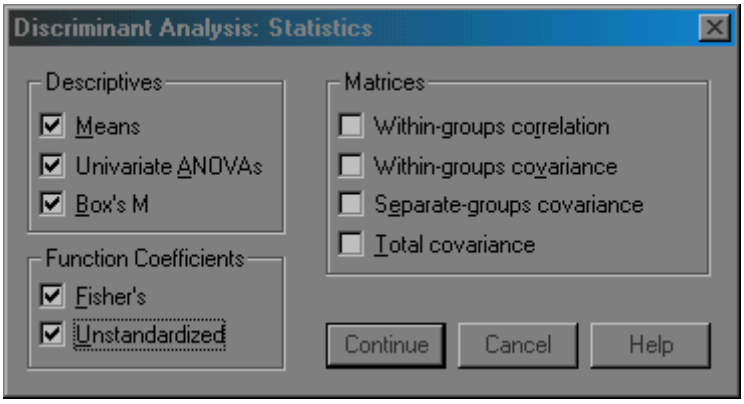
Dans « Grouping Variable » (i.e. les critère de regroupement), il faut indiquer la variable à expliquer en la sélectionnant dans la partie de droite puis en cliquant sur la flèche qui pointe vers la droite. SPSS demande alors de définir l'intervalle, c'est-à-dire les différentes modalités que la variable peut prendre.

Dans « Independents » (i.e. les variables explicatives), il faut indiquer les variables métriques que l'on souhaite intégrer à l'analyse. Il est important de choisir « Use stepwise method » (i.e. la méthode pas à pas).

Trois options s'offrent alors à nous : « Statistics… », « Method… » et « Classify… ». On ne touchera pas aux différentes options de « Méthod… »

1. Statistics…

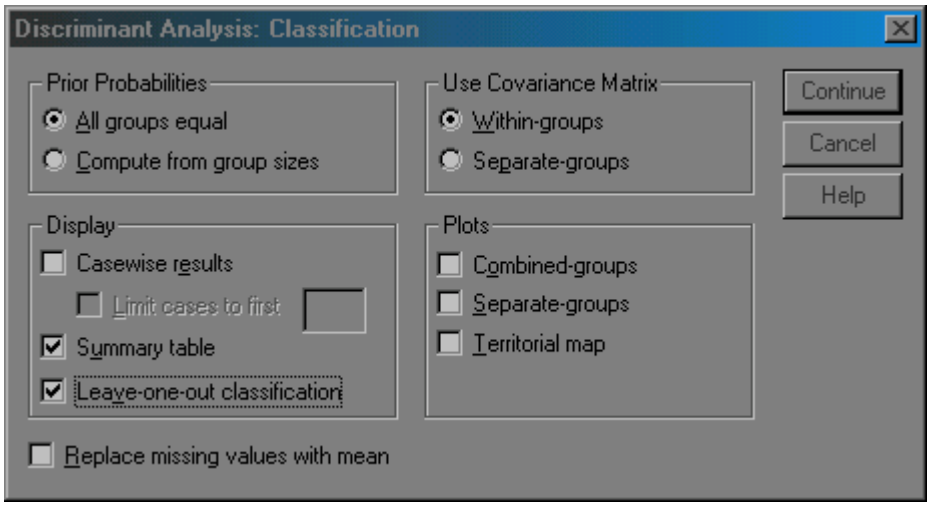
La boîte de dialogue « Discriminant Analysis : Statistics » apparaît.



Dans la boîte qui apparaît, il convient de cocher « Means » (moyennes), « Univariate ANOVAs » (ANOVA à 1 facteur) et « Box's M » (Test de Box) dans « Descriptives » et « Fischer's » ainsi que « Unstandardized » dans « Function Coefficients ».

2. Classify…

La boîte de dialogue « Discriminant Analysis : Classification » apparaît.



Dans la boîte qui apparaît, il convient de cocher « Summary Table » (option qui permet l'affichage de la matrice de confusion) et « Leave-one-out classification » dans « Display ».

Analyse des résultats

Une analyse discriminante se déroule en 3 étapes :

1. On vérifie l'existence de différences entre les groupes.
2. On valide l'étude.
3. On vérifie le pouvoir discriminant des axes.
4. On juge la qualité de la représentation du modèle.

La 3ème étape peut être passée dans la plupart des cas.

1. Vérification de l'existence de différences entre les sous-groupes.

On vérifie s'il existe bien des différences entre les groupes grâce à trois indicateurs : la moyenne ou la variance, le test du F et le Lambda de Wilks. Ils s'interprètent de la façon suivante :

	En cas d'influence	En absence d'influence
Moyenne ou variance	Différence	Similitude
Test du F	F élevé Sig F tend vers 0,000	F faible SIG F >= 0,01 ou 0,05
Lambda de Wilks	<= 0,90	Tend vers 1

Cette première analyse permet de déterminer quelles sont les variables qui sont les plus discriminantes entre les groupes.

Les moyennes et écart-types s'observent dans le tableau « Group Statistics ». Les variables « Years with current employes », « Years at current adress », « Debt to income ration » et « Credit card debt » dans l'exemple ci-dessous semblent être les variables les plus discriminantes.

Group Statistics

Previously defaulted		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
No	Age in years	35,5145	7,70774	517	517,000
	Level of education	1,6596	,90443	517	517,000
	Years with current employer	9,5087	6,66374	517	517,000
	Years at current address	8,9458	7,00062	517	517,000
	Household income in thousands	47,1547	34,22015	517	517,000
	Debt to income ratio (x100)	8,6793	5,61520	517	517,000
	Credit card debt in thousands	1,2455	1,42231	517	517,000
	Other debt in thousands	2,7734	2,81394	517	517,000
Yes	Age in years	33,0109	8,51759	183	183,000
	Level of education	1,9016	,97279	183	183,000
	Years with current employer	5,2240	5,54295	183	183,000
	Years at current address	6,3934	5,92521	183	183,000
	Household income in thousands	41,2131	43,11553	183	183,000
	Debt to income ratio (x100)	14,7279	7,90280	183	183,000
	Credit card debt in thousands	2,4239	3,23252	183	183,000
	Other debt in thousands	3,8628	4,26368	183	183,000
Total	Age in years	34,8600	7,99734	700	700,000
	Level of education	1,7229	,92821	700	700,000
	Years with current employer	8,3886	6,65804	700	700,000
	Years at current address	8,2786	6,82488	700	700,000
	Household income in thousands	45,6014	36,81423	700	700,000
	Debt to income ratio (x100)	10,2606	6,82723	700	700,000
	Credit card debt in thousands	1,5536	2,11720	700	700,000
	Other debt in thousands	3,0582	3,28755	700	700,000

Le test du F et du Lambda de Wilks s'observe dans le tableau « Tests of Equality of Group Means ».

L'examen du F dans notre exemple nous confirme que ce sont bien les variables « Years at current address », « Credit card debt in thousands », « Years with current employer », et « Debt to income ratio (x100) » qui sont les plus discriminantes.

De plus, d'après le test du Lambda de Wilks, seule la variable « Debt to income ratio (x100) » semble avoir une influence.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Age in years	,981	13,482	1	698	,000
Level of education	,987	9,301	1	698	,002
Years with current employer	,920	60,759	1	698	,000
Years at current address	,973	19,402	1	698	,000
Household income in thousands	,995	3,533	1	698	,061
Debt to income ratio (x100)	,848	124,889	1	698	,000
Credit card debt in thousands	,940	44,472	1	698	,000
Other debt in thousands	,979	15,142	1	698	,000

2. Vérification de la validité de l'étude.

On estime la validité d'une analyse discriminante à partir de indicateurs :

- Le test de Box.
- La corrélation globale.
- Le Lambda de Wilks.

On observe le test de Box grâce au tableau « Test Results ».

Test Results

Box's M		364,962
F	Approx.	36,182
	df1	10
	df2	552413,8
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Le M doit être le plus élevé possible. La significativité du test de F doit tendre vers 0. S'il est supérieur à 0,05, l'analyse n'est pas valide.

La corrélation globale se mesure quant à elle se retrouve dans le tableau « Eigenvalues » (Valeurs propres).

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,395[a]	100,0	100,0	,532

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

On observe notamment la colonne « Canonical Correlation » (Corrélation Canonique). Plus elle est proche de 1, meilleur est le modèle.

Le Lambda de Wilks s'observe quant à lui dans le tableau « Wilks' Lambda ».

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,717	231,524	4	,000

Plus la valeur du Lambda de Wilks (deuxième colonne) est faible, plus le modèle est bon. On observe également sa significativité : plus elle est tend vers 0, meileur, plus le modèle est bon.

3. Estimation des coefficients de la fonction discriminante.

On observe le pouvoir discriminant des axes grâce au tableau « Canonical Discriminant Function Coefficients ».

Canonical Discriminant Function Coefficients

	Function
	1
Years with current employer	-,120
Years at current address	-,037
Debt to income ratio (x100)	,075
Credit card debt in thousands	,312
(Constant)	,058

Unstandardized coefficients

Ce tableau permet d'obtenir la fonction discriminante. Dans notre exemple, la fonction est égale à :

0,058 – 0,12*(Years with current employer) – 0,037*(Years at current adress) + 0,075*(Debet to income ratio) + 0,312*(Credit card ddebt in thousands)

4. Qualité de la représentation.

on observe la qualité de la représentation : on s'assure que la fonction discriminante classifie bien les individus en sous-groupes. Pour cela, on analyse la matrice de confusion qui regroupe les individus bien classés et les mal classés :

Groupes prévus (ou théoriques)

Groupes réels (ou observés)	Groupe 1	Groupe 2	Total
Groupe 1	**22**	4	26
Groupe 2	4	**18**	22
Total	26	22	48

Ainsi, dans notre exemple, 22 éléments du groupe 1 ont été bien reclassés grâce à la fonction discriminante et 4 l'ont mal été. De même, pour le groupe 2, 4 individus ont été mal reclassés et 18 bien reclassés. Au total, c'est donc 40 individus (22 + 18) qui ont été correctement reclassés soit 83% de réussite (40 / 48 = 83%).

Sous SPSS, la matrice de confusion s'observe dans le tableau « Classification Results ».

Classification Results[b,c]

			Predicted Group Membership		
		Previously defaulted	No	Yes	Total
Original	Count	No	391	126	517
		Yes	42	141	183
		Ungrouped cases	96	54	150
	%	No	75,6	24,4	100,0
		Yes	23,0	77,0	100,0
		Ungrouped cases	64,0	36,0	100,0
Cross-validated[a]	Count	No	391	126	517
		Yes	43	140	183
	%	No	75,6	24,4	100,0
		Yes	23,5	76,5	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. 76,0% of original grouped cases correctly classified.

c. 75,9% of cross-validated grouped cases correctly classified.

La note (b.) nous indique le pouvoir de reclassement de la fonction discriminante, ici 76,0%. On peut retrouver ce chiffre en additionnant les observations bien reclassées (ici 398 et 138 soit un total de 536) et en les divisant par le nombre total d'observations classées (dans le cas présent 700 soit 517 + 183)

Il existe une dernière étape qui consiste à observer les mal-classés et savoir si c'est dû à un atypisme ou à une défaillance de la fonction discriminante. S'il s'agit d'un atypisme, il convient de les enlever et de recommencer l'étude.