Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye

Année académique 2023-2024

***

DEVOIR D'ANALYSE DES DONNEES I

Durée : 02h30mn

Questions de cours (4 points)

1. Qu'est-ce que l'analyse factorielle et quels sont les principaux avantages qu'elle offre par rapport à une série d'analyses descriptives univariées sur un ensemble de variables ?

2. Quelle sont les différentes étapes qui entrent en jeu lors de la réalisation d'une analyse factorielle ?

3. Quelles sont les principales limites des méthodes d'analyse factorielle et comment les dépasser ?

Exercice 1 (4 points)

Vous disposez des données fictives suivantes composées de 3 variables quantitatives et 5 individus :

Individu	Variable 1	Variable 2	Variable 3
1	1	2	3
2	2	3	4
3	3	4	5
4	4	5	6
5	5	6	7

La matrice d'inertie, associée au nuage des individus, des données centrées et réduites, est la suivante :

$$\begin{pmatrix} 1 & 0,75 & 0,5 \\ 0,75 & 1 & 0,75 \\ 0,5 & 0,75 & 1 \end{pmatrix}$$

Les valeurs propres de cette matrice sont :

$$\lambda_1 = 2, \quad \lambda_2 = 1 \quad et \ \lambda_3 = 0$$

Les vecteurs propres associés sont :

$$V_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \quad V_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \quad et \ V_3 = \begin{pmatrix} -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}$$

Consignes

1. Calculer les pourcentages d'inertie expliquer par les deux premiers axes factoriels.
2. Calculer les coordonnées (facteurs) des individus sur le premier axe factoriel.
3. Calculez la qualité de représentation (cos^2) des individus sur le deuxième axe factoriel.
4. Calculez la contribution des individus sur le deuxième axe factoriel.

Exercice 2 (12 points)

L'ensemble de données de prédiction du diabète est une base de données médicales de patients comprenant leurs caractéristiques telles que l'âge, l'indice de masse corporelle (IMC), l'hypertension, les maladies cardiaques, le taux d'HbA1c[1] et la glycémie.

Dans le cadre de ce travail pratique, cet ensemble de données est utilisé pour explorer les divers facteurs médicaux déterminants dans la probabilité de développer un diabète. Sur cette base, les professionnels de la santé peuvent ainsi identifier les patients susceptibles de développer un diabète et élaborer des plans de traitement personnalisés.

La base de données contient 100 000 individus décrits par un ensemble de sept (07) variables. Les détails portant sur ces variables sont consignés dans le tableau 1 ci-dessous.

[1] Hémoglobine liée au glucose (appelée HbA1c ou hémoglobine glyquée)

Tableau 1 : Description des variables

Code	Libellé
age	L'âge du patient
hypertension	0 indique que le patient est atteint d'hypertension et 1 signifie qu'il ne l'est pas.
heart_disease	0 indique que le patient est atteint d'une maladie cardiaque et 1 signifie qu'il ne l'est pas.
bmi	L'IMC (Body Mass Index) est une mesure de la graisse corporelle basée sur le poids et la taille.
HbA1c_level	Le taux d'HbA1c (hémoglobine A1c) est une mesure de la glycémie moyenne d'une personne au cours des 2 ou 3 derniers mois.
blood_glucose_level	La glycémie indique la quantité de glucose dans le sang à un moment donné.
diagnostic	Le diabète est la variable cible à prédire, la valeur 1 indiquant la présence de diabète et celle 0 indiquant l'absence de diabète.

Dans un premier temps, une étude statistique préalable sur les différentes variables a donné les résultats consignés dans les tableaux suivants.

Tableau 2 : Moyennes des variables pour l'ensemble de la population considérée

diagnostic	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level
Négatif	40,12	0,06	0,03	26,89	5,40	132,85
Positif	60,95	0,25	0,15	31,99	6,93	194,10
Total	**41,89**	**0,07**	**0,04**	**27,32**	**5,53**	**138,06**

Tableau 3 : Médianes des variables pour l'ensemble de la population considérée

diagnostic	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level
Négatif	40	0	0	27,32	5,80	140,00
Positif	62	0	0	29,97	6,60	160,00
Total	**43**	**0**	**0**	**27,32**	**5,80**	**140,00**

Tableau 4 : Ecart-type des variables pour l'ensemble de la population considérée

diagnostic	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level
Négatif	22,31	0,24	0,17	6,37	0,97	34,25
Positif	14,55	0,43	0,36	7,56	1,08	58,64
Total	**22,52**	**0,26**	**0,19**	**6,64**	**1,07**	**40,71**

Dans un second temps, une analyse factorielle discriminante est réalisée sur un échantillon d'apprentissage représentant 70% de la population d'étude. Les principaux résultats obtenus sont les suivants.

Consignes

1. Quelle est le pourcentage d'inertie discriminante expliqué par le premier facteur ou axe factoriel ?
2. Analyser les différents résultats obtenus.
3. Soit les données de trois (03) patients participants à l'étude consignées dans le tableau suivant.

Variables	Patient 1	Patient 2	Patient 3
age	32,00	40	51
hypertension	0	1	1
heart_disease	0	0	1
bmi	27,32	53,44	48,92
HbA1c_level	5,0	5,7	9,0
blood_glucose_level	100	126	280

a) Déterminer les scores de ces patients

b) Proposez un diagnostic pour chacun d'entre eux en justifiant votre choix.

4. Sur la base de l'échantillon test, la matrice de confusion suivante a été obtenue.

		Prédiction	
		Négatif	Positif
Réalité	Négatif	20 081	192
	Positif	792	1 167

Déterminer le pouvoir d'affectation de cette AFD.

Tableau 5 : Probabilités à priori des différents groupes

Négatif	Positif
0,915	0,085

Tableau 6 : Coefficient linéaire de la discrimination

	LD1
age	0,01053372
hypertension	0,72597760
heart_disease	0,90352018
bmi	0,03209271
HbA1c_level	0,61260134
blood_glucose_level	0,01713233

Figure 1 : Histogramme de la distribution des scores suivant le premier facteur discriminant

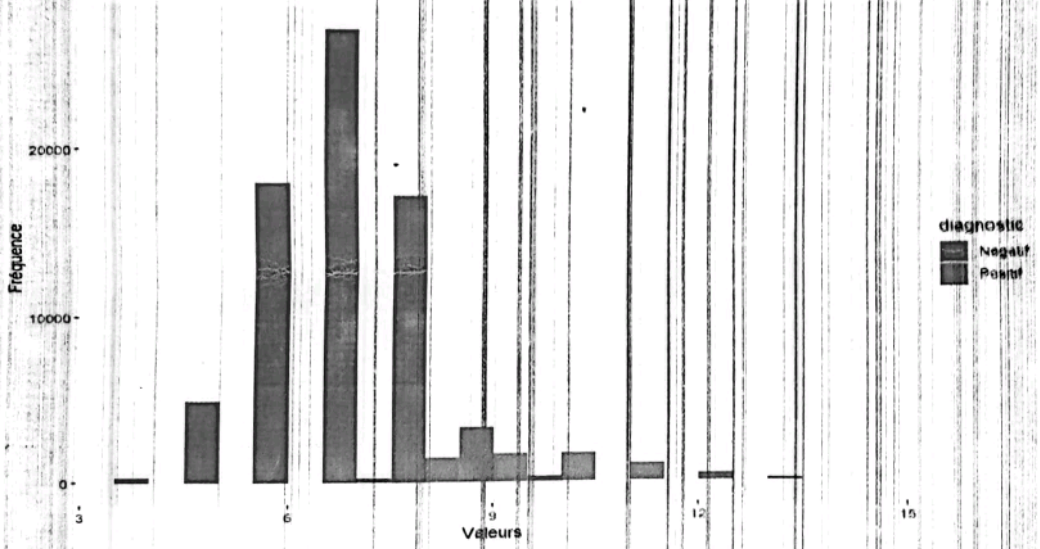

Figure 2 : Courbe de densité des scores suivant le premier facteur discriminant