Analyse Exploratoire Multidimensionnelle
TP5 : Analyse Discriminante

Consignes : on utilise le logiciel R

Tableau 1: Qualité des vins de Bordeaux

Année	temp	inso	chal	pluie	Qualité
1924	3064	1201	10	361	2
1925	3000	1053	11	338	3
1926	3155	1133	19	393	2
1927	3085	970	4	467	3
1928	3245	1258	36	294	1
1929	3267	1386	35	225	1
1930	3080	966	13	417	3
1931	2974	1189	12	488	3
1932	3038	1103	14	677	3
1933	3318	1310	29	427	2
1934	3317	1362	25	326	1
1935	3182	1171	28	326	3
1936	2998	1102	9	349	3
1937	3221	1424	21	382	1
1938	3019	1230	16	275	2
1939	3022	1285	9	303	2
1940	3094	1329	11	339	2
1941	3009	1210	15	536	3
1942	3227	1331	21	414	2
1943	3308	1366	24	282	1
1944	3212	1289	17	302	2
1945	3316	1444	25	253	1

1946	3061	1175	12	261	2
1947	3478	1317	42	259	1
1948	3126	1248	11	315	2
1949	3458	1508	43	286	1
1950	3252	1361	26	346	2
1951	3052	1186	14	443	3
1952	3270	1399	24	306	1
1953	3198	1259	20	367	1
1954	2904	1164	6	311	3
1955	3247	1277	19	375	1
1956	3083	1195	5	441	3
1957	3043	1208	14	371	3

On considère le tableau 1 ci-dessus où l'on relie la qualité des vins de Bordeaux entre 1924 et 1957 à des caractéristiques météorologiques. La variable qualitative à expliquer est la qualité des vins de Bordeaux et prend trois modalités : modalités: 1 = bon, 2 = moyen, 3 = médiocre. Les variables explicatives de la qualité du vin de Bordeaux sont

Temp =: somme des températures moyennes journalières (°c) ; Inso = : durée d'insolation (heures),

Chal = : nombre de jours de grande chaleur ; Pluie = : huateur des pluies (mm).

On se propose de mettre en œuvre l'analyse discriminante à l'aide de ce tableau de données. Les résultats issus du logiciel SPSS sont donnés dans l' annexe: **Analyse discriminante**

1. Tracer et analyser les boîtes à moustaches de chacune des variables explicatives par rapport aux modalités de la variable de groupe « qualité des vins »

2. Estimer le pouvoir discriminant de chacune des variables explicatives par rapport à la variable Qualité des vins : Le rapport de la variance inter-classes (variance expliquée) sur variance totale (On peut passer si besoin)

3. Tester l'égalité des moyennes de chacune des variables explicatives à travers les groupes de la variable dépendante. (Test de Lambda de Wilks par exemple ou autre) :
On posera convenablement les hypothèses nulle et alternative

4. Tester l'égalité des matrices de variance-covariance de chacun des groupes de la variable dépendante. (Test de BOX de l'égalité des matrices de covariances par exemple ou autre)
On posera convenablement les hypothèses nulle et alternative

5. Analyse discriminante linéaire (ou Analyse Factorielle Discriminante)
   Combien de facteurs discriminants peut-on construire ?
   Estimer le pouvoir discriminant de chacun des facteurs issus de l'AFD
   Tester le pouvoir discriminant de chacun des facteurs issus de l'AFD
   Analyser les corrélations entre les facteurs discriminantes et la variable explicative.
   Estimer le taux de bon reclassement (matrice de confusion)
   A quelle classe de vin appartiendrait l'année 1958 dont les informations sont les suivantes :
      temp=3000 ; Inso=1100 ; Chal=20 ; Pluie=300 ?

6. Effectuer l'analyse discriminante probabiliste (ou quadratique) sur les vins de Bordeaux.

Exercice - 2 - : *Analyse discriminante et parties de Hockey*

Un parieur veut essayer de déterminer un modèle qui lui permettrait de gagner plus souvent ses paris lorsqu'il mise sur les résultats des parties de hockey de la ligue nord américaine de hockey. Il décide d'utiliser l'analyse discriminante afin de prévoir si une équipe va gagner, perdre ou faire match nul. Les variables retenues sont :

RV : rang de l'équipe visiteuse au classement général

RL : rang de l'équipe locale au classement général

NPGV : nombre de parties gagnées par l'équipe visiteuse (parmi les 10 dernières parties)

NPGL : nombre de parties gagnées par l'équipe locale (parmi les 10 dernières parties)

NPGD : nombre de parties gagnées à domicile (par l'équipe locale)

NPGE : nombre de parties gagnées à l'extérieur (par l'équipe visiteuse)

PRED : prédiction (1 si l'équipe locale gagne, 2 si l'équipe locale perd, 3 si l'équipe locale fait match nul).

Les résultats des dernières parties sont ci-contre:

PRED	RV	RL	NPGV	NPGL	NPGD	NPGE
2	10	26	5	1	1	8
2	11	22	6	3	6	5
3	13	14	5	5	6	4
2	5	4	5	5	12	8
1	16	17	4	4	5	8
2	1	15	7	2	6	10
1	26	12	1	5	5	1
2	7	23	5	1	8	8
2	10	6	5	6	10	8
1	19	3	3	8	14	5
1	11	2	4	6	12	4
1	15	13	5	4	7	7
2	8	21	4	3	5	7
1	23	6	3	5	10	4
1	3	7	7	4	9	7
1	26	16	2	4	9	1
1	12	1	3	7	13	7
1	24	14	2	5	9	2
3	5	10	8	7	8	7
1	20	22	4	4	3	6
3	6	23	5	3	5	4

Deux analyses discriminantes ont été faites en utilisant comme variable dépendante la variable PRED.

Modèle 1 : Utilisation de toutes les variables explicatives : RV RL NPGV NPGL NPGD NPGE
Modèle 2 : Utilisation de quatre variables explicatives : RV RL NPGL NPGE

Question 1 : En utilisant chacun des deux modèles :

a) Déterminer le pouvoir discriminant de chacune des variables explicatives utilisées ;

b) Quelles variables explicatives le parieur devrait-il retenir par modèle et pourquoi ?

c) Déterminer le pouvoir discriminant de chaque facteur issu de chacun des deux modèles.

Question 2 : Parmi les deux modèles proposés, lequel semble le plus performant ?

Question 3 : Le parieur décide d'utiliser l'un des deux modèles pour prédire les 7 parties suivantes :

Visiteur	Local	Valeurs des variables explicatives
San Jose	Anaheim	RV=21, RL=25, NPGV=3, NPGL=2, NPGD=3, NPGE=6
Montréal	New-Jersey	RV=18, RL=10, NPGV=4, NPGL=5, NPGD=8, NPGE=2
Vancouver	Edmonton	RV=17, RL=19, NPGV=4, NPGL=5, NPGD=9, NPGE=5
NY Rangers	Philadelphie	RV=15, RL=6, NPGV=2, NPGL=6, NPGD=10, NPGE=7
Ottawa	Québec	RV=26, RL=1, NPGV=1, NPGL=7, NPGD=13, NPGE=1
St-Louis	Détroit	RV=5, RL=2, NPGV=5, NPGL=8, NPGD=14, NPGE=8
Chicago	Calgary	RV=4, RL=9, NPGV=5, NPGL=4, NPGD=10, NPGE=12

a) En utilisant le modèle 1, quelles équipes devraient gagner, perdre ou faire match nul ? Justifiez et indiquez, pour chaque partie, la probabilité que le choix soit correct.

b) En utilisant le modèle 2, quelles équipes devraient gagner, perdre ou faire match nul ? Justifiez et indiquez, pour chaque partie, la probabilité que le choix soit correct.

c) Les résultats réels de ces parties sont : New-Jersey gagnant, Vancouver gagnant, Québec gagnant, Philadelphie gagnant, Anaheim gagnant, Calgary gagnant, et match nul entre Détroit et St-Louis. Sur les 7 parties, combien y-a-t-il eu de bonnes prévisions selon le modèle utilisé ? Quelle est maintenant le modèle le plus performant compte tenu de ces résultats ?