M1 SES Enquêtes statistiques et économétrie appliquée

Examen de la 2ème session, juin 2015 (durée : 2 heures)

Exercice 1

Rappel : dans le modèle logistique, nous cherchons à expliquer une variable Y, qui vaut 0 ou 1, à partir d'une variable explicative X (ou d'un vecteur de variables explicatives également noté X). Nous modélisons pour cela $\pi(x) = \mathbb{P}(Y = 1|X = x)$ par

$$\pi(x) = \frac{\exp(\beta_0 + \langle \beta, x \rangle)}{1 + \exp(\beta_0 + \langle \beta, x \rangle)}$$

où $\langle \beta, x \rangle = \beta_1 x$ si X est une variable simple et $\langle \beta, x \rangle = \beta_1 x_1 + \beta_2 x_2$ si X est un vecteur à deux composantes. Ce modèle est équivalent à $\log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \langle \beta, x \rangle$ avec β_0 et β inconnus. Enfin, dans le cas où X prend également les valeur 0 ou 1, on montre que $\hat{\beta}_0 = \log(n_{21}/n_{11})$ et $\hat{\beta}_1 = \log((n_{11}n_{22})/(n_{12}n_{21}))$ où n_{ij} représente l'effectif observé pour $i = 1, 2$ et $j = 1, 2$ du tableau de contingence correspondant :

	X=0	X=1
Y=0	n_{11}	n_{12}
Y=1	n_{21}	n_{22}

Nous traitons un problème de défaut bancaire. La variable `default` est la variable à expliquer. Nous disposons ici d'un échantillon de taille 10000 et deux variables explicatives `student` et `balance`.

- `default` : `Yes` (ou 1) si le client fait défaut sur sa dette et `No` (ou 0) sinon.
- `student` : `Yes` (ou 1) si le client est un étudiant et `No` (ou 0) sinon
- `balance` : montant moyen mensuel d'utilisation de la carte de crédit

1. On considère pour commencer un modèle de régression logistique simple où on cherche à expliquer `default` en fonction de `student`.

   (a) À l'aide du tableau de contingence ci dessous, calculer "à la main" les coefficients estimés du modèle logistique.

   ```
          student   No  Yes
   default
   No               6850 2817
   Yes               206  127
   ```

   (b) Donner l'équation du modèle logistique simple ajusté.

   (c) Calculer le rapport de rapport de chances (odds ratio). Que peut-on en conclure ?

2. Nous avons utilisé le logiciel R pour ajuster le même modèle logistique multiple où on cherche à expliquer `default` en fonction de `student` et de `balance`.
   À l'aide des sorties de R ci-dessous, donner l'équation du modèle logistique ajusté avec les coefficients estimés pour les `"student=Yes"` et pour les `"student=No"`.

   ```
   glm(formula = default ~ student + balance,
   family = binomial(link = "logit"), data = Default)

   Coefficients:
                Estimate Std. Error z value Pr(>|z|)
   (Intercept) -10.749496  3.692e-01 -29.116  < 2e-16 ***
   studentYes  -0.7148776  1.475e-01  -4.846 1.26e-06 ***
   balance      0.0057381  2.318e-04  24.750  < 2e-16 ***
   ```

Exercice 2 : ACP

On considère les notes sur 20 en maths, physique, français et anglais de 15 élèves.

Simple Statistics

	MATHS	PHYSIQ	FRAN	ANGL
Mean	10.30000000	9.786666667	10.73333333	11.26666667
StD	3.86744656	2.880443088	3.16152465	2.91465917

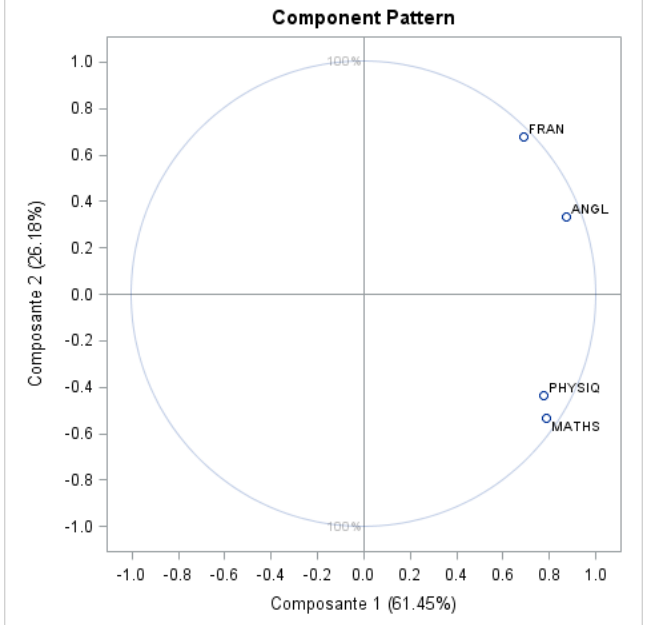
Correlation Matrix

		MATHS	PHYSIQ	FRAN	ANGL
MATHS	MATHS	1.0000	0.7092	0.1311	0.5785
PHYSIQ	PHYSIQ	0.7092	1.0000	0.3306	0.4012
FRAN	FRAN	0.1311	0.3306	1.0000	0.7350
ANGL	ANGL	0.5785	0.4012	0.7350	1.0000

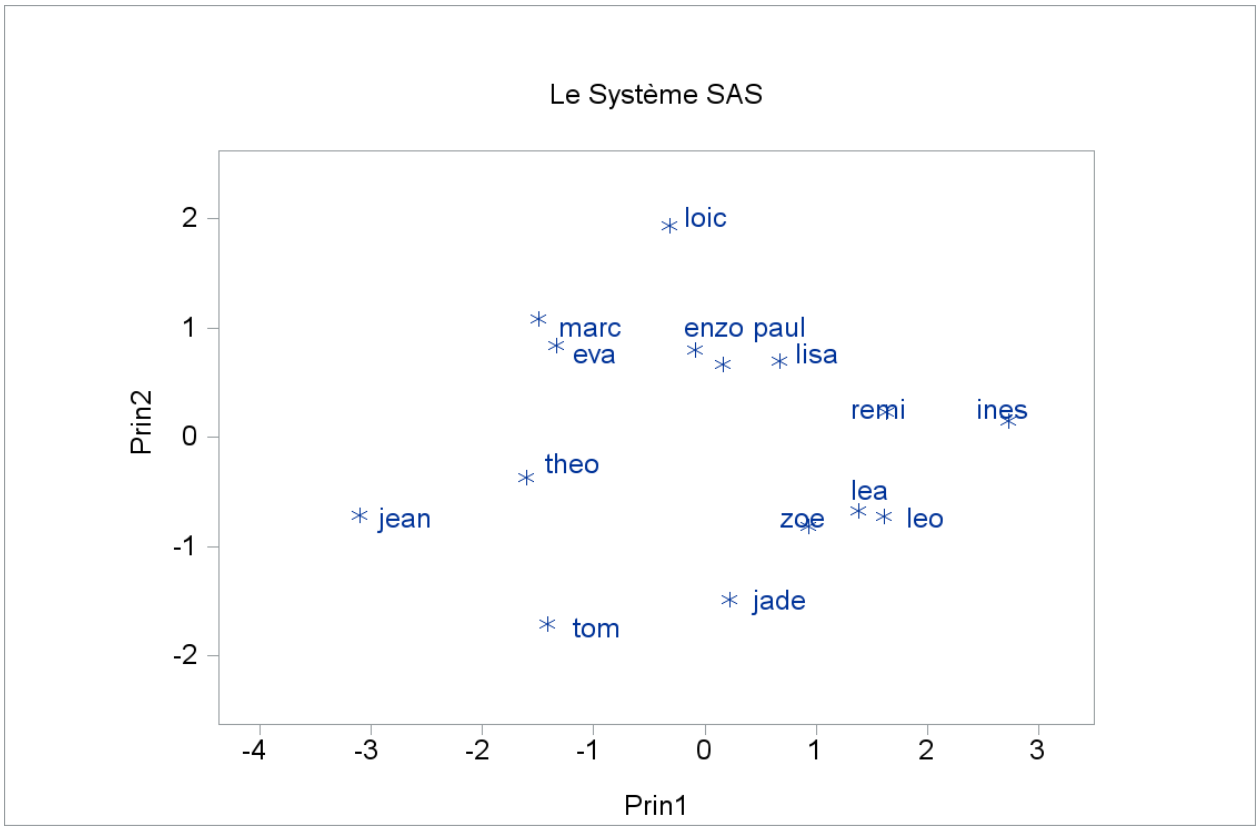
Eigenvalues of the Correlation Matrix

	Valeur propre	Différence	Proportion	Cumulé
1	2.45814791	1.41085051	0.6145	0.6145
2	1.04729740	0.61425114	0.2618	0.8764
3	0.43304625	0.37153781	0.1083	0.9846
4	0.06150844		0.0154	1.0000

Corrélation composantes / variables



	MATHS	PHYSIQ	FRAN	ANGL
Prin1	0.78662	0.77843	0.68705	0.87257
Prin2	-0.53465	-0.43396	0.67924	0.33431
Prin3	-0.27766	0.44570	0.22646	-0.32561
Prin4	0.13523	-0.08409	0.12376	-0.14434



Questions ACP

1. Donner les individus et les variables.
2. Y'a t-il un effet taille ? (Justifier)
3. Tableau des « valeurs propres de la matrice de corrélation » Combien de composantes principales allez-vous conserver et selon quels critères ?
4. Interprétez le premier axe et le second axe.
5. Représentation dans le plan principal. Caractérisez la position de Loic et Jean.

Exercice 3 : Chi2

La distribution suivante a été dressée par Haberman (1978) à partir de données fournies par le National Opinion Research Center de l'Université de Chicago. Les variables sont le nombre d'années de scolarité X et l'attitude face à l'avortement Y.

X/Y	Pour	Indifférent	Contre
Moins de 8 ans	31	23	56
Entre 9 et 12 ans	171	89	177
Plus de 12 ans	116	39	74

Questions

1. Quelles sont les variables étudiées, quelle est leur nature ?
2. Donner la distribution, en pourcentages, de l'opinion des personnes qui ont été scolarisées moins de 8 ans. .
3. Donner la distribution de la scolarisation, en pourcentages, pour les personnes contre l'avortement.
4. Donner la distribution marginale de l'opinion sur l'avortement, en effectifs et en pourcentages.
5. .On effectue un test d'indépendance du chi 2 entre les deux variables.
   a) Donner les hypothèses du test d'indépendance
   b) Donner les conditions d'application du test. Sont-elles vérifiées ?
   c) Donner la statistique du Chi2 et sa loi sous l'hypothèse nulle.
   d) On a calculé la valeur observée du chi2 qui est de 17,71. La p-valeur associée est 0,0014. Que pouvez-vous conclure ?

Le tableau suivant donne les résultats d'une procédure Freq effectuée sur le logiciel SAS.

Fréquence Attendu Pourcentage Pctage en ligne Pctage en col.	Table de scola par opinion			
scola	**opinion**			
	pour	**indiff**	**contre**	**Total**
inf8	31	23	56	110
	45.077	21.405	43.518	
	3.99	2.96	7.22	14.18
	28.18	20.91	50.91	
	9.75	15.23	18.24	
9a12	171	89	177	437
	179.08	85.035	172.89	
	22.04	11.47	22.81	56.31
	39.13	20.37	40.50	
	53.77	58.94	57.65	
plus12	116	39	74	229
	93.843	44.561	90.597	
	14.95	5.03	9.54	29.51
	50.66	17.03	32.31	
	36.48	25.83	24.10	
Total	318	151	307	776
	40.98	19.46	39.56	100.00