3.2 Régression logistique

Bibliographie :
Howell, D.C., Méthodes Statistiques en Sciences Humaines, De Boeck, Paris Bruxelles, 1998.
Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.

3.2.1 La régression logistique

La régression logistique peut être vue comme une extension de la régression linéaire au cas où la variable dépendante est dichotomique. Plus précisément, sur un échantillon de n individus statistiques, on a observé :

- p variables numériques ou dichotomiques X_1, X_2, ..., X_p (variables indépendantes ou explicatives)
- une variable dichotomique Y (variable dépendante, ou "à expliquer").

Dans le cas le plus simple, on cherche à expliquer une variable dichotomique Y par une variable numérique X. On dispose donc d'un tableau de données sous la forme :

	s1	s2		sn
Y	1	0	...	0
X	x1	x2		xn

Exemple : On considère un échantillon de 30 sujets pour lesquels on a relevé :
- d'une part le niveau des revenus (variable numérique)
- d'autre part la possession ou non d'un nouvel équipement électro-ménager.

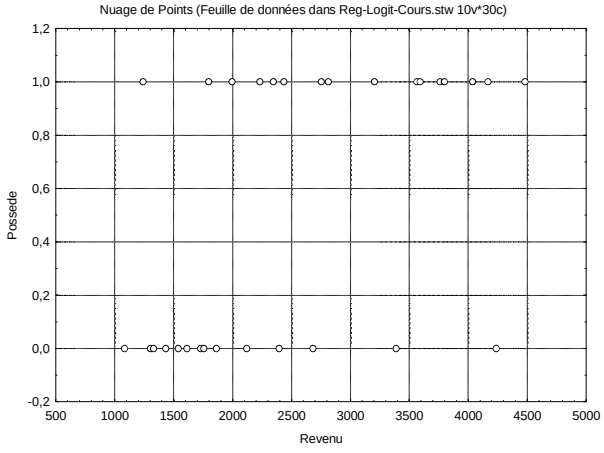
On a obtenu les données suivantes :

Revenu	1085	1304	1331	1434	1541	1612	1729	1759	1863	2121	2395	2681	3390	4237	1241
Possède	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Revenu	1798	1997	2234	2346	2436	2753	2813	3204	3564	3592	3762	3799	4037	4168	4484
Possède	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

3.2.1.1 Principe de la méthode

Ces données peuvent être représentées à l'aide d'un nuage de points, qui a l'allure suivante :



On cherche un modèle permettant d'estimer Y ("Possède") connaissant X ("Revenu"). Plutôt que de rechercher un modèle mathématique donnant pour une valeur donnée X exactement la valeur 0 ou la valeur 1, il peut sembler pertinent de rechercher un modèle produisant des valeurs comprises entre 0 et 1 qui seront interprétées comme des probabilités. Par exemple :

$\hat{Y} = 0{,}1$ signifie que : il y a 10% de chances que Y=1

Cependant, la droite de régression de la variable Y par rapport à la variable X ne constitue pas un bon modèle car les valeurs estimées ne seront pas limitées à 0 et 1.

Pour passer d'une variable prenant ses valeurs dans [0, 1] à une variable prenant ses valeurs dans [0, +∞[, on introduit le rapport de chances ou cote :

$$p_1 = \frac{P(Y=1)}{1-P(Y=1)}$$

Ainsi, si P(Y=1)=0,9, le rapport de chances vaut p1 = 0,9/0,1=9 : on a 9 fois plus de chances d'observer Y=1 que Y=0.
De même, si P(Y=1)=0,2, le rapport de chances vaut p1 = 0,2/0,8=1/4 : on a 4 fois plus de chances d'observer Y=0 que Y=1.

Pour passer d'une quantité (le rapport de chances) variant dans [0, +∞[à une quantité prenant n'importe quelle valeur réelle, on applique une nouvelle transformation, en prenant le logarithme népérien du rapport. On obtient ainsi la transformation logit :

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

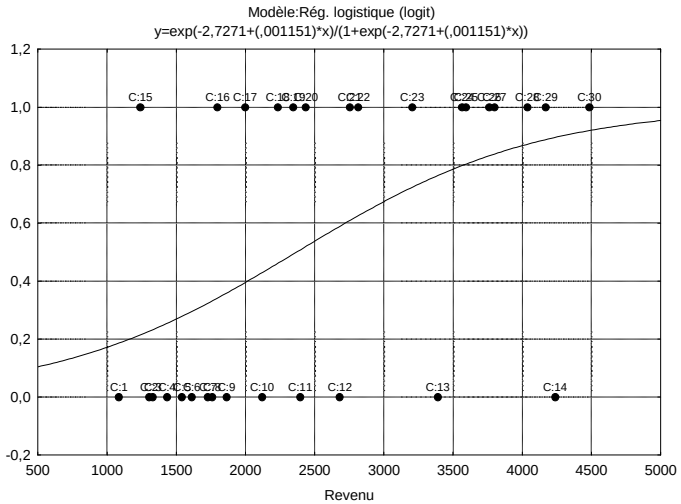
Ainsi,

- si P = 0,9, logit(P) = ln 9 = 2,1972
- si P = 0,5, logit(P) = ln 1 = 0
- si P = 0,2, logit(P) = ln(1/4) = -1,3863.

A partir d'une "valeur logit" y, on peut facilement revenir à la probabilité P correspondante en appliquant la transformation :

$$P = \frac{e^y}{1+e^y}$$

On ajuste alors logit(P) par une fonction affine, ce qui revient à déterminer une "sigmoïde" qui passe au mieux par les points expérimentaux :



L'équation correspondant à cet ajustement est :

$$\text{logit}(Y) = -2{,}7271 + 0{,}001151\, X$$

Exemple d'utilisation de cette équation : à partir de quel revenu a-t-on 90% de chances de tirer un sujet possédant l'équipement envisagé ?

P = 0,9 correspond à P/(1-P) = 0,9/0,1 = 9 d'où logit(P) = 2,1972.
Or : 2,1972 = -2,7271 + 0,001151 X donne X = (2,1972 + 2,7271)/0,001151, c'est-à-dire : X=4278.

Remarque : Cette équation n'est pas obtenue par une "simple" régression linéaire, mais par des méthodes itératives. D'une part, il n'est pas envisageable de faire les calculs manuellement, d'autre part, il faudra, dans certains cas, "aider" les logiciels en indiquant des valeurs initiales plausibles pour les coefficients.

3.2.1.2 Aides à l'interprétation. Evaluation de la qualité du modèle obtenu.

La qualité du modèle peut être évaluée en comparant les résultats obtenus avec ceux du modèle "constant" qui attribuerait la probabilité 14/30 à la valeur 0 et 16/30 à la valeur 1. Une fonction de vraisemblance est évaluée dans les deux cas, et la différence des deux fonctions suit une loi du khi-2 à 1 degré de liberté lorsqu'il n'y a qu'une seule variable indépendante.

Sur notre exemple, on obtient :
Chi-deux = 7,636181 ; dl = 1 ; p = ,0057242
Le revenu est donc un prédicteur significatif de la variable Y.

Une autre aide à l'interprétation courante est le rapport de cotes ou odds-ratio (OR). En particulier, la contribution de la variable X à la variation de Y est calculée par :

$$\text{OR} = \exp(\text{Coefficient de X dans le modèle})$$

Ainsi, sur notre exemple, l'odds-ratio correspondant au coefficient 0,001151 est : $e^{0{,}001151}$=1,0012. Autrement dit, une augmentation du revenu de 1 unité se traduit par une multiplication de la probabilité par 1,0012.

D'une manière générale, l'odds-ratio est défini comme le rapport de deux rapports de chances. Ainsi, l'odds-ratio relatif à l'étendue des valeurs observées est défini de la manière suivante :

- On calcule le rapport de chances relatif à la plus grande valeur observée du revenu :

Pour X = 4484, P_1=0,919325 et $\frac{P_1}{1-P_1} = 11{,}3954$

- On calcule le rapport de chances relatif à la plus petite valeur observée du revenu :

Pour X = 1085, P_2=0,185658 et $\frac{P_2}{1-P_2} = 0{,}2280$

- L'odds-ratio est obtenu comme quotient des deux rapports précédents :

$$\text{OR} = \frac{\frac{P_1}{1-P_1}}{\frac{P_2}{1-P_2}} = \frac{11{,}3954}{0{,}2280} = 49{,}98$$

On évalue également un Odds-ratio comparant valeurs observées et valeurs prévues. Pour cela, on définit deux classes dans les valeurs prévues : celles inférieures à 0,5 et celles supérieures à 0,5 et on forme le tableau de contingence croisant les valeurs observées (0 ou 1) avec les classes ainsi définies. Sur notre exemple, on obtient :

Obs	Prév. < 0,5	Prév. > 0,5
0	10	4
1	5	11

Le rapport est alors obtenu en formant le rapport ad/bc (produit des effectifs des cases d'accord divisé par le produit des effectifs des cases de désaccord.
On obtient ainsi :

$$\mathrm{OR} = \frac{10 \times 11}{5 \times 4} = 5{,}50$$

Signification approximative : si la valeur prévue est supérieure à 0,5, on a 5,5 fois plus de chances d'observer Y=1 que Y=0.

3.2.2 La régression logistique avec Statistica

Source : Howell. p. 633, ex. 15.31 a 15.33

La feuille de données Harass contient des données légèrement modifiées relatives à 343 cas créés pour répliquer les résultats d'une étude sur le harcèlement sexuel (Brooke et Perot 1991). Les variables sont :
- l'âge
- l'état-civil (1 = marié(e), 2 = célibataire) (NB étonnant, n'est-ce pas l'inverse? cf données)
- l'idéologie féministe
- la fréquence du comportement
- le caractère agressif du comportement
- le fait qu'il ait été ou non signale (0 = non, 1 = oui).

1) Utiliser un programme de régression logistique et examiner la probabilité qu'un sujet signale un cas de harcèlement sexuel sur la base des VI.

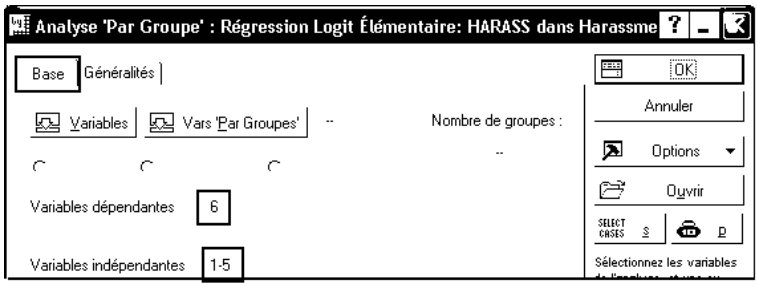
2) Même question, mais en n'utilisant que le prédicteur dichotomique relatif à l'état civil. Faire une table de contingence, calculer les rapports de chances et comparer ces résultats à ceux de la régression logistique. (résultats son significatifs, mais cela importe peu, selon Howell).

3) Apparemment, la fréquence du comportement n'est pas liée à la probabilité de voir la victime signaler le cas de harcèlement. Peut-on en imaginer les raisons ?

Ouvrez le classeur Harassment.stw.

Un premier résultat peut être obtenu à l'aide du menu : Statistiques, Analyses par groupes, Modèles linéaires/non-linéaires avancés, Estimation non linéaire, Régression Logit élémentaire.

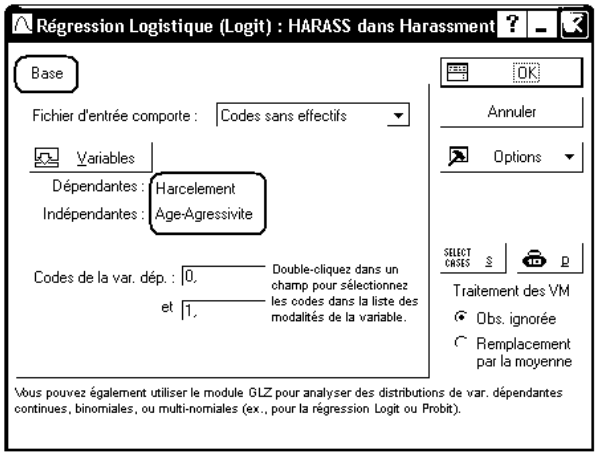
Indiquez "Harcelement" comme variable dépendante et les 5 autres variables comme variables idépendantes.



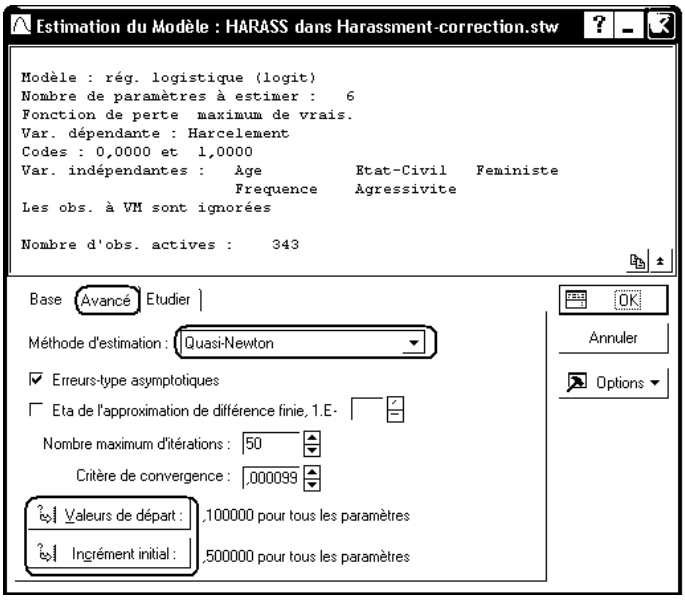
Lorsqu'on laisse les options par défaut (minimum de résultats), on obtient la feuille de résultats suivante :

N=343	Modèle: Rég. logistique (logit) Nbre de 0 : 174 1 : 169 Var dép. : Harcelement Perte : Max vraisemblance (MC-er. posit. à Perte finale= 219,99193498 Chi²(5)=35,442 p=,00000					
	Const.B0	Age	Etat-Civil	Feministe	Frequence	Agressivite
Estimat.	-1,7317	-0,0137	-0,0723	0,0070	-0,0464	0,4878
Erreur-type	1,4296	0,0129	0,2338	0,0146	0,1525	0,0949
t(337)	-1,2113	-1,0614	-0,3091	0,4771	-0,3043	5,1409
niveau p	0,2266	0,2893	0,7575	0,6336	0,7611	0,0000
-95%CL	-4,5439	-0,0391	-0,5321	-0,0218	-0,3464	0,3011
+95%CL	1,0804	0,0117	0,3876	0,0358	0,2536	0,6744
Chi² de Wald	1,4672	1,1265	0,0955	0,2277	0,0926	26,4292
niveau p	0,2258	0,2885	0,7573	0,6333	0,7609	0,0000
Odds ratio (unité)	0,1770	0,9864	0,9303	1,0070	0,9547	1,6287
-95%CL	0,0106	0,9617	0,5874	0,9784	0,7072	1,3514
+95%CL	2,9460	1,0118	1,4734	1,0364	1,2887	1,9629
Odds r. (étendue)		0,4644	0,9303	1,3985	0,8306	80,6394
-95%CL		0,1121	0,5874	0,3509	0,2501	15,0336
+95%CL		1,9244	1,4734	5,5731	2,7579	432,5462

On peut aussi utiliser le menu Statistiques, Modèles linéaires/non-linéaires avancés, Estimation non linéaire, Régression Logit: On indique de la même façon la variable dépendante et les variables indépendantes :



On peut ensuite choisir un algorithme d'estimation et éventuellement indiquer manuellement les valeurs initiales des coefficients bi, ce qui est souvent utile, si les plages de variations des VI sont très différentes de l'intervalle [0, 1] (et n'est pas prévu par le menu précédent).



Le tableau de résultats produit par la méthode précédente est alors accessible par le bouton "Synthèse : paramètres et erreurs-types" du dialogue des résultats.

L'équation de la courbe de régression est :

$$\text{logit}\, P = -1{,}7317 - 0{,}013698\,\text{Age} - 0{,}072251\,\text{EtatCivil} + 0{,}0069870\,\text{Feministe} - 0{,}046408\,\text{Frequence} + 0{,}4878\,\text{Agressivite}$$

Le khi-2 correspondant au modèle vaut 35,442, et il est significatif au seuil de 1%. En revanche, seule la variable Agressivite semble avoir un rôle explicatif supérieur à celui que le hasard est susceptible de produire.

Les odds-ratio unitaires correspondant aux différentes variables sont :

	Modèle: Rég. logistique (logit) Nbre de 0 : 174 1 : 169 (HARASS) Var dép. : Harcelement Perte : Max vraisemblance (MC-er. posit. à Perte finale= 219,99193498 Chi²(5)=35,442 p=,00000					
N=343	Const.B0	Age	Etat-Civil	Feministe	Frequence	Agressivite
Odds ratio (unité)	0,1770	0,9864	0,9303	1,0070	0,9547	1,6287

On voit que seules les variables Feministe et Agressivite possèdent des odds-ratio unitaires supérieurs à 1 et que seul celui de Agressivite est nettement différent de l'unité.

On peut également afficher le tableau des valeurs observées et des valeurs prévues de la variable dépendante :

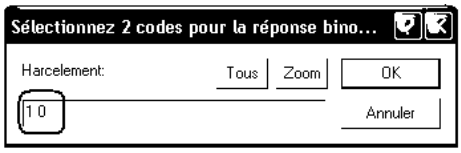
	Modèle : (HARASS dans Harassment-correction.stw Var. Dép. : Harcelement		
	Observée	Prév.	Résidus
1	0,0000	0,6431	-0,6431
2	0,0000	0,7312	-0,7312
3	1,0000	0,8696	0,1304
4	1,0000	0,3080	0,6920

Sous l'onglet Résidus, on peut obtenir le calcul de l'odds-ratio pour le modèle :

	Classification d'obs. (HARASS) Odds ratio : 2,1051 Pourc. corrigé : 59,18%		
Observée	Prév. 0,000000	Prév. 1,000000	% Corrigé
0,000000	111	63	63,79310
1,000000	77	92	54,43787

On peut également utiliser le menu : Statistiques, Modèles linéaires/non-linéaires avancés, Modèles linéaires/non linéaires généralisés, puis l'item Modèle logit dans l'onglet Base ou les items : Régression simple (ou multiple), Distribution: Binomiale et Fonction de liaison : logit de l'onglet Avancé.

Lorsqu'on indique les variables et leur rôle, il est important de préciser que c'est le code "1" de la variable Harcelement qui doit être assimilé à la modalité "succès" de la variable binomiale, faute de quoi les résultats seraient inversés :



On retrouve ainsi les résultats obtenus par les deux autres méthodes, mais avec une présentation différente. On peut également obtenir des résultats supplémentaires, tels que l'évolution des valeurs des coefficients à chaque itération de l'algorithme :

	Harcelement - Historique itérations (HARASS) Distribution : BINOMIALE Fonction de Liaison : LOGIT					
Effet	Niveau Effet	Colonne	Itérat. 0	Itérat. 1	Itérat. 2	Itérat. 3
Ord.Orig		1	0,000	-1,556	-1,727	-1,732
Age		2	0,000	-0,012	-0,014	-0,014
Etat-Civil		3	0,000	-0,066	-0,072	-0,072
Feministe		4	0,000	0,006	0,007	0,007
Frequence		5	0,000	-0,044	-0,046	-0,046
Agressivite		6	0,000	0,438	0,486	0,488
Vraisembl.			-237,749	-220,156	-219,992	-219,992

On peut également noter que l'on obtient des résultats légèrement différents lorsque l'on indique "Etat-Civil" comme variable catégorielle.

3.2.3 Un exemple de régression logistique issu d'un article.

Réf. : Factors Influencing Adolescents Engagement in Risky Internet Behavior, ALBERT KIENFIE LIAU, Ph.D., ANGELINE KHOO, Ph.D., and PENG HWAANG, Ph.D., CYBERPSYCHOLOGY & BEHAVIOR, Volume 8, Number 6, 2005, pp 513-520.

Dans l'article cité supra les auteurs se sont intéressés aux facteurs liés à la prise de risques dans le comportement sur Internet pour des adolescents de Singapour. Ils identifient notamment comme conduite à risques le fait de rencontrer physiquement une personne qu'ils ont d'abord connu "online".

Dans les résultats de leur étude, les auteurs indiquent notamment :

1045 (93.0% of the total sample) adolescents reported having used the Internet, and 827 (73.6%) adolescents reported having chatted on the Internet. The study focused on this group of 827 adolescents who have experienced chatting on the Internet. These adolescents have a mean age = 14.42 (SD = 1.33) and are 51.4% girls. (...)
A total of 169 adolescents (16.2% of Internet users, or 20.4% of those who chat) reported having met someone in real life that they first encountered online.

A series of multiple logistic regression analyses was used to examine the factors that influence adolescents' engagement in risky internet behavior, in particular, meeting in person with someone encountered online. Odds ratios (OR) were calculated to approximate relative risk and are presented with 99% confidence intervals. Age was a significant predictor of the risky behavior (OR = 1.26, 99% CI (1.06, 1.48), $p < 0.0001$) but gender was not a significant predictor; 80 out of the 169 (47.3%) adolescents were girls. For ease of interpretation, the frequency of use of the Internet variable was dichotomized so that 1 = "at least once a day" and 0 = "less than once a day." Controlling for age, frequency of use of the Internet was a significant predictor of the risky behavior (OR = 1.68, 99% CI (1.07, 2.65), $p < 0.01$). Parents' educational background and whether parents lived together were not significant predictors of the risky behavior. All subsequent analyses include age and frequency of use as covariates in order to control for the influence of these factors. The following factors were examined as predictors of the risky behavior: frequency of chatting and gaming behavior, parental supervision, communication with parents, type of personal information given out, amount of inappropriate messages received, whether inappropriate websites have been visited, and type of internet advice heard. Significant and marginally significant predictors of the risky behavior are reported in Table 2.

TABLE 2. SIGNIFICANT AND MARGINALLY SIGNIFICANT PREDICTORS OF THE RISKY INTERNET BEHAVIOR—MEETING IN PERSON SOMEONE ENCOUNTERED ONLINE

Predictor	*OR*	*99% CI*
Frequency of Internet activities	3.13**	1.75, 5.55
Frequency of chatting	1.77*	1.07, 2.91
Frequency of gaming		
Parental supervision		
Rules for Internet use		
Not allowed to meet in person someone encountered online	0.49**	0.30, 0.81
Not allowed to talk to strangers in chatrooms	0.46*	0.23, 0.93
Not allowed to give out personal information	0.62†	0.39, 1.01
People usually at home when arrive from school	1.56†	1.06, 1.48
Communication with parents		
Tell parents about receiving pornographic junk mail	0.49†	0.22, 1.06
Giving out personal information		
Phone number	2.17*	1.15, 4.09
Photograph	2.68*	1.16, 6.18
Favorite band, music	1.67*	1.03, 2.90
Receiving inappropriate message		
Met someone on the Internet who asked for personal information	4.16**	2.42, 6.67
Sent pornography from someone met only on the Internet	1.80†	0.97, 3.34
Received unwanted sexual comments on the Internet	2.59**	1.58, 4.23
Received pornographic junk mail in e-mail or Instant Messaging	1.90**	1.19, 3.04
Visiting Inappropriate websites		
Accidentally ended up in a pornographic website	1.68*	1.04, 2.73
Purposely visited a pornographic website	2.39**	1.33, 4.28
Accidentally ended up in a website with violent/gruesome images	1.60*	1.01, 2.54
Accidentally ended up in a hate website	1.44†	0.90, 2.33
Heard of the following Internet safety advice		
Never arrange to meet anyone	0.55*	0.33, 0.90
Do not download anything	1.88*	1.06, 3.17

**$p < 0.0001$.
*$p < 0.01$.
†$p < 0.05$.

3.3 Introduction à l'analyse discriminante

3.3.1 Présentation de la méthode

3.3.1.1 Position du problème

On dispose de n observations sur lesquelles on a relevé :

- les valeurs d'une variable catégorielle comportant quelques modalités (2, 3, ...) : c'est le groupe ou diagnostic.
- les valeurs de p variables numériques : X1, X2, ..., Xp : ce sont les prédicteurs.

On se pose des questions telles que :

- dans quelle mesure la valeur de Y est-elle liée aux valeurs de X1, X2, ..., Xp ?
- Etant donné d'autres observations, pour lesquelles X1, X2, ..., Xp sont connues, mais Y ne l'est pas, est-il possible de prévoir Y (le groupe), et avec quel degré de certitude ?

Exemples de situations où une telle méthode peut être intéressante :

Exemple 1. On étudie les différentes espèces de poissons peuplant un lac, mais la détermination exacte de l'espèce suppose que l'on sacrifie l'animal. Peut-on se contenter de relever différents paramètres concernant les poissons prélevés, et déduire l'espèce à partir de ces paramètres avec un degré de certitude raisonnable ?

Exemple 2. Pour déterminer le type d'utilisation de parcelles agricoles, on peut évidemment faire des relevés sur le terrain. Mais pourrait-on utiliser les informations données par des images satellites ?

La méthode est également utilisée sans que l'on ait un objectif de prédiction; on souhaite seulement déterminer les prédicteurs les plus liés au groupe d'appartenance.

3.3.1.2 Précautions et limites de la méthode

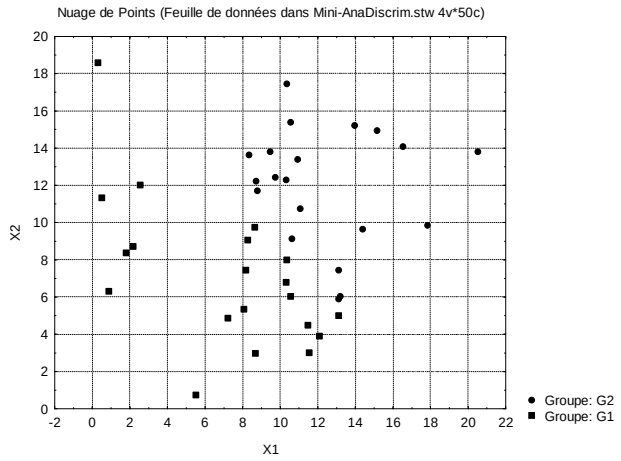
Comme dans le cas de la régression linéaire, l'emploi de cette méthode suppose que les variables prédictrices possèdent des propriétés de régularité satisfaisantes : distribution normale (voire multinormale) des variables Xi dans les différentes populations.

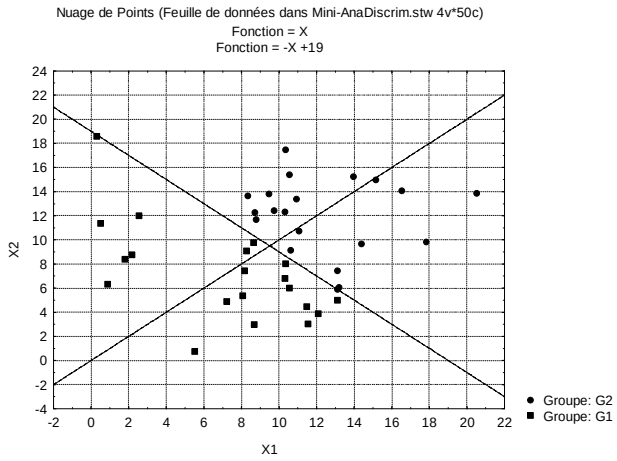
Par ailleurs (comme pour la régression linéaire), l'analyse discriminante peut conduire à des résultats incorrects si les variables Xi sont trop fortement corrélées entre elles.

3.3.2 Analyse discriminante sur un mini-exemple

3.3.2.1 Présentation de l'exemple

On a relevé les valeurs de deux variables X1 et X2 sur 40 individus statistiques répartis en deux groupes. Le nuage de points représentant ces observations est le suivant :

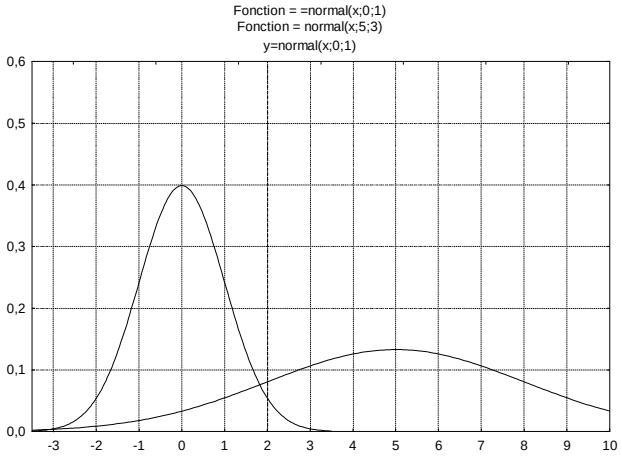




Prise isolément, aucune des deux variables X1 et X2 ne permet de différencier les deux groupes G1 et G2. Cependant, on voit bien que les deux groupes occupent des régions du plan bien spécifiques.
On voit cependant intuitivement que notre problème pourrait être résolu en considérant une variable abstraite, combinaison linéaire de X1 et X2 (approximativement X1 + X2) définie de façon que :

- la variance (dispersion) intra-groupes soit la plus petite possible
- la variance inter-groupes (variance calculée à partir des points moyens pondérés des groupes) soit la plus grande possible.

Ainsi, sur notre exemple, la droite d'équation Y= - X + 19 semble séparer correctement les deux groupes et il semblerait que c'est en projetant les points sur la droite Y=X que l'on obtiendra une dispersion minimale dans les groupes et maximale entre les groupes.
Remarque : Dans notre exemple, les deux groupes présentent à peu près la même dispersion de valeurs. Cependant, dans d'autres situations, l'un des groupes peut être nettement plus dispersé que l'autre.
Considérons la situation suivante, où l'on a représenté la distribution des valeurs issues des deux groupes sur le "facteur discriminant". On souhaite par exemple, affecter la valeur x=2 à l'un des deux groupes. Pour la distance "habituelle" (euclidienne), cette valeur est plus près du centre du premier groupe (valeur 0) que du centre du second groupe (valeur 5). Cependant, x=2 a plus de chances d'être une observation provenant du scond groupe qu'une observation provenant du premier groupe.

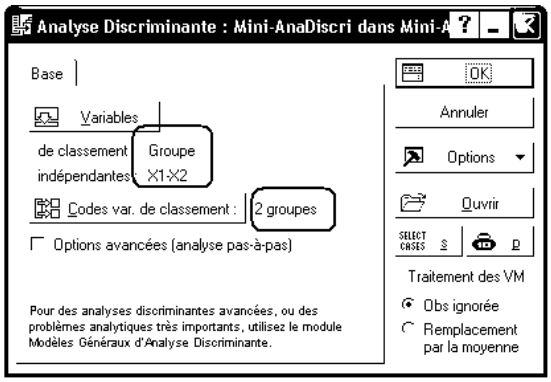


Pour résoudre ce problème, on introduit une distance particulière : la **distance de Mahalanobis** pour évaluer la distance entre un point et le centre d'un groupe. Pour calculer cette distance, on fait intervenir les écarts réduits entre x et les centres de groupes. On aura ainsi :

$$d_1^2(\bar{x}_1,2)=\left(\frac{2-0}{1}\right)^2=4 \quad ; \quad d_2^2(\bar{x}_2,2)=\left(\frac{2-5}{3}\right)^2=3$$

3.3.2.2 Traitement de l'exemple précédent avec Statistica

Ouvrez le fichier Mini-AnaDiscrim.stw
Faites une analyse discriminante (menu Statistiques - Techniques exploratoires multivariées - Analyse discriminante) en indiquant les codes G2 et G1 comme codes pour la variable catégorielle "Groupe", X1 et X2 comme variables indépendantes.



L'onglet Avancé nous donne accès aux boutons suivants :

Synthèse (variables dans le modèle) :

Synthèse de l'Analyse Discriminante (Mini-AnaDiscri dans Mini-AnaDiscrim.stw
Vars dans le modèle : 2; Classmt : Groupe (2 grps)
Lambda Wilk : ,38021 F approx. (2,37)=30,158 p< ,0000

N=40	Wilk (Lambda)	Partiel (Lambda)	F d'exc. (1,37)	niveau p	Tolér.	1-Tolér. (R^2)
X1	0,676419	0,562090	28,82580	0,000004	0,838237	0,161763
X2	0,668372	0,568857	28,04266	0,000006	0,838237	0,161763

Cette feuille donne les résultats d'un test

Distances inter-groupes, qui fournit trois feuilles de résultats :

Distance entre les centroïdes des deux groupes

Groupe	Dist. de Mahalanobis au Carré (Mini-AnaDiscri dans Mini-AnaDiscrim.stw	
	G2	G1
G2	0,000000	6,194525
G1	6,194525	0,000000

Un test statistique concernant la séparation des deux groupes

Groupe	Valeurs F ; dl 2,37 (Mini-AnaDiscri dans Mini-AnaDiscrim.stw	
	G2	G1
G2		30,15756
G1	30,15756	

Groupe	niveau p (Mini-AnaDiscri dans Mini-AnaDiscrim.stw	
	G2	G1
G2		1,6998E-8
G1	1,6998E-8	

Réaliser une analyse canonique, qui donne accès à un autre ensemble de résultats

Notamment, le bouton "Coefficients des variables canoniques" produit deux feuilles de résultats, dont la définition de la première variable canonique.

Variable	Coefficients bruts des Variables Canoniques
	Comp_1
X1	-0,241220
X2	-0,254916
Constte	4,776475
V.Propre	1,630138
Prop.Cum	1,000000

Ici, la première variable est C1 = -0,24122 X1 - 0,25916 X2 + 4,776475.

L'onglet Scores canoniques, puis le bouton "Scores canoniques de chaque observation" donnent la valeur de la variable canonique sur chaque observation. On voit ainsi que, sauf exception, les observations classées dans le groupe G2 ont des scores négatifs pendant que celles classées dans le groupe G1 ont des scores positifs.

En cliquant sur le bouton Annuler, on revient aux résultats de l'analyse discriminante proprement dite. L'onglet Classification donne accès aux résultats suivants :

Fonctions de classification

	Fonctions de classif. ; classement: Groupe	
Variable	G2 p=,50000	G1 p=,50000
X1	1,4380	0,83765
X2	1,5504	0,91594
Constte	-18,8231	-6,93504

La fonction discriminante linéaire du groupe G2 est :
F2 = 1,4380 X1 + 1,5504 X2 - 18,8231
Celle du groupe G1 est :
F1 = 0,83765 X1 + 0,91594 X2 - 6,93504
La méthode classe un élément dans le groupe G1 si F1 > F2 et dans G2 dans le cas contraire.

Matrice de classification

Ce tableau est encore appelé *Matrice de confusion*. Il croise la classification observée avec la classification calculée par la méthode.

	Matrice de Classification Lignes : classifications observées Colonnes : classifications prévues		
Groupe	% Correct	G2 p=,50000	G1 p=,50000
G2	90,00000	18	2
G1	95,00000	1	19
Total	92,50000	19	21

Classification d'observations

	Classification d'observations Classif. incorrectes indiquées par '		
Observation	Classif. Observée	1 p=,50000	2 p=,50000
1	G2	G2	G1
2	G1	G1	G2
3	G1	G1	G2
4	G2	G2	G1
5	G2	G2	G1
* 6	G2	G1	G2
7	G2	G2	G1

Ce tableau donne pour chaque observation, le groupe le plus probable (selon le calcul), ainsi que le second candidat. Il indique également le classement calculé des valeurs qui n'étaient pas classées a priori :

	Classification d'observations Classif. incorrectes indiquées par '		
Observation	Classif. Observée	1 p=,50000	2 p=,50000
40	G2	G2	G1
41	---	G2	G1
42	---	G1	G2
43	---	G1	G2
44	---	G2	G1
45	---	G1	G2
46	---	G1	G2

Distances de Mahalanobis au carré

Observation	Dist. Mahalanobis Carrées aux Centroïdes de Group Classif. incorrectes indiquées par *		
	Classif. Observée	G2 p=,50000	G1 p=,50000
1	G2	1,02516	3,21197
2	G1	9,43294	0,46701
3	G1	4,30475	0,81343
4	G2	1,71089	3,08567
5	G2	0,20940	6,53698
* 6	G2	2,95094	2,71562
7	G2	0,48679	4,17369

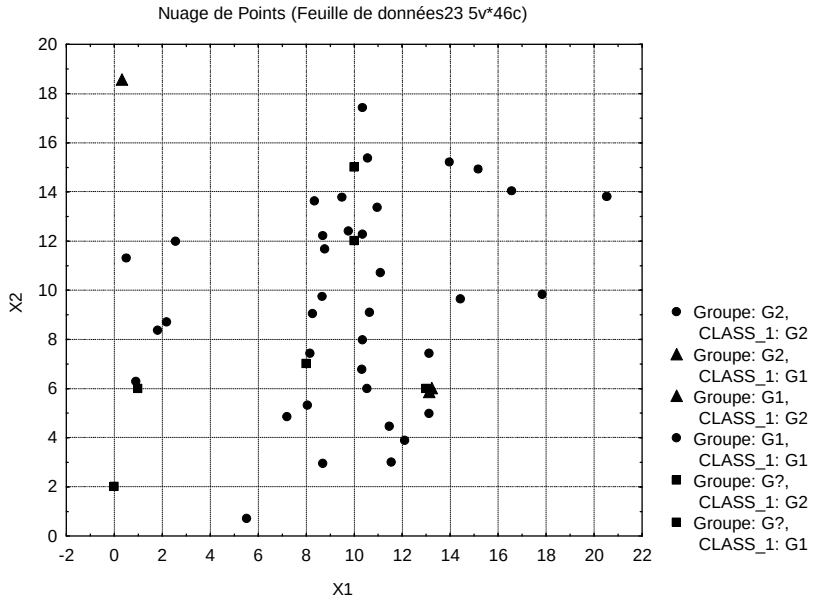
Probabilités a posteriori

Observation	Probabilités a posteriori Classif. incorrectes indiquées par *		
	Classif. Observée	G2 p=,50000	G1 p=,50000
1	G2	0,749022	0,250978
2	G1	0,011174	0,988826
3	G1	0,148595	0,851405
4	G2	0,665386	0,334614
5	G2	0,959449	0,040551
* 6	G2	0,470618	0,529382
7	G2	0,863356	0,136644

En fait pour chaque observation la méthode calcule une probabilité d'appartenance à chacun des deux groupes et affecte l'observation au groupe le plus probable.

Enregistrer les scores

Ce bouton permet de générer une feuille de données avec la classification produite par la méthode, et éventuellement, les variables et la classification initialement observées. Cette feuille de données peut être utilisée pour produire un nuage de points tel que le suivant :



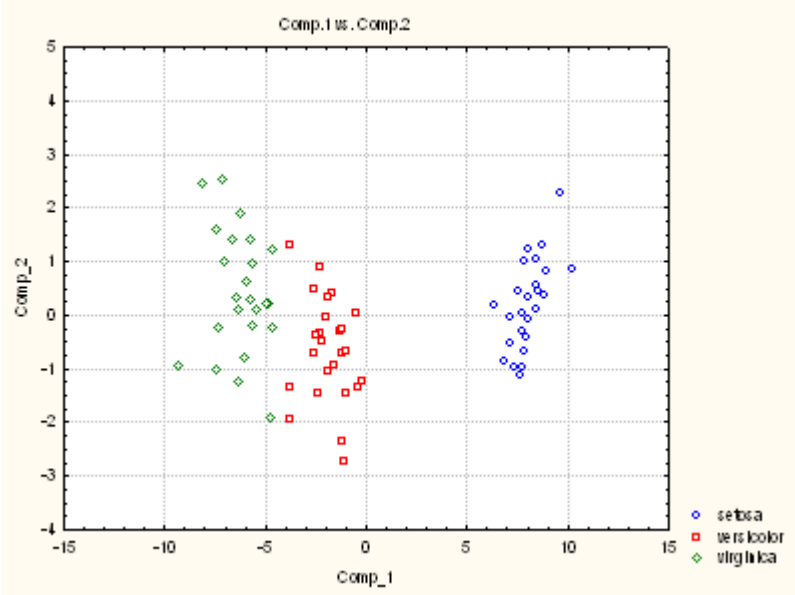
Dans ce graphique, les points bien classés sont représentés par des cercles, les points mal classés par des triangles, et les points supplémentaires par des carrés. La couleur (rouge ou noir) correspond au groupe calculé.

3.3.3 Les iris de Fisher

Ouvrez le classeur Iris.stw. Il s'agit d'un exemple, initialement proposé par Fisher, et utilisé comme données de référence par la plupart des logiciels de statistiques.

On a noté, pour 150 iris, l'espèce (setosa, versicolor, virginica) et 4 variables numériques : la longueur et la largeur des sépales, la longueur et la largeur des pétales. Pour chaque espèce, on

dispose de 50 observations. Les 25 premières observations de chaque espèce vont contituer l'ensemble d'apprentissage, tandis que les 25 observations restantes seront classifiées à l'aide des résultats de l'analyse discriminante. La classification ainsi obtenue pourra ainsi être comparée aux données réelles.
Procédez de même à une analyse discriminante sur ces données. Comme nous avons ici 4 variables numériques et 3 groupes, nous aurons deux facteurs discriminants, et Statistica nous permet de construire un graphique représentant les observations selon les valeurs de leurs scores canoniques :



3.4 Analyse et régression PLS

3.4.1 Position du problème

On a observé sur un échantillon de n individus statistiques :
- d'une part, p variables indépendantes ou explicatives : X_1, X_2, ..., X_p
- d'autre part, q variables dépendantes, ou "à expliquer" : Y_1, Y_2, ..., Y_q.

On souhaite établir entre les variables indépendantes et les variables explicatives une relation linéaire du type :

$$Y_1 = b_{10} + b_{11}X_1 + ... + b_{1p}X_p + \varepsilon_1$$
$$Y_2 = b_{20} + b_{21}X_1 + ... + b_{2p}X_p + \varepsilon_2$$
$$...$$
$$Y_q = b_{q0} + b_{q1}X_1 + ... + b_{qp}X_p + \varepsilon_q$$

On dispose déjà d'un outil s'appliquant à ce type de problème : la régression linéaire multiple. Cependant, la régression linéaire classique présente les inconvénients suivants :
- Elle "met en compétition" les différentes variables X_i, et elle est très sensible aux collinéarités entre les Xi, et même inutilisable si l'une des variables X_i est combinaison linéaire des autres variables.
- Elle ne peut pas être utilisée si le nombre d'observations (n) est inférieur au nombre de prédicteurs (p).

Une façon de contourner ces problèmes consiste à faire d'abord une ACP sur les prédicteurs, puis de réaliser la régression des variables dépendantes sur les variables principales ainsi définies. Mais le résultat n'est pas facilement interprétable par l'utilisateur.

L'idée de la régression PLS est de procéder de façon analogue à la régression sur composantes principales, mais en formant des composantes ou *variables latentes* tenant compte des variables à expliquer.

3.4.2 Le principe de la régression PLS sur un mini-exemple

Considérons les données suivantes (1 variable dépendante Y, 4 variables explicatives X_j, 3 sujets observés) :

	Y	X_1	X_2	X_3	X_4
s1	12	8	2	7	6
s2	10	2	12	5	7
s3	5	15	6	5	5

Afin d'éliminer les effets dus aux unités avec lesquelles sont mesurés les X_j, on introduit les variables Z_j, variables centrées réduites associées aux X_j.

Ainsi, les variables Z_j sont ici données par :

Y	Z_1	Z_2	Z_3	Z_4
0,8321	-0,0512	-0,9272	1,1547	0,0000
0,2774	-0,9734	1,0596	-0,5774	1,0000
-1,1094	1,0246	-0,1325	-0,5774	-1,0000

La première composante, ou variable latente P_1 est obtenue en pondérant les Z_j proportionnellement aux coefficients de corrélation $w_j=r(Y, X_j)$.

Sur notre exemple, les coefficients de corrélation valent :

	Y
X_1	-0,7247
X_2	-0,1653
X_3	0,7206
X_4	0,6934

On divise ces coefficients par un même nombre, de manière que la somme des carrés des poids soit égale à 1. On obtient ainsi les poids suivants :

$$w_1=-0,582 ; w_2 = -0,133 ; w_3 = 0,578 ; w_4 = 0,556$$

La variable latente P1 a donc pour valeur :

$$P_1 = - 0,582 * Z_1 - 0,133 * Z_2 + 0,578 * Z_3 + 0,556 * Z_4.$$

Sur les 3 observations, elle prend les valeurs suivantes :

	P_1
s1	0,8206
s2	0,6481
s3	-1,4687

La régression de Y par rapport à P_1 conduit à l'équation :

$$Y = 2,7640 P_1 + 9$$

et les valeurs estimées de Y sont :

	Y	Y estimé	Résidus
s1	12	11,2682	0,7318
s2	10	10,7915	-0,7915
s3	5	4,9404	0,0596

D'où un coefficient de détermination :

$$R^2(Y, Y \text{ estimé}) = 0,955$$

Il serait ensuite possible de recommencer la même méthode à partir des résidus de Y, pour produire une deuxième variable latente, et améliorer la qualité de l'estimation.

3.4.3 Un exemple de régression PLS avec Statistica

Dans l'ouvrage : M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. pp. 792-795, Hervé Abdi donne l'exemple suivant, que l'on trouve également sur son site, à partir de la page

http://www.utdallas.edu/~herve/#Articles.

On veut prévoir l'évaluation subjective d'un ensemble de 5 vins. Les variables dépendantes que nous voulons prédire sont son appréciation générale et la façon dont il s'accorde avec la viande et les desserts. Les prédicteurs sont le prix, le taux de sucre, le taux d'alcool, et l'acidité.

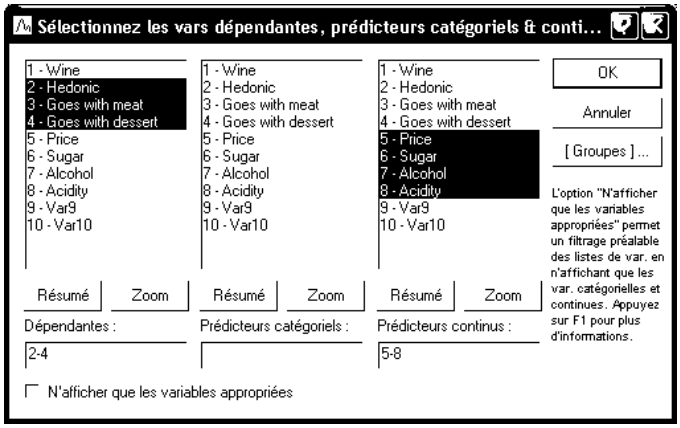
Les données sont les suivantes :

	1 Wine	2 Hedonic	3 Goes with meat	4 Goes with dessert	5 Price	6 Sugar	7 Alcohol	8 Acidity
1	1	14	7	8	7	7	13	7
2	2	10	7	6	4	3	14	7
3	3	8	5	5	10	5	12	5
4	4	2	4	7	16	7	11	3
5	5	6	2	4	13	3	10	3

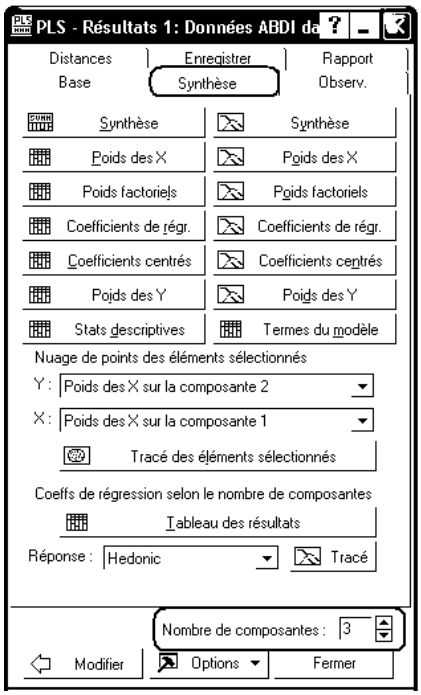
Ouvrez le fichier PLS-Abdi.stw.

La régression PLS est accessible à partir du menu : Statistiques - Modèles linéaires/non-linéaires avancés - Modèles généraux PLS - Modèles linéaires généraux.

Sélectionnez les variables comme suit :



La fenêtre de dialogue "Résultats" permet d'indiquer le nombre de variables latentes souhaité et comporte différents onglets :



L'onglet "Base" est entièrement repris dans l'onglet "Synthèse".
Le bouton "Synthèse" produit le résultat suivant :

	Synthèse de la PLS (Données ABDI dans PLS-Abdi.stw) Réponses : Hedonic Goes with meat Goes with dessert Options : NO-INTERCEPT AUTOSCALE						
	Augmente R² de Y	Moyenne R² de Y	Augmente R² de X	Moyenne R² de X	R² de Hedonic	R² de Goes with meat	R² de Goes with dessert
Comp 1	0,6333	0,6333	0,7045	0,7045	0,7053	0,9374	0,2572
Comp 2	0,2206	0,8540	0,2790	0,9835	0,7071	0,9851	0,8697
Comp 3	0,1044	0,9583	0,0165	1,0000	1,0000	1,0000	0,8750

Ce tableau nous donne le pourcentage de variance de chacune des variables dépendantes expliqué, pris en compte par le modèle, en séparant l'apport de chacune des composantes (colonnes R2 de Hedonic, R2 de Goes with meat, R2 de Goes with dessert). Il donne également le pourcentage global pour l'ensemble des 3 variables dépendantes (R2 de Y), obtenu simplement comme moyenne des 3 pourcentages précédents. Il indique également le pourcentage de variance des prédicteurs pris en compte par les composantes.

Le bouton "Poids des X" conduit au tableau suivant, qui donne l'expression des composantes en fonction des prédicteurs :

	Poids des prédicteurs (Données ABDI dans PLS-Abdi.stw Réponses : Hedonic Goes with meat Goes with dessert Options : NO-INTERCEPT AUTOSCALE			
	Price	Sugar	Alcohol	Acidity
Compo 1	-0,5137	0,2010	0,5705	0,6085
Compo 2	0,2343	0,9611	0,1267	0,0734
Compo 3	-0,3747	0,1291	-0,8069	0,4380

Ainsi, on a, sur les données centrées réduites :

Compo 1 = -0,51 * Price + 0,20 * Sugar + 0,57 * Alcohol + 0,61 * Acidity

Le bouton "Poids Factoriels" donne l'expression des prédicteurs en fonction des composantes :

Pds Fac. X (Données ABDI dans PLS-Abdi.stw)
Réponses : Hedonic Goes with meat Goes with desser
Options : NO-INTERCEPT AUTOSCALE

	Price	Sugar	Alcohol	Acidity
Comp 1	-0,5678	0,0142	0,5933	0,6032
Comp 2	0,3302	0,9638	-0,0136	-0,0268
Comp 3	-0,3496	0,1613	-0,8220	0,4222

Ainsi, en données centrées réduites :

Price = -0,57 * Compo1 + 0,33 * Compo2 - 0,350 * Compo3

Les boutons Coefficients de régression et Coefficients de régression centrés donnent les résultats de la régression (utilisant le modèle PLS). Les variables dépendantes estimées y sont exprimées en fonction des variables de départ.

Coefficient de régression PLS (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Ord.Ori	Price	Sugar	Alcohol	Acidity
Hedonic	48,5000	-1,0000	0,7500	-4,0000	2,7500
Goes with meat	-8,9167	-0,0333	0,2750	1,0000	0,1750
Goes with dessert	-3,8542	0,0417	0,5937	0,5000	0,0937

PLS coefficients de régression centrés (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Price	Sugar	Alcohol	Acidity
Hedonic	-1,0607	0,3354	-1,4142	1,2298
Goes with meat	-0,0745	0,2593	0,7454	0,1650
Goes with dessert	0,1250	0,7510	0,5000	0,1186

Ainsi, par exemple, en données non centrées réduites, on a :

Hedonic estimé = 48,5 - Price + 0,75 * Sugar - 4 * Alcohol + 2,75 * Acidity

(et il s'agit d'une valeur exacte, puisque R2=1 pour cette variable).

Le bouton "Poids des Y" donne l'expression des variables dépendantes (centrées réduites) en fonction des composantes :

Poids des réponses (Données ABDI dans PLS-Abdi.stw
Réponses : Hedonic Goes with meat Goes with dessert
Options : NO-INTERCEPT AUTOSCALE

	Hedonic	Goes with meat	Goes with dessert
Comp 1	0,6093	0,7024	0,3680
Comp 2	-0,0518	0,2684	0,9619
Comp 3	0,9672	-0,2181	-0,1301

L'onglet "Observ." donne quant à lui des tableaux des valeurs observées, valeurs prévues et résidus des variables dépendantes sur les différents individus statistiques observés :

	Valeurs prévues (Données ABDI dans PLS-Abdi.stw) Réponses : Hedonic Goes with meat Goes with desser Options : NO-INTERCEPT AUTOSCALE		
	Hedonic	Goes with meat	Goes with dessert
1	14,0000	7,0000	7,7500
2	10,0000	7,0000	5,7500
3	8,0000	5,0000	6,0000
4	2,0000	4,0000	6,7500
5	6,0000	2,0000	3,7500

Il donne également les scores des individus sur les composantes, calculés soit à partir des variables prédictives, soit à partir des variables dépendantes :

	Valeurs des prédicteurs et réponses (Données ABDI dans PLS-Abdi.stw Réponses : Hedonic Goes with meat Goes with dessert Options : NO-INTERCEPT AUTOSCALE					
	Comp. X 1	Comp. X 2	Comp. X 3	Comp. Y 1	Comp. Y 2	Comp. Y 3
1	1,4952	0,9663	0,2937	1,9451	0,7611	0,6191
2	1,7789	-1,0239	-0,2380	0,9347	-0,5305	-0,5388
3	0,0000	0,0000	0,0000	-0,2327	-0,6084	0,0823
4	-1,4181	1,1040	-0,2724	-0,9158	1,1575	-0,6139
5	-1,8560	-1,0464	0,2167	-1,7313	-0,7797	0,4513

3.5 Analyse de segmentation

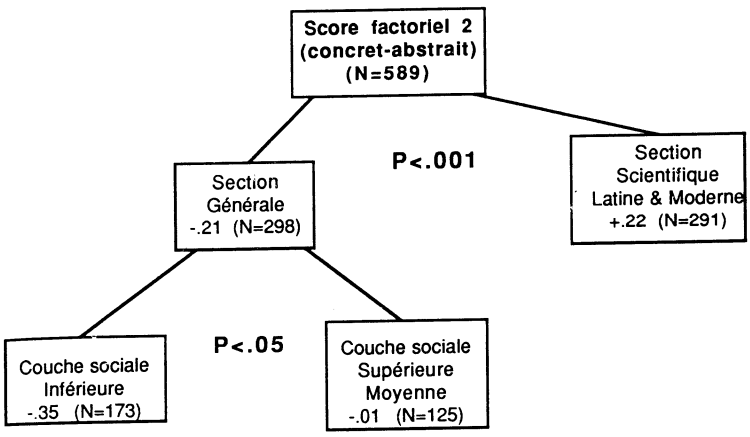
Bibliographie:
Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.
Doise, W., Clémence A., Lorenzi-Cioldi, F., Représentations sociales et analyses de données, Presses Universitaires de Grenoble, Grenoble, 1992
Idams : http://ead.univ-angers.fr/~statidams/

3.5.1 But de la méthode

On a observé sur un échantillon d'individus statistiques une variable dépendante numérique ou qualitative Y et plusieurs variables numériques ou catégorielles X1, X2, ..., Xp.

La segmentation vise à expliquer la variable Y à l'aide d'une ou plusieurs variables quantitatives ou qualitatives. Elle permet également de créer des groupes d'individus ou d'observations homogènes.

Le résultat est fourni sous la forme d'un arbre de décision binaire du type suivant :



3.5.2 Rappel : décomposition de l'inertie

Du point de vue de la variable dépendante, l'inertie totale est la somme des carrés des écarts à la moyenne générale :

$$I = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

On suppose les observations réparties en g groupes (j=1, 2, ..., g). Pour chacun des groupes, on a une moyenne du groupe : $\bar{y}_j$, un effectif nj et une inertie intra-groupe : $I_j = \sum_{i'=1}^{n_j} \left(y_{i'} - \bar{y}_j\right)^2$.

Une relation fondamentale est donnée par le théorème de Huygens : l'inertie totale est la somme des inerties intra-groupes et de l'inertie des points moyens des groupes, pondérés par l'effectif des groupes.

$$I = \sum_{j=1}^{g} I_j + \sum_{j=1}^{g} n_j \left(\bar{y}_j - \bar{y}\right)^2$$

$$\underset{\text{totale}}{\text{Inertie}} = \sum_{\text{les groupes}} \underset{\text{dans}}{\text{Inertie}} + \underset{\text{pondérés par}}{\text{Inertie des points moyens}}$$
$$\text{les effectifs des groupes}$$

Exemple : Soient les 4 observations suivantes, réparties en deux groupes A et B :

Groupe	A	B	A	B
Y	1	2	3	4

La moyenne générale est donnée par : $\bar{y} = 2{,}5$. L'inertie totale vaut :

$$\text{Inertie totale} = (1-2{,}5)^2 + (2-2{,}5)^2 + (3-2{,}5)^2 + (4-2{,}5)^2 = 5$$

Les inerties des deux groupes A et B (inerties intra-groupes) sont données par :

$$I_A = (1-2)^2 + (3-2)^2 = 2 \qquad I_B = (2-3)^2 + (4-3)^2 = 2$$

L'inertie des points moyens pondérés, ou inertie inter-groupes vaut :

$$I_{Inter} = 2\times(2-2{,}5)^2 + 2\times(3-2{,}5)^2 = 1$$

On vérifie bien que :

$$\text{Inertie totale} = I_A + I_B + I_{Inter} = 2 + 2 + 1 = 5$$

3.5.3 Principe de la méthode :

L'inertie inter-groupes mesure l'hétérogénéité entre les différents groupes alors que l'inertie intra-groupes mesure l'homogénéité à l'intérieur des groupes. Pour obtenir des groupes les plus distincts possibles, il faut une inertie inter-groupes le plus élevée possible. L'inertie intra-groupe sera alors faible et donc les individus d'un même groupe seront homogènes.

1) Au départ, on dispose d'un seul segment contenant l'ensemble des individus.
2) A la première étape, la procédure de construction de l'arbre examine une par une toutes les variables explicatives. Pour chaque variable, elle passe en revue toutes les divisions possibles (de la forme Xj< A et Xj > A si Xj est numérique, regroupement des modalités en deux sous-ensembles si Xj est catégorielle). Pour chaque division, l'inertie inter-groupes est calculée.
3) La division choisie est celle qui maximise l'inertie inter-groupes.
4) On recommence la procédure dans chacun des deux groupes ainsi définis.

Critères d'arrêt :

On peut utiliser comme critères d'arrêt de l'algorithme de segmentation :
- La taille des groupes (classes) à découper
- Le rapport entre l'inertie intra et la variance totale
- Des tests statistiques (tests de Student de comparaison de moyennes, tests du Khi deux)

3.5.4 Exemple d'analyse de segmentation

Source : http://lib.stat.cmu.edu/datasets/

Determinants of Wages from the 1985 Current Population Survey

Summary:

The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine (i) whether wages are related to these characteristics and (ii) whether there is a gender gap in wages.

Based on residual plots, wages were log-transformed to stabilize the variance. Age and work experience were almost perfectly correlated (r=.98). Multiple regression of log wages against sex, age, years of education, work experience, union membership, southern residence, and occupational status showed that these covariates were related to wages (pooled F test, $p < .0001$). The effect of age was not significant after controlling for experience. Standardized residual plots showed no patterns, except for one large outlier with lower wages than expected. This was a male, with 22 years of experience and 12 years of education, in a management position, who lived in the north and was not a union member. Removing this person from the analysis did not substantially change the results, so that the final model included the entire sample.

Adjusting for all other variables in the model, females earned 81% (75%, 88%) the wages of males ($p < .0001$). Wages increased 41% (28%, 56%) for every 5 additional years of education ($p < .0001$). They increased by 11% (7%, 14%) for every additional 10 years of experience ($p < .0001$). Union members were paid 23% (12%, 36%) more than non-union members ($p < .0001$). Northerns were paid 11% (2%, 20%) more than southerns ($p =.016$). Management and professional positions were paid most, and service and clerical positions were paid least (pooled F-test, $p < .0001$). Overall variance explained was $R2 = .35$.

In summary, many factors describe the variations in wages: occupational status, years of experience, years of education, sex, union membership and region of residence. However, despite adjustment for all factors that were available, there still appeared to be a gender gap in wages. There is no readily available explanation for this gender gap.

Authorization: Public Domain

Reference: Berndt, ER. The Practice of Econometrics. 1991. NY: Addison-Wesley.

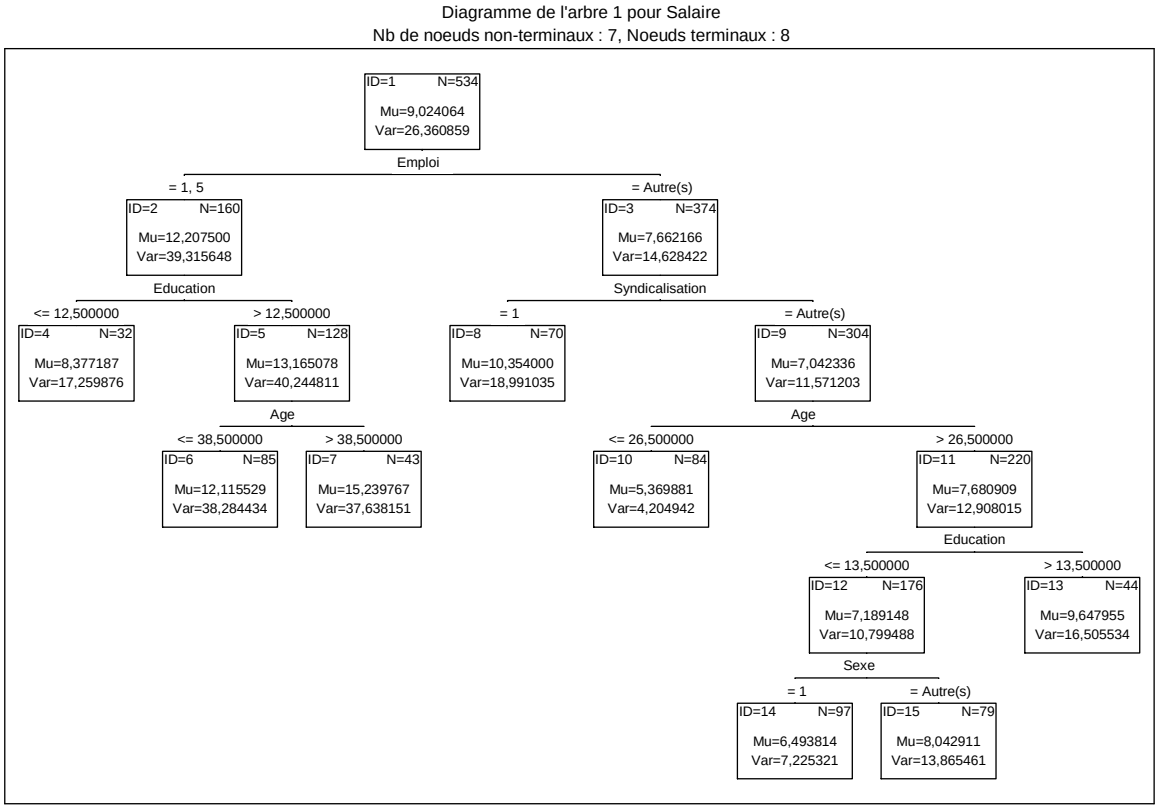
Description: The datafile contains 534 observations on 11 variables sampled from the Current Population Survey of 1985. This data set demonstrates multiple regression, confounding, transformations, multicollinearity, categorical variables, ANOVA, pooled tests of significance, interactions and model building strategies.

Variable names in order from left to right:
EDUCATION: Number of years of education.
SOUTH: Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).
SEX: Indicator variable for sex (1=Female, 0=Male).
EXPERIENCE: Number of years of work experience.
UNION: Indicator variable for union membership (1=Union member, 0=Not union member).
WAGE: Wage (dollars per hour).
AGE: Age (years).
RACE: Race (1=Other, 2=Hispanic, 3=White).
OCCUPATION: Occupational category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other).
SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).
MARR: Marital Status (0=Unmarried, 1=Married)

Ces données (avec des noms de variables francisés) se trouvent dans la feuille de données du classeur CPS_85_Wages.stw.

La variable à expliquer est évidemment la variable WAGE.
On constate que les variables SOUTH, SEX, UNION, RACE, OCCUPATION, SECTOR, MARR sont des variables catégorielles, alors que les variables EDUCATION, EXPERIENCE, AGE peuvent être considérées comme numériques.

On présente ci-dessous l'arbre obtenu en indiquant "Salaire" comme variable dépendante, Education, Expérience et Age comme variables numériques, Localisation, Sexe, Syndicalisation, Origine Ethnique, Emploi, Secteur et Situation de famille comme variables catégorielles et en adoptant la règle d'arrêt : on ne segmente pas les groupes d'effectifs inférieurs à 100 :

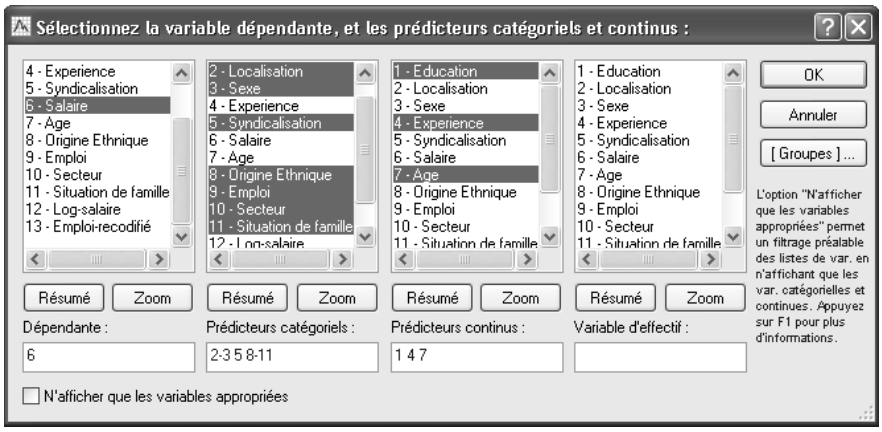


3.5.5 Traitements sous Statistica

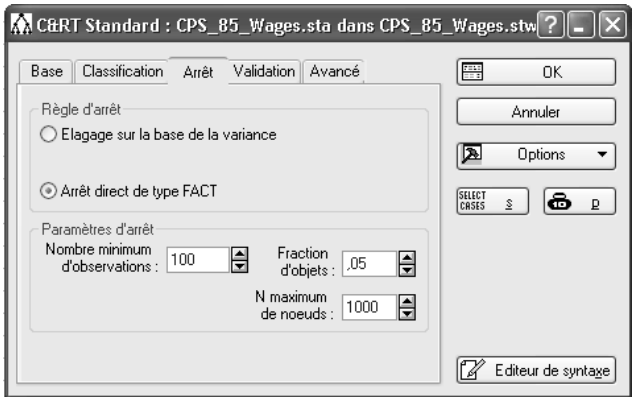
Ouvrir le classeur CPS_85_Wages.stw.

Utiliser le menu Statistiques - Data Mining - Modèles d'arbres de classification et de régression - C&RT Standard.
N.B. "C&RT" signifie : classification and regression trees. L'aide de Statistica indique : "Modèles d'Arbres de Classification et de Régression (GC&RT)",

Sous l'onglet "Standard", indiquer "Salaire" comme variable dépendante, Education, Expérience et Age comme variables numériques (prédicteurs continus), Localisation, Sexe, Syndicalisation, Origine Ethnique, Emploi, Secteur et Situation de famille comme variables catégorielles (prédicteurs catégoriels) :



Sous l'onglet "Arrêt", indiquer 100 comme nombre minimum d'observations.



Cliquer sur le bouton OK puis cliquer sur le bouton "Diagramme de l'arbre" du dialogue "Résultat". Vous devriez obtenir le graphique représenté ci-dessus.

3.5.6 Retrouver avec Statistica les résultats indiqués dans la présentation de l'exemple

Observez la colonne 12 : les logarithmes des salaires y ont été calculés.
Observez également la colonne 13 : la variable Emploi y a été recodifiée afin que les codes numériques des types d'emploi soient classés dans le même ordre que les salaires moyens des différents types.

Calculez le coefficient de corrélation entre Age et Expérience : on trouve effectivement r=0,977.

Faites la première régression indiquée : Log-Salaire (v12) est la variable dépendante, Sexe, Age, Education, Expérience, Syndicalisation, Localisation et Emploi-recodifié (v1 à v5, v7 et v13) sont les prédicteurs. On trouve effectivement F(7, 526)=39,56 et donc un effet significatif de l'ensemble de ces variables.

Calcul de l'effet de l'âge après contrôle de l'expérience : on peut, par exemple, calculer les résidus de la régression de Log-Salaire par rapport à Expérience, puis coller ces résidus dans une colonne supplémentaire de la feuille de données et tester le résultat de la régression de ces résidus par rapport à Age. Le coefficient b* vaut alors 0,09, significatif à 3% seulement.

La valeur atypique citée dans le texte pourra être mise en évidence sur l'un des graphiques de la première régression multiple (par exemple en demandant le tracé des bandes de prévision). Il s'agit de l'observation n° 200.

Les indications de l'avant-dernier paragraphe pourront être retrouvées en effectuant une nouvelle régression linéaire multiple en prenant Log-Salaire (v13) comme variable dépendante, Education, Localisation, Sexe, Expérience, Syndicalisation, Emploi-recodifié (v1 à v5 et v13) comme variables prédictrices.

On trouve bien un coefficient de détermination voisin de 0,35. On remarque que les coefficients des variables prédictrices sont alors tous significativement différents de 0. Pour retrouver les pourcentages indiqués, il faut reconvertir les effets (linéaires, additifs) des coefficients bi sur Log-Salaire en effets multiplicatifs sur Salaire (qui est l'exponentielle de Log-Salaire).

Par exemple, le coefficient bi pour Expérience est b4 = 0,011. Pour 10 ans d'expérience supllémentaires (toutes choses égales par ailleurs), Log-Salaire augmente de 0,1103. L'effet multiplicatif sur le salaire est obtenu en calculant l'exponentielle de cette valeur : exp(0,1103)=1,1166. L'effet se traduit donc par une augmentation du salaire de 11,66%.

Pour l'ensemble des prédicteurs, les calculs menés sous Excel conduisent au tableau suivant :

	b*	Err-Type	b	Err-Type	t(527)	valeur p	b* Coef	% Variation
OrdOrig.			0,6457	0,1193	5,4100	0,0000%		
Education	0,3788	0,0409	0,0764	0,0083	9,2611	0,0000%	0,3822	46,55%
Localisation	-0,0923	0,0358	-0,1070	0,0415	-2,5751	1,0294%	-0,1070	-10,15%
Sexe	-0,1791	0,0365	-0,1895	0,0387	-4,9009	0,0001%	-0,1895	-17,26%
Experience	0,2587	0,0382	0,0110	0,0016	6,7643	0,0000%	0,1103	11,66%
Syndicalisation	0,1407	0,0362	0,1932	0,0497	3,8904	0,0113%	0,1932	21,32%
Emploi-recodifié	0,2288	0,0386	0,0870	0,0147	5,9227	0,0000%		

Enfin, concernant l'effet du sexe, on pourra faire une troisième régression multiple en prenant l'ensemble des prédicteurs de la régression précédente, à l'exception du sexe. On calcule les résidus de cette régression et on les place dans une nouvelle colonne de la feuille de données. Enfin, on réalise une régression linéaire en utilisant ces résidus comme variable dépendante et le sexe comme prédicteur. On obtient alors (en ajoutant le traitement sous Excel) :

	b*	Err-Type	b	Err-Type	t(532)	valeur p	b* Coef	% Variation
OrdOrig.			0,0810	0,0252	3,2135	0,14%		
Sexe	-0,2015	0,0425	-0,1765	0,0372	-4,7442	0,00%	-0,1765	-16,18%

Il est assez remarquable que cet écart entre les sexes se retrouve, avec des valeurs presque identiques, dans l'ensemble des régressions qui ont été faites.

Remarque. Les résultats ainsi obtenus sont proches de ceux annoncés dans la présentation de l'exemple, sans être strictement identiques. Les différences proviennent probablement de la façon de traiter les variables catégorisées, notamment la variable Emploi.

4 Bibliographie :

Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.
Doise, W., Clémence A., Lorenzi-Cioldi, F., Représentations sociales et analyses de données, Presses Universitaires de Grenoble, Grenoble, 1992
Escofier, B., Pagès J., Analyses factorielles simples et multiples, Dunod, Paris, 1998.
Bry, X., Analyses factorielles simples, Economica, Paris, 1995.
Bry, X., Analyses factorielles multiples, Economica, Paris, 1996.
Morineau A., Morin S., Pratique du traitement des enquêtes - Exemple d'utilisation du système SPAD, Cisia-Ceresta, Montreuil, 2000
Croutsche, J.-J., Pratiques statistiques en gestion et études de marchés, Editions ESKA, Paris, 1997
Howell, D.C., Méthodes Statistiques en Sciences Humaines, De Boeck, Paris Bruxelles, 1998.
Saporta, G., Probabilité Analyse des données et statistique. Editions Technip , 1990

Articles :
Flament C., Milland L., Un effet Guttman en ACP, Mathématiques & Sciences humaines (43e année, n° 171, 2005, p. 25-49)
Hahn A., Eirmbter W. H., Jacob R., Le sida : savoir ordinaire et insécurité, traduction française de Herrmann M.
Costarelli, S., Callà, R.-M.. Self-directed negative affect: The distinct roles of ingroup identification and outgroup derogation, Current research in Social Psychology, Volume 10 No 2, 2004.
Gil-Monte P. R., Ma Peiró J., A study on significant sources of the burnout syndrome in workers at occupational centres for mentally disabled, , Psychology in Spain, 1997, Vol. 2. No 1, 116-123.
Page Web : http://www.psychologyinspain.com/content/full/1997/6bis.htm

Sites internet :
Site Eurostat de l'Union Européenne : http://epp.eurostat.ec.europa.eu/portal/
Site d'Hervé Abdi : http://www.utdallas.edu/~herve/#Articles.
Idams : http://ead.univ-angers.fr/~statidams/
Statlib - datasets archive : http://lib.stat.cmu.edu/datasets/
Psychologie Sociale : http://www.psychologie-sociale.org/reps2.php?article=7
Site pour télécharger ce polycopié et les fichiers d'exemples : http://geai.univ-brest.fr/~carpenti/

5 Table des matières