SPAD•D

VERSION 5.0

Procédures de décision

CISIA-CERESTA - 261 rue de Paris - 93556 MONTREUIL Cedex

tel : +33 (0)1 55 82 15 15 - Fax : +33 (0)1 43 63 21 00

e-mail : cisia@fr.inter.net - Web : http://www.cisia.com

SPAD•D

Option Décision du logiciel SPAD

MANUEL DE REFERENCE

SPAD•D ®

Manuel de référence

*Le **logiciel** décrit dans le manuel est diffusé dans le cadre d'un accord de licence d'utilisation et de non divulgation, et ne peut être utilisé ou copié qu'en conformité avec les stipulations de l'accord. Toute copie du programme sur cassette, disque ou autre support à des fins autres que l'usage personnel du programme par le licencié est interdite par la loi. Les informations figurant dans ce **manuel** sont sujettes à révision sans préavis et ne présentent aucun engagement de la part du CISIA.*

ISBN 2-906711-20-9

Centre **I**nternational de **S**tatistique et d'**I**nformatique **A**ppliquées
261 rue de Paris, 93556 Montreuil Cedex (France)
Tel : 01 55 82 15 15 – Fax : 01 43 63 21 00
e-mail : cisia@fr.inter.net – Web : http://www.cisia.com/

SPAD•D

Option DECISION du logiciel SPAD

La première partie de ce manuel est consacrée pour l'essentiel aux notices de référence des procédures de SPAD dédiées aux modèles statistiques.

Dans la seconde partie, on trouvera des précisions et des exemples concernant les aspects peut-être moins connus de certaines méthodes : la fonction de score telle qu'on peut la construire dans SPAD, le modèle log-linéaire avec recherche des meilleurs modèles, et enfin l'utilisation des réseaux de neurones pour faire de l'analyse discriminante.

Pour une présentation plus complète des méthodes, le lecteur est renvoyé sur l'ouvrage "*Statistique Exploratoire Multidimensionnelle*" de L. Lebart, A. Morineau et M. Piron, paru chez Dunod (1995). On y trouvera également une bibliographie abondante et récente.

SOMMAIRE

Introduction

Procédures d'ajustements et modèles

L'option SPAD•D de SPAD concerne des procédures dites de "décision". L'utilisateur choisit un modèle de dépendance entre les variables et utilise les observations pour estimer les paramètres inconnus du modèle. La statistique classique permet d'évaluer la qualité des estimations. Le modèle peut ensuite être utilisé pour extrapoler ou pour prévoir des observations non réalisées. Prévoir une valeur est le support de la décision (par exemple prévoir à quelle classe appartient un individu). Dans le cadre de certaines hypothèses de travail, le modèle peut être aussi utilisé pour tester la réalité des dépendances supposées.

Les modèles utilisés ici sont linéaires, non pas au titre des dépendances, mais du point de vue des coefficients inconnus à estimer. Si la variable à prévoir est du type continue, on parle de régression multiple. Dans cette famille on trouvera les méthode d'analyse de la variance et de la covariance utilisées en particulier pour le traitement des plans d'expérience.

Si la variable à prévoir est de type nominal, on parlera d'analyse discriminante. Deux techniques sont disponibles dans SPAD : l'analyse discriminante linéaire à 2 groupes, et l'analyse discriminante neuronale à k groupes. La discrimination entre deux groupes d'individus est d'une grande facilité d'interprétation, et présente l'avantage d'une analogie de forme avec la régression multiple. Une application particulièrement intéressante de la discriminante est le "scoring" auquel une procédure particulière est consacrée. La procédure neuronale permet de discriminer k groupes. Elle est particulièrement adaptée à l'analyse de relations non-linéaires. La méthode de calcul est basée sur l'utilisation d'un réseau de neurones multicouche.

Le choix d'un modèle d'ajustement linéaire est une opération délicate pour le statisticien. Il peut être guidé dans cette voie par des analyses exploratoires de type factorielles et des classifications permettant de décrire les dépendances entre paramètres. Ensuite il restera à choisir, dans un champ plus restreint de modèles, celui ou ceux qu'il faudra retenir. La procédure FUWIL peut être utilisée dans cette phase de choix. `

Les modèles log-linéaires permettent d'étudier les relations entre variables nominales par analyse des fréquences dans les tableaux de contingence à entrées multiples. La procédure d'estimation des modèles log-linéaires constitue dans SPAD un complément important aux procédures d'analyse des correspondances.

LISTE DES PROC

Ajustements et modèles

VAREG	*régressions multiples et analyses de variance*
DIS2G	*analyse discriminante à 2 groupes*
SCORE	*création et étude d'une fonction score*
NEURO	*analyse discriminante neuronale*
FUWIL	*sélection des ajustements optimaux*
LOGLI	*modèles log-linéaires*
MLGEN	*modèle linéaire général*

VAREG (régression et analyse de variance)

Il s'agit d'une procédure générale d'ajustement de modèles linéaires permettant de réaliser une très grande variété d'analyses statistiques. Citons:

- les régressions simples et multiples
- les analyses de la variance pour traiter tout plan d'expérience
  nombre quelconque de facteurs et de niveaux par facteur
  nombre quelconque d'interactions d'ordre 2 ou 3
  nombre de répétions quelconques, non égales
  blocs équilibrés, carrés latins, etc
- les analyses de covariance sans limitation sur le nombre de covariables
  sans limitation sur les interactions de facteurs
  sans limitation sur le nombre d'observations par case, etc.

Les modèles à ajuster sont écrits par l'utilisateur avec une notation "algébrique" très simple, telle que : V4 = V1 + V2 + V7 + V1*V2. La notation "V1*V2" indique l'introduction de l'interaction des facteurs V1 et V2.

Quelle que soit l'analyse, on peut sélectionner un sous groupe d'individus, et on peut faire intervenir une pondération dans tous les calculs. L'estimation d'un coefficient est toujours accompagnée de l'estimation de son écart-type, de la valeur du *t* de Student, de la probabilité critique correspondante et de son expression en terme de valeur-test. La procédure contient en option un traitement automatique des données manquantes.

Chaque test *F* de Fisher associé à la décomposition de la variance est édité avec sa probabilité critique et la valeur-test correspondante pour faciliter l'évaluation des effets des facteurs et des interactions présents dans l'analyse. On édite les statistiques classiques des ajustements: somme des carrés d'écarts, estimation de la variance résiduelle, coefficient de corrélation multiple, test global de nullité de tous les coefficients et valeur-test associée.

On peut éditer les statistiques usuelles caractérisant les variables entrant dans le modèle, ainsi que la matrice des corrélations et la matrice des covariances. En sortie de la procédure, on peut créer un fichier "texte" contenant les principaux résultats, en particulier les coefficients de l'ajustement. Ce fichier est utilisé par le *module graphique* disponible sur la version PC.

DIS2G (analyse discriminante à 2 groupes)

La procédure DIS2G réalise une analyse discriminante linéaire à 2 groupes selon la méthode classique de Fisher. Dans sa forme usuelle la variable "y" à prédire est une variable nominale à 2 modalités, et les variables "x" sont des variables continues. Dans le cas de 2 modalités, l'analyse discriminante est formellement équivalente à une régression, ce qui justifie certains calculs réalisés par la procédure.

L'analyse peut être réalisée en utilisant comme variables "x" les axes factoriels d'une analyse préalable. Dans le cas d'une analyse en **composantes principales**, cette procédure permet de choisir les axes les plus intéressants pour établir la fonction discriminante. On pourra par exemple éliminer des axes éloignés, porteurs de fluctuations aléatoires ou sans intérêt pour la discrimination. Après calcul des coefficients relatifs aux axes factoriels, le programme les transforme pour exprimer la fonction discriminante finale sur les variables d'origine.

Dans le cas d'une analyse des **correspondances multiples**, le programme établit la fonction discriminante sur les axes choisis par l'utilisateur, puis calcule les coefficients attribués aux modalités des variables nominales de l'analyse. Cette méthode permet donc de réaliser de façon naturelle les analyses discriminantes sur variables nominales.

La procédure DIS2G possède plusieurs méthodes de validation des résultats. Une méthode consiste à scinder l'échantillon en deux parties: l'une pour calculer la fonction discriminante (échantillon *"d'apprentissage"*), l'autre pour évaluer la qualité de la discrimination (échantillon *"test"*). L'échantillon test peut être construit automatiquement par tirage au hasard, ou défini par l'utilisateur soit par liste soit par filtre logique sur les variables.

Le programme intègre de plus une procédure de validation des résultats par "*bootstrap*". On obtient ainsi des estimations sans biais pour les coefficients, les écarts-types des coefficients, les corrélations de la fonction discriminante avec les variables d'origine, les pourcentages de bien et mal classés et les écarts-types associés à ces pourcentages.

Pour tout individu *anonyme* (dont on ne connaît pas le groupe), la procédure calcule la probabilité d'appartenance à chaque groupe.

Dans tous les cas on peut, avant de calculer la fonction discriminante, introduire des probabilités a priori d'appartenance aux groupes, ainsi qu'une matrice de coûts a priori. On peut sélectionner aisément les individus qui participeront aux calculs et les munir d'un poids de redressement (voir la procédure SELEC).

La procédure fournit de nombreuses éditions de résultats, en particulier les affectations des individus dans les groupes, accompagnées de la probabilité d'appartenance, les statistiques des variables pour chaque groupe, ainsi que les matrices de corrélations, et les histogrammes superposés des distributions des individus dans les groupes.

La procédure DIS2G crée deux fichiers de résultats permettant de communiquer avec l'extérieur. Un premier fichier contient le reclassement des individus par la fonction discriminante en 4 catégories: bien ou mal classé dans chaque groupe. Ce fichier est un fichier interne de SPAD, donc archivable et récupérable pour tout traitement statistique interne au logiciel.

Un second fichier, de type texte, contient les coefficients de la fonction discriminante. Ce fichier est utilisé comme véhicule des résultats vers la procédure SCORE lorsque l'on veut étudier une fonction de score.

SCORE (création et étude de scores)

La procédure SCORE est exécutable après une analyse discriminante à 2 groupes réalisée sur des variables nominales (procédure CORMU suivie de DIS2G). Le score attribué à un individu s'obtient en additionnant les coefficients associés aux modalités de l'individu.

Les coefficients sont automatiquement récupérés après l'analyse discriminante. Il est possible cependant d'étudier l'effet d'une modification des coefficients ou l'effet des arrondis sur les valeurs en introduisant soi-même les coefficients corrigés de la fonction. Tous les calculs seront réalisés avec ces coefficients.

L'introduction d'une tolérance d'erreur de classement permet de définir trois zones de décision sur la fonction de score: la *zone verte* du côté des scores forts, la *zone rouge* du côté des scores faibles, et la zone intermédiaire ou *zone d'indécision*. Un graphique permet d'apprécier comment la zone d'indécision diminue quand la tolérance d'erreur augmente.

Le tableau des coefficients est édité en rangeant les variables dans l'ordre décroissant de leur participation maximale au score. Dans chaque variable, les modalités sont rangées dans l'ordre décroissant de leur contribution au score.

Indépendamment de la pondération générale pour tous les calculs (procédure SELEC) on peut utiliser ici une pondération sur les deux groupes de la variable à discriminer pour rendre les effectifs représentatifs de ceux de la population.

Une abaque permet de lire, pour chaque valeur du score, l'estimation de la *probabilité conditionnelle* d'appartenir à chaque groupe. Des graphiques fournissent également la répartition des groupes et l'estimation de la répartition de la population en fonction de la valeur du score.

La procédure crée un fichier de résultats contenant les scores calculés pour chaque individu. Ces données sont archivables dans SPAD et donc réutilisables pour tout traitement ultérieur (par exemple pour des graphiques).

NEURO (analyse discriminante neuronale)

La procédure NEURO réalise une analyse discriminante pour une variable nominale à k groupes. La méthode est basée sur l'utilisation d'un réseau de neurones multicouche. Un réseau est composé de plusieurs couches, chaque couche étant constituée de neurones.

La variable "y" à prédire est une variable nominale à k modalités et les variables explicatives "x" sont continues. Sur la couche de sortie, il y a autant de neurones que de groupes k à discriminer. Sur les couches d'entrée, il y a autant de neurones que de variables continues "x".

Comme pour DIS2G, l'analyse peut être réalisée en utilisant comme variables explicatives, les axes factoriels d'une analyse préalable. (On fera par exemple une analyse des correspondances multiples afin de réaliser une discriminante neuronale sur variables qualitatives).

La procédure de calcul se divise en 2 phases :

- la propagation (ou relaxation)
  Les neurones d'une couche inférieure sont reliés aux neurones d'une couche supérieure par des connexions appelées "poids" et chaque neurone de chaque couche est affecté d'un "biais". Les poids et les biais sont pris au hasard lors de la première présentation des individus. On calcule les valeurs de chaque neurone avec le réseau de connexions calculé au temps t (à l'itération t). Les neurones de sortie seront combinaisons linéaires des neurones d'entrée avec les poids et les biais calculés.
- la rétropropagation
  Elle consiste à modifier les poids et les biais en fonction de l'erreur calculée afin que les données d'entrée fournissent de meilleures prévisions d'affectations.

Pour valider les résultats, on cherche les meilleurs poids et biais en utilisant le fichier d'apprentissage. On évalue la qualité de la discrimination en utilisant un échantillon-test. L'échantillon-test peut être calculé soit par tirage au hasard d'individus-tests, soit par liste, soit par filtre logique sur les variables.

Pour tout individu anonyme, la procédure calcule la classe de sortie, c'est à dire le groupe d'affectation.

La procédure NEURO, comme la procédure DIS2G, crée deux fichiers de résultats permettant de communiquer avec l'extérieur. Un premier fichier contient le reclassement des individus par groupe. C'est un fichier interne à SPAD, donc archivable et récupérable pour des traitements ultérieurs.

Le second fichier, de type texte, contient les poids et biais de la dernière itération. Le fichier peut être récupéré pour affiner les résultats par des itérations supplémentaires.

FUWIL (sélection des ajustements optimaux)

La procédure FUWIL est utilisée pour aider au choix des "meilleures" variables à introduire dans un modèle d'ajustement linéaire. Elle servira aussi bien dans le cas de la *régression multiple* que dans le cas de la *discriminante à deux groupes*, formellement équivalente à une régression.

L'utilisateur dispose de trois critères de comparaison globale des ajustements:

- le coefficient "R^2"de corrélation multiple
- le coefficient de corrélation multiple "corrigé"
- le coefficient "Cp" de Mallows

La notion de "meilleur" ajustement est relative ici au critère choisi. En fait l'utilisateur fera intervenir en général beaucoup d'autres considérations au moment de la sélection du modèle final (en particulier les valeurs-tests des coefficients).

Le programme édite les ajustements avec une seule variable, de la meilleure à la moins bonne. Puis il édite les ajustements à deux variables, du meilleur couple de variables au moins bon. Ensuite il édite les ajustements à trois variables, etc.

Ces éditions sont accompagnées de la valeur du critère global, ainsi que des principales informations sur les coefficients: valeurs, écarts-types, probabilités critiques et valeurs-tests. L'utilisateur dispose ainsi des éléments essentiels pour le choix de son modèle. Un graphique de synthèse montre comment le critère globale évolue d'un ajustement à l'autre.

LOGLI (modèles log-linéaires)

La procédure LOGLI met en oeuvre des modèles log-linéaires sur deux types de tableaux en entrée :

- soit un tableau de contingence multidimensionnel fourni par l'utilisateur (sous une forme adaptée au logiciel) ;
- soit le fichier "individus x variables" et dans ce cas le tableau de contingence à entrées multiples est construit automatiquement par la procédure.

Le programme accepte des modèles contenant jusqu'à 7 variables et des interactions jusqu'à l'ordre 7.

Le modèle à ajuster est de la forme suivante :

$$\text{Log } y = V1 + V2 + V7 + V1*V2$$

La notation "V1*V2" indique l'introduction de l'interaction des variables V1 et V2.

Tous les modèles considérés seront "hiérarchiques", autrement dit ils doivent contenir les effets principaux quand ils contiennent une interaction.

Les modèles peuvent être introduits de plusieurs façons :

- les modèles sont définis par l'utilisateur.
- tous les modèles possibles sont calculés et édités.
- les modèles sont sélectionnés par une méthode "pas à pas".
- les modèles sont construits par une méthode combinatoire.

Dans la recherche automatique de modèles, le nombre de variables explicatives de chaque modèle est limité à quatre.

Tous les modèles peuvent être calculés pour un sous-groupe d'individus et on peut faire intervenir une pondération dans les calculs.

Pour chaque modèle, on édite la statistique du maximum de vraisemblance, l'estimation du χ^2 de Pearson, et la statistique AIC (critère de l'information d'Akaike). On peut éditer les fréquences estimées en regard des fréquences observées et calculer les coefficients de chaque modèle.

MLGEN (modèle linéaire général)

Cette procédure assure les calculs et les éditions d'un ajustement des moindres carrés sur un modèle linéaire comprenant un terme constant. Elle permet d'effectuer les régressions multiples, les analyses de variance et de covariance avec facteurs hiérarchisés (ou emboîtés), et interactions d'ordre un ou deux.

En option, il est possible d'obtenir le calcul et l'impression des coefficients de régression avec pour chacun, son écart-type et le test de sa nullité, valable dans le contexte où le terme aléatoire est supposé engendré par une loi de Laplace-Gauss.

La procédure réalise une analyse de la variance pour tester l'existence de l'effet de chacun des facteurs du modèle. Pour ces tests, le carré moyen de référence est, par défaut, le carré moyen résiduel. Trois type de somme des carrés des écarts sont possibles qui considèrent plusieurs conditions d'ajustement des facteurs. Le programme peut aussi tester un facteur par rapport à un autre facteur dont le carré moyen est pris comme référence en lieu et place du carré moyen résiduel.

En option, la procédure permet le calcul d'estimations ou de contrastes, ainsi que celui des moyennes ajustées des niveaux de facteurs du modèle. Ces moyennes ajustées peuvent faire l'objet de comparaisons deux à deux.

Les éditions sont nombreuses et pour la majorité d'entre elles, optionnelles.

L'édition de base, après une présentation sommaire des variables concourant à la définition du modèle, fournit le tableau de la régression multiple qui permet, grâce à un test de Fisher et au coefficient de corrélation multiple (Somme des carrés des écarts du modèle/somme totale des carrés des écarts), de juger de la pertinence et de l'intérêt du modèle linéaire général analysé et donne également l'estimation de la variance commune des résidus. Ensuite, pour chaque type de somme des carrés des écarts retenu, la procédure édite le tableau classique de l'analyse de variance. Pour chaque facteur, il est possible d'éditer le nombre de degrés de liberté, la somme des carrés des écarts, le carré moyen, la statistique de Fisher correspondante, ainsi que la probabilité critique qui lui est associée. Cette statistique qui prend, comme carré moyen de référence, le carré moyen de la résiduelle, considère que chaque facteur est à effet fixe.

De façon optionnelle, on peut obtenir l'édition de la matrice X'X, de la matrice inverse généralisée, des solutions des équations normales du modèle, de la forme générale des fonctions estimables et enfin des fonctions liées au calcul, pour chaque facteur, de la somme des carrés des écarts dont on demandait le calcul.

Si une estimation ou un contraste ou une moyenne ajustée a été demandée, outre l'édition des résultats découlant directement de la demande, la procédure permet d'éditer les coefficients des fonctions assurant le calcul de ces entités.

Notice de référence

1. Présentation

1.1 Objet

Cette procédure effectue une analyse linéaire discriminante à deux groupes par la méthode classique de Fisher. La procédure permet en particulier:

- **de discriminer à partir de coordonnées factorielles**. On peut ainsi effectuer une discrimination sur des données nominales (après une analyse des correspondances multiples), sur des fréquences ou des données disjonctives (après une analyse des correspondances) ou sur tout ou partie des facteurs d'une analyse en composantes principales non normée (LCORR = NON).
- **de réaliser des "estimations bootstrap"** du biais et de la précision des principaux résultats de la discrimination : coefficients, probabilités individuelles de classement, pourcentages globaux de classement.
- **de modifier les coûts et les probabilités a priori** de classement dans les groupes.

Dans la suite on utilisera le vocabulaire suivant. Les **individus-de-base** sont les individus (ou lignes du tableau) utilisés pour calculer la fonction linéaire discriminante. On dit aussi qu'ils constituent *l'échantillon d'apprentissage* de la règle de décision. Ces individus ou lignes doivent être déclarés "actifs" dans la procédure SELEC.

Les **individus-tests** ne participent pas au calcul de la fonction discriminante. Cependant leur numéro de groupe est connu. Il pourra donc être comparé au numéro de groupe calculé par la fonction discriminante, et servira à l'évaluation de la règle de décision. Les individus tests sont déclarés "illustratifs" dans la procédure SELEC, ou définis dans DIS2G par tirage au hasard parmi les individus de base (paramètres LEVAL et PRCT).

Les **individus-anonymes** sont, par définition, les seuls dont le numéro de groupe n'est pas connu. Le programme est utilisé pour estimer le numéro du groupe auquel chaque individu anonyme appartient. Ces individus doivent être déclarés "illustratifs" dans SELEC, et de plus doivent posséder le code 0 (correspondant à "donnée manquante") pour la variable de groupe.
Dans de nombreux cas, on utilisera la possibilité de "filtrage" dans SELEC (LSELI=FILT) pour déclarer les individus illustratifs sachant qu'ils sont codés 0 pour la variable de groupe.

La procédure peut créer deux fichiers. D'une part un fichier formaté, de type "NCOEF" qui contient essentiellement les coefficients de la fonction discriminante, les coefficients de la régression équivalente et leurs écarts-types, dans l'ordre des variables intervenant dans l'analyse.

Si les variables sont nominales (analyse après une correspondance multiple), il s'agit des coefficients attribués aux modalités des variables. Ce fichier est utilisé en particulier pour le passage de résultats entre les procédures DIS2G et SCORE pour l'étude d'une fonction de score.

Le second fichier est de type "NGRO" et contient la partition des individus dans les 4 classes suivantes:

Classe 1	Groupe 1	et	bien classé
Classe 2	Groupe 1	et	mal classé
Classe 3	Groupe 2	et	mal classé
Classe 4	Groupe 2	et	bien classé.

Ce fichier ne peut être fabriqué que si tous les individus de l'échantillon ont participé à la construction de la fonction discriminante. L'archivage par ESCAL de ces résultats permet l'analyse ultérieure de la discrimination.

1.2 Editions

La procédure permet d'imprimer au préalable les statistiques descriptives sur les variables du modèle dans chacun des 2 groupes.

On présente ensuite les résultats de l'analyse discriminante: tableaux de classement, fonction discriminante, résultats de la "régression équivalente", édition des affectations des individus. Dans le cas de la discrimination sur coordonnées factorielles, la règle de décision est finalement exprimée en fonction des variables ou modalités d'origine. Les résultats de la "régression équivalente" sont donnés à titre indicatif ; en effet les hypothèses classiques de "normalité" n'ont pas de sens ici.

Si une validation "bootstrap" est demandée, les résultats de la discrimination sont réédités avec les estimations "bootstrap". En particulier, le biais et la précision des classements globaux sont édités en vis-à-vis des classements directs. Pour les individus-anonymes, la procédure calcule leur probabilité bootstrap d'affectation.

Si une évaluation sur des individus tests est demandée, la procédure réédite les résultats de la discrimination relatifs à ces individus. Si l'affectation d'individus anonymes est demandée, seule l'édition des affectations est fournie.

1.3 Paramètres

Les paramètres de la procédure se divisent en trois catégories:

Les paramètres de fonctionnement indiquent le type des données sur lesquelles s'effectuent les calculs, soit des coordonnées factorielles soit les données brutes (LVEC), et fixent les probabilités a priori et les coûts de classification éventuels (PROB1, COUT1).

Les paramètres d'évaluation de la discrimination permettent de gérer le traitement des individus tests et des individus anonymes (LEVAL, PRCT). Enfin la procédure de ré-échantillonnage du bootstrap fournit une autre forme de validation (LBOOT).

Les paramètres d'édition (LEDIV et LEDIN) gèrent les éditions des statistiques sur les variables et des affectations des individus dans les groupes.

1.4 Après les paramètres

A la suite des paramètres se trouvent définis les modèles de discrimination à ajuster. Chaque modèle correspond à une commande. La liste des modèles doit se terminer par le mot-clé FIN.

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que tous les paramètres prennent leur valeur par défaut, on codera le mot-clef NOPAR à la place de la liste des paramètres repérés par (3).

(1) PROC DIS2G Discrimination linéaire à deux groupes

(2) *titre donné à l'analyse discriminante*

(3) LVEC (0 ou NON) utilisation de coordonnées factorielles.
- 0 ou NON : calcul sur données brutes.
- 1 ou OUI : calcul sur coordonnées factorielles.

LEDIN (0 ou NON) édition des affectations dans les groupes.
- 0 ou NON : pas d'édition.
- 1 ou OUI : édition des affectations.

LEDIV (0 ou NON) édition des statistiques sur les variables.
- 0 ou NON : pas d'édition.
- 1 ou STAT : statistiques par groupe.
- 2 ou CORR : édition STAT + corrélations.
- 3 ou COVA : édition CORR + covariances.

LEVAL (0 ou NON) évaluations selon les types d'individus.
- 0 ou NON : pas d'évaluation.
- 1 ou TEST : évaluation sur individus-tests.
- 2 ou ANO : affectation d'individus-anonymes

PRCT (0 ou NON) pourcentage d'individus-tests à tirer au hasard.
LBOOT (0 ou NON) validation par le bootstrap.
PROB1 (50 ou UNIF) probabilité a priori de classification dans groupe 1.
COUT1 (50 ou UNIF) coût de classification dans le groupe 1.

(4) *définition des modèles à ajuster*.
FIN fin de la liste des modèles.

3. Présentation détaillée des paramètres

LVEC **utilisation de coordonnées factorielles.**

- *valeurs possibles* : 0 ou NON (calcul sur données brutes continues)
  1 ou OUI (calcul sur coordonnées factorielles)
- *valeur par défaut* : 0 ou NON

Si LVEC = 0 le calcul s'effectue directement sur les variables, qui doivent être de type continu (à l'exception de la variable de groupe). Si LVEC=1 le calcul est réalisé sur les coordonnées factorielles lues sur le fichier NGUS d'une analyse précédente. Dans ce cas on vérifiera avec soin la cohérence des statuts des individus (actifs et illustratifs) pour l'analyse factorielle et pour l'analyse discriminante. L'analyse factorielle sera une analyse des correspondances multiples (CORMU), une analyse en composantes principales (COPRI) ou une analyse des correspondances (CORBI).

Le calcul des coefficients des variables d'origine est prévu dans le programme DIS2G (sauf dans le cas d'une analyse en composantes principales *normée* : cas LCORR=OUI de COPRI).

LEDIN **édition des affectations des individus dans les groupes.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des affectations)
- *valeur par défaut* : 0 ou NON

Chaque individu est affecté au groupe le plus probable. On obtient une ligne d'édition par individu, précisant également la probabilité d'appartenance au groupe.

LEDIV **édition des statistiques sur les variables**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou STAT (édition des statistiques classiques)
  2 ou CORR (édition STAT et corrélations)
  3 ou COVA (édition CORR et covariances)
- *valeur par défaut* : 0 ou NON

L'édition la plus courte (LEDIV=STAT) fournit les moyennes et écarts-types des variables dans chaque groupe et dans l'échantillon de base. Avec LEDIV=CORR, on ajoute l'édition des matrices de corrélation dans chaque groupe et globale. Avec LEDIV=COVA, on ajoute encore l'édition des matrices de covariances.

LEVAL **évaluation selon le type d'individus (test ou anonyme)**

- *valeurs possibles* : 0 ou NON (pas d'évaluation)
  1 ou TEST (validation sur individus-tests)
  2 ou ANO (affectation d'individus anonymes)
- *valeur par défaut* : 0 ou NON

Le paramètre LEVAL, joint au paramètre PRCT, permet de gérer différentes utilisations de la fonction discriminante sur les individus tests et les individus anonymes.

Si LEVAL = 1 (ou TEST) des individus tests sont utilisés pour évaluer la fonction discriminante en comparant leur groupe connu (1 ou 2) au groupe estimé par la règle de décision. Ces individus tests peuvent être introduits de deux façons. Si PRCT = 0, ils sont sélectionnés dès la procédure SELEC comme individus "illustratifs". Si PRCT est non nul, ils constituent le pourcentage PRCT d'individus tirés au hasard (par DIS2G) parmi les individus de base.

Si LEVAL = 2 (ou ANO) la procédure DIS2G est utilisée pour prévoir le groupe auquel appartient chaque individu anonyme. Sont considérés comme anonymes les individus codés 0 (donnée manquante) pour la variable de groupe, et déclarés "illustratifs" dans la procédure SELEC.
Si LEVAL = 0 (ou NON), et quelle que soit la valeur de PRCT, il n'y a aucun traitement d'individus anonymes ou tests.

PRCT **pourcentage d'individus-tests tirés au hasard**

- *valeurs possibles* : 1 à 99 (pourcentage de tirages au hasard)
- *valeur particulière*: 0 ou NON (pas de tirage)
- *valeur par défaut* : 0 ou NON

Ce paramètre n'est pris en compte que si LEVAL est non nul. Le pourcentage de tirage d'individus-tests est généralement compris entre 15 % et 30 %, selon la taille de l'échantillon de base. Ce pourcentage doit être un nombre entier.

La définition conjointe des paramètres LEVAL et PRCT permet les cas de figure suivants:

LEVAL = 0 (et PRCT quelconque) : il n'y a aucune évaluation des calculs sur des individus tests, ni sur des individus anonymes.

LEVAL = 1 et PRCT = 0 : il y a des *individus-tests* permettant d'évaluer la procédure en estimant (sans biais) les probabilités de classement. Sont individus-tests les individus illustratifs dans SELEC.

LEVAL = 1 et 1 < PRCT < 99 : analogue au cas précédant, mais les *individus-tests* sont ici tirés au hasard dans l'échantillon de base (individus déclarés actifs dans SELEC).

LEVAL = 2 et PRCT = 0 : la procédure est utilisée pour prévoir le groupe auquel appartient un *individu anonyme*. Est anonyme tout individu déclaré illustratif dans SELEC, et codé 0 pour la variable indicatrice de groupe.

LEVAL = 2 et 1 < PRCT < 99 : ce cas cumule les deux possibilités. Il y a prévision du groupe pour les *individus anonymes* (LEVAL=2) qui sont les individus déclarés illustratifs dans SELEC et codés 0. De plus il y a évaluation des probabilités de classement à partir d'*individus-tests*, qui sont tirés au hasard (PRCT non nul) parmi les individus de base (actifs dans SELEC).

LBOOT **validation par bootstrap**

- *valeurs possibles* : de 0 à 1000 (nombre de tirages bootstrap)
- *valeur particulière*: 0 ou NON (pas de re-échantillonnage bootstrap)
- *valeur par défaut* : 0 ou NON

En général, de bonnes estimations bootstrap sont atteintes à partir de LBOOT=50 tirages (la limite supérieure autorisée est LBOOT = 1000 tirages)

PROB1 **probabilité a priori de classement dans le groupe 1**

- *valeurs possibles* : de 1 à 99 (probabilité d'appartenir au groupe 1)
- *valeur particulière*: 50 ou UNIF (probabilités égales dans les 2 groupes)
- *valeur par défaut* : 50 ou UNIF

Il s'agit plus précisément d'un entier représentant une probabilité multipliée par 100. La valeur attribuée au groupe 2 vaut: 100 - PROB1.

COUT1 **coût a priori de classement dans le groupe 1**

- *valeurs possibles* : de 1 à 99 (coût de classification dans le groupe 1)
- *valeur particulière*: 50 ou UNIF (coûts égaux dans les 2 groupes)
- *valeur par défaut* 50 ou UNIF

La valeur par défaut (COUT1=UNIF) attribue le même coût a priori de classement dans les 2 groupes. La somme des coûts est arbitrairement fixée à 100. Le coût a priori de classement dans le groupe 2 vaut donc 100-COUT1.

4. Définition du modèle de discrimination

On définit, après les paramètres, le ou les modèles de discrimination. Un modèle s'écrit sous la forme (par exemple):

V12 = V1 + V2 + V8 + V10

ou encore:

V12 = V1 -- V4 + V10 + V14 -- V16

Dans cette écriture symbolique, V12 à gauche du signe égal, désigne la variable endogène du modèle ou variable indicatrice de groupe. Elle est obligatoirement nominale à 2 modalités. Elle est codée 1 et 2, avec éventuellement des codes 0; les 0 désignent des *individus anonymes* qui de plus doivent avoir été déclarés illustratifs dans SELEC.

A droite du signe égal, se trouvent les variables exogènes du modèle. Ces variables, toujours de type continu, seront soit toutes des variables d'origine (LVEC=0), soit toutes des facteurs lus sur un fichier NGUS (cas LVEC=1).

Dans le cas ou LVEC = 1 ou OUI, une écriture de la forme:

V3 = F1 + F3 + F5

V3, à gauche du signe égal, désigne la variable numéro 3 du fichier archive des données (c'est une variable nominale à 2 modalités). Le symbole F3 à droite du signe égal désigne ici, puisque LVEC=1, le facteur numéro 3 (lu sur le fichier NGUS). De même F1 et F5 désignent les facteurs 1 et 5.

Les variables à droite du signe égal ne peuvent être séparées que par des signes "+", ou décrites par une liste à l'aide du signe " --". Tous les blancs sont facultatifs.

Pour les variables continues, à droite du signe égal, les **données manquantes** (repérées par la valeur TEST de la procédure ARDON) sont traitées de la manière suivante. S'il s'agit d'un *individu de base* (donc déclaré actif dans SELEC), chaque valeur manquante est remplacée par la moyenne du groupe auquel l'individu appartient. Pour un *individu anonyme* (déclaré illustratif), chaque valeur manquante est remplacée par la moyenne générale calculée sur les individus actifs. Notons qu'il est souvent préférable d'éliminer les individus présentant des données manquantes.

5. Exemples de commande

5.1 Premier exemple

Ce premier exemple concerne une analyse discriminante effectuée sur des variables brutes. On fait figurer la procédure SELEC associée à la définition des paramètres de DIS2G.

```
----+----1----+----2----+----+----3----+----4----+----5----
    PROC SELEC
=== SELECTION DES VARIABLES ET STATUT DES INDIVIDUS ===

LZERO = NOREC, LSELI = FILT
NOMI ACT  18
CONT ACT  5--10
FIN

V18 < > 0

    PROC DIS2G
====  ANALYSE DISCRIMINANTE  ====

LEDIV = STAT, LEDIN = OUI, LEVAL = ANO, PRCT = 20, LBOOT = 50

V18 = V5 + V7--V9
V18 = V5 + V10
FIN
----+----1----+----2----+----+----3----+----4----+----5----
```

La procédure SELEC sélectionne la variable nominale V18 (à 2 modalités) et les variables continues V5 à V10. On notera que le statut actif ou illustratif des variables est indifférent pour DIS2G. Dans le but de prévoir le groupe inconnu des individus anonymes (codés 0 dans V18), on déclare ces individus illustratifs: le filtre (LSELI=FILT) retient en actifs les individus qui ne sont pas codés 0. On notera que le paramètre LZERO=NOREC empêche le recodage automatique des 0 de la variable V18 dans une catégorie *donnée manquante* qui sinon serait codée "3".

Les 2 premiers paramètres de DIS2G concernent les éditions. LEDIV=STAT commande l'impression des statistiques des variables dans chaque groupe. Les matrices de corrélations et de covariances ne seront pas imprimées. Les individus n'étant pas trop nombreux, on demande l'impression de l'affectation prévue pour *tous* les individus (LEDIN=OUI).

LEVAL=ANO commande la prévision du groupe d'appartenance de tout individu anonyme (illustratif pour SELEC et codé 0 dans V18).

Le paramètre PRCT=20 déclenche une procédure de validation par *individus tests* : 20 % des individus actifs pour SELEC sont tirés au hasard et considérés comme *individus tests*. Le reste constitue l'*échantillon d'apprentissage*, ensemble des individus de base utilisés pour calculer la fonction discriminante.

Enfin LBOOT=50 commande les estimations bootstrap des principaux éléments de l'analyse discriminante à partir de 50 opérations de re-échantillonnage. Pour tous les

calculs, les probabilités a priori d'appartenance aux groupes et les coûts de mauvaise classification sont considérés uniformes par défaut. On trouve ensuite l'écriture de 2 modèles ayant la même variable de groupe V18. La liste des modèles s'achève avec le mot-clé FIN.

5.2 Deuxième exemple

Cet exemple montre comment effectuer une analyse discriminante sur données qualitatives, en passant par le moyen d'une analyse des correspondances multiples intermédiaire.

```
----+----1----+----2----+----+----3----+----4----+----5----
PROC SELEC
=== SELECTION POUR LES CORRESP. MULTIPLES ===

NOPAR
NOMI ACT  19--25
NOMI ILL  4
FIN

   PROC CORMU
==== ANALYSE DES CORRESPONDANCES MULTIPLES ====
NAXE = 5

    PROC DIS2G
==== ANALYSE DISCRIMINANTE  ====

LVEC = OUI, LBOOT = 20

V4 = F1--F5
FIN
----+----1----+----2----+----+----3----+----4----+----5----
```

On effectue une analyse des correspondances multiples sur les variables nominales actives numéros 19 à 25. La variable nominale à 2 modalités V4, qui sera la variable de groupe de l'analyse discriminante, est introduite en variable illustrative. On notera que, par défaut dans SELEC, les données manquantes sont automatiquement recodées (LZERO=REC) pour effectuer l'analyse factorielle. On suppose ici qu'il n'y a pas de code 0 (données manquantes) dans la variable de groupe V4.

L'analyse discriminante met en jeu les coordonnées factorielles (LVEC=OUI). Le modèle estime la catégorie de la variable V4 à l'aide des 5 premiers axes factoriels. Il n'y a ni individus tests, ni individus anonymes, ni validation par bootstrap.

5.3 Troisième exemple

```
----+----1----+----2----+----+----3----+----4----+----5----
    PROC SELEC
=== SELECTION POUR LES CORRESP. MULTIPLES ===
LSELI = FILT, LZERO=REC
NOMI ACT  19--25
NOMI ILL  5
FIN
```

```
V5 < > 0

    PROC CORMU
==== ANALYSE DES CORRESPONDANCES MULTIPLES ====
NAXE = 10

PROC DIS2G
==== ANALYSE DISCRIMINANTE  ====
LVEC = OUI, LEVAL = ANO

V5 = F1--F5
V5 = F2--F5 F7 F8
FIN
----+----1----+----2----+----+----3----+----4----+----5----
```

On effectue l'analyse des correspondances multiples sur les mêmes variables nominales actives numéros 19 à 25. La variable de groupe à 2 modalités est ici V5. Les individus ayant le code 0 pour la variables de groupe V5 sont mis en illustratifs pour CORMU à l'aide d'un, filtre dans SELEC. Pour qu'ils soient des individus anonymes dans DIS2G, il faut qu'ils conservent leur code 0. On code donc LZERO=NOREC dans SELEC.

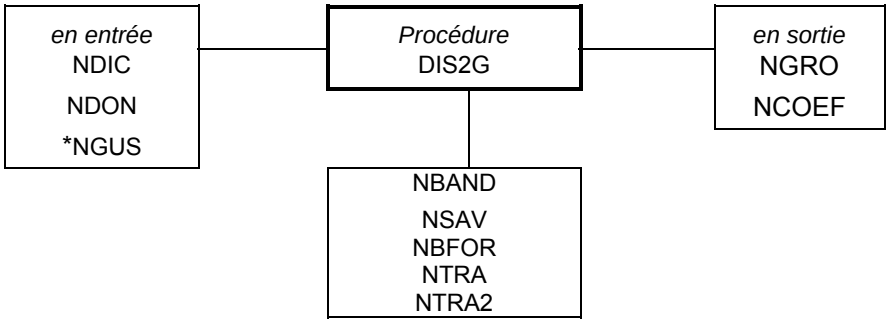
Les paramètres de DIS2G spécifient que l'on travaille sur coordonnées factorielles (LVEC=OUI), et que l'on utilisera des individus anonymes (LEVAL=ANO). Les autres paramètres prennent leurs valeurs par défaut.

6. Fichiers nécessaires à l'exécution

- en lecture NDIC (dictionnaire utile)
  NDON (données utiles)
  NGUS (coordonnées factorielles si LVEC=1)

- de travail NSAV, NBAND, NBFOR, NTRA, NTRA2

- de sortie NGRO (partition des individus bien et mal classés)
  NCOEF (coefficients de la fonction discriminante)



En entrée le fichier NGUS n'est nécessaire que si LVEC=1, c'est-à-dire si la fonction discriminante est calculée sur des axes factoriels.

En sortie le fichier formaté NCOEF contient les coefficients de la fonction discriminante. NGRO contient la partition des individus en bien et mal classés dans chaque groupe (seulement si tous les individus participent au calcul de la fonction).

Le fichier NCOEF est utilisé par SCORE. Le fichier NGRO est utilisable par DECLA pour caractériser les groupes bien et mal classés.

1. Présentation

1.1 Objet

Cette procédure réalise la sélection des "meilleurs" ajustements pour une régression ou pour une analyse linéaire discriminante à deux groupes. Le critère de sélection peut être le "R2", le "R2 ajusté" ou le "Cp" de Mallows.

L'algorithme de sélection est une transcription de l'algorithme "leaps and bounds" de Furnival et Wilson. (Technometrics, 174, Vol.16, pp.499-511).

1.2 Editions

Soit n le nombre de meilleurs ajustements demandés et p le nombre des variables "explicatives" du modèle. La procédure édite les n meilleurs ajustements, pour toutes les tailles de modèles, de 1 à p - 1 variables (l'ajustement avec les p variables est unique). Pour chaque régression, la procédure fournit la valeur du critère (le R2 , le R2 ajusté ou le Cp), le F de Fisher associé au R2, la probabilité critique associée à ce F, et la valeur-test correspondante.

La liste des variables du modèle est ensuite éditée avec les coefficients estimés, les tests de nullité, la probabilité critique et la valeur-test associée. Enfin un graphique représentant l'évolution du critère en fonction du nombre de variables dans les modèles fournit une synthèse rapide des sélections.

Dans le cas d'une analyse discriminante, on édite les résultats de la régression "équivalente" à l'analyse discriminante à 2 groupes. On utilisera cependant cette présentation à titre indicatif seulement, puisque les hypothèses classiques de normalité ne sont pas admissibles dans ce cas.

Pour le critère du R2, toutes les sélections imprimées sont optimales. Pour les deux autres critères, les sélections ne sont pas toujours optimales (le R2 ajusté et le Cp de Mallows varient de façon non monotone en fonction du nombre de variables). On repère qu'une sélection n'est pas optimale si la procédure n'édite pas les coefficients des variables (seuls les noms des variables et la valeur du critère sont imprimés). Dans ce cas l'ajustement sélectionné, s'il n'est pas optimal pour le critère, est cependant meilleur que les ajustements qui n'ont pas été calculés.

1.3 Paramètres

Il y a trois paramètres pour la procédure:

- **Le paramètre de fonctionnement** (LMODE) indique s'il s'agit d'une régression ou d'une analyse discriminante.

• **Le nombre de meilleurs ajustements demandé** (NMREG) pour chaque taille de modèle.

• **Le choix du critère de sélection** (LCRIT)

1.4 Après les paramètres

A la suite des paramètres on introduit les modèles (de régression ou de discrimination) à ajuster. Chaque modèle correspond à une instruction de commande. La liste des modèles doit se terminer par le mot-clé FIN.

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que tous les paramètres prennent leur valeur par défaut, on codera le mot-clé NOPAR à la place de la liste des paramètres repérée par (3).

(1) PROC FUWIL Sélections optimales de variables

(2) *titre donné à l'analyse*

(3) LMODE (1 ou REG) type d'analyse (régression ou discrimination).

1 ou REG : régression multiple.
2 ou DISC : analyse discriminante à 2 groupes.

NMREG (3) nombre de meilleurs ajustements par taille de modèle.

LCRIT (1 ou R2) critère de sélection des ajustements.

1 ou R2 : coefficient de corrélation multiple
2 ou R2AJ : corrélation multiple ajusté.
3 ou CP : coefficient Cp de Mallows.

(4) *définition du ou des modèles à ajuster.*
FIN fin de la liste des modèles.

3. Présentation détaillée des paramètres

LMODE **type d'analyse**

- *valeurs possibles* : 1 ou REG (régression)
  2 ou DISC (discriminante à 2 groupes)
- *valeur par défaut* : 1 ou REG

NMREG **nombre de meilleurs ajustements par taille de modèle**

- *valeurs possibles* : supérieures à 1
- *valeur par défaut* : 3

On donnera à ce paramètre une valeur choisie en général entre 1 et 10. La valeur par défaut (3) fournit le plus souvent des éditions suffisantes.

LCRIT **critère de sélection des ajustements**

- *valeurs possibles* : 1 ou R2 (carré de la corrélation multiple)
  2 ou R2AJ (corrélation multiple ajusté)
  3 ou Cp (Cp de Mallows)
- *valeur par défaut* : 1 ou R2

On notera que les paramètres de la procédure (en particulier le critère de sélection) sont les mêmes pour tous les modèles demandés.

4. Les modèles à ajuster

Après avoir défini les paramètres, on introduit les modèles pour lesquels on recherche les variables optimales. Un modèle s'écrit sous la forme d'une "équation" du type suivant (à titre d'exemple):

4.1 Exemples

- V3 = V1 + V2 + V6 + V9 + V12 + V14
- V3 = V1 + V10 + V14 -- V17

On notera la contrainte suivante : le modèle doit avoir au moins trois variables explicatives, et ne doit pas contenir plus de 50 variables.

Dans cette écriture symbolique, V3 à gauche du signe égal désigne la variable numéro 3 comme variable "endogène" du modèle (notée habituellement y, et dite variable "à expliquer"). A droite du signe égal se trouvent les variables "exogènes" ou "explicatives" du modèle. Le "terme constant" sera automatiquement présent dans tous les modèles. Les variables doivent être séparées par des signes "+", ou engendrées par une liste à l'aide du signe "--". Les blancs sont facultatifs.

Dans le cas de cette procédure FUWIL, les variables exogènes (à droite du signe égal) doivent toujours être de type continu. (Cette restriction n'existe pas dans la procédure plus classique VAREG).

Ceci implique que si l'on veut faire intervenir une variable explicative nominale (cas d'une analyse de variance par exemple), il est nécessaire de procéder au préalable à son codage disjonctif. Il faut prendre garde alors à ne créer que k-1 colonnes pour une variable à k modalités (pour que le modèle soit de plein rang).

De plus, la variable endogène "y" doit être de type continu dans le cas de la régression, et de type nominal (à 2 modalités) dans le cas de la discrimination.

Dans la procédure FUWIL, les *données manquantes* sont traitées de la manière suivante. Une donnée manquante pour la variable endogène "y" n'est pas admise et provoque l'arrêt de la procédure. Pour les variables "explicatives", qui sont toujours de type continu ici, toute donnée manquante est remplacée par la moyenne des valeurs connues de la variable. Il sera souvent préférable d'éliminer les lignes comportant des données manquantes avant d'exécuter FUWIL.

5. Exemples de commande

5.1 Premier exemple

```
----+----1----+----2----+----+----3----+----4----+----5----
    PROC SELEC
=== SELECTION DES VARIABLES  ===
NOPAR

CONT ACT  1--6 9--20
FIN
    PROC FUWIL
==== SELECTION OPTIMALE DE VARIABLES POUR LA REGRESSION  ====

NOPAR

V10 = V5 + V6 + V11--V15
V10 = V6 + V12--V18
FIN
----+----1----+----2----+----+----3----+----4----+----5----
```

Les variables V1 à V6 et V9 à V20 sont continues. Leur statut (illustratif ou actif) ne sera pas pris en compte par la procédure FUWIL.

La procédure FUWIL va effectuer des régressions (LMODE=REG par défaut) et fournira les 3 meilleurs ajustements pour chaque taille de modèle (NMREG=3 par défaut). Le critère de sélection sera le R2 usuel (LCRIT=R2 par défaut).

Les 2 lignes suivantes définissent les modèles utilisés. Le mot clé FIN termine nécessairement la liste des modèles.

5.2 Second exemple

```
----+----1----+----2----+----+----3----+----4----+----5----
    PROC SELEC
=== SELECTION DES VARIABLES  ===

NOPAR

NOMI ACT  7 8
CONT ACT  1--6  9--20
FIN

    PROC FUWIL
==== SELECTION OPTIMALE POUR LA DISCRIMINANTE A 2 GROUPES  ====

LMODE = DISC

V7 = V2 + V5 + V12--V19
V8 = V2 + V5 + V11 + V14 + V19
FIN
----+----1----+----2----+----+----3----+----4----+----5----
```

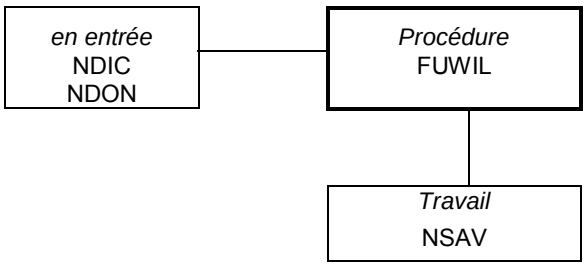
V7 et V8 sont des variables nominales à 2 modalités. Elles sont les variables endogènes "y" des modèles. Il est très important de noter que ces variables ne doivent pas contenir de *données manquantes.* La première donnée manquante rencontrée stopperait la procédure. Les variables V1 à V6 et V9 à V20 sont déclarées continues. Le statut (illustratif ou actif) de ces variables ne sera pas pris en compte.

La procédure FUWIL va réaliser des analyses discriminantes (LMODE=DISC) et fournira les 3 meilleurs ajustements pour chaque taille de modèle (NMREG=3 par défaut). Le critère de sélection des meilleurs ajustements sera le R2 (LCRIT=R2 par défaut).

6. Fichiers nécessaires à l'exécution

- en lecture NDIC (dictionnaire utile)
  NDON (données utiles)

- de travail NSAV



1. Présentation

L'écriture de la procédure LOGLI est une adaptation du programme d'analyse de modèles log-linéaires réalisé par Noboru Ohsumi, de l'Institut de Mathématique Statistique de Tokyo.

1.1 Objet

L'analyse log-linéaire est une méthode de modélisation des tableaux de contingence à plusieurs entrées. A partir d'un tableau, la procédure LOGLI permet d'ajuster plusieurs modèles log-linéaires *hiérarchiques* (une interaction n'est spécifiée qu'en présence des termes principaux). Deux types de données sont acceptés en entrée : des données "individus x variables" d'une part et des tables de contingence d'autre part. Dans le cas de données individuelles, on peut utiliser une variable poids (définie dans la procédure SELEC) pour calculer les fréquences. Plusieurs possibilités sont offertes pour sélectionner les modèles à ajuster :

- modèle(s) défini(s) par l'utilisateur (le nombre de variables dans chaque modèle est limité à 7).
- tous les modèles possibles (dans ce cas, le nombre de variables est limité à 4).
- sélection d'un modèle par une méthode combinatoire fondée sur la minimisation du critère de l'information d'AKAIKE (AIC) ainsi que sur la décomposition de la statistique du CHI2 du rapport de vraisemblance (le nombre de variables est limité à 4).
- sélection d'un modèle par une méthode "pas à pas" ascendante fondée sur la minimisation du critère de l'information d'AKAIKE (le nombre de variables est limité à 4).

En entrée, la procédure utilise les fichiers NDIC et NDON créés par SELEC.

1.2 Editions

L'édition commence par l'écriture du modèle à estimer, suivie éventuellement par les tris à plat des variables nominales du modèle. Apparaissent ensuite, à la demande de l'utilisateur, le tableau de contingence avec les effectifs estimés par le modèle, les informations sur la convergence de l'algorithme et le nombre d'itérations de l'ajustement. Une synthèse des statistiques d'ajustement est imprimée (estimation du CHI-2 de Pearson, du CHI2 du rapport de vraisemblance ainsi que leurs degrés de liberté et les probabilités associées). Puis l'étape fournit l'identification des coefficients des modalités des facteurs et des interactions éventuelles. Il est également possible d'obtenir la matrice de structure du modèle. Dans le cas d'une sélection automatique d'un modèle, le tableau récapitulatif de tous les modèles traités est imprimé.

1.3 Paramètres

La procédure comporte sept paramètres ayant chacun une valeur par défaut. On les regroupe en deux types :

Les paramètres de calcul qui fixent le mode de calcul : la méthode choisie (LMOD), le type des données (LTAB), la génération du dictionnaire des variables d'un tableau de contingence (LDICO), le nombre maximal d'itérations (MAXIT), le seuil de convergence (MAX), la valeur choisie pour remplacer les fréquences nulles ou négatives (VADD).

Le paramètre d'édition (LEDIT) qui détermine la sortie, pour chaque modèle, du tri à plat des variables du modèle, des tableaux de contingence avec les effectifs estimés ainsi que des informations sur la convergence de l'algorithme avec le nombre d'itérations de l'ajustement. Le tableau des variations successives de la statistique AIC fondée sur le rapport de vraisemblance peut être imprimé dans le cas de sélection de modèles.

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que tous les paramètres prennent leur valeur par défaut, on codera le mot-clé NOPAR à la place de la liste des paramètres repérée par (3).

(1) PROC LOGLI

(2) titre de la procédure

(3) LMOD (1 ou LIST) sélection des modèles à ajuster
- 1 ou LIST : modèle(s) défini(s) par l'utilisateur
- 2 ou TOT : tous les modèles possibles
- 3 ou COMBI : modèle sélectionné par une méthode combinatoire
- 4 ou PAP : modèle sélectionné par une méthode "pas à pas"

LEDIT (2 ou STAND) choix des éditions de résultats
- 1 ou MIN : édition minimale
- 2 ou STAND : édition standard
- 3 ou TOT : édition complète
- 4 ou DENSE : édition condensée

MAXIT (10) nombre maximum d'itérations de convergence

LTAB (1) type du tableau des données en entrée
- 1 : tableau individus x variables
- 2 : tableau de contingence

LDICO (0 ou NON) introduction des libellés des variables du tableau de contingence en entrée (si LTAB = 2)
- 0 ou NON : pas d'introduction
- 1 ou OUI : définie par l'utilisateur

MAX (0.001) seuil de convergence

VADD (0.5) valeur remplaçant toute fréquence nulle ou négative

(4) (Si LTAB = 2) définition des lignes et des colonnes de fréquence

(5) (Si LDICO = 1) dictionnaire des libellés des variables

(6) (Si LMOD = 1) liste du ou des modèles

FIN fin de la liste des modèles

3. Présentation détaillée des paramètres

LMOD **choix de la méthode**

- *valeurs possibles* : 1 ou LIST (modèle(s) défini(s) par l'utilisateur)
  2 ou TOT (tous les modèles possibles)
  3 ou COMBI (méthode combinatoire)
  4 ou PAP (méthode "pas à pas")
- *valeur par défaut* : 1 ou LIST (modèle(s) défini(s) par l'utilisateur)

  - Si LMOD = 1 ou LIST, le modèle est défini par l'utilisateur qui ne peut en spécifier que 10 au plus. Le nombre de variables dans chaque modèle est limité à 7.

  - Si LMOD = 2 ou TOT, tous les modèles possibles sont édités (dans ce cas, le nombre de variables est limité à 4).

  - Si LMOD = 3 ou COMBI, le modèle est sélectionné par une méthode combinatoire fondée sur la minimisation du critère de l'information d'AKAIKE (AIC) ainsi que sur la décomposition de la statistique du CHI2 du rapport de vraisemblance. Le nombre de variables est limité à 4.

  - Enfin pour LMOD = 4 ou PAP, la sélection d'un modèle se fait par une méthode "pas à pas" ascendante fondée sur la minimisation du critère de l'information d'AKAIKE (AIC). Le nombre de variables est limité à 4.

LEDIT **choix de l'option pour l'édition des résultats**

- *valeurs possibles* : 1 ou MIN (édition minimale)
  2 ou STAND (édition standard)
  3 ou TOT (édition complète)
  4 ou DENSE (édition condensée)
- *valeur par défaut* : 2 ou STAND (édition standard des résultats)

  - Si LEDIT = 1 ou MIN, on n'édite que le modèle traité avec la table des statistiques d'ajustement.

  - Si LEDIT = 2 ou STAND, on obtient l'édition pour LEDIT = 1, complétée par le tri à plat des variables du modèle, le tableau de contingence avec les effectifs estimés, les informations sur la convergence et les coefficients estimés du modèle. Les variations successives de la statistique AIC fondée sur le rapport de vraisemblance est imprimée dans le cas de sélection de modèles (LMOD = 3 ou 4).

- Si LEDIT = 3 ou TOT, l'édition obtenue par LEDIT = 2 est assortie de la matrice de structure du modèle.

- Si LEDIT = 4 ou DENSE, on obtient les mêmes éditions que dans le cas LEDIT=2, mais les informations (effectifs observés et estimés) sont rassemblées dans un seul tableau.

MAXIT **nombre maximum d'itérations**

- *valeurs possibles* : de 1 à 999999
- *valeur par défaut* : 10

Il s'agit du nombre maximum d'itérations de l'algorithme d'ajustement d'un modèle.

LTAB **type du tableau des données en entrée**

- *valeurs possibles* : 1 ("individus x variables")
  2 (tableau de contingence)
- *valeur par défaut* : 1 ("individus x variables")

- Si LTAB = 1, le tableau des données est de type "individus x variables". Dans ce cas les fréquences calculées utilisent le poids éventuellement défini dans la procédure SELEC.

- Si LTAB = 2, le tableau des données est un tableau de contingence (à entrées multiples). Dans ce cas, la description du tableau de contingence s'effectue à l'aide des deux commandes MLIG et MCOL qui suivent immédiatement la liste des paramètres.

LDICO **introduction des libellés des variables du tableau de contingence en entrée (si LTAB = 2)**

- *valeurs possibles* : 0 ou NON (pas d'introduction)
  1 ou OUI (introduction définie par l'utilisateur)
- *valeur par défaut* : 0 ou NON (pas d'introduction)

Ce paramètre a un sens dans le cas où les données sont entrées sous forme d'un tableau de contingence (LTAB = 2). Dans le cas où le tableau de données est de type "individus x variables", ce paramètre est ignoré.

- Si LDICO = 0 ou NON, les libellés des variables du tableau de contingence seront générés automatiquement par la procédure LOGLI.
- Si LDICO = 1 ou OUI, les libellés des variables du tableau de contingence sont introduits par l'utilisateur à la suite des deux commandes MLIG et MCOL qui suivent immédiatement la liste des paramètres (voir l'exemple 3).

MAX **seuil de convergence**

- *valeurs possibles* : de 10^{-7} à 0.1
- *valeur par défaut* : 0.001

Cette valeur indique le seuil en dessous duquel l'algorithme d'ajustement s'arrête.

VADD **valeur remplaçant toute fréquence nulle**

- *valeurs possibles* : toute valeur positive de 0.001 à 1.0
- *valeur par défaut* : 0.5

Les effectifs nuls ne sont pas autorisés dans tout tableau de contingence traité, ils sont donc remplacés par la valeur du paramètre VADD. Il est à noter que toute fréquence négative éventuelle sera aussi remplacée par cette valeur VADD.

4. Définition des lignes et colonnes du tableau de contingence (si°LTAB°=°2)

Dans le cas LTAB = 2, le tableau de contingence fourni par l'utilisateur doit être décrit par les deux commandes suivantes.

MLIG = liste des nombres de modalités des variables en ligne ; les nombres sont séparés par un espace ou une virgule

MCOL = liste des nombres de modalités des variables en colonne ; les nombres sont séparés par un espace ou une virgule

Ainsi un tableau à deux entrées (une variable en ligne avec 5 modalités et une variable en colonne à 22 modalités) sera décrit par :

MLIG = 5
MCOL = 22

Mais les lignes comme les colonnes d'un tableau de contingence peuvent être issues du croisement de plusieurs variables nominales selon un ordre que l'utilisateur doit préciser.
Considérons par exemple, un tableau de contingence à 6 lignes et à 8 colonnes qui est le tableau croisé de 4 variables (2 en lignes et 2 en colonnes).
Les lignes sont issues du croisement de deux variables V1 et V2 ayant respectivement 2 et 3 modalités (A1, A2 pour V1 ; B1, B2 et B3 pour V2) et sont structurées de la manière suivante :

A1 B1
A1 B2
A1 B3
A2 B1
A2 B2
A2 B3

Cette structure croisée en ligne sera décrite par MLIG = 2 3. Dans la liste, on remarque que c'est la deuxième variable qui varie en premier. Ainsi la deuxième variable V2 à 3 modalités varie avant la première variable V1 à 2 modalités.

Pour le même tableau, il faut également décrire la structure en colonne. Si elle provient du croisement de 2 variables avec respectivement 2 et 4 modalités (X1, X2 pour la première variable ; Y1, Y2, Y3 et Y4 pour la seconde), les 8 colonnes doivent être indiquées par MCOL = 2 4 si l'ordre de variation est le suivant :

X1	X1	X1	X1	X2	X2	X2	X2
Y1	Y2	Y3	Y4	Y1	Y2	Y3	Y4

Remarque :
Le nombre de termes dans la liste de MLIG doit être égal au nombre de variables composant les lignes (2 dans le deuxième exemple). Le nombre de termes pour MCOL doit être égal au nombre de variables composant les colonnes (2 dans le deuxième exemple).
Le produit des termes de MLIG doit être égal au nombre de lignes du tableau (2*3 = 6 dans le deuxième exemple). De même, le produit des termes de MCOL doit être égal au nombre de colonnes du tableau (2*4 = 8 dans le deuxième exemple).

5. Définition du dictionnaire des variables du tableau de contingence (si°LTAB°=°2°et°LDICO°=°1)

Si LTAB = 2, l'utilisateur peut spécifier les libellés des variables en choisissant l'option LDICO = 1. Cette option permet d'obtenir des listages de résultats plus faciles à lire. Le format d'enregistrement des variables nominales composant le tableau de contingence est le format étendu (type large) décrit dans la procédure ARDIC. De plus, l'ordre d'introduction des libellés dépend des instructions définissant les variables du modèle (voir l'exemple 3).

6. Définition des modèles log-linéaires (si°LMOD°=°1)

6.1 Cas individus x variables (LTAB = 1)

Les variables mentionnées dans un modèle sont sélectionnées parmi l'ensemble des variables retenues dans l'étape SELEC qui précède. Mais, même si certaines variables ne sont pas nommées dans le modèle, elles participent à la définition du tableau de contingence analysé.

Par exemple si les variables V1, V2, V3 sont déclarées dans SELEC, le modèle V1 + V2, sélectionné par LOGLI ou défini par l'utilisateur, n'est pas indépendant de V3 et s'écrit dans les listages de la procédure LOGLI sous la forme :
V1 + V2 / V3 (lire V1 + V2 sachant V3). Ceci signifie que l'ajustement porte sur le logarithme des effectifs du tableau de contingence croisant V1, V2 et V3 à l'aide uniquement des variables V1 et V2 présentes dans le modèle.

Dans le cas LMOD = 1, on définit le ou les modèles log-linéaires hiérarchiques que l'on veut étudier. Un modèle s'écrit sous la forme suivante :

*V1 +...+ Vi +Vj + Vk +...+ Vn + Vi*Vj +...+ Vi*Vj*Vk +...*

Les variables V1, Vi, Vj, Vk, Vn sont les variables exogènes du modèle. Ce sont nécessairement des variables nominales en nombre au plus égal à 7, auxquelles LOGLI adjoint systématiquement une constante dans tout modèle traité. Quant à la variable endogène du modèle, il s'agit du logarithme des effectifs composant le tableau de contingence des variables du modèle. Ce tableau est construit automatiquement par LOGLI dans le cas de données de type "individus x variables".

Dans l'écriture d'un modèle, l'usage du symbole de suite "--" est licite pour décrire une suite de variables. Le caractère "*", placé entre 2 variables, annonce une interaction :

Vi*Vj : interaction d'ordre 2 (entre les variables Vi et Vj)
Vi*Vj*Vk : interaction d'ordre 3. Le signe d'interaction ne peut pas être utilisé conjointement avec le symbole de suite "--".

La procédure LOGLI n'accepte que des modèles hiérarchiques : il est interdit de spécifier une interaction sans avoir mentionné dans le modèle les termes principaux associés.

Par exemple, si l'on désire prendre en compte une interaction d'ordre 2 entre les variables V1 et V2, les termes V1 et V2 doivent obligatoirement figurer dans le modèle.

Ecritures incorrectes : *V1 + V1*V2,*
*V2 + V1*V2*
*V1*V2*

Ecriture correcte : *V1 + V2 + V1*V2*

Attention :
Il est interdit d'insérer une ligne de commentaires en lieu et place d'un modèle. En revanche, il est possible de mettre un commentaire à droite d'un modèle.

Toutes les variables spécifiées dans un modèle doivent avoir été retenues lors de la précédente étape SELEC. Après l'écriture des modèles à ajuster, on doit introduire l'instruction FIN.

6.2 Cas d'un tableau de contingence (LTAB = 2)

Dans le cas d'un tableau de contingence introduit par l'utilisateur (LTAB=2), la syntaxe des variables constituant un tableau de contingence diffère de celle utilisée précédemment. Les n variables en ligne d'un tableau croisé sont mentionnées dans tout modèle par L1, L2, ..., Ln. Et les p variables en colonne sont mentionnées par C1, C2, ..., Cp. Hormis cette différence de syntaxe, les règles d'écriture des modèles sont les mêmes que les précédentes (voir l'exemple 3).

Attention :
Dans le cas d'un tableau de contingence (LTAB = 2), les colonnes du tableau sont considérées comme des *fréquences* et déclarées comme telles dans SELEC. Et les lignes du tableau sont des individus pour SELEC.

Dans le cas de données "individus x variables" (LTAB = 1), les variables doivent être *nominales* et déclarées comme telles dans SELEC.

7. Exemples de commande

Trois exemples concernant l'utilisation de la procédure LOGLI sont présentés ci-dessous. Dans chaque cas, la procédure SELEC qui doit précéder LOGLI permet de créer les fichiers NDIC et NDON nécessaires à la procédure LOGLI.

7.1 Premier exemple

```
----+----1----+----2----+----3----+----4----+----5----+----
PROC SELEC
==== SELECTION DES DONNEES UTILES ====
LSELI = TOT, IMASS = UNIF, LZERO = NOREC, LEDIT = COURT
NOMI ACT 2,8,11
FIN

PROC LOGLI
=== MODELE LOG-LINEAIRE ===
LMOD = 1 LEDIT = 1 MAXIT = 15 LTAB = 1
V2+V8+V11
V2+V8+V11+V2*V8+V8*V11
V2+V8+V11+V2*V8+V2*V11+V8*V11+V2*V8*V11
FIN
STOP     : FIN DU FICHIER DE COMMANDES
----+----1----+----2----+----3----+----4----+----5----+----
```

Dans cet exemple, l'utilisateur travaille sur un fichier de données "individus x variables". Les variables nominales 2, 8 et 11 ont été sélectionnées par l'étape SELEC. On a pris soin de coder LZERO = NOREC, de sorte que les individus présentant une valeur manquante seront éliminés de l'analyse par la procédure LOGLI. Les modèles sont définis par l'utilisateur (LMOD = 1).

(1) log F = M + var2 + var8 + var11
(2) log F = M + var2 + var8 + var11 + var2*var8 + var8*var11

(3) log F = M + var2 + var8 + var11 + var2*var8 + var2*var 11 + var8*var11 + var2*var8*var11

La constante M est la constante du modèle ajoutée automatiquement par la procédure. De plus l'utilisateur demande seulement l'édition de la table des statistiques d'ajustement (LEDIT =1) après au plus 15 itérations de calcul (MAXIT).

7.2 Exemple 2

```
----+----1----+----2----+----3----+----4----+----5----+----
PROC SELEC
==== SELECTION DES VARIABLES ====
LSELI = TOT, IMASS = UNIF, LZERO = NOREC, LEDIT = COURT
NOMI ACT 1,6,9
FIN

PROC LOGLI
=== MODELE LOG-LINEAIRE ===
LMOD = 2, LEDIT = 4, LTAB = 1

STOP     : FIN DU FICHIER DE COMMANDES
----+----1----+----2----+----3----+----4----+----5----+----
```

Dans l'exemple 2, l'utilisateur demande de traiter tous les modèles possibles (LMOD =2). La procédure analysera 19 modèles log-linéaires. Seront édités les tris à plat des 3 variables, le tableau de contingence, les informations sur la convergence, les estimations des coefficients du maximum de vraisemblance de chaque modalité, le tableau des modèles triés selon la statistique AIC, la matrice de structure du modèle ainsi que la table des statistiques d'ajustement de chaque modèle (LEDIT = 4).

7.3 Exemple 3

L'exemple suivant indique comment faire une analyse log-linéaire sur un tableau de fréquences.

```
----+----1----+----2----+----3----+----4----+----5----+----6----+---
PROC SELEC     : SELECTION DES FREQUENCES POUR MODELES LOG-LINEAIRES
MODELE LOG-LINEAIRE
LSELI = LIST, IMASS = UNIF, LZERO = NOREC, LEDIT = COURT
FREQ ACT 11--15

ACT 41--64
FIN

PROC LOGLI
=== MODELE LOG-LINEAIRE ===
LMOD = 1, LEDIT = 1, MAXIT = 5, LTAB = 2, LDICO = OUI
MLIG= 2 4 3      : 3 VARIABLES EN LIGNE A 2, 4 ET 3 MODALITES
MCOL= 5          : 1 VARIABLE EN COLONNE A 5 MODALITES
   2 SEXE
 HOM HOMME
 FEM FEMME
```

```
   4 STATUT MATRIMONIAL
   C CELIBATAIRE
   M MARIE
   D DIVORCE
   V VEUF
   3 RESIDENCE
  CR COMMUNE RURALE
  CU COMMUNE URBAINE
 AGG AGGLOMERATION
   5 AGE
AGE1 DE 15 A 35
AGE2 DE 36 A 45
AGE3 DE 46 A 55
AGE4 DE 56 A 65
AGE5 PLUS DE 65
L1 + L2 + L1*L2 + C1 : ECRITURE DU MODELE
FIN
STOP     : FIN DU FICHIER DE COMMANDES
----+----1----+----2----+----3----+----4----+----5----+----6----+---
```

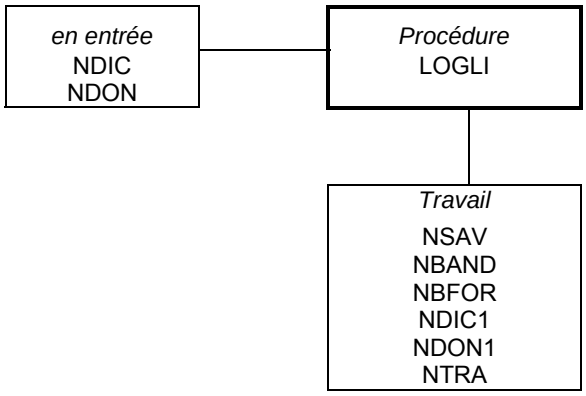
Dans cet exemple, la procédure SELEC sélectionne 5 colonnes et 24 lignes d'un tableau de fréquences. Dans LOGLI, le tableau de contingence retenu est décrit selon sa structure en ligne et en colonne. Les 24 lignes sont issues du croisement de 3 variables ayant dans l'ordre 2, 4 et 3 modalités (MLIG = 2 4 3). Les 5 colonnes sont les modalités d'une seule variable (MCOL = 5). On remarque que le produit du nombre total de modalités est égal à 2*4*3*5 =120, qui est exactement 24*5, le produit du nombre de lignes par le nombre de colonnes. Si ce n'était pas le cas, la procédure demanderait à l'utilisateur une vérification de sélection des fréquences ou des nombres de modalités de chaque variable. Les libellés des 4 variables constituant le tableau de contingence sont insérés à la suite des commandes MLIG et MCOL (LDICO = OUI).
Dans ce dictionnaire, les variables apparaissent dans le même ordre que dans les instructions MLIG et MCOL, en commençant obligatoirement par les variables lignes (de MLIG). De plus, l'utilisateur demande une édition minimale (LEDIT = 1) du modèle suivant (LMOD = 1) :

log F = M + Sexe + Statut Matrimonial + Sexe*Statut Matrimonial + Age

Il est à noter que, dans l'écriture du modèle, les variables sont repérées par un préfixe (Li ou Cj) indiquant s'il s'agit de lignes ou de colonnes dans le tableau de contingence traité. On remarquera que la variable Résidence ne figure pas dans le modèle car aucune mention à L3 n'est faite dans la spécification du modèle qui suit le dictionnaire des variables.

8. Fichiers nécessaires à l'exécution

- en lecture NDIC (dictionaire utile)
  NDON (données utiles)



Procédure MLGEN

1. Présentation

1.1 Objet

Cette procédure, proche à bien des égards de la procédure VAREG, assure les calculs et les éditions d'un ajustement des moindres carrés sur un modèle linéaire comprenant un terme constant. Elle permet d'effectuer les régressions multiples, les analyses de variance et de covariance avec facteurs hiérarchisés (ou emboîtés) et interactions d'ordre un ou deux. Elle accepte de traiter des dispositifs déséquilibrés, des modèles linéaires généraux. En option, il est possible d'obtenir le calcul et l'impression des coefficients de régression avec, pour chacun, son écart-type et le test de sa nullité, valable dans le contexte où le terme aléatoire est supposé engendré par une loi de Laplace-Gauss.

La procédure réalise une analyse de la variance pour tester l'existence de l'effet de chacun des facteurs du modèle. Pour ces tests, le carré moyen de référence est, par défaut, le carré moyen résiduel. Trois types de sommes des carrés des écarts sont possibles qui considèrent plusieurs conditions d'ajustement des facteurs. Sur demande, le programme teste un facteur par rapport à un autre facteur dont le carré moyen est pris comme référence en lieu et place du carré moyen résiduel.

En option, il est permis de demander le calcul d'estimations ou de contrastes, ainsi que celui des moyennes ajustées des niveaux de facteurs du modèle. En outre, ces moyennes ajustées peuvent faire l'objet de comparaisons deux à deux, ou de comparaisons à un niveau de référence, à un témoin.

Le programme est susceptible d'écrire un fichier contenant les variables ayant participé à la définition du modèle, les résidus du modèle et les variables illustratives choisies au départ par l'utilisateur. A partir de ce fichier, il est loisible de faire appel aux procédures graphiques de SPAD pour, notamment, examiner les résidus.

Les enregistrements présentant des **données manquantes** au niveau des variables définissant le modèle ne sont pas retenus par la procédure **MLGEN**. Seul un manque partiel d'information sur les variables à expliquer est toléré. En effet, dans l'optique de l'analyse de la variance, il n'est pas raisonnable, sauf dans certains cas, de générer artificiellement de l'information, des degrés de liberté, en remplaçant les variables continues non renseignées par une moyenne, ou en créant, pour les variables nominales une modalité "valeur inconnue".

Le programme a été écrit par Yves-Marie Chatelin, dans le cadre d'un partenariat de développement entre le CISIA•CERESTA et l'Institut de l'Elevage (149 rue de Bercy – 75595 PARIS Cedex 12).

1.2 Editions

Les éditions résultant de la présente procédure sont nombreuses et, pour la majorité d'entre elles, optionnelles.

L'édition de base, après une présentation sommaire des variables concourant à la définition du modèle, fournit le tableau de la régression multiple qui permet, grâce à un test de Fisher et au coefficient de corrélation multiple (Somme des carrés des écarts du modèle/somme totale des carrés des écarts), de juger de la pertinence et de l'intérêt du modèle linéaire général analysé et donne également l'estimation de la variance commune des résidus.

Arrive ensuite, pour chaque type de sommes des carrés des écarts retenu, le tableau classique de l'analyse de la variance. Pour chaque facteur, il est possible de connaître le nombre de degrés de liberté, la somme des carrés des écarts, le carré moyen, la statistique de Fisher correspondante, ainsi que la probabilité critique qui lui est associée. Soulignons que cette statistique qui prend, comme carré moyen de référence, le carré moyen résiduel, considère que chaque facteur est à effet fixe.

A sa demande, l'utilisateur peut obtenir l'édition de la matrice X'X, de la matrice inverse généralisée, des solutions des équations normales du modèle, de la forme générale des fonctions estimables et enfin des fonctions liées au calcul, pour chaque facteur, de la somme des carrés des écarts de type I, II ou III, ou, plus précisément, de celles dont il a sollicité le calcul.

Lorsqu'une estimation, un contraste, ou une moyenne ajustée a été demandé, outre l'édition des résultats découlant directement de la requête, il est possible d'obtenir, à la demande, les coefficients des fonctions assurant, à partir des solutions des équations normales, le calcul de ces entités.

1.3 Paramètres

La procédure MLGEN fait appel à douze paramètres généraux, quatre paramètres de fonctionnement, sept paramètres d'édition et un paramètre d'archivage.

Un paramètre (LPOND) offre le choix entre deux types de pondération des individus, une pondération de type fréquence, chaque enregistrement pouvant représenter plusieurs individus, ou une pondération de type poids, un enregistrement, bien que représentant, pour ce qui est du calcul des différentes sommes des carrés des écarts, plusieurs individus, ne comptant que pour un dans le calcul du total des degrés de liberté. Trois paramètres (MSC1, MSC2, MSC3), non exclusifs, demandent le calcul, pour les facteurs, des sommes des carrés des écarts de type respectivement I, II ou III.

Les sept paramètres suivants permettent d'obtenir l'édition de la matrice X'X (LXPX), de la matrice inverse généralisée (LINVG), des solutions des équations normales du modèle (LSOL), de la forme générale des fonctions estimables (LFES) et, pour terminer, des fonctions associées au calcul, pour chaque facteur, de la somme des carrés des écarts de type I (LESF1), II (LESF2), III (LESF3), pour autant que le calcul des sommes des carrés des écarts du type désigné ait été sollicité à l'aide du paramètre *ad hoc* (MSC1, MSC2, MSC3).

Le dernier paramètre (LARCH) déclenche l'archivage, dans une base SPAD, des variables servant à définir le modèle, des résidus de celui-ci et des variables nominales et/ou continues illustratives. Ces dernières, ne participant pas à l'analyse, sont retenues par l'utilisateur qui envisage de les faire intervenir ultérieurement, par exemple, lors de l'examen graphique des résidus.

1.4 Après les paramètres

Une fois introduits les paramètres, il convient de définir précisément le modèle linéaire général souhaité, sachant que la procédure n'accepte qu'un modèle, mais que celui-ci peut concerner, si cela est désiré, plusieurs variables continues à expliquer. Ce modèle peut être complété par des demandes, formulées à l'aide de mots clefs, de calculs de tests de facteurs (TFACT), d'estimations (ESTIM), de contrastes (CONTRA) et de moyennes ajustées (MOYAJ). Ces différentes requêtes sont explicitées au moyen d'autres mots clefs (NOM, FACT, CREF, TSC, NDL, COEFMU, COEF, DIESTI, LEDIT, COMPM, COMPJ) qui sont examinés plus loin. La spécification du modèle et des calculs complémentaires s'achève impérativement par le mot clef FIN.

Le fait de ne pouvoir soumettre qu'un seul modèle ne devrait pas être un problème. Dans le cas d'un dispositif expérimental, ce modèle est unique, déterminé qu'il est par le dispositif expérimental et, dans le cas de recherche de modèle, l'écriture de celui-ci est si aisé qu'il ne faut pas hésiter à faire appel, autant de fois que cela est nécessaire, à la présente procédure.

2. Instructions de commande

La valeur par défaut de chaque paramètre fait suite, entre parenthèses, au nom de celui-ci.

(1) **PROC MLGEN** Modèle linéaire général

(2) *titre de l'analyse*

Paramètres de fonctionnement

(3) **LPOND** **Pondération des individus**
1 : pondération de type fréquence
2 : pondération de type poids

MSC1 **Calcul des sommes des carrés des écarts de type I**
(1 ou OUI) 0 ou NON : le calcul n'est pas réalisé
1 ou OUI : le calcul est réalisé

MSC2 **Calcul des sommes des carrés des écarts de type II**
(0 ou NON) 0 ou NON : le calcul n'est pas réalisé
1 ou OUI : le calcul est réalisé

MSC3 **Calcul des sommes des carrés des écarts de type III**
(1 ou OUI) 0 ou NON : le calcul n'est pas réalisé
1 ou OUI : le calcul est réalisé

Paramètres d'édition.

LXPX Edition de la matrice X'X
(0 ou NON) 0 ou NON : pas d'édition
1 ou OUI : édition de la matrice X'X

LINVG Edition de la matrice inverse généralisée
(0 ou NON) 0 ou NON : pas d'édition
1 ou OUI : édition de la matrice inverse généralisée

LSOL Edition des solutions des équations normales du modèle
(0 ou NON) 0 ou NON : pas d'édition
1 ou OUI : édition des solutions

LFES Edition de la forme générale des fonctions estimables
(0 ou NON) 0 ou NON : pas d'édition
1 ou OUI : édition

LESF1 Edition des fonctions associées au calcul, pour chaque facteur, des sommes des
(0 ou NON) carrés des écarts de type I
0 ou NON : pas d'édition
1 ou OUI : édition

LESF2 Edition des fonctions associées au calcul, pour chaque facteur, des sommes des
(0 ou NON) carrés des écarts de type II
0 ou NON : pas d'édition
1 ou OUI : édition

LESF3 Edition des fonctions associées au calcul, pour chaque facteur, des sommes des
(0 ou NON) carrés des écarts de type III
0 ou NON : pas d'édition
1 ou OUI : édition

Paramètre d'archivage

LARCH **Archivage des variables actives et illustratives des résidus**
(0 ou NON) 0 ou NON : pas d'archivage
1 ou OUI : archivage

Ecriture du modèle.

(4) **MODEL** Définition obligatoire du modèle

TFACT Introduction du test d'un facteur

ESTIM Introduction de la définition d'une estimation

CONTR Introduction de la définition d'un contraste

MOYAJ Introduction de la demande d'une moyenne ajustée

NOM Introduction du nom attribué à une estimation ou à un contraste

FACT **Introduction d'une liste de facteurs :**
- ou à tester,
- ou auxquels sont affectés des coefficients définissant une estimation ou un contraste,
- ou dont des moyennes ajustées sont demandées.

CREF Introduction du choix d'un carré moyen de référence pour le test d'un facteur, un contraste ou une moyenne ajustée

TSC Introduction du type de sommes des carrés des écarts du carré moyen de référence désigné par le paramètre CREF

NDL (1) Introduction du nombre de contrastes testés simultanément (nombre de degrés de liberté)

COEFMU (0.0) Introduction du coefficient affecté au terme moyen Mu (ordonnée à l'origine) dans la définition d'une estimation ou d'un contraste

COEF (0.0) Introduction de la liste des coefficients attribués aux modalités des facteurs servant à définir une estimation ou un contraste

DIESTI (1.0) Introduction du diviseur commun aux différents coefficients définissant une estimation.

LEDIT **Edition des coefficients de la fonction associée au calcul, suivant le cas,**
(0 ou NON) **d'une estimation, d'un contraste ou d'une moyenne ajustée.**
0 ou non : pas d'édition.
1 ou oui : édition.

COMPM **Comparaison deux à deux des moyennes ajustées des niveaux d'un facteur.**
(0 ou NON) 0 : pas de comparaison.
1 : comparaison avec test de la ppds.
2 : comparaison avec test de bonferroni.
3 : comparaison avec test de scheffé.
4 : comparaison avec test de tukey-kramer.

COMPJ Comparaison des moyennes ajustées des niveaux d'un facteur, à la moyenne
(0 ou NON) ajustée de l'un d'entre eux pris comme référence. Il s'agit du test bilatéral de dunnett.

FIN Fin de la description du modèle et de la définition des options qui lui sont associées.

3. Présentation détaillée des paramètres

3.1 Paramètres de fonctionnement.

LPOND **Pondération des individus.**

valeurs possibles : 1 (la pondération est de type fréquence)
2 (la pondération est de type poids)
valeur par défaut : 1

Ce paramètre offre le choix entre deux types de pondération des individus.

Si LPOND = 1, chaque individu pèse le poids qui lui a été attribué, sachant que ce poids, un entier, doit être supérieur à 0. Par défaut, lorsque aucune variable de pondération n'a été désignée, ce poids est égal à 1. Le nombre de degrés de libertés total du modèle est égal à la somme des poids des individus moins un. Les moyennes ajustées tiennent compte également de cette pondération. Tout individu dont le poids est inférieur à 1 est exclu de l'analyse.

Si LPOND = 2, La variable de pondération ne joue pas sur le poids des individus qui reste, pour tous, de 1. Seule la matrice X'X et la matrice X'Y sont affectées au niveau de leur calcul et, par voie de conséquence, les solutions des équations normales ainsi que les moyennes ajustées. Il est à noter que seuls les individus nantis d'un poids strictement positif sont retenus dans l'analyse.

En définitive, la seule différence entre ces deux types de pondération est le degré de liberté total du modèle, et, subséquemment, celui de la résiduelle, qui n'est affecté que par la pondération de type fréquence (LPOND=1).

MSC1 **Calcul des sommes des carrés des écarts de type I.**

- *valeurs possibles* : 0 ou NON (pas de calcul)
  1 ou OUI (calcul des sommes des carrés de type I)
- *valeur par défaut* : 1 ou OUI

Dans le cas des sommes des carrés des écarts de type I, chaque facteur n'est ajusté que des facteurs qui le précèdent dans le modèle tel qu'il a été défini par l'utilisateur. Cela permet d'obtenir une analyse progressive de ces facteurs, une sorte d'analyse ascendante puisque le test d'un facteur est indépendant de tous les facteurs situés après lui.

Soit le modèle : Y= Mu+V1+V2+V3(V1)+V1*V2

V1 est sans ajustement,
V2 est ajusté de V1,
V3(V1) est ajusté de V1 et V2.
V1*V2 est ajusté de V1, V2 et V3(V1).

Il convient de souligner que, hormis le cas où le modèle est équilibré et sans covariables, les résultats obtenus avec les sommes des carrés des écarts de type I sont tributaires de l'ordre dans lequel les facteurs sont introduits dans le modèle.

MSC2 **Calcul des sommes des carrés des écarts de type II.**

- *valeurs possibles* : 0 ou NON (pas de calcul)
  1 ou OUI (calcul des sommes des carrés de type II)
- *valeur par défaut* : 0 ou NON

Avec les sommes des carrés des écarts de type II, chaque facteur est ajusté de tous les autres à l'exception de ceux à la constitution desquels il participe ou plus précisément de ceux dans lesquels il est "contenu". Un facteur est "contenu" dans un autre s'il :

1 hiérarchise cet autre facteur. Par exemple, le facteur V_j est "contenu" dans le facteur hiérarchisé $V_i(V_j)$

2 définit, avec d'autres facteurs ou d'autres variables, une interaction. Ainsi, le facteur V_j est "contenu" dans l'interaction V_i*V_j, laquelle est contenue dans l'interaction $V_i*V_j*V_k$

Soit le modèle : Y= Mu+V1+V2+V3(V1)+V1*V2

V1 est ajusté de V2, mais ni de V3(V1), ni de V1*V2,
V2 est ajusté de V1 et de V3(V1), mais pas V1*V2,
V3(V1) est ajusté de V1, V2 et V1*V2,
V1*V2 est ajusté de V1, V2 et V3(V1).

L'interprétation de telles sommes des carrés des écarts est délicate, aussi, cette option est-elle rarement retenue. Elle est utilisée pour tester, en un seul passage de la procédure, des effets marginaux et d'interaction

MSC3 **Calcul des sommes de carrés des écarts de type III.**

- *valeurs possibles* : 0 ou NON (pas de calcul)
  1 ou OUI (calcul des sommes de carrés de type III)
- *valeur par défaut* : 1 ou OUI

Dans le cas de sommes des carrés des écarts de type III, chaque facteur est ajusté de tous les autres présents dans le modèle. Il s'agit là de l'option la plus couramment adoptée.

Il est parfaitement possible de demander les trois types de sommes des carrés des écarts. Si l'utilisateur ne choisit aucuns de ces trois types de sommes des carrés des écarts, les types I et III seront d'office calculés par la procédure. Dans le cas général, seule la somme des carrés des écarts du dernier facteur introduit dans le modèle est identique pour les trois types de sommes des carrés.

Ce n'est que si le dispositif analysé est parfaitement équilibré et sans covariables, que les trois types de sommes des carrés des écarts fournissent rigoureusement les mêmes résultats.

3.2 Paramètres d'édition.

LXPX **Edition de la matrice X'X ainsi que des matrices X'Y et Y'Y**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des matrices X'X, X'Y et Y'Y)
- *valeur par défaut* : 0 ou NON

LINVG **Edition de la matrice inverse généralisée X'X$^-$.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition de la matrice inverse généralisée)
- *valeur par défaut* : 0 ou NON

Le nombre de paramètres du modèle qu'il est possible d'estimer est égal au nombre de degrés de liberté de ce modèle. Pour rendre inversible la matrice X'X et obtenir une matrice inverse généralisée, il faut introduire des équations supplémentaires, en nombre égal à la différence entre la dimension et le rang de la matrice X'X. Ces équations supplémentaires, ou contraintes, consistent, par exemple, à poser :

- que la somme des coefficients de régression associés à un facteur est nulle,
- ou, ce qui est la contrainte posée dans la procédure MLGEN, que le coefficient de régression lié au dernier niveau de chaque facteur est nulle.

LSOL **Edition des solutions des équations normales.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des solutions)
- *valeur par défaut* : 0 ou NON

Dans la mesure où la matrice X'X n'est généralement pas de plein rang, les solutions des équations normales du modèle dépendent des contraintes supplémentaires posées pour générer la matrice inverse généralisée. Il s'en suit que ces solutions sont, pour la plupart, biaisées. Il en est, le plus souvent, tout autre des estimations obtenues à partir de ces solutions. C'est notamment le cas pour les moyennes ajustées.

LFES **Edition de la forme générale des fonctions estimables.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition de la forme générale des fonctions estimables)
- *valeur par défaut* : 0 ou NON

Le terme de fonction ne s'applique à la matrice opératrice par laquelle est multipliée la matrice des solutions des équations normales pour obtenir une somme des carrés des écarts, un contraste, ou la moyenne ajustée du niveau d'un facteur du modèle.

Seules les fonctions dont le résultat ne dépend pas des contraintes supplémentaires posées pour obtenir la matrice inverse généralisée de la matrice de X'X sont estimables, c'est-à-dire valablement calculables. Cette option permet demander l'édition de la forme que doit respecter toute fonction qu'elle soit destinée au calcul d'une somme de carrés des écarts, d'un contraste, d'une estimation ou d'une moyenne ajustée. Si cette forme n'est pas satisfaite, le calcul associé n'est pas réalisé.

LESF1 **Edition des fonctions destinées au calcul, pour chaque facteur, de la somme des carrés des écarts de type I.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des fonctions)
- *valeur par défaut* : 0 ou NON

Une telle édition, bien entendu, n'est possible que si le calcul des sommes des carrés des écarts de type I a été auparavant demandé.

LESF2 **Edition des fonctions destinées au calcul, pour chaque facteur, de la somme des carrés des écarts de type II.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des fonctions)
- *valeur par défaut* : 0 ou NON

Cette édition n'est réalisable que si le calcul de ce type de sommes des carrés des écarts a été au préalable demandé.

LESF3 **Edition des fonctions destinées au calcul, pour chaque facteur, de la somme des carrés des écarts de type III.**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des fonctions)
- *valeur par défaut* : 0 ou NON

L'édition n'a bien évidemment lieu, que si un tel calcul a été sollicité.

3.3 Paramètre d'archivage.

LARCH **Archivage des variables et des résidus.**

- *valeurs possibles* : 0 ou NON (pas d'archivage)
  1 ou OUI (Archivage des variables et des résidus)
- *valeur par défaut* : 0 ou NON

Au moment de la définition du modèle, il est suggéré à l'utilisateur d'archiver les variables actives dans l'analyse ainsi que les résidus du modèle. Comme, en outre, il lui est proposé de désigner des variables illustratives qui seront également archivées, il a tout loisir de constituer une base exploitable, pourquoi pas, par les procédures graphiques de SPAD.

Bien entendu, le nom de la base d'archivage doit être stipulée auparavant.

4. Définition du modèle linéaire général.

Modèle **Définition du modèle linéaire général.**

L'écriture du modèle, un seul par procédure, suit l'introduction des paramètres de fonctionnement, d'édition et d'archivage. Cette définition comporte deux parties, l'une obligatoire qui correspond à l'écriture du modèle proprement dit, l'autre, facultative, permet de demander un certain nombre de calculs complémentaires à savoir :

- des tests de facteurs,
- des estimations,
- des contrastes,
- des moyennes ajustées, avec éventuellement comparaisons multiples.

4.1 Définition du modèle linéaire général

Le modèle s'écrit sous la forme suivante :

- $V_5\ V_9\text{--}V_{12} = V_2 + V_3 + \ldots + V_n + V_l(V_m) + V_i{*}V_j + \ldots + V_i{*}V_j{*}V_k + \ldots$

Les variables se trouvant à gauche du signe "=" sont les **variables à expliquer** (parfois dénommées **endogènes**). Elles doivent être, cela est indispensable, continues. Il est possible d'en proposer plusieurs, sachant que toutes celles qui correspondent strictement à la même matrice X'X, c'est-à-dire qui ont leurs éventuelles valeurs manquantes pour les mêmes individus, sont traitées simultanément. Les variables qui ne satisfont pas cette règle sont, elles, traitées dans un autre temps. L'usage du symbole de suite "--" est autorisé pour introduire une suite de variables.

A la droite du signe "=" se trouvent les **facteurs explicatifs** et/ou les **covariables** (ou **exogènes**). Ceux-ci peuvent être des facteurs à effet simple, des facteurs hiérarchisés (ou emboîtés) ou des interactions d'ordre un ou d'ordre deux. Ils font intervenir des variables généralement nominales, mais, avec certaines limitations, les variables continues ou covariables sont acceptées. Ces facteurs doivent être séparés les un des autres par le signe "+".

Un facteur hiérarchisé est décrit par deux éléments, la variable hiérarchisée et la variable hiérarchisante placée à la suite et entre parenthèses.

- $V_l(V_m)$ est un facteur hiérarchisé. La variable V_l est hiérarchisée par la variable V_m.

La variable hiérarchisée et la variable hiérarchisante sont obligatoirement des variables nominales. La variable hiérarchisante doit impérativement figurer par ailleurs dans le modèle en tant que facteur à effet simple. Au contraire, la variable hiérarchisée ne doit pas apparaître autre part dans ce modèle. Ainsi, la procédure n'admet pas d'interactions impliquant une variable hiérarchisée.

Une interaction est identifiée par le caractère "*" placé entre le nom des deux variables contribuant à définir l'interaction.

- V_i*V_j est une interaction d'ordre un (entre les variables V_i et V_j),
- $V_i*V_j*V_k$ est une interaction d'ordre deux.

Une interaction d'ordre un fait intervenir au moins une variable nominale et peut concerner aussi une variable continue avec, alors, une variable nominale.

Par contre, une interaction d'ordre deux ne doit être définie que par des variables nominales.

Ces limites relatives aux facteurs possibles peuvent paraître restrictives, mais, dans la grande majorité des cas il n'en est rien. En particulier, une interaction entre deux variables continues, non autorisée par la procédure MLGEN et à déconseiller, peut facilement être introduite par le truchement d'une variable résultant du produit de ces deux variables. S'agissant des interactions d'un ordre supérieur à deux, il faut être doté d'une très forte imagination pour les interpréter.

Bien entendu, il est interdit à une variable à expliquer d'apparaître au niveau d'un facteur explicatif.

Il va de soit, enfin, que toutes les variables figurant dans le modèle doivent avoir été préalablement retenues lors de la précédente étape SELEC.

4.2 Calculs complémentaires

Chaque demande de calcul complémentaire est introduite par un mot clef. Elle est ensuite précisée par une série d'informations elles-mêmes identifiées par un mot clef.

TFACT **Introduction du test d'un facteur.**

Tous les facteurs du modèle sont réputés à effet fixe. Par défaut, ces facteurs se testent donc par rapport au carré moyen résiduel. Toutefois, il peut se faire que, pour un facteur particulier, l'utilisateur souhaite recourir à un autre carré moyen de référence, celui d'une interaction dans laquelle ce facteur est impliqué, par exemple. Ce mot clef permet de réaliser un tel test. Pour préciser la demande il faut, à l'aide de trois mots clefs spécifiques, mentionnés entre parenthèses, indiquer :

- le facteur à tester (FACT).
- le facteur dont le carré moyen est le carré moyen de référence (CREF). Cette mention, comme la précédente est évidemment obligatoire.
- le type de sommes des carrés des écarts de ce carré moyen de référence (TSC). Cette information n'est pas obligatoire. Si elle n'est pas fournie, la procédure opte pour le type de sommes des carrés dont le numéro est le plus élevé parmi ceux dont le calcul a été requis (paramètres MSC1, MSC2, MSC3).

Dans le cas où le modèle est hiérarchisé, il importe de faire attention à l'ordre des facteurs si seules les sommes des carrés de type I sont demandées.

Si l'utilisateur dépouille un dispositif en Spit-plot, la présente option est intéressante qui permet d'indiquer, pour les facteurs du premier étage, le bon carré moyen de référence qui n'est évidemment pas le carré moyen résiduel.

Dans l'exemple suivant :

- TFACT FACT=V3 CREF=V3*V6 TSC=3

Le facteur V3 est à tester par rapport au carré moyen de l'interaction V3*V6 en prenant les sommes des carrés des écarts de type III. En fait, l'option TSC=3 est superflue puisqu'il s'agit le l'option par défaut.

ESTIM **Introduction de la définition d'une estimation.**

L'utilisateur peut souhaiter calculer une combinaison linéaire des moyennes ajustées des niveaux de certains facteurs du modèle. Il peut vouloir connaître, par exemple, la différence entre un niveau d'un facteur et la moyenne de plusieurs autres niveaux de ce facteur. Cette estimation doit être définie en indiquant, par le truchement de six mots clefs désignés entre parenthèses, :

- le nom désignant cette estimation (NOM). Ce nom est limité à vingt caractères.
- les facteurs aux niveaux desquels sont attribués des coefficients (FACT).
- le coefficient à attribuer, facultativement, au terme moyen Mu (COEFMU). Par défaut, ce coefficient est nul.
- les coefficients à attribuer aux niveaux des facteurs spécifiés par le mot clef FACT (COEF). Par défaut, le coefficient d'un niveau non renseigné est nul.
- le diviseur commun à tous les coefficients (DIESTI). Ce diviseur, dont la valeur par défaut est un, permet de ne pas introduire des coefficients fractionnaires, il suffit qu'il soit le plus petit dénominateur commun de tous ces coefficients qui seront alors indiqués au moyen de leurs numérateurs.
- l'édition des coefficients associés à l'estimation peut être réclamée par l'intermédiaire du mot clef LEDIT dont la valeur par défaut est 0 ou NON.

Le résultat de l'estimation est testé à zéro en utilisant, comme référence, le carré moyen résiduel.

Dans l'exemple suivant :

- ESTIM NOM='1/2(V3.1+V3.2)-V3.3' FACT= V3 COEF= 1 1 -2 DIESTI= 2

la procédure calcule, sous le nom de "1/2(V3.1+V3.2)-V3.3", la différence entre la moyenne des moyennes ajustées des niveaux 1 et 2 du facteur V3 et la moyenne ajustée du niveau 3 de ce même facteur.
Le diviseur 2 permet d'entrer :

- COEF 1 1 -2

plutôt que :

- COEF 0.5 0.5 -1

Une seule estimation peut être définie à la suite du mot clef ESTIM, mais il est possible de faire plusieurs fois appel à celui-ci dans un modèle.

CONTR Introduction de la définition d'un contraste.

Pour tester à zéro une combinaison linéaire des solutions des équations normales, un contraste, il suffit de faire appel à cette instruction CONTR.

S'il s'agit simplement d'un contraste, le recours au calcul d'une estimation peut assurer ce test de nullité mais, comme cela a été dit plus haut, uniquement en référence au carré moyen résiduel. La demande est à préciser, à l'aide de huit mots clefs dont certains sont facultatifs. Il s'agit en effet :

- de dénommer, en au plus vingt caractères, le contraste (NOM).

- d'indiquer le nombre degrés de liberté du contraste (NDL). Un contraste n'a qu'un degré de liberté. Cependant cette procédure autorise le test simultané de plusieurs contrastes dès lors qu'ils sont séparément estimables. Ce mot clef permet de stipuler le nombre de ces contrastes associés dans un même test. Il est à remarquer que cette extension de la notion de contraste est déjà exploitée pour le test des facteurs qui fait appel effectivement à un jeu de contrastes. Si l'estimabilité des contrastes individuels est indispensable, il est bon également qu'ils soient deux à deux orthogonaux. Mais cette dernière condition n'est pas expressément obligatoire. Par défaut, ce nombre de degrés de liberté prend la valeur un.

- de préciser les facteurs dont les niveaux seront nantis de coefficients (FACT).

- de déclarer ces coefficients qui par défaut sont initialisés à zéro (COEF). Evidemment ces deux mots-clefs et les deux séries d'informations qu'ils annoncent sont indispensables.

- de signaler s'il faut affecter un coefficient au terme constant Mu (COEFMU). Si cette information n'est pas fournie, le coefficient est nul.

- de désigner, si ce n'est pas la résiduelle, le facteur dont le carré moyen doit servir de référence pour les tests (CREF).

- de notifier, lorsque l'option précédente a été prise, le type de sommes de carrés des écarts qui doit être retenu pour les tests (TSC).

- de demander, éventuellement, l'édition des coefficients servant au calcul du contraste (LEDIT).

Dans l'exemple suivant :

- CONTRA NOM='1/2(V3.1+V3.2)-V3.3' NDL=1 FACT= V3 COEF= 1 1 -2 > CREF=V2*V3 TSC=3 LEDIT=1

la procédure compare les niveaux un et deux du facteur V3 au niveau 3 de ce même facteur et elle utilise, comme carré moyen de référence pour le test de cette différence, le carré moyen, de type III, de l'interaction V2*V3. Les coefficients assurant cette comparaison sont édités.

Un seul contraste est décrit à la suite du mot clef CONTRA qui, par contre, peut être invoqué à plusieurs reprises dans les instructions associées à un modèle.

MOYAJ **Introduction de la demande d'une moyenne ajustée.**

Dans le cas d'un modèle déséquilibré, l'analyse de variance ne repose pas sur les moyennes brutes des niveaux des facteurs, mais sur les moyennes ajustées du déséquilibre.

Ces moyennes ajustées peuvent être éditées. C'est l'objet de cette instruction qui est à préciser au moyen de cinq mots clefs. Il importe :

- d'indiquer le facteur, un seul par requête, dont les moyennes ajustées sont souhaitées (FACT).

- de désigner, le cas échéant, le facteur, autre que la résiduelle, dont le carré moyen doit servir de référence dans les tests des moyennes à zéro, ou, si elles ont été demandées, des comparaisons deux à deux des moyennes ajustées des niveaux du facteur (CREF).

- de signaler le type de sommes des carrés des écarts à considérer dans les tests évoqués ci-dessus (TSC).

- de dire s'il faut procéder à la comparaison deux à deux des moyennes ajustées des niveaux du facteur (COMPM). Quatre tests sont proposés :

  - un simple test basé sur la plus petite différence décelable qui ne tient pas compte du nombre de comparaisons réalisées. Si le risque α au niveau de chaque comparaison ne bouge pas, par contre, celui portant sur l'ensemble de ces comparaisons croit dans des proportions importantes. Les résultats de ce test ne sont à considérer que si le test global du facteur est significatif.

  - le test avec ajustement de Bonferroni. L'ajustement considère que, pour être dite significative, la différence entre deux moyennes doit satisfaire à la condition :
    $\left|\overline{Y}_i - \overline{Y}_j\right| \Big/ \left(s_i + s_j\right) \geq t(\varepsilon, \nu)$, s_i et s_j étant les écarts-types des moyennes des deux niveaux du facteur, ν étant le nombre de degrés de liberté associé au carré moyen de référence
    avec : $\varepsilon = \alpha \Big/ \left(k(k-1)/2\right)$, k étant le nombre de niveaux du facteur.

    Le risque individuel est donc divisé par le nombre de comparaisons possible, $k(k-1)/2$. L'expérience montre que ce test est très conservateur.

  - le test avec ajustement de Scheffé. Avec ce test, deux moyennes sont significativement différentes si :

    $$\left|\overline{Y}_i - \overline{Y}_j\right| \Big/ \left(s_i + s_j\right) \geq \sqrt{(k-1)\mathrm{F}(\alpha, k-1, \nu)}$$

    Ce test est réputé manquer de puissance.

- le test avec ajustement de Tukey-Kramer qui considère que deux moyennes différent significativement si :

$$\left|\overline{Y}_i - \overline{Y}_j\right| \Big/ \left(s_i + s_j\right) \geq q(\alpha, k, \nu)$$

$q(\alpha, k, \nu)$ étant la valeur critique, au seuil α, d'une distribution du range studentisé de k variables normales indépendantes, avec ν degrés de liberté.

- de demander, si désirée, la comparaison des moyennes ajustées des niveaux du facteur à la moyenne ajustée de l'un de ces deux niveaux pris comme témoin (COMPJ). Le test réalisé est le test bilatéral de Dunnett qui tient compte du fait que ces comparaisons, qui utilisent toutes la même référence, ne sont pas indépendantes.

- d'appeler, si le fait est désiré, l'édition des coefficients qui concourent au calcul des moyennes ajustées.

Une demande de comparaison multiple des moyennes ajustées des niveaux d'un facteur n'est pas suivie d'effet dès lors que l'une de ces moyennes ajustées n'est pas estimable.

Dans l'exemple suivant :

- MOYAJ FACT= V2*V3 COMPM=2

les moyennes ajustées des différents niveaux de l'interaction V2*V3 sont calculées et comparées deux à deux selon le test de Bonferroni. Le carré moyen de référence est le carré moyen résiduel puisque c'est là l'option par défaut.

A la suite de l'instruction MOYAJ, un seul facteur peut faire l'objet d'une demande de moyennes ajustées. Toutefois, cette dernière peut être renouvelée plusieurs fois durant la description d'un modèle.

NOM **Introduction du nom attribué à une estimation ou à un contraste.**

Ce nom, obligatoire, ne comporte pas plus de vingt caractères.

FACT **Introduction d'une liste de facteurs**

- *valeurs possibles* : tout nom de facteur du modèle.

Ce mot clef est utilisé dans toutes les demandes de calcul complémentaire pour annoncer le (ou les) facteur(s) concerné(s). Il va sans dire qu'il est indispensable.

CREF **Introduction du choix d'un carré moyen de référence pour le test d'un facteur, un contraste ou une moyenne ajustée.**

- *valeurs possibles* : tout nom de facteur du modèle.

Ce mot clef introduit le carré moyen de référence lorsque ce dernier n'est pas celui de la résiduelle. Comme le mot clef suivant, il est facultatif.

TSC **Introduction du type de sommes des carrés des écarts du carré moyen de référence désigné par le paramètre CREF.**

- *valeurs possibles* : 1 (Sommes des carrés des écarts de type I)
  2 (Sommes des carrés des écarts de type II)
  3 (Sommes des carrés des écarts de type III)
- *valeur par défaut* : le numéro du type de somme des carrés des écarts calculé le plus élevé.

Dans le cas où un carré moyen de référence a été désigné, ce mot clef, optionnel, indique le type de sommes des carrés des écarts à considérer.

NDL **Introduction du nombre de degrés de liberté d'un contraste.**

- *valeurs possibles* : un nombre de 1 à 5.
- *valeur par défaut* : 1

Un contraste peut, en fait, être une association de plusieurs contrastes qui seront testés conjointement. Ce mot clef, dont la valeur par défaut est un, permet de donner le nombre de contrastes individuels dans l'association.

COEFMU **Introduction du coefficient affecté au terme moyen Mu dans la définition d'une estimation ou d'un contraste.**

- *valeurs possibles* : un nombre réel positif.
- *valeur par défaut* : 0

COEF **Introduction de la liste des coefficients attribués aux modalités des facteurs utilisés pour définir une estimation ou un contraste.**

- *valeurs possibles* : un nombre réel positif.
- *valeur par défaut* : 0

DIESTI **Introduction du diviseur commun des coefficients définissant une estimation.**

- *valeurs possibles* : un nombre réel positif.
- *valeur par défaut* : 1

LEDIT **Edition des coefficients de la fonction associée au calcul, suivant le cas, d'une estimation, d'un contraste ou d'une moyenne ajustée.**

- *valeurs possibles* : 0 ou NON (Pas d'édition)
  1 ou OUI (Edition des coefficients)
- *valeur par défaut* : 0 ou NON

COMPM **Comparaison deux à deux des moyennes ajustées des niveaux d'un facteur.**

- *valeurs possibles* : 0 (Pas de comparaison)
  1 (Comparaison deux à deux et test de la PPDS)
  2 (Comparaison deux à deux, test de Bonferroni)
  3 (Comparaison deux à deux, test de Scheffé)
  4 (Comparaison deux à deux et Tukey-Kramer)
- *valeur par défaut* : 0

COMPJ **Comparaison des moyennes ajustées des niveaux d'un facteur à la moyenne ajustée de l'un de ces niveaux pris comme référence (test bilatéral de Dunnett)**

- *valeurs possibles* : 0 (Pas de comparaison)
  J (indice du niveau pris comme référence)
- *valeur par défaut* : 0

Il convient de signaler que le test de Dunnett est particulièrement gourmand en temps. L'utilisateur doit être un peu patient.

Le recours à ces cinq derniers mots clefs n'est pas obligatoire.

FIN **Fin de la description du modèle et de la définition des options qui lui sont associées.**

5. Exemple de commandes

Un exemple de commandes liées à la procédure MLGEN est présenté ci-après, il comporte également les commandes de la procédure SELEC précédente.

```
----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8
PROC SELEC
Sélection des individus et des variables
:----- PARAMETRES
LEDIT = COURT, LESTI = NON, LZERO = NOREC, LSELI = TOT, IMASS = UNIF, >
TIRER = NON
: ---- VARIABLE(S) RETENUE(S)
NOMI ILL  3
NOMI ACT  1,2,4
CONT ILL  7
CONT ACT  8--10
FIN

PROC MLGEN
Modèle linéaire général
:----- PARAMETRES

LPOND = 0 MSC1 = 1 MSC2 = 0 MSC3= 1 LXPX = 1 LINVG = 1 LSOL = 1 LFES = 1 >
LESF1 = 0 LESF2 = 0 LESF3 = 1 LARCH = 1
MODEL V8--V10 = V1 + V2 + V4 + V1*V4 + V2*V4
TFACT FACT= V4 CREF= V1*V4 TSC= 3
ESTIM NOM= 'Pare 2 - Pare 1' FACT= V2 COEF= -1 1
ESTIM NOM= 'Trait 2+3 - Trait 1' FACT= V4 COEF= -2 1 1 DIESTI= 2>
LEDIT= 1
CONTR NOM= 'Trait 2+3 - Trait 1' FACT= V4 NDL= 1 COEF= -2 1 1 >
      CREF= V1*V4 TSC= 3 LEDIT= 1
MOYAJ FACT= V1 COMPM= 0 LEDIT= 1
MOYAJ FACT= V2 COMPM= 1
MOYAJ FACT= V1*V4 COMPM= 2
MOYAJ FACT= V4 COMPJ= 1
FIN
:=============================================================================
FBDON = 'C:\STAGE\RESIDU.SBA'
NDICA = 'C:\SPAD3\Tmp\XFY7ZAI.999'
NDONA = 'C:\SPAD3\Tmp\XFY7ZAO.999'
PROC EBASE
========= Ecriture de fichier BASE =========
FIN
STOP
----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8
```

Les variables utilisées pour écrire le modèle linéaire général sont au nombre de six, trois nominales (V1,V2 et V4) et trois continues (V8, V9 et V10). Ces dernières sont les variables à expliquer. En outre, deux variables, une nominale (V3) et une continue (V7) sont choisies comme illustratives. Ce choix s'explique, car l'archivage des variables et des résidus est demandé puisque LARCH = 1.

Le calcul des sommes des carrés des écarts de type I et de type III est requis (MSC1 = 1 et MSC3 = 1).

L'utilisateur demande l'édition de la matrice X'X (LXPX = 1), de la matrice inverse généralisée (LINVG = 1), des solutions des équations normales (LSOL = 1), de la forme générale des fonctions estimables (LFES = 1), et des fonctions estimables pour les sommes des carrés des écarts de type I et de type III (LESF1 =1 et LESF3 = 1).

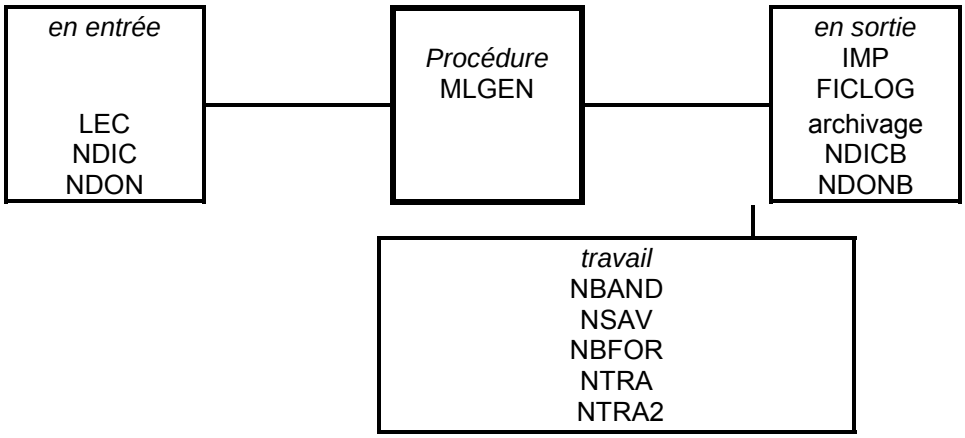
Le test du facteur V4 par rapport à l'interaction V1*V4 est demandé (TFACT FACT= V4 ...).

Deux estimations (ESTIM NOM= ...) et un contraste (CONTR NOM= ...) sont sollicités

Enfin quatre moyennes ajustées sont demandées (MOYAJ FACT=), dont deux avec comparaison multiple (COMPM = 1 ....COMPM = 2) et une avec comparaison à une référence (CPMP = 1).

6. Fichiers nécessaires à l'exécution de la procédure.

- en lecture — LEC (paramètres); NDIC (dictionnaire utile); NDON (données utiles)
- de travail — NBAND; NSAV; NBFOR; NTRA; NTRA2
- en sortie — IMP (résultats); FICLOG (rapport d'exécution)
- d'archivage — NDICB; NDONB — ces deux fichiers n'interviennent que si LARCH = 1.



7. Des exemples de sorties

Dans le présent exemple, fictif, qui concerne une expérience mettant en jeu 90 vaches laitières, trois variables sont à expliquer :

- la production laitière journalière moyenne exprimée en kilogrammes,
- le taux butyreux moyen exprimé en grammes par kilogramme, il s'agit là de la teneur du lait en matière grasse,
- le taux protéique moyen exprimé également en grammes par kilogramme

L'essai s'étant déroulé sur deux campagnes laitières, l'année est le premier facteur introduit dans le modèle.

Le troupeau expérimental comporte des primipares et des multipares, le pare est donc le second facteur pris en compte.

L'essai consiste à comparer, par rapport à un lot d'animaux nourris classiquement, deux niveaux d'un aliment à étudier. Ce facteur, le traitement, est le troisième considéré.

Les sorties ci-après résultent de l'exécution du fichier de commande fourni plus haut en 6 à titre d'exemple.

7.1 Description des variables

Le début de la sortie décrit sommairement les protagonistes du modèle en fournissant une correspondance entre l'intitulé des variables qui sera utilisé ultérieurement dans les tableaux de résultats et leur libellé en clair. Il en est de même pour les libellés des niveaux des variables nominales.

```
Modèle linéaire général

Présentation des données

Les variables à expliquer :
        libellé
  V8    Lait brut
  V9    Taux butyreux
  V10   Taux protéique

Les variables nominales explicatives :
        libellé               modalité   nom    libellé
  V1    Année
                                    1   AA_1   1987
                                    2   AA_2   1988
  V2    Pare
                                    1   AB_1   primipare
                                    2   AB_2   multipare
  V4    Traitement
                                    1   AD_1   témoin
                                    2   AD_2   niveau bas
                                    3   AD_3   niveau haut
```

7.2 Les individus

Le fichier des données comporte 90 individus, mais comme le taux protéique moyen n'est pas connu pour deux animaux, l'analyse de variance se déroule en deux temps. Un premier qui concerne la production laitière et le taux butyrique, c'est-à-dire les variables V8 et V9, un second qui prend en compte le taux protéique, la variable V10.

```
Les individus

Nombre d'individus dans le fichier de données :    90

Pour cause de données manquantes, les variables à expliquer ne peuvent
être traitées en même temps.
Il y a en effet  2 sous-fichiers homogènes.

+ sous-fichier  1, correspondant à 90 individus et relatifs à :
        V8   V9

+ sous-fichier  2, correspondant à 88 individus et relatifs à :
        V10
```

7.3 Le modèle

En plus des trois facteurs à effet simple déjà évoqués, une interaction d'ordre un entre le pare et le traitement (V2*V4) est introduite, car il est possible que le traitement ne réponde pas de la même manière avec les vaches primipares et les vaches multipares. De même une interaction entre l'année et le traitement (V1*V4) a été prise en compte.

```
Le modèle étudié :

  V8,
  V9,
  V10 = Mu + V1 + V2 + V4 + V1*V4 + V2*V4
```

Pour la production laitière, par exemple, le modèle linéaire s'écrit comme suit

- $V8=MU+V1+V2+V4+V1*V4+V2*V4+\varepsilon$

7.4 L'analyse de la variance

La suite concerne l'analyse de la variance proprement dite. La première variable traitée est la production laitière moyenne, la variable V8.

7.4.1 Le tableau de la régression multiple

Le premier tableau édité est celui de la régression multiple

```
Variable expliquée : V8    Lait brut

Tableau de la régression multiple
Source ddl   Sommes des carrés     Carrés moyens      F     P>F

Modèle   8        501.43652344       62.67956543   3.32  0.0026
Erreur  81       1529.54333496       18.88325119
Total   89       2030.97985840

         R²   Coef de variation Ecart-type erreur      V8 moyen

  0.246894          12.22309685        4.34548616   35.55143356
```

Ce tableau exhibe le pouvoir explicatif du modèle au travers du F de Fisher et du coefficient de corrélation multiple. Il fournit également le carré moyen résiduel et la moyenne brute de la variable expliquée.

7.4.2 Le tableau d'analyse de la variance

Après, vient le tableau de l'analyse de la variance qui décompose la somme des carrés des écarts du modèle fourni par le tableau précédent. Dans le cas présent il y en a en fait deux car il a été demandé le calcul des sommes des carrés des écarts de type I et de type III, c'est-à-dire deux analyses distinctes. Le modèle étant déséquilibré, il y a en effet deux fois plus de multipares que de primipares, ces deux tableaux diffèrent pour ce qui est des sommes des carrés des écarts, des carrés moyens et des F de Fisher. Seule la dernière ligne, celle qui correspond à l'interaction V2*V4, est la même dans les deux cas puisqu'avec les sommes des carrés des écarts de type I c'est le seul facteur à être ajustés de tous les autres.

```
Tableau de l'analyse de la variance
Source ddl     SCE type   I       Carrés moyens       F     P>F

V1       1       0.76381326          0.76381326    0.04  0.8356
V2       1      75.86797333         75.86797333    4.02  0.0458
V4       2     236.54899597        118.27449799    6.26  0.0031
V1*V4    2      31.35746002         15.67873001    0.83  0.4429
V2*V4    2     156.89582825         78.44791412    4.15  0.0189

Source ddl     SCE type III       Carrés moyens       F     P>F

V1       1       0.76380986          0.76380986    0.04  0.8356
V2       1      75.86801910         75.86801910    4.02  0.0458
V4       2     345.10998535        172.55499268    9.14  0.0003
V1*V4    2      31.35746765         15.67873383    0.83  0.4429
V2*V4    2     156.89587402         78.44793701    4.15  0.0189
```

Il ressort de ces deux tableaux que le facteur V2, le pare, et le facteur V4, le traitement, ont un effet respectivement significatif (probabilité de 0.0458) et hautement significatif (probabilité de 0.0031 avec les sommes des carrés de type I, de 0.0003 avec les sommes des carrés de type III). Cet effet du traitement est différent suivant le pare, l'interaction V2*V4 est significative (probabilité de 0.0189).

De ce fait, il n'est pas possible de conclure au vue de cette unique analyse de la variance et il faudra procéder à une analyse séparée pour les primipares et les multipares.

7.5 Les solutions

L'édition des solutions des équations normales est optionnelle. Dans le cas d'une analyse de la variance, l'examen des ces solutions offre peu d'intérêt car elles dépendent des contraintes supplémentaires introduites pour obtenir l'inverse généralisée de la matrice X'X. Elles sont donc biaisées. Ce tableau permet de constater que les contraintes posées consistent à mettre à zéro la solution correspondant à la dernière modalité de chacune des variables impliquées dans le facteur.

Pour les facteurs simples, la solution associée à un niveau de ce facteur correspond à l'écart entre celui-ci et le dernier niveau du facteur.

Les solutions des équations normales du modèle

Paramètre			Estimation		T(param.=0)	P>|T|	Ec-type estim
Mu			37.24946213	B	29.69	0.0000	1.25443387
V1	1		0.02605388	B	0.02	0.9841	1.58674741
	2		0.00000000	B	.	.	.
V2	1		1.32753372	B	0.79	0.4383	1.68299937
	2		0.00000000	B	.	.	.
V4	1		-1.02991486	B	-0.58	0.5703	1.77403724
	2		-2.39277649	B	-1.35	0.1777	1.77403712
	3		0.00000000	B	.	.	.
V1*V4	1	1	-1.20205963	B	-0.54	0.6002	2.24399948
	1	2	1.67664027	B	0.75	0.4635	2.24399948
	1	3	0.00000000	B	.	.	.
	2	1	0.00000000	B	.	.	.
	2	2	0.00000000	B	.	.	.
	2	3	0.00000000	B	.	.	.
V2*V4	1	1	-6.84204483	B	-2.87	0.0052	2.38012075
	1	2	-2.98355103	B	-1.25	0.2110	2.38012075
	1	3	0.00000000	B	.	.	.
	2	1	0.00000000	B	.	.	.
	2	2	0.00000000	B	.	.	.
	2	3	0.00000000	B	.	.	.

NB : Les solutions suivies de la lettre "B" ont des valeurs qui dépendent de la contrainte retenue pour calculer l'inverse généralisée de la matrice "X'X". Elles sont donc biaisées.

7.6 Test de facteurs

L'utilisateur a demandé que le facteur V4, le traitement, soit testé par rapport à l'interaction année*traitement (V1*V4)

```
TFACT FACT= V4 CREF= V1*V4 TSC= 3
```

Le résultat correspondant à cette requête est le suivant :

```
Tests des facteurs
Test du facteur   : V4

Le test suivant se réfère au carré moyen de type III associé à V1*V4

Source  ddl   SCE type III       Carrés moyens       F     P>F

V4        2   345.10998535        172.55499268   11.01  0.0835
```

Au vue de ce test, il s'avère que le traitement (V4) n'a pas d'effet significatif (probabilité de 0.0835) lorsque le carré moyen de référence est celui de l'interaction année*traitement.

7.7 Estimations

```
ESTIM NOM= 'Pare 2 - Pare 1' FACT= V2 COEF= -1 1
ESTIM NOM= 'Trait 2+3 - Trait 1' FACT= V4 COEF= -2 1 1 DIESTI= 2>
LEDIT= 1
```

Deux estimations ont été souhaitées. La première concerne le facteur V2, le pare, et correspond à la différence entre les primipares et les multipares. La seconde compare la moyenne des niveaux 2 et 3 du traitement au résultat obtenu pour le niveau 1 de ce traitement.

```
Estimations
Intitulé              Estimation   ec.type estim  t(E=0)  P>|t|

Pare 2 - Pare 1          1.94766         0.97168    2.00 0.0458
Trait 2+3 - Trait 1      3.52885         1.03062    3.42 0.0011
```

Comme il était permis de l'espérer, les multipares, toutes choses égales par ailleurs, donnent significativement plus de lait que les primipares (1.95 Kg). S'agissant du traitement, il s'avère qu'en moyenne les animaux recevant l'aliment étudié produisent plus (3.52 Kg).

Le test à zéro de ces estimations se réfère au carré moyen résiduel comme c'est le cas également pour le test des facteurs dans l'analyse de la variance. Cela explique que ,pour la différence entre primipares et multipares, la probabilité associée aux deux tests soit la même (0.0458).

7.8 Contrastes

L'utilisateur a souhaité tester la différence entre la moyenne les niveaux 2 et 3 du traitement et le niveau 1 de celui-ci en référence avec le carré moyen de l'interaction année*traitement (V1*V4).

```
CONTR NOM='Trait 2+3-Trait 1'FACT= V4 NDL= 1 COEF=-2 1 1>
      CREF= V1*V4 TSC= 3 LEDIT= 1
```

```
Les contrastes
Résultat du contraste .. : "Trait 2+3 - Trait 1"

Le test suivant se réfère au carré moyen de type III associé à V1*V4

Intitulé         ddl  SC contraste Carrés moyens      F     P>F

Trait 2+3-Trait 1  1  221.38291931  221.38291931  14.12  0.0004
```

La procédure fournit la somme des carrés des écarts associée au contraste, le carré moyen, le F de Fisher et la probabilité afférente. La valeur du contraste n'est pas indiquée, pour l'obtenir, il faut faire appel à une estimation.

7.9 Moyennes ajustées

7.9.1 Coefficients associés au calcul des moyennes ajustées

```
Moyennes ajustées

Coefficients associés au calcul des moyennes ajustées du facteur V4

V4                            1      2      3

                        Coefficients

Mu                        1.000  1.000  1.000

V1           1            0.500  0.500  0.500
             2            0.500  0.500  0.500

V2           1            0.500  0.500  0.500
             2            0.500  0.500  0.500

V4           1            1.000  0.000  0.000
             2            0.000  1.000  0.000
             3            0.000  0.000  1.000

V1*V4        1  1         0.500  0.000  0.000
             1  2         0.000  0.500  0.000
             1  3         0.000  0.000  0.500
             2  1         0.500  0.000  0.000
             2  2         0.000  0.500  0.000
             2  3         0.000  0.000  0.500

V2*V4        1  1         0.500  0.000  0.000
             1  2         0.000  0.500  0.000
             1  3         0.000  0.000  0.500
             2  1         0.500  0.000  0.000
             2  2         0.000  0.500  0.000
             2  3         0.000  0.000  0.500
```

Cette édition permet d'obtenir les coefficients par lesquels sont multipliées solutions des équation normale pour obtenir la moyenne ajustée de chaque niveau du facteur considéré.
Il est ainsi possible de comprendre, en examinant ce tableau, pourquoi, dans certains cas, une moyenne ajustée n'est pas estimable.

En effet ces coefficients doivent respecter la forme générale des fonctions estimable dont la forme peut être fournie par l'option LFES=1.

7.9.2 Simple édition des moyennes ajustées

```
Les tests suivants se réfèrent au carré moyen résiduel

Simple édition des moyennes ajustées

Variable expliquée : V8

V4             moyenne ajust        écart-type  t(M=0)   p>|t|

  1                 32.87429           0.84150   39.07  0.0000
  2                 34.88002           0.84150   41.45  0.0000
  3                 37.92625           0.84150   45.07  0.0000
```

S'il n'a pas été demandé de comparaisons multiples entre les moyennes ajustées des niveaux du facteur, outre ces moyennes ajustées, la procédure fournit l'écart-type de chacune d'elles , ainsi que leur test à zéro. Cet écart-type, pour un niveau donné, est bien entendu différent de l'écart-type des résultats obtenus par les individus appartenant à ce niveau qui pourrait être calculé par une procédure de description statistique.

7.9.3 Comparaisons multiples sans ajustement

```
Test de comparaisons multiples sans ajustement (PPDS)

Variable expliquée : V8
                                T pour H0moy.aju(i)=moy.aju(j)
                                P>|T|
V4            moyenne ajust i/j  1       2       3

  1                32.87429   1  .

  2                34.88002   2 1.6854   .
                                0.0918
  3                37.92625   3 4.2451  2.5597   .
                                0.0001  0.0119

Attention : les comparaisons multiples peuvent sous-estimer le risque
alpha !
```

Dans le cas où une comparaison multiple a été demandée sans que soit demandé le recours à une méthode d'ajustement pour contrôler le risque α (COMPM=0), le test de la PPDS est réalisé. Alors sont indiquées, en plus des moyennes ajustées, la valeur des tests de Student, de même que les probabilités associées.

7.9.4 Comparaisons multiples avec ajustement de Bonferroni

```
Test de comparaisons multiples avec l'ajustement de Bonferroni

Variable expliquée : V8
                        P>|T|  H0 : moy.aju(i)=moy.aju(j)
V1*V4 moyenne ajust i/j  1       2       3      4      5     6

 1  1      32.28629  1   .
 1  2      35.73137  2  0.5431   .
 1  3      37.93929  3  0.0151  1.0000   .
 2  1      33.46230  4  1.0000  1.0000  0.1124   .
 2  2      34.02868  5  1.0000  1.0000  0.2734  1.0000   .
 2  3      37.91323  6  0.0158  1.0000  1.0000  0.1172 0.2844 .
```

Pour les comparaisons multiples avec ajustement, ici celui de Bonferroni, la procédure fournit, pour chaque comparaison, la probabilité afférente au test d'égalité des moyennes ajustées des deux niveaux impliqués.

7.9.5 Comparaisons à une référence (test bilatéral de Dunnett)

```
Test de comparaisons multiples par rapport une référence (test bilatéral
de Dunnett)

Variable expliquée : V8
                     différence  P>|t| H0:moy.aju(i)=référence
V4   moyenne ajust   i   M(i)-référ        référence : V4  1

 1       32.87429    1
 2       34.88002    2      2.00573             0.1782
 3       37.92625    3      5.05196             0.0001
```

Si les comparaisons se font par rapport à un même niveau pris comme référence, l'écart entre la moyenne ajustée de chaque niveau et la moyenne ajustée du niveau de référence est donné en sus de la moyenne ajustée et de la probabilité liée au test de l'égalité des deux moyennes.

7.10 Les matrices X'X et X'Y

```
Matrices X'X et X'Y

Correspondance titres lignes/titres colonnes

lignes                 colonnes

Mu                   : Niv001

V1          1        : Niv002
            2        : Niv003

V2          1        : Niv004
```

Pour des raisons pratiques, les colonnes sont simplement désignées par leur numéro. La correspondance entre intitulé des lignes et intitulé des colonnes est donc indiquée.

			Niv001	Niv002	Niv003	Niv004	Niv005
Mu			90	45	45	30	60
V1	1		45	45	0	15	30
	2		45	0	45	15	30
V2	1		30	15	15	30	0
	2		60	30	30	0	60
V4	1		30	15	15	10	20
	2		30	15	15	10	20
	3		30	15	15	10	20
V1*V4	1	1	15	15	0	5	10
	1	2	15	15	0	5	10
	1	3	15	15	0	5	10
	2	1	15	0	15	5	10
	2	2	15	0	15	5	10
	2	3	15	0	15	5	10
V2*V4	1	1	10	5	5	10	0
	1	2	10	5	5	10	0
	1	3	10	5	5	10	0
	2	1	20	10	10	0	20
	2	2	20	10	10	0	20
	2	3	20	10	10	0	20
V8			3199.629	1603.960	1595.670	1027.590	2172.040
V9			3745.401	1852.220	1893.180	1236.640	2508.760

Suite de la matrice X'X

			Niv006	Niv007	Niv008	Niv009	Niv010

7.11 Forme générale des fonctions estimables

Forme générale des fonctions estimables

			Coefficients
Mu			L01
V1	1		L02
	2		L01-L02
V2	1		L04
	2		L01-L04
V4	1		L06
	2		L07
	3		L01-L06-L07
V1*V4	1	1	L09
	1	2	L10
	1	3	L02-L09-L10
	2	1	L06-L09
	2	2	L07-L10
	2	3	L01-L02-L06-L07+L09+L10

```
V2*V4          1  1     L15
               1  2     L16
               1  3     L04-L15-L16
               2  1     L06-L15
               2  2     L07-L16
               2  3     L01-L04-L06-L07+L15+L16
```

Le précédent tableau indique la forme que doit respecter une fonction pour que le résultat qu'elle donne soit estimable. Sa présentation est condensée, elle peut être développée pour l'expliciter.

		col 1	Col 2	Col 3	col 4	col 5	col 6	col 7	col 8
V1	1	0	L02	0	0	0	0	0	0
	2	L01	-L02	0	0	0	0	0	0
V2	1	0	0	0	L04	0	0	0	0
	2	L01	0	0	-L04	0	0	0	0
V3	1	0	0	0	0	0	L06	0	0
	2	0	0	0	0	0	0	L07	0
	3	L01	0	0	0	0	-L06	-L07	0

L01 désigne le coefficient qui , en première colonne, correspond à la ligne Mu. Ce coefficient se retrouve, toujours en première colonne, au niveau des lignes associées à la dernière modalité de chaque facteur.

L02 est le coefficient lié à la première modalité de V1 et à la deuxième colonne, celle qui correspond à cette première modalité. Ce coefficient apparaît, avec le signe moins, dans cette même colonne, au niveau de la ligne associée à la dernière modalité de la variable V1.

Le coefficient ne figure pas dans ce tableau dans la troisième colonne. Sa valeur est en effet nulle car cette colonne correspond à la dernière modalité de la variable V1.

Les coefficients de toute fonction doivent vérifier les règles définies par ce tableau, pour que celle-ci soit estimable.

7.12 Fonctions estimables

```
Fonctions estimables de type III

Fonction estimable de type III pour le facteur V1

                       Coefficients

Mu                      0

V1             1       +L02
               2       -L02

V2             1        0
               2        0

V4             1        0
               2        0
               3        0
```

```
V1*V4         1  1     +0.3333*L02
              1  2     +0.3333*L02
              1  3     +0.3333*L02
              2  1     -0.3333*L02
              2  2     -0.3333*L02
              2  3     -0.3333*L02

V2*V4         1  1      0
              1  2      0
              1  3      0
              2  1      0
              2  2      0
              2  3      0

Fonction estimable de type III pour le facteur V2

                       Coefficients

Mu                      0

V1            1         0
              2         0

V2            1        +L04
              2        -L04

V4            1         0
```

Ce tableau est élaboré suivant les principes évoqués plus haut à propos de l'édition de la forme générale des fonctions estimables. Il fournit, pour chaque facteur du modèle, et pour le type de sommes des carrés des écarts demandé, la valeur des coefficients de la fonction servant au calcul de la somme des carrés des écarts associée à ce facteur.

7.13 Coefficients associés au calcul d'un contraste

```
Coefficients associés au calcul du contraste : Trait 2+3 - Trait 1

                       Coefficients

Mu                      0

V1            1         0
              2         0

V2            1         0
              2         0

V4            1        -2.0000*L01
              2        +L01
              3        +L01

V1*V4         1  1     -L01
              1  2     +0.5000*L01
              1  3     +0.5000*L01
              2  1     -L01
              2  2     +0.5000*L01
              2  3     +0.5000*L01
```

```
V2*V4          1  1      -L01
               1  2      +0.5000*L01
               1  3      +0.5000*L01
               2  1      -L01
               2  2      +0.5000*L01
               2  3      +0.5000*L01
```

Le contraste testant l'écart entre la moyenne ajustée du niveau 1 du traitement et la moyenne des moyennes ajustées des niveaux 2 et 3 de ce traitement a été demandé. Ce tableau donne les coefficients de la fonction impliquée dans le calcul de ce contraste. Ces coefficients sont plus nombreux que ceux indiqués, par l'utilisateur, pour définir le contraste, car la variable V4 intervient également dans deux interactions, V1*V4 et V2*V4.

1. Présentation

1.1 Objet

Cette procédure permet de discriminer une variable nominale à n classes par une méthode neuronale. Cette méthode est basée sur l'utilisation d'un réseau de neurones multicouche. Elle est particulièrement adaptée à la prise en compte de relations non linéaires.

Le programme a été réalisé à partir d'une version préliminaire écrite par J. Proriol. Pour une description détaillée de ce programme, on consultera la référence suivante°:

> Proriol J. (1991) *MLP: Programme de réseau neuronal multicouche*. Revue de Modulad n° 8 pp 23-29

Un réseau comprend plusieurs couches, chaque couche étant constituée de neurones. Dans la couche d'entrée, il y a autant de neurones que de variables explicatives; dans la couche de sortie, il y a autant de neurones que de classes à discriminer. Les couches intermédiaires appelées couches cachées, sont des paramètres du réseau.

Le nombre de couches cachées et le nombre de neurones des couches cachées sont choisis par l'utilisateur. On prendra rarement plus de deux couches cachées. Un choix judicieux du nombre de neurones par couche cachée évitera la création d'une couche supplémentaire.

Les neurones d'une couche inférieure seront reliés au neurones d'une couche supérieure par des connexions appelées POIDS et chaque neurone de chaque couche est affectée d'un BIAIS. C'est le calcul de ces poids et biais qui permet l'affectation des individus dans les classes. Les poids et les biais sont pris au hasard lors de la première présentation des individus. Le paramètre LPOIDS permet d'utiliser des poids et des biais déjà calculés et de les améliorer.

Les poids et les biais seront ajustés alors par un apprentissage utilisant la méthode de *rétropropagation du gradient*. C'est un algorithme qui se divise en deux parties :

- *La propagation ou relaxation*

  On calcule les valeurs de chaque neurone avec le réseau de connexions calculé à l'itération t (on dit aussi au temps t). Les neurones de sortie seront combinaisons linéaires des neurones d'entrée. Poids et biais sont les coefficients de ces combinaisons linéaires.

- *La rétropropagation*

  Elle consiste à modifier les poids et les biais en fonction de l'erreur observée afin que les données d'entrée fournissent de meilleurs résultats.

L'estimation des poids est un processus itératif avec correction en fonction de l'erreur rencontrée. L'erreur est appelée *coût*. Elle est calculée après un certain nombre de présentations de l'échantillon d'apprentissage. Le nombre de présentations du fichier d'apprentissage est fixé par le paramètre NLECT. Un paramètre permet également d'arrêter la procédure quand la décroissance du coût devient très faible. L'apprentissage nécessite le choix du nombre de couches du réseau et du nombre de neurones dans ces couches. On peut également ajuster les paramètres epsilon et êta dont dépendent la correction apportée aux poids et biais à chaque itération. Il est possible de permuter les individus de l'échantillon.

Les résultats du classement peuvent être sauvegardés sur un fichier de type "NGRO". Deux variables figureront sur le fichier NGRO: la variable classe de sortie (à k modalités), et la variable croisant la classe de sortie et le fait d'être bien classé ou mal classé (à 2 x k modalités).

Remarque : dans cette procédure, les calculs sont réalisés en mémoire centrale de l'ordinateur , ce qui accélère les calculs mais peut apporter des contraintes sur la taille des données traitées.

Discriminante sur variables continues

Le choix de la variable de groupe et des variables explicatives est fait dans l'étape SELEC précédente.

La variable de groupe (les classes de sortie à estimer):

- sera déclarée NOMI ILL dans SELEC
- sera la seule variable NOMInale ILLustrative

Les variables explicatives seront déclarées CONT ACT dans SELEC dans le cas d'une analyse discriminante qui n'est pas précédée d'une analyse factorielle.

Discriminante sur coordonnées factorielles

L'algorithme pourra également être utilisé pour discriminer une variable à partir des coordonnées sur les axes factoriels d'une analyse effectuée au préalable. Le paramètre LVEC indique que l'on utilise des axes factoriels pour réaliser l'analyse.

La variable de groupe sera déclarée NOMInale ILLustrative dans l'étape SELEC précédant l'analyse factorielle. Ce sera la seule variable nominale illustrative présente.

Echantillon d'apprentissage, échantillon-test et anonymes

Il est possible de tester la qualité de la fonction discriminante sur des individus-tests n'ayant pas participé au calcul. On compare alors la classe de sortie obtenue avec la classe observée et on édite le pourcentage de bon classement.
Trois méthodes permettent d'effectuer des tests.

Au moment de l'apprentissage, on peut sélectionner des individus-tests

- par tirage au hasard dans l'échantillon d'apprentissage (paramètre PRCT)
- par déclaration en individus illustratifs dans la procédure SELEC précédant NEURO.

Après le processus d'apprentissage,

La procédure NEURO permet également de tester une fonction discriminante neuronale déjà calculée. Les individus actifs seront alors les individus-tests. Les paramètres de la fonction (poids, biais, nombre de couches, nombre de neurones, ...) sont lus sur le fichier NCOEF.

Les individus anonymes sont des individus dont on ignore la classe d'appartenance. La procédure NEURO permet de leur affecter des classes de sortie. Ces individus anonymes peuvent être sélectionnés par deux méthodes.

- soit les individus anonymes sont les individus actifs. L'affectation des individus anonymes est le seul calcul effectué. Les paramètres de la fonction discriminante calculés lors d'un apprentissage précédent sont alors lus sur NCOEF.
- soit les individus anonymes sont définis comme individus illustratifs dans la procédure SELEC précédente. Les individus actifs formeront alors soit l'échantillon d'apprentissage, soit un échantillon-test. Dans ce dernier cas, les paramètres de la fonction discriminante sont lus sur NCOEF.

1.2 Editions

Pour l'apprentissage:
La procédure édite les coûts et le tracé de la fonction coût, un tableau individuel des valeurs des neurones de sortie avec la classe d'entrée et la classe de sortie. La classe de sortie est la classe pour laquelle la valeur du neurone de sortie est la plus grande. La procédure édite également la matrice de confusion, le pourcentage de bien classés et la pureté de chaque classe, les histogrammes des valeurs des neurones de sortie.

Pour l'échantillon-test :
On obtient les sorties précédentes, à l'exception de celles relatives aux coûts.

Pour les anonymes :
On obtient le tableau des classes d'affectation et les neurones de sortie.

1.3 Paramètres

Les paramètres de la procédure se divisent en trois catégories:

- **Les paramètres de fonctionnement** indiquent le type des données sur lesquelles s'effectuent les calculs: coordonnées factorielles ou données brutes (LVEC). Ils permettent de définir des pondérations des classes (LPOND). Ils fixent le type de calcul à effectuer selon le statut des individus, apprentissage, test ou individus anonymes (LEVAL, LILL, PRCT).
- **Les paramètres de définition du réseau de neurones** indiquent le nombre de couches cachées (NCACH) et le nombre de neurones par couche cachée (NNEUR). Ces paramètres n'ont d'effet que pour le processus d'apprentissage.
- **Les paramètres d'ajustement des calculs** permettent la permutation de l'échantillon d'apprentissage avant le processus (LPERM), ou en cours de calcul (NPERI). Ils définissent l'arrêt du processus par nombre d'itérations (NLECT) ou seuil de décroissance du coût (LSTOP). Ils règlent la fonction d'ajustement (TEMPE, ETAVA, EPSIN, LEPSI).
- **Les paramètres d'édition** (LHIST et LEDIT) gèrent les éditions des résultats des calculs et des affectations des individus dans les groupes.

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que tous les paramètres prennent leur valeur par défaut, on codera le mot-clef NOPAR à la place de la liste des paramètres repérée par (3).

(1) PROC NEURO — Analyse discriminante neuronale

(2) — *titre de l'analyse*

(3) LEVAL (0 ou APP) — statut des individus actifs
- 0 ou APP : échantillon d'apprentissage
- 1 ou TEST : échantillon-test
- 2 ou ANO : individus anonymes

LILL (0 ou NON) — statut des individus illustratifs
- 0 ou NON : Pas utilisés
- 1 ou TEST : Individus tests
- 2 ou ANO : Individus anonymes

LVEC (0 ou NON) — discriminante à partir de coordonnées factorielles.
- 1 ou OUI
- 0 ou NON

PRCT (0.) — pourcentage d'individus tirés au hasard dans l'échantillon d'apprentissage pour constituer un échantillon-test

NCACH (1) — nombre de couches cachées

NNEUR (5) — nombre de neurones par couche cachée
- LIST : liste de nombres

LSTOP (1 ou OUI) — arrêt si décroissance relative du coût <0.05%

NLECT (30) — nombre maximum de présentation de l'échantillon d'apprentissage.

LPOID (0 ou TIRE) poids de départ
0 ou TIRE : tirés au hasard
1 ou LU : lus sur le fichier NCOEF

LNORM (0 ou NORM) transformation initiale des variables
0 ou NORM : variables centrées réduites
1 ou MINMAX : valeurs mises entre -1 et 1 sans normalisation

LPOND (1 ou UNIF) pondération des classes de sortie
1 ou UNIF : équipondérées
2 ou LIST : liste de pondération des classes

LPERM (0 ou NON) permutation au hasard des individus avant apprentissage.
1 ou OUI : permutation
0 ou NON : pas de permutation

NTIRA (0 ou NON) périodicité de permutation des individus .

TEMPE (.5) réglage de la fonction permettant de cadrer les données entre 0.1 et 1.

NPERI 10) périodicité du calcul du coût.

ETAVA (.5) poids de la correction au temps t-1 dans la correction au temps t

EPSIN (0.5) importance de la correction à apporter à chaque itération.

LEPSI (0 ou NON) réduction de EPSIN au cours de l'apprentissage.
0 ou NON : pas de réduction de EPSIN
1 ou OUI : réduction de EPSIN

LHIST (0 ou NON) édition des histogrammes des valeurs de sortie
0 ou NON : pas d'édition
1 ou OUI : édition

LEDIT (0 ou NON) édition des valeurs de sortie et des classes d'affectation des individus.
0 ou NON : pas d'édition
1 ou OUI : édition

(4) *(Si LVEC = OUI)Liste de sélection des axes factoriels*
(5) *(Si LPOND = LIST) Liste des pondérations (coûts) pour chaque classe*
(6) *(Si NNEUR = LIST) Liste des nombres de neurones par couche cachée*

3. Présentation détaillée des paramètres

LEVAL **statut des individus actifs**

- *valeurs possibles* : 0 ou APP (échantillon d'apprentissage)
  1 ou TEST (échantillon-test)
  2 ou ANO (individus " anonymes ")
- *valeur par défaut* : 0 ou APP

Ce paramètre conditionne le type de calcul qui sera effectué sur les individus actifs.

- Si LEVAL = APP, les individus actifs forment l'échantillon d'apprentissage. Ils servent à calculer la fonction discriminante selon la méthode neuronale. Les paramètres de cette fonction, appelés poids et biais, peuvent être conservés sur le fichier NCOEFB. Ils serviront ultérieurement, soit à continuer le processus d'apprentissage, soit à faire des tests, soit à affecter des individus anonymes à des classes. Les individus illustratifs pourront être considérés comme individus-tests ou individus anonymes selon la valeur du paramètre LILL.
- Si LEVAL = TEST, les individus actifs sont utilisés pour tester une fonction discriminante déjà calculée. Les poids et biais estimés lors du processus d'apprentissage sont lus sur le fichier NCOEF. La procédure calcule les classes estimées et les compare aux classes réelles. La variable représentant les classes ne devra comporter aucune donnée manquante. Les individus illustratifs pourront être considérés comme individus-anonymes.
- Si LEVAL = ANO, les individus actifs sont des individus "anonymes", dont on ne connaît pas la classe de sortie. La procédure calcule les classes estimées.

Dans tous les cas, les classes de sortie estimées sont conservées sur le fichier NGRO.

LILL **statut des individus illustratifs**

- *valeurs possibles* : 0 ou NON (pas utilisés)
  1 ou TEST (échantillon-test)
  2 ou ANO (individus anonymes)
- *valeur par défaut*: 0 ou NON

Ce paramètre conditionne le type de calcul qui sera effectué sur les individus illustratifs. Les valeurs possibles dépendent du paramètre LEVAL. La classe d'affectation sera conservée sur le fichier NGRO.

- Si LILL = NON, les individus illustratifs ne seront pas pris en compte dans les calculs.

- Si LILL = TEST, les individus illustratifs serviront d'échantillon-test. Ceci n'est possible que si LEVAL = APP.
- Si LILL = ANO, les individus illustratifs seront considérés comme anonymes et on calculera leur classe d'affectation. Ceci est possible si LEVAL = APP ou LEVAL = TEST.

LVEC **calcul sur coordonnées factorielles**

- *valeurs possibles* : 0 ou NON (calcul sur données brutes)
  1 ou OUI (calcul sur coordonnées factorielles)
- *valeur par défaut* : 0 ou NON

- Si LVEC = OUI, les neurones d'entrée seront les coordonnées factorielles d'une analyse, lues sur le fichier NGUS. L'analyse factorielle sera une analyse des correspondances multiples (CORMU), une analyse en composantes principales (COPRI), ou une analyse des correspondances simples (CORBI). On pourra en particulier réaliser une analyse discriminante à partir de variables nominales, après une analyse des correspondances multiples. La variable définissant les classes aura été déclarée nominale illustrative dans l'analyse factorielle. Ce sera la seule variable nominale illustrative.
- Si LVEC = NON, les neurones d'entrée seront les variables CONTinues ACTives de l'étape SELEC précédant NEURO. La variable définissant les classes sera la seule variable nominale illustrative dans SELEC.

PRCT **Pourcentage d'individus-tests dans l'échantillon d'apprentissage**

- *valeurs possibles* : 0. à 99.
- *valeur par défaut* :0.

Ce paramètre n'a de sens que quand LEVAL = APP, c'est-à-dire quand les individus actifs servent à l'apprentissage de la fonction discriminante. Il est alors possible de sélectionner par tirage au hasard un pourcentage PRCT d'individus dans l'échantillon d'apprentissage, c'est-à-dire parmi les individus actifs. Ces individus seront exclus du processus d'apprentissage et serviront à tester la capacité de prévision de la fonction discriminante.

Une autre méthode de sélection d'un échantillon-test est de déclarer les individus illustratifs et de préciser LILL = TEST. Les deux méthodes sont compatibles.

Si PRCT = 0., il n'y a pas tirage d'individus-tests dans l'échantillon d'apprentissage.

NCACH **nombre de couches cachées**

- *valeurs possibles*: 1 à 5
- *valeur par défaut* :1

La procédure autorise jusqu'à 5 couches cachées. Il est toutefois souhaitable de ne pas dépasser 2 couches cachées, à la fois pour des raisons d'efficacité de l'algorithme et de réservation de mémoire centrale.

NNEUR **nombre de neurones dans la couche cachée, s'il n'y a qu'une couche cachée**

- *valeurs possibles* : 1 à 100
- *valeur particulière* : LIST (Liste de nombres de neurones si NCACH>1)
- *valeur par défaut* : 5

  - Si NNEUR > 0 et NCACH = 1, le réseau aura une couche cachée dont le nombre de neurones est NNEUR.
  - Si NNEUR = LIST, l'utilisateur précisera le nombre de neurones de chaque couche cachée dans une liste.

LSTOP **décroissance minimale du coût**

- *valeurs possibles* : 1 ou OUI (arrêt de l'apprentissage si la décroissance relative du coût est inférieure à .05%)
  0 ou NON (arrêt de l'apprentissage au bout de NLECT itérations)
- *valeur par défaut* :0 ou NON

Le coût est une mesure des écarts entre les classes réelles et les classes prévues pour l'ensemble de l'échantillon. Ce paramètre permet d'arrêter la procédure d'apprentissage quand la décroissance relative du coût devient inférieure à 0.05.

NLECT **nombre de présentations de l'échantillon d'apprentissage**

- *valeurs possibles* : positives
- *valeur par défaut* : 100

Le programme effectue au plus NLECT présentations de l'échantillon d'apprentissage pour calculer les poids. Ce paramètre permet d'arrêter la procédure d'apprentissage, même si la décroissance relative du coût n'est pas inférieure à 0.05%.

LPOID **initialisation des poids et des biais**

- *valeurs possibles* : 0 ou TIRE (tirage au hasard)
  1 ou LU (poids et biais lus sur le fichier NCOEF)
- *valeur par défaut* : 0 ou TIRE

  - Si LPOID = TIRE, lors de la présentation du premier individu de l'échantillon d'apprentissage, les poids et les biais sont tirés au hasard .
  - Si LPOID = LU, on commence la phase d'apprentissage avec les poids et les biais calculés précédemment lors de la présentation d'un autre échantillon d'apprentissage. Ces poids et ces biais sont lus sur le fichier NCOEF.

Ce paramètre permet de ne présenter qu'une petite partie de l'échantillon d'apprentissage et de ne continuer les calculs que si le processus d'affectation n'est pas suffisamment précis. Le processus d'apprentissage étant assez long et nécessitant une place importante en mémoire centrale, il peut être intéressant de scinder les opérations. Cela permet également de représenter des individus mal classés pour ajuster la fonction discriminante.

LNORM **méthode de normalisation des variables**

- *valeurs possibles* : 0 ou NORM (centrage-réduction des variables et application de la fonction sigmoïde)
  1 ou MINMAX (utilisation du minimum et du maximum pour amener les valeurs entre -1 et 1)
- *valeurs par défaut* : 0 ou NORM

L'algorithme suppose que les valeurs des neurones d'entrée soient comprises entre -1 et +1. La procédure propose 2 méthodes.

  - Si LNORM = NORM, on centre et on réduit chacune des variables explicatives, puis on utilise la fonction sigmoïde pour mettre les valeurs entre -1 et +1.
  - Si LNORM = MINMAX, les variables sont mises entre -1 et +1 par la formule :
    {Valeur - (max+min)/2 } / { (max - min)/2 }.

LPOND **pondération des classes de sortie**

- *valeurs possibles* : 1 ou UNIF (équipondérées)
  2 ou LIST (pondération des classes)
- *valeur par défaut* :1 ou UNIF

Les pondérations entrent en jeu dans le calcul des corrections à apporter aux paramètres de la fonction calculée. A chaque itération, la correction tiendra d'autant plus compte de l'erreur de mauvais classement dans la classe de sortie que la pondération affectée à cette classe sera forte.

- Si LPOND = UNIF, les coefficients de pondération de chaque classe seront égaux à 1.
- Si LPOND = LIST, l'utilisateur précisera à la suite des paramètres une liste de coefficients de pondération. Cette liste contiendra autant d'éléments qu'il y a de classes de sortie.

LPERM **paramètre de permutation des individus**

- *valeurs possibles* : 1 ou OUI (permutation de l'échantillon d'apprentissage au hasard au début du processus)
  0 ou NON (pas de permutation)
- *valeur par défaut* : 0 ou NON

Le processus d'apprentissage étant dépendant de l'ordre dans lequel on présente les individus, il est possible de permuter au hasard les individus de l'échantillon d'apprentissage avant de commencer le processus.

NTIRA **périodicité de permutation de l'échantillon d'apprentissage**

- *valeurs possibles* : de 1 à NLECT
- *valeur particulière* : 0 ou NON (aucune permutation)
- *valeur par défaut* : 0 ou NON

Pour rendre les calculs indépendants de l'ordre de présentation de l'échantillon d'apprentissage, ou du moins pour en diminuer les conséquences, il est proposé de permuter les individus au cours de l'apprentissage. NTIRA représente le nombre de lectures de l'échantillon d'apprentissage entre deux permutations au hasard. Par exemple, si NLECT = 100 et NTIRA = 20, l'échantillon d'apprentissage sera lu au plus 100 fois et son ordre sera permuté toutes les 20 lectures.

Si NTIRA = NLECT (nombre maximum de lectures de l'échantillon), il n'y a pas réordonnancement de l'échantillon d'apprentissage au cours du processus.

TEMPE **"température" de la fonction sigmoïde**

- *valeurs possibles* : réelles entre 0.1 et 1.0
- *valeur par défaut* : 0.5

Sur chaque couche, les valeurs calculées des neurones sont mises entre -1 et +1 par la fonction sigmoïde. Au moment de la rétropropagation, on utilise la dérivée de cette fonction pour effectuer les corrections. Cette dérivée dépend de la valeur de TEMPE. Si TEMPE est proche de 1, les corrections seront fortes, l'apprentissage rapide, mais avec le risque d'une estimation insuffisamment précise. Si TEMPE est proche de 0, les corrections seront faibles, l'apprentissage lent, mais avec moins de risques d'oscillations autour de la bonne valeur.

Rappelons que la fonction sigmoïde est:

(1 - EXP(-X / t))/(1 + EXP(-X / t))

TEMPE est la température t de cette fonction

NPERI **périodicité du calcul du coût**

- *valeurs possibles* : 1 à NLECT
- *valeur par défaut* : 10

NPERI est le nombre de présentations de l'échantillon d'apprentissage entre chaque calcul du coût. Si NPERI = 10, le coût est calculé chaque fois que 10 itérations sont faites. Le fait de ne pas calculer le coût à chaque présentation permet d'économiser du temps de calcul.

ETAVA **constante agissant lors de la correction des poids et biais**

- *valeurs possibles* : réelles entre 0.0 et 1.0
- *valeur par défaut* : 0.5

A la correction des poids et biais calculée au temps t, on ajoute la correction des poids obtenue au temps t-1 multipliée par ETAVA. ETAVA est appelé pas du gradient. Il permet d'augmenter la correction donc la vitesse de convergence de l'algorithme.

- Si le pas est faible les variations des poids et biais seront faibles.
- Si le pas est très fort, les variations des poids et biais seront très importantes. Voulant se rapprocher d'un état d'équilibre, on risque alors de le dépasser. On risque ainsi soit la divergence, soit l'oscillation du réseau. En particulier, si le coût augmente au cours du processus, il est conseillé de réduire ETAVA.

EPSIN **importance de la correction apportée à chaque itération**

- *valeurs possibles* : réelles entre 0.0 et 1.0
- *valeur par défaut* : 0.5

EPSIN est le facteur multiplicatif appliqué à l'erreur pour calculer la correction des poids et biais. Comme ETAVA, EPSIN guide la rapidité des variations et la convergence des résultats. Si le coût augmente au cours du processus, EPSIN est le premier paramètre à faire varier. Il est conseillé de diviser alors sa valeur par 10.

LEPSI **réduction de EPSIN au cours de l'apprentissage**

- *valeurs possibles* : 1 ou OUI (réduction de EPSIN)
  0 ou NON (EPSIN constant)
- *valeur par défaut* : 0 ou NON

Ce paramètre permet de diminuer l'importance de la correction au fur et à mesure de l'apprentissage. Quand on s'approche de la fonction à estimer, cette méthode a pour but d'éviter l'oscillation du réseau.

LHIST **édition des histogrammes des valeurs de sortie**

- *valeurs possibles* : 1 ou OUI (édition des histogrammes)
  0 ou NON (pas d'édition)
- *valeur par défaut* : 0 ou NON

Les valeurs de chaque neurone de la couche de sortie fournissent une mesure de la qualité de la prévision. Ces valeurs sont comprises entre -1 et 1. La classe de sortie est le neurone dont la valeur est la plus élevée. Une situation idéale serait la valeur 1 pour le neurone de la classe de sortie (et c'est la bonne classe) et -1 pour les autres neurones.

LEDIT **édition des classes de sortie et des valeurs de sortie pour chaque individu**

- *valeurs possibles* : 1 ou OUI (édition des classes)
  0 ou NON (pas d'édition)
- *valeur par défaut* : 0 ou NON

Si LEDIT = OUI, la procédure édite, pour chaque individu, la classe d'origine, la classe de sortie et les valeurs des neurones de sortie.

4. Liste de sélection des axes factoriels (si LVEC = OUI)

Syntaxe

FACT *liste de facteurs*

Si l'on effectue une analyse discriminante neuronale à partir des coordonnées factorielles d'une analyse effectuée préalablement, (LVEC = OUI), on indique ici la liste des facteurs qui formeront la couche d'entrée.

Exemple:

```
FACT F1 -- F5, F7 -- F10
```

Les neurones d'entrée seront les coordonnées des axes factoriels de 1 à 5 et de 7 à 10.

5. Liste des pondérations (coûts) par classe (si LPOND = LIST)

Syntaxe

POND *liste de réels*

Il est possible d'attribuer des pondérations à chacune des classes de sortie, c'est à dire de considérer que les coûts de mauvais classement ne sont pas les mêmes selon les classes. Si LPOND=LIST, l'instruction POND permet de donner la liste de ces pondérations. Elle doit contenir autant de nombres réels qu'il y a de modalités dans la variable à discriminer.

Exemple: si la variable à discriminer a 4 modalités, on pourra écrire

```
POND  .35 .35 .20 .10
```

6. Liste des nombres de neurones par couche cachée (si NNEUR = LIST)

Syntaxe

NEUR *liste de numéros*

Les paramètres principaux du réseau sont les nombres de couches et les nombres de neurones de chaque couche. Ce sont les principaux éléments à faire varier pour ajuster le modèle de fonction discriminante.

Si l'on ne désire qu'une couche cachée, le paramètre NNEUR permet d'en indiquer le nombre de neurones. Si l'on désire plusieurs couches cachées, on indique le nombre de couches par le paramètre NCACH. L'instruction NEUR permet alors de préciser le nombre de neurones dans chaque couche cachée.

Exemple:
Si l'on souhaite tester un modèle avec deux couches cachées, la première de 4 neurones et la deuxième de 6 neurones, on écrira:

```
NEUR 4 6
```

7. Exemples de commande

7.1 Premier exemple

Cet exemple montre un processus d'apprentissage pour la discrimination d'une variable nominale en fonction de 6 variables continues.

```
----+----1----+----2----+----+----3----+----4----+----5----
PROC SELEC
SELECTION DES VARIABLES
NOPAR
CONT ACT 7--12  : VARIABLES EXPLICATIVES
NOMI ILL 4   : VARIABLE A DISCRIMINER
FIN

NCOEFB='result'
PROC NEURO
ANALYSE DISCRMINANTE NEUROLALE
LEVAL = APP PRCT = 20. NNEUR = 7 NCACH = 1 NLECT = 50
----+----1----+----2----+----+----3----+----4----+----5----
```

La variable à discriminer est la variable 4. Les variables explicatives sont les variables 7, 8, 9, 10, 11 et 12. LEVAL = APP indique qu'il y a processus d'apprentissage. PRCT = 20. indique que 20% des individus sont tirés au hasard pour former l'échantillon-test. NCACH = 1 indique que le réseau n'a qu'une couche cachée et NNEUR = 7 précise qu'elle se compose de 7 neurones. NLECT = 50 demande l'arrêt du processus d'apprentissage après 50 itérations.

Notons que les poids de départ sont tirés au hasard (LPOID = TIRE par défaut), qu'il y aura permutation des individus au hasard avant le premier passage (LPERM = OUI par défaut), mais qu'il n'y aura pas de nouvelle permutation au cours du processus (NTIRA = 0 par défaut).
Il n'y aura pas d'édition des histogrammes des valeurs de sortie (LHIST = NON par défaut), ni des valeurs de sortie et des classes d'affectation des individus (LEDIT = NON par défaut).
Les seules éditions seront donc la matrice de confusion et les pourcentages de bien classés et mal classés.

7.2 Deuxième exemple

Dans cet exemple, un calcul a déjà été réalisé et il n'y a aucun processus d'apprentissage. On affecte des classes de sortie à un échantillon-test ainsi qu'à des individus anonymes. Les poids et biais, paramètres de la fonction discriminante calculés et conservés lors d'un passage précédent sur le fichier NCOEFB, sont lus sur le fichier NCOEF.

```
----+----1----+----2----+----+----3----+----4----+----5----
PROC SELEC
SELECTION DES VARIABLES
LSELI = FILT
CONT ACT 7--12  : VARIABLES EXPLICATIVES
NOMI ILL 4   : VARIABLE A DISCRIMINER
FIN
V4 > 0

NCOEF = 'result'
PROC NEURO
ANALYSE DISCRMINANTE NEUROLALE
LEVAL = TEST LILL = ANO LPOID = LU LEDIT = OUI
----+----1----+----2----+----+----3----+----4----+----5----
```

La variable de groupe est la variable 4. Les variables continues explicatives sont les variables 7 à 12. Ces variables doivent évidemment correspondre aux variables sélectionnées dans le processus d'apprentissage effectué précédemment.

Les individus actifs de la sélection (pour lesquels V4>0) seront considérés comme des individus-tests (LEVAL = TEST) et on éditera les pourcentages de bien classés dans chacune des classes.

Les individus illustratifs (pour lesquels V4 = 0) seront considérés comme des individus anonymes (LILL = ANO). On notera que tous les paramètres de définition et de fonctionnement du réseau n'ont aucun sens puisqu'il n'y a pas de processus d'apprentissage. Ils sont lus sur le fichier NCOEF.

Pour les individus-tests comme pour les individus-anonymes, la valeur des neurones de sortie et la classe d'affectation seront listés (LEDIT = OUI).

7.3 Troisième exemple

Dans cet exemple, on effectue une analyse discriminante neuronale à partir des coordonnées factorielles d'une analyse des correspondances multiples. La procédure NEURO est donc précédée d'une procédure SELEC et d'une procédure CORMU.

```
----+----1----+----2----+----+----3----+----4----+----5----
PROC SELEC
=== SELECTION DES VARIABLES =
LZERO  =  REC
NOMI ACT 29--34
NOMI ILL 4
FIN
```

```
PROC CORMU
== ANALYSE DES CORRESPONDANCES MULTIPLES ===
NOPAR

PROC NEURO
ANALYSE DISCRIMINANTE NEURONALE SUR FACTEURS
LEVAL = APP LVEC = OUI NNEUR = LIST NCACH = 3 LPOND = LIST
FACT F1--F10, F15
POND .4 .2 .1 .2 .1
NEUR  4 7 8
----+----1----+----2----+----+----3----+----4----+----5----
```

Les variables explicatives sont les variables nominales actives 29 à 34. La variable à discriminer est la variable 4. Il y a processus d'apprentissage (LEVAL = APP). Les neurones d'entrées sont des coordonnées factorielles (LVEC = OUI). Les facteurs choisis sont précisés par l'instruction FACT.

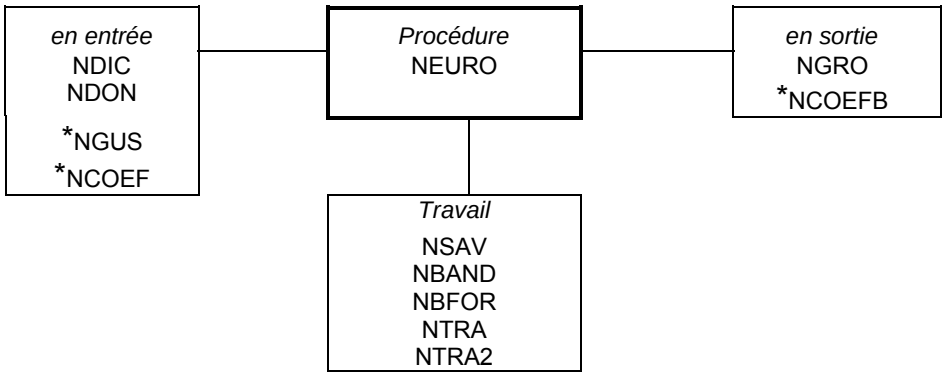
Ce sont les dix premiers axes factoriels et le quinzième. Le réseau de neurones compte 3 couches cachées (NCACH = 3). Le nombre de neurones de chaque couche cachée est précisé dans l'instruction NEUR: 4 pour la première couche, 7 pour la deuxième et 8 pour la dernière.

Des pondérations sont à prendre en compte dans les calculs d'erreur (LPOND = LIST). Ces pondérations sont précisées dans l'instruction POND. On y indique 5 réels inférieurs à 1 (la variable 4 présente 5 modalités).

8. Fichiers nécessaires à l'exécution

- en lecture NDIC (dictionnaire des variables utiles)
  NDON (données utiles)
  NGUS (coordonnées factorielles si LVEC = OUI)
  NCOEF (contient les poids et les biais définissant les connexions, si LPOID = LU)

- en écriture NGRO (2 partitions: classes de sortie et bien ou mal classé)
  NCOEFB (poids et biais calculés si LEVAL = APP)



1. Présentation

Les calculs et les éditions de résultats proposés dans cette procédure sont inspirés des travaux réalisés par J-M. Plumyène, du Groupe de Recherche Opérationnelle du Crédit Lyonnais. L'écriture de la procédure SCORE a grandement bénéficié de son expérience et de ses conseils.

1.1 Objet

La procédure SCORE est exécutée après une analyse discriminante concernant deux groupes (proc DIS2G). Elle calcule une *fonction de score* qui est une modification de la fonction discriminante, destinée à faciliter son interprétation et son utilisation. On introduit en particulier des "*zones*" de décision (rouge, verte, centrale) à partir d'une *tolérance d'erreur de classement*, et on édite divers résultats permettant d'apprécier les performances de la discrimination. Dans la suite, les deux groupes à discriminer sont appelés:

- groupe des scores "*forts*" ou groupe "1"
- groupe des scores "*faibles*" ou groupe "2".

C'est l'utilisateur qui définit en entrée la correspondance entre les modalités 1 et 2 de la variable de groupe, et les attributs "*fort*" et "*faible*" des scores.

La fonction discriminante fournie par la procédure DIS2G doit avoir été calculée préalablement sur des variables nominales, c'est à dire en fait sur les axes factoriels résultant d'une analyse des correspondances multiples (analyse discriminante "qualitative"). Le passage des résultats entre DIS2G et SCORE est réalisé au moyen d'un fichier contenant les coefficients (fichier de type NCOEF).

L'utilisateur a cependant la possibilité d'introduire lui-même en paramètre l'ensemble des coefficients d'une fonction linéaire discriminante calculée par ailleurs (par exemple, des estimations "bootstrap").

La procédure SCORE transforme les coefficients en utilisant les deux règles suivantes:

- Coefficient minimum de chaque variable: pour chaque variable nominale, le plus petit coefficient est mis à la valeur 0. Ainsi le score minimum possible d'un individu est 0. Il est obtenu pour un individu qui, *pour chaque variable,* possèderait la modalité affectée du coefficient transformé 0.

- Maximum possible de la fonction score: la valeur du *score maximum* possible est un choix de l'utilisateur (par exemple 1000). Ce maximum correspond à la somme des coefficients transformés les plus grands dans chaque variable.

Le score attribué à un individu s'obtient en additionnant les coefficients transformés associés aux modalités de l'individu. La fonction de score transformée classe les individus de la même façon que la fonction discriminante initiale.

D'autre part la procédure SCORE peut être utilisée pour tester le comportement d'un jeu de coefficients transformés, modifiés manuellement par l'utilisateur. En général il s'agira des coefficients transformés fournis par une première exécution de la procédure, et "arrondis" à des valeurs proches pour en faciliter l'utilisation.

L'utilisateur peut définir un taux appelé *"tolérance d'erreur de classement"* noté TEC. Ce taux permet de calculer des zones sur l'échelle de la fonction de score:

- une **zone "verte"**, qui correspond à la région des scores élevés, où l'on s'attend à trouver la plupart des individus du groupe 1. Dans cette zone, un mal classé est un individu du groupe 2 affecté d'un score "trop" élevé.La frontière est calculée pour contenir un taux de mal classés ne dépassant pas TEC.

- une **zone "rouge"** du côté des scores faibles, contenant essentiellement des individus du groupe 2, donc bien classés, et un pourcentage ne dépassant pas TEC d'individus du groupe 1, donc mal classés.

- une **zone intermédiaire**, entre les frontières des zones rouge et verte, où l'affectation à un groupe est laissée indécise. Cette zone d'indécision diminue quand l'utilisateur augmente la tolérance de mauvais classement TEC.

La procédure SCORE crée un fichier de type "NGUS" qui contient, pour chaque individu, la valeur de son score transformé (fonction de score calculée par la procédure). On peut ainsi archiver les scores (par la procédure ESCAL) et procéder ultérieurement à des analyses, tabulations, calculs de moyennes ou graphiques.

1.2 Editions

Coefficients transformés

On édite les coefficients de la fonction discriminante initiale et, en parallèle, les coefficients transformés, ou coefficients de la fonction de score. Le tableau permet en particulier le calcul du score de tout nouvel individu par addition des coefficients qui le concernent.

Le tableau des coefficients est édité une seconde fois, les variables étant rangées dans l'ordre de leur participation maximale à la fonction de score. Pour chaque variable, les modalités sont également rangées dans l'ordre décroissant de leur contribution au score.

La normalisation des coefficients permet d'apprécier la contribution de chaque modalité dans le calcul d'un score. On utilise ces valeurs également pour comparer deux fonctions de score.

Pondération des groupes

Il arrive que l'échantillon utilisé pour calculer la fonction discriminante (et la fonction de scores) ne soit pas "représentatif" de la population au sens suivant: la répartition des deux groupes dans l'échantillon n'est pas à l'image de la répartition dans la population globale. En effet, on est parfois amené à équilibrer ou égaliser les effectifs de chaque groupe pour le calcul de la fonction discriminante.

Dans ce cas l'utilisateur doit annoncer quelle est la répartition réelle (paramètre PRG2). On édite alors le poids à affecter aux individus d'un groupe pour que l'échantillon soit représentatif de la population (tableau "Redressement de l'échantillon").

On notera que ce redressement, qui concerne un groupe par rapport à l'autre, est indépendant d'un éventuel poids attribué à chaque individu de l'échantillon au moment de la discrimination. Une pondération peut en effet être choisie lors de la procédure SELEC (paramètre IMASS); elle opérera dans le calcul de la fonction discriminante (procédure DIS2G) et sera transmise dans la procédure SCORE: les effectifs publiés dans SCORE sont alors des effectifs pondérés.

Distributions dans les zones de scores

En fonction du taux d'erreur de classement demandé (TEC), la procédure fournit une courbe des seuils des zones Rouge, Verte et intermédiaire en fonction du TEC. Cette courbe permet de visualiser l'étendue de la zone d'indécision en fonction de la tolérance d'erreur. Un tableau récapitulatif donne la répartition des échantillons dans les 3 zones correspondant au TEC choisi.

On édite ensuite la distribution des deux groupes en fonction d'un découpage des scores en tranches. Un autre tableau concerne la répartition de la population, éventuellement redressée si un des groupes est pondéré.

Ce tableau contient également une estimation du "*rapport d'efficacité*" de la fonction de score. Pour chaque tranche de scores, on calcule le nombre d'individus du groupe des scores forts pour un individu du groupe des scores faibles. Lorsque les tranches de scores sont de petite amplitude, ce rapport sera en général très variable. Il est stabilisé par lissage.

La procédure SCORE édite des graphiques par tranche de valeurs des scores, où apparaissent les frontières des zones rouge et verte. Un graphique donne la répartition de la population (éventuellement redressée si un groupe est pondéré) en fonction des tranches de scores. Un graphique analogue concerne les échantillons des deux groupes. Un troisième graphique concerne la courbe du rapport d'efficacité.

Probabilités conditionnelles

Pour chaque valeur du score, on calcule la probabilité conditionnelle d'appartenir au groupe des scores faibles sous les hypothèses classiques de l'analyse discriminante (en particulier: normalité, égalité des covariances).
Cette probabilité, qui est forte du coté des scores faibles, est une fonction décroissante du score. La courbe est éditée, et contient au moins la zone d'indécision.

1.3 Paramètres

Les paramètres de la procédure se divisent en trois catégories:

- **Les paramètres relatifs aux données** permettent de choisir la variable de groupe (NUMGR), de sélectionner la modalité qui doit correspondre aux scores faibles (MOG2) et de fixer le pourcentage de redressement des groupes (PRG2).

- **Les paramètres d'édition.** Le maximum de la fonction score SXTOT et le nombre de classes de scores (NCLAS) utilisées pour les graphiques, sont essentiellement des paramètres d'édition.

- **Les paramètre des options.** LCOEF gère l'entrée des coefficients de la fonction discriminante. La tolérance d'erreur de classement TEC permet de contrôler la zone d'indécision et LABAQ détermine le mode de calcul des probabilités conditionnelles.

1.4 Après les paramètres

Après la liste des paramètres, on placera la liste des coefficients de la fonction score dans les cas LCOEF = 2 et 3. Cette liste est terminée par le mot-clef FIN.

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre.

(1) PROC SCORE Analyse d'une fonction de score.

(2) *titre de l'analyse*

(3) NUMGR numéro d'origine (sur NDICA) de la variable de groupe à 2 groupes (pas de valeur par défaut).

MODG2 modalité (1 ou 2) attribuée au groupe des scores faibles (pas de valeur par défaut).

LCOEF (1 ou EXT) mode de lecture des coefficients.

- 1 ou EXT : lecture sur le fichier NCOEF créé par DIS2G.
- 2 ou LU : lecture après les paramètres, et normalisation.
- 3 ou FORCE : lecture après les paramètres, sans normalisation.

PRG2 (0 ou PROP) pourcentage d'individus du groupe des scores faibles dans la population.

- PROP ou 0 : pourcentage identique à celui de l'échantillon.

TEC (10.) tolérance d'erreur de classement (en %) dans les groupes.

- MAX ou -1 : assure une zone intermédiaire vide.

SXTOT (1000.0) maximum possible de la fonction de score.

NCLAS (50) nombre de classes de découpage des scores.

LABAQ (NIACT) abaque des probabilités conditionnelles.

- 0 ou NON : pas de calcul de l'abaque.
- N : calcul de l'abaque en un nombre limité N d'opérations.

(4) *(si* LCOEF = 2 ou 3*) introduction de la liste des coefficients.*
FIN

3. Présentation détaillée des paramètres

NUMGR **numéro d'origine (sur NDICA) de la variable de groupe.**

- *valeurs possibles* : valeur entière positive (inférieure à NQEXA)
- ***pas de valeur par défaut***

La variable de groupe est une variable nominale à 2 modalités. On reprend ici la variable indicatrice de groupe utilisée dans la procédure préalable d'analyse discriminante DIS2G. Cette variable doit être déclarée illustrative sur le fichier NDIC, c'est-à-dire en sortie de SELEC. Plus généralement, les variables actives doivent être ici celles ainsi déclarées pour CORMU.

MODG2 **modalité (1 ou 2) attribuée au groupe des scores faibles**

- *valeurs possibles* : 1 ou 2
- ***pas de valeur par défaut***

En fonction du codage de la variable indicatrice de groupe, l'utilisateur choisit d'attribuer l'étiquette "groupe des scores faibles" soit à la modalité 1, soit à la modalité 2 de la variable indicatrice de groupe. Par exemple MODG2 = 1 signifie que les individus codés "1" sont ceux qui appartiennent au groupe des scores faibles.

LCOEF **mode de lecture des coefficients**

- *valeurs possibles* : 1 ou EXT
  2 ou LU et normalisés
  3 ou FORCE et non normalisés
- *valeur par défaut* :1 (lecture sur fichier)

Si LCOEF = 1 ou EXT, les coefficients de la fonction discriminante sont lus directement sur le fichier "NCOEF" créé par la procédure DIS2G.

Si LCOEF = 2 ou LU, les coefficients de la fonction discriminante sont introduits après les paramètres, selon un mode décrit plus bas. Ils sont ensuite transformés par la procédure de normalisation de la fonction de score.

Si LCOEF = 3 ou FORCE, les coefficients lus sont considérés comme déjà transformés pour le calcul de la fonction de score. Cette option permet de tester l'effet d'une modification manuelle des coefficients transformés résultant d'un précédent passage. On prendra garde dans ce cas à conserver la cohérence entre les valeurs des coefficients que l'on introduit et la norme de la fonction de score (qui varie entre 0 et SXTOT).

PRG2 **coefficient de redressement du groupe 2**

- *valeurs possibles* : 1. à 99.
- *valeurs particulières:* 0 ou PROP (% dans la population identique à l'échantillon)
- *valeur par défaut:* 0 ou PROP

Si la proportion d'individus du groupe des scores faibles dans la population est égale à la proportion observée dans l'échantillon, on indique PRG2 = PROP. Sinon PRG2 sera égal au pourcentage *réel* d'individus du groupe des scores faibles dans la population totale. Les individus seront alors pondérés pour que l'échantillon soit à l'image de la population.

TEC **tolérance d'erreur de classement**

- *valeurs possibles* : 1. à 99.
- *valeur par défaut* :10.
- *valeur particulière*: MAX

Ce taux est utilisé pour calculer les seuils des zones définies sur l'échelle des scores. La valeur MAX assure le TEC le plus grand possible.

Du côté des valeurs faibles des scores, on trouvera au plus TEC pour cent d'individus du groupe des scores forts: ce sont donc des mal classés.

Du côté des scores élevés on trouvera au plus TEC pour cent d'individus du groupe des scores faibles. Ce sont des individus classés à tort du côté des scores forts.

> Si un individu a un score inférieur au premier seuil, on le classera dans le groupe des scores faibles: c'est ce qu'on appellera la ***zone ROUGE***. Si le score est supérieur au deuxième seuil, on classera l'individu dans le groupe des scores forts. On dira que l'on est dans la ***zone VERTE***. On appellera ***zone intermédiaire*** (ou d'indécision), la zone comprise entre les deux seuils, zone où il est plus difficile d'attribuer une étiquette à un individu.

En diminuant le taux de tolérance d'erreur TEC, on adopte une stratégie prudente d'attribution des étiquettes, mais en contrepartie on élargit la zone d'indécision.

Un taux de tolérance trop élevé peut entraîner le chevauchement des zones rouge et verte conduisant à une situation absurde. Dans ce cas un message alerte l'utilisateur et la valeur de TEC est ramenée à sa borne maximale.

SXTOT **maximum de la fonction score**

- *valeurs possibles* :réelle positive
- *valeur par défaut* : 1000.

La fonction de score peut prendre ses valeurs entre 0. et SXTOT (même si ces valeurs extrêmes ne sont pas atteintes dans l'échantillon). On choisira en général SXTOT = 1000. (valeur par défaut) ou parfois 100. Le choix détermine la précision d'affichage de la fonction discriminante et donc dépend de la taille des échantillons.

NCLAS **nombre de classes de découpage des scores**

- *valeurs possibles* : entières ≥ à 5
- *valeur par défaut* : 50

Il s'agit d'un paramètre d'édition. Les tableaux de résultats et les graphiques des scores sont découpés en classes de scores: chaque classe de scores occupe une ligne du tableau de résultat ou une ligne du graphique.

Si SXTOT = 1000 et NCLAS = 50, les scores sont présentés par tranche de largeur 20. Une ligne de résultat concerne 20 unités de scores. Les tableaux et les graphiques sont édités sur NCLAS=50 lignes. Ce nombre de classes, donc la finesse du découpage, est choisi en fonction de la taille des échantillons.

LABAQ **abaque des probabilités conditionnelles**

- *valeurs possibles:* rélles positives
- *valeur par défaut :* NIACT, taille de l'échantillon actif
- *valeur particulière* : 0 ou NON (pas de calcul de l'abaque)

Lorsque l'échantillon est de taille faible, la courbe des probabilités conditionnelles peut être imprécise. On demandera alors le calcul de l'abaque pour toutes les valeurs possibles des scores. Le nombre des opérations est égal au produit cumulé des nombres de modalités des variables de base.

- Si la taille de l'échantillon est plus petite que le produit cumulé des nombres de modalités des variables de base, la valeur par défaut NIACT (nombre d'individus actifs dans l'échantillon) conduit au calcul des probabilités conditionnelles pour les scores de chaque individu de l'échantillon. Une courbe des probabilités conditionnelles est ensuite construite à partir de ces valeurs.

- Si la taille de l'échantillon est plus grande que le produit cumulé des nombres de modalités des variables de base, on calcule l'abaque exacte.

4. Introduction des coefficients

Dans le cas LCOEF = 2 ou LU, les coefficients de la fonction discriminante sont introduits avec les commandes de la procédure.

Dans le cas LCOEF = 3 ou FORCE, ce sont des coefficients déjà transformés et éventuellement modifiés manuellement qui sont introduits.

Dans les deux cas, les coefficients sont lus après les paramètres de commande. Ils sont écrits en format libre, séparés par des blancs ou des virgules. La liste des coefficients peut s'étendre sur plusieurs lignes, à condition d'annoncer la continuation par le symbole ">".

Le mot-clé FIN annonce la fin de la liste.

Les coefficients doivent être introduits en respectant l'ordre des variables du modèle de discrimination et, pour chaque variable, en respectant l'ordre de ses modalités. Pour cela on se référera au listage fourni par la procédure DIS2G qui a précédé.

Il est commode de grouper les coefficients d'une même variable sur une ligne, afin de contrôler plus facilement leur introduction.

5. Exemples de commande

5.1 Premier exemple

Cet exemple montre l'enchaînement des procédures, à partir de la sélection des variables de l'analyse discriminante.

```
----+----1----+----2----+----+----3----+----4----+----5----
   PROC SELEC
====  SELECTION DES VARIABLES POUR L'ANALYSE  ====
NOPAR
NOMI ACT   19 -- 25
NOMI ILL   4
FIN

   PROC CORMU
====  ANALYSE DES CORRESPONDANCES MULTIPLES ====
NAXE = 10

   PROC DIS2G
====  ANALYSE DISCRIMINANTE SUR LES 7 PREMIERS AXES  ====
LVEC = OUI
V4 = F1--F7

   PROC SCORE
====  ANALYSE DES SCORES  ====
NUMGR = 4, MODG2 = 2, PRG2 = 8.5

STOP
----+----1----+----2----+----+----3----+----4----+----5----
```

La variable du groupe à 2 modalités est la variable 4. Elle doit être déclarée illustrative dans la procédure SELEC. Les variables exogènes du modèle doivent être déclarées actives. Ce sont ici les variables nominales 19 à 25.

Avec la procédure CORMU, on calcule les NAXE = 10 premiers axes factoriels de l'analyse des correspondances multiples des variables actives.

La procédure DIS2G effectue l'analyse discriminante. Le paramètre LVEC = OUI indique que l'analyse est effectuée sur des axes factoriels. Le modèle spécifie que la variable de groupe est V4 (à gauche du signe égal). Les variables exogènes, à droite du signe égal, notées F1 à F7, sont les 7 premiers facteurs de l'analyse. Les résultats obtenus sont exprimés finalement sur les variables originales du modèle, c'est à dire les variables nominales 19 à 25, par la procédure DIS2G.

Enfin la procédure SCORE calcule les résultats et les édite. On y précise que la modalité de la variable de groupe (ici V4 car NUMGR = 4), représentant le groupe des scores faibles, est la deuxième (MODG2 = 2). Pour "redresser" l'échantillon, on indique que la population contient en fait 8,5 % d'individus dans le groupe des scores faibles (PRG2 = 8.5).

Notons que les coefficients de la fonction discriminante sont directement lus sur le fichier NCOEF créé par DIS2G (LCOEF = EXT par défaut). Pour fixer les seuils des zones rouge et verte, on décide de tolérer des taux d'erreur de classement égaux à 10% dans les 2 groupes (TEC = 10.0 par défaut).

5.2 Deuxième exemple

Dans cet exemple, les coefficients de la fonction discriminante sont lus avec les paramètres de commande (LCOEF = LU). Ils viennent après les paramètres, comme une suite ininterrompue de valeurs numériques. Pour faciliter la lecture, il y a passage à une ligne suite pour chaque nouvelle variable. La liste se termine par le mot-clé FIN.

```
----+----1----+----2----+----+----3----+----4----+----5----
   PROC SCORE
====  ANALYSE DES SCORES  ====
NUMGR = 4, MODG2 = 2, PRG2 = 8.5, TEC = 5., LCOEF = LU
-1.718   .048   2.409   -1.388           > : VARIABLE 1
2.143   .922   -6.906   -9.178           > : VARIABLE 2
-8.924   -.598   8.820   5.358   9.677   > : VARIABLE 3
-1.055   2.611   5.070   9.570           > : VARIABLE 5
6.334  .264   -3.574                     > : VARIABLE 6
-11.934   2.850   4.704                   > : VARIABLE 7
-5.228   -1.825   2.899   4.920          > : VARIABLE 8
6.202   -.316   -7.425                   > : VARIABLE 9
-1.573   1.195                           > : VARIABLE 10
2.200   -17.229                           : VARIABLE 11

FIN : FIN DE LA LISTE DES COEFFICIENTS

----+----1----+----2----+----+----3----+----4----+----5----
```

5.3 Cas de coefficients imposés

On peut imaginer que cet exemple est exécuté après l'exemple 1 ou après l'exemple 2. Après avoir obtenu les valeurs numériques des scores transformés, on teste des valeurs modifiées manuellement par arrondi à une valeur voisine, ou modifiés pour respecter un certain ordre sur des catégories ordonnées. La procédure permettra de comparer les résultats obtenus avec les résultats initiaux.

```
----+----1----+----2----+----+----3----+----4----+----5----
   PROC SCORE
====  ANALYSE DES SCORES  ====
NUMGR = 4, MODG2 = 2 , PRG2 = 8.5 , TEC = 5., LCOEF = FORCE
   0    20    30    10         > : VARIABLE 1
  80    70    20     0         > : VARIABLE 2
   0    50   130   100   140   > : VARIABLE 3
 130     0                     > : VARIABLE 4
   0    30    50    80         > : VARIABLE 5
  80    30     0               > : VARIABLE 6
   0   110   120               > : VARIABLE 7
   0    30    60    80         > : VARIABLE 8
 100    50     0               > : VARIABLE 9
   0    20                     > : VARIABLE 10
 130     0                       : VARIABLE 11

FIN

----+----1----+----2----+----+----3----+----4----+----5----
```

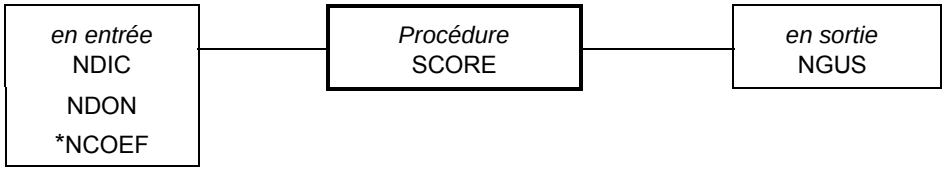
Avec l'option LCOEF = FORCE, les valeurs numériques introduites ici ne seront pas modifiées. L'utilisateur doit donc s'assurer d'un certaine cohérence de ces valeurs avec la normalisation demandée de la fonction discriminante. Par exemple ici, le score maximum possible est fixé par défaut à SXTOT = 1000. On peut vérifier que le score maximum, atteint en ajoutant les maxima dans chaque variable, n'est pas très éloigné de cette valeur:

$$30 + 80 + 140 + 130 + 80 + 80 + 120 + 80 + 100 + 20 + 130 = 990$$

Si les résultats ne différent pas sensiblement, l'utilisateur pourra remplacer les valeurs calculées par ces nouveaux coefficients plus faciles à manipuler.

6. Fichiers nécessaires à l'exécution

- en lecture NDIC (dictionnaire utile)
  NDON (données utiles)
  NCOEF (coefficients de la discriminante)



NCOEF n'est utilisé que si LCOEF = 1 ou EXT.

Le fichier NGUS contient les scores individuels calculés par la procédure. Ces scores sont archivables par ESCAL comme des coordonnées factorielles sur un seul axe (AXIS(1)).

1. Présentation

1.1 Objet

Cette étape gère les calculs et les éditions d'un ajustement des moindres-carrés sur un modèle linéaire comprenant un terme constant. Elle permet d'effectuer les régressions multiples, les analyses de variance et de covariance avec interactions jusqu'à l'ordre 3. A chaque coefficient de la régression est associé le test de nullité, valable dans le contexte classique où le terme aléatoire est supposé engendré par une loi de Laplace-Gauss.

Si les observations sont caractérisées par différents critères nominaux ou "facteurs", le programme exécutera une analyse de la variance pour tester l'existence de l'effet de chacun des facteurs. S'il y a plusieurs critères, on peut introduire dans les modèles et tester d'éventuelles interactions entre couples et triplets de facteurs. Les estimations peuvent prendre en compte les répétitions d'observations dans les plans d'expérience.

On peut demander l'écriture d'un fichier contenant les coefficients de la régression, récupérables pour réaliser en particulier des représentations graphiques des résultats d'analyse de variance. Dans la version micro-ordinateur P.C., ce fichier sera utilisé par le module graphique d'interprétation des analyses de variance.

Le traitement des **données manquantes** est guidé par le paramètre LSUPR pour les données continues, et par l'option REC ou NOREC de la procédure SELEC précédente pour les variables nominales.

1.2 Editions

On édite les statistiques sommaires sur les variables du modèle (tri-à-plat des variables nominales, moyenne, écart-type, minimum et maximum des variables continues). L'étape fournit l'identification des coefficients du modèle: coefficient des variables continues (endogènes), des modalités des facteurs et des interactions éventuelles. Il est ensuite possible d'obtenir l'édition de la matrice des variances-covariances ou celle de la matrice des corrélations.

L'étape imprime les coefficients, l'estimation de leur écart-type, la statistique de Student correspondante, la probabilité critique ainsi que la valeur-test associée. On trouve également la somme des carrés des écarts, le coefficient de corrélation multiple, et l'estimation de la variance commune des résidus. On effectue enfin le test de nullité simultanée de tous les coefficients (test d'une variable endogène "y" constante).

Dans le cas d'une analyse de la variance, on obtient de plus les sommes des carrés d'écarts suivant leur source (résiduelle, critère ou interaction), ainsi que les statistiques de Fisher, les probabilités critiques et valeurs-tests associées. Dans le cas d'observations répétées, on édite la variance "de répétabilité" ainsi que les estimations obtenues en tenant compte de cette variance.

1.3 Paramètres

Les paramètres de la procédure VAREG sont au nombre de six: un paramètre (LSUPR) permet de choisir le mode de traitement des données manquantes (soit l'abandon, soit l'affectation à la moyenne générale). Les deux suivants contrôlent l'édition des statistiques sommaires (LSTAT) et de la matrice des variances-covariances (LMAT). Un paramètre (LREP) permet de prendre en compte les répétitions. Le paramètre LCOF commande l'écriture du fichier externe des coefficients. Enfin le paramètre LED permet de choisir le type d'édition des libellés associés aux coefficients.

1.4 Après les paramètres

A la suite de ces paramètres se trouvent définis le ou les modèles de régressions, chaque instruction correspondant à un ajustement. La liste se termine impérativement par le mot-clé FIN.

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que tous les paramètres prennent leur valeur par défaut, on codera le mot-clé NOPAR à la place de la liste des paramètres repérée par (3).

(1) PROC VAREG Régression et analyse de la variance

(2) *titre de la procédure*

(3) LSUPR (1 ou OUI) suppression des individus présentant des données manquantes pour les variables continues.
- 0 ou NON : individus sont conservés.
- 1 ou OUI : individus retirés de l'analyse.

LSTAT (0 ou NON) statistiques sommaires sur les variables des ajustements.
- 0 ou NON : pas d'édition.
- 1 ou OUI : édition des statistiques.

LMAT (0 ou NON) édition de la matrice de variance-covariance.
- 0 ou NON : pas d'édition.
- 1 ou COVA : édition de la matrice de variance-covariance.
- 2 ou CORR : édition des corrélations.

LREP (0 ou NON) calcul tenant compte des répétitions.
- 0 ou NON : pas de répétitions.
- 1 ou OUI : répétitions en séquence.
- 2 ou TRI : répétitions en désordre (à trier)

LCOF (0 ou NON) sauvegarde des coefficients dans NCOVA
- 0 ou NON : pas de fichier de sauvegarde.
- 1 ou OUI : sauvegarde sur fichier ASCII.

LED (0 ou NON) édition avec libellés pour les coefficients.
- 0 ou NON : pas de rappel des libellés.
- 1 ou OUI : libellés en clair.

(4) *définition du ou des modèles de régression (obligatoire).*
FIN fin de la liste des modèles.

3. Présentation détaillée des paramètres

LSUPR **élimination des individus présentant des données manquantes**

- *valeurs possibles* : 0 ou NON (les individus sont conservés)
  1 ou OUI (les individus sont retirés de l'analyse)
- *valeur par défaut* : 1 ou OUI

Ce paramètre détermine le mode de traitement des données continues manquantes.

Si LSUPR = OUI, les individus présentant des données manquantes pour une des variables du modèle (endogène ou exogène) seront éliminées de l'analyse.

Si LSUPR = NON, les données exogènes manquantes seront remplacées par la moyenne de la variable correspondante.
Attention : si LSUPR = NON, vérifier que la variable endogène ne présente pas de données manquantes.

Quelle que soit la valeur de LSUPR, le traitement d'un individu présentant des données manquantes pour une variable nominale se fait par le choix de l'option REC ou NOREC dans l'étape SELEC précédente.

- Si on a choisi REC, les valeurs manquantes seront traitées comme une modalité particulière.
- Si on a choisi NOREC, les individus présentant des données manquantes seront éliminés.

LSTAT **édition des statistiques sommaires sur les variables du modèle**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des statistiques)
- *valeur par défaut* : 0 ou NON

Si LSTAT = 1, on obtiendra un tri-à-plat des variables nominales du modèle, ainsi que diverses statistiques concernant les variables continues (moyenne, écart-type, minimum et maximum).

LMAT **édition de la matrice des variances-covariances**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou COVA (édition de la matrice de variance-covariance)
  2 ou CORR (édition de la matrice de corrélations)
- *valeur par défaut* : 0 ou NON

LREP **calcul tenant compte des répétitions**

- *valeurs possibles* : 0 ou NON (il n'y a pas de répétitions)
  1 ou OUI (répétitions en séquence)
  2 ou TRI (répétitions en désordre)
- *valeur par défaut* : 0 ou NON

Ce paramètre concerne le traitement des plans d'expérience. Lorsqu'il y a des répétitions, la variance des observations peut être estimée sur les répétitions d'observations plutôt que sur l'ensemble des observations. Il n'est pas nécessaire que le nombre de répétitions soit le même partout.

On code LREP = OUI si les répétitions sont les unes en dessous des autres en lignes du tableau des données. Sinon on codera LREP = TRI. Cette dernière option est plus coûteuse en mémoire et temps de calcul.

LCOF **sauvegarde des coefficients sur NCOVA**

- *valeurs possibles* : 0 ou NON (pas de sauvegarde)
  1 ou OUI (création du fichier)
- *valeur par défaut* : 0 ou NON

Si LCOF = 1, on crée un fichier dont le nom imposé est 'NCOVA'. Il s'agit d'un fichier texte contenant les valeurs des coefficients et les libellés des variables auxquelles ils se rapportent. Ce fichier est essentiellement structuré pour servir d'entrée au moniteur de graphiques de la version micro-ordinateur (pour les interprétations des analyses de variance).

Si la procédure VAREG traite plusieurs modèles dans la même exécution, le fichier NCOVA ne contient que les résultats du *dernier* modèle traité.

LED **édition avec libellés pour les coefficients**

- *valeurs possibles* : 0 ou NON (pas de rappel des libellés)
  1 ou OUI (libellés en clair)
- *valeur par défaut* : 0 ou NON

Si LED = 1, les éditions des coefficients, des facteurs et des interactions sont faites avec le rappel des libellés en clair des éléments auxquels ils se rapportent.

4. Définition du modèle de régression

On définit, après les paramètres, le ou les modèles de régression. Un modèle s'écrit sous la forme suivante:

- *Vp = V1 + ... + Vn + Vi*Vj + ... + Vi*Vj*Vk + ...*

Vp est la **variable endogène** du modèle. Elle doit être continue, et se trouver seule à gauche du signe "=". D'autre part V1, Vn, Vi, Vj, Vk sont les **variables exogènes**. L'usage du symbole de suite "--" est licite pour introduire une suite de variables.

Le caractère "*", placé entre 2 variables, annonce une interaction:

- Vi*Vj : interaction d'ordre 2 (entre les variables Vi et Vj)
- Vi*Vj*Vk : interaction d'ordre 3.

Les interactions peuvent être définies entre variables continues, nominales, ou mixant continues et nominales. Le signe d'interaction ne peut pas être utilisé conjointement avec le symbole de suite "--".
Toutes les variables spécifiées dans un modèle doivent avoir été retenues lors de la précédente étape SELEC.
Après l'écriture des modèles à ajuster, on doit introduire l'instruction FIN.

5. Exemples de commande

On fournit ici un exemple de l'utilisation de la procédure VAREG. Pour la compréhension des modèles, nous avons choisi de présenter également l'étape SELEC précédente.

5.1 Différents ajustements (régressions)

```
----+----1----+----2----+----3----+----4----+----5----+--
   PROC SELEC
====  SELECTION DES VARIABLES ===

LZERO = NOREC

NOMI ACT 1--4
NOMI ILL 5, 7
CONT ACT 6
CONT ILL 10--13
FIN

   PROC VAREG
=== REGRESSION, ANALYSE DE LA VARIANCE ===

LSUPR = 0, LMAT = CORR

V10 = V6 + V11--V13
V6  = V1--V3 + V1*V2 + V1*V2*V3
V13 = V1+ V10--V12
FIN
----+----1----+----2----+----3----+----4----+----5----+--
```

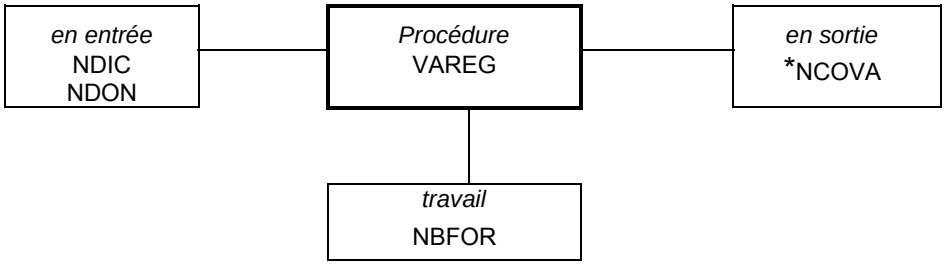
On demande le traitement successif de trois modèles. Dans tous les cas, la variable endogène est une variable continue, dont le statut (actif ou illustratif) est indifférent.

On a pris soin de coder LZERO = NOREC lors de l'étape SELEC. Pour chacun des ajustements on demande que les données continues exogènes manquantes soient remplacées par leur moyenne (LSUPR = 0). Rappelons qu'il est interdit d'avoir une donnée manquante dans la variable endogène (continue).

On demande l'édition de la matrice des corrélations (LMAT =CORR). On n'obtiendra pas l'édition des statistiques sommaires des variables des modèles, puisque LSTAT = 0 par défaut. Le premier modèle est un modèlede régression multiple puisque l'on n'y trouve que des variables continues. Le second est un modèle d'analyse de la variance avec des interactions, l'une d'ordre 2 entre V1 et V2, et l'autre d'ordre 3 entre V1, V2, V3. Enfin, le troisième modèle est un modèle d'analyse de la covariance, puisque l'on y retrouve une variable nominale et des variables continues. Après la description des différents modèles se trouve le mot-clé FIN.

6. Fichiers nécessaires à l'exécution

- en lecture NDIC (dictionnaire utile)
  NDON (données utiles)
- de travail NBFOR
- de sortie NCOVA, seulement si LCOF = OUI



Exemples pratiques commentés

Discriminante à 2 groupes et score

1. L'essentiel en quelques procédures

Notre objectif est de proposer une **stratégie** pour l'élaboration d'une fonction de score. Une fonction de score est un instrument de décision souvent utilisé de façon routinière. Il faut donc apporter le plus grand soin à sa fabrication, étudier ses propriétés statistiques, s'assurer de sa stabilité avant de la mettre en circulation.

Les données du problème sont essentiellement les suivantes: il y a **deux** groupes à discriminer et la discrimination doit reposer sur des caractéristiques **nominales** des individus (sexe, catégorie d'âge, etc). Le fil directeur des opérations est le suivant :

- *Choisir les caractéristiques (nominales) de l'individu qui doivent entrer dans la détermination du score.*
- *Transformer ces caractéristiques nominales en paramètres continus par le biais d'une analyse des correspondances multiples.*
- *Sélectionner parmi les paramètres continus obtenus (les facteurs) ceux qui seront utilisés pour participer à la construction de la fonction discriminante.*
- *Quand la fonction discriminante est établie sur les paramètres continus, l'exprimer en fonction des caractéristiques nominales des individus.*
- *Etudier les propriétés et évaluer les "qualités" de cette fonction discriminante.*

Chacune de ces opérations est un travail statistique parfois délicat, nécessitant souvent plusieurs tentatives avant que le résultat soit estimé satisfaisant. Notre propos est d'illustrer la mise en oeuvre de SPAD.N pour la construction raisonnée d'une fonction discriminante.

Dans la **première partie** de cette note, on décrit l'enchaînement minimum des procédures conduisant à l'établissement d'une fonction de score. Il s'agit là de la suite des calculs à effectuer en principe pour obtenir le résultat.

Dans la réalité, l'utilisateur procède à petits pas, faisant des arrêts pour réfléchir sur des résultats intermédiaires, des détours pour obtenir une information qui lui manque pour continuer, et même des retours en arrière si son chemin aboutit à des résultats qui ne lui conviennent pas.

La **seconde partie** de cette note veut donner une certaine idée de cette démarche. Il est cependant exclu de consigner ici toutes les hésitations, les essais infructueux, les réflexions que l'utilisateur en situation de recherche doit faire avant d'obtenir le résultat qui le satisfera.

Dans la **troisième partie**, on édite les principaux résultats obtenu sur un exemple d'application, afin d'illustrer brièvement chaque étape et servir de support aux commentaires et interprétations.

1.1 CORMU : analyse des correspondances multiples

L'utilisateur a choisi les variables nominales qui serviront à la discrimination des deux groupes d'individus. Il a choisi également les individus qui constitueront l'échantillon pour le calcul des coefficients de la fonction discriminante.

La procédure CORMU effectue l'analyse des correspondances multiples pour laquelle les variables actives sont les variables choisies plus haut. La variable indicatrice de groupe est introduite dans l'analyse comme variable illustrative (ou supplémentaire).

CORMU calcule les coordonnées factorielles de chaque individu. L'ensemble des coordonnées situe les individus les uns par rapport aux autres de la même façon que l'ensemble des modalités des variables nominales (ils permettent la reconstitution de la configuration du nuage). CORMU remplace donc les variables nominales par un ensemble de variables continues équivalentes. Ces variables, les coordonnées factorielles, sont copiées sur un fichier dont le nom clé est NGUS.

1.2 DIS2G, SCORE: analyse discriminante et fonction de score

Dans le cas de deux groupes, l'analyse discriminante classique (discrimination linéaire de Fisher) fournit la combinaison linéaire des variables qui sépare au mieux les deux groupes d'individus. Un individu est affecté à un groupe si la fonction prend pour lui une valeur supérieure à un certain seuil. Sinon il est affecté à l'autre groupe. La procédure DIS2G réalise le calcul des coefficients de la fonction discriminante à partir des coordonnées factorielles des individus (ce sont des variables continues). Le fichier NGUS est donc utilisé en entrée de la procédure DIS2G.

Cette fonction discriminante est ensuite transformée en un système équivalent de coefficients attribués aux modalités des variables nominales. Un individu est affecté à un groupe si la somme des coefficients correspondant à ses modalités est supérieure à un certain seuil. Les coefficients attribués aux modalités sont copiés sur un fichier de résultats dont le mot clé est NCOEF.

La procédure SCORE relit et utilise les coefficients de la fonction de score copiés sur le fichier NCOEF. Un tableau de résultats permet d'apprécier la contribution de chaque modalité dans le calcul du score d'un individu.

Si l'utilisateur fixe en pourcentage une tolérance d'erreur de classement notée "TEC", la procédure définit 3 zones sur l'échelle des scores:

- **zone verte**: affectation de l'individu dans la zone 1, cette zone contient TEC pour cent d'individus mal classés (individus du groupe 1 classés en 2)
- **zone rouge**: affectation de l'individu dans la zone 2, cette zone contient TEC pour cent d'individus mal classés (individus du groupe 2 classés en 1)
- **zone centrale** ou intermédiaire, ou zone d'indécision pas d'affectation

Tableaux et graphiques fournissent la répartition des individus dans les groupes en fonction de la valeur de la fonction de score. Une évaluation est faite pour la population entière, en tenant compte d'un éventuel redressement si la répartition dans les groupes 1 et 2 dans l'échantillon n'est pas à l'image de la répartition dans la population.

On trouvera dans l'***encart 1*** un exemple d'enchaînement de procédures réalisant une analyse de ce type. On y a spécifié les fichiers intermédiaires servant à la communication de résultats entre les étapes de calcul (NDICA, NDIC, etc).

Dans le cas d'une exécution globale et définitive, il n'est pas utile de spécifier ces fichiers. Le logiciel les crée automatiquement pour la durée de l'exécution et les efface aussitôt après. Il est rare cependant de procéder en une seule exécution, car les options à fixer pour un calcul dépendent en général des résultats des procédures qui précèdent. Il est même courant d'avoir à exécuter plusieurs fois une même procédure avant de trouver le jeu de paramètres qui satisfait l'utilisateur.

Dans ces conditions, la sauvegarde des fichiers intermédiaires, assurée par l'écriture de leur nom dans l'encart 1, permet de reprendre les calculs à n'importe quel endroit, sans avoir à ré-exécuter les procédures précédentes. Cette stratégie est mise en oeuvre dans la présentation plus détaillée de la partie II.

Encart 1 : commande globale

```
: ---------- CALCUL ET EDITION D'UNE FONCTION DE SCORE ----------

LISTP = OUI, NXLPA = 60, LERFA = OUI      : Paramètres généraux

: ----------------------------------------  LECTURE ET ARCHIVAGE
NDICZ = 'CREDIT.LAD', NDONZ = 'CREDIT.DAD'      : entrée: fichiers initiaux
NDICA = 'NDICA.SCO', NDONA = 'NDONA.SCO'        : sortie: fichiers créés
Proc ARDIC                                      : Lecture des libellés
---- exemple pour un score ----
NQEXA = 14
Proc ARDON                                      : Lecture des données
---- exemple pour un score ----
NIDI = 0, NQEXA = 14, NIEXA = 468, NLFOR = -1

: ----------------------------------------  SELECTION ET CORRESPONDANCES
NDICA = 'NDICA.SCO', NDONA = 'NDONA.SCO'        : entrée = données archivées
NDIC = 'NDIC.SCO', NDON = 'NDON.SCO'    : sortie = données sélectionnées
Proc SELEC                                      : Sélection pour correspondances
---- exemple pour un score ----
IMASS = 15
NOMI ACT 3--10, 12
NOMI ILL 1, 2, 11
CONT ILL 13, 14
FIN

NDIC = 'NDIC.SCO', NDON = 'NDON.SCO'    : entrée = données sélectionnées
NGUS = 'NGUS.SCO'                               : sortie = coord. factorielles
Proc CORMU                                      : Correspondances multiples
---- exemple pour un score ----
NAXE = 20

: ----------------------------------------  APPEL DES CALCULS
NDIC = 'NDIC.SCO', NDON = 'NDON.SCO', NGUS = 'NGUS.SCO' : fichiers en entrée
NCOEF = 'NCOEF.SCO'                             : Sortie = coefficients
Proc DIS2G                                      : Analyse discriminante
---- exemple pour un score ----
LVEC = OUI
V1 = V1 -- V12                                  : Modèle à estimer
FIN

NDIC = 'NDIC.SCO', NDON = 'NDON.SCO', NCOEF = 'NCOEF.SCO': fichiers en entrée
Proc SCORE                                      : Fonction de score
---- exemple pour un score ----
NUMGR = 1, MODG2 = 2, TEC = 10.0 , PRG2 = 15.0 , LABAQ = 1000

STOP : ----------------------------------  Fin de l'analyse
```

2. La pratique

2.1 Jusqu'à l'analyse des correspondances multiples

Il arrive que l'on soit amené à *redresser* l'échantillon pour que certaines répartitions en pourcentage dans une ou plusieurs variables nominales soient respectées. On donne ici l'exemple d'un redressement portant sur les variables V2 et V3. Les listes associées sont les pourcentages à respecter.

Le redressement peut porter sur la variable indicatrice de groupe elle-même, la variable V1 dans l'exemple. Il faut noter une pratique assez courante lorsqu'un des 2 groupes à discriminer est rare par rapport à l'autre (pièces défectueuses, mauvais clients, etc). Dans ce cas, on calcule la fonction discriminante, donc les scores, en équilibrant les deux groupes par élimination au hasard d'individus dans le groupe trop abondant. Il faut alors introduire un redressement en fin de calcul, lorsque l'on évalue le comportement des scores sur la population globale et non sur l'échantillon. Ce paramètre de redressement apparaîtra dans la procédure SCORE. Deux types de redressement sont donc utilisables ici.

Les procédures SELEC et CORMU qui suivent, réalisent l'**analyse des correspondances** multiples où les variables actives sont les variables qui doivent figurer dans la fonction discriminante. La variable de groupe V1, ainsi que d'autres variables éventuelles, sont déclarées illustratives.

On ne commente pas ici comment se fait le choix des variables nominales participant à la fonction discriminante. Il résulte de considérations *a priori* faites par le spécialiste du domaine étudié d'une part, modifiées éventuellement au vu des résultats statistiques obtenus (validations faites dans DIS2G, et va-et-vient "SELEC-CORMU-DIS2G"). Ce choix doit être guidé par les aides à l'interprétation de DIS2G.

Encart 2 : correspondance multiple

```
: ------ "1_CORMU" = ANALYSE DES CORRESPONDANCES MULTIPLES ------

LISTP = OUI, NXLPA = 120, LERFA = OUI     : Paramètres généraux

: ----------------------------------------  REDRESSEMENT
NDICA = '0_NDICA', NDONA = '0_NDONA'    : entrée = fichiers archivés
NDICB = '1_NDICA', NDONB = '1_NDONA'    : sortie = données redressées
Proc REDRE                                          : redressement de l'échantillon
---- exemple pour un score ----
NOPAR
 V2  15.0 50.0 25.0 10.0
 V3  30.0 40.0 20.0 10.0
FIN

: ----------------------------------------  SELECTION ET CORRESPONDANCES
NDICA = '1_NDICA', NDONA = '1_NDONA'    : entrée = données redressées
NDIC = '1_NDIC', NDON = '1_NDON'                    : sortie = données sélectionnées
Proc SELEC                                          : Sélection pour correspondances
---- exemple pour un score ----
IMASS = 15
 NOMI ACT 3--10, 12
 NOMI ILL 1, 2, 11
 CONT ILL 13, 14
FIN

NDIC = '1_NDIC', NDON = '1_NDON'                    : entrée = données sélectionnées
NGUS = '1_NGUS'                                             : sortie = coord. factorielles
Proc CORMU                                          : Correspondances multiples
---- exemple pour un score ----
NAXE = 20
STOP : ----------------------------------  Fin de "1_CORMU"
```

2.2 Archivage des coordonnées factorielles (ESCAL)

Lorsque le choix des variables actives est arrêté (après éventuellement plusieurs essais "SELEC-CORMU-DIS2G"), il est bon d'archiver les coordonnées factorielles qui en résultent. Ces nouvelles variables constituent un intermédiaire de calcul pour la fonction discriminante, mais seront utiles aussi au moment des validations statistiques.

Dans l'ignorance du nombre de coordonnées réellement utiles, on archivera le plus grand nombre possible (mais raisonnable) d'axes factoriels.

Encart 3 : archivage des coordonnées

```
: --------- "2_ARCHI"  ARCHIVAGE DES COORDONNEES ---------

LISTP = OUI, NXLPA = 120, LERFA = OUI     : Paramètres généraux

: ----------------------------------------  ARCHIVAGE DES AXES
NGUS = '1_NGUS'                                              : entrée : coordonnées (de CORMU)
NDICA = '1_NDICA', NDONA = '1_NDONA'    : entrée : les données initiales
NDICB = '2_NDICA', NDONB = '2_NDONA'    : sortie : ajout des coordonnées
Proc ESCAL                                                   : archivage des résultats
---- exemple pour un score ----
LVEC = OUI, LEDIT = COURT

PROG
  COPY
  TRANS V1--V15, AXIS(1)--AXIS(20)                           : transfert données et coordonnées
RUN

STOP : ----------------------------------  Fin de "2_ARCHI"
```

2.3 Faut-il prendre toutes les coordonnées factorielles (SELEC, FUWIL)

Si l'on retient toutes les coordonnées factorielles pour effectuer l'analyse discriminante, on utilise toutes les informations contenues dans les variables nominales actives. On sait qu'en régression comme en discrimination il est souvent risqué d'utiliser beaucoup de variables dans le modèle à ajuster, des colinéarités entre variables étant cause d'instabilité des résultats. Ce danger est écarté ici car les variables factorielles sont non corrélées.

On trouvera parfois intérêt à abandonner les derniers axes factoriels. En effet, parmi les composantes de la dispersion des observations, on trouve en général sur les derniers axes les résultats d'effets non systématiques (le "hasard"). Les supprimer constitue une sorte de "lissage" des données. Cette opération peut améliorer l'aptitude des données à entrer dans le modèle de discrimination s'il est confirmé que ces axes ne sont pas des directions séparant les groupes.

La procédure FUWIL peut être utile pour choisir les axes factoriels à retenir. Elle indique les meilleures discriminations possibles, quelque soit le nombre de variables factorielles utilisées. Dans le cas présent, les variables factorielles étant non corrélées, la contribution d'une variable est indépendante des autres variables présentes dans le modèle de discrimination, ce qui va simplifier le choix.

Encart 4 : choix des axes

```
: ------- "3_FUWIL"  CHOIX DES AXES POUR LA DISCRIMINANTE -------

LISTP = OUI, NXLPA = 120, LERFA = OUI     : Paramètres généraux

: ---------------------------------------  SELECTION DES VARIABLES
NDICA = '2_NDICA', NDONA = '2_NDONA'    : entrées : données et coordonnées
Proc SELEC                                          : Sélection pour FUWIL
---- exemple pour un score ----
IMASS = 15

NOMI ILL 1
CONT ILL 16 -- 35
FIN
: ---------------------------------------  SELECTION ET CALCULS
Proc FUWIL                                          : comparaison des discriminantes
---- exemple pour un score ----
LMODE = DISC

V1 = V16 -- V35                                                 : les axes sont les var. 16 à 35
FIN

STOP : ---------------------------------  Fin de "3_FUWIL"
```

Pour comparer les résultats, on peut utiliser le critère purement géométrique que constitue le R2 ou coefficient de corrélation multiple. La valeur "1 - R2" mesure la distance, dans l'espace des observations, entre la variable à prévoir (la variable indicatrice de groupe) et le sous espace engendré par les axes factoriels retenus pour l'analyse. La discriminante est donc "meilleure" quand le R2 augmente. Cependant le R2 augmente systématiquement dès qu'on ajoute une variable, quelle qu'elle soit.

Si les premiers ajouts de variables font croître sensiblement le R2, l'augmentation est de moins en moins forte ensuite. On pourrait arrêter l'introduction de nouvelles variables dès que l'accroissement correspond à la contribution d'une variable purement "aléatoire". Le graphique final édité par FUWIL peut guider dans le choix du nombre de variables à retenir.
On dispose par ailleurs dans FUWIL d'un critère qui permet d'apprécier l'importance d'une variable dans la discrimination. Il s'agit de la "valeur-test" associée au coefficient de cette variable: plus la valeur-test est grande (en valeur absolue), moins il est "probable" que le coefficient de cette variable est nul. Par construction, il s'agit d'un critère de type statistique, mais que l'on utilise ici dans un cadre purement *descriptif*, car les conditions de validité statistique ne sont pas satisfaites.

On peut donc ranger les variables dans l'ordre des valeurs-tests décroissantes avant de choisir où s'arrêter.

(NB: Les éventuelles variations des valeurs-tests dans le listage de FUWIL sont dues aux difficultés de calcul numérique: il faut transformer la probabilité souvent très faible d'une distribution à chaque fois différente, en une valeur de loi de Laplace-Gauss.)

Encart 5 : rangement des axes

Variable	Coefficient	Valeur-Test
axe1	-.9118	11.9181
axe2	-.6398	7.7796
axe5	.4104	4.2976
axe8	.3774	3.6194
ax12	-.4158	3.5055
ax15	.3940	2.9681
axe7	-.2710	2.6180
axe4	-.2047	2.2376
ax18	.2978	2.0620
ax10	-.1986	1.7985
ax14	.2047	1.6080
ax16	-.2186	1.6354
axe9	.1513	1.4442
ax13	-.1540	1.2566
ax17	.0917	.6497
ax20	.1455	.5549
axe6	.0499	.4988
ax11	.0387	.3400
axe3	.0154	.1720
ax19	.0210	.1108

Considérons les axes 1 à 20 rangés en fonction des valeurs-tests comme l'indique l'encart 5. Les axes 12 et 15, situés assez haut dans la liste, sont peut-être discriminants. L'axe 18, un des derniers axes (de dispersion très faible), sera peut-être exclu du modèle. L'axe 14, situé au delà de l'axe 18, est donc vraisemblablement à éliminer, ce qui suggère de s'arrêter à l'axe 12 plutôt qu'à l'axe 15.

De plus on peut se demander s'il est préférable de travailler avec les axes 1 à 12, ou avec les seuls axes les plus discriminants, c'est-à-dire 1, 2, 5, 8 et 12. On vérifiera toujours, compte tenu du caractère expérimental de la démarche, que les résultats sont peu affectés quand on augmente ou diminue un peu le nombre d'axes.

Il faut noter que, s'il est important de minimiser le nombre de variables dans le cas de variables quelconques, il n'y a pas d'inconvénient à ajouter des variables non corrélées. Il y a parfois avantage à reconstituer au mieux les "vraies" distances entre les observations pour calculer le plan de séparation des deux groupes. En fait l'expérience du praticien peut être plus déterminante ici que des considérations théoriques pour arrêter le modèle final.

2.4 Analyse discriminante (DIS2G)

La procédure DIS2G met en oeuvre différentes techniques de validation des résultats et fournit des aides à l'interprétation qui lui sont propres et dont on ne donne pas le détail ici. Cependant l'utilisateur devra en tirer profit avant de poursuivre le calcul des scores, car les opérations de validation sont très importantes en analyse discriminante.

Encart 6 : discriminante

```
: ---------- "4_DIS2G"  ANALYSE DISCRIMINANTE ----------

LISTP = OUI, NXLPA = 120, LERFA = OUI                    : Paramètres généraux

: ------------------------------------------------- APPEL DES CALCULS
NDIC = '1_NDIC', NDON = '1_NDON', NGUS = '1_NGUS'        : fichiers en entrée
NCOEF = '4_NCOEF'                                              : Sortie : coefficients
NGRO = '4_NGRO'                                                      : Sortie : classement
Proc DIS2G
---- exemple pour un score ----
LVEC = OUI, LEDIN = non, LEDIV = non                     : paramètres courants

V1 = V1 -- V15                                                 : pour vérification
V1 = V1 + V2 + V4 + V5 + V8 + V12                        : pour vérification
V1 = V1 -- V12                                                 : pour archivage
FIN

STOP : ---------------------------------------- Fin de "4_DIS2G"
```

La procédure crée un autre fichier de résultats, appelé NGRO, qui contient la variable à 4 modalités résultant de la discrimination de chaque individu:

"1" = du groupe 1 et classé 1 "2" = du groupe 1 et classé 2
"3" = du groupe 2 et classé 1 "4" = du groupe 2 et classé 2

L'utilisateur a la possibilité d'ajuster plusieurs modèles de discrimination au cours du même passage dans DIS2G afin d'effectuer des comparaisons. Seuls les résultats du dernier modèle seront sauvegardés sur NCOEF et NGRO.

2.5 Fonction de score (SCORE)

La procédure SCORE cherche les coefficients de la fonction discriminante sur le fichier NCOEF créé par DIS2G. Elle utilise en entrée les données sélectionnées pour l'analyse discriminante, donc en fait les fichiers NDIC et NDON créés par SELEC pour l'analyse des correspondances multiples CORMU. Les valeurs individuelles de la fonction discriminante sont copiées sur un fichier de type NGUS.

Encart 7: scores

```
: ---------- "5_SCORE"  FONCTION DE SCORE  ----------

LISTP = OUI, NXLPA = 120, LERFA = OUI              : Paramètres généraux

: ------------------------------------------------- APPEL DES CALCULS
NDIC = '1_NDIC', NDON = '1_NDON', NCOEF = '4_NCOEF': fichiers en entrée
NGUS = '5_NGUS'                                                : sortie = les scores
Proc SCORE                                                     : calculs et éditions
---- exemple pour un score ----
NUMGR = 1, MODG2 = 2, TEC = 10. , PRG2 = 15. , LABAQ = 1000

STOP : ---------------------------------------- Fin de "5_SCORE"
```

Les coefficients de la fonction score constituent des notes attachées aux modalités des variables nominales. La somme des notes concernant un individu constitue son score. Pour faciliter sa tâche, l'utilisateur peut être tenté d'arrondir les notes

attachées aux modalités. De plus, lorsque les modalités sont ordonnées (catégories d'âge, de revenu, etc), on peut souhaiter avoir des notes qui reflètent cet ordre.

La procédure SCORE permet de prendre en entrée des coefficients transformés manuellement. En comparant les résultats obtenus avec les résultats initiaux, on saura si les manipulations effectuées étaient autorisées. Notons que l'optimum fourni par les méthodes d'estimation est en général assez "plat": autrement dit, des modifications "raisonnables" des coefficients le plus souvent modifient peu les résultats.

Encart 8 : scores (bis)

```
: ---------- "5_SCORE"  FONCTION DE SCORE (bis) ----------

LISTP = OUI, NXLPA = 120, LERFA = OUI              : Paramètres généraux

: ------------------------------------------------- APPEL DES CALCULS
NDIC = '1_NDIC', NDON = '1_NDON'                          : Entrée sans NCOEF
NGUS = '5_NGUS'                                                  : Sortie = les scores

Proc SCORE
---- exemple pour un score ----
NUMGR = 1, MODG2 = 2, TEC = 10. , PRG2 = 15. , LABAQ = 1000, LCOEF = FORCE

  225 130  0  70    >
   0  70 110  50  20 >
  10  0   >
  40  0  10  80 >
  160  10  0   >
  70  20  0   >
   0  50  20  10  >
  130  0  90    >
  180  0

STOP : ------------------------------------------- Fin de "5_SCORE"
```

On montre dans l'encart 8 un second appel à la procédure SCORE avec lecture des coefficients modifiés manuellement. Dans ce cas, l'option "LCOEF=FORCE" interdit de modifier les coefficients fournis par l'utilisateur. Le fichier NGUS créé ici remplacera le fichier précédent pour les utilisations ultérieures.

Le cas échéant, l'utilisateur peut souhaiter introduire des coefficients autres que ceux fournis par la méthode classique d'estimation. Il pourra par exemple tester les estimations "bootstrap" des coefficients usuels. On utilise alors l'option "LCOEF=LU".

2.6 Archivage des résultats (ESCAL)

Cette étape intermédiaire est consacrée à l'archivage des résultats de l'analyse discriminante. On archive dans le même passage, pour chaque individu :

- son classement croisé: classe d'origine, classe d'affectation
- la valeur de la fonction discriminante

Les fichiers résultant, contiendront donc toutes les variables déjà connues, plus une variable nominale à 4 modalités et une variable continue. Cet archivage

dans des fichiers SPAD permet de disposer de toutes les fonctionnalités du logiciel pour l'étude ou l'utilisation ultérieures de la fonction discriminante.

Encart 9: archivage des scores

```
: ------ "6_ARCHI"  ARCHIVAGE DU RECLASSEMENT ET DES SCORES ------

LISTP = OUI, NXLPA = 120, LERFA = OUI     : Paramètres généraux

: ----------------------------------------  ARCHIVAGE
NGRO = '4_NGRO'                                          : entrée = reclassement
NGUS = '5_NGUS'                                          : entrée = les scores individuels
NDICA = '2_NDICA', NDONA = '2_NDONA'    : entrée = les données
NDICB = '6_NDICA', NDONB = '6_NDONA'    : sortie = ajout des résultats
Proc ESCAL                                          : archivage des résultats
---- exemple pour un score ----
LVEC = OUI, LCLAS = OUI, LEDIT = COURT

PROG
  COPY
  TRANS V1--V35                                          : transfert données et coordonnées

DICO CLUST(1)                                          : transfert du classement croisé
  4 Classement croisé
DIS1 Bon classé Bon
DIS2 Bon classé Mauv
DIS3 Mauv classé Bon
DIS4 Mauv classé Mauv

DICO AXIS(1)                                          : transfert des scores individuels
  1 Fonction de score
SCOR fonction de score

RUN

STOP : ----------------------------------  Fin de "6_ARCHI"
```

2.7 Exemple d'illustration des résultats

On utilise en entrée les fichiers contenant les résultats de DIS2G et de SCORE. La procédure TABLE effectue les tabulations de la variable "classement croisé" avec les autres variables nominales disponibles. Dans chaque case des tableaux, on trouvera la moyenne et l'écart-type des scores des individus de la case.

Encart 10 : illustrations

```
: ------- "7_DESCRI"  DESCRIPTION STATISTIQUE DES SCORES --------

LISTP = OUI, NXLPA = 120, LERFA = OUI                : Paramètres généraux

: ----------------------------------------  TABULATIONS
NDICA = '6_NDICA ', NDONA = '6_NDONA'                : Entrée = données complétées
Proc TABLE                                                    : tabulation des scores
---- exemple pour un score ----
IMASS = 15, LETYP = OUI
  LIG = V36, COL = V2 -- V12, MOY = V37
FIN

: ----------------------------------------  SELECTION ET DESCRIPTION
NDICA = '6_NDICA ', NDONA = '6_NDONA'                : Entrée = données complétées
Proc SELEC
---- exemple pour un score ----
LZERO = NOREC, LEDIT = COURT, IMASS = 15
NOMI ACT 1--12, 36
CONT ACT 13, 14, 16--35, 37
FIN

: ----------------------------------------  Décrire le classement
Proc DEMOD
---- exemple pour un score ----
NOPAR
V36

: ----------------------------------------  Décrire la fonction score
Proc DESCO
---- exemple pour un score ----
NOPAR
V37

: ----------------------------------------  Exemples de graphiques
Proc GRAPH
---- exemple pour un score ----
LVEC = NON
POINTS = IND, X = V23, Y = V37, IDENT = V36 = '1..4' >
    ZOOM = 3.5, DOUBLE = NO, WIDTH=CHAR=120, HEIGHT=CHAR=80
POINTS = IND, X = V3, Y = V37, IDENT = V1 >
    ZOOM = 3.5, DOUBLE = NO, WIDTH=CHAR=120, HEIGHT=CHAR=80
FIN

STOP : ----------------------------------      Fin de "7_DESCRI
```

On demande ensuite la caractérisation statistique des individus dans chacun des 4 groupes du classement croisé résultant de la discriminante: "1" classé "1"; "1" classé "2"; "2" classé "1" et "2" classé "2". La procédure DEMOD fournit une caractérisation automatique en tenant compte de toutes les variables disponibles.

La procédure DESCO permet de caractériser selon le même principe une variable continue, ici la fonction score proprement dite.

3. Un exemple

On présente dans cette partie quelques extraits des listages de résultats obtenus par exécution des commandes décrites dans la partie II de cette note

EXTRAIT 01

```
---- exemple pour un score ----
1 . Type de client                                          (   2 MODALITES )
   2 . Age du client                                        (   4 MODALITES )
   3 . Situation familiale                                  (   4 MODALITES )
   4 . Ancienneté                                           (   5 MODALITES )
   5 . Domiciliation du salaire                             (   2 MODALITES )
   6 . Domiciliation de l'épargne                           (   4 MODALITES )
   7 . Profession                                           (   3 MODALITES )
   8 . Moyenne en cours                                     (   3 MODALITES )
   9 . Moyenne des mouvements                               (   4 MODALITES )
  10 . Cumul des débits                                     (   3 MODALITES )
  11 . Autorisation de découvert                            (   2 MODALITES )
  12 . Interdiction de chéquier                             (   2 MODALITES )
  13 . continue 1                                           ( CONTINUE )
  14 . continue 2                                           ( CONTINUE )
-------------------------------------------------------------------------------------------------------------------------------
```

EXTRAIT 02

```
INSTRUCTIONS LUES
          V2  15.0 50.0 25.0 10.0
          V3  30.0 40.0 20.0 10.0
          FIN
RESULTATS APRES 10 ITERATIONS
                              --------- POURCENTAGES ---------
                             INITIAUX    DEMANDES     OBTENUS
-------------------------------------------------------------------------------------------------------------------------------
   2 . Age du client
AGE1 -  moins de    23 ans        18.80       15.00       14.95     *********
AGE2 -  de 23 à     40 ans        32.05       50.00       49.74     ******************************
AGE3 -  de 40 à     50 ans        26.07       25.00       24.97     ***************
AGE4 -  plus de     50 ans        23.08       10.00       10.34     *******
-------------------------------------------------------------------------------------------------------------------------------
   3 . Situation familiale
CELB - célibataire                36.32       30.00       30.25     ******************
MARI -    marié                   47.22       40.00       40.08     ************************
DIVO -   divorcé                  13.03       20.00       19.99     ************
VEUF -    veuf                     3.42       10.00        9.69     ******
-------------------------------------------------------------------------------------------------------------------------------
```

EXTRAIT 03

```
                     ---- exemple pour un score ----
-------------------------------------------------------------------------------------------------------------------------------
   1 . Type de client                                       (   2 MODALITES )
   2 . Age du client                                        (   4 MODALITES )
   3 . Situation familiale                                  (   4 MODALITES )
   4 . Ancienneté                                           (   5 MODALITES )
   5 . Domiciliation du salaire                             (   2 MODALITES )
   6 . Domiciliation de l'épargne                           (   4 MODALITES )
   7 . Profession                                           (   3 MODALITES )
   8 . Moyenne en cours                                     (   3 MODALITES )
   9 . Moyenne des mouvements                               (   4 MODALITES )
  10 . Cumul des débits                                     (   3 MODALITES )
  11 . Autorisation de découvert                            (   2 MODALITES )
  12 . Interdiction de chéquier                             (   2 MODALITES )
  13 . continue 1                                           ( CONTINUE )
  14 . continue 2                                           ( CONTINUE )
  15 . coefficient de ponderation                           ( CONTINUE )
-------------------------------------------------------------------------------------------------------------------------------
```

EXTRAIT 05

TRI-A-PLAT DES QUESTIONS ACTIVES

MODALITES IDENT LIBELLE	AVANT APUREMENT EFF.	POIDS	APRES APUREMENT EFF.	POIDS	HISTOGRAMME DES POIDS RELATIFS
3 . Situation familiale					
CELB - célibataire	170	141.57	170	141.57	******************
MARI - marié	221	187.56	221	187.56	************************
DIVO - divorcé	61	93.53	61	93.53	************
VEUF - veuf	16	45.34	16	45.34	******
4 . Ancienneté					
ANC1 - anc. 1 an ou moins	199	211.28	199	211.28	****************************
ANC2 - anc. de 1 à 4 ans	47	48.47	47	48.47	*******
ANC3 - anc. de 4 à 6 ans	69	73.32	69	73.32	**********
ANC4 - anc. de 6 à 12 ans	66	71.16	66	71.16	**********
ANC5 - anc. plus 12 ans	87	63.77	87	63.77	*********
5 . Domiciliation du salaire					
Soui - domicile salaire	316	304.77	316	304.77	**
Snon - non dimicile salaire	152	163.23	152	163.23	*********************
6 . Domiciliation de l'épargne					
EPA0 - pas d'épargne	370	371.36	372	372.78	**
EPA1 - moins de 10KF épargn	58	62.17	60	63.84	*********
EPA2 - de 10 à 100KF épargn	32	28.04	36	31.38	*****
EPA3 - plus de 100KF épargn	8	6.44	=== VENTILEE ===		
7 . Profession					
CADR - cadre	77	62.72	77	62.72	*********
EMPL - employé	237	247.02	237	247.02	*********************************
AUTR - autre	154	158.26	154	158.26	**********************
8 . Moyenne en cours					
ENC1 - moins de 2KF encours	98	107.51	98	107.51	**************
ENC2 - de 2 à 5 KF encours	308	302.79	308	302.79	**************************************
ENC3 - plus de 5 KF encours	62	57.69	62	57.69	********
9 . Moyenne des mouvements					
MOU1 - moins 10 KF mouvt	154	161.07	154	161.07	**********************
MOU2 - de 10 à 30KF mouvt	71	77.96	71	77.96	**********
MOU3 - de 30 à 50KF mouvt	129	124.90	129	124.90	*****************
MOU4 - plus de 50KF mouvt	114	104.07	114	104.07	**************
10 . Cumul des débits					
DEB1 - moins de 40 débits	171	164.54	171	164.54	**********************
DEB2 - de 40 à 100 débits	161	153.39	161	153.39	********************
DEB3 - plus de 100 débits	136	150.08	136	150.08	********************
12 . Interdiction de chéquier					
Coui - chéquier autorisé	415	403.12	415	403.12	**
Cnon - chéquier interdit	53	64.88	53	64.88	*********

EXTRAIT 06

COORDOONNEES, CONTRIBUTIONS ET COSINUS CARRES DES MODALITES ACTIVES SUR LES AXES 1 A 5

MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
3 . Situation familiale																	
CELB - célibataire	3.36	2.31	.22	-.18	.08	.71	-.38	.6	.5	.1	10.8	3.3	.02	.01	.00	.22	.06
MARI - marié	4.45	1.50	-.32	-.06	-.10	-.26	.22	1.7	.1	.2	1.9	1.5	.07	.00	.01	.05	.03
DIVO - divorcé	2.22	4.00	.18	.04	.52	-.01	.37	.3	.0	3.8	.0	2.1	.01	.00	.07	.00	.03
VEUF - veuf	1.08	9.32	.27	.74	-.94	-1.12	-.50	.3	2.9	5.9	8.7	1.8	.01	.06	.10	.14	.03
CONTRIBUTION CUMULEE =								2.9	3.5	10.1	21.5	8.7					
4 . Ancienneté																	
ANC1 - anc. 1 an ou moins	5.02	1.22	.48	-.02	.54	-.14	-.06	4.3	.0	9.1	.6	.1	.19	.00	.24	.02	.00
ANC2 - anc. de 1 à 4 ans	1.15	8.66	.19	.38	-.08	1.10	.43	.2	.8	.0	8.9	1.5	.00	.02	.00	.14	.02
ANC3 - anc. de 4 à 6 ans	1.74	5.38	.05	-.08	-.61	.08	.36	.0	.1	4.0	.1	1.5	.00	.00	.07	.00	.02
ANC4 - anc. de 6 à 12 ans	1.69	5.58	-.80	-.17	-.41	.41	-1.28	4.1	.2	1.8	1.9	19.1	.11	.01	.03	.03	.29
ANC5 - anc. plus 12 ans	1.51	6.34	-.88	.05	-.58	-.93	.90	4.5	.0	3.1	8.3	8.4	.12	.00	.05	.14	.13
CONTRIBUTION CUMULEE =								13.2	1.2	18.1	19.7	30.6					
5 . Domiciliation du salaire																	
Soui - domicile salaire	7.24	.54	-.32	-.23	-.15	.07	.19	2.9	1.9	1.1	.2	1.8	.19	.10	.04	.01	.07
Snon - non dimicile salaire	3.88	1.87	.60	.44	.29	-.13	-.36	5.4	3.6	2.0	.4	3.4	.19	.10	.04	.01	.07
CONTRIBUTION CUMULEE =								8.3	5.6	3.0	.6	5.2					
6 . Domiciliation de l'épargne																	
EPA0 - pas d'épargne	8.85	.26	.33	-.14	-.16	-.13	-.09	3.6	.9	1.4	1.0	.5	.42	.08	.10	.07	.03
EPA1 - moins de 10KF épargn	1.52	6.33	-.64	.57	.34	.54	1.03	2.4	2.4	1.1	2.9	11.1	.06	.05	.02	.05	.17
EPA2 - de 10 à 100KF épargn	.74	13.91	-2.57	.55	1.21	.50	-1.05	18.8	1.1	6.7	1.2	5.7	.47	.02	.10	.02	.08
CONTRIBUTION CUMULEE =								24.7	4.4	9.2	5.1	17.3					
7 . Profession																	
CADR - cadre	1.49	6.46	-.62	.67	.09	-1.44	-.16	2.2	3.2	.1	19.8	.2	.06	.07	.00	.32	.00
EMPL - employé	5.86	.89	-.14	-.13	-.41	.43	.13	.4	.5	6.2	6.8	.7	.02	.02	.19	.20	.02
AUTR - autre	3.76	1.96	.47	-.06	.61	-.09	-.14	3.1	.1	8.7	.2	.5	.11	.00	.19	.00	.01
CONTRIBUTION CUMULEE =								5.7	3.8	14.9	26.8	1.5					
8 . Moyenne en cours																	
ENC1 - moins de 2KF encours	2.55	3.35	.51	1.34	-.26	.14	.34	2.5	22.4	1.1	.3	2.0	.08	.54	.02	.01	.03
ENC2 - de 2 à 5 KF encours	7.19	.55	.21	-.58	-.10	-.17	-.09	1.2	11.6	.5	1.3	.4	.08	.61	.02	.05	.01
ENC3 - plus de 5 KF encours	1.37	7.11	-2.06	.52	1.03	.61	-.16	22.3	1.8	9.0	3.3	.2	.60	.04	.15	.05	.00
CONTRIBUTION CUMULEE =								26.1	35.8	10.5	4.9	2.6					
9 . Moyenne des mouvements																	
MOU1 - moins 10 KF mouvt	3.82	1.91	.67	.11	.49	.14	.01	6.6	.2	5.7	.5	.0	.24	.01	.13	.01	.00
MOU2 - de 10 à 30KF mouvt	1.85	5.00	.29	-.38	.04	.38	-.39	.6	1.3	.0	1.7	2.0	.02	.03	.00	.03	.03
MOU3 - de 30 à 50KF mouvt	2.97	2.75	-.37	-.40	-.40	.17	.82	1.5	2.3	2.9	.5	13.6	.05	.06	.06	.01	.24
MOU4 - plus de 50KF mouvt	2.47	3.50	-.81	.59	-.31	-.71	-.70	6.2	4.2	1.5	8.0	8.3	.19	.10	.03	.14	.14
CONTRIBUTION CUMULEE =								14.9	8.0	10.2	10.8	23.9					
10 . Cumul des débits																	
DEB1 - moins de 40 débits	3.91	1.84	-.02	-.72	.48	-.48	.12	.0	9.8	5.5	5.7	.4	.00	.28	.12	.12	.01
DEB2 - de 40 à 100 débits	3.64	2.05	-.23	-.37	-.47	.32	-.35	.7	2.4	5.0	2.4	3.0	.03	.07	.11	.05	.06
DEB3 - plus de 100 débits	3.56	2.12	.26	1.17	-.04	.19	.23	.9	23.5	.0	.9	1.3	.03	.64	.00	.02	.02
CONTRIBUTION CUMULEE =								1.7	35.7	10.5	8.9	4.7					
12 . Interdiction de chéquier																	
Coui - chéquier autorisé	9.57	.16	-.10	-.08	.18	-.06	.11	.4	.3	1.9	.2	.8	.06	.04	.20	.02	.07
Cnon - chéquier interdit	1.54	6.21	.61	.49	-1.10	.38	-.67	2.2	1.8	11.6	1.4	4.8	.06	.04	.20	.02	.07
CONTRIBUTION CUMULEE =								2.5	2.0	13.4	1.7	5.6					

EXTRAIT 07

COORDOONNEES ET VALEURS-TEST DES MODALITES SUR LES AXES 1 A 5

MODALITES				COORDONNEES					VALEURS-TEST				
IDEN	LIBELLE	EFF.	P.ABS	DISTO	1	2	3	4	5	1	2	3	4	5
	3 . Situation familiale													
CELB	célibataire	170	141.57	2.31	.22	-.18	.08	.71	-.38	3.1	-2.6	1.2	10.1	-5.4
MARI	marié	221	187.56	1.50	-.32	-.06	-.10	-.26	.22	-5.6	-1.1	-1.7	-4.6	3.9
DIVO	divorcé	61	93.53	4.00	.18	.04	.52	-.01	.37	1.9	.4	5.7	-.1	4.0
VEUF	veuf	16	45.34	9.32	.27	.74	-.94	-1.12	-.50	1.9	5.3	-6.7	-8.0	-3.5
	4 . Ancienneté													
ANC1	anc. 1 an ou moins	199	211.28	1.22	.48	-.02	.54	-.14	-.06	9.3	-.3	10.6	-2.7	-1.2
ANC2	anc. de 1 à 4 ans	47	48.47	8.66	.19	.38	-.08	1.10	.43	1.4	2.8	-.6	8.1	3.1
ANC3	anc. de 4 à 6 ans	69	73.32	5.38	.05	-.08	-.61	.08	.36	.5	-.8	-5.6	.7	3.3
ANC4	anc. de 6 à 12 ans	66	71.16	5.58	-.80	-.17	-.41	.41	-1.28	-7.3	-1.6	-3.8	3.8	-11.7
ANC5	anc. plus 12 ans	87	63.77	6.34	-.88	.05	-.58	-.93	.90	-7.6	.5	-5.0	-7.9	7.7
	5 . Domiciliation du salaire													
Soui	domicile salaire	316	304.77	.54	-.32	-.23	-.15	.07	.19	-9.5	-6.9	-4.5	2.0	5.6
Snon	non dimicile salaire	152	163.23	1.87	.60	.44	.29	-.13	-.36	9.5	6.9	4.5	-2.0	-5.6
	6 . Domiciliation de l'épargne													
EPA0	pas d'épargne	370	371.36	.26	.33	-.15	-.16	-.13	-.09	14.0	-6.2	-7.0	-5.7	-3.6
EPA1	moins de 10KF épargn	58	62.17	6.53	-.60	.57	.35	.57	1.03	-5.0	4.8	2.9	4.9	8.7
EPA2	de 10 à 100KF épargn	32	28.04	15.69	-2.55	.56	1.15	.63	-1.03	-13.9	3.1	6.3	3.4	-5.6
EPA3	plus de 100KF épargn	8	6.44	71.72	-2.20	.50	1.18	-.55	-.49	-5.6	1.3	3.0	-1.4	-1.2
	7 . Profession													
CADR	cadre	77	62.72	6.46	-.62	.67	.09	-1.44	-.16	-5.3	5.7	.7	-12.3	-1.3
EMPL	employé	237	247.02	.89	-.14	-.13	-.41	.43	.13	-3.2	-2.9	-9.4	9.7	3.0
AUTR	autre	154	158.26	1.96	.47	-.06	.61	-.09	-.14	7.2	-1.0	9.4	-1.4	-2.2
	8 . Moyenne en cours													
ENC1	moins de 2KF encours	98	107.51	3.35	.51	1.34	-.26	.14	.34	6.0	15.9	-3.0	1.6	4.0
ENC2	de 2 à 5 KF encours	308	302.79	.55	.21	-.58	-.10	-.17	-.09	6.2	-16.9	-3.1	-4.9	-2.6
ENC3	plus de 5 KF encours	62	57.69	7.11	-2.06	.52	1.03	.61	-.16	-16.7	4.2	8.3	5.0	-1.3
	9 . Moyenne des mouvements													
MOU1	moins 10 KF mouvt	154	161.07	1.91	.67	.11	.49	.14	.01	10.5	1.7	7.7	2.3	.1
MOU2	de 10 à 30KF mouvt	71	77.96	5.00	.29	-.38	.04	.38	-.39	2.8	-3.7	.4	3.7	-3.8
MOU3	de 30 à 50KF mouvt	129	124.90	2.75	-.37	-.40	-.40	.17	.82	-4.8	-5.2	-5.2	2.2	10.7
MOU4	plus de 50KF mouvt	114	104.07	3.50	-.81	.59	-.31	-.71	-.70	-9.4	6.8	-3.6	-8.2	-8.1
	10 . Cumul des débits													
DEB1	moins de 40 débits	171	164.54	1.84	-.02	-.72	.48	-.48	.12	-.3	-11.4	7.6	-7.6	1.9
DEB2	de 40 à 100 débits	161	153.39	2.05	-.23	-.37	-.47	.32	-.35	-3.5	-5.6	-7.1	4.8	-5.3
DEB3	plus de 100 débits	136	150.08	2.12	.26	1.17	-.04	.19	.23	3.8	17.3	-.6	2.9	3.4
	12 . Interdiction de chéquier													
Coui	chéquier autorisé	415	403.12	.16	-.10	-.08	.18	-.06	.11	-5.3	-4.2	9.6	-3.3	5.8
Cnon	chéquier interdit	53	64.88	6.21	.61	.49	-1.10	.38	-.67	5.3	4.2	-9.6	3.3	-5.8
	1 . Type de client													
BON	bon client	237	210.99	1.22	-.52	-.32	.01	-.09	.17	-10.1	-6.3	.1	-1.8	3.4
MAUV	mauvais client	231	257.02	.82	.42	.26	-.01	.07	-.14	10.1	6.3	-.1	1.8	-3.4
	2 . Age du client													
AGE1	moins de 23 ans	88	69.96	5.69	.42	-.11	.24	.19	-.27	3.8	-1.0	2.2	1.7	-2.5
AGE2	de 23 à 40 ans	150	232.79	1.01	-.01	-.04	-.05	.26	-.05	-.2	-.9	-1.1	5.5	-1.1
AGE3	de 40 à 50 ans	122	116.86	3.00	-.16	-.01	.12	-.20	.23	-2.0	-.2	1.5	-2.5	2.9
AGE4	plus de 50 ans	108	48.40	8.67	-.18	.39	-.39	-1.03	.09	-1.3	2.9	-2.9	-7.6	.6
	11 . Autorisation de découvert													
Doui	découvert autorisé	202	215.96	1.17	.12	.25	-.22	.19	-.06	2.4	5.0	-4.3	3.8	-1.2
Dnon	découvert interdit	266	252.04	.86	-.10	-.22	.18	-.16	.05	-2.4	-5.0	4.3	-3.8	1.2

EXTRAIT 08

CORRELATIONS ENTRE LES VARIABLES CONTINUES ET LES FACTEURS POUR LES AXES 1 A 5

VARIABLES		CARACTERISTIQUES				CORRELATIONS				
NUM . (IDEN)	LIBELLE COURT	EFF.	P.ABS	MOYENNE	EC.TYPE	1	2	3	4	5
13 .	(VAR1) continue 1	468	468.00	526.38	171.36	-.56	.01	-.33	-.31	.06
14 .	(VAR2) continue 2	468	468.00	494.52	165.17	-.84	-.39	.04	-.16	.04

EXTRAIT 09

DICTIONNAIRE : ---- exemple pour un score ----

1 .	Type de client	(2 MODALITES)
2 .	Age du client	(4 MODALITES)
3 .	Situation familiale	(4 MODALITES)
4 .	Ancienneté	(5 MODALITES)
5 .	Domiciliation du salaire	(2 MODALITES)
6 .	Domiciliation de l'épargne	(4 MODALITES)
7 .	Profession	(3 MODALITES)
8 .	Moyenne en cours	(3 MODALITES)
9 .	Moyenne des mouvements	(4 MODALITES)
10 .	Cumul des débits	(3 MODALITES)
11 .	Autorisation de découvert	(2 MODALITES)
12 .	Interdiction de chéquier	(2 MODALITES)
13 .	continue 1	(CONTINUE)
14 .	continue 2	(CONTINUE)
15 .	coefficient de ponderation	(CONTINUE)
16 .	AXE FACTORIEL NUMERO 1	(CONTINUE)
17 .	AXE FACTORIEL NUMERO 2	(CONTINUE)
18 .	AXE FACTORIEL NUMERO 3	(CONTINUE)
19 .	AXE FACTORIEL NUMERO 4	(CONTINUE)
20 .	AXE FACTORIEL NUMERO 5	(CONTINUE)
21 .	AXE FACTORIEL NUMERO 6	(CONTINUE)
22 .	AXE FACTORIEL NUMERO 7	(CONTINUE)
23 .	AXE FACTORIEL NUMERO 8	(CONTINUE)
24 .	AXE FACTORIEL NUMERO 9	(CONTINUE)
25 .	AXE FACTORIEL NUMERO 10	(CONTINUE)
26 .	AXE FACTORIEL NUMERO 11	(CONTINUE)
27 .	AXE FACTORIEL NUMERO 12	(CONTINUE)
28 .	AXE FACTORIEL NUMERO 13	(CONTINUE)
29 .	AXE FACTORIEL NUMERO 14	(CONTINUE)
30 .	AXE FACTORIEL NUMERO 15	(CONTINUE)
31 .	AXE FACTORIEL NUMERO 16	(CONTINUE)
32 .	AXE FACTORIEL NUMERO 17	(CONTINUE)
33 .	AXE FACTORIEL NUMERO 18	(CONTINUE)
34 .	AXE FACTORIEL NUMERO 19	(CONTINUE)
35 .	AXE FACTORIEL NUMERO 20	(CONTINUE)

EXTRAIT 10

```
         ESTIMATION AVEC CONSTANTE ET  20 VARIABLES        DDL(STUDENT) = 447

            R**2=     .414   F(R2)=    15.7724  PROBA=  .0000    V-TEST= 13.2406

            VAR     COEFFICIENT        STUDENT       PROBA       VALEUR-TEST
            AXE1        -.9118         12.9389         .0000        11.9181
            AXE2        -.6398          8.0553         .0000         7.7796
            AXE3         .0154           .1721         .8635          .1720
            AXE4        -.2047          2.2451         .0252         2.2376
            AXE5         .4104          4.3449         .0000         4.2976
            AXE6         .0499           .4992         .6179          .4988
            AXE7        -.2710          2.6295         .0088         2.6180
            AXE8         .3774          3.6481         .0003         3.6194
            AXE9         .1513          1.4467         .1487         1.4442
            AX10        -.1986          1.8027         .0721         1.7985
            AX11         .0387           .3402         .7338          .3400
            AX12        -.4158          3.5318         .0005         3.5055
            AX13        -.1540          1.2584         .2089         1.2566
            AX14         .2047          1.6113         .1078         1.6080
            AX15         .3940          2.9845         .0030         2.9681
            AX16        -.2186          1.6388         .1020         1.6354
            AX17         .0917           .6502         .5159          .6497
            AX18         .2978          2.0681         .0392         2.0620
            AX19         .0210           .1108         .9118          .1108
            AX20         .1455           .5553         .5790          .5549
```

EXTRAIT 11

```
    EVOLUTION DU : R**2 EN FONCTION DU NOMBRE DE VARIABLES SELECTIONNEES

                .025                                                                                .414
               --I----------------------------------------------------------------------------------I--  R**2
  1 VARIABLE(S) I *              *                                *                                   I
  2 VARIABLE(S) I                                                      **             *             I
  3 VARIABLE(S) I                                                                 2 *                I
  4 VARIABLE(S) I                                                                     * 2            I
  5 VARIABLE(S) I                                                                       2*           I
  6 VARIABLE(S) I                                                                        2*          I
  7 VARIABLE(S) I                                                                         *2         I
  8 VARIABLE(S) I                                                                          *2        I
  9 VARIABLE(S) I                                                                           3        I
 10 VARIABLE(S) I                                                                           2*       I
 11 VARIABLE(S) I                                                                           2*       I
 12 VARIABLE(S) I                                                                            3       I
 13 VARIABLE(S) I                                                                            3       I
 14 VARIABLE(S) I                                                                            2*      I
 15 VARIABLE(S) I                                                                             3      I
 16 VARIABLE(S) I                                                                             3      I
 17 VARIABLE(S) I                                                                             3      I
 18 VARIABLE(S) I                                                                             3      I
 19 VARIABLE(S) I                                                                             3      I
 20 VARIABLE(S) I                                                                             *      I
               ----------------------------------------------------------------------------------
                                    SYMBOLE "S" ==> SUPERIEUR A 10 REG.
```

EXTRAIT 12

```
COMMANDE  1
          V1 = V1 -- V15

TABLEAU DE CLASSEMENT
---------------------
                              POURCENTAGES DES CLASSEMENTS
                              BIEN CLASSES     MAL CLASSES      TOTAL

       GROUPES D'ORIGINE ----------------------------------------------
                         BON    199.00           38.00          237.00
                                 ( 83.97)        ( 16.03)        (100.00)
                          ---------------------------------------------
                         MAUV    166.00           65.00          231.00
                                 ( 71.86)        ( 28.14)        (100.00)
                          ---------------------------------------------
                        TOTAL    365.00          103.00          468.00
                                 ( 77.99)        ( 22.01)        (100.00)


COMMANDE  2
          V1 = V1 + V2 + V4 + V5 + V8 + V12

TABLEAU DE CLASSEMENT
                              POURCENTAGES DES CLASSEMENTS
                              BIEN CLASSES     MAL CLASSES      TOTAL

       GROUPES D'ORIGINE ----------------------------------------------
                         BON    195.00           42.00          237.00
                                 ( 82.28)        ( 17.72)        (100.00)
                          ---------------------------------------------
                         MAUV    173.00           58.00          231.00
                                 ( 74.89)        ( 25.11)        (100.00)
                          ---------------------------------------------
                        TOTAL    368.00          100.00          468.00
                                 ( 78.63)        ( 21.37)        (100.00)
```

EXTRAIT 13

```
COMMANDE  3
          V1 = V1 -- V12


TABLEAU DE CLASSEMENT
                              POURCENTAGES DES CLASSEMENTS
                              BIEN CLASSES     MAL CLASSES      TOTAL

       GROUPES D'ORIGINE ------------------------------------------------
                         BON    202.00           35.00          237.00
                                 ( 85.23)        ( 14.77)        (100.00)
                          -------------------------------------------------
                        MAUV    165.00           66.00          231.00
                                 ( 71.43)        ( 28.57)        (100.00)
                          -------------------------------------------------
                       TOTAL    367.00          101.00          468.00
                                 ( 78.42)        ( 21.58)        (100.00)


IDENTIFICATEURS DES INDIVIDUS MAL CLASSES

 GROUPE    BON :
 013   023   028   039   047   052   053   054   075   089   092   094   105   106   128   137   139   149   150   164
 168   171   190   204   208   212   229   240   242   243   262   270   276   277   284

 GROUPE   MAUV :
 027   032   051   060   063   072   085   095   112   129   133   156   157   158   169   173   174   180   206   234
 250   257   264   287   303   308   313   323   327   329   332   337   342   346   349   352   353   355   356   357
 359   361   385   392   393   394   397   403   405   416   420   422   428   431   437   439   441   442   444   446
 449   456   462   463   467   468


 FONCTION LINEAIRE DISCRIMINANTE

   FACTEURS    . CORRELATIONS  . COEFFICIENTS  . COEFFICIENTS  .     ECARTS    .    T    DE    .  PROBABILITE
...............    FACTEURS    .   FONCTION    .  REGRESSION   .     TYPES     .    STUDENT    .
NUMERO . NOM  . AVEC  F.L.D   . DISCRIMINANTE .               .  (RESULTATS TYPE REGRESSION)
       .      . (SEUIL=  .09) .               .               .               .               .
.....................................................................................................................
   1   . F  1 .    -.472      .    -3.006464  .     -.920575  .       .0716   .     12.8608   .      .0000
   2   . F  2 .    -.294      .    -2.109493  .     -.645924  .       .0807   .      8.0067   .      .0000
   3   . F  3 .     .006      .      .050892  .      .015583  .       .0911   .       .1710   .      .8643
   4   . F  4 .    -.082      .     -.674985  .     -.206680  .       .0926   .      2.2316   .      .0261
   5   . F  5 .     .158      .     1.353283  .      .414373  .       .0959   .      4.3186   .      .0000
   6   . F  6 .     .018      .      .164436  .      .050350  .       .1015   .       .4962   .      .6200
   7   . F  7 .    -.096      .     -.893561  .     -.273607  .       .1047   .      2.6136   .      .0093
   8   . F  8 .     .133      .     1.244243  .      .380985  .       .1051   .      3.6261   .      .0003
   9   . F  9 .     .053      .      .498918  .      .152768  .       .1062   .      1.4380   .      .1511
  10   . F 10 .    -.066      .     -.654849  .     -.200514  .       .1119   .      1.7919   .      .0738
  11   . F 11 .     .012      .      .127736  .      .039113  .       .1157   .       .3382   .      .7354
  12   . F 12 .    -.129      .    -1.370991  .     -.419796  .       .1196   .      3.5104   .      .0005
   CONSTANTE  .              .     -.125337  .     -.043380  .       .0367   .      1.1821   .      .2378
.....................................................................................................................
        R2 =     .39017     F  =   24.25911     PROBA =     .00000
        D2 =    2.54869     T2 =  298.14710     PROBA =     .00000
.....................................................................................................................
```

EXTRAIT 14

FONCTION LINEAIRE DE FISHER RECONSTITUEE A PARTIR DES VARIABLES D'ORIGINE

VARIABLES NUMERO	NOM	COEFFICIENTS FONCTION DISCRIMINANTE	COEFFICIENTS REGRESSION	ECARTS TYPES (RESULTATS TYPE REGRESSION)	T DE STUDENT	PROBABILITE
3	CELB	5.958140	.802118	.4293	1.8684	.0624
3	MARI	.467870	1.101930	.3195	3.4490	.0006
3	DIVO	-6.531761	-2.098393	.5767	3.6388	.0003
3	VEUF	-2.703831	-2.734483	.6714	4.0731	.0001
4	ANC1	-1.385484	-1.386593	.2473	5.6065	.0000
4	ANC2	2.613549	-.045623	.9559	.0477	.9620
4	ANC3	4.578146	1.463101	.6742	2.1702	.0305
4	ANC4	1.597571	.506278	.5153	.9825	.3264
4	ANC5	-.382814	2.381410	.4742	5.0214	.0000
5	Soui	-.422235	1.035891	.1339	7.7351	.0000
5	Snon	-.727415	-1.934087	.2500	7.7351	.0000
6	EPA0	.399745	-.297430	.1044	2.8482	.0046
6	EPA1	-1.795711	.684169	.6658	1.0276	.3047
6	EPA2	-1.093578	1.704108	.7183	2.3723	.0181
6	EPA3	2.715195	3.130641	.5575	5.6155	.0000
7	CADR	7.801423	2.269371	.5701	3.9806	.0001
7	EMPL	-.375002	.131350	.1979	.6638	.5072
7	AUTR	-.891039	-1.104337	.3248	3.3998	.0007
8	ENC1	2.754292	-2.086161	.3549	5.8779	.0000
8	ENC2	-.029124	.379077	.1323	2.8663	.0044
8	ENC3	-.995652	1.898147	.4500	4.2179	.0000
9	MOU1	-1.657008	-1.815740	.2932	6.1934	.0000
9	MOU2	1.036308	-.211874	.6621	.3200	.7491
9	MOU3	-.292530	2.138614	.4153	5.1490	.0000
9	MOU4	-1.036646	.402188	.3567	1.1276	.2601
10	DEB1	2.719447	2.082835	.3019	6.8990	.0000
10	DEB2	-4.311916	-.709685	.3493	2.0316	.0428
10	DEB3	.790287	-1.558173	.2691	5.7906	.0000
12	Coui	.769765	.728917	.0880	8.2824	.0000
12	Cnon	-9.281121	-4.528937	.5468	8.2825	.0000
CONSTANTE		-.125337	-.043380			

Extrait 15
En tête de listage, on rappelle le libellé de la variable indicatrice de groupe, déjà utilisée dans DIS2G. On indique ensuite quelle modalité a été choisie par l'utilisateur pour être affectée au groupe des "scores forts". Les individus dans ce groupe devraient avoir les valeurs élevées du score.

Extrait 16
Le tableau liste en parallèle les coefficients de la fonction linéaire discriminante attribués aux modalités et les coefficients de la fonction de score. Les deux fonctions affectent les individus de la même façon dans les deux groupes.

Extrait 17
La tolérance d'erreur de classement (TEC) est un pourcentage porté en abscisse du graphique. Le graphique indique comment varient les zones rouge (R), verte (V) et centrale en fonction de la valeur du TEC.

Ce graphique peut être utilisé pour comparer deux fonctions discriminantes. La forme "idéale" de ce graphique est facile à imaginer. Pour une fonction de score particulière, l'utilisateur consultera ce graphique pour choisir une valeur de TEC qui convienne à l'application.

Extrait 18
La courbe est tracée en fonction des scores portés en ordonnée. Pour chaque valeur du score (en fait ici pour une tranche de valeurs), elle fournit l'estimation de la probabilité d'appartenir au groupe des scores faibles (groupe 2). Cette probabilité décroît quand le score augmente.
Il s'agit de la probabilité conditionnelle théorique, calculée sous les hypothèses classiques de travail en analyse discriminante. Les résultats sont déjà édités, sous une autre forme, par la procédure DIS2G.

Le graphique fournit une abaque de référence, éditée entre les bornes où son utilisation peut être intéressante. Si la taille de l'échantillon est assez grande, la forme de la courbe apparaît dès que l'on effectue le calcul pour les individus de l'échantillon (paramètre LABAQ). Pour les petits échantillons, on demandera le calcul point par point de la courbe.

Extrait 19
L'échantillon est redressé pour tenir compte d'une sur-représentation volontaire dans l'échantillon des individus du groupe des scores faibles. Ainsi pour assurer une répartition 15%-85% on voit qu'il faut attribuer le poids 0.145 à chaque individu du groupe 2. Le reste des calculs concernant la population globale tiendra compte automatiquement de ce redressement.

Extrait 20
Ce tableau synthétique donne la répartition des échantillons dans les zones rouge, verte, et centrale, pour une tolérance d'erreur de classement fixée à 10%. On peut contrôler ici la pertinence de choix de la valeur TEC. Le tableau indique les valeurs du score qui définissent les frontières des zones.

Extrait 21
Le graphique fournit la répartition de la population globale en fonction des scores. Cette distribution est estimée en tenant compte de la correction due à la sur-représentation des individus du groupe 2 dans l'échantillon.
On lit que, pour un TEC de 10%, on définit une zone rouge qui contiendra environ 14% de la population, et une zone verte qui en contiendra 19%. Il reste donc environ 67% dans la zone centrale.

Extrait 22
Ce graphique fournit la distribution des deux groupes définis par la variable indicatrice de groupe, en fonction du score des individus.
On imagine quelle serait la représentation correspondant à une discrimination parfaite. La forme du graphique est une indication d'une bonne ou d'une médiocre discrimination.

Extrait 23

A partir de cet extrait commence une nouvelle exécution de SCORE, pour laquelle les coefficients sont "forcés" manuellement. L'utilisateur est averti éventuellement du non-respect du score maximum annoncé.

Extrait 24
Le tableau fournit les coefficients de la fonction score, non pas dans l'ordre naturel des variables, mais rangés en ordre décroissant à partir du maximum dans chaque variable. Cette édition est utilisée pour étudier la contribution relative des variables au calcul du score de l'individu.

Ainsi on voit que la plus forte contribution au score est obtenue par la variable "situation familiale", pour la modalité "célibataire". Les autres modalités sont rangées par contribution décroissante au score. La seconde variable sera "interdiction de chéquier", etc. On voit que les caractéristiques citées en queue de liste ont peu d'effet sur le score final d'un individu.

Extrait 25
Ce graphique est analogue à celui de l'extrait 17. Il ne doit pas en différer de façon appréciable si les coefficients forcés par l'utilisateur ne l'éloignent pas de la solution discriminante. Il s'agit d'une édition de contrôle.

Extrait 26
Le tableau de redressement est identique au tableau de l'extrait 19. La distribution de l'échantillon dans les 3 zones définies par TEC = 10% ne devrait pas différer beaucoup de celle éditée dans l'extrait 20.

Dans le dernier tableau, on appelle "*rapport d'efficacité*" dans une certaine zone de scores le nombre d'individus du groupe 2 pour un individu du groupe 1. On édite ici la valeur de ce rapport à proximité des frontières de zone. Ainsi dans la zone rouge, à proximité du seuil frontière score=390, on observe au mieux 3 "groupe-2"pour un "groupe-1". Dans la zone verte, à proximité de la frontière score=635, on observe au pire 7 "groupe-2" pour un "groupe-1". En moyenne ce rapport augmente quand on monte sur l'échelle des scores.

Extrait 27
Ce tableau numérique fournit la répartition des échantillons dans les zones en fonction du score. Il correspond au graphique de l'extrait 30 qui suit.

Extrait 28

Ce tableau numérique correspond aux courbes des extraits 29 et 31. Il s'agit de la répartition de la population en fonction du score. Cette répartition est estimée en tenant compte de la sur-représentation d'un groupe. On trouve dans ce tableau le calcul du rapport d'efficacité par tranche de scores, ainsi qu'une valeur lissée (par moyenne mobile) qui indique plus clairement la tendance.

Extrait 29
La répartition de la population est à comparer à l'extrait 21.

Extrait 30
Les répartitions des groupes sont à comparer avec celles de l'extrait 22.

Extrait 31
Le graphique fournit la courbe du rapport d'efficacité et son lissage.

EXTRAIT 15

```
ETUDE DE LA FONCTION SCORE
POUR LA VARIABLE DE GROUPE :

    Type de client
    MODALITE 1 : " GROUPE DES SCORES FORTS   " =  bon client           211.
    MODALITE 2 : " GROUPE DES SCORES FAIBLES " =   mauvais    client   257.

    ENSEMBLE                                                          468.
```

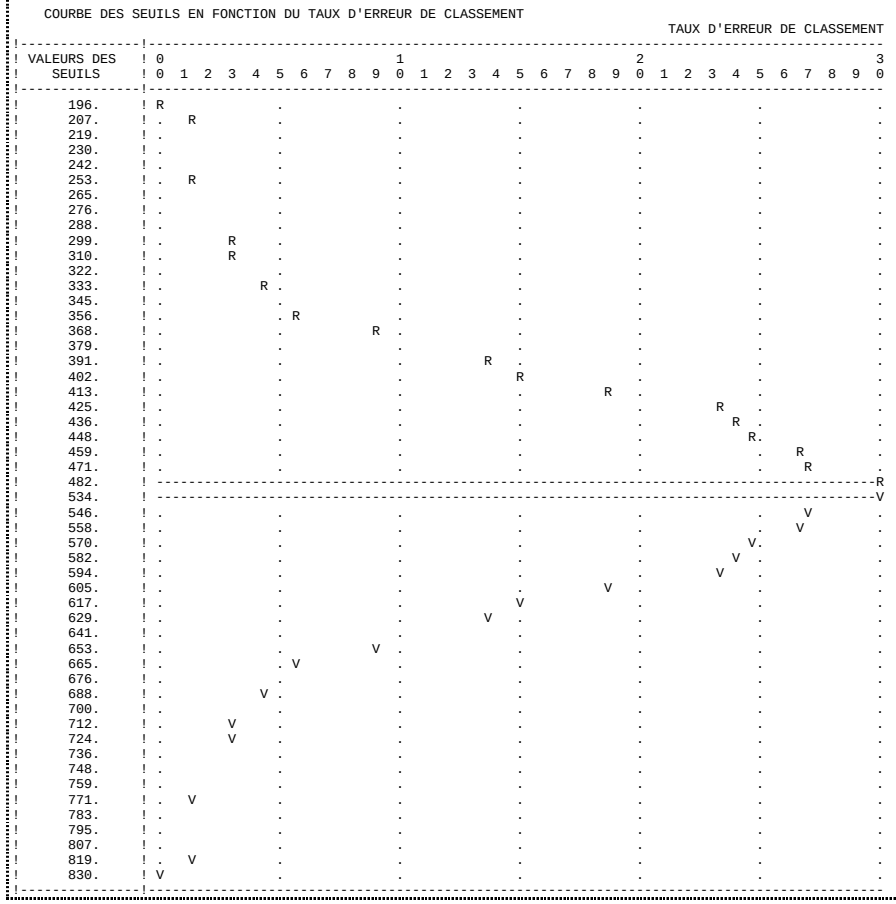
EXTRAIT 16

COEFFICIENTS DES FONCTIONS DISCRIMINANTE ET SCORE

IDEN	LIBELLES	COEFFICIENTS FONCTION DISCRIMINANTE	COEFFICIENTS TRANSFORMES (SCORE)
3	. Situation familiale		
CELB	- célibataire	5.958	225.09
MARI	- marié	.468	126.15
DIVO	- divorcé	-6.532	.00
VEUF	- veuf	-2.704	68.99
4	. Ancienneté		
ANC1	- anc. 1 an ou moins	-1.385	.00
ANC2	- anc. de 1 à 4 ans	2.614	72.07
ANC3	- anc. de 4 à 6 ans	4.578	107.48
ANC4	- anc. de 6 à 12 ans	1.598	53.76
ANC5	- anc. plus 12 ans	-.383	18.07
5	. Domiciliation du salaire		
Soui	- domicile salaire	-.422	5.50
Snon	- non dimicile salaire	-.727	.00
6	. Domiciliation de l'épargne		
EPA0	- pas d'épargne	.400	39.57
EPA1	- moins de 10KF épargn	-1.796	.00
EPA2	- de 10 à 100KF épargn	-1.094	12.65
EPA3	- plus de 100KF épargn	2.715	81.30
7	. Profession		
CADR	- cadre	7.801	156.66
EMPL	- employé	-.375	9.30
AUTR	- autre	-.891	.00
8	. Moyenne en cours		
ENC1	- moins de 2KF encours	2.754	67.58
ENC2	- de 2 à 5 KF encours	-.029	17.42
ENC3	- plus de 5 KF encours	-.996	.00
9	. Moyenne des mouvements		
MOU1	- moins 10 KF mouvt	-1.657	.00
MOU2	- de 10 à 30KF mouvt	1.036	48.54
MOU3	- de 30 à 50KF mouvt	-.293	24.59
MOU4	- plus de 50KF mouvt	-1.037	11.18
10	. Cumul des débits		
DEB1	- moins de 40 débits	2.719	126.72
DEB2	- de 40 à 100 débits	-4.312	.00
DEB3	- plus de 100 débits	.790	91.95
12	. Interdiction de chéquier		
Coui	- chéquier autorisé	.770	181.14
Cnon	- chéquier interdit	-9.281	.00

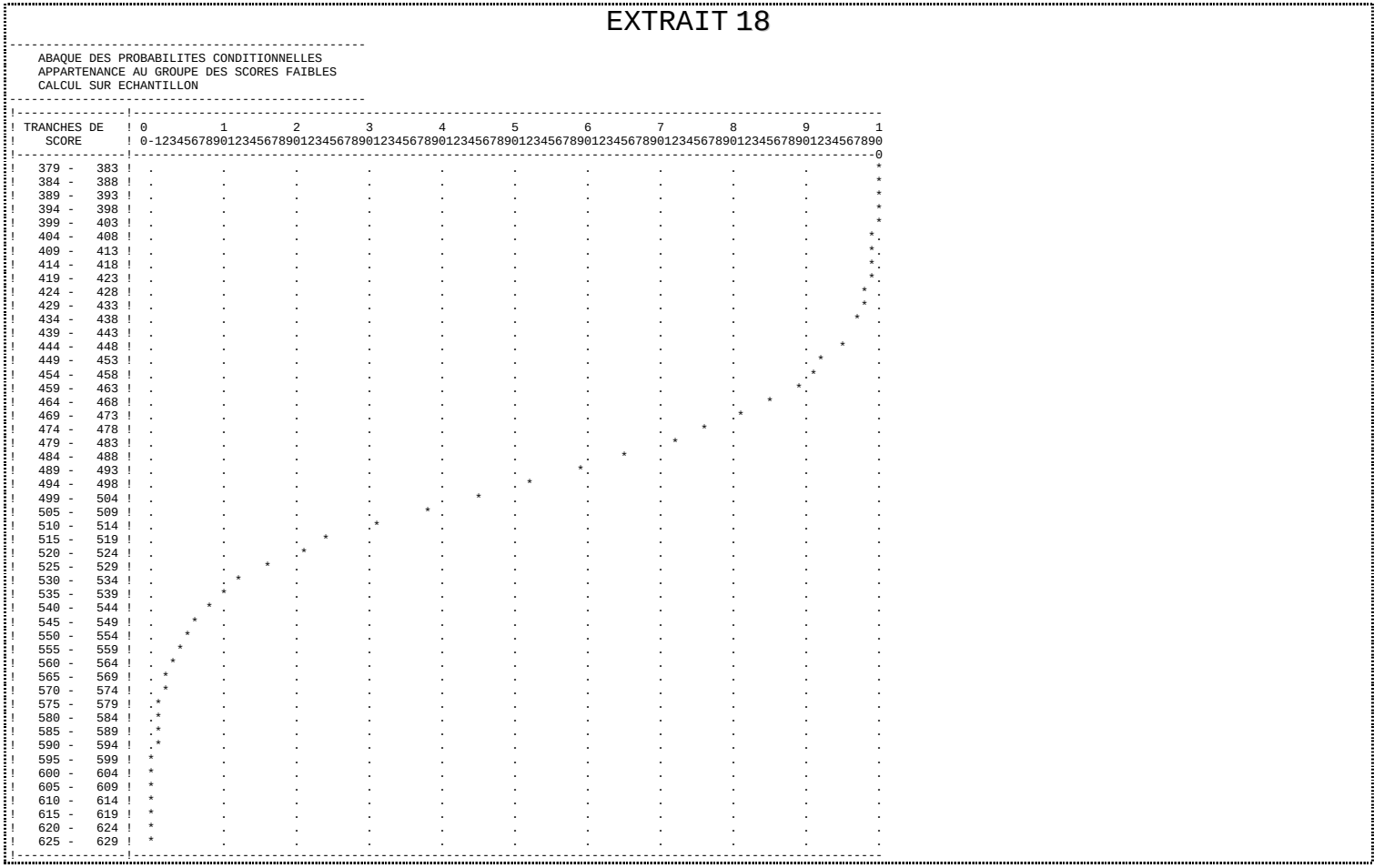
EXTRAIT 17

COURBE DES SEUILS EN FONCTION DU TAUX D'ERREUR DE CLASSEMENT



EXTRAIT 18

```
------------------------------------------------
   ABAQUE DES PROBABILITES CONDITIONNELLES
   APPARTENANCE AU GROUPE DES SCORES FAIBLES
   CALCUL SUR ECHANTILLON
------------------------------------------------
```



EXTRAIT 19

REDRESSEMENT DE L'ECHANTILLON :

POUR UN POURCENTAGE DE GROUPE DES SCORES FAIBLES THEORIQUE DE : 15.00 %
POIDS ATTRIBUE A UN INDIVIDU DU GROUPE DES SCORES FAIBLES : .145

	EFFECTIF ECHANTILLON	POURCENTAGE ECHANTILLON	POURCENTAGE THEORIQUE
GROUPE DES SCORES FORTS	211.	45.08	85.00
GROUPE DES SCORES FAIBLES	257.	54.92	15.00
ENSEMBLE	468.	100.00	100.00

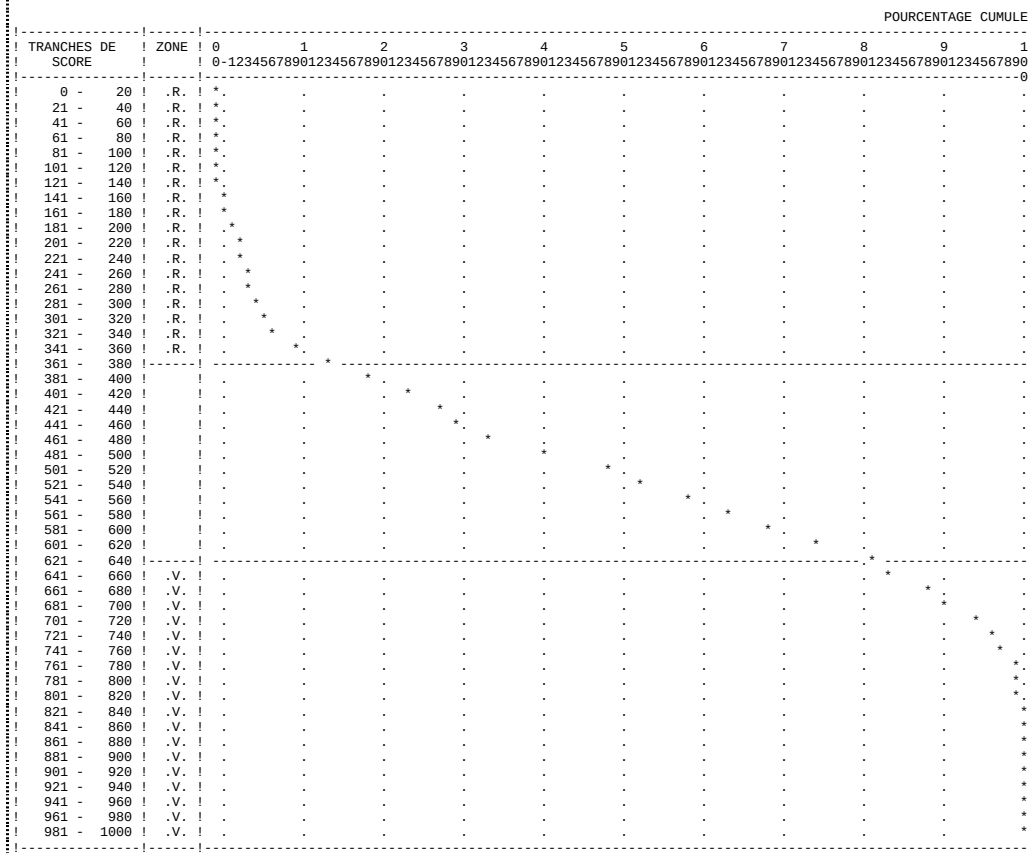
EXTRAIT 20

DISTRIBUTION DE L'ECHANTILLON DANS LES ZONES DE CLASSEMENT
POUR LE TAUX D'ERREUR DE CLASSEMENT TOLERE : 10.00 %

SEUIL DES SCORES 0 --- 380. --- 629. --- 1000.	ZONE ROUGE	ZONE ORANGE	ZONE VERTE	
GR. SCORES FORTS				
EFFECTIFS	20.	135.	56.	211.
POURCENTAGES	9.36	63.90	26.74	100.00
GR. SCORES FAIBLES				
EFFECTIFS	71.	160.	26.	257.
POURCENTAGES	27.67	62.30	10.04	100.00

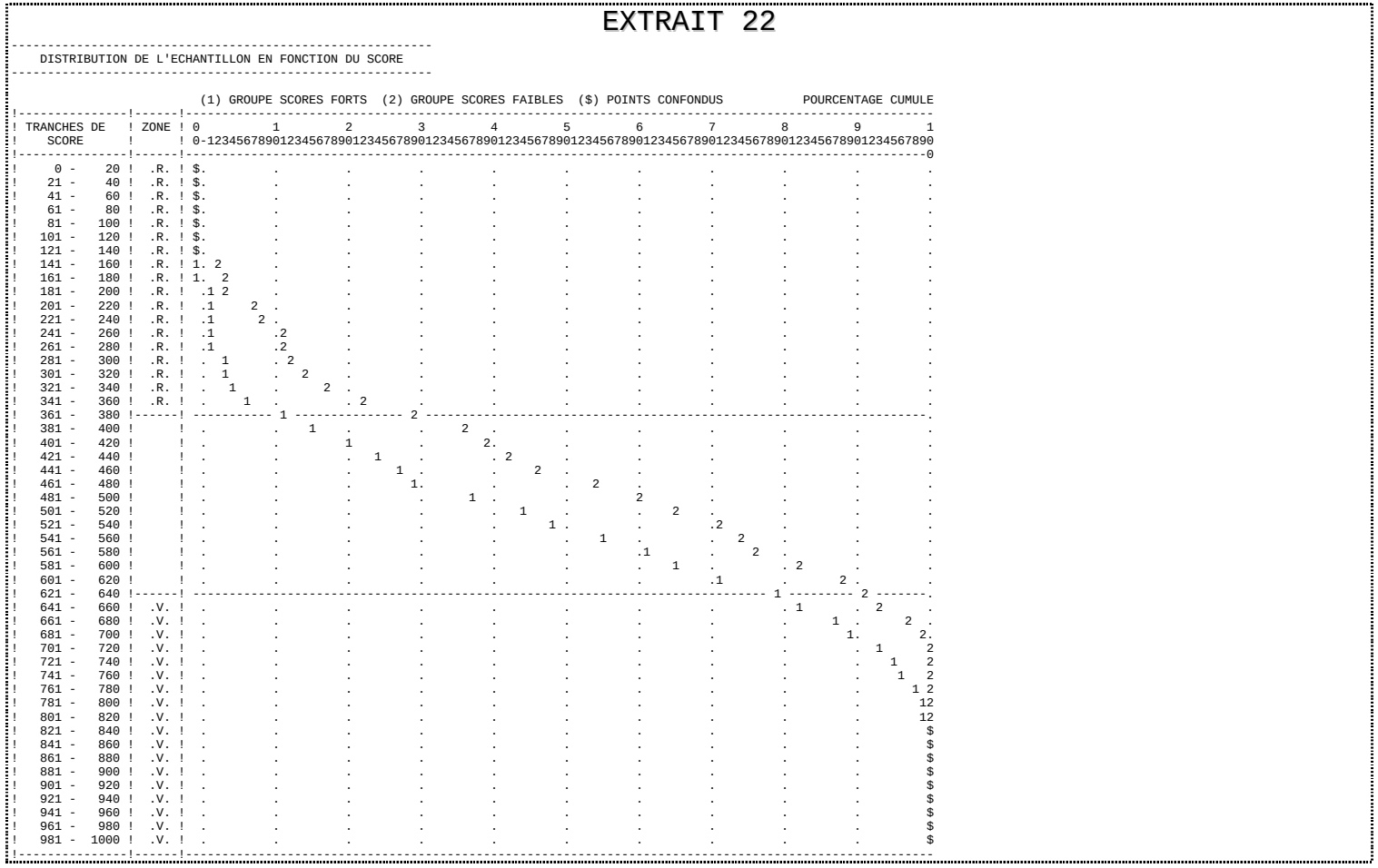
EXTRAIT 21

COURBE DE LA POPULATION EN FONCTION DU SCORE



EXTRAIT 22

DISTRIBUTION DE L'ECHANTILLON EN FONCTION DU SCORE



(1) GROUPE SCORES FORTS (2) GROUPE SCORES FAIBLES ($) POINTS CONFONDUS

TRANCHES DE SCORE	ZONE	0 1 2 3 4 5 6 7 8 9 1 / 0-1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890	POURCENTAGE CUMULE
0 - 20	.R.	$.	$
21 - 40	.R.	$.	$
41 - 60	.R.	$.	$
61 - 80	.R.	$.	$
81 - 100	.R.	$.	$
101 - 120	.R.	$.	$
121 - 140	.R.	$.	$
141 - 160	.R.	1. 2	$
161 - 180	.R.	1. 2	$
181 - 200	.R.	.1 2	$
201 - 220	.R.	.1 2	$
221 - 240	.R.	.1 2	$
241 - 260	.R.	.1 .2	$
261 - 280	.R.	.1 .2	$
281 - 300	.R.	. 1 . 2	$
301 - 320	.R.	. 1 . 2	$
321 - 340	.R.	. 1 . 2	$
341 - 360	.R.	. 1 . 2	$
361 - 380		1 2	$
381 - 400		. 1 . 2 .	$
401 - 420		. 1 2.	$
421 - 440		. 1 . 2	$
441 - 460		. 1 . 2	$
461 - 480		. 1. 2	$
481 - 500		. 1 . 2	$
501 - 520		. 1 . 2	$
521 - 540		. 1 . .2	$
541 - 560		. 1 . 2	$
561 - 580		.1 . 2	$
581 - 600		1 . 2	$
601 - 620		.1 2 .	$
621 - 640		1 2	$
641 - 660	.V.	. 1 . 2	$
661 - 680	.V.	1 . 2	$
681 - 700	.V.	1. 2.	$
701 - 720	.V.	. 1 2	$
721 - 740	.V.	1 2	$
741 - 760	.V.	1 2	$
761 - 780	.V.	1 2	$
781 - 800	.V.	12	$
801 - 820	.V.	12	$
821 - 840	.V.		$
841 - 860	.V.		$
861 - 880	.V.		$
881 - 900	.V.		$
901 - 920	.V.		$
921 - 940	.V.		$
941 - 960	.V.		$
961 - 980	.V.		$
981 - 1000	.V.		$

EXTRAIT 23

```
ETUDE DE LA FONCTION SCORE
POUR LA VARIABLE DE GROUPE :
    Type de client
                                                                  EFFECTIF
    MODALITE 1 : " GROUPE DES SCORES FORTS   " =  bon client         211.
    MODALITE 2 : " GROUPE DES SCORES FAIBLES " =   mauvais    client 257.
    ENSEMBLE                                                      468.
** ATTENTION ** (TRCO1-800)    1015.000 (1)    1000.000 (2)
LA SOMME (1) DES COEFFICIENTS MAXIMAUX FORCES EST
SUPERIEURE A LA SOMME SXTOT (2)
```

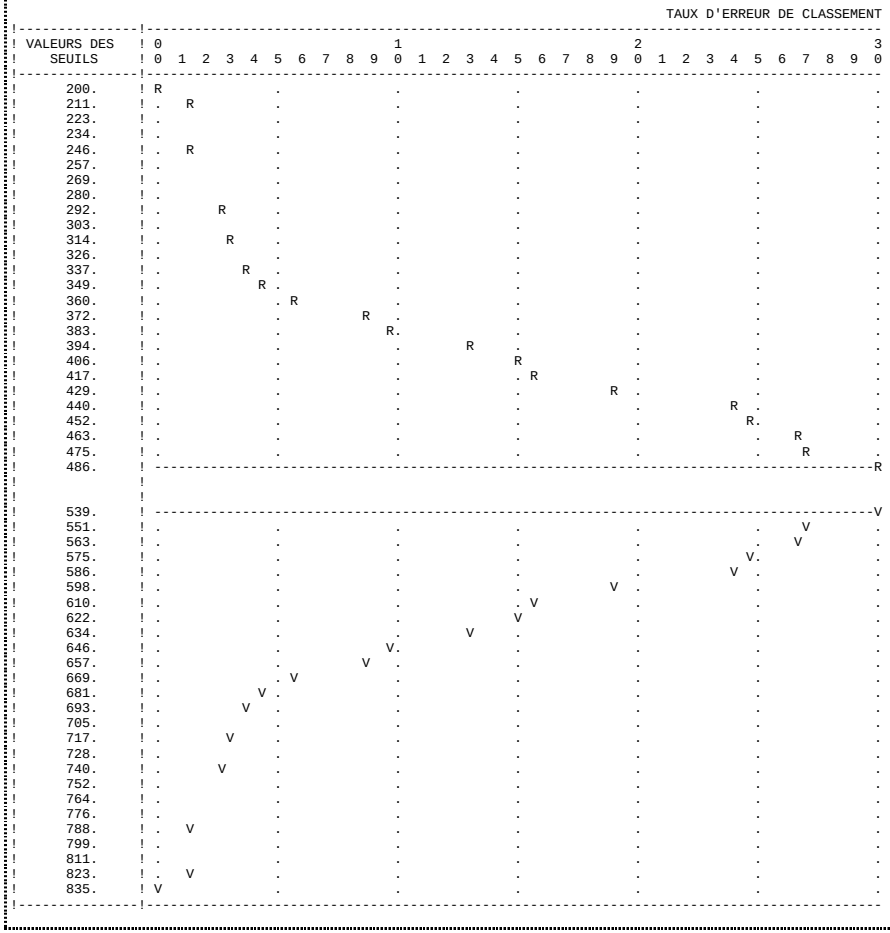
EXTRAIT 24

COEFFICIENTS REORDONNES DE LA FONCTION SCORE

IDEN	LIBELLES	COEFFICIENTS DU SCORE	HISTOGRAMMES DES POINTS DE SCORE
3	. Situation familiale		
CELB	- célibataire	225.00	**********
MARI	- marié	130.00	******
VEUF	- veuf	70.00	***
DIVO	- divorcé	.00	*
12	. Interdiction de chéquier		
Coui	- chéquier autorisé	180.00	********
Cnon	- chéquier interdit	.00	*
7	. Profession		
CADR	- cadre	160.00	********
EMPL	- employé	10.00	*
AUTR	- autre	.00	*
10	. Cumul des débits		
DEB1	- moins de 40 débits	130.00	******
DEB3	- plus de 100 débits	90.00	****
DEB2	- de 40 à 100 débits	.00	*
4	. Ancienneté		
ANC3	- anc. de 4 à 6 ans	110.00	*****
ANC2	- anc. de 1 à 4 ans	70.00	***
ANC4	- anc. de 6 à 12 ans	50.00	**
ANC5	- anc. plus 12 ans	20.00	*
ANC1	- anc. 1 an ou moins	.00	*
6	. Domiciliation de l'épargne		
EPA3	- plus de 100KF épargn	80.00	****
EPA0	- pas d'épargne	40.00	**
EPA2	- de 10 à 100KF épargn	10.00	*
EPA1	- moins de 10KF épargn	.00	*
8	. Moyenne en cours		
ENC1	- moins de 2KF encours	70.00	***
ENC2	- de 2 à 5 KF encours	20.00	*
ENC3	- plus de 5 KF encours	.00	*
9	. Moyenne des mouvements		
MOU2	- de 10 à 30KF mouvt	50.00	**
MOU3	- de 30 à 50KF mouvt	20.00	*
MOU4	- plus de 50KF mouvt	10.00	*
MOU1	- moins 10 KF mouvt	.00	*
5	. Domiciliation du salaire		
Soui	- domicile salaire	10.00	*
Snon	- non dimicile salaire	.00	*

EXTRAIT 25

COURBE DES SEUILS EN FONCTION DU TAUX D'ERREUR DE CLASSEMENT



EXTRAIT 26

REDRESSEMENT DE L'ECHANTILLON :

POUR UN POURCENTAGE DE GROUPE DES SCORES FAIBLES THEORIQUE DE : 15.00 %
POIDS ATTRIBUE A UN INDIVIDU DU GROUPE DES SCORES FAIBLES : .145

	EFFECTIF ECHANTILLON	POURCENTAGE ECHANTILLON	POURCENTAGE THEORIQUE
GROUPE DES SCORES FORTS	211.	45.08	85.00
GROUPE DES SCORES FAIBLES	257.	54.92	15.00
ENSEMBLE	468.	100.00	100.00

DISTRIBUTION DE L'ECHANTILLON DANS LES ZONES DE CLASSEMENT
POUR LE TAUX D'ERREUR DE CLASSEMENT TOLERE : 10.00 %

SEUIL DES SCORES	ZONE ROUGE	ZONE ORANGE	ZONE VERTE	
	0---------	390.--------	635.-------	1000.
GR. SCORES FORTS				
EFFECTIFS	21.	134.	56.	211.
POURCENTAGES	9.75	63.59	26.66	100.00
GR. SCORES FAIBLES				
EFFECTIFS	86.	145.	26.	257.
POURCENTAGES	33.47	56.61	9.92	100.00

ETUDE DE L'EFFICACITE: NOMBRE DE " FORTS " POUR 1 " FAIBLE "
POUR LE TAUX D'ERREUR DE CLASSEMENT TOLERE DE 10.00 %

	ZONE ROUGE	ZONE ORANGE ETENDUE	ZONE VERTE
SEUIL DES SCORES	0--------	390.-------- 25. % ---	635.----- 1000.
EFFICACITE	MAX = 3/1	SAUT = 4	MIN = 7/1

EXTRAIT 27

REPARTITION DE L'ECHANTILLON SELON LES TRANCHES DE SCORE

TRANCHES DES SCORES	ZONE	GROUPE DES SCORES FORTS EFF.	POURC.	POURC. CUMULES	GROUPE DES SCORES FAIBLES EFF.	POURC.	POURC. CUMULES
0 - 20	.R.	0.	.00	.00	0.	.00	.00
21 - 40	.R.	0.	.00	.00	0.	.00	.00
41 - 60	.R.	0.	.00	.00	0.	.00	.00
61 - 80	.R.	0.	.00	.00	0.	.00	.00
81 - 100	.R.	0.	.00	.00	0.	.00	.00
101 - 120	.R.	0.	.00	.00	0.	.00	.00
121 - 140	.R.	0.	.00	.00	0.	.00	.00
141 - 160	.R.	0.	.00	.00	5.	1.97	1.97
161 - 180	.R.	0.	.00	.00	3.	.97	2.94
181 - 200	.R.	2.	.74	.74	0.	.00	2.94
201 - 220	.R.	0.	.23	.97	10.	3.73	6.67
221 - 240	.R.	0.	.23	1.20	3.	1.30	7.97
241 - 260	.R.	0.	.00	1.20	9.	3.34	11.31
261 - 280	.R.	0.	.00	1.20	0.	.00	11.31
281 - 300	.R.	3.	1.19	2.39	1.	.28	11.59
301 - 320	.R.	1.	.41	2.80	4.	1.47	13.06
321 - 340	.R.	3.	1.32	4.12	11.	4.26	17.33
341 - 360	.R.	3.	1.49	5.61	10.	3.89	21.22
361 - 380	.R.	8.	3.75	9.36	24.	9.36	30.58
381 - 400		12.	5.49	14.85	12.	4.75	35.33
401 - 420		8.	4.00	18.85	6.	2.43	37.76
421 - 440		11.	5.11	23.96	10.	3.75	41.51
441 - 460		6.	2.63	26.58	18.	6.98	48.49
461 - 480		3.	1.20	27.79	11.	4.32	52.81
481 - 500		13.	6.20	33.99	14.	5.46	58.27
501 - 520		22.	10.47	44.46	16.	6.38	64.65
521 - 540		4.	1.92	46.38	15.	5.76	70.41
541 - 560		13.	6.00	52.38	9.	3.56	73.97
561 - 580		16.	7.75	60.13	5.	2.08	76.05
581 - 600		8.	3.69	63.82	12.	4.60	80.65
601 - 620		11.	5.40	69.22	19.	7.29	87.94
621 - 640		12.	5.61	74.82	9.	3.53	91.47
641 - 660	.V.	13.	6.23	81.05	5.	1.88	93.34
661 - 680	.V.	10.	4.71	85.76	8.	2.94	96.29
681 - 700	.V.	6.	2.82	88.58	4.	1.57	97.86
701 - 720	.V.	5.	2.55	91.13	4.	1.56	99.42
721 - 740	.V.	6.	2.64	93.78	1.	.32	99.75
741 - 760	.V.	5.	2.40	96.17	0.	.00	99.75
761 - 780	.V.	3.	1.22	97.39	0.	.00	99.75
781 - 800	.V.	4.	2.02	99.41	1.	.25	100.00
801 - 820	.V.	0.	.00	99.41	0.	.00	100.00
821 - 840	.V.	1.	.59	100.00	0.	.00	100.00
841 - 860	.V.	0.	.00	100.00	0.	.00	100.00
861 - 880	.V.	0.	.00	100.00	0.	.00	100.00
881 - 900	.V.	0.	.00	100.00	0.	.00	100.00
901 - 920	.V.	0.	.00	100.00	0.	.00	100.00
921 - 940	.V.	0.	.00	100.00	0.	.00	100.00
941 - 960	.V.	0.	.00	100.00	0.	.00	100.00
961 - 980	.V.	0.	.00	100.00	0.	.00	100.00
981 - 1000	.V.	0.	.00	100.00	0.	.00	100.00
ENSEMBLE		211.	100.00	100.00	257.	100.00	100.00

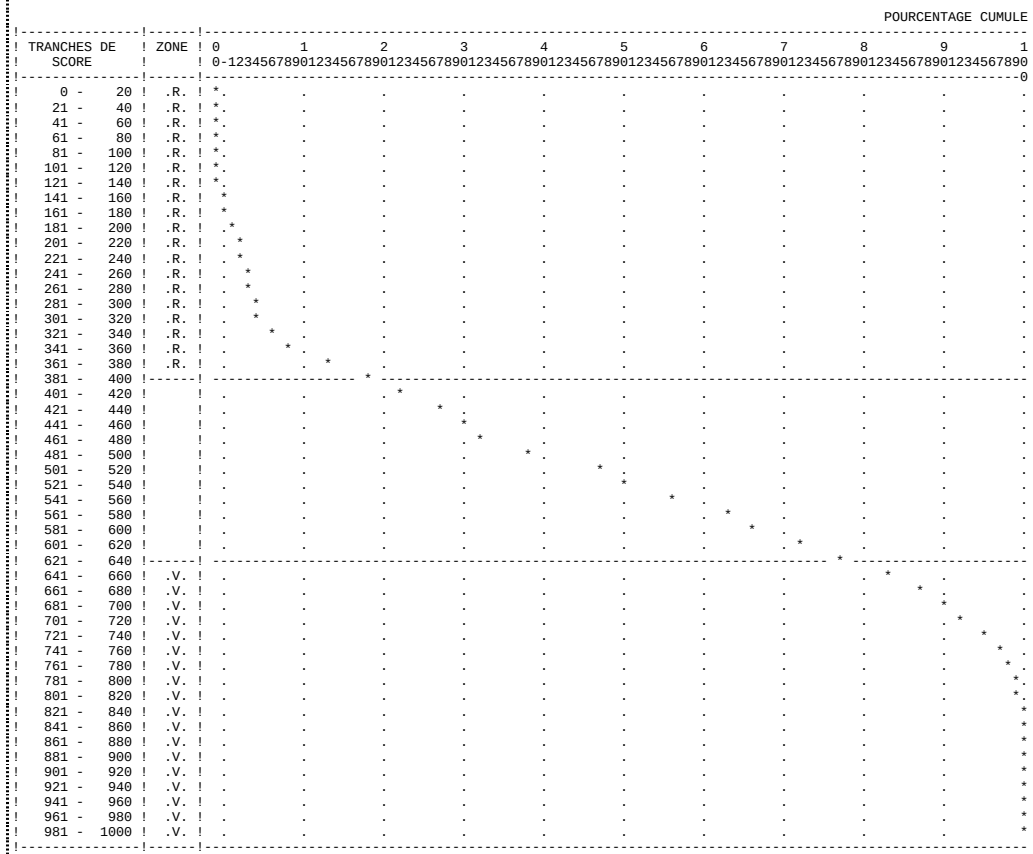
EXTRAIT 28

EVALUATION DANS LA POPULATION THEORIQUE

TRANCHES DES SCORES	ZONE	POPULATION % CUMULE	NB DE " FORTS " POUR 1 "FAIBLE"	VALEUR LISSEE	% DE FAIBLES PAR TRANCHES
0 - 20	.R.	.00	0 /1	0 /1	.00
21 - 40	.R.	.00	0 /1	0 /1	.00
41 - 60	.R.	.00	0 /1	0 /1	.00
61 - 80	.R.	.00	0 /1	0 /1	.00
81 - 100	.R.	.00	0 /1	0 /1	.00
101 - 120	.R.	.00	0 /1	0 /1	.00
121 - 140	.R.	.00	0 /1	0 /1	.00
141 - 160	.R.	.30	0 /1	0 /1	100.00
161 - 180	.R.	.44	0 /1	0 /1	100.00
181 - 200	.R.	1.07	0 /1	1 /1	.00
201 - 220	.R.	1.83	0 /1	1 /1	74.10
221 - 240	.R.	2.22	1 /1	1 /1	49.86
241 - 260	.R.	2.72	0 /1	1 /1	100.00
261 - 280	.R.	2.72	0 /1	1 /1	.00
281 - 300	.R.	3.77	24 /1	2 /1	3.94
301 - 320	.R.	4.34	1 /1	2 /1	38.78
321 - 340	.R.	6.10	1 /1	3 /1	36.35
341 - 360	.R.	7.95	2 /1	3 /1	31.58
361 - 380	.R.	12.54	2 /1	3 /1	30.57
381 - 400		17.92	6 /1	3 /1	13.25
401 - 420		21.69	9 /1	4 /1	9.68
421 - 440		26.59	7 /1	4 /1	11.47
441 - 460		29.87	2 /1	4 /1	31.92
461 - 480		31.54	1 /1	5 /1	38.82
481 - 500		37.63	6 /1	6 /1	13.44
501 - 520		47.49	9 /1	5 /1	9.71
521 - 540		49.99	1 /1	5 /1	34.61
541 - 560		55.62	9 /1	6 /1	9.47
561 - 580		62.51	21 /1	7 /1	4.52
581 - 600		66.34	4 /1	7 /1	18.02
601 - 620		72.03	4 /1	7 /1	19.25
621 - 640		77.32	9 /1	7 /1	10.00
641 - 660	.V.	82.90	18 /1	9 /1	5.05
661 - 680	.V.	87.34	9 /1	9 /1	9.94
681 - 700	.V.	89.97	10 /1	9 /1	8.98
701 - 720	.V.	92.38	9 /1	10 /1	9.73
721 - 740	.V.	94.67	46 /1	14 /1	2.12
741 - 760	.V.	96.71	0 /1	16 /1	.00
761 - 780	.V.	97.75	0 /1	16 /1	.00
781 - 800	.V.	99.50	45 /1	21 /1	2.17
801 - 820	.V.	99.50	0 /1	30 /1	.00
821 - 840	.V.	100.00	0 /1	86 /1	.00
841 - 860	.V.	100.00	0 /1	139 /1	.00
861 - 880	.V.	100.00	0 /1	85 /1	.00
881 - 900	.V.	100.00	0 /1	58 /1	.00
901 - 920	.V.	100.00	0 /1	0 /1	.00
921 - 940	.V.	100.00	0 /1	0 /1	.00
941 - 960	.V.	100.00	0 /1	0 /1	.00
961 - 980	.V.	100.00	0 /1	0 /1	.00
981 - 1000	.V.	100.00	0 /1	0 /1	.00

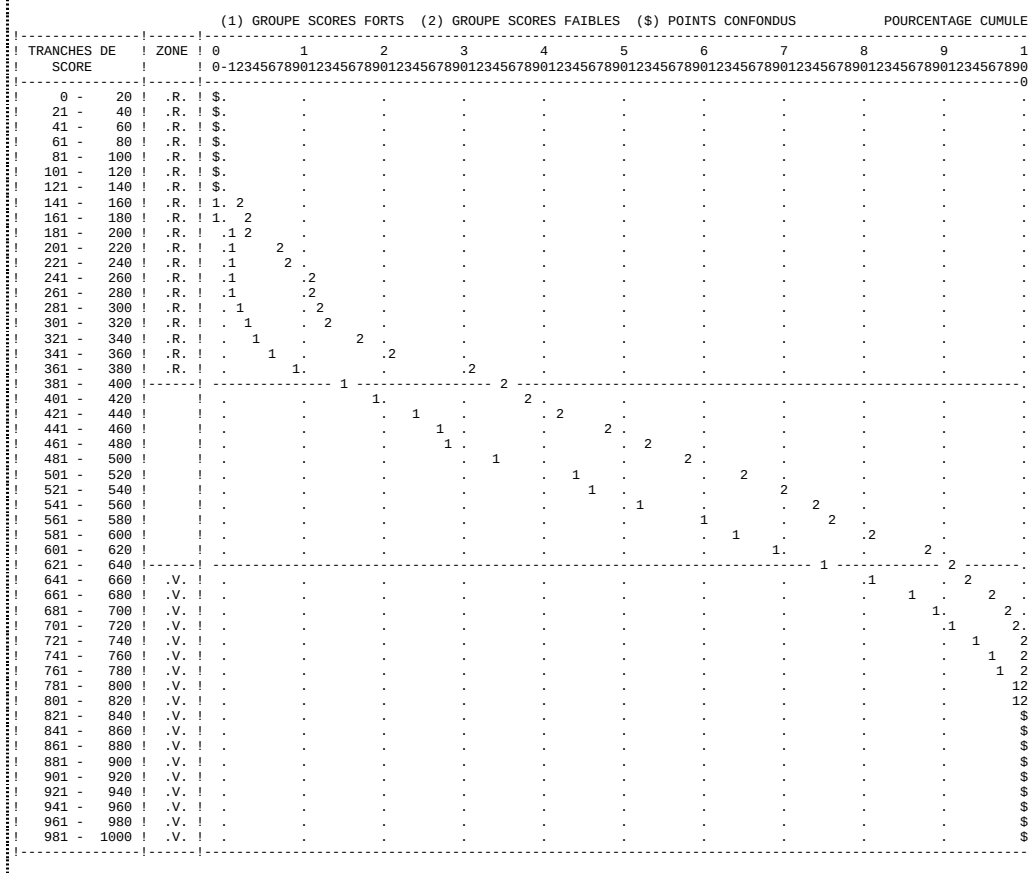
EXTRAIT 29

COURBE DE LA POPULATION EN FONCTION DU SCORE



EXTRAIT 30

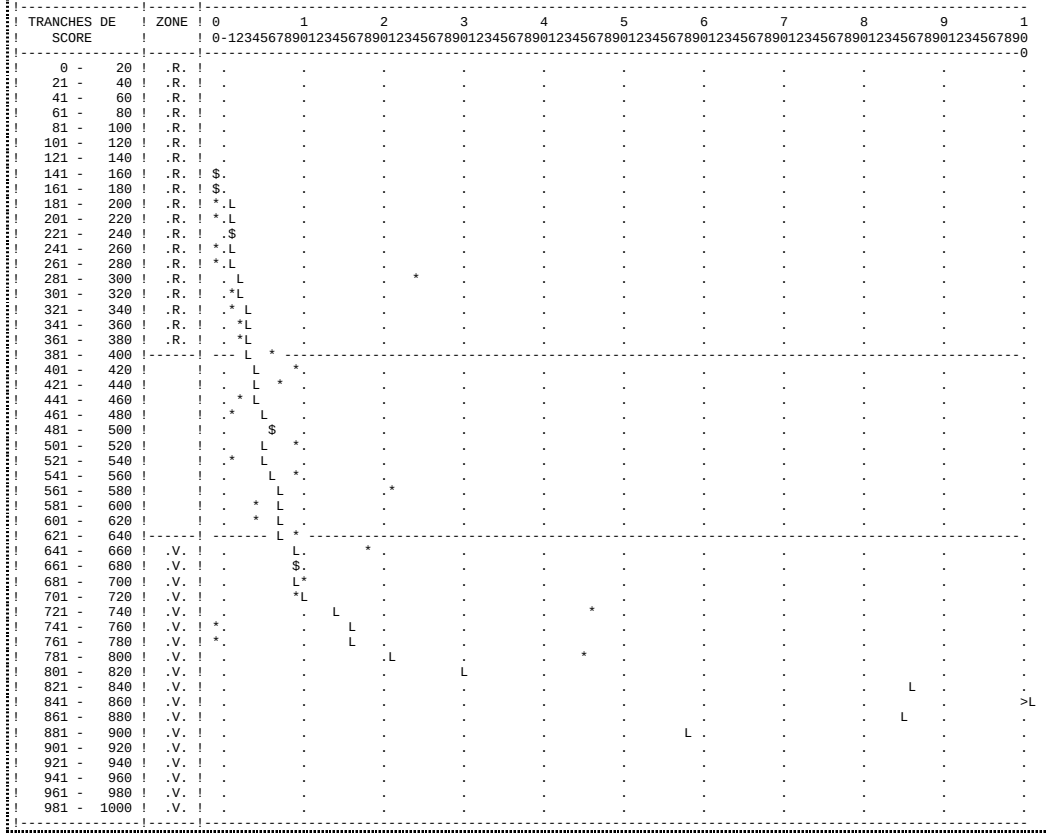
DISTRIBUTION DE L'ECHANTILLON EN FONCTION DU SCORE



EXTRAIT 31

COURBE D'EFFICACITE DANS LA POPULATION EN FONCTION DU SCORE

(*) AVANT LISSAGE (L) APRES LISSAGE ($) CONFONDUS (>) HORS CADRE EFFICACITE = NB DE " FORTS " POUR 1 " FAIBLE "



Le modèle log-linéaire

1. Généralités

Les modèles log-linéaires permettent d'analyser des tableaux de contingence multidimensionnelles (tableaux à n entrées), d'étudier plus précisément les liaisons entre les variables qualitatives qui constituent ces tableaux. On peut s'intéresser, par exemple, à la dépendance entre les variables dans leur ensemble ou à l'indépendance entre certaines variables conditionnellement à une ou plusieurs autres. Les modèles log-linéaires présentent les avantages de souplesse et d'interprétabilité propres à l'analyse de variance et à la régression.

Une hypothèse (sur certaines liaisons entre variables) est émise. Elle est traduite par l'écriture du modèle correspondant. Les paramètres de ce modèle sont alors estimés en utilisant, par exemple, la méthode du maximum de vraisemblance. L'adéquation du modèle est ensuite testée en vue d'accepter ou de rejeter cette hypothèse. Le problème essentiel est le nombre de paramètres inconnus du modèle qu'on est conduit à estimer.

Dans le modèle multinominal complet le nombre de paramètres à estimer devient vite très important quand le nombre de variables augmente. Par exemple, p variables binaires incluses dans le modèle entraînent l'estimation de 2^p-1 paramètres. En pratique une estimation efficace de ces paramètres nécessite un échantillon de taille énorme, en général non disponible. On dispose à l'autre extrême du modèle d'indépendance conditionnelle.

Les modèles log-linéaires résultent d'une reparamétrisation du modèle multinominal complet en vue de diminuer le nombre de paramètres à estimer.

1.1 Variables qualitatives, données individuelles et tableaux de°contingence

Les variables qualitatives expriment l'appartenance à une catégorie ou à une modalité.

Certaines sont purement nominales: par exemple la catégorie socio-professionnelle d'un actif (ouvrier, cadre, ...). D'autres sont ordinales lorsque l'ensemble des catégories est muni d'un ordre (par exemple: très résistant, assez résistant, peu résistant).

On utilise ces variables à partir d'un tableau de données individuelles ou à partir d'un tableau de contingence.

Dans un tableau de données individuelles, les lignes désignent les individus et les colonnes les variables. On inscrit dans les cellules les modalités des variables (tableau de codage « réduit »). Par exemple, si on dispose de dix individus et des deux variables qualitatives X et Y suivantes :

X variable à deux modalités désignant la catégorie socio professionnelle-C.S.P. (i représente la modalité de X):

i = 1 ouvrier
i = 2 cadre

Y variable à quatre modalités désignant la situation familiale S.F. (j représente la modalité de Y):

j = 1 célibataire
j = 2 marié
j = 3 divorcé
j = 4 veuf

Le tableau aura la forme suivante :

TABLEAU 1.1
Tableau « individus - variables » codage réduit

N° Ind.	X C.S.P.	Y S.F.
1	1	3
2	1	1
3	2	1
4	1	2
5	2	4
6	2	3
7	1	1
8	1	4
9	2	4
10	1	4

Il est facile d'obtenir, à partir de ce tableau, le tableau de contingence associé qui fournit la ventilation des dix individus selon les deux variables qualitatives.

Les lignes du tableau de contingence désignent les modalités de la variable X et les colonnes désignent les modalités de la variable Y. Le tableau comporte les effectifs n_{ij} c'est-à-dire le nombre d'individus appartenant simultanément aux modalités i et j des deux variables.

On obtient le tableau de contingence suivant :

TABLEAU 1.2
Tableau de contingence

X \ Y	**j = 1**	**j = 2**	**j = 3**	**j = 4**	$n_{i.}$
i = 1	2	1	1	2	*6*
i = 2	1	0	1	2	*4*
n_{j}	*3*	*1*	*2*	*4*	*n = 10*

On note :

$$n_{i.} = \sum_j n_{ij}$$

$$n_{.j} = \sum_i n_{ij}$$

$$n = \sum_i \sum_j n_{ij}$$

Les $n_{i.}$ et les $n_{.j}$ représentent respectivement les marges en ligne et en colonne. Par exemple $n_{1.}$ désigne le nombre d'ouvriers et $n_{.1}$ le nombre de célibataires parmi les dix individus. On désigne le nombre total d'individus par n.

1.2 Le modèle log-linéaire

On dispose d'un tableau de contingence et on veut savoir quelles sont les liaisons entre les variables. Dans ce cas de figure (tableau à deux entrées), le problème est bien connu. On se pose la question de savoir si les deux variables sont indépendantes. Et pour cela on opère un test du KHI-2. Le problème est plus délicat lorsque la dimension du tableau est supérieure à deux. On peut en effet se demander si les variables sont indépendantes dans leur ensemble ou si certaines sont indépendantes conditionnellement à une ou plusieurs autres. Autrement dit quelle est la structure du tableau? C'est l'objet du modèle log-linéaire de tenter de répondre à cette question.

Un modèle est l'écriture d'une formule permettant de reconstituer au moins approximativement le tableau de contingence à partir de l'information contenue dans les données proprement dites, marges et effectifs dans les cellules.

Lorsque l'on utilise le modèle log-linéaire pour étudier un tableau de contingence T, on se donne un ensemble d'hypothèses quant aux liaisons entre les variables du tableau. Dans l'écriture du modèle se trouvent inscrites ces hypothèses liant les données. Tester si le modèle est adéquat revient alors à tester si les hypothèses sont acceptables. Sous ces hypothèses et à partir des fréquences observées dans le tableau T, on construit un tableau estimé $\hat{T}$.

$\hat{T}$ étant calculé, il est possible de le confronter au tableau observé T. Si $\hat{T}$ est suffisamment proche de T, dans un sens qui sera à préciser, on ne rejettera pas les hypothèses qui ont conduit à construire $\hat{T}$.

Le modèle est « acceptable » si les fréquences calculées par le modèle sont « proches » des fréquences observées.

Si le modèle est « acceptable », on peut procéder à l'estimation des paramètres de ce modèle.

Les objectifs poursuivis avec les modèles log-linéaires sont différents de ceux poursuivis en régression.

En effet, un des objectifs de la régression est la prédiction de la valeur de la variable dépendante Y à partir des variables exogènes X_q. Cette prédiction s'effectue grâce à l'équation du modèle linéaire :

$$Y = \beta_0 + \sum_q \beta_q X_q + \varepsilon$$

Dans ce cas le premier intérêt est la liaison entre Y et les variables X_q et non pas les liaisons entre toutes les variables X_q.

Dans un modèle log-linéaire, la variable dépendante n'est pas une variable au sens usuel. Les quantités à prévoir peuvent être vues comme les probabilités d'appartenance à chaque cellule du tableau. Ainsi, la « variable dépendante » du modèle est en réalité l'ensemble de ces probabilités.

2. Les modèles log-linéaires

2.1 Cas de deux variables nominales

2.1.1 Introduction

Considérons le cas simple de deux variables qualitatives X et Y:
X est une variable à trois modalités représentant trois opinions politiques majeures (i désigne la modalité de X) :

i = 1 Gauche
i = 2 Centre
i = 3 Droite

Y est une variable à trois modalités représentant trois partis politiques (j désigne la modalité de Y):

j = 1 Parti démocrate
j = 2 Parti indépendant
j = 3 Parti républicain

Nous disposons d'un échantillon de n individus pour lesquels nous connaissons le parti et l'opinion politique. On construit le tableau de contingence avec en ligne les modalités i de la variable X et en colonne les modalités j de la variable Y.

TABLEAU 2.1
Croisement opinion × parti
Fréquences observées n_{ij}

Y / X	**Dém. j=1**	**Ind. j=2**	**Rép. j=3**	$\mathbf{n_{i.}}$
Gauche i=1	n_{11}	n_{12}	n_{13}	$\mathbf{n_{1.}}$
Centre i=2	n_{21}	n_{22}	n_{23}	$\mathbf{n_{2.}}$
Droite i=3	n_{31}	n_{32}	n_{33}	$\mathbf{n_{3.}}$
$\mathbf{n_{.j}}$	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	$\mathbf{n_{.3}}$	**n**

On pose que la fréquence observée n_{ij} dans chaque cellule du tableau est la réalisation d'une variable aléatoire N_{ij}. A chaque cellule du tableau correspond alors une variable aléatoire. On s'intéresse à l'espérance mathématique de ce terme aléatoire: c'est la « fréquence espérée », notée $f_{ij} = E(N_{ij})$

Un modèle log-linéaire exprime les relations qui sont supposées exister entre les fréquences espérées (qui sont par ailleurs des valeurs inconnues). Les relations sont posées a priori par le statisticien qui cherchera à voir ensuite si les données observées sont compatibles ou non avec le modèle choisi.

2.1.2 Les modèles log-linéaires à deux variables

Rappelons que tous les modèles log-linéaires concernent les fréquences espérées f_{ij}, c'est-à-dire des inconnues qui devront être estimées à partir des fréquences observées n_{ij} par une technique statistique appropriée.

Avant d'aborder les problèmes d'estimation des f_{ij}, familiarisons-nous avec les modèles log-linéaires en décrivant les plus simples d'entre eux et en étudiant les propriétés contenues dans leur écriture même.
Pour rendre plus concret certains développements, nous remplacerons ici les fréquences espérées inconnues par des valeurs numériques. Nous verrons clairement les propriétés et les relations entre ces termes inconnus en lisant les valeurs numériques (égalités, proportionnalité, somme nulle, etc...).

Modèle sans aucun effet des variables

Ce modèle est le plus simple que l'on puisse proposer. Il stipule que toutes les cellules (i,j) de la table ont même fréquence espérée. Ainsi, le modèle s'écrit, pour chaque cellule : $f_{ij} = \lambda$

ou en passant aux logarithmes : $\mathrm{Log} f_{ij} = \mu$

Le paramètre μ apparaît comme un effet global. Il est indépendant des indices i et j. Sa valeur, étant constante, est aussi obtenue en prenant la moyenne sur i et j des deux membres de l'égalité (II.1) : $\mu = \mathrm{Moy}_{ij}\{\mathrm{Log} f_{ij}\}$

Le paramètre μ représente ainsi l'apport moyen des modalités i et j des variables X et Y.
Soit un tableau de données d'une population de 972 individus pour lequel ce modèle serait « parfait ». Ceci signifie que les fréquences f_{ij} vérifient exactement l'écriture du modèle. (Tableau 2.2).

TABLEAU 2.2
Modèle « aucun effet des deux variables »
Fréquences espérées f_{ij}

Y / X	**Dém. j=1**	**Ind. j=2**	**Rép. j=3**	$f_{i.}$
Gauche i=1	108	108	108	***324***
Centre i=2	108	108	108	***324***
Droite i=3	108	108	108	***324***
$f_{.j}$	***324***	***324***	***324***	***972***

Dans cet exemple, il est facile de voir que pour tout i et tout j :

$$f_{ij} = 108$$

soit

$$\mathrm{Log}f_{ij} = 4{,}682$$

Cette relation est donc l'écriture du modèle log-linéaire sans aucun effet des deux variables : toutes les fréquences espérées dans les cellules sont égales.

Modèle avec effet de la variable ligne

Avec ce modèle, les fréquences espérées pour chaque modalité de la variable ligne X ne sont pas égales. L'effet de la variable ligne se manifeste par le fait que les effectifs marginaux dans les lignes n'étant plus égaux, les effectifs dans les cellules en dépendent d'une certaine façon.

D'une façon plus précise, le modèle correspondant s'écrit :

$$\mathrm{Log}f_{ij} = \mu + \mu_i(X)$$

avec μ *l'effet global et* $\mu_i(X)$ *l'effet spécifique de la modalité i de la variable X.*

L'effet provenant de la variable ligne X se traduit par l'ensemble des trois paramètres $\mu_i(X)$. Nous disposons d'un paramètre pour chaque modalité de la variable X.

D'après (II.3) : $\mu_i(X) = \mathrm{Log}f_{ij} - \mu$

Le paramètre $\mu_1(X)$ est indépendant de l'indice j. Sa valeur est obtenue en prenant la moyenne sur j des deux membres de l'égalité ci-dessus:

$$\mu_i(X) = \mathrm{Moy}_j\{\mathrm{Log}f_{ij}\} - \mu$$

Le paramètre $\mu_i(X)$ représente ainsi l'apport moyen de la modalité i de la variable X moins le paramètre μ.

Considérons les données du tableau II.3 pour lesquelles le modèle est « parfait ». Les valeurs des paramètres sont données par les formules suivantes :

$$\mu = \frac{\mathrm{Log}101 + \mathrm{Log}68 + \mathrm{Log}155}{3}$$
$$\mu_1(X) = \mathrm{Log}101 - \mu$$
$$\mu_2(X) = \mathrm{Log}68 - \mu$$
$$\mu_3(X) = \mathrm{Log}155 - \mu$$

Ce qui donne :

$$\mu = 4.63$$
$$\mu_1(X) = -0.01$$
$$\mu_2(X) = -0.41$$
$$\mu_3(X) = 0.42$$

TABLEAU 2.3
Modèle avec effet ligne
Fréquences espérées fij

X \ Y	**Dém. j=1**	**Ind. j=2**	**Rép. j=3**	$f_{i.}$
Gauche i=1	101	101	101	***303***
Centre i=2	68	68	68	***204***
Droite i=3	155	155	155	***465***
$f_{.j}$	***324***	***324***	***324***	***972***

On peut vérifier (aux erreurs d'arrondis près) par exemple la valeur de f_{11} en calculant l'expression suivante :

$$f_{11} = \exp(\mu + \mu_1(X))$$

On vérifie aussi l'égalité suivante, qui exprime une contrainte sur les coefficients :

$$\mu_1(X) + \mu_2(X) + \mu_3(X) = 0$$

D'une façon générale, on vérifiera sur tous les exemples que chaque fois qu'un terme apparaît dans un modèle (ici le facteur ligne X), des contraintes sur les coefficients correspondants sont satisfaites (ici la somme des coefficients $\mu_i(X)$ est nulle).

Modèle avec effet ligne et effet colonne

Ce modèle s'écrit de la manière suivante :

$$\mathrm{Log} f_{ij} = \mu + \mu_i(X) + \mu_j(Y)$$

avec μ *l'effet global,* $\mu_i(X)$ *l'effet de la ligne i,* $\mu_j(Y)$ *l'effet de la colonne j.*

L'expression de $\mu_j(Y)$ se déduit de la formule (II.4):

$$\mu_j(Y) = \mathrm{Moy}_i\{\mathrm{Log} f_{ij}\} - \mu$$

Le paramètre $\mu_j(Y)$ représente l'apport moyen de la modalité j moins le paramètre μ.

Considérons les données du tableau 2.4 pour lesquelles le modèle est « parfait ».

Tableau 2.4
Modèle avec effet ligne et effet colonne.
Fréquences espérées f_{ij}

Y / X	**Dém. j=1**	**Ind. j=2**	**Rép. j=3**	$f_{i.}$
Gauche i=1	106	130	251	***487***
Centre i=2	28	34	66	***128***
Droite i=3	77	95	185	***357***
$f_{.j}$	***211***	***260***	***501***	***972***

Ainsi

$$\mu = \frac{\text{Log}106 + \text{Log}130 + \text{Log}251 + \text{Log}28 + \text{Log}34}{9} + \frac{\text{Log}66 + \text{Log}77 + \text{Log}95 + \text{Log}185}{9}$$

$$\mu_1(\text{X}) = \frac{\text{Log } 106 + \text{Log } 130 + \text{Log } 251}{3} - \mu$$

$$\mu_2(\text{X}) = \frac{\text{Log } 28 + \text{Log } 34 + \text{Log } 166}{3} - \mu$$

$$\mu_3(\text{X}) = \frac{\text{Log } 77 + \text{Log } 95 + \text{Log } 185}{3} - \mu$$

Ce qui donne les valeurs suivantes (si les valeurs espérées étaient connues)

Effet global	$\mu = 4.47$		
Effet ligne	$\mu_1(\text{X}) = 0.55$	$\mu_2(\text{X}) = -0.79$	$\mu_3(\text{X}) = 0.24$
Effet colonne	$\mu_1(\text{Y}) = -0.36$	$\mu_2(\text{Y}) = -0.15$	$\mu_3(\text{Y}) = 0.51$

On peut vérifier par exemple la valeur de f_{11} (aux erreurs d'arrondis près) en calculant :

$$f_{11} = \exp(\mu + \mu_1(\text{X}) + \mu_1(\text{Y}))$$

On vérifie les deux contraintes sur les coefficients des facteurs entrés dans le modèle :

$$\mu_1(\text{X}) + \mu_2(\text{X}) + \mu_3(\text{X}) = 0$$
$$\mu_1(\text{Y}) + \mu_2(\text{Y}) + \mu_3(\text{Y}) = 0$$

On note que ce modèle exprime l'indépendance entre la variable ligne X et la variable colonne Y. On peut vérifier sur le tableau 2.4 la relation:

$$f_{ij} = \frac{f_{i.} f_{.j}}{f_{..}} \quad \text{pour tout i et tout j}$$

Le modèle saturé

Il s'agit d'un modèle qui traduit l'existence d'une liaison entre les variables X et Y. On parlera d'interaction entre X et Y.

$$\mathrm{Log}\,f_{ij} = \mu + \mu_i(X) + \mu_j(Y) + \mu_{ij}(XY)$$

La quantité $\mathrm{Log}\,f_{ij}$ s'exprime sous la forme d'une somme de différents effets :

- effet global (μ),
- effet principal dû à la variable X ($\mu_i(X)$),
- effet principal dû à la variable Y ($\mu_j(Y)$)
- effet dû à l'interaction entre les deux variables X et Y ($\mu_{ij}(XY)$).

Les paramètres $\mu_{ij}(XY)$ s'obtiennent par la formule du modèle :

$$\mu_{ij}(XY) = \mathrm{Log}\,f_{ij} - \mu - \mu_i(X) - \mu_j(Y)$$

En remplaçant μ, $\mu_i(X)$ et $\mu_j(Y)$ par leurs formules respectives et

$$\begin{aligned}\mu_{ij}(XY) = {} & \mathrm{Log}\,f_{ij} \\ & - \mathrm{Moy}_j\{\mathrm{Log}\,f_{ij}\} - \mathrm{Moy}_i\{\mathrm{Log}\,f_{ij}\} \\ & + \mathrm{Moy}_{ij}\{\mathrm{Log}\,f_{ij}\}\end{aligned}$$

Soit le tableau de données suivant pour lequel les fréquences f_{ij} (supposées connues) ne vérifient aucune relation particulière. (Noter bien qu'il s'agit de fréquences espérées et non d'un réel tableau d'observations).

TABLEAU 2.5
Modèle avec interaction
Fréquences espérées f_{ij}

Y / X	**Dém. j=1**	**Ind. j=2**	**Rép. j=3**	$f_{i.}$
Gauche i=1	90	2	60	***152***
Centre i=2	60	18	222	***300***
Droite i=3	50	80	390	***520***
$f_{.j}$	***200***	***100***	***672***	***972***

On peut appliquer les définitions des paramètres pour la cellule (i = 1 ; j = 1) du tableau 2.5.

$$\mu = \frac{\text{Log90} + \text{Log2} + \text{Log60} + \text{Log60} + \text{Log18} + \text{Log222} + \text{Log50} + \text{Log80} + \text{Log390}}{9}$$

$$\mu_1(\text{X}) = \frac{\text{Log90} + \text{Log2} + \text{Log60}}{3} - \mu$$

$$\mu_1(\text{Y}) = \frac{\text{Log90} + \text{Log60} + \text{Log50}}{3} - \mu$$

$$\mu_{11}(\text{XY}) = \text{Log90} - \frac{\text{Log90} + \text{Log60} + \text{Log50}}{3} - \frac{\text{Log90} + \text{Log2} + \text{Log60}}{3} + \mu$$

On obtient les valeurs suivantes

$$\mu = 3.993$$
$$\mu_1(\text{X}) = -0.897$$
$$\mu_1(\text{Y}) = 0.176$$
$$\mu_{11}(\text{XY}) = 1.228$$

L'ensemble des valeurs des coefficients est donné dans le tableau suivant :

TABLEAU 2.6
Valeurs des coefficients

μ = **3.99**	$\mu_1(\mathbf{Y}) =$ **0.18**	$\mu_2(\mathbf{Y}) =$ **-1.33**	$\mu_3(\mathbf{Y}) =$ **1.16**
$\mu_1(\mathbf{X}) =$ **–0.90**	$\mu_{11}(\text{XY}) = 1.23$	$\mu_{12}(\text{XY}) = -1.07$	$\mu_{13}(\text{XY}) = -0.17$
$\mu_2(\mathbf{X}) =$ **0.14**	$\mu_{21}(\text{XY}) = -0.21$	$\mu_{22}(\text{XY}) = 0.10$	$\mu_{23}(\text{XY}) = 0.11$
$\mu_3(\mathbf{X}) =$ **0.76**	$\mu_{31}(\text{XY}) = -1.02$	$\mu_{32}(\text{XY}) = 0.96$	$\mu_{33}(\text{XY}) = 0.05$

On vérifie la valeur de f_{11} (aux erreurs d'arrondis près) en calculant :

$$f_{11} = \exp(\mu + \mu_1(\text{X}) + \mu_1(\text{Y}) + \mu_{11}(\text{XY}))$$

On vérifie également les égalités suivantes :

$$\sum_i \mu_{ij}(\text{XY}) = 0$$
$$\sum_j \mu_{ij}(\text{XY}) = 0$$

Elles s'ajoutent aux relations sur les coefficients des effets principaux :

$$\sum_i \mu_i(\text{X}) = 0$$
$$\sum_j \mu_j(\text{Y}) = 0$$

Remarque
Le choix du modèle saturé ou du modèle non saturé est donc lié au choix d'une hypothèse sous-jacente sur les variables, en l'occurrence ici de l'hypothèse d'indépendance.

Notion d'indépendance et d'interaction

On applique deux types de vaccins (V1, V2) contre la grippe sur une population de mille individus. On observe ensuite si les individus ont attrapé la maladie ou non. On dispose alors d'un tableau de contingence à deux entrées. La variable colonne V indique le type de vaccin et la variable ligne M indique si l'individu est malade ou non.

Dans le cas du tableau 2.7, les variables V et M sont indépendantes. Les fréquences f_{ij} vérifient la formule d'indépendance :

$$f_{ij} = \frac{f_{i.} f_{.j}}{f_{..}}$$

Le tableau 2.8 donne le logarithme des fréquences espérées f_{ij}.

TABLEAU 2.7
Indépendance entre lignes et colonnes
Valeurs des f_{ij}

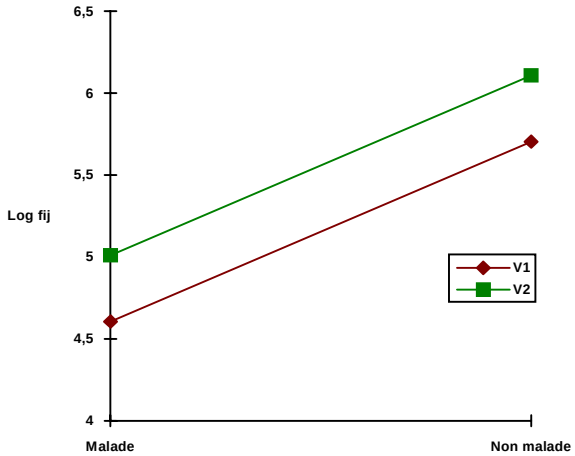
V / M	**Vaccin V1**	**Vaccin V2**	***TOTAL***
Malade	100	150	***250***
Non Malade	300	450	***750***
TOTAL	***400***	***600***	***1000***

TABLEAU 2.8
Indépendance sur les logarithmes
Valeurs des Log f_{ij}

V / M	**Vaccin V1**	**Vaccin V2**	***TOTAL***
Malade	4,60	5,01	***9,62***
Non Malade	5,70	6,11	***11,81***
TOTAL	***10,31***	***11,12***	***21,43***

La figure 2.1 indique en ordonnée, pour chaque vaccin utilisé, le logarithme du nombre d'individus malades (Malade en abscisse) et le logarithme du nombre d'individus non malades (Non Malade en abscisse).

FIGURE 2.1
Parallélisme et indépendance



Les lignes joignant les points ne servent qu'à mettre en évidence le sens des variations (il ne s'agit pas de courbes de variation). Le parallélisme des lignes sur les graphiques traduit le fait que les deux traitements ont le même effet sur la maladie (profils identiques en ligne).

Une technique communément utilisée dans l'analyse des tableaux de contingence est l'examen des « odds ratios ». Dans notre exemple concernant vaccin et maladie, le « odds » des malades ayant reçu le vaccin V_1 est n_{11}/n_{12}. Celui des non malades ayant reçu le vaccin V_1 est n_{21}/n_{22}. L'expression du « odds-ratio » est :

$$\frac{n_{11} / n_{12}}{n_{21} / n_{22}} = \frac{n_{11}\, n_{22}}{n_{12}\, n_{21}}$$

S'il y a indépendance entre lignes et colonnes du tableau de contingence alors, pour tout i et tout j :

$$n_{ij} = n_{i.} n_{.j}$$

et dans ce cas le odds ratio est égal à 1.

C'est le cas des données du tableau 2.7 qui vérifient l'équation :

$$\frac{n_{11} n_{22}}{n_{12} n_{21}} = 1 \qquad \text{soit} \quad \frac{n_{12}}{n_{11}} = \frac{n_{22}}{n_{21}}$$

En prenant le logarithme des deux membres, on démontre l'égalité des coefficients directeurs des deux droites:

$$\mathrm{Logn}_{12} - \mathrm{Logn}_{11} = \mathrm{Logn}_{22} - \mathrm{Logn}_{21}$$

Si l'on pense que l'effet de la modalité i de la variable X peut être différent selon la modalité j de la variable Y, on introduit dans le modèle un terme μ_{ij} qui exprime l'interaction.

$$\mathrm{Logf}_{ij} = \mu + \mu_i(\mathrm{X}) + \mu_j(\mathrm{Y}) + \mu_{ij}(\mathrm{XY})$$

Dans le cas de non indépendance, on n'observe plus le parallélisme des lignes sur le graphique (comme le montre la figure 2.2.).

Soit les données du tableau 2.9.
Les deux vaccins n'ont pas le même effet sur la maladie. Seulement 11,8% des individus vaccinés avec V1 sont malades contre 60% pour les individus vaccinés avec V2.
Si on prend le logarithme des fréquences espérées, on obtient le tableau 2.10

TABLEAU 2.9
Cas de non indépendance
Valeurs des f_{ij}

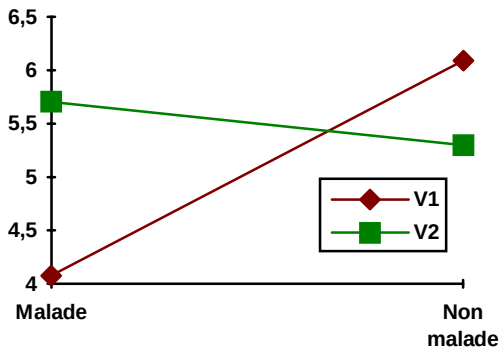
V / M	Vaccin V1	Vaccin V2	*TOTAL*
Malade	59	300	***359***
Non Malade	441	200	***641***
TOTAL	***500***	***500***	***1000***

TABLEAU 2.10
Cas de non indépendance
Valeurs des Log f_{ij}

V / M	Vaccin V1	Vaccin V2	*TOTAL*
Malade	4,08	5,70	***9,78***
Non Malade	6,09	5,20	***11,39***
TOTAL	***10,17***	***11,00***	***21,17***

La figure 2.2 indique en ordonnée, pour chaque vaccin utilisé, le logarithme du nombre d'individus malades (Malade en abscisse) et le logarithme du nombre d'individus non malades (Non Malade en abscisse). Le non parallélisme des deux trajectoires indique la présence d'une interaction entre maladie et vaccin et traduit le fait que les deux vaccins n'ont pas le même effet sur la maladie.

FIGURE 2.2
Traduction graphique d'une interaction



2.1.3 Remarques

Modèles hiérarchiques et écriture symbolique

Il est possible d'écrire le modèle contenant l'interaction entre les deux facteurs $\mu_{ij}(XY)$ sans les effets ligne et colonne $\mu_i(X)$ et $\mu_j(Y)$:

$$\mathrm{Log}\, f_{ij} = \mu + \mu_{ij}(XY)$$

Ce modèle est dit « non-hiérarchique ». Il est classique de ne prendre en compte que les modèles dits « hiérarchiques ». Ils sont plus simples à étudier, à estimer et à interpréter. Ils vérifient le principe suivant:

> Si dans le modèle, un terme μ contient dans son écriture un ensemble de lettres L représentant différentes variables (par exemple L = {XY}) alors le modèle doit contenir tous les termes μ contenant les sous ensembles de L (les sous ensembles {X} et {Y}).

Ainsi, si le modèle contient le terme $\mu_{ij}(XY)$ alors il doit contenir les termes $\mu_i(X)$ et $\mu_j(Y)$. Pour écrire ce modèle, il suffit donc de dire qu'il contient l'interaction [XY].

Ce principe permet d'utiliser une écriture « symbolique » du modèle.
Considérons le cas d'un modèle à trois variables X, Y et Z. Le modèle s'écrit sous la forme [Y] [XZ], s'il contient les termes $\mu_j(Y)$ et $\mu_{ik}(XZ)$. Parce qu'il est hiérarchique, il doit aussi contenir tous les termes μ contenant les sous ensembles de (Y) et de (XZ), c'est-à-dire les termes $\mu_i(X)$ et $\mu_k(Z)$. D'où l'écriture développée du modèle :

$$\mathrm{Log}\, f_{ijk} = \mu + \mu_i(X) + \mu_j(Y) + \mu_k(Z) + \mu_{ik}(XZ)$$

Les tableaux suivants présentent respectivement des modèles log-linéaires « hiérarchiques » possibles avec leur écriture symbolique et des modèles log-linéaires non « hiérarchiques » pour des tableaux de contingence à deux entrées.

TABLEAU 2.11
Exemples de modèles hiérarchiques et écriture symbolique

MODELES HIERARCHIQUES	ECRITURE SYMBOLIQUE
$\mathrm{Log}\, f_{ij} = \mu + \mu_i(X)$	[X]
$\mathrm{Log}\, f_{ij} = \mu + \mu_j(Y)$	[Y]
$\mathrm{Log}\, f_{ij} = \mu + \mu_i(X) + \mu_j(Y)$	[X] [Y]
$\mathrm{Log}\, f_{ij} = \mu + \mu_i(X) + \mu_j(Y) + \mu_{ij}(XY)$	[XY]

TABLEAU 2.12
Exemples de modèles non hiérarchiques

MODELES NON HIERARCHIQUES
$\mathrm{Log}f_{ij} = \mu + \mu_{ij}(XY)$
$\mathrm{Log}f_{ij} = \mu + \mu_{i}(X) + \mu_{ij}(XY)$
$\mathrm{Log}f_{ij} = \mu + \mu_{j}(Y) + \mu_{ij}(XY)$

Nombre de paramètres indépendants et modèle saturé

Les paramètres du modèle sont liés par des contraintes. Nous avons vu qu'ils vérifiaient les relations suivantes

$$\sum_{i=1}^{I} \mu_{i}(X) = 0$$

Sur les I paramètres, seulement (I-1) sont « indépendants »: il suffit d'en connaître (I-1), le dernier s'en déduit. De même

$$\sum_{j=1}^{J} \mu_{j}(Y) = 0$$

Sur les J paramètres, (J-1) sont indépendants. On a également les contraintes sur les paramètres de l'interaction:

$$\sum_{i=1}^{I} \mu_{ij}(XY) = 0$$

$$\sum_{j=1}^{J} \mu_{ij}(XY) = 0$$

Sur les IJ paramètres, (I-1)(J-1) sont indépendants.
Pour le modèle le plus complet :

$$\mathrm{Log}f_{ij} = \mu + \mu_{i}(X) + \mu_{j}(Y) + \mu_{ij}(XY)$$

le nombre de paramètres indépendants se calcule de la façon suivante :

Effet moyen :	1
Effet ligne :	I-1
Effet colonne :	J-1
Interactions :	(I-1)(J-1)
TOTAL:	1 + (I-1) + (J-1) + (I-1)(J-1) = IJ

Le tableau de contingence contient IJ effectifs et le modèle contient IJ paramètres indépendants (à estimer). Ce modèle est dit saturé: il contient autant de données que d'inconnues. Un tel modèle saturé s'adaptera toujours exactement aux données du tableau.

2.2 Indépendance

Soient G_1 et G_2 deux groupes de variables.
Les variables de G_1 et celles de G_2 sont indépendantes si et seulement si toutes les interactions comprenant au moins une variable de G_1 et une variable de G_2 sont nulles.

Prenons un exemple illustrant cette propriété. Soient G_1 = {X, Y} et G_2 = {Z}.
Si X et Y prises ensemble sont indépendantes de Z, le modèle s'écrit :

$$\mathrm{Log} f_{ijk} = \mu + \mu_i(X) + \mu_j(Y) + \mu_k(Z) + \mu_{ij}(XY)$$

On constate que les termes $\mu_{ik}(XZ)$, $\mu_{jk}(YZ)$ et $\mu_{ijk}(XYZ)$ sont nuls.

Indépendance mutuelle

Des variables sont mutuellement indépendantes si et seulement si toutes leurs interactions, de quelque ordre que ce soit, sont identiquement nulles.

Utilisons l'exemple précédent pour illustrer cette propriété.
Si les variables X, Y et Z sont mutuellement indépendantes, le modèle s'écrit :

$$\mathrm{Log}\ f_{ijk} = \mu + \mu_i(X) + \mu_j(Y) + \mu_k(Z)$$

Dans ce modèle, tous les termes d'interaction ont disparu.

Indépendance conditionnelle

Considérons trois groupes de variables G_1, G_2 et G_3.
Les variables de G_1 sont conditionnellement indépendantes de celles de G_2 connaissant les variables de G_3 si et seulement si les interactions comprenant au moins une variable de G_1 et une variable de G_2 sont nulles ainsi que les interactions comprenant au moins une variable de G_1, une variable de G_2 et une variable de G_3.

Soient par exemple G_1 = {X}, G_2 = {Y} et G_3 = {Z}

Si X et Y sont conditionnellement indépendants connaissant Z, le modèle s'écrit :

$$\mathrm{Log} f_{ijk} = \mu + \mu_i(X) + \mu_j(Y) + \mu_k(Z) + \mu_{ik}(XZ) + \mu_{jk}(YZ)$$

On constate que les termes $\mu_{ij}(XY)$ et $\mu_{ijk}(XYZ)$ sont nuls.

3. La procédure LOGLI

A partir d'un tableau qui peut être soit un tableau individus x variables, soit un tableau de contingence, la procédure **LOGLI de SPAD** permet d'ajuster un ou des modèles log-linéaires.

3.1 Cas d'un tableau individus-variables

Dans un tableau individus-variables, la ligne p concerne l'individu p et la colonne q concerne la modalité de la $q^{ième}$ variable.
Dans l'exemple suivant, on dispose de 30 individus et de quatre variables V_1, V_2, V_3 et V_4. Chaque variable possède deux modalités.

3.1.1 Premier exemple

Pour écrire un modèle où interviennent les fréquences indicées sur les trois premières variables, on sélectionne dans la procédure **SELEC** les trois variables V_1, V_2 et V_3.

On décide d'étudier un modèle où n'interviennent que les effets des variables V_1 et V_2, c'est-à-dire :

$$\mathrm{Log}f_{ijk} = \mu + \mu_i(V1) + \mu_j(V2)$$

Le modèle (3.1) est dépendant de la variable V_3 parce qu'elle a été sélectionnée même si celle-ci n'est pas présente dans l'écriture du modèle.

TABLEAU 3.1
Tableau individus x variables

Individus	V_1	V_2	V_3	V_4
1	1	1	1	1
2	1	1	1	2
3	1	1	2	1
4	1	2	1	1
5	1	2	1	1
6	1	1	2	2
7	1	2	2	1
8	2	2	1	1
9	2	1	1	2
10	2	1	2	1
11	1	2	1	2
12	1	2	2	2
13	2	2	2	1
14	2	1	2	2
15	2	2	1	2
16	2	2	2	2
17	1	1	2	1
18	1	1	1	1
19	2	2	1	2
20	1	2	2	1
21	1	2	2	1
22	2	2	1	2
23	1	2	2	1
24	1	1	1	1
25	2	2	1	2
26	1	1	1	1
27	1	2	1	2
28	1	1	2	2
29	2	2	2	2
30	2	2	2	2

Sous forme symbolique le modèle peut s'écrire

$$\mathrm{Log}(Y) = M + V1 + V2$$

La procédure **LOGLI** doit être précédée de la procédure **SELEC** qui assure la sélection des données utiles à l'analyse.

Dans l'exemple, la procédure **SELEC** a sélectionné les trois variables V_1, V_2 et V_3 et la totalité des individus de l'échantillon. Le bilan de la sélection est inscrit dans le fichier résultats **Etude-1.lst**. (Encart 3.2).

ENCART 3.1
Fichier paramètre Etude-1.pad

```
PROC SELEC
Sélection des données utiles
LSELI=TOT,IMASS=UNIF,LZERO    =    NOREC,LEDIT    =    COURT,    TIRER    =    NON
      NOMI ACT 1,2,3
      FIN

PROC LOGLI
Modèle Log-linéaire
LMOD=1,LEDIT   =   2,MAXIT   =   10,LTAB   =   1,MAX   =   0.001,   VADD   =   0.5
      V1+V2
      FIN

STOP              : Fin du fichier des commandes
```

ENCART 3.2
Résultats de la procédure SELEC

```
3 QUESTIONS  ACTIVES        6 MODALITES ASSOCIEES
------------------------------------------------
   1 . Variable N° 1       (   2 MODALITES )
   2 . Variable N° 2       (   2 MODALITES )
   3 . Variable N° 3       (   2 MODALITES )
------------------------------------------------
INDIVIDUS
--------------------- NOMBRE ---------- POIDS ---
 RETENUS    .........    NITOT    =    30               PITOT    =           30.000
 ACTIFS     .........    NIACT    =    30               PIACT    =           30.000
 SUPPLEMENTAIRES.. NISUP =  0    PISUP =     .000
------------------------------------------------
```

Les résultats de la procédure LOGLI indiquent tout d'abord le modèle défini par l'utilisateur. Pour l'exemple

```
      MODELE   1
      V1+V2 / V3
```

La notation V_1+V_2/V_3 signifie que l'ajustement porte sur le logarithme des effectifs du tableau de contingence croisant V_1, V_2 et V_3 à l'aide uniquement des variables V_1 et V_2 présentes dans le modèle.
Vient ensuite une édition de statistiques sommaires sur les variables du modèle, sur les variables V_1 et V_2 pour l'exemple. (Encart 3.3).

ENCART 3.3
Statistiques sur les variables du modèle

```
EDITION DE STATISTIQUES SOMMAIRES SUR LES VARIABLES DU MODELE
NOMBRE DE VARIABLES NOMINALES :  2
NOMBRE DE MODALITES ASSOCIEES :  4

TRI A PLAT DES VARIABLES DU MODELE :
---------------+-----------+------------------------------
    MODALITES  |           |
IDENT -LIBELLE |EFF. POIDS |HISTOGRAMME DES POIDS RELATIFS
---------------+-----------+------------------------------
1.Variable N° 1
V1-1-Mod. N°1  |18   18.00 | ******************
V1-2-Mod. N°2  |12   12.00 | *************
---------------+-----------+------------------------------
2.Variable N° 2
V2-1-Mod. N°1  |12   12.00 | *************
V2-2-Mod. N°2  |18   18.00 | ******************
---------------+-----------+------------------------------
```

Sont ensuite calculées les estimations $\hat{f}_{ijk}$ des fréquences espérées prédites par le modèle (3.1).

Si dans la procédure **LOGLI** le paramètre LEDIT = 2, les estimations sont affichées dans le tableau de contingence correspondant aux données en dessous de la fréquence observée n_{ijk}.

Le tableau à trois dimensions est édité de la façon suivante: on *fixe* la modalité de la *première* variable V_1. On construit le tableau de contingence à deux dimensions ayant pour *ligne* les modalités de la *deuxième* variable V_2 et pour *colonne* les modalités de la *troisième* variable V_3. On dispose alors d'autant de tableaux à deux dimensions que de modalités de la variable V_1.

Dans l'encart, les effectifs marqués par une * sont les effectifs estimés.

ENCART 3.4
Edition des tableaux de contingence avec les estimations des effectifs
Type d'édition LEDIT = 2

```
VARIABLE 1 : Variable N° 1
MODALITE  1 : V1-1-Modalité N°    1
      LIGNE: VARIABLE 2 : Variable N° 2
    COLONNE: VARIABLE 3 : Variable N° 3
                             1          2
                          V3-1       V3-2
   1  (observé)                          5.0                  4.0          9.0
Modalité N°    1 (*)       3.6        3.6    7.2
   2  (observé)                          4.0                  5.0          9.0
Modalité N°    2 (*)       5.4        5.4   10.8
                           9.0                            9.0             18.0
                           9.0        9.0   18.0
```

```
VARIABLE 1 : Variable N° 1
MODALITE  2 : V1-2-Modalité N°    2
      LIGNE: VARIABLE 2 : Variable N° 2
    COLONNE: VARIABLE 3 : Variable N° 3
                             1          2
                          V3-1       V3-2
   1  (observé)                          1.0                  2.0          3.0
Modalité N°    1 (*)       2.4        2.4    4.8
   2  (observé)                          5.0                  4.0          9.0
Modalité N°    2 (*)       3.6        3.6    7.2
                           6.0                            6.0             12.0
                           6.0        6.0   12.0
```

Si dans la procédure **LOGLI** le paramètre LEDIT = 4, les résultats sont rassemblés dans un seul tableau à trois colonnes.

La première colonne indique la fréquence observée n_{ijk}, la deuxième colonne la fréquence estimée $\hat{f}_{ijk}$ et la troisième colonne indique le triplet de modalités (ijk) considéré.

ENCART 3.5
Edition des estimations des effectifs - Type d'édition LEDIT = 4

```
 A=VARIABLE 1 : Variable N° 1
 B=VARIABLE 2 : Variable N° 2
 C=VARIABLE 3 : Variable N° 3
+-------------+-------------+--------------+
| observation |    ajuste   |    A   B   C |
+-------------+-------------+--------------+
|         5.0 |         3.6 |    1   1   1 |
|         1.0 |         2.4 |    2   1   1 |
|         4.0 |         5.4 |    1   2   1 |
|         5.0 |         3.6 |    2   2   1 |
|         4.0 |         3.6 |    1   1   2 |
|         2.0 |         2.4 |    2   1   2 |
|         5.0 |         5.4 |    1   2   2 |
|         4.0 |         3.6 |    2   2   2 |
+-------------+-------------+--------------+
```

Les estimations $\hat{f}_{ijk}$ sont obtenues par l'algorithme de DEMING-STEPHAN.

D'après le modèle (3.1) les estimations $\hat{f}_{ijk}$ doivent vérifier les relations suivantes. On note n l'effectif total de l'échantillon et n_{ijk} l'effectif observé dans la cellule (ijk).

$$\hat{f}_{...} = n$$

$$\hat{f}_{i..} = n_{i..}$$

$$\hat{f}_{.j.} = n_{.j.}$$

Ces relations sont à la base de l'algorithme encore appelé algorithme « d'ajustement itératif ». Elles sont vérifiées pour notre exemple dès la deuxième itération. (Encart 3.6).

ENCART 3.6
Convergence de l'algorithme

```
CONVERGENCE ..... NOMBRE D'ITERATIONS =   2
deviation maximale calculee sur les marges:
ITERATION     1           2
             3.00000      .00000
```

Sont indiquées ensuite les informations statistiques concernant le modèle proposé.

ENCART 3.7
Informations statistiques concernant le modèle

```
** INFORMATIONS STATISTIQUES **
ESTIMATION DU KHI-2 DE PEARSON           = 2.454
ESTIMATION DU RAPPORT DE VRAISEMBLANCE   = 2.605
PROBABILITE DU KHI-2 DE PEARSON          =  .783
PROBABILITE DU RAPPORT DE VRAISEMBLANCE  =  .761
DEGRE DE LIBERTE                         = 5
AIC BASE SUR L'ESTIMATION DE PEARSON     =-7.546
AIC BASE SUR LE RAPPORT DE VRAISEMBLANCE =-7.395
```

La probabilité du KHI-2 de Pearson égale à 0.783 et la probabilité du rapport de vraisemblance égale à 0.761. On ne peut pas rejeter le modèle proposé aux seuils usuels.
Pour estimer les paramètres, on utilise la méthode du maximum de vraisemblance. On trouve les résultats de l'encart 3.8.

Ce qui donne avec les notations utilisées ici:

$$\hat{\mu} = 1.2809$$
$$\hat{\mu}_1(V_1) = 0.2027$$
$$\hat{\mu}_2(V_1) = -0.2027$$
$$\hat{\mu}_1(V_2) = -0.2027$$
$$\hat{\mu}_2(V_2) = 0.2027$$

ENCART 3.8
Estimation des coefficients par la méthode du maximum de vraisemblance

```
COEFFICIENTS(METHODE          DU          MAXIMUM          DE          VRAISEMBLANCE)
CONSTANTE                      1.2809
VARIABLE    V1
V  1 = Variable N° 1
         Modalité N°    1         .2027
         Modalité N°    2        -.2027
VARIABLE    V2
V  2 = Variable N° 2
       Modalité N°     1        -.2027
       Modalité N°     2         .2027
```

Les relations suivantes sont vérifiées :

$$\sum_i \hat{\mu}_i(V_1) = 0$$
$$\sum_j \hat{\mu}_j(V_2) = 0$$

Un récapitulatif des principaux résultats de l'analyse est donné à la fin du fichier résultats. Ce récapitulatif est intéressant dans le cas où plusieurs modèles sont estimés simultanément. (Encart 3.9).

Remarque
Si on change l'ordre de sélection des variables dans SELEC ou si on change l'ordre d'écriture des variables dans LOGLI, les résultats sont inchangés.

ENCART 3.9
Récapitulatif des résultats

```
LISTE DES MODELES TRAITES
MODELE    1
V1+V2 / V3
NUMERO  DEGRE  ESTIMATION ESTIMATION
 DU     DE      CHI-2 DE  RAPPORT    AIC    AIC
MODELE  LIBERTE  PEARSON   VRAIS.  PEARSON.VRAIS.
                 (PROB.)                                                (PROB.)
 1       5      2.454     2.605   -7.546  -7.395
             (  .7835 )(  .7606 )
```

Noter l'importance de la sélection des variables NOMI ACT dans SELEC: cette commande définit les indices du modèle, même si toutes les variables ne sont pas nommées dans les modèles.

3.1.2 Deuxième exemple

On décide de ne sélectionner que les variables V_1 et V_2 afin que le modèle ne porte que sur deux indices (tableau de contingence simple). On inscrit dans la commande de SELEC « sélection des variables nominales des modèles » les variables suivantes: V_1 et V_2.

Le modèle devient : $\mathrm{Log}f_{ij} = \mu + \mu_i(V1) + \mu_j(V2)$

ENCART 3.10
Résultats de la procédure SELEC

```
2 QUESTIONS  ACTIVES           4 MODALITES ASSOCIEES
   1 . Variable N° 1           (   2 MODALITES )
   2 . Variable N° 2           (   2 MODALITES )

INDIVIDUS
--------------------------- NOMBRE ------ POIDS ---
 RETENUS ..... NITOT =  30    PITOT =  30.000
 ACTIFS ..... NIACT =  30    PIACT =  30.000
 SUPPLEMENTAIRES .... NISUP =   0    PISUP =    .000
```

Le modèle défini par l'utilisateur est rappelé:

```
      MODELE  1
      V1+V2
```

L'écriture V_1+V_2 signifie que l'ajustement porte sur le logarithme des effectifs du tableau de contingence croisant V_1 etV_2.

Si LEDIT = 2 dans la procédure **LOGLI**, on obtient les estimations des fréquences espérées $\hat{f}_{ij}$ dans le tableau de contingence à deux dimensions croisant les variables V_1 et V_2.

ENCART 3.11
Edition des tableaux de contingence avec les estimations des effectifs
Type d'édition LEDIT = 2

```
TABLEAU DE CONTINGENCE
LES EFFECTIFS MARQUES PAR (*) SONT LES EFFECTIFS ESTIMES

      LIGNE : VARIABLE 1 : Variable N° 1
    COLONNE : VARIABLE 2 : Variable N° 2

                              1          2
                           V2-1       V2-2
   1 (observe)              9.0        9.0       18.0
Modalité N°    1 (*)       7.2       10.8       18.0
   2 (observe)              3.0        9.0       12.0
Modalité N°    2 (*)       4.8        7.2       12.0
                           12.0       18.0       30.0
                           12.0       18.0       30.0
```

En *ligne* est indiquée la modalité de la *première* variable V_1 et en *colonne* la modalité de la *deuxième* variable V_2.

Si LEDIT = 4 dans la procédure **LOGLI,** les estimations des fréquences sont rassemblées dans le tableau suivant :

ENCART 3.12
Edition des estimations des effectifs
Type d'édition LEDIT = 4

```
TABLEAU DE CONTINGENCE
 A=VARIABLE 1 : Variable N° 1
 B=VARIABLE 2 : Variable N° 2
+-------------+-------------+----------+
| observation |    ajuste   |    A   B |
+-------------+-------------+----------+
|         9.0 |         7.2 |    1   1 |
|         3.0 |         4.8 |    2   1 |
|         9.0 |        10.8 |    1   2 |
|         9.0 |         7.2 |    2   2 |
+-------------+-------------+----------+
```

Les informations statistiques concernant le modèle proposé diffèrent du modèle précédent (qui faisait intervenir la variable V_3 dans le tableau de contingence).

ENCART 3.13
Informations statistiques concernant le modèle

```
** INFORMATIONS STATISTIQUES **
ESTIMATION DU CHI-2 DE PEARSON           =     1.875
ESTIMATION DU RAPPORT DE VRAISEMBLANCE   =     1.931
PROBABILITE DU CHI-2 DE PEARSON          =      .171
PROBABILITE DU RAPPORT DE VRAISEMBLANCE  =      .165
DEGRE DE LIBERTE                         =     1
AIC BASE SUR L'ESTIMATION DE PEARSON     =     -.125
AIC BASE SUR LE RAPPORT DE VRAISEMBLANCE =     -.069
```

La probabilité du rapport de vraisemblance est plus faible (0,165 contre 0,761) mais assez élevée pour ne pas rejeter le modèle proposé aux seuils usuels.

Rappel important
Lorsque l'on travaille à partir d'un tableau individus-variables (LTAB = 1), les variables mentionnées dans un modèle sont sélectionnées parmi l'ensemble des variables retenues dans l'étape SELEC qui précède. Même si certaines variables ne sont pas nommées dans le modèle, elles participent à la définition du tableau de contingence analysé et interviennent par conséquent dans les estimations et donc dans les résultats de l'analyse.

3.1.3 Troisième exemple

Cette partie concerne l'édition des tableaux de contingence affichant les estimations des fréquences espérées.

On considère n variables V_1, V_2,..., V_n dans l'ordre d'apparition dans le fichier. Le tableau de contingence à n dimensions sera décomposé en $m_1*m_2*...*m_{n-2}$ tableaux à deux dimensions avec m_i le nombre de modalités de la variableV_i.

Les tableaux à deux dimensions qui sont édités concernent en *ligne* les modalités de *l'avant dernière* variable $V_{(n-1)}$ et en *colonne* les modalités de la *dernière* variable V_n.

Par exemple on considère cinq variables V_1, V_2, V_3, V_4 et V_5 à deux modalités chacune.

Les $m_1*m_2*m_3$ = 8 tableaux de contingence (2*2) édités concernent en ligne les modalités de la variable V_4 et en colonne les modalités de la variable V_5.
Le schéma suivant affiche la modalité fixée des variables V1, V2 et V3 pour chaque tableau édité.

TABLEAU 3.2
Dans quel ordre sont édités les tableaux de contingence

	V1	**V2**	**V3**
Tableau N°1	1	1	1
Tableau N°2	2	1	1
Tableau N°3	1	2	1
Tableau N°4	2	2	1
Tableau N°5	1	1	2
Tableau N°6	2	1	2
Tableau N°7	1	2	2
Tableau N°8	2	2	2

La variable V_1 change de modalité à chaque tableau. La variable V_2 change de modalité tous les deux tableaux, ce qui correspond au nombre de modalités de la variable V_1.
La variable V_3 change de modalité tous les quatre tableaux, ce qui correspond au nombre de modalités de la variable V_1 multiplié par le nombre de modalités de la variable V_2.

A partir du tableau individus-variables (3.1), on sélectionne les quatre variables suivantes V_1, V_2, V_3 et V_4 et on étudie le modèle log-linéaire :

$$\mathrm{Log}(Y) = M + V1 + V2 + V3 + V4$$

Il s'écrit aussi :

$$\mathrm{Log}f_{ijkm} = \mu + \mu_i(V_1) + \mu_j(V_2) + \mu_k(V_3) + \mu_m(V_4)$$

On dispose de quatre tableaux édités de la façon suivante : en ligne est indiquée la modalité de l'avant dernière variable V_3 et en colonne est indiquée, la modalité de la dernière variable V_4.

ENCART 3.14
Edition des tableaux de contingence

```
VARIABLE 1 : Variable N° 1
MODALITE  1 : V1-1-Modalité N°    1
VARIABLE 2 : Variable N° 2
MODALITE  1 : V2-1-Modalité N°    1

      LIGNE: VARIABLE 3 : Variable N° 3
    COLONNE: VARIABLE 4 : Variable N° 4
                              1          2
                           V4-1       V4-2
   1 (observe)              4.0        1.0        5.0
Modalité N°    1 (*)       1.9        1.8        3.7
   2 (observe)              2.0        2.0        4.0
Modalité N°    2 (*)       1.8        1.8        3.6
                           6.0        3.0        9.0
                           3.7        3.6        7.4
```

```
VARIABLE 1 : Variable N° 1
MODALITE  2 : V1-2-Modalité N°    2
VARIABLE 2 : Variable N° 2
MODALITE  1 : V2-1-Modalité N°    1

      LIGNE: VARIABLE 3 : Variable N° 3
    COLONNE: VARIABLE 4 : Variable N° 4

                             1         2
                          V4-1      V4-2
   1 (observe)              .5       1.0        1.5
Modalité N°    1 (*)      1.3       1.3        2.6
   2 (observe)             1.0       1.0        2.0
Modalité N°    2 (*)      1.3       1.2        2.5
                          1.5       2.0        3.5
                          2.6       2.5        5.1
```

```
VARIABLE 1 : Variable N° 1
MODALITE  1 : V1-1-Modalité N°    1
VARIABLE 2 : Variable N° 2
MODALITE  2 : V2-2-Modalité N°    2

      LIGNE: VARIABLE 3 : Variable N° 3
    COLONNE: VARIABLE 4 : Variable N° 4

                             1         2
                          V4-1      V4-2

   1 (observe)             2.0       2.0        4.0
Modalité N°    1 (*)      2.7       2.7        5.4
   2 (observe)             4.0       1.0        5.0
Modalité N°    2 (*)      2.7       2.6        5.2
                          6.0       3.0        9.0
                          5.4       5.2       10.6
VARIABLE 1 : Variable N° 1
MODALITE  2 : V1-2-Modalité N°    2
VARIABLE 2 : Variable N° 2
MODALITE  2 : V2-2-Modalité N°    2

      LIGNE: VARIABLE 3 : Variable N° 3
    COLONNE: VARIABLE 4 : Variable N° 4

                             1         2
                          V4-1      V4-2
   1 (observe)             1.0       4.0        5.0
Modalité N°    1 (*)      1.9       1.8        3.7
   2 (observe)             1.0       3.0        4.0
Modalité N°    2 (*)      1.8       1.8        3.6
                          2.0       7.0        9.0
                          3.7       3.6        7.4
```

Remarque concernant les données manquantes

Si on totalise les marges des quatre tableaux édités on obtient la valeur 30,5

9+3,5+9+9 = 30,5

au lieu du nombre total d'individus, c'est-à-dire 30.

Ceci est dû à la présence d'une cellule nulle. Aucun individu ne possède simultanément la modalité 2 pour V_1, la modalité 1 pour V_2, la modalité 1 pour V_3 et la modalité 1 pour V_4 (i.e $n_{2111} = 0$). LOGLI ajoute à cette cellule la valeur de 0,5 pour pouvoir effectuer les calculs.

Cette valeur de substitution peut être modifiée avec le paramètre VADD dans la procédure LOGLI. VADD accepte toute valeur positive. Toute valeur négative éventuelle dans une table de contingence sera automatiquement remplacée par la valeur VADD.

3.2 Cas d'un tableau de contingence

Il est facile d'obtenir à partir du tableau VI.1 le tableau de contingence associé qui fournit la ventilation des trente individus selon les quatre variables qualitatives V_1, V_2, V_3 et V_4.

Plusieurs écritures sont possibles pour représenter ce tableau à quatre dimensions sur une feuille de papier.

3.2.1 Création du tableau de contingence

Par exemple, pour la modalité i de V_1 et la modalité j deV_2 fixées, on peut définir en *ligne* les modalités de la variable *V_3* et en *colonne* les modalités de la variable *V_4*.
On dispose alors des quatre tableaux suivants

TABLEAU 1

$V_1 = 1$ $V_2 = 1$	$V_4 = 1$	$V_4 = 2$
$V_3 = 1$	4	1
$V_3 = 2$	2	2

TABLEAU 2

$V_1 = 2$ $V_2 = 1$	$V_4 = 1$	$V_4 = 2$
$V_3 = 1$	0	1
$V_3 = 2$	1	1

TABLEAU 3

$V_1 = 1$ $V_2 = 2$	$V_4 = 1$	$V_4 = 2$
$V_3 = 1$	2	2
$V_3 = 2$	4	1

TABLEAU 4

$V_1 = 2$ $V_2 = 2$	$V_4 = 1$	$V_4 = 2$
$V_3 = 1$	1	4
$V_3 = 2$	1	3

On peut disposer ces quatre tableaux les uns à côtés des autres dans le sens de la longueur. On obtient le tableau 3.3. Il comporte une variable en ligne V_3 et trois variables en colonne (V_2, V_1, V_4).

TABLEAU 3.3
Alignement horizontal des quatre tableaux

	TABLEAU 1		TABLEAU 2		TABLEAU 3		TABLEAU 4	
	$V_2 = 1$	**$V_2 = 1$**	**$V_2 = 1$**	**$V_2 = 1$**	**$V_2 = 2$**	**$V_2 = 2$**	**$V_2 = 2$**	**$V_2 = 2$**
	$V_1 = 1$	**$V_1 = 1$**	**$V_1 = 2$**	**$V_1 = 2$**	**$V_1 = 1$**	**$V_1 = 1$**	**$V_1 = 2$**	**$V_1 = 2$**
	$V_4 = 1$	**$V_4 = 2$**	**$V_4 = 1$**	**$V_4 = 2$**	**$V_4 = 1$**	**$V_4 = 2$**	**$V_4 = 1$**	**$V_4 = 2$**
$V_3 = 1$	4	1	0	1	2	2	1	4
$V_3 = 2$	2	2	1	1	4	1	1	3

On peut encore considérer le tableau de contingence comportant deux variables en ligne (par exemple V_1 et V_2) et deux variables en colonne (par exemple V_3 et V_4). On obtient le tableau 3.4.

TABLEAU 3.4
Deux variables en ligne et deux variables en colonne

V_3 = 1 **V_4 = 1**	**V_3 = 1** **V_4 = 2**	**V_3 = 2** **V_4 = 1**	**V_3 = 2** **V_4 = 2**
4	1	2	2
0	1	1	1
2	2	4	1
1	4	1	3

On peut aussi les disposer les uns en dessous des autres. On obtient le tableau 3.5. Il comporte trois variables en ligne (V_1, V_2, V_3) et une variable en colonne V_4.

TABLEAU 3.5
Alignement vertical des quatre tableaux

	V_4 = 1	**V_4 = 2**
TABLEAU 1		
V_1 = 1 **V_2 = 1** **V_3 = 1**	4	1
V_1 = 1 **V_2 = 1** **V_3 = 2**	2	2
TABLEAU 2		
V_1 = 2 **V_2 = 1** **V_3 = 1**	0	1
V_1 = 2 **V_2 = 1** **V_3 = 2**	1	1
TABLEAU 3		
V_1 = 1 **V_2 = 2** **V_3 = 1**	2	2
V_1 = 1 **V_2 = 2** **V_3 = 2**	4	1
TABLEAU 4		
V_1 = 2 **V_2 = 2** **V_3 = 1**	1	4
V_1 = 2 **V_2 = 2** **V_3 = 2**	1	3

3.2.2 Comment introduire un tableau de contingence dans SPAD ?

Il est possible de saisir les tableaux 3.3, 3.4 et 3.5 dans SPAD afin d'effectuer une analyse log-linéaire.

La présentation des résultats dépend du tableau choisi.

Saisie du tableau 3.3

Pour le tableau 3.3 les huit combinaisons entre les modalités des variables V_2, V_1 et V_4 sont introduites comme étant huit variables continues (fréquences) et les deux modalités de la variable V_3 sont introduites comme étant deux individus (lignes du tableau individus x variables).

Dans le fichier de commande de la procédure LOGLI, on enregistre alors huit fréquences (colonnes) actives et deux lignes (individus) actives.

Avec LOGLI, les variables inscrites en colonne sont renommées en variable C_i et les variables inscrites en ligne sont renommées en variable L_j de la façon suivante.

Dans l'exemple on dispose de trois variables colonne V_2, V_1 et V_4. La modalité de la variable V_4 variant dans le tableau de contingence *toutes* les colonnes, la variable V_4 devient C_3. La modalité de la variable V_1 variant *toutes les deux* colonnes, V_1 devient C_2. La modalité de la variable V_2 variant *toutes les quatre* colonnes, V_2 devient C_1.

En résumé :

$V_4 := C_3$
$V_1 := C_2$
$V_2 := C_1$

Dans la liste des variables colonne C_i, on remarque que c'est la dernière variable (dans notre exemple C_3) qui varie avant toutes les autres. Ceci est également valable pour les variables ligne L_j.s'il y a plusieurs variables empilées en ligne).

Les lignes sont construites à l'aide uniquement de la variable V_3 qui devient la variable L_1. Ainsi $V_3 := L_1$

Saisie du tableau 3.5

Pour le tableau 3.5 il suffit de procéder à l'identification suivante:

$V_4 = C_1$
$V_3 = L_3$
$V_1 = L_2$
$V_2 = L_1$

Saisie du tableau 3.4

Le tableau 3.4 comporte deux variables ligne et deux variables colonne. La variable V_2 change de modalité toutes les lignes et la variable V_4 change de modalité toutes les colonnes.

Ainsi $V_2 = L_2$
$V_1 = L_1$
$V_4 = C_2$
$V_4 = C_1$

3.2.3 Résultats de la procédure LOGLI avec le tableau 3.3

Après avoir introduit le tableau 3.3, il faut choisir les variables intervenant dans l'étude du modèle.

On mentionne les variables colonne C_j sélectionnées pour l'étude dans le fichier paramètre de la procédure SELEC à la suite de l'intitulé

FREQ ACT

Par exemple si on désire sélectionner les trois variables colonne C_1, C_2 et C_3 on note

FREQ ACT 1--3

De même on mentionne les variables ligne L_i sélectionnées pour l'étude à la suite de l'intitulé ACT

Par exemple si on désire sélectionner les deux variables ligne L_1 et L_2, on note

ACT 1--2

Si l'on souhaite étudier le modèle suivant:

$$\mathrm{Log}f_{ijkm} = \mu + \mu_i(V_1) + \mu_j(V_2) + \mu_k(V_3) + \mu_m(V_4)$$

il suffit d'inscrire dans la commande « la définition du ou des modèles » le modèle correspondant en remplaçant V_1 par C_2, V_2 par C_1, V_3 par L_1 et V_4 par C_3.

Ce qui donne

$$\mathrm{Log}(Y) = M + L1 + C1 + C2 + C3$$

On obtient le fichier paramètre suivant:

ENCART 3.15
Fichier paramètre
(Tableau de contingence 3.3)

```
PROC SELEC
Sélection des données utiles
LSELI   =   LIST,   IMASS   =   UNIF,   LZERO   =   NOREC,   LEDIT   =   COURT
         FREQ ACT 1--8
         FIN
: Lignes actives
         ACT 1--2
         FIN
PROC LOGLI
Modèle Log-linéaire
LMOD = 1,  LEDIT = 2, MAXIT = 10, LTAB = 2,
LDICO = 0, MAX = 0.001, VADD = 0.5
MLIG = 2
MCOL = 2 2 2
         L1+C1+C2+C3
         FIN
STOP                    : Fin du fichier des commandes
```

Dans la procédure SELEC sont indiquées les fréquences actives et les lignes actives du tableau de contingence.

Dans la procédure LOGLI, le paramètre MLIG fournit la liste des nombres de modalités des variables en ligne et le paramètre MCOL fournit la liste des nombres de modalités des variables en colonne; les nombres sont séparés par un espace ou une virgule.

Exemple :
Si l'on dispose de trois variables ligne L_1, L_2 et L_3 et de deux variables colonne C_1 et C_2, les paramètres MLIG et MCOL affichent les nombres de modalités des variables dans l'ordre suivant:

MLIG = Nombre de modalités de L_1
Nombre de modalités de L_2
Nombre de modalités de L_3
MCOL = Nombre de modalités de C_1
Nombre de modalités de C_2.

La procédure SELEC sélectionne les trois variables colonne et la variable ligne.

ENCART 3.16
Résultats de la procédure SELEC
(Tableau de contingence 3.3)

```
8 FREQUENCES ACTIVES
   1 . Variable N° 1                    ( CONTINUE )
   2 . Variable N° 2                    ( CONTINUE )
   3 . Variable N° 3                    ( CONTINUE )
   4 . Variable N° 4                    ( CONTINUE )
   5 . Variable N° 5                    ( CONTINUE )
   6 . Variable N° 6                    ( CONTINUE )
   7 . Variable N° 7                    ( CONTINUE )
   8 . Variable N° 8                    ( CONTINUE )

POIDS DES INDIVIDUS: Poids des individus (somme des
                     fréquences actives).
------------------- NOMBRE ------------ POIDS ---
 RETENUS .... ..... NITOT =  2     PITOT =  30.000
 ACTIFS ...... .... NIACT =  2     PIACT =  30.000
 SUPPLEMENTAIRES .  NISUP =  0     PISUP =    .000
-------------------------------------------------
```

Si l'on ordonne les variables ligne et les variables colonne de la façon suivante: les variables ligne précédant les variables colonne

L1 C1 C2 C3

les tableaux de contingence sont édités de la même manière que dans le cas d'un tableau individus-variables.
En effet, les tableaux de contingence édités indiquent en ligne les modalités de l'avant dernière variable (dans l'exemple C2) et en colonne

les modalités de la dernière variable (dans l'exemple C3) pour une modalité de L1 et de C1 fixées.

La variable L1 change de modalité à chaque tableau et la variable C1 change de modalité tous les deux tableaux, ce qui correspond au nombre de modalités de la variable L1.

On obtient les éditions suivantes :

ENCART 3.17
Edition des tableaux de contingence
(Tableau de contingence 3.3)

```
VARIABLE : L1
MODALITE :  1
VARIABLE : C1
MODALITE :  1
      LIGNE: VARIABLE : C2
    COLONNE: VARIABLE : C3
                              1          2
   1 (observe)              4.0        1.0        5.0
     (estime)               1.9        1.8        3.7
   2 (observe)               .5        1.0        1.5
     (estime)               1.3        1.3        2.6
                            4.5        2.0        6.5
                            3.2        3.1        6.4
```

```
VARIABLE : L1
MODALITE :  2
VARIABLE : C1
MODALITE :  1
      LIGNE: VARIABLE : C2
    COLONNE: VARIABLE : C3
                              1          2
   1 (observe)              2.0        2.0        4.0
     (estime)               1.8        1.8        3.6
   2 (observe)              1.0        1.0        2.0
     (estime)               1.3        1.2        2.5
                            3.0        3.0        6.0
                            3.1        3.0        6.1

VARIABLE : L1
MODALITE :  1
VARIABLE : C1
MODALITE :  2
      LIGNE: VARIABLE : C2
    COLONNE: VARIABLE : C3
                              1          2
   1 (observe)              2.0        2.0        4.0
     (estime)               2.7        2.7        5.4
   2 (observe)              1.0        4.0        5.0
     (estime)               1.9        1.8        3.7
                            3.0        6.0        9.0
                            4.6        4.5        9.1

VARIABLE : L1
MODALITE :  2
VARIABLE : C1
MODALITE :  2
      LIGNE: VARIABLE : C2
    COLONNE: VARIABLE : C3 :
                              1          2
   1 (observe)              4.0        1.0        5.0
     (estime)               2.7        2.6        5.2
   2 (observe)              1.0        3.0        4.0
     (estime)               1.8        1.8        3.6
                            5.0        4.0        9.0
                            4.5        4.4        8.9
```

Pour obtenir des listages de résultats plus faciles à lire, l'utilisateur peut spécifier les libellés des variables en choisissant l'option LDICO = 1 dans la procédure LOGLI.
Les libellés des variables constituant le tableau de contingence sont insérés à la suite des commandes MLIG et MCOL et respectent l'ordre suivant:
L_1 C_1 C_2 C_3

Etant donné que :

$L_1 = V_3$
$C_1 = V_2$
$C_2 = V_1$
$C_3 = V_4$

le fichier paramètre s'écrit :

ENCART 3.18
Fichier paramètre avec spécification des libellés (Tableau de contingence 3.3)

```
    PROC SELEC
    Sélection des données utiles
    LSELI = LIST, IMASS = UNIF, LZERO = NOREC,
    LEDIT = COURT
             FREQ ACT 1--8
             FIN
    : Lignes actives
             ACT 1--2
             FIN
    PROC LOGLI
    Modèle Log-linéaire
    LMOD = 1,  LEDIT = 2, MAXIT = 10, LTAB = 2,
    LDICO = 1, MAX = 0.001,   VADD = 0.5
       MLIG = 2
       MCOL = 2 2 2
        2   Variable N°3
    V3-1   Modalité 1 de V3
    V3-2   Modalité 2 de V3
        2   Variable N°2
    V2-1   Modalité 1 de V2
    V2-2   Modalité 2 de V2
        2   Variable N°1
    V1-1   Modalité 1 de V1
    V1-2   Modalité 2 de V1
        2   Variable N°4
    V4-1   Modalité 1 de V4
    V4-2   Modalité 2 de V4

             L1+C1+C2+C3
             FIN
    STOP                    : Fin du fichier des commandes
```

Le chiffre présent avant chaque variable indique le nombre de modalités. Dans l'exemple, chaque variable possède deux modalités. Les tableaux de contingence deviennent plus faciles à lire ainsi que la valeur des paramètres:

ENCART 3.19
Un tableau de contingence avec libellés des variables - (Tableau de contingence 3.3)

```
VARIABLE L1 :   Variable N°3
MODALITE  1 : V3-1-  Modalité 1 de V3
VARIABLE C1 :   Variable N°2
MODALITE  1 : V2-1-  Modalité 1 de V2
      LIGNE: VARIABLE C2 :   Variable N°1
    COLONNE: VARIABLE C3 :   Variable N°4
                              1          2
                           V4-1       V4-2
   1 (observe)              4.0        1.0        5.0
Modalité 1 de V1 (*)        1.9        1.8        3.7
   2 (observe)               .5        1.0        1.5
  Modalité 2 de V1 (*)      1.3        1.3        2.6
                            4.5        2.0        6.5
                            3.2        3.1        6.4
```

ENCART 3.20
Edition des estimations des paramètres - (Tableau de contingence 3.3)

```
COEFFICIENTS (METHODE DU MAXIMUM DE VRAISEMBLANCE)

CONSTANTE                        .6118

VARIABLE    L1
L1 =   Variable N°3
  L1 .   Modalité 1 de V3        .0164
  L1 .   Modalité 2 de V3       -.0164

VARIABLE    C1
C1 =   Variable N°2
  C1 .   Modalité 1 de V2       -.1823
  C1 .   Modalité 2 de V2        .1823
VARIABLE    C2
C2 =   Variable N°1
  C2 .   Modalité 1 de V1        .1823
  C2 .   Modalité 2 de V1       -.1823
VARIABLE    C3
C3 =   Variable N°4
  C3 .   Modalité 1 de V4        .0164
  C3 .   Modalité 2 de V4       -.0164
```

3.2.4 Utilisation des tableaux de contingence 3.4 et 3.5

Si on saisit les tableaux 3.4 ou 3.5 au lieu de 3.3, on obtient les mêmes résultats mais présentés différemment.

Pour le tableau 3.5, on a vu que

$V_4 = C_1$
$V_3 = L_3$
$V_1 = L_2$
$V_2 = L_1$

Les résultats sont présentés en tenant compte de cette nouvelle affectation.

Le modèle s'écrit :

$L_1+L_2+L_3+C_1$

Pour éditer les tableaux de contingence, il suffit de considérer la suite :

L1 L2 L3 C1

(les variables ligne précédant les variables colonne). Les modalités de l'avant dernière variable (L3) sont indiquées en ligne tandis que les modalités de la dernière variable (C1) sont indiquées en colonne pour une modalité de L1 et L2 fixées.

La variable L1 change de modalité à chaque tableau et la variable L2 change de modalité tous les deux tableaux, ce qui correspond au nombre de modalités de L1.

Le premier tableau édité est le suivant:

ENCART 3.21
Edition du premier tableau - (Tableau de contingence 3.4)

```
VARIABLE L1
 MODALITE  1
 VARIABLE L2
 MODALITE  1
      LIGNE: VARIABLE L3
    COLONNE: VARIABLE C1
                            1          2
   1 (observe)           4.0        1.0        5.0
     (estime)            1.9        1.8        3.7
   2 (observe)           2.0        2.0        4.0
     (estime)            1.8        1.8        3.6
                         6.0        3.0        9.0
                         3.7        3.6        7.4
```

Pour le tableau 3.4, on a vu que $V_2 = L_2$
$V_1 = L_1$
$V_4 = C_2$
$V_4 = C_1$

En considérant la suite de variables L1 L2 C1 C2 et en procédant de la même façon que précédemment, on obtient comme premier tableau de contingence:

ENCART 3.22
Edition du premier tableau - (Tableau de contingence 3.5)

```
 VRIABLE L1
 MODALITE  1
 VARIABLE L2
 MODALITE  1

      LIGNE: VARIABLE C1
    COLONNE: VARIABLE C2
                              1          2
   1 (observe)              4.0        1.0        5.0
     (estime)               1.9        1.8        3.7

   2 (observe)              2.0        2.0        4.0
     (estime)               1.8        1.8        3.6

                            6.0        3.0        9.0
                            3.7        3.6        7.4
```

4. Une étude

4.1 Description des données

On désire étudier l'influence de la dose de radiations reçue (mesurée en Rad) et de l'âge des individus au moment de l'éclatement de la bombe atomique au Japon sur le risque de mourir de la leucémie.

4.1.1 Le tableau de contingence

Les données se présentent sous la forme d'un tableau de contingence croisant trois variables:

DOSE (L1): Variable continue regroupée en 6 classes
AGE (L2) : Variable continue regroupée en 5 classes
ETAT(C1) : Variable réponse comprenant 2 modalités:
* ML: Mort de la leucémie
* NML: Non mort de la leucémie

On dispose d'un échantillon de 105634 individus représentant un peu moins du tiers des survivants après les deux bombardements nucléaires d'août 1945 au Japon.

TABLEAU 4.1
Tableau de contingence croisant les trois variables

		ETAT (C1)	
DOSE (L1)	**AGE (L2)**	ML	NML
Absent de la ville	0 à 9 ans	0	5015
	10 à 19 ans	5	5973
	20 à 34 ans	2	5669
	35 à 49 ans	3	6158
	50 ans et plus	3	3695
0 à 9 Rad	0 à 9 ans	7	10752
	10 à 19 ans	4	11811
	20 à 34 ans	8	10828
	35 à 49 ans	19	12645
	50 ans et plus	7	9053
10 à 49 Rad	0 à 9 ans	3	2989
	10 à 19 ans	6	2620
	20 à 34 ans	4	2798
	35 à 49 ans	3	3566
	50 ans et plus	3	2415
50 à 99 Rad	0 à 9 ans	1	694
	10 à 19 ans	1	771
	20 à 34 ans	1	797
	35 à 49 ans	2	972
	50 ans et plus	2	655
100 à 199 Rad	0 à 9 ans	4	418
	10 à 19 ans	3	792
	20 à 34 ans	3	596
	35 à 49 ans	1	694
	50 ans et plus	2	393
200 Rad et plus	0 à 9 ans	11	387
	10 à 19 ans	6	820
	20 à 34 ans	7	624
	35 à 49 ans	10	608
	50 ans et plus	6	289
SOMME		***137***	***105497***

Nous utiliserons les notations suivantes :

n_{ijk} désigne l'effectif de la case (i,j,k) du tableau de contingence avec :

i : la ième modalité de la variable L1 Dose (i = 1,...,6)
j : la jème modalité de la variable L2 Age (j = 1,...,5)
k: la kième modalité de la variable C1 Etat (k = 1,2)

n: nombre d'observations ($n = \sum_{i,j,k} n_{ijk}$). Ici n = 105497

On suppose que n_{ijk} est réalisation du vecteur aléatoire f_{ijk} suivant une loi multinomiale. $P = (p_{ijk})$ désigne le vecteur inconnu des probabilités.

4.1.2 Distribution des données

Pour obtenir la distribution des données, on effectue le tri à plat des trois variables. Pour cela on crée à partir du tableau de contingence 4.1 un tableau individus-variables à 6*5*2 = 60 lignes et à quatre colonnes. Les trois premières colonnes indiquent respectivement la modalité i; la modalité j et la modalité k. La quatrième colonne indique pour le triplet (i,j,k) l'effectif n_{ijk}. Le tableau ainsi créé se trouve en annexe 4.1.

La procédure STATS de SPAD donne les tris à plat dans l'encart VII.1

Pour équilibrer les effectifs dans les modalités, on peut procéder à un recodage de la variable L1 en regroupant les modalités 3, 4, 5 et 6. On utilise pour cela la procédure ESCAL qui se trouve en annexe 4.2. On obtient une nouvelle variable L1 à trois modalités (absent de la ville - de 0 à 9 rad. - de 10 rad. et plus).

ENCART 4.1
Tri à plat des trois variables

```
                    -- EFFECTIFS --          -- POIDS --
               Absolu  %/Total  %/Expr. Absolu  %/Total  %/Expr. HIST. DES POIDS
    1 . var L1: Dose
L101-L1 mod 1  10  16.67  16.67  26523.00  25.11  25.11 ***********
L102-L1   mod   2      10      16.67      16.67      55134.00      52.19      52.19   ************************
L103-L1 mod 3  10  16.67  16.67  14407.00  13.64  13.64 *******
L104-L1 mod 4  10  16.67  16.67   3896.00   3.69   3.69 **
L105-L1 mod 5  10  16.67  16.67   2906.00   2.75   2.75 **
L106-L1 mod 6  10  16.67  16.67   2768.00   2.62   2.62 **
ENSEMBLE       60 100.00 100.00 105634.00 100.00 100.00
   2 . var L2 : Age
L201-L2 mod 1  12  20.00  20.00  20281.00  19.20  19.20 *********
L202-L2 mod 2  12  20.00  20.00  22812.00  21.60  21.60 **********
L203-L2 mod 3  12  20.00  20.00  21337.00  20.20  20.20 **********
L204-L2 mod 4  12  20.00  20.00  24681.00  23.36  23.36 ***********
L205-L2 mod 5  12  20.00  20.00  16523.00  15.64  15.64 ********
ENSEMBLE       60 100.00 100.00 105634.00 100.00 100.00
   3 . var C1: Etat
C101-C1 mod 1  30  50.00  50.00    137.00   0.13   0.13 *
C102-C1 mod 2  30  50.00  50.00 105497.00  99.87  99.87 **********************************
ENSEMBLE       60 100.00 100.00 105634.00 100.00 100.00
```

Le tri à plat de la variable L1 recodée est le suivant :

ENCART 4.2
Tri à plat de la variable L1 recodée

```
             ------ EFFECTIFS -------     ------ POIDS -------
            Absolu  %/Total  %/Expr.  Absolu  %/Total  %/Expr.  HIST. DES POIDS
5 . L1 RECODEE
AO01 - absent   10   16.67   16.67   26523.00  25.11  25.11 ***********
AO02 - de 0 à 9 10   16.67   16.67   55134.00  52.19  52.19 ************************
AO03 - 10 et +  40   66.67   66.67   23977.00  22.70  22.70 **********
ENSEMBLE        60  100.00  100.00  105634.00 100.00 100.00
```

Le tableau de contingence devient :

TABLEAU 4.2
Tableau de contingence avec la variable L1 recodée

		ETAT (C1)	
DOSE (L1)	**AGE (L2)**	ML	NML
Absent de la ville	0 à 9 ans	0	5015
	10 à 19 ans	5	5973
	20 à 34 ans	2	5669
	35 à 49 ans	3	6158
	50 ans et plus	3	3695
0 à 9 Rad	0 à 9 ans	7	10752
	10 à 19 ans	4	11811
	20 à 34 ans	8	10828
	35 à 49 ans	19	12645
	50 ans et plus	7	9053
10 Rad. et plus	0 à 9 ans	19	4488
	10 à 19 ans	16	5003
	20 à 34 ans	15	4815
	35 à 49 ans	16	5840
	50 ans et plus	13	3752
SOMME		***137***	***105497***

4.2 Etude

4.2.1 Recherche manuelle du « meilleur » modèle

Dans la mesure où il s'agit d'expliquer les variations sur une variable (C1: l'individu est mort ou non de la leucémie à la suite de la bombe atomique au Japon) en fonction des deux autres L1: la dose de radiation reçue et L2 l'âge de l'individu, tous les modèles pertinents incluront la totalité des effets relatifs aux variables L1 et L2, c'est-à-dire le terme d'interaction L1L2. De cette façon, l'analyse prend en compte les liaisons qui existent entre les deux variables explicatives.
L'analyse effectuée peut être synthétisée dans le tableau qui suit :

TABLEAU 4.3
Tableau récapitulatif des modèles étudiés

MODELE	d.d.l	G^2	P. critique	Part de G^2 expliquée en %
[1] (L1L2)(C1)	14	96.207	0.000	--
[2] (L1L2)(L1C1)	12	16.699	0.1613	82.64%
[3] (L1L2)(L2C1)	10	94.021	0.000	2.27%
[4] (L1L2)(L1C1)(L2C1)	8	14.914	0.0608	84.49%
[5] Modèle saturé (L1L2C1)	0	0	1	100%

Chaque modèle s'exprime en termes d'effets qui s'exercent sur la variable C1, indiquant si l'individu est mort ou non de la leucémie. Il s'agit de sélectionner le modèle qui, tout en étant le plus simple possible, s'ajuste correctement au tableau initial 4.2.

Le modèle [1] qui forme le modèle de base suppose que ni l'âge ni la dose de radiation reçue n'influence le risque de mourir de la leucémie.
Les modèles [2] et [3] autorisent l'effet de l'une ou l'autre des variables. Le modèle [4] permet une influence des deux variables l'âge et la dose de radiation reçue.
Le modèle [5] qui est le modèle saturé dispose en plus de l'interaction entre les trois variables. Ce qui stipule que la liaison entre par exemple la dose de radiation reçue et le risque de mourir de la leucémie dépend de l'âge de l'individu.
Quatre aspects méritent d'être soulignés:

- Le modèle [1] est très éloigné de la réalité.
- Les seuls modèles non saturés acceptables au sens du Khi-2 sont les modèles [2] et [4].
- Ces deux modèles incluent l'effet (L1C1). On conclut donc que la dose de radiation reçue influe sur le risque de mourir de la leucémie.
- Le modèle [4] comporte en plus le terme d'interaction (L2C1) c'est-à-dire une liaison entre l'âge de l'individu et le risque de mourir de la leucémie. Cependant dans le tableau VII.3 l'examen des parts de G^2 expliquées en pourcentage pour les deux modèles [2] et [3] indique que l'influence prédominante est exercée par la dose de radiation reçue (82.64% contre 2.27%).

Les modèles [4] et [5] sont emboîtés il est alors possible de tester l'apport du paramètre L1L2C1:

Soit H_0: L1L2C1 = 0

$$\begin{aligned} G^2(\text{Modèle}[4]/\text{Modèle}[5]) &= G^2(\text{Modèle}[4]) - G^2(\text{Modèle}[5]) \\ &= 14.914 \end{aligned}$$

Sous H_0, $G^2(\text{Modèle}[4]/\text{Modèle}[5])$ suit asymptotiquement un Khi-2 à 8 degré de liberté. Or

$$P(\chi_8^2 \geq 14.914) = 0.0608$$

On ne peut pas rejeter H_0 avec un seuil de 0.05. Par conséquent le modèle [4] semble être aussi bon que le modèle [5].

On peut également tester l'apport du paramètre (L2C1) avec les modèles [2] et [4]

$$\begin{aligned} G^2(\text{Modèle}[2]/\text{Modèle}[4]) &= G^2(\text{Modèle}[2]) - G^2(\text{Modèle}[4]) \\ &= 16.699 - 14.914 \\ &= 1.785 \end{aligned}$$

Sous Ho, $G^2(\text{Modèle}[2]/\text{Modèle}[4])$ suit asymptotiquement un Khi-2 à 4 degré de liberté. Or

$$P(\chi_4^2 \geq 1.785) = 0.78$$

On ne peut pas rejeter H_0 avec un seuil de 0.05. Par conséquent le modèle [2] est aussi bon que le modèle [4].

On choisit le modèle le plus simple, c'est-à-dire le modèle [2].

$$\text{Modèle}\,2 : \text{Log}f_{ijk} = \mu + \mu_i(L1) + \mu_j(L2) + \mu_k(C1) + \mu_{ij}(L1L2) + \mu_{ik}(L1C1)$$

On conclut en définitive que la dose de radiation reçue influence le risque de mourir de la leucémie (présence du terme L1C1). Par contre l'âge de l'individu semble ne pas avoir d'influence (le terme L2C1 n'est pas significativement différent de zéro). L'absence du terme L1L2C1 indique que la liaison entre la dose reçue et l'état de l'individu est de même intensité quelque soit son âge.
La procédure LOGLI de SPAD calcule les fréquences espérées estimées par le modèle [2] : (Tableau 4.4).
On peut désormais procéder à l'examen des paramètres du modèle choisi. Le tableau 4.5 ci-dessous donne les estimations des paramètres, les écarts type et les valeurs standardisées. Ces dernières sont obtenues en divisant l'estimation du paramètre par l'écart type correspondant.

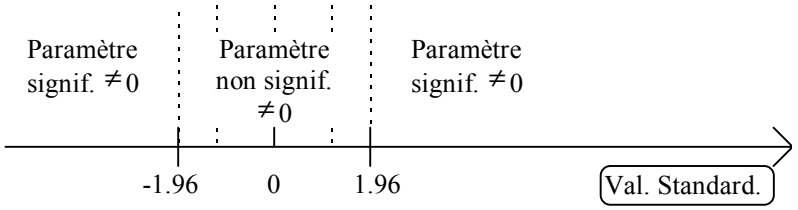
TABLEAU 4.4
Estimation des fréquences espérées par le modèle [2]

		ETAT (C1)	
DOSE (L1)	**AGE (L2)**	ML	NML
Absent de la ville	0 à 9 ans	2.6	5012.9
	10 à 19 ans	3	5975
	20 à 34 ans	2.9	5668.1
	35 à 49 ans	3.1	6157.9
	50 ans et plus	1.9	3696.1
0 à 9 Rad	0 à 9 ans	8.8	10750.2
	10 à 19 ans	9.6	11805.4
	20 à 34 ans	8.8	10827.2
	35 à 49 ans	10.3	12653.7
	50 ans et plus	7.4	9052.6
10 Rad. et plus	0 à 9 ans	14.8	4492.2
	10 à 19 ans	16.5	5002.5
	20 à 34 ans	15.9	4814.1
	35 à 49 ans	19.3	5836.70
	50 ans et plus	12.4	3752.6
SOMME		***137.3***	***105497.2***

TABLEAU 4.5
Estimation des paramètres, écarts-type, valeurs standardisées

Paramètres indépendants	Estimation	Ecart-type	Valeur standardisée
μ	5.3734		
$\mu_1(L1)$	-0.6050	0.0954	-6.34
$\mu_2(L1)$	0.3726	0.0697	5.34
$\mu_1(L2)$	-0.0367	0.00674	-5.44
$\mu_2(L2)$	0.0889	0.00642	13.85
$\mu_3(L2)$	0.0297	0.00654	4.54
$\mu_4(L2)$	0.1735	0.00621	27.94
$\mu_1(C1)$	-3.4008	0.0548	-62.06
$\mu_{11}(L1L2)$	-0.00320	0.00991	-0.32
$\mu_{12}(L1L2)$	0.0468	0.00936	5
$\mu_{13}(L1L2)$	0.0532	0.00954	5.58
$\mu_{14}(L1L2)$	-0.00768	0.00917	-0.84
$\mu_{21}(L1L2)$	0.0183	0.00837	2.19
$\mu_{22}(L1L2)$	-0.0137	0.00801	-1.71
$\mu_{23}(L1L2)$	-0.0410	0.00821	-4.99
$\mu_{24}(L1L2)$	-0.0289	0.00778	-3.71
$\mu_{11}(L1C1)$	-0.3905	0.0954	-4.09
$\mu_{21}(L1C1)$	-0.1542	0.0697	-2.21

Sous l'hypothèse nulle stipulant la nullité d'un paramètre la valeur standardisée correspondante suit asymptotiquement la loi normale centrée réduite. On peut tester la significativité de chaque paramètre en appliquant la règle ci-dessous :



L'adéquation du modèle est confirmée par l'examen des valeurs standardisées. Les paramètres sont quasiment tous significativement différents de zéro.
A partir du tableau des fréquences espérées estimées croisant les variables L1 dose de radiation reçue et C1 l'état du malade, calculons les différents odds-ratios

TABLEAU 4.6
Tableau des fréquences espérées estimées
croisant les variables dose L1 et état C1

	ML	**NML**
Absent de la ville	13.5	26510
0 à 9 Rad.	44.9	55089.1
10 Rad. et plus	78.9	23898.1

Premier odds-ratio:

$$\frac{\text{individus morts ayant subi 0 à 9 Rad/individus vivants ayant subi 0 à 9 Rad}}{\text{individus morts absents de la ville/individus vivants absents de la ville}} = \frac{44.9/55089.1}{13.5/26510} = 1.60$$

Ainsi les individus ayant reçu de 0 à 9 Rad. ont 1.60 fois plus de risque de mourir de la leucémie que les individus ayant été absents de la ville.

Deuxième odds-ratio:

$$\frac{\text{individus morts ayant subi 10 Rad. et plus / individus vivants ayant subi 10 Rad. et plus}}{\text{individus morts absents de la ville / individus vivants absents de la ville}} = \frac{78.9 / 23898.1}{13.5 / 26510} = 6.48$$

Ainsi les individus ayant reçu de 10 Rad et plus ont 6.48 fois plus de risque de mourir de la leucémie que les individus ayant été absents de la ville.

4.2.2 Utilisation des algorithmes de sélection de modèle de la procédure LOGLI de SPAD

On peut utiliser les algorithmes de sélection de modèle de la procédure LOGLI de SPAD.
La méthode combinatoire et la méthode pas à pas sélectionnent précisément le modèle [2] (4.1) trouvé précédemment.

ENCART 4.3
Tableau récapitulatif de la méthode combinatoire

```
 LISTE DES MODELES TRAITES
 MODELE    1
 M / L1 L2 C1
 MODELE    2
 L1 / L2 C1
 MODELE    3
 L2 / L1 C1
 MODELE    4
 C1 / L1 L2
 MODELE    5
 L1+L2+L1*L2 / C1
 MODELE    6
 L1+C1+L1*C1 / L2
 MODELE    7
 L2+C1+L2*C1 / L1
 MODELE    8
L1+L2+C1+L1*L2+L1*C1+L2*C1+L1*L2*C1
```

```
LE MODELE SELECTIONNE EST LE SUIVANT :
L1+L2+C1+L1*L2+L1*C1
...........
NUMERO  DEGRE  ESTIMATION   ESTIMATION
 DU      DE    CHI-2 DE     RAPPORT DE       AIC         AIC
MODELE LIBERTE PEARSON     VRAISEMBLANCE   PEARSON    VRAISEMBLANCE
............
   1     29   142944.391    162376.250    142886.391     162318.250
                                         ( 0.0000 )      ( 0.0000 )
   2     27   108867.312    146390.828    108813.312     146336.828
                                         ( 0.0000 )      ( 0.0000 )
   3     25   139361.469    160556.672    139311.469     160506.672
                                         ( 0.0000 )      ( 0.0000 )
   4     28    19018.836     18037.662     18962.836      17981.662
                                         ( 0.0000 )      ( 0.0000 )
   5     15   105085.766    144434.766    105055.766     144404.766
                                         ( 0.0000 )      ( 0.0000 )
   6     24     1913.379      1972.768      1865.379       1924.768
                                         ( 0.0000 )      ( 0.0000 )
   7     20    17219.215     16215.932     17179.215      16175.932
                                         ( 0.0000 )      ( 0.0000 )
   8      0      ----          ----          ----           ----
   9     12       16.717        16.699        -7.283         -7.301
                                         ( 0.1606 )      ( 0.1613 )
```

Si on décide d'étudier la totalité des modèles log-linéaires hiérarchiques possibles, on constate que le modèle qui possède le plus petit critère d'AIC est également le modèle 2 (4.1). Pour cet exemple, la méthode combinatoire et la méthode pas à pas ont sélectionné le modèle possédant le plus petit critère d'Akaike.

La discrimination par réseau de neurones

1. Présentation

1.1 Introduction

Les réseaux de neurones se sont développés dans plusieurs domaines durant ces dernières années, et apparaissent aujourd'hui comme un outil universel. Ils ont renouvelé la discipline connue sous le nom de *reconnaissance de formes* qui recouvre beaucoup d'applications industrielles (notamment en temps réel) en discrimination. Les réseaux peuvent être considérés comme des modèles statistiques d'une grande flexibilité. En statistique, ils sont utilisés en tant que classifieur (analyse discriminante), système d'identification des classes (classification automatique), et estimateurs *non-paramétriques* de régression *non linéaire*.
Les réseaux de neurones sont fondés au départ, sur des analogies biologiques et sur la modélisation mathématique des mécanismes de perception visuelle et auditive. Ils ont acquis depuis une certaine autonomie. Les travaux récents montrent que les techniques fondamentales de l'inférence statistique, telles que les tests d'hypothèses et l'estimation, s'appliquent aux réseaux de neurones. Des articles de synthèse montrent qu'il existe une complémentarité entre les réseaux et la statistique.

Malgré leur flexibilité et leur universalité, les réseaux de neurones ont des limitations en modélisation statistique qui apparaissent, notamment à propos du compromis "biais/variance". En effet, pour un problème donné et des échantillons de taille fixe, le biais d'un estimateur décroît et sa variance augmente. Un réseau *sous-dimensionné* aura un nombre de degrés de liberté trop faible, l'erreur aura donc un biais important et une variance faible. Un réseau *sur-dimensionné* aura un grand nombre de degrés de liberté et l'optimisation à partir d'échantillons d'apprentissage différents conduira à des estimations pouvant être très différentes, ce qui correspond à une variance importante. On ne peut donc pas diminuer le biais sans augmenter sa variance.
On présente ici le modèle le plus répandu dans le cadre de discrimination, le Réseau Perceptron Multicouche, RPM. Le réseau Perceptron le plus simple a uniquement deux couches de neurones: l'une des couches représente les entrées du système, l'autre les sorties. Les RPM plus complexes ont trois couches ou plus. Nous nous intéressons à l'algorithme d'apprentissage en tant que procédure d'optimisation, et plus particulièrement à l'algorithme de rétropropagation du gradient. On trouve une présentation détaillée des réseaux de neurones dans par exemple [Chabanon et *al*. 90]. Dans le cadre neuronal, L'écriture de ces procédures d'optimisation a donné naissance à diverses formes de l'algorithme de rétropropagation du gradient.

1.2 Réseaux de neurones multicouches

L'architecture des réseaux multicouches est organisée en niveaux ou couches de neurones. Les connexions interneuronales s'effectuent d'une couche à l'autre et jamais sur une même couche. Chaque neurone ou cellule d'une couche est directement connecté à tous les neurones de la couche suivante.

Le perceptron est un réseau comprenant une seule couche de connexions et donc deux couches de neurones; l'une des couches représente les entrées du système et l'autre les sorties. Dans le cadre de la discrimination, ce réseau appartient à la famille des algorithmes supervisés.

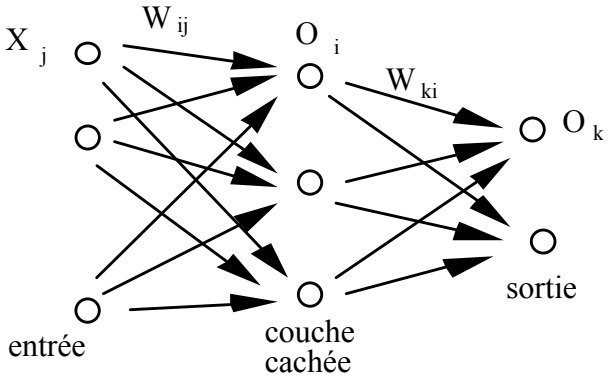


figure 1. Exemple d'un réseau de neurones.

Le réseau de la *figure 1* est dit *réseau multicouche*. Ce réseau est entièrement défini par les connexions de voisinage W_{ij} et la fonction de transition (de transfert). Avec les connexions de voisinage, tous les neurones j de couche c de ce réseau produisent une réponse O_i sur les neurones i de la couche (c + 1). Cette réponse, ou entrée des neurones i, s'obtient en calculant une somme pondérée des sorties O_j des neurones de la couche auxquels ils sont connectés. Cette somme est ensuite transformée par une fonction non linéaire dérivable g :

$$O_i = g\left(\sum_{j=1}^{K(c)} W_{ij} O_j - \theta_i\right) \qquad i = 1.....K(c+1)$$

avec

θ_i = seuil du neurone i

W_{ij} = poids de la connexion entre les neurones i et j

g(x) fonction de transition (en général une sigmoïde)

O_i et O_j sorties des neurones i et j

K(c) et K(c + 1) : nombre de neurones des couches c et (c + 1).

La fonction de transition est souvent la fonction sigmoïde :

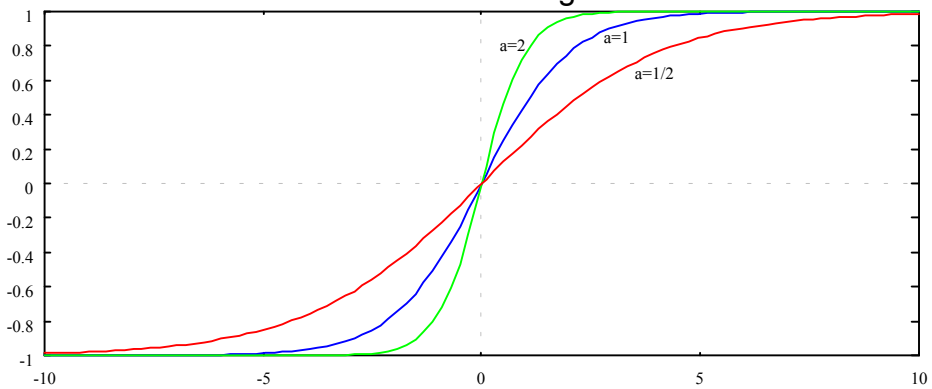


figure 2 La fonction sigmoïde $g(x) = \dfrac{(1 - e^{-ax})}{(1 + e^{-ax})}$.

L'algorithme de rétropropagation du gradient de l'erreur se décompose de la manière suivante :

- Propagation d'un événement à travers le réseau :

Les individus se présentent séquentiellement en entrée par un vecteur X à p dimensions. Les neurones calculent leur réponse et le réseau évolue pour arriver à un certain état O qui est comparé à la réponse désirée. L'état interne se caractérise par $e_i = \sum_j W_{ij} X_j - \theta_i$. Lorsque e_i est grand (respectivement petit), alors $O_i = g(e_i)$ est grand (respectivement petit), il y a activation (respectivement inhibition).

- Rétropropagation du gradient d'une erreur :

Le mécanisme d'apprentissage repose sur la minimisation d'une fonction d'erreur par un algorithme adaptatif de type gradient. Le réseau ajuste les poids W_{ij} de connexions entre les neurones pour réaliser la correspondance souhaitée entre l'état O obtenue et la réponse désirée d. Cette fonction évalue l'écart entre la sortie calculée et la sortie désirée sur la dernière couche du réseau (couche de sortie). La fonction d'erreur choisie est en général la fonction moyenne quadratique.

Pour chaque individu présenté t, le réseau cherche à minimiser une erreur E commise entre la réponse effective $O_k^{(t)}$ des K neurones de la couche de sortie et la réponse désirée $d_k^{(t)}$:

$$E = \frac{1}{N}\sum_{t=1}^{N} E_t = \frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{K}(O_k^{(t)} - d_k^{(t)})^2 \qquad (1)$$

avec

N le nombre des individus dans l'ensemble d'apprentissage
K le nombre de neurones de la couche de sortie
$O_k^{(t)}$ est la valeur de neurone k de la couche de sortie obtenue lors de la présentation de l'individu t.
$d_k^{(t)}$ la valeur désirée de l'individu t à la sortie du neurone k.

Pour ce faire, le gradient de E est rétropropagé en modifiant la valeur des poids :

$$W_{ij}(t) = W_{ij}(t-1) + \Delta W_{ij}(t)$$

$$\text{avec } \Delta W_{ij}(t) = -\varepsilon \frac{\partial E}{\partial W_{ij}}$$

où ε est un paramètres d'apprentissage du réseau.

Ces deux étapes sont répétées. L'ensemble des individus d'apprentissage est présenté plusieurs fois et de manière séquentielle jusqu'à l'obtention d'un minimum acceptable de la fonction d'erreur.

1.3 Construction des réseaux neuronaux

Abordons le Réseau Perceptron Multicouches dont les neurones sont organisés en une couche d'entrée, une couche de sortie et une ou plusieurs couches intermédiaires. Leur utilisation repose sur un théorème qui, sous des conditions de régularité relativement modestes, affirme qu'un réseau à tois couches de neurones (une seule couche cachée) peut donner une approximation aussi bonne que possible d'une fonction quelconque de plusieurs variables. La qualité de cette approximation augmente en fonction du nombre de neurones employés.

1.3.1 Architecture du Réseau Perceptron Multicouches

Le nombre de neurones dans la couche d'entrée ne dépend que du nombre de variables explicatives, et le nombre de neurones dans la couche de sortie ne dépend que du nombre de classes à discriminer. Un réseau peut avoir une ou plusieurs couches cachées.

1.3.1.1 Le nombre de couches cachées

Il n'existe pas de méthode générale réellement convaincante pour fixer de façon idéale le nombre de couches cachées. Un réseau utilise une couche cachée de neurones pour créer sa propre représentation interne en fonction du problème à résoudre. Cette couche est alors considérée comme un niveau de prétraitement de l'information avant la décision finale.

Ainsi, pour s'approcher empiriquement du nombre optimal de couches cachées, il est conseillé de respecter la complexité du problème et la nature des données. La détermination de l'architecture du réseau dépend aussi de la capacité de l'ordinateur.

Plus un réseau a de couches cachées (donc plus de connexions), plus le traitement nécessite de mémoire et de temps de calcul. En général, si l'on effectue rigoureusement une sélection de variables explicatives, l'emploi d'un RPM à une ou deux couches cachées suffit.

La sélection des variables peut être effectuée en utilisant des critères statistiques. Pour cela, on peut utiliser la procédure DEMOD ou FUWIL de SPAD, ainsi que la segmentation par arbre binaire de SPAD.S.

Le nombre de neurones dans la couche cachée

La détermination du nombre de neurones dans la couche cachée est délicate.

En général, la qualité de l'estimation augmente en fonction du nombre de neurones utilisés. Mais l'utilisation de nombreux neurones entraînera l'emploi d'un système d'apprentissage complexe qui pénalise le réseau. Pour un problème donné et des échantillons de taille fixe, un réseau « sous-estimé » (le nombre de neurones dans la couche cachée ou le nombre de couches cachées est insuffisant), aura un nombre de degrés de liberté trop faible; la fonction d'erreur aura une composante de biais importante et un terme de variance faible; le modèle est stable, mais a une performance relativement faible.

Par contre, un réseau « sur-estimé » (un grand nombre des couches cachées ou un grande nombre de neurones dans la couche cachée), possède un grand nombre de degrés de liberté et l'optimisation à partir d'échantillons d'apprentissage différents conduira à des solutions pouvant être différentes, ce qui correspond à une composante de variance importante. Le modèle neuronal devient instable.

Sous l'hypothèse que le RPM ne dispose que d'une seule couche cachée, on peut utiliser la technique de rééchantillonnage, en faisant varier le nombre de neurones dans la couche cachée pour la sélection du réseau optimal. Un nombre trop élevé de neurones entraîne d'inévitables redondances et donc pénalise le temps de calcul inutilement employé.

1.3.1.2 Sélection d'un meilleur réseau de neurones

Le dilemme « qualité/complexité » mentionné précédemment joue ici son rôle; le meilleur réseau n'est pas forcément celui qui minimise la fonction d'erreur sur l'ensemble des individus d'apprentissage, mais plutôt celui qui offre une bonne « généralisation » (ou prévision). L'idée générale est qu'on choisit un réseau par compromis entre l'erreur obtenue et le nombre de paramètres du réseau. Un réseau produit une erreur acceptable sur l'ensemble des individus d'apprentissage et possède un nombre faible de connexions, ce qui assure une meilleure généralisation. On peut mettre en œuvre certaines techniques statistiques, notamment celle de la *validation croisée* , pour sélectionner le meilleur réseau de neurones.

Etude du coefficient d'apprentissage ε

Remarquons que la rétropropagation du gradient descendant est une approche assez intuitive. Il s'agit d'ajuster les paramètres de la fonction d'erreur dans le sens de la descente (c'est-à-dire dans le sens de la réduction de l'erreur globale) jusqu'à l'approximation de la fontière de séparation entre les classes. La mesure utilisée est la dérivée première (c'est-à-dire la pente de la fonction).
Le coefficient ε d'apprentissage représente l'ampleur du déplacement que l'algorithme réalise dans le sens de la pente. Si ε a une valeur faible, les déplacements successifs sont faibles. Cela implique une augmentation de l'apprentissage mais, en contrepartie, une convergence précise vers le minimum (en effet, à chaque pas la pente est réajustée, permettant ainsi de coller très bien à la fonction d'erreur). Si ε a une valeur grande, il y a risque d'oscillation autour du minimum. On prend en général une valeur comprise entre 0 et 1.

1.4 ANNEXE : algorithme de rétropropagation du gradient

A chaque élément n de l'échantillon est associé un vecteur de représentation $x^{(n)}$ et les K valeurs $y_1^{(n)},\ldots,y_k^{(n)}$ caractérisant la sortie désirée. On notera :

- $g_i^{(c)}$ la fonction de transition du neurone i de la couche c.
- $s_i^{(c)}(n)$ la valeur de la sortie du neurone i de la couche c pour l'individu présenté n.

- $e_i^{(c)}(n)$ la valeur de l'entrée du neurone i de la couche c pour l'élément présenté n.
- $W_{ij}^{(c)}(n)$ le poids de la connexion entre le neurone i de la couche $(c+1)$ et le neurone j de la couche c pour l'élément présenté n.
- $\boldsymbol{n}_c$ le nombre de neurones de la couche c et par NC le nombre de couches de ce réseau.

Nous supposons que tous les neurones ont une fonction de transition continue et dérivable. Sous cette hypothèse, l'apprentissage d'un RPM est supervisé et utilise un algorithme de rétropropagation du gradient de l'erreur qui se décompose de la manière suivante :

Propagation des entrées d'un individu à travers le réseau :

La relation entre les sorties de la couche c et l'entrée du neurone i de la couche $(c+1)$ est une relation linéaire égale à

$$e_i^{(c+1)}(n) = \sum_{j=1}^{n_c} W_{ij}^{(c)} . s_j^{(c)}(n)$$

La relation entre la sortie et l'entrée du neurone i de la couche $(c+1)$ est donnée par la fonction de transition $g_i^{(c+1)}$, d'où

$$s_i^{(c+1)}(n) = g_i^{(c+1)}\left(e_i^{(c+1)}(n)\right) = g_i^{(c+1)}\left(\sum_{j=1}^{n_c} W_{ij}^{(c)} . s_j^{(c)}(n)\right)$$

Ajustement des poids par la méthode du gradient :

Si toutes les fonctions de transition sont égales, alors le réseau est déterminé uniquement par les poids et par le choix de fonction de transition g. Les valeurs de ces poids contribuent à l'excitation (positivité des poids) ou à l'inhibition (négativité) des neurones. Dans la phase rétropropagation, l'ajustement des poids du réseau se fait de la façon suivante :

$$W_{ij}^{(c)}(t+1) = W_{ij}^{(c)}(t) + \varepsilon \frac{\partial E(n)}{\partial W_{ij}^{(c)}}$$

Il faut calculer pour chaque poids $W_{ij}^{(c)}$ le gradient de E, d'où

$$\frac{\partial E}{\partial W_{ij}^{(c-1)}} = \frac{\partial E}{\partial e_i^{(c)}} . \frac{\partial e_i^{(c)}}{\partial W_{ij}^{(c-1)}} = \frac{\partial E}{\partial e_i^{(c)}} . \frac{\partial}{\partial W_{ij}^{(c-1)}}\left(\sum_{l=1}^{n_{c-1}} W_{il}^{(c-1)} . s_l^{(c-1)}\right) = \frac{\partial E}{\partial e_i^{(c)}} s_j^{(c-1)}$$

Pour le neurone i de la couche de sortie, il faut calculer $\frac{\partial E}{\partial e_i^{(NC)}}$ qui ne dépend que de la fonction de coût E. Par contre pour les neurones des autres couches nous avons la relation :

$$\frac{\partial E}{\partial e_i^{(c)}} = \sum_{l=1}^{n_{c+1}} \frac{\partial E}{\partial e_l^{(c+1)}} . \frac{\partial e_l^{(c+1)}}{\partial e_i^{(c)}}$$

et comme

$$\frac{\partial e_l^{(c+1)}}{\partial e_i^{(c)}} = \frac{\partial}{\partial e_i^{(c)}} \left(\sum_{j=1}^{n_c} W_{lj}^{(c)} . s_j^{(c)} \right)$$

et que seul $s_i^{(c)}$ dépend de $e_i^{(c)}$, nous avons:

$$\frac{\partial e_l^{(c+1)}}{\partial e_i^{(c)}} = W_{li}^{(c)} . g'^{(c)}_i$$

d'où

$$\frac{\partial E}{\partial e_i^{(c)}} = \left(\sum_{l=1}^{n_{c+1}} \frac{\partial E}{\partial e_l^{(c+1)}} . W_{li}^{(c)} \right) g'^{(c)}_i$$

et

$$\frac{\partial E}{\partial W_{ij}^{(c-1)}} = \left(\sum_{l=1}^{n_{c+1}} \frac{\partial E}{\partial e_l^{(c+1)}} . W_{li}^{(c)} \right) g'^{(c)}_i . s_j^{(c-1)}$$

Ainsi pour les couches cachées les poids de ces couches se calculent en fonction des poids de la couche suivante. Donc la modification de la fonction de coût n'entraîne que la modification du mode de calcul des poids de la dernière couche.

2. Exemple

2.1 Les données

L'échantillon est constitué de PME de l'industrie, observées en 1990 et réparties en 2 populations :

- **les entreprises défaillantes** qui ont déposé leur bilan en 1991, 1992 ou 1993 et sont donc observées 1 an ou 2 ans ou 3 ans avant que la défaillance ne survienne ; elles sont au nombre de 809 ;
- **les entreprises non défaillantes** qui n'ont jamais déposé leur bilan ; beaucoup plus nombreuses que les précédentes, elles ont fait l'objet d'un tirage aléatoire assurant la représentativité de l'échantillon au niveau secteur et taille des entreprises ; elles sont au nombre de 1381.

Pour les exercices comptables 1986 à 1989 et 1991-1992, on a constitué des échantillons analogues qui serviront d'échantillons de validation.

La détection précoce des défaillances d'entreprise repose ici sur des variables économiques et financières. Le large choix initial de 50 variables intègre les principaux aspects à l'analyse des bilans du point de vue de la bonne santé de l'entreprise.

Dans le but de construire un outil de détection précoce de la défaillance d'entreprise, on peut appliquer au même fichier de données individuelles économiques et financières l'analyse discriminante linéaire de Fisher, la régression logistique et un réseau de neurones multicouche. Cet exemple est tiré des travaux de [Bardos et Zhu 95].

2.2 La sélection préalable des variables

2.2.1 L'algorithme de Furnival et Wilson (Proc FUWIL de SPAD.D)

On peut sélectionner les meilleurs sous-ensembles de variables par un modèle de régression linéaire multiple suivant trois critères possibles : R^2 maximum, R^2 ajusté maximum, C_p de Mallows minimum.

Parmi p variables, on peut choisir 2^p-1 sous-ensembles non vide. Pour chaque taille j de sous-ensemble de variables, le programme édite les meilleurs choix de sous-ensemble pour les trois critères ci-dessus. Appliqué ainsi ce programme serait très coûteux en temps calcul. On limite considérablement le nombre d'opérations en utilisant l'algorithme « leaps and bounds » de Furnival et Wilson. Parmi les 50 variables, on se ramènera à 16 variables et même à 8 variables seulement.

2.3 L'analyse linéaire discriminante (Proc DIS2G)

Dans le cas de deux groupes N et D, après avoir sélectionné les k variables les plus discriminantes, on s'intéresse à la règle de décision suivante :

Si y^N et y^D sont les points moyens de chacun des 2 groupes, on affecte l'entreprise "a" au groupe N si et seulement si $d(a, y^N) \leq d(a, y^D)$, c'est-à-dire, si et seulement si, la distance de "a" à y^N est inférieure à la distance de "a" à y^D.

Ceci se traduit par l'inégalité : $f(a) = (y^N - y^D)' \; T^{-1} \; (a - \frac{y^N + y^D}{2}) \geq 0$, quand on choisit pour métrique T^{-1} avec les formules les notations suivantes :

- *a est le vecteur des k variables a_1, a_2 ... a_k de l'entreprise "a";*
- *T est la matrice de variance covariance totale ;*
- *$(y^N - y^D)' \; T^{-1}$ est le vecteur α des k coefficients de la fonction linéaire discriminante ;*
- *$(y^N - y^D)' \; T^{-1}(- \frac{y^N+y^D}{2}) = \beta$ est la constante ;*
- f *est la fonction score ; $f(a) = \alpha_1 a_1 + ... + \alpha_k a_k + \beta$ est le score de l'entreprise a.*
- f*(x) = 0 n'est autre que l'équation de l'hyperplan qui sépare au mieux les nuages de points N et D ;*
- *$p = \frac{y^N+y^D}{2}$ est le vecteur des k valeurs pivots p_1, p_2 ... p_k.*

2.4 La régression logistique (Proc LOGLI de SPAD.D)

La régression logistique est une méthode dans laquelle la variable endogène Y correspond au codage : 0 si la firme est défaillante, 1 sinon ; X est la matrice des variables exogènes, à savoir une batterie de variables sélectionnées par la procédure FUWIL.
Pour l'entreprise i, on suppose que :

$$Y_i = \begin{cases} 0 & \text{si } \beta + X_i \, \alpha + u_i \leq 0 \\ 1 & \text{sinon} \end{cases}$$

Où β est une constante et α est le vecteur colonne des coefficients de la combinaison linéaire à estimer. Les u_i sont les perturbations supposées indépendantes, de moyenne nulle, de variance 1 ; elles sont supposées suivre une loi logistique de fonction de répartition :

$$F(x) = \frac{1}{1 + e^{-x}}$$

X_i est le vecteur de variables qui caractérisent l'entreprise i, et $s = \alpha + X_i \, \beta$ est son « score ». Compte tenu de ces hypothèses :

$$P(Y_i = 0 / X_i) = F\left[-\beta - X_i \, \alpha\right] = \frac{1}{1+e^{\beta + X_i \, \alpha}}$$

$$P(Y_i = 1 / X_i) = 1 - F\left[-\beta - X_i \, \alpha\right] = \frac{1}{1+e^{-\beta - X_i \, \alpha}}$$

Les paramètres α et β sont estimés par la méthode du maximum de vraisemblance. La fonction $\beta + X\alpha$ est considérée comme la fonction discriminante. L'hypothèse probabiliste du modèle permet de calculer la probabilité pour chaque observation d'être codée 0 ou 1.

2.5 Comparaison des résultats : les pourcentages de bons classements

La règle de décision associée à un score permet d'affecter chaque entreprise à un des deux groupes. Toute entreprise a donc un groupe auquel elle appartient réellement et un groupe auquel elle est affectée. Le décompte des affectations correctes, c'est-à-dire correspondant au groupe réel, fournit les pourcentages de bons classements.

Le pourcentage de bons classements dépend du seuil de décision. Dans la phase de construction de la fonction linéaire on prend pour seuil 0, qu'il s'agisse de la fonction linéaire de Fisher ou de la régression logistique. Pour mieux connaître l'efficacité de la méthode, il est nécessaire de calculer les pourcentages de bons classements dans chaque groupe. Ceux-ci sont calculés par catégorie sur **l'échantillon de base**, mais aussi sur tout autre échantillon appelé **échantillon-test**. On a ainsi une mesure de la qualité de la fonction score.

2.6 Sensibilité de l'algorithme au coefficient ε

Les réseaux de neurones souffrant fréquemment de « sur-apprentissage » on a scindé l'échantillon de base en deux : l'échantillon d'apprentissage (suivant le cas il représente les 2/3 ou les 9/10 de l'échantillon de base) et son complément l'échantillon test. On utilise aussi la validation croisée. Cette pratique permet de sélectionner les réseaux les plus intéressants, et de les tester ensuite sur d'autres tableaux de données.

On utilise la fonction sigmoïde et l'apprentissage s'effectue avec une fonction d'erreur quadratique (encore appelée coût). La vitesse de convergence est alors sensible au coefficient ε.

Le tableau ci-dessous montre comment évoluent le pourcentage de bien classés et le coût en fonction de ε. Le choix du paramètre ε est délicat. Quand ε a une valeur grande (par exemple ε=0.5), on constate une oscillation autour de minimum. Quand ε a une valeur petite (par exemple ε=0.0001), les déplacements successifs sont faibles.

POURCENTAGES DE BIENS CLASSÉS ET COUT EN FONCTION DE ε

Coefficient ε	Pourcentage de bien classés		Coût	
	APP.	Test	APP.	Test
0.5	72.23	73.10	0 .559557	0.548841
0.1	72.44	71.55	0 .447912	0.444563
0.05	71.54	69.72	0.424521	0.439866
0.01	74.86	74.23	0.346475	0.367107
0.005	77.29	72.25	0.332132	0.380004
0.001	74.10	74.93	0.336935	0.333717
0.0005	75.76	71.41	0.329560	0.365224
0.0001	74.10	73.38	0.337683	0.344127

Réseaux utilisés : RPM comportant trois couches. Il y a 8 neurones dans la couche d'entrée, 10 neurones dans la couche cachée et 2 neurones dans la couche sortie.

Les résultats sont obtenus après 2000 présentations de l'échantillon d'apprentissage de taille.

2.7 Les résultats avec cinquante variables

On construit un réseau multicouche en entrant les 50 variables dans le réseau. Nous avons donc un RPM comportant 50 neurones d'entrée, 10 neurones dans la couche cachée et 2 neurones dans la couche sortie. Ce nombre total d'individus est 2190, 721 entreprises sont tirées aléatoirement pour constituer l'échantillon test. On a testé 3 méthodes : l'analyse discriminante linéaire de Fisher (ADL), le réseau de neurones multicouche (RPM), la segmentation.

POURCENTAGES DE BONS CLASSEMENTS PAR LES TROIS MÉTHODES
ε = 0,004 ; 200 présentations de l'échantillon d'apprentissage

	Echantillon d'apprentissage			Echantillon test		
	A.D. Linéaire	RPM	Segmen-tation	A.D. Linéaire	RPM	Segmen-tation
Défaillantes	75,18 %	61,54 %	63,50 %	72,83 %	50,0 %	61,15 %
Non défaillantes	74,20 %	95,12 %	79,34 %	71,94 %	86,71 %	80,61 %
Ensemble	74,69 %	82,43 %	73,35 %	72,39 %	73,27 %	72,20 %

Nous constatons que, sur l'échantillon d'apprentissage, les méthodes d'ADL et de segmentation donnent quasiment les mêmes taux de biens classés (1% de moins pour la segmentation), que le réseau multicouche donne un taux plus élevé qui offre une « généralisation » aussi bonne que celle de l'ADL. En contrepartie, le temps de calcul est beaucoup plus long que celui de l'ADL. Remarquons que le phénomène de sur-apprentissage apparaît dans presque tous les apprentissages utilisant les RPM.

2.8 Résultats avec seize variables parmi 50

Ici on utilise 16 variables parmi les 50 variables pour définir des typologies d'entreprises. Ces 16 variables couvrent un large champs de critères (parfois non linéaires) ayant trait à la défaillance d'entreprises.
Le réseau de neurones comporte une couche cachée à 3 neurones. On pratique une validation croisée avec 10 échantillons (les échantillons tests étant 1/10 de l'échantillon).

MOYENNES DES POURCENTAGES DE BONS CLASSEMENTS PAR VALIDATION CROISÉE

	Échantillon d'Apprentissage (%)		Échantillon Test (%)	
	ADL	RPM	ADL	RPM
défaillantes	71,00	67,50	69,40	60,54
non défaillantes	73,20	82,72	72,90	78,32
ensemble	72,45	77,10	71,60	71,61

Les résultats sont comparables aux précédents. N'utilisant que 16 variables, le temps de calcul est plus avantageux. Tout se passe comme si ces 16 variables résumaient bien l'information du fichier.

2.9 Résultats avec huit variables

Nous cherchons à construire un modèle neuronal moins complexe. On choisit de prendre en entrée, les 8 variables fournissent la meilleure discrimination au sens de la procédure FUWIL.

Les réseaux utilisés sont multicouches. Ils comportent 8 neurones en entrée, 2 neurones en sortie, le nombre des neurones dans la couche cachée peut varier entre 2 et 10. Nous utilisons la technique de l'échantillon test pour choisir le réseau optimal.

Le meilleur modèle neuronal est un réseau Perceptron en 3 couches comportant 3 neurones dans la seule couche cachée, car la fonction d'erreur sur l'échantillon test vaut 0,423197 qui est minimale.

POURCENTAGE DE BIENS CLASSÉS ET COÛT EN FONCTION DU NOMBRE DE NEURONES DE LA COUCHE CACHÉE

	Pourcentage de biens classés		Coût	
Nombre de neurones	APP.	Test.	APP.	Test.
2	73,41	72,39	0,455704	0,471251
3	74,72	72,39	0,379565	0,423197
4	73,75	72,82	0,422183	0,455237
5	76,52	70,85	0,406102	0,514710
6	78,88	72,25	0,402298	0,529572
7	80,06	72,25	0,393130	0,550276
8	78,6	70,7	0,421659	0,579803
9	80,4	69,72	0,391037	0,599482
10	80,61	73,52	0,385257	0,524924

Nous constatons aussi que le réseau de neurones donne un taux de biens classés un peu plus élevé (4 % de plus sur l'échantillon test) que celui de l'ADL et que celui de la segmentation. De plus, comme nous utilisons de moins de variables, le réseau converge vers le minimum assez rapidement et donne une bonne « généralisation ».

L'analyse discriminante linéaire pratiquée sur des fichiers où les variables ont été préalablement sélectionnées aboutit, en un temps calcul très bref, à des fonctions de peu de variables (entre 7 et 10 suivant le cas) avec des pourcentages de bons classements supérieurs à 70 %. En utilisant les résultats de cette sélection, on peut gagner du temps sur les deux autres méthodes : la segmentation et les réseaux de neurones.

Références

[Bardos et zhu 95] **Bardos Mireille et Zhu Wenhua** (1995) " Comparaison de l'analyse discriminante linéaire et des réseaux de neurones - Application à la détection de défaillance d'entreprises ", *Congrès international ANSEG*, 2.

[Chabanon et al. 90] **Chabanon C. et Dubuisson B.** (1990) " Méthodes non probabilistes ", Dans : *Analyse discriminante sur variables continues,* Collection Didactique INRIA.

[Lebart et al. 95] **Lebart Ludovic, Morineau Alain et Piron Marie** (1995) " Statistique exploratoire multidimensionnelle ", *Dunod,* Paris.