Table des matières

Tiré de : *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

Introduction

L'objectif de l'analyse multivariée est d'étudier les interrelations entre plusieurs variables figurant dans une base de données et, si possible, d'en généraliser les conclusions par inférence statistique. L'analyse multivariée réunit un grand nombre de méthodes, souvent complexes, qui tentent de donner une image simplifiée des multiples relations entre les variables d'une enquête ou d'une base de données.

L'importance accordée aux méthodes de l'analyse multivariée correspond à une demande sociale. Les sociétés actuelles font face à une pléthore d'informations spécialisées rendues plus facilement accessibles par le biais de l'Internet. Les collectes des données pullulent et une grande partie de cette information, qui coûte très cher à colliger aux gouvernements et aux entreprises, n'est pas traitée correctement. Faute de connaissances et de moyens, une grande partie de l'information recueillie est laissée en jachère.

La plupart des méthodes d'analyse multivariée sont nées avant la Deuxième Guerre mondiale, mais leur utilisation posait de multiples problèmes : les calculs à effectuer étaient longs, innombrables et fastidieux ; ces méthodes étaient dès lors très peu utilisées. L'apparition d'ordinateurs et de microordinateurs de plus en plus performants allait changer radicalement cette situation ; des logiciels puissants et faciles à utiliser ont rendu ces méthodes accessibles aux chercheurs et aux praticiens.

Ce livre s'adresse aux étudiants, aux chercheurs et aux praticiens de la recherche. Souvent, pour l'étudiant qui doit réaliser une enquête pour un travail de session ou réunir des données pour une thèse, l'analyse des données est un exercice périlleux et semé d'embûches. Pour le chercheur et le praticien, dont la connaissance de l'analyse des données n'est pas la formation première, le traitement statistique des données risque d'être superficiel ou, parfois, carrément erroné. Nous proposons ici une approche pratique et empirique qui allie l'analyse statistique à l'usage d'un logiciel statistique facile d'accès.

L'interaction entre, d'une part, le traitement informatique des données et, d'autre part, l'utilisation d'une méthode statistique et l'interprétation des résultats facilite un apprentissage opérationnel de ***l'art de l'analyse multivariée.***

CHAPITRE

1

L'analyse des données

L'analyse des données peut se définir comme l'ensemble des méthodes permettant une étude approfondie d'informations quantitatives. Selon Jean de Lagarde : « Le propre de l'analyse des données, dans son sens moderne, est justement de raisonner sur un nombre quelconque de variables, d'où le nom d'analyse multivariée qu'on lui donne souvent[1]. » Pour certains, le rôle principal de l'analyse des données est « de mettre en relief les structures pertinentes de grands ensembles de données[2] ».

La plupart des méthodes de l'analyse des données sont nées dans les années 1930[3] et certaines d'entre elles ont été élaborées bien avant. La philosophie sous-jacente à la création de ces méthodes est « que tout progrès, dans un domaine quelconque, ne peut être réalisé que grâce à des méthodes appropriées[4] ». Cette vision très instrumentale de la science, doit bien être tempérée aujourd'hui.

---

1. J. de Lagarde (1995), *Initiation à l'analyse des données*, Paris, Dunod, p. 2.
2. J.-P. Crauser, Y. Harvatopoulos et P. Sarnin (1989), *Guide pratique de l'analyse des données*, Paris, Éditions d'Organisation, p. 9.
3. Voir à ce sujet : J.-M. Bouroche et G. Saporta (1980), *L'analyse des données*, Paris, Presses universitaires de France, p. 3.
4. L. Festinger et D. Katz (1963), *Les méthodes de recherche dans les sciences sociales*, Paris, Presses universitaires de France, p. 3.

L'analyse des données, telle qu'on la connaît aujourd'hui, s'inscrit dans la convergence :

- de disciplines particulières des sciences de la gestion ou des sciences sociales ;
- des méthodes de la statistique appliquée ;
- et de l'existence de logiciels très performants de traitement des données.

Dans l'analyse des données, on distingue habituellement :

- l'analyse univariée, qui porte sur l'étude des variables prises une à une dans la présentation et l'interprétation ;
- l'analyse bivariée, qui a pour objectif d'examiner les relations de deux variables en même temps ;
- enfin, l'analyse multivariée, qui vise l'étude de plusieurs variables en même temps.

Dans ce livre, nous allons présenter seulement les méthodes de l'analyse bivariée et de l'analyse multivariée.

Ce livre se veut une présentation systématique des principales méthodes d'analyse des données. Nous nous en tiendrons donc à l'exposé de ces méthodes sans tenir compte ni de ce qui précède (la formulation d'une problématique, d'hypothèses, etc.)[5] ni de la suite (c'est-à-dire l'établissement de politiques ou de stratégies appropriées).

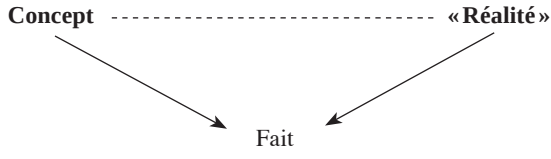
1. LES PROBLÈMES ÉPISTÉMOLOGIQUES DE L'ANALYSE DES DONNÉES

L'épistémologie se définit, au sens strict, comme un discours sur la science (ou les sciences) et, au sens plus large, comme l'étude de la production des sciences au sein des groupes et de la société globale. Le rôle de l'épistémologie est d'examiner de façon critique « les principes, les hypothèses générales, les conclusions des différentes sciences pour en apprécier la valeur et la portée objective[6] ».

5. Voir à ce sujet : J. Stafford (1996), *La recherche touristique. Introduction à la recherche quantitative par questionnaire*, Sainte-Foy, Presses de l'Université du Québec.
6. G. Durozoi et A. Roussel (1987), *Dictionnaire de philosophie*, Paris, Nathan, p. 113.

Dans cette perspective, l'analyse des données peut se définir comme un système de connaissance orienté vers l'établissement des faits dans l'ensemble social. L'objectif principal de l'analyse des données est la recherche et la définition des faits. On peut résumer cette démarche par un schéma reliant trois notions essentielles.

LA NATURE ÉPISTÉMOLOGIQUE DE LA RECHERCHE DES FAITS



Dans ce schéma, nous posons comme principe qu'il n'y a pas d'adéquation automatique entre une réalité et un concept ; la médiation entre les deux éléments du schéma se fera grâce à la notion de fait. Donnons tout de suite un exemple : le propriétaire d'un hôtel veut connaître la satisfaction de ses clients. Dans ce cas, la satisfaction n'est pas une réalité tangible, une donnée immédiate ; pour connaître la satisfaction des clients, il faudra passer par le biais d'une mesure. À la limite, on pourrait se contenter d'une seule question[7] :

Indiquez votre satisfaction face à cet hôtel (entourez le chiffre qui correspond à votre réponse) :

1	2	3	4
Très insatisfait	Insatisfait	Satisfait	Très satisfait

Ici, la « réalité » « satisfaction des clients » se résume aux réponses obtenues par cette mesure de la satisfaction.

La réponse, ou plutôt la moyenne des réponses obtenue devient un fait. La formulation de la question et la structure de l'échelle des réponses possibles (de 1 à 3, de 1 à 4, de 1 à 5 ou de 1 à *n* où : 1 = très insatisfait et *n* = très satisfait) nous amènent au cœur même des problèmes épistémologiques et méthodologiques de l'analyse des données.

7. Ce n'est pas l'idéal ; la mesure serait plus juste si l'on posait plusieurs questions (en découpant le concept de satisfaction en plusieurs parties) portant sur l'accueil, le prix, la propreté de la chambre, la salle de bain, l'éclairage, etc. L'échelle (la réponse à la question) pourrait être de 1 à 10 (et non de 1 à 4) ; il existe plusieurs façons d'établir ou de mesurer la notion de satisfaction.

1.1. LES FAITS SONT DES REPRÉSENTATIONS

Depuis Platon et Kant, on admet une relative autonomie du monde des idées ; ainsi, ce qu'on perçoit comme étant la réalité se rapporte à des idées, à des énoncés, et non aux choses elles-mêmes. La connaissance est faite de représentations, c'est-à-dire des idées, des mots, des propositions (hypothèses et conjectures), des schémas et des mesures. Une grande partie du travail scientifique porte sur la production, la structure et le contenu de ces représentations. Comme le signale Jean-Michel Besnier : « [I]l n'est pas de connaissance sans le truchement de signes pour interpréter le réel et […], par conséquent, le mécanisme de production de ces représentations et de ces signes peut seul donner les clés de la compréhension du pouvoir de l'homme de s'assimiler, ce qui n'est pas lui[8]. »

Bruno Jarroson[9] montre qu'il y a plusieurs façons de percevoir le réel ; il distingue quatre théories possibles :

1. « le réalisme fort », qui suppose que la réalité a une autonomie propre de la pensée de l'homme et qu'elle est totalement connaissable par des méthodes appropriées ;
2. « le réalisme faible », qui admet l'autonomie de la réalité, mais est très réservé sur la capacité de connaître facilement cette réalité ;
3. « le positivisme instrumentaliste », qui ne s'intéresse qu'au comment et non au pourquoi ; pour le positiviste, il s'agit de savoir comment mesurer, tout le reste appartient à la métaphysique ;
4. « le positivisme idéaliste » : pour le positiviste idéaliste, « seule existe l'idée du monde en son esprit. L'accord que je rencontre avec les autres sur les mesures que je fais n'est qu'une idée[10]. »

Nous le voyons ici, il existe plusieurs théories de la représentation ; ces théories ont une influence sur le mode de pensée des scientifiques, sur les façons de travailler et de comprendre leur rôle dans l'établissement des connaissances.

8. J.-M. Besnier (1996), *Les théories de la connaissance*, Paris, Flammarion, p. 15.
9. B. Jarrosson (1992), *Invitation à la philosophie des sciences*, Paris, Seuil, p. 128.
10. *Ibid.*, p. 131.

1.2. Les faits : la construction de la réalité

Pour qu'un fait soit reconnu comme scientifique, c'est-à-dire accepté comme étant valide dans notre culture, il faut le construire (le cerner, le mesurer, le reproduire) d'une certaine façon. Comme le souligne Gérard Fourez : « On n'observe donc pas passivement, mais on structure ce qu'on veut observer en utilisant les notions qui paraissent utiles en vue d'avoir une observation adéquate, c'est-à-dire qui répond au projet que l'on a. Et c'est alors qu'on dit qu'on observe des "faits"[11]. »

De façon pratique, la connaissance se bâtit dans une forme de bricolage sophistiqué entre les idées-propositions et les idées-mesures, dans un va-et-vient qui est rarement transparent. Ainsi : « La leçon est générale : un énoncé se produit en même temps que l'objet qu'il qualifie, et sa production s'instrumentalise dans toute une série d'opérations qui font parler l'objet de connaissance et le contraignent à reconnaître qu'il est réellement ce que l'énoncé dit qu'il est[12]. »

Dans la plupart des sciences sociales, le travail de la connaissance se fait surtout à partir de l'étude des relations entre plusieurs variables. Les liens entre la problématique et la mesure sont formalisés à l'aide d'un modèle. Le modèle est une représentation plus ou moins structurée de la réalité observée ; il peut être qualitatif ou quantitatif. La notion de modèle est l'élément commun de la plupart des disciplines scientifiques. Selon Bernard Walliser et Charles Prou : « À partir de ces matériaux [les données] et de considérations théoriques, toutes les disciplines construisent des modèles empiriques sous forme de simples récits permettant une lecture cohérente d'un phénomène observé ou d'une évolution passée, ou de systèmes d'équations permettant de simuler un processus concret[13]. »

1.3. Le caractère inachevé de la représentation

Nous l'avons vu, la connaissance est une construction, mais c'est une construction inachevée ! La représentation de la réalité n'est jamais totale, jamais définitive : ce serait la fin de toute science. Qu'on le veuille ou

---

11. G. Fourez (1992), *La construction des sciences*, Montréal, De Boeck-ERPI, p. 32.
12. M. Callon et B. La Tour (1991), *La science telle qu'elle se fait*, Paris, La Découverte, p. 17.
13. B. Walliser et C. Prou (1988), *La science économique*, Paris, Seuil, p. 62.

non, toute représentation, aussi scientifique soit-elle, comporte des zones d'ombre, des obscurités, un non-dit qui démontrent son caractère évanescent, temporaire. Selon Raymond Boudon : « Toute théorie produite par la connaissance ordinaire comme par la connaissance scientifique est entourée d'un halo d'*a priori* relevant de divers types : logiques, mais aussi épistémologiques, ontologiques, linguistiques, etc.[14]. »

Le caractère inachevé de la science nous amène à plus de modestie ; il ne s'agit pas de rejeter les savoirs, mais « de prendre acte du caractère inéluctablement parcellaire, fragile et contingent de toute science et de ses rapports avec les pratiques éventuelles qu'elle fonde[15] ».

La fragilité des représentations, il faut faire bien attention, ne concerne que la science en train de se faire à travers une recherche spécifique. Bruno Latour fait une distinction entre la « science faite » et la recherche ; la science est sûre, objective, froide ; un fait, c'est ce qu'on ne discute pas. La recherche, la science en train de se faire, est incertaine, risquée, subjective, chaude ; un fait est déjà construit[16]. Quand la recherche est terminée, le plus souvent, la science devient action sur le monde : ordinateur ou avion, médicament ou politique économique[17].

1.4. La validité des représentations scientifiques

La représentation n'épuise pas le réel ; il y a toujours un écart entre le modèle formulé et la réalité elle-même. L'oubli de cet écart « conduit à l'illusion ontologique de l'unité, de l'identité, de la stabilité et de la permanence du sens[18] ». L'ampleur et la forme de cet écart sont une mesure de la validité.

On peut évaluer la validité des représentations par une approche générale et une approche instrumentale, la première approche étant une condition de la deuxième. L'approche générale repose sur « le point de vue falsificationniste » de Karl Popper : « Les théories scientifiques se distinguent des mythes uniquement par ceci qu'elles sont critiquables et

14. R. Boudon (1990), *L'art de se persuader*, Paris, Fayard, p. 272.
15. J.-M. Lévy-Leblond (1996), *La pierre de touche. La science à l'épreuve…*, Paris, Gallimard, p. 50.
16. B. Latour (1995), *Le métier de chercheur : regard d'un anthropologue*, Paris, Éditions de l'INRA, p. 11-14.
17. Voir à ce sujet : N. Journet (1996), « Comment peut-on être relativiste ? », *Sciences Humaines*, n° 67.
18. F. Laplantine (1996), *La description ethnographique*, Paris, Nathan, p. 35.

modifiables, à la lumière de la critique. Mais elles ne peuvent être vérifiées ni rendues probables[19]. » Pour Raymond Boudon, cette vision de la science a un côté tragique, car « elle implique en effet que toute théorie scientifique est condamnée d'avance, que l'histoire de la science est celle d'une suite d'échecs se corrigeant les uns les autres[20] ».

Habituellement, on tient compte de la validité externe et de la validité interne des modèles[21]. La validité externe se rapporte aux liens entre les concepts choisis et le problème étudié ; la plupart du temps, il y a une pléthore de problèmes et peu de concepts réellement opérationnels (mesurables). Pour certains, la seule règle de validité est de suivre les procédures usuelles de la recherche, décrites dans tous les manuels.

La validité interne se concentre le plus souvent dans l'analyse des écarts entre les résultats du modèle et la réalité observée. La validité interne s'évalue à partir de l'étude méticuleuse des résidus[22]. On peut aussi aborder le problème de la validité interne par le biais de la cohérence formelle de la représentation présentée (refus des contradictions), par sa simplicité ou son caractère esthétique. Ainsi : « On a expliqué quelque chose lorsqu'on parvient à relier, dans un discours cohérent, la représentation qu'on s'est donnée d'un phénomène aux autres représentations que l'on a et auxquelles on tient[23]. »

Il n'est pas facile de vérifier la validité des représentations formulées ; dans la plupart des cas, cette question est abordée, soit de façon nébuleuse, en fonction des diverses théories (paradigmes) qui portent sur la teneur de la réalité étudiée, soit de façon strictement instrumentale, par l'utilisation de tests statistiques, sans tenir vraiment compte du contexte historique et social. Dans un certain sens, le « vrai » est plus un besoin humain qu'une notion démontrable.

---

19. K. Popper (1990), *Le réalisme et la science*, Paris, Hermann, p. 26.
20. R. Boudon (1995), *Le juste et le vrai. Études sur l'objectivité des valeurs et de la connaissance*, Paris, Fayard, p. 520.
21. Voir à ce sujet : E. Carmines et A. Zeller (1979), *Reliability and Validity Assessment*, Newbury Park, Sage University Paper, nº 17.
22. Voir l'élaboration des tests de validité interne des modèles par Pearson, Fisher et Gosset : A. Desrosières (1993), *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte, p. 352-355.
23. G. Fourez (1992), *La construction des sciences*, Montréal, De Boeck-ERPI, p. 221.

1.5. LA PLACE DE L'ANALYSE DES DONNÉES DANS LE PROCESSUS DE RECHERCHE

L'analyse des données correspond à certaines étapes bien spéciales du processus de recherche. Voici par exemple les principales étapes d'une recherche faite à l'aide d'un questionnaire fermé (dans ce type de questionnaire, les réponses sont définies à l'avance)[24] :

1. Problème de base
2. Questions posées (problématique)
3. Hypothèses
4. Conceptualisation
5. Spécification des concepts
6. Choix des indicateurs
7. Formation des indices (questions)
8. Prétest
9. Rédaction finale du questionnaire
10. Échantillon
11. Collecte des données
12. Codification des réponses
13. Traitement informatique des données
14. Analyse des données
15. Interprétation des données
16. Rédaction du rapport de recherche

Dans cette longue liste, les étapes 1 à 3 servent à bien cerner le problème étudié. Les étapes 4 à 10 ont pour objectif de construire les principaux outils qui seront utiles pour l'enquête. L'étape 11 correspond au travail « sur le terrain » ; c'est une étape strictement empirique. Les étapes 12 à 15 résument ce que l'on appelle habituellement l'analyse des données.

La codification des réponses sert à établir des catégories, donc des mesures selon les variables utilisées. Le traitement informatique des données permet la saisie numérique des informations et, par la suite, la création des tableaux et l'étude des relations entre les principales

24. Voir à ce sujet : U. Sekaran (1992), *Research Methods for Business*, New York, Wiley, chapitre 4.

variables de l'enquête. L'analyse des données proprement dite regroupe surtout un ensemble de méthodes statistiques susceptibles de « faire parler » les données. Enfin, l'interprétation des données nous ramène au problème du « sens », du caractère explicatif des informations présentées. L'interprétation des données est aussi liée aux aspects « utiles » de ces informations.

Le rapport de recherche est du domaine de l'écriture ; comme l'écrit Michel Volle : « [L]e style mathématique est le meilleur pour décrire un édifice logique achevé, mais le style littéraire est le seul capable d'expliquer les choix sur lequel cet édifice est fondé, ou de présenter une pensée en évolution, dont la communication peut avoir un intérêt même si elle n'est pas formalisée[25]. »

2. LES TYPES DE DONNÉES

Les faits comme représentation, comme construction sociale, prennent des formes diverses. On distingue tout d'abord les données primaires des données secondaires. Habituellement, les données primaires sont construites par le chercheur dans un but bien précis ; par exemple, si l'on fait une recherche sur la satisfaction des usagers face à un produit ou un service, les questions posées se rapporteront directement à ce produit ou à ce service. Les enquêtes qualitatives ou quantitatives réalisées à l'aide de sondages aléatoires ou non produisent des données primaires. Dans tous les cas, c'est le chercheur ou l'équipe de recherche qui décide de la forme que prendront les variables.

Les données secondaires sont des données recueillies par des gouvernements[26] ou des organismes officiels internationaux ou nationaux[27]. Elles découlent de décisions politiques et administratives prises à un haut niveau. Les principaux objectifs visés sont d'avoir des données objectives et comparables d'un lieu à un autre : ce sont l'indice du chômage, l'indice des prix à la consommation, les indices qui touchent les secteurs de la santé et de l'éducation, etc.

---

25. M. Volle (1980), *Le métier de statisticien*, Paris, Hachette, p. 226.
26. L'Institut de la statistique du Québec, Statistique Canada, l'INSEE pour la France, etc.
27. L'Organisation de coopération et de développement économiques (OCDE), l'Organisation mondiale du tourisme (OMT) ; ces organismes peuvent aussi avoir un caractère privé.

En théorie, les données secondaires sont produites à des fins de gestion sociale, du bien commun ; elles jouent aussi, en même temps, un rôle de contrôle social ; elles ont enfin une coloration idéologique et politique (les données primaires aussi, d'ailleurs). La naissance des statistiques officielles est fortement liée à la construction et à la consolidation des États modernes[28] ; elles affichent une neutralité factice qui laisse souvent dans l'ombre les mécanismes réels de leur construction.

Les données primaires et secondaires résultent les unes comme les autres d'un processus de recherche, mais elles diffèrent dans l'organisation même de ce processus. Les données primaires dépendent en très grande partie de l'équipe de recherche ; les données secondaires résultent d'un cheminement bureaucratique et politique. Les données primaires tirent leur légitimité de l'autorité scientifique d'un ou plusieurs chercheurs ; les données secondaires sont aussi produites par des chercheurs patentés, mais elles bénéficient, en plus, de l'appui du système étatique, donc de l'autorité légalement constituée.

Très souvent, les données primaires et secondaires sont complémentaires : le taux de chômage peut s'expliquer, en partie par la situation économique, mais aussi par des dimensions démographiques, sociales, psychologiques et politiques. Une véritable étude de marché doit tenir compte non seulement de la demande d'un bien révélée par un sondage, mais aussi de la situation économique générale, du revenu disponible, etc.

2.1. La notion de variable

Avant d'aborder les types de données, il faut définir au préalable la notion de variable, qui joue un rôle central dans toutes les recherches en sciences sociales et en sciences de la gestion. Au plan strictement sémantique, le terme « variable » suppose qu'une réponse à une question donnée peut varier (dans un certain écart) d'un individu à un autre. Donc : « Si la caractéristique mesurée peut prendre différentes valeurs, on dit alors que cette caractéristique est une variable[29]. »

---

28. Voir à ce sujet : E. Brian (1998), « Du bon observateur au statisticien d'État », *Les cahiers de Science et vie*, n° 48.
29. N. Lemieux, G. Roy et J.-G. Savard (1991), *Méthodes quantitatives*, Laval, Études Vivantes, p. 13.

Au plan mathématique, une variable est perçue comme un ensemble de règles qui permettent de ranger les éléments d'un ensemble donné dans des catégories définies au départ. À partir de cette définition, toute variable est comprise dans le sens d'une norme de classification[30]. Une variable est donc un critère par lequel on classe des individus dans des catégories. Par exemple, si on demande dans un sondage : quelle était la destination de votre dernier voyage ? La réponse pourrait être :

1	*2*	*3*	*4*
1. Au Québec 2. Ailleurs au Canada 3. Aux États-Unis 4. En Europe 5. Autres destinations	1. Au Québec 2. En Ontario 3. Dans les provinces de l'Est 4. Dans les provinces de l'Ouest 5. En Colombie-Britannique 6. Aux États-Unis 7. En Europe 8. Autres destinations	1. Canada 2. États-Unis 3. Amérique du Sud 4. Europe 5. Asie 6. Afrique 7. Océanie	Autres catégories

On voit dans cet exemple que le chercheur doit décider de fermer (colonne n° 1) ou d'ouvrir (colonne 3) l'éventail possible des réponses. Dans le premier cas, on trouvera à peu près 80 % des répondants dans les catégories 1-2-3. Dans le deuxième cas, les catégories 2-3-4-5 regrouperont très peu de répondants. Dans le troisième cas, les catégories 3-5-6-7 ne compteront que très peu de répondants. Il y a donc un choix à faire pour se rapprocher de la « réalité » et maximaliser l'utilisation de chacune des catégories utilisées.

2.2. Les types de variables

On classifie les variables selon leur degré d'abstraction et leur pouvoir explicatif. Le tableau 1.1 présente ces types de variables.

30. Voir à ce sujet : V. Papillon et R. Turcotte (1981), *Probabilités et statistique*, Montréal, Modulo, p. 1-2.

Tableau 1.1

LES TYPES DE VARIABLES SELON LEUR DEGRÉ D'ABSTRACTION ET LEUR POUVOIR EXPLICATIF

Types de variables	*Degré d'abstraction*	*Pouvoir explicatif*
Variables factuelles concernant : • la personne • son environnement • ses comportements	Faible	Faible
Les variables reliées aux opinions de la personne	Moyen	Moyen
Les variables reliées aux attitudes de la personne	Élevé	Élevé

Il y a donc ici deux axes : le continuum abstrait/concret et le continuum explicatif/descriptif. Le couple abstrait/concret renvoie aux difficultés de définir la ou les variables au plan méthodologique ; une variable abstraite suppose la médiation d'un appareillage plus complexe dans l'observation de la réalité. Le continuum explicatif/descriptif nous ramène aux étapes de la problématique de la recherche et de la formulation des hypothèses ; la question est : quelles variables expliquent le mieux le problème étudié ? Nous verrons des exemples plus loin.

2.2.1. Les variables factuelles

Essayons de définir un peu mieux les variables de notre typologie. Les variables factuelles sont celles que l'on qualifie de « faits » dans la vie quotidienne en ce sens qu'elles nous semblent indiscutables. On distingue les variables factuelles qui identifient :

- la personne ;
- son environnement ;
- ses comportements économiques, politiques, sociologiques et psychologiques.

Les variables factuelles qui identifient la personne sont, par exemple :

- l'âge ;
- le sexe ;
- le niveau de scolarité ;
- la profession ;
- le revenu.

Les variables factuelles reliées à l'environnement abordent, le plus souvent :

- la situation familiale ;
- l'emploi ;
- les conditions de travail ;
- l'habitat de la personne (locataire ou propriétaire), le lieu de l'habitation, le type de maison ou d'appartement, etc. ;
- aussi, cela le cas, les relations familiales, de travail, de voisinage, etc.

Les variables factuelles qui concernent les comportements peuvent être très diversifiées :

- les comportements économiques, en particulier ceux qui se rattachent aux objets de consommation courante (alimentation, vêtements, transport, etc.) ;
- les comportements politiques (vote, contribution/participation à des partis, etc.) ;
- les comportements sociologiques et psychologiques (loisirs, tourisme, sport, participation à des clubs, groupes communautaires, etc.).

Ici, nous parlons de comportements observables, mesurables directement, et nous excluons les idéologies, les intérêts et les valeurs ; ces derniers éléments font partie soit des opinions, soit des attitudes.

Pour les non-spécialistes, les variables factuelles semblent tangibles et concrètes, d'accès facile à peu de frais ; il s'agit là, bien sûr, d'une illusion. Par exemple, l'âge paraît être un fait brut sur lequel il est difficile d'épiloguer ; pourtant comme le signale Rémi Lenoir : « Si l'âge de l'état civil et les divisions qu'il rend possibles sont des notions sociales, les catégories qu'il permet de distinguer ne forment pas pour autant des groupes sociaux[31]. » Dans les enquêtes par questionnaire, les personnes âgées sont définies par la catégorie d'âge des 65 ans et plus ; il s'agit d'une convention juridique qui définit les ayants droit à une pension de retraite. La catégorie d'âge pourrait bien être de 60 ans et plus ou 70 ans et plus.

---

31. R. Lenoir (1990), « Objet sociologique et problème social », dans P. Champagne, R. Lenoir, D. Merllié et L. Pinto, *Initiation à la pratique sociologique*, Paris, Dunod, p. 61.

On se rend compte que les découpages que l'on retrouve dans les variables factuelles sont la plupart du temps des conventions très commodes qui « structurent » la réalité et lui donnent une consistance qu'elle n'a pas réellement. Les mêmes difficultés se retrouvent dans la définition des catégories socioprofessionnelles. Une catégorie telle que « cadres supérieurs » peut cacher des situations très différentes selon la taille de l'entreprise. « La variabilité de l'instrument de mesure n'est donc pas liée seulement à celle des conditions techniques de sa mise en œuvre dans des enquêtes, mais elle est fonction également des objets auxquels on l'applique[32]. »

Les variables factuelles ne sont pas des faits bruts ; ce sont des représentations souvent fragiles d'une réalité mouvante. Elles nous donnent un aperçu de cette réalité ; elles ne peuvent être ni exhaustives, ni définitives. Il faut les accepter comme telles et vivre dans un certain flou définitionnel qui est le propre de la condition humaine.

2.2.2. Les opinions

Les opinions et les attitudes sont des variables abstraites qu'on ne peut saisir directement. Elles représentent des concepts qui exigent, le plus souvent, une longue élaboration théorique et méthodologique[33]. Guy Serraf définit l'opinion de la façon suivante : « L'opinion est une expression manifeste, en termes d'adhésion ou de refus, vis-à-vis d'une proposition socialement soutenue par un groupe ou une fraction du public, dans le contexte des courants culturels d'une société à un moment donné dans un certain milieu[34]. » Plus simplement, une opinion correspond à ce que pense une personne sur un sujet, un bien ou un service quelconque.

L'opinion a un lien plus direct avec l'action que l'attitude (ce que nous verrons plus loin). Selon Roger Mucchielli : « Il ne faut pas sous-estimer ce que pensent les gens, ce que croient les autres. C'est en le sachant qu'on peut comprendre ce qu'ils font. C'est en comprenant ce qu'on leur fait croire que nous comprenons une part importante de leurs

32. D. Merllié (1990), « La construction statistique », dans P. Champagne, R. Lenoir, D. Merllié et L. Pinto, *Initiation à la pratique sociologique*, Paris, Dunod, p. 126.
33. Voir à ce sujet : L.-R. Baker (1995), *Explaining Attitudes*, Cambridge, Cambridge University Press.
34. G. Serraf (1985), *Dictionnaire méthodologique du marketing*, Paris, Éditions d'Organisation, p. 169.

comportements[35]. » La connaissance des opinions peut donc permettre d'expliquer et de prévoir certains comportements.

Certains contestent la réalité d'une « opinion publique » ; pour Jean-Louis Besson « le premier effet d'un sondage d'opinion est de postuler l'existence d'une opinion publique et de valider ce postulat[36] ». Pierre Bourdieu montre bien la fausse neutralité de certains sondages d'opinion qui ne font que traduire le vide de certaines institutions publiques[37]. Il faut voir la notion d'opinion comme un concept utile et provisoire qui nous permet de cerner quelque peu des réalités souvent fuyantes.

2.2.3. Les attitudes

Les attitudes sont plus abstraites et complexes que les opinions : « [L'attitude] signifie une disposition interne de l'individu sous-tendant sa perception et ses réactions vis-à-vis d'un objet ou d'une simulation[38]. » Dans l'attitude profonde, on retrouve les motivations et les principaux éléments de la personnalité de l'individu. Les attitudes sont le fruit de la socialisation dans la famille et le groupe social. Elles ont une certaine permanence et sont orientées soit positivement, soit négativement.

Dans l'axe abstrait/concret et l'axe descriptif/explicatif, les attitudes apparaissent comme des variables latentes qui influencent les comportements. La figure 1.1 montre la place des variables selon les axes.

Habituellement, au plan empirique, les liens entre les attitudes, les opinions et les comportements suivent la séquence suivante[39] :

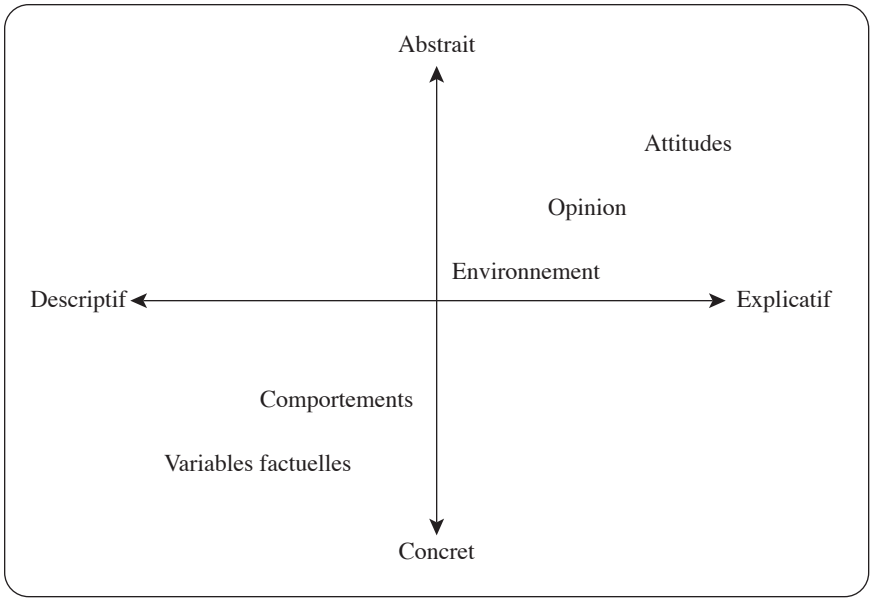
---

35. R. Mucchielli (1972), *Opinions et changement d'opinions*, Paris, Entreprise moderne d'édition, p. 5.
36. J.-L. Besson (1992), « Les statistiques : vraies ou fausses », *Autrement*, nº 5, p. 27.
37. Voir à ce sujet : P. Bourdieu (1972), « Les Doxosophes », *Minuit*, Éditions de Minuit, nº 1.
38. C. Tapia et P. Roussay (1991), *Les attitudes*, Paris, Éditions d'Organisation, p. 15.
39. Voir à ce sujet : A. Eagly et S. Chaiken (1995), *The Psychology of Attitudes*, New York, Harcourt Brace Jovanovich ; A. Channouf, J. Py et A. Somat (1996), « Prédire des comportements à partir des attitudes : nouvelles perspectives », dans J.-C. Deschamps et J.-L. Beauvois, *Des attitudes aux attributions*, Grenoble, Presses universitaires de Grenoble.

Une attitude…	***qui s'exprime dans une opinion…***	***qui s'actualise dans un comportement.***
La personne aime les choses sucrées.	La personne aime le chocolat.	La personne achète/consomme la marque de chocolat X.
La personne aime les choses sucrées.	La personne n'aime pas le chocolat, mais aime les caramels.	La personne achète/consomme la marque de caramels X.

Figure 1.1

LES TYPES DE VARIABLES SELON L'AXE ABSTRAIT/CONCRET ET L'AXE DESCRIPTIF/EXPLICATIF



Si les attitudes sont relativement stables, les liens avec les opinions exprimées et les comportements effectifs peuvent bien ne pas l'être. La personne peut refuser d'exprimer son opinion ou différer ses comportements pour des raisons économiques, sociologiques, psychologiques ou même politiques.

40. P. De Baty (1967), *La mesure des attitudes*, Paris, Presses universitaires de France, p. 13.

2.2.4 La mesure des opinions et des attitudes

La mesure des opinions et des attitudes vise à connaître :

- « la direction de l'attitude » : l'opinion ou l'attitude peut être positive ou négative ;
- « l'intensité de l'attitude[40] » : l'opinion ou l'attitude va être forte ou faible, s'exprimer à divers degrés.

La mesure des opinions et des attitudes se fait par le biais d'items ou de propositions ; ces items et propositions représentent les questions posées. Les réponses sont retenues sous forme d'échelles où la direction et l'intensité sont mesurées en même temps.

Exemple de proposition

Indiquez votre accord ou votre désaccord face à la proposition suivante :

LES HOMMES SONT PAREILS AUX FEMMES

Encerclez le chiffre qui correspond à votre réponse

Entièrement en désaccord	Partiellement en désaccord	Partiellement en accord	Totalement en accord
1	2	3	4

Le chiffre 1 indique que l'on ne croit pas aux différences homme/femme ; le chiffre 4 indique que l'on croit qu'il y a des différences entre les sexes.

L'échelle peut avoir de 2 à 10 degrés selon le problème étudié et la taille de l'échantillon ; ce dernier élément est la contrainte la plus importante. Les items et les propositions peuvent se mesurer par d'autres qualificatifs que « Totalement en désaccord / Totalement d'accord » ; cela peut-être :

Pas du tout intéressé / Très intéressé ;

Très insatisfait / Très satisfait, etc.

La meilleure technique consiste à employer la même expression pour la proposition (ou l'item) et pour la réponse (l'échelle)[41].

---

41. Voir à ce sujet : M. Henerson, L. Morris et C. Fitz-Gibbon (1987), *How to Measure Attitudes*, Beverly Hills, Sage.

2.2.5 Attitudes, opinions et comportements

Pour différencier ces trois notions, donnons trois exemples de questions :

A- C'est très important d'être bien habillé

Pas du tout important	Peu important	Important	Très important
1	2	3	4

B- C'est très important de porter des vêtements d'une marque reconnue (Lacoste, Yves St-Laurent, etc.)

Pas du tout important	Peu important	Important	Très important
1	2	3	4

C- Dans la dernière année, avez-vous acheté un ou plusieurs vêtements portant la marque suivante ?

1- Lacoste	1- Oui	2- Non
2- Yves St-Laurent	1- Oui	2- Non
3- _____	1- Oui	2- Non

Dans l'exemple A, nous avons une proposition qui porte sur l'attitude ; celle-ci est plus vague que l'opinion et a un faible ancrage dans la réalité. L'exemple B présente une proposition portant sur une opinion ; il y a ici un fort ancrage dans la réalité – celle-ci est représentée par des marques prestigieuses. L'exemple C est une question comportementale ; il s'agit de bien saisir le comportement effectif de la personne en ce qui concerne la consommation de vêtement de luxe.

On pourrait s'attendre à ce qu'une personne en accord avec la proposition A, en accord avec la proposition B, réponde oui aux questions C_1, C_2, …, C_n. Mais ce n'est pas si simple, car il y a souvent un hiatus, un décalage entre ce que la personne désire profondément, la façon dont elle interprète ce désir, et la concrétisation (ici dans l'achat) de celui-ci.

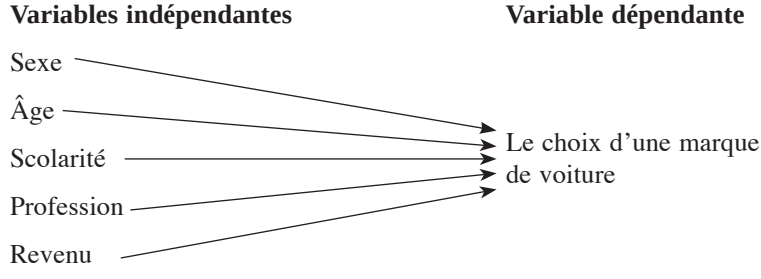
2.2.6 La chaîne causale

Le terme de cause est une notion bien délicate à manier dans les sciences sociales et les sciences de la gestion. Il s'agit bien sûr d'une causalité limitée ; la principale contrainte vient de notre capacité de représenter l'objet à étudier, donc de le mesurer. L'analyse des données tient compte des variables indépendantes et des variables dépendantes.

La variable indépendante représente habituellement un facteur causal, un élément qui semble déterminant dans l'explication d'un phénomène. Dans certains cas, la variable indépendante est celle qui survient en premier lieu, dans le temps ; il y a donc une certaine antériorité. La variable indépendante comporte des aspects historiques et logiques.

La variable dépendante est celle qui est influencée, celle dont on observe les variations. Voici une définition plus spécifique de la variable dépendante : « Toute grandeur dont le niveau peut être connu lorsqu'on connaît le niveau atteint par une autre grandeur dite variable indépendante. On dit alors que la variable dépendante est fonction de la variable indépendante[42]. » Les variables dépendantes sont les variables manipulées, expliquées par les autres variables ; elles représentent ce que l'on cherche à savoir.

En règle générale, dans les enquêtes par questionnaire fermé, les variables indépendantes sont les principales variables socioéconomiques. Par exemple, l'achat d'une voiture (la marque, le prix) varie en fonction de variables indépendantes bien connues :



Ici, les variables sexe, âge, scolarité, profession et revenu sont les grandes variables de segmentation qui permettent de comprendre le comportement des acheteurs.

À l'aide de tests, le chercheur vérifie si les relations entre les variables indépendantes et la variable dépendante sont significatives au plan statistique. Il arrive que les relations observées entre les variables indépendantes usuelles (les variables socioéconomiques) ne soient pas significatives. Si tel est le cas, à ce moment de l'enquête, il faudra, si nécessaire, puiser dans le vivier des autres variables de l'enquête, si elles existent.

---

42. J.-J. Pinty et C. Gaultier (1971), *Dictionnaire pratique de mathématiques et statistiques en sciences humaines*, Paris, Éditions Universitaires, p. 290.

À cette étape, l'analyse causale se fait en tenant compte des opinions formulées :

Variable indépendante		**Variable dépendante**
Les opinions des répondants face au prix et à la fiabilité d'une voiture	→	Le choix d'une marque de voiture

ou des attitudes face à l'automobile :

Variable indépendante		**Variable dépendante**
Attitude des répondants face à l'automobile	→	Le choix d'une marque de voiture

Nous pouvons rendre l'analyse plus complexe en essayant de cerner les motivations profondes des répondants :

Variable indépendante		**Variable dépendante**
Attitude des répondants face à l'automobile	→	Les opinions des répondants face au prix et à la fiabilité d'une voiture

On le voit ici, ce qui était auparavant une variable indépendante (les opinions) peut devenir, pour les besoins de l'analyse, une variable dépendante.

Dans la recherche d'une explication plus approfondie des données, la désignation des variables indépendantes et dépendantes devient toute relative. À ce moment de l'analyse, il faut faire un retour en arrière vers les objectifs de départ et les hypothèses de la recherche.

L'explication de beaucoup de comportements économiques et sociaux a sa source dans les motivations et les attitudes profondes des répondants. Malheureusement, la recherche des attitudes est un domaine complexe et mouvant[43] qui exige de très bonnes connaissances en méthodologie et en psychosociologie et une longue expérience des enquêtes sur le terrain.

43. Voir à ce sujet : S. Oskamp (1991), *Attitudes and Opinions*, Englewood Cliffs, Prentice Hall.

Ces quelques exemples montrent que, dans l'analyse des données, « il n'existe pas de connaissance absolue, vraie en soi, indépendante de la manière de l'acquérir[44] ». Les résultats obtenus par l'analyse des données n'existent qu'en fonction des méthodes utilisées.

3. LES ÉCHELLES DE MESURE

Les niveaux de mesure sont le parachèvement de la problématique et de la conceptualisation du problème de la recherche. Il s'agit en définitive de faire correspondre un concept à une mesure ; c'est dans cette opération que la démarche de recherche devient empirique. Mesurer, c'est relier des nombres à des entités plus ou moins abstraites : l'âge, le sexe, la satisfaction, l'intérêt…

L'analyse des données est basée en grande partie sur les principes des mathématiques et plus particulièrement de la statistique appliquée. Les nombres possèdent certaines propriétés mathématiques dont il faut tenir compte. Ces propriétés sont les suivantes :

1. la propriété de classer des individus dans des catégories ;
2. la propriété d'établir un ordre de préséance, un ordre hiérarchique entre ces catégories ;
3. la propriété de fixer des intervalles égaux dans cet ordre hiérarchique construit en fonction de la deuxième propriété ;
4. la propriété de fixer une origine 0 à cet ordre hiérarchique (en plus d'avoir des intervalles égaux).

3.1. LES ÉCHELLES DE MESURE ET LES PROPRIÉTÉS DES NOMBRES

Les variables utilisées dans les recherches en sciences sociales et en sciences de la gestion possèdent une ou plusieurs de ces propriétés. Nous présentons ces échelles avec leurs propriétés dans le tableau 1.2 :

44. G. Fourez (1992), *La construction des sciences*, Montréal, De Boeck-ERPI, p. 168.

Tableau 1.2

LES ÉCHELLES DE MESURE ET LES PROPRIÉTÉS DES NOMBRES

Échelles	***Propriétés***			
	Classement	Ordre	Distance	Zéro absolu
Nominale	Oui	Non	Non	Non
Ordinale	Oui	Oui	Non	Non
Intervalles	Oui	Oui	Oui	Non
Rapport	Oui	Oui	Oui	Oui

Nous voyons dans le tableau 1.2 :

- que l'échelle nominale permet de classer les individus dans des catégories ;
- que l'échelle ordinale permet de classer les individus dans des catégories et, en plus, d'établir un ordre hiérarchique entre ces catégories ;
- que l'échelle par intervalles possède les propriétés des deux premières échelles ; en plus, les intervalles de l'échelle sont égaux ;
- que l'échelle de rapport possède toutes les propriétés des nombres ; c'est donc la plus achevée des mesures.

Voyons tout de suite des exemples de chacune de ces échelles.

3.1.1. L'échelle nominale

L'échelle nominale a pour principale propriété de classer les individus d'un ensemble donné (population ou échantillon) dans des catégories données. Donnons des exemples :

- Le sexe des personnes se répartit comme suit :
  1. Femme
  2. Homme
- La destination des dernières vacances de quatre jours et plus hors du domicile habituel :
  1. Québec
  2. Canada (à l'exception du Québec)
  3. États-Unis
  4. Europe
  5. Autre

Dans le premier exemple, la population étudiée se divise en deux catégories ; dans le deuxième exemple, cette même population se découpe en cinq parties.

Les catégories nominales reposent, la plupart du temps, sur des conventions culturelles ; en ce sens, dans le deuxième exemple, le Québec pourrait s'appeler X1 et le Canada X2, et ainsi de suite, sans que cela change grand-chose au classement initial des personnes. Il s'agit bien sûr d'une mesure rudimentaire ; c'est le plus faible niveau de mesure accessible. L'échelle nominale consiste en fait à énumérer les possibilités et à classer les individus selon ces possibilités.

Le classement des individus dans des catégories doit répondre à des règles assez strictes :

1. les catégories doivent être exhaustives, c'est-à-dire tenir compte de toutes les possibilités (ou du moins des principales) ;
2. les catégories doivent être mutuellement exclusives en ce sens qu'une personne ne peut être classée à la fois dans deux catégories (ou plus) ;
3. les individus de la population étudiée doivent être classés dans les catégories avec le minimum d'erreur possible.

Ces règles incontournables s'appliquent à toutes les échelles de mesure.

3.1.2. L'échelle ordinale

Dans l'échelle nominale, chacune des catégories de la variable est équivalente aux autres ; dans le cas de l'échelle ordinale, une catégorie peut être plus petite ou plus grande qu'une autre : il y a une gradation dans les catégories utilisées. Voici des exemples :

- La satisfaction face à un service :
  1. Très insatisfait
  2. Insatisfait
  3. Satisfait
  4. Très satisfait

- L'utilité d'un produit :
  1. Inutile
  2. Peu utile
  3. Utile
  4. Très utile
- L'achat d'un bien de consommation :
  1. Jamais
  2. Rarement
  3. Souvent
  4. Très souvent
- Le niveau de scolarité :
  1. Primaire
  2. Secondaire
  3. Collégial
  4. Universitaire

Dans tous les cas présentés, on remarque que 4 est plus grand que 3, 3 est plus grand que 2 et 2 est plus grand que 1 ; il y a donc une relation d'ordre qui est transitive. Si ce postulat hiérarchique est reconnu, il s'agit bel et bien d'une échelle ordinale. L'échelle ordinale possède donc deux des principales propriétés des nombres : classer les individus dans des catégories et établir un ordre valable entre ces catégories – deux opérations naturellement simultanées.

3.1.3. L'échelle par intervalles

L'échelle par intervalles possède les propriétés des échelles nominales et ordinales, auxquelles elle ajoute des intervalles égaux dans les différents niveaux gradués de l'échelle de mesure. Donnons des exemples :

- Le revenu du ménage :
  1. 20 000 $ et moins
  2. 20 001 $ à 40 000 $
  3. 40 001 $ à 60 000 $
  4. 60 001 $ à 80 000 $
  5. 80 001 $ à 100 000 $
  6. 100 001 $ et plus

- La scolarité :
  1. 7 années et moins
  2. 8 à 14 années
  3. 15 à 21 années
  4. 22 années et plus

Il s'agit ici d'une façon plus abstraite de mesurer la scolarité, car elle ne tient pas compte des niveaux scolaires habituels.

Au plan pratique, il est rare que les « barreaux » inférieurs et supérieurs de l'échelle par intervalles soient réellement égaux ; nous avons affaire, la plupart du temps, en sciences sociales et en sciences de la gestion, à des échelles à intervalles quasi égaux. Cette légère entorse aux propriétés des nombres n'invalide pas nécessairement ce type d'échelle, qui a tout de même des qualités mathématiques supérieures à celles des échelles ordinales.

3.1.4. L'échelle de rapport

L'échelle de rapport possède les mêmes propriétés des nombres que les trois premières échelles ; s'ajoutent à ces propriétés les éléments suivants :

- le zéro dans l'échelle est absolu et a un sens, le sens d'absence de quelque chose ;
- les proportions calculées, dans l'échelle même, ont aussi un sens quelconque.

Donnons des exemples de cette fameuse échelle :

- Les dépenses alimentaires du ménage par semaine :
  1. 0
  2. 1 $ à 50 $
  3. 51 $ à 100 $
  4. 101 $ à 150 $
  5. 151 $ à 200 $
  6. 201 $ à 250 $
  7. 251 $ à 300 $
  8. etc.

Ici, il semble impossible que le ménage dépense 0 $ pour se nourrir ; l'échelle est bien construite au plan technique, mais elle n'a aucun sens aux plans économique et sociologique.

- L'âge du répondant à une enquête sur les opinions politiques :
  1. 0
  2. 1 an à 10 ans
  3. 11 ans à 20 ans
  4. 21 ans à 30 ans
  5. 31 ans à 40 ans
  6. 41 ans à 50 ans
  7. 51 ans à 60 ans
  8. 61 ans à 70 ans
  9. etc.

Dans cet exemple, on peut se poser des questions sur la valeur des catégories 1, 2 et 3 au sujet de l'opinion politique des enfants ! Au niveau des proportions, on peut affirmer qu'une personne de 40 ans a l'équivalent de 2 × 20 ans, mais cette expression n'a aucun sens aux plans psychologique et sociologique.

Une personne qui n'a ni revenu, ni âge et qui ne consomme pas de biens alimentaires, cela n'a pas beaucoup de sens. C'est pour cela qu'une échelle de rapport est si difficile à construire dans les sciences sociales et les sciences de la gestion ! Le zéro absolu est une denrée rare dans ces disciplines. Toute personne, même la plus démunie, possède un certain degré de revenu, d'intelligence, de satisfaction à l'égard d'un bien ou service, d'intérêt pour la politique, etc.

3.2. LES ÉCHELLES DE MESURE ET LES OPÉRATIONS STATISTIQUES

Nous avons vu que les échelles de mesure ne possèdent pas toutes les propriétés des nombres ; elles ont donc des qualités mathématiques différentes. Par exemple, à la question portant sur la destination des dernières vacances de quatre jours et plus hors du domicile habituel, nous avions les catégories de réponses suivantes :

1. Québec
2. Canada (à l'exception du Québec)
3. États-Unis
4. Europe
5. Autre

La moyenne des destinations n'a ici aucun sens ; la seule opération mathématique possible serait le mode (la catégorie qui a la fréquence absolue ou relative la plus élevée).

On se rend compte que le type d'échelle de mesure conditionne fortement les opérations mathématiques possibles. Le tableau 1.3 présente ces diverses opérations selon le type d'échelle de mesure utilisée.

Tableau 1.3

LES OPÉRATIONS MATHÉMATIQUES POSSIBLES SELON LE TYPE D'ÉCHELLE

Échelles	*Les calculs statistiques utilisables*	*Les tests des relations entre les variables*
Nominale	– Fréquence absolue et relative – Mode	– Khi carré – Coefficient de contingence – Coefficient phi – Lambda – Régression logistique
Ordinale	Ceux de l'échelle nominale plus : – Médiane – Mesures de positions	Ceux de l'échelle nominale plus : – Corrélation de rang – Autres tests non paramétriques – Régression logistique ordinale
Intervalles	Ceux des deux premières échelles plus : Mesures de tendance centrale et de dispersion (moyenne, écart-type…)	Ceux des deux premières échelles plus : – Analyse de variance – Corrélation de Pearson – Régression simple et multiple
Rapport	Tous	Tous

Il est important de combiner les échelles avec les opérations statistiques qui sont acceptables, car la solidité de nos résultats dépendra de notre capacité de respecter les propriétés mathématiques de la mesure. On ne peut à la fois s'appuyer sur les règles mathématiques et faire le contraire dans la pratique de l'analyse des données.

3.2.1. Les relations entre les échelles de mesure

Il y a une relation d'inclusion entre les différents niveaux de mesure. L'échelle de rapport possède toutes les propriétés des autres échelles en plus des siennes propres. L'échelle d'intervalles cumule les propriétés des

échelles nominale et ordinale en plus d'avoir des caractéristiques bien à elle. L'échelle ordinale a les propriétés d'ordonner et de classer (échelle nominale). L'échelle nominale est le niveau de mesure le plus primitif. Il y a donc une complexité croissante de l'échelle nominale à l'échelle de rapport.

Une même variable peut être mesurée par des échelles différentes (mais ce n'est pas toujours possible pour toutes les variables : par exemple, la variable sexe restera toujours au niveau nominal). Donnons un exemple. À la question :

Consommez-vous du vin à la maison ?

On aura les réponses suivantes, selon les échelles utilisées :

- Échelle nominale :
  1. Oui
  2. Non
- Échelle ordinale :
  1. Jamais
  2. Rarement
  3. Souvent
  4. Très souvent
- Échelle d'intervalles : nous allons ici changer la formulation de la question :

  Combien de fois avez-vous consommé du vin à la maison dans le dernier mois ?
  1. Jamais
  2. 1 à 5 fois
  3. 6 à 10 fois
  4. 11 à 15 fois
  5. 16 à 20 fois
  6. 21 fois et plus

Dans le cas de l'échelle nominale, nous savons que la personne consomme (ou non) du vin. Pour l'échelle ordinale, la mesure est un peu plus précise. Par l'échelle par intervalles, nous avons une représentation plus juste de la consommation de vin dans la population étudiée. Dans cette dernière échelle, nous avons une unité de temps et des quantités, ce

qui donne une plus grande précision aux données recueillies. Il semble évident que pour une étude de marché, notre client (Société des alcools, centre d'alimentation ou dépanneur) s'intéressera beaucoup plus à la mesure par intervalles qu'aux autres niveaux de mesure.

3.2.2. Les niveaux de mesure et les méthodes d'analyse des données

Dans le tableau 1.4, nous avons regroupé les niveaux de mesure et les méthodes d'analyse des données selon que l'on désire faire une étude descriptive ou explicative des informations recueillies.

Dans le tableau 1.4, plusieurs méthodes sont proposées ; elles seront choisies en fonction du niveau de mesure (les échelles) et du choix entre une approche descriptive ou explicative et selon le nombre de variables à l'étude.

Tableau 1.4

LES NIVEAUX DE MESURE ET LES MÉTHODES D'ANALYSE DES DONNÉES SELON UNE APPROCHE « DESCRIPTIVE » OU « EXPLICATIVE »

Échelle	*Approche « descriptive »*	*Approche « explicative »*
Nominale	Étude des fréquences Analyse factorielle	Tableaux croisés Analyse discriminante Régression logistique
Ordinale	Étude des fréquences Mesures de position Analyse factorielle	Tableaux croisés Analyse discriminante Régression logistique
Intervalles	Étude des fréquences Mesures de position Mesures de tendance centrale et de dispersion Analyse factorielle	Tableaux croisés Analyse discriminante Analyse de variance Régression logistique Régression simple et multiple
Rapport	Étude des fréquences Mesures de position Mesures de tendance centrale et de dispersion Analyse factorielle	(Tableaux croisés) Analyse discriminante Analyse de variance (Régression logistique) Régression simple et multiple

Dans l'étude des fréquences, des mesures de tendance centrale, de dispersion et de positionnement, les variables sont prises une à une : c'est une analyse univariée. Dans les tableaux croisés, l'analyse de variance, la corrélation, la régression simple, on procède à une analyse bivariée : on

y étudie la relation entre deux variables. Lorsque l'analyse (analyse des correspondances, analyse en composantes principales, analyse discriminante et régression) vise à étudier les relations entre deux ou plus de deux variables, on parlera d'analyse multivariée. Nous verrons ces méthodes de l'analyse des données dans les chapitres qui vont suivre.

CHAPITRE

2

Le traitement des données par ordinateur

Le développement de l'analyse des données, telle qu'on la connaît aujourd'hui, est étroitement lié aux découvertes réalisées dans le domaine de l'informatique. Si l'histoire des machines à calculer date de plusieurs siècles[1], celle des ordinateurs modernes est relativement récente. C'est dans la foulée du projet Manhattan[2] qu'a été conçu le premier supercalculateur permettant de traiter plusieurs milliers d'informations en quelques secondes.

Après 1980, l'utilisation des puces de silicium permit la construction de microordinateurs et de logiciels accessibles à tous à peu de frais. Ces diverses inventions simplifient à l'extrême l'analyse des données (basée sur des procédures statistiques et des calculs interminables). Ainsi, en

1. Voir à ce sujet : J. Marguin (1994), *Histoire des instruments et des machines à calculer : trois siècles de mécanique pensante*, Paris, Hermann. Il faut souligner que c'est en 1889 que Herman Hollerith inventa la fameuse carte perforée ; voir à ce sujet : C. Reyraud (1998), « L'essor des machines à compter », *Les cahiers de Science et Vie*, nº 48.
2. Voir à ce sujet : (1992), « Le projet Manhattan : histoire de la première bombe atomique », *Les cahiers de Science et vie*, hors-série, nº 7.

2000, réaliser une analyse en composantes principales ou une analyse de régression multiple devient à la portée de n'importe quel étudiant de première année d'université !

Comme il arrive souvent, ce rapide développement a entraîné quelques inconvénients dans l'utilisation de la microinformatique pour l'analyse des données. Le principal obstacle vient, pour certains, de la confusion entre le traitement et l'analyse des données[3].

Le néophyte en ce domaine a tendance à donner à l'ordinateur un pouvoir symbolique ; il écrira dans un rapport de recherche : « l'ordinateur nous indique – nous montre – telle ou telle chose ». Ce genre de démarche tend à occulter le travail même de construction théorique des données ; il va donc à l'encontre des théories actuelles de la connaissance.

Un autre obstacle, moins important, vient de l'utilisation quasi automatique de certaines méthodes sans tenir compte tellement de la structure des données et des objectifs de l'étude à réaliser. Cet obstacle résulte de la grande facilité à employer les logiciels spécialisés dans l'analyse des données.

Quoi qu'il en soit, les microordinateurs et les logiciels de traitement des données sont là pour rester ; ils sont une véritable bénédiction pour ceux qui doivent utiliser les méthodes de l'analyse des données dans leur travail quotidien. Il faut aborder l'analyse des données par son côté ludique, allier la logique de la démarche scientifique à l'imagination et à la curiosité. Comme le souligne William Fox : « Être joueur est cependant essentiel et complète la discipline et la rigueur. Vous apprendrez et vous obtiendrez davantage des statistiques, et vos analyses seront fructueuses, si votre approche est celle du jeu[4]. »

1. LES LOGICIELS DE TRAITEMENT DES DONNÉES

Les logiciels de traitement des données sont nombreux et, pour la plupart, très bien construits et très faciles à utiliser. Nous allons citer ici les plus importants :

3. Voir à ce sujet : R. Bertrand (1986), *L'analyse statistique des données*, Sainte-Foy, Presses de l'Université du Québec, p. 11.
4. W. Fox (1999), *Statistiques sociales*, Québec, Presses de l'Université Laval, p. 26.

- Le logiciel Excel, produit par Microsoft, est sûrement le plus connu et le plus utilisé ; la version la plus récente contient une partie des procédures statistiques utilisées dans les analyses des données.
- StatBox et Question, mis au point par la firme Grimmer Logiciels, sont des logiciels conçus spécialement pour l'analyse des données d'enquête ; ces logiciels fonctionnent à partir du logiciel Excel de Microsoft.
- Le Sphinx, dont le concepteur est Jean Moscarola, professeur à Grenoble, est un logiciel utilisé surtout pour la recherche marketing.
- Minitab est un logiciel statistique puissant qui propose un grand nombre de procédures statistiques.
- Le logiciel SAS (système d'analyse statistique) a été conçu au départ pour le calcul économique et les modèles de régression ; par la suite, on l'a adapté de façon à y inclure les méthodes les plus connues de l'analyse des données.
- Le logiciel SPSS (Statistical Package for the Social Sciences) a été créé, au tout début, pour les besoins des psychologues. Avec le temps (cette entreprise existe depuis 1965), on a intégré un grand nombre de procédures statistiques tout en facilitant le travail de manipulation des données.

Dans l'ensemble, tous les logiciels statistiques se valent. À la longue, de perfectionnement en perfectionnement, ils finissent par tous se ressembler ! Quatre éléments vont surtout jouer dans l'achat d'un logiciel de traitement des données :

- L'apprentissage : le logiciel dans lequel on a appris le traitement des données a une certaine longueur d'avance sur les autres (on évite de réapprendre le maniement d'un logiciel).
- L'accessibilité : le produit est-il accessible dans notre ville ou dans notre région ?
- La maniabilité : la simplicité dans l'entrée des données et dans les commandes générales et particulières.
- Enfin, le coût du logiciel.

Dans ce livre, nous allons utiliser le logiciel SPSS sous Windows, car il arrive premier pour tous les critères énoncés ci-dessus. Après avoir comparé les logiciels cités plus haut, le logiciel SPSS nous semble le plus performant ; c'est véritablement la « Rolls Royce » des logiciels de traitement des données.

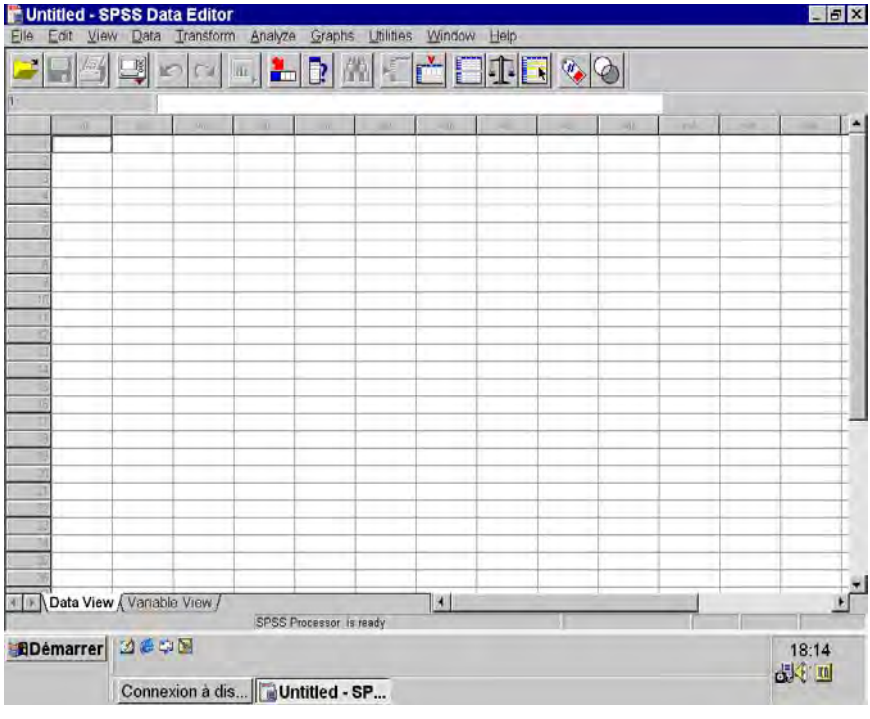
2. LE FONCTIONNEMENT DU LOGICIEL SPSS

Le logiciel SPSS fonctionne à partir de fenêtres et de menus. Chacun des menus présente plusieurs commandes et chacune des commandes comprend des sous-commandes qui précisent la commande principale. Ce logiciel ressemble donc à l'emboîtement des poupées russes.

La figure 2.1 reproduit la fenêtre d'application. Cette fenêtre est un tableau où les lignes correspondent à des observations et les colonnes, à des variables.

Figure 2.1

LA FENÊTRE D'APPLICATION



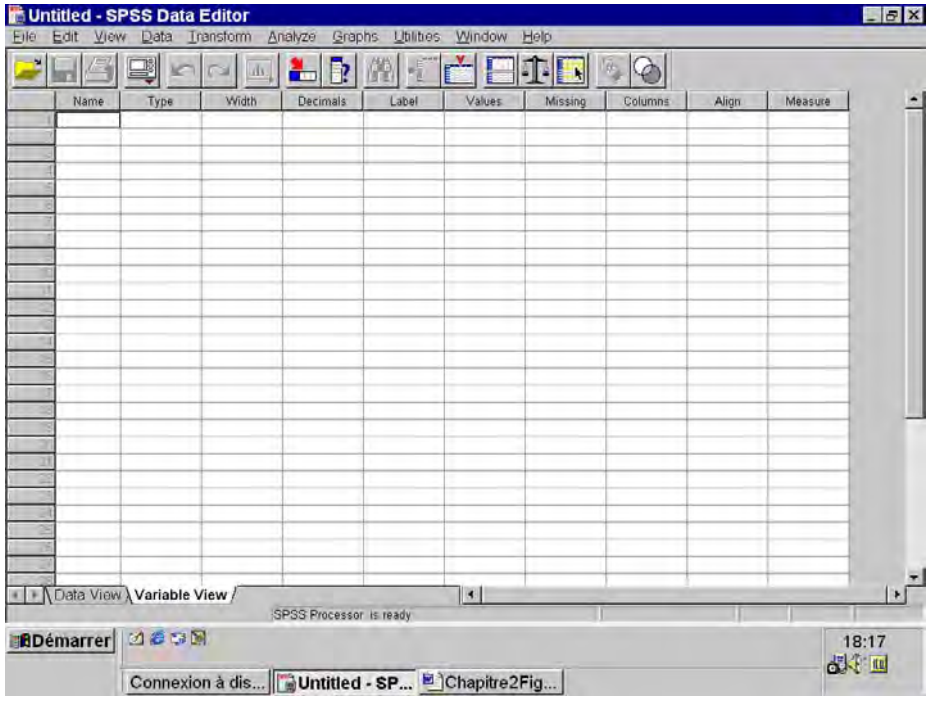
Par exemple, dans une enquête quantitative par sondage, les lignes représentent les répondants et les colonnes, les questions posées. Chacune des fenêtres contient des menus déroulants ; ces menus seront présentés plus loin dans ce chapitre.

La figure 2.2 nous montre la fenêtre servant à la définition des variables.

Chaque variable sera donc définie par dix colonnes contenant les caractéristiques particulières de chacune des variables.

Figure 2.2

LA FENÊTRE SERVANT À LA DÉFINITION DES VARIABLES

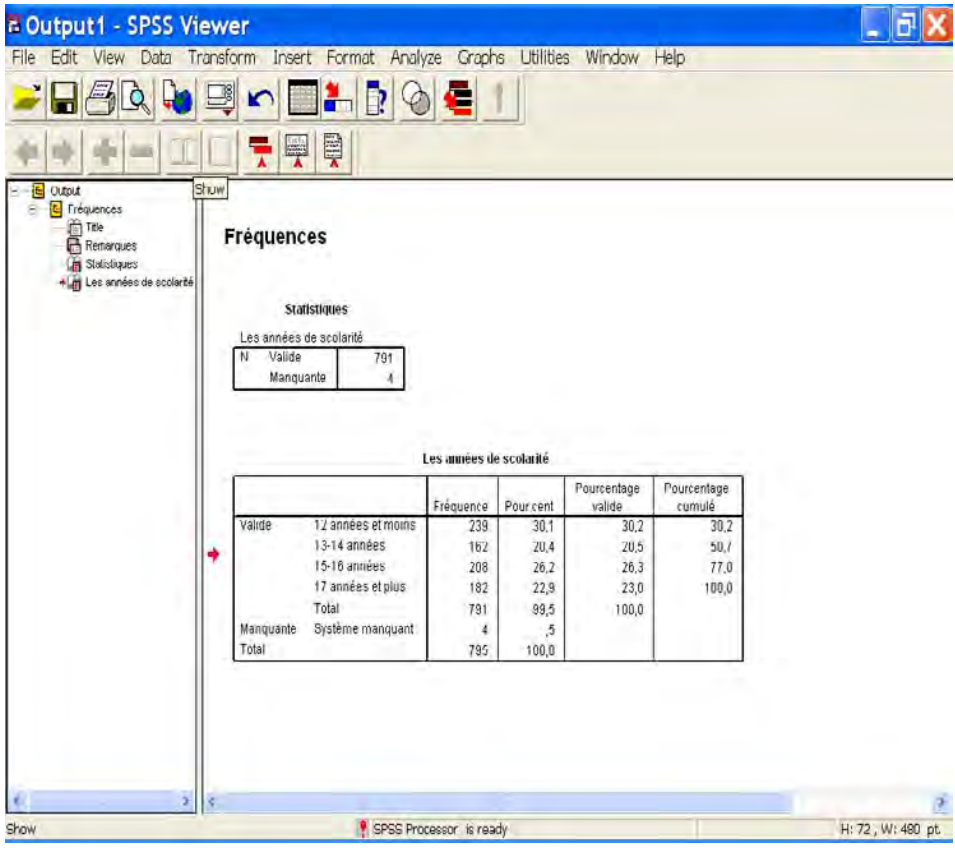


Chacune des variables doit donc répondre à une forme de questionnaire interne : nom, type de variable (numérique ou non), etc. Nous verrons ces caractéristiques en détail dans la section 3 de ce chapitre.

Quand une commande est exécutée, apparaît automatiquement une fenêtre qui montre les résultats obtenus. Nous voyons cette fenêtre et ces résultats dans la figure 2.3.

Figure 2.3

LA FENÊTRE DES RÉSULTATS



Dans le rectangle de gauche, nous avons la table des matières des résultats qui apparaissent dans la surface de droite. Avec les versions récentes de SPSS, tous les éléments qui apparaissent dans le tableau de la figure 2.3 peuvent être modifiés.

Les menus proposés par le logiciel font partie intégrante des fenêtres présentées aux figures 2.1 et 2.2.

Voyons maintenant chacun de ces menus :

- **File** est le menu qui concerne le fichier de travail ; il permet de créer un fichier SPSS, de le sauvegarder et aussi, si nécessaire, de créer des copies du fichier principal.
- **Edit** ou édition contient les commandes servant à couper, copier et coller du texte, ainsi que les fonctions de recherche et les options très nombreuses de ce logiciel.
- **View** porte sur l'organisation même des fenêtres et des infos-bulles (que nous verrons plus loin).
- **Data** est un menu très important, car il permet de définir des variables et d'insérer de nouvelles informations et de nouvelles variables si besoin est.
- **Transform** joue aussi un rôle essentiel, qui est de transformer les variables selon les besoins de l'analyse des données.
- **Analyze** renferme les principales procédures statistiques, les plus connues et les plus utilisées dans tous les domaines des sciences sociales et des sciences de la gestion.
- **Graphs** est le menu qui permet de créer des graphiques de toutes les formes possibles.
- **Utilities** propose deux façons d'afficher les informations : par le nom des variables ou par leur contenu.
- **Window** donne un accès facile et rapide aux fenêtres d'applications, de définition des variables et aux fenêtres des résultats de l'application des commandes.
- Enfin, **Help** fournit des indications sur les façons d'utiliser les commandes de SPSS et sur les diverses procédures statistiques.

Dans la fenêtre d'application (figure 2.1) et dans la fenêtre servant à la définition des variables (figure 2.2), apparaissent sous la barre des menus principaux les infobulles.

Les infobulles sont des boutons qui permettent un accès facile et rapide à certaines commandes contenues ou non dans les menus principaux.

Figure 2.4

LES INFOBULLES


Ouvrir un fichier


Sauver un fichier


Imprimer


Rappeler la commande la plus récente


Annuler la dernière opération


Aller à un graphique


Accéder à une case particulière


Obtenir une information à propos d'une variable


Trouver des données


Insérer une ligne


Insérer une colonne


Diviser les données en deux groupes


Donner un poids à certaines variables


Sélectionner une case


Faire apparaître ou disparaître les étiquettes des variables


Créer des ensembles de variables

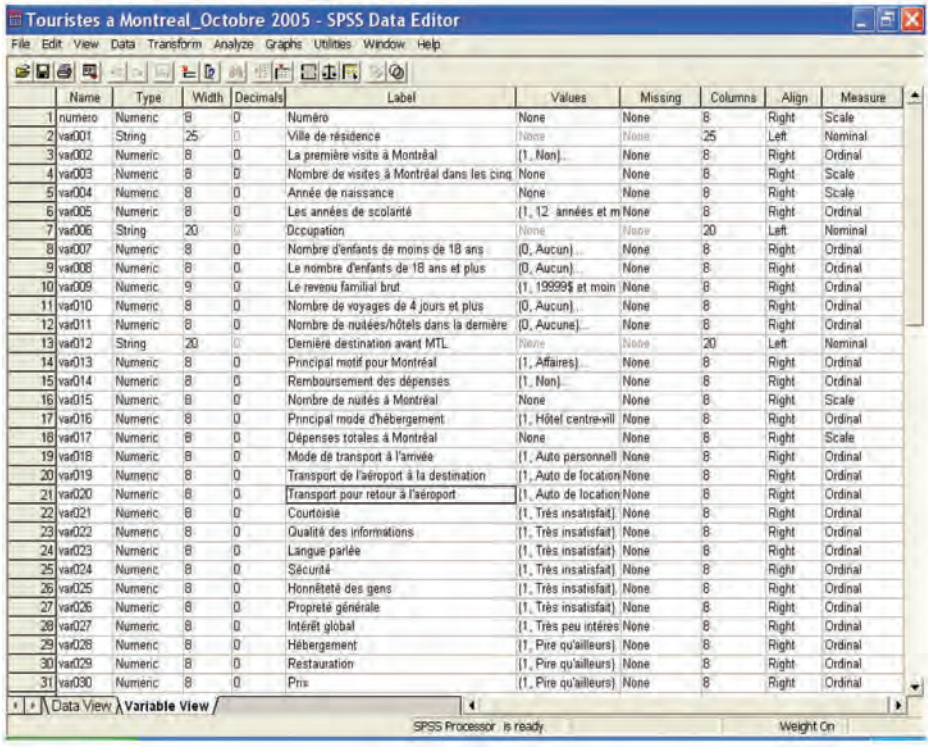
3. LES PRINCIPALES COMMANDES

Les principales commandes de SPSS concernent plus particulièrement la définition des variables et la saisie des données ; sans ces opérations essentielles, l'analyse des données est impossible.

La définition des variables se fait à partir de la fenêtre servant à cette opération (voir la figure 2.2). Les variables sont définies à partir de dix éléments ; ces éléments apparaissent à l'en-tête des colonnes.

Figure 2.5

LA DÉFINITION DES VARIABLES



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	numero	Numeric	8	0	Numéro	None	None	8	Right	Scale
2	var001	String	25	0	Ville de résidence	None	None	25	Left	Nominal
3	var002	Numeric	8	0	La première visite à Montréal	{1, Non}...	None	8	Right	Ordinal
4	var003	Numeric	8	0	Nombre de visites à Montréal dans les cinq	None	None	8	Right	Scale
5	var004	Numeric	8	0	Année de naissance	None	None	8	Right	Scale
6	var005	Numeric	8	0	Les années de scolarité	{1, 12 années et m	None	8	Right	Ordinal
7	var006	String	20	0	Occupation	None	None	20	Left	Nominal
8	var007	Numeric	8	0	Nombre d'enfants de moins de 18 ans	{0, Aucun}...	None	8	Right	Ordinal
9	var008	Numeric	8	0	Le nombre d'enfants de 18 ans et plus	{0, Aucun}...	None	8	Right	Ordinal
10	var009	Numeric	9	0	Le revenu familial brut	{1, 19999$ et moin	None	8	Right	Ordinal
11	var010	Numeric	8	0	Nombre de voyages de 4 jours et plus	{0, Aucun}...	None	8	Right	Ordinal
12	var011	Numeric	8	0	Nombre de nuitées/hôtels dans la dernière	{0, Aucune}...	None	8	Right	Ordinal
13	var012	String	20	0	Dernière destination avant MTL	None	None	20	Left	Nominal
14	var013	Numeric	8	0	Principal motif pour Montréal	{1, Affaires}...	None	8	Right	Ordinal
15	var014	Numeric	8	0	Remboursement des dépenses	{1, Non}...	None	8	Right	Ordinal
16	var015	Numeric	8	0	Nombre de nuités à Montréal	None	None	8	Right	Scale
17	var016	Numeric	8	0	Principal mode d'hébergement	{1, Hôtel centre-vill	None	8	Right	Ordinal
18	var017	Numeric	8	0	Dépenses totales à Montréal	None	None	8	Right	Scale
19	var018	Numeric	8	0	Mode de transport à l'arrivée	{1, Auto personnell	None	8	Right	Ordinal
20	var019	Numeric	8	0	Transport de l'aéroport à la destination	{1, Auto de location	None	8	Right	Ordinal
21	var020	Numeric	8	0	Transport pour retour à l'aéroport	{1, Auto de location	None	8	Right	Ordinal
22	var021	Numeric	8	0	Courtoisie	{1, Très insatisfait}	None	8	Right	Ordinal
23	var022	Numeric	8	0	Qualité des informations	{1, Très insatisfait}	None	8	Right	Ordinal
24	var023	Numeric	8	0	Langue parlée	{1, Très insatisfait}	None	8	Right	Ordinal
25	var024	Numeric	8	0	Sécurité	{1, Très insatisfait}	None	8	Right	Ordinal
26	var025	Numeric	8	0	Honnêteté des gens	{1, Très insatisfait}	None	8	Right	Ordinal
27	var026	Numeric	8	0	Propreté générale	{1, Très insatisfait}	None	8	Right	Ordinal
28	var027	Numeric	8	0	Intérêt global	{1, Très peu intéres	None	8	Right	Ordinal
29	var028	Numeric	8	0	Hébergement	{1, Pire qu'ailleurs}	None	8	Right	Ordinal
30	var029	Numeric	8	0	Restauration	{1, Pire qu'ailleurs}	None	8	Right	Ordinal
31	var030	Numeric	8	0	Prix	{1, Pire qu'ailleurs}	None	8	Right	Ordinal

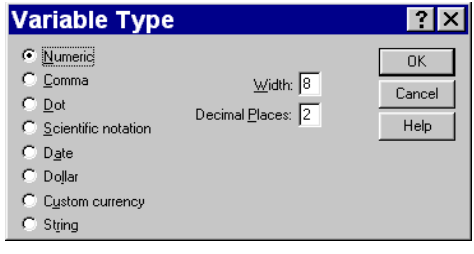
Dans la première colonne de la figure 2.5, nous devons indiquer le nom de la variable ; ce nom doit commencer par une lettre et ne pas avoir plus de huit caractères sans accent ni majuscule. Dans l'exemple de la figure 2.5 (une enquête portant sur les touristes à Montréal), chacune des variables est définie par un numéro d'ordre (qui correspond à sa place exacte dans le questionnaire) ; ainsi, nous avons les variables « var001 », « var002 », « var003 », etc. Cette façon de procéder est très utile, car, si on tente de donner un nom de huit caractères à chacune des variables, cette tâche deviendra très compliquée après la quinzième variable (ou avant). De toute manière, la variable sera définie plus longuement, et en français, dans la colonne Label.

En cliquant sur les trois points (…) à l'intérieur de la case dans la colonne Type, on fait apparaître la boîte de dialogue servant à définir le type de la variable (figure 2.6).

Figure 2.6

LA BOÎTE DE DIALOGUE SERVANT À DÉFINIR LE TYPE DE LA VARIABLE

La définition automatique « par défaut » de la variable.



Cette boîte de dialogue résume les réponses à donner pour les colonnes Type, Width et Decimals.

Dans la liste de gauche de la boîte de dialogue (figure 2.6), on doit choisir entre plusieurs possibilités :

- **NUMERIC** : à choisir si les réponses sont codifiées de façon numérique ;
- **COMMA** : si les valeurs possèdent des virgules, un signe plus ou moins, etc.
- **DOT** : si les données ont des points comme séparateurs de chiffres ;

- **SCIENTIFIC NOTATION** : cette possibilité autorise la notation scientifique, la plus utilisée étant le E intercalaire (par exemple $123^{E}3$) ;
- **DATE** : si l'information est une date (exemple : la date de naissance) ;
- **DOLLAR** : le signe de dollar avec des virgules et un point ;
- **CUSTOM CURRENCY** : à utiliser pour d'autres types de monnaie (euro, rouble, dinar, etc.) ;
- **STRING** : l'entrée de données en lettres, par exemple, le nom d'une ville, d'un produit, etc. Il est fortement suggéré, si l'on veut construire des tableaux avec cette variable et utiliser des procédures statistiques, de traduire les variables-lettres en données numériques (Montréal = 1, Québec = 2, etc.).

Dans la figure 2.6, nous voyons deux carrés à droite dans la boîte à dialogue :

- **WIDTH** sert à indiquer le nombre de caractères ; si ce n'est pas mentionné, le logiciel inscrit 8 par défaut.
- **DECIMAL PLACES** indique le nombre de décimales de la variable à définir (le nombre maximal étant de 16 décimales). Par la suite, on doit cliquer sur **OK** pour exécuter la commande ; cette commande d'exécution apparaîtra dans toutes les boîtes de dialogue.

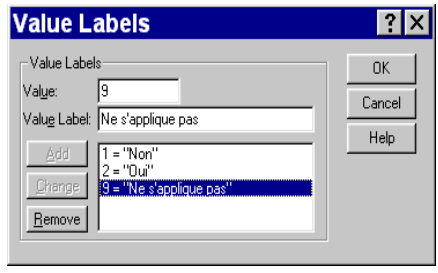
Les cinquième, sixième et septième colonnes de la figure 2.5 sont très importantes dans la définition des variables. C'est dans la sortie des résultats que les propriétés définies dans ces trois colonnes apparaîtront. Voyons ces trois colonnes. La colonne cinq, Label, sert à définir la variable en utilisant un plus grand nombre de caractères (le maximum est de 120 caractères), y compris les majuscules et les accents. Par exemple, pour la variable « var002 », nous avons écrit : « La première visite à Montréal ». Le contenu de ce champ correspond donc à l'étiquette que prendra la variable.

Les étiquettes des différentes valeurs de la variable apparaissent dans la sixième colonne. En cliquant sur les trois points (…) dans la case correspondante (ligne « var002 » et colonne Values), apparaît la boîte de dialogue de la codification (figure 2.7).

Figure 2.7

LA BOÎTE DE DIALOGUE DE LA CODIFICATION

Le premier rectangle contient la valeur numérique de l'une des catégories de la variable. Le deuxième rectangle, le contenu sémantique de cette catégorie. En cliquant sur ADD, on relie les deux éléments.



Dans la figure 2.7, le premier rectangle désigne le code de la variable (ici 9) ou **VALUE**.

Le deuxième rectangle, **VALUE LABEL**, renferme l'étiquette en lettres de la valeur de cette variable (ici : « Ne s'applique pas »). En cliquant sur le bouton **ADD**, on enregistre la commande. Pour détruire un code qui ne fait pas l'affaire, il faut cliquer sur le bouton **REMOVE**, puis changer le code et cliquer ensuite sur le bouton **CHANGE**.

La colonne sept, Missing, désigne les valeurs manquantes de la variable. La figure 2.8 reproduit la boîte de dialogue des valeurs manquantes de la variable.

Figure 2.8

LA BOÎTE DE DIALOGUE DES VALEURS MANQUANTES DE LA VARIABLE

Il existe ici trois possibilités :

1. **Aucune valeur manquante**
2. **Valeurs manquantes discrètes**
3. **Intervalle plus une valeur manquante discrète**

Le premier rectangle (à gauche) contient la liste des variables source.

Le deuxième rectangle (à droite) désigne la ou les variables sélectionnées pour créer un tableau de fréquence

Le premier bouton, **NO MISSING VALUES**, indique qu'il n'y a aucune valeur manquante pour cette variable. Le deuxième bouton indique qu'il s'agit d'une ou plusieurs valeurs manquantes discrètes : **DISCRETE MISSING VALUES**. On ne peut désigner ici que trois valeurs manquantes discrètes.

Le troisième bouton sert à indiquer l'intervalle des valeurs manquantes, **RANGE PLUS ONE OPTIONAL DISCRETE MISSING VALUE** (on pourrait désigner, par exemple, les valeurs de 9 à 20). Le dernier rectangle sert à mentionner (si elle existe) une valeur discrète en plus de l'intervalle mentionné plus haut.

Les trois dernières colonnes ont un sens purement décoratif :

- Columns désigne le nombre de caractères vraiment utilisés.
- Align sert à aligner les nombres à l'intérieur de la case (et de la colonne).
- Measure indique si les valeurs de la variable sont nominales ou métriques (elles sont métriques par défaut).

L'entrée des données se fait en fonction de la définition des variables et, bien sûr, à partir des réponses aux questionnaires. Dans la figure 2.9, nous avons l'entrée des données pour les vingt-neuf premiers répondants et les dix premières variables.

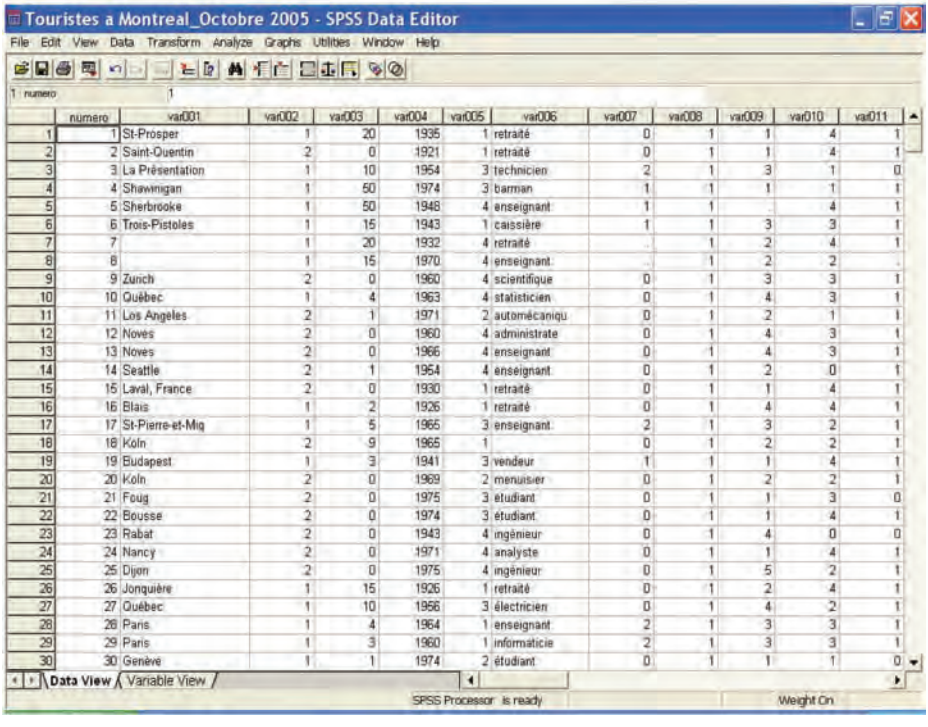
Comment lire ces données ? Nous allons partir du premier répondant (ligne 1) de la fenêtre. Voyons les onze premières colonnes de la ligne 1 :

1. 1 : numéro 1 – le premier questionnaire codifié ;
2. « St-Prosper » (var001) : la ville de résidence ;
3. 1 (var002) : « La première visite à Montréal » ; 1 = « Non » ;
4. 20 (var003) : « Nombre de visites à Montréal dans les cinq dernières années » ; le répondant n° 1 est venu 20 fois à Montréal ;
5. 1935 (var004) : « Année de naissance » ;
6. 1 (var005) : « Les années de scolarité » où 1 correspond à « 12 années et moins » ;
7. Retraité (var006) : « Occupation » ;

8. 0 (var007) : « Nombre d'enfants de moins de 18 ans » ;
9. 1 (var008) : « Nombre d'enfants de 18 ans et plus » ;
10. 1 (var009) : « Le revenu familial brut », où 1 correspond à l'intervalle : 19 999 $ ou moins ;
11. 4 (var010) : « Nombre de voyages de 4 jours et plus » ; ici 4 signifie « 4 voyages ou plus ».

Figure 2.9

L'ENTRÉE DES DONNÉES



	numero	var001	var002	var003	var004	var005	var006	var007	var008	var009	var010	var011
1	1	St-Prosper	1	20	1935	1	retraité	0	1	1	4	1
2	2	Saint-Quentin	2	0	1921	1	retraité	0	1	1	4	1
3	3	La Présentation	1	10	1954	3	technicien	2	1	3	1	0
4	4	Shawinigan	1	50	1974	3	barman	1	1	1	1	1
5	5	Sherbrooke	1	50	1948	4	enseignant	1	1	.	4	1
6	6	Trois-Pistoles	1	15	1943	1	caissière	1	1	3	3	1
7	7	.	1	20	1932	4	retraité	.	1	2	4	1
8	8	.	1	15	1970	4	enseignant	.	1	2	2	.
9	9	Zurich	2	0	1960	4	scientifique	0	1	3	3	1
10	10	Québec	1	4	1963	4	statisticien	0	1	4	3	1
11	11	Los Angeles	2	1	1971	2	automécaniqu	0	1	2	1	1
12	12	Noves	2	0	1960	4	administrate	0	1	4	3	1
13	13	Noves	2	0	1966	4	enseignant	0	1	4	3	1
14	14	Seattle	2	1	1954	4	enseignant	0	1	2	0	1
15	15	Laval, France	2	0	1930	1	retraité	0	1	1	4	1
16	16	Blais	1	2	1926	1	retraité	0	1	4	4	1
17	17	St-Pierre-et-Miq	1	5	1965	3	enseignant	2	1	3	2	1
18	18	Köln	2	9	1965	1	.	0	1	2	2	1
19	19	Budapest	1	3	1941	3	vendeur	1	1	1	4	1
20	20	Koln	2	0	1969	2	menuisier	0	1	2	2	1
21	21	Foug	2	0	1975	3	étudiant	0	1	1	3	0
22	22	Bousse	2	0	1974	3	étudiant	0	1	1	4	1
23	23	Rabat	2	0	1943	4	ingénieur	0	1	4	0	0
24	24	Nancy	2	0	1971	4	analyste	0	1	1	4	1
25	25	Dijon	2	0	1975	4	ingénieur	0	1	5	2	1
26	26	Jonquière	1	15	1926	1	retraité	0	1	2	4	1
27	27	Québec	1	10	1956	3	électricien	0	1	4	2	1
28	28	Paris	1	4	1964	1	enseignant	2	1	3	3	1
29	29	Paris	1	3	1960	1	informaticie	2	1	3	3	1
30	30	Genève	1	1	1974	2	étudiant	0	1	1	1	0

4. L'ÉTUDE DES FRÉQUENCES ET LA TRANSFORMATION DES VARIABLES

Il est nécessaire de faire une première analyse des fréquences avant de transformer les variables. Donnons un exemple, celui de la variable « var003 », « Nombre de visites à Montréal depuis 5 ans ». Voici les principales commandes (voir aussi la figure 2.10) :

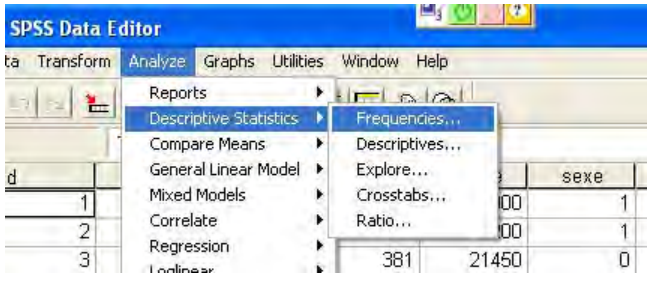
A) dans le menu **Analyze** ;

B) choisir la commande **Descriptive Statistics** ;

C) puis la sous-commande **Frequencies**.

Figure 2.10

LE CHEMINEMENT POUR LES TABLEAUX DE FRÉQUENCE



À l'appel de la sous-commande **Frequencies**, apparaît une boîte de dialogue (figure 2.11).

Figure 2.11

LA COMMANDE DES FRÉQUENCES

Le premier rectangle (à gauche) contient la liste des variables source.

Le deuxième rectangle (à droite) désigne la ou les variables sélectionnées pour créer un tableau de fréquences.

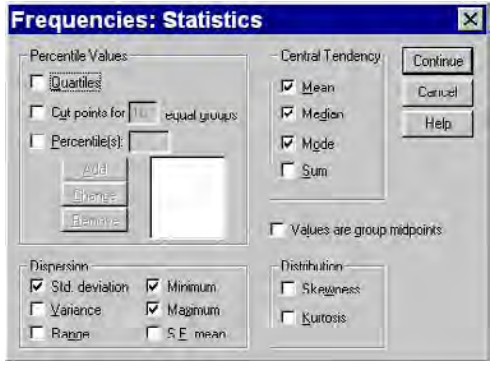
Dans le rectangle de gauche, nous avons la liste des variables du fichier des données. Dans le rectangle de droite, il faut faire « glisser » la variable pour laquelle nous voulons créer un tableau des fréquences. Il s'agit de « noircir » la variable choisie dans le rectangle de gauche et de cliquer sur la flèche au milieu ; la variable sera sélectionnée et apparaîtra dans le rectangle de droite. Pour faire passer la variable de droite à gauche, il faut à nouveau cliquer sur la flèche.

En bas de la boîte de dialogue **Frequencies…**, nous avons trois boutons. Le premier produit une sous-boîte de dialogue portant sur les statistiques usuelles ; le deuxième concerne la production d'un graphique et le troisième, la forme du tableau des fréquences. Pour obtenir des statistiques, il faut cliquer sur le bouton **STATISTICS** (voir la figure 2.11).

Le bouton **STATISTICS** génère une sous-boîte de dialogue qui est présentée à la figure 2.12.

Figure 2.12

LA SOUS-COMMANDE DES PROCÉDURES STATISTIQUES



Dans la figure 2.12, nous avons choisi, en cliquant dans les carrés, les procédures statistiques :

- **MEAN** (moyenne) ;
- **MEDIAN** (médiane) ;
- **MODE** (mode) ;
- **STD. DEVIATION** (écart-type) ;
- **MINIMUM** (valeur la plus petite) ;
- **MAXIMUM** (valeur la plus grande).

Par la suite, il faut cliquer sur le bouton **CONTINUE** pour revenir à la boîte de dialogue principale. De retour à la boîte de dialogue principale (voir figure 2.11), on clique sur **OK** pour exécuter la commande.

4.1. L'ANALYSE DES FRÉQUENCES

Les résultats de la commande des fréquences sont présentés dans l'encadré 2.1.

Encadré 2.1

LES RÉSULTATS DE LA COMMANDE DES FRÉQUENCES

Statistiques

N	Valide	794
	Manquante	1
Moyenne		7,55
Médiane		3,00
Mode		0
Ecart-type		12,169
Minimum		0
Maximum		60

Nombre de visites à Montréal dans les cinq dernières années

		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide	0	257	32,3	32,4	32,4
	1	51	6,4	6,4	38,8
	2	85	10,7	10,7	49,5
	3	68	8,6	8,6	58,1
	4	33	4,2	4,2	62,2
	5	50	6,3	6,3	68,5
	6	19	2,4	2,4	70,9
	7	12	1,5	1,5	72,4
	8	13	1,6	1,6	74,1
	9	3	,4	,4	74,4
	10	48	6,0	6,0	80,5
	12	3	,4	,4	80,9
	15	38	4,8	4,8	85,6
	20	41	5,2	5,2	90,8
	22	4	,5	,5	91,3
	25	4	,5	,5	91,8
	30	19	2,4	2,4	94,2
	35	2	,3	,3	94,5
	40	8	1,0	1,0	95,5
	42	2	,3	,3	95,7
	44	2	,3	,3	96,0
	50	27	3,4	3,4	99,4
	52	2	,3	,3	99,6
	55	1	,1	,1	99,7
	60	2	,3	,3	100,0
	Total	794	99,9	100,0	
Manquante	Système manquant	1	,1		
Total		795	100,0		

Dans le petit tableau, nous avons les statistiques usuelles demandées :

N	Valide	794	: nombre de réponses
	Manquante	1	: absence de réponse
Moyenne		7,55	: moyenne
Médiane		3	: médiane
Mode		0	: mode
Écart-type		12,169	: écart-type
Minimum		0	: valeur minimale
Maximum		60	: valeur maximale

Au bas de l'encadré 2.1, apparaît le tableau des fréquences. La lecture du tableau se fait comme suit :

- Colonne 1 : le nombre des visites à Montréal (de 0 à 60) ;
- Colonne 2 : la fréquence absolue ; ainsi, 257 personnes ne sont jamais venues à Montréal auparavant ;
- Colonne 3 : la fréquence relative en pourcentage et incluant les valeurs manquantes ;
- Colonne 4 : la fréquence relative en pourcentage à l'exclusion des valeurs manquantes ;
- Colonne 5 : la fréquence cumulative en pourcentage ; ainsi, on peut écrire que 70,9 % des répondants ont effectué six voyages ou moins à Montréal dans les cinq dernières années.

4.2. LA TRANSFORMATION DES VARIABLES

Ce tableau (encadré 2.1) est beaucoup trop long si nous voulons créer des tableaux croisés ; par exemple, relier la variable « Nombre de visites à Montréal depuis 5 ans » avec des variables indépendantes (« pays d'origine », « âge », « revenus », etc.). Nous devons donc réduire le nombre de catégories de cette variable. Idéalement, nous pouvons construire une échelle d'intervalles égaux ; ainsi, cela donne :

1. 0 Aucun voyage
2. 1-2 voyages
3. 3-4 voyages
4. 5-6 voyages

5. 7-8 voyages
6. 9-10 voyages
7. 11 voyages et plus.

En considérant les données, on se rend compte, par exemple, que la catégorie 5, « 7 ou 8 voyages », ne compte que 25 répondants, ce qui n'est pas beaucoup. Après plusieurs essais, nous réduisons le tableau à cinq catégories :

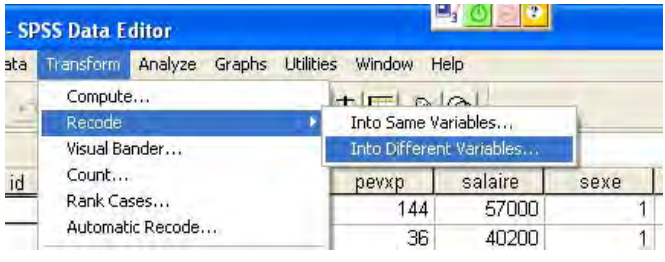
1. 0 Aucun voyage
2. 1-2 voyages
3. 3-4 voyages
4. 5-10 voyages
5. 11 voyages et plus.

Pour recoder var003, il faut suivre les étapes suivantes :

- utiliser le menu (figure 2.13)
- puis la commande **Recode** ;
- ensuite, choisir la sous-commande **Into Different Variables**.

Figure 2.13

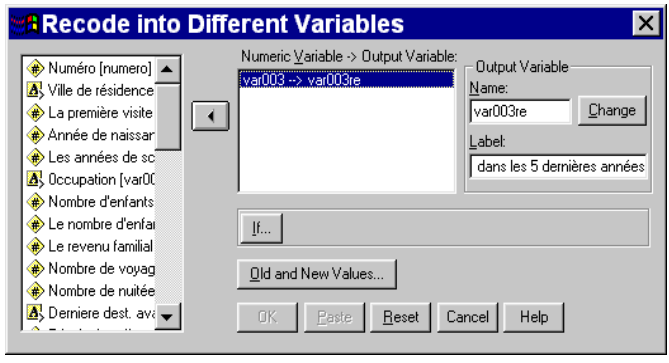
LE CHEMINEMENT POUR RECODER UN TABLEAU DE FRÉQUENCES



À ce moment, une boîte de dialogue apparaît à l'écran (figure 2.14).

Figure 2.14

LA BOÎTE DE DIALOGUE DE LA COMMANDE RECODE

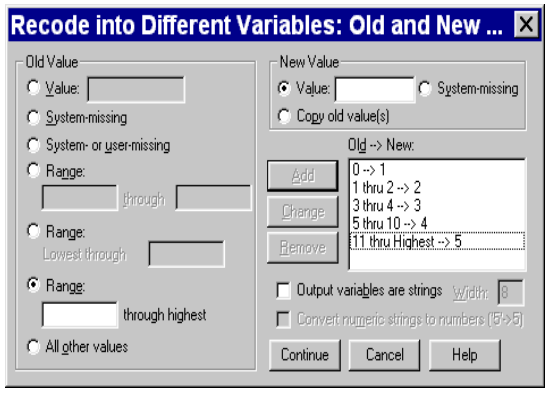


Dans le rectangle de gauche, nous avons la liste des variables du questionnaire ; dans le rectangle du centre, nous avons la variable à transformer. Nous créons une nouvelle variable ; il faut donc lui donner un nouveau nom. Dans le premier petit rectangle de droite, nous devons donner un nouveau nom à l'ancienne variable ; ici ce nouveau nom sera « var003re » ; le petit rectangle du dessous permet de définir l'étiquette de la nouvelle variable. On doit ensuite cliquer sur le bouton **CHANGE** ; à ce moment apparaîtra, dans le rectangle du centre : « var003 → var003re ».

Pour modifier les catégories de la variable, on doit faire appel à une autre boîte de dialogue ; on l'obtient en cliquant sur le bouton **OLD AND NEW VALUES**. Cette sous-boîte de dialogue est reproduite à la figure 2.15.

Figure 2.15

LA BOÎTE DE DIALOGUE DE LA TRANSFORMATION DE LA VARIABLE



La partie de gauche de cette boîte de dialogue est réservée aux anciennes valeurs et celle de droite, aux nouvelles valeurs de la variable. Pour indiquer que 0 (aucun voyage) prendra désormais le code 1, on active le bouton placé devant **VALUE** dans le rectangle de gauche (**OLD VALUE**) et on inscrit 0 dans le rectangle ; on active ensuite le premier bouton à droite et on écrit 1 dans le rectangle de droite (**NEW VALUE**) ; enfin, on clique sur le bouton **ADD** et on obtient dans le grand rectangle de droite : « 0 ⟶ 1 ».

Pour le code 2, nous retournons à gauche (dans la partie **OLD VALUE**), on active le quatrième bouton (**RANGE**) et l'on inscrit 1 dans le premier rectangle et 2 dans le deuxième rectangle – on active ensuite le premier bouton de droite (dans la partie **NEW VALUE**) et l'on inscrit 2 dans le rectangle – ensuite on clique sur le bouton **ADD** et nous obtenons dans le grand rectangle de droite : « 1 thru 2 ⟶ 2 ».

Il faut faire la même chose pour les codes 3 (3-4) et 4 (5-10) ; à ce moment, on obtiendra dans le rectangle de droite : « 3 thru 4 ⟶ 3 » et « 5 thru 10 ⟶ 4 ». Pour obtenir la dernière catégorie 5, nous revenons à gauche dans le sixième cercle où l'on écrit dans le rectangle adjacent 11 (11 voyages et plus). Dans le principal rectangle de droite (**NEW VALUE**), nous pouvons vérifier le résultat de la recodification ; il devrait se présenter ainsi :

Old ⟶ New :

0 ⟶ 1

1 thru 2 ⟶ 2

3 thru 4 ⟶ 3

5 thru 10 ⟶ 4

11 thru Highest ⟶ 5

À gauche de la flèche, nous avons les anciennes valeurs de la variable et à droite, les nouvelles valeurs (voir la figure 2.15).

Nous devons cliquer sur le bouton **CONTINUE** pour revenir à la boîte de dialogue principale (figure 2.14). Parvenu dans cette boîte, **Recode into Different Variables**, on clique sur **OK** pour faire exécuter la commande.

La variable « var003 » recodée (qui porte le nom « var003re ») sera placée à la toute fin du fichier des données. Il est possible d'obtenir de nouveaux tableaux des fréquences pour cette variable. Ces tableaux sont présentés dans l'encadré 2.2.

Encadré 2.2

LA VARIABLE « VAR003 » TRANSFORMÉE

Tableau A: Nombre de visites à Montréal dans les cinq dernières années

				Pourcentage	Pourcentage
Valide	1	257	32,3	32,4	32,4
	2	136	17,1	17,1	49,5
	3	101	12,7	12,7	62,2
	4	145	18,2	18,3	80,5
	5	155	19,5	19,5	100,0
	Total	794	99,9	100,0	
Manquante	Système manquant	1	,1		
Total		795	100,0		

Tableau B: Nombre de visites à Montréal dans les cinq dernières années

				Pourcentage	Pourcentage
Valide	0	257	32,3	32,4	32,4
	1-2	136	17,1	17,1	49,5
	3-4	101	12,7	12,7	62,2
	5-10	145	18,2	18,3	80,5
	11 et +	155	19,5	19,5	100,0
	Total	794	99,9	100,0	
Manquante	Système manquant	1	,1		
Total		795	100,0		

Tableau C: Nombre de visites à Montréal dans les cinq dernières années

			Pourcentage	Pourcentage
Valide	0 (Aucune)	257	32,4	32,4
	1-2 visites	136	17,1	49,5
	3-4 visites	101	12,7	62,2
	5-10 visites	145	18,3	80,5
	11 visites et plus	155	19,5	100,0
	Total	794	100,0	

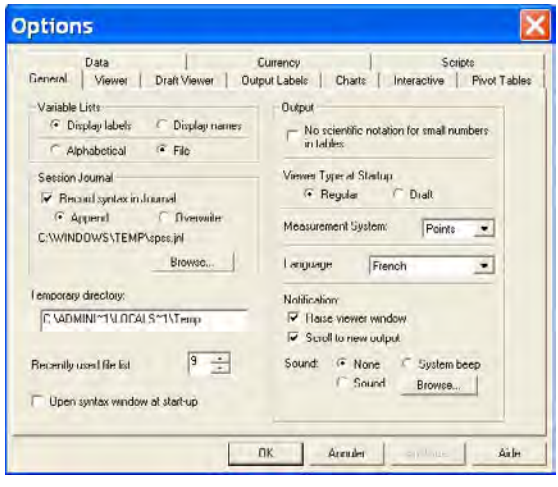
Dans le tableau A, nous avons la variable recodifiée. Nous pouvons remarquer, dans ce tableau, que les codes apparaissent (de 1 à 5) mais que la définition de chacun des codes n'apparaît pas. Pour créer le tableau B, il faut revenir à la fenêtre de définition des variables (voir les figures 2.2 et 2.5) et dans la colonne **LABEL**, donner un nouveau nom à la variable « var003re » ; on doit aussi donner de nouvelles étiquettes aux cinq catégories en utilisant la colonne **VALUES** (voir la figure 2.7).

Dans les chapitres qui vont suivre, nous verrons, pour chacune des méthodes d'analyse multivariée choisies, toutes les étapes du traitement des données avec le logiciel SPSS Windows.

Figure 2.16

COMMENT OBTENIR DES RÉSULTATS EN FRANÇAIS ?

Le cheminement est le suivant : dans le menu **Edit**, sélectionnez **Options**, puis choisissez l'onglet **General**, et dans la section **Language,** choisissez **French** (choix de dix langues).



Pour continuer, voici quelques références intéressantes au sujet du logiciel SPSS :

BAILLARGEON, G. et F. OUELLET (1999). *Traitement des données avec SPSS pour Windows*, Trois-Rivières, SMG.

FIELD, A. (2003). *Discovering Statistics using SPSS for Windows*, Londres, Sage.

GEORGE, D. et P. MALLERY (1999). *SPSS for Windows Step by Step*, Boston, Allyn and Bacon.

GREEN, S., N. SALKIND et T. AKEY (1997). *Using SPSS for Windows : Analyzing and Understanding Data*, Upper Saddle River, Prentice Hall.

HOWITT, D. et D. CRAMER (1997). *A Guide to Computing Statistic with SPSS for Windows*, Boston, Prentice Hall.

KINNEAR, P. et C. GRAY (2000). *SPSS for Windows Made Simple*, East Sussex, Psychology Press.

NORUSIS, M. (1997). *SPSS for Windows* : *Boase System User's Guide*, Chicago, SPSS.

NORUSIS, M. (1998). *Guide to Data Analysis, SPSS 8.0*, Upper Saddle River, Prentice Hall.

PLAISENT, M., P. BERNARD, E. MORIN, C. ZUCCARO, E. CHÉRON et P. BODSON (1999). *SPSS 9.0 Windows. Guide d'autoformation*, Sainte-Foy, Presses de l'Université du Québec.

RODEGHIER, M. (1996). *Surveys with Confidence* : *A Practical Guide to Survey Research Using SPSS*, Chicago, SPSS.

SHANNON, D. et M. DAVENPORT (2001). *Using SPSS to Solve Statistical Problems : A Self-Instruction Guide*, Upper Saddle River, Prentice Hall.

CHAPITRE

3

L'analyse factorielle en composantes principales

L'analyse factorielle est un excellent exemple pour illustrer l'étude multivariée des données. C'est une approche qui vise à réduire un grand nombre d'informations sur un sujet donné à un petit nombre d'éléments plus facilement interprétables.

On distingue habituellement l'analyse factorielle de l'analyse en composantes principales. L'analyse factorielle tente généralement de vérifier une ou plusieurs hypothèses ; c'est donc une approche « confirmatoire[1] ». On suppose que dans cet amas de données existe une structure sous-jacente qui confirmera les avancées théoriques. Elle peut aussi servir à mesurer la validité de certaines échelles d'opinions ou d'attitudes[2].

L'analyse en composantes principales est habituellement une analyse factorielle exploratoire ; les résultats de l'analyse seront de nouvelles hypothèses permettant d'élargir et de mieux comprendre le problème

1. Voir à ce sujet : R. Bertrand (1986), *L'analyse statistique des données*, Sainte-Foy, Presses de l'Université du Québec, p. 18.
2. Voir à ce sujet : J.-J. Bernier (1985), *Théorie des tests*, Chicoutimi, Gaëtan Morin, p. 255.

étudié. L'analyse en composantes mène donc à l'analyse factorielle bien qu'il s'agisse de deux démarches distinctes, mais complémentaires. En fait, il s'agit de deux expressions de la même méthode.

1. OBJECTIFS ET ASPECTS THÉORIQUES

L'analyse factorielle en composantes principales a surtout trois objectifs :

1. étudier les interrelations entre un assez grand nombre de variables ;
2. à partir de cette étude, regrouper ces variables dans des groupes limités appelés facteurs ou composantes ;
3. établir entre ces groupes de variables une hiérarchie basée essentiellement sur la valeur explicative de chacun d'eux (il est à noter que la méthode permet aussi d'établir une hiérarchie des variables dans chacune des composantes).

En bref, l'analyse factorielle considère quatre types de relations :

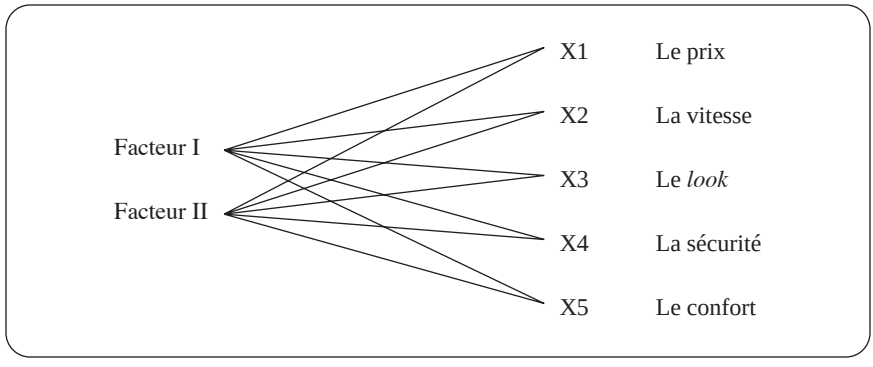
1. les relations des variables entre elles ;
2. les relations des variables aux facteurs ;
3. les relations entre les variables d'un même facteur ;
4. les relations entre les différents facteurs.

Au plan pratique, l'analyse factorielle en composantes principales essaie de répondre à des questions simples ; par exemple, au sujet d'une marque de voiture, parmi les qualités suivantes : le prix, la vitesse, le *look*, la sécurité, le confort, lesquelles sont les plus importantes ? On pose ces questions à un échantillon de clients possibles ; ils doivent noter chacune de ces qualités de 1 à 10, où 1 = Pas du tout important et 10 = Très important.

On aura donc les relations suivantes :

Figure 3.1

LES RELATIONS DES VARIABLES AUX FACTEURS[3]



On voit (dans la figure 3.1) que les variables sont reliées à tous les facteurs. L'analyse subséquente permettra, par exemple, d'arriver aux résultats suivants :

La sécurité Le confort	Facteur I	Ces deux variables expliqueraient à 80 % l'importance que cette population accorde à une marque de voiture.
Le prix La vitesse Le *look*	Facteur II	Regroupe 20 % de l'importance que l'on donne à une marque de voiture

Dans la publicité, les qualités de sécurité et de confort sont les arguments les plus importants à utiliser pour maximiser les ventes d'une marque donnée[4].

Dans le facteur I, la sécurité est jugée plus importante que le confort ; dans le facteur II, il y a aussi une gradation, un ordre hiérarchique : le prix, la vitesse, le *look*. Les facteurs dégagés permettent d'indiquer les éléments (ici les qualités) les plus importants d'une situation tout en conservant

3. Nous nous inspirons ici de la démarche de J.O. Kim et C. Mueller (1978), *Introduction to Factor Analysis*, Beverly Hills, Sage University Paper n° 13, p. 25 ss.
4. Il s'agit d'un exemple imaginaire.

tous les éléments de l'ensemble. L'analyse en composantes principales, dans cet exemple, a bien rempli son rôle : réduire les données et donner une certaine explication aux choix effectués par les répondants.

L'analyse en composantes principales doit respecter certaines contraintes :

- le nombre des variables doit être suffisant (cinq variables ou plus) ;
- la forme des réponses aux questions (les items) doit être la même (par exemple, cinq choix de réponse) ; dans le cas contraire, les variables doivent être réduites et normalisées ;
- on doit avoir dix fois plus de cas qu'il y a de variables impliquées ; par exemple, 10 variables $\times$ 10 cas donnent une taille *n* égale à 100.

1.1. Les étapes de l'analyse en composantes principales

Les principales étapes[5] de l'analyse en composantes principales sont :

1. la recherche des variables similaires ; celles-ci doivent faire partie d'un même ensemble : mesure de la satisfaction, de l'intérêt, etc. ;
2. la matrice des corrélations entre les variables choisies ;
3. la diagonalisation de la matrice $\lambda_1, \lambda_2, \ldots \lambda_n$;
4. la matrice des saturations, qui permet de dégager les facteurs ;
5. la rotation, qui désigne les facteurs les plus importants selon leur degré d'inertie (de variance expliquée) ;
6. la définition « littéraire » des facteurs ;
7. la lecture des tests les plus importants ;
8. l'interprétation des résultats au plan des décisions et de l'action.

Voyons certaines de ces étapes à partir d'un exemple. On veut évaluer la qualité d'un restaurant à partir des questions suivantes (les variables de satisfaction) :

5. Il faut noter que certaines de ces étapes peuvent se réaliser simultanément.

1. le prix ;
2. le menu ;
3. le service ;
4. le goût des aliments ;
5. l'ambiance du restaurant ;
6. les portions.

Pour chacune des questions, les réponses suivantes sont proposées :

1. Très mauvais
2. Mauvais
3. Plus ou moins bon
4. Bon
5. Très bon

Le tableau 3.1 pésente la matrice des corrélations entre les variables.

Tableau 3.1

LA MATRICE DES CORRÉLATIONS* ENTRE LES VARIABLES

	1 Prix	***2 Menu***	***3 Service***	***4 Goût***	***5 Ambiance***	***6 Portion***
1 Prix	1	0,551	0,577	0,648	0,289	0,216
2 Menu	0,551	1	0,556	0,509	0,318	0,275
3 Service	0,577	0,556	1	0,686	0,257	0,196
4 Goût	0,648	0,509	0,686	1	0,332	0,291
5 Ambiance	0,289	0,318	0,257	0,332	1	0,740
6 Portion	0,216	0,275	0,196	0,291	0,740	1

* Dans ce tableau, plus le chiffre est élevé, plus la relation entre les variables est forte (1 = relation « totale » et 0 = relation nulle).

On voit dans cette matrice que le prix est plus fortement relié au goût (0,648), au service (0,577) et au menu (0,551) et plus faiblement relié à l'ambiance (0,289) et à la portion (0,216). Nous retrouvons à peu près les mêmes caractéristiques pour les variables menu, service et goût ; les deux dernières variables semblent assez marginales.

1.2. LA DIAGONALISATION ET LES SATURATIONS DES VARIABLES

La deuxième étape est la « diagonalisation » ; il s'agit, à partir de la matrice de corrélation initiale, de dégager des facteurs. Cela se fait par la multiplication et la transposition de la matrice originale selon la formule suivante[6] :

$$\begin{array}{lll} (a_1a_1 + b_1b_1) & (a_1a_2 + b_1b_2) & (a_1a_3 + b_1b_3) \\ (a_2a_1 + b_2b_1) & (a_2a_2 + b_2b_2) & (a_2a_3 + b_2b_3) \\ (a_3a_1 + b_3b_1) & (a_3a_2 + b_3b_2) & (a_3a_3 + b_3b_3) \end{array} \quad \begin{array}{lll} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{array} \quad \begin{array}{ll} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{array}$$

$$R = F^1 \times F$$

Ici, R correspond à la matrice des corrélations, F^1 à la matrice transposée et F à la matrice des facteurs[7].

La diagonalisation suppose de longs calculs assez pénibles. Ces fameux calculs sont présentés au tableau 3.2[8].

La matrice A : dans la matrice des corrélations initiales, nous remplaçons, dans la diagonale, le coefficient 1 par le coefficient le plus élevé de la colonne. On fait par la suite l'addition de toutes les colonnes, ce qui donne ici : 16,938 ; on extrait la racine carrée de 16,938 = 4,1156. Ensuite, nous divisons le total de chacune des colonnes par 4,1156 ; le résultat nous donne le premier facteur (première ligne et première colonne de la matrice B).

6. Cette formule est tirée de : J.-P. Guilford (1954), *Psychometric Methods*, New York, McGraw-Hill, p. 480.
7. Voir à ce sujet : P. Horst (1965), *Factorial Analysis of Data Matrices*, New York, Holt, Rinehart and Winston.
8. Nous nous inspirons ici de la méthode de Thurstone présentée par : G. Thomson (1950), *L'analyse factorielle des aptitudes humaines*, Paris, Presses universitaires de France, p. 24-27 et p. 174-177.

Tableau 3.2

LA DIAGONALISATION DES VARIABLES ET LA TRANSPOSITION DES MATRICES

Matrice A		***1***	***2***	***3***	***4***	***5***	***6***	
	1	(0,648)	0,551	0,577	0,648	0,289	0,216	Le coefficient le plus
	2	0,551	(0,556)	0,556	0,509	0,318	0,275	élevé doit être placé
	3	0,577	0,556	(0,686)	0,686	0,257	0,196	dans les diagonales
	4	0,648	0,509	0,686	(0,686)	0,332	0,291	
	5	0,289	0,318	0,257	0,332	(0,740)	0,740	
	6	0,216	0,275	0,196	0,291	0,740	(0,740)	
	TOTAL	2,929	2,765	2,958	3,152	2,676	2,458	16,938
Saturation I	Facteur I 16,938 = 4,1156	0,7112	0,6718	0,7187	0,7659	0,6502	0,5972	
Matrice B	0,7112	(0,5058)	0,4778	0,5111	0,5447	0,4624	0,4247	Matrice du premier
	0,6718	0,4778	(0,4513)	0,4828	0,5145	0,4368	0,4012	facteur
	0,7187	0,5111	0,4828	(0,5165)	0,5504	0,4673	0,4292	
	0,7659	0,5447	0,5145	0,5504	(0,5866)	0,4980	0,4574	
	0,6502	0,4624	0,4368	0,4673	0,4980	(0,4228)	0,3753	
	0,5972	0,4247	0,4012	0,4292	0,4574	0,3883	(0,3566)	
Matrice C		(0,1422)	0,0732	0,0659	0,1033	–0,1734	–0,2087	Première matrice
		0,0992	(0,1047)	0,0732	0,0055	–0,1188	–0,1262	résiduelle A – B
		0,1369	0,0732	(0,1695)	0,1356	–0,2103	–0,2332	
		0,1033	–0,0055	0,1356	(0,994)	–0,166	–0,1664	
		–0,1734	–0,1188	–0,2103	–0,166	(0,3172)	0,3647	
		–0,2087	–0,1262	–0,2332	–0,1664	0,3517	(0,3834)	
		0,0995	0,0000	0,0000	0,0011	0,0004	0,0136	
Matrice D		(0,2087)	0,0732	0,0659	0,1033	–0,1734	–0,2087	Le coefficient le plus
		0,0992	(0,1262)	0,0732	0,0055	–0,1188	–0,1262	plus élevé de chaque
		0,1369	0,0732	(0,2332)	0,1356	–0,2103	–0,2332	colonne (sans tenir
		0,1033	–0,0055	0,1356	(0,1664)	–0,166	–0,1664	compte du signe)
		–0,1734	–0,1188	–0,2103	–0,166	0,3517	0,3647	doit être placé dans
		–0,2087	–0,1262	–0,2332	–0,1664	0,3517	(0,3647)	la diagonale
	TOTAL	0,9302	0,5231	0,9514	0,7432	1,3719	1,4639	Somme ; abstraction faite du signe

Tableau 3.2 (suite)

LA DIAGONALISATION DES VARIABLES ET LA TRANSPOSITION DES MATRICES

Sommation des totaux des six colonnes = 5,9837 = 5,9837 = 2,4462								
		1	***2***	***3***	***4***	***5***	***6***	
Saturation II	Facteur II	0,3802	0,2184	0,3889	0,3038	0,5608	0,5984	
Matrice E		(0,2087)	0,0732	0,0659	0,1033	0,1734	0,2087	On reproduit la
		0,0992	(0,1262)	0,0732	0,0055	0,1188	0,1262	matrice D sans
		0,1369	0,0732	0,2332	0,1356	0,2103	0,2332	les signes
		0,1033	0,0055	0,1356	0,1664	0,1660	0,1664	
		0,1734	0,1188	0,2103	0,1660	0,3517	0,3647	
		0,2087	0,1262	0,2332	0,1664	0,3517	0,3647	
Matrice F		0,3802	0,2184	0,3889	0,3038	0,5608	0,5984	Matrice du second
	0,3802	(0,1445)	0,083	0,1478	0,1155	0,2132	0,2275	facteur sans
	0,2184	0,0830	(0,0477)	0,0849	0,0663	0,1225	0,1307	les signes
	0,3889	0,1478	0,0849	(0,1512)	0,1181	0,2180	0,2327	
	0,3038	0,1155	0,0663	0,1181	(0,0923)	0,1704	0,1819	
	0,5608	0,2132	0,1225	0,2180	0,1704	(0,3145)	0,3356	
	0,5984	0,2275	0,1307	0,2327	0,1819	0,3356	(0,3581)	
Matrice G		0,0642	0,0098	–0,0819	–0,0122	–0,0398	–0,0188	Seconde matrice
		0,0162	0,0785	–0,0117	–0,0608	–0,0037	–0,0045	résiduelle E – F
		–0,0109	–0,0117	0,0820	0,0175	–0,0077	0,0005	
		–0,0122	–0,0608	0,0175	0,0741	–0,0044	–0,0155	
		–0,0398	–0,0040	–0,0077	–0,0044	0,0372	0,0291	
		–0,0188	–0,0045	0,0005	–0,0155	0,0161	0,0068	
		–0,0013	0,0073	–0,0013	–0,0013	0,0023	–0,0024	

La matrice B est produite par la multiplication des coefficients de la première ligne et de la première colonne. Par exemple, pour la deuxième colonne, on aura : 0,7112 × 0,7112 = 0,5058, 0,6718 × 0,7112 = 0,4778, etc.

La **matrice C** est la première matrice résiduelle. Elle est produite par la soustraction de la matrice A moins la matrice B. La somme des colonnes de la matrice C devrait être de zéro ou près de zéro. C'est une méthode de contrôle des colonnes. On voit que les résultats de la colonne 1 donnent 0,0995, ce qui est relativement élevé, mais rien n'est parfait !

Dans la **matrice D**, nous suivons la même démarche que dans la matrice A : les totaux des colonnes servent à calculer le deuxième facteur (qui est aussi le deuxième niveau de saturation des coefficients). Les **matrices E, F et G** ne sont que la répétition des calculs effectués dans les premières matrices.

Il faut noter qu'afin de continuer les opérations de saturation, dans les matrices D, E, F et G on ne tient pas compte du signes positif ou négatif. Cette astuce, recommandée par Thurstone, permet de dégager le facteur suivant (et les autres facteurs s'il y a lieu)[9].

Les étapes de diagonalisation et de saturation des coefficients vont de pair. Par la suite, on peut établir les poids variables et les poids facteurs. Les poids variables s'obstiennent par la formule :

$$X_i = A_{i1}F_1 + A_{i2}F_2 + ... + A_{in}F_n$$

On appelle les poids variables « qualité de la représentation ». Les poids facteurs sont appelés « inertie », « valeur propre » ou « variance expliquée ».

La formule de la variance expliquée est[10] :

$$Fj = \sum_{i=1}^{p} W_{ji} X_i = W_{j1} X_1 + W_{j2} X_2 + ... + W_{jp} X_p$$

où :

W = les coefficients de scores factoriels ;

p = le nombre de variables.

Dans notre exemple, nous parvenons aux résultats du tableau 3.3 pour les deux premiers facteurs ; il faudrait continuer les calculs avec les matrices pour dégager autant de facteurs qu'il y a de variables dans l'étude des composantes principales. Ces résultats sont présentés au tableau 3.3.

9. G. Thomson, *op. cit.*, p. 30.
10. M. Norusis (1992), *SPSS for Windows Professional Statistics*, Chicago, SPSS, p. 48.

Tableau 3.3

L'APPROXIMATION DES DEUX PREMIERS FACTEURS

Variables	*Approximation des facteurs*		*Communauté ou poids variables*
	I	II	
1	0,7112	0,3802	0,6503
2	0,6718	0,2184	0,4990
3	0,7187	0,3889	0,6678
4	0,7609	0,3038	0,6713
5	0,6502	0,5608	0,7372
6	0,5972	0,5984	0,7147
« Poids des facteurs », inertie ou variance expliquée	2,8320	1,1083	3,9403
	72 %	28 %	100 %

Le calcul de la communauté, pour la variable 1, par exemple, se calcule $0{,}7112^2 + 0{,}3802^2 = 0{,}6503$; par la suite, on procède de la même façon pour les variables 2 à 5. Ce qui veut dire que 65 % de la variance de la variable 1 est expliquée par les deux premiers facteurs.

Les poids facteurs vont être la résultante de la sommation des coefficients au carré d'un facteur. Par exemple pour le facteur I, on aura :

$$0{,}7112^2 + 0{,}6718^2 + 0{,}7187^2 + 0{,}7609^2 + 0{,}6502^2 + 0{,}5972^2 = 2{,}8320.$$

Le total 2,8320 nous donne le poids du facteur I. En faisant les mêmes calculs, nous obtenons 1,1083 pour le poids du facteur II. Ainsi : « La somme des carrés des saturations obtenues pour chaque composante indique la proportion de la variance totale de l'information originale imputable à chacune d'elles[11]. »

11. J.-B. Racine et H. Reymond (1973), *L'analyse quantitative en géographie*, Paris, Presses universitaires de France, p. 167.

Dans cet exemple, si l'on arrête la saturation après deux facteurs, la variance expliquée (le degré d'inertie ou la valeur propre) sera pour, le facteur I :

$$\frac{2{,}8320}{3{,}9403} = 0{,}7187\,;$$

en arrondissant et en multipliant par 100, on aura 72 %. Pour le deuxième facteur, on aura :

$$\frac{1{,}1083}{3{,}9403} = 0{,}28 \text{ ou } 28\,\% \text{ (voir le tableau 3.3).}$$

1.3. LA ROTATION DES FACTEURS

En considérant le tableau 3.3, on se rend compte qu'à toutes fins pratiques, il n'existe qu'un seul facteur puisque les coefficients de chacune des variables sont tous plus élevés sur le facteur I. Afin de pallier à cet inconvénient où l'on se retrouve avec autant de facteurs qu'il y a de variables, on va effectuer ce que l'on appelle une rotation orthogonale.

Cette méthode orthogonale « repose sur la maximalisation de la somme des variances des carrés des saturations dans chaque colonne ; il s'ensuit une augmentation de certaines saturations et la diminution des autres[12] ». La formule est :

$$MPF \times MSC = MR$$

où :

MPF = matrice des premiers facteurs ;

MSC = matrice de transformation sinus, cosinus ;

MR = matrice avec rotation.

Nous verrons plus loin qu'il existe d'autres méthodes que la rotation orthogonale.

Grâce à cette méthode de rotation d'axes, chacune des variables de l'ensemble aura le poids le plus élevé possible sur un facteur et le poids le plus faible possible sur les autres facteurs. Il s'agit donc de maximaliser

12. H. Laforge (1981), *Analyse multivariée*, Saint-Laurent, Études vivantes, p. 184.

les corrélations faibles de façon à opérer une discrimination entre les facteurs et de parvenir à une explication plus intéressante des relations entre les variables.

La rotation orthogonale (varimax) appliquée aux données du tableau 3.3 donne la structure finale présentée au tableau 3.4.

Tableau 3.4

LA ROTATION ORTHOGONALE (VARIMAX) APPLIQUÉE À LA MESURE DE LA SATISFACTION D'UN RESTAURANT

Variables	*Facteurs (composantes)*	
	I	II
Goût	0,85	0,07
Service	0,84	0,19
Prix	0,82	0,12
Menu	0,74	0,21
Portion	0,13	0,93
Ambiance	0,20	0,91

Dans cet exemple, le premier facteur regroupe les qualités : goût, service, prix et menu (la place d'une variable sur un facteur est liée à son coefficient : ici, la portion à 0,93 sur le facteur II et l'ambiance à 0,91 sur le facteur II, elles « appartiennent » donc à ce facteur). Si l'on conserve les pourcentages de variance expliquée dans le tableau 3.3, ces quatre items représenteraient 72 % de la satisfaction globale (le facteur I). Les items portion et ambiance (le facteur II) totalisent 28 % de la variance expliquée ; ils sont donc moins importants.

Nous pouvons aussi remarquer que dans le facteur I, il y a une hiérarchie des variables (même si, dans ce cas, les différences sont minimes) ; nous avons la variable goût (0,85) suivie de la variable service (0,84) et du prix (0,82). La variable menu a un coefficient un peu plus faible (0,74). Nous retrouvons un ordre semblable dans le facteur II.

Les étapes 6 (la définition littéraire des facteurs), 7 (la lecture des tests les plus importants) et 8 (l'interprétation des résultats) seront abordées d'une façon plus approfondie dans la partie 3 de ce chapitre. Nous verrons aussi, dans la partie 2 ci-dessous, que ces interminables calculs sont exécutés par le logiciel SPSS sous Windows en quelques picosecondes.

2. LES COMMANDES AVEC LE LOGICIEL SPSS

Le traitement des données par la méthode de l'analyse factorielle en composantes principales est très maniable avec le logiciel SPSS sous Windows. Ce logiciel, pour certaines étapes du traitement des données, propose un grand choix de solutions possibles. Par exemple :

- il offre sept méthodes pour l'extraction des facteurs : **PRINCIPAL COMPONENTS**, **UNWEIGHTED LEAST SQUARES**, **GENERALIZED LEAST SQUARES**, **MAXIMUM LIKEHOOD**, **PRINCIPAL AXIS FACTORING**, **ALPHA FACTORING** et **IMAGE FACTORING**.
- SPSS dispose de cinq méthodes de rotation : **VARIMAX**, **DIRECT OBLIMIN**, **QUARTIMAX**, **EQUAMAX** et **PROMAX**.
- Pour le calcul des facteurs, il existe trois méthodes : **REGRESSION**, **BARTLETT** et **ANDERSON-RUBIN**.

Nous ne pouvons, dans ce livre, présenter chacune de ces méthodes ; nous nous contenterons des plus utilisées dans la recherche actuelle.

2.1. LES COMMANDES PRINCIPALES

Dans un premier exemple, nous voulons mesurer la satisfaction des clients de l'Auberge Frontenac. À partir d'un échantillon aléatoire de la clientèle, nous demandons aux personnes choisies d'exprimer leur satisfaction (ou leur insatisfaction) face aux éléments suivants :

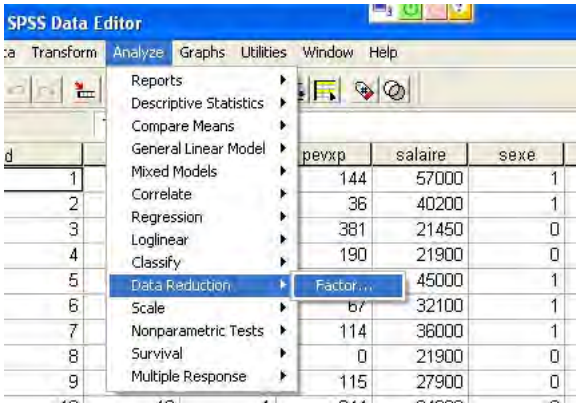
- la restauration ;
- l'hébergement ;
- l'attrait du site ;
- l'ambiance ;
- le prix ;
- l'accueil ;
- le stationnement ;
- la propreté.

Chacune de ces qualités est mesurée par une échelle ordinale allant de 1 à 5, où 1 = Très insatisfait et 5 = Très satisfait.

Dans la figure 3.2, nous avons la boîte de dialogue principale ; pour obtenir cette boîte, on doit utiliser le menu **Analyze** ; dans le menu déroulant, on doit choisir **Data Reduction,** puis **Factor**.

Figure 3.2

LE CHEMINEMENT POUR OBTENIR UNE ANALYSE FACTORIELLE



En cliquant dans le menu **Factor**, on fait apparaître une nouvelle fenêtre.

Figure 3.3

LA BOÎTE DE DIALOGUE PRINCIPALE POUR L'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES

Nous voyons dans la figure 3.3 deux rectangles ; celui de gauche contient toutes les variables utilisées dans l'enquête. Le rectangle au centre droit regroupe toutes les variables de satisfaction qui seront utilisées dans l'analyse factorielle en composantes principales (la flèche au centre permet le passage d'un rectangle à l'autre). En bas de la boîte principale, nous distinguons cinq boutons ouvrant des boîtes de dialogue secondaires.

À droite de la boîte principale, nous avons cinq boutons :

OK : la commande pour exécuter (à la toute fin) l'analyse factorielle ;

PASTE : correspond à la commande « Coller » ;

RESET : correspond à la commande « Restaurer » ;

CANCEL : correspond à la commande « Annuler » ;

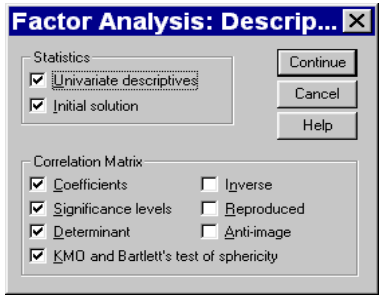
HELP : correspond à la commande « Aide ».

Le petit rectangle du milieu, sous **SELECTION VARIABLE**, permet de réaliser une analyse factorielle pour un segment de la population. Par exemple, la variable 75 nous indique la région où habitent les répondants ; dans le rectangle de gauche, nous sélectionnons la variable 75 ; par la suite, à l'aide du bouton **VALUE**, apparaît une miniboîte de dialogue, **Factor Analysis** : **Set Value** ; à ce moment, nous indiquons que Montréal = code 1. L'analyse factorielle sera réalisée uniquement pour les clients qui habitent la région de Montréal.

Passons maintenant à la boîte de dialogue **Descriptives**, qui est reproduite à la figure 3.4.

Figure 3.4

LA BOÎTE DE DIALOGUE DES STATISTIQUES DESCRIPTIVES



Cette figure propose certaines statistiques, des matrices et des tests. Nous retrouvons les statistiques suivantes :

- les statistiques univariées (**UNIVARIATE DESCRIPTIVES**), qui comprennent la moyenne, l'écart-type et le nombre d'observations valides dans l'échantillon ;

- la structure initiale (**INITIAL SOLUTION**), qui présente la qualité de la représentation initiale, les valeurs propres (la variance expliquée par le modèle).

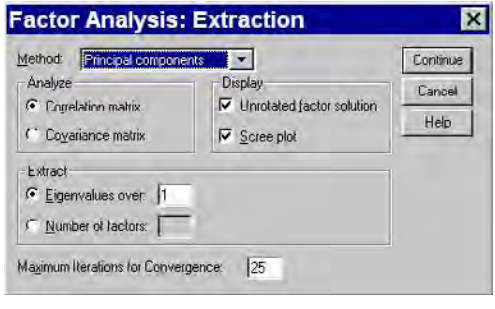
Dans les matrices et les tests statistiques, nous avons :

- la matrice des coefficients de corrélation des variables prises en compte (**COEFFICIENTS**) ;
- le déterminant de la matrice de corrélation (**DETERMINANT**) ;
- les tests KMO et de Bartlett (**KMO AND BARTLETT'S TEST OF SPHERICITY**) ;
- le seuil de signification du test de Bartlett (**SIGNIFICANCE LEVELS**) ;
- l'inverse (**INVERSE**) ;
- la matrice reconstituée (**REPRODUCED**) ;
- l'anti-image (**ANTI-IMAGE**).

Pour extraire les facteurs, on fait appel à la boîte de dialogue **Extraction,** reproduite à la figure 3.5.

Figure 3.5

LA BOÎTE DE DIALOGUE DE L'EXTRACTION DES FACTEURS



Le logiciel nous propose ici sept méthodes d'extraction des facteurs (voir plus haut) ; celle qui apparaît dans la boîte à dialogue de la figure 3.5 est la méthode en composantes principales (« Principal components »). L'analyse souhaitée peut porter sur la matrice de corrélation ou la matrice de covariance (**CORRELATION MATRIX** et **COVARIANCE MATRIX**). Il est possible de produire (**DISPLAY**) la structure factorielle avant la rotation (**UNROTATED FACTOR SOLUTION**) et un graphique des valeurs propres de

chacun des facteurs (**SCREE PLOT**). La commande **EXTRACT** permet de n'accepter que les facteurs dont la valeur propre est égale ou supérieure à 1 (**EIGENVALUES OVER**) ; il est possible d'indiquer ici d'autres valeurs que 1. Il est possible aussi de décider à l'avance le nombre de facteurs pour l'analyse des données (**NUMBER OF FACTORS**). Enfin, nous pouvons indiquer le nombre d'itérations nécessaires pour parvenir à une solution factorielle (**MAXIMUM ITERATIONS FOR CONVERGENCE**) ; par défaut, le nombre maximal d'itérations est fixé à 25.

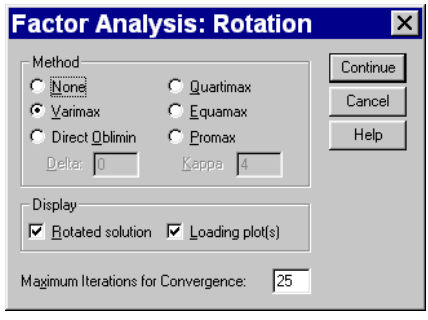
Les méthodes de rotation des facteurs sont présentées à la figure 3.6. Nous avons le choix entre cinq méthodes :

- **VARIMAX** ;
- **QUARTIMAX** ;
- **EQUAMAX** ;
- **DIRECT OBLIMIN** ;
- **PROMAX ;**

La plus connue est la méthode de rotation **VARIMAX.**

Figure 3.6

LA BOÎTE DE DIALOGUE DE LA ROTATION DES FACTEURS

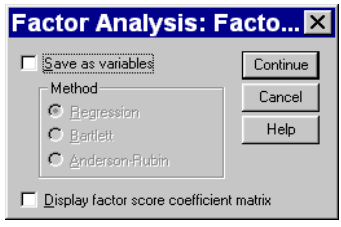


Le logiciel permet d'afficher (**DISPLAY**) la structure factorielle après rotation (**ROTATED SOLUTION**) et le graphique des facteurs choisis (**LOADING PLOT(S)**). On peut indiquer (en bas de la boîte de dialogue) le nombre d'itérations souhaitées pour la saturation des facteurs ; le nombre par défaut est de 25.

À la figure 3.7, nous avons la boîte de dialogue des variables et des scores factoriels.

Figure 3.7

LA BOÎTE DE DIALOGUE DES VARIABLES ET SCORES FACTORIELS

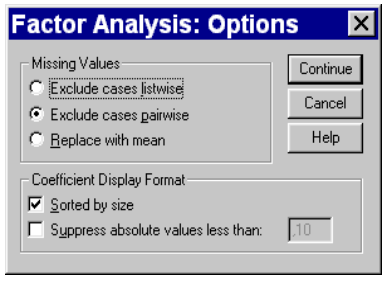


Il est possible de sauvegarder les variables créées pour chacun des facteurs par la commande **SAVE AS VARIABLES**. Trois méthodes sont suggérées : **REGRESSION, BARTLETT** et **ANDERSON-RUBIN.** Le bouton du bas permet de présenter la matrice des coefficients factoriels (**DISPLAY FACTOR SCORE COEFFICIENT MATRIX**).

Enfin, le logiciel SPSS propose aussi des options ; celles-ci apparaissent à la figure 3.8.

Figure 3.8

LA BOÎTE DE DIALOGUE DES OPTIONS



Les premières concernent les valeurs manquantes (s'il y a lieu). Nous avons ici trois possibilités :

- exclure toutes les observations incomplètes, c'est-à-dire que le répondant doit avoir complété TOUTES les questions qui touchent la satisfaction : les huit questions dans l'exemple (**EXCLUDE CASES LISTWISE**) ;

- exclure seulement les réponses non valides : refus, oubli, ne s'applique pas ou ne sais pas (**EXCLUDE CASE PAIRWISE**) ;
- remplacer les réponses manquantes par la moyenne de l'ensemble des répondants (**REPLACE WITH MEAN**).

Les dernières commandes de base sont reliées au tableau des facteurs après rotation ; la première commande permet un classement des variables dans les facteurs en débutant par les variables qui ont les coefficients les plus élevés (**SORTED BY SIZE**). La deuxième commande permet de supprimer, pour chacun des facteurs, les coefficients qui sont inférieurs à la valeur mentionnée (**SUPPRESS ABSOLUTE VALUES LESS THAN**).

2.2. LES TESTS STATISTIQUES

Avant d'analyser et d'interpréter la structure factorielle, il est bon de faire une lecture des principaux tests. Le premier test à considérer est le déterminant de la matrice de corrélation (il est placé sous la matrice de corrélation). Il doit être le plus petit possible sans être égal à zéro ; ce dernier cas indiquerait qu'une ou plusieurs variables ont une corrélation parfaite avec une ou plusieurs autres variables. Ce phénomène rendrait impossibles les saturations de la matrice de corrélation. Il indique aussi que ces variables hautement corrélées sont des « doublets » et font donc double emploi. Dans le cas de l'Auberge Frontenac, le déterminant est égal à 0,003, ce qui est acceptable (voir la sortie des résultats dans l'encadré 3.1, tableau B sous la matrice de corrélation).

Encadré 3.1

LES RÉSULTATS DE L'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES DE LA SATISFACTION DES CLIENTS DE L'AUBERGE FRONTENAC

TABLEAU A

Statistiques descriptives

	Moyenne	Ecart-type	n analyse	N manquantes
Restauration	4,0075	,89398	534	66
Hébergement	3,7914	1,03880	441	159
Attrait/site	3,8162	1,00682	506	94
Ambiance	3,7866	1,06815	478	122
Prix	4,5567	,70542	600	0
Accueil	4,5567	,80074	600	0
Stationnement	4,6644	,58550	587	13
Propreté	4,5436	,76242	596	4

TABLEAU B

Matrice de corrélation[a]

		Restauration	Hébergement	Attrait/site	Ambiance	Prix	Accueil	Stationnement	Propreté
Corrélation	Restauration	1,000	,576	,423	,414	,294	,304	,178	,248
	Hébergement	,576	1,000	,470	,479	,152	,302	,249	,343
	Attrait/site	,423	,470	1,000	,370	,248	,261	,128	,219
	Ambiance	,414	,479	,370	1,000	,226	,375	,145	,328
	Prix	,294	,152	,248	,226	1,000	,284	,133	,150
	Accueil	,304	,302	,261	,375	,284	1,000	,329	,607
	Stationnement	,178	,249	,128	,145	,133	,329	1,000	,386
	Propreté	,248	,343	,219	,328	,150	,607	,386	1,000
Signification (unilatérale)	Restauration		,000	,000	,000	,000	,000	,000	,000
	Hébergement	,000		,000	,000	,001	,000	,000	,000
	Attrait/site	,000	,000		,000	,000	,000	,002	,000
	Ambiance	,000	,000	,000		,000	,000	,001	,000
	Prix	,000	,001	,000	,000		,000	,001	,000
	Accueil	,000	,000	,000	,000	,000		,000	,000
	Stationnement	,000	,000	,002	,001	,001	,000		,000
	Propreté	,000	,000	,000	,000	,000	,000	,000	

a. Déterminant = ,003

TABLEAU C

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,790
Test de sphéricité de Bartlett	Khi-deux approximé	789,225
	ddl	28
	Signification de Bartlett	,000

TABLEAU D

Qualité de représentation

	Initial	Extraction
Restauration	1,000	,635
Hébergement	1,000	,630
Attrait/site	1,000	,557
Ambiance	1,000	,508
Prix	1,000	,206
Accueil	1,000	,671
Stationnement	1,000	,514
Propreté	1,000	,718

Méthode d'extraction : Analyse en composantes principales.

Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

TABLEAU E

Variance totale expliquée

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	3,220	40,246	40,246	3,220	40,246	40,246	2,523	31,541	31,541
2	1,220	15,252	55,498	1,220	15,252	55,498	1,917	23,957	55,498
3	,907	11,333	66,831						
4	,751	9,392	76,223						
5	,604	7,556	83,779						
6	,551	6,883	90,662						
7	,401	5,012	95,674						
8	,346	4,326	100,000						

Méthode d'extraction : Analyse en composantes principales.

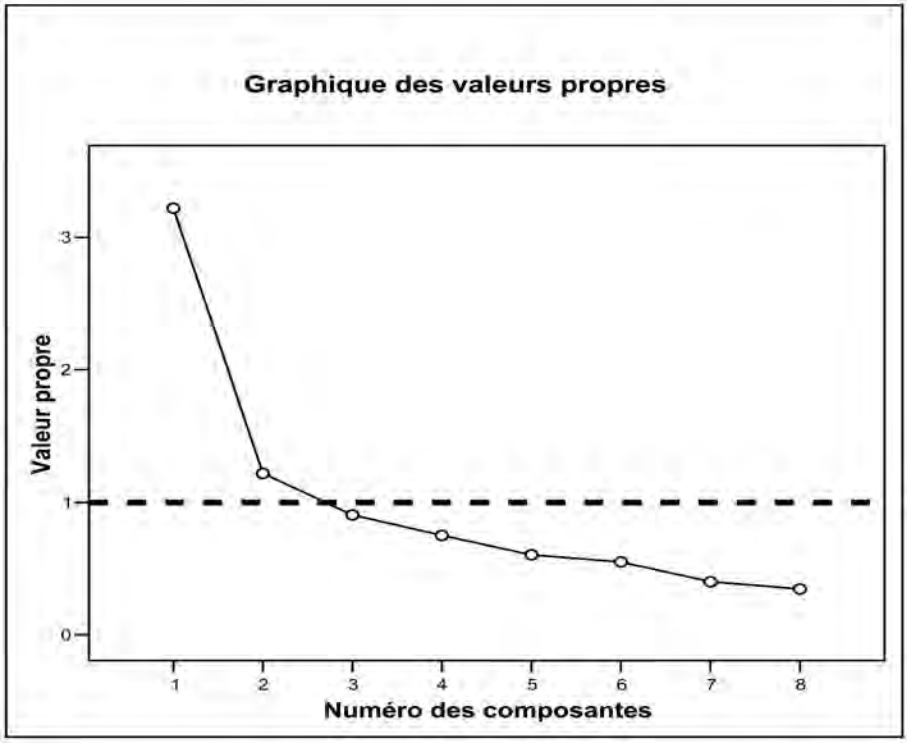


TABLEAU F

Matrice des composantes[a]

	Composante	
	1	2
Hébergement	,740	-,287
Restauration	,703	-,376
Accueil	,691	,440
Ambiance	,684	-,200
Propreté	,655	,538
Attrait/site	,627	-,405
Prix	,444	-,096
Stationnement	,463	,548

Méthode d'extraction : Analyse en composantes principales.
a. 2 composantes extraites.

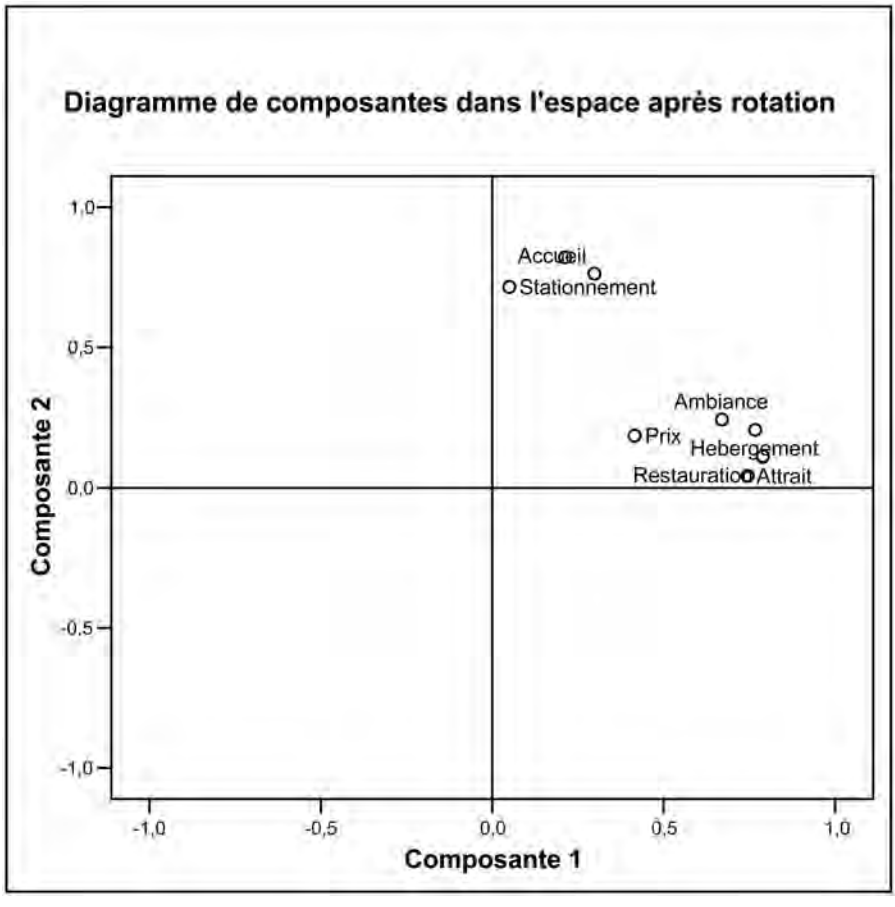
TABLEAU G

Matrice des composantes après rotation[a]

	Composante	
	1	2
Restauration	,789	,111
Hébergement	,767	,205
Attrait/site	,745	,043
Ambiance	,670	,242
Prix	,415	,184
Propreté	,212	,821
Accueil	,298	,763
Stationnement	,050	,716

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

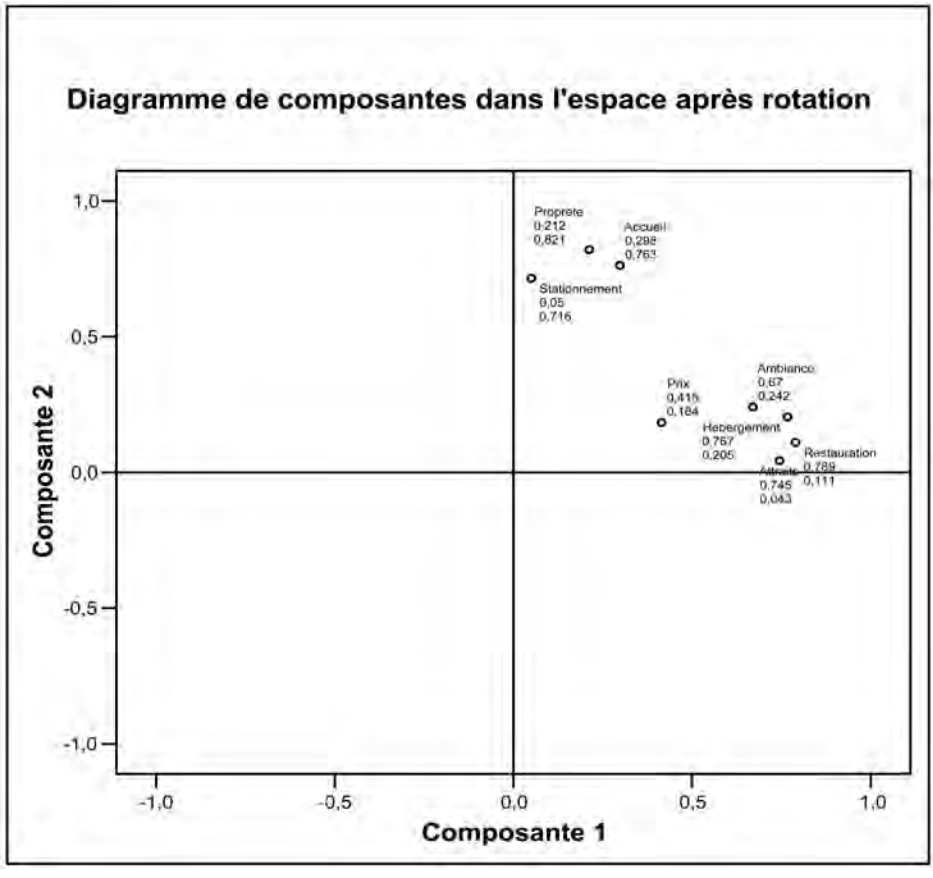
TABLEAU H



Matrice de tranformation des composantes

Composante	1	2
1	,807	,590
2	-,590	,807

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.



Le test de Kaiser-Meyer-Olkin est une mesure généralisée de la corrélation partielle entre les variables de l'étude. Cette mesure est basée sur la moyenne des coefficients de corrélation qui sont situés dans la diagonale de la matrice anti-image. La formule est[13] :

$$KMO = \frac{\sum\sum_{i \neq 1} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

La lecture du test *KMO* se fait de la façon suivante :

- 0,90 et plus = très grande validité ;
- 0,89 à 0,80 = grande validité ;
- 0,79 à 0,70 = validité moyenne ;
- 0,69 à 0,60 = validité faible ;
- 0,59 à 0,50 = validité au seuil limite ;
- 0,49 et moins = invalide.

Dans l'analyse factorielle en composantes principales de l'Auberge Frontenac (voir l'encadré 3.1, tableau C), la validité (test KMO) est donc moyenne (avec 0,79).

Le dernier test, le test de sphéricité de Bartlett, « permet de juger de l'inégalité des racines latentes, c'est-à-dire de l'absence significative de sphéricité du modèle mentionné. Si le modèle s'avère sphérique, on peut présumer que les corrélations entre les variables sont voisines de zéro et donc qu'il n'y a pas intérêt à remplacer les variables par des composantes[14]. » Le test de Bartlett est un test d'hypothèse, une forme approchée du khi carré. Le calcul se fait à partir du rapport *r* de la moyenne géométrique à la moyenne arithmétique des valeurs propres ; les formules utilisées sont[15] :

$$X^2 = -(n - v - \tfrac{1}{2}) \ln r^3$$

où les degrés de liberté sont :

$$DL = \tfrac{1}{2} (t - k + 2) (t - k - 1)$$

13. M. Norusis (1992), *SPSS for Windows Professional Statistics*, Chicago, SPSS, p. 52.
14. H. Laforge (1981), *Analyse multivariée*, Saint-Laurent, Études vivantes, p. 173.
15. *Ibid.*, p. 174.

où :

v = nombre de variables ;

t = nombre de composantes prises en compte ;

k = nombre de composantes non prises en compte ;

n = nombre de cas.

On doit donc formuler deux hypothèses :

H_0 = la matrice de corrélation est égale à une matrice identité[16] où, par exemple, une matrice 3×3 =

$$I\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

À toutes fins pratiques, cela signifie que les corrélations entre variables sont faibles ou nulles.

H_1 = que la matrice de corrélation est différente d'une matrice identité et qu'il est justifié de rechercher des composantes (facteurs).

Dans ce test, il est inutile de faire appel à une table de décision. On doit considérer seulement la signification du test. La valeur observée doit être égale ou inférieure à 0,05. Dans le cas de l'Auberge Frontenac (voir l'encadré 3.1, tableau C), la signification est égale à 0,000 ; ce qui signifie que l'hypothèse H_0 est rejetée et qu'il faut accepter H_1. Nous pouvons donc poursuivre l'étude des composantes principales de la satisfaction.

2.3. L'ÉTUDE DES PRINCIPAUX RÉSULTATS DE L'ANALYSE EN COMPOSANTES PRINCIPALES

Dans l'encadré 3.1, nous avons les principaux résultats de l'analyse factorielle ; voyons ces tableaux :

- au tableau A, nous avons la moyenne, l'écart-type, le nombre des répondants et les valeurs manquantes pour chacun des items (question) de l'échelle de satisfaction ;

16. Voir à ce sujet : V. Papillon (1993), *Vecteurs, matrices et nombres complexes*, Montréal, Modulo.

- le tableau B présente la matrice des corrélations entre les variables avec le déterminant de la matrice ; on remarque ici que les variables accueil, stationnement et propreté ont des coefficients de corrélation plus faibles avec les autres variables ;
- au tableau C, nous avons le test KMO avec la valeur 0,79 et le test de Bartlett avec le niveau de signification du test 0,000 ;
- le tableau D donne la qualité de la représentation (poids variables) pour chacune des variables ;
- au tableau E, nous avons la variance totale expliquée (inertie ou valeurs propres) de chacun des facteurs ;
- au tableau F, nous avons la matrice après les saturations, mais avant la rotation ; la sommation des coefficients au carré pour chacune des variables donne les poids variables (communauté) qui apparaissent dans le tableau D par exemple, pour la restauration (tableau F) $0{,}703^2 + -0{,}376^2 = 0{,}635$ (voir le résultat dans le tableau D) ; la sommation des coefficients pour chacune des colonnes du tableau F, par exemple pour la première composante (facteur) : $0{,}74^2 + 0{,}703^2 + 0{,}691^2 + 0{,}684^2 + 0{,}655^2 + 0{,}627^2 + 0{,}444^2 + 0{,}463^2 = 3{,}22$; ce total, le poids facteur (variance expliquée), apparaît dans la première ligne, deuxième colonne du tableau E ;
- au tableau G, nous avons la matrice des composantes après la procédure de rotation varimax (nous verrons dans un instant comment déterminer les composantes) ;
- Tableau H : la matrice de transformation des composantes.

2.4. Le choix des composantes et la présentation du tableau des composantes principales

Le tableau G de l'encadré 3.1 est très important, car c'est à partir de ces données que nous parviendrons à l'étape ultime (et gratifiante) de notre analyse de la satisfaction.

Le tableau 3.5 illustre la façon de présenter les composantes.

Tableau 3.5

L'ANALYSE EN COMPOSANTES PRINCIPALES DE LA SATISFACTION DES CLIENTS DE L'AUBERGE FRONTENAC PAR LA MÉTHODE DE ROTATION VARIMAX

Composantes et variables 1	*Coefficients* 2	*Variance en %*	
		*Réelle** 3	*Interne* 4
Composante I : Les qualités maîtresses de l'Auberge			
1. Restauration	0,789	40,2	72,5
2. Hébergement	0,767		
3. Attrait/site	0,745		
4. Ambiance	0,670		
5. Prix	0,415		
Composante II : Les qualités « techniques » de l'Auberge			
1. Propreté	0,821	15,2	27,5
2. Accueil	0,763		
3. Stationnement	0,716		
TOTAL :		55,4	100

* Nous présentons dans cette colonne la variance exprimée avant la rotation des composantes (voir le tableau E de l'encadré 3.1).

Le tableau 3.5 dresse le bilan final de l'analyse en composantes principales de la satisfaction des clients de l'Auberge Frontenac (c'est souvent le seul tableau, en ce qui concerne l'analyse en composantes principales, qui est présenté dans le rapport de recherche).

Voici comment lire ce tableau :

- dans la colonne 1, nous avons les composantes et les variables ; la composante I regroupe les variables qui sont les plus importantes dans la satisfaction des clients ;
- dans la colonne 2, nous avons les coefficients placés par ordre de grandeur ;
- dans la colonne 3 est présentée la variance expliquée par le modèle factoriel ; ainsi, la composante I représente 40,2 % de la variance et la composante II, 15,2 % ; le modèle lui-même « explique » à 55,4 % la satisfaction des usagers : cela veut dire

que 44,6 % de la satisfaction des clients reste inexpliquée par les variables choisies ; dans ce cas, il aurait fallu ajouter de nouvelles variables (ce qui rend plus difficile la collecte des données sur le terrain) ;

- dans la dernière colonne, nous avons la variance interne, qui se calcule comme suit : 40,2/55,4 = 0,725 × 100 = 72,5 % et 15,2/55,4 = 0,275 × 100 = 27,5 %.

La plupart des logiciels statistiques adoptent, pour la sélection des facteurs, le « critère de Kaiser » ; ce critère conserve les facteurs dont la variance expliquée (valeur propre ou inertie) est égale ou supérieure à 1[17] ; nous pouvons voir au tableau E de l'encadré 3.1 que la composante 3 a une valeur de 0,907 ; cette composante est donc rejetée, ainsi que celles qui suivent (de 4 à 8). Le graphique des valeurs propres (un graphique situé au-dessus du tableau F dans l'encadré 3.1) ne fait que reproduire les deux premières colonnes du tableau E (de l'encadré 3.1). Ce graphique[18] présente, en abscisse, les valeurs propres (variance expliquée ou inertie) et, en ordonnée, les composantes : sous la barre de 1,0, nous avons les composantes-variables rejetées.

L'analyse des composantes principales de la satisfaction des clients de l'Auberge Frontenac tend à montrer (voir le tableau 3.5) que « l'image » de l'Auberge repose essentiellement sur certaines qualités : la restauration, l'hébergement, l'attrait du site et, dans une moindre mesure, l'ambiance et le prix. Il s'agit donc de maintenir le niveau de ces qualités (de la première composante) dans la gestion quotidienne et de s'en servir, si besoin est, dans la promotion et la publicité.

Il y a une hiérarchie entre les composantes et une autre entre les variables d'une même composante. Par exemple, dans la première composante, la variable prix a un coefficient assez bas par rapport aux autres variables. On peut constater que la deuxième composante est moins importante pour l'Auberge Frontenac. Nous pouvons penser que les variables propreté, accueil et stationnement sont des qualités nécessaires, mais non déterminantes de la satisfaction de la clientèle.

---

17. Voir à ce sujet : G. Ferguson (1971), *Statistical Analysis in Psychology*, New York, McGraw-Hill, p. 421-425.

18. R. Cattell a développé une manière assez complexe, bien qu'en partie intuitive, de choisir les composantes à partir de graphiques ; voir à ce sujet : R. Cattell (1952), *Factor Analysis*, New York, Harper.

2.5. Les diverses méthodes de rotation

Le logiciel SPSS propose plusieurs méthodes de rotation des axes. Une première famille, la méthode de rotation orthogonale, suppose une certaine indépendance des composantes entre elles. La deuxième famille, la rotation oblique, part de l'hypothèse qu'il existe des liens entre les facteurs, par exemple entre la perception de l'espace et la mémoire dans un test d'intelligence. Fondamentalement, dans les faits, les composantes sont toujours liées d'une certaine façon et l'hypothèse de l'indépendance des composantes n'est qu'une manière commode de cerner la structure factorielle des variables[19].

Dans la famille de rotation orthogonale, nous avons les méthodes suivantes : Varimax, Quartimax et Equamax. Dans la méthode varimax[20], on tend à maximaliser les poids facteurs, souvent aux dépens des poids variables ; il s'agit d'augmenter la somme des variances des carrés des poids facteurs. Cette approche accorde une plus grande importance aux colonnes qu'aux lignes du tableau. Certaines saturations seront plus élevées sur une composante que sur les autres ; par cette méthode, on parvient à une structure des composantes plus tranchée, donc plus facile à lire et à interpréter. C'est la méthode de rotation la plus utilisée en recherche appliquée.

La méthode de rotation Quartimax vise, au contraire de la méthode Varimax, à augmenter sensiblement le poids des variables. On donne une plus grande place aux poids variables qu'aux poids facteurs, ce qui diminue la possibilité d'obtenir plusieurs coefficients élevés pour chacune des variables. Cette méthode tend à augmenter les poids variables indûment sur une composante en négligeant les autres. Dans ces conditions et dans certains cas, la structure finale est plus difficile à trouver avec cette méthode.

La méthode Equamax est un savant mélange des méthodes Varimax et Quartimax. Comme le souligne H. Laforge « elle vise à la simplification simultanée des lignes et des colonnes de la matrice des saturations[21] ». Le principal résultat est souvent de réduire les poids facteurs sur la ou les premières composantes.

---

19. Voir à ce sujet : J. Van de Geer (1971), *Introduction to the Multivariate Analysis for the Social Sciences*, San Francisco, W.H. Freeman, p. 152.
20. D'après J.-H. Kaiser, méthode mise au point en 1958.
21. H. Laforge (1981), *Analyse multivariée*, Saint-Laurent, Études vivantes, p. 184.

Dans la famille de rotation oblique, on compte surtout deux méthodes : Oblimin et Promax. Cette approche repose sur des hypothèses de relations entre les composantes qu'il faut formuler au préalable ; elle convient bien à une étude confirmatoire. Dans la méthode Direct Oblimin, il faut définir le paramètre delta et dans la méthode Promax, il faut aussi donner une valeur au paramètre kappa. Ce sont des méthodes complexes et exigeantes qui supposent un cadre théorique très étoffé.

Dans les tableaux 3.6 à 3.9, nous avons les rotations effectuées avec les différentes méthodes définies plus haut.

Tableau 3.6

LA MÉTHODE QUARTIMAX

	Composantes	
	1	***2***
Restauration	0,796	
Hébergement	0,782	
Attrait/site	0,746	
Ambiance	0,689	
Prix	0,430	
Propreté		0,799
Accueil		0,734
Stationnement		0,708

Tableau 3.7

LA MÉTHODE EQUAMAX

	Composantes	
	1	***2***
Restauration	0,789	
Hébergement	0,767	
Attrait/site	0,745	
Ambiance	0,670	
Prix	0,415	
Propreté		0,821
Accueil		0,763
Stationnement		0,716

Tableau 3.8

LA MÉTHODE OBLIMIN

	Composantes	
	1	***2***
Restauration	0,795	
Hébergement	0,793	
Attrait/site	0,737	
Ambiance	0,706	
Prix	0,444	
Propreté		0,846
Accueil		0,806
Stationnement		0,712

Tableau 3.9

LA MÉTHODE PROMAX

	Composantes	
	1	***2***
Restauration	0,795	
Hébergement	0,793	
Attrait/site	0,737	
Ambiance	0,706	
Prix	0,444	
Propreté		0,847
Accueil		0,811
Stationnement		0,708

Nous avons conservé dans ces tableaux, pour l'étude de la structure finale des composantes après rotation, l'exemple de l'étude de la satisfaction des clients de l'Auberge Frontenac (pour la méthode de rotation Varimax, voir le tableau 3.5).

Dans ces six tableaux (3.5 à 3.9), la structure factorielle reste inchangée en ce sens que ce sont les mêmes variables, dans le même ordre, qui se retrouvent dans les composantes. La structure par la méthode Varimax et la structure par la méthode Equamax sont identiques ; de plus, les coefficients sont les mêmes pour chacune des variables.

Dans la structure Quartimax (tableau 3.6), nous pouvons constater un certain affaiblissement des coefficients dans le deuxième facteur. Pour toutes ces méthodes, seuls les coefficients attachés à chacune des variables changent quelque peu.

Nous avons vu dans cette partie du chapitre 3 les principales commandes de l'analyse en composantes principales (avec un exemple) avec le logiciel SPSS Windows. Dans celle qui suit, nous allons étudier un exemple de l'analyse factorielle en composantes principales.

3. UN EXEMPLE D'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES

Dans cette partie, nous allons voir un exemple concret d'analyse factorielle en composantes principales. Cet exemple porte sur l'intérêt à suivre des cours de formation chez les employés d'une chaîne hôtelière. Dans le questionnaire soumis aux employés, douze cours étaient proposés :

- cours de langue ;
- cours de gestion ;
- cours portant sur les métiers de l'hôtellerie ;
- cours sur l'accueil ;
- cours de cuisine ;
- cours de relations humaines ;
- cours sur les attraits touristiques régionaux ;
- cours d'informatique ;
- cours sur la restauration ;
- cours sur l'industrie touristique ;
- cours sur les destinations touristiques ;
- cours sur la gestion du personnel.

Le répondant devait répondre, pour chacun des cours, en fonction de l'échelle suivante :

1. Pas du tout intéressé ;
2. Peu intéressé ;
3. Intéressé ;
4. Très intéressé.

Les principaux résultats de cette analyse factorielle en composantes principales sont donnés dans l'encadré 3.2.

Encadré 3.2

L'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES DE L'INTÉRÊT À SUIVRE DES COURS CHEZ LES EMPLOYÉS D'UNE CHAÎNE HÔTELIÈRE

TABLEAU A

Statistiques descriptives

	Moyenne	Ecart-type	n analyse
Cours de langue	2,8293	1,10408	287
Cours de gestion	2,3449	1,13872	287
Métiers de l'hôtellerie	2,3833	1,09965	287
Accueil	2,3136	1,10905	287
Cuisine	1,9373	1,09492	287
Relations humaines	2,6446	1,12776	287
Attraits touristiques	2,5052	1,12775	287
Informatique	2,4983	1,18815	287
Restauration	2,2056	1,09172	287
Industrie touristique	2,3937	1,07500	287
Destinations touristiques	2,4983	1,11842	287
Gestion du personnel	2,5087	1,18222	287

Matrice de corrélation[a]

		Cours de langue	Cours de gestion	Métiers de l'hôtellerie	Accueil	Cuisine	Relations humaines	Attraits touristiques	Informatique	Restauration	Industrie touristique	Destinations touristiques	Gestion du personnel
Corrélation	Cours de langue	1,000	,400	,414	,341	,196	,428	,421	,412	,287	,419	,466	,316
	Cours de gestion	,400	1,000	,310	,268	,219	,371	,302	,563	,235	,389	,345	,521
	Métiers de l'hôtellerie	,414	,310	1,000	,497	,409	,316	,416	,290	,624	,475	,458	,361
	Accueil	,341	,268	,497	1,000	,186	,458	,549	,279	,368	,500	,488	,395
	Cuisine	,196	,219	,409	,186	1,000	,172	,198	,150	,640	,199	,214	,195
	Relations humaines	,428	,371	,316	,458	,172	1,000	,529	,477	,315	,543	,490	,606
	Attraits touristiques	,421	,302	,416	,549	,198	,529	1,000	,461	,319	,701	,709	,425
	Informatique	,412	,563	,290	,279	,150	,477	,461	1,000	,231	,514	,481	,531
	Restauration	,287	,235	,624	,368	,640	,315	,319	,231	1,000	,384	,380	,371
	Industrie touristique	,419	,389	,475	,500	,199	,543	,701	,514	,384	1,000	,761	,579
	Destinations touristiques	,466	,345	,458	,488	,214	,490	,709	,481	,380	,761	1,000	,537
	Gestion du personnel	,316	,521	,361	,395	,195	,606	,425	,531	,371	,579	,537	1,000
Signification (unilatérale)	Cours de langue		,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	Cours de gestion	,000		,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	Métiers de l'hôtellerie	,000	,000		,000	,000	,000	,000	,000	,000	,000	,000	,000
	Accueil	,000	,000	,000		,001	,000	,000	,000	,000	,000	,000	,000
	Cuisine	,000	,000	,000	,001		,002	,000	,005	,000	,000	,000	,000
	Relations humaines	,000	,000	,000	,000	,002		,000	,000	,000	,000	,000	,000
	Attraits touristiques	,000	,000	,000	,000	,000	,000		,000	,000	,000	,000	,000
	Informatique	,000	,000	,000	,000	,005	,000	,000		,000	,000	,000	,000
	Restauration	,000	,000	,000	,000	,000	,000	,000	,000		,000	,000	,000
	Industrie touristique	,000	,000	,000	,000	,000	,000	,000	,000	,000		,000	,000
	Destinations touristiques	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000		,000
	Gestion du personnel	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	

[a] Déterminant = ,002

TABLEAU B

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,873
Test de sphéricité de Bartlett	Khi-deux approximé	1727,484
	ddl	66
	Signification de Bartlett	,000

TABLEAU C

Qualité de représentation

	Initial	Extraction
Cours de langue	1,000	,401
Cours de gestion	1,000	,748
Métiers de l'hôtellerie	1,000	,662
Accueil	1,000	,592
Cuisine	1,000	,759
Relations humaines	1,000	,564
Attraits touristiques	1,000	,748
Informatique	1,000	,699
Restauration	1,000	,812
Industrie touristique	1,000	,751
Destinations touristiques	1,000	,736
Gestion du personnel	1,000	,628

Méthode d'extraction : Analyse en composantes principales.

TABLEAU D

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	5,575	46,459	46,459	5,575	46,459	46,459	3,393	28,272	28,272
2	1,477	12,305	58,763	1,477	12,305	58,763	2,583	21,525	49,797
3	1,049	8,739	67,502	1,049	8,739	67,502	2,125	17,705	67,502
4	,738	6,150	73,652						
5	,641	5,338	78,991						
6	,576	4,799	83,789						
7	,483	4,025	87,814						
8	,410	3,418	91,232						
9	,322	2,684	93,916						
10	,262	2,180	96,096						
11	,246	2,054	98,149						
12	,222	1,851	100,000						

Méthode d'extraction : Analyse en composantes principales.

TABLEAU E

Matrice des composantes[a]

	Composante		
	1	2	3
Industrie touristique	,822	-,164	-,218
Destinations touristiques	,804	-,137	-,265
Attraits touristiques	,766	-,161	-,368
Gestion du personnel	,730	-,193	,241
Relations humaines	,715	-,228	,016
Métiers de l'hôtellerie	,673	,446	-,096
Informatique	,670	-,327	,379
Accueil	,663	,050	-,386
Cours de langue	,623	-,061	,093
Cuisine	,414	,727	,245
Restauration	,604	,665	,062
Cours de gestion	,595	-,179	,602

Méthode d'extraction : Analyse en composantes principales.

a. 3 composantes extraites.

TABLEAU F

Matrice des composantes après rotation[a]

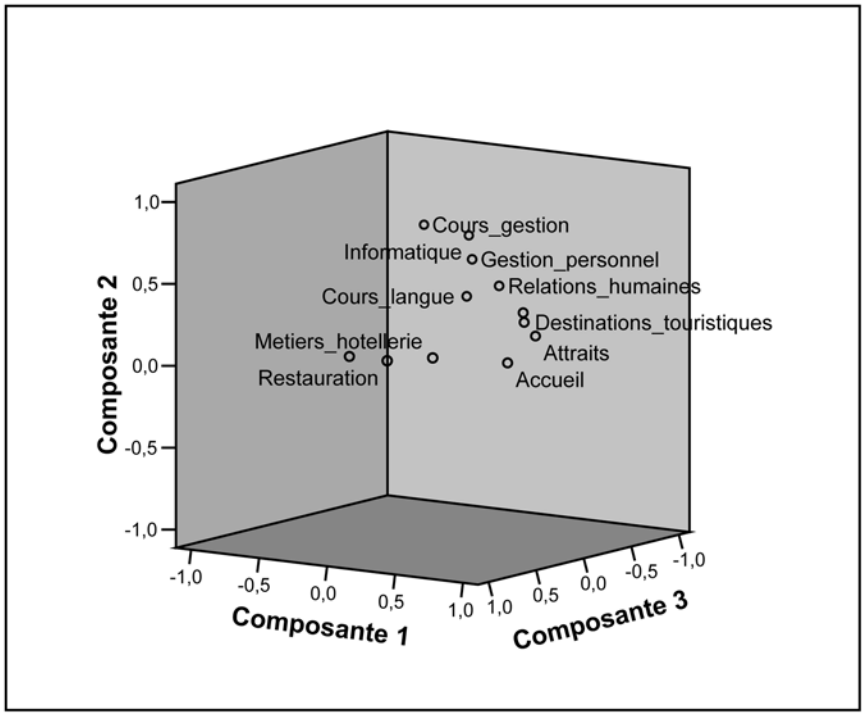
	Composante		
	1	2	3
Attraits touristiques	,827	,232	,102
Destinations touristiques	,781	,319	,156
Industrie touristique	,768	,374	,146
Accueil	,724	,079	,248
Relations humaines	,547	,507	,084
Cours de gestion	,059	,845	,176
Informatique	,291	,783	,037
Gestion du personnel	,400	,666	,158
Cours de langue	,397	,446	,210
Cuisine	-,010	,132	,861
Restauration	,261	,132	,852
Métiers de l'hôtellerie	,459	,141	,656

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.

a. La rotation a convergé en 6 itérations.

Le tableau A présente les statistiques descriptives ; la moyenne de la valeur la plus élevée (2,8293) va au cours de langue, la moyenne de la variable la plus faible (1,9373) va au cours de cuisine. Le déterminant de la matrice est de 0,002146 ; le test KMO est de 0,873 et la signification statistique du test de Bartlett est de 0,000 (voir le tableau B).

Diagramme de composantes dans l'espace après rotation



Le tableau C présente la qualité de la représentation. Au tableau D, nous avons les poids des composantes :

- la première composante exprime 46,4 % de la variance expliquée ;
- la deuxième composante, 12,3 % de cette variance ;
- et la dernière composante, 8,7 % de la variance.

Ce modèle factoriel « explique » 67,5 % de l'ensemble de l'intérêt à suivre des cours de formation. Il reste donc 32,5 de la variance du modèle qui reste inexpliquée ; il faudrait donc, pour accroître cette variance, ajouter des cours, mais lesquels ?

Le tableau E présente la matrice des composantes avant la rotation. Dans le tableau F apparaît la structure factorielle finale avec les trois composantes et les coefficients les plus élevés pour chacune des variables

(après la rotation Varimax). Enfin, un graphique à trois dimensions représente les trois composantes.

3.1. LA PRÉSENTATION DES RÉSULTATS

Il faut tout d'abord présenter les résultats de la structure factorielle finale ; ces résultats apparaissent au tableau 3.10.

Tableau 3.10

L'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES DE L'INTÉRÊT À SUIVRE DES COURS CHEZ LES EMPLOYÉS D'UNE CHAÎNE HÔTELIÈRE (ROTATION VARIMAX)

Composantes et variables	*Coefficients*	*Variance en %*	
		*Réelle**	*Interne*
Composante I : Le volet « tourisme »			
1. Attraits touristiques	,827	46,4	68,8
2. Destinations touristiques	,781		
3. Industrie touristique	,768		
4. Accueil	,724		
5. Relations humaines	,547		
Composante II : Le volet « gestion »			
1. Cours de gestion	,845	12,3	18,2
2. Informatique	,783		
3. Gestion du personnel	,666		
4. Cours de langue	,446		
Composante III : Le volet « hôtellerie »			
1. Cuisine	,861	8,7	13
2. Restauration	,852		
3. Métiers de l'hôtellerie	,656		
TOTAL :		67,4	100

* Nous présentons, dans cette colonne, la variance exprimée avant la rotation des composantes (voir le tableau D de l'encadré 3.2).

Il faut noter que le tableau 3.10 est construit à partir des données du tableau F de l'encadré 3.2. L'analyse en composantes principales a dégagé trois facteurs. La première composante regroupe les variables qui touchent le tourisme au sens large du terme. Il y a un ordre hiérarchique entre les variables de la première composante : les attraits, suivis par

les destinations, l'industrie et l'accueil ; la variable « cours de relations humaines » ne cadre pas très bien avec les autres ; aussi son coefficient est plus faible que les autres. La première composante « explique » près de 69 % de la variance interne, ce sont donc des cours d'un grand intérêt pour les répondants.

La deuxième composante réunit les cours de gestion, à l'exception du cours de langue, qui détonne dans cet ensemble ; le coefficient du cours de langue est plus faible que les autres avec 0,446. Cette deuxième composante « explique » 18,2 % de l'intérêt accordé aux cours de formation par cette population.

Enfin, la troisième composante contient les cours de formation propres à l'hôtellerie ; on peut penser que c'est une composante marginale. Avec cette structure factorielle, les responsables de la formation des employés possèdent des éléments pour établir les programmes de cours.

L'étape suivante de l'analyse des données dans le cadre de l'analyse factorielle en composantes principales de l'intérêt à suivre des cours chez les employés d'une chaîne hôtelière est de considérer les tests statistiques. Ces tests sont présentés dans le tableau 3.11.

Tableau 3.11

LES RÉSULTATS DE DIVERS TESTS POUR L'ANALYSE EN COMPOSANTES PRINCIPALES DE L'INTÉRÊT À SUIVRE DES COURS CHEZ LES EMPLOYÉS D'UNE CHAÎNE HÔTELIÈRE

Déterminant de la matrice de corrélation	*Test de Kaiser-Meyer-Olkin*	*Test de Bartlett*
.002146	.873	.000

Au tableau 3.11, nous voyons que le déterminant de la matrice de corrélation est relativement faible sans être égal à zéro (,002146) ; la structure factorielle présentée au tableau 3.10 passe donc le premier test. Le test KMO est égal à ,873, ce qui indique une grande validité. Enfin, le test de Bartlett nous indique que l'hypothèse H_0 doit être rejetée (le résultat du test doit être égal ou inférieur à 0,05) ; la recherche des composantes est donc justifiée (voir le tableau B de l'encadré 3.2).

3.2. L'ÉTUDE DE LA VALIDITÉ DE L'ÉCHELLE DE MESURE

Dans certains cas, il est important de bien mesurer la validité de l'échelle de mesure que nous utilisons. Il s'agit de savoir si l'échelle utilisée est cohérente et si toutes les variables présentes dans l'échelle jouent un rôle dans l'explication du problème étudié. Les tests de validité ne sont que des indicateurs parmi d'autres. En recherche appliquée, il est assez rare que l'on utilise des échelles au sens strict du terme. Les tests de validité deviennent très importants lorsqu'une mesure est répétée dans le temps et qu'elle permet de discriminer entre certains comportements : les tests d'intelligence, la mesure du stress, les tests de sélection, etc.

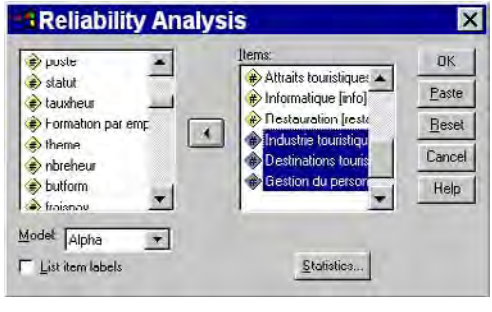
Le logiciel SPSS propose cinq modèles d'analyse de validité des échelles utilisées :

- le test « Alpha » de Cronbach a pour objectif de mesurer la cohérence interne de l'échelle ; il repose sur les corrélations moyennes entre les variables (ou items) contenues dans l'échelle ;
- le test « Split-half » sépare l'échelle en deux et présente la corrélation dans les deux parties ;
- le test de Gutmann donne plusieurs mesures de la validité de l'échelle en utilisant six paramètres lambda ; il compare les limites minimales à une validité expérimentale ;
- le test « ParallelX » correspond à un test d'hypothèse ; on suppose que les variables ont des variances quasi égales et des résidus semblables ;
- le test « Strictly Parallel » est une variante du test « Parallel » ; ici, l'hypothèse est que les variables ont les mêmes moyennes.

Les commandes avec SPSS sont les suivantes :

Figure 3.9

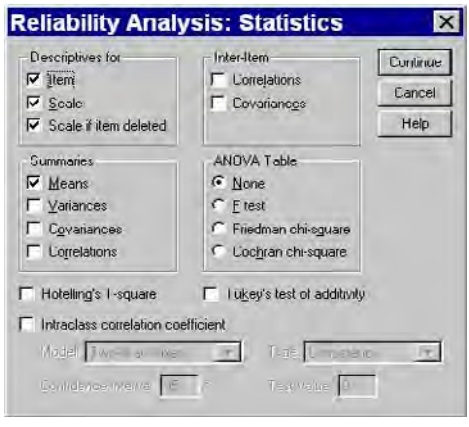
LA BOÎTE DE DIALOGUE PRINCIPALE DE L'ANALYSE DE LA VALIDITÉ



Nous voyons à la figure 3.9 la boîte de dialogue principale de l'analyse de la validité. Le rectangle de gauche contient les variables sources (celles du questionnaire) ; dans le rectangle du milieu, nous avons les variables (items) qui font partie de l'échelle de mesure de l'intérêt à suivre des cours. Le petit rectangle de gauche permet de choisir la méthode que l'on veut utiliser ; ici, la méthode « Alpha » de Cronbach. C'est la méthode la plus simple et la plus courante.

Figure 3.10

LA BOÎTE DE DIALOGUE DES STATISTIQUES POUR L'ANALYSE DE VALIDITÉ



À la figure 3.10, nous avons les statistiques disponibles pour l'analyse de la validité ; ces statistiques sont nombreuses et variées. Nous avons :

- les statistiques descriptives des items ;
- les principales statistiques (moyennes, variances, covariances et corrélations) ;
- la cohérence inter-items qui produit des matrices des corrélations et des covariances entre les variables ;
- **ANOVA TABLE**, qui propose des tests de moyennes (le test F de Fisher, le khi-carré de Friedman et le khi-carré de Cochran) ;
- des études interclasses ainsi que les tests de Hotelling et de Tukey sont également disponibles.

L'encadré 3.3 présente les principaux résultats de l'étude de validité de l'échelle de la mesure de l'intérêt à suivre des cours chez les employés d'une chaîne hôtelière.

Encadré 3.3

L'ÉTUDE DE LA VALIDITÉ DE L'ÉCHELLE DE LA MESURE DE L'INTÉRÊT À SUIVRE DES COURS CHEZ LES EMPLOYÉS D'UNE CHAÎNE HÔTELIÈRE

Statistiques de fiabilité

Alpha de Cronbach	Alpha de Cronbach basé sur des éléments normalisés	Nombre d'éléments
,891	,891	12

Statistiques sur les éléments

	Moyenne	Ecart-type	N
Cours de langue	2,8293	1,10408	287
Cours de gestion	2,3449	1,13872	287
Métiers de l'hôtellerie	2,3833	1,09965	287
Accueil	2,3136	1,10905	287
Cuisine	1,9373	1,09492	287
Relations humaines	2,6446	1,12776	287
Attraits touristiques	2,5052	1,12775	287
Informatique	2,4983	1,18815	287
Restauration	2,2056	1,09172	287
Industrie touristique	2,3937	1,07500	287
Destinations touristiques	2,4983	1,11842	287
Gestion du personnel	2,5087	1,18222	287

Statistiques récapitulatives sur les éléments

	Moyenne	Minimum	Maximum	Plage	Maximum/Minimum	Variance	Nombre d'éléments
Moyenne des éléments	2,422	1,937	2,829	,892	1,460	,049	12

La matrice de covariance est calculée et utilisée dans l'analyse.

Statistiques complètes sur les éléments

	Moyenne de l'échelle en cas de suppression d'un élément	Variance de l'échelle en cas de suppression d'un élément	Corrélation complète des éléments corrigés	Carré de la corrélation multiple	Alpha de Cronbach en cas de suppression de l'élément
Cours de langue	26,2334	71,187	,546	,364	,885
Cours de gestion	26,7178	71,210	,524	,430	,887
Métiers de l'hôtellerie	26,6794	70,135	,610	,524	,882
Accueil	26,7491	70,587	,578	,429	,884
Cuisine	27,1254	74,593	,359	,429	,895
Relations humaines	26,4181	69,370	,636	,507	,881
Attraits touristiques	26,5575	68,604	,680	,624	,878
Informatique	26,5645	69,491	,590	,477	,883
Restauration	26,8571	71,249	,550	,595	,885
Industrie touristique	26,6690	68,166	,747	,681	,875
Destinations touristiques	26,5645	67,960	,725	,669	,876
Gestion du personnel	26,5540	68,388	,655	,566	,880

Statistiques d'échelle

Moyenne	Variance	Ecart-type	Nombre d'éléments
29,0627	82,583	9,08754	12

Dans cet encadré nous avons cinq tableaux :

1. les principales statistiques de fiabilité (dont le fameux alpha de Cronbach) ;
2. les statistiques descriptives des variables étudiées ;
3. les statistiques récapitulatives sur les éléments ;
4. les statistiques complètes sur les éléments ;
5. les statistiques d'échelle (score).

Dans le quatrième tableau, nous avons les principaux éléments du test alpha de Cronbach. Dans la deuxième colonne, « Moyenne de l'échelle en cas de suppression d'un élément », nous avons la moyenne du score (la moyenne de l'échelle est de 29,0627) quand une variable est enlevée de l'ensemble de l'échelle. Par exemple, pour la première variable « Cours de langue », nous avons la moyenne 29,0627 – la moyenne de cette variable (voir le deuxième tableau de l'encadré 3.3), 2,8293 = 26,2334. Dans la troisième colonne, nous faisons les mêmes calculs avec la variance (on élimine à chaque fois une variable).

La quatrième colonne, « Corrélation complète des éléments corrigés », correspond à la corrélation d'un item (variable) avec la somme des scores de ceux qui restent (la somme des scores des autres variables). La cinquième colonne, « Carré de la corrélation multiple », indique la variance d'une variable expliquée par les autres variables de l'échelle. Par exemple, la variable « Cours de gestion » doit 0,430 % de sa variance expliquée aux autres variables de l'échelle.

Enfin, la dernière colonne contient les fameux coefficients alpha. Ces coefficients se lisent en tenant compte du coefficient global alpha qui apparaît à la dernière ligne de l'encadré 3.3 ; ici ce coefficient est de 0,8914. Le coefficient global alpha doit être le plus élevé possible : plus il est élevé, plus la validité est forte.

Les coefficients de la sixième colonne nous informent de la valeur que prendrait le coefficient global alpha si cette variable n'était pas prise en compte. Par exemple, si la variable « Cuisine » était rejetée de l'échelle, le coefficient global alpha serait de 0,895 (il est de 0,891 en comptabi-

lisant cette variable : voir le tableau de la page 97). Donc, toute variable supérieure au coefficient global alpha devrait être théoriquement enlevée de l'échelle. Dans ce dernier cas, la différence n'est pas très grande avec ou sans la variable « Cuisine ».

La meilleure façon de procéder est d'enlever la variable dont le coefficient alpha est trop élevé (plus élevé que le coefficient alpha global) et d'effectuer une nouvelle étude de validité ; si le coefficient global change peu ou pas, nous pouvons conserver cette variable.

La lecture du test alpha de Cronbach se fait à deux niveaux :

- premièrement, le coefficient global alpha doit être élevé (supérieur à 0,60) ; plus il est élevé, plus la validité est forte ; ici le coefficient est égal à 0,891 ;
- deuxièmement, il permet d'évaluer l'importance de chacune des variables de l'échelle et, éventuellement, d'éliminer les variables peu compatibles avec l'ensemble.

Les tests de validité doivent être appliqués scrupuleusement quand on a affaire à une « véritable » échelle au sens psychosociologique (mesure du stress, de l'intelligence, de la satisfaction, etc.). Dans les autres cas, l'utilité de ces tests est beaucoup plus limitée. Dans notre dernier exemple, l'éventail des cours offerts peut dépendre du cadre institutionnel ou organisationnel ; il est donc impossible de proposer une panoplie complète de cours sans tenir compte des coûts et des besoins des individus et des entreprises ; la mesure porte sur des éléments empiriques et non sur des concepts abstraits.

Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

CHAPITRE

4

L'analyse factorielle des correspondances

L'analyse factorielle des correspondances est une méthode qui sert à représenter graphiquement un tableau croisé. Elle vise à réunir les informations les plus utiles de façon à donner une image claire de l'association de deux variables. Dans l'analyse des correspondances, les lignes représentent les catégories d'une première variable et les colonnes, les catégories d'une deuxième variable.

Dans l'analyse factorielle en composantes principales, les colonnes sont nécessairement des variables et les lignes, des individus ; les principaux résultats reposent sur les corrélations entre ces variables. Malgré certaines ressemblances, l'analyse factorielle des correspondances se démarque donc de l'analyse factorielle en composantes principales.

1. OBJECTIFS ET ASPECTS THÉORIQUES

Le principal objectif de l'analyse factorielle des correspondances est d'étudier simultanément, par le biais de leurs catégories, la relation entre deux variables. Il s'agit de présenter visuellement les principales liaisons entre les catégories des deux variables.

Ces liaisons sont analysées selon les oppositions :

- centre/périphérie ;
- éloignement/proximité ;
- ressemblance/dissemblance ;
- attraction/répulsion.

La carte des correspondances doit être interprétée en termes de territoire, de géographie de plan, où les distances entre les catégories expriment l'un ou l'autre des qualificatifs propres aux couples des oppositions.

La première étape consiste à établir les profils lignes et les profils colonnes. Ces profils se calculent (à partir des données brutes) en divisant chaque terme par le total de cette ligne ou de cette colonne. La deuxième étape consiste à mesurer les ressemblances/dissemblances entre les profils par la distance du khi-carré à partir de la formule de Pythagore (distance euclidienne)[1] :

$$s(i,i') = \sqrt{\sum_i \left(a_{ij} - a_{i'j}\right)^2}$$

cette formule devient :

$$d(i,i') = \sqrt{\sum_i \left(\frac{a_{ij} - a_{i'j}}{a}\right)^2}$$

La formule de distance du khi-carré servira donc à mesurer les systèmes d'opposition des éléments étudiés.

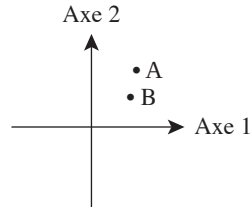
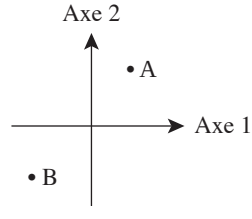
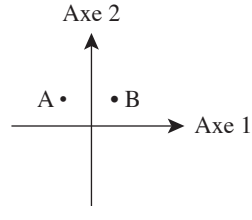
1. S.-E. Clausen (1998), *Applied Correspondence Analysis*, New York, Sage, p. 11-12. Voir aussi à ce sujet : B. Escofier et J. Pages (1998), *Analyses factorielles simples et multiples*, Paris, Dunod, p. 58-59.

L'analyse factorielle des correspondances permet aussi de définir des facteurs en fonction de la contribution à l'inertie d'une ligne ou d'une colonne ; le degré d'inertie correspond, en quelque sorte, à la variance expliquée. Selon Jean-Jacques Lambin : « Un facteur est retenu pour l'analyse s'il possède un taux d'inertie expliqué significativement supérieur à ce qu'apporte en moyenne une variable, c'est-à-dire 100 %/*p* si *p* est le nombre de colonnes du tableau de fréquences[2]. »

L'interprétation des résultats se fait en fonction de l'image projetée. La carte des résultats de l'analyse factorielle des correspondances se fait comme suit[3] :

Figure 4.1

LE TABLEAU DE LECTURE DES RÉSULTATS DE LA CARTE DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES

Attraction	
Répulsion	
Pas de conclusion	

2. J.-J. Lambin (1990), *La recherche marketing*, Paris, McGraw-Hill, p. 302.
3. Nous nous inspirons ici de : M. Tenenhaus (1994), *Méthodes statistiques en gestion*, Paris, Dunod, p. 171.

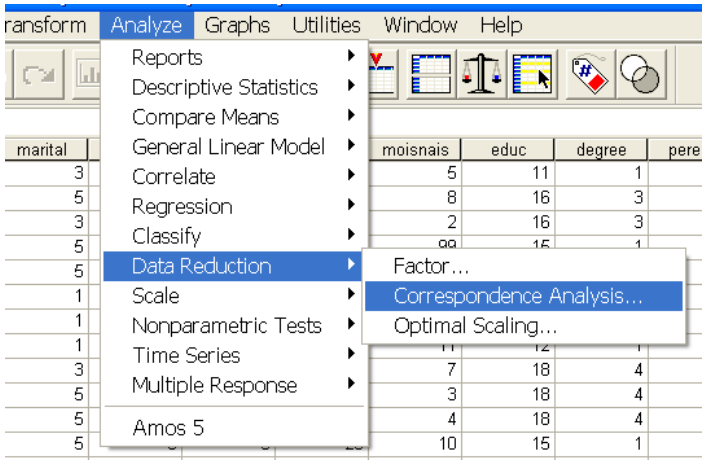
Comme l'indique cette table de lecture (figure 4.1), quand A et B sont dans le même quadrant, il y a attraction ; cela indique que les effectifs qui correspondent aux deux catégories sont plus nombreux que si les effectifs étaient distribués de façon proportionnelle. De la même façon, quand A et B sont dans des quadrants opposés, cela montre que les catégories de l'une ou l'autre des variables se repoussent. Quand A et B sont dans des quadrants adjacents, l'interprétation des résultats est plus difficile. Une concentration au centre de la carte représente la moyenne des catégories de chacune des variables impliquées. D'après Jean de Lagarde : « L'origine correspond au point neutre, c'est-à-dire à l'indépendance complète des deux caractères ou, en d'autres termes, à des proportions identiques dans chaque classe (ligne ou colonne)[4]. »

2. LES COMMANDES AVEC LE LOGICIEL SPSS

Le cheminement pour parvenir à la méthode de l'analyse des correspondances est illustré à la figure 4.2.

Figure 4.2

LE CHEMINEMENT POUR L'ANALYSE DES CORRESPONDANCES



Nous devons utiliser le menu **Analyze** ; dans le menu déroulant, on doit choisir **Data Reduction** puis **Correspondence Analysis**.

4. J. Lagarde (1995), *Initiation à l'analyse des données*, Paris, Dunod, p. 65.

Dans la fenêtre qui apparaît (figure 4.3), nous avons à gauche, dans un rectangle, les variables de la recherche ; à droite de la fenêtre, nous retrouvons deux petits rectangles intitulés **ROW** et **COLUMN** destinés à recevoir les variables du modèle. Sous ces rectangles, nous pouvons distinguer deux boutons **DEFINE RANGE** ; un clic sur un de ces boutons fait apparaître deux nouvelles fenêtre : **Correspondence Analysis : Define Row Range** et **Correspondence Analysis : Define Column Range** (figures 4.4 et 4.5).

Figure 4.3

LA BOÎTE PRINCIPALE POUR L'ANALYSE DES CORRESPONDANCES

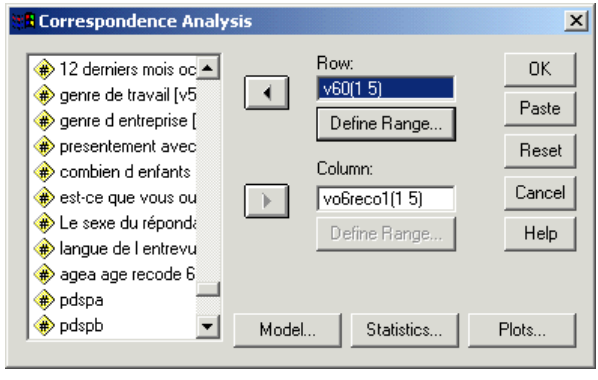


Figure 4.4

LA DÉFINITION DE LA PREMIÈRE VARIABLE

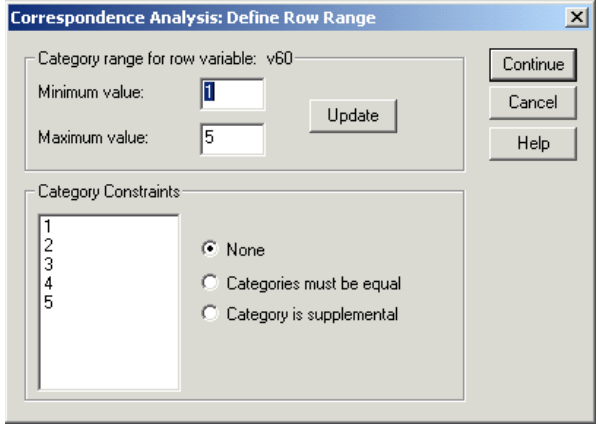
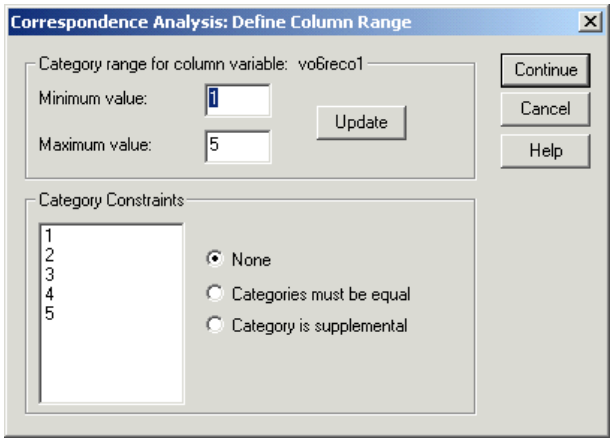


Figure 4.5

LA DÉFINITION DE LA DEUXIÈME VARIABLE



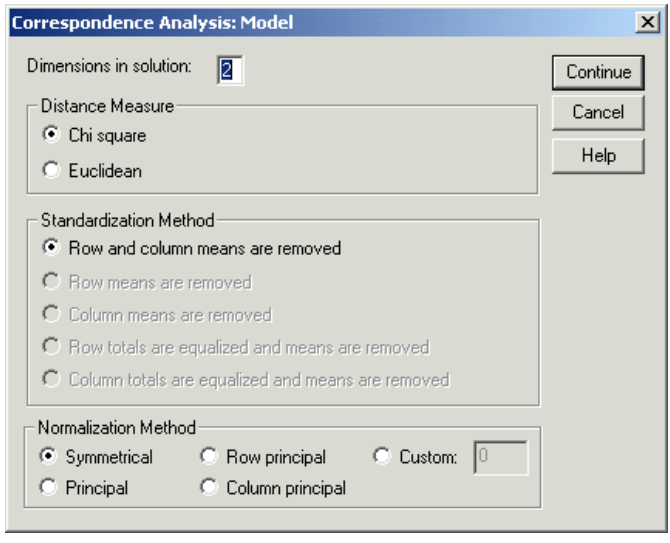
Il faut indiquer à ce moment le nombre de catégories de chacune des variables : la valeur la plus faible (**MINIMUM VALUE**, ici : 1) et la valeur la plus élevée (**MAXIMUM VALUE**, ici : 5) ; par la suite, il faut cliquer sur le bouton **UPDATE**. Dans ce modèle, la variable 60 correspond au revenu brut familial (avec cinq catégories) et la variable « v06reco » (avec cinq catégories) représente le nombre de semaines de vacances prises par le répondant.

Le bouton **CONTINUE** permet de retourner à la fenêtre principale **Correspondence Analysis** (figure 4.3). Nous pouvons remarquer, en bas de cette figure à gauche, trois autres boutons. En cliquant sur le bouton **MODEL**, nous voyons apparaître une nouvelle fenêtre **Correspondence Analysis : Model** (figure 4.6).

Cette fenêtre donne accès à quatre commandes différentes : **DIMENSIONS IN SOLUTION**, **DISTANCE MEASURE**, **STANDARDIZATION METHOD** et **NORMALIZATION METHOD**. Les choix par défaut apparaissent à la figure 4.6. En cliquant à nouveau sur le bouton **CONTINUE**, nous revenons à la fenêtre principale (figure 4.3).

Figure 4.6

LES STATISTIQUES



Le deuxième bouton, **STATISTICS…**, nous amène à une nouvelle fenêtre, **Correspondence Analysis : Statistics** ; ici plusieurs traitements statistiques sont possibles ; nous utiliserons seulement les trois principaux (voir la figure 4.7). Le dernier bouton sert à la construction de graphiques : **Correspondence Analysis : Plots** (voir la figure 4.8) : dans ce cas, la construction de cinq graphiques est possible. Dans tous les cas, il faut cliquer sur le bouton **CONTINUE** pour revenir à la fenêtre principale (figure 4.3). Pour exécuter l'ensemble des commandes contenues dans les figures, il faut cliquer sur le bouton **OK**.

Figure 4.7

LES PRINCIPAUX TABLEAUX

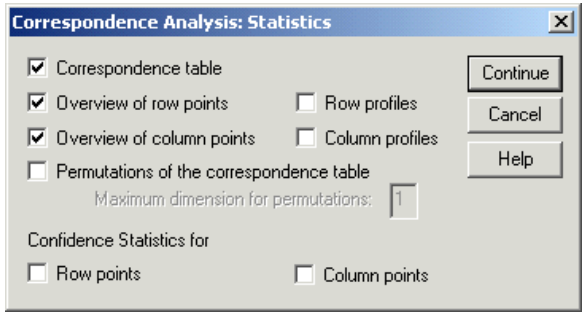
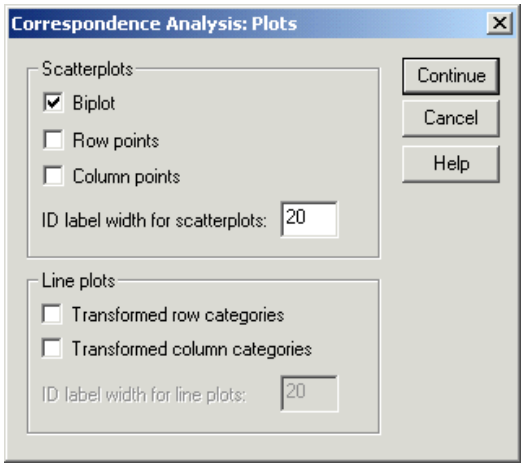


Figure 4.8

LES GRAPHIQUES



3. UN EXEMPLE D'ANALYSE

Notre analyse factorielle des correspondances repose sur la relation entre deux variables :

- le « revenu brut familial » (voir le tableau 4.1) ;
- le « nombre de semaines de vacances » (voir le tableau 4.2).

Dans ces deux tableaux, les variables ont chacune cinq catégories.

Tableau 4.1

LE REVENU BRUT FAMILIAL

	Pourcentage	*pourcentage cumulatif*
20 000 $ et –	29,0	29,0
20 001 – 40 000 $	33,5	62,5
40 001 – 60 000 $	22,2	84,7
60 001 – 80 000 $	8,5	93,2
80 001 $ et plus	6,8	100,0
Total	**100,0**	

Tableau 4.2

LE NOMBRE DE SEMAINES DE VACANCES

	Pourcentage	***pourcentage cumulatif***
1 semaine	16,2	16,2
2 semaines	26,4	42,6
3 semaines	17,3	59,8
4 semaines	20,5	80,3
5 semaines et plus	19,7	100,0
Total	**100,0**	

L'encadré 4.1 contient les principaux résultats de l'analyse factorielle des correspondances. Le premier tableau, intitulé Correspondence Table (tableau des correspondances), contient les données brutes qui résultent du croisement entre les deux variables choisies.

Encadré 4.1

LES RÉSULTATS DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES ; LE REVENU BRUT FAMILIAL ET LE NOMBRE DE SEMAINES DE VACANCES

Tableau des correspondances

	Nombre de semaines de vacances					
Revenu brut familial	1sem	2sem	3sem	4sem	5sem+	Marge active
20 000$ et -	26	26	10	7	8	77
20 001-40 000$	36	50	32	25	35	178
40 001-60 000$	13	45	29	42	33	162
60 001-80 000$	8	14	10	20	14	66
80 001$ et +	2	7	11	14	18	52
Marge active	85	142	92	108	108	535

Résumé

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie		Valeur singulière de confiance	
					Expliqué	Cumulé	Ecart-type	Corrélation 2
1	,314	,098			,864	,864	,040	,000
2	,104	,011			,095	,959	,046	
3	,068	,005			,041	1,000		
4	,004	,000			,000	1,000		
Total		,114	60,867	,000[a]	1,000	1,000		

a. 16 degrés de liberté

Source : Data Theory Scaling System Group (D75S), Faculty of Social and Behavioral Sciences, Leider University, The Netherlands

Caractéristiques des points lignes[a]

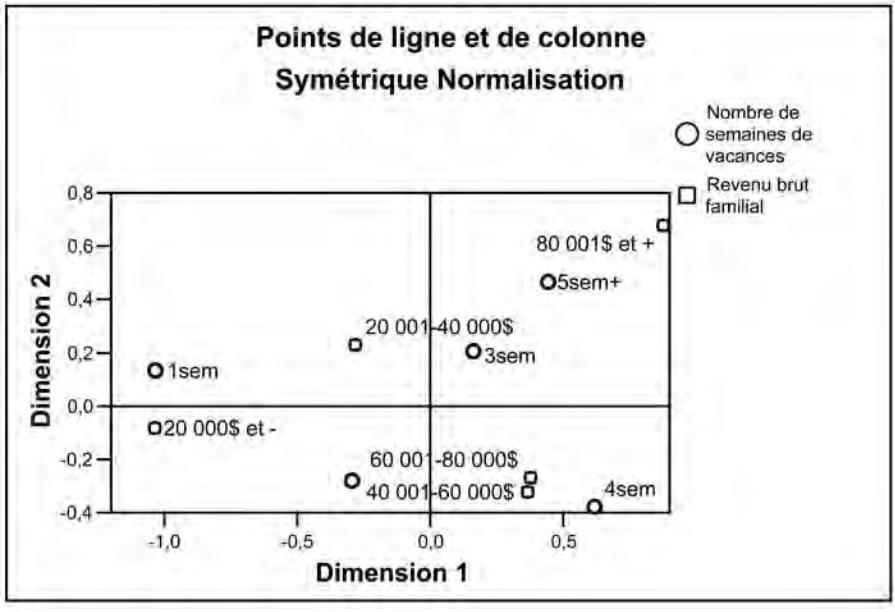
Revenu brut familial	Masse	Score dans la dimension		Inertie	Contribution				
		1	2		De point à inertie de dimension		De dimension à inertie de point		
					1	2	1	2	Total
20 000$ et -	,144	-1,035	-,082	,049	,492	,009	,989	,002	,992
20 001-40 000$	,333	-,282	,231	,011	,084	,171	,780	,173	,952
40 001-60 000$	,303	,367	-,323	,017	,130	,304	,751	,192	,944
60 001-80 000$	,123	,377	-,269	,009	,056	,086	,609	,102	,711
80 001$ et +	,097	,876	,678	,028	,238	,430	,829	,164	,993
Total actif	1,000			,114	1,000	1,000			

a. Normalisation principale symétrique

Caractéristiques des points colonnes[a]

Nombre de semaines de vacances	Masse	Score dans la dimension		Inertie	Contribution				
		1	2		De point à inertie de dimension		De dimension à inertie de point		
					1	2	1	2	Total
1sem	,159	-1,033	,134	,055	,540	,028	,972	,005	,977
2sem	,265	-,294	-,280	,011	,073	,201	,660	,199	,859
3sem	,172	,162	,207	,003	,014	,071	,493	,266	,759
4sem	,202	,618	-,379	,028	,246	,279	,851	,106	,957
5sem+	,202	,443	,465	,017	,126	,421	,732	,267	,999
Total actif	1,000			,114	1,000	1,000			

a. Normalisation principale symétrique



Dans le premier tableau (« Tableau des correspondances ») nous avons les données observées ; dans le deuxième (« Résumé ») sont présentés :

- dans la colonne 1, les principaux facteurs du modèle ;
- dans la colonne 2, les valeurs singulières qui sont les racines des valeurs propres ;
- dans la colonne 3, les valeurs propres (inertie) ;
- dans la colonne 4, le résultat du khi-carré (60,867) ;
- dans la colonne 5, le degré de signification du test ou le degré d'erreur (ici il est de 0,000).

  On peut conclure ici que l'hypothèse H_0 est rejetée et qu'il y a une relation certaine entre les deux variables ;
- dans la colonne 6, nous avons la proportion d'inertie attribuable à chacun des facteurs (la variance expliquée par chacun d'eux) ; le premier facteur retient 86,4 % de l'inertie totale et le deuxième facteur, 9,5 % ; la somme cumulée donne 95,9 % (voir la colonne 7).

Les deux tableaux suivants, « Caractéristiques des points lignes » et « Caractéristiques des points colonnes », présentent une analyse en fonction des lignes et des colonnes du tableau croisé. Ils permettent de se faire une idée de l'importance de chacune des catégories des deux variables étudiées dans le modèle d'analyse factorielle des correspondances.

Enfin, le graphique réunissant les deux variables se nomme « Points de ligne et de colonne. Normalisation symétrique ». Nous pouvons y lire :

- que la catégorie 20 000 $ et moins se rapproche de une semaine de vacances ; à l'autre extrémité, la catégorie 80 001 $ et plus est très près de la catégorie de cinq semaines de vacances et plus ;
- les valeurs moyennes se retrouvent au centre de la carte ;
- dans l'ensemble, la carte ne fait qu'illustrer les éléments les plus visibles du tableau croisé.

CHAPITRE

5

Analyse bivariée

Tableau de contingence et khi-carré

Dans la plupart des études empiriques, nous cherchons à montrer les principales relations entre les variables utilisées ; selon les hypothèses formulées, nous tentons de démontrer l'influence de certaines variables sur les autres.

1. OBJECTIFS ET ASPECTS THÉORIQUES

L'objectif de l'analyse bidimensionnelle est d'étudier les liens entre deux variables d'une enquête. Dans un premier temps, nous nous attendons à ce que les variables de segmentation (les variables factuelles ou socio-économiques) aient une influence sur les composants, les choix, les goûts des consommateurs. Par exemple, nous pouvons faire l'hypothèse que les revenus des répondants vont avoir un effet sur le nombre de nuitées à l'hôtel ou le nombre de repas consommés au restaurant dans les trois derniers mois.

Habituellement, trois types de liaison entre variables peuvent être envisagés[1] :

- **La liaison nulle** : il n'y a aucune relation entre les deux variables ; par exemple, entre la taille des répondants et leur niveau de scolarité.
- **La relation quasi totale entre les variables** ; par exemple, le chauffage d'une pièce métallique et son degré de dilatation ; dans ce cas, chacune des transformations d'un caractère a un effet direct et proportionnel sur l'autre caractère.
- **La liaison relative** : une variable en influence une autre, mais dans certaines limites ; par exemple, la scolarité influence les départs en vacances, mais cette relation n'est pas proportionnelle : la différence entre les niveaux de scolarité primaire et secondaire est de 30 % dans les départs en vacances ; elle est de 19 % entre le secondaire et le collégial, de 10 % entre le collégial et le niveau universitaire et, enfin, de seulement 3 % entre le baccalauréat et la maîtrise.

La liaison relative correspond à ce que Raymond Boudon appelle « une théorie de l'implication faible[2] ». Selon lui : « Les relations d'implication réciproque qu'on peut établir entre les différents éléments des systèmes institutionnels sont donc assimilables, non à des implications strictes de type logique (si A, alors B), mais à des implications faibles de type stochastique (si A, alors plus souvent B)[3]. » La réalité sociale permet rarement d'observer des implications strictes.

Dans la plupart des enquêtes, nous avons affaire à des variables qualitatives (nominales ou ordinales) ou à un mixage de variables qualitatives et quantitatives (par intervalles ou de rapport). Il faut donc choisir des tests qui conviennent à ce genre de variables. Pour mesurer l'indépendance des variables, il faut utiliser des tests d'hypothèses. Les tests d'hypothèses nous aident à interpréter les données et à prendre des décisions. Ces tests nous permettent de déterminer si les relations entre deux variables données sont dues au hasard ou sont réellement significatives. Ces tests statistiques visent à vérifier des hypothèses. Une hypothèse se

1. Voir à ce sujet : B. Py (1987), *Statistique descriptive*, Paris, Economica, p. 186.
2. R. Boudon (1971), *Les mathématiques en sociologie*, Paris, Presses universitaires de France, p. 19.
3. R. Boudon et F. Bourricaud (1982), *Dictionnaire critique de la sociologie*, Paris, Presses universitaires de France, p. 578.

formule en supposant des relations ou l'absence de relation entre les deux variables choisies. On a donc deux hypothèses. L'hypothèse H_0 présume qu'il n'y a pas de relation entre les variables. L'hypothèse H_1, au contraire, affirme qu'il y a une relation entre les deux variables. Les résultats des tests permettront de trancher entre ces deux hypothèses et d'en tirer les conséquences pour l'interprétation des données.

Pour l'étude des relations entre les variables, le khi-carré (ou khi deux) est le test le plus utilisé. Le khi-carré est d'abord et avant tout destiné à l'examen de la relation entre deux variables qualitatives, nominales ou ordinales. C'est en même temps un test à large spectre qui s'adresse à plusieurs types de variables, étant donné qu'il est toujours possible de regrouper des données quantitatives selon des classes (ce qui implique évidemment une certaine perte d'information). « Contrairement à la corrélation qui exige des variables quantitatives, cette méthode s'applique à toutes les variables, quelle que soit leur nature[4]. » Le test du khi-carré va donc mesurer la liaison statistique entre deux variables.

Le test du khi-carré est un test d'hypothèse ; il fonctionne essentiellement dans la comparaison entre une fréquence observée et une fréquence théorique. La formule générale est la suivante :

Le khi-carré comme test d'indépendance :

$$\chi^2 = \sum\sum \frac{\left(F_0 - F_{th}\right)^2}{F_{th}}$$

Où :

F_0 : la fréquence observée ; ce sont les résultats obtenus sur le terrain.

F_{th} : la fréquence théorique ; elle est obtenue par le calcul : (Total de la colonne $\times$ Total de la rangée) / Grand total.

Le test du khi-carré se doit d'obéir à certaines contraintes :

1. Le calcul du test doit toujours se faire à partir des données brutes.
2. Il faut que les catégories de chacune des variables soient exhaustives et mutuellement exclusives.
3. Les fréquences théoriques doivent être égales ou supérieures à cinq.

4. J. Rose (1993), *Le hasard au quotidien*, Paris, Seuil, p. 178.

Le khi-carré doit aussi tenir compte de la largeur du tableau de contingence ; nous allons donc calculer le « degré de liberté ». Celui-ci correspond à : dl = (Nombre de lignes – 1) × (Nombre de colonnes – 1). Nous devons ajouter un autre élément pour la compréhension du test : il s'agit du risque d'erreur alpha (α) que nous sommes prêts à accepter : « Il s'agit de définir quel risque nous prenons quand nous affirmons que les variables sont liées dès que le χ^2 calculé est supérieur à χ^2 théorique[5]. » Le khi-carré théorique apparaît dans la table de distribution du khi-carré de Pearson (il faut noter que cette table est intégrée au logiciel statistique SPSS).

La valeur du khi-carré calculée doit être supérieure à celle du khi-carré théorique. Quand le khi-carré calculé est supérieur à la valeur critique (selon le degré de liberté), nous devons retenir l'hypothèse d'une dépendance (d'un lien) entre les deux variables étudiées. Par convention le seuil est de 0,10, 0,05 ou 0,01 (ou moins). Normalement, dans les tableaux, pour faciliter la lecture des résultats, nous allons retrouver seulement les principaux éléments du test du khi-carré :

1. le khi-carré calculé ;
2. le degré de liberté (dl) ;
3. enfin, le risque d'erreur ; celui-ci pourrait varier entre 0,10 et 0,000.

Les principales étapes dans le déroulement de la méthode sont :

1. la formulation des hypothèses H_0 et H_1 ;
2. le calcul du khi-carré ;
3. la détermination du seuil de décision acceptable ;
4. la conclusion sur l'hypothèse définie à l'étape 1 ;
5. l'interprétation « littéraire » des données en tenant compte des résultats du test.

Nous verrons ces étapes à la section 3.

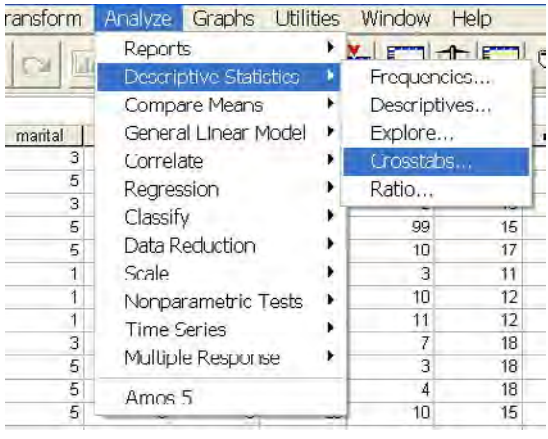
5. J.-P. Crauser, Y. Harvatopoulos et P. Sarnin (1989), *Guide pratique d'analyse des données*, Paris, Les Éditions d'Organisation, p. 93.

2. LES COMMANDES AVEC SPSS

Le cheminement pour parvenir à la fenêtre principale de dialogue pour la commande d'un tableau croisé est :

Figure 5.1

LE CHEMINEMENT POUR OBTENIR UN TABLEAU CROISÉ

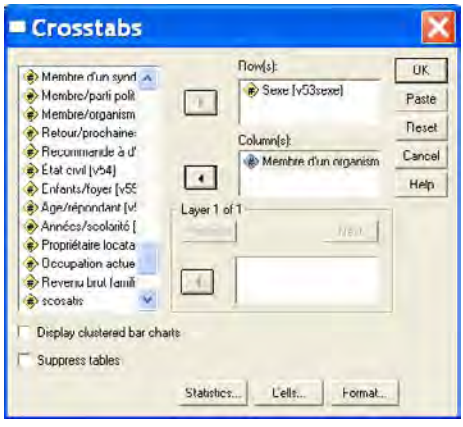


Dans le rectangle de gauche de la figure 5.2, nous retrouvons les variables de l'enquête. Dans le premier rectangle de droite (**ROW**), nous devons faire glisser la variable indépendante ; ici, c'est la variable « Sexe ». Dans le deuxième rectangle de droite (**COLUMN**), il faut introduire la variable dépendante : « Membre d'un organisme communautaire ». Le troisième rectangle servira, dans les tableaux à triples entrées, à insérer une variable de contrôle. Dans ces trois rectangles, il est possible de faire appel à plusieurs variables à la fois.

Dans la fenêtre principale, sous le grand rectangle des variables, se trouvent deux petits carrés. Le premier (**DISPLAY CLUSTERED BAR CHARTS**) permet la création de diagrammes en bâtons juxtaposés ; le diagramme ainsi créé contiendra la variable indépendante et la variable dépendante. Le deuxième carré (**SUPPRESS TABLES**) va supprimer les tableaux statistiques pour ne conserver que les résultats des tests statistiques à la condition que ceux-ci aient été demandés au préalable. Nous pouvons activer ces commandes en cliquant à l'intérieur des carrés.

Figure 5.2

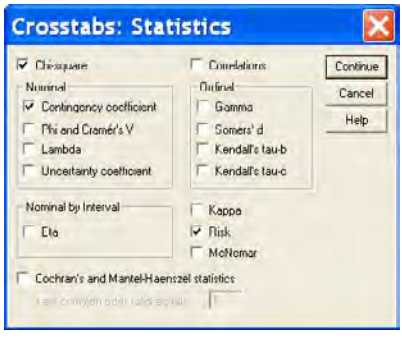
LA BOÎTE PRINCIPALE POUR LES TABLEAUX CROISÉS



Tout en bas de la fenêtre principale, nous avons accès à trois types de commandes. Ces commandes amènent l'ouverture de trois fenêtres secondaires. La première (**STATISTICS**) est reproduite à la figure 5.3.

Figure 5.3

LES STATISTIQUES

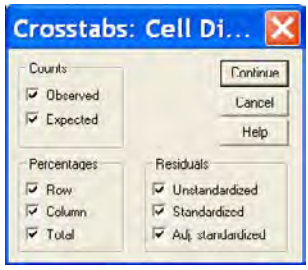


Nous voyons à la figure 5.3 que quinze tests sont disponibles. L'utilisation de l'un ou de l'autre de ces tests dépendra de la structure des données (variables, nominales, ordinales ou par intervalles). Habituellement, le test le plus utilisé est le khi-carré (**CHI-SQUARE**). Nous verrons, à la section 3, dans quelles conditions il faut utiliser ces tests.

La deuxième commande en bas de la fenêtre principale s'appelle **CELLS** ; elle permet l'ouverture de la deuxième fenêtre secondaire, représentée à la figure 5.4.

Figure 5.4

LES CELLULES DES TABLEAUX



En activant chacun des carrés de la figure 5.4, nous allons créer huit informations par cellule du tableau croisé. Ces informations se divisent en trois groupes :

1. Les effectifs (**COUNTS**) :
   – observés (**OBSERVED**),
   – théoriques (**EXPECTED**).
2. Les pourcentages (**PERCENTAGES**) :
   – en ligne (**ROW**),
   – en colonne (**COLUMN**),
   – total ligne et total colonne (**TOTAL**).
3. Les résidus (**RESIDUALS**) :
   – non standardisés (**UNSTANDARDIZED**),
   – standardisés (**STANDARDIZED**),
   – standardisés ajustés (**ADJ. STANDARDIZED**).

Enfin, la dernière commande, **FORMAT**, permet l'ouverture d'une troisième fenêtre secondaire, présentée à la figure 5.5.

Figure 5.5

L'ORDRE DES CARACTÈRES DES VARIABLES

Dans la fenêtre de la figure 5.5, seulement deux options sont possibles :

- Dans la première option, **ASCENDING** (favorisée par convention), l'ordre des catégories de la variable en ligne (habituellement la variable indépendante) est respecté (un ordre croissant) ;

- dans la deuxième option, **DESCENDING**, l'ordre des catégories de la variable en ligne est décroissant, c'est-à-dire qu'il débute par la valeur la plus forte jusqu'à la plus faible.

3. UN EXEMPLE D'ANALYSE

Dans l'encadré 5.1, nous présentons les résultats obtenus avec le logiciel SPSS à partir de la commande **CROSSTABS** (voir la figure 5.2). Tel que demandé (voir la figure 5.4), nous avons huit informations dans la première cellule du tableau. Voyons ces informations :

Encadré 5.1

UN TABLEAU CROISÉ AVEC TOUTES LES POSSIBILITÉS DES CELLULES

Tableau croisé

Tableau croisé Sexe * Membre d'un organisme communautaire

			Membre d'un organisme communautaire		
			Oui	Non	Total
Sexe	Féminin	Effectif	299	691	990
		Effectif théorique	253,6	736,4	990,0
		% dans Sexe	30,2%	69,8%	100,0%
		% dans Membre d'un organisme communautaire	90,6%	72,1%	76,9%
		% du total	23,2%	53,6%	76,9%
		Résidu	45,4	-45,4	
		Résidu standardisé	2,8	-1,7	
		Résidu ajusté	6,9	-6,9	
	Masculin	Effectif	31	267	298
		Effectif théorique	76,4	221,6	298,0
		% dans Sexe	10,4%	89,6%	100,0%
		% dans Membre d'un organisme communautaire	9,4%	27,9%	23,1%
		% du total	2,4%	20,7%	23,1%
		Résidu	-45,4	45,4	
		Résidu standardisé	-5,2	3,0	
		Résidu ajusté	-6,9	6,9	
Total		Effectif	330	958	1288
		Effectif théorique	330,0	958,0	1288,0
		% dans Sexe	25,6%	74,4%	100,0%
		% dans Membre d'un organisme communautaire	100,0%	100,0%	100,0%
		% du total	25,6%	74,4%	100,0%

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	47,118[b]	1	,000		
Correction[a] pour la continuité	46,085	1	,000		
Rapport de vraisemblance	54,039	1	,000		
Test exact de Fisher				,000	,000
Association linéaire par linéaire	47,082	1	,000		
Nombre d'observations valides	1288				

a. Calculé uniquement pour un tableau 2x2

b. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 76,35.

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Coefficient de contingence	,188	,000
Nombre d'observations valides		1288	

a. L'hypothèse nulle n'est pas considérée.

b. Utilisation de l'erreur standard asymptotique dans l'hypothèse nulle.

Estimation du risque

	Valeur	Intervalle de confiance de 95%	
		Inférieur	Supérieur
Odds Ratio pour Sexe (Féminin / Masculin)	3,727	2,509	5,537
Pour cohorte Membre d'un organisme communautaire = Oui	2,903	2,053	4,105
Pour cohorte Membre d'un organisme communautaire = Non	,779	,736	,824
Nombre d'observations valides	1288		

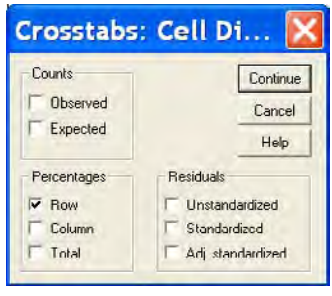
1. Le premier nombre comprend l'effectif réel de la population, soit 299 personnes de sexe féminin qui ont répondu « oui » à la question, « Êtes-vous membre d'un organisme communautaire ? »
2. Le deuxième nombre correspond à l'effectif théorique calculé pour cette cellule, soit : 330 × 990 / 1288 = 253,6.
3. Le premier pourcentage, 30,2 %, est le résultat de la lecture en ligne : 299/990 = 0,302 ou 30,2 % ; ici 30,2 % des femmes font partie d'un organisme communautaire.

4. Le deuxième pourcentage est le résultat de la lecture en colonne : 299/330 = 0,906 ou 90,6 % ; dans ce cas, 90,6 % de ceux qui ont répondu « oui » sont de sexe féminin.
5. Le troisième pourcentage renvoie au total de la population ; ainsi, 299/1288 = 0,232 ou 23,2 % ; 23,2 % des répondantes sont membres d'un organisme communautaire.
6. La sixième information, le troisième nombre, est le résidu, c'est-à-dire l'écart entre l'effectif réel (observé) et l'effectif théorique : 299 – 253,6 = 45,4.
7. La septième information contient le résidu standardisé[6].
8. La huitième information désigne le résidu standardisé et ajusté[7].

La présentation des données en ligne, dans la construction des tableaux croisés, facilite la lecture des résultats. Dans la figure 5.6, nous avons adopté cette manière de faire : il s'agit tout simplement de désactiver toutes les options à l'exception de **ROW**. Le nouveau tableau croisé ainsi créé est présenté à l'encadré 5.2.

Figure 5.6

LA LECTURE HORIZONTALE D'UN TABLEAU



Encadré 5.2

UN TABLEAU CROISÉ SIMPLIFIÉ ; UNE LECTURE HORIZONTALE

Tableau croisé Sexe * Membre d'un organisme communautaire

% dans Sexe

		Membre d'un organisme communautaire		
		Oui	Non	Total
Sexe	Féminin	30,2%	69,8%	100,0%
	Masculin	10,4%	89,6%	100,0%
Total		25,6%	74,4%	100,0%

6. Voir à ce sujet : S.J. Haberman (1978), *Analysis of Qualitative Data*, New York, Academic Press.
7. *Idem*.

Dans l'encadré 5.1, nous avons les résultats du test du khi-carré ; ces résultats apparaissent sous le tableau principal sous vocable « Khi-carré de Pearson ». La valeur du khi-carré est de 47,118, le nombre de degrés de liberté est de 1 et la « signification asymptotique bilatérale » est de 0,000. Ces résultats montrent (selon la table du khi-carré) que l'hypothèse H_0 doit être rejetée ; cela indique que la différence observée entre les femmes et les hommes (voir encadré 5.2) est significative au plan statistique. En d'autres mots, les femmes sont plus susceptibles que les hommes d'être membres d'un organisme communautaire.

Afin de mieux comprendre ce test du khi-carré, nous allons examiner un autre exemple. Dans une recherche, nous avons les données suivantes :

Tableau 5.1

LA PUBLICITÉ ET L'ACHAT D'UN FORFAIT-VOYAGE

Achat du forfait-voyage / *A vu la publicité dans le dernier mois*	***Oui***	***Non***	**Total**
Oui	80	120	200
Non	15	85	100
Total	95	205	**300**

Nous avons au tableau 5.1 une variable indépendante (les répondants ont vu la publicité du forfait-voyage dans le dernier mois : oui/non) et une variable dépendante (les répondants ont acheté le forfait-voyage : oui/non). Nous pouvons dès lors formuler deux hypothèses (test d'hypothèses) :

1. l'hypothèse H_0 : la publicité n'a aucune influence sur l'achat du forfait-voyage ;
2. l'hypothèse H_1 : la publicité a une influence sur l'achat du forfait voyage.

Nous allons maintenant appliquer la formule :

$$\chi^2 = \sum\sum \frac{(F_o - F_{th})^2}{F_{th}}$$

Nous nous rappelons que la fréquence théorique correspond à : (Total de la colonne × Total de la ligne) / Grand total ; le degré de liberté se calcule par : (Nombre de lignes – 1) × (Nombre de colonnes – 1) ; ici, le total ligne et le total colonne ne sont pas pris en compte.

Au tableau 5.2, nous allons déconstruire cette formule afin de faciliter la compréhension.

Tableau 5.2

LE CALCUL DU KHI-CARRÉ À PARTIR DES DONNÉES DU TABLEAU 5.1

Fréquence observée ***(F_o)***	**Fréquence théorique** ***(F_{th})***	$F_o - F_{th}$	$(F_o - F_{th})^2$	$\frac{(F_0 - F_{th})^2}{F_{th}}$
80	63,3	16,7	278,9	$\frac{278,9}{63,3} = 4,4$
120	136,6	–16,6	275,6	$\frac{275,6}{136,6} = 2,0$
15	31,6	-16,6	275,6	$\frac{275,6}{31,6} = 8,7$
85	68,3	16,7	278,9	$\frac{278,9}{68,3} = 4,1$
	KHI-CARRÉ			**19,2**

Le degré de liberté sera égal à (2 – 1) (2 – 1) = 1.

L'étape suivante consiste à consulter la table du khi-carré ; celle-ci nous indique les probabilités de nous tromper si H_1 est vraie.

Selon cette table[8], avec un degré de liberté, et selon la marge d'erreur, les seuils limites sont les suivants :

8. Sur le débat sémantique entre le khi-carré comme test unilatéral ou bilatéral, voir : D.C. Howell (1998), *Méthodes statistiques en sciences humaines*, Bruxelles, De Boeck-Université, p. 177 et suivantes.

Tableau 5.3

LES SEUILS LIMITES DU KHI-CARRÉ POUR UNE PROBABILITÉ D'ERREUR DONNÉE*

Probabilité d'erreur	*Seuils limites*
0,01	6,635
0,02	5,412
0,05	3,841
0,10	2,706

* Voir la table A de lecture du khi-carré en annexe, page 242.

Au tableau 5.2, le khi-carré est égal à 19,2 ; il est donc largement supérieur à la limite de 6,635 avec une probabilité d'erreur de 0,01 (ou 1 %). Nous devons donc rejeter l'hypothèse nulle H_0 et admettre que la publicité a, dans ce cas, une influence sur l'achat d'un forfait-voyage.

Le tableau à analyser (à publier dans le rapport final) sera le suivant.

Tableau 5.4

LA PUBLICITÉ ET L'ACHAT D'UN FORFAIT-VOYAGE EN POURCENTAGE

Achat du forfait-voyage / *A vu la publicité dans le dernier mois*	*Oui*	*Non*	**Total**
Oui	40	60	100
Non	15	85	100

Khi-carré : 19,2 ; dl : 1 ; signification : 0,000.

Nous voyons dans ce tableau que 40 % de ceux qui ont vu la publicité dans le dernier mois ont acheté le forfait-voyage. Le test du khi-carré étant significatif, nous pouvons croire que, dans ce cas, la publicité influence l'achat d'un forfait-voyage.

3.1. LES AUTRES TESTS STATISTIQUES POUR LES TABLEAUX CROISÉS

Comme nous l'avons déjà vu à la figure 5.3, d'autres tests statistiques sont disponibles selon la structure des données du tableau croisé[9]. Soulignons ces quelques tests :

- Le khi-carré : quand le tableau croisé est formé de plus de deux lignes et de deux colonnes, seuls le khi-carré de Pearson et le khi-carré du rapport de vraisemblance peuvent s'appliquer. Pour un tableau avec seulement deux lignes et deux colonnes, le test exact de Fisher et le test khi-carré de Yates corrigé peuvent être utilisés. Il faut noter que lorsque les deux variables du tableau croisé sont des variables quantitatives, le test de variables Mantel-Haenszel (linéaire par linéaire) sera considéré.
- Le test des corrélations de Pearson s'adresse seulement à des variables quantitatives.
- Pour les variables nominales, en plus du khi-carré, s'appliquent les tests suivants : le coefficient de contingence, les coefficients phi et V de Cramer.
- Le test du risque sera utilisé seulement dans un tableau croisé avec deux lignes et deux colonnes ; de plus, ce test doit évaluer le risque (ou la probabilité) d'appartenir ou non à une catégorie de l'une ou de l'autre des variables.
- Tous les autres tests dépendent de conditions très explicites liées à la structure des variables impliquées dans le tableau croisé.

Nous allons maintenant étudier deux tests (les plus utilisés) : le coefficient de contingence et le test du risque.

3.2. LE COEFFICIENT DE CONTINGENCE

Le test de contingence peut être considéré comme une approximation du *r* carré de Pearson. Pour être pris en considération, il doit pouvoir répondre à certaines conditions :

9. SPSS est un programme *self-service* où chaque chercheur sélectionne la partie de traitement des données qui lui convient dans le cadre de sa propre recherche.

1. les deux variables se distribuent normalement dans la population ;
2. les deux variables ont chacune plusieurs catégories (trois ou plus) ;
3. la taille de l'échantillon est relativement grande ;
4. le khi-carré est significatif.

Le khi-carré est un test d'hypothèse qui nous indique s'il y a une relation (ou non) entre les deux variables. Le test de contingence nous donne une mesure de l'intensité de cette relation ; cette mesure se situe entre zéro et un. Le zéro nous montre une relation nulle ou très faible ; à l'autre extrême, un nous indique une relation totale entre les deux variables.

La formule de calcul du coefficient de contingence est :

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Ici, *n* correspond à la taille de l'échantillon ou de la population. En prenant l'exemple de l'encadré 5.1, nous voyons que le khi-carré est égal à 47,118 ; nous aurons donc :

$$C = \sqrt{\frac{47,118}{1288 + 47,118}} = 0,1878$$

Le calcul du coefficient de contingence doit tenir compte d'un facteur de correction qui dépend de la taille du tableau en termes de lignes et de colonnes (voir la table B du facteur de correction du test de contingence en annexe, page 243). Le nouveau coefficient de contingence obtenu correspond à :

$$\frac{\text{Coefficient de contingence}}{\text{Facteur de correction}} = \text{Nouveau coefficient de contingence}$$

Ainsi, nous aurons :

$$\frac{0,1878}{0,707} = 0,2656$$

Si nous considérons la table de lecture du tableau 5.5 ci-dessous, nous avons donc une association « moyenne » entre la variable indépendante et la variable dépendante.

Tableau 5.5

LE TABLEAU DE LECTURE DU TEST DE CONTINGENCE

La forme d'association	*Les valeurs du test CC*
Une association nulle ou très faible	0,0 et 0,10
Une association faible	0,11 et 0,20
Une association moyenne	0,21 et 0,30
Une association forte	0,31 et 0,40
Une association très forte	0,41 et plus

3.3. LA NOTION DE RISQUE OU DE CHANCE

À la figure 5.3, nous voyons en bas à droite la commande **RISK**. Cette commande s'applique uniquement au tableau croisé à deux variables ayant chacune seulement deux catégories ; il faut aussi, bien sûr, que la notion de risque ou de chance puisse s'appliquer.

Dans l'encadré 5.1, nous avons, au dernier tableau, les résultats de la commande **RISK**. À partir du tableau croisé de l'encadré 5.2, le calcul du test se fait comme suit :

1. 30,2 / 69,8 = 0,432664756
2. 10,4 / 89,6 = 0,116071428
3. 0,432664756 / 0,116071428 = 3,727

Ce dernier résultat veut dire que les femmes sont 3,7 fois plus susceptibles que les hommes d'être membres d'un organisme communautaire. Dans le tableau d'estimation du risque, nous avons aussi les intervalles de confiance à 95 %. Les résultats de calcul du risque pour chacune des cohortes sont aussi présentés ; par exemple, pour la cohorte « Membre d'un organisme communautaire = Oui » la valeur est 2,903 ; le calcul se fait comme suit :

- Total des « Non » 25,6 / Total des « Oui » 74,4 = 2,9

Pour la cohorte « Membre d'un organisme communautaire = Non » nous aurons :

- « Non » de sexe féminin 69,8 / « Non » de sexe masculin 89,6 = 0,779

3.4. LES TABLEAUX CROISÉS À TROIS VARIABLES

Les tableaux croisés à trois variables nous permettent d'approfondir l'analyse des données. À la figure 5.7, nous avons ajouté, dans le dernier rectangle de droite (**LAYER 1 OF 1**), la nouvelle variable indépendante de contrôle « Membre d'un syndicat **».**

Figure 5.7

UN TABLEAU CROISÉ À TROIS VARIABLES

Les résultats de cette commande sont présentés à l'encadré 5.3.

Encadré 5.3

UN TABLEAU CROISÉ À TROIS VARIABLES

Tableau croisé Sexe * Membre/organisme communautaire * Membre d'un syndicat

% dans Sexe

Membre d'un syndicat			Membre/organisme communautaire		Total
			Oui	Non	
Oui	Sexe	Féminin	21,6%	78,4%	100,0%
		Masculin	16,7%	83,3%	100,0%
	Total		19,4%	80,6%	100,0%
Non	Sexe	Féminin	30,7%	69,3%	100,0%
		Masculin	9,4%	90,6%	100,0%
	Total		26,1%	73,9%	100,0%

Tests du Khi-deux

Membre d'un syndicat		Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Oui	Khi-deux de Pearson	,355[b]	1	,552		
	Correction pour la continuité[a]	,110	1	,740		
	Rapport de vraisemblance	,358	1	,550		
	Test exact de Fisher				,607	,372
	Association linéaire par linéaire	,351	1	,554		
	Nombre d'observation valides	93				
Non	Khi-deux de Pearson	47,288[c]	1	,000		
	Correction pour la continuité[a]	46,191	1	,000		
	Rapport de vraisemblance	55,341	1	,000		
	Test exact de Fisher				,000	,000
	Association linéaire par linéaire	47,248	1	,000		
	Nombre d'observation valides	1195				

a. Calculé uniquement pour un tableau 2x2

b. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 8,13.

c. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 66,84.

Mesures symétriques

Membre d'un syndicat			Valeur	Signification approximée
Oui	Nominal par Nominal	Coefficient de contingence	,062	,552
	Nombre d'observations valides		93	
Non	Nominal par Nominal	Coefficient de contingence	,195	,000
	Nombre d'observations valides		1195	

a. L'hypothèse nulle n'est pas considérée.

b. Utilisation de l'erreur standard asymptotique dans l'hypothèse nulle.

Nous pouvons constater au deuxième tableau de l'encadré 5.3, « Tests du khi-carré », que dans la catégorie « Oui » de la variable « Membre d'un syndicat », les différences observées entre les hommes et les femmes ne sont pas significatives (la « Signification asymptotique (bilatérale) » est de 0,552) ; nous devons donc, dans ce cas, rejeter l'hypothèse d'une influence du sexe sur le fait d'être membre ou non d'un organisme communautaire.

Nous voyons que dans la catégorie « Non » de la variable « Membre d'un syndicat » les différences observées entre les hommes et les femmes sont significatives (la « Signification asymptotique (bilatérale) » est de 0,000). Pour les personnes membres d'un syndicat, les hommes ou les femmes, il n'y a pas de différences significatives en ce qui concerne la participation à un organisme communautaire.

À l'inverse, pour les hommes et les femmes qui ne sont pas membres d'un syndicat, il y a une différence significative relativement à la participation à un organisme communautaire : 30,7 % des femmes pour 9,4 % des hommes. Nous voyons que l'utilisation d'une variable de contrôle amène un raffinement de l'analyse des données ; à ce moment, la principale difficulté est de bien choisir cette variable de contrôle.

Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

CHAPITRE

6

Analyse bivariée

L'analyse de variance

Lorsqu'on confronte une variable quantitative à une variable qualitative (nominale ou ordinale), on recourt très généralement à la comparaison de moyennes ou à l'analyse de variance (ANOVA). L'analyse de régression est aussi adaptée. On y suppose tout comme dans ANOVA une relation linéaire entre la variable quantitative et une ou plusieurs variables qualitatives de type binaire (1,0). L'analyse par régression simple aboutit en pratique à une analyse de variance (ANOVA) avec une variable quantitative et une variable muette de type binaire (1,0). Si la variable qualitative comporte plus de deux catégories, l'analyse de variance à partir de la régression requiert une régression multiple.

L'analyse de variance sera présentée à partir du fichier Salempl.sav. Ce fichier comporte une variable « salaire », le logarithme naturel de salaire : « Lsal », une variable « ndiplom » identifiant 4 niveaux d'instruction (1, 2, 3, 4), 4 variables muettes (« niv1 », « niv2 » , « niv3 », « niv4 ») représentant chacune un niveau d'instruction par le chiffre un et posant égal à zéro chacun des autres cas.

1. REPÈRES THÉORIQUES

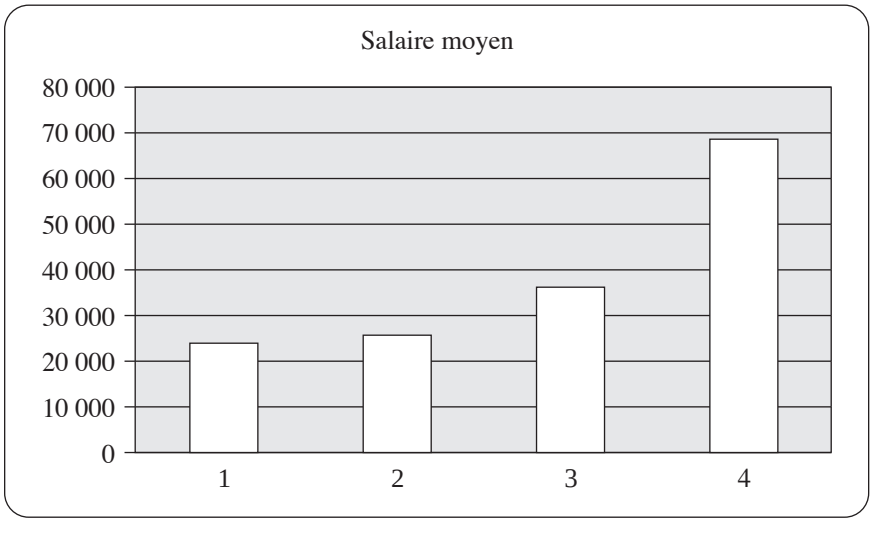
Les logiciels d'analyse statistique offrent diverses alternatives pour étudier la relation entre une variable quantitative et une variable qualitative. Ces méthodes ont toutes en commun le fait qu'elles font une comparaison des moyennes résultant de la répartition des données de la variable quantitative selon les catégories que comporte la variable qualitative. L'analyse dite de variance (conjointement à la régression) est la version la plus générale ; elle dispense en pratique de recourir aux autres méthodes.

L'analyse de variance permet de confronter les données d'une variable quantitative aux données d'une variable qualitative comportant deux catégories ou plus. On se demande par exemple dans quelle mesure le revenu observé (variable quantitative) est associé aux différents niveaux (observés) d'instruction (variable qualitative ordinale). Dans certaines conditions, on peut généraliser la conclusion relative à une relation entre le revenu et le niveau d'instruction.

Le graphique ci-dessous présente, sur la base des données du fichier Salempl.sav, le salaire moyen selon le niveau d'instruction.

Figure 6.1

LES SALAIRE SELON LE NIVEAU D'INSTRUCTION (FICHIER SALEMPL.SAV)



Les différences de moyennes que permet de visualiser le graphique sont-elles dues aux particularités aléatoires de l'échantillon ou reflètent-elles aussi des différences réelles dans les quatre populations correspondant aux quatre niveaux d'instruction ?

Le principe du raisonnement est le suivant :

> Plus les différences entre les moyennes dans l'échantillon sont importantes, plus il est difficile d'admettre que ces différences résultent simplement du hasard et plus on est porté à admettre qu'il existe des différences entre les moyennes de populations (correspondant aux différents niveaux d'instruction).

Par ailleurs, on sera plus confiant sur ce type de conclusion si la variation autour des moyennes observées est petite.

Sur la base de ce raisonnement, on calcule la somme des carrés **entre** les groupes (SCG). SCG est censé mesurer l'importance des différences entre les moyennes. On calcule par ailleurs la somme des carrés **dans** les groupes (SCE). SCE mesure la variation dans les groupes.

$$\text{Variation entre les groupes} = \text{SCG} = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{\bar{X}}\right)^2 \qquad (1)$$

$$\text{Variation dans les groupes} = \text{SCE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2 \qquad (2)$$

Dans ces expressions, $\bar{\bar{X}}$ représente la moyenne des X sur l'ensemble des cas constituant l'échantillon. $\bar{X}_i$, pour sa part, représente la moyenne calculée sur les observations correspondant à la catégorie i.

L'ordre de grandeur de SCG et de SCE est affecté par le nombre de groupes (k) et la taille de l'échantillon. C'est pourquoi il vaudrait mieux calculer des quantités uniformisées, qu'on appelle **variances.**

$$\text{Variance entre les groupes} = \text{CMG} = \text{SCG} / (k - 1) \qquad (3)$$

$$\text{Variance dans les groupes} = \text{CME} = \text{SCE} / (N - k) \qquad (4)$$

où

N = nombre total d'observations

k = nombre de catégories

La figure suivante montre le détail du calcul de CME et CMG à partir de la variable « Lsal ».

Tableau 6.1

LE CARRÉ MOYEN ENTRE LES GROUPES (CMG) – DÉTAIL DU CALCUL

Niveau de diplomation	***Nombre de cas***	***Salaire moyen***	***ln (salaire moyen)***			
	n_i		$\bar{X}_i$	$(\bar{X}_i - \bar{X})^2$	$n_i(\bar{X}_i - \bar{X})^2$	
1	53	23 840	10,08	0,08	4,09	
2	196	25 516	10,15	0,04	8,62	
3	186	36 135	10,50	0,02	3,55	
4	39	68 115	11,13	0,60	23,25	
					SCG	***CMG***
Toutes catégories (1, 2, 3, 4)	N		$\bar{X}$		$\sum_i^k n_i\left(\bar{X}_i - \bar{X}\right)^2$	$\dfrac{\sum_i^k n_i\left(\bar{X}_i - \bar{X}\right)^2}{(k-1)}$
	474,00	31 470	10,36		39,52	13,17

En divisant CMG par CME, on obtient le *F* calculé. Le tableau 6.2 présente les calculs principaux qui conduisent au *F* calculé.

Tableau 6.2

LE TABLEAU D'ANALYSE DE VARIANCE

	Somme des carrés	***Degrés de liberté***	***Carré moyen***	***F***	***Sig***
	SCG	***DF***	***CMG***		
	$\sum_i^k n_i\left(\bar{X}_i - \bar{X}\right)^2$	$k-1$	$\dfrac{\sum_i^k n_i\left(\bar{X}_i - \bar{X}\right)^2}{(k-1)}$ F = CMG/CME		
Entre les groupes	39,52	3	13,17	176,09	0,00
	SCE		CME		
	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$	$N-k$	$\text{CME} = \dfrac{\text{SCE}}{N-k}$		
Dans les groupes	35,16	470	0,07		
	SCT = SCG + SCE				
	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$	$N-1$			
Total	74,67	473			

En divisant la variance **entre** les groupes (13,17) par la variance **dans** les groupes (0,07), on obtient le *F* calculé : 176,09.

Intuitivement, plus le numérateur (variance entre les groupes) du *F* calculé est grand et le dénominateur est petit, plus on sera porté à renoncer à l'hypothèse (nulle) selon laquelle, au niveau des quatre populations (correspondant aux quatre niveaux d'instruction), les revenus sont en moyenne égaux, et à envisager qu'il existe des différences selon le niveau d'instruction.

La conclusion peut s'obtenir de façon plus rigoureuse sur le plan statistique en confrontant le *F* calculé à la table de Fisher. À $p = 0{,}05$, 3 degrés de liberté au numérateur et 470 degrés de liberté au dénominateur, il apparaît que *F* calculé dépasse de loin le chiffre repère de la table de Fisher (2,68 si $v_1 = 3$ et $v_2 = (120)$ avec $p = 0{,}05$; **2,69 si** $v_1 = 3$ et v_2 tend vers l'infini avec $p = 0{,}05$).

De façon plus formelle,

Si $F_{calc} \geq F_{table}$, on rejette l'hypothèse nulle.

Si $F_{calc} < F_{table}$, on accepte l'hypothèse nulle.

On peut consulter la table des valeurs théoriques de *F* de Fisher en tenant compte des degrés de liberté.

Degrés de liberté du numérateur : $k - 1$

Degrés de liberté du dénominateur : $N - k$

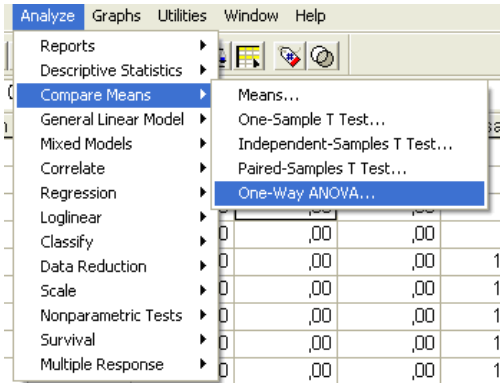
On peut, de manière équivalente, tester l'hypothèse de l'inexistence d'une relation entre les deux variables en examinant la valeur *p* correspondant à *F* calculé. Si cette valeur *p* est inférieure à 0,05, on rejette l'hypothèse nulle.

2. LES COMMANDES AVEC SPSS ET LE TRAITEMENT D'UN EXEMPLE

Avec SPSS, on peut obtenir le tableau de variance (ANOVA) de deux manières.

La première manière recourt à **One-Way ANOVA**.

On procède comme suit :



En sélectionnant **One-Way ANOVA**, on ouvre une nouvelle fenêtre qu'on remplira, pour l'exemple traité, de la façon suivante :

Figure 6.2

LES COMMANDES POUR L'ANALYSE DE VARIANCE : ONE-WAY ANOVA

On remarquera que la variable « Niveau de diplomation » (« ndiplom ») contient les quatre catégories identifiées par les chiffres 1, 2, 3, 4.

Le listing[1] donne les résultats détaillés antérieurement.

1. À partir de maintenant, les encadrés seront présentés conformément à la sortie de listing de manière à nous rapprocher le plus possible des conditions de travail avec le logiciel.

Encadré 6.1

LE TABLEAU D'ANALYSE DE VARIANCE D'APRÈS ONE-WAY ANOVA

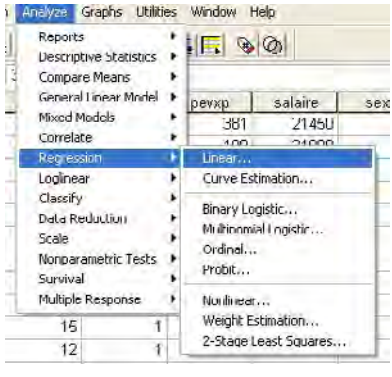
ANOVA

lsal

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	39,516	3	13,172	176,085	,000
Within Groups	35,158	470	,075		
Total	74,675	473			

Le même tableau peut s'obtenir à partir d'une régression multiple où la variable dépendante sera « Lsal » et où les variables indépendantes seront les variables muettes « niv1 », « niv3 », « niv4 ». « Niv1 » = 0 sauf pour les cas où le niveau de diplomation (« ndiplom ») = 1. « Niv1 » prendra dans ces cas la valeur 1. « Niv3 » = 0 sauf pour les cas où le niveau de diplomation (« ndiplom ») = 3. « Niv3 » prendra dans ces cas la valeur 1. « Niv4 » = 0 sauf pour les cas où le niveau de diplomation (« ndiplom ») = 4. « Niv2 », construit de la même manière, a été omis par choix méthodologique pour neutraliser un problème dit de colinéarité lié au fait que « niv1 », « niv2 », « niv3 », « niv4 » sont strictement complémentaires (la somme des 4 variables donnant une colonne de 1).

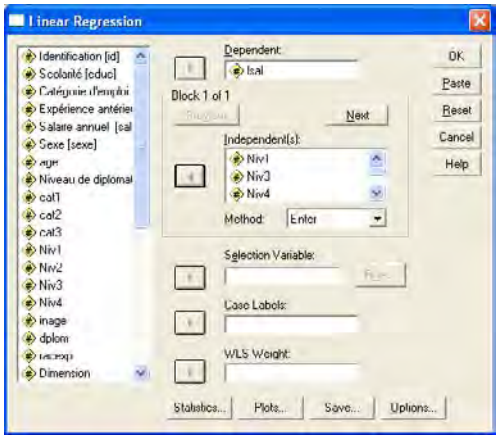
On procède comme suit :



À partir de **Linear,** on remplit la fenêtre suivante (figure 6.3) :

Figure 6.3

LES COMMANDES POUR L'ANALYSE DE VARIANCE À PARTIR DE LA RÉGRESSION



Sous des titres différents, on retrouvera dans le listing le même tableau ANOVA.

Encadré 6.2

L'ANALYSE DE VARIANCE À PARTIR DE LA RÉGRESSION LINÉAIRE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39,516	3	13,172	176,085	,000[a]
	Residual	35,158	470	,075		
	Total	74,675	473			

a. Predictors: (Constant), Niv4, Niv1, Niv3

b. Dependent Variable: lsal

C'est dire que la procédure ANOVA est sensible aux conditions que doit respecter la régression multiple !

CHAPITRE

7

Analyse bivariée

Corrélation et régression simple

Les instruments d'exploration de données empiriques se diversifient et s'affinent lorsque les deux variables sont quantitatives. On les utilisera, par exemple, pour examiner à partir d'une base de données la relation entre la variable « salaire » et la variable « années de scolarité ». La représentation graphique, la corrélation, la régression simple sont ici généralement privilégiées. Les trois approches sont en pratique indissociables, la corrélation explorant la relation entre deux variables en supposant qu'elles sont liées par une relation « linéaire ».

1. LA CORRÉLATION BIVARIÉE SIMPLE (CORRÉLATION DE PEARSON)

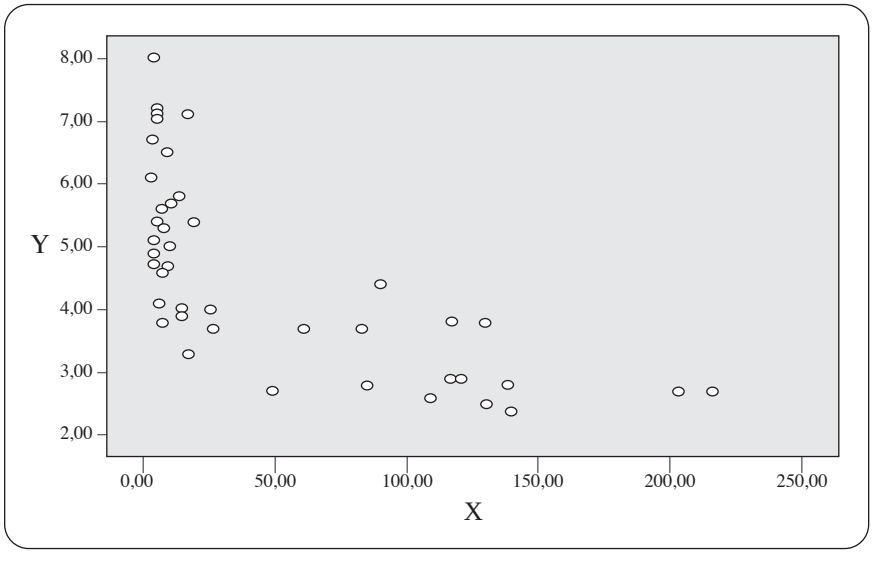
Pour présenter la corrélation bivariée simple (corrélation de Pearson) et sa relation à la régression simple, on fera appel à un fichier de données (Fecond.xls sous forme Excel ; Fecond.sav sous forme SPSS) qui exploite des informations provenant du rapport annuel 2003 du PNUD. Pour un certain nombre de pays, les taux de fertilité « Fertility Rate » (*Y*)

seront confrontés à une approximation du nombre moyen de personnes à charges d'un médecin (*X* exprime le nombre de médecins par 100 000 habitants).

Le graphique suivant laisse entrevoir une régularité entre *Y* et *X*.

Figure 7.1

***TAUX DE FERTILITÉ ET NOMBRE DE MÉDECINS PAR 100 000 HABITANTS DANS DIFFÉRENTS PAYS** D'APRÈS LE RAPPORT DU PNUD 2003*



1.1. REPÈRES THÉORIQUES

La corrélation bivariée simple (corrélation de Pearson) tente de donner une synthèse de la régularité que l'on devine dans le graphique en supposant qu'une droite est capable de « rassembler » au mieux (« le plus près possible de la droite ») les divers points du graphique.

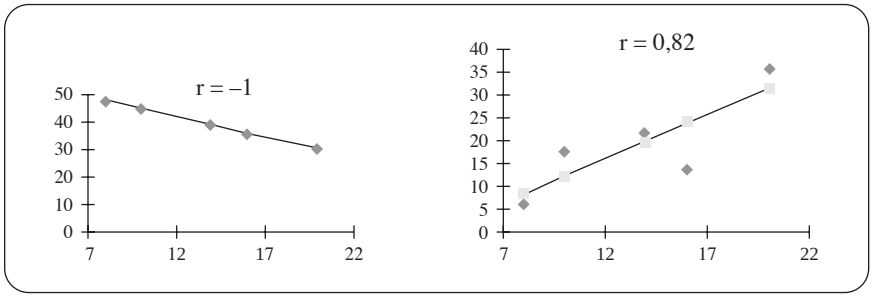
Le coefficient de corrélation simple prend ses valeurs à l'intérieur de l'intervalle –1 et +1.

$$-1 \leq r \leq 1$$

Le signe positif ou négatif du coefficient de corrélation (*r*) correspond à l'orientation de la pente de la droite autour de laquelle se regroupent les divers points du nuage de points.

Figure 7.2

LA RELATION LINÉAIRE SOUS-JACENTE À LA CORRÉLATION SIMPLE



Le coefficient ne prend sa signification que pour des ensembles de données susceptibles d'être résumées graphiquement autour d'une droite. La figure 7.1 des données (Y_i, X_i) évoque une situation dont la régularité risque d'échapper partiellement au coefficient de corrélation linéaire simple.

L'expression qui définit le coefficient de corrélation linéaire est la suivante :

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}$$

Les deux tableaux suivants permettent de suivre le détail du calcul du coefficient de corrélation (*r*).

Tableau 7.1

DÉTAILS DU CALCUL DU COEFFICIENT DE CORRÉLATION SIMPLE (PREMIÈRE PARTIE)

	Y_i	X_i	$y_i = Y_i - \overline{Y}$	$x_i = X_i - \overline{X}$	$x_i y_i$	y_i^2	x_i^2
Mexique	2,5	130	-2,12	82,32	-174,45	4,49	6776,44
Colombie	2,6	109	-2,02	61,32	-123,81	4,08	3760,04
Brésil	2,2	158	-2,42	110,32	-266,88	5,85	12170,31
Paraguay	3,8	117	-0,82	69,32	-56,78	0,67	4805,14
République Dominicaine	2,7	216	-1,92	168,32	-323,03	3,68	28331,34
Équateur	2,8	138	-1,82	90,32	-164,30	3,31	8157,55
El Salvador	2,9	121	-1,72	73,32	-126,05	2,96	5375,70
Bolivie	3,8	130	-0,82	82,32	-67,43	0,67	6776,44
Honduras	3,7	83	-0,92	35,32	-32,46	0,84	1247,44
Guatemala	4,4	90	-0,22	42,32	-9,27	0,05	1790,91
Nicaragua	3,7	61	-0,92	13,32	-12,24	0,84	177,40
Haïti	4	25	-0,62	-22,68	14,04	0,38	514,42
Venezuela	2,7	203	-1,92	155,32	-298,08	3,68	24124,04
Jamaïque	2,4	140	-2,22	92,32	-204,87	4,92	8522,83
Pérou	2,9	117	-1,72	69,32	-119,17	2,96	4805,14
Cap Vert	3,3	17	-1,32	-30,68	40,47	1,74	941,31
Algérie	2,8	85	-1,82	37,32	-67,89	3,31	1392,72
Botswana	3,7	26	-0,92	-21,68	19,93	0,84	470,06
Maroc	2,7	49	-1,92	1,32	-2,53	3,68	1,74
Ghana	4,1	6	-0,52	-41,68	21,64	0,27	1737,29
Lesotho	3,8	7	-0,82	-40,68	33,32	0,67	1654,93
Togo	5,3	8	0,68	-39,68	-27,02	0,46	1574,57
Cameroun	4,6	7	-0,02	-40,68	0,78	0,00	1654,93
Zimbabwe	3,9	14	-0,72	-33,68	24,22	0,52	1134,40
Kenya	4	14	-0,62	-33,68	20,85	0,38	1134,40
Ouganda	7,1	5	2,48	-42,68	-105,88	6,15	1821,66
Madagascar	5,7	11	1,08	-36,68	-39,65	1,17	1345,48
Gambie	4,7	4	0,08	-43,68	-3,53	0,01	1908,02
Nigeria	5,4	19	0,78	-28,68	-22,40	0,61	822,59
Mauritanie	5,8	14	1,18	-33,68	-39,77	1,39	1134,40
Érythrée	5,4	5	0,78	-42,68	-33,33	0,61	1821,66
Sénégal	5	10	0,38	-37,68	-14,35	0,15	1419,85
Guinée	5,8	13	1,18	-34,68	-40,95	1,39	1202,76
Bénin	5,7	10	1,08	-37,68	-40,73	1,17	1419,85
Tanzanie	5,1	4	0,48	-43,68	-21,00	0,23	1908,02
Côte d'Ivoire	4,7	9	0,08	-38,68	-3,13	0,01	1496,21
Zambie	5,6	7	0,98	-40,68	-39,90	0,96	1654,93
Angola	7,2	5	2,58	-42,68	-110,15	6,66	1821,66
Tchad	6,7	3	2,08	-44,68	-92,97	4,33	1996,38
Guinée-Bissau	7,1	17	2,48	-30,68	-76,11	6,15	941,31
République Centrafricaine	4,9	4	0,28	-43,68	-12,27	0,08	1908,02
Éthiopie	6,1	3	1,48	-44,68	-66,17	2,19	1996,38
Mozambique	5,6	6	0,98	-41,68	-40,88	0,96	1737,29
Mali	7	5	2,38	-42,68	-101,62	5,67	1821,66
Burkina Faso	6,7	3	2,08	-44,68	-92,97	4,33	1996,38
Niger	8	4	3,38	-43,68	-147,68	11,43	1908,02
Sierra Leone	6,5	9	1,88	-38,68	-72,75	3,54	1496,21
	4,62	**47,68**	**0,00**	**0,00**	**-3119,21**	**110,47**	**164610,21**
						10,51	**405,72**

$r = -0,731$

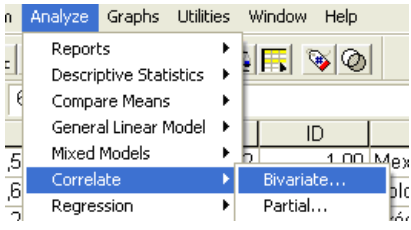
Tableau 7.2

DÉTAILS DU CALCUL DU COEFFICIENT DE CORRÉLATION (DEUXIÈME PARTIE)

$\bar{Y} = \left(\sum Y_i\right)/N$	$\bar{X} = \left(\sum X_i\right)/N$	$\sum y_i x_i$	$\sum y_i^2$	$\sum x_i^2$
4,62	47,68	–3119,21	110,47	164610,21
			$\sqrt{\sum y_i^2}$	$\sqrt{\sum x_i^2}$
			10,51	405,72
			$r = \dfrac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = 0,731$	

1.2. COMMANDES SPSS ET LE TRAITEMENT D'UN EXEMPLE

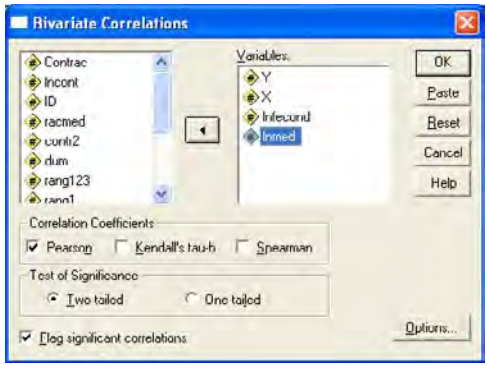
Pour accéder au calcul du coefficient de corrélation en SPSS, on procède comme suit :



En sélectionnant **Bivariate**, on fait apparaître une nouvelle fenêtre qu'on peut compléter comme indiqué ci-dessous :

Figure 7.3

COMMANDES SPSS POUR LA CORRÉLATION BIVARIÉE : FENÊTRE PRINCIPALE



Dans la nouvelle fenêtre, « Y », « X », « lnfecond », « lnmed » ont été introduits sous le terme **VARIABLES** (N.B. : on aurait pu introduire plus de variables). Les variables « lnfecond » et « lnmed » correspondent au logarithme naturel de *Y* et de *X*. Les corrélations seront calculées pour les différents couples de variables qui peuvent être formés à partir des quatre variables.

L'extrait ci-joint du listing présente les corrélations obtenues.

Encadré 7.1

LA MATRICE DE CORRÉLATION D'APRÈS SPSS

Correlations

		Y	X	lnfecond	lnmed
Y	Pearson Correlation	1	-,731**	,987**	-,802**
	Sig. (2-tailed)		,000	,000	,000
	N	47	47	47	47
X	Pearson Correlation	-,731**	1	-,787**	,921**
	Sig. (2-tailed)	,000		,000	,000
	N	47	47	47	47
lnfecond	Pearson Correlation	,987**	-,787**	1	-,835**
	Sig. (2-tailed)	,000	,000		,000
	N	47	47	47	47
lnmed	Pearson Correlation	-,802**	,921**	-,835**	1
	Sig. (2-tailed)	,000	,000	,000	
	N	47	47	47	47

**. Correlation is significant at the 0.01 level (2-tailed).

Le tableau de corrélations est strictement symétrique. On peut, par exemple, lire les corrélations de *Y* avec « Y », « X », « lnfecond », « lnmed » de façon horizontale ou verticale. Les corrélations calculées ont été évaluées d'un point de vue statistique. Le listing indique le niveau de significativité de deux manières, l'une chiffrée, l'autre par « * ». En fait, le coefficient de corrélation se prête relativement mal à une évaluation statistiquement rigoureuse. Par contre, la transformation suivante permet une évaluation à partir de la table de Fisher.

$$F_{calc} = \frac{r^2 (N-2)}{(1-r^2)}$$

où *N* indique le nombre d'individus considérés dans l'échantillon.

Si $F_{calc} > F_{table}$, on rejette l'hypothèse H_0 de l'absence de lien statistique entre *Y* et *X*.

Dans le cas de la corrélation simple, le F_{table} est choisi dans la table en fonction des degrés de liberté 1 et $(N - 2)$.

Le *F* calculé associé au coefficient de corrélation simple figure dans le listing produit en demandant une régression simple entre *Y* et *X*.

En examinant le tableau de corrélation (encadré 7.1), on remarquera que la corrélation entre *Y* et *X* vaut –0,731, mais –0,835 pour le couple de variables « lnfecond » et « lnmed ». La droite sous-jacente à la corrélation rassemble mieux les points du nuage de points dans un graphique associant les données de « lnfecond » et « lnmed » que dans un graphique associant *Y* et *X*. C'est une indication qu'une courbe résumerait plus adéquatement le nuage des couples Y_i, X_i.

2. L'ANALYSE DE RÉGRESSION SIMPLE

Pour identifier la droite résumant « au mieux » le nuage des couples Y_i, X_i, on recourt à **l'analyse de régression simple**. L'analyse de régression simple est un cas particulier de l'analyse de régression multiple. L'analyse de régression multiple prend en considération plusieurs variables indépendantes pour rendre compte de la variable dépendante (*Y*). L'analyse de régression simple se limite à une variable indépendante (*X*). L'analyse de régression simple est apparemment moins complexe que l'analyse de régression multiple. Elle échappe en particulier au problème dit de « colinéarité ou proximité statistique » entre les variables indépendantes introduites dans une régression multiple. Par contre, comme on se limite à une seule variable indépendante dans la régression simple, les résidus (les différences entre les Y_i observés et les Y_i calculés, désignés aussi par YC_i dans le texte) présentent fréquemment de fortes régularités. La régression simple sera présentée ici comme instrument exploratoire des relations bivariées, complémentaire à la corrélation.

À partir du critère de « minimisation de la somme des carrés des résidus », la régression simple permet d'identifier la droite qui rassemble « au mieux » autour d'elle le nuage de points représentant les couples d'observations (Y_i, X_i). C'est la droite sous-jacente à la corrélation. On se reportera au chapitre sur la régression multiple pour une présentation plus complète de la régression.

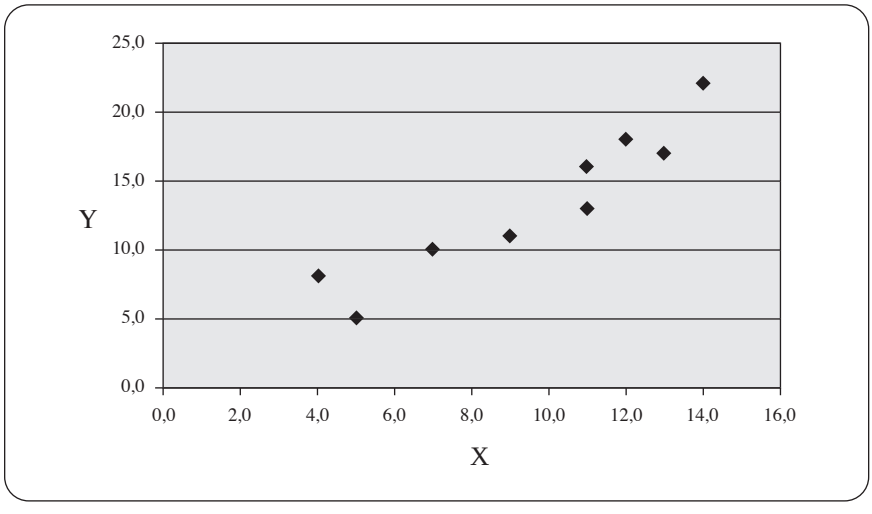
La régression simple sera tout d'abord présentée à partir d'un exemple simplifié qui permettra de suivre par le détail la logique des calculs conduisant aux « estimations » du modèle sous-jacent à l'analyse de régression. L'exemple, fictif, comporte 10 couples d'observations

(Y_i, X_i). Disons que X_i représente le nombre de cigarettes qu'une personne (i) a fumées le jour où l'information a été recueillie. Y_i représente le nombre de quintes de toux pour cette même personne (i) durant la journée d'observation. Nous reviendrons ensuite à l'exemple tiré de Fecond.sav utilisé pour la présentation de la corrélation.

Figure 7.4

UN EXEMPLE SIMPLIFIÉ POUR PRÉSENTER L'ANALYSE DE RÉGRESSION

ID	X	Y
1	4,0	8,0
2	5,0	5,0
3	7,0	10,0
4	9,0	11,0
5	11,0	13,0
6	12,0	18,0
7	13,0	17,0
8	14,0	22,0
9	12,0	18,0
10	11,0	16,0



Quelle est la droite qui rassemblerait « au mieux » les divers points du nuage de points ?

2.1. REPÈRES THÉORIQUES

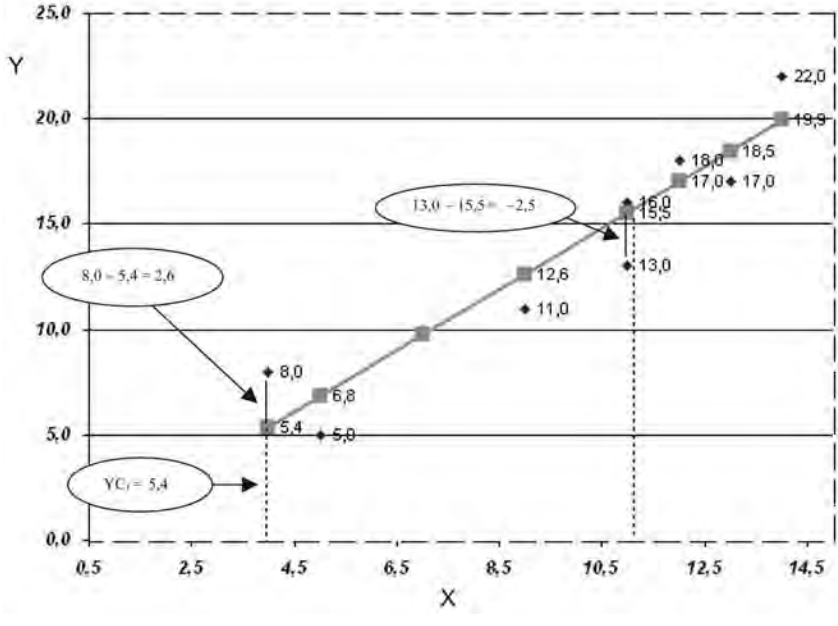
Pour approximer cette droite, les statisticiens utilisent habituellement le critère de minimisation de la somme des carrés des résidus. Un résidu correspond à la distance verticale positive ou négative entre le Y_i observé

et le Y_i calculé, communément désigné par le symbole $\hat{Y}_i$ (que nous désignerons aussi par YC_i). La figure suivante permet de visualiser les résidus en supposant disponible la droite de régression.

Figure 7.5

LA DROITE DE RÉGRESSION SUR LA BASE DE L'EXEMPLE TRAITÉ

ID	X	Y	YC	Résidus
1	4,0	8,0	5,4	2,6
2	5,0	5,0	6,8	-1,8
3	7,0	10,0	9,7	0,3
4	9,0	11,0	12,6	-1,6
5	11,0	13,0	15,5	-2,5
6	12,0	18,0	17,0	1,0
7	13,0	17,0	18,5	-1,5
8	14,0	22,0	19,9	2,1
9	12,0	18,0	17,0	1,0
10	11,0	16,0	15,5	0,5



La droite sur le graphique correspond de fait à la droite calculée par régression sur la base de la minimisation de la somme des carrés des résidus.

Au départ, cette droite n'est pas disponible, mais on dispose de la formule mathématique de la minimisation de la somme des carrés des résidus.

$$Min \sum (Y_i - \hat{Y}_i)^2$$

où

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Cette dernière expression représentant la droite de régression sur laquelle se situent les $\hat{Y}_i$ calculés (ou YC_i).

$\hat{\beta}_1$ représente la pente de la droite tandis que $\hat{\beta}_0$ en représente l'ordonnée à l'origine.

À partir du traitement de la formule de minimisation de la somme des carrés des résidus, il est possible de calculer les coefficients β_1 et $\hat{\beta}_0$, et par la suite les $\hat{Y}_i$ et les résidus $r_i = Y_i - \hat{Y}_i$.

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Le tableau suivant présente le détail des calculs.

Tableau 7.3

LA RÉGRESSION SIMPLE : LES CALCULS DE BASE

ID	X	Y	$x_i = X_i - \overline{X}$	$y_i = Y_i - \overline{Y}$	$x_i y_i$	x_i^2	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X$	$r_i = Y_i - \hat{Y}_i$	r_i^2	y_i^2
1	4,0	8,0	-5,80	-5,80	33,64	33,64	5,36	2,64	6,95	33,64
2	5,0	5,0	-4,80	-8,80	42,24	23,04	6,82	-1,82	3,31	77,44
3	7,0	10,0	-2,80	-3,80	10,64	7,84	9,73	0,27	0,07	14,44
4	9,0	11,0	-0,80	-2,80	2,24	0,64	12,64	-1,64	2,68	7,84
5	11,0	13,0	1,20	-0,80	-0,96	1,44	15,55	-2,55	6,48	0,64
6	12,0	18,0	2,20	4,20	9,24	4,84	17,00	1,00	1,00	17,64
7	13,0	17,0	3,20	3,20	10,24	10,24	18,45	-1,45	2,12	10,24
8	14,0	22,0	4,20	8,20	34,44	17,64	19,91	2,09	4,37	67,24
9	12,0	18,0	2,20	4,20	9,24	4,84	17,00	1,00	1,00	17,64
10	11,0	16,0	1,20	2,20	2,64	1,44	15,55	0,45	0,21	4,84
	$\overline{X} = \frac{\sum X_i}{10}$	$\overline{Y} = \frac{\sum Y_i}{10}$			$\sum x_i y_i$	$\sum x_i^2$			$\sum r_i^2$	$\sum y_i^2$
	9,80	13,80			153,60	105,60			28,18	251,60

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{153,60}{105,60} = 1,45455$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 13,80 - 1,45455 \times 9,80 = -0,45455$$

$$R^2 = 1 - \frac{\sum r_i^2}{\sum y_i^2} = 1 - \frac{28,18}{251,60} = 0,89$$

C'est en utilisant les couples **(Y_i, X_i)** et **($\hat{Y}_i$, X_i)** qu'a été construit le graphique 7.4 du nuage de points et de la droite de régression autour de laquelle les points se regroupent.

Le tableau 7.3 présente aussi le calcul de R^2. ***R^2* est connu sous le nom de coefficient de détermination**. Il prend ses valeurs entre 0 et 1. Si la variable indépendante rend « bien » compte de la variable dépendante, la somme des carrés des résidus ($\sum r_i^2$) sera relativement peu importante et, par conséquent, le poids de la fraction dans l'expression qui définit R^2 sera relativement faible. R^2 sera dans ce cas relativement proche de un. Par contre, R^2 se rapproche de zéro dans la mesure où la variable indépendante rend « peu » compte de la variable dépendante.

Dans le cas de la régression simple, le coefficient de détermination est égal au carré du coefficient de corrélation.

$R^2 = r^2$

Cette égalité ne se maintient pas lorsqu'il s'agit d'une régression multiple.

À partir des informations figurant dans le tableau 7.3, il est possible de construire le tableau d'analyse de variance (ANOVA).

TABLEAU ANOVA, F ET R^2 EN RÉGRESSION SIMPLE

	Somme des carrés	***Degrés de liberté***	***Variance***	***F***	***Sig.***
Expliqué par la régression	$\sum \hat{y}_i^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$		$\frac{\sum y_i^2}{1}$	$F_{calc} = \frac{\sum y_i^2}{1.s^2} = \frac{R^2.(n-2)}{(1-R^2)}$	
	223,42	$k = 1$	223,42	63,42	0,00
Résiduel ou inexpliqué	$\sum r_i^2 = \sum \left(Y_i - \hat{Y}\right)^2$		$s^2 = \frac{\sum y_i^2}{n-2}$		
	28,18	$n - 2 = 8$	3,52		
Total	$\sum y_i^2 = \sum \left(Y_i - \bar{Y}\right)^2$				
	251,60	$n - 1 = 9$			

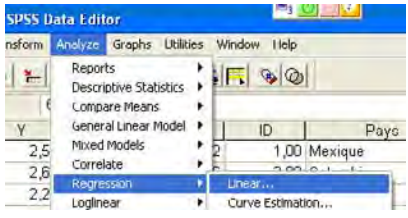
Le *F* calculé (63,42), statistiquement très significatif (Sig. : 0,00), permet d'écarter l'hypothèse nulle de la non-incidence de *X* sur *Y*.

L'ensemble des calculs qui viennent d'être évoqués à titre d'introduction à l'analyse de régression simple est pris en charge par SPSS.

2.2. LES COMMANDES SPSS ET LE TRAITEMENT D'UN EXEMPLE

Revenons maintenant à l'exemple traité lors de la présentation de la corrélation (fichier Fecond.sav) : les taux de fertilité (Y_i) confrontés à une mesure du nombre moyen de personnes à charge d'un médecin (X_i).

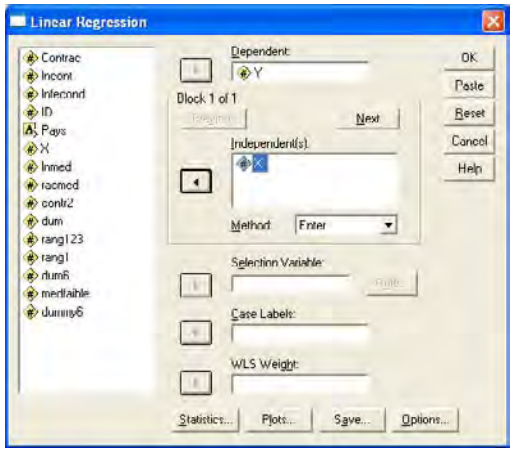
Pour faire appel à la procédure de régression simple en SPSS, on procède comme suit :



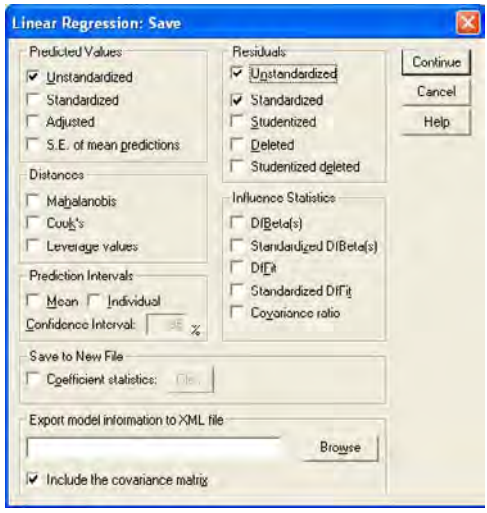
Après avoir cliqué sur **Linear**, on complétera comme indiqué la fenêtre qui s'ouvre.

Figure 7.6

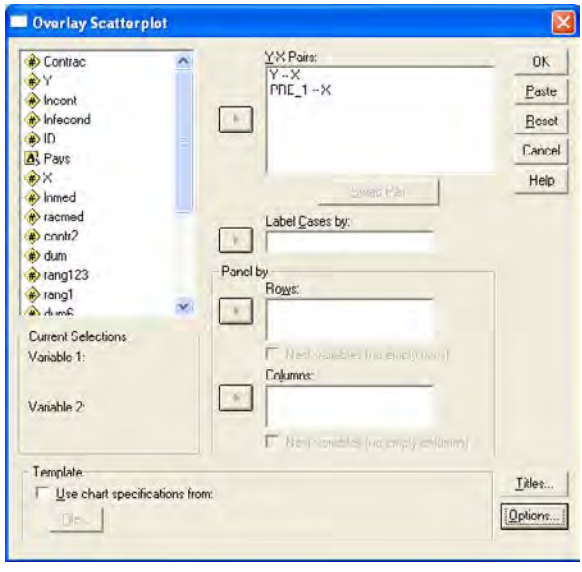
COMMANDE SPSS POUR LA RÉGRESSION LINÉAIRE : FENÊTRE PRINCIPALE



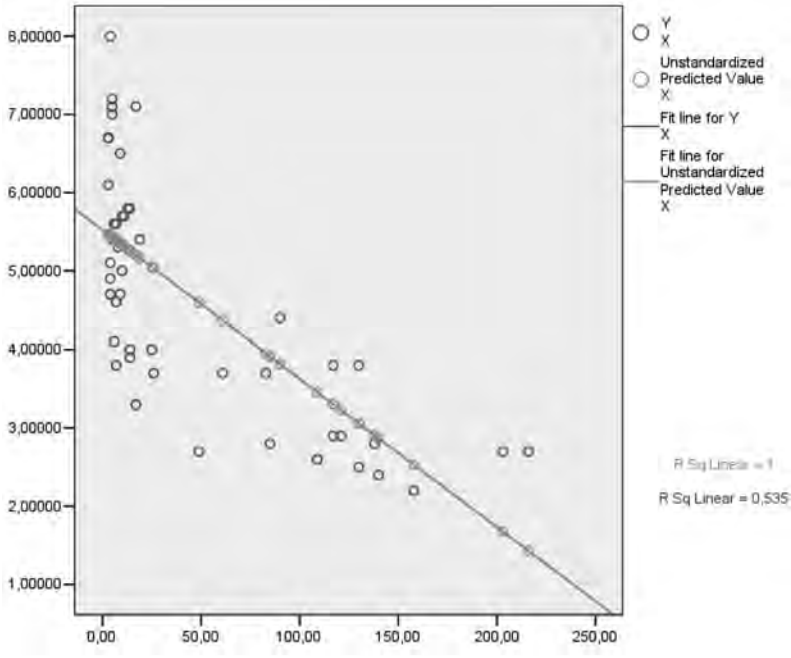
À partir de **SAVE** en bas de la figure précédente, on demande les $\hat{Y}_i$ calculés (« prévus ») par la régression et les résidus sous forme non standardisée et sous forme standardisée. Les variables demandées viendront s'ajouter au fichier de données.



En recourant à la procédure graphique **Graphs, Scatter/dot, Overlay Scatter**,



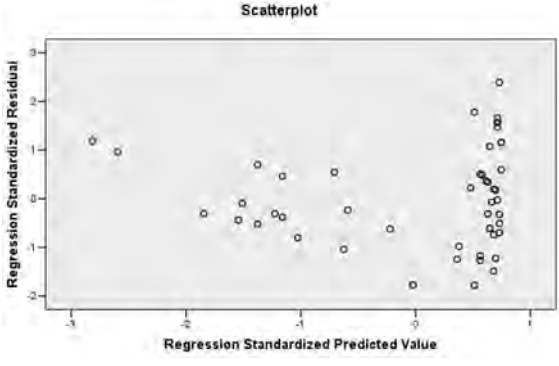
on obtient le graphique suivant :



La droite de régression n'arrive pas à capter une partie de la régularité du nuage de points. Le graphique des résidus standardisés en fonction des $\hat{Y}_i$ calculés standardisés confirme d'ailleurs cette impression.

Figure 7.7

GRAPHIQUE DES RÉSIDUS STANDARDISÉS EN FONCTION DES $\hat{Y}_i$ CALCULÉS STANDARDISÉS



Dans le listing de la régression simple de *Y* sur *X*, on retrouvera les renseignements permettant de faire le lien entre la corrélation et l'analyse de régression simple.

Encadré 7.2

EXTRAITS DU LISTING DE RÉGRESSION SIMPLE

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,731(a)	,535	,525	1,06840

a Predictors: (Constant), X
b Dependent Variable: Y

$R^2 = 0{,}535$ = le carré de la corrélation dans le cas de la régression simple

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	59,11	1,00	59,11	51,78	0,00
	Residual	51,37	45,00	1,14		
	Total	110,47	46,00			

a Predictors: (Constant), X
b Dependent Variable: Y

$$F_{calculé} = \frac{R^2(N-2)}{(1-R^2)} = \frac{r^2(N-2)}{(1-r^2)}$$

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5,52	0,20		27,60	0,00
	X	-0,02	0,00	-0,73	-7,20	0,00

a Dependent Variable: Y

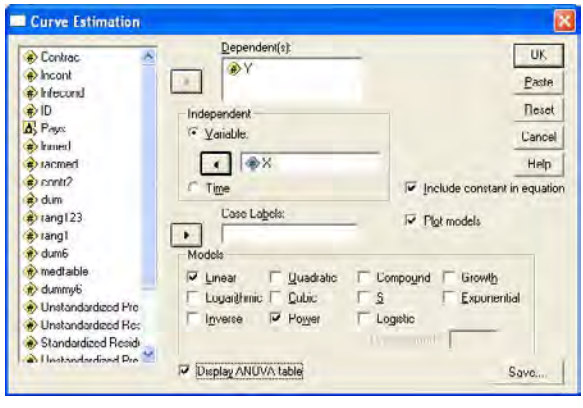
Pente de la droite de régression = -0,02

Le programme SPSS permet d'explorer un certain nombre de « courbes » dont l'expression mathématique se prête à une transformation de linéarisation et, par ce biais, à des estimations par régression obtenues sur la base du critère de minimisation de la somme des carrés des résidus. On accède à la procédure de la manière suivante :

En cliquant sur **Curve Estimation**, il est possible d'explorer quelques modèles.

Figure 7.8

LA COMMANDE SPSS CURVE ESTIMATION



N.B. : Avec le bouton droit de la souris, il est possible de faire apparaître l'expression mathématique du modèle, en amenant la souris sur le nom du modèle.

Dans le cas traité, deux modèles ont été pris en considération, le modèle linéaire (**LINEAR**) et le modèle Puissance (**POWER**).

Le graphique de la droite et de la courbe estimées a été demandé (**PLOT MODELS**).

L'extrait de listing montre une partie des résultats des estimations et le graphique demandé.

Encadré 7.3

LES RÉSULTATS DES ESTIMATIONS À PARTIR DE LA PROCÉDURE « REGRESSION… CURVE ESTIMATIONS »

Model Summary and Parameter Estimates
Dependent Variable: Y

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	0,54	51,78	1,00	45,00	0,00	5,52	-0,02
Power	0,70	103,73	1,00	45,00	0.00	8,12	-0,21

The independent variable is X.

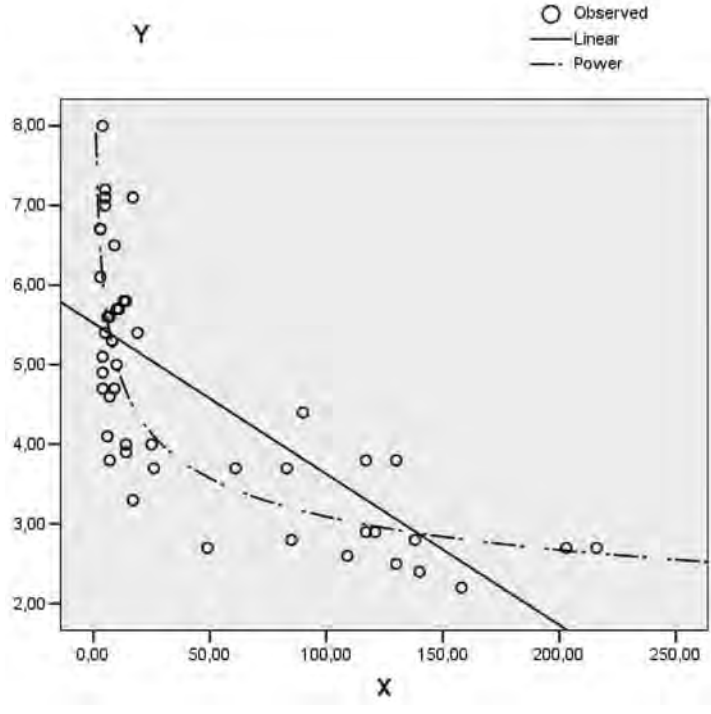
0,54 est le carré de la correlation entre *Y* et *X*.

0,70 est la carré de la corrélation entre *Lnfecond* et *Lnmed*.

Comparativement à la droite estimée, la courbe estimée donne une image « plus adéquate » de la régularité du nuage de points.

Figure 7.9

GRAPHIQUE DE COURBES ESTIMÉES (MODÈLE LINÉAIRE ET MODÈLE PUISSANCE) – ÉTUDE DE CAS



Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

CHAPITRE

8

La régression multiple

La régression linéaire multiple généralise l'approche adoptée dans la régression linéaire simple. La régression linéaire simple renvoie à un modèle où une variable dépendante (Y_i) est interprétée selon une relation linéaire en fonction d'une variable indépendante (X_i) et d'un terme d'erreur (ε_i).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

Dans la régression multiple, le nombre de variables indépendantes est supérieur ou égal à 2, mais inférieur au nombre de situations (observations) considérées.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_{k_3} X_{ki} + \varepsilon_i$$

avec $k < n$

où n désigne le nombre de situations (observations) considérées et k, le nombre de variables indépendantes (terme constant exclu).

En pratique, le nombre de variables indépendantes est relativement limité en raison des problèmes d'estimation et d'interprétation qui accompagnent l'augmentation du nombre de variables indépendantes figurant dans le modèle étudié.

1. REPÈRES THÉORIQUES

Les caractéristiques du modèle sous-jacent à l'analyse de régression sont déterminantes à la fois dans le choix des modalités de traitement et dans l'interprétation des résultats empiriques. Une attention particulière à une formulation adéquate du modèle et à ses implications s'impose dès lors avant d'entreprendre le traitement multivarié par régression.

Le modèle : $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_{k_3} X_{ki} + \varepsilon_i$

On remarquera tout d'abord que le recours à l'indice *i*, $i = 1, 2, 3..., n$ permet l'écriture succincte de *n* équations.

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + ... + \beta_{k_3} X_{k1} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + ... + \beta_{k_3} X_{k2} + \varepsilon_2$$

…

Dans ces équations, le terme de gauche correspond à la variable dépendante. « Les variables dépendantes » Y_i correspondent à un même phénomène considéré dans *n* situations, par exemple, le revenu pour la personne *1*, pour la personne *2*, …, pour la personne *n*. Les variables dépendantes (Y_i) sont en principe des variables quantitatives continues.

Les variables indépendantes X_{1i}, X_{2i}… (nous les désignerons aussi sous le vocable de « variable explicative ») sont censées « rendre compte » de la variable dépendante Y_i.

Pour chaque situation *i* considérée, les variables indépendantes et le terme constant (de valeur 1), pondérés par les paramètres $\beta_0, \beta_1, \beta_2 ... \beta_k$ respectifs, et le terme d'erreur sont associés à chaque variable dépendante (Y_i) selon une relation additive.

Ce type de relation peut paraître contraignant. Il donne néanmoins accès à une grande variété de modèles pour une double raison. Par transformations mathématiques, de nombreux modèles peuvent se ramener au modèle type sous-jacent à l'analyse de régression. Par ailleurs, les variables indépendantes X_{1i}, X_{2i}... peuvent contourner la contrainte dite de « linéarité ». X_{1i}, par exemple peut correspondre au carré de la variable Z_{1i} : $X_{1i} = Z_{1i}^{\ 2}$; ou encore X_{2i} peut correspondre au produit de deux variables : $X_{2i} = Z_{1i} * Z_{2i}$. Compte tenu des deux exemples, le modèle de référence correspondrait en pratique à l'expression suivante :

$$Y_i = \beta_0 + \beta_1 (Z_{1i}^{\ 2}) + \beta_2 (Z_{1i} \times Z_{2i}) + ... + X_{ki} + \varepsilon_i$$

Les variables indépendantes

Le contenu des variables indépendantes dépend beaucoup de la démarche de recherche en cours. S'agit-il simplement d'un intérêt pour la prévision où l'on se préoccupe avant tout d'obtenir les « meilleures prévisions » de Y_i compte tenu d'un certain nombre de variables indépendantes et de certains critères de performance prévisionnels ? S'agit-il plutôt d'une démarche explicative s'appuyant sur un corpus théorique bien établi ou d'une analyse s'appuyant sur une argumentation de « bon sens » ? Dans les sciences humaines, l'avancement des connaissances se réalise fréquemment par un va-et-vient entre une réflexion théorique et une interrogation empirique, l'observation des données aidant à formuler progressivement les questions d'ordre théorique. L'introduction de variables explicatives dans un modèle se fait dès lors très souvent à travers un processus de quasi-tâtonnement. L'expérience révèle tout de même que les modélisations les moins problématiques pour le travail empirique correspondent aux démarches menées analytiquement avec le maximum de rigueur.

Les variables indépendantes sont de types variés. Elles peuvent être quantitatives (par exemple l'âge ou le nombre d'années de scolarité). Elles peuvent être qualitatives (par exemple la distinction homme-femme ou l'appartenance à une minorité). Dans les variables indépendantes, on peut retrouver des variables « décalées dans le temps ». À titre d'exemple, des dépenses consacrées à des voyages de tourisme à l'étranger peuvent dépendre des dépenses faites auparavant pour ce type de voyages. Certaines variables indépendantes peuvent avoir un très faible contenu explicatif (par exemple la séquence du temps exprimée en jours, en trimestres ou en années) ou être introduites à titre instrumental pour permettre de mieux dégager l'effet spécifique d'une variable indépendante

sur la variable dépendante. Certaines variables explicatives peuvent être « endogènes », c'est-à-dire s'expliquer elles-mêmes par une équation où elles figurent comme variable dépendante et dans laquelle on peut éventuellement retrouver une ou plusieurs autres variables explicatives endogènes. Très souvent, les variables indépendantes prennent leurs valeurs indépendamment du modèle. Elles sont alors dites exogènes. On disposera par exemple du nombre d'années de scolarité (X_{1i}) ou de l'âge (X_{2i}) pour chaque personne appartenant à l'échantillon considéré.

Notre présentation de la régression multiple renvoie à une modélisation ne comportant que des variables indépendantes exogènes prenant des valeurs données (sans interférence aléatoire) selon la situation considérée. Par ailleurs, ne seront pas abordés les problèmes spécifiques au traitement des séries temporelles par régression. On se référera pour cet aspect au livre de Jean Stafford et Bruno Sarrasin, *La Prévision-prospective en gestion* et aux classiques en économétrie (le livre de Peter Kennedy, *A Guide to Econometrics*, est pratiquement un incontournable pour une entrée en matière).

Les paramètres $\beta_0, \beta_1, \beta_2, \ldots \beta_k$

Les paramètres $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ prennent des valeurs uniques qui se maintiennent constantes, quelle que soit la situation considérée. Ces valeurs sont *a priori* inconnues. La démarche de traitement des données se chargera d'extraire des estimations (approximations) de ces paramètres. Il revient néanmoins à la démarche de recherche de préciser des attentes *a priori* sur le signe affectant les paramètres et, le cas échéant, sur certains ordres de grandeur de ces paramètres. La formulation des attentes *a priori* dépend de l'état d'avancement des connaissances dans le domaine de recherche. Dans de nombreuses situations, la formulation de l'attente *a priori* est une question de « bon sens ». Parfois, il est difficile de préciser l'attente *a priori* parce que la variable indépendante recouvre plusieurs possibilités de réactions. Dans les cas où domine la préoccupation prévisionnelle, il arrive aussi que l'attente *a priori* ne soit pas énoncée.

Le paramètre β_i associé à une variable indépendante, par exemple β_2 associé à X_{2i}, traduit l'effet spécifique ou partiel de la variable indépendante concernée sur la variable dépendante Y_i, compte tenu des autres variables indépendantes insérées dans le modèle. Ainsi, l'incidence de X_{2i} sur Y_i varie selon les valeurs prises par X_{2i} et compte tenu de β_2 qui, lui, reste invariable. Voici un exemple fictif :

Tableau 8.1

L'INCIDENCE SPÉCIFIQUE D'UNE VARIABLE INDÉPENDANTE SUR LA VARIABLE DÉPENDANTE

i	β_2	X_{2i}	***Incidence partielle sur*** Y_i
1	3	15	$3 \times 15 = 45$
2	3	20	$3 \times 20 = 60$
3	3	34	$3 \times 34 = 102$
4	3	26	$3 \times 26 = 78$

La capacité d'un paramètre de traduire, à la manière d'une pondération, l'impact partiel spécifique d'une variable indépendante implique que chaque variable indépendante est distincte des autres variables indépendantes. Lorsque l'information que traduit une variable explicative peut être totalement ou partiellement obtenue par une autre variable indépendante, le paramètre qui lui est associé perd sa capacité de traduire l'effet spécifique partiel de la variable concernée, en raison même du caractère hybride de la variable. Il est dès lors de première importance dans la formulation d'un modèle d'identifier des variables indépendantes qui, par leur contenu notionnel, comportent le minimum de chevauchements avec d'autres variables indépendantes. Les chevauchements de contenu peuvent subvenir tant avec des variables quantitatives qu'avec des variables qualitatives. À titre d'exemple, les variables quantitatives suivantes comportent un risque de chevauchement : le nombre de jours non travaillés, le nombre de jours de congé, le nombre de jours d'absence pour maladie. Il s'agirait, dans le cas donné à titre d'exemple, de reformuler les variables de telle manière que chacune renvoie à un contenu distinct. Pour leur part, les variables qualitatives sont construites à partir de catégories exhaustives et exclusives ; c'est par exemple le cas d'une variable identifiant, dans un ensemble de personnes, les personnes qui vivent en ville, une variable identifiant celles qui vivent en périphérie urbaine et une variable identifiant celles qui vivent à la campagne. Comme il s'agit de catégories exhaustives et exclusives, l'appartenance ou la non-appartenance à deux catégories détermine automatiquement l'appartenance ou la non-appartenance à la troisième catégorie. Lors de la formulation du modèle, il s'agira de rompre cette stricte complémentarité par l'omission d'une des trois catégories sous peine de rendre impossible le calcul des valeurs approchées (estimations) des paramètres à partir des données empiriques.

Malgré les précautions prises pour éviter le chevauchement notionnel de variables indépendantes, il arrive fréquemment que des variables indépendantes soient statistiquement fortement corrélées entre elles. Ce problème dit de « colinéarité » risque de perturber la précision des estimations et exige un traitement spécifique.

Les termes d'erreur

Les termes d'erreur ε_i sont des variables aléatoires couvrant les phénomènes non explicitement pris en considération dans l'analyse ou encore des erreurs de mesure ou d'approximation. Le terme d'erreur se justifie aussi par le caractère imprévu qu'implique tout comportement humain.

Les termes d'erreur sont régis par les hypothèses suivantes.

- La moyenne de chaque terme d'erreur est égale à zéro : $E(\varepsilon_i) = 0$.
- La variance (inconnue) σ^2 de chaque terme d'erreur est constante.
- Les termes d'erreur ne sont pas corrélés entre eux.
- Les termes d'erreur ne sont pas corrélés avec les variables indépendantes.
- Les fluctuations de chaque terme d'erreur sont distribuées selon une distribution normale : $\varepsilon_i \sim N(0, \sigma^2)$.

Ces hypothèses sont requises pour que le traitement par régression aboutisse à des estimations permettant des inférences fiables. Elles sont congruentes avec la notion même du terme d'erreur, censé regrouper des facteurs secondaires non identifiés qui, « en moyenne », devraient se neutraliser.

La figure 8.1 résume le modèle et les hypothèses qui guident l'interprétation des résultats de régression multiple.

Sous les hypothèses énoncées, chaque Y_i prend ses valeurs en suivant une distribution normale.

$$Y_i \sim N(E(Y_i), \sigma^2)$$

avec $E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_k X_{ki}$

Figure 8.1

LE MODÈLE DE RÉGRESSION MULTIPLE

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_{l_3} X_{ki} + \varepsilon_i$ avec $k < n$
$\varepsilon_i \sim N(0, \sigma^2)$
ε_i non corrélé avec ε_j avec $i \neq j$
ε_i non corrélé avec X_{1i}, X_{2i} ... X_{ki}

Les estimateurs par minimisation de la somme des résidus au carré

Dans le cadre des hypothèses résumées dans la figure précédente, l'approche par minimisation de la somme des carrés des résidus (*ordinary least square*) permet d'aboutir analytiquement aux variables aléatoires :

$\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_k$ dont les moyennes sont respectivement les paramètres inconnus β_0, β_1, β_2, …, β_k.

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

$$E(\hat{\beta}_2) = \beta_2$$

…

$$E(\hat{\beta}_k) = \beta_k$$

À partir d'une base de données de n observations ($n > k$) pour Y_i, X_{1i}, X_{2i}, …, X_{ki}, les programmes de traitement de données par régression (SPSS, SAS, STATA, MINITAB, EXCEL) calculent des valeurs pour $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_k$ et des « erreurs-types » correspondantes $S_{\hat{\beta}_1}, S_{\hat{\beta}_2}, .., S_{\hat{\beta}_i}$. Ces informations permettront de construire des intervalles de confiance entre les bornes desquels se situent respectivement les paramètres β_0, β_1, β_2, …, β_k.

$$\Pr(\hat{\beta}_i - t_{0,025} \times S_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{0,025} \times S_{\hat{\beta}_i}) = 0,95$$

la table de Student (t) se lisant avec $n - k - 1$ degrés de liberté. Si les degrés de liberté dépassent 30, on recourt à la table Normale centrée réduite ($N_{0,025} = 1,96$).

Si l'une des bornes est négative et l'autre positive, le paramètre β_i se situe dans un intervalle où figure la valeur zéro, ce qui ne permet pas d'exclure l'hypothèse que la variable indépendante qui lui est associée n'exerce aucun effet sur la variable dépendante Y_i.

Ce type d'interprétation peut se lire plus directement. L'intervalle de confiance peut se réécrire comme suit :

$$\Pr(-t_{0,025} < \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} < t_{0,025}) = 0,95$$

Sous l'hypothèse $\beta_i = 0$,

$$\Pr(-t_{0,025} < \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} < t_{0,025}) = 0,95$$

En d'autres termes, si $\left|\frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}\right| < t_{0,025}$ avec $n - k - 1 < 30$

ou si $\left|\frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}\right| < 1,96$ avec $n - k - 1 \geq 30$,

on ne peut exclure, au risque de se tromper 5 fois sur 100, l'hypothèse que la variable indépendante associée au paramètre β_i n'exerce aucune influence sur la variable dépendante Y_i.

2. UN EXEMPLE POUR LES TESTS SUR LES COEFFICIENTS ESTIMÉS

Prenons un exemple à partir des données (**fichier : Salempl.sav**) dont un extrait figure dans le tableau suivant.

Tableau 8.2

UN EXEMPLE DE FICHIER DE DONNÉES POUR LE TRAITEMENT PAR RÉGRESSION MULTIPLE

Id	Educ	Travcat	Pevxp	Salaire	Sexe	Age	Ndiplom	Cat1	Cat2	Cat3	Niv1	Niv2	Niv3	Niv4	PRE_1	RES_1
1	15	3	144	57000	1	51	3	0	0	1	0	0	1	0	41065	15935
2	16	1	36	40200	1	45	3	1	0	0	0	0	1	0	43782	-3582
3	12	1	381	21450	0	74	2	1	0	0	0	1	0	0	31865	-10415
4	8	1	190	21900	0	56	1	1	0	0	1	0	0	0	13478	8422
5	15	1	138	45000	1	48	3	1	0	0	0	0	1	0	40993	4007
6	15	1	67	32100	1	45	3	1	0	0	0	0	1	0	40136	-8036
7	15	1	114	36000	1	47	3	1	0	0	0	0	1	0	40703	-4703
8	12	1	0	21900	0	37	2	1	0	0	0	1	0	0	27266	-5366
9	15	1	115	27900	0	57	3	1	0	0	0	0	1	0	40715	-12815
10	12	1	244	24000	0	57	2	1	0	0	0	1	0	0	30211	-6211
11	16	1	143	30300	0	53	3	1	0	0	0	0	1	0	45073	-14773
12	8	1	26	28350	1	37	1	1	0	0	1	0	0	0	11498	16852
13	15	1	34	27750	1	43	3	1	0	0	0	0	1	0	39737	-11987
14	15	1	137	35100	0	54	3	1	0	0	0	0	1	0	40981	-5881
15	12	1	66	27300	1	41	2	1	0	0	0	1	0	0	28063	-763
16	12	1	24	40800	1	39	2	1	0	0	0	1	0	0	27556	13244
17	15	1	48	46000	1	41	3	1	0	0	0	0	1	0	39906	6094
18	16	3	70	103750	1	47	3	0	0	1	0	0	1	0	44192	59558
19	12	1	103	42300	1	41	2	1	0	0	0	1	0	0	28509	13791
20	12	1	48	26250	0	63	2	1	0	0	0	1	0	0	27845	-1595
21	16	1	17	38850	0	40	3	1	0	0	0	0	1	0	43552	-4702
22	12	1	315	21750	1	63	2	1	0	0	0	1	0	0	31068	-9318
23	15	1	75	24000	0	38	3	1	0	0	0	0	1	0	40232	-16232
24	12	1	124	16950	0	70	2	1	0	0	0	1	0	0	28763	-11813
25	15	1	171	21150	0	61	3	1	0	0	0	0	1	0	41391	-20241
26	15	1	14	31050	1	37	3	1	0	0	0	0	1	0	39496	-8446
27	19	3	96	60375	1	49	4	0	0	1	0	0	0	1	56567	3808
28	15	1	43	32550	1	40	3	1	0	0	0	0	1	0	39846	-7296
29	19	3	199	135000	1	59	4	0	0	1	0	0	0	1	57810	77190
30	15	1	54	31200	1	42	3	1	0	0	0	0	1	0	39979	-8779
31	12	1	83	36150	1	39	2	1	0	0	0	1	0	0	28268	7882
32	19	3	120	110625	1	49	4	0	0	1	0	0	0	1	56857	53768
33	15	1	68	42000	1	42	3	1	0	0	0	0	1	0	40148	1852
34	19	3	175	92000	1	54	4	0	0	1	0	0	0	1	57521	34479
35	17	3	18	81250	1	42	3	0	0	1	0	0	1	0	47585	33665
36	8	1	52	31350	0	40	1	1	0	0	1	0	0	0	11812	19538
37	12	1	113	29100	1	49	2	1	0	0	0	1	0	0	28630	470
38	15	1	49	31350	1	41	3	1	0	0	0	0	1	0	39918	-8568
39	16	1	46	36000	1	43	3	1	0	0	0	0	1	0	43902	-7902
40	15	1	23	19200	0	70	3	1	0	0	0	0	1	0	39604	-20404
41	12	1	52	23550	0	42	2	1	0	0	0	1	0	0	27894	-4344
42	15	1	90	35100	1	43	3	1	0	0	0	0	1	0	40413	-5313
43	12	1	46	23250	1	39	2	1	0	0	0	1	0	0	27821	-4571
44	8	1	50	29250	1	40	1	1	0	0	1	0	0	0	11788	17462
45	12	2	307	30750	1	65	2	0	1	0	0	1	0	0	30972	-222
46	15	1	165	22350	0	63	3	1	0	0	0	0	1	0	41319	-18969
47	12	1	228	30000	0	65	2	1	0	0	0	1	0	0	30018	-18
48	12	2	240	30750	1	56	2	0	1	0	0	1	0	0	30163	587
49	15	1	93	34800	1	45	3	1	0	0	0	0	1	0	40449	-5649

Le fichier **Salempl.sav** rassemble les fiches individuelles de 474 personnes concernant le salaire et un certain nombre de variables qui pourraient « rendre compte » des différences de salaire constatées.

On se propose, en un premier temps, d'étudier la relation suivante :

$$Salaire_i = \beta_0 + \beta_1 \times Educ_i + Pevxp_i + \varepsilon_i$$

où *Salaire* désigne le salaire annuel brut avant impôts,

Educ, le nombre d'années de scolarité.

Pevxp, la durée (en mois) de l'expérience antérieure de travail.

Le traitement des données à partir de SPSS a donné les résultats suivants :

Encadré 8.1

RÉGRESSION MULTIPLE
L'ÉVALUATION STATISTIQUE DES COEFFICIENTS ESTIMÉS

Variable dépendante: Salaire			
Variables indépendantes	$\hat{\beta}_i$	$S_{\hat{\beta}_i}$	$\left\|\frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}\right\|$
Terme constant	-20978,30	3087,26	6,80
Educ	4020,34	210,65	19,09
Pevxp	12,07	5,81	2,08

$n - k - 1 = 474 - 2 - 1 = 471$

On se référera en conséquence à la table normale centrée réduite :

$$\Pr(\hat{\beta}_1 - t_{0,025} \times S_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{0,025} \times S_{\hat{\beta}_1}) = 0,95$$

$$\hat{\beta}_1 - 1,96 \times S_{\hat{\beta}_1} = 4020,34 - 1,96 \times 210,65$$

$$\hat{\beta}_1 + 1,96 \times S_{\hat{\beta}_1} = 4020,34 + 1,96 \times 210,65$$

Les deux bornes sont positives. On exclut l'hypothèse que $\beta_1 = 0$.

Sur la base des résultats obtenus, on peut aussi exclure l'hypothèse que $\beta_2 = 0$.

Les mêmes conclusions auraient pu être obtenues

en considérant : $\left|\frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}\right|$

19,09 et 2,08 sont tous deux supérieurs à 1,96. Au risque de se tromper 5 fois sur 100, on peut donc écarter les hypothèses que $\beta_1 = 0$ et $\beta_2 = 0$.

Si les hypothèses à la base de la régression sont respectées, on serait dès lors porté à accepter l'incidence des années de formation scolaire ou académique (*Educ*) et de l'expérience professionnelle antérieure (*Pevxp*) sur les salaires pour des contextes semblables à celui qui a donné lieu aux observations recueillies. Toutefois, au stade actuel, la conclusion est trop hâtive, car il y aurait lieu d'examiner comment se distribuent les résidus r_i et d'envisager l'incidence éventuelle d'autres variables indépendantes.

Les résidus r_i (à ne pas confondre avec les termes d'erreur) s'obtiennent en soustrayant de Y_i les $\hat{Y}_i$ calculés à partir de la régression.

$$r_i = Y_i - \hat{Y}_i$$

où $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \times X_{1i} + \hat{\beta}_2 \times X_{2i} + ... \hat{\beta}_k \times X_{ki}$

On peut obtenir ces valeurs automatiquement à partir des logiciels de type SPSS. Dans les deux dernières colonnes du tableau de données, on retrouvera les $\hat{Y}_i$ sous la rubrique *PREV_1* et les résidus r_i sous la rubrique *RES_1*.

Par exemple, pour la personne identifiée par $Id = 1$,

$$\hat{Y}_i = -20978{,}3036 + 4020{,}34334 \times 15 + 12{,}07129 \times 144 = 41065$$

et

$$r_i = 57000 - 41065 = 15935.$$

En fait, les logiciels livrent un ensemble de résultats qui permettent une interprétation nuancée :

- des estimations des paramètres du modèle théorique considéré,
- des valeurs prévues qu'on pourrait construire pour la variable dépendante compte tenu des estimations obtenues,
- des résidus,

et un ensemble de résultats connexes permettant d'évaluer dans quelle mesure la régression a été effectuée dans le respect des conditions requises à l'inférence statistique.

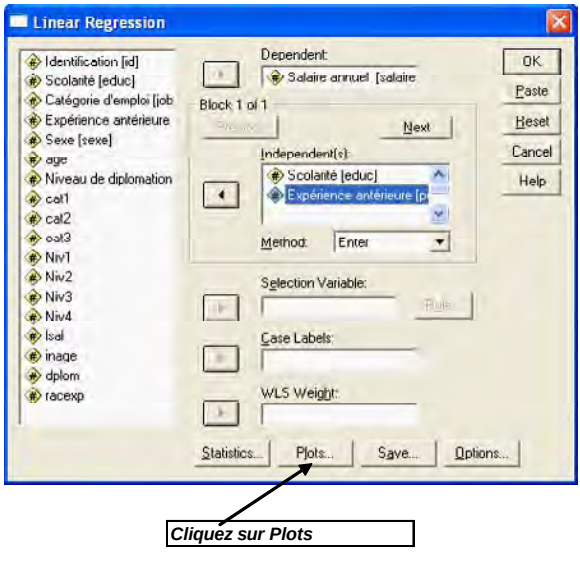
3. LA RÉGRESSION MULTIPLE AVEC SPSS

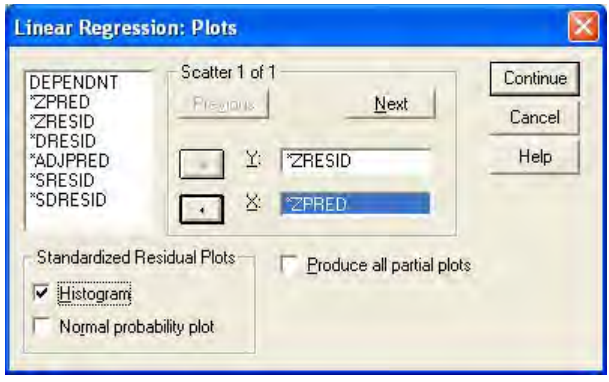
Un exemple avec deux variables explicatives

En supposant le fichier SPSS **Salempl.sav** ouvert, on demande la régression selon la procédure suivante. À partir de l'onglet ***Analyze***, on sollicite ***Regression***, puis ***Linear***. Dans la fenêtre qui s'ouvre, on complète les cases respectives tel qu'indiqué à la figure 8.2. On introduit les variables dans les cases appropriées en les sélectionnant, puis en cliquant sur la flèche ➤.

Figure 8.2

LES COMMANDES SPSS POUR LA RÉGRESSION MULTIPLE





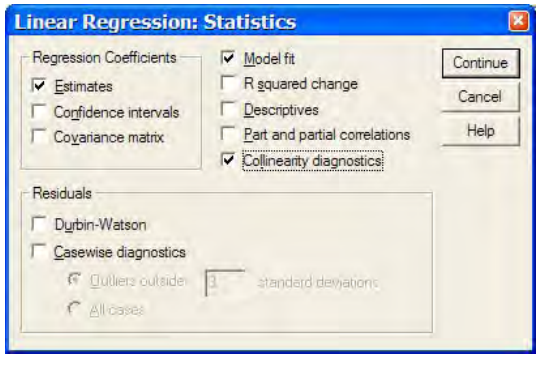
En remplissant les cases tel qu'indiqué, on demande le graphique entre les résidus normalisés (*ZRESID*) et *ZPRED* : les valeurs prédites normalisées de la variable dépendante (le salaire prévu normalisé).

D'autre part, on demande l'histogramme des résidus normalisés.

À gauche de **PLOTS**, cliquez sur **STATISTICS** et, dans la fenêtre qui s'ouvre, demandez les tests de colinéarité.

Figure 8.3

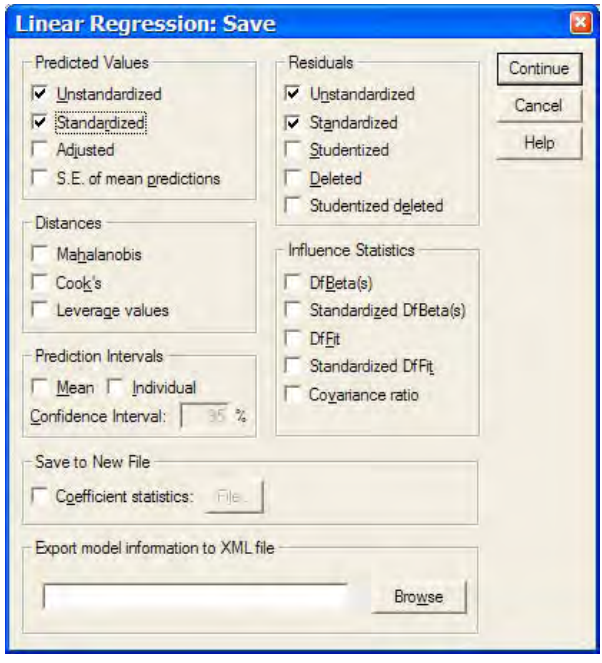
LA COMMANDE SPSS POUR LES TESTS DE COLINÉARITÉ



À droite de **PLOTS**, sollicitez **SAVE** et demandez que soient ajoutés dans votre base de données les prévisions (*PRED_1*) et les prévisions normalisées (*ZPRED_1*) relatives à la variable dépendante, les résidus (*RES_1*), et les résidus normalisés (*ZRES_1*).

Figure 8.4

LA COMMANDE SPSS SAVE EN RÉGRESSION MULTIPLE



Les principaux résultats

Les principaux résultats du traitement des données figurent dans les encadrés suivants.

Encadré 8.2

COEFFICIENT DE DÉTERMINATION : R^2 ET $R^2_{AJUSTÉ}$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,664[a]	,441	,439	12788,694

a. Predictors: (Constant), Expérience antérieure, Scolarité

b. Dependent Variable: Salaire annuel

Le coefficient de détermination R^2 compare les valeurs estimées (prédites) de la variable dépendante à ses valeurs observées, à l'aide de la somme des écarts à la moyenne. R^2 prend ses valeurs entre 1 et 0. La somme des résidus au carré fait baisser R^2, indiquant par là même l'importance relative de ce dont « ne rendent pas compte » les variables explicatives.

$R^2_{ajusté,}$ plus faible que R^2, tient compte de la perte d'information liée aux degrés de liberté.

Dans le cas de la régression qui sert d'exemple, $R^2 = 0{,}44$: les variables indépendantes retenues ont conjointement un pouvoir « explicatif » relativement limité.

Figure 8.5

RÉGRESSION MULTIPLE : R^2 ET $R^2_{AJUSTÉ}$

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum r_i^2}{\sum y_i^2}$$

$$R^2_{ajusté} = 1 - \frac{\sum r_i^2/(n-k-1)}{\sum y_i^2/(n-1)}$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,664[a]	,441	,439	12788,694

a. Predictors: (Constant), Expérience antérieure, Scolarité

b. Dependent Variable: Salaire annuel

Moyennant une transformation adéquate, R^2 permet de tester l'hypothèse : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, c'est-à-dire l'hypothèse qu'aucune variable explicative n'exerce une incidence sur la variable dépendante.

La transformation requise est la suivante :

$$F_{calculé} : \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

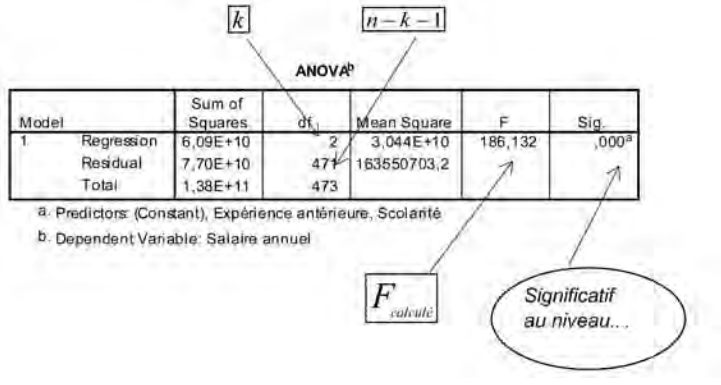
$F_{calculé}$ est confronté à la table F selon k degrés de liberté au numérateur et $n - k - 1$ degrés de liberté au dénominateur.

Si, au seuil de signification retenu (α), $F_{calculé} > F_{\alpha;k,n-k-1}$, on rejette l'hypothèse que $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ pour lui préférer l'hypothèse qu'au moins un $\beta_i = 0$, c'est-à-dire qu'au moins une variable explicative a une incidence sur la variable dépendante. C'est dire que le test F effectué à partir de R^2 envisage une hypothèse très minimale !

$F_{calculé}$ et le résultat de sa confrontation avec la table F figurent dans le tableau d'analyse de variance (ANOVA).

Encadré 8.3

RÉGRESSION MULTIPLE : ANALYSE DE VARIANCE



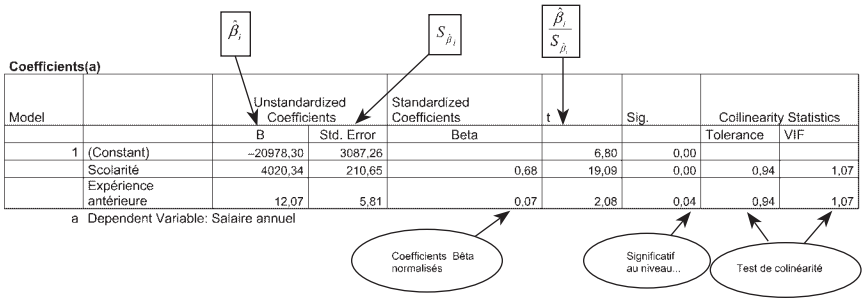
Dans le cas présenté en exemple, on peut accepter, sans grand risque (Sig. = 0,000) de se tromper, qu'au moins une variable explicative a une incidence sur la variable *Salaire*.

L'encadré suivant apporte plus de précisions.

À l'aide des $\hat{\beta}_i$ calculés et des erreurs-types correspondantes $S_{\hat{\beta}_i}$, il est possible de tester séparément l'hypothèse de la non-incidence de chaque variable explicative sur la variable dépendante.

Encadré 8.4

RÉGRESSION MULTIPLE : LES ESTIMATIONS $\hat{\beta}_i$ DES PARAMÈTRES β_I



Chaque coefficient estimé (4020,34 ; 12,07) répond aux attentes positives *a priori* et est statistiquement significatif à un niveau (0,00 ; 0,038) plus exigeant que le seuil traditionnel de 0,05 utilisé en sciences sociales, laissant entendre que la scolarité (*Educ*) et l'expérience professionnelle antérieure (*Pevexp)* ont chacune une incidence spécifique sur la variable *Salaire*.

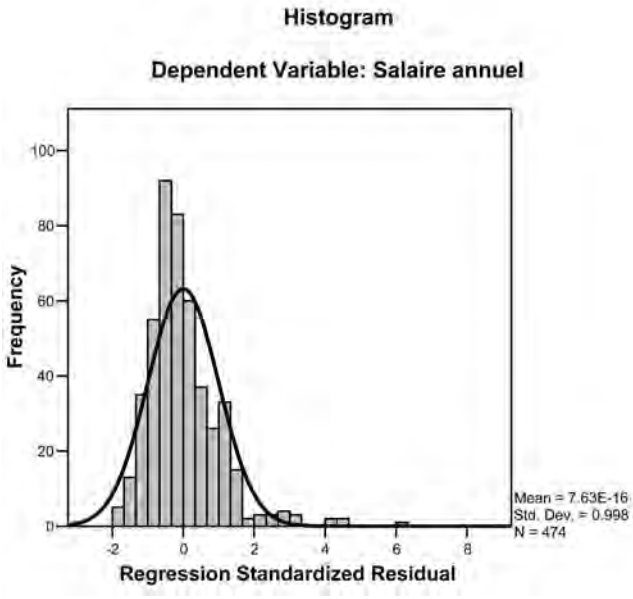
Diagnostic de « validité »

Les conclusions qui viennent d'être évoquées laissent supposer qu'ont été respectées les hypothèses à la base de la régression « classique ». Le diagnostic de « validité » de l'interprétation statistique donnée aux estimations obtenues par régression se fait à l'aide d'un examen détaillé des résidus.

L'histogramme des résidus permet de voir si la distribution des résidus se rapproche d'une distribution normale. C'est relativement le cas pour les résidus obtenus lors du traitement par régression de l'exemple traité. De toute manière, les estimations restent « robustes » (« résistent assez bien ») à la relative non normalité des résidus.

Figure 8.6

RÉGRESSION MULTIPLE : HISTOGRAMME DES RÉSIDUS

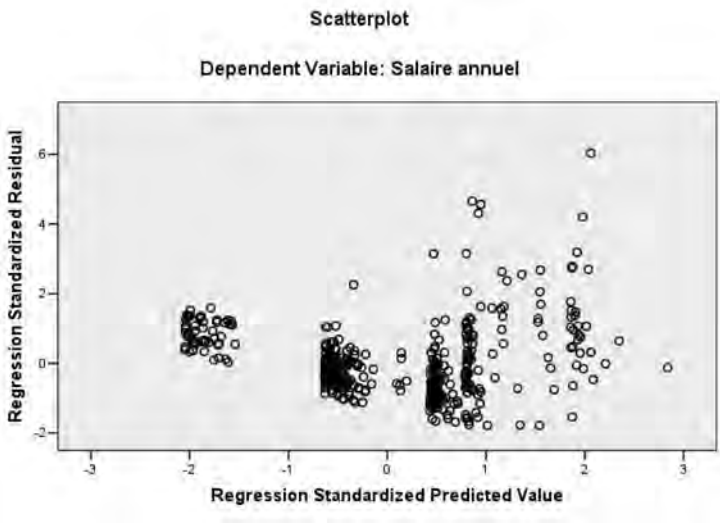


Variance constante des termes d'erreur et transformation de variable

Le graphique entre les ***résidus normalisés*** et les ***valeurs prévues normalisées*** de la variable dépendante (le salaire prévu normalisé) permet de déceler certaines transgressions importantes des hypothèses relatives aux termes d'erreur. Si les résidus normalisés se distribuent de manière relativement uniforme autour de zéro et sans une trop grande dispersion au-delà des repères 2 et –2, on considérera généralement que l'hypothèse de la variance constante des termes d'erreurs est respectée. Cela ne semble pas le cas à l'examen du graphique suivant : on y discerne une dispersion en cône.

Figure 8.9

LE GRAPHIQUE ENTRE LES RÉSIDUS NORMALISÉS ET LES VALEURS PRÉDITES NORMALISÉES DE LA VARIABLE DÉPENDANTE



Cette situation peut trouver son origine dans le fait que la variable dépendante n'a pas été transformée adéquatement.

Les transformations les plus fréquentes (nous les désignons par la lettre *Z*) sont :

$$Z_{1i} = 1 / Y_i$$
$$Z_{2i} = \sqrt{Y_i}$$
$$Z_{3i} = 1 / \sqrt{Y_i}$$

$Z_{4i} = ln(Y_i)$ ou en base 10 $Z_{5i} = log(Y_i)$

Si on utilise *Lsal*, c'est-à-dire le logarithme naturel de *Salaire* comme variable dépendante, la forme de cône tend à s'atténuer (sans disparaître !) dans le graphique qui associe les résidus normalisés et la variable dépendante prédite normalisée.

Encadré 8.5

RÉGRESSION MULTIPLE : LES ESTIMATIONS AVEC LA VARIABLE LSAL

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,699[a]	,489	,486	,28477

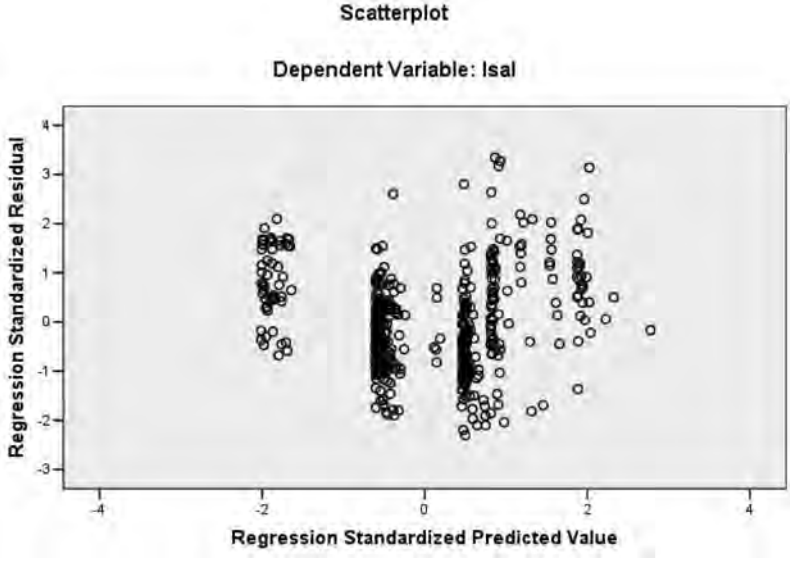
a. Predictors: (Constant), Expérience antérieure, Scolarité

b. Dependent Variable: lsal

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	9,014	,069		131,129	,000
	Scolarité	,098	,005	,711	20,884	,000
	Expérience antérieure	,000	,000	,057	1,683	,093

a. Dependent Variable: lsal



Par contre, le coefficient estimé associé à *Pevexp* (l'expérience antérieure) devient non significatif (Sig. = 0,093, supérieur à 0,05). Une transformation de la variable *Pevexp* a été envisagée dans une tentative de « remédier » à la situation. L'encadré suivant présente les résultats des estimations lorsque *Lsal* est « régressé » sur *Educ* (les années de scolarité) et *Racexp* (la racine de *Pevexp*). Les coefficients estimés des deux variables explicatives sont cette fois tous deux statistiquement significatifs.

Encadré 8.6

RÉGRESSION MULTIPLE : LES RÉSULTATS AVEC LSAL RÉGRESSÉ SUR EDUC ET RACEXP

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,700[a]	,490	,488	,28428

a. Predictors: (Constant), racexp, Scolarité

b. Dependent Variable: lsal

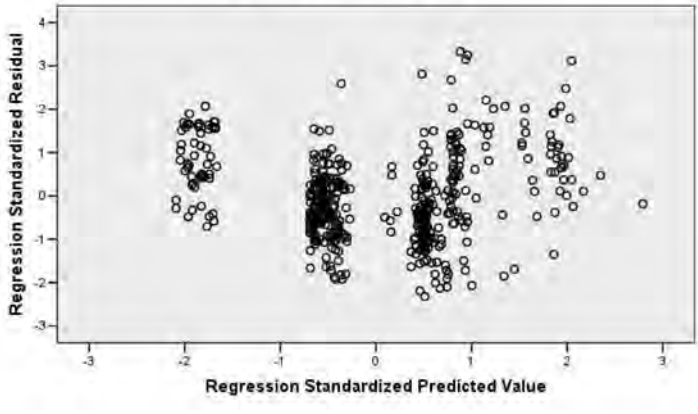
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8,993	,071		127,499	,000
	Scolarité	,098	,005	,709	21,212	,000
	racexp	,005	,003	,071	2,109	,036

a. Dependent Variable: lsal

Scatterplot

Dependent Variable: lsal



Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

4. RÉGRESSION MULTIPLE

Un exemple avec trois variables explicatives

Bien que relativement satisfaisants, les résultats obtenus peuvent être remis en cause parce qu'ils n'ont pas pris en considération certaines variables explicatives. L'encadré suivant présente les résultats d'une régression où la variable *Lsal* est « expliquée » par *Educ*, *Racexp* et *Inage (Inage* = 1/*Age)*. Tout comme dans l'exemple précédent, l'introduction de la variable concernant l'âge a fait l'objet d'une transformation dans la perspective de dégager des « coefficients estimés statistiquement significatifs ».

Encadré 8.7

RÉGRESSION MULTIPLE AVEC TROIS VARIABLES INDÉPENDANTES : LA RÉGRESSION DE LSAL SUR EDUC, RACEXP, INAGE

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,705[a]	,497	,494	,28260

a. Predictors: (Constant), inage, Scolarité, racexp

b. Dependent Variable: lsal

Coefficients[a]

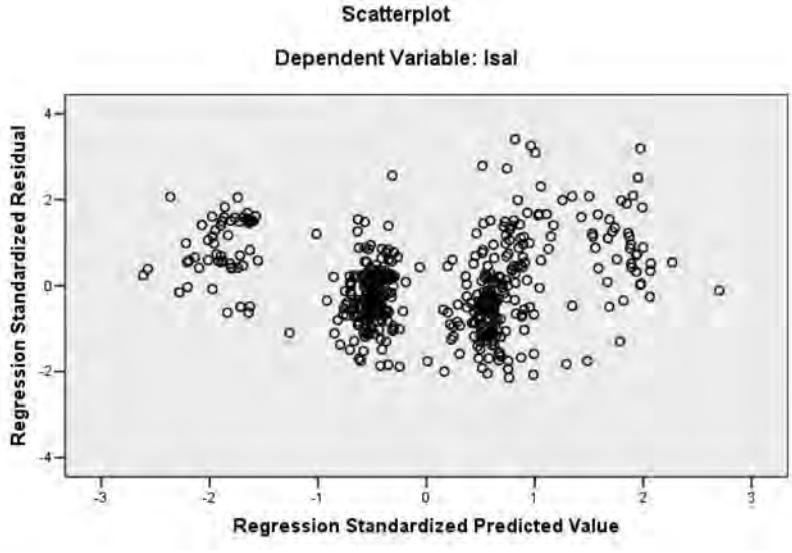
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	8,677	,142		61,233	,000		
	Scolarité	,097	,005	,704	21,134	,000	,964	1,038
	racexp	,014	,004	,185	3,329	,001	,348	2,876
	inage	,112	,044	,143	2,571	,010	,347	2,878

a. Dependent Variable: lsal

L'ajout de la variable *Inage* aux deux variables indépendantes *Educ* et *Racexp* améliore les résultats : légèrement en termes de R^2, mais très nettement en termes de précision des estimations (voir les colonnes *t* et Sig). Ces résultats sont obtenus dans le cadre d'une distribution relativement « satisfaisante » des résidus, même s'il subsiste encore une certaine régularité en forme de cône.

Figure 8.10

***RÉGRESSION MULTIPLE* :**
LES RÉSIDUS DANS LE CAS DE LA RÉGRESSION DE LSAL SUR EDUC, RACEXP, INAGE



Il est possible, comme on le verra ultérieurement, d'atténuer cette situation par l'introduction de variables explicatives supplémentaires.

Les résultats des estimations ne semblent pas, non plus, « trop » affectés par des problèmes de ***colinéarité.***

Les deux dernières colonnes de la deuxième partie de l'encadré 8.7 (où figurent les coefficients estimés) présentent deux « statistiques » permettant d'évaluer la proximité linéaire de deux ou plusieurs variables explicatives.

Les deux mesures (*Tolerance* et *Vif*) sont bâties à partir de régressions artificielles (sans prétention d'interprétation) où une variable explicative du modèle étudié est « régressée » sur les autres variables explicatives du modèle.

Ce type de régression permet de saisir la relation « multiple » qui unit une variable explicative aux autres variables explicatives. C'est ce type de relation qui peut menacer la précision de certains résultats estimés.

Prenons un exemple. En régressant *Inage* sur *Educ* et *Racexp*, on obtient les résultats suivants :

Encadré 8.8

RÉGRESSION INSTRUMENTALE ET COLINÉARITÉ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,808[a]	,653	,651	,29873

a. Predictors: (Constant), racexp, Scolarité

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2,825	,074		38,111	,000
	Scolarité	,007	,005	,038	1,370	,171
	racexp	-,079	,003	-,800	-28,976	,000

a. Dependent Variable: inage

Le niveau de tolérance se calcule selon l'expression :

Tolerance = $1 - R^2$; $1 - 0{,}653 = 0{,}347$

Il est généralement admis qu'un niveau de *Tolérance* inférieur à 0,20 risque de provoquer de fortes imprécisions dans les estimations des coefficients associés aux variables affectées de colinéarité. Il s'agit d'une règle de travail pratique. En fait, si les estimations dans la régression de base (voir l'encadré 8.7) restent statistiquement assez précises, on accepte les résultats en dépit de la présence de colinéarité.

Vif utilise R^2 selon l'expression suivante : $Vif = 1/(1 - R^2)$.

Vif (dont la valeur positive s'éloigne de 1 au fur et à mesure que R^2 augmente) conduit aux mêmes conclusions que *Tolerance,* étant donné que tous deux exploitent la même information.

En pratique, lorsque *Tolerance* ou *Vif* laissent soupçonner des problèmes de colinéarité, il vaut mieux demander explicitement les régressions (artificielles) d'une variable explicative sur les autres variables explicatives. En effet, en plus de R^2, on pourra aussi s'aider du *t* de Student pour mieux discerner les variables impliquées dans les problèmes de colinéarité. Si on se reporte au tableau précédent, on voit que *Racexp*

a un lien de colinéarité ($t = |-28{,}976|$) avec *Inage*, ce qui ne semble pas le cas pour *Educ* ($t = |1{,}37|$). La colinéarité entre *Inage* et *Racexp* ne semble pas trop affecter les résultats des estimations lorsqu'on régresse *Lsal* sur *Educ*, *Racexp* et *Inage*. La colinéarité « n'empêche donc pas » d'accepter les résultats.

Que faire si la colinéarité menace la précision des estimations ?

Le choix d'une solution adéquate n'est pas aisé.

- On peut essayer de rompre la colinéarité par la transformation d'une ou plusieurs variables explicatives.
- On peut essayer de recourir à des variables qui constituent d'assez bonnes approximations de variables impliquant des problèmes de colinéarité.
- On peut aussi, ce que l'on fait très souvent, omettre une variable explicative sur la base du fait que la variable ou les variables qui lui sont colinéaires capteront de toute manière une bonne part de la variable omise.
- On peut tenter de réduire les problèmes de colinéarité par l'adjonction d'observations, ce qui n'est évidemment pas souvent possible.

Quelle qu'elle soit, la solution adoptée impose de nuancer l'interprétation des résultats obtenus parce qu'elle affecte le plus souvent l'estimation de l'impact spécifique d'une ou plusieurs variables explicatives.

5. UN EXEMPLE AVEC VARIABLES INDÉPENDANTES QUANTITATIVES ET QUALITATIVES

Les variables indépendantes qualitatives

Dans la base données **Salempl.sav**, la variable *Travcat* distingue trois catégories d'employés : des « cadres », des « employés » et des « employés subalternes ». Ces catégories sont exhaustives et exclusives. Il s'agit d'une variable qualitative où les chiffres n'expriment pas une distance séparant une catégorie, par exemple 1, de la catégorie qui la suit, la catégorie 2. Les chiffres correspondent à des étiquettes d'identification. Dans ce contexte, un paramètre de type β_i associé à *Travcat* n'aurait pas de sens. On crée

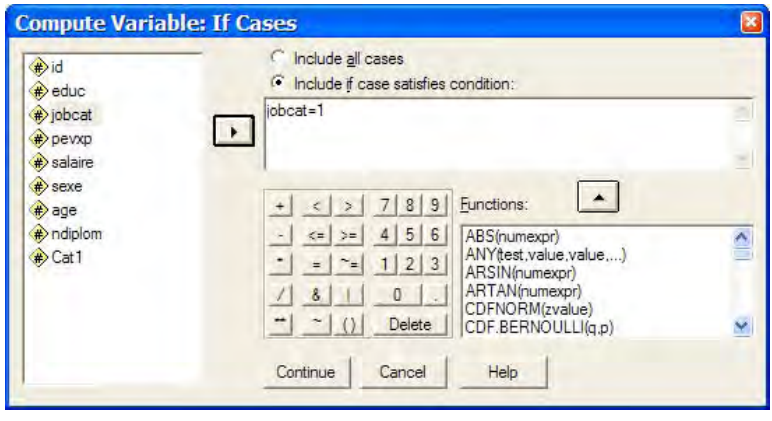
alors trois variables *Cat1*, *Cat2* et *Cat3*. *Cat1* prend la valeur 1 chaque fois que *Travcat* vaut 1 et la valeur zéro dans les autres cas. *Cat2* prend la valeur 1 chaque fois que *Travcat* prend la valeur 2 et la valeur zéro dans les autres cas. *Cat3* prend la valeur 1 si *Travcat* = 3 et la valeur zéro si *Travcat* ≠ 3. On voit au tableau 8.2 comment se présentent dans la base de données les trois variables *Cat1, Cat2* et *Cat3*.

De toute évidence, les trois catégories sont strictement complémentaires. Dès qu'on connaît pour une personne le contenu de deux catégories, on connaît automatiquement le contenu de la troisième.

Il s'agit d'un cas de « colinéarité parfaite » rendant impossible l'estimation des β_i affectant chacune des variables *Cat1, Cat2, Cat3*. ***Une des trois variables devra obligatoirement être omise pour casser la stricte complémentarité qui les lie.***

En SPSS, on peut recourir à la procédure ***Recode*** pour créer les données de *Cat1*, *Cat2*, *Cat3* ou encore créer les variables *Cat1, Cat2, Cat3* de la manière suivante.

Après avoir sollicité l'onglet ***Transform***, puis ***Compute*** dans le menu déroulant, on crée une variable *Cat1* égale à zéro (hormis les *missing* que comporterait *Jobcat* !), puis on sollicite une seconde fois ***Compute*** et on impose à la variable *Cat1* la valeur 1 à l'aide de la condition (bouton : **if**…) tel qu'indiqué dans la fenêtre ci-dessous :



On procède de même pour *Cat2* et *Cat3*.

Le fichier **Salempl.sav** contient aussi la variable *Sexe* identifiant les hommes par le chiffre 1 et les femmes par le chiffre 0.

À l'aide de la base de données, on estime par régression multiple les paramètres du modèle :

$$lsal_i = \beta_0 + \beta_1 Educ_i + \beta_2 Racxp_{ii} + \beta_3 Inage_i + \beta_4 Cat3_i + \beta_5 Cat2_i + \beta_6 Sexe_i + \varepsilon_i$$

Encadré 8.9

RÉGRESSION MULTIPLE : RÉSULTATS AVEC VARIABLES INDÉPENDANTES QUANTITATIVES ET VARIABLES INDÉPENDANTES QUALITATIVES

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,873[a]	,763	,760	,19469

a. Predictors: (Constant), Sexe, inage, cat3, cat2, Scolarité, racexp

b. Dependent Variable: lsal

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	9,447	,107		88,499	,000		
	Scolarité	,042	,004	,305	9,710	,000	,514	1,944
	racexp	-,001	,003	-,015	-,341	,734	,278	3,604
	inage	,065	,031	,082	2,063	,040	,318	3,141
	cat2	,200	,045	,117	4,423	,000	,726	1,378
	cat3	,566	,030	,544	18,950	,000	,615	1,626
	Sexe	,174	,022	,218	7,883	,000	,664	1,506

a. Dependent Variable: lsal

Le coefficient concernant l'expérience antérieure (*Racexp*) est non significatif (Sig. : 0,734). La tolérance est relativement basse : 0,278 (pour *Inage* aussi d'ailleurs : 0,318).

La régression (artificielle) de *Racxp* sur *Educ*, *Inage, Cat2*, *Cat3* et *Sexe* donne les résultats suivants :

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,825[a]	,681	,677	59,406

a. Predictors: (Constant), Sexe, inage, cat3, cat2, Scolarité

b. Dependent Variable: Expérience antérieure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	446,496	19,629		22,747	,000
	Scolarité	-3,741	1,315	-,103	-2,844	,005
	inage	-141,607	5,701	-,685	-24,837	,000
	cat2	97,533	13,529	,216	7,209	,000
	cat3	-5,010	9,097	-,018	-,551	,582
	Sexe	28,063	6,332	,134	4,432	,000

a. Dependent Variable: Expérience antérieure

Visiblement, la variable *Racexp* est linéairement associée à la scolarité *Educ*, à *Inage*, à *Cat2*, à *Sexe* de manière statistiquement significative (Sig. très nettement inférieur à 0,05). C'est un cas typique de **colinéarité** présentant un risque de perturber les estimations.

Le modèle finalement retenu exclut la variable *Racexp*, non parce que l'expérience professionnelle antérieure n'a pas d'incidence sur la variable *Lsal*, mais en raison de la difficulté (colinéarité) à extraire une mesure statistiquement satisfaisante de l'incidence spécifique de cette variable sur la variable dépendante.

$$lsal_i = \beta_0 + \beta_1 Educ_i + \beta_3 Invage_i + \beta_4 Cat3_i + \beta_5 Cat2_i + \beta_6 Sexe_i + \varepsilon_i$$

Les résultats figurant dans l'encadré suivant mettent en évidence des coefficients statistiquement très satisfaisants (Voir les *t* et les Sigs.), conformes aux attentes *a priori*, peu affectés par des problèmes de colinéarité et donnant lieu à des résidus relativement conformes aux hypothèses concernant les termes d'erreur.

Encadré 8.10.

RÉGRESSION MULTIPLE
RÉGRESSION DE LSAL SUR EDUC, INAGE, CAT2, CAT3, SEXE

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,873[a]	,763	,760	,19450

a. Predictors: (Constant), Sexe, inage, cat3, cat2, Scolarité

b. Dependent Variable: lsal

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56,970	5	11,394	301,182	,000[a]
	Residual	17,705	468	,038		
	Total	74,675	473			

a. Predictors: (Constant), Sexe, inage, cat3, cat2, Scolarité

b. Dependent Variable: lsal

Coefficients[a]

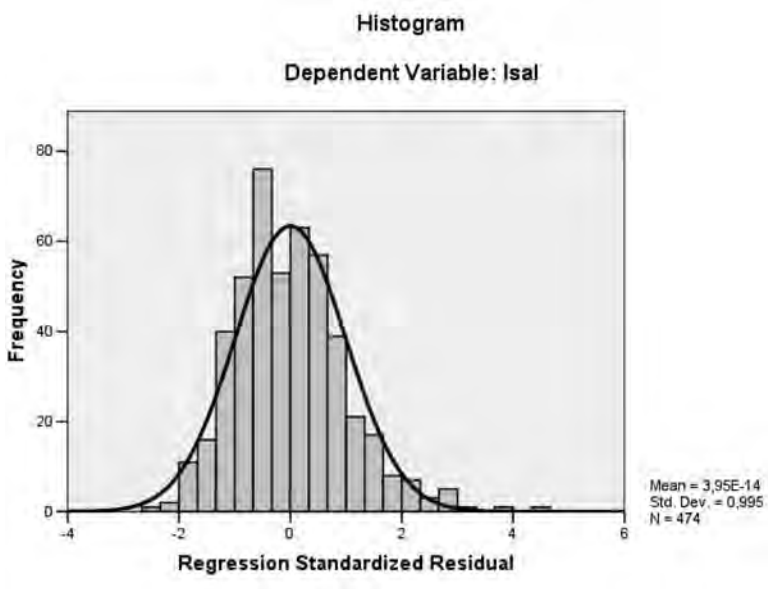
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9,418	,064		146,540	,000
	Scolarité	,042	,004	,306	9,785	,000
	inage	,073	,019	,093	3,927	,000
	cat2	,197	,044	,115	4,451	,000
	cat3	,566	,030	,545	19,015	,000
	Sexe	,171	,021	,215	8,257	,000

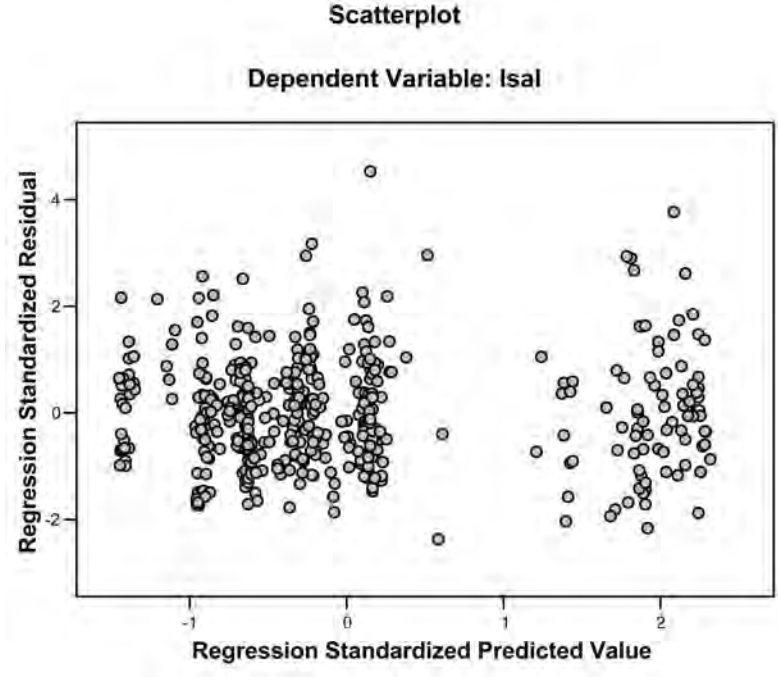
a. Dependent Variable: lsal

Casewise Diagnostics[a]

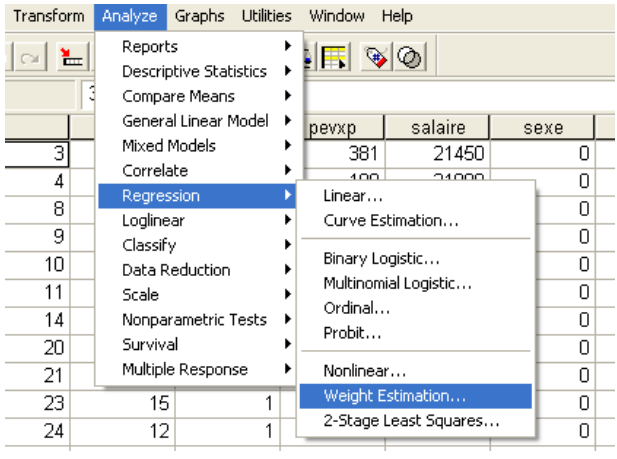
Case Number	Std. Residual	lsal	Predicted Value	Residual
29	3,768	11,81	11,0802	,73279
72	3,171	10,90	10,2800	,61677
218	4,528	11,29	10,4090	,88076

a. Dependent Variable: lsal





On remarquera que cette fois, la forme en cône des résidus a disparu. Il arrive cependant que les résidus continuent à se disperser **en forme de cône** même si on a introduit dans le modèle les variables explicatives jugées opportunes ou effectué les transformations adéquates de la variable dépendante. On se trouve alors devant un cas dit d'« hétérocédasticité rémanente », ce qui laisse supposer que la variance des termes d'erreur « persiste » à rester non constante. Cette situation exige un traitement spécifique (régression « pondérée »). La « pondération » se fait à l'aide d'une variable censée atténuer la forme de cône des résidus. On y recourt en faisant appel aux instructions SPSS suivantes :



Weight Estimation permet d'ouvrir une fenêtre complémentaire, que nous ne détaillerons pas dans le cadre de cette présentation.

Points « aberrants » ?

Il resterait toutefois à s'interroger sur la présence de quelques (3) résidus standardisés dont l'ordre de grandeur est particulièrement élevé (3,768 ; 3,171 ; 4,528 ; voir l'encadré 8.10, « Casewise Diagnostics »). Il se peut que la prise en considération de variables explicatives supplémentaires fasse disparaître ces points « aberrants ». Ce n'est pas toujours le cas. Il s'agit alors de rechercher la cause de cette situation particulière et d'y remédier. La cause peut dans certains cas être captée par une variable muette identifiant par 1 la présence de cette cause et par 0 son absence. C'est, par exemple, un événement particulier qui perturbe une série temporelle (le 11 septembre 2001). Très souvent, il s'agit de neutraliser ou de corriger une imprécision ou une erreur dans l'information recueillie.

Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

La suppression de l'observation donnant lieu à un point aberrant peut être envisagée comme solution limite. Il s'agit cependant d'une mesure à utiliser avec prudence. Fréquemment, l'élimination d'une observation donnant lieu à un résidu « aberrant » provoque techniquement l'apparition d'autres points « aberrants ».

6. LA CONTRIBUTION MARGINALE D'UNE VARIABLE EXPLICATIVE

Coefficients bêta et corrélation partielle

Dans le modèle de référence à la base de l'analyse de régression multiple, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_i$, chaque paramètre associé à une variable indépendante, par exemple β_2 associé à X_{2i}, traduit l'effet spécifique ou partiel de la variable indépendante sur la variable dépendante Y_i compte tenu des autres variables indépendantes insérées dans le modèle.

Les paramètres β_i sont tributaires de la manière de définir chaque variable indépendante. C'est aussi le cas des estimations $\hat{\beta}_i$. On ne peut dès lors utiliser directement les estimations $\hat{\beta}_i$ pour comparer les incidences spécifiques de chaque variable indépendante sur la variable dépendante.

6.1. LE RECOURS À DES VARIABLES CENTRÉES RÉDUITES ET LES COEFFICIENTS NORMALISÉS BÊTA

Une manière de tenter de contourner cet inconvénient est de travailler avec des variables centrées réduites (en écart à la moyenne et divisées par leur écart-type).

Dans cette perspective, le modèle de référence devient :

$$y_i = \gamma_1 x_{1i} + \gamma_2 x_{2i} + ... + \gamma_k x_{ki} + \varepsilon / \sigma_i$$

où $y_i = \dfrac{Y_i - \bar{Y}}{S_{Y_i}}$, $x_i = \dfrac{X_i - \bar{X}}{S_{X_i}}$ et $\gamma_i = \beta_i \dfrac{S_{X_{i_i}}}{S_{Y_i}}$.

Les coefficients normalisés estimés $\hat{y}_i$ présentent toutefois un défaut majeur. Ils sont instables alors même que les $\hat{\beta}_i$ ne changent pas. L'exemple suivant en témoigne.

Encadré 8.12

RÉGRESSION MULTIPLE AVEC VARIABLES CENTRÉES RÉDUITES

Y_i	X_{1i}	X_{2i}	$Pred\_1_i$
46	5	15	46,63306
52	7	17	51,2837
43	9	12	42,98613
64	13	22	63,38589
65	14	23	65,71122
67,61	18	23	67,61352

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	16,509	1,893		8,721	,013
	X1	,476	,195	,181	2,437	,135
	X2	1,850	,161	,851	11,476	,008

a. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	16,509	1,528		10,802	,002
	X1	,476	,121	,217	3,915	,030
	X2	1,850	,126	,814	14,656	,001

a. Dependent Variable: Y

Deux régressions ont été effectuées, l'une avec 5 observations, l'autre avec 6 observations. La valeur de Y_6 : 67,61 est en fait une prévision obtenue à partir des estimations sur 5 observations. Les estimations ($\hat{\beta}_i$) effectuées sur 6 observations sont identiques à celles qu'on obtient à partir de 5 observations. On observera par contre que les coefficients normalisés dits « *bêta* » varient d'un tableau à l'autre. Cette instabilité est liée au fait que les coefficients normalisés bêta sont tributaires de l'échantillon utilisé. Les comparaisons sur la base des coefficients normalisés bêta ne se prêtent donc pas à des généralisations.

6.2. LA CORRÉLATION PARTIELLE

La corrélation partielle est l'instrument le plus couramment utilisé pour comparer l'intensité des associations de chaque variable indépendante avec la variable dépendante. Tout comme la corrélation simple, la corrélation partielle prend ses valeurs entre –1 et 1, la proximité de 0 indiquant une association relativement faible. À la différence de la corrélation simple,

la corrélation partielle capte le lien entre une variable indépendante et la variable dépendante, compte tenu des autres variables prises en considération. La corrélation partielle traduit la contribution « marginale » d'une variable indépendante à rendre compte de la variable dépendante lorsqu'on introduit cette variable à la suite des autres variables explicatives. Le changement de R^2 sert de base au calcul de la corrélation partielle.

Considérons deux modèles, l'un comportant $k - 1$ variables indépendantes et l'autre k variables indépendantes. Le traitement par régression multiple donnera lieu à deux R^2 : R^2_{k-1} et R^2_k.

$1 - R^2_{k-1}$ résume la part inexpliquée de la variable indépendante dont ne rendent pas compte les $k - 1$ variables indépendantes.

$\Delta R^2 = R^2_k - R^2_{k-1}$ exprime pour sa part la contribution de la variable k à la réduction de la part « inexpliquée » de la variable dépendante.

La corrélation partielle compare ΔR^2 et $1 - R^2_{k-1}$ à partir de l'expression :

$$\sqrt{\frac{\Delta R^2}{1 - R^2_{k-1}}}$$ affectée du signe de $\hat{\beta}_k$.

Sur la base des relations qui lient les définitions de R^2 et du $t_{\hat{\beta}_k}$ de Student, la corrélation partielle peut aussi être obtenue à partir de l'expression :

$$\frac{t_{\hat{\beta}_k}}{\sqrt{t^2_{\hat{\beta}_k} + (n - k - 1)}}$$

ou, de façon équivalente en termes de F :

$$\frac{\sqrt{F_{\hat{\beta}_k}}}{\sqrt{F_{\hat{\beta}_k} + (n - k - 1)}} \text{ avec } F_{\hat{\beta}_k} = t^2_{\hat{\beta}_k}$$

Quelles que soient les modalités de calcul utilisées, les 3 approches en termes de R^2, du t de Student ou de F conduisent strictement au même résultat. En fait, il s'agit de manipulations d'un même ensemble de données à partir de définitions (R^2, t, F) qui renvoient à un même noyau notionnel. *Les résultats livrés par les logiciels du type SPSS ne mettent pas explicitement cette interdépendance en évidence. Il revient à la démarche*

de recherche d'interpréter les différents résultats obtenus non comme des informations strictement nouvelles, mais comme des modalités différentes de lire un même noyau d'informations.

L'exemple donné dans les tableaux suivants permet à la fois de suivre le calcul de la corrélation partielle selon les 3 approches et de mettre en évidence l'interdépendance des résultats livrés lors du traitement des données.

À la suite de l'exemple développé dans ce chapitre à partir des données de **Salempl.sav**, trois régressions ont été effectuées. La variable *Lsal* a tout d'abord été « régressée » sur 5 variables indépendantes (*Educ, Racexp, Inage, Cat2, Cat3*). La seconde régression a pris 4 variables indépendantes en considération : *Educ, Racexp, Cat2, Cat3*. La variable *Inage* a été omise. La troisième régression a exclu *Racexp* et retenu comme variables explicatives : *Educ, Inage, Cat2, Cat3*.

Les R^2 issus des trois régressions figurent dans le tableau suivant :

R^2 (5 var)	R^2 (5 var-Inage)	R^2 (5 var-Racexp)
0,73	0,72	0,73

À l'aide de ces R^2 et du tableau des « coefficients estimés » livré par le SPSS lors du traitement impliquant 5 variables indépendantes, il est possible de suivre par le détail le lien qui associe chaque corrélation partielle aux R^2 ou aux *t* de Student respectifs.

Tableau 8.3

LA RELATION ENTRE R^2, LE T DE STUDENT ET LA CORRÉLATION PARTIELLE

$R^2_{(5\ var)}$	$R^2_{(5\ var\text{-}Image)}$	$R^2_{(5\ var\text{-}Racexp)}$
0,73	0,72	0,73
	$R^2_{(5\ var)} - R^2_{(5\ var\text{-}Ima}$	$R^2_{(5\ var)} - R^2_{(5\ var\text{-}Racexp)}$
	0,01	0,00
	$1-R^2_{(5\ var\text{-}Image)}$	$1-R^2_{(5\ var\text{-}Racexp)}$
	0,28	0,27
	$=(0,01/0,28)^{0,5}$	$=(0,00/0,27)^{0,5}$
	0,18	0,11

Coefficients		Unstandardized Coefficients		t	Sig.	Correlations	
Model		B	Std. Error			Zero-order	Partial
	(Constant)	9,16	0,11	85,82	0,00		
	Scolarité	0,05	0,00	12,21	0,00	0,78	0,49
	racexp	0,01	0,00	2,31	0,02	-0,06	0,11
	inage	0,13	0,03	4,05	0,00	0,12	0,18
	cat2	0,29	0,05	6,21	0,00	-0,01	0,28
	cat3	0,59	0,03	18,88	0,00	0,79	0,66
Dependent Variable: lsal							

	$\frac{t}{\sqrt{t^2+(n-k-1)}}$	$\frac{2,31}{\sqrt{2,31^2+(474-4-1)}}$	0,11
		$\frac{4,05}{\sqrt{4,05^2+(474-4-1)}}$	0,18

Il est aussi possible d'obtenir la corrélation partielle entre une variable indépendante et une variable dépendante à partir du rapport *F*.

Le rapport *F*, qui est couramment utilisé pour montrer l'intérêt à ajouter une ou plusieurs variables explicatives supplémentaires au modèle, se définit comme le rapport entre « la variance additionnelle expliquée par l'introduction de *r* variables indépendantes supplémentaires et la variance inexpliquée » :

$$F = \frac{\Delta R^2 / r}{(1 - R_k^2)/(n-k-1)} \quad \text{où} \quad \Delta R^2 = R_k^2 - R_{k-r}^2$$

Dans le cas présent ($r = 1$), on passe d'une régression à 4 variables indépendantes à une régression à 5 variables indépendantes.

Si on veut s'interroger sur l'opportunité d'inclure *Inage* à titre de cinquième variable explicative, le recours au rapport *F* donnera :

$$F_{\hat{\beta}_{Inage}} = \frac{(0,73-0,72)/1}{(1-0,73)/(474-5-1)} = 16,36$$

Ce résultat n'est rien d'autre que le carré de $t_{\hat{\beta}_{Inage}} = 4,04475$.

Pour évaluer l'intérêt d'ajouter *Inage* aux 4 autres variables explicatives :

- il suffit de considérer $t_{\hat{\beta}_{Inage}} = 4,04475$ et de le confronter à la table de Student avec $n - k - 1$ degrés de liberté, en pratique dans le cas présent à la table normale centrée réduite (en fait, il suffit de consulter Sig.),
- ou de prendre le carré du *t* de Student et de le confronter à la table *F* avec *r* ($r = 1$ dans le cas présent) degrés de liberté « au numérateur » et $n - k - 1$ (374 – 5 – 1 dans le cas présent) degrés de liberté « au dénominateur ».

Étant donné que $F_{\hat{\beta}_k} = t^2_{\hat{\beta}_k}$, il est équivalent de calculer le coefficient de corrélation partielle entre *Inage* et *Lsal* en utilisant $t_{\hat{\beta}_{Inage}}$ ou $F_{\hat{\beta}_{Inage}}$.

7. MÉTHODES « MÉCANIQUES » POUR DÉGAGER UNE ÉQUATION DE RÉGRESSION MULTIPLE

R^2, *la corrélation partielle, le F « calculé » associé à l'ajout (ou au retrait) d'une variable explicative supplémentaire* sont les outils de base utilisés pour sélectionner « mécaniquement » dans un ensemble de variables explicatives celles qui sont le plus aptes à rendre compte de la variable dépendante.

La méthode « progressive » (« Forward ») introduit comme première variable explicative la variable capable de fournir à partir d'une régression simple le R^2 le plus élevé. Les autres variables sont successivement introduites en fonction de leur capacité à augmenter R^2. Le processus s'arrête lorsque la dernière variable introduite n'est pas à même d'apporter une contribution statistiquement significative.

La méthode « Backward » ou à rebours part du modèle entier et élimine progressivement les variables explicatives les plus inaptes (statistiquement) à rendre compte de la variable dépendante. Le processus s'arrête dès que toutes les variables retenues contribuent de façon statistiquement significative à expliquer la variable dépendante.

La méthode « pas à pas » (« Stepwise ») s'apparente à la méthode progressive, mais elle introduit à chaque étape un contrôle supplémentaire permettant de « renvoyer » certaines variables du fait qu'elles sont devenues « provisoirement superflues » lors de l'ajout de nouvelles variables. Le processus s'arrête lorsqu'il n'est plus possible de faire entrer une variable capable d'améliorer de manière « statistiquement significative » la performance de la régression en termes de R^2.

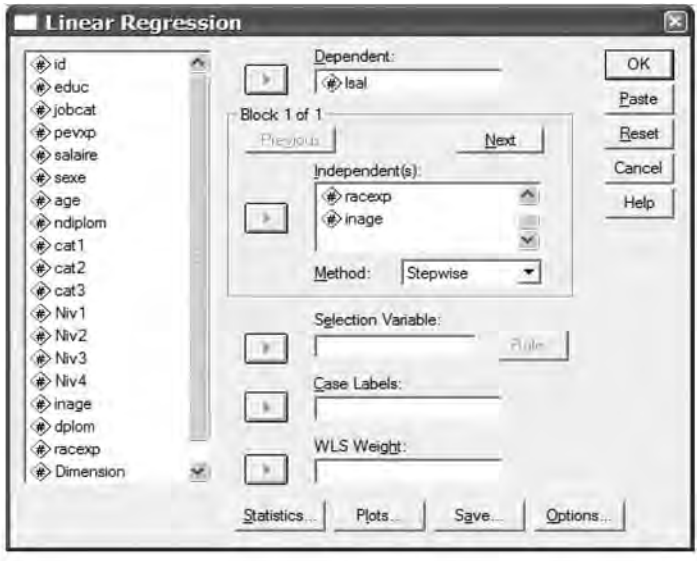
Ces méthodes peuvent se révéler utiles lorsqu'on dispose non pas d'un modèle théorique clairement défini, mais de bases de données comportant une gamme variée de variables potentiellement « explicatives ». En sciences sociales, l'interrogation bivariée et multivariée d'une base de données à partir d'hypothèses de « bon sens » ou d'hypothèses « partielles » s'inspirant des travaux existants constitue un outil souvent très utile dans le processus de réflexion. Par contre, ces méthodes sont aveugles parce qu'elles sont guidées par des principes « mécaniques » de performance statistique. Elles n'apportent aucune réponse aux problèmes de colinéarité. Le choix final des variables explicatives, le choix d'une tentative de solution en présence de colinéarité critique relève de la personne qui mène la recherche.

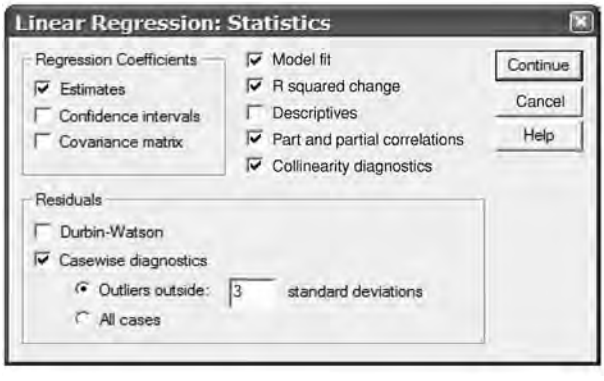
Dans le cas qui a servi d'exemple dans ce chapitre, la variable *Racexp* a été éliminée en raison de la colinéarité qui l'associait à d'autres variables. La variable *Sexe* a été retenue dans le modèle final. On pourrait douter de l'opportunité de ce choix dans un contexte où la discrimination selon le *Sexe* ne serait « apparemment » pas en cause. Par contre, la « découverte » de l'incidence partielle « très significative » de la variable *Sexe* sur la variable de salaire (*Lsal*) ouvre de nouvelles pistes d'interrogation et d'interprétation que la recherche sera amenée à approfondir.

La figure suivante montre comment faire appel à la procédure *Stepwise* en SPSS.

Figure 8.11

LA PROCÉDURE* STEPWISE *EN SPSS





Les extraits des résultats présentés par SPSS montrent assez bien le cheminement de la logique de sélection des variables indépendantes. Dans l'encadré 8.13, on peut y suivre dans la colonne *R Square change*, la croissance progressivement plus faible de R^2 au fur et à mesure que sont introduites des variables indépendantes supplémentaires.

Encadré 8.13

LA RÉGRESSION STEPWISE *ET LES CHANGEMENTS EN TERMES DE* R^2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	,787[a]	,619	,618	,24554	,619	766,560	1	472	,000
2	,837[b]	,700	,699	,21810	,081	127,273	1	471	,000
3	,865[c]	,748	,747	,19996	,048	90,299	1	470	,000
4	,869[d]	,755	,753	,19747	,007	12,949	1	469	,000
5	,873[e]	,763	,760	,19450	,008	15,419	1	468	,000

a. Predictors: (Constant), cat3
b. Predictors: (Constant), cat3, Sexe
c. Predictors: (Constant), cat3, Sexe, Scolarité
d. Predictors: (Constant), cat3, Sexe, Scolarité, cat2
e. Predictors: (Constant), cat3, Sexe, Scolarité, cat2, inage

Dans l'encadré ci-dessous, on remarquera par ailleurs que l'introduction des variables indépendantes se réalise selon l'ordre décroissant des corrélations partielles disponibles (Colonne *Partial Correlation*). Toutes les variables retenues ont une incidence statistiquement très significative sur *Lsal*. La variable *Racexp* est exclue au terme du processus. Le coefficient estimé qui lui est associé n'est pas statistiquement significatif (Sig. = 0,734). La variable, on le sait, comporte un problème de colinéarité avec certaines autres variables indépendantes (*Tolerance* = 0,278 *et Vif* = 3,604).

Encadré 8.14

LA RÉGRESSION STEPWISE *ET LA CORRÉLATION PARTIELLE*

Excluded Variables[f]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	Scolarité	,348[a]	10,901	,000	,449	,634
	Sexe	,300[a]	11,282	,000	,461	,902
	cat2	,079[a]	2,774	,006	,127	,987
	inage	,102[a]	3,625	,000	,165	,999
	racexp	-,026[a]	-,903	,367	-,042	,998
2	Scolarité	,283[b]	9,503	,000	,401	,603
	cat2	,000[b]	-,017	,987	-,001	,912
	inage	,115[b]	4,662	,000	,210	,997
	racexp	-,109[b]	-4,235	,000	-,192	,928
3	cat2	,092[c]	3,598	,000	,164	,799
	inage	,069[c]	2,933	,004	,134	,947
	racexp	-,051[c]	-2,052	,041	-,094	,860
4	inage	,093[d]	3,927	,000	,179	,897
	racexp	-,085[d]	-3,343	,001	-,153	,782
5	racexp	-,015[e]	-,341	,734	-,016	,278

a. Predictors in the Model: (Constant), cat3
b. Predictors in the Model: (Constant), cat3, Sexe
c. Predictors in the Model: (Constant), cat3, Sexe, Scolarité
d. Predictors in the Model: (Constant), cat3, Sexe, Scolarité, cat2
e. Predictors in the Model: (Constant), cat3, Sexe, Scolarité, cat2, inage
f. Dependent Variable: lsal

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations		
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1	(Constant)	10,212	,012		821,315	,000			
	cat3	,818	,030	,787	27,687	,000	,787	,787	,787
2	(Constant)	10,099	,015		678,038	,000			
	cat3	,720	,028	,693	26,058	,000	,787	,768	,658
	Sexe	,239	,021	,300	11,282	,000	,517	,461	,285
3	(Constant)	9,624	,052		185,734	,000			
	cat3	,559	,030	,538	18,332	,000	,787	,646	,424
	Sexe	,197	,020	,248	9,915	,000	,517	,416	,229
	Scolarité	,039	,004	,283	9,503	,000	,697	,401	,220
4	(Constant)	9,556	,055		175,286	,000			
	cat3	,556	,030	,535	18,454	,000	,787	,649	,422
	Sexe	,170	,021	,214	8,094	,000	,517	,350	,185
	Scolarité	,044	,004	,323	10,273	,000	,697	,429	,235
	cat2	,158	,044	,092	3,598	,000	-,012	,164	,082
5	(Constant)	9,418	,064		146,540	,000			
	cat3	,566	,030	,545	19,015	,000	,787	,660	,428
	Sexe	,171	,021	,215	8,257	,000	,517	,357	,186
	Scolarité	,042	,004	,306	9,785	,000	,697	,412	,220
	cat2	,197	,044	,115	4,451	,000	-,012	,202	,100
	inage	,073	,019	,093	3,927	,000	,122	,179	,088

a. Dependent Variable: lsal

8. « PRÉVISIONS CONDITIONNELLES » À PARTIR DES ESTIMATIONS OBTENUES PAR RÉGRESSION MULTIPLE

À l'aide des $\hat{\beta}_i$ estimés, on calcule les valeurs estimées (prédites) de la variable dépendante à partir de l'expression

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \times X_{1i} + \hat{\beta}_2 \times X_{2i} + ... + \hat{\beta}_k \times X_{ki}$$

Ces « prévisions » sont conditionnelles aux valeurs des différentes variables indépendantes. Le programme SPSS ne fournit pas seulement les « prévisions » pour les situations qui ont servi au calcul des coefficients $\hat{\beta}_i$. Il suffit d'ajouter des observations pour les variables indépendantes intervenant dans la régression tout en omettant de donner des informations pour la variable dépendante et SPSS calculera automatiquement pour la variable dépendante des « prévisions » conditionnelles aux valeurs imposées aux variables indépendantes en se servant des estimations $\hat{\beta}_i$ préalablement obtenues.

Conjointement à chaque prévision conditionnelle, le programme SPSS permet de calculer deux types d'intervalles de confiance.

- *L'intervalle de confiance « moyen »* : qui tient compte du modèle retenu mais fait abstraction du terme d'erreur. Les bornes en sont définies par LMCL_1 et UMCL_1.

- *L'intervalle dit « individuel »* tient de plus compte de la variance du terme d'erreur. Les bornes sont désignées par SPSS sous les vocables LICL_1 et UICL_1.

Dans le cas qui sert d'exemple, les prévisions conditionnelles et les bornes des deux intervalles à l'intérieur desquelles il est « légitime » de penser que se situe la « vraie » prévision, sont exprimées en logarithmes naturels. Pour les interpréter directement en termes de salaire, il s'agit d'en demander l'exponentielle.

La variable *Salprevu* a été obtenue par la transformation suivante :

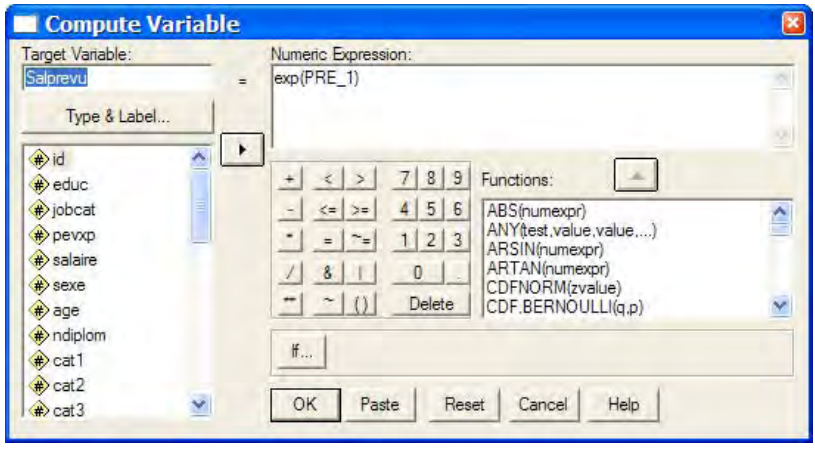


Tableau 8.4

LA RÉGRESSION MULTIPLE : PRÉVISIONS CONDITIONNELLES SUR LA BASE DES ESTIMATIONS

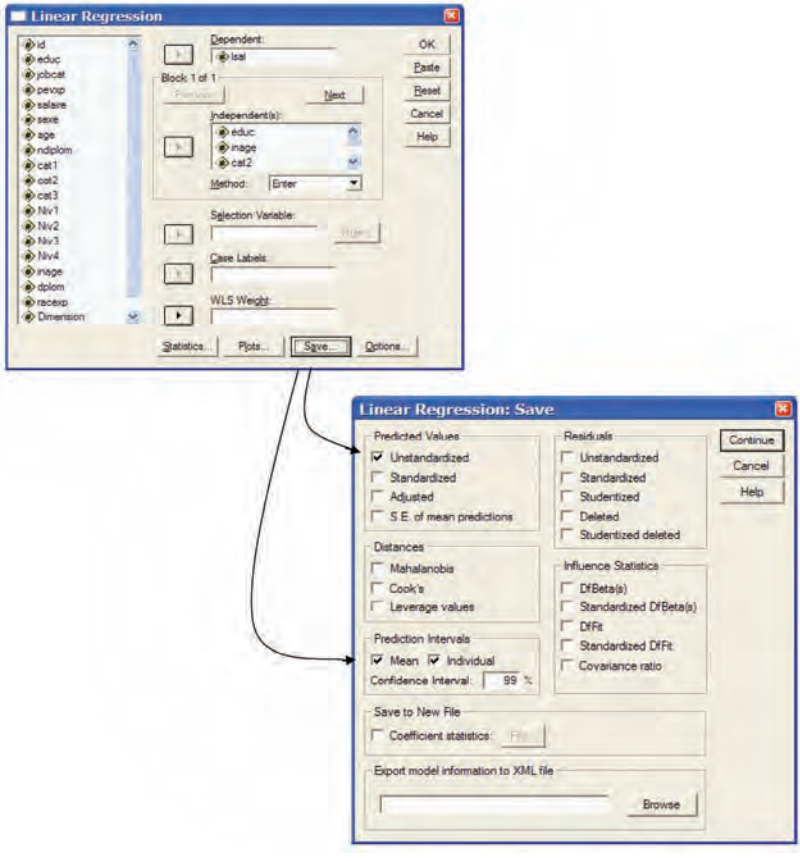
$$Salprevu_i = e^{PREV\_1_i}$$

id	educ	jobcat	pevxp	salaire	sexe	age	ndiplom	cat1	cat2	cat3	lsal	inage	dplom	racexp	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1	Salprevu
475	12	1	163	.	1	70	2	1	0	0	.	1,43	2	12,77	10,2	10,1	10,3	9,7	10,7	26889
476	15	1	86	.	1	42	3	1	0	0	.	2,38	3	9,27	10,4	10,4	10,4	9,9	10,9	32713
477	16	1	19	.	1	37	3	0	0	1	.	2,7	3	4,36	11,0	11,0	11,1	10,5	11,5	61544
478	16	3	19	.	1	37	3	0	0	1	.	2,7	3	4,36	11,0	11,0	11,1	10,5	11,5	61544
479	19	3	129	.	1	49	4	0	0	1	.	2,04	4	11,36	11,1	11,0	11,2	10,6	11,6	66539
481	12	1	163	.	0	70	2	1	0	0	.	1,43	2	12,77	10,0	10,0	10,1	9,5	10,5	22658
482	15	1	86	.	0	42	3	1	0	0	.	2,38	3	9,27	10,2	10,2	10,3	9,7	10,7	27566
483	16	1	19	.	0	37	3	0	0	1	.	2,7	3	4,36	10,9	10,8	10,9	10,3	11,4	51860
484	16	3	19	.	0	37	3	0	0	1	.	2,7	3	4,36	10,9	10,8	10,9	10,3	11,4	51860
485	19	3	129	.	0	49	4	0	0	1	.	2,04	4	11,36	10,9	10,9	11,0	10,4	11,4	56069

Les instructions qui ont permis d'obtenir les prévisions PRE_1 et les intervalles de prédiction « moyens » et « individuels » sont consignées dans la figure suivante :

Figure 8.12

LA RÉGRESSION MULTIPLE :
COMMANDES SPSS POUR OBTENIR LES PRÉVISIONS CONDITIONNELLES ET LES INTERVALLES DE PRÉVISION



CHAPITRE

9

La régression logistique

Lorsque la variable dépendante est une variable qualitative (nominale ou ordinale), l'approche par régression « classique » telle que présentée au chapitre précédent est inadéquate. À partir de valeurs données consignées dans les variables indépendantes, la régression classique renvoie à une estimation quantitative de la valeur dépendante. Lorsque la variable dépendante est qualitative, le problème est d'un autre ordre. La question qui se pose est la suivante : à quelle catégorie de la variable qualitative renvoient les valeurs prises par les variables explicatives ? Comme on ne se situe pas dans un monde de certitude absolue, la question peut être reformulée en probabilité d'occurrence. Selon quelles probabilités les valeurs prises par les variables explicatives renvoient-elles aux différentes catégories de la variable qualitative dépendante ?

Les fonctions qui associent les variables explicatives aux probabilités d'occurrence des catégories d'une variable qualitative dépendante doivent au minimum respecter certains critères. Elles doivent faire en sorte que la variable dépendante prenne ses valeurs entre 0 et 1, puisqu'il s'agit de probabilités. Elles doivent tenir compte du fait qu'elles renvoient à des catégories exhaustives et exclusives. Les probabilités relatives à l'occurrence de chaque catégorie sont soumises à la contrainte que la somme des probabilités est égale à 1.

Diverses fonctions peuvent respecter ces critères. La fonction dite « logistique » est d'usage courant en raison de ses caractéristiques et parce qu'elle se plie relativement bien à l'estimation de ses paramètres.

Trois cas seront pris en considération :

- le cas où la variable dépendante qualitative comporte deux catégories exhaustives et exclusives (par exemple, s'il s'agit de vacanciers, rester à la maison ou entreprendre un voyage) ;
- le cas où la variable dépendante qualitative nominale comporte plus de deux catégories (par exemple, résider en ville, en périphérie ou à la campagne) ;
- le cas où la variable qualitative dépendante est ordinale (par exemple, être mécontent, indifférent ou content).

Les trois cas correspondent à des situations typiques « d'analyse discriminante » (selon laquelle on cherche à prédire l'appartenance à une catégorie en fonction d'un certain nombre de caractéristiques). La régression dite « logistique » est de fait une approche privilégiée pour effectuer une analyse discriminante. On remarquera que la régression « classique » peut aussi être assimilée à une méthode d'analyse discriminante dans la mesure où il est loisible de regrouper les « prévisions » quantitatives obtenues en catégories exhaustives et exclusives.

1. VARIABLE DÉPENDANTE BINAIRE ET RELATION LOGISTIQUE

Dans le cadre de la présentation de l'analyse de régression logistique binaire, nous utiliserons les données figurant au tableau suivant.

La variable binaire Y : le fait d'avoir (L_2 ; $L_2 = 1$) ou de ne pas avoir (L_1 ; $L_1 = 0$) un plan d'affaires sera confrontée à deux variables indépendantes, une variable quantitative (X_1 : le chiffre d'affaires par agent) et une variable qualitative binaire X_2 (le chiffre 1 correspondant à l'utilisation de la publicité à la télévision et le chiffre 0, à la non-utilisation de la télévision comme support de publicité).

Tableau 9.1

AGENCES DE VOYAGES ET PLAN D'AFFAIRES

Y	Agences	Ventes	Vendeurs	Années	X1	X2
1	1	3604	6	4	360,4	0
0	2	3400	8	2	242,86	0
1	3	3857	6	3	321,42	0
0	4	3666	7	4	203,67	0
0	5	4490	9	1	187,08	1
0	6	3260	6	4	203,75	0
1	7	6148	10	3	279,45	1
0	8	5810	11	2	181,56	1
0	9	4370	7	3	168,08	1
1	10	4070	6	4	290,71	0
0	11	4150	9	1	207,5	1
0	12	2610	7	1	326,25	0
1	13	4721	7	3	262,28	0
0	14	6050	11	4	177,94	1
0	15	6001	10	2	157,92	1
1	16	4522	6	2	251,22	0
1	17	5812	9	4	264,18	1
0	18	3240	6	3	202,5	0
0	19	4610	8	4	177,31	1
1	20	5984	8	2	213,71	1
1	21	4172	7	4	298	0
0	22	5140	9	3	160,63	1
0	23	3870	8	3	215	0
0	24	2760	6	3	276	0
1	25	7623	9	4	211,75	1
1	26	6793	8	3	188,69	1
0	27	2810	6	2	281	0
1	28	5041	9	2	360,07	0
0	29	4970	10	1	207,08	1
1	30	6353	9	4	244,35	1
0	31	3290	7	3	205,63	0
1	32	3604	6	2	300,33	0
1	33	7850	11	2	245,31	1
1	34	8366	11	2	220,16	1
0	35	5460	10	4	160,59	1
0	36	3630	7	3	181,5	1
0	37	4240	8	3	176,67	1
0	38	2940	6	3	183,75	0
1	39	5147	8	4	257,35	1
1	40	5896	8	2	245,67	1
.	.	.	.	.	240	1
.	.	.	.	.	240	0
.	.	.	.	.	250	1
.	.	.	.	.	250	0

P_{L2} représente la probabilité d'occurrence d'avoir un plan d'affaires, P_{L1}, la probabilité de ne pas avoir de plan d'affaires. $P_{L2} + P_{L1} = 1$. Dès lors, $P_{L1} = 1 - P_{L2}$ ou encore $P_{L2} = 1 - P_{L1}$. Dès qu'on connaît une probabilité, on connaît automatiquement l'autre probabilité. Prenons arbitrairement P_{L2} comme point de repère.

1.1. REPÈRES THÉORIQUES

Pour obtenir une représentation adéquate des probabilités de L_2 en fonction de X_1 et de X_2, il s'agit d'adopter une fonction telle que toutes les valeurs prédites pour P_{L2} soient enfermées entre les bornes 0 et 1. Une fonction très généralement retenue pour capter ce genre de situation est la fonction logistique. Il existe bien sûr un certain arbitraire dans le choix de cette fonction. La fonction logistique n'est pas l'unique fonction possible, loin de là. Néanmoins, la fonction logistique présente un certain nombre de propriétés qui facilitent l'interprétation des résultats des estimations et des simulations subséquentes. Par ailleurs, cette fonction se prête à des extensions lui permettant d'intégrer de façon heureuse le traitement de relations multivariées et l'analyse de variables dépendantes qualitatives nominales comportant plus de deux alternatives.

Pour P_{L2}, le modèle logistique de base se présente comme suit :

$$P_{L2} = \frac{1}{1+e^{-Z}} = \frac{e^{Z}}{1+e^{Z}}$$

où Z représente une fonction linéaire de plusieurs variables indépendantes.

$$P_{L2} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3 \ldots)}}$$

Cette dernière relation est transformée sous forme de *Logit*.

Reprenons l'expression

$$P_{L2} = \frac{1}{1+e^{-Z}} = \frac{e^{Z}}{1+e^{Z}}$$

et transformons-la comme suit :

$$1-P_{L2} = 1-\frac{1}{1+e^{-Z}} = \frac{e^{-Z}}{1+e^{-Z}}$$

En confrontant les deux dernières relations, on obtient par transformation :

$$\frac{P_{L2}}{1-P_{L2}} = e^{Z}$$

Le rapport

$$\frac{P_{L2}}{1-P_{L2}}$$

est connu en langue anglaise sous le nom de *Odd*.

$$Odd \equiv \frac{P_{L2}}{1-P_{L2}} \equiv \Omega$$

Si on prend le logarithme de *Odd*, on obtient l'expression connue sous le nom de *Logit*.

Par définition,

$$Logit(P_{L2}) \equiv \ln\left(\frac{P_{L2}}{1-P_{L2}}\right) \equiv \ln\Omega$$

Si on recourt à cette définition, la relation logistique peut s'écrire :

$$Logit(P_{L2}) = \ln\left(\frac{P_{L2}}{1-P_{L2}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots$$

Odd se prête à une analyse de sensibilité relativement simple.

Soit

$$\Omega_{L2} = e^{\beta_0 + \beta_1 (X_1 + 1) + \beta_2 X_2 + \beta_3 X_3 \ldots}$$

et

$$\Omega^*_{L2} = e^{\beta_0 + \beta_1 (X_1 + 1) + \beta_2 X_2 + \beta_3 X_3 \ldots}$$

Il apparaît que l'impact de la variation d'une unité de X_1 sur Ω_{L2} est égal à e^{β_1}.

Ce que traduisent les deux expressions suivantes :

$$\ln(\Omega^*) = \ln\Omega + \beta_1$$

et

$$\Omega^* = \Omega \cdot e^{\beta_1}$$

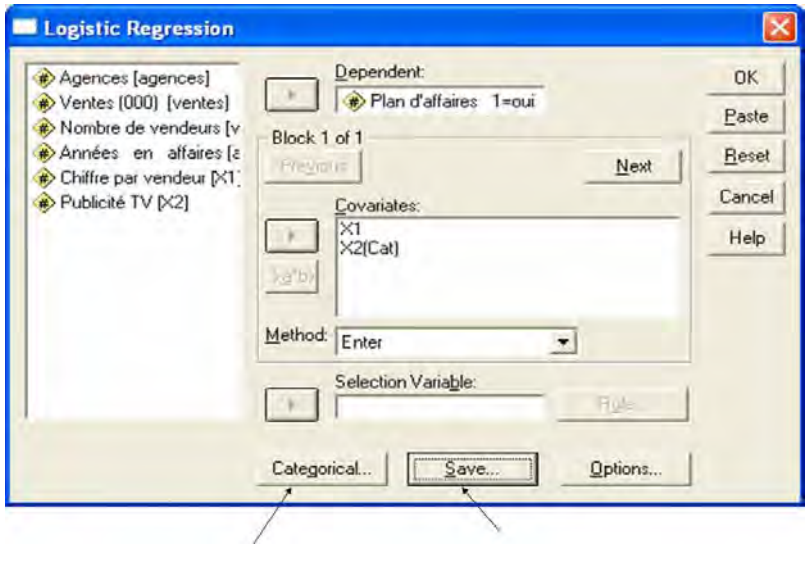
L'estimation du modèle *Logit* s'effectue généralement par la méthode du **maximum de vraisemblance.** Il s'agit dans un premier temps de définir la fonction de vraisemblance *L*, *L* représentant la probabilité d'observer les données d'échantillon sous l'hypothèse que le modèle est vrai. La procédure revient à choisir à l'aide d'un processus itératif les estimations des paramètres qui permettent de maximiser la fonction *L*. Cette démarche a été analysée par différents auteurs (Gouriéroux, Greene, Maddala, Kennedy, etc.). Elle est accessible, sous différentes versions, à partir de plusieurs logiciels statistiques courants, dont SAS, SPSS, LIMDED et STATA. Les logiciels fournissent en même temps des informations « omnibus » permettant d'évaluer globalement la qualité des estimations obtenues, des informations plus spécifiques sur les coefficients estimés, leur erreur type et les tests associés, et des informations permettant d'évaluer la distribution des résidus. La structure des différents tests et leur signification dans l'interprétation des résultats seront détaillées lors de la présentation des estimations. Les logiciels permettent aussi d'effectuer des prévisions conditionnelles à des valeurs données pour les variables explicatives.

1.2. LA LOGISTIQUE BINAIRE AVEC SPSS

La figure suivante montre comment se fait l'accès à la « régression logistique binaire » avec SPSS.

Figure 9.1

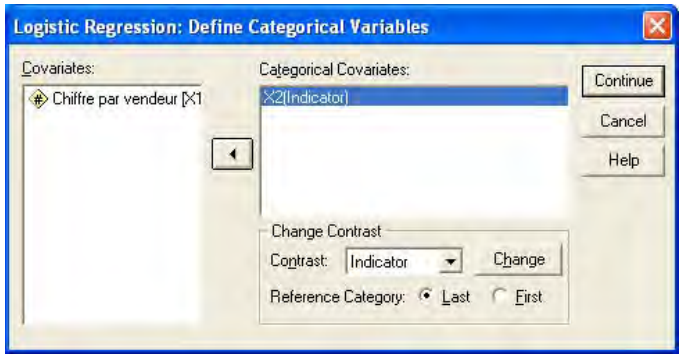
LA LOGISTIQUE BINAIRE AVEC SPSS



En cliquant sur **CATEGORICAL**, il est possible de préciser que la variable X_2 (le recours à la télévision comme support publicitaire) est une variable qualitative nominale.

Figure 9.2

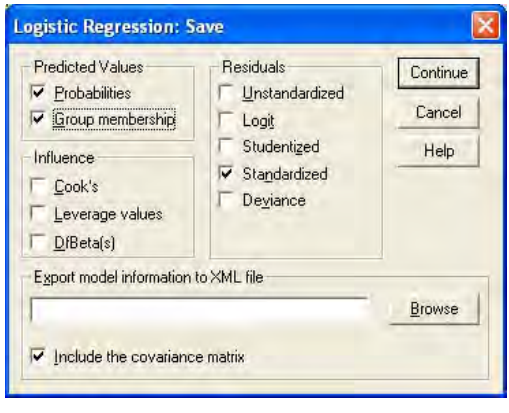
LA LOGISTIQUE BINAIRE AVEC SPSS :
LES VARIABLES QUALITATIVES INDÉPENDANTES



En cliquant sur **SAVE**, il est possible, entre autres, de récupérer les résidus normalisés, les probabilités estimées par la procédure et la catégorie d'appartenance (L_2, L_1) la plus probable.

Figure 9.3

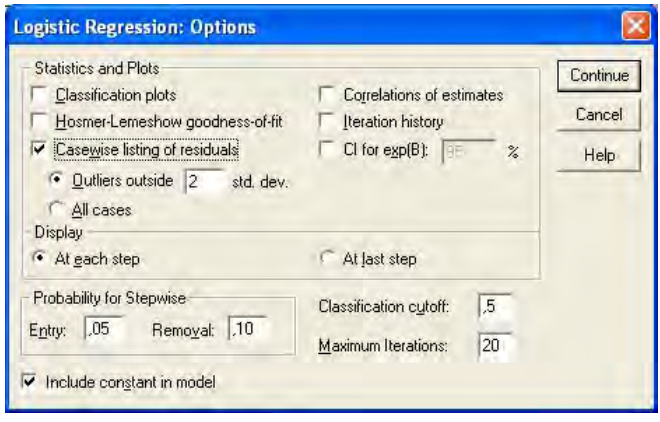
LA LOGISTIQUE BINAIRE AVEC SPSS :
RÉCUPÉRATION DE VARIABLES CALCULÉES



En cliquant sur **OPTIONS**, on dispose par ailleurs d'informations complémentaires, en particulier sur la présence de résidus dont l'ordre de grandeur est particulièrement élevé et risque de ce fait de perturber la précision des estimations.

Figure 9.4

LA LOGISTIQUE BINAIRE AVEC SPSS :
RÉCUPÉRATION D'INFORMATIONS COMPLÉMENTAIRES



1.3. LES RÉSULTATS DES ESTIMATIONS

1.3.1. Les tests « omnibus »

L'estimation des paramètres de la régression logistique est réalisée *par maximisation de la fonction de vraisemblance L.* Cette démarche consiste à rechercher les estimations qui maximisent la probabilité d'obtenir l'échantillon observé des *Y.* En pratique, cette maximisation s'effectue à partir du logarithme naturel de cette fonction (LL : *Log Likehood*). En multipliant LL par –2, on obtient une distribution qui suit assez bien la distribution du khi-carré.

Dans un premier temps, on calcule les valeurs de $-2LL_{(1)}$, en ne retenant que le terme constant comme variable indépendante. Dans un second temps, on calcule $-2LL_{(2)}$ en tenant compte de l'ensemble des variables indépendantes considérées dans l'analyse. L'introduction des variables indépendantes en plus du terme constant réduit l'ordre de grandeur de –2LL. La différence entre $-2LL_{(1)}$ et $-2LL_{(2)}$, appelée le *khi-carré du modèle* (*model Chi-square*), est confrontée à la table de khi-carré. Si le test est significatif au niveau 0,05, il est admis qu'au moins une variable indépendante exerce une influence sur la variable dépendante.

Dans le listing, ces résultats se présentent comme suit :

Encadré 9.1

LA LOGISTIQUE BINAIRE : LES INFORMATIONS « OMNIBUS »

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	55,051	-,200
	2	55,051	-,201

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 55,051

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than ,001.

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	X1	X2(1)
Step 1	1	32,522	-6,579	,031	-1,930
	2	29,556	-9,922	,048	-3,103
	3	29,150	-11,664	,056	-3,797
	4	29,136	-12,052	,058	-3,962
	5	29,136	-12,068	,058	-3,968
	6	29,136	-12,068	,058	-3,968

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 55,051

d. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	25,915	2	,000
	Block	25,915	2	,000
	Model	25,915	2	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29,136[a]	,477	,638

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

En ne tenant compte que du terme constant, $-2LL_{(1)} = 55{,}051$ après deux itérations.

En tenant compte de X_1 et de X_2 et du terme constant, $-2LL_{(2)}$ = 29,136 après six itérations (N.B. : on remarquera que –2LL est toujours positif).

Le *khi-carré du modèle* (*model Chi-square*) appelé aussi *Log Likehood ratio* (LR) = 55,051 – 29,136 = 25,925.

Confronté à la table du khi-carré, *le khi-carré du modèle 25,925 (deux degrés de liberté, deux variables explicatives) est significatif au niveau 0,000. On accepte donc qu'au moins une des variables X_1 et X_2 exerce une influence sur la variable dépendante.*

En divisant le *khi-carré du modèle* 25,925 par 55,051 (c'est-à-dire $-2LL_{(1)}$), on obtient un *pseudo-Rcarré* : 0,471. Les valeurs *Cox and Snell RSquare* (0,477) et *Nagelkerke RSquare* (0,638) sont aussi des *pseudo-Rcarré*. Ils ont tous les trois en commun le fait qu'ils manipulent fondamentalement l'information contenue dans les –2LL. À ce titre, on ne peut les comparer de manière stricte au R^2 de la régression multiple développée dans le chapitre précédent.

1.3.2. Coefficients estimés et tests de Wald

Si on divise chaque coefficient estimé par son erreur type et qu'on prend la valeur absolue du résultat calculé, on obtient le Wald (*Z* dans le listing). *Z* peut être confronté à la table normale si on dispose d'un échantillon relativement large, ou à la table de Student pour un échantillon relativement petit. Mis au carré, *Z* suit approximativement le khi-carré avec un degré de liberté (on teste l'hypothèse de nullité d'un seul paramètre).

Sur le listing, les résultats se présentent comme suit :

Encadré 9.2

LOGISTIQUE BINAIRE : COEFFICIENTS ESTIMÉS ET WALD

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1	,058	,018	10,570	1	,001	1,060
	X2(1)	-3,968	1,652	5,773	1	,016	,019
	Constant	-12,068	3,657	10,887	1	,001	,000

a. Variable(s) entered on step 1: X1, X2.

On remarquera que sous la colonne Wald figure Z^2, ce qui signifie que SPSS se réfère à la table du khi-carré avec un degré de liberté (on teste l'hypothèse qu'un paramètre = 0 contre l'hypothèse qu'il soit différent de zéro). Tous les trois coefficients estimés passent le seuil de significativité (Sig. ≤ 0,05).

Il existe une autre manière de tester la significativité des coefficients estimés : le *Likehood Ratio* (LR).

Si on demande successivement la logistique en laissant figurer dans les variables indépendantes d'abord le seul terme constant, puis le terme constant et X_1, puis le terme constant, X_1 et X_2, ont obtient les –2LL suivants :

- –2LL (terme constant) = 55,051
- –2LL (terme constant et X_1) = 39,258
- –2LL (terme constant, X_1, X_2) = 29,136

Le *Likehood Ratio* (LR) dans le cas de la prise en considération de X_1 = 55,051 – 39,258 = **15,793**.

Si on tient compte de X_2 en plus de X_1, le LR spécifique à X^2 = 39,258 – 29,136 = **10,122**.

Les LR sont confrontés au khi-carré avec un degré de liberté (on considère l'hypothèse de la nullité d'un paramètre).

Dans le cas d'échantillons relativement grands,

LR ≈ Z^2 de Wald.

Dans le cas d'échantillons relativement petits, LR diverge du Z^2 de Wald. C'est la raison pour laquelle les statisticiens préfèrent le LR. En pratique, les conclusions restent généralement les mêmes, en dépit des divergences chiffrées. C'est le cas dans l'exemple traité.

Encadré 9.3

LOGISTIQUE BINAIRE :
LES TESTS LOG LIKEHOOD RATIO PAR VARIABLE EXPLICATIVE

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	29,136[a]	,000	0	.
X1	55,047	25,911	1	,000
X2	39,258	10,122	1	,001

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

1.3.3. Autres contrôles de validité des résultats

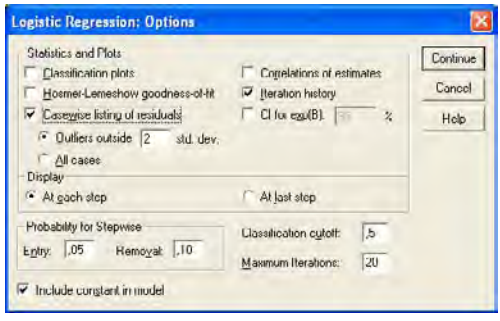
Parallèlement à la régression classique, les estimations des paramètres de la logistique par maximum de vraisemblance sont sensibles à la répartition des résidus et aux problèmes de colinéarité entre les variables indépendantes.

L'examen des résidus se réalise de manière similaire à ce qui se pratique dans la régression classique. Pour ce faire, il y aurait lieu de demander au logiciel SPSS de créer une nouvelle variable où figureraient les résidus. Le plus pratique est de demander les résidus sous forme normalisée (*Zres.*) et d'inviter SPSS à identifier les cas où *Zres.* se situe en dehors des bornes –2 et 2. Dans les cas où *Zres.* se situe en dehors des bornes mentionnées, il convient de rechercher la cause de cette situation « exceptionnelle ». Ce n'est pas toujours possible. Il est par contre risqué d'éliminer automatiquement l'observation problématique. Ce faisant, on fait souvent survenir d'autres cas apparemment problématiques.

Pour identifier les cas où *Zres.* se situe en dehors des bornes critiques, on ouvre *Options* dans le fenêtre de commande de **Binary Logistic** et on sollicite les commandes comme suit :

Figure 9.5

LA LOGISTIQUE BINAIRE :
IDENTIFICATION DES VALEURS EXTRÊMES DES RÉSIDUS



Selon le listing produit, la douzième observation donne lieu à un résidu normalisé (–4,380) dont la valeur est nettement en deçà de –2. On est devant un cas où il convient de s'interroger sur les raisons de cette situation. En tout état de cause, éliminer l'observation provoque l'apparition de trois nouveaux cas en dehors des bornes repères (–2 et 2). Cela ne semble pas une démarche opportune dans le cas présent.

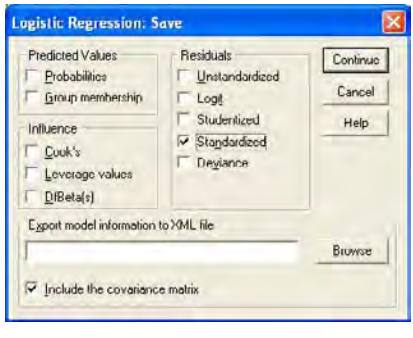
Casewise List[b]

Case	Selected Status[a]	Observed Plan d'affaires 1=oui 0=non	Predicted	Predicted Group	Temporary Variable Resid	Temporary Variable ZResid
12	S	0**	,950	1	-,950	-4,380

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Pour obtenir la variable des résidus normalisés (*Zres_1*), on sollicite la clé **SAVE** dans le tableau de commande **Binary Logistic**.



Une nouvelle variable (*Zre_1*) apparaît alors dans le fichier de données. Il s'agira d'en examiner les données à la recherche de régularités qui pourraient avoir perturbé les estimations.

1.4. PRÉVISIONS CONDITIONNELLES

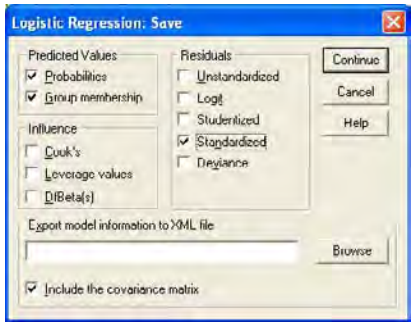
SPSS permet de calculer des « prévisions » à partir des données correspondant aux variables indépendantes. Sur la base du modèle estimé, SPSS est aussi à même de calculer des prévisions conditionnelles pour des situations où l'on ne dispose pas d'observations pour la variable dichotomique dépendante.

Pour obtenir des valeurs « prévues », on procède comme suit :

Après avoir sollicité la clé **SAVE** dans le fenêtre de commande **Binary Logistic**, on ouvre les commandes tel qu'énoncé :

Figure 9.6

LA RÉCUPÉRATION DES PRÉVISIONS EN LOGISTIQUE BINAIRE



Les commandes permettent de récupérer les probabilités calculées d'occurrence et les occurrences calculées pour les deux catégories (dans l'exemple, avoir un plan d'affaires ; ne pas avoir un plan d'affaires). En fait, chaque fois que la probabilité calculée atteint ou dépasse 0,50, le programme « décide » qu'on est devant une situation prévoyant la présence d'un plan d'affaires. Dans les situations où l'on dispose d'observations pour la variable dichotomique Y (avoir ou ne pas avoir un plan d'affaires), il est ainsi possible de constater dans quelle mesure le modèle estimé permet de rejoindre la réalité observée. Dans les cas où l'on ne dispose pas d'observations pour la variable dichotomique Y, on dispose de prévisions dont la valeur dépend du modèle retenu et de la qualité des estimations obtenues.

Ci-dessous figure un extrait du fichier de données avec les prévisions demandées (*PRE_1*, *PGR_1*) et la variable *ZRE_1* contenant les résidus normalisés.

Encadré 9.4

PRÉVISIONS À PARTIR DE LA LOGISTIQUE BINAIRE

Y	Agences	Ventes	Vendeurs	Années	X1	X2	PRE_1	PGR_1	ZRE_1
1	34	8366	11	2	220,16	1	0,67855	1	0,68828
0	35	5460	10	4	160,59	1	0,06179	0	-0,25663
0	36	3630	7	3	181,5	1	0,18197	0	-0,47165
0	37	4240	8	3	176,67	1	0,14376	0	-0,40976
0	38	2940	6	3	183,75	0	0,00477	0	-0,06924
1	39	5147	8	4	257,35	1	0,94843	1	0,23317
1	40	5896	8	2	245,67	1	0,90308	1	0,3276
.	.	.	.	.	240	1	0,87013	1	.
.	.	.	.	.	240	0	0,11242	0	.
.	.	.	.	.	250	1	0,92302	1	.
.	.	.	.	.	250	0	0,18479	0	.

2. VARIABLE DÉPENDANTE NOMINALE COMPORTANT PLUS DE DEUX CATÉGORIES

Jusqu'à maintenant, l'approche logistique a été envisagée par rapport à une variable dépendante nominale binaire (la présence ou l'absence d'une caractéristique ou d'un comportement). Qu'en est-il si le nombre de catégories de la variable qualitative nominale est supérieur à deux (trois candidats à une élection par exemple) ? La généralisation de l'approche logistique se fait en termes certes plus complexes que dans la logistique binaire, mais selon un canevas d'interprétation similaire.

Le cas d'une variable nominale comportant trois catégories « exhaustives et exclusives » sera détaillé dans la présentation. La structure de modélisation et d'analyse est semblable lorsque le nombre de catégories est supérieur à trois. Il faut seulement tenir compte d'un nombre plus élevé d'équations, chaque catégorie (moins une) requérant une équation spécifique.

Dans tous les cas, il est entendu que les catégories exclusives et exhaustives correspondent à des distinctions « qui ont un sens ». Il est de la responsabilité du chercheur d'assurer cette condition préliminaire.

Dans le cadre de la présentation de l'analyse de régression logistique concernant une variable nominale comportant trois catégories, nous utiliserons un extrait d'un fichier directement disponible dans le programme

SPSS (**Clinton.sav**). Ce fichier présente certaines caractéristiques (âge, sexe, niveau d'éducation) de personnes qui se sont prononcées sur leur préférence à voir Clinton, Bush ou Perot à la présidence des États-Unis en 1992. La variable nominale dépendante (*Pres92*) comporte des catégories correspondant respectivement aux choix de Perot, de Clinton et de Bush. Dans le cas de l'exemple traité, S1 identifie une réponse favorable à Perot, S2, une réponse favorable à Bush et S3, une réponse favorable à Clinton.

Les probabilités de S1, S2 et S3 sont liées par la contrainte :

$$P_{S1} + P_{S2} + P_{S3} = 1$$

2.1. REPÈRES THÉORIQUES

Si l'on choisit S1 comme catégorie de référence, et en se rappelant que $P_{S1} + P_{S2} + P_{S3} = 1$, les équations logistiques définissant les probabilités de S2, S3 et S1 sont les suivantes.

$$P_{S3} = \frac{P_{S1},\, e^{\beta_{30}+\beta_{31}X_1+\beta_{32}X_2+\beta_{33}X_3\ldots}}{1 + e^{\beta_{20}+\beta_{21}X_1+\beta_{23}X_3\ldots} + e^{\beta_{30}+\beta_{31}X_1+\beta_{32}X_2+\beta_{33}X_3\ldots}}$$

$$P_{S2} = \frac{P_{S1},\, e^{\beta_{20}+\beta_{21}X_1+\beta_{22}X_2+\beta_{23}X_3\ldots}}{1 + e^{\beta_{20}+\beta_{21}X_1+\beta_{23}X_3\ldots} + e^{\beta_{30}+\beta_{31}X_1+\beta_{32}X_2+\beta_{33}X_3\ldots}}$$

$$P_{S1} = \frac{1}{1 + e^{\beta_{20}+\beta_{21}X_1+\beta_{23}X_3\ldots} + e^{\beta_{30}+\beta_{31}X_1+\beta_{32}X_2+\beta_{33}X_3\ldots}}$$

En prenant S1 comme référence, les logits se présentent comme suit :

$$\ln\left(\frac{P_{S2}}{P_{S1}}\right) = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3\ldots$$

$$\ln\left(\frac{P_{S3}}{P_{S1}}\right) = \beta_{30} + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_3\ldots$$

avec la contrainte :

$$P_{S1} = 1 - P_{S2} + P_{S3}$$

On remarquera qu'en raison de cette dernière contrainte, seuls deux logits sont énoncés.

On remarquera aussi qu'on aurait pu formuler les logits en choisissant S2 ou S3 comme référence.

L'interprétation des rapports *Odd* ou des *Logit* (qui expriment les rapports *Odd* sous forme de logarithmes) n'est plus immédiate. Qu'on se rappelle : dans une situation binaire, une augmentation du numérateur (P_{L2}) de *Odd* se traduisait automatiquement par une diminution du dénominateur ($1 - P_{L2}$), ce qui n'est plus le cas dans la situation « polytomique » envisagée.

En même temps, il existe une difficulté systématique à interpréter le sens des coefficients estimés, ceux-ci impliquant toujours une référence à la catégorie repère (S1 dans le cas présent).

Pour contourner ces difficultés d'interprétation des résultats, il est généralement proposé, une fois disponibles les estimations des paramètres, de procéder à des simulations en faisant varier systématiquement une à une chaque variable indépendante selon la modalité suivante. On procède tout d'abord à l'estimation des paramètres. S'il s'agit d'une variable indépendante qualitative binaire (1,0), on impose ensuite la valeur 0 à tous les individus de l'échantillon et, en maintenant inchangées les observations des autres variables indépendantes, on récupère la probabilité « prévue » (relative à la variable dépendante) pour chacun des individus. Pour la même variable indépendante, on impose ensuite la valeur 1 à tous les individus de l'échantillon. On compare ensuite les deux moyennes des probabilités « prévues » obtenues. S'il s'agit d'une variable indépendante quantitative, on impose, après estimation des paramètres du modèle, la moyenne des observations à chaque individu et on calcule les probabilités individuelles « prévues » résultantes pour la variable dépendante. On augmente ensuite d'une unité la moyenne imposée à chaque individu et on recalcule les probabilités individuelles « prévues » résultantes pour la variable dépendante. Ici aussi, on procède à une comparaison des valeurs moyennes « prévues » obtenues. La procédure ne dispense évidemment pas d'examiner les tests statistiques qui aident à interpréter la qualité des estimations obtenues.

Méthodologiquement, la procédure implique aussi que les attentes *a priori* soient formulées sur des variations de probabilités d'occurrence concernant les catégories captées par la variable dépendante.

2.2. LA RÉGRESSION LOGISTIQUE « POLYTOMIQUE » AVEC SPSS

La figure suivante montre comment se fait l'accès à la « régression logistique polytomique ou *multinomiale*[1] ». On notera que la procédure est utilisable dans le cas d'une variable dépendante binaire. Elle donne alors les mêmes résultats que la procédure **Logistic Binary**.

Figure 9.7

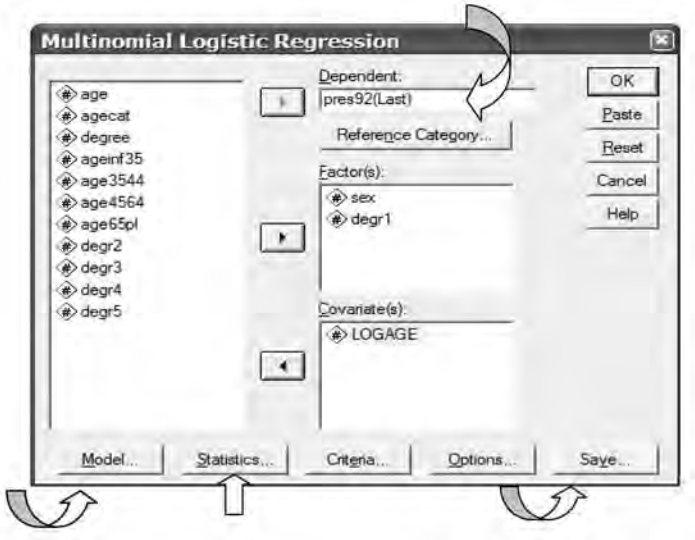
LA COMMANDE MULTINOMIAL LOGISTIC

En sollicitant **Multinomial Logistic**, on ouvre une nouvelle fenêtre qui permet de préciser la variable dépendante (*Pres92*) et les variables indépendantes prises en considération. Dans la case **FACTOR(S)** ont été insérées deux variables explicatives (*Sex* et *Degr1*). *Sex* identifie les hommes par le chiffre 1 et les femmes par le chiffre 2. *Degr1* identifie par le chiffre 1 les individus ayant atteint ou dépassé le *high school* ; les autres sont identifiés par 0. Dans la case **COVARIATE(S)** figure la variable quantitative *Logage,* qui donne le logarithme naturel de l'âge des individus figurant dans l'échantillon. Les flèches ajoutées à la fenêtre indiquent les clés à activer pour donner des instructions complémentaires.

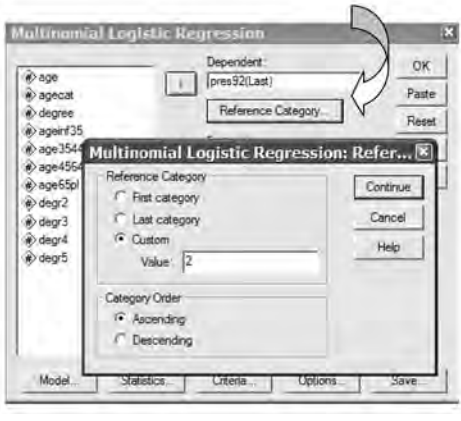
1. Le terme **Multinomial Logistic** utilisé par SPSS comporte une certaine ambiguïté, étant donné que « l'analyse multinomiale » renvoie généralement à l'analyse de l'impact conjoint de plusieurs variables explicatives sur une variable indépendante. La commande SPSS **Multinomial Logistic** concerne l'estimation des paramètres de la logistique polytomique**.**

Figure 9.8

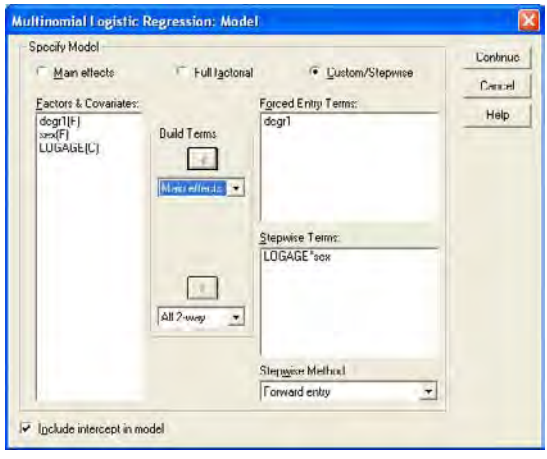
LA LOGISTIQUE POLYTOMIQUE : LA COMMANDE D'INSERTION DES VARIABLES



En cliquant sur **REFERENCE CATEGORY**, il est possible de préciser la catégorie de référence. Le chiffre 2 indiqué correspond à *Perot*.

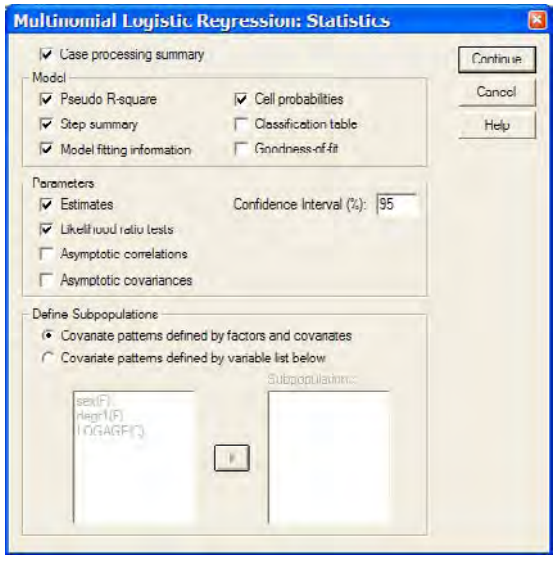


En activant la clé **MODEL**, on obtient la figure suivante :

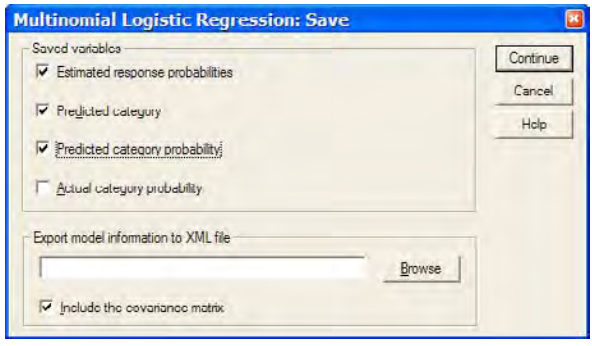


Il y est précisé qu'en plus de *Degr1, Logage* et *Sex* interviendront dans le modèle selon une relation interactive multiplicative, ce qui revient à capter séparément les hommes selon leur âge et à capter les femmes selon leur âge.

Le bouton **STATISTICS** permet de préciser les statistiques désirées :



Le bouton **SAVE** permet de récupérer les prévisions sur la base des observations correspondant aux données qui ont servi aux estimations, mais aussi des prévisions conditionnelles à des données supplémentaires fournies pour les variables indépendantes retenues dans le modèle.



2.3. *Les résultats des estimations*

Le listing fournit des renseignements parallèles à ceux obtenus dans le cas de la logistique binaire : des informations « omnibus », des renseignements spécifiques sur les coefficients estimés dans les diverses équations, des renseignements complémentaires permettant d'évaluer la qualité des estimations obtenues, des prévisions en termes de probabilité d'occurrence et en termes d'occurrence (identification de la catégorie prévue).

Les tests « omnibus »

Le khi-carré du modèle obtenu par la soustraction entre les –2LL vaut 96,142. Significatif au niveau 0,000, il laisse entendre qu'au moins une variable indépendante exerce une influence sur la variable dépendante. Les *pseudo-Rcarré* suggèrent néanmoins une performance limitée du modèle, comme c'est souvent le cas avec ce genre de modélisation et d'estimations.

Encadré 9.5

La logistique polytomique : les tests « omnibus »

Model Fitting Information

Model	Model Fitting Criteria -2 Log Likelihood	Likelihood Ratio Tests Chi-Square	df	Sig.
Intercept Only	1156,395			
Final	1059,983	96,412	6	,000

Pseudo R-Square

Cox and Snell	,051
Nagelkerke	,059
McFadden	,026

Estimation des paramètres

Les coefficients estimés concernent les logit des catégories Clinton et Bush, la catégorie Perot ayant été prise comme référence.

Encadré 9.6

LA LOGISTIQUE POLYTOMIQUE : LES COEFFICIENTS ESTIMÉS ET LES WALD

Parameter Estimates

VOTE FOR CLINTON, BUSH, PEROT[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Bush	Intercept	-3,417	,948	12,980	1	,000			
	[degr1=,00]	-,336	,336	1,015	1	,314	,713	,369	1,377
	[degr1=1,00]	0[b]	.	.	0	.	.	.	.
	[sex=1] * LOGAGE	1,182	,223	28,135	1	,000	3,262	2,107	5,048
	[sex=2] * LOGAGE	1,273	,225	31,992	1	,000	3,570	2,297	5,549
Clinton	Intercept	-2,614	,914	8,185	1	,004			
	[degr1=,00]	-,886	,321	7,639	1	,006	,412	,220	,773
	[degr1=1,00]	0[b]	.	.	0	.	.	.	.
	[sex=1] * LOGAGE	1,119	,216	26,919	1	,000	3,063	2,007	4,674
	[sex=2] * LOGAGE	1,320	,218	36,710	1	,000	3,744	2,443	5,739

a. The reference category is: Perot.

b. This parameter is set to zero because it is redundant.

Tous les coefficients estimés sont statistiquement significatifs à l'exception du coefficient affectant *Degr1* dans l'équation relative à la catégorie Bush. Comme il a déjà été souligné, l'interprétation du signe de ces coefficients est malaisée, parce que cette interprétation est tributaire de la catégorie qui sert de référence (dans le cas présent, *Perot*).

Encadré 9.7

***LA LOGISTIQUE POLYTOMIQUE : SIMULATIONS SUR LA VARIABLE* DEGR1**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Estimated Cell Probability for Response Category: 1	1847	,24	,35	,2833	,03582
Estimated Cell Probability for Response Category: 2	1847	,03	,22	,0873	,04095
Estimated Cell Probability for Response Category: 3	1847	,51	,72	,6295	,06158
Estimated Cell Probability for Response Category: 1	1847	,30	,46	,3675	,03742
Estimated Cell Probability for Response Category: 2	1847	,05	,35	,1558	,06518
Estimated Cell Probability for Response Category: 3	1847	,34	,58	,4766	,07032
Valid N (listwise)	1847				

2.4. PRÉVISIONS CONDITIONNELLES ET SIMULATIONS

Si, après estimation des paramètres, on suppose que toutes les personnes de l'échantillon sont de niveau *high school* ou plus (*Degr1* = 1 pour tous), la probabilité moyenne d'occurrence de la catégorie 3 (*Clinton*) se situe à 0,6295. Par contre, si on suppose que toutes les personnes constituant l'échantillon ont un niveau de formation inférieur au *high school* (*Degr1* = 0 pour tous), cette probabilité moyenne tombe à 0,4766. C'est dire combien les personnes plus scolarisées étaient en faveur de Clinton.

Les résultats de cette simulation ont été obtenus en recopiant les données initiales de *Logage* et *Sex* en bas de la dernière observation du fichier initial. Par ailleurs, en un premier temps, *Degr1* a été posé égal à 1 au-delà des observations initiales. Après avoir estimé le modèle et avoir obtenu les prévisions conditionnelles de probabilités en tenant compte des données ajoutées, la variable *Degr1* a été posée égale à 0 au-delà de la dernière observation du fichier initial, ce qui a permis de nouvelles prévisions de probabilités conditionnelles suite à l'estimation du modèle. Après sélection des prévisions conditionnelles situées au-delà du bloc initial d'observations, il a suffi ensuite de demander les moyennes des variables prévues.

Ci-dessous figurent deux extraits de fichier montrant l'organisation des données pour les simulations évoquées.

Encadré 9.8

LA LOGISTIQUE POLYTOMIQUE
SIMULATIONS : CONFIGURATION DES DONNÉES

Pres92	Age	Sex	Degr1	Logage	ID
1	40	1	0	3,69	1845
3	36	1	0	3,58	1846
1	33	2	0	3,5	1847
.	.	1	**1**	4,37	1848
.	.	1	**1**	3,47	1849
.	.	2	**1**	3,91	1850
.	.	2	**1**	4,03	1851
.	.	2	**1**	3,93	1852

Pres92	Age	Sex	Degr1	Logage	ID
1	40	1	0	3,69	1845
3	36	1	0	3,58	1846
1	33	2	0	3,5	1847
.	.	1	**0**	4,37	1848
.	.	1	**0**	3,47	1849
.	.	2	**0**	3,91	1850
.	.	2	**0**	4,03	1851
.	.	2	**0**	3,93	1852

Au-delà de *ID* = 1847, *Degr1* = 1 dans le premier tableau, *Degr1* = 0 dans le second tableau.

La séquence des commandes se présente comme suit en partant de l'agencement des données où *Degr1* = 1 au-delà de *ID* = 1847.

Analyze ; Regression ; Multinomial Regression…

SAVE… ESTIMATED RESPONSE PROBABILITIES…

Trois nouvelles variables sont ajoutées : *EST1_1, EST2_1, EST3_1*.

Modification de *Degr1* : *Degr1* = 0 au-delà de *ID* = 1847.

Analyze ; Regression ; Multinomial Regression…

SAVE… ESTIMATED RESPONSE PROBABILITIES.

Trois nouvelles variables sont ajoutées : *EST1_2, EST2_2, EST3_2*.

DATA ; SELECT CASES… *if ID>1847*

Analyze ; Descriptive Statistics ; Descriptives.

Dans la fenêtre **Variable(s)**,

faire figurer *EST1_1, EST2_1, EST3_1, EST1_2, EST2_2, EST3_2*

de manière à obtenir les moyennes.

3. VARIABLE DÉPENDANTE ORDINALE

Lorsque la variable dépendante est ordinale (être indifférent, satisfait, très satisfait par exemple), on pourrait penser à recourir à l'analyse de régression classique. La démarche serait par contre inadéquate du fait que l'ordre selon lequel s'ordonnent les catégories n'implique pas une distance univoque entre elles. Il est équivalent, par exemple de traduire l'ordre qui associe « être indifférent, satisfait, très satisfait » par les chiffres 1, 2, 3 ou par les chiffres 1, 3, 8.

Le recours à la régression logistique polynomiale (***Regression, Multinomial Logistic***) peut être envisagé, à condition toutefois de négliger l'ordre qui associe les catégories d'une variable ordinale.

3.1. REPÈRES THÉORIQUES

Les statisticiens ont préféré mettre au point une modélisation logistique adaptée au cas où la variable dépendante est ordinale. La modélisation recourt à une seule équation, mais en l'assortissant de repères complémentaires qui permettent d'obtenir, à la suite des estimations, des prédictions de probabilités d'occurrence des catégories et des prédictions d'occurrence de ces mêmes catégories.

La démarche recourt au calcul des valeurs *d'une variable latente instrumentale exprimant un score*. Une analogie permettra d'évoquer le sens de la démarche. Dans un passé relativement récent, les évaluations du niveau d'assimilation de contenus de cours en milieu universitaire se faisaient en termes quantitatifs. On est ensuite passé (en la simplifiant dans ce document) à une évaluation qualitative (A, B, C, D, E). Beaucoup d'enseignants ont continué leur habitude d'évaluer en termes quantitatifs et se sont donnés des règles pour traduire leurs résultats selon la grille qualitative ordinale désormais de règle. Prenons le cas fictif d'une évaluation quantitative dépendant d'exercices d'assimilation ($X1$), et de deux tests ($X2$ et $X3$). Le score Z sera calculé selon la formule $Z = \beta_1 X1 + \beta_2 X2 + \beta_3 X3$. Une fois disponible le score Z pour chaque étudiant, il faut établir des règles de traduction pour aboutir à l'évaluation qualitative. Par exemple, on décide qu'au-delà de la borne 90, il s'agit d'un A. Entre 80 et 90, il s'agit d'un B. À la limite inférieure, en dessous de 60, il s'agit d'un E. Le passage de l'évaluation quantitative à l'évaluation qualitative se fera à l'aide de bornes (4 bornes dans le cas des 5 catégories A, B, C, D, E).

Pour aboutir aux probabilités d'occurrence des catégories d'une variable qualitative ordinale, on adopte une démarche similaire. Dans un premier temps est définie une fonction « score » : $Z = \beta_1 X1 + \beta_2 X2 + \beta_3 X3$ … où $X1$, $X2$, $X3$ sont des variables indépendantes et où les bêta sont des paramètres à estimer. En même temps, des seuils sont à définir de manière à classer les scores obtenus ($n - 1$ seuils s'il y a n catégories).

Prenons le cas de trois catégories (A, B, C). Deux seuils seront définis, soit S1 et S2.

La probabilité d'occurrence pour A correspondra à Pr ($Z <$ S1).

La probabilité d'occurrence pour B, à Pr ($Z \geq$ S1 et $Z <$ S2).

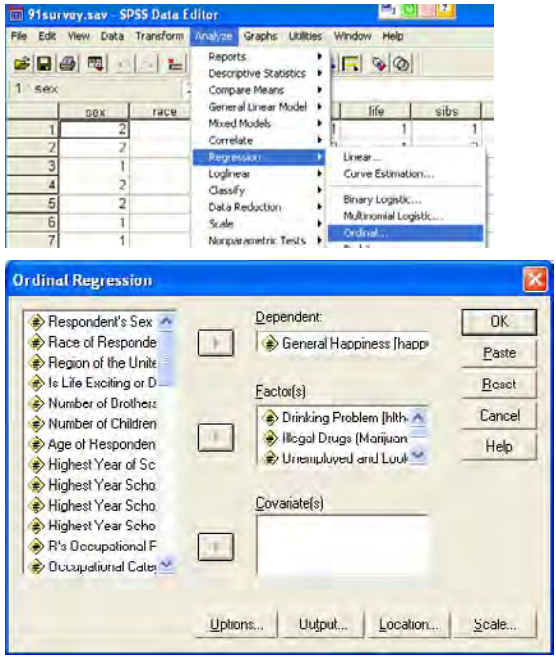
La probabilité d'occurrence pour C, à Pr ($Z \geq$ S2).

3.2. LA RÉGRESSION LOGISTIQUE ORDINALE AVEC SPSS

La figure suivante montre comment se fait l'accès à la « régression logistique ordinale ». Le fichier de données utilisé est un extrait du fichier 91SURVEY.SAV disponible dans le SPSS. La variable dépendante ordinale (*Happy*) comporte trois catégories selon le degré de satisfaction énoncé. Cette variable est confrontée à trois variables indépendantes qualitatives : problèmes de drogue, problèmes d'alcool, problèmes d'emploi (chômage). La procédure autorise aussi le recours à des variables indépendantes quantitatives.

Figure 9.9

LA COMMANDE POUR LA RÉGRESSION LOGISTIQUE ORDINALE



3.3. LES RÉSULTATS DES ESTIMATIONS

Tests « omnibus »

L'introduction des trois variables indépendantes se traduit par une réduction importante du (–2LL). Le khi-carré du modèle (81,249 – 48,887 = 32,363) est statistiquement très significatif (Sig. = 0,000), ce qui indique qu'au moins une variable indépendante exerce une influence sur la variable dépendante ordinale.

Encadré 9.9

LOGISTIQUE ORDINALE : TESTS OMNIBUS

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	81,249			
Final	48,887	32,363	3	,000

Link function: Logit.

Pseudo R-Square

Cox and Snell	,033
Nagelkerke	,039
McFadden	,018

Link function: Logit.

Tous les coefficients estimés sont statistiquement significatifs, ce qui permet d'écarter les hypothèses de non-influence pour chacune des variables indépendantes.

On remarquera que deux seuils estimés (–0,721 et 2,254) sont fournis à la manière de deux constantes. Ces seuils servent de bornes pour répartir les scores estimés (non donnés par le listing) selon des intervalles correspondant aux trois catégories de la variable indépendante (*Happy*).

Encadré 9.10

LOGISTIQUE ORDINALE : ESTIMATION DES PARAMÈTRES ET DES SEUILS

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[happy = 1]	-,721	,072	101,672	1	,000	-,862	-,581
	[happy = 2]	2,254	,111	410,192	1	,000	2,036	2,473
Location	[hlth4=1]	1,150	,501	5,267	1	,022	,168	2,132
	[hlth4=2]	0[a]	.	.	0	.	.	.
	[hlth5=1]	1,561	,379	16,965	1	,000	,818	2,303
	[hlth5=2]	0[a]	.	.	0	.	.	.
	[work1=1]	,658	,276	5,676	1	,017	,117	1,199
	[work1=2]	0[a]	.	.	0	.	.	.

Link function: Logit.

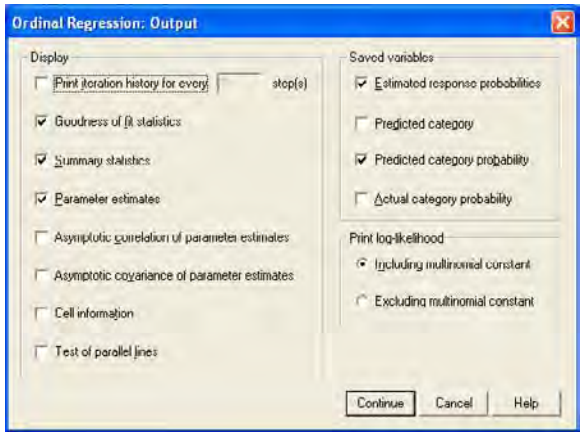
a. This parameter is set to zero because it is redundant.

3.4. LES PRÉVISIONS CONDITIONNELLES

La fenêtre ci-dessous a permis d'obtenir les prévisions des probabilités d'occurrence des trois catégories et les occurrences « prévues » compte tenu des probabilités d'occurrence les plus performantes.

Figure 9.10

LOGISTIQUE ORDINALE : L'OBTENTION DE PRÉVISIONS



Les résultats des « prévisions » sont consignés dans de nouvelles variables ajoutées au fichier de données. On remarquera toutefois qu'au stade actuel de la programmation de la procédure SPSS de traitement de la logistique ordinale, l'obtention de prévisions conditionnelles sur la base des estimations mais utilisant des données supplémentaires relatives aux variables indépendantes est difficilement accessible. Cette situation sera sans nul doute corrigée dans un avenir très prochain.

Épilogue

Voici terminé ce long périple à travers le foisonnement des méthodes de l'analyse multivariée. Chacune des méthodes de l'analyse des données correspond à une façon de construire la réalité, de saisir et d'interpréter cette réalité. L'utilisation de ces méthodes est souvent le fruit d'une longue histoire ; elle résulte à la fois des découvertes de la statistique fondamentale et appliquée, dont plusieurs remontent au XVII^{e} siècle, et des innovations technologiques récentes en informatique. Face à la profusion de données chiffrées de toutes sortes, il est important de disposer d'outils puissants et faciles à utiliser, qui permettent d'analyser en profondeur et de mieux comprendre l'évolution sociale et économique. Certains outils, dont l'analyse factorielle, l'analyse de variance, la régression « classique », la régression « logistique » et, plus généralement, l'ensemble des développements que proposent l'économétrie et la statistique appliquée sont largement utilisés dans certains secteurs de recherche (économie, biologie, environnement, médecine, etc.). Les sciences sociales y ont recours de plus en plus fréquemment. Il faut reconnaître qu'à première vue, « l'infrastructure mathématique » sous-jacente peut décourager certains utilisateurs. Par contre, les logiciels, y compris SPSS, ont progressivement rendu leur utilisation accessible. En pratique, le déploiement détaillé de l'articulation mathématique n'est pas nécessairement requis pour une

utilisation judicieuse. Il existe divers types de connaissance. L'utilisation adéquate des outils d'analyse multivariée requiert certes une connaissance de base de la logique des instruments utilisés, mais aussi (surtout ?) une connaissance pratique jumelée à un art de l'interprétation.

Bibliographie

AGRESTI, A. et B. FINLAY (1997). *Statistical Methods for the Social Sciences*, Upper Saddle River, Prentice Hall.

BACHELARD, G. (1967). *La formation de l'esprit scientifique*, Paris, Vrin.

BAKER, L.-R. (1995). *Explaining Attitudes*, Cambridge, Cambridge University Press.

BAILLARGEON, G. (2004). *Méthodes statistiques avec application en gestion, production, marketing, relations industrielleset comptables,* Trois-Rivières, SMG

BENZÉCRI, J.-P. *et al*. (1976). *L'analyse des données*. Tome I : *La taxinomie*. Tome II : *L'analyse des correspondances,* Paris, Dunod.

BESNIER, J.-M. (1996). *Les théories de la connaissance*, Paris, Flammarion, coll. « Dominos », nº 105.

BESSON, J.-L. (1992). « Les statistiques : vraies ou fausses », *Autrement*, nº 5.

BOUDON, R. (1971). *Les mathématiques en sociologie*, Paris, Presses universitaires de France.

BOUDON, R. (1990). *L'art de se persuader*, Paris, Fayard.

BOUDON, R. (1995). *Le juste et le vrai. Études sur l'objectivité des valeurs et de la connaissance*, Paris, Fayard.

BOUDON, R. et F. BOURRICAUD (1982). *Dictionnaire critique de la sociologie*, Paris, Presses universitaires de France.

BOURDIEU, P. (1972). « Les Doxosophes », *Minuit*, Éditions de Minuit, nº 1.

BOUROCHE, J.-M. et G. SAPORTA (1980). *L'analyse des données*, Paris, Presses universitaires de France.

BRIAN, E. (1998). « Du bon observateur au statisticien d'État », *Les cahiers de Science et Vie*, nº 48.

CALLON, M. et B. LATOUR(1991), *La science telle qu'elle se fait*, Paris, Éditions La Découverte.

CARMINES, E. et A. ZELLER(1979). *Reliability and Validity Assessment*, Newbury Park, Sage University Paper, nº 17.

CATELL, R. (1952). *Factor Analysis*, New York, Harper.

CATHELAT, B. (1990). *Socio-styles-système ; les styles de vie : théorie, méthodes et applications*, Paris, Les Éditions d'Organisation.

CHALMERS, A. (1991). *La fabrication de la science*, Paris, Éditions La Découverte.

CHANNOUF, A., J. PY et A. SOMAT (1996). « Prédire des comportements à partir des attitudes : nouvelles perspectives », dans J.-C. Deschamps et J.-L. Beauvois, *Des attitudes aux attributions*, Grenoble, Presses universitaires de Grenoble.

CLAUSEN, S.-E. (1998). *Applied Correspondence Analysis*, New York, Sage.

CRAUSER, J.-P., Y. HARVATOPOULOS et P. SAMIN (1989). *Guide pratique de l'analyse des données*, Paris, Les Éditions d'Organisation.

D'ASTOUS, A. (1993). *L'analyse des données issues d'une enquête*, Montréal, Guérin-Universitaire.

DE BATY, P. (1967). *La mesure des attitudes*, Paris, Presses universitaires de France.

DESROSIÈRES, A. (1993). *La politique des grands nombres. Histoire de la raison statistique*, Paris, Éditions La Découverte.

DOMETRIUS, N. (1992). *Social Statistics Using SPSS*, New York, Harper.

DROESBEKE, J.-J. et L. LEBART (2001). *Enquêtes, modèles et applications*, Paris, Dunod.

DUROZOI, G. et A. ROUSSEL (1987). *Dictionnaire de philosophie*, Paris, Nathan.

EAGLY, A. et CHAIKEN (1995). *The Psychology of Attitudes*, New York, Harcourt Brace Jovanovich.

ESCOFIER, B. et J. PAGES (1998). *Analyses factorielles simples et multiples*, Paris, Dunod.

FENNETEAU, H. et C. BIALÈS (1993). *Analyse statistique des données. Application et cas pour le marketing*, Paris, Ellipses.

FIELD, A. (2003). *Discovering Statistics Using SPSS for Windows*, Londres, Sage.

FOUREZ, G. (1992). *La construction des sciences*, Montréal, ERPI.

FOX, W. (1999). *Statistiques sociales*, Québec, Les Presses de l'Université Laval.

FREES, E. (1996). *Data Analysis Using Regression Models : The Business Perspective,* Upper Saddle River, Prentice Hall.

GEORGE, D. et P. MALLERY (1999). *SPSS for Windows Step by Step*, Boston, Allyn and Bacon.

GOURIÉROUX, Ch. (1990). *Écomométrie des modèles qualitatifs*, Paris, Economica.

GOURIÉROUX, Ch. (2000). *Econometrics of Qualitative Dependent Variables*, Cambridge, Cambridge University Press.

GREEN, B.S. et N.J. SALKIND (2004). *Using SPSS for Windows and Macintosh : Analysing and Understanding Data*, New York, Prentice Hall.

GROUPE CHADULE (1997). *Initiation aux pratiques statistiques en géographie*, Paris, Armand Colin.

HABERMAN, S.J. (1978). *Analysis of Qualitative Data*, New York, Academic Press.

HAIR, J., R. ANDERSON, R. TATHAM et W. BLACK ((1992). *Multivariate Data Analysis*, New York, Macmillan.

HENERSON, M., L. MORRIS et C. FITZ-GIBBON (1987). *How to Measure Attitudes*, Beverly Hills, Sage.

HOSMER, D.W. et S. LEMESHOW (1989). *Applied Logistic Regression*, New York, Wiley.

HOWELL, D. (1998). *Méthodes statistiques en sciences humaines*, Bruxelles, De Boeck Université.

HOWITT, D. et D. CRAMER (1997). *A Guide to Computing Statistics with SPSS for Windows*, New York, Prentice Hall.

JOURNET, N. (1996). « Comment peut-on être relativiste ? », *Sciences humaines*, nº 67.

JUDGE, G.G. (1982). *Theory and Practise of Econometrics*, New York, Wiley.

KENNEDY, P. (2003). *A Guide to Econometrics*, Cambridge, MA, MIT Press.

KIM, J.O. et C. MUELLER (1978). *Introduction to Factor Analysis*, Beverly Hills, Sage University Paper nº 13.

KINNEAR, P. et C. GRAY (2000). *SPSS for Windows Made Simple*, Hove (R.-U.), Psychology Press.

KLEINBAUM, D. (1994). *Logistic Regression. A Self-Learning Text*, New York, Springer

KLEINBAUM, D.G., L.L. KUPPER et K.E. MULLER (1988). *Applied Regression Analysis and Other Multivariable Methods*, Boston, PWS-Kent.

KUHN, T. (1972). *La structure des révolutions scientifiques*, Paris, Flammarion.

LAFORGE, H. (1981). *Analyse multivariée*, Saint-Laurent, Études Vivantes.

LAGARDE, J. (1995). *Initiation à l'analyse des données*, Paris, Dunod.

LAPLANTINE, F. (1996). *La description ethnographique*, Paris, Nathan.

LATOUR, B. (1995). *Le métier de chercheur : regard d'un anthropologue*, Paris, Éditions de l'INRA.

LENOIR, R. (1990). « Objet sociologique et problème social », dans P. Champagne, R. Lenoir, D. Merllié et L. Pinto, *Initiation à la pratique sociologique*, Paris, Dunod.

LÉVY-LEBLOND, J.-M. (1996). *La pierre de touche. La science à l'épreuve…*, Paris, Gallimard.

MARGUIN, J. (1994). *Histoire des instruments et des machines à calculer : trois siècles de mécanique pensante*, Paris, Hermann.

MENDENHALL, W. et T. SINCICH (1996). *A Second Course in Statistics : Regression Analysis*, Upper Saddle River, Prentice Hall.

MERLLIÉ, D. (1990). « La construction statistique », dans P. Champagne, R. Lenoir, D. Merllié et L. Pinto, *Initiation à la pratique sociologique*, Paris, Dunod.

MOSCAROLA, J. (1990). *Enquête et analyse des données*, Paris, Vuibert.

NORUSIS, M.J. (2005). *SPSS 13.0 Guide to Data Analysis*, Upper Saddle River, NJ, Prentice Hall.

NORUSIS, M.J. (2005), *SPSS 13.0 Statistical Procedure Companion*, Upper Saddle River, NJ, Prentice Hall.

NORUSSIS, M.J. (2005). *SPSS 13.0 Advanced Statistical Procedures Companion*, Upper Saddle River, NJ, Prentice Hall.

OSKAMP, S. (1991). *Attitudes and Opinions*, Englewood Cliffs, Prentice Hall.

OUELLET, F. et G. BAILLARGEON (2005). *Analyse de données avec SPSS*, *Version 12.0*, Trois-Rivières, SMG.

PAGANO, M. et K. GAUVREAU (2000). *Principles of Biostatistics*, Belmont, CA, Duxbury Press.

PAPILLON, V. (1993). *Vecteurs, matrices et nombres complexes*, Montréal, Modulo.

PAPILLON, V. et R. TURCOTTE (1981). *Probabilités et statistique*, Montréal, Modulo.

PÉTRY, F. (2003). *Guide pratique d'introduction à la régression en sciences sociales,* Québec, Les Presses de l'Université Laval.

PINTY, J.-J. et C. GAULTIER (1971). *Dictionnaire pratique de mathématiques et statistiques en sciences humaines*, Paris, Éditions Universitaires.

PLAISENT, M., P. BERNARD, C. ZUCCARO et N. DAGHFOUS (2004). *SPSS pour Windows 12.0 : guide d'utilisation,* Sainte-Foy, Presses de l'Université du Québec.

POPPER, K. (1990). *Le réalisme et la science*, Paris, Hermann.

RODEGHIER, M. (1996). *A Practical Guide to Survey Research Using SPSS. Surveys with Confidence,* Chicago, SPSS Inc.

ROSSI, J.-P. (1991). *La recherche en psychologie*, Paris, Dunod.

STAFFORD, J. (1996). *La recherche touristique. Introduction à la recherche quantitative par questionnaire*, Sainte-Foy, Presses de l'Université du Québec.

STAFFORD, J. et Br. SARRASIN (2005*). La prévision-prospective en gestion*, Sainte-Foy, Presses de l'Université du Québec.

TAPIA, C. et P. ROUSSAY (1991). *Les attitudes*, Paris, Éditions d'Organisation.

THOMSON, G. (1950). *L'analyse factorielle des aptitudes humaines,* Paris, Presses universitaires de France.

VAN DE GEER, J. (1971). *Introduction to the Multivariate Analysis for the Social Sciences*, San Francisco, W.H. Freeman.

VEDRINE, J.-P. (1991). *Le traitement des données en marketing*, Paris, Éditions d'Organisation.

VERBEEK, M. (2004). *A Guide to Modern Econometrics*, New York, Wiley.

VOLLE, M. (1980). *Le métier de statisticien*, Paris, Hachette.

WALLISER, B. et C. PROU (1988). *La science économique*, Paris, Éditions du Seuil.

WONNACOTT, R. et T. WONNACOTT (1979). *Econometrics*, New York, Wiley.

Tables statistiques

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Sainte-Foy, Québec G1V 2M2 • Tél.: (418) 657-4399 – www.puq.ca

Tiré de: *L'analyse multivariée avec SPSS,* Jean Stafford et Paul Bodson, ISBN 2-7605-1392-0 • D1392N

Table A

TABLEAU DE LECTURE DU KHI-DEUX (DISTRIBUTION DU KHI-DEUX SELON LA LOI DE K. PEARSON)

DL	P = 0,20	0,10	0,05	0,02	0,01
01	1,642	2,706	3,841	5,412	6,635
02	3,219	4,605	5,991	7,824	9,210
03	4,642	6,251	7,815	9,837	11,345
04	5,989	7,779	9,488	11,668	13,277
05	7,289	9,236	11,070	13,388	15,086
06	8,558	10,645	12,592	15,033	16,812
07	9,803	12,017	14,067	16,662	18,475
08	11,030	13,362	15,507	18,168	20,090
09	12,242	14,684	16,919	19,679	21,666
10	13,442	15,987	18,307	21,161	23,209
11	14,631	17,275	19,675	22,618	24,725
12	15,812	18,549	21,026	24,054	26,217
13	16,985	19,812	22,362	25,472	27,688
14	18,151	21,064	23,685	26,873	29,141
15	19,311	22,307	24,996	28,259	30,578
16	20,465	23,542	26,926	29,633	32,000
17	21,615	24,769	27,587	30,995	33,409
18	22,760	25,989	28,869	32,346	34,805
19	23,900	27,204	30,144	33,687	36,191
20	25,038	28,412	31,410	35,020	37,566
21	26,171	29,615	32,343	36,343	38,932
22	27,301	30,813	33,924	37,659	40,289
23	28,429	32,007	35,172	38,968	41,638
24	29,553	33,196	36,415	40,270	42,980
25	30,675	34,382	37,652	41,566	44,314
26	31,795	35,563	38,885	42,856	45,642
27	32,912	36,741	40,140	44,140	46,963
28	34,027	37,916	41,337	45,419	48,278
29	35,139	39,087	42,557	46,693	49,588
30	36,250	40,256	43,773	47,962	50,892

Table B

LES FACTEURS DE CORRECTION POUR LE COEFFICIENT DE CONTINGENCE C EN FONCTION DE LA TAILLE DES TABLEAUX ÉTUDIÉS

Taille du tableau	Facteur	Taille du tableau	**Facteur**	Taille du tableau	Facteur
2 × 2	0.707	3 × 9	0.843	6 × 6	0.913
2 × 3	0.685	3 × 10	0.846	6 × 7	0.930
2 × 4	0.730	4 × 4	0.866	6 × 8	0.936
2 × 5	0.752	4 × 5	0.863	6 × 9	0.941
2 × 6	0.765	4 × 6	0.877	6 × 10	0.945
2 × 7	0.774	4 × 7	0.888	7 × 7	0.926
2 × 8	0.779	4 × 8	0.893	7 × 8	0.947
2 × 9	0.783	4 × 9	0.898	7 × 9	0.952
2 × 10	0.786	4 × 10	0.901	7 × 10	0.955
3 × 3	0.816	5 × 5	0.894	8 × 8	0.935
3 × 4	0.786	5 × 6	0.904	8 × 9	0.957
3 × 5	0.810	5 × 7	0.915	8 × 10	0.961
3 × 6	0.824	5 × 8	0.920	9 × 9	0.943
3 × 7	0.833	5 × 9	0.925	9 × 10	0.966
3 × 8	0.838	5 × 10	0.929	10 × 10	0.949

Table C

TABLE DE LA LOI DE STUDENT

ν	P=0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
∞	0,12566	0,25335	0,38532	0,52440	0,67449	0,84162	1,03643	1,28155	1,64485	1,95996	2,32634	2,57582

Table D (1)

TABLE DE LA LOI DE FISHER

ν_2	$\nu_1 = 1$		$\nu_1 = 2$		$\nu_1 = 3$		$\nu_1 = 4$		$\nu_1 = 5$	
	$P = 0{,}05$	$P = 0{,}01$	$P = 0{,}05$	$P = 0{,}01$	$P = 0{,}05$	$P = 0{,}01$	$P = 0{,}05$	$P = 0{,}01$	$P = 0{,}05$	$P = 0{,}01$
1	161,4	4 052	199,5	4 999	215,7	5 403	224,6	5 625	230,2	5 764
2	18,51	98,49	19,00	99,00	19,16	99,17	19,25	99,25	19,30	99,30
3	10,13	34,12	9,55	30,81	9,28	29,46	9,12	28,71	9,01	28,24
4	7,71	21,20	6,94	18,00	6,59	16,69	6,39	15,98	6,26	15,52
5	6,61	16,26	5,79	13,27	5,41	12,06	5,19	11,39	5,05	10,97
6	5,99	13,74	5,14	10,91	4,76	9,78	4,53	9,15	4,39	8,75
7	5,59	12,25	4,74	9,55	4,35	8,45	4,12	7,85	3,97	7,45
8	5,32	11,26	4,46	8,65	4,07	7,59	3,84	7,01	3,69	6,63
9	5,12	10,56	4,26	8,02	3,86	6,99	3,63	6,42	3,48	6,06
10	4,96	10,04	4,10	7,56	3,71	6,55	3,48	5,99	3,33	5,64
11	4,84	9,65	3,98	7,20	3,59	6,22	3,36	5,67	3,20	5,32
12	4,75	9,33	3,88	6,93	3,49	5,95	3,26	5,41	3,11	5,06
13	4,67	9,07	3,80	6,70	3,41	5,74	3,18	5,20	3,02	4,86
14	4,60	8,86	3,74	6,51	3,34	5,56	3,11	5,03	2,96	4,69
15	4,54	8,68	3,68	6,36	3,29	5,42	3,06	4,89	2,90	4,56
16	4,49	8,53	3,63	6,23	3,24	5,29	3,01	4,77	2,85	4,44
17	4,45	8,40	3,59	6,11	3,20	5,18	2,96	4,67	2,81	4,34
18	4,41	8,28	3,55	6,01	3,16	5,09	2,93	4,58	2,77	4,25
19	4,38	8,18	3,52	5,93	3,13	5,01	2,90	4,50	2,74	4,17
20	4,35	8,10	3,49	5,85	3,10	4,94	2,87	4,43	2,71	4,10
21	4,32	8,02	3,47	5,78	3,07	4,87	2,84	4,37	2,68	4,04
22	4,30	7,94	3,44	5,72	3,05	4,82	2,82	4,31	2,66	3,99
23	4,28	7,88	3,42	5,66	3,03	4,76	2,80	4,26	2,64	3,94
24	4,26	7,82	3,40	5,61	3,01	4,72	2,78	4,22	2,62	3,90
25	4,24	7,77	3,38	5,57	2,99	4,68	2,76	4,18	2,60	3,86
26	4,22	7,72	3,37	5,53	2,98	4,64	2,74	4,14	2,59	3,82
27	4,21	7,68	3,35	5,49	2,96	4,60	2,73	4,11	2,57	3,78
28	4,20	7,64	3,34	5,45	2,95	4,57	2,71	4,07	2,56	3,75
29	4,18	7,60	3,33	5,42	2,93	4,54	2,70	4,04	2,54	3,73
30	4,17	7,56	3,32	5,39	2,92	4,51	2,69	4,02	2,53	3,70
40	4,08	7,31	3,23	5,18	2,84	4,31	2,61	3,83	2,45	3,51
60	4,00	7,08	3,15	4,98	2,76	4,13	2,52	3,65	2,37	3,34
120	3,92	6,85	3,07	4,79	2,68	3,95	2,45	3,48	2,29	3,17
∞	3,84	6,64	2,99	4,60	2,60	3,78	2,37	3,32	2,21	3,02

Table D (2)

TABLE DE LA LOI DE FISHER

ν_2	$\nu_1 = 6$		$\nu_1 = 8$		$\nu_1 = 12$		$\nu_1 = 24$		$\nu_1 = \infty$	
	$P = 0,05$	$P = 0,01$	$P = 0,05$	$P = 0,01$	$P = 0,05$	$P = 0,01$	$P = 0,05$	$P = 0,01$	$P = 0,05$	$P = 0,01$
1	234,0	5 859	238,9	5 981	243,9	6 106	249,0	6 234	254,3	6 366
2	19,33	99,33	19,37	99,36	19,41	99,42	19,45	99,46	19,50	99,50
3	8,94	27,91	8,84	27,49	8,74	27,05	8,64	26,60	8,53	26,12
4	6,16	15,21	6,04	14,80	5,91	14,37	5,77	13,93	5,63	13,46
5	4,95	10,67	4,82	10,27	4,68	9,89	4,53	9,47	4,36	9,02
6	4,28	8,47	4,15	8,10	4,00	7,72	3,84	7,31	3,67	6,88
7	3,87	7,19	3,73	6,84	3,57	6,47	3,41	6,07	3,23	5,65
8	3,58	6,37	3,44	6,03	3,28	5,67	3,12	5,28	2,93	4,86
9	3,37	5,80	3,23	5,47	3,07	5,11	2,90	4,73	2,71	4,31
10	3,22	5,39	3,07	5,06	2,91	4,71	2,74	4,33	2,54	3,91
11	3,09	5,07	2,95	4,74	2,79	4,40	2,61	4,02	2,40	3,60
12	3,00	4,82	2,85	4,50	2,69	4,16	2,50	3,78	2,30	3,36
13	2,92	4,62	2,77	4,30	2,60	3,96	2,42	3,59	2,21	3,16
14	2,85	4,46	2,70	4,14	2,53	3,80	2,35	3,43	2,13	3,00
15	2,79	4,32	2,64	4,00	2,48	3,67	2,29	3,29	2,07	2,87
16	2,74	4,20	2,59	3,89	2,42	3,55	2,24	3,18	2,01	2,75
17	2,70	4,10	2,55	3,79	2,38	3,45	2,19	3,08	1,96	2,65
18	2,66	4,01	2,51	3,71	2,34	3,37	2,15	3,00	1,92	2,57
19	2,63	3,94	2,48	3,63	2,31	3,30	2,11	2,92	1,88	2,49
20	2,60	3,87	2,45	3,56	2,28	3,23	2,08	2,86	1,84	2,42
21	2,57	3,81	2,42	3,51	2,25	3,17	2,05	2,80	1,81	2,36
22	2,55	3,76	2,40	3,45	2,23	3,12	2,03	2,75	1,78	2,31
23	2,53	3,71	2,38	3,41	2,20	3,07	2,00	2,70	1,76	2,26
24	2,51	3,67	2,36	3,36	2,18	3,03	1,98	2,66	1,73	2,21
25	2,49	3,63	2,34	3,32	2,16	2,99	1,96	2,62	1,71	2,17
26	2,47	3,59	2,32	3,29	2,15	2,96	1,95	2,58	1,69	2,13
27	2,46	3,56	2,30	3,26	2,13	2,93	1,93	2,55	1,67	2,10
28	2,44	3,53	2,29	3,23	2,12	2,90	1,91	2,52	1,65	2,06
29	2,43	3,50	2,28	3,20	2,10	2,87	1,90	2,49	1,64	2,03
30	2,42	3,47	2,27	3,17	2,09	2,84	1,89	2,47	1,62	2,01
40	2,34	3,29	2,18	2,99	2,00	2,66	1,79	2,29	1,51	1,80
60	2,25	3,12	2,10	2,82	1,92	2,50	1,70	2,12	1,39	1,60
120	2,17	2,96	2,01	2,66	1,83	2,34	1,61	1,95	1,25	1,38
∞	2,09	2,80	1,94	2,51	1,75	2,18	1,52	1,79	1,00	1,00

Table E

TABLE DE DURBIN-WATSON (À 95 %)

T	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,74	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78