Statistique Canada Statistics Canada

TECHNIQUES D'ENQUÊTE

juin 1984

volume 10

numéro 1

ÉDITION SPÉCIALE

analyse des données d'enquête — problèmes et méthodes

Une revue préparée par Statistique Canada

Canada

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010300475

SYMPOSIUM INTERNATIONAL SUR LA STATISTIQUE DES PETITES RÉGIONS

OTTAWA, CANADA

DU 22 AU 24 MAI 1985

Depuis quelques années, la demande d'informations fiables sur les petites régions a beaucoup augmenté dans de nombreux pays. Des chercheurs au service d'universités et de bureaux de la statistique ont effectué d'importants travaux d'ordre théorique et pratique. Le symposium international sera donc l'occasion pour présenter et discuter les conclusions de ces travaux ainsi que d'autres idées et points de vue.

Organisé par Statistique Canada, le Laboratoire de recherche en statistique et probabilité de l'Université Carleton et le Département de mathématiques et statistique de l'Université de Montréal, le symposium devrait rassembler plus de 200 personnes provenant de divers pays et organisations (universités, administrations, entreprises). Vingt conférenciers présenteront des communications sur les sujets suivants:

L'estimation synthétique
Les autres méthodes de modélisation pour l'estimation
Les méthodes démographiques
L'application aux domaines social et économique
Les divers aspects des politiques
Les pratiques organisationnelles

En outre, on prévoit des séances où ceux qui le désirent pourront présenter une communication dont le résumé devra avoir été soumis le 31 janvier 1985 au plus tard. Comme il est prévu de publier les actes du symposium peu après sa tenue, le texte final de la communication devra parvenir le 30 avril 1985 au plus tard. Voici la liste des membres du comité d'organisation:

R. Platek - Statistique Canada
J.N.K. Rao - Université Carleton
C.E. Sarndal - Université de Montréal
M.P. Singh - Statistique Canada

Veuillez envoyer vos résumés à l'attention de: R. Platek, Statistique Canada, 4-C7, Immeuble Jean-Talon, Parc Tunney, Ottawa (Ontario), K1A 0T6, Canada.

TECHNIQUES D'ENQUÊTE

Juin 1984 Vol. 10 No. 1

Édition spéciale

Une revue préparée par Statistique Canada

TABLE DES MATIÈRES

Préface..... 0

Sur le besoin de paramètres analytiques pour les échantillons complexes
LESLIE KISH..... 1

Introduction aux modèles linéaires et aux modèles linéaires généralisés
DAVID A. BINDER..... 10

L'ajustement des séries infra-annuelles aux répères annuels
PIERRE A. CHOLETTE..... 39

Analyse des dépenses consacrées à l'énergie
LOUISE A. HESLOP..... 54

Régression logistique et analyse de données de l'enquête sur la population active
S. KUMAR and J.N.K. RAO..... 68

Application de modèles linéaires et log-linéaires aux données d'échantillons complexes
ROBERT E. FAY..... 90

Application de la méthode des moindres carrées et de techniques connexes aux plans de sondage complexes
WAYNE A. FULLER.....107

Bibliographie sélective pour l'analyse des données d'enquêtes complexes.....131

8-3200-501
No. de références
Z - 079

ISSN: 0714-0045

TECHNIQUES D'ENQUÊTE

Juin 1984 Vol. 10 No. 1

Édition spéciale

Une revue préparée par Statistique Canada

Comité de rédaction:	R. Platek	- Président
	M.P. Singh	- Rédacteur en chef
	G.J.C. Hole	
	C. Patrick	
	P.F. Timmons	
	H. Lee	- Rédacteur adjoint

Politique de la rédaction:

La revue "Techniques d'enquête" veut donner aux personnes qui s'intéressent aux aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquêtes: les problèmes de conception causés par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages Statistique Canada, 4e étage, Édifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi.

PRÉFACE

Ce numéro renferme des communications présentées lors d'un colloque intitulé **Analyse de données d'enquête - aspects et méthodes** qui a eu lieu à Statistique Canada le jeudi 3 mai 1984.

Ce colloque a été parrainé par le Comité d'étude des méthodes de Statistique Canada et le Laboratoire de recherche en statistique et probabilités des universités Carleton et Ottawa. Il avait pour objet de démontrer comment des réalisations récentes en matière d'analyse de données provenant d'enquêtes complexes pourraient être appliquées aux études analytiques effectuées à Statistique Canada.

Il a débuté par des observations du Statisticien en chef, Martin B. Wilk, qui a souligné l'importance que Statistique Canada attache à l'amélioration de ses moyens de recherche et de développement et aux efforts communs des praticiens et des universitaires à ce sujet. Le colloque comprenait deux séances: une le matin, présidée par Leslie Kish de l'Institut pour la recherche sociale de l'Université de Michigan, au cours de laquelle des communications de Statistique Canada ont été présentées par D. Binder, P. Cholette, L. Heslop, et S. Kumar, en plus d'un aperçu concernant l'analyse présentée par le président.

La séance de l'après-midi, présidée par le sous-statisticien en chef, Ivan Fellegi, s'est ouverte par de brèves observations du président, et des communications présentées par R. Fay, U.S. Bureau of the Census et W. Fuller, de l'Université de l'État d'Iowa. La séance s'est terminée par une discussion générale des faits nouveaux concernant l'analyse de données, sous la direction de J.N.K. Rao, de l'Université Carleton. Plus de 200 représentants d'universités et de ministères et organismes fédéraux et provinciaux ont participé au colloque.

On trouvera également à la fin une bibliographie sélective sur le sujet établie par l'équipe de projet de l'analyse des données d'enquêtes complexes.

SUR LE BESOIN DE PARAMÈTRES ANALYTIQUES POUR LES ÉCHANTILLONS COMPLEXES[1]

Leslie Kish[2]

Mon propos ici est de souligner l'importance et l'urgence de découvrir des expressions algébriques utiles pour le calcul de paramètres analytiques adaptés aux plans de sondage complexes. Je voudrais m'adresser aux statisticiens mathématiques car ce problème devrait les intéresser puisqu'il en présente toutes les caractéristiques. c'est-à-dire qu'il est important. non encore résolu et soluble.

Il n'existe pas à l'heure actuelle d'analyse mathématique adéquate des problèmes les plus importants et les plus difficiles de la pratique de l'échantillonnage Les manuels portent presque exclusivement sur le calcul de bonnes estimations d'agrégats. de moyennes de variables et de moyennes de quotients. Il est parfois question de la différence entre deux valeurs d'un de ces paramètres. mais ce problème n'est abordé que de façon passagère et sporadique. C'est à cela que se limitent nos outils statistiques pour l'étude d'échantillons complexes.

Depuis la naissance de la théorie de l'échantillonnage. les méthodes probabilistes ont acquis progressivement un rôle privilégié dans le cours des travaux d'échantillonnage. et ces méthodes sont appliquées à des plans de sondage qui sont à la fois économiques et complexes. Le volume de données de haute qualité recueillies au moyen d'enquêtes n'a cessé d'augmenter. grâce à ces techniques. et les spécialistes veulent soumettre leurs données empiriques à des analyses de plus en plus compliquées. Toutefois on ne dispose pas de la théorie statistique mathématique requise pour faire des jugements valables. Pour utiliser les paramètres analytiques habituels. on doit supposer que les éléments choisis sont indépendants· mais cette indépendance est absente dans les plans de sondage complexes. Ainsi. un chercheur peut se

---

1 Exposé présenté au symposium.

2 Leslie Kish. Institute for Social Research. The University of Michigan

voir contraint à renoncer à une analyse qu'il considère indispensable. Par contre, s'il est trop impatient ou mal renseigné pour faire preuve d'une telle prudence, il utilisera peut-être les formules conçues pour l'échantillonnage aléatoire simple qu'il peut trouver dans les manuels de statistique, ce qui conduit souvent à des erreurs très graves.

J'espère que les statisticiens mathématiques se rendront compte de l'importance des problèmes non encore résolus dans la recherche de paramètres analytiques pour les données recueillies à partir de plans de sondage complexes. Cette lacune cause plus d'erreurs fondamentales que tout autre écart des hypothèses habituelles

Ces problèmes sont importants, ils n'ont pas encore été résolus et ils sont intéressants. Sommes-nous actuellement en mesure de proposer des solutions? Je crois que oui pour trois raisons. Premièrement, la théorie statistique a beaucoup progressé au cours des dernières années. Deuxièmement, l'augmentation rapide de la quantité d'ordinateurs électroniques et de leur qualité fait que le temps est propice pour la résolution de quelques-uns de ces problèmes. Troisièmement, on peut constater un nouvel intérêt à l'égard d'une méthode générale qui fait espérer des progrès rapides aboutissant à des approximations utiles. Au Survey Research Center, nous sommes en train d'élaborer cette méthode de calcul de la variance des coefficients de régression et d'autres paramètres pour lesquels des formules n'ont pas encore été mises au point.

Cette méthode me fait penser à la manière dont Alexandre a "résolu" le problème du noeud gordien. D'un point de vue théorique, je ne sais pas si cette méthode résout le problème ou le contourne. Mais, comme elle promet de bonnes approximations de variances qu'il est très important de mesurer, les praticiens devraient l'accueillir avec enthousiasme et intérêt. Cette méthode permet de calculer des estimations d'intervalles de confiance de certains paramètres analytiques pour lesquels il n'existe pas de formule actuellement.

Tout ce qui précède est tiré intégralement d'un exposé que j'ai présenté au cours d'une séance conjointe de l'American Statistical Association et l'Institute of Mathematical Statistics en 1957. La situation n'a guère évolué depuis. La technique sur laquelle nos espoirs de trancher le noeud gordien étaient fondés en 1957 est souvent utilisée aujourd'hui: c'est la méthode des duplications successives par blocs (balanced repeated replication) (Kish et Frankel 1970, 1974). Toutefois, mon discours passionné sur les lacunes de la

théorie des distributions de statistiques analytiques doublement complexes n'a pas mobilisé les enthousiasmes des statisticiens mathématiques. Je comprends maintenant pourquoi il en est ainsi. la sagesse étant le fruit de l'expérience. Premièrement, les problèmes que les statisticiens, comme les autres scientifiques. tentent de résoudre sont choisis non en fonction d'un critère de nécessité, mais en fonction d'un critère de faisabilité à un moment défini. (La mise au point de bombes nucléaires est un bon exemple.) Deuxièmement, les problèmes de la théorie des distributions de paramètres d'échantillons complexes semblent trop difficiles à surmonter. Troisièmement, les solutions comporteraient elles-mêmes trop de paramètres pour être utiles. Le point de vue que j'ai exprimé en 1978 (Kish 1978) et qui est encore le mien aujourd'hui est donc plus modéré: "Les nouvelles méthodes de calcul permettent d'obtenir des approximations des variances qui semblent satisfaisantes pour des besoins pratiques. Toutefois il serait préférable d'avoir une théorie mathématique des distributions de paramètres analytiques (tels que les coefficients de régression) qui est fondée non sur des hypothèses d'indépendance. mais sur des corrélations complexes entre les observations recueillies pour un échantillon. On peut oser espérer qu'un jour il y aura de nouvelles découvertes. mais les résultats ne seront pas universellement utilisables. à cause des complexités mathématiques. et surtout à cause du fait que le nombre de paramètres nécessaire sera trop grand pour les applications pratiques.

Je veux maintenant décrire sans ambages sept points importants au sujet des échantillons complexes. Ces points ne sont pas tous généralement connus ou acceptés. mais je vous demande d'y croire, de les mettre en pratique et de les enseigner, comme je le fais moi-même.

1. Les effets des plans de sondage complexes doivent être analysés <u>séparément</u> pour les estimations ponctuelles et les jugements de probabilité (comme. par exemple. les intervalles de confiance ou les tests d'hypothèses). Dans le cas des estimations ponctuelles, nous avons. pour tous les plans de sondage. une méthode cohérente pour évaluer les paramètres à l'aide des estimateurs semblables pondérés selon la probabilité (estimateurs H-T). En revanche. les jugements de probabilité tels que les intervalles de confiance sont très sensibles aux effets du plan de sondage. surtout dans l'échantillonnage en grappes.

a. "Les paramètres d'un échantillon (moyennes, coefficients de régression etc.) approchent la valeur de la population mère à mesure que la taille de l'échantillon augmente.

b. La convergence est généralement ralentie par les effets du plan.

c. Les effets du plan varient pour différents paramètres, pour différentes variables et pour différents plans de sondage" (Kish et Frankel 1974).

L'article cité plus haut présente également les arguments les plus convaincants à l'appui de ces propositions: la preuve en est d'ailleurs abondante (voir par exemple, Verma et coll. 1980). Néanmoins, deux statisticiens célèbres ont trahi complètement notre pensée dans des descriptions de notre article: "Les auteurs arrivent à la conclusion importante que les estimations d'intervalles de confiance d'un paramètre inconnu ne sont pas sensibles numériquement à la non-indépendance des observations introduite par les techniques d'enquête complexes telles que l'échantillonnage stratifié en grappes". Hélas, cette erreur est citée par d'autres théoriciens qui n'ont pas lu notre réponse à l'intention des praticiens des méthodes d'enquête: "Ces auteurs n'ont pas du tout compris le message principal que nous ne cessons de répéter: les estimations d'intervalles de confiance sont très sensibles numériquement au manque d'indépendance entre les observations dans les techniques d'enquête complexes telles que l'échantillonnage stratifié en grappes" (Kish et Frankel, 1974).

Le fait que cette erreur soit partagée tant par des non-statisticiens que par des théoriciens de l'échantillonnage pose des problèmes à nous, praticiens des techniques d'enquête. Pour cette raison, nous travaillons à l'heure actuelle à fournir une explication plus claire de notre point de vue.

2. Faut-il tenir compte des erreurs d'échantillonnage dans le calcul de paramètres analytiques à partir des données d'enquêtes complexes? Ou est-ce que quelques-uns d'entre nous se sont consacrés à la solution d'un problème négligeable, voire insignifiant? Je comprends mieux saint Sébastien quand les flèches d'une foule de païens enragés lui pleuvaient sur le corps (même un acuponcteur se plaindrait). Au premier rang, il y a les spécialistes des études de marché et des maisons de sondage qui ne nous écoutent pas, bien que certains aient appris à mettre un $\sqrt{\ }$ entre 2 et (pq/n). Au deuxième rang certains démographes écrivent qu'à cause de leurs grands échantillons et de leurs grandes erreurs de mesure, ils n'ont pas le temps de s'occuper des

erreurs d'échantillonnage. Au troisième rang, on retrouve les spécialistes en psychologie mathématique, en économétrie et en biométrie, qui, malheureusement, tirent leurs modèles linéaires directement de la statistique mathématique. Au quatrième rang, ce qui est encore plus malheureux, il y a les statisticiens mathématiques, qui ont tendance à oublier que "la fin ne justifie pas leurs moyennes," à invoquer l'hypothèse selon laquelle les observations sont indépendantes et distribuées identiquement ou à utiliser des formules d'exorcisme bayésiennes pour chasser les mauvais esprits hors du plan de sondage. Au cinquième rang, ce qui est le plus décourageant, on retrouve les théoriciens de l'échantillonnage qui construisent des théorèmes prouvant que, dans les modèles complètement spécifiés de superpopulations arbitraires, il n'est pas nécessaire de se soucier de la provenance et de la méthode de sélection des unités, ni de pondérer les observations pour tenir compte des inégalités des probabilités de sélection. Ces théoriciens ont même réussi à convaincre certains praticiens qu'ils peuvent rester sur l'Olympe avec leurs modèles sans jamais descendre sur terre, là où se trouve la population.

Ces remarques forcément brèves vous permettent de constater que je suis extrémiste pour plusieurs raisons: a) les effets du plan mesurés pour les paramètres analytiques révèlent généralement que les modèles des échantillons les mieux stratifiés sont mal définis: b) on observe souvent que les poids de sélection ont des effets sur les échantillons et c) il existe des relations entre les variables explicatives et les variables expliquées dans chaque individu, mais chaque individu fait partie de populations réelles et ces dernières ont des effets sur les plans de sondage (j'examine ces points en détail dans un ouvrage intitulé Statistical Design for Social Research, sur lequel je travaille actuellement et qui sera publié par Wiley en 1985.)

Ma philosophie est catégorique, mais, en pratique, je suis moins dogmatique. Je reconnais qu'en pratique: a) il n'est jamais possible de dénombrer complètement une population cible et, par conséquent, on doit toujours construire des modèles pour faire des inférences: b) l'échantillonnage probabiliste est trop coûteux et impraticable pour la plupart des expériences et c) malgré le manque de randomisation dans la sélection d'unités ou les analyses, on parvient souvent à des résultats fiables grâce à la prudence, à la répétition de l'échantillonnage, à la qualité du plan de sondage, à la propriété d'additivité et, aussi, un peu au jeu du hasard.

3. Les paramètres analytiques reposent sur la notion de sous-classes et de leur comparaison. Au cours des trois dernières décennies, beaucoup d'ouvrages et d'articles utiles ont été publiés à propos des variances et des effets du plan de sondage au niveau des sous-classes. On peut y trouver de grandes quantités de résultats empiriques, plusieurs observations pratiques basées sur ces résultats (Kish 1980, Kish et coll. 1976, Verma et coll. 1980) ainsi que des théories récentes (Rust 1984, chapitre 6).

a) On doit distinguer les domaines proprement dits des classes recoupées, qui sont plus fréquentes et constituent notre centre d'intérêt ici.

b) La probabilité de sélection demeure constante pour les classes recoupées, mais la taille des échantillons est très variable.

c) Les estimations de totaux et de moyennes calculées à partir d'échantillons complexes doivent être conservées sous la forme de quotients et d'expressions conditionnelles.

d) Les effets du plan au niveau des classes recoupées s'approchent d'une valeur presque égale à 1 proportionnellement à mesure que la taille des sous-classes par grappe primaire s'approche de 1. Ce modèle approximatif doit être utilisé avec prudence et sous certaines réserves, mais il est préférable à toutes les échappatoires respectables qu'on trouve habituellement au sujet des effets du plan et selon lesquelles ces effets sont simplement soit égaux à 1 ou à une autre constante, soit identiques pour l'ensemble de l'échantillon. Un modèle agrégé peut souvent s'avérer meilleur qu'une série de calculs séparés et très variables.

4. Dans les comparaisons de moyennes appariées, les effets du plan sont souvent supérieurs à 1, mais beaucoup moins élevés que la somme de la variance de ces deux moyennes. Ce phénomène est attribuable à des covariances positives (et donc à une sorte d'additivité) et il a été observé régulièrement et souvent dans des comparaisons de classes recoupées et des enquêtes périodiques (Kish 1965, 14.1: voir aussi les publications susmentionnées).

5. En ce qui concerne la production de paramètres analytiques, plusieurs méthodes exploitent les possibilités des ordinateurs électroniques. Les méthodes linéarisées tayloriennes (méthodes delta), qui comprennent des programmes de différentiation automatisée, la méthode des duplications successives par blocs (balanced repeated replication) et la technique du jackknife se sont toutes révélées des outils importants pour l'estimation de la variance et des

effets du plan pour les échantillons complexes (Kish et Frankel 1970 et 1974; Woodruff et Causey 1978). La méthode bootstrap pourrait également se répandre à l'avenir (Rao 1984).

Les paramètres analytiques comportent généralement des effets du plan supérieurs à 1 et ces résultats sont significatifs dans tous les sens de ce mot, mais ces effets du plan sont également moins élevés que ceux calculés pour les moyennes. Certaines régularités ont été mises en évidence dans les relations entre les effets du plan sur divers coefficients et les effets du plan sur les moyennes.

Pour nous aider à voir clair, il nous faut non seulement plus de travaux empiriques, mais aussi plus de résultats fondés sur la théorie de l'échantillonnage et l'analyse de modèles. Je dois avouer que je suis déçu que, depuis nos premiers travaux, nous n'avons pas vu paraître plus de publications sur la théorie et les modèles qui seraient directement applicables à l'inférence à partir de données réelles. Il est nécessaire d'étudier les fondements empiriques des effets du plan, mais pour satisfaire notre besoin intellectuel de comprendre, il nous faut aussi une théorie mieux étayée et de meilleurs modèles. Même nos besoins pratiques ne sont pas complètement satisfaits par les travaux empiriques sur les effets du plan parce que les résultats dépendent à la fois des variables, du type d'estimations calculées pour le plan de sondage utilisé et de la population mère dans laquelle les données ont été recueillies. Cette source de variation à quatre dimensions est trop complexe et ce qui nous manque, c'est une théorie qui permet d'élaborer des modèles qui simplifient la réalité.

6. L'analyse de données qualitatives est un domaine important qui s'étend rapidement et plusieurs travaux ont été faits sur l'application de ces méthodes aux données d'enquêtes complexes (Fay 1982; Landis et coll. 1982; Koch et coll. 1975). Ces travaux abordent également des questions d'analyse de la variance qui étaient le point de départ de quelques-uns des premiers modèles, mais n'ont pas été approfondies par la suite (Kempthorne et Wilk 1955; Tukey et Cornfield).

7. Quant à l'avenir, j'ai bon espoir que de nouveaux éléments de théorie donneront naissance à des applications pratiques, sauf dans deux domaines. Premièrement, la statistique mathématique n'a pas et ne pourra pas nous donner une théorie complète des distributions qui soit directement utilisable parce

qu'il y a trop de paramètres dans la double complexité des paramètres analytiques d'enquêtes complexes. Deuxièmement, les concepteurs de modèles sont incapables de faire disparaître cette complexité. Ils pourront toutefois nous aider à trouver des méthodes d'inférence plus exactes et compréhensives et améliorer l'utilisation et la présentation des paramètres analytiques d'enquêtes complexes.

BIBLIOGRAPHIE

[1] Fay, R. (1982). Contingency table analysis for complex sample designs: CPLX. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.

[2] Kish, L. (1957). Some unsolved problems of complex samples. Communication présentée à une séance conjointe de l'Amercian Statistical Association etde l'Institute for Mathematical Statistics.

[3] Kish, L., et Frankel, M.R. (1970). Balanced repeated replications for standard errors. JASA, 65, pp. 1071-1094.

[4] Kish, L., et Frankel, M.R. (1974). Inference from complex samples. JRSS (B), 36, pp. 1-74.

[5] Kish, L. (1980). Design and estimation for domains. The Statistician (Londres) 29, pp. 209-222.

[6] Kish, L., Groves, R.M, et Krotki (1976). Sampling errors for fertility. Occasional paper 17, Londres: World Fertility Surveys, 61 pages.

[7] Koch, G., Freeman, D., et Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, pp. 53-59.

[8] Landis, J.R., Lepkowski, J., Eklund, S., et Stehouwer, S. (1982). A statistical methodology for analyzing data from a complex sample survey. Vital and Health Statistics, Series 2 - no. 92. DHHS Publ. no. 82-1366. Public Health Service, Washington, U.S. Government Printing Office.

[9] Rao, J.N.K. (1984). Bootstrap inference with stratified samples. (non encore publié).

[10] Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys, The University of Michigan, thèse de Ph.D.

[11] Verma, V., Scott, C., et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. JRSS (a), 143, pp. 431-473.

INTRODUCTION AUX MODÈLES LINÉAIRES ET AUX MODÈLES LINÉAIRES GÉNÉRALISÉS

David A. Binder[1]

Cette étude décrit brièvement les modèles statistiques à une variable, les modèles de régression linéaire et les modèles linéaires généralisés. Des examples d'une analyse de variance à deux facteurs et d'une régression logistique sont présentés.

1. INTRODUCTION

Notre propos ici est de donner un aperçu général de certaines notions utilisées en statistique pour construire des modèles applicables à des données.

L'estimation de moyennes et de proportions à partir de données sur un échantillon prélevé dans une population est devenue une pratique courante. À la section 2, nous abordons brièvement cette notion et nous décrivons les méthodes d'estimation d'intervalles de confiance.

Les modèles de régression linéaire et d'analyse de la variance sont souvent utilisés pour simplifier des données multidimensionnelles à l'aide d'un schéma comprenant un petit nombre de paramètres. Ces techniques sont des outils puissants pour l'analyste qui veut approfondir sa compréhension d'un ensemble de données complexes. Nous examinons ces méthodes à la section 3.

Les notions utilisées dans les méthodes de régression linéaire peuvent être étendues à tout un éventail de problèmes, grâce aux modèles linéaires généralisés décrits par Nelder et Wedderburn (1972). Ce type d'extension est particulièrement utile quand la variable dépendante est qualitative au lieu de continue. La section 4 porte sur la structure de ces modèles.

Enfin, la section 5 offre une description très brève des diagnostics qui permettent de vérifier si un modèle est inadéquat et de déceler les multicollinéarités.

[1] David A. Binder, Division des méthodes d'enquête-institutions et agriculture, Statistique Canada.

2. MODÈLES À UNE VARIABLE

2.1 Modèles fondés sur la loi binomiale

Supposons que nous prélevons un échantillon dans une grande population et que nous observons une caractéristique de chaque unité choisie. Si la taille de l'échantillon est n, nous pouvons représenter les observations à l'aide de la notation $Y_1, Y_2, \ldots, Y_n$. Nous recueillons ces données parce que nous voulons faire des inférences sur la population à partir de cet échantillon. Par exemple, notre population pourrait être les personnes qui résident au Canada, et nos données peuvent être définies de la manière suivante:

$$Y_j = \begin{cases} 1 \text{ si la personne } j \text{ est née au Canada} \\ 0 \text{ si la personne } j \text{ est née à l'extérieur du Canada.} \end{cases}$$

Nous voulons nous servir de cet échantillon pour déduire des informations sur la proportion de la population qui est née au Canada.

Si un échantillon aléatoire simple de n = 5,000 résidents est prélevé et que la proportion réelle de personnes nées au Canada est p = 0.85, le nombre de personnes dans notre échantillon qui sont nées au Canada sera une variable aléatoire qui suit une loi binomiale exprimée par la fonction

$$f(y) = \binom{5000}{y}(.85)^y(.15)^{n-y};\ y = 0, 1, \ldots, 5000.$$

Comme nous savons que p = 0.85, nous pouvons fournir une description complète des propriétés de $Y = \Sigma Y_j$, le nombre total de personnes nées au Canada dans notre échantillon. Par contre, dans la plupart des analyses statistiques, nous ne connaissons pas toutes les caractéristiques de la population mère et nous devons utiliser notre échantillon pour faire des inférences au sujet de cette population. Supposons que nous ignorons la valeur de p dans notre exemple. On peut alors prédire que le nombre de personnes nées au Canada dans notre échantillon sera une variable binomiale aléatoire dont la distribution correspond à la fonction

$$f(y) = \binom{5000}{y}p^y(1-p)^{5000-y};\ y = 0, 1, \ldots, 5000.$$

L'estimateur habituel de p, pour ce genre de données, est $\hat{p} = \bar{Y} = \Sigma Y_i/5000$. On définit $s(\hat{p}) = \{\hat{p}(1-\hat{p})/(5000)\}^{\frac{1}{2}}$, où $s(\hat{p})$ est notre estimation de l'erreur-type de $\hat{p}$. Or, $\hat{p} \pm 1.96\ s(\hat{p})$ est un intervalle aléatoire qui contient la vraie valeur (inconnue) de p avec une probabilité de 95 %. Si nous remplaçons le chiffre 1.96 par une autre valeur, cet intervalle deviendrait plus petit ou plus grand et le niveau de confiance ne serait plus de 95 %. Pour déterminer le niveau de confiance, on doit consulter une table des probabilités associées à l'intervalle approprié sous la courbe de la loi normale réduite.

Nous venons d'illustrer l'application de la loi binomiale à un échantillon aléatoire simple prélevé dans une grande population. Dans cet exemple, toute inférence basée sur l'échantillon s'étend à la population mère. Toutefois, dans bien des cas, nous voulons faire des déductions sur d'autres populations qui, à notre avis, présentent des caractéristiques semblables. Par exemple, le nombre de décès au Canada dans un groupe d'âge et sexe au cours d'une année peut être conçu comme étant un résultat particulier provenant d'une distribution binomiale dans laquelle chaque individu a une probabilité égale de mourir et tous les décès sont essentiellement des événements indépendants. Si la probabilité de décéder est constante sur une période de quelques années, on peut utiliser le nombre de décès au cours d'une année pour estimer le nombre de décès au cours d'autres années, bien que les populations mères ne soient pas identiques. (Les compagnies d'assurance-vie et leurs actuaires utilisent ce type d'hypothèse dans leurs calculs.) Si chaque décès individuel est un événement indépendant, les hypothèses sur l'invariabilité de la probabilité de décéder peuvent être vérifiées à l'aide de la loi binomiale.

Il est important de souligner que les modèles linéaires généralisés décrits à la section 4 permettent de remplacer l'hypothèse de probabilités constantes pour tous les individus par un schéma dans lequel les probabilités peuvent être liées à d'autres facteurs tels que l'âge, le sexe, l'état de santé, la consommation de tabac, le poids, etc.

2.2 Modèles fondés sur la loi normale

Une des distributions importantes utilisées dans l'élaboration de modèles pour expliquer des données est la loi normale qui est exprimée par la fonction

$$f(y) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp\{-\frac{1}{2\sigma^2}(y - \mu)^2\}; \; -\infty < y < \infty.$$

La moyenne de la population est μ et c'est souvent ce paramètre qu'on veut estimer. La variance de la population est σ^2.

Si nous observons des données $Y_1, Y_2, \ldots, Y_n$ dans la population, l'estimateur habituel de μ est $\hat{\mu} = \bar{Y} = \Sigma Y_j/n$. L'estimateur de l'erreur-type de $\hat{\mu}$ est $s(\hat{\mu}) = s/n^{\frac{1}{2}}$, où

$$s^2 = \Sigma (Y_j - \hat{\mu})^2/(n - 1).$$

Comme dans le cas de la loi binomiale, l'intervalle de confiance de 95 % est $\hat{\mu} \pm 1.96s(\hat{\mu})$ pour les grands échantillons. Cet intervalle est aléatoire et a une probabilité de 95 % de couvrir la vraie valeur de μ. Quand un échantillon est petit ($n < 60$), on peut remplacer la valeur 1.96 par la valeur appropriée de la distribution du t de Student pour accroître l'exactitude des intervalles estimés. Pour obtenir différents niveaux de confiance, il suffit de substituer à la valeur 1.96 le chiffre qui correspond au percentile approprié de la loi normale réduite ou de la distribution du t de Student.

Dans certains cas, il n'est pas réaliste de supposer que la variance est constante, surtout dans les modèles linéaires décrits à la section 3. Une manière simple d'étendre l'application de la loi normale est de représenter la variance de X_i par σ_i^2, où $\sigma_i^2 = \sigma^2/w_i$. Nous supposons ici que $w_1, w_2, \ldots, w_n$ sont des poids connus. L'estimateur de la moyenne est alors $\hat{\mu} = \Sigma w_j Y_j/\Sigma w_j$, une moyenne pondérée des observations. L'erreur-type de cette moyenne est $s(\hat{\mu}) = s/(\Sigma w_j)^{\frac{1}{2}}$, où

$$s^2 = \Sigma w_j(Y_j - \hat{\mu})^2/(n - 1).$$

L'estimation d'intervalles de confiance pour μ est analogue au cas simple. Il est important de souligner que les poids $w_1, \ldots, w_n$ sont liés à la spécification des modèles basés sur la loi normale et n'ont habituellement aucun

rapport avec les poids d'échantillonnage utilisés dans les plans de sondage complexes élaborés pour des populations finies. Quand un analyste ajuste des modèles à des données recueillies à partir d'un plan de sondage complexe conçu pour une population finie, il peut décider d'inclure à la fois des poids théoriques et des poids d'échantillonnage dans le calcul d'une estimation.

2.3 Modèles appartenant à la famille exponentielle

Dans les modèles décrits plus haut, les lois binomiale et normale peuvent être considérées comme des cas spéciaux d'un ensemble beaucoup plus général de distributions connu sous le nom de "famille exponentielle". La forme générale que nous utiliserons pour exprimer cette distribution est:

$$f(y_j) = \exp[\kappa_j \{y_j\theta - b(\theta)\} + c(y_j, \kappa_j)],$$

où les valeurs de y_j ne dépendent pas de θ.

Nous supposons que $\kappa_j = \kappa w_j$, où les poids $w_1, \ldots, w_n$ sont connus. Dans un grand nombre de cas, la valeur de κ est également connue.

Exemple 1 (proportion binomiale)

Soit $\bar{y}_j = y_j/n_j$ une proportion calculée pour un échantillon de n_j observations d'une population binomiale. On peut donc écrire:

$$f(\bar{y}_j) = \binom{n_j}{n_j\bar{y}_j} p^{n_j\bar{y}_j}(1 - p)^{n_j(1 - \bar{y}_j)} : \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \ldots, 1,$$

$$E(\bar{y}_j) = p, \ Var(\bar{y}_j) = p(1 - p)/n_j,$$

$$\theta = \log[p/(1 - p)],$$

$$\kappa_j = n_j,$$

$$b(\theta) = \log(1 + e^{\theta}).$$

Exemple 2 (Variable normale)

Soit y_j une variable qui suit une loi normale et qui a pour moyenne μ et pour variance σ_j^2. Ainsi, on peut écrire:

$$f(y_j) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\frac{y_j - \mu}{\sigma_j})^2\}; \; -\infty < y_j < \infty$$

$$E(y_j) = \mu, \qquad Var(y_j) = \sigma_j^2,$$

$$\theta = \mu,$$

$$\kappa_j = 1/\sigma_j^2,$$

$$b(\theta) = \mu^2/2.$$

Exemple 3 (Moyenne d'une variable de Poisson)

Soit y_j une variable de Poisson qui a pour moyenne $n_j\lambda$. Définissons ensuite $\bar{y}_j = y_j/n_j$. On a alors:

$$f(\bar{y}_j) = e^{-n_j\lambda} (n_j\lambda)^{n_j\bar{y}_j}/(n_j\bar{y}_j)!; \qquad \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \ldots,$$

$$E(\bar{y}_j) = \lambda, \qquad Var(\bar{y}_j) = \lambda/n_j$$

$$\theta = \log \lambda,$$

$$\kappa_j = n_j,$$

$$b(\theta) = e^{\theta}.$$

Exemple 4 (χ^2)

Soit y_j une variable qui suit une loi de type $\sigma^2\chi^2_{\nu_j}/\nu_j$. Ce genre de variable est souvent utilisée dans les modèles d'analyse et de décomposition de la variance, où y_j est le carré moyen. On a alors:

$$f(y_j) = y_j^{(\nu_j - 2)/2} (\frac{\nu_j}{2\sigma^2})^{\nu_j/2} \exp\{-y_j\nu_j/(2\sigma^2)\}/\Gamma(\nu_j/2); \; y_j \geq 0,$$

$$E(y_j) = \sigma^2, \qquad Var(y_j) = 2\sigma^4/\nu_j,$$

$$\theta = -1/\sigma^2,$$

$$\kappa_j = \nu_j/2,$$

$$b(\theta) = -\log(-\theta).$$

Comme on peut le noter dans ces exemples, la famille exponentielle comprend un grand nombre de distributions fréquentes. En général:

$$E(y_j) = b'(\theta) = \mu, \qquad Var(y_j) = b''(\theta)/\kappa_j = V_j$$

où $b'(\cdot)$ et $b''(\cdot)$ représentent la dérivée première et la dérivée seconde de $b(\cdot)$.

Si les valeurs de $y_1, \ldots, y_n$ sont indépendantes, l'estimation du maximum de vraisemblance de θ provient de la solution de:

$$\hat{\mu} = \Sigma\, \kappa_j y_j / \Sigma\, \kappa_j = \Sigma\, w_j y_j / \Sigma\, w_j$$

où $\hat{\mu} = b'(\hat{\theta})$. Il doit donc exister une grande famille de modèles où une moyenne pondérée d'un échantillon constitue un estimateur efficace de la moyenne de la population. L'estimateur de la variance de $\hat{\mu}$ est

$$\hat{V}(\hat{\mu}) = (\Sigma\, \kappa_j^2 \hat{V}_j)/(\Sigma\, \kappa_j)^2$$

$$= b''(\hat{\theta})/(\Sigma\, \kappa_j).$$

Pour les grands échantillons, l'intervalle de confiance de 95 % de μ est $\mu \pm 1.96\,\{\hat{V}(\hat{\mu})\}^{\frac{1}{2}}$, pourvu que le modèle soit vrai.

Dans les cas où $\kappa_j = \kappa w_j$ est connu seulement jusqu'à la valeur de la constante de proportionnalité, κ, (comme, par exemple, dans un modèle fondé sur la loi normale), il est nécessaire d'estimer la valeur de κ. L'estimation du maximum de vraisemblance dépend de la solution à l'équation:

$$\Sigma\, w_j[y_j\theta - b(\theta) + \frac{\partial c(y_j, \kappa_j)}{\partial \kappa_j}] = 0.$$

L'autre moyen d'estimer $\hat{V}(\hat{\mu})$ sans biais est d'utiliser l'estimateur suivant qui est lié moins étroitement à un modèle:

$$\hat{V}_1(\hat{\mu}) = \frac{\Sigma\, w_j(y_j - \hat{\mu})^2}{(n-1)(\Sigma\, w_j)}.$$

Cet estimateur offre une autre façon de construire les intervalles de confiance de $\hat{\mu}$. La principale hypothèse sur laquelle repose la validité de cette méthode est que $Var(y_j) \propto 1/w_j$.

3. MODÈLES LINÉAIRES

3.1 Analyse de la variance à un facteur

Le modèle de l'analyse de la variance à un facteur constitue une extension simple des modèles à une variable basés sur la loi normale qui sont décrits à la section 2.2. Dans l'analyse de la variance, nous observons une caractéristique de chaque unité échantillonnée, mais nous définissons également des sous-populations. Ces sous-populations peuvent être des groupes d'âge et sexe, des groupes de secteurs d'activité/professions, etc. Le modèle peut être représenté par l'expression

$$y_{ij} = \mu_i + \varepsilon_{ij};\ i = 1, \ldots, I;\ j = 1, \ldots, n_i,$$

où les μ sont les moyennes réelles qui varient d'une sous-population à une autre, et, par hypothèse, les ε sont indépendants et suivent une loi normale avec une variance égale à $\sigma^2_{ij} = \sigma^2/w_{ij}$, où les w_{ij} sont des poids connus. Dans la plupart des analyses, ces poids sont constants.

L'estimateur habituel de μ_i dans ce modèle est

$$\hat{\mu}_i = \sum_j w_{ij}y_{ij}/\sum_j w_{ij}.$$

Selon les hypothèses de ce modèle, les moyennes estimées sont indépendantes, obéissent à une loi normale et $E(\hat{\mu}_i) = \mu_i$ et $Var(\hat{\mu}_i) = \sigma^2/\sum_j w_{ij}$. Ces propriétés permettent d'établir des intervalles de confiance pour les moyennes individuelles.

Une manière différente mais équivalente de formuler ce modèle est

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

où $\Sigma\Sigma\, w_{ij}\alpha_i = 0$ et

$$\mu = \Sigma\Sigma\, w_{ij}\mu_i/\Sigma\Sigma\, w_{ij}$$

$$\alpha_i = \mu_i - \mu.$$

L'extension de cette représentation est particulièrement utile pour les modèles d'analyse de la variance à deux facteurs ou plus, ce qui fait l'objet des sections 3.2 et 3.3. Un des problèmes les plus importants dans ce genre de modèle est de vérifier si toutes les moyennes sont égales. Autrement dit, on veut savoir si $\mu_1 = \mu_2 = \ldots = \mu_I$ ou $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$. Il existe des logiciels standard (SAS, SPSS, etc.) qui comprennent ces tests d'hypothèses dans leurs sous-programmes d'analyse de la variance (ANOVA). Une question connexe est: quelles sous-populations ont une moyenne égale s'il s'avère que les moyennes ne sont pas toutes égales? Quand on ne peut plus ajouter d'autres facteurs à l'analyse comme, par exemple, dans une analyse de la variance à deux facteurs, il faut alors procéder à une comparaison multiple de moyennes. Un grand nombre de logiciels statistiques permettent d'exécuter ce type d'analyse.

3.2 Analyse de la variance à deux facteurs

Les données présentées au tableau 1 sont tirées de l'enquête sur la fécondité menée en 1975 au Sri Lanka (voir Little: 1982). La moyenne qui figure

dans chaque case indique le nombre d'enfants mis au monde selon la durée du mariage et le niveau d'instruction.

Les moyennes des colonnes et des rangées semblent indiquer que le nombre moyen d'enfants augmente en fonction de la durée du mariage et diminue en fonction du niveau d'instruction. Le modèle d'analyse de la variance à deux facteurs peut être exprimé sous la forme suivante:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

où les valeurs de ε sont considérées comme étant indépendantes, distribuées selon une loi normale et possédant une variance égale à $\sigma^2_{ijk} = \sigma^2/w_{ijk}$. Les w sont des poids connus. Dans la plupart des analyses, ces poids sont constants. Pour estimer les paramètres de ce modèle, il est nécessaire de fixer des contraintes à ces paramètres afin d'éviter qu'ils ne soient pas uniques. Les conditions habituellement ajoutées au modèle sont:

$$\sum_i \sum_j \sum_k w_{ijk}\alpha_i = 0,$$

$$\sum_i \sum_j \sum_k w_{ijk}\beta_j = 0,$$

$$\sum_i \sum_k w_{ijk}\gamma_{ij} = 0,$$

$$\sum_j \sum_k w_{ijk}\gamma_{ij} = 0.$$

Les estimateurs sont définis par le système d'équations

$$\sum_i \sum_j \sum_k w_{ijk}(y_{ijk} - \hat{\mu}_{ij}) \frac{\partial \hat{\mu}_{ij}}{\partial \hat{\theta}_\ell} = 0$$

où $\hat{\theta}_1$. $\hat{\theta}_2$. ... correspondent aux estimations des paramètres $\hat{\mu}$, $\hat{\alpha}_i$, etc. Les valeurs des α et des β représentent les effets principaux, tandis que celles des γ sont les interactions entre les deux facteurs. On obtient ainsi les estimateurs suivants:

$$\hat{\mu} = \bar{y}_{...},$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} - \sum_j \sum_k w_{ijk}\hat{\beta}_j / \sum_j \sum_k w_{ijk},$$

$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} - \sum_i \sum_k w_{ijk}\hat{\alpha}_i / \sum_i \sum_k w_{ijk},$$

$$\gamma_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j.$$

où $\bar{y}_{ij.}$, $\bar{y}_{i..}$, etc. sont les moyennes pondérées appropriées.

Or, dans le modèle additif, $\mu_{ij} = \mu + \alpha_i + \beta_j$. Nous avons tracé à la figure 1 les courbes formées par les moyennes dans chaque case du tableau 1. Selon le modèle additif, toutes ces courbes doivent être parallèles. Si on ajuste le modèle additif aux données du tableau 1, on obtient les moyennes ajustées qui figurent au tableau 2. Comme nous pouvons le constater, l'effet du niveau d'instruction est considérablement affaibli après l'ajustement de ce modèle. La raison est que les femmes les plus instruites n'étaient pas mariées aussi longtemps que les autres, de sorte que c'est le nombre d'années depuis le premier mariage qui devient le facteur important. Toutefois, comme le révèle l'analyse de la variance résumée au tableau 3, tous les effets principaux et les interactions ont un caractère significatif. Le modèle additif doit donc être rejeté. Mais seulement 0.4% de la variation totale est expliquée par les interactions entre le niveau d'instruction et la durée du mariage, tandis que 49.7% de cette variation est expliquée par le modèle additif. Nous pouvons conclure que le modèle additif nous a permis de mieux comprendre les données et que l'effet du niveau d'instruction n'est pas aussi prononcé qu'il le semblait à l'origine.

3.3 Modèle de régression

Les modèles d'analyse de la variance décrits plus haut peuvent être considérés comme des cas spéciaux du modèle de régression linéaire multiple défini par l'équation

$$y_j = \beta_0 X_{0j} + \beta_1 X_{1j} + \ldots + \beta_r X_{rj} + \varepsilon_j,$$

où X_{0j}, X_{1j}, ... X_{rj} sont des constantes connues et β_0, β_1, ..., β_r sont des coefficients inconnus. Nous supposons que les ε sont indépendants, suivent une loi normale et ont pour variance $\sigma_j^2 = \sigma^2/w_j$, où les w sont des poids connus. Par exemple, dans une analyse de la variance à un facteur, on pourrait utiliser la notation suivante:

$X_{0j} = 1$ pour toutes les valeurs de j
$X_{ij} = 1$ si la $j^{ième}$ unité appartient à la $i^{ième}$ sous-population
$= - a_i/a_I$ si la $j^{ième}$ unité appartient à la I^{ième} sous-population
$= 0$ autrement

pour i = 1, ..., I - 1, où a_i est la somme des poids pour toutes les unités de la $i^{ième}$ sous-population. Dans ce modèle:

$$\mu_i = \beta_0 + \beta_i \quad \text{pour } i = 1, \ldots, I - 1,$$

$$\mu = \beta_0 - (a_1\beta_1 + \ldots + a_{I-1}\beta_{I-1})/a_I.$$

Ainsi, $\mu = \beta_0$ et $\alpha_i = \beta_i$ pour i = 1, .., I - 1.

Une formule de régression semblable peut également être définie pour les analyses à deux facteurs et plus.

Dans la version générale du modèle de régression, les estimateurs de β_0, ..., β_r sont $\hat{\beta}_0$, ..., $\hat{\beta}_r$, les solutions de l'expression

$$\Sigma\, w_j(y_j - \hat{y}_j)X_{ij} = 0, \qquad i = 0, 1, \ldots, r$$

où $\hat{y}_j = \hat{\beta}_0 X_{0j} + \hat{\beta}_1 X_{ij} + \ldots + \hat{\beta}_r X_{rj}$.

Si on veut vérifier des hypothèses, construire des modèles et établir des intervalles de confiance pour les β, il faut calculer la matrice des covariances des $\hat{\beta}$. On doit résoudre l'équation

$$\mathrm{Var}(\hat{\underset{\sim}{\beta}}) = \sigma^2 A^{-1}$$

où A est une matrice dont le (k,ℓ)ième élément est $\sum_j w_j X_{kj} X_{\ell j}$. Pour estimer σ^2, nous utilisons la formule $\hat{\sigma}^2 = \sum_j w_j (y_j - \hat{y}_j)^2/(n - r - 1)$.

Un grand nombre de logiciels comprennent des commandes faciles pour faire des tests d'hypothèses sur $\hat{\beta}$ à partir de la matrice des covariances estimées $\hat{\sigma}^2 A^{-1}$ et des valeurs critiques de la distribution appropriée de la variable F (comme, par exemple, PROC REG, PROC ANOVA et PROC GLM dans le logiciel SAS).

On peut illustrer ce genre de modèle à l'aide des données présentées au tableau 4, qui est tiré d'une étude de Koch, Gillings et Stokes (1980) et indique le nombre annuel de visites chez les médecins par habitant, aux États-Unis, en 1973 selon la taille de la ville (deux catégories: 1. RSMS = région statistique métropolitaine standard (Standard Metropolitan Statistical Area) et 2. Non-RSMS) et le niveau d'instruction (trois catégories). Ces données proviennent de la Health Interview Survey de 1973, enquête basée sur un échantillon probabiliste complexe. La figure 3 offre une représentation graphique de ces données.

Un modèle de régression et quelques tests statistiques ont permis d'obtenir un modèle réduit qui a la forme suivante:

$$E(Y_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j},$$

où X_{1j} = 1 si la $j^{\text{ième}}$ personne habite une RSMS
= 0 autrement

X_{2j} = 1 si le revenu de la famille de la $j^{\text{ième}}$ personne était inférieur à \$5,000 ou si la $j^{\text{ième}}$ personne avait plus de douze années d'instruction, dans le cas d'un chef de famille.
= 0 autrement.

Les paramètres estimés étaient $\hat{\beta}_0 = 4.18$ (erreur-type de 0.11), $\hat{\beta}_1 = 0.65$ (erreur-type de 0.11) et $\hat{\beta}_2 = 1.12$ (erreur-type de 0.09). Les erreurs-types calculées ici ne sont pas celles décrites plus haut parce que les auteurs ont utilisé la matrice de dimension 18 x 18 qui contient les covariances estimées de l'enquête susmentionnée, pour obtenir les erreurs-types. Cette méthode ne suppose pas l'indépendance des termes d'erreurs et elle est souvent utilisée dans l'analyse de données d'enquêtes complexes.

Les résultats sont résumés au tableau 5 et sont illustrés à la figure 4.

On peut constater que l'ajustement du modèle est assez bon. Nous avons réduit des données qui comprenaient dix-huit paramètres à trois résumés statistiques et nous obtenons également des erreurs-types moins élevées qu'auparavant.

4. MODÈLES LINÉAIRES GÉNÉRALISÉES

4.1 Régression avec une variable dépendante dichotomique

Un des problèmes qui se posent souvent dans l'utilisation des modèles linéaires décrits à la section 3 est que les erreurs sont supposées suivre une loi normale. Il est vrai que des analyses semblables à celles abordées à la section 3 sont faisables même si les erreurs n'ont pas une distribution normale, mais seulement si la variance des erreurs satisfait à la condition $\sigma_j^2 = \sigma^2/w_j$ et s'il n'y a pas de corrélation entre les erreurs. Dans un tel cas, les estimateurs linéaires que nous avons décrits produisent des estimations centrées à variance minimale pour les paramètres du modèle, mais il existe peut-être de meilleurs estimateurs (c'est-à-dire des estimateurs non linéaires). La recherche de meilleurs estimateurs a mené à la conception de modèles linéaires généralisés (voir Nelder et Wedderburn; 1972) et d'estimateurs robustes (voir Huber: 1973).

Par exemple, supposons que la variable dépendante y_j peut prendre seulement deux valeurs, 0 ou 1. Nous voulons construire un modèle pour exprimer $p_j = Pr(Y_j = 1)$ en fonction de l'expression linéaire $X_{0j}\beta_0 + X_{1j}\beta_1 + \ldots + X_{rj}\beta_r$. Trois méthodes sont généralement utilisées pour résoudre ce problème. La première approche est de calculer $\hat{\beta}_0, \ldots, \hat{\beta}_r$ à l'aide de l'estimateur habituellement employé dans un modèle de régression classique. Cette technique est analogue à une analyse discriminante dans laquelle les variables $X_{0j}, \ldots, X_{rj}$ sont considérées non comme des constantes fixes et connues, mais comme des variables aléatoires (qui suivent une loi normale à plusieurs caractères et ont une matrice des covariances constante) dont la moyenne dépend de la valeur de $\hat{Y}_j$. Le principal inconvénient de cette méthode est que $\hat{Y}_j = X_{0j}\hat{\beta}_0 + \ldots + X_{rj}\hat{\beta}_r$ ne peut pas être utilisé directement pour prédire la valeur de p_j. En outre, les X_{ij} sont souvent des variables qualitatives (telles que la province, la profession, etc.), ce qui annule l'hypothèse de normalité à plusieurs caractères.

Deux autres techniques très répandues sont l'analyse des probits et la régression logistique. Dans l'analyse des probits, on suppose que $p_j = \Phi(\sum_i X_{ij}\beta_i)$, où Φ est la fonction cumulative d'une variable aléatoire normale réduite. Dans la régression logistique, on suppose que

$$\theta_j = \log[p_j/(1 - p_j)] = \sum_i X_{ij}\beta_i.$$

Ces deux méthodes sont des outils analytiques importants, et on les retrouve dans un grand nombre de logiciels statistiques (tels que SAS et BMDP). Pour avoir une idée générale de ces deux approches, définissons

$$\eta_j = g(p_j) = \sum_i X_{ij}\beta_i.$$

Dans l'analyse des probits, $\eta_j = \Phi^{-1}(p_j)$, alors que dans la régression logistique, $\eta_j = \log[p_j/(1 - p_j)]$. Pour calculer l'estimateur du maximum de vraisemblance de $\beta_0, \ldots, \beta_r$, il faut résoudre l'équation

$$\sum_j \frac{(y_j - \hat{p}_j)X_{ij}}{\hat{p}_j(1 - \hat{p}_j)g'(\hat{p}_j)} = 0, \qquad \text{pour } i = 0, \ldots, r,$$

où $g(\hat{p}_j) = \sum_i X_{ij}\hat{\beta}_i$. Ce système doit souvent être résolu par itération. Pour l'analyse des probits, on peut écrire

$$g'(p_j) = \frac{1}{\phi[\Phi^{-1}(p_j)]}$$

où $\phi(\cdot)$ est la fonction de densité de la loi normale réduite. Dans le cas de la régression logistique,

$$g'(p_j) = [p_j(1 - p_j)]^{-1}$$

de sorte que l'estimation du paramètre recherché est contenue dans la solution de l'équation

$$\sum_j (y_i - \hat{p}_j)X_{ij} = 0, \qquad \text{pour } i = 0, \ldots, r.$$

La matrice des covariances de $\hat{\beta}_0, \ldots, \hat{\beta}_r$ est A^{-1}, où A est une matrice dont le $(k, \ell)^{\text{ième}}$ élément est

$$A_{k\ell} = \sum_j \frac{X_{kj}X_{\ell j}}{p_j(1 - p_j)\{q'(p_j)\}^2}$$

Ce résultat peut être utilisé pour établir des intervalles de confiance, faire des tests d'hypothèses et construire des modèles.

Dans la régression logistique, la covariance se ramène à l'expression

$$A_{k\ell} = \sum_j p_j(1 - p_j)X_{kj}X_{\ell j}.$$

Pour illustrer l'utilité de ces modèles, nous examinerons une analyse non publiée faite par Dolson et Morin à partir de données de l'enquête canadienne sur la santé et l'invalidité. La variable dépendante indiquait si une personne était classée comme handicapée possible après l'exécution du test de sélection 2 de l'enquête supplémentaire sur les personnes handicapées qui a été menée en même temps que l'enquête sur la population active de janvier 1983. (Pour une description détaillée de cette enquête, voir Dolson et Morin: 1983) L'analyse a été limitée aux hommes âgés de 15 à 64 ans. Parmi les 13,897 répondants, 14.4 % (résultat non pondéré) ont été sélectionnés. Comme on peut le voir au tableau 6, le taux de sélection a été ventilé en fonction du groupe d'âge, de la catégorie d'activité sur le marché du travail (deux catégories: actif et inactif) et d'une variable indiquant si un enquêté-substitut a répondu à la place d'une autre personne (trois catégories ont été établies: il n'y a pas eu d'enquêté-substitut, l'enquêté-substitut était un homme ou l'enquêté-substitut était une femme). (Le tableau 6 contient également les valeurs ajustées à partir du modèle décrit plus bas.) Ces données

sont illustrées à la figure 5.

Le modèle ajusté a permis de réduire le nombre de paramètres de trente à onze. Le modèle final était défini par l'équation

$$\log[p_{ijk}/(1 - p_{ijk})] = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij},$$

où $\Sigma\ \alpha_i = \Sigma\ \beta_j = \Sigma\ \gamma_k = 0$, $\sum_j \delta_{ij} = 0$, $\sum_i \delta_{ij} = 0$, pour la $i^{\text{ième}}$ groupe d'âge, la $j^{\text{ième}}$ catégorie d'activité sur le marché du travail et le $k^{\text{ième}}$ indice de recours à un enquêté-substitut (deux possibilités, selon qu'il a y eu ou qu'il n'y a pas eu d'enquêté-substitut). Les valeurs estimées des paramètres sont les suivantes.

Paramètre	Modalités	Estimation
μ		-1.43
α	15 à 24 ans	-1.12
	25 à 34 ans	-0.571
	35 à 44 ans	0.0143
	45 à 54 ans	0.629
	55 à 64 ans	1.05
β	Actifs	-0.576
	Inactifs	0.576
γ	- enquêté-substitut	0.0859
	+ enquêté-substitut	-0.0859
δ	actifs de 15 à 24 ans	0.385
	actifs de 25 à 34 ans	0.0938
	actifs de 35 à 44 ans	-0.175
	actifs de 45 à 54 ans	-0.243
	actifs de 55 à 64 ans	-0.0612
	inactifs de 15 à 24 ans	-0.385
	inactifs de 25 à 34 ans	-0.0938
	inactifs de 35 à 44 ans	0.175
	inactifs de 45 à 54 ans	0.243
	inactifs de 55 à 64 ans	0.0612

Les valeurs ajustées sont illustrées à la figure 6.

On peut voir que, même quand on tient compte de l'âge et de la catégorie d'activité, le recours à un enquêté-substitut a un effet sur les taux de sélection. Cet effet ne semble pas dépendre du sexe de l'enquête-substitut.

Par ailleurs, il n'y a pas d'interaction entre l'indice de recours à un enquêté-substitut et les variables âge et catégorie d'activité. Les résultats n'indiquent pas nécessairement la présence systématique d'un biais attribuable au recours à un enquêté-substitut, mais ils révèlent qu'un tel biais est possible. Sans une étude spéciale comme, par exemple, un programme de réinterview des enquêtés-substituts, il est impossible de conclure avec certitude que ce biais existe.

4.2 Modèles linéaires généralisés

Dans la section précédente, nous avons décrit un éventail de modèles linéaires fondés sur la loi binomiale et dont l'analyse des probits et la régression logistique constituent des cas spéciaux. Nous étendons maintenant ces modèles à la famille exponentielle selon l'analyse proposée par Nelder et Wedderburn (1972).

Comme à la section 2.3, nous supposons que y_j a une distribution exprimée par la fonction

$$f(y_j) = \exp[\kappa_j \{y_j \theta_j - b(\theta_j)\} + c(y_j, \kappa_j)],$$

où $\mu_j = E[Y_j] = b'(\theta_j)$ et $V_j = Var[Y_j] = b''(\theta_j)/\kappa_j$.

Nous définissons $\eta_j = g(\mu_j) = \sum_i X_{ij}\beta_i$ comme étant la composante linéaire du modèle, où $g(\cdot)$ est une fonction connue.

Pour obtenir les estimations du maximum de vraisemblance de $\underset{\sim}{\beta}$, il faut résoudre le système d'équations

$$\sum_j \frac{(y_j - \hat{\mu}_j)X_{ij}}{\hat{V}_j[g'(\hat{\mu}_j)]} = 0.$$

Nelder et Wedderburn (1972) ont démontré qu'un moyen raisonnable d'estimer $\underset{\sim}{\beta}$ est d'effectuer une série de régressions par la méthode des moindres carrés pondérés, où les valeurs des poids et des variables dépendantes sont révisées après chaque itération. Cette technique a été baptisée "méthode des moindres

carrés repondérés itérativement". En particulier, les poids à la $t^{\text{ième}}$ itération ont la valeur

$$\hat{w}_j^{(t)} = \frac{1}{\hat{V}_j^{(t)}[g'(\hat{\mu}_j^{(t)})]^2}$$

et on obtient la valeur des variables dépendantes à la $t^{\text{ième}}$ itération à partir de l'équation

$$\hat{Z}_j^{(t)} = g(\hat{\mu}_j^{(t)}) + g'(\hat{\mu}_j^{(t)})(y_j - \hat{\mu}_j^{(t)}).$$

Le résultat de la $(t + 1)^{\text{ième}}$ itération de $\hat{\underset{\sim}{\beta}}$ est donc la solution de l'équation

$$\sum_j \hat{w}_j^{(t)}[\hat{Z}_j^{(t)} - \sum_\ell X_{\ell j}\hat{\beta}_\ell^{(t+1)}]\, X_{kj} = 0.$$

La matrice des covariances estimées de $\hat{\underset{\sim}{\beta}}$ est A^{-1}, où le $(k, \ell)^{\text{ième}}$ élément de la matrice A est

$$A_{k\ell} = \sum_j \hat{w}_j X_{kj} X_{\ell j}.$$

On peut donc utiliser les programmes habituels de régression par la méthode des moindres carrés pour exécuter ces modèles linéaires généralisés.

Par exemple, dans la méthode très répandue d'analyse des tableaux de contingence à partir de modèles dits "log-linéaires", le point de départ est un modèle fondé sur la loi de Poisson, où $\log \mu_j = \sum_i X_{ij}\beta_i$. Dans ce modèle,

$$V_j = \mu_j,$$

$$g(\mu_j) = \log \mu_j,$$

de sorte qu'on obtient la solution des moindres carrés pondérés itérativement

en faisant les définitions suivantes:

$$\hat{w}_j^{(t)} = \hat{\mu}_j^{(t)},$$

$$\hat{Z}_j^{(t)} = \log \hat{\mu}_j^{(t)} + \frac{y_j - \hat{\mu}_j^{(t)}}{\hat{\mu}_j^{(t)}}.$$

Ainsi, des modèles semblables à ceux décrits à la section 3 peuvent être analysés de manière analogue dans le cadre d'un modèle linéaire généralisé.

5. DIAGNOSTICS

Les méthodes de régression linéaire sont maintenant connues depuis plus d'un siècle (voir Hocking (1983) pour un résumé des recherches faites au cours des vingt-cinq dernières années). Les travaux faits ces dernières années ont surtout porté sur les difficultés qui se posent lorsqu'il y a des multicollinéarités dans les variables (qui entraînent des variances élevées dans les estimations des paramètres) et lorsqu'un modèle est inadéquat. Quelques diagnostics qui permettent de déceler ces deux genres de problèmes sont maintenant offerts dans les logiciels SAS et SPSS-X.

Les méthodes décrites plus haut étendent la régression linéaire à une grande variété de problèmes. Les nouveaux diagnostics applicables à ce genre de modèle sont présentés dans un article de Landwehr, Pregibon et Shoemaker (1984).

Dans beaucoup d'analyses statistiques, le modèle proposé n'est qu'une approximation de la réalité. Par conséquent, l'utilisateur de ces modèles doit prendre en compte ces outils diagnostiques au cours de son analyse.

BIBLIOGRAPHIE

[1] Dolson, D. et Morin, J.-P. (1983). Disability data development project: Analysis of screening questionnaires. Document technique de la Division de la santé, Statistique Canada.

[2] Hocking, R.R. (1983). Developments in linear regression methodology: 1959-1982 (comprend une analyse). Technometrics 25, pp. 219-249.

[3] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, pp. 799-821.

[4] Koch, G.G., Gillings, D.B. et Stokes, M.E. (1980). Biostatistical implications of design, sampling and measurement to health science data analysis. Ann. Rev. Public Health 1, pp. 163-225.

[5] Landwehr, J.M., Pregibon, D. et Shoemaker, A. (1984). Graphical methods for assessing logistic regression models (comprend une analyse). J. Amer. Statist. Assoc. 79, pp. 61-83.

[6] Little. R.J.A. (1982). Direct standardization: A tool for teaching linear models for unbalanced data. Amer. Statist. 36, pp. 38-43.

[7] Nelder, J.A. et Wedderburn, R.W.M. (1972). Generalized linear models. J. Roy. Statist. Soc., Ser. A, 135, pp. 370-384.

Tableau 1: Nombre moyen d'enfants nés, selon la durée du mariage et le niveau d'instruction. Sri Lanka, 1975 (source: Little: 1982)

Nombre d'années depuis le premier mariage		Niveau d'instruction (années d'études)				
		0	1 à 5	6 à 9	10 et plus	Rangées
0 à 4	Moyenne	0.96	0.88	0.95	0.92	0.92
	Total	112	376	442	351	1281
5 à 9	Moyenne	2.54	2.46	2.39	2.39	2.44
	Total	172	442	362	255	1231
10 à 14	Moyenne	3.87	3.91	3.73	3.14	3.76
	Total	197	482	293	145	1117
15 à 19	Moyenne	5.13	4.97	4.61	4.13	4.84
	Total	239	461	262	95	1057
20 à 24	Moyenne	6.22	5.87	5.22	4.47	5.79
	Total	292	377	184	40	893
25 et plus	Moyenne	6.92	6.55	6.23	5.97	6.65
	Total	501	548	161	22	1232
Colonnes	Moyenne	5.17	4.24	3.26	2.30	3.94
	Total	1513	2686	1704	908	6811

Tableau 2: Interactions influant sur le nombre moyen d'enfants (données du tableau 1)

Nombre d'années depuis le premier mariage		Niveau d'instruction (années d'études)				
		0	1 à 5	6 à 9	10 et plus	Rangées
0 à 4	Moyenne brute	0.96	0.88	0.95	0.92	0.92
	Moy. ajustée	1.31	1.07	0.86	0.71	1.02
	Interaction	-0.35	-0.19	0.09	0.21	
5 à 9	Moyenne brute	2.54	2.46	2.39	2.39	2.44
	Moy. ajustée	2.78	2.54	2.33	2.18	2.49
	Interaction	-0.24	-0.08	0.06	0.21	
10 à 14	Moyenne brute	3.87	3.91	3.73	3.14	3.76
	Moy. ajustée	4.06	3.82	3.61	3.46	3.77
	Interaction	-0.19	0.09	0.12	-0.32	
15 à 19	Moyenne brute	5.13	4.97	4.61	4.13	4.84
	Moy. ajustée	5.11	4.87	4.66	4.51	4.82
	Interaction	0.02	0.10	-0.05	-0.38	
20 à 24	Moyenne brute	6.22	5.87	5.22	4.47	5.79
	Moy. ajustée	6.01	5.77	5.56	5.41	5.72
	Interaction	0.21	0.10	-0.34	-0.94	
25 et plus	Moyenne brute	6.92	6.55	6.23	5.97	6.65
	Moy. ajustée	6.82	6.58	6.37	6.22	6.53
	Interaction	0.10	-0.03	-0.14	-0.25	
Colonnes		5.17	4.24	3.26	2.30	3.94
		4.23	3.99	3.78	3.63	3.94

Tableau 3: Analyse de la variance des données du tableau 1

Source	Somme des carrés (SC)	Proportion du total des SC	Degrés de liberté	Carré moyen	F	Niveau de signification
Effets principaux						
Durée du mariage	27402.684	0.493	5	5480.537	1340.990	.000
Niveau d'instr.	225.535	0.004	3	75.178	18.395	.000
Interactions						
Durée du mariage x niveau d'instr.	206.965	0.004	15	13.798	3.376	.000
Variation résiduelle	27729.848	0.499	6787	4.986		
Total	55565.031		6810			

Tableau 4: Nombre annuel de visites chez un médecin par habitant selon la taille de la localité, le revenu de la famille et le niveau d'instruction du chef de famille, États-Unis, 1973

Niveau d'instruction (années d'études)	Revenu de la famille		
	0 à 4,999	5,000 à 14,999	15,000 ou plus
RSMS			
Moins de 12 ans	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
12 ans	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
Plus de 12 ans	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
NON-RSMS			
Moins de 12 ans	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
12 ans	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
Plus de 12 ans	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)

Note: Les chiffres entre parenthèses indiquent l'erreur-type des estimations.

Tableau 5: Nombre estimé de visites chez un médecin, données du tableau 4 et valeurs ajustées

Niveau d'instruction (années d'études)		Revenu de la famille		
		0 à 4,999	5,000 à 14,999	15,000 ou plus
RSMS				
Moins de 12 ans	valeur orig.	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
	valeur ajustée	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Écart	0.20	-0.10	-0.01
12 ans	valeur orig.	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
	valeur ajustée	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Écart	0.22	0.15	-0.13
Plus de 12 ans	valeur orig.	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
	valeur ajustée	5.95 (0.07)	5.95 (0.07)	5.95 (0.07)
	Écart	0.36	0.13	-0.29
NON-RSMS				
Moins de 12 ans	valeur orig.	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
	valeur ajustée	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Écart	-0.22	-0.04	0.24
12 ans	valeur orig.	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
	valeur ajustée	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Écart	0.06	0.14	0.31
Plus de 12 ans	valeur orig.	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)
	valeur ajustée	5.30 (0.11)	5.30 (0.11)	5.30 (0.11)
	Écart	-0.72	-0.24	-0.82

Tableau 6: Valeurs brutes et ajustées des taux de sélection après le test 2 de l'enquête canadienne sur la santé et les personnes handicapées, hommes âgés de 15 à 64 ans, selon la catégorie d'activité sur le marché du travail et l'indice de recours à un enquêté-substitut, Canada, janvier 1983 (résultats non pondérés)

Âge		Aucun enquêté-substitut	Enquêté-substitut: homme	Enquêté-substitut: femme
		Actifs		
15 - 24	valeurs brutes	.065(.0067)	.055(.0143)	.056(.0069)
	val. ajustées	.065(.0051)	.056(.0044)	.056(.0044)
	Écart	.000	-.001	.000
25 - 34	valeurs brutes	.085(.0058)	.058(.0252)	.069(.0069)
	val. ajustées	.085(.0048)	.071(.0046)	.071(.0046)
	Écart	.000	-.013	-.002
35 - 44	valeurs brutes	.113(.0079)	.029(.0290)	.094(.0086)
	val. ajustées	.111(.0064)	.093(.0059)	.093(.0059)
	Écart	.002	-.064	.001
45 - 54	valeurs brutes	.180(.0109)	.082(.0351)	.154(.0120)
	val. ajustées	.177(.0088)	.153(.0083)	.153(.0083)
	Écart	.003	-.071	.001
55 - 64	valeurs brutes	.284(.0150)	.207(.0752)	.250(.0183)
	val. ajustées	.283(.0124)	.249(.0124)	.249(.0124)
	Écart	.001	-.042	.001
		Inactifs		
15 - 24	valeurs brutes	.104(.0127)	.071(.0190)	.074(.0084)
	val. ajustées	.104(.0078)	.079(.0065)	.079(.0065)
	Écart	.000	-.008	-.005
25 - 34	valeurs brutes	.146(.0239)	.367(.1450)	.227(.0365)
	val. ajustées	.192(.0213)	.167(.0194)	.167(.0194)
	Écart	-.046	.200	.060
35 - 44	valeurs brutes	.348(.0372)	.455(.1501)	.324(.0544)
	val. ajustées	.359(.0309)	.320(.0299)	.320(.0299)
	Écart	-.011	.135	.004
45 - 54	valeurs brutes	.534(.0361)	.625(.1712)	.454(.0505)
	val. ajustées	.525(.0293)	.483(.0301)	.483(.0301)
	Écart	.009	.142	-.029
55 - 64	valeurs brutes	.571(.0220)	.563(.1240)	.591(.0420)
	val. ajustées	.585(.0194)	.543(.0217)	.543(.0217)
	Écart	-.014	.020	.048

NOTE: Les chiffres entre parenthèses sont les erreurs-types.

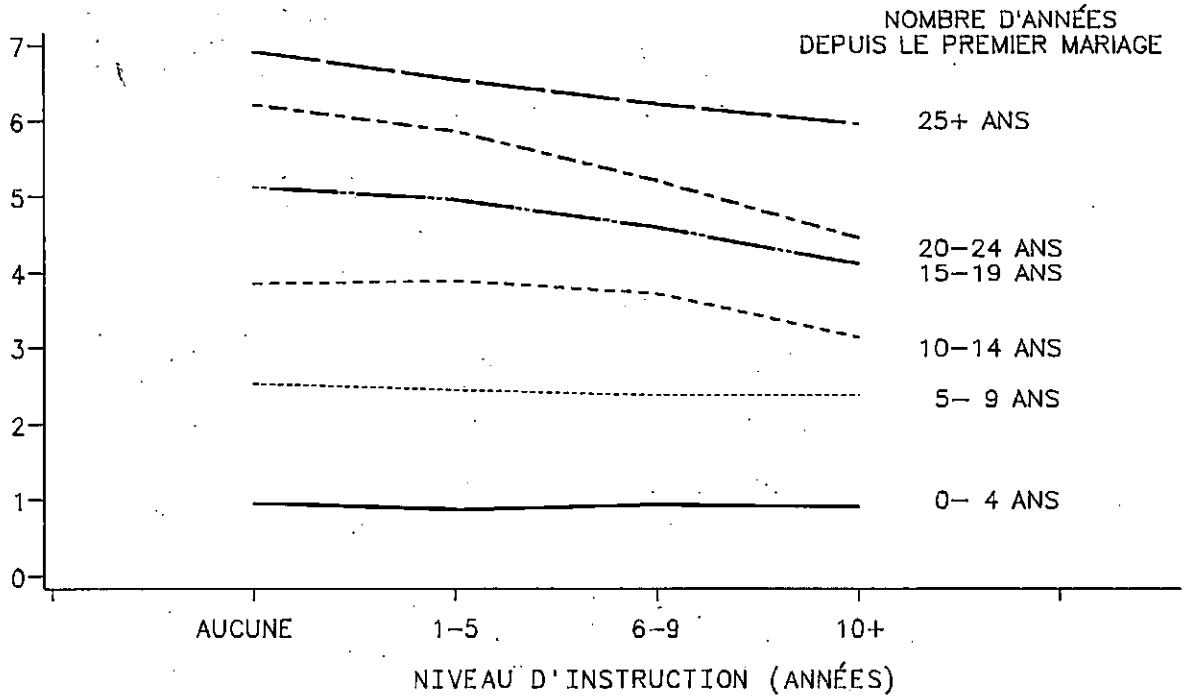

Figure 1: Moyennes observées de l'enquête sur la fécondité au Sri Lanka, 1975. Source de données: Little (1982).

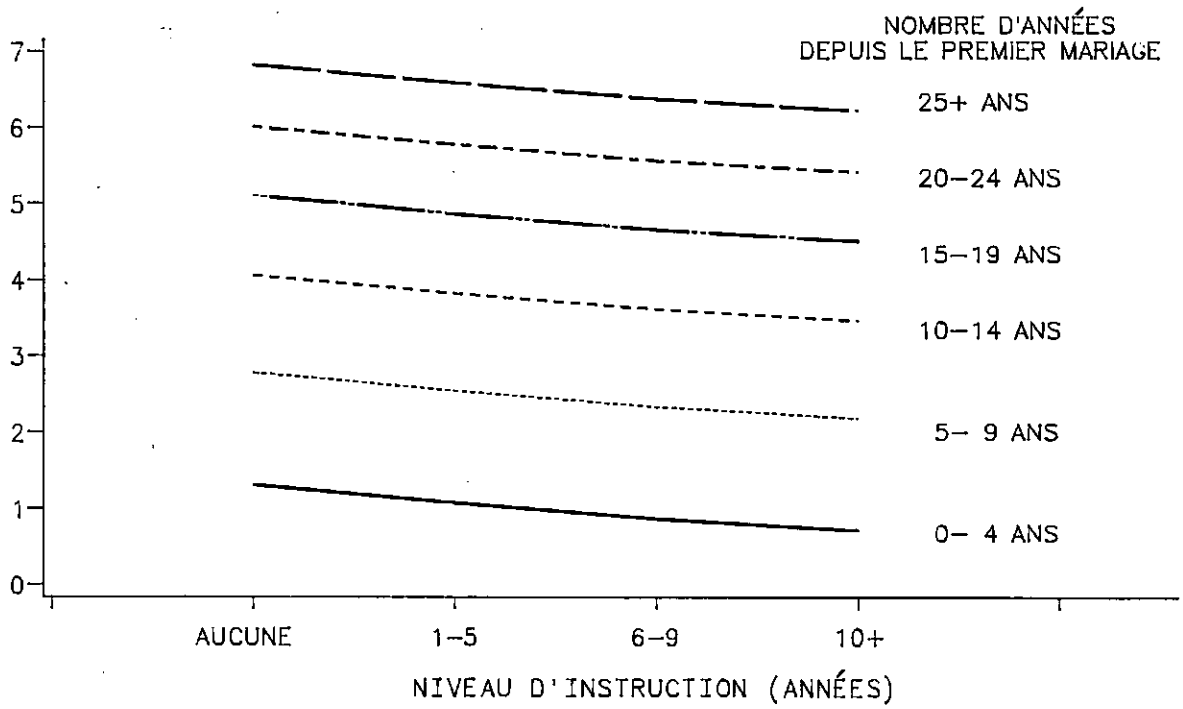

Figure 2: Moyennes ajustées de l'enquête sur la fécondité au Sri Lanka, 1975. Source de données: Little (1982).

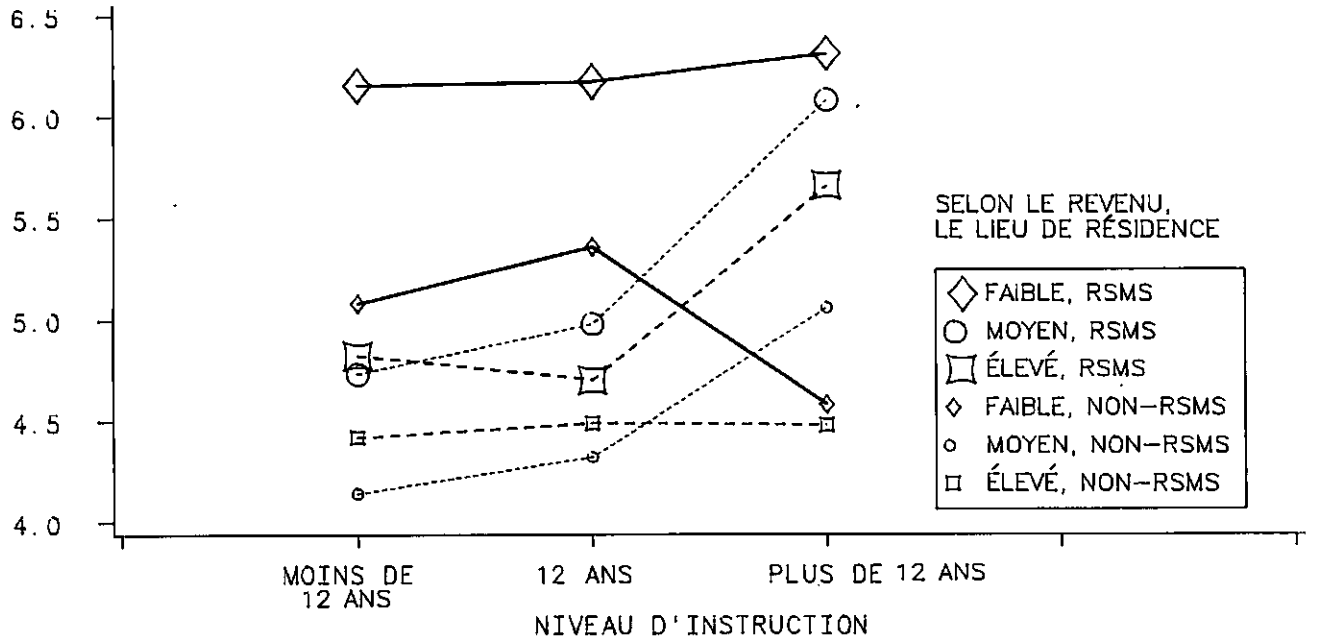

Figure 3: Moyenne observée du nombre annuel de visites chez les médecins par habitant, États-Unis 1973.

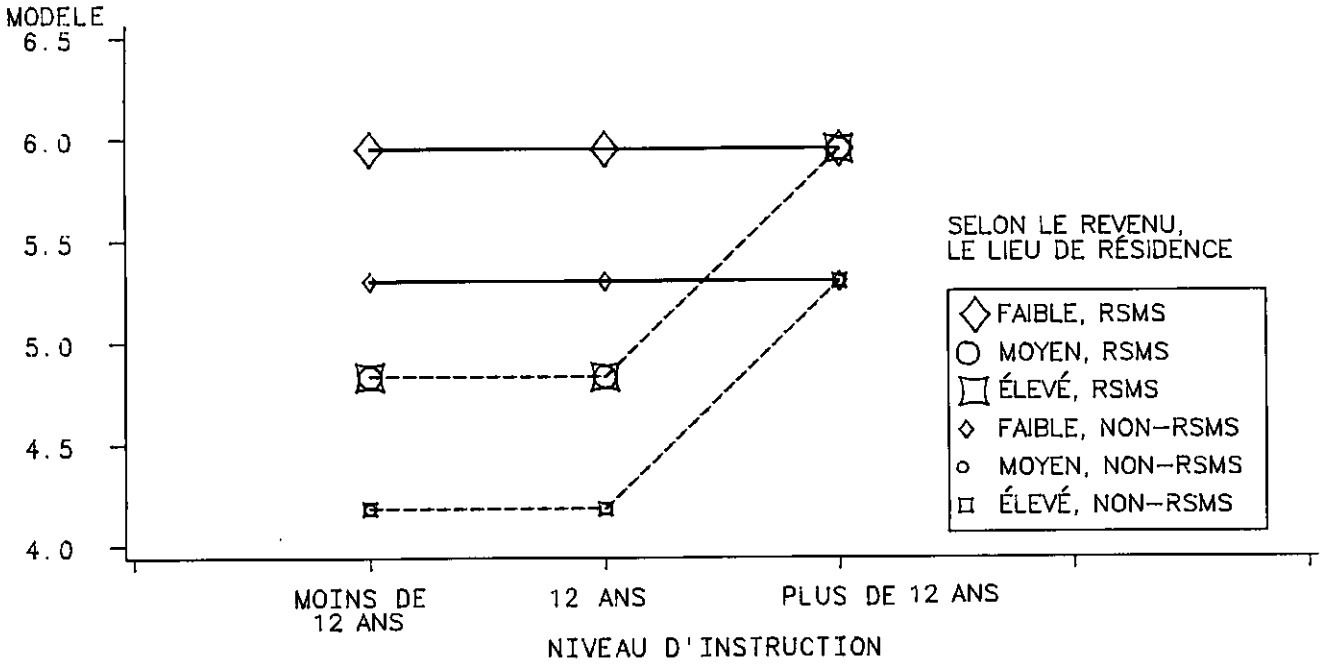

Figure 4: Moyenne prédite selon le modèle du nombre annuel de visites chez les médécins par habitant, États-Unis, 1973.

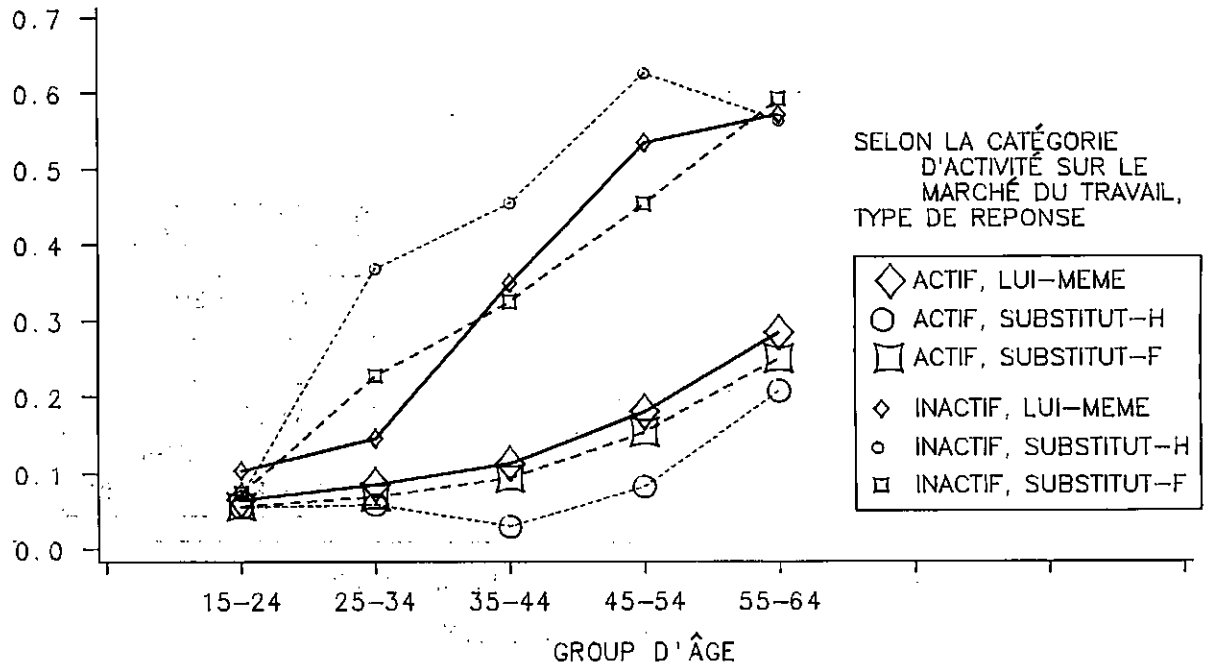

Figure 5: Taux de sélection observées, Enquête canadienne sur la santé et l'invalidité, janvier 1983, hommes 15-64.

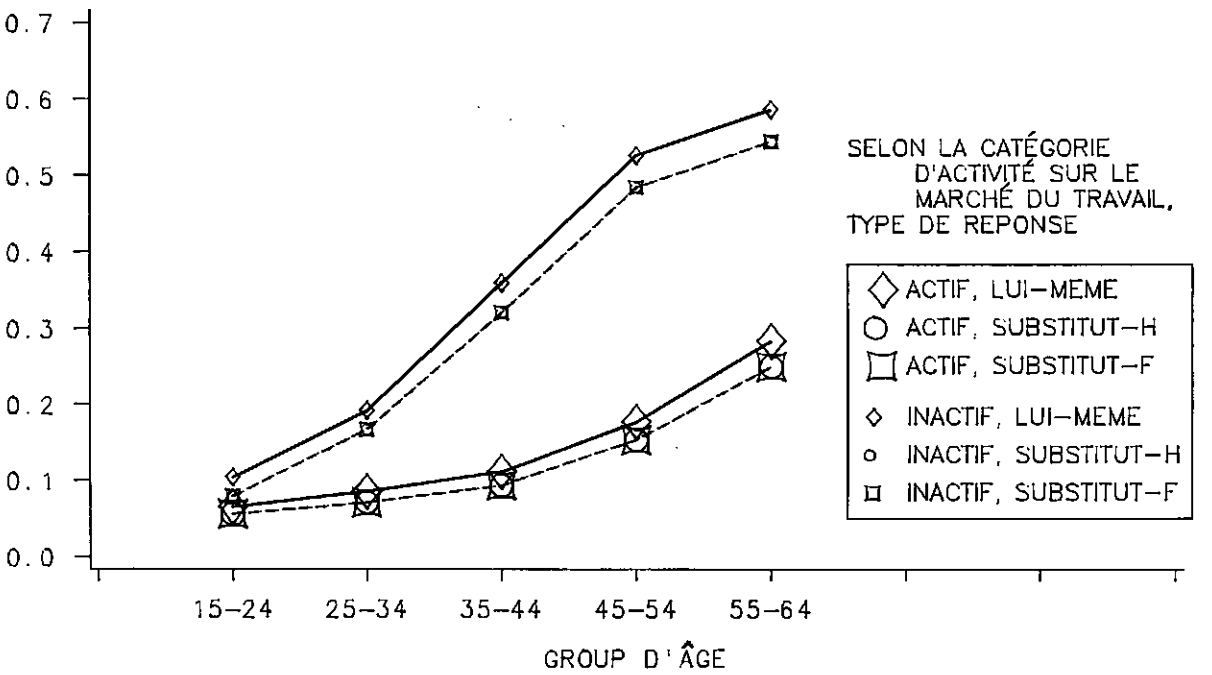

Figure 6: Taux de sélection prédits, Enquête canadienne sur la santé et l'invalidité, janvier 1983, hommes 15-64.

L'AJUSTEMENT DES SÉRIES INFRA-ANNUELLES AUX RÉPÈRES ANNUELS

Pierre A. Cholette[1]

Ce travail propose une modification de la méthose d'étalonnage de Denton (1971) pour l'ajustement des séries infra-annuelles aux totaux annuels. Ces totaux proviennent de source plus fiable et constituent des repères ou des jalons annuels. La série ajustée selon la méthode modifiée s'avère plus parallèle à la série non ajustée que ce n'est le cas avec la méthode originale. Des variantes additive et proportionnelle de la méthode sont exposées. Elles s'adaptent facilement aux séries de flux, de stock et d'indice. On trouve aussi quelques recommendations relatives à l'étalonnage préliminaire des données courantes et à la gestion des estimés "historiques" de la série.

1. INTRODUCTION

Dans un grand nombre de cas, le statisticien obtient des données infra-annuelles d'une série à partir d'une source de données (telle un échantillon); et, les valeurs annuelles repères correspondantes à partir d'une autre source de données plus fiable (telle un recensement). Les totaux annuels des observations infra-annuelles ne sont généralement pas égaux aux valeurs annuelles repères. De telles séries nécessitent l'ajustement aux jalons annuels, c'est-à-dire l'étalonnage.

La solution proposée par Denton (1971) (et généralisée par Fernandez en 1981) consiste à trouver une série infra-annuelle qui épouserait le plus possible le mouvement de la série infra-annuelle disponible et dont les sommes (ou moyennes) annuelles correspondraient aux repères annuels plus fiables. Le niveau de la série résultante serait ainsi donné par les repères annuels, tandis que son mouvement serait gouverné par la série infra-annuelle originale. En d'autres mots, la série ajustée devrait être la plus parallèle possible à la série originale, tout en satisfaisant aux jalons annuels. Ce travail recommande une modification à la spécification de Denton qui rend les séries originale et ajustée encore plus parallèles.

---

[1] Pierre A. Cholette, Séries chronologique recherche et analyse.

Nous suivons le modèle de Ehrenberg (1982) pour la présentation des textes scientifiques. Le lecteur se verra exposé les illustrations et résultats d'abord; et, les détails méthodologiques, ensuite.

2. ILLUSTRATION DE LA MÉTHODE

La figure 1 montre les corrections (x_t - z_t) apportées à la série originale z_t selon la solution additive (avec premières différences) de Denton et selon la solution correspondante proposée dans ce travail. Puisque ces corrections devront être ajoutées à la série infra-annuelle originale z_t, la série ajustée x_t sera tout à fait parallèle à la série originale si et seulement si les corrections sont constantes. Dans la figure, cela se produit seulement pour les corrections associées à la méthode proposée dans ce travail.

La figure 1 montrait un cas trivial et idéal qui admettait la solution des corrections constantes: tous les écarts annuels moyens, différences entre les jalons annuels et les totaux annuels de la série originale (divisées par le nombre de mois par année), étaient constants. La figure 2 propose un cas plus réaliste, où les cinq écarts annuels moyens fluctuent autour de 200. Comme dans le premier exemple, les corrections calculées selon la méthode présentée ici sont beaucoup plus constantes, surtout dans la première année.

Comme expliqué plus bas, la méthode de Denton minimise non seulement le changement dans les corrections (pour les rendre les plus constantes possible) mais aussi la taille de la première correction. Cela se vérifie dans les figures 1 et 2, où les premières corrections avoisinent zéro. Par contre, la solution alternative minimise seulement le changement dans les corrections. Graphiquement, cela consiste à tracer, à travers les écarts annuels moyens, une courbe qui soit la plus plate possible et qui recouvre aussi les mêmes surfaces annuelles que les écarts annuels moyens.

3. CONSERVATION DU PARALLÉLISME DES SÉRIES ORIGINALE ET AJUSTÉE

Reprenant la formulation additive avec premières différences de Denton ainsi que sa notation, la série recherchée x_t minimise la fonction objective suivante

$$p(x) = \sum_{t=1}^{n} (\Delta x_t - \Delta z_t)^2 = \sum_{t=1}^{n} (\Delta(x_t - z_t))^2, \; x_0 = z_0, \tag{1}$$

où z_t représente la série infra-annuelle originale au temps t. Cette fonction est minimisée sujette aux contraintes d'égalité entre les sommes annuelles des valeurs obtenues et les jalons disponibles y_i:

$$\sum_{t=(i-1)k+1}^{ik} x_t = y_i, \; i = 1, 2, \ldots, m. \tag{2}$$

où k est le nombre de "mois" par année.

Denton justifie l'hypothèse $x_0 = z_0$ en prétendant qu'il est légitime de supposer l'égalité des dernières valeurs observée et ajustée, antérieures à l'intervalle d'estimation. La fonction objective (1) signifierait donc que la série ajustée x_t devrait avoir la même pente que la série originale z_t : et, que par conséquent la pente des différences entre les deux séries devrait être minimisée (sujette aux contraintes). Mais, on peut récrire la fonction objective (1) de la manière suivante en substituant $x_0 = z_0$:

$$p(x) = (x_1 - z_1)^2 + \sum_{t=2}^{n} (\Delta(x_t - z_t))^2. \tag{3}$$

Cette transformation met clairement en évidence que l'hypothèse $x_0 = z_0$ implique la minimisation de la taille de la première correction. Comme illustré aux figures 1 et 2, la minimisation de la première correction tire la courbe de correction vers zéro en début de série. Cela produit une ondulation dans la première année qui se répercute sur les autres années. Cette ondulation dans les corrections empêche, par définition, le parallélisme maximum des séries observée et ajustée.

La spécification proposée ici s'abstient tout simplement de l'hypothèse $x_0 = z_0$ et donne la fonction objective suivante:

$$p(x) = \sum_{t=2}^{n} (\Delta(x_t - z_t))^2. \tag{4}$$

sujette aux mêmes contraintes de l'équation (2).

En algèbre linéaire, la fonction objective contrainte s'écrit

$$\underline{u}(\underline{x}, \underline{q}) = (\underline{x} - \underline{z})'\underline{A}(\underline{x} - \underline{z}) - 2\,\underline{q}'(\underline{y} - \underline{B}'\underline{x}), \tag{5}$$

où les vecteurs matrices impliqués valent:

$$\underset{n\times 1}{x} = \begin{vmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{vmatrix}, \quad \underset{n\times 1}{\underline{z}} = \begin{vmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{vmatrix}, \quad \underset{m\times 1}{\underline{y}} = \begin{vmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{vmatrix}, \quad \underset{m\times 1}{\underline{q}} = \begin{vmatrix} q_1 \\ q_2 \\ \cdot \\ \cdot \\ \cdot \\ q_m \end{vmatrix}. \tag{6}$$

$$\underset{n\times n}{\underline{A}} = \underline{D}'\underline{D}, \quad \underset{(n-1)\times n}{\underline{D}} = \begin{vmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{vmatrix}, \tag{7}$$

$$\underset{n\times m}{\underline{B}} = \begin{vmatrix} \underline{j} & 0 & \cdots \\ 0 & \underline{j} & \cdots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{vmatrix}, \quad \underset{k\times 1}{\underline{j}} = \begin{vmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \end{vmatrix}, \quad (n = km). \tag{8}$$

Le vecteur $\underline{q}$ contient les multiplicateurs de Lagrange. Les variables n (= mk), m et k désignent respectivement le nombre d'observations et le nombre d'années dans la série et le nombre de mois par année

Les équations normales associées à la fonction objective (5) sont

$$\begin{aligned} d\underline{u}/d\underline{x} &= (\underline{A} + \underline{A}')(\underline{x} - \underline{z}) + 2\,\underline{B}\,\underline{q} = \underline{0} \\ d\underline{u}/d\underline{q} &= 2(\underline{B}' - \underline{y}) = \underline{0} \end{aligned} \tag{9}$$

et débouchent sur la solution

$$\begin{vmatrix} \underline{x} \\ \underline{q} \end{vmatrix} = \begin{vmatrix} \underline{A} & \underline{B} \\ \underline{B}' & \underline{0} \end{vmatrix}^{-1} \begin{vmatrix} \underline{A} & \underline{0} \\ \underline{0} & \underline{I} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{y} \end{vmatrix} = \underset{(n+m)\times(n+m)}{W} \begin{vmatrix} \underline{z} \\ \underline{y} \end{vmatrix}. \tag{10}$$

La substitution de l'identité $\underline{y} = \underline{B}'\underline{z} + \underline{r}$, $\underline{r}$ renfermant les écarts annuels. donne

$$\begin{vmatrix} \underline{x} \\ \underline{q} \end{vmatrix} = \begin{vmatrix} \underline{A} & \underline{B} \\ \underline{B}' & \underline{0} \end{vmatrix}^{-1} \begin{vmatrix} \underline{A} & \underline{0} \\ \underline{B}' & \underline{I} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{r} \end{vmatrix} = \begin{vmatrix} \underset{n\times n}{\underline{I}} & \underset{n\times m}{\underline{W}_x} \\ \underset{m\times n}{\underline{0}} & \underset{m\times m}{\underline{W}_1} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{r} \end{vmatrix} \Rightarrow \underline{x} = \underline{z} + \underline{W}_x \underline{r}. \tag{11}$$

Cette reformulation de la solution réduit le temps de calcul dans l'application des poids calculés comparativement à la formulation (10). À remarquer aussi qu'une fois obtenus, les poids W_x peuvent servir pour un nombre quelconque de séries ayant le même nombre d'observations. Nous recommandons en outre (Cholette, 1978, section 6: 1979, 4.3) de calculer W_x pour un intervalle quinquennal et de l'utiliser à la manière d'une moyenne mobile (se mouvant d'une année à la fois) pour les séries de cinq ans et plus. En plus d'économiser les calculs, ce procédé n'engendre que deux révisions des estimés (ceteris paribus) lorsque d'autres années d'observations s'ajoutent à la série.

Denton résout l'inversion de l'équation (10) par partie. Cela est impossible ici, car la matrice A est singulière. Par contre, l'ensemble de la matrice n'est pas singulier et s'inverse.

La méthode exposée ici utilise en fait la solution que Boot, Feibes et Lisman (1967) proposaient pour interpoler entre des données annuelles en l'absence d'information infra-annuelle. La solution (11) consiste exactement à interpoler entre les écarts annuels avec la méthode de ces auteurs et à ajouter les estimés obtenus (les corrections) à la série infra-annuelle originale.

4. VARIANTE PROPORTIONNELLE

La méthode proportionnelle maintenant présentée est aussi une variante de la méthode proportionnelle de Denton, dont on a retiré $x_0 = z_0$. Comme dans la

section 2, la fonction objective minimise toujours la somme des différences quadratiques de pente entre les séries infra-annuelles originale et recherchée (z_t et x_t). Mais chaque terme de la somme est pondéré par la valeur de l'observation infra-annuelle correspondante:

$$p(x) = \sum_{t=2}^{n} (\Delta(x_t - z_t)/z_t)^2 = \sum_{t=2}^{n} (\Delta(x_t/z_t))^2. \tag{12}$$

Cette variante convient aux séries à forte saisonnalité, lorsqu'on juge que les mois de creux saisonnier ne peuvent être aussi responsables de l'écart annuel que les mois de sommet saisonnier: La taille de chaque correction est proportionnelle au niveau de l'observation, comme illustré dans la figure 3. Les observations faibles écopent de corrections plus petites que les observations (saisonnièrement) fortes, même si les corrections proportionnelles z_t/z_t minimisées sont aussi constantes que les écarts annuels le permettent. À noter qu'avec la variante proportionnelle, toutes les observations doivent être positives et que les valeurs ajustées seront aussi toutes positives.

On peut également démontrer (Cholette, 1978, section 3: 1979, 3) que la variante proportionnelle est une approximation linéaire de la méthode fortement non linéaire de préservation des taux de croissance (Smith, 1977: Helfand et al., 1978) qui aurait la fonction objective suivante:

$$p(x) = \sum_{t=2}^{n} (x_t/x_{t-1} - z_t/z_{t-1})^2 \tag{13}$$

L'approximation est exacte en régime d'écarts annuels proportionnels constants sur l'intervalle d'estimation.

En algèbre linéaire, la fonction objective contrainte associée à la méthode proportionnelle s'écrit

$$\underline{u}(\underline{x}, \underline{g}) = (\underline{x} - \underline{z})' \, \underline{Z}^{-1}\underline{A} \, \underline{Z}^{-1} \, (\underline{x} - \underline{z}) - 2 \, \underline{g}'(\underline{y} - \underline{B}'\underline{x}), \tag{14}$$

où Z^{-1} est une matrice diagonale dont les éléments sont $1/z_1$, $1/z_2$. ... La solution a la même structure que la variante additive ($Z^{-1}A\,Z^{-1}$ remplaçant A dans (11)) et s'écrit:

$$\begin{vmatrix} \underline{x} \\ \underline{g} \end{vmatrix} = \begin{vmatrix} \underline{Z}^{-1}\underline{A} & \underline{Z}^{-1}\underline{B} \\ \underline{B}' & \underline{0} \end{vmatrix}^{-1} \begin{vmatrix} \underline{Z}^{-1}\underline{A} & \underline{Z}^{-1}\underline{0} \\ \underline{B}' & \underline{I} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{r} \end{vmatrix} = \begin{vmatrix} \underline{I} & \underline{W}_x \\ \underline{0} & \underline{W}_1 \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{r} \end{vmatrix}. \tag{15}$$

Contrairement à ceux de la variante additive, on doit cependant calculer les poids W_x de la solution proportionnelle pour chaque série et même chaque intervalle d'application d'une série donnée.

5. SÉRIES DE STOCKS ET D'INDICE

Les variantes additive et proportionelle de la méthode présentée ci-dessus sont conçues pour des séries de flux, dont la valeur annuelle correspond à la somme des valeurs infra-annuelles. On peut très facilement adapter les solutions trouvées aux séries de stock ou cumulatives, dont la valeur annuelle n'est associée qu'à une seule valeur infra-annuelle (habituellement celle du dernier mois); ainsi qu'aux séries d'indice, dont la valeur annuelle correspond à la moyenne des valeurs infra-annuelles. Pour une série trimestrielle cumulative par exemple, il suffit tout simplement de redéfinir le vecteur $\underline{j}$, composante de $\underline{B}$, comme ceci

$$\underset{1\times 4}{\underline{j}'} = [0\ 0\ 0\ 1]:$$

et, pour une série mensuelle d'indice comme ceci

$$\underset{1\times 12}{\underline{j}'} = [1/12\quad 1/12\quad \ldots\quad 1/12].$$

6. DISCUSSION

6.1 Données historiques

Il y a selon nous beaucoup de confusion quant à l'interprétation de l'hypothèse $x_0 = z_0$ de Denton. L'auteur écrit à ce sujet: "On suppose qu'il n'y a pas d'ajustement à faire à la série originale pour les années extérieures à l'intervalle allant des années 1 à m inclusivement." (p. 100, au dessus de l'équation 3.2, notre traduction).

Si on laisse ces années intactes parce qu'elles n'ont jamais eu de repères annuels, la solution proposée par Denton est défendable: il n'en découle aucune correction pour les années -1 et 0: et. de petites corrections graduellement introduites au début de l'année 1. (On se souvient que $x_0 = z_0$ implique la minimisation de la première correction.) La série ajustée résultante est donc continue, comme illustré à la figure 4 par la courbe ADEB.

Par contre, si on laisse intactes les premières années parce qu'elles ont déjà été ajustées aux jalons et qu'on les considère maintenant comme "historiques", nous ne sommes pas d'accord avec l'hypothèse $x_0 = z_0$. Généralement en effet, celle-ci produira une discontinuité entre les années 0 et 1 comme illustré dans figure 4 par la courbe A'CDEB. Les années -1 et 0 ont déjà reçu des corrections de taille voisine de CD, tandis que le début de l'année 1 reçoit des corrections les plus petites possibles.

Pour rendre immuables les données historiques après un certain nombre d'années, deux solutions sont possibles. La première consiste à spécifier explicitement la contrainte d'immuabilité dans la fonction objective qui devient

$$p(x) = ((x_1 - z_1) - (x_0 - z_0))^2 + \sum_{t=2}^{n} (\Delta(x_t - z_t))^2, \qquad (16)$$

où $(x_0 - z_0)$ est connu et égal à la dernière correction utilisée pour l'année historique 0. Cette correction est généralement différente de zéro (Cholette, 1979b, 1983). Cette spécification équivaut à déterminer le point de départ de la courbe de correction.

Une deuxième solution, moins spécifique mais aussi efficace, consiste à appliquer la méthodologie déjà proposée dans ce travail (version additive ou proportionnelle) à la manière d'une moyenne mobile se mouvant annuellement. Avec un intervalle quinquennal d'application, les estimés deviennent automatiquement définitifs après deux années de révision: et, après une année, en régime triennal (Cholette, 1978, section 6 a; 1974, 4.3). La série ajustée aux jalons résultante est également continue, comme l'illustre la courbe A'CB de la figure 4.

6.2 Mise en oeuvre

Les praticiens de l'étalonnage ont tendance à soumettre au programme d'étalonnage les années de données déjà ajustées aux jalons suivie d'une année de

données non ajustées (accompagnées de leurs jalons annuels). Pour des méthodologistes, il est évident qu'il faut toujours soumettre les données non ajustées (avec leurs jalons). La soumission de données ajustées provoquera généralement un mouvement saisonnier artificiel dans la série ajustée résultante (Cholette, 1978, section 6b).

6.3 Étalonnage préliminaire des données courantes

Finalement une dernière remarque s'impose. Pendant une année (incomplète) en cours, on ne peut pas calculer des taux de croissance, par exemple, entre le segment de la série ajustée aux jalons (AB) avec le segment non ajusté (CD). Ce faire produit généralement une discontinuité BC entre les deux segments AB et CD, comme illustré dans la figure 5 par la courbe ABCD.

Deux issues s'offrent alors. Premièrement, on fait les comparaisons inter-temporelles en se basant seulement sur les données non ajustées. Deuxièment, on effectue un ajustement préliminaire des données courantes, en répétant la dernière correction BC pour l'année en cours. (À noter que l'incorporation de l'année incomplète courante dans la fonction objective (4) (ou 12) produirait des valeurs ajustées préliminaires identiques.) On peut ensuite comparer le segment ajusté AB avec le segment préliminairement ajusté BE, comme illustré dans la figure 5 par le courbe ABE. Nous favorisons cette deuxième alternative.

6.4 Rapport avec les autres méthodes

On pourrait qualifier d'univariées les méthodes d'étalonnage de Denton (1971), de Denton modifiée (présentée ici), de Glejser (1966), de Boot, Feibes et Lisman (1967), de Lisman et Sandee (1964), et de Bassie (1939). En effet, ces méthodes ne font intervenir que la série considérée et ses jalons annuels dans le processus d'étalonnage. Les méthodes de Friedman (1962), Chow et Lin (1971) Somermeyer, Jansen et Louter (1976), et de Wilcox (1983) sont au contraire multivariées. Des séries auxilières y sont utilisées dans le calcul de la série recherchée.

Par exemple, Chow et Lin (1971) ont proposé une méthode pour obtenir la série infra-annuelle désirée à partir de totaux annuels et de série apparentées. Le mouvement de la série résultante est le plus semblable possible aux

mouvements des séries apparentées (et la série trouvée satisfait aux contraintes annuelles). Fernandez (1981) note que la méthode de Chow et Lin peut produire des discontinuités de mouvements entre les années. Il propose alors une synthèse des méthodes de Chow et Lin et de Denton (1971). La méthode combinée élimine les discontinuités inter-annuelles mais repose également sur l'hypothèse $x_0 = z_0$. Comme illustré plus haut, cette hypothèse engendre souvent des fluctuations artificielles dans la série calculée. Nous aurions tendance à croire qu'il est possible de s'abstenir de cette hypothèse dans le cas de la méthode de Fernandez comme pour la méthode de Denton.

7. RÉSUMÉ ET CONCLUSIONS

Denton (1971) voulait garder les séries originale et ajustée aux jalons annuels aussi parallèles que le permettaient les écarts annuels. Ce travail a proposé une modification de la méthode d'étalonnage qui rend les séries originale et ajustée plus parallèles l'une à l'autre que ce n'est le cas avec la méthode originale. Cette amélioration vaut autant pour les variantes additive que proportionnelle. Nous pensons que la méthode généralisée multivariée de Fernandez pourrait être améliorée dans le même sens.

On peut tout aussi facilement adapter la méthode proposée aux séries de flux, de stock et d'indice.

Avant de procéder à des comparaisons inter-temporelles entre les données ajustées et les nouvelles données courantes, il est essentiel d'ajuster aux jalons de manière préliminaire les données courantes (de la façon proposée).

La mise en oeuvre suggérée, à la façon d'une moyenne mobile quinquennale, laissera automatiquement intactes les estimations passées après deux années de révision.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Baldwin, A. (1978), "New Benchmarking Algorithms using Quadratic Minimization," National Product Division, Statistique Canada, Document de recherche.

[2] Bassie, B.L. (1939), "Interpolation Formulae for the Adjustment of Index Numbers," Proceedings of the Annual Meetings of the American Statistical Association.

[3] Boot, J.C.G., Feibes, W., Lisman, J.H.C. (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 16, No. 1, pp.65-75.

[4] Cholette, P.A. (1978), "Comparaison et évaluation de quelques méthodes d'ajusment de séries infra-annuelles aux repères annuels," Séries chronologique recherche et analyse, Statistique Canada, document de recherche 78-03-001B.

[5] Cholette, P.A. (1979a), "Adjustment Methods of Sub-Annual Series to Yearly Benchmarks," Proceedings of the Computer Science and Statistics, 12th Annual Symposium on the Interface, J.F. Gentleman Ed., University of Waterloo, pp. 358-36.

[6] Cholette, P.A. (1979b), "A Note on 'Freezing' Past Estimates when Benchmarking," Séries chronologique recherche et analyse, Statistique Canada, document de recherche 79-06-002E.

[7] Cholette, P.A. (1982), "Programme d'ajustement quadratique minimum (PAQM-I) de séries aux totaux annuels - Manuel des utilisateurs," Séries chronologique recherche et analyse, Statistique Canada, 82-11-003B.

[8] Cholette, P.A. (1983), "Étalonnage de séries en régime de jalons bi-annuels et de connaissance du point d'arrivée," Séries recherche et analyse, Statistique Canada, document de recherche 83-05-002B.

[9] Chow, G.C., Lin, An-loh (1971), "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series," Review of Economics and Statistics, Vol. 53, No. 4, pp. 372-375.

[10] Dagum, E.B. (1977), "Comparison of Various Interpolation Procedures for Benchmarking Economic Time Series," Séries chronologique reserche et analyse, Statistique Canada, document de recherché 77-05-006E.

[11] Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization," J.A.S.A., Vol. 66, No. 333, pp. 99-102.

[12] Fernandez, R.B. (1981), "A Methodoloqical Note on the Estimation of Time Series," Review of Economic and Statistics, Vol. 63, pp. 471-476.

[13] Friedman, M. (1962), "The Interpolation of Time Series by Related Series," J.A.S.A., Vol. 57, No. 300, pp. 729-757.

[14] Glejser, H. (1966), "Une méthode d'évaluation de données mensuelles à partir d'indices trimestriels ou annuels," Cahiers Economiques de Bruxelles, No. 19, 1er trimestre, pp. 45-64.

[15] Helfand, S.D. Monsour, N.J, Traqer, M.L. (1978), "Historical Revision of Current Business Survey Estimates," U.S. Bureau of the Census, (Document de recherche).

[16] Huot, G. (1975), "Quadratic Minimization of Montlhly Estimates to Annual Totals," Séries chronoloqique recherche et analyse, Statistique Canada, Document de recherche 75-11 010E.

[17] Lisman, J.H.C., Sandee, J. (1964), "Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 13, No. 2, pp. 87-90.

[18] Smith, P. (1977), "Alternative Method for Step Adjustment," Groupes des analyses de conjoncture, Statistique Canada, document de recherche.

[19] Somermeyer, W.H, Jansen, R., Lauter, A.S. (1976), "Estimatinq Quarterly Values from Annually Known Variables in Quarterly Relationships," J.A.S.A, Vol. 71, No 355, pp. 588-595.

[20] Wilcox, J.A. (1983), "Disaqqreqating Data Usinq Related Series," Journal of Business and Economic Statistics, Vol. 1, No 3, pp. 187-191.

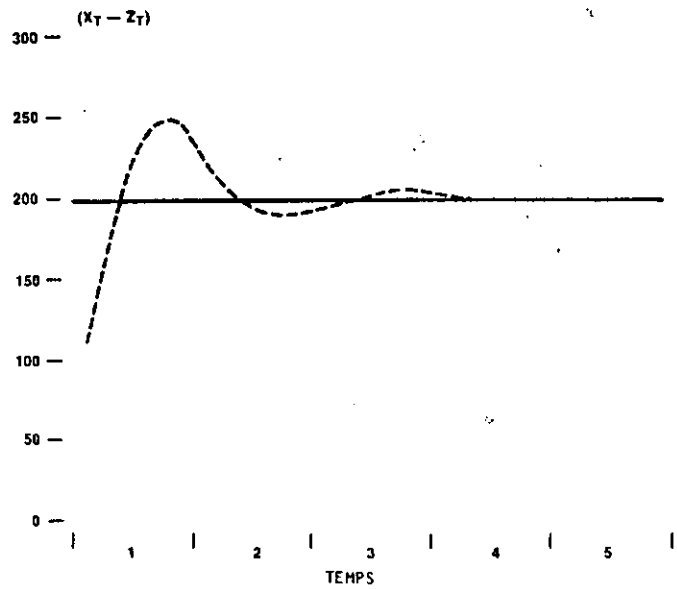

Figure 1: Corrections (x_t - z_t) apportées à la série non ajustée aux jalons selon la méthode de Denton (ligne en tirets) et selon la méthode d'étalonnage proposée dans ce travail (continue), en régime idéal d'écarts annuels constants.

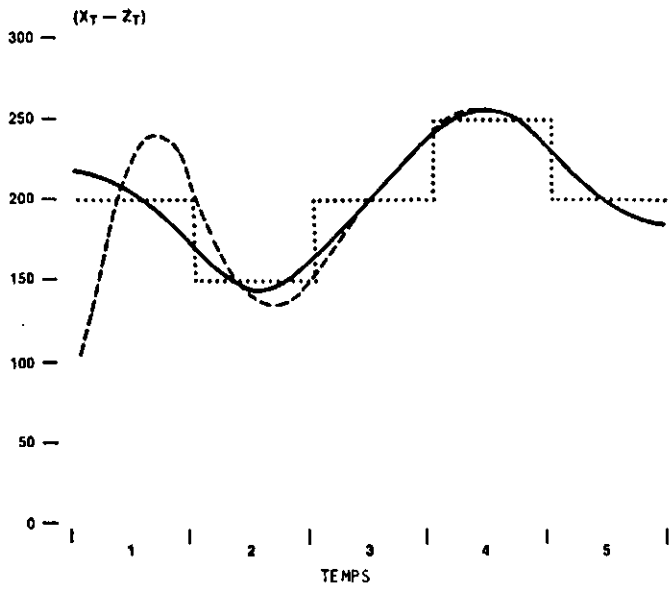

Figure 2: Corrections (x_t - z_t) apportées à la série non ajustée aux jalons selon la méthode de Denton (ligne en tirets) et selon la méthode d'étalonnage proposée dans ce travail (continue), en régime d'écarts annuels moyens variables (pointillée).

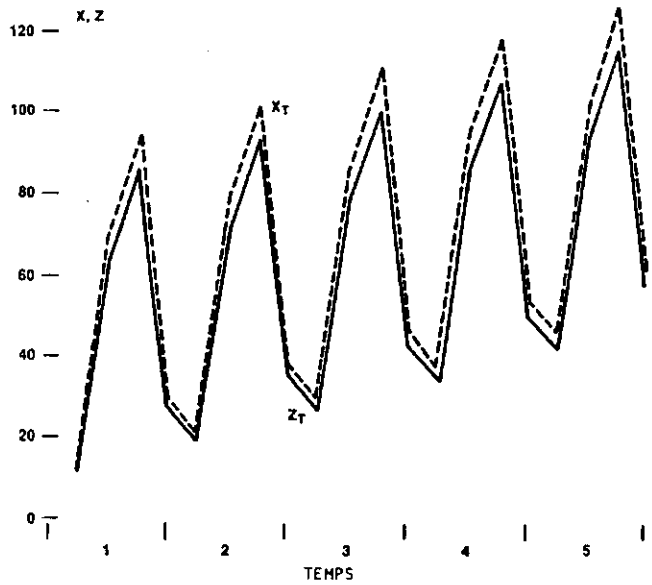

Figure 3: Série originale (ligne continue) et série ajustée aux jalons (tirets) selon la variante proportionnelle de la méthode d'étalonnage proposée dans ce travail (en régime d'écarts annuels proportionnels constants).

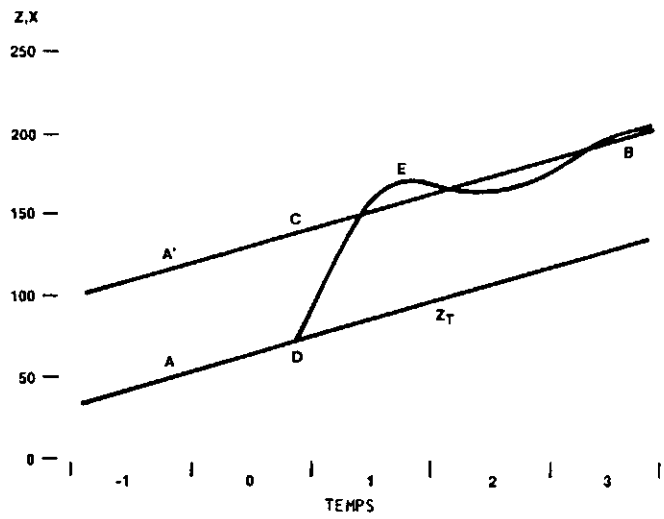

Figure 4: Séries ajustées aux jalons selon la méthode de Denton, en l'absence de jalons pour les années -1 et 0 (courbe ADEB) et en présence de jalons annuels et d'ajustement préalable pour les années -1 et 0 (A'CDEB); et selon la méthode proposée dans ce travail, appliquée à la manière d'une moyenne mobile, en présence de jalons annuels pour les années -1 et 0 (A'CB).

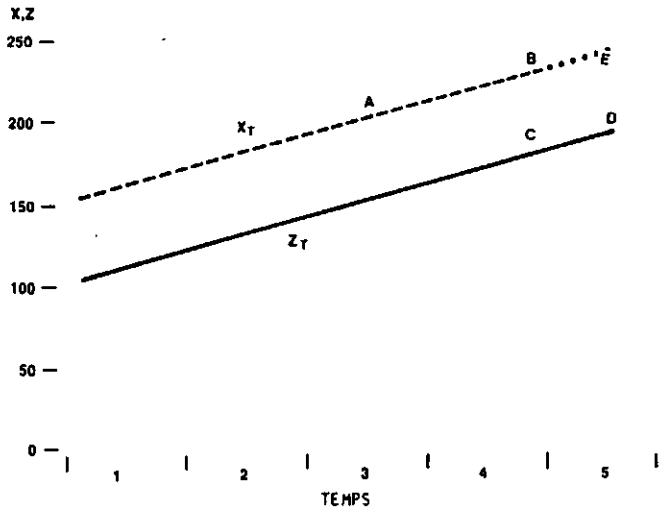

Figure 5: Continuité entre la série ajustée aux jalons (ligne en tirets) et la série préliminairement ajustée (pointillée) et discontinuité BC entre les séries ajustée (tirets) et non ajustée (continue).

ANALYSE DES DÉPENSES CONSACRÉES À L'ÉNERGIE

Louise A. Heslop[1]

À l'aide des données chronologiques de l'enquête sur les dépenses des familles, on effectue actuellement des analyses des dépenses de consommation d'énergie pour la maison et le transport, pour les années 1969 à 1982. Dans le présent article, on décrit brièvement certaines des méthodes d'analyse utilisées, on présente des résultats sommaires et on s'intéresse aux changements survenus dans la consommation dans la mesure où ils peuvent influer sur l'orientation des politiques. Devant un ensemble de données si complexe et un sujet si difficile et si polyvalent, il faut tâcher de réduire les flux d'information, tout en augmentant le contenu informatif de chaque facteur qui intervient dans les analyses, tant en entrée qu'en sortie.

1. LA QUESTION ÉNERGÉTIQUE

Pour certains, le problème de la conservation de l'énergie ne se pose plus. Il n'y a pas de pénurie (peut-être n'y en a-t-il jamais eue), et les prix se sont stabilisés.

La question énergétique a dominé l'actualité des années 1970, bouleversant profondément l'ordre économique mondial et engendrant des affrontements à l'échelle internationale. Au pays, cette question a eu de lourds effets sur les relations fédérales-provinciales, sur les liens entre le monde des affaires et le gouvernement et sur les budgets familiaux. Elle a entraîné une restructuration de l'industrie manufacturière, du secteur de l'automobile, etc.

Bien qu'apparemment relégués au second plan, les problèmes liés à la consommation et aux prix de l'énergie demeurent prioritaires pour les consommateurs, les entreprises et le gouvernement. La conservation de l'énergie ne soulève plus les passions, mais elle demeure un problème crucial.

La recherche dont fait état cet article a été organisée en consultation avec des responsables des politiques aux ministères de la Consommation et des

[1] Louise A. Heslop, Division de la recherche et de l'analyse, Statistique Canada.

Corporations et de l'Énergie, des Mines et des Ressources, où l'on continue de mener des programmes de recherche actifs sur l'utilisation et la conservation de l'énergie par les consommateurs. La structure du projet tient compte des intérêts, des orientations et des contraintes de ces ministères.

Par ailleurs, au cours des cinq dernières années, un groupe international de spécialistes des sciences sociales a entrepris un programme de recherche et d'échange d'information sur les comportements des consommateurs et les utilisations de l'énergie. À titre de membre du groupe, l'auteur est bien au fait des problèmes et des perspectives examinés par ce groupe, ainsi que de l'état actuel de ses travaux et de ses techniques de recherche.

2. PROBLÈMES LIÉS À LA RECHERCHE SUR LES QUESTIONS ÉNERGÉTIQUES

Le problème principal qu'on rencontre dans les études sur la consommation d'énergie, c'est peut-être celui d'obtenir des mesures raisonnablement fiables de la consommation à partir d'échantillons suffisamment vastes et représentatifs. Un chercheur serait comblé si, en outre, il pouvait disposer de telles données pour une certaine période, en particulier si celle-ci englobait le fameux choc pétrolier de 1973. Les données sur les dépenses des familles recueillies par la Division des revenus et des dépenses des familles de Statistique Canada correspondent assez bien à cette description pour au moins soulever l'intérêt du chercheur. Il s'agit de données provenant d'une série d'enquêtes rétrospectives réalisées pour les années 1969, 1972, 1974, 1976, 1978 et 1982. La période intéressante est donc couverte, l'échantillon est vaste et les méthodes d'échantillonnage employées garantissent une bonne représentativité pour les secteurs étudiés, habituellement les centres urbains. En outre, les données comprennent un grand nombre d'autres variables utiles dans toute étude de la consommation d'énergie, par exemple le mode d'occupation des logements, certaines caractéristiques des logements, le nombre de propriétaires de véhicules et d'appareils ménagers, les caractéristiques des familles, les dépenses consacrées à d'autres catégories de biens et de services de consommation, etc.

Dans la plupart des enquêtes sur les dépenses de consommation, on demande aux répondants de fournir des renseignements de mémoire ou par référence à des

dossiers. L'exactitude des données ainsi obtenues pose des problèmes au niveau des individus, mais ces problèmes sont atténués lorsqu'on dispose d'échantillons très grands. Dans la plupart des enquêtes indépendantes, toutefois, il en coûteraît trop cher de former des échantillons aussi vastes. Heureusement, les données de l'enquête sur les dépenses des familles proviennent d'échantillons de très grande taille.

Il n'est arrivé qu'une seule fois au Canada qu'on tente dans une enquête importante d'effectuer une vérification indépendante de dossiers, en se procurant les documents relatifs à chaque ménage auprès des fournisseurs, avec la permission du chef de ménage. Toutefois, on n'a pu obtenir des dossiers sur la consommation d'électricité que pour moins de la moitié de l'échantillon, et sur la consommation de mazout et de gaz naturel, pour à peu près le tiers de l'échantillon seulement. Ce procédé de vérification des dossiers assure un degré élevé d'exactitude, élimine les problèmes liés au besoin de se rappeler des événements passés, surtout s'ils remontent à une période éloignée, et empêche l'introduction d'un biais de réponse. Il est toutefois impossible, en pratique, de le mettre en oeuvre pour de vastes échantillons couvrant l'ensemble du pays.

Bien que l'enquête sur les dépenses des familles fasse appel à la mémoire des répondants, les renseignements sur les dépenses consacrées à l'énergie ne sont probablement pas entachés d'un biais aussi grand que dans une enquête portant expressément sur la consommation d'énergie, car les répondants ne sont pas sensibilisés à l'objet de l'enquête. Par ailleurs, les données ont été recueillies selon les mêmes méthodes avant et après la crise de l'énergie, ce qui réduit encore la possibilité de biais de réponse. Les données de cette enquête offrent donc une occasion unique d'observer un ensemble très vaste d'échantillons pendant une période cruciale.

Les données posent quand même certains problèmes, dont certains sont attribuables à la méthode d'échantillonnage et d'autres proviennent de la complexité inhérente à toute enquête sur la consommation d'énergie. En raison des changements apportés aux catégories de dépenses et à leur contenu, en particulier aux catégories autres que l'énergie, il nous a fallu manipuler considérablement les données afin d'en assurer l'uniformité d'une période à l'autre. Il est impossible de connaître les dépenses consacrées à l'énergie à la maison par les familles qui ne paient pas cette énergie directement, par exemple les

locataires d'appartements situés dans des immeubles à compteur central et les chambreurs.

Certains chercheurs ont imputé des valeurs à ces ménages en se fondant sur leurs loyers, mais nous avons choisi de ne pas procéder ainsi et de restreindre plutôt notre étude aux ménages qui sont en mesure de surveiller et de contrôler leur propre consommation d'énergie. Il s'agit en effet des groupes de consommateurs auxquels s'adressera tout programme gouvernemental visant à modifier la consommation.

Plusieurs facteurs font qu'il est difficile d'étudier et d'infléchir les schémas de consommation d'énergie des ménages :

- les caractéristiques des biens : elles peuvent limiter la capacité du ménage d'agir rapidement et augmenter le coût des changements (ces caractéristiques comprennent la taille de la maison, le nombre et le type d'appareils ménagers, la taille et le nombre de véhicules, etc.) On note dans certaines études que les caractéristiques des maisons peuvent expliquer à elles seules 24% de la consommation d'énergie au foyer. La taille du ménage peut être incluse au nombre des caractéristiques des biens.

- la possibilité de changement : il n'est pas toujours possible de changer la quantité et les types de combustibles utilisés, car on n'a pas toujours accès, selon la situation ou les montants à investir, à la technologie ou aux combustibles voulus; le chauffage au gaz naturel, par exemple, est inaccessible aux résidents des régions rurales : on ne peut pas, dans ce cas, modifier instantanément le type de combustible utilisé.

- facteurs exogènes : ils influent sur la quantité d'énergie nécessaire pour obtenir un même rendement dans des situations différentes (au nombre de ces facteurs figurent les conditions météorologiques, les distances à parcourir en ville, etc.).

3. COMBINAISON DES DONNÉES D'ENTRÉE ET OPTIMISATION DES RÉSULTATS

Devant un ensemble de données si complexe et un sujet si difficile et si polyvalent, il faut tâcher de réduire les flux d'information, tout en augmentant le contenu informatif de chaque facteur qui intervient dans les analyses,

tant en entrée qu'en sortie. Il y a plusieurs facons de procéder. Nous avons choisi les suivantes :

A) Construction de variables d'entrée complexes, afin de s'en tenir à l'étude des facteurs les plus significatifs.

(i) On a produit des variables complexes discontinues en combinant les consommations d'énergie à la maison et pour le transport, mais en ne considérant pas ces deux catégories de dépenses comme des variables continues. On a pris plutôt des tranches de ces dernières afin de former des groupes types de consommateurs, dont on peut ensuite étudier les caractéristiques pour mettre en lumière leurs différences. Pour créer ces groupes, on a divisé les dépenses des deux catégories en quartiles, on a combiné les deux quartiles centraux, puis on a formé une matrice à neuf cases à partir des deux ensembles de trois tranches résultantes (voir le tableau 1; source : McDougall, Ritchie et Claxton). Il est intéressant de comparer entre elles et avec la case centrale les cases qui se trouvent aux angles de cette matrice. Ces groupes types ont été utilisés dans le cadre d'une étude antérieure réalisée pour le ministère de la Consommation et des Corporations; la comparaison entre les résultats obtenus avec les données sur les dépenses des familles et les autres ensembles de données utilisés dans l'étude du MCC se révèle donc spécialement intéressante. On voudra aussi étudier comment les caractéristiques de ces groupes évoluent avec le temps. Par exemple, est-ce que la catégorie des "souris d'église" demeure formée de Canadiens à faible revenu (frugalité involontaire), ou décèle-t-on une tendance vers l'adoption de modes de vie moins énergivores. Au tableau 2, on compare les caractéristiques de trois groupes types (souris d'église, lièvres et hippopotames) pour deux années différentes. On donne d'abord, dans le cas des "souris d'église", des renseignements relatifs à une gamme possible de variables d'analyse, pour les années 1974 et 1978. Afin de simplifier la présentation, seul le rang de la case parmi l'ensemble des groupes types est indiqué. Traditionnellement, ceux qui consomment le moins d'énergie sont ceux qui, en général, sont les plus dépourvus: revenus les plus bas, faible niveau d'instruction, âge avancé. On peut voir que les "souris d'église" possèdent bel et bien ces caractéristiques en 1974. Ce groupe type se situe également au niveau le plus faible pour toutes les catégories de dépenses indiquées. Bien qu'on y retrouve les personnes les plus âgées, le groupe ne vient pas au rang le plus bas pour le nombre d'enfants

très jeunes. Cela tient sans doute au fait que le groupe inclut à la fois des personnes âgées et des ménages monoparentaux (dont les chefs de famille sont probablement des femmes) comprenant de jeunes enfants. On notera aussi que ce groupe possède le moins grand nombre de salariés à plein temps. En 1978, la situation générale demeure la même, sauf que le groupe n'est plus le plus âgé. En fait, le groupe le plus âgé est celui de la case située immédiatement à droite dans la matrice des groupes types. Il semble donc qu'en 1978, les personnes les plus âgées consomment une quantité relativement plus grande d'énergie à la maison. Peut-être ce groupe était-il en meilleure position financière en 1978 qu'en 1974, mais peut-être aussi a-t-il été incapable de contenir les dépenses consacrées à l'énergie en raison de la hausse des prix.

En 1974, les "hippopotames" possèdent aussi les caractéristiques prévues. Il semble qu'il s'agisse de personnes d'âge moyen ayant un grand nombre d'enfants de cinq à seize ans, les familles ayant atteint leur taille maximale. Ce sont eux qui dépensent le plus dans la plupart des catégories. Ils ont en outre le niveau d'instruction le plus élevé. En 1978, toutefois, ce groupe ne se situe plus au sommet sur le plan de l'instruction, ni sur celui des dépenses consacrées au logement. La raison en est peut-être que ceux qui ont les plus grandes maisons et qui sont les plus instruits ont commencé à modifier leurs habitations en vue de réaliser des économies d'énergie.

On note également une évolution chez les "lièvres". En 1974, ils formaient le groupe le plus jeune, avec de très petites familles. En 1978, ce groupe semble être constitué davantage de jeunes familles ayant de jeunes enfants. Le changement le plus marqué chez ce groupe est la baisse considérable des dépenses consacrées au tabac et aux boissons alcooliques.

Le niveau de signification de ces changements peut être évalué à l'aide de tests statistiques appropriés. Cette analyse avait pour but d'isoler certains groupes types dans la population. On peut ainsi mettre en lumière certaines caractéristiques des modes de vie qui sont très utiles pour adapter les programmes d'économie d'énergie à chaque groupe.

D'autres analyses pourraient porter non pas sur les niveaux des dépenses, mais sur les pourcentages correspondants. Cela permettrait de dresser le profil de ceux pour qui les dépenses consacrées à l'énergie représentent le fardeau le plus lourd.

(ii) On peut produire des variables d'entrée complexes continues qui

éliminent les effets de variables dont le poids est très lourd, mais qui échappent en grande partie ou totalement au contrôle des consommateurs.

Les dépenses consacrées à l'énergie à la maison peuvent être examinées en fonction de certains facteurs, mais comme l'un des principaux facteurs qui conditionnent ces dépenses est la taille de la maison, cet élément peut être absorbé dans la variable d'entrée pour permettre l'examen d'autres facteurs plus intéressants (du point de vue de l'élaboration des politiques). Ainsi, au lieu d'étudier les dépenses totales consacrées à l'énergie, on s'intéresse aux dépenses par pièce. En poussant un peu plus loin, on peut éliminer l'effet des variations climatiques et météorologiques d'une année à l'autre en examinant les dépenses/pièce/degré-jour. Les degrés-jours sont pris en compte par ville et par année. Les données portant sur les degrés jours par année et pour chaque ville furent obtenues d'Environnement Canada. Le tableau 3 montre comment les chiffres varient à mesure que la complexité du facteur étudié s'accroît, encore une fois pour deux années. La comparaison des deux années et les différences dans les taux de changements d'une année à l'autre suggèrent l'importance de raffiner la mesure afin d'améliorer la compréhension du phénomène.

B) Construction de variables de sortie sommaires pour l'examen de la structure des données - exemple de coefficients de régression.

On présente aux tableaux 4 à 6 certains résultats de l'analyse de régression. Trois modèles sont examinés. D'un modèle à l'autre, la variable dépendante devient plus complexe. Il est ainsi possible de contrôler les facteurs qui ont une incidence marquée sur la consommation d'énergie et de mieux voir dans quelle mesure les autres variables peuvent intervenir de facon significative.

On n'a pas tenté dans ces analyses d'aborder le problème du plan de sondage complexe. Une telle étude sera effectuée plus tard au moyen de la technique de linéarisation de Taylor, et les résultats seront comparés. On présente toutefois, pour l'année 1974, les résultats obtenus avec un échantillon pondéré et un échantillon non pondéré. Comme on peut le constater, les valeurs des coefficients changent très peu et leur degré de signification demeure le même. À cause des contraintes mentionnées et du fait que des différences très légères peuvent produire des résultats significatifs en raison de la très grande taille des échantillons, ces résultats préliminaires doivent être

examinés avec la plus grande prudence. Pour les besoins de cet article, nous ne noterons que les variables significatives au niveau de 0.01 et au-delà, et nous n'en indiquerons que le signe.

Des variables fictives sont incorporées à la liste des variables indépendantes dans le premier et le second modèle pour la ville, et dans les trois modèles pour le type de logement. Les variables non indiquées sont, en ce qui concerne la ville, Ottawa, et en ce qui concerne le type de logement, la maison individuelle.

En 1974, la taille du logement, certaines villes, les dépenses totales, l'âge du chef et la taille de la famille, ainsi que certains types de logements, sont des variables significatives. Les grandes familles dont les dépenses totales sont élevées et qui vivent à Saint-Jean (T.-N.) dans des maisons individuelles sont celles qui consomment le plus. La consommation est moins forte dans les villes de l'ouest que dans celles de l'est. Elle est par ailleurs plus forte dans les maisons individuelles que dans tous les autres types de logements, bien que pour les duplex, la différence ne soit pas significative lorsqu'on tient compte du nombre de pièces. Les résultats non pondérés sont semblables aux résultats pondérés.

Lorsqu'on prend les dépenses par pièce comme variable dépendante et qu'on retranche le nombre de pièces de la liste des variables indépendantes, le schéma global demeure le même. Toutefois, la taille de la famille n'est plus une variable significative (elle est probablement étroitement liée à la taille du logement seulement), et le niveau d'instruction du chef de famille devient significatif, avec coefficient négatif, c'est-à-dire que moins les personnes sont instruites, plus elles consomment, toutes choses étant égales par ailleurs. Enfin, le duplex devient une variable significative avec coefficient positif, c'est-à-dire qu'une fois soustrait l'effet du nombre de pièces, il se consomme plus d'énergie dans les duplex que dans les maisons individuelles.

Dans le modèle 3, on tient compte des conditions climatiques en incorporant les degrés-jours dans la variable dépendante et en retranchant la liste des villes de l'ensemble des variables indépendantes. Il convient de noter que si la valeur des coefficients baisse de façon si prononcée, c'est parce que le nombre de degrés-jours dans ces villes se situe entre 4000 et 7000. La faible valeur des coeficients ne signifie pas, par conséquent, qu'ils ne sont pas importants. Les dépenses totales demeurent un facteur significatif, tout

comme l'instruction du chef de famille et la variable maison en rangée. Un fait important à noter est la baisse de la valeur du coefficient R^2 ajusté. En fait, les variables indépendantes qui demeurent dans l'équation ne contribuent pas beaucoup à expliquer la variation de la variable dépendante. Il faudrait chercher d'autres variables plus utiles.

En comparant les résultats non pondérés de 1974 et de 1978, dans le cas du modèle 1, on note un changement au niveau du paramètre Vancouver, ainsi qu'une variation de l'importance des maisons jumelées et des duplex par rapport aux maisons individuelles.

Dans le cas du modèle 2, le changement principal réside encore une fois dans l'effet des types de logement.

Enfin, dans le cas du modèle 3, seules les maisons en rangée se démarquent des maisons individuelles. Le niveau d'instruction du chef de famille est encore important, mais, en 1978, l'âge du chef est significatif avec un coefficient positif. On observe une certaine amélioration du coefficient R^2 pour 1978, mais celui-ci demeure très faible.

Cette comparaison entre deux années, du point de vue de l'élaboration des politiques, laisse croire que la qualité des maisons individuelles s'est améliorée au Canada. Sur le plan des méthodes, la comparaison montre l'importance d'un choix judicieux de la variable dépendante.

Comme on l'a mentionné précédemment, on procédera à de nombreuses autres analyses au moyen des techniques de régression disponibles afin de raffiner ces résultats et de tenir compte du plan de sondage.

Les données sur les dépenses des familles, on l'a vu, ont leurs limites, mais elles contiennent une manne de renseignements importants qu'il vaut la peine d'explorer.

RÉFÉRENCE

[1] McDougall, Gordon H.G., Ritchie, J.R. Brent et Claxton, John D. (1979), Energy Conservation and Conservation Patterns in Canadian Households: Overview, Behavioral Energy Research Group, 203-2053 Main Mall, University of British Columbia.

Tableau 1: Groupes types de consommateurs d'énergie

		Consommation d'énergie à la maison			
		Faible 127 mil. kJ	Moyenne 127-222 mil. kJ	Élevée 222 mil. kJ	Total
Consommation d'essence pour automobiles	Faible 1136 litres	SOURIS D'ÉGLISE 4.5% de l'échantillon	9.8% de l'échantillon	OURS 2.5% de l'échantillon	16.8
	Moyenne 1136-4545 litres	14.5% de l'échantillon	CASTOR 33.7% de l'échantillon	12.3% de l'échantillon	60.5
	Élevée 4546 litres	LIÈVRE 4.0% de l'échantillon	12.6% de l'échantillon	HIPPOPOTAME 6.1% de l'échantillon	22.7
	Total	23.0	56.1	20.9	100.0

Source : Voir la référence à la fin du texte.

Tableau 2: Rang parmi les groupes types

	Souris d'église		Hippopotames		Lièvres	
	1974	1978	1974	1978	1974	1978
Instruction du chef (faible à élevée)	1	1	9	7	7	8
Âge (âgé à jeune)	1	2	6	6	9	9
Salariés à plein temps (faible à élevé)	1	1	8.5	9	7	6.5
Taille de la famille (faible à élevée)	1	1	9	9	4	4
Enfants de moins de cinq ans (faible à élevé)	3	1	4	2	1.5	7
Enfants de cinq à quinze ans (faible à élevé)	1	2.5	7	6.5	5	2.5
Aliments achetés à des magasins (faible à élevé)	1	1	9	9	4	4
Aliments consommés à des restaurants, etc. (faible à élevé)	1	1	9	9	6	6
Logement (faible à élevé)	1	1	9	7	4	3
Habillement (faible à élevé)	1	1	9	9	6	5
Soins personnels (faible à élevé)	1	1	9	9	5	4
Frais médicaux (faible à élevé)	1	1	8	8	4	4
Tabac et alcools (faible à élevé)	1	1	9	9	7	4
Lecture, loisirs, éducation (faible à élevé)	1	1	9	8	8	9

Tableau 3: Dépenses moyennes consacrées à l'énergie à la maison

	1974	1978	Changement en %
Dépenses moyennes ($) consacrées à l'énergie de la maison	451	764	+69
Dépenses moyennes par pièce ($) consacrées à l'énergie à la maison	73	121	+66
Dépenses moyennes par pièce et degré -jours consacrées à l'énergie à la maison	.019	.029	+53

Tableau 4: Résultats de l'analyse de régression
Modèle 1 - Dépenses consacrées à l'énergie à la maison.

	1974 Non pondéré	1974 Pondéré	1978 Non pondéré
Coordonnée à l'origine	197.3 A	225.4 A	298.0 A
Nombre de pièces	13.9 A	12.0 A	4.2 C
Ville - Saint-Jean (T.-N.)	193.9 A	204.9 A	341.1 A
Halifax	75.5 A	73.9 B	162.0 A
Montréal	12.2	22.7	-16.6
Toronto	-10.2	-3.0	50.5
Winnipeg	-127.1 A	-125.4 A	-72.2 C
Edmonton	-244.9 A	-243.2 A	-195.8 A
Vancouver	-22.9	-17.5	-71.9 C
Dépenses totales	.006 A	.006 A	.01A
Âge du chef de famille	1.2 A	0.8 B	3.6 A
Taille de la famille	13.2 A	12.1	21.6 B
Instruction du chef de famille	0.7	0.6	-3.6
Type de logement -			
maisons jumelées	-50.9 B	-49.0 A	-23.8
maisons en rangée	-81.2 A	-88.9 A	-119.7 B
duplex	-12.3	-13.7	-84.6 C
R^2 ajusté =	0.43	0.34	0.38
F =	118.5(.0001)	79.7(.0001)	74.6(.0001)

Notez: A = probabilité inférieure à 0.0001, B = probabilité inférieure à 0.001, C = probabilité inférieure à 0.01

Tableau 5: Résultats de l'analyse de régression
Modèle 2 - Dépenses par pièce

	1974 Non pondéré	1974 Pondéré	1978 Non pondéré
Coordonnée à l'origine	76.2 A	77.3 A	99.8 A
Ville - Saint-Jean (T.-N.)	30.4 A	32.0 A	74.8 A
Halifax	16.8 A	16.3 B	31.6 A
Montréal	4.5	6.7	6.5
Toronto	-3.5	-1.7	10.1
Winnipeg	-17.6 A	-16.3 A	-0.9
Edmonton	-37.9 A	-36.8 A	-26.4 A
Vancouver	0.3	0.8	-6.7
Dépenses totales	2.2×10^{-4} B	2.5×10^{-4} B	6.9×10^{-4} A
Âge du chef de famille	0.015	-0.03	0.33 B
Taille de la famille	0.6	0.04	-0.63
Instruction du chef de famille	-1.9 A	-1.4 B	-4.0 A
Type de logement -			
maisons jumelées	-6.5 C	-7.1 B	3.1
maisons en rangée	-11.5 A	-11.8 A	-11.0
duplex	6.1 C	6.6 C	3.24
R^2 ajusté =	.31	.19	.24
F =	73.85(.0001)	38.9(.0001)	41.4(.0001)

Notez: A = probabilité inférieure à 0.0001, B = probabilité inférieure à 0.001
C = probabilité inférieure à 0.01

Tableau 6: Résultats de l'analyse de régression
Modèle 3 - Dépenses par pièce par degré-jour

	1974 Non pondéré	1974 Pondéré	1978 Non pondéré
Coordonnée à l'origine	.017 A	.019 A	.02 A
Dépenses totales	8.01×10^{-8} B	9.4×10^{-8} A	1.4×10^{-7} A
Âge du chef de famille	1.8×10^{-5}	-7.0×10^{-6}	9.9×10^{-5} A
Taille de la famille	-1.4×10^{-5}	-18.4×10^{-5}	27.0×10^{-5}
Instruction du chef de famille	-5.3×10^{-4} A	-4.7×10^{-4} A	-7.8×10^{-4} B
Type de logement -			
maisons jumelées	3.4×10^{-4}	-7.5×10^{-4}	24.8×10^{-4}
maisons en rangée	-23×10^{-4} C	-35.9×10^{-4} A	-38.8×10^{-4} B
duplex	16.9×10^{-4}	6.3×10^{-4}	11.6×10^{-4}
R^2 ajusté =	.01	.02	0.03
F =	5.6(.0001)	6.6(.0001)	9.5(.0001)

Notez: A = probabilité inférieure à 0.0001, B = probabilité inférieure à 0.001
C = probabilité inférieure à 0.01

RÉGRESSION LOGISTIQUE ET ANALYSE DE DONNÉES DE L'ENQUÊTE SUR LA POPULATION ACTIVE

S. Kumar et J.N.K. Rao[1]

Les tests habituels du khi-carré (X^2) et du rapport de vraisemblance (G^2) pour l'analyse de régressions logistiques comprenant une variable binaire sont rectifiés en fonction du plan de sondage. Ces corrections reposent sur certains effets généralisés du plan de sondage. Les variables modifiées sont utilisées pour analyser des données de l'enquête sur la population active au Canada (EPA) menée en octobre 1980. La variable de Wald, qui permet aussi de prendre en compte le plan de sondage, est également examinée pour vérifier la qualité de l'ajustement du modèle et des hypothèses concernant les paramètres du modèle théorique. À l'aide de données de l'EPA, des indices diagnostiques des régressions logistiques sont appliqués à la recherche des proportions extrêmes dans les cases d'un tableau et des points prépondérants dans l'espace factoriel, une fois effectuées les corrections nécessaires pour tenir compte du plan de sondage.

1. INTRODUCTION

Les spécialistes des sciences sociales, des sciences du comportement et des sciences de la santé utilisent beaucoup les modèles de régression logistique pour étudier la variation de proportions binomiales (voir, par exemple, les ouvrages de Cox (1970) et McCullagh et Nelder (1983)). Mais quand un plan de sondage comprend des grappes et des strates, les méthodes statistiques applicables aux proportions binomiales sont souvent inadéquates pour l'analyse des données d'enquête. Par exemple, les tests habituels du khi-carré (X^2) ou du rapport de vraisemblance (G^2) font beaucoup grossir le risque de première espèce (seuil de signification). Il est donc nécessaire de corriger les méthodes classiques en fonction du plan de sondage si on veut faire des inférences valables à partir des données d'une enquête. Dans le présent exposé,

[1] S. Kumar, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada et J.N.K. Rao, département des mathématiques et de la statistique, Université Carleton.

nous recourons à deux corrections simples de X^2 et de G^2 qui reposent sur certains effets généralisés du plan de sondage. pour analyser des données de l'enquête sur la population active au Canada (EPA) menée en octobre 1980 (section 3). On examine également la variable de Wald qui renferme aussi une correction pour tenir compte du plan de sondage.

En plus de formuler des tests statistiques généraux, il est essentiel de préciser des mécanismes diagnostiques pour déceler les proportions extrêmes dans les cases d'un tableau et les points prépondérants dans l'espace factoriel. La recherche de variables diagnostiques pour le modèle linéaire classique a été décrite en détail dans les ouvrages statistiques (voir le livre récent de Cook et Weisberg (1982)). Pregibon (1981) a proposé dernièrement des méthodes semblables pour la régression logistique avec des proportions binomiales. À la section 4, quelques-unes de ces techniques sont appliquées aux données de l'EPA d'octobre 1980 une fois effectuées les corrections nécessaires pour tenir compte du plan de sondage.

2. RÉSULTATS THÉORIQUES

Supposons que la population mère est répartie en I cases (domaines) en fonction du niveau d'un facteur ou plus et que $\hat{N}_i$ est l'estimation de la taille du $i^{\text{ième}}$ domaine $N_i (i = 1, 2, \ldots, I: \Sigma N_i = N)$ évaluée à l'aide des données d'une enquête. L'estimation correspondante du total d'une variable binaire N_{i1} (0, 1) pour le $i^{\text{ième}}$ domaine est représentée par N_{i1}. Le quotient $\hat{p}_i = \hat{N}_{i1}/\hat{N}_i$ sert à estimer la proportion $\pi_i = N_{i1}/N_i$ à l'échelle de la population.

Un modèle logit des proportions π_i a la forme $\pi_i = f_i(\underset{\sim}{\beta})$, où

$$\ln\{f_i/(1 - f_i)\} = \text{logit } f_i = \underset{\sim}{x}_i'\underset{\sim}{\beta}\ , \ i = 1, \ldots, I. \tag{1}$$

Dans l'équation (1), $\underset{\sim}{x}_i$ et un vecteur de dimension s de constantes connues calculées à partir des niveaux des facteurs et $\underset{\sim}{\beta}$ est un vecteur de dimension s de paramètres inconnus. Si un échantillon binomial indépendant est prélevé dans chaque domaine, les estimations du maximum de vraisemblance (EMV) pro-

viennent des solutions aux équations de vraisemblance suivantes:

$$X'D(\underset{\sim}{n}/n)\hat{\underset{\sim}{f}} = X'D(\underset{\sim}{n}/n)\hat{\underset{\sim}{q}}, \tag{2}$$

où $X' = (\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_I)$, $D(\underset{\sim}{n}/n) = \text{diag}\ (n_1/n, \ldots, n_I/n)$, $\hat{\underset{\sim}{f}} = \underset{\sim}{f}(\hat{\underset{\sim}{\beta}}) = (\hat{f}_1, \ldots, \hat{f}_I)'$, et $\hat{\underset{\sim}{q}}$ est le vecteur des proportions de l'échantillon $q_i = n_{i1}/n_i$, où n_i est la taille de l'échantillon tiré du $i^{\text{ième}}$ domaine ($\Sigma\, n_i = n$). Pour les plans de sondage généraux, il n'y a pas d'EMV à cause de difficultés relatives à la formulation de fonctions de vraisemblance convenables. On utilise souvent des "pseudo-EMV" de $\underset{\sim}{\beta}$ ou de $\underset{\sim}{f}$ en remplaçant n_i/n dans l'équation (2) par l'estimation de la taille relative du domaine, $w_i = \hat{N}_i/\hat{N}$, et $\hat{q}_i$ par l'estimation $\hat{p}_i$ calculée à partir des données d'enquête:

$$X'D(\underset{\sim}{w})\hat{\underset{\sim}{f}} = X'D(\underset{\sim}{w})\hat{\underset{\sim}{p}}. \tag{3}$$

Les estimations qui proviennent de cette équation, $\hat{\underset{\sim}{\beta}}$ et $\hat{\underset{\sim}{f}} = \underset{\sim}{f}(\hat{\underset{\sim}{\beta}})$, sont asymptotiquement (c'est-à-dire dans de grands échantillons) convergentes. On peut également récrire l'équation (3) sous la forme suivante:

$$X'\hat{\underset{\sim}{N}}_1(m) = X'\hat{\underset{\sim}{N}}_1, \tag{4}$$

où $\hat{\underset{\sim}{N}}_1$ est le vecteur des totaux estimés $\hat{N}_{i1}$ et $\hat{\underset{\sim}{N}}_1(m)$ est le vecteur de pseudo-EMV, $\hat{N}_{i1}(m) = \hat{N}_i\hat{f}_i$, des totaux N_{i1}. On obtient les valeurs de l'estimateur $\hat{\underset{\sim}{\beta}}$, et, par conséquent, celles de $\hat{\underset{\sim}{f}}$ et de $\hat{\underset{\sim}{N}}_1(m)$, au moyen de l'équation (3) ou (4) par une série de calculs itératifs.

2.1 Estimation des variances et des covariances

Soit $\hat{V}$ la matrice des covariances estimées de $\hat{\underset{\sim}{p}}$; la matrice des covariances estimées de $\hat{\underset{\sim}{\beta}}$ est

$$\hat{D}(\hat{\underset{\sim}{\beta}}) = (X'\hat{\Delta}X)^{-1}(X'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})X)(X'\hat{\Delta}X)^{-1} \tag{5}$$

dans les grands échantillons, où $\hat{\Delta} = \text{diag}\ (w_1\ \hat{f}_1(1 - \hat{f}_1),\ \ldots,\ w_I \hat{f}_I(1 - \hat{f}_I))$. Les éléments diagonaux de la matrice définie en (5) sont les estimations de la variance des estimations $\hat{\beta}_i$. De même, la matrice des covariances estimées du vecteur des résidus $\underset{\sim}{r} = \hat{\underset{\sim}{p}} - \hat{\underset{\sim}{f}}$ est

$$\hat{D}(\underset{\sim}{r}) = A\hat{V}A'. \tag{6}$$

où

$$A = I - D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})X(X'\hat{\Delta}X)^{-1}X'D(\underset{\sim}{w}). \tag{7}$$

Les éléments diagonaux $\hat{V}_{ii}(\underset{\sim}{r})$ de la matrice définie en (6) permettent de calculer les résidus centrés réduits r_i/e.-t.(r_i), qui sont utiles pour déceler les cases contenant des proportions extrêmes.

2.2 Tests de la validité de l'ajustement

La variable khi-carré normalement utilisée pour vérifier la validité de l'ajustement du modèle (1) est

$$X^2 = n \sum_{i=1}^{I} \frac{(\hat{p}_i - \hat{f}_i)^2\, w_i}{\hat{f}_i(1 - \hat{f}_i)} = \sum_{i=1}^{I} X_i^2. \tag{8}$$

Le test du rapport de vraisemblance repose sur la fonction

$$G^2 = 2n \sum_{i=1}^{I} w_i \left\{ \hat{p}_i \ln \frac{\hat{p}_i}{\hat{f}_i} + (1 - \hat{p}_i) \ln \frac{(1 - \hat{p}_i)}{(1 - \hat{f}_i)} \right\} = \sum_{i=1}^{I} G_i^2. \tag{9}$$

Soulignons le fait que la variable G_i^2 est définie quand $\hat{p}_i = 0$ et 1 et vaut $-2nw_i \ln(1 - \hat{f}_i)$ et $-2nw_i \ln \hat{f}_i$ respectivement. Dans l'échantillonnage binomial indépendant, il est bien connu que X^2 et G^2 suivent asymptotiquement une loi

de χ^2 à I - s degrés de liberté, mais pour les plans de sondage généraux, cette propriété n'est pas valable. En fait, X^2 (ou G^2) a la distribution asymptotique d'une somme pondérée, $\Sigma\ \delta_i Z_i$, de variables χ^2, Z_i, ayant chacune un degré de liberté, où les poids δ_i (i = 1, .., I - s) sont les valeurs propres d'une matrice des "effets généralisés du plan de sondage" calculée à partir de l'expression $\Sigma_0^{-1}\ \Sigma_\phi$, où

$$\Sigma_\phi = G'D(\hat{\underset{\sim}{f}})^{-1}D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})^{-1}\hat{V}D(\hat{\underset{\sim}{f}})^{-1}D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})^{-1}G, \tag{10}$$

$$\Sigma_0 = \frac{1}{n}\ G'\hat{\Delta}^{-1}G \tag{11}$$

et G est une matrice quelconque I×(I - s) de rang I - s pour laquelle G'X = 0: autrement dit, la matrice G est orthogonale par rapport à X. Dans un échantillon binomial, $\Sigma_0^{-1}\ \Sigma_\phi$ se ramène à I, la matrice unité.

On peut apporter une modification simple à X^2 ou G^2 (Roberts, 1984) en considérant $X_c^2 = X^2/\delta.$ ou $G_c^2 = G^2/\delta.$ comme une variable χ^2 à I - s degrés de liberté si le modèle est vrai, où

$$(I - s)\delta. = n \sum_{i=1}^{I} \hat{V}_{ii}(r)w_i/[\hat{f}_i(1 - \hat{f}_i)]. \tag{12}$$

La variable corrigée X_c^2 ou G_c^2 devrait être acceptable, sauf dans les cas où les δ_i ont un grand coefficient de variation (CV). Dans une meilleure solution, qui repose sur l'approximation de Satterthwaite, la variable $X_S^2 = X_c^2(1 + a^2)$ ou $G_S^2 = G_c^2/(1 + a^2)$ est considérée comme une variable χ^2 à $(I - s)/(1 + a^2)$ degrés de liberté, où

$$a^2 = \Sigma\ (\delta_i - \delta_.)^2/[(I - s)\delta_.^2] \tag{13}$$

est le carré du CV des δ_i et

$$\Sigma\ \delta_i^2 = \sum_{i=1}^{I} \sum_{j=1}^{I} \hat{V}_{ij}^2(r)(nw_i)(nw_j)\ /\ [\hat{f}_i\hat{f}_j(1 - \hat{f}_i)(1 - \hat{f}_j)]\ , \tag{14}$$

où $\hat{V}_{ij}(r)$ est le $(i, j)^{\text{ième}}$ élément de $\hat{D}(r)$. Les variables X_S^2 et G_S^2 renferment une correction pour tenir compte de la variation des δ_i.

Une variable de Wald pour vérifier la validité de l'ajustement du modèle (1) est

$$X_W^2 = \hat{\underset{\sim}{\nu}}' G \Sigma_{\hat{\phi}}^{-1} G' \hat{\underset{\sim}{\nu}}, \tag{15}$$

où $\hat{\underset{\sim}{\nu}}$ est le vecteur de logits $\hat{\nu}_i$ = logit $\hat{p}_i$. La variable X_W^2 suit une loi de χ^2 à I - s degrés de liberté dans les grands échantillons. La variable X_W^2 n'est pas définie si $\hat{p}_i = 0$ ou 1 pour une valeur de i. En outre, cette fonction devient instable lorsqu'une valeur de $\hat{p}_i$ est proche de un (voir la section 3) ou quand le nombre de degrés de liberté de $\hat{V}$ n'est pas grand en comparaison de I - s (Fay, 1983).

2.3 Hypothèses emboîtées

Supposons que la matrice X comprend deux partitions (X_1, X_2), où X_1 est de dimension I×r et X_2 de dimension I×u (r + u = s); le modèle (1) peut s'écrire

$$\underset{\sim}{\nu} = X\underset{\sim}{\beta} = X_1\underset{\sim}{\beta}_1 + X_2\underset{\sim}{\beta}_2, \tag{16}$$

où $\underset{\sim}{\beta}_1$ est de dimension r×1 et $\underset{\sim}{\beta}_2$ de dimension u×1. Nous voulons souvent vérifier l'hypothèse nulle H: $\underset{\sim}{\beta}_2 = 0$ dans le modèle (16). Les pseudo-EMV pour H proviennent du système d'équations

$$X_1' D(w) \hat{\hat{\underset{\sim}{f}}} = X_1' D(w) \hat{\underset{\sim}{p}} \tag{17}$$

où $\hat{\hat{\underset{\sim}{f}}} = f(\hat{\hat{\underset{\sim}{\beta}}})$ et on obtient encore une fois les solutions par une série de calculs itératifs. Les variables habituelles des tests du khi-carré et du rapport de vraisemblance de H: $\underset{\sim}{\beta}_2 = 0$ sont

$$X^2(2|1) = n \sum_{i=1}^{I} \frac{w_i(\hat{f}_i - \hat{\hat{f}}_i)^2}{\hat{\hat{f}}_i(1 - \hat{\hat{f}}_i)} \tag{18}$$

et

$$G^2(2|1) = 2n \sum_{i=1}^{I} w_i \left\{ \hat{f}_i \ln \frac{\hat{f}_i}{\hat{\hat{f}}_i} + (1 - \hat{f}_i) \ln \frac{(1 - \hat{f}_i)}{(1 - \hat{\hat{f}}_i)} \right\} \tag{19}$$

respectivement. Dans l'échantillonnage binomial, $X^2(2|1)$ et $G^2(2|1)$ suivent asymptotiquement une loi de χ^2 à u degrés de liberté quand l'hypothèse H est vraie, mais, dans les plans de sondage généraux, cette propriété n'existe pas. En fait, $X^2(2|1)$ ou $G^2(2|1)$ a asymptotiquement la même distribution qu'une somme pondérée, $\Sigma\delta_i(H)Z_i$, de variables indépendantes Z_i de type χ_1^2, où les poids $\delta_i(H)$ $(i = 1, \ldots, u)$ sont les valeurs propres de la matrice des effets du plan de sondage.

$$(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}(\tilde{X}_2'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})\tilde{X}_2), \tag{20}$$

où

$$\tilde{X}_2 = [I - X_1(X_1'\Delta X_1)^{-1}X_1'\Delta]X_2, \tag{21}$$

(Roberts, 1984). Dans un échantillon binomial, la matrice des effets du plan (20) se ramène à I, comme dans le test de la validité de l'ajustement à la section précédente.

On peut obtenir une simple modification de $X^2(2|1)$ ou $G^2(2|1)$ en considérant $X_c^2(2|1) = X^2(2|1)/\delta.(H)$ ou $G_c^2(2|1)/\delta.(H)$ comme suivant une loi de χ^2 à u degrés de liberté si l'hypothèse H est vraie, où

$$u\ \delta.(H) = n \sum_{i=1}^{I} \tilde{V}_{ii}(r) w_i / \hat{\hat{f}}_i(1 - \hat{\hat{f}}_i) \tag{22}$$

et $\tilde{V}_{ii}(r)$ est le $i^{\text{ième}}$ élément de la diagonale de la matrice des covariances des résidus, $r_i(H) = \hat{f}_i - \hat{\hat{f}}_i$, définie par l'expression

$$\tilde{V}(r) = D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})\tilde{X}_2A\tilde{X}_2'D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}}), \tag{23}$$

où

$$A = (\tilde{X}_2'\Delta\tilde{X}_2)^{-1}[\tilde{X}_2'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})\tilde{X}_2](\tilde{X}_2'\Delta\tilde{X}_2)^{-1}. \tag{24}$$

Les résidus centrés réduits $(\hat{f}_i - \hat{\hat{f}}_i)/[\tilde{V}_{ii}(r)]^{\frac{1}{2}}$ peuvent aussi être calculés. Comme dans le test de la validité de l'ajustement, on peut également produire une meilleure approximation à l'aide de la méthode de Satterthwaite.

Une variable de Wald pour H: $\underset{\sim}{\beta}_2 = 0$ est définie par l'équation

$$X_W^2(2|1) = \hat{\beta}_2'[\hat{D}(\hat{\underset{\sim}{\beta}}_2)]^{-1}\hat{\underset{\sim}{\beta}}_2, \tag{25}$$

où $\hat{D}(\hat{\underset{\sim}{\beta}}_2)$ est la sous-matrice principale de (5) qui correspond à $\hat{\underset{\sim}{\beta}}_2$. En vertu de l'hypothèse H, $X^2(2|1)$ suit asymptotiquement une loi de χ^2 à u degrés de liberté. En particulier, si β_2 est scalaire, on peut considérer $\hat{\beta}_2/\text{e.t.}(\hat{\beta}_2)$ comme une variable de type N(0,1) si l'hypothèse H: $\beta_2 = 0$ est vraie ou $\hat{\beta}_2^2/\text{Var}(\hat{\beta}_2)$ comme une variable χ^2 à un degré de liberté.

2.4 Indices diagnostiques

Il est souhaitable de faire une évaluation critique de l'ajustement du modèle logit en repérant les cases contenant des proportions extrêmes et les points prépondérants dans l'espace factoriel. Pour ce faire, le vecteur des résidus et une matrice-projection de l'espace factoriel sont des ressources utiles. Toutefois, contrairement au modèle linéaire classique, ce modèle produit des résidus qui peuvent être définis sur différentes échelles. Le choix naturel qui permet de prendre en compte le plan de sondage est le vecteur des résidus centrés réduits $e_i = r_i/[\hat{V}_{ii}(r)]^{\frac{1}{2}}$ présenté à la section 2.1. Étant donné que les e_i ont approximativement une distribution de type N(0,1) dans le modèle (1), le nombre de résidus e_i dont la valeur absolue dépasse 1.96, 2.33 et 2.58 a pour espérance mathématique 0.05I, 0.02I et 0.01I respectivement, où I est le nombre de résidus (cases). Ces probabilités théoriques offrent des critères généraux pour déceler les valeurs extrêmes. Si on fait abstraction du plan de sondage et qu'on utilise les résidus centrés réduits dans un échantillon binomial, on risque de tirer des conclusions erronées.

Par contre, les résidus centrés réduits e_i perdent leur fiabilité si $\hat{p}_i$ est égal à un ou est proche de un. À l'instar de Pregibon (1981), nous proposons l'utilisation de composantes de X_c^2 ou G_c^2, notamment $\tilde{X}_i = X_i/\delta_\bullet^{\frac{1}{2}}$ ou $\tilde{G}_i = G_i/\delta_\bullet^{\frac{1}{2}}$, $i = 1, \ldots, I$, pour analyser les résidus en contournant cette difficulté. Dans un cas ou dans l'autre, les valeurs élevées de ces composantes devraient donner une idée d'où se trouvent les cases mal expliquées par le modèle. Il est utile d'examiner ces composantes à l'aide de graphiques diagnostiques (c'est-à-dire des courbes illustrant la relation entre $\tilde{X}_i$ et i et entre $\tilde{G}_i$ et i). Les représentations de $\tilde{X}_i$ ou $\tilde{G}_i$ sur une échelle gaussienne (qui montrent la répartition de valeurs ordonnées parmi les quantiles de la loi normale centrée réduite) est un autre moyen utile de voir si le modèle est exact (c'est-à-dire si les points forment tous une ligne droite).

Pregibon (1981) a proposé l'utilisation des éléments diagonaux m_{ii} de la matrice-projection

$$M = I - \hat{V}_b^{\frac{1}{2}} X(X'\hat{V}_b X)^{-1} X'\hat{V}_b^{\frac{1}{2}}$$

$$= I - H \text{ (par example)} \qquad (26)$$

pour déceler les points prépondérants, où $\hat{V}_b$ est la matrice des covariances estimées dans un échantillon binomial: autrement dit, on peut utiliser $\mathrm{diag}[\hat{p}_1(1 - \hat{p}_1)/(nw_1), \ldots, \hat{p}_I(1 - \hat{p}_I)/(nw_I)]$ pour des données d'enquête. La matrice M est normalement nécessaire pour résoudre les équations de vraisemblance (4) par la méthode des moindres carrés itérativement pondérés et les petites valeurs de m_{ii} indiquent les points extrêmes dans l'espace factoriel. Ici encore, un graphique diagnostique (de m_{ii} en fonction de i) serait utile. On peut souligner le fait qu'on n'est pas obligé de tenir compte des effets du plan de sondage sur m_{ii} parce qu'il repose sur une "pseudo-EMV" conçue pour un échantillon binomial. Un autre graphique utile qui résume bien les informations contenues dans les graphiques diagnostiques de $\tilde{X}_i$ en fonction de i et de m_{ii} en fonction de i est le diagramme de dispersion des coordonnées $\tilde{X}_i^2/X_c^2 = X_i^2/X^2$ et h_{ii}, où h_{ii} est le $i^{\text{ième}}$ élément de la diagonale de la matrice (26) (voir Pregibon, 1981).

Les variables diagnostiques e_i, $\tilde{X}_i$ ou $\tilde{G}_i$ et m_{ii} permettent de déceler les points extrêmes, mais non d'évaluer l'effet de ces points sur divers aspects

de l'ajustement tels que les estimations des paramètres, $\hat{\underset{\sim}{\beta}}$, les valeurs ajustées, $\hat{\underset{\sim}{f}}$, et les mesures de la validité de l'ajustement $X^2/\delta_.$ ou $G^2/\delta_.$ ou d'autres. Comme Preqibon (1981), nous proposons trois indices qui mesurent l'effet des cases (points) extrêmes sur la qualité de l'ajustement.

1) Indice de la sensibilité des coefficients: Soit $\hat{\beta}_j(-\ell)$ la valeur de la pseudo-EMV de β_j calculée sans les données de la $\ell^{\text{ième}}$ case. L'indice $\Delta_j(\ell) = \hat{\beta}_j - \hat{\beta}_j(-\ell)/e.t.(\hat{\beta}_j)$ mesure la sensibilité du $j^{\text{ième}}$ coefficient au $\ell^{\text{ième}}$ point. Les qraphiques diaqnostiques de la relation entre $\Delta_j(\ell)$ et ℓ pour chaque j sont utiles, mais la quantité de qraphiques à examiner peut devenir rébarbative si le nombre de coefficients dans le modèle est élevé.

2) Indice de la sensibilité des valeurs ajustées: Le fait que les estimations des coefficients peuvent chanqer beaucoup quand on supprime le $\ell^{\text{ième}}$ point (ou la $\ell^{\text{ième}}$ case) ne signifie pas nécessairement qu'il existe de qrands écarts entre les valeurs ajustées $\hat{\underset{\sim}{f}}$ et $\hat{\underset{\sim}{f}}(-\ell)$, le vecteur des valeurs ajustées obtenues sans la $\ell^{\text{ième}}$ case; autrement dit, $\|\hat{\underset{\sim}{f}} - \hat{\underset{\sim}{f}}(-\ell)\|$ peut être faible. On utilise donc $[G^2 - \tilde{G}^2(-\ell)]/\delta_.$ ou $[X^2 - \tilde{X}^2(-\ell)]/\delta_.$ pour évaluer l'effet du $\ell^{\text{ième}}$ point sur les valeurs ajustées, où $\tilde{G}^2(-\ell)$ et $\tilde{X}^2(-\ell)$ sont la solution des équations (9) et (8) respectivement quand on remplace $\hat{f}_i = f_i(\hat{\underset{\sim}{\beta}})$ par $\hat{f}_i(-\ell) = f_i(\beta(-\ell))$.

3) Indice de la sensibilité de la validité de l'ajustement: La valeur de $[G^2 - G^2(-\ell)]/\delta_.$ ou $[X^2 - X^2(-\ell)]/\delta_.$, où $G^2(-\ell)$ et $X^2(-\ell)$ sont respectivement le rapport de vraisemblance et la variable khi-carré calculés sans la $\ell^{\text{ième}}$ case, mesure la sensibilité de la validité de l'ajustement. (Souliqnons le fait que $G^2(-\ell) \neq \tilde{G}^2(-\ell)$.)

3. APPLICATION DES RÉSULTATS THÉORIQUES À L'EPA

Nous avons appliqué les techniques décrites plus haut à quelques données de l'enquête sur la population active au Canada (EPA) menée en octobre 1980. L'échantillon était composé d'hommes âqés de 15 à 64 ans qui faisaient partie de la population active et n'étudiaient pas à temps plein. Nous avons choisi deux facteurs, l'âqe et le niveau d'instruction, pour expliquer la variation des taux de chômage à l'aide de modèles logit. On a réparti l'échantillon en

groupes d'âge en divisant l'intervalle [15, 64] en dix groupes dont le $j^{ième}$ était l'intervalle [10 + 5j, 14 + 5j], j = 1, 2, ..., 10, et la valeur au milieu de chaque intervalle, A_j, a été utilisée comme l'âge de toutes les personnes dans un même groupe d'âge. Pareillement, on a défini des niveaux d'instruction, E_k, en attribuant à chaque personne un nombre d'années d'instruction correspondant à la médiane dans sa catégorie respective; on a ainsi obtenu six niveaux d'instruction, soit 7, 10, 12, 13, 14 et 16 années. La classification croisée de l'âge et du niveau d'instruction remplit donc un tableau à deux dimensions composé de I = 60 cases contenant chacune une proportion π_{jk}.

L'EPA repose sur un plan d'échantillonnage par grappes stratifié à plusieurs degrés. Deux degrés d'échantillonnage sont définis pour les régions urbaines autoreprésentatives (AR) et trois ou quatre degrés pour les régions non autoreprésentatives (NAR), dans chaque province. Les estimations de l'enquête, $\hat{p}_{jk}$, subissent une correction pour tenir compte de la stratification a posteriori en fonction des projections du recensement relatives à la composition par âge et par sexe au niveau provincial. La matrice des covariances estimées $\hat{V}$ des estimations $\hat{p}_{jk}$ est calculée à partir de plus de 450 unités au premier degré d'échantillonnage (UPÉ), de sorte que le nombre de degrés de liberté de $\hat{V}$ est grand en comparaison de I = 60.

3.1 Tests d'hypothèses

Les graphiques diagnostiques de la relation entre les logits $\hat{\nu}_{jk}$ et l'âge A_j pour chaque niveau d'instruction E_k montrent que, pour une valeur donnée de k, $\hat{\nu}_{jk}$ augmente généralement en fonction de l'âge, atteint un sommet et diminue par la suite (en d'autres mots, les courbes sont convexes et ascendantes jusqu'à un maximum). Par conséquent, le modèle suivant pourrait expliquer la variation des π_{jk}:

$$\nu_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k + \beta_4 E_k^2,$$

$$j = 1, \ldots, 10; \; k = 1, \ldots, 6. \qquad (27)$$

Des recherches en sociologie ont aussi révélé que ce genre de modèle est justifié (Bloch et Smith, 1977). À partir des résultats présentés à la section 2, on a obtenu les valeurs suivantes pour les variables utilisées dans le test de la validité de l'ajustement:

$$X^2 = 98.9 \qquad G^2 = 101.2$$
$$X^2/\delta_. = 52.5 \qquad G^2/\delta_. = 53.7, \quad \delta_. = 1.88.$$

Étant donné que X^2 et G^2 dépassent $\chi^2_{0.05}(55) = 73.3$, la valeur critique du χ^2 à $I - s = 55$ degrés de liberté pour laquelle le risque de première espèce est de 5%, on rejetterait le modèle si on ne tenait pas compte du plan de sondage. Par contre, la valeur de $X^2/\delta_.$ ou $G^2/\delta_.$ indique que le modèle est acceptable à un niveau de signification (seuil P) d'approximativement 0.52. La variable X^2_S, une fois rectifiée en fonction de $\chi^2_{0.05}(55)$, est égale à 47.7, valeur non significative. En outre, dans le cas présent, où s (= 5) est relativement faible en comparaison de I (= 60), le facteur de rectification simple $\bar{d}$, la moyenne des effets du plan de sondage dans une case (voir Fellegi, 1980), est très proche de $\bar{\delta}$: en effet $\bar{d} = 1.905$, tandis que $\bar{\delta} = 1.88$ (voir Rao et Scott (1984) pour un exposé théorique).

La variable de Wald X^2_S n'est pas définie dans ce cas parce que $\hat{p}^2_{jk} = 1$ dans deux cases, mais nous avons modifié légèrement les totaux estimés pour nous assurer que $\hat{p}_{jk} < 1$ dans toutes les cases et nous avons ensuite calculé X^2_W. Les valeurs de X^2_W qu'on a obtenues sont toutes grandes en comparaison de $X^2/\delta_.$ (elles sont au moins trente fois supérieures à $X^2/\delta_.$) et varient beaucoup (de 1715 à 3061). La variable de Wald est donc très instable dans ce test de la validité de l'ajustement. Si on élimine deux cases où $\hat{p}_{jk} = 1$, $X^2 = 68.4 < \chi^2_{0.05}(53) = 71.0$, ce qui indique que le modèle (27) est acceptable. Mais supprimer des cases simplement pour justifier l'emploi d'une certaine variable dans un test ne constitue pas une démarche souhaitable. Une autre difficulté liée à X^2_W, que Fay (1983) a déjà signalée, ne se présente pas dans ce test parce que le nombre de degrés de liberté de $\hat{V}$ est grand en comparaison du nombre de cases dans le tableau.

Le tableau 1 présente les valeurs des pseudo-EMV, de leurs erreurs types (e.t.) et des erreurs types correspondantes dans l'échantillonnage binomial

aléatoire calculées dans le modèle (27), ainsi que la variable de Wald $X_W^2(2|1)$ et la variable de type G^2 $G^2(2|1)/\delta.(H)$ pour les hypothèses H_i: $\beta_i = 0$, $i = 1, 2, 3, 4$ à l'intérieur du modèle (27). Tel que prévu, les vraies valeurs des erreurs types sont plus élevées que celles des erreurs types binomiales correspondantes. L'hypothèse H_4: $\beta_4 = 0$ (autrement dit, l'hypothèse selon laquelle le coefficient de E_i^2 est nul) ne peut être rejeté au seuil de 5% à partir de la variable de Wald ou de la variable G^2. En revanche, le coefficient β_2 de A_i^2 est très significatif. Quand on examine le niveau de signification des coefficients individuels, on compare les valeurs de $X_W^2(2|1)$ ou $G^2(2|1)/\delta_.(H)$ et $\chi^2_{0.05}(1) = 3.84$, la valeur critique du χ^2 à un degré de liberté pour laquelle le risque de première espèce est de 5%.

Nous avons aussi vérifié les hypothèses emboîtées suivantes dans le cadre du modèle (27): H_{34}: $\beta_3 = \beta_4 = 0$ (ce qui revient à dire que le niveau d'instruction n'a aucun effet); H_{24}: $\beta_2 = \beta_4 = 0$ (hypothèse d'absence d'effets quadratiques). Les résultats de H_{34} et H_{24} sont très significatifs:

$$G^2(2|1)/\delta_.(H_{34}) = 282.2/1.64 = 172.1,\ X_W^2(2|1) = 165.6 \text{ for } H_{34};$$

$$G^2(2|1)/\delta_.(H_{24}) = 242.2/2.28 = 106.3,\ X_W^2(2|1) = 162.1 \text{ pour } H_{24}, \text{ alors que}$$
$\chi^2_{0.05}(2) = 5.99$.

Tableau 1: Pseudo-EMV $\hat{\beta}_i$, e.t. $(\hat{\beta}_i)$, $X_W^2(2|1) = \hat{\beta}_i^2/\text{var}\ (\hat{\beta}_i)$ et $G^2(2|1)/\delta_.(H_i)$ pour les données de l'EPA à l'intérieur du modèle (27).

	$\hat{\beta}_i$	e.t.$(\hat{\beta}_i)$ Vraie Valeur	e.t.$(\hat{\beta}_i)$ Valeur Binomiale	$X_W^2(2\|1)$	$G^2(2\|1)/\delta_.(H_i)$
0	-2.76	0.557		24.6	
1	0.209	0.0132	0.012	250.6	168.4
2	-0.00217	0.000173	0.000136	157.3	102.1
3	0.0913	0.0891	0.068	1.04	1.01
4	0.00276	0.00411	0.0030	0.45	0.46

Contrairement à ce qui s'est produit dans le test de la validité de l'ajustement, la variable de Wald est stable dans le test d'hypothèses emboîtées et atteint des valeurs proches des résultats correspondants obtenus pour $G^2(2|1)/\delta_{\bullet}(H)$.

Le test de la validité de l'ajustement et ces tests d'hypothèses emboîtées permettent de construire un modèle simple comprenant seulement quatre paramètres:

$$\nu_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \tag{28}$$

où $\hat{\beta}_0 = -3.10$, $\hat{\beta}_1 = 0.211$, $\hat{\beta}_2 = 0.00218$ et $\hat{\beta}_3 = 0.1509$ et les erreurs types sont 0.247, 0.0130, 0.000172 et 0.0115 respectivement. Nous utilisons le modèle (28) à la section 3.2 pour définir des indices diagnostiques de la régression logistique.

3.2 Indices diagnostiques

Nous décrivons maintenant des exemples de l'utilisation des indices diagnostiques décrits à la section 2.4.

(i) Analyse des résidus

Les soixante cases du tableau à deux dimensions ont été numérotées en ordre lexicographique et les résidus centrés réduits e_i ont été calculés pour le modèle (28) qui a été formulé après une série de tests d'hypothèses. Parmi les soixante valeurs de e_i, celles des cases numéros 6 et 54, où $\hat{p}_{jk} = 1$, étaient très élevées, soit 166.6 et 6.2 respectivement. Parmi les autres e_i, les résidus numéros 7, 27 et 59 étaient égaux à 3.84, 2.73 et 2.52 respectivement, tandis que l'espérance mathématique du nombre de $|e_i|$ dépassant 2.33 dans le cadre du modèle (28) est d'environ 0.02 x 60 = 1.2. Par conséquent, il y a lieu de penser que les cases 7 et 27 contiennent peut-être des proportions extrêmes.

Une représentation graphique de $\tilde{G}_i$ sur une échelle fonctionnelle normale est présentée à la figure 1; la courbe de $\tilde{\chi}_i$ n'est pas incluse pour économiser

de l'espace et parce qu'elle ressemble à celle de $\tilde{G}_i$. La figure 1 révèle une allure linéaire dont aucun point ne s'éloigne beaucoup. Le graphique diagnostique de $\tilde{G}_i$ présenté à la figure 2 concorde avec la figure 1. Il semble donc qu'aucune case ne contient des proportions extrêmes quand on utilise $\tilde{G}_i$ et G_c^2 pour l'analyse des résidus.

(ii) Repérage des cases prépondérantes

Le graphique diagnostique de m_{ii} est présenté à la figure 3, où les points des cases 1 et 6 se trouvent bien loin de l'ensemble des points. La figure 4 montre la relation entre $\tilde{X}_i^2/X_c^2 = X_i^2/X^2$ et h_{ii}, où une droite ayant une pente égale à -1 a pour équation $X_i^2/X^2 + h_{ii} = 3\ \mathrm{moy}(h_{ii}^*)$. Ici $h_{ii}^* = h_{ii} + X_i^2/X^2$ et les valeurs de h_{ii}^* voisines de un correspondent aux cases qui sont soit extrêmes, soit prépondérantes, voire les deux (Pregibon, 1981) et se trouvent au-dessus de la droite tracée à la figure 3. Il est clair que les cases 1 et 6 et, dans une certaine mesure, les cases 7 et 58 devraient être examinées de plus près.

(iii) Indice de la sensibilité des coefficients

Les graphiques diagnostiques de l'indice qui mesure la sensibilité des coefficients aux différentes cases (relation entre $\Delta_j(\ell)$ et ℓ) sont présentés aux figures 5, 6, 7 et 8 pour β_0, β_1, β_2 et β_3 respectivement. Il ressort clairement de ces graphiques que les cases 2 et 3 engendrent de l'instabilité dans $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$, tandis $\hat{\beta}_3$ est perturbé par la case 7.

(iv) Indice de la sensibilité des valeurs ajustées

La figure 9 présente le graphique de la relation entre $[G^2 - \tilde{G}^2(-\ell)]/\delta_{\bullet} = c$ et ℓ et permet d'évaluer l'effet des cases individuelles sur les valeurs ajustées. Cette figure comprend un grand sommet correspondant aux cases 2 et 3 et un sommet plus bas pour la case 7. Comme Cook (1977) et Pregibon (1981) l'ont déjà souligné, on peut comparer c et le pourcentage lié à $\chi^2(s)$ ($s = 4$ dans le modèle (28)) pour avoir une idée approximative du contour de la région de confiance où la pseudo-EMV est ramenée quand la $i^{\text{ième}}$ case est supprimée. Dans ce cas, la valeur 2.1 pour la case 2 correspond à peu près au contour de 78% de la région de confiance.

(v) Indice de la sensibilité de la validité de l'ajustement

La figure 10 montre la relation entre $[G^2 - G^2(-\ell)]/\delta_{\bullet}$ et ℓ. Le graphique de $[X^2 - X^2(-\ell)]/\delta_{\bullet}$ est semblable et, par conséquent, on ne le présente pas ici; ce dernier indice est d'ailleurs le moins préférable des deux (Pregibon,

1981). La figure 10 révèle des sommets importants pour les cases 2, 3, 7, 27, 39 et 54 (valeurs ≥ 3), la valeur la plus élevée étant celle de la case 7, qui a un indice égal à 5.4. Si on supprime la case 7 et qu'on calcule de nouveau la variable corrigée $G_c^2(-\ell) = G^2(-\ell)/\delta_{.}(-\ell)$, où $\delta_{.}(-\ell)$ est la valeur correspondante de $\delta_{.}$, on obtient un résultat égal à 48.43 avec 55 degrés de liberté, alors que $G^2/\delta_{.} = 55.3$ avec 56 degrés de liberté.

Dans l'ensemble, nos résultats révèlent qu'il serait peut-être opportun de supprimer les cases 7, 2 et 3 mais nous croyons que leur effet sur les estimations n'est pas assez important pour justifier une telle décision. Après tout, on veut expliquer la variation entre les proportions contenues dans toutes les cases, à moins qu'il n'existe un grand désaccord entre les données et le modèle ajusté, qui est attribuable à certaines cases seulement.

REMERCIEMENTS

Nous aimerions remercier M. Gratton de Statistique Canada qui a fait les graphiques présentés dans cette analyse.

BIBLIOGRAPHIE

[1] Bloch, F.E. et Smith, S.P. (1977). Human capital and labour market employment. J. Human Resources 12, p. 550-559.

[2] Cook, R.D. (1977). Detection of influential observations in linear regression. J. American Statistical Association 72, p. 169-174.

[3] Cook, R.D. et Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, Londres.

[4] Cox, D.R. (1970). Analysis of Binary Data. Chapman and Hall, Londres.

[5] Fay, R.E. (1983). Replication approaches to the log-linear analysis of data from complex samples. Manuscrit non publié.

[6] Fellegi. I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. J. American Statistical Association 75, p. 261-268.

[7] McCullagh, P. et Nelder. J.A. (1983). Generalized Linear Models. Chapman and Hall, Londres.

[8] Pregibon, D. (1981). Logistics regression diagnostics. Ann. Statist. 9, p. 705-724.

[9] Rao, J.N.K. et Scott, A.J. (1984). On simple adjustments to chisquared tests with survey data: log-linear and logit models. Manuscrit non publié.

[10] Roberts. G. (1984). On chi-squared tests for logit models with cell proportions estimated from survey data. Manuscrit non publié. Université Carleton.

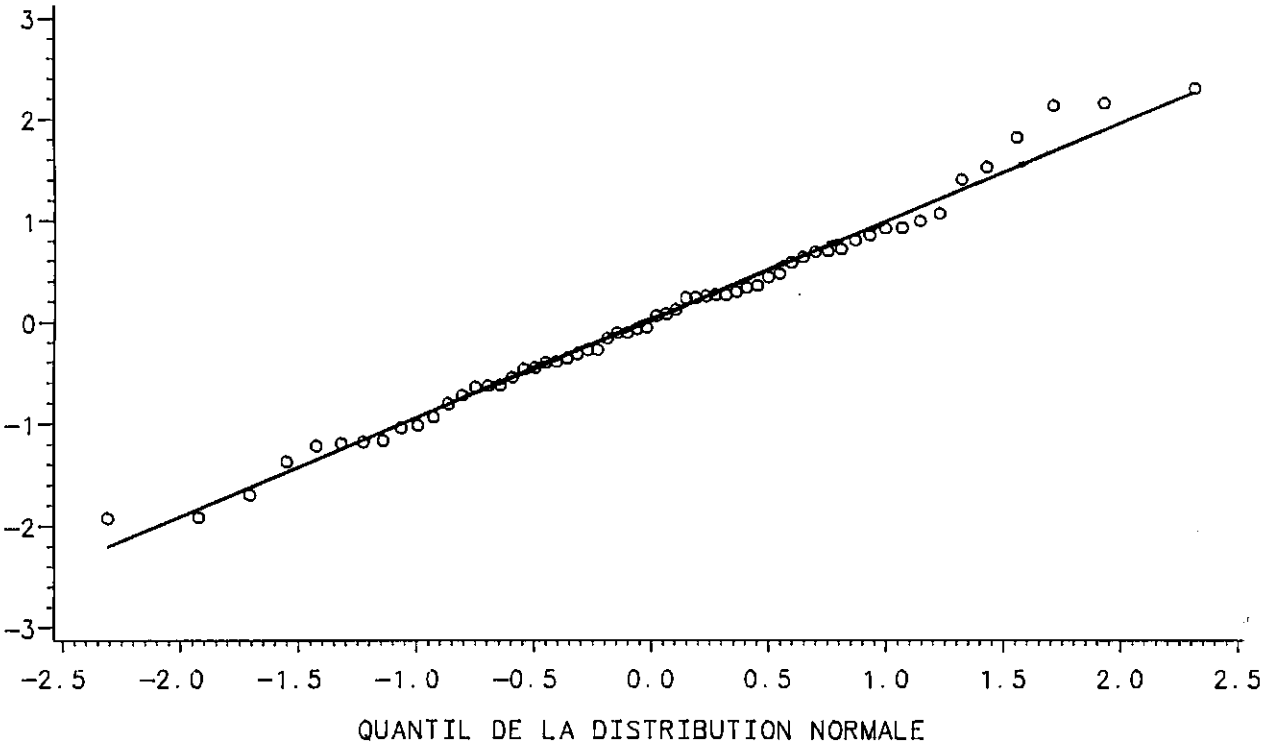

Figure 1: Représentation de $\widetilde{G}_i$ sur l'échelle fonctionnelle normale

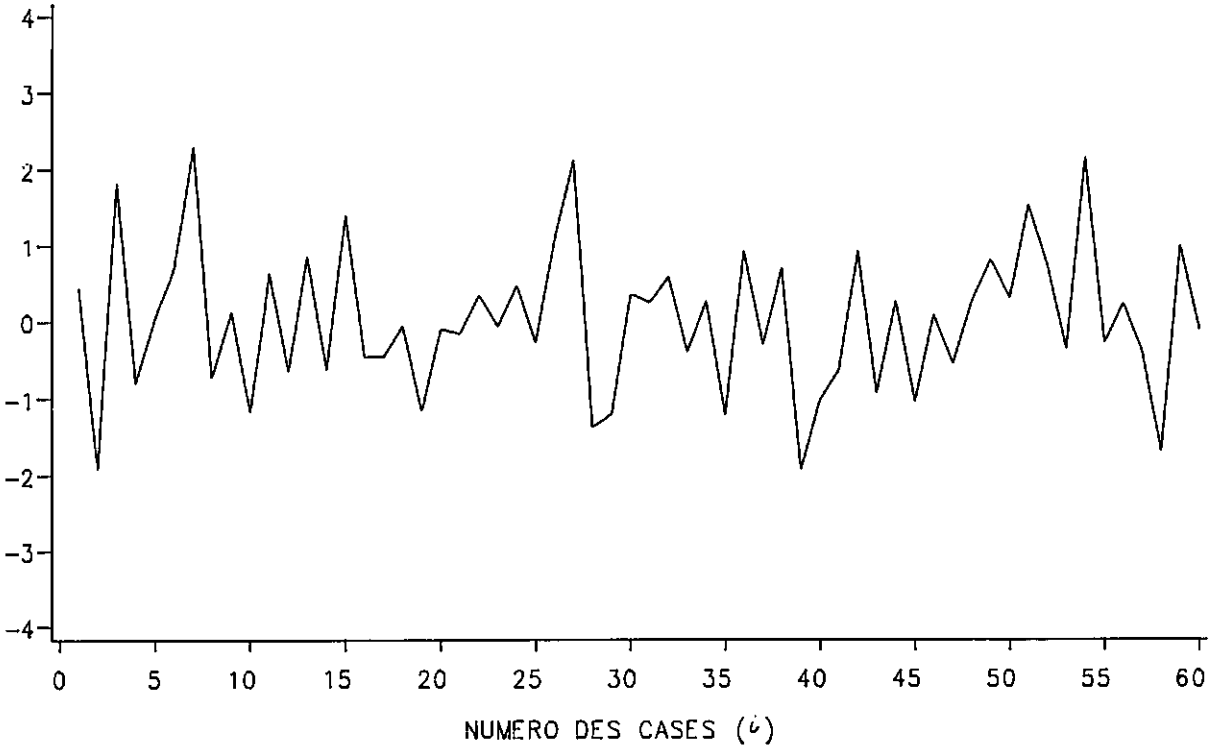

Figure 2: Graphique diagnostique de $\widetilde{G}_i$

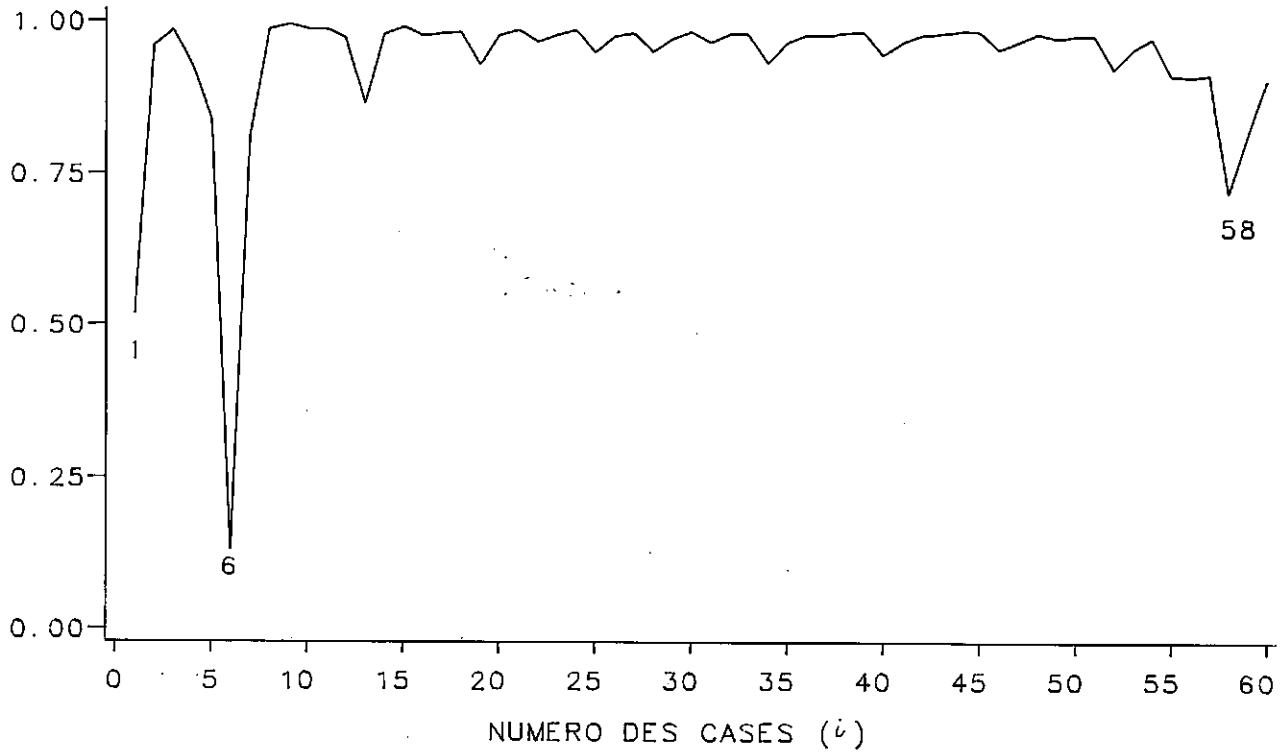

Figure 3: Graphique diagnostique: de m_{ii}

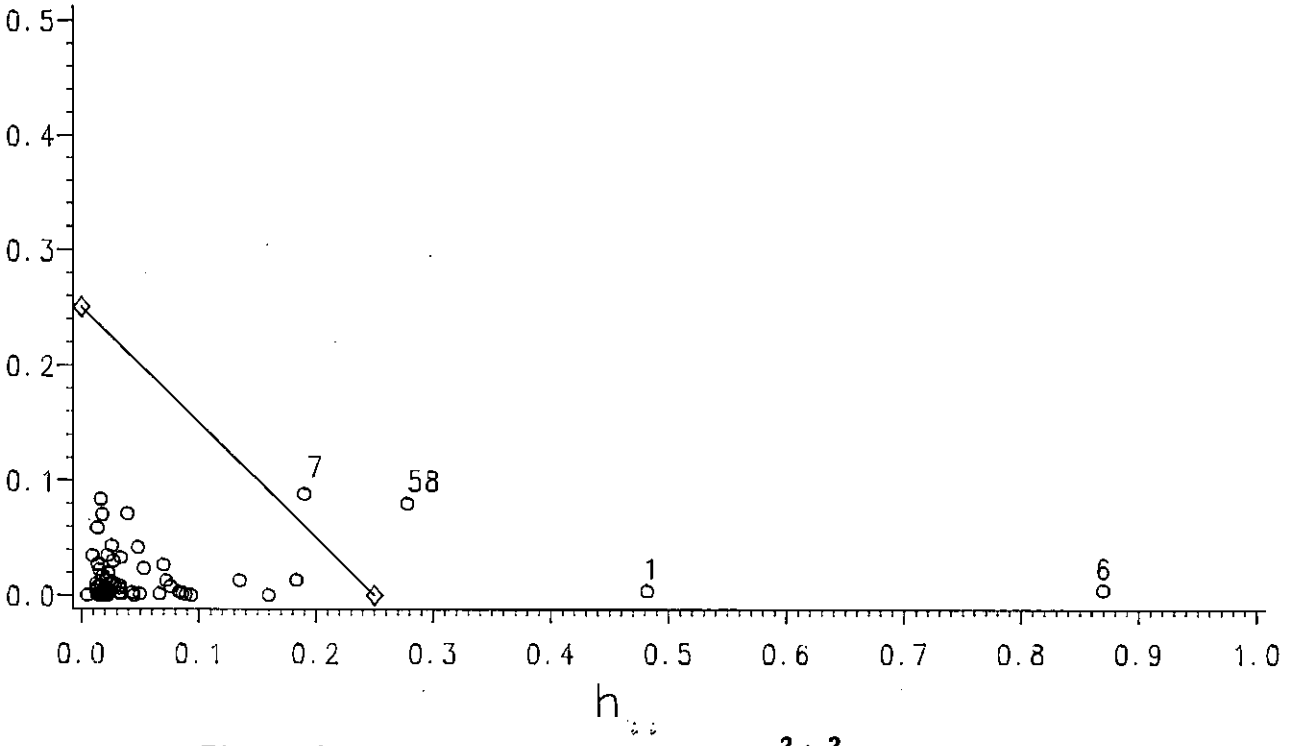

Figure 4: Diagramme de dispersion χ_i^2/χ^2 en fonction de h_{ii}

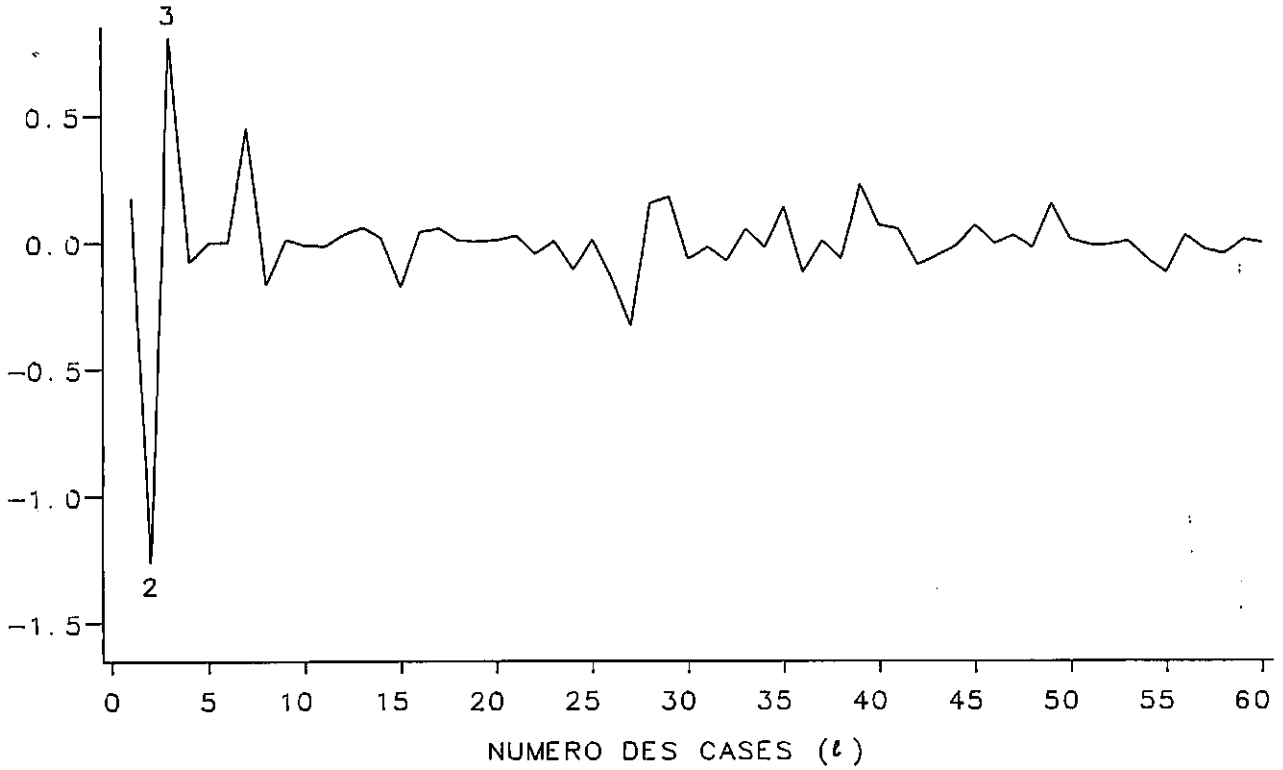

Figure 5: Graphique diagnostique de of $\{\hat{\beta}_0 - \hat{\beta}_0(-\ell)\}/E.T.(\hat{\beta}_0)$

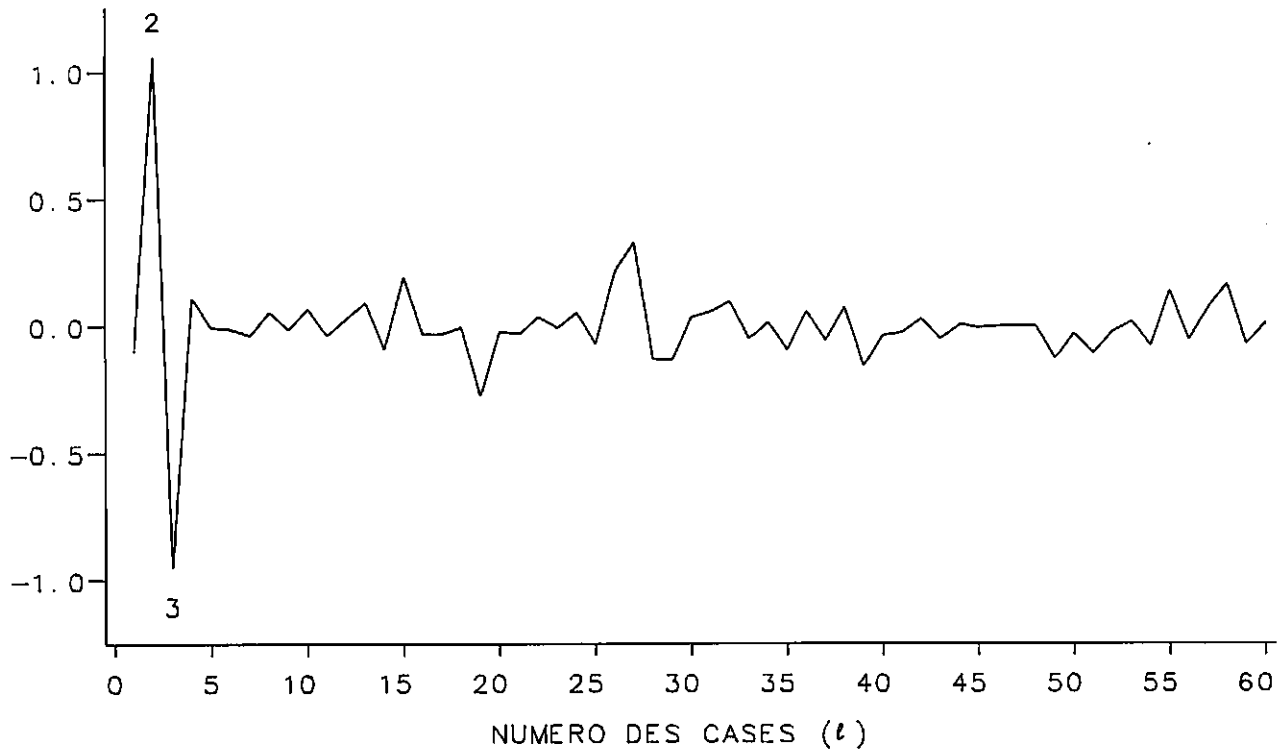

Figure 6: Graphique diagnostique de $\{\hat{\beta}_1 - \hat{\beta}_1(-\ell)\}/e.t.(\hat{\beta}_1)$

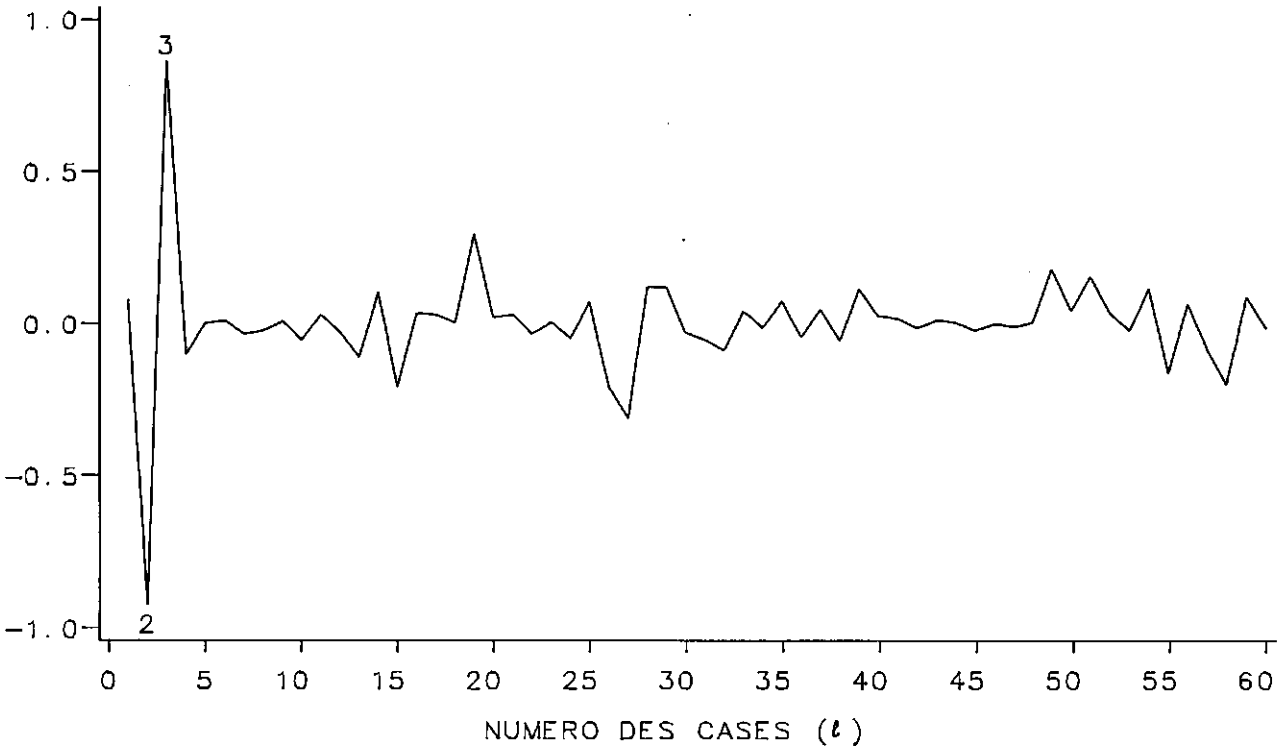

Figure 7: Graphique Diagnostique de $\{\hat{\beta}_2 - \hat{\beta}_2(-\ell)\}/e.t.(\hat{\beta}_2)$

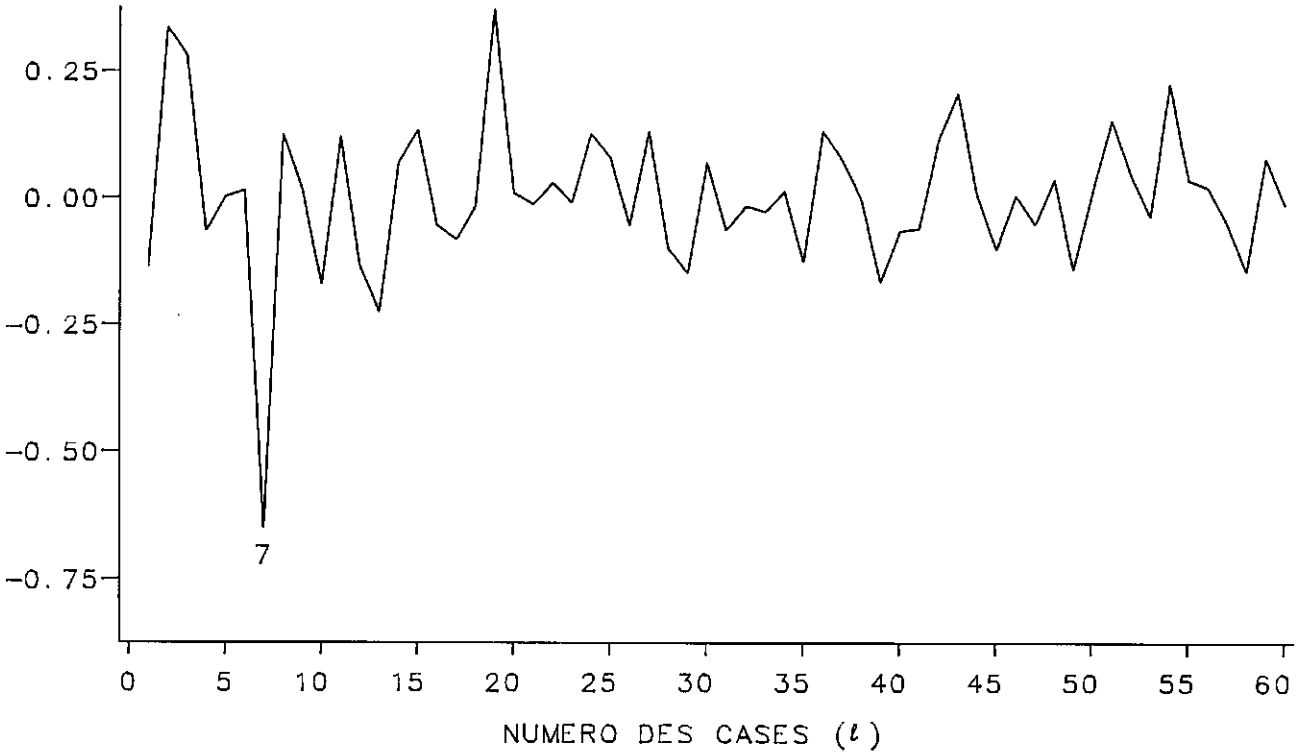

Figure 8: Graphique Diagnostique de $\{\hat{\beta}_3 - \hat{\beta}_3(-\ell)\}/e.t.(\hat{\beta}_3)$

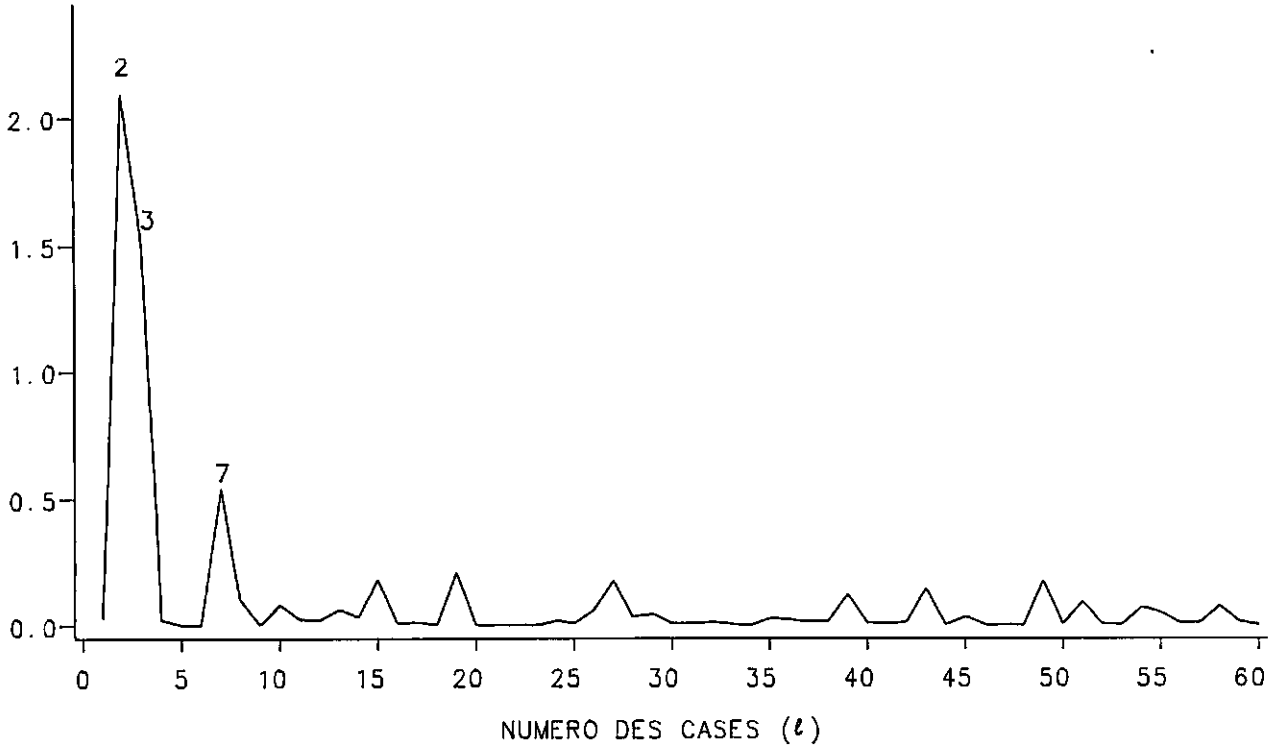

Figure 9: Graphique Diagnostique de $\{G^2 - \tilde{G}^2(-\ell)\}/\hat{\hat{\delta}}$

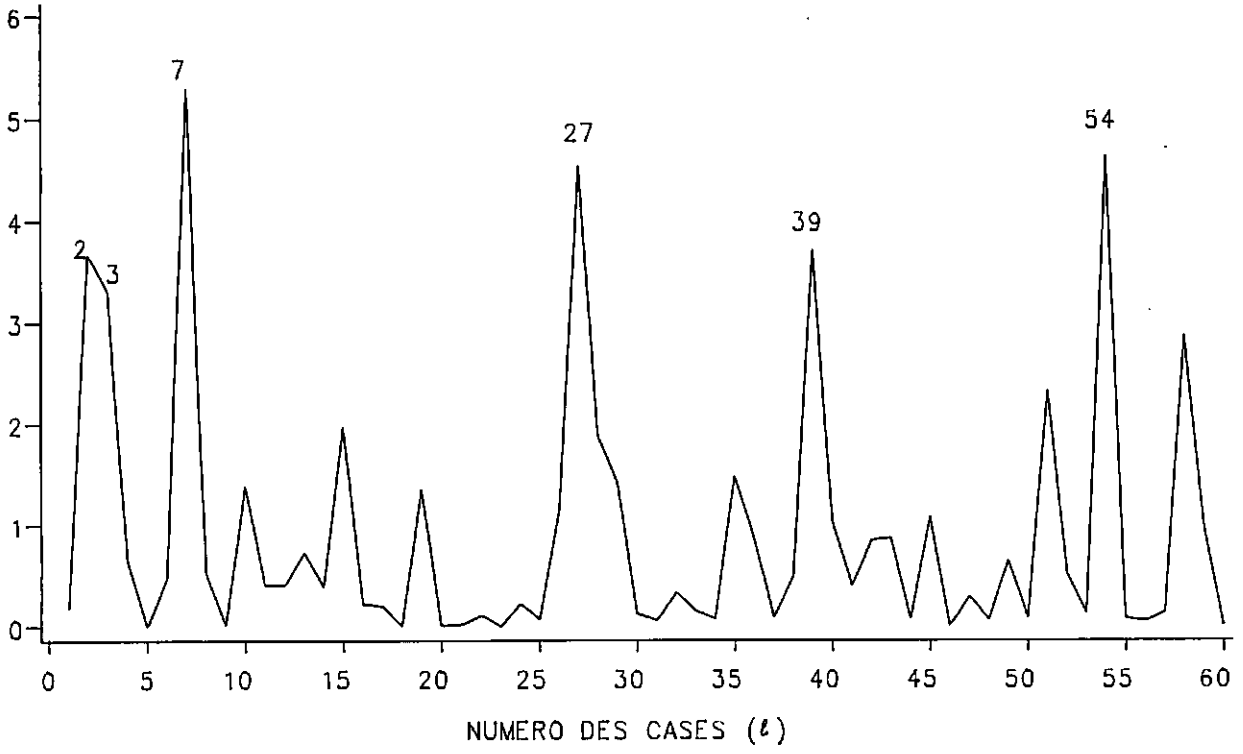

Figure 10: Graphique Diagnostique de $\{G^2 - G^2(-\ell)\}/\hat{\hat{\delta}}$

APPLICATION DE MODÈLES LINÉAIRES ET LOG-LINÉAIRES AUX DONNÉES D'ÉCHANTILLONS COMPLEXES

Robert E. Fay[1]

La plupart des enquêtes menées par des organismes comme Statistique Canada ou le U.S. Bureau of the Census reposent sur des plans de sondage complexes. Les techniques d'inférence statistique basées sur le plan de sondage, qui sont généralement utilisées par ce genre d'organisme pour calculer des moyennes et des totaux, peuvent également s'étendre à l'estimation des paramètres de modèles analytiques. La plus grande partie de cette étude porte sur l'application des méthodes d'inférence basées sur le plan de sondage aux modèles théoriques, mais elle présente également des arguments justifiant le recours à des procédés basés sur le modèle dans certains cas, ce qui explique le fait que ces deux formes d'inférence soient utilisées par l'organisme dont l'auteur fait partie.

Cette étude décrit brièvement l'expérience acquise dans l'extension des techniques d'inférence basées sur le plan de sondage à l'analyse de régression linéaire. Récemment, la méthode des "poids d'échantillonnage artificiels" (replicate weighting) a été appliquée à l'estimation de la variance dans diverses enquêtes menées par le Census Bureau. Jusqu'à présent, cette méthode a servi avant tout à calculer la variance de variables statistiques simples, mais elle facilite aussi l'évaluation des variances dans pratiquement n'importe quel modèle analytique complexe. Enfin, on décrit des techniques relatives aux modèles log-linéaires et on résume les travaux faits sur ce sujet.

1. INTRODUCTION

Statistique Canada a joué un rôle important dans l'élaboration d'un grand nombre de procédés pour l'application de méthodes analytiques aux données d'enquête. Le présent exposé a pour objet de résumer et de partager l'expérience acquise par le U.S. Bureau of the Census dans ce domaine.

Les techniques d'inférence "basées sur le plan de sondage" (aussi appelées méthodes "classiques") prédominent dans l'analyse et la production des données

[1] Robert E. Fay, Statistical Methods Division, U.S. Bureau of the Census, Washington, D.C.

de la plupart des organismes statistiques gouvernementaux comme Statistique Canada et le U.S. Bureau of the Census et de la plupart des grandes maisons de sondage privées. Cette forme d'inférence statistique repose sur le procédé de randomisation utilisé pour prélever un échantillon dans une population finie. Les principes appliqués à la construction d'intervalles de confiance et aux tests d'hypothèses découlent de la théorie des grands échantillons et de la randomisation plutôt que d'un modèle précis. Les manuels classiques, comme ceux de Cochran [4], de Kish [17] et de Hansen, Hurwitz et Madow [14] présentent les éléments de cette théorie. Dans un article récent, Hansen, Madow et Tepping [15] ont décrit comment ces techniques sont supérieures aux méthodes "basées sur un modèle" pour l'inférence à partir de données d'enquête: Särndal [25] et Cassel, Särndal et Wretman [3] ont examiné d'une manière un peu différente le choix entre les méthodes basées sur un modèle et celles basées sur le plan de sondage. Dans l'ensemble, les premiers travaux sur les techniques d'inférence basées sur le plan de sondage visaient l'estimation de totaux, de proportions, de moyennes et de ratios à l'échelle de la population, et la plupart des recherches correspondantes sur les méthodes basées sur un modèle mettent également l'accent sur ces statistiques fondamentales.

Par contre, les schémas analytiques les plus répandus, tels que les modèles de régression linéaire, les modèles log-linéaires et les modèles linéaires généralisés, ont d'abord été formulés en fonction de modèles stochastiques explicites fondés, par exemple, sur la loi normale ou la loi multinomiale. Les méthodes d'inférence "classiques" ont progressivement assumé le sens de procédés d'inférence statistique fondés sur une distribution hypothétique (l'adjectif "classique" peut englober le terme "bayésien" dans la présente analyse). Les méthodes d'estimation "robustes" ne comprennent pas de restrictions précises concernant la distribution de la population, mais elles comportent souvent des hypothèses qu'on ne retrouve pas généralement dans les techniques d'enquête. Ainsi, par exemple, les termes d'erreur du modèle sont supposés indépendants et considérés comme provenant d'une population symétrique.

Beaucoup de chercheurs qui connaissent l'un ou l'autre de ces modèles analytiques les ont appliqués directement à des données d'enquête sans tenir compte des effets possibles du plan de sondage sur la validité des inférences basées sur les hypothèses distributionnelles courantes. Notre propos ici,

bien entendu, est essentiellement d'examiner des méthodes "basées sur le plan de sondage" qui permettent effectivement de prendre en compte les effets du plan. Toutes les autres sections de cet exposé portent sur les méthodes "basées sur le plan de sondage", mais la section suivante présente d'abord certaines notions théoriques et pratiques qui interviennent dans le choix d'une de ces deux techniques. On explique aussi comment ce problème se pose dans les opérations du Census Bureau.

La troisième section décrit brièvement l'expérience du Census Bureau avec les méthodes de régression linéaire basées sur le plan de sondage. La quatrième section résume un procédé qui est utilisé dans la mise en oeuvre informatique des méthodes de duplication et repose sur des "poids d'échantillonnage artificiels" (replicate weights). Bien que cette technique soit conçue principalement pour calculer la variance des statistiques habituellement estimées dans les enquêtes, elle facilite également le calcul des erreurs types dans les modèles complexes. Cette solution générale peut s'avérer particulièrement utile pour l'utilisation de modèles relativement moins répandus, c'est-à-dire les schémas autres que les modèles linéaires. les modèles log-linéaires et d'autres modèles linéaires généralisés. Enfin, on résume quelques travaux récents sur les modèles log-linéaires et on mentionne des logiciels conçus pour l'ajustement de ces modèles.

2. SÉLECTION D'UNE TECHNIQUE D'INFÉRENCE BASÉE SUR LE PLAN DE SONDAGE OU SUR UN MODÈLE

Le choix entre une méthode d'inférence basée sur le plan de sondage et une méthode basée sur un modèle peut dépendre de plusieurs facteurs tels que les effets de la stratifications et l'existence ou l'ampleur de la dépendance entre les valeurs échantillonnées (effet de grappe). DuMouchel et Duncan [6] ont énuméré les principales questions soulevées par ce choix dans leur étude sur la nécessité d'intégrer les poids d'échantillonnage dans les régressions linéaires.

Si $\underset{\sim}{Y}$ représente un vecteur colonne d'observations Y_i et que $\underset{\sim}{X} = \{X_{ij}\}$, $j = 1, \ldots, p$ représente une série de variables explicatives de $\underset{\sim}{Y}$, le modèle

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{2.1}$$

où $\underset{\sim}{\varepsilon} = \{\varepsilon_i\}$ est un vecteur de résidus $\varepsilon_i \sim N(0, \sigma^2)$ qui sont indépendants et identiquement distribués, a pour estimateur du maximum de vraisemblance de $\underset{\sim}{\beta}$

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{Y}. \tag{2.2}$$

Dans le calcul des estimations d'une enquête, un poids W_i est affecté à chaque unité i de l'échantillon en fonction de l'inverse de la probabilité de sélection, et des facteurs de correction pour les cas de non-réponse et l'estimation par le quotient sont souvent utilisés. Si $\underset{\sim}{W}$ est une matrice diagonale de poids W_i, l'estimateur

$$\hat{\underset{\sim}{\beta}}_W = (\underset{\sim}{X}^T\underset{\sim}{W}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{W}\underset{\sim}{Y} \tag{2.3}$$

donne un résultat qui est rectifié pour tenir compte du plan de sondage et des poids d'échantillonnage. Dans le modèle stochastique qui justifie le choix de l'estimateur (2.2) ou, plus généralement, s'il n'existe aucune corrélation entre les valeurs de ε_i et que ces erreurs ont une espérance mathématique nulle et des variances égales, l'estimateur (2.3) a une plus grande variance d'échantillonnage que (2.2). En revanche, si ces conditions ne sont pas satisfaites (en particulier celle relative à l'espérance mathématique des ε_i), l'estimateur (2.3) produit quand même une estimation corrigée en fonction du plan de sondage pour le paramètre de la population $\underset{\sim}{\beta}^*$ et correspond à l'application de (2.2) aux valeurs contenues dans l'ensemble de la population finie, mais le calcul de (2.2) sans pondération des unités de l'échantillon n'assure pas une estimation convergente de $\underset{\sim}{\beta}^*$.

DuMouchel et Duncan vont plus avant dans leur étude du choix entre les estimateurs (2.2), dont la variance est plus faible dans le modèle simple, et (2.3), qui est convergent même si les hypothèses du modèle s'avèrent fausses. Ils citent d'autres auteurs qui ont pris l'un ou l'autre des deux partis dans cette controverse; leur article est un excellent exposé des différents points de vue sur cette question. Ils proposent également un test qui peut être exé-

cuté par les logiciels habituels pour la régression linéaire et qui vise à déterminer s'il y a une différence marquée entre les résultats de régressions pondérées et non pondérées. Si ce test nous oblige à rejeter l'hypothèse selon laquelle (2.2) et (2.3) sont des estimateurs convergents d'un même ensemble de coefficients, on doit utiliser l'estimateur (2.3) parce qu'il converge vers la valeur de $\underset{\sim}{\beta}^*$ pour la population. Si cette hypothèse n'est pas rejetée, les auteurs préfèrent (2.2) parce que sa variance est (généralement) plus faible.

Si un chercheur rejette (2.2) à cause des résultats du test proposé par DuMouchel et Duncan et qu'il calcule plutôt (2.3), le motif de ce choix est relativement clair: (2.3) est choisi au lieu de (2.2) parce que (2.3) demeure convergent si le modèle ne peut être accepté. Si ce test permet d'"accepter" l'hypothèse nulle et qu'on utilise (2.2) et les erreurs types correspondantes calculées dans le cadre du modèle, il faut néanmoins être prudent vis-à-vis des interprétations naïves faites au sujet du paramètre de la population $\underset{\sim}{\beta}^*$ à partir de (2.2) et des intervalles de confiance construits à l'aide de cet estimateur. Dans bien des cas, le choix de (2.3) et du degré de fiabilité qu'il offre constitue le moyen le plus "sûr" d'interpréter les données et d'estimer $\underset{\sim}{\beta}^*$ quand on pense que le modèle peut être inexact, même si un test de l'hypothèse selon laquelle il n'existe aucune différence significative entre les analyses pondérées et non pondérées ne permet pas de rejeter le modèle.

L'article de DuMouchel et Duncan met clairement en évidence les facteurs les plus essentiels dans le choix entre les techniques d'inférence basées sur un modèle et celles basées sur le plan de sondage, à savoir l'efficacité si le modèle est bien spécifié et la convergence si les hypothèses du modèle sont inacceptables. Il serait opportun de faire deux observations concernant ces méthodes. D'abord, bien que l'omission des poids d'échantillonnage rende les estimateurs non convergents dans tout procédé basé sur le plan de sondage et soit justifiée seulement dans les techniques basées sur un modèle, les techniques basées sur un modèle ne font pas toutes abstraction des informations contenues dans ces poids.

Rubin [24] a formulé une explication concise de ce dernier point dans son compte rendu de l'article de Hansen, Madow et Tepping [15]. Faisant référence aux notions plus développées élaborées par Rosenbaum et Rubin [22], Rubin a

souligné le fait qu'une interprétation bayésienne complète des données observées repose non seulement sur les relations fonctionnelles et distributionnelles qui existent dans l'ensemble de la population (comme dans des modèles tels que (2.1) à l'échelle de la population), mais aussi sur le processus par lequel les observations de l'échantillon sont recueillies. (Dans un plan de sondage randomisé, , la "propension" à faire partie de l'échantillon est une notion équivalente à celle de la probabilité de sélection et au "coefficient de propension" (propensity score) de Rosenbaum et Rubin [22].) À partir de cette notion, Rubin [23] a présenté une justification intéressante, dans une optique bayésienne, de l'utilisation de la randomisation dans la sélection d'un échantillon, démarche qui est défendue fermement par les partisans des méthodes basées sur le plan de sondage, mais qui a été accueillie avec une sorte de mépris par beaucoup de défenseurs des méthodes basées sur un modèle. Par conséquent, Rubin préconise l'application de techniques d'inférence basées sur un modèle et une analyse sérieuse des effets du processus de sélection ou de la propension d'inclusion dans l'échantillon; dans certains cas, ces principes conduiraient soit à (2.2), soit à (2.3) ou à d'autres estimateurs possibles.

La deuxième observation concerne le fait que DuMouchel et Duncan ont explicitement limité leur analyse à la question de la pondération dans les échantillons aléatoires simples stratifiés. Un problème qui est aussi important dans un grand nombre de cas est celui de l'effet de grappe, c'est-à-dire la dépendance entre les unités échantillonnées qui est attribuable à leur inclusion simultanée dans l'échantillon en vertu du plan de sondage comme, par exemple, quand des personnes appartenant à un ménage échantillonné ou des personnes appartenant à des ménages voisins sont simultanément choisies pour l'échantillon. Dans les échantillons autopondérés, (dont toutes les unités ont un poids égal), les méthodes basées sur le plan de sondage et celles basées sur un modèle peuvent souvent produire les mêmes estimations pour les paramètres d'un modèle analytique, mais l'évaluation de la fiabilité de ces résultats peut être fort différente, sauf si les dépendances découlant de l'effet de grappe sont explicitement incorporées dans le procédé d'inférence basé sur un modèle. Contrairement au problème de l'utilisation des poids dans les échantillons aléatoires simples stratifiés, où une méthode d'inférence basée sur un modèle peut être justifiée si les termes d'erreur ont exactement

le comportement prévu dans le modèle théorique, toute dépendance entre les observations causée par l'effet de grappe (quand elle est assez prononcée) contredit automatiquement toute hypothèse d'indépendance des erreurs qui peut faire partie d'un modèle trop simplifié. Ainsi, les modèles qui font abstraction des effets de grappe quand on sait que ceux-ci existent sont forcément des représentations inexactes des données en question.

Les techniques d'inférence basées sur le plan de sondage constituent la norme au U.S. Bureau of the Census; toutefois, en pratique, les deux formes d'inférence sont unies dans l'ajustement de modèles. La plupart des chercheurs ont tendance à appliquer des techniques basées sur le plan de sondage pour faire des inférences concernant des relations à l'échelle nationale à partir d'échantillons complexes. Quand la variation des poids d'échantillonnage est modérée ou nulle et qu'on peut supposer que les effets de grappe sont faibles, les méthodes d'inférence basées sur un modèle semblent être bien acceptées. L'intérêt manifesté à l'égard des procédés basés sur un modèle dans ces cas découle sans doute plus d'un choix pratique que d'un point de vue philosophique: les méthodes basées sur un modèle sont plus accessibles et mieux connues que les techniques correspondantes basées sur le plan de sondage. (L'auteur a observé des cas où la variation des poids et les effets de grappe étaient tels que les méthodes basées sur le plan de sondage reproduisaient exactement les mêmes conclusions que celles basées sur un modèle, ce qui justifie l'utilisation de techniques basées sur un modèle dans ce genre de conditions. Toutefois, quand les poids d'échantillonnage varient beaucoup ou qu'il existe des effets de grappes dans certaines caractéristiques, il est facile de trouver des exemples où ces deux formes d'inférence entraînent des résultats très différents et où les méthodes d'inférence basées sur un modèle sont fort douteuses.)

Dans certains domaines d'activité précis du Census Bureau, il semble que ce soit exclusivement les méthodes basées sur un modèle qui sont utilisées. En particulier, les méthodes d'imputation de données manquantes, dont certaines reposent sur des modèles paramétriques explicites, sont caractérisées par l'absence d'un rôle pour les poids renfermés dans le plan de sondage. Un autre domaine d'étude, celui des estimations relatives aux petites régions ou aux petits secteurs, comporte souvent un mélange de procédés d'inférence basés sur le plan de sondage et un modèle. L'application des techniques d'inférence

au Census Bureau semble donc concorder avec le critère établi par DuMouchel et Duncan: il faut choisir un estimateur efficace (et simple) si le modèle hypothétique est vraisemblable ou un estimateur convergent si le modèle est inacceptable. On utilise le plus souvent des méthodes basées sur le plan de sondage pour faire des inférences à l'échelle nationale, tandis que les procédés basés sur un modèle servent souvent d'outils pour des analyses moins strictes ou des analyses dans lesquelles on espère qu'un modèle théorique est exact (quand des données sont manquantes).

3. MÉTHODE UTILISÉE AU CENSUS BUREAU POUR L'INFÉRENCE EN FONCTION DU PLAN DE SONDAGE DANS LES RÉGRESSIONS LINÉAIRES

En général, dans les travaux statistiques, la régression linéaire est sans doute la technique analytique la plus répandue. La plupart des données recueillies par le Census Bureau, en particulier celles provenant d'études démographiques qui révèlent des caractéristiques des personnes ou des logements, sont qualitatives. Le Census Bureau utilise donc la régression linéaire, sous quelque forme que ce soit, relativement moins qu'ailleurs.

Fuller [13] a élaboré les notions de base pour l'application des techniques d'inférence à la régression linéaire à partir de méthodes (de linéarisation) fondées sur le développement de Taylor. Ses résultats sont intégrés au programme informatique SUPER CARP [16], dont la mise au point à été en partie financée par le U.S. Bureau of the Census. Nous pouvons affirmer que nous avons utilisé ce logiciel avec succès, mais nous l'avons appliqué à seulement un petit nombre de problèmes jusqu'à présent. Le document de travail de Moore [26] offre probablement l'illustration la plus accessible de l'utilisation de SUPER CARP au Census Bureau.

Dans la section suivante, on explique la manière de définir les méthodes de duplication en fonction de poids d'échantillonnage artificiels (replicate weights) et on présente aussi des réflexions préliminaires sur la conception de logiciels pour exécuter les calculs nécessaires, quoiqu'on n'ait pas encore essayé de les mettre en oeuvre. Cette méthode n'est pas très différente de celle utilisée dans SUPER CARP, mais les méthodes de duplication aboutissent souvent à des valeurs un peu plus élevées des erreurs types que les techniques

de linéarisation. L'avantage de ce type de logiciel est qu'il permettrait d'exploiter les méthodes de duplication élaborées pour quelques-unes de nos enquêtes et de mieux tenir compte des effets des estimateurs complexes que les programmes fondés sur les procédés de linéarisation.

4. CALCUL DES VARIANCES CORRIGÉES POUR TENIR COMPTE DU PLAN DE SONDAGE À L'AIDE DE POIDS D'ÉCHANTILLONNAGE ARTIFICIELS

Les méthodes de duplication telles que le jackknife, la séparation en deux parties égales et le bootstrap représentent les principales solutions de rechange aux techniques de linéarisation pour calculer des estimations de la variance de paramètres non linéaires en fonction du plan de sondage. Kish et Frankel [18] ont publié un premier compte rendu des applications des méthodes de duplication à ce problème, et beaucoup de travaux ont été réalisés par la suite.

L'utilisation des méthodes de duplication pour l'estimation de la variance a eu des hauts et des bas. La linéarisation est une technique puissante, bien entendu, et les propriétés mises en évidence par Binder [1] favorisent son application à un grand éventail de modèles analytiques. Toutefois, les enquêtes du Census Bureau reposent en général sur des estimateurs assez complexes, et l'évaluation de l'effet total de la variance d'échantillonnage de ces estimateurs s'est souvent avérée une tâche qui nécessite beaucoup de temps de la part de spécialistes en statistique et, surtout, en programmation informatique. Récemment, les calculs de la variance dans diverses enquêtes ont pu être exécutés par des méthodes de duplication fondées sur des "poids d'échantillonnage artificiels" (replicate weights). Cette méthode est essentiellement un procédé général pour calculer la variance d'un grand nombre de statistiques dans les enquêtes et simplifier l'estimation de la variance des paramètres analytiques complexes.

L'utilisation de poids d'échantillonnage artificiels n'est pas une découverte récente. Quelques-unes des premières applications de cette technique sont résumées en [5], qui décrit également l'expérience acquise dans ce domaine par le U.S. Bureau of Labor Statistics, le Bureau of the Census et Westat, Inc. Pour un plan de sondage (qui peut être complexe) et un estima-

teur (qui peut être complexe), l'algorithme de cette méthode représente la variance sous la forme de données qui sont associées au fichier des données d'enquête plutôt que comme un ensemble de formules de variance (qui peuvent être complexes) pour lesquelles il faut composer un programme informatique. Les méthodes de duplication bien connues, comme la séparation de l'échantillon en deux parties égales et équilibrées et le jackknife, peuvent être formulées en fonction de poids d'échantillonnage artificiels, mais l'algorithme mentionné plus haut permet également de reproduire un gamme très étendue de plans de ré-échantillonnage (voir [7]). Le document [10] montre qu'il existe un plan de ré-échantillonnage (plus exactement, un nombre infini de plans de ré-échantillonnage) qui correspond essentiellement à n'importe lequel des estimateurs de la variance bien connus pour les estimations de totaux dans une population. De cette manière, on peut, par exemple, reproduire les expressions de la variance pour les plans de sondage à plusieurs degrés, les estimateurs de Yates-Grundy et ainsi de suite. La représentation des relations de variance complexes sous la forme de données rend le calcul des variances accessible à un plus grand nombre d'utilisateurs de données.

Dans un grand nombre d'enquêtes, des poids W_{i0} sont affectés à chaque unité i, de sorte que, pour toute variable X_i, l'estimation d'un total correspond à la somme pondérée du produit de la variable et des poids des unités d'échantillonnage:

$$\hat{X}_0 = \sum_i W_{i0} X_i. \tag{4.1}$$

Le résultat de la méthode des poids d'échantillonnage artificiels est un nouvel ensemble de poids W_{ir}, r = 1, ..., R, pour chaque unité d'échantillonnage i, à partir desquels on peut calculer une autre estimation d'un total:

$$\hat{X}_r = \sum_i W_{ir} X_i. \tag{4.2}$$

L'estimateur de la variance est

$$\hat{Var}(\hat{X}_0) = \sum_{r=1}^{R} d_r(\hat{X}_r - \hat{X}_0)^2 \qquad (4.3)$$

pour une valeur prédéterminée de d_r qui est indépendante du choix de la variable X. (Par exemple, dans un échantillon séparé en deux parties égales équilibrées, une manière simplifiée d'estimer la variance, abstraction faite des effets des estimations complexes qui peuvent influer sur les poids W_{i0}, consisterait à fixer les poids W_{ir} égaux soit à $2W_{i0}$, soit à 0, selon que l'unité i fait partie ou non du demi-échantillon, et de fixer $d_r = 1/R$ pour chaque r.) Plus généralement, pour une fonction lisse S des estimations pondérées de totaux à l'échelle de la population $\hat{X}_0^{(1)}, \ldots, \hat{X}_0^{(k)}$, qui ont toutes la forme (4.1),

$$\hat{Var}\{S(\hat{X}_0^{(1)}, \ldots, \hat{X}_0^{(r)})\} = \sum_{r=1}^{R} d_r\{S(\hat{X}_r^{(1)}, \ldots, \hat{X}_r^{(k)}) - S(\hat{X}_0^{(1)}, \ldots, \hat{X}_0^{(k)})\}^2 \qquad (4.4)$$

Dans l'équation (4.4), S peut être un des estimateurs extrêmement complexes qui sont souvent utilisés dans les enquêtes et peuvent inclure des corrections pour tenir compte des cas de non-interview et des estimations directes ou itératives par le quotient. En outre, si ces formes d'estimation complexe sont intégrées dans les poids W_i, elles peuvent également être incluses dans le calcul des valeurs de W_{ir}. Bref, le calcul de la variance dans cette méthode comprend trois étapes ou phases distinctes:

1. Détermination des poids d'échantillonnage artificiels de base $W_{ir}*$ pour le schéma de pondération simple sans biais des données (schéma de Horwitz-Thompson) qui correspond aux poids de base $W_{i0}*$.
2. Calcul des poids d'échantillonnage artificiels finals, W_{ir}, par l'application des mêmes rectifications pour les cas de non-interview et les estimations par le quotient, $W_{ir}*$, que dans les premières méthodes d'estimations utilisées pour obtenir les W_{i0} à partir des $W_{i0}*$.
3. Estimation de la variance de paramètres simples ou complexes à l'aide de l'équation (4.4).

La modularité de ces trois phases est un des grands avantages de cette technique; des programmes généraux peuvent servir à l'exécution des phases 1 et 2 ou des programmes spéciaux peuvent être mis au point si des problèmes particuliers se posent. Pour une même enquête, il est nécessaire d'exécuter les phases 1 et 2 seulement une fois. Les programmes conçus pour la phase 3 ne renferment pas de corrections en fonction du plan de sondage ou de l'estimateur et peuvent donc être utilisés au besoin par tout utilisateur qui dispose des poids artificiels W_{ir} produits à la deuxième phase.

La plupart des applications de cette méthode au Census Bureau ont porté sur l'estimation de la variance de caractéristiques fondamentales telles que des moyennes, des totaux ou des proportions dans des enquêtes, mais l'équation (4.4) est également utile pour des travaux analytiques. Cette méthode permet de représentés tous les effets des plans de sondage et des estimateurs complexes, mais, en pratique, l'application des techniques de linéarisation se limite souvent à des cas simples et assez fréquents. Par ailleurs, bien qu'on puisse élaborer des logiciels pour appliquer la linéarisation aux méthodes analytiques répandues telles que la régression linéaire, les modèles log-linéaires, les modèles linéaires généralisés, etc., la formule (4.4) permet de calculer des variances dans des modèles analytiques plus spécialisés pour lesquels il n'existe pas de programmes de linéarisation, car (4.4) nécessite seulement l'exécution d'algorithmes complets en fonction des nouvelles estimations produites à partir des poids artificiels.

5. MÉTHODES D'INFÉRENCE BASÉES SUR LE PLAN DE SONDAGE POUR LES MODÈLES LOG-LINÉAIRES

Les modèles log-linéaires, qui expriment le logarithme des fréquences théoriques des valeurs d'une variable qualitative sous la forme d'une fonction linéaire de paramètres inconnus, comprennent les modèles factoriels pour les classements recoupés de données qualitatives et les modèles logistiques, qui définissent la relation entre une variable qualitative dépendante ou plus et une combinaison quelconque de variables explicatives qualitatives et continues. Bishop, Fienberg et Holland [2] ont publié un des premiers ouvrages sur ce domaine qui se développe très rapidement.

Beaucoup de modèles log-linéaires, surtout ceux conçus pour des classements

recoupés complets de données qualitatives, comprennent un grand nombre de paramètres. Les trois problèmes d'inférence les plus fréquents sont:

1. Le calcul de l'erreur type et des intervalles de confiance pour chaque paramètre estimé.
2. La vérification de l'importance de l'effet d'ensembles précis de paramètres sur la qualité de l'ajustement d'un modèle.
3. La vérification de la validité globale de l'ajustement du modèle.

Dans un échantillon aléatoire simple, les principes classiques de la méthode du maximum de vraisemblance offrent une solution à ces problèmes, quoique le test du khi-carré de Pearson soit, à juste titre, plus répandu que le test du khi-carré lié au rapport de vraisemblance pour résoudre le troisième problème.

Koch, Freeman et Freeman [9] ont appliqué la méthode des moindres carrés pondérés (MCP) aux échantillons complexes pour obtenir une solution à chacun des trois principaux problèmes d'inférence décrits plus haut. Cette méthode s'est avérée très utile en général, mais, dans certains cas, elle est limitée par la nécessité de produire des estimations très précises corrigées en fonction du plan de sondage pour la covariance des estimations calculées à partir de l'échantillon avant que ses propriétés asymptotiques puissent reproduire les résultats des MCP. (Pour d'autres observations sur les limites des MCP, voir [8] et [11].)

Fellegi [12] a conçu un des premiers tests pour évaluer les résultats de la méthode des MCP dans des cas particuliers. Plus récemment, Rao et Scott [20], [21] ont formulé et étendu un ensemble de techniques connexes pour des tests d'une classe générale de modèles comprenant les modèles log-linéaires. Statistique Canada a participé activement à l'élaboration de ces méthodes.

Un test moins bien connu unit la technique du jackknife et la variable khi-carré [11] pour résoudre le problème général des tests d'hypothèses modifiés en fonction du plan de sondage. Ce test repose sur la duplication de l'échantillon, et l'équation (4.4) et une expression semblable pour l'approximation du biais d'ordre premier (comme d'habitude dans le jackknife) sont utilisées pour faire des inférences approximatives concernant la distribution postulée par l'hypothèse nulle dans les tests habituels du khi-carré appliqués

directement aux estimations pondérées d'une enquête. Cette méthode ressemble beaucoup aux procédés élaborés par Rao et Scott. Aucune comparaison exhaustive des avantages relatifs du test par la méthode du jackknife et des tests proposés par Rao et Scott n'a été faite jusqu'à présent, mais il semble à première vue qu'ils soient tous bons et que ni l'un, ni l'autre ne soit supérieur. (D'autres observations à ce sujet sont présentées dans [11].)

Les tests par la méthode du jackknife semblent toutefois un peu plus faciles à appliquer, surtout quand un tableau contient un grand nombre de cases. Un programme écrit en FORTRAN et intitulé CPLX ([8] le décrit et [9] en explique le mode d'utilisation) comprend des tests par le jackknife pour l'ajustement de modèles factoriels log-linéaires aux classements recoupés et est maintenant à la disposition du public. Ce programme calcule également des erreurs types pour les paramètres de modèles log-linéaires par une méthode de duplication, ce qui permet de résoudre le premier des trois problèmes d'inférence examinés plus haut. CPLX est très pratique quand les autres variances des paramètres d'une enquête sont aussi estimées à l'aide de techniques de duplication, comme celles fondées sur les poids d'échantillonnage artificiels décrits à la section précédente, mais ce n'est pas là une condition obligatoire pour l'utilisation de ce programme et beaucoup de spécialistes à l'intérieur et à l'extérieur du Census Bureau l'ont appliqué à divers types d'analyses.

L'auteur espère pouvoir éventuellement intégrer la méthodologie de Rao et Scott dans un programme comme CPLX afin que les utilisateurs de données puissent aussi disposer de cette technique. Pour le moment, toutefois, la version actuelle de CPLX devrait aider les chercheurs qui veulent appliquer des procédés d'inférence basés sur le plan de sondage aux données d'enquête.

BIBLIOGRAPHIE

[1] Binder, D.A. (1983). On the variances of asymptotically normal estimators from Complex Samples. International Statistical Review 51, pp. 279-292.

[2] Bishop, Y.M.M., Fienberg, S.E. et Holland, P.W. (1975). Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

[3] Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: John Wiley.

[4] Cochran, W.G. (1977). Sampling Techniques (3e édition). New York: John Wiley.

[5] Dippo, C.S., Fay, R.E. et Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Communication présentée à la réunion annuelle de l'American Statistical Association, section des techniques d'enquête.

[6] DuMouchel, W.H. et Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association 78, pp. 535-543.

[7] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphie, PA: Society for Industrial and Applied Mathematics.

[8] Fay, R.E. (1982). Contingency Tables for Complex Designs: CPLX. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 44-53.

[9] Fay, R.E. (1983). CPLX - Contingency Tables Analysis for Complex Sample Designs, Program Documentation. Document non publié, Washington, D.C.: U.S. Bureau of the Census.

[10] Fay, R.E. (1984). Some Properties of Estimates of Variance Based on Replication Methods. Communication présentée à la réunion annuelle de l'American Statistical Association, section sur les techniques d'enquête.

[11] Fay, R.E. (1984). A Jackknifed Chi-Square Test for Complex Samples. À paraître dans la revue Journal of the American Statistical Association.

[12] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. Journal of the American Statistical Association 75, pp. 261-268.

[13] Fuller, W.A. (1975). Regression Analysis for Sample Survey. Sankhyā C 37, pp. 117-132.

[14] Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). Sample Survey Methods and Theory, Volumes I and II. New York: John Wiley.

[15] Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association 78, pp. 776-793.

[16] Hidiroglou, M.A., Fuller W.A. et Hickman, R.D. (1978). Super Carp (3e édition). Ames, IO: Statistical Laboratory, Iowa State University.

[17] Kish, L. (1965). Survey Sampling. New York: John Wiley.

[18] Kish, L. et Frankel, M.R. (1974). Inference from Complex Samples. Journal of the Royal Statistical Society, Ser. B 36, pp. 1-37.

[19] Koch, G.G.. Freeman, D.H. et Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Samples. International Statistical Review 43, pp. 59-78.

[20] Rao, J.N.K. et Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. Journal of the American Statistical Association 76, pp. 221-230.

[21] Rao, J.N.K. et Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Annals of Statistics 12. pp. 46-60.

[22] Rosenbaum, P.R.R. et Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies. Biometrika 70, pp. 41-55.

[23] Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics 6, pp. 34-58.

[24] Rubin, D.B. (1983). Comment: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys. Journal of the American Statistical Association 78, pp. 803-805.

[25] Särndal, C.-E. (1978). Design-Based and Model-Based Inference in Survey Sampling. Scandinavian Journal of Statistics 5, pp. 27-52.

[26] U.S. Bureau of the Census (1982). Preliminary Evaluation Results. Memorandum No. 31: Evaluating the Public Information Campaign for the 1980 Census - Results of the 1980 KAP Survey. Document rédigé par Jeffrey C. Moore, Washington, D.C.

APPLICATION DE LA MÉTHODE DES MOINDRES CARRÉS ET DE TECHNIQUES CONNEXES AUX PLANS DE SONDAGE COMPLEXES

Wayne A. Fuller[1]

1. INTRODUCTION ET DESCRIPTION DU MODÈLE

Supposons qu'un échantillon de grappes d'unités élémentaires est prélevé dans une population finie divisée en L strates. Le nombre total de grappes (unités primaires d'échantillonnage) dans l'échantillon, n, est défini par l'expression

$$n = \sum_{h=1}^{L} n_h, \tag{1}$$

où $n_h \geq 2$ est le nombre de grappes tirées de la $h^{\text{ième}}$ strate. Un vecteur colonne de caractéristiques

$$\underset{\sim}{Y}_{hij} = (Y_{hij1}, Y_{hij2}, \ldots, Y_{hijp})' \tag{2}$$

est observé pour la $j^{\text{ième}}$ unité élémentaire appartenant à la $i^{\text{ième}}$ grappe de la $h^{\text{ième}}$ strate. La forme du vecteur $\underset{\sim}{Y}_{hij}$ est assez générale. Ainsi, certains éléments de ce vecteur peuvent être des puissances de produits d'autres composantes du même vecteur. Un élément peut aussi, et ce sera souvent le cas, être identiquement égal à l'unité. Les totaux pour chaque grappe sont calculés à l'aide de la formule

$$\underset{\sim}{Y}_{hi.} = \sum_{j=1}^{m_{hi}} \underset{\sim}{Y}_{hij}, \tag{3}$$

où m_{hi} est le nombre d'unités élémentaires dans la $hi^{\text{ième}}$ grappe.

Nous voulons nous pencher sur le comportement de fonctions continues

---

[1] Wayne A. Fuller, département de statistique, Iowa State University.

localement d'une fonction linéaire du vecteur des moyennes de chaque grappe

$$\hat{\underset{\sim}{\theta}} = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{i=1}^{n_h} \underset{\sim}{Y}_{hi.} \tag{4}$$

où les W_h sont des poids fixes. Souvent, ces poids valent

$$W_h = N_h N^{-1} \tag{5}$$

où N_h est le nombre de grappes dans la $h^{\text{ième}}$ strate et N est le nombre total de grappes au niveau de la population. Pour les poids définis dans la formule (5), la fonction linéaire en (4) est l'estimateur centré habituellement utilisé pour calculer la moyenne par grappe pour une population finie. Un autre groupe de poids qui présente beaucoup d'intérêt est l'ensemble de poids unitaires

$$W_h = n^{-1} n_h. \tag{6}$$

Notre modèle nous permet d'examiner des fonctions de la moyenne par unité élémentaire. L'estimateur habituel de la moyenne d'une des variables Y au niveau des unités élémentaires est le rapport entre la moyenne de la variable au niveau des grappes et le nombre moyen d'unités par grappe. Le nombre moyen d'unités par grappe est la moyenne au niveau des grappes d'une variable Y qui est identiquement égale à l'unité.

On peut facilement étendre notre modèle pour inclure diverses formes de sous-échantillonnage à l'intérieur des grappes. Étant donné que ce genre d'extension n'augmente guère la généralité du développement et qu'elle compliquerait beaucoup les notations, nous limiterons notre description à l'échantillonnage à un degré à l'intérieur des strates.

Notre analyse repose fondamentalement sur le théorème central limite suivant pour les échantillons tirés dans une population finie.

Théorème 1. Soit $\{\xi_r : r = 1, 2, \ldots\}$ une série de populations finies stratifiées. Définissons la population de la $h^{\text{ième}}$ strate de la $r^{\text{ième}}$ population

comme un échantillon aléatoire de taille $N_{rh} \geq N_{r-1.h}$ choisi à partir d'une population infinie à p dimensions admettant des moments absolus d'ordre $2 + \delta$, où $\delta > 0$, qui sont bornés en vertu de l'inégalité $M_\delta < \infty$. Soit $\underset{\sim}{\Sigma}_{rh}$ la matrice des covariances de la $rh^{\text{ième}}$ population infinie. Soit $L_r \geq L_{r-1}$ le nombre de strates dans chaque population finie et supposons qu'un échantillon aléatoire simple de n_{rh} unités ($n_{rh} \geq 2$ et $n_{rh} \geq n_{r-1.h}$) est tiré dans la $h^{\text{ième}}$ strate. Soit $f_{rh} = N_{rh}^{-1} n_{rh}$ une matrice triangulaire dans lequel

$$0 \leq f_{rh} < M_{fu} < 1,$$

où M_{fu} est un nombre fixe. Soit $Y_{rhi.}$ le total correspondant à la $i^{\text{ième}}$ grappe tirée dans la $h^{\text{ième}}$ strate de la $r^{\text{ième}}$ population et définissons

$$\hat{\underset{\sim}{\theta}}_r = \sum_{h=1}^{L_r} W_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \underset{\sim}{Y}_{rhi.},$$

$$\underset{\sim}{\theta}_{rf} = \sum_{h=1}^{L_r} W_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}} Y_{rhi.},$$

$$\underset{\sim}{\theta}_r = \sum_{h=1}^{L_r} W_{rh} \underset{\sim}{\mu}_{.h..},$$

où $\underset{\sim}{\theta}_{rf}$ est un paramètre de la population finie et $\mu_{.h..}$ est la moyenne de la population infinie utilisée pour produire la $h^{\text{ième}}$ strate de la population finie. Nous supposons également que

$$0 < M_{SL} < \left| n_r \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \underset{\sim}{\Sigma}_{rh} \right| < M_{SU} < \infty,$$

où les bornes inférieure et supérieure M sont des nombres fixes: on suppose que

$$n_r = \sum_{h=1}^{L_r} n_{rh} \longrightarrow \infty,$$

$$\sup_h \left[\sum_{t=1}^{L_r} W_{rt}^2 n_{rt}^{-1}\right]^{-1} W_{rh}^2 n_{rh}^{-2} \longrightarrow 0,$$

à mesure que $r \longrightarrow \infty$, ou W_{rh} est une matrice trianqulaire de poids. On peut écrire

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\}]^{-\frac{1}{2}} (\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}]^{-\frac{1}{2}} (\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

où

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\} = \sum_{h=1}^{L_r} W_{rh}^2 (1 - f_{rh}) n_{rh}^{-1} \hat{\underset{\sim}{\Sigma}}_{rh}.$$

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\} = \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \hat{\underset{\sim}{\Sigma}}_{rh}.$$

$$\hat{\underset{\sim}{\Sigma}}_{rh} = (n_{rh} - 1)^{-1} \sum_{i=1}^{n_{rh}} (\underset{\sim}{Y}_{rhi.} - \bar{\underset{\sim}{Y}}_{rh..})(\underset{\sim}{Y}_{rhi.} - \bar{\underset{\sim}{Y}}_{rh..})'.$$

$$\bar{\underset{\sim}{Y}}_{rhi.} = n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \underset{\sim}{Y}_{rhi.}.$$

La démonstration de ce théorème découle des théorèmes 1 et 2 de Fuller (1975) et elle est applicable aux échantillons à plusieurs degrés. (Voir aussi Krewski et Rao (1981) et Isaki et Fuller (1982)).

La plupart des cas que nous examinerons portent sur des fonctions continues de $\hat{\underset{\sim}{\theta}}$.

Corollaire 1. Supposons que les hypothèses du théorème 1 sont justes. Soit $\underset{\sim}{g}(\underset{\sim}{\theta})$ une fonction à valeurs vectorielles qui dépend de $\underset{\sim}{\theta}$, où $\underset{\sim}{g}(\underset{\sim}{\theta})$ est continue et admet des dérivées premières continues pour $\underset{\sim}{\theta}$ dans la sphère $|\underset{\sim}{\theta} - \underset{\sim}{\theta}_r| \leq \delta$ pour tout r. où $\delta > 0$ est une valeur fixe. Soit $\underset{\sim}{G}(\underset{\sim}{\theta})$ une matrice non singulière contenant les dérivées premières de $\underset{\sim}{g}(\underset{\sim}{\theta})$, où le $ij^{\text{ième}}$ élément de $\underset{\sim}{G}(\underset{\sim}{\theta})$ est

$$\frac{\partial g_i(\underset{\sim}{\theta})}{\partial \theta_j},$$

$g_i(\underset{\sim}{\theta})$ est le $i^{\text{ième}}$ élément de $\underset{\sim}{g}(\underset{\sim}{\theta})$ et θ_j est le $j^{\text{ième}}$ élément de $\underset{\sim}{\theta}$. On peut écrire

$$[\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_T)\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_T - \underset{\sim}{\theta}_{Tf}\}\underset{\sim}{G}'(\hat{\underset{\sim}{\theta}}_T)]^{-\frac{1}{2}}[\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_T) - \underset{\sim}{g}(\underset{\sim}{\theta}_{Tf})] \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}).$$

$$[\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_T)\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_T - \underset{\sim}{\theta}_T\}\underset{\sim}{G}'(\hat{\underset{\sim}{\theta}}_T)]^{-\frac{1}{2}}[\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_T) - \underset{\sim}{g}(\underset{\sim}{\theta}_T)] \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}).$$

Le corollaire 1 s'applique à l'estimateur taylorien de la variance de la distribution approximative de $\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_T) - \underset{\sim}{g}(\underset{\sim}{\theta}_T)$. Il est également possible d'évaluer la variance à l'aide d'estimateurs bien définis pour des échantillons répétés. Quelques-unes des techniques utilisées sont la méthode des duplications successives par blocs (balanced repeated replication: voir McCarthy (1969)), les méthodes jackknife (voir Miller (1974)) et les méthodes bootstrap (voir Efron (1979, 1981)). Bien que ces méthodes puissent être adaptées à la structure de l'échantillon, cette adaptation n'est pas toujours directe (voir Rao et Wu (1983)).

Une classe de fonctions continues de $\hat{\underset{\sim}{\theta}}$ qui mérite une attention spéciale est celle qu'on obtient quand $\hat{\underset{\sim}{\theta}}$ est la variable dépendante dans un ajustement par la méthode des moindres carrés généralisés.

Corollaire 2. Supposons que les hypothèses du théorème 1 sont justes. Définissons θ de manière à satisfaire la condition

$$\underset{\sim}{\theta} = \underset{\sim}{h}(\underset{\sim}{\alpha}),$$

où α est un vecteur de dimension $\underset{\sim}{k}$ ($k \leq p$), $\underset{\sim}{h}(\underset{\sim}{\alpha})$ est une fonction continue de $\underset{\sim}{\alpha}$ et il existe des dérivées première et seconde pour tout $\underset{\sim}{\alpha}$ dans une sphère ouverte contenant la vraie valeur $\underset{\sim}{\alpha}_T$ pour tout r. Définissons l'espace des paramètres de $\underset{\sim}{\alpha}$ comme un sous-ensemble borné ouvert de l'espace euclidien à k dimensions. Soit $\hat{\underset{\sim}{\alpha}}_T$ le vecteur qui minimise

$$[\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{h}(\underset{\sim}{\alpha}_r)]'\hat{\underset{\sim}{V}}^{-1}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}[\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{h}(\underset{\sim}{\alpha}_r)].$$

On a alors

$$[\hat{\underset{\sim}{V}}\{\underset{\sim}{\alpha}_r\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\alpha}}_r - \underset{\sim}{\alpha}_r) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}).$$

où $$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\alpha}}_r\} = [\underset{\sim}{H}(\hat{\underset{\sim}{\alpha}}_r)\hat{\underset{\sim}{V}}^{-1}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}\underset{\sim}{H}'(\hat{\underset{\sim}{\alpha}}_r)]^{-1},$$

et $\underset{\sim}{H}(\hat{\underset{\sim}{\alpha}}_r)$ est la matrice des dérivées premières de $\underset{\sim}{h}(\underset{\sim}{\alpha})$ par rapport à $\underset{\sim}{\alpha}$ au point $\hat{\underset{\sim}{\alpha}}$.

2. MOYENNES, QUOTIENTS ET RÉGRESSIONS

Une application élémentaire du théorème 1 est l'estimation de la moyenne par grappe et l'établissement de limites de confiance approximatives pour cette moyenne. Souvent, le paramètre clé pour l'estimateur de la moyenne est la moyenne au niveau des grappes pour la population finie, ce qui nous oblige à inclure la correction d'échantillonnage pour population finie $(1 - f_h)$ dans l'estimateur de la variance.

Une application un peu plus compliquée est l'estimation de la différence entre les valeurs de la moyenne par grappe obtenues pour deux domaines. Si nous définissons

Y_{hij1} = observation de la caractéristique d'intérêt si l'unité hij appartient au domaine 1
= 0 autrement,

Y_{hij2} = observation de la caractéristique d'intérêt si l'unité hij appartient au domaine 2
= 0 autrement,

Y_{hij3} = 1 si l'unité hij appartient au domaine 1
= 0 autrement et

Y_{hij4} = 1 si l'unité hij appartient au domaine 2
= 0 autrement,

l'estimation de la différence entre les valeurs de la moyenne par unité calculée pour les deux domaines est

$$g(\hat{\underset{\sim}{\theta}}) = g(\bar{\underset{\sim}{Y}}_{...}) = \bar{Y}^{-1}_{...3}\bar{Y}_{...1} - \bar{Y}^{-1}_{...4}\bar{Y}_{...2}. \tag{7}$$

Deux méthodes sont souvent utilisées pour calculer l'estimateur taylorien de la variance. Dans la première, l'estimateur du corollaire 1 est calculé directement à partir des matrices $G(\hat{\theta}_r)$ et $V\{\hat{\theta}_r - \theta_r\}$ ou $V\{\hat{\theta}_r - \theta_{rf}\}$. Une démarche algébriquement identique consiste à définir les observations de la manière suivante:

$$\underset{\sim}{Z}(\underset{\sim}{Y}_{hi.}, \hat{\underset{\sim}{\theta}}) = \hat{\underset{\sim}{Z}}_{hi} = \underset{\sim}{G}(\hat{\underset{\sim}{\theta}})(\underset{\sim}{Y}_{hi.} - \bar{\underset{\sim}{Y}}_{h..}) \tag{8}$$

et de calculer l'estimateur stratifié habituel de la variance de la moyenne par grappe pour $\hat{\underset{\sim}{Z}}_{hi}$.

$$\hat{\underset{\sim}{V}}\{\hat{\bar{\underset{\sim}{Z}}}_{...}\} = \hat{\underset{\sim}{V}}\{\underset{\sim}{g}(\bar{\underset{\sim}{Y}}_{...})\}$$

$$= \sum_{h=1}^{L} W_h^2(1 - f_h)n_h^{-1}(n_h - 1)^{-1} \sum_{i=1}^{n_h} (\hat{\underset{\sim}{Z}}_{hi} - \hat{\bar{\underset{\sim}{Z}}}_{h.})(\hat{\underset{\sim}{Z}}_{hi} - \hat{\bar{\underset{\sim}{Z}}}_{h.})', \tag{9}$$

où

$$\hat{\bar{\underset{\sim}{Z}}}_{..} = \sum_{h=1}^{L} W_h \hat{\bar{\underset{\sim}{Z}}}_{h.},$$

$$\hat{\bar{\underset{\sim}{Z}}}_{h.} = n_h^{-1} \sum_{i=1}^{n_h} \hat{\underset{\sim}{Z}}_{hi}.$$

Par exemple, l'algorithme de calcul (9) est mis en oeuvre dans SUPER CARP. (Voir Hidiroglou et coll.. 1980, p. 32.)

L'analyste peut vouloir tirer des conclusions statistiques sur la population finie qui a été échantillonnée ou sur la superpopulation, quand il travaille avec des grandeurs telles que les différences entre des moyennes.

Une des utilisations analytiques les plus fréquentes des données d'enquête est l'ajustement d'équations de régression. On peut même exprimer la différence entre les moyennes de différents domaines sous la forme d'un coefficient de régression. Le vecteur de coefficients de régression a la forme de $\underset{\sim}{g}(\hat{\underset{\sim}{\theta}})$, variable décrite à la section précédente, mais il peut être avantageux de décomposer le vecteur $\underset{\sim}{Y}$ défini à la section 1 en plusieurs partitions et de fournir des expressions explicites pour les coefficients de régression. L'équation de régression peut s'écrire de la manière suivante:

$$Y_{hij} = \underset{\sim}{X}'_{hij}\underset{\sim}{\beta} + e_{hij}, \tag{10}$$

où Y_{hij} est la variable dépendante et $\underset{\sim}{X}_{hij}$ est un vecteur de dimension k contenant les variables explicatives. L'estimateur des moindres carrés pondérés de $\underset{\sim}{\beta}$ est

$$\hat{\underset{\sim}{\beta}}_W = \left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}W_{hij}\underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}W_{hij}Y_{hij}. \tag{11}$$

Les poids W_{hij} peuvent être une fonction de hij, mais nous supposons ici que les poids sont fixes en ce sens qu'ils dépendent seulement de la numérotation des unités élémentaires. On exclut ainsi (sauf à titre d'approximation) l'utilisation de poids qui varient en fonction des autres unités de l'échantillon.

Si on adopte quelques hypothèses non restrictives sur les moments de la superpopulation à laquelle la population finie appartient, on peut appliquer le théorème 1 à l'estimateur défini par la formule (11). Soit π_{hij} la probabilité de sélection: l'estimateur $\hat{\underset{\sim}{\beta}}_W$ est un estimateur convergent du vecteur

$$\underset{\sim}{B}_f = \left[\sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}W_{hij}\pi_{hij}\underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}W_{hij}\pi_{hij}Y_{hij}. \tag{12}$$

pour la population finie. Il découle de (12) que l'estimateur (11) est un estimateur convergent du coefficient de régression pour la population finie

quand les valeurs de W_{hij} sont proportionnelles à l'inverse des probabilités de sélection. L'erreur comprise dans $\hat{\underset{\sim}{\beta}}_W$ quand on l'utilise pour estimer $\underset{\sim}{B}_f$ est

$$\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{B}_f = [\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij}]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} v_{hij} \tag{13}$$

où $\quad v_{hij} = Y_{hij} - \underset{\sim}{X}'_{hij} \underset{\sim}{B}_f.$

En vertu du théorème 1 et du corollaire 1, un estimateur convergent de la variance de la distribution approximative de $\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{\beta}$ est

$$\hat{V}\{\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{\beta}\} = \hat{\underset{\sim}{A}}^{-1} \hat{\underset{\sim}{G}} \hat{\underset{\sim}{A}}^{-1}, \tag{14}$$

où

$$\hat{\underset{\sim}{A}} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij},$$

$$\hat{\underset{\sim}{G}} = (n-1)(n-k)^{-1} \sum_{h=1}^{L} n_h (n_h - 1)^{-1} \sum_{i=1}^{n_h} \hat{\underset{\sim}{d}}_{hi} \hat{\underset{\sim}{d}}'_{hi.},$$

$$\hat{\underset{\sim}{d}}_{hi.} = \sum_{j=1}^{m_{hi}} \hat{\underset{\sim}{d}}_{hij},$$

$$\hat{d}_{hij} = W_{hij} \underset{\sim}{X}_{hij} \hat{v}_{hij},$$

$$n = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi},$$

$$\hat{v}_{hij} = Y_{hij} - \underset{\sim}{X}'_{hij} \hat{\underset{\sim}{\beta}}_W,$$

et $\underset{\sim}{\beta}$ est l'homologue de $\underset{\sim}{B}_f$ au niveau de la superpopulation. Cette forme particulière de l'estimateur de la variance a été proposée par Fuller (1975) et elle est utilisée par SUPER CARP.

Une des questions que les statisticiens se posent souvent quand ils ana-

lysent des enquêtes est : "Faut-il tenir compte des poids d'échantillonnage dans l'ajustement d'une équation de régression?" Comme pour la plupart des questions de ce genre. la réponse dépend des circonstances. Le fait que cette question soit posée indique généralement que le chercheur veut faire des inférences sur une population plus grande que la population finie qui a été échantillonnée. Cela ne signifie pas pour autant que la superpopulation d'intérêt est parfaitement définie ou définissable. Il en découle toutefois que le chercheur envisage la population finie comme étant générée par une superpopulation dans laquelle un modèle linéaire peut décrire la réalité. Une expression quantitative de l'hypothèse selon laquelle il n'est pas nécessaire de prendre en compte les poids d'échantillonnage est l'hypothèse suivante concernant la superpopulation:

$$H_0: \underset{\sim}{\theta}_\pi = \underset{\sim}{\theta}_{(1)}, \tag{15}$$

où les $\underset{\sim}{\theta}$ sont les homologues de (12) au niveau de la superpopulation et

$$\underset{\sim}{\theta}_\pi = \left[\sum_{h=1}^{L}\sum_{i=1}^{N_h} E_\xi \sum_{j=1}^{m_{hi}} \{\underset{\sim}{X}_{hij}\pi_{hij}\underset{\sim}{X}'_{hij}\}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{N_h} E_\xi\{\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}\pi_{hij}Y_{hij}\},$$

$$\underset{\sim}{\theta}_{(1)} = \left[\sum_{h=1}^{L}\sum_{i=1}^{N_h} E_\xi\{\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}\underset{\sim}{X}'_{hij}\}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{N_h} E_\xi\{\sum_{j=1}^{m_h} \underset{\sim}{X}_{hij}Y_{hij}\}, \tag{16}$$

et E_ξ est l'opérateur d'espérance mathématique par rapport à la superpopulation. Cette hypothèse est vérifiable. Il semble que le moins qu'on puisse faire, c'est d'effectuer un test de cette hypothèse si on fait une analyse non pondérée d'un échantillon tiré avec des probabilités de sélection inégales.

Si l'hypothèse nulle comprend également l'hypothèse selon laquelle l'estimateur calculé en tenant compte des poids unitaires est l'estimateur à variance minimale, le test de cette hypothèse repose sur la variable

$$F^k_{n-L-2k} = k^{-1}\hat{\underset{\sim}{\delta}}'_2\hat{\underset{\sim}{V}}^{-1}_{22}\hat{\underset{\sim}{\delta}}_2, \tag{17}$$

où

$$(\hat{\delta}_1', \hat{\delta}_2')' = [\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{Z}_{hij}\underset{\sim}{Z}_{hij}']^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{Z}_{hij}Y_{hij},$$

$$\underset{\sim}{Z}_{hij}' = (\underset{\sim}{X}_{hij}', \underset{\sim}{X}_{hij}'W_{hij}),$$

et

$$\hat{\underset{\sim}{V}} = \begin{pmatrix} \hat{\underset{\sim}{V}}_{11} & \hat{\underset{\sim}{V}}_{12} \\ \hat{\underset{\sim}{V}}_{21} & \hat{\underset{\sim}{V}}_{22} \end{pmatrix} \tag{18}$$

est défini dans l'équation (14), sauf que $\underset{\sim}{Z}_{hij}$ remplace $\underset{\sim}{X}_{hij}$. Tel que l'indique la notation, le critère de ce test suit approximativement la loi du F de Snedecor avec k et n - L - 2k degrés de liberté.

Exemple 1. Le tableau 1 contient des observations recueillies sur trente-sept aires agricoles par le Statistical Reporting Service du département de l'agriculture du gouvernement des États-Unis dans la région centre-nord de l'Iowa en 1978. Deux séries de chiffres sur le nombre d'hectares de soja figurent dans ce tableau. La première série a été obtenue au moyen d'interviews sur place dans le cadre d'une enquête menée au mois de juin (June Enumerative Survey). La deuxième série provient d'une classification de données transmises par le satellite Landsat et analysées par un système de répartition mis au point par le Statistical Reporting Service. Ces chiffres sont tirés d'une étude qui avait pour objet la construction d'un estimateur du nombre total d'acres de soja par la méthode de régression. Nous utilisons ces données ici pour illustrer le calcul de paramètres de régression à partir de données d'enquête. L'échantillon ressemble de très près à un échantillon stratifié où les strates correspondent aux territoires numérotés dans la colonne des comtés. L'inverse de la fraction de sondage est indiqué dans la colonne des poids. L'ajustement par l'estimateur (11) d'une équation pour la régression du nombre d'hectares déclaré dans les interviews par rapport au nombre d'hectares mesuré par satellite conduit au résultat suivant:

$$\hat{Y} = \underset{(8.332)}{-11.845} + \underset{(0.0922)}{1.1602}X,$$

où les nombres entre parenthèses sont les erreurs types évaluées à partir de la matrice des covariances estimées qui est calculée à l'aide de l'équation (14). Le logiciel SUPER CARP a été utilisé pour exécuter tous les calculs nécessaires. Si on introduit les poids unitaires dans les équations (11) et (14), on obtient les résultats suivants pour l'équation de régression et les erreurs types respectivement:

$$\hat{Y} = \underset{(9.282)}{-3.927} + \underset{(0.0963)}{1.0850}X.$$

Si on calcule la variable F définie par l'équation (17), le résultat est

$$F^2_{23} = 2.81.$$

À première vue, cette valeur est suffisamment grande pour mettre en doute l'égalité des deux coefficients. À cause de la petite taille de cet échantillon et de la structure des poids, ce test est presque l'équivalent d'une comparaison de deux droites, soit une droite pour le premier comté et une droite moyenne pour les autres comtés. Dans ce petit échantillon, les écarts de la droite du premier comté sont faibles. Par conséquent, les estimations des erreurs types des coefficients des deux variables additionnées sont peu élevées. Ce phénomène est abordé de nouveau à la section 3. Si on calcule la variable F habituellement utilisée pour la régression dans un test où on suppose que les variances des erreurs sont homogènes et qu'on fait abstraction de la stratification, on obtient

$$F^2_{33} = 0.68.$$

Bien que cette variable ne suive pas la loi du F de Snedecor, elle nous permet de croire plus allégrement que les deux régressions pondérées faites plus haut conduisent à la même équation.

Le tableau 2 contient les erreurs types des coefficients de régression estimés par différentes méthodes. Les erreurs types estimées de la constante ont à peu près le comportement qu'on pourrait prévoir. La régression pondérée

pour échantillons stratifiés produit la plus petite erreur type estimée, tandis que la régression pour échantillons stratifiés faite à partir des poids unitaires et la régression des moindres carrés ordinaires ont les deuxième et troisième erreurs types les moins élevées. N'oublions pas que ces erreurs types sont des estimations. Les deux régressions pour échantillons stratifiés produisent un résultat convergent si on accepte le schéma de stratification. L'estimateur pondéré a la plus petite variance parce que les observations appartenant à la strate 1, celle qui a le poids le plus élevé, sont plus proches de la droite estimée que les points correspondant aux autres strates.

L'estimation de l'erreur type par la méthode des moindres carrés ordinaires indique un manque de convergence dans cette estimation du modèle stratifié. Si on envisage l'échantillon comme un échantillon de comtés divisé en grappes, les estimations de l'erreur type de la constante sont d'environ 30% à 40% plus élevées que celles calculées pour l'échantillon stratifié.

Les erreurs types estimées de la pente ont toutefois un comportement différent. L'erreur type estimée la plus faible est associée à l'estimation calculée pour un échantillon en grappes à partir des poids unitaires, alors que l'erreur type la plus élevée correspond aux moindres carrés ordinaires. En gros, la variation de la pente d'une grappe à l'autre est relativement moins grande que la variation à l'intérieur des grappes. Étant donné qu'il existe une corrélation inverse entre les poids et la variabilité observée, les estimateurs pondérés ont de faibles variances estimées. Cet échantillon est petit, mais il suffit de démontrer que les poids unitaires ne produisent pas toujours des variances plus faibles que les poids d'échantillonnage et que la stratification et la division en grappes peuvent avoir des effets assez complexes sur les variances estimées des coefficients de régression.

3. QU'EST-CE QU'UN GRAND ÉCHANTILLON?

Les raisonnements présentés dans les sections précédentes sont fondés sur les propriétés des estimateurs des paramètres et de la variance dans les grands échantillons. Si on utilise la loi normale limite pour établir des intervalles de confiance, la taille d'échantillon requise pour une bonne approximation dépend de la nature de la population mère. Par exemple, si la

caractéristique observée est rare et se mesure avec une variable binaire (supposons que la probabilité d'obtenir une valeur non nulle est inférieure à 0.05), il est nécessaire de prélever un très grand échantillon (plus de 1,400 unités dans le cas d'un échantillon aléatoire simple (Cochran, 1977, p. 58)) pour une approximation basée sur la loi normale. Une variable qui suit une loi binomiale et a une faible probabilité d'être non nulle est seulement un exemple des éléments fortement asymétriques que les praticiens de l'échantillonnage rencontrent souvent dans les populations qu'ils étudient.

Des grandeurs telles que le chiffre d'affaires brut d'entreprises, le nombre d'employés dans des entreprises, le nombre d'animaux par ferme et le revenu familial sont des valeurs qui appartiennent à des populations asymétriques dont il faut prélever de grands échantillons avant que la distribution de la moyenne tende vers une forme gaussienne. En revanche, la distribution de la moyenne de variables telles que la taille des familles peut être approximativement normale quand la taille de l'échantillon est petite (moins de cent unités).

L'utilisation du développement de Taylor est une méthode semi-non paramétrique en ce sens que cette approximation est valable, dans les grands échantillons, si on adopte quelques hypothèses très peu restrictives au sujet de la population. Les propriétés des grands échantillons sont assurées s'il n'y a pas de points isolés dans notre espace échantillon. Cette méthode peut s'avérer piètre dans les cas où à cause de la distribution génératrice et de la taille de l'échantillon, une observation ou plus est isolée du nuage de points principal. Nous examinerons maintenant le problème d'estimer la variance du vecteur de coefficients de régression utilisé pour évaluer l'effet de la pondération sur ces coefficients dans l'exemple présenté plus haut sur la culture du soja. Le vecteur original est

$$(1, X, XW, W)$$

et l'hypothèse à vérifier est que les coefficients de XW et W sont nuls. Pour illustrer les problèmes que pose l'estimation de la variance du vecteur de coefficients calculés pour l'ensemble de données sur la culture du soja, nous définissons un vecteur qui est orthogonal dans la métrique des poids unitaires. La matrice des observations des variables indépendantes

transformées contient les résidus obtenus dans la régression de chaque variable, sauf la première, par rapport aux éléments qui la précèdent dans le vecteur original. Les valeurs des variables de régression transformées ($X - \bar{X}$, RWX, RW) figurent au tableau 3. On a retenu seulement quelques chiffres pour faciliter la lecture de ce tableau.

La régression de Y par rapport à (1, $X - \bar{X}$, RWX, RW) a pour résultat:

$$\hat{Y} = \underset{(2.24)}{95.34} + \underset{(0.093)}{1.085}(X - \bar{X}) + \underset{(0.044)}{0.093} \times 10^{-2}RWX - \underset{(0.023)}{0.015}RW,$$

où les erreurs types estimées ont été calculées pour un échantillon stratifié avec les poids unitaires à partir de l'expression (14). Si on utilise la méthode des moindres carrés ordinaires, on obtient

$$\hat{Y} = \underset{(3.37)}{95.34} + \underset{(0.113)}{1.085}(X - \bar{X}) + \underset{(0.086)}{0.093} \times 10^{-2}RWX - \underset{(0.034)}{0.015}RW.$$

Dans les résultats des méthodes tayloriennes, l'erreur type estimée du coefficient de RWX est à peu près la moitié de celle qu'on obtient avec la méthode des moindres carrés ordinaires. Ce fait est attribuable à la répartition des données.

La valeur absolue de la première observation de RWX est beaucoup plus grande que celle de toutes les autres observations. Cette seule observation représente 67% de la somme des carrés pour RWX. Dans l'approximation taylorienne, la variance des écarts dans l'échantillon, c'est-à-dire la variable $\hat{d}_{hij}$ dans l'expression (14), est utilisée pour estimer la variance du paramètre d'intérêt. Les écarts de la courbe de régression, $\hat{v}$, figurent dans la dernière colonne du tableau 3. La valeur de $\hat{v}$ pour la première observation est une des plus faibles dans cette colonne. Le carré moyen des résidus est de 421. Le produit (RWX)($\hat{v}$) pour la première observation est égal à -1113. Ce produit est comparable aux résultats obtenus pour les observations 3, 33 et 36. Ainsi, la première observation est à l'origine d'environ 67% de la somme des carrés pour RWX, mais seulement d'à peu près 15% de la somme des carrés pour (RWX)($\hat{v}$). Cette différence est due au fait que $\hat{v}^2$ pour la première ob-

servation représente moins d'un dixième du carré moyen des autres observations. En outre, l'écart quadratique pour la première observation présente un biais vers le bas parce que la méthode des moindres carrés oblige le plan estimé à se rapprocher d'une observation qui est séparée des autres. Par conséquent, si la variance de l'erreur sur toutes les observations est identique, la méthode taylorienne produira une estimation de la variance du coefficient de RWX qui comporte un biais vers le bas.

Cette méthode conduit-elle à une sous-estimation de la variance pour cet échantillon? On l'ignore. Si on utilise la méthode paramétrique des moindres carrés ordinaires, on accorde beaucoup d'importance à l'observation extrême dans le calcul de l'estimation combinée de la variance de l'erreur. Il est impossible de savoir si cette méthode est exacte parce que notre estimation de la variance de l'observation isolée provient d'un estimateur à un degré de liberté. Dans ce genre de circonstances, la plupart des chercheurs spécialistes préfèrent supposer que la variance du point isolé est la même que celle des autres points au lieu d'évaluer la faible variance observée du point extrême.

Dans un modèle non paramétrique, une seule observation contient peu d'information sur la variabilité à l'intérieur de la population dans laquelle cette observation a été tirée. À toutes fins utiles, une observation qui est isolée de toutes les autres est essentiellement une seule observation. Dans la forme complète d'un modèle paramétrique, l'observation isolée est membre à part entière de l'ensemble de données parce qu'elle est censée avoir été produite par le même mécanisme générateur qui est à l'origine des autres observations. Pour des données comme celles qui figurent au tableau 3, le résultat qu'on obtient par des méthodes paramétriques est très étroitement lié à certaines hypothèses au sujet de la variance des erreurs.

Dans l'estimation de la variance, une des mesures de la grandeur numérique de l'échantillon est le nombre de degrés de liberté pour l'ensemble des grappes. Ainsi, par exemple, la matrice des covariances estimées pour une variable aléatoire vectorielle de dimension k est singulière, sauf si

$$\sum_{h=1}^{L} (n_h - 1) > k.$$

Pour établir des intervalles de confiance approximatifs, il semble raisonnable d'utiliser la loi du t de Student avec au plus $\Sigma(n_h - 1)$ degrés de liberté. Étant donné que la variance d'une variance estimée est une fonction du quatrième moment de la population, les variances estimées sont extrêmement peu fiables. Le coefficient de variation des carrés est de $2^{\frac{1}{2}}$ dans le cas de la loi normale et cette valeur est beaucoup plus grande pour un grand nombre d'autres distributions souvent utilisées.

Si les variances des erreurs dans les strates sont inégales ou si des poids inégaux sont affectés aux estimations de différentes strates, la variance de l'estimateur de la variance peut être très différente de la valeur obtenue par un simple calcul du nombre de degrés de liberté des erreurs. Le tableau 4 a été construit à partir des données du tableau 1 afin d'illustrer les effets de ces facteurs sur la variance estimée. Dans la première colonne, nous supposons que la stratification n'a aucun effet; autrement dit, nous supposons que la variance dans chaque strate est égale à la variance de la population. Nous supposons que la population mère suit une loi normale, ce qui nous permet de formuler une expression explicite pour la variance de la variance. La variance de la variance estimée des erreurs dans le calcul d'une moyenne est alors proportionnelle à $(26.6)^{-1}$ si l'échantillon est stratifié, mais elle est proportionnelle à 36^{-1} si on applique les principes de l'échantillonnage aléatoire simple (ÉAS). Le nombre réel de degrés de liberté pour l'échantillon stratifié est d'un peu moins de vingt-sept parce que la taille des échantillons à l'intérieur des strates est inégale. Si on utilise les poids d'échantillonnage du tableau 1 et l'estimateur habituel de la variance pour échantillons stratifiés, la variance de la variance estimée est proportionnelle à $(4.6)^{-1}$. Ce chiffre est très bas à cause du poids élevé de la première strate. Si la variance dans la première strate est la moitié de la variance dans les autres strates, le nombre réel de degrés de liberté pour l'estimateur de la variance est de 12.4. Dans la dernière colonne du tableau 4, nous indiquons le nombre réel de degrés de liberté pour un échantillon aléatoire simple dans le cas où la variance de l'échantillon aléatoire simple est deux fois celle de l'échantillon stratifié. Ce résultat montre comment la stratification peut réduire à la fois la variance d'une moyenne estimée et la variance de la variance de cette moyenne.

Bien qu'il nous soit impossible de préciser le nombre requis de degrés de

liberté dans les erreurs pour le calcul de nos approximations, il est clair qu'on ne peut pas se limiter à un petit nombre de degrés de liberté, surtout si les poids sont inégaux.

Le corollaire 1 repose sur l'utilisation d'une approximation linéaire d'une fonction non linéaire des moyennes de l'échantillon pour reproduire le comportement de cette fonction non linéaire. Pour que cette approximation donne de bons résultats, la courbure de la fonction non linéaire doit être faible en comparaison de l'erreur type des moyennes de l'échantillon. Par exemple, si cette fonction est quadratique, c'est-à-dire qu'elle a la forme

$$g(\bar{Y}) = \alpha_1\bar{Y} + \alpha_2\bar{Y}^2,$$

l'approximation linéaire est

$$g(\bar{Y}) \doteq \alpha_1\mu + \alpha_2\mu^2 + (\alpha_1 + 2\alpha_2\mu)(\bar{Y} - \mu).$$

L'espérance mathématique de $g(\bar{Y})$ est

$$E\{g(\bar{Y})\} = \alpha_1\mu + \alpha_2[\mu^2 + V\{\bar{Y}\}].$$

Pour que l'approximation linéaire soit bonne, il faut que la valeur de $V\{\bar{Y}\}$ et (ou) celle de α_2 soient faibles.

En somme, on peut appliquer la théorie des grands échantillons avec assurance:

1. si on dispose d'un nombre raisonnable d'observations, c'est-à-dire un nombre suffisant pour qu'aucune observation ne soit trop éloignée des principales concentrations d'observations (c'est là une autre façon de dire que la distribution des erreurs tayloriennes est telle que leur moyenne suit approximativement une loi normale),
2. si le nombre réel de degrés de liberté dans les erreurs pour l'estimateur de la variance est raisonnable et
3. si la courbure de la fonction non linéaire des moyennes de l'échantillon est faible en comparaison de l'erreur type des moyennes de l'échantillon.

REMERCIEMENTS

Cette étude a été réalisée en partie grâce à une subvention accordée en vertu du Research Agreement 58-319T-1-0054X du Statistical Reporting Service du département de l'agriculture du gouvernement des États-Unis. Je remercie Nancy Hasabelnaby pour les calculs qu'elle a exécutés et Carol Francisco pour ses observations.

BIBLIOGRAPHIE

[1] Cochran, W. G. (1977). Sampling Techniques, 3e édition, Wiley, New York.

[2] Efron, B. (1979). Bootstrap method: Another look at the jackknife. Ann. Statist. 7, 1-26.

[3] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68, pp. 589-599.

[4] Fuller, W. A. (1975), Regression analysis for sample survey. Sankhya Series C 37, pp 117-132.

[5] Fuller W. A. et Hidiroglou, M. A. (1978). Regression estimation after correcting for attenuation. J. Amer. Statist. Assoc. 73, pp. 99-104.

[6] Hidiroglou, M. A., Fuller W. A. et Hickman, R. D. (1980). Super Carp, département de statistique, Iowa State University, Ames, Iowa.

[7] Isaki, C. et Fuller, W. A. (1982). Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. 77, pp. 89-96.

[8] Kish, L. et Frankel, M. R. (1974). Inference from complex samples. J. Roy. Statist. Soc. B 36, pp. 1-22.

[9] Krewski, D. et Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist. 9, pp. 1010-1019.

[10] McCarthy, P. J. (1965). Stratified sampling and distribution-free confidence intervals for the median. J. Amer. Statist. Assoc. 60, pp. 772-783.

[11] McCarthy, P. J. (1969). Pseudo-replication: Half-samples. Rev. Int. Statist. Inst. 37, pp. 239-264.

[12] Miller, R. G., Jr. (1974). The jackknife - a review. Biometrika 61, pp. 1-15.

[13] Rao, J. N. K. et Wu, C. F. J. (1984). Bootstrap with stratified samples. Technical Report No. 19 of the Laboratory for Research in Statistics and Probability. Université Carleton, Ottawa, Canada.

Tableau 1. Superficie des cultures de soja selon deux méthodes d'observation

Comtés	Aires	Poids	Hectares de soja	
			Interviews (Y)	Satellite (X)
1	1	502	8.09	24.75
1	2		106.03	98.10
1	3		103.60	112.50
2	1	212	6.47	43.20
2	2		63.82	80.10
3	1	188	43.50	61.65
3	2		71.43	92.70
3	3		42.49	74.25
4	1	190	105.26	98.10
4	2		76.49	99.45
4	3		174.34	152.10
5	1	134	95.67	57.60
5	2		76.57	66.15
5	3		93.48	91.80
6	1	189	37.84	34.65
6	2		131.12	97.65
6	3		124.44	116.10
7	1	172	144.15	136.35
7	2		103.60	99.45
7	3		88.59	99.90
7	4		115.58	123.30
8	1	114	99.15	85.50
8	2		124.56	121.50
8	3		110.88	77.40
8	4		109.14	102.60
8	5		143.66	133.65
9	1	193	91.05	75.15
9	2		132.33	85.95
9	3		143.14	112.05
9	4		104.13	81.90
9	5		118.57	80.55
10	1	93	102.59	117.90
10	2		29.46	39.15
10	3		69.28	72.00
10	4		99.15	99.45
10	5		143.66	155.25
10	6		94.49	85.50

Tableau 2. Erreurs types estimées des coefficients de régression calculés par différentes méthodes

Méthode	Erreur type estimée	
	$\hat{\beta}_0$	$\hat{\beta}_1$
Moindres carrés ordinaires	10.747	0.1116
Échantillon stratifié; poids d'échantillonnage	8.332	0.0922
Échantillon en grappes; poids d'échantillonnage	11.121	0.0823
Échantillon stratifié; poids unitaires	9.282	0.0963
Échantillon en grappes; poids unitaires	13.256	0.1071

Tableau 3: Valeurs des données de régression transformées

Grappes Stratifiées	Poids	$X - \bar{X}$	$10^{-2}RWX$	RW	$\hat{v}$
1	502	-67	-195	167	6
1	502	7	25	336	6
1	502	21	68	369	-15
2	212	-48	1	1	-37
2	212	-11	4	24	-19
3	188	-30	10	-7	-20
3	188	1	5	7	-26
3	188	-17	8	-1	-35
4	190	7	4	12	3
4	190	8	4	13	-28
4	190	61	-3	38	14
5	134	-34	28	-53	34
5	134	-25	23	-51	6
5	134	0	5	-47	-3
6	189	-57	13	-20	3
6	189	6	4	11	29
6	189	25	2	20	3
7	172	45	-9	8	1
7	172	8	3	-6	-1
7	172	8	2	-6	-16
7	172	32	-5	3	-14
8	114	-6	10	-67	8
8	114	30	-22	-66	-2
8	114	-14	18	-68	28
8	114	11	-5	-67	1
8	114	42	-32	-65	5
9	193	-16	7	4	13
9	193	-6	6	9	43
9	193	21	3	22	26
9	193	-10	6	7	19
9	193	-11	6	6	35
10	114	26	-24	-90	-21
10	114	-52	63	-84	-16
10	114	-19	26	-87	-9
10	114	8	-4	-89	-6
10	114	64	65	-93	-16
10	114	-6	12	-88	3

Tableau 4. Efficacité de l'estimateur de la variance dans différentes situations

Méthode	Nombre équivalent de degrés de liberté	
	$V_{ÉAS} = V_{st}$	$V_{ÉAS} = 2V_{st}$
Échantillonnage aléatoire simple (ÉAS)	36	9
Éch. strat., poids unitaires, var. égales	26.6	26.6
Éch. strat., poids inégaux, var. égales	4.8	4.8
Éch. strat., poids inégaux, $\sigma_1^2 = 0.5\sigma^2$	13.9	13.9

BIBLIOGRAPHIE SÉLECTIVE POUR L'ANALYSE DES DONNÉES D'ENQUÊTES COMPLEXES[1]

[1] Bellhouse, D.R. (1982). Commentaires présentés à la Session on Data Analysis from Complex Designs. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 54-55.

[2] Bickel, P.J. et Freedman, D.A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. Ann. Statist., 12, pp. 470-482.

[3] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimates from Complex Surveys. Intl. Statist. Review, 51, pp. 279-292.

[4] Binder, D.A. (1982). Non-Parametric Bayesian Models for Samples from Finite Populations. J.R. Statist. Soc. B, 44, pp. 388-393.

[5] Binder, D.A., Gratton, M., Hidiroglou, M.A., Kumar, S. et Rao, J.N.K. (1984). Analyse de données qualitatives d'enquêtes complexes: quelques expériences canadiennes. Technique d'enquête - à paraître.

[6] Brier, S.E. (1980). Analysis of Contingency Tables under Cluster Sampling. Biometrika, 67, pp. 591-596.

[7] Cohen, J.E. (1976). The Distribution of the Chi-squared Statistics under Cluster Sampling from Contingency Tables. J. Amer. Statist. Ass., 71, pp. 665-670.

[8] Choi, J.W. (1981). A Further Study on the Analysis of Categorical Data from Weighted Cluster Sample Survey. Proc. Amer. Statist. Ass., Section on Survey Methods Research, pp. 15-20.

---

[1] Préparé par l'équipe de Projet de l'analyse des données d'enquêtes complexes dont les membres sont D. Binder, M. Gratton, M. Jeays, G. Kriger, S. Kumar, D. Paton, C. Patrick et A. van Baaren.

[9] Cowan, J. et Binder, D.A. (1978). The Effect of a Two Stage Sample Design on Tests of Independence in a 2 by 2 Table. Survey Methodology, 4, pp. 16-28.

[10] Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. J. Roy. Statist. Soc. B, 31, pp. 195-223.

[11] Fay, R.E. (1979). On Adjusting the Pearson Chi-Square Statistics for Clustered Sampling. Proc. Amer. Statist. Ass., Social Statist. Section, pp. 402-406.

[12] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part I: Descriptions and Applications of the Method. Document non publié.

[13] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part II: Asymptotic Theory. Document non publié.

[14] Fay, R. (1982). Contingency Table Analysis for Complex Survey Designs: CPLX. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 44-53.

[15] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. J. Amer. Statist. Ass., 75, pp. 261-268.

[16] Fuller, W.A. (1975). Regression Analysis for Survey Data. Sankhyā C, 37, pp. 117-132.

[17] Gross, S.T. (1980). Median Estimation in Sample Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 181-184.

[18] Grizzle, J.E., Starmer, C.F. et Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, 25, pp. 489-504.

[19] Hidiroglou, M.A. (1983). Approximations de la distribution d'une somme pondérée de variables khi-carré. Statistique Canada, Ottawa, Ontario, Canada.

[20] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1980). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

[21] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1980). MINI CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

[22] Hidiroglou, M.A. et Rao, J.N.K. (1983). Tests khi-carré pour l'analyse d'observations classées de l'enquête santé Canada. Rapport technique. Statistique Canada.

[23] Holt, D. et Scott, A.J. (1981). Regression Analysis using Survey Data. The Statistician, 30, pp. 169-178.

[24] Holt, D., Scott, A.J. et Ewings, P.D. (1980). Chi-Squared Tests with Survey Data. J.R. Statist. Soc. A, 143, pp. 302-320.

[25] Holt, D., Smith, T.M.F. et Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. J.R. Statist. Soc. A, 143 pp. 474-487.

[26] Imrey, P.B., Koch, G.G. et Stokes, M.E. (1981). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Review. Intl. Statist. Rev., 49, pp. 265-283 (en collaboration avec J.N. Darroch, D.H. Freeman, Jr. et H.D. Tolley).

[27] Imrey, P.B., Koch, G.G. et Stokes, M.E. (1982). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part II: Data Analysis. Intl. Statist. Rev., 50, pp. 35-64 (en collaboration avec J.N. Darroch, D.H. Freeman, Jr. et H.D. Tolley).

[28] Imrey, P.B., Sobel, E. et Francis, M. (1980). Modeling Contingency Tables from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 212-217.

[29] Kish, L. et Frankel, M.R. (1974). Inference from Complex Sample Surveys. J. Roy. Statist. Soc. B, 36, pp. 1-37.

[30] Koch, G.G., Freeman, D.H., Jr. et Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. Intl. Statist. Review, 43, pp. 59-78.

[31] Konijn, H.S. (1962). Regression Analysis in Sample Surveys. J. Amer. Statist. Ass., 57, pp. 590-606.

[32] Landis, J.R., Lepkowski, J.M., Eklund, S.A. et Stehouwer, S.A. (1982). A Statistical Methodology for Analyzing Data from a Complex Survey. The First National Health and Nutrition Examination Survey: National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 92. Washington, D.C.

[33] Lepkowski, J.M. (1982). The Use of OSIRIS IV to Analyse Complex Sample Survey Data. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 38-43.

[34] Lepkowski, J.M., Bromberg, J. et Landis, J.R. (1981). Program for the Analysis of Multivariate Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 8-15.

[35] McCarthy, P.J. (1965). Stratified Sampling and Distribution-Free Confidence Intervals for a Median. J. Amer. Statist. Ass., 60, pp. 772-783.

[36] Nathan, G. (1969). Tests of Independence in Contingency Tables from Stratified Samples. New Developments in Survey Sampling, pp. 578-600. (rédacteurs, N.L. Johnson, et H. Smith,). Wiley: New York.

[37] Nathan, G. (1971). A Simulation Comparison of Tests for Independence in Stratified Cluster Sampling. Bull. Int. Statist. Inst., 44(2), pp. 274-280.

[38] Nathan, G. (1972). On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples. J. Amer. Statist. Ass., 67, pp. 917-920.

[39] Nathan, G. (1973). Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples. National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 53, Washington, D.C.

[40] Nathan, G. (1975). Tables of Independence in Contingency Tables from Stratified Samples. Sankhyā C, 37, pp. 77-87.

[41] Nathan, G. (1981). L'inférence statistique basée sur des plans d'échantillonnage complexes. Techniques d'enquête, 7, pp. 109-130.

[42] Nathan, G. et Holt, D. (1980). The Effect of Survey Design on Regression Analysis. J. Roy. Statist. Soc. B, 42, pp. 377-386.

[43] Pfefferman, D. et Nathan, G. (1977). Regression Analysis of Data from Complex Surveys. Bull. Intl. Statist. Inst., 41(3), pp. 21-42.

[44] Rao, J.N.K. (1975). Analytic Studies of Sample Survey Data. Survey Methodology (Supplementary Issue).

[45] Rao, J.N.K. (1983). Some Current Topics in Sample Survey Theory. Exposé presenté à Iowa State University Stat. Lab. 50 Anniversary Conference, juin, 1983.

[46] Rao, J.N.K. et Hidiroglou, M.A. (1981). Chi-Squared Tests for the Analysis of Categorical Data from the Canada Health Survey. Bull. Intl. Statist. Inst., 49(2), pp. 699-718.

[47] Rao, J.N.K. et Scott, A.J. (1979). Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 58-66.

[48] Rao, J.N.K. et Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys - Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. J. Amer. Statist. Ass., 76, pp. 221-230.

[49] Rao, J.N.K. et Scott, A.J. (1984). On Chi-Square Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Ann. Statist., 12, pp. 46-60.

[50] Särndal, C.E. (1982). Implications of Survey Design for Generalized Regression Estimation of Linear Functions. J. of Statist. Planning and Inference, 7, pp. 155-170.

[51] Schuster, J.J. et Downing, D.J. (1976). Two-Way Contingency Tables for Complex Sampling Schemes. Biometrika, 63, pp. 271-276.

[52] Scott, A.J. et Holt, D. (1982). The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. J. Amer. Statist. Ass., 77, pp. 848-854.

[53] Scott, A.J. et Rao, J.N.K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. Current Topics in Survey Sampling. (rédacteurs, D. Krewski, R. Platek et J.N.K. Rao.)

[54] Sedransk, J. et Meyer, J. (1978). Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling. J.R. Statist. Soc. B, 40, pp. 239-252.

[55] Shah, B.V. (1978). SUDAAN: Survey Data Analysis Software. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 146-151.

[56] Shah, B.V. (1981). Development of Survey Data Analysis Software. Research Triangle Institute, Research Triangle Park, North Carolina, U.S.A.

[57] Shah, B.V., Holt, M.M. et Folsom, R.E. (1977). Inference about Regression Models from Sample Survey Data. Bull. Intl. Statist. Inst., 41(3), pp. 43-57.

[58] Tepping, B.J. (1968). Variance Estimation in Complex Surveys. Proc. Amer. Statist. Ass., Social Statistics Section, pp. 11-18.

[59] Tomberlin, T.J. (1979). The Analysis of Contingency Tables of Data from Complex Samples. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 152-157.

[60] Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. J. Amer. Statist. Ass., 47, pp. 635-646.

TECHNIQUES D'ENQUÊTE

Décembre 1983 Vol. 9 No. 2

Préparé par Statistique Canada

TABLE DES MATIÈRES

Modèles de coût pour la répartition optimale d'échantillons à plusieurs degrés
WILLIAM D. KALSBEEK, OPHELIA M. MENDOZA et DAVID V. BUDESCU..... 169

Évaluation de l'application d'estimateurs composites à l'enquête sur la population active du Canada
S. KUMAR et H. LEE..... 196

L'enquête sur la consommation de carburant des automobiles
D. ROYCE..... 222

Estimation par régression de la population à l'échelon infraprovincial au Canada
RAVI B.P. VERMA, K.G. BASAVARAJAPPA et ROSEMARY K. BENDER..... 242

Une bibliographie pour l'estimation pour les petites régions..... 267

8-3200-501
No. de références
Z - 079

ISSN: 0714-0045