Recueil du symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

# MÉTHODES D'ANALYSE DES DONNÉES D'ENQUÊTES COMPLEXES SUR LA SANTÉ BASÉES SUR LE PLAN D'ÉCHANTILLONNAGE OU SUR UN MODÈLE : UNE ÉTUDE DE CAS

Risto Lehtonen[1,2] , Kari Djerf[2], Tommi Härkänen[3] et Johanna Laiho[2]

## RÉSUMÉ

Le présent article traite de l'analyse des données d'enquêtes complexes sur la santé par des méthodes multivariées de modélisation. Nous nous concentrons sur les méthodes basées sur le plan d'échantillonnage et celles basées sur un modèle qui visent à tenir compte des effets de la mise en grappes, de la stratification et de la pondération. Nous nous intéressons avant tout aux effets de la mise en grappes. Les méthodes étudiées incluent la modélisation linéaire généralisée fondée sur la pseudo-vraisemblance et les équations d'estimation généralisées, les modèles linéaires mixtes estimés par le maximum de vraisemblance restreints et les techniques hiérarchiques bayésiennes basées sur les méthodes de Monte Carlo par chaîne de Markov (MCCM). Nous comparons empiriquement ces méthodes sur des données provenant d'une enquête comprenant une entrevue sur la santé et un examen physique réalisée en Finlande en 2000 (Health-2000 Study). Les données de la Health-2000 Study ont été recueillies au moyen d'entrevues personnelles, de questionnaires et d'examens cliniques. L'enquête a été réalisée auprès d'un échantillon en grappes stratifié à deux degrés. Le plan d'échantillonnage a produit des corrélations intra-grappes positives pour nombre de variables étudiées. Nous avons sélectionné en vue d'une étude plus approfondie les variables pression sanguine systolique et morbidité chronique, tirées des volets de l'entrevue sur la santé et de l'examen clinique. Dans de nombreux cas, les diverses méthodes ont produit des résultats numériques comparables et appuyé des conclusions statistiques similaires. Celles qui ne tenaient pas compte de la complexité du plan d'échantillonnage ont parfois produit des conclusions contradictoires. Nous discutons aussi de l'application des méthodes lors de l'utilisation de logiciels statistiques standards.

MOTS CLÉS : Analyse multivariée de données d'enquête; plan d'échantillonnage complexe; équations d'estimation généralisées; modèles mixtes; techniques bayésiennes.

## 1. INTRODUCTION

La conception d'une étude comportant une interview sur la santé et un examen physique posent souvent des difficultés d'origine diverse. Le plan peut prévoir un échantillonnage à plusieurs degrés susceptible de causer une corrélation intra-grappe positive pour certaines variables étudiées. On peut utiliser des probabilités d'inclusion variables et devoir procéder à une repondération pour tenir compte de la non-réponse complète non ignorable. On peut traiter la non-réponse partielle par des méthodes d'imputation. Pour obtenir des résultats fiables, l'analyste doit s'efforcer de tenir compte de ces divers aspects complexes durant l'analyse. Il peut, pour cela, se servir des outils statistiques intégrés dans les progiciels statistiques standards. Néanmoins, le choix des méthodes et des outils appropriés pour un contexte analytique donné peut être difficile. L'objectif du présent article est de discuter des méthodes existantes, en nous appuyant sur notre expérience, dans le contexte de l'analyse multivariée de données d'enquête complexes sur la santé.

Nous considérons les méthodes basées sur le plan d'échantillonnage et celles basées sur un modèle qui visent à tenir compte des effets de la mise en grappes, de la stratification et de la pondération. Nous nous préoccupons avant tout des effets de la mise en grappes. Les méthodes étudiées incluent la modélisation linéaire généralisée basée sur l'estimation par pseudo-vraisemblance et les équations d'estimation généralisées, les modèles linéaires mixtes estimés par maximum de vraisemblance restreint (MVRE) et les techniques hiérarchiques bayésiennes basées sur les méthodes de Monte Carlo à chaîne de Markov (MCCM). Nous comparons empiriquement les résultats obtenus par

1 University of Jyväskylä
2 Statistics Finland
3 National Public Health Institute

ces méthodes sur des données provenant d'une grande enquête comprenant une interview sur la santé et un examen physique réalisée en Finlande en 2000 (Health-2000 Study). Le plan de sondage de la Health-2000 Study, et les variables sélectionnées pour la présente étude sont résumés à la section 2. À la section 3, nous décrivrons les méthodes de modélisation utilisées. Les résultats sont présentés à la section 4 et la discussion, à la section 5.

# 2. DONNÉES D'ENQUÊTE

## 2.1 Plan d'échantillonnage et collecte des données

La Health-2000 Study (Aromaa et Koskinen 2002) a été réalisée en 2000 par un consortium dirigé par l'Institut de santé publique de la Finlande. La population cible de la phase principale de l'étude englobait la population à domicile de 30 ans et plus de la Finlande. La taille de cette population était de 3,3 millions de personnes. Les données ont été recueillies auprès d'un échantillon de personnes par des entrevues personnelles au domicile des répondants, des examens physiques qui ont eu lieu dans des centres de santé locaux et des questionnaires à remplir soi-même. Étant donné les examens physiques, le plan d'échantillonnage devait inclure une mise en grappes régionales. On a utilisé un plan d'échantillonnage stratifié à deux degrés où les districts des centres de santé (englobant une ou plusieurs municipalités) étaient les unités primaires d'échantillonnage (c.-à-d. les grappes régionales). En tout, la population comptait 249 de ces grappes. On a d'abord formé 15 strates avec certitude (les 15 villes les plus grandes) qui représentent des grappes ayant une probabilité de sélection de 1. On a ensuite réparti les 234 autres grappes en cinq strates régionales couvrant l'ensemble (partie continentale) de la Finlande. Puis, on a sélectionné, en tout, 65 grappes à partir de ces strates par échantillonnage systématique PPT où les probabilités d'inclusion étaient proportionnelles à la taille de la population cible dans une grappe. Donc, les nombres totaux de strates et de grappes échantillonnées au premier degré étaient de 20 et 80, respectivement.

L'échantillon de deuxième degré (environ 8 000 personnes de 30 ans et plus) a été réparti proportionnellement à la taille des strates. On a suréchantillonné les personnes de 80 ans et plus en appliquant une probabilité de sélection égale au double de celle utilisée pour les groupes d'âge moins avancés. Enfin, on a sélectionné des individus dans chaque strate par échantillonnage systématique à partir d'une base de sondage implicitement stratifiée. Environ 88 % des personnes échantillonnées ont été interviewées, 80 % ont subi l'examen physique complet et 5 % ont subi un examen abrégé à leur domicile. Les renseignements les plus essentiels sur la santé et la capacité fonctionnelle ont été recueillis auprès de 83 % des sujets.

Nous avons utilisé un ensemble de données contenant un minimum de révisions, d'imputations et d'autres facteurs confusionnels. L'ensemble de données créé pour l'étude comprend 5 954 observations. Les paramètres de l'étude sont tels qu'elle compte, en tout, 2 495 UPE, dont 65 réparties entre les cinq strates à deux degrés et 2 430 réparties entre les 15 strates sélectionnées avec certitude (où chaque unité représente une « grappe »). Donc, dans la plupart des cas considérés, nous disposons de 2 475 degrés de liberté pour l'estimation de la variance fondée sur le plan d'échantillonnage. Pour la construction des poids, nous avons également émis l'hypothèse simplifiée selon laquelle la non-réponse complète est ignorable. Nous avons rééchelonné les poids d'analyse de sorte que la moyenne des poids soit égale à l'unité. Le coefficient de variation des poids est assez faible, compte tenu du suréchantillonnage, soit 15 %. Statistics Finland s'est chargée de l'établissement du plan d'échantillonnage, de l'échantillonnage de premier degré, de la réalisation des entrevues sur la santé et de l'élaboration des méthodes de pondération. L'institution de l'assurance sociale de la Finlande a procédé au tirage de l'échantillon de personnes de second degré. Enfin, l'Institut national de santé publique s'est chargé de la collecte des données cliniques et de la coordination globale de l'étude. Il convient de souligner que les résultats présentés ici pourraient différer de ceux figurant dans d'autres publications concernant la Health 2000 Study.

## 2.2 Effets de plan

Les variables sélectionnées pour l'étude étaient, pour le volet de l'examen physique, la pression artérielle systolique et la pression artérielle diastolique, qui sont des variables continues, et pour le volet de l'entrevue sur la santé, la maladie chronique (1 oui, 0 autrement) et l'état de santé autoévalué comme étant bon ou très bon (1 oui,

0 autrement), qui sont des variables binaires. En outre, nous avons sélectionné une variable de dénombrement du volet de l'entrevue sur la santé, à savoir le nombre de visites chez le médecin à cause d'une maladie l'année précédente. Les variables étudiées semblent présenter une corrélation intra-grappe positive (tableau 1). Dans le tableau, la première estimation de l'effet de plan tient compte de tous les aspects complexes du plan d'échantillonnage (pondération, stratification, mise en grappes) et la deuxième, de la mise en grappes et de la stratification uniquement. Les estimations les plus importantes sont celles calculées pour les moyennes des variables étudiées continues. Nous avons obtenu une estimation de l'effet de plan s'approchant de l'unité pour l'estimateur de proportion de la variable binaire indiquant l'état de santé autoévalué. La contribution de la pondération aux estimations de l'effet de plan est faible.

Nous avons sélectionné deux variables de l'étude, à savoir la pression artérielle systolique et la maladie chronique, pour les analyses finales. Nous avons choisi ces variables parce qu'elles couvraient les deux modes de collecte de données, à savoir l'entrevue sur la santé et l'examen physique. Elles sont de types différents (continue et binaire) et suggèrent l'application de modèles logistiques en plus des modèles linéaires. En outre, elles indiquent toutes deux l'existence d'un effet de grappes assez prononcé.

**Tableau 1**. Estimations de l'effet de plan des moyennes ou des proportions de certaines variables étudiées

| Variable | Moyenne ou proportion | Effet global de plan | Effet de plan dû à la mise en grappes et à la stratification |
|---|---|---|---|
| Pression artérielle diastolique | 80,87 | 2,86 | 2,79 |
| Pression artérielle systolique | 133,14 | 2,41 | 2,36 |
| Maladie chronique | 0,52 | 1,75 | 1,71 |
| Visite chez un médecin à cause d'une maladie l'année précédente | 4,50 | 1,45 | 1,42 |
| État de santé autoévalué comme étant bon ou très bon | 0,62 | 1,23 | 1,20 |

# 3. MÉTHODES MULTIVARIÉES DE MODÉLISATION

## 3.1 Variables explicatives

Nous avons ajusté des modèles comportant les principaux effets et interactions des variables explicatives disponibles. Nous nous intéressons aux variables explicatives ayant trait à la condition physique et aux caractéristiques socioéconomiques. Nous avons calculé les chiffres corrigés en fonction de l'âge et du sexe dans la plupart des cas. Notre ensemble initial de variables explicatives comprenait des variables démographiques, à savoir l'âge (en années) et le sexe (1 masculin, 2 féminin), une variable socioéconomique, à savoir le niveau de scolarité (en années mises pour achever les études) et une variable ayant trait à la condition physique, à savoir le tour de taille (en centimètres). À titre d'illustration, nous avons aussi conçu de nouvelles variables par regroupement des observations en trois groupes ayant à peu près la même taille (trois parties) selon les valeurs d'une variable explicative particulière. Nous avons appliqué ce groupement aux observations sur le niveau de scolarité et sur le tour de taille.

## 3.2 Méthodes multivariées

La modélisation linéaire généralisée par la méthode dite de la nuisance ou agrégée (Skinner et coll., 1989; Lehtonen et Pahkinen, 1996) permet d'obtenir des estimations convergentes en ce qui concerne le plan de sondage et des résultats de tests asymptotiquement valides. Cette approche comprend l'utilisation de l'estimation par la « pseudo »-vraisemblance (MPV) et les techniques d'estimation connexes, les estimateurs empiriques de l'erreur-type de type « sandwich » et le critère de Wald basé sur le plan de sondage. Nous utilisons les techniques basées sur les équations d'estimation généralisées (EEG) avec estimation multivariée par quasi-vraisemblance (MQV) comme solution plus avancée (Liang et Zeger, 1986; Diggle, Liang et Zeger, 1994). Dans le cas des EEG, nous émettons

l'hypothèse que la structure de corrélation des observations dans une grappe est interchangeable. Dans ces méthodes, les principaux efforts d'inférence portent sur les coefficients du modèle; en tant que telles, les corrélations intra-grappes ne présentent pas nécessairement d'intérêt scientifique. Comme autre solution, nous utilisons des modèles multiniveaux ou mixtes (Goldstein, 1995; McCullogh et Searle, 2001) à effets fixes ou aléatoires. Nous intégrons les effets aléatoires particuliers à la grappe dans un modèle afin de tenir compte des effets de la mise en grappes. Lorsqu'on adopte cette approche, les corrélations intra-grappes et les inférences au niveau de la grappe présentent souvent un intérêt scientifique. Pour compléter notre exercice, nous appliquons aussi les techniques hiérarchiques bayésiennes à un modèle à effets aléatoires simples (Gelman et coll., 1995). Nous comparons empiriquement les méthodes, en calculant des estimations ponctuelles, ainsi que leur erreur-type et leur effet du plan, et la valeur de t et de statistiques similaires.

## 3.3 Options analytiques

Pour comparer les méthodes choisies, nous formulons un ensemble d'options analytiques auxquelles nous ferons référence dans notre exercice d'analyse multivariée (tableau 2). Dans le cas des options 1 et 2, basées sur un modèle à effets fixes, nous utilisons l'approche des équations d'estimation généralisées (EEG). La méthode des EEG en posant que la structure des corrélations est indépendante (option 1) est en rapport avec la méthode MPV type où l'on suppose que les observations sont indépendantes dans les grappes pour l'estimation des coefficients de régression, mais peuvent être corrélées à l'intérieur des grappes dans l'estimation de la matrice des covariances des coefficients de régression estimés (au moyen d'un estimateur « sandwich », ou robuste, ou empirique de la variance; par exemple, voir Lehtonen et Pahkinen, 1996 p. 271). Dans le cas de la méthode des EEG où l'on suppose que la structure des corrélations est interchangeable (option 2), les observations peuvent être corrélées à l'intérieur des grappes pour l'estimation des coefficients de régression et de la matrice des covariances des coefficients de régression estimés, et l'on utilise l'estimation du MQV. Dans le cas de l'estimation par la méthode des EEG, nous voulons aussi déterminer si nous obtiendrons des résultats d'estimation et de test assez approchants avec un logiciel « basé sur le plan de sondage » (tel que SUDAAN) et un logiciel davantage « axé sur un modèle » (tel que SAS). Nous supposons que cette comparaison pourrait présenter un intérêt pratique pour les utilisateurs de la base de données Health-2000. Les méthodes EEG sont disponibles dans toutes les procédures de modélisation SUDAAN (référence Internet 1) et dans la procédure GENMOD de SAS (référence Internet 2). SUDAAN et SAS permettent aussi d'intégrer des poids d'élément dans l'analyse.

L'option 3 est basée sur un modèle mixte comprenant à la fois des effets fixes et aléatoires. Par souci de simplicité, nous adoptons un modèle des composantes de la variance où des coordonnées à l'origine aléatoires particulières à la grappes sont incluses en plus des effets fixes. Nous utilisons la méthode du maximum de vraisemblance restreint (MVRE pour maximum de vraisemblance restreint ou résiduel) pour l'estimation des composantes de la variance. De nouveau, nous utilisons un estimateur robuste de l'erreur-type du genre « sandwich » pour estimer les effets fixes. Dans le cas de l'option 4, nous estimons le modèle à effet aléatoire par les méthodes de Monte Carlo par chaîne de Markov (MCCM). Nous introduisons également des poids d'élément dans les exercices de modélisation sous les options 3 et 4. Ces analyses peuvent être réalisées au moyen de logiciels statistiques standards. Nous utilisons la procédure MIXED de SAS pour l'option 3 et WinBUGS (référence Internet 3) pour l'option 4 comportant l'analyse hiérarchique bayésienne. Nous postulons un modèle linéaire pour la variable dépendante continue et un modèle logistique pour la variable dépendante binaire.

**Tableau 2.** Options analytiques utilisées dans l'exercice de modélisation multivariée

| Option | Formulation du modèle et méthode d'estimation | Visant à tenir compte de la... | | |
|---|---|---|---|---|
| | | Pondération | Stratification | Mise en grappes |
| Option 0 | Option de référence<br>Modèle à effets fixes, MV, estimation de l'E.-T. basée sur un modèle | Non | Non | Non |
| Option 1 | Modèle à effets fixes, MPV, estimation robuste de l'E.-T. | Oui | Oui | Oui |
| Option 2 a) et b) | Modèle à effets fixes, MQV, estimation robuste de l'E.-T.<br>a) Application SUDAAN, b) Application SAS | Oui | a) Oui<br>b) Non | Oui |
| Option 3 | Modèle mixte, MVRE, estimation robuste de l'E.-T. | Oui | Non | Oui |
| Option 4 | Modèles bayésien à effets aléatoires, MCCC | Oui | Oui | Oui |

Abréviations :
E.-T. : Erreur-type
MV : Estimation par le maximum de vraisemblance
MPV : Estimation par le maximum de pseudo-vraisemblance
MQV : Estimation par le maximum de quasi-vraisemblance
MVRE : Estimation par le maximum de vraisemblance restreint
MCCM : Méthode de Monte Carlo par chaîne de Markov

Dans le cas de l'option de référence, qui sert de méthode de référence, nous postulons un modèle à effets fixes types et supposons que l'échantillonnage est aléatoire simple avec remise. Cette option ne tient pas compte de tous les aspects complexes de l'échantillonnage et correspond à une analyse basée sur une estimation classique, par les MCO ou le MV, des coefficients de régression et l'utilisation d'estimateurs de l'erreur-type basée sur un modèle. Par définition, la statistique d'effet du plan pour cette option est égale à un. L'analyse peut être réalisée à l'aide de tout progiciel statistique standard (comme SAS et SPSS).

Faut-il pondérer ou ne pas pondérer en cas d'enquête analytique complexe? Il ne semble pas exister de solution unique à cet important problème théorique et pratique (voir, par exemple, Pfeffermann et coll., 1998, pour une discussion). Toutefois, nous adoptons une position proche du raisonnement fondé sur le plan de sondage et intégrons les poids (rééchelonnés pour que la moyenne soit égale à l'unité) dans toutes les options, sauf celle de référence. Cette décision est également motivée par des considérations concernant la convergence des plans de sondage et la protection contre un malfonctionnement éventuel du modèle. Le prix en est une légère perte d'efficacité, mais l'effet est faible.

L'ajustement des modèles de régression par les méthodes des EEG est discuté, par exemple, dans Horton et Lipsitz (1999) qui passent en revue les logiciels pour l'estimation par les EEG et couvrent les options offertes par SAS, Stata, SUDAAN et S-Plus, et dans Ziegler et coll. (1998) qui présentent une revue très étendue de la littérature sur la méthodologie des EEG. Enfin, la modélisation multiniveaux par la procédure MIXED de SAS est discutée, par exemple, dans Singer (1998).

## 3.4 Modèles

Nous pouvons écrire l'équation d'un modèle mixte linéaire de façon succincte sous la forme

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \boldsymbol{\varepsilon} \qquad (1)$$

où $\mathbf{y}$ représente le vecteur des mesures de la variable dépendante de dimensions $n \times 1$, $\mathbf{X}$ est la matrice de plan d'échantillonnage de dimensions $n \times p$ pour la partie fixe du modèle, $\boldsymbol{\beta}$ est le vecteur des paramètres fixes de dimensions $p \times 1$ correspondant, $\mathbf{Z}$ est la matrice de plan d'échantillonnage de dimensions $n \times q$ pour la partie aléatoire du modèle, $\mathbf{v}$ est le vecteur des effets aléatoires de dimensions $q \times 1$ correspondant, et $\boldsymbol{\varepsilon}$ est le vecteur

résiduel de dimensions $n \times 1$. Nous émettons les hypothèses standards selon lequelles $\mathbf{v}$ et $\boldsymbol{\varepsilon}$ suivent la loi normale avec

$$\mathrm{E}\begin{bmatrix}\mathbf{v}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix} \text{ et } \mathrm{Var}\begin{bmatrix}\mathbf{v}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}$$

où $\mathbf{G}$ est la matrice des variances-covariances de $\mathbf{v}$ de dimensions $q \times q$ et $\mathbf{R}$ est la matrice équivalente pour $\boldsymbol{\varepsilon}$. Donc, la variance de $\mathbf{y}$ est $\mathrm{Var}(\mathbf{y}) = \mathbf{ZGZ}^T + \mathbf{R}$. Maintenant, nous modélisons $\mathrm{Var}(\mathbf{y})$ en établissant la matrice de plan d'échantillonnage à effets aléatoires $\mathbf{Z}$ et en spécifiant les structures de covariance de $\mathbf{G}$ et $\mathbf{R}$. Un modèle à effets aléatoires simples est un cas spécial de la spécification générale où $\mathbf{Z}$ contient des variables nominales, $\mathbf{G}$ contient les composantes de la variance dans une structure diagonale et $\mathbf{R} = \sigma^2\mathbf{I}_n$, où $\mathbf{I}_n$ représente la matrice d'identité de dimensions $n \times n$. Le modèle linéaire général $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ est un autre cas spécial où $\mathbf{Z} = 0$ et $\mathbf{R} = \sigma^2\mathbf{I}_n$. Pour des renseignements plus détaillés, consulter, par exemple, McCullogh et Searle (2001).

Dans le modèle (1), nous nous intéressons tout spécialement aux paramètres bêta de la partie fixe du modèle. Par conséquent, nous insérons nos variables explicatives dans la matrice de plan d'échantillonnage de la partie fixe du modèle (1). La matrice de plan d'échantillonnage $\mathbf{Z}$ de la partie aléatoire de (1) permet de modéliser les effets de grappes. Les coordonnées à l'origine aléatoires représentent les paramètres de la partie aléatoire dans notre simple exercice de modélisation. Donc, le modèle (1) se simplifie en

$$y_{hik} = \mathbf{x}_{hik}^T\boldsymbol{\beta} + v_{hi} + \varepsilon_{hik}, \tag{2}$$

où les $\mathbf{x}_{hik}$ sont les vecteurs x et $h$ désigne la strate, $i$ désigne la grappe dans la strate et $k$ désigne l'élément dans la grappe. Dans notre modèle, nous supposons que les essais aléatoires propres à la grappe $v_{hi}$ et les résidus $\varepsilon_{hik}$ sont mutuellement indépendants et suivent une loi normale de moyenne nulle et de variance $\sigma_v^2$ ou $\sigma_e^2$, respectivement. Pour la pression artérielle systolique (variable dépendante continue), nous utilisons des modèles linéaires à effets fixes et des modèles linéaires à effets mixtes. Pour la maladie chronique (variable dépendante binaire), nous utilisons des modèles logistiques binomiaux à effets fixes.

Pour l'estimation par la méthode EEG, nous considérons un modèle linéaire généralisé de la forme

$$E_m(F(\mathbf{y})) = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

où $F$ représente la fonction de lien. Nous spécifions une fonction de lien linéaire pour nos variables dépendantes continues et une fonction de lien logistique pour les variables dépendantes binaires. En EEG, nous modélisons la structure de covariance des observations dans les grappes par $\mathbf{V}_{hi} = \varphi\mathbf{A}_{hi}^{1/2}\mathbf{R}(\alpha)\mathbf{A}_{hi}^{1/2}$ où $\mathbf{A}_{hi}$ est une matrice diagonale des fonctions de variance, $\alpha$ représente la corrélation « provisoire » des paires d'observations dans une grappe, et $\varphi$ est le paramètre d'échelle. Pour une structure de corrélations indépendantes, la valeur de $\alpha$ est fixée à zéro et la matrice des corrélations « provisoires » $\mathbf{R}(\alpha)$ est réduite à une matrice d'identité. Pour une structure de corrélations interchangeables, $\alpha$ représente les éléments non diagonaux de $\mathbf{R}(\alpha)$. Pour des renseignements supplémentaires, consultez Liang et Zeger (1986) et Diggle et coll. (1994).

Nous estimons les paramètres des modèles (1) à (3) comme suit. Nous estimons les effets aléatoires de (1) par la méthode du maximum de vraisemblance restreint (MVRE) et les effets fixes par les moindres carrés généralisés (MCG), respectivement, en utilisant la procédure MIXED de SAS. On peut intégrer des poids dans la procédure d'estimation. On peut aussi obtenir l'estimation empirique (robuste) de la matrice des covariances des paramètres estimés du modèle. Nous estimons les modèles EEG par la méthode multivariée du maximum de la quasi-vraisemblance (McCullagh et Nelder 1989, chapitre 9) en utilisant la procédure GENMOD de SAS. Dans GENMOD, on peut utiliser une variable de pondération comme poids du paramètre d'échelle, de sorte que ce dernier

soit divisé par la valeur de la variable de pondération pour chaque observation (ce qui revient effectivement à travailler avec des EEG pondérées). On peut obtenir des estimations basées sur un modèle, ainsi qu'empiriques (robustes), de la matrice des covariances des paramètres estimés du modèle. Le logiciel SUDAAN utilise des équations normales pondérées de type Horvitz-Thompson pour les modèles linéaires à effets fixes et des équations pondérées de la vraisemblance pour les modèles non linéaires. Dans les modèles EEG, SUDAAN utilise un algorithme de Newton-Raphson modifié pour l'estimation des paramètres et calcule l'estimation empirique (robuste) de la matrice des covariances par la méthode implicite de linéarisation de Taylor (Binder, 1983). Toutes les procédures permettent de calculer le critère de Wald et des statistiques de test connexes (voir, par exemple, Lehtonen et Pahkinen, 1996). Dans SUDAAN, nous utilisons les procédures REGRESS et LOGISTIC.

**Modèles bayésiens**

Nous appliquons ici la méthode bayésienne complète (Gelman et coll., 1995). Nous choisissons les distributions a priori de sorte qu'elles soient vagues, car nous ne supposons l'existence d'aucun renseignement a priori. Nous spécifions un modèle à effets aléatoires dans lequel il faut tenir compte de la stratification en 20 strates, représentée par l'indice *h*, au moyen des termes d'effets aléatoires propres à ces strates $u_h$. En outre, chaque grappe *i* possède son propre terme d'effets aléatoires $v_{hi}$ qui est commun aux sujets tirés à partir de cette grappe. Nous pouvons alors exprimer un modèle par régression linéaire pour une variable dépendante continue $y_{hik}$ et des covariables $\mathbf{x}_{hik}$ sous la forme

$$y_{hik} = \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h + v_{hi} + \varepsilon_{hik} \tag{4}$$

où les $\varepsilon_{hik}$ sont les termes d'erreur individuels des sujets *k*. Nous supposons que les effets aléatoires $u_h$ et $v_{hi}$, ainsi que les termes d'erreur $\varepsilon_{hik}$ sont indépendants et suivent une loi de distribution normale de moyenne nulle et de variance $\sigma_u^2$, $\sigma_v{}^2$ ou $\sigma_\varepsilon{}^2$, respectivement. Dans les 15 strates autoreprésentatives, les grappes ne contiennent qu'un seul élément, si bien qu'on peut omettre les effets aléatoires $v_{hi}$ en restructurant les termes de vraisemblance

$$p(y_{hik} \mid \mathbf{x}_{hik}, \boldsymbol{\theta}) = \Phi(y_{hik} \mid \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h + v_{hi}, \sigma_\varepsilon^2)$$

comme suit :

$$p(y_{hik} \mid \mathbf{x}_{hik}, \boldsymbol{\theta}) = \Phi(y_{hik} \mid \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h, \sigma_v^2 + \sigma_\varepsilon^2),$$

où $\Phi$ représente la fonction de densité normale de la réponse étant donné la moyenne et la variance. L'avantage de cette méthode est que le nombre d'effets aléatoires du modèle reste faible et que l'estimation par les méthodes MCCM est possible.

Nous pouvons appliquer une construction semblable pour les variables dépendantes pondérées variant dans {0,1} en considérant

$$\mathrm{F}(\Pr[y_{hik} = 1 \mid \mathbf{x}_{hik}, \theta]) = \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h + v_{hi}, \tag{5}$$

où la fonction de lien *F* est logistique. Dans les strates autoreprésentatives, nous omettons entièrement les effets aléatoires au niveau de la grappe et nous déterminons la densité de probabilité de $y_{hik}$ au moyen du prédicteur linéaire $\mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h$.

L'imputation multiple (Tanner et Wong 1987) est utilisé pour augmenter les valeurs des données manquantes. Par conséquent, les poids ne sont pas utilisés dans les analyses. Tous les 8 028 individus ont été utilisé dans les analyses. Par contre, les variables pression sanguine systolique, tour de taille, maladie chronique et éducation avaient respectivement 1 692, 1 739, 1 047 et 1 130 valeurs manquantes. D'autres variables dans les analyses n'avaient quand à elles aucune données manquantes. Dans l'inférence bayésienne, les valeurs des données manquantes et les paramètres du modèle sont traités de la même façon. Les données manquantes des covariantes catégoriques tour de

taille et éducation, sont imputées à partir des distributions prédictives basées sur des modèles de régression logistique multinomial, qui eux, comprennent l'âge, le sexe, la langue et le district des centres de santé, comme variables explicatives. Les valeurs manquantes des réponses sont imputées en utilisant les modèles d'analyse (4) et (5) comme distributions prédictives.

Dans l'estimation, 10 000 itérations de MCCM sont exécutées en plus de 1 000 itérations de « burn-in ». Nous présentons les espérances a posteriori des paramètres du modèle, ainsi que les écarts-types, les erreurs par Monte Carlo (MC) et certains quantiles. En vue d'une comparaison aux autres options, nous calculons aussi la valeur de la statistique t.

# 4. RÉSULTATS

## 4.1 Modèles linéaires pour la pression artérielle systolique

Notre premier exercice est celui de la modélisation de la variation de la pression artérielle systolique. La variable explicative principale est le tour de taille. À titre d'illustration, nous l'utilisons sous forme de variable à trois catégories (RCIRCUM). Dans l'analyse, nous corrigeons pour les effets de l'âge et du sexe (l'âge étant considéré comme une variable explicative continue). L'analyse préliminaire indique que la pression artérielle systolique moyenne a tendance à augmenter avec l'âge et, pour un groupe d'âge particulier, à augmenter avec le tour de taille. La moyenne globale est plus élevée pour les hommes que pour les femmes. Nous examinons la structure de la variation de la valeur moyenne de la pression artérielle systolique au moyen d'un modèle linéaire d'analyse de la covariance sous les cinq options analytiques, y compris les effets principaux de toutes les variables explicatives, ainsi que leurs termes d'interaction par paires. Si nous ne tenons pas compte des aspects complexes du plan d'échantillonnage, l'analyse sous l'option de référence donne à penser qu'un modèle raisonnable inclut tous les effets principaux et les effets d'interaction par paires du tour de taille avec l'âge et le sexe. Nous obtenons aussi un modèle comparable sous les options les plus raisonnables qui tiennent compte des aspects complexes du plan d'échantillonnage. Cependant, sous ces options, la signification de l'effet d'interaction du sexe et du tour de taille est plus faible que pour l'option de référence (tableau 3).

**Tableau 3.** Test de l'effet d'interaction du sexe et du tour de taille dans la modélisation de la pression artérielle systolique sous les options analytiques

| | DDL | Valeur de F | Prob. |
|---|---|---|---|
| Option de référence | 2 | 5,51 | 0,0041 |
| Option 1 | 2 | 5,24 | 0,0054 |
| Option 2 a | 2 | 4,69 | 0,0093 |
| Option 2 b | 2 | 4,65 | 0,0096 |
| Option 3 | 2 | 4,72 | 0,0090 |

Examinons maintenant plus en détail l'interaction du sexe et du tour de taille. Les résultats sont résumés au tableau 4. De nouveau, l'option de référence produit des résultats plus libéraux que les autres options. Les options 1 à 3 donnent des estimations ponctuelles et des erreurs-types estimées qui concordent étroitement. Cependant, pour les options 2a, 2b et 3, la valeur de la statistique t est légèrement plus faible que pour l'option 1. Sous l'option 1, l'estimation est basée sur la méthode MPV, où il est tenu compte des effets de grappes dans l'estimation des erreurs-types (mais non dans l'estimation des paramètres à effets fixes). Dans les options 2a, 2b et 3, il est tenu compte des effets de grappes dans l'estimation des paramètres à effets fixes ainsi que dans l'estimation des erreurs-types, soit par la méthode des EEG avec structure de corrélation interchangeable ou par ajustement d'un modèle mixte linéaire à coordonnées à l'origine aléatoires propres à la grappe. Comparativement aux autres options, l'option 4, où l'on utilise les techniques hiérarchiques bayésiennes pour un modèle à effets aléatoires, donne des estimations ponctuelles et des estimations de l'erreur-type un peu plus faibles.

**Tableau 4.** Test de l'effet de l'interaction du sexe (masculin) et du tour de taille (catégorie moyenne) dans la modélisation de la pression artérielle systolique sous les options analytiques

| | Estimation | Erreur-type | Effet de plan | Test t (valeur studentisée) | Prob. |
|---|---|---|---|---|---|
| **Options de référence** | | | | | |
| sexe (1).rcircum(2) | 0,019 | 0,0088 | 1,00 | 2,14 | 0,0324 |
| **Option 1** (SUDAAN/REGRESS) | | | | | |
| sexe(1).rcircum(2) | 0,018 | 0,0091 | 1,06 | 2,02 | 0,0430 |
| **Option 2a** (SUDAAN/REGRESS) | | | | | |
| sexe(1).rcircum(2) | 0,017 | 0,0091 | 1,05 | 1,87 | 0,0614 |
| **Option 2b** (SAS/GENMOD) | | | | | |
| sexe(1).rcircum(2) | 0,017 | 0,0089 | 1,02 | 1,90 | 0,0573 |
| **Option 3** (SAS/MIXED) | | | | | |
| sexe(1).rcircum(2) | 0,017 | 0,0089 | 1,02 | 1,93 | 0,0532 |
| **Option 4** (WinBUGS) | | | | | |
| sexe(1).rcircum(2) | 0,0093 | 0,0083 | - | 1,12 | - |

## 4.2 Modèles logistiques pour la morbidité chronique

Nous nous tournons maintenant vers la modélisation non linéaire multivariée. La variable dépendante, c'est-à-dire la morbidité chronique, est binaire et demande une modélisation logistique. Nous utilisons l'âge et le sexe comme variables explicatives en plus du niveau de scolarité, qui est inclus à titre de variable d'intérêt de fond. De nouveau, à titre d'illustration, nous utilisons cette variable de niveau de scolarité sous forme de variable explicative à trois catégories (REDUC). Nous intégrons l'âge dans le modèle sous forme de variable explicative continue, mais le sexe et le niveau de scolarité sous forme de variables nominales. L'analyse préliminaire montre que la prévalence de la morbidité chronique a tendance à augmenter avec l'âge et, pour un groupe d'âge particulier, à diminuer lorsque le niveau de scolarité augmente. La moyenne globale est plus élevée pour les femmes que pour les hommes.

Nous sélectionnons les options analytiques 2 et 4 en plus de l'option de référence pour un examen plus approfondi. Il semble que, dans ces conditions d'analyse, l'option de référence, qui ne tient compte d'aucun aspect complexe du plan d'échantillonnage, est celle qui donne les résultats les plus prudents. Les options 2a et 2b, basées sur le plan d'échantillonnage avec EEG, donnent des résultats rapprochés, tandis que l'option 4, basée sur les techniques bayésiennes, produit des résultats un peu plus prudents que les options 2a et 2b.

**Tableau 5.** Test de l'effet du sexe dans la modélisation de la morbidité chronique sous certaines options analytique

| | Estimation | Erreur-type | Effet de plan | Test t (valeur studentisée) | Prob. |
|---|---|---|---|---|---|
| **Options de référence** | | | | | |
| Effet du sexe | 0,1042 | 0,0570 | 1,00 | 1,83 | 0,0675 |
| **Option 2a** (SUDAAN/REGRESS) | | | | | |
| Effet du sexe | 0,1125 | 0,0542 | 0,91 | 2,08 | 0,0380 |
| **Option 2b** (SAS/GENMOD) | | | | | |
| Effet du sexe | 0,1129 | 0,0532 | 0,87 | 2,12 | 0,0337 |
| **Option 4** (WinBUGS) | | | | | |
| Effet du sexe | 0,1289 | 0,0546 | - | 2,36 | - |

Toutes les options appropriées examinées pour la modélisation de la morbidité chronique appuient un même modèle, proposant au moins une certaine signification statistique pour toutes les variables explicatives, c'est-à-dire le niveau de scolarité, le sexe et l'âge. De ces options, l'option 4 semble être la plus prudente. Cependant, l'option de référence donne un léger appui à un modèle dont serait exclue la variable de sexe (tableau 5).

## 5. DISCUSSION

Les enquêtes sur la santé sont souvent des enquêtes complexes dont le plan d'échantillonnage comporte une stratification, une mise en grappes et une sélection avec probabilités inégales, et dont le plan d'estimation comporte des scénarios d'imputation et de repondération. Ces enquêtes demandent l'élaboration de stratégies d'analyse statistiquement valables et applicables en pratique. Notre but est d'étudier les méthodes disponibles et d'appliquer les outils de calcul correspondants dans un cadre analytique assez simple pour les données d'une enquête sur la santé recueillies récemment en Finlande dans le contexte de la Health-2000 Study. Il semble que, dans une mesure raisonnable, on puisse tenir compte des principaux aspects complexes du plan d'échantillonnage grâce à l'utilisation d'outils de calcul appropriés disponibles dans les progiciels statistiques utilisés couramment.

Nous nous concentrons sur les méthodes qui visent à tenir compte des aspects complexes les plus importants du plan d'échantillonnage. Nous nous intéressons tout spécialement aux effets de grappes. Par conséquent, nous créons un cadre analytique dont sont exclus les scénarios complexes de repondération et d'imputation. La variable de poids est un poids de plan d'échantillonnage simple n'ayant subi que des manipulations mineures. Nous sélectionnons deux variables étudiées différentes et un ensemble raisonnable de variables explicatives permettant de construire des modèles suffisamment souples. Les deux variables étudiées présentent l'une et l'autre une corrélation intra-grappe positive assez forte (2,41 pour la variable continue de prestation artérielle systolique et 1,75 pour la variable binaire de morbidité chronique).

Les options analytiques couvrent les méthodes utilisées couramment pour les enquêtes complexes. Les différences principales tiennent au cadre d'inférence, à la formulation des modèles, à la stratégie d'estimation et à l'application logicielle. Nous avons sélectionné des méthodes d'inférence fréquentiste ainsi que bayésiennes. Dans le premier groupe, nous appliquons des modèles à effets fixes et des modèles mixtes, linéaires ainsi que non linéaires. Nous utilisons des techniques d'estimation conçues pour être suffisamment souples pour s'adapter à des situations de modélisation de complexité variable. En outre, nous choisissons les outils de calcul de sorte que les méthodes soient disponibles dans les progiciels statistiques standards. Donc, nous utilisons des produits logiciels rentrant dans les catégories « logiciel pour l'analyse basée sur le plan d'échantillonnage », « logiciel pour l'analyse basée sur un modèle » et « logiciel pour l'analyse bayésienne ».

Nous obtenons des résultats numériques qui concordent étroitement et des conclusions inférentielles comparables sous les diverses options analytiques. Ceci donne à penser que, dans une mesure raisonnable, n'importe laquelle des méthodes examinées (sauf la méthode de référence ne tenant pas compte des aspects complexes du plan de sondage) permet de tenir compte des effets de grappes. Les méthodes fondées sur le plan d'échantillonnage utilisant l'estimation par la pseudo-vraisemblance ou les équations d'estimation généralisées couvrent un grand nombre de situations analytiques types pour la modélisation linéaire généralisée. L'utilisation de modèles mixes multiniveaux sous l'approche basée sur un modèle permettrait une modélisation plus complexe des structures de covariance.

L'analyse fondée sur les techniques bayésiennes donne des résultats comparables à ceux de l'analyse fréquentiste. L'inférence bayésienne est souple lors de l'analyse de données présentant des valeurs manquantes, car il est facile d'intégrer dans le modèle des modèles de mesure pour l'imputation multiple.

**Remerciements :** Les auteurs remercient l'équipe du projet Health 2000 de leur avoir permis d'utiliser les données pour la présente étude.

# RÉFÉRENCES


Aromaa, A. et Koskinen, S. (eds.) (2002), Health and functional capacity in Finland. Baseline results of the Health 2000 health examination survey, Helsinki, Publications of the National Public Health Institute, B3/2002. (In Finnish with English summary).

Binder, D. (1983), On the variances of asymptotically normal estimators from complex samples, *International Statistical Review*, 51, pp. 279-292.

Diggle, P.J., Liang, K.-Y. et Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford, Oxford University Press.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Goldstein, H. (1995), *Multilevel Statistical Models, 2nd Edition*, London, Arnold and New York, John Wiley & Sons.

Horton, N.J. et Lipsitz, S.R. (1999), Review of software to fit generalized estimating equation regression models, *The American Statistician*, 53, pp. 160-169.

Lehtonen, R. et Pahkinen, E. (1996), *Practical Methods for Design and Analysis of Complex Surveys, Revised Edition*, Chichester, John Wiley & Sons.

Liang, K.-Y. et Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika* 73, pp. 13-22.

McCullagh, P. et Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London, Chapman and Hall.

McCullogh, C.E. et Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, New York, John Wiley & Sons.

Pfeffermann, D., Skinner, C.J., Goldstein, H., Holmes, D.J. et Rasbash, J. (1998), Weighting for unequal selection probabilities in multilevel models (With discussion), *Journal of the Royal Statistical Society*, Series B, 60, pp. 23-40.

Singer, J.D. (1998), Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models, *Journal of Educational and Behavioral Statistics*, 24, pp. 323-355.

Skinner, C., Holt, T. et Smith, T.M.F. (eds.) (1989), *Analysis of Complex Surveys*, Chichester, John Wiley & Sons.

Tanner, M. A. et Wong W. H. (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Society*, 82, pp. 528-550.

Ziegler, A., Kastner, C. et Blettner, M. (1998), The generalized estimating equations: an annotated bibliography. *Biometrical Journal*, 40, pp. 115-139.


**Références Internet**


Référence Internet 1: The SAS web site, http://www.sas.com
Référence Internet 2: The SUDAAN web site, http://www.rti.org/sudaan
Référence Internet 3: The WinBUGS web site, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml