Mai 2007 **Numéro 40**

Cahiers
de l'IMA

Les données d'enquêtes : Spécificités et méthodes d'analyse

André Berchtold

Groupe de Recherche sur la Santé des Adolescents
& Institut de Mathématiques Appliquées

Institut de Mathématiques Appliquées
Faculté des S.S.P.
Université de Lausanne
Anthropole
CH-1015 Lausanne

Les données d'enquêtes : Spécificités et méthodes d'analyse

André Berchtold[1]

Résumé

Cet article décrit les particularités des données issues d'enquêtes et donne des lignes directrices pour les analyser correctement au moyen d'outils statistiques. Il est destiné à tous les chercheurs travaillant à partir de données d'enquêtes, mais qui ne sont pas des spécialistes de l'analyse statistique. Nous étudions tout d'abord l'influence du plan de sondage (pondérations, strates, grappes) sur les résultats des analyses quantitatives, puis nous comparons au moyen d'un exemple numérique deux approches statistiques différentes : l'approche classique dans laquelle le plan de sondage est un élément exogène au modèle statistique employé et l'approche multi-niveaux dans laquelle le plan de sondage est un élément constitutif du modèle.

Mots clés : Enquête, plan de sondage, pondération, strate, grappe, variance, effet plan, approche classique, approche multi-niveaux.

[1]Institut Universitaire de Médecine Sociale et Préventive, Groupe de Recherche sur la Santé des Adolescents, Centre Hospitalier Universitaire Vaudois, Suisse,
& Institut de Mathématiques Appliquées, Faculté des Sciences Sociales et Politiques, Université de Lausanne, Suisse.
Email : Andre.Berchtold@unil.ch, Web : http ://www.andreberchtold.com

Table des matières

1 Introduction

Les données collectées au moyen d'enquêtes sont généralement obtenues à l'aide d'un plan de sondage complexe pouvant inclure différentes strates, des effets de grappes et des pondérations (Kish, 1965, Cochran, 1977, Tillé, 2001, Groves et al., 2004). Durant l'analyse statistique, les éléments du plan de sondage doivent être pris en compte afin d'obtenir des estimations exactes des caractéristiques de la population étudiée. Une utilisation incorrecte du plan de sondage n'influencera pas forcément la valeur des coefficients estimés, mais elle conduira dans la majorité des cas à des erreurs d'estimation des variances, ce qui implique que les intervalles de confiance et les résultats des tests d'hypothèses seront faux (Crockett, 2004).

Quatre éléments influencent la magnitude de l'erreur due à l'échantillonnage (Groves et al., 2004) :

1. La possibilité d'inclure ou non avec une probabilité connue non-nulle chaque membre de la population dans l'échantillon (pondérations).
2. Le fait que la composition de l'échantillon soit ou non contrôlée par rapport à certains facteurs clés (stratification).
3. Le fait de tirer les observations à inclure dans l'échantillon indépendamment les unes des autres ou au contraire par groupes (effet de grappe).
4. La taille de l'échantillon.

Si le dernier point est tout-à-fait classique avec des conséquences bien connues en statistique, les trois premiers se révèlent souvent critiques.

La prise en compte du plan de sondage d'une enquête peut s'effectuer de deux façons différentes (Goldstein, 2003). L'approche classique consiste à estimer les variances à l'aide de formules tenant compte du plan de sondage de l'enquête et notamment des strates et des phénomènes de corrélation inter-observations dus aux grappes. La seconde approche consiste à intégrer directement au sein des modèles statistiques le plan de sondage. Ce dernier consistant généralement en une structure hiérarchique, comme par exemple des élèves appartenant à des classes qui appartiennent elles-mêmes à des écoles, nous parlons alors de modèles multi-niveaux (Hox, 2002).

Les deux approches reflètent des problématiques fondamentalement différentes. Dans l'approche classique (dite "design based"), nous disposons d'un échantillon provenant d'une population finie dont chaque individu avait la possibilité d'être inclu aléatoirement dans l'échantillon. Les variables ne sont pas aléatoires et nous cherchons à estimer la vraie valeur de paramètres de la population, comme par exemple le score moyen des élèves à un test de mathématiques. Le plan de sondage étant entièrement exogène au modèle, il faut en tenir compte au moment de l'estimation. Dans l'approche multi-niveaux (dite "model-based") en revanche, nous cherchons à modéliser la distribution des variables au sein de la population et le modèle dépend des hypothèses faites sur cette distribution. Le score d'un élève est vu comme une somme d'effets en provenance des différents niveaux de la hiérarchie. Il y a donc une prise en compte directe du fait de se trouver dans une classe particulière et une école particulière.

Dans leur version de base, de nombreux logiciels statistiques ne permettent pas de tenir compte de façon correcte de plans de sondage complexes et ne sont donc pas adaptés à l'analyse des données d'enquêtes. Différentes comparaisons numériques des logiciels à disposition pour l'analyse des données d'enquêtes ont été publiées par le passé (Cohen, 1997, Brogan, 1998, 2005). Dans un cadre plus général, il faut aussi citer le remarquable travail de Mitchell (2005) qui compare de façon exhaustive SPSS, SAS et Stata. Notre but n'est pas ici d'effectuer une étude supplémentaire de ce type, mais nous citerons tout-de-même différents programmes tout au long de l'article. SPSS, S-Plus, R, SAS et Stata sont des logiciels statistiques généralistes. WesVar et SUDAAN s'utilisent pour l'analyse classique des données d'enquêtes, alors que GLLAMM, LISREL, MLwiN et HLM sont adaptés à l'analyse multi-niveaux. Cette liste n'est pas exhaustive et de nouveaux logiciels sont développés régulièrement.

L'article est organisé de la façon suivante : La section 2 aborde d'un point de vue théorique les problèmes statistiques liés à l'analyse des données d'enquête. La section 3 présente une comparaison détaillée des approches classique et multi-niveaux sur la base d'un exemple numérique, puis une conclusion résume notre propos.

2 Aspects théoriques de l'analyse statistique des données d'enquête

2.1 Pondérations

Dans un plan de sondage un tant soit peu complexe, les individus de la population étudiées se voient généralement assigner chacun une probabilité d'inclusion dans l'échantillon, cette probabilité pouvant différer d'un individu à l'autre. Ces probabilités d'inclusion sont ensuite prises en compte dans la base de données de par l'assignation de pondérations différentes à chaque individu de l'échantillon. Dans les logiciels statistiques, quatre types de pondérations peuvent être proposées pour modifier l'importance respective de chaque observation au sein d'une base de données :

1. **Pondérations d'échantillonnage :** Ce sont les pondérations utilisées dans les échantillons issus d'enquêtes. Elles représentent la probabilité d'inclusion d'un sujet de la population étudiée dans l'échantillon et elles sont proportionnelles à l'inverse du taux d'échantillonnage. Par exemple, si 2 personnes sont échantillonnées dans une population de taille 10, le taux d'échantillonnage vaut $2/10 = 0.2$ et les pondérations correspondantes valent $1/0.2 = 5$. En fonction de l'utilisation prévue, leur total peut être normalisé pour être égal à la taille de l'échantillon plutôt qu'à celle de la population et elles prennent alors souvent des valeurs non-entières.
2. **Pondérations de fréquence :** Elles indiquent combien de données identiques chaque observation d'un fichier représente réellement et ne prennent que des valeurs entières.
3. **Pondérations analytiques :** Elles sont utilisées lorsque chaque observation d'un fichier est constituée d'une moyenne de plusieurs autres observations. Si ces moyennes sont calculées à partir de nombres différents de valeurs, leur précision varie et les pondérations analytiques sont utilisées pour obtenir des résultats corrects. Elles sont proportionnelles à l'inverse de la variance et prennent souvent des valeurs non-entières.
4. **Pondérations d'importance :** Ces pondérations indiquent simplement l'importance relative accordée à chaque observation. Il n'y a pas de façon standard de les calculer, ni même de définition stricte. Elles peuvent être entières ou non.

Les pondérations ne sont pas utilisées pour augmenter artificiellement la taille d'un échantillon, mais seulement pour attribuer une importance différente à chaque observation. Cela implique que

le nombre de données pondérées doit toujours être exactement égal au nombre de données non-pondérées. Par ailleurs, il n'est pas possible de convertir un type de pondérations en un autre. Des pondérations d'échantillonnage par exemple ne peuvent pas être traitées comme des pondérations de fréquence et vice versa.

Au niveau du traitement statistique, le principal problème relatif aux pondérations tient au fait que la plupart des logiciels généralistes ne gèrent pas tous les types de pondérations. En particulier, peu de logiciels savent utiliser les pondérations d'échantillonnage. Parmi les principaux logiciels statistiques généralistes du marché, Stata est le seul à pouvoir utiliser les quatre types de pondérations. Par ailleurs, même si un logiciel peut utiliser un certain type de pondérations, il n'est pas du tout certain que ce type puisse être utilisé dans toutes les procédures. Dans certains cas, des modules additionnels permettent d'ajouter certaines fonctionalités d'analyse d'enquêtes à un logiciel, comme par exemple la toolbox "Complex samples" pour SPSS.

Comme alternative aux logiciels statistiques standards, il existe des logiciels développés spécifiquement pour l'analyse classique des données d'enquêtes, notamment SUDAAN et WesVar. Bien entendu, ces logiciels traitent correctement les pondérations d'échantillonnage. En revanche, ils sont plus limités quant aux types d'analyses proposées et à la gestion des fichiers de données. En ce qui concerne les logiciels d'analyse multi-niveaux, tant LISREL que GLLAMM, HLM et MLwiN autorisent l'utilisation des pondérations d'échantillonnage dans la majorité des modèles.

Au vu de la difficulté à analyser des données pondérées, nous pouvons être tentés de faire abstraction des pondérations lors des analyses. Cette pratique est en réalité totalement déconseillée. Brogan (1998) identifie au moins quatre problèmes liés à cette façon de faire :

1. Les estimations ponctuelles des paramètres de la population risquent d'être biaisées.
2. Les variances des estimations ponctuelles seront sous-estimées.
3. Les intervalles de confiance calculés autour des estimations ponctuelles des paramètres de la population seront trop étroits.
4. Les tests d'hypothèses seront trop enclins à rejeter l'hypothèse nulle.

Les deux derniers points découlent directement du deuxième.

Une variante consiste à traiter les pondérations d'échantillonnage comme un autre type de pondérations qui est lui proposé par le logiciel utilisé, des pondérations de fréquence par exemple. Dans certains cas, cela peut permettre d'obtenir des estimations ponctuelles non-biaisées pour les paramètres de la population, ce qui répond au point 1 ci-dessus. En revanche, les variances continueront généralement à être sous-estimées, ce qui impliquera toujours des erreurs au niveau des intervalles de confiance et des tests d'hypothèses.

Un autre problème peut aussi survenir si nous traitons les pondérations d'échantillonnage comme des pondérations de fréquences. Ces dernières étant par définition entières, des logiciels comme SPSS commencent parfois par arrondir les pondérations à l'entier le plus proche avant d'effectuer les calculs. Cela résulte en deux effets néfastes : les estimations ponctuelles se retrouvent faussées et l'effectif total pondéré n'est plus égal à l'effectif total non-pondéré. Il est aussi parfois possible de simuler l'utilisation de pondérations (Ellis & Hesterberg, 1999), mais il s'agit d'une approche compliquée ne garantissant pas toujours l'obtention de résultats. Il est donc préférable d'opter pour un logiciel directement adapté aux analyses envisagées.

Lorsque les données sont analysées à l'aide d'un modèle multi-niveaux, le problème des pondérations reste entier et il est aussi nécessaire d'en tenir compte lors des calculs. Cependant, au contraire de l'analyse classique, il n'existe pas encore à l'heure actuelle de méthode standard pour tenir compte de ces pondérations. Les deux principales approches sont dues à Pfeffermann et al. (1998) et Asparouhov (2006). Les principaux logiciels du marché reposent sur l'une ou l'autre de ces deux méthodes qui donnent des résultats légèrement différents, sans qu'il ne soit possible de dire laquelle est la meilleure (Chantala et al., 2005, 2006).

2.2 Stratification et effets de grappes

Afin de tenir compte de la complexité de la réalité et s'assurer que toutes les composantes utiles de la population seront bien représentées au sein de l'échantillon, les enquêtes utilisent généralement des plans de sondage complexes à plusieurs niveaux dans lesquels la population est tout d'abord répartie en différentes strates avant que l'échantillon proprement dit ne soit sélectionné. La stratification est généralement bénéfique, car elle permet de réduire la variance des estimations.

Par ailleurs, certains groupes particuliers de sujets sont parfois systématiquement inclus dans leur intégralité dans l'échantillon, ce qui induit des effets de grappes avec des observations partageant certaines caractéristiques communes et étant donc corrélées entre-elles. Cela est le cas notamment lors d'enquêtes menées au sein des écoles, dans lesquelles tous les élèves de certaines classes sont inclus dans l'échantillon.

Afin que les estimations soient correctes, les analyses statistiques doivent tenir compte de la stratification et des grappes. Cependant, tout comme dans le cas des pondérations, les logiciels généralistes permettent rarement d'inclure de tels éléments, la seule exception étant Stata. En revanche, les logiciels SUDAAN et WesVar spécialisés dans l'analyse classique des données d'enquêtes utilisent correctement le plan de sondage. En ce qui concerne l'approche multi-niveaux, son principe même fait qu'il est possible de tenir compte du plan de sondage.

2.3 Estimation des variances

Dans l'approche classique, plus le plan de sondage est complexe, plus les variances des statistiques calculées deviennent difficiles à estimer. Les logiciels reposent alors sur deux principes différents : la linéarisation et le rééchantillonnage (Groves et al., 2004, Brogan, 2005). La linéarisation consiste à transformer, au moyen d'une approximation par une série de Taylor, une variance difficile à estimer car généralement constituée de différents termes étant eux-mêmes des estimations issues de l'enquête, en une somme d'éléments. Il s'agit donc d'obtenir une formule analytique approximant le plus possible la vraie variance.

Les méthodes de rééchantillonnage, soit équilibré, soit jacknife, consistent à extraire répétitivement un grand nombre de sous-échantillons à partir de l'échantillon total de l'enquête, puis à calculer la statistique d'intérêt sur chacun d'eux et finalement à en déduire la variance. La différence entre l'approche équilibrée et l'approche jacknife tient à la façon dont sont construits les sous-échantillons. Dans la méthode équilibrée, chaque échantillon est constitué d'une moitié des données disponibles. Dans l'approche jacknife, chaque échantillon est constitué de toutes les données originales à l'exception d'une seule.

Les logiciels statistiques adaptés à l'analyse classique des données d'enquêtes proposent géné-

ralement les deux méthodes d'estimation de la variance. Celles-ci produisant des estimations très proches, il n'y a pas de raison particulière de choisir l'une plutôt que l'autre.

2.4 Effets du plan de sondage sur les estimations

Les effets des différentes composantes du plan de sondage sur les estimations des modèles statistiques peuvent être quantifiés à l'aide de l'effet plan (*design effect*) (Kish, 1965). Ce coefficient est défini comme

$$deff = 1 + (n - 1) \cdot \rho \tag{1}$$

où n est la taille moyenne des groupes d'observations pour lesquels on effectue le calcul et où ρ est la corrélation intra-classe. Un plan de sondage à plusieurs niveaux, avec éventuellement des effets de grappes, tendant à fournir un échantillon de données moins différentes les unes des autres que ce qui aurait été obtenu par un échantillonnage aléatoire simple à un niveau, l'échantillon obtenu au moyen du plan de sondage complexe contient moins d'information que sa taille nominale ne le laisse supposer. L'effet plan s'interprète alors comme le coefficient mettant en relation la taille nominale de l'échantillon et la taille correspondante pour un échantillon obtenu par tirages simples. Supposons par exemple que nous ayons interrogé 50 classes comportant en moyenne 20 élèves pour un total de 1000 élèves. Si l'effet plan vaut 3, cela signifie que nos 1000 observations ne comportent pas plus d'information que si nous avions interrogé 1000/3= 333 élèves parfaitement au hasard.

La corrélation intra-classe mesure la similarité des observations au sein d'un même groupe. Plus cette corrélation est importante, plus l'hypothèse d'indépendance entre les observations est remise en cause et donc plus les méthodes statistiques traditionnelles basées sur cette hypothèse risquent de donner des résultats erronés. Une valeur de 0.1 signifie par exemple que deux observations d'un même groupe ont environ 10% de chance de plus de prendre la même valeur que si elles avaient été sélectionnées parfaitement au hasard dans l'ensemble de la population. Selon les cas, la corrélation intra-classe peut être calculée pour un paramètre précis du modèle statistique employé ou pour l'ensemble de ce dernier. Il faut noter que la mesure *deff* utilise non-seulement la corrélation intra-classe, mais aussi la taille moyenne des groupes dans son calcul. Ainsi, même si la corrélation

intra-classe est très petite, l'effet plan peut être élevé lorsque la taille des groupes est grande. Par exemple, pour une corrélation intra-classe de 0.02, la mesure *deff* vaudra 1.18 pour des groupes de taille 10, mais 2.98 pour des groupes de taille 100.

L'effet plan s'interprète aussi comme le rapport attendu entre la variance d'un estimateur obtenu à partir d'un plan de sondage complexe et celle du même estimateur obtenu à partir d'un plan de sondage simple. On utilise cependant alors de préférence la mesure *deft* définie comme la racine carrée de *deff* et qui donne le rapport attendu entre les écarts-types. Si *deft* vaut 1.73, alors un écart-type de 3.46 calculé à partir de données issues d'un plan de sondage complexe correspond à un écart-type de 3.46 / 1.73 = 2 pour un échantillon de même taille, mais issu de tirages aléatoires simples.

Les mesures *deff* et *deft* sont des indicateurs permettant de déterminer l'influence du plan de sondage sur les résultats. Ils prennent généralement des valeurs supérieures ou égales à 1, la valeur 1 signifiant l'absence total d'influence du plan de sondage. Il n'existe pas de test statistique permettant de déterminer à partir de quelle valeur le plan de sondage a vraiment une influence significative sur les résultats et doit donc obligatoirement être pris en compte dans le calcul des modèles statistiques. Dans une approche conservatrice, il faudrait donc toujours prendre en compte la totalité du plan de sondage dans les calculs. Certains auteurs se basant sur les simulations de Monte-Carlo présentées par Muthen & Satorra (1995) pensent que les effets plan inférieurs à 2 n'ont pas besoin d'être pris en compte, car l'implication sur les tests statistiques ne produit que peu de changement pour un risque de 5%. Pour une discussion plus approfondie de ces questions, le lecteur peut se référer à Kalton (1983) et Lê, Brick & Kalton (2002).

3 Comparaison des approches classiques et multi-niveaux

Nous utilisons ici des données provenant de l'enquête SMASH 2002 (Narring et al., 2004), une enquête multi-centres menée en Suisse en 2002 concernant la santé des adolescents. L'échantillon comprend 7126 adolescents âgés de 16 à 20 ans répartis en 579 classes. Il a été construit de façon à donner une image la plus fidèle possible de l'ensemble des adolescents vivant en Suisse encore

en formation (école ou apprentissage), à l'exclusion des jeunes déjà sortis de la filière éducative. L'enquête a été menée au sein de classes entières d'élèves qui partagent certainement des caractéristiques communes, ce dont il faut tenir compte. Une pondération d'échantillonnage est associée à chacune des classes. Par ailleurs, l'enquête a été menée au niveau national, ce qui implique qu'elle concerne des adolescents provenant de trois régions (Suisse alémanique, Suisse romande, Tessin) ayant une langue et une culture différentes. La prise en compte correcte du plan de sondage doit donc faire intervenir la pondération d'échantillonnage de chaque classe, le fait d'appartenir à une classe particulière (effet de grappe) et le fait d'appartenir à une région particulière (stratification).

Notre objectif est d'estimer le pourcentage de jeunes vivant en Suisse et âgés de 16 à 20 ans consommant du cannabis. Nous disposons d'une variable dépendante dichotomique prenant la valeur 0 pour les jeunes déclarant ne jamais avoir consommé de cannabis durant leur vie et 1 sinon. Les facteurs explicatifs retenus sont l'âge et un indicateur de la probabilité d'avoir déjà été saoul. L'âge concerne individuellement chaque adolescent et prend des valeurs entières comprises entre 16 et 20 ans. Le second facteur nommé *ClSaoul* est calculé comme la probabilité moyenne qu'un sujet d'une classe donnée ait été saoul au moins une fois dans sa vie. Il prend la même valeur pour tous les adolescents d'une classe et représente donc une notion de risque de groupe.

Deux approches différentes sont proposées pour expliquer la consommation de cannabis en tenant compte du plan de sondage. Dans un premier temps nous utilisons l'approche classique, puis nous construisons un modèle multi-niveaux. Les calculs ont été effectués dans Stata 9.0 en utilisant la fonction intégrée *svy :logistic* pour l'approche classique et la fonction additionnelle *GLLAMM 6.0* pour l'approche multi-niveaux.

3.1 Approche classique

Dans l'approche classique, le plan de sondage est spécifié a priori, puis un modèle est estimé en tenant compte de ce plan de façon à ce que les variances ne soient pas sous-estimées, ce qui impliquerait des erreurs au niveau des tests d'hypothèses et des intervalles de confiance. Le modèle statistique lui-même est une régression logistique pour la variable Cannabis avec deux facteurs

explicatifs, *Age* et *ClSaoul* :

$$\pi_i = \beta_0 + \beta_1 Age_i + \beta_2 ClSaoul_i + e_i \tag{2}$$

où i est un indice représentant chaque adolescent de l'échantillon, π est le logit pour la catégorie de référence "Avoir consommé du cannabis", *Age* est l'âge en année et *ClSaoul* la probabilité d'avoir été saoul au moins une fois dans sa vie calculée au niveau de l'ensemble de la classe à laquelle appartient l'adolescent i.

Nous avons estimé quatre fois le modèle, tout d'abord en faisant totalement abstraction du plan de sondage, puis en ajoutant successivement les pondérations d'échantillonnage, les grappes et les strates. A chaque fois, les trois paramètres du modèle (la constante et les coefficients des deux variables explicatives) sont fortement significatifs ($p<0.005$). Au niveau des estimations ponctuelles des paramètres, le premier modèle ne tenant pas compte des pondérations d'échantillonnage donne des résultats différents des trois autres modèles : -2.97 contre -3.44 pour la constante, 0.11 contre 0.13 pour la variable *Age* et 2.03 contre 2.19 pour *ClSaoul*. Cette différence met en évidence la nécessité de prendre en compte les pondérations lors de l'estimation du modèle.

Le Tableau 1 donne les mesures *deff* et *deft* pour chacun des trois éléments explicatifs du modèle. Son analyse montre la nécessité de tenir compte du plan de sondage. La mesure *deft* indique que les écarts-types seraient sous-évalués d'un facteur allant de 1.76 à 2.02 si les pondérations n'étaient pas prises en compte. Par ailleurs, les mesures données pour les deux derniers modèles sont quasiment identiques. Cela implique que la prise en compte de l'effet dû aux régions (les strates) en plus de ceux dus aux classes (les grappes) et aux pondérations ne change pratiquement rien. Nous pouvons donc éliminer le niveau des régions pour nous ramener à une structure à deux niveaux, les élèves et les classes. Toutefois, ainsi que noté dans la section 2.4, on préfère parfois adopter une approche conservatrice et utiliser quand même l'intégralité du plan de sondage.

Les valeurs de la mesure *deft* données dans le Tableau 1 indiquent que les écarts-types obtenus sans tenir compte intégralement du plan de sondage sont systématiquement inférieurs à ceux obtenus avec le plan de sondage, ce qui se traduit par des intervalles de confiance trop étroits et des

Tab. 1 Mesures de l'effet plan (design effect) dans l'approche classique.

	Pondérations seulement		Pondérations et classes		Pondérations, classes et régions	
Variable	*deff*	*deft*	*deff*	*deft*	*deff*	*deft*
Age	3.09	1.76	3.84	1.96	3.84	1.96
ClSaoul	4.08	2.02	2.83	1.68	2.84	1.68
Constante	3.31	1.82	4.05	2.01	4.05	2.01

p-valeurs trop petites. Dans certaines situations, cela peut impliquer qu'une variable apparemment significative se révèle en fait être non-significative lorsque l'on prend en compte correctement le plan de sondage. Nous pouvons encore noter que pour la variable *ClSaoul*, les mesures *deff* et *deft* diminuent lorsque les classes son prises en compte en plus des seules pondérations. Cela s'explique par le fait que cette variable est justement calculée au niveau des classes et que le fait de découper l'ensemble de la population étudiée en différents groupes avant d'effectuer l'échantillonnage a généralement un effet bénéfique sur le calcul des variances.

Sans tenir compte des strates (voir annexe A.1), l'estimation du modèle (2) donne

$$\pi_i = -3.44 + 0.13 Age_i + 2.19 ClSaoul_i + e_i$$

En prenant comme catégorie de référence un adolescent de 16 ans appartenant à une classe dont aucun des membres n'a jamais été saoul, le logit vaut

$$\pi = -3.44 + 0.13 \cdot 16 + 2.19 \cdot 0 = -1.36$$

et la probabilité correspondante d'avoir consommé du cannabis vaut

$$p = \frac{1}{1 + \exp^{-\pi}} = \frac{1}{1 + \exp^{1.36}} = 20.42\%$$

Une augmentation de l'âge d'une année équivaut à une augmentation du logit de 0.13 et une augmentation de 10% du nombre d'adolescents de la classe ayant déjà été saouls fait augmenter le

logit de 0.219. Au maximum, pour un adolescent de 20 ans appartenant à une classe dont tous les membres ont déjà été saouls, le logit vaut

$$\pi = -3.44 + 0.13 \cdot 20 + 2.19 \cdot 1 = 1.35$$

ce qui équivaut à une probabilité d'avoir déjà consommé du cannabis égale à 79.41%.

3.2 Approche multi-niveaux

Dans une approche multi-niveaux, les éléments hiérarchiques (classes et régions) sont directement intégrés au sein du modèle. Nous commençons par écrire une équation utilisant uniquement les variables du niveau le plus désagrégé, le niveau des élèves (niveau 1) :

$$\pi_{ijk} = \beta_{0jk} + \beta_{1jk} Age_{ijk} + e_{ijk} \tag{3}$$

où i est l'indice de l'élève, j celui de sa classe et k celui de sa région. Cette équation a des paramètres β_{0jk} et β_{1jk} différents pour chaque combinaison d'une classe et d'une région.

Au niveau des classes (niveau 2), nous construisons un modèle utilisant la variable explicative de ce niveau pour expliquer la variation de ces paramètres :

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k} ClSaoul_{jk} + u_{0jk} \tag{4}$$

Ce modèle postule que dans chaque région (indice k) le paramètre β_{0jk} de la classe j est expliqué de façon linéaire par la variable $ClSaoul$. De la même façon, les paramètres β_{1jk} sont expliqués au niveau 2 par le modèle suivant :

$$\beta_{1jk} = \gamma_{10k} + \gamma_{11k} ClSaoul_{jk} + u_{1jk} \tag{5}$$

Selon le même principe, les paramètres γ_{00k}, γ_{01k}, γ_{10k} et γ_{11k} sont eux-mêmes expliqués au

niveau des régions (niveau 3) par les modèles suivants :

$$\gamma_{00k} = \gamma_{000} + u_{00k} \tag{6}$$

$$\gamma_{01k} = \gamma_{010} + u_{01k} \tag{7}$$

$$\gamma_{10k} = \gamma_{100} + u_{10k} \tag{8}$$

$$\gamma_{11k} = \gamma_{110} + u_{11k} \tag{9}$$

Comme nous ne considérons aucune variable explicative au niveau 3, les modèles (6) à (9) ne sont constitués que d'une constante affectée par un terme d'erreur représentant la variabilité inter-région. Bien entendu, il est possible d'utiliser un nombre quelconque de facteurs explicatifs à chacun des niveaux du modèle.

Les termes d'erreur e_{ijk}, u_{0jk}, u_{1jk}, u_{00k}, u_{01k}, u_{10k} et u_{11k} sont tous distribués aléatoirement avec une moyenne nulle. Les termes d'erreur d'un même niveau suivent une distribution normale multivariée. Ils sont généralement hétéroscédastiques et corrélés entre-eux. D'un niveau à l'autre, les termes d'erreur sont en revanche supposés être non-corrélés.

En introduisant les équations (6) à (9) du niveau des régions (niveau 3) dans celles du niveau des classes (niveau 2), nous obtenons :

$$\beta_{0jk} = \gamma_{000} + u_{00k} + (\gamma_{010} + u_{01k})\, ClSaoul_{jk} + u_{0jk} \tag{10}$$

$$\beta_{1jk} = \gamma_{100} + u_{10k} + (\gamma_{110} + u_{11k})\, ClSaoul_{jk} + u_{1jk} \tag{11}$$

Ensuite, en introduisant les équation (10) et (11) dans l'équation (3) du niveau 1, nous obtenons le modèle final suivant :

$$\begin{aligned} \pi_{ijk} \quad = \quad & \gamma_{000} + \gamma_{100} Age_{ijk} + \gamma_{010} ClSaoul_{jk} + \gamma_{110} Age_{ijk} ClSaoul_{jk} \\ & + e_{ijk} + u_{0jk} + u_{1jk} Age_{ijk} + u_{00k} + u_{10k} Age_{ijk} + u_{01k} ClSaoul_{jk} + u_{11k} Age_{ijk} ClSaoul_{jk} \end{aligned} \tag{12}$$

Le terme $\gamma_{000} + \gamma_{100} Age_{ijk} + \gamma_{010} ClSaoul_{jk} + \gamma_{110} Age_{ijk} ClSaoul_{jk}$ est la partie fixe ou déterministe du modèle. La seconde partie du modèle regroupe tous les termes aléatoires avec e_{ijk} pour le niveau 1, $u_{0jk}+u_{1jk}Age_{ijk}$ pour le niveau 2 et $u_{00k}+u_{10k}Age_{ijk}+u_{01k}ClSaoul_{jk}+u_{11k}Age_{ijk}ClSaoul_{jk}$ pour le niveau 3.

Deux différences importantes apparaissent entre ce modèle multi-niveaux et le modèle classique de la section précédente. Tout d'abord, un terme d'interaction entre les deux variables explicatives ($Age_{ijk}ClSaoul_{jk}$) a fait son apparition, tant dans la partie fixe que dans la partie aléatoire du modèle, sans avoir été spécifié explicitement dans les équations (3) à (9). Ensuite, des relations complexes apparaissent entre les différents niveaux du modèle, non seulement entre les variables explicatives, mais aussi entre les termes d'erreur et les variables explicatives. La probabilité d'avoir consommé du cannabis est expliquée ici par une somme de termes fixes et aléatoires en provenance de chacun des niveaux de la hiérarchie.

Dans le modèle (12) construit ci-dessus, nous avons supposé que les données obéissaient à une hiérarchie à trois niveaux : les élèves (niveau 1), les classes (niveau 2) et les régions (niveau 3). Tout comme dans le cas de l'approche classique, il convient de se demander si une structure aussi complexe est bien nécessaire. Pour cela, nous utilisons à nouveau la mesure *deff* représentant l'effet plan sur les estimations. Dans un premier temps, il est possible d'estimer la corrélation intra-classe en calculant un modèle dans lequel tous les facteurs explicatifs ont été supprimés, c'est-à-dire le modèle suivant :

$$\pi_{ijk} = \gamma_{000} + e_{ijk} + u_{0jk} + u_{00k} \tag{13}$$

Si ce modèle n'explique en rien la variabilité des observations, il procure en revanche une décomposition de leur variance en trois termes distincts correspondant chacun à l'un des niveaux de la hiérarchie. Soit σ_1^2 la variance des termes d'erreur e_{ijk} du niveau 1, σ_2^2 la variance des termes d'erreur u_{0jk} du niveau 2 et σ_3^2 celle des termes d'erreur u_{00k} du niveau 3. La corrélation intra-classe pour le niveau 1 à l'intérieur du niveau 2 se calcule alors comme

$$\rho_{1|2} = \frac{\sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$$

Cette corrélation représente le degré de similarité des élèves au sein d'une même classe. Le numérateur et le dénominateur comprennent non-seulement la variance du terme d'erreur du niveau 2, mais aussi celle du terme d'erreur du niveau 3, car des élèves appartenant à la même classe appartiennent aussi a fortiori à la même région. Cette corrélation tient donc compte de l'intégralité de la hiérarchie. Il est cependant également correct d'exclure σ_3^2 du numérateur de la corrélation. Dans ce cas, la corrélation intra-classe ne s'interprète plus comme le degré de similarité des élèves au sein des classes, mais simplement comme la part de la variance totale qui est attribuable aux classes (Hox, 2002).

L'estimation du modèle (13) donne les résultats suivants :

$$\sigma_1^2 = 1 \ , \quad \sigma_2^2 = 0.1887 \ , \quad \sigma_3^2 = 0.0093$$

Il est à noter que la variance du terme d'erreur du niveau 1, σ_1^2, n'est généralement pas donnée explicitement par les logiciels statistiques, car dans le cas de modèles logistiques, elle s'interprète comme un simple paramètre d'échelle fixé à 1 (Hox, 2002). La corrélation intra-classe vaut alors

$$\rho_{1|2} = \frac{0.1887 + 0.0093}{1 + 0.1887 + 0.0093} = \frac{0.1980}{1.1980} = 0.1653$$

et nous obtenons

$$\textit{deff} = 1 + (12.31 - 1)\, 0.1653 = 2.87$$

où 12.31 est le nombre moyen d'élèves par classe calculé sur le nombre total d'observations pondérées (7126 / 579 = 12.31). Si les données avaient été collectées à l'aide d'un plan de sondage simple sans stratification ni grappes, il aurait ainsi suffi de 7126 / 2.87 = 2483 données pondérées pour obtenir des estimations offrant la même précision que celle que nous pouvons espérer obtenir ici.

En prenant la racine carrée de *deff*, nous obtenons la mesure *deft* = 1.69 qui nous indique que les écarts-types calculés à partir des données issues du plan de sondage complexe sont en moyenne 1.69 fois plus grands que ceux qui seraient obtenus avec un échantillon de même taille issu de tirages aléatoires simples dans l'ensemble de la population.

Afin de vérifier l'utilité du niveau des régions dans la modélisation, nous avons ensuite considéré le modèle suivant :

$$\pi_{ijk} = \gamma_{00} + e_{ij} + u_{0j} \tag{14}$$

Il s'agit d'un modèle à deux niveaux, les élèves et les classes, sans facteurs explicatifs, où e_{ij} et u_{0j} sont les termes d'erreur des deux niveaux avec des variances estimées respectivement à $\sigma_1^2 = 1$ et $\sigma_2^2 = 0.1918$. La corrélation intra-classe vaut alors

$$\rho_{1|2} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{0.1918}{1 + 0.1918} = 0.1609$$

et nous obtenons *deff* = 2.82 et *deft* = 1.68.

Les mesures *deff* et *deft* étant presque identiques pour les modèles à deux et trois niveaux, il est inutile de tenir compte du niveau des régions. Par ailleurs, en utilisant les log-vraisemblances des deux modèles, il est aussi possible de tester leur différence au moyen de la statistique de déviance. Les modèles à 3 et 2 niveaux ayant des log-vraisemblances égales respectivement à -4793.7301 et -4794.0813, nous obtenons

$$Deviance = -2(-4794.0813 - (-4793.7301)) = 0.7025 \sim \chi_1^2$$

où le nombre de degré de liberté est égal à la différence du nombre de paramètres entre les deux modèles. Le seuil de rejet correspondant à un risque de première espèce de 5% vaut 3.84, ce qui correspond à une p-valeur de 0.4020. L'hypothèse d'égalité des deux modèles est donc très nettement acceptée et il est possible de ne conserver que la structure hiérarchique à deux niveaux en excluant les régions. Le modèle (12) peut alors être simplifié en supprimant tous les éléments relatifs aux régions de façon à se ramener au modèle à deux niveaux suivant :

$$\pi_{ij} = \gamma_{00} + \gamma_{10}\, Age_{ij} + \gamma_{01}\, ClSaoul_j + \gamma_{11}\, Age_{ij} ClSaoul_j + e_{ij} + u_{0j} + u_{1j}\, Age_{ij} \tag{15}$$

où $\gamma_{00} + \gamma_{10} Age_{ij} + \gamma_{01} ClSaoul_j + \gamma_{11} Age_{ij} ClSaoul_j$ est la partie fixe et $e_{ij} + u_{0j} + u_{1j} Age_{ij}$ la partie aléatoire.

Ce modèle n'est cependant pas encore optimal, car en l'estimant nous constatons que les quatre paramètres fixes du modèle sont statistiquement égaux à zéro (voir annexe A.2). Nous décidons alors de le simplifier en supprimant le facteur explicatif le moins significatif, c'est-à-dire l'effet d'interaction entre les variables *Age* et *ClSaoul*. Le modèle final ainsi obtenu s'écrit

$$\pi_{ij} = -3.8098 + 0.1454\, Age_{ij} + 2.2382\, ClSaoul_j + e_{ij} + u_{0j} + u_{1j}\, Age_{ij} \tag{16}$$

avec

$$Var(u_{0j}) = 16.2159\,, \quad Var(u_{1j}) = 0.0530\,, \quad Cov(u_{0j}, u_{1j}) = -0.9267$$

Ainsi que nous pouvons le constater à la lecture du listing complet (annexe A.3), les effets fixes du modèle sont maintenant significatifs, mais en revanche le modèle est statistiquement différent du modèle précédent en terme de déviance. En effet

$$Deviance = -2(-4648.71 - (-4645.2426)) = 6.9348 \sim \chi^2_1$$

Le seuil de rejet du test étant fixé à 3.84, l'égalité des deux modèles est nettement rejetée avec une p-valeur égale à 0.0085. Nous décidons toutefois de conserver le modèle (16) en raison de la meilleure interprétabilité de ses coefficients fixes.

En ce qui concerne les termes aléatoires, tant u_{0j} que u_{1j} semblent avoir une variance nulle lorsque l'on compare les estimations données ci-dessus avec les écarts-types correspondant au moyen d'un test de Wald. Cependant, ce test assume que le paramètre est distribué de façon normale, ce qui n'est certainement pas le cas avec les variances. Il est préférable alors de recalculer le modèle en supprimant un des termes aléatoires et de refaire un test de déviance. En pratique, nous avons supprimé le paramètre aléatoire u_{1j} associé à l'âge au niveau 2. Le listing complet est donné en annexe A.4. La statistique de comparaison des deux modèles vaut alors

$$Deviance = -2(-4652.9461 - (-4648.71)) = 8.4722 \sim \chi^2_2$$

Nous avons ici une loi du chi-2 à 2 degrés de liberté, car deux paramètres, la variance de u_{1j} et la

covariance entre u_{0j} et u_{1j}, ont été supprimés. Le seuil de rejet du test étant fixé à 5.99, l'égalité des deux modèles est rejetée avec une p-valeur égale à 0.0145. Il est donc possible d'affirmer que l'effet de l'âge varie significativement d'une classe à l'autre.

Selon le modèle (16), la valeur moyenne du logit basée sur les seuls effets fixes pour un jeune de 16 ans appartenant à une classe dans laquelle personne n'a été saoul vaut

$$\pi_{ijk} = 3.8098 + 0.1454 \cdot 16 + 2.2382 \cdot 0 = -1.4834$$

ce qui équivaut à une probabilité de 18.49% d'avoir déjà consomé du cannabis. Si nous considérons maintenant un jeune de 20 ans appartenant à une classe dans laquelle tout le monde a déjà été saoul, le logit vaut 1.3364 pour une probabilité d'avoir déjà consommé du cannabis égale à 79.19%.

Ces chiffres sont proches des résultats obtenus à partir de l'approche classique (respectivement 20.42% et 79.41%). Toutefois, il faut également prendre en compte la partie aléatoire du modèle. La variance de u_{1j} indique que l'effet d'un même âge peut être très différent d'une classe à l'autre. Par ailleurs, la forte variabilité de u_{0j} (16.2159) implique qu'il existe une forte variabilité de la probabilité d'avoir consommé du cannabis, cela même entre deux classes pour lesquelles la variable *ClSaoul* prend la même valeur. Il est donc probable que d'autres facteurs explicatifs devraient être pris en compte afin d'améliorer la modélisation de la variable dépendante.

4 Conclusion

Ainsi que nous l'avons montré, l'analyse statistique des données issues d'enquêtes nécessite la prise en compte du plan de sondage, ce qui ne peut pas toujours être réalisé à l'aide des logiciels statistiques courants. La seule exception notable à cette constatation est Stata qui intègre d'origine un ensemble complet de procédures pour l'analyse des données d'enquêtes. Pour les autres logiciels généralistes, il est toujours nécessaire d'ajouter au moins un module additionnel, comme par exemple *Complex samples* pour SPSS, *SAS/STAT* ou *SUDAAN* pour SAS et le *Survey package* (Lumley, 2004) pour R. Le module additionnel *GLLAMM* ajoute à Stata la modélisation multi-niveaux et les modèles d'équations structurelles (Skrondal & Rabe-Hesketh, 2003). L'autre possibilité consiste

à utiliser certains des logiciels spécialisés cités dans l'introduction de l'article. Au vu de la rapidité de développement de nouveaux logiciels et de nouvelles procédures statistiques, il est probable que beaucoup plus d'outils seront d'ici peu à même de traiter correctement les données d'enquêtes et de proposer un ensemble cohérent d'analyses. Il est donc inutile de dresser ici une liste exhaustive de l'offre actuelle et nous laissons au lecteur le soin de s'informer. Parmi les sources d'information fiables en la matière, nous pouvons citer

- *Techniques d'enquête* (http ://www.statcan.ca/francais/ads/12-001-XPB/index_f.htm) ;
- l'*Association for Survey Computing* (http ://www.asc.org.uk/) ;
- le site de l'UCLA dédié aux enquêtes (http ://statcomp.ats.ucla.edu/survey/).

Le choix d'une approche statistique, soit classique, soit multi-niveaux, doit être fait en tenant compte de différents facteurs parfois contradictoires, mais il convient de préciser que les deux approches sont a priori tout aussi correctes l'une que l'autre. Au-delà de la simple disponibilité des outils informatiques eux-mêmes, il faut être conscient que la complexité du plan de sondage a une influence sur les temps de calcul des modèles et que les logiciels actuels ne permettent que rarement d'estimer en un laps de temps raisonnable des modèles multi-niveaux ayant plus de trois niveaux et/ou comportant beaucoup de facteurs explicatifs. Cette limitation est nettement moindre dans le cas de l'approche classique. Si nous excluons les contingences matérielles, la grande différence entre les approches classique et multi-niveaux tient à la façon de prendre en compte le plan de sondage. Dans l'approche classique, il est un élément exogène au modèle et nous cherchons à obtenir un résultat comportant le moins de variabilité possible au niveau des paramètres. Dans l'approche multi-niveaux, le plan de sondage est au contraire endogène au modèle, ce qui se traduit par des résultats dans lesquels les coefficients sont en grande partie aléatoires. L'approche multi-niveaux est ainsi plus riche du point de vue des possibilités de modélisation, mais ses résultats sont plus complexes à interpréter (Berchtold, 2007). De plus, Kreft (1996) ajoute une limitation concernant l'utilisation des modèles multi-niveaux. Du moment que ces modèles prennent en compte les interactions existant à tous les niveaux de la structure des données, il est probable qu'ils soient souvent surentraînés en regard du jeu de données utilisé et qu'ils soient donc moins facilement généralisables que des modèles plus simples.

Une dernière question importante a trait à l'importance respective des différents éléments du plan de sondage. En ce qui concerne les pondérations d'échantillonnage, il semble indispensable de les utiliser puisque leur non prise en compte peut engendrer des erreurs dès le niveau de l'estimation ponctuelle des paramètres. Aucun modèle ne peut donc a priori s'en affranchir. Dans le cas des effets de grappes et de la stratification, la réponse doit être plus nuancée. Nous savons que la prise en compte de ces éléments va influencer le calcul des variances, des intervalles de confiance et des p-valeurs des tests d'hypothèses. De ce fait, la prise en compte du plan de sondage se traduit par des résultats plus conservateurs. Lorsque les effets des corrélations intra-classes et du design peuvent être considérés comme faibles, il est alors probable que la prise en compte des strates et des grappes ne modifiera que peu les résultats, et si les paramètres obtenus sans cela sont fortement significatifs, les résultats pourront être considérés comme fiables.

Remerciements

Ce document a été écrit suite aux demandes de plusieurs chercheurs et aux discussions très constructives que j'ai eues avec eux. Je tiens donc à remercier ici tout particulièrement Jean-Philippe Antonietti, André Jeannin, Pierre-André Michaud et Joan-Carles Suris.

L'enquête SMASH est une enquête multidisciplinaire multi-centres conduite en 2002 par l'Institut de médecine sociale et préventive de Lausanne (*), the Institute for Psychology, Psychology of Development and Developemental Disorders, University of Berne, Switzerland (**) et la Sezione Sanitaria, Dipartimento della sanità e della socialità, Canton Ticino (***) : Véronique Addor*, Françoise Alsaker**, Andrea Bütikofer**, Chantal Diserens*, Laura Inderwildi Bonivento***, André Jeannin*, Guy van Melle*, Pierre-André Michaud*, Françoise Narring**, Joan-Carles Suris*, Annemarie Tschumper**.

A Estimation des modèles classique et multi-niveaux

Les listings suivants ont été générés à l'aide de Stata 9.0 en utilisant la fonction intégrée *svy :logistic* pour l'approche classique et la fonction additionnelle *GLLAMM 6.0* pour l'approche multi-niveaux.

A.1 Approche classique : Prise en compte des pondérations d'échantillonnage et des classes

```
Survey: Logistic regression

Number of strata   =         1                  Number of obs      =      7278
Number of PSUs     =       579                  Population size    = 7125.8229
                                                Design df          =       578
                                                F(   2,    577)    =     47.79
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
    cannabis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1257709   .0417144     3.02   0.003     .0438407    .2077011
     clsaoul |    2.19147   .2384827     9.19   0.000     1.723071    2.659868
       _cons |  -3.442066   .7581541    -4.54   0.000    -4.931139   -1.952993
------------------------------------------------------------------------------

------------------------------------------------------------
             |             Linearized
    cannabis |      Coef.   Std. Err.       Deff      Deft
-------------+----------------------------------------------
         age |   .1257709   .0417144     3.84013   1.95963
     clsaoul |    2.19147   .2384827     2.83278   1.68309
       _cons |  -3.442066   .7581541     4.04525   2.01128
------------------------------------------------------------
```

A.2 Approche multi-niveaux : Modèle à deux niveaux incluant tous les facteurs explicatifs

```
log likelihood = -4645.2426

Robust standard errors
------------------------------------------------------------------------------
      q80_a1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3944172    .268475     1.47   0.142    -.1317842    .9206186
     clsaoul |   8.495547   5.951398     1.43   0.153     -3.16898    20.16007
    interact |  -.3504966   .3369919    -1.04   0.298    -1.010988    .3099954
       _cons |  -8.244142   4.742058    -1.74   0.082     -17.5384     1.05012
------------------------------------------------------------------------------

Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (classno)

    var(1): 15.987676 (11.426507)
    cov(2,1): -.91027815 (.64191511) cor(2,1): -.99951169

    var(2): .05187847 (.03607697)
------------------------------------------------------------------------------
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q80_a1       |
         age |   .3944172    .268475     1.47   0.142    -.1317842    .9206186
     clsaoul |   8.495547   5.951398     1.43   0.153     -3.16898    20.16007
    interact |  -.3504966   .3369919    -1.04   0.298    -1.010988    .3099954
       _cons |  -8.244142   4.742058    -1.74   0.082     -17.5384     1.05012
-------------+----------------------------------------------------------------
cla1_1       |
       const |   3.998459   1.428864     2.80   0.005     1.197938    6.798981
-------------+----------------------------------------------------------------
cla1_2       |
         age |  -.0071171   .0063383    -1.12   0.261      -.01954    .0053058
-------------+----------------------------------------------------------------
cla1_2_1     |
       _cons |  -.2276572   .0792691    -2.87   0.004    -.3830219   -.0722926
------------------------------------------------------------------------------
```

A.3 Approche multi-niveaux : Modèle à deux niveaux sans effet d'interaction

```
log likelihood = -4648.71

Robust standard errors
------------------------------------------------------------------------------
      q80_a1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1453852   .0509069     2.86   0.004     .0456095    .2451608
     clsaoul |   2.238222   .2468234     9.07   0.000     1.754457    2.721987
       _cons |  -3.809843   .9260128    -4.11   0.000    -5.624795   -1.994891
------------------------------------------------------------------------------

Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (classno)

    var(1): 16.215922 (11.192797)
    cov(2,1): -.92668637 (.63005651) cor(2,1): -.99962018

    var(2): .05299732 (.03547814)
------------------------------------------------------------------------------
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q80_a1       |
         age |   .1453852   .0509069     2.86   0.004     .0456095    .2451608
     clsaoul |   2.238222   .2468234     9.07   0.000     1.754457    2.721987
       _cons |  -3.809843   .9260128    -4.11   0.000    -5.624795   -1.994891
-------------+----------------------------------------------------------------
cla1_1       |
       const |     4.0269   1.389754     2.90   0.004     1.303033    6.750767
-------------+----------------------------------------------------------------
cla1_2       |
         age |   .0063444   .0070514     0.90   0.368    -.0074762    .0201649
-------------+----------------------------------------------------------------
cla1_2_1     |
       _cons |   -.230124   .0771168    -2.98   0.003    -.3812702   -.0789778
------------------------------------------------------------------------------
```

A.4 Approche multi-niveaux : Modèle à deux niveaux sans effet d'interaction et sans effet aléatoire de l'âge

```
log likelihood = -4652.9461

Robust standard errors
------------------------------------------------------------------------------
      q80_a1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1331405   .0438802     3.03   0.002     .0471369     .219144
     clsaoul |   2.219148   .2514646     8.82   0.000     1.726286    2.712009
       _cons |  -3.592761   .8013533    -4.48   0.000    -5.163384   -2.022137
------------------------------------------------------------------------------

Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (classno)

    var(1): .04827428 (.03222945)
------------------------------------------------------------------------------
------------------------------------------------------------------------------
      q80_a1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q80_a1       |
         age |   .1331405   .0438802     3.03   0.002     .0471369     .219144
     clsaoul |   2.219148   .2514646     8.82   0.000     1.726286    2.712009
       _cons |  -3.592761   .8013533    -4.48   0.000    -5.163384   -2.022137
-------------+----------------------------------------------------------------
cla1_1       |
       const |   .2197141   .0733441     3.00   0.003     .0759623    .3634658
------------------------------------------------------------------------------
```

Références

Asparouhov T. (2006) General Multilevel Modeling with Sampling Weights. *Communications in Statistics : Theory and Methods*, 35, 439-460.

Berchtold A. (2007) Key elements in the statistical analysis of surveys. *International Journal of Public Health*, 52, 117-119.

Brogan D.J. (1998) Pitfalls of using Standard Statistical Software Packages for Survey Sample Data. In *Encyclopedia of Biostatistics*, Vol. 5, edited by Peter Armitage & Theodore Colton, John Wiley & Sons, New York.

Brogan D.J. (2005) Sampling error estimation for survey data. In *Household Sample Surveys in Developing and Transition Countries*, Chapter XXI, United Nations.

Chantala K., Suchindran C.M. & Blanchette D. (2005) Adjusting for Unequal Selection Probability in Multilevel Models : A Comparison of Software Packages. Atelier sur les modèles de variables latentes et les données d'enquêtes en sciences sociales et en santé, Université de Montréal. www.stata.com/meeting/4nasug/Chantala.ppt

Chantala K., Blanchette D. & Suchindran C.M. (2006) Software to Compute Sampling Weights for Multilevel Analysis. Carolina Population Center, UNC at Chapel Hill. Disponible sur le site http ://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights

Cochran W.G. (1977) *Sampling Techniques*, 3rd edition. Wiley, New York.

Cohen S.B. (1997) An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data. *The American Statistician*, 51, 285-292.

Crockett A. (2004) Weighting the Social Surveys. UK Data Archive and Institute for Social and Economic Research.

Ellis S.J., Hesterberg T.C. (1999) Computation of Weighted Functional Statistics Using Software That Does Not Support Weights. Mathsoft Research Report No. 85. http ://www.insightful.com/Hesterberg/articles/tech85-weightedFunc.ps

Goldstein H. (2003) *Multilevel Statistical Models*, 3rd edition. Hodder Arnold, London.

Groves R.M., Fowler F.J., Couper M.P., Lepkowski J.M., Singer E. & Tourangeau R. (2004) *Survey Methodology*. Wiley Series in Survey methodology, John Wiley & Sons, New Jersey.

Hox J.J. (2002) *Multilevel Analysis : Techniques and Applications*. Lawrence Erlbaum Associates, London.

Kalton G. (1983) *Introduction to Survey Sampling.* Quantitative Applications in the Social Sciences, 35. Sage Publications.

Kish L. (1965) *Survey Sampling.* John Wiley & Sons, New York.

Kreft Ita G.G. (1996) Are multilevel techniques necessary ? An overview, including simulation studies. California Stata University, Los Angeles. http ://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html

Lê T., Brick J.M., Kalton G. (2002) Decomposing Design Effects. Proceedings of the Survey Research Methods Section, American Statistical Association. http ://www.amstat.org/sections/SRMS/Proceedings/y2002/Files/JSM2002-000761.pdf

Lumley T. (2004) Analysis of complex survey samples. *Journal of Statistical Software*, 9(8). http ://www.jstatsoft.org/v09/i08/paper.pdf

Mitchell N.M. (2005) Strategically using General Purpose Statistics Packages : A Look at Stata, SAS and SPSS Technical Report Series, Report Number 1, Version Number 1. Statistical Consulting Group : UCLA Academic Technology Services. Available at http ://www.ats.ucla.edu/stat/technicalreports/

Muthen B.O., Satorra A. (1995) Complex Sample Data in Structural Equation Modeling. *Sociological Methodology*, 25, 267-316.

Narring F., Tschumper A., Inderwildi Bonivento L., Jeannin A., Addor V., Bütikofer A., Suris J., Diserens C., Alsaker F., Michaud P. (2004) Santé et styles de vie des adolescents âgés de 16 à 20 ans en Suisse (2002). SMASH 2002 : Swiss multicenter adolescent study on health 2002. Lausanne. Lausanne : Institut universitaire de médecine sociale et préventive ; Bern : Institut für Psychologie ; Bellinzona : Sezione sanitaria.

Pfeffermann D., Skinner C.J., Holmes D.J., Goldstein H. & Rasbash J. (1998) Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society, series B*, 60, 123-140.

Skrondal A. & Rabe-Hesketh S. (2003) Some applications of generalized linear latent and mixed models in epidemiology : Repeated measures, measurement error and multilevel modeling. *Norsk Epidemiologi*, 13, 265-278.

Tillé Y. (2001) *Théorie des sondages : Echantillonnage et estimation en populations finies.* Dunod, Paris.