Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 137-147
Statistique Canada, N° 12-001 au catalogue

Études cas-témoins basées sur la population

Alastair Scott [1]

Résumé

Nous discutons de méthodes d'analyse des études cas-témoins pour lesquelles les témoins sont sélectionnés selon un plan de sondage complexe. La méthode la plus simple est l'approche du sondage standard basée sur des versions pondérées des équations d'estimation pour la population. Nous examinons aussi des méthodes plus efficaces et comparons leur degré de robustesse aux erreurs de spécification du modèle dans des cas simples. Nous discutons également brièvement des études familiales cas-témoins, pour lesquelles la structure intragrappe présente un intérêt en soi.

Mots clés : Études cas-témoins; échantillonnage sélectif; échantillonnage rétrospectif; pondération.

1. Introduction

L'étude cas-témoins, dans laquelle des échantillons distincts sont tirés parmi les « cas » (disons, les personnes présentant une maladie d'intérêt) et parmi les « témoins » (personnes n'ayant pas la maladie), est l'une des méthodes d'étude les plus fréquentes en recherche sur la santé. En fait, Breslow (1996) a qualifié ce genre d'étude de « pivot de l'épidémiologie ». Nous nous concentrerons ici sur les applications biostatistiques, mais le plan de base représente une stratégie d'échantillonnage efficace dans les situations où les cas sont rares et est aussi utilisé fréquemment dans de nombreux autres domaines (commerce, sciences sociales, écologie, études de marché, par exemple). On a notamment assisté dans la littérature économétrique au développement parallèle de la plupart de la théorie de l'échantillonnage basé sur les choix (voir, par exemple, Manski et McFadden 1981; Cosslett 1981).

Il existe deux types fondamentalement différents d'études cas-témoins, à savoir les études avec cas-témoins appariés, dans lesquelles chaque cas est apparié à un ou à plusieurs témoins, et les études avec cas-témoins non appariés, dans lesquelles les échantillons de cas et de témoins sont tirés indépendamment, quoiqu'il puisse exister un vague « appariement fréquentiste », l'échantillon de témoins étant réparti entre des strates définies par des variables démographiques de base de façon telle que la distribution de ces variables dans l'échantillon de témoins soit la même que celle prévue dans l'échantillon de cas. Nous ne nous intéressons ici qu'aux études sans appariement et, plus précisément, uniquement à la catégorie restreinte d'études sur la population dans lesquelles les témoins (et, à l'occasion, les cas également) sont sélectionnés en utilisant des méthodes standard d'échantillonnage.

Une excellente introduction à l'échantillonnage cas-témoins et à ses points forts et ses inconvénients éventuels est donnée dans Breslow (1996, 2004). L'un des plus grands défis que doit relever toute personne qui conçoit une étude de ce genre consiste à s'assurer que les témoins soient réellement sélectionnés à partir de la même population, selon les mêmes protocoles, que les cas. Comme l'a dit Miettinen (1985), les cas et les témoins « devraient être représentatifs de la même expérience fondamentale » [traduction]. Le fait qu'au début, des mesures adéquates n'aient pas été prises à cet égard lors de certaines études a suscité des doutes au sujet de l'échantillonnage cas-témoins chez de nombreux chercheurs. Une discussion approfondie des principes qui devraient régir la sélection des témoins figure dans Wacholder, McLaughlin, Silverman et Mandel (1991). Puisque l'essence de l'échantillonnage tient aux méthodes suivies pour tirer des échantillons représentatifs à partir d'une population cible, il est devenu naturel de penser aux méthodes de sondage pour obtenir les témoins. De plus en plus fréquemment, au cours des quelque 25 dernières années, les témoins (et parfois les cas également) ont été sélectionnés en utilisant des plans de sondage complexes stratifiés à plusieurs degrés. Le lecteur trouvera au chapitre 9 de Korn et Graubard (1999) un bon historique de cette évolution.

L'analyse de ce genre d'études est un sujet tout indiqué pour le présent article, car Joe Waksberg lui-même a été l'un des principaux artisans de l'adoption des méthodes de sondage (et de la composition aléatoire, en particulier) pour l'obtention des témoins (voir, par exemple, Waksberg 1998, ainsi que DiGaetano et Waksberg 2002).

2. Exemples

Nous commençons par deux exemples en vue d'illustrer le genre de problèmes que nous voulons résoudre. Le premier est typique des études à grande échelle réalisées par le National Cancer Institute, aux employés duquel nous devons la plupart des progrès réalisés dans le domaine. Joe Waksberg, et ses collègues de Westat ont eu une

1. Alastair Scott, Department of Statistics, University of Auckland, Auckland 1, Nouvelle-Zélande. Courriel : a.scott@auckland.ac.nz.

influence considérable sur le choix des méthodes d'échantillonnage adoptées pour réaliser ces études (voir Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg 1984 qui donnent aussi une description de plusieurs études similaires), de sorte qu'il s'agit d'un point de départ naturel.

Exemple 1

En 1977 – 1978, le National Cancer Institute et l'Environmental Protection Agency des États-Unis ont réalisé une étude cas-témoins sur la population afin d'examiner les effets des rayonnements ultraviolets sur le cancer de la peau de type non-mélanome au cours d'une période d'un an (Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg 1984; Fears et Gail 2000). L'étude a été réalisée dans huit régions géographiques où l'intensité des rayonnements solaires ultraviolets différait. Dans chaque région, un échantillon de personnes atteintes d'un cancer de la peau de type non-mélanome âgées de 20 à 74 ans et un échantillon de témoins sélectionnés dans la population générale ont été interviewés par téléphone afin d'obtenir des renseignements sur les facteurs de risque. Pour chaque région, un échantillon aléatoire simple de 450 cas et un échantillon supplémentaire de 50 cas du groupe des 20 à 49 ans ont été sélectionnés en vue d'une prise de contact. Pour les témoins, 500 ménages ont été échantillonnés dans chaque région selon la méthode de composition aléatoire de Mitofsky-Waksberg (Waksberg 1978). On s'est efforcé, dans la mesure du possible, d'interviewer tous les adultes de 65 à 74 ans, ainsi qu'une personne de chaque sexe de 20 à 64 ans choisie au hasard. En outre, un deuxième échantillon de Mitofsky-Waksberg contenant de 500 à 2 100 ménages a été sélectionné et des renseignements ont été recueillis au sujet de tous les adultes de 65 à 74 ans. Cela a donné un échantillon d'environ 3 000 cas et un échantillon d'environ 8 000 témoins, le taux d'échantillonnage des cas étant environ 300 fois celui des témoins, selon l'âge.

Le deuxième exemple revêt une importance particulière pour moi, car il correspond à la première incursion que Chris Wild et moi-même avons faite dans ce domaine.

Exemple 2

L'étude de la méningite à Auckland a été exécutée à la demande du ministère de la Santé et du Conseil de recherche sur la santé de la Nouvelle-Zélande en vue d'examiner les facteurs de risque de méningite chez les jeunes enfants parmi lesquels la progression de la maladie prenait des proportions épidémiques (voir Baker, McNicholas, Garrett, Jones, Stewart, Koberstein et Lennon 2000). L'étude avait pour population cible l'ensemble des enfants de moins de neuf ans dans la région d'Auckland entre 1997 et 2000.

Tous les cas de méningite relevés dans le groupe d'âge cible au cours des trois années qu'a duré l'étude ont été inclus, ce qui a donné environ 250 cas. Un nombre comparable de témoins ont été sélectionné parmi les autres enfants faisant partie de la population étudiée selon un plan d'échantillonnage complexe à plusieurs degrés. Au premier degré d'échantillonnage, 300 îlots de recensement (contenant chacun environ 70 ménages) ont été sélectionnés avec probabilité proportionnelle au nombre de maisons dans l'îlot. Au deuxième degré, un échantillon systématique de 20 ménages a été tiré dans chaque îlot sélectionné et les enfants provenant de ces ménages ont été choisis pour l'étude avec une probabilité variant selon l'âge et l'ethnicité, déterminée de façon qu'elle corresponde à la fréquence prévue parmi les cas. Les probabilités de sélection sont présentées plus loin au tableau 1 (IP signifie originaire des îles du Pacifique). La taille des échantillons de grappes varie de 1 à 6, et environ 250 témoins ont été sélectionnés en tout. Cela correspond à une fraction d'échantillonnage d'environ 1 pour 400, en moyenne, de sorte que les cas ont été échantillonnés à un taux d'environ 400 fois celui des témoins.

Ces deux études sont assez représentatives de celles dont nous voulons discuter. Elles illustrent aussi les deux principales méthodes d'échantillonnage utilisées, à savoir la composition aléatoire et l'échantillonnage aréolaire. Une vive discussion des mérites relatifs de ces deux stratégies figure dans Brogan, Denniston, Liff, Flagg, Coates et Brinton (2001), ainsi que dans DiGaetano et Waksberg (2002).

Tableau 1
Probabilités de sélection

ÂGE	MAORI	ÎLES DU PACIFIQUE	AUTRE
≤ 1 an	0,29	0,70	0,10
≤ 3 ans	0,15	0,50	0,07
≤ 5 ans	0,15	0,31	0,04
≤ 8 ans	0,15	0,17	0,04

3. Conditions générales

Supposons que nous ayons une variable de réponse binaire, Y, avec $Y = 1$ dénotant un cas et $Y = 0$ dénotant un témoin, et un vecteur de variables explicatives éventuelles, $\mathbf{x}$. Nous supposons que la valeur de Y est connue pour chacune des N unités d'une population cible donnée, mais qu'au moins certaines composantes de $\mathbf{x}$ sont inconnues. Nous stratifions la population en cas et en témoins, tirons un échantillon dans chaque strate d'après les variables que nous connaissons pour toutes les unités, et mesurons les valeurs des covariables manquantes pour les unités échantillonnées (en pratique, l'échantillon de témoins est souvent tiré à partir de l'ensemble de la population, plutôt que parmi les unités pour lesquelles $Y = 0$. Si la

proportion de cas est faible, la différence est négligeable. Sinon, il est simple d'adapter les résultats qui suivent à cette variante – pour un développement rigoureux, voir Lee, Scott et Wild 2006). Habituellement, nous voulons ensuite utiliser les données d'échantillon pour ajuster un modèle de régression binaire de la probabilité marginale qu'une unité soit un cas ayant la forme d'une fonction des covariables. Le modèle utilisé est presque toujours logistique avec

$$\text{logit}\,\{P(Y = 1 \mid \mathbf{x})\} = \log\left(\frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})}\right) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1 \qquad (1)$$

disons, où β_0 et $\boldsymbol{\beta}_1$ sont des paramètres inconnus et, tout au long de l'article, nous supposerons que nous avons affaire au modèle (1). Les extensions à des modèles de régression plus généraux sont simples en principe (voir Scott et Wild 2001b), mais les expressions résultantes sont un peu moins élégantes que le modèle logistique.

Comment devrions-nous nous y prendre pour ajuster le modèle (1) sachant les données d'échantillon? Les méthodes efficaces sont faciles à appliquer en cas d'échantillonnage aléatoire simple ou stratifié, mais nous nous intéressons ici à des méthodes d'échantillonnage plus complexes. Très souvent, on omet tout bonnement de tenir compte de l'échantillonnage complexe, ce qui risque d'entraîner tous les problèmes qui se posent habituellement lorsqu'on ne prend pas en compte la structure du plan d'échantillonnage. Des probabilités de sélection variables pourraient fausser la structure de la moyenne de sorte que les estimations produites par les programmes standard risquent d'être non cohérentes. La corrélation intragrappe pourrait réduire la taille effective d'échantillon, de sorte que les erreurs-types produites couramment seraient trop faibles, les intervalles de confiance, trop courts, les valeurs p, trop faibles, et ainsi de suite. Une stratégie simple qu'adoptent certains chercheurs pour réduire au minimum les effets consiste à garder petit le nombre de sujets dans chaque grappe (voir Graubard, Fears et Gail 1989, par exemple). Cela réduit l'effet de plan, donc l'incidence sur la mise en grappes, mais le remède peut être coûteux. Dans les sections qui suivent, nous examinons certains moyens éventuels d'utiliser les plans d'échantillonnage standard, plus rentables.

4. Approche de la pondération de sondage

Une option évidente consiste à suivre l'approche standard des équations d'estimations pondérées qui est intégrée dans la plupart des progiciels contemporains d'analyse de données d'enquête (voir Binder 1983). Supposons d'abord que nous ayons des données provenant de la population finie complète. Si nous émettons l'hypothèse que cette population finie est tirée à partir d'une superpopulation dans laquelle le modèle logistique conditionnel (1) est vérifié, alors nous pourrions estimer $\boldsymbol{\beta}$ en résolvant les équations d'estimation relatives à l'ensemble de la population, c'est-à-dire sous recensement, suivantes

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_1^N \mathbf{x}_i(y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})) = 0, \qquad (2)$$

où $p_1(\mathbf{x};\boldsymbol{\beta}) = e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1}/(1 + e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1})$. (Il s'agit des équations de vraisemblance si l'on suppose que les unités de population sont échantillonnées indépendamment à partir d'une superpopulation, mais que les estimateurs résultants sont convergents sous des structures de population nettement plus réalistes, à condition que le modèle (1) soit vérifié marginalement – voir Rao, Scott et Skinner 1998 pour une discussion plus approfondie.)

Maintenant, pour toute valeur fixée de $\boldsymbol{\beta}$, $\mathbf{S}(\boldsymbol{\beta})$ dans l'équation (2) est simplement un vecteur de totaux de population. Nous pouvons donc l'estimer d'après l'échantillon, disons par

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_{\text{échantillon}} w_i\mathbf{x}_i(y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})), \qquad (3)$$

où w_i est l'inverse de la probabilité de sélection, peut-être corrigée de la non-réponse et de la poststratification. Fixer $\hat{\mathbf{S}}(\boldsymbol{\beta})$ égal à 0 nous donne notre estimateur, $\hat{\boldsymbol{\beta}}$. Nous pourrions appliquer la linéarisation ou le jacknife directement à $\hat{\boldsymbol{\beta}}$ pour obtenir les erreurs-types. Nous pouvons aussi développer $\hat{\mathbf{S}}(\hat{\boldsymbol{\beta}})$ autour de la valeur réelle, $\boldsymbol{\beta}$, et obtenir comme matrice de covariance estimée l'estimateur « sandwich »

$$\hat{\text{Cov}}\{\hat{\boldsymbol{\beta}}\} \approx \mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}\,\hat{\text{Cov}}\{\hat{S}(\hat{\boldsymbol{\beta}})\}\,\mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}, \qquad (4)$$

où $\mathbf{J}(\boldsymbol{\beta}) = -\partial\mathbf{S}/\partial\boldsymbol{\beta}^T = \sum_{\text{échantillon}} w_i\,p_1(\mathbf{x}_i;\boldsymbol{\beta})\,p_0(\mathbf{x}_i;\boldsymbol{\beta})\,\mathbf{x}_i\mathbf{x}_i^T$ avec $p_0 = 1 - p_1$. Puisque $\hat{\mathbf{S}}(\boldsymbol{\beta})$ est un vecteur de totaux, $\hat{\text{Cov}}\{\hat{\mathbf{S}}(\boldsymbol{\beta})\}$ devrait être disponible d'office pour tout plan de sondage standard. De nos jours, la plupart des grands progiciels statistiques (par exemple, SAS (PROC SURVEYLOGISTIC), SPSS (CSLOGISTIC), STATA (SVY:LOGIT), SUDAAN (LOGISTIC)) peuvent traiter de façon courante la régression logistique en cas d'échantillonnage complexe et de pondération. Par conséquent, il est assez simple de produire des estimations pondérées et de faire les inférences connexes.

Strictement parlant, dans notre cadre fondé sur un modèle, les probabilités de sélection sont souvent elles-mêmes des variables aléatoires basées sur une population finie hypothétiquement générée à partir du modèle. Nous pouvons tenir compte de ce fait en utilisant les résultats de Rao (1973), mais la correction est d'ordre $1/N$ et peut être ignorée dans la plupart des études de grande taille.

L'inconvénient de la pondération est, en général, qu'elle a tendance à être inefficace si les poids sont très variables (une règle empirique parfois proposée est que $w_{\max} / w_{\min}$ ne devrait pas être supérieur à 10). Dans les études cas-témoins, la variation des poids est à peu près aussi extrême qu'elle peut l'être. Par exemple, le ratio de $w_{\max}$ par rapport à $w_{\min}$ est approximativement de 300 pour 1 dans l'exemple 1, et de 1 000 pour 1 dans l'exemple 2. Et des ratios encore plus extrêmes ne sont pas inhabituels. Aucun spécialiste chevronné de l'échantillonnage ne s'étonnerait de constater que la pondération n'est guère efficace dans ces circonstances.

Pouvons-nous trouver une solution plus efficace? La réponse est assurément affirmative dans certains cas. Des méthodes de vraisemblance entièrement efficaces ont été élaborées dans des situations où les cas et les témoins sont sélectionnés par échantillonnage aléatoire simple ou stratifié, et ces méthodes peuvent être nettement plus efficaces que les méthodes de pondération. Nous passons ces résultats en revue à la section suivante.

5. Révision : Cas simple

Examinons d'abord la situation la plus simple où les cas et les témoins sont sélectionnés par échantillonnage aléatoire simple et où nous ne disposons d'information au niveau de la population sur aucune des covariables à l'étape de l'élaboration du plan de sondage. Pour ce cas, il existe des méthodes semi-paramétriques entièrement efficaces d'estimation du maximum de vraisemblance bien établies. En outre, ces méthodes sont très faciles à appliquer en utilisant les logiciels standard (Prentice et Pyke 1979) (les méthodes sont semi-paramétriques, parce que la vraisemblance complète dépend de la distribution inconnue des covariables, ce que nous voulons en général modéliser).

En fait, il nous suffit d'ajuster le modèle (1) en utilisant un programme standard de régression logistique sans aucune pondération. Plus précisément, la résolution de l'équation non pondérée

$$\sum_{\text{échantillon}} \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \qquad (5)$$

produit des estimations efficaces de tous les coefficients, sauf l'ordonnée à l'origine. Fait peut-être plus important, toutes les erreurs-types et les inférences résultantes que nous obtenons au moyen du programme standard sont également valides, de nouveau à l'exception de tout résultat où intervient l'ordonnée à l'origine. Il est assez simple de corriger les inférences dans lesquelles intervient l'ordonnée à l'origine, à condition que nous connaissions les fractions d'échantillonnage, mais nous ne nous intéressons souvent qu'aux autres coefficients de toute façon.

Les résultats s'étendent directement à l'échantillonnage aléatoire stratifié, à condition d'inclure dans le modèle une ordonnée à l'origine distincte pour chaque strate. De nouveau, il est possible d'obtenir des estimateurs semi-paramétriques efficaces de tous les coefficients, sauf les ordonnées à l'origine de strate en traitant simplement les données à l'aide d'un programme de régression logistique (non pondérée) ordinaire. De nouveau, les erreurs-types estimées et les inférences connexes sont également valides. Comme dans le cas de l'échantillonnage aléatoire simple, nous pouvons corriger les résultats pour les ordonnées à l'origine de strate à condition que nous connaissions les fractions d'échantillonnage dans les strates, mais, encore une fois, celles-ci ne présentent habituellement que peu d'intérêt.

Donc, dans ces situations simples, les estimations du maximum de vraisemblance sont plus faciles à calculer que les estimations pondérées, et elles sont aussi plus efficaces. Dans quelle mesure le sont-elles? Cela dépend du nombre de covariables, de la grandeur de leur coefficient et du ratio des fractions d'échantillonnage, mais la différence est souvent importante (ainsi, les estimations pondérées ont une efficacité d'environ 50 % dans l'exemple 2 de l'introduction, et de moins de 20 % dans l'exemple du cancer du cerveau que nous examinons à la section 8. Lawless, Kalbfleisch et Wild (1999) discutent de situations où l'efficacité est encore plus faible).

Enfin, nous notons que les estimations du maximum de vraisemblance présentent encore un autre avantage par rapport aux estimations pondérées : elles ont tendance à avoir de nettement meilleures propriétés sur petit échantillon, surtout quand l'efficacité des estimations pondérées est faible. Essentiellement, la pondération réduit la taille effective d'échantillon et cette dernière est l'élément qui détermine le moment où la théorie asymptotique commence à donner une bonne approximation (voir Scott et Wild 2001a pour plus de précisions). Manifestement, le prix de l'adhésion stricte aux poids de population peut être très lourd.

6. Échantillonnage plus complexe

Dans les deux exemples de la section 2, les témoins ont été obtenus par échantillonnage complexe à plusieurs degrés plutôt que par échantillonnage aléatoire simple. Comme nous l'avons mentionné dans l'introduction, cette situation est de plus en plus fréquente dans les études cas-témoins à grande échelle (à l'occasion, comme dans l'exemple 1, les cas sont également sélectionnés selon un plan d'échantillonnage complexe). Il est possible de dériver des estimateurs semi-paramétriques efficaces pour l'échantillonnage stratifié à plusieurs degrés, en supposant que les unité primaire

d'échantillonnage sont sélectionnées indépendamment dans les strates (hypothèse qui est de toute façon celle faite dans tous les progiciels offrant l'approche de pondération par les poids de sondage), mais cela nous oblige à construire des modèles multivariés pour le vecteur des réponses dans une unité primaire d'échantillonnage. Le lecteur trouvera des renseignements détaillés dans Neuhaus, Scott et Wild (2002, 2006). À moins que nous ne nous intéressions à la structure intragrappe en soi (comme dans les études familiales cas-témoins considérées à la section 9, par exemple), l'exercice demande beaucoup trop d'efforts pour être faisable, du moins dans les analyses de routine.

Pouvons-nous faire quelque chose de plus simple sans perdre trop d'efficacité? Naturellement, nous pouvons toujours nous rabattre sur les estimations pondérées. Cependant, elles sont tout aussi inefficaces pour les plans de sondage complexes que pour le cas simple examiné à la section précédente. En fait, nous pouvons obtenir de sensiblement meilleurs résultats sans trop de complications supplémentaires.

Revenons un instant à la situation de la section précédente où nous avons un échantillon aléatoire simple de taille n_1 provenant de la strate de cas et un échantillon aléatoire simple indépendant de taille n_0 provenant de la strate de témoins. Ici, toutes les unités de la strate ℓ ont un poids $w_i \propto W_\ell / n_\ell$, où W_ℓ dénote la proportion de la population dans la strate, pour $\ell = 0,1$. Si nous divisons tout au long par N et posons que $p_0(\mathbf{x};\boldsymbol{\beta}) = 1 - p_1(\mathbf{x};\boldsymbol{\beta})$, alors nous pouvons réécrire l'équation (3) pour l'estimateur pondéré sous la forme

$$W_1 \frac{\sum_{\text{cas}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - W_0 \frac{\sum_{\text{témoins}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (6)$$

De même, nous pouvons écrire l'équation (5) pour l'estimateur efficace du maximum de vraisemblance sous la forme

$$\omega_1 \frac{\sum_{\text{cas}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \omega_0 \frac{\sum_{\text{témoins}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}, \quad (7)$$

où $\omega_\ell = n_\ell /(n_0 + n_1)$, pour $\ell = 0,1$. Ces deux expressions sont des cas particuliers de l'ensemble général d'équations d'estimation

$$\lambda_1 \frac{\sum_{\text{cas}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \lambda_0 \frac{\sum_{\text{témoins}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (8)$$

À mesure que $n_0, n_1 \to \infty$, sous des contraintes faibles quant à la façon dont la population finie est générée à partir de la superpopulation, la solution de (8) converge presque certainement vers la solution $\boldsymbol{\beta}^*$ de

$$\lambda_1 E_1\{\mathbf{X}\, p_0(\mathbf{X};\boldsymbol{\beta}^*)\} - \lambda_0 E_0\{\mathbf{X}\, p_1(\mathbf{X};\boldsymbol{\beta}^*)\} = \mathbf{0}, \quad (9)$$

où $E_\ell\{\cdot\}$ dénote l'espérance conditionnelle sachant que $Y = \ell$ pour $\ell = 0,1$. Si le modèle (1) est vérifié, alors l'équation (8) a pour solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ et $\beta_0^* = \beta_0 + b_\lambda$ avec $b_\lambda = \log(\lambda_1 W_0 / \lambda_0 W_1)$ pour toute valeur positive de λ_0, λ_1 [voir Scott et Wild (1986) pour des détails de la preuve]. Donc, la solution de l'équation (8) produit des estimateurs convergents pour tous les coefficients de régression, sauf le terme constant pour tout $\lambda_\ell > 0$ ($\ell = 0,1$). Comme dans le cas simple, il est facile de corriger les inférences au sujet du terme constant, à condition de connaître la proportion de cas dans la population.

Maintenant, passons à des plans d'échantillonnage plus complexes. Puisque le premier membre de l'équation (9) ne fait intervenir que deux moyennes de sous-population, nous pouvons encore estimer ces moyennes pour tout plan de sondage standard. Cela suggère un estimateur, disons $\hat{\boldsymbol{\beta}}_\lambda$, pour les plans d'échantillonnage généraux qui satisfait

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \lambda_1 \hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \lambda_0 \hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \mathbf{0}, \quad (10)$$

où $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ est l'estimateur sur échantillon de la moyenne de sous-population $E_\ell\{\mathbf{X}(1 - p_\ell(\mathbf{X};\boldsymbol{\beta}))\}$ ($\ell = 0,1$). La matrice de covariance de $\hat{\boldsymbol{\beta}}_\lambda$ peut alors être obtenue à l'aide d'arguments standard de linéarisation, ce qui nous mène à une matrice de covariance estimée (« sandwich »)

$$\hat{\mathrm{Cov}}\{\hat{\boldsymbol{\beta}}_\lambda\} \approx \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1} \hat{\mathrm{Cov}}\{\hat{\mathbf{S}}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)\} \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1}, \quad (11)$$

avec $\mathbf{J}_\lambda(\boldsymbol{\beta}) = (-\partial \hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T)$ et $\hat{\mathrm{Cov}}\{\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta})\} = \lambda_1^2 \hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta})\} + \lambda_0^2 \hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta})\}$. Ici, $\hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})\}$ dénote l'estimation par sondage habituelle qui devrait être disponible systématiquement pour tout plan de sondage standard, puisque $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ est simplement une moyenne estimée.

Tout cela peut aussi être exécuté facilement au moyen de n'importe quel progiciel capable de traiter la régression logistique sous plan de sondage complexe, simplement en spécifiant le vecteur approprié de poids. Plus précisément, supposons que

$$\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta}) = \frac{\sum_{i \in S_\ell} w_i \, \mathbf{x}_i \, (y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta}))}{\sum_{i \in S_\ell} w_i}, \quad (12)$$

où S_1 dénote la sous-population de cas (c'est-à-dire l'ensemble des unités pour lesquelles $Y = 1$) et S_0 dénote la sous-population de témoins (l'ensemble des unités pour lesquelles $Y = 0$). Alors, l'équation d'estimation (9) peut s'écrire sous la forme

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_{\text{échantillon}} w_i^* \, \mathbf{x}_i (y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})) = \mathbf{0}, \quad (13)$$

avec $w_i^* \propto \lambda_\ell w_i / \sum_{i \in S_\ell} w_i$ pour les unités comprises dans $S_\ell (\ell = 0,1)$. Autrement dit, nous devons simplement rééchelonner les poids des cas et des témoins séparément, de sorte que la somme de ceux des cas soit proportionnelle à λ_1 et que la somme de ceux des témoins soit proportionnelle à λ_0, puis les introduire, avec la spécification habituelle de la structure du plan de sondage (strates, unités primaires d'échantillonnage), dans le programme de notre choix. Soulignons que le choix de la constante de proportionnalité n'a pas d'incidence sur le résultat.

Il nous reste à décider des bonnes valeurs de λ_1 et λ_0. Nous pouvons souvent réaliser des gains importants en utilisant les poids d'échantillon ($\lambda_\ell = n_\ell / n$) plutôt que les poids de population ($\lambda_\ell = W_\ell$). Scott et Wild (2002) ont fait état de gains d'efficacité de 50 % ou plus dans l'exemple 2 et dans des simulations basées sur cette population. Les gains devenaient plus importants à mesure que s'intensifiait la force de la relation et qu'augmentait l'effet de la mise en grappe. De surcroît, la couverture des intervalles de confiance était plus proche de la valeur nominale en cas de pondération par les poids d'échantillon dans les simulations.

L'utilisation des poids d'échantillon est la stratégie la plus efficace disponible lorsqu'on a affaire à des échantillons aléatoires simples de cas et de témoins, mais pour des plans de sondage plus complexes, elle n'est plus entièrement efficace. Nous pourrions nous attendre à ce que des poids basés sur une forme de tailles d'échantillon équivalentes donnent de meilleurs résultats. Cette approche produit effectivement certains gains d'efficacité dans des simulations limitées décrites dans Scott et Wild (2001a). Toutefois, les gains sont relativement faibles, du moins lorsque l'effet de plan de l'échantillon de témoins est inférieur à 2, puisque $\mathrm{Cov}\{\hat{\boldsymbol{\beta}}_\lambda\}$ est une fonction de $\boldsymbol{\lambda}$ très plate près de son minimum. Les considérations relatives à la robustesse dont nous discutons à la section 8 pourraient jouer un rôle plus important dans le choix de $\boldsymbol{\lambda}$.

Les avantages qu'offre la pondération d'échantillon peuvent dépendre en grande partie du problème à l'étude. Korn et Graubard (1999, page 327) font remarquer que, dans leur expérience, la stratégie de pondération par les poids d'échantillon produit rarement de gros gains d'efficacité. De toute évidence, la poursuite des travaux, tant empiriques que théoriques, est nécessaire ici. Quoi qu'il en soit, il semble prudent d'ajuster le modèle en utilisant systématiquement les poids d'échantillon ainsi que les poids de population. Si les estimations des coefficients sont semblables, alors nous pouvons porter un jugement en nous basant sur les erreurs-types estimées. Par contre, des écarts significatifs entre les estimations des coefficients indiquent que le modèle a été mal spécifié. Si nous sommes incapables de corriger les déficiences du modèle, alors nous devons bien réfléchir à ce que nous essayons vraiment d'estimer. Nous examinons cette question à la section 8.

7. Échantillonnage stratifié

Le compromis proposé à la section précédente (c'est-à-dire utiliser la pondération standard par les poids de sondage dans les sous-populations définies d'après la situation de cas ou de témoin, mais combiner les sous-populations en utilisant les proportions d'échantillon) semble donner d'assez bons résultats en pratique, mais elle est entièrement *ad hoc*. Pourrions-nous obtenir de meilleurs résultats en adoptant une approche plus systématique?

Dans le cas particulier de l'échantillonnage aléatoire stratifié, où des échantillons indépendants de cas et de témoins sont tirés dans chaque strate, il existe des méthodes entièrement efficaces bien établies et faciles à appliquer. En particulier, si notre modèle comprend une ordonnée à l'origine distincte pour chaque strate, alors la régression logistique non pondérée ordinaire (avec un simple ajustement pour les ordonnées à l'origine de strate si l'on veut les obtenir) est la méthode semi-paramétrique efficace du maximum de vraisemblance (Prentice et Pyke 1979). Il est assez facile de l'étendre à des plans stratifiés plus généraux. Notre modèle est maintenant

$$\mathrm{logit}\{P(Y = 1 \mid \mathbf{x}, \text{ Strate h})\} = \beta_{0h} + \mathbf{x}^T \beta_1, \qquad (14)$$

et l'équivalent stratifié de l'équation d'estimation (7) est

$$\sum_h \left(\lambda_{1h} \frac{\sum_{\text{cas}} \mathbf{x}_i p_{0h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{1h}} - \lambda_{0h} \frac{\sum_{\text{témoins}} \mathbf{x}_i p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{0h}} \right) = \mathbf{0}. \qquad (15)$$

À mesure que $n_{0h}, n_{1h} \to \infty$, la solution de (7) converges presque certainement vers la solution de

$$\sum_h (\lambda_{1h} E_{1h}\{\mathbf{X}\, p_{0h}(\mathbf{X}; \boldsymbol{\beta})\} - \lambda_{0h} E_0\{\mathbf{X}\, p_{1h}(\mathbf{X}; \boldsymbol{\beta})\}) = \mathbf{0}, \qquad (16)$$

avec l'extension évidente de la notation utilisée pour le cas non stratifié. Si le modèle (13) est vérifié, alors l'équation (8) a pour solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ et $\beta_{0h}^* = \beta_{0h} + b_{\lambda h}$ avec $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$. Puisque l'équation (14) ne fait intervenir que des moyennes de strate, nous pouvons estimer facilement ces dernières en utilisant les données provenant de tout plan de sondage raisonnable, par exemple par

$$\hat{\boldsymbol{\mu}}_{\ell h}(\boldsymbol{\beta}) = \frac{\sum_{i \in S_{\ell h}} w_{ih} \mathbf{x}_{ih} (y_{ih} - p_1(\mathbf{x}_{ih}; \boldsymbol{\beta}))}{\sum_{i \in S_{\ell h}} w_{ih}}.$$

La substitution de ces estimateurs aux moyennes d'échantillon dans l'équation (14) donne l'équation d'estimation

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_h \sum_{i \in S_h} w_{ih}^* \mathbf{x}_i (y_i - p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \qquad (17)$$

avec $w_{ih}^* \propto \lambda_{\ell h} w_{ih} / \sum_{i \in S_{\ell h}} w_{ih}$ pour les unités comprises dans $S_{\ell h} (\ell = 0,1; h = 1, ..., H)$. Ce modèle peut être ajusté dans tout programme standard d'analyse de données d'enquête en introduisant ces poids et l'information appropriée sur le plan de sondage. Notons que nous devons faire attention à la façon dont nous incluons ce que nous appelons des « strates » dans la spécification du plan. Si les unités primaires d'échantillonnage sont emboîtées dans les « strates », comme c'est le cas des régions géographiques dans l'exemple 1, il n'y a pas de problème et les strates doivent être incluses de la façon standard. Toutefois, si les unités primaires d'échantillonnage recoupent les « strates », comme c'est le cas de l'âge dans l'exemple 1, et de l'âge et de l'ethnicité dans l'exemple 2, il ne s'agit plus de strates au sens habituel du terme en échantillonnage. Elles ne devraient pas être incluses dans les spécifications du plan, mais simplement être traitées par la pondération.

Parfois, nous voulons modéliser la contribution des variables de strate en utilisant une courbe paramétrique lisse au lieu de les inclure à l'aide de variables muettes. Par exemple, nous pourrions fort bien vouloir inclure une fonction linéaire de l'âge dans notre modèle, tant dans l'exemple 1 que dans l'exemple 2. La méthode de pondération par les poids de sondage et la pondération de compromis proposée à la section 6 s'appliquent l'une et l'autre, et aucun nouveau développement théorique n'est nécessaire. Par contre, les méthodes plus efficaces ne sont guères aussi simples. Des méthodes entièrement efficace ont été élaborées dans la situation où des échantillons aléatoire simple de cas et de témoins sont tirés dans chaque strate (voir Scott et Wild 1997, ainsi que Breslow et Holubkov 1997), mais les équations d'estimation résultantes ne sont pas des combinaisons linéaires des moyennes de strates, et il n'existe aucune manière évidente de les généraliser à des plans d'échantillonnage plus complexes. Néanmoins, il existe un moyen un peu moins efficace, mais facile à étendre. Si nous modifions le modèle (14) en incluant $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$ comme correction, c'est-à-dire si nous supposons que

$$\text{logit}\{P^*(Y = 1 \mid \mathbf{x}, \text{ Strate h})\} = b_{\lambda h} + \beta_{0h} + x^T \beta_1, \qquad (18)$$

alors l'équation (15) produit des estimations convergentes, entièrement efficaces, de tous les coefficients, y compris $\beta_{0h} (h = 1, ..., H)$. L'introduction des mêmes corrections dans les modèles ne contenant pas de terme β_{0h} et où le vecteur $\mathbf{x}$ comprend des fonctions de la variable de stratification produit des estimateurs convergents pour tous les coefficients avec une efficacité habituellement élevée (quoi que non totale) (voir Fears et Brown 1986, ainsi que Breslow et Cain 1988). Cela se généralise à des plans de sondage arbitraires immédiatement. Il nous suffit d'utiliser l'équation (16) en remplaçant p_{1h} par p_{1h}^* défini en fixant $\text{logit}(p_{1h}^*) = b_{\lambda h} + \mathbf{x}^T \beta$. Alors, tout programme d'analyse de données d'enquête qui permet d'appliquer des corrections peut être utilisé pour ajuster le modèle et fournir des estimations des erreurs-types, *etc*.

Quel est notre gain d'efficacité dans ce cas-ci? Nous avons exécuté plusieurs simulations, dont certaines sont décrites dans Scott et Wild (2002). La plupart des scénarios sont fondés sur l'étude de la méningite de l'exemple 2 et nous fixons le ratio de la fraction d'échantillonnage de strate la plus grande à la plus faible dans l'échantillon de témoins à environ 10 pour 1. Sans aucune mise en grappes, le gain d'efficacité dû à l'utilisation de la méthode de correction (qui est le maximum de vraisemblance complète dans ce cas-ci) comparativement à la méthode *ad hoc* n'a jamais été supérieur à 10 %. Les efficacités relatives sont demeurées à peu près les mêmes lors de l'introduction d'une mise en grappes sur l'ensemble des strates. Quand nous sommes passés à la mise en grappes emboîtée dans les strates, les gains ont disparus progressivement à mesure que l'effet de plan augmentait et la méthode *ad hoc* est, en fait, devenue plus efficace que la méthode de correction, lorsque la valeur de l'effet de plan a atteint environ 1,5.

Comme nous l'avons mentionné plus haut, il est possible de produire des estimateurs semi-paramétriques entièrement efficaces si nous sommes prêts à modéliser la structure de dépendance à l'intérieur des unités primaires d'échantillonnage. Nous avons commencé à exécuter certaines simulations. Les premiers résultats donnent à penser que le travail supplémentaire que demande la modélisation ne vaudra presque jamais la peine si nous nous intéressons uniquement aux paramètres du modèle marginal (1). Notre conclusion provisoire est que les méthodes *ad hoc* partiellement pondérées (avec les poids d'échantillon) sont faciles à utiliser et donnent de suffisamment bons résultats pour la plupart des objectifs pratiques couverts par notre expérience, mais il s'agit toutefois d'un autre domaine où il conviendrait de poursuivre les travaux empiriques. Nous soulignons cependant que, pour certains problèmes, comme l'étude familiale cas-témoins dont il est question à la section 9, le comportement intragrappe est intéressant en soi. Il faut alors recourir à des méthodes plus perfectionnées.

8. Robustesse

Il doit y avoir un piège quelque part. Que se passe-t-il si le modèle est incorrect? Quel est alors le prix du gain d'efficacité?

Par construction, l'estimateur pondéré en fonction de la population estime toujours l'approximation logistique linéaire que nous obtiendrions si nous disposions de données pour l'ensemble de la population. Par contre, ce que l'estimateur plus efficace pondéré en fonction de l'échantillon estime dépend des tailles d'échantillons particulières utilisées. Certaines personnes considéreraient cet élément à lui seul comme une raison suffisamment valable d'utiliser l'estimateur pondéré d'après la population et je soupçonne que fort peu d'entre elles jugeraient entièrement satisfaisant que la cible de leur inférence dépende du choix arbitraire de la taille d'échantillon.

Notre estimateur général $\hat{\boldsymbol{\beta}}_\lambda$ satisfaisant (10) converge vers la solution de l'équation (9), disons $\mathbf{B}_\gamma$, avec $\gamma = \lambda_0/(\lambda_0 + \lambda_1)$, qui dépend du modèle réel et de la distribution des covariables, ainsi que de γ. Dans Scott et Wild (2002), nous avons examiné ce qui arrive à $\mathbf{B}_\gamma$ lors d'écarts faibles par rapport au modèle hypothétique (nous nous intéressons aux petits écarts, car en principe, les grands sont décelés par les procédures courantes de vérification des modèles qui devraient alors être améliorés en conséquence). Pour simplifier, supposons que nous ajustions un modèle linéaire ne contenant qu'une seule variable explicative pour le logarithme du rapport de cotes, mais que le modèle réel soit quadratique, disons

$$\text{logit}\{P(Y = 1 \mid x)\} = \beta_0 + \beta_1 x + \delta x^2 \qquad (19)$$

où δ est petit.

De toute évidence, la pente réelle de l'échelle logit, $\beta_1 + 2\delta x$, varie lorsque nous nous déplaçons le long de la courbe. Pour tout $0 < \gamma < 1$, $\mathbf{B}_{\gamma 1}$ est égal à la pente réelle à un point donné sur la courbe. Dénotons cette valeur par $x = x_\gamma$. Soit x_0 la valeur attendue de x dans la population de témoins et soit x_1 la valeur attendue de x dans la population de cas. Nous supposons que $\beta_1 > 0$, de sorte que $x_0 < x_1$. Il s'avère que x_γ est toujours compris entre x_0 et x_1, et que x_γ augmente quand la valeur de γ passe de 0 à 1. Rappelons que la pondération par les poids de sondage correspond à $\gamma = W_0$ et que la pondération par les poids d'échantillon correspond à $\gamma = \omega_0 = n_0/n$. Habituellement, W_0 est sensiblement plus grand que ω_0, de sorte que l'utilisation des poids de sondage donnent une estimation de la pente pour des valeurs plus grandes de x, où la probabilité d'un cas est plus élevée, tandis que la pente estimée d'après la pondération d'échantillon s'approche davantage de la valeur moyenne de x dans la population. La figure 1, adaptée de Scott et Wild (2002), illustre la position dans deux scénarios, l'un avec une courbure positive et l'autre, une courbure négative, basés approximativement sur l'exemple 2. Nous choisissons une valeur de δ telle que celui-ci serait décelé à l'aide d'un test standard du rapport de vraisemblance dans environ 50 % des cas si nous sélectionnions des échantillons aléatoires simples de taille $n_0 = n_1 = 200$ à partir de la population.

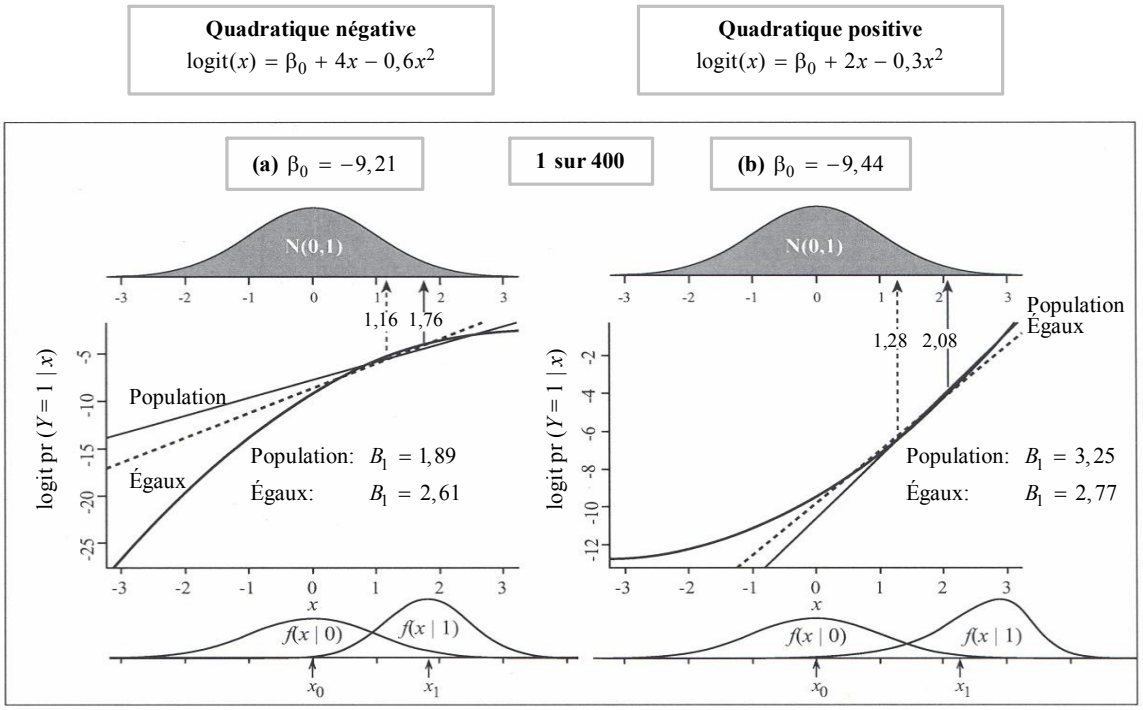

Figure 1. Comparaison entre les poids de population et les poids égaux.

Dans les deux scénarios, la valeur de β_0 est fixée de sorte que la proportion de cas dans la population soit de 1 sur 400, c'est-à-dire $W_0 = 0,9975$. La densité globale de x est représentée à la partie supérieure du graphique et les densités conditionnelles pour les cas et les témoins le sont au bas du graphique. Les valeurs de x_γ et $\mathbf{B}_{\gamma 1}$ sont données pour $\gamma = W_0$ (étiquetées « Population ») et $\gamma = 0,5$ (étiquetées « Égaux »). La seconde valeur correspond à la pondération d'échantillon si nous tirons des nombres égaux de cas et de témoins. Manifestement, dans les deux scénarios, la pondération d'échantillon produit une estimation de la pente appropriée pour des valeurs de x située plus à l'extrémité de la queue supérieure de la distribution (c'est-à-dire pour les personnes à haut risque) que dans le cas de la pondération égale.

Notons que, si nous sélectionnions des échantillons aléatoires simples de taille $n_0 = n_1 = 200$ à partir de la population de la figure 1 (a), l'efficacité relative de la pondération d'échantillon ne serait que d'environ 16 %, et le biais de petit échantillon serait de 0,24. Dans ce cas, même si nous prenions la valeur de population comme cible, la pondération par les poids de sondage produirait une erreur quadratique moyenne plus grande que la pondération d'échantillon.

Un plus grand nombre de résultats sont présentés dans Scott et Wild (2002), où nous examinons aussi l'effet des covariables omises. Celles-ci s'avèrent avoir un effet semblable, mais un peu plus faible, que l'omission d'un terme quadratique.

Quelle est la valeur de γ qu'il convient d'utiliser? Cela dépend clairement de l'utilisation que nous voulons faire du modèle résultant. Si notre principal intérêt est d'utiliser le modèle pour estimer des rapports de cotes à des valeurs de x où la probabilité d'un cas est élevée, et que l'échantillon est suffisamment grand pour que la variance et le biais de petit échantillon soient moins importants, nous pourrions utiliser les poids de population. Pour les tailles d'échantillon plus petites, ou si nous nous intéressons à des valeurs de x plus proche de la moyenne de population, les poids d'échantillon conviendraient mieux. Parfois, une valeur intermédiaire entre les poids de population et les poids d'échantillon pourrait représenter un compromis raisonnable. Par exemple, l'élagage des poids à 10 pour 1 (c'est-à-dire fixer $\gamma \approx 0,91$) dans l'exemple, au lieu de 1 pour 1 (pondération d'échantillon) ou 400 pour 1 (pondération de population) donne une efficacité de 70 % et un biais de petit échantillon de 0,04. Les valeurs correspondantes pour la pondération de population sont 16 % et 0,24. La valeur de $x_{0,91}$ est située presque exactement à mi-chemin entre $x_{0,5}$ et $x_{0,9975}$.

9. Études familiales cas-témoins

Si nous nous intéressons principalement aux paramètres du modèle marginal (1), alors les méthodes dont nous avons discuté aux sections précédentes sont faciles à appliquer et raisonnablement efficaces. L'utilisation de méthodes entièrement efficaces requiert la construction de modèles paramétriques de la dépendance intragrappe et l'effort supplémentaire que cela demande en vaut rarement la peine. Cependant, il existe des situations où la structure de dépendance présente un intérêt intrinsèque. En particulier, il est de plus en plus fréquent que les épidémiologistes généticiens étoffent les données d'une étude cas-témoins standard au moyen d'information sur les réponses et les covariables fournie par des membres de la famille, afin d'essayer d'obtenir des renseignements sur le rôle de la génétique et de l'environnement. Cette approche peut être considérée comme un échantillonnage en grappes stratifié, où les familles sont les grappes et, dans ce cas, la structure intragrappe est de toute première importance. L'exemple qui suit est assez typique.

Exemple 3

Wrensch, Lee, Miike, Newman, Barger, Davis, Wiencke et Neuhaus (1997) ont réalisé une étude cas-témoins sur la population du gliome, forme la plus fréquente de tumeur maligne du cerveau, dans la région de la baie de San Francisco. Ils ont recueilli des renseignements sur tous les cas de gliome diagnostiqués durant un intervalle de temps particulier et sur un échantillon comparable de témoins sélectionnés par la méthode de composition aléatoire. Ils ont également recueilli des renseignements sur la situation de tumeur du cerveau et sur les covariables auprès des membres de la famille des sujets sélectionnés dans l'échantillon cas-témoins original. L'étude portait sur 476 familles comptant un cas de tumeur du cerveau et 462 familles comptant un témoin.

Nous pourrions utiliser les méthodes dont nous venons de discuter pour ajuster un modèle marginal de la probabilité de devenir une victime du gliome, mais les chercheurs s'intéressaient avant tout à l'estimation des caractéristiques intrafamiliales. Une approche consisterait à ajuster un modèle logistique mixte comprenant un ou plusieurs effets familiaux aléatoires.

Notons que, strictement parlant, le plan d'échantillonnage de l'exemple 3 n'est pas compris dans ce plan d'étude cas-témoins. Ici, la stratification est reliée à la variable réponse, mais n'est pas entièrement déterminée par cette dernière. La strate 1 contient les 476 familles dans lesquelles un cas a été diagnostiqué durant un petit intervalle de temps déterminé, tandis que la strate 2 contient les 1 942 490 autres familles, dont certaines comprennent des victimes du cancer du cerveau.

Dans Neuhaus et coll. (2006), nous élaborons des méthodes semi-paramétriques efficaces pour l'échantillonnage stratifié à plusieurs degrés dans des situations où la stratification dépend de la réponse, éventuellement d'une façon non spécifiée qui doit être modélisée, et les observations dans une unité primaire d'échantillonnage sont reliées au moyen d'un modèle paramétrique. Le calcul des estimations requiert la résolution des $p+1$ équations d'estimation, où p est la dimension du vecteur de paramètres. La matrice de covariance peut aussi être estimée facilement en utilisant une analogue de l'inverse de la matrice d'information observée. La procédure complète peut être exécutée à l'aide d'une routine de maximisation raisonnablement générale, mais demande néanmoins une certaine expertise en calcul.

Nous pourrions aussi ajuster les mêmes modèles en utilisant des estimateurs pondérés par les poids de sondage, ce qui offre l'énorme avantage de ne nécessiter aucun logiciel spécialisé. Dans notre exemple, les familles comprenant un cas auraient un poids de 1 et les familles comprenant un témoin auraient un poids de $1\,942\,490/462 \approx 4\,200$. Étant donné cette grande différence, nous pourrions nous attendre à ce que les estimations pondérées soient très inefficaces. Malheureusement, il s'avère presque impossible d'ajuster un modèle intéressant pour lequel les estimations pondérées convergent. L'un des problèmes est que les estimations pondérées sont fondées presque entièrement sur l'échantillon de témoins et que l'on possède fort peu d'information au sujet des effets familiaux dans les familles de témoins (un autre problème est que nous ne possédons pas d'information sur l'âge des membres de la famille et que toute spécification du modèle sans la variable d'âge était exagérément incorrecte). Donc, nous avons dû recourir à une simulation, qui est loin d'être achevée à ce stade. Il semble cependant qu'ici l'efficacité des estimations pondérées soit inférieure à 10 % des estimations par la méthode semi-paramétrique du maximum de vraisemblance. Plus de détails sont donnés dans Neuhaus et coll. (2002, 2006).

Bien que nos simulations en soient à un stade très précoce, il est possible de tirer quelques conclusions provisoires. La principale est que les grandeurs intrafamiliales sont fort mal estimées, même en utilisant des méthodes entièrement efficaces. Les plans d'étude familiale cas-témoins, où l'information sur les membres de la famille est obtenue à titre de supplément à un plan cas-témoins standard, ne fournissent tout simplement pas suffisamment d'information pour estimer les paramètres qui intéressent les épidémiologistes généticiens, à moins que les associations soient extrêmement (voire déraisonnablement) fortes (il convient de souligner que tous les épidémiologistes généticiens ne sont pas d'accord sur ce point). L'utilisation de variantes plus efficaces est néanmoins possible. Ainsi, si nous pouvions identifier les familles contenant plus d'un cas, il serait alors possible d'atteindre une efficacité sensiblement plus grande en suréchantillonnant fortement ces familles. Essentiellement, nous considérerions la famille comme l'unité d'échantillonnage, définirions une « famille-cas » comme contenant plusieurs cas individuels, puis sélectionnerions un échantillon cas-témoins de familles. Il s'agit d'un domaine important où de nombreux travaux restent à accomplir.

10. Conclusion

L'étude cas-témoins sur la population est l'un des domaines où la pratique a devancé la théorie. Autant que je sache, le seul ouvrage où le sujet est abordé en profondeur est celui de Korn et Graubard (1999, chapitre 9). Un aspect auquel a été accordée une attention théorique raisonnablement importante dans la littérature est la stratification. Des méthodes efficaces en vue d'intégrer des variables de stratification dans l'analyse ont été élaborées, entre autres, par Scott et Wild (1997), Breslow et Holubkov (1997), ainsi que Lawless et coll. (1999), dans des circonstances où les variables peuvent prendre uniquement un ensemble fini de valeurs. Breslow et Chatterjee (1999) ont examiné le meilleur moyen d'utiliser ce genre d'information à l'étape de la conception de l'étude. L'extension de tous ces travaux (analyses ainsi que conception) à des situations où nous possédons de l'information sur des variables continues, comme l'âge, pour tous les membres de la population est un domaine où les travaux doivent se poursuivre. Bien que l'échantillonnage à plusieurs degrés soit d'usage répandu, l'effet de la mise en grappes a fait couler nettement moins d'encre. Font exception Graubard et coll. (1989), Fears et Gail (2000), ainsi que Scott et Wild (2001a). Le présent article suscitera peut-être d'autres travaux portant sur ce sujet important. En particulier, puisque le problème se résume essentiellement à l'estimation de deux moyennes de population (voir l'équation (8)), il devrait être possible d'appliquer une grande partie des connaissances sur les plans de sondage efficaces à la résolution de ce problème.

Remerciements

Je tiens à remercier les examinateurs, ainsi que Barry Graubard et Graham Kalton, dont la discussion réfléchie d'une version antérieure du présent article a fait progresser considérablement ma compréhension du sujet. Enfin, j'aimerais remercier tout spécialement mes collaborateurs de longue date Chris Wild, avec lequel ont été réalisés presque tous les travaux qui sous-tendent le présent article, et Jon Rao, auquel je dois essentiellement tout mon savoir sur l'analyse des données d'enquête.

Bibliographie

Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V. et Lennon, D. (2000). Household crowding: A major risk factor for epidemic meningococcal disease in Auckland children. *Pediatric Infectious Disease Journal*, 19, 983-990

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Breslow, N.E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91, 14-28.

Breslow, N.E. (2004). Case-control studies. Dans *Handbook of Epidemiology*. (Éds. W. Aherns et I. Pigeot). New York : Springer. 287-319.

Breslow, N.E., et Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.

Breslow, N.E., et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Applied Statistics*, 48, 457-468.

Breslow, N.E., et Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society*, B, 59, 447-461.

Brogan, D.J., Denniston, M.M., Liff, J.M., Flagg, E.W., Coates, R.J. et Brinton, L.A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, 153, 1119-1127.

Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica*, 49, 1289-1316.

DiGaetano, R., et Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology*, 155, 771-775.

Fears, T.R., et Brown, C.C. (1986). Logistic regression models for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42, 955-960.

Fears, T.R., et Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to nonmelanoma skin cancer. *Biometrics*, 56, 190-198.

Graubard, B.I., Fears, T.R. et Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics*, 45, 1053-1071.

Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. et Waksberg, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.

Hartge, P., Brinton, L.A., Cahill, J.I., West, D., Hauk, M., Austin, D., Silverman, D. et Hoover, R.N. (1984). Design and methods in a multi-center case-control interview study. *American Journal of Public Health*, 74, 52-56.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York : John Wiley & Sons, Inc.

Lawless, J.F., Kalbfleisch, J.D. et Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society*, B, 61, 413-38.

Lee, A.J., Scott, A.J. et Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 95 (A paraître).

Manski, C.F., et McFadden, D. (Éds) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. New York : John Wiley & Sons, Inc.

Miettinen, O.S. (1985). The case-control study: Valid selection of subjects. *American Journal of Epidemiology*, 135, 1042-1050.

Prentice, R.L., et Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Neuhaus, J., Scott, A.J. et Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23-37.

Neuhaus, J., Scott, A.J. et Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics*, 62, sous presse.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

Rao, J.N.K., Scott, A.J. et Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.

Scott, A.J., et Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society*, B, 48, 170-182.

Scott, A.J., et Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 83, 57-72.

Scott, A.J., et Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics*, 50, 57-71.

Scott, A.J., et Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference*, 96, 3-27.

Scott, A.J., et Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *Journal of the Royal Statistical Society*, B, 64, 207-220.

Wacholder, S., McLaughlin, J.K., Silverman, D.T. et Mandel, J.S. (1991). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, 135, 1019-1028.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Waksberg, J. (1998). Random digit dialing sampling for case-control studies. Dans *Encyclopedia of Biostatistics*. (Éds. P.Armitage et T. Colton). New York : John Wiley & Sons, Inc., 3678-3682.

Wrensch, M., Lee, M., Miike, R., Newman, B., Barger, G., Davis, R., Wiencke, J. et Neuhaus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology*, 145, 581-93.