ESTIMATION D'UNE DISTRIBUTION POUR UN DOMAINE

Franck ARNAUD

Insee, Unité Méthodes Statistiques

Etant donné les contraintes budgétaires actuelles pesant sur la réalisation d'enquêtes par sondages, il est probable que nous assistions à un large recours aux techniques spécifiques d'estimation dans les domaines. Ces dernières permettent en effet de tirer le maximum d'information des enquêtes disponibles. Les méthodes d'estimation sur petits domaines portent sur les petites populations, pour lesquelles peu d'information est disponible. Pour pallier ce problème, on a recours à des techniques sortant du cadre de la théorie pure des sondages mais qui allient théorie des sondages et statistique dite « classique ». Il existe une abondante littérature sur l'estimation de paramètres simples, essentiellement des fonctions de totaux. Nous rappellerons cette littérature car elle nous permettra d'introduire les problématiques propres aux domaines, que nous retrouverons ensuite dans nos travaux sur l'estimation de la distribution d'une variable aléatoire dans un domaine. Il existe en effet à notre connaissance peu de travaux sur ce sujet. Certains auteurs se sont en effet penchés sur l'estimation de fonction de répartition, dans un cadre d'enquêtes par sondages, mais hors de toute référence aux domaines. Et leurs techniques se révèlent inadaptées au cas où peu d'information est disponible. Nous verrons comment notre méthode tente de concilier la nécessité de données importantes avec la contrainte pratique de la faible taille du domaine d'étude.

1. Introduction

Après un bref rappel des méthodes classiques de la théorie des sondages, nous aborderons la problématique propre de l'estimation dans les domaines.

1.1 Notation

1.1.1 Echantillonnage

Nous nous intéressons à une population U de taille N. Nous sélectionnons des échantillons s selon un plan de sondage p, une probabilité sur U, fixée.

On note π_k et π_{kl} les probabilités d'inclusion d'ordre 1 et 2 du plan p. On ne considère qu'une seule variable d'intérêt, y, qui prend la valeur y_k pour l'individu $k \in U$ (le terme d'individu est à prendre au sens statistique du terme). C'est une hypothèse simplificatrice importante dans la mesure où, dans

la pratique, une enquête comporte souvent non pas une unique question d'intérêt mais plusieurs. Nous y reviendrons.

1.1.2 Estimation

Pi-estimateur On cherche à estimer, dans un premier temps, des fonctions simples de $\{y_k / k \in U\}$ tels que le total ou la moyenne de y :

$$t_y = \sum_{k \in U} y_k \qquad \bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

Pour ce faire, on peut prendre le π-estimateur :

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

dont on sait qu'il est sans biais, et dont la variance vaut :

$$\boldsymbol{V}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_k \pi_l} y_k y_l$$

Nous disposons bien sûr d'estimations de variance et d'approximations de variance, nous permettant d'avoir des estimateurs de variance utilisables en pratique. C'est très classique quand le plan de sondage n'est pas trop compliqué. Pour les plans complexes de type à plusieurs degrés ou stratifiés dans tous les sens, la difficulté n'est pas beaucoup plus grande que pour les cas simples, car ils se décomposent en plans simples.

Estimateurs avec information auxiliaire Il existe d'autres estimateurs comme celui de Hájek :

$$\tilde{y} = \frac{\sum_{k \in S} \frac{y_k}{\pi_k}}{\sum_{k \in S} \frac{1}{\pi_k}}$$

L'objectif n'étant pas ici de faire un cours de sondage, on ne s'étend pas plus sur les propriétés de cet estimateur. D'autant plus qu'il incorpore que peu d'information auxiliaire, contrairement aux estimateurs par le quotient ou par la régression :

$$\hat{t}_{yQ} = \frac{t_x}{\hat{t}_{x\pi}} \hat{t}_{y\pi} \qquad \hat{t}_{yR} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})' \hat{\mathbf{b}}$$

où $\hat{\mathbf{b}}$ est le coefficient estimé de la régression de y sur l'information auxiliaire x.

L'avantage principal de l'estimateur par la régression est que sa variance s'exprime comme la variance du π-estimateur des résidus de y sur x : autrement dit, $\mathbf{V}(\hat{t}_{yR}) = \mathbf{V}(\hat{t}_{e\pi})$, où $\hat{t}_{e\pi} = \sum_{k \in S} \frac{e_k}{\pi_k}$ et $e_k = y_k - \hat{\mathbf{b}} x_k$. Par exemple, si y est bien lié linéairement à x, alors l'introduction de la variable auxiliaire lors de l'estimation est très intéressante du point de vue de la précision.

Après cette brève introduction et rappel des notions, nous pouvons nous attaquer à la spécificité des domaines.

1.2 Domaines

L' "Estimation dans les domaines" s'intéresse à la connaissance d'un estimateur d'une variable d'intérêt pour une population de petite taille. Cette population peut être délimitée géographiquement, par exemple l'agglomération de Brest, socialement (les utilisateurs de voiture fonctionnant au GPL), ou de toute autre manière, non ambiguë.
Face à une demande spécifique concernant une variable d'intérêt pour une petite sous-population, deux cas de figure sont envisageables :

- Soit une étude récente est disponible sur le sujet, auquel cas il ne reste plus qu'à l'exploiter
- Soit on ne dispose d'absolument aucune étude et il va falloir mettre en place une enquête, ce qui est très coûteux

C'est l'aspect rédhibitoire de ce dernier coût qui est à l'origine de ce champ de la théorie des sondages qu'est l'estimation dans les domaines. Nous nous plaçons en effet dans le cas où une enquête sur le thème d'intérêt ou un thème proche existe. Le plus souvent, l'enquête n'aura pas été calibrée pour la sous-population d'intérêt mais par exemple au niveau France entière. Et c'est précisément l'objet de cette branche de la théorie des sondages qu'est l'estimation dans les domaines de permettre de tirer de l'information pertinente et la plus précise possible de l'information déjà disponible.

Il existe plusieurs méthodes, que nous allons voir dans les sections suivantes. Mais avant cela, nous distinguons deux types de variables d'intérêt différentes. Il existe en effet une sorte de dichotomie importante. Certaines variables sont assez faciles à estimer, telles que les moyennes ou les totaux, les coefficients de corrélation. En réalité, d'ailleurs, ce n'est pas leur estimation qui nous intéresse tant que l'estimation de leur variance ! Et d'un autre coté, il existe certaines caractéristiques plus difficilement accessibles, que nous pouvons qualifier de "non linéaires", au sens où la technique de linéarisation ne leur est pas adaptée : on pense essentiellement aux fonctions de la fonction de répartition, telles que le plus grand intervalle de probabilité α .

Face à ces deux niveaux de difficultés : des contraintes. Pour la première classe, les statistiques simples, on peut espérer réussir avec des méthodes usuelles, éventuellement adaptées à la marge, même quand l'échantillon dans le domaine est de petite taille. Pour les autres statistiques, plus complexes, il sera difficile de faire quoi que ce soit avec un échantillon de taille trop petite. Dans la suite de l'exposé, nous conserverons cette différence entre les deux niveaux de difficultés.

2. Paramètres linéaires

Pour les paramètres linéaires, une abondante littérature existe, qui nous fournit plusieurs solutions pour parvenir d'une part à un estimateur cohérent, d'autre part à une estimation de la variance de cet estimateur.

2.1 Inefficacité de la théorie classique

Supposons que nous souhaitons estimer la moyenne $\bar{y}_d = \frac{1}{N_d}\sum_{k\in U_d} y_k$ d'un critère y dans le domaine d , noté U_d . Un première idée, étant donné l'échantillon s a tiré avec un plan de sondage de probabilités d'inclusion π_k , est de considérer :

$$\hat{y}_{d\pi} = \frac{1}{N_d}\sum_{k\in S_d}\frac{y_k}{\pi_k}$$

Nous savons alors que, sous la condition usuellement vérifiée $\forall k \in U_d$, $\pi_k > 0$, cet estimateur est sans biais pour la moyenne $\bar{y}_d$. Son approximation de variance, pour un sondage aléatoire simple au taux $f = n/N$, est $\mathbf{V}_{SAS}(\hat{y}_{d\pi}) = N_d^2 \left[\frac{1-f}{n} P_d (S_{yd}^2 + Q_d \bar{y}_d^2) \right]$ où $S_{yd}^2 = \frac{1}{N_d - 1} \sum_{k \in U_d} (y_k - \bar{y}_d)^2$, $P_d = N_d / N$ et $Q_d = 1 - P_d$. Autrement dit, on retrouve une expression attendue, $N_d^2 \frac{1-f}{n} P_d S_{yd}^2$, à un facteur près. La raison est simple $n_d = \#S_d = \#S \cap U_d$ est aléatoire, et cette inconnue rajoute de la variance. Pour avoir un estimateur de meilleure qualité, on peut utiliser l'estimateur de Hájek :

$$\tilde{y}_d = \frac{\sum_{k \in S_d} \frac{y_k}{\pi_k}}{\sum_{k \in S_d} \frac{1}{\pi_k}}$$

Néanmoins, ces estimateurs se heurtent à une autre contrainte, plus fondamentale : S_d peut être très petit. Pensons que si le domaine est de petite taille, S_d peut avoir seulement quelques observations : il est hors de question de faire des proportions avec 5 ou 10 observations ! Nous allons voir dans les sections suivantes deux remèdes possibles à ce problème.

En ce qui concerne l'estimation dans les domaines, la première approche, historiquement, est basée sur un modèle. Ce n'est que dans les années 80 qu'est apparue l'approche dite assistée par un modèle.

2.2 Un peu d'histoire

Au début des années 70, devant l'incapacité de la théorie des sondages à donner des résultats très puissants du type "meilleur estimateur", une nouvelle approche a émergé. Elle consistait à quitter le champ de la théorie des sondages classique, où l'aléatoire porte sur l'échantillon tiré, pour revenir dans le champ de la statistique classique, ie en considérant l'échantillon tiré selon un plan déterministe, dégénéré, où l'aléatoire vient donc des observations. Autrement dit, dans $\{y_k / k \in U\}$, la théorie des sondages postule que les y_k sont fixés conditionnellement à S qui est aléatoire. Et dans la statistique classique, on considère que S est fixé, et que ce sont les y_k qui sont aléatoires.

L'approche basée sur un modèle s'est développé autour de Royall dans les années 1970. Un des premiers modèles considérés est très simple : $y_k = x_k \beta + e_k$, où les e_k sont les iid centrées et de variance proportionnelle à x_k^2. On en déduit un estimateur optimal en un certain sens de la moyenne de y_k.

Hormis leur différence de définition, quelles sont les différences entre ces deux approches ? Pour les identifier, on raisonne dans un espace regroupant les deux notions, c'est-à-dire avec deux sources d'aléa : d'une part le plan, comme dans les sondages classiques, et d'autre part le modèle, comme pour les statisticiens universitaires classiques.

Dans l'approche basée sur un modèle, on cherche des estimateurs qui ont de bonnes propriétés sous le modèle. Auquel cas on ne sait rien des propriétés de l'estimateur sous le plan. Et toute la qualité de notre estimateur final dépend de la validité du modèle : à modèle faux, estimateur faux. Ce n'est pas très robuste. Dans l'approche assistée par un modèle, c'est le plan qui préside à la construction de l'estimateur, et donc l'estimateur n'est pas trop mauvais au sens où on l'entend "naturellement" : si on retire l'échantillon un grand nombre de fois selon le même plan de sondage, la moyenne des estimations finira par converger vers "la vraie valeur".

Le problème du statisticien public apparaît alors clairement: pour utiliser un modèle, il faut qu'il soit en adéquation avec la réalité. C'est très subjectif, d'autant plus que des hypothèses du modèle découle la forme de l'estimateur. Les hypothèses faites ne sont absolument pas sans effets sur le résultat final. Dans ces conditions, étant donné la nécessaire objectivité des statisticiens publics, il est difficile de souscrire à cette approche. A moins que le modèle n'ait été validé de toute part, ce qui n'est pas la norme. Nous verrons néanmoins qu'il est difficile de nous en passer, parfois, et qu'il peut être d'une précieuse aide.

2.3 Estimation assistée par un modèle

Nous avons vu qu'historiquement ce n'était pas la première méthode. Nous la plaçons toutefois en tête en raison de sa simplicité et de sa robustesse.

Nous avons déjà considéré les versions domaines des π-estimateurs et estimateurs de Hájek. Nous pouvons de même exhiber des versions domaines des estimateurs par le quotient ou la régression. Mais tous butent sur le même problème : ils reposent sur $\{y_k / k \in S_d\}$, qui est trop petit. Les estimateurs dont les seuls y_k utilisés appartiennent au domaine sont dits directs. La conclusion logique de cette définition est : les estimateurs directs sont inutiles pour les estimations dans les domaines. L'estimation synthétique permet de dépasser ce paradoxe.

2.3.1 Estimateur synthétique

Une première idée, dont la portée est limitée, pour estimer la moyenne d'une variable d'intérêt dans un domaine consisterait à prendre la moyenne de la variable d'intérêt pour la population de l'étude dont nous disposons, typiquement la France entière, en considérant que ces deux chiffres sont proches. Par exemple, si le maire de Brest nous demande d'estimer son taux de chômage, nous pouvons toujours considérer qu'après tout, la ville de Brest est proche de la population française en terme de taux de chômage, et proposer comme estimation du taux de chômage de Brest le taux pour la France entière. On se rend bien compte que cette estimation est très grossière : on risque de faire une grosse erreur. Mais nous sommes sur la bonne voie. Nous gagnerons certainement en utilisant de manière efficace de l'information auxiliaire, comme les tranches d'âge ! Et en effet, une hypothèse plus robuste est que le taux de chômage, pour chaque tranche d'âge, est relativement identique pour toute la France et pour Brest. Avec cette hypothèse, nous pouvons construire un bien meilleur estimateur, dit synthétique par opposition à direct : un estimateur synthétique fait intervenir de l'information non seulement du domaine mais aussi des autres domaines.

Cette idée générale, remplacer une certaine quantité au niveau du domaine par son équivalent au niveau d'une population plus vaste, est la clé de l'estimation dite synthétique, par opposition à l'estimation directe.

2.3.2 Application

Voyons concrètement ce que donne ce principe d'estimation synthétique. Pour cela, nous allons introduire la notion de groupe : c'est le niveau auquel une bonne information auxiliaire est disponible. Dans notre exemple brestois, le groupe est une tranche d'âge. Les groupes sont indicés par $g = 1,...,G$ et notés $U_1,...,U_g,...,U_G$ (ou $U_{.g}$ s'il y a un doute). L'intersection du domaine d et du groupe g est U_{dg}. On note :

$$N_d = \#U_d \qquad N_g = N_{.g} = \#U_g = \#U_{.g} \qquad N_{dg} = \#U_{dg} \qquad N = \sum_{g=1}^{G} N_{.g} = \sum_{d=1}^{D} N_d = \sum_{d=1}^{D}\sum_{g=1}^{G} N_{dg}$$

De même pour l'échantillon :

$$s_{dg} = U_{dg} \cap s \qquad s_d = U \cap s \qquad s_{\dot{g}} = U_{.g} \cap s \quad \text{de cardinaux} \quad n_{dg}, n_d, n_{.g} = n_g$$

Avec ces notations, nous allons adapter nos anciens estimateurs. Initialement, nous utilisions les π-estimateurs du total $t_y = \sum_{k\in U} y_k$ qui était $\hat{t}_{y\pi} = \sum_{k\in S} \frac{y_k}{\pi_k}$. Maintenant, comme $t_y = \sum_{g=1}^{G} t_{yg}$, où $t_{yg} = \sum_{k\in U_g} y_k$, nous avons notre nouvel estimateur $\hat{t}_{y\pi}^{G} = \sum_{g=1}^{G}\sum_{k\in S_{dg}} \frac{y_k}{\pi_k}$. On adapte de même les autres estimateurs :

	Version groupe
$\hat{t}_{y,1} = \sum_{k\in S_d} w_k(S) y_k$	$\hat{t}_{y,1}^{(G)} = \sum_{g=1}^{G}\sum_{k\in S_{dg}} w_{kS} y_k$
$\hat{t}_{y,2} = N_d \frac{\sum_{k\in S_d} w_k(S) y_k}{\sum_{k\in S_d} w_k(S)}$	$\hat{t}_{y,2}^{(G)} = \sum_{g=1}^{G} N_{dg} \frac{\sum_{k\in S_{dg}} w_k(S) y_k}{\sum_{k\in S_{dg}} w_k(S)}$

Mais cette notion de groupe ne résout pas notre problème, bien au contraire, puisque nous avons découpé notre domaine U_d en plus petits sous-domaines encore, avec les groupes. C'est précisément là qu'intervient l'estimation synthétique, en supposant que :

$$\forall d, \quad \forall g, \quad S_{dg} \approx S_g$$

dans un sens que nous allons préciser maintenant. Notre hypothèse de proximité des taux de chômage par tranche d'âge entre Brest et la France entière peut en effet se généraliser sous la forme suivante :

$$\forall d, \quad \frac{\sum_{k\in S_{dg}} w_k(S) y_k}{\sum_{k\in S_{dg}} w_k(S) x_k} \approx \frac{\sum_{k\in S_g} w_k(S) y_k}{\sum_{k\in S_g} w_k(S) x_k}$$

Appliqué à l'estimateur de Hájek de la moyenne de y, $\bar{y}_d = N_d^{-1}\sum_{k\in U_d} y_k$, nous obtenons l'estimateur synthétique suivant

$$\tilde{\bar{y}}_{d,SYN}^{(G)} = \frac{1}{N_d}\sum_{g=1}^{G} N_{dg} \frac{\sum_{k\in S_g} \frac{y_k}{\pi_k}}{\sum_{k\in S_{gg}} \frac{1}{\pi_k}}$$

Les estimations de variance sont fastidieuses mais tout à fait possibles, avec la méthode usuelle, qui consistent à trouver une approximation de variance (linéarisation,…), puis à en estimer les paramètres inconnus. L'avantage de cette technique d'estimation non pas basée mais assistée par un modèle est que les performances des estimateurs, évalués selon le plan de sondage, ne dépendent pas crucialement

des hypothèses. Pour être plus précis, l'estimateur est toujours sans biais, mais il est de variance d'autant plus faible que les hypothèses sont justes.

2.3.3 Estimateurs composites

Néanmoins, ces estimateurs ne tiennent que peu compte des particularités locales du domaine. Si les arsenaux de Brest ont fermé, notre estimateur du taux de chômage ne le fera pas apparaître. Pour améliorer l'adéquation à la réalité, on peut penser trouver un juste milieu entre un estimateur direct de qualité douteuse et un estimateur synthétique, fiable, mais pour estimer une quantité dont on espère qu'elle est proche de ce qui nous intéresse. Un compromis entre ces deux principes d'estimation, direct et synthétique, consiste donc à prendre :

$$\hat{t}_{d,COMP} = \gamma_d \hat{t}_{d,SYN} + (1-\gamma_d)\hat{t}_{d,DIR}$$

Le paramètre γ_d est alors un réel dans [0 ; 1], à préciser. Nous voudrions bien sûr qu'il soit optimal en un certain sens. Nous nous heurtons alors encore une fois à un problème de subjectivité : pour résoudre ce problème, il faut considérer un modèle. C'est un point clé de l'estimation dans les domaines : étant donné la faible information disponible, le modèle est la solution pour prétendre à des résultats fiables. Nous devons trouver une liaison entre les différents domaines, au regard de la variable d'intérêt, qui nous permettent de donner une bonne estimation.

2.4 Estimation basée sur un modèle

On suppose alors les $\{y_k / k \in U\}$ aléatoires : on note ξ la loi jointe de $\{y_k / k \in U\}$. Nous disposons donc de deux sources d'aléatoires : le plan p et le modèle ξ. Les espérances, variances et écarts quadratiques moyens seront suffixés par p pour le plan et ξ pour le modèle. Une hypothèse importante est que le modèle et le plan sont indépendants, on parle alors de plans non informatifs, ou neutres[1]. Dans cette optique, t_y et $\bar{y}$ sont des variables aléatoires. Nous ne parlons donc plus d'estimateurs de ces quantités, mais de prédicteurs. Ou alors d'estimateurs de $\mathbf{E}t_y$ ou $\mathbf{E}\bar{y}$.

Pour l'estimation d'une variable d'intérêt du type moyenne, nous présentons deux approches principales : une bayésienne et une reposant sur l'emploi de modèles mixtes.

2.4.1 Méthodes bayésiennes

L'article séminal est celui de Fay et Herriot de 1979, dans lequel ils présentent leur estimateur composite. L'intérêt principal de l'article vient de ce qu'il est le premier à poser un modèle de lien entre les différents domaines, dont se déduit naturellement le coefficient de moyenne γ_d de l'estimateur composite.

L'approche de Fay et Herriot Ils considèrent le modèle :

$$\forall d = 1,...,D, \forall k \in S_d, \quad y_{dk} \,|\, \theta_d \sim \mathcal{N}_{ind}(\theta_d, \sigma_d^2) \quad \text{et} \quad \boldsymbol{\theta} = (\theta_1,...,\theta_D)' \sim N(X\boldsymbol{\beta}, \tau^2 Id_D)$$

[1] C'est le cas le plus souvent, mais pas toujours : part exemple si l'on réalise deux sondages, et que les individus tirés au second tour dépendent des variables récoltées au premier passage. Une autre hypothèse souvent faite en pratique est celle de l'indépendance de y_k.

où les y_{dk} sont indépendants deux à deux (entre domaines et au sein d'un même domaine). La matrice $X = (\mathbf{x}_1^{'},...,\mathbf{x}_D^{'})^{'}$ de taille $D \times p$ $(D > p)$ et de rang p contient toute l'information auxiliaire, que l'on suppose connue au niveau du domaine. De même, les σ_d^2 sont connus, tandis que $\boldsymbol{\beta} \in \mathbb{R}^p$ et τ^2 sont à estimer. Les paramètres $\boldsymbol{\beta}$ et τ^2 sont des hyperparamètres du modèle bayésien.

Du modèle découle la forme de l'estimateur bayésien :

$$\forall d = 1,...,Q \quad \delta_d^B = \frac{\sigma_d^2}{\sigma_d^2 + n_d \tau^2} \bar{\mathbf{x}}_d{}' + \frac{n_d \tau^2}{\sigma_d^2 + n_d \tau^2} \bar{y}_d$$

Dans cette expression, il reste deux paramètres inconnus, β et τ^2, à estimer. Fay et Herriot utilisent la méthode ANOVA pour τ^2. Et pour β, on prend le BLUP calculé avec le τ^2 juste obtenu. L'estimateur que nous obtenons finalement est composite : c'est une moyenne entre l'estimateur direct $\bar{y}_{Sd}$ et l'estimateur synthétique $x'_d \beta$. Le poids accordé aux deux composantes dépend de l'ajustement du modèle de régression : l'estimateur synthétique a un poids d'autant plus important que le modèle ajuste bien les données. C'est une idée très importante, que nous réutiliserons pour l'estimation de la fonction de répartition.

Ghosh et Meeden Ghosh et Meeden suivent une approche similaire à Fay et Herriot, à cela près qu'ils tiennent compte du plan de sondage. Implicitement, Fay et Herriot modélisait l'échantillon, s, à la manière des statisticiens "classiques", sans tenir compte de la particularité de la sélection. C'est ce qu'essaient de corriger Ghosh et Meeden, en considèrant le modèle suivant :

$$\forall d = 1,..., D, \forall k \in U_d, \quad y_{dk} \,|\, \theta_d \sim N_{ind}(\theta_d, \sigma^2) \qquad \boldsymbol{\theta} = (\theta_1,...,\theta_D)^{'} \sim N(\mu \mathbf{1}_D, \tau^2 Id_D)$$

Dans ces deux lois, σ^2, τ^2 et μ sont inconnus.
L'estimateur bayésien, pour ce modèle, est :

$$\mathrm{E}(\bar{y}_d \mid \mathrm{y}_{S_d}) = \frac{1}{N_d} \{ n_d \bar{y}_{sd} + (B_d \mu + (1 - B_d) \bar{y}_{sd} (N_d - n_d)) \} \qquad (3)$$

où $B_d = M/(M + n_d)$ avec $M = \sigma^2 / \tau^2$.Comme pour l'estimateur de Fay et Herriot, nous devons maintenant pallier l'ignorance de σ^2, τ^2 et μ. Pour cela, ils proposent eux-aussi des estimateurs de type ANOVA pour σ^2 et τ^2. Et ils en déduisent une estimation de μ, qui n'est en fait qu'un BLUP (mais qui dépend des composants de variance, estimés).

L'approche de Ghosh et Meeden améliore le modèle de Fay et Herriot, en cela qu'il tient plus compte du plan de sondage. Ces deux modèles font appel à la théorie bayésienne. Mais nous pouvons aussi les réinterpréter en termes de modèles mixtes.

2.4.2 Modèles mixtes

Il n'est pas question de faire ici un exposé sur les modèles mixtes. Nous nous contentons de les présenter brièvement.

Présentation Les modèles mixtes sont l'extension du modèle linéaire lorsque non seulement l'espérance de la variable d'intérêt est linéaire en un paramètre inconnu, β, mais aussi sa variance :

$$\exists \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{\theta} \in \mathbb{R}^r \ / \mathrm{E(y)} = X\boldsymbol{\theta} \qquad \mathbf{V(y)} = \sum_{j=1}^{r} \theta_j F_j$$

Généralement, on écrit un modèle mixte sous la forme :

$$y = X\boldsymbol{\beta} + Z_1\mathbf{v}_1 + ... + Z_m\mathbf{v}_m + \varepsilon \tag{4}$$

où $y \in \mathbb{R}^n$ est un vecteur des $\{y_k / k \in S\}$, $X : n \times p$ est la matrice des covariables[2], elle est connue, $\boldsymbol{\beta}$ est un paramètre de $\mathbb{R}^p$ inconnu. Les $\mathbf{v}_i \in \mathbb{R}^{q_i}$ $(i = 1,...,q)$ et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ sont des erreurs aléatoires, que l'on suppose non corrélées. Avec $Z = (Z_1,...,Z_m)$ et $\mathbf{v} = (\mathbf{v}'_1,...,\mathbf{v}'_m)'$, le modèle s'écrit $\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{v} + \varepsilon$. Nous supposerons dans la suite que les $\mathbf{v}_i$ sont non corrélés deux à deux.

On note $R = \mathbf{V}(\boldsymbol{\varepsilon}) : n \times n$ et $G_i = \mathbf{V}(\mathbf{v_i}) : q_i \times q_i$. Auquel cas ($q = \sum_{i=1}^{m} q_i$) :

$$G = \mathrm{V(v)} = \mathrm{diag}(G_i) : q \times q \quad \text{et} \quad V = \mathbf{V(y)} = ZGZ' + R = \sum_{i=1}^{q} Z_i G_i Z_i^{'} + R$$

La seconde hypothèse des modèles mixtes est qu'il existe $\boldsymbol{\theta} = (\theta_1,...,\theta_r)'$ tels que :

$$V = \sum_{j=1}^{r} \theta_j F_j$$

où les F_j sont des matrices connues.

Dans la plupart des cas, les G_i et R sont des matrices proportionnelles à l'identité, à différents niveaux de hiérarchies, comme nous le verrons par la suite. Dans ce cas, les composants de variance $\boldsymbol{\theta}$ sont les variances de chacun des termes.

Il existe plusieurs méthodes pour estimer les paramètres linéaires et les composants de variance, nous ne les détaillerons pas ici. Le lecteur intéressé trouvera toutes les précisions dans le livre de Searle, McCulloch et Casella "Variance Components", chez Wiley.

Prévision Nous souhaitons, par la suite, prévoir une variable aléatoire de la forme $t = \boldsymbol{\lambda}'\boldsymbol{\beta} + \boldsymbol{\mu}'\mathbf{v} \in \mathbb{R}$, où $\boldsymbol{\lambda} \in \mathbb{R}^p$ et $\boldsymbol{\mu} \in \mathbb{R}^q$ sont connus. Pour cela, nous procédons en deux étapes :

- Supposons que les composants de variance $\boldsymbol{\theta}$ soient connus, alors nous trouvons le BLUP de t, ou plutôt, le pseudo-BLUP, noté $\hat{t}(\theta)$. Notons qu'one ne trouve de BLUP de t qu'à l'expresse condition que $\lambda'\boldsymbol{\beta}$ soit estimable.
- Nous estimons ensuite les composantes $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$, et nous prenons pour prédicteur de t $\tilde{t} = \hat{t}(\hat{\boldsymbol{\theta}})$.

L'estimateur final $\tilde{t}(\hat{\boldsymbol{\theta}})$ n'est bien sûr plus le BLUP de t, on l'appelle néanmoins le BLUP empirique. Il est important, et difficile, d'évaluer les performances de ces deux estimateurs. Nous renvoyons notamment à [18], [24], [25] pour les précisions.

Nous donnons en illustration quelques modèles, pour voir les optiques choisies par les auteurs

Littérature

[2] Nous ne précisons rien de plus : si elle est de plein rang, alors c'est une régression. Sinon, c'est une analyse de variance ou de covariance, et on prend une inverse généralisée ou des contraintes.

Fay et Herriot, 1979 Nous avons déjà introduit ce modèle. Les auteurs ne l'avaient pas introduit comme un modèle mixte, mais comme un exemple d'approche Bayes empirique. Nous l'écrivons toutefois ici sous une forme de modèle mixtes :

$$\forall d\ \forall k \in S_d, \quad y_{dk} = \theta_d + \varepsilon_{dk} \quad \text{où} \quad \theta_d = \mathbf{x}'_d\,\boldsymbol{\beta} + v_d \Rightarrow y_{dk} = \mathbf{x}'_d\,\boldsymbol{\beta} + v_d + \varepsilon_{dk}$$

où les deux erreurs $\varepsilon_{dk} \sim (0, \sigma^2)$ et $v_d \sim (0, \sigma^2)$ sont non corrélées. Ils supposent les σ_d^2 connus, auquel cas il ne resterait plus à estimer que le coefficient de régression $\boldsymbol{\beta}$ et le composant de variance τ^2.

Dempster, Rubin et Tsutakawa, 1981 Leur modélisation est hiérarchique, ou en plusieurs niveaux (les multivel models de Golstein), en ce sens où l'on résonne par "niveaux". En premier lieu, au niveau individuel $\forall d\ \forall k \in S_d, \quad y_{dk} = \mathbf{x}'_{dk}\,\boldsymbol{\beta}_d + \varepsilon_{dk}$ .Puis au niveau du domaine : $\boldsymbol{\beta}_d = \boldsymbol{\beta} + \mathbf{v}_d$
On arrive donc au modèle final :

$$\forall d\ \forall k \in S_d, \quad y_{dk} = \mathbf{x}'_{dk}\,\boldsymbol{\beta} + \mathbf{x}'_{dk}\,\mathbf{v}_d + \varepsilon_{dk}$$

Battese, Harter et Fuler, 1988 C'est le premier modèle mixte présenté comme tel :

$$\forall d\ \forall k \in S_d, \quad y_{dk} = \mathbf{x}'_{dk}\,\boldsymbol{\beta} + v_d + \varepsilon_{dk}$$

Prasad et Rao, 1990 Ils prennent le modèle de Battese, Harter et Fuller, en travaillant, comme Kott, non pas sur un estimateur de modèle, mais sur un estimateur de plan (l'estimateur de Hájek).

Approche modèle et plan Kott avait déjà proposé cette solution qui consistait, à partir du modèle, à déterminer, dans une classe d'estimateurs de plan, l'estimateur optimal sous le modèle. Mais ses résultats ne semblaient pas convaincants. En 1999, Prasad et Rao [30] ont repris cette idée pour le modèle simple, en s'intéressant à l'estimateur de Hájek.

Nous avons déjà considéré le modèle :

$$\forall d = 1,..., D, \quad \forall k \in S_d, \quad y_{dk} = \mathbf{x}'_{dk}\,\boldsymbol{\beta} + v_d + \varepsilon_{dk} \tag{9}$$

où les v_d sont iid $(0, \tau^2)$, indépendant des ε_{dk} iid $(0, \sigma^2)$. La matrice $\boldsymbol{X} : n \times p$ est connue, $\boldsymbol{\beta}$ ne l'est pas, nous devons l'estimer, de même que les composants de variance $\boldsymbol{\theta} = (\sigma^2\ \tau^2)'$. Dans la section précédente, nous travaillions directement sur ce modèle. Mais ici, nous nous restreignons à une certaine classe d'estimateurs : considérant les poids de sondage π_k, nous réduisons le modèle (9). Soit en effet $\bar{y}_d = \sum_{k \in S_d} y_k / \pi_k / \sum_{l \in S_d} 1/\pi_l$ l'estimateur de Hájek de $\bar{y}$, alors nous considérons le modèle :

$$\forall d = 1,..., D, \quad \tilde{y}_d = \tilde{\mathbf{x}}_d{}'\boldsymbol{\beta} + v_d + \tilde{\varepsilon}_d$$

où les variables $\tilde{\varepsilon}_d$ sont indépendantes, centrées, de variance $\sigma_d^2 = \sigma^2 \sum_{k \in S_d} w_k^2$ où $w_k = \pi_k^{-1} / \sum_{l \in S_d} \pi_l^{-1}$.

Nous connaissons la forme du speudo-BLUP de $\tilde{\mathbf{x}}'_d\boldsymbol{\beta} + v_d$:

$$\hat{t} = \tilde{\mathbf{x}}'_d\,\hat{\boldsymbol{\beta}} + (1 - \gamma_d)(\tilde{y}_d - \tilde{\mathbf{x}}'_d\,\hat{\boldsymbol{\beta}})$$

où $\gamma_d = \dfrac{\sigma^2}{\sigma_d^2 + \tau^2}$ et $\hat{\boldsymbol{\beta}}$ est l'estimateur GLS de $\boldsymbol{\beta}$. Comme précédemment, on estime ensuite les composants de variance, et on remplace dans la speudo-BLUP $\hat{\sigma}^2$ et $\hat{\tau}^2$: on obtient ainsi l'estimateur BLUP speudo-empirique. Dont nous devons approcher puis estimer la précision.

En procédant ainsi, on espère se protéger contre une mauvaise adéquation du modèle aux données. D'un point de vue historique, cette préoccupation est récente. En statistique d'enquêtes, la robustesse est la capacité d'estimateurs de modèle à résister à de mauvaises spécifications du modèle. L'estimateur par régression classique est robuste, dans ce sens, puisqu'il est approximativement sans biais, sous le plan, et consistent, sous le plan, sous de simples conditions sur le plan de sondage (et plus spécifiquement sur les probabilités d'inclusion aux premiers ordres) et sur la population (typiquement, convergence des deux premiers moments). Pour plus de détails, [22], [26], [4],[40].

En conclusion, il semblerait que les modèles mixtes sont un outil redoutablement adapté au traitement des domaines. D'autant plus que nous avons réussi à intégrer l'information du plan de sondage et l'information auxiliaire avec la modélisation mixte. Cette approche a connue un très fort développement récemment, et pas seulement pour l'estimation dans les domaines, mais aussi pour d'autres problèmes liés, tels que les mesures répétées dans le temps.

Néanmoins, l'estimation des fonctions dites linéaires est d'un ordre de difficulté différent de l'estimation des fonctions non linéaires pour un domaine. La contrepartie est que l'information obtenue est beaucoup plus riche : connaître la distribution est beaucoup plus intéressante que connaître uniquement la moyenne, par exemple.

3. Paramètres non linéaires

Nous nous intéressons donc ici à l'estimation de la fonction de répartition d'une variable aléatoire. Nous procédons en deux étapes : dans la première, nous présentons la méthode de Chambers et Dunstan d'estimation de la fonction de répartition, qui est valide dès que la taille du domaine n'est pas trop réduite, ainsi que celle de Rao, Kovar et Mantel. Ces deux méthodes ne sont pas propres aux domaines, mais nous croyons utiles de les citer sans développer complètement la méthodologie des auteurs pour plusieurs raisons. En premier lieu, pour rappeler l'opposition modèle et plan. Mais aussi pour montrer l'adaptation nécessaire des techniques pour l'estimation dans un domaine, qui est, dans le cas de l'estimation de la distribution, bien plus importante que dans le cas d'estimation de moyenne, où nous avons vu que, si les techniques différaient, elles étaient sur le fond relativement similaires, mais seulement adaptées aux domaines. Autrement dit, c'est dans le cas le plus complexe que se révèle toute la difficulté et la spécificité de l'estimation dans les domaines. Nous proposons alors notre modélisation, développée avec Jean-Claude Deville au Laboratoire de Statistiques d'Enquêtes du Centre de Recherche en Economie et Statistiques de l'Insee, qui permet de s'affranchir de la contrainte sur la taille de l'échantillon. Mais avant cela, nous commençons par présenter le problème, pour expliciter sa spécificité.

3.1 Présentation

La fonction de répartition de la variable d'intérêt y dans la population U est :

$$\forall x \in \mathbb{R}, \quad F_U(x) = \frac{1}{N} \sum_{k \in U} \mathbf{1}_{\{y_k \leq x\}}$$

$F_U(x)$ n'étant est rien d'autre que la moyenne de la variable $\mathbf{1}_{\{y_k \leq x\}}$, nous pouvons utiliser l'estimateur de Hájek, dont on sait qu'il est meilleur pour estimer les proportions (ce qui est le cas ici : $F_U(x)$ est la proportion d'individus ayant y_k inférieur à x) :

$$\forall x \in \mathbb{R}, \quad \widetilde{F}(x) = \frac{\sum_{k \in S} \frac{1}{\pi_k} \mathbf{1}_{\{y_k \leq x\}}}{\sum_{k \in S} \frac{1}{\pi_k}}$$

Une approche très similaire consiste à utiliser un noyau au lieu de l'indicatrice. Mais on gagne peu avec un tel estimateur, par rapport à $\tilde{\tilde{F}}$. D'autant que se pose encore le problème du choix de la fenêtre (on sait que le problème du choix de noyau est négligeable, en pratique).

$\tilde{\tilde{F}}$ est donc le meilleur estimateur que nous pouvons obtenir sans utiliser ni information auxiliaire ni modèle de superpopulation. Une étape supplémentaire a été proposée par Chambers et Dunstan, qui suivent une approche basée sur un modèle que nous étudierons dans la section suivante.

Rao, Kovar et Mantel ont proposé plusieurs estimateurs de plan, en considérant que, puisque $F(t)$ n'est rien d'autre qu'une moyenne, on peut l'estimer plus efficacement en introduisant de l'information auxiliaire, via un estimateur par la régression.

3.2 Estimateur de Chambers et Dunstan

3.2.1 Présentation

Leur idée [6] est la suivante : ils commencent par décomposer la fonction entre ce qui est connu et ce qui ne l'est pas ($\bar{S}$ est $U \setminus s$) :

$$F_U(t) = \frac{1}{N}\left[\sum_{k\in S} 1_{\{y_k \le x\}} + \sum_{k\in \bar{S}} 1_{\{y_k \le x\}}\right]$$

Dans cette expression, une fois récolté l'échantillon, le premier terme est connu. Il ne nous reste plus que le second à estimer. Ils font pour cela l'hypothèse que les variables d'intérêts sont liés à des variables auxiliaires selon un modèle de régression simple $\forall k \in U, \quad y_k = x_k \boldsymbol{\beta} + v(x_k)u_k$. Ils dérivent alors leur estimateur en approximant le second morceau de $F_U(t)$ en respectant le modèle de superpopulation. Pour les détails, nous renvoyons les lecteurs intéressés à [6].
L'estimateur de F_U est alors :

$$\forall x \in \mathbb{R}, \qquad \hat{F}_{CD}(x) = \frac{1}{N}\left[\sum_{k\in S} \mathbf{1}_{\{y_k \le t\}} + \sum_{k\in S}\sum_{l\in S} w_l \mathbf{1}_{\left\{\frac{y_l - x_l\hat{\beta}}{v(x_l)} \le \frac{t - x_k\hat{\beta}}{v(x_k)}\right\}}\right]$$

où $\hat{\boldsymbol{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ est l'estimateur GLS classique (le choix de la pondération, le choix de $\mathbf{V}$, n'est pas très important : ce n'est pas ce qui compte au premier ordre, cf [35])
Pour ce qui est de ses performances, on montre facilement que si les y_k sont exactement égaux aux $x_k\boldsymbol{\beta}$ (et si la somme des w_k vaut 1), alors $\hat{F}_{CD} \equiv F_U$. Les propriétés à distance finie de l'estimateur ne sont pas évidentes pour le modèle linéaire, on a donc peu d'espoir pour la formulation générale, sans spécification particulière du modèle de superpopulation autre que les hypothèses formulées par Chambers et Dunstan.

Néanmoins, l'estimateur CD repose sur un modèle, ce qui est assez sensible, peu robuste. Rao, Kovar et Mantel ont proposé des estimateurs de plan, dont nous avons vu qu'ils étaient préférables, dans un premier temps.

3.3 Estimateurs de Rao, Kovar et Mantel

Les auteurs reprennent une remarque de Chambers et Dunstan : "it is not obvious how either the radio or regression estimation concept can be sensibly extended to estimation of $F(t)$ under a design-based

approach". Comme nous l'avons déjà precise, pour chaque $t \in \mathbb{R}$, $F(t)$ n'est rien d'autre que la moyenne de la variable d'intérêt $\mathbf{1}_{\{y_k \le t\}}$. Par conséquent, pour peu que l'on dispose d'une information auxiliaire, on peut l'utiliser pour améliorer l'estimation de F, à l'aide de techniques classiques.

Plaçons-nous en effet dans le modèle de régression considéré par Chambers et Dunstan : $y_k = x_k^{'}\beta + \sigma_k^2 u_k$, où les u_k sont iid $(0,\sigma^2)$ (souvent $\sigma_k^2 = v(x_k)$). Nous pouvons alors incorporer l'information auxiliaire $1_{\{x_k^{'}\hat{\beta} \le t\}}$ dans l'estimation de F. Nous disposons alors de l'estimateur par le quotient (ratio) $\hat{F}_q(t)$, défini par :

$$\hat{F}_q(t) = \frac{1}{N}\sum_{k\in U} \mathbf{1}_{\{x_k^{'}\hat{\beta} \le t\}} \frac{\sum_{k\in S} \frac{1}{\pi_k}\mathbf{1}_{\{y_k \le t\}}}{\sum_{k\in S} \frac{1}{\pi_k}\mathbf{1}_{\{x_k^{'}\hat{\beta} \le t\}}}$$

et de l'estimateur par régression $\hat{F}_r(t)$:

$$\hat{F}_r(t) = \frac{1}{N}\left[\sum_{k\in S}\frac{1}{\pi_k}1_{\{y_k \le t\}} + \left(\sum_{k\in U} 1_{\{x_k^{'}\hat{\beta} \le t\}} - \sum_{k\in S}\frac{1}{\pi_k}1_{\{x_k^{'}\hat{\beta} \le t\}}\right)\right]$$

Dans cette expression, $\hat{\beta}$ est l'estimateur π-BLUP : $\hat{\beta} = \left(\sum_{k\in S}\frac{x_k x_k^{'}}{\pi_k \sigma_k^2}\right)^{-1}\sum_{k\in S}\frac{x_k y_k}{\pi_k \sigma_k^2}$. On peut bien sûr prendre le BLUP (les résultats sont similaires).

Comme toujours, nous souhaitons éviter le recours explicite au modèle, ou tout du moins, nous espérons que nos estimations ont de bonnes propriétés indépendamment du modèle. C'est le cas, puisque, sous le plan , nos estimateurs $\hat{F}_q(t)$ et $\hat{F}_r(t)$ sont sans biais (approximativement pour le premier). Pour continuer avec les propriétés simples, nous pouvons remarquer que $\hat{F}_d$ et $\hat{F}_q$ sont calés sur la variable auxiliaire : si y_k est réellement égal à $x_k^{'}\boldsymbol{\beta}$, alors $\hat{F}_d, \hat{F}_q \equiv F$.

Si nous espérons que nos estimateurs sont bons indépendamment du modèle, nous pouvons néanmoins nous poser la question de leur performance lorsque le modèle est bon. Et là, nous avons un problème, car même si $\mathbf{E}_m(y_k) = x_k^{'}\beta$ et $\mathbf{E}_m(\hat{\beta}) = \beta$, nous n'avons pas pour autant $\mathbf{E}\left(\mathbf{1}_{\{y_k \le t\}}\right) = \mathbf{1}_{\{x_k\beta \le t\}}$, et donc nos estimateurs ne sont pas sans biais , même pas asymptotiquement, sous le modèle. Les auteurs proposent de remédier à ce problème (sans pour autant sacrifier les bonnes propriétés sous le plan).
Nous renvoyons le lecteur à [32] pour les hypothèses complètes permettant de dériver l'estimateur :

$$\hat{F}_{rm}(t) = \frac{1}{N}\left[\sum_{k\in S}\frac{1}{\pi_k}1_{\{y_k \le t\}} + \left(\sum_{k\in U}\hat{G}_k - \sum_{k\in S}\frac{1}{\pi_k}\hat{G}_{kc}\right)\right]$$

où : $\hat{G}_k(t) = \frac{\sum_{l\in S}\frac{1}{\pi_l}\mathbf{1}_{\left\{\frac{y_l - x_l^{'}\beta}{\sigma_l^2} \le \frac{t - x_k^{'}\beta}{\sigma_k^2}\right\}}}{\sum_{l\in S}\frac{1}{\pi_l}}$ et $\hat{G}_{kc}(t) = \frac{\sum_{l\in S}\frac{\pi_k}{\pi_{lk}}\mathbf{1}_{\left\{\frac{y_l - x_l^{'}\beta}{\sigma_l^2} \le \frac{t - x_k^{'}\beta}{\sigma_k^2}\right\}}}{\sum_{l\in S}\frac{\pi_k}{\pi_{kl}}}$

$\hat{F}_{rm}(t)$ est sans biais sous le modèle et sous le plan, approximativement. Cet estimateur n'a aucune raison pour être croissant, mais dans un soucis de cohérence, nous en prenons une version croissante.

L'inconvénient majeur de cet estimateur est le recours aux probabilités d'inclusion au second ordre, qui sont souvent inconnues. Nous allons maintenant comparer cet estimateur à celui de Chambers et Dunstan.

3.4 Comparaisons

Les comparaisons sont sans appel : l'estimateur de Chambers et Dunstan est bien meilleur que l'estimateur simple (de Hájek) ou que ceux proposés par Rao, Kovar et Mantel. Tout du moins, lorsque le modèle est bien spécifié. Mais, dans la pratique, on commence par chercher un modèle crédible, et une fois trouvé, on peut être sûr que l'estimateur de CD construit avec ce modèle donne des estimations plus fiables que les autres.

Avec ces différents estimateurs de la fonction de répartition, nous avons une estimation de la distribution de la variable d'intérêt dans toute le population. Si nous voulons le même genre d'information pour le domaine, il est difficile de procéder de la même manière, du fait du manque de données : que faire avec quelques points ? Nous pouvons chercher une réponse dans l'estimation synthétique. Ainsi, notre estimateur de la fonction de répartition du domaine tiendrait compte de l'estimation dans le domaine et de l'estimation pour toute la population, et le degré de mélange dépendrait de la ressemblance entre la population et le domaine. C'est cette idée que nous développons dans la section suivante.

3.5 Estimateur par famille exponentielles généralisées

3.5.1 Introduction

Nous souhaitons estimer la distribution d'une variable d'intérêt dans un domaine U_d de la population U. Nous pouvons voir ce problème comme l'estimation d'une loi conditionnelle : soit en effet (X,Y) un couple de variables aléatoires, où X est la variable d'intérêt, quantitative, et Y la variable du domaine, catégorielle. On note $\{1,...,D\}$ les différents domaines de U. En notant $f_{X,Y}$ la densité du couple, nous nous intéressons à l'estimation de $f_d = f_{X|Y=d}$

Nous pouvons alors voir le problème sous deux aspects différents. Nous pouvons en effet écrire :

$$f_d(x) = \frac{p_d(x)}{P_d} f_X(x),$$

où $p_d(x) = \mathbf{P}(Y = d \mid X = x)$ et $P_d = \mathbf{P}(Y = d)$. Ou alors, si F et F_d sont les lois marginales et conditionnelle à $Y = d$ de X, $dF_d = \frac{dF_d}{dF} dF$, auquel cas nous identifions dF_d / dF à p_d / P_d.

A notre connaissance, aucune technique spécifique pour l'estimation de la distribution de la variable d'intérêt dans un domaine. Nous exposons ici une technique, basée sur les propriétés des familles exponentielles.

Nous proposons de modéliser cette densité inconnues, $g_d = \frac{dF_d}{dF} = \frac{p_d}{P_d}$ à l'aide d'une famille exponentielle généralisée, $e^{\theta_d . \phi_d - \Phi_d(\theta_d) - \nu}$. Nous fixons en effet, pour chaque d, un jeu de fonctions $\phi_d : \mathbb{R} \to \mathbb{R}^{q_d}$. $\Phi_d(\theta_d)$ est alors une constante de normalisation :

$$\Phi_d(\theta_d) = \log \int_R e^{\theta_d . \phi_d - \nu} dF$$

et v est une fonction assurant que $\sum_{d=1}^{D} \mathbf{P}(Y = d \mid X) = 1$, autrement dit : $v = \log \sum_{d=1}^{D} e^{\theta_d . \phi_d - \Phi_d(\theta_d)}$.

3.5.2 Modélisation : approche synthétique

Présentation La modélisation de $dF_d / dF = p_d / P_d$ que nous utilisons est la suivante :

$$g_d = e^{\theta_d . \phi_d - \Phi_d(\theta_d) - v} \tag{10}$$

où $\Phi_d(\theta_d) = \log \int e^{\theta_d . \phi_d - v} dF$ et v est telle que $\sum_{d=1}^{D} P_d g_d = 1$. Tel quel, ce modèle n'est pas identifiable. Notons en effet $\Phi(v) = \log \int e^{\theta . \phi - v} dF$ pour marquer la dépendance en v, alors $\forall c \in \mathbb{R}$, $\Phi(v + c) - (v + c) = \Phi(v) - v$. Nous rendons le modèle identifiable en imposant une contrainte linéaire sur v, du type $\int e^{-v} dF = 1$.

On note $G_d = g_d dF$ l'approximation, mathématique, de F_d. Nous verrons dans la section suivante l'estimation de (10), c'est-à-dire la seconde approximation, statistique. Ce modèle (10) tire parti des facilités de calcul des familles exponentielles :

$$\forall \theta_d \quad \begin{cases} \Phi'_d(\theta_d) = \dfrac{\partial \Phi_d}{\partial \theta_d} = \int \phi_d g_d dF = \mathbf{E}_{G_d}(\phi_d) \\ \Phi''_d(\theta_d) = \dfrac{\partial^2 \Phi_d}{\partial \theta_d \partial \theta_d^t} = \int (\phi - \mathbf{E}_{G_d} \phi)^{\otimes 2} dG_d = \mathbf{V}_{Gd}(\phi_d) \end{cases}$$

On cherche alors θ_d qui minimise la distance de g_d à dF_d / dF. Notre objectif étant la distribution G_d plus que g_d, notre critère de distance est l'information de Kullback :

$$\min_{\theta_d} \text{ID}\ (dF_d / dG_d) \text{ où ID}\ (d\mathbf{P} / d\mathbf{Q}) = \int \log \frac{d\mathbf{P}}{d\mathbf{Q}} \frac{d\mathbf{P}}{d\mathbf{Q}} d\mathbf{Q}$$

Par conséquent, θ_d est solution du programme d'optimisation :

$$\max_{\theta_d} \underbrace{\int (\theta_d . \phi_d - \Phi_d(\theta_d) - v) dF_d}_{= L_d}$$

Une condition nécessaire pour θ_d maximise L_d est :

$$\mathbf{E}_{F_d}(\phi_d) = \mathbf{E}_{G_d}(\phi_d) \tag{11}$$

Si $\phi_d(x) = (x, x^2)^t$ cette équation signifie que les deux premiers moments sont les mêmes, dans le domaine et avec le modèle.

Estimation Nous présentons cette phase en deux temps : dans un premier temps, nous procédons comme si v était connu. Puis nous intégrons l'ignorance de v, par un système itératif.

Supposons donc connue la fonction de correction globale de v. Dans l'équation estimante (11) :

$$\frac{1}{N_d}\sum_{k\in U_d}\phi_d(y_k)=\int\phi_d\,dF_d=\int\phi_d\,dG_d=\int\phi_d\,g_d\,dF=\frac{1}{N}\sum_{k\in U}\phi_d(y_k)g_d(y_k) \qquad (12)$$

F et F_d sont inconnues, et donc Φ_d et $g_d=e^{\theta_d.\phi_d-\Phi_d(\theta_d)-v}$ aussi. Remarquons que g_d est inconnue de deux façons différentes : par θ_d et part Φ_d. Pour l'instant seul Φ_d nous importe. Et nous estimons, en remplaçant dF par $d\hat{F}$, et $dG_d=g_d\,dF$ par $d\hat{G}_d=\hat{g}_d\,d\hat{F}$.

Avec $\hat{F}_d$ et $\hat{F}$, on estime Φ_d et g_d (à θ_d près, donc) :

$$\forall\theta_d,\quad \hat{\Phi}_d(\theta_d)=\log\int e^{\theta_d.\phi_d-v}d\hat{F}=\log\sum_{k\in S}w_k e^{\theta_d.\phi_d(y_k)-v(y_k)} \quad \text{d'où} \quad \hat{g}_d=e^{\theta_d.\phi_d-\hat{\Phi}_d(\theta_d)-v} \qquad (13)$$

Pour le membre de gauche de (11), $\mathbf{E}_{F_d}(\phi_d)=N_d^{-1}\sum_{k\in U_d}\phi_d(y_k)$, nous pouvons utiliser l'estimateur direct $\mathbf{E}_{\hat{F}_d}(\phi_d)=\sum_{k\in S_d}w_{kd}\phi_d(y_k)$. Il est direct car les seuls y_k nécessaires à son expression appartiennent au domaine. Nous pourrions aussi utiliser des estimateurs à la Fay et Herriot [21], comme nous l'avons vu précédemment.

Finalement, au lieu de prendre $\mathbf{E}_{\hat{F}_d}(\phi_d)$ pour estimer $\mathbf{E}_{F_d}(\phi_d)$, on utilise un estimateur synthétique de modèle, du type $\gamma_d\hat{\phi}_d+(1-\gamma_d)\hat{\phi}$ où $\hat{\phi}_d=\mathbf{E}_{\hat{F}_d}(\phi_d)$ et $\hat{\phi}=\mathbf{E}_{\hat{F}}(\phi_d)$. On note $\hat{\phi}_d^S$ cet estimateur synthétique. Dans la pratique, les poids γ_d sont aléatoires. Pour l'estimation de variance, on aboutit à la même formule que précédemment, à cela près que les variances ne sont plus de plan mais de modèle. Ce qui pose le problème de l'estimation de la variance de $\hat{\phi}_d^S$, complexe lui-aussi [30] [18].

Nous en tirons alors une estimation de l'équation estimante (11) :

$$\hat{\phi}_d^S=\int\phi_d\hat{g}_d\,d\hat{F}=\sum_{k\in S}w_k\phi_d(y_k)\exp(\theta_d.\phi_d(y_k)-\hat{\Phi}_d(\theta_d)-v(y_k))=\mathbf{E}_{\hat{G}_d}(\phi_d) \qquad (14)$$

Soit finalement :

$$\hat{\phi}_d^S=\frac{\sum_{k\in S}w_k\phi_d(y_k)e^{\theta_d.\phi_d(y_k)-v(y_k)}}{\sum_{k\in S}w_k e^{\theta_d.\phi_d(y_k)-v(y_k)}}$$

D'où nous tirons une estimation de $\hat{\theta}_d$.

Néanmoins, dans la pratique, v est inconnue. Nous procédons alors par étapes. Partant de $\hat{v}^{(0)}=0$, on estime un premier $\hat{\theta}_d^{(1)}$, avec (14). Nous obtenons donc le premier terme de g_d, $\exp(\hat{\theta}_d^{(1)}.\phi_d)$. Nous cherchons alors $\Phi_d(\hat{\theta}_d)$ et $\hat{v}$ tels que :

- $g_d=e^{\hat{\theta}_d.\phi_d-\Phi_d(\hat{\theta}_d)-\hat{v}}$ soit une densité par rapport à F
- g_d vérifie $\sum_{d=1}^{D}P_d g_d\equiv 1$

Nous traduisons ces deux contraintes par le système :

$$\begin{cases} \forall d, & \sum_{k\in S} w_k e^{\hat{\theta}_d.\phi_d(y_k)-\hat{\Phi}_d(\hat{\theta}_d)-\hat{v}(y_k)} = 1 \\ \forall k \in S, & \sum_{d=1}^{D} P_d e^{\hat{\theta}_d.\phi_d(y_k)-\hat{\Phi}_d(\hat{\theta}_d)-\hat{v}(y_k)} = 1 \end{cases}$$

Nous voyons que ce système n'est pas linéaire en $\hat{\Phi}_d(\hat{\theta}_d)$ ni en $v(y_k)$. Nous retrouvons un problème similaire au raking ratio : caler le tableau des $e^{\hat{\theta}_d.\phi_d(y_k)}$ sur les marges 1, avec les deux systèmes de poids w_k et P_d. Le calage sur marges est très rapide en calcul. Les multiplicateurs nous donnent alors $\Phi_d(\hat{\theta}_d)$ et $v(y_k)$.

La fonction v obtenue ne vérifie a priori pas la contrainte d'identification du modèle, $f(v) = c$, où f est une fonction équivariante par translation (si c est une constante, $f(v+c) = f(v) + c$). Par exemple, $f(v) = -\log\int e^{-v} dF$ et $c = 0$. Pour ce faire, il suffit de prendre $v - f(v) + c$ et $\Phi_d(\hat{\theta}_d) + f(v) - c$. Après cela, nous disposons de $\hat{\theta}_d^{(1)}$, $\hat{\Phi}_d^{(1)}(\hat{\theta}_d)$ et $\hat{v}_d^{(1)}$. On recommence jusqu'à convergence.

Finalement, l'estimateur de la fonction de répartition est :

$$\forall x \in \mathbb{R}, \qquad \hat{F}_d(x) = \sum_{k\in S} w_k \mathbf{1}_{\{y_k \le x\}} e^{\hat{\theta}_d.\phi_d(y_k)-\hat{\Phi}_d(\hat{\theta}_d)-\hat{v}(y_k)} \tag{16}$$

Il est bien positif, croissant, et vérifie $\hat{F}_d(-\infty) = 0$, $\hat{F}_d(+\infty) = 1$.

Pour l'estimation de la densité, $dF_d / d\lambda$, on peut utiliser deux méthodes. Dans la première, on utilise la formule $dF_d = \frac{dF_d}{dF} dF$. Le premier morceau est connu, nous nous sommes efforcés de l'estimer dans la section précédente, et pour le second, nous pouvons utiliser un estimateur de la densité classique, car il y a suffisamment d'observations. On obtient donc finalement :

$$\forall x \in \mathbb{R}, \qquad \hat{f}_d(x) = e^{\hat{\theta}_d.\phi_d(x)-\hat{\Phi}_d(\hat{\theta}_d)-\hat{v}(x)} \sum_{k\in S} w_k K_h(x - y_k)$$

où $K_h = h^{-1}K(./h)$ avec K un noyau de Parzen et h la fenêtre (dont le choix est ardu). Un autre méthode consiste à régulariser l'estimation (12), à l'aide d'un noyau régularisant (cf [9], section 11).

Estimation de variance Pour des raisons de lisibilité, on omet l'indice de domaine pour θ, ϕ et Φ. On cherche une estimation de variance de $\hat{\theta}$, après convergence de v. On procède par linéarisation. Soit θ^* la solution de l'équation estimante (11), alors :

$$\begin{aligned} \Phi'(\hat{\theta}) &= \Phi'(\theta^*) + \Phi''(\theta^*)(\hat{\theta} - \theta^*) + o(\hat{\theta} - \theta^*) \\ \Rightarrow \hat{\theta} - \theta^* &\approx \left[\Phi''(\theta^*)\right]^{-1}\left[\Phi'(\hat{\theta}) - \Phi'(\theta^*)\right] \end{aligned}$$

D'où l'approximation de la variance :

$$A\mathbf{V}(\hat{\theta}) = \left[\Phi''(\theta^*)\right]^{-1} \mathbf{V}\left(\Phi'(\hat{\theta})\right) \left[\Phi''(\theta^*)\right]^{-1}$$

Estimation de $\mathbf{V}\left(\Phi'(\hat{\theta})\right)$ Pour la clarté de l'exposé, on note $g_\theta = e^{\theta.\phi - \Phi(\theta) - v}$, pour le domaine d. Alors, pour estimer $\mathbf{V}\left(\Phi'(\hat{\theta})\right)$, on procède comme suit : $\Phi'(\hat{\theta}) = \int \phi e^{\hat{\theta}.\phi - \Phi(\hat{\theta}) - v} dF$, or par

construction[3] : $g_{\theta^*} dF \approx dF_d$; ensuite ; comme précédemment, nous estimons $\mathbf{E}_{F_d}(\phi)$ par $\hat{\phi}^S$. Finalement, nous prenons $\hat{\phi}^S$ pour pallier l'ignorance de $\mathbf{E}_{g_{\hat{\theta}} dF}(\phi)$.

Nous nous ramenons donc à l'estimation de la variance de modèle de l'estimateur synthétique $\hat{\phi}^S$, qui dépend du modèle de superpopulation utilisé. On trouvera dans [3], [28] et [29] plusieurs modélisations possibles. Signalons néanmoins l'intérêt particulier de [30], où les auteurs introduisent le concept de robustesse : l'estimateur construit, basé sur le plan, résiste aux mauvaises spécifications du modèle.

Estimation de $\mathbf{\Phi}^{"}(\boldsymbol{\theta}^*)$ Comme précédemment, $\Phi^{"}(\theta^*)$ est égale à la variance de ϕ sous la vraie loi $g_{\theta^*} dF$ du modèle. Pour l'estimation, on approche $g_{\theta^*} dF$ par $d\hat{F}_d$, d'où l'on estime $\Phi^{"}(\theta^*)$ par $\mathbf{V}_{\hat{F}_d}(\phi)$:

$$\widehat{\Phi^{"}(\theta^*)} = \mathbf{V}_{\hat{G}_d}(\phi) = \int \left[\phi - \mathbf{E}_{\hat{F}_d}(\phi)\right]^{\otimes 2} d\hat{F}_d \qquad (17)$$

$$= \sum_{k \in S_d} w_{kd} \left(\phi(y_k) - \sum_{l \in S_d} w_{ld} \phi(y_l) \right)^{\otimes 2}$$

$$= \sum_{k \in S_d} w_{kd} \phi(y_k)^{\otimes 2} - \left(2 - \sum_{k \in S_d} w_{kd} \right) \left(\sum_{k \in S_d} w_{kd} \phi(y_k) \right)^{\otimes 2}$$

Le plus souvent, on utilise les poids de Hájek, $w_{kd} = \pi_k^{-1} \Big/ \sum_{k \in S_d} \pi_k^{-1}$, auquel cas $\sum_{k \in S_d} w_{kd} = 1$ et :

$$\widehat{\Phi^{"}(\theta^*)} = \mathbf{V}_{\hat{F}_d}(\phi) = \sum_{k \in S_d} w_{kd} \phi(y_k)^{\otimes 2} - \left(\sum_{k \in S_d} w_{kd} \phi(y_k) \right)^{\otimes 2}$$

Pour un SAS, $w_{kd} = n_d^{-1}$ d'où :

$$\widehat{\Phi^{"}(\theta^*)} = \frac{1}{n_d} \sum_{k \in S_d} \phi(y_k)^{\otimes 2} - \left(\frac{1}{n_d} \sum_{k \in S_d} \phi(y_k) \right)^{\otimes 2}$$

3.5.3 Choix des variables

Avec l'estimation de variance de $\hat{\theta}_d$, il est possible de tester la nullité du paramètre. Si en effet la valeur absolue du ratio $\hat{\theta}_d \big/ \hat{\mathbf{V}}(\hat{\theta}_d)$ est trop faible, on ne peut considérer que la distribution du domaine diffère sensiblement de celle de toute la population : la modélisation est inutile.

De la même manière, on peut tester si une des composantes de $\hat{\theta}_d$ est nulle, ce qui correspond à une inefficacité de la fonction associée pour améliorer l'estimation de la distribution. Avec cette méthode, on arrive à sélectionner les variables influentes.

3.5.4 Application aux quantiles

[3] Ou plus précisément : $g_{\theta^*} dF \approx dF_d$, et $\hat{\theta} \approx \theta^* \Rightarrow g_{\hat{\theta}} \approx g_{\theta^*}$ d'où $g_{\hat{\theta}} dF \approx dF_d$

Une application immédiate de ce qui précède est l'estimation de fractiles. Soit en effet $\alpha \in]0;1[$, on cherche $t_\alpha = F_d^{-1}(\alpha)$. On retient l'estimateur $\hat{t}_\alpha = \hat{F}_d^{-1}(\alpha)$, si l'on utilise l'estimateur régularisé $\hat{F}_d$ de F_d. Sinon, on peut prendre l'inverse de P. Lévy : $\hat{t}_\alpha = \inf\left\{t \in \mathbb{R} / \hat{F}_d(t) \geq \alpha\right\}$

L'estimation de variance de cet estimateur repose sur le $\hat{F}_d$ choisi. Pour $\hat{F}_d = \sum_{k \in S_d} w_{kd} \mathbf{1}_{\{y_k \leq .\}}$, Deville [9] montre qu'on peut prendre pour linéarisée du fractiles $\hat{t}_\alpha$ la variable :

$$z_k = -\frac{1}{\hat{f}_d(\hat{t}_\alpha)}\left[\mathbf{1}_{\{y_k \leq \hat{t}_\alpha\}} - \alpha\right]$$

Pour notre modèle, on suit [6] où les auteurs font l'hypothèse suivante :

$$\hat{t}_\alpha = t_\alpha + \frac{\alpha - \hat{F}_d(t_\alpha)}{f(t_\alpha)} + o_p(\sqrt{n}) \tag{18}$$

D'où l'on tire une estimation de la variance de $\hat{t}_\alpha$:

$$A\mathbf{V}(\hat{t}_\alpha) = \mathbf{V}\left(\frac{\hat{F}_d(t_\alpha)}{f(t_\alpha)}\right) = \frac{\mathbf{V}\left(\hat{F}_d(t_\alpha)\right)}{\left[f(t_\alpha)\right]^2} \tag{19}$$

Et finalement l'estimation de variance :

$$\hat{V}(\hat{t}_\alpha) = \frac{\hat{V}\left(\hat{F}_d(x)\right)\Big|_{x=\hat{t}_\alpha}}{\left[\hat{f}_d(\hat{t}_\alpha)\right]^2} \tag{20}$$

Conclusion

Ce dernier estimateur présente l'avantage d'utiliser à la fois le domaine et le reste de l'échantillon. C'est un gros avantage par rapport à des versions locales, ou directes, d'estimateurs plus classiques comme celui de Chambers et Dunstan. Par ailleurs, il ne fait pas l'hypothèse de modèles forte, qui serait par ailleurs d'autant plus difficile à vérifier que la taille de l'échantillon est petite. Enfin, via notre estimateur, nous pouvons réemployer des techniques utilisées pour l'estimation de fonctions d'intérêt plus simples, puisque l'estimation du paramètre de dimension finie repose sur une équation estimante, qui égalise deux moyennes de domaines. La suite de ce travail consiste en partie en la réalisation de simulations, permettant de comparer sur des exemples précis les performances de différents estimateurs. Nous aimerons aussi voir comment ces différentes techniques peuvent être appliquées à des cas pratiques. Il existe un précédent à l'Insee, pour l'estimation de la non-inscription électorale dans trois zones urbaines de Bretagne.

Références

[1] Pascal Ardilly. *Les techniques de sondage*. Technip, 1994.

[2] Ketty Attal-Toubert and Olivier Sautory. Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle. *Document de travail de l'Unité de Méthodologie Statistique*, 1998.

[3] Georges E. Batesse, Rachel M. Harter, and Wayne A. Füller. An error-components model for prediction of contry crop using servey satellite data. *Journal of the American Statistical Association*, 1998.

[4] K. R. W Brewer. A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 1979.

[5] Claes-Magnus Cassel, Carl-Erik Särndal, and Håkan Wretman. *Foundations of Inference in Survey sampling*. Krieger plublishing compagny, 1977.

[6] Dunstan Chambers. Estimating distribution functions from survey data. *Biometrika*, 1986.

[7] William G. Cochran. *Sampling techniques*. John Wiley and Sons, 1977.

[8] A.P. Dempster, D.B. Rubin, and R.K. Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 1981.

[9] Jean-Claude Deville. Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquêtes*, 25:219-230, 1999.

[10] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 1992.

[11] Jean-Claude Deville and Yves Tillé. Variance reduction using balanced sampling : the cube method. *a paraître*, 2000.

[12] Dorfman. A comparison of design-based ans model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 1993.

[13] Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators - Part I : The Bayes case. *Journal of the American Statistical Association*, 1971.

[14] Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators - Part II : The empirical Bayes case. *Journal of the American Statistical Association*, 1972.

[15] Bradley Efron and Carl Morris. Stein's estimation rule and its competitors - an empirical bayes approach. *Journal of the American Statistical Association*, 68:117-130, mars 1973.

[16] Risto Lehtonen et Ari Veijanen. Estimateurs de régression généralisés logistiques. *Techniques d'enquêtes*, 24:53-58, 1998.

[17] Victor M. Estevao et Carl-Erik Särndal. The use of auxiliary information in design-based estimation for domains. *Techniques d'enquêtes*, 1999.

[18] Raghu N. Kackar et David A. Harville. Approximation for standard errors of estimation of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 1984.

[19] E.L. Lehmann et George Casella. *Theory of point estimation*. Springer-Verlag, 1998.

[20] Malay Ghosh et J.N.K Rao. Small area estimation : an appraisal. *Statistical science*, 1994.

[21] Robert E. Fay and R.A. Herriot. Estimates of incomes for small places : an application of james-stein procedures do census data. *Journal of the American Statistical Association*, pages 269-277, juin 1979.

[22] David Firth and F.E. Bennett. Robust models in probability sampling. *Journal of the Royall Statistical society, Series B*, 1998.

[23] Malay Gosh and Glen Meeden. Empirical bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 1986.

[24] David A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *JASA*, 1977.
[25] David A. Harville and Daniel R. Jeske. Mean squares error of estimation or prediction under a general linear model. *JASA*, 1992.
[26] Cary T. Isaki and Wayne A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 1982.
[27] Philipp S. Kott. Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires. *Techniques d'enquêtes*, 1989.
[28] Fernando A.S. Moura and David Holt. Production d'estimations régionales à partir de modèles multiniveaux. *Techniques d'enquêtes*, 1999.
[29] N.G.N Prasad and J.N.K Rao. The estimation of lean squared error of small-area estimators. *Journal of the American Statistical Association*, 1990.
[30] N.G.N Prasad and J.N.K Rao. Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires, *Techniques d'enquêtes*, 1999.
[31] J.N.K Rao. Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquêtes*, 1999.
[32] Mantel Rao, Kovar. On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 1990.
[33] Christian Robert. *L'anlayse statistique bayésienne*. Economica, 1992.
[34] Richard M. Royall. On finite population sampling theory under certain linear regression models. *Biometrika*, 57:377-387, Août 1970.
[35] Carl-Erik Särndal. On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 1980.
[36] Carl-Erik Särndal, Bengt Swennsson, and Jan Wretman. *Model assisted survey sampling*. Springer-Verlag, 1992.
[37] Shayle R. Searle. *Linear models*. John Wiley, 1971.
[38] George Casella et Charles E. McCulloch Shayle R. Searle. *Variance component*. Wiley series in probability and mathematical statistics. John Wiley and Sons, 1992.
[39] Yves Tillé. *Théorie des sondages*. Dunod, 2001.
[40] Roger L. Wright. Finite population sampling with multivariate information. *Journal of the American Statistical Association*, 1983.