Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Propriétés théoriques et empiriques d'estimateurs par la régression fondés sur un test de décision assistés par modèle

par Jun Shao, Eric Slud, Yang Cheng, Sheng Wang
et Carma Hogue

Juin 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 89-114
Statistique Canada, N° 12-001-X au catalogue

Propriétés théoriques et empiriques d'estimateurs par la régression fondés sur un test de décision assistés par modèle

Jun Shao, Eric Slud, Yang Cheng, Sheng Wang et Carma Hogue[1]

Résumé

En 2009, deux enquêtes importantes réalisées par la division des administrations publiques du U.S. Census Bureau ont été remaniées afin de réduire la taille de l'échantillon, d'économiser des ressources et d'améliorer la précision des estimations (Cheng, Corcoran, Barth et Hogue 2009). Sous le nouveau plan de sondage, chaque strate habituelle, définie par l'État et le type d'administration publique, qui contient un nombre suffisant d'unités (administrations publiques) est divisée en deux sous-strates en fonction de la masse salariale totale de chaque unité afin de tirer un plus petit échantillon de la sous-strate des unités de petite taille. L'approche assistée par modèle est adoptée pour estimer les totaux de population. Des estimateurs par la régression utilisant des variables auxiliaires sont obtenus soit pour chaque sous-strate ainsi créée soit pour la strate originale en regroupant des deux sous-strates. Cheng, Slud et Hogue (2010) ont proposé une méthode fondée sur un test de décision qui consiste à appliquer un test d'hypothèse pour décider quel estimateur par la régression sera utilisé pour chaque strate originale. La convergence et la normalité asymptotique de ces estimateurs assistés par modèle sont établies ici sous un cadre asymptotique fondé sur le plan de sondage ou assisté par modèle. Nos résultats asymptotiques suggèrent aussi deux types d'estimateurs de variance convergents, l'un obtenu par substitution des quantités inconnues dans les variances asymptotiques et l'autre en appliquant la méthode du bootstrap. La performance de tous les estimateurs des totaux et des estimateurs de leur variance est examinée au moyen d'études empiriques. L'Annual Survey of Public Employment and Payroll (ASPEP) des États-Unis est utilisé pour motiver et illustrer notre étude.

Mots-clés : Normalité asymptotique; bootstrap; estimateur fondé sur un test de décision; probabilité proportionnelle à la taille; stratification; estimation de la variance.

1 Introduction

L'Annual Survey of Public Employment and Payroll (ASPEP) des États-Unis fournit des estimations courantes de l'emploi et de la rémunération à temps plein et à temps partiel dans les administrations publiques d'État et locales par fonction (par exemple, enseignement primaire et secondaire, enseignement supérieur, services de police, services de protection contre l'incendie, administration financière, services judiciaires et juridiques, *etc.*). Cette enquête a pour champ d'observation les administrations publiques d'État et locales (89 526 selon le Census of Governments de 2007), qui englobent les comtés, les villes, les cantons, les administrations appelées « districts spéciaux » et les districts scolaires. L'ASPEP, qui est la seule source de données sur l'emploi dans le secteur public par fonction administrative et catégorie d'emploi, fournit des données sur le nombre et la rémunération des employés à temps plein et à temps partiel, ainsi que le nombre d'heures travaillées par les employés à temps partiel. Habituellement, la collecte des données débute en mars et se poursuit pendant environ sept mois, en prenant la période de paye incluant le 12 mars comme période de référence.

Soit U la population finie de N unités subdivisée en H strates, $U_1, \ldots, U_H$, où U_h contient N_h unités et $N_1 + \cdots + N_H = N$. Le plan de sondage habituel de l'ASPEP est un plan avec probabilité

1. Jun Shao, Statistics Department University of Wisconsin, Madison WI , Courriel: shao@stat.wisc.edu; Eric Slud, Center for Statistical Research and Methodology, US Census Bureau, Washington DC and Mathematics Department, University of Maryland, College Park, MD, Courriel: eric.v.slud@census.gov; Yang Cheng, Demographic Statistical Methods Division, US Census Bureau, Washington DC, Courriel: yang.cheng@census.gov; Sheng Wang, Mathematica Policy Research, Princeton NJ, Courriel : swang@mathematica-mpr.com; et Carma Hogue, Governments Division, US Census Bureau, Washington DC, Courriel: carma.ray.hogue@census.gov.

proportionnelle à la taille (PPT), où les strates sont construites en se basant sur l'État et le type d'administration publique, à savoir le comté, le sous-comté (grande ou petite ville), le district spécial ou le district scolaire. La taille de chaque unité (administration publique) est mesurée par la masse salariale totale, et l'échantillonnage est effectué indépendamment dans les diverses strates. En 2009, on a élaboré un plan d'échantillonnage modifié, qui comprend la division de certaines strates U_h en deux sous-strates, U_{h1} et U_{h2} contenant N_{h1} et N_{h2} unités, respectivement, où U_{h1} contient les unités de petite taille (Cheng et coll. 2009). L'idée était d'économiser des ressources et de réduire le fardeau de réponse en sélectionnant dans U_{h1} un échantillon plus petit sous le plan modifié que sous le plan habituel. Soit S_{hj} un échantillon PPT de taille n_{hj} provenant de U_{hj}, $j = 1, 2$, $n_{h1} + n_{h2} = n_h$. Notons que n_{h1} peut encore être plus grand que n_{h2}, parce que N_{h1} est habituellement beaucoup plus grand que N_{h2}.

Pour l'unité $i \in U$, soit y_i une variable étudiée clé (p. ex., l'emploi à temps plein, la rémunération à temps plein, l'emploi à temps partiel, la rémunération à temps partiel, les heures travaillées à temps partiel), x_i une variable auxiliaire, disons la même variable que y_i provenant du recensement le plus récent, et soit z_i la covariable utilisée comme variable de taille dans l'échantillonnage PPT. Les valeurs des covariables x_i et z_i sont observées pour tout $i \in U$, tandis que y_i est observée uniquement pour chaque unité *i* échantillonnée.

L'estimateur de Horvitz-Thompson du total inconnu $Y = \sum_{i \in U} y_i$ est

$$\hat{Y}_{\mathrm{HT}} = \sum_h \sum_j \sum_{i \in S_{hj}} y_i / \pi_i , \tag{1.1}$$

où π_i est la probabilité d'inclusion d'ordre un de l'unité i dans S_{hj}, une fonction connue des z_i. Pour utiliser la variable auxiliaire x_i et accroître la précision de l'estimation de Y, l'approche assistée par modèle (Särndal, Swensson et Wretman 1992) a été adoptée. L'application de la régression dans chaque échantillon S_{hj} conduit à l'estimateur par la régression de Y de la forme

$$\hat{Y}_{\mathrm{reg},2} = \sum_h \sum_j \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right], \tag{1.2}$$

où $X_{hj} = \sum_{i \in U_{hj}} x_i$, $\hat{Y}_{hj} = \sum_{i \in S_{hj}} y_i / \pi_i$, $\hat{X}_{hj} = \sum_{i \in S_{hj}} x_i / \pi_i$, $\hat{N}_{hj} = \sum_{i \in S_{hj}} 1/\pi_i$, et

$$\hat{\beta}_{hj} = \frac{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right) y_i / \pi_i}{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2 / \pi_i}.$$

Autrement, la combinaison des deux sous-strates S_{h1} et S_{h2} donne l'estimateur par la régression suivant. (Un examinateur fait remarquer correctement que $\hat{Y}_{\mathrm{reg},1}$ dans (1.3) n'est pas l'estimateur groupé que l'on utiliserait si les droites de régression dans la strate h étaient combinées mais que les deux sous-strates ne l'étaient pas; cependant, il *est* l'estimateur naturel lorsque non seulement les droites de régression, mais aussi les sous-strates sont combinées.)

$$\hat{Y}_{\text{reg},1} = \sum_h \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right], \tag{1.3}$$

où $\hat{Y}_h = \sum_j \hat{Y}_{hj}$, $\hat{X}_h = \sum_j \hat{X}_{hj}$, $\hat{N}_h = \sum_j \hat{N}_{hj}$, et

$$\hat{\beta}_h = \frac{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right) y_i / \pi_i}{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right)^2 / \pi_i}.$$

Puisque $\hat{Y}_{\text{reg},1}$ ainsi que $\hat{Y}_{\text{reg},2}$ sont des estimateurs assistés par modèle, ils sont convergents sous échantillonnage répété, que le modèle de régression soit ou non vérifié. Si les droites de régression par les moindres carrés dans les deux sous-strates U_{hj} sont les mêmes, $\hat{Y}_{\text{reg},1}$ peut être plus efficace que $\hat{Y}_{\text{reg},2}$. Par ailleurs, si les droites de régression sont différentes, $\hat{Y}_{\text{reg},2}$ peut être plus efficace que $\hat{Y}_{\text{reg},1}$.

Cheng et coll. (2010) ont proposé une méthode fondée sur un test de décision qui consiste à appliquer un test d'hypothèse pour décider s'il faut combiner S_{h1} et S_{h2}. À l'intérieur de la strate h, on teste l'hypothèse d'égalité des pentes des droites de régression dans U_{h1} et U_{h2}. Soit

$$\hat{\alpha}_{hj} = \frac{\hat{Y}_{hj} - \hat{\beta}_{hj} \hat{X}_{hj}}{\hat{N}_{hj}}, \hat{\sigma}^2_{xe,hj} = \frac{n_{hj}}{\hat{N}^2_{hj}} \sum_{i \in S_{hj}} \left(x_i - \frac{\hat{X}_{hj}}{\hat{N}_{hj}} \right)^2 \frac{\left(y_i - \hat{\alpha}_{hj} - \hat{\beta}_{hj} x_i \right)^2}{\pi_i^2},$$

$$\hat{\sigma}^2_{xhj} = \sum_{i \in S_{hj}} \frac{\left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2}{\pi_i \hat{N}_{hj}}, t_h = \sqrt{n_h - 4} \left(\hat{\beta}_{h1} - \hat{\beta}_{h2} \right) \Big/ \sqrt{n_h \sum_{j=1}^{2} \frac{\hat{\sigma}^2_{xe,hj}}{n_{hj} \hat{\sigma}^4_{xhj}}}.$$

Si $|t_h| > t_{1-\tau/2, n_h - 4}$, où $t_{1-\tau/2,\nu}$ est le $(1 - \tau/2)^{\text{e}}$ quantile de la distribution t avec ν degrés de liberté, alors nous rejetons l'hypothèse d'une pente commune et nous utilisons $\hat{\beta}_{hj}$ (et fixons $\zeta_h = 1$). Ici, τ est un seuil de signification nominal fixé par défaut à 0,05, mais nous considérerons d'autres choix de la valeur de τ à la section consacrée aux simulations. La définition de la statistique de test faisant intervenir $n_h - 4$ degrés de liberté est un choix légèrement artificiel conçu afin de rendre les probabilités de rejet d'un échantillon modéré plus proches de la valeur nominale, mais la théorie asymptotique en grand échantillon justifiant ce test est donnée à la partie (c) du théorème 1. Si $|t_h| \leq t_{1-\tau/2, n_h - 4}$, alors nous acceptons l'hypothèse d'une pente commune, nous combinons les sous-strates S_{h1} et S_{h2}, et nous utilisons $\hat{\beta}_h$ (en fixant $\zeta_h = 0$). Les tests sont effectués de manière indépendante dans les diverses strates $h = 1, \ldots, H$. L'estimateur de Y fondé sur le test de décision est alors

$$\hat{Y}_{\text{dec}} = \sum_h \sum_j \zeta_h \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right] + \sum_h (1 - \zeta_h) \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right]. \tag{1.4}$$

Puisque les deux droites de régression ayant une pente commune peuvent avoir des ordonnées à l'origine différentes, on pourrait tester une hypothèse supplémentaire concernant les ordonnées à l'origine pour décider s'il faut combiner les deux sous-strates. Cependant, des points de population (x_i, y_i) se

trouvant sur deux droites de régression de sous-strate parallèles, mais non identiques seraient discontinus autour du seuil entre les deux sous-strates U_{h1} et U_{h2}, ce qui ne semble ne se produire que rarement dans les situations pratiques. Par exemple, dans l'ASPEP, Cheng et coll. (2010) ont étudié les pentes et les ordonnées à l'origine de sous-strates dans les ensembles de donnés des recensements des administrations publiques de 2002 et de 2007, et ont constaté que l'hypothèse d'une ordonnée à l'origine commune ne pouvait jamais être rejetée lorsque l'hypothèse d'une pente commune ne pouvait pas l'être. Donc, l'estimateur fondé sur un test de décision donné dans (1.4) dépend uniquement du test de l'hypothèse d'égalité des pentes des droites de régression des sous-strates.

Les estimateurs à deux degrés étudiés ici sont des cas particuliers de procédures nommées antérieurement estimateurs après un test préliminaire (*preliminary test estimators*). Il existe une littérature abondante traitant de l'utilisation de ce genre de procédures dans les enquêtes, y compris une bibliographie de Bancroft et Han (1977), un livre publié par Saleh (2006) et un traitement proposé par Fuller (2009, section 6.7). Une idée de Saleh (2006) consiste à estimer les coefficients par une combinaison convexe des coefficients estimés à partir des strates distinctes en faisant dépendre les proportions d'une statistique de test. Les estimateurs lissés de ce genre pourraient être plus efficaces que nos procédures fondées sur un test de décision. Si les ordonnées à l'origine et les pentes propres aux strates étaient considérées comme aléatoires, on pourrait aussi essayer d'appliquer à l'estimation une approche bayésienne empirique fondée sur un modèle.

Les estimateurs fondés sur un test de décision (1.4) sont nouveaux, parce qu'ils sont assistés par modèle et convergents sous le plan dans le contexte des sondages, et utilisent explicitement les tailles de population de sous-strate connues. Dans un esprit à peu près semblable, Rao et Ramachandran (1974) avaient effectué antérieurement une comparaison exacte des estimateurs par le ratio distincts et combinés sous un modèle de ratio similaire au modèle de régression considéré dans le présent article.

L'objectif de l'article est d'illustrer certaines propriétés asymptotiques et empiriques des estimateurs de Y décrits plus haut et des estimateurs de leur variance. La convergence et la normalité asymptotique de $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, et $\hat{Y}_{\text{dec}}$ sont établies à la section 2, dans le contexte de la théorie asymptotique fondée sur le plan de sondage ou assistée par modèle. Bien que les résultats asymptotiques d'ordre un favorisent $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{reg},1}$ pourrait être meilleur quand certaines tailles d'échantillon de sous-strate n_{h2} sont modérées, un effet asymptotique d'ordre deux. L'avantage de l'estimateur fondé sur un test de décision $\hat{Y}_{\text{dec}}$ tient à l'adaptation en vue d'être proche de $\hat{Y}_{\text{reg},1}$ ou de $\hat{Y}_{\text{reg},2}$ selon celui qui est le meilleur. Comme l'indique la discussion du paragraphe (III) de la section 4.4, les simulations montrent que l'avantage de cette adaptabilité est de réduire l'EQM d'une quantité allant jusqu'à quelques pour cent sous des conditions de paramétrisation raisonnables, et de plus grandes quantités sous des conditions plus étranges.

L'estimation de la variance de l'estimateur fondé sur un test de décision est traitée à la section 3. Même si la théorie asymptotique exposée à la section 2 laisse entendre que des estimateurs convergents de variance sont obtenus par substitution des quantités inconnues dans les formules de variance asymptotique, nous étudions aussi les estimateurs bootstrap de la variance proposés dans Cheng et coll. (2010), qui ont généralement de meilleures propriétés en échantillon fini que les estimateurs par substitution. Les résultats empiriques sont présentés à la section 4, les interprétations et les conclusions étant formulées à la sous-section 4.4. Toutes les preuves techniques sont données en annexe.

2 Convergence et normalité asymptotique

Afin d'examiner les propriétés asymptotiques, nous considérons la population U comme l'une d'une série de populations $\{U^{(m)}, m = 1, 2, \ldots\}$, où le nombre d'unités dans $U^{(m)}$ tend vers l'infini quand $m \to \infty$. Nous ne traitons ici que le cas de strates desquelles est tiré un grand échantillon n_h ; autrement dit, nous supposons que, pour chaque strate h, la taille de l'échantillon n_h dépend de m et tend vers l'infini quand $m \to \infty$, mais nous omettons l'indice m pour simplifier la notation. Tous les processus limites sont considérés pour $m \to \infty$. À l'instar d'auteurs tels que Isaki et Fuller (1982) et Deville et Särndal (1992), nous donnons à ces conditions le nom de cadre asymptotique de *superpopulation*. Sous le cadre fondé sur le plan de sondage considéré à la section 2.1, les vecteurs d'attributs dans les populations sous-jacentes ne doivent pas être considérés comme des vecteurs aléatoires. Cependant, sous le cadre assisté par modèle considéré à la section 2.2, des modèles de régression hypothétiques sont associés aux vecteurs d'attributs.

Puisque chaque estimateur est une somme d'estimateurs indépendants construits dans chaque strate, pour simplifier, nous présentons les résultats asymptotiques pour le cas où $H = 1$. Les résultats et les conclusions s'appliquent directement au cas d'une valeur fixe de H et peuvent aussi être étendus à la situation où H tend vers l'infini. (Il est habituel que les grandes enquêtes contiennent de nombreuses strates, quoique dans l'ASPEP, le nombre de strates définies selon le type d'administration publique qui ont été subdivisées en sous-strates était un peu inférieur à 100.) Puisque nous considérons seulement le cas $H = 1$, nous omettons l'indice h désignant la strate à la présente section, p. ex., $n_{hj} = n_j$, $n_h = n$, $N_{hj} = N_j$ et $N_h = N$. En outre, pour $j = 1, 2$, les estimateurs $\hat{\beta}_j$ et $\hat{\beta}$ sont définis par les formules présentées après les équations (1.2) et (1.3) avec l'indice inférieur h supprimé, considérées conjointement avec

$$\hat{\mu}_{xj} = \hat{X}_j \big/ \hat{N}_j, \quad \hat{\alpha}_j = \hat{Y}_j \big/ \hat{N}_j - \hat{\beta}_j \hat{\mu}_{xj}, \quad \hat{\sigma}^2_{xj} = \hat{N}_j^{-1} \sum_{i \in S_j} \pi_i^{-1} \left(x_i - \hat{\mu}_{xj}\right)^2$$

$$\hat{\sigma}^2_{xe,j} = n_j \sum_{i \in S_j} \left(x_i - \hat{\mu}_{xj}\right)^2 \left(y_i - \hat{\alpha}_j - \hat{\beta}_j x_i\right)^2 \Big/ \left(\pi_i^2 \hat{N}_j^2\right).$$

De surcroît, pour simplifier, nous n'examinons les résultats asymptotiques que sous échantillonnage avec remise. Les résultats peuvent être appliqués au cas de l'échantillonnage sans remise si la fraction d'échantillonnage n/N est négligeable.

2.1 Cadre asymptotique fondé sur le plan de sondage

Premièrement, nous établissons la normalité asymptotique de $\hat{Y}_{\text{reg},1}$ et $\hat{Y}_{\text{reg},2}$ sous échantillonnage répété, c'est-à-dire quand y_i et x_i sont fixes pour $i \in U$, et S_j est un échantillon PPT aléatoire.

Théorème 1 Supposons que S_1 et S_2 sont des échantillons PPT indépendants tirés avec remise de U_1 et U_2, respectivement, où l'unité $i \in U_j$ possède la probabilité $p_{ij} = z_i \big/ \sum_{i \in U_j} z_i > 0$ d'être sélectionnée, et le poids d'échantillonnage $\pi_i^{-1} = 1 \big/ \left(n_j p_{ij}\right)$ pour $j = 1, 2$, et que les quatre conditions qui suivent sont vérifiées, à mesure que l'indice séquentiel de population m tend vers ∞.

(C1) Il existe des constantes φ_j et ω_j telles que $\sqrt{n/n_j} \to \varphi_j$ et $N_j/N \to \omega_j$.

(C2) Pour $j = 1, 2,$ il existe des constantes μ_{yj}, μ_{xj} et β_j telles que

$$\bar{Y}_j = Y_j/N_j = \sum_{i \in U_j} y_i/N_j \to \mu_{yj}, \bar{X}_j = X_j/N_j = \sum_{i \in U_j} x_i/N_j \to \mu_{xj}$$

existent, de même que les limites $N_j^{-1} \sum_{i \in U_j} (x_i - \mu_{xj})^2 \to \sigma_{xj}^2 > 0,$ et en outre,

$$\left(\sqrt{n_j}/N_j\right) \sum_{i \in U_j} x_i \left(y_i - Y_j/N_j - \beta_j \left(x_i - X_j/N_j\right)\right) \to 0 \text{ quand } n, N \to \infty.$$

(C3) Les limites $D_{N_j} = \sum_{i \in U_j} p_{ij} b_{ij} b_{ij}^T / N_j^2 \to D_j$ existent, où pour $i \in U_j$,

$$b_{ij} = \left[1/p_{ij} - N_j, x_i/p_{ij} - X_j, y_i/p_{ij} - Y_j\right]^T,$$

v^T désigne la transposée vectorielle, et D_j est définie positive. La limite $\sigma_{xe,j}^2 = \lim N_j^{-2} \sum_{i \in U_j} (x_i - \mu_{xj})^2 (y_i - \alpha_j - \beta_j x_i)^2 / p_{ij}$ existe aussi, pour $\alpha_j = \mu_{yj} - \beta_j \mu_{xj}$.

(C4) Les éléments de $\Lambda_j = \sum_{i \in U_j} p_{ij} c_{ij} c_{ij}^T / N_j^4$ forment une séquence bornée, où pour $i \in U_j$, $c_{ij} = \left[(1/p_{ij} - N_j)^2, (x_i/p_{ij} - X_j)^2, (y_i/p_{ij} - Y_j)^2\right]^T$.

Alors, quand $m \to \infty,$ les conclusions qui suivent sont vérifiées.

(a) Pour $j = 1, 2, \hat{\mu}_{xj} \to_P \mu_{xj}, \hat{\mu}_{yj} \to_P \mu_{yj}, \hat{\beta}_j \to_P \beta_j, \hat{\alpha}_j \to_P \alpha_j,$ et $\hat{\sigma}_{xj}^2 \to_P \sigma_{xj}^2,$ où $\to_P$ désigne la convergence en probabilité.

(b) L'estimateur pour la strate combinée $\hat{\beta}$ possède l'expression exacte

$$\hat{\beta} = \frac{\sum_{j=1}^2 \hat{\beta}_j \hat{\sigma}_{xj}^2 \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)\left(\hat{Y}_2 - \hat{Y}_1\right) \hat{N}_1 \hat{N}_2 / \left(\hat{N}_1 + \hat{N}_2\right)}{\sum_{j=1}^2 \hat{\sigma}_{xj}^2 \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)^2 \hat{N}_1 \hat{N}_2 / \left(\hat{N}_1 + \hat{N}_2\right)} \tag{2.1}$$

et la limite en probabilité

$$\beta = \frac{\sum_{j=1}^2 \beta_j \sigma_{xj}^2 \omega_j + (\mu_{x2} - \mu_{x1})(\mu_{y2} - \mu_{y1}) \omega_1 \omega_2}{\sum_{j=1}^2 \sigma_{xj}^2 \omega_j + (\mu_{x2} - \mu_{x1})^2 \omega_1 \omega_2}.$$

(c) $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right) \to_d N\left(0, \sigma_{xe,j}^2 / \sigma_{x,j}^4\right),$ où $\to_d$ désigne la convergence en loi, et $\hat{\sigma}_{xe,j}^2 \to_P \sigma_{xe,j}^2$.

(d) Pour $k = 1, 2$,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},k} - Y\right)/N \to_d N\left(0, \sigma_k^2\right) \tag{2.2}$$

où $\sigma_k^2 = \sum_{j=1}^{2} a_{kj}^T D_j a_{kj}$ et

$$a_{1j} = \omega_j \varphi_j \left[-\left(\mu_y - \beta \mu_x\right), -\beta, 1\right]^T, \quad a_{2j} = \omega_j \varphi_j \left[-\left(\mu_{yj} - \beta_j \mu_{xj}\right), -\beta_j, 1\right]^T,$$

$\mu_x = \omega_1 \mu_{x1} + \omega_2 \mu_{x2}$, $\mu_y = \omega_1 \mu_{y1} + \omega_2 \mu_{y2}$, et D_j est donnée dans la condition (C3).

Les conditions (C1) à (C4) du théorème 1 fournissent une formulation générale du cadre de superpopulation pour l'inférence statistique sous le plan de sondage en grand échantillon, dans laquelle les coefficients de régression selon l'enquête estiment des paramètres descriptifs bien définis de la population servant de base de sondage. Les résultats des parties (a) à (b) montrent que les limites en probabilité β_j, α_j de $\hat{\beta}_j, \hat{\alpha}_j$ possèdent l'interprétation classique de pentes et d'ordonnées à l'origine de droites des moindres carrés de superpopulation. (Ces paramètres de pente et d'ordonnée à l'origine conservent aussi leur interprétation sous un modèle habituelle sous le modèle (2.7) présenté à la section 2.2.) La théorie asymptotique pour $\hat{\beta}_j$ dans la conclusion (c) nous permet de déduire le comportement en grand échantillon de $\hat{Y}_{\text{dec}}$ à partir de celui fourni dans (d) pour $\hat{Y}_{\text{reg},k}$.

Sous les conditions supplémentaires

$$\beta_1 = \beta_2, \alpha_1 = \alpha_2, \tag{2.3}$$

il découle clairement de la partie (b) du théorème 1 que $\beta_j = \beta$, et $\sigma_1^2 = \sigma_2^2$ dans (2.2), de sorte que $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$ sont tous les trois asymptotiquement les mêmes jusqu'à des restes d'ordre plus faible que $N/\sqrt{n}$, comme nous allons le montrer maintenant. En outre, si $\beta_1 \neq \beta_2$, alors $\hat{Y}_{\text{reg},2} - \hat{Y}_{\text{dec}}$ continue d'être $o_P\left(N/\sqrt{n}\right)$, et le test d'égalité des pentes aboutit au rejet, c.-à-d. $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$, et par conséquent $\hat{Y}_{\text{dec}}$ suit la même loi asymptotique que $\hat{Y}_{\text{reg},2}$, qui est plus efficace que $\hat{Y}_{\text{reg},1}$ selon le résultat de la section 2.2.

Théorème 2 Supposons que l'on formule les mêmes hypothèses (C1) à (C4) que pour le théorème 1.

(a) Quand la condition (2.3) est vérifiée, alors quand $m \to \infty$

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1\right) \to_d N\left(0, \sigma_d^2\right), \qquad \sigma_d^2 = \sum_{j=1}^{2} \frac{\sigma_{xe,j}^2}{\varphi_j^2 \sigma_{xj}^4}, \tag{2.4}$$

et les estimateurs $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$ suivent tous une loi asymptotiquement normale et sont équivalents au sens où

$$\frac{n}{N^2}\left[\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right)^2 + \left(\hat{Y}_{\text{reg},2} - \hat{Y}_{\text{dec}}\right)^2\right] \to_P 0. \tag{2.5}$$

(b) Quand $\beta_1 \neq \beta_2$, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$ et $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \to_d N\left(0, \sigma_2^2\right)$.

Une étude plus perfectionnée du comportement asymptotique des estimateurs $\hat{Y}_{\text{dec}}$ peut être entreprise dans l'esprit de Saleh (2006), comme dans le cas des versions contiguës ou de Pitman pour les modèles statistiques hors du contexte des sondages, en supposant que $\sqrt{n}(\beta_1 - \beta_2) \to r$ pour une constante r. Sous cette hypothèse, on peut montrer que $\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2} = o_P\left(N/\sqrt{n}\right)$ et, par conséquent, que les trois estimateurs centrés et réduits $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)$, $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)$ et $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)$ suivent tous la même loi normale asymptotique de moyenne 0. En outre,

$$P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to \Phi\left(-z_{\tau/2} + r/\sigma_d\right) + \Phi\left(-z_{\tau/2} - r/\sigma_d\right), \tag{2.6}$$

où σ_d^2 est donné dans (2.4), et $z_{\tau/2}$ et Φ sont, repectivement, le point de pourcentage et la fonction de répartition de la loi normale centrée réduite. Donc, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$ possède une limite différente de 1. En particulier, dans (2.6), la limite est égale à τ quand $\beta_1 = \beta_2$ (c.-à-d. quand $r = 0$).

2.2 Cadre asymptotique assisté par modèle

À la présente section, nous examinons le comportement des estimateurs $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ sous le modèle probabiliste hypothétique selon lequel les triplets (x_i, y_i, z_i) dans la population finie, $i \in U_j$, sont indépendants et identiquement distribués (iid), où les variables de taille $z_i > 0$ sont utilisées pour définir les probabilités de sélection PPT avec remise $p_{ij} = z_i \Big/ \sum_{i' \in U_j} z_{i'}$, et où x_i et y_i suivent le modèle

$$y_i = \alpha_j + \beta_j x_i + \varepsilon_i,\ i \in U_j, \tag{2.7}$$

avec α_j et β_j représentant les paramètres ordonnée à l'origine et pente inconnus pour la régression dans la strate U_j. Nous supposons que les erreurs $\varepsilon_i, i \in U_j$, sont iid de moyenne 0 et de variance finie σ_ε^2, et qu'elles sont indépendantes de (x_i, z_i), et que les variables x_i pour $i \in U_j$ ont une variance finie. En outre, pour permettre l'échantillonnage PPT, nous supposons que $\max_{i \in U_j} n_j p_{ij} < 1$ avec la probabilité s'approchant de 1 quand m est grand, c.-à-d. quand n_j, N_j sont grands.

À la présente section, les propriétés asymptotiques des estimateurs $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ sont considérées en regard du modèle de régression et de l'échantillonnage répété. En vertu du théorème 1, les estimateurs assistés par modèle $\hat{Y}_{\text{reg},1}$ et $\hat{Y}_{\text{reg},2}$ sont encore convergents et asymptotiquement normaux pour les triplets (x_i, y_i, z_i) iid à l'intérieur des strates, puisque les conditions (C1) à (C4) sont satisfaites sous les hypothèses de moments sur $z_i, 1/z_i$, même si le modèle (2.7) est incorrect. Cependant, les estimateurs $\hat{Y}_{\text{reg},k}$ sont efficaces quand le modèle (2.7) est correct.

Théorème 3 Supposons que l'on a le modèle (2.7) ainsi que la condition (C1), avec $E\left(x_i^4\right) < \infty, E\left(\varepsilon_i^4\right) < \infty, E\left(z_i\right) < \infty,$ et $E\left(\left(1 + x_i^4\right)/z_i^3\right) < \infty$. Alors, toutes les conclusions du théorème 1 et du théorème 2 sont encore vérifiées. En particulier, quand $\beta_1 \neq \beta_2, \sigma_1^2,$ la variance

asymptotique de $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N$ est plus grande que σ_2^2, la variance asymptotique de $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N$. En outre,

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \rightarrow_d N\left(0, (1-\pi)\sigma_1^2 + \pi\sigma_2^2\right), \tag{2.8}$$

où π est la limite de $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$.

Notons que, dans (2.8), π est égal à 1 quand $\beta_1 \neq \beta_2$ et égal à τ quand $\beta_1 = \beta_2$.

Selon le théorème 3, sous le modèle (2.7), les trois estimateurs définis dans (1.2) à (1.4) ont tous la même efficacité asymptotique quand $\alpha_1 = \alpha_2$ et $\beta_1 = \beta_2$ (condition (2.3)). De surcroît, $\hat{Y}_{\text{reg},1}$ est asymptotiquement pire que $\hat{Y}_{\text{reg},2}$ quand $\beta_1 \neq \beta_2$. Donc, pourquoi n'utiliserions-nous pas systématiquement $\hat{Y}_{\text{reg},2}$?

Les assertions du théorème 3 sont des résultats asymptotiques d'ordre un. Un résultat asymptotique d'ordre deux, plus affiné, sous les conditions du théorème 3 et la condition (2.3) quand les tailles z_i sont toutes égales est que, jusqu'à un terme d'ordre $n_1^{-2} + n_2^{-2}$,

$$\text{eqm}\left(\frac{\hat{Y}_{\text{reg},1}}{N}\right) - \frac{\sigma_\varepsilon^2}{n} \leq \left[\text{eqm}\left(\frac{\hat{Y}_{\text{reg},2}}{N}\right) - \frac{\sigma_\varepsilon^2}{n}\right]\left[1 - \frac{n_1 n_2 \left(\bar{X}_1 - \bar{X}_2\right)^2}{nD_n}\right], \tag{2.9}$$

où l'eqm est l'erreur quadratique moyenne conditionnellement aux x_i, $\bar{X}_j = N_j^{-1}\sum_{i \in U_j} x_i$, et

$$D_n = \sum_{j=1}^{2}\sum_{i \in U_j}\left(x_i - \bar{X}_j\right)^2 + \frac{n_1 n_2 \left(\bar{X}_1 - \bar{X}_2\right)^2}{n}.$$

Le résultat (2.9) indique que, lorsque les poids sont égaux et que $\beta_1 = \beta_2$ et $\alpha_1 = \alpha_2$, la performance en échantillon fini de $\hat{Y}_{\text{reg},1}$ pourrait être meilleure que celle de $\hat{Y}_{\text{reg},2}$ pour des valeurs modérées de n_1 et n_2. Voir les résultats des simulations à la section 4. La preuve de (2.9) est un cas particulier d'un résultat plus général donné dans Slud (2012) et est donc omise.

Dans les applications, nous ne savons pas si $\beta_1 = \beta_2$. Donc, l'estimateur fondé sur un test de décision $\hat{Y}_{\text{dec}}$ est une procédure adaptative pour sélectionner un bon estimateur. Compte tenu de (2.8), la performance de $\hat{Y}_{\text{dec}}$ est proche (un peu moins bonne) de celle de $\hat{Y}_{\text{reg},2}$ quand $\beta_1 \neq \beta_2$, et est proche (un peu moins bonne) de celle de $\hat{Y}_{\text{reg},1}$ quand $\alpha_1 = \alpha_2$ et $\beta_1 = \beta_2$. Ces constatations sont également corroborées par les résultats des simulations à la section 4.

3 Estimation de la variance

Il est d'usage de communiquer une estimation de la variance ou de l'erreur-type pour chaque estimation d'après des données d'enquête. L'estimation de la variance est également essentielle pour l'inférence statistique lorsqu'on établit un intervalle de confiance pour un paramètre d'intérêt inconnu.

Les résultats asymptotiques de la section 2 suggèrent un estimateur de variance pour $\hat{Y}_{\text{reg},k}$ obtenu en substituant dans (2.2) des estimateurs pour les quantités inconnues dans σ_k^2. Puisque la variance totale est une somme de H variances intrastrate, sans perte de généralité, nous considérons une strate $(H = 1)$. Pour $j = 1, 2,$ soit

$$\hat{D}_{n_j} = \sum_{i \in S_j} \frac{\hat{b}_{ij}\hat{b}_{ij}^T}{(n_j - 1)\hat{N}_j}, \quad \hat{b}_{ij} = \left[1/p_{ij} - \hat{N}_j, x_i/p_{ij} - \hat{X}_j, y_i/p_{ij} - \hat{Y}_j\right]^T, \quad i \in S_j,$$

$$\hat{a}_{1j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-\left(\bar{y}_j - \hat{\beta}_j \bar{x}_j\right), -\hat{\beta}_j, 1\right]^T, \quad \hat{a}_{2j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-\left(\bar{y} - \hat{\beta}\bar{x}\right), -\hat{\beta}, 1\right]^T,$$

$$\bar{y}_j = \hat{Y}_j/\hat{N}_j, \quad \bar{x}_j = \hat{X}_j/\hat{N}_j, \quad \bar{y} = \sum_{j=1}^{2} \hat{Y}_j/\left(\hat{N}_1 + \hat{N}_2\right), \quad \bar{x} = \sum_{j=1}^{2} \hat{X}_j/\left(\hat{N}_1 + \hat{N}_2\right).$$

Alors, sous les conditions du théorème 1,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{2} \hat{a}_{kj}^T \hat{D}_{n_j} \hat{a}_{kj} \to_P \sigma_k^2, \quad k = 1, 2.$$

C'est-à-dire que $\hat{\sigma}_k^2$ est convergent pour σ_k^2. Les résultats des théorèmes 2 et 3 montrent aussi que $\hat{\sigma}_2^2$ est un estimateur de variance convergent pour l'estimateur fondé sur un test de décision $\hat{Y}_{\text{dec}}$, parce que nous avons soit $\sigma_1^2 = \sigma_2^2$ soit $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$.

Cependant, ces estimateurs de variance obtenus par substitution peuvent ne pas donner d'aussi bons résultats lorsque la valeur de n_1 ou de n_2 est modérée (voir la section 4). Une autre méthode est celle du bootstrap proposée par Cheng et coll. (2010). Soit $\hat{\theta}$ l'estimateur pris en considération. L'estimateur bootstrap de sa variance peut être obtenu comme il suit.

1. Tirer un échantillon bootstrap S_j^* de taille n_j par échantillonnage aléatoire simple avec remise à partir de S_j, où S_1^* et S_2^* sont obtenus de manière indépendante. S'il existe k_j unités autoreprésentatives (AR) dans S_j, comme il est discuté à la section 4.1 qui suit, on tire alors des échantillons de tailles $n_j - k_j$ avec remise, avec $j = 1, 2$.

2. Utiliser les poids de sondage et les données observées provenant de l'ensemble de données originales pour former un ensemble de données bootstrap $S_1^* \cup S_2^*$. À partir de cet ensemble de données, calculer l'analogue bootstrap $\hat{\theta}^*$ de $\hat{\theta}$.

3. Répéter indépendamment les étapes qui précèdent B fois pour obtenir $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$. La variance d'échantillon de $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ est l'estimateur bootstrap de la variance de $\hat{\theta}$.

Sous les conditions des théorèmes 1 et 2, les estimateurs bootstrap de la variance de $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$ sont des estimateurs convergents. La preuve pour le bootstrap est similaire aux preuves des théorèmes et est donc omise.

4 Résultats des simulations pour $H = 1$

La théorie en grand échantillon présentée plus haut ne convient pas pour indiquer si les résultats asymptotiques décrivent adéquatement le comportement des estimateurs $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$, et des estimateurs de leur variance dans des échantillons de taille modérée, ou si $\hat{Y}_{\text{reg},1}$ et $\hat{Y}_{\text{dec}}$ fournissent jamais des améliorations utiles de l'erreur quadratique moyenne dans des échantillons de taille moyenne. Nous présentons certains résultats de simulations pour étudier ces questions, ainsi que les problèmes de petit échantillon qui se posent lorsqu'on applique ces méthodes dans le contexte de l'ASPEP.

Dans les simulations, les valeurs dans la population servant de base de sondage U sont soit générées sous un modèle soit tirées des recensements des administrations publiques de 2002 et 2007 en utilisant les poids de sondage de l'ASPEP de 2007. Le premier jeu de simulations (présenté dans les tableaux 4.1 à 4.6) résume le comportement moyen sur de nombreuses populations servant de bases de sondage générées par un modèle. Dans le deuxième jeu de simulations portant sur des données artificielles, résumé au tableau 4.8, la population servant de base de sondage demeure fixe tout au long de la simulation. Toutes les populations servant de bases de sondage sont constituées d'une seule strate ($H = 1$) décomposée en deux sous-strates ($j = 1, 2$) selon que la valeur d'une variable de taille se situe en-dessous ou au-dessus d'un quantile particulier, habituellement le quantile 0,8. Dans toutes les simulations décrites à la présente section, l'échantillonnage des populations servant de bases de sondage est effectué selon un plan PPT avec remise.

4.1 Considérations concernant les petits échantillons

Avant de décrire les simulations, nous discutons de certaines caractéristiques particulières de l'échantillonnage PPT avec remise (PPTAR) qui, lorsqu'il est appliqué dans des conditions où les échantillons sont petits et les variables de taille ne sont pas équilibrées, requiert une approche de calcul spéciale. Des résultats numériquement irréguliers peuvent être obtenus lorsque les petits échantillons sélectionnés sont utilisés par strate, puis soumis au bootstrap pour estimer les variances.

Les poids $\pi_i^{-1} = 1/(n_j p_{ij})$ dans l'échantillonnage PPTAR ne sont tous supérieurs à 1 que si les probabilités de tirage simples $p_{ij} = z_i / \sum_{i' \in U_j} z_{i'}$ sont toutes inférieures à $1/n_j$. Pour éviter les résultats anormaux en petit échantillon et pour que les plans PPTAR imitant les plans PPT sans remise demeurent pertinents, toute unité $i \in U_j$ avec $n_j p_{ij} \geq 1$ est rendue *autoreprésentative* (AR), c.-à-d. qu'elle est échantillonnée avec certitude mais une seule fois, et si ces unités sont au nombre de k_j, alors les probabilités $\{p_{ij} : i \in U_j, n_j p_{ij} < 1\}$ sont renormalisées pour tirer un échantillon PPTAR de taille $n_j - k_j$. S'il reste des probabilités renormalisées $\geq 1/(n_j - k_j)$, leurs unités deviennent aussi autoreprésentatives et une nouvelle renormalisation est effectuée. Ce processus est répété aussi souvent qu'il est nécessaire. Donc, les petits échantillons dont les distributions des variables de taille sont très inégales pourraient ne pas être compatibles avec l'échantillonnage PPTAR, situation qui se présente dans certains cas de données réelles de l'ASPEP examinés plus loin.

Nous aurions pu faire un autre choix, mais nous nous conformons à la pratique de l'ASPEP consistant à inclure toutes les unités autoreprésentatives dans l'ajustement des estimateurs par régression pondérés

par les poids de sondage $\hat{\beta}_2$ et $\hat{\beta}$. Cependant, sous ce choix, l'échantillonnage PPTAR suivi par le rééchantillonnage bootstrap des petits échantillons peut donner lieu à un comportement très imprévisible, qui doit être reconnu quand on résume le comportement des estimateurs bootstrap de la variance. Le problème tient au fait que, quand un petit nombre m d'unités non autoreprésentatives sont échantillonnées selon un plan PPTAR, en plus d'un ensemble d'unités autoreprésentatives, puis sont traitées par la méthode du bootstrap, la probabilité que l'échantillon bootstrap contienne seulement une unité non autoreprésentative unique peut être étonnamment grande, ce qui donne lieu à une très forte variabilité du bootstrap. Ce phénomène a été observé dans les simulations présentées plus loin, pour une sous-strate de grande taille contenant 20 éléments ou moins et des variables de taille ayant une distribution très asymétrique, dans les cas de variables x_i lognormales ou de l'ASPEP.

4.2 Données artificielles générées par un modèle

Toutes les populations artificielles servant de bases de sondage ont été générées au moyen de N = 2 000 triplets (x_i, y_i, z_i) iid satisfaisant l'équation (2.7), pour U_1 constituée des N_1 = 1 600 d'entre-eux pour lesquels la valeur de x_i était inférieure à leur 80^e percentile empirique $c = \left(x_{(1,600)} + x_{(1,601)}\right)/2$, et U_2 constituée des 400 autres. Dans la plupart des cas, les variables z_i ont été générées comme $N(30 + x_i, 100)$ variables conditionnées pour qu'elles soient positives (ce qui a nécessité à l'occasion une resimulation dans les modèles lognormaux de x_i ci-dessous) et étaient conditionnellement indépendantes de y_i sachant x_i. (Cependant, dans certains cas, des échantillons non pondérés ont été tirés en prenant les z_i identiquement égales.) Des échantillons PPT avec remise stratifiés de tailles $(n_1, n_2) = (100, 50), (100, 20)$, ou $(50, 20)$ ont été tirés dans des exécutions de simulation successives, en utilisant les variables de taille z_i, à partir de la même base de sondage.

Les modèles générant (x_i, y_i) sont indexés comme il suit. Dans les modèles dont le préfixe est **M1**, les variables indépendantes x_i suivent une loi Gamma $(4; 0,1)$ dont le quantile 0,8 est égal à 55,2, tandis que dans les modèles **M2**, les variables x_i suivent une loi lognormale (1;6,25) dont le quantile 0,8 est égal à 22,3. Les populations **M1**, et les modèles **M2** avec le suffixe **E** ont une variance conditionnelle de 100 pour y_i sachant x_i, tandis que les modèles **M2** sans le suffixe **E** ont une variance conditionnelle de $20x_i$. Les moyennes conditionnelles $E(y_i | x_i)$ sont toutes linéaires, égales à $20 + 1,5x_i$ dans les modèles indicés **H0** et égales à $20 + x_i + 0,5(x_i - c)I_{[j=2]}$ dans la sous-strate U_j dans les modèles **H1**. Les ordonnées à l'origine des modèles de régression sont choisies de manière que les droites se coupent à $x = c$, que les pentes soient égales ou non (voir la discussion à la section 1). Le tableau 4.1 donne la moyenne et l'écart-type (É.T.) pour les totaux Y générés à partir des attributs de la population servant de base de sondage $\{y_i\}_{i=1}^{2\,000}$ sous les divers modèles. Les variables x_i ainsi que les totaux Y ont une distribution à queue plus longue sous les modèles lognormaux.

Tableau 4.1
Moyennes et écarts-types des totaux Y sous les modèles de simulation

	Gamma		Lognormaux			
Modèle	**M1.H0**	**M1.H1**	**M2.H0**	**M2.H0E**	**M2.H1**	**M2.H1E**
E(Y)	160 000	123 177	225 603	225 603	173 485	173 485
É.T.(Y)	1 414,2	653,5	94 380	94 368	62 362	62 344

Modèles de population simulés

M1.H0 : $x_i \sim \text{Gamma}(4;0,1)$ (paramètre de forme 4, paramètre d'échelle 10), $y_i \sim N(20 + 1,5x_i;100)$ (variance 100), tout $i \in U$.

M1.H1 : $x_i \sim \text{Gamma}(4;0,1)$, $y_i \sim N\left(20 + x_i + 0,5(x_i - c)I_{[x_i \geq c]};100\right)$, tout i.

M2.H0 : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N(20 + 1,5x_i;20x_i)$, tout i.

M2.H0E : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N(20 + 1,5x_i;100)$, tout i.

M2.H1 : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N\left(20 + x_i + 0,5(x_i - c)I_{[x_i \geq c]};20x_i\right)$, tout i.

M2.H1E : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N\left(20 + x_i + 0,5(x_i - c)I_{[x_i \geq c]};100\right)$, tout i.

Les résultats des simulations et les résultats bootstrap présentés dans les tableaux 4.2 à 4.5 ont été générés suivant le plan de sondage et de présentation des résultats qui suit. Pour chaque type de population, 60 populations servant de bases de sondage distinctes ont été générées, et 50 expériences d'échantillonnage indépendantes ont été exécutées avec chacune de ces populations. Dans les cas où les résultats de l'échantillonnage pondéré et non pondéré ont été comparés, ces échantillons ont été tirés indépendamment l'un de l'autre à partir du même ensemble de 60 populations servant de bases de sondage. Donc, on disposait de 3 000 répliques indépendantes pour le calcul de la moyenne Monte Carlo des résultats statistiques, pour trois tailles d'échantillon stratifiées différentes, et 400 itérations bootstrap ont été effectuées pour chaque échantillon généré de cette façon.

Tableau 4.2
É.T. empiriques et estimés et couverture de l'IC, d'après les simulations du modèle M1

		M1.H0			M1.H1		
Tailles	**Stat.**	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$
100,50	É.T._{MC}	1 785,5	1 794,3	1 788,0	1 817,6	1 773,5	1 774,4
	É.T._{S}	1 757,1	1 751,5	1 755,6	1 794,6	1 735,2	1 735,8
	É.T._{B}	1 752,4	1 762,0	1 758,4	1 788,1	1 742,9	1 747,0
	PC_{S}	94,47	94,37	94,50	93,93	93,73	93,77
	PC_{B}	94,60	94,53	94,67	93,93	94,03	94,07
100,20	É.T._{MC}	1 930,0	1 944,8	1 934,0	2 008,4	1 944,4	1 960,4
	É.T._{S}	1 888,3	1 876,6	1 884,1	1 944,4	1 861,0	1 866,5
	É.T._{B}	1 878,8	1 901,4	1 895,8	1 936,1	1 885,6	1 897,9
	PC_{S}	94,20	93,83	94,13	93,53	93,20	93,07
	PC_{B}	93,80	94,00	93,97	93,60	93,83	93,97
50,20	É.T._{MC}	2 583,5	2 610,7	2 593,5	2 591,3	2 522,8	2 535,4
	É.T._{S}	2 509,2	2 490,8	2 505,1	2 562,2	2 465,0	2 474,5
	É.T._{B}	2 498,5	2 538,0	2 522,9	2 550,3	2 508,5	2 525,6
	PC_{S}	93,70	93,13	93,57	93,97	93,63	93,43
	PC_{B}	93,63	93,73	93,87	93,83	93,77	94,10

Tableau 4.3
É.T. empiriques et estimés et couverture de l'IC, d'après les simulations du modèle M2

		M2.H0			M2.H1		
Tailles	**Stat.**	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$
100,50	É.T._{MC}	3 400,1	3 475,4	3 406,8	3 481,9	3 483,8	3 482,2
	É.T._{S}	3 420,6	3 400,0	3 417,0	3 537,8	3 405,0	3 463,7
	É.T._{B}	3 590,0	3 715,2	3 623,4	3 852,0	3 921,9	3 898,4
	PC_{S}	95,10	93,43	94,83	95,03	93,40	94,13
	PC_{B}	95,67	95,77	95,77	95,63	95,77	95,70
100,20	É.T._{MC}	5 655,2	6 184,0	5 698,6	5 853,0	6 181,1	5 955,6
	É.T._{S}	5 644,9	5 575,7	5 640,9	5 798,3	5 587,3	5 697,3
	É.T._{B}	5 565,1	6 687,3	5 857,8	5 907,8	6 838,0	6 466,6
	PC_{S}	93,83	88,47	93,40	92,77	88,30	90,70
	PC_{B}	92,33	93,67	93,37	92,63	94,33	94,17
50,20	É.T._{MC}	5 773,2	6 319,2	5 833,9	5 934,2	6 230,6	6 009,8
	É.T._{S}	5 800,2	5 677,2	5 785,8	6 012,6	5 755,4	5 919,2
	É.T._{B}	5 728,5	6 825,2	6 086,0	6 102,2	6 978,1	6 522,1
	PC_{S}	94,60	88,67	93,97	94,07	89,37	92,27
	PC_{B}	93,40	94,23	94,27	93,47	95,03	94,80

Tableau 4.4
É.T. pour $\hat{Y}_{\text{HT}}$ *vs* $\hat{Y}_{\text{dec}}$, et couverture des intervalles de confiance percentiles bootstrap pour $\hat{Y}_{\text{dec}}$, pour $\tau = 0,05$ *vs* $0,20$, pour les modèles M1 et M2, H0 et H1

		$\hat{Y}_{\text{dec}}, \tau = 0,05$		$\hat{Y}_{\text{HT}}$	$\hat{Y}_{\text{dec}}, \tau = 0,20$	
Modèle	**Échantillons**	É.T_{MC}	PC_{PB}	É.T._{HT}	É.T_{MC}	PC_{PB}
M1.H0	100,50	1 788,0	94,23	2 774,0	1 745,5	94,60
	100,20	1 934,0	93,50	3 032,6	1 915,9	94,10
	50,20	2 593,5	93,17	3 000,7	2 500,1	94,43
M1.H1	100,50	1 774,4	93,70	2 387,3	1 737,3	94,43
	100,20	1 960,4	93,27	2 678,9	1 948,0	93,23
	50,20	2 535,4	93,90	3 035,0	2 509,8	94,23
M2.H0	100,50	3 406,8	95,20	4 160,0	3 398,8	94,83
	100,20	5 698,6	91,13	6 720,2	5 705,7	92,57
	50,20	5 833,9	92,60	7 080,0	5 979,8	92,17
M2.H1	100,50	3 482,2	95,13	4 393,6	3 423,9	94,03
	100,20	5 955,6	92,07	7 413,1	5 917,3	92,40
	50,20	6 009,8	92,33	7 840,4	6 105,6	92,17

Tableau 4.5
Comparaisons des estimations de l'É.T. et de la couverture de l'IC pour H0 et H1 pour trois modèles lognormaux, pondérés (W) et non pondérés (U) dans M2, et pondérés (E) dans M2.E. Les couvertures en % des IC sont données pour les É.T. ainsi que les intervalles percentiles bootstrap

Modèle	Taille	Stat	É.T.	$\widehat{\text{É.T.}}_S$	$\widehat{\text{É.T.}}_B$	PC_S	PC_B	PC_{PB}
H0.W	100,50	$\hat{Y}_{reg,1}$	3 400,1	3 420,6	3 590,0	95,10	95,67	94,93
		$\hat{Y}_{reg,2}$	3 475,4	3 400,0	3 715,2	93,43	95,17	95,33
		$\hat{Y}_{dec}$	3 406,8	3 417,0	3 623,4	94,83	95,77	95,20
H0.U		$\hat{Y}_{reg,1}$	5 481,6	3 674,8	5 571,9	81,43	93,50	92,07
		$\hat{Y}_{reg,2}$	5 782,8	3 646,6	6 076,3	80,13	93,67	91,90
		$\hat{Y}_{dec}$	5 525,5	3 669,0	5 726,8	81,07	93,83	92,20
H0.E		$\hat{Y}_{reg,1}$	1 888,8	1 930,1	1 904,7	94,73	94,53	94,23
		$\hat{Y}_{reg,2}$	1 888,6	1 911,1	1 893,2	94,43	94,30	94,20
		$\hat{Y}_{dec}$	1 892,9	1 926,5	1 905,0	94,67	94,57	94,20
H0.W	50,20	$\hat{Y}_{reg,1}$	5 773,2	5 800,2	5 728,5	94,60	93,40	92,00
		$\hat{Y}_{reg,2}$	6 319,2	5 677,2	6 825,2	88,67	94,23	92,60
		$\hat{Y}_{dec}$	5 833,9	5 785,8	6 086,0	93,97	94,27	92,60
H0.U		$\hat{Y}_{reg,1}$	10 000,3	5 136,5	9 905,6	71,10	90,73	89,80
		$\hat{Y}_{reg,2}$	11 192,8	5 085,0	12 806,8	68,70	92,90	89,37
		$\hat{Y}_{dec}$	10 134,1	5 120,7	11 245,9	70,73	92,37	90,27
H0.E		$\hat{Y}_{reg,1}$	2 811,4	2 831,6	2 769,5	94,13	94,00	93,93
		$\hat{Y}_{reg,2}$	2 811,9	2 753,8	2 741,1	93,47	93,77	93,30
		$\hat{Y}_{dec}$	2 817,4	2 821,8	2 777,0	93,83	93,90	93,77
H1.W	100,50	$\hat{Y}_{reg,1}$	3 481,9	3 537,8	3 852,0	95,03	95,63	95,27
		$\hat{Y}_{reg,2}$	3 483,8	3 405,0	3 921,9	93,40	95,77	95,10
		$\hat{Y}_{dec}$	3 482,2	3 463,7	3 898,4	94,13	95,70	95,13
H1.U		$\hat{Y}_{reg,1}$	5 631,4	3 774,8	5 614,6	80,90	92,33	91,07
		$\hat{Y}_{reg,2}$	5 838,3	3 699,6	6 010,5	79,13	92,73	91,37
		$\hat{Y}_{dec}$	5 727,0	3 732,8	5 870,5	80,40	92,93	91,63
H1.E		$\hat{Y}_{reg,1}$	2 005,5	2 094,2	2 019,1	95,60	94,97	94,60
		$\hat{Y}_{reg,2}$	1 909,9	1 908,2	1 892,5	94,83	94,77	94,17
		$\hat{Y}_{dec}$	1 931,9	1 941,7	1 934,6	94,97	95,20	94,83
H1.W	50,20	$\hat{Y}_{reg,1}$	5 934,2	6 012,6	6 102,2	94,07	93,47	91,97
		$\hat{Y}_{reg,2}$	6 230,6	5 755,4	6 978,1	89,37	95,03	92,23
		$\hat{Y}_{dec}$	6 009,8	5 919,2	6 522,1	92,27	94,80	92,33
H1.U		$\hat{Y}_{reg,1}$	9 315,8	5 350,9	10 040,0	74,17	93,10	90,57
		$\hat{Y}_{reg,2}$	10 583,8	5 229,6	12 476,8	71,23	94,57	90,87
		$\hat{Y}_{dec}$	9 989,6	5 295,4	11 479,5	72,53	94,33	91,47
H1.E		$\hat{Y}_{reg,1}$	3 096,1	3 137,7	2 795,6	94,63	93,43	93,37
		$\hat{Y}_{reg,2}$	2 880,6	2 766,8	2 745,7	93,10	93,40	93,47
		$\hat{Y}_{dec}$	2 977,3	2 929,2	2 882,0	93,77	93,77	93,77

Nous avons calculé les quantités qui suivent pour chaque combinaison de modèles, pondérations et tailles d'échantillon : les biais en pourcentage de $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{dec}}$ (avec $\tau = 0{,}05$ dans tous les tableaux, sauf le tableau 4.4 où $\tau = 0{,}05$ ou 0,20) en tant qu'estimateurs de Y; les écarts-types (É.T.) Monte Carlo, É.-T._{MC}, de ces trois estimateurs; les É.T. estimés des estimateurs, en utilisant les estimateurs de l'É.T. par substitution $(\widehat{\text{É.T.}}_S)$ et bootstrap $(\widehat{\text{É.T.}}_B)$, respectivement, décrits à la section 3; la probabilité de couverture, PC_u, des intervalles de confiance à 95 % nominaux pour $Y : \hat{Y} \pm 1{,}960 \cdot \widehat{\text{É.T.}}_u$, où $\hat{Y}$ est l'un des trois estimateurs de Y, et $u = S$ ou B; et les intervalles de confiance bootstrap percentiles (et leur pourcentage de couverture PC_{BP}) obtenus d'après les quantiles 0,025 et 0,975 empiriques des (400) valeurs bootstrap de chacun des trois estimateurs $\hat{Y}$ de Y. En outre, nous avons calculé les biais empiriques des estimations de Horvitz-Thompson $\hat{Y}_{\text{HT}}$ dans (1.1) et leurs écarts-types empiriques É.T._{HT}. (De ces quantités calculées, seuls les biais ne sont pas présentés, puisqu'ils étaient tous nettement inférieurs à 0,5 % sauf pour le modèle **M2**.**H1**.**U**, et même dans ce cas, la valeur la plus importante du biais était de l'ordre de 1 %.) Deux autres statistiques, calculées et présentées au tableau 4.6 pour chacun des estimateurs $\hat{Y}$ de Y, sont les erreurs-types sur l'ensemble des populations servant de bases de sondage générées aléatoirement des estimations Monte Carlo et bootstrap intrapopulation des É.T. des estimateurs $\hat{Y}$.

Tableau 4.6
Erreurs-types sur l'ensemble des populations des É.T. empiriques et bootstrap estimés pour les estimateurs $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, et $\hat{Y}_{\text{dec}}$, pour certains modèles et pondérations

		$\hat{Y}_{\text{reg},1}$		$\hat{Y}_{\text{reg},2}$		$\hat{Y}_{\text{dec}}$	
Modèle	**Tailles**	**É.T.**	$\widehat{\text{É.T.}}_B$	**É.T.**	$\widehat{\text{É.T.}}_B$	**É.T.**	$\widehat{\text{É.T.}}_B$
M1.**H0**	100,50	198	35	196	35	197	35
	50,20	210	52	208	51	210	51
M1.**H1**	100,50	204	39	183	40	184	41
	50,20	319	57	298	62	302	62
M2.**H0**	100,50	404	345	450	383	405	351
	50,20	825	518	1,075	916	889	631
M2.**H0**.**E**	100,50	187	49	185	45	184	47
	50,20	294	85	293	71	298	82
M2.**H1**	100,50	409	409	410	421	408	414
	50,20	767	624	946	929	841	730
M2.**H1**.**E**	100,50	208	59	196	46	204	50
	50,20	258	141	261	82	239	102
M2.**H1**.**U**	100,50	1 676	1 351	1 773	1 539	1 726	1 467
	50,20	2 397	2 543	3 425	3 454	3 102	3 159

4.3 Données réelles du recensement des administrations publiques

Nos simulations fondées sur l'échantillonnage répété à partir de bases de sondage contenant des données réelles s'appuient sur un ensemble de données nationales au niveau des États rassemblées par Yang Cheng. La base de sondage de l'ASPEP réalisée auprès des administrations publiques pour l'année de référence 2007, qui était aussi une année de recensement, est la même que celle du fichier du

recensement des administrations publiques (*Census of Governments*) de 2007. Notre ensemble de données contient les valeurs des variables de l'ASPEP de 2002 et de 2007 (nombre d'employés, rémunération et heures travaillées à temps plein et à temps partiel) tirées des recensements de ces années, ainsi que les poids de sondage de 2007 et les variables indicatrices de présence dans l'échantillon pour l'ASPEP. Un poids égal à 1 signifie que l'administration publique en question était autoreprésentative, au sens où elle a été choisie avec certitude en vue d'être incluse dans l'ASPEP. La variable de taille z_i pour l'échantillonnage PPT dans l'ASPEP est égale à la somme des masses salariales à temps plein et à temps partiel provenant du recensement le plus récent, de sorte que nous nous limitons à l'examen des 53 402 administrations publiques figurant dans le fichier pour lesquelles la valeur de cette variable était positive. Le tableau 4.7 donne les administrations publiques de type sous-comté et district spécial (les seules qui sont subdivisées en sous-strates de petites et de grandes unités) dans neuf États, ainsi que les nombres d'unités autoreprésentatives et les nombres d'unités échantillonnées en 2007. Comme il est mentionné à la sous-section 4.1, le nombre final d'unités autoreprésentatives (AR) pour l'échantillonnage PPT avec remise peut dépasser le nombre d'unités sélectionnées initialement en vue d'être incluses avec certitude, et les nombres plus élevés, qui correspondent à la taille de l'échantillon effectivement sélectionné en 2007, sont indiqués dans les colonnes AR du tableau 4.7. L'inspection de ce tableau montre que plusieurs combinaisons État-type d'administration publique ont une population nulle dans une sous-strate ou ne contiennent qu'un nombre trop faible d'unités non autoreprésentatives pour être utile dans la simulation d'échantillons répétés. À titre de règle empirique, nous prenons 15 comme nombre minimal d'unités non autoreprésentatives et nous recommandons que les paires de sous-strates contenant un nombre plus faible d'unités non autoreprésentatives dans la strate des grandes unités soient fusionnées sans recourir à la stratégie fondée sur un test de décision étudiée dans le présent article.

Tableau 4.7
Population de recensement, tailles d'échantillon de l'ASPEP et nombre d'administrations publiques de types sous-comté et district spécial autoreprésentatives par sous-strate en 2007, pour neuf États choisis

	Sous-comté					District spécial				
	Petites unités		Grandes unités			Petites unités		Grandes unités		
	Pop.	Éch.	Pop.	Éch.	AR	Pop.	Éch.	Pop.	Éch.	AR
AL	378	15	55	45	26	0	0	400	102	64
CA	0	0	475	104	86	1 595	39	107	107	107
CO	0	0	265	34	18	627	16	65	55	33
FL	317	16	81	54	36	0	0	330	48	24
GA	461	17	49	36	20	0	0	293	70	32
MO	980	25	101	101	101	799	27	106	66	42
NY	1 473	25	69	69	69	606	16	33	23	4
PA	2 409	55	123	81	31	921	21	37	37	37
WI	1 702	36	129	71	44	281	16	61	40	20

Pour neuf combinaisons d'administration publique par type comprenant 15 unités non autoreprésentatives ou plus et au moins 17 unités non autoreprésentatives non échantillonnées de la strate des grandes unités (sauf pour les États AL, CO, et GA pour lesquels il existait respectivement 9, 10 et 11 unités non autoreprésentatives non échantillonnées), le tableau 4.8 donne les résultats pour les

estimateurs fondés sur un test de décision et les estimations de la variance dans les paires de sous-strates. Dans chacune des combinaisons État-type d'administration publique, 3 000 échantillons PPTAR stratifiés ayant les tailles indiquées ont été tirés de la base de sondage de l'ASPEP et du recensement des administrations publiques décrites plus haut, avec x_i et y_i désignant, respectivement, la masse salariale des employés à temps plein de l'administration publique concernée telle qu'enregistrée aux recensements des administrations publiques de 2002 et de 2007, et z_i désignant la masse salariale totale (temps plein plus temps partiel) en 2002. Pour chaque échantillon simulé, on a calculé les estimateurs $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ et estimé les variances empiriques. La variance de $\hat{Y}_{\text{dec}}$ a également été estimée par les méthodes de la formule de substitution et du bootstrap comme dans les simulations basées sur des données artificielles. (Mais il convient de souligner que, comme il a été décrit plus haut, dans chaque échantillon de sous-strate, les échantillons bootstrap ont été tirés uniquement parmi les unités non autoreprésentatives.) Les résultats sont présentés au tableau 4.8. Les efficacités relatives des estimateurs par la régression stratifiés combinés et distincts peuvent être évaluées d'après le ratio correspondant des É.T. donné dans la colonne 5 du tableau. Les autres É.T. présentés sont les estimateurs empiriques, par substitution et bootstrap de l'écart-type de $\hat{Y}_{\text{dec}}$.

Tableau 4.8
Sommaire des simulations par échantillonnage répété à partir de la base de sondage de l'ASPEP de 2007. La masse salariale totale des employés à temps plein (Y) est exprimée en multiples de 100 millions de dollars, et les estimations de l'É.T. données dans les colonnes 6 à 8 sont exprimées en unités de 1 million de dollars $\hat{Y}_{\text{dec}}$. $\text{É.T.}_1/\text{É.T.}_2$ dans la colonne 5 est le ratio de l'É.T. empirique de $\hat{Y}_{\text{reg},1}$ à celui de $\hat{Y}_{\text{reg},2}$.

État	Strate	Y	Taille	$\text{É.T.}_1/\text{É.T.}_2$	$\widehat{\text{É.T.}}$	$\widehat{\text{É.T.}}_S$	$\widehat{\text{É.T.}}_B$
AL	Sous-comté	1,2	25,46	2,14	4,90	3,67	5,71
CA	Distr. spécial	4,3	30,90	0,98	29,4	21,2	26,8
CO	Distr. spécial	0,6	25,55	1,14	3,77	2,58	3,00
FL	Sous-comté	4,3	25,54	1,16	11,9	9,4	12,2
GA	Sous-comté	1,5	25,38	1,15	4,38	3,26	4,88
MO	Distr. spécial	0,6	40,70	2,13	2,99	2,20	2,99
NY	Sous-comté	23,6	35,52	1,53	13,6	12,0	14,1
PA	Sous-comté	3,0	40,70	1,12	7,28	5,79	7,60
WI	Sous-comté	1,4	40,70	2,06	5,00	4,45	5,17

4.4 Discussion des résultats des simulations

L'exposé qui suit est un résumé et une interprétation des résultats des tableaux, ainsi que d'autres résultats non présentés.

I) Bon nombre des simulations au moyen de données artificielles servent à confirmer les résultats théoriques en grand échantillon des théorèmes. Nous avons déjà mentionné que, dans les tableaux 4.2 et 4.3, les biais des trois estimateurs de Y $\left(\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}\right)$ sont généralement faibles. Dans le tableau 4.2, qui se rapporte aux modèles avec variables indépendantes et poids reliés à la loi Gamma dans les modèles **M1**, les estimateurs de variance par substitution et par bootstrap de chaque estimateur de Y sont assez précis et proches l'un de l'autre, et les intervalles de confiance ont tous une couverture proche de la couverture nominale. Sous **M1**.**H0** ainsi que **M1**.**H1**, pour les plus petites tailles d'échantillon n_2, on

note une tendance des estimateurs $\widehat{\text{É.T.}}_S$ et $\widehat{\text{É.T.}}_B$ à sous-estimer légèrement les écarts-types réels ou empiriques, mais $\widehat{\text{É.T.}}_B$ semble suivre l'écart-type de plus près que $\widehat{\text{É.T.}}_S$ pour $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$.

II) La distribution des valeurs de la variable x_i lognormale dans les modèles **M2** est beaucoup plus dispersée et asymétrique que dans les modèles **M1**, mais les résultats des simulations appuient néanmoins la théorie asymptotique quand $n_2 = 50$, quoique pas si $n_2 = 20$. Les intervalles de confiance de Y fondés sur l'estimateur par substitution en ce qui concerne $\hat{Y}_{\text{reg},2}$ ont une probabilité de couverture beaucoup trop faible lorsque l'on utilise l'estimateur de variance par substitution. Dans le tableau 4.3, pour chaque type d'estimateur de Y, l'estimateur de variance par substitution présente une tendance prononcée à sous-estimer la variance (empirique) réelle et l'estimateur par le bootstrap, à la surestimer.

Le tableau 4.5 clarifie le fait que le comportement extrême des estimateurs de variance sous les modèles **M2** résulte partiellement de ce que les distributions des variables indépendantes et de y_i sont dispersées et asymétriques, et partiellement de ce que la variable de taille utilisée dans les pondérations PPT présente aussi ces propriétés. Les cas désignés par le suffixe **W** dans ce tableau sont les mêmes que dans le tableau 4.3. Les cas portant le suffixe **E** ont les mêmes variables (x_i, z_i) que dans le tableau 4.3, mais les variances conditionnelles de y_i sachant x_i ont la valeur constante de 100; grâce à ce changement, le comportement irrégulier des estimateurs de l'écart-type disparaît. Cependant, lorsque les variances de y_i conditionnelles sont les mêmes que dans le modèle de base **M2**, mais que l'échantillonnage PPTAR est *non* pondéré, c.-à-d. lorsque toutes les variables z_i sont remplacées par la valeur 1, les estimateurs empiriques et bootstrap de l'écart-type sont très proches et très grands, tandis que l'estimateur de variance par substitution est trop faible, et ce d'un facteur spectaculairement grand variant de $1/2$ à $3/4$. Ce phénomène étrange s'observe de la même façon pour les trois estimateurs de Y. (Cependant, une variante de l'échantillonnage non pondéré dans le modèle **M1** ne modifie pas matériellement les résultats par rapport à ceux présentés au tableau 4.2.)

III) Un objectif des simulations était de savoir s'il existe jamais un avantage, en ce qui concerne l'erreur quadratique moyenne (EQM), à utiliser $\hat{Y}_{\text{reg},1}$ plutôt que $\hat{Y}_{\text{reg},2}$, faute de quoi il y aurait fort peu de raisons d'utiliser $\hat{Y}_{\text{dec}}$. En effet, les théorèmes en grand échantillon disent que le terme principal de variance en grand échantillon est toujours optimal pour $\hat{Y}_{\text{reg},2}$ (parce qu'il est le même que pour $\hat{Y}_{\text{reg},1}$ sous l'hypothèse nulle ou parce qu'il est strictement meilleur sous le modèle (2.7) avec des pentes distinctes). Toutefois, nous avons indiqué après le théorème 3, dans la borne (2.9), que $\hat{Y}_{\text{reg},1}$ peut avoir une EQM d'ordre deux plus petite que $\hat{Y}_{\text{reg},2}$, et les colonnes **H0** des tableaux 4.2 et 4.3 révèlent un avantage faible mais consistant de $\hat{Y}_{\text{reg},1}$ par rapport à $\hat{Y}_{\text{reg},2}$ en ce qui concerne l'écart-type, avantage qui est plus prononcé pour **M2**. Cet avantage disparaît sous la version fixe **M1**.**H1**, mais curieusement, pas sous **M2**.**H1**. L'avantage léger, mais réel, de $\hat{Y}_{\text{reg},1}$ en ce qui concerne l'EQM conditionnelle quand les pentes dans les sous-strates sont très proches de l'égalité est discuté plus en détail par Slud (2012).

Les estimateurs $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ considérés ici sont du type régression et il pourrait être intéressant de comparer le comportement de leur EQM dans les populations simulées à celles de l'estimateur plus simple de Horvitz-Thompson $\hat{Y}_{\text{HT}}$ dans (1.1). Tous ces estimateurs sont presque sans biais, de sorte que les EQM sont essentiellement les mêmes que les variances, et une comparaison des troisième et cinquième colonnes du tableau 4.4 montre que les variances de $\hat{Y}_{\text{HT}}$ sont considérablement plus grandes que celles de $\hat{Y}_{\text{dec}}$. La

différence est moins prononcée pour les échantillons de plus grande taille, mais même dans ce cas, elle est de 30 % à 55 %. L'avantage de $\hat{Y}_{\text{dec}}$ reste encore très prononcé dans le modèle **M2**, où les variances sous le modèle et l'asymétrie de la distribution sont plus importantes, mais moins que dans le modèle **M1**.

IV) La définition de $\hat{Y}_{\text{dec}}$ contient le seuil de signification nominal arbitraire τ, qui dans tous les tableaux sauf le tableau 4.4 a été fixé à 0,05. Comme le laisse entendre la théorie en grand échantillon, les propriétés de l'estimateur fondé sur un test de décision sont comprises entre celles de $\hat{Y}_{\text{reg},1}$ et de $\hat{Y}_{\text{reg},2}$, et de plus grandes valeurs de τ rendent $\hat{Y}_{\text{dec}}$ plus souvent égal à $\hat{Y}_{\text{reg},1}$. Comme le montre la comparaison des colonnes 6 et 7 du tableau 4.4, le choix $\tau = 0{,}20$ semble aboutir, dans les modèles simulés, à des écarts-types de $\hat{Y}_{\text{dec}}$ très légèrement plus faibles sous le modèle **M1**, tandis que sous le modèle **M2**, l'écart-type est plutôt plus grand pour les petites tailles d'échantillon. La conclusion est faible, parce que les différences sont relativement petites comparativement aux différences d'écart-type observées d'une population servant de base de sondage à l'autre. Nous préférons laisser une plus petite valeur de τ dicter le groupement fréquent de sous-strates, sauf quand il existe des différences prononcées de pente estimée entre les sous-strates. Cette constatation selon laquelle de plus grands seuils de signification τ n'améliorent pas les propriétés de $\hat{Y}_{\text{dec}}$ diffère de celle de Saleh (2006) voulant que de plus grands seuils de signification soient très avantageux dans d'autres contextes de tests préliminaires.

V) Le tableau 4.6 renseigne sur la variabilité des estimateurs de l'écart-type des estimateurs de Y selon la population servant de base de sondage. Les estimateurs bootstrap de la variance semblent moins susceptibles de varier d'une population servant de base de sondage à l'autre, parce que la moyenne réalisée par le bootstrap les stabilise. Dans ce tableau, la principale constatation semble être que la variabilité entre les populations servant de bases de sondage est modérée, sauf sous le modèle **M2** non pondéré, où elle est remarquablement grande. Ce résultat semble expliquer l'inflation extrême des variances sous **M2**.**U** observées dans le tableau 4.5.

VI) Dans de nombreuses applications bootstrap avec statistique suivant approximativement une loi normale, la mauvaise couverture des intervalles de confiance fondés sur la théorie normale due à la non-normalité de la statistique obtenue par bootstrap peut être atténuée en utilisant les intervalles bootstrap percentiles (BP) (Shao et Tu 1995, section 4.1). Dans les présentes simulations, le tableau 4.4 (colonnes 4 et 6) donne les pourcentages de couverture des intervalles BP pour $\hat{Y}_{\text{dec}}$ dans les conditions où les tableaux 4.2 et 4.3 donnent les couvertures des IC sous la théorie de la loi normale basées sur l'écart-type estimé par bootstrap. Quelle qu'en soit la raison, les tableaux montrent que, sous la théorie de la loi normale, PC_B a systématiquement tendance à être légèrement inférieur à la valeur nominale mais néanmoins légèrement supérieur à la couverture des intervalles BP, PC_{BP}. Donc, nos simulations indiquent que, dans ces conditions, la préférence va à l'intervalle plus simple $\hat{Y}_{\text{dec}} \pm 1,96 \cdot \widehat{\text{É.T.}}_B$.

VII) Il reste à tirer les leçons des simulations portant sur des données réelles du recensement des administrations publiques présentées à la section 4.3. Le premier commentaire qui s'impose est que l'étalement et l'asymétrie de la distribution des variables indépendantes x_i correspondant à la masse salariale des employés à temps plein et de la variable de taille z_i correspondant à la masse salariale totale sont très importants, et ressemblent davantage à ceux observés pour les modèles lognormaux **M2** que pour les modèles Gamma **M1**. Le tableau 4.8 indique (dans la colonne 5) un avantage constant de $\hat{Y}_{\text{reg},2}$ par rapport à $\hat{Y}_{\text{reg},1}$ en ce qui concerne l'EQM, sauf dans le cas CA-district spécial, bien que la différence

soit faible dans le cas CO-district spécial et dans les cas FL, GA et PA-sous-comté. Il convient de souligner que, dans presque tous ces exemples, l'estimateur bootstrap de l'écart-type pour $\hat{Y}_{\text{dec}}$ est plus précis que l'estimateur par la formule de substitution, malgré les nombres assez faibles d'unités non autoreprésentatives échantillonnées et non échantillonnées et (dans plusieurs cas, comme le montre le tableau 4.7) des nombres relativement élevés d'unités autoreprésentatives. Les estimations de l'écart-type par substitution sont systématiquement trop petites, tandis que les estimations bootstrap sont habituellement légèrement élevées (c.-à-d. qu'en général $\widehat{\text{É.T.}}_S < \widehat{\text{É.T.}} < \widehat{\text{É.T.}}_B$). L'erreur relative de $\widehat{\text{É.T.}}_B$ par rapport à $\widehat{\text{É.T.}}$ ne dépasse pas environ 5 % dans ces exemples, sauf dans les cas (AL, CO, GA) où les unités non autoreprésentatives non échantillonnées sont particulièrement peu nombreuses dans la sous-strate de grandes unités.

Les sous-strates de grandes unités dans l'ASPEP ont habituellement une petite population totale dans la base de sondage et contiennent souvent un nombre relativement grand d'unités autoreprésentatives. Bien que nos simulations aient montré que cela n'invalide pas complètement les inférences faites au moyen de $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ ou $\hat{Y}_{\text{dec}}$, ces statistiques ont des distributions assez différentes de celles prévues par la théorie en grand échantillon, et de futures subdivisions des sous-strates permettraient peut-être d'obtenir des sous-strates de grandes unités un peu plus importantes en vue d'obtenir des inférences statistiques se comportant de la manière attendue.

Plus généralement, les résultats des simulations indiquent que l'estimateur fondé sur un test de décision avec l'estimateur des intervalles défini d'après les variances bootstrap se comporte bien et peut être recommandé, sauf pour des populations très dispersées et asymétriques ou des populations pour lesquelles les tailles d'échantillon de grandes unités sont plus petites que 20 à 25.

Remerciements

Le présent article décrit les travaux de recherche et analyses des auteurs et est diffusé en vue d'informer les parties intéressées et de favoriser la discussion. Les conclusions n'engagent que les auteurs et n'ont pas été approuvées par le Census Bureau. Nous tenons à remercier trois examinateurs et un rédacteur associé de leurs commentaires et suggestions utiles qui nous ont permis d'améliorer l'article. Les travaux de recherche de Jun Shao ont été financés partiellement par la bourse NSF Grant DMS-1007454.

Annexe

Preuve du théorème 1. Sous échantillonnage PPT, $\pi_i = n_j p_{ij}$ pour l'unité $i \in U_j$, et à chaque tirage avec remise, l'indice échantillonné $i_t \in U_j, t = 1, \ldots, n_j$ possède $P(i_t = i) = p_{ij}$ pour chaque $i \in U_j$. En calculant les moyennes et les variances (sous échantillonnage répété) de $\hat{N}_j$, $\hat{X}_j$, $\hat{Y}_j$, $N_j^{-1} \sum_{i \in S_j} x_i y_i / \pi_i$ et $N_j^{-1} \sum_{i \in S_j} x_i^2 / \pi_i$, nous constatons que les variances sont d'ordre n_j^{-1} au moyen

des limites données dans (C2) et (C3) et des bornes données dans (C4). Les assertions de la partie (a) s'ensuivent directement.

Pour l'assertion (b), nous avons, en vertu de la définition de $\hat{\beta}$, que

$$\hat{\beta} = \frac{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right) y_i/\pi_i}{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right)^2/\pi_i}$$

$$= \frac{N^{-1}\left(\sum_{j=1}^{2}\hat{\beta}_j\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)\left(\hat{\mu}_{x1} - \hat{\mu}_{x2}\right)\left(\hat{\mu}_{y1} - \hat{\mu}_{y2}\right)\right)}{N^{-1}\left(\sum_{j=1}^{2}\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)\left(\hat{\mu}_{x1} - \hat{\mu}_{x2}\right)^2\right)},$$

d'où l'égalité (2.1) dans (b) découle immédiatement par substitution des limites de la partie (a) ainsi que des limites $N_j/N \to \omega_j$.

Soit Σ_N la matrice diagonale par blocs avec deux blocs diagonaux D_{N_1} et D_{N_2}, et pour $j = 1, 2$, soit

$$\begin{aligned}
\Omega_{1j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{1}{p_{ij}} - N_j\right), \qquad \Omega_{2j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{x_i}{p_{ij}} - X_j\right), \\
\Omega_{3j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{y_i}{p_{ij}} - Y_j\right), \qquad \Omega_{4j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\frac{x_i - \mu_{x,j}}{p_{ij}}\left(y_i - \alpha_j - \beta_j x_i\right).
\end{aligned} \tag{A.1}$$

Puisque S_1 et S_2 sont indépendants, $\{\Omega_{k1}\}_{k=1}^4$ est indépendant de $\{\Omega_{k2}\}_{k=1}^4$. Notons que, ici et tout au long de la présente preuve, les sommes sur $i \in S_j$ utilisées pour définir $\hat{X}_j$, $\hat{Y}_j$, Ω_{kj}, et les estimateurs de variance doivent être interprétés comme étant des sommes *avec multiplicité* compte tenu du plan d'échantillonnage PPT avec remise. La condition (C4) permet d'appliquer le théorème central limite de Liapounov pour montrer que

$$\Sigma_N^{-1/2}\left[\Omega_{11}, \Omega_{21}, \Omega_{31}, \Omega_{12}, \Omega_{22}, \Omega_{32}\right]^T \to_d N\left(0, I_6\right), \ \Omega_{4j} \to_d N\left(0, \sigma_{xe,j}^2\right), \tag{A.2}$$

où I_6 est la matrice identité de dimensions 6×6, et $\sigma_{xe,j}^2$ est donné dans l'énoncé de (d). Les limites qui définissent les variances asymptotiques dans (A.2) existent conformément à (C3).

Preuve de (c). Il est facile de vérifier d'après la définition que

$$\begin{pmatrix}\hat{\beta}_j - \beta_j \\ \hat{\alpha}_j - \alpha_j\end{pmatrix} = \frac{1}{\hat{N}_j\hat{\sigma}_{xj}^2}\sum_{i\in S_j}\begin{pmatrix}x_i - \hat{\mu}_{xj} \\ \hat{\sigma}_{xj}^2 - (x_i - \hat{\mu}_{xj})\hat{\mu}_{xj}\end{pmatrix}\frac{y_i - \alpha_j - \beta_j x_i}{\pi_i}.$$

Puisqu'il a été établi dans (a) que $\hat{\sigma}_{xj}^2 \to_P \sigma_{xj}^2$ et $\hat{N}_j/N_j \to_P 1$, il s'ensuit que la distribution limite de $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right)$ est la même que celle de

$$\sqrt{n}_j \left(N_j \sigma_{xj}^2\right)^{-1} \sum_{i \in S_j} \left(x_i - \mu_{xj}\right)\left(y_i - \alpha_j - \beta_j x_i\right)/\pi_i ,$$

qui est clairement la même que celle de $\sigma_{xj}^{-2}\Omega_{4j}$ dans (A.1). La première assertion de (c) découle immédiatement de (A.2). La convergence de $\hat{\sigma}_{xe,j}^2$ s'ensuit en notant en vertu de (a) que

$$\hat{\sigma}_{xe,j}^2 - N_j^{-2} \sum_{i \in S_j} \frac{\left(x_i - \mu_{xj}\right)^2}{\pi_i p_{ij}} \left(y_i - \alpha_j - \beta_j x_i\right)^2 \to_p 0. \tag{A.3}$$

Le deuxième terme du premier membre de (A.3) contient une variance d'échantillonnage PPT avec remise calculée de manière qu'elle soit bornée par $1/n_j$ conformément à (C4), dont l'espérance en vertu de (C3) converge vers $\sigma_{xe,j}^2$.

Preuve de (d). De (1.2) et (a), il découle que $\left(\hat{Y}_{\text{reg},2} - Y\right)/N \to_p 0,$ qui peut aussi être considéré comme la représentation

$$\begin{aligned}
\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N &= \frac{\sqrt{n}}{N} \sum_{j=1}^{2} \left[\frac{N_j \hat{Y}_j}{\hat{N}_j} - Y_j + \hat{\beta}_j \left(X_j - \frac{N_j \hat{X}_j}{\hat{N}_j}\right)\right] \\
&= \frac{\sqrt{n} N_1^2}{\sqrt{n_1} N \hat{N}_1} \left[\left(-\bar{Y}_1 + \hat{\beta}_1 \bar{X}_1\right)\Omega_{11} - \hat{\beta}_1 \Omega_{21} + \Omega_{31}\right] \\
&\quad + \frac{\sqrt{n} N_2^2}{\sqrt{n_1} N \hat{N}_2} \left[\left(-\bar{Y}_2 + \hat{\beta}_2 \bar{X}_2\right)\Omega_{12} - \hat{\beta}_2 \Omega_{22} + \Omega_{32}\right] \\
&= d_{n1}^T \bar{\Omega}_1 + d_{n2}^T \bar{\Omega}_2,
\end{aligned}$$

où la deuxième égalité découle des définitions notationnelles de Ω_{kj} de même que $\pi_i = n_j p_{ij}$, $\hat{Y}_j = \sum_{i \in S_j} y_i/\pi_i$, $\hat{X}_j = \sum_{i \in S_j} x_i/\pi_i$, et la troisième de

$$d_{nj} = \frac{\sqrt{n} N_j^2}{\sqrt{n_j} N \hat{N}_j} \left[-\bar{Y}_j + \hat{\beta}_j \bar{X}_j, -\hat{\beta}_j, 1\right]^T, \quad \bar{\Omega}_1 = \left[\Omega_{11}, \Omega_{21}, \Omega_{31}\right]^T, \quad \bar{\Omega}_2 = \left[\Omega_{21}, \Omega_{22}, \Omega_{32}\right]^T.$$

En vertu de (A.2), $\bar{\Omega}_1 = O_p(1)$ et $\bar{\Omega}_2 = O_p(1)$. En vertu de la condition (C2), $d_{nj}^T = a_{2j}^T + o_p(1)$. Par conséquent, en vertu de (A.2), de la condition (C3) et de la méthode delta,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1) \to_d N\left(0, \sigma_2^2\right),$$

où la variance asymptotique $\sigma_2^2 = \sum_{j=1}^{2} a_{2j}^T D_j a_{2j}$ est systématiquement estimée par

$$\frac{n}{N^2} \sum_{j=1}^{2} \sum_{i \in S_j} \frac{1}{\pi_i^2} \left(y_i - \hat{\beta}_j x_i - \left(\hat{Y}_j - \hat{\beta}_j \hat{X}_j\right)/\hat{N}_j\right)^2,$$

qui est en accord avec la formule (9) de Cheng et coll. (2010). La preuve que $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N \to_d N\left(0, \sigma_1^2\right)$ est similaire.

Preuve du théorème 2. En vertu de la conclusion (c) du théorème 1,

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1 - \beta_2 + \beta_1\right) \to_d N\left(0, \sum_{j=1}^{2} \sigma_{xe,j}^2 / \left(\varphi_j^2 \sigma_{xj}^4\right)\right). \tag{A.4}$$

La conclusion (2.4) dans la partie (a) de ce théorème s'ensuit directement.

Dans la preuve du théorème 1, nous avons montré que

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.5}$$

où les vecteurs constants a_{kj} (et μ_x, μ_y) ont été définis dans la partie (d) du théorème 1. De même,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N = a_{11}^T \bar{\Omega}_1 + a_{12}^T \bar{\Omega}_2 + o_p(1). \tag{A.6}$$

Quand (2.3) est vérifiée, $\beta_j = \beta$ (en vertu de la partie (b) du théorème 1) et $\mu_y - \beta\mu_x = \sum_{j=1}^{2} \omega_j \left(\mu_{yj} - \beta_j \mu_{xj}\right) = \mu_{y2} - \beta\mu_{x2}$, de sorte que $a_{1j} = a_{2j}$ pour $j = 1, 2$. Il découle directement de (A.5) et (A.6) que $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right)/N \to_p 0$, et donc que les estimateurs $\hat{Y}_{\text{reg},k}$ suivent la même loi asymptotique, qui est normale comme nous l'avons montré à la partie (d) du théorème 1. Enfin, la définition de $\hat{Y}_{\text{dec}}$ implique que $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},1} \text{ ou } \hat{Y}_{\text{reg},2}\right) = 1$, et (A.5) et (A.6) impliquent que

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.7}$$

ce qui achève la preuve de (2.5) dans (a).

Preuve de (b). Si $\beta_1 \neq \beta_2$, alors (A.4) implique que $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$, c.-à-d. que le test t pour l'égalité de $\hat{\beta}_j$ donne lieu au rejet avec certitude à la limite. Alors (A.7) continue d'être vérifiée, et la loi asymptotique de $\hat{Y}_{\text{dec}}$ demeure la même que celle de $\hat{Y}_{\text{reg},2}$.

Preuve du théorème 3. Dans ce théorème, les hypothèses (C2) à (C4) sont remplacées par les hypothèses selon lesquelles les triplets iid (y_i, x_i, z_i) satisfont les conditions de moments et le modèle (2.7). Les assertions dans (C2) à (C4) restent alors vérifiées lorsque la probabilité tend vers 1 quand n, N sont grands, ce qui est établi à l'aide de la loi (forte) des grands nombres.

Outre les conclusions des théorèmes 1 et 2, il reste à montrer que $\hat{Y}_{\text{reg},2}$ possède une plus petite variance asymptotique que $\hat{Y}_{\text{reg},1}$. Soit $\vartheta = (\vartheta_1, \vartheta_2)$ et

$$F_j(\vartheta) = [-\vartheta_1, -\vartheta_2, 1]\, D_j\, [-\vartheta_1, -\vartheta_2, 1]^T.$$

Selon la définition de σ_1^2 et σ_2^2 dans (2.2), il suffit de montrer que $F_j(\vartheta)$ prend sa valeur minimale à $\vartheta = (\alpha_j, \beta_j)$. Nous allons maintenant prouver cela pour $j = 1$. La preuve pour $j = 2$ est similaire. Soit $m_{ii'}$ l'élément (i, i') de D_1. Puisque D_1 est symétrique et définie positive sous la condition (C3), $m_{12} = m_{21}$ et il existe un $\theta^* = (\theta_1^*, \theta_2^*)$ unique tel que $F_1(\theta^*) = \min_\vartheta F_1(\vartheta)$ et $\partial F_1(\vartheta)/\partial\vartheta^T|_{\vartheta=\theta^*} = 0$. Cela implique que θ^* est la solution des deux équations suivantes :

$$m_{11}\vartheta_1 + m_{12}\vartheta_2 = m_{13}, \quad m_{12}\vartheta_1 + m_{22}\vartheta_2 = m_{23} \tag{A.8}$$

Par conséquent, il suffit de montrer que $\theta^* = (\alpha_1, \beta_1)$. Puisque D_1 est définie positive, le système d'équations (A.8) possède une solution unique. Étant donné la définition de D_1,

$$\begin{aligned} m_{11}\alpha_1 + m_{12}\beta_1 &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)^2 p_{i1}\alpha_1 + \sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{x_i}{p_{i1}} - X_1\right)p_{i1}\beta_1\right] \\ &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

et

$$\begin{aligned} m_{13} &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{y_i}{p_{i1}} - Y_1\right)p_{i1}\right] \\ &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 + \beta_1 x_i + \varepsilon_i - N_1\alpha_1 p_{i1} - \beta_1 p_{i1} X_1)\right] \\ &= \lim_{N1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

où la dernière égalité découle de l'hypothèse que ε_i est indépendant de x_i et z_i, et est de moyenne 0 et de variance finie, et chacune des séquences z_i, $1/z_i$, et x_i/z_i est iid avec une espérance finie. Par conséquent, $m_{11}\alpha_1 + m_{12}\beta_1 = m_{13}$. On prouve de même que $m_{12}\alpha_1 + m_{22}\beta_2 = m_{23}$. Par conséquent, (α_1, β_1) est la solution unique du système d'équations (A.8), c.-à-d. que $F_1(\vartheta)$ prend sa valeur minimale à $\vartheta = (\alpha_1, \beta_1)$. D'où, $\sigma_2^2 < \sigma_1^2$. Cela termine la preuve du théorème 3.

Bibliographie

Bancroft, T., et Han, C.-P. (1977). Inference based on conditional specifications: A note and a bibliography. *International Statistical Review*, 45, 117-127.

Cheng, Y., Corcoran, C., Barth, J. et Hogue, C. (2009). An estimation procedure for the new public employment survey design. Washington, DC: American Statistical Association. *Survey Research Methods Section*, American Statistical Association, 3032-3046.

Cheng, Y., Slud, E. et Hogue, C. (2010). Variance estimation for decision-based estimators with application to the annual survey of public employment and payroll. *Government Statistics Section of the American Statistical Association*. Vancouver: American Statistical Association.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Isaki, C., et Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Rao, J.N.K., et Ramachandran, V. (1974). Comparison of the separate and combined ratio estimators. *Sankhyā*, C, 36, 151-156.

Saleh, A.K. Md. (2006). *Theory of Preliminary Test and Stein-type Estimation, with Applications*. Hoboken: Wiley-Interscience.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J., et Tu, D. (1995) *The Jackknife and Bootstrap*. New York: Springer.

Slud, E.V. (2012). Moderate-sample behavior of adaptively pooled stratified regression estimators. U.S. Census Bureau preprint.