Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Note brève sur l'estimation fondée sur les quantiles et les expectiles dans les échantillons à probabilités inégales

par Linda Schulze Waltrup et Göran Kauermann

Date de diffusion : le 22 juin 2016

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 STATCAN.infostats-infostats.STATCAN@canada.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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 191-199
Statistique Canada, nº 12-001-X au catalogue

Note brève sur l'estimation fondée sur les quantiles et les expectiles dans les échantillons à probabilités inégales

Linda Schulze Waltrup et Göran Kauermann[1]

Résumé

L'estimation des quantiles est une question d'intérêt dans le contexte non seulement de la régression, mais aussi de la théorie de l'échantillonnage. Les expectiles constituent une solution de rechange naturelle ou un complément aux quantiles. En tant que généralisation de la moyenne, les expectiles ont gagné en popularité ces dernières années parce qu'en plus d'offrir un portrait plus détaillé des données que la moyenne ordinaire, ils peuvent servir à calculer les quantiles grâce aux liens étroits qui les associent à ceux-ci. Nous expliquons comment estimer les expectiles en vertu d'un échantillonnage à probabilités inégales et comment les utiliser pour estimer la fonction de répartition. L'estimateur ajusté de la fonction de répartition obtenu peut être inversé pour établir les estimations des quantiles. Nous réalisons une étude par simulations pour examiner et comparer l'efficacité de l'estimateur fondé sur des expectiles.

Mots-clés : Quantiles; expectiles; probabilité proportionnelle à la taille; approche fondée sur le plan de sondage; variable auxiliaire; fonction de répartition.

1 Introduction

Ces dernières années, l'estimation des quantiles et la régression quantile ont connu de nouveaux développements découlant des travaux de Koenker (2005). L'idée principale est d'estimer une fonction de répartition cumulative inversée, qu'on appelle généralement la fonction quantile $Q(\alpha) = F^{-1}(\alpha)$ pour $\alpha \in (0,1)$, où le quantile 0,5, $Q(0{,}5)$, la médiane, joue un rôle central. Pour le dépistage de données d'enquête à partir d'un échantillon à probabilités inégales et à probabilités connues d'inclusion, Kuk (1988) montre comment estimer les quantiles en tenant compte des probabilités d'inclusion. L'idée principale est d'estimer une fonction de répartition de la variable d'intérêt et de l'inverser pour obtenir la fonction quantile. Chambers et Dunstan (1986) proposent un estimateur fondé sur un modèle pour la fonction de répartition. Rao, Kovar et Mantel (1990) proposent un estimateur fondé sur le plan de sondage et faisant appel à des données auxiliaires pour la fonction de répartition cumulative. Chen, Elliott et Little (2010) et Chen, Elliott et Little (2012) ont également récemment proposé des approches bayésiennes allant dans le même sens.

L'estimation des quantiles résulte de la minimisation d'une fonction de perte L_1, comme l'a montré Koenker (2005). Si la perte L_1 est remplacée par la fonction de perte L_2, on obtient ce qu'on appelle des « expectiles », une notion présentée par Aigner, Amemiya et Poirier (1976) et par Newey et Powell (1987). Pour $\alpha \in (0,1)$, on obtient la fonction expectile $M(\alpha)$ qui, comme la fonction quantile $Q(\alpha)$, définit de façon unique la fonction de répartition cumulative $F(y)$. Les expectiles sont relativement faciles à estimer et suscitent un certain intérêt depuis quelques temps; voir par exemple Schnabel et Eilers (2009), Pratesi, Ranalli et Salvati (2009), Sobotka et Kneib (2012) et Guo et Härdle (2013). Ils ne sont toutefois pas faciles à interpréter, et sont donc moins acceptés et utilisés en statistique que les quantiles; voir Kneib (2013). Les quantiles et les expectiles sont reliés, c'est-à-dire qu'il existe une fonction de transformation unique et

1. Linda Schulze Waltrup, Administration des affaires et sciences sociales, Université Louis-et-Maximilien de Munich, Ludwigstraße 33, 80539 Munich, Allemagne. Courriel : lschulze_waltrup@stat.uni-muenchen.de; Göran Kauermann, Administration des affaires et sciences sociales, Université Louis-et-Maximilien de Munich, Ludwigstraße 33, 80539 Munich, Allemagne. Courriel : goeran.kauermann@stat.uni-muenchen.de.

inversible $h_y : [0,1] \to [0,1]$ de sorte que $M(h(\alpha)) = Q(\alpha)$; voir Yao et Tong (1996) et De Rossi et Harvey (2009). Cette relation peut être exploitée pour estimer les quantiles à partir d'un ensemble d'expectiles ajustés. Schulze Waltrup, Sobotka, Kneib et Kauermann (2014) ont utilisé ce principe et montré de façon empirique que les quantiles ainsi obtenus peuvent être plus efficients que les quantiles empiriques, même lorsque ces derniers font l'objet d'un lissage (voir Jones 1992). Ce résultat pourrait s'expliquer intuitivement par le fait que les expectiles tiennent compte de toutes les données, alors que les quantiles fondés sur la fonction de répartition empirique ne tiennent compte que des données de gauche (ou de droite). Autrement dit, la médiane est définie par la moitié gauche (ou droite) des données, alors que la moyenne (expectile de 50 %) est une fonction tenant compte de tous les points de données. Dans la présente note, nous utilisons ces constatations comme point de départ pour montrer comment les expectiles peuvent être estimés pour des échantillons à probabilités inégales et comment obtenir une fonction de répartition ajustée à partir d'expectiles ajustés.

La présentation de l'article est la suivante. À la section 2, on présente les éléments de notation utiles et on discute de la régression quantile dans un échantillonnage à probabilités inégales. Ce sujet est approfondi à la section 3, où l'on présente l'estimation des expectiles. À la section 4, on exploite la relation entre les expectiles et les quantiles pour montrer comment dériver les quantiles à partir d'expectiles ajustés. La section 5 présente des simulations pour illustrer le gain d'efficacité découlant de l'utilisation des quantiles dérivés d'expectiles; l'article se termine par une discussion à la section 6.

2 Estimation des quantiles

Considérons une population finie de N éléments et une variable d'enquête continue Y. On s'intéresse aux quantiles de la fonction de répartition cumulative $F(y) = \sum_{i=1}^{N} 1\{Y_i \leq y\}/N$, et on définit comme

$$Q(\alpha) = \inf\left\{\arg\min_q \sum_{i=1}^{N} w_\alpha(Y_i - q)\,|\,Y_i - q\,|\right\} \tag{2.1}$$

la fonction quantile de Y (voir Koenker 2005), où

$$w_\alpha(\varepsilon) = \begin{cases} \alpha & \text{pour } \varepsilon > 0 \\ 1-\alpha & \text{pour } \varepsilon \leq 0. \end{cases}$$

L'argument « inf » de l'expression (2.1) est nécessaire pour une population finie puisque « arg min » n'est pas unique. On tire un échantillon de la population selon des probabilités d'inclusion connues π_i, $i = 1, \ldots, N$. En notant $y_1, \ldots, y_n$ l'échantillon obtenu, on estime la fonction quantile en remplaçant (2.1) par la version avec échantillon pondéré

$$\hat{Q}_N(\alpha) = \inf\left\{\arg\min_q \sum_{j=1}^{n} \frac{1}{\pi_j} w_{\alpha,j}\,|\,y_j - q\,|\right\} \tag{2.2}$$

avec $w_{\alpha,j} = w_\alpha(y_j - q)$, selon la définition ci-dessus. Il est facile de voir que la somme en (2.2) est une estimation sans biais par rapport au plan de la somme dans $Q(\alpha)$ donnée en (2.1). Néanmoins, parce qu'on

admet « arg min », il s'ensuit que $\hat{Q}_N(\alpha)$ n'est pas sans biais pour $Q(\alpha)$. Examinons donc les énoncés de cohérence pour $\hat{Q}_N(\alpha)$ comme suit. Soit $R_i(q) = w_\alpha(y_i - q)|y_i - q|$ et

$$\bar{R}_N(q) := \frac{1}{N}\sum_i R_i(q).$$

On tire un échantillon à partir de $R_i(q), i = 1, \ldots, N$ en appliquant un plan de sondage cohérent de sorte que

$$\bar{r}_n(q) := \frac{1}{N}\sum_{j=1}^{n}\frac{1}{\pi_j} r_j(q)$$

converge par rapport au plan pour $\bar{R}_N(q)$, où $r_j(q)$ désigne l'échantillon de $R_i(q)$. Soulignons que $r_j(q)$ et donc $\bar{r}_n(q)$, $R_i(q)$ et $\bar{R}_N(q)$ dépendent aussi de α, qui a été supprimé de la notation par souci de lisibilité. Soit q_0 la valeur minimale de $\bar{R}_N(q)$, qui n'est pas nécessairement unique en raison de la structure finie de la population. On peut admettre l'argument « inf », c'est-à-dire $q_0 = \inf\{\arg\min \bar{R}_N(q)\}$, mais par souci de simplicité, on suppose un modèle de superpopulation (voir Isaki et Fuller 1982) en considérant la population finie comme un échantillon d'une superpopulation infinie. Pour cette dernière, on présume que la variable d'enquête Y a une fonction de répartition cumulative continue, de sorte que q_0 donne un quantile α unique. Pour $\delta > 0$, on obtient

$$P(\bar{r}_n(q_0) < \bar{r}_n(q_0 - \delta)) \Leftrightarrow P\left(\frac{1}{N}\sum_{j=1}^{n}\frac{1}{\pi_j}\{r_j(q_0) - r_j(q_0 - \delta)\} < 0\right).$$

Soulignons que l'argument dans l'énoncé de probabilité est une estimation convergente par rapport au plan de sondage pour $\bar{R}_N(q_0) - \bar{R}_N(q_0 - \delta)$, dont la valeur est inférieure à zéro puisque q_0 correspond à la valeur minimale de $\bar{R}_N(\cdot)$. En conséquence, la probabilité tend vers un au sens de la convergence par rapport au plan de sondage définie par Isaki et Fuller (1982). Il en va bien sûr de même pour $\delta < 0$. En vertu de cet énoncé, on peut conclure que la valeur estimée minimale $\hat{q}_0 = \arg\min \sum_{j=1}^{n} 1/\pi_j \; r_j(q)$ est une estimation convergente par rapport au plan de sondage pour q_0 de sorte que $\hat{Q}_N(\alpha)$ en (2.2) converge aussi par rapport au plan de sondage pour $Q_N(\alpha)$. Il est facile de montrer que $\hat{Q}_N(\alpha)$ est l'inverse de la fonction de répartition cumulative pondérée normalisée

$$\hat{F}_N(y) := \frac{\sum_{j=1}^{n} 1\{y_j \le y\}/\pi_j}{\sum_{j=1}^{n} 1/\pi_j}$$

selon la notation utilisée par Kuk (1988). Soulignons que $\hat{F}_N(y)$ correspond à l'estimation de Hajek (1971) pour la fonction de répartition cumulative (voir aussi Rao et Wu 2009) et n'est donc pas une estimation de Horvitz-Thompson. Par conséquent, $\hat{Q}_N(\alpha)$ n'est pas sans biais par rapport au plan de sondage. Néanmoins, $\hat{F}_N(y)$ est une fonction de répartition valide et peut donc être considérée comme une version normalisée de l'estimateur de Lahiri ou de Horvitz-Thompson de la fonction de répartition (voir Lahiri 1951), désignée par

$$\hat{F}_L(y) := \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j 1\{y_j \le y\}.$$

Kuk (1988) propose de remplacer $\hat{F}_L(\cdot)$ par d'autres estimations de la fonction de répartition : au lieu d'estimer la fonction de répartition elle-même, il suggère d'estimer la proportion complémentaire $\hat{S}_R(y)$ qui mène ensuite à l'estimation $\hat{F}_R(y)$ définie par

$$\hat{F}_R(y) = 1 - \hat{S}_R(y) = 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j 1\{y_j > y\}.$$

Directement à partir de ces définitions, on peut exprimer $\hat{F}_R(\cdot)$ en termes de $\hat{F}_N(\cdot)$ par

$$\hat{F}_R = 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j + \hat{F}_L \quad \text{et} \quad \hat{F}_L = \frac{\sum_{j=1}^{n} 1/\pi_j}{N}\hat{F}_N. \tag{2.3}$$

Kuk (1988) montre que, en vertu d'un échantillonnage à probabilités inégales, l'estimation de la médiane dérivée de $\hat{F}_R$ est plus efficace que celles qui sont dérivées de $\hat{F}_N$ et $\hat{F}_L$ en termes d'estimation de l'erreur quadratique moyenne. Soulignons que les estimateurs $\hat{F}_N$, $\hat{F}_L$ et $\hat{F}_R$ coïncident dans le cas d'un échantillonnage aléatoire simple sans remise où $\pi_j = \pi = n/N$.

3 Estimation des expectiles

Les expectiles sont une solution de rechange aux quantiles. La fonction expectile $M(\alpha)$ est donc définie par remplacement de la perte L_1 dans l'expression (2.1) par la perte L_2 pour donner

$$M(\alpha) = \arg\min_m \left\{\sum_{i=1}^{N} w_\alpha (Y_i - m)(Y_i - m)^2\right\}. \tag{3.1}$$

Soulignons que $M(\alpha)$ est continue en α même pour des populations finies. En outre, $M(0{,}5)$ est égale à la valeur moyenne $\bar{Y} = \sum_{i=1}^{N} Y_i/N$. À partir de l'échantillon $y_1, \ldots, y_n$ avec probabilités d'inclusion $\pi_1, \ldots, \pi_n$, on peut estimer $M(\alpha)$ en remplaçant la somme de l'expression (2.2) par sa version d'échantillon, c'est-à-dire

$$\hat{M}(\alpha) = \arg\min_m \left\{\sum_{j=1}^{n} \frac{1}{\pi_j} w_{\alpha,j} (y_j - m)^2\right\}$$

avec $w_{\alpha,j}$ correspondant à la définition ci-dessus. Il est facile de voir que la somme dans $\hat{M}(\alpha)$ est une estimation sans biais par rapport au plan de la somme dans $M(\alpha)$. L'estimation elle-même n'est toutefois pas sans biais par rapport au plan comme pour la fonction quantile susmentionnée. Toutefois, on peut utiliser les mêmes arguments que pour $Q_N(\alpha)$ en (2.2) pour établir la convergence par rapport au plan de sondage.

4 Des expectiles à la fonction de répartition

La fonction quantile $Q(\alpha)$ et la fonction expectile $M(\alpha)$ définissent toutes deux de façon unique une fonction de répartition $F(.)$. Tandis que $Q(\alpha)$ est une simple inversion de $F(.)$, la relation entre $M(\alpha)$ et $F(.)$ est plus complexe. Selon Schnabel et Eilers (2009) et Yao et Tong (1996), on peut établir la relation

$$M(\alpha) = \frac{(1-\alpha)\,G(M(\alpha)) + \alpha\{M(0{,}5) - G(M(\alpha))\}}{(1-\alpha)\,F(M(\alpha)) + \alpha\{1 - F(M(\alpha))\}}, \tag{4.1}$$

où $G(m)$ est la fonction génératrice des moments définie par $G(m) = \sum_{i=1}^{N} Y_i\, 1\{Y_i \le m\}/N$. L'expression (4.1) donne la relation unique de la fonction $M(\alpha)$ à la fonction de répartition $F(.)$. Il faut maintenant résoudre (4.1) pour $F(.)$, c'est-à-dire exprimer la répartition $F(.)$ en termes de la fonction expectile $M(.)$. Cela n'est apparemment pas possible sous une forme analytique, mais on peut effectuer le calcul numériquement. Pour ce faire, on évalue la fonction ajustée $\hat{M}(\alpha)$ selon un ensemble dense de valeurs $0 < \alpha_1 < \alpha_2 \ldots < \alpha_L < 1$, en désignant les valeurs ajustées par $\hat{m}_l = \hat{M}(\alpha_l)$. On définit aussi des bornes à gauche et à droite par $\hat{m}_o = \hat{m}_1 - c_0$ et $\hat{m}_{L+1} = \hat{m}_L + c_{L+1}$, où c_0 et c_L sont des constantes définies par l'utilisateur. Par exemple, on peut définir $c_0 = \hat{m}_2 - \hat{m}_1$ et $c_{L+1} = \hat{m}_L - \hat{m}_{L-1}$. Ce faisant, on dérive les valeurs ajustées pour la fonction de répartition cumulative $F(.)$ à $\hat{m}_l$, que l'on écrit $\hat{F}_l := \hat{F}(\hat{m}_l) = \sum_{j=1}^{l} \hat{\delta}_j$ pour les échelons non négatifs $\hat{\delta}_j \ge 0, j = 1, \ldots, L$ avec $\sum_{j=1}^{L} \hat{\delta}_j \le 1$. On définit $\hat{\delta}_{L+1} = 1 - \sum_{l=1}^{L} \hat{\delta}_l$ pour faire de $\hat{F}(.)$ une fonction de répartition. En supposant une répartition uniforme entre les points de support $\hat{m}_l$ de l'ensemble dense, on peut exprimer la fonction de génération des moments $G(.)$ par simple intégration séquentielle comme

$$\hat{G}_l := \hat{G}(\hat{m}_l) = \int_{-\infty}^{m_l} x\, d\hat{F}(x) = \sum_{j=1}^{l} \hat{d}_j \hat{\delta}_l,$$

où $\hat{d}_j = (\hat{m}_j - \hat{m}_{j-1})/2$ sous la contrainte que $\hat{G}_{L+1} = \hat{M}(0{,}5)$ et $\hat{M}(0{,}5) = \sum_{j=1}^{n}(y_j/\pi_j)\big/\sum_{j=1}^{n}(1/\pi_j)$. Avec les échelons $\hat{\delta}_l, l = 1, \ldots, L$, on peut maintenant réécrire l'expression (4.1) comme

$$\hat{m}_l = \frac{(1-\alpha)\sum_{j=1}^{l} \hat{d}_j \hat{\delta}_j + \alpha\left(\hat{M}(0{,}5) - \sum_{j=1}^{l} \hat{d}_j \hat{\delta}_j\right)}{(1-\alpha)\sum_{j=1}^{l} \hat{\delta}_j + \alpha\left(1 - \sum_{j=1}^{l} \hat{\delta}_j\right)}, \quad l = 1, \ldots, L,$$

que l'on résout ensuite pour $\hat{\delta}_1, \ldots, \hat{\delta}_L$. Il s'agit d'un exercice numérique relativement direct sur le plan conceptuel. On peut consulter Schulze Waltrup et coll. (2014) pour les détails. Une fois qu'on a calculé $\hat{\delta}_1, \ldots, \hat{\delta}_L$, on obtient une estimation pour la fonction de répartition cumulative, qu'on écrit $\hat{F}_N^M(y) = \sum_{l:\hat{m}_l < y} \hat{\delta}_l$. On peut aussi inverser $\hat{F}_N^M(.)$, ce qui donne une fonction quantile ajustée que l'on désigne $\hat{Q}_N^M(\alpha)$.

Comme le montre Kuk (1988), à la fois théoriquement et empiriquement, $\hat{F}_R(.)$ est plus efficiente que $\hat{F}_N(.)$. On exploite cette relation en l'appliquant à $\hat{F}_N^M(.)$ pour obtenir l'estimateur

$$\hat{F}_R^M := 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j + \frac{\sum_{j=1}^{n} 1/\pi_j}{N}\hat{F}_N^M .$$

Dans la section qui suit, on compare les quantiles calculés à partir de l'estimateur fondé sur les expectiles $\hat{F}_R^M$ avec les quantiles calculés à partir de $\hat{F}_R$. Soulignons que $\hat{F}_R^M$ et $\hat{F}_R$ ne sont pas des fonctions de répartition appropriées puisqu'elles ne sont pas normalisées pour prendre des valeurs situées entre 0 et 1.

5 Simulations

On a réalisé une petite étude par simulations pour illustrer l'efficacité des estimations fondées sur les expectiles. On utilise ci-dessous la méthode d'échantillonnage de Midzuno (voir Midzuno 1952); les probabilités d'inclusion π_j sont définies comme étant proportionnelles à une mesure de la taille x, selon le module « *sampling* » de Tillé et Matei (2015) dans le logiciel R. On examine deux ensembles de données, que Kuk (1988) a aussi utilisés. Le premier ensemble de données (Logements) comprend deux variables fortement corrélées (corrélation de 0,97), soit le nombre d'unités de logement (X) et le nombre d'unités louées (Y); voir aussi Kish (1965). Le deuxième ensemble de données (Villages) comprend de l'information sur la population (X) et sur le nombre de personnes travaillant dans des entreprises familiales (Y) dans 128 villages de l'Inde; voir Murthy (1967). Dans le deuxième ensemble de données, la corrélation entre Y et X s'établit à 0,54. Afin de comparer les résultats de la simulation à ceux de Kuk (1988), on a choisi un échantillon de la même taille, soit $n = 30$ (pour une population totale de $N = 270$ en ce qui concerne les données sur les logements et de $N = 128$ pour les données sur les villages).

On compare les quantiles définis par l'inversion de $\hat{F}_R$ avec les quantiles définis par l'inversion de $\hat{F}_R^M$. Le tableau 5.1 présente la racine de l'erreur quadratique moyenne (REQM) et l'efficience relative pour certains quantiles. On constate que la médiane pour les données relatives aux villages et, pour les données relatives aux logements, les quantiles supérieurs dérivés des expectiles ont une efficience accrue. En outre, le gain d'efficience n'est pas uniforme; on constate en effet une perte d'efficience dans les quantiles inférieurs.

Tableau 5.1
Comparaison de l'erreur quadratique moyenne fondée sur 500 répliques

	α	quantiles $\sqrt{\text{EQM}\left(\hat{Q}_R(\alpha)\right)}$	quantiles dérivés des expectiles $\sqrt{\text{EQM}\left(\hat{Q}_R^M(\alpha)\right)}$	efficience relative $\frac{\sqrt{\text{EQM}\left(\hat{Q}_R^M(\alpha)\right)}}{\sqrt{\text{EQM}\left(\hat{Q}_R(\alpha)\right)}}$
Logements	0,1	2,57	2,76	1,07
	0,25	1,77	1,97	1,11
	0,5	2,45	2,35	0,96
	0,75	3,15	2,91	0,92
	0,9	4,20	3,43	0,82
Villages	0,1	5,52	6,65	1,21
	0,25	11,41	10,31	0,90
	0,5	12,29	11,69	0,95
	0,75	16,24	15,41	0,95
	0,9	13,31	18,34	1,38

Pour mieux comprendre, on a réalisé une simulation à l'aide d'un échantillon plus grand de taille $n = 100$ sélectionné à partir de populations de tailles $N = 1\,000$ et $N = 10\,000$. On a tiré Y et X d'une loi log-normale standard bivariée avec $\mu = 0$ et $\sigma = 1$. Les variables Y et X sont tirées de façon que la corrélation entre les variables soit égale à 0,9. On a encore une fois calculé la racine de l'erreur quadratique moyenne pour une gamme de valeurs de α; l'efficience relative de l'approche fondée sur les expectiles est illustrée à la figure 5.1. Pour une meilleure présentation visuelle, les graphiques donnent une version lissée de l'efficience relative. On constate une diminution de la racine de l'erreur quadratique moyenne dans les deux cas, soient $N = 1\,000$ et $N = 10\,000$. On peut conclure que les expectiles peuvent être facilement ajustés dans un échantillonnage à probabilités inégales et que la relation entre les expectiles et la fonction de répartition peut être exploitée numériquement pour calculer les quantiles avec une efficience accrue. Ce gain d'efficience ne se vérifie que pour les quantiles supérieurs, c'est-à-dire pour les valeurs de α dont la borne inférieure est strictement positive. Soulignons toutefois que le plan de sondage est tel que les grandes valeurs de Y sont échantillonnées selon une probabilité supérieure, puisque le plan de sondage vise à obtenir des estimations plus fiables pour le côté droit de la fonction de répartition, c'est-à-dire pour les quantiles supérieurs. Si on s'intéresse aux quantiles inférieurs, il faut utiliser un plan de sondage différent en attribuant une probabilité d'inclusion accrue aux personnes ayant une valeur Y faible. Dans ce cas, on observerait un comportement correspondant au reflet de celui qui est illustré à la figure 5.1 en ce qui concerne α.

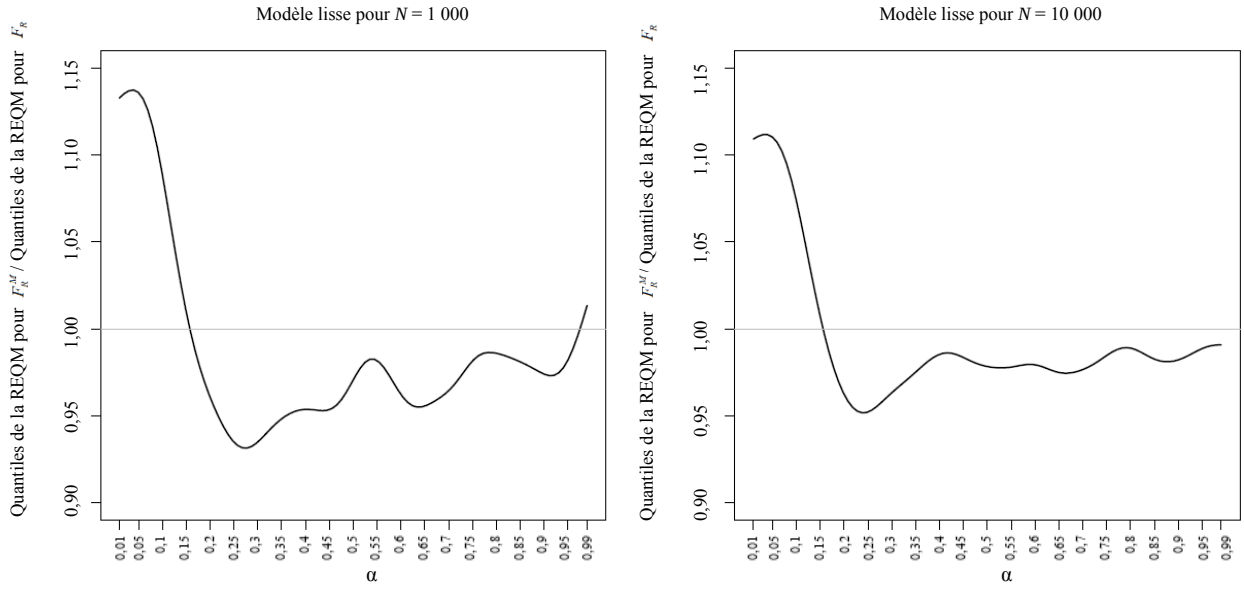

Figure 5.1 Racine de l'erreur quadratique moyenne (REQM) relative des quantiles et des quantiles dérivés des expectiles pour le plan de sondage avec probabilité proportionnelle à la taille (PPT), calculée à partir de 500 répliques (à gauche : $N = 1\,000$; à droite : $N = 10\,000$).

6 Discussion

À la section 4, on a augmenté la boîte à outils des expectiles à l'estimation des fonctions de répartition dans le cadre d'un échantillonnage à probabilités inégales. On a défini les expectiles pour des échantillons

à probabilités inégales. Quand on compare les quantiles fondés sur $\hat{F}_R$ avec les quantiles reposant sur l'estimateur fondé sur les expectiles $\hat{F}_R^M$, on constate que l'estimateur proposé offre un bon rendement en comparaison des méthodes existantes. Le calcul des expectiles empiriques est mis en œuvre dans le logiciel libre R (voir R Core Team 2014) et se trouve dans le module expectreg du logiciel R mis au point par Sobotka, Schnabel, et Schulze Waltrup (2013). Le calcul de l'estimateur de la fonction de répartition fondé sur les expectiles $\hat{F}_N^M$ fait aussi partie du module expectreg du logiciel R. Le calcul de $\hat{F}_R^M$ est toutefois plus exigeant que celui de $\hat{F}_R$ parce qu'il comporte trois étapes : il faut d'abord calculer les expectiles pondérés selon la méthode exposée à la section 3, puis estimer $\hat{F}_R^N$ et enfin, dériver $\hat{F}_R^M$ à partir de $\hat{F}_R^N$ (voir la section 4). Dans la simulation log-normale, il faut environ 2-3 secondes pour calculer $\hat{F}_R^M$ pour $N = 1\ 000$, tandis que l'effort pour calculer $\hat{F}_R$ est négligeable.

Remerciements

Les deux auteurs remercient la *Deutsche Forschungsgemeinschaft* DFG (KA 1188/7-1) pour le soutien financier.

Bibliographie

Aigner, D.J., Amemiya, T. et Poirier, D.J. (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, 17(2), 377-396.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 36, 1, 25-37.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2012). Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales. *Techniques d'enquête*, 38, 2, 221-233.

De Rossi, G., et Harvey, A. (2009). Quantiles, expectiles and splines. Nonparametric and robust methods in econometrics. *Journal of Econometrics*, 152(2), 179-185.

Guo, M., et Härdle, W. (2013). Simultaneous confidence bands for expectile functions. *AStA - Advances in Statistical Analysis*, 96(4), 517-541.

Hajek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, part one". *The Foundations of Survey Sampling*, 236.

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

Jones, M. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44(4), 721-727.

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

Kneib, T. (2013). Beyond mean regression (with discussion and rejoinder). *Statistical Modelling*, 13(4), 275-385.

Koenker, R. (2005). *Quantile Regression, Econometric Society Monographs*. Cambridge: Cambridge University Press.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75(1), 97-103.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute*, (33), 133-140.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Newey, W.K., et Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819-847.

Pratesi, M., Ranalli, M. et Salvati, N. (2009). Nonparametric M-quantile regression using penalised splines. *Journal of Nonparametric Statistics*, 21(3), 287-304.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienne, Autriche : R Foundation for Statistical Computing.

Rao, J., et Wu, C. (2009). Empirical likelihood methods. *Handbook of Statistics*, 29B, 189-207.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

Schnabel, S.K., et Eilers, P.H. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53(12), 4168-4177.

Schulze Waltrup, L., Sobotka, F., Kneib, T. et Kauermann, G. (2014). Expectile and quantile regression - David and Goliath? *Statistical Modelling*, 15, 433-456.

Sobotka, F., et Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56(4), 755-767.

Sobotka, F., Schnabel, S. et Schulze Waltrup, L. (2013). *Expectreg: Expectile and Quantile Regression*. Avec la contribution de P. Eilers, T. Kneib et G. Kauermann, R package version 0.38.

Tillé, Y., et Matei, A. (2015). *Sampling: Survey Sampling. R package, version 2.7*. https://cran.r-project.org/web/packages/sampling/index.html.

Yao, Q., et Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, 6(2-3), 273-292.