

## ESTIMATEURS DE CALAGE ROBUSTES AVEC POIDS CONTRAINTS

Pierre Duchesne<sup>1</sup>

### RÉSUMÉ

Nous considérons l'utilisation d'estimateurs de calage en présence de valeurs aberrantes. Une extension de la classe des estimateurs de calage de Deville et Särndal (1992) reposant sur les estimateurs QR de Wright (1983) est obtenue. Comme application, cette classe d'estimateurs nous permet de considérer des estimateurs de calage robustes. En choisissant une métrique adéquate, nous obtenons des poids robustes qui sont complètement bornés à un intervalle spécifié à l'avance. Les estimateurs robustes considérés reposent sur des estimateurs avec un haut point de rupture. Dans le cas particulier où la métrique choisie est la métrique quadratique, l'estimateur que nous suggérons est une généralisation d'une proposition de Lee (1991). Une brève étude de simulation illustre la nouvelle méthodologie.

MOTS CLÉS: Estimateurs de calage; pondération négative; robustesse.

### ABSTRACT

We consider the use of calibration estimators when outliers are present. We extend the class of estimators of Deville and Särndal (1992) using the QR estimators of Wright (1983). As an application, this class of estimators allows construction of robust calibration estimators. Using an adequate metric, we obtain robust weights which are completely bounded in a given interval. We consider robust estimators with high breakdown point. When the metric is quadratic, the suggested estimator is a generalisation of an estimator proposed by Lee (1991). A brief simulation study illustrates the new methodology.

KEYS WORDS: Calibration estimator; Negative weights; Robustness.

### 1. INTRODUCTION

On s'intéresse au problème des valeurs aberrantes en sondages où l'on estime le total d'une variable d'intérêt  $y$  dans une population finie. Lee (1995) fait un survol des motivations, de la terminologie et des développements entourant la robustesse en théorie des sondages. Nous privilégions l'utilisation d'estimateurs de calage en présence de valeurs aberrantes. Avec une métrique adéquate, ils permettent d'éviter le problème de la pondération négative, situation indésirable en pratique. Les estimateurs de calage de Deville et Särndal (1992) sont asymptotiquement équivalents à l'estimateur par régression généralisé (GREG). Ce dernier repose essentiellement sur l'estimateur des moindres carrés généralisés, et donc est sensible aux valeurs

aberrantes. D'un autre côté, les estimateurs robustes du total connus sont susceptibles d'avoir des poids négatifs. Notre construction s'inspire des estimateurs QR de Wright (1983). Ils suggèrent une extension pour avoir des estimateurs de calage robustes avec poids contraints.

Parmi les premières alternatives robustes de l'estimation du total plusieurs reposent sur les GM-estimateurs. Cependant, beaucoup d'attention a été consacré à des estimateurs possédant en plus de bonnes propriétés de robustesse globale mesurée par le point de rupture, qui mesure le pourcentage de valeurs aberrantes dans l'échantillon que l'estimateur peut tolérer tout en donnant tout de même une bonne estimation d'une certaine caractéristique de la population. Nous nous proposons de construire des

---

<sup>1</sup> Pierre Duchesne, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3J7, duchesne@DMS.Umontreal.CA.

estimateurs de calage moins sensibles aux valeurs aberrantes reposant sur des estimateurs robustes avec haut point de rupture.

## 2. ESTIMATEURS DE CALAGE ROBUSTES

### 2.1 Estimateurs de calage

Soit une population finie  $U = \{1, 2, \dots, N\}$  de taille  $N$  dont nous désirons estimer le total  $t_y = \sum_U y_k$  pour une variable d'intérêt  $y$  positive. On suppose que la population est une réalisation d'un modèle de superpopulation  $\xi$  tel que sous le modèle  $y_k$  et  $x_k$  sont reliés par le modèle de superpopulation  $y_k = x_k' \beta + \sqrt{v_k} \varepsilon_k$ ,  $k = 1, 2, \dots, N$ , où les  $\varepsilon_k$  sont centrés en 0, de variance constante et non-corrélés. On suppose que l'information auxiliaire est unitaire, c'est-à-dire que  $x_k$  est connu de source sûre,  $\forall k \in U$ .

Considérons les estimateurs de calage du total  $t_y$  développés dans Deville et Särndal (1992). Ils peuvent s'écrire comme  $\sum_s w_k y_k$ . On cherche des poids  $w_k$  aussi près que possible des poids d'échantillonnage  $d_k = 1/\pi_k$  mais en respectant les contraintes d'étalonnage (CE)  $\sum_s w_k x_k = t_x$ , où  $x_k$  est l'information auxiliaire de total  $t_x$  connu. Le GREG est un exemple important avec les poids

$$w_k = d_k \left\{ 1 + (t_x - \hat{t}_{x\pi})^t T_s^{-1} x_k / c_k \right\},$$

où  $T_s = \sum_s d_k x_k x_k' / c_k$ , obtenu en minimisant la métrique quadratique  $\sum_s c_k (w_k - d_k)^2 / d_k$ , où les constantes  $c_k$  sont des facteurs de pondération. Puisque les poids-g  $g_k = w_k / d_k$  du GREG ne sont pas bornés en général, d'autres métriques sont proposées afin de les borner pour qu'ils satisfassent certaines restrictions applicables à la fourchette des valeurs (RAFV). Ceci permet en particulier d'éviter la présence indésirable de poids négatifs.

### 2.2 Estimateurs QR et estimateurs restreints QR

Supposons que nous disposons des constantes  $q_k > 0$ ,  $r_k \geq 0$ ,  $\forall k \in U$ . Les estimateurs QR de Wright sont définis à partir des constantes  $q_k$  et  $r_k$  par la relation  $\hat{t}_{QR} = t_x' \hat{B}_q + \sum_s r_k e_k$ , où  $\hat{B}_q$  admet une forme pondérée par les  $q_k$

$$\hat{B}_q = \left( \sum_s q_k x_k x_k' \right)^{-1} \sum_s q_k x_k y_k, \quad (2.2.1)$$

où  $e_k = y_k - x_k' \hat{B}_q$ . Le GREG est un estimateur QR

obtenu en posant  $(q_k, r_k) = (d_k / c_k, d_k)$ . Les estimateurs QR sont des estimateurs de calage obtenus de la résolution de la minimisation de la métrique quadratique suivante sujet aux CE

$$\min \frac{1}{2} \sum_s \frac{(w_k - r_k)^2}{q_k}, \text{ tel que } \sum_s w_k x_k = t_x. \quad (2.2.2)$$

Ainsi les poids  $w_k$  sont choisis aussi près que possible des  $r_k$  sujet aux CE et les  $q_k$  sont des facteurs de pondération. La solution au problème (2.2.2) est donnée par  $w_k = r_k + (t_x - \hat{t}_{xr})^t \left( \sum_s q_k x_k x_k' \right)^{-1} q_k x_k$ . Cependant, rien ne garantit que les poids de l'estimateur QR soient positifs. Afin de restreindre les poids à un intervalle  $[L, U]$ , nous considérons le problème de programmation convexe

$$\min \sum_s G(w_k; q_k, r_k), \text{ tel que } \sum_s w_k x_k = t_x \\ \text{et } w_k \in [L, U]. \quad (2.2.3)$$

La fonction  $G(w; q, r)$  est supposée strictement convexe et dérivable en  $w$  pour  $q$  et  $r$  fixés. Nous notons  $g(u; q, r) = G'(u; q, r)$  et  $h(u; q, r) = g^{-1}(u; q, r)$ . On suppose de plus que  $h(0; q, r) = r$  et  $h'(0; q, r) = q$ . Nous appelons les estimateurs obtenus les estimateurs de calage restreints QR.

La métrique qui retiendra notre attention afin que les poids satisfassent les RAFV est une modification du cas 7 de Deville et Särndal (1992). Nous l'appelons la métrique quadratique restreinte. La fonction  $G$  correspondant au choix de cette métrique est

$$G(w_k; q_k, r_k) = \begin{cases} (w_k - r_k)^2 / 2q_k & \text{si } w_k \in [L, U] \\ \infty & \text{sinon} \end{cases}$$

La fonction  $h$  correspondant au choix de cette métrique est alors donné par

$$h(x_k' \lambda; q_k, r_k) = \begin{cases} L & r_k + q_k x_k' \lambda < L \\ r_k + q_k x_k' \lambda & r_k + q_k x_k' \lambda \in [L, U] \\ U & r_k + q_k x_k' \lambda > U \end{cases}$$

où le paramètre  $\lambda$  est obtenu de la résolution de l'équation suivante

$$\sum_s h(x_k' \lambda; q_k, r_k) x_k = t_x. \quad (2.2.4)$$

L'estimateur final est  $\hat{t}_{yRQR} = \sum_s h(x_k' \lambda_\infty; q_k, r_k) y_k$  où  $\lambda_\infty$  est solution de (2.2.4), calculé selon la méthode de Newton

$$\lambda_{v+1} = \lambda_v - \left( \sum_s h'(x_k' \lambda_v; q_k, r_k) x_k x_k' \right)^{-1} \\ \left( \sum_s h(x_k' \lambda_v; q_k, r_k) x_k - t_x \right)$$

avec  $\lambda_0 = 0$ .

### 2.3 Méthodes de réduction de poids

Lee (1995) discute de diverses propositions reposant sur la méthode de réduction de poids dans un tirage aléatoire simple. Une fois les observations aberrantes détectées, ces méthodes consistent à réduire les poids des observations extrêmes. Traduit dans le langage des estimateurs de calage, commençons par considérer la situation où nous ne disposons pas d'information auxiliaire et où la seule contrainte est  $\sum_s w_k = N$ . Ce cas sert à motiver notre démarche. Considérons l'estimateur QR avec  $q_k = r_k$ . Les poids minimisant (2.2.2) où la seule contrainte est  $\sum_s w_k = N$  sont données par  $w_k = C_s(r)r_k$ , où  $C_s(r) = N/\sum_s r_k$ , de sorte que l'estimateur QR obtenu devient

$$\hat{t}_{yQR} = C_s(r) \sum_s r_k y_k. \quad (2.3.1)$$

Intuitivement, on raisonne de la manière suivante: Une observation extrême représente peu d'unités comme elle dans la population et par conséquent son poids devrait peut-être être réduit. Afin de satisfaire les CE, ceci suggère de trouver des poids  $w_k$  les plus près possible des poids d'échantillonnage  $d_k$  pour les unités qui ne sont pas aberrantes mais aussi près que possible d'un facteur de réduction  $r$  pour les unités aberrantes. Plus précisément, notons  $s = s_1 \cup s_2$ , où  $s_1$  de cardinalité  $n_1$  représente les unités qui ne sont pas déclarées aberrantes, alors que  $s_2 = s - s_1$  de cardinalité  $n_2 = n - n_1$  représente les unités aberrantes de  $s$ . Le facteur de réduction  $r$  satisfera  $r \leq d_k$ ,  $\forall k \in s_2$ . Par exemple, considérons l'estimateur (2.3.1) avec  $q_k = r_k = d_k I_{k1} + r(1 - I_{k1})$ , où  $I_{k1}$  est la variable indicatrice de l'appartenance à  $s_1$ . Ainsi, les constantes  $q_k$  et  $r_k$  sont réduites pour les unités de  $s_2$  afin de refléter que les unités de  $s_2$  sont extrêmes. L'estimateur (2.3.1) devient

$$\hat{t}_{yQR} = C_s(B) \left( \sum_{s_1} d_k y_k + r \sum_{s_2} y_k \right) = C_s(B) \hat{y}_{yB}.$$

Dans le cas du tirage aléatoire simple,  $d_k = N/n$  et  $\hat{t}_{yB}$  est l'estimateur de Bershad (1960) discuté dans Lee (1995). Ce dernier discute de d'autres méthodes de réduction de poids ainsi que du choix du  $r$ .

### 2.4 Estimateurs de calage robustes en présence d'information auxiliaire

Afin d'obtenir les constantes  $q_k$ , nous devons orienter notre choix d'estimateur robuste de régression vers un estimateur pouvant s'écrire sous une forme pondérée de la forme (2.2.1). L'estimateur robuste que nous

allons utiliser est celui proposé par Coakley et Hettmansperger (1993), qui possède entre autres un haut point de rupture. Cependant, comme il ne possède pas une forme pondérée, nous effectuons à la manière de Simpson et Chang (1997) une repondération de cet estimateur. Si nous le notons

$$\hat{B}_{CH}, \text{ on pose alors } q_k = \frac{d_k}{c_k} \hat{u}_k, \text{ où}$$

$$\hat{u}_k = \frac{\psi\left(\frac{y_k - x_k' \hat{B}_{CH}}{\sigma h_k \sqrt{c_k}}\right)}{\left(\frac{y_k - x_k' \hat{B}_{CH}}{\sigma h_k \sqrt{c_k}}\right)'}.$$

et  $h_k = \min\left(1, \frac{\pi_k T}{x_k / \text{med}(x_k)}\right)$ . La fonction  $\psi$  est la fonction de Huber

$$\psi_{Hub}(x; c) = \begin{cases} c & \text{si } x > c \\ x & \text{si } |x| \leq c \\ -c & \text{si } x < -c \end{cases}.$$

Dans les applications, on a choisit  $c = 1.345$ ,  $T = 1.4$  et on estime  $\sigma$  comme dans Coakley et Hettmansperger (1993).

Ayant déterminé les constantes  $q_k$ , nous devons maintenant déterminer les  $r_k$ . Si  $r_k = d_k$  alors l'estimateur est sous des conditions générales un estimateur ADU (asymptotic design-unbiased). Cependant ce choix donne un estimateur sensible aux valeurs aberrantes. Lee (1991) suggère le choix  $r_k = \theta d_k$ . Le choix de  $\theta$  permet de contrôler le biais de l'estimateur. La discussion de la section précédente nous amène à suggérer des constantes  $r_k$  proches des  $d_k$  pour les bonnes unités et graduellement réduites pour les observations suspectes. Nous suggérons le choix  $r_k = d_k u_k^*$  où

$$u_k^* = \frac{\psi\left(\frac{K_{n,N} (y_k - x_k' \hat{B}_q)}{\sigma h_k \sqrt{c_k}}\right)}{K_{n,N} \left(\frac{y_k - x_k' \hat{B}_q}{\sigma h_k \sqrt{c_k}}\right)'},$$

où  $\hat{B}_q$  s'obtient avec le choix de  $q_k$  précédent et  $K_{n,N} = (1 - n/N)^2$ .

Ces choix des constantes  $q_k$  et  $r_k$  fournissent un estimateur QR. Cependant, afin d'obtenir un estimateur avec des poids restreints, nous considérons ces constantes et résolvons le problème d'optimisation (2.2.3) avec la métrique quadratique restreinte.

## 3. BRÈVE ÉTUDE EMPIRIQUE

Nous avons entrepris une étude de simulation de Monte Carlo. Dans quatre populations,  $B=2000$

échantillons ont été tirés par tirage aléatoire simple pour différentes tailles échantillonnelles. L'objectif principal est de savoir si on peut obtenir des estimateurs possédant de bonnes propriétés empiriques (biais, erreur quadratique moyenne) satisfaisant en plus les CE et les RAFV. La première population de taille 51 provient de Mosteller et Tukey (1977, p.560) et concerne la population américaine en 1960 et en 1970 pour les états. Nous la notons POPUSA. La seconde provient de Singh et Chaudhary (1986, p. 177) de taille 34 et concerne la superficie de champs ensemencés en 1971 et en 1974 que nous nommons AREA. La troisième population est la population MU284 avec  $x=S82$  et  $y=P85$  tirée du livre de Särndal et al. (1992). Finalement la dernière population est MU281 avec  $x=REV84$  et  $y=RMT85$  tirée du livre de Särndal et al. (1992). Pour les fins de l'étude, nous considérons le GREG ainsi que le cas 7 de Deville et Särndal (1992). Nous notons ces estimateurs GREG/U et GREG/R. Nous avons choisi  $c_k \equiv 1$  pour les populations POPUSA et AREA, alors que nous avons optés pour  $c_k = x_k$  pour les populations MU284 et MU281. Nous avons également considéré l'estimateur robuste introduit

dans la section (2.4) selon la métrique quadratique et la métrique quadratique restreinte. Parmi les mesures descriptives considérées, nous avons calculé le poids minimum (MIN) ainsi que le poids maximum (MAX). Nous avons également calculé dans quelle pourcentage tous les poids respectaient les contraintes choisies, que l'on retrouve dans la colonne RAFV. Nous avons calculé les biais relatifs, la valeur moyenne, la variance (en millions) et l'erreur quadratique moyenne (en millions) de Monte Carlo avec les formules  $BRMC = t_y^{-1}(EMC - t_y) \times 100$ ,  $EMC = B^{-1} \sum_{i=1}^B \hat{t}_i$ ,  $VMC = B^{-1} \sum_{i=1}^B (\hat{t}_i - EMC)^2$ ,  $MSEMC = B^{-1} \sum_{i=1}^B (\hat{t}_i - t_y)^2$ , respectivement.

Les résultats décrits dans les tableaux 1-4 montrent que tous les estimateurs robustes, au prix d'un biais plus élevé par rapport au GREG/U, sont tout de même plus efficaces en terme d'erreur quadratique moyenne. De plus les estimateurs robustes avec poids contraints se comportent de manière semblable aux estimateurs robustes QRROB/U, mais évitent le problème de la pondération négative.

Table 1: Échantillonnage dans POPUSA avec les contraintes sur les poids [0.20, 32] pour  $n=10$  et [0.20, 16] lorsque  $n=15$ .

ESTIMATEURS	VARMC	MSEMC	BRMC	MIN	MAX	RAFV
n=10						
GREG/U	34.90	34.92	-0.07	-6.24	26.75	86.7
GREG/R	35.29	35.30	-0.04	0.20	32.00	100.0
QRROB/U	27.46	28.44	-0.49	-15.68	40.10	83.1
QRROB/R	28.40	29.22	-0.45	0.20	32.00	100.0
n=15						
GREG/U	21.90	21.95	-0.10	-3.13	15.32	94.7
GREG/R	22.12	22.15	-0.09	0.20	16.00	100.0
QRROB/U	15.15	16.66	-0.60	-4.48	16.06	90.5
QRROB/R	15.32	16.78	-0.59	0.20	16.00	100.0

Table 2: Échantillonnage dans AREA avec les contraintes sur les poids [0.20, 14] pour  $n=10$  et [0.20, 6] lorsque  $n=15$ .

ESTIMATEURS	VARMC	MSEMC	BRMC	MIN	MAX	RAFV
n=10						
GREG/U	1.334	1.700	8.92	-3.35	14.94	86.6
GREG/R	1.295	1.629	8.53	0.20	14.00	100.0
QRROB/U	1.050	1.445	9.27	-4.73	15.38	87.7
QRROB/R	1.053	1.443	9.20	0.20	14.00	100.0
n=15						
GREG/U	0.940	1.178	7.18	-1.40	7.03	93.0
GREG/R	0.928	1.154	7.01	0.20	6.00	100.0
QRROB/U	0.566	0.973	9.41	-1.59	8.90	94.2
QRROB/R	0.566	0.971	9.38	0.20	6.00	100.0

Table 3: Échantillonnage dans MU284 avec les contraintes sur les poids [0.20, 16] pour  $n=30$  et [0.20, 7] lorsque  $n=60$ .

ESTIMATEURS	VARMC	MSEMC	BRMC	MIN	MAX	RAFV
n=30						
GREG/U	2.833	2.925	-3.64	-6.83	23.30	89.8
GREG/R	2.813	2.910	-3.73	0.20	16.00	100.0
QRROB/U	0.693	1.564	-11.19	-9.75	25.84	86.3
QRROB/R	0.694	1.555	-11.13	0.20	16.00	100.0
n=60						
GREG/U	1.473	1.489	-1.49	-1.19	10.03	90.1
GREG/R	1.467	1.484	-1.57	0.20	7.00	100.0
QRROB/U	0.313	0.965	-9.68	-2.54	10.99	86.2
QRROB/R	0.313	0.968	-9.70	0.20	7.00	100.0

Table 4: Échantillonnage dans MU281 avec les contraintes sur les poids [0.20, 25] pour n=30 et [0.20, 9] lorsque n=60.

ESTIMATEURS	VARMC	MSEMC	BRMC	MIN	MAX	RAFV
n=30						
GREG/U	17.33	17.35	-0.26	-38.97	34.56	86.0
GREG/R	17.40	17.41	-0.24	0.20	25.00	100.0
QRROB/U	12.52	12.97	1.27	-55.35	36.95	69.3
QRROB/R	12.53	13.03	1.33	0.20	25.00	100.0
n=60						
GREG/U	7.57	7.57	-0.10	-12.77	15.34	86.4
GREG/R	7.58	7.58	-0.09	0.20	9.00	100.0
QRROB/U	5.49	6.18	1.56	-20.95	21.23	69.4
QRROB/R	5.48	6.19	1.58	0.20	9.00	100.0

## RÉFÉRENCES

- Bershad, M. A. (1960), Some observations on outliers, Unpublished Memorandum, Statistical Research Division, U.S. Bureau of Census.
- Coakley, C. W. et Hettmansperger, T.P. (1993), A bounded influence, high breakdown, efficient regression estimator, *Journal of the American Statistical Association* **88**, 872-880.
- Deville, J.-C. et Särndal, C. E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association* **87**, 376-382.
- Lee, H. (1991), Model-based estimators that are robust to outliers, in *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of Census.
- Lee, H. (1995), Outliers in business surveys, in Cox, Binder, Chinnappa, Christianson, Colledge et Kott, eds, *Business Surveys Methods*, New-York:Wiley.
- Mosteller, F. et Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley Publishing Company.
- Särndal, C. E., Swensson, B. et Wretman, J. H. (1992), *Model Assisted Survey Sampling*, New-York: Springer-Verlag.
- Simpson, D. G. et Chang, Y.-C. I. (1997), Reweighted approximate GM-estimators: asymptotics and residual-based graphics, *Journal of Statistical Planning and Inference* **57**, 273-293.
- Singh, D. et Chaudhary, F. S. (1986), *Theory and Analysis of Sample Survey Designs*, India:Wiley.
- Wright, R. L. (1983), Finite population sampling with multivariate auxiliary information, *Journal of the American Statistical Association* **78**, 879-884.