

## SPÉCIFICATION DU PARAMÈTRE DE LA STRUCTURE DE VARIANCE DU MODÈLE DANS L'ESTIMATEUR DE RÉGRESSION GÉNÉRALISÉ

Eric Pelletier et Eric Rancourt <sup>1</sup>

### RÉSUMÉ

Dans les enquêtes par sondage on effectue souvent l'estimation à l'aide d'information auxiliaire. Un estimateur de choix permettant de tirer profit de cette information est l'estimateur de régression généralisé (GREG). Cet estimateur fournit un cadre de travail flexible et assez vaste; c'est d'ailleurs une des principales options disponibles dans le Système généralisé d'estimation (SGE) de Statistique Canada. Entre autres, il est possible de spécifier plusieurs structures de variance pour le modèle sous-jacent. Cependant les caractéristiques du GREG ne sont pas complètement connues pour toutes ces spécifications. Cet article s'intéresse à l'impact de la spécification du paramètre de la structure de variance du modèle. De plus, on présente différents traitements de données aberrantes qui permettent d'atténuer les effets qu'elles ont lors du calcul de l'estimation et de l'estimation de la variance.

MOTS-CLÉS: Distance de Cook; traitement de données aberrantes; estimation de variance.

### ABSTRACT

In sample surveys, estimation is often performed with the aid of auxiliary information. A commonly used estimator which can utilise this information is the generalized regression (GREG) estimator. This estimator provides good flexibility and can be used in many situations; it is indeed one of the main options available in Statistics Canada's Generalized Estimation System (GES). Within GES, it is possible to specify different variance structures for the underlying model. However, the characteristics of the GREG estimator are not completely known for some specifications. This paper is interested in the impact of the parameter specification for the model variance structure. Also, different outlier treatments which could reduce their effect on point estimation and variance estimation are presented.

KEY WORDS: Cook's distance; Outlier treatment; Variance estimation.

### 1. INTRODUCTION

Lorsqu'on désire estimer un paramètre d'une population finie, il est avantageux de pouvoir utiliser de l'information connue comme source auxiliaire. Dans le cas des sondages, l'estimateur de régression généralisé, présenté dans Särndal, Swensson et Wretman (1992), offre la possibilité de profiter d'une telle information auxiliaire. Cet estimateur, de même qu'un estimateur de sa variance, sont disponibles dans le Système généralisé d'estimation (SGE) de Statistique Canada pour lequel la théorie est décrite dans Estevao, Hidiroglou et Särndal (1994). Par le biais d'un modèle, il est possible d'y spécifier des paramètres tels que les variables auxiliaires, le niveau du modèle (groupe modèle) et la structure de variance

du modèle. Ceci étant fait, on peut alors obtenir les estimations désirées pour tout ensemble de domaines d'intérêt. La possibilité de pouvoir spécifier toute structure de variance offre une très grande flexibilité à l'utilisateur. Par contre, elle mène également à la spécification d'estimateurs dont les propriétés – et surtout celles de l'estimateur de variance – sont jusqu'à maintenant mal connues.

Dans cet article, on s'intéresse au comportement de l'estimateur de variance de l'estimateur de régression généralisé (GREG) sous plusieurs modèles de structure de variance de la population. Nous verrons également que la spécification de cette structure peut conduire directement à des méthodes de traitement des données aberrantes.

---

<sup>1</sup> Eric Pelletier et Eric Rancourt, Division des méthodes d'enquêtes-entreprises, 11<sup>e</sup> étage édifice R.H. Coats, Statistique Canada, Ottawa, Ontario, K1A 0T6

La section 2 présente la notation utilisée dans cet article, de même qu'un survol des diverses utilisations de la structure de variance dans la littérature. À la section 3 se trouve une étude de comportement des estimateurs sous différentes spécifications de la variance de la population. La section 4 développe ensuite le sujet du traitement des données aberrantes, suivie d'un exemple d'application à la section 5.

## 2. ESTIMATION ET STRUCTURE DE VARIANCE

Soit une population  $U = \{1, \dots, k, \dots, N\}$ . On tire un échantillon  $s$  de taille  $n$  pour laquelle on désire estimer le total  $T_Y = \sum_U y_k$  de la variable  $Y$ . Les unités sont tirées avec une probabilité  $\pi_k$  et  $a_k = 1/\pi_k$  est le poids de sondage. Un vecteur de variables auxiliaires,  $\mathbf{x}_k$ , est disponible pour les unités de l'échantillon et on connaît les totaux  $T_X = \sum_U x_k$ . L'estimateur utilisé est l'estimateur GREG, dont le modèle sous-jacent est

$$y_k = \mathbf{x}'_k \beta + \varepsilon_k \quad \text{où } E(\varepsilon_k) = 0 \text{ et } V(\varepsilon_k) = \sigma^2 c_k.$$

La structure de variance est représentée par le facteur  $c_k$ . L'estimateur GREG est

$$\hat{Y}_{GREG} = \sum_s a_k g_k y_k = \left( \sum_U \mathbf{x}_k \right)' \hat{\mathbf{B}} + \sum_s a_k e_k$$

avec

$$g_k = 1 + \left( \sum_U \mathbf{x}_k - \sum_s a_k \mathbf{x}_k \right) \left( \sum_s \frac{a_k \mathbf{x}'_k \mathbf{x}_k}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k},$$

$$\hat{\mathbf{B}} = \left( \sum_s \frac{a_k \mathbf{x}_k \mathbf{x}'_k}{c_k} \right)^{-1} \sum_s \frac{a_k \mathbf{x}_k y_k}{c_k} \quad \text{et } e_k = y_k - \mathbf{x}'_k \hat{\mathbf{B}}.$$

L'estimateur de la variance est

$$\hat{V}(\hat{Y}_{GREG}) = N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_s \left( g_k e_k - \frac{1}{n} \sum_s g_k e_k \right)^2.$$

On voit donc que la spécification du  $c_k$  peut revêtir une certaine importance. Par exemple, si  $c_k = 1$ , on a une structure de variance constante qui ne dépend pas des variables auxiliaires. Dans le cas d'une seule variable auxiliaire, si on spécifie  $c_k = x_k$ , on retrouve

$$\text{l'estimateur par ratio } \hat{Y} = \sum_U x_k \sum_s a_k y_k / \sum_s a_k x_k.$$

On pourrait également spécifier n'importe quelle valeur pour  $c_k$  et même, des valeurs différentes pour des sous-ensembles de la population.

Dans Särndal, Swensson et Wretman (1992), on trouve les estimateurs pour diverses spécifications du  $c_k$  telles que  $c_k = cte$ ,  $c_k = x_k$  ou  $c_k = x_k^2$ . Par ailleurs, Wright (1983) propose d'estimer le paramètre  $p$ , où  $c_k = x_k^p$ . C'est dans le cas de l'échantillonnage avec probabilités proportionnelles à la taille que l'on retrouve quelques détails sur la spécification du paramètre  $p$ . Dans Brewer (1963), on mentionne que  $p$  est presque toujours entre 0 et 2 et on y présente un estimateur. Rao (1978) traite des poids de sondage optimaux qui dépendent de  $p$  et on y retrouve l'estimateur de Brewer (1963) pour le cas de  $p = 1$ . D'autre part, Cochran (1977) mentionne que  $p$  est habituellement entre 1 et 2 dans la plupart des applications. Dans le cas de l'échantillonnage de Poisson, Pelletier (1996) et Särndal (1996) présentent la structure de variance optimale  $c_k = 1/(a_k - 1)$ .

## 3. ÉTUDE DU COMPORTEMENT DU GREG SOUS DIFFÉRENTES STRUCTURES DE VARIANCE

Cette section présente une étude par simulation du comportement de l'estimateur de variance de l'estimateur GREG sous différentes spécifications de la structure de variance.

### 3.1 Création de populations

La première étape a été de créer les différentes populations utilisées lors des simulations. Les populations ont été créées selon le modèle suivant :  $y_k = 1.5x_k + \varepsilon_k$  où  $E(\varepsilon_k) = 0$  et  $E(\varepsilon_k^2) = x^p \sigma^2$ . Ces populations suivent les modèles utilisés dans Lee, Rancourt et Särndal (1994).

Ainsi, quatre populations distinctes ont été créées avec  $P = 0, 1, 2, 3$ . On s'intéresse à différentes structures de variance pour l'estimateur :  $p = 0, 1, 2, 3$ , c'est-à-dire  $c_k = 1$ ,  $c_k = x_k$ ,  $c_k = x_k^2$  et  $c_k = x_k^3$ . On obtient donc 16 combinaisons possibles qui seront évaluées.

**Tableau 1: Biais relatif de l'estimateur de la variance  
(et de l'estimateur ponctuel) en %**

Population	$c_k=1$	$c_k=x_k$	$c_k=x_k^2$	$c_k=x_k^3$
$P=0$	-1,99 (0,002)	0,08 (0,003)	7,67 (0,008)	34,86 (0,024)
$P=1$	-2,32 (0,003)	-0,308 (0,007)	3,02 (0,023)	14,30 (0,072)
$P=2$	-6,75 (-0,205)	-3,66 (-0,029)	-0,00 (0,159)	7,04 (0,418)
$P=3$	-3,37 (0,869)	-2,18 (0,585)	-0,78 (0,392)	0,36 (0,183)

### 3.2 Simulations

Pour toutes les simulations effectuées, des échantillons de taille  $n=25$  ont été tirés parmi chacune des populations de taille  $N=100$ . D'autres fractions de sondages plus petites auraient pu être utilisées comme celles employées dans Hidioglou et Srinath (1981). Pour chacune des simulations, 100 000 échantillons ont été tirés. Pour pouvoir comparer les différentes combinaisons entre elles, plusieurs variables ont été mesurées. Ainsi, la moyenne et la variance de  $\hat{Y}_{GREG}$  sur les 100 000 échantillons ainsi que la moyenne de  $\hat{V}(\hat{Y}_{GREG})$  sur les 100 000 échantillons ont été calculées. Également, le taux de recouvrement de l'intervalle de confiance, le biais relatif de l'estimateur et le biais relatif de l'estimateur de la variance ont aussi été considérés. Le tableau 1 résume quelques résultats des différentes combinaisons.

En tout premier lieu, on remarque que le biais de l'estimateur ponctuel est toujours en deçà de 1% en valeur absolue, ce qui montre que la spécification de la structure de variance ne semble pas cruciale dans le cas de l'estimateur ponctuel. Pour ce qui est maintenant du biais relatif de l'estimateur de la variance, on note que, dans la plupart des cas, la meilleure combinaison possible est celle située sur la diagonale (c'est-à-dire une spécification correcte). On remarque aussi que pour une sous-spécification de  $c_k$ , par exemple  $p=0$  pour la population  $P=2$ , le biais relatif est négatif (-6,75). On est alors en présence d'une sous-estimation de la variance. Il semble donc

préférable d'utiliser une valeur de  $p$  plus grande ou égale à la vraie valeur théorique de  $P$  selon la population utilisée. Également, plus on s'éloigne de la diagonale, plus l'erreur est grande, principalement avec  $p=3$ . La mauvaise spécification de la structure de variance peut donc avoir une grande influence sur l'estimateur de variance. Enfin, signalons que les taux de recouvrement oscillent entre 91% et 96% à l'exception de la population  $P=3$  où la couverture de l'intervalle de confiance est de l'ordre de 62% à 66%.

Ainsi, comme on vient de le voir, la spécification du  $c_k$  a un impact sur l'estimateur de variance. On doit donc s'en préoccuper et tenter de déterminer la valeur de  $P$ . Par contre, son estimation peut s'avérer compliquée, comme on y fait allusion dans Brewer (1963) où l'idée d'une approche empirique est lancée.

### 4. TRAITEMENTS DES DONNÉES ABERRANTES

Comme on l'a vu à la section précédente, il est possible de spécifier une structure de variance différente pour certaines unités. En effet, si  $c_k = x_k$ , on a donc une valeur de  $c_k$  différente pour chaque unité  $k$ . Ainsi, on a donc la possibilité d'atténuer l'effet des données aberrantes à l'aide de la spécification du  $c_k$ . Si on spécifie une valeur de  $c_k$  très grande pour les données aberrantes et plus petite pour les autres, on diminuera grandement leur impact car on obtient un modèle où la variabilité « permise »

est plus grande pour les données aberrantes. Comme chaque unité possède sa propre structure de variance, il est donc possible de séparer la population en sous-populations ayant un modèle similaire et une structure de variance différente. Pour ce faire, trois méthodes ont été considérées pour traiter les valeurs aberrantes :

1. Mise à l'infini du  $c_k$  des données aberrantes
2. Utilisation de la distance de Cook (Cook, 1979)
3. La méthode de Ghangurde (Ghangurde, 1989)

#### 4.1 Mise à l'infini du $c_k$ des données aberrantes

Pour cette première méthode, il s'agit de mettre une valeur de  $c_k$  très grande pour les unités aberrantes et  $c_k = x_k^p$ , par exemple, pour les autres. À noter que la valeur du  $c_k$  est relative à l'échelle des données utilisées. Donc, pour les besoins des simulations, deux valeurs de  $c_k$  ont été considérées,  $c_k = 100$  et  $c_k = 500$ , ce qui est amplement suffisant étant donné l'étendue des données utilisées.

#### 4.2 Utilisation de la distance de Cook

La distance de Cook est une mesure d'influence des données de l'échantillon sur l'estimation des paramètres du modèle. Plus de détails se trouvent dans Cook (1979). Soit  $D_i =$  la distance de Cook. Pour cette méthode-ci, trois cas distincts ont été étudiés :

- Spécifier  $c_k = D_i$  pour les données aberrantes et  $c_k = x_k^p$  pour les autres
- Spécifier  $c_k = D_i / \min(D_i)$  pour les données aberrantes et  $c_k = x_k^p$  pour les autres
- Spécifier  $c_k = D_i / \min(D_i)$  pour toutes les données de l'échantillon.

L'avantage de ces méthodes comparativement à celle de la section 4.1 est qu'elles dépendent de l'échantillon et ne requièrent pas d'intervention externe lors du calcul des estimations. De plus, dans le dernier cas, les données aberrantes n'ont pas à être détectées préalablement. À noter que dans le second cas, si l'étendue des  $x_k$  est grande, il est possible que la correction apportée aux données aberrantes

( $c_k = D_i / \min(D_i)$ ) soit plus petite que le  $c_k$  originalement prévu ( $c_k = x_k^p$ ).

#### 4.3 La méthode de Ghangurde

La méthode de Ghangurde, (Ghangurde, 1989) n'utilise pas la spécification du  $c_k$  pour traiter les données aberrantes, mais puisqu'elle sépare l'échantillon en deux groupes, les données aberrantes et non aberrantes, elle est étudiée pour fins de comparaison. La particularité de cette méthode est dans l'estimation du « bêta » qui est différente de l'estimateur GREG habituel :

$$\hat{B} = \frac{\sum_{s-o} y_k + W \sum_o y_k}{\sum_{s-o} x_k + W \sum_o x_k}$$

où  $s-o$  étant l'ensemble des données non aberrantes,  $o$  étant l'ensemble des données aberrantes et  $W$  est un poids attribué aux données aberrantes. Deux cas particuliers sont intéressants : Si  $W = 0$ , il s'agit d'une repondération, les données aberrantes étant exclues et si  $W = 1$ , aucun traitement particulier n'est appliqué aux données aberrantes.

#### 4.4 Simulations

Pour effectuer les simulations dans le but de comparer les différents traitements pour données aberrantes, les combinaisons de populations ainsi que de structure de variance suivantes ont été utilisées :

- Population  $P = 0$  avec  $c_k = 1$
- Population  $P = 1$  avec  $c_k = x_k$
- Population  $P = 2$  avec  $c_k = x_k^2$

Encore une fois, les populations sont de taille  $N = 100$ , chaque échantillon est de taille  $n = 25$  et 100 000 échantillons ont été tirés. Pour les besoins des simulations, 5 données aberrantes ont été créées

et préalablement identifiées. Le tableau 2 présente des résultats pour la population  $P = 1$ ; des résultats similaires ont été obtenus pour les populations  $P = 0$  et  $P = 2$ .

**Tableau 2: La variance et le biais relatif de l'estimateur  
de la variance en % pour la population  $P = 1$**

Méthodes	Variance	Biais relatif
Sans aucune correction	301 693	-10,20
$c_k$ à l'infini: $c_k=100$	293 806	-12,58
$c_k$ à l'infini: $c_k=500$	339 219	-3,16
Cook: $c_k=D_i$	334 674	-31,69
Cook: $c_k=D_i / \min(D_i)$	358 999	0,42
Cook: tous les $c_k=D_i / \min(D_i)$	321 543	-4,97
Ghangurde	256 490	-2,78

En tout premier lieu, on remarque que lorsqu'aucune correction n'est effectuée, le biais relatif de l'estimateur de la variance est élevé. Pour ce qui est de la méthode de mettre le  $c_k$  à l'infini, comme mentionné à la section 4.1, la valeur du  $c_k$  est une valeur arbitraire. Ainsi, avec la population  $P = 1$ , seulement le cas  $c_k = 500$  fonctionne assez bien en terme du biais relatif. Une spécification plus grande du  $c_k$  serait meilleure, mais demanderait quand même une intervention lors de l'utilisation de cette méthode.

Pour ce qui est de la méthode dérivée à partir de la distance de Cook, le premier cas présenté ( $c_k = D_i$  pour les données aberrantes seulement) ne fonctionne pas du tout. Pour ce qui est des deux autres cas utilisant la distance de Cook ajustée, les deux fonctionnent relativement bien (toujours en terme du biais relatif). En plus, dans le dernier cas les données aberrantes n'ont pas à être détectées préalablement. Finalement, pour la méthode de Ghangurde, elle fonctionne très bien. Par contre, elle est un peu plus complexe que la simple spécification du  $c_k$  dans l'estimateur GREG.

## 5. EXEMPLE D'APPLICATION

Un exemple d'application pour le traitement des données aberrantes à Statistique Canada est l'Enquête sur l'Emploi, la Rémunération et les Heures (EERH).

Dans cette enquête, les heures sont estimées en fonction de l'emploi et de la rémunération à l'aide du modèle suivant :

$$\text{Heures}_k = \beta_1 \text{Emplois}_k + \beta_2 \text{Rémunération}_k + \varepsilon_k$$

où  $E(\varepsilon_k) = 0$  et  $V(\varepsilon_k) = \sigma^2 c_k$ . La détection des données aberrantes se fait à l'aide de la distance de Cook et une donnée est considérée aberrante si la distance de Cook est plus grande que 5. Pour ce qui est de la structure de variance, elle est définie de la façon suivante :

$$c_k = \begin{cases} \text{Emplois} & \text{si } k \text{ est une donnée non aberrante} \\ 10\,000 & \text{si } k \text{ est une donnée aberrante} \end{cases}$$

Pour le traitement des données aberrantes, cette enquête utilise donc la méthode décrite à la section 4.1, c'est-à-dire mettre le  $c_k$  à l'infini. Cette méthode donne des résultats satisfaisants mais étant

donné qu'il s'agit d'une valeur arbitraire, une mise à jour de cette valeur doit être faite à chaque année.

## 6. CONCLUSION

En ce qui a trait à la spécification du  $c_k$ , comme on l'a vu précédemment, il peut y avoir un impact sur

l'estimation de la variance si la structure de variance est mal spécifiée. Également, la spécification du  $c_k$  peut servir à définir des estimateurs optimaux et c'est ce que l'on retrouve habituellement sur le sujet dans la littérature. Dans le Système généralisé d'estimation (SGE), il est possible pour l'utilisateur de spécifier les  $c_k$  pour chaque unité, ce qui permet une grande souplesse dans le choix de l'estimateur. Il est néanmoins de la responsabilité de l'utilisateur de spécifier correctement la structure de variance qui correspond aux données étudiées.

D'autre part, la spécification de la structure de variance permet d'estimer correctement la variance en présence de données aberrantes sans affecter les estimations ponctuelles. Comme on l'a vu, une des méthodes qui donne de bons résultats est la distance de Cook ajustée. Cette méthode semble prometteuse car elle réduit de beaucoup le biais relatif de l'estimateur de variance. Cependant, il reste encore à évaluer ses propriétés plus en profondeur.

## 7. RÉFÉRENCES

- Brewer, K. R. W. (1963). Ratio estimation and finite populations : Some results deducible from the assumption of an underlying stochastic process, *The Australian Journal of Statistics*, 5, 93-105.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley, New York.
- Cook, R. D. (1979). Influential observation in linear regression, *Journal of the American Statistical Association*, 74, 169-174.
- Estevao, V., Hidiroglou, M. A. et Särndal, C.-E. (1995). Methodological principles for a Generalized estimation system at Statistics Canada, *Journal of Official Statistics*, 11, 181-204.
- Ghangurde, P. D. (1989). Outliers in sample surveys, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 736-739.
- Hidiroglou, M. A. et Srinath K. P. (1981). Some estimators of a population total from simple random samples containing large units, *Journal of the American Statistical Association*, 76, 690-695.
- Lee, H., Rancourt, E. et Särndal, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- Pelletier, E. (1996). *Étude comparative d'estimateurs par la régression pour plusieurs plans d'échantillonnage à probabilités inégales*, Mémoire de maîtrise, Université de Montréal.
- Rao, J. N. K. (1978). Sampling designs involving unequal probabilities of selection and robust estimation of a finite population total. *Contributions to survey sampling and applied statistics*, Academic Press, Inc., New York.
- Särndal, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York : Springer-Verlag.
- Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information, *Journal of the American Statistical Association*, 78, 879-884.