Estimation robuste en population finie pour des modèles GLM et GLMM

Cyril Favre Martinoz[1], David Haziza[2] et Nikos Tzavidis[3]

28 avril 2014

[1] *Laboratoire de Statistique d'Enquête, Crest/Ensai, Campus de Ker Lann, 35170 Bruz, France,* cyril.favremartinoz@ensai.fr

[2] *Département de mathématiques et de statistique, Université de Montréal, Montréal, Canada, H3C 3J7* David.Haziza@umontreal.ca

[3] *Université de Southampton, Southampton, Royaume-Uni,* N.TZAVIDIS@soton.ac.uk

Dans les enquêtes auprès des entreprises, il est courant de collecter des variables économiques dont la distribution est fortement asymétrique. Dans ce contexte, on est souvent confronté à la présence de valeurs influentes dans l'échantillon tiré. Ces dernières sont habituellement de très grandes valeurs dont la présence dans l'échantillon tend à rendre les estimateur classiques très instables. Dans une optique basée sur le modèle, Chambers (1986) a proposé une version robuste du BLUP dans le cas d'un modèle linéaire. En utilisant le concept de biais conditionnel Beaumont et Al. (2013) ont proposé une version robuste du BLUP dont la forme est similaire à celle proposée par Chambers (1986). En pratique, il n'est pas rare de rencontrer des variables dichotomiques, ou de comptages, qui rendent le modèle linéaire inadéquat. Nous commençons par présenter une généralisation de ces résultats dans le cas d'un modèle linéaire généralisée. Ces résultats serviront de point de départ à l'estimation robuste dans un contexte d'estimation sur les petits domaines.

Dans le cas d'une estimation sur petits domaines avec un modèle linéaire à effets mixtes entre la variable d'intérêt et l'information auxiliaire, plusieurs auteurs Sinha & Rao (2009), Chambers et Al. (2014) et Dongmo Jiongo et Al. (2013) ont développé des techniques d'estimation robuste. La généralisation au cas d'un modèle linéaire généralisé à effets mixtes sera également discutée.

Mots clés : Biais conditionnel ; Estimation robuste ; Estimateur BLUP ; Valeurs influentes ; Modèle linéaire généralisé.

Dans une approche basée sur le modèle dans une population finie, les valeurs de la variable d'intérêt y sont issus d'un certain modèle. On note X, la matrice de taille $N \times p$ contenant l'information des p variables auxiliaires connus pour les N-unités de la population U. On désigne par x_i^T le vecteur représentant la $i^{ème}$ ligne de la matrice X. On suppose qu'un échantillon non informatif s est sélectionné dans la population finie U, celui-ci sera fixé dans la suite de l'inférence. Ainsi on raisonne conditionnellement à l'échantillon tiré. On souhaite prédire une fonction des variables aléatoires Y issues de la population à l'aide des variables aléatoires observées dans l'échantillon s. Le paramètre d'intérêt considéré dans la suite est le total aléatoire $\theta = \sum_{i \in U} Y_i$. Pour prédire ce total, nous faisons l'hypothèse que la variable aléatoire Y possède une distribution issue de la famille exponentielle : $f_{y_i|\mathbf{x_i}}(y_i) = exp\left(\frac{y_i\gamma_i - b(\gamma_i)}{A(\phi)} + c(y_i, A(\phi))\right)$ avec $\mu_i = E(y_i) = \frac{\partial b(\gamma_i)}{\partial \gamma_i}$ et $V(\mu_i) = \frac{\partial^2 b(\gamma_i)}{\partial^2 \gamma_i}$.

De plus, nous supposons que la variable Y dépend des variables auxiliaires X suivant un modèle linéaire généralisé : $E(Y) = F(\mathbf{x_i^T}\beta)$ où F est une fonction de lien classique des modèles GLM.

On définit la log-vraisemblance par :

$$l(\beta, \mathbf{y}, \mathbf{X}) = log \prod_{i \in U} f_{y_i|\mathbf{x_i}}(y_i).$$

Le maximum de vraisemblance $\tilde{\beta}$ vérifie alors l'équation estimante suivante :

$$\sum_{i \in U} \frac{Y_i - F(\boldsymbol{x_i^T\beta})}{A(\phi)V(\mu_i)} \frac{\partial F(u)}{\partial u}(\mathbf{x_i^T}\beta)\boldsymbol{x_i} = 0.$$

Malheureusement, $\tilde{\beta}$ ne peut être estimé, car l'équation estimante requiert les valeurs de Y_i sur l'ensemble de la population, or celles-ci ne sont connues que sur l'échantillon s. On considère donc l'équation estimante rapportée à l'échantillon s :

$$\sum_{i \in S} \frac{Y_i - F(\boldsymbol{x_i^T\beta})}{A(\phi)V(\mu_i)} \partial F(u)(\mathbf{x_i^T}\beta)\boldsymbol{x_i} = 0.$$

et on note : $g(x_i, y_i, \beta) = \frac{Y_i - F(\boldsymbol{x_i^T\beta})}{A(\phi)V(\mu_i)} \frac{\partial F(u)}{\partial u}(\mathbf{x_i^T}\beta)\boldsymbol{x_i}$ et $\boldsymbol{H}(\boldsymbol{x_i}, \boldsymbol{\beta}) = \frac{\partial g(u)}{\partial u}(\boldsymbol{x_i^T\beta})$.

Royall et Al. (1976) ont proposé le meilleur estimateur linéaire sans biais (BLUP) du paramètre $\theta = \sum_{i \in U} Y_i$, qui prend la forme suivante :

$$\hat{\theta}^{BLUP} = \sum_{i \in S} Y_i + \sum_{i \in U \setminus S} F(\mathbf{x_i^T}\hat{\beta}).$$

Nous allons maintenant déterminer le biais conditionnel associé à l'estimateur $\hat{\theta}^{BLUP}$. Dans le cas d'une approche sous le modèle, le biais conditionnel associé à l'estimateur $\hat{\theta}^{BLUP}$ pour l'unité i est défini par : $B_i(y_i; \beta) = E(\hat{\theta}^{BLUP} - \theta | s; Y_i = y_i)$.

L'estimateur $\hat{\theta}^{BLUP}$ est non linéaire, nous procédons à un développement de Taylor afin de déterminer le biais conditionnel de cet estimateur.

$$F(x_i^T\hat{\beta}) \;=\; F(\mathbf{x_i^T}\beta) + \frac{dF(u)}{du}(\mathbf{x_i^T}\beta)\mathbf{x_i^T}(\hat{\beta}-\beta) + O_p\left(\frac{1}{n^{1/2}}\right). \tag{1}$$

En suivant la démarche de Fuller (2011, page 65), nous avons :

$$\hat{\beta}-\beta \;=\; \left(\sum_{j\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta})\right)^{-1}\sum_{k\in S}\boldsymbol{g}(y_k,\boldsymbol{x_k},\boldsymbol{\beta}) + o_p\left(\frac{1}{n^{1/2}}\right). \tag{2}$$

En combinant (1) and (2), nous obtenons :

$$F(x_i^T\hat{\beta}) = F(x_i^T\beta) + \frac{dF(u)}{du}(\mathbf{x_i^T}\beta)\mathbf{x_i^T}\left(\sum_{j\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta})\right)^{-1}\sum_{k\in S}\boldsymbol{g}(y_k,\boldsymbol{x_k},\boldsymbol{\beta}) + O_p\left(\frac{1}{n^{1/2}}\right)$$

Il faut maintenant distinguer deux types d'unités, les unités qui sont sélectionnées dans l'échantillon et les unités non sélectionnées.

Si l'unité est sélectionnée,$i \in S$, le biais conditionnel est :

$$\begin{aligned}
B_i^{BLUP}(I_i=1) &= E_m(\hat{\theta}^{BLUP}-\theta|s,Y_i=y_i)\\
&\approx E_m(\sum_{j\in S}Y_j + \sum_{j\in U\setminus S}F(\mathbf{x_j^T}\beta) - \sum_{j\in U}Y_j \,|\, s,Y_i=y_i)\\
&+ E_m(\sum_{j\in U\setminus S}\frac{dF(u)}{du}(\mathbf{x_j^T}\beta)\mathbf{x_j^T}\left(\sum_{j\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta})\right)^{-1}\sum_{k\in S}\boldsymbol{g}(y_k,\boldsymbol{x_k},\boldsymbol{\beta}) \,|\, s,Y_i=y_i)\\
&\approx y_i + \sum_{j\in S,\, j\neq i}F(\mathbf{x_j^T}\beta) + \sum_{j\in U\setminus S}F(\mathbf{x_j^T}\beta) - y_i - \sum_{j\in U,\, j\neq i}F(\mathbf{x_j^T}\beta)\\
&+ \sum_{j\in U\setminus S}\frac{dF(u)}{du}(\mathbf{x_j^T}\beta)\mathbf{x_j^T}\left(\sum_{j\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta})\right)^{-1}E_m(\sum_{k\in S}\boldsymbol{g}(y_k,\boldsymbol{x_k},\boldsymbol{\beta}) \,|\, s,Y_i=y_i)\\
&\approx \sum_{j\in U\setminus S}\frac{dF(u)}{du}(\mathbf{x_j^T}\beta)\mathbf{x_j^T}\left(\sum_{j\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta})\right)^{-1}E_m(\sum_{k\in S}\boldsymbol{g}(y_k,\boldsymbol{x_k},\boldsymbol{\beta}) \,|\, s,Y_i=y_i)
\end{aligned}$$

Voici quelques cas particuliers importants, pour lesquels on peut simplifier l'expression du biais conditionnel :

1) le cas linéaire : $F = Id$ et$g(x_i,y_i,\beta) = (y_i - x_i^T\beta)$

$$B_i^{BLUP}(I_i=1) \;=\; E_m(\hat{\theta}^{BLUP}-\theta|s,Y_i=y_i)$$

$$= \sum_{j\in U\setminus S} \frac{dF(u)}{du}(\mathbf{x_j^T}\beta)\mathbf{x_j^T}(\sum_{k\in S}\mathbf{x_k}\mathbf{x_k^T})^{-1}E_m(\sum_{j\in S}(y_j-\mathbf{x_j^T}\beta)\mathbf{x_j}\,|\,s,Y_i=y_i)$$
$$= \sum_{j\in U\setminus S}\mathbf{x_j^T}(\sum_{k\in S}\mathbf{x_k}\mathbf{x_k^T})^{-1}\mathbf{x_i}\left(y_i-\mathbf{x_i^T}\beta\right)$$

2) le cas de la régression logistique : $g(x_i,y_i,\beta)=\left((y_i-F(x_i^T\beta)\right)\mathbf{x_i}$ et :

$$B_i^{BLUP}(I_i=1) = \sum_{j\in U\setminus S}\frac{dF(u)}{du}(\mathbf{x_j^T}\beta)\mathbf{x_j^T}(\sum_{k\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta}))^{-1}E_m(\sum_{j\in S}(y_j-\mathbf{x_j^T}\beta)\mathbf{x_j}\,|\,s,Y_i=y_i)$$
$$= \sum_{j\in U\setminus S}F(\mathbf{x_j^T}\beta)(1-F(\mathbf{x_j^T}\beta))\mathbf{x_j^T}\left(\sum_{k\in S}F(\mathbf{x_k^T}\beta)(1-F(\mathbf{x_k^T}\beta))\mathbf{x_k}\mathbf{x_k^T}\right)^{-1}\mathbf{x_i}\left((y_i-F(\mathbf{x_i^T}\beta)\right)$$

avec $F(\mathbf{x_i^T}\beta)=\frac{exp(\mathbf{x_i^T}\beta)}{1+exp(\mathbf{x_i^T}\beta)}$.

3) le cas de la régression de Poisson :

$$B_i^{BLUP}(I_i=1) = E_m(\hat{\theta}^{BLUP}-\theta|s,Y_i=y_i)$$
$$= \sum_{j\in U\setminus S}\frac{dF(u)}{du}(\mathbf{x_j^T}\beta)\mathbf{x_j^T}(\sum_{k\in S}\boldsymbol{H}(\boldsymbol{x_j},\boldsymbol{\beta}))^{-1}E_m(\sum_{j\in S}(y_j-\mathbf{x_j^T}\beta)\mathbf{x_j}\,|\,s,Y_i=y_i)$$
$$= \sum_{j\in U\setminus S}F(\mathbf{x_j^T}\beta)\mathbf{x_j^T}\left(\sum_{k\in S}F(\mathbf{x_k^T}\beta)\mathbf{x_k}\mathbf{x_k^T}\right)^{-1}\mathbf{x_i}\left((y_i-F(\mathbf{x_i^T}\beta)\right)$$

where $F(x_i^T\beta)=exp(x_i^T\beta)$.

Si l'unité i n'est pas sélectionnée, i.e $i\in U\setminus S$, le biais conditionnel est donné par :

$$B_i^{BLUP}(I_i=0)=-(y_i-F(\mathbf{x_i^T}\beta)).$$

Le biais conditionnel est inconnu, mais il peut être estimé en remplaçant le paramètre inconnu β par son estimateur issu de l'échantillon $\hat{\beta}$. Par exemple, dans le cas d'un modèle logistique, le biais conditionnel $B_i^{BLUP}(I_i=1)$ associé à l'estimateur BLUP de l'unité i est donné par :

$$\hat{B}_i^{BLUP}(I_i=1)=\sum_{j\in U\setminus S}F(\mathbf{x_j^T}\hat{\beta})(1-F(\mathbf{x_j^T}\hat{\beta}))\mathbf{x_j^T}\left(\sum_{k\in S}F(\mathbf{x_k^T}\hat{\beta})\mathbf{x_k}\mathbf{x_k^T}\right)^{-1}\mathbf{x_i}\left((y_i-F(\mathbf{x_i^T}\hat{\beta})\right).$$

On peut montrer que l'erreur de prédiction de l'estimateur BLUP peut se décomposer de la façon suivante :

$$\hat{\theta}^{BLUP}-\theta=\sum_{i\in U\setminus S}B_i^{BLUP}(I_i=0)+\sum_{i\in S}B_i^{BLUP}(I_i=1).$$

En suivant la démarche de Beaumont et Al. (2013), nous proposons la forme suivante pour l'estimateur robuste :

$$\hat{\theta}^{RBLUP} = \hat{\theta}^{BLUP} - \sum_{i \in S} B_i^{BLUP}(I_i = 1) + \sum_{i \in S} \psi_c \left(B_i^{BLUP}(I_i = 1) \right),$$

où $\psi_c(.)$ est la fonction de Huber.

On souhaite maintenant déterminer le biais conditionnel de l'unité i associé à l'estimateur robuste $\hat{\theta}^{RBLUP}$, $B_i^R(I_i = 1) = E_m(\hat{\theta}^{RBLUP} - \theta | s, Y_i = y_i)$.

On peut montrer que ce biais conditionnel peut s'écrire de la façon suivante :

$$B_i^R(I_{1i} = 1) = B_i^{BLUP}(I_i = 1) + n\bar{\Delta}(c),$$

où

$$\bar{\Delta}(c) = \frac{1}{n} \sum_{i \in S} \left[\psi \left(B_i^{BLUP}(I_i = 1) \right) - B_i^{BLUP}(I_i = 1) \right]$$

Ce biais conditionnel est inconnu, mais il peut être estimé par :

$$\hat{B}_i^R(I_{1i} = 1) = \hat{B}_i^{BLUP}(I_i = 1) + \sum_{i \in S} \left[\psi \left(\hat{B}_i^{BLUP}(I_i = 1) \right) - \hat{B}_i^{BLUP}(I_i = 1) \right].$$

Notons $\hat{B}_{min} = \min \left(\hat{B}_i^{BLUP}(I_{1i} = 1); c \right)$ et $\hat{B}_{max} = \max \left(\hat{B}_i^{BLUP}(I_{1i} = 1); c \right)$, on peut montrer que la valeur c qui minimise $\max\{\hat{B}_i^R(I_{1i} = 1 | i \in S\}$ appelée c_{minmax} , engendre l'estimateur suivant :

$$\hat{\theta}^{RBLUP}(c_{minmax}) = \hat{\theta}^{BLUP} - \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max}). \tag{3}$$

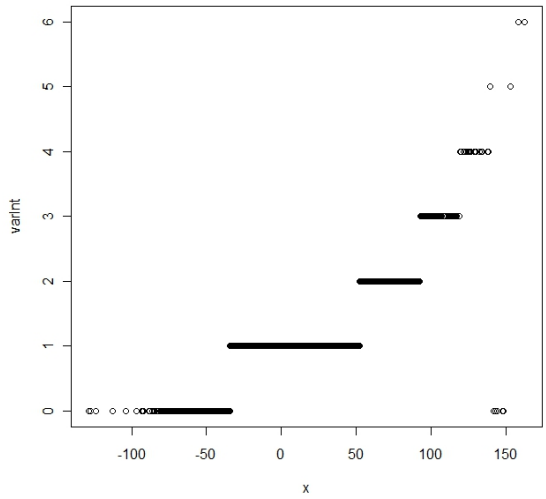
Afin de tester l'efficacité de cet estimateur, nous allons le comparer à un autre estimateur robuste proposé par Cantoni et Ronchetti (2001)

$$\hat{\theta}^{RCantoni} = \sum_{i \in S} Y_i + \sum_{i \in U \setminus S} F(\mathbf{x_i^T} \hat{\beta}^R).$$

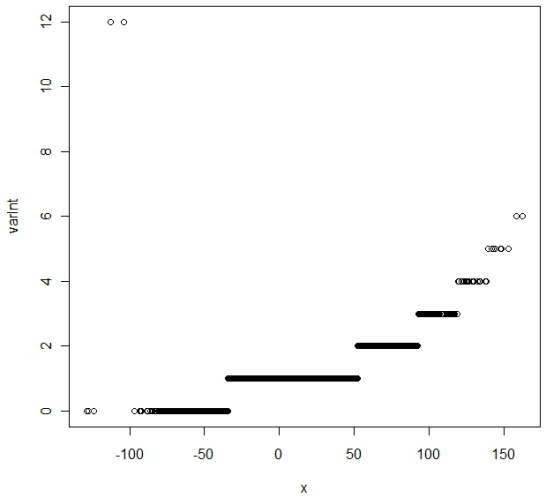
où le coefficient $\hat{\beta}^R$est estimé à l'aide de M-estimateur.

Etude par simulations:

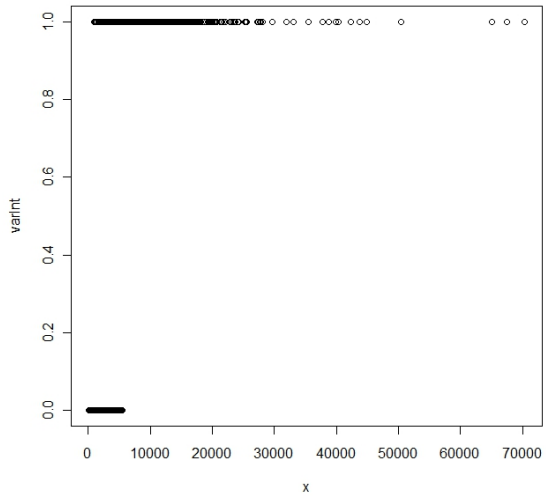
On génére deux jeux de deux populations pour le cas d'une régression logistique et d'une régression de Poisson. Chaque population est contaminée par une proportion assez faibles d'unités influentes. Les quatre populations sont présentées sur le graphique ci-dessous.



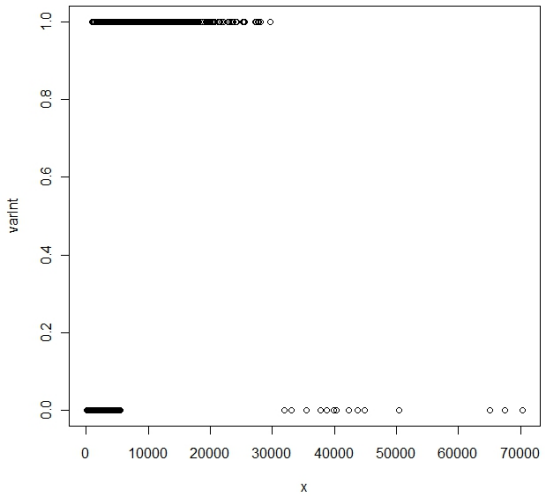
(a) Population 1



(b) Population 2



(c) Population 3



(d) Population 4

Figure 1: Boxplot

Pour comparer les deux estimateurs, on détermine le biais relatif (RB) et l'efficacité relative (RE) par Monte Carlo à l'aide des expressions suivantes :

$$RB_{MC}\left(\hat{\theta}_p^R\right) = \frac{E_{MC}\left(\hat{\theta}_p^R\right)}{\theta} \times 100,$$

où

$$E_{MC}\left(\hat{\theta}_p^R\right) = \frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_p^R - \theta\right),$$

et

$$RE_{MC}\left(\hat{\theta}_p^R, \hat{\theta}\right) = \frac{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_p^R - \theta\right)^2}{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_p - \theta\right)^2} \times 100.$$

Population	Taille de l'échantillon	$\hat{\theta}^{RBLUP}$	$\hat{\theta}^{RCantoni}$
1	100	0.14(93)	6.4(631)
	500	0.10(93)	5.9(2500)
2	100	−0.28(71)	5.7(340)
	500	−0.20(73)	5.2(130)
3	100	0.05(101)	0.2(115)
	500	0.03(100)	0.14(113)
4	100	0.14(93)	0.8(86)
	500	0.06(95)	0.9(96)

Table 1: Biais en pourcentage et efficacité en parenthèse des estimateurs robustes par rapport à l'estimateur BLUP

A travers cette étude par simulation, on constate que l'estimateur robuste (3) est au moins aussi efficace que l'estimateur $BLUP$. De plus, le biais de cet estimateur est relativement faible, moins de 1% dans tous les scénarii envisagés.

References

[1] J-F Beaumont, David Haziza, and Anne Ruiz-Gazen. A unified approach to robust estimation in finite population sampling. *Biometrika*, 2013.

[2] Ray Chambers, Hukum Chandra, Nicola Salvati, and Nikos Tzavidis. Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):47–69, 2014.

[3] R.L. Chambers. Outlier robust finite population estimation. *Journal of the American Statistical Association*, pages 1063–1069, 1986.

[4] Wayne A Fuller. *Sampling statistics*, volume 560. John Wiley & Sons, 2011.

[5] V Dongmo Jiongo, D Haziza, and P Duchesne. Controlling the bias of robust small-area estimators. *Biometrika*, 100(4):843–858, 2013.

[6] Sanjoy K Sinha and JNK Rao. Robust small area estimation. *Canadian Journal of Statistics*, 37(3):381–399, 2009.