Estimation robuste dans les enquêtes: une approche unifiée

David Haziza

Département de mathématiques et de statistique
Université de Montréal
et
Crest-Ensai

Séminaire de Méthodologie de l'INSEE
Paris, France

2 juillet 2013

PLAN

Cette présentation est essentiellement basée sur deux articles:

(1) Beaumont, Haziza et Ruiz-Gazen (2013, *Biometrika*)

(2) Favre Martinoz, Haziza et Beaumont (2013, *Soumis pour publication*).

- Qu'est-ce qu'une unité influente?
- Comment mesurer l'influence d'une unité?
- Comment traiter les unités influentes à l'étape de l'estimation?
- Étude empirique
- Estimation pour des domaines

Unité influente vs. erreur de mesure

- Une unité influente fait partie de la population
- Ce n'est pas une erreur de mesure
  - erreur grossière, erreur d'unité
  - les erreurs de mesures sont habituellement détectées à l'étape de la vérification (editing) et sont traitées soit manuellement soit par imputation
- Dans la suite, on supposera que toutes les erreurs de mesures ont été traitées à l'étape de la vérification $\rightarrow$ il n'en reste plus à l'étape de l'estimation.

Unité influente vs. erreur de mesure

- Quelles sont les raisons conduisant à des unités influentes?
  - La distribution des variables collectées est fortement asymétrique (queues lourdes)
  - Certaines unités peuvent être associées à des poids extrêmes
  - La base de sondage peut contenir des erreurs importantes
- Une unité influente échantillonnée peut représenter d'autres unités similaires dans la portion non-échantillonnée
- Problème particulièrement important dans les enquêtes auprès des entreprises

Configuration

- Avant de définir le concept d'unité influente, on définit celui de configuration.
- Une configuration est un quadruplet, qui consiste en
  (1) une variable d'intérêt
  (2) un paramètre que l'on cherche à estimer
  (3) un plan de sondage
  (4) un estimateur
- Exemples de configurations:
  - (Revenu, revenu total, échantillonnage stratifié aléatoire simple, estimateur d'Horvitz-Thompson)
  - (Revenu, revenu total, échantillonnage stratifié aléatoire simple, estimateur par calage)
  - (Revenu, revenu total, échantillonnage stratifié Poisson, estimateur d'Horvitz-Thompson)

Unité influente

Définition

Une unité est dite influente si, étant donné une configuration, elle a un impact significatif sur l'erreur due à l'échantillonnage, $\hat{\theta} - \theta$.

- Les estimateurs usuels (estimateur d'Horvitz-Thompson et estimateurs par calage) sont sensibles à la présence d'unités influentes.
- Inclure ou exclure une unité influente de l'échantillon peut avoir un très grand impact sur les estimations $\rightarrow$ très grande variance.
- Ce n'est pas un problème de biais car, en l'absence d'erreurs non dues à l'échantillonnage, les estimateurs usuels sont (asymptotiquement) sans biais.
- Une unité peut être très influente par rapport à une configuration et n'avoir aucune influence par rapport à une autre configuration.
- La modification d'un seul élément du quadruplet peut avoir un impact très important sur l'influence d'une unité.

Unité influente

- Objectif: réduire l'influence des unités qui ont une grande influence $\rightarrow$ estimateurs biaisés mais plus stables .
- Traitement des unités influentes: compromis entre biais et variance.
- Estimateur robuste: estimateur dont l'erreur quadratique moyenne est inférieure à celle des estimateurs non robuste (par ex., Horvitz-Thompson).
- Souhait: s'il n'y a pas d'unité influente, on aimerait que les estimateurs robuste ne soient pas beaucoup moins efficaces que les estimateurs non robustes correspondant.

Mesure d'influence: le biais conditionnel

- U: population finie de taille N.
- θ: paramètre à estimer.
- S: échantillon de taille n tiré selon un plan de sondage $p(S)$.
- π_i: probabilité d'inclusion de l'unité i.
- $d_i = \pi_i^{-1}$: poids de sondage de l'unité i.
- I_i: variable indicatrice de selection telle que $I_i = 1$ si $i \in S$ et $I_i = 0$, sinon.
- $\hat{\theta}$: estimateur sans biais par rapport au plan de sondage pour θ

$$B\left(\hat{\theta}\right) = E_p(\hat{\theta}) - \theta = 0.$$

- Interprétation: moyenne de l'erreur due à l'échantillonnage prise sur tous les échantillons possibles tirés de la population finie.

Mesure d'influence: le biais conditionnel

- Soit i une unité échantillonnée (i.e., $I_i = 1$).
- Le biais conditionnel associé à l'unité i par rapport à $\hat{\theta}$ est défini selon

$$B_{1i} = E_p(\hat{\theta}|I_i = 1) - \theta.$$

- Interprétation de B_{1i}: moyenne de l'erreur due à l'échantillonnage prise sur tous les échantillons qui contiennent l'unité i.
- Biais conditionnel B_{1i}: mesure de l'influence d'une unité échantillonnée.
- voir Moreno-Rebollo et al. (1999) et Beaumont et al. (2013).

Mesure d'influence: le biais conditionnel

- Soit i une unité non-échantillonnée (i.e., $I_i = 0$). Le biais conditionnel associé à l'unité i par rapport à $\hat{\theta}$ est défini selon

$$B_{0i} = E_p(\hat{\theta}|I_i = 0) - \theta.$$

- Interprétation of B_{0i}: moyenne de l'erreur due à l'échantillonnage prise sur tous les échantillons qui ne contiennent pas l'unité i.
- Biais conditionnel B_{0i}: mesure de l'influence d'une unité non-échantillonnée.
- Contrairement aux unités échantillonnées, rien ne peut être fait pour réduire l'influence des unités non-échantillonnées à l'étape de l'estimation car leur biais conditionnel est inconnu et ne peut être estimé au moyen des observations dans l'échantillon.

Biais conditionnel: estimateur d'Horvitz-Thompson

- $\theta = t_y = \sum_{i \in U} y_i$: total sur la population de la variable d'intérêt y.
- Estimateur d'Horvitz-Thompson de t_y:

$$\hat{t}_{HT} = \sum_{i \in S} d_i y_i.$$

Go to

- Biais conditionnel de l'unité échantillonnée i par rapport à $\hat{t}_{HT}$:

$$\begin{aligned} B_{1i}^{HT} &= E_p(\hat{t}_{HT} | I_i = 1) - t_y \\ &= (d_i - 1) y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j \\ &= \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j. \end{aligned}$$

Un fichier de données typique...

Unit	y_1	y_2	...	y_p	d_i
1	y_{11}	y_{21}	...	y_{p1}	d_1
2	y_{12}	y_{22}	...	y_{p2}	d_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
n	y_{1n}	y_{2n}	...	y_{pn}	d_n

Remarques

- Le biais conditionnel dépend des probabilités d'inclusion d'ordre deux, π_{ij}.
- Le biais conditionnel est, en général, inconnu $\rightarrow$ il faut l'estimer!
- Si $\pi_i = 1$, alors

$$B_{1i}^{HT} = 0.$$

Biais conditionnel: cas particuliers

- Échantillonnage aléatoire simple sans remise: $\pi_i = n/N$ pour tout i et $\pi_{ij} = n(n-1)/N(N-1)$, $i \neq j$.

$$B_{1i}^{HT} = \frac{N}{(N-1)}\left(\frac{N}{n} - 1\right)(y_i - \bar{Y}_U),$$

où $\bar{Y}_U = t_y/N$ est la moyenne de la population.

- Échantillonnage de Poisson: $\pi_{ij} = \pi_i \pi_j$, $i \neq j$

$$B_{1i}^{HT} = (d_i - 1)y_i = (d_i - 1)(y_i - 0)$$

Lien entre le biais conditionnel et l'erreur due à l'échantillonnage

- Pour l'échantillonnage de Poisson, l'erreur due à l'échantillonnage peut s'écrire comme:

$$\hat{t}_{HT} - t_y = \sum_{i \in S} B_{1i}^{HT} + \sum_{i \in U-S} B_{0i}^{HT}.$$

- Interprétation: le biais conditionnel de l'unité i (échantillonnée ou non) est une mesure de sa contribution à l'erreur due à l'échantillonnage .
- Pour l'échantillonnage aléatoire simple sans remise et les plans à grande entropie:

$$\hat{t}_{HT} - t_y \approx \sum_{i \in S} B_{1i}^{HT} + \sum_{i \in U-S} B_{0i}^{HT}$$

pourvu que la taille de la population N soit grande.

Lien entre le biais conditionnel et la variance de $\hat{t}_{HT}$

- Pour un plan de sondage arbitraire, la variance de $\hat{t}_{HT}$ est donnée par

$$\begin{aligned} V_p\left(\hat{t}_{HT}\right) &= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_i y_j \\ &= \sum_{i\in U} B_{1i}^{HT} y_i \end{aligned}$$

Estimation du biais conditionnel

- En général, le biais conditionnel d'une unité échantillonnée, B_{1i}^{HT}, est inconnu et doit être estimé.
- Rappel:

$$B_{1i}^{HT} = \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j.$$

- Exception: Échantillonnage de Poisson + estimateur d'Horvitz-Thompson $\rightarrow B_{1i}^{HT} = (d_i - 1) y_i$
- Comment doit-on estimer le biais conditionnel? De manière robuste ou non robuste?
- Comme nous le verrons empiriquement, il y a peu de différences entre les deux stratégies...

Estimation du biais conditionnel

- Un estimateur non robuste du biais conditionnel est donnée par

$$\hat{B}_{1i}^{HT} = \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) y_j,$$

  pourvu que $\pi_{ij} > 0$ pour tout $j \in U$.

- On a

$$E_p \left(\hat{B}_{1i}^{HT} | I_i = 1 \right) = B_{1i}^{HT}.$$

- $\hat{B}_{1i}^{HT}$ est un estimateur conditionnellement sans biais de $B_{1i}^{HT} \Rightarrow$ inconditionnellement sans biais
- Peut être calculé pour la plupart des plans de sondage

Estimation du biais conditionnel

Échantillonnage aléatoire simple sans remise:

- Rappel:

$$B_{1i}^{HT} = \frac{N}{(N-1)}\left(\frac{N}{n}-1\right)(y_i - \bar{Y}_U)$$

- Estimateur non robuste de B_{1i}^{HT}:

$$\hat{B}_{1i}^{HT} = \frac{n}{n-1}\left(\frac{N}{n}-1\right)(y_i - \bar{y}),$$

où $\bar{y} = \sum_{i\in S} y_i/n$ désigne la moyenne de l'échantillon.

- Estimateur alternatif de B_{1i}^{HT}:

$$\hat{B}_{1i}^{HT} = \frac{N}{N-1}\left(\frac{N}{n}-1\right)(y_i - m),$$

où m désigne la médiane de l'échantillon.

Traitement à l'étape du plan de sondage

- Meilleur traitement: prévention !
- Étapes du plan de sondage: l'impact des unités influentes peut être contrôlé au moyen d'un plan de sondage approprié.
- Exemple: plan stratifié aléatoire simple sans remise avec une strate exhaustive (take all stratum):
  - Plan de sondage utilisé dans les enquêtes auprès des entreprises.
  - Habituellement, les strates exhaustives contiennent les grandes unités.
  - Les unités appartenant à la strate exhaustive ont une probabilité d'inclusion égale à 1 $\rightarrow$ leur influence est égale à 0 et elles ne contribuent pas à la variance des estimateurs.
- Malgré une bonne planification à l'étape du plan de sondage, le problème des unités influentes n'est jamais complètement éliminé:
  - Les enquêtes recueillent généralement plusieurs dizaines (voire plusieurs centaines) de variables alors qu'un nombre limité de variables est utilisé à des fins de stratification.
  - Problème des "Stratum jumpers".

Traitement des valeurs influentes à l'étape de l'estimation

- On considère la classe des estimateurs robustes:

$$\hat{t}_{HT}^{R}(K) = \hat{t}_{HT} + \Delta(K),$$

où $\Delta(K)$ est une variable aléatoire.

- K: seuil à déterminer
- Plusieurs estimateurs robustes appartiennent à cette classe.
- On cherche la valeur de $\Delta(K)$ qui minimise

$$\max_{i \in S} \left\{ |\hat{B}_{1i}^{R}| \right\},$$

où $\hat{B}_{1i}^{R}$ désigne le biais conditionnel associé à l'unité i par rapport à $\hat{t}_{HT}^{R}(K)$

Traitement des valeurs influentes à l'étape de l'estimation

- Un mauvais choix du seuil K peut grandement affecter les propriétés des estimateurs robustes $\Rightarrow$ peuvent même exhiber une erreur quadratique moyenne plus grande que celle des estimateurs non robustes!
- Premier critère: "Le doigt mouillé"
- Deuxième critère: Déterminer K qui minimise l'erreur quadratique moyenne estimée de l'estimateur robuste:
  - Requiert de l'information historique et/ou un modèle
  - Requiert souvent des hypothèses simplificatrices
  - En général, complexe à implémenter
  - voir, par exemple, Hulliger (1995), Kokic et Bell (1994), Rivest et Hurtubise (1995).

Traitement des valeurs influentes à l'étape de l'estimation

- On a

$$B_{1i}^R = E_p\{\hat{t}_{HT}^R(K)|I_i = 1\} - t_y = B_{1i}^{HT} + E_p(\Delta(K)|I_i = 1)$$

- Estimateur de B_{1i}^R:

$$\hat{B}_{1i}^R = \hat{B}_{1i}^{HT} + \Delta(K)$$

- $E_p\left(\hat{B}_{1i}^R \mid I_i = 1\right) = B_{1i}^R \rightarrow \hat{B}_{1i}^R$ conditionnellement sans biais pour B_{1i}^R
- Beaumont et al. (2013): la valeur $\Delta(K)$ qui minimise $\max_{i\in S}\left\{|\hat{B}_{1i}^R|\right\}$ est donnée par

$$\Delta(K_{opt}) = -\frac{1}{2}(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}),$$

où $\hat{B}_{min}^{HT} = \min_{i\in S}(\hat{B}_{1i}^{HT})$ et $\hat{B}_{max}^{HT} = \max_{i\in S}(\hat{B}_{1i}^{HT})$.

Traitement des valeurs influentes à l'étape de l'estimation

- Ce principe de type min-max conduit à l'estimateur robuste

$$\begin{aligned}\hat{t}_{HT}^{R}(K_{opt}) &= \hat{t}_{HT} + \Delta(K_{opt}) \\ &= \hat{t}_{HT} - \frac{1}{2}(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}).\end{aligned}$$

- Sous certaines conditions de régularité, on a

$$\hat{t}_{HT}^{R}(K_{opt}) - t_y = O_p(Nn^{-1/2}).$$

$\rightarrow \hat{t}_{HT}^{R}(K_{opt})$ est convergent par rapport au plan de sondage.

Cas particuliers de $\hat{t}_{HT}^{R}(K)$: estimateur winsorisé standard

- Soit $\tilde{y}_i$ la valeur de la variable y après winsorisation. On a

$$\tilde{y}_i = \begin{cases} y_i \text{ if } d_i y_i \leq K \\ \frac{K}{d_i} \text{ if } d_i y_i > K \end{cases}$$

- L'estimateur winsorisé standard est donné par

$$\hat{t}_{stand}(K) = \sum_{i \in S} d_i \tilde{y}_i = \hat{t}_{HT} + \Delta(K),$$

où

$$\Delta(K) = -\sum_{i \in S} \max\left(0,\, d_i y_i - K\right).$$

- Écriture alternative: $\hat{t}_{stand}(K) = \sum_{i \in S} \tilde{d}_i y_i$, où

$$\tilde{d}_i = d_i \frac{min\left(y_i,\, \frac{K}{d_i}\right)}{y_i}$$

Peut être plus petit que 1!

.

Cas particuliers de $\hat{t}_{HT}^{R}(K)$: estimateur de Dalén-Tambay

- Dalén (1987) et Tambay (1988) ont étudié une winsorisation alternative:

$$\tilde{y}_i = \begin{cases} y_i \text{ si } d_i y_i \leq K \\ \frac{K}{d_i} + \frac{1}{d_i}(y_i - \frac{K}{d_i}) \text{ si } d_i y_i > K \end{cases}$$

- L'estimateur winsorisé de Dalén-Tambay est donné par

$$\hat{t}_{DT}(K) = \sum_{i \in S} d_i \tilde{y}_i = \hat{t}_{HT} + \Delta(K),$$

où

$$\Delta(K) = -\sum_{i \in S} \frac{(d_i - 1)}{d_i} \max(0,\, d_i y_i - K)\,.$$

- Écriture alternative: $\hat{t}_{DT}(K) = \sum_{i \in S} \tilde{d}_i y_i$, où

$$\tilde{d}_i = 1 + (d_i - 1)\frac{min\left(y_i,\, \frac{K}{d_i}\right)}{y_i} \geq 1$$

Cas particuliers de $\hat{t}_{HT}^{R}(K)$: estimateur BHR

- Beaumont et al. (2013) ont proposé l'estimateur robuste suivant:

$$\hat{t}_{BHR}(K) = \hat{t}_{HT} - \sum_{i \in S} \hat{B}_{1i}^{HT} + \sum_{i \in S} \psi_K \left(\hat{B}_{1i}^{HT} \right) = \hat{t}_{HT} + \Delta(K),$$

où

$$\Delta(K) = - \sum_{i \in S} \left\{ \hat{B}_{1i}^{HT} - \psi_K \left(\hat{B}_{1i}^{HT} \right) \right\}$$

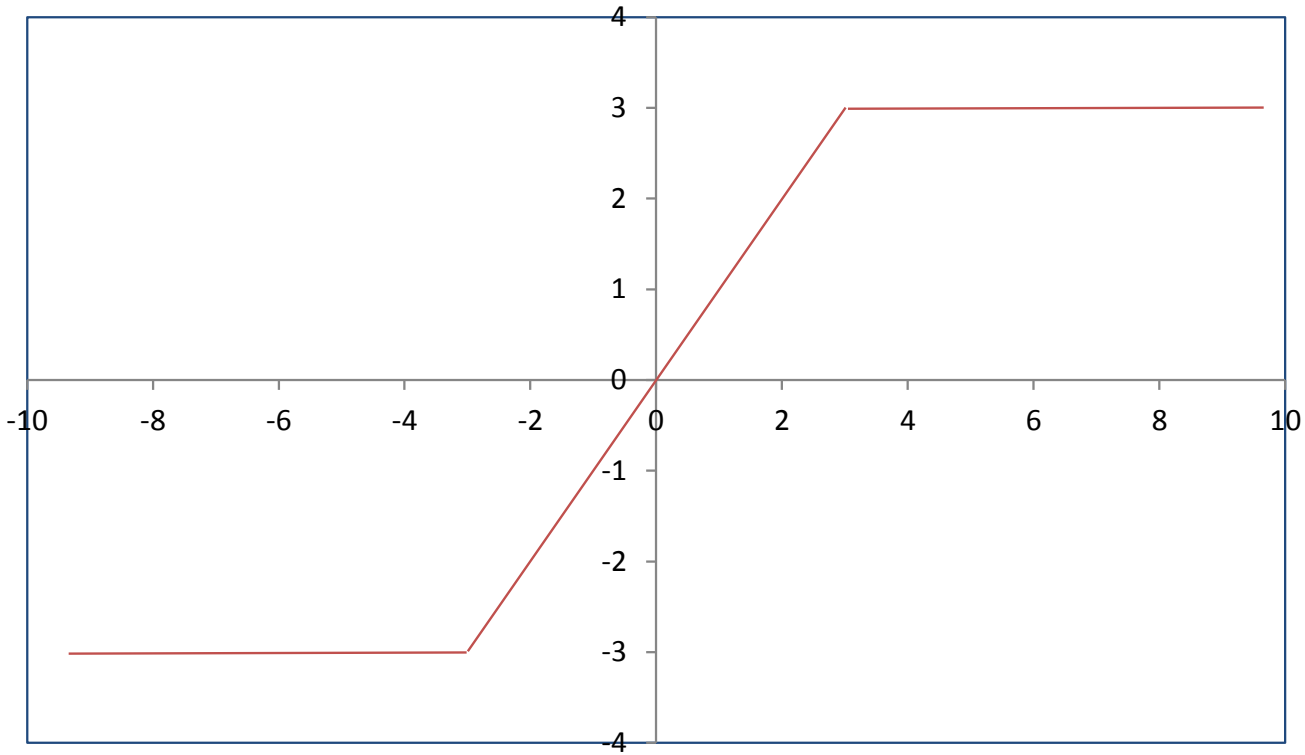
et $\psi_K(\cdot)$ est la fonction de Huber donnée par

$$\psi_K(t) = \begin{cases} K & \text{if } t > K \\ t & \text{if } |t| \leq K \\ -K & \text{if } t < -K \end{cases}$$

- Rôle de la fonction ψ: réduire l'influence des unités qui ont une grande influence.

Fonction de Huber $\psi_3(t)$ (avec $K = 3$)

- Remarque: $0 \leq \frac{\psi_K(t)}{t} \leq 1$.



Cas particuliers de $\hat{t}_{HT}^{R}(K)$: estimateur BHR

- L'estimateur BHR peut s'écrire comme

$$\hat{t}_{BHR}(K) = \sum_{i \in S} d_i \tilde{y}_i,$$

où

$$\tilde{y}_i = y_i - \frac{\alpha_i}{d_i} \hat{B}_{1i}^{HT}, \qquad \alpha_i = 1 - \psi_K(\hat{B}_{1i}^{HT}) / \hat{B}_{1i}^{HT}.$$

- Écriture alternative:

$$\hat{t}_{BHR}(K) = \sum_{i \in S} \tilde{d}_i y_i,$$

où

$$\tilde{d}_i = d_i - \frac{\alpha_i}{y_i} \hat{B}_{1i}^{HT}.$$

Récapitulons...

- Trois techniques pour traiter les valeurs influentes à l'étape de l'estimation $\rightarrow \hat{t}_{stand}(K), \hat{t}_{DT}(K)$ and $\hat{t}_{BHR}(K)$
- Appartiennent tous à la classe: $\hat{t}_R(K) = \hat{t}_{HT} + \Delta(K)$.
- Si $\Delta(K)$ est déterminé de manière à $\max_{i \in S} \left\{ |\hat{B}_{1i}^R| \right\}$, on obtient

$$\hat{t}_{stand}(K_{opt}) = \hat{t}_{DT}(K_{opt}) = \hat{t}_{BHR}(K_{opt}) = \hat{t}_{HT} - \frac{1}{2}(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}).$$

- La valeur K_{opt} varie d'une méthode à l'autre
- La valeur K_{opt} est obtenue en résolvant:

$$\Delta(K) = -\frac{1}{2}(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT})$$

- Il existe toujours une solution à l'équation précédente; voir Beaumont et al. (2013) et Favre Martinoz et al. (2013) mais la solution n'est pas toujours unique!

Choix de K

Unité	K_1 $\hat{t}_{HT}^R(K_1)$	K_2 $\hat{t}_{HT}^R(K_2)$	... ...	K_q $\hat{t}_{HT}^R(K_q)$
1	$\hat{B}_{11}^R(K_1)$	$\hat{B}_{11}^R(K_2)$	...	$\hat{B}_{11}^R(K_q)$
2	$\hat{B}_{12}^R(K_1)$	$\hat{B}_{12}^R(K_2)$	...	$\hat{B}_{12}^R(K_q)$
.	.	.	...	.
.	.	.	...	.
n	$\hat{B}_{1n}^R(K_1)$	$\hat{B}_{1n}^R(K_2)$	...	$\hat{B}_{1n}^R(K_2)$
	$\max\left\{\hat{B}_{1i}^R(K_1)\right\}$	$\max\left\{\hat{B}_{1i}^R(K_2)\right\}$	...	$\max\left\{(\hat{B}_{1i}^R(K_2)\right\}$

- Winsorisation standard: $\hat{t}_{stand}(K) = \sum_{i\in S} d_i \tilde{y}_i$, où

$$\tilde{y}_i = \begin{cases} y_i \text{ if } d_i y_i \leq K \\ \frac{K}{d_i} \text{ if } d_i y_i > K \end{cases}$$

- EASSR: $\hat{B}_{1i}^R = \frac{n}{n-1}\left(\frac{N}{n} - 1\right)(y_i - \bar{y}) - \sum_{i\in S} \max\left(0,\, d_i y_i - K\right).$

Étude par simulation

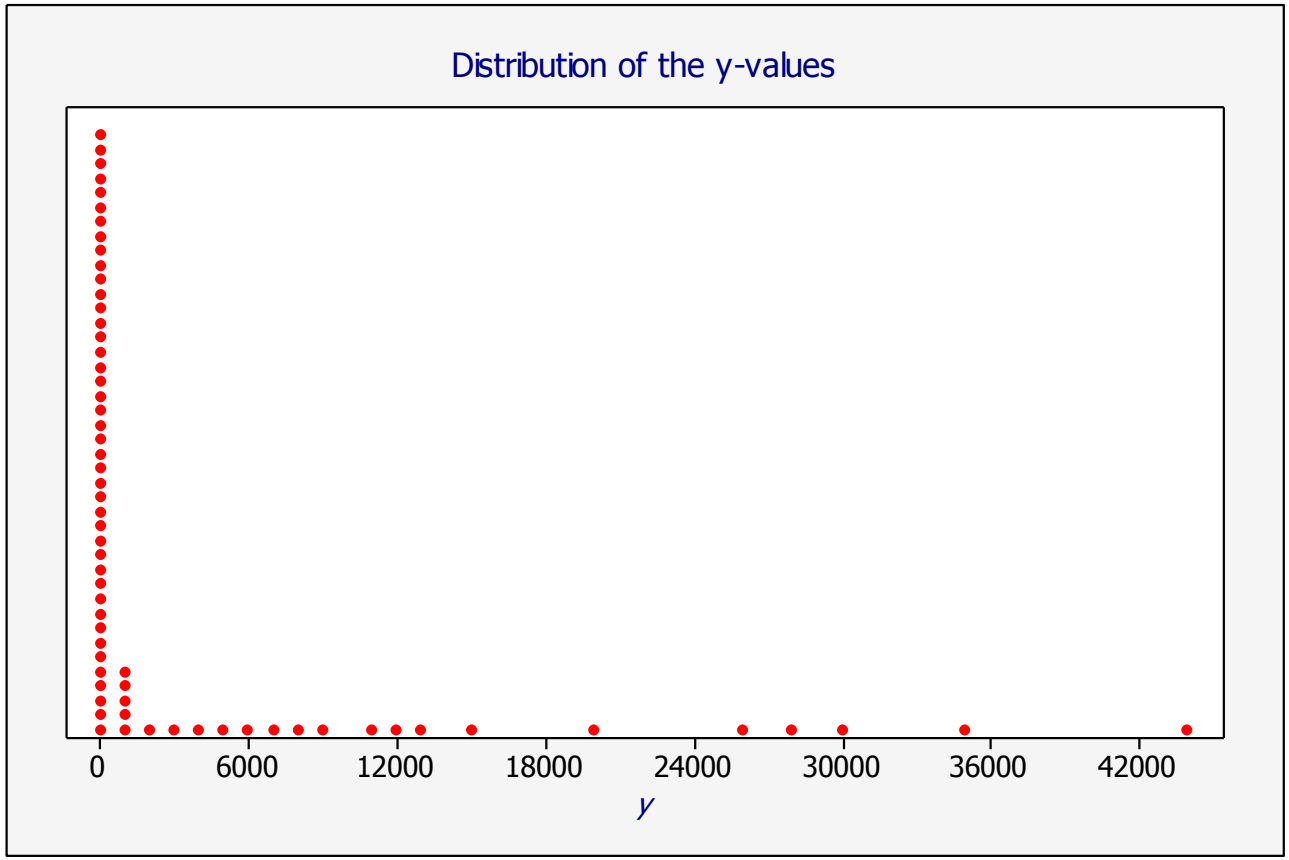
- On a généré une population de taille $N = 5,000$ avec une variable y
  - Population: mélange de deux lois log-normales

$$y_0 \sim LN(5,1) \quad \text{et} \quad y_1 \sim LN(8,2)$$

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.01$$

- Dans chaque population, on a tiré $R = 25,000$ échantillons selon un plan aléatoire simple sans remise, de taille $n = 50; 100; 500$

Distribution de la variable y



Étude par simulation

On a calculé trois estimateurs:

- L'estimateur d'Horvitz-Thompson: $\hat{t}_{HT} = \sum_{i \in S} d_i y_i$
- L'estimateur robuste proposé :

$$\hat{t}_R(K_{opt}) = \hat{t}_{HT} - \frac{1}{2}(\hat{B}^{HT}_{min} + \hat{B}^{HT}_{max})$$

  - avec le biais conditionnel estimé de manière non robuste:

$$\hat{B}^{HT}_{1i} = \frac{N}{(N-1)}\left(\frac{N}{n} - 1\right)(y_i - \bar{y}).$$

  - avec le biais conditionnel estimé de manière robuste:

$$\hat{B}^{HT}_{1i} = \frac{N}{(N-1)}\left(\frac{N}{n} - 1\right)(y_i - m).$$

Étude par simulation

Mesures Monte Carlo:

- Biais relatif Monte Carlo (en %):

$$RB_{MC}\left(\hat{\theta}\right) = 100 \times \left[E_{MC}\left(\hat{\theta}\right) - t_y\right] / t_y.$$

- Efficacité relative Monte Carlo (en %) par rapport à l'estimateur d'Horvitz-Thompson:

$$RE = 100 \times EQM_{MC}\left(\hat{\theta}\right) / EQM_{MC}\left(\hat{t}_{HT}\right).$$

Étude par simulation

n	$\hat{t}_{HT}$	$\hat{t}_R(K_{opt})$ (non robuste)	$\hat{t}_R(K_{opt})$ (robuste)
50	0.11	-12.22	-13.01
	(100)	(44.0)	(43.2)
100	0.16	-10.27	-10.70
	(100)	(50.7)	(50.5)
500	0.01	-5.50	-5.60
	(100)	(79.2)	(79.3)

- $\hat{t}_{HT} = \sum_{i \in S} d_i y_i$
- $\hat{t}_R(K_{opt}) = \hat{t}_{HT} - \frac{1}{2}(\hat{B}^{HT}_{min} + \hat{B}^{HT}_{max})$

Estimation pour des domaines

- En pratique, des estimations pour des sous-populations (domaines) sont toujours requises.
- La population est divisée en D domaines disjoints, $U_1, \ldots, U_d, \ldots, U_D$ de taille $N_1, \ldots, N_d, \ldots, N_D$, respectivement. On a

$$\bigcup_{d=1}^{D} U_d = U \text{ et } \sum_{d=1}^{D} N_d = N.$$

- Soit $S_d = S \bigcap U_d$.
- On cherche à estimer les totaux sur les domaines, $t_1, \ldots, t_d, \ldots, t_D$, où $t_d = \sum_{i \in U_d} y_i, d = 1, \ldots, D$. On a la relation:

$$t_y = \sum_{i \in U} y_i = \sum_{d=1}^{D} t_d$$

Estimation pour des domaines

- On peut estimer t_d par un estimateur de type Horvitz-Thompson:

$$\hat{t}_d = \sum_{i \in S_d} d_i y_i, \quad d = 1, \ldots, D.$$

- On a la relation:

$$\hat{t}_{HT} = \sum_{d=1}^{D} \hat{t}_d.$$

- Le système des estimations $\left\{\hat{t}_{HT}, \hat{t}_1, \ldots, \hat{t}_d, \ldots, \hat{t}_D\right\}$ est dit cohérent

Estimation pour des domaines

- En présence de valeurs influentes, on peut utiliser dans chaque domaine n'importe quelle méthode de traitement des valeurs influentes afin de réduire leur impact. On obtient alors D estimateurs robustes

$$\hat{t}_{R,1}, \ldots, \hat{t}_{R,d}, \ldots, \hat{t}_{R,D}.$$

- On définit l'estimateur agrégé, $\hat{t}_{agr}$, selon

$$\hat{t}_{agr} = \sum_{d=1}^{D} \hat{t}_{R,d}.$$

- Le système des estimations $\left\{\hat{t}_{agr}, \hat{t}_{R,1}, \ldots, \hat{t}_{R,d}, \ldots, \hat{t}_{R,D}\right\}$ est cohérent
- On s'attend à ce que l'estimateur agrégé $\hat{t}_{agr}$ soit biaisé de manière importante, car il s'exprime comme la somme d'estimateurs (négativement) biaisés; voir Hidiroglou and Rivest (2004).

Estimation pour des domaines

- Afin d'éviter un biais trop important au niveau global, on peut, dans un premier temps, obtenir D estimateurs robustes

$$\hat{t}_{R,1}, \ldots, \hat{t}_{R,d}, \ldots, \hat{t}_{R,D}.$$

- Deuxième temps: indépendamment, on obtient un estimateur robuste au niveau global: $\hat{t}_{R,0}$.
- Problème: le système des estimations $\left\{\hat{t}_{R,0}, \hat{t}_{R,1}, \ldots, \hat{t}_{R,d}, \ldots, \hat{t}_{R,D}\right\}$ n'est pas cohérent, en général. On a

$$\hat{t}_{R,0} \neq \sum_{d=1}^{D} \hat{t}_{R,d},$$

en général.

Estimation pour des domaines

- Solution (Favre Martinoz et al., 2013): déterminer des estimations robustes finales

$$\hat{t}^*_{R,1}, \ldots, \hat{t}^*_{R,d}, \ldots, \hat{t}^*_{R,D}$$

aussi proche que possible des estimations robustes initiales

$$\hat{t}_{R,1}, \ldots, \hat{t}_{R,d}, \ldots, \hat{t}_{R,D}$$

de manière à satisfaire l'équation de calage

$$\sum_{d=1}^{D} \hat{t}^*_{R,d} = \hat{t}_{R,0}. \tag{1}$$

Estimation pour des domaines

- Si on utilise la distance du chi-deux généralisé (méthode linéaire dans CALMAR) on cherche des estimations robustes finales $\hat{t}^*_{R,d}$, telles que

$$\sum_{d=1}^{D} \frac{1}{q_d} \frac{\left\{\hat{t}^*_{R,d} - \hat{t}_{R,d}\right\}^2}{\hat{t}_{R,d}}$$

  est minimisée sous la contrainte de calage (1).
- q_d: coefficient associé à l'estimation $\hat{t}^*_{R,d}$
- Solution:

$$\hat{t}^*_{R,d} = \hat{t}_{R,d} + q_d \left\{ \frac{\left(\hat{t}_{R,0} - \sum_{g=1}^{D} \hat{t}_{R,g}\right)}{\sum_{g=1}^{D} q_g} \right\}$$

Étude par simulation

- On a généré une population de taille $N = 5000$ comprenant 5 strates
- Dans chaque strate, on a généré une variable d'intérêt y selon une loi log-normale de paramètres log(2000) et 1.5.
- De la population, on a tiré 5000 échantillons selon un plan stratifié aléatoire simple sans remise.
- Dans la strate h, un échantillon aléatoire simple sans remise, S_h, de taille n_h a été tiré.
- Dans cet exemple, domaine = strate

Strate	1	2	3	4	5
N_h	2000	1500	1000	400	100
f_h	0.01	0.05	0.1	0.2	0.8

Étude par simulation

		$\hat{t}_{agr}$	$\hat{t}_{R,0}$	$\hat{t}_{R,0}$
Estimateur global		$-8.2(94)$	$-3.9(79)$	$-3.9(79)$
		$\hat{t}_{R,h}$	$\hat{t}^*_{R,h}$	
			$q_h = 1\ \forall h$	$q_h = n_h^{-1}(1-f_h)$
Strate	1	$-18.9(51)$	$-15.0(52)$	$-4.7(56)$
	2	$-8.8(72)$	$-4.5(72)$	$-4.8(72)$
	3	$-5.8(95)$	$-3.3(96)$	$-3.0(96)$
	4	$-8.2(74)$	$-3.7(77)$	$-3.3(79)$
	5	$-1.6(97)$	$2.9(116)$	$-0.8(97)$

- $\hat{t}_{R,h} = \hat{t}_{HT,h} - \frac{1}{2}(\hat{B}^{HT}_{h,min} + \hat{B}^{HT}_{h,max})$
- $\hat{t}_{HT,h} = \frac{N_h}{n_h}\sum_{i\in S_h} y_i.$

Remarques finales

- Minimiser le plus grand biais conditionnel estimé est une bonne alternative à minimiser l'erreur quadratique moyenne estimée, tout en étant beaucoup plus simple à implémenter en pratique
- S'il n'y a pas d'unités influentes, l'estimateur robuste proposé est à peine moins efficace que l'estimateur d'Horvitz-Thompson.
- Estimation de l'erreur quadratique moyenne de l'estimateur robuste : Bootstrap
- Extensions: estimation pour petits domaines, échantillonnage à deux phases, non-réponse totale (ajustement des poids), non-réponse partielle (imputation)