Triple robustesse en présence de données imputée dans les enquêtes

David Haziza

Département de mathématiques et de statistique
Université de Montréal

En collaboration avec Valéry Dongmo Jiongo et Pierre Duchesne
Statistique Canada et Université de Montréal

Les Journées de Méthodologie Statistique
Paris, France

25 janvier 2012

Plan de la présentation

- Introduction
- Un exemple
- Biais conditionnel
- Estimateur robuste proposé
- Étude par simulation
- Conclusions et perspectives

Unités influentes

- Observations atypiques avec des poids de sondage potentiellement élevés; par exemple, Beaumont et Rivest (2009)

Unités influentes

- Observations atypiques avec des poids de sondage potentiellement élevés; par exemple, Beaumont et Rivest (2009)
- Les unités influentes font partie de la population.
- Les statistiques d'enquête sont généralement sensibles à la présence d'unités influentes

Unités influentes

- Observations atypiques avec des poids de sondage potentiellement élevés; par exemple, Beaumont et Rivest (2009)
- Les unités influentes font partie de la population.
- Les statistiques d'enquête sont généralement sensibles à la présence d'unités influentes
- Inclure ou exclure une valeur aberrante dans les estimations peut avoir un impact considérable sur les résultats

Unités influentes

- Observations atypiques avec des poids de sondage potentiellement élevés; par exemple, Beaumont et Rivest (2009)
- Les unités influentes font partie de la population.
- Les statistiques d'enquête sont généralement sensibles à la présence d'unités influentes
- Inclure ou exclure une valeur aberrante dans les estimations peut avoir un impact considérable sur les résultats
- Les unités influentes sont fréquentes pour des variables telles que le *Revenu* car la distribution est fortement asymétrique (à droite)

Unités influentes

- Observations atypiques avec des poids de sondage potentiellement élevés; par exemple, Beaumont et Rivest (2009)
- Les unités influentes font partie de la population.
- Les statistiques d'enquête sont généralement sensibles à la présence d'unités influentes
- Inclure ou exclure une valeur aberrante dans les estimations peut avoir un impact considérable sur les résultats
- Les unités influentes sont fréquentes pour des variables telles que le *Revenu* car la distribution est fortement asymétrique (à droite)
- Réduire l'impact des unités influentes:
  - Sur l'erreur due à l'échantillonnage: à l'aide d'un bon plan de sondage; par exemple, l'échantillonnage stratifié avec une strate à tirage complet

Unités influentes

- Observations atypiques avec des poids de sondage potentiellement élevés; par exemple, Beaumont et Rivest (2009)
- Les unités influentes font partie de la population.
- Les statistiques d'enquête sont généralement sensibles à la présence d'unités influentes
- Inclure ou exclure une valeur aberrante dans les estimations peut avoir un impact considérable sur les résultats
- Les unités influentes sont fréquentes pour des variables telles que le *Revenu* car la distribution est fortement asymétrique (à droite)
- Réduire l'impact des unités influentes:
  - Sur l'erreur due à l'échantillonnage: à l'aide d'un bon plan de sondage; par exemple, l'échantillonnage stratifié avec une strate à tirage complet
  - Sur l'erreur de non-réponse: construire des classes d'imputation de sorte que dans une classe, les unités ont approximativement le même revenu; par exemple, Little (1986); Haziza et Beaumont (2007).

Unités influentes

- En présence d'unités influentes, l'estimateur imputé d'un total
  - est (approximativement) sans biais si le modèle d'imputation et/ou le modèle de non-réponse est bien spécifié

Unités influentes

- En présence d'unités influentes, l'estimateur imputé d'un total
  - est (approximativement) sans biais si le modèle d'imputation et/ou le modèle de non-réponse est bien spécifié
  - peut avoir une très grande variance

Unités influentes

- En présence d'unités influentes, l'estimateur imputé d'un total
  - est (approximativement) sans biais si le modèle d'imputation et/ou le modèle de non-réponse est bien spécifié
  - peut avoir une très grande variance
- Réduire l'influence des unités ayant une grande influence conduit à des estimateurs plus stables mais biaisés

Unités influentes

- En présence d'unités influentes, l'estimateur imputé d'un total
  - est (approximativement) sans biais si le modèle d'imputation et/ou le modèle de non-réponse est bien spécifié
  - peut avoir une très grande variance
- Réduire l'influence des unités ayant une grande influence conduit à des estimateurs plus stables mais biaisés
- Traitement des unités influentes: compromis entre biais et variance

Unités influentes

- En présence d'unités influentes, l'estimateur imputé d'un total
  - est (approximativement) sans biais si le modèle d'imputation et/ou le modèle de non-réponse est bien spécifié
  - peut avoir une très grande variance
- Réduire l'influence des unités ayant une grande influence conduit à des estimateurs plus stables mais biaisés
- Traitement des unités influentes: compromis entre biais et variance
- Nous cherchons une approche systématique pour construire des estimateurs robustes à la présence d'unités influentes: utilisation du biais conditionnel

Cadre de travail

- U: population finie de taille N

Cadre de travail

- U: population finie de taille N
- But: estimer le total dans la population U de la variable d'intérêt y,

$$Y = \sum_{i \in U} y_i$$

Cadre de travail

- U: population finie de taille N
- But: estimer le total dans la population U de la variable d'intérêt y,

$$Y = \sum_{i \in U} y_i$$

- s: échantillon de taille n tiré selon le plan de sondage $p(s)$

Cadre de travail

- U: population finie de taille N
- But: estimer le total dans la population U de la variable d'intérêt y,

$$Y = \sum_{i \in U} y_i$$

- s: échantillon de taille n tiré selon le plan de sondage $p(s)$
- Absence de non-réponse: estimateur par dilatation

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i$$

  - $d_i = 1/\pi_i$: poids de sondage de l'unité i
  - π_i: probabilité d'inclusion dans l'échantillon pour l'unité i

Cadre de travail

- Non-réponse à la variable y: certaines valeurs de y sont manquantes

Cadre de travail

- Non-réponse à la variable y: certaines valeurs de y sont manquantes
- Estimateur imputé:

$$\hat{Y}_I = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y_i^*$$

  - $r_i = 1$ si l'unité i répond à la variable y et $r_i = 0$, sinon.
  - y_i^*: valeur imputée utilisée pour remplacer la valeur manquante y_i

Cadre de travail

- Non-réponse à la variable y: certaines valeurs de y sont manquantes
- Estimateur imputé:

$$\hat{Y}_I = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y_i^*$$

  - $r_i = 1$ si l'unité i répond à la variable y et $r_i = 0$, sinon.
  - y_i^*: valeur imputée utilisée pour remplacer la valeur manquante y_i
- Comment doit-on imputer en présence d'unités influentes?

Un exemple: étude par simulation

- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y

Un exemple: étude par simulation

- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y
  - Population: mélange de deux distributions log-normales (moyenne et écart type dans l'échelle log)

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

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.05; 0.1$$

Un exemple: étude par simulation

- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y
  - Population: mélange de deux distributions log-normales (moyenne et écart type dans l'échelle log)

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

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.05; 0.1$$

- Dans chaque population, nous avons tiré un EASSR de taille $n = 100$

Un exemple: étude par simulation

- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y
  - Population: mélange de deux distributions log-normales (moyenne et écart type dans l'échelle log)

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

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.05; 0.1$$

- Dans chaque population, nous avons tiré un EASSR de taille $n = 100$
- Dans chaque échantillon, nous avons généré la non-réponse selon un mécanisme de non-réponse uniforme avec une probabilité de 70%

Un exemple: étude par simulation

- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y
  - Population: mélange de deux distributions log-normales (moyenne et écart type dans l'échelle log)

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

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.05; 0.1$$

- Dans chaque population, nous avons tiré un EASSR de taille $n = 100$
- Dans chaque échantillon, nous avons généré la non-réponse selon un mécanisme de non-réponse uniforme avec une probabilité de 70%
- Méthodes d'imputation:
  - l'imputation par la moyenne: chaque valeur manquante est remplacée par la moyenne des répondants

Un exemple: étude par simulation

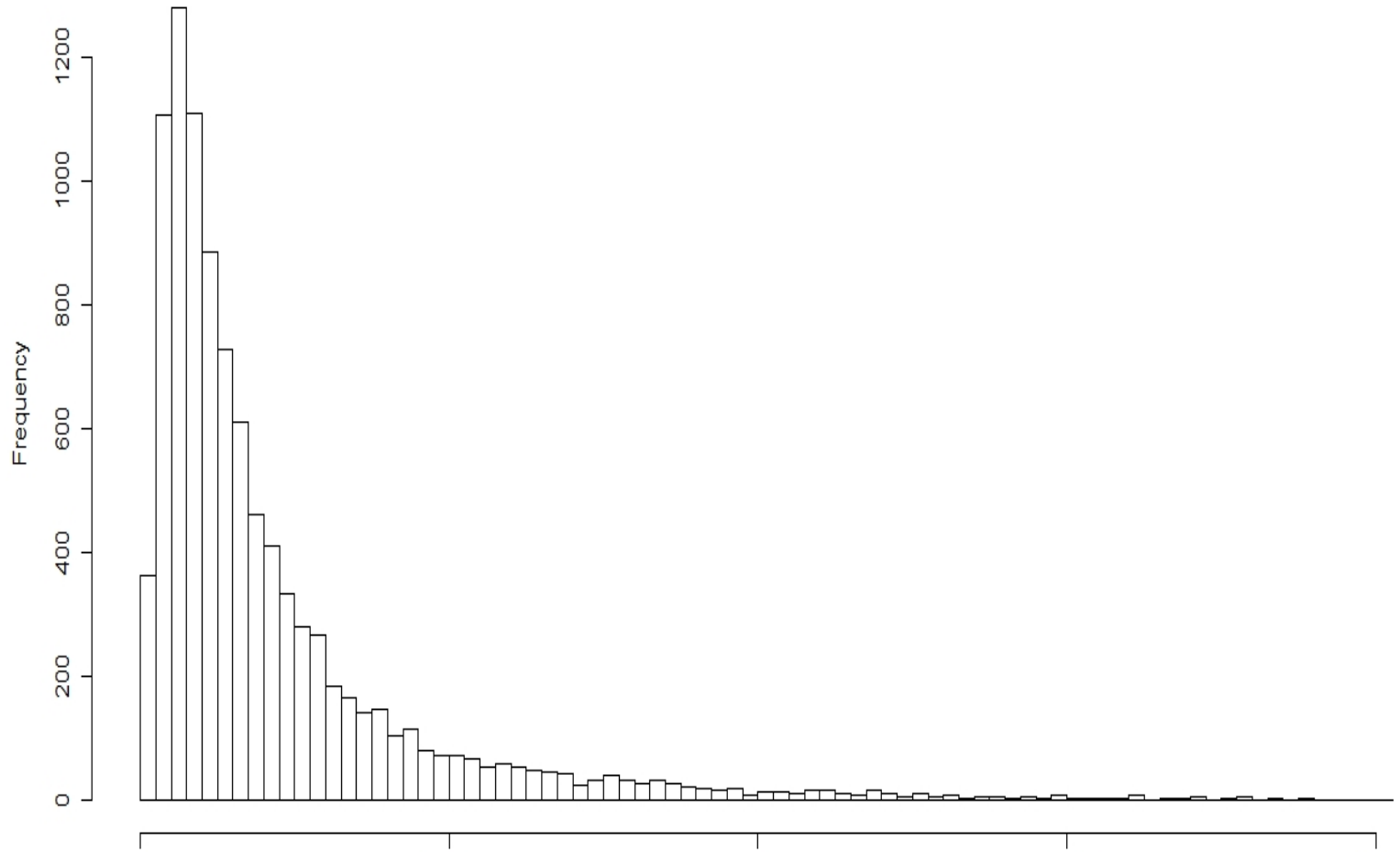
- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y
  - Population: mélange de deux distributions log-normales (moyenne et écart type dans l'échelle log)

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

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.05; 0.1$$

- Dans chaque population, nous avons tiré un EASSR de taille $n = 100$
- Dans chaque échantillon, nous avons généré la non-réponse selon un mécanisme de non-réponse uniforme avec une probabilité de 70%
- Méthodes d'imputation:
  - l'imputation par la moyenne: chaque valeur manquante est remplacée par la moyenne des répondants
  - l'imputation par la médiane: chaque valeur manquante est remplacée par la médiane des répondants

Distribution de y: distribution contaminée avec un taux de 10%



Résultats

- Biais relatif Monte carlo:

$$BR(\hat{Y}_I) = \frac{E_{MC}\left(\hat{Y}_I - Y\right)}{Y} \times 100$$

- Efficacité relative:

$$ER = \frac{MSE_{MC}(\hat{Y}_I^{(\text{médiane})})}{MSE_{MC}(\hat{Y}_I^{(\text{moyenne})})} \times 100$$

	Contamination à 10%		Contamination à 5%	
	BR	ER	BR	ER
imputation par la moyenne	0.3	100.0	0.3	100.0
imputation par la médiane	-14.0	88.6	-11.9	94.4

Imputation par la régression déterministe

- Imputation par la régression déterministe: motivée par le modèle d'imputation

$$y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i,$$
$$E_m(\epsilon_i) = 0, Cov_m(\epsilon_i, \epsilon_j) = 0 \quad \text{si } i \neq j, V_m(\epsilon_i) = \sigma^2 c_i$$

Imputation par la régression déterministe

- Imputation par la régression déterministe: motivée par le modèle d'imputation

$$y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i,$$
$$E_m(\epsilon_i) = 0, Cov_m(\epsilon_i, \epsilon_j) = 0 \quad \text{si } i \neq j, V_m(\epsilon_i) = \sigma^2 c_i$$

- Valeur imputée:

$$y_i^* = \mathbf{z}_i^T \hat{\mathbf{B}}_r,$$

où

$$\hat{\mathbf{B}}_r = \left(\sum_{i \in s} \omega_i r_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \sum_{i \in s} \omega_i r_i c_i^{-1} \mathbf{z}_i y_i$$

Imputation par la régression déterministe

- Imputation par la régression déterministe: motivée par le modèle d'imputation

$$y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i,$$
$$E_m(\epsilon_i) = 0, Cov_m(\epsilon_i, \epsilon_j) = 0 \quad \text{si } i \neq j, V_m(\epsilon_i) = \sigma^2 c_i$$

- Valeur imputée:

$$y_i^* = \mathbf{z}_i^T \hat{\mathbf{B}}_r,$$

où

$$\hat{\mathbf{B}}_r = \left(\sum_{i \in s} \omega_i r_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \sum_{i \in s} \omega_i r_i c_i^{-1} \mathbf{z}_i y_i$$

  - $\omega_i = 1$: imputation par la régression déterministe non-pondérée

Imputation par la régression déterministe

- Imputation par la régression déterministe: motivée par le modèle d'imputation

$$y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i,$$
$$E_m(\epsilon_i) = 0, Cov_m(\epsilon_i, \epsilon_j) = 0 \quad \text{si } i \neq j, V_m(\epsilon_i) = \sigma^2 c_i$$

- Valeur imputée:

$$y_i^* = \mathbf{z}_i^T \hat{\mathbf{B}}_r,$$

où

$$\hat{\mathbf{B}}_r = \left(\sum_{i \in s} \omega_i r_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \sum_{i \in s} \omega_i r_i c_i^{-1} \mathbf{z}_i y_i$$

  - $\omega_i = 1$: imputation par la régression déterministe non-pondérée
  - $\omega_i = d_i$: imputation par la régression déterministe pondérée (par les poids de sondage)

Imputation par la régression doublement robuste

- $\omega_i = d_i(1 - \hat{p}_i)/\hat{p}_i$: Imputation par la régression déterministe doublement robuste

Imputation par la régression doublement robuste

- $\omega_i = d_i(1 - \hat{p}_i)/\hat{p}_i$: Imputation par la régression déterministe doublement robuste
- $\hat{Y}_I$ est approximativement sans biais et convergent pour Y si
  - le modèle d'imputation $y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i$ est bien spécifié (approche par modèle d'imputation)

    ou
  - le modèle de non-réponse utilisé pour obtenir $\hat{p}_i$ est bien spécifié (approche par modèle de non-réponse)
  - par exemple, Haziza and Rao (2006) et Kim and Haziza (2010)

Imputation par la régression doublement robuste

- $\omega_i = d_i(1 - \hat{p}_i)/\hat{p}_i$: Imputation par la régression déterministe doublement robuste
- $\hat{Y}_I$ est approximativement sans biais et convergent pour Y si
  - le modèle d'imputation $y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i$ est bien spécifié (approche par modèle d'imputation)

    ou
  - le modèle de non-réponse utilisé pour obtenir $\hat{p}_i$ est bien spécifié (approche par modèle de non-réponse)
  - par exemple, Haziza and Rao (2006) et Kim and Haziza (2010)
- Les procédures doublement robustes offrent une protection contre une mauvaise spécification de l'un modèle ou l'autre

Imputation par la régression doublement robuste

- $\omega_i = d_i(1 - \hat{p}_i)/\hat{p}_i$: Imputation par la régression déterministe doublement robuste
- $\hat{Y}_I$ est approximativement sans biais et convergent pour Y si
  - le modèle d'imputation $y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i$ est bien spécifié (approche par modèle d'imputation)

    ou
  - le modèle de non-réponse utilisé pour obtenir $\hat{p}_i$ est bien spécifié (approche par modèle de non-réponse)
  - par exemple, Haziza and Rao (2006) et Kim and Haziza (2010)
- Les procédures doublement robustes offrent une protection contre une mauvaise spécification de l'un modèle ou l'autre
- Cependant, $\hat{Y}_I$ n'est pas robuste à la présence de valeurs aberrantes

Imputation par la régression doublement robuste

- $\omega_i = d_i(1-\hat{p}_i)/\hat{p}_i$: Imputation par la régression déterministe doublement robuste
- $\hat{Y}_I$ est approximativement sans biais et convergent pour Y si
  - le modèle d'imputation $y_i = \mathbf{z}_i^T\beta + \epsilon_i$ est bien spécifié (approche par modèle d'imputation)

    ou
  - le modèle de non-réponse utilisé pour obtenir $\hat{p}_i$ est bien spécifié (approche par modèle de non-réponse)
  - par exemple, Haziza and Rao (2006) et Kim and Haziza (2010)
- Les procédures doublement robustes offrent une protection contre une mauvaise spécification de l'un modèle ou l'autre
- Cependant, $\hat{Y}_I$ n'est pas robuste à la présence de valeurs aberrantes
- Estimateur robuste naïf:

$$\hat{Y}_I^{\text{naïf}} = \sum_{i\in s} d_i r_i y_i + \sum_{i\in s} d_i(1-r_i)\mathbf{z}_i^T\hat{\mathbf{B}}_r^{\text{rob}}$$

Influence d'une unité influente

- L'erreur totale peut se décomposer comme

$$\hat{Y}_I - Y = \underbrace{\left(\hat{Y}_\pi - Y\right)}_{\substack{\text{erreur} \\ \text{due à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_\pi\right)}_{\substack{\text{erreur} \\ \text{de non-réponse}}}$$

Influence d'une unité influente

- L'erreur totale peut se décomposer comme

$$\hat{Y}_I - Y = \underbrace{\left(\hat{Y}_\pi - Y\right)}_{\substack{\text{erreur} \\ \text{due à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_\pi\right)}_{\substack{\text{erreur} \\ \text{de non-réponse}}}$$

- Une unité influente peut avoir un impact sur l'erreur due à l'échantillonnage et sur l'erreur de non-réponse

Influence d'une unité influente

- L'erreur totale peut se décomposer comme

$$\hat{Y}_I - Y = \underbrace{\left(\hat{Y}_\pi - Y\right)}_{\substack{\text{erreur} \\ \text{due à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_\pi\right)}_{\substack{\text{erreur} \\ \text{de non-réponse}}}$$

- Une unité influente peut avoir un impact sur l'erreur due à l'échantillonnage et sur l'erreur de non-réponse
- Comment mesurer l'influence (ou l'impact) d'une unité?

Influence d'une unité influente

- L'erreur totale peut se décomposer comme

$$\hat{Y}_I - Y = \underbrace{\left(\hat{Y}_\pi - Y\right)}_{\substack{\text{erreur} \\ \text{due à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_\pi\right)}_{\substack{\text{erreur} \\ \text{de non-réponse}}}$$

- Une unité influente peut avoir un impact sur l'erreur due à l'échantillonnage et sur l'erreur de non-réponse
- Comment mesurer l'influence (ou l'impact) d'une unité?
  Absence de non-réponse: le biais conditionnel; Moreno-Rebollo, Munoz-Reyez et Munoz-Pichardo (1999), Beaumont, Haziza et Ruiz-Gazen (2009).

Influence d'une unité influente

- L'erreur totale peut se décomposer comme

$$\hat{Y}_I - Y = \underbrace{\left(\hat{Y}_\pi - Y\right)}_{\substack{\text{erreur} \\ \text{due à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_\pi\right)}_{\substack{\text{erreur} \\ \text{de non-réponse}}}$$

- Une unité influente peut avoir un impact sur l'erreur due à l'échantillonnage et sur l'erreur de non-réponse
- Comment mesurer l'influence (ou l'impact) d'une unité?
  Absence de non-réponse: le biais conditionnel; Moreno-Rebollo, Munoz-Reyez et Munoz-Pichardo (1999), Beaumont, Haziza et Ruiz-Gazen (2009).
- Comment construire un estimateur robuste en présence d' unités influentes?

Influence d'une unité influente

- L'erreur totale peut se décomposer comme

$$\hat{Y}_I - Y = \underbrace{\left(\hat{Y}_\pi - Y\right)}_{\substack{\text{erreur} \\ \text{due à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_\pi\right)}_{\substack{\text{erreur} \\ \text{de non-réponse}}}$$

- Une unité influente peut avoir un impact sur l'erreur due à l'échantillonnage et sur l'erreur de non-réponse
- Comment mesurer l'influence (ou l'impact) d'une unité?
  Absence de non-réponse: le biais conditionnel; Moreno-Rebollo, Munoz-Reyez et Munoz-Pichardo (1999), Beaumont, Haziza et Ruiz-Gazen (2009).
- Comment construire un estimateur robuste en présence d' unités influentes?
  Absence de non-réponse: en utilisant le concept de biais conditionnel; Beaumont, Haziza et Ruiz-Gazen (2011).

Biais conditionnel: erreur due à l'échantillonnage

- Influence d'une unité échantillonnée sur l'erreur due à l'échantillonnage:

$$B_i^{\pi}(I_i = 1) = E_p(\hat{Y}_{\pi} - Y | I_i = 1) = (d_i - 1)y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j$$

Biais conditionnel: erreur due à l'échantillonnage

- Influence d'une unité échantillonnée sur l'erreur due à l'échantillonnage:

$$B_i^{\pi}(I_i = 1) = E_p(\hat{Y}_{\pi} - Y | I_i = 1) = (d_i - 1)y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j$$

- Exemple: échantillonnage aléatoire simple sans remise

$$B_i^{\pi}(I_i = 1) = \frac{N}{(N-1)} (\frac{N}{n} - 1)(y_i - \bar{Y})$$

Biais conditionnel: erreur due à l'échantillonnage

- Influence d'une unité échantillonnée sur l'erreur due à l'échantillonnage:

$$B_i^{\pi}(I_i = 1) = E_p(\hat{Y}_{\pi} - Y | I_i = 1) = (d_i - 1)y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j$$

- Exemple: échantillonnage aléatoire simple sans remise

$$B_i^{\pi}(I_i = 1) = \frac{N}{(N-1)} (\frac{N}{n} - 1)(y_i - \bar{Y})$$

- Influence d'une unité non-échantillonnée sur l'erreur due à l'échantillonnage:

$$B_i^{\pi}(I_i = 0) = E_p(\hat{Y}_{\pi} - Y | I_i = 0) = -(d_i - 1) B_i^{\pi}(I_i = 1)$$

Biais conditionnel: erreur due à l'échantillonnage

- Influence d'une unité échantillonnée sur l'erreur due à l'échantillonnage:

$$B_i^{\pi}(I_i = 1) = E_p(\hat{Y}_{\pi} - Y | I_i = 1) = (d_i - 1)y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j$$

- Exemple: échantillonnage aléatoire simple sans remise

$$B_i^{\pi}(I_i = 1) = \frac{N}{(N-1)} (\frac{N}{n} - 1)(y_i - \bar{Y})$$

- Influence d'une unité non-échantillonnée sur l'erreur due à l'échantillonnage:

$$B_i^{\pi}(I_i = 0) = E_p(\hat{Y}_{\pi} - Y | I_i = 0) = -(d_i - 1) B_i^{\pi}(I_i = 1)$$

- $B_i^{\pi}(I_i = 1)$: inconnu $\Rightarrow$ doit être estimé

Biais conditionnel: erreur due à l'échantillonnage

- Erreur due à l'échantillonnage:

$$\hat{Y}_\pi - Y = \sum_{i \in s} B_i^\pi (I_i = 1) + \sum_{i \in U - s} B_i^\pi (I_i = 0)$$

Biais conditionnel: erreur due à l'échantillonnage

- Erreur due à l'échantillonnage:

$$\hat{Y}_\pi - Y = \sum_{i \in s} B_i^\pi (I_i = 1) + \sum_{i \in U - s} B_i^\pi (I_i = 0)$$

- B_i^π: contribution d'une unité à l'erreur due à l'échantillonnage

Biais conditionnel: erreur due à l'échantillonnage

- Erreur due à l'échantillonnage:

$$\hat{Y}_\pi - Y = \sum_{i \in s} B_i^\pi (I_i = 1) + \sum_{i \in U - s} B_i^\pi (I_i = 0)$$

- B_i^π: contribution d'une unité à l'erreur due à l'échantillonnage
- Absence de non-réponse (Beaumont, Haziza et Ruiz-Gazen, 2011): l'estimateur robuste est donné par

$$\hat{Y}_\pi^R = \hat{Y}_\pi - \sum_{i \in s} \hat{B}_{\pi i} (I_i = 1) + \sum_{i \in s} \psi(\hat{B}_{\pi i} (I_i = 1))$$

Biais conditionnel: erreur due à l'échantillonnage

- Erreur due à l'échantillonnage:

$$\hat{Y}_\pi - Y = \sum_{i \in s} B_i^\pi (I_i = 1) + \sum_{i \in U - s} B_i^\pi (I_i = 0)$$

- B_i^π: contribution d'une unité à l'erreur due à l'échantillonnage
- Absence de non-réponse (Beaumont, Haziza et Ruiz-Gazen, 2011): l'estimateur robuste est donné par

$$\hat{Y}_\pi^R = \hat{Y}_\pi - \sum_{i \in s} \hat{B}_{\pi i} (I_i = 1) + \sum_{i \in s} \psi(\hat{B}_{\pi i} (I_i = 1))$$

- Choix de $\psi(t)$: par exemple, la fonction de Huber est donnée par

$$\psi(t) = \begin{cases} c & \text{si } t > c \\ t & \text{si } |t| \leq c \\ -c & \text{si } t < -c \end{cases}$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- Influence d'une unité répondante sur l'erreur de non-réponse:

$$B_{mi} = E_m(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, \mathbf{r}, Y_i = y_i) = d_i(g_i - 1)\left(y_i - \mathbf{z}_i^T \beta\right),$$

où

$$g_i = 1 + \left(\sum_{j \in s} d_j(1 - r_j)\mathbf{z}_j^T\right)\left(\sum_{j \in s} \omega_j r_j c_j^{-1} \mathbf{z}_j \mathbf{z}_j^T\right)^{-1} \frac{\omega_i c_i^{-1} \mathbf{z}_i}{d_i}$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- Influence d'une unité répondante sur l'erreur de non-réponse:

$$B_{mi} = E_m(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, \mathbf{r}, Y_i = y_i) = d_i(g_i - 1)\left(y_i - \mathbf{z}_i^T \beta\right),$$

où

$$g_i = 1 + \left(\sum_{j \in s} d_j(1 - r_j)\mathbf{z}_j^T\right)\left(\sum_{j \in s} \omega_j r_j c_j^{-1} \mathbf{z}_j \mathbf{z}_j^T\right)^{-1} \frac{\omega_i c_i^{-1} \mathbf{z}_i}{d_i}$$

- Influence d'une unité non-répondante sur l'erreur de non-réponse:

$$B_{mi} = E_m(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, \mathbf{r}, Y_i = y_i) = -d_i\left(y_i - \mathbf{z}_i^T \beta\right)$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- L'influence d'une unité répondante i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- L'influence d'une unité répondante i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- L'influence d'une unité répondante i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - son résidu $y_i - \mathbf{z}_i^T \beta$ est grand

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- L'influence d'une unité répondante i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - son résidu $y_i - \mathbf{z}_i^T \beta$ est grand
- Erreur de non-réponse:

$$\hat{Y}_I - \hat{Y}_\pi = \sum_{i \in s} B_{mi}$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- L'influence d'une unité répondante i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - son résidu $y_i - \mathbf{z}_i^T \beta$ est grand
- Erreur de non-réponse:

$$\hat{Y}_I - \hat{Y}_\pi = \sum_{i \in s} B_{mi}$$

- B_{mi}: contribution de l'unité i à l'erreur de non-réponse

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle d'Imputation)

- L'influence d'une unité répondante i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - son résidu $y_i - \mathbf{z}_i^T \beta$ est grand
- Erreur de non-réponse:

$$\hat{Y}_I - \hat{Y}_\pi = \sum_{i \in s} B_{mi}$$

- B_{mi}: contribution de l'unité i à l'erreur de non-réponse
- B_{mi}: inconnu $\Rightarrow$ doit être estimé

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence d'une unité répondante sur l'erreur de non-réponse:

$$B_{qi}(r_i = 1) = E_q(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, r_i = 1) \approx d_i(1 - p_i)\tilde{g}_i(y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p),$$

où

$$\tilde{g}_i = 1 + \left(\sum_{j \in s} d_j(1 - p_j)\mathbf{z}_j^T \right) \left(\sum_{j \in s} \omega_j p_j c_j^{-1} \mathbf{z}_j \mathbf{z}_j^T \right)^{-1} \frac{\omega_i c_i^{-1} \mathbf{z}_i}{d_i}$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence d'une unité répondante sur l'erreur de non-réponse:

$$B_{qi}(r_i = 1) = E_q(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, r_i = 1) \approx d_i(1 - p_i)\tilde{g}_i(y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p),$$

où

$$\tilde{g}_i = 1 + \left(\sum_{j \in s} d_j(1 - p_j)\mathbf{z}_j^T \right) \left(\sum_{j \in s} \omega_j p_j c_j^{-1} \mathbf{z}_j \mathbf{z}_j^T \right)^{-1} \frac{\omega_i c_i^{-1} \mathbf{z}_i}{d_i}$$

- Influence d'une unité non-répondante sur l'erreur de non-réponse:

$$B_{qi}(r_i = 0) = E_q(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, r_i = 0) \approx -d_i p_i \tilde{g}_i(y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p)$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - sa probabilité de réponse p_i est faible

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - sa probabilité de réponse p_i est faible
  - son résidu $y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p$ est grand

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - sa probabilité de réponse p_i est faible
  - son résidu $y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p$ est grand
- Erreur de non-réponse:

$$\hat{Y}_I - \hat{Y}_\pi \approx \sum_{i \in s_r} B_{qi}(r_i = 1) + \sum_{i \in s - s_r} B_{qi}(r_i = 0)$$

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - sa probabilité de réponse p_i est faible
  - son résidu $y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p$ est grand
- Erreur de non-réponse:

$$\hat{Y}_I - \hat{Y}_\pi \approx \sum_{i \in s_r} B_{qi}(r_i = 1) + \sum_{i \in s - s_r} B_{qi}(r_i = 0)$$

- B_{qi}: contribution de l'unité i à l'erreur de non-réponse

Biais conditionnel: influence sur l'erreur de non-réponse (approche par Modèle de Non-réponse)

- Influence de l'unité i sur l'erreur de non-réponse est grande si
  - son poids de sondage d_i est grand
  - son facteur d'ajustement g_i est grand (pourrait indiquer que $\mathbf{z}_i$ est grand)
  - sa probabilité de réponse p_i est faible
  - son résidu $y_i - \mathbf{z}_i^T \hat{\mathbf{B}}_p$ est grand
- Erreur de non-réponse:

$$\hat{Y}_I - \hat{Y}_\pi \approx \sum_{i \in s_r} B_{qi}(r_i = 1) + \sum_{i \in s - s_r} B_{qi}(r_i = 0)$$

- B_{qi}: contribution de l'unité i à l'erreur de non-réponse
- B_{qi}: inconnu $\Rightarrow$ doit être estimé

Estimateur robuste sous l'approche par Modèle d'Imputation

- Similaire à l'estimateur robuste dans le cas de données complètes:

$$\begin{aligned}
\hat{Y}_{Im}^{R} &= \hat{Y}_I - \sum_{i \in s_r} (\hat{B}_{\pi i}(I_i = 1) + \hat{B}_{mi}) + \sum_{i \in s_r} \psi(\hat{B}_{\pi i}(I_i = 1) + \hat{B}_{mi}) \\
&= \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) \mathbf{z}_i^T \hat{\mathbf{B}}_{Rob} + \sum_{i \in s_r} \left[\psi(\hat{B}_{\pi i} + \hat{B}_{mi}) - \hat{B}_{\pi i} \right] \\
&= \hat{Y}_I^{\text{naïf}} + \text{terme de correction}
\end{aligned}$$

Estimateur robuste sous l'approche par Modèle d'Imputation

- Similaire à l'estimateur robuste dans le cas de données complètes:

$$\begin{aligned}\hat{Y}_{Im}^{R} &= \hat{Y}_I - \sum_{i\in s_r}(\hat{B}_{\pi i}(I_i=1)+\hat{B}_{mi}) + \sum_{i\in s_r}\psi(\hat{B}_{\pi i}(I_i=1)+\hat{B}_{mi}) \\ &= \sum_{i\in s} d_i r_i y_i + \sum_{i\in s} d_i(1-r_i)\mathbf{z}_i^T\hat{\mathbf{B}}_{Rob} + \sum_{i\in s_r}\left[\psi(\hat{B}_{\pi i}+\hat{B}_{mi}) - \hat{B}_{\pi i}\right] \\ &= \hat{Y}_I^{\text{naïf}} + \text{terme de correction}\end{aligned}$$

- Absence de non-réponse: $s_r = s \Rightarrow \hat{B}_{mi} = 0 \Rightarrow \hat{Y}_{Im}^{R}$ se réduit à celui proposé dans Beaumont, Haziza and Ruiz-Gazen (2011) en l'absence de non-réponse.

Estimateur robuste sous l'approche par Modèle d'Imputation

- Similaire à l'estimateur robuste dans le cas de données complètes:

$$\begin{aligned}\hat{Y}_{Im}^{R} &= \hat{Y}_I - \sum_{i\in s_r}(\hat{B}_{\pi i}(I_i=1)+\hat{B}_{mi}) + \sum_{i\in s_r}\psi(\hat{B}_{\pi i}(I_i=1)+\hat{B}_{mi}) \\ &= \sum_{i\in s} d_i r_i y_i + \sum_{i\in s} d_i(1-r_i)\mathbf{z}_i^T\hat{\mathbf{B}}_{Rob} + \sum_{i\in s_r}\left[\psi(\hat{B}_{\pi i}+\hat{B}_{mi}) - \hat{B}_{\pi i}\right] \\ &= \hat{Y}_I^{\text{naïf}} + \text{terme de correction}\end{aligned}$$

- Absence de non-réponse: $s_r = s \Rightarrow \hat{B}_{mi} = 0 \Rightarrow \hat{Y}_{Im}^{R}$ se réduit à celui proposé dans Beaumont, Haziza and Ruiz-Gazen (2011) en l'absence de non-réponse.
- Nous obtenons un estimateur robuste similaire sous l'approche par Modèle de Non-réponse

Retour sur l'étude par simulation précédente...

- Nous avons généré $R = 10,000$ populations de taille $N = 10,000$ comportant une variable y
  - Population: mélange de deux distributions log-normales (moyenne et écart type dans l'échelle log)

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

$$y_i = (1-\delta_i)y_{0i} + \delta_i y_{1i} \quad \text{et} \quad Prob(\delta_i = 1) = 0.05; 0.1.$$

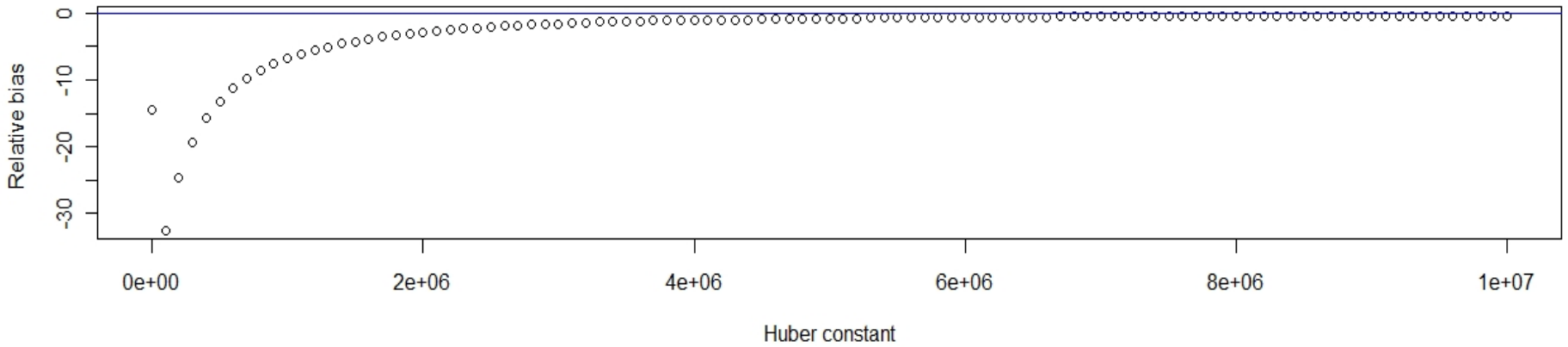
- Dans chaque population, nous avons tiré un EASSR de taille $n = 100$
- Dans chaque échantillon, nous avons généré la non-réponse selon un mécanisme de non-réponse uniforme: : i.e., $p_i = 0.7$ pour tout i
- Méthodes d'imputation:
  - l'imputation par la moyenne: chaque valeur manquante est remplacée par la moyenne des répondants
  - l'imputation par la médiane: chaque valeur manquante est remplacée par la médiane des répondants

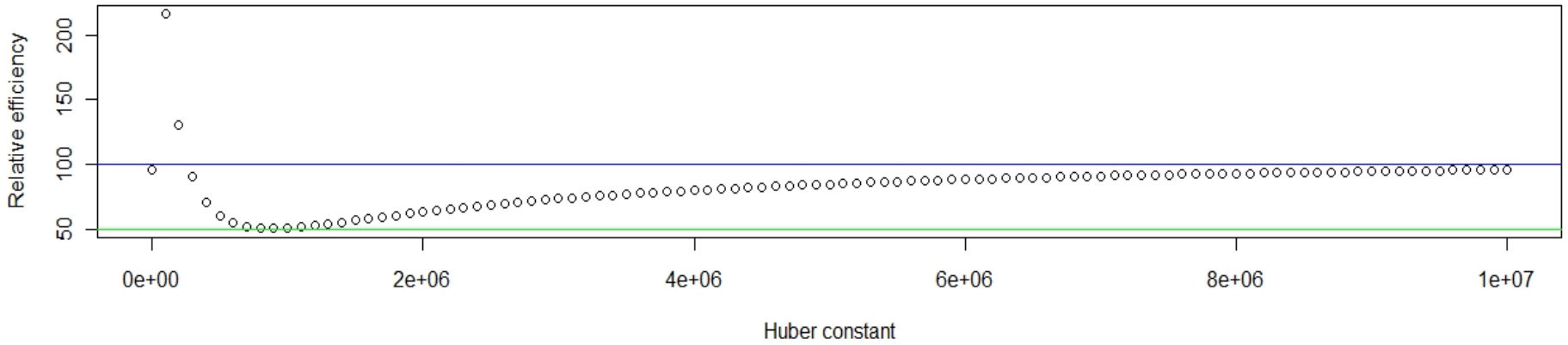
Résultats

Table: **Biais et EQM Monte Carlo**

	Contamination à 10%		Contamination à 5%	
	BR	ER	BR	ER
Estimateur naïf				
moyenne	0.3	100.0	0.3	100.0
médiane	-14.0	88.6	-11.9	94.4
Estimateur proposé				
$c = 8.5e5$	-7.9	44.5	-2.6	56.9
$c = 3.6e6$	-0.9	69.6	0.0	86.3

BR et ER vs. c, 10% de contamination





Conclusions et perspectives

- Nous avons commencé avec une procédure doublement robuste pour la rendre robuste à la présence d' unités influentes

$\Rightarrow$ Procédure Triplement Robuste

Conclusions et perspectives

- Nous avons commencé avec une procédure doublement robuste pour la rendre robuste à la présence d' unités influentes

$$\Rightarrow \text{ Procédure Triplement Robuste}$$

- Choix de la constante de réglage : trouver c qui minimise l'EQM estimée

Conclusions et perspectives

- Nous avons commencé avec une procédure doublement robuste pour la rendre robuste à la présence d' unités influentes

$$\Rightarrow \text{ Procédure Triplement Robuste}$$

- Choix de la constante de réglage : trouver c qui minimise l'EQM estimée
- L'estimation de la variance nécessite des investigations: utiliser l'approche renversée de Shao et Steel (1999)

Conclusions et perspectives

- Nous avons commencé avec une procédure doublement robuste pour la rendre robuste à la présence d' unités influentes

$$\Rightarrow \text{ Procédure Triplement Robuste}$$

- Choix de la constante de réglage : trouver c qui minimise l'EQM estimée
- L'estimation de la variance nécessite des investigations: utiliser l'approche renversée de Shao et Steel (1999)
- Étude de la méthode proposée pour des paramètres plus complexes