

Une nouvelle approche de pondération et d'inférence pour des échantillons tirés d'une population finie

Jean-François Beaumont

Statistique Canada

Société Statistique du Canada, Vancouver

31 mai 2009 – 3 juin 2009

Sommaire

- Description du problème
 - Variabilité des poids de sondages et poids extrêmes
- Idée principale
 - Lisser les poids de sondage au moyen d'un modèle
- Justification (empirique + théorique)
- Extensions / Applications
 - Unités sauteuses de strate, analyse de données (éq. d'est.), ajustement de poids pour la non-réponse et calage

Contexte des sondages

- Population finie : U
- Objectif:
 - Estimation de paramètres de la population finie
- Exemple:
 - Vecteur de totaux de population : $\mathbf{T}_y = \sum_{k \in U} \mathbf{y}_k$
 - \mathbf{y} est le vecteur des variables d'intérêt
 - \mathbf{Y} : matrice des valeurs de \mathbf{y} dans la population

Plan de sondage

- Sélection d'un éch. aléat. S à partir d'un plan de sondage

- Un plan de sondage est défini par:

1) L'ensemble de tous les échantillons possibles

2) La probabilité $p(S = s | \mathbf{Z})$ de sélect. chaque échant. S

- \mathbf{Z} : matrice des variables \mathbf{Z} du plan (e.g., indicateurs de strate, mesure de taille, ...)

- \mathbf{I} : vect. aléat. des indicateurs d'inclusion dans l'échantillon

→ connaître \mathbf{I} est équivalent à connaître $S = \{k \in U : I_k = 1\}$

- **Plan de sondage** : $F(\mathbf{I} | \mathbf{Z})$

Théorie fondée sur le plan

- Jusqu'à maintenant, on a défini 3 quantités: **I**, **Z** and **Y**
- **Inférence fondée sur la plan:** $F(\mathbf{I} | \mathbf{Z}, \mathbf{Y}) = F(\mathbf{I} | \mathbf{Z})$
 - Inférence est faite par rapport au plan de sondage
CONNU: on considère seulement **I** comme étant aléatoire
- Probabilité de sélection:

$$\pi_k(\mathbf{Z}) = \Pr(I_k = 1 | \mathbf{Z}, \mathbf{Y}) = E_p(I_k | \mathbf{Z}, \mathbf{Y})$$


Estimateur classique fondé sur le plan

- Poids de sondage: $w_k \equiv w_k(\mathbf{Z}) = 1/\pi_k(\mathbf{Z})$

- Estimateur de Horvitz-Thompson (HT)

$$\hat{\mathbf{T}}_y^{HT} = \sum_{k \in S} w_k \mathbf{y}_k = \sum_{k \in U} w_k \mathbf{y}_k I_k$$

- Est. HT est sans biais sous p : $\mathbf{E}_p \left(\hat{\mathbf{T}}_y^{HT} \mid \mathbf{Z}, \mathbf{Y} \right) = \mathbf{T}_y$

- Aucune hypothèse \ modèle est requise pour cette propriété  **Approche non-paramétrique**

Quel est le problème avec cette théorie?

- L'estimateur HT est sans biais sous le plan mais il peut être instable quand
 - **les poids de sondages sont faiblement associés aux variables d'intérêt et**
 - **sont fortement dispersés (avec possiblement des poids extrêmes)**
- Voir Rao(1966) et Basu (1971)
- Problème d'efficacité mais pas de validité

Une première solution

- **Hypothèse:** Les poids de sondage ne sont pas associés aux variables y
- Si cette hypothèse est vraie, les poids de sondage n'apportent aucune information sur les paramètres à estimer et peuvent donc être jetés
- Ceci a conduit Rao (1966) à suggérer l'estimat.:

$$\hat{\mathbf{T}}_y^{RAO} = N \frac{\sum_{k \in S} \mathbf{y}_k}{n} = \sum_{k \in S} (N/n) \mathbf{y}_k$$

Une première solution

- L'estim. de Rao n'est pas sans biais sous p mais
 - Son biais devrait être petit si l'hypothèse tient
 - Sa variance (anticipée) est plus petite que la variance (anticipée) de l'estimateur HT
- L'estimateur de Rao est très relié à:

$$\hat{\mathbf{T}}_y^{FS} = \sum_{k \in S} \left(\hat{N}/n \right) \mathbf{y}_k, \quad \hat{N}/n = \frac{\sum_{k \in S} w_k}{n}$$

⇒ \hat{N}/n est un poids complètement lissé

Étude par simulation

- Population: 50,000 unités
- Variable du plan: $z_k = 0.5 + \exp(\mu_z = 30)$

- Trois variables d'intérêt:

$$y_k^{(i)} = 30 + \beta^{(i)} z_k + N(0, 2000) , \quad i = 1, 2, 3$$

- Coefficients de corrélation:

$$\rho_{yz}^{(1)} = 0 ; \rho_{yz}^{(2)} = \sqrt{0.01} ; \rho_{yz}^{(3)} = \sqrt{0.8}$$

- Plan de sondage: ppt (Rao-Sampford) de taille 500

Étude par simulation

- Biais relatif d'un estimateur (BR):

$$\text{BR} = \frac{\text{biais sous le plan d'un estimateur}}{T_y} \times 100\%$$

- Efficacité relative d'un estimateur (ER):

$$\text{ER} = \frac{\text{EQM sous le plan d'un estimateur}}{\text{EQM sous le plan de l'estimateur HT}} \times 100\%$$

- BR et ER sont approchés en choisissant 50,000 échantillons
- **C'est une simulation fondée sur le plan**

Résultats de la simulation: BR (%)

Estimateur	$\rho_{yz}^{(1)} = 0$ (pas de cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (cor. faible)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (cor. forte)
HT	0.06	-0.01	-0.02
FS (Rao)	-0.77	12.05	73.34

Résultats de la simulation: ER (%)

Estimateur	$\rho_{yz}^{(1)} = 0$ (pas de cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (cor. faible)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (cor. forte)
HT	100	100	100
FS (Rao)	45.0	145	43095

Alternatives aux estimateurs HT ou FS

- **Winsorization des poids de sondage**

- **Problème:** un seuil de winsorization approprié pour une variable d'intérêt peut ne pas l'être pour une autre



Comment atteindre un compromis dans les enquêtes à objectifs multiples?

- Les gains d'efficacité, s'il y en a, sont habituellement modestes
- Pourquoi? Ne s'attaque pas au vrai problème (i.e., les grands poids ne sont pas nécessairement problématiques) et seulement quelques poids sont modifiés

Alternatives aux estimateurs HT ou FS

- **Approche de prédiction (fondé sur un modèle)**
- Inférence: $F(Y | Z, I)$
 - Distribution n'est pas contrôlé par le statisticien d'enquête: **un modèle est requis**
 - Note: $T_y = \sum_{k \in U} y_k$ est une variable aléatoire dans cette approche
- Royall (1970, 1976) a proposé le BLUP (“Best Linear Unbiased Predictor”) de T_y

Alternatives aux estimateurs HT ou FS

- Le BLUP est obtenu en trouvant des poids qui
 - 1) minimisent la var. sous le modèle de l'erreur de prédiction
 - 2) sous la contrainte que le prédicteur résultant est sans biais sous le modèle (i.e., l'espérance sous le modèle de l'erreur de prédiction est égale à 0)
- Equivalent au calage (Deville and Särndal, 1992) mais n'utilise pas les poids de sondage (Chambers, 1996)
- **Problème:** Spécifier et valider un modèle dans les enquêtes à plusieurs variables peut être une tâche énorme


Lissage de poids

- Le lissage est utilisé pour réduire l'instabilité de l'estimateur HT:

$$\tilde{\mathbf{T}}_y^{SHT} = \mathbf{E}_\xi \left(\hat{\mathbf{T}}_y^{HT} \mid \mathbf{I}, \mathbf{Y} \right) = \sum_{k \in S} \tilde{w}_k \mathbf{y}_k$$

- **Poids lissé:** $\tilde{w}_k = E_\xi (w_k \mid \mathbf{I}, \mathbf{Y})$
- **L'idée est d'enlever le bruit des poids et de garder seulement la portion "utile"**
- Pourquoi conditionner sur \mathbf{I} et \mathbf{Y} quand on évalue l'espérance?

Lissage de poids

- **Problème:** $\tilde{w}_k = E_{\xi}(w_k | \mathbf{I}, \mathbf{Y})$ est inconnu!
- **Solution:** Modéliser les poids de sondage pour obtenir un poids lissé \hat{w}_k
- Estimateur HT lissé:
$$\hat{\mathbf{T}}_y^{SHT} = \sum_{k \in S} \hat{w}_k \mathbf{y}_k$$
- Un estimateur sans biais évident est: $\hat{w}_k = w_k$
 **Conduit à l'estimateur HT**
- Il est robuste quant au biais mais inefficace: **une seule observation est utilisée pour estimer \tilde{w}_k**

Un modèle possible

- Modèle (requis seulement pour les unités $k \in S$):

$$w_k = 1 + \exp(\mathbf{h}'_k \boldsymbol{\beta} + \varepsilon_k) \quad , \quad \mathbf{h}_k = \mathbf{h}(y_k)$$

- Alternativement, il peut être écrit sous la forme

$$\ln(w_k - 1) = \mathbf{h}'_k \boldsymbol{\beta} + \varepsilon_k$$

- $\boldsymbol{\beta}$ peut être estimé en utilisant les moindres carrés ordinaires **NON PONDÉRÉS**
- Pourquoi? Parce qu'on conditionne sur **I**

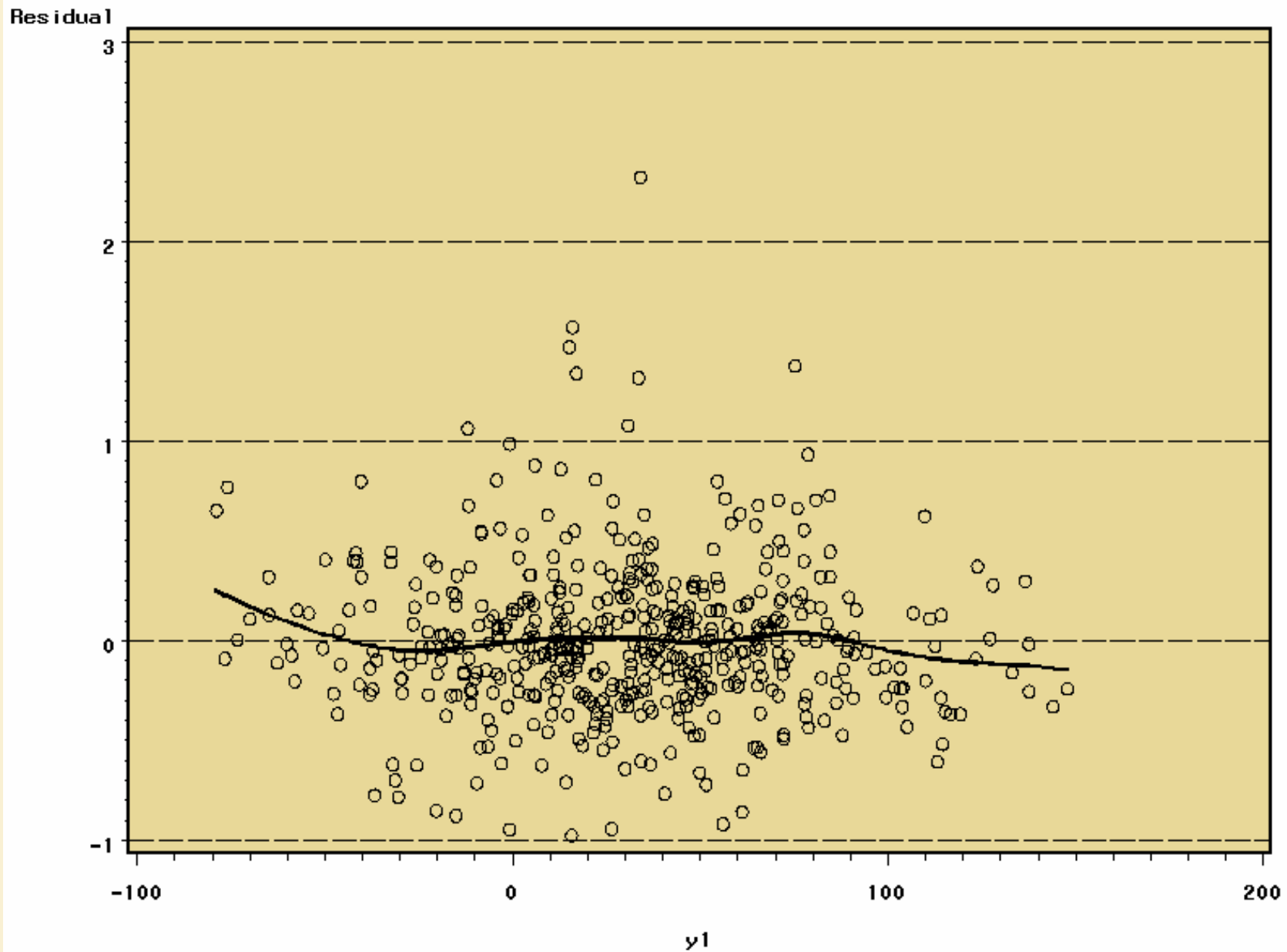
Un modèle possible

- En supposant la normalité des erreurs:

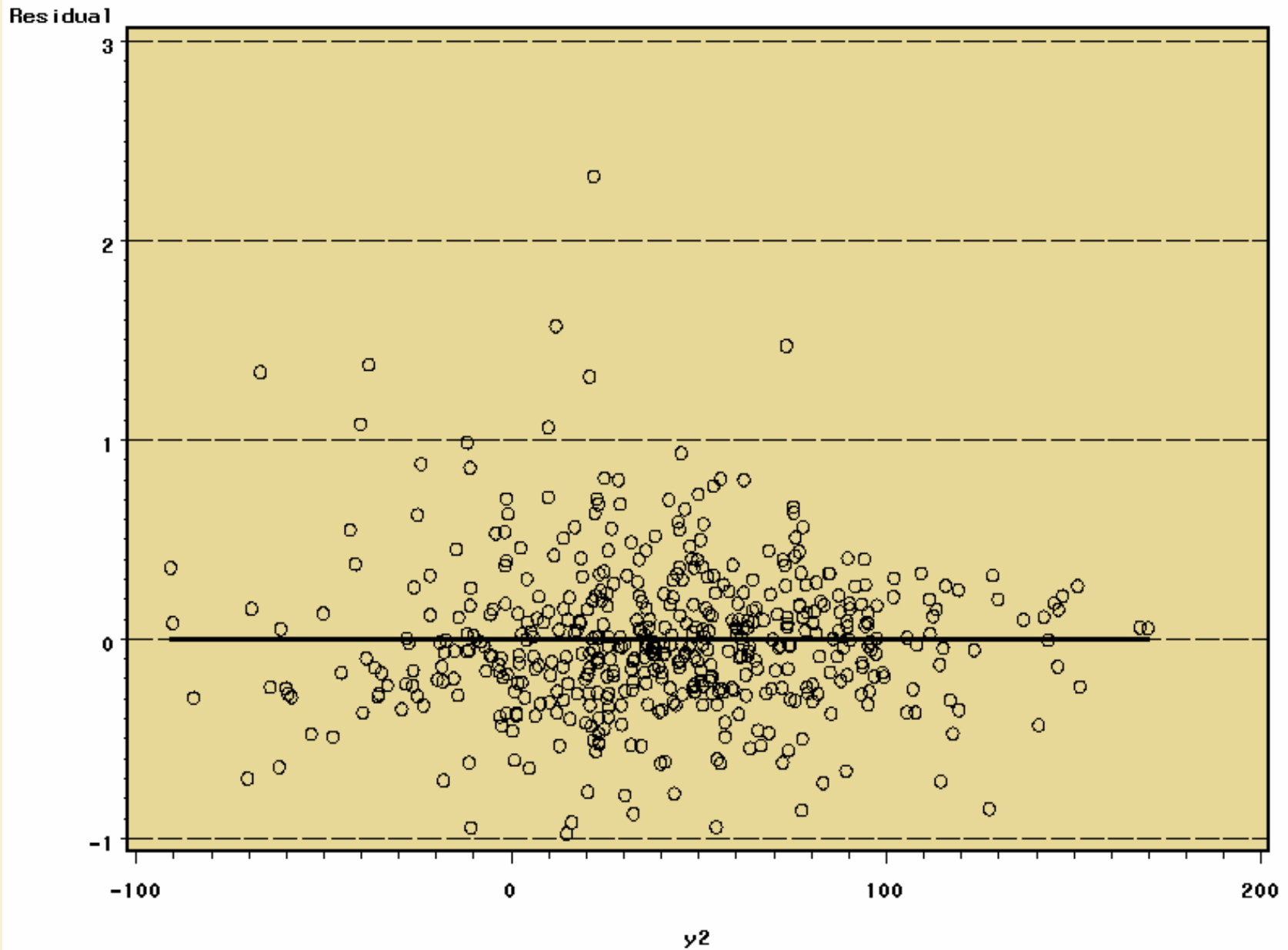
$$\tilde{w}_k = E_{\xi}(w_k \mid \mathbf{I}, \mathbf{Y}) = 1 + \exp\left(\mathbf{h}'_k \boldsymbol{\beta} + \sigma_{\varepsilon,k}^2 / 2\right)$$

- Le poids lissé estimé \hat{w}_k est obtenu en estimant $\boldsymbol{\beta}$ et $\sigma_{\varepsilon,k}^2$
- $\sigma_{\varepsilon,k}^2$ pourrait être estimé en calculant la variance des résidus à l'intérieur de groupes homogènes
- Beaumont (2008) évite l'hypothèse de normalité

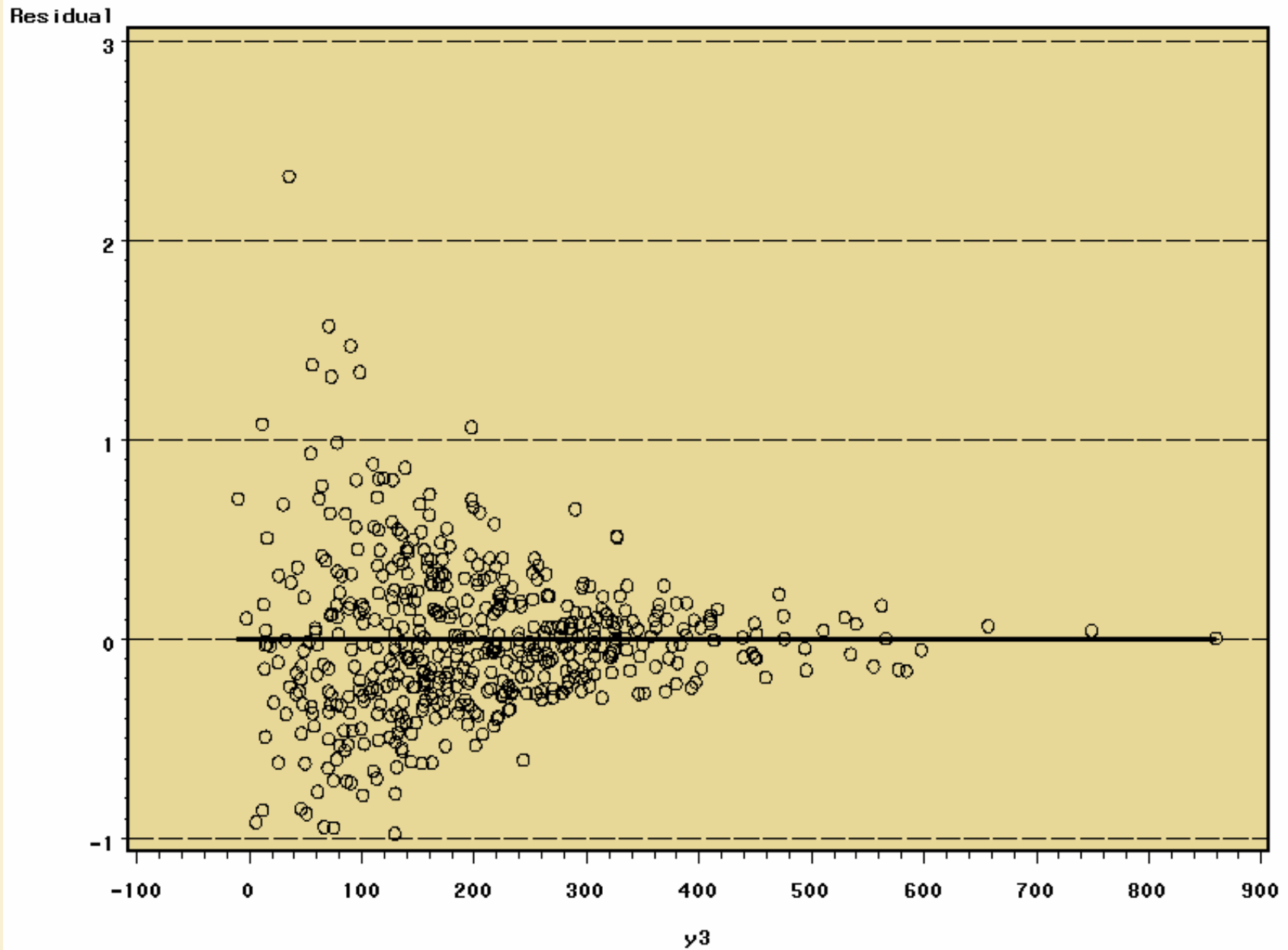
Graph. des résidus pour l'éch. 1 vs $y^{(1)}$



Graph. des résidus pour l'éch. 1 vs $y^{(2)}$



Graph. des résidus pour l'éch. 1 vs $y^{(3)}$



Résultats de la simulation: BR (%)

Estimateur	$\rho_{yz}^{(1)} = 0$ (pas de cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (cor. faible)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (cor. forte)
HT	0.06	-0.01	-0.02
FS (Rao)	-0.77	12.05	73.34
SHT	-2.78	-2.01	-2.23

Résultats de la simulation: ER (%)

Estimateur	$\rho_{yz}^{(1)} = 0$ (pas de cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (cor. faible)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (cor. forte)
HT	100	100	100
FS (Rao)	45.0	145	43095
SHT	60.8	59.3	96.2

Deux cas particuliers extrêmes de l'estimateur SHT

■ Cas 1: Pas d'association

■ Modèle: $w_k = \beta + \varepsilon_k$

■ Conduit à l'estimateur FS (poids lissé = poids moyen)

■ Cas 2: Association parfaite


■ Modèle: $w_k = g(\mathbf{y}_k)$  **No error term**

■ Conduit à l'estimateur HT (poids lissé = poids de sondage)

Justification théorique

- Le lissage de poids conduit à un biais sous le plan (devrait être petit si le modèle est bien spécifié)

$$\mathbf{E}_p \left(\tilde{\mathbf{T}}_y^{SHT} \mid \mathbf{Z}, \mathbf{Y} \right) \neq \mathbf{T}_y$$

- Approche proposée d'inférence: $F(\mathbf{I}, \mathbf{Z} \mid \mathbf{Y})$
- Similaire à l'approche fondée sur le plan puisqu'elle conditionne sur \mathbf{Y}  permet d'éviter la modélisation de \mathbf{Y}
- L'estimateur HT reste sans biais sous cette approche fondée sur le plan et le modèle

Justification théorique

- \tilde{w}_k **connu**: Par le théorème de Rao-Blackwell,

- $E_{\xi p} \left(\lambda' \tilde{\mathbf{T}}_y^{SHT} \mid \mathbf{Y} \right) = \lambda' \mathbf{T}_y \implies \text{Aucun biais}$

- $\text{var}_{\xi p} \left(\lambda' \tilde{\mathbf{T}}_y^{SHT} \mid \mathbf{Y} \right) \leq \text{var}_{\xi p} \left(\lambda' \hat{\mathbf{T}}_y^{HT} \mid \mathbf{Y} \right)$

- \tilde{w}_k **inconnu**: Si un modèle linéaire tient et est utilisé pour estimer \tilde{w}_k ,

- $E_{\xi p} \left(\lambda' \hat{\mathbf{T}}_y^{SHT} \mid \mathbf{Y} \right) = \lambda' \mathbf{T}_y \implies \text{Aucun biais}$

- $\text{var}_{\xi p} \left(\lambda' \hat{\mathbf{T}}_y^{SHT} \mid \mathbf{Y} \right) \leq \text{var}_{\xi p} \left(\lambda' \hat{\mathbf{T}}_y^{HT} \mid \mathbf{Y} \right)$

Estimation de la variance

- Estimer $\text{var}_{\xi p} \left(\lambda' \hat{\mathbf{T}}_y^{SHT} \mid \mathbf{Y} \right)$
 - Dépend de la validité du modèle pour les poids de sondage

- Une alternative est d'estimer l'EQM sous le plan:

$$E_p \left\{ \left(\lambda' \hat{\mathbf{T}}_y^{SHT} - \lambda' \mathbf{T}_y \right)^2 \mid \mathbf{Z}, \mathbf{Y} \right\}$$

- Ne requiert pas la validité d'un modèle
- A bien fonctionné dans une étude par simulations

Unités sauteuses de strate

- Qu'est-ce qu'une unité sauteuse de strate?
 - Une unité qu'on place dans une strate, selon l'information de la base de sondage (\mathbf{Z}), mais qui aurait été placée dans une autre strate, selon l'information de collecte (\mathbf{Z}^{col})
- Quel est le problème?
 - **Inefficacité:** Un grand poids de sondage peut être malencontreusement assigné à une unité avec une grande valeur de y

Un exemple simple

Strate de collecte	Strate du plan	Nombre d'unités	Poids de sondage
G	G	9	1
G	P	1	31
P	P	40	31
Somme sur les unités de l'échantillon		50	1280

Lissage de poids pour les unités sauteuses de strate

- Hypothèse:

$$F(\mathbf{Y} | \mathbf{Z}^{col}, \mathbf{Z}, \mathbf{I}) = F(\mathbf{Y} | \mathbf{Z}^{col}, \mathbf{I})$$

- Implique que:

$$F(\mathbf{Z} | \mathbf{Z}^{col}, \mathbf{Y}, \mathbf{I}) = F(\mathbf{Z} | \mathbf{Z}^{col}, \mathbf{I})$$

- Modèle:

$$\tilde{w}_k = E_{\xi}(w_k | \mathbf{I}, \mathbf{Z}^{col}, \mathbf{Y}) = g(\mathbf{z}_k^{col})$$

- Le poids lissé est simplement la moyenne des poids de sondage à l'intérieur de sa strate de collecte (Beaumont and Rivest 2007, 2008)

Un exemple simple

Strate de collecte	Strate du plan	Nombre d'unités	Poids de sondage	Poids lissé	Poids lissé (avec contrainte)
G	G	9	1	4	1 (1.0)
G	P	1	31	4	4 (4.1)
P	P	40	31	31	31 (31.7)
Somme sur les unités de l'échantillon		50	1280	1280	1253 (1280)

Efficités relatives bootstrap pour l'Enquête sur le milieu du travail et les employés

Estimateur	y_1	y_2	y_3	y_4	y_5
Lissé	41.9	31.4	73.3	246.9	46.4
M-estimat.	40.7	43.4	95.3	100	89.1
Poids Winsorizés	108.3	107.8	112.9	155.8	112.5
Fondé sur le plan	100	100	100	100	100

Équations d'estimation

- On peut être intéressé à estimer les paramètres d'un modèle au sujet de la distribution de y étant donné \mathbf{X}
- Équation d'estimation pondérée habituelle (Binder, 1983):

$$\sum_{k \in S} w_k u_k (y_k, \mathbf{x}_k; \boldsymbol{\beta}) = \mathbf{0}$$

- Peut être inefficace dû à l'utilisation des poids de sondage

Équations d'estimation

- Pour améliorer l'efficacité, Pfeiffermann and Sverchkov (1999) ont suggéré:

$$\sum_{k \in S} \frac{w_k}{E_{\xi}(w_k | I_k = 1, \mathbf{x}_k)} u_k(y_k, \mathbf{x}_k; \boldsymbol{\beta}) = \mathbf{0}$$

- Cette idée pourrait être combinée au lissage des poids pour encore plus de gains d'efficacité:

$$\sum_{k \in S} \frac{E_{\xi}(w_k | I_k = 1, \mathbf{x}_k, y_k)}{E_{\xi}(w_k | I_k = 1, \mathbf{x}_k)} u_k(y_k, \mathbf{x}_k; \boldsymbol{\beta}) = \mathbf{0}$$

Ajustement des poids pour la non-réponse

- Estimateur HT repondéré:

$$\hat{\mathbf{T}}_y^{RHT} = \sum_{k \in S_r} w_k a_k \mathbf{y}_k$$

- Lisser les ajustements de poids a_k :

$$\tilde{a}_k = E_{\xi}(a_k | w_k, \mathbf{y}_k), \text{ for } k \in S_r$$

- Conduit à l'estimateur HT repondéré lissé:

$$\hat{\mathbf{T}}_y^{SRHT} = \sum_{k \in S_r} w_k \hat{a}_k \mathbf{y}_k$$

Calage

- Le calage consiste à trouver des poids w_k^C près des poids de sondage et qui satisfont l'équation de calage

$$\sum_{k \in S} w_k^C \mathbf{x}_k = \mathbf{T}_x$$

- **Cas 1:** Estimer $\tilde{w}_k^C = E_\xi(w_k^C | \mathbf{I}, \mathbf{Y})$
 - L'équation de calage ne tient plus
- **Cas 2:** $\tilde{w}_k = E_\xi(w_k | \mathbf{I}, \mathbf{X}, \mathbf{Y})$ et ensuite caler
 - Possiblement moins efficace mais le calage est préservé

Conclusion

- **Idée principale:** Lisser les poids d'enquête pour en extraire la portion utile
- **Avantage principal:** le lissage améliore l'efficacité des estimateurs
- **Inconvénient principal:** la validité d'un modèle est requise (implique des diagnostics de validation du modèle)
 - Méthodes non-paramétriques?

Bibliographie

- **Basu** (1971, *Foundations of Statistical Inference*, Ed. Godambe & Sprott)
- **Beaumont** (2008, *Biometrika*)
- **Beaumont & Rivest** (2008, *Handbook of Statistics*, vol. 29, Ed. Pfeffermann & Rao)
- **Beaumont & Rivest** (2007, *Proc. of the Survey Methods Section, SSC*)
- **Binder** (1983, *International Statistical Review*)
- **Chambers** (1996, *Journal of Official Statistics*)
- **Chambers, Dorfman & Wehrly** (1993, *Journal of the American Statistical Association*)
- **Deville & Särndal** (1992, *Journal of the American Statistical Association*)
- **Pfeffermann & Sverchkov** (1999, *Sankhya*, Series B)
- **Rao** (1966, *Sankhya*, Series A)
- **Royall** (1970, *Biometrika*)
- **Royall** (1976, *Journal of the American Statistical Association*)

Thanks - Merci

For more
information
please contact

Pour plus
d'information,
veuillez contacter

Jean-François Beaumont

Jean-Francois.Beaumont@statcan.gc.ca