A New Approach to Weighting and Inference for Samples Drawn From a Finite Population

Jean-François Beaumont
Statistics Canada
Statistical Society of Canada, Vancouver
May 31 – June 3, 2009

Overview

- Description of the problem
  - Variability of design weights and extreme design weights
- Main idea
  - Smooth design weights through modelling
- Justification (empirical + theoretical)
- Extensions / Applications
  - Stratum jumpers, data analysis (estimating equations), nonresponse weight adjustment and calibration

Survey sampling set-up

- Finite population : U
- Objective:
  - Estimation of finite population parameters
- Example:
  - Vector of population totals : $\mathbf{T}_{\mathbf{y}} = \sum_{k \in U} \mathbf{y}_k$
  - $\mathbf{y}$ is the vector of variables of interest
  - $\mathbf{Y}$: matrix of population $\mathbf{y}$-values

Sampling design

- Selection of a random sample S using a sampling design
- A sampling design is defined by:

  1) The set of all possible samples

  2) The probability $p(S = s \mid \mathbf{Z})$ **of selecting each sample *S***

- $\mathbf{Z}$: matrix of design variables $\mathbf{z}$ (e.g., strata indicators, size measure, …)
- $\mathbf{I}$: random vector of sample inclusion indicators

⇒ knowing $\mathbf{I}$ is equivalent to knowing $S = \{k \in U : I_k = 1\}$

- **Sampling design :** $F(\mathbf{I} \mid \mathbf{Z})$

Design-based theory

- So far, 3 quantities have been defined: **I**, **Z** and **Y**
- Design-based inference: $F(\mathbf{I} \mid \mathbf{Z}, \mathbf{Y}) = F(\mathbf{I} \mid \mathbf{Z})$
  - Inference is made with respect to the **KNOWN** sampling design: only **I** is taken to be random
- Selection probability:

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

Basic design-based estimator

- Design weight: $w_k \equiv w_k(\mathbf{Z}) = 1/\pi_k(\mathbf{Z})$
- Horvitz-Thompson (HT) estimator:

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

- HT estimator is *p*-unbiased: $\mathbf{E}_p\left(\hat{\mathbf{T}}_{\mathbf{y}}^{HT} \mid \mathbf{Z}, \mathbf{Y}\right) = \mathbf{T}_{\mathbf{y}}$
- No assumption/model is required for this property ⇨ **Nonparametric approach**

What's the problem with this theory?

- The HT estimator is design-unbiased but it can be quite unstable when
  - **the design weights are weakly associated with the variables of interest and**
  - **are widely dispersed (with perhaps extreme weights)**
- See Rao(1966) and Basu (1971)
- Problem of efficiency but not of validity

A first solution

- **Assumption:** The design weights are not associated with the **y**-variables
- If this assumption is true, the design weights do not bring any information about the parameters to be estimated and can thus be thrown away
- This led Rao (1966) to suggest the estimator:

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

A first solution

- Rao's estimator is not design-unbiased but
  - Its bias should be small if the assumption holds
  - Its (model-design) variance is smaller than the (model-design) variance of the HT estimator
- Rao's estimator is closely related to:

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

$\Rightarrow$ $\hat{N}/n$ is a Fully Smoothed (FS) weight

Simulation study

- Population: 50,000 units
- Design variable: $z_k = 0.5 + \exp(\mu_z = 30)$
- Three variables of interest:

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

- Coefficients of correlation:

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

- Sampling design: pps (Rao-Sampford) of size 500

Simulation study

- Relative bias of an estimator (RB):

$$\text{RB} = \frac{(\text{design})\ \text{bias of an estimator}}{T_y} \times 100\%$$

- Relative efficiency of an estimator (RE):

$$\text{RE} = \frac{(\text{design})\ \text{MSE of an estimator}}{(\text{design})\ \text{MSE of the HT estimator}} \times 100\%$$

- RB and RE are approximated by selecting 50,000 samples
- **This is a design-based simulation experiment**

Simulation results: RB (%)

Estimator	$\rho_{yz}^{(1)} = 0$ (no cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (weak cor.)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (strong cor.)
HT	0.06	-0.01	-0.02
FS (Rao)	**-0.77**	**12.05**	**73.34**

Simulation results: RE (%)

Estimator	$\rho_{yz}^{(1)} = 0$ (no cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (weak cor.)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (strong cor.)
HT	100	100	100
FS (Rao)	**45.0**	**145**	**43095**

Alternatives to HT or FS estimators

- **Winsorization of design weights**
- **Issue:** an appropriate winsorization cut-off for one variable of interest may not be appropriate for another ⇒ a compromise is needed in multipurpose surveys : How to achieve it?
- The efficiency gains, if any, are usually modest
- Why? It does not address the real problem (i.e., large weights are not necessarily a problem) and only few weights are modified

Alternatives to HT or FS estimators

- **Prediction approach (model-based approach)**
- Inference: $F(\mathbf{Y} \mid \mathbf{Z}, \mathbf{I})$
  - Distribution is not controlled by the survey statistician: **a model is required**
  - Note: $\mathbf{T}_y = \sum_{k \in U} \mathbf{y}_k$ is a random variable in this approach
- Royall (1970, 1976) proposed the BLUP (Best Linear Unbiased Predictor) of $\mathbf{T}_y$

Alternatives to HT or FS estimators

- The BLUP is obtained by finding weights that
  1) minimize the model variance of the prediction error
  2) under the constraint that the resulting predictor is model-unbiased (i.e., the model expectation of the prediction error is equal to 0)
- Equivalent to calibration (Deville and Särndal, 1992) but does not use the design weights (Chambers, 1996)
- **Issue:** Specifying and validating the model in multipurpose surveys may be a huge task

Weight smoothing

- Smoothing is used to reduce the instability of the HT estimator:

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

- Smoothed weight: $\tilde{w}_k = E_{\xi}(w_k \mid \mathbf{I}, \mathbf{Y})$
- **The idea is to remove noise from the weights and only keep the "useful" portion**
- Why conditioning on **I** and **Y** when taking the expectation?

Weight smoothing

- **Issue:** $\tilde{w}_k = E_\xi(w_k \mid \mathbf{I}, \mathbf{Y})$ is unknown!
- **Solution:** Model the design weights to obtain an estimated smoothed weight $\hat{w}_k$
- Smoothed HT estimator: $\hat{\mathbf{T}}_{\mathbf{y}}^{SHT} = \sum_{k \in S} \hat{w}_k \mathbf{y}_k$
- An obvious unbiased estimator is: $\hat{w}_k = w_k$

  ⇒ **Leads to HT estimator**
- It is bias-robust but inefficient: only one observation is used to estimate $\tilde{w}_k$

A possible model

- Model (required for sample units $k \in S$ only):

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

- It can alternatively be written as

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

- $\boldsymbol{\beta}$ can be estimated using **UNWEIGHTED** OLS

- Why? Because conditioning is on **I**

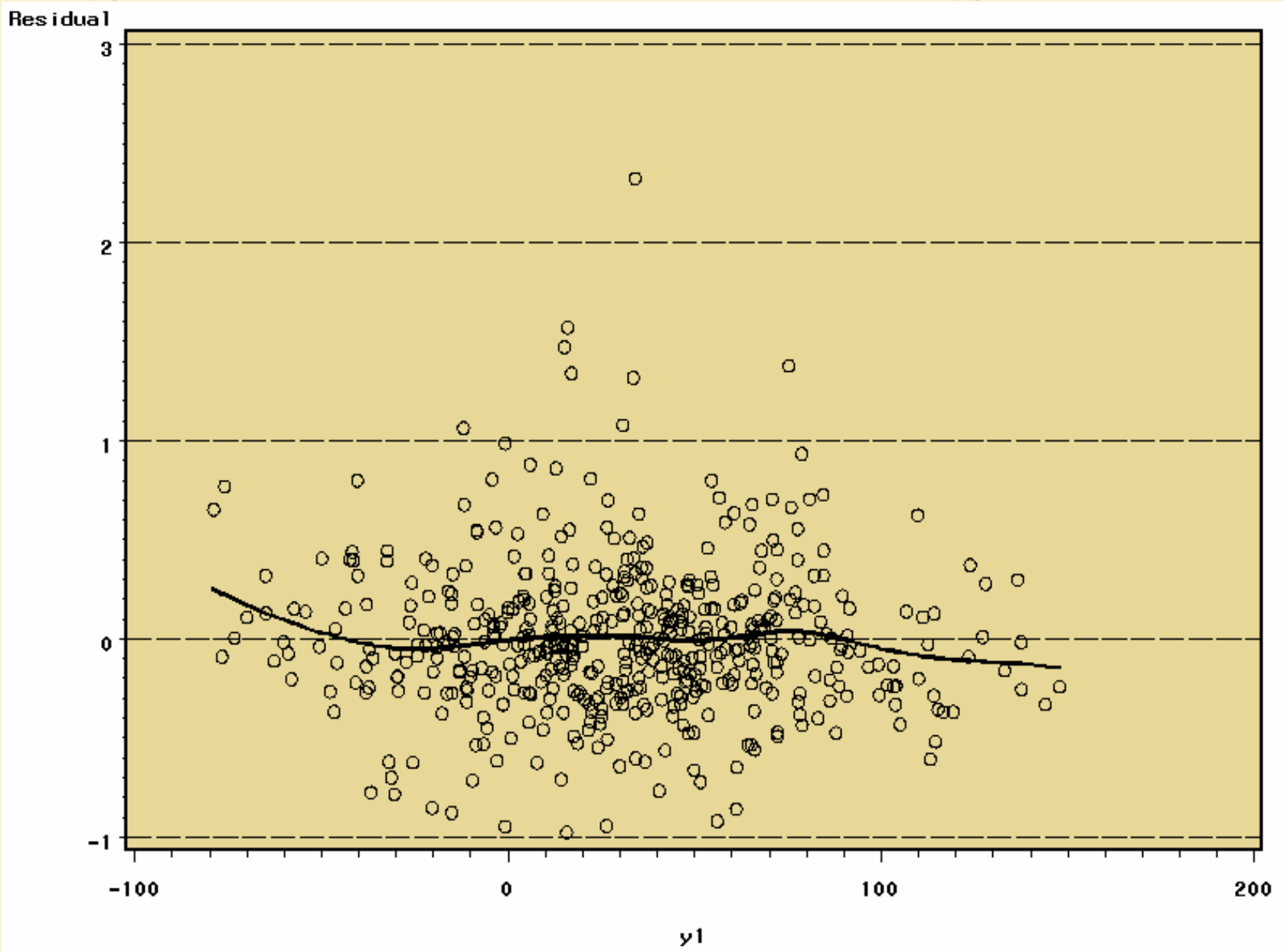
A possible model

- Assuming normality of the errors:

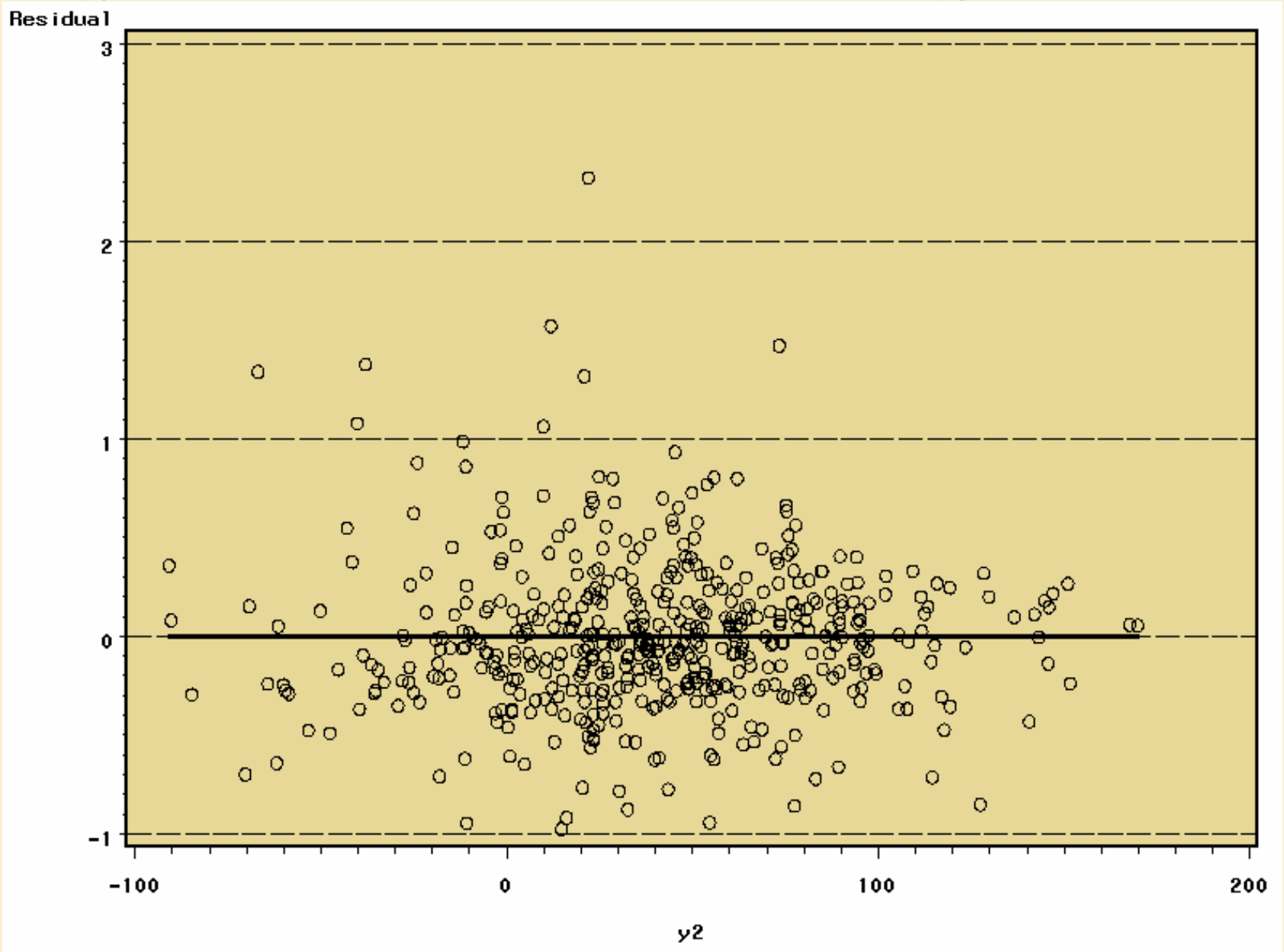
$$\tilde{w}_k = E_\xi(w_k \mid \mathbf{I}, \mathbf{Y}) = 1 + \exp\left(\mathbf{h}_k'\boldsymbol{\beta} + \sigma^2_{\varepsilon,k}/2\right)$$

- The estimated smoothed weight $\hat{w}_k$ is obtained by estimating $\boldsymbol{\beta}$ and $\sigma^2_{\varepsilon,k}$
- $\sigma^2_{\varepsilon,k}$ could be estimated by computing the variance of residuals within homogeneous groups
- Beaumont (2008) avoids the normality assumption

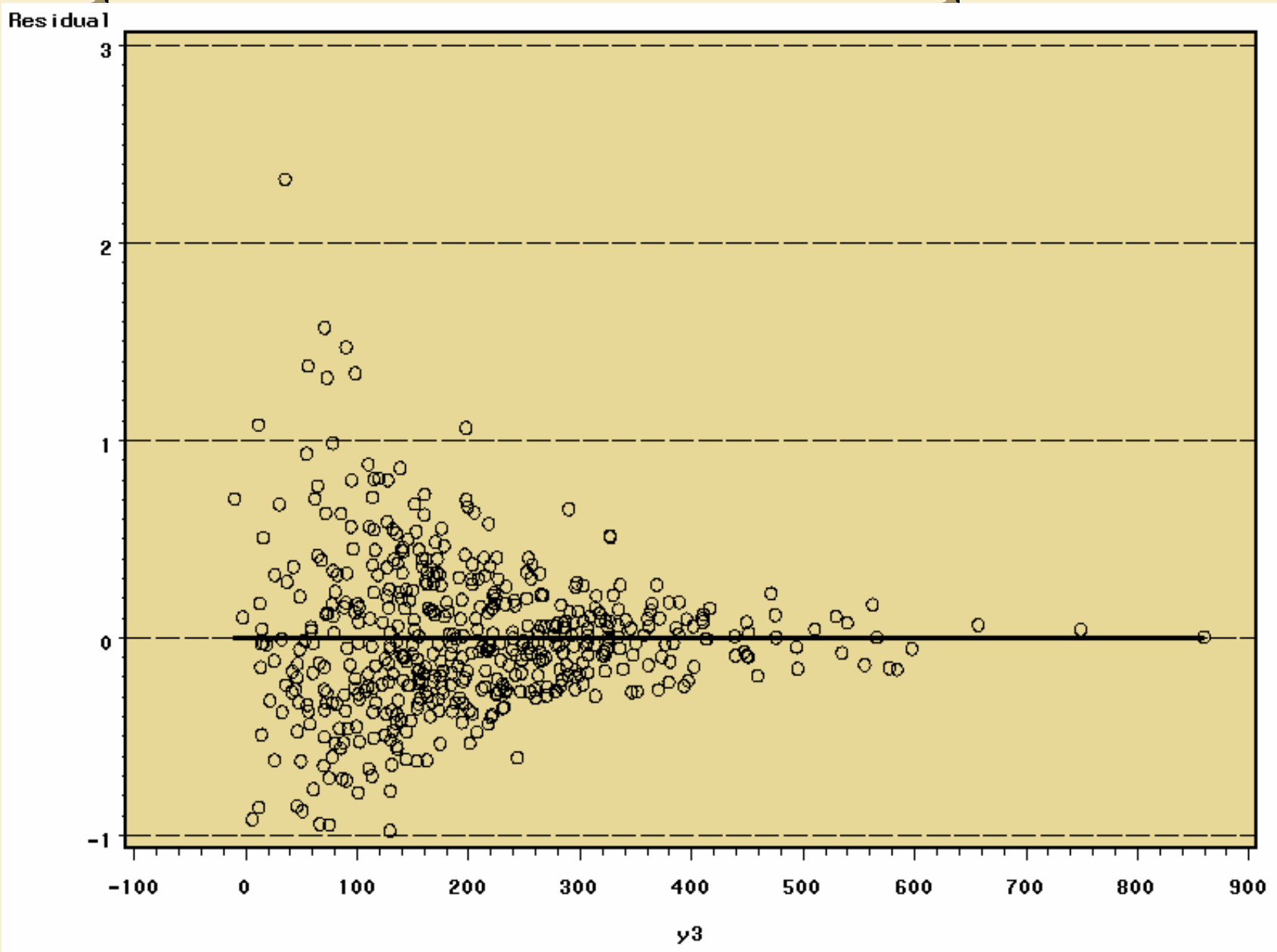
Graph of residuals for sample 1 vs $y^{(1)}$



Graph of residuals for sample 1 vs $y^{(2)}$



Graph of residuals for sample 1 vs $y^{(3)}$



Simulation results: RB (%)

Estimator	$\rho_{yz}^{(1)} = 0$ (no cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (weak cor.)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (strong cor.)
HT	0.06	-0.01	-0.02
FS (Rao)	-0.77	12.05	73.34
SHT	**-2.78**	**-2.01**	**-2.23**

Simulation results: RE (%)

Estimator	$\rho_{yz}^{(1)} = 0$ (no cor.)	$\rho_{yz}^{(2)} = \sqrt{0.01}$ (weak cor.)	$\rho_{yz}^{(3)} = \sqrt{0.8}$ (strong cor.)
HT	100	100	100
FS (Rao)	45.0	145	43095
SHT	**60.8**	**59.3**	**96.2**

Two extreme special cases of the SHT estimator

- **Case 1: No association**
  - Model: $w_k = \beta + \varepsilon_k$
  - Leads to the FS estimator (smoothed weight = average weight)
- **Case 2: Perfect association**
  - Model: $w_k = g(\mathbf{y}_k)$ ⇒ **No error term**
  - Leads to the HT estimator (smoothed weight = design weight)

Theoretical justification

- Weight smoothing leads to design bias (but should be small if the model is properly specified)

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

- Proposed approach to inference: $F(\mathbf{I}, \mathbf{Z} \mid \mathbf{Y})$
- Similar to design-based approach as it conditions on $\mathbf{Y}$ ⇨ allows us to avoid modelling $\mathbf{Y}$
- The HT estimator remains unbiased under this model-design approach

Theoretical justification

- $\tilde{w}_k$ **known:** By the Rao-Blackwell theorem,
  - $E_{\xi p}\left(\boldsymbol{\lambda}'\tilde{\mathbf{T}}_{\mathbf{y}}^{SHT} \mid \mathbf{Y}\right) = \boldsymbol{\lambda}'\mathbf{T}_{\mathbf{y}}$ ⇒ **No bias**
  - $\mathrm{var}_{\xi p}\left(\boldsymbol{\lambda}'\tilde{\mathbf{T}}_{\mathbf{y}}^{SHT} \mid \mathbf{Y}\right) \leq \mathrm{var}_{\xi p}\left(\boldsymbol{\lambda}'\hat{\mathbf{T}}_{\mathbf{y}}^{HT} \mid \mathbf{Y}\right)$
- $\tilde{w}_k$ **unknown:** If a linear model holds and is used to estimate $\tilde{w}_k$,
  - $E_{\xi p}\left(\boldsymbol{\lambda}'\hat{\mathbf{T}}_{\mathbf{y}}^{SHT} \mid \mathbf{Y}\right) = \boldsymbol{\lambda}'\mathbf{T}_{\mathbf{y}}$ ⇒ **No bias**
  - $\mathrm{var}_{\xi p}\left(\boldsymbol{\lambda}'\hat{\mathbf{T}}_{\mathbf{y}}^{SHT} \mid \mathbf{Y}\right) \leq \mathrm{var}_{\xi p}\left(\boldsymbol{\lambda}'\hat{\mathbf{T}}_{\mathbf{y}}^{HT} \mid \mathbf{Y}\right)$

Variance estimation

- Estimate $\mathrm{var}_{\xi p}\left(\boldsymbol{\lambda}'\hat{\mathbf{T}}_{\mathbf{y}}^{SHT} \mid \mathbf{Y}\right)$
  - Depends on the validity of the model for the design weights
- An alternative is to estimate the design MSE:

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

  - Does not need the validity of the model
  - Worked well in a simulation study

Stratum jumpers

- What's a stratum jumper?
  - A unit that is put in some design stratum based on frame information ($\mathbf{Z}$) but that would have been put in another stratum based on collection information ($\mathbf{Z}^{col}$)
- What's the problem?
  - **Inefficiency:** A unit with a large y-value can be inadvertently assigned a large design weight

A simple example

Collection stratum	Design stratum	Number of units	Design weight
L	L	9	1
L	S	1	31
S	S	40	31
Sum over the sample units		50	1280

Weight smoothing for stratum jumpers

- Assumption:

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

- Implies that:

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

- Model:

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

- Smoothed weight is simply the average of the design weights within collection strata (Beaumont and Rivest 2007, 2008)

A simple example

Collection stratum	Design stratum	Number of units	Design weight	Smoothed weight	Smoothed weight (with constraint)
L	L	9	1	4	1 (1.0)
L	S	1	31	4	4 (4.1)
S	S	40	31	31	31 (31.7)
Sum over the sample units		50	1280	1280	1253 (1280)

Bootstrap relative efficiencies for the Workplace and Employee Survey

Estimator	y_1	y_2	y_3	y_4	y_5
Smoothed	41.9	31.4	73.3	246.9	46.4
M-estimat.	40.7	43.4	95.3	100	89.1
Winsor. weights	108.3	107.8	112.9	155.8	112.5
Design-based	100	100	100	100	100

Estimating equations

- We may be interested in estimating model parameters about the distribution of y given $\mathbf{x}$
- Usual weighted estimating equation (Binder, 1983):

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

- May be inefficient due to using design weights

Estimating equations

- To increase efficiency, Pfeffermann and Sverchkov (1999) suggested:

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

- This idea could be combined with weight smoothing for more efficiency gains:

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

Nonresponse weight adjustment

- Reweighted HT estimator:

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

- Smooth the weight adjustment a_k :

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

- This leads to the Smoothed Reweighted HT est.:

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

Calibration

- Calibration consists of finding weights w_k^C close to the design weights and satisfying the calibration equation

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

- **Case 1:** Estimate $\tilde{w}_k^C = E_\xi(w_k^C \mid \mathbf{I}, \mathbf{Y})$
  - The calibration equation is lost
- **Case 2:** $\tilde{w}_k = E_\xi(w_k \mid \mathbf{I}, \mathbf{X}, \mathbf{Y})$ and then calibrate
  - Possibly less efficient but calibration is preserved

Conclusion

- **Main idea:** Smooth survey weights so as to extract their useful portion
- **Main advantage:** smoothing increases efficiency
- **Main disadvantage:** the validity of a model is required (implies model validation diagnostics)
  - Nonparametric methods?

References

- **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