Variance Estimation for Survey-Weighted Data Using Bootstrap Resampling Methods: 2013 Methods-of-Payment Survey Questionnaire

Heng Chen and Q. Rallye Shen

2017 STATA

The views expressed in this presentation are those of the author. No responsibility for them should be attributed to the Bank of Canada.

June 9, 2017

Outline

- Basic concepts
- Calibration estimator in survey: related to incomplete data estimator in econometrics
- Variance estimation of the calibration estimator
- Result

Basic Concepts

- Let U be a finite population of size N and the population total be $T_y = \sum_{i \in U} y_i$.

- A sample s is selected according to a sampling design $p(s)$ with inclusion probabilities, $\pi_1, ..., \pi_N$.

- Calibration estimator:

$$\widehat{T}_y = \sum_{i \in s} w_i y_i$$

where w_i is the calibrated weight.

We want to estimate $Var\left(\widehat{T}_y\right)$.

Calibrated Weight

- Each unit i is assigned a **design weight**:

$$\left\{ d_i = \pi_i^{-1}; i \in s \right\}.$$

- **Calibration adjustment**: The weight d_i are adjusted to match known(imputed) population counts.

$$d_i \overset{calibration}{\longrightarrow} w_i$$

where w_i from minimizing

$$\sum_{i \in s} d_i \left[\left(\frac{w_i}{d_i} \right) \log(\frac{w_i}{d_i}) - \left(\frac{w_i}{d_i} \right) + 1 \right]$$

subject to $\sum_{i \in s} w_i x_i = \mathbf{X}$.

Problem

Recall

$$Var(\widehat{T}_y) = Var\left(\sum_{i \in s} w_i y_i\right)$$

where there are two sources of randomness in $\widehat{T}_y$:

1. Sampling variation in the s;
2. Sampling variation in the calibrated weight w_i.

However, people usually ignore the sampling variation in w_i, and this is problematic.

Problem (cont.)

- The estimated variance can be written as

$$\widehat{Var}(\widehat{T}_y) \approx \sum_{i,j\in s} \frac{d_{ij} - d_i d_j}{d_{ij}} \frac{y_i - \widehat{\beta} x_i}{d_i} \frac{y_i - \widehat{\beta} x_i}{d_j}$$

where $\widehat{\beta}$ is $\left[\sum_s x_i x_i'\right]^{-1} \left[\sum_s y_i x_i'\right]$.

- Compare to

$$\widehat{Var}^*(\widehat{T}_y) \approx \sum_{i,j\in s} \frac{w_{ij} - w_i w_j}{w_{ij}} \frac{y_i}{w_i} \frac{y_j}{w_j}$$

which ignores the sampling variation in w_i.

Bootstrap in STATA

- Bootstrap is easy to implement.

1. Use the *ipfraking* and *bsweights* commands in Stata (Kolenikov, 2010, 2014).
2. Generate replicate calibrated weights instead of recomputing the statistics for each resample.

					weights	Replicate weights		
Unit	y_1	y_2	...	y_p	ω	ω^{R1}	ω^{R2}	ω^{R3}
1	y_{11}	y_{21}	...	y_{p1}	ω_1	$\omega 1^{R1}$	$\omega 1^{R2}$	$\omega 1^{R3}$
2	y_{12}	y_{22}	...	y_{p2}	ω_2	$\omega 2^{R1}$	$\omega 2^{R2}$	$\omega 2^{R3}$
.	.	.		.	.	.	.	.
.	.	.		.	.	.	.	.
.	.	.		.	.	.	.	.
n	y_{1n}	y_{2n}	...	y_{pn}	ω_n	ωn^{R1}	ωn^{R2}	ωn^{R3}

Implementation in Stata

- Step 1: Input the initial weight (d).
- Step 2: Generate the calibrated weight (w) using the *ipfraking* command.

$$\text{Steps: } d_i \overset{calibration}{\longrightarrow} w_i$$

- Step 3: Generate the replicate raking weights using the *bsweights* command.
- Step 4: Declare the bootstrap survey environment in Stata: `svyset [pw=], vce(bootstrap) bsrw()`

Result

	Ignoring w	**Considering w**
Cash on Hand		
Mean	1.03	1.03
Variance	1.51	0.85
Usage of CTC		
Mean	0.93	0.93
Variance	2.10	1.24

Note: The numbers in the second and third columns are divided by the numbers in the first column. CTC stands for the contactless feature of a credit card.