Variance Estimation for Calibration to Estimated Control Totals

Siyu Qing

Coauthor with
Michael D. Larsen
Associate Professor of Statistics

Tuesday, 11/05/2013

Outline

- A. Background
- B. Calibration Technique
- C. Calibration Weighting with Estimated Control Totals
- D. Variance Estimation
- E. Simulation
- E. Discussion

What is calibration weighting?

- Let $\{d_k\}$ be original survey (design) weights.
- $t_x = \sum_U x_k$ is a known total in the population with indices U; x_k can be a vector
- The calibrated weights $\{w_k\}$ are "close" to $\{d_k\}$ but satisfy a set of calibration equations:

$$\sum_s w_k x_k = \sum_U x_k.$$

- The "closeness" is measured by a distance function:
  1. Linear calibration: $E_p\{\sum_s (w_k - d_k)^2 / d_k q_k\}$
  2. Raking: $E_p\{\sum_s [w_k log(w_k/d_k) - w_k + d_k]\}$
  3. Logit method and others.
- The calibration estimator $\hat{t}_{yw} = \sum_s w_k y_k$.

Calibration Estimator Properties

- The calibration weights can be written as $w_k = d_k F_k(\boldsymbol{x}_k' \widehat{\boldsymbol{\lambda}}_s)$, where $F_k(\cdot)$ comes from the inverse of distance function.
- The linear calibration estimator

$$\hat{t}_{yl} = \sum_s w_k y_k = \hat{t}_{Ay} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{Ax})' \hat{\boldsymbol{B}}_s$$

- $AV(\hat{t}_{yw}) = AV(\hat{t}_{yl})$.
- Calibration weigthing can reduce mean squared error.
- There are various ways to compute the weights, including in the survey and TeachingSampling packages in R.

Calibration on Estimated Control Totals

- In calibration only $\boldsymbol{t}_x$ are needed from outside source.
- Sometimes $\boldsymbol{t}_x$ are not available and one may seek estimated totals $\hat{\boldsymbol{t}}_{Cx}$ instead.
- The calibration constraint equation becomes:

$$\sum_s w_k \boldsymbol{x}_k = \hat{\boldsymbol{t}}_{Cx}.$$

- Calibration estimator with estimated control totals $\hat{t}_{yW}$ (CEEC).
- Assuming *Analytic Survey* and *Control Total Survey* are independent.

Formula of the CEEC

- The linear CEEC

$$\hat{t}_{yL} = \sum\nolimits_s w_k y_k = \hat{t}_{Ay} + (\hat{\boldsymbol{t}}_{Cx} - \hat{\boldsymbol{t}}_{Ax})' \hat{\boldsymbol{B}}_s$$

, where $\hat{\boldsymbol{B}}_s = \boldsymbol{T}_s^{-1} \sum\nolimits_s d_k q_k \boldsymbol{x}_k y_k$, and $\boldsymbol{T}_s = \sum\nolimits_s d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k{\prime}$.

- The general CEEC

$$\hat{t}_{yW} = \sum_s w_k y_k = \sum_s d_k F_k(\boldsymbol{x}_k' \widehat{\gamma}_s) y_k$$

where $\widehat{\gamma}_s$ can be computationally solved from calibration constraints.

Assumptions for the CEEC

1. $\max ||\boldsymbol{x}_k||$ and $F_k''(0)$ are both uniformly bounded by $K_1 < \infty$ and $K_2 < \infty$, respectively.
2. $\lim N^{-1} \boldsymbol{t}_x$ exists.
3. $N^{-1}(\hat{\boldsymbol{t}}_{Ax} - \boldsymbol{t}_x) \to 0$ in design probability.
4. $n^{1/2} N^{-1}(\hat{\boldsymbol{t}}_{Ax} - \boldsymbol{t}_x) \xrightarrow{d} MVN(0, \boldsymbol{\Sigma}_A)$.
5. $C_\lambda = \bigcap_{k \in U} \{\boldsymbol{\lambda} : \boldsymbol{x}_k' \boldsymbol{\lambda} \in \mathrm{Im}_k(d_k)\}$ is a convex domain as well as an open neighborhood of 0.
6. $N^{-1}(\hat{\boldsymbol{t}}_{Cx} - \boldsymbol{t}_x) \to 0$ in design probability.
7. $n^\alpha N^{-1}(\hat{\boldsymbol{t}}_{Cx} - \boldsymbol{t}_x) \xrightarrow{d} MVN(0, \boldsymbol{\Sigma}_C)$ with $\alpha \geq 1/2$, so the control total survey is at least as accurate as the analytic survey we conduct.

Properties of the CEEC

- $\hat{t}_{yL}$ and $\hat{t}_{yl}$ has exactly the same expectation.
- $N^{-1}(\hat{t}_{yL} - \hat{t}_{yl}) = O_p(n^{-\alpha})$
- $\hat{t}_{yW}$ is design consistent as well as asymptotically design unbiased.
- $N^{-1}(\hat{t}_{yW} - \hat{t}_{Ay}) = O_p(n^{-1/2})$ and $N^{-1}(\hat{t}_{yW} - \hat{t}_{yL}) = O_p(n^{-1})$.
- $n^{1/2}N^{-1}(\hat{t}_{yW} - \hat{t}_{yl}) = O_p(n^{1/2-\alpha})$.

Variance Estimation for the CEEC

- Extra variation brought by the estimated control totals.
- Variance might be underestimated by traditional methods.
- the contribution of the bias to the MSE is likely to be small.
- The key is to precisely estimate the inflation part of the variance.
- The asymptotic variance of $\hat{t}_{yW}$ is:

$$\begin{aligned} AV(\hat{t}_{yW}) &= AV(\hat{t}_{yw}) + \boldsymbol{B}' V(\hat{\boldsymbol{t}}_{Cx})\boldsymbol{B} \\ &= \sum\sum_U \Delta_{kl} d_k E_k d_l E_l + \boldsymbol{B}' V(\hat{\boldsymbol{t}}_{Cx})\boldsymbol{B}, \end{aligned}$$

where $\boldsymbol{B} = (\sum_U \boldsymbol{x}_k \boldsymbol{x}_k')^{-1}(\sum_U \boldsymbol{x}_k y_k)$.

A traditional variance estimator (Naive)

- Traditional variance estimator considers the estimated control totals as they were true population totals.
- Deville and Särndal, 1992, *JASA* first defined the calibration estimator as well as gave a variance estimator.
- The formula is:

$$\widehat{V}_{naive}(\hat{t}_{yW}) = \sum\sum_{s}(\Delta_{kl}/\pi_{kl})(w_k e_k)(w_l e_l)$$

  where $w_k = d_k F_k(\boldsymbol{x}_k' \widehat{\gamma}_s)$, and $e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}_s$ is the sample fit residual.
- Whether this formula is a good estimate depends on the accuracy of control total survey.

Taylor linearization variance estimator (TL)

- The Taylor linearization variance estimator is:

$$\begin{aligned}\hat{V}_{TL}(\hat{t}_{yW}) &= \sum\sum_s \check{\Delta}_{kl} w_k e_{ks} w_l e_{ls} + \hat{\boldsymbol{B}}_s' \widehat{V}(\hat{\boldsymbol{t}}_{Cx}) \hat{\boldsymbol{B}}_s \\ &\stackrel{\text{def}}{=} \widehat{V}_{naive} + \widehat{V}_{inf}\end{aligned}$$

- $E_p(\hat{\boldsymbol{B}}_s' \widehat{V}(\hat{\boldsymbol{t}}_{Cx}) \hat{\boldsymbol{B}}_s) = \boldsymbol{B}' V(\hat{\boldsymbol{t}}_{Cx}) \boldsymbol{B} + O_p(n^{-2\alpha})$.
- $\widehat{V}(\hat{\boldsymbol{t}}_{Cx})$ needs to be specified also from the outside source.
- Taylor Linearization method is likely to be faster than Jackknife methods.

Multivariate normal jackknife method (MVNJ)

- Dever and Valliant, 2010, *Survey Methodology* first used this method in variance estimation of poststratification to populatoin control totals. It is a delete-one jackknife method.
- The replicates $\hat{\boldsymbol{t}}_{Cx(j)} = \hat{\boldsymbol{t}}_{Cx} + c_n\ \hat{\epsilon}_{(j)} \sqrt{1/(n-1)}$ where $\hat{\epsilon}_{(j)} \overset{i.i.d}{\sim} MVN(0, \widehat{V}(\hat{\boldsymbol{t}}_{Cx}))\ \ (j = 1, 2, ..., n)$ and $c_n = \sqrt{1/(n-1)}$.
- The replicates of $\hat{t}_{yL}$: $\hat{t}_{yL(j)} = \hat{t}_{Ay(j)} + (\hat{\boldsymbol{t}}_{Cx(j)} - \hat{\boldsymbol{t}}_{Ax(j)})'\hat{\boldsymbol{B}}_{s(j)}$
- The MVNJ variance estimator is:

$$\widehat{V}_{MVNJ}(\hat{t}_{yW}) = \frac{n-1}{n} \sum_{j=1}^{n} (\hat{t}_{yL(j)} - \hat{t}_{yL})^2.$$

- $E_p\{\widehat{V}_{MVNJ}\} = E_p[\frac{n-1}{n} \sum_{j=1}^{n} (\hat{t}^{*}_{yL(j)} - \hat{t}_{yL})^2] + \boldsymbol{B}' V(\hat{\boldsymbol{t}}_{Cx})\boldsymbol{B} + O_p(n^{-2\alpha})$

Fuller two-phase jackknife after Taylor linear. (F2TL)

- $\hat{t}_{yL} = \sum_s d_k a_{ks} E_k + \hat{\boldsymbol{t}}'_{Cx}\boldsymbol{B}$ and $\hat{\theta} - \hat{\boldsymbol{t}}'_{Cx}\boldsymbol{B} = O_p(n^{-1})$, where $\hat{\theta} = \hat{\boldsymbol{t}}'_{Cx}\hat{\boldsymbol{B}}_s$.
- Fuller jackknife method is used to estimate $V(\hat{\theta})$.
- Let $\widehat{V}(\hat{\boldsymbol{t}}_{Cx})$ be $m \times m$ matrix, and λ_1, λ_2,...,λ_m be its eigenvalues with $\boldsymbol{q}_1$, $\boldsymbol{q}_2$,...,$\boldsymbol{q}_m$ their corresponding eigenvectors.
- The replicates are: $\hat{\boldsymbol{t}}_{Cx(j)} = \hat{\boldsymbol{t}}_{Cx} + c_m\, \lambda_j^{1/2}\boldsymbol{q}_j$ and $\hat{\theta}_{(j)} = \hat{\boldsymbol{t}}'_{Cx(j)}\hat{\boldsymbol{B}}_{s(j)}$, where $c_m = (m-1)^{-1/2}m^{1/2}$.
- Then the F2TL variance estimator is:

$$\widehat{V}_{F2TL}(\hat{t}_{yW}) = \widehat{V}_{naive}(\hat{t}_{yW}) + \frac{m-1}{m}\sum_{j=1}^{m}(\hat{\theta}_{(j)} - \hat{\theta})^2.$$

- $E_p(\widehat{V}_{F2TL})$ is equal to $E_p(\widehat{V}_{naive}) + E_p[\frac{m-1}{m}\sum_{j=1}^{n}(\hat{\theta}^*_{(j)} - \hat{\theta})^2] + \boldsymbol{B}'V(\hat{\boldsymbol{t}}_{Cx})\boldsymbol{B} + O_p(n^{-2\alpha})$.

Simulation procedure

1. SDR 2010 is chosen as our populatoin. The population contains 27297 individuals, which were divided by us into 54 clusters.
2. At first stage np=30 PSU's (U_1, U_2,...,U_{30}) are sampled without replacement out of 54 clusters.
3. Secondly, from within each PSU sampled, we selected 50 individuals using simple random sampling without replacement.
4. Choose salary as the parameter we want to estimate and then choose m=20 auxiliary variables ($X_1, X_2, ..., X_{20}$).
5. For each X_i, calculate the PSU totals for each sampled PSU: ($\hat{t}_{i1}$, $\hat{t}_{i2}$,..., $\hat{t}_{i30}$)
6. Estimate population totals of X_i using PSU totals, then consider it as estimated control totals.
7. Calculate the calibrated estimator and its variance estimation with four methods mentioned above.

Simulation Result

Variance estimator	RBVE (%)	Cover (%)	MeanSE	StdSE
	Simulation 1. np=20, linear calibration.			
HT	10.40	94.5	277,002	23,373
Naive	-96.31	30.4	49,782	10,246
TL	5.88	95.4	271,200	23,739
MVNJ	9.71	93.7	272,788	48,803
F2TL	6.60	96.0	271,668	28,489
	Simulation 2. np=30, raking.			
HT	9.56	95.2	190,940	10,888
Naive	-95.68	31.5	37,586	5,388
TL	4.44	94.1	186,402	11,032
MVNJ	7.46	94.4	187,470	27,049
F2TL	4.92	94.4	186,537	15,280

Summary

- The CEEC is certainly a reasonable estimate when the population control totals are unknown. In simulation 2, the estimates are close to the true values.
- Overall, all the improved variance estimators we give are acceptable in simulation 2. They certainly mitigate the bias of the naive estimator.

Future plans

-

Thanks!

Variance Estimation for Calibration to Estimated Control Totals

Siyu Qing
The George Washington University
Department of Statistics