BRR Estimation of Variance of Survey Estimates Weight-adjusted for Nonresponse

Eric Slud[1,2] and Yves Thibaudeau[1]

[1]Stat. Res. Div., Census Bureau [2]Math. Dept., Univ. of MD

Objective: to evaluate theoretically the bias of Balanced Replication Variance estimates of survey-weighted nonresponse-adjusted totals with misspecified nonresponse adjustment cells.

Method: large-sample formulas under superpopulation quasi-randomization model (Oh & Scheuren 1983) and reasonable assumptions on attributes and split-PSU intersections with true and working adjustment cells.

Rationale

Large complex surveys generally involve

- nonresponse adjustments, based on adjustment cells, using ratio, raking, or calibration estimators
- difficulty in specifying joint inclusion probabilities to obtain variances of survey weighted estimators
- replication-based variance estimators

Justification of BRR (e.g. Krewski-Rao 1981) generally given for full response, not *misspecified* nonresponse adjustment.

Nonresp. adjustment bias treated by Särndal & Lündstrom 2005.

Effect of erroneous adjustment on BRR not treated before.

Framework & Notation

Large frame $\mathcal{U}$, size N, (balanced) split-PSU's $\mathcal{U}_{kH}$, $H = 1, 2$

Adjustment cells C_m , $m = 1, \ldots, M$, partition $\mathcal{U}$

Stratified Simple Random Sample $\mathcal{S} = \cup_{k,H} \mathcal{S}_{kH}$

— attributes y_i, single & joint inclusion probabilities π_i , π_{ij}

— sampling fraction f **small**, same in all PSU's; $n = fN$ **large**

r_i the $\{0,1\}$ valued indicator of unit i response
assumed random, independent : $\phi_i = 1/E(r_i)$

Assume $1/\phi_i = \rho_l$ when $l = l(i) \Leftrightarrow i \in B_l$ **true response cells**

Partitions $\mathcal{U} = B_1 \cup B_2 \cup \cdots \cup B_L = C_1 \cup C_2 \cup \cdots \cup C_M$.

Estimator $\hat{Y} \equiv \sum_{m=1}^{M} \sum_{\mathcal{S} \cap C_m} \hat{c}_m \frac{r_i}{\pi_i} y_i$, Adjustmt $\hat{c}_m = \frac{\sum_{\mathcal{S} \cap C_m} \pi_i^{-1}}{\sum_{\mathcal{S} \cap C_m} r_i \pi_i^{-1}}$

Ratio & Regression Estimators

Calibration and regression estimators for the predictor variables

$$\mathbf{x}_i \;=\; (\, I_{[i \in C_1]}\,, I_{[i \in C_2]}\,, \ldots, I_{[i \in C_M]}\,)$$

Denote $\quad m(i) = m \iff i \in C_m.$

Regression $\qquad \widehat{\beta}_m \;\equiv\; \sum_{i \in \mathcal{S} \cap C_m} \frac{r_i\, y_i}{\pi_i} \Big/ \sum_{i \in \mathcal{S} \cap C_m} \frac{r_i}{\pi_i}$

Residuals $\qquad \widehat{e}_i \;\equiv\; y_i - \widehat{\beta}_{m(i)}$

Estimator $\tilde{\phi}_i$ of $\phi_i = 1/E(r_i)$ can be

- $\widehat{c}_{m(i)}$ based on cells C_m or
- based on detailed (e.g., *logistic regression*) model with demographic/geographic covariates.

BRR Variance Estimator

Let $t = 1, \ldots, R$ index **replicate factors** $(f_{it},\ i \in \mathcal{U})$.

$$f_{it} \; = \; 1 \; + \; 0.5\,(-1)^H\, a_{kt} \quad \text{if} \quad i \in \mathcal{U}_{kH} \;\; , \quad a_{kt} \; = \; \pm 1$$

$$\textstyle\sum_{t=1}^R a_{kt} \; = \; R \;\; , \quad \sum_{t=1}^R a_{kt}\, a_{k't} \; = \; 0 \qquad \text{if} \quad k \neq k'$$

Replicate Adjustment Factor: $\quad \hat{c}_m^{(t)} \; = \; \dfrac{\sum_{i \in \mathcal{S} \cap C_m} (f_{it}/\pi_i)}{\sum_{i \in \mathcal{S} \cap C_m} (f_{it}\, r_i/\pi_i)}$

Replicate Survey Estimator: $\quad \hat{Y}^{(t)} \; = \; \sum_m \sum_{\mathcal{S} \cap C_m} \dfrac{f_{it}\, r_i}{\pi_i}\, \hat{c}_m^{(t)}\, y_i$

BRR Estimator of $V(\hat{Y})$: $\quad \hat{V}_{\mathrm{BRR}} \; = \; 4\, R^{-1} \sum_{t=1}^R (\hat{Y}^{(t)} \, - \, \hat{Y})^2$

$$\approx \; f^{-2} \sum_k \Big[\sum_{i \in \mathcal{S}_{k,1}} (\hat{\beta}_{m(i)} + r_i\, \hat{c}_{m(i)}\, \hat{e}_i) \; - \; \sum_{i \in \mathcal{S}_{k,2}} (\hat{\beta}_{m(i)} + r_i\, \hat{c}_{m(i)}\, \hat{e}_i) \Big]^2$$

Inclusion Prob Variance Estimators

Särndal-Lündstrom (2005) approximate formula (based on linearization & approx. correct adjustment)

$$\hat{V}_{LS} \;=\; \sum_{i,j \in \mathcal{S}} \left(\frac{\pi_{ij}}{\pi_i\,\pi_j} \;-\; 1\right) \frac{y_i\,y_j}{\pi_{ij}} + \sum_m \sum_{i \in \mathcal{S} \cap C_m} (\widehat{c}_m - 1)\,\frac{\widehat{e}_i^2}{\pi_i^2}$$

Could also replace $\widehat{c}_{m(i)}$ by $\tilde{\phi}_i$: if that is available a more accurate linearization formula is

$$\hat{V}(\hat{Y}) \;=\; \sum_{m=1}^{M} \sum_{i \in \mathcal{S} \cap C_m} \pi_i^{-2}\,\widehat{c}_m^2\,(\widehat{e}_i/\tilde{\phi}_i)^2\,(\tilde{\phi}_i - 1)$$

$$+ \;\sum_{i,j \in \mathcal{S}} \left(\frac{\pi_{ij}}{\pi_i\,\pi_j} \;-\; 1\right) (\pi_{ij})^{-1}\,\left(\widehat{\beta}_{m(i)} + \frac{\widehat{c}_{m(i)}}{\tilde{\phi}_i}\,\widehat{e}_i\right)\left(\widehat{\beta}_{m(j)} + \frac{\widehat{c}_{m(j)}}{\tilde{\phi}_j}\,\widehat{e}_j\right)$$

Superpopulation Framework

- r_i assumed independent $\text{Binom}(1, \rho_{l(i)})$, $l(i) = l \Leftrightarrow i \in B_l$.
- y_i assumed independent $\sim (\mu_k, \sigma^2)$ for $i \in \mathcal{U}_{kH}$ (with unif bounded third absolute moments)
- True response cells B_l, adjustment cells C_m, half-PSU's $\mathcal{U}_{kH}$ have limiting intersections

$$N^{-1} \,\#(\mathcal{U}_{kH} \cap B_l \cap C_m) \approx \nu(l, m, k, H)$$

joint prob. mass function on $(1 : L)\times(1 : M)\times(1 : K)\times(1 : 2)$

Problem: to Compare $\hat{V}(\hat{Y})$, $\hat{V}_{LS}$, $E(\hat{V}_{\mathbf{BRR}})$

— In our setting, $f\,\hat{V}(\hat{Y})/N$, $f\,\hat{V}_{LS}/N$ have limits.
— $\hat{V}_{\text{BRR}}$ consistent when $L = M$, $B_m = C_m$.
— in general $f\,\hat{V}_{\text{BRR}}/N \not\to$; examine only $(f/N)\,E(\hat{V}_{\text{BRR}})$.

Limiting Parameter Values

Approx. distribution of cells $B_l \cap C_m$ and half-PSU for randomly chosen $i \in \mathcal{U}$ makes (l, m, k, H) jointly ν-distributed.

$$\widehat{c}_m \;\longrightarrow\; c_m \;\equiv\; 1/E_\nu(\rho_l \,|\, m)$$

$$\widehat{\beta}_m \;\longrightarrow\; \beta_m^0 \;\equiv\; E_\nu(\rho_l \, \mu_k \,|\, m)/E_\nu(\rho_l \,|\, m)$$

Limits for Inclusion-Prob Var Estimators

$$f \, \widehat{V}_{LS}/N \;\longrightarrow\; \sum_{l,m,k,H} \{\sigma^2 \, c_m + (c_m - 1)\,(\mu_k - \beta_m^0)^2\} \, \nu(l, m, k, H)$$

$$\lim_N \text{Bias}(\widehat{Y}/N) \;\longrightarrow\; \sum_{l,m,k,H} (\beta_m^0 - \mu_k) \, \nu(l, m, k, H)$$

Limits $f \, \widehat{V}(\widehat{Y})/N$, $f \, E(\widehat{V}_{\text{BRR}})/N$ more complicated.

Two Special Cases related to Cell Intersections and PSU's

(A) For all $k, l, m, \quad \nu(l, m, k, 1) = \nu(l, m, k, 2)$.
Says Half-PSU's are perfectly asymptotically balanced across all intersections of PSU's, true and adjustment cells.

(B) For all $k, l, m, H, \quad \nu(l|m) = \nu(l|m, k, H)$.

True cell label conditionally indep. of half-PSU given adj. cell.

Proposition. In the superpopulation setting above,

Under **(A)**, $\quad (f/N)\,(E(\hat{V}_{\text{BRR}}) - \hat{V}(\hat{Y})) \to 0$.

Under **(B)**: $(f/N)\,(\hat{V}(\hat{Y}) - \hat{V}_{LS}) \to 0$ and $\text{Bias}(\hat{Y}/N) \to 0$;

also $\max_k \frac{1}{N}|\#\mathcal{U}_{k1} - \#\mathcal{U}_{k2}| \to 0 \Rightarrow \frac{f}{N}(E(\hat{V}_{\text{BRR}}) - \hat{V}(\hat{Y})) \to 0$.

When half-PSU H is chosen 'randomly' for each i (regardless of k, l, m), then BRR is large-sample unbiased.

Computational Examples

Numerical examples with $\nu(l, m, k, H)$ arrays defined to satisfy **(A)** and nearly **(B)**, then violate **(A)** more and more strongly.

Data on Four $\nu(\cdot)$ Arrays, $L = M = 10$, $K = 5$

Examp	avrsp	missp	SDcond	bias
1	.800	.159	.0039	.001
2	.800	.116	.0025	.001
3	.800	.121	.0080	.002
4	.800	.069	.0040	.001

`avrsp` = Average response $E_\nu(\rho_l)$

`missp` = *Misspecification of cells* $\text{Var}_\nu^{1/2}(\rho_l c_m)$

`SDcond` = average over (k, H) of SD$(\nu(l|m, k, H))$ (*measures violation of* **(B)**)

`bias` = bias of $\hat{Y}/N$, for $\underline{\mu} = (\frac{3}{4}, \frac{7}{8}, 1, \frac{9}{8}, \frac{5}{4})$.

Comparison of Large-Sample Variances in Examples

Parameter ω measures **imbalance**: $\nu(H|l,m,k) = \frac{1}{2}(1 \pm \omega)$
with random signs $\pm$ applied independently for each (k,l,m)

Table of $V \cdot f/N$ **Values, where** $\sigma^2 = 0.2$, $n = fN = 5000$

Examp	SDcond	ω	V_{SL}	V_{tru}	V_{brr}
1	.0039	0	.258	.258	.258
		0.10	.258	.258	.276
2	.0025	0	.262	.262	.262
		0.10	.262	.262	.296
3	.0080	0	.285	.291	.285
		0.05	.285	.291	.297
		0.10	.285	.291	.411
4	.0040	0	.264	.265	.264
		0.01	.264	.265	.294
		0.05	.264	.265	.311

Illustration with SIPP 1996

Survey of Income & Program Participation self representing strata (approx. 60% of sample in 1996 panel) had split-PSU design.

2 PSU's sampled for each non-SR stratum, then split. Systematic sample within PSU, by HU; split by alternate index.

Variances for weighted survey estimators calculated via BRR (**VPLX**). **Inclusion probabilities unrealistic:** systematic sampling & Wave 1 nonresponse adjustment.

Next compare BRR (VPLX) variances vs. `ppswr` inclusion prob. formulas, at both person & HH level, for SR strata wave 1 totals.

Item	π-Est	VPLX.SD	V_{LS}	PPSWR	HH.PPS
Foodst	15378514	481500	216117	217054	390471
SocSec	20572397	300225	262270	261587	279827
UnEmp	3789512	126464	127137	118941	136608
DIV	10878183	206557	198058	191773	204829

Summary & Conclusions

BRR bias for complex surveys under misspecified response models studied theoretically, showing for large survey-samples:

(1) for half-PSU index $\mathtt{H}$ closely balanced across cells intersected with PSU's, BRR variance estimator is remarkably **un**biased.

(2) **imbalances** of a few percent (independently over cell intersections with PSU's) **can inflate BRR variance from a few percent to a lot** (40-50% or greater), depending on misspecification and PSU & cell intersection patterns.

Caveats: the superpopulation model here oversimplifies:

- independent responses likelier for HH than person units.
- attributes homoscedastic with means allowed to depend on PSU but not on true response or adjustment cells.

References

1. Fay, R. (1984) ASA, SRMS Proc. pp. 495-500.
2. Fay, R. (1989) ASA, SRMS Proc. pp. 212-217.
3. Kish, L. and Frankel, M. (1970) JASA.
4. Kim, Jae and Kim, Jay (2007) Canadian Jour. Stat. **35**, pp. 501-514.
5. Krewski, D. and Rao, J.N.K. (1981) Ann. Statist.
6. Oh, H. and Scheuren, F. (1983) paper in: *Incomplete Data in Sample Surveys*, vol. 2, 143-184.
7. Särndal, C.-E. and Lündstrom, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley.
8. Slud, E. and Bailey, L. (2007) FCSM