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

Eric Slud & Yves Thibaudeau, `eric.v.slud@census.gov`

Objective: understand bias of Balanced Repeated Replication Variance of survey-weighted nonresponse-adjusted estimates with misspecified nonresponse adjustments.

Method: linearized large-sample formulas and simulation under superpopulation model with reasonable assumptions on attributes, split-PSU's, and pattern of response probabilities.

Rationale

Large complex surveys generally involve

- nonresponse adjustments, based on adjustment cells (using ratio adjustment, raking or calibration)
- difficulty in specifying joint inclusion probabilities adjusted for nonresponse
- replication-based variance estimators

Justifications of BRR (e.g. Krewski-Rao 1981) for complete response, not *misspecified* nonresponse adjustment.

Nonresp. adjustment bias treated by Särndal & Lündstrom 2005.

Effect of erroneous adjustment on BRR was 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 random response indicator of unit i

assumed independent with : $E(r_i) = 1/\phi_i = \rho_l$ when $i \in B_l$

$$\mathcal{U} = \underbrace{B_1 \cup B_2 \cup \cdots \cup B_L}^{\texttt{true resp. cells}} = \overbrace{C_1 \cup C_2 \cup \cdots \cup C_M}^{\texttt{working cells}}$$

Survey Weighted Total Estimator

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

is also `regression estimator` with predictors

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

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

Residuals $\quad \widehat{e}_i \equiv y_i - \widehat{\beta}_m$ for $i \in C_m$

Could replace factors $\widehat{c}_m$ by $\tilde{\phi}_i = 1/(\texttt{predictors})$ from *logistic regression* model.

(Fay-Method) BRR Variance Estimator

Replicate factors $f_{it} = .5, 1.5$ indexed by $t = 1 \ldots R,\ i \in \mathcal{U}$

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

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

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

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

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

Inclusion Prob Variance Estimators

Särndal-Lündstrom (2005) approximate formula

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

With $\widehat{c}_m$ replaced for $i \in C_m$ by $\tilde{\phi}_i$: we have a more accurate new linearization formula

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

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

Superpopulation Framework

- r_i assumed indep. Binom$(1, \rho_l)$, $i \in B_l$
- y_i assumed indep. $\sim (\mu_k, \sigma^2)$ for $i \in \mathcal{U}_{kH}$
- True resp. cells B_l, working cells C_m, $\frac{1}{2}$-PSU's $\mathcal{U}_{kH}$ have limiting intersection proportions

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

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

- As $N \to \infty$, $f \, \widehat{V}(\widehat{Y})/N$ and $f \, \widehat{V}_{SL}/N$ have limits.
- With K finite: $\frac{f}{N} \, \widehat{V}_{\text{BRR}} \not\to$; examine only $\frac{f}{N} \, E(\widehat{V}_{\text{BRR}})$.

Limiting Parameter Values

Half-PSU and cell indices (l, m, k, H) approx. $\nu(\cdot)$-distributed for $i \in B_l \cap C_m \cap \mathcal{U}_{kH}$ for randomly chosen in $\mathcal{U}$.

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

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

Limits for Bias & Variance Expressions

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

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

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

Properties of Cell Intersections & PSU's

(A) For all k, l, m, $\quad \nu(l,m,k,1) = \nu(l,m,k,2)$.

Half-PSU's perfectly asymptotically balanced across 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 conditionally indep. of half-PSU given adj. cell.

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

Under **(B)**: $\frac{f}{N}\,(\hat{V}(\hat{Y}) - \hat{V}_{SL}) \rightarrow 0$ and $\text{Bias}(\hat{Y}/N) \rightarrow 0$,

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

If H is chosen randomly, independently for each i then BRR is large-sample unbiased.

Computations & Simulations: Design

$L = M = 10,\ K = 20$, 5 distinct PSU's in blocks of 4 each

PSU attrib. means $\mu_k = 1.5 \ldots 2.5$, $\sigma = .8$

Response probabilities ρ_l spaced $0.6 \ldots 1.0$, avg. $= 0.8$

Example $\nu(l,\, m,\, k,\, H)$ **Arrays,** **quantified by:**

`missp` = *Misspecification of cells* $\mathrm{Var}_\nu^{1/2}(\rho_l c_m)$, .07 to .16

`SDcond` = average over (l, m) of $\mathrm{SD}(\{\nu(l|m, k, H)\}_{k,H})$

(*measures violation of* **(B)**) , ranging 0 to .01

`imbalance` parameter $\omega = 0,\ 0.1$, $\nu(H|l, m, k) = \frac{1}{2}\,(1 \pm \omega)$

random signs $\pm$ indep. for all (k, l, m)

Table of $V\,n/\,N^2$ **Values, where** $n = 4000$, $\omega = 0.1$

Simulations done with 1000 iterations.

	Theoretical		Simulated		Simulated	
Examp	VY	Vbrr	VY.mean	VB.mean	VY.sd	VB.sd
a	.832	0.864	0.832	0.863	.047	.282
b	.841	0.917	0.839	0.934	.049	.312
c	.851	1.023	0.850	1.034	.050	.325

NOTES. (1) Linearized approximation used for BRR, has relative error in range $(-.001, 0)$.

(2) Simulations corroborate formulas. BRR more biased and has larger SE when PSU's are fewer .

BRR vs Incl Prob SE's in SIPP 1996

In *Survey of Income & Program Participation* 1996 panel, self representing strata (60% of sample) had split-PSU design. Systematic sample within PSU, by HU; split by alternate index.

Survey uses BRR: **inclusion probabilities thought unrealistic** due to systematic sampling & Wave 1 nonresponse adjustment.

Table: SD's for SIPP 1996 *SR strata* Wave 1 totals, estimated from BRR vs. Household `ppswr` incl. prob.'s.

Item	Total/10^7	HHpps.SE	BRR.SE
Foodst	1.538	390471	481500
SocSec	2.057	279827	300225
UnEmp	0.379	136608	126464
Divorce	1.088	204829	206557

Summary & Conclusions

Studied BRR bias for complex surveys under misspecified response models, showing for large samples:

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

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

(3) More strata/PSU's, less bias in BRR variances.

Caveat: superpopulation model oversimplifies attributes by PSU.

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. Krewski, D. and Rao, J.N.K. (1981) Ann. Statist.
5. Oh, H. and Scheuren, F. (1983) paper in: *Incomplete Data in Sample Surveys*, vol. 2, 143-184.
6. Särndal, C.-E. and Lündstrom, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley.
7. Slud, E. and Bailey, L. (2007) FCSM