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

Eric Slud\textsuperscript{1,2} and Yves Thibaudeau\textsuperscript{1}
\textsuperscript{1}Stat. Res. Div., Census Bureau \quad \textsuperscript{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.

USCENSUSBUREAU
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 $\mathcal{S} = \bigcup_{k,H} \mathcal{S}_{kH}$
— attributes $y_i$, single & joint inclusion probabilities $\pi_i$, $\pi_{ij}$
— 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 $\hat{Y} \equiv \sum_{m=1}^M \sum_{S \cap C_m} \hat{c}_m \frac{r_i}{\pi_i} y_i$, Adjustmt $\hat{c}_m = \frac{\sum_{S \cap C_m} \pi_i^{-1}}{\sum_{S \cap C_m} r_i \pi_i^{-1}}$.
Ratio & Regression Estimators

Calibration and regression estimators for the predictor variables

\[ x_i = (I_{i \in C_1}, I_{i \in C_2}, \ldots, I_{i \in C_M}) \]

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

Regression

\[ \hat{\beta}_m \equiv \frac{\sum_{i \in S \cap C_m} r_i y_i}{\sum_{i \in S \cap C_m} r_i} / \frac{\sum_{i \in S \cap C_m} \pi_i}{\sum_{i \in S \cap C_m} \pi_i} \]

Residuals

\[ \hat{e}_i \equiv y_i - \hat{\beta}_{m(i)} \]

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

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

USCENSUSBUREAU
BRR Variance Estimator

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

\[
f_{it} = 1 + 0.5 (-1)^H a_{kt} \quad \text{if} \quad i \in U_{kH} , \quad a_{kt} = \pm 1
\]

\[
\sum_{t=1}^{R} a_{kt} = R , \quad \sum_{t=1}^{R} a_{kt} a_{k't} = 0 \quad \text{if} \quad k \neq k'
\]

Replicate Adjustment Factor: \( \hat{c}_{m}^{(t)} = \frac{\sum_{i \in S \cap C_m} (f_{it}/\pi_i)}{\sum_{i \in S \cap C_m} (f_{it} r_i/\pi_i)} \)

Replicate Survey Estimator: \( \hat{Y}(t) = \sum_{m} \sum_{S \cap C_m} \frac{f_{it}r_i}{\pi_i} \hat{c}_{m}^{(t)} y_i \)

**BRR Estimator of** \( V(\hat{Y}) \): \( \hat{V}_{BRR} = 4 R^{-1} \sum_{t=1}^{R} (\hat{Y}(t) - \bar{Y})^2 \)

\[
\approx f^{-2} \sum_{k} \left[ \sum_{i \in S_{k,1}} (\hat{\beta}_m(i) + r_i \hat{c}_{m}(i) \hat{e}_i) - \sum_{i \in S_{k,2}} (\hat{\beta}_m(i) + r_i \hat{c}_{m}(i) \hat{e}_i) \right]^2
\]
Inclusion Prob Variance Estimators

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

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

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

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

\[
+ \sum_{i,j \in S} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \left( \pi_{ij} \right)^{-1} \left( \hat{\beta}_{m(i)} + \frac{\hat{c}_{m(i)}}{\tilde{\phi}_i} \hat{e}_i \right) \left( \hat{\beta}_{m(j)} + \frac{\hat{c}_{m(j)}}{\tilde{\phi}_j} \hat{e}_j \right)
\]
Superpopulation Framework

• $r_i$ assumed independent $\text{Binom}(1, \rho_{l(i)})$, $l(i) = l \iff i \in B_l$.

• $y_i$ 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 $\hat{V}(\hat{Y}), \hat{V}_{LS}, E(\hat{V}_{BRR})$

— In our setting, $f \hat{V}(\hat{Y})/N, f \hat{V}_{LS}/N$ have limits.

— $\hat{V}_{BRR}$ consistent when $L = M$, $B_m = C_m$.

— in general $f \hat{V}_{BRR}/N \not\rightarrow$; examine only $(f/N) E(\hat{V}_{BRR})$. 

USCENSUSBUREAU
**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 $\nu$-distributed.

$$ \hat{c}_m \to c_m \equiv 1/E_\nu(\rho_l | m) $$

$$ \hat{\beta}_m \to \beta_m^0 \equiv E_\nu(\rho_l \mu_k | m)/E_\nu(\rho_l | m) $$

**Limits for Inclusion-Prob Var Estimators**

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

$$ \lim_{N} \text{Bias}(\hat{Y}/N) \to \sum_{l,m,k,H} (\beta_m^0 - \mu_k) \nu(l, m, k, H) $$

Limits $f \hat{V}(\hat{Y})/N$, $f E(\hat{V}_{BRR})/N$ more complicated.
Two Special Cases related to Cell Intersections and PSU’s

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

Says Half-PSU’s are perfectly asymptotically balanced across all intersections of PSU’s, true and adjustment cells.

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

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

Proposition. In the superpopulation setting above,

Under (A), \( \left( \frac{f}{N} \right) (E(\hat{V}_{BRR}) - \hat{V}(\hat{Y})) \to 0 \).

Under (B): \( \left( \frac{f}{N} \right) (\hat{V}(\hat{Y}) - \hat{V}_{LS}) \to 0 \) and \( \text{Bias}(\hat{Y}/N) \to 0 \);

also \( \max_k \frac{1}{N}|\#U_{k1} - \#U_{k2}| \to 0 \Rightarrow \left( \frac{f}{N} \right) (E(\hat{V}_{BRR}) - \hat{V}(\hat{Y})) \to 0 \).

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

USCENUSBUREAU
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, $L = M = 10$, $K = 5$

<table>
<thead>
<tr>
<th>Examp</th>
<th>avrsp</th>
<th>missp</th>
<th>SDcond</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.800</td>
<td>.159</td>
<td>.0039</td>
<td>.001</td>
</tr>
<tr>
<td>2</td>
<td>.800</td>
<td>.116</td>
<td>.0025</td>
<td>.001</td>
</tr>
<tr>
<td>3</td>
<td>.800</td>
<td>.121</td>
<td>.0080</td>
<td>.002</td>
</tr>
<tr>
<td>4</td>
<td>.800</td>
<td>.069</td>
<td>.0040</td>
<td>.001</td>
</tr>
</tbody>
</table>

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 $\text{SD}(\nu(l|m,k,H))$ (measures violation of (B))

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

USCENSUSBUREAU
Comparison of Large-Sample Variances in Examples

Parameter \( \omega \) 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 \( \sigma^2 = 0.2, n = fN = 5000 \)

<table>
<thead>
<tr>
<th>Examp</th>
<th>SDcond</th>
<th>( \omega )</th>
<th>( V_{SL} )</th>
<th>( V_{tru} )</th>
<th>( V_{brr} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.0039</td>
<td>0</td>
<td>.258</td>
<td>.258</td>
<td>.258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>.258</td>
<td>.258</td>
<td>.276</td>
</tr>
<tr>
<td>2</td>
<td>.0025</td>
<td>0</td>
<td>.262</td>
<td>.262</td>
<td>.262</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>.262</td>
<td>.262</td>
<td>.296</td>
</tr>
<tr>
<td>3</td>
<td>.0080</td>
<td>0</td>
<td>.285</td>
<td>.291</td>
<td>.285</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.05</td>
<td>.285</td>
<td>.291</td>
<td>.297</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>.285</td>
<td>.291</td>
<td>.411</td>
</tr>
<tr>
<td>4</td>
<td>.0040</td>
<td>0</td>
<td>.264</td>
<td>.265</td>
<td>.264</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>.264</td>
<td>.265</td>
<td>.294</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.05</td>
<td>.264</td>
<td>.265</td>
<td>.311</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Item</th>
<th>π-Est</th>
<th>VPLX.SD</th>
<th>V_LS</th>
<th>PPSWR</th>
<th>HH.PPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foodst</td>
<td>15378514</td>
<td>481500</td>
<td>216117</td>
<td>217054</td>
<td>390471</td>
</tr>
<tr>
<td>SocSec</td>
<td>20572397</td>
<td>300225</td>
<td>262270</td>
<td>261587</td>
<td>279827</td>
</tr>
<tr>
<td>UnEmp</td>
<td>3789512</td>
<td>126464</td>
<td>127137</td>
<td>118941</td>
<td>136608</td>
</tr>
<tr>
<td>DIV</td>
<td>10878183</td>
<td>206557</td>
<td>198058</td>
<td>191773</td>
<td>204829</td>
</tr>
</tbody>
</table>
Summary & Conclusions

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

(1) for half-PSU index $H$ closely balanced across cells intersected with PSU’s, BRR variance estimator is remarkably unbiased.

(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