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 & Lünstrom 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 $S = \bigcup_{k,H} S_{kH}$
— 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 random response indicator of unit $i$
assumed independent with: $E(r_i) = 1/\phi_i = \rho_l$ when $i \in B_l$

true resp. cells working cells
$\mathcal{U} = B_1 \cup B_2 \cup \cdots \cup B_L = C_1 \cup C_2 \cup \cdots \cup C_M$
Survey Weighted Total Estimator

\[ \hat{Y} = \sum_{m=1}^{M} \sum_{S \cap C_m} \hat{c}_m \frac{r_i}{\pi_i} y_i, \quad \text{Adjustmt} \quad \hat{c}_m = \frac{\sum_{S \cap C_m} \pi_i^{-1}}{\sum_{S \cap C_m} r_i \pi_i^{-1}} \]

is also regression estimator with predictors

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

**Regression**

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

**Residuals**

\[ \hat{e}_i \equiv y_i - \hat{\beta}_m \quad \text{for} \quad i \in C_m \]

Could replace factors \( \hat{c}_m \) by \( \tilde{\phi}_i = 1/(\text{predictors}) \) from *logistic regression* model.
(Fay-Method) BRR Variance Estimator

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

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

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

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

Särndal-Lundstrom (2005) approximate formula

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

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

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

$$+ \sum_{i,j \in S} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \frac{1}{\pi_{ij}} \left( \beta_m(i) \frac{\tilde{c}_m(i) \hat{e}_i}{\tilde{\phi}_i} \right) \left( \beta_m(j) \frac{\tilde{c}_m(j) \hat{e}_j}{\tilde{\phi}_j} \right)$$
Superpopulation Framework

- $r_i$ assumed indep. $\text{Binom}(1, \rho_l)$, $i \in B_l$
- $y_i$ 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** $\hat{V}(\hat{Y}), \hat{V}_{SL}, E(\hat{V}_{BRR})$

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

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

\[
\hat{c}_m \to c_m \equiv 1/E\nu(\rho_l | m)
\]

\[
\hat{\beta}_m \to \beta^0_m \equiv E\nu(\rho_l \mu_k | m)/E\nu(\rho_l | m)
\]

Limits for Bias & Variance Expressions

\[
\frac{f}{N} \hat{V}_{SL} \to \sum_{l, m, k, H} \{\sigma^2 c_m + (c_m - 1) ( \mu_k - \beta^0_m)^2 \} \nu(l, m, k, H)
\]

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

Limits \(\frac{f}{N} \hat{V}(\hat{Y}), \frac{f}{N} E(\hat{V}_{BRR})\) more complicated.
Properties of Cell Intersections & PSU’s

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

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

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

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

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

Under (B): \( \frac{f}{N} (\hat{V}(\hat{Y}) - \hat{V}_{SL}) \rightarrow 0 \) and \( \text{Bias}(\hat{Y}/N) \rightarrow 0, \)
and \( \max_k \frac{1}{N}|\#U_{k1} - \#U_{k2}| \rightarrow 0 \Rightarrow \frac{f}{N} (E(\hat{V}_{BRR}) - \hat{V}(\hat{Y})) \rightarrow 0. \)

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 $\rho_l$ spaced $0.6 \ldots 1.0$, avg. = 0.8

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

$\text{missp} = \text{Misspecification of cells} \ Var_{\nu}^{1/2}(\rho_l c_m)$, .07 to .16

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

(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.

<table>
<thead>
<tr>
<th>Examp</th>
<th>VY</th>
<th>Vbrr</th>
<th>VY.mean</th>
<th>VB.mean</th>
<th>VY.sd</th>
<th>VB.sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.832</td>
<td>0.864</td>
<td>0.832</td>
<td>0.863</td>
<td>.047</td>
<td>.282</td>
</tr>
<tr>
<td>b</td>
<td>.841</td>
<td>0.917</td>
<td>0.839</td>
<td>0.934</td>
<td>.049</td>
<td>.312</td>
</tr>
<tr>
<td>c</td>
<td>.851</td>
<td>1.023</td>
<td>0.850</td>
<td>1.034</td>
<td>.050</td>
<td>.325</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Item</th>
<th>Total/10^7</th>
<th>HHpps.SE</th>
<th>BRR.SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foodst</td>
<td>1.538</td>
<td>390471</td>
<td>481500</td>
</tr>
<tr>
<td>SocSec</td>
<td>2.057</td>
<td>279827</td>
<td>300225</td>
</tr>
<tr>
<td>UnEmp</td>
<td>0.379</td>
<td>136608</td>
<td>126464</td>
</tr>
<tr>
<td>Divorce</td>
<td>1.088</td>
<td>204829</td>
<td>206557</td>
</tr>
</tbody>
</table>
Summary & Conclusions

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

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

(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


