Sparse and Efficient Replication Variance Estimation for Complex Surveys

Jae Kwang Kim\textsuperscript{1}

Iowa State University

April 4th, 2012

Westat

\textsuperscript{1}Joint work with Changbao Wu at University of Waterloo
1. Introduction

2. Fully Efficient Replication Weights

3. Sparse and Efficient Replication Weights

4. Balanced Sampling

5. Simulation Studies

6. Concluding Remarks
1 Introduction

2 Fully Efficient Replication Weights

3 Sparse and Efficient Replication Weights

4 Balanced Sampling

5 Simulation Studies

6 Concluding Remarks
Variance Estimation

- Construct confidence intervals
- Conduct hypothesis tests
- Provide descriptive measures on accuracies of estimates
- Measure the efficiency of the sampling design
Traditional Approaches

- Inclusion probabilities $\pi_i = P(i \in S)$ and $\pi_{ij} = P(i, j \in S)$
- Basic design weights $w_i = 1/\pi_i$
- The Horvitz-Thompson estimator of $Y = \sum_{i=1}^{N} y_i$

\[ \hat{Y} = \sum_{i \in S} w_i y_i \]  

(1)

- Variance estimation

\[ \hat{V} = \sum_{i \in S} \sum_{j \in S} \Omega_{ij} y_i y_j, \]  

(2)

where $\Omega_{ij} = (\pi_{ij} - \pi_i \pi_j)/(\pi_{ij} \pi_i \pi_j)$.

- Nonlinear statistics: linearization (first order Taylor series approximation)
Replication Methods

- Use jackknife, bootstrap, or BRR
- Enlarge the data set with additional columns of replication weights
- Work the same way for linear or nonlinear statistics
- Preserve confidentiality (by not releasing certain design information)
Issues with Replication Weights

- **Validity**
  - Asymptotic Unbiasedness
  - Jackknife method is valid for restricted scenarios
  - Bootstrap method needs to be modified to fit into special designs
  - No replication method is valid for arbitrary sampling designs

- **Efficiency**
  - Small variance
  - Replication variance estimators are often approximations to the fully efficient $\hat{V}$
  - Efficiency often depends on sample sizes, number of replication weights, and design features (such as sampling fractions)

- **Sparsity**
  - Want to reduce the size of replication.
  - The number of jackknife replication weights depends on the sample size
  - Bootstrap methods typically require at least 1000 sets of weights (StatCan uses 500!)
An Algebraic Construction of Replication Weights

Two research problems in this talk

1. How to construct a valid replication variance estimator?
2. How to reduce the replication size without sacrificing the efficiency (for some key items)?
1 Introduction

2 Fully Efficient Replication Weights

3 Sparse and Efficient Replication Weights

4 Balanced Sampling

5 Simulation Studies

6 Concluding Remarks
Basic Settings

- Consider an arbitrary sampling design with first and second order inclusion probabilities $\pi_i$ and $\pi_{ij}$ and basic design weights $w_i$.
- Develop variance estimator $v(\hat{Y})$ for $\hat{Y} = \sum_{i \in S} w_i y_i$.
- The analytic variance estimator

$$\hat{V} = \sum_{i \in S} \sum_{j \in S} \Omega_{ij} y_i y_j,$$

where $\Omega_{ij} = (\pi_{ij} - \pi_i \pi_j) / (\pi_{ij} \pi_i \pi_j)$, is viewed as fully efficient.
- Write $\hat{V} = y' \Delta y$, where $\Delta = (\Omega_{ij})$ is a $n \times n$ matrix and $y = (y_1, y_2, \cdots, y_n)'$
An Algebraic Construction of Replication Weights

\[ \mathbf{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})', \quad k = 1, \ldots, L \]

\[ \hat{Y}^{(k)} = \sum_{i \in S} w_i^{(k)} y_i \]

Replication variance estimator

\[ \hat{V}_R = \sum_{k=1}^{L} h_k \left( \hat{Y}^{(k)} - \hat{Y} \right)^2 \quad (3) \]

for some known constants \( h_k \)

Want to construct \( \mathbf{w}^{(k)} \) so that \( \hat{V}_R = \hat{V} \) for any \( y \)-variable.
Construction of Weights (Cont’d)

Approach 1: Fay’s Idea

- Decomposition of $\Delta$:

$$
\Delta = \sum_{k=1}^{p} \lambda_k \delta_k \delta'_k
$$

(4)

where $p = \text{rank}(\Delta) \leq n$

- $\hat{V} = \hat{V}_R$ for any $y$-variable if $L = p$ and

$$
\mathbf{w}^{(k)} = \mathbf{w} + \left(\frac{\lambda_k}{h_k}\right)^{1/2} \mathbf{\delta}_k,
$$

where $\mathbf{w} = (w_1, \ldots, w_n)'$ is the set of basic design weights.

- Algebraically equivalent because

$$
\begin{align*}
   h_k(\hat{Y}^{(k)} - \hat{Y})^2 &= \mathbf{y}'\{h_k(\mathbf{w}^{(k)} - \mathbf{w})(\mathbf{w}^{(k)} - \mathbf{w})'\}\mathbf{y} \\
   &= \mathbf{y}'\{\lambda_k \mathbf{\delta}_k \mathbf{\delta}'_k\}\mathbf{y}
\end{align*}
$$
Approach 2: Direct construction for some designs

- In many single-stage designs, we can write \( \hat{V} = y' \Delta y \) where

\[
\Delta = \Delta_0 - \Delta_0 Z (Z' \Delta_0 Z)^{-1} Z' \Delta_0
\]  

where \( \Delta_0 = \text{diag}\{\lambda_1, \cdots, \lambda_n\} \), \( \lambda_i \geq 0 \) for all \( i = 1, 2, \cdots, n \) and \( Z' = (z_1, \cdots, z_n) \) is a \( q \times n \) matrix with full row rank.

- After some matrix algebra, we can get

\[
y' \Delta y = (y - Z \hat{\beta})' \Delta_0 (y - Z \hat{\beta}) \\
= \sum_{k=1}^{n} \lambda_k (y_k - z_k' \hat{\beta})^2
\]

where \( \hat{\beta} = (Z' \Delta_0 Z)^{-1} Z' \Delta_0 y \).
Construction of Weights (Cont’d)

- Deville(1999)’s approximation for variance estimation under unequal probability of fixed-size sampling:

\[ \hat{V} = c \sum_{i \in S} (1 - \pi_i) \left( \frac{y_i}{\pi_i} - \tilde{Y} \right)^2 \]

where \( c = (1 - \sum_{i \in S} b_i^2)^{-1} \), \( b_i = (1 - \pi_i) / \sum_{k \in S} (1 - \pi_k) \) and \( \tilde{Y} = \sum_{k \in S} b_i (y_i / \pi_i) \).

- Deville’s approximation uses \( \lambda_i = c \pi_i^{-2} (1 - \pi_i) \) and \( z_i = \pi_i \).
Construction of Weights (Cont’d)

- Bredit and Chauvet (2011) approximation for balanced sampling with constraint $\sum_{i \in S} x_i / \pi_i = \sum_{i \in U} x_i$:

$$\hat{V} = \frac{n}{n - q} \sum_{i \in S} (1 - \pi_i) \pi_i^{-2} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = x_i' \hat{B}$ and

$$\hat{B} = \left\{ \sum_{k \in S} (1 - \pi_k) \pi_k^{-2} x_k x_k' \right\}^{-1} \sum_{k \in S} (1 - \pi_k) \pi_k^{-2} x_k y_i.$$

- Bredit and Chauvet (2011) approximation uses

$$\lambda_i = c \pi_i^{-2} (1 - \pi_i)$$

and $z_i = x_i$, where $c = n/(n - q)$. 
Jackknife for the designs with $\Delta$ in (5).

1. Recall that we can write $\hat{V} = \sum_{k=1}^{n} \lambda_k \left( y_k - z'_k \hat{\beta} \right)^2$

2. Thus, we want to construct $\hat{Y}^{(k)} = \sum_{i=1}^{n} w^{(k)}_i y_i$ such that

$$h_k \left( \hat{Y}^{(k)} - \hat{Y} \right)^2 = \lambda_k \left( y_k - z'_k \hat{\beta} \right)^2$$

which is achieved when

$$\sum_{i=1}^{n} w^{(k)}_i y_i = \sum_{i=1}^{n} w_i y_i - \left( y_k - z'_k \hat{\beta} \right) \left( \frac{\lambda_k}{h_k} \right)^{1/2}.$$

3. Jackknife-type replication weights

$$w^{(k)}_i = \begin{cases} w_i + \left( \frac{\lambda_k}{h_k} \right)^{1/2} \left\{ -1 + z'_k (Z' \Delta_0 Z)^{-1} z_i \lambda_i \right\} & \text{if } i = k \\ w_i + \left( \frac{\lambda_k}{h_k} \right)^{1/2} \left\{ z'_k (Z' \Delta_0 Z)^{-1} z_i \lambda_i \right\} & \text{otherwise.} \end{cases}$$
Delete-a-Group Jackknife construction

1. Partition $S$ into $L$ groups: $S = S_1 \cup S_2 \cup \cdots S_L$.

2. We can write

$$\hat{V} = \sum_{j \in S} \lambda_j \left( y_j - z'_j \hat{\beta} \right)^2 = \sum_{k=1}^{L} \sum_{j \in S_k} \lambda_j \left( y_j - z'_j \hat{\beta} \right)^2$$

3. Thus, we want to construct $\hat{Y}(k) = \sum_{i=1}^{n} w_i^{(k)} y_i$ such that

$$h_k \left( \hat{Y}(k) - \hat{Y} \right)^2 = \sum_{j \in S_k} \lambda_j \left( y_j - z'_j \hat{\beta} \right)^2$$

which is not always possible.

4. Instead, construct $\hat{Y}(k) = \sum_{i=1}^{n} w_i^{(k)} y_i$ such that

$$h_k \left( \hat{Y}(k) - \hat{Y} \right)^2 = \left\{ \sum_{j \in S_k} \lambda_j^{1/2} (2\delta_j^{(k)} - 1) \left( y_j - z'_j \hat{\beta} \right) \right\}^2$$

where $\delta_j^{(k)}$ are IID from Bernoulli(0.5).
Delete-a-Group Jackknife construction (Cont’d)

- Note that

\[
E^* \left[ \sum_{j \in S_k} \lambda_j^{1/2} (2\delta_j^{(k)} - 1)(y_j - z_j'\hat{\beta}) \right]^2 = \sum_{j \in S_k} \lambda_j (y_j - z_j'\hat{\beta})^2
\]

where the expectation is with respect to the distribution for \(\delta_i^{(k)}\).

- Delete-a-group Jackknife replication weights

\[
w_i^{(k)} = \begin{cases} 
    w_i + \left\{ -q_{ik} + \sum_{j \in S_k} q_{jk}z_j'(Z'\Delta_0 Z)^{-1} z_i\lambda_i \right\} & \text{if } i \in S_k \\
    w_i + \left\{ \sum_{j \in S_k} q_{jk}z_j'(Z'\Delta_0 Z)^{-1} z_i\lambda_i \right\} & \text{otherwise.}
\end{cases}
\]

where \(q_{jk} = (\lambda_j/h_k)^{1/2} (2\delta_j^{(k)} - 1)\).

- Thus, it is essentially the same as applying BRR within each random group.
1 Introduction

2 Fully Efficient Replication Weights

3 Sparse and Efficient Replication Weights

4 Balanced Sampling

5 Simulation Studies

6 Concluding Remarks
Efficiency and Sparsity

- An initial $L$ sets of replication weights $\mathbf{w}^{(k)}$
  - Weights from bootstrap, jackknife or BRR
  - Weights from the algebraic construction
- $\hat{V}_R$ is (assumed to be) fully efficient
- $L$ is very large
- **Goal:** Reduce $L$ to a much smaller number $L_0$ without major loss of efficiency
- **Strategy:** Random sampling (random grouping) plus weight calibration
Random Sampling of Weights

- View $\hat{V}_R = \sum_{k=1}^{L} c_k \left( \hat{Y}^{(k)} - \hat{Y} \right)^2$ as a population total
- Select $w^{(kj)}$, $j = 1, \cdots, L_0$ from $w^{(k)}$, $k = 1, \cdots, L$ by simple random sampling without replacement
- Use $\hat{V}_R^{(1)} = \sum_{j=1}^{L_0} h_{kj} (L/L_0) \left( \hat{Y}^{(kj)} - \hat{Y} \right)^2$
- $\hat{V}_R^{(1)}$ is still a valid variance estimator:
  $$E^*(\hat{V}_R^{(1)}) = \hat{V}_R$$
Random Sampling of Weights (Cont’d)

- Under the algebraic construction of replication weights:

\[ \hat{V}_R = \sum_{k=1}^{L} \lambda_k (\delta'_k y)^2 \]

- \( \hat{V}_R \) is a **population total** and \( \lambda_k \) can be viewed as a **size variable**
- Select \( w^{(k)} \) with inclusion probability \( \eta_k \) and use

\[ \hat{V}_R^{(2)} = \sum_{j=1}^{L_0} \frac{h_{kj}}{\eta_{kj}} \left( \hat{Y}^{(k_j)} - \hat{Y} \right)^2 \]

- \( \hat{V}_R^{(2)} \) is still valid and is more efficient than \( \hat{V}_R^{(1)} \)
Random Sampling of Weights (Cont’d)

- Random selection of $L_0$ sets of weights ensures validity of replication variance estimator using the smaller number sets of weights
- None of $\hat{V}_R^{(1)}$ or $\hat{V}_R^{(2)}$ is as efficient as $\hat{V}_R$
- Calibration over the selected smaller number sets of weights can improve efficiency
Weight Calibration

Step 1

- Collect all key variables \( z = (z_1, \cdots, z_m)' \) from the survey
- Compute the fully efficient variance-covariance matrix

\[
\hat{V}_1(\hat{Z}) = \sum_{k=1}^{L} h_{k1} (\hat{Z}^{(k)} - \hat{Z}) (\hat{Z}^{(k)} - \hat{Z})'
\]

using the initial large number \( L \) sets of replication weights, where

\[
\hat{Z}^{(k)} = \sum_{i \in S} w_i^{(k)} z_i
\]
Weight Calibration (Cont’d)

Step 2

- Construct an initial set of $L_0$ replication weights such that the resulting variance estimator is also asymptotically unbiased.
- Let $w_0^{(k)} = (w_{10}^{(k)}, \cdots, w_{n0}^{(k)})'$, $k = 1, \cdots, L_0$
- Less efficient variance estimator for $\hat{Z}$:

\[
\hat{V}_0(\hat{Z}) = \sum_{k=1}^{L_0} h_{k0}(\hat{Z}_0^{(k)} - \hat{Z})(\hat{Z}_0^{(k)} - \hat{Z})'
\]

where $\hat{Z}_0^{(k)} = \sum_{i \in S} w_{i0}^{(k)} z_i$
Weight Calibration (Cont’d)

Step 3

- Find $\hat{T}^{(k)} = \sum_{i \in S} w_{i0}^*(k) z_i$ such that

$$
\sum_{k=1}^{L_0} h_{k0}(\hat{T}^{(k)} - \hat{Z})(\hat{T}^{(k)} - \hat{Z})' = \hat{V}_1(\hat{Z}).
$$

To find $\hat{T}^{(k)}$ satisfying (6), we use the following steps:

1. Decompose $\hat{V}_1(\hat{Z})$ into $\hat{V}_1(\hat{Z}) = \sum_{k=1}^{p_0} \alpha_k q_k q_k'$ for some $\alpha_k > 0$ and $m$-dimensional vector $q_k$, $k = 1, \cdots, p_0$. ($p_0 < m$)

2. Construct

$$
\hat{T}^{(k)} = \begin{cases} 
\hat{Z} + (\alpha_k/h_{k0})^{1/2} q_k, & \text{if } k = 1, \cdots, p_0 \\
\hat{Z}, & \text{if } k = p_0 + 1, \cdots, L_0
\end{cases}
$$
Weight Calibration (Cont’d)

Step 4

• Find

\[ w^*_0(k) = (w^*_{10}, \ldots, w^*_{n0})', \quad k = 1, \ldots, L_0 \]

that minimizes

\[ D \left( w^*_0(k), w^{(k)}_0 \right) = \sum_{i \in S} \tau_i \left( w^*_i(k) - w^{(k)}_i \right)^2 / w^{(k)}_i \]

subject to

\[ \sum_{i \in S} w^*_i(k) z_i = \hat{T}^{(k)}. \]

• The resulting replicate \( \hat{Y}^*(k) = \sum_{i \in S} w^*_i(k) y_i \) can take the form

\[ \hat{Y}^{(k)} = \hat{Y}_0^{(k)} + \left( \hat{T}^{(k)} - \hat{Z}_0^{(k)} \right)' \hat{\beta}^{(k)} \]

where \( \hat{\beta}^{(k)} = \left\{ \sum_{i \in S} w^{(k)}_{i0} z_i z_i' / \tau_i \right\}^{-1} \sum_{i \in S} w^{(k)}_{i0} z_i y_i / \tau_i. \)
Weight Calibration (Cont’d)

Step 4 (Cont’d)

- The choice of $\tau_i = w_i^{-1}$ is proposed (to make $\text{Cov} (\hat{e}, \hat{Z}) = 0$). In this case,

\[
\hat{V}_0^* = \sum_{k=1}^{L_0} c_{k0} \left( \hat{Y}_0^{*(k)} - \hat{Y} \right)^2
\]

\[
= \sum_{k=1}^{L_0} c_{k0} \left\{ \hat{e}_0^{(k)} - \hat{e}_0 + \left( \hat{T}^{(k)} - \hat{Z} \right)' \hat{\beta} \right\}^2
\]

\[
= \hat{V}_0 (\hat{e}) + \hat{\beta}' \hat{V}_1 (\hat{Z}) \hat{\beta}.
\]

(7)

The first component is not fully efficient but the second component is fully efficient.

- In practice, negative replication weights need to be avoided. Need to change the objective function $D(w_0^{*(k)}, w_0^{(k)})$ in the calibration weighting for replication.
Properties of the Calibrated Weights

- $\hat{V}_0(\hat{Y})$: Less efficient variance estimator based on $L_0$ replicates.
- $\hat{V}_1(\hat{Y})$: Fully efficient variance estimator based on $L$ replicates. ($L >> L_0$)
- $\hat{V}_0^*(\hat{Y})$: proposed calibration variance estimator based on $L_0$ replicates.

Results

1. $V\{\hat{V}_1(\hat{Y})\} \leq V\{\hat{V}_0^*(\hat{Y})\} \leq V\{\hat{V}_0(\hat{Y})\}$.
2. If $\hat{Y} = \sum_{i=1}^{m} a_i \hat{Z}_i$, then $V\{\hat{V}_0^*(\hat{Y})\} \div V\{\hat{V}_1(\hat{Y})\}$. 
Balanced sampling

- Sampling design satisfying

\[
\sum_{i \in S} \frac{1}{\pi_i} x_i = \sum_{i \in U} x_i. \tag{8}
\]

- We are interested in creating replication variance method for

\[ \hat{Y}_{HT} = \sum_{i \in S} \pi_i^{-1} y_i. \]

- Write \( \hat{Y}_{HT} = \sum_{i \in S} \pi_i^{-1} y_i \) as

\[
\hat{Y}_{reg} = \hat{Y}_{HT} + \left( X - \hat{X}_{HT} \right)' \hat{B}
\]

for some \( \hat{B} \).

- Note that \( \hat{Y}_{reg} = \hat{Y}_{HT} \), regardless of \( \hat{B} \), under (8).
**Balanced sampling**

- The choice of

\[
\hat{B} = \left\{ \sum_{i \in S} \pi_i^{-2} (1 - \pi_i) x_i x_i' \right\}^{-1} \sum_{i \in S} \pi_i^{-2} (1 - \pi_i) x_i y_i
\]

leads to

\[
V\left(\hat{Y}_{\text{reg}}\right) = E \left\{ \sum_{i \in S} \pi_i^{-2} (1 - \pi_i) (y_i - \hat{y}_i)^2 \right\}
\]

where \( \hat{y}_i = x_i' \hat{B} \).

- Bredit and Chauvet (2011) approximation

\[
\hat{V} = \frac{n}{n - q} \sum_{i \in S} (1 - \pi_i) \pi_i^{-2} (y_i - \hat{y}_i)^2.
\]
Balanced sampling

- Jackknife variance estimator

\[ \hat{V}_{JK} = \sum_{k=1}^{n} h_k \left( \hat{Y}^{*}(k) - \hat{Y}_{HT} \right)^2, \tag{9} \]

where \( \hat{Y}^{*}(k) = \hat{Y}(k) + (\hat{X} - \hat{X}^{(k)}_{HT})' \hat{B}, \)

\( (\hat{X}^{(k)}_{HT}, \hat{Y}^{(k)}_{HT}) = \sum_{i \in S^{(k)}} \pi_i^{-1} (x_i', y_i), \ h_k = (1 - \pi_k) n / (n - q), \) and \( S^{(k)} = S \cap \{k\}^c. \)

- Note that

\[ \hat{Y}^{*}(k) - \hat{Y}_{HT} = \left( \hat{Y}(k) - \hat{Y}_{HT} \right) + (\hat{X} - \hat{X}^{(k)}_{HT})' \hat{B} \]

\[ = -\pi_k^{-1} \left( y_k - x_k' \hat{B} \right). \]
1 Introduction

2 Fully Efficient Replication Weights

3 Sparse and Efficient Replication Weights

4 Balanced Sampling

5 Simulation Studies

6 Concluding Remarks
Family Expenditure Survey; N=2246; treated as a population

\( x_1: \# \text{ Children}; x_2: \# \text{ Youth}; x_3: \# \text{ people}; x_4: \text{ annual income}; y_1: \text{ total annual expenditure}; y_2: \text{ expenditure on clothing}; y_3: \text{ expenditure on furnishings and equipment} \)

Rao-Sampford PPS sampling method; known \( \pi_i \) and \( \pi_{ij} \)

\[
\begin{align*}
\theta_1 &= Y_1 = \sum_{i=1}^{N} y_{i1} \\
\theta_2 &= \frac{\bar{Y}_2}{\bar{Y}_3} = \left( N^{-1} \sum_{i=1}^{N} y_{i2} \right) / \left( N^{-1} \sum_{i=1}^{N} y_{i3} \right)
\end{align*}
\]

\( z_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, y_{i2}, y_{i3})' \)

Relative bias and relative efficiency, based on 2000 runs

\[
R_B = \frac{1}{B} \sum_{b=1}^{B} \frac{\hat{V}(b) - V}{V} \quad \text{and} \quad RE = \frac{\{MSE(\hat{V}_L)\}^{1/2}}{\{MSE(\hat{V})\}^{1/2}},
\]
Relative Bias ($\%$) of Replication Variance Estimators for $\theta = Y_1$

<table>
<thead>
<tr>
<th>$L_0$</th>
<th>$n$</th>
<th>$\hat{V}_L$</th>
<th>$\hat{V}_R^{(1)}$</th>
<th>$\hat{V}_R^{(2)}$</th>
<th>$\hat{V}_R^{(3)}$</th>
<th>$\hat{V}_R^{(4)}$</th>
<th>$\hat{V}_R^{(C)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>50</td>
<td>−7.39</td>
<td>−7.74</td>
<td>−7.58</td>
<td>−6.48</td>
<td>−6.00</td>
<td>0.92</td>
</tr>
<tr>
<td>100</td>
<td>2.49</td>
<td>1.39</td>
<td>−0.63</td>
<td>1.80</td>
<td>2.95</td>
<td>2.13</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>−1.16</td>
<td>−1.09</td>
<td>−1.65</td>
<td>−1.96</td>
<td>2.96</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>1.73</td>
<td>−1.39</td>
<td>−2.02</td>
<td>0.58</td>
<td>2.91</td>
<td>−0.07</td>
</tr>
<tr>
<td>150</td>
<td>−1.16</td>
<td>1.62</td>
<td>0.04</td>
<td>−1.93</td>
<td>−0.74</td>
<td>−1.63</td>
<td></td>
</tr>
</tbody>
</table>
Relative Bias (\%) of Replication Variance Estimators for 
\( \theta = \bar{Y}_2 / \bar{Y}_3 \)

<table>
<thead>
<tr>
<th>( L_0 )</th>
<th>( n )</th>
<th>( \hat{V}_L )</th>
<th>( \hat{V}_R^{(1)} )</th>
<th>( \hat{V}_R^{(2)} )</th>
<th>( \hat{V}_R^{(3)} )</th>
<th>( \hat{V}_R^{(4)} )</th>
<th>( \hat{V}_R^{(C)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>50</td>
<td>-4.01</td>
<td>-3.45</td>
<td>-1.74</td>
<td>-5.37</td>
<td>-5.56</td>
<td>-2.20</td>
</tr>
<tr>
<td>100</td>
<td>-1.06</td>
<td>-0.39</td>
<td>3.37</td>
<td>0.39</td>
<td>-0.96</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>-0.12</td>
<td>-3.08</td>
<td>0.83</td>
<td>-7.16</td>
<td>-5.59</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>-1.18</td>
<td>-0.83</td>
<td>0.60</td>
<td>-2.12</td>
<td>-1.02</td>
<td>-1.31</td>
</tr>
<tr>
<td>150</td>
<td>-0.12</td>
<td>0.19</td>
<td>1.79</td>
<td>1.53</td>
<td>0.31</td>
<td>1.03</td>
<td></td>
</tr>
</tbody>
</table>
Relative Efficiency of Replication Variance Estimators for $\theta = Y_1$

<table>
<thead>
<tr>
<th>$L_0$</th>
<th>$n$</th>
<th>$\hat{V}_L$</th>
<th>$\hat{V}_R^{(1)}$</th>
<th>$\hat{V}_R^{(2)}$</th>
<th>$\hat{V}_R^{(3)}$</th>
<th>$\hat{V}_R^{(4)}$</th>
<th>$\hat{V}_R^{(C)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>50</td>
<td>1.00</td>
<td>0.84</td>
<td>1.05</td>
<td>0.89</td>
<td>0.93</td>
<td>1.34</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>1.00</td>
<td>0.55</td>
<td>0.80</td>
<td>0.92</td>
<td>0.93</td>
<td>1.34</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>1.00</td>
<td>0.47</td>
<td>0.70</td>
<td>0.83</td>
<td>0.76</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
<td>0.97</td>
<td>0.98</td>
<td>1.61</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>1.00</td>
<td>0.57</td>
<td>0.72</td>
<td>0.95</td>
<td>0.96</td>
<td>1.28</td>
</tr>
</tbody>
</table>
Relative Efficiency of Replication Variance Estimators

for $\theta = \bar{Y}_2 / \bar{Y}_3$

<table>
<thead>
<tr>
<th>$L_0$</th>
<th>$n$</th>
<th>$\hat{V}_L$</th>
<th>$\hat{V}_R^{(1)}$</th>
<th>$\hat{V}_R^{(2)}$</th>
<th>$\hat{V}_R^{(3)}$</th>
<th>$\hat{V}_R^{(4)}$</th>
<th>$\hat{V}_R^{(C)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>50</td>
<td>1.00</td>
<td>0.78</td>
<td>0.80</td>
<td>0.69</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.00</td>
<td>0.54</td>
<td>0.57</td>
<td>0.35</td>
<td>0.49</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>1.00</td>
<td>0.50</td>
<td>0.52</td>
<td>0.42</td>
<td>0.55</td>
<td>0.99</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>1.00</td>
<td>0.74</td>
<td>0.77</td>
<td>0.60</td>
<td>0.64</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>1.00</td>
<td>0.60</td>
<td>0.61</td>
<td>0.35</td>
<td>0.50</td>
<td>0.99</td>
</tr>
</tbody>
</table>
1 Introduction

2 Fully Efficient Replication Weights

3 Sparse and Efficient Replication Weights

4 Balanced Sampling

5 Simulation Studies

6 Concluding Remarks
Remarks

- Construction of sparse and efficient replication weights is a task at the data file preparation stage.
- Issues with computational implementation need to be dealt with by those who produce public-use data files, not the data users.
- The algebraic construction of fully efficient replication weights can serve as a starting point.
- The proposed strategy, i.e. random selection plus weight calibration, on producing sparse and efficient replication weights can be applied to any initial weights from jackknife, bootstrap or BRR methods.
- Identify and include all key variables at the calibration stage.
- To produce a small number (say 30-50) sets of replication weights which provide valid and efficient variance estimation is an important research problem with both theoretical and practical significance.
Dedication

This work is dedicated to the memory of

Professor Randy Sitter