Chapter 11: Two-phase sampling

Jae-Kwang Kim

Iowa State University

Fall, 2014
Introduction

Two-phase sampling for stratification

Two-phase regression estimator

Repeated Survey
Motivation

- Use of auxiliary variable $x$
  1. Design stage: stratification, PPS (or $\pi ps$) sampling
  2. Estimation stage: ratio estimation, regression estimation

- Need to observe $x_i$ in the population (design stage), or need to know the total of $x_i$ (estimation stage)

- What if $x$ is not available? - Use a sampling to observe $x$ first.
Basic structure

- [Phase 1] Select a (simple) random sample $A_1$. Observe $x_i \in A_1$.
- [Phase 2] Treat $A_1$ as if the population and make a well chosen sampling design to select a sample $A_2$ from $A_1$ using the information $x_i$ that are observed in $A_1$. Observe $y_i \in A_2$.

Sometimes called double sampling.
Notation

- \( \pi^{(1)}_i = Pr (i \in A_1) \): (first-order) inclusion probability under the first-phase sampling
- \( \pi^{(2)}_{i|A_1} = Pr (i \in A_2 \mid A_1) \): (first-order) conditional inclusion probability under the second-phase sampling given the first-phase sample
- \( \pi_i = Pr (i \in A_2) \): (first-order) inclusion probability under the two-phase sampling

\[
\pi_i = \sum_{A_1; \ i \in A_1} \pi^{(2)}_{i|A_1} P_1 (A_1) = E_1 \left\{ \pi^{(2)}_{i|A_1} I (i \in A_1) \right\}
\]
If $\pi_{i|A_1}^{(2)} = \pi_i^{(2)}$ (invariance), then $\pi_i = \pi_i^{(1)} \pi_i^{(2)}$.

If the invariance does not hold, then we cannot compute the first-order inclusion probability $\pi_i$.

Cannot use the HT estimator

Example:

1. Phase one: SRS of size $n$.
2. Phase two: $\pi_{ps}$ sampling of size $r$ with $\pi_i \propto x_i$.

Note that

$$\pi_{i|A_1}^{(2)} = \frac{rx_i}{\sum_{k \in A_1} x_k}$$

and $\pi_i$ cannot be computed from one realization of $A_1$.  

Remedy

- Use $\pi^*$-estimator:

$$\hat{Y}^* = \sum_{i \in A_2} \frac{y_i}{\pi_i^{(1)} \pi_i^{(2)}} \equiv \sum_{i \in A_2} \frac{y_i}{\pi^*_i}$$

- Properties
  - Unbiased for $Y = \sum_{i=1}^{N} y_i$
  - Variance

$$V\left(\hat{Y}^*\right) = V\left\{\sum_{i \in A_1} \frac{y_i}{\pi_i^{(1)}}\right\} + E\left\{\sum_{i \in A_1} \sum_{j \in A_1} \left(\pi_{ij|A_1}^{(2)} - \pi_i^{(2)} \pi_j^{(2)}\right) \frac{y_i}{\pi_i^*} \frac{y_j}{\pi_j^*}\right\}$$
1 Introduction

2 Two-phase sampling for stratification

3 Two-phase regression estimator

4 Repeated Survey
Basic setup: Wish to perform a stratified sampling, but the stratum indicator variables $x_i = (x_{i1}, \cdots, x_{iH})$ are not available in the population frame.

Two-phase sampling for stratification

1. Perform a SRS of size $n$ from the finite population and obtain $\sum_{i \in A_1} x_i = (n_1, n_2, \cdots, n_H)$ where $n = \sum_{h=1}^{H} n_h$.

2. Among the $n_h$ elements, select $r_h$ elements by SRS independently across the strata.
Point estimation

\[ \hat{Y}_{tp} = \sum_{h=1}^{H} w_h \bar{y}_{h2} \]

where \( w_h = n_h / n \) and \( \bar{y}_{h2} = r_h^{-1} \sum_{i \in A_2} x_i h y_i \).

Variance

\[ V(\hat{Y}_{tp}) = \left( \frac{1}{n} - \frac{1}{N} \right) S^2 + E \left\{ \sum_{h=1}^{H} \left( \frac{n_h}{n} \right)^2 \left( \frac{1}{r_h} - \frac{1}{n_h} \right) s_{h1}^2 \right\} \]

\[ = E \left\{ n^{-1} \sum_{h=1}^{H} w_h (\bar{y}_{h1} - \bar{y}_1)^2 + \sum_{h=1}^{H} r_h^{-1} w_h^2 s_{h1}^2 \right\} \]

where

\[ s_{h1}^2 = \frac{1}{n_h - 1} \sum_{i \in A_1} x_i h (y_i - \bar{y}_{h1})^2 \]
Variance estimation

\[ \hat{V}(\hat{Y}_{tp}) = n^{-1} \sum_{h=1}^{H} w_h (\bar{y}_{h2} - \hat{Y}_{tp})^2 + \sum_{h=1}^{H} r_h^{-1} w_h^2 s_{h2}^2 \]

Variance comparison

\[ V(\hat{Y}_{SRS}) - V(\hat{Y}_{tp}) = E \left\{ \left( \frac{1}{r} - \frac{1}{n} \right) \sum_{h=1}^{H} w_h (\bar{y}_{h1} - \bar{y}_1)^2 + \sum_{h=1}^{H} \left( \frac{1}{r} - \frac{w_h}{r_h} \right) w_h s_{h1}^2 \right\} \]

Two sources for the gain of efficiency:
1. Use \( n \) elements for the between-stratum variances.
2. An optimal choice of \( r_h \) can improve the efficiency.
Optimal allocation: Minimize

\[ V = \frac{1}{n} \left\{ \left( S^2 - \sum_{h=1}^{H} W_h S_h^2 \right) + \sum_{h=1}^{H} W_h S_h^2 \frac{1}{\nu_h} \right\} \]

subject to \( C = n \left( c_1 + \sum_{h=1}^{H} c_{2h} W_h \nu_h \right) \), where \( \nu_h = r_h/n_h \).
Solution

\[ \frac{r_h^*}{n^*} = W_h \left( \frac{c_1}{c_{2h}} \right)^{1/2} \left( \frac{S_h^2}{S^2 - \sum_{h=1}^{H} W_h S_h^2} \right)^{1/2}. \]

If \( c_{2h} = c_2, S_h = S_w, \) and \( \phi = S^2/S_w^2 \) then the optimal solution is

\[ \frac{r^*}{n^*} = \left( \frac{c_1}{c_2} \right)^{1/2} \left( \frac{1}{\phi - 1} \right)^{1/2}. \]
1 Introduction

2 Two-phase sampling for stratification

3 Two-phase regression estimator

4 Repeated Survey
Basic setup:

1. Phase 1: Simple random sampling of size $n$ to get $A_1$, observe $x_i, i \in A_1$
2. Phase 2: From $A_1$, simple random sampling of size $r$ to get $A_2$, observe $(x_i, y_i), i \in A_2$

Regression estimator

$$\bar{y}_{reg,tp} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)' \hat{B}$$
Two-phase regression estimator

- Property
  1. Taylor linearization
     \[ \bar{y}_{reg, tp} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)' B + O_p (r^{-1}) \]
  2. Approximately unbiased
  3. Variance
     \[ V (\bar{y}_{reg, tp}) \approx \left( \frac{1}{n} - \frac{1}{N} \right) B' S_{xx} B + \left( \frac{1}{r} - \frac{1}{N} \right) S_{ee} \]
Note that

\[ S_y^2 = S_e^2 + B' S_{xx} B \]

Variance comparison

\[ V(\bar{y}_2) - V(\bar{y}_{\text{reg},tp}) = \left( \frac{1}{r} - \frac{1}{n} \right) B' S_{xx} B \geq 0. \]

If \( \text{Corr}(x, y) \to 1 \), the gain is high.
1 Introduction

2 Two-phase sampling for stratification

3 Two-phase regression estimator

4 Repeated Survey
Motivation: sampling for the same population over time.

Several parameters

1. Difference of the means between time: $\bar{Y}_2 - \bar{Y}_1$
2. Overall mean: $\left( \bar{Y}_1 + \bar{Y}_2 \right) / 2$
3. Most recent mean: $\bar{Y}_2$

Best sampling design

1. For $\theta_1 = \bar{Y}_2 - \bar{Y}_1$: full overlap sampling
2. For $\theta_2 = \left( \bar{Y}_1 + \bar{Y}_2 \right) / 2$: no overlap sampling
3. For $\theta_3 = \bar{Y}_2$: partial replacement sampling
Partial overlap sampling: Let $\bar{Y}_2$ be the parameter of interest.

1. At time $t = 1$: Select a SRS of size $n$. Let $A_1$ be the realized sample at $t = 1$.
2. At time $t = 2$: partition the population $U$ into two strata: $A_1$ and $U \sim A_1$. From $A_1$, select a SRS of size $n_m$ to get $A_{2m}$. From $U \sim A_1$, select a SRS of size $n_u = n - n_m$ to get $A_{2u}$. The final sample at $t = 2$ is $A_2 = A_{2m} \cup A_{2u}$.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Pop’n size</th>
<th>Sample size</th>
<th>Estimator for $Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched</td>
<td>$n$</td>
<td>$n_m$</td>
<td>$\hat{Y}_m$</td>
</tr>
<tr>
<td>Unmatched</td>
<td>$N - n$</td>
<td>$n_u$</td>
<td>$\hat{Y}_u$</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>$n$</td>
<td>$\alpha \hat{Y}_u + (1 - \alpha) \hat{Y}_m$</td>
</tr>
</tbody>
</table>
Lemma

Consider the composite estimator

\[ \hat{Y}_\alpha = \alpha \hat{Y}_u + (1 - \alpha) \hat{Y}_m \]

for some constant \( \alpha \).

- If \( \hat{Y}_u \) and \( \hat{Y}_m \) are unbiased, then \( \hat{Y}_\alpha \) is unbiased.
- Variance of \( \hat{Y}_\alpha \) is minimized at

\[ \alpha^* = \frac{V \left( \hat{Y}_m \right) - \text{Cov} \left( \hat{Y}_u, \hat{Y}_m \right)}{V \left( \hat{Y}_u \right) + V \left( \hat{Y}_m \right) - 2 \text{Cov} \left( \hat{Y}_u, \hat{Y}_m \right)} \]
Two-phase sampling approach

- How to choose $\hat{Y}_u$ and $\hat{Y}_u$?
- $x$: observation at $t = 1$, $y$: observation at $t = 2$
- Estimators

$$\hat{Y}_u = \frac{1}{n_u} \sum_{i \in A_{2u}} y_i \equiv \bar{y}_u$$

$$\hat{Y}_m = \bar{y}_m + (\bar{x}_1 - \bar{x}_m) b$$
Repeated Survey

Two-phase sampling approach

- Variances and covariance

\[
V \left( \hat{Y}_u \right) = n_u^{-1} S^2 \\
V \left( \hat{Y}_m \right) = n_m^{-1} (1 - \rho^2) S^2 + n^{-1} \rho^2 S^2 \\
Cov \left( \hat{Y}_u, \hat{Y}_m \right) = 0
\]
Two-phase sampling approach (Cont’d)

- Optimal composite estimation

$$\alpha^* = \frac{nn_u - n_u^2 \rho^2}{n^2 - n_u^2 \rho^2}$$

- Variance of the optimal estimator

$$V \left( \hat{Y}_{\alpha^*} \right) = \frac{n - n_u \rho^2}{n^2 - n_u^2 \rho^2} S^2$$
Two-phase sampling approach (Cont’d)

• Optimal allocation

\[
\frac{n_u}{n} = \frac{1}{1 + \sqrt{1 - \rho^2}}, \quad \frac{n_m}{n} = \frac{\sqrt{1 - \rho^2}}{1 + \sqrt{1 - \rho^2}}
\]

• Variance under optimal allocation

\[
V (\hat{Y}_\alpha^*) = \frac{S^2}{2n} \left(1 + \sqrt{1 - \rho^2}\right).
\]

• If \( \rho \to 1 \), then \( V (\hat{Y}_\alpha^*) \to S^2/(2n) \).

• If \( \rho \to 0 \), then \( V (\hat{Y}_\alpha^*) \to S^2/n \).