Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

Dealing with Nonresponse Using Follow-up

7th International Total Survey Error Workshop (ITSEW 2013)

M.A. Hidiroglou and V. Estevao
Statistics Canada
June 3, 2013

Overview

1. Introduction
2. Estimation with no follow-up sample
3. Follow-up sample: design and estimation
4. Hansen and Hurtwitz (1946)
5. Conclusion

1. Introduction

- With *some loss* in generality, assume sample s of size n selected by SRWSWOR from population U of size N.
- Suppose $n_1 < n$ units respond to the survey, while the remaining ones do not.
- Nonresponse may lead to bias in estimates
- Two different approaches for handling nonresponse (Bethlehem 1988)
    1. Reweigh the data: use auxiliary information in estimation
    2. Follow-up intensively a sample of the nonrespondents

2. Estimation no follow-up sample

No auxiliary data available

- Estimator of total $Y = \sum_{k \in U} y_k$ using responding units only

$$\hat{Y}_1 = \frac{N}{n_1} \sum_{k \in s_1} y_k$$

- Bias can be evaluated in one of two ways:
  - Fixed response approach (Cochran 1977)
  - Random response approach (Hartley 1946)

2. Estimation no follow-up sample

- **Fixed response approach**
  - Consider two groups:
    - Responding group U_1 with mean $\bar{Y}_1$
    - Non responding group U_2 with mean $\bar{Y}_2$
    - Bias: $B\left(\hat{Y}_1\right) = N_2\left(\bar{Y}_1 - \bar{Y}_2\right)$
- **Random response approach**
  - Response probability (unknown): $\theta_k > 0,\ k \in U$
  - Approximate bias: $B\left(\hat{Y}_1\right) = N \dfrac{\sum_U y_k\ \theta_k}{\sum_U \theta_k} - Y$

2. Estimation no follow-up sample

Auxiliary data available

Info-s : $\boldsymbol{x}_k$ is known for all $k \in s \rightarrow \hat{Y}_{1s} = \sum_{k \in s_1} w_{ks} y_k$

Info-U : Info-s and $\sum_{k \in U} \boldsymbol{x}_k$ known $\rightarrow \hat{Y}_{1U} = \sum_{k \in s_1} w_{kU} y_k$

where w_{ks} and w_{kU} are appropriate regression weights

- Calibrate from response sample to full sample/universe
  - Info U: Fuller, Loughlin and Baker (1994)
  - Info-s and U: Lundström and Särndal (1999)

2. Estimation no follow-up sample

- Assuming response probabilities θ_k
- Bias of $\hat{Y}_{1s}$ or $\hat{Y}_{1U}$ is approximately

$$-\sum_U (1-\theta_k)\left(y_k - \boldsymbol{x}_k^T \boldsymbol{B}_{U\theta}\right)$$

where

$$\boldsymbol{B}_{U\theta} = \left(\sum_U \theta_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \left(\sum_U \theta_k \boldsymbol{x}_k y_k\right)$$

2. Estimation no follow-up sample

- Unbiased if $\theta_k^{-1} = 1 + \boldsymbol{\lambda}^T \boldsymbol{x}_k$
  - Fuller, Loughlin and Baker (1994) or
    Lundström and Särndal (1999)
- How does one verify in practice! Can't!
- Above condition satisfied if (Fuller, et al. 1994):

i. Include in $\boldsymbol{x}_k$ *dummy variables* that define subgroups

ii Response probabilities θ_k are constant in each subgroup.

2. Estimation no follow-up sample

- What about if models differ between respondents and nonrespondents

$$y_k = \begin{cases} \boldsymbol{x}_k^T \boldsymbol{\beta}_1 + e_k & \text{if } k \in U_1 \text{ (respondents)} \\ \boldsymbol{x}_k^T \boldsymbol{\beta}_2 + e_k & \text{if } k \in U_2 \text{ (nonrespondents)} \end{cases}$$

- Model bias: $B_\zeta\left(\hat{Y}_{1U}\right) \doteq \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right)^T \sum_{U_2} \boldsymbol{x}_k$
- Need to follow-up

3. Follow-up sample: design and estimation

- Set-up
  - Respondents (n_1) in sample s: put in first group ($h = 1$)
  - Split nonrespondent portion of sample s into (L-1) response homogeneity groups ($h = 2, \ldots L$)
  - Select follow-up sample in each nonresponse group ($h = 2, \ldots L$)

n_h units in h-th nonreponse group

m_h units sampled in h-th nonreponse group

b_h units respond to FU in h-th sampled nonreponse group

3. Follow-up sample: design and estimation

- Parameter of interest is the population total

$$Y = \sum_{k \in U} y_k$$

- Estimator is

$$\hat{Y} = \frac{N}{n}\left(\sum_{k \in s_1} y_k + \sum_{h=2}^{L} \frac{n_h}{m_h} \frac{m_h}{b_h} \sum_{k \in s_{2h}} y_k \right)$$

$$= N\left(w_1 \, \overline{y}_1 + \sum_{h=2}^{L} w_h \, \overline{y}_{2h} \right)$$

- Estimator does not include auxiliary data (but could)

3. Follow-up sample: design and estimation

- Assume nonrespondents are missing completely at random

$$E\left(\hat{Y}\right) = N\left(W_1 \bar{Y}_1 + \sum_{h=2}^{L} W_h \bar{Y}_h\right) = N\bar{Y} = Y$$

$$Var\left(\hat{Y}\right) = N\left\{\left(\frac{1}{v} - 1\right) S^2 + \sum_{h=2}^{L} \frac{1}{v}\left(\frac{1}{v_h\, r_h^*} - 1\right) W_h\, S_h^2\right\}$$

- Anticipated response rates r_h^* for $h = 2, \ldots\ L$
- Sampling fractions v and v_h to be determined

$$v = n / N \quad \text{and} \quad v_h = m_h / n_h \quad \text{for } h = 2, \ldots\ L$$

3. Follow-up sample: design and estimation

- Costs

c_0 contact cost for each of the initial units n units

c_1 processing cost for each of the n_1 respondents

c_{2h} contact cost for each of the m_h units in follow-up sample

c_{3h} processing cost of each of the b_h respondents in follow-up

- Overall cost is random so we work with expected cost

$$C^* = N\,v\left(c_0 + W_1\,c_1 + \sum_{h=2}^{L} W_h\,v_h\,(\,c_{2h} + r_h^*\,c_{3h}) \right)$$

3. Follow-up sample: design and estimation

- Allocation problem

$$\text{Minimize} \quad N\left\{\left(\frac{1}{\nu}-1\right)S^2+\sum_{h=1}^{L}\frac{1}{\nu}\left(\frac{1}{\nu_h\, r_h^*}-1\right)W_h\, S_h^2\right\}$$

with respect to ν and ν_h $(h=2,\ldots L)$

$$\text{subject to} \quad N\nu\left(c_0+W_1\,c_1+\sum_{h=2}^{L}W_h\,\nu_h\,(c_{2h}+r_h^*\,c_{3h})\right)=C^*$$

$$\text{and} \qquad 0<\nu\le 1,\quad 0<\nu_h\le 1$$

3. Follow-up sample: design and estimation

- Solution to problem
  - by nonlinear programming techniques – Trust region method
  - use of Proc OPTMODEL in SAS 9.3
  - Closed form expression as in Rao (1973) – ignoring bounds

$$v_h = \sqrt{\frac{S_h^2 \left(c_0 + W_1 c_1 \right)}{r_h^* \left(c_{2h} + r_h^* c_{3h} \right) \left(S^2 - \sum_{h=2}^{L} W_h S_h^2 \right)}}$$

4. Hansen and Hurtwitz (1946)

- Survey of 40,000 retail stores
  - initial contact by mail and follow-up by interview
  - one response group and one nonresponse group ($h = 2$)
  - $c_0 = 0.1,\ c_1 = 0.4$ and $c_2 = 4.1$ and $c_3 = 0.4$ (US \$)
  - H&H looked at the reverse problem

$$\text{Minimize} \quad (N c_0 + N c_1 W_1) \nu + N c_2 W_2 \nu\, \nu_2$$

with respect to ν and ν_2

$$\text{Subject to} \quad \left(\frac{1}{\nu} - 1 \right) S^2 + W_2\, S_2^2\, \frac{1}{\nu} \left(\frac{1}{\nu_2} - 1 \right) = V$$

4. Hansen and Hurtwitz (1946)

- Our problem formulation with their data:

$$\text{Minimize} \quad \left(\frac{1}{v}-1\right)+W_2\frac{1}{v}\left(\frac{1}{v_2}-1\right)+W_2\frac{1}{v}\frac{1}{v_2}\left(\frac{1}{r_2^*}-1\right)$$

with respect to v and v_2

$$\text{Subject to} \quad (N c_0+N c_1 W_1)v+N(c_2+r_2^* c_3)W_2 v\, v_2=C^*$$

- They assumed 100% response to follow-up ($r_2^*=1$)
- We consider two cases: ($r_2^*=1$) and ($r_2^*=0.5$)

4. Hansen and Hurtwitz (1946)

	$r_2^* = 1$			$r_2^* = 0.5$		
C*	ν	ν_2	Min Var	ν	ν_2	Min Var
1500	0.033	0.365	54.9	0.026	0.528	90.6
2000	0.045	0.365	40.9	0.035	0.528	67.7
2500	0.056	0.365	32.5	0.044	0.528	54.0
3000	0.067	0.365	27.0	0.052	0.528	44.8
3500	0.078	0.365	23.0	0.061	0.528	38.3
4000	0.089	0.365	20.0	0.070	0.528	33.4
4500	0.100	0.365	17.6	0.078	0.528	29.5
5000	0.111	0.365	15.8	0.087	0.528	26.5

5. Conclusion

- Non-response usually handled at estimation stage with no follow-up
- However, nonresponse bias could be present even with calibration
- Follow-up of non-respondents should reduce nonresponse bias
- Incorporation of auxiliary data and estimated response probabilities in the estimation (given follow-up) can further attenuate nonresponse bias