A test of sample matching using a pseudo-web sample

Golshid Chatrchi and Jack Gambino

INPS Conference, March 16, 2017
Outline

• Introduction
• Sample matching
• Pseudo-web sample
• Simulation results
• Carrot project: an experiment
Introduction

• With increasing levels of nonresponse in household surveys, there is renewed interest in alternatives to the traditional way of conducting surveys.

• Can we use non-probability samples in a probabilistic way? How about the self selection bias?
• Bethlehem (2014)
  "Due to (high) nonresponse, probability sampling surveys more and more resemble self-selection surveys."

• Rivers (2007)
  "There is no logical difference between the type of modeling assumptions needed for nonresponse adjustments and those needed for self-selected samples."
Sample Matching (SM)

- Rivers (2007) proposed the application of Sample Matching.
- The variable of interest is not measured directly from $s$. 

Web Panel
SM- population of interest

• Let $U$ be a population of size $N$.
• A probability sample $s$ of size $n$ is drawn using a sample design $p(s)$.
• Let $\pi_i$ be the probability of selection of unit $i \in U$.
• Variable of interest is $y$.
• Let $x_i$ be the auxiliary variables in the entire population $U$ or for the sample $s$. 
SM-panel

• Let $n^*$ be the size of panel.
• Let $x_1^*, \ldots, x_n^*$ be the auxiliary variables in the panel.
• Let $y_1^*, \ldots, y_n^*$ be the values of the measurements in the panel.
• Let $z_i$ be an indicator of responding to the web-panel survey.
• We assume that $z_i = 1, i = 1, \ldots, n^*$. 
SM- mechanism

• Let $d(a,b)$ be a measure of distance between $a$ and $b$.

• For each unit $i$ in sample $s$, we find a set of pairs $(y_{i}^*, x_{i}^*)$ on the panel where $d(x_{i}, x_{i}^*)$ is small.

• We select one unit at random from the set and substitute $y_{i}$ with $y_{i}^*$. 
SM- mechanism

- Estimator of total: \[ \hat{T} = \sum_{i \in S} \frac{y_i^*}{\pi_i} \]
SM- assumptions

There are three main assumptions in Rivers’ paper:

1. “iid” data generating process \((y_i, x_i, z_i)\)
2. The panel covers all relevant portions of the population \(U\).
3. Ignorable selection

\[ F_{Y|X}(y|x) = F_{Y^*|X^*}(y|x) \forall x, y \]

The conditional distribution of \(Y\) on \(X\) in the population is identical to that in the panel.
Pseudo-web sample

- Two different household surveys are used to simulate the SM methodology:
  - 2011 National Household Survey (NHS)
  - 2011 Canadian Labour Force Survey (LFS)
- NHS respondents are considered as the population of the study. A probability sample is selected from the NHS.
- LFS respondents are treated as a pseudo-web sample.
National Household Survey (NHS)

- Statistics Canada conducted the NHS in May 2011 as a replacement for the long census questionnaire.

- The survey was designed to collect social and economic data about the Canadian population.

- NHS respondents~ 6.7 million persons (“population size”)
Labour Force Survey (LFS)

• The LFS is a household survey carried out monthly by Statistics Canada.

• The goal of the survey is to provide information on major labour market trends such as unemployment rates.

• May 2011 LFS respondents ~127,000 persons (“Panel size”)
Why NHS and LFS?

- Demographic information from both surveys can be used as auxiliary information.
- NHS is large enough to be considered as our population.
- Both surveys were conducted in May 2011.
- Both surveys collect information on the labour force status and we can evaluate the method using NHS data.
Variables

- Variables of interest
  1. employed
  2. unemployed
  3. not in Labour force
  6. not applicable
     (less than 15 years old)

- Matching variables \((x_i, x^*_i)\)
  geographical variables, sex, age, education
Simulation

- Random sample from NHS
- Sample size: 5000, 10000, 25000
- R=1000 simulated samples
- Matching variables:
  - Age/sex/province
  - Age/sex/education
- Variable of interest: respondent was employed during the reference week
  
  \[ y = \begin{cases} 
  1 & \text{if respondent was employed} \\
  0 & \text{otherwise} 
  \end{cases} \]
Simulation

- Two performance measures are considered:
  - Absolute bias (AB)
    \[ AB = \left| \left( \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}(r) \right) - \theta \right| \]
  - Root mean square error (RMSE)
    \[ RMSE = \sqrt{\frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}(r) - \theta)^2} \]
Simulation 1

- Matching variable: province*age*sex
- \( R=1000 \) simulated samples
- SRS

<table>
<thead>
<tr>
<th>( n )</th>
<th>( AB )</th>
<th>( RMSE )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>0.036</td>
<td>0.694</td>
</tr>
<tr>
<td>10000</td>
<td>0.028</td>
<td>0.520</td>
</tr>
<tr>
<td>25000</td>
<td>0.004</td>
<td>0.321</td>
</tr>
<tr>
<td>50000</td>
<td>0.071</td>
<td>0.230</td>
</tr>
<tr>
<td>100000</td>
<td>0.659</td>
<td>0.912</td>
</tr>
<tr>
<td>150000</td>
<td>1.225</td>
<td>1.463</td>
</tr>
</tbody>
</table>
Simulation 2

- Matching variable: age*sex*level of education
- R=1000 simulated samples
- SRS

<table>
<thead>
<tr>
<th>$n$</th>
<th>$AB$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>1.002</td>
<td>1.197</td>
</tr>
<tr>
<td>10000</td>
<td>0.951</td>
<td>1.313</td>
</tr>
<tr>
<td>25000</td>
<td>0.676</td>
<td>0.730</td>
</tr>
</tbody>
</table>
Simulation 3

- Matching variable: province*age*sex
- \( R = 1000 \) simulated samples
- Stratified sampling with power allocation \((q=0.5)\)

\[
n_h = n \frac{M_h^q}{\sum_{h=1}^{L} M_h^q}
\]

- \( M_h \) is total number of persons with employment income

<table>
<thead>
<tr>
<th>( n )</th>
<th>( AB )</th>
<th>( RMSE )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>0.335</td>
<td>0.640</td>
</tr>
<tr>
<td>10000</td>
<td>0.303</td>
<td>0.530</td>
</tr>
<tr>
<td>25000</td>
<td>0.021</td>
<td>0.327</td>
</tr>
</tbody>
</table>
Lessons learned

• Sample size
• Matching variables
• Sampling mechanism

Rivers (2007)

• “Sample matching is nearly unbiased if the panel is five times the size of the target sample.”
• “The plausibility of this assumption depends largely on the extent and relevance of the matching variables.”
• “Matching from a sufficiently large and diverse panel yields results similar to a SRS.”
Limitations of the method

- Survey data don’t have the same characteristics as the panel data
  - self-selected
  - coverage
- Variable of interest (LFS) is a complex derived variable.
  - Imputation impact
Carrot Project: an experiment

- Carrot Rewards app*
  - incentive-based digital platform
  - originally, a wellness app for making healthy choices
- Register using basic demographic information
- Register rewards card (gas card, movie card, AEROPLAN miles)
- Receive mini surveys
- Complete tasks and collect reward points

* non-governmental application developed by Social Change Rewards (www.carrotinsights.com)
Carrot Project: an experiment

• Three mini surveys were tested using content from the Canadian Community Health Survey (CCHS).

• **Survey #1:** Demographics + Alcohol consumption
• **Survey #2:** Exposure to second hand smoke
• **Survey #3:** Neighbourhood environment
• **Surveys #2 and #3** were only sent to respondents of the first survey.
Carrot Project: an experiment

- **Survey #1** was sent to around 41K users
  Response rate: 28%
- **Survey #2 and #3** was sent to around 11.5K users
  Response rate: 65%

![Age distribution chart](chart.png)

- 18-34: 29%, 21% (Carrot App, BC population)
- 35-64: 62%, 52%
- 65+: 10%, 27%
Carrot Project: an experiment

- **Goal**: compare CCHS estimates to Carrot sample matched estimates on the same variables.
Where are we heading?
Thank you

For more information please contact:

Golshid Chatrchi
Golshid.Chatrchi@canada.ca

Jack Gambino
Jack.Gambino@canada.ca

Merci

Pour plus d’information, veuillez contacter:

Golshid Chatrchi
Golshid.Chatrchi@canada.ca

Jack Gambino
Jack.Gambino@canada.ca