Combining data from a probability and nonprobability sample to reduce survey costs and burden

STATISTICS CANADA

ONE HUNDRED YEARS AND COUNTING

Jean-François Beaumont

ITACOSM 2019, Florence
June 5-7, 2019

Context

- Since Neyman (1934), **probability surveys** have been the standard in National Statistical Offices (NSO)
- **Why?**
  - **Nonparametric approach**: Its validity does not depend on model assumptions (design-based inference)
- **In practice…**
  - Requires assumptions about nonsampling errors
  - Known to be accurate in general

Winds of change …

- Other types of data sources are more and more considered
- **Four main reasons:**
  - Decline of survey response rates ➡ bias
  - **High data collection costs + burden on respondents**
  - Desire to have "real time" statistics (Rao, 2019)
  - Proliferation of nonprobability sources (ex.: Web panel surveys, administrative data, social medias, …)
    - Less costly, typically larger sample size

Issues with nonprobability surveys

- **Bias** (selection, coverage)
  - Becomes dominant as the sample size n increases (Meng, 2018)
  - Large sample size is not a guarantee of high quality estimates...
- Measurement errors (ex.: Web panel surveys administered to volunteers)

A relevant question in the current context

- How can data from a nonprobability sample be used to
  - minimize **data collection costs** and **burden on respondents** of a probability survey
  - while preserving a **valid statistical inference framework** and an **acceptable quality**?

In what follows …

- Model-based data integration methods
  - Calibration
  - Statistical matching (sample matching)
  - Weighting by the inverse propensity score
- Few results

Notation

- Nonprobability sample: s_{NP}
  - Subset of U
  - Contains a variable of interest y_k , **assumed to be measured without errors**: $y_k \Rightarrow \mathbf{Y}$
  - Indicator of inclusion in s_{NP} : $\delta_k \Rightarrow \boldsymbol{\delta}$
- Probability sample: s_P
  - Subset of U drawn randomly
  - Survey weight: w_k (e.g., $w_k = 1/\pi_k$)
  - Does not contain y_k

Model-based approaches

- Objective:
  - Reduce burden and costs by eliminating collection of some variables of interest in s_P
- Naïve estimator of the total $\theta = \sum_{k \in U} y_k$:

$$\hat{\theta}^{NP} = N \frac{\sum_{k \in s_{NP}} y_k}{n^{NP}}$$

  - Uses only s_{NP} but can be very biased (Bethlehem, 2016)
- Data integration methods
  - Reduce bias by combining both samples through a vector of common auxiliary variables $\mathbf{x}_k$: $\mathbf{x}_k \Rightarrow \mathbf{X}$
  - Inferences are valid if model assumptions hold

Model-based approaches

- Important assumption for all three methods: **Noninformative selection**

$$F(\mathbf{Y} \mid \boldsymbol{\delta}, \mathbf{X}) = F(\mathbf{Y} \mid \mathbf{X}) \Rightarrow \Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) = \Pr(\delta_k = 1 \mid \mathbf{X})$$

- A rich vector of auxiliary variables, as predictive as possible of both y_k **and** δ_k, makes this assumption more realistic
- Key for removing selection/coverage bias
- A large multipurpose probability survey may be useful to find a rich set of auxiliary variables (beyond age, sex and region)

Calibration of s_{NP}

- **Idea** (Royall, 1970; Brewer, 1963):
  - Model the relationship between y_k and $\mathbf{x}_k$ using s_{NP} and a linear model

$$E\left(y_k \middle| \mathbf{X}\right) = \mathbf{x}_k'\boldsymbol{\beta}$$

  - BLUP of the total θ : $\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} y_k + \sum_{k \in U - s_{NP}} \mathbf{x}_k'\hat{\boldsymbol{\beta}}$

- The BLUP can be written as a calibration estimator:

$$\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} w_k^C y_k$$

with w_k^C that satisfy the calibration equation:

$$\sum_{k \in s_{NP}} w_k^C \mathbf{x}_k = \mathbf{T}_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$$

Calibration of s_{NP}

- If $\mathbf{T}_{\mathbf{x}}$ is unknown, it can be replaced with a design-unbiased estimator (e.g., Elliott and Valliant, 2017):

$$\hat{\mathbf{T}}_{\mathbf{x}} = \sum_{k \in s_P} w_k \mathbf{x}_k$$

- **Remarks:**
  - Linear model ⟺ calibration
  - Bias-Variance tradeoff
  - If many auxiliary variables, variable selection techniques (e.g., LASSO) can be useful (Chen, Valliant and Elliott, 2018)

Calibration of s_{NP}

- **Poststratification model**:
  - $E\left(y_k \middle| \mathbf{X}\right) = \mu_h \quad , \quad k \in U_h$
  - Natural when auxiliary variables are categorical
  - BLUP of θ: $\hat{\theta}^{BLUP} = \sum_{h=1}^{H} \hat{N}_h \hat{\mu}_h$
- **Reduction of selection bias:**
  - Consider a large number of poststrata (e.g., crossing many categorical variables)
  - Regression trees could be useful to avoid overfitting

Statistical matching

- **Idea**:
  - Model the relationship between y_k and $\mathbf{x}_k$ using s_{NP}
  - Predict (impute) y_k , $k \in s_P$, by y_k^{imp}
- Predictor of the total θ : $\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp}$
- For linear models, $y_k^{imp} = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ and, in most cases,
  - statistical matching is identical to calibration on estimated totals $\hat{\mathbf{T}}_{\mathbf{x}}$
  - Ex.: poststratification model

Statistical matching

- Donor imputation is often considered (ex.: Rivers, 2007)
  - Nonparametric method
  - Does not require a linear model
- Fractional donor imputation (Kim and Fuller, 2004) is an alternative
  - More efficient
  - Does not have impact in terms of bias reduction

Statistical matching

- Linear regression, donor and fractional donor imputation are all special cases of **linear imputation:**
  (Beaumont and Bissonnette, 2011)

$$y_k^{imp} = \sum_{l \in s_{NP}} \omega_{kl} y_l \quad , \quad k \in s_P$$

- $\hat{\theta}^{SM}$ can be rewritten in a weighted form:

$$\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp} = \sum_{k \in s_{NP}} W_k \, y_k$$

Weighting by the inverse PS

- **Idea:**
  - Model the relationship between δ_k and $\mathbf{x}_k$
  - Estimate the **participation probability** $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right)$ by $\hat{p}_k$
  - Assumption: $p_k > 0$
  - Estimator: $\hat{\theta}^{PS} = \sum_{k \in s_{NP}} w_k^{PS} y_k$, where $w_k^{PS} = 1/\hat{p}_k$
- **Main advantage:**
  - Simplify the modelling effort when there are many variables of interest (only one participation indicator to model)

Weighting by the inverse PS

- **Parametric model** (ex.: logistic):

$$p_k(\boldsymbol{\alpha}) = g(\mathbf{x}_k;\boldsymbol{\alpha}) = \{1+\exp(-\mathbf{x}_k'\boldsymbol{\alpha})\}^{-1}$$

- Estimated probability: $\hat{p}_k = g(\mathbf{x}_k;\hat{\boldsymbol{\alpha}})$
- How to estimate $\boldsymbol{\alpha}$ such that $\hat{\theta}^{PS}$ is unbiased?
- **Maximum likelihood (logistic):**
  - $\sum_{k\in s_{NP}} \mathbf{x}_k - \sum_{k\in U} p_k(\boldsymbol{\alpha})\mathbf{x}_k = \mathbf{0}$
  - Requires knowing $\mathbf{x}_k$ for the entire population

Weighting by the inverse PS

- **Chen, Li and Wu (2019):**
  - $\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$
  - Requires knowing $\mathbf{x}_k$ for a probability sample
- **Alternative** (Iannacchione, Milne and Folsom, 1991)**:**
  - $\sum_{k \in s_{NP}} \dfrac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{0}$
  - **Calibration property:**

$$\sum_{k \in s_{NP}} w_k^{PS} \mathbf{x}_k = \hat{\mathbf{T}}_{\mathbf{x}}$$

Weighting by the inverse PS

- Formation of homogeneous classes with respect to $\hat{p}_k$
  - For units of the nonprobability sample in a given class h:

$$w_k^{PS} = \frac{\hat{N}_h}{n_h^{NP}}$$

  - Equivalent to a poststratified estimator
- **Some remarks:**
  - Choice of auxiliary variables (or homogeneous classes) is the key to reduce selection bias
  - Regression trees?

Application to real data

- **Nonprobability sample:**
  - Web panel of about 155 000 volunteers
- **Probability sample:**
  - CCHS (health survey of about 25 000 respondents)
- **Auxiliary variables:**
  - Health region, age, sex, marital status, education
- **Methods:**
  - Statistical matching using donor imputation (with hierarchical classes)
  - Calibration (raking on marginals)

Variable	Estimates of proportions			
	CCHS (±1.96*s.e.)	Naive	Calibration	Statistical Matching
High blood pressure	19.3% (±0.8%)	14.3%	22.1%	28.6%
Very strong sense of belonging to the community	19.5% (± 0.8%)	8.4%	10.9%	14.8%
Somewhat weak sense of belonging to the community	22.1% (± 1.0%)	36.4%	33.6%	30.2%
Excellent health	23.3% (±0.9%)	7.8%	8.9%	11.7%
Very good health	35.9% (±1.0%)	29.4%	33.8%	33.0%
Excellent mental health	33.5% (±1.1%)	13.7%	17.0%	21.4%
Fair mental health	6.0% (±0.5%)	17.1%	13.1%	11.4%

Conclusions from results

- Both statistical matching and calibration reduced bias of the nonprobability sample
- Statistical matching seemed to achieve slightly larger bias reduction
  - Accounted for interactions between variables
- Some bias persisted. Two possible reasons:
  - Matching variables not sufficiently associated with the health variables of interest that we considered
  - Measurement errors