Small Area Estimation

1 - Small area estimation problem

2 - Estimation for domains - **Direct estimators** – estimation for planned domains

3 – Coefficient of Variation and Minimum level of precision

4- Estimation for unplanned domains and/or where the sample size is not enough for the minimum level of precision

Recap

- Target parameters in the domain:

- Total of the study variable $t = \sum_{k \in U} y_k$
- Mean of the study variable $\bar{y} = \sum_{k \in U} y_k / N$
- At risk of poverty rate $P_0 = \frac{1}{N} \sum_{i=1}^{N} I(y_i < z).$
- Poverty gap $P_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{G_i}{z}.$

More on direct estimators for domains

There are more complex formulations of the direct estimator, other than HT estimator:

- ratio estimator
- regression estimator
- **calibration estimator**

These estimators are based on the use of auxiliary information, the HT estimator is based only on the y_i values of the sampled units

Auxiliary information

Auxiliary: providing supplementary or additional help and support to the estimation process.

- The auxiliary information about the population in the domain may include one or more known variables to which the variable of interest is **approximately** related.
- Suppose that the population total for the X variable is known: $\mathbf{t}_x = \sum_{i=1}^{N} \mathbf{x}_i$

Calibration estimators

we explain here the basic theory and use of calibration estimators proposed by Deville and Särndal (1992), which incorporate the use of auxiliary data

$\{\mathbf{x}_i,\ i = 1, \ldots, N\}$ and $\mathbf{t}_x = \sum_{i=1}^{N} \mathbf{x}_i$

J.C. Deville and C.E. Särndal, *Calibration estimators in survey sampling*, Journal of the American Statistical Association **87** (1992), 376–382.

Recap what learnt on the HT estimator

Suppose we are interested in estimating the population total $t_y = \sum_{i=1}^{N} y_i$. We draw a sample $s = \{1, 2, \ldots, n\} \subset U$ using a probability sampling design P, where the first and second order inclusion probabilities are $\pi_i = Pr(i \in s)$ and $\pi_{ij} = Pr(i, j \in s)$ respectively. An estimate of t_y is the Horvitz-Thompson (HT) estimator

$$\widehat{t}_{HT} = \sum_{i \in s} d_i y_i,$$

where $d_i = 1/\pi_i$ is the sampling weight, defined as the inverse of the inclusion probability for unit i.[1] An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design P. Its variance under P is given as

$$V_p(\widehat{t}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{1.1}$$

Note that we have seen different expressions of the variance of HT estimator, all derived by (1.1)

Please note that d_i and $\frac{1}{\pi_i}$ are used interchangeably

Calibration estimators

Ideally we would like that:

$$\sum_{i \in s} d_i \mathbf{x}_i = \mathbf{t}_x,$$

that is the survey **weights** d_i - positive values associated with the observations (rows) in your dataset (**sample**) - ensure that x_i sample values represent the X population total. But often times this is not true.

Calibration estimators

The idea behind calibration estimators is to find weights $w_i, i = 1, \ldots n$ close to d_i based on a *distance function* such that:

$$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{t}_x$$

when applied to x_i sample values represent the X known population total $\mathbf{t}_x = \sum_{i=1}^{N} \mathbf{x}_i$

Calibration estimators

In other words calibrating (rectifying the graduation of) the sampling weights on the basis of a distance function, we adjust sampling weights to meet benchmark constraints and range restrictions.

Assumption: $\mathbf{t}_x = \sum_{i=1}^{N} \mathbf{x}_i$ is known without error and it is a relevant constraint

Calibration estimator of a total t_y

We wish to find weights w_i similar to d_i so as to preserve the *unbiasedeness* of the HT estimator.
Once the w_i is found, the calibration estimator for t_y is:

$$\hat{t}_c = \sum_{i \in s} w_i y_i$$

How to find w_i and which is the appropriate distance function?

Distance functions

Recall our constraint, $\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{t}_x.$

we want to find w_i close to d_i based on a distance function $D(w, d)$ subject to the constraint . This is an optimization problem where we wish to minimize Q
Using the method of Lagrange multiplyers

$$Q(w_1, \ldots, w_n, \boldsymbol{\lambda}) = \sum_{i \in s} D(w_i, d_i) - \boldsymbol{\lambda} \left(\sum_{i \in s} w_i \mathbf{x}_i - \mathbf{t}_x \right)$$

Sum of the distances

Sum of the differences

Distance function

	$D(w,d)$
1. Chi-squared distance	$(w-d)^2/2qd$
2. Modified minimum entropy distance	$q^{-1}(w\log(w/d)-w-d)$
3. Hellinger distance	$2(\sqrt{w}-\sqrt{d})^2/q$
4. Minimum entropy distance	$q^{-1}(-d\log(w/d)+w-d)$
5. Modified chi-squared distance	$(w-d)^2/2qw$

The choice of distances depends on the statistician and on the problem. It is unimportant for large samples.

A common choice is Chi-squared distance

q is a tuning that can be manipulated to obtain an optimal minimum of Q

Generalized REGression Estimator

The resulting calibration estimator of t_y, differentiating with respect to w_i, is

GREG estimator

$$\begin{aligned}\hat{t}_c &= \sum_{i \in s} w_i y_i \\ &= \hat{t}_{y_{HT}} + \sum_{i \in s} d_i q_i \mathbf{x}_i^{\mathrm{T}} \mathbf{T}_s^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_{x_{HT}}) y_i \\ &= \hat{t}_{y_{HT}} + \hat{\mathbf{B}} (\mathbf{t}_x - \hat{\mathbf{t}}_{x_{HT}}),\end{aligned}$$

Where $\hat{\mathrm{B}} = \mathbf{T}_s^{-1} \sum_{i \in s} d_i q_i \mathbf{x}_i y_i$ and $\mathbf{T}_s = \sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$

slope of regression line cov(y,x)/var(x)

Confidence Interval for *GREG*

$\widehat{t_c}$ is an approximately designed-unbiased estimator of t_y

its variance is estimated by:

$$v(\widehat{t_c}) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(d_i (y_i - \widehat{\mathbf{B}} \mathbf{x}_i) \right) \left(d_j (y_j - \widehat{\mathbf{B}} \mathbf{x}_j) \right)$$

Approximate 100(1-a)% confidence interval based on t student tables is obtained on the estimated variance

$$\widehat{t_c} \pm t_{(n-1, a/2)} \cdot \sqrt{v(\widehat{t_c})}$$

Remarks

There are three major **advantages of calibration** approach
in survey sampling.

1 - the **calibration** approach leads to consistent **estimates**.

2 - it provides an important class of technique for
the efficient combination
of data sources.

3 - **calibration** approach has computational **advantage** to calculate **estimates**.

Remarks

There are also **limitations of calibration** approach in survey sampling.

1 - a limitation of the calibration estimator is that it relies on an **implicit linear relationship** between the study variable, y , and the auxiliary variable x (all calibration estimators are asymptotically equivalent to the GREG)

2 - if there exists a **non-linear relationship** between y and x , the calibration estimator does not perform as well as the HT estimator, that is, if we ignore the auxiliary variable altogether

3 - another limitation of the calibration estimator previously mentioned is that the **weights can take on negative and/or extremely large values**. Deville and Särndal recognized this issue and showed how to restrict the weights to fall within a certain range.

- Examples of application of the calibration estimator during the R lab