EU Twinning Project on Statistics in Jordan

دائرة الإحصاءات العامة
Department of Statistics

Small Area Estimation
-An simplified Introduction-

Andreas Berg, PhD

Amman, 26-02-2024

Small Area Estimation

- Introduction
- Horvitz-Thompson
- Ratio and difference estimation
- GREG

Small Area Estimation

- Introduction

Small Area Estimation

- We don‘t need small area estimation

- If we rely on small are estimation we probably made already some mistakes in planning our survey regarding

- Overall sample size
- Wrong choice of areas/domains
- Mismatch between planned and to be published area results
- Post-correction of results to be published (can we even publish results in subcategories/deeper disaggregation level?)
- Serious underestimation of non-response

- Example

- A sample size of 1000 people has been considered to get an estimate for whole Jordan
- To answer a question of expenditure for eating out
- After sampling one decides to regionalise the estimates
- Governorate of Jarash by chance gets only a net sampling size of 5
- With classical methods sample size of 5 the results for Jarash will be unreliable and unpublishable

- In general: Target is to improve our estimates in terms of accuracy

- This can be done by

- Sample size
- Sampling design
- Estimation method

- Best: as a combination of all the three items

- Small are estimation is a tool which can - but doesn‘t guarantee - in some cases improve the quality/reliability of our results if sample sizes are too small for classical/established estimation procedures.
- When we need to use small area estimation the
- Sample size is already fixed
- The sample is already drawn
- So small area estimation is the last resort to improve – if needed – the reliability of survey results

- We need small area estimation if classical estimation results are unreliable. When are they unreliable? Depends on the quality standards you like to achieve.
- For instance in terms of setting threashholds for

- Relative/absolute standard errors
- MSE
- ..

- What are classical estimators?

- In terms of survey sampling methods, we like to use the selection probabilities for selecting a unit from a population to create our estimation. Because the selection probabilities are defined by the the sampling design, resp. The chosen sampling method, we call these estimates also **design-based**

- Horvitz-Thompson estimator
- Ratio estimator
- Difference estimator

(General) Regression estimator

- A typical small area estimator consists of a combination of

- Design-based estimator
- And a synthetic estimator

Estimation

- Horvitz-Thompson estimator

Estimation

In survey sampling we often sample with different selection probabilities.

A simple **unbiased** estimator for the population total or population mean taking in account unequal selection probabilities is the so called **Horvitz Thompson estimator**

For the population total:

$$\hat{Y} = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i$$

With selection probability π_i and sample units y_i

Example: simple random sampling without replacement: $\pi_i = \frac{n}{N}$

Estimation

- Ratio estimator

Estimation

If auxiliary information is available we can in general greatly improve the quality of our estimation by using estimators which include this information

Ratio estimator:

$$\hat{Y} = \frac{X}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} y_i$$

with y_i , x_i sample units and X population total of an auxiliary variable

Estimation

Properties of the ratio estimator:

- The accuracy improves with the correlation between X and Y
- Not unbiased, but at least approximatively so (n large)

Rule of thumb: when comparing with Horvitz-Thompson we can expect a noticeable improvement of acccuracy when $n \geq 30$ and correlation coefficient $\geq 0{,}6$

Estimation

Example: circumference of Pumpkins

Estimation

- Difference estimator

Estimation

Whereas the ratio estimator relies on a significant multiplicative relationship the difference estimator shows his strengths with significant **additive** relationships with auxiliary variables

difference estimator:

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i + \left(X - \frac{N}{n}\sum_{i=1}^{n} x_i\right)$$

Correction of the sampling estimator by the difference of the total value and the estimated total of the auxiliary variable

Estimation

- Properties of the difference estimator:

- Auxiliary variables should be similar to the target variable in terms of dimension and functionality
- Unbiased

- Rule of thumb: when comparing with Horvitz-Thompson we can expect a noticeable improvement of acccuracy when $\rho_{xy} > 0{,}5\,\frac{sd(y)}{sd(x)}$

Estimation

Example: Turnover of the last 50 years

Estimation

- Regression estimator

Estimation

A further improvement can be expected by using a (generalised) regression estimator (GREG)

The regression estimator uses the linear relationship between the target variable y and an auxialiary variable x (here: mean estimator):

$$\bar{y}_r = \bar{y} + b(\bar{X} - \bar{x})$$

$\bar{y}$ and $\bar{x}$ are Horvitz-Thompson estimates for the sample variables, and $\bar{X}$ the true population value of the auxiliary variable.

We can interprete this estimator as a correction of the sample mean in relation to the auxiliary variable

The parameter b can be calculated in various ways, usually estimated trough the sample valuesw

Estimation

- Properties of the regression estimator:

- It combines the difference estimator and the ratio estimator
- For small samples this estimator is particuliar sensitive to outliers!

Estimation

Example: Weight and height

Estimation

The generalised regression estimator (**G**eneralized **Reg**ression Estimator, GREG) includes additionally certain different sample weights and can be displayed as:

$$\hat{t}_{GREG} = \sum_{i=1}^{n} w_i y_i + \widehat{\beta}' \left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{n} w_i x_i \right)$$

with

$$\hat{\beta} = \left(\sum_{i=1}^{n} w_i x_i x'_i\right)^{-1} \left(\sum_{i=1}^{n} w_i x_i y_i\right)$$

In fact we are only estimating directly the regression parameter and not the target variable itself, which is then derived by the regression relationship

Estimation

The variance of the GREG-estimator relys highly on the correlation between the target variable and the auxiliary variable(s).

A simplified version of the variance for the GREG-Estimator in the case of only one auxiliary variable leads to

$$V(\hat{t}_{GREG}) = \frac{S_Y^2}{n} \cdot \left(1 - \frac{n}{N}\right) \cdot (1 - \rho^2)$$

Estimation

- Small area estimation

Estimation

- Small area: what does that mean?

- Small area refers to a stratum/subgroup etc. where only very few sample units exist

- This can occur if
  - The original sample was not planned for this kind of subgroups
    - Those subgroup sample areas are called then „unplanned domains“
  - a high non-response rate leaves us with few data points in this stratum/subgroup

- Example: sample n=1000 pps on governate strata

Estimation

Small area: what does that mean?

- Area is hereby not necessary a geographical unit
- Example: sample n=200 of salamanders

Estimation

Small area: what is the consequence for estimation?

Few sample units means

- Inaccurate estimators (since variance is expected to be very large)
- If only 1 or even no sampling units exist, classical estimation is impossible

- -> results from classical estimates are unreliable and cannot be calculated or published!

Estimation

Idea of small area estimation:
„Borrowing strength“ by:

- Using/Adding auxiliary or proxy variables which are available
  - On a higher aggregated level,
  - Example: muncipalities – länder,
  - Drawback: with this choice, the special properties of the target municipality are thereby often levelled

  - Not in the target area but in an area with very similar properties and high correlation to the target area
  - Sea resort town – all holiday resort towns (including ski resort towns), small village – neighbouring village(s)

  - from the same area (compare to regression estimation)

Estimation

Idea of small area estimation:

Crucial for successful small area estimation:

- Auxiliary or proxy variables possess a high correlation with the target variable

Estimation

small area estimation is largely model-based:

- There is an error which has to be taken in account according to model-misspecification
- In practice: the model is believed to be „true“ and no specific misspecification error will be introduced,
- Although: contradictory to one of the most important assertion of model-based statistics:

  - **„All models are wrong, but some are useful“**

Estimation

2 simple and popular models:

1) Unit-level model

$$y_d = x_d'\beta + e_{i,d} \text{ mit } e_{i,d} \sim \text{iid N}(0; \sigma_e^2)$$

With d domain and auxiliary information available for every sample unit

Estimation

2 simple and popular models:

2) Area-level model

$$y_d = x_d'\beta + e_d \quad \text{mit } e_d \sim \text{iid N}(0; \frac{\sigma_e^2}{n_d})$$

With d domain and auxiliary information available only as a total for the area

Important: the regression parameter β will be calculated according to the aggregated areas for stabilisation purposes

Estimation

small area estimation is largely model-based:

- Since aggregated data of the domains or even the whole population is used to extract an estimate for the target domain, we call this type of estimators **synthetic estimators**
- If, for all domains d the relationship between the auxiliary variable and the target variable remained equal, then this type of synthetic estimators would be unbiased and efficient.
- This is rarely the case

Estimation

small area estimation is largely model-based:

- Therefore a domain-specific factor u_d will be added to the equation which leads to

$y_d = x'_d\beta + u_d + e_{i,d}$ mit $u_d \sim$ iid N(0; σ_u^2) und $e_{i,d} \sim$ iid N(0; σ_e^2)
For the unit-level model and

$y_d = x'_d\beta + u_d + e_d$ mit $u_d \sim$ iid N(0; σ_u^2) und $e_d \sim$ iid N(0; $\frac{\sigma_e^2}{n_d}$)
For the area-level model

Estimation

small area estimation is largely model-based:

- **Battese, Harter und Fuller** (1988) introduce for the **Unit-Level-Model** an **EBLUP** which is the most popular approach in the literature for the mean of y:

$$\hat{\bar{y}}_d^{BHF} = \overline{X'_d}\hat{\beta} + \hat{u}_d \quad with$$

$$\hat{u}_d = \hat{\gamma}_d\left(\bar{y}_d - \overline{x'_d}\hat{\beta}\right) \quad and \quad \gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}} .$$

Estimation

small area estimation is largely model-based:

After a little cosmetic we can display this estimator in the following form:

$$\hat{\bar{y}}_d^{BHF} = \hat{\gamma}_d\left(\bar{y}_d + (\bar{X}_d - \bar{x}_d)'\hat{\beta}\right) + (1-\hat{\gamma}_d)\,\overline{X'_d}\hat{\beta}$$

As a composite estimator by a weighted sum of a direct GREG-estimator and a synthetic estimator.

Using our sampling estimation notation, the first sum can be written as

$$\sum_{i=1}^{n_d} w_{i,d}y_{i,d} + \widehat{\beta}'\left(\sum_{i=1}^{N_d} X_{i,d} - \sum_{i=1}^{n_d} w_{i,d}x_{i,d}\right)$$

Estimation

small area estimation is largely model-based:

Properties of the BHF estimator:

- If the fraction of the variance of u_d in relation to the overall model variance is large, we can assume a large difference i domains regarding the relationshoip between target and auxiary variables
- Together with a large domain sample size n_d yields a high weight factor for the direct GREG estimation component co the synthetic estimation component
- Since the synthetic estimation component is usually biased, the complete composite estimation estimator will be biase
- Therefore for quality assessment purposes we don‘t compare variances but the Mean Squared Error (MSE), or the Re Root Mean Squared Error (RRMSE), respectively.
- The estimation of MSE or RRMSE is highly complicated, usually based on simulation procedures

Estimation

small area estimation is largely model-based:

- The most popular **area-level-model estimator** was introduced by **Fay und Herriot** (1979)
- This estimator corresponds mainly to the Battese, Harter, Fuller estimator for the unit-level-model.

Estimation

small area estimation is largely model-based:

Properties of the FH-estimator:

- Since auxiliary variables are only available on domain level, we cannot use the GREG-estimator for the direc estimation part.
- Therefore the GREG will be replaced by the Horvitz-Thompson estimator.
- The error terms e_d regarding the regressions model $y_d = x_d'\beta + u_d + e_d$ measure only the errors in the sums, for the singel units.
- If the underlying model is not appropriate we can get serious bias introduced by the synthetic component (va also for the BHF-estimator)

Estimation

Conclusion:

- The (Relative) Root Mean Squared Error (RRMSE) of the introduced composite estimators can be significant smaller than the (relative) Standard Error of the direct estimator.
- Even with a domain sample size of one or zero, we can still estimate the domain total/mean (by exclusively making use of the synthetic estimation part)

Estimation

Literature:

- Rao, J. N. K. (2005): *Small Area Estimation*
- Rao, J. N. K., Molina, I. (2015): *Small Area Estimation*

Area-Level:

- Fay, R. E., Herriot, R. A. (1979): *Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data* in Journal of the American Statistical Association, 74, S. 269-277

Unit-Level:

- Battese, G. Harter, R., Fuller, W. A. (1988): *An Error-Components Modell for Predictions of County Crop Area Using Survey and Satellite Data* in Journal of the American Statistical Association, 83, S. 28-36