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 – **Indirect estimators**

Recap

- Target parameters in the domain:

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

Inference framework

Further we have, depending on the "reference framework" for inference:

— **Design Based Approach:** Estimator properties are assessed with respect to the sampling design (see previous example). This framework is used for small area estimation, mainly because of its simplicity.

— **Model Assisted Approach:** In practice, the values of Y are typically defined by assuming a model for the distribution of Y given X. That is, practitioners have been willing to use models in order to identify optimal strategies for estimating T_Y. However, their assessment of these strategies remain design-based (Särndal, Swensson and Wretman, 1992).

— **Model Based Approach:** design-unbiasedness is no longer a requirement, the alternative property we require of the estimator under this approach is that it be model-unbiased $E\left(\hat{T}_Y - T_Y \mid \mathbf{S}, \mathbf{X}\right) = 0$. given the sample S and aux info X.

What are we modelling?

We are modelling the relationship between an outcome and the auxiliary variables

note that:

- unplanned domains=geographical domains= areas

- notation:

Y_{ij} outcome= the value of the study variable (*income survey data unit j, individual or household, in area i*)

$\hat{\theta}_i^{dir}$ outcome= survey direct estimator (*per capita income in area i, total income in area i*)

What are we modelling?

The models are classified into two broad types:

1 Aggregate level (or area-level) models that relate the small area outcome (means, totals) to area-specific auxiliary variables. Such models are essential if unit level data are not available

2 Unit level models that relate the outcome (unit values of the study variable) to unit-specific auxiliary variables

The use of *explicit models* offers several advantages

What are we modelling?

Advantages both for are-level and unit-level models:

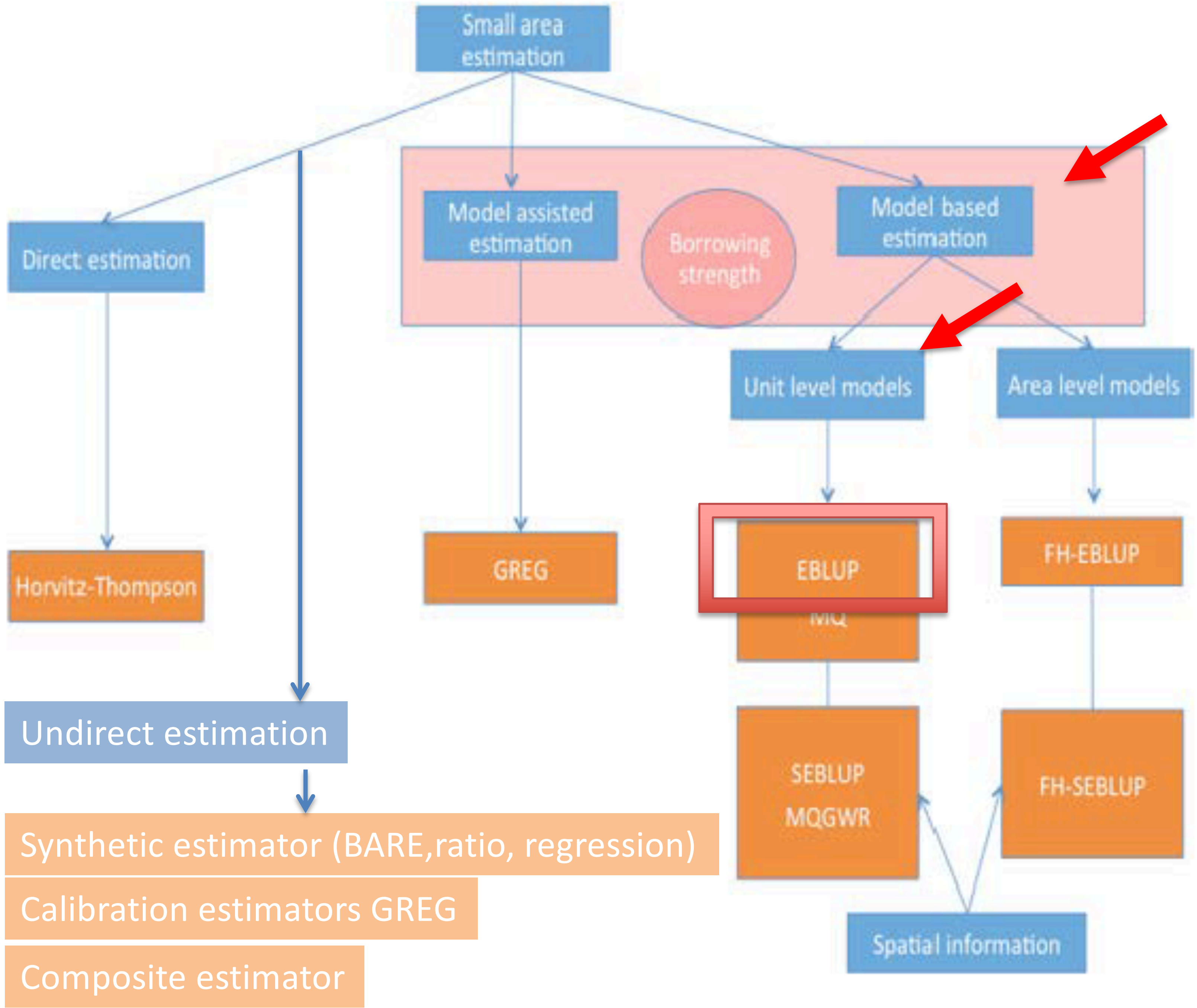
1 Model diagnostics can be used to find suitable models that fit the data well

2 Area-specic measures of precision can be associated with each small area estimate, solving the problem of instability seen for synthetic and composite estimators

3 Linear mixed models as well as nonlinear models can be used.

4 Complex data structures, such as spatial dependence and time series structures, can also be handled

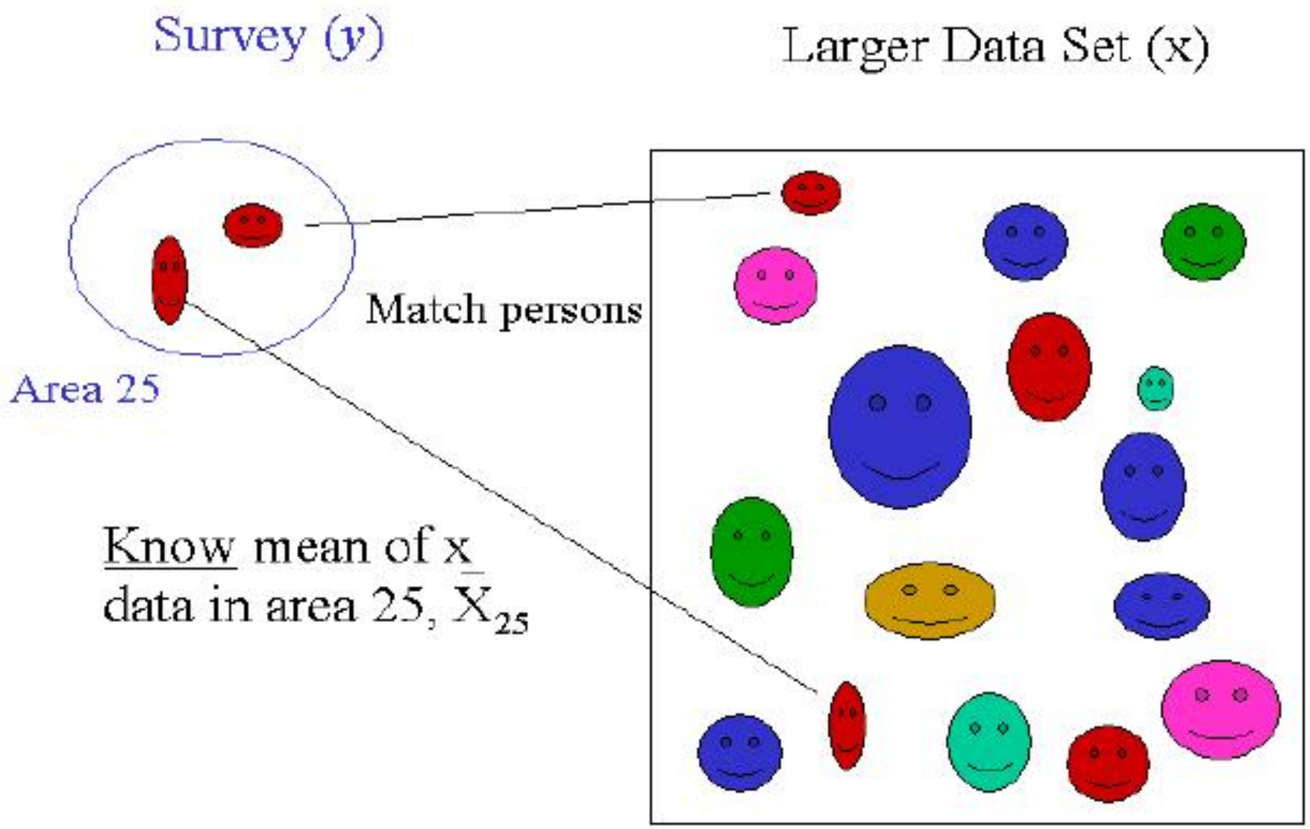
5 Methodological developments for random effects models can be utilized to achieve accurate small area inferences


Small area estimation
Direct estimation
Model assisted estimation
Borrowing strength
Model based estimation
Unit level models
Area level models
Horvitz-Thompson
GREG
EBLUP
FH-EBLUP
SEBLUP
MQGWR
FH-SEBLUP
Spatial information
Undirect estimation
Synthetic estimator (BARE,ratio, regression)
Calibration estimators GREG
Composite estimator

Unit Level Approach

1 - outcome : y_{ij} , auxiliary variable *X* available at unit level (j) from a larger data set
Ex: unit=individual, area=province: y_{ij} income of unit, X household size

Pisa
province



EBLUP: Unit level approach

- $\mathbf{y}$ the vector for the y variable for the population Ω
- $\mathbf{y} = [\mathbf{y}'_s, \mathbf{y}'_r]'$, where $\mathbf{y}_s$ is the vector of the observed units (the sampled ones) and $\mathbf{y}_r$ is the vector of the non observed units ($N - n$, $r = 1, \ldots, N - n$)
- $\mathbf{X}$ is the covariates matrix and is considered know for all the population units
- Subscript i refers to small areas (e.g. $\mathbf{y}_{s_i}$ is the vector of observed variables in area i)

$$\boxed{\mathbf{y} = [\mathbf{y}'_s, \mathbf{y}'_r]'}$$

EUSILC: …you observe income of the units in the sample

There are other units in the area i …but they are not included in the sample: you do not observe their income

EBLUP: Unit level approach

- Model for the y variable (known as superpopulation model)

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

  - that can be alternatively write as

$$y_{ij} = \boldsymbol{x}_{ij}\beta + u_i + e_{ij}$$

In addition to the assumptions already made, we require that the model holds for both the population and the sample.

$$\boldsymbol{X}\beta \qquad \boldsymbol{Z}\boldsymbol{u}$$

Base-line part of the model

Differences among areas

EBLUP: Unit level approach

Starting with the linear regression model for grouped individuals:

$$y_{dj} = \beta_{0d} + \beta_{1d} \cdot x_{dj} + \varepsilon_{dj}$$

for the groups $d = 1, \ldots, D$ and the individuals j, we assume that $\beta_{0d} = \beta_0 + u_{0d}$ and $\beta_{1d} = \beta_1 + u_{1d}$. For the random effects u_{0d} and u_{1d} we further assume

$$\mathrm{E}(u_{0d}) = \mathrm{E}(u_{1d}) = 0 \quad \text{and}$$
$$\mathrm{V}(u_{0d}) = \sigma^2_{u0}, \quad \mathrm{V}(u_{1d}) = \sigma^2_{u1}, \quad \mathrm{Cov}(u_{0d}, u_{1d}) = \sigma_{u01}.$$

This yields

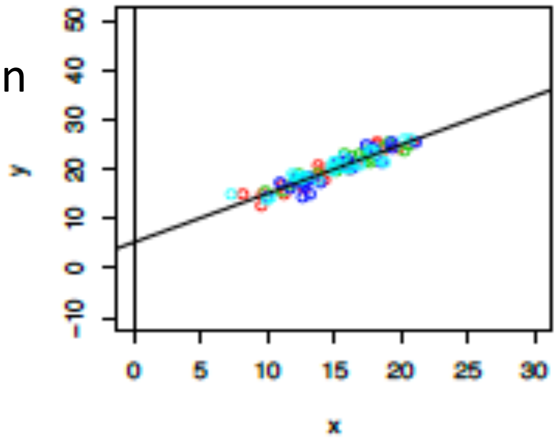
$$y_{dj} = \overbrace{\beta_0 + \beta_1 \cdot x_{dj}}^{\boldsymbol{X}\beta} + \left(\overbrace{u_{0d} + u_{1d} \cdot x_{dj}}^{\boldsymbol{Z}\boldsymbol{u}} + \varepsilon_{0dj}\right)$$
$$\mathrm{V}(\varepsilon_{0dj}) = \sigma^2_{\varepsilon 0} \quad .$$

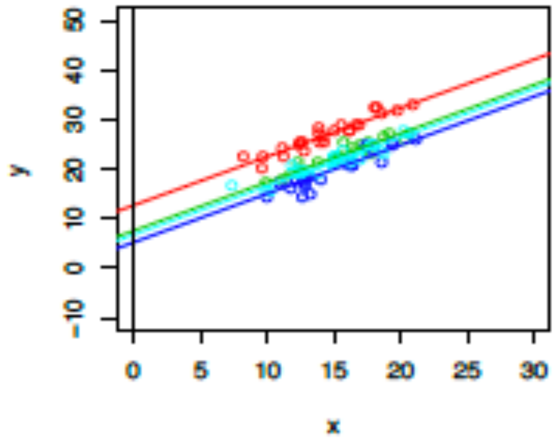
Moreover, we require independence between ε_{0dj} and both random effects.

EBLUP: Unit level approach

Different models (colours are identifying groups of units:

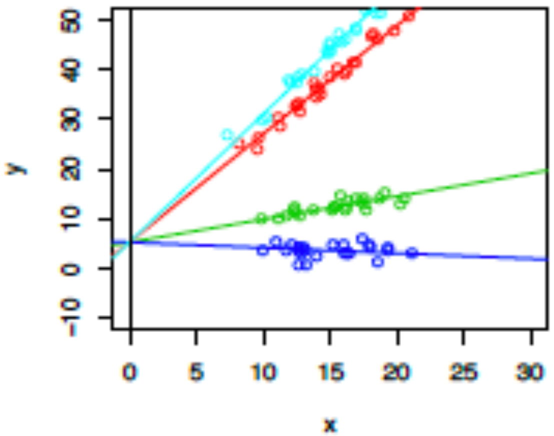
Regression

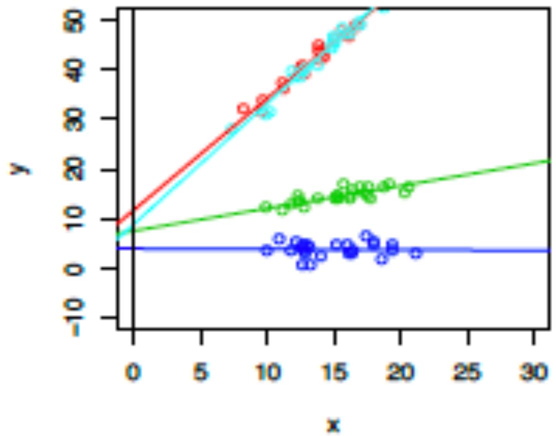




Different intercepts

Different slopes





Different Intercepts and slopes

EBLUP: Unit level approach

Random intercept model:

- Assume that only the intercept is a random component on the second level.
- This is the by far most common choice in small area estimation.
- It implies $u_{1d} \equiv 0$ and hence

$$y_{dj} = \mathbf{x}'_{dj}\beta + u_d + \varepsilon_{dj}$$

- The covariance of any two units follows as:

$$\mathrm{Cov}\left(y_{dj}, y_{d'j'}\right) = \begin{cases} \sigma^2_{u0} + \sigma^2_{\varepsilon 0}, & \text{if } d = d' \text{ and } j = j', \\ \sigma^2_{u0}, & \text{if } d = d' \text{ and } j \neq j', \\ 0,, & \text{if } d \neq d' \text{ and } j \neq j' \end{cases}$$

- Hence, two units from the same group d will be correlated, whereas units from different groups are independent.

Battese Harter Fuller Model

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988) *An error-components model for prediction of county crops using survey and satellite data*. Journal American Statistical Association 83 28-36.

- This yields

$$y_{dj} = \mathbf{x}_{dj}'\beta + u_d + \varepsilon_{dj}, \quad d = 1, \ldots, D, j = 1, \ldots, N_d, \qquad (1)$$

$$u_d \overset{iid}{\sim} N(0, \sigma_u^2)$$

$$\varepsilon_{dj} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

  with $\mathbf{x}_{dj} = (x_{dj1}, \ldots, x_{djp})'$ as a $p \times 1$ column vector of the covariates for the j-th unit within the d-th area.

- We estimate β by GLS, i.e

$$\widehat{\beta} = \left[\sum_{d=1}^{D}\sum_{j=1}^{n_d}\mathbf{x}_{dj}(\mathbf{x}_{dj} - \gamma_d\bar{\mathbf{x}}_d)'\right]^{-1}\left[\sum_{d=1}^{D}\sum_{j=1}^{n_d}(\mathbf{x}_{dj} - \gamma_d\bar{\mathbf{x}}_d)'y_{dj}\right]$$

Battese Harter Fuller Model

- Under model (1) the small area means are $\boxed{\mu_d = \overline{\mathbf{X}}'_d\beta + u_d + \overline{\varepsilon}_d}$ (cf. Pfeffermann (2013))
- This may be approximated as $\mu_d \approx \overline{\mathbf{X}}'_d\beta + u_d$
- For small sampling fractions the empirical best linear unbiased predictor follows as

$$\begin{aligned} \widehat{\mu}_d^{BHF} &= \overline{\mathbf{X}}'_d\widehat{\beta} + \widehat{u}_d \\ \widehat{u}_d &= \widehat{\gamma}_d\left(\overline{y}_d - \overline{\mathbf{x}}'_d\widehat{\beta}\right) \\ \widehat{\gamma}_d &= \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \frac{\widehat{\sigma}_\varepsilon^2}{n_d}}. \end{aligned} \tag{2}$$

- Upper-case notation refers to population values and lower-case notation to sample values; a hat indicates that the variable is estimated and a bar denotes the mean

Battese Harter Fuller Model

$$\widehat{\mu}_d^{\mathrm{BHF}} = \widehat{\gamma}_d(\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)'\widehat{\beta}) + (1 - \widehat{\gamma}_d)\overline{\mathbf{X}}_d'\widehat{\beta} \quad (3)$$

- Expressions (2) and (3) are equivalent
- As (3) indicates, the BHF-estimator may be viewed as a composite estimator of
  - the survey regression estimator (multilevel-GREG under SRS) $\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)'\widehat{\beta}$
  - and the regression-synthetic component $\overline{\mathbf{X}}_d'\widehat{\beta}$

  with weights $\widehat{\gamma}_d$ and $(1 - \widehat{\gamma}_d)$.

$$\widehat{\gamma}_d = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \frac{\widehat{\sigma}_\varepsilon^2}{n_d}}.$$

- Unlike the composite estimators in the previous lecture, the weights of the EBLUP emerge as a simple ratio of the estimated variance components.

Battese Harter Fuller Model

For finite populations with non-negligible sampling fractions equation (2) or (3) have to be replaced by:

$$\widehat{\mu}_d^{BHF} = \frac{1}{N_d}\left[\sum_{j \in S_d} y_{dj} + \sum_{j \notin S_d} \widehat{y}_{dj}\right] \text{ with} \tag{4}$$

$$\widehat{y}_{dj} = \mathbf{x}_{dj}'\widehat{\beta} + \widehat{u}_d \tag{5}$$

Equation (4) is a general representation of an empirical best predictor (EBP) as well. The EBP generally comprises two parts

1. The sum of the observations for the sampled units and
2. The sum of the predictions for the non-sampled units.

BHF model: MSE

$\hat{\theta}_i$ statistics of interest at area *i* level (or group *d* level):
say the area mean, area poverty rates , estimated by BHF model

Next step is to derive an MSE estimator

- $MSE(\hat{\theta}_i) \approx g_{1i}(\boldsymbol{\sigma}) + g_{2i}(\boldsymbol{\sigma}) + g_{3i}(\boldsymbol{\sigma}))$
- $g_{1i}(\boldsymbol{\sigma}) = \alpha_r' \boldsymbol{Z}_r \boldsymbol{T}_s \boldsymbol{Z}_r' \alpha_r$
- $g_{2i}(\boldsymbol{\sigma}) = [\alpha_r' b X_r - \alpha_r' \boldsymbol{Z}_r \boldsymbol{T}_s \boldsymbol{Z}_s' \boldsymbol{R}_s' \boldsymbol{X}_s](\boldsymbol{X}_s' \boldsymbol{V}^{-1} \boldsymbol{X}_s)^{-1} [\boldsymbol{X}_r' \alpha_r - \boldsymbol{X}_s' \boldsymbol{R}_s^{-1} \boldsymbol{Z}_s \boldsymbol{T}_s \boldsymbol{Z}_r' \alpha_r]$
- $g_{3i}(\boldsymbol{\sigma}) = tr\{(\nabla(\alpha_r' \boldsymbol{Z}_r \boldsymbol{\Sigma}_u \boldsymbol{Z}_s' \boldsymbol{V}_s^{-1})') \boldsymbol{V}_s (\nabla(\alpha_r' \boldsymbol{Z}_r \boldsymbol{\Sigma}_u \boldsymbol{Z}_s' \boldsymbol{V}_s^{-1})')' E[(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})']\}$
- $\boldsymbol{T} = \boldsymbol{\Sigma}_u - \boldsymbol{\Sigma}_u \boldsymbol{Z}_s' (\boldsymbol{\Sigma}_{es} + \boldsymbol{Z}_s \boldsymbol{\Sigma}_u \boldsymbol{Z}_s')^{-1} \boldsymbol{Z}_s \boldsymbol{\Sigma}_u$
- $\boldsymbol{\sigma} = (\sigma_e^2, \sigma_u^2 / \sigma_e^2)'$

BHF model: MSE

Finally, the estimator for the MSE of $\hat{\theta}_i$ is

$$\widehat{MSE}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}) + g_{2i}(\hat{\sigma}) + 2g_{3i}(\hat{\sigma})$$

- $\hat{\sigma}$ is an unbiased estimator for σ

Remark: it is possible to obtain an estimate of the MSE using alternative techniques, such as bootstrap and jackknife

BHF model: recap

- If $\widehat{\sigma}_u^2$ is small, a small value $\widehat{\gamma}_d$ results as well and more weight will be given to the regression synthetic component.
  The same holds if the area specific sample sizes n_d is very small.
- The weighting factor will tend to 1 and more weight will be given to the survey regression estimators, if the random effects variance is large The same holds for large sample sizes n_d.
- The BHF-estimator is not design-consistent for general survey designs $\longrightarrow$ Survey regression estimator does not consider design-weights
- The EBLUP is not model unbiased when conditioning on u_d, as this implies assuming fixed intercepts in the different areas (see Pfeffermann 2013).

Statistical Science
2013, Vol. 28, No. 1, 40–68
DOI: 10.1214/12-STS395

- Examples of application of the BHF EBLUP estimator during the R lab

Pros and cons of the model will be discussed after examples and applications to real data