Recap on small area estimation

Monica Pratesi - based on April 2016 short course by S. Marchetti

University of Pisa, Italy

Jean Monnet Chair SAMPIEU - a.y. 2018/2019

Outline

1. Introduction to Small Area Estimation

2. Direct Estimators

3. Synthetic Estimators

4. Composite Estimators

Part I

Introduction to Small Area Estimation

Introduction to Small Area Estimation

- Problem: demand from official and private institutions of statistical data referred to a given population of interest
- Possible solutions:
  - Census
  - Sample survey

Sample surveys have been recognized as cost-effectiveness means of obtaining information on wide-ranging topics of interest at frequent interval over time

Introduction to Small Area Estimation

- Population of interest (or target population): population for which the survey is designed
- Domain: sub-population of the population of interest, they could be planned or not in the survey design
  - Geographic areas (e.g. Regions, Districts, Ward, Villages)
  - Socio-demographic groups (e.g. Gender, Age, Religion)
  - Other sub-populations (e.g. the set of firms belonging to a industry subdivision)

→we don't know the reliability of estimates of the domains/area that have not been planned in the survey design

Introduction to Small Area Estimation

- Often estimates based only on sample data are not reliable for some areas/domains of interest
- In these cases we have two choices:
  - oversampling over that areas
  - applying statistical techniques that allow for reliable estimates in that areas/domains

Small Domain or Small Area

Geographical area or domain where direct estimators do not reach a minimum level of precision

Introduction to Small Area Estimation: Example

- US Survey sample sizes with an equal probability of selection method, sample of 10,000 persons

Table : Sample and population size by State (US 1994)

State	1994 Population (thousands)	Sample size
California	31,431	1207
Texas	18,378	706
New York	18,169	698
⋮	⋮	⋮
DC	570	22
Wyoming	476	18

- Customer satisfaction for a government service:
  - California 24.86% → 95% C.I. $[22.4\%, 27.3\%]$ and CV=0.05 → reliable
  - Wyoming 33.33% → 95% C.I. $[11.5\%, 55.1\%]$ and CV=0.33 → unreliable

Introduction to Small Area Estimation: Example

- Target population: farmers in Tanzania Mainland
- Variable of interest: Production of maize in 2015
- Survey sample: Annual Agricultural Sample Survey (AASS), designed to obtain reliable estimate at Regional level in Tanzania Mainland
  - planned design domains: Regions
  - unplanned design domains: e.g. Districts, Wards, Villages

- Planned sample size of AASS in Morogoro: e.g. 500 farmers
  - Kilombero district 28 farmers → need SAE
  - Kilosa district 7 farmers → need SAE
  - ...

Part II

Direct Estimators

Definitions

- Direct estimator: an estimator based only on area specific sample data
- Design-based estimation: the main focus is on the design unbiasedness. Estimators are unbiased with respect to the randomization that generates survey data
- Finite population $\Omega = 1, \ldots, N$
- y: variable of interest, with y_i value of the i-th unit of the population
- Statistics of interest: e.g. total, $Y = \sum_{\Omega} y_i$ or mean, $\bar{Y} = Y/N$
- Sample $s = 1, \ldots, n$
- $p(s)$: probability of selecting the sample s from population Ω. $p(s)$ depends on know design variables such as stratum indicator and size measures of clusters

Definitions

- Consider an estimator $\hat{\theta}$ of θ
- Design bias: $Bias(\hat{\theta}) = E_p[\hat{\theta}] - \theta$
- Design variance: $V(\hat{\theta}) = E_p[(\hat{\theta} - \theta)^2]$
- Design Mean Squared Error: $V(\hat{\theta}) + B(\hat{\theta})^2$

Design-based properties

1. Design-unbiasedness: $E_p[\hat{\theta}] = \sum p(s)\hat{\theta}_s = \theta$
2. Design-consistency: $\hat{\theta} \to \theta$ in probability

Estimation of Means: Direct Estimation

- Sample data $\{y_i\}, i \in s$
- Direct estimator for the mean θ:

$$\hat{\theta}^{dir} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$$

- $w_i = \pi_i^{-1}$, the basic design weight
- π_i is the probability of selecting the unit i in sample s

Remark: weights w_i are independent from y_i

Small Area Estimation

- Let partitioning population Ω into m areas:

$$\Omega = \cup_{i=1}^{m} \Omega_i$$

- $\Omega_i = 1, \ldots, N_i$, population of the domain i
- $s_i = 1, \ldots, n_i$, sample of the domain i
- Statistics of interest for the variable y:

$$\theta_i = \frac{1}{N_i} \sum_{j \in \Omega_i} y_j$$

Small Area Estimation: Direct Estimator

- Sample data $\{y_{ij}\}, j \in s_i, i = 1, \ldots, m$
- Direct estimator of the mean for area i:

$$\hat{\theta}_i^{dir} = \frac{\sum_{j \in s_i} w_{ij} y_{ij}}{\sum_{j \in s_i} w_{ij}}$$

- $w_{ij} = \pi_{ij}^{-1}$ is the weight for unit j in area i
- π_{ij} is the inclusion probability of unit j in area i
- The case of the simple random sampling within area i (SRS):
    - $\pi_{ij} = \pi_i = \frac{\binom{1}{1}\binom{N_i-1}{n_i-1}}{\binom{N_i}{n_i}} = \frac{n_i}{N_i} \quad \rightarrow \quad w_{ij} = \pi_i^{-1} = \frac{N_i}{n_i}$
    - $\hat{\theta}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \frac{\sum_{j=1}^{n_i} \frac{N_i}{n_i} y_{ij}}{\sum_{j=1}^{n_i} \frac{N_i}{n_i}} = \frac{\frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}}{n_i \frac{N_i}{n_i}} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ (that is the sample mean)

Small Area Estimation: Direct Estimator

- $\hat{\theta}_i^{dir}$ is design unbiased
- Sampling variance:
  $v(\hat{\theta}_i^{dir}) = \left(\sum_{j=1}^{n_i} (w_{ij}^2 - w_{ij})(y_{ij} - \bar{y}_i)^2 \right) / \left(\sum_{j=1}^{n_i} w_{ij} \right)^2$
  - SRS: $v(\hat{\theta}_i^{dir}) = (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$, $S_i^2 = \frac{\sum_{j \in s_i} (y_{ij} - \bar{y}_i)^2}{n_i}$
- The magnitude of the variance depends on: w_{ij}, S_i^2 and n_i
- If n_i is small the design variance is likely to be large

Remark: In such a situation, estimation of variance is even more problematic

Direct estimation, example

- Synthetic population of 1.6mln farms generated using Tanzania Agricultural Census 2007/08
- Target variable: production of *maize*
- Sample: stratified simple random sample
    - Strata: 99 districts
- Distribution of sample and population sizes of the 99 districts

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n_i	3	24	50	38.5	50	50
N_i	29	485	6048	16320.0	25970	105000

Direct estimation, example

Direct estimates of maize production (kg) at district level

	District	n_i	Estimate	SD	CV%
1	Arusha 1	50	692.6	135.2	19.5
2	Arusha 3	50	625.3	104.3	16.7
3	Arusha 4	50	653.5	105.5	16.1
4	Arusha 5	28	699.6	187.6	26.8
5	Arusha 6	50	613.5	111.2	18.1
6	Arusha 7	50	587.2	96.9	16.5
...					
98	Tanga 8	50	721.2	123.0	17.1
99	Urban West 1	50	275.6	41.9	15.2

Direct estimation, example

- As the sample size decrease, the CV increase

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$CV\%$	15.1	16.6	17.9	24.1	24.0	66.0

- U.S Census Bureau want majority of the CVs of key estimates to be $< 30\%$ (`www.census.gov/quality/standards/standardf1.html`)
- In this example 21% of estimates have a $CV > 30\%$
- Often, National Statistical Institute consider a very good estimates those estimates with a $CV \leq 16\%$
- In this example only 7% of estimates have a $CV \leqslant 16\%$

Part III

Synthetic Estimators

Synthetic Estimators

- Synthetic assumption: small areas have same characteristic as the large area (e.g. maize production for different districts is the same as that for Tanzania)
- Advantages of synthetic estimator:
  - Simple and intuitive
  - Applies to general sampling designs
  - Borrow strength from similar
  - Provides estimates for areas with no sample from the sample survey

Synthetic Estimation with no auxiliary variable (dummy estimator)

- Implicit model assumed:

$$y_j = \alpha + \varepsilon_j, \quad j \in \Omega$$

- Synthetic estimator for the mean:

$$\hat{\theta}_i^{syn} = \hat{\theta}^{dir} = \frac{\sum_{j \in s} w_j y_j}{\sum_{j \in s} w_j}$$

- $E_p[\hat{\theta}_i^{syn} - \theta_i] \approx \theta - \theta_i$, the bias relative to the parameter θ_i is small if $\theta_i \approx \theta$
- This synthetic estimator is very efficient if the small area mean is approximately equal to the overall mean
- It can be heavily biased for areas exhibiting strong individual effects which can lead to large MSE

Synthetic Estimation with auxiliary variables

- Area-specific auxiliary information available, $\mathbf{X}_i$
- Implicit model assumed:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \varepsilon_{ij}, \quad j \in \Omega_i$$

- Synthetic estimator:

$$\hat{\theta}_i^{syn} = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}}$$

- $\hat{\boldsymbol{\beta}} = (\sum_{j\in s} w_j \mathbf{x}_j \mathbf{x}_j^T / c_j)^{-1} (\sum_{j\in s} w_j \mathbf{x}_j y_j / c_j)$

Synthetic Estimation with auxiliary variables

- $E_p[\hat{\theta}_i^{syn} - \theta_i] \approx \bar{\mathbf{X}}_i^T \boldsymbol{\beta} - \theta_i$, expected bias
- $\boldsymbol{\beta} = (\sum_{j \in \Omega} \mathbf{x}_j \mathbf{x}_j' / c_j)^{-1} (\sum_{j \in \Omega} \mathbf{x}_j y_j / c_j)$
- The relative bias is small if both of the following conditions are satisfied

    i. $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, where
    $\boldsymbol{\beta}_i = (\sum_{j \in \Omega_i} \mathbf{x}_{ij} \mathbf{x}_{ij}^T / c_{ij})^{-1} (\sum_{j \in \Omega_i} \mathbf{x}_{ij} y_{ij} / c_{ij})$

    ii. $\theta_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta}_i$
- The synthetic estimator will be very efficient when the small area i does not exhibit strong individual effect with respect to the regression coefficient
- If $c_j = \nu^T \mathbf{x}_j$, then $\hat{\theta}_i^{syn}$ add up to GREG estimator (i.e. $\hat{\theta}^{GREG} = \bar{\mathbf{X}}^T \hat{\beta}$)

Synthetic Estimation, MSE

- MSE of synthetic estimator

$$MSE(\hat{\theta}_i^{syn}) = E_p[(\hat{\theta}_i^{syn} - \theta_i)]^2 = E_p[(\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir})^2] - V_p(\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir}) + V_p(\hat{\theta}_i^{syn})$$

- Estimator of the MSE

$$mse(\hat{\theta}_i^{syn}) = (\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir})^2 - v(\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir}) + v(\hat{\theta}_i^{syn})$$

- $mse(\hat{\theta}_i^{syn})$ is approximately unbiased, but very unstable and can take negative values
- An alternative is $mse(\hat{\theta}_i^{syn}) \approx (\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir})^2 - v(\hat{\theta}_i^{dir})$
- Many other alternatives exist in literature

Synthetic estimator, example

- Synthetic population of 1.6mln farms generated using Tanzania Agricultural Census 2007/08
- Target variable: production of *maize*
- Sample: stratified simple random sample
  - Strata: 99 districts
- Distribution of sample and population sizes of the 99 districts

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n_i	3	24	50	38.5	50	50
N_i	29	485	6048.0	16320.0	25970	105000

Synthetic estimator, example

- Estimate the average production of maize in 99 districts in Tanzania Mainland
- Sample data available for each districts with sample size n_i
- y_{ij}: production of maize (kg) of farm j in district i (sample data)
- $x_{1,ij}$ and $x_{2,ij}$ are the agricultural surface and cost of farm j in district i (sample data)
- The district average of agricultural surface ($\bar{X}_{1,i}$) and cost ($\bar{X}_{2,i}$) is considered known, $i = 1, \ldots, 99$ (Census data)
- Assumption of linear relation between target and predictors

$$y_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,ij} + \hat{\beta}_2 x_{2,ij} + \varepsilon_{ij} \, , \tag{1}$$

$\varepsilon_{ij} \overset{i.i.d.}{\sim} (0, \sigma^2_\varepsilon)$

Synthetic estimator, example

- From (1) estimate β_0, β_1 and β_2 with OLS
- The synthetic estimator of the average production of maize for district i is
$$\hat{\theta}_i^{syn} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_{1,i} + \hat{\beta}_2 \bar{X}_{2,i}$$
- Note: it is *not* necessary to know $x_{1,ij}$ and $x_{2,ij}$ for all the farms in the district i
- Remark: $(1/99) \sum_{i=1}^{99} \hat{\theta}_i^{syn} <,=,> \hat{\theta}^{dir}$
    - A simple solution is a ratio adjustment

$$\hat{\theta}_i^{syn-a} = \frac{\hat{\theta}_i^{syn}}{\sum_{i=1}^{99} \hat{\theta}_i^{syn}}$$

Synthetic estimator, example

Synthetic estimates of maize production (kg) at district level

	District	n_i	Estimate	SD	CV%
1	Arusha 1	50	672.0	20.6	3.1
2	Arusha 3	50	682.5	57.3	8.4
3	Arusha 4	50	652.5	1.0	0.2
4	Arusha 5	28	670.6	29.1	4.3
5	Arusha 6	50	658.0	44.4	6.8
6	Arusha 7	50	660.6	73.4	11.1
...					
98	Tanga 8	50	770.6	49.4	6.4
99	Urban West 1	50	282.2	6.6	2.3

Synthetic estimator, example

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$CV\%, dir$	15.1	16.6	17.9	24.1	24.0	66.0
$CV\%, syn$	0.2	4.3	7.9	12.4	16.1	95.1

- 9% of synthetic estimates have a $CV > 30\%$
- 75% of synthetic estimates have a $CV \leqslant 16\%$

Part IV

Composite Estimators

Composite Estimators

A composite estimator is an estimator that combine direct and synthetic estimator:

$$\hat{\theta}_i^{com} = \phi_i \hat{\theta}_i^{dir} + (1 - \phi_i)\hat{\theta}_i^{syn}$$

where

- $\hat{\theta}_i^{dir}$ is a direct estimator for the i-th small area
- $\hat{\theta}_i^{syn}$ is a synthetic estimator for the i-th small area
- ϕ_i is a suitably chosen weight, $0 \leq \phi_i \leq 1$

The aim of the composite estimator is to balance the potential bias of the synthetic estimator against the instability of the design-based estimator

The Choice of ϕ_i

Sample size dependent estimate

$$\phi_i = \begin{cases} 1 & \text{if } (N/n)n_i \geq \delta N_i \\ (N/n)n_i/(\delta N_i) & \text{otherwise} \end{cases}$$

where δ is subjectively chosen ($\delta \in [2/3, 3/2]$ in most practical situation)

- Consider a SRS of size n from a population of N units and $\delta = 1$
- If $n_i \geq (N_i/N)n$ then the composite estimator reduces to direct estimator
- If $n_i < (N_i/N)n$ then the composite estimator assign a weight of $(N/n)(n_i/N)$, that is an increasing function of the sample rate within the area

The Choice of ϕ_i

Optimal ϕ_i

a. Minimize the $MSE(\hat{\theta}_i^{com})$ with respect to ϕ_i assuming $COV(\hat{\theta}_i^{dir}, \hat{\theta}_i^{syn}) \approx 0$

- the optimal solution is given by

$$\phi_i^* = \frac{MSE(\hat{\theta}_i^{syn})}{MSE(\hat{\theta}_i^{syn}) + V(\hat{\theta}_i^{dir})}$$

- the parameter ϕ_i can be estimated by

$$\hat{\phi}_i^* = \frac{mse(\hat{\theta}_i^{syn})}{(\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir})^2}$$

Note: $\hat{\phi}_i^*$ is very unstable

The Choice of ϕ_i (James-Stein method)

b. Minimize $m^{-1}\sum_{i=1}^{m} MSE(\hat{\theta}_i^{com})$ with respect to a common weight $\phi_i = \phi \ \forall \ i = 1, \dots, m$

  - the optimal solution is given by

$$\phi^* = \frac{\sum_{i=1}^{m} MSE(\hat{\theta}_i^{syn})}{\sum_{i=1}^{m}(MSE(\hat{\theta}_i^{syn}) + V(\hat{\theta}_i^{dir}))}$$

  - the parameter ϕ can be estimated by

$$\hat{\phi}^* = 1 - \frac{\sum_{i=1}^{m} \hat{v}(\hat{\theta}_i^{dir})}{\sum_{i=1}^{m}(\hat{\theta}_i^{syn} - \hat{\theta}_i^{dir})^2}$$

- The use of a common weight may not be reasonable if the individual variances vary considerably.

MSE of composite estimator

- Suppose small area means θ_i are the parameter of interest
- Let $T_i = g(\theta_i)$ be a specified transformation of θ_i such that $\hat{T}_i = g(\hat{\theta}_i dir) \overset{ind}{\sim} N(T_i, \psi_i^2)$
- ψ_i is considered known
- Assume that a prior guess of T_i, say T_i^0 is available $\forall\ i = 1, \ldots, m$
- T_i^0 can be a least square predictor or $T_i^0 = m^{-1} \sum_{i=1}^m \hat{T}_i = \hat{T}_.$
- Let $\delta_i = T_i / \psi_i$ and $\hat{\delta}_i = \hat{T}_i / \psi_i$ so that $\hat{\delta}_i \overset{ind}{\sim} N(\delta_i, 1)$
- Let $\delta_i^0 = T_i^0 / \psi_i$ be the guess of δ_i
- The transformed composite estimator is than

$$\hat{T}_i^{com} = T_i^0 + \left(1 - \frac{m-2}{S}\right)(\hat{T}_i - T_i^0), \quad m \geq 3$$

$S = \sum_{i=1}^m (\hat{T}_i - T_i^0)^2 / \psi_i^2$

MSE of composite estimator

- Write $\hat{T}_i^{com}$ as

$$\hat{T}_i^{com} = \hat{T}_i + \frac{m-2}{S}(T_i^0 - \hat{T}_i)$$

- So if T_i^0 is fixed we have

$$\begin{aligned} MSE(\hat{T}_i^{com}) &= E_p[\hat{T}_i + \frac{m-2}{S}(T_i^0 - \hat{T}_i) - T_i]^2 \\ &= E_p\left[\psi_i^2 - 2\psi_i^2 \frac{m-2}{S}\left(1 - \frac{2(\hat{T}_i - T_i^0)^2}{\psi_i^2 S}\right)\right. \\ &\left. + \frac{(m-2)^2}{S^2}(T_i^0 - \hat{T}_i)^2\right] \end{aligned}$$

MSE of composite estimator

- An unbiased estimator of $MSE(\hat{T}_i^{com})$ is

$$mse(\hat{T}_i^{com}) = \psi_i^2 + 2\psi_i^2 \frac{m-2}{S}\left(1 - \frac{2(\hat{T}_i - T_i^0)^2}{\psi_i^2 S}\right) + \frac{(m-2)^2}{S^2}(T_i^0 - \hat{T}_i)^2$$

- $mse(\hat{T}_i^{com})$ can take negative value, so a better estimator is $mse^+(\hat{T}_i^{com}) = \max(0, mse(\hat{T}_i^{com}))$
- The CV derived from $mse(\hat{T}_i^{com})$ can be quite large, that is, it can be very unstable

Model-based small area estimators overcome this problem. However, they can be biased in the design-based framework

Recap Composite estimation

- The estimator $\hat{T}_i^{com}$ is a composite estimtor

$$\hat{T}_i^{com} = \hat{\phi}\hat{T}_i + (1-\hat{\phi})T_i^0$$

- $\hat{\phi} = 1 - \frac{m-2}{S}$
- The more common transform $g(\cdot)$ are the following
  - $g(\theta_i) = \arcsin\theta_i$
  - $g(\theta_i) = \theta_i$
  - $g(\theta_i) = \ln\theta_i$

Comparison Between Direct, Synthetic and Composite Estimator

Empirical comparison of small area estimation methods for the Italian Labor Force Survey (LFS)

- Performance of small area estimators are studied by simulating sample from 1981 Population Census. Samples are drown following the LFS design (two stages with stratification)
- 400 sample replicates, each of identical size of the LFS sample
- 14 Health Service Areas (HSA) of the Friuli Region are considered to be small areas

Comparison Between Direct, Synthetic and Composite Estimator

Index used to evaluate the performances of the estimators

- Absolute Relative Bias

$$ARB = \frac{1}{14}\sum_{i=1}^{14}\frac{1}{400}\sum_{h=1}^{400}\left|\frac{\hat{\theta}_i^{(h)} - \theta_i}{\theta_i}100\right|$$

- Relative Root MSE

$$RRMSE = \frac{1}{14}\sum_{i=1}^{14}\left(\frac{\sqrt{\frac{1}{400}\sum_{h=1}^{400}(\hat{\theta}_i^{(h)} - \theta_i)^2}}{\theta_i}100\right)$$

Comparison Between Direct, Synthetic and Composite Estimator

ARB and RRMSE for Direct, Synthetic and Composite estimators

Table : Estimators performances

Estimator	ARB	RRMSE
Direct	2.39	31.08
Synthetic	8.97	23.80
Composite	6.00	23.57

Note: the RRMSE of Direct estimator is approximatively 30% higher than Synthetic and Composite estimator

Part V

Model-based estimators

Model-based estimators

- Synthetic and composite estimators are based on implicit models that provide a link to related small areas through supplementary data
- Small area model-based estimators explicit small area models that make specific allowance for between area variation
- In this framework models involve random area-specific effects that account for between area variation beyond that explained by auxiliary variables included in the model
- The success of any model-based method depends on the availability of good auxiliary data
- Subject matter specialists or end users should have influence on the choice of models, particularly on the choice of auxiliary variables

Model-based estimators

The use of explicit models offers several advantages:

1. Model diagnostics can be used to find suitable models that fit the data well
2. Area-specific 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 entertained. Complex data structures, such as spatial dependence and time series structures, can also be handled
4. Methodological developments for random effects models can be utilized to achieve accurate small area inferences

Model-based estimators

The models may be classified into two broad types:

1. Aggregate level (or *area-level*) models that relate the small area means to area-specific auxiliary variables. *Such models are essential if unit level data are not available*
2. Unit level models that relate the unit values of the study variable to unit-specific auxiliary variables

   In this technical workshop we will focus on area-level models

Area-level model-based estimators

Framework

- Population Ω divided into m (small) areas
- Availability of sample data on target variable, y
- m parameter of interest (e.g. mean), θ_i, $i = 1, \ldots, m$
- From sample → m direct estimates, $\hat{\theta}_i^{dir}$, $i = 1, \ldots, m$
- and m *MSE* estimates, $mse(\hat{\theta}_i^{dir}) \approx \psi_i^2$, $i = 1, \ldots, m$
    - Here ψ_i^2 is the $MSE(\hat{\theta}_i^{dir})$ and is often considered known
    - In real application we know only its estimates $mse(\hat{\theta}_i^{dir})$
- From other data sources → m p-vector of auxiliary variables, $\mathbf{x}_i$ $\forall$ $i = 1, \ldots, m$

The Fay-Herriot Model

Assumptions

1. $\hat{\theta}_i^{dir} = \theta_i + e_i$
2. $e_i \overset{iid}{\sim} N(0, \psi_i^2)$
3. $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i$
4. $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$
5. $u_i \perp e_i \; \forall \; i = 1, \ldots, m$

From (1) and (3) follow the Fay-Herriot (FH) model

$$\hat{\theta}_i^{dir} = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i$$

$\boldsymbol{\beta}$ is the p-vector of regression parameters

Note: this is a special case of the general linear mixed model with diagonal covariance structure

The Fay-Herriot Model

- Under above mentioned assumptions

$$\begin{aligned} E_m[\hat{\theta}_i^{dir}] &= E_m[\mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i] = \mathbf{x}_i^T \boldsymbol{\beta} \\ MSE_m(\hat{\theta}_i^{dir}) &= V_m(\hat{\theta}_i^{dir}) = E_m[(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \boldsymbol{\beta})^2] \\ &= E_m[(\mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i - \mathbf{x}_i^T \boldsymbol{\beta})^2] \\ &= E_m[u_i^2 + e_i^2 + 2u_i e_i] = \sigma_u^2 + \psi_i^2 \end{aligned}$$

- Under Normality of u_is and e_is and under FH model

$$\hat{\theta}_i^{dir} \sim N(\mathbf{x}_i \boldsymbol{\beta}, \sigma_u^2 + \psi_i^2)$$

The Fay-Herriot Model

The Best Linear Unbiased Predictor (BLUP) is obtained minimizing $MSE_m(\hat{\theta}_i^{dir})$

- $\hat{\theta}_i^{dir} = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i = \mathbf{x}_i^T \mathbf{a} + \mathbf{b}$, $\hat{\theta}_i^{dir}$ is a linear estimator
- $E_m[\hat{\theta}_i^{dir}] = \mathbf{x}_i^T \boldsymbol{\beta} = E_m[\theta_i]$, $\hat{\theta}_i^{dir}$ is an unbiased estimator (under FH model)
- $\min_{\hat{\theta}_i^{dir}} MSE_m(\hat{\theta}_i^{dir}) \rightarrow$

$$\tilde{\theta}_i^{BLUP} = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}) = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + u_i \qquad (2)$$

so $\tilde{\theta}_i^{BLUP}$ is the *best* linear unbiased predictor

The Fay-Herriot Model

The BLUP can be rewritten as follows

$$\tilde{\theta}_i^{BLUP} = \gamma_i \hat{\theta}_i^{dir} + (1 - \gamma_i)\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}$$

- $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}$
- σ_u is unknown
- $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}^{dir}$
- ψ_i^2 is assumed known (actually is the estimated MSE of direct estimate)
- $\tilde{\theta}_i^{BLUP}$ is a composite estimator

The Fay-Herriot Model

- Using the joint distribution $f(\hat{\theta}_i^{dir}, u_i)$ under the Normality assumption we can get the Restricted Maximum Likelihood (REML) estimates of σ_u, say $\hat{\sigma}_u$
- so $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\theta}}^{dir}$
- where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and $\hat{\boldsymbol{\theta}}^{dir} = [\hat{\theta}_1^{dir}, \ldots, \hat{\theta}_m^{dir}]^T$
- and $\hat{\mathbf{V}} = \text{diag}\{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\psi_1^2}, \ldots, \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\psi_m^2}\}$
- $\hat{u}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\psi_i^2}(\hat{\theta}_i^{dir} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
- Given that
  - $\hat{u}_i$ need both the estimates of $\hat{\sigma}_u$ and $\hat{\boldsymbol{\beta}}$ and,
  - $\hat{\boldsymbol{\beta}}$ need the estimates of $\hat{\sigma}_u$
  - then there is need of an iterative algorithm to obtain these estimates (e.g. Fisher scoring algorithm)

The Fay-Herriot Model

- Pluggin in $\hat{\beta}$ and $\hat{u}_i$ into (2) we get the Empirical BLUP (EBLUP)

$$\hat{\theta}_i^{EBLUP} = \mathbf{x}_i^T \hat{\beta} + \hat{u}_i = \mathbf{x}_i^T \hat{\beta} + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \hat{\beta}) \qquad (3)$$

- The EBLUP in equation 3 can be rewritten as follows

$$\boxed{\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{\theta}_i^{dir} + (1 - \hat{\gamma}_i)\mathbf{x}_i^T \hat{\beta}}$$

- $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}$
- $\hat{\gamma}_i$s are known as *shrinkage* factors

MSE of BLUP under FH model

- The BLUP ca be expressed as

$$\tilde{\theta}_i^{BLUP} = \underbrace{\mathbf{x}_i^T \boldsymbol{\beta} + \gamma_i(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \boldsymbol{\beta})}_{\text{Estimator when } \boldsymbol{\beta} \text{ is known}} + (1-\gamma_i)\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

- The MSE of BLUP is

$$\begin{aligned} MSE_m(\tilde{\theta}_i^{BLUP}) &= MSE_m[\mathbf{x}_i^T \boldsymbol{\beta} + \gamma_i(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \boldsymbol{\beta}) + (1-\gamma_i)\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\ &= V_m(\mathbf{x}_i^T \boldsymbol{\beta} + \gamma_i(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \boldsymbol{\beta})) + V_m((1-\gamma_i)\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= g_{1i}(\sigma_u) + g_{2i}(\sigma_u) \end{aligned}$$

- it can be shown that $\mathbf{x}_i^T \boldsymbol{\beta} + \gamma_i(\hat{\theta}_i^{dir} - \mathbf{x}_i^T \boldsymbol{\beta})$ and $(1-\gamma_i)\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are uncorrelated

MSE of BLUP under FH model

- $g_{1i}(\sigma_u) = V_m(\mathbf{x}_i^T\boldsymbol{\beta} + \gamma_i(\hat{\theta}_i^{dir} - \mathbf{x}_i^T\boldsymbol{\beta})) = \sigma_u^2 - \frac{\sigma_u^4}{\sigma_u^2+\psi_i^2} = \gamma_i\psi_i^2$
- $g_{2i}(\sigma_u) = V_m((1-\gamma_i)\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})) = (1-\gamma_i)^2\mathbf{x}_i\left(\frac{\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^T}{\sigma_u^2+\psi_i^2}\right)^{-1}\mathbf{x}_i$
- $g_{1i}(\sigma_u)$ is the variance of the BLUP estimator when all the parameters are known
- $g_{2i}(\sigma_u)$ accounts for the variability in the estimator $\tilde{\boldsymbol{\beta}}$

Remark: at this stage σ_u^2 is considered known

MSE of EBLUP under FH model

- The error in the EBLUP estimator can be decomposed as follow

$$\hat{\theta}_i^{EBLUP} - \theta_i = (\tilde{\theta}_i^{BLUP} - \theta_i) + (\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})$$

- Then

$$\begin{aligned} MSE_m(\hat{\theta}_i^{EBLUP}) &= E_m[(\hat{\theta}_i^{EBLUP} - \theta_i)^2] \\ &= E_m[((\tilde{\theta}_i^{BLUP} - \theta_i) + (\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP}))^2] \\ &= MSE_m(\tilde{\theta}_i^{BLUP}) + E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2] \\ &+ 2E_m[(\tilde{\theta}_i^{BLUP} - \theta_i)(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})] \end{aligned}$$

MSE of EBLUP under FH model

- Under Normality assumptions made on u_is and e_is

$$E_m[(\tilde{\theta}_i^{BLUP} - \theta_i)(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})] = 0$$

- So the $MSE_m(\hat{\theta}_i^{EBLUP})$ reduces to

$$MSE_m(\hat{\theta}_i^{EBLUP}) = MSE_m(\tilde{\theta}_i^{BLUP}) + E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2]$$

- It follows that $MSE_m(\hat{\theta}_i^{EBLUP}) > MSE_m(\tilde{\theta}_i^{BLUP})$
- The term $E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2]$ is intractable. Often heuristic approximation are used.

MSE of EBLUP under FH model

- Using Taylor linearization to approximate $E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2]$ it is possible to show the follow

$$E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2] \approx \left(\frac{\partial\gamma}{\partial\sigma_u^2}\right)^2 \bar{\mathbf{V}}(\hat{\sigma_u^2}) = g_{3i}(\sigma_u^2)$$

- $\bar{\mathbf{V}}(\hat{\sigma_u^2})$ is the asymptotic variance of $\hat{\sigma_u^2}$

$$\bar{\mathbf{V}}(\hat{\sigma_u^2}) = 2\left[\sum_{i=1}^{m} \frac{1}{(\sigma_u^2 + \psi_i^2)^2}\right]^{-1}$$

obtained from the Fisher information in REML (or ML) estimation

- So the third term in $MSE_m(\hat{\theta}_i^{EBLUP})$ is

$$g_{3i}(\sigma_u^2) = \frac{\psi_i^4}{(\sigma_u^2 + \psi_i^2)^3} 2\left[\sum_{i=1}^{m} \frac{1}{(\sigma_u^2 + \psi_i^2)^2}\right]^{-1}$$

MSE of EBLUP under FH model

- The MSE of EBLUP is

$$MSE_m(\hat{\theta}_i^{EBLUP}) = MSE_m(\tilde{\theta}_i^{BLUP}) + E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2]$$

- $MSE_m(\tilde{\theta}_i^{BLUP}) = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2)$
- $E_m[(\hat{\theta}_i^{EBLUP} - \tilde{\theta}_i^{BLUP})^2] \approx g_{3i}(\sigma_u^2)$
- Then $MSE_m(\hat{\theta}_i^{EBLUP}) \approx g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2)$
- Next step is to estimate $MSE_m(\hat{\theta}_i^{EBLUP})$

Estimation of MSE of EBLUP under FH model

- Let's plugging in the formula of $MSE_m(\hat{\theta}_i^{EBLUP})$ the REML estimate of σ_u^2

$$mse_1(\hat{\theta}_i^{EBLUP}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + g_{3i}(\hat{\sigma}_u^2)$$

- $E_m[mse_1(\hat{\theta}_i^{EBLUP})] = E_m[g_{1i}(\hat{\sigma}_u^2)] + E_m[g_{2i}(\hat{\sigma}_u^2)] + E_m[g_{3i}(\hat{\sigma}_u^2)]$
- where
  - $E_m[g_{2i}(\hat{\sigma}_u^2)] \approx g_{2i}(\sigma_u^2)$
  - $E_m[g_{3i}(\hat{\sigma}_u^2)] \approx g_{3i}(\sigma_u^2)$
  - but $E_m[g_{1i}(\hat{\sigma}_u^2)] \not\approx g_{1i}(\sigma_u^2)$
- There is need to evaluate the bias of $g_{1i}(\hat{\sigma}_u^2)$ so to obtain an approximately correct estimator of the MSE

Estimation of MSE of EBLUP under FH model

- Recall that $g_{1i}(\sigma_u^2) = \psi_i^2 \gamma_i = \psi_i^2 \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}$
- and that $g_{1i}(\hat{\sigma}_u^2) = \psi_i^2 \hat{\gamma}_i = \psi_i^2 \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}$
- A Taylor expansion of $g_{1i}(\hat{\sigma}_u^2)$ is

$$
\begin{aligned}
g_{1i}(\hat{\sigma}_u^2) &= g_{1i}(\sigma_u^2) + (\hat{\sigma}_u^2 - \sigma_u^2)\frac{\partial g_{1i}(\sigma_u^2)}{\partial \sigma_u^2} + \frac{1}{2}(\hat{\sigma}_u^2 - \sigma_u^2)^2 \frac{\partial g_{1i}(\sigma_u^2)}{\partial \sigma_u^2 \partial \sigma_u^2} \\
&= \psi_i^2 \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2} + (\hat{\sigma}_u^2 - \sigma_u^2)\frac{\psi_i^4}{(\sigma_u^2 + \psi_i^2)^2} - (\hat{\sigma}_u^2 - \sigma_u^2)^2 \frac{\psi_i^4}{(\sigma_u^2 + \psi_i^2)^3}
\end{aligned}
$$

Estimation of MSE of EBLUP under FH model

- Under the condition $E[\hat{\sigma}_u^2] = \sigma_u^2$, the expected value of $g_{1i}(\hat{\sigma}_u^2)$ is

$$\begin{aligned} E_m[g_{1i}(\hat{\sigma}_u^2)] &\approx \psi_i^2 \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2} + \frac{\psi_i^4}{(\sigma_u^2 + \psi_i^2)^2} E_m[(\hat{\sigma}_u^2 - \sigma_u^2)] \\ &- \frac{\psi_i^4}{(\sigma_u^2 + \psi_i^2)^3} E_m[(\hat{\sigma}_u^2 - \sigma_u^2)^2] \\ &= \psi_i^2 \gamma_i + 0 - \frac{\psi_i^4}{(\sigma_u^2 + \psi_i^2)^3} \bar{\mathbf{V}}(\hat{\sigma}_u^2) \\ &= g_{1i}(\sigma_u^2) - g_{3i}(\sigma_u^2) \end{aligned}$$

Estimation of MSE of EBLUP under FH model

So the bias of the MSE estimator, $E_m[mse_1(\hat{\theta}_i^{EBLUP}) - MSE(\hat{\theta}_i^{EBLUP})]$, is

$$\begin{aligned}&E_m[g_{1i}(\hat{\sigma}_u^2)] + E_m[g_{2i}(\hat{\sigma}_u^2)] + E_m[g_{3i}(\hat{\sigma}_u^2)] - (g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2))\\ &= g_{1i}(\sigma_u^2) - g_{3i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) - g_{1i}(\sigma_u^2) - g_{2i}(\sigma_u^2) - g_{3i}(\sigma_u^2)\\ &= -g_{3i}(\sigma_u^2)\end{aligned}$$

Finally, a correct estimator of the MSE of EBLUP is

$$mse(\hat{\theta}_i^{EBLUP}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2)$$

Recap on mse of EBLUP

- $g_{1i}(\hat{\sigma}_u^2) = \hat{\gamma}_i \psi_i^2$ is the leading term of the *mse*
- $g_{2i}(\hat{\sigma}_u^2) = (1 - \hat{\gamma}_i)^2 \mathbf{x}_i \left(\frac{\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T}{\hat{\sigma}_u^2 + \psi_i^2} \right)^{-1} \mathbf{x}_i$
- $g_{3i}(\hat{\sigma}_u^2) = \frac{\psi_i^4}{(\hat{\sigma}_u^2 + \psi_i^2)^3} 2 \left[\sum_{i=1}^m \frac{1}{(\hat{\sigma}_u^2 + \psi_i^2)^2} \right]^{-1}$

The case of out of sample areas

- Population is divided into m small areas
- A sample is available in $m - k$ areas $\Longrightarrow$ in k areas there are not observation
- We call the k areas *out of sample areas* $(j = 1, \ldots, k)$
- In this case the EBLUP under FH model reduce to a synthetic estimator

$$\hat{\theta}_j^{OUT} = \mathbf{x}_j^T \hat{\boldsymbol{\beta}} \quad j = 1, \ldots, k$$

Remark: Synthetic estimation is possible unless there are auxiliary variables in out of sample areas

Conclusions

- Few data requirements
- In many applications the method can reduce the MSE of direct estimates
- Area-level models are used as a standard technique to obtain small area statistics
- For out of sample areas, where there are no sample observations, the method provides only model based synthetic estimates instead of estimates that result from the combination of direct estimates (collected data) and model based estimates
- Pros and cons of EBLUP in TZ will be discussed at the end