Small Area Estimation (Part I)

Short Course – Institute of Statistics, Republic of Albania

Tobias Schoch, Feburary 17 and 18, 2022

About

This slide deck has been prepared for the "Short Course on Small Area Estimation with R" at the Institute of Statistics of the Republic of Albania on February 17 and 18, 2022.

Contact

University of Applied Sciences Northwestern Switzerland
School of Business – Institute ICC
Prof. Dr. Tobias Schoch
Riggenbachstrasse 16
CH-4600 Olten
Switzerland

E-Mail: tobias.schoch@fhnw.ch

Phone: +41 62 957 21 02

Outline

1 Introduction

- **Overview of the documentation**

    - **Part I** *(this slide deck)* provides the theoretical background to Small Area Estimation (SAE)

    - Part II presents two useful R packages for SAE and illustrates their application

    - Part III is on the transfer of SAE to Instat

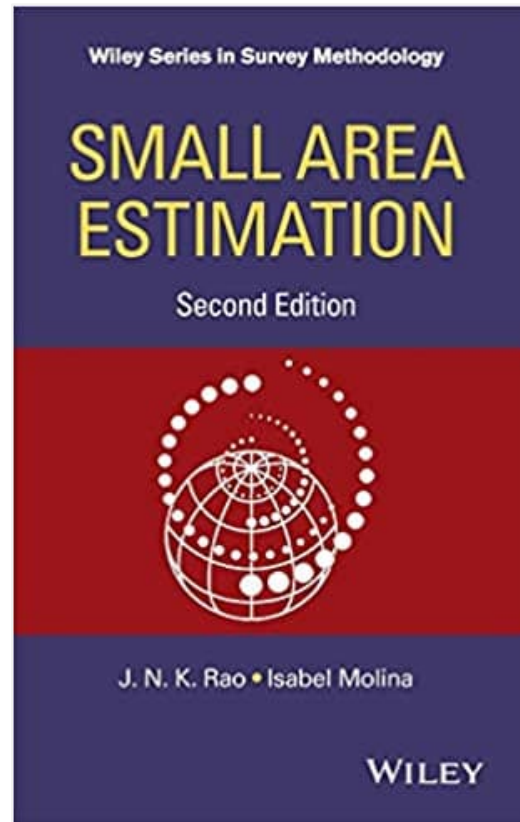- **Theory**

    - This short course introduces only the mathematical concepts essential to understanding the basic principles of SAE. We do not seek to be mathematically rigorous or to be complete (in any sense).

    - For the mathematical details, please refer to the books (see next slide).
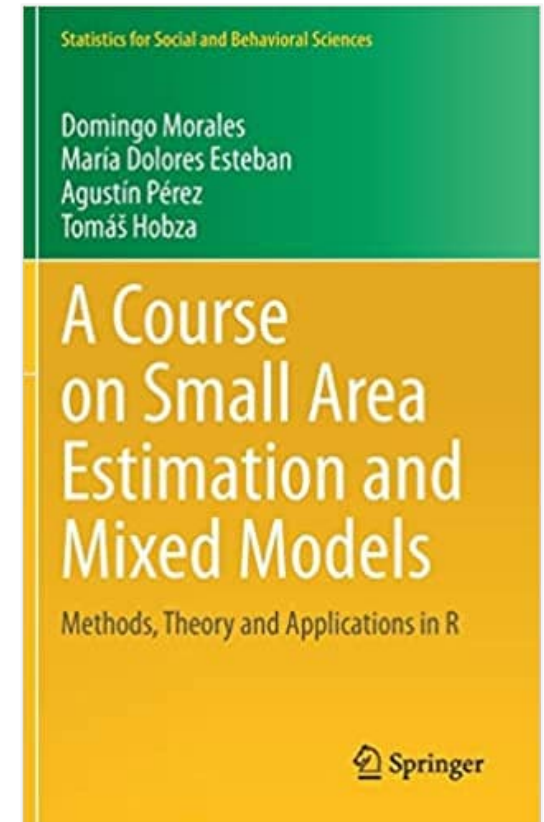
1 Introduction (ctd.)
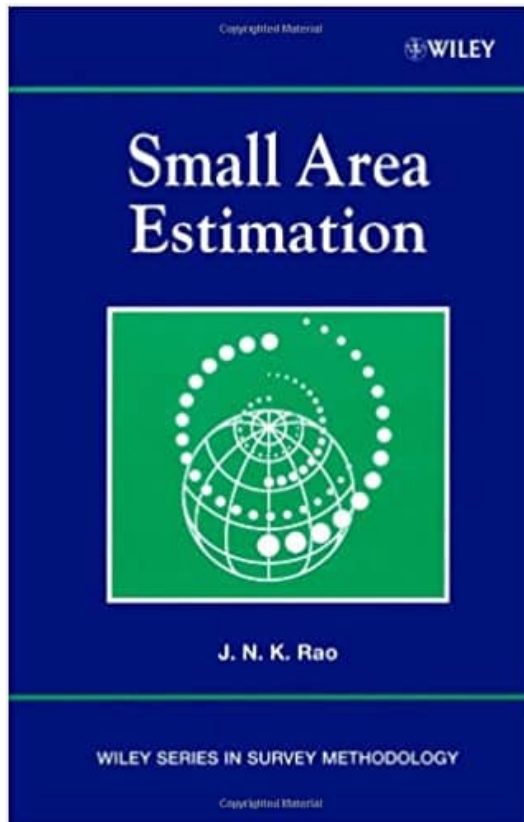
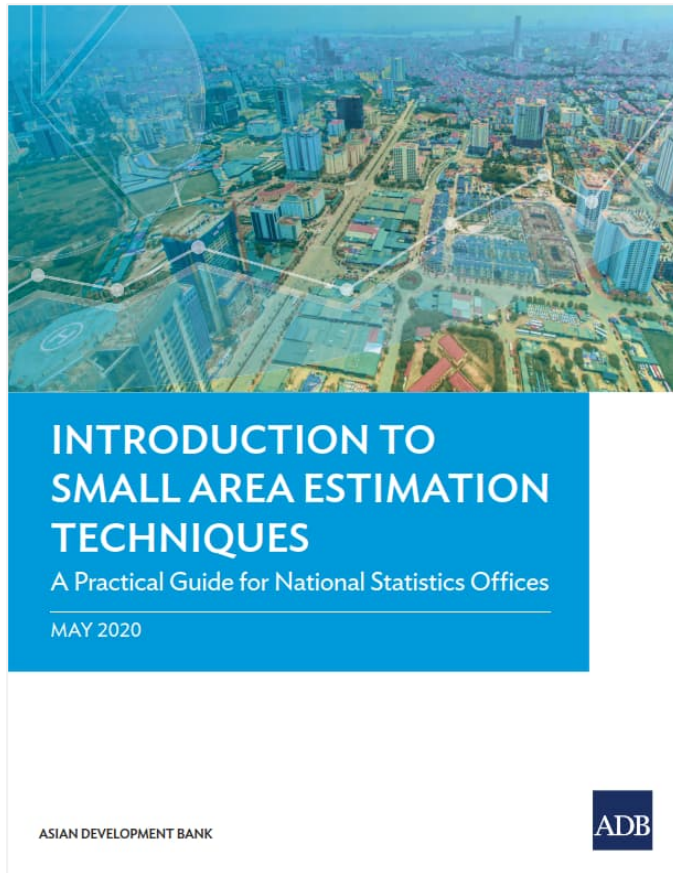## Useful books

Rao (2003)　　　　Rao & Molina (2015)　　　　Morales et al. (2021)

1 Introduction (ctd.)

Useful books (ctd.)

Asian Development Bank (2020)

https://www.adb.org/publications/small-area-estimation-guide-national-statistics-offices

1 Introduction (ctd.)

Documentation of the files in the archive "SAEcourse.zip"

- `/lecture`
  - [slide decks of the Parts I, II, and III as pdf-files]
  - `data.R` (SAIPE 2005 data used in the slide deck of Parts I and II)
  - `snippets.R` (code snippets used in Part II of the slide deck)

- `/software`
  - `methods.R` (functions that extend package `sae`; useful but not necessary)

- `/application`
  - `application.pdf` (and `.md`)
  - `application.R` (R code)
  - `application_solution.pdf` (and `.md`)
  - `datLCS.txt` and `auxLCS.txt` (data sets)

1 Introduction (ctd.)

How to download the archive "SAEcourse.zip"?

- Download https://github.com/tobiasschoch/SAEcourse/zipball/master

- Extract the zip-archive and copy the content to your R working directory

2 Motivation: Small Area Income and Poverty Estimates (SAIPE)

- US Census Bureau, 2005 (https://www.census.gov/programs-surveys/saipe.html)

- **Goal:** Estimate **child poverty** for the $i = 1, ..., 51$ **US states** (poverty rates; families with 5–17 years old children)

**Illinois
14.6%**

Illinois, IL

Estimate: 14.6%
90% Confidence Interval: 14.0 to 15.2%

Picture: 2019 data,
https://www.census.gov/programs-surveys/saipe.html
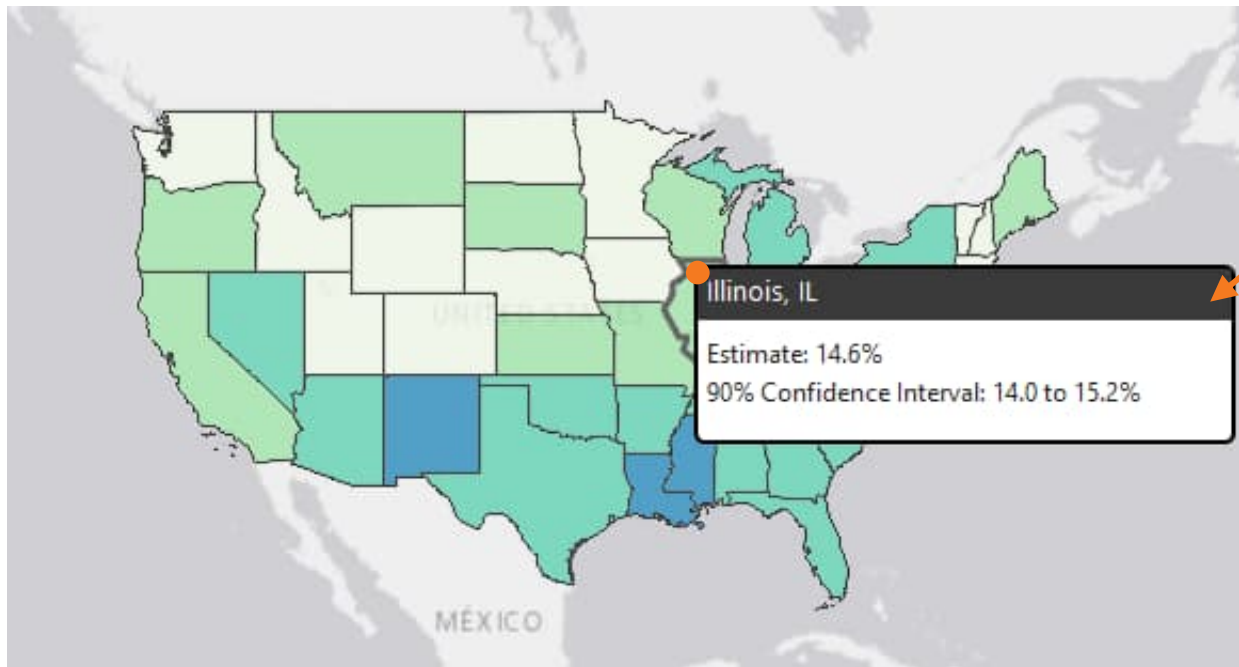
2 Motivation: SAIPE (ctd.)

- **Direct estimator** of poverty rates (for each state)

$$\hat{\theta}_i = \frac{\text{\# families with children in poverty}}{\text{\# families}}, \qquad i = 1, \ldots, 51$$

  based on the Current Population Survey (CPS) of the Census Bureau

- **CPS**

  – Household survey (60 000 households; annual survey; personal and telephone interviews; panel with 3 waves)

  – Multistage stratified random sample

- The direct estimator $\hat{\theta}_i$ has **variance** v_i, i.e., $v_i = var(\hat{\theta}_i)$

- The **coefficient of variation** is $cv_i = \dfrac{v_i}{\hat{\theta}_i}$
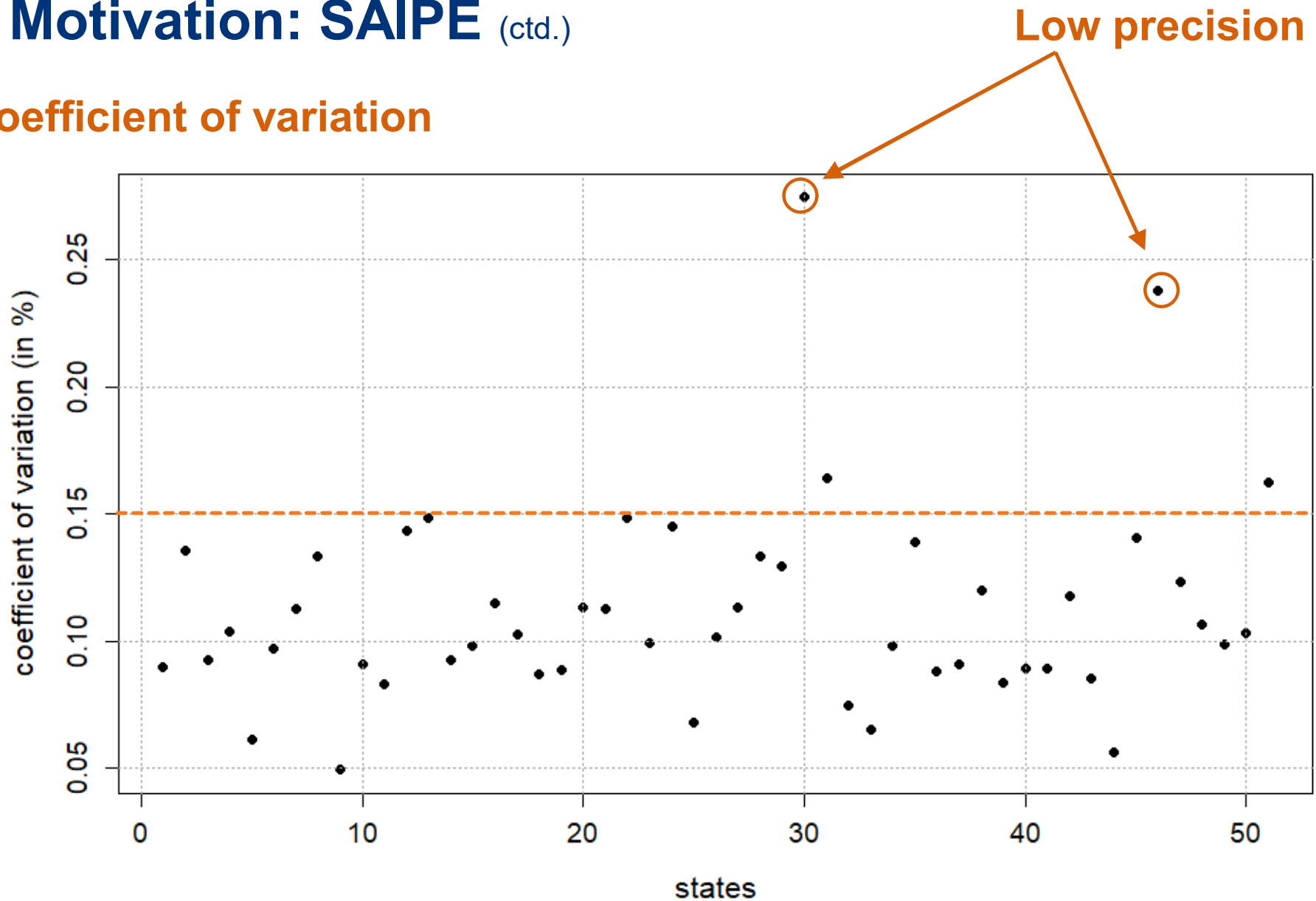
2 Motivation: SAIPE (ctd.)

What counts as an area?

- **Geographical**: regions, provinces, municipalities, healthcare service areas, etc.

- **Socio-economic** groups (e.g., gender, age, etc.)

- Other **sub-populations**/ domains of interest (e.g., firms grouped by economic activities such as agriculture, mining, construction, etc.)

Key characteristic of SAE problems

The sample size in an area/ domain is so small that the direct estimator is considered unreliable (has excessively large variance or does not meet other quality standards)

2 Motivation: SAIPE (ctd.)

Wrap-up

- Survey data (CPS)

- Estimator of poverty ratio $\widehat{\theta}_i$ and its variance, v_i

- Design-based approach

- **Issue with SAIPE data:** very poor precision for 2 states; fair precision in majority of states, and good precision in 3 – 5 states

Ok, we publish only the estimates for states with acceptable precision; that is, we ignore the "bad" states. ⇨ **No, this is unacceptable**.

Can we do better?

- Is a **model-based** approach feasible? meaningful? desired?

- What does "**better**" mean—in this context?

Note: estimator $\hat{\theta}_i^{EBLUP}$ will be introduced later

2 Motivation: SAIPE (ctd.)

Yes, we can do better: EBLUP instead of direct estimator

State	$\hat{\theta}_i$	$\hat{\theta}_i^{EBLUP}$	$v_i = var(\hat{\theta}_i)$	$\text{MSE}(\hat{\theta}_i^{EBLUP})$	Ratio
CA	16.4	16.4	1.0	0.9	1.2
TX	20.3	20.0	1.3	1.0	1.2
FL	13.1	13.6	1.4	1.1	1.3
NH	4.3	4.6	1.4	1.1	1.3
…	…	…	…	…	…
LA	21.4	23.0	3.6	2.1	1.7
MS	28.6	26.4	3.8	2.2	1.7
Total			*113.3*	*77.7*	

SAE estimator
EBLUP

precision of SAE estimator
mean square error

2 Motivation: SAIPE (ctd.)

Doing better

- **component-wise**, i.e., for <u>most</u> areas, we have

$$\mathrm{MSE}\left(\hat{\theta}_i^{EBLUP}\right) < \mathrm{var}\left(\hat{\theta}_i\right) \qquad (i = 1, \dots, n)$$

- **compound error** (or in total)

$$\sum_{i=1}^{n} \mathrm{MSE}\left(\hat{\theta}_i^{EBLUP}\right) < \sum_{i=1}^{n} \mathrm{var}\left(\hat{\theta}_i\right)$$

EBLUP differs conceptionally from the direct estimator in that it explicitly takes the compound error into account.

EBLUP is not unbiased as an estimator of θ_i, therefore we compute the MSE not the variance

3 Models in Survey Sampling

- The use of models has long been controversial in survey sampling. "But recent years have brought a change: Models are now well-engrained in the design-based philosophy and practice as well." (Särndal, 2011)

- **Paradigm 1: Model-assisted approach**

  - Design-based inference is the centerpiece. The model remains in the background; its characteristics are not (or only loosely) incorporated into making inference.

  - The full potential of a "correct" model is not realized.

- **Paradigm 2: Model-based approach**

  - The model is the centerpiece of statistical inference (purists deny the importance of the sampling design for making inference).

3 Models in Survey Sampling (ctd.)

Some more notation (design-based approach)

- Population U of size N

- Random sample s from U with sample size n sampling design $p(s)$

- Sample inclusion probabilities π_j (or weights $w_j = 1/\pi_i$)

- Variable of interest y_j $(j = 1, \ldots, n)$

- Estimator of the population y-total (Horvitz-Thompson estimator, HTE) and population y-mean

US population

CPS

multistage

stratified

$$\hat{Y} = \sum_{j \in s} \frac{y_j}{\pi_j} = \sum_{j \in s} w_j y_j \qquad \text{and} \qquad \hat{\theta} = \frac{\hat{Y}}{N}$$

3 Models in Survey Sampling (ctd.)

Model-assisted approach

- Population U, sample s, … (as before)

- **+ model** ❶

- **+ auxiliary information** ❷ (level of population)

Example

- Simple random sample of $n = 100$ US counties (population $N = 3141$)

- Variables observed in sample s

  - `farmpop`   Farm population (farmers & family)

  - `numfarm`   Number of farms

  - …   [additional variables]

- Goal: What is **total of farm population** in the US?

---

3 Models in Survey Sampling (ctd.)

Example (ctd.)

- Regression **model** ① (population level)

$$\text{farmpop}_j = b \cdot \text{numfarm}_j + e_j, \qquad j = 1, \dots, 3141$$

  - where b is an unknown regression coefficient; e_j is random error

  - Assumption: `farmpop` is linearly related to `numfarm`

- The total number of farms in the population

$$\sum_{j=1}^{N} \text{numfarm}_j$$

  is known ⇨ **Auxiliary information** ② (e.g., from a register).

3 Models in Survey Sampling (ctd.)

Example (ctd.)

- The **generalized regression estimator** (GREG) of the total farm population is defined as

$$\widehat{\text{Total}}_{GREG} = \sum_{j \in s} \frac{\text{farmpop}_j}{\pi_j} + \hat{b} \left(\sum_{j \in s} \frac{\text{numfarm}_j}{\pi_j} - \sum_{j=1}^{N} \text{numfarm}_j \right)$$

Horvitz-Thompson estimator of `farmpop`

Estimated coefficients

Horvitz-Thompson estimator of `numfarm`

Auxiliary information = 2'015'128

3 Models in Survey Sampling (ctd.)

Example (ctd.)

- How does $\widehat{\mathrm{Total}}_{GREG}$ compare to the HTE?

  – Estimated standard error of HTE is 329'787; for the GREG 168'400 (see Appendix)

  – **GREG is almost twice as efficient!**

- **Wrap-up:** Estimator $\widehat{\mathrm{Total}}_{GREG}$

  – is a nearly unbiased estimator of the population total

  – can be much more efficient than HTE

  – gives valid estimates even if the model is not "ideal"; the contribution of the assisting model vanishes as n grows;

  – [can be very inefficient if the model is completely misspecified]

3 Models in Survey Sampling (ctd.)

Model-based approach

- Let $\widehat{\mathrm{farmpop}}_j$ be the **prediction** for unit j under the estimated regression model

- The model-based estimator of the total farm population is

$$\widehat{\mathrm{Total}}_{Pred} = \underbrace{\sum_{j \in s} \mathrm{farmpop}_j}_{\substack{\text{Sample total of} \\ \mathtt{farmpop}}} + \underbrace{\sum_{j \in U \setminus s} \widehat{\mathrm{farmpop}}_j}_{\substack{\text{Total of the \underline{non-sample} part} \\ \text{of the population (prediction)}}}$$

- Sampling design is ignored; total is estimated by sum (not HTE)

3 Models in Survey Sampling (ctd.)

Model-based approach

- The **prediction** for the j-th county is

$$\widehat{\text{farmpop}}_j = \hat{b} \cdot \text{numfarm}_j,$$

thus, the total over the **non-sample** part becomes

$$\sum_{j \in U \setminus s} \widehat{\text{farmpop}}_j = \hat{b} \cdot \sum_{j \in U \setminus s} \text{numfarm}_j$$

and since $\sum_{j=1}^{N} \text{numfarm}_j$ (auxiliary information) is a known quantity, $\sum_{j \in U \setminus s} \text{numfarm}_j$ is easy to compute

- The variance of $\widehat{\text{Total}}_{Pred}$ is computed only with respect to the model (the sampling design is ignored).

3 Models in Survey Sampling (ctd.)

Wrap-up

- The **model-/ prediction-based estimator** exploits the model's full potential in terms of efficiency.

  - If the model is "good", the efficiency advantage over the GREG estimator tends to be large.

  - If the model is "bad", the estimator tends to be heavily biased.

- The **model-assisted estimator** (GREG) is—to some extent— robust against model misspecification. This "insurance" against model failure comes at the cost of a lower efficiency (premium or penalty).

3 Models in Survey Sampling (ctd.)

...Let's return to our example on poverty rates in the US

- For estimating poverty rates, should we use the GREG or prediction-based estimator using

  – a separate model for each state? (so, 51 models in total)

  – one overall model instead?

- **Neither is a good idea**

  – Specifying, estimating and maintaining 51 different models is usually not worth the effort.

  – One overall regression model attempts to treat all states the same. This is not sensible. The coefficient of variation was very small for some states. Why would you want to lump "good" and "bad" estimates/ states into the same pot?

Whish list (for our application on poverty rates by state in the US)

- Estimates **for all** states are required (no exclusion)

- **Manageable** approach (i.e., not 51 separate models)

- Estimates with **high and low coefficient of variation** should be treated differently

- Exploit as much **potential of the model-based** approach as possible

- Maintain some of the **robustness** that the design-based/ model-assisted approach offers (protection against model misspecification)

- Statistical inference should be **computationally manageable** (e.g., no Markov chain Monte-Carlo methods as in Bayesian statistics)

- Methods should not be very "data-hungry" (to be explained in a second)

⇨ **Small Area Estimation**

4 Overview on Small Area Estimation

Overview: 2 types of models

- **Area-level model** $y_i = \dots$ for $i = 1, \dots, n$ our focus

- **Unit-level model** $y_{ij} = \dots$ for $j = 1, \dots, n_i, \;\; i = 1, \dots, n$

   where index

  - i refers to the areas

  - j refers to the units that are embedded in the i-th area

Discussion

- Area-level models are less "data-hungry".
- Unit-level models can be more efficient—however, they require much more attention in model building ($\Rightarrow$ tricky). Also, access to unit-level data might be restricted due to confidentiality reasons.

4 Overview on Small Area Estimation (ctd.)

Overview: What distinguishes SAE from other approaches?

- Model-based approach

- "Explicit linking models based on **random area-specific effects** that account for between area variation **beyond that is explained by auxiliary variables** included in the model will be called 'small area model'" (Rao, 2003, p. 4)

- "Borrowing strength" + shrinkage estimators

5 Area-Level Model

Characteristics of interest

- Our focus is on area-level **means** and **totals** (i.e., statistics that are linear in the observations)

- *Side note:* The R package `emdi` provides functions to compute transformations (e.g., logarithm, etc.) ⇨ SAE models for indicators, which are usually non-linear functions, $g(\theta_i)$

Basic area-level model = Fay-Herriot model

Required data

- Direct estimators and variances (survey data)

- Auxiliary information (area-level data, any data source will do fine, e.g., register data, tax data, etc.)

Note: $\hat{\theta}_i = y_i$

5.1 Model Specification and EBLUP

The **Fay-Herriot** model is defined as

$$y_i = x_i^T \beta + u_i + e_i, \qquad i = 1, \ldots, n,$$

where

- y_i direct estimator $\hat{\theta}_i$, (e.g., estimated total or mean in the i-th area)

- x_i explanatory variables (area-specific), $x_i \in \mathbb{R}^q$

- β regression coefficient, unknown, $\beta \in \mathbb{R}^q$

- u_i area-specific random effect $u_i \sim N(0, \ v_i)$, where the v_i's are the known variances of the direct estimator $\hat{\theta}_i$

- e_i model error $e_i \sim N(0, \ \sigma^2)$, where the variance $\sigma^2 > 0$ is unknown

$\Rightarrow$ **β and σ^2 must be estimated**

5.1 Model Specification and EBLUP (ctd.)

- We have **2 random effects** u_i and e_i (for the areas $i = 1, \ldots, n$)

  - Isn't there an identifiability problem? I mean, u_i and e_i are area-specific—so, how can we "distinguish" them from each other?

  - Recall that $u_i \sim N(0, \ v_i)$ and the variances v_i are known! Thus, no issue.

- FH approach to SAE is **model-based**, but the direct estimators $\hat{\theta}_i$ are design-based

- **Normality assumptions**

  - $u_i \sim N(0, \ v_i)$ is justified by the central limit theorem ($\hat{\theta}_i$ is a mean or total)

  - $e_i \sim N(0, \ \sigma^2)$, where σ^2 must be estimated $\Rightarrow$ we (the model builder) must check the validity of this assumption (see diagnostic plots, later)

5.1 Model Specification and EBLUP

A different point of view

For the areas $i = 1, \ldots, n$, the FH-model can be expressed as a **hierarchical Bayes model**

$$y_i \mid \Theta_i = \theta_i \sim N(\theta_i, v_i) \qquad \text{sampling model}$$

$$\Theta_i \sim N\left(x_i^T \beta, \sigma^2\right) \qquad \text{linking model}$$

- In Bayesian jargon, the linking model is called the prior or a-priori distribution of the random variables Θ_i.
- The linking model (prior) assumes that the Θ_i's have the same parent distribution ($\Rightarrow$ borrowing strength)
- Clearly, β and σ^2 are unknown in practice; thus, we follow an **empirical Bayes** approach and estimate β and σ^2

5.1 Model Specification and EBLUP (ctd.)

The **best linear unbiased predictor** (BLUP) is defined for all areas $i = 1, \dots, n$ as

$$\tilde{\theta}_i^{BLUP} = \gamma_i \hat{\theta}_i + (1 - \gamma_i) x_i^T \beta \qquad \text{with} \qquad \gamma_i = \frac{\sigma^2}{\sigma^2 + v_i}$$

- $\tilde{\theta}_i^{BLUP}$ is a smooth blend of the direct estimator $\hat{\theta}_i$ and the linear predictor $x_i^T \beta$.

- The parameters σ^2 and β are assumed known ($\Rightarrow$ rare in practice).

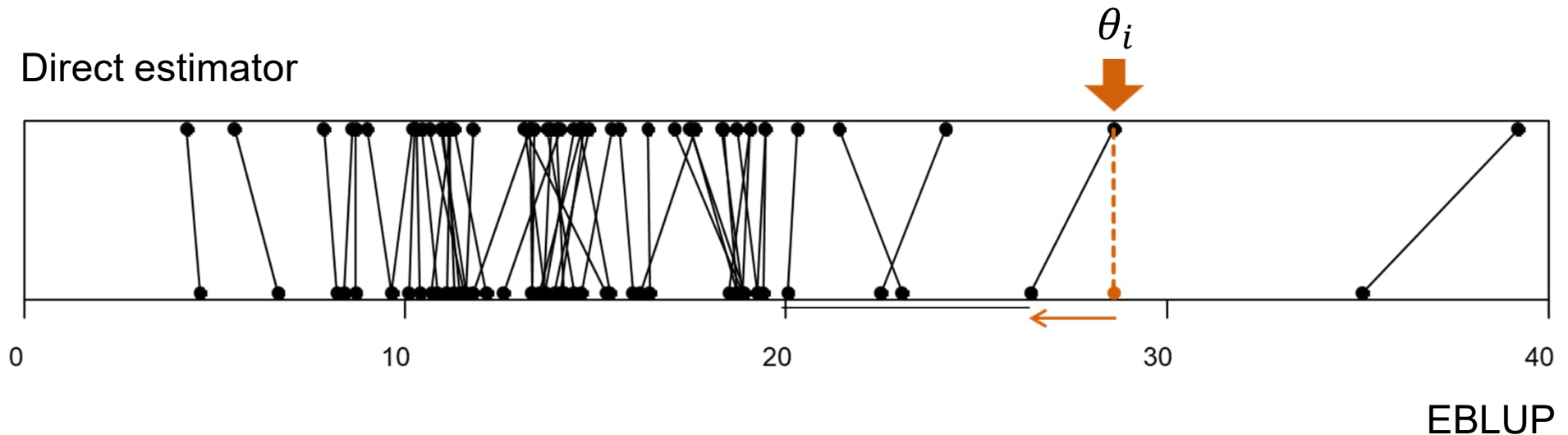The **empirical best linear unbiased predictor** (EBLUP) is

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) x_i^T \hat{\beta} \qquad \text{with} \qquad \hat{\gamma}_i = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + v_i}$$

where $\hat{\sigma}^2$ and $\hat{\beta}$ are estimators of σ^2 and β, respectively.

5.1 Model Specification and EBLUP (ctd.)

Shrinkage (SAIPE data)

- Direct estimator of θ_i 28.7
- EBLUP of θ_i 26.4

5.1 Model Specification and EBLUP (ctd.)

Two areas with different variances of the direct estimator

- Let $\hat{\sigma}^2 = 10$ (variance of model error)

- Thus $\hat{\gamma}_i = \dfrac{10}{10 + v_i}$

"High precision" area with $v_i = 1$

$$\hat{\gamma}_i = \frac{10}{10 + \mathbf{1}} = \frac{\mathbf{10}}{\mathbf{11}}$$

Thus

$$\tilde{\theta}_i^{BLUP} = \frac{\mathbf{10}}{\mathbf{11}} \hat{\theta}_i + \frac{\mathbf{1}}{\mathbf{11}} x_i^T \beta$$
$$\approx \mathbf{0.9}\, \hat{\theta}_i + \mathbf{0.1}\, x_i^T \beta$$

"Low precision" area with $v_i = \mathbf{10}$

$$\hat{\gamma}_i = \frac{10}{10 + \mathbf{10}} = \frac{\mathbf{1}}{\mathbf{2}}$$

Thus

$$\tilde{\theta}_i^{BLUP} = \frac{\mathbf{1}}{\mathbf{2}} \hat{\theta}_i + \frac{\mathbf{1}}{\mathbf{2}} x_i^T \beta$$
$$= \mathbf{0.5}\, \hat{\theta}_i + \mathbf{0.5}\, x_i^T \beta$$

5.1 Model Specification and EBLUP (ctd.)

Behavior of EBLUP as a function of $\hat{\sigma}^2$ and v_i. For ease of reference, we print $\hat{\gamma}_i$ again

$$\hat{\gamma}_i = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + v_i}.$$

An asymptotic point of view is helpful

- if $\hat{\sigma}^2 \to \infty$, then $\hat{\gamma}_i \to 1$ (model variance increases $\Rightarrow$ bad model)

- if $v_i \to 0$, then $\hat{\gamma}_i \to 1$ (precision of direct estimator, $\hat{\theta}_i$, increases)

As a consequence,

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) x_i^T \hat{\beta} \quad \to \quad \hat{\theta}_i$$

- $\hat{\theta}_i^{EBLUP}$ approaches the direct estimator $\hat{\theta}_i$

- SAE model is irrelevant

5.2 Estimation

Our model

$$y_i = \beta_0 + \beta_1 \, \text{prIRS}_i + \beta_2 \, \text{nfIRS}_i + \beta_3 \, \text{prCensus}_i + u_i + e_i,$$

with variables

```
    prIRS      nfIRS    prCensus          yi       vi      state
1  23.1582    14.6657    11.8464     19.4400   3.0371         AL
2  15.1852    10.9282     7.6478     11.0042   2.2330         AK
3  19.5926    19.3337     9.0293     17.4417   2.6003         AZ
```

Explanatory variables

prIRS   poverty rate (tax data)

nfIRS   non filer rate (tax data)

prCensus poverty rate (census)

Survey estimates (CPS data)

yi direct estimator of poverty rate

vi variance of direct estimator

5.2 Estimation (ctd.)

Regression coefficients

Weighted least squares estimator $\hat{\beta}$ with weights $(v_i + \sigma^2)^{-1}$

Variance estimator

- Moment-type estimator of Fay & Herriot (FH) $\hat{\sigma}^2_{FH}$

- Maximum likelihood (MLE) $\hat{\sigma}^2_{MLE}$ ⇨ asymptotically most efficient, but biased

- Restricted maximum likelihood estimator (REML) $\hat{\sigma}^2_{REML}$ ⇨ unbiased

Discussion

- There exist other variance estimators, e.g., simple moment estimator of Prasad & Rao (1990) ⇨ rarely used in practice

- The variance estimators do not have close-form expressions ⇨ iterative computation

5.2 Estimation (ctd.)

Discussion (ctd.)

- Asymptotically (as $n \to \infty$), we have $\hat{\sigma}^2_{REML} = \hat{\sigma}^2_{MLE} \leq \hat{\sigma}^2_{FH}$

- **In practice,**

  - the difference between the variance estimators in small

  - $\hat{\sigma}^2_{REML}$ is the method of choice

- Iterative computation of $\hat{\beta}$ and $\hat{\sigma}^2$

5.2 Estimation (ctd.)

Significance of the estimated regression coefficients

```
              Estimate     Std.Err    t value     Pr(>|t|)
(Intercept)  -4.15645     1.53397    -2.7096     0.006736 **
prIRS         0.22610     0.15124     1.4949     0.134934
nfIRS         0.87004     0.14354     6.0615     1.349e-09 ***
prCensus      0.43653     0.18175     2.4019     0.016312 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- All coefficients (except `prIRS`) are significantly different from zero at least at the 5% level of significance. We keep variable `prIRS` because the US Census Bureau kept it in the model.

- Model selection can be done with the help of **information criteria** (AIC, BIC, etc.); see Marhuenda et al. (2014)

5.3 Statistical Inference

Mean square error of the EBLUP

Second-order analytic approximation of the mean square error of the EBLUP (Prasad & Rao, 1990; granted some regularity conditions)

$$\mathrm{MSE}\big(\hat{\theta}_i^{EBLUP}\big) \approx g_{1i}(\sigma^2) + g_{2i}(\sigma^2) + g_{3i}(\sigma^2)$$

- $g_{1i}(\sigma^2)$ and $g_{1i}(\sigma^2)$ account for the uncertainty of the BLUP

- $g_{3i}(\sigma^2)$ reflects the uncertainty due to estimation of σ^2 (depends estimator, MLE, REML, etc.)

Estimator

The MSE can be estimated by

$$\widehat{\mathrm{MSE}}\big(\hat{\theta}_i^{EBLUP}\big) \approx g_{1i}(\hat{\sigma}^2) + g_{2i}(\hat{\sigma}^2) + g_{3i}(\hat{\sigma}^2)$$

5.3 Statistical Inference (ctd.)

Other MSE estimators

- Bootstrap

- Jackknife

5.3 Statistical Inference (ctd.)

Confidence intervals

Normal-theory confidence intervals (CI) of $\hat{\theta}_i$ (for $i = 1, \dots, n$), where $\hat{\theta}_i$ is shorthand notation for the EBLUP, are defined as

$$\hat{\theta}_i \pm z_{\alpha/2}\, s(\hat{\theta}_i),$$

where

- $z_{\alpha/2}$ is the $\left(1 - \dfrac{\alpha}{2}\right)$ quantile of the standard normal distribution,

- $s(\hat{\theta}_i)$ is an estimate of the standard error of $\hat{\theta}_i$, for instance, we may use the estimated MSE of $\hat{\theta}_i$ as $s^2(\hat{\theta}_i)$.

For instance, $\alpha = 0.05$ gives the 95% CI.

The validity of the confidence intervals depends on whether the normal distribution holds.

5.4 Extensions

For ease of reference, we show the FH model again

$$y_i = x_i^T \beta + u_i + e_i, \qquad i = 1, \dots, n,$$

Extensions

- **Multivariate** Fay-Herriot model $\boldsymbol{y}_i = \left(y_{1i}, \dots, y_{qi}\right)^T \in \mathbb{R}^q$ in place of y_i

- **Spatial correlation** models: The random effects $e = (e_1, \dots, e_n)^T$ follow a first-order simultaneous autoregressive (SAR) process, $e = \rho W e + \epsilon$, where ρ is an autoregression parameter, $\epsilon \sim N(0 \cdot 1_n, \sigma_1^2 I_n)$, and I_n and W are, respectively, the $(n \times n)$ identity matrix and a $(n \times n)$ proximity matrix; 1_n is the n-vector of ones

- **Time series** and cross-sectional methods (e.g., spatio-temporal FH model: time-specific random effects account for temporal changes)

6 Unit-Level Model (Big Picture)

The **basic unit-level model** is defined as

$$y_{ij} = x_{ij}^T \beta + u_i + e_{j'}, \qquad j = 1, \dots, n_{i'}, \qquad i = 1, \dots n,$$

where

- y_{ij} observation of unit j in the area i,
- x_{ij} explanatory variables (unit-specific),
- β regression coefficient,
- u_i area-specific random effect
- e_j individual random effect

Basic area-level model is a "special case" of the basic unit-level model

Note that

$$\theta_i \approx y_i = \bar{x}_i \beta + u_i$$

where $\bar{x}_i$ is the area-specific mean of the x_{ij}'s   (if $N_i \gg n_i$)

Literature

Asian Development Bank (2020) Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices, Manila (Philippines).

Battese, Harter & Fuller (1988) An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, J. Am. Stat. Assoc.

Fay & Herriot (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, J. Am. Stat. Assoc.

Morales, Esteban, Pérez & Hobza (2021) A Course in Small Area Estimation and Mixed Models: Methods, Theory and Applications in R, Cham: Springer Nature.

Marhuenda, Morales & Pardo (2014) Information criteria for Fay-Herriot model selection. Computational Statistics and Data Analysis 70, p. 268-280.

Prasad & Rao (1990) The Estimation of the Mean Squared Error of Small-Area Estimators, J. Am. Stat. Assoc.

Rao (2003) Small Area Estimation, John Wiley and Sons: Hoboken (NJ).

Rao & Molina (2015) Small Area Estimation, 2nd ed. John Wiley and Sons: Hoboken (NJ).

Särndal (2011) Models in Survey Sampling, in: Carlson, Nyqvist & Villani (eds.), Official Statistics: Methodology and Applications: In Honour of Daniel Thorburn, Stockholm, p. 15-27.

A Code for GREG estimator of population total

```
library(robsurvey)

library(survey)

# Data and design object

data(counties)

dn <- svydesign(ids = ~1, fpc = ~fpc, weights = ~weights, data = counties)

# Horvitz-Thompson estimator (HTE) of farmpop

HTE <- svytotal(~farmpop, dn)

# Step 1 of GREG estimator of the total of farmpop (in argument 'population',
# we define the auxiliary information; i.e., N = 3141 and the population total of
# numfarm which is 2015128; the result of calibrate is a new design object)

dn_greg <- calibrate(dn, ~numfarm, population=c(3141, 2015128))

# Step 2 of GREG estimator (we compute the GREG, given the calibrated design)

GREG <- svytotal(~farmpop, dn_greg)

# Comparison of HTE and GREG (we compute the ratio of the standard errors)

SE(HTE) / SE(GREG)
```