Estimation for domains and small areas with design-based and model-based methods - Topic 1

Risto Lehtonen
University of Helsinki

B-N-U Summer School, 23-27 August 2009, Kyiv, Ukraine

Lecture topics

- Topic 1: Introduction and framework
- Topic 2: Design-based methods
- Topic 3: Model-based methods
- Topic 4: Computation
Topic 1
INTRODUCTION AND FRAMEWORK

World-wide trend

- An increasing need in society for reliable statistics for regional and other population subgroups or domains
- Challenge for scientific research and official statistics
Lively SAE Research

- Applied small area estimation (SAE) research in Europe
  - EURAREA Project (2001-2004), EU’s FP5
    http://www.statistics.gov.uk/eurarea/
  - AMELI Project (2008-2011), EU’s FP7
    http://www.uni-trier.de/index.php?id=24474&no_cache=1
  - SAMPLE Project (2008-2011), EU’s FP7
    http://www.sample-project.eu/

SAE Conferences

- EWORSAE
  European Working Group on Small Area Estimation
  http://sae.wzr.pl/

  SAE2005 (University of Jyväskylä)
  SAE2007 (University of Pisa)
  SAE2009 (University of M. Hernandez, Elche)
  Forthcoming:
  SAE2011 (University of Trier)
  SAE2013 (Poland)
What is estimation for domains and small areas?

- Domains of interest: Well-defined population subgroups
- Examples, social survey
  - Regional areas constructed by administrative criteria: county, municipality,…
  - Demographic criteria: sex and age grouping
  - Demographic breakdown within regional areas
- Examples, business survey
  - Grouping of enterprises into domains according to the type of industry

Estimation for domains, or domain estimation for short, refers to the estimation of population quantities, such as:

- Totals
- Means
- Proportions
- Medians, Quantiles, Percentiles…

for the desired population subgroups called domains.
Special case - SAE

Small area estimation, SAE

- Estimation for domains whose sample size is small or very small (even zero)

- Alternative definition (Partha Lahiri): Small area = Domain of interest, for which the sample size is not adequate to produce reliable direct estimates

Typical estimation task

Specify and identify the domains of interest
The number $D$ of domains of interest $U_d$ is usually large

Specify the target parameters for study variable $y$
Domain totals $t_d = \sum_{k \in U_d} y_k$
Domain means $\bar{Y}_d = t_d / N_d, d = 1, ..., D$
where $N_d$ is domain size
### Typical estimation task

**Data management tasks**
- Specify data sources, Merging of data from different sources
- Data cleaning
- Nonresponse adjustment, Imputation, Reweighting...

**Specify the estimator of domain parameters**

**Specify variance estimator**

**Computing and quality assurance**

**Publication**

---

### Examples

- Estimation of regional number of ILO unemployed by sex and age group, based on Labour Force Survey
- Estimation of median household disposable income by municipality, based on sample survey such as EU SILC
- Estimation of regional poverty indicators, such as regional poverty rate, based on sample survey
Issues concerning domain estimation procedure

- Types of domains of interest?
  - Planned domains / Unplanned domains
- Available data infrastructure?
  - Sample survey data
  - Level of auxiliary information
- Type of domain estimator?
  - Direct / Indirect
  - Design based / Model-based
- Type of model?
  - Linear / Non-linear
  - Fixed effects models / Mixed models
- Accuracy measures?
  - Variance estimators / MSE estimators

Two main domain structures

- Planned domains
  - The most important domains are defined as strata in the sampling design
  - Domain sample sizes are fixed
  - Domain sample sizes are controlled by allocation scheme
  - Small sample sizes can be avoided

- Unplanned domains
  - Domain sample sizes are not fixed but are random
  - Small domain sample sizes can occur
  - Most common case in practice
Planned domains

$U$ Population

$U_d$ Population domain $d$

Domains = Strata

$s_d$ Sample in domain $d$

Sample size $n_d$ in domain $d$ is fixed

$d = 1,\ldots, D$

Unplanned domains

$U$ Population

$s$ Sample

$U_d$ Population domain $d$

$s_d = s \cap U_d$ Sample in domain $d$

Sample size $n_d$ in domain $d$ is random

$d = 1,\ldots, D$
Direct and indirect estimation

Direct estimation
- **Direct** domain estimator uses values of the variable of interest \( y \) only from the time period of interest and only from units in the domain of interest (Federal Committee on Statistical Methodology, 1993)
- Often in connection to **planned** domain structures

Indirect estimation
- **Indirect** domain estimator uses values of the variable of interest \( y \) from a domain and/or time period other than the domain and time period of interest
- Often in connection to **unplanned** domain structures

Example: Direct HT estimator

Design-based Horvitz-Thompson estimator

\[ \hat{t}_{dHT} = \sum_{k \in s_d} y_k / \pi_k \]

of domain total \( t_d \) only uses \( y \)-values from \( s_d \)

Therefore \( \hat{t}_{dHT} \) is direct estimator
Example: Direct GREG estimator

Direct design-based model-assisted generalized regression GREG estimator

\[ \hat{i}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_j} (y_k - \hat{y}_k) / \pi_k \]

uses linear models fitted separately in each domain:

\[ Y_k = x_k' \beta_d + \varepsilon_k , \quad k \in U_d , \quad d = 1, \ldots, D \]

where \( \beta_d \) is domain-specific, and \( \hat{y}_k = x_k' \hat{\beta}_d \) are fitted \( y \)-values calculated for every \( k \in U_d \).

Example: Indirect GREG estimator

Indirect design-based GREG estimator

\[ \hat{i}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_j} (y_k - \hat{y}_k) / \pi_k \]

uses a linear fixed-effects model

\[ Y_k = x_k' \beta + \varepsilon_k , \quad k \in U \]

with vector \( \beta \) common for all domains and \( \hat{y}_k = x_k' \hat{\beta} \) calculated for every \( k \in U \).
Example: Indirect SYN estimator

Indirect model-based synthetic SYN estimator

\[ \hat{d}_{SYN} = \sum_{k \in U_d} \hat{y}_k \]

uses a linear fixed-effects model

\[ Y_k = x'_k \beta + \epsilon_k, \quad k \in U \]

with vector \( \beta \) common for all domains

and \( \hat{y}_k = x'_k \hat{\beta} \) calculated for every \( k \in U \)

“Borrow strength”

- In general, indirect estimators are attempting to “borrow strength” from other domains and/or in a temporal dimension
- For domains with small sample size, this is a well justified goal
- The concept of “borrowing strength” is often used in model-based small area estimation
Approaches for domain estimation and SAE

- Design-based methods
- Model-based and model-dependent methods
- Bayesian methods
  - Empirical Bayes
  - Hierarchical Bayes
- Poverty mapping
  - World Bank, Peter Lanjouw, Chris Elbers,…
  - PovMap Software

Design-based methods

- Estimation approach where the randomness is introduced by the sampling design
- Statistical properties of estimators are evaluated under the sampling design
- Estimators are constructed so that the complexities of the sampling design are accounted for
- Design weights are usually incorporated in the estimation procedure
Model-based and model-dependent methods

- Estimation approach where the randomness is introduced by an assumed superpopulation model
- Statistical properties of estimators are evaluated under the model
- Sampling design and design weights are not necessarily an issue
  - Certain differences between model-based and model-dependent methods

Examples of domain estimation methods

- Design-based methods
  - Horvitz-Thompson HT estimators
  - Generalized regression GREG family estimators
  - Model-free calibration estimators
  - Model calibration MC estimators
  - Restriction estimators (Kaja Sõstra)

- Model-based and model-dependent methods
  - Synthetic SYN estimators
  - Empirical best linear unbiased predictor EBLUP type estimators
  - Empirical best predictor EBP type estimators
The role of models

- **GREG estimators** use models as an assisting tool
  - GREG estimators are **model-assisted**

- **SYN estimators** rely exclusively on the model
  - SYN estimators are **model-dependent**

Comparison of estimators

Design-based properties of different estimators of $t_d$ are often compared in terms of:

**Bias:** \[ \text{Bias}(\hat{t}_d) = E(\hat{t}_d) - t_d \]

**Precision:** \[ \text{Var}(\hat{t}_d) = E(\hat{t}_d - E(\hat{t}_d))^2 \]

**Accuracy:** \[ \text{MSE}(\hat{t}_d) = E(\hat{t}_d - t_d)^2 = \text{Var}(\hat{t}_d) + \text{Bias}^2(\hat{t}_d) \]
Design-based properties of estimators

<table>
<thead>
<tr>
<th></th>
<th>Design-based methods</th>
<th>Model-based methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HT, GREG, MC</td>
<td>SYN, EBLUP, EB</td>
</tr>
<tr>
<td>Bias</td>
<td>Design unbiased</td>
<td>Design biased</td>
</tr>
<tr>
<td></td>
<td>(approximately) by</td>
<td>Bias does not</td>
</tr>
<tr>
<td></td>
<td>the construction</td>
<td>necessarily approach</td>
</tr>
<tr>
<td></td>
<td>principle</td>
<td>zero with increasing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sample size</td>
</tr>
<tr>
<td>Precision (Variance)</td>
<td>Large variance for</td>
<td>Small variance for</td>
</tr>
<tr>
<td></td>
<td>small domains</td>
<td>small domains</td>
</tr>
<tr>
<td></td>
<td>Variance decreases</td>
<td>Variance decreases</td>
</tr>
<tr>
<td></td>
<td>with increasing</td>
<td>with increasing</td>
</tr>
<tr>
<td></td>
<td>sample size</td>
<td>sample size</td>
</tr>
<tr>
<td>Accuracy (Mean Squared</td>
<td>MSE = Variance</td>
<td>MSE = Variance</td>
</tr>
<tr>
<td>Error, MSE)</td>
<td>(or nearly so)</td>
<td>+ squared Bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy can be poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>if the bias is</td>
</tr>
<tr>
<td></td>
<td></td>
<td>substantial</td>
</tr>
<tr>
<td>Confidence intervals</td>
<td>Valid design-based CI</td>
<td>Valid design-based CI</td>
</tr>
<tr>
<td></td>
<td>can be constructed</td>
<td>not necessarily</td>
</tr>
<tr>
<td></td>
<td></td>
<td>obtained</td>
</tr>
</tbody>
</table>

Natural application areas of estimation approaches by domain sample size

<table>
<thead>
<tr>
<th>ESTIMATION APPROACH</th>
<th>DOMAIN SAMPLE SIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minor</td>
</tr>
<tr>
<td>Model-based</td>
<td></td>
</tr>
<tr>
<td>Synthetic SYN</td>
<td>++</td>
</tr>
<tr>
<td>EBLUP, EBP</td>
<td>+++</td>
</tr>
<tr>
<td>Design-based</td>
<td></td>
</tr>
<tr>
<td>Horvitz-Thompson HT</td>
<td>0</td>
</tr>
<tr>
<td>GREG, MC</td>
<td>+</td>
</tr>
</tbody>
</table>

Applicability codes:

- 0: Not at all
- +: Low
- ++: Medium
- +++: High
Calculation tools

- SAS Software
  - Procedure SURVEYMEANS

- Project EURAREA
  - SAS Macro for standard estimators
  - SAS Macro EBLUPGREG for advanced estimators

- Program DOMEST
  - Developed by Dr Ari Veijanen (Statistics Finland and University of Helsinki)

- R programs

  - VLISS-Virtual Laboratory in Survey Sampling
    http://mathstat.helsinki.fi/VLISS

Selected literature - Design-based


Selected literature - Design-based


Selected literature - Model-based

Selected literature - Model-based


Estimation for domains and small areas with design-based and model-based methods - Topic 2

Risto Lehtonen
University of Helsinki

B-N-U Summer School, 23-27 August 2009, Kyiv, Ukraine

Topic 2
DESIGN-BASED METHODS
Issues to be discussed

- Data infrastructure
- Direct HT estimators
- Direct and indirect GREG estimators
- Assisting models
- Variance estimation
- Case study
  - GREG and model calibration MC
  - Lehtonen, Särndal and Veijanen (2009)
- Main source
  - Lehtonen and Veijanen (2009), Handbook of Statistics, Vol. 29B, Chapter 31

Register-driven data infrastructure

- “Register” countries
  - Administrative and statistical registers are available for statistical purposes
  - Official statistics production mainly based on statistical registers
  - Sample surveys completing register sources
  - Unique ID codes in different data sources
  - Micro-merging of data from different sources
- Unit-level models and unit-level auxiliary data are often used for estimation for domains and small areas
Survey-driven data infrastructure

- “Survey” countries
  - Weaker availability of administrative registers for statistical purposes
  - Official statistics production mainly based on sample surveys
  - Unique ID code systems not (yet) well developed
  - Data from different sources can only be merged at an aggregate level
- **Area-level models** and **area-level auxiliary data** are often used for estimation for domains and small areas

Approach chosen here

- Register-driven data infrastructure
  - Combined use of sample survey data and auxiliary register data
- Why?
- This is for flexibility
- Additional points
  - Many countries in Europe and elsewhere are using, or are turning towards, this option
  - Most research is carried out under this option
Components of estimation procedure in a register-driven data infrastructure

- Sample survey data
  - Access to unit-level sample survey data
  - Model specification
  - Model fitting

- Auxiliary data
  - Access to auxiliary unit-level population data

- Domain estimation
  - Merging of sample data and auxiliary data
  - Calculation of fitted values for all population elements
  - Calculation of domain estimators and accuracy measures

Design-based estimation for domains: HT and model-free calibration

- Design-based Horvitz-Thompson estimator
  - Direct estimation for planned domains structures

- Model-free calibration estimators
  - Direct estimation for planned domain structures
  - Särndal (2007)
  - Estevao and Särndal (2004)
    - Restriction estimator
  - Plikusas & Pumputis (2007)
    - Nonlinear calibration
  - Särndal (2007)
  - Lehtonen and Veijanen (2009)
Design-based estimation for domains: GREG and model calibration

- Model-assisted generalized regression
  - GREG family estimators
    - Direct or indirect estimation procedures under planned or unplanned domain structures
    - Särndal, Swensson and Wretman (1992)
    - Lehtonen, Särndal and Veijanen (2003, 2005)

- Model calibration MC family estimators
  - Indirect estimation under unplanned domain structures
  - Wu and Sitter (2001)
  - Lehtonen, Särndal and Veijanen (2009)

Requirements for auxiliary data

- GREG family estimators
  - “Standard” GREG using linear fixed-effects model: Aggregate-level data
  - Advanced GREG using nonlinear models or mixed models: Unit-level data

- Model-free calibration estimators
  - Aggregate-level data

- Model calibration MC family estimators using nonlinear models
  - Unit-level data
EXAMPLE 1. Direct HT and GREG under planned domains


- **Section 3.5. Computational example with direct estimation under a planned domain structure**

Variables

- **Study variable** $y$
  - Disposable household income

- **Auxiliary x-variables**
  - EDUC: the number of household members who had higher education
  - EMP: the number of months in total the household members were employed during last year

- All three variables were determined using administrative registers
**Sampling design**

- Population: \( N = 431,000 \) households
- Household sampling:
  - Stratified \( \pi \)PS (PPS-WOR)
- Size variable in PPS-WOR: Number of household members
- Strata: \( D = 12 \) NUTS4 regions (domains)
- Proportional allocation
  - Domain sample sizes are fixed
- Sample size: 1000 households

**HT estimator**

HT estimator

\[
\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k = \sum_{k \in s_d} y_k / \pi_K
\]

For variance estimation, we approximated the design by with-replacement type PPS
(SAS Procedure SURVEYMEANS)

Approximate variance estimator

\[
\hat{V}_A(\hat{t}_{dHT}) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} \left( n_d a_k y_k - \hat{t}_{dHT} \right)^2
\]
Direct GREG estimator

\[ \hat{y}_k = \sum_{i \in U_j} \hat{y}_k + \sum_{i \in s_j} a_i e_k = \sum_{k \epsilon s_j} a_k g_{dk} y_k \]

Variance estimator

\[ \hat{\nu}^2 (\hat{t}_{dGREG}) = \sum_{k \epsilon s_j} \sum_{l \epsilon s_j} (a_k a_l - a_{kl}) g_{dk} g_{dl} e_k e_l \]

g-weights are

\[ g_{dk} = I_{dk} + I_{dk} \left( t_{dx} - \hat{t}_{dx} \right)^T \hat{M}_d x_k \]

with 

\[ \hat{M}_d = \sum_{i \epsilon s_j} a_i x_i x_i' \] and 

\[ I_{dk} = \{ k \epsilon U_d \} \]

Residuals are

\[ e_k = y_k - \hat{y}_k \]

Assisting models in GREG

Direct GREG estimator with linear assisting model

\[ Y_k = \beta_{0d} + \beta_{1d} EMP_k + \epsilon_k \] (column 2), or

\[ Y_k = \beta_{0d} + \beta_{1d} EMP_k + \beta_{2d} EDUC_k + \epsilon_k \] (column 3)

NOTE: Domain-specific intercepts and slopes
Quality measures of estimators

MARE
Absolute relative error of an estimator in domain \( d \)

\[
|\hat{t}_d - t_d| / t_d
\]

MARE in domain group:
The mean of absolute relative errors over domains in the group

MCV
The mean coefficient of variation of the estimate over domain group

The coefficient of variation is calculated as \( s.e(\hat{t}_d) / \hat{t}_d \)
where \( s.e \) refers to the estimated standard error of an estimator

Table 1. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and calibration (GREG) estimators of totals for minor, medium-sized and major domains by using various amounts of auxiliary information in a planned domains case.

<table>
<thead>
<tr>
<th>Auxiliary information</th>
<th>HT 1 None</th>
<th>HT 2 Domain sizes and domain totals of EMP</th>
<th>GREG 3 Domain sizes and domain totals of EMP and EDUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain sample size class</td>
<td>MARE %</td>
<td>MCV %</td>
<td>MARE %</td>
</tr>
<tr>
<td>Minor ( 8 \leq n_j \leq 33 )</td>
<td>11.5</td>
<td>11.9</td>
<td>5.8</td>
</tr>
<tr>
<td>Medium ( 34 \leq n_j \leq 45 )</td>
<td>7.6</td>
<td>9.0</td>
<td>3.7</td>
</tr>
<tr>
<td>Major ( 46 \leq n_j \leq 277 )</td>
<td>12.5</td>
<td>5.2</td>
<td>4.3</td>
</tr>
</tbody>
</table>
EXAMPLE 2. Direct HT and indirect GREG under unplanned domains

- Section 4.2. Computational example with direct and indirect estimation under an unplanned domain structure
- Household sampling: πPS
  - No stratification
  - Domain sample sizes are random
- Details on data as in Example 1

HT estimator

\[ \hat{i}_{dHT} = \sum_{k \in d} a_k y_k \]

Variance estimator

\[ \hat{V}_U(\hat{i}_{dHT}) = \frac{n}{n-1} \sum_{k \in d} \left( a_k y_{dk} - \hat{i}_{d} / n \right)^2 \]

with extended domain variables \( y_{dk} = I\{k \in U_d\} y_k \)
**GREG estimator**

\[ \hat{i}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k = \sum_{k \in s} a_k g_{dk} y_k \]

Variance estimator

\[ \hat{\text{V}} \left( \hat{i}_{dGREG} \right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \]

\[ g_{dk} = I_{dk} + \left( t_{dx} - \hat{t}_{dx} \right)' \hat{M}^{-1} x_k \text{ and } \hat{M} = \sum_{l \in s} a_i x_i x_l' \]

---

**Assisting model in GREG**

The indirect GREG estimator was assisted by a common model

\[ Y_k = \beta_0 + \beta_1 \text{EMP}_k + \epsilon_k \text{ (column 2)} \]

fitted to the whole sample, and domain sizes and domain totals of EMP were assumed known

NOTE: Common intercept and slope for all domains
Table 2. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of HT and indirect GREG estimators of totals for minor, medium-sized and major domains by using various amounts of auxiliary information in an unplanned domains case.

<table>
<thead>
<tr>
<th>Auxiliary information</th>
<th>Domain sample size class</th>
<th>HT</th>
<th>GREG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>MARE %</td>
<td>MCV %</td>
</tr>
<tr>
<td>Minor</td>
<td>$8 \leq n_j \leq 33$</td>
<td>11.5</td>
<td>28.3</td>
</tr>
<tr>
<td>Medium</td>
<td>$34 \leq n_j \leq 45$</td>
<td>7.6</td>
<td>20.3</td>
</tr>
<tr>
<td>Major</td>
<td>$46 \leq n_j \leq 277$</td>
<td>12.5</td>
<td>9.6</td>
</tr>
</tbody>
</table>

What can be learned?

- Planned domains, direct estimators
  - GREG better than HT in terms of accuracy

- Unplanned domains, indirect estimators
  - GREG again better than HT in terms of accuracy

- Use of auxiliary data makes sense!

- Planned vs. unplanned case
  - Accuracy better in planned domains case

- Stratification for important domains of interest makes sense!
  - An issue of the survey planning stage
Case Study 1
GREG and MC estimators


Study problem

- Comparison of accuracy of two design-based SAE approaches
  - Generalized regression (GREG) estimation for domain totals
  - Model-calibration (MC) estimation for domain totals
- NOTE: Both GREG and MC are nearly design unbiased
- NOTE: For a linear assisting model GREG and MC coincide
More details…

- Response variable: Binary (values 0 or 1)
- Assisting model: Logistic fixed-effects model
- Unequal probability sampling design
  - PPS-WOR
- Assumed data infrastructure
  - Unit-level sample and population data
- Domain structure: Unplanned
- Estimator type: Indirect

Specific question of interest

- Accuracy performance of logistic GREG and logistic MC in estimation for domain totals of binary response variable
- Accuracy measured in Monte Carlo simulation experiments by relative root mean squared error RRMSE
- Simulation design
  - Design-based simulation
Options for comparison

- Parametrization of logistic model in LGREG and LMC
  - Common models
  - Models with domain-specific parameters

- Level of calibration in MC
  - Calibration at the population level
  - Calibration at the domain level
  - Calibration at an intermediate level

Logistic GREG (LGREG) and logistic MC (LMC) estimators by the level of model calibration and model type

<table>
<thead>
<tr>
<th>Level of model calibration</th>
<th>Model type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Common model formulation for all domains</td>
</tr>
<tr>
<td>Population level</td>
<td>LMC-P1</td>
</tr>
<tr>
<td>Domain level</td>
<td>LMC-D1</td>
</tr>
<tr>
<td>None</td>
<td>LGREG1</td>
</tr>
</tbody>
</table>
**Notation**

$U = \{1, 2, \ldots, k, \ldots, N\}$ Population (fixed, finite)
$U_1, \ldots, U_d, \ldots, U_D$ Domains of interest

Non-overlapping, identified domains
Domain size $N_d$ assumed known

PPS-WOR sampling with sample size $n$
$s$ Sample from $U$
$s_d = s \cap U_d$ Random part of $s$ falling in domain $d$

$\pi_k = n \frac{x_{1k}}{\sum_{k \in U} x_{1k}}$ Inclusion probability for $k \in U$ in PPS
$x_1$ Size variable in PPS
$a_k = 1/\pi_k$ Sampling weight for $k \in s$

We observe values $y_k$ of binary response variable $y$ for $k \in s$

---

**Logistic fixed-effects models 1**

Common model formulation for all domains

$$E_{m}(y_k) = \frac{\exp(x_k' \beta)}{1 + \exp(x_k' \beta)}$$

where

$x_k = (1, x_{1k}, \ldots, x_{pk})'$, $k \in U$

$\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$

$\beta_j$ are fixed effects common for all domains

$j = 0, \ldots, p$
Logistic fixed-effects models 2

Model formulation with domain-specific intercepts

$$E_m(y_k) = \frac{\exp(x_k'\beta)}{1 + \exp(x_k'\beta)}$$

where

- $$x_k = (l_{1k}, \ldots, l_{Dk}, x_{tk}, \ldots, x_{pk})', k \in U$$
- $$l_{dk} = 1$$ if $$k \in U_d$$, $$l_{dk} = 0$$ otherwise, $$d = 1, \ldots, D$$
- $$\beta = (\beta_{01}, \ldots, \beta_{0D}, \beta_1, \ldots, \beta_p)'$$
- $$\beta_{0d}$$ are domain-specific intercepts, $$d = 1, \ldots, D$$
- $$\beta_j$$ are common slopes, $$j = 1, \ldots, p$$

More on logistic GREG...

- Assisting models discussed
  - Logistic fixed-effects and mixed models
  - General framework: Generalized linear mixed models (GLMM)
Generalized linear mixed models

Model formulation with domain-specific random terms

\[ E_m(y_k | u_d) = f(x'_k (\beta + u_d)), \ d = 1, \ldots, D \]

where

- \( f(.) \) refers to the chosen functional form
- \( x_k = (1, x_{1k}, \ldots, x_{pk})' \)
- \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)' \) are fixed effects
- \( u_d = (u_{0d}, \ldots, u_{pd})' \) are random effects

Fitted values are \( \hat{y}_k = f(x'_k (\hat{\beta} + \hat{u}_d)), \ k \in U \)

GREG estimators of domain totals

Target parameters \( t_d = \sum_{U_d} y_k, \ d = 1, \ldots, D \)

Domain totals of binary response variable \( y \)

GREG estimators

\[ \hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in S_d} a_k \hat{e}_k \]

where

- \( a_k = 1/\pi_k, \ \hat{e}_k = y_k - \hat{y}_k \)
- \( \hat{y}_k = \frac{\exp(x'_k \hat{\beta})}{1 + \exp(x'_k \hat{\beta})}, \ k \in U \)
MC estimators of domain totals 1

Calibration at the population level

MC estimators \( \hat{f}_{dMC} = \sum_{k \in s} w_k y_k \)

Find weights \( w_k \) to satisfy calibration equation
\[ \sum_{k \in s} w_k z_k = \sum_{k \in U} z_k = z_U \text{ where } z_k = (1, \hat{y}_k)' \]

Minimize
\[ \sum_{k \in s} \frac{(w_k - a_k)^2}{a_k} - \lambda' \left( \sum_{k \in s} w_k z_k - z_U \right) \quad (1) \]

under calibration constraints
\[ \sum_{k \in s} w_k = N \quad \text{and} \quad \sum_{k \in s} w_k \hat{y}_k = \sum_{k \in U} \hat{y}_k \]

Using Lagrange multiplier \( \lambda \), the equation (1) is minimized by weights
\[ w_k = a_k g_k, \quad g_k = 1 + \lambda' z_k \]

where
\[ \lambda = \left( \sum_{k \in U} z_k - \sum_{k \in s} a_k z_k \right)' \left( \sum_{k \in s} a_k z_k z_k' \right)^{-1} \]

and
\[ z_k = (1, \hat{y}_k)' \]
MC estimators of domain totals 2
Calibration at the domain level

MC estimators \( \hat{t}_{dMC} = \sum_{k \in s_d} w_{dk} y_k \)

Find weights \( w_{dk} \) to satisfy calibration equation
\[
\sum_{k \in s_d} w_{dk} z_k = \sum_{k \in U_d} z_k = z_{U_d} \text{ where } z_k = (1, \hat{y}_k)' 
\]

Minimize
\[
\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k} - \lambda'(\sum_{k \in s_d} w_{dk} z_k - z_{U_d}) \quad (2)
\]
under calibration constraints
\[
\sum_{k \in s_d} w_{dk} = N_d \quad \text{and} \quad \sum_{k \in s_d} w_{dk} \hat{y}_k = \sum_{k \in U_d} \hat{y}_k
\]

Equation (2) is minimized by weights
\[
w_{dk} = a_k g_{dk}, \quad g_{dk} = 1 + \lambda'_d z_k
\]
where
\[
\lambda_d = \left( \sum_{k \in U_d} z_k - \sum_{k \in s_d} a_k z_k \right)'(\sum_{k \in s_d} a_k z_k z_k')^{-1}
\]
and
\[
z_k = (1, \hat{y}_k)'
\]
Monte Carlo simulation

Design-based simulation setting

Artificial finite population with one million elements
\( D = 100 \) domains
(Minor / Medium sized / Major)

Size \( N_d \) of domain \( d \) proportional to
\[ \exp(u), \ u \sim U(0,2.9) \]

Population generating model: Logistic mixed model

The binary random variable \( Y_k \) was defined by
\[ P(Y_k = 1) = \frac{\exp(\eta_k)}{1 + \exp(\eta_k)} \]
with \( \eta_k = (u_{0d} - 5) + (1 + u_{1d})x_{1k} + (1 + u_{2d})x_{2k} \)

where
- \( x_1 \) size variable in PPS, \( x_1 \sim U(1,11) \)
- \( x_2 \sim U(-5,5) \), independent of \( x_1 \)
- Random effects \( N(0,\sigma_{\epsilon}^2) \), \( \sigma_{\epsilon_1}^2 = 9, \sigma_{\epsilon_2}^2 = 0.125 \)
- \( \text{Corr}(u_{1i},u_{2i}) = -0.5, \ i = 1,2 \)

Binary \( y_k \) : If a random number from \( U(0,1) \) was smaller than the computed value \( P(Y_k = 1) \), then \( y_k = 1 \), otherwise \( y_k = 0 \)
Monte Carlo simulation

Correlations: 
\[ \text{corr}(y, x_1) = 0.41 \]
\[ \text{corr}(y, x_2) = 0.32 \]
\[ \text{corr}(x_1, x_2) = -0.001 \]

Unequal probability sampling design

\[ K = 1,000 \] independent with-replacement samples
drawn with PPS-WOR

Size variable in PPS-WOR: \( x_i \)
Sample size \( n = 10,000 \) elements

Monte Carlo simulation

Models

\[ P(Y_k = 1) = \frac{\exp(\eta_k)}{1 + \exp(\eta_k)} \]

Common models

Model 1: \( \eta_k = \beta_0 + \beta_1 x_{ik} \)
Model 2: \( \eta_k = \beta_0 + \beta_2 x_{2k} \)
Model 3: \( \eta_k = \beta_0 + \beta_1 x_{ik} + \beta_2 x_{2k} \)

Models with domain-specific terms

Model 4: \( \eta_k = \beta_{0d} + \beta_1 x_{1k}, \ d = 1,\ldots,D \)
Model 5: \( \eta_k = \beta_{0d} + \beta_2 x_{2k}, \ d = 1,\ldots,D \)
Model 6: \( \eta_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k}, \ d = 1,\ldots,D \)
Quality measure of estimators

Accuracy measure
Relative root mean squared error

\[
\text{RRMSE} \left( \hat t_d \right) = \frac{1}{K} \sqrt{\frac{1}{K} \sum_{v=1}^{K} \left( \hat t_d(s_v) - t_d \right)^2 / t_d}, \quad d = 1, \ldots, D
\]

RRMSE figures are averaged over domain sample size classes:
- Minor (20–69) 47 domains
- Medium (70–119) 19 domains
- Major (120–) 34 domains

Accuracy of LGREG and LMC estimators

Common model type (example)
MC at the population level

<table>
<thead>
<tr>
<th>Linear part of model</th>
<th>Estimator</th>
<th>Mean RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 + \beta_1 x_{1d} )</td>
<td>LMC-P1</td>
<td>41.4 22.4 19.5</td>
</tr>
<tr>
<td>( \beta_0 + \beta_2 x_{2d} )</td>
<td>LMC-P1</td>
<td>28.2 13.8 14.6</td>
</tr>
<tr>
<td>( \beta_0 + \beta_1 x_{1d} + \beta_2 x_{2d} )</td>
<td>LMC-P1</td>
<td>24.8 10.9 14.3</td>
</tr>
<tr>
<td>( \beta_0 + \beta_1 x_{1d} )</td>
<td>LGREG1</td>
<td>28.9 13.4 16.3</td>
</tr>
<tr>
<td>( \beta_0 + \beta_2 x_{2d} )</td>
<td>LGREG1</td>
<td>28.2 13.8 14.6</td>
</tr>
<tr>
<td>( \beta_0 + \beta_1 x_{1d} + \beta_2 x_{2d} )</td>
<td>LGREG1</td>
<td>24.8 10.9 14.3</td>
</tr>
</tbody>
</table>
### Accuracy of LGREG and LMC estimators

#### Common model type

**MC at the domain level**

<table>
<thead>
<tr>
<th>Linear part of model</th>
<th>Estimator</th>
<th>Mean RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Minor (20–69)</td>
</tr>
<tr>
<td>$\beta_0 + \beta_1 x_{1k}$</td>
<td>LMC-D1</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td>LGREG1</td>
<td>28.9</td>
</tr>
<tr>
<td>$\beta_0 + \beta_2 x_{2k}$</td>
<td>LMC-D1</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>LGREG1</td>
<td>28.2</td>
</tr>
<tr>
<td>$\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k}$</td>
<td>LMC-D1</td>
<td>22.1</td>
</tr>
<tr>
<td></td>
<td>LGREG1</td>
<td>24.8</td>
</tr>
</tbody>
</table>

---

### Accuracy of LGREG and LMC estimators

#### Domain-specific intercepts

**MC at the domain level**

<table>
<thead>
<tr>
<th>Linear part of model</th>
<th>Estimator</th>
<th>Mean RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Minor (20–69)</td>
</tr>
<tr>
<td>$\beta_{0i} + \beta_1 x_{1i}$</td>
<td>LMC-D2</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td>LGREG2</td>
<td>24.9</td>
</tr>
<tr>
<td>$\beta_{0i} + \beta_2 x_{2i}$</td>
<td>LMC-D2</td>
<td>25.9</td>
</tr>
<tr>
<td></td>
<td>LGREG2</td>
<td>26.0</td>
</tr>
<tr>
<td>$\beta_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i}$</td>
<td>LMC-D2</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td>LGREG2</td>
<td>19.8</td>
</tr>
</tbody>
</table>
What can be learned?

- If the explanatory power of the assisting model of GREG is “weak”, MC (at the domain level) can improve accuracy.
- If the explanatory power of the model is “strong”, MC (at the domain level) and GREG indicate similar accuracy.
- GREG tends to be more sensitive to the model choice than MC.

Additional references

Estimation for domains and small areas with design-based and model-based methods - Topic 3

Risto Lehtonen
University of Helsinki

B-N-U Summer School, 23-27 August 2009, Kyiv, Ukraine

Topic 3
MODEL-BASED METHODS
Issues to be discussed

- Commonly used model-based or model-dependent estimator types
  - SYN, EBLUP, pseudo EBLUP
- The role of models
  - GLMM framework
- Estimation of accuracy
  - MSE estimation
- Case study
  - Accounting for unequal probability sampling in EBLUP type estimators

EBLUP type estimators for domain totals

- Empirical Best Linear Unbiased Predictor
- Based on mixed models theory
  - Jiang and Lahiri (2006)
  - Lehtonen, Särndal and Veijanen (2003, 2005)
  - Lehtonen, Myrskylä, Särndal and Veijanen (2007)
    - Comparison of GREG and SYN & EBLUP
  - Torabi and Rao (2008)
    - Comparison of GREG and EBLUP
**Options for EBLUP**

- Area-level models
  - Fay and Herriot (1979)
  - Used under limited data availability

- Unit-level models and unit-level auxiliary information
  - Used under advanced data infrastructures allowing access to unit-level population data
  - Recent research area
  - Option used here

**EURAREA project**

- Extensive applied research on:
  - Synthetic estimators
  - EBLUP estimators

- MSE estimation

- Application to real data sets from different countries

- SAS macros available for users

http://www.statistics.gov.uk/eurarea/
Statistical properties

- Model-based / model-dependent methods
  - Statistical properties depend on goodness of the model
  - All models are wrong!
  - Estimators can be seriously design biased
- Accuracy can be good if model is well specified
- Often used methods for cases where domain sample sizes are small
  - Design-based methods fail

Special problem

- How to account for unequal probability sampling?
  - Stratified sampling with non-proportional allocation
  - PPS type sampling designs
- The role of design weights
- The role of design variables in the model
Model-dependent methods

- EBLUP type estimators
- Design biased
- Bias can be reduced by incorporating design variables in the model

Model-based methods

- Pseudo EBLUP type estimators
- Design consistency by using design weights

- Pseudo EBLUP estimators
  - Torabi and Rao (2005)

- EBLUPW estimators
  - Weighted EBLUP
  - Lehtonen, Myrskylä, Särndal and Veijanen (2007)
Generalized linear mixed models

Model formulation with **domain-specific** random terms
\[ E_m(y_k | u_d) = f(x'_k (\beta + u_d)), \quad d = 1, \ldots, D \]

where
\[ f(.) \] refers to the chosen functional form
\[ x_k = (1, x_{tk}, \ldots, x_{pk})' \]
\[ \beta = (\beta_0, \beta_1, \ldots, \beta_p)' \] are fixed effects
\[ u_d = (u_{od}, \ldots, u_{pd})' \] are random effects

Fitted values are \[ \hat{y}_k = f(x'_k (\hat{\beta} + \hat{u}_d)), \quad k \in U \]

---

Special cases

(1) Linear mixed models, random intercepts
\[ E_m(y_k | u_d) = x'_k \beta + u_{od}, \quad d = 1, \ldots, D \]
Fitted values are \[ \hat{y}_k = x'_k \hat{\beta} + \hat{u}_{od}, \quad k \in U \]

(2) Logistic mixed models
\[ E_m(y_k | u_d) = \frac{\exp(x'_k \beta + u_d)}{1 + \exp(x'_k \beta + u_d)} \]
Fitted values are \[ \hat{y}_k = \frac{\exp(x'_k \hat{\beta} + \hat{u}_d)}{1 + \exp(x'_k \hat{\beta} + \hat{u}_d)}, \quad k \in U \]
**EBLUP estimators for domain totals**

Target parameters (domain totals)

\[ t_d = \sum_{U_d} y_k, \quad d = 1, \ldots, D \]

EBLUP estimators

\[ \hat{t}_{dEBLUP} = \sum_{k \in s_d} y_k + \sum_{k \in U_d - s_d} \hat{y}_k, \quad d = 1, \ldots, D \]

Underlying model:

Generalized linear mixed model

---

**Case study 2**

**EBLUP and EBLUPW estimators**

**Study problem**

How to account for unequal probability sampling in EBLUP type estimators?

- PPS-WOR sampling design

**Options considered here:**

- The inclusion of the PPS size variable in the model - model-dependent estimator
- The inclusion of design weights in the estimation procedure - model-based estimator
EBLUP type estimators of domain totals

- Continuous study variable \( y \)

- **EBLUP estimators**
  - Linear mixed model
  - Estimation with GLS and REML
  - **Weights are ignored**

- **EBLUPW estimators**
  - Mixed linear model
  - Estimation with GWLS and a weighted modification of REML
  - **Weights are included**

---

Monte Carlo simulation

- Artificial finite population
  - One million elements
  - \( D = 100 \) domains

- Population generating model
  - Linear mixed model

- \( K = 1000 \) πPS samples
  - \( n = 10,000 \) elements

- Correlations
  - \( corr(y, x_1) = 0.44 \)
  - \( corr(y, x_2) = 0.45 \)
  - \( corr(x_1, x_2) = 0.001 \)

\[
y_k = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + (\beta_2 + u_{2d})x_{2k} + \epsilon_k
\]

- \( x_1 \) size variable in πPS

- \( \beta_0 = \beta_1 = \beta_2 = 1, \sigma_{u0}^2 = 1, \sigma_{u1}^2 = \sigma_{u2}^2 = 0.125 \)
### Quality measures of estimators

- **Bias**
  - Absolute relative bias ARB (%)

- **Accuracy**
  - Relative root mean squared error RRMSE (%)

\[
ARB(\hat{t}_d) = \left| \frac{1}{K} \sum_{v=1}^{K} \hat{t}_d(s_v) - t_d \right| / t_d
\]

\[
RRMSE(\hat{t}_d) = \sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{t}_d(s_v) - t_d)^2 / t_d}
\]

### EBLUP under Model 1

<table>
<thead>
<tr>
<th>Model and estimator</th>
<th>Average ARB (%)</th>
<th>Average RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain size class</td>
<td>Domain size class</td>
</tr>
<tr>
<td></td>
<td>Minor (20-69)</td>
<td>Medium (70-119)</td>
</tr>
<tr>
<td>Model 1 ( y_k = \beta_k + u_d + e_k )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EBLUP</strong></td>
<td>19.7</td>
<td>19.5</td>
</tr>
<tr>
<td><strong>EBLUPW</strong></td>
<td>3.7</td>
<td>3.1</td>
</tr>
</tbody>
</table>
## EBLUP under Model 2

<table>
<thead>
<tr>
<th>Model and estimator</th>
<th>Average ARB (%)</th>
<th>Average RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain size class</td>
<td>Domain size class</td>
</tr>
<tr>
<td>Minor (20-69)</td>
<td>Medium (70-119)</td>
<td>Major (120+)</td>
</tr>
<tr>
<td>EBLUP</td>
<td>4.0</td>
<td>3.6</td>
</tr>
<tr>
<td>EBLUPW</td>
<td>3.6</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Model 2 \( y_k = \beta_0 + u_d + \beta_k x_{ik} + \varepsilon_k \)

## EBLUP under Model 3

<table>
<thead>
<tr>
<th>Model and estimator</th>
<th>Average ARB (%)</th>
<th>Average RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain size class</td>
<td>Domain size class</td>
</tr>
<tr>
<td>Minor (20-69)</td>
<td>Medium (70-119)</td>
<td>Major (120+)</td>
</tr>
<tr>
<td>EBLUP</td>
<td>19.6</td>
<td>19.6</td>
</tr>
<tr>
<td>EBLUPW</td>
<td>3.4</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Model 3 \( y_i = \beta_0 + u_d + \beta_2 x_{2i} + \varepsilon_i \)
## EBLUP under Model 4

<table>
<thead>
<tr>
<th>Model and estimator</th>
<th>Average ARB (%)</th>
<th>Average RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain size class</td>
<td>Domain size class</td>
</tr>
<tr>
<td>Minor (20-69)</td>
<td>Medium (70-119)</td>
<td>Major (120+)</td>
</tr>
<tr>
<td>Minor (20-69)</td>
<td>Medium (70-119)</td>
<td>Major (120+)</td>
</tr>
</tbody>
</table>

**Model 4**  
\[ y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \epsilon_k \]

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Minor (20-69)</th>
<th>Medium (70-119)</th>
<th>Major (120+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBLUP</td>
<td>3.6</td>
<td>3.2</td>
<td>1.9</td>
</tr>
<tr>
<td>EBLUPW</td>
<td>3.3</td>
<td>2.7</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>5.2</td>
<td>6.0</td>
<td>5.6</td>
</tr>
</tbody>
</table>

**What can be learned?**

- **Bias can be large for a misspecified model**

- **Unequal probability sampling could be accounted for with two options**
  - Inclusion of the size variable into the model for model-dependent EBLUP
  - Use of a weighted version of EBLUP

- **The squared bias component can still dominate the MSE**
  - Can be difficult to obtain proper confidence intervals
References


References (contd.)


Estimation for domains and small areas with design-based and model-based methods - Topic 4

Risto Lehtonen
University of Helsinki

B-N-U Summer School, 23-27 August 2009, Kyiv, Ukraine
Issues to be discussed

- SAS software
  - SAS Procedures
  - SAS Macros
- SAS macro EBLUPGREG
- Program Domest
- R programs

SAS procedure SURVEYMEANS

- Design-based estimation for domains
- Estimation of domain totals and means
- BY statement
  - Proper estimation for planned domains with stratification
- DOMAIN statement
  - Proper estimation for unplanned domains
SYN and EBLUP estimation

- SAS procedure MIXED
  - Fitting of mixed linear models

- SAS procedure GLIMMIX
  - Fitting of generalized linear mixed models

- MSE estimation requires additional programming

- SAS macro EBLUPGREG

SAS macros

- EURAREA Project
  - http://www.statistics.gov.uk/eurarea/

- Design-based methods
  - GREG with linear fixed-effects models

- Model-based methods
  - Standard estimators
  - Estimators with spatial or temporal effects
  - Estimators for cross-classifications

- Proper MSE estimation
EURAREA Project

- The EURAREA "Standard" Estimators and performance criteria
  Office for National Statistics, UK
- Area-level Composite Estimator with Time-Varying Area Effect
  Office for National Statistics, UK
- EBLUPGREG: Unit-level Composite Estimator with Spatial or Temporal Effects
  Statistics Finland
- Unit-level Composite Estimator with Spatial Effects
  ISTAT, Italy
- Small Area Estimation with Sampling Weights
  INE, Spain / UMH, Spain
- Cross Classifications with Two-way and Three-way tables
  ISTAT, Italy

SAS macro EBLUPGREG

- GREG Generalised regression estimator
- SYN Synthetic estimator
- EBLUP Empirical linear unbiased predictor
- EBLUP with spatial correlation structure
- EBLUP with autocorrelated time effect
- EBLUP with autocorrelated time effect autocorrelation parameter fixed
- EBLUP with fixed time trend
- EBLUP with fixed categorical time effect
- EBLUP with time varying area effects
Program DOMEST

- Stand-alone Java program for estimation of totals and means for domains and small areas
- Developed by Dr. Ari Veijanen (Statistics Finland and University of Helsinki)
- Estimators
  - HT and Hájek estimator
  - GREG and SYN with linear model
  - GREG and SYN with linear mixed model
  - EBLUP with linear mixed model

Risto Lehtonen
R programs

- R programs are under development in different EU funded projects
  - SAMPLE project (EU FP7)
  - AMELI project (EU FP7)
- Others may exist…

THANK YOU FOR ATTENTION!

Questions, comments…?