HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

# ANALYSIS OF EUROPEAN DATA BY SMALL AREA METHODS

## Reweighting estimates from European sample surveys

**University of Pisa, 18 May 2016**

Risto Lehtonen, University of Helsinki

# Lecture topics

- Topic 1: Preliminaries
- Topic 2: Traditional GREG and calibration methods
- Topic 3: Extensions
- Topic 4: CASE STUDY: Perceived income for regional domains in Finland
- ANNEX Notation and inferential principles

# Topic 1: Preliminaries

- Important EC regulated surveys by NSIs
  - LFS
  - SILC      Quality reports
  - HBS

- Others
  - European Social Survey (academy-driven)
  - PISA survey (OECD)

- What might be the common properties of these types of surveys?

# European sample surveys – 1

- **Some properties**
  - Surveys are implemented in different statistical data infrastructures: survey-driven, register-driven
  - Multi-stage probability sampling designs are often used
    - Stratification, clustering, unequal probability sampling

    *Proper analysis requires methods to account for sampling complexities*
  - Observed data are contaminated by non-sampling errors
    - Nonresponse, measurement errors

    *Methods are needed to account for data contamination*
  - Published statistics are under high precision requirements, also for domain and small area estimates

    *Methods are needed to reduce standard errors*

# European sample surveys – 2

- **Weighting and reweighting**
- In a probability-based survey, a *design weight* is associated with each sampled unit
- The design weight can be interpreted as the number of typical units in the survey population that each sampled units represents
- Estimates can be calculated using the design weights or *estimation weights* obtained by adjusting the design weights
- Common adjustments include those that *account for nonresponse* and that *incorporate auxiliary information*

- Statistics Canada Quality Guidelines

# European sample surveys – 3

- **Weighting**
  - Accounting for stratification and unequal probability sampling with *design weight*
  - Design weight = inverse of inclusion probability
- **Reweighting for nonresponse**
  - Adjusting for *selection bias* caused by unit nonresponse
  - Lundström & Särndal (2005) Estimation in Surveys with Nonresponse. New York. Wiley
- **Reweighting to improve precision of estimates**
  - Calibration and generalized regression estimation, adapted for the estimation for domains and small areas
  - This is the topic of this mini course

# Estimation for domains

- The estimation of quantities for population subgroups called *domains* (small or large)
  - Total number of people in poverty for counties (SILC data)
  - Mean disposable income by municipality (SILC data)
  - Proportion of ILO unemployed in sex-age groups within provinces (LFS data)

- Small area estimation, SAE
  - Estimation for domains whose **sample size** is small or very small (even zero)
  - Alternative definition:

    Small area = Domain of interest for which the sample size is not adequate to produce reliable **direct estimates**

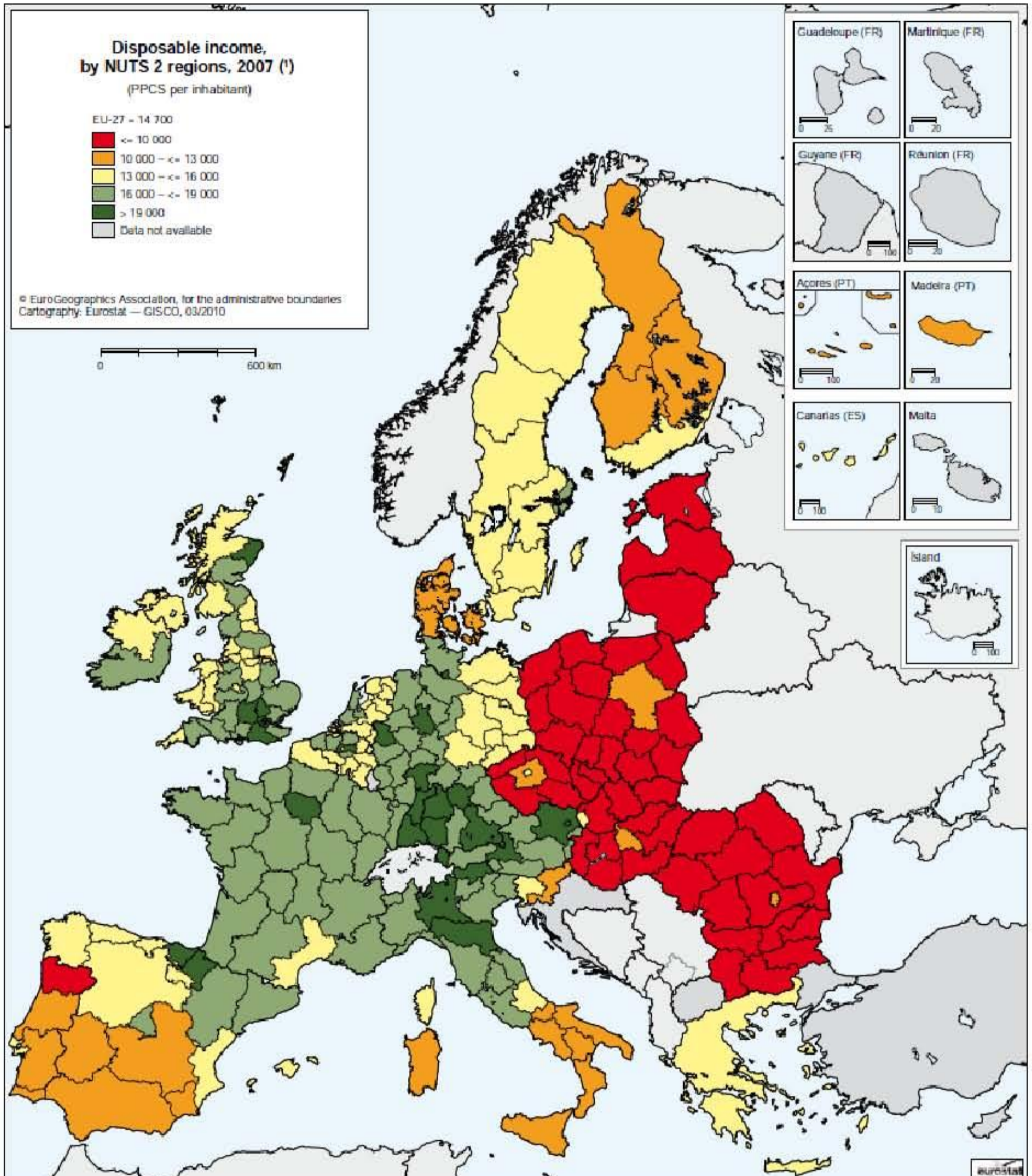
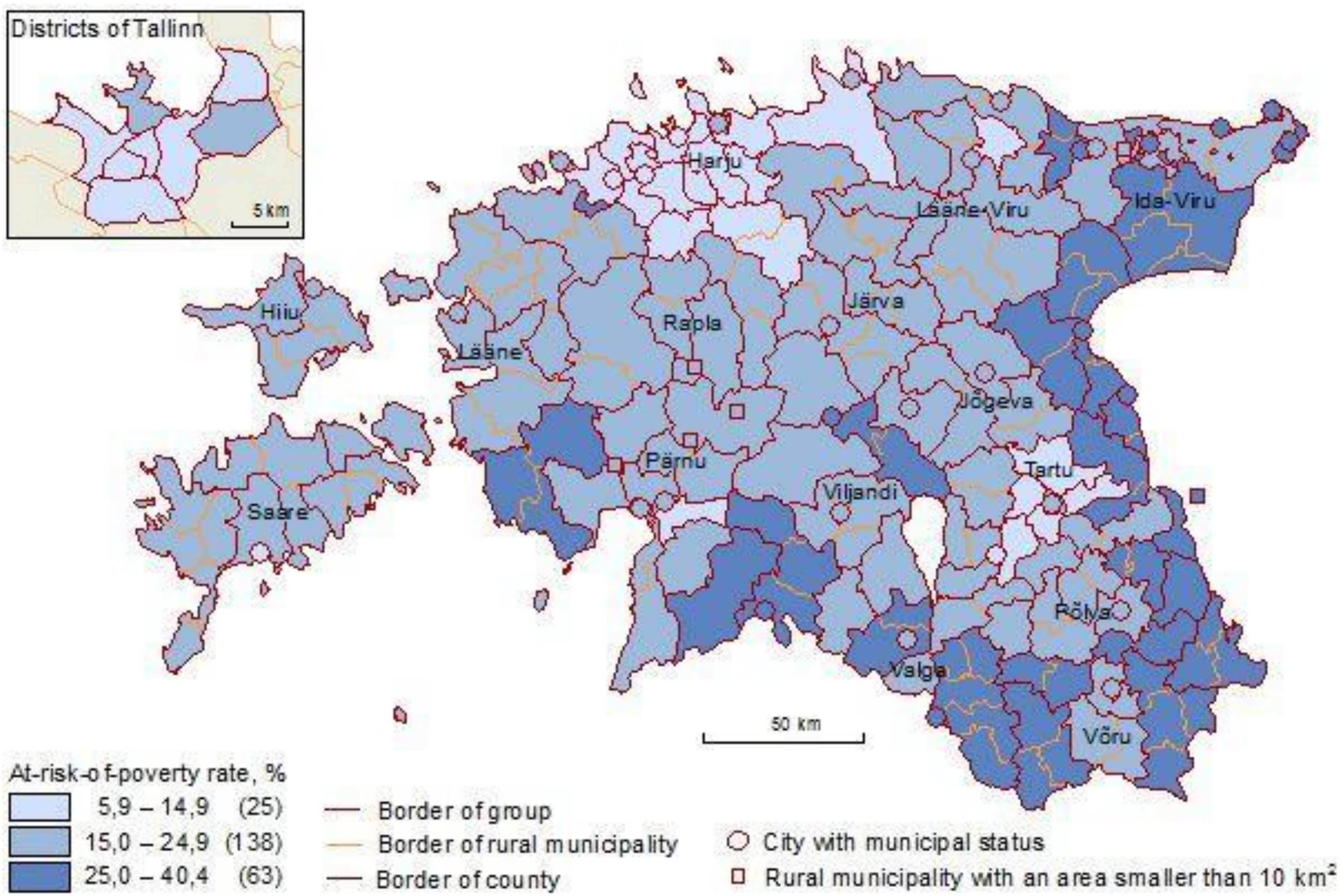



**Figure 1. Disposable income by NUTS 2 regions in 2007 in the European Union**

*Source: Eurostat Regional Yearbook 2010, p.93, Section on* *__Household Accounts__*. *Information about the metadata is available at* [*http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/EN/reg_ecohh_esms.htm*](http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/EN/reg_ecohh_esms.htm)

# Poverty map: Estonia

## World Bank 2014 – Regional poverty rates based on SILC data

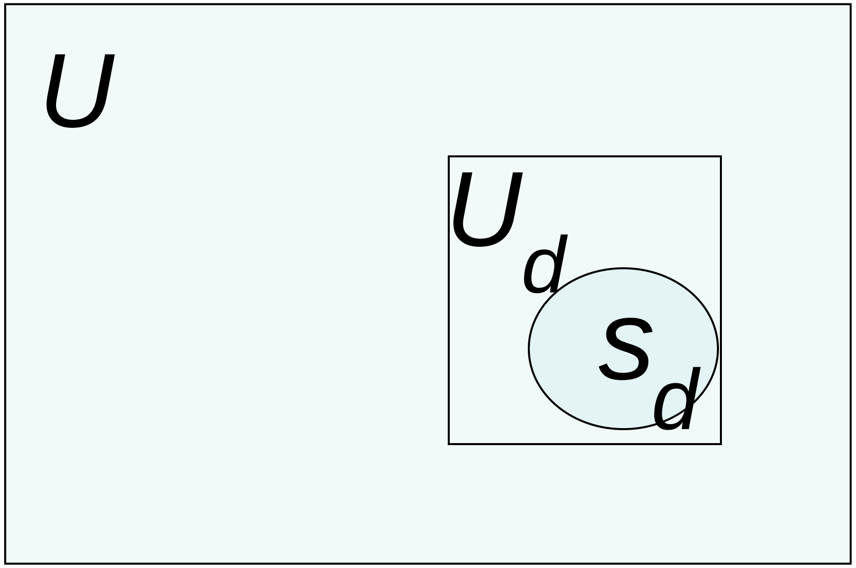


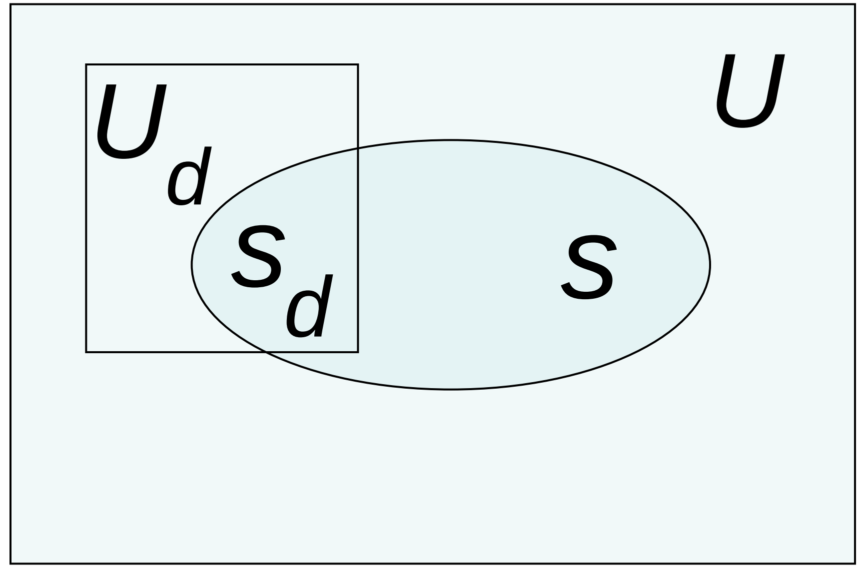
# Estimation for domains: Important aspects

- Type of domains of interest
  - Planned domains / Unplanned domains
- Type of domain estimator
  - Direct estimator / Indirect estimator
- Availability of auxiliary (population) data
  - Unit-level / Aggregate-level (area-level)
  - Sources: Census, Admin. registers, Statistical registers
- Type of model
  - Linear models/ Generalized linear models
  - Fixed-effects models / Mixed models
- Accuracy measures
  - Variance estimators / MSE estimators
- Computation tools
  - R (packages survey and sae), SAS (SURVEY procedures)
  - R package ReGenesees (ISTAT)

# Two main domain structures

- **Planned domains**
  - Most important domains are defined as strata
  - Strata are independent sub-populations
  - Domain sample sizes can be fixed in advance
  - Domain sample sizes are controlled by allocation scheme
  - Small sample sizes can be avoided if desired
- **Unplanned domains**
  - Domain sample sizes are not fixed but are random
  - Small domain sample sizes can occur
  - Most common case in small area estimation practice




**Planned domains**

$U$ Population

$U_d$ Population domain $d$, $d = 1,\ldots,D$

Domains = Strata

Several ($= D$) independent samples

Sample $s_d \subset U_d$ drawn in domain $d$

Sample size $n_d$ is **fixed** by sampling design

**Unplanned domains**

$U$ Population

A single sample $s$ is drawn

$s \subset U$ Sample

$U_d$ Population domain $d$, $d = 1,\ldots,D$

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

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

# Direct and indirect estimator

- **Direct estimator for domains**
  - *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 estimator for domains**
  - *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

# Domain type and estimator type

| Domain type | Estimator type | |
|---|---|---|
| | **Direct** | **Indirect** |
| **Planned** | Typical set-up | More rarely |
| **Unplanned** | More rarely | Typical set-up |

# “Borrow strength”

- *Indirect estimators* are attempting to “borrow strength” from other (similar) domains and/or in a temporal dimension
- For domains with small sample sizes, this is a well justified goal – Why?
- The concept of “borrowing strength” is often used in *model-based* small area estimation
  - Jon Rao (2015)
- Borrowing strength also is possible for *design-based model assisted* estimators
  - Lehtonen & Veijanen (2009)
- NOTE: Principles of design-based and model-based inference are summarized (very briefly) in ANNEX

# Key properties of estimators

Source: Lehtonen and Veijanen (2009)

Table 1
Design-based properties of model-assisted and model-dependent estimators for domains and small areas

| | Design-based model-assisted methods | Model-dependent methods |
|---|---|---|
| | GREG and calibration estimators | Synthetic and EBLUP estimators |
| Bias | Design unbiased (approximately) by the construction principle | Design biased<br>Bias does not necessarily approach zero with increasing domain sample size |
| Precision (Variance) | Variance may be large for small domains<br>Variance tends to decrease with increasing domain sample size | Variance can be small even for small domains<br>Variance tends to decrease with increasing domain sample size |
| Accuracy (MSE) | MSE = Variance (or nearly so) | MSE = Variance + squared bias<br>Accuracy can be poor if the bias is substantial |
| Confidence intervals | Valid design-based intervals can be constructed | Valid design-based intervals not necessarily obtained |

EXAMPLE. Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005): Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* 7, 649-673.

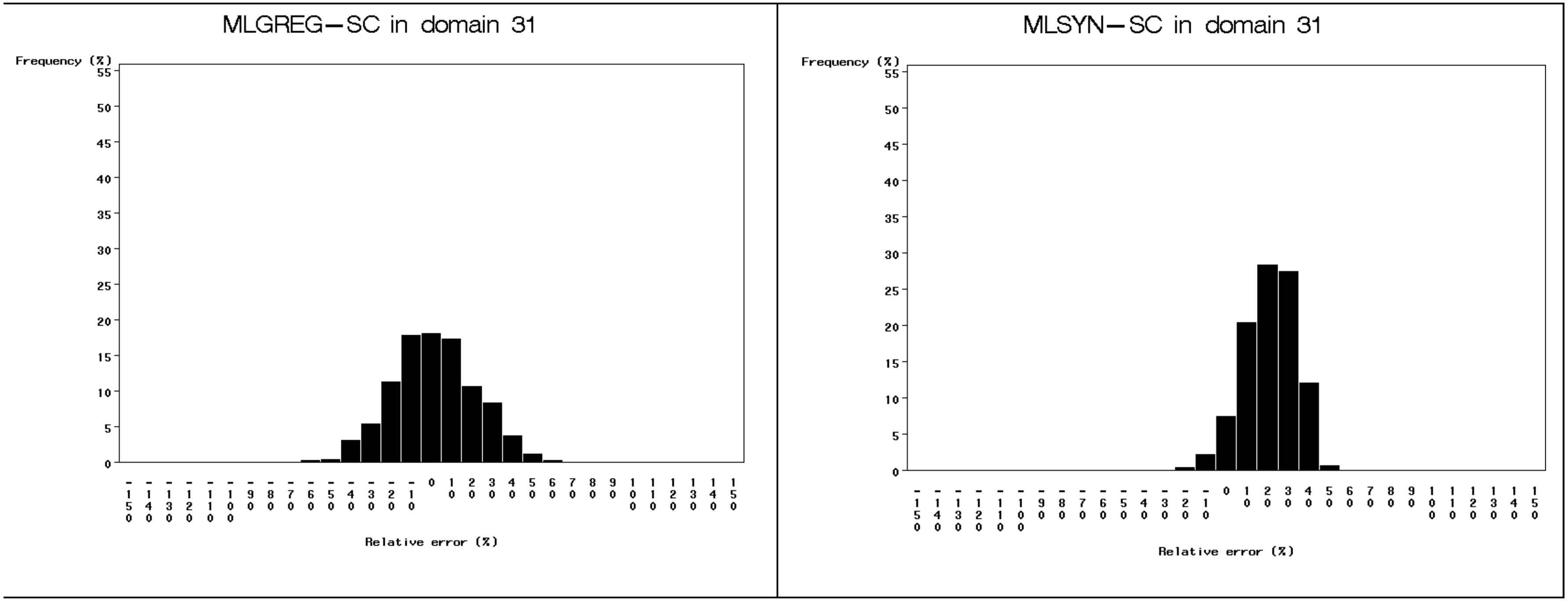



**FIGURE 1** Distribution of relative error (%) of design-based MLGREG (left-hand side) and model-based MLSYN (right-hand side) estimators in domain 31 of the generated LFS population. (Design-based simulation experiment, 1,000 independent simple random samples of 12,000 elements from population of three million elements and 84 domains)

Relative error of an estimator $\hat{t}_d$ for sample $s_i$, $i = 1,\ldots,1000$, in domain *d* is defined as

$$\mathrm{RE}(\hat{t}_d) = (\hat{t}_d(s_i) - t_d)/t_d,\; d = 1,\ldots,84$$

# Discussion

- Previous example:

- MLGREG: *design-based generalized regression* (GREG) estimator assisted by logistic mixed model
- MLSYN: *model-based synthetic* estimator with the same underlying logistic mixed model formulation as GREG

- Which one is:
  - Design unbiased?
  - More accurate?

- NOTE: Trade-off between bias and accuracy!

# Topic 2: Traditional GREG and calibration methods

- **Generalized regression (GREG) estimators** and **calibration methods** provide *design-based* methods for the estimation of population and sub-population parameters
- GREG and calibration estimators are (approximately) design unbiased
- Estimation of precision (variance and standard error) of estimators is straightforward
- Basic goal: Improvement of precision over "standard" methods (e.g. *Horvitz-Thompson (HT) estimator*) by incorporating auxiliary data in the estimation procedure
- GREG and calibration methods are extensively used in official statistics (e.g. Statistics Finland and ISTAT)
- Särndal, Swensson & Wretman (1992)
- Deville & Särndal (1992)

# Extended GREG family

- *Traditional GREG estimators* (Särndal et al.) are designed for population total of continuous study variable
  - Assisting models: Linear fixed-effects models
  - Therefore, this GREG is called *linear GREG estimator*
  - Auxiliary data: Population totals of auxiliary variables
  - Examples: Regression estimation, ratio estimation and post-stratification for totals of continuous study variable
- *Extended family of GREG estimators* is designed for population cell frequencies or totals of binary, polytomous and count variables
  - Assisting models: Generalized linear mixed models (GLMMs)
  - Auxiliary data: Values of auxiliary variables at the *unit level* for all population elements
  - Example: Logistic GREG estimator for population frequencies of polytomous study variable (Lehtonen & Veijanen (1998)

# NOTE on auxiliary data

- **"Survey" countries**
  - Auxiliary data are often available at *aggregate level* (population totals and frequency distributions)
  - Sample survey data and auxiliary data cannot be merged at the unit level
- **"Register" countries**
  - Auxiliary data from statistical registers are available at the *unit level* (values of auxiliary variables for all population elements) and can be micro-merged with sample survey data by using identification keys
  - This option involves more flexible estimation than for "survey" countries
- Many countries in Europe and elsewhere have developed, or are turning towards, register-driven data infrastructures

# Traditional linear GREG estimator

- *GREG = Generalized regression estimator*

- Robinson P.M. and Särndal C.-E. (1983) Asymptotic properties of the generalized regression estimator in probability sampling, *Sankhyā Ser. B,* 45, 240–248.

- Särndal, C.E. (1980) On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 67, 639–650.

- Särndal C.-E., Swensson B. and Wretman J. (1992) *Model-Assisted Survey Sampling*. New York: Springer.

# GREG principle - 1

**Difference estimator** of population total $t$ of $y$ (Särndal 1980)

By assuming known $y_k^0$, $k \in U$, write the unknown population total as

$$t = \sum_{k \in U} y_k = \sum_{k \in U} y_k^0 + \sum_{k \in U} (y_k - y_k^0)$$

where $y_k$ is (unknown) population value of study variable $y$

Assume sample $s$ that includes y-values $y_k$ and x-values $x_k$

Difference estimator: Estimate the second sum from sample using HT:

$$\hat{t}_{DIFF} = \sum_{k \in U} y_k^0 + \sum_{k \in s} a_k (y_k - y_k^0)$$

where $a_k = 1/\pi_k$ is design weight, $k \in s \subset U$

NOTE: $U$ refers to population, $s$ refers to sample

# GREG principle - 2

In practice, such a magic $y_k^0,\ k \in U,$ rarely exists...

Instead, let us use sample data, modelling and auxiliary information

Assume known auxiliary variable vector $\mathbf{x}_k = (1, x_k)'$ for every $k \in U$

Specify linear fixed-effects model:

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k = \beta_0 + \beta_1 x_k + \varepsilon_k,\ Var(\varepsilon_k) = \sigma^2, \text{ where } \boldsymbol{\beta} = (\beta_0, \beta_1)'$$

Fit the model to sample data $s$ and obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Calculate **predicted values** $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ for every $k \in U$

by using the estimated model and x-data $\mathbf{x}_k,\ k \in U$

We obtain **model-assisted GREG estimator** of the total $t$:

$$\hat{t}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} a_k (y_k - \hat{y}_k)$$

# GREG principle - 3

Simple **variance estimator** of $\hat{t}_{GREG}$ under SRS sampling:

Assisting model: $y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$

GREG estimator: $\hat{t}_{GREG} = \sum_{k \in U} \hat{y}_k + \frac{N}{n} \sum_{k \in s} (y_k - \hat{y}_k)$

where $N$ is population size and $n$ is sample size

Variance estimator: $\hat{V}_{SRS}(\hat{t}_{GREG}) = \hat{V}_{SRS}(\hat{t}_{HT})(1 - \hat{\rho}_{yx}^2)$

where $\hat{V}_{SRS}(\hat{t}_{HT})$ is variance estimator of SRS-based (HT) estimator

$$\hat{t}_{HT} = \sum_{k \in s} a_k y_k = \frac{N}{n} \sum_{k \in s} y_k$$

and $\hat{\rho}_{yx}$ is sample correlation of y and x

Think about **efficiency improvement** of $\hat{t}_{GREG}$ when compared to $\hat{t}_{HT}$

# NOTES on GREG

- Linear GREG estimators are called *model assisted* because models are explicitly specified and used as assisting tools in incorporating auxiliary x-data in the estimation process
- Even finding a good model for y-variable is important for efficiency improvement, the interest is not in the model itself but in the target indicator (total in this case)
- NOTE: Models can involve several x-variables

- **Expected gain in GREG estimation:**
- *Improved efficiency* (decrease of standard error relative to Horvitz-Thompson estimator) if y-variable and x-variable are correlated

- NOTE: In addition to efficiency improvement GREG is often used in adjusting for *unit nonresponse*

# Traditional calibration estimator

- *Calibration estimators*

- Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *JASA* 87, 376–382.

- Estevao V.M. and Särndal C.-E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology* 2, 213-221.

- Särndal C.-E. (2007) The calibration approach in survey theory and practice. *Survey Methodology* 33, 99–119.

# Calibration principle

Aim: Estimation of population total $t = \sum_{k \in U} y_k$ from sample $s \subset U$

Assume again access to auxiliary x-data $\mathbf{x}_k = (1, x_k)'$, $k \in U$

Assume sample $s$ that includes y-values $y_k$ and x-values $x_k$

Construct weights $w_k = a_k g_k$ that fulfil **calibration equation**:

$$\sum_{k \in s} w_k \begin{pmatrix} 1 \\ x_k \end{pmatrix} = \sum_{k \in U} \begin{pmatrix} 1 \\ x_k \end{pmatrix} = \begin{pmatrix} N \\ t_x \end{pmatrix} = \begin{pmatrix} N \\ \sum_{k \in U} x_k \end{pmatrix}$$

where $a_k = 1/\pi_k$ is design weight and $g_k$ is g-weight for element $k \in s$

NOTE: This means that $\sum_{k \in s} a_k g_k = N$ and $\sum_{k \in s} a_k g_k x_k = t_x$

Use the *calibrated weights* to estimate the y-variable total:

$$\hat{t}_{CAL} = \sum_{k \in s} w_k y_k$$

# NOTES on calibration

- Traditional calibration (Deville & Särndal 1992) is called *model-free calibration* because models are not explicitly specified to obtain calibration weights
- Only x-data are needed (both in sample and in population)
- NOTE: Calibration can involve several x-variables
- NOTE: In *model calibration*, models are used explicitly

- **Expected gains in calibration:**
- *Calibration property*: Coherence of sample estimates of x-variable totals with known population totals
- *Improved efficiency* (decrease of standard error relative to Horvitz-Thompson estimator) if y-variable and x-variable are correlated

- NOTE: Calibration can be used for nonresponse adjustment

# RECALL: Direct and indirect estimators for domains

- **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

# IMPORTANT NOTE

- **Planned domains**
- Domains of interest coincide with the strata
- Domain sample sizes are fixed in the sampling design
- In estimation for domains, the domains of interest can be treated as independent sub-populations
- *Standard GREG and calibration estimators for the whole population can be applied separately for each domain*

- **Unplanned domains**
- A single sample is drawn from population
- Domain sample sizes are not under control but are random
- Both small and large domain sample sizes can realize
- *Additional methods must be introduced to account for these complexities*

# SIMPLE EXAMPLE: Ratio-type GREG

Assume continuous y-variable and one continuous auxiliary x-variable

Domains of interest $U_d,\ d = 1,\ldots,D$

Assisting linear fixed-effects models in two cases:

a) Planned domains case: $y_k = \beta_d x_k + \varepsilon_k,\ \ k \in U_d,\ d = 1,\ldots,D$

b) Unplanned domains case: $y_k = \beta x_k + \varepsilon_k,\ \ k \in U$

NOTE: Intercept parameters $\beta_{0d} = \beta_0 = 0$

NOTE: Models a) and b) are different. In what essential way?

For both domain types, let us construct a GREG estimator of domain total of y-variable

# a) Direct GREG estimator for domains

Assisting model: $y_k = \beta_d x_k + \varepsilon_k, \quad k \in U_d,\ d = 1,\ldots,D$

By noting that $\hat{\beta}_d = \dfrac{\hat{t}_{dHT}}{\hat{t}_{dxHT}}$ and $\hat{y}_k = \hat{\beta}_d x_k$ we have:

$$\hat{t}_{dRAT} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

$$= \hat{t}_{dHT} + \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}} (t_{dx} - \hat{t}_{dxHT})$$

$$= t_{dx} \times \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}}\ ,\ d = 1,\ldots,D$$

which is standard textbook form of *ratio estimator*

Why this GREG estimator is direct?

NOTE: Auxiliary information needed: x-totals $t_{dx}$ for domains

# b) Indirect GREG estimator for domains

Assisting model: $y_k = \beta x_k + \varepsilon_k, \quad k \in U$

By noting that $\hat{\beta} = \dfrac{\hat{t}_{HT}}{\hat{t}_{xHT}}$ and $\hat{y}_k = \hat{\beta} x_k$ we have

$$\hat{t}_{dRAT} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

$$= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{xHT}} (t_{dx} - \hat{t}_{dxHT})$$

which is standard textbook form of *regression estimator*

using aggregate auxiliary information

Why this GREG estimator is indirect?

NOTE: Auxiliary information needed: x-totals $t_{dx}$ for domains

# Simple example

- **Hypothetical example in estimation of domain totals**
- Demonstration of direct and indirect GREG estimators a) and b) and comparison with Horvitz-Thompson (HT) estimator

- Population: $N$=966 units, sample: $n$=100
- Sampling with simple random sampling without replacement (SRSWOR)

- *Planned domains*: domains are taken as strata and a sample is drawn from each stratum with proportional allocation such that the total sample size $n$=100
- *Unplanned domains*: A single sample of $n$=100 is drawn

- Study variable y, explanatory (auxiliary) variable x
- Correlation cor(y,x)=0.83
- Varies between domains: range 0.15 to 0.96

# **Results** (sorted by domain sample size)

| Domain ID | Population size $N_d$ | Sample size $n_d$ | Population total $t_d$ | Estimates of domain totals | | |
|---|---|---|---|---|---|---|
| | | | | Direct HT $\hat{t}_{dHT}$ | a) Direct GREG $\hat{t}_{dGREG}$ | b) Indirect GREG $\hat{t}_{dGREG}$ |
| 7 | 46 | 3 | 835 | 553 | 748 | 740 |
| 8 | 47 | 3 | 1022 | 632 | 1029 | 1054 |
| 9 | 40 | 3 | 884 | 638 | 803 | 839 |
| 1 | 69 | 5 | 1299 | 911 | 1231 | 1215 |
| 5 | 86 | 8 | 1738 | 1572 | 1585 | 1586 |
| 3 | 94 | 10 | 1839 | 1735 | 1956 | 1945 |
| 4 | 86 | 12 | 1865 | 2378 | 1904 | 1901 |
| 2 | 120 | 13 | 2533 | 2663 | 2622 | 2623 |
| 10 | 174 | 17 | 3594 | 3611 | 3707 | 3704 |
| 6 | 204 | 26 | 4663 | 5693 | 4738 | 4729 |
| All | 966 | 100 | | | | |

# Indirect GREG estimator for domains - 1

**Indirect GREG estimator** of domain totals

$$t_d = \sum_{k \in U_d} y_k,\ d = 1,\dots,D$$

Assume known vector values of auxiliary x-data with $J$ variables

$$\mathbf{x}_k = (1, x_{1k}, \dots x_{Jk})',\ k \in U$$

Assisting linear fixed-effects model:

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k,\ Var(\varepsilon_k) = \sigma^2,\ k \in U$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_J)'$ are beta coefficients common for all domains

Parameter $\boldsymbol{\beta}$ is estimated from the sample $s$ by

weighted least squares with weights $a_k = 1/\pi_k$ :

$$\hat{\boldsymbol{\beta}} = \left( \sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} a_k \mathbf{x}_k y_k$$

# Indirect GREG estimator for domains - 2

Fitted values

$$\hat{y}_k = \mathbf{x}'_k \hat{\boldsymbol{\beta}}, \ k \in U$$

and sample residuals

$$e_k = y_k - \hat{y}_k, \ k \in s$$

are incorporated into **indirect GREG estimator**

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k, \ \ d = 1, \ldots, D$$

NOTE: This GREG is indirect since all y-values in the sample contribute to the predicted values $\hat{y}_k, \ k \in U$

# Some notes on efficiency

- The model is not domain specific but is specified for the whole population
- This means borrowing strength for given (possibly small) domain from other "similar" (possibly larger) domains
- Efficiency improves if explanatory power of x-variables in the model is good involving small residuals

EXAMPLE: Variance estimator of (direct) $\hat{t}_{dGREG}$

$$\hat{V}_1\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) e_k e_l \text{ where } e_k = y_k - \hat{y}_k$$

Compare with variance estimator of (direct) HT estimator

$$\hat{V}_1\left(\hat{t}_{dHT}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) y_k y_l$$

Think about possible efficiency improvement over HT

# Examples of assisting models

Linear fixed-effects models:

Common model with $J$ x-variables for all domains

$$y_k = \beta_0 + \beta_1 x_k + \ldots + \beta_J x_{Jk} + \varepsilon_k,\ k \in U$$

Domain-specific fixed intercepts and common slopes

$$y_k = \beta_{01} I_{1k} + \beta_{02} I_{2k} + \ldots + \beta_{0D} I_{Dk} + \beta_1 x_k + \ldots + \beta_J x_{Jk} + \varepsilon_k,\ k \in U$$

where $I_{dk} = I\{k \in U_d\}$ (domain membership indicator)

NOTE: Completely domain specific model involves direct GREG estimator for domains:

$$y_k = \beta_{01} I_{1k} + \beta_{02} I_{2k} + \ldots + \beta_{0D} I_{Dk} + \beta_{1d} x_k + \ldots + \beta_{Jd} x_{Jk} + \varepsilon_k,\ k \in U_d$$

# GREG as calibration estimator

Indirect GREG can be written as a weighted sum of observations incorporating *calibrated weights* (g-weights) $w_k = a_k g_{dk}$:

$$\hat{t}_{dGREG} = \sum_{k \in s_d} w_k y_k = \sum_{k \in s_d} a_k g_{dk} y_k$$

where $g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$ are *extended* g-weights

$I_{dk} = I\{k \in U_d\}$ is domain membership indicator

such that $I_{dk} = 1$ if $k \in U_d$, 0 otherwise

$\hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$ NOTE: Extends over the whole sample $s$

NOTE: **Calibration property** holds for all x-variables $x_j$, $j = 1,\ldots,J$:

$$\hat{t}_{dx_j GREG} = \sum_{k \in s_d} a_k g_{dk} x_{jk} = \sum_{k \in U_d} x_{jk} = t_{dx_j}$$

# Variance estimator of indirect GREG with g-weights

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

where $e_k = y_k - \hat{y}_k$ are sample residuals

$$g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k \text{ with } \hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$$

Extended g-weights $g_{dk}$ are used

The whole sample data set $s$ is used to estimate variance for given domain $d$

NOTE: $\hat{V}\left(\hat{t}_{dGREG}\right)$ requires weights $a_{kl} = 1/\pi_{kl}$

where $\pi_{kl}$ are second-order inclusion probabilities

They are intractable for practical variance estimation

# More practical variance estimator

**Approximate variance estimator** of GREG

by using *extended residuals*:

$$\hat{V}_U\left(\hat{t}_{dGREG}\right) = \frac{n}{n-1}\sum_{k\in s}\left(a_k e_{dk} - \hat{t}_{dHTe} / n\right)^2$$

where $n$ is the total sample size and $a_k = 1/\pi_k$ (design weights)

$e_{dk} = I\{k \in U_d\}e_k$ are extended residuals, where $e_k = y_k - \hat{y}_k$

NOTE: $e_{dk} = e_k$ if $k \in s_d$ and $e_{dk} = 0$ if $k \notin s_d$

$\hat{t}_{dHTe} = \sum_{k\in s_d} a_k e_k$ is HT estimator of residual total in domain $d$

NOTE: This form resembles variance estimator for PPSWR and is used in some software (e.g. RDomest software)

# Indirect GREG – textbook form

Since assisting model in traditional GREG estimator is linear, GREG estimation does not require unit-level information on $\mathbf{x}_k$

It is enough to have access to the vector $\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k$ of domain totals of auxiliary x-variables in the population and the corresponding HT estimates $\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} \mathbf{x}_k$ in the sample

Standard textbook form:

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}, \text{ where } \hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

# EXAMPLE: GREG estimation for domains with real data

- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

- Section 4.2. Computational example with direct and indirect estimation under an unplanned domain structure

- Summary leaflet: Comparison of results of direct HT estimator with direct GREG and indirect GREG estimators

# Data

- Real data from statistical registers of Statistics Finland

- Population: $N$ = 431,000 households from Western Finland
- Domains: $D$ = 12 NUTS4 regions (domains)

- Household sampling: πPS (PPS-WOR)
- Size variable in PPS-WOR: Number of household members (obtained from statistical register)

- Sample size: $n$ = 1000 households

# Finland

# Variables

- **Study variable** *y*
  - Disposable household income
- **Auxiliary x-variables (known for all HHs)**
  - EMP: the number of months in total the household members were employed during last year
  - EDUC: the number of household members who had higher education
  - Variables are derived from administrative registers
  - Domain sizes in population and domain totals of EMP and EDUC are assumed known

- NOTE: We have access to population values of our study variable y and auxiliary x-variables
- This gives option to compare results with true values

# Quality measures of estimators

ARE Absolute relative error of an estimator $\hat{t}_d$ in domain *d*

$$\text{ARE}(\hat{t}_d) = |\hat{t}_d - t_d| / t_d,\ d = 1,\ldots,D$$

where $t_d$ is known true total

MARE: Mean ARE calculated in three domain size classes

MCV Mean coefficient of variation of the estimate in three domain size classes

Coefficient of variation is calculated as

$$CV(\hat{t}_d) = s.e(\hat{t}_d) / \hat{t}_d$$

# Estimators of domain totals
## a) Direct GREG for planned domains

- HT estimator and variance estimators
- Direct GREG estimator and variance estimators

Parameter: Domain totals $t_d = \sum_{k \in U_d} y_k,\ d = 1,\ldots,12$

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

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

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}_d$$

$$\hat{V}_A\left(\hat{t}_{dGREG}\right) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} \left(n_d a_k e_k - \hat{t}_{dHTe}\right)^2$$

# Assisting models in GREG

## Planned domains

Direct GREG estimator with linear fixed-effects assisting model and domain-specific terms

$y_k = \beta_{0d} + \beta_{1d}\mathrm{EMP}_k + \varepsilon_k$ (column 2), or

$y_k = \beta_{0d} + \beta_{1d}\mathrm{EMP}_k + \beta_{2d}\mathrm{EDUC}_k + \varepsilon_k$ (column 3)

NOTE: Domain-specific intercepts and slopes

Therefore, this GREG is direct

**Table 2.** Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and direct calibration (GREG) estimators of totals for minor, medium-sized and major domains by using various amounts of auxiliary information for **planned domains**.

| | HT | | GREG | | | |
|---|---|---|---|---|---|---|
| | Auxiliary information | | | | | |
| | 1<br>None | | 2<br>Domain sizes and domain totals of EMP | | 3<br>Domain sizes and domain totals of EMP and EDUC | |
| Domain sample size class | MARE % | **MCV %** | MARE % | **MCV %** | MARE % | **MCV %** |
| Minor<br>$8 \le n_d \le 33$ | 11.5 | **11.9** | 5.8 | **7.7** | 6.4 | **6.8** |
| Medium<br>$34 \le n_d \le 45$ | 7.6 | **9.0** | 3.7 | **8.0** | 3.6 | **8.1** |
| Major<br>$46 \le n_d \le 277$ | 12.5 | **5.2** | 4.3 | **4.7** | 5.2 | **3.7** |

# Estimators of domain totals
## b) Indirect GREG for unplanned domains

- HT estimator and variance estimators
- Indirect GREG estimator and variance estimators

Parameter: Domain totals $t_d = \sum_{k \in U_d} y_k,\ d = 1,\ldots,12$

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

$$\hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{n}{n-1} \sum_{k \in s} \left(a_k y_{dk} - \hat{t}_{dHT} / n\right)^2$$

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}$$

$$\hat{V}_U\left(\hat{t}_{dGREG}\right) = \frac{n}{n-1} \sum_{k \in s} \left(a_k e_{dk} - \hat{t}_{dHTe} / n\right)^2$$

# Assisting models in GREG

## Unplanned domains

Indirect GREG estimator is assisted by a linear fixed-effects model

$$y_k = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k$$

fitted to the whole sample

NOTE: Common intercept and slope for all domains

Therefore, this GREG is indirect

**Table 3.** 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 for **unplanned domains**.

| | **HT** | | **GREG** | |
|---|---|---|---|---|
| | Auxiliary information | | | |
| | 1 None | | 2 Domain sizes and domain totals of EMP | |
| Domain sample size class | MARE % | **MCV %** | MARE % | **MCV %** |
| Minor $8 \leq n_d \leq 33$ | 11.5 | **28.3** | 7.6 | **9.0** |
| Medium $34 \leq n_d \leq 45$ | 7.6 | **20.3** | 3.8 | **8.1** |
| Major $46 \leq n_d \leq 277$ | 12.5 | **9.6** | 4.1 | **5.0** |

# Lessons learned from examples a) and b)

- 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
  - For both HT and GREG, accuracy tends to be better in planned domains case
- Stratification for important domains of interest makes sense! This is an issue of the survey planning stage!
- However, the unplanned case and indirect methods are much more common in practice

# TOPIC 3: Extensions

- Traditional *linear GREG* and *model-free calibration* methods use *linear fixed-effects models for continuous study variables*
- More general model families are needed to cover *binary*, *polytomous* and *count* type study variables
- Generalized linear (fixed-effects) models (GLM)
  Nelder & Wedderburn (1972) JRSS-A
  McCullagh & Nelder (1982) Generalized Linear Models. Springer.
- Generalized linear mixed models (GLMM) family models
  Demidenko (2005) Mixed Models: Theory and Applications. Wiley.
- These model types are used in *extended family of GREG estimators for domains* and *model calibration estimators for domains*

# EXAMPLE: Assisting model in GREG and model calibration - 1

**Linear mixed model** for continuous study variable $y$

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k,\ k \in U_d,\ d = 1,\ldots,D$$

where $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$

$u_d$ are domain-level random intercepts

$u_d \sim N(0, \sigma_u^2)$, $\varepsilon_k \sim N(0, \sigma^2)$, $u_d$ and $\varepsilon_k$ independent

Estimate $\boldsymbol{\beta}$ and $\sigma_u^2$ from the data

Calculate estimates $\hat{u}_d,\ d = 1,\ldots,D$ and calculate fitted values

$$\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d,\ k \in U_d,\ d = 1,\ldots,D$$

Used in linear mixed model assisted GREG estimator (MGREG)

(Lehtonen, Särndal and Veijanen (2003)

# EXAMPLE: Assisting model in GREG and model calibration - 2

**Logistic fixed - effects model**

for binary response variable $y$

$$E_m(y_k) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta})}$$

Estimate $\boldsymbol{\beta}$ from the data

Calculate fitted values $\hat{y}_k = \dfrac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})}, \quad k \in U$

Used in logistic model assisted GREG estimator (LGREG)

(Lehtonen and Veijanen (1998)

# EXAMPLE: Assisting model in GREG and model calibration - 3

**Logistic mixed model** for binary response variable *y*

$$E_m(y_k \mid u_d) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}, \quad k \in U_d,\ d = 1,\dots,D$$

where $u_d$ are domain-level random intercepts, $u_d \sim N(0, \sigma_u^2)$

Estimate $\boldsymbol{\beta}$ and $\sigma_u^2$ from the data

Calculate estimates $\hat{u}_d,\ d = 1,\dots,D$ and calculate fitted values:

$$\hat{y}_k = \frac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}, \quad k \in U_d,\ d = 1,\dots,D$$

Used in logistic mixed model assisted GREG estimator (MLGREG) (Lehtonen, Särndal and Veijanen (2005)

# GLMM assisted GREG estimator

- For any assisting GLMM for GREG the formulation of GREG estimator for domain total and mean or proportion remain the same. The difference is in obtaining predicted y-values

MGREG estimator for domain total $t_d$ of continuous y

Assisting model: Linear mixed model

Predicted values: $\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d,\ k \in U_d,\ d = 1,\ldots,D$

MLGREG for domain proportion $p_d$ of binary y

Assisting model: Logistic mixed model:

Predicted values: $\hat{y}_k = \dfrac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)},\quad k \in U_d,\ d = 1,\ldots,D$

For MGREG and MLGREG the estimator is of the same form:

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

# Recall: Data requirements

- Traditional linear GREG estimator and model-free calibration estimator
  - Unit-level x-vectors not necessarily needed
  - Known domain totals of x-variables only are needed
  - Applicable in ”survey” countries in particular
- Extended GREG family estimators and model calibration estimators
  - Unit-level x-data are needed for all units in population
  - Applicable in ”register” countries
  - Applicable also in ”survey” countries if for example census data (or population data from another reliable register source) can be merged with sample survey data at the unit level

# Estimation of the model

- GLMMs can be fitted for example by:
  - R packages `nlme` or `lme4 (glmer` function) using maximum likelihood
  - SAS procedures GLIMMIX (using ML) or MIXED (using REML or ML)

- Some methodological references
  - Datta (2009)
  - Jiang and Lahiri (2006)
  - Rao (2003, 2015)

# NOTE on the role of models

- The role of model differs in *model-assisted design-based* estimators and *model-based* estimators

  - Model assisted (GREG) uses models as assisting tools
    - This is to avoid design bias
    - Cost to be paid is poor accuracy in small domains

  - Model-based (SYN, EBLUP, EBP) rely solely on models
    - A benefit is better accuracy in small domains
    - Cost to be paid is the risk of design bias

    - NOTE: Recall trade-off between bias and accuracy!

# Model calibration

- Idea: Extension of model-free calibration beyond linear models for continuous study variables to cover nonlinear models for continuous variables and GLMs and GLMMs for binary, polytomous and count type study variables

- Calibration principle in domain estimation: Calibration of totals of *model predictions* estimated from sample to agree with population totals of model predictions
- NOTE: difference w.r.t. model-free calibration

- Model calibration: Wu and Sitter (2001), Montanari and Ranalli (2005, 2009)
- Model calibration for domains: Lehtonen and Veijanen (2016a,b)

# Calibration estimators for totals

Domain totals $t_d = \sum_{k \in U_d} y_k,\ d = 1,\ldots,D$

Calibration estimators

$$\hat{t}_d = \sum_{k \in s_d} w_k y_k = \sum_{k \in s_d} a_k g_k y_k$$

$a_k = 1/\pi_k$ design weight

$g_k$ method-specific g-weight for element $k \in s$

$w_k$ method-specific *calibration weight* for element $k$

$\pi_k$ inclusion probability for element $k$

$s_d \subset U_d$ planned domains case

$s_d = s \cap U_d$ unplanned domains case

# Calibration weights for model-free calibration

Calibration estimator $\hat{t}_{dMFC} = \sum_{k \in s_d} w_k^{MFC} y_k$ for domain total $t_d$

*Calibration equation* for model-free calibration

$$\sum_{k \in s_d} w_k \mathbf{x}_k = \sum_{k \in U_d} \mathbf{x}_k = \left( N_d, \sum_{k \in U_d} x_{1k}, \ldots, \sum_{k \in U_d} x_{Jk} \right)'$$

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{Jk})'$ *calibration vector* for $k \in U$

Minimize chi-square distance to design weights $a_k = 1/\pi_k$

$$\sum_{k \in s_d} \frac{(w_k - a_k)^2}{a_k}$$

Calibration weights for unit $k \in s_d$, $d = 1, \ldots, D$:

$$w_k^{MFC} = a_k \left( 1 + \left( \sum_{i \in U_d} \mathbf{x}_i - \sum_{i \in s_d} a_i \mathbf{x}_i \right)' \left( \sum_{i \in s_d} a_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \mathbf{x}_k \right)$$

# General case

Calibration weights $w_k$ minimize

$$\sum_{k \in s_d} \frac{(w_k - a_k)^2}{a_k} - \boldsymbol{\lambda}' \left( \sum_{k \in s_d} w_k \mathbf{z}_k - \sum_{k \in U_d} \mathbf{z}_k \right)$$

where $a_k = 1/\pi_k$, $d = 1, \ldots, D$

$\mathbf{z}_k$ is **method-specific** vector of calibration variables

Calibrated weights are defined in:

$w_k = a_k(1 + \boldsymbol{\lambda}'\mathbf{z}_k)$, where $\boldsymbol{\lambda}$ is the Lagrange coefficient

$$\boldsymbol{\lambda}' = \left( \sum_{i \in U_d} \mathbf{z}_i - \sum_{i \in s_d} a_i \mathbf{z}_i \right)' \left( \sum_{i \in s_d} a_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1}$$

# Model calibration equation

$$\sum_{k \in s_d} w_k^{MC} \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k = \left( N_d, \sum_{k \in U_d} \hat{y}_k \right)'$$

where calibration vector is $\mathbf{z}_k = (1, \hat{y}_k)'$

$\hat{y}_k$ are predicted values of y calculated for every $k \in U$

by using the model fitted with the entire sample data set

**Semi-direct MC estimator**: $\hat{t}_{dMC} = \sum_{k \in s_d} w_k^{MC} y_k, \quad d = 1, \ldots, D$

Examples of assisting GLMMs in MC

Linear mixed model:

Predicted values: $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}} + \hat{u}_d, \; k \in U_d, \; d = 1, \ldots, D$

Logistic mixed model:

Predicted values: $\hat{y}_k = \dfrac{\exp(\mathbf{x}_k' \hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(\mathbf{x}_k' \hat{\boldsymbol{\beta}} + \hat{u}_d)}, \quad k \in U_d, \; d = 1, \ldots, D$

# Properties

**Model-free calibration**: Multi-purpose weighting

- No explicit model statement (linear model assumed)

- Calibration of x-variable totals at the domain level

- Coherence property with x-variable totals is met

- MFC estimators of domain totals are of **direct type**

**Model calibration**: Single-purpose weighting

- Explicit model statement

- Calibration of y-prediction totals at the domain level

- Coherence property for x-variable totals is not met

- MC estimators of domain totals are of **semi-direct** type

    - modelling for the whole sample

    - calibration at the domain level

## TAXONOMY: Statistical calibration methods in survey sampling

| | Model-free (linear) calibration MFC | Model calibration MC | Hybrid calibration HC |
|---|---|---|---|
| **Weight calibration** | Calibration to reproduce known population totals of auxiliary variables | Calibration to the population total of predictions derived via specified model | Combination of MC and MFC, depending on modeling and coherence requirements |
| **Typical study variable** | Continuous | Continuous, binary, polytomous, count | |
| **Level of auxiliary data** | Aggregate level | Unit level | Unit level<br>Aggregate level |
| **Model specification** | Linear relationships<br>(No explicit model statement) | Many options<br>e.g. Generalized linear (mixed) models family | |
| **Main aims** | Coherence with published statistics<br>“Multi-purpose” weighting<br>Accuracy improvement | Accuracy improvement<br>Flexible modelling | Accuracy improvement<br>Flexible modelling<br>Coherence with published statistics |
| **Selected literature** | Deville & Särndal (1992)<br>Estevao & Särndal (1999)<br>Särndal (2007)<br>Lehtonen & Veijanen (2009) | Wu & Sitter (2001)<br>Wu (2003)<br>Montanari & Ranalli (2005)<br>Lehtonen & Veijanen (2012, 2016a,b) | Montanari & Ranalli (2009)<br>Lehtonen & Veijanen (2015) |

# Simulation experiment: Summary

Synthetic register population $U$ of one million elements and $D = 40$ domains

Auxiliary x-variables:

$x_1$, $x_2$ continuous variables

$x_c$ categorical variable with 5 classes (treated as continuous $x_3$ in models)

Domain size $N_d$ in domain $U_d$ determined by $exp(R)$, $R \sim Uniform(2,5)$

Response variable $y$ was created by a mixed model with fixed and random effects

Random intercept $u_d$ and random slopes $u_{d1}$, $u_{d2}$ and $u_{d3}$, all following $N(0,0.04)$ were associated with each domain $d$

After creating x-variable values, the values of response variable $y$ were created in each domain $d$ by linear mixed model:

$$y_k = 1 + (1.25 + u_{d1})x_{1k} + (0.75 + u_{d2})x_{2k} + (5 + u_{d3})x_3 + u_d + \varepsilon_k,\ k \in U_d,\ d = 1,\ldots,D$$

Errors $\varepsilon \sim N(0,5)$

Sampling: 1,000 independent SRSWOR samples of size $n = 4000$ elements

# Properties of population data

Properties of classes of $x_c$ in the population.

| Class | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Share of population (%) | 6.7 | 13.3 | 20.0 | 26.7 | 33.3 |
| Mean of $y$ | 17.3 | 23.2 | 28.8 | 34.8 | 40.5 |

Correlation coefficients of variables in the population.
The categorical variable $x_c$ is here treated as quantitative $(= x_3)$.

| | $x_2$ | $x_3$ | $y$ |
|---|---|---|---|
| $x_1$ | 0.34 | 0.00 | 0.49 |
| $x_2$ | 1 | 0.40 | 0.61 |
| $x_3$ | 0.40 | 1 | 0.69 |

# Assisting models in MC

**Linear mixed model** with domain-level random intercepts $u_d$

$y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k$ for $k \in U_d$, $d = 1,\ldots,40$

$\mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k})'$ continuous variables, known for all $k \in U$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ vector of fixed effects

$u_d \sim N(0, \sigma_u^2)$, $\varepsilon_k \sim N(0, \sigma^2)$, $u_d$ and $\varepsilon_k$ independent

Estimate $\boldsymbol{\beta}$ and $\sigma_u^2$ from the $n$ element sample $s$ (by ML or REML)

Calculate estimates $\hat{u}_d$, $d = 1,\ldots,40$

Calculate fitted values $\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d$, $k \in U_d$, $d = 1,\ldots,40$

**Special case :**

Model: $\mathbf{x}_k = (1, x_{1k}, x_{2k})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$

# Estimators

| Estimators for domain total parameters $t_d = \sum_{k \in U_d} y_k,\ d = 1,\ldots,40$ |
| --- |
| Design-based direct estimators |
| Direct HT estimator<br>$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$ where $a_k = 1/\pi_k$ |
| Direct model-free calibration estimator<br>$\hat{t}_{dMFC} = \sum_{k \in s_d} w_k^{MFC} y_k$ |
| Model assisted design-based model calibration estimator |
| Semi-direct model calibration estimator<br>$\hat{t}_{dMC} = \sum_{k \in s_d} w_k^{MC} y_k$ |

# Quality measures of estimators

- **Design bias**
  - Absolute relative bias ARB (%)
- **Accuracy**
  - Relative root mean squared error RRMSE (%)

$$ARB(\hat{t}_d) = \left| \frac{1}{1000} \sum_{k=1}^{1000} \hat{t}_d(s_k) - t_d \right| / t_d$$

$$RRMSE(\hat{t}_d) = \sqrt{\frac{1}{1000} \sum_{k=1}^{1000} (\hat{t}_d(s_k) - t_d)^2} / t_d$$

Averages calculated over domain sample size classes (minor/medium/major)

NOTE: Estimators considered are nearly design unbiased

# Comparison scenario

- Accuracy comparison of design-based direct estimators and semi-direct estimators
  - HT against calibration methods
  - Model-free calibration MFC against model calibration MC
- NOTE: Information supply
  - MFC and MC: Supply of similar auxiliary information but in different form!
  - HT: No auxiliary information

**Table 4.** Mean relative root mean squared error (RRMSE) (%) of design-based estimators of domain totals over domain sample size classes.

| Estimator | Assisting model & domain-level calibration scheme | Expected domain sample size | | |
|---|---|---|---|---|
| | | Minor 13-20 | Medium 20-50 | Major >50 |
| *Direct estimators* | | | | |
| HT | None | 24.00 | 13.23 | 7.59 |
| Model-free calibration | Calibration: $\mathbf{z}_k = (1, x_{1k}, x_{2k})'$ | 5.90 | 2.96 | 1.70 |
| *Semi-direct estimators* | | | | |
| Model: $y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + u_d + \varepsilon_k,\ k \in U_d,\ d = 1,...,40$ | | | | |
| Model calibration | Calibration: $\mathbf{z}_k = (1, \hat{y}_k)'$ | 5.66 | 2.94 | 1.70 |

# Conclusions for this example

- Calibration improves accuracy substantially over the HT

- Under same auxiliary information supply, semi-direct model calibration MC outperforms direct model-free calibration MFC in accuracy in minor domains

- General points:
- Incorporation of auxiliary information in the estimation procedure by using flexible modeling is helpful in improving precision of domain and small area estimates over the standard methods
- This is true for both the planned domains case and the unplanned domains case

# CASE STUDY: Estimation of mean of “Perceived income” for regional domains

- Source: Master Thesis in Statistics
- Nico Maunula (2012). Small Area Estimation Methods with Application to Perceived Income for Domains in Finland in 2009. Master’s Thesis, University of Helsinki. (In Finnish)

# Study problem

- Estimation of mean perceived income for regions in Finland
- Regions: $D$ = 70 NUTS4 areas
- Target population: $N$ about 4,3 million

- Sizes of regions vary:
  - Smallest: about 2000 persons
  - Largest: about 1 million persons

# EU-SILC data of Finland (2009)

- Sample size $n$ = 11,000 households
- Interview data (CAPI)
- Respondent: Household head
- Stratified unequal probability sampling
- Reweighting to adjust for unit nonresponse
- Model-free calibration for final weights

- Domains are of unplanned type

  Smallest domain sample size: 10

  Largest domain sample size: 2425

# Auxiliary data

- Auxiliary data are taken from statistical registers covering the target population
- Registers maintained by Statistics Finland

- Auxiliary data were merged with sample survey data at the unit level by using unique identification keys
  - Personal ID number

# Study variable

- HS120: **Ability to make ends meet**

- Represents ”experienced” (perceived) income (contrasted with “actual” income)
  - A subjective wellbeing indicator

- Ordinal level measurement with 6 levels
  - 1 = lowest, 6 = highest
  - Treated as continuous variable in modelling
  - Mean = 4.3 in SILC data

  - NOTE: Why “perceived income” This is because it is not available in administrative registers!

## HS120: Ability to make ends meet

*SOCIAL EXCLUSION (Non-monetary household deprivation indicators)*
*Cross-sectional and longitudinal*
*Reference period: current*
*Unit: household*
*Mode of collection: household respondent*

| Values | |
|---|---|
| 1 | with great difficulty |
| 2 | with difficulty |
| 3 | with some difficulty |
| 4 | fairly easily |
| 5 | easily |
| 6 | very easily |

| Flags | |
|---|---|
| 1 | filled |
| -1 | missing |

The household respondent’s assessment of the level of difficulty experienced by the household in making ends meet.
A household may have different source of income and more than one household member may contribute to it. Thinking of the household’s total monthly income, the idea is with which level of difficulty the household is able to pay its usual expenses.

# Auxiliary variables

- Variables (for HH head) from statistical registers
  - Gender
  - Age group (4 age groups)
  - Education (3 classes)
  - Actual (register) income
  - Socio-economic status (6 classes)
  - Stage in life of household-dwelling unit (5 classes)

- Categorical variables are transformed to indicator (dummy) variables
- 16 x-variables in the regression model
- All variables statistically significant
- R squared = 15%

# Models

Linear fixed-effects model

$$y_k = \beta_0 + \beta_1 x_k + \ldots + \beta_{16} x_{16k} + \varepsilon_k, \ , \ k \in U, \ \varepsilon_k \sim N(0, \sigma^2)$$

where beta coefficients are common for all domains

Linear mixed model

$$y_k = \beta_0 + u_d + \beta_1 x_k + \ldots + \beta_{16} x_{16k} + \varepsilon_k, \ k \in U_d, \ d = 1, \ldots, 70$$

with domain-level random intercepts $u_d$

$u_d \sim N(0, \sigma_u^2), \ \varepsilon_k \sim N(0, \sigma^2), \ u_d$ and $\varepsilon_k$ independent

# Estimators

Population mean for domain $d$: $\bar{y}_d = t_d / N_d, \; d = 1,\ldots,70$

HT estimator for domain means

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k, \; d = 1,\ldots,70$$

$$\hat{\bar{y}}_{dHT} = \hat{t}_{dHT} / N_d$$

wnere $N_d$ are known domain sizes in population

GREG estimators for domain means

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) \; , \; d = 1,\ldots,70$$

$$\hat{\bar{y}}_{dGREG} = \hat{t}_{dGREG} / N_d$$

where $w_k = a_k g_k$ are final calibrated weights (g-weights)

# GREG estimators

GREG assisted by linear fixed-effects model

Model fitted by ML

Predicted values

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \ldots + \hat{\beta}_{16} x_{16k},\ k \in U$$

MGREG assisted by linear mixed model

Model fitted by REML

Predicted values

$$\hat{y}_k = \hat{\beta}_0 + \hat{u}_d + \hat{\beta}_1 x_k + \ldots + \hat{\beta}_{16} x_{16k},\ k \in U_d,\ d = 1,\ldots,D$$

# Variance estimators
## (unplanned domains)

HT estimator for domain means

$$\hat{V}_U\left(\hat{\bar{y}}_{dHT}\right) = \hat{V}_U\left(\hat{t}_{dHT}\right) / N_d^2$$

$$= \frac{n}{N_d^2(n-1)} \sum_{k \in s} \left( w_k y_{dk} - \hat{t}_{dHT} / n \right)$$

where $y_{dk} = I\{k \in U_d\} y_k$ are extended y-variables

GREG estimators for domain means

$$\hat{V}_U\left(\hat{\bar{y}}_{dGREG}\right) = \hat{V}_U\left(\hat{t}_{dGREG}\right) / N_d^2$$

$$= \frac{n}{N_d^2(n-1)} \sum_{k \in s} \left( w_k e_{dk} - \hat{t}_{dHTe} / n \right)^2$$

where $e_{dk} = I\{k \in U_d\} e_k$ are extended residuals

# Quality indicators

Standard error of domain mean estimate $\hat{\bar{y}}_d$

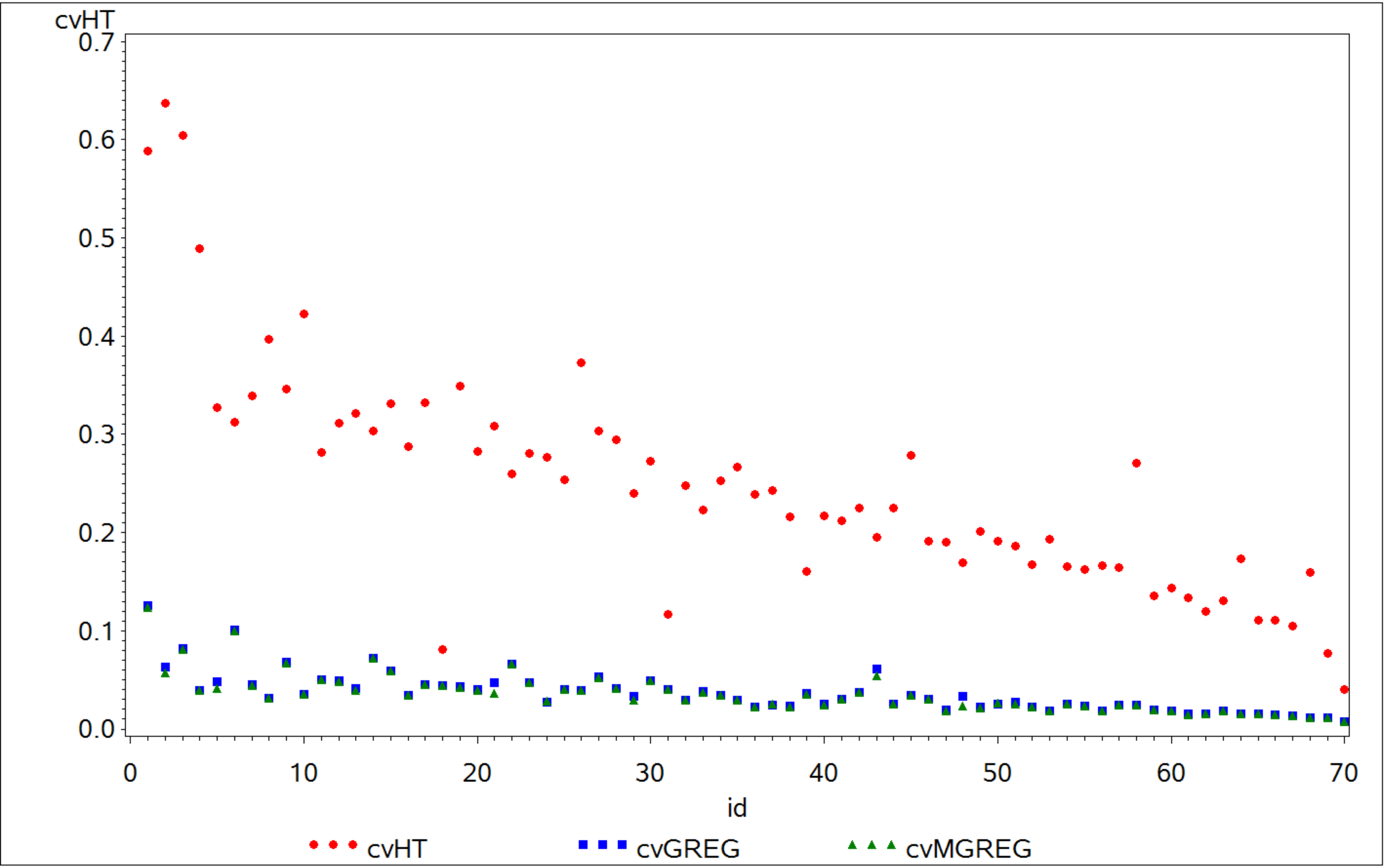
$$\text{s.e}\left(\hat{\bar{y}}_d\right) = \sqrt{\hat{V}\left(\hat{\bar{y}}_d\right)}, \quad d = 1,\ldots,70$$

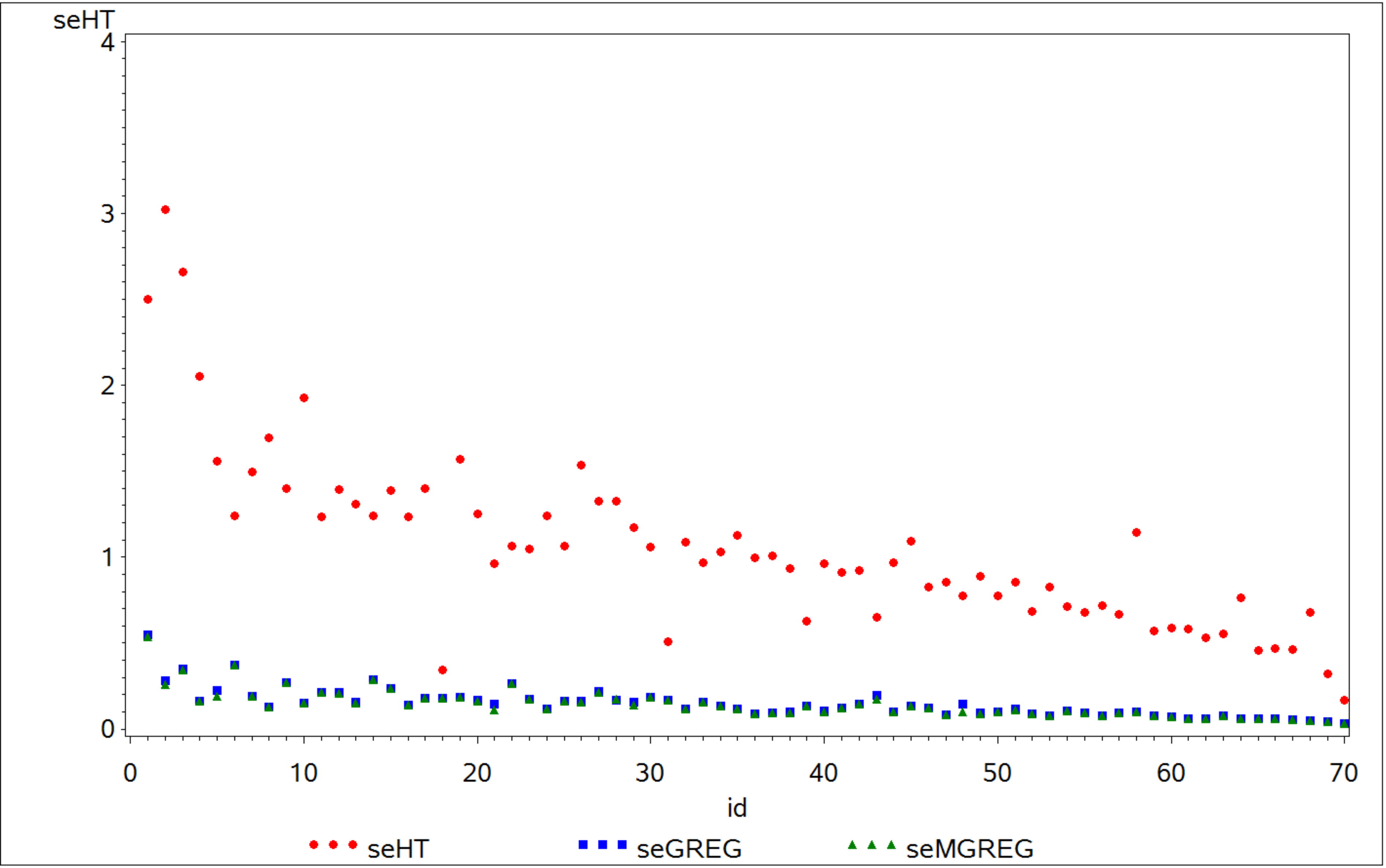
Coefficient of variation of domain mean estimate $\hat{\bar{y}}_d$

$$\text{cv}\left(\hat{\bar{y}}_d\right) = \frac{\text{s.e}\left(\hat{\bar{y}}_d\right)}{\hat{\bar{y}}_d} \quad d = 1,\ldots,70$$

**Table 5.** Average coefficient of variation of HT, GREG and MGREG estimates of domain totals by domain sample size class. Sample size n = 11,000, D=70 NUTS3 unplanned domains.

| | Domain sample size class | | | |
|---|---|---|---|---|
| | Minor | Medium-sized | Major | All |
| | Average domain sample size | | | |
| | 34 | 72 | 325 | 152 |
| **Direct estimator** | | | | |
| Design-based HT | 37.2 | 24.9 | 15.4 | 24.8 |
| **Indirect estimators** | | | | |
| *Model-assisted* | | | | |
| GREG | 5.7 | 3.7 | 1.9 | 3.6 |
| MGREG | 5.5 | 3.6 | 1.9 | 3.5 |


cvHT
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
id
cvHT
cvGREG
cvMGREG


seHT
4
3
2
1
0
0
10
20
30
40
50
60
70
id
seHT
seGREG
seMGREG

# Points for discussion

- Strategies in sampling design phase
- Strategies in estimation phase
- Share of labor between sampling design and estimation design
- NOTE: Key feature: Clever use of auxiliary data and modelling!

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

# Main literature


Deville J.-C. and Särndal C.-E. (1992) Calibration estimators in survey sampling. *JASA* 87, 376-382.

Estevao V. M. and Särndal C.-E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology* 2, 213-221.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology Journal, 24, 51–55.*

Lehtonen, R. and Veijanen, A. (2009) Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C. R. and Pfeffermann D. (Eds.) *Handbook of Statistics Vol. 29B. Sample Surveys. Inference and Analysis*. Amsterdam: Elsevier, 219–249.

Lehtonen R. and Veijanen A. (2012) Small area poverty estimation by model calibration*. Journal of the Indian Society of Agricultural Statistics* 66, 125-133 (Special issue on small area estimation).

Lehtonen R. and Veijanen A. (2015) Small area estimation by calibration methods. Invited paper, World Statistics Congress of the ISI, Rio de Janeiro, August 2015.

Lehtonen R. and Veijanen A. (2016a) Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) *Analysis of Poverty Data by Small Area Estimation*. Chichester: Wiley.

Lehtonen R. and Veijanen A. (2016b) Estimation of poverty rate and quintile share ratio for domains and small areas. In: Alleva G. and Giommi A. (Eds.) *Topics in Theoretical and Applied Statistics*. New York: Springer.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003) The effect of model choice in estimation for domains, including small domains. *Survey Methodology* 29, 33–44.

Lehtonen R., Särndal C.-E. and Veijanen A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition,* 7, 649–673.

Montanari G. E. and Ranalli M. G. (2005) Nonparametric model calibration estimation in survey sampling. *JASA* 100, 1429–1442.

Montanari G.E. and Ranalli M.G. (2009) Multiple and ridge model calibration. Proceedings of Workshop on Calibration and Estimation in Surveys 2009. Statistics Canada.

Robinson P.M. and Särndal C.-E. (1983) Asymptotic properties of the generalized regression estimator in probability sampling, *Sankhyā Ser. B,* 45, 240–248*.*

Rueda M., Sánchez-Borrego I., Arcos A. and Martínez S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika,* 71, 33–44.

Särndal, C.E. (1980) On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 67, 639–650.

Särndal C.-E. (2007) The calibration approach in survey theory and practice. *SMJ* 33, 99–119.

Särndal C.-E., Swensson B. and Wretman J. (1992) *Model-Assisted Survey Sampling*. New York: Springer

Wu C. and Sitter R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *JASA* 96, 185–193. (with corrigenda)

Wu C. (2003) Optimal calibration estimators in survey sampling. *Biometrika* 90, 937–951.

# ANNEX Notation

*Fixed and finite population* $U = \{1, 2, ..., k, ..., N\}$ and *sample* $s \subset U$

Variable of interest *y* with values $y_k$, $k \in U$ regarded as *fixed but unknown*

*Auxiliary variable* vector $\mathbf{x}_k$ known for all units $k \in U$

*Sample inclusion indicator* $Z_k$, $k \in U$, represents how many times element *k* is included in sample *s*

WOR sampling: $Z_k = 1$ if $k \in s$, 0 otherwise

*Inclusion probability* $\pi_k = P\{Z_k = 1\}$, $0 < \pi_k \leq 1$, $k \in U$

*Sample selection probability*

$$p(s) = P\{Z_k = 1, k \in s, Z_l = 0, l \notin s, k \neq l\}$$

$p(s)$ is called *sampling design*

# Design-based (randomization-based) inference

The source of randomness is the *sampling design* $p(s)$

Inference is based on assumed hypothetical repeated sampling under design $p(s)$ from the fixed population $U$

The random variables used for inference are the $Z_k,\ k \in U$

Example: *Horvitz-Thompson* (1952) estimator of population total $t = \sum_{k \in U} y_k$ :

$$\hat{t}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} Z_k \frac{y_k}{\pi_k}$$

# Model-based inference

**Model-based (prediction-based) inference**

The values $y_k$, $k \in U$, are assumed to be realizations of random vectors that follow a stochastic model

Let $Y_k$ represent the r.v. generating the value $y_k$ for unit *k*

Example: The ratio model $Y_k = \beta x_k + \varepsilon_k$, where $\varepsilon_k$ are i.i.d with mean 0 and variance $x_k \sigma^2$

*Prediction estimator* (Brewer 1963) of population total $t = \sum_{k \in U} y_k$ :

$$\hat{t}_{pred} = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{\beta} x_k,$$

where $\hat{\beta} = \sum_{k \in s} Y_k / \sum_{k \in s} x_k$ is the BLU estimator of $\beta$ under the model