THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Estimation Theory for Sample Surveys

An Introduction to Design-Based and Model-Assisted Analysis

Diego Zardetto
World Bank STC for research

Course Overview

- Inference Approaches in Official Statistics

- An Introduction to Design-Based and Model-Assisted Analysis
  - Basic Notation and Formulas
  - The Importance of Weighting
  - Horvitz-Thompson Estimators and their Variance Estimators
  - Calibration Estimators and their Variance Estimators
  - Handling Nonresponse by Calibration
  - Appendix - Selected Topics

Inference Approaches in Official Statistics

- Design-Based Inference (randomization approach)
  - The reference approach in National Statistical Institutes (NSI)
  - Includes Model-Assisted methods (e.g. Calibration Estimators) as special cases
  - Naturally belongs to the Frequentist inferential framework
- Model-Based Inference (prediction approach)
  - In NSIs typically used to complement analyses when the Design-Based approach would fail
    - e.g. to treat Nonresponse, frame imperfections, measurement errors, non-probability samples, …
  - Can adopt either Frequentist or Bayesian inferential frameworks

Design-Based Inference (1/2)

- Finite population values $y_1, ..., y_N$ and parameters θ are non-random quantities (i.e. fixed and error-free)
- Randomness arises only from probability sampling
  - Samples are drawn by means of rigorously random algorithms
  - Each unit in the population has a known, non-zero probability of being selected in the random sample
    - Data sampling is entirely controlled
- Ideally, the STATISTICIAN is the one and the only RANDOMIZER
  - Ideally means ignoring all non-sampling errors (e.g. list problems, total and item nonresponse, measurement errors…)

Design-Based Inference (2/2)

- Statistical properties of estimators $\hat{\theta}(s)$ (like bias and efficiency) depend on the probability distribution induced by the sampling design $p(\mathrm{s})$
- Estimators $\hat{\theta}(s)$ invariably involve survey weights tied to sample units
  - weights may either depend only on the sampling design or incorporate further auxiliary information on the target population
- Design-Based methods allow to:
  - Build unbiased estimators (or nearly so)
    - even if samples are not naively representative, because we can adjust for unequal inclusion probabilities!
  - Exploit probability theory to assess the quality of obtained estimates

Model-Assisted Inference

- Key distinction: interest variable $\mathbf{Y}$ and auxiliary variables $\mathbf{X}$
- Relations between $\mathbf{Y}$ and $\mathbf{X}$ are generated by Nature (i.e. by real-world, domain-specific phenomena which are unknown)
- Auxiliary information about the target population is available from sources external to the survey at hand
  - Can use this information to describe relations between $\mathbf{Y}$ and $\mathbf{X}$ through a statistical model
- Model-Assisted inference is a suite of methods to improve the quality of Design-Based inferences by hinging upon available auxiliary information in a systematic and rigorous way
  - build more efficient (but still nearly unbiased) estimators
  - reduce bias (from nonresponse, frame imperfections, …)
- Note: the model is assisting only (i.e. descriptive): no stochastic structure ever assumed!

Model-Based Inference (1/2)

- Finite population values $y_1, ..., y_N$ and parameters θ are realizations of random variables that follow some unknown stochastic model (the superpopulation model)
- NATURE is the one and the only RANDOMIZER
  - A statistical model is a (human interpretable) guess made by the statistician about the true, unknown data generating mechanism adopted by Nature
- Model-Based inference requires two steps:

  1) A model condensing assumptions on the probability distribution of $\mathbf{Y}$ and $\mathbf{X}$, as well as on their dependency structure, is fitted and tested against observed data

  2) The fitted model is used to predict unobserved values, i.e. values y_k for units k which do not belong to the sample:

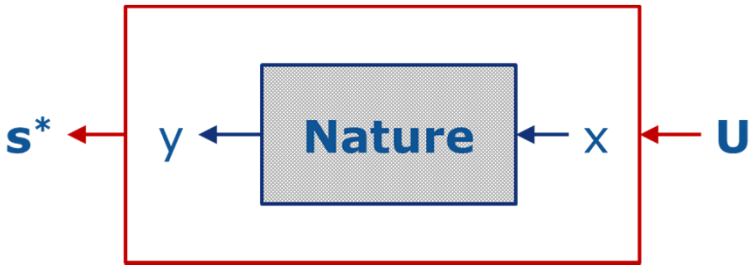
$$y_k \ \ unobserved \ \rightarrow \ \hat{y}_k = \hat{y}(\mathbf{x}_k) \ \ predicted$$

Model-Based Inference (2/2)

- The Model-Based approach can be applied to both probability and non-probability samples
  - For non-probability samples, model-based estimation is the only viable choice
- Model-Based inference treats the sample as fixed and ignores the sampling design
  - Strictly speaking, this is correct only for self-weighting designs
- Model-Based methods allow to build estimators that are unbiased under the adopted model
  - No definitive protection against bias exists
  - Bias can always be lurking, due to model misspecification

Inference in Official Statistics – Pictorial Synopsis

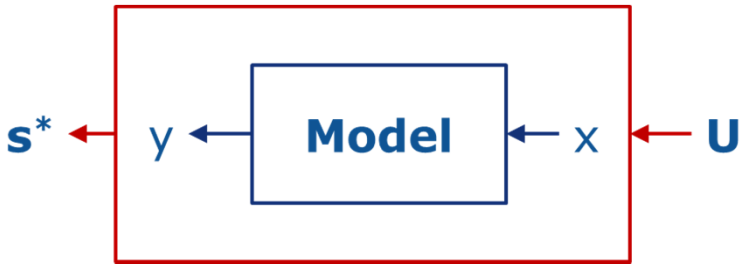
Design-Based



$$\hat{Y}_{HT} = \sum_{k \in s} d_k y_k = \sum_{k \in s} (1/\pi_k) y_k$$

Design Weights

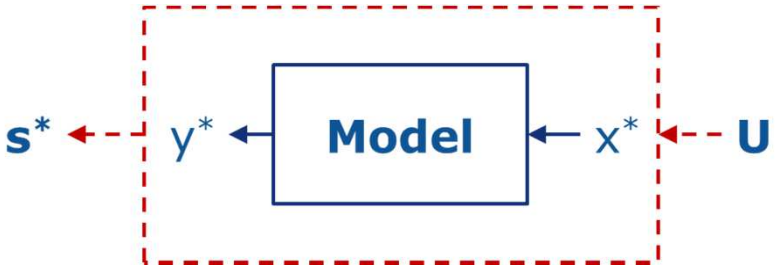
Model-Assisted



$$\hat{Y}_{CAL} = \sum_{k \in s} w_k y_k = \sum_{k \in s} (g_k d_k) y_k$$

Adjusted Weights

Model-Based



$$\hat{Y}_{\xi} = \sum_{k \in s} y_k + \sum_{k \in (U-s)} \hat{y}_k = \sum_{k \in s} y_k + \sum_{k \in (U-s)} \hat{y}(\mathbf{x}_k)$$

Predictions

An Introduction to Design-Based and Model-Assisted Analysis

Basic Notation and Formulas (1/5)

- Finite target population U of units:

$$U = \{1, 2, ..., N\}$$

- Probability sample s drawn from U:

$$s = \{1, 2, ..., n\} \qquad s \in S$$

- Sampling design:

$$p : S \to [0,1] \qquad p(s) = \Pr(s \text{ is selected})$$

- First order inclusion probabilities:

$$\pi_k = \Pr(k \subset s) = \sum_{s \supset k} p(s)$$

- Second order inclusion probabilities:

$$\pi_{kj} = \Pr(\{k, j\} \subset s) = \sum_{s \supset \{k,j\}} p(s) \qquad k = j \Rightarrow \pi_{kj} = \pi_{kk} = \pi_k$$

Basic Notation and Formulas (2/5)

- Estimator of a population parameter θ:

$$\hat{\theta} = \hat{\theta}(s)$$

- Design Expectation:

$$E_D(\hat{\theta}) = \sum_{s \in S} \hat{\theta}(s) p(s)$$

- Design Bias:

$$B_D(\hat{\theta}) = E_D(\hat{\theta}) - \theta$$

- Design Variance:

$$V_D(\hat{\theta}) = E_D([\hat{\theta} - E_D(\hat{\theta})]^2)$$

- Design Covariance:

$$Cov_D(\hat{\theta}_1, \hat{\theta}_2) = E_D[\hat{\theta}_1 \hat{\theta}_2 - E_D(\hat{\theta}_1) E_D(\hat{\theta}_2)]$$

Basic Notation and Formulas (3/5)

- Design Mean Squared Error:

$$MSE_D(\hat{\theta}) = E_D[(\hat{\theta} - \theta)^2] = B_D(\hat{\theta})^2 + V_D(\hat{\theta})$$

- Sample Membership Indicators (random variables):

$$\delta_k = \delta_k(s) = \begin{cases} 1 & if \quad k \in s \\ 0 & if \quad k \notin s \end{cases} \quad \forall k \in U$$

- Notable relations with inclusion probabilities:

$$E_D(\delta_k) = \sum_{s \in S} p(s)\delta_k(s) = \sum_{s \supset k} p(s) = \pi_k$$

$$E_D(\delta_k \delta_j) = \pi_{kj}$$

$$V_D(\delta_k) = \pi_k(1 - \pi_k)$$

$$Cov_D(\delta_k, \delta_j) = \pi_{kj} - \pi_k \pi_j = \Delta_{kj}$$

Basic Notation and Formulas (4/5)

- Sample Membership Indicators (matrix notation):

$$\boldsymbol{\delta}(s) = \begin{pmatrix} \delta_1(s) \\ \vdots \\ \delta_N(s) \end{pmatrix} \qquad \boldsymbol{\Delta} = V_D\left[\boldsymbol{\delta}(s)\right] = \begin{pmatrix} \Delta_{11} & \cdots & \Delta_{1N} \\ \vdots & \ddots & \vdots \\ \Delta_{N1} & \cdots & \Delta_{NN} \end{pmatrix}$$

- Sample size (random, in general):

$$n(s) = \sum_{k \in U} \delta_k(s)$$

- Average sample size:

$$E_D(n) = \sum_{k \in U} E_D(\delta_k) = \sum_{k \in U} \pi_k$$

Basic Notation and Formulas (5/5)

- Survey variables (non-random values):

$$\{y_1, ..., y_q\}(interest) \quad \{x_1, ..., x_p\}(auxiliary)$$

- Observed survey data:

$$(k, y_{k1}, ..., y_{kq}, x_{k1}, ..., x_{kp}), \qquad k \in s$$

Two Fundamental Definitions (1/2)

Sampling Design: $p : S \rightarrow [0,1]$ $p(s) = \Pr(s \text{ is selected})$

1) Probability Sampling Design:

$$\forall k \in U \qquad \pi_k = \Pr(k \subset s) = E_D(\delta_k) > 0$$

Each unit in the population must have a strictly positive inclusion probability

2) Measurable Sampling Design:

$$\forall k \in U,\ \forall j \in U \qquad \pi_{kj} = \Pr(\{k, j\} \subset s) = E_D(\delta_k \delta_j) > 0$$

Each pair of units in the population must have a strictly positive second order (i.e. joint) inclusion probability

Two Fundamental Definitions (2/2)

NON-probability sampling designs

- Building unbiased estimators is impossible in a design-based approach
  - Samples can never be “representative” of the whole population, as units with zero inclusion probability will be never selected (e.g. cut-off sampling in business surveys)

NON-measurable sampling designs

- Building unbiased *variance* estimators is impossible in a design-based approach
  - Even if unbiased estimators of population parameters exist (because of probability sampling) we will not be able to assess their precision and build valid confidence intervals (e.g. systematic sampling)

Goals of Design-Based Survey Sampling

- Given a sampling design, build "good estimators" of population parameters
- A "good estimator" should be:
  - Unbiased, or nearly so: substantial bias leads to poor estimates (on average) and prevents from building valid confidence intervals
  - Efficient: small coefficient of variation (for a nearly unbiased estimator) means that, for most samples, the estimator is likely to produce an estimate near the true value
- Small bias and small variance are often conflicting objectives in practice

The Importance of Weighting

- Suppose you want to estimate the population total of variable y:

$$Y = \sum_{k \in U} y_k$$

- If you try with the naive estimator:

$$\hat{Y}_{naive} = \sum_{k \in s} y_k$$

- …you soon realize it's biased:

$$E(\hat{Y}_{naive}) = E(\sum_{k \in s} y_k) = E(\sum_{k \in U} y_k \delta_k) = \sum_{k \in U} y_k \pi_k \neq Y$$

- Removing bias is straightforward: simply introduce weighted (aka Horvitz-Thompson, aka expansion) estimators:

$$\hat{Y}_{HT} = \sum_{k \in s} d_k y_k = \sum_{k \in s} y_k (1/\pi_k)$$

Direct Weights and HT Estimators

- Weights appearing in HT estimators are named "direct weights" or "design weights":

$$d_k = 1/\pi_k$$

- Intuitively one can think that each sampled unit "represents" a number of population units given by its direct weight
- HT estimators are unbiased by construction (neglecting non-response)

$$B(\hat{Y}_{HT}) = 0$$

- Moreover they are linear functions of sample membership indicators, hence computing (formally) variance is easy:

$$V(\hat{Y}_{HT}) = \sum_{k \in U} \sum_{j \in U} d_k y_k \Delta_{kj} d_j y_j$$

Estimating HT Estimators Variance (1/2)

- Formal expression for HT estimators variance relies on whole population, so variance true value is unknown
- Consequently to asses HT estimators efficiency we need to estimate it from the sample…
- …let's try again with a weighted estimator (as done before when estimating a total):

$$\hat{V}(\hat{Y}_{HT}) = \sum_{k \in s} \sum_{j \in s} d_k y_k \left(\frac{\Delta_{kj}}{\pi_{kj}} \right) d_j y_j = \sum_{k \in s} \sum_{j \in s} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}} \right) d_k y_k d_j y_j$$

- It is easy to prove that above estimator (due again to Horvitz & Thompson again) is unbiased for the variance of HT

Estimating HT Estimators Variance (2/2)

- A more concise formula for the unbiased variance estimator of the HT estimator of the total is as follows:

$$\hat{V}(\hat{Y}_{HT}) = \sum_{k \in s} \sum_{j \in s} \breve{\Delta}_{kj} \breve{y}_k \breve{y}_j$$

where:

$$\breve{y}_k = \frac{y_k}{\pi_k} = y_k \cdot d_k \qquad \text{and} \qquad \breve{\Delta}_{kj} = \frac{\Delta_{kj}}{\pi_{kj}} = \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right)$$

- For fixed-size sampling a possible, still unbiased, alternative is due to Yates & Grundy & Sen:

$$\hat{V}_{YGS}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{k \in s} \sum_{j \in s} \breve{\Delta}_{kj} (\breve{y}_k - \breve{y}_j)^2$$

A Very Important Example: SRSWOR (1/2)

- When dealing with simple random sampling without replacement we are able to compute all previous formulas explicitly:

$$p(s) = \binom{N}{n}^{-1} \qquad \pi_k = p(s)\binom{N-1}{n-1} = n/N = f \qquad \pi_{kj} = p(s)\binom{N-2}{n-2} = \frac{n(n-1)}{N(N-1)}$$

$$\hat{Y}_{HT} = \sum_{k \in s} d_k y_k = N \cdot \bar{y} \qquad \hat{V}(\hat{Y}_{HT}) = N^2 \frac{1-f}{n} S_y^2$$

- here f denotes the sampling fraction (aka finite population correction) and we used the sample mean and the sample variance of y:

$$\bar{y} = \sum_{k \in s} y_k / n \qquad S_y^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2$$

A Very Important Example: SRSWOR (2/2)

- Note that the HT variance estimator can also be expressed as:

$$\hat{V}(\hat{Y}_{HT}) = (1-f)\cdot\frac{n}{n-1}\sum_{k\in s}(\breve{y}_k - \bar{\breve{y}})^2 = n\cdot(1-f)\,S^2_{\breve{y}}$$

$$\breve{y}_k = y_k \cdot d_k$$

- Most statistical software use the above formula, which can be easily modified in order to take into account stratification and clustering in wor sampling designs
- Clusters: substitute weighted y values for units with weighted y totals into clusters
- Strata: sum over variances computed treating strata as independent samples
- The above expression is also the basic tool to build approximate variance formulas when dealing with multistage sampling

Why Real-Word Surveys Are So Difficult

- Complex selection schemes
  - e.g. [multiphase / multistage / stratified / cluster / pps / systematic / wor / mixed / …] sampling
  - computing first order inclusion probabilities becomes difficult (only approximations available for second order)
- Complex estimators
  - e.g. [non linear / non analytic / model dependent / …]
  - main statistical properties (MSE, bias, consistency...) hard or even impossible to investigate from a theoretical standpoint
  - variance estimation a big issue
- Non sampling errors
  - e.g. [nonresponse / missing values / inconsistencies / …]
  - removing or reducing nonresponse bias requires: 1) understanding non response mechanism and 2) accurate auxiliary information
  - assessing imputation variance is very hard (depends on imputation technique)

Computing HT Variance for Multistage Designs (1/2)

- The HT variance formula obtained for SRSWOR can be suitably modified to cope with multistage, stratified, cluster sampling designs
- Complexity arising from multistage sampling is handled by the Ultimate Cluster Approximation (Kalton 1979):
  - as long as the first stage sampling fraction is small, contribution to variance arising from stages other than the first can be neglected

$$\hat{V}_{multi}(\hat{Y}_{HT}) = (1 - f_{psu}) \cdot \nu_1(\breve{y}_{psu}) + f_{psu}(1 - f_{ssu})\nu_2(\breve{y}_{ssu}) + \ldots$$

- Therefore we need only to deal with PSUs. Let's write y_{spk} (d_{spk}) for y observed value (direct weight) for unit k in PSU p in stratum s and call n_s (f_s) the number (fraction) of sampled PSUs in stratum s
- Define y weighted total for PSUs in stratum s:

$$\breve{y}_{sp} = \sum_{k \in p} d_{spk} y_{spk}$$

- and their mean:

$$\bar{\breve{y}}_s = (1/n_s) \sum_{p=1}^{n_s} \breve{y}_{sp}$$

Computing HT Variance for Multistage Designs (2/2)

- Now take the old SRSWOR expression and: 1) substitute weighted y values for units with weighted y totals for PSUs, 2) sum resulting expressions over all strata:

$$\hat{V}_{multi}(\hat{Y}_{HT}) \approx \sum_s (1-f_s) \cdot \frac{n_s}{n_s - 1} \sum_{p=1}^{n_s} (\breve{y}_{sp} - \bar{\breve{y}}_s)^2 = \sum_s n_s \cdot (1-f_s)\, S^2_{\breve{y}_{sp}}$$

- Above formula is a good approximation for the HT variance estimator under multistage WithOut Replacement sampling with equal probabilities and small first stage sampling fractions
- Most statistical software use it with $f_s=0$ also for estimating HT variance for With Replacement sampling (thus implicitly using Hansen-Hurwitz rather than HT estimators). This works correctly both for equal and unequal (pps) inclusion probabilities
- Using the formula above with $f_s=0$ for pps WithOut Replacement sampling results in conservative variance estimates. Alternatively, one must find approximate formulas for second order inclusion probabilities

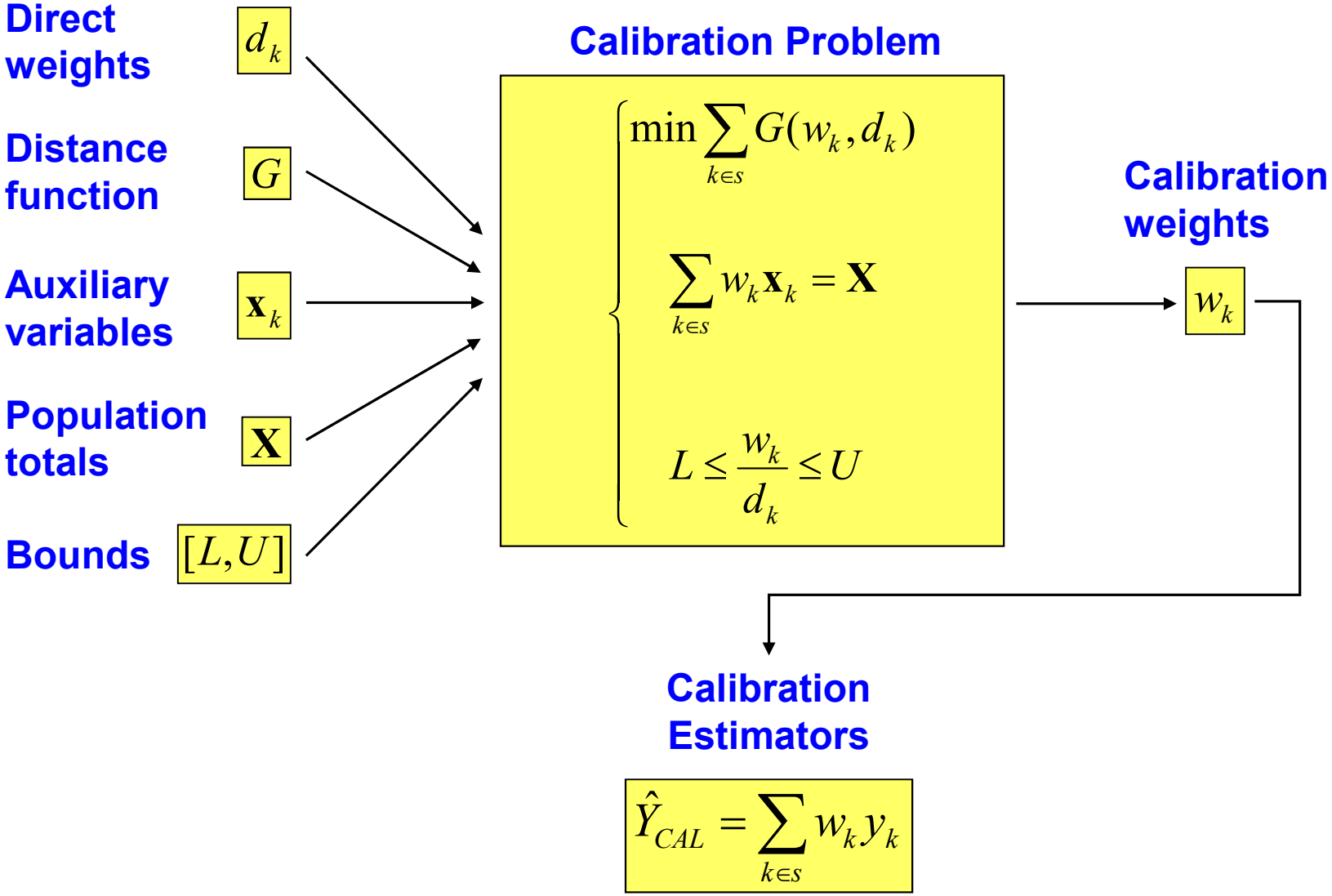
The Calibration Approach to Survey Sampling

What do we mean by calibration?

- General definition
  - a method to improve the quality of inferences by using available auxiliary information on the target population in a systematic and rigorous way
- Operative definition
  - a method to compute weights (calibration weights) in such a way that:

  1) a specified set of constraints (calibration equations) involving auxiliary variables is satisfied
  2) weights can be used to compute weighted (though non linear) estimators (calibration estimators) of arbitrary population parameters
  3) calibration estimators are nearly unbiased and more efficient than HT

The Calibration Problem



The Calibration Problem: Some Comments (1/2)

- From a mathematical point of view, calibration is a constrained optimization problem:
  - calibration weights are obtained by minimizing an appropriate distance function from direct weights…
  - …subject to calibration constraints ensuring that the calibrated estimates of the totals of a set of auxiliary variables exactly match the corresponding known population totals
- From a statistical perspective, calibration generates a whole new class of estimators: the Calibration Estimators
- An important property of Calibration Estimators is **UNIVERSALITY**: since the calibration problem knows nothing about y
  - one can use the same set of calibration weights w_k to estimate arbitrary interest variables y
- Universality is often a fundamental requirement in Official Statistics, as sample surveys are typically multipurpose:
  - ***Calibrate once, estimate whatever you need!***

The Calibration Problem: Some Comments (2/2)

- Because sample size is greater than the number of auxiliary variables, calibration constraints alone are unable to select a unique set of calibration weights
- Distance minimization picks out a specific solution, but still we are left with the freedom to choose among many distinct distance functions G
- In order to ensure a solution, G must be twice differentiable w.r.t. w and strictly convex in a neighborhood of $w=d$, with $G(d,d)=0$
- Calibration weights will be a (complicated) function of direct weights, auxiliary variables and population totals (as well as of bounds, if any)
- Only for Euclidean G will that function be expressible in analytic closed-form
- When G is the Euclidean distance, calibration estimators are identical to GREG estimators
- There exists a family of distances G such that calibration estimators are asymptotically equivalent to GREG for “big” sample sizes n

Calibration: Why Minimizing a Distance?

- The idea of imposing calibration constraints seems rather natural: it's good to make our estimates consistent with known aggregates…
- …but why do we need to minimize a distance between direct and calibration weights?
- The reason is that we do want to modify as little as possible a good property of HT estimators: unbiasedness!
- Asking for near unbiasedness of calibration estimators (whatever the choice of variable y)…

$$B(\hat{Y}_{CAL}) = E(\hat{Y}_{CAL} - Y) = E(\hat{Y}_{CAL} - \hat{Y}_{HT}) = E\left(\sum_{k \in s} y_k (w_k - d_k)\right) \approx 0$$

- …evidently translates into requiring small deviations from direct weights:

$$w_k - d_k \approx 0 \qquad \forall k \in s$$

How Does Calibration Improve Efficiency? (1/4)

- Generalized Regression Estimator basic theory provides a useful starting point
- Given the descriptive linear assisting model ξ with p regressors $x_1, \ldots, x_p$:

$$\xi: \quad y_k \sim \mathbf{x}_k \cdot \boldsymbol{\beta} + \varepsilon_k$$

$$E_\xi(\varepsilon_k) = 0, \quad V_\xi(\varepsilon_k) = \sigma^2 < \infty, \quad Cov_\xi(\varepsilon_k, \varepsilon_j) = 0 \quad \forall k, j \in U$$

- the GREG estimator for the total of y reads:

$$\hat{Y}_{GREG} = (\sum_{k \in U} \mathbf{x}_k) \cdot \hat{\boldsymbol{\beta}} + \sum_{k \in s} d_k (y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}) = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \hat{\boldsymbol{\beta}}$$

- where the weighted estimator for β has the familiar Least Squares expression arising in ordinary regression theory:

$$\hat{\boldsymbol{\beta}} = (X^t D X)^{-1} (X^t D Y) = \left(\sum_{k \in s} d_k \mathbf{x}_k^t \cdot \mathbf{x}_k \right)^{-1} \cdot \left(\sum_{k \in s} d_k \mathbf{x}_k^t y_k \right) = \mathbf{T}^{-1} \cdot \mathbf{t}$$

- with D being the diagonal matrix of sample direct weights and X (Y) the matrix (vector) of sample values for the auxiliary variables (study variable)

How Does Calibration Improve Efficiency? (2/4)

- GREG estimator can be seen as the sum of an HT estimator plus a regression adjustment term, which is proportional to the difference between population totals and HT estimates of predictors (auxiliary variables)
- Moreover, since the weighted estimator of β is linear in y, it follows that the GREG can be expressed as a weighted sum of y_k:

$$\hat{Y}_{GREG} = \sum_{k \in s} w_k y_k = \sum_{k \in s} g_k (d_k y_k)$$

- for some weights w_k (or ratios $g_k = w_k / d_k$) that depend (with a complicated expression) on direct weights and x values, but not on y:

$$g_k = w_k / d_k = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \mathbf{T}^{-1} \cdot \mathbf{x}_k^t$$

- Not only is GREG a weighted estimator (though non-linear, due to the expression of the sample estimate of β), what's more the weights it involves happen to be calibrated!

$$\hat{\mathbf{X}}_{GREG} = \sum_{k \in s} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k = \mathbf{X}$$

How Does Calibration Improve Efficiency? (3/4)

- Provided both target population and sample sizes are big, GREG main statistical properties are under theoretical control
- Whatever the assisting model ξ, GREG is asymptotically unbiased with relative bias of order:

$$B(\hat{Y}_{GREG})/Y = O(n^{-1})$$

- Moreover if we call R^2 the determination coefficient of the census least squares fit to the assisting model ξ, we can express GREG's variance as follows:

$$V(\hat{Y}_{GREG}) = \left[1 - R^2 + O(n^{-1/2})\right] \cdot V(\hat{Y}_{HT})$$

- Hence the efficiency gain of a GREG estimator, compared to the simple HT estimator, depends on the goodness of the assisting model ξ in fitting the population scatter of variables $y, x_1, \ldots, x_p$
- The higher the power of the auxiliary variables $x_1, \ldots, x_p$ to predict the study variable y, the smaller will be GREG's variance

How Does Calibration Improve Efficiency? (4/4)

- Now we are in a good position to understand how does calibration improve efficiency
- The key point is that one can prove, under mild conditions on distance functions G and whatever the choice of variable y, that:

$$\hat{Y}_{CAL} - \hat{Y}_{GREG} = O_p(n^{-1})$$

- namely, all Calibration estimators converge in probability to the GREG estimator when the sample size grows
- Therefore, as a consequence of GREG properties, we can expect that a calibration procedure will yield efficient estimates provided that the auxiliary variables explain/predict well the study variable(s)
- Notice also that, being GREG asymptotically unbiased, the same will hold true for every calibration estimator, no matter how badly one can choose the auxiliary variables:

$$B(\hat{Y}_{CAL}) / Y = O(n^{-1})$$

GREG: Useful Alternative Expressions

- HT + Regression Adjustment

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \hat{\boldsymbol{\beta}}$$

- Population Projection of Predictions + Residuals

$$\hat{Y}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} d_k e_k$$

$$\hat{y}_k = \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}$$

$$e_k = y_k - \hat{y}_k$$

- Calibration Weighted Total

$$\hat{Y}_{GREG} = \sum_{k \in s} w_k y_k$$

$$w_k = g_k d_k$$

Estimating Calibration Estimators Variance (1/4)

- Till now we learned that both Calibration and GREG estimators are nearly unbiased and more efficient then HT
- Anyway we still must face the problem of estimating their variance (otherwise we cannot build meaningful confidence intervals around estimates)
- Both estimators are non linear functions of sample membership indicators, so an exact variance formula cannot be obtained
- Taylor linearization technique is widely used to provide approximate variance formulas for nonlinear estimators:
  1) A complex function of HT estimators get expanded in power series around expected (=true) values till first order
  2) Higher order contribution to variance are discarded, even without any warranty that their importance is actually negligible
  3) Usual variance formulas for HT estimators are applied to the linearized estimator

Estimating Calibration Estimators Variance (2/4)

- By adding and subtracting β wherever you get it's estimator in GREG formula, GREG can be re-expressed as:

$$\hat{Y}_{GREG} = (\sum_{k \in U} \mathbf{x}_k) \cdot \boldsymbol{\beta} + \sum_{k \in s} d_k (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta}) + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

- First term is a constant not contributing to sampling variance, third term can be proved to be of smaller order compared to the second, which equals the HT estimator of the population sum of residuals to the census least-squares fit of model ξ. Thus:

$$\hat{Y}_{GREG,lin} = \sum_{k \in s} d_k (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta}) = \sum_{k \in s} d_k E_k$$

- Because we don't know β we must substitute population residuals E_k with their sample estimates e_k. At the end we obtain:

$$\hat{V}(\hat{Y}_{GREG}) \approx \hat{V}(\hat{Y}_{GREG,lin}) = \sum_{k \in s} \sum_{j \in s} \left(\frac{\Delta_{kj}}{\pi_{kj}} \right) d_k e_k d_j e_j$$

$$e_k = y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}$$

Estimating Calibration Estimators Variance (3/4)

- Although our last formula gives a good variance estimator, a better one has been proposed by Sarndal (1982):

$$\hat{V}(\hat{Y}_{GREG}) \approx \hat{V}(\hat{Y}_{GREG,lin}) = \sum_{k \in s} \sum_{j \in s} \left(\frac{\Delta_{kj}}{\pi_{kj}} \right) w_k e_k w_j e_j$$

$$w_k(\mathbf{d}) = g_k(\mathbf{d}) d_k$$

- Notice that it relies on GREG "calibrated" weights, which (as we already stressed) are complex functions of direct weights $\mathbf{d}=(d_1,\ldots,d_n)$
- Notice also that above formula equals the ordinary HT variance estimator for the population total of the g-expanded residuals:

$$\hat{V}(\hat{Y}_{GREG}) \approx \hat{V}(\hat{Y}_{GREG,lin}) = \hat{V}\left(\sum_{k \in s} d_k (g_k e_k) \right)$$

- Last remark explains why this formula is so widely used in survey analysis software: it allows you to use your ordinary HT variance estimation program on the new linearized variable $g_k e_k = (w_k/d_k) e_k$

Estimating Calibration Estimators Variance (4/4)

- Once we have our preferred variance estimator for GREG, we can directly use it also for Calibration estimators (again by means of asymptotic equivalence)
- The final cookbook recipe for estimating Calibration estimators variance is as follows:

1) Solve by a computer program your calibration problem; take at hand the calibration weights w_k
2) Compute weighted regression coefficients by projecting your interest variable y on the auxiliary variables $x_1, \ldots, x_q$ and using direct weights d_k
3) Use regression coefficients from step 2) to compute estimated residuals e_k; use calibration weights from step 1) to compute g-expanded residuals $g_k\, e_k = (\, w_k / d_k \,)\, e_k$
4) Treat the g-expanded residuals from step 3) as an ordinary variable of interest; compute its total's variance letting do the work to your ordinary HT variance estimation program

Calibration Metrics

- Back to the Calibration Problem: the most popular distance functions are Euclidean (aka Linear), Logarithmic (aka Raking) and Logit. Each one leads to calibration weights with their own distinctive features
- Euclidean distance

$$G = \sum_{k \in s} (w_k - d_k)^2 \Big/ (2 d_k)$$

- If bounds constraints are not imposed, the Euclidean (or Unbounded Linear) distance leads to Calibration Estimators identical to GREG
- Calibration weights can sometimes be less than 1 or even negative; for estimation purposes this is not a problem, though it may seem strange to naive users
- Anyway it is possible to search for g-weights falling in a given interval $0 \leq L < U < \infty$ by simply truncating the Euclidean distance (obtaining thus the Bounded Linear distance)
- Remember that imposing additional bounds constraints can sometimes prevent from finding a numerical solution to the Calibration Problem

Unbounded Linear Calibration (1/2)

- Euclidean unbounded distance:

$$G = \sum_{k \in s} (w_k - d_k)^2 \Big/ (2 d_k)$$

- allows to get an analytic closed-form solution to the calibration problem:

$$\begin{cases} \min \sum_{k \in s} G(w_k, d_k) \\ \sum_{k \in s} w_k x_{kj} = X_j \end{cases}$$

- Introducing a set of Lagrange multipliers (1 for each auxiliary variable) one is led to the unconstrained minimization of a new function Q:

$$Q = G + \sum_j \lambda_j (X_j - \sum_{k \in s} w_k x_{kj})$$

- Setting to zero first derivatives of Q w.r.t. w_k and λ_j one obtains:

$$\begin{cases} 0 = \partial Q / \partial w_k = \partial G / \partial w_k - \sum_j \lambda_j x_{kj} = g_k - 1 - \boldsymbol{\lambda}^t \mathbf{x}_k^t \\ 0 = \partial Q / \partial \lambda_j = X_j - \sum_{k \in s} w_k x_{kj} \end{cases}$$

Unbounded Linear Calibration (2/2)

- Substituting first equation into second (via $w_k = g_k d_k$) one gets for $\boldsymbol{\lambda}$:

$$\sum_i \left(\sum_{k \in s} d_k x_{jk}^t x_{ki} \right) \lambda_i = X_j - \hat{X}_j^{HT}$$

- Above equation can be expressed in matrix notation as follows:

$$\mathbf{T} \cdot \boldsymbol{\lambda} = (\mathbf{X} - \hat{\mathbf{X}}_{HT})^t$$

- where:

$$\mathbf{T} = \sum_{k \in s} d_k \mathbf{x}_k^t \cdot \mathbf{x}_k$$

- So that the following expression for $\boldsymbol{\lambda}$ stems:

$$\boldsymbol{\lambda} = \mathbf{T}^{-1} \cdot (\mathbf{X} - \hat{\mathbf{X}}_{HT})^t$$

- And finally the g-weights formula reads:

$$g_k = 1 + \boldsymbol{\lambda}^t \cdot \mathbf{x}_k^t = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \mathbf{T}^{-1} \cdot \mathbf{x}_k^t$$

- It is easy to recognize this is equal to the GREG g-weights expression

Raking and Logit Calibration (1/2)

- Raking distance

$$G = \sum_{k \in s} \{ w_k \ln(w_k / d_k) - w_k + d_k \}$$

- Logit distance

$$G = \sum_{k \in s} \left\{ (g_k - L) \ln\left(\frac{g_k - L}{1 - L} \right) - (g_k - U) \ln\left(\frac{g_k - U}{1 - U} \right) \right\}$$

- For both above distances one cannot end up with an analytic closed-form expression for the calibration weights. Anyway it is easy to prove that the following implicit formula holds for the g-weights:

$$g_k = F(\mathbf{x}_k \cdot \boldsymbol{\lambda}) \qquad where \qquad F = \left(\partial G / \partial w \right)^{-1}$$

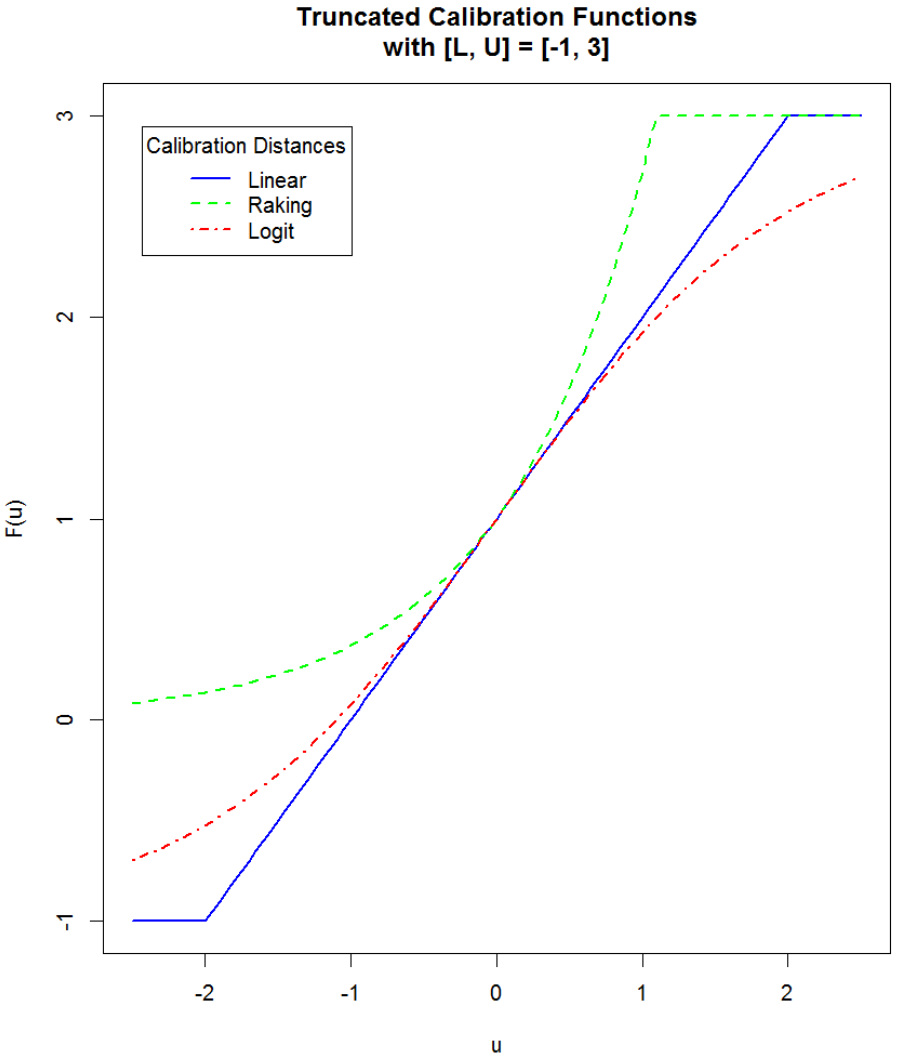
  with $\boldsymbol{\lambda}$ implicitly defined by the calibration constraints

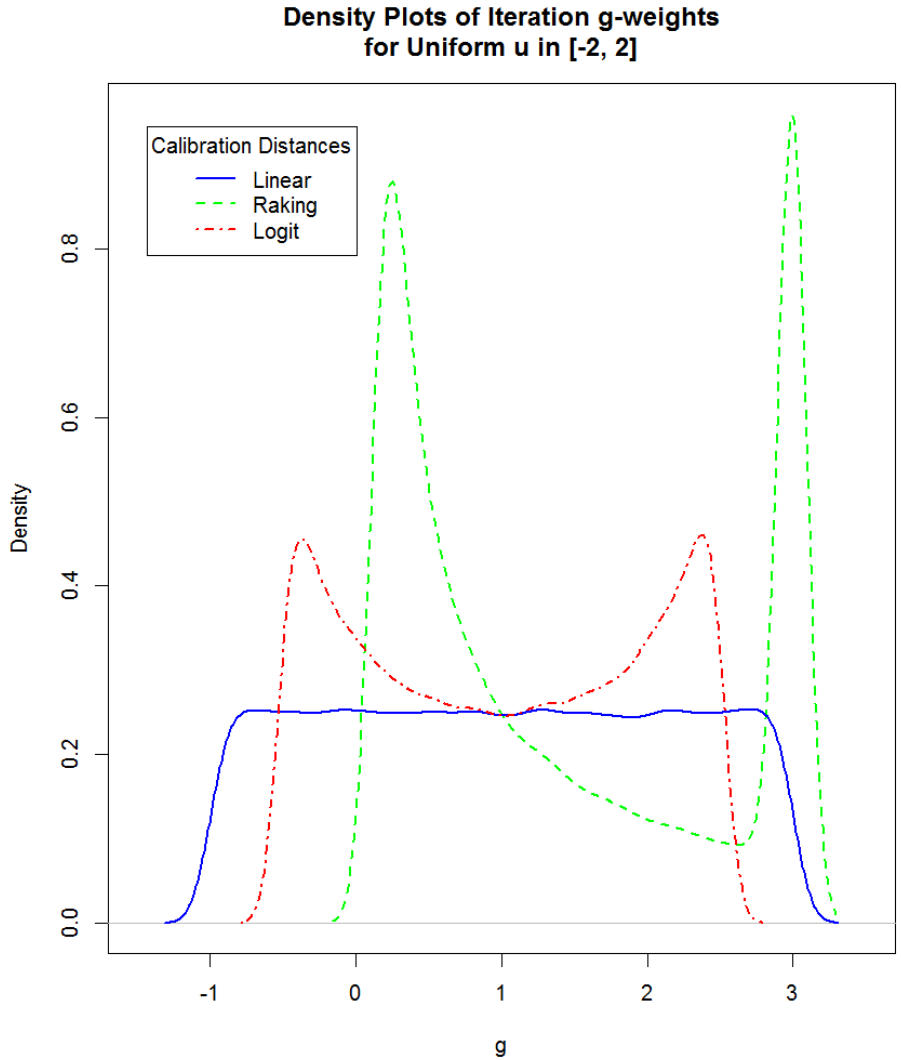
- Given its central role, the function F, namely the inverse of G's first order derivative, is called the Calibration Function

Raking and Logit Calibration (2/2)

- Notice that the Calibration Function associated to the Euclidean distance is $F(u)=1+u$, thus making clear the origin of the 'Linear' alias
- For the Raking distance you get $F(u)=e^u$, hence calibration weights are surely positive. Because same weight may be found to be unexpectedly larger than the others, sometimes the Bounded Raking distance is preferred
- The 'Raking' alias comes from the fact that, when known population totals are the marginal distributions of two categorical variables, you end-up with the Raking Estimator
- The Logit distance automatically builds g-weights constrained to fall into the interval $-\infty<L<U<\infty$
- For that reason Logit Calibration is a very popular choice in National Statistical Agencies (mainly in Europe): it helps to control the size of calibration weights, hence allowing those weights to be used in a wide variety of statistical analyses

Calibration Metrics and g-weights Features





Calibration Estimators: Familiar Examples (1/4)

- Suppose the Euclidean unbounded metric is selected, so g-weights are:

$$g_k = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \mathbf{T}^{-1} \cdot \mathbf{x}_k^t$$

- we want to derive explicitly the expressions of calibration estimators for some simple auxiliary information

- Total number of units in the population

- This is the simplest case of auxiliary information, the only auxiliary variable being $x_k=1$ for all k, thus:

$$T = \hat{N} \quad \Rightarrow \quad g_k = N / \hat{N}$$

- and the calibration estimator looks as Hajek ratio estimator of the total:

$$\hat{Y}_{CAL} = N\left(\frac{\hat{Y}}{\hat{N}}\right)$$

- Notice that, for SRSWOR, calibration and HT estimators are the same because direct weights perfectly estimate N

Calibration Estimators: Familiar Examples (2/4)

- Population counts in a one-way classification
- Population U is partitioned in P non overlapping and exhaustive subpopulations U_p with $p=1,...,P$. The total number of units N_p belonging to every subpopulation p is known
- The auxiliary vector can be represented as follows:

$$\mathbf{x}_k = (\gamma_{k1},...,\gamma_{kP}) \quad where \quad \gamma_{kp} = \begin{cases} 1 & if \quad k \in U_p \\ 0 & if \quad k \notin U_p \end{cases}$$

- $\mathbf{T}$ is a diagonal PxP matrix whose p-th diagonal element equals the HT estimator of the total of units in subpopulation U_p:

$$\mathbf{T} = \begin{pmatrix} \hat{N}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{N}_P \end{pmatrix}$$

- The only contribution to the matrix product $\mathbf{T}^{-1}\,\mathbf{x}^{t}_{k}$ arises from the only nonzero component of $\mathbf{x}_k$, namely the one referring to the subpopulation $p(k)$ to which unit k belongs

Calibration Estimators: Familiar Examples (3/4)

- Thus one obtains:

$$g_k = 1 + (N_{p(k)} - \hat{N}_{p(k)}) / \hat{N}_{p(k)} = N_{p(k)} / \hat{N}_{p(k)}$$

- that is calibration amounts to rescaling direct weights by factors depending only on the subpopulation. The resulting estimator has the same expression as the Post-stratification Estimator:

$$\hat{Y}_{CAL} = \sum_{p=1}^{P} N_p \left(\frac{\hat{Y}_p}{\hat{N}_p} \right)$$

- that is a sum of partition means estimators, each weighted by the known partition total
- For SRSWOR, denoting by n_p the sample units belonging to the p-th partition subset s_p, we get:

$$\hat{Y}_{CAL} = \sum_{p=1}^{P} N_p \left(\sum_{k \in s_p} y_k / n_p \right) = \sum_{p=1}^{P} N_p \bar{y}_{s_p}$$

Calibration Estimators: Familiar Examples (4/4)

- Total of a single numeric variable
- Population total X of a single numeric variable x (e.g. income) is supposed to be known. This time $\mathbf{T}$ equals the HT estimate for the population total of x^2, thus:

$$g_k = 1 + x_k (X - \hat{X}) / \hat{X}^2$$

- so that the calibration estimator is:

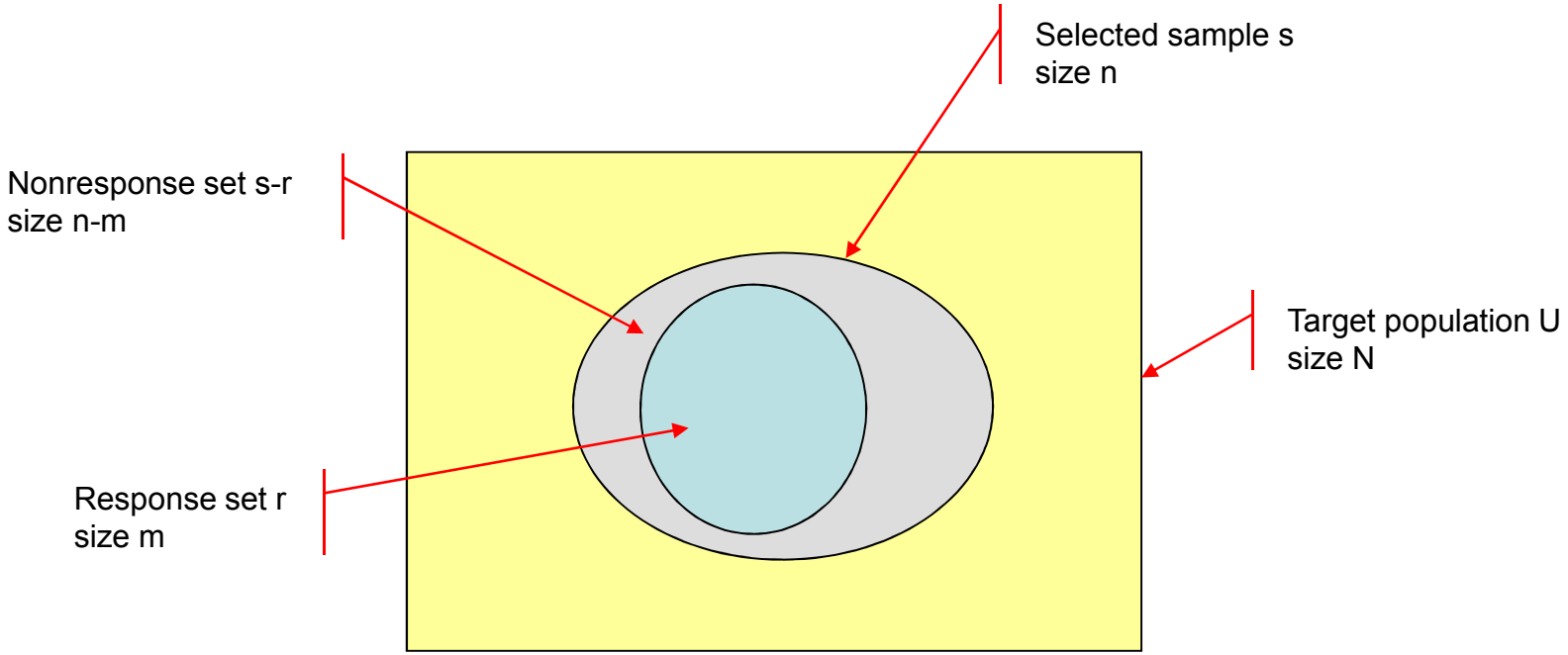
$$\hat{Y}_{CAL} = \hat{Y} + (X - \hat{X}) \frac{\hat{XY}}{\hat{X}^2}$$

- The latter expression is easily recognized as the GREG for a model with only one predictor and no intercept:

$$\hat{\beta} = \frac{\hat{XY}}{\hat{X}^2}$$

Effects of Nonresponse on Estimates (1/5)

- Till now we supposed that survey variables were observed for all the n units belonging to the selected random sample s
- In real surveys this is almost never the case: due to the nonresponse phenomenon we can actually collect information only on $m < n$ units belonging to a subset r of the planned sample s



Effects of Nonresponse on Estimates (2/5)

- The response set r (as well as the nonresponse set s-r) is a random set. It can be thought as the outcome of two subsequent random experiments:
  1) Sampling phase: select a random sample s from U under the given design
  2) Response phase: for each unit k in s include it into r with a specified response probability φ_k
- Response probabilities are unknown, if they were not we could easily modify our previous theoretical results to deal with nonresponse
- The simple rule would be: substitute everywhere sample inclusion probabilities π_k with response set inclusion probabilities p_k:

$$p_k = \Pr(k \subset r) = \Pr(k \subset s)\Pr(k \subset r|s) = \pi_k \varphi_k$$

- This rule would lead e.g. to the two-phase extension of HT estimators:

$$\hat{Y}_{2P} = \sum_{k \in r} y_k (1/p_k) = \sum_{k \in r} y_k (d_k/\varphi_k)$$

Effects of Nonresponse on Estimates (3/5)

- What would happen if, being response probabilities unknown, we simply keep using old theory in analyzing a real survey?
- This would lead to many undesired effects, the worst being nonresponse bias
- Understanding how nonresponse bias can arise is simple:
  - nonresponse probability is not uniform over population units: it is higher for units belonging to specific subpopulations (e.g. in social surveys, for older persons, metropolitan residents, …)
  - subpopulations which have, on average, higher nonresponse rates will be under-represented in the response set compared to what would happen in the sample
  - thus standard HT estimates for the size of those subpopulation will be downward biased
  - HT estimates for variables that covary with the ones defining high nonresponse subpopulations will be biased too (under-estimated or over-estimated depending on whether they are positively or negatively correlated)

Effects of Nonresponse on Estimates (4/5)

- A useful (though somewhat unrealistic) example of nonresponse bias:
  - Suppose you want to estimate the total income I for population U
  - Suppose you are sampling by SRSWOR from a population containing nearly the same number of males and females, $N_m \sim N_f$
  - Suppose nonresponse probability for males is significantly higher than for females (say twice, $\varphi_m = \varphi_f / 2$)
  - Suppose also that incomes are higher, on average, for males than for females (say twice, $I_m / N_m \sim 2\, I_f / N_f$)
  - How much bias would you expect if you decide to use a standard HT estimator?

$$E(\hat{I}_{HT}) = E(\sum_{k \in r} i_k / \pi_k) = E(\sum_{k \in U} \delta_k(r) i_k / \pi_k) = \sum_{k \in U} i_k / \pi_k E(\delta_k(r)) =$$

$$= \sum_{k \in U} i_k \varphi_k = \sum_{k \in F} i_k \varphi_f + \sum_{k \in M} i_k (\varphi_f / 2) = \varphi_f (\frac{1}{2} I_M + I_F) \cong 2\varphi_f I_F \cong$$

$$\cong \frac{2}{3} \varphi_f I$$

Effects of Nonresponse on Estimates (5/5)

- Another effect of nonresponse is a loss of efficiency in estimates: variance is increased due to the reduction of the effective sample size
- The latter may be considered a minor disturbance compared to nonresponse bias
  - we could fight nonresponse variance simply by some degree of 'over-sampling', at the price of some additional data collection cost
- Thus we are left with the following conclusion:
  - nonresponse affects all real-word surveys
  - nonresponse bias is its most dangerous drawback
  - trying to reduce nonresponse bias is mandatory
  - to do this we must modify our previous (ideal) design-based survey theory
  - we have to understand how to modify it

Reducing Nonresponse Bias: How? (1/2)

- As we said, we must change our previous theory in order to reduce nonresponse bias
- This can be done in 2 ways, both relying on auxiliary information:
  - The two-phase approach: exploit auxiliary information to estimate (i.e. model) unknown response probabilities

    $\varphi_k \ \ unknown \ \ \mapsto \ \ \hat{\varphi}_k = \hat{\varphi}_k(x_1,...,x_q) \ \ estimated$

    then use two-phase extended HT estimators

    $\hat{Y}_{2P} = \sum_{k \in r} y_k (d_k / \hat{\varphi}_k)$

  - The calibration approach: exploit auxiliary information and nonresponse patterns to identify 'good calibration variables'

    $x_1,...,x_q \ \ \mapsto \ \ \mathbf{x}_{good}$

    then use calibration estimators to reduce both estimators variance and nonresponse bias

    $\hat{Y}_{CAL} = \sum_{k \in r} w_k y_k$

Reducing Nonresponse Bias: How? (2/2)

- Comparing 2P and CAL estimators one can find a link between the two techniques:
  - calibration g-weights can be thought as proxy-values for inverse response probabilities: $g_k = w_k/d_k \rightarrow 1/\varphi_k$
- The two-phase approach has been very popular in the last 30 years but today the calibration approach is generally preferred
- The most important advantages of the CAL approach over the 2P one are the following:

2P

  - it is necessary to explicitly build a nonresponse model and then to use it to estimate nonresponse probabilities
  - a subsequent calibration step is required to increase estimators efficiency

CAL

  - a good description of the nonresponse mechanism in terms of correlated variables is enough
  - a single calibration step can reduce both nonresponse bias and estimators variance

The Response Homogeneity Group Model (1/3)

- The Response Homogeneity Group model (RHG) is perhaps the most popular implementation of the 2P approach to nonresponse bias reduction
- RHG key assumption is that the population consists of non-overlapping subpopulations (the Groups) such that:
  - all units within each group respond with the same probability
  - different groups may have different response probabilities
  - response/nonresponse outcomes are independent for all the units
- In formulas:

$$\begin{cases} U = G_1 \bigcup G_2 \bigcup ... \bigcup G_g \\ G_i \bigcap G_j = \varnothing \qquad for \quad i \neq j \\ \varphi_k = const = \Phi^j \qquad \forall k \in G_j \\ \Phi^i \neq \Phi^j \qquad for \quad i \neq j \end{cases}$$

The Response Homogeneity Group Model (2/3)

- The modeling effort for the RHG model lies in the problem of forming groups G_j that give nearly constant response probability Φ^j within each group
- Those groups must be built with the aid of some auxiliary variables, whose values we have to know, at least, for every unit belonging to the planned sample (thus also for nonrespondents)
- Moreover since response probabilities are unknown, groups get formed by using observed response rates inside groups as estimates:

$$\begin{cases} \forall k \in G_j \\ \varphi_k = \Phi^j \quad unknown \quad \mapsto \quad \hat{\varphi}_k = \hat{\Phi}^j \quad estimated \\ \hat{\Phi}^j = m_j / n_j \quad where \quad m_j = \left\| G_j \cap r \right\| \quad and \quad n_j = \left\| G_j \cap s \right\| \end{cases}$$

- From the above formulas one recognizes that under the RHG model (and given the actual response rates for the groups) the response set is seen as Stratified SRSWOR sample drawn from the planned sample, with strata given by the Groups

The Response Homogeneity Group Model (3/3)

- Very often Groups for the RHG model are built by collapsing in some 'smart' fashion real design-strata
- Here 'smart collapsing' means a collapsing such that the overall response rates inside each obtained Group are found nearly constant
- A technique used in Istat for the survey on the small and medium enterprises (PMI) is as follows:
  a) Compute response rates inside real design-strata (since strata are a big number, some of those rates may be 0 due to nonresponse)
  b) Compute the deciles of the distribution from point a) and classify enterprises inside the 10 obtained cells
  c) Treat enterprises belonging to each cell as a distinct RHG group
  d) Attach to each enterprise an estimated response probability $\hat{\varphi}_k$ given by the overall response rate for its group
- At the end, each enterprise gets a new nonresponse adjusted weight: $\tilde{w}_k = d_k / \hat{\varphi}_k$
- This weight is further used as an initial weight in a subsequent calibration step (whose major aim is variance reduction)

Handling Nonresponse by Calibration (1/2)

- To understand why calibration should succeed in reducing nonresponse bias, one has first to define when calibration variables are 'good'
- A good auxiliary vector (to calibrate on) should be one that:
  a) is able to explain the variation of response propensity
  b) covaries with the main study variables
  c) Identifies the most important estimation domains
- Fulfilling requirement a) is of crucial importance for achieving an effective reduction of nonresponse bias for all possible variables of interest
- Property b) is the basic condition (already discussed) to ensure a variance reduction when using calibration estimators; it also helps to remove the nonresponse bias for the covariates (a smaller set of study variables)
- Principle c) gives a contribution to both bias and variance reduction in domain estimates

Handling Nonresponse by Calibration (2/2)

- Because nonresponse mechanism is unknown, it is impossible to prove theoretically that requirement a) will do the job
- Anyway Monte Carlo simulation studies (in which nonresponse is fully under control) give strong support
- To get an intuitive insight of why should requirement a) work one can think as follows:
  - calibrating on variables that are strongly correlated with (non)response behavior exactly removes nonresponse bias from the estimates of the population totals of the auxiliary variables
  - thus it is likely (actually true for most of the practical applications) that calibration at least decreases nonresponse bias for study variables covarying with the auxiliary ones

Different Levels of Auxiliary Information (1/2)

- When using calibration to fight nonresponse bias, 2 distinct levels of auxiliary information can be used
- Info-U
  - the population total of variables $\mathbf{x} = (x_1, \ldots, x_q)$ is known:

$$\mathbf{X}_U = \sum_{k \in U} \mathbf{x}_k$$

- Info-S
  - auxiliary variables values are known for every unit belonging to the selected sample (thus also for nonrespondents), so one can compute:

$$\mathbf{X}_S = \sum_{k \in s} d_k \mathbf{x}_k$$

- In both cases the values of the auxiliary variables $\mathbf{x}_k$ must be known for every respondent unit k belonging to the response set r
- When powerful auxiliary information is available, it is also possible to mix InfoU and InfoS variables: the resulting vector can be denoted as Info-US

Different Levels of Auxiliary Information (2/2)

- When we discussed calibration as a tool to gain precision we only dealt with auxiliary information of the Info-U kind
- Info-S is something new: we are using $\mathbf{X}_S$ as it was a population total while it is actually an HT estimate (though an estimate built by using both respondents and nonrespondents)
  - the key point is that calibrating the respondent weights on an unbiased estimate of a population total (which $\mathbf{X}_S$ is) turns to be enough to soften nonresponse bias
- Notice that Info-S calibration will be less effective then Info-U in decreasing estimators variance, since some additional variability is introduced into the calibration constraints via the benchmark values $\mathbf{X}_S$
- For this reason, after having performed an Info-S calibration step to handle nonresponse bias, almost always a subsequent Info-U calibration step is carried out to gain efficiency

1-Step Calibration vs 2-Step Calibration (1/2)

- The calibration problem for Info-U and Info-S reads:

$$\begin{cases} \min \sum_{k \in r} G(w_k^{U,S}, d_k) \\ \sum_{k \in r} w_k^{U,S} \mathbf{x}_k = \mathbf{X}_{U,S} \end{cases}$$

- When auxiliary information fulfills simultaneously all conditions a) b) and c), the solution of the above problem can achieve the goal of reducing both nonresponse bias and variance
- After this one-step calibration (CAL1S) procedure the final weights will be simply:

$$w_k^{U,S} = g_k^{U,S} d_k$$

- Otherwise, when only condition a) holds a subsequent calibration is required to gain efficiency (no matter if you are using Info-U or Info-S)
- This leads to the widely diffused two-step calibration technique

1-Step Calibration vs 2-Step Calibration (2/2)

- The two-step calibration (CAL2S) technique is summarized as follows:
  - Step1. Calibrate direct weights to reduce nonresponse Bias
  - Step2. Calibrate Step1 output weights to reduce estimators Variance
- CAL2S final weights will be:

$$w_k^{fin} = g_k^{Step2} w_k^{Step1} = g_k^{Step2} g_k^{Step1} d_k$$

- Notice that CAL2S final weights will satisfy exactly only Step2 calibration constraints, while slightly violating those imposed in Step1
- This 'soft' violation of Step1 calibration constraints can be shown to have only a negligible impact: nonresponse bias doesn't resurrect anymore
- Notice also that, for the same given set of auxiliary variables, CAL2S shows a big advantage over CAL1S, namely it needs a substantially lower computational burden (indeed complexity of calibration algorithms grows more than linearly with the number of auxiliary variables)
- As a consequence, CAL2S is sometimes preferred to CAL1S even when powerful auxiliary information fulfilling a) b) and c) is available

Appendix – Selected Topics

Estimating 2-stage Variance in Practice (1/2)

- Under 2-stage stratified cluster sampling without replacement and equal inclusion probabilities, HT variance estimator formula has general structure

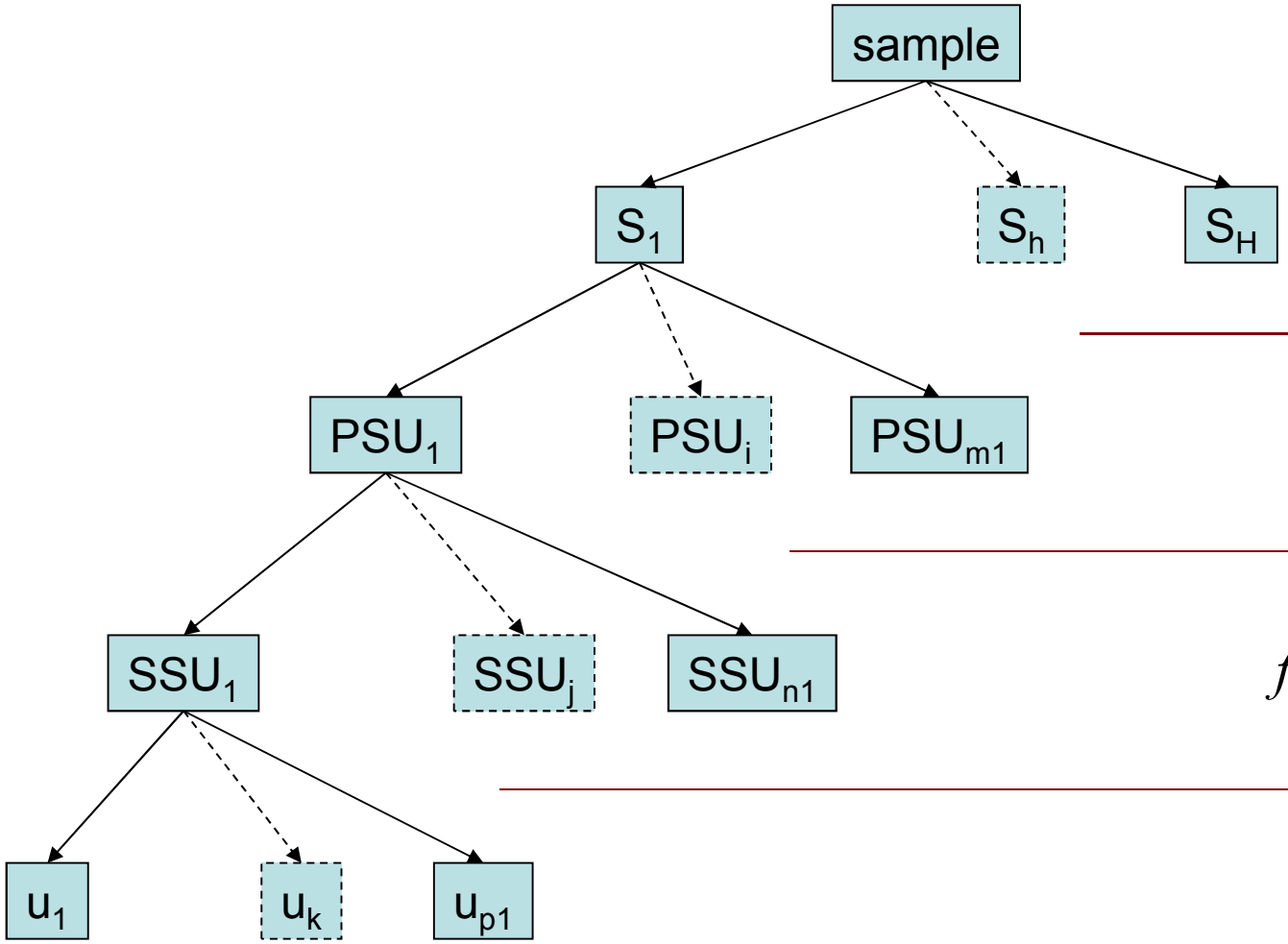
$$\hat{V}_{2S}(\hat{Y}_{HT}) = \sum_{strat} \left\{ \sum_{psu} \left[\left(1 - f_{strat}^{PSU}\right) v_1(\breve{y}_{psu}) + \sum_{ssu} f_{strat}^{PSU} \left(1 - f_{psu}^{SSU}\right) v_2(\breve{y}_{ssu}) \right] \right\}$$

- where:
  - f_{strat}^{PSU} is the sampling fraction of PSUs in stratum $strat$
  - f_{psu}^{SSU} is the sampling fraction of SSUs inside sampled PSU psu
  - $\breve{y}_{clus}$ stands for the weighted total of y inside cluster $clus$
  - the functional form of v_i depends on the adopted sampling scheme at stage i

- This formula can be computed by representing the sample as a tree and then summing all the contribution attached to its nodes

Estimating 2-stage Variance in Practice (2/2)

Must sum contributions from all the nodes of the sampling tree



$$\sum_S V_S$$

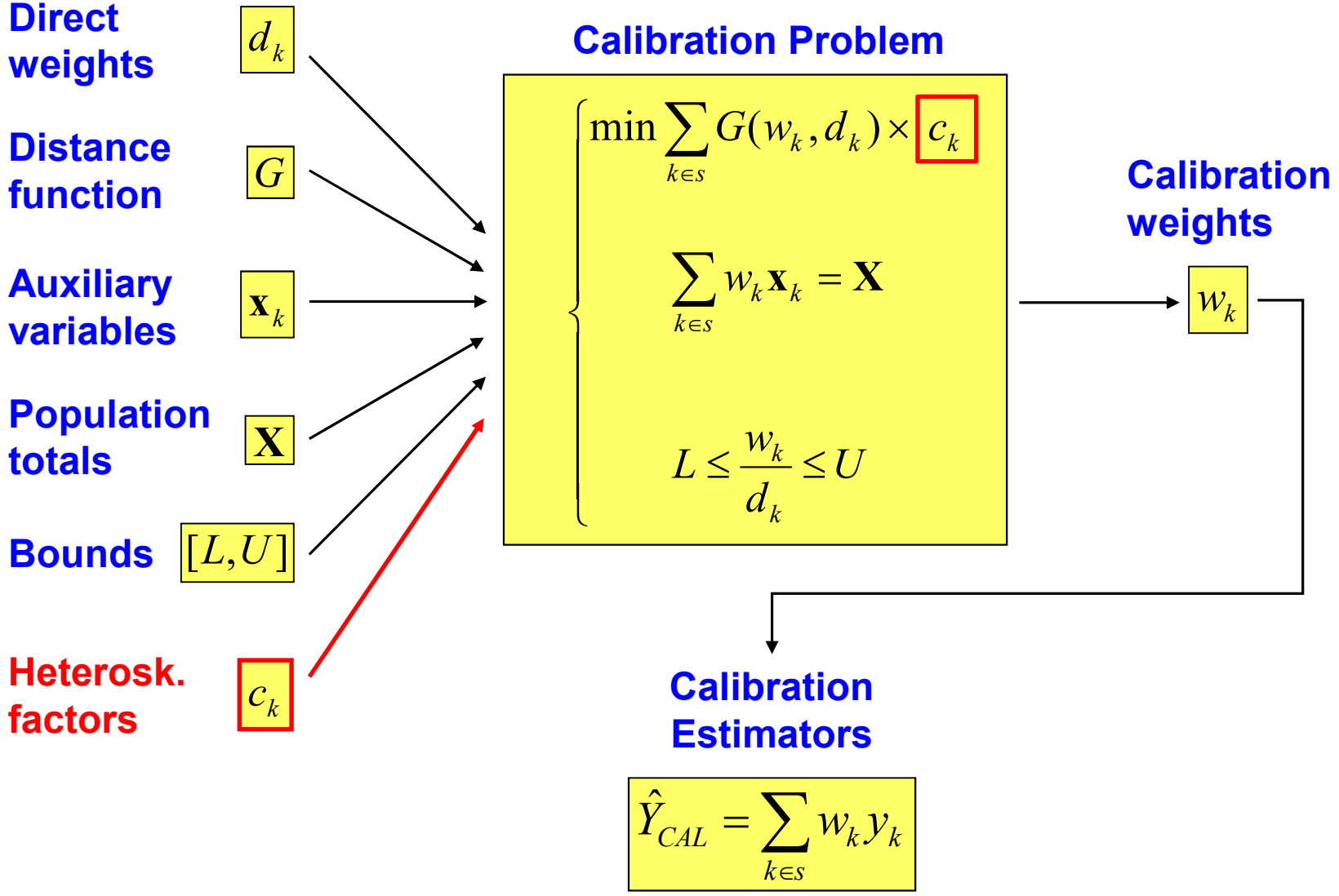
$$V_S$$

$$\left(1 - f_{strat}^{PSU}\right) v_1(\breve{y}_{psu})$$

$$f_{strat}^{PSU}\left(1 - f_{psu}^{SSU}\right) v_2(\breve{y}_{ssu})$$

$$\breve{y}_k = d_k y_k$$

Calibration with Heteroskedasticity (1/5)



Calibration with Heteroskedasticity (2/5)

- A more general version of the calibration problem can incorporate a set of multiplicative factors c_k inside the distance function
- Such c_k factors must be:
  - Strictly positive
  - Uncorrelated with direct weights d_k
- Most applications will use $c_k = 1$ for all sample units k, which leads back to the ordinary calibration problem
  - This is generally the standard choice in social surveys
- The final effect of using any nontrivial (i.e. non-constant) set of c_k values will be that:

  on average, calibrated weights w_k of units with higher values of c_k will tend to stay closer to their corresponding initial weights d_k than would happen for units with lower c_k

Calibration with Heteroskedasticity (3/5)

- Thus, with a smart choice of factors c_k, we may try to prevent the calibration algorithm to change too much the weights of selected influent units
  - This opportunity is often exploited in enterprise surveys
- Interest variables in enterprise surveys may be highly skewed, so that few big firms can account for a significant fraction of the population total Y
- Even if these big firms are censused, i.e. their direct weight is 1, calibration could inflate their weight, thus generating influent outliers, which could impair estimation

$$k: \ (y_k \ is \ big) \ AND \ (w_k \ is \ big)$$

- To contrast this risk, a good choice is to use the "enterprise size" (e.g. as measured by the number of employees) to set the c_k values:

$$c_k = empnum_k$$

Calibration with Heteroskedasticity (4/5)

- To understand why the c_k are called heteroskedasticity factors one has to resort to the GREG
- Indeed, if the linear assisting model ξ underlying the calibration problem is heteroskedastic

$$\xi: \quad y_k \sim \mathbf{x}_k \cdot \boldsymbol{\beta} + \varepsilon_k$$

$$E_\xi(\varepsilon_k) = 0, \quad V_\xi(\varepsilon_k) = c_k \cdot \sigma^2 < \infty, \quad Cov_\xi(\varepsilon_k, \varepsilon_j) = 0 \quad \forall k, j \in U$$

- the GREG estimator for the total of y still reads:

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \hat{\boldsymbol{\beta}}$$

- but now the estimator for β has the Generalized Least Squares expression:

$$\hat{\boldsymbol{\beta}} = (X^t DC^{-1} X)^{-1} (X^t DC^{-1} Y) = \left(\sum_{k \in s} \frac{d_k}{c_k} \mathbf{x}_k^t \cdot \mathbf{x}_k \right)^{-1} \cdot \left(\sum_{k \in s} \frac{d_k}{c_k} \mathbf{x}_k^t y_k \right) = \mathbf{T}^{-1} \cdot \mathbf{t}$$

- where C is the diagonal matrix of the c_k

Calibration with Heteroskedasticity (5/5)

- Since the weighted estimator of β is still linear in y, the GREG can still be expressed as a weighted sum of y_k:

$$\hat{Y}_{GREG} = \sum_{k \in s} w_k y_k = \sum_{k \in s} g_k (d_k y_k)$$

- but now the g-weights explicitly depend on the heteroskedasticity factors c_k:

$$g_k = w_k / d_k = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \mathbf{T}^{-1} \cdot \mathbf{x}_k^t \, c_k^{-1}$$

- Now it's easy to prove that Unbounded Linear Calibration with nontrivial c_k values yields exactly the same g-weights as the Heteroskedastic GREG expression above

- Similarly, for Raking and Logit distances nontrivial c_k values imply:

$$g_k = F\left(\mathbf{x}_k \cdot \boldsymbol{\lambda} \, c_k^{-1}\right) \quad where \quad F = \left(\partial G / \partial w\right)^{-1}$$

The Ratio Estimator as a Calibration Estimator

- The Ratio Estimator of the Total long predates the theory of Calibration
- It exploits as auxiliary information the population total X of a single numeric variable x which is believed to be positively correlated with the interest variable y

$$\hat{Y}_R = X\left(\frac{\hat{Y}}{\hat{X}}\right)$$

- Interestingly enough, while the Ratio Estimator can be shown to be a specific case of Calibration, this necessarily requires a heteroskedastic calibration model, with $c_k = x_k$

$$c_k = x_k \quad \Rightarrow \quad \hat{\beta} = \left(\sum_{k\in s} \frac{d_k}{c_k} x_k^{\,2}\right)^{-1} \cdot \left(\sum_{k\in s} \frac{d_k}{c_k} x_k y_k\right) = \frac{\hat{Y}}{\hat{X}}$$

$$\hat{Y}_{GREG} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}}) \cdot \hat{\boldsymbol{\beta}} = \hat{Y} + (X - \hat{X}) \cdot \frac{\hat{Y}}{\hat{X}} = X\left(\frac{\hat{Y}}{\hat{X}}\right) = \hat{Y}_R$$

Taylor Linearization Variance in a Nutshell (1/3)

- Complex population parameter, non linear function of population totals

$$\theta = f(Y_1,\ldots,Y_m)$$

- Natural estimator (generally biased): same function of HT estimators

$$\hat{\theta} = f(\hat{Y}_1,\ldots,\hat{Y}_m)$$

- Expand it by Taylor series to first order around true totals $\mathbf{Y}=(Y_1,\ldots,Y_m)$

$$\hat{\theta} \approx \theta + \sum_{j=1}^{m} \left.\frac{\partial f}{\partial \hat{Y}_j}\right|_{\mathbf{Y}} (\hat{Y}_j - Y_j) = \hat{\theta}_{lin}$$

- Using explicit expressions for HT estimators you get

$$\hat{\theta}_{lin} = \sum_{k\in s} d_k z_k + const \qquad z_k = \sum_{j=1}^{m} \left.\frac{\partial f}{\partial \hat{Y}_j}\right|_{\mathbf{Y}} y_{jk}$$

- Here *const* means constant (maybe unknown) terms not depending on the sample. Notice that the linearized estimator is unbiased for θ

Taylor Linearization Variance in a Nutshell (2/3)

- Now you are left with the problem of estimating the variance of an HT estimator for the total of a new variable z, namely the linearized variable derived by θ (aka Woodruff transform)

$$V(\hat{\theta}) \approx V(\sum_{k \in s} d_k z_k) = \sum_{k \in s} \sum_{j \in s} d_k z_k \Delta_{kj} d_j z_j$$

- Because the linearized variable depends on unknown population totals, it must be substituted with the same expression computed using sample estimates

$$\tilde{z}_k = \sum_{j=1}^{m} \left. \frac{\partial f}{\partial \hat{Y}_j} \right|_{\hat{\mathbf{Y}}} y_{jk}$$

- So you end with the approximate variance estimator

$$\hat{V}(\hat{\theta}) \approx \hat{V}(\sum_{k \in s} d_k \tilde{z}_k) = \sum_{k \in s} \sum_{j \in s} d_k \tilde{z}_k \left(\frac{\Delta_{kj}}{\pi_{kj}} \right) d_j \tilde{z}_j$$

Taylor Linearization Variance in a Nutshell (3/3)

- Notice that the linearized variable associated to a complex estimator can also be expressed as follows:

$$\tilde{z}_k = \sum_{j=1}^{m} \left.\frac{\partial f}{\partial \hat{Y}_j}\right|_{\hat{\mathbf{Y}}} y_{jk} = \sum_{j=1}^{m} \left.\frac{\partial f}{\partial \hat{Y}_j}\right|_{\hat{\mathbf{Y}}} \frac{\partial \hat{Y}_j}{\partial (d_k)} = \left.\frac{df}{d(d_k)}\right|_{\mathbf{d}}$$

- i.e. it equals the total derivative of the complex estimator w.r.t. the direct weights, evaluated at $\mathbf{d}=(d_1,\ldots,d_n)$
- This formula can be useful whenever it's easier to think the complex estimator as a complex function of direct weights, rather then of HT estimators. GREG and Calibration estimators are good examples

Taylor Linearization Examples

- Ratio Estimator (numerator and denominator are both supposed HT estimators)

$$\hat{R} = \frac{\hat{Y}}{\hat{X}}$$

- corresponding linearization variable (Woodruff transform)

$$\tilde{z}_k = \left.\frac{\partial \hat{R}}{\partial \hat{Y}}\right|_{\hat{X},\hat{Y}} y_k + \left.\frac{\partial \hat{R}}{\partial \hat{X}}\right|_{\hat{X},\hat{Y}} x_k = \frac{1}{\hat{X}}(y_k - \hat{R}x_k)$$

- Ratio Estimator of a Total (population total X for the denominator variable is supposed to be known from external sources)

$$\hat{Y}_R = X\left(\frac{\hat{Y}}{\hat{X}}\right) = X\hat{R}$$

- corresponding linearization variable (Woodruff transform)

$$\tilde{z}_k = \left.\frac{\partial \hat{Y}_R}{\partial \hat{Y}}\right|_{\hat{X},\hat{Y}} y_k + \left.\frac{\partial \hat{Y}_R}{\partial \hat{X}}\right|_{\hat{X},\hat{Y}} x_k = \frac{X}{\hat{X}}(y_k - \hat{R}x_k)$$

Exercise: GREG Taylor Linearization (1/2)

- Start with GREG estimator expressions

$$\hat{Y}_{GREG} = (\sum_{k \in U} \mathbf{x}_k) \cdot \hat{\boldsymbol{\beta}} + \sum_{k \in s} d_k (y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}) = \sum_{k \in s} g_k (d_k y_k)$$

- using available formulas for the regression coefficients estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s} d_k \mathbf{x}_k^t \cdot \mathbf{x}_k \right)^{-1} \cdot \left(\sum_{k \in s} d_k \mathbf{x}_k^t y_k \right) = \mathbf{T}^{-1} \cdot \mathbf{t}$$

- first obtain g-weights expression

$$g_k = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \cdot \mathbf{T}^{-1} \cdot \mathbf{x}_k^t$$

- then compute explicitly GREG linearization variable (a là Woodruff)

$$\tilde{z}_k = \frac{d}{d(d_k)} \hat{Y}_{GREG} \Big|_{\mathbf{d}} = g_k y_k + \sum_{j \in s} d_j y_j \frac{d}{d(d_k)} g_j \Big|_{\mathbf{d}}$$

Exercise: GREG Taylor Linearization (2/2)

- start computing the derivative of g_j w.r.t. d_k

$$\frac{d}{d(d_k)} g_j = -\mathbf{x}_k \mathbf{T}^{-1} \mathbf{x}_j^t + (\mathbf{X} - \hat{\mathbf{X}}_{HT}) \left[-\mathbf{T}^{-1} \frac{d\mathbf{T}}{d(d_k)} \mathbf{T}^{-1} \right] \mathbf{x}_j^t$$

- then compute the derivative of $\mathbf{T}$ w.r.t d_k

$$\frac{d}{d(d_k)} \mathbf{T} = \mathbf{x}_k^t \mathbf{x}_k$$

- lastly substitute and simplify to get final result

$$\tilde{z}_k = g_k y_k + \sum_{j \in s} d_j y_j \frac{d}{d(d_k)} g_j \Big|_{\mathbf{d}} = g_k y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}} - (g_k - 1)\mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}$$

$$= g_k (y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}})$$

$$= g_k e_k$$

- Above expression leads directly to Sarndal proposal for the GREG variance estimator