MULTIPLE-FRAME SAMPLING

LUIS AMBROSIO FLORES
The required information

Sustainable development

Economic
- Agriculture
- Remaining sectors

Environmental
- Natural resources, uses and conservation
- Land
- Water
- Air

Social
- Labour
- Households
- Family budgets
- Poverty

Macro-economy
- Aggregated values
- Economic counts and balance

Micro-economy
- Farm economic
THREE COUNTRY EXAMPLES

Agricultural Master Sampling Frame
• A dual frame for agricultural surveys
  – Area frame of segments
    • Guatemala: 190100 segments with geometric limits
    • Ecuador: 352254 segments with geometric limits
    • Costa Rica: 120360 segments with identifiable physical boundaries
  – List frame of farms

Household Master Sampling Frame
• An area sampling frame of Enumeration Areas (EAs) with identifiable physical boundaries
The Multiple-frame Master Sampling of Costa Rica

MULTIPLE-FRAME SAMPLING
• The number of single frames is $Q$

• We select a set of samples $\{S^q; q = 1, 2, \cdots, Q\}$

  independently from each single frame

• The inclusion probabilities $\pi^q_i$ differ from a single frame to another
MULTIPLE-FRAME ESTIMATORS

• **Biased**: Typically, a population unit (e.g., a farm) is covered by two or more single frames (e.g., area and list frames) and, as a result, the weight estimator is biased

\[
\hat{Y}_p = \sum_{q=1}^{Q} \sum_{i=1}^{S_q} w_i^q y_i
\]

• Alternative unbiased approaches
  
  • Adjusted-weight estimators

\[
\hat{Y}_p = \sum_{q=1}^{Q} \sum_{i=1}^{S_q} \tilde{w}_i^q y_i
\]

• Multiplicity-adjusted estimators

\[
\hat{Y}_p = \sum_{q=1}^{Q} \sum_{i=1}^{S_q} w_i^q \tilde{y}_i^q
\]
Multiplicity of a population unit: Is the number of sampling units with which it is associated,

\[ m_j = \sum_{q=1}^{Q} m^q_j \]

Multiplicity-adjusted value of the survey variable in the sampling unit

\[ \hat{\gamma}^q_i = \sum_{j=1}^{P} \alpha^q_{ij} y_j \]

where

\[ \alpha^q_{ij} = \frac{I^q_{ij}}{m_j} \]

\[ I^q_{ij} = 1 \quad \text{if the population unit is associated with the sampling unit} \]

The estimator is:

\[ \hat{Y} = \sum_{q=1}^{Q} \sum_{i=1}^{S^q} W^q_i \hat{\gamma}^q_i \]

\[ V\hat{Y}_{P} = \sum_{q=1}^{Q} \sum_{i=1}^{A^q} \sum_{i'=1}^{A^q} \left( \frac{\pi^q_{ii'}}{\pi^q} - \pi^q_{ii'} \frac{\hat{\gamma}^q_{Pi}}{\pi^q} \frac{\hat{\gamma}^q_{Pi'}}{\pi^q_{i'}} \right) \]
MULTIPLE-FRAME REGRESSION ESTIMATORS

The model is:

\[
\tilde{y}_i^q = \tilde{x}_i^q \beta + \tilde{\varepsilon}_i^q
\]

The Multiplicity-adjusted GREG estimator is:

\[
\hat{Y}_{MGREG} = \sum \sum \tilde{x}_i^q \hat{\beta}_w
\]

where

\[
\hat{\beta}_w = (\tilde{X}^T D_w \tilde{X})^{-1} \tilde{X}^T D_w \tilde{Y}
\]

The asymptotic design-variance is:

\[
\hat{\hat{\Sigma}}_{MGREG} = \hat{\hat{\Sigma}} \left( \sum Q \sum S^q w_i^q \hat{\varepsilon}_i^q \right)
\]

where

\[
\hat{\varepsilon}_i^q = \tilde{y}_i^q - \tilde{x}_i^q \hat{\beta}_w
\]
We consider the data set \((y_j, x_j)\), where

- \(y_j\) is the survey variable
- \(x_j\) are register data

The MGREG estimator is

\[
\hat{Y}_{MGREG} = \sum_{q=1}^{Q} \sum_{i=1}^{S_q} w_i^q \tilde{x}_i^q \beta_{iW}^q
\]

\[
\hat{VY}_{MGREG} = \hat{V} \sum_{q=1}^{Q} \sum_{i=1}^{S_q} w_i^q \hat{e}_i^q + \sum_{q=1}^{Q} \hat{\beta}_{iW}^q \left( \hat{V} \sum_{i=1}^{S_q} w_i^q \tilde{x}_i^q \right) \hat{\beta}_{iW}^q
\]

\[
\hat{e}_i^q = \tilde{y}_i^q - \tilde{x}_i^q \hat{\beta}_{iW}^q
\]
SMALL AREA ESTIMATION

The population is partitioned into \( d = 1, 2, \ldots, D \) small areas.

In a big secondary sample (e.g. a register) we observe \( X_j \).

The estimator is

\[
\hat{Y}_{dreg, bc} = \sum_{j=1}^{n_j} \hat{w}_j \delta_j(d) x_j \hat{B}
\]

where \( \delta_j(d) = 1 \) if unit “\( j \)” is from small area “\( d \)”.

The asymptotic variance is

\[
V_{d} \hat{p} \lim \left( \hat{Y}_{dreg, bc} - Y_d \right) = V \sum_{j=1}^{n_j} \hat{w}_j \delta_j(d) \left( y_j - x_j \beta \right) + \beta^T \left( \text{Var} \sum_{j=1}^{n_j} \hat{w}_j \delta_j(d) x_j \right) \beta
\]
ESTIMATION IN TIME

- We want to estimate the survey variable total, \( y_t \), using the sample of the period “t” and the estimates of the previous periods, \( \{ \hat{y}_t; t = t-1, t-2, \ldots, 1 \} \).

- We consider the model \( \hat{y}_t = y_t + u_t \), where \( \hat{y}_t \) is a unbiased multiple-frame estimator of \( y_t \) and \( u_t = \hat{y}_t - y_t \) is the sampling error.

- The Best Linear Unbiased Predictor (BLUP) is \( \hat{y}_{BLUP} = GV^{-1} \hat{y} \) where \( \hat{y} = y + Vu = G + R \).

- Its variance is \( V\hat{y}_{BLUP} = (R^{-1} + G^{-1})^{-1} \).

- The estimate change series is \( \Delta y = Cy_{BLUP} = CGV^{-1} \hat{y} \) and \( \text{Var}\Delta y = CV\text{Var}\hat{y}_{BLUP}C^T \).

- The predictor is \( \hat{y}_{T+h} = \sum_{i=1}^{T} a_i \hat{y}_{tBLUP} \) where,

  \[
  a = (\text{Var}\hat{y}_{BLUP})^{-1} \left[ C + 1^T (\text{Var}\hat{y}_{BLUP})^{-1} 1 \right] \left( 1 - 1^T (\text{Var}\hat{y}_{BLUP})^{-1} C \right)
  \]
Comparison of the domain means:

- The asymptotic distribution of the estimator \( \hat{Y} = [\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_d, \ldots, \hat{Y}_D]^{T} \) is

\[
\sqrt{n}(\hat{Y} - \mu) \rightarrow N(0, \text{Var}\hat{Y})
\]

- We want to test the hypothesis \( H_0: \mu_1 = \mu_2 = \cdots = \mu_d = \cdots = \mu_D \)

- If \( H_0 \) is true, then \( \sqrt{nR\hat{Y}} \rightarrow AN\left(0, R\text{Var}\hat{Y}R^{T}\right) \) where \( R \) has raws

\[
[1 \ 0 \cdots -1 \cdots 0]
\]

- The hypothesis is refused if

\[
\left(R\hat{Y}\right)^{T}\left(R\text{Var}\hat{Y}R^{T}\right)^{-1}R\hat{Y} > \chi^2_{D,1-\alpha}
\]
ANALYSIS OF COMPLEX SURVEYS: LINEAR MODELS

• We assume that the finite population is a iid sample from a superpopulation generated by the linear model
  \[ y_j = x_j \beta + \varepsilon_j \]

• The model in terms of sampling units is
  \[ \tilde{y}_i^q = \tilde{x}_i^q \beta + \tilde{\varepsilon}_i^q \]

• The weighted estimator
  \[ \hat{\beta}_w = \left( \tilde{X}^T D_w \tilde{X} \right)^{-1} \tilde{X}^T D_w \tilde{y} \]
  is consistent

• Its covariance matrix is
  \[ \hat{V}_w = \hat{M}_{xw}^{-1} \hat{V}_{bb} \hat{M}_{xw}^{-1}, \text{ where } \hat{M}_{xw} = \frac{1}{n_N} \tilde{X}^T D_w \tilde{X} \]

• The asymptotic distribution of
  \[ \hat{\beta}_w \]
  is
  \[ \frac{\hat{\beta}_w - \beta}{\sqrt{\hat{V}_w}} \rightarrow \mathcal{N}(0, I) \]
  and can be used for hypothesis testing

MULTIPLE-FRAME SAMPLING
• We assume that the finite population is a sample from a superpopulation generated by the generalized linear model $f(y_j, \theta)$.

• The weight estimator $(\hat{\theta}_W - \theta_0) = \left[\frac{1}{N} \sum_{q=1}^{Q} \sum_{i=1}^{s_q} \tilde{w}_i \frac{\partial^2 I(\tilde{y}_i, \theta_0)}{\partial \theta \partial \theta^T} \right]^{-1} \frac{1}{N} \sum_{q=1}^{Q} \sum_{i=1}^{s_q} \tilde{w}_i \frac{\partial I(\tilde{y}_i, \theta_0)}{\partial \theta} + O_p\left(\frac{1}{n_N}\right)$ is consistent.

• Its asymptotic variance is $\hat{V}(\hat{\theta}_W - \theta_0) = \hat{V}(\hat{\theta}_W | F_N) + \hat{V}(\theta_N - \theta_0)$.

• The asymptotic distribution of $\hat{\theta}_W - \theta_N | F_N$ is $\frac{\hat{\theta}_W - \theta_N | F_N}{\sqrt{\hat{V}(\hat{\theta}_W | F_N)}} \rightarrow N(0, I)$ and can be used for hypothesis testing.
Thank you

(LUIS AMBROSIO luis.ambrosio@upm.es)