Project Survey Sampling

DRUILHE, LAROSE & SAUE

December 2024

Table of Contents

Part I: Three different estimators

Data Overview

The population under consideration consists of the 554 *communes* in the Haute-Garonne department of France with fewer than 10,000 inhabitants in 1999. The key variables include:

- `CODE_N`: Code of the *commune*.
- `COMMUNE`: Name of the *commune*.
- `BVQ_N`: Code of the *Bassin de vie quotidienne* (local life area).
- `POPSDC99`: Population of the *commune* in 1999.
- `LOG`: Number of dwellings or housing units (auxiliary variable).
- `stratlog`: Stratification variable based on `LOG`, with 4 categories:
  - 1 if $\texttt{LOG} < 100$,
  - 2 if $100 \leq \texttt{LOG} < 300$,
  - 3 if $300 \leq \texttt{LOG} < 1000$,
  - 4 if $\texttt{LOG} \geq 1000$.
- `LOGVAC`: Number of empty dwellings (variable of interest).

The poststratified estimator: Definition

Definition

The poststratified estimator of the total Y is defined as:

$$\hat{Y}_{st} = \sum_{q=1}^{Q} N_q \cdot \bar{y}_q$$

where:

- N_q: Population size of stratum q
- $\bar{y}_q$: Sample mean of the variable of interest y (e.g., `LOGVAC`) within stratum q
- Q: Total number of strata.

The poststratified estimator: Statistics

Statistics on the whole population:

- $N = 554$
- $Y = 10768$
- $S_y^2 = 1104.5$

Statistics on the strata:

	$\mathbf{q = 1}$	$\mathbf{q = 2}$	$\mathbf{q = 3}$	$\mathbf{q = 4}$
N_q	221	169	110	54
Y_q	895	1807	3341	4725
S_{yq}^2	11.06569	47.13095	459.7589	4184.142

The poststratified estimator: Estimators

Horvitz-Thompson estimator (SRWOR)

- $\hat{Y}_{HT} = 10914$
- $SE(\hat{Y}_{HT}) = 1906.75$

Poststratified estimator

- $\hat{Y}_{st} = 11195$
- $SE(\hat{Y}_{st}) = 1037.2$

The poststratified estimator: Estimators

Poststratified estimator (computed with R)

- $\hat{Y}_{st} = 11195$
- $SE(\hat{Y}_{st}) = 1037.2$

Poststratified estimator (manually computed)

- $\hat{Y}_{st} = 11195.167$
- $SE(\hat{Y}_{st}) = 1186.57$

The poststratified estimator: Simulations

We draw 1000 samples, here are the results:

SRSWOR (HT)

- Monte Carlo Mean: 10787.18
- Monte Carlo SD: 2057.044
- Monte Carlo CV: 19.0693

Poststratified

- Monte Carlo Mean: 10763.04
- Monte Carlo SD: 1489.289
- Monte Carlo CV: 13.837

The poststratified estimator: Simulations

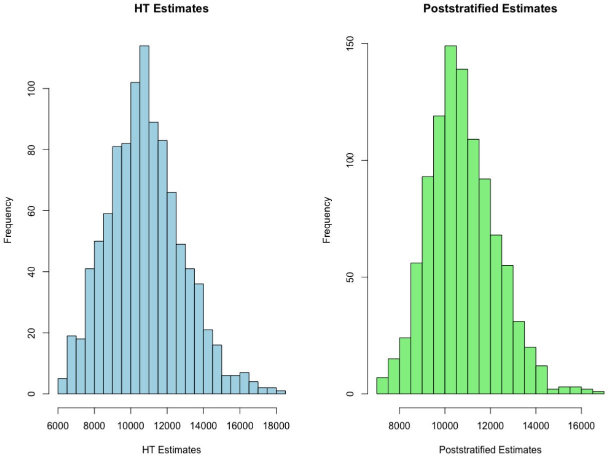


Figure: Histogram of the 1000 samples : HT vs Poststratified

The regression estimator: Simulations

Analysis of the results:

- These results align with expectations, as poststratification typically improves the efficiency of the estimates by reducing variance when appropriate auxiliary information is available.

The ratio estimator: Definition

Definition

The ration estimator of the total Y is defined as:

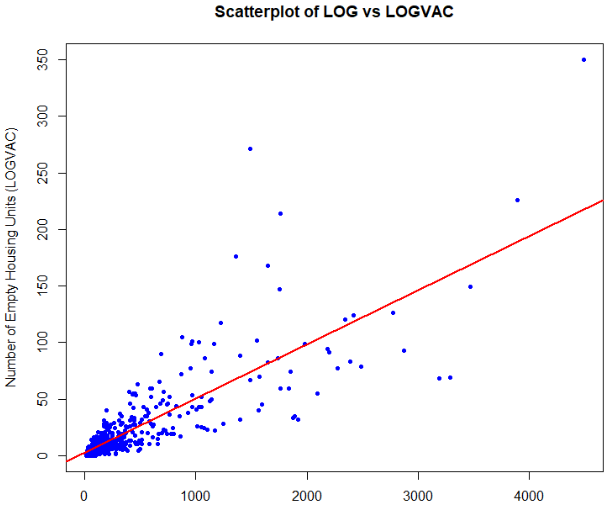
$$\hat{Y}_R = R \cdot X$$

where:

- $R = \sum y / \sum x$ (sample ratio)
- X is the total of x in the population (197314)

The ratio estimator : Preliminary results

Checking for linearity:



The ratio estiamtor: The one sample case

Let's compute the Ratio Estimator.
First we compute the ratio $\sum y / \sum x$ for the sample using the following code:

```
est.ratio← svyratio( LOGVAC,  LOG, ech.si)
```

Then we use the ratio to predict the total Y for the population.

```
predict(est.ratio, total = 197314)
```

Next we verify by computing 'manually' and find the same result as with the built-in function, namely:

$$\hat{Y}_{ratio} = 11681.32$$

$$SE(\hat{Y}_{ratio}) = 875.523$$

The ratio estimator: Simulations

We draw 1000 samples, here are the results:

SRSWOR (HT)

- Monte Carlo Mean: 10787.18
- Monte Carlo SD: 2057.044
- Monte Carlo CV: 19.0693

Ratio

- Monte Carlo Mean: 10866.34
- Monte Carlo SD: 1250.761
- Monte Carlo CV: 11.51042

The ratio estimator: Simulations

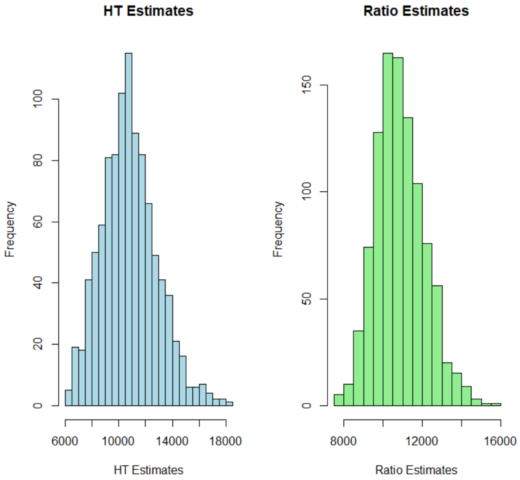


Figure: Histogram of the 1000 samples : HT vs Ratio

The regression estimator: Definition

The regression estimator for the total, $\hat{Y}_{\text{reg}}$, is given by:

$$\hat{Y}_{\text{reg}} = \sum_{i=1}^{n} w_i y_i + \sum_{j=1}^{p} (\bar{X}_j - \hat{\bar{X}}_j)\beta_j$$

Where:

- w_i: Original sampling weight for unit i,
- y_i: Value of the variable of interest (`LOGVAC`),
- $\bar{X}_j$: Known population mean of auxiliary variable j (`LOG`),
- $\hat{\bar{X}}_j$: Sample mean of auxiliary variable j,
- β_j: Regression coefficient for auxiliary variable j, calculated as:

$$\beta_j = \frac{\text{Cov}(y, x_j)}{\text{Var}(x_j)}$$

The regression estimator

The calibration of survey weights is performed using the `calibrate` function:

```
ech.si.cal ← calibrate(ech.si, ~LOG, c(554, 197314))
```

Here:

- `ech.si` is the original survey design object containing the sample data.
- `~LOG` specifies the calibration variable, which in this case is `LOG`.
- `c(554, 197314)` represents the known population totals for the calibration variable.

The regression estimator: Estimator

The total number of empty housing units is estimated using the `svytotal` function:

```
total_empty_units ← svytotal(~LOGVAC, ech.si.cal)
```

Here:

- `~LOGVAC` specifies the variable for which the total is to be calculated (empty housing units).
- `ech.si.cal` is the calibrated survey design object obtained from Step 1.

We obtain :

$$\hat{Y}_{reg} = 9916.5$$
$$SE(\hat{Y}_{reg}) = 720.69$$

The regression estimator: 1. Input Known Data

We compute the regression estimator manually too.

- Known population totals for the auxiliary variable: $T_X = \sum_{i=1}^{N} x_i = c(554, 197314)$,
- Variable of interest: $y_i = \text{LOGVAC}$,
- Auxiliary variable: $x_i = \text{LOG}$.

The regression estimator: 2. Compute Sample Statistics

$$\text{Sample mean of } x : \hat{\bar{X}} = \frac{\sum w_i x_i}{\sum w_i}$$

$$\text{Sample mean of } y : \hat{\bar{Y}} = \frac{\sum w_i y_i}{\sum w_i}$$

The regression estimator: 3. Calculate regression coefficients

$$\mathrm{Cov}(y,x) = \frac{\sum w_i (y_i - \hat{\hat{Y}})(x_i - \hat{\hat{X}})}{\sum w_i}$$

$$\mathrm{Var}(x) = \frac{\sum w_i (x_i - \hat{\hat{X}})^2}{\sum w_i}$$

$$\beta = \frac{\mathrm{Cov}(y,x)}{\mathrm{Var}(x)}$$

The regression estimator: 4. Adjust for known totals

$$\hat{Y}_{\text{reg}} = \sum w_i y_i + (T_X - \hat{\bar{X}} \cdot N) \cdot \beta$$

Where:

- N: Total population size.
- T_X: Known total for the auxiliary variable (`LOG`).

Computing manually, we obtain $\hat{Y}_{reg} = 16071.62$, which is not at all the same as the one obtaining automatically with R.

The regression estimator: Simulations

We obtain :

SRSWOR (HT)

- Monte Carlo Mean: 10787.18
- Monte Carlo SD: 2057.044
- Monte Carlo CV: 19.0693

Regression

- Monte Carlo Mean: 10816.03
- Monte Carlo SD: 1262.62
- Monte Carlo CV: 11.6736

The regression estimator: Simulations

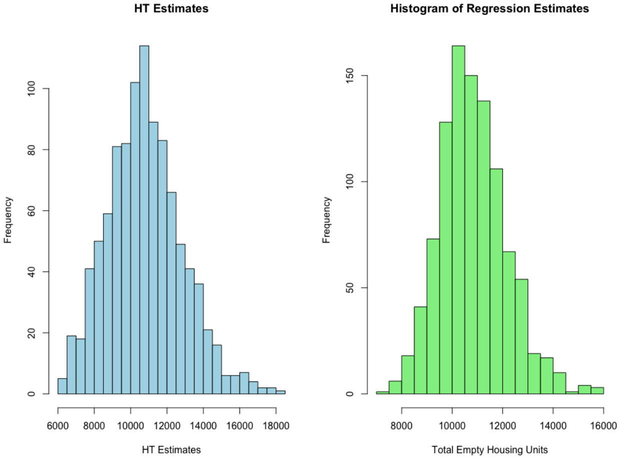


Figure: Histogram of the 1000 samples : HT vs Regression

The regression estimator: Simulations

Analysis of the results:

- If the regression estimator uses a well-chosen auxiliary variable x that is strongly correlated with y, it should have a lower variance than the HT estimator, as it leverages this relationship to improve the precision of the estimate.
- Here, the correlation between LOG and $LOGVAC$ is equal to 0.82, so it is not surprising that we have lower variance and a better approximation using the regression estimator instead of the Horvitz-Thompson one.

Part II: Imputation Methods and illustration with the Canadian LFS

Overview of the LFS

- **Purpose:** Provides monthly data on employment, unemployment, and labour market trends in Canada.
- **Key Indicators:** Unemployment rate, employment rate, participation rate.
- **Target Population:** Non-institutionalized civilians aged 15+ (excludes reserves, military, remote regions).
- **Sample Design:**
  - Two-stage sampling: Primary Sampling Units (PSUs) and dwellings.
  - Six-month rotation for efficiency in change estimation.
- **Collection Methods:** Telephone, in-person, and Internet questionnaires.

Nonresponse Challenges

Nonresponse in Surveys:

- Nonresponse occurs when data is not collected from all sampled units.
- A significant challenge for large-scale surveys like the LFS.

Why It Matters:

- Nonresponse introduces **bias** and increases **variance**, reducing survey accuracy.
- Results in distorted survey estimates, particularly for key statistics like unemployment rates.
- Response rates have steadily declined over the years.

Types of Nonresponse

1. **Unit Nonresponse:**
   - Entire sampled units fail to respond.
   - Example: A household refuses to participate in the LFS.
2. **Item Nonresponse:**
   - Some questions remain unanswered.
   - Example: Missing data for income or employment status in the LFS.

LFS Implications:

- Unit nonresponse affects coverage and sample size.
- Item nonresponse reduces data completeness for specific variables.

Effects of Nonresponse

Nonresponse Bias:

- Occurs when respondents and nonrespondents do not have the same characteristics with respect to the variables of interest.
- Example: High-income households less likely to report income, biasing average income estimates.

Nonresponse Variance:

- Smaller sample size increases variability of estimates. Variance of estimators is generally greater than that of estimators that would have been obtained if there were no nonresponse.

Objective of Nonresponse Treatment:

- Reduce bias and possibly control variance using methods like weighting and imputation.

Factors Influencing Unit Nonresponse

Reasons for Unit Nonresponse:

- **Accessibility Issues:** Difficulty in contacting sampled units
- **Amenability Issues:** Refusal to cooperate after contact is made.

Key Factors Influencing Response:

- **Type of Unit:** Establishments respond less than individuals. (Amenability)
- **Technology:** Devices like caller ID hinder contact but don't reduce eventual cooperation. (Accessibility)
- **Survey Topic:** Interest in the topic increases participation. (Amenability)
- **Mode of Collection:** Face-to-face surveys have the highest response rates; mail surveys the lowest. (Both)

Auxiliary Information in Surveys

Role of Auxiliary Information:

- Essential for estimation and modeling in the presence of nonresponse.
- Improves the quality of statistics by supporting efficient sampling, reducing coverage errors, and addressing nonresponse.

Types of Auxiliary Variables:

- **Design Variables:** Used for stratification or proportional-to-size sampling.
- **Calibration Variables:** Used at the estimation stage with known population totals (e.g., census counts).
- **Nonresponse Treatment Variables:** Applied for weighting or imputation.

Nonresponse Mechanisms

Three Mechanisms:

- **MCAR (Missing Completely At Random):** Probability of missingness is unrelated to observed or unobserved data.
- **MAR (Missing At Random):** Missingness depends only on observed auxiliary variables.
- **NMAR (Not Missing At Random):** Missingness depends on the variable of interest or unobserved data.

Example in LFS:

- Employment status is NMAR if unemployed individuals are less likely to respond.

Weighting for Unit Nonresponse: Overview

Concept:

- Response probabilities p_i are estimated using auxiliary variables z_i available from the sampling frame or past survey responses.
- Assumes p_i is modeled as $f(z_i, \gamma)$ using regression methods.

Key Formulas:

- Adjusted weight for respondent i:

$$w_i^* = \frac{d_i}{\hat{p}_i}, \quad \text{where } \hat{p}_i \text{ is the estimated response probability.}$$

- PSA estimator for the population total Y:

$$\hat{Y}_{PSA} = \sum_{i \in s_r} w_i^* y_i = \sum_{i \in s_r} \frac{d_i}{\hat{p}_i} y_i.$$

Here, s_r represents the set of survey respondents, and d_i is the initial design weight.

Estimating Response Probabilities in LFS

Parametric Estimation in LFS:

- The LFS uses logistic regression models to estimate response probabilities $\hat{p}_i$.
- Example logistic model:

$$p_i = \frac{e^{{z_i}^\top \gamma}}{1 + e^{{z_i}^\top \gamma}}, \quad \hat{p}_i = \frac{e^{{z_i}^\top \hat{\gamma}}}{1 + e^{{z_i}^\top \hat{\gamma}}}.$$

- Auxiliary variables z_i include:
  - Geographic information (e.g., urban vs. rural).
  - Demographic variables (e.g., age, household size).
  - Survey history (e.g., prior response status).

Imputation for Item Nonresponse

Why Impute?

- Nonresponse occurs when respondents fail to provide answers to specific items (e.g., wages, hours worked).
- Imputation addresses missing values to:
  - Create a complete dataset for analysis.
  - Use a single set of sampling weights for estimation.
  - Ensure consistency across users analyzing LFS data.
- Examples in LFS:
  - Imputing missing income or weekly hours worked.
  - Logical rules (e.g., deducing age from birth year) for missing demographic data.

Advantages and Cautions for Imputation

Advantages:

- Facilitates application of point estimation methods.
- Ensures complete datasets for timely publication of LFS statistics.
- Reduces bias introduced by excluding nonrespondents.

Cautions:

- Imputed data may overstate accuracy of estimates.
- May distort relationships between variables.
- Variance is underestimated if imputed values are treated as observed.

Classification of Imputation Methods

Main Groups of Methods:

- **Deterministic Methods:**
  - Regression imputation, ratio imputation, mean imputation.
  - Previous value and nearest-neighbor imputation.
- **Random Methods:**
  - Random hot-deck imputation.
  - Residual-based methods (e.g., regression or ratio imputation with residuals).

Alternative Classification:

- **Donor Methods:** Use observed values from similar respondents.
- **Predicted Value Methods:** Use functions of respondent values to generate imputations.

Overview of Imputation in the LFS

Purpose of Imputation: Steps in Data Processing:

1. Phase I editing: Validation of demographic and household data.
2. Phase II editing: Resolution of refusals and "Don't Know" responses.
3. Hot-deck imputation: Replacing missing values with donor values.
4. Post-imputation processing: Finalizing imputed data for analysis.

Hot-Deck Imputation in LFS

Concept:

- Missing values are replaced using data from a randomly selected donor in the same imputation class.
- Imputation classes are defined using socio-demographic and survey variables.

Pre-Processing for Hot-Deck Imputation:

- Records are divided into:
  - Group A: Valid and consistent donors.
  - Group B: Valid but inconsistent, not used as donors.
  - Group C: Recipients requiring imputation.
- Suspicious or extreme values (e.g., earnings) are flagged and adjusted.
- Temporary path variables (TPATH) are assigned to guide imputation.

Imputation for Item Nonresponse

Procedure:

- Donors are selected within imputation classes.
- Each imputation class is defined by crossing variables such as:
  - Labour force status, province, age group, occupation, sex, etc.
- Missing values are filled using donor values that satisfy consistency rules.

Constraints:

- Each class must have at least three donors.
- Number of donors must exceed number of recipients in the class.

Imputation for Person and Household Nonresponse

Whole Record Imputation:

- Used when item imputation is insufficient, or no survey data is available for a person/household.
- Previous month's data (if available) is combined with current data to impute missing values.

Imputation Classes:

- Defined using variables such as:
  - Province, labour force status, occupation, age group, sex, education, etc.

Constraints and Adjustments:

- If donor pools are insufficient, imputation classes are collapsed by removing the least important variable.
- Donors are selected based on validity and consistency rules.

Post-Imputation Processing in LFS

Final Steps:

- Duplicates of donor records are removed.
- Derived variables are recalculated (e.g., weekly/hourly earnings, labour force status).
- Flags are set to indicate that imputation has occurred.

Benefits:

- Ensures the completeness of the dataset.
- Minimizes the impact of nonresponse on published statistics.

Evaluation of Imputation

Criteria for Evaluation:

- **Bias Control:** Ensure imputed values do not introduce bias.
- **Variance Preservation:** Maintain data variability.
- **Consistency:** Ensure logical relationships between variables are preserved.

Validation:

- Regular simulation studies.
- Historical comparisons to validate accuracy.

Conclusion

Summary of Key Points:

- The LFS provides essential labour market data but faces challenges due to unit and item nonresponse.
- Nonresponse introduces bias and increases variance, necessitating methods like weighting and imputation.
- Imputation techniques, such as hot-deck and whole record imputation, ensure data completeness and accuracy.
- Post-imputation processing and evaluation steps are crucial to validate the quality of imputed data.