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

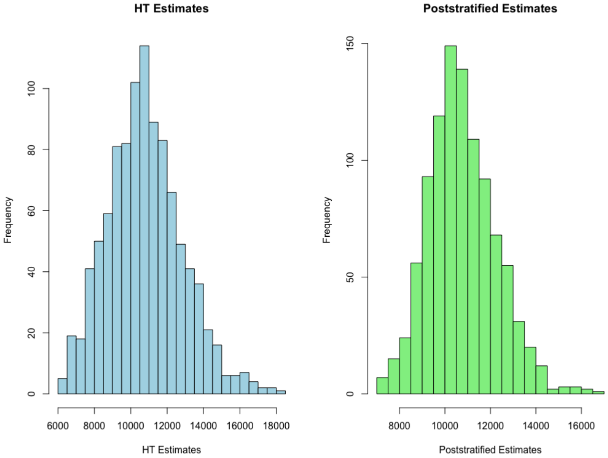


Figure: Histogram of the 1000 samples : HT vs Poststratified

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:

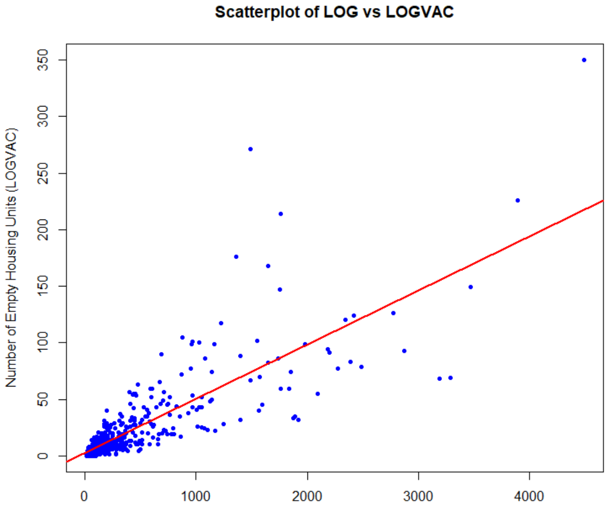
$$\hat{Y}_R = R \cdot X$$

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

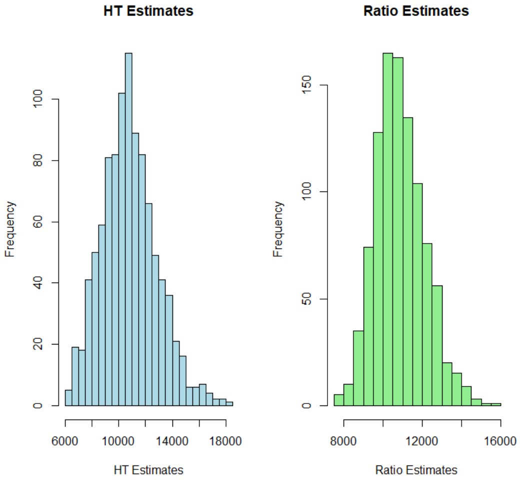


Figure: Histogram of the 1000 samples : HT vs Ratio

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

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

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

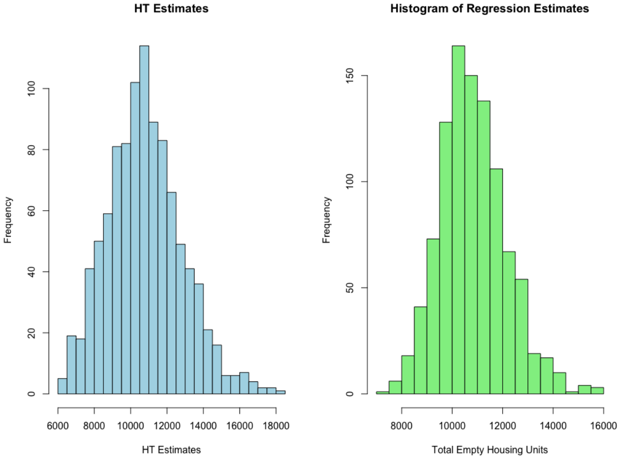


Figure: Histogram of the 1000 samples : HT vs Regression

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: Methods for nonresponse treatment, theory 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 occurs when data is not collected from all sampled units.
- Response rates have steadily declined over the years.
- A significant challenge for large-scale surveys like the LFS.
- In 2015, about 12% of sampled households did not respond to the LFS questionnaire each month.

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.

Types of Nonresponse

1. **Unit Nonresponse:** entire sampled units fail to respond.
2. **Item Nonresponse:** some questions remain unanswered.

Nonresponse in LFS:

- **Item nonresponse:** Missing data for specific items, e.g., income or employment status.
- **Unit nonresponse:**
    - **Household nonresponse:** Happens when an entire household fails to respond to the survey, resulting in no data collected for that unit.
    - **Person nonresponse:** Occurs when an individual within a sampled household does not provide any data, even though other household members may have responded.

Nonresponse Mechanisms

1. **MCAR (Missing Completely at Random):**
   - The probability of absence is the same for all observations.
   - Example: A survey is lost in the mail.
2. **MAR (Missing at Random):**
   - The probability absence is linked to one or more observed variables, the missing data are say missing data randomly.
   - Example: Younger respondents are less likely to report income.
3. **MNAR (Missing Not at Random):**
   - The probability of the absence of a variable depends on the variable itself or other variables not observed.
   - Example: High earners avoid reporting income.

Methods to Handle Nonresponse

Two Main Approaches:

- **Imputation**
- **Weighting**

Key Difference: Imputation fills in missing data, while weighting adjusts the influence of responding units to correct for nonresponse bias.

Imputation for Item Nonresponse

Definition: Imputation replaces missing responses to specific survey items with plausible values to ensure dataset completeness.

Key Features:

- Addresses **item nonresponse** (e.g., missing income) and sometimes **unit nonresponse**.
- Uses observed data to generate plausible values, preserving representativeness.

Why Impute?

- Imputation creates a **complete data file**, enabling full analysis.
- Unlike weighting adjustments, imputation allows the use of a **single sampling weight**.
- Ensures consistent results across analysts performing **identical analyses**.
- Facilitates the application of complete data estimation methods for point estimates (but not variance estimates).

Warnings

- Imputed data are **artificial** and may give a false impression of accuracy.
- Imputation can distort relationships between variables.
- Treating imputed values as observed can lead to **underestimation of variance**, especially with high nonresponse rates.

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

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

Concept: A "recipient" is matched with a "donor" based on specific characteristics or variables. The donor's value is then used to impute the recipient's missing value.

Theory:

- **Assumption:** Respondents with similar characteristics have similar values for the variable of interest.
- **Matching:** The process involves selecting matching variables (e.g., demographics, geography) to identify suitable donors.

Hot-Deck Imputation (2)

Advantages:

- Preserves the distribution and relationships within the dataset.
- Simple to implement and uses existing survey data (no external sources needed).

Challenges:

- Quality of imputation depends on the selection of matching variables.
- May introduce bias if donor pool is not representative.

Imputation pre-processing in LFS

Pre-Processing for Hot-Deck Imputation in LFS:

- Records are divided into:
    - **Group A:** Valid and consistent donors.
    - **Group B:** Valid but inconsistent, not used as donors.
    - **Group C:** Recipients requiring imputation.
- Derive imputation matching variables.
- Identify outlier earnings and finalize Groups A, B, and C.

Imputation for Item Nonresponse

Procedure:

- Each imputation class is defined by crossing 18 categorical variables such as:
- Random hot-deck imputation within classes is used to fill-in missing values
- In a given imputation class, each recipient is imputed by selecting a series of donors using SRSWOR

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.

Constraints: same as before

Next steps: The remaining nonrespondents households are treated by adjusting the design weights of responding households,

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 that response probabilities can be parametrically modeled that is: $p_i = f(z_i, \gamma)$

Key Formulas:

- Adjusted weight for respondent i:

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

- example: 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

- In the Labour Force Survey (LFS), response probabilities are estimated using a **uniform nonresponse model** within predefined classes.

- Each nonresponse class c is assumed to have a constant response probability p_c, which is estimated as:

$$\hat{p}_c = \frac{\text{design-weighted sum of } \textbf{responding households} \text{ in class } c}{\text{design-weighted sum of } \textbf{all households} \text{ in class } c}.$$

Nonresponse Classes in LFS

Purpose: Reduce bias by grouping households with similar response probabilities.

Key Features:

- Classes assume constant response probabilities and are homogeneous with respect to main variables of interest.
- Separate classes for Aboriginal or high-income strata.
- Other classes defined by crossing socio-demographic variables

Outcome: Ensures reliable adjustments, minimizing bias and variability.

Nonresponse Adjustment Factor:

- The nonresponse adjustment factor for households in class c is calculated as the inverse of the estimated response probability:

$$a_{cl}^{NA} = \frac{1}{\hat{p}_c}.$$

- This adjustment factor is applied to the weights of responding households to account for nonresponse.

Conclusion

- Nonresponse in surveys poses significant challenges, introducing bias and increasing variance.
- Effective nonresponse treatment methods include:
  - **Imputation:** Addresses item and unit nonresponse by filling in missing data with plausible values.
  - **Weighting:** Adjusts survey weights to correct for unit nonresponse and ensure representativeness.
- The Labour Force Survey (LFS) demonstrates the application of these methods, leveraging auxiliary information and nonresponse classes for robust adjustments.

Key Takeaway: Combining imputation and weighting effectively reduces bias and improves the quality of survey estimates, ensuring the reliability of results for policymaking and analysis.