Response Burden Reduction Through the Use of Administrative Data and Robust Sampling

Maria Caterina Bramati

Abstract

There are several reasons why robust regression techniques are useful tools in sampling design. First of all, when stratified samples are considered, one needs to deal with three main issues: the sample size, the strata bounds determination and the sample allocation in the strata. Since the target variable y, objective of the survey, is unknown, it is used some auxiliary information x known for the entire population from which the sample is drawn. Such information is helpful as it is strongly correlated with the target y, but of course some discrepancies between them may arise. The use of auxiliary information, combined with the choice of the appropriate statistical model to estimate the relationship with the variable of interest y, is crucial for the determination of the strata bounds, the size of the sample and the sampling rates according to a chosen precision level of the estimates, as it has been shown by Rivest (2002). Nevertheless, this regression-based approach is highly sensitive to the presence of contaminated data. Indeed, the influence of outlying observations in both y and x has an explosive impact on the variances with the effect of strong departures from the optimum sample allocation. Therefore, we expect increasing sample sizes in the strata, wrong allocation of sampling units in the strata and some errors in the strata bounds determination. Since the key tool for stratified sampling is the measure of scale of y conditional to the knowledge of some auxiliary x, a robust approach based on S-estimator of regression is proposed in this paper. The aim is to allow for robust sample size and strata bounds determination, together with the optimal sample allocation. To show the advantages of the proposed method, an empirical illustration is provided for Belgian business surveys in the sector of Construction.

M.C. Bramati (✉)
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza, University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy
e-mail: mariacaterina.bramati@uniroma1.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7__12,

It is considered a skewed population framework, which is typical for businesses, with a stratified design with one *take-all* stratum and $L - 1$ strata. Simulation results are also provided.

1 Introduction

The presence of outliers can strongly bias the sampling design and hence the survey results. In particular, it could induce a wrong computation of the number of statistical units to sample, usually overestimating it.

In what follows we focus on the stratified sampling design, which has been proven to be the most efficient surveying technique under some basic assumptions (see Tillé 2001) and it is currently in use at several NSIs for business surveys.

For instance, suppose that in the stratification variable X some outliers arise. Outliers are observations arbitrarily far from the majority of the data. They are often due to mistakes, like editing, measurement and observational errors. Intuitively, when outliers are present in a given stratum for the stratification variable X they affect both the location and scale measures for X. Therefore, it is clear that a higher dispersion than the "true" one will be observed in that stratum.

Such a situation will bias the outcome of the HL method. For instance, the sample size would be bigger than it should be, given the fact that observations seem to be more distant (in average) than they are in the reality. Moreover, the strata bounds and the sample allocation would be both biased. This is clear when we consider the Neyman allocation, for example, which is based on within-stratum dispersion. Since the principle is to survey more units in the strata in which the auxiliary variable is more dispersed within the stratum, outliers might have the effect of increasing enormously and unduly the sample size in each stratum.

For this reason we build two robust versions of the HL method, the *naive robust* and the *robust* HL sampling strategy which we compare through a simulation study.

2 The Problem

We focus on simple stratified samples with one take-all stratum and several take-some strata. This because we deal with

- skewed distributions (small number of units accounts for a large share of the study variables)
- availability of administrative information, providing a list of the statistical units of the target population (i.e. tax declaration, social security registers)
- survey burdens for firms and costs for NSIs
- data quality (administrative sources and survey collection)
- compliance requirements established by EUROSTAT

Now, it is known that there exists a discrepancy between the auxiliary variable X used for stratification and the survey variable Y. Therefore, the strategy suggested by Rivest (2002) is to recover such discrepancy by the use of a regression model.

Of course, the auxiliary information X is only a proxy for the target variable Y, which requires to estimate the *discrepancy* between Y and X, as suggested by Rivest (2002) with the *modified* HL algorithm.

In the business survey literature, the relationship existing between Y and X is often modeled by a log-linear regression relationship. Let X and Y be continuous random variables and $f(x)$, $x \in \mathbb{R}$ the density of X. The data $x_1, \dots, x_N$ are considered as N independent realizations of the random variable X.

Since stratum h consists of the population units with an X-value in the interval $(b_{h-1}, b_h]$, the stratification process uses the values of $E(Y|b_h \geq X > b_{h-1})$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h, for $h = 1, \dots, L-1$.

This model considers the regression relationship between Y and X expressed by

$$\log Y = \alpha + \beta_{\log} \log X + \varepsilon,$$

where ε is assumed to be a 0-mean random variable, normally distributed with variance $\sigma^2_{\log}$ and independent from X, whereas α and $\beta_{\log}$ are the parameters to be estimated.

However this approach presents some weaknesses

1. s^2_{yh} is unknown, which makes crucial the use of the auxiliary information X
2. the number L of strata is selected by the user
3. the administrative records are often of low quality (errors)

We can distinguish three main sources of anomalies, listed below

- erroneous records in the surveyed data (Y) (**vertical outliers**)
- quality issues in the administrative registers (X) (**leverage**)
- outliers in both variables (X, Y) (good/bad **leverages**)

The presence of such anomalies makes unreliable the conditional mean and variance of $Y|X$, therefore affecting the sample size and strata bounds determination as well as the sample allocation.

In what follows we propose a possible alternatives to the Rivest (2002) modified HL algorithm. Strata bounds and sizes are derived minimizing the conditional variance in each stratum after a re-weighting of the information according to the degree of *outlyingness*. We refer to this approach as to the *robust modified HL* algorithm.

3 The Robust Modified HL Algorithm

Supposing that a log-linear relationship exists between the survey variable Y and the auxiliary one X, then consider the S-estimator of regression as in Rousseeuw and Yohai (1984) as

$$S(x, y) = \arg\min_{\beta} s(r_1(\beta), \dots, r_N(\beta))$$

where $r_i(\beta)$ are the regressions residuals and s is scale measure which solves

$$\frac{1}{N}\sum_{i=1}^{N}\rho\left(\frac{r_i(\beta)}{s}\right) = b$$

for a conveniently chosen ρ function and a constant b. This estimator is robust with respect to both vertical outliers and leverage points. Then, with some straightforward calculations (expanding $\rho(\cdot)$), the following approximation holds

$$\mathrm{Var}[Y|b_h \geq X > b_{h-1}] \approx e^{\sigma^2}\psi_h/W_h - (\phi_h/W_h)^2,$$

where

$$W_h = \int_{b_{h-1}}^{b_h} \omega(x^\beta)f(x)dx \tag{1}$$

$$\phi_h = \int_{b_{h-1}}^{b_h} x^\beta\omega(x^\beta)f(x)dx \tag{2}$$

$$\psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta}\omega(x^\beta)f(x)dx, \tag{3}$$

β and σ are the parameters of the log-linear model in the previous section, and $\omega(x) = \rho'(x)/x$ is the weighting function.

The problem then reduces to solving for bounds $b_1, \ldots, b_h, \ldots, b_L$ which minimize n using the Neyman allocation scheme. In symbols, under the loglinear specification the objective function is

$$n_{\hat{t}_{\mathrm{ystrat}}} = N_L + \frac{\left(\sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}\right)^2}{(c\sum x_i^\beta/N)^2 + \sum_{h=1}^{L-1}\frac{(e^{\sigma^2}\psi_h - \phi_h^2/W_h)}{N}} \tag{4}$$

where *robust* moments W_h, ϕ_h and ψ_h are those in (3), β and σ are the parameters of the log-linear model estimated by robust regression (S-estimator or LTS).

Then, the Sethi's iterations are run for a given L and precision c, computing the optimal strata bounds and sample size.

4 Simulation Study

The aim of the simulation study is to compare the performance of the two robust sampling strategies proposed in this paper with respect to Rivest (2002)'s based on classical LS regression.

Table 1 Summary of results comparing Robust modified HL method versus modified HL (Rivest 2002), target precision: 1 %

Design	Relative efficiency	Relative sample size
No outliers	0.10	100
Long-tailed Cauchy	0.00	0.29
Long-tailed *t*	0.08	10
Vertical outliers 15 %	0.99	10
Leverage points 15 %	0.00	10
Vertical outliers 30 %	0.99	1.43
Leverage points 30 %	0.00	1.43

Simulations are performed using the business sampling frame of the Structural Business Survey in 2002, where we consider as target variable (y) the value added of enterprises in the industry of *Constructions* which are stratified by the economic-size class. The number strata $h = 1, \ldots, 6$ is set according to the common practice in SBS, with 1 take-all stratum and 5 take-some strata. The auxiliary information x is on the turnover (from the VAT register). Then, population is generated from

$$\log y_i = \beta \log x_i + \varepsilon_i$$

with a choice of $\beta = 0.75$.

Then we consider the following designs

1. no outliers: $\varepsilon_i \sim \mathcal{N}(0, 1)$
2. long-tailed errors: $\varepsilon_i \sim \text{Cauchy}_1$
3. long-tailed errors: $\varepsilon_i \sim \text{t}_3$
4. vertical outliers: $\delta\%$ of $\varepsilon_i \sim \mathcal{N}\left(5\sqrt{\chi^2_{1;0.99}}, 1.5\right)$
5. bad leverage points: $\delta\%$ of $\varepsilon_i \sim \mathcal{N}(10, 10)$ and corresponding $X \sim \mathcal{N}(-10, 10)$.

The contamination level, i.e. the percentage of outliers in the data, is set to $\delta = 15$ and 30 %. Then the three procedures are used to compute the strata bounds, sizes and allocation

- generalized HL method (Rivest 2002)
- robust generalized HL method

at 1 % precision and compared by means of relative MSE of the Horvitz–Thompson estimator for the mean. In Table 1 are displayed the main results.

References

Rivest, L.P.: A generalization of Lavallée and Hidiroglou algorithm for stratification in business surveys. Techniques d'enquêtes **28**, 207–214 (2002)

Rousseeuw, P.J., Yohai, V.J.: Robust regression by means of S-estimators. In: Franke, J., Hardle, W., Martin Robust, D. (eds.) Nonlinear Time Series. Lecture Notes in Statistics. vol. 26, pp. 256–272. Springer, Berlin (1984)

Tillé, Y.: Théorie des sondages. Dunod, Paris (2001)