Stratified Sampling for time dependent survey variables

Maria Caterina Bramati
Sapienza University of Rome, e-mail: mariacaterina.bramati@uniroma1.it

Abstract

The temporal dimension is often neglected in survey designs since, in general, samples are drawn independently across statistical units and across time. When the information on the target variable Y is collected during a time interval, then it is possible to study the temporal pattern of the variable itself in order to improve the sampling design. Ideally, not only the dispersion across population units is taken into account, but also the variability through time and the ability of forecasting future values of the target variable conditional to the values observed in the past. Nowadays, NSIs make more extensive use of administrative data, like the VAT registry, which contain also records concerning the past. This historical dimension allows to account for the variability through time of some auxiliary information as proxies for the target variable one.

Keywords: Temporal variability, sample, stratification

1. Introduction

The time dimension is generally not accounted for in classical sampling theory, since the working hypothesis is independent samples across units and time.

However, there are surveys for which target variables are characterized by a strong time dependence. The typical examples are monthly surveys for short terms statistics (STS), generally prices, which use a sample of firms interviewed on a regular basis throughout the year. This scheme corresponds to a panel of statistical units which is fixed during a time interval, generally a year, and might vary in the next time-frame for a few firms which are dropped and replaced, whilst the majority of the sample (usually consisting in large firms) does not change.

In the common practice of NSIs, the sampling scheme is not reviewed so often (at least not on a monthly basis) also for a matter of coherence and comparability of survey variables. Therefore, it is often possible to construct series of observations of a target variable Y for each of the n sampling units i=1,…,n in the time periods t=1,…, T.

In the classical approach to optimal sampling in the Neyman sense, the sample is allocated proportionally to the cross-sectional dispersion of the target variable, which means that the higher the standard deviation of the variable, the larger the sample size is. When the stratification is based on variables such as the economic size of firms, this implies that large firms are constantly surveyed all months through the years, with consequent heavy response burdens.

When the information on the target variable Y is collected during a time interval, then it is possible to study the temporal pattern of the variable itself in order to improve the sampling design. Ideally, not only the dispersion across population units is taken into account, but also the variability through time and the ability of forecasting future values of the target variable conditional to the values observed in the past. Intuitively, the higher the variance in time of the target variable is, the greater the sample size needed for a fixed precision level of estimates. On the other hand, the higher the correlation (positive or negative) of present values with past values of Y is, the higher the ability to forecast the future and the smaller the sample size needed at a given precision of estimates. In particular, when dealing with stratified designs based on prices, price variation *across units* alone might not be able to capture the proper variance structure of the target variable, which is known to be autocorrelated, and therefore better described *across the time dimension*. This approach to stratification and to sample allocation clearly has the advantage to reduce sample sizes since firms with almost zero price variations in time are less likely to be regularly surveyed.

The framework we consider in what follows is the stratified sample where we focus on a general **temporal stratification** algorithm with optimal strata bounds and sample sizes determination à la Hidiroglou-Lavallée when discrepancies between the target and the auxiliary information are captured by a regression model with first order autocorrelated errors.

2. Sampling design in time

In what follows the working hypothesis is a dependency pattern through time of the population Y, which is related to some auxiliary information X through a linear regression equation

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

where i=1,…,N is the population unit index. The error ε_i is assumed to be a white noise with constant variance σ_ε^2.

In what follows we assume that the statistical units are not cross-correlated, i.e. are independent, but have a specific autocorrelation structure ρ which varies across units i=1,…,N.

In other words, the error is autocorrelated of order 1, i.e.

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$$

With where $|\rho|$ <1, and η_t i.i.d. white noise. Then, the variance of the process is given by $\gamma(0) = \frac{\sigma_\varepsilon^2}{1-\rho^2}$ and autocovariances $\gamma(k) = \rho^k\gamma(0)$, k=1,…

This implies that y_t follows an AR(1) since

$$y_t - \rho y_{t-1} = \beta\,(x_t - \rho x_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1.}$$

Therefore

$$y_t = \rho y_{t-1} + \beta\,(x_t - \rho x_{t-1}) + \eta_t$$

If X does not vary in time, then the regression coefficient of the auxiliary variable becomes $\beta\,(1-\rho)$. However, we assume that the auxiliary variable X which is used to stratify the population, has the same autocorrelation structure as the target variable.
Indeed, if X is a good proxy for the target variable Y, the autocorrelation structure should be the same, therefore ρ could be estimated using the auxiliary information X.

Then the strategy should follows the generalized HL algorithm, but with the appropriate changes in the estimates of the regression model. The logic is: stratify considering homogenous strata in terms of variance (cross-section) and similar autocorrelation, with the attention that higher autocorrelation should lead to survey less, because of persistency. On the opposite, when units show a low persistence, then the variability is more due to the random error and therefore need to be sampled more often.

When the linear model is assumed in a stratified design with Neyman allocation, it is shown that

$$n = nW_L + \frac{\left(\sum_{h=1}^{L-1} W_h s_h\right)}{\dfrac{yc^2}{(1-\rho^2)} + 1/N \sum_{h=1}^{L-1} W_h s_h^2}$$

the optimal sample size for the estimator of the total is proportional to the stratum variance s_h and to the persistence ρ of the AR process. If ρ is close to 1, which means long memory, then the sample size should decrease because the dispersion in the stratum does not vary much from the previous period. On the opposite, for ρ is close to 0 there is not persistence and the sample size boils down to the classical form, as Hidiroglou and Lavallée (1988) shows.

As business surveys deal with skewed populations, strata h=1,…,H are built grouping according to the strength of autocorrelation through time, leading to 3 strata: a take-all grouping the units with low autocorrelation ($|\rho| \approx 0$, variability mainly due to the white noise), a take-0 grouping the units strongly persistent (high autocorrelation $|\rho| \approx 1$), and those with a medium autocorrelation level which are placed in the take-some stratum. When stratification is performed on skewed populations, the loglinear regression model results a more suitable formulation of the relationship between the target and the auxiliary information, as pointed out by Rivest (2002). The nonlinearities implied by the model require more efforts on the computation of the algorithm which is part of the agenda for further research studies.

3. Conclusions

The focus of this research is to propose a stratification algorithm for surveys repeated in time where the target variable shows correlation in time stronger than across individuals. The resulting procedure suggests larger samples when the variability across units is predominant on the autocorrelation, whereas when the process is characterized by strong persistence in time, then it is not necessary to sample the same units and therefore the

procedure suggests smaller samples. This intuitive result is shown using a linear regression model, but some extensions to other models are possible.

References

Lavallée, P., Hidiroglou, M. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.

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