

# Generalized Variance Function: Theory and Empirics

## *La Funzione Generalizzata della Varianza Campionaria: Teoria e Pratica*

A. Pavone

ISTAT – Via Magenta 2, 00184 Roma, e-mail: pavone@istat.it

Aldo Russo

Università di Roma Tre – Via C. Segre 2, 00145 Roma, e-mail: aldoruss@uniroma3.it

**Riassunto:** Per esprimere il grado di precisione di una stima predetta da un'indagine è indispensabile valutare il corrispettivo errore standard. Un modo semplice per predire gli errori standard è di ricorrere alle funzioni generalizzate della varianza campionaria. Tuttavia, nel caso di stime costituite da totali di variabili quantitative, questi modelli non sono molto evoluti.

Il presente lavoro ha come obiettivo di colmare alcune di queste lacune, per quanto riguarda la rilevazione sui conti economici delle imprese.

**Keywords:** Variance Approximation, Reliability of Estimates, Business Surveys.

## 1. Introduction

Concise presentation of sampling errors in survey reports is a challenge when computing and printing standard errors for each estimate could be costly of time-consuming. Estimates of sampling error indirectly is one solution.

One method uses *generalized variance function* (GVF) models which are based on estimators of sampling variability. As a rule, they are obtained in two steps: (i) for each domain of interest, the survey estimates and some of corresponding variances are directly obtained, and then (ii) the relationship between estimates with their sampling errors are modelled.

The principal advantage of this approach is quick estimates of sampling error for any user-specified variables. In addition GVF leads to more stable and precise variance estimates than direct ones (Valliant 1987). This contribution attempts to extend the GVF forms in the case of economic-estimated totals for business surveys. Furthermore, we present empirical results to predict sampling variances for the account estimates. The applied work has showed many of domains with a small sampling error, compared to other domains with same total level of survey characteristics. We consider ways in which to address this problem based on an appropriate treatment of the design effect.

## 2. A methodology overview

Let  $\hat{Y}$  denote an estimator of the population total  $Y$  and with  $\varepsilon(\hat{Y})$  the relative standard error of the  $\hat{Y}$

$$\varepsilon(\hat{Y}) = \frac{\sqrt{Var(\hat{Y})}}{Y} \quad (1)$$

where  $Var(\hat{Y})$  is the variance of  $\hat{Y}$ .

Graphical inspection shows that the relative standard error is a decreasing function of the magnitude of  $\hat{Y}$ . A shape that exhibits this propriety is

$$\varepsilon(\hat{Y}) = \beta_1 \hat{Y}^{-\beta_2} \exp^e \quad (2)$$

where  $e$  is a residual term serially uncorrelated, with zero mean and constant variance. The parameters  $\beta_1, \beta_2$  are unknown and have to be estimated. They depend by the population, the sampling design, the estimator, and the  $Y$ .

There is a nonlinear model in the parameters and must make some suitable transformation. Convert this nonlinear regression equation into a linear one by taking the logarithm of both sides, so that

$$\ln(\varepsilon(\hat{Y})) = \ln(\beta_1) - \beta_2 \ln(\hat{Y}) + e \quad (3)$$

An attractive aspect of this log-linear model is that the slope coefficient  $\beta_2$  measures the elasticity of  $\varepsilon(\hat{Y})$  with respect to  $\hat{Y}$ , that is, the percentage change in relative standard error for a given percentage change in the expectation of the  $\hat{Y}$ .

This model can be reformulated in term of standard error: the parameters of this equivalent model are in biunivocal correspondence with the original model. If we break up the relative standard error into two components, the model (3) can be written in the following way

$$\ln(SE(\hat{Y})) = \ln(\beta_1) + (1 - \beta_2) \ln(\hat{Y}) + e \quad (4)$$

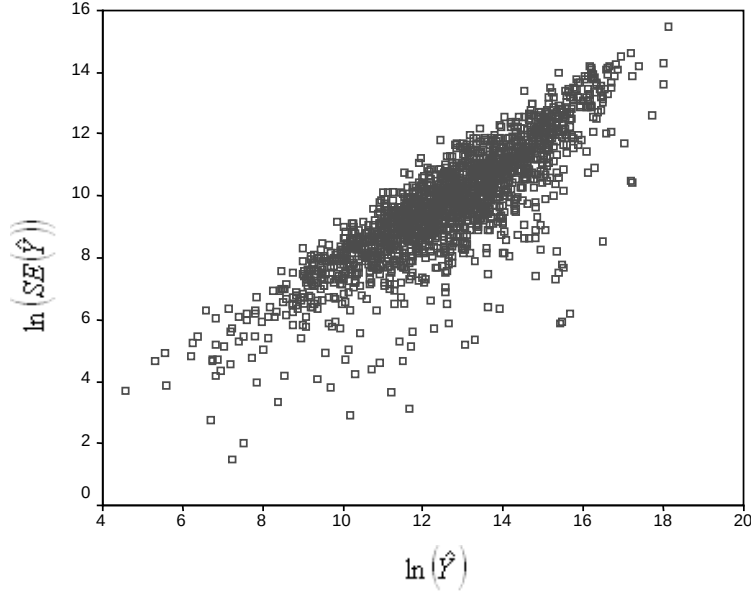
where  $SE(\hat{Y}) = \sqrt{Var(\hat{Y})}$ .

Keep in mind that, if the design effects (that is defined as the ratio of the sampling variance of the statistic under the actual sampling design divided by the variance that would be expected for a simple random sample of the same size) vary greatly among domains, this general GVF prediction model is not exhaustive. In this case integrate the right-hand side of the GVF model with a proxy for the design effects (*deff*).

### 3. Empirical application

Here we describe how we approximate standard error using sampling variances and totals directly estimated to 462 ATECO domains for a following set of 4 economic variables: *Annual turnover*; *Value added*; *Labour cost*; *Investments*. Data that refer to year 1999 were collected by the National Statistics Institute of Italy, merging the exhaustive survey SCI (on large enterprises  $\geq 100$  employees) with the sample survey PMI (on small and medium enterprises, up to 99 employees) to overlapping domains.

To get a first approximation of the graphical evaluation of potential bias in model prediction we plotted  $\ln(SE(\hat{Y}))$  versus the  $\ln(\hat{Y})$  (figure 1). The scatter plot inspection shows that a linear relationship does exist.



**Figure 1:** *Scatter plot*

Nevertheless emerges the presence of several domain observations that exhibit small sampling error compared to other domains with same total level of survey characteristics. The reasons for this anomaly depends on the different types of surveys' designs. In fact, for each overlapping domain, both total and variance is estimated combining the PMI sample characteristics with SCI census ones. This produces high variability in the *deff*.

As Wolter (1985) suggests, the specification of GVF model is consistent with a approximately constant *deff*, for a group of variables. This assumption is violated in the examined data; indeed, the logarithm of the *deff* ranges from -9,39 e 1,70.

To evaluate how much the *deff* impacts on the variance prediction the right-hand side of the GVF model (4) it is integrated for it

$$\ln(SE(\hat{Y})) = c + \sum_{j=2}^5 a_j \lambda_j + b_1 \ln(\hat{Y}) + b_2 \ln(deff) + e \quad (5)$$

where:  $\ln(\beta_1) = c$ ;  $(1 - \beta_2) = b_1$ ;  $\lambda_j$  represent one-zero indicator of the last four economic variables and the reference category is enclosed in the constant  $c$ .

Table shows the results of an ordinary least squares fit of the model in (5) above.

Note that introducing the  $\ln(deff)$  in the estimated model leads to a faultless fit. However to be realistic, a user has no idea of the magnitude of the *deff* for each domain. Without this important information the user cannot insert the survey-derived total accounts into the fitted GVF model to generate a model-based prediction of the variance.

To bypass this problem consists select homogenous partitions of domains, identified in terms of their level on the *deff*. These partitions are obtained by using the CHAID technique, acronym for Chi-squared Automatic Interaction Detector. Setting to 100 a minimum number of domains, for partition, the CHAID algorithm generated 11 groups (domain).

The alternative fitted model is

$$\ln(SE(\hat{Y})) = c + \sum_{j=2}^5 a_j \lambda_j + \sum_{i=2}^{11} q_i \xi_i + b_1 \ln(\hat{Y}) + e \quad (6)$$

where the variables  $\xi_i$  are indicators that the item falls into the last 11 partitions above indicated. Note that each variable indicator corresponds to a specific ATECO group domain effects and could be easily included in the survey report.

The adjusted  $R^2$ , fitting on the GVF model (6), shows a decrease of less than 4% .

**Table 1:** *Least square linear regression for model (5) and (6)*

| Predictor Variables | Coefficients | Std. error | Student's T | Adjusted $R^2$ | Predictor Variables | Coefficients | Std. error | Student's T | Adjusted $R^2$ |
|---------------------|--------------|------------|-------------|----------------|---------------------|--------------|------------|-------------|----------------|
| $c$                 | 0,047        | 0,080      | 0,585       |                | $c$                 | -1,947       | 0,124      | -15,691     |                |
| $\ln(\hat{Y})$      | 0,878        | 0,006      | 158,446     | 0,715          | $\ln(\hat{Y})$      | 0,785        | 0,008      | 100,108     | 0,716          |
| $\ln(deff)$         | 0,771        | 0,008      | 98,799      | 0,956          | $\xi_{11}$          | 3,238        | 0,066      | 49,054      | 0,741          |
| $\lambda_3$         | -0,327       | 0,028      | -11,747     | 0,958          | $\xi_{10}$          | 2,851        | 0,056      | 50,816      | 0,770          |
| $\lambda_2$         | -0,213       | 0,026      | -8,151      | 0,959          | $\lambda_3$         | -0,760       | 0,039      | -19,674     | 0,790          |
| $\lambda_4$         | -0,129       | 0,030      | -4,228      | 0,960          | $\xi_9$             | 2,614        | 0,061      | 42,837      | 0,805          |
|                     |              |            |             |                | $\xi_8$             | 2,468        | 0,061      | 40,688      | 0,818          |
|                     |              |            |             |                | $\xi_7$             | 2,333        | 0,058      | 40,018      | 0,832          |
|                     |              |            |             |                | $\xi_6$             | 2,219        | 0,058      | 38,347      | 0,847          |
|                     |              |            |             |                | $\xi_5$             | 2,091        | 0,059      | 35,564      | 0,863          |
|                     |              |            |             |                | $\xi_4$             | 1,913        | 0,062      | 31,016      | 0,877          |
|                     |              |            |             |                | $\xi_3$             | 1,679        | 0,060      | 28,055      | 0,893          |
|                     |              |            |             |                | $\xi_2$             | 1,322        | 0,059      | 22,384      | 0,915          |
|                     |              |            |             |                | $\lambda_2$         | -0,404       | 0,037      | -11,050     | 0,922          |
|                     |              |            |             |                | $\lambda_4$         | -0,052       | 0,043      | -1,211      | 0,922          |

## References

- Kish, L. (1965) *Survey Sampling*. New York: John Wiley.
- Russo A. (1987) *Sulla Presentazione degli Errori di Campionamento mediante Modelli: Il Metodo dei Modelli Regressivi*, Quaderni di Discussione, n° 4, Roma, Istat.
- Valliant, R. (1987) "Generalized Variance Functions in Stratified Two-Stage Sampling", *Journal of the American Statistical Association*, 82, 499-508.
- Wolter, K. M. (1985) *Introduction to Variance Estimation*. New York, Springer-Verlag.