

On a Mixed Model approach to Regression Estimation¹

Su un Approccio Mixed Model nella Stima per Regressione

Giorgio E. Montanari - M. Giovanna Ranalli

Dipartimento di Scienze Statistiche, Università degli Studi di Perugia

giorgio@stat.unipg.it - giovanna@stat.unipg.it

Riassunto: in questo lavoro si propone l'impiego di modelli lineari a componenti di varianza per la costruzione di stimatori per regressione di parametri descrittivi di una popolazione finita. La metodologia proposta consente di conciliare la maggiore efficienza asintotica dello stimatore ottimo con la maggiore stabilità dello stimatore per regressione generalizzata.

Keywords: Finite Population, Auxiliary Information, Design based approach.

1. Introduction

Sample surveys are commonly conducted nowadays to gather information in the most diverse fields of interest. Auxiliary information is also usually available to increase the precision of the estimation procedure and the task for statisticians is to properly incorporate such information by means of more careful modeling. This has often been coupled with the need of keeping *design-based* properties of the resulting estimation procedures. Properties such as design consistency are usually sought as a protection against model misspecification.

Generalized regression estimators (GREG; Särndal *et al.*, 1992) have been proposed within a *model-assisted* approach to inference and gained much popularity. Also the Optimal estimator (OPT; Montanari, 1987) firstly introduced as a complete design-based estimator has been shown to implicitly rely on some linear model which accounts for the sampling plan (Montanari and Ranalli, 2002). In this work we explore the possibility of employing estimators based on Linear mixed models (LMM; Searle *et al.*, 1992) as a compromise between the two and a tool to better understand their behavior.

To look at this in more detail we consider a finite population $U = \{1, \dots, N\}$. Let y_i and $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{qi})$ be the values of the survey variable y and of a q -dimensional auxiliary variable $\mathbf{x}$ associated with unit i . The population mean vector $\bar{\mathbf{x}} = \sum_{i \in U} \mathbf{x}_i / N$ of $\mathbf{x}$ is assumed known, e.g. from administrative registers or census data. The unknown y population mean, $\bar{Y} = \sum_{i \in U} y_i / N$, has to be estimated by means of a sample s of size n drawn from U according to a sampling design with inclusion probabilities π_i and π_{ij} , with $i, j \in U$. Let $\hat{\bar{Y}} = N^{-1} \sum_{i \in s} y_i / \pi_i$ and $\hat{\bar{\mathbf{x}}} = N^{-1} \sum_{i \in s} \mathbf{x}_i / \pi_i$ be the Horvitz-Thompson estimators (HT) of $\bar{Y}$ and $\bar{\mathbf{x}}$, respectively. The estimators we consider in this work will take into account the knowledge of $\bar{\mathbf{x}}$ in the form of a regression estimator,

$$\hat{\bar{Y}}_{reg} = \hat{\bar{Y}} + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}}) \hat{\boldsymbol{\beta}}_{reg}, \quad (1)$$

¹ The present paper is financially supported by a grant COFIN2002 from MIUR, Italy.

with estimates of β sought by assuming different working models ξ for the relation between y and x and accounting for the sampling plan. In particular GREG is derived under a linear regression model for which $y = X\beta + \epsilon$, with $E_\xi(\epsilon) = \mathbf{0}$ and $V_\xi(\epsilon) = \sigma_\epsilon^2 \mathbf{I}$, where y is the N -vector of population values of y , X is the $N \times q$ -matrix of values of x with x_i on its i -th row and $\mathbf{I}$ is an $N \times N$ identity matrix (we won't consider here more complex variance structures only for ease of notation). Estimates of β are given by $\hat{\beta}_{\text{GREG}} = (X_s' \Pi_s^{-1} X_s)^{-1} X_s \Pi_s^{-1} y_s$, where subscript s denotes matrices with sample values and $\Pi_s = \text{diag}\{\pi_i\}_{i \in s}$. Estimates of β for OPT are given by $\hat{\beta}_{\text{OPT}} = (X_s' W_{ss} X_s)^{-1} X_s' W_{ss} y_s$, where $W_{ss} = \{(\pi_{ij} - \pi_i \pi_j) / \pi_i \pi_j \pi_{ij}\}_{i,j \in s}$. Although no explicit model is assumed for this estimator, it has been shown that this formulation relies on a regression model for which $y = [X \ D]\beta + \epsilon$, with $E_\xi(\epsilon) = \mathbf{0}$ and $V_\xi(\epsilon) = \Sigma$, where D is an $N \times k$ -matrix which assembles the values of a set d of *Design Balanced Variables* (DBV's) and Σ is a covariance matrix which is related to W to account for the sampling plan through first and second order inclusion probabilities. A DBV is any non null auxiliary variable whose mean is estimated without error by the HT estimator; for example, in simple random sampling any vector proportional to the unit vector $\mathbf{1}$ is a DBV, while for stratified random sampling the indicator variables of stratum membership of population units are a set of linearly independent DBV's (for more details on the covariance matrix Σ and other examples of DBV's see Montanari and Ranalli, 2002).

Therefore OPT automatically inserts in the working model information about the sampling design; this feature on one side allows a better fit of the data and explains the asymptotic superiority of this estimator over the GREG. On the other side, it provides an explanation of its instability for finite size samples, especially when an inadequate number of residual degrees of freedom is available for estimating model parameters. In particular, this concern may be relevant in stratified designs with a few observations per stratum. In the following Section an estimator of the population mean based on linear mixed models is introduced to take advantage of introduction of DBV's in the working model to increase the efficiency only when they are really needed.

2. A Mixed Model approach

An estimator which represents a compromise between GREG and OPT is the one that allows the coefficients of the DBV's in the working model to be random variables. LMM is based on a linear mixed model in which the fixed component is given by the auxiliary variables x and the random component by the DBV's d ; in particular

$$y = X\beta + Du + \epsilon, \quad (2)$$

where $E_\xi([u', \epsilon']') = [0', 0']'$ and $\text{Cov}_\xi([u', \epsilon']') = \text{blockdiag}\{\sigma_u^2 \mathbf{I}, \Sigma\}$. Introduction of random coefficients for D has the advantage of requiring the estimation of a single extra parameter σ_u^2 , commonly referred to as a *variance component* (Searle *et al.*, 1992).

To better understand the implications of such a model let us consider the case of a stratified random sample with proportional allocation of the units to the strata. As mentioned earlier, the indicator variables of stratum membership are DBV's for this sampling plan, while the covariance structure Σ simplifies to $\sigma_\epsilon^2 \mathbf{I}$ (Montanari and

Ranalli, 2002). In this case OPT fits a different intercept for each stratum, which would likely fit the data well in a large sample, but would also not be adequate for small stratum sample sizes. On the other hand, GREG does not account for stratum membership at all. The random component in model (2) allows the fitting of a random intercept for each stratum, thereby accounting for within-stratum correlation and, at the same time, for the randomness due to the stratification employed.

The variance component provides a measure of the covariance between the values of y in unit i and i' belonging to the same strata h , i.e. $Cov_{\xi}(y_{hi}, y_{hi'}) = \sigma_u^2$. The correlation coefficient is given by $Corr_{\xi}(y_{hi}, y_{hi'}) = \sigma_u^2 / (\sigma_{\epsilon}^2 + \sigma_u^2)$ and relative large values of $Corr_{\xi}(\cdot)$ indicate considerable within-stratum correlation and therefore that the DBV's should be considered in the analysis. Otherwise, relative small values of it indicate that the information coming from the stratification could be ignored. In fact, in the working model for GREG the assumption is that $\sigma_u^2 = 0$, which implies that $\mathbf{u} = \mathbf{0}$ and that all the random effects drop out. In the model for OPT, on the other hand, $\sigma_u^2 = \infty$ and $\mathbf{u}$ is no longer random but rather a vector of unknown fixed constants.

The estimator LMM can still be written as in (1) with estimates of β given by $[\hat{\beta}'_{LMM}, \hat{\mathbf{u}}'_{LMM}]' = ([\mathbf{X}_s' \mathbf{D}_s] \Pi_s^{-1} [\mathbf{X}_s' \mathbf{D}_s] + \hat{\sigma}_{\epsilon}^2 / \hat{\sigma}_{\mathbf{u}}^2 \mathbf{G})^{-1} [\mathbf{X}_s' \mathbf{D}_s] \Pi_s^{-1} \mathbf{y}_s$, with $\mathbf{G} = \text{diag}\{[\mathbf{0}'_p \mathbf{1}'_k]\}$. *Maximum Likelihood* and *Residual Maximum Likelihood* (REML) can be used to get the estimates $\hat{\sigma}_{\mathbf{u}}^2$ and $\hat{\sigma}_{\epsilon}^2$ (Searle *et al.*, 1992). Under assumptions similar to those in Breidt *et al.* (2003), LMM can be shown to be design consistent for $\bar{Y}$.

3. Simulation studies and final remarks

A small simulation study has been conducted to explore finite sample performance of GREG, OPT and LMM. A finite population of 1,200 units partitioned into $H=20$ strata of equal size was considered. The values of an auxiliary variable x were generated through a Chi-squared random variable with 8 degrees of freedom. They were randomly assigned to the strata, to simulate a stratification based on other characters independent of x . Given the values of x , six populations of y values were generated, according to the following models:

- | | |
|---|---|
| $P_1: y_{hi} = 100(1+0.15\epsilon_{hi})$ | (total independence between y , x and h); |
| $P_2: y_{hi} = (10+3h)(1+0.15\epsilon_{hi})$ | (dependence between y and h); |
| $P_3: y_{hi} = 3x_{hi} (1+0.15\epsilon_{hi})$ | (linear dependence between y and x); |
| $P_4: y_{hi} = (12+4h+5x_{hi})(1+0.15\epsilon_{hi})$ | (linear dependence between y , x and h); |
| $P_5: y_{hi} = x_{hi}^2 (1+0.15\epsilon_{hi})$ | (quadratic dependence between y and x); |
| $P_6: y_{hi} = 20h \sqrt{x_{hi}} (1+0.15\epsilon_{hi})$ | (non linear dependence between y , x and h); |

where y_{hi} and x_{hi} are the values of y and x for unit i ($i=1, \dots, 60$) in stratum h ($h = 1, \dots, 20$) and $\epsilon_{hi} \sim N(0,1)$, $\forall i, h$. For each population 1,000 proportional stratified random samples of size 40, 80 and 240 (two, four and twelve units per stratum, respectively) were selected and the estimators calculated. GREG fits a model with two parameters, while OPT fits twenty-one parameters – one intercept for each stratum and a common slope for x . For LMM, the effective number of degrees of freedom employed is given by $df_{LMM} = 2+(H-1) n_h / (n_h + \sigma_{\epsilon}^2 / \sigma_{\mathbf{u}}^2)$, where n_h is the sample size in stratum h . The function `lme()` in the S-PLUS package has been employed to fit the model for LMM.

The performance of each estimator has been assessed by means of $Smse(\hat{\bar{Y}}_{reg}) = M\hat{SE}(\hat{\bar{Y}}_{reg}) / M\hat{SE}(\hat{\bar{Y}}_{AOPT})$, where $\hat{\bar{Y}}_{AOPT}$ is the asymptotic approximation of OPT, i.e. $\hat{\bar{Y}}_{AOPT} = \bar{\bar{Y}} + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}})\boldsymbol{\beta}_{AOPT}$, where $\boldsymbol{\beta}_{AOPT} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ with $\mathbf{W} = \{(\pi_{ij} - \pi_i\pi_j)/\pi_i\pi_j\}_{i,j \in U}$ is the limit in probability of $\hat{\boldsymbol{\beta}}_{OPT}$. This estimator has the smallest design variance in the class of regression estimators and therefore is a gauge of the best we can expect from them. Table 1 reports the achieved values of $Smse(\cdot)$ of each estimator from populations $P_1, P_2, \dots, P_6$. The most noticeable feature is that LMM has a performance always almost as good as the most efficient between GREG and OPT; being a compromise between the two we can't expect it to do better than any of those, but for each population and each sample size it provides the best out of them. In this sense it can be considered as a data driven procedure. In particular, when accounting for the stratification is important (populations P_2, P_4 and P_6) LMM gets closer to OPT by estimating a larger value for σ_u^2 , while, in the cases for which the information provided by x is sufficient to explain the behavior of y , LMM gets closer to GREG by estimating small values for σ_u^2 . More detailed results concerning the variance components and the estimated coefficients for each estimator are available from the authors.

The use of mixed models in a model assisted framework allows a better understanding of the role of the sampling design in the estimation procedure. Another useful application of LMM currently under study by the authors is in the field of calibration estimation. In fact, when a large amount of calibration yields extremely erratic sampling weights, LMM allows relaxing the calibration constraints for some variables by considering them in the random component of model (2).

Table 1: Scaled mean squared errors of estimators ($M\hat{SE}(\hat{\bar{Y}}_{AOPT})=1$).

Estimator n	HT			GREG			OPT			LMM		
	40	80	240	40	80	240	40	80	240	40	80	240
P_1	1.00	1.00	1.00	1.04	1.01	1.01	1.11	1.02	1.00	1.05	1.01	1.01
P_2	1.00	1.00	1.00	2.31	1.97	1.89	1.13	1.03	1.01	1.17	1.03	1.01
P_3	5.71	5.29	4.91	1.09	1.01	1.02	1.16	1.03	1.02	1.09	1.01	1.02
P_4	4.30	4.00	3.75	1.29	1.16	1.15	1.15	1.03	1.01	1.19	1.04	1.01
P_5	7.88	7.40	6.76	1.10	1.06	1.02	1.15	1.05	1.02	1.11	1.05	1.02
P_6	2.72	2.62	2.41	2.22	1.80	1.80	1.15	1.04	1.02	1.22	1.04	1.02

References

- Breidt F. J., Claekens, G., Opsomer, J. D. (2003) Model-assisted estimation for complex surveys using penalized splines. Manuscript.
- Montanari G. E. (1987) Post-sampling efficient QR-prediction in large-sample surveys, *International Statistical Review*, 55, 191-202.
- Montanari G. E., Ranalli, M. G. (2002) Asymptotically Efficient Generalised Regression Estimators, *Journal of Official Statistics*, 18, 577-590.
- Särndal C. E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Searle S.R., Casella, G., McCulloch C. (1992) *Variance components*, Wiley, New York.