

Model based efficient stratification methods

Metodi di Stratificazione Efficiente Basati sul Modello

Enrico Fabrizi,

DMSIA, Università di Bergamo
Via dei Caniana 2, 24127 Bergamo
enrico.fabrizi@unibg.it

Carlo Trivisano

Dipartimento di Statistica “P. Fortunati”, Università di Bologna
Via Belle Arti 42, 40126 Bologna

Riassunto: In questo lavoro vengono proposte alcune estensioni del noto algoritmo di Dalenius e Hodges per la stratificazione univariata di una popolazione finita. Tali estensioni si basano su metodi di regressione non parametrici. Viene considerata una situazione informativa in cui si dispone di più variabili di stratificazione e una proxy della variabile obiettivo è nota su un campione. Le performance degli algoritmi sono comparate con uno studio di simulazione basato su pseudo-popolazioni contemplando diverse forme per relazione tra variabile obiettivo e variabili di stratificazione.

Keywords: Dalenius and Hodges rule; one-way optimal stratification; Additive Models; MARS; Boosted Regression Trees

1. Introduction

Let's suppose we want to estimate, say, the population total t_y of a target variable y , by means of the Horwitz-Thompson estimator associated to stratified sampling ($\hat{t}_{y,s}$). Whenever there is some latitude in the definition of strata, that is when stratification is not driven by practical considerations, and we have a single target variable y , it is sensible to look for an efficient stratification in terms of $V(\hat{t}_{y,s})$. This problem is known as one-way optimal stratification. In the simplest case, a univariate auxiliary stratification variable x (known for each unit in the population) is used, and stratification consists in the definition of an ordered sequence of $H-1$ boundary points, such that $x_{INF} < x_1 < x_2 < \dots < x_{H-1} < x_{SUP}$ which partition the domain of x and hence the N observations of population U into H groups. In general stratification can be based on several auxiliary variables ($\mathbf{x}$). A classical theoretical result in the field of optimal one-way stratification, under Neyman allocation and the assumption of stratification based on the target variable ($x \equiv y$), is widely known as Dalenius or $cum\sqrt{f}$ rule (see Cochran 1977, p. 127-133). Practical application of the $cum\sqrt{f}$ rule rely on the assumption that either ($x \equiv y$) or the relationship linking x and y is known. In practice, none of these assumptions hold, and the optimal stratification will be only approximated, the quality of the approximation depending on how strongly the selected stratification variable is correlated with the target variable y or on the adequacy of the assumption of the relationship between y and x . We consider the following

information setting. Suppose that a matrix $\mathbf{X} = \{x_{ij}, i \in U, j = 1, \dots, p\}$ of potential stratification variables is known. Of course, the realized values of target variable $\mathbf{y}_U = (y_1, \dots, y_N)$ are unknown before the survey is conducted but we assume that a proxy y^* of the target variable $\mathcal{Y}$ is observed for a random sample $S \subset U$: that is $\mathbf{y}_S^* = \{y_i^* : i \in S\}$ is observed. This situation may arise for instance in repeated surveys or when data from some pilot survey on a similar subject are available. In particular, in the case of pilot surveys we have that $y_i = y_i^*, i \in S$. Our basic idea is to predict the values of y_i for each unit i in the population using the information contained in $d_S = \{(y_i^*, x_{1i}, \dots, x_{pi}), i \in S\}$ and then use this approximated target variable to stratify the population in an efficient way. Note that “efficient” is used in its informal meaning, and that the stratification methods we propose are not optimal in a mathematical sense. To obtain the predicted values $\hat{y}_i, i \in U$ we consider general regression models and non-parametric fitting methods, that provide flexible tools for obtaining good predictions for the y_i ’s either when the relationship between $\mathcal{Y}$ and the auxiliary variables is linear or nonlinear. In section 2 the generalization of the Dalenius and Hodges rule is described in some detail, while section 3 outlines the simulation exercise used to compare the methods, describes shortly the main results and offers some conclusions and suggestions for further research.

2. Generalizations of the Dalenius and Hodges rule

The Dalenius and Hodges rule is based on two basic elements: i) an auxiliary variable x highly correlated with y , that is, a linear approximation of y ; ii) the $\text{cum}\sqrt{f}$ rule.

A possible extension of the algorithm may be described as follows. Use the sample information d_S to fit the linear model:

$$y_i^* = \alpha + \mathbf{x}_i^T \beta + e_i, i \in U$$

with $E(e_i | \mathbf{x}_i) = 0$, $V(e_i | \mathbf{x}_i) = \sigma^2$ and use the estimated parameters to predict y_i with $\hat{y}_i = \hat{\alpha} + \mathbf{x}_i^T \hat{\beta} \quad \forall i \in U$ ($i \notin S$ if $y^* \equiv y$). Eventually apply the ordinary $\text{cum}\sqrt{f}$ rule on $\hat{\mathbf{y}}_N = (\hat{y}_1, \dots, \hat{y}_N)$. We will refer to this stratification method as Linear Prediction Dalenius-Hodges rule (LPDH). The assumption of a linear relationship between the stratification and the target variable is realistic in many cases, but may fail in others. Thus, a natural generalization of the LPDH rule consists in assuming the more general model

$$y_i^* = g(\mathbf{x}_i) + e_i \quad E(e_i | \mathbf{x}_i) = 0 \quad V(e_i | \mathbf{x}_i) = \sigma^2, \quad (2.1)$$

where g is some unknown regression function. This model is to be fitted by some non parametric algorithm. Once the predicted values $\hat{y}_i \quad \forall i \in U$ are obtained, they can be used for stratification with the $\text{cum}\sqrt{f}$ rule.

Assume that the general model (2.1) is fitted on the basis of the sample data d_S . We consider three different fitting methods that are very popular in the applied non-parametric literature: Additive Regression Models, Multivariate Adaptive Regression Splines and BOOSTed regression trees. We selected these three methods among the

many proposed in the literature because they can be interpreted as generalizations of the linear model, on which the LPDH rule is based. The predicted values of y_i , for all $i \in U$, will then be used to construct stratification rules based on the $\text{cum}\sqrt{f}$ algorithm to be compared with the LPDH on the basis of the simulation experiment described in the next section. A detailed description of these three regression methods is out of the scope of this paper; for an excellent and comprehensive introduction to all the three see for instance the book of Hastie *et al.* (2003, ch. 9, 10).

3. A simulation experiment

We compare the LPDH rule with its generalizations based on regression models (2.1) by means of a simulation exercise that can be described by the following steps:

- i) we generate synthetic populations characterized by different assumptions on g ;
- ii) we generate the population values of y^* , $\mathbf{y}_U^* = (y_1^*, \dots, y_N^*)$ according to a predetermined value of $\text{Corr}_U(\mathbf{y}_U, \mathbf{y}_U^*)$ (with $\text{Corr}_U(\cdot, \cdot)$ denoting the descriptive covariance in population U);
- iii) R simple random samples of fixed size n are drawn from the synthetic populations, d_s is assumed to be observed ($S = 1, \dots, R$);
- iv) the stratification algorithms are applied and their performances compared in terms of

$$\text{Deff}(\hat{t}_{y,s_M=m}) = V(\hat{t}_{y,s_M=m}) \left[V(\hat{t}_{y,srs}) \right]^{-1} \quad (3.1)$$

where $\hat{t}_{y,s_M=m}$ is the stratification estimator of the t_y based on stratification method m . Neyman allocation of the samples to strata (assuming variances within strata known) is assumed. We consider the five synthetic populations discussed in Banks *et al.* (2003). All populations are characterized by the general structure $y_i = g(\mathbf{x}_i) + u_i$, where $g: \mathbb{R}^p \rightarrow \mathbb{R}$ is a deterministic function and u_i is a zero-mean disturbance term. The population size is set to be moderately large: $N = 20,000$. In fact, in many situations efficient methods are to be applied to the stratification of sub-populations. The number of auxiliary variables is set to 6 ($p = 6$) and they are all assumed to be uniformly distributed over the unit interval: $\mathbf{x}_{Uj} \sim \text{Uni}(0,1)$, with $\mathbf{x}_{Uj}$ being the $N \times 1$ column vector with the values of the j -th auxiliary variable. These auxiliary variables are set to be equally and mildly correlated: $\text{Corr}_U(\mathbf{x}_{Uj}, \mathbf{x}_{Uj'}) \cong 0.3 \quad \forall j \neq j'$. As regards the error term we set: $V_U(\mathbf{u}) = (1 - \rho_2^2) \rho_2^{-2} V_U\{f(\mathbf{X})\}$, where $\mathbf{u} = (u_1, \dots, u_N)$. As a consequence, $\text{Corr}_U\{\mathbf{y}, f(\mathbf{X})\} = \rho_2$. We set this parameter ρ_2 to 0.9. The relationship between y and $\mathbf{x}$ in the five synthetic populations ranges from exact linearity to non linear cases. Marginally, the population densities of y range from the exactly symmetrical case to moderately skewed situations. As far the Monte Carlo simulation exercise is concerned, all experiments are based on 2000 independent simple random samples from the synthetic populations. As anticipated, it is assumed that in each sample, a “proxy” variable y^* such that $y^* = y + w$ is observed, where w is a zero-mean disturbance. The values of w are generated in order to satisfy the following conditions: i) $\bar{\mathbf{w}}_U = 0$, ii) $V_U(\mathbf{w}) = \phi^{-2}(1 - \phi^2) V_U(\mathbf{y})$ with $\phi = 0.9$ so that $\text{Corr}_U(\mathbf{y}^*, \mathbf{y}) = 0.9$.

We consider two different sample sizes: $n=100$ and $n=1000$. The problem of determining the number of strata should in principle be also faced. We consider the case of $H=3,6,12$. These values for H have been chosen considering that, when model (2.1) holds and g is linear, little reduction in variance is to be expected beyond $H=6$ (Cochran, 1977 p. 132). We also consider the cases $H=3,12$ in order to verify whether this holds when in model (2.1) g is non linear.

Detailed presentation and analysis of results may be found in Fabrizi and Trivisano (2004). In summary simulation results show that linear prediction is a good basis for the definition of a stratification rule whenever the available sample is small ($n=100$) or the relationship holding between y and $\mathbf{x}$ is not far from being linear. When a large sample is available for “training” the prediction algorithm non-parametric regression algorithms perform closely to the LPDH rule when the relationship between y and $\mathbf{x}$ is linear or approximately linear and better in the non-linear cases. However, not all these methods perform equally well: boosted trees seem to offer the best basis for a modified Dalenius and Hodges stratification rule. As far as the sensitivity of the results to the selected number of strata, we note that we have large gains in precision passing from the simple random sampling to stratification with three strata; smaller, but sometimes still relevant improvements going from three to six strata, while the gains in efficiency in the case with twelve strata with respect to the one with six are minor for all populations and all methods. It should be emphasized that our methods can be applied to generalization of rules other than the Dalenius and Hodges, such as the Ekman’s rule. When stratifying highly skewed populations an analogous generalization of the Lavallée and Hidioglou (1988) can, in principle, be proposed.

Developments in many directions of this research are conceivable: the simulation experiment we considered is fairly general but other and more general situations may be considered, as for instance the inclusion of categorical stratification variables, the stratification of extremely skewed populations, or the case in which the relationship between y and $\mathbf{x}$ is affected by heteroskedastic errors. Moreover, we did not deal with the problem of multiple-ways optimal stratification which is a topic of relevance in many applied situations.

References

- Banks, D.L., Olszewski, R.T., Maxon, R.A. (2003). Comparing Methods for Multivariate Nonparametric Regression. *Communication in Statistics, Simulation and Computation*, 32, 541-571.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed, New York: Wiley.
- Fabrizi E., Trivisano C. (2004) Optimal Stratification Based on Non-parametric Regression Methods, paper presented at the 2004 Joint Statistical Meeting, Toronto, ON (available from the authors).
- Hastie, T., Tibshirani, R., Friedman, J. (2002). *The Elements of Statistical Learning*, New York: Springer.
- Lavallée, P., and Hidioglou, M.A. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33-43.