How optimal subsampling depends on guessed parameter values

Sulla dipendenza dai valori nominali dei parametri nel campionamento ottimale

Laura Deldossi and Chiara Tommasi

Abstract For huge amounts of data, subsampling is useful to downside the data volume to get an inferential goal. We suggest to select a sample of observations applying the theory of optimal design. Assuming a relationship between a response variable and some covariates, the idea is to draw a sample from a large dataset so that it contains the majority of the information about the unknown model parameters. For nonlinear models this optimal selection method depends on the unknown parameters and proper values should be guessed. We analyze how the optimal sample depends on these nominal values in the logistic regression model.

Abstract *Quando si dispone di una ingente quantità di dati, ridurne la dimensione - estraendone un campione - può essere utile ai fini di una analisi inferenziale. Se si assume esista un modello che spieghi la relazione tra una variabile di risposta e alcune covariate, la selezione del campione può essere eseguita secondo la teoria del disegno ottimo, per ottenere un campione che preservi la maggior parte dell'informazione sui parametri incogniti del modello. Quando il modello è non lineare, il disegno ottimo dipende dai parametri ignoti ed è quindi necessario fissare dei valori nominali per poterlo utilizzare. In questo lavoro, assumendo un modello di regressione logistica, ci proponiamo di studiare quanto il metodo di selezione basato sul disegno ottimo sia condizionato dalla scelta dei valori nominali.*

Key words: Logistic model, Optimal sample selection, D-optimality, A-optimality

Laura Deldossi
Department of Statistical Science, Università Cattolica del Sacro Cuore, Milano
e-mail: laura.deldossi@unicatt.it

Chiara Tommasi
Department of Economics, Management and Quantitative Methods, University of Milan
e-mail: chiara.tommasi@unimi.it

1 Introduction

Big Dataset are a huge amount of data that are automatically accrued. Since they arise from observational studies, the quality of the Big Data information might not be very good. In addition, in large-sample studies, if the inferential goal is to test the effect of an explanatory variable, then the p-value often leads to the rejection of the null hypothesis. That is, even very small effects can become statistically significant because of the increased power due to the huge amount of data. From here, the idea of selecting a subsample of the Big Dataset to achieve an inferential goal. This topic has been already studied by [5], [3], [8], [9] and [1], among others.

To get a subsample of data, we apply the theory of optimal design instead of considering the most commonly used sampling schemes. Indeed, the connection between the sampling and experimental design had been already explored by [11], [12], [13], [4] and [6], among others.

In [2] we propose an optimal subsample selection strategy – which is called the "Optimal Design Based" (ODB) method - consisting of two steps. First, we identify the "most informative" values of the explanatory variables according to an optimality criterion (these optimal "theoretical" values are not necessarily present in the observed Big Dataset). Then, we select the observations from the full data set that are closer to these "theoretical" optimal values. Hence, this "optimal-sampling" approach enables us to select the most "informative" observations from the Big Dataset.

A selection strategy that is based on D-optimality and linear models is the Information-Based Optimal Subdata Selection (IBOSS) method that was proposed by [8]. The ODB method, unlike IBOSS, can be based on any optimality criterion (herein, we consider the D- and A-criteria) and can be applied also to nonlinear models.

Since in nonlinear models optimal designs depend on the unknown model parameters, herein we study the dependence of our selection procedure on nominal values of the parameters in a logistic regression model.

As a measure of the quality of the optimal sample based on a guessed nominal value, with respect to the ODB sample obtained from the true parameter value, we use the so called design efficiency, which is defined in Section 2.1.

The remainder of this paper is organized as follows. In Section 2 we briefly describe the Optimal Design Based sampling method and the related notation. In Section 3, we develop a simulation study to compare the ODB sample based on a nominal value with the ODB sample based on the true value of the parameter.

2 A sampling rule based on an optimality criterion

We assume that the relationship between the binary response Y and a set of p covariates $x_1, \ldots, x_p$ is a logistic regression model:

$$P(Y_i = 1; \mathbf{x}_i, \theta) = \frac{e^{\mathbf{x}_i^T \theta}}{1 + e^{\mathbf{x}_i^T \theta}}, \quad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ and $\theta = (\theta_1, \ldots, \theta_p)^T$. We are interested in estimating the parameter vector θ as precisely as possible. It is well known that the asymptotic covariance matrix of the maximum likelihood estimator is the inverse of the Fisher Information matrix

$$I(\theta) = \sum_{i=1}^{N} I(\mathbf{x}_i; \theta), \tag{2}$$

where the typical element

$$[I(\mathbf{x}_i; \theta)]_{r,s} = E_Y \left(-\frac{\partial^2 \log P(Y; \mathbf{x}_i, \theta)}{\partial \theta_r \partial \theta_s} \right).$$

When some values of the covariates are repeated then (2) can be written as

$$I(\theta) = N \sum_{i=1}^{k} I(\mathbf{x}_i; \theta) \omega_i,$$

where $\mathbf{x}_i$ are the distinct values of the covariates and ω_i are their frequencies, $i = 1, \ldots, k$.

$$\xi = \begin{Bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_k \\ \omega_1 & \cdots & \omega_k \end{Bmatrix}, \quad 0 \leq \omega_j \leq 1, \ \sum_{j=1}^{k} \omega_j = 1$$

is the so called approximate design and

$$I(\xi; \theta) = \sum_{i=1}^{k} I(\mathbf{x}_i; \theta) \omega_i \propto I(\theta)$$

is the information matrix of the design ξ. The available data are observational, but if it had been possible, the best choice for the values $\mathbf{x}_i$ and ω_i for $i = 1, \ldots, k$ would have been according to an optimality criterion $\Phi[I(\xi; \theta)]$ for precise estimation of θ.

Given a model and a concave optimality criterion, it is always possible to compute an optimum design as

$$\xi_\Phi^*(\theta) = \arg\max_{\xi} \Phi[I(\xi; \theta)] = \begin{Bmatrix} \mathbf{x}_1^* & \cdots & \mathbf{x}_j^* & \cdots & \mathbf{x}_k^* \\ \omega_1^* & \cdots & \omega_j^* & \cdots & \omega_k^* \end{Bmatrix}. \tag{3}$$

A wise "data generator" would have generated $N\omega_j^*$ responses at $\mathbf{x}_j^*$, $j = 1, \ldots, k$ and this would have been the "ideal" combination of covariates values to be applied to obtain a precise parameter estimation.

Given a large dataset, we aim at selecting a sample s of $n < N$ observations emulating the optimal design $\xi_\Phi^*(\theta)$. More specifically, since the Φ-optimum design (3) gives the best combination of covariate values for the precise estimation of θ, we suggest to select the $n\omega_j^*$ observations that are as closest as possible to

$\mathbf{x}_j^*$ for $j = 1, \ldots, k$. As a measure of closeness, the Euclidean, Mahalanobis or any other distance can be applied. When $n\omega_j^*$ is not an integer number, then a suitable rounding-off rule can be applied (see for instance [7]). This selection method is called optimal design based (ODB) sampling scheme; for more detail see [2]. The ODB sample is denoted by s_θ because it depends on θ through the optimal design $\xi_\Phi^*(\theta)$.

2.1 A- and D-efficiencies

In this paper, we consider the well-known A- and D-optimality criteria defined as $\Phi_A[I(\xi;\theta)] = -\mathrm{Tr}\left[I(\xi;\theta)^{-1}\right]$ and $\Phi_D[I(\xi;\theta)] = |I(\xi;\theta)|^{1/p}$ respectively. The A-optimum design minimizes the total variation of θ as $\xi_A^*(\theta) = \arg\min_\xi \mathrm{Tr}\left[I(\xi;\theta)^{-1}\right] = \arg\max_\xi \Phi_A[I(\xi;\theta)]$. The D-optimum design minimizes the generalized variance of θ as $\xi_D^*(\theta) = \arg\min_\xi |I(\xi;\theta)|^{-1/p} = \arg\max_\xi \Phi_D[I(\xi;\theta)]$.

Let θ_0 denote the true value for the parameter vector θ. The "goodness" of any sample s with respect to the ODB sample s_{θ_0} can be measured through the following ratios

$$0 \le \mathrm{Eff}_A(s;\theta_0) = \frac{\mathrm{Tr}[I(s_{\theta_0};\theta_0)^{-1}]}{\mathrm{Tr}[I(s;\theta_0)^{-1}]} \le 1 \quad \text{and} \quad 0 \le \mathrm{Eff}_D(s;\theta_0) = \frac{|I(s_{\theta_0};\theta_0)|^{-1/p}}{|I(s;\theta_0)|^{-1/p}} \le 1$$

which are called A- and D-efficiencies, respectively.

3 Simulation experiments

This section concerns some results of a simulation study that aims at analyzing how the ODB sampling approach depends on the nominal value of the parameter vector. $N = 100000$ data have been generated from the logistic model (1) with $p = 3$ and $\theta = \theta_0 = (0.5, 0.5, 0.5)^T$. To find the ODB sample s_θ it is necessary to know the parameter value, to be able to compute $\xi_\Phi^*(\theta)$ in (3). In real-life problems, however, we do not know the true value θ_0 and a nominal value θ must be guessed.

To explore how the choice of nominal value influences the ODB selection rule, we considered θ such that $(\theta - \theta_0)$ takes values accordingly to a 3^3 factorial design. In other words, let θ_l be the l-th component of θ, $l = 1, 2, 3$, we fixed $\theta_l = \theta_{0l} + k$ with $k = -1, 0, 1$. In this way we are able to measure the loss in efficiency of s_θ with respect to s_{θ_0}, for guessed θ-values which differ from θ_0 in all directions.

In Table 1, A- and D-efficiencies of s_θ with respect to s_{θ_0} are given for different values of θ, when the covariates are generated from three independent $U(-1, 1)$ random variables. Table 2 reports the same efficiencies when $(X_1, X_2, X_3) \sim N(\mathbf{0}, I_3)$.

Table 1 A- and D-efficiencies of s_θ with respect to s_{θ_0} with $\theta_0 = (0.5, 0.5, 0.5)^T$, when the covariates are generated from three independent $U(-1;1)$.

Guessed value θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$	Guessed value θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$
-0.5, -0.5, -0.5	1.00	1.00	+0.5, +0.5, +1.5	0.791	0.818
-0.5, -0.5, +0.5	0.92	0.919	+0.5, +1.5, -0.5	0.777	0.791
-0.5, -0.5, +1.5	0.752	0.743	+0.5, +1.5, +0.5	0.793	0.815
-0.5, +0.5, -0.5	0.939	0.925	+0.5, +1.5, +1.5	0.618	0.731
-0.5, +0.5, +0.5	0.925	0.921	+1.5, -0.5, -0.5	0.748	0.746
-0.5, +0.5, +1.5	0.772	0.785	+1.5, -0.5, +0.5	0.778	0.784
-0.5, +1.5, -0.5	0.748	0.742	+1.5, -0.5, +1.5	0.615	0.725
-0.5, +1.5, +0.5	0.775	0.784	+1.5, +0.5, -0.5	0.779	0.786
-0.5, +1.5, +1.5	0.622	0.722	+1.5, +0.5, +0.5	0.792	0.816
+0.5, -0.5, -0.5	0.938	0.916	+1.5, +0.5, +1.5	0.612	0.730
+0.5, -0.5, +0.5	0.937	0.924	+1.5, +1.5, -0.5	0.621	0.725
+0.5, -0.5, +1.5	0.775	0.793	+1.5, +1.5, +0.5	0.618	0.730
+0.5, +0.5, -0.5	0.936	0.919	+1.5, +1.5, +1.5	0.603	0.731

Table 2 A- and D-efficiencies of s_θ with respect to s_{θ_0} with $\theta_0 = (0.5, 0.5, 0.5)^T$, when the covariates are generated from three independent $N(0,1)$.

θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$	θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$
-0.5, -0.5, -0.5	1.00	1.00	+0.5, +0.5, +1.5	0.984	0.952
0.5, -0.5, +0.5	1.00	1.00	+0.5, +1.5, -0.5	0.970	0.949
-0.5, -0.5, +1.5	0.977	0.952	+0.5, +1.5, +0.5	0.942	0.926
-0.5, +0.5, -0.5	1.00	1.00	+0.5, +1.5, +1.5	0.875	0.867
-0.5, +0.5, +0.5	1.00	1.00	+1.5, -0.5, -0.5	0.968	0.946
-0.5, +0.5, +1.5	0.967	0.940	+1.5, -0.5, +0.5	0.996	0.955
-0.5, +1.5, -0.5	0.952	0.932	+1.5, -0.5, +1.5	0.921	0.871
-0.5, +1.5, +0.5	0.936	0.931	+1.5, +0.5, -0.5	0.994	0.946
-0.5, +1.5, +1.5	0.889	0.822	+1.5, +0.5, +0.5	0.951	0.952
+0.5, -0.5, -0.5	1.00	1.00	+1.5, +0.5, +1.5	0.894	0.873
+0.5, -0.5, +0.5	1.00	1.00	+1.5, +1.5, -0.5	0.911	0.876
+0.5, -0.5, +1.5	0.975	0.941	+1.5, +1.5, +0.5	0.868	0.850
+0.5, +0.5, -0.5	1.00	0.996	+1.5, +1.5, +1.5	0.698	0.861

3.1 Comments

Comparing the results in Tables 1 and 2, we can observe that the loss in efficiency due to a wrong nominal value of the parameter is always less than 0.4.

Furthermore, it seems that the ODB selection rule is less influenced by the choice of the nominal value when the covariates follow a Gaussian distribution.

However, the "true" ODB sample s_{θ_0} is less informative when the covariates are generated by a normal distribution than when they come from a uniform random variable. This becomes clear if we compare the efficiency of s_{θ_0} with respect to the optimal design ξ^*_Φ (which does not depend on the distribution of the covariates). In the uniform case,

$$\frac{\mathrm{Tr}[I(\xi_A^*;\theta_0)^{-1}]}{\mathrm{Tr}[I(s_{\theta_0};\theta_0)^{-1}]} = 0.925 \qquad \frac{|I(\xi_D^*;\theta_0)|^{-1/3}}{|I(s_{\theta_0};\theta_0)|^{-1/3}} = 0.923$$

while in the Gaussian case,

$$\frac{\mathrm{Tr}[I(\xi_A^*;\theta_0)^{-1}]}{\mathrm{Tr}[I(s_{\theta_0};\theta_0)^{-1}]} = 0.308 \qquad \frac{|I(\xi_D^*;\theta_0)|^{-1/3}}{|I(s_{\theta_0};\theta_0)|^{-1/3}} = 0.362.$$

This means that the per-unit information contained in s_{θ_0} is larger when data come from the uniform distribution; see [2] for a detailed motivation.

In conclusion, when data follow a Gaussian distribution the ODB selection method seems not depending strongly on the nominal value of the parameters, even if the ODB sub-samples are less informative than in the uniform case.

References

1. Campbell, T. and Broderick, T.: Automated scalable Bayesian inference via Hilbert coresets. J. Mach. Learn. Res., **20**, 1–38 (2019)
2. Deldossi, L. and Tommasi, C.: Big Data and model-based survey sampling. http://arxiv.org/abs/2002.04255
3. Drovandi, C.C., Holmes, C., McGree, J.M., Mengersen, K., Richardson, S., Ryan, E.G.: Principles of Experimental Design for Big Data Analysis. Stat. Sci. **32**(3), 385-404 (2017)
4. Fedorov V.V.: Optimal design with bounded density: optimization algorithms of the exchange type. J. Statist. Plann. Inference, **22**, 1-13 (1989)
5. Ma, P. and Sun, X.: Leveraging for Big Data Regression. Comput. Statist.**7**(1), 70–76 (2015)
6. Pronzato L.: On the sequential constructions of optimum bounded designs. J. Statist. Plann. Inference, **136**, 2783-2804 (2006)
7. Pulkesheim, F. and Rieder, S.: Efficient rounding of approximate designs, Biometrika, **79**(4) 763–770 (1992)
8. Wang H. and Yang M. and Stufken J.: Information-Based Optimal Subdata Selection for Big Data Linear Regression. J. Amer. Statist. Assoc., **114**(525), 393-405 (2019)
9. Wang H. and Zhu R. and Ma P.: Optimal Subsampling for Large Sample Logistic Regression. J. Amer. Statist. Assoc., **113**(522), 829-844 (2018)
10. Wynn H.P.: Results in the theory and construction of D-optimum experimental designs, J. of the Royal Stat. Soc. Ser. B, 34, 133–147 (1972)
11. Wynn H.P.: Minimax Purposive Survey Sampling Design. J. Amer. Statist. Assoc., **72**(359), 655-657 (1977)
12. Wynn H.P.: Optimum designs for finite populations sampling. In: Gupta, S., Moore, D. S. (Eds.), Statistical Decision Theory and Related Topics II, 471-478 (1977)
13. Wynn H.P.: Optimum submeasures with applications to finite population. in Gupta, S., Berger, J. (Eds.), Statistical Decision Theory and Related Topics III. Proc. 3rd Purdue Symp., **2**, 485-495 (1982)