Variable selection with LASSO regression for complex survey data

Amaia Iparragirre[1], *Thomas Lumley*[2], *Irantzu Barrio*[3], *Inmaculada Arostegui*[4]

[1]amaia.iparragirre@ehu.eus, Department of Mathematics, University of the Basque Country (UPV/EHU)

[2]t.lumley@auckland.ac.nz, Department of Statistics, University of Auckland (UoA)

[3]irantzu.barrio@ehu.eus, Department of Mathematics, University of the Basque Country (UPV/EHU) & Basque Center for Applied Mathematics (BCAM)

[4]inmaculada.arostegui@ehu.eus, Department of Mathematics, University of the Basque Country (UPV/EHU) & Basque Center for Applied Mathematics (BCAM)

Complex survey data are becoming increasingly relevant in a number of fields, including social and health sciences. In this framework, the finite population of interest for the study is usually sampled following a complex sampling design, which may include techniques such as stratification, clustering, or a combination of them in different stages of the sampling scheme. In this context, a sampling weight is assigned to each sampled unit, indicating the number of units that this observation represents in the finite population. Due to these particularities, complex survey data do not satisfy independence and identically distributed conditions, and hence, validity of traditional statistical techniques should be checked before applying them to data collected from complex surveys.

LASSO regression models are one of the most commonly used methods for variable selection. In this context, a tuning parameter must be previously selected to fit the models. Cross-validation is the most widely used validation technique in practice to select the optimal value of this parameter in order to minimize the error of the model to be fitted.

Nevertheless, applying LASSO regression models to complex survey data could be challenging for several reasons, including the fact that traditional validation techniques need to be updated in order to work properly with this type of data. In complex survey framework, other approaches, different to the traditional validation techniques are usually used to define partially independent subsets of the sample. Those approaches are known as "replicate weights" methods. However, to our knowledge, they have never been used in a LASSO regression context. The goal of this work is two-fold. On the one hand, we analyze the performance of replicate weights methods to select the tuning parameter for fitting LASSO regression models to complex survey data. On the other hand, we propose new replicate weights methods for the same purpose. In particular, we propose a new design-based cross-validation method as a combination of the traditional cross-validation and replicate weights. The performance of all these methods has been analyzed and compared by means of an extensive simulation study to the traditional cross-validation technique to select the tuning parameter for LASSO regression models. The results suggest a considerable improvement when the new proposal design-based cross-validation is used instead of the traditional cross-validation.

Keywords: complex survey data, LASSO regression, replicate weights.