A regression clustering method for the prediction of the pro capita disposal income in municipalities

Paolo Chirico

1 Introduction

The aim of *regression clustering* (Bin Zhang, 2003) is segmenting a number of units in some clusters in order to detect a good regression model in each cluster. Then regression clustering is suitable when, given some explicative variables (regressors), a single regression model doesn't fit well all the units, but different regression models might fit well partitions of the data (see also Sarstedt and Schwaiger (2006)). In this paper a regression clustering procedure is adapted to a particular regression to predict the pro capita disposal income (PCDI) in municipalities. The particularity of this regression consist in: it is a two-level regression (municipalities and provinces) and the parameters estimation is run at the provincial level under some assumptions.

2 Main Results

The *PCDI* of a municipality is assumed explainable by some municipal indices (regressors) in a regressive model, but a single model for every municipality in a country or region would be a bit efficient: the regression errors may be too large. It is more flexible to assume the existence of K regressive models explaining the municipal *PCDI* in K clusters of provinces. So:

$$y_{ijk} = \mathbf{x}_{ijk}\beta_k + \varepsilon_{ijk} \tag{1}$$

where y_{ijk} is the *PCDI* of the i^{th} municipality in the j^{th} province of the k^{th} cluster; $\mathbf{x} = [1, X_1, X_2,]$ is the vector of the regressors and β is the vector of the correspondent parameters; ε is a random error. The distributional features of ε are inferred by the following model, assumed for the individual disposable income:

Paolo Chirico
Dep. of Applied Statistics e Mathematics, Turin University, e-mail: paolo.chirico@unito.it

$$y_{hijk} = \mathbf{x}_{ijk}\beta_k + \varepsilon_{hijk} \quad (2)$$

where h indicates the individual; ε_{hijk} is a random error and includes all the individual factors determining y. Therefore it is assumed independent of each other error as well as of the regressors. No distributional form is assumed about ε_{hijk}, but only $E(\varepsilon_{hijk}) = 0$ and $Var(\varepsilon_{hijk}) = \sigma_k^2$.

As $y_{ijk} = \sum y_{hijk}/n_{ijk}$ then $\varepsilon_{ijk} = \sum \varepsilon_{hijk}/n_{ijk}$; generally $n_{ijk} > 1000$ so ε_{ijk} is approximately $N(0, \sigma_k^2/n_{ijk})$. Moreover ε_{ijk} is independent of each other error and of the regressors as well. Consequently the model for the provincial PCDI is:

$$y_{jk} = \mathbf{x}_{jk}\beta_k + \varepsilon_{jk} \quad (3)$$

where $y_{jk} = \sum y_{hijk}/n_{jk}$ and $\varepsilon_{jk} = (\sum \varepsilon_{hijk}/n_{jk}) \sim N(0, \sigma_k^2/n_{jk})$.

The *PCDIs* of the municipalities are unknown (they are the object of the prediction), but the *PCDIs* of the provinces are. So the parameters in β_k are estimated through provincial data by the *WLS* method.
The clusters are determined by a segmentation oriented to the efficiency of the local regression models. This procedure of "regression clustering" (see Introduction), is a model-based version of the K-means clustering method. It is characterized by the following steps:

1. estimation of a global regression model on all provinces;
2. hierarchical classification on the residual of the global model (dendogramme);
3. choice of the number, K, of the clusters according to the dendogramme and assignment of provinces to the K clusters;
4. estimation of the K local model (one for each cluster) and computation of the prediction error for each provinces in each K local model;
5. assignment of each provinces to the closest local model (where the prediction error is the smallest);
6. repetition of the steps 4. and 5. until the composition of the K clusters doesn't change.

References

Sarstedt M., Schwaiger M. (2006). Model Selection in Mixture Regression Analysis: A Monte Carlo Simulation Study. *Data Analysis Machine Learning and Applications*, Springer Berlin Heidelberg, 61-68.

Zhang B. (2003). Regression Clustering. In: ICDM03, Third IEEE International Conference on Data Mining, 451.