A Clusterwise regression method for the prediction of the disposal income in municipalities

Paolo Chirico
DIGSPES Alessandria
University of Eastern Piedmont, Italty
paolo.chirico@uniupo.it

Abstract The paper illustrates a *clusterwise regression* procedure applied to the prediction of per capita disposal income (*PCDI*) in Italian municipalities. The municipal prediction is derived from the provincial *PCDI* taking into account the discrepancy between municipality and province in some indicators like per capita taxable income, per capita bank deposits, employment rate, etc. The relation between *PCDI* and indicators is shaped by a regression model. A single regression model doesn't fit very well all territorial units, but different regression models do it in groups of them. The aim of clusteriwise regression is just that: detecting clusters where the correspondent regression models explain the data better than an overall regression model does. The application of the procedure to a real case shows that a significative reduction of the regression standard error can be achieved.

1 Introduction

The present work originates from a study of Unioncamere Piemonte (2009) about the prediction of the per capita disposal income (*PCDI*) in the Piedmont municipalities. More specifically Unioncamere Piemonte intended to predict the *PCDI* of the Piedmont municipalities by means of a regression model using some municipal indicators like "per capita taxable income", "per capita bank deposits", etc. Formally:

$$y_{ij} = \mathbf{x'}_{ij}\beta + \varepsilon_{ij} \tag{1}$$

where y_{ij} is the *PCDI* of the i^{th} municipality in the j^{th} province; $\mathbf{x'}_{ij}$ is the vector of regressors; β is the vector of the correspondent coefficients; ε_{ij} is the residual regression error.
Unioncamere knew the indicators for every Piedmont municipality but didn't know the *PCDIs*, even for a sample of municipalities, so that the model parameters couldn't be estimated on municipal data. On the other hand, all data were known at provincial level (the provincial *PCDIs* were provided by an external research institute). Therefore, the model parameters were estimated using the model (1) at provincial level; the *Ordinary Least Squares* estimation method was adopted considering all provinces on the same level of importance.

This paper proposes an evolution of that model in order to:

This is a post-peer-review, pre-copyedit version of the article published in Classification and Data Mining, 2013, Springer

- formalize better the regression errors and have municipal predictions consistent with the provincial *PCDI* (Section 2);
- reduce the prediction errors by means of a *clusterwise regression* procedure (Section 3).

2 The basic model

Let's assume that the municipal *PCDIs* can be explained by some municipal indicators with a linear regression model like (1). The regression error ε_{ij} can be viewed as:

$$\begin{aligned}\varepsilon_{ij} &= y_{ij} - \mathbf{x'}_{ij}\beta = [\sum_h y_{hij}]/n_{ij} - \mathbf{x'}_{ij}\beta \\ &= \sum_h [y_{hij} - \mathbf{x'}_{ij}\beta]/n_{ij} = \sum_h \varepsilon_{hij}/n_{ij}\end{aligned} \tag{2}$$

where ε_{hij} is the difference between the disposal income of the generic h^{th} resident and the expected *PCDI* in its municipality; n_{ij} is the municipal population.
According to its definition, ε_{hij} is a random error and includes all individual factors determining the individual disposal income. At first every ε_{hij} is assumed *independent* of every other error and regressor, and *identically distributed* with $E(\varepsilon_{hij}) = 0$ and $Var(\varepsilon_{hij}) = \sigma^2$. Such statements are clearly hard, but, at the moment, let's view them as a way to formalize better the features of ε_{ij}. Since $\varepsilon_{ij} = \sum_h \varepsilon_{hij}/n_{ij}$ and generally $n_{ij} > 1000$, ε_{ij} can be assumed Gaussian. Now the model (1) can be better specificated as:

$$y_{ij} = \mathbf{x'}_{ij}\beta + \varepsilon_{ij} \tag{3}$$

with $\varepsilon_{ij} \sim N(0, \sigma^2/n_{ij})$

As the provincial *PCDI* is $y_j = \sum_{hj} y_{hij}/n_j$, then:

$$y_j = \mathbf{x'}_j\beta + \varepsilon_j \tag{4}$$

with $\varepsilon_j \sim N(0, \sigma^2/n_j)$

In our case the *PCDIs* of the municipalities are unknown, even for a sample of municipalities, so that the model (3) is not useful for the parameters estimation. Nevertheless the *PCDIs* of the provinces are known so that the model parameters can be estimated through provincial data (model 4). Since the provincial regression errors have different variances, each of them equal to σ^2/n_j, the *Weighted Least Squares* (*WLS*) estimation method should be used:

$$\beta = (\mathbf{X'NX})^{-1}\mathbf{X'Ny} \tag{5}$$

where $\mathbf{X}$ is the data matrix of provincial regressors; $\mathbf{y}$ is the vector of provincial *PCDIs*; $\mathbf{N}$ is the diagonal matrix of provincial populations.
Now let's reconsider the assumptions about ε_{hij}. If the assumptions about mean and variance can be acceptable, their independence seems not realistic, in particular among the individual errors in a same municipality. Nevertheless these assumptions have only one effect on the modeling: the adoption of the *WLS* method for the models estimations. That means the models have to fit better the

provinces with more population, and that seems reasonable.

According with the model (3), the prediction of the municipal *PCDI* should be $\widehat{y}_{ij} = \mathbf{x}'_{ij}\widehat{\boldsymbol{\beta}}$ since the prediction of the municipal error, $\widehat{\varepsilon}_{ij}$, is generally assumed equal to zero. Nevertheless the provincial average of the municipal errors, $\widehat{\varepsilon}_j$, is known before predicting the municipal errors, $\widehat{\varepsilon}_{ij}$; indeed it is known by the estimation of the provincial models (4): $\widehat{\varepsilon}_j = y_j - \mathbf{x}'_j\widehat{\boldsymbol{\beta}}$.
A way to take into account this information is to predict every municipal errors in a province equal to their provincial average: $\widehat{\varepsilon}_{ij} = \widehat{\varepsilon}_j$. Consequently the municipal *PCDI* prediction becames:

$$\widehat{y}_{ij} = \mathbf{x}'_{ij}\widehat{\boldsymbol{\beta}} + (y_j - \mathbf{x}'_j\widehat{\boldsymbol{\beta}}) = y_j + (\mathbf{x}'_{ij} - \mathbf{x}'_j)\widehat{\boldsymbol{\beta}} \tag{6}$$

Therefore the prediction of the municipal *PCDI* can be viewed as an adjustment of the provincial *PCDI* on the basis of the differences between the municipal indicators and the provincial ones. Moreover, the formula (6) assures that the average of all municipal predictions is equal to the known provincial *PCDI*.

3 From a single model to *k* models

The detection of a suitable provincial model (4) (and its estimation) only on the basis of the data of the eight Piedmont provinces would have led to an overfitting model. To get over this problem, the model was initially generalized to the Italian provinces and was therefore estimated using the data of 87 Italian provinces (some provinces were excluded from the analysis because not all the requested data were available). The regression results are reported in Table 1 and 2.

Table 1 Regressors and coefficients

regressor	coefficient	sign.
intercept	5.710,91	***
per capita taxable income	0,59	***
employment rate	69,38	***
per capita banc deposit	0,18	***
rate of graduates	- 266,40	***
oldness index	14,94	***

Table 2 Quality Indices

Index	value
R^2	0,962
$\overline{R}^2$	0,959
$\widehat{\sigma}$	486.879,5

We can note an unexpected results: the negative contribution of ”rate of graduates”. It doesn't mean that the relationship between *PCDI* and ”rate of graduates” is negative, indeed their correlation is positive, although very low (0,152). It means that the contribution of the ”rate of graduates” to the prediction of *PCDI* with the others predictors is negative; it concerns the role of the ”rate of graduates” in explaining what it is not explained by the others predictors.

The R^2 and the $\overline{R}^2$ are very high, and that is understandable since the high correlation between *PCDI* and the regressor "per capita taxable income" (0,958). All regressors are significant at 1% level (***) and each one of them improves the Akaike's Information Criterion (AIC) and the Schwarz' Criterion (SC) if added after the other regressor.
Nevertheless, even if the R^2 and the $\overline{R}^2$ are very high, we can't state that the model fits the data very well. Indeed the value of the standard regression error, $\widehat{\sigma}$, is not realistic (486.879 euros!). According to the assumpion in the section 2, σ is the standard deviation of ε_{hij} and can be viewed as a measure of the average difference between the individual disposal income and the expected *PCDI* in the correspondent municipality. If the model fits well the data, the value of $\widehat{\sigma}$ should be realistic. Therefore, an overall model like (4) is not good for every Italian province. On the other hand, *K* groups (*clusters*) of provinces may be fitted quite well by *K* local regression models, like:

$$y_{jk} = \mathbf{x}'_{jk}\beta_{\mathbf{k}} + \varepsilon_{jk} \tag{7}$$

with $\varepsilon_{jk} \sim N(0, \sigma_k^2/n_{jk})$, $k = 1,..,K$

The detection of such locals model and the corresponding partition concerns the *clusterwise regression*.

3.1 Clusterwise regression

The aim of *clusterwise regression*, (CR), also named *regression clustering* by other authors (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. The origins of CR can be founded in the works of Bock (1969) and Spaeth (1979), whose original algorithms can be viewed as a special case of k-means clustering with a criterion based on the minimization of the squared residuals instead of the classical within-class dispersion (Preda and Saporta , 2005).
More specifically, if $G = \{G(1), G(2), ..., G(n)\}$ identifies a partition of *n* units in *K* clusters, and:

$$V(K, G, \beta_1, ..., \beta_K) = \sum_k \sum_{G(i)=k} (y_i - \mathbf{x}'_i\beta_k)^2 \tag{8}$$

is the sum of the squared residuals of the *K* local regressions, the basic algorithm of CR consist on iterating the following two steps:

a) for given G, $V(K, G, \beta_1, ..., \beta_K)$ is minimized by the LS-estimators of the $\beta_1, ..., \beta_K$;
b) for given $\beta_1, ..., \beta_K$, $V(K, G, \beta_1, ..., \beta_K)$ is minimized by assigning each unit to the cluster where the corresponding regression error is minimum; that identifies a new partition G.

Like in k-means clustering (MacQueen, 1967) the algorithm in converging, because, the sequence of $V(K, G, \beta_1, ..., \beta_K)$ is, clearly, monotonically non-increasing. But, unlike k-means clustering, the algorithm converges to a local optimal solution, that depends on the initial partitions and not necessarily is the global optimal solution. Therefore, it would be better to simulate several initial partition in order to choose the best final partition! Since its development, numerous adaptations and extensions of CR have been proposed; DeSarbo and Cron (1988) extended clusterwise regression to the case of multiple response variables and repeated measures on subjects and proposed a simulated annealing algorithm for solving the resulting optimization problem. As reported in (Brusco et al. ,

2008), mixture-model formulations of CR have been proposed by numerous authors (DeSarbo and Cron , 1988), (Henning, 2000) that assume the response variable measures are obtained from a mixture of K conditional densities (usually normal) that arise in unknown proportions.

Obviously, the bigger the number of clusters, the better the fit of data, but that doesn't mean necessary better partition of data. About this issue, DeSarbo and Cron (1988) suggest to adopt the Akaike's Information Criterion, while Henning (2000) suggest to adopt the Schwarz' Criterion.
A correlated issue is the problem ovefitting, that has been analyzed recently by Brusco et al. (2008).

3.2 Four models for PCDI prediction

To detect the local models (7) for *PCDI* prediction, the basic algorithm of CR, with WLS estimation method, was adopted. According with DeSarbo and Cron (1988) and Henning (2000), partitions in 2, 3, 4, 5 clusters were tried, in order to detect the most suitable solution. For every partition in K cluster, several random initial partition were used.
The Table 3 reports some quality indices of the final (optimal) partitions in different number of clusters.

Table 3 Quality Indices for each partition

num.clusters	1	2	3	4	5
AIC	2.531,3	2.440,8	2.373,2	2.314,2	2.299,2
SC	2.546,1	2.472,9	2.422,5	2.380,8	2.383,1
logL	-1.259,7	-1.207,4	-1.166,6	-1.130,1	-1.115,6
min $\widehat{\sigma}$	486.679	187.114	128.378	104.720	98.436
max $\widehat{\sigma}$	486.679	254.027	181.329	155.416	146.775

The partitions in 4 clusters is better according to Schwarz' Criterion, while the partition in 5 clusters is better according to Akaike's Criterion. In such partition the local regression standard errors are less than in 4-clusters partition, but the improvement is not very significant, so the partition in 4 clusters was preferred. The table 4 report the local regression results of that partition.

Table 4 The four local regressions

	cluster 1	cluster 2	cluster 3	cluster 4	overall
intercept	3.488,45	4.662,35	4.543,81	4.113,12	5.710,91
per capita taxable income	0,42	0,50	0,40	0,22	0,59
employment rate	146,23	87,08	106,45	182,05	69,38
per capita banc deposit	0,05	0,19	0,25	0,21	0,18
rate of graduates	-144,06	-139,31	-179,61	-183,39	-266,40
oldness index	16,09	14,51	19,31	21,11	14,94
provinces	18	31	19	20	88
R^2	1,00	1,00	1,00	1,00	0,99
$\overline{R}^2$	0,99	1,00	1,00	1,00	0,959
$\widehat{\sigma}$	143.368	111.928	104.720	155.416	486.879

Now the standard regression errors of the local regressions are clearly lower, and consequently more realistic than the standard regression error of the overall regression. Both the R^2s and the $\overline{R}^2$s are very high and could be a sign of overfitting, but it is not the case. Indeed the same indices of the overall model are high too and not for the presence of overfitting, as explained in section 2.

3.3 The municipal PCDI prediction

Properly, the clusterwise regression described in the last subsection has concerned the provincial models, not the municipal ones. Then the extension of the clustering to the municipal predictions requires the assumption that the *PCDIs* of all municipalities of a province are explained by the model of their province:

$$y_{ijk} = \mathbf{x}'_{ijk}\beta_k + \varepsilon_{ijk} \tag{9}$$

with $\varepsilon_{ijk} \sim N(0, \sigma_k^2 / n_{jk})$.
Therefore, the *PCDIs* of the municipality *i* of the province *j* belonging to the cluster *k* will be predict by the following formula:

$$\widehat{y}_{ijk} = y_{jk} + (\mathbf{x}'_{ijk} - \mathbf{x}'_{jk})\widehat{\beta}_k \tag{10}$$

Obviously some municipal *PCDI*s might be explained better by the model of another cluster than by its cluster model. Nevertheless there isn't way to known exactly which model is the best for every municipality. Then, in absence of further information, the assumption in (9) can be reasonable at least for middle-big municipalities that are not too different from the profile of its province.

4 Final Considerations

The paper describe a case where the clusterwise regression can be useful to detect a number of suitable regression models in a heterogeneous population. All the methodology can be viewed like a way to predict the municipal *PCDIs* in case of: *(i)* the *PCDIs* are explainable by some regressors; *(ii)* the *PCDIs* are not known at municipal level, but are known for territorial aggregations; *(iii)* the territorial aggregations are heterogeneous. Obviously the number of territorial aggregations has to be enough numerous for being segmented in clusters where regression models are drawn.
The explained methodology joins in a series of proposals about the Italian municipal disposal incomes, that includes Marbach (1985), Frale (1998), Bollino and Pollinori (2005), quoting only some authors. Here, as in Marbach, the municipal disposal income is derived from the provincial disposal income, but in Marbach the provincial disposal income is object of prediction; here is exogenous. As in Bollino and Pollinori, the regressive models are heteroscedastic and the estimated provincial errors are used for the prediction of the municipal errors. Those proposals illustrate procedures very articulated, but don't handle the problem of the heterogeneity by means of a model-based approach. The present proposal does it by clusterwise regression.
Finally the provincial *PCDIs* are exogenous data in the models as well as all the regressors. The present paper doesn't consider how they are calculated. Actually the most of them are estimated. For example, the Bank of Italy estimates the *PCDIs* at regional level by a sample survey; private research institutes provide estimations of the *PCDIs* at provincial level, but their methods are not exactly known.
Obviously the quality of the municipal predictions (10) depends on the quality of exogenous data too!

References

Bock H. H. (1969). The equivalence of two extremal problems and its application to the iterative classification of multivariate data. *Lecture note, Mathematisches Forschungsinstitut Oberwolfach*.

Bollino C.A., Pollinori P. (2005). Il valore aggiunto su scala comunale: La Regione Umbria 2001-2003. *Quaderni del Dipartimento di economia, Finanza e Statistica*, No. 15/2005.

Brusco, M.J., Cradit, J., Steinley, D., Fox, G.J. (2008) Cautionary Remarks on the Use of Clusterwise Regression. *Multivariate Behavioral Reseach*, 43, 29-49.

DeSarbo, W. S., Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5, 249282.

DeSarbo, W. S., Oliver, R. L., Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear regression. *Psychometrika*, 54, 70736.

Frale C. (1998). Stime comunali del reddito disponibile: la provincia di Udine, *Osservatorio permanente dell'economia del Friuli venezia Giulia*, No. 3.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17, 273296.

Marbach G. editor. (1985). Il reddito nei comuni italiani 1982. *Quaderni del Banco di Santo spirito*, UTET, Torino.

MacQueen J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1, 281-297.

Preda, C., Saporta, G. (2005): Clusterwise PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 49, 99108.

Unioncamere Piemonte (2009). *Geografia dei redditi 2009, Osservatorio sul reddito disponibile e prodottoin Piemonte*.

Spaeth H. (1979). Clusterwise linear regression. *Computing* 22, 367-373.

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