Statistica & Applicazioni - Vol. VII, n. 1, 2009, pp. 45-61

A NEW OLS-BASED PROCEDURE FOR CLUSTERWISE LINEAR REGRESSION

Antonella Plaia*
Salvatore Bologna**

SUMMARY

Data heterogeneity, within a (linear) regression framework, often suggest the use of a Clusterwise Linear Regression (CLR) procedure, which implies, among other things, the estimate of the appropriate number of clusters as well as the cluster membership of each unit. The approaches to the estimation of a CLR model are essentially based on the Ordinary Least Square (OLS) criterion or the likelihood criterion. In this paper, in a context of OLS approach, we propose an estimation of the model making use of an algorithm based on a threshold criterion for the determination coefficient of each cluster, to identify the appropriate number of clusters, and of a modified Spath's algorithm, to estimate the cluster membership of each sample unit. A simulation design and an application to a real data-set show that the procedure outperforms other algorithms commonly used in literature.

Keywords: *Linear regression, Cluster analysis, Monte Carlo simulation.*

1. INTRODUCTION

The classical linear regression model assumes that a linear statistical relationship between J explanatory variables X_j and a dependent variable Y holds for all data sample, that is data are homogeneous with respect to a single linear regression relationship. It can happen that data-set presents a group of outliers whose cut off leaves homogeneity among the other units. It can also happen that data heterogeneity implies that some kind of segmentation or clustering exists among the sample units so that more than a single regression relationship should be used to fit the data. In this case data-set can appear partitioned in two or more regimes with respect to the level of a variable (Switching Regression) and, thus, the problem becomes to identify the change point(s) among regimes and to estimate the parameters (Quandt, 1958); or, also, data-set can appear such that two or more linear regression functions can be necessary to summarize the underlying structure of the data, but no change points exist with respect to the predictor(s). In the last case one can apply regression in a finite mixture context. Regression modelling for finite mixtures, also known as

* Dipartimento di Scienze Statistiche e Matematiche ''S. Vianelli'' - Università degli Studi di Palermo, Viale delle Scienze - 90128 PALERMO (e-mail: plaia@unipa.it).
** Dipartimento di Scienze Statistiche e Matematiche ''S. Vianelli'' - Università degli Studi di Palermo, Viale delle Scienze - 90128 PALERMO (e-mail: bologna@unipa.it).

Clusterwise Linear Regression, CLR, (Spath, 1979) or Latent Class Regression, is relevant in biology (animals and plants can be grouped according to linear relationships among their properties) and economics. In marketing, for example, consumers can be grouped according to the kind of the relationship between utility and price (Figure 1, artificial data, can represent two different groups of consumers).

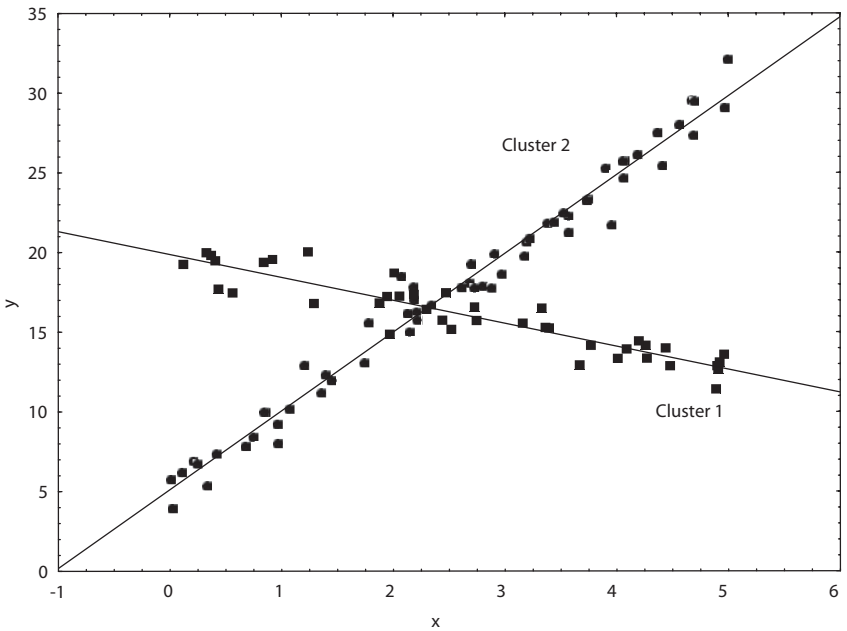

FIGURE 1. - *Two different relationships between utility and price*

The approaches to the estimation of a regression model for finite mixture are essentially based on the Ordinary Least Square (OLS) criterion (Spath, 1979) or on the likelihood criterion (see, for example, DeSarbo and Cron, 1988). The number of clusters is usually unknown, and thus the inference on the model implies the estimation of the appropriate number of clusters together with the cluster membership of each unit and the regression coefficients.

In particular, in order to estimate the cluster membership of each unit, (Spath, 1979) proposed the so called Exchange algorithm. The final result of the estimation of the model, according to Spath's procedure, depends on the choice of an initial partition of the data, on the starting observation and on the choice of the minimum number of observations in each cluster; for these reasons, the procedure exhibits prone to provide locally optimum solutions. The maximum likelihood approach, essentially, overcomes this problem and provides globally optimum solutions. Most of the literature, since DeSarbo and Cron (1988), followed the likelihood approach.

In this paper, in a context of OLS approach, we propose an estimation of the model based on an algorithm based on a threshold criterion for the determination

coefficient of each cluster, to identify the appropriate number of clusters, and on a modified Spath's algorithm, to estimate the cluster membership of each sample unit. We show that this procedure, similarly to the likelihood one, provides a global optimum solution. In addition, the proposed procedure is comparable to the likelihood approach, according to the goodness of fit, and presents a higher speed of convergence with respect to the procedures based on the maximum likelihood.

In the following sections the literature on the problem will be reviewed, the new algorithms will be described and their performance will be evaluated and compared with other methods presented in the literature, by means of both a factorial design of Montecarlo simulations and an application to a real data-set analyzed by Cook and Weisberg (1994) and Pena, Rodriguez, and Tiano (2003).

2. THE LITERATURE

A model specifying clusterwise linear relationships can be written as:

$$y_i = \sum_{c=1}^{C} \sum_{j=0}^{J} a_{ic} \beta_j^c x_{ij} + \varepsilon_i \tag{1}$$

where:
y_i is the $i-th$ observation of the dependent variable Y, $i = 1, 2, \cdots, n$;
x_{ij} is the $i-th$ observation of the $j-th$ predictor, $j = 0, \cdots, J$, with $x_{i0} = 1$;
β_j^c is the $j-th$ regression coefficient in the $c-th$ cluster, $c = 1, 2, \cdots, C$;
a_{ic} is the indicator function of the $c-th$ cluster:

$$a_{ic} = \begin{cases} 1 & \text{if the } i-th \text{ unit belongs to the } c-th \text{ cluster} \\ 0 & \text{otherwise;} \end{cases}$$

ε_i is the determination of a Normal random variable: $N(0, \sigma^2)$.

The fundamental papers are the ones by Spath (1979), DeSarbo and Cron (1988) and, more recently, Hennig (2003).

In order to estimate the regression parameters and the cluster membership of each unit Spath (1979) proposed the so called *Exchange Algorithm*. With a known number of clusters C, he considers a starting $C-$cluster random partition and estimates the regression coefficients in model (1) by OLS; then, starting from a random point (unit) in the sample, at each iteration moves *a single unit*, from a cluster to another one, if such movement cuts down the following objective function (defined as the sum of the squares of errors over all clusters):

$$Z_C = \sum_{c=1}^{C} \sum_{i=1}^{n_{c,C}} \left(y_i - \sum_{J=0}^{J} \hat{\beta}_j^c x_{ij} \right)^2, \tag{2}$$

where $n_{c,C}$ is the dimension of the $c-th$ cluster. In Figure 1, for example, one square unit can be moved from cluster 1 to cluster 2 if this reduces the objective function (2). According to this approach, in order to properly estimate β_j^c, the num-

ber of different units belonging to each cluster must necessarily exceed the number of parameters to be estimated (that is $n_{c,C} > J + 1$) and thus, for each cluster, a minimum number of different units n_c^* is usually required. The final partition, as the same Spath asserts, depends on such a minimum value and on the starting configuration.

DeSarbo and Cron (1988) consider an approach to the estimation of the model based on the likelihood function. The model is specified as a mixture of Gaussian distributions and, given the number C of groups and the observed data, under suitable restrictions, the parameters can be estimated by the (log)likelihood function. As far as the cluster memberships, a probabilistic clustering can be obtained by the estimation of the posterior probabilities (McLachlan and Basford, 1987). A partition of the units into C nonoverlapping clusters can be obtained by assigning each unit to the cluster to which it has the highest estimated posterior probability of belonging.

DeSarbo and Cron propose then to use the Akaike Information Criterion (AIC) to choose the number of groups. AIC is defined as:

$$AIC(C) = -2\log L(C) + 2k(C)$$

where $k(C)$ is the effective number of parameters estimated in a C-clusterwise regression solution, but the performance of this proposal is not treated. Cleaver and Wedel (2001) and Wedel and DeSarbo (2002) use Consistent Akaike Information Criterion (CAIC), defined as:

$$CAIC(C) = -2\log L(C) + \log(n+1)k(C)$$

to find the number of clusters that best represents the data.

Interesting suggestions come from Hennig (2000) who considers the identifiability problem in mixture model related to relabeling of components and overfitting (if two or more components have the same parameters the data generating process can be represented by a smaller model with fewer components), and the slowness of convergence (to local maximum) of the EM algorithm, usually adopted in this context. Leisch (2004a), in the R-package FlexMix (Leisch, 2004b) tries to avoid overfitting vanishing prior probabilities by automatically removing components whose prior λ_c falls below a user-specified threshold.

A Bayesian approach can be found in Viele and Tong (2002), who extend computational methods and consistency results of mixture models to mixture of linear regressions.

Hennig's FPC (Henning, 2003), differently from the approaches described above, considers a single cluster at a time, assuming the rest of the data to be outliers with respect to this cluster, and uses Schwarz's Criterion defined as:

$$BIC(C) = -2\log L(C) + \log nk(C)$$

or CAIC to estimate the number C of clusters.

As far as the choice of the appropriate number of clusters, it is worth noting that specific tests are not available, and the procedures usually employed are to be seen as a guide rather than an absolute criteria.

The clusterwise linear regression problem has also been dealt widely in a fuzzy framework. A reference paper is the one by Jajuga (1986).

3. A NEW APPROACH TO CLUSTERWISE LINEAR REGRESSION

Within a regression framework, given a $(J+1)$-dimensional n-sample from a dataset with, in general, data heterogeneity, we assume that the underlying structure of the data is a clusterwise linear regression structure and we aim at determining simultaneously the most appropriate number of clusters C and the best partition of the sample given C.

To solve this problem we propose an OLS approach that develops according to a sequential procedure based on the following conditions:

A1 the coefficient of determination in each cluster of the partition must be not less than a 'fixed threshold value' which ensures that each estimated regression model fit the data, in the respective cluster, in a satisfactory way. We can write:

$$R^2_{c,C} \geq \delta,\ c = 1, 2, \ldots C, \tag{3}$$

where $R^2_{c,C}$ is the coefficient of determination of the model fitted in the $c-th$ cluster in a generic partition into C groups, δ is a fixed threshold value (for example $\delta = 0.8$), $C = 1, 2, \ldots C_{max}$ represents the number of clusters to be identified and C_{max} the maximum number of groups allowed. Note that the choice of the most appropriate threshold value, as well as of C_{max}, is usually left to the judgment of the operator;

A2 the number $n_{c,C}$ of different units in each cluster must be at least equal to a value $n^* \geq J+1$;

A3 for every fixed $C = 2, \ldots, C_{max}$ we will choose, among the partitions into C groups, the one that achieves the minimization of (2), by using a modified Spath's algorithm.

3.1 *The Determination Coefficient Threshold Criterion (DCTC)*

We can say that the problem of 'clusterwise linear regression' essentially consists in finding an appropriate number of clusters of the observations such that units inside each cluster satisfy a linear regression relationship. In other words we look for an 'admissible' classification in the sense that a linear statistical model between y and x_j's holds for each cluster. By means of the new procedure of clustering we aim, on one side, at determining a sampling partition into C groups such that the sum of the squares of errors (2) computed over all clusters is minimized and, on the other, at guaranteeing that all the estimated linear regression models fit the units in the respective cluster in the best way. We consider a partition to be satisfactory when a 'good' coefficient of determination is associated to each cluster regression model.

Among the classifications that satisfy conditions A1, A2 and A3, we will select the one with the smallest number of clusters, say $\hat{C}$, by considering that increasing the number of clusters we could achieve a better fit, but get a trivial solution (a lot of smaller clusters with $R^2_{c,C} \cong 1$ and a Z_C value closer to zero).

Actually, more than a single partition, say $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_m$, could minimize (2) and satisfy A1 and A2: therefore a 'natural' optimality condition is the maximization of the sum of the coefficients of determination over the clusters, weighted with cluster sizes, that is:

$$\bar{R}^2_h = \left(\sum_{c=1}^{\hat{C}} \frac{n_{c,\hat{C}}}{n} R^2_{c,\hat{C}} | \hat{P}_h \right), h = 1, 2, \ldots, m,$$

where $R^2_{c,\hat{C}}$, $c = 1, 2, \ldots, \hat{C}$ is the determination coefficient for the $c - th$ cluster of the partition.

In other words, we will choose

$$\hat{h} = argmax(\bar{R}^2_h). \tag{4}$$

We call Determination Coefficient Threshold Criterium (DCTC) the procedure that allows to satisfy condition (4) and, as a measure of goodness of fit of the model (1), we propose:

$$\bar{R}^2 = \max_h \left(\bar{R}^2_h \right). \tag{5}$$

3.2 *The All Substitution at the Same Time method (ASST)*

To estimate the regression parameters and the cluster membership of each unit, given C, we propose 'a modified Spath's algorithm' that improves the speed of convergence of Spath's algorithm. Starting from a $(J + 1)$ - dimensional n-sample, we consider a random starting partition of units but, differently from Spath, at each iteration, performs *all the exchanges* which cut down the objective function (2), respecting the constraint on the minimum number per cluster. We call this algorithm ''All Substitution at the Same Time'' (ASST).

The algorithm acts in accordance with the following points:

1. consider a random starting partition[1] of units in C clusters (C is given);
2. estimate model (1) parameters inside each cluster by OLS;
3. starting from the first unit of the sample, find all the *exchanges* which cut down the objective function (2), respecting the constraint (A2) on the minimum number per cluster: the model parameters will be estimated again after all these *exchanges;*

[1] In case we have additional information at our disposal, we can use them and start from a not-merely random partition.

4. repeat steps 2-3 until no more units can change cluster, that is the partition becomes stable.

3.3 *The proposed procedure*

We now illustrate, in detail, the CLR procedure.

STEP 1: I. Set $C = 1$ and estimate model (1) parameters.
II. Compute sample coefficient of determination $R^2_{1,1}$. If $R^2_{1,1} \geq \delta$, we conclude that a single linear regression relationship holds between y_i and $x'_{ij}s$. If $R^2_{1,1} < \delta$ go on to STEP 2.

STEP 2: I. Set $C = 2$ and choose a starting random bipartition, respecting condition A2.
II. Estimate the model (1) parameters.
III. Apply ASST algorithm in order to optimize the bipartition.
IV. Compute the two coefficients of determination $R^2_{1,2}$ and $R^2_{2,2}$. If both of them are greater than or equal to δ (condition (3)) , we conclude that $C = 2$ is the most appropriate number of clusters; so the optimal partition (in the sense of condition (3)) has been found. If at least one coefficient of determination is less than δ, go on to STEP 3.

STEP 3: I. Choose an initial random bipartition for the cluster whose coefficient of determination is less than δ (if there are more than one, consider only the first one), again respecting A2.
II. Estimate regression parameters in each cluster.
III. Apply ASST algorithm in order to optimize the C-partition (ASST reshuffle all the units, as the optimal C-partition is not necessarily obtained by splitting one of the clusters of the optimal $(C-1)$-partition).
IV. Compute the coefficients of determination $R^2_{1,C}$, $R^2_{2,C}$, ..., $R^2_{C,C}$ ($C = 3$ the first time we pass through STEP 3). If they are all greater than or equal to δ, we conclude that C is the most appropriate number of clusters, since the optimal partition (in the sense of condition (3)) has been found. If at least one coefficient of determination is less than δ, go back to the beginning of STEP 3, as long as $C < C_{max}$.

Figure 2 shows in detail the behavior of the procedure with reference to a simulation with 6 clusters of different sizes, 2 predictors and $\delta = 0.8$: it displays the trend of the weighted mean $\bar{R}^2$ (5) and of R^2 inside each cluster along the iterations of a single simulation run. Starting with a number of cluster $C = 2$, ASST finds the partition which satisfies condition (5).

This partition doesn't satisfy condition (3), therefore the number of cluster is increased to $C = 3$. In the same way the figure shows that condition (3) is not satisfied also for $C = 3$ and $C = 4$. With $C = 5$ we get the optimal partition in the sense of condition (4) (condition (3) is satisfied in each cluster). Moreover, the figure shows that since iteration 56 all $R^2_{c,C} > \delta$, but at the end of the simulation run, by applying

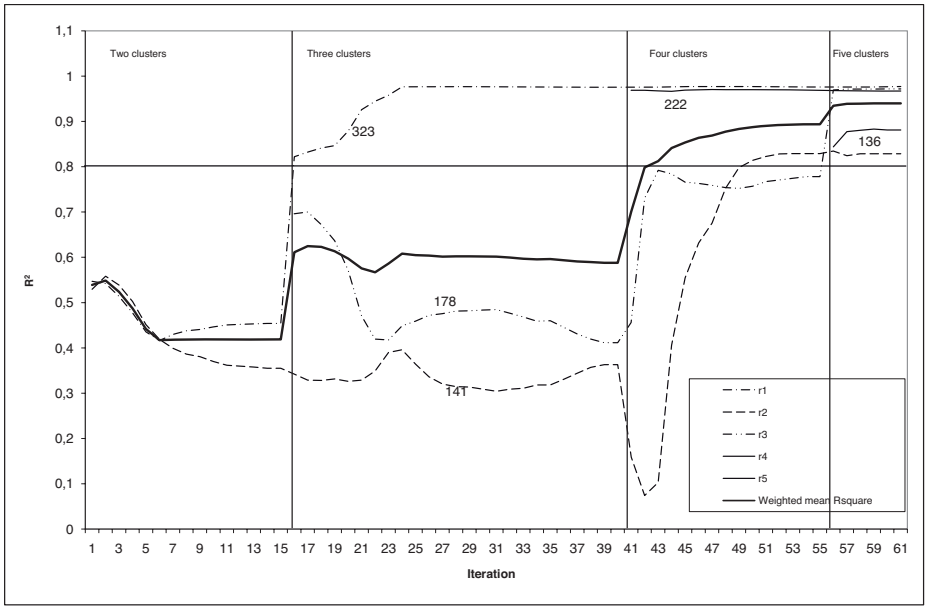

FIGURE 2. - *Coefficients of determination during the application of ASST algorithm*

ASST, not only each $R^2_{c,C} > \delta$, but also condition (5) is satisfied. Of course, for $C > 5$ it is always possible to find a partition for which condition (3) is satisfied but, since we was aiming at selecting the partition with the smallest number of groups which satisfies (3), $\hat{C} = 5$ is the solution.

The performance of the proposed procedure will be compared to DeSarbo and Cron (1988) likelihood approach (as implemented in Hennig, 2004) and to Leisch's flexible likelihood approach (implemented in the R-package Flexmix).

4. MONTECARLO SIMULATION

In order to study the performance of the procedure, a complete $2x3^2$ factorial design (Table 1) has been considered. For each parameter level set, 5 response values of the dependent variable were randomly generated and 5 different starting partitions were considered for the STEP 2 of the procedure described above; the complete design was made of 450 runs, each considering a sample of size 1000. The results of Montecarlo simulations show that the approach proposed in this paper allows, on one hand, to identify the appropriate minimal dimension C of the partition (which satisfies (3)) and, on the other, to identify an optimal partition that, given C, satisfies (4). Besides, as shown in Figure 2, $\bar{R}^2$ is a measure increasing with C.

4.1 *Description*

Each component β_j^c in (1) and each explanatory variable were randomly and independently generated from a Uniform distribution (Table 2) while, in order to compute the dependent variable 5 replications, the values of the random component ε in (1) were generated from a univariate Normal distribution. The simulation program was written in Mathematica (Wolfram-Research, 2005). The same data-sets have also been used to study the performance of DeSarbo88 mixture of linear regression (as implemented in Hennig, 2004) and Leisch (2004b) R-package FlexMix, a general framework for fitting discrete mixtures of regression models which uses EM algorithm for parameter estimation.

TABLE 1. - *Simulation parameter levels*

Cluster sizes	*N. of clusters (C*)*	*N. of predictors (J)*
a: same size	2	2
b: arith. progression	4	4
	6	6

TABLE 2. - *Random distribution for coefficients and explanatory variables*

Coefficient	*Distribution*	*Explanatory variable*	*Distribution*
β_0	U[-100;100]		
β_1	U[-50;50]	x_1	U[0;20]
β_2	U[-10;10]	x_2	U[10;14]
β_3	U[-5;5]	x_3	U[0;100]
β_4	U[-5;5]	x_4	U[0;40]
β_5	U[-10;10]	x_5	U[10;20]
β_6	U[-5;5]	x_6	U[30;60]

Algorithm outline:

1. generate parameters β_j^c;
2. generate data, that is ε, x and y according to model (1);
3. verify whether a single regression model (namely a single cluster) provide a good coefficient of determination (greater than or equal to a fixed δ). If so, consider $\widehat{C} = 1$ as *optimal solution and skip to step 9;*
4. generate a random starting partition in two clusters;
5. estimate model (5) parameters inside each cluster, by OLS;
6. optimize the partition by applying ASST algorithm;
7. compute the coefficient of determination in each cluster. If one (at least) of these coefficients is lower than δ, divide the corresponding cluster (only the first one

if more) into two random groups and go to step 5, otherwise save this optimal partition;
8. repeat steps 4-7 for 5 times (to check the influence of the random starting partition);
9. repeat steps 2-8 for 5 times regenerating the random component ε;
10. repeat steps 1-9 for each parameter level set.

At the end of step 7 an *optimal solution is found.*

The measures collected at the end of each trial are:
i. the estimated number of clusters $\widehat{C}$;
ii. the coefficient of determination of the regression model in each cluster;
iii. the size of each cluster together with cluster membership of each unit (estimated partition);
iv. the number of iterations to get the solution.

4.2 *Results and comments*

Before discussing the results, let's make a comment. It can happen that fewer clusters (than the actual simulated ones) guarantee a good fit. For example, the left side of Figure 3 shows the true classification of the units: four clusters should be estimated ($\widehat{C}$=4), with coefficient of determination $R_1^2 = 0.9721$, $R_2^2 = 0.9751$, $R_3^2 = 0.9787$ and $R_4^2 = 0.9805$; nevertheless Figure 3b shows that two regression relationships, with coefficients of determination $R_1^2 = 0.8434$ and $R_2^2 = 0.8394$, guarantee a good fit with $\delta \simeq 0.8$. This is not to be considered a misclassification, but only the attempt to reduce an ''overfitting'', according to the aim of finding the minimum number of clusters $\hat{C}$ (that is of linear regression relationships) that fit the

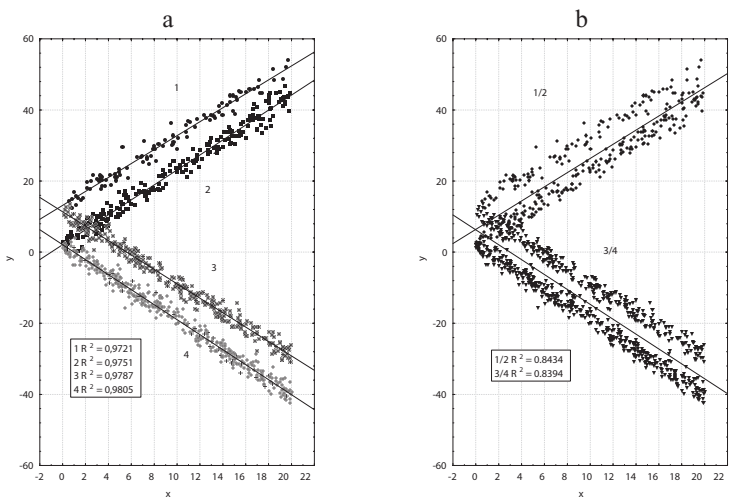

FIGURE 3. - *Example of possible misclassification*

TABLE 3. - *Percentage of success and overestimate (in italics) in estimating the number of clusters $\hat{C}$, by cluster sizes, real number of clusters C∗ and number of regressors (I: De Sarbo & Cron, II: Leisch, III: DCTC.)*

cluster sizes			$\hat{C}=2$			$\hat{C}=3$			$\hat{C}=4$			$\hat{C}=5$			$\hat{C}=6$			$\hat{C}=7$		
	C*	J	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
a	2	2	96	100	88	*4*		*12*												
		4	100	100	92			8												
		6	100	100	100															
	4	2		12	40		44	28	72	32	32	*20*	*12*		8					
		4						52	64	92	44	*24*	*8*	*4*	*8*			*4*		
		6						20	76	96	60	*8*	*4*	*16*	*16*		4			
	6	2					4	44		56	20		36	32	96	4	4	4		
		4									24		28	36	60	60	20	*40*	*12*	*20*
											16		24	8	64	60	68	*36*	*8*	*16*
b	2	2	100	100	100															
		4	100	100	100															
		6	96	100	96			*4*				*4*								
	4	2			60			20	60	88	20	*40*	*12*							
		4			16		4	60	60	92	24	*24*	*4*		*12*			*4*		
		6						16	88	100	80	*12*		*4*						
	6	2						24		48	48		52	28	92			*8*		
		4			16			4		20	12		80	40	64		20	*36*		*8*
		6						6		40	38		28	30	80	32	14	*20*		*12*

TABLE 4. - *Weigthed mean coefficient of determination (%), by cluster sizes, real number of clusters C∗ and number of regressors (I: De Sarbo & Cron, II: Leisch, III: DCTC.)*

cluster sizes			$\hat{C}=2$			$\hat{C}=3$			$\hat{C}=4$			$\hat{C}=5$			$\hat{C}=6$			$\hat{C}=7$		
	C∗	J	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
a	2	2	94	93	94	96		94												
		4	94	94	93			93			96									
		6	94	94	94															
	4	2		82	86		93	92	94	94	90	93	95		92					
		4						91	96	96	95	96	97	95	96			95		
		6						87	93	93	92	91	94	91	93		94			
	6	2					89	89		91	91		94	94	94	94	95	78		
		4									86		94	91	95	96	95	84	96	83
		6									90		94	93	95	96	93	82	96	87
b	2	2	93	93	94															
		4	93	93	94															
		6	91	91	91			91				90								
	4	2			86			93	95	96	92	96	95							
		4			84		95	94	96	97	97	97	96		96			63		
		6						90	96	96	95	96		96						
	6	2						90		94	93		94	94	95			84		
		4			85			92		97	89		93	92	95		94	84		92
		6						82		89	92		95	95	94	94	92	89		87

data in a good way and satisfies (3). Conversely, it is also important to control the number of overestimations. According to this, the assessment of the actual distance among the final regression models could be considered (this will be the object of a further development of this research).

Table 3 shows the percentage of success and overestimate (*in italics*) in estimating the number of clusters, with the three approaches:

I DeSarbo and Cron (1988) likelihood approach, where the number of clusters is estimated by BIC, as implemented in the R-package FPC;
II Leisch flexible likelihood approach (Leisch, 2004a), with number of clusters estimated by BIC;
III Determination Coefficient Threshold Criterion (DCTC), proposed in this paper.

The whole simulation plan shows that overfitting is low for both the II (3.4%) and the III (4%) approaches, while the I procedure overestimates the number of clusters in 16% of runs.

Table 4 shows the weighted mean coefficients of determination over all the clusters. As it can be seen, all values are very high. Actually, with $C = 4$, $J = 4$ and clusters of the same size, the I approach finds (fortunately only once) $\hat{C} = 7$ and $\bar{R}^2 = 0.63$.

Approaches II and III get comparable solutions but, as Table 5 shows, they differ in the number of iterations (steps) to converge. We notice that, for similar numbers of the explanatory variables, clusters and their size, the number of iterations with the II approach is always much greater than the III one's. Anyhow, in general, approach II needs up to 200 iterations while the III one up to 58 (Table 6).

The random starting partition, as well as the number of the explanatory variables, the number of clusters and their size, do not affect significatively the performance of the algorithms.

Finally, Figure 2 shows how ASST algorithm guarantees for condition (4) to be satisfied (as anticipated in Section 3.3). It means that, among the C-size partitions that satisfy condition (3), we will select the one with the maximum weighted (by the cluster size) average coefficient of determination. The figure shows what happens to the coefficients of determination during a single simulation: the example refers to $J = 2$ and 6 clusters with unequal sizes. The estimated number of cluster is $\hat{C} = 5$, as the best solution with $\hat{C} = 4$ does not satisfy condition (3). While the coefficient of determination in some cluster can also decrease during the application of the algorithm (due to the random choice of the bipartition), the weighted mean coefficient of determination over the clusters (4) only grows.

5. AN APPLICATION TO REAL DATA

Besides the simulated data, we compared the three methods also by means of their application to a real data-set widely analyzed in literature. Figure 4 shows the ethanol data-set (Cook and Weisberg, 1994), presented also in (Pena *et al.*, 2003) which

TABLE 5. - *Mean number of iterations to converge (II: Leisch, III: DCTC)*

cluster sizes			$\hat{C}=2$		$\hat{C}=3$		$\hat{C}=4$		$\hat{C}=5$		$\hat{C}=6$		$\hat{C}=7$	
	C*	J	II	III	II	III	II	III	II	III	II	III	II	III
a	2	2	44	12		25								
		4	40	11		21								
		6	36	12										
	4	2	17	13	51	16	65	24	76					\cr a
		4	4				20	41	21	90	24			
		6				23	33	17	81	29		27		
	6	2		74	20	53	27	80	19	145	23			
		4						17	40	30	48	35	67	11
		6						31	42	41	49	22	69	15
b	2	2	13	13										
		4	14	12										
		6	35	12		17								
	4	2		18		19	50	11	92					
		4		19	200	16	40	20	57					
		6				22	28	20		26				
	6	2				13	69	19	68	10				
		4		10		25	49	14	49	24		25		15
		6				52	39	25	64	17	59	41		10

TABLE 6. - *Frequency distribution of the number of iterations to converge (II: Leisch, III: DCTC)*

	II	III
0-15	97	247
6-30	82	154
31-45	127	41
46-60	62	8
61-100	50	0
101-150	20	0
151-200	12	0

relates the equivalence ratio, that is the richness of the air-ethanol mix in an engine, E, against the concentration of nitrogen oxide plus the concentration of nitrogen dioxide (normalized by the work of the engine), NO_2. The drawings clearly indicate two different dependency relationships. The procedure proposed in this paper perfectly finds the two linear regression functions and classifies the 88 units of the data-set, as shown in Figure 4 on the top-right. The figure on the top-left shows the partition got by Leisch's flexible likelihood approach; the solution differs from DCTC's one only in the classification of a few units at the intersection of the two lines, but FlexMix needs about 27 iterations to converge, against 15 of DCTC. Really worse solutions are obtained by DeSarbo and Cron likelihood approach, which usually finds more than two clusters, as shown in the bottom of Figure 4.

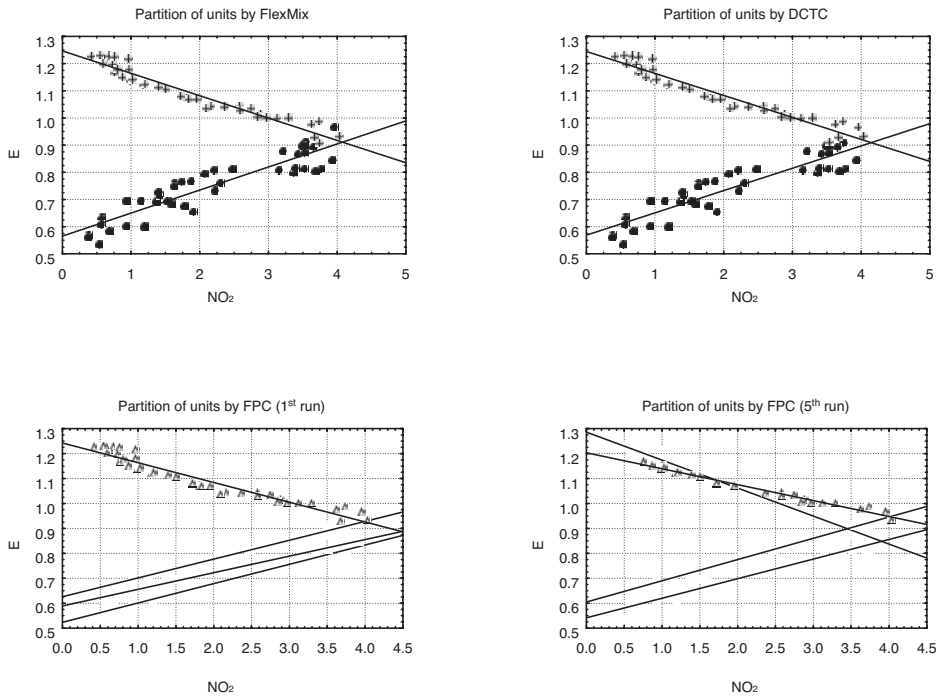

FIGURE 4. - *Cluster estimates for the ethanol data by flexible likelihood (FlexMix), DCTC and likelihood (FPC) approach*

6. CONCLUSIONS

In this paper, in a context of OLS approach, we have proposed a procedure for the estimation of a CLR model, using an algorithm based on a threshold criterion of the determination coefficient for each cluster (DCTC), in order to identify the appropria-

te number of clusters, and a modified Spath's algorithm (ASST) for estimating the cluster membership of each sample unit.

In the paper it is shown, by simulation and an application to real data, that this procedure outperforms, in speed of convergence, the Leisch's flexible likelihood approach (implemented in the R-package FlexMix) and the DeSarbo and Cron's likelihood approach (as implemented by Hennig in the R-package FPC), both for the speed of convergence and in reducing ''overfitting''. The simulation results, as well as an application to a real data-set, show that the procedure presented satisfies the optimality condition, consisting in the maximization of the sum of the coefficients of determination over the clusters, reduces overfitting (with respect to DeSarbo and Cron, 1988) and converges faster than Leisch's algorithm (Leisch, 2004a).

ACKNOWLEDGEMENTS

Work partially supported by a University of Palermo grant (ex 60%).

RIASSUNTO

In un contesto di regressione lineare, una eterogeneità dei dati suggerisce spesso l'uso di un modello Clusterwise Linear Regression (CLR). La stima di tale modello implica, fra l'altro, la stima dell'appropriato numero di cluster insieme con l'appartenenza di ogni unità ad un cluster: gli approcci a tale problema presenti in letteratura sono basati, essenzialmente, sul criterio dei Minimi Quadrati Ordinari (OLS) e sul criterio di verosimiglianza. In questo lavoro, nel contesto di un approccio OLS, viene proposta una procedura di stima del suddetto modello basata sulla ''applicazione congiunta'' di due algoritmi: un algoritmo caratterizzato da un vincolo di soglia minima del coefficiente di determinazione associato ad ogni cluster, per individuare l'appropriato numero di cluster, e un algoritmo di Spath modificato per stimare il cluster di appartenenza di ogni unità campionaria. La procedura soddisfa una condizione di ottimalità che consiste nella massimizzazione della somma dei coefficienti di determinazione dei cluster stimati. Sia il piano di simulazione che l'applicazione ad un data-set reale (the ethanol data-set (Cook and Weisberg, 1994)) considerati nel lavoro, mostrano che la procedura di stima proposta nel contesto dell'approccio OLS ha una migliore performance rispetto ad ''alcuni'' algoritmi, comunemente usati in letteratura e basati su un approccio di verosimiglianza, sia per la velocità di convergenza, sia in termini di riduzione della sovrastima del numero di cluster.

REFERENCES

Cleaver G., Wedel M. (2001). Identifying random-scoring respondents in sensory research using finite mixture regression models. *Food Quality and Preference*, **12**, 373-384.

Cook R.D., Weisberg S. (1994). *An introduction to regression graphics*. John Wiley & Sons.

DeSarbo W.S., Cron W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**(2), 249-282.

Hennig C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, **17**, 273-296.

Hennig C. (2003). Clusters, outliers and regression: fixed point clusters. *Journal of Multivariate Analysis*, **86**, 183-212.

Hennig C. (2004). R package fpc: Fixed point clusters, clusterwise regression and discriminant plo [Computer software manual]. Available from http://www.math.uni-hamburg.de/home/hennig.

Jajuga K. (1986). Linear fuzzy regression. *Fuzzy Sets and Systems*, **20**, 343-353.

Leisch F. (2004a). Flexmix: a general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**(8), 1-18.

Leisch F. (2004b). R package exmix: Flexible mixture modeling. version 1.1-0 [Computer software manual]. Available from http://www.ci.tuwien.ac.at/ leisch/FlexMix.

McLachlan G.J., Basford K.E. (1987). *Mixture models*. Marcel Dekker.

Pena D., Rodriguez J., Tiao G.C. (2003). Identifying mixtures of regression equations by the sar procedure. In *Proceedings of the Seventh Valencia International Meeting* (pp. 327-348).

Quandt R.E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **53**, 873-880.

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

Viele K., Tong B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, **12**, 315-330.

Wedel M., DeSarbo W. (2002). Market segment derivation and profiling via a finite mixture model framework. *Marketing Letters*, **13**(1), 17-25.

Wolfram-Research (2005). Mathematica 5 [Computer software manual].