Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n1p13

A new procedure of regression clustering based on Cook's D

By Jayakumar, Sulthan

Published: 26 April 2015

Electronic Journal of Applied Statistical Analysis
Vol. 08, Issue 01, April 2015, 13-27
DOI: 10.1285/i20705948v8n1p13

A new procedure of regression clustering based on Cook's D

D.S. Jayakumar*and A. Sulthan

Jamal Institute of Management, Tiruchirappalli, India

Published: 26 April 2015

Clustering is an extremely important task in a wide variety of application domains especially in management and social science research. Usually, clustering methods work based on some distance metric among the observation or it may use Co-variance and correlation structure among the variables. If all the given variables depend on a single variable, then the procedure of clustering the observations is said to be regression clustering. In this paper, an iterative procedure of regression clustering method was proposed by using the famous Cook's D distance. At first, the Cook's D distance should be calculated for the entire sample, then fix a Cut-off distance proposed by (Bollen and Jackman, 1990) as $4/(n-K-1)$.The authors fixed this Cut-off point as structural break in the sample, observations above the cut-off are considered as Influential which are grouped as Influential cluster and repeat the same procedure for the remaining observations, until there are no influential observations in the last cluster. At each iteration, Chow's F-test (1960) was used to check the discrimination between the influential cluster and the non-influential cluster. Moreover, control charts also used to graphically visualizes the iterations and the clustering process .Finally Chow's test of equality of several regression equation helps firmly to establish the cluster discrimination and validity. This paper employed this procedure for clustering 220 customers of a famous four-wheeler in India based on 19 different attributes of the four wheeler and its company.

keywords: Distance metric, Correlation structure, Cook's distance, Structural Break, Influential observation, Influential cluster, Chow's F- test

*Corresponding author: samjaya77@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

The regression quality cannot be converted into any kind of distance measure between the elements of the data set, though such a measure is necessary for classical clustering algorithms (Duran and Odell, 1974). Clustering is a regression of individual outcomes on explanatory variables of which some are observed on a more aggregate level (Moulton, 1986) ,Moulton (1990) . As in the general setting of model-based clustering, there are also two different approaches for regression clustering in the literature. One is the random partition regression clustering. The discussion can be found in (DeSarbo and Cron, 1988), Quandt and Ramsey (1978) among others. Another one is the fixed partition regression clustering discussed in (Bock, 1969) (Bock, 1996) (Späth, 1979) (Späth, 1982). Multiple Linear Regression model relaxes the homoscedasticity assumption and allows the error terms to be heteroscedastic and correlated within groups are so-called clusters. (Späth, 1979). Besides the regression clustering approach, many clustering methodologies exist in the literature such as outlier clustering and some authors may apply clustering as a method to identify the outliers.Outliers are the set of objects that are considerably dissimilar from the remainder of the data (Jiawei and Kamber, 2001). Outlier detection is an extremely important problem with a direct application in a wide variety of application domains, including fraud detection (Bolton and Hand, 2002), identifying computer network intrusions and bottlenecks (Lane and Brodley, 1999), criminal activities in e-commerce and detecting suspicious activities (Chiu and Fu, 2003). Different approaches have been proposed to detect outliers, and a good survey can be found in (Knox and Ng, 1998) (Knorr et al., 2000) (Hodge and Austin, 2004).Clustering is a popular technique used to group similar data points or objects in groups or clusters (Jain and Dubes, 1988) and it is an important tool for outlier analysis. Several clustering-based outlier detection techniques have been developed. Most of these techniques rely on the key assumption that normal objects belong to large and dense clusters, while outliers form very small clusters (Loureiro et al., 2004) (Niu et al., 2007).It has been argued by many researchers whether clustering algorithms are an appropriate choice for outlier detection. For example, in (Zhang and Wang, 2006), the authors reported that clustering algorithms should not be considered as outlier detection methods. This might be true for some of the clustering algorithms, such as the k-means clustering algorithm (MacQueen et al., 1967). This is because the Cluster means produced by the k-means algorithm is sensitive to noise and outliers (Van der Laan et al., 2003).Similarly, that the case is different for the Partitioning Around Medoids (PAM) algorithm Kaufman and Rousseeuw (1990). PAM attempts to determine k partitions for n objects. The algorithm uses the most centrally located object in a cluster (called medoid) instead of the cluster mean. PAM is more robust than the k-means algorithm in the presence of noise and outliers. This is because the medoids produced by PAM are robust representations of the cluster centers and are less influenced by outliers and other extreme values than the means (Van der Laan et al., 2003) Kaufman and Rousseeuw (1990) (Dudoit and Fridlyand, 2002). Furthermore, PAM is a data-order independent algorithm (Hodge and Austin, 2004), and it was shown in (Bradley et al., 1999) that the medoids produced by PAM provide better class separation than the means produced by the k-means clustering algorithm. PAM starts by selecting an initial set of medoids (cluster centers) and iteratively replaces each one of the selected medoids by one of the none-selected medoids in the data set as long as the sum of dissimilarities of the objects to their closest medoids is improved. The process is iterated until the criterion function converges. These approaches can be easily implemented which includes the PAM, such as (Kaufman and Rousseeuw, 1990) (Ng and Han, 2002) (Zhang and Couloigner, 2005).

As discussed in (Loureiro et al., 2004) (Niu et al., 2007) (Zhang and Wang, 2006), there is no single universally applicable or generic outlier detection approach. Therefore,many approaches have been proposed to detect outliers. These approaches can be classified into four major categories based on the techniques used (Zhang and Wang, 2006), which are: distribution-based, distance-based, density-based and clustering-based approaches. Distribution-based approaches (Hawkins, 1980) (Barnett and Lewis, 1994) (Rousseeuw and Leroy, 2005) develop statistical models (typically for the normal behavior) from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects that have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications (Zhang and Wang, 2006).In the distance-based approach (Knox and Ng, 1998) (Knox and Ng, 1998) (Ramaswamy et al., 2000) (Angiulli and Pizzuti, 2005), outliers are detected as follows. Given a distance measure on a feature space, a point q in a data set is an outlier with respect to the parameters M and d, if there are less than M points within the distance d from q, where the values of M and d are decided by the user. The problem with this approach is that it is difficult to determine the values of M and d. Density-based approaches (Breunig et al., 2000)(Papadimitriou et al., 2003) compute the density of regions in the data and declare the objects in low dense regions as outliers. In (Breunig et al., 2000), the authors assign an outlier score to any given data point, known as Local Outlier Factor (LOF), depending on its distance from its local neighborhood. A similar work is reported in (Papadimitriou et al., 2003).Clustering-based approaches (Loureiro et al., 2004) (Loureiro et al., 2004) (Gath and Geva, 1989) (Van Cutsem and Gath, 1993) (Jiang et al., 2001) (Acuna and Rodriguez, 2004), consider clusters of small sizes as clustered outliers. In these approaches, small clusters (i.e., clusters containing significantly less points than other clusters) are considered outliers. The advantage of the clustering-based approaches is that they do not have to be supervised. Moreover, clustering-based techniques are capable of being used in an incremental mode (i.e., after learning the clusters, new points can be inserted into the system and tested for outliers). (Van Cutsem and Gath, 1993) present a method based on fuzzy clustering. In order to test the absence or presence of outliers, two hypotheses are used. However, the hypotheses do not account for the possibility of multiple clusters of outliers. (Jiang et al., 2001) presented a two-phase method to detect outliers. In the first phase, the authors proposed a modified k-means algorithmto cluster the data, and then, in the second phase, an Outlier-Finding Process (OFP) is proposed. The small clusters are selected and regarded as outliers by using minimum spanning trees. In (Loureiro et al., 2004) clustering methods have been applied. The key idea is to use the size of the resulting clusters as indicators of the presence of outliers. The authors use a hierarchical clustering technique. A similar approach was reported in (Almeida et al., 2007). (Acuna and Rodriguez, 2004) performed the PAM algorithm followed by the technique (henceforth, the method will be termed PAMST). The separation of a cluster A is defined as the smallest dissimilarity between two objects; one belongs to Cluster A and the other does not. If the separation is large enough, then all objects that belong to that cluster are considered outliers. In order to detect the clustered outliers, one must vary the number k of clusters until obtaining clusters of small size and with a large separation from other clusters. In (Yoon et al., 2007), the authors proposed a clustering- based approach to detect outliers. The K-means clustering algorithm is used. As mentioned in (Lane and Brodley, 1999)(Laan, 2003), the k-means is sensitive to outliers, and hence may not give accurate results. Later,Jayakumar and Thomas (2013) proposed a new approach of clustering the sample observation based on multivariate outlier detection by using the T-square distance. In this paper, a new method of regression clustering was proposed based on Cook's D distance and it is discussed in the subsequent sections.

2 Proposed Approach

Here, we proposed a new approach of regression clustering based on Cooks distance. In statistics, Cooks distance is a measure introduced by Cook, R.Dennis, and it is a measure of change in the regression co-efficients that would occur if an observation was omitted, thus revealing which observations are the most influential in affecting the regression equation. It is affected by both the case being an outlier on Y-space and on the set of the predictors in X-space. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a large Cook's distance are considered to merit closer examination in the analysis. Formally, the cooks D is defined as

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{(-i)})^T (X^T X)(\widehat{\beta} - \widehat{\beta}_{(-i)})}{(p+1)\widehat{\sigma_e^2}} \tag{1}$$

From (1) where $\widehat{\beta}_{(-i)}$ is the vector of estimated regression co-effcients with the i^{th} observation deleted, p is the no.of predictors and $\widehat{\sigma_e^2}$ is the estimated error variance for the full dataset. Removing the i^{th} observation should keep $\widehat{\beta}_{(-i)}$ close to $\widehat{\beta}$, unless the i^{th} observation is an influential observation. Based on the above said distance measure, first, calculate the Cook's distance from (1) by fitting a multiple linear regression model using OLS for the n observations based on p independent variables, where $\widehat{\beta}$. Secondly, fix a cut-off distance by using (Bollen and Jackman, 1990) $4/(n-K-1)$, observations above the cut-off are considered as Influential cluster and named as Influential cluster-1.Repeat the same procedure for remaining observations excluding the Influential observations in cluster-1. Repeat the process, until there are no influential observations in the last cluster. The basic structure of the proposed method is as follows:

Step1-Calculate the Cook's distance from (1) by fitting a multiple regression model for n observations based on p independent variables.

Step2-Identify the observations which are above the cut-off distance and consider those observations belong to Influential cluster-1.

Step3-Again fit a multiple regression model for Influential cluster 1 and for the remaining sample size. Check the equality of two regression equations by using chow's F-test.If two regression equations are equal, then stop the iteration and it shows there is no discrimination between the influential cluster-1 and the remaining sample size. If the regression equations are not equal at 5% or 1% significance level, then the influential cluster-1 is different from the remaining sample size, then continues step.4

Step4-Repeat step no.1 and 2 for the remaining observations and ascertain the Influential cluster- 2.

Step5-Continue the iteration process, until there is no influential observations in the remaining sample size.

Step6- If 'r' clusters are explored, and then scrutinize the overall discriminant validity among the clusters by fitting 'r' multiple regression equations.

Step7-Then apply Chow's F-test for checking the equality of several regression equations and if the regression equations are equal, there is no discrimination among the clusters. Similarly, if the regression equations are not equal at 5% or 1% significance level, then each cluster is different among each other and this shows we achieved the discriminant validity among the clusters.

The discriminant validity among "r" clusters (includes '(r-1)'Influential cluster and a Non-influential cluster) through the regression clustering approach and the application of Chow's F-test explores many interesting facts of the structural break in survey data. This will be discussed in the next section.

3 Chow Test and Cluster Validation

In survey data also, there often contains a structural break, due to the cluster effect or group effect of the similar observations. In order to test a structural break, we often use the Chow test, and it uses an F-test to determine whether a single regression is more efficient than two or more separate regressions involving splitting the data into several clusters. In multiple linear regression analysis, the structural break could occur at a known point Z.The point Z breaks the given sample into influential cluster and non-influential cluster. Based on the previous section, the authors adopted cut-off Cook's D as $4/(n-K-1)$ proposed by (Bollen and Jackman, 1990) as breaking point (Z) and it helps to segregate the influential observations as influential cluster and the remaining observations are non-influential. In multiple linear regression analysis, if we have just a single regression equation to fit the data points, it can be expressed as

$$Y = X\beta + e \tag{2}$$

In the second case, where there is a structural break at Z, we have two separate models, expressed as:

Influential cluster regression equation

$$-Y_1 = X_1\beta_1 + e_1 \tag{3}$$

Non-Influential cluster regression equation

$$-Y_2 = X_2\beta_2 + e_2 \tag{4}$$

This suggests that (3) applies for the observations above the break at Z, then (4) applies for the observations below the structural break at Z .If the parameters in the above same, i.e. $\beta_1 = \beta_2$, then models from (3) and (4) can be expressed as a single model as in (2), where there is a single regression. The Chow test basically tests whether the single regression or the two separate regressions fit the data best. The stages in running the Chow test are:

1. At first, run the regression using all the observation without the structural break and calculate the SSE $(\widehat{e}'\widehat{e})$. 2. Run two separate regressions on the observations, one for the influential cluster and another for the non-influential cluster, then collecting the SSE in both cases, giving SSE$_1$ $(\widehat{e_1}'\widehat{e_1})$and SSE$_2$ $(\widehat{e_2}'\widehat{e_2})$ 3. Using these three values, calculate the test statistic from the following formula:

$$F = \frac{\left(\widehat{e}'\widehat{e} - (\widehat{e_1}'\widehat{e_1} + \widehat{e_2}'\widehat{e_2})\right)/K}{\left(\widehat{e_1}'\widehat{e_1} + \widehat{e_2}'\widehat{e_2}\right)/n - 2K} \sim F(K, n-2K)$$

Where K is the estimated no.of parameters.

4. Find the critical values in the F- tables; in this case it has $F(K, n-2K)$degrees of freedom. If the test result is statistically significant at 5% or 1% level, then reject the null hypothesis $(\beta_1 = \beta_2)$ and conclude that the sample of observations having structural break. The acceptance of alternative hypothesis $(\beta_1 \neq \beta_2)$under chow's F-test helps us to drawn the discriminant validity between the Influential regression cluster and the non-influential regression cluster.

5. The overall discriminant validity among the Influential clusters can be tested if there 'r'clusters in the sample, then we have to use the same Chow's F-test under the null hypothesis $(\beta_1 = \beta_2 = \beta_3 = \beta_4 = ... = \beta_{r+1} = \beta_r)$the test statistic is given as

$$F = \frac{\left(\widehat{e}'\widehat{e} - \sum\limits_{j=1}^{r} \widehat{e_j}'\widehat{e_j}\right)/K}{\left(\sum\limits_{j=1}^{r} \widehat{e_j}'\widehat{e_j}\right)/n - rK} \sim F(k, n-rK)$$

Where K is the estimated no.of parameters and *'r'* is the no.of clusters. If the test result is statistically significant at 5% or 1% level, then accept the alternative hypothesis $(\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq ..... \neq \beta_{r+1} \neq \beta_r)$and this shows the clusters are distinctive among each other and we achieve the overall discriminant validity.

4 Results and Discussion

In this section, we investigated the effectiveness of our proposed approach on the survey data collected from the famous four wheeler users' in India .The data comprised of 19 different attributes about the four wheeler company and the data was collected from 220 four wheeler users. A well-structured questionnaire was prepared and distributed to 240 four wheeler customers and the questions were anchored at five point likert scale from 1 to 5.After the data collection is over, only 220 completed questionnaires were used for analysis. The aim of this article is to describe the proposed clustering approach not the application of the theoretical concept. The following table shows the results extracted from the analysis by using SAS JMP V10 and IBM SPSS V22.

Table 1: Iteration Summary for Identification of Influential clusters

Iteration	**Cut-off Cook's D** $4/(n-K-1)$	**Pooled Sample**		**Observations < Cut-off Cook's D**		**Observations > Cut-off Cook's D**		**Influential cluster size**	**Chow's F test**	
		(n)	$\widehat{e}'\widehat{e}$	***(n_1)***	$\widehat{e_1}'\widehat{e_1}$	***(n_2)***	$\widehat{e_2}'\widehat{e_2}$		**F-ratio**	***d.f***
										(K,n-2K)
1	0.0199	220	21.7215	196	4.93543	24	1.68965	**24**	21.827*	**(19,182)**
2	0.0226	196	4.93543	177	1.28598	19	0	**19**	23.599*	**(19,158)**
3	0.0253	177	1.28598	155	.09959	22	.28512	**22**	17.138*	**(19,139)**
4	0.03448	**155**	0	-	-	-	**-**	**-**	-	**-**

K (no. of estimated parameters) =19 *p-value <0.01 d.f- degrees of freedom

Table-1 visualizes the iteration summary of the identification of the influential clusters by using the Cook's D distance. At first iteration, 220 observations, a dependent and 18 independent variables were used to fit a linear multiple regression model and calculate the Cook's D distance for all observation. Among 220 observations, the Cook's D for 196 observations were below the Cut-off D (0.0199) and the remaining no. of observations (24) are above the cut-off.Therefore, we consider the 24 observations as Influential cluster-1.The result of the Chow's F-test in this iteration is also significant at 1% level and this shows there is some discrimination between the influential cluster-1 and the remaining non-influential observation. Then repeat the iteration process to the next stage by fitting a linear multiple regression model and calculate the Cook's D distance based on 196 observations (220-24) for the same variables in iteration 2.Likewise, if we continue the iteration process for the remaining stages, the iteration reached the limit in the 4^{th} step with 155 observations as non-influential cluster. At the iteration no.4, the result of the linear multiple regression model for 155 observations reveals the sum of the squared error is 0, and then it is impossible to calculate the Cook's D, So stop the iteration process. Hence based on 4 iterations, we identified three different Influential cluster at 1% significance level with ($\boldsymbol{n}$=24), ($\boldsymbol{n}$=19), ($\boldsymbol{n}$=22) and ($\boldsymbol{n}$=155) observations as non-influential cluster respectively. The iteration and identification of influential clusters were explained with the help of the following control charts.

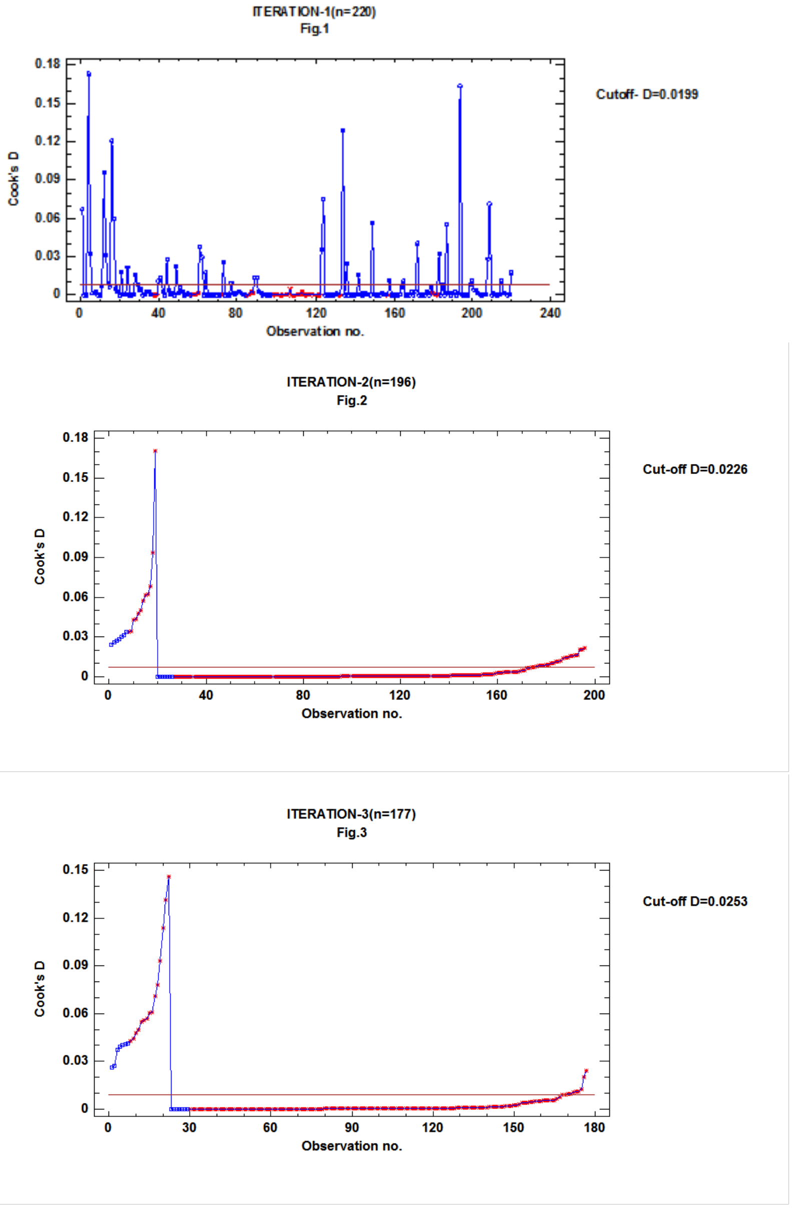

Figure 1: Transition in model selection at 50% Inflation level of error variance

Table 2: Iteration Summary for Identification of Influential clusters

Chow's F test	**Pooled Sample**		**Influential cluster-1**		**Influential cluster-2**		**Influential cluster-3**		**Non Influential cluster**		**F-ratio**	***d.f*** ***(K,n-2K)***
	(n)	$\widehat{e}'\widehat{e}$	*(n_1)*	$\widehat{e_1}'\widehat{e_1}$	*(n_2)*	$\widehat{e_2}'\widehat{e_2}$	*(n_3)*	$\widehat{e_3}'\widehat{e_3}$	*(n_4)*	$\widehat{e_4}'\widehat{e_4}$		
	220	21.7215	24	1.6896	19	0	22	.2851	155	0	75.785*	(19,182)

K (no. of estimated parameters) =19 *p-value <0.01 d.f- degrees of freedom

Table-2 describes the results of the Chow's F-statistic of several clusters which helps us to finalize the discriminant validity among the influential clusters. The result of the test statistic confirms that the fitted multiple regression equation based on each cluster are significantly different among the influential clusters at 1% significant level. This show the clusters are different in the regression plane and we achieved the overall discriminant validity among the clusters. The following 2-D and 3-D surface plot visualizes the summary of membership of each observation in influential cluster as well as in non-influential cluster based on Cook's D.

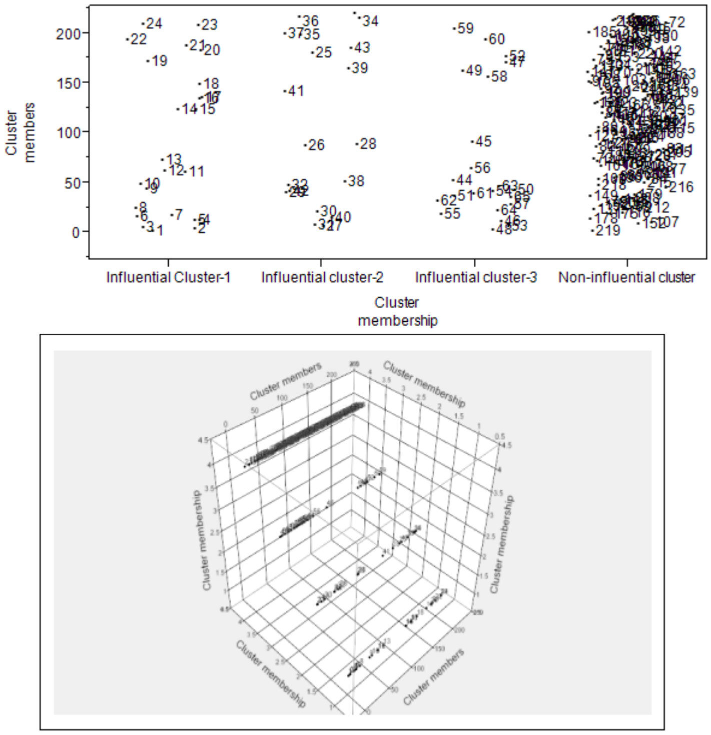

Figure 2: Membership of observations in influential and Non-influential clusters

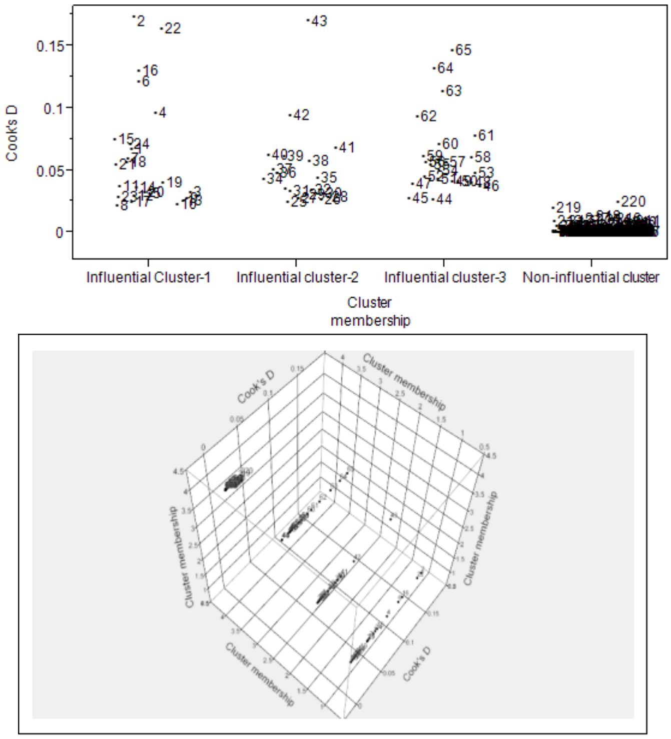

Figure 3: Membership of observations in influential and Non-influential clusters based on Cook's D

5 Conclusion

In this paper a new method of regression clustering was proposed based on influential observations. Though several regression clustering procedures available in the literature, the proposed technique gives a unique idea to cluster the sample observations in a survey study based on the influential observations. The feature of the proposed clustering technique was elaborately discussed and the authors also highlighted the application of the technique in a survey research. Based on the results derived, the proposed technique gives more insights to the researcher to cluster the sample observation and introduce the concept of structural break in survey data due to the cluster effect or group effect of the similar observations. Finally the authors enlighten an idea for further research by using step-wise regression procedure to identify the influential clusters with different sub-sets of independent variables and this issue will raise a new exploration of hybrid chow test which can used to test the equality of several regression equations with different sub-set of regressors.

References

Acuna, E. and Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez.*

Almeida, J., Barbosa, L., Pais, A., and Formosinho, S. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 87(2):208–217.

Angiulli, F. and Pizzuti, C. (2005). Outlier mining in large high-dimensional data sets. *Knowledge and Data Engineering, IEEE Transactions on*, 17(2):203–215.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*, volume 3. Wiley New York.

Bock, H. (1969). The equivalence of two extremal problems and its application to the iterative classification of multivariate data. In *WorkshopMedizinische Statistik.*

Bock, H.-H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. *Clustering and classification*, pages 377–453.

Bollen, K. A. and Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. *Modern methods of data analysis*, pages 257–291.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, pages 235–249.

Bradley, P. S., Fayyad, U. M., and Mangasarian, O. L. (1999). Mathematical programming for data mining: formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, pages 93–104. ACM.

Chiu, A. L.-m. and Fu, A.-C. (2003). Enhancements on local outlier detection. In

Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International, pages 298–307. IEEE.

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

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7):research0036.

Duran, B. S. and Odell, P. L. (1974). Cluster analysis, a survey, volume 100 of lectures notes in economics and mathematical systems.

Gath, I. and Geva, A. B. (1989). Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions. *Pattern Recognition Letters*, 9(2):77–86.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

Jain, A. and Dubes, R. (1988). Algorithms for clustering data. 1988. *Michigan State University: Prentice Hall*.

Jayakumar, G. D. S. and Thomas, B. J. (2013). A new procedure of clustering based on multivariate outlier detection. *Journal of Data Science*, 11(1):69–84.

Jiang, M.-F., Tseng, S.-S., and Su, C.-M. (2001). Two-phase clustering process for outliers detection. *Pattern recognition letters*, 22:691–700.

Jiawei, H. and Kamber, M. (2001). Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, 5.

Kaufman, L. and Rousseeuw, P. (1990). Finding groups in data, 1990. *New York*.

Knorr, E. M., Ng, R. T., and Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The VLDB JournalThe International Journal on Very Large Data Bases*, 8(3-4):237–253.

Knox, E. M. and Ng, R. T. (1998). Algorithms for mining distancebased outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer.

Lane, T. and Brodley, C. E. (1999). Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security (TISSEC)*, 2(3):295–331.

Loureiro, A., Torgo, L., and Soares, C. (2004). Outlier detection using clustering methods: a data cleaning application. In *Proceedings of KDNet Symposium on Knowledge-based systems for the Public Sector*.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3):385–397.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pages 334–338.

Ng, R. T. and Han, J. (2002). Clarans: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5):1003–1016.

Niu, K., Huang, C., Zhang, S., and Chen, J. (2007). Oddc: outlier detection using distance distribution clustering. In *Emerging Technologies in Knowledge Discovery and Data Mining*, pages 332–343. Springer.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral. In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pages 315–326. IEEE.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, volume 29, pages 427–438. ACM.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John Wiley & Sons.

Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing*, 22(4):367–373.

Späth, H. (1982). A fast algorithm for clusterwise linear regression. *Computing*, 29(2):175–181.

Van Cutsem, B. and Gath, I. (1993). Detection of outliers and robust estimation using fuzzy clustering. *Computational statistics & data analysis*, 15(1):47–61.

Van der Laan, M., Pollard, K., and Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584.

Yoon, K.-A., Kwon, O.-S., and Bae, D.-H. (2007). An approach to outlier detection of software measurement data using the k-means clustering method. In *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on*, pages 443–445. IEEE.

Zhang, J. and Wang, H. (2006). Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and information systems*, 10(3):333–355.

Zhang, Q. and Couloigner, I. (2005). A new and efficient k-medoid algorithm for spatial clustering. In *Computational Science and Its Applications–ICCSA 2005*, pages 181–189. Springer.