Identification of Influential Variables by using Cluster Approach

Dr.M.ChandraSekhar Reddy
Associate Professor, Department of Statistics, College of Natural Sciences, Arba Minch University,
Arba Minch, P.O.Box: 21, Ethiopia.

Abstract

The availability of sophisticated data storage facilities at less cost make us to get huge amount of data in all the domains like manufacturing and health care. This creation of huge amount of data requires highly efficient infrastructure in terms of high end computers with high computational power. The main aim this paper was to compare clustering approach with traditional approaches of influential variable analysis. Hence, from empirical data traditional methods of finding influential variables some time might not give exact results. In that case we can go for cluster analysis and then profiling the results through scatterplots will give more results.

Keywords:-Cluster approach, Graphical approach, Influential variable.

1. Introduction

The availability of sophisticated data storage facilities at less cost make us to get huge amount of data in all the domains like manufacturing and health care. This creation of huge amount of data requires highly efficient infrastructure in terms of high end computers with high computational power. There is always great need to identify very key critical variable which we call as influential variable to take necessary action. For example, in the manufacturing domain, we need to find at least top three or four influential variables for production / output and this will help them to increase the production after the required steps. According to Miyako Sagawa in evolutionary multi-objective optimization, variation operators are crucially important to produce improving solutions, hence leading the search towards the most promising regions of the solution space. In the literature we can find large number of approaches to find the critical or influential variable identification where some of them are numerical approaches and some of them are non-numerical or graphical approaches. Graphical methods are the very first approach to find whether really makes sense to the independent variable and it is easy to implement and understand. However, there is no very common approach that fits in to all types of issues and domains. In this paper, we briefly described the current approaches for identification of critical variable and explained the use of clustering approach in this case. It starts with some basic graphical methods and then modeling methods. These results were compared with clustering approach.

2. Traditional Approaches of Influential Variable Analysis

2.1. Partial regression plot

In a simple linear regression model, a scatter plot of dependent and independent variable give some idea about the relationship between the variables. When performing a linear regression with a single independent variable, a scatter plot of the response variable against the independent variable provides a good indication of the nature of the relationship. If there is more than one independent variable, things become more complicated. Although it can still be useful to generate scatter plots of the response variable against each of the independent variables, this plot do not consider the influence of other variables that already exist in the data. Hence, partial regression plots are very useful to identify the impact of adding a new variable in to the model and gives clear picture of its influence on the dependent variable. This plot should be constructed for each variable separately in the following way:

- Construct a regression model and calculate the residuals between the dependent variable and all independent variables without including a particular variable say X_i
- Second step is to build a regression model between omitted variable X_i and all other independent variable and obtain the residuals.
- Now plot the residuals constructed in step 1 against step 2 and observe the pattern

In the literature, Velleman and Welsch explained the properties of these plots in the literature and also difference between partial regression plots and partial residual plots. Partial regression plots are most generally applied to identify the impact at individual variable considering the partial impact of other variables; whereas partial residual plots are most common the cases of in finding out the relation between just two variables only.

In the following Figure 1, we can see the partial regression plot of three variables on a single dependent variable and we can see how they impact individually.

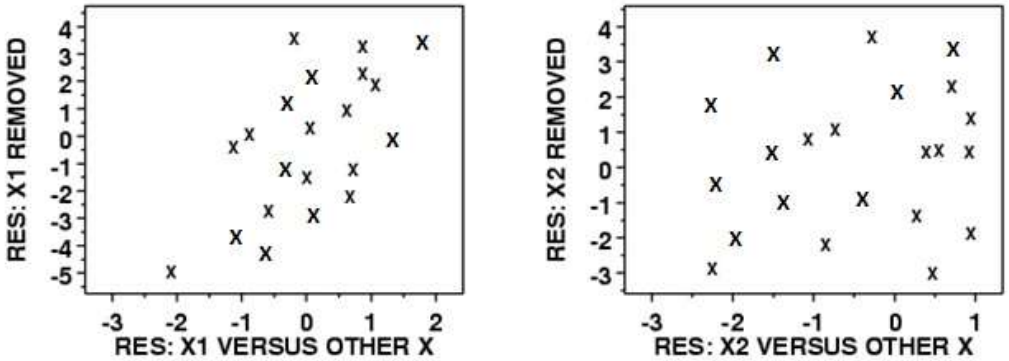


Figure1: Partial regression plot of three variables on a single dependent variable

2.2. Partial residual plot

In the usual regression model with more independent variables, it is more complicated to find the impact of individual observations and in the case one can go for partial residual plot. Given the fact of other independent variable, partial residual plot will give the relation between X and Y variables. The steps involved in constructing the partial regression plots are:

- Calculate the residual values from the usual regression model.
- find the sum of residuals and product of regression coefficient of X_i and X_i
- plot the two values calculated in the above two steps

We need to take care that these plots give a different meaning if the dependent and independent variables are highly correlated. In other words, the high correlation values cannot treat as high impact on influence variables.

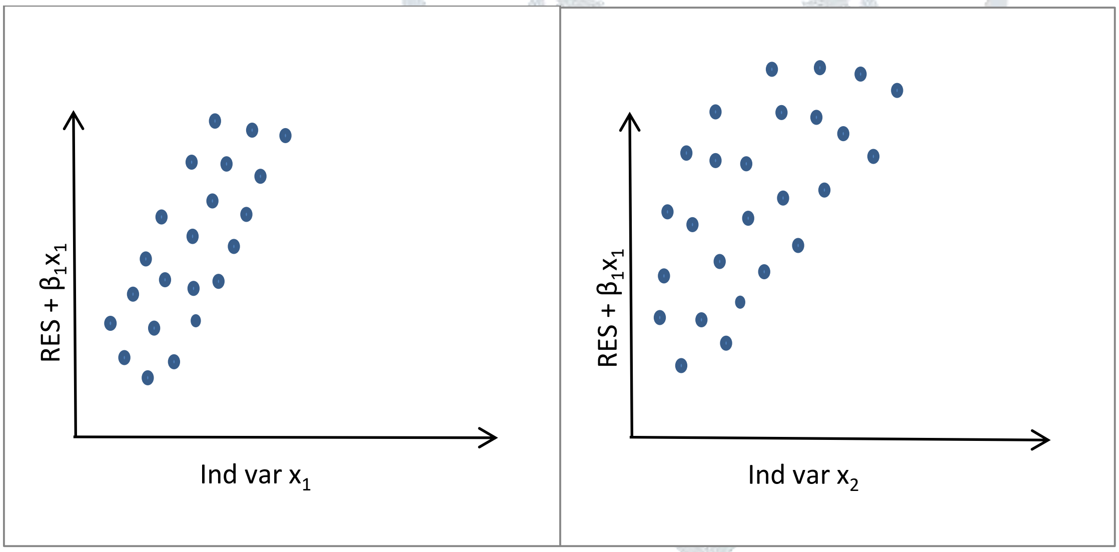


Figure 2: Partial residual plots

2.3. Influence of variables using R^2 value and p-value

In the usual regression model, we can calculate the R^2 value and find the variability explained by the independent variables in the model. By using this measure we can identify the amount of variability added by a new variable in the model. Say for example, in the first step we can calculate the R^2 value by using dependent variable Y and an independent variable X_1. Find the R^2 value from this model. Now compute the same regression model by using X_1 and X_2 on the dependent variable Y. Now the changed R^2 value indicates the influence that it gives on the Y variable. But this is just an indication only and not always true. R^2 value might change due to some other reasons as well. In addition to this, we can use the p-value of each of the independent variables also as an indication of how influence it causes to the model. At the same time, when the numbers of independent variables are large, then p-value might be slightly misleading to take the decision of influencing variables.

2.4. Other methods of finding the influential variables

There are some other methods also used in the literature to find the influencing variable and one of such methods is stepwise regression. Here each variable is added or deleted based on its influence on the dependent variable. Here we may not get any direct measure but we can say whether a particular variable is important or not.

3. Cluster based approach for influential variable analysis

The basic ideology behind clustering is grouping the similar kind of objects together based on the distance between them. In other words, all the items in the single cluster are very closely located in terms of their distance between each other. The actual separation between the points using clustering methods will tell us the property of the clusters. Say for example in Figure 3, the three clusters divided shows cases about their properties that they have in unique.

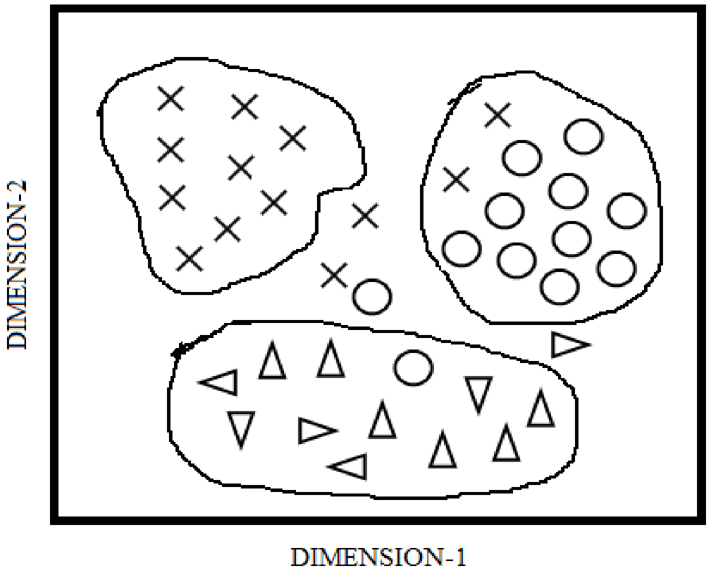


Figure 3: Three groups by clustering approach

3.1 Methods of clustering

There are several approaches that exist in the literature for doing the clustering and we discuss the most popular ones among them.

Hierarchical clustering

In these methods, a step-by-step cluster formation is carried out and it is stopped once the required numbers of clusters are formed. The first approach is agglomerative approach or bottom-up approach in which we start with each item forming its own group. Then it is slowly merged into its nearest point and so on to form clusters (see Figure 4).

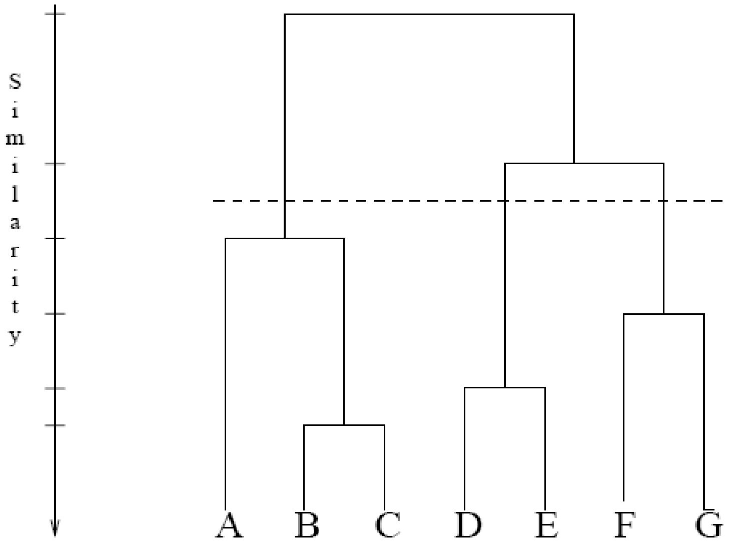


Figure 4: Agglomerative or bottom-up approach of clustering

The second approach is divisive approach or top down approach in which we start with all items in a single cluster. Then it is divided into two clusters based on the largest separation. This procedure will continue until the required clusters are formed.

K-means clustering

K-means clustering a numerical method to divide the entire group in to k clusters by minimizing the sum of squares of distances between data and corresponding cluster center point. The most commonly used distance measures are: Euclidean distance also called as 2-norm distance and Manhattan distance also call as 1-norm distance. The basic approach of k-means clustering is outlined as follows in Figure 5.

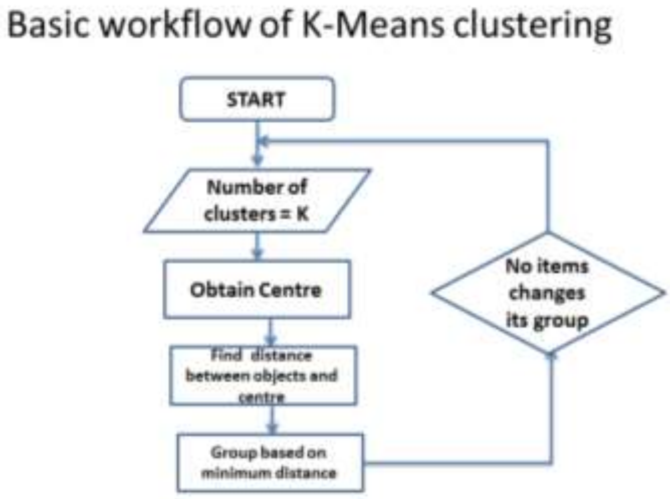

Basic workflow of K-Means clustering
START
Number of clusters = K
Obtain Centre
Find distance between objects and centre
Group based on minimum distance
No items changes its group

Figure 5: The basic approach of k-means clustering

3.2 Cluster profiling

Profiling of clusters is nothing but studying the different properties of points within clusters which can reveal the similarities between the clusters. The cluster profiling usually reveals some hidden properties of the items in a cluster such as the unique properties of cluster which cannot be identified with a naked eye. In the following section with we will describe the profiling in a detailed based on empirical data.

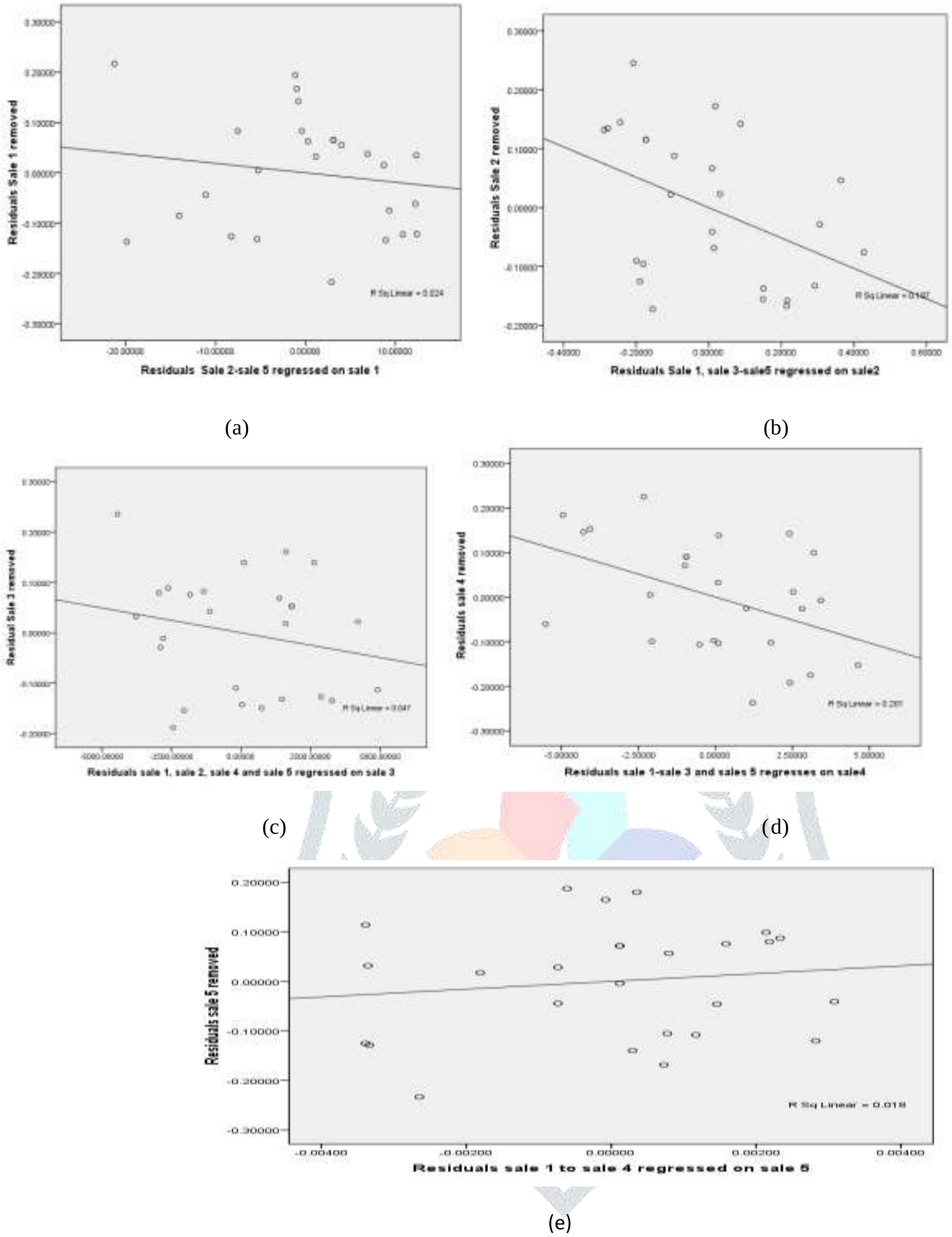
4. Results based on empirical data

For the current study for the sake demonstration we took a hypothetical super market sale data related to five commodities and the target is to find the best influential variable on performance index of the super market. We applied the multiple linear regression on the above data with performance index as the dependent variable and the results are given as follows in Table 1.

Table 1: Estimated regression parameters of the multiple linear regression models

Variable	Unstandardized $\hat{\beta}$	Std. Error($\hat{\beta}$)	Standardized $\hat{\beta}$	T value	P-value
Constant	5.246	2.020	-.153	2.597	.017
Sales 1	-.002	.003	-.440	-.701	.491
Sales 2	-.258	.120	-.192	-2.144	.045
Sales 3	-9.839E-6	.000	-.452	-.998	.330
Sales 4	-.021	.009	.134	-2.240	.037
Sales 5	7.838	12.887	-.153	.608	.550

From the Table 2, we can see that sales 2(p-value=0.045) and sales 4(p-value=0.037) are important predictors of performance indices of supermarket and they have influential on the index variable based on p-value approach.



(a)

(b)

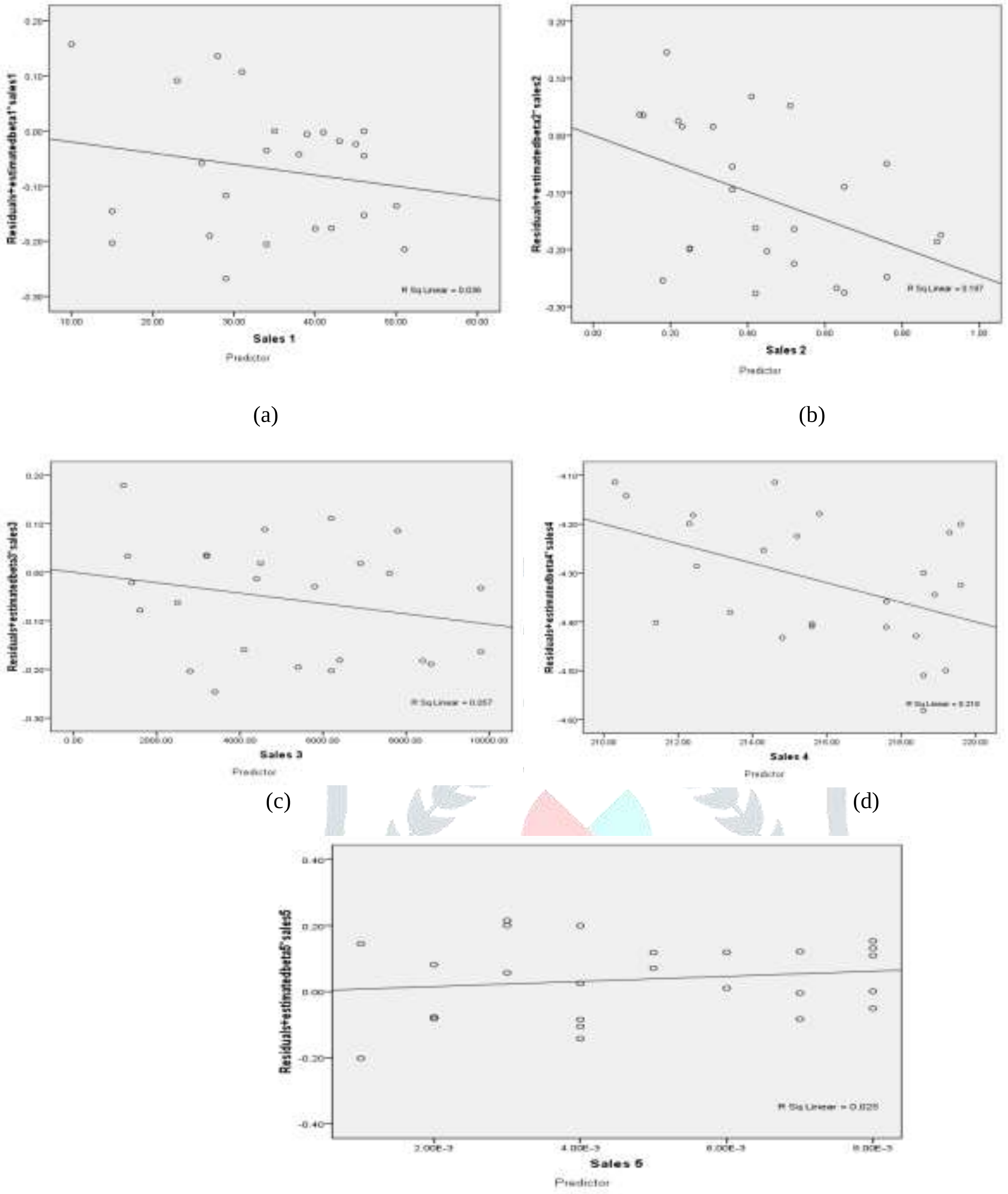
(c)

(d)

(e)

Figure 6: Partial Regression plots

The plots in the Figure 6 (a), (c) and (e) shows approximately the horizontal reference line around zero, which indicates that the coefficient of sale 1, sale 3 and sale 5 are not significantly different from zero. Hence, from Partial Regression plots on Figure 6 and Figure 8 we can see that sale 2 and sale 4 seems to be influential variables on the performance indices of supermarket. Moreover, partial residual plots in Figure 7 show the relationship between a given independent variable sales 1, sales 2, sales 3, sales 4, and sales 5 and the response variable performance indices of supermarket given that other independent variables are also in the model. Hence, from partial residual plots the independent variables sales 2 and sales 4 appeared have additional useful information in predicting performance indices of supermarket (Figure 7 (b) and (d)). However, the independent variables sales 1, sales 3, and sales 5 appeared be have no additional useful information in predicting performance indices of supermarket as points are approximately showed horizontal band Figure 7 (a), (c) and (e)).



(a) (b)

(c) (d)

(e)

Figure 7: Partial Residual plots

Long bar represents predictors that contributes the most new information to the model. Hence, adding predictors sales 2 and 4 in the model contributes highest percentage as indicated by longer bars (Figure 8), indicating two predictors were influential in predicting the performance indices of super market. Furthermore, stepwise regression is used as a modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a non-significant variable is found, it is removed from the model. By specifying two default significance levels: one for adding variables and one for removing variables as 0.15, the predictors sales 2 and sales 4 were selected as influential predictors of performance indices for super market data(results are not shown here).

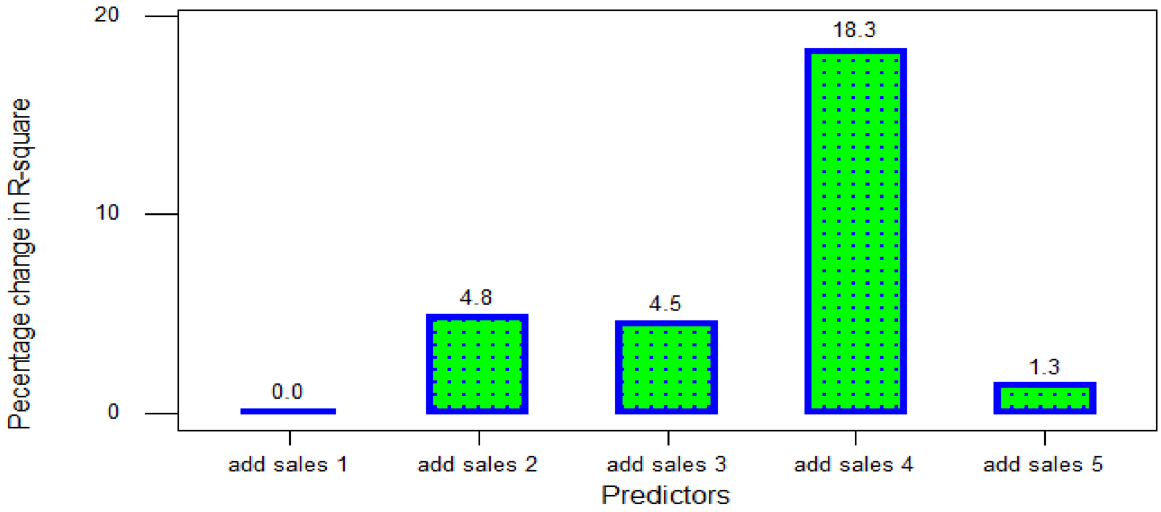


Figure 8: Shows incremental impact of including predictors in the Model using R-square change

Now we will go for clustering approach and by using K-means clustering, we divide the data into 4 clusters and results are further analyzed (profiling) and the simple scatter plots as follows.

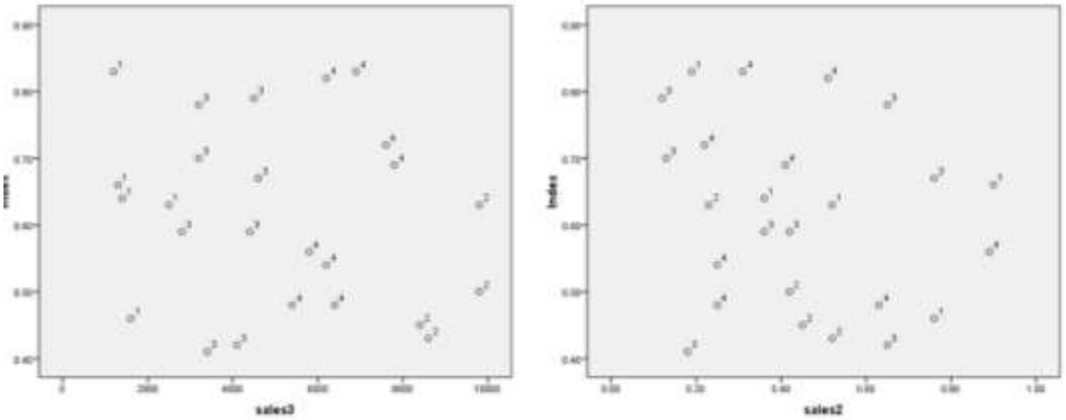


Figure 9:

From the above Figure 9, we can see the sales3 has clearly differentiated the clusters in a better way and hence it is also a vital variable of importance to the index variable.

5. Conclusions

Traditional methods of finding influential variables some time might not give exact results. In that case we can go for cluster analysis and then profiling the results through scatterplots will give more results. In this way we can identify the critical variables from manufacturing or health care domains.

References

1) Amos, C.I., Chen, W.V., Lee, A., Li, W., Kern, M., Lundesten, R., Batliwalla, F., Wener, M., Remmers, E., Kastner, D.A., Criswell, L. A., Seldin, M.F. and Gregersen P.K. (2006). Highdensity SNP analysis of 642 Caucasian families with rheumatoid Arthritis identifies two new linkage regions on 11p 12 and 2q33.Genes Immun.7277-286

2) Bache, I., Nielsen, N. M., Rostgaard, K., Tommerup, N. and Frisch, M. (2007). Autoimmune diseases in a danish cohort of 4,866 carriers of constitutional structural chromosomal rearrangements. Arthitis Rheum. 56 2402-2409.

3) Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate –a practicle and powerful approach to multiple testing .J.Roy. Statist.Soc.Ser.B 57 289-300. MR 1325392

4) Breiman, L. (2001). Random forests. Machine learning 455-32.

5) Dash, M. and Liu, H.(1997). Feature selection for classification. Intelligent Data Analysis**1**131-156.

6) Ding, Y., Cong, L., Ionita-Laza, I., Lo, S.H. and Zheng, T. (2007). Constructing gene Association networks for rheumatoid arthritis using the backward genetype-traitassociation (BGTA) algorithm. *BMC Proceedings* **1**131- 156.

7) Tom Ryan (1997), "*Modern Regression Methods",* John Wiley.

8) Neter, Wasserman, and Kunter (1990), "*Applied Linear Statistical Methods*",3rd ed., Irwin.

9) Draper and Smith (1998), "Applied Regression Analysis", 3rd.ed., John Wiley.

10) Cook and Weisberg (1982), "*Residuals and Influence in Regression*", Chapman and Hall.

11) Belsley, Kuh, and Welsch (1980), "*Regression Diagnostics",* John Wiley.

12) Velleman and Welsch (1981), "*Efficient Computing of Regression Diagnostics'*"", *The American Statistician,* Vol.35, No.4, pp. 234-242.