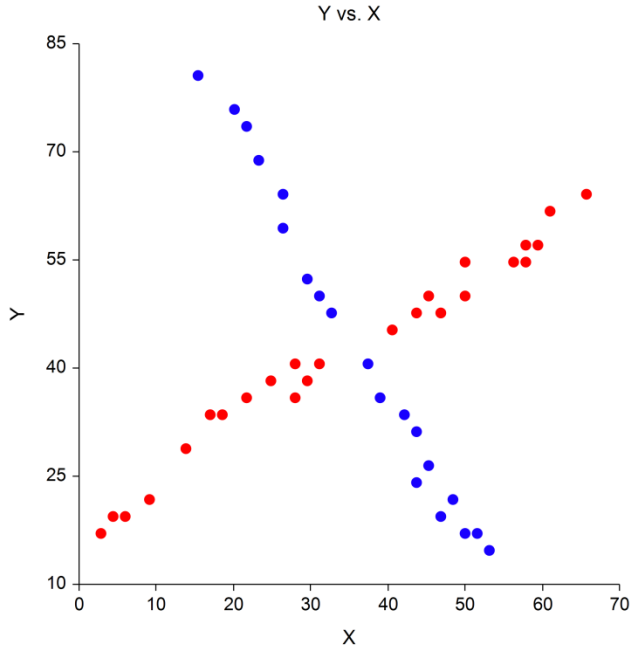
Chapter 449

Regression Clustering

Introduction

This algorithm provides for clustering in the multiple regression setting in which you have a dependent variable *Y* and one or more independent variables, the *X's.* The algorithm partitions the data into two or more clusters and performs an individual multiple regression on the data within each cluster. It is based on an exchange algorithm described in Spath (1985).

The following chart shows data that were clustered using this algorithm. Notice how the two clusters intersect.



Regression Exchange Algorithm

This algorithm is fairly simple to describe. The number of clusters, K, for a given run is fixed. The rows are randomly sorted into the groups to form K initial clusters. An exchange algorithm is applied to this initial configuration which searches for the rows of data that would produce a maximum decrease in a least-squares penalty function (that is, maximizing the increase in R-squared at each step). The algorithm continues until no beneficial exchange of rows can be found.

Our experience with this algorithm indicates that its success depends heavily upon the initial-random configuration. For this reason, we suggest that you try many different configurations. In one test, we found that the optimum resulted from only one in about fifteen starting configurations. Hence, we suggest that you repeat the process twenty-five or thirty times. The program lets you specify the number of repetitions.

Number of Clusters

A report is provided the gives the value of R-squared for each of the values of *K*. Select the value of *K* (number of clusters) that seems to maximize R-squared while minimizing *K*. Also, you should look at the plots of Y versus each X to help in determining the number of clusters. For example, the plot of the data on the previous page would suggest 2, 3, or 4 clusters.

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable. One variable must be a dependent variable that will be regressed on the independent variables.

The data used in our tutorial, a portion of which is given in the following table, were generated with a large X pattern. They are plotted in the scatter plot that was shown above. The data are contained in the RegClus dataset.

RegClus Dataset (Subset)

Y	X
80.58823	15.4088
75.88235	20.12579
73.52941	21.69811
68.82353	23.27044
17.05882	2.830189
19.41177	4.402516
19.41177	5.974843
21.76471	9.119497
.	.
.	.
.	.

Missing Values

Rows with missing values are removed from the analysis.

Example 1 – Regression Clustering

This section presents an example of how to run a cluster analysis of the data found in the RegClus dataset. This is a bivariate set of data generated to exhibit a large X pattern.

Setup

To run this example, complete the following steps:

1 Open the RegClus example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **RegClus** and click **OK**.

2 Specify the Regression Clustering procedure options

- Find and open the **Regression Clustering** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Y Dependent Variable....................................**Y**
X's Independent Variables..............................**X**
Number of Random Starts..............................**50**
Random Seed..**55499** (for reproducibility)
Maximum Clusters..**4**
Reported Clusters..**2**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Iteration Detail

Iteration Detail

		R-Squared	
Number of Clusters	**Replication**	**Value**	**Bar**
2	1	0.997218	|||||||||||||||||||||||||||||||||||||||||||||||||||
2	2	0.961698	||||||||||||||||||||||||||||||||||||||||||||||||||
2	3	0.961698	||||||||||||||||||||||||||||||||||||||||||||||||||
2	4	0.997218	|||||||||||||||||||||||||||||||||||||||||||||||||||
2	5	0.961698	||||||||||||||||||||||||||||||||||||||||||||||||||
.	.	.	.
.	.	.	.
.	.	.	.

3	1	0.998364	||||||||||||||||||||||||||||||||||||||||||||
3	2	0.998332	||||||||||||||||||||||||||||||||||||||||||||
3	3	0.997952	||||||||||||||||||||||||||||||||||||||||||||
3	4	0.997899	||||||||||||||||||||||||||||||||||||||||||||
3	5	0.997952	||||||||||||||||||||||||||||||||||||||||||||
.	.	.	.
.	.	.	.
.	.	.	.
4	1	0.999507	||||||||||||||||||||||||||||||||||||||||||||
4	2	0.998169	||||||||||||||||||||||||||||||||||||||||||||
4	3	0.999489	||||||||||||||||||||||||||||||||||||||||||||
4	4	0.998388	||||||||||||||||||||||||||||||||||||||||||||
4	5	0.999457	||||||||||||||||||||||||||||||||||||||||||||
.	.	.	.
.	.	.	.
.	.	.	.

This report displays the progress of the program through the various replications.

Number of Clusters

This column displays the number of clusters for the configurations presented on this row.

Replication

This column displays a sequence number for this replication.

R-Squared: Value

This is the R-Squared that would result from fitting a separate regression of Y on X within each cluster. As this value approaches one, the fit of the regression is better.

R-Squared: Bar

This is a bar chart of the R-Squared Value. This helps you visually determine the optimum value for the number of clusters.

Iteration Summary

Iteration Summary

Number of Clusters	Maximum R-Squared Value	Maximum R-Squared Bar
2	0.997218	||||||||||||||||||||||||||||||||||||||||||||
3	0.998393	||||||||||||||||||||||||||||||||||||||||||||
4	0.999570	||||||||||||||||||||||||||||||||||||||||||||

This report is the identical to the Iteration Detail report, except that only the row with the maximum value of R-Squared is displayed for each number of clusters. This report should help you determine the number of clusters by finding the first value of *K* where there is a large jump in the R-Squared value.

In this example, there is no jump. The value of *K* selected would be two.

Regression Coefficients

Regression Coefficients (2 Clusters)

	Cluster	
Variable	**1**	**2**
Intercept	18.26343	110.6421
X	0.6797758	-1.866411

This report displays the coefficients of each regression equation for each cluster. For example, since we selected two clusters, there are two regression equations. These are

$$Y = 18.26343 + 0.6797758X$$

and

$$Y = 110.6421 - 1.866411X$$

Cluster Detail

Cluster Detail (2 Clusters)

Row	Cluster	Y
1	2	80.58823
2	2	75.88235
3	2	73.52941
4	2	68.82353
5	1	17.05882
6	1	19.41177
7	1	19.41177
8	1	21.76471
9	1	28.82353
10	1	33.52941
.	.	.
.	.	.
.	.	.

This report displays the cluster to which each row is assigned. The value of the dependent variable is also displayed to help you quickly identify a particular row. The cluster ID number may be stored directly on the database for further analysis and plotting.

Scatter Plot using Cluster Numbers

Once the cluster numbers are stored, you may use them as a grouping variable in the Scatter Plot procedure. This will provide a plot such as this:

Scatter Plot