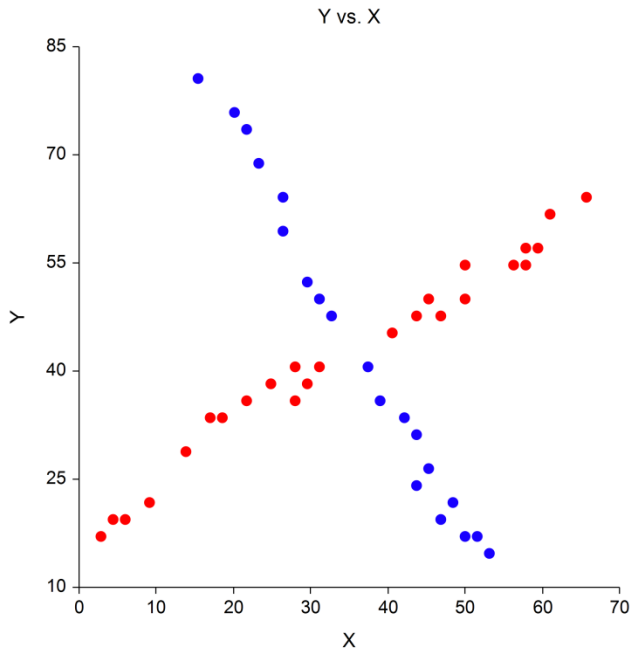
Chapter 449

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 actually 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.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable

Specify a single, dependent variable. Remember that the dependent variable is predicted by the independent variables.

Independent Variables

X's: Independent Variables

Specify one or more independent variables. These are used to predict the dependent variable.

Include Intercept

Specifies whether you want to include the Y-intercept term in the regression model. Under most circumstances, you would.

Clustering Options

Number of Random Starts

This option specifies the number of different random configurations that should be run for each value of *K*. We suggest that about twenty-five repetitions be run since each initial configuration is completely random and the algorithm often converges to a non-optimal local optimum.

Because of execution time, you might want to set this value to three or four until you have found an appropriate value for *K* and then reset this value to twenty-five for a second run.

Maximum Iterations

This option sets the number of internal iterations before the algorithm is aborted. It is possible for a set of data to put the algorithm into an infinite loop. This option prevents this.

Minimum Rows Per Cluster

The third box lets you specify the minimum number of rows per cluster. Remember that in regression analysis, each cluster must contain at least one more row than there are independent variables.

Zero Exponent

This is the exponent of the value used as zero by the regression algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

This box supplies the negative exponent. A value of 5 represents 1E-5 which is 0.00001.

Clustering Options – Number of Clusters

Minimum Clusters

The minimum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters. The actual number of clusters used is set above by the Number Clusters option.

Maximum Clusters

The maximum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters. The actual number of clusters used is set above by the Number Clusters option.

Reported Clusters

The is the number of clusters to be reported on. Although the program can find results for a range of cluster sizes, this option sets the size that is used in the detail and data storage sections.

Format Options

Label Variable

This is an optional variable containing identification for each row. These labels are used to enhance the interpretability of the reports.

Reports Tab

The following options control the formatting of the reports.

Select Reports

Iteration Detail Report - Cluster Report

Specify whether to display the indicated reports.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Storage Tab

These options let you specify where to store the cluster number of each row on the current database.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the value of *K* specified by the Reported Clusters option.

Warning: Any data already in this variable are replaced by the cluster number. Be careful not to specify variables that contain important data.

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.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Regression Clustering window.

1 Open the RegClus dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **RegClus.NCSS**.
- Click **Open**.

2 Open the Regression Clustering window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Regression Clustering** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Regression Clustering window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**. “Y” will appear in the Interval Variables box.
- Double-click in the **X’s: Independent Variables** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**. “X” will appear in the X’s: Independent Variables box.
- Enter **5** for **Number of Random Starts**.
- Enter **4** for **Maximum Clusters**.
- Enter **2** for **Reported Clusters**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Iteration Detail Section (each run may differ slightly)

Iteration Detail Section

<table>
<tr><th>Number of Clusters</th><th>Replication Number</th><th>R-Squared Value</th><th>R-Squared Bar</th></tr>
<tr><td>2</td><td>1</td><td>0.997218</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>2</td><td>2</td><td>0.997218</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>2</td><td>3</td><td>0.997218</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>2</td><td>4</td><td>0.997218</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>2</td><td>5</td><td>0.961698</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>1</td><td>0.998170</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>3</td><td>2</td><td>0.998170</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>3</td><td>3</td><td>0.997952</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>3</td><td>4</td><td>0.997952</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>3</td><td>5</td><td>0.998343</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>4</td><td>1</td><td>0.999457</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>4</td><td>2</td><td>0.999015</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>4</td><td>3</td><td>0.999489</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>4</td><td>4</td><td>0.998505</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
<tr><td>4</td><td>5</td><td>0.998975</td><td>|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII</td></tr>
</table>

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 Number

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 Section

Iteration Summary Section

Number of Clusters	R-Squared Value	R-Squared Bar
2	0.997218	||||||||||||||||||||||||||||||||||||||||
3	0.998343	||||||||||||||||||||||||||||||||||||||||
4	0.999489	||||||||||||||||||||||||||||||||||||||||

This section is the identical to the Iteration Detail Section 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 Coefficient Section

Regression Coefficient Section

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

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

$$Y = 110.6421 - (1.866411)\,X$$

and

$$Y = 18.26343 + (0.6797758)\,X$$

Cluster Section

Cluster Section

Row	Cluster Number	Y
1	1	80.58823
2	1	75.88235
3	1	73.52941
4	1	68.82353
5	2	17.05882
6	2	19.41177
7	2	19.41177
8	2	21.76471
.	.	.
.	.	.
.	.	.

This report displays a report of which cluster each row is assigned to. The value of the dependent variable is also displayed to help you quickly identify a particular row. The cluster 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 program. This will provide a plot such as this: