Two Step Cluster Analysis

Brawijaya Professional Statistical Analysis
BPSA MALANG
Jl. Kertoasri 66 Malang
(0341) 580342
081 753 3962

TwoStep Cluster Analysis

The TwoStep Cluster Analysis procedure is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- The ability to create clusters based on both categorical and continuous variables.
- Automatic selection of the number of clusters.
- The ability to analyze large data files efficiently.

Clustering Principles

In order to handle categorical and continuous variables, the TwoStep Cluster Analysis procedure uses a likelihood distance measure which assumes that variables in the cluster model are independent. Further, each continuous variable is assumed to have a normal (Gaussian) distribution and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but you should try to be aware of how well these assumptions are met.

The two steps of the TwoStep Cluster Analysis procedure's algorithm can be summarized as follows:

Step 1. The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree in a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A node that contains multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the data file.

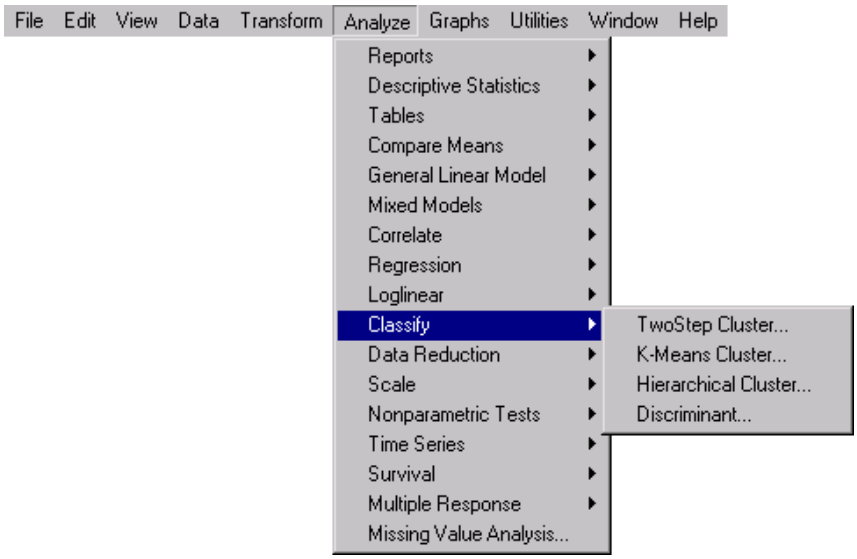
Step 2. The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

Using TwoStep Cluster Analysis to Classify Motor Vehicles

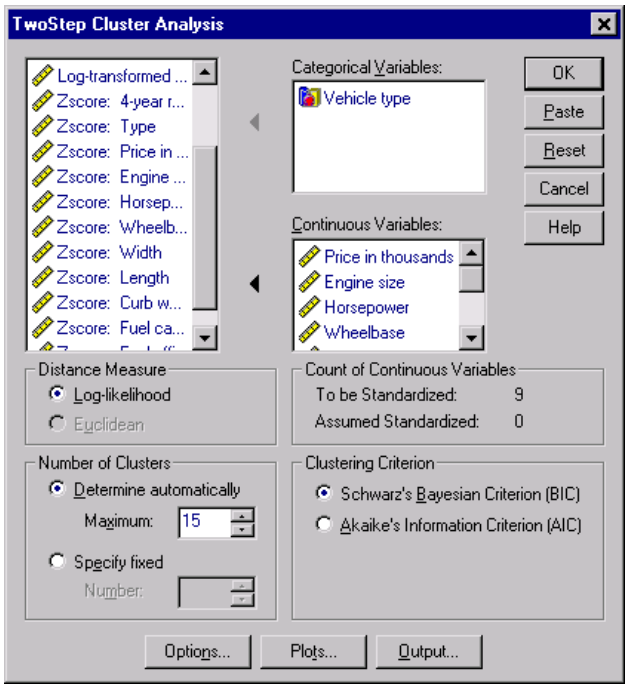
Car manufacturers need to be able to appraise the current market to determine the likely competition for their vehicles. If cars can be grouped according to available data, this task can be largely automatic using cluster analysis.

Information for various makes and models of motor vehicles is contained in car_sales.sav .Use the TwoStep Cluster Analysis procedure to group automobiles according to their prices and physical properties.

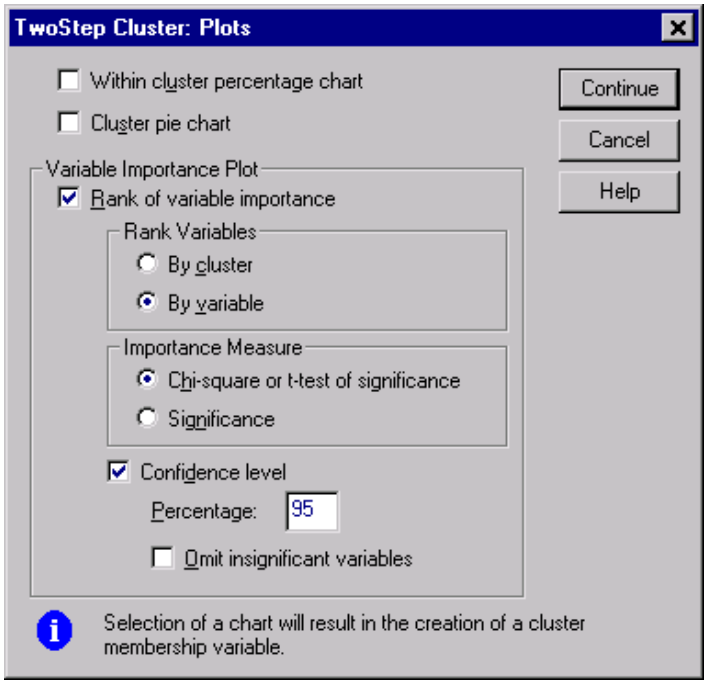
Running the Analysis



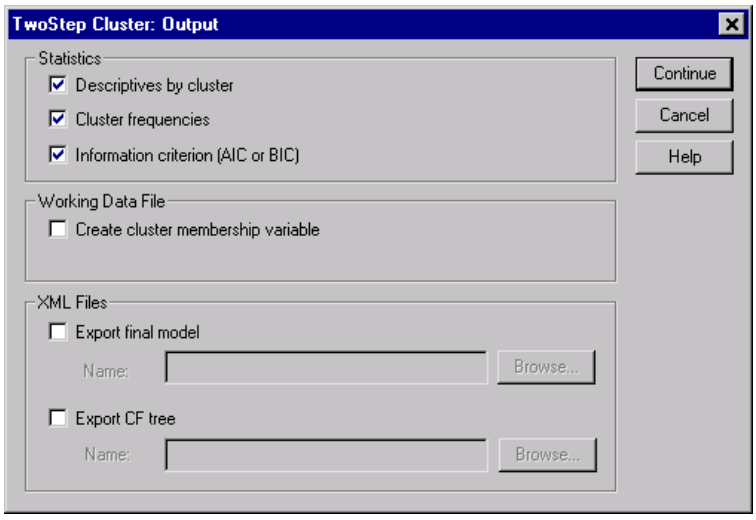
Analyze
Classify
TwoStep Cluster...



- To run a TwoStep Cluster Analysis analysis, from the menus choose:
- Select *Vehicle type* as a categorical variable.
- Select *Price in thousands* through *Fuel efficiency* as continuous variables.
- Click **Plots**.



- Select **Rank of variable importance**.
- Select **By variable** in the Rank Variables group.
- Select **Confidence level**.
- Click **Continue**, then click **Output** in the TwoStep Cluster Analysis dialog box.



- Select **Information criterion (AIC or BIC)** in the Statistics group.
- Click **Continue**.
- Click **OK** in the TwoStep Cluster Analysis dialog box.

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are with respect to the change at the two clusters.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

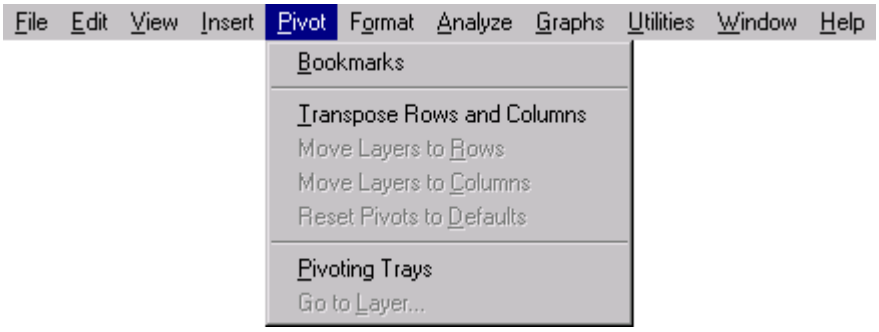
- The Auto-clustering table summarizes the process by which the number of clusters is chosen.
- The clustering criterion (in this case the BIC) is computed for each potential number of clusters. Smaller values of the BIC indicate better models, and in this situation, the "best" cluster solution has the smallest BIC.
- However, there are clustering problems in which the BIC will continue to decrease as the number of clusters increases, but the improvement in the cluster solution, as measured by the BIC Change, is not worth the increased complexity of the cluster model, as measured by the number of clusters. In such situations, the changes in BIC and changes in the distance measure are evaluated to determine the "best" cluster solution.
- A good solution will have a reasonably large Ratio of BIC Changes and a large Ratio of Distance Measures.

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	62	40.8%	39.5%
	2	39	25.7%	24.8%
	3	51	33.6%	32.5%
	Combined	152	100.0%	96.8%
Excluded Cases		5		3.2%
Total		157		100.0%

The cluster distribution table shows the frequency of each cluster. Of the 157 total cases, 5 were excluded from the analysis due to missing values on one or more of the variables. Of the 152 cases assigned to clusters, 62 were assigned to the first cluster, 39 to the second, and 51 to the third.

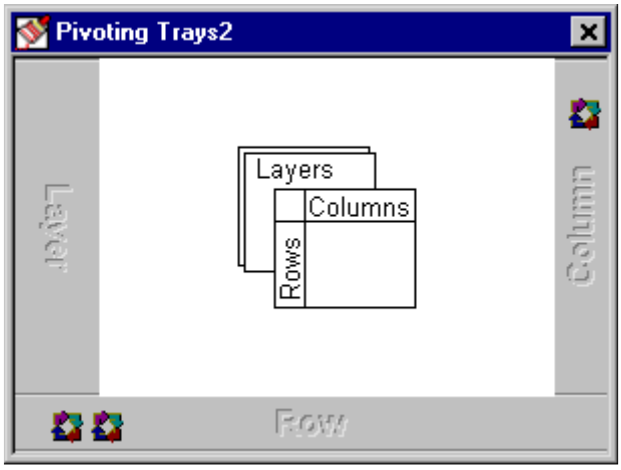
Cluster Profiles



The Centroids table displays the mean and standard deviation for the cases in each cluster. It is currently too wide for easy viewing, so you are going to pivot it.

- In the output window, double-click the Centroids table to activate it.
- From the Viewer menus choose:

Pivot
Pivoting Trays...



- Drag the Cluster ID icon to the Columns tray.
- Drag the Continuous Variables and Statistics icons (in that order, left to right) to the Rows tray.
- Close the Pivoting Trays window.
- Deactivate the table by clicking outside of its boundaries.

		Cluster			
		1	2	3	Combined
Price in thousands	Mean	19.61671	26.56182	37.29980	27.33182
	Std. Deviation	7.644070	10.185175	17.381187	14.418669
Engine size	Mean	2.194	3.559	3.700	3.049
	Std. Deviation	.4238	.9358	.9493	1.0498
Horsepower	Mean	143.24	187.92	232.96	184.81
	Std. Deviation	30.259	39.049	54.408	56.823
Wheelbase	Mean	102.595	112.972	109.022	107.414
	Std. Deviation	4.0799	9.6537	5.7644	7.7178
Width	Mean	68.539	72.744	72.924	71.089
	Std. Deviation	1.9366	4.1781	2.1855	3.4647
Length	Mean	178.235	191.110	194.688	187.059
	Std. Deviation	9.6534	14.4415	10.3512	13.4712
Curb weight	Mean	2.83742	3.96759	3.57890	3.37618
	Std. Deviation	.310867	.671766	.297204	.636593
Fuel capacity	Mean	14.979	22.064	18.443	17.959
	Std. Deviation	1.8699	4.2894	2.0445	3.9376
Fuel efficiency	Mean	27.24	19.51	23.02	23.84
	Std. Deviation	3.578	2.910	2.060	4.305

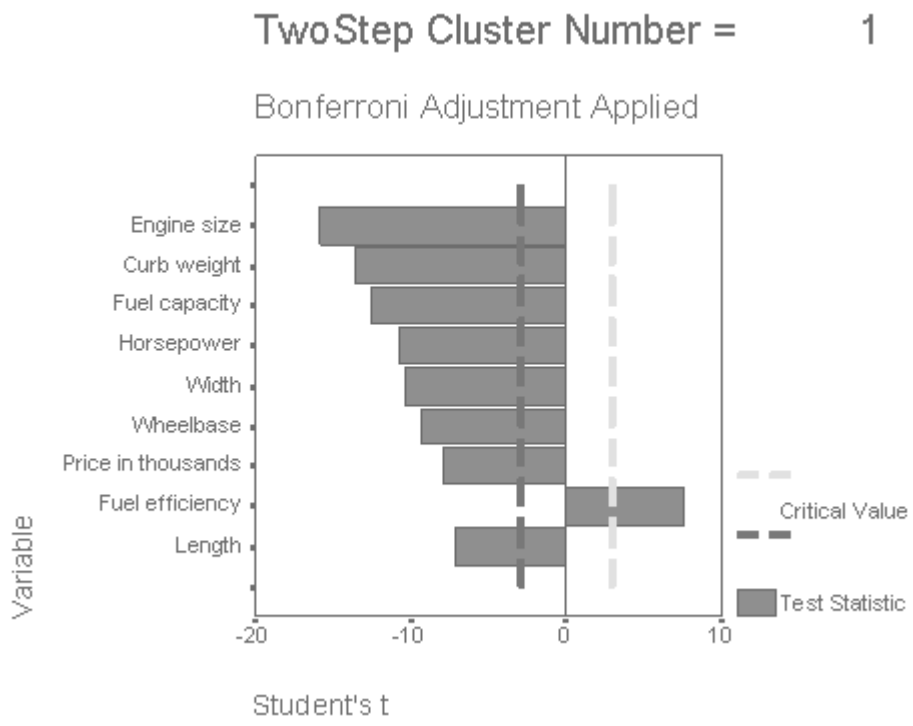
The centroids show that the clusters are well separated by the continuous variables.

- Motor vehicles in cluster 1 are cheap, small, and fuel efficient.
- Motor vehicles in cluster 2 are moderately priced, heavy, and have a large gas tank, presumably to compensate for their poor fuel efficiency.
- Motor vehicles in cluster 3 are expensive, large, and are moderately fuel efficient.

		Automobile		Truck	
		Frequency	Percent	Frequency	Percent
Cluster	1	61	54.5%	1	2.5%
	2	0	.0%	39	97.5%
	3	51	45.5%	0	.0%
	Combined	112	100.0%	40	100.0%

- The cluster frequency table by *Vehicle type* further clarifies the properties of the clusters.
- Cluster 2 is comprised entirely of trucks. Clusters 1 and 3 contain automobiles, save for a single truck in Cluster 1.
- Examination of the data file reveals this to be the Toyota RAV4.

Attribute Importance

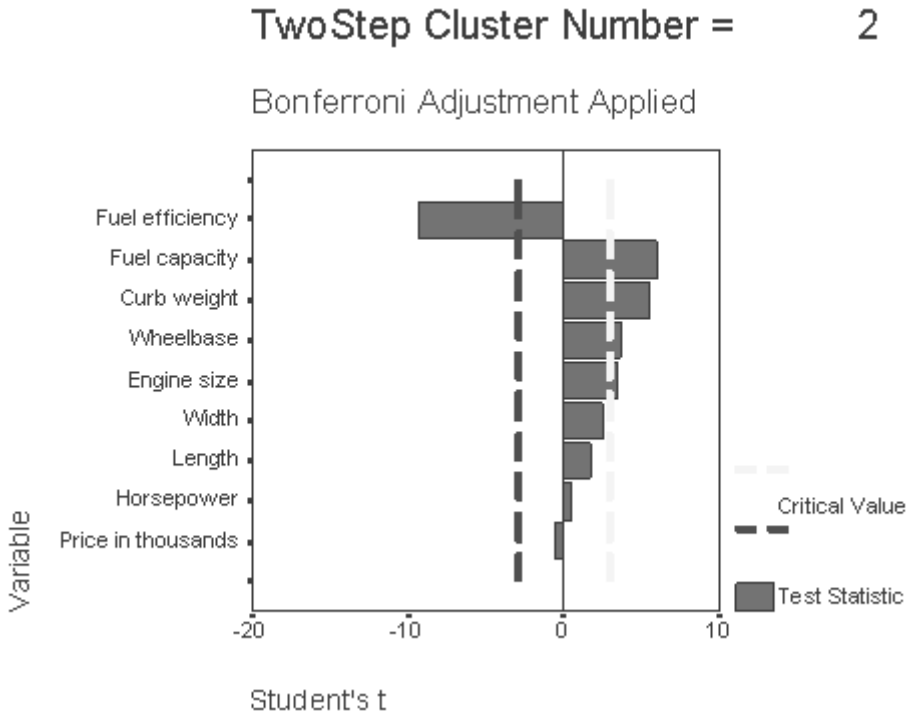


The "by variable" importance charts are produced with a separate chart for each cluster. The variables are lined up on the Y axis, in descending order of importance.

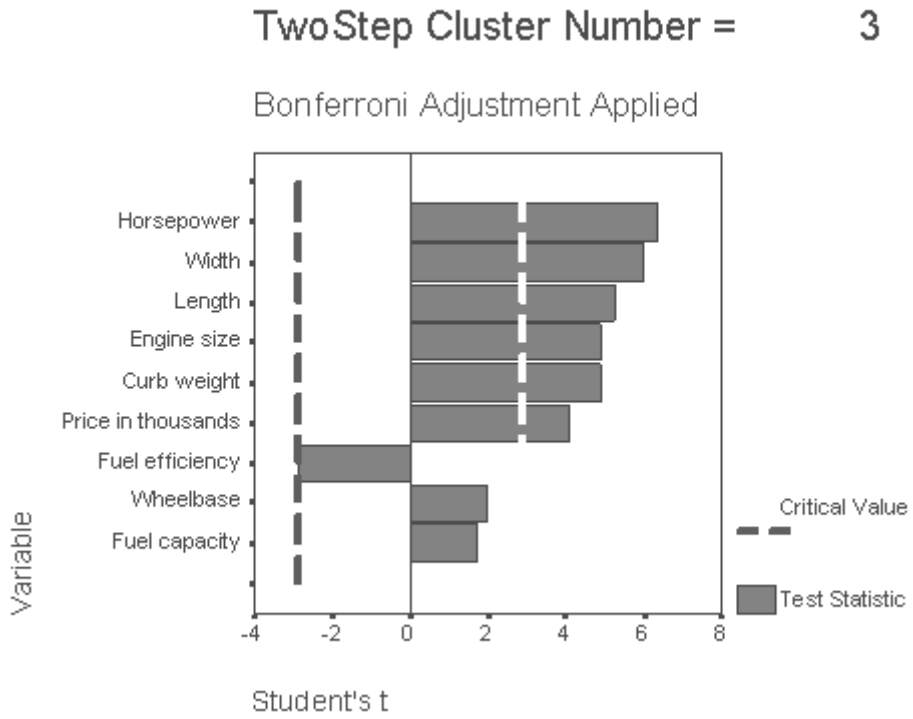
The dashed vertical lines mark the critical values for determining the significance of each variable. For a variable to be considered significant, its *t* statistic must exceed the dashed line in either a positive or negative direction. Since the importance measures for all of the variables exceed the critical value in this chart, you can conclude that all of the continuous variables contribute to the formation of the first cluster.

A negative *t* statistic indicates that the variable generally takes smaller than average values within this cluster, while a positive *t* statistic indicates the variable takes larger than

average values. Thus, for Cluster 1, *Fuel efficiency* takes larger than average values while all of the other variables take smaller than average values. These results confirm the trends observed in the Centroids table.



The chart for Cluster 2 shows that *Width*, *Length*, *Horsepower*, and *Price in thousands* are not important to the formation of this cluster



The chart for Cluster 3 shows that *Wheelbase* and *Fuel capacity* are not important to the formation of this cluster, while *Fuel efficiency* is just barely significant.

Summary

Using the TwoStep Cluster Analysis procedure, you have separated the motor vehicles into three fairly broad categories. In order to obtain finer separations within these groups, you should collect information on other attributes of the vehicles. For example, you could note the crash test performance or the options available.

Related Procedures

The TwoStep Cluster Analysis procedure is useful for finding natural groupings of cases or variables. It works well with categorical and continuous variables, and can analyze very large data files.

- If you have a small number of cases, and want to choose between several methods for cluster formation, variable transformation, and measuring the dissimilarity between clusters, try the Hierarchical Cluster Analysis procedure. The Hierarchical Cluster Analysis procedure also allows you to cluster variables instead of cases.
- The K-Means Cluster Analysis procedure is limited to scale variables, but can be used to analyze large data and allows you to save the distances from cluster centers for each object.