Tutorial overview

In this tutorial, we show how to perform a regression analysis.

Our dataset consists in engine cars description. We want to predict “mpg” consumption from cars characteristics such as weight, horsepower, …

In this tutorial, we use the following components:

<table>
<thead>
<tr>
<th>TAB</th>
<th>Operator (Component)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data visualization</td>
<td>View dataset</td>
<td>View the dataset in a grid</td>
</tr>
<tr>
<td>Feature selection</td>
<td>Define status</td>
<td>Define attributes status</td>
</tr>
<tr>
<td>Regression</td>
<td>Multiple linear regression</td>
<td>Perform the regression analysis</td>
</tr>
</tbody>
</table>

Loading the dataset

- Open the existing stream diagram

  1 – Choose File/Open… in the TANAGRA main menu.

  2 – Get « autompg.bdm » which is in the « Dataset » sub-directory.

- Add a component for data visualization

  1 – Add View dataset to the diagram. Click on DATA VISUALIZATION components tab, select **View dataset** component. Drag this component over the diagram and drop on the “Dataset” node.

  2 – Select the “View dataset” component in the diagram (if it is not already selected) and right click to show the popup menu. Choose the View command, dataset will be visible in the right frame.
There are 392 cars in the dataset, described with 6 continuous attributes: number of cylinder, displacement, horsepower, weight, acceleration and consumption (miles per gallon).
Define status of attributes for regression

1 – Click on the “Dataset” component in the diagram and add a Define Status component with the shortcut in the tool bar. A dialog box appears automatically, you may define the status of each attribute.

2 – Ensure that “Input” tab is selected, choose the five first attributes of the list, and click on the arrow button to add them to the “Input” list. These attributes are “exogenous” attributes of the analysis.
3 – Activate the “Target” tab. Select “mpg” attribute and add this one to the target list. This attribute is the “endogenous” attribute; we want to predict its values from exogenous ones.

4 – Click on the OK button to validate this selection and close this dialog box.
5 – Double-click this **Define Status** component, you can verify on the right side of the window your attribute selection.

![Image of Define Status component](image.png)

**Perform the regression analysis**

1 – Add **Multiple linear regression component** (REGRESSION TAB) to the stream diagram, after the node «Define status 1».

2 – There is no parameter to specify in this analysis, select **View** in the popup menu of “Multiple linear regression 1” to run the analysis. Results appear in the right frame.
Coefficient of determination is rather good (0.70), the most significant attributes seem to be “horsepower” and “weight”. I think nevertheless that there is a very strong colinearity between the exogenous attributes, it would be necessary to study more in detail this regression.