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:

TAB	**Operator (Component)**	**Comment**
Data visualization	View dataset	View the dataset in a grid
Feature selection	Define status	Define attributes status
Regression	Multiple linear regression	Perform the regression analysis

Loading the dataset

➢ Open the existing stream diagram

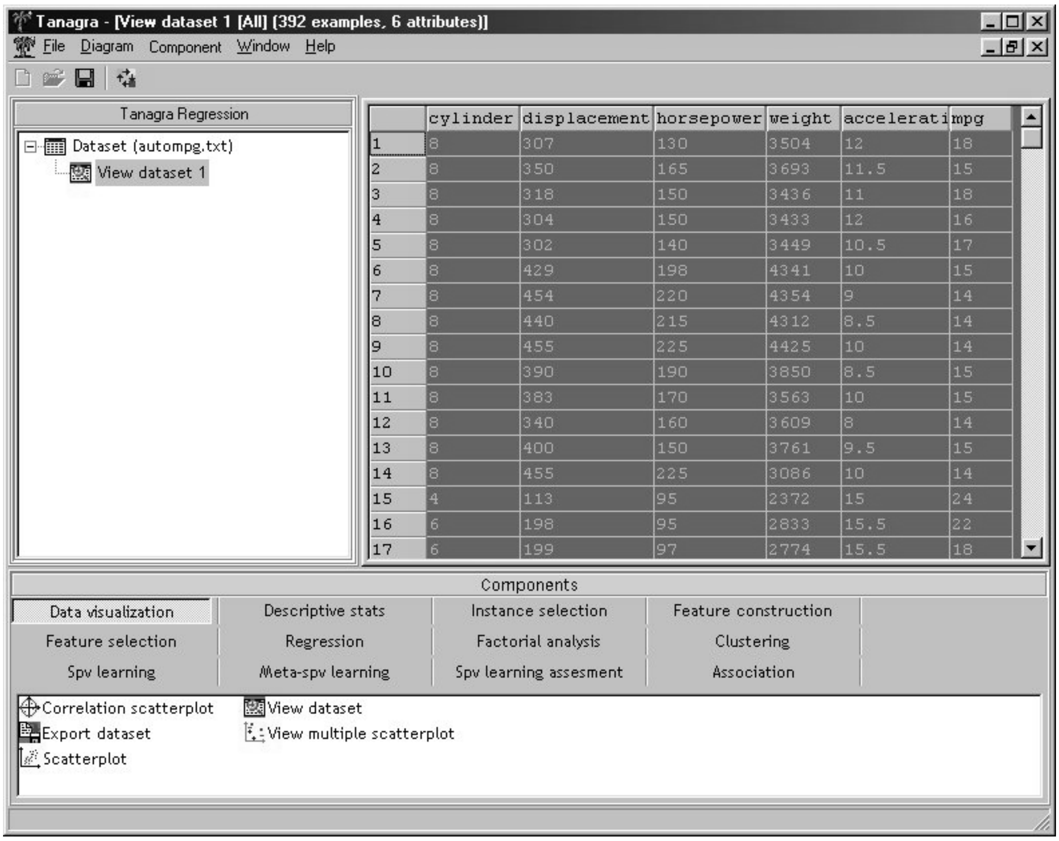
1 – Choose *File/Open…* in the TANAGRA main menu.

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.

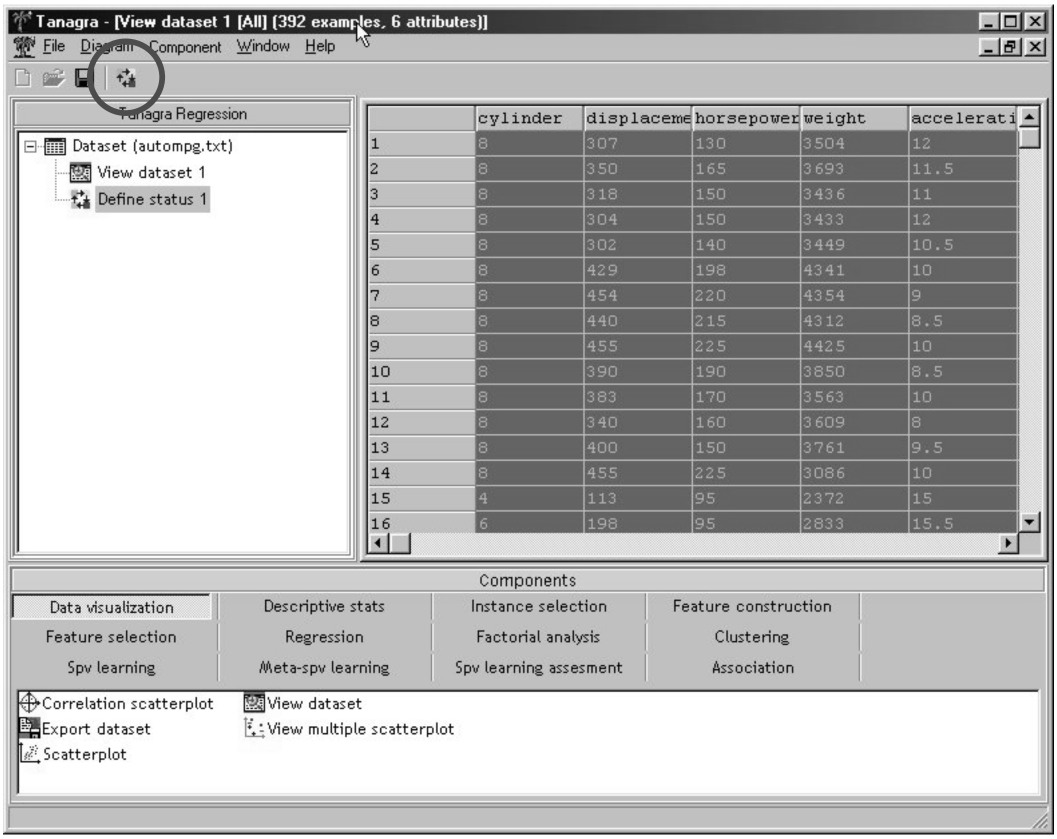


	cylinder	displacement	horsepower	weight	accelerati	mpg
1	8	307	130	3504	12	18
2	8	350	165	3693	11.5	15
3	8	318	150	3436	11	18
4	8	304	150	3433	12	16
5	8	302	140	3449	10.5	17
6	8	429	198	4341	10	15
7	8	454	220	4354	9	14
8	8	440	215	4312	8.5	14
9	8	455	225	4425	10	14
10	8	390	190	3850	8.5	15
11	8	383	170	3563	10	15
12	8	340	160	3609	8	14
13	8	400	150	3761	9.5	15
14	8	455	225	3086	10	14
15	4	113	95	2372	15	24
16	6	198	95	2833	15.5	22
17	6	199	97	2774	15.5	18

There are 392 cars in the dataset, described with 6 continuous attributes: number of cylinder, displacement, horsepower, weight, acceleration and consumption (miles per gallon).

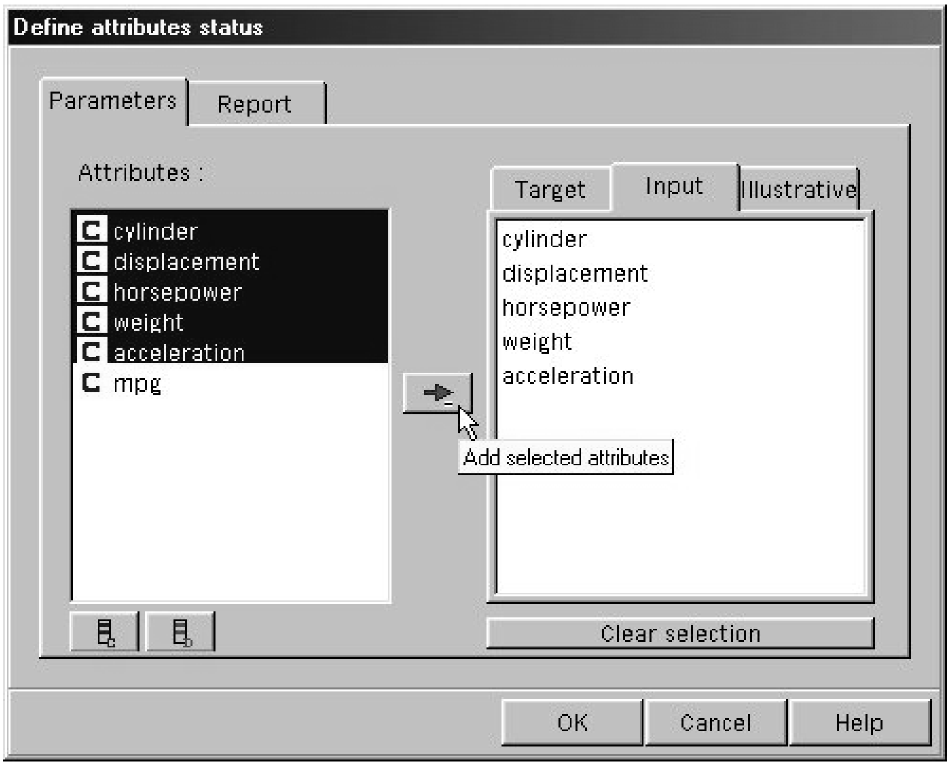
Define status of attributes for regression

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.



	cylinder	displaceme	horsepower	weight	accelerati
1	8	307	130	3504	12
2	8	350	165	3693	11.5
3	8	318	150	3436	11
4	8	304	150	3433	12
5	8	302	140	3449	10.5
6	8	429	198	4341	10
7	8	454	220	4354	9
8	8	440	215	4312	8.5
9	8	455	225	4425	10
10	8	390	190	3850	8.5
11	8	383	170	3563	10
12	8	340	160	3609	8
13	8	400	150	3761	9.5
14	8	455	225	3086	10
15	4	113	95	2372	15
16	6	198	95	2833	15.5

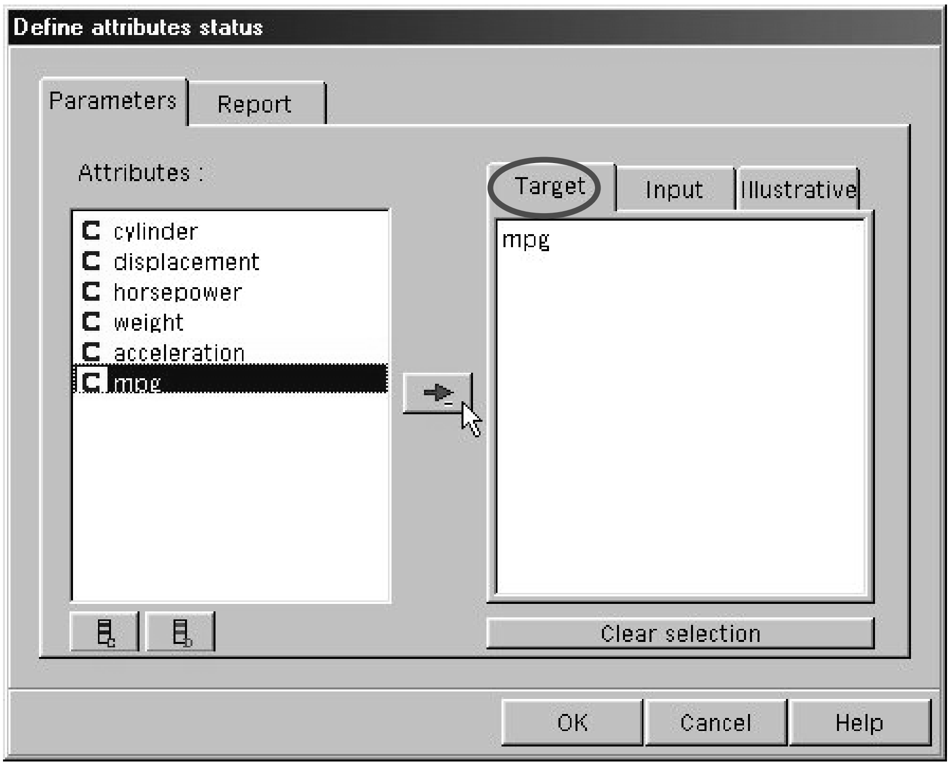
2 – Ensure that “Input” tab is selected, choose the five first attributes of the list, and click on the arrow button to add them to the “Input” list. These attributes are “exogenous” attributes of the analysis.



Attributes: cylinder, displacement, horsepower, weight, acceleration, mpg

Input: cylinder, displacement, horsepower, weight, acceleration

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.

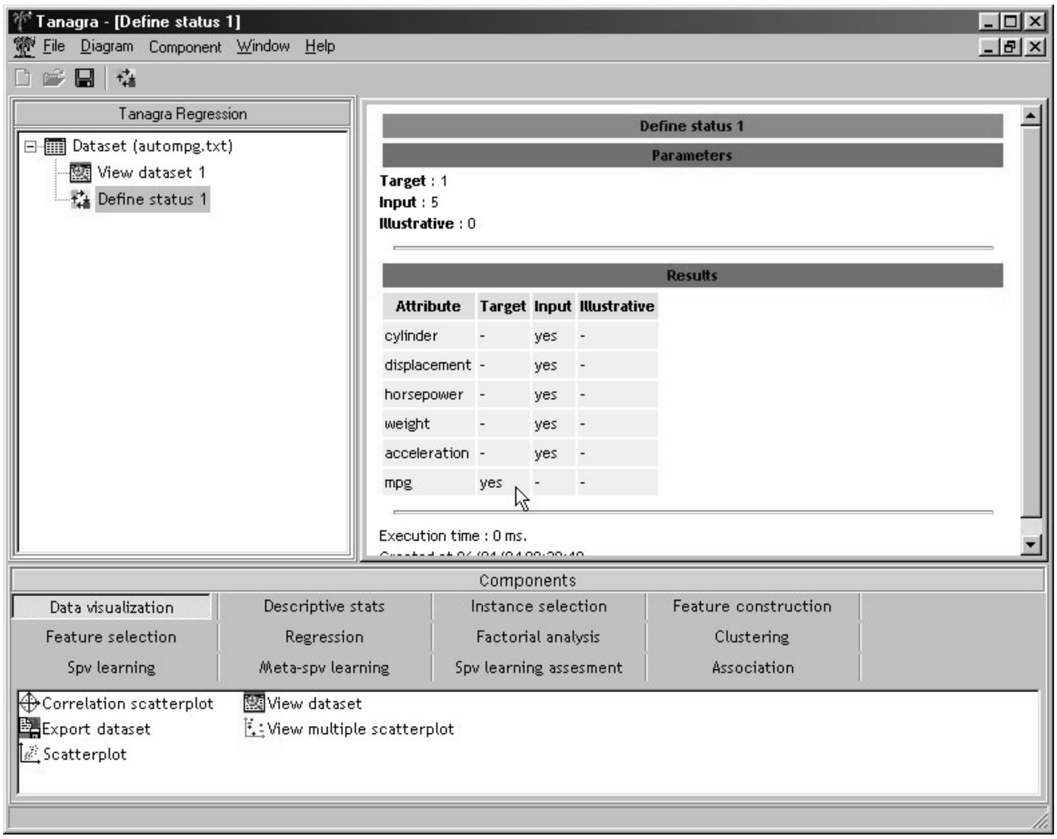


Attributes: cylinder, displacement, horsepower, weight, acceleration, mpg

Target: mpg

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.



Define status 1

Parameters

Target : 1
Input : 5
Illustrative : 0

Results

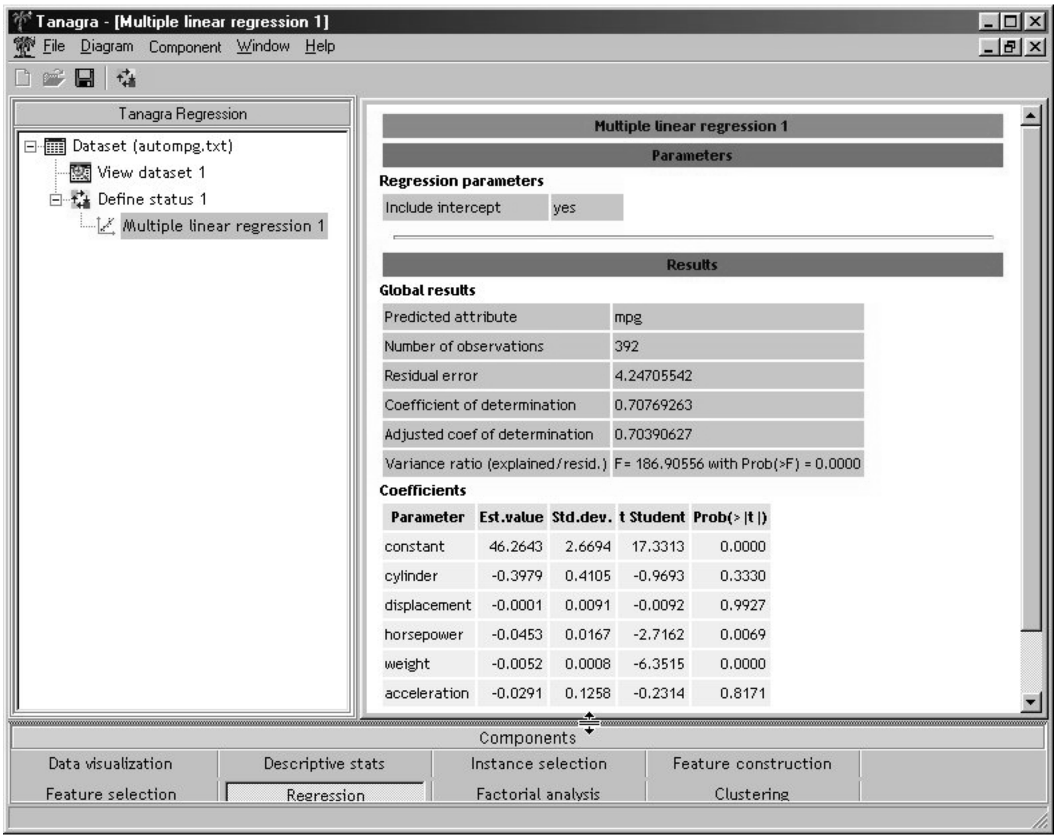
Attribute	Target	Input	Illustrative
cylinder	-	yes	-
displacement	-	yes	-
horsepower	-	yes	-
weight	-	yes	-
acceleration	-	yes	-
mpg	yes	-	-

Execution time : 0 ms.

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.



Multiple linear regression 1

Parameters

Regression parameters

Include intercept	yes

Results

Global results

Predicted attribute	mpg
Number of observations	392
Residual error	4.24705542
Coefficient of determination	0.70769263
Adjusted coef of determination	0.70390627
Variance ratio (explained/resid.)	F= 186.90556 with Prob(>F) = 0.0000

Coefficients

Parameter	Est.value	Std.dev.	t Student	Prob(>|t|)
constant	46.2643	2.6694	17.3313	0.0000
cylinder	-0.3979	0.4105	-0.9693	0.3330
displacement	-0.0001	0.0091	-0.0092	0.9927
horsepower	-0.0453	0.0167	-2.7162	0.0069
weight	-0.0052	0.0008	-6.3515	0.0000
acceleration	-0.0291	0.1258	-0.2314	0.8171

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.