Subject

Computing the correlation coefficient between two or more variables.

Computing a statistical indicator and sorting the results according to this indicator is a recurring task of the data miner. In this tutorial, we show how to quickly set up the calculation of the linear correlation (1) of an endogenous variable with exogenous variables in order to detect relevant attributes; (2) between exogenous variables in order to detect collinearities.

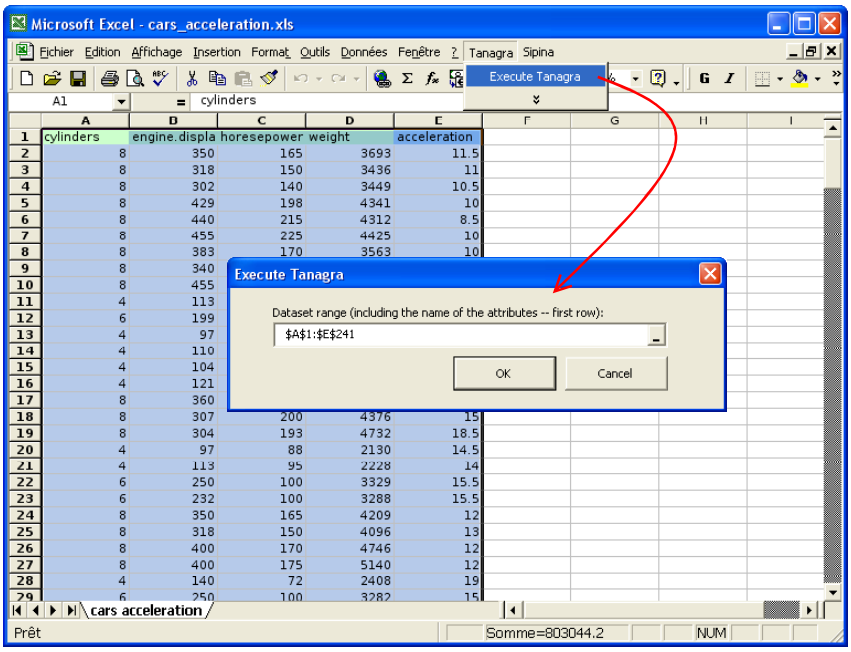
Dataset

We use the CARS_ACCELERATION.XLS dataset: ACCELERATION is the endogenous attribute (acceleration time from 0 to x mph).

Correlation coefficient

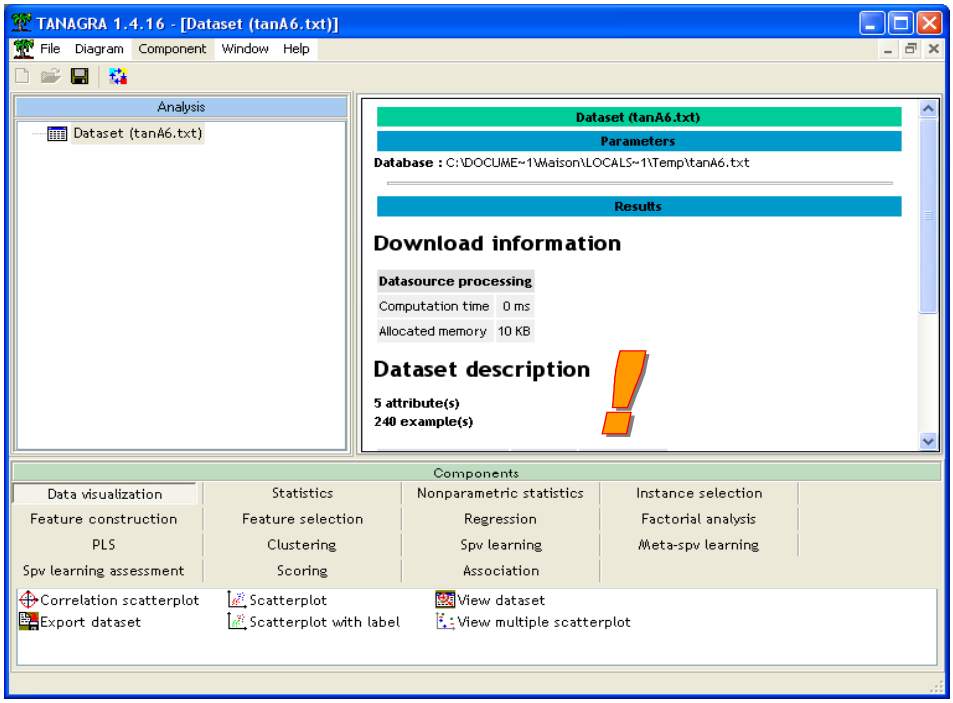
Creating a new diagram

The simplest way in order to create a diagram is to load the dataset in the EXCEL spreadsheet. Then, we select the data range and we click on the menu TANAGRA/EXECUTE TANAGRA[1].



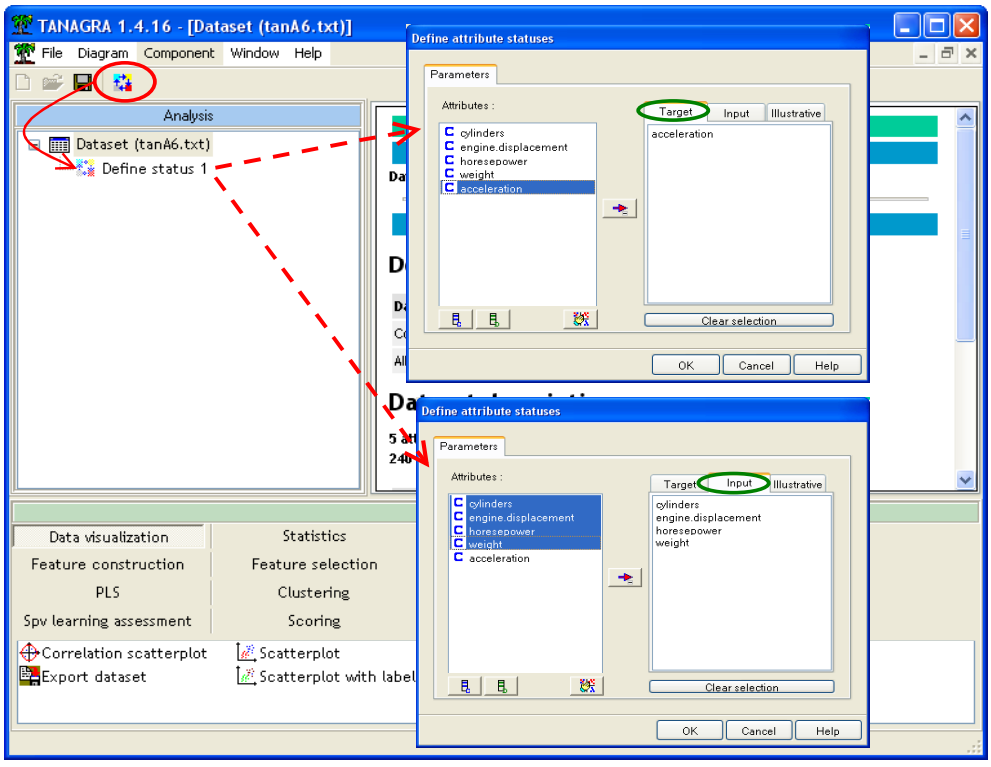
We check the range selection. If it is right, we click on the OK button. TANAGRA is automatically started and a new stream diagram is ready.

[1] The EXCEL add-in TANAGRA.XLA is available since the version 1.4.11. See the tutorial on the web site for the installation of this add-in in your spreadsheet.



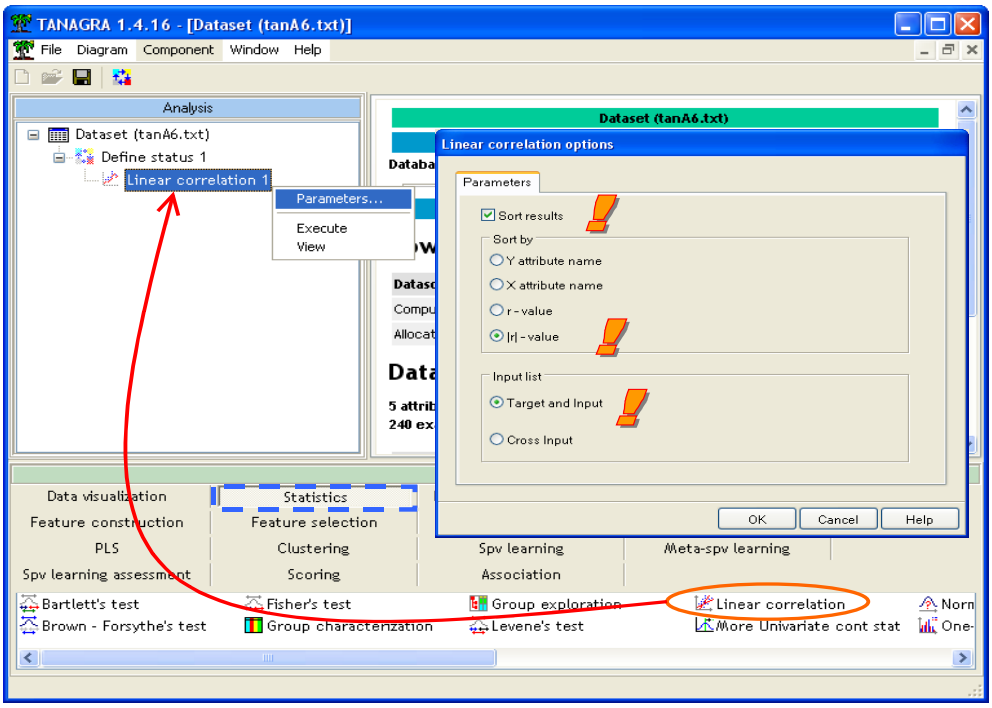
Correlation between the variable of interest and the others

We want to measure the relation between ACCELERATION and the other variables. We add a component in the diagram, using the short cut into the toolbar. We set ACCELERATION as TARGET and all the others variables as INPUT.

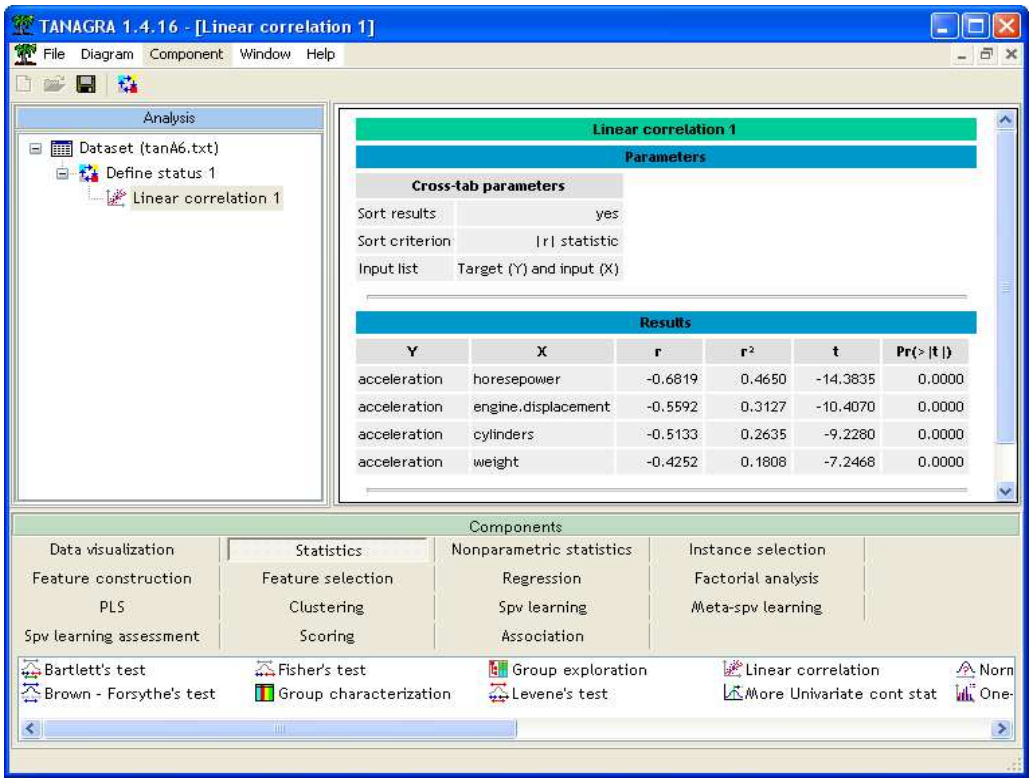


Then we insert the LINEAR CORRELATION component (STATISTICS tab).

We activate the PARAMETERS contextual menu. We state that the results must be sorted according to the absolute value of the correlation coefficient.



We do not modify the INPUT LIST option here. We validate the parameters specification. By clicking on the VIEW menu, we obtain the following results.

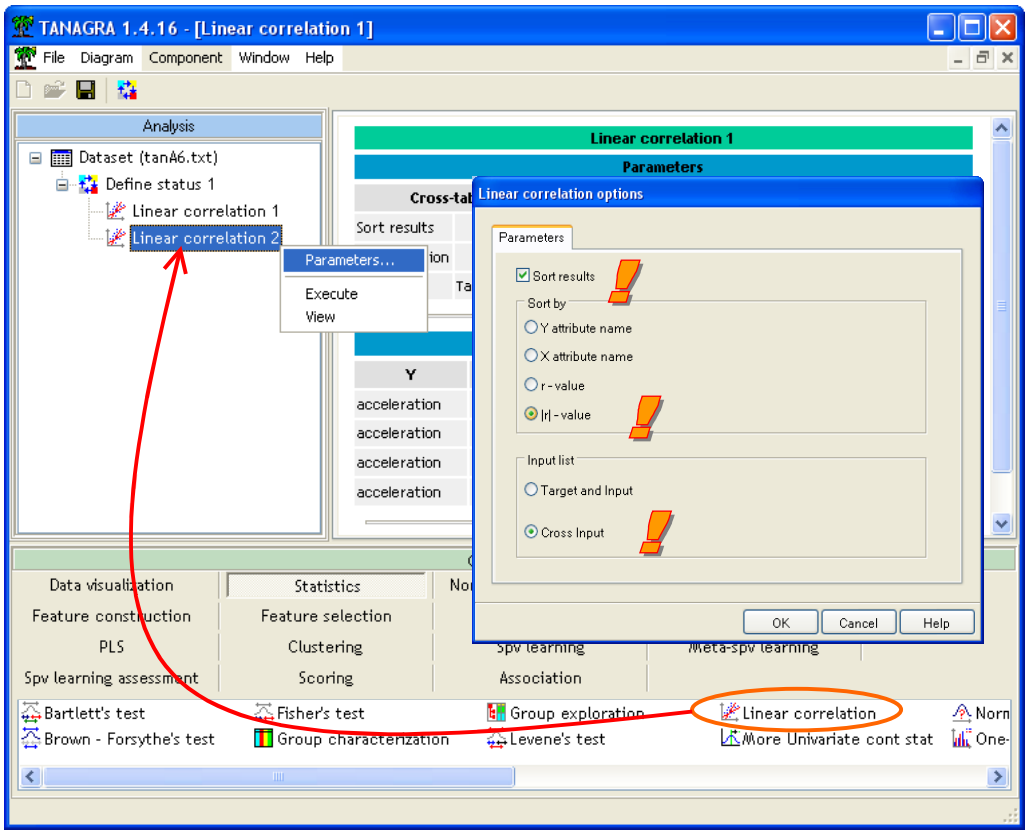


HORSEPOWER is the most correlated variable to ACCELERATION. All the correlations are significant at 1% level.

Cross-correlations

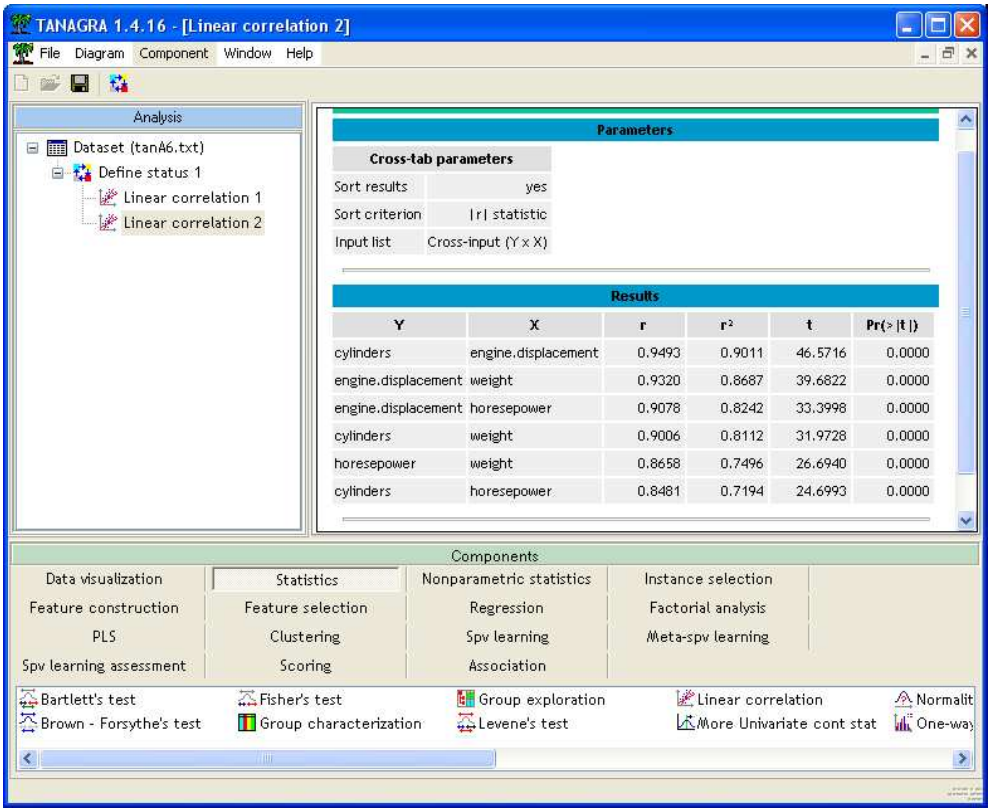
In order to check the collinearities between the candidates' variables, we want to compute the linear correlations between the INPUT attributes.

We insert again the LINEAR CORRELATION component below DEFINE STATUS 1 into the diagram. We activate the menu PARAMETERS, we always want that the results are sorted, but in addition, we state now that calculation must be applied between the INPUT variables.



Even if the number of INPUT variable is large, the computation is rather quick. The possibility to sort the results according to a statistical criterion is useful.

We obtain the following results.



ACCELERATION does not appear in this calculation. We note that CYLINDERS (number of cylinders) and ENGINE.DISPLACEMENT (engine size) are strongly related. This is not a surprise.

We see also that ENGINE.DISPLACEMENT and HORSEPOWERE are highly correlated.

Partial correlation

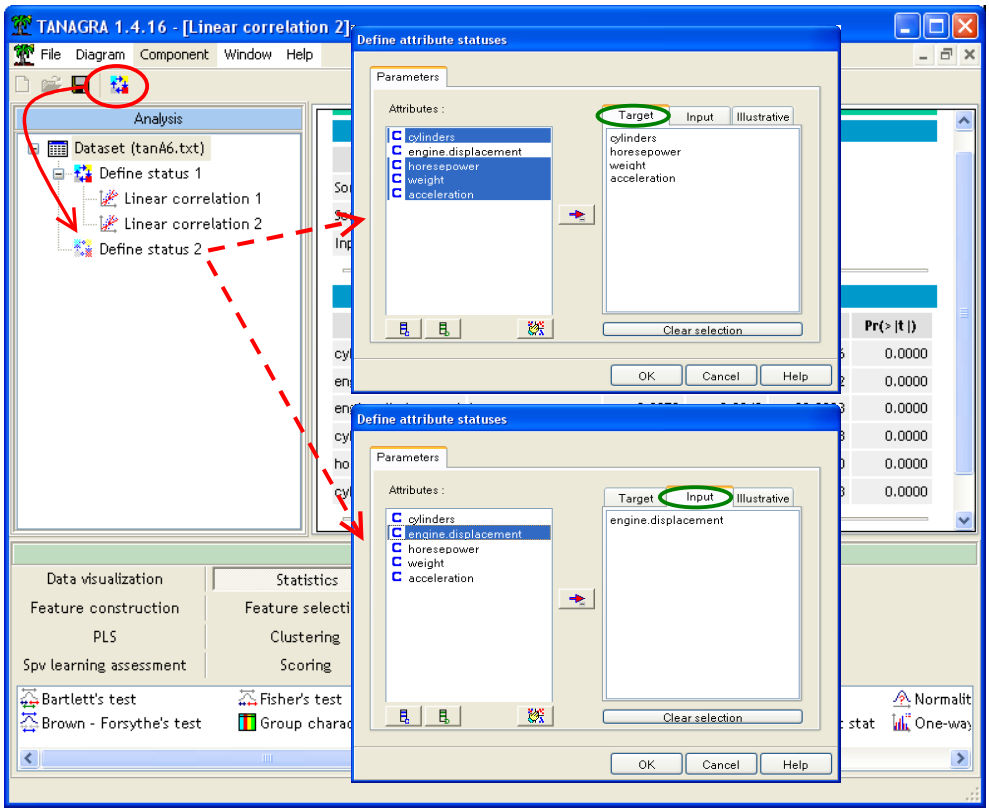
The negative correlation (-0.4252) between ACCELEREATION and WEIGHT seems unusual. It means that the larger is the weight of a car, the faster is acceleration.

This result becomes clear when we see that the WEIGHT is highly correlated to ENGINE.DISPLACEMENT, which is anyway highly correlated to all the variables. We can think that this last variable hides and/or disturbs the correlation between the various variables. We want to proceed again to all analysis by controlling the role of this variable. This is the idea of the partial correlation (see. http://www2.chass.ncsu.edu/garson/pa765/partialr.htm).

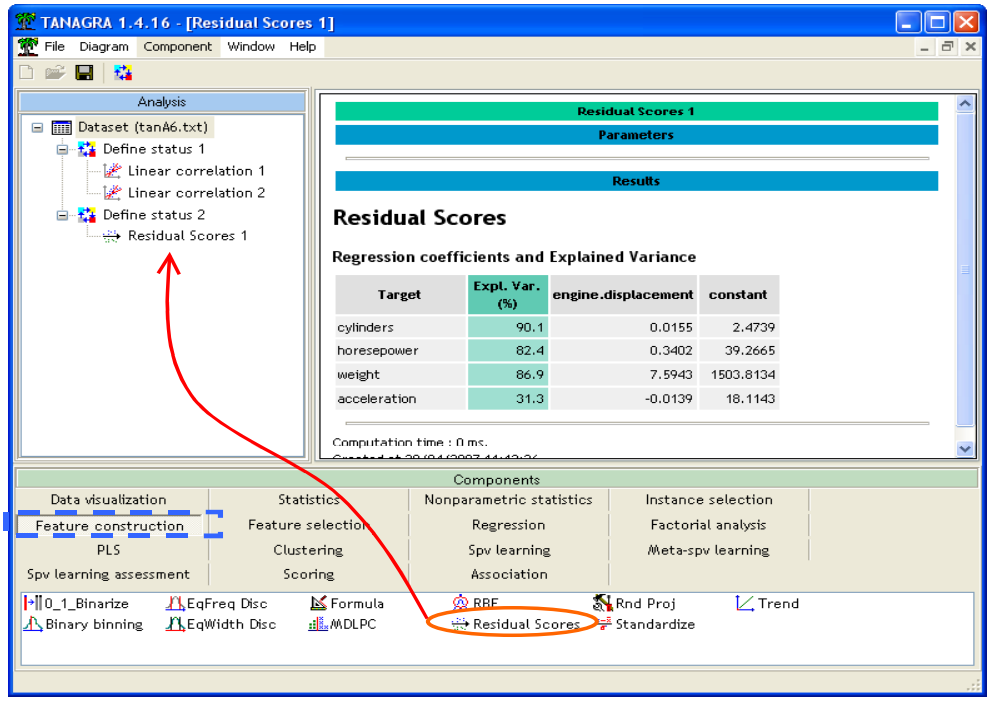
We proceed according to the following steps: (1) we remove to all the variables the effect of the control variable (ENGINE.DISPLACEMENT), by computing the residuals of the linear regression of each variable on this control variable; (2) we perform the above analyses on the residuals corresponding to each variable, where the effect of ENGINE.DISPLACEMENT is removed.

Computing the residuals variables

We insert the DEFINE STATUS component in the root of the diagram. We set as TARGET: CYLINDERS, HORSEPOWER, WEIGHT and ACCELERATION; we set as INPUT ENGINE.DISPLACEMENT.



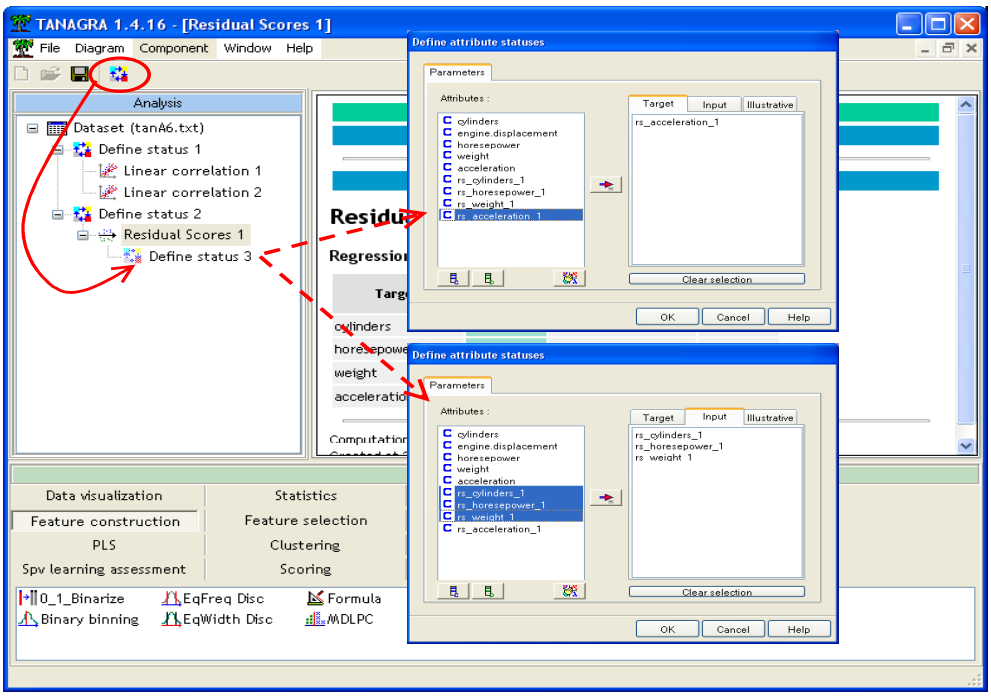
The idea is to remove to the TARGET variables, the effect of the INPUT(s) by computing the residuals of the linear (multiple) regression. We place for this purpose RESIDUAL SCORES component (FEATURE CONSTRUCTION tab). We activate the VIEW menu to obtain the results.



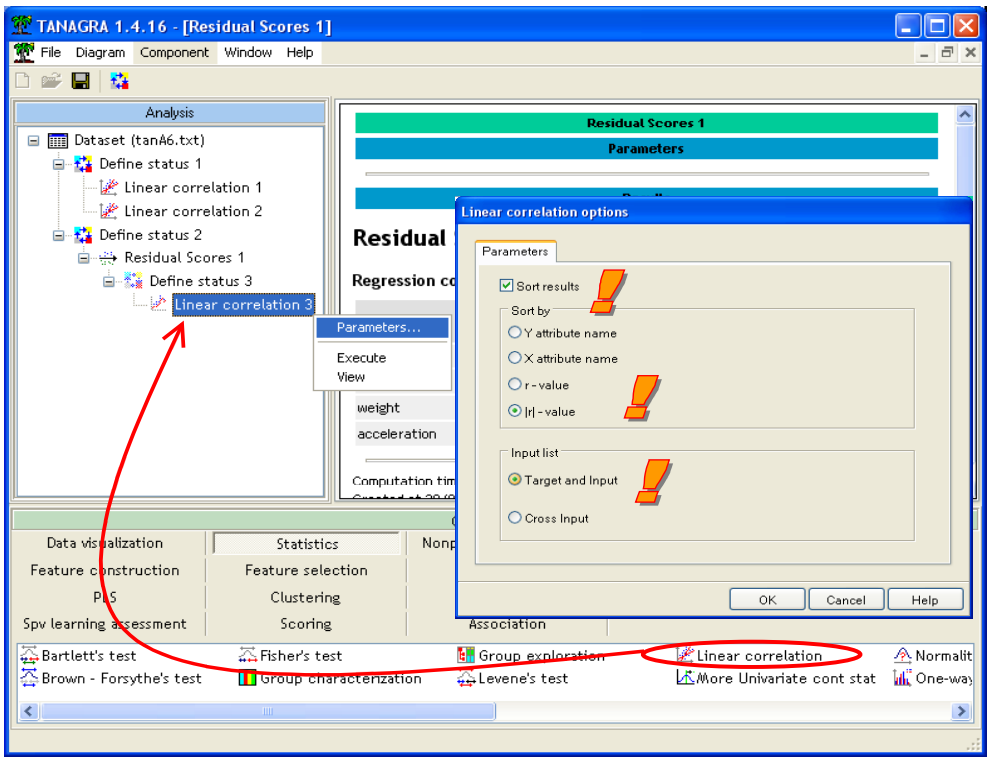
On all regressions, excepted ACCELERATION, the proportion of variance explained is larger than 80%.

Computation of the partial correlation

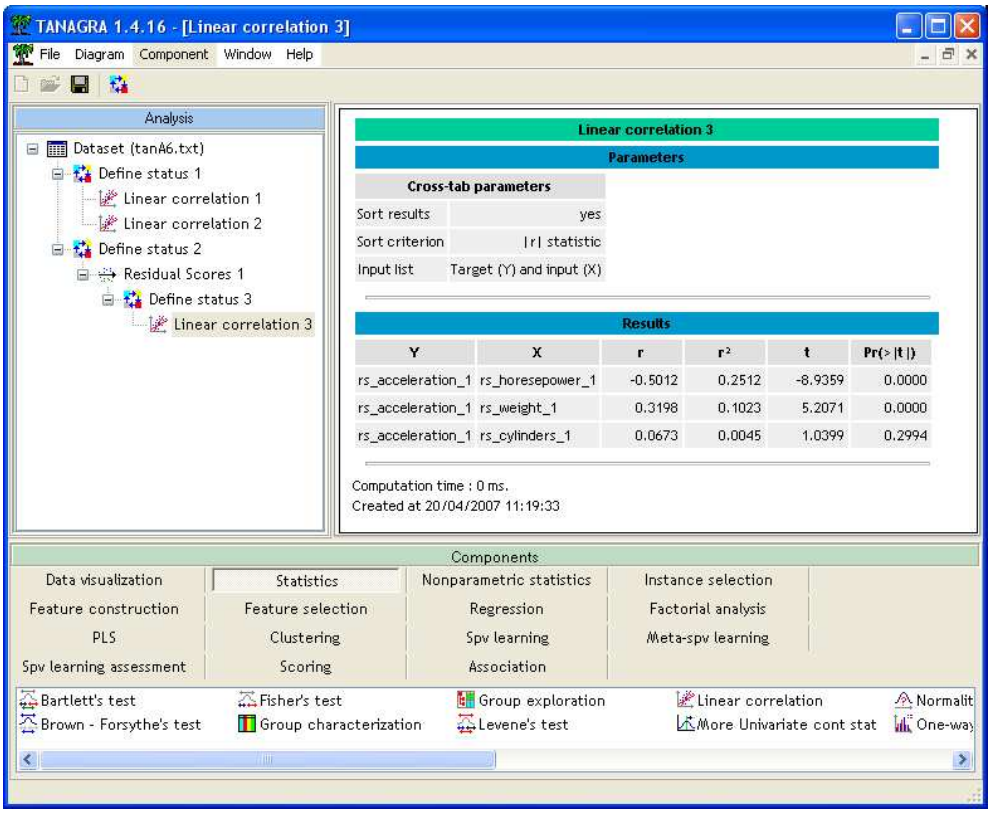
In order to compute partial correlations, we add the DEFINE STATUS component into the diagram. We set as TARGET the residuals of ACCELERATION i.e. RS_ACCELERATION_1, and as INPUT the other residuals (RS_CYLINDERS_1, RS_HORSEPOWER_1 et RS-WEIGHT_1).



Then we insert the LINEAR CORRELATION component. We sort the results according the absolute value of the correlation coefficient.



We obtain the following results.



HORSEPOWER is again the most correlated variable to ACCELERATION. It seems natural, the larger is the horsepower, and the smaller is the time needed to reach a certain speed. And two cars with the same engine displacement can have different horsepower.

The positive correlation (0.3198) between the ACCELERATION and the WEIGHT is natural. The heavy cars are not very quick. It is well known. The statistical indicator confirms this fact now. The negative original correlation computed above was a spurious correlation.

Conclusion

In this tutorial, we show how to compute correlations between a large number of variables, and how to sort the results in order to display first the most relevant results.