

# 1 Subject

## Computing semi-partial correlation with Tanagra.

The semi-partial correlation measures the additional information of an independent variable ( $X$ ), compared with one or several control variables ( $Z_1, \dots, Z_p$ ), that we can use for the explanation of a dependent variable ( $Y$ ).

We can compute the semi-partial correlation in various ways.

The square of the semi-partial correlation can be obtained with the difference between the square of the multiple correlation coefficient of regression  $Y / X, Z_1, \dots, Z_p$  (including  $X$ ) and the same quantity for the regression  $Y / Z_1, \dots, Z_p$  (without  $X$ ).

We can also obtain the semi-partial correlation by computing the residuals of the regression  $X / Z_1, \dots, Z_p$ ; then, we compute the correlation between  $Y$  and these residuals. In other words, we seek to quantify the relationship between  $X$  and  $Y$ , by removing the effect of  $Z$  on the latter. The semi-partial correlation is an asymmetrical measure.

In this tutorial, we show the different ways of producing the semi-partial correlation. We compare the results with the dedicated tool of TANAGRA (SEMI-PARTIAL CORRELATION).

## 2 Dataset

We want to explain the consumption of vehicles ( $Y$ : CONSUMPTION) from horsepower ( $X$ : HORSEPOWER), by controlling the engine size ( $Z_1$ : ENGINE.SIZE) and weight ( $Z_2$ : WEIGHT) effect. The aim is to determine the additional information of HORSEPOWER compared to the control variables.

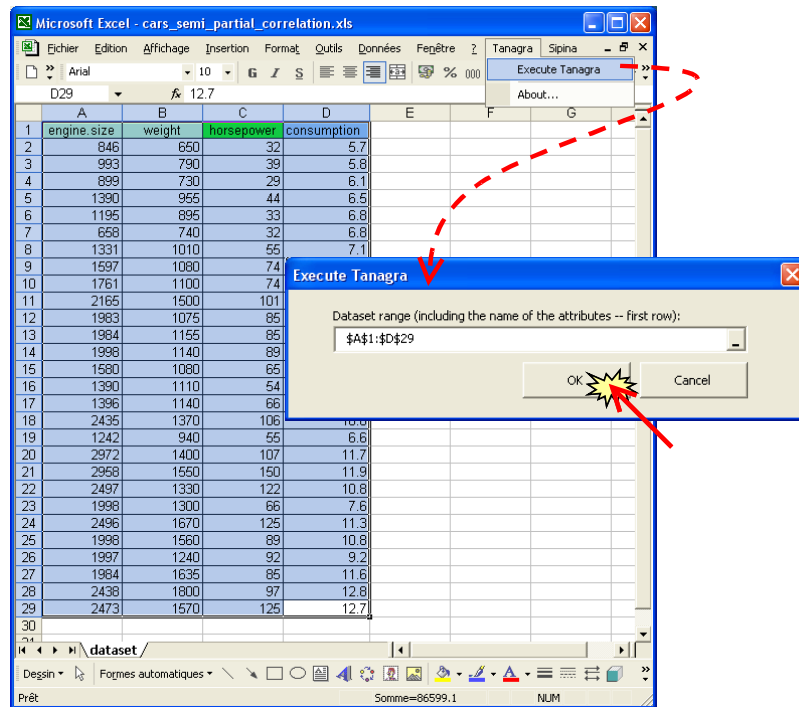
## 3 Computing the semi-partial correlations

### 3.1 Dataset importation

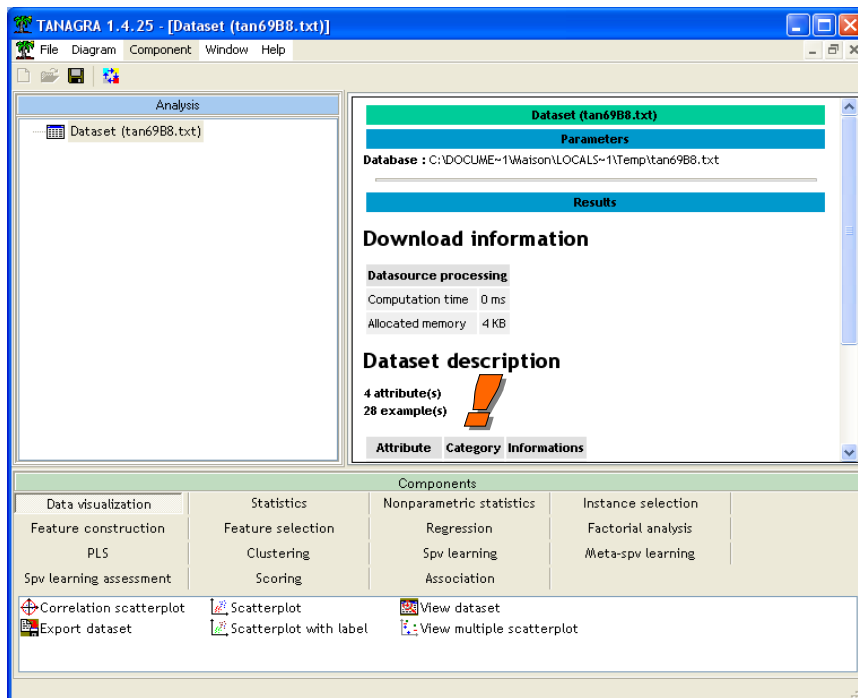
The simplest way in order to create a diagram is to load the dataset in the EXCEL spreadsheet ([http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cars\\_semi\\_partial\\_correlation.xls](http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cars_semi_partial_correlation.xls)). We select the data range and we click on the menu TANAGRA/EXECUTE TANAGRA<sup>1</sup>. After checking the range selection, we click on OK. Tanagra is automatically launched and the dataset transferred.

---

<sup>1</sup> The EXCEL add-in TANAGRA.XLA is available since the version 1.4.11. See the tutorial on the web site for the installation and the utilization of this add-in in your spreadsheet.



We have 28 observations and 4 variables.



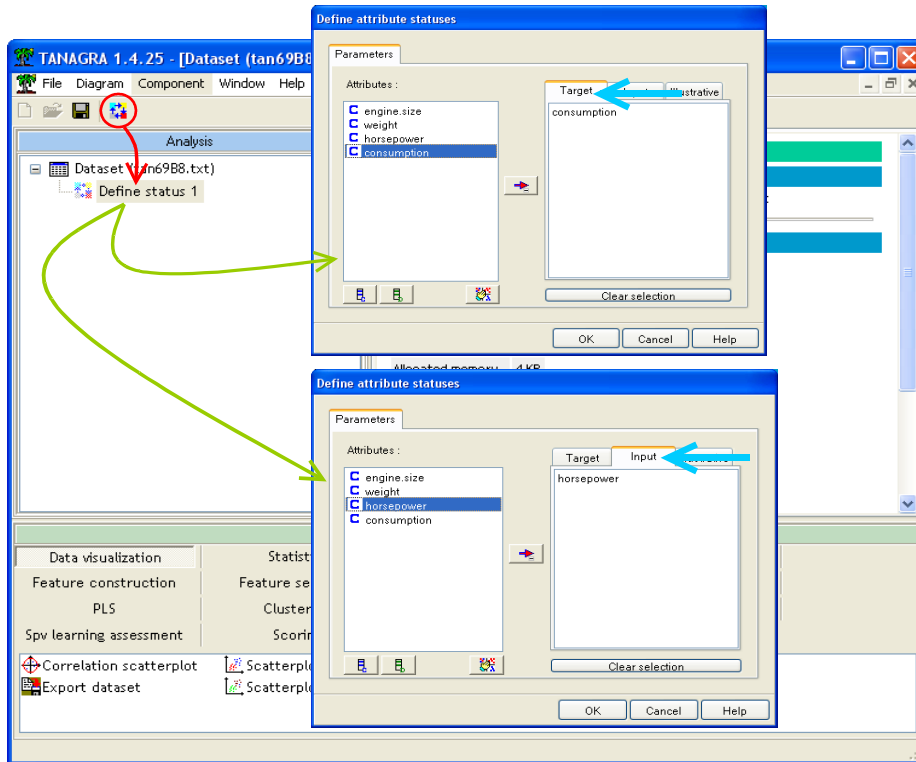
## 3.2 Simple linear regression and correlation

### 3.2.1 Regression and R-square

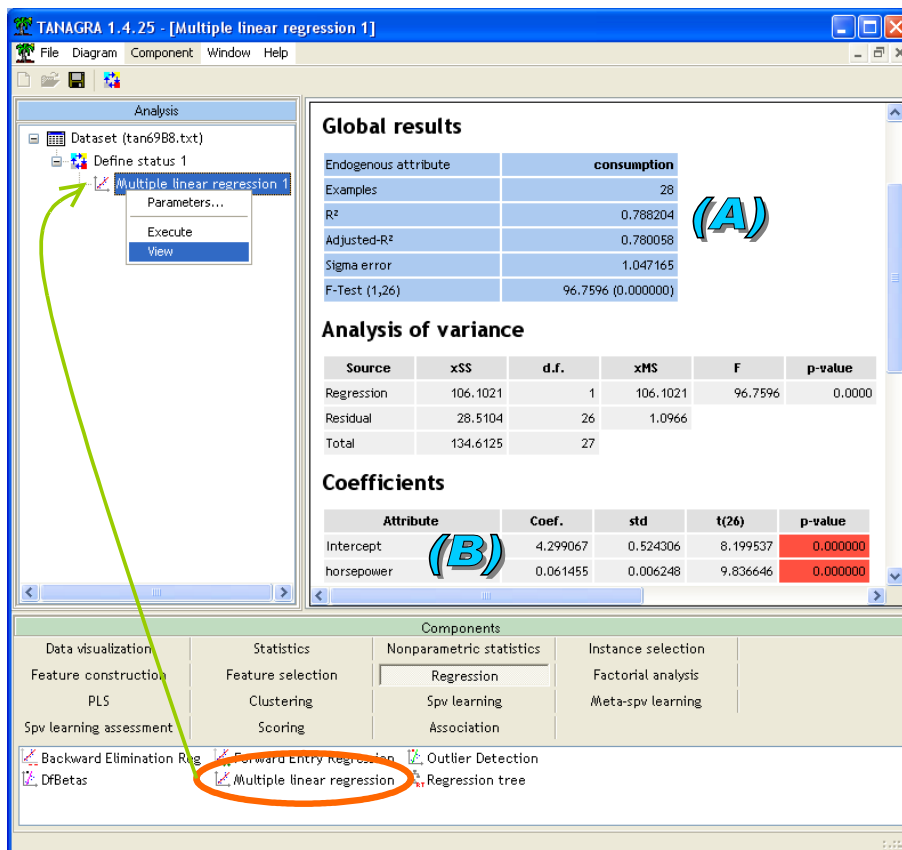
In a first step, we try to evaluate the direct association between HORSEPOWER and CONSUMPTION using a simple linear regression<sup>2</sup>.

<sup>2</sup> [http://en.wikipedia.org/wiki/Linear\\_regression](http://en.wikipedia.org/wiki/Linear_regression)

We insert the DEFINE STATUS component into the diagram; we use the shortcut into the toolbar. We set COSUMPTION as TARGET, HORSEPOWER as INPUT.



We insert next the MULTIPLE LINEAR REGRESSION (REGRESSION tab). We click on the VIEW menu in order to obtain the results.



We highlight two main results: (A) the R-square = 0.7882 i.e. 78.82% of the variance of CONSUMPTION is explained by the regression, it is rather a good result; (B) the HORSEPOWER is highly significant, the t statistic is 9.8366 with a p-value < 0.0001.

In despite of this encouraging result, we note that there are other variables in our dataset. Perhaps, they are useful for the explanation of the CONSUMPTION. We analyze deeply this way later (section 3.2.2).

### 3.2.2 Correlation

Another way to analyze the association between CONSUMPTION and HORSEPOWER is to calculate the Pearson's correlation coefficient. The square of the correlation coefficient can be also interpreted as the proportion of variance explained. We insert the LINEAR CORRELATION (STATISTICS tab) component into the diagram.

The screenshot shows the TANAGRA 1.4.25 interface. The main window displays the results of a linear correlation analysis. The 'Parameters' section shows 'Cross-tab parameters' with 'Sort results' set to 'non' and 'Input list' set to 'Target (Y) and input (X)'. The 'Results' table is as follows:

Y	X	r	r <sup>2</sup>	t	Pr(>  t )
consumption	horsepower	0.8878	0.7882	9.8366	0.0000

Below the table, it indicates 'Computation time : 0 ms.' and 'Created at 14/06/2008 09:47:02'. In the bottom panel, the 'Linear correlation' component is highlighted with a red circle. A green arrow points from the 'View' button in the 'Analysis' tree to the 'Linear correlation' component in the bottom panel.

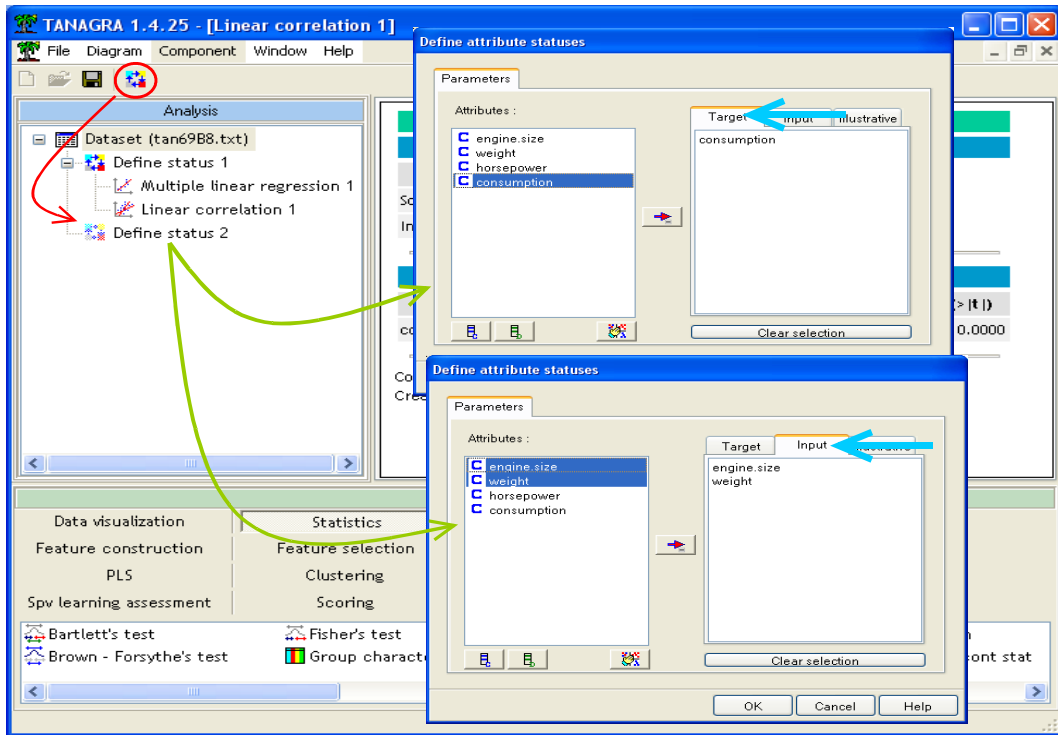
The correlation is positive  $r = 0.8878$ , confirming the sign of regression coefficient. Its square is equal to the coefficient of determination of the simple regression  $r^2 = 0.7882$ . The correlation is significant, we find again the t test value of the coefficient of the regression ( $t = 9.8366$ ). [Testing the significance of the coefficient of the predictor in the simple linear regression and testing the significance of the correlation coefficient are equivalent.](#)

### 3.3 Comparison of regressions

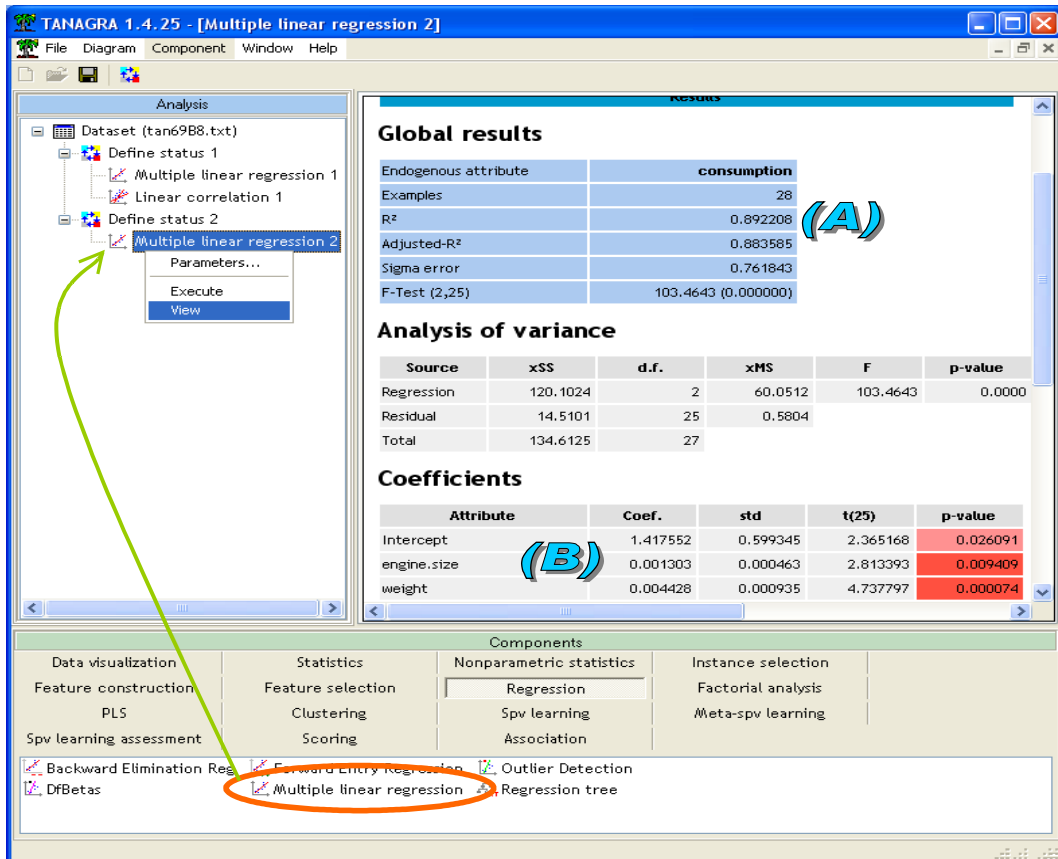
This first analysis is not very satisfactory. Indeed, we know that power is heavily dependent on the engine size, the association between horsepower and weight of cars is often high. We must therefore ask ourselves the following question: What is the additional information provided by the horsepower, compared to the engine size and weight, which would be useful to explain consumption? The semi-partial correlation helps us to answer this question.

### 3.3.1 Regression Y / Z1, Z2

We want to compute the regression  $Y/Z1, Z2$ . We insert the DEFINE STATUS component into the diagram. We set CONSUMPTION as TARGET, ENGINE.SIZE (Z1) and WEIGHT (Z2) as INPUT.



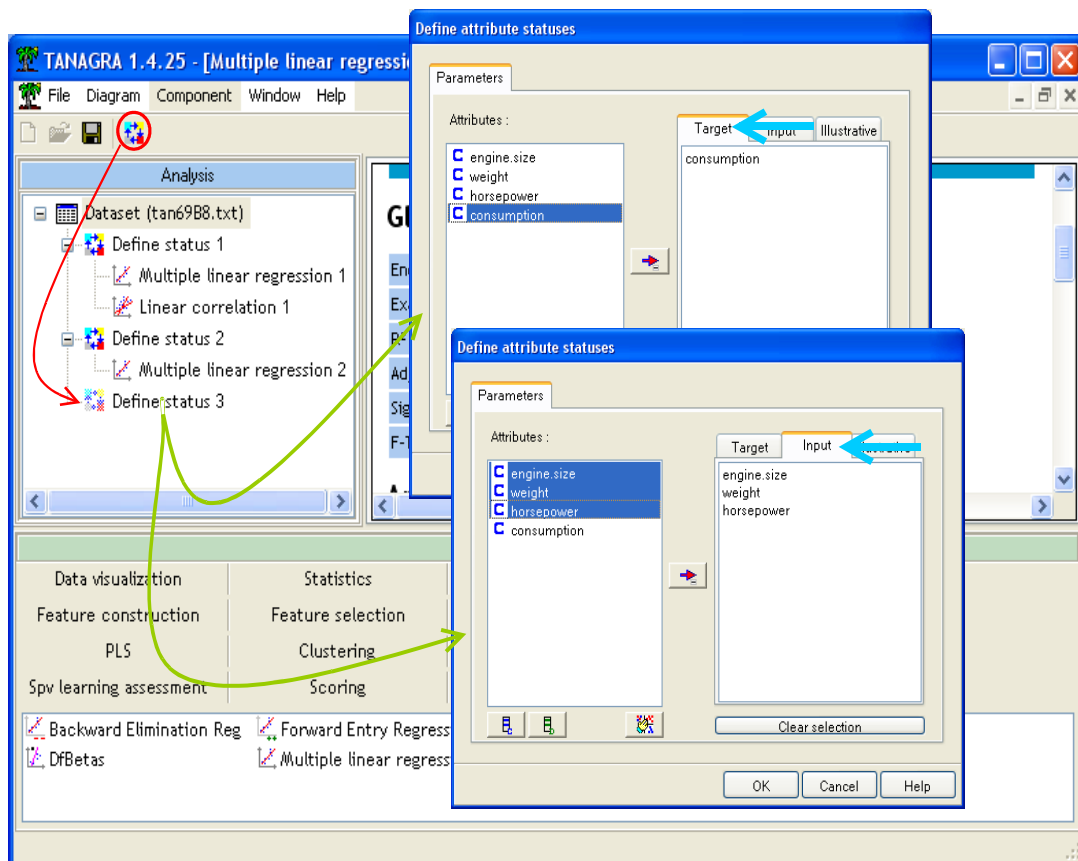
We add the MULTIPLE LINEAR REGRESSION component.



The R-square is high: 89.22% of the variance is explained by the regression (A). All the Z1 and Z2 variables are significant ( $p\text{-value} < 0.01$ ) (B). This regression seems more powerful than the regression using HORSEPOWER only above ( $R\text{-square} = 0.7882$ ; *but the comparison is not really relevant, we have not the same degrees of freedom*).

### 3.3.2 Regression Y / X, Z1, Z2

We want to add now the X (HORESPower) in the preceding regression. We want to compare the R-square, is it significantly improved in this new regression? We insert again the DEFINE STATUS component into the diagram, we set CONSUMPTION as TARGET, the 3 other variables as INPUT.



We perform a new regression analysis.

The screenshot shows the TANAGRA 1.4.25 interface. The 'Analysis' pane on the left shows a tree structure with 'Multiple linear regression 3' selected, indicated by a green arrow. The main window displays the following results:

### Global results

Endogenous attribute	consumption
Examples	28
R <sup>2</sup>	0.899113 (A)
Adjusted-R <sup>2</sup>	0.886502
Sigma error	0.752238
F-Test (3,24)	71.2965 (0.000000)

### Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	121.0318	3	40.3439	71.2965	0.0000
Residual	13.5807	24	0.5659		
Total	134.6125	27			

### Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	1.702048	0.632052	2.692891	0.012712
engine.size (B)	0.000494	0.000780	0.633038	0.532695
weight	0.004229	0.000936	4.518384	0.000141
horsepower	0.018251	0.014240	1.281612	0.212223

The 'Components' pane at the bottom shows 'Multiple linear regression' selected in the 'Regression' category, circled in orange.

The new R-square is 0.899113 (A). The gap between the R-square of the two regressions is not very important i.e.  $d^2 = 0.899113 - 0.892208 = 0.006905$ . **The square root of this gap is the semi-partial correlation i.e.  $d = 0.0831$ .**

The value does not seem relevant. But we cannot define a test of significance now. We see that later (section 4).

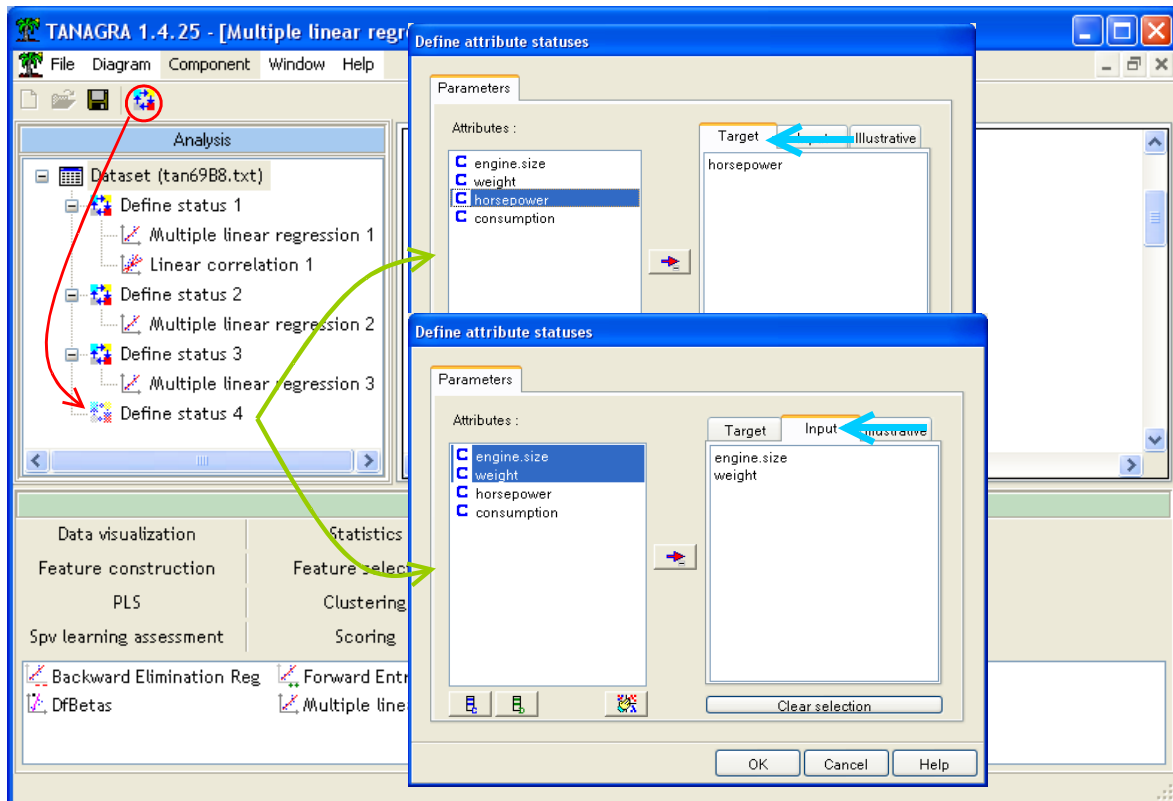
Strangely, two predictors do not seem relevant in the regression (B). In fact, it is because they are highly collinear.

### 3.4 Residuals of the regression $X/Z_1, \dots, Z_p$ and correlation

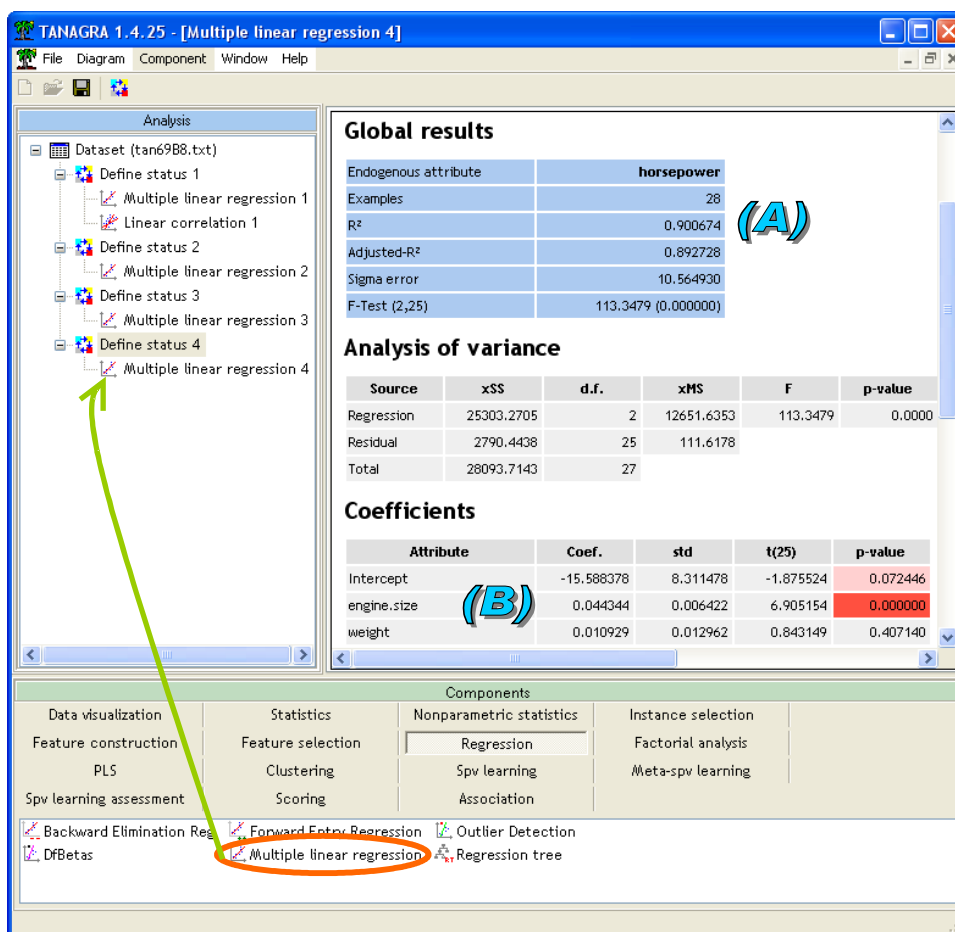
We can calculate the semi-partial correlation in another way. In a first step, we calculate the residuals of the regression of  $X/Z_1, Z_2$ ; we remove from  $X$  the information given by the control variables. Then we calculate the correlation between the residuals and the dependent variable  $Y$ . The obtained coefficient is the semi-partial correlation.

#### 3.4.1 Residuals of $X/Z_1, Z_2$

In order to obtain the residuals, we perform the regression  $X / Z_1, Z_2$ . We insert the DEFINE STATUS component into the diagram. We set HORSEPOWER as TARGET, ENGINE.SIZE and WEIGHT as INPUT.



The regression gives the following results.

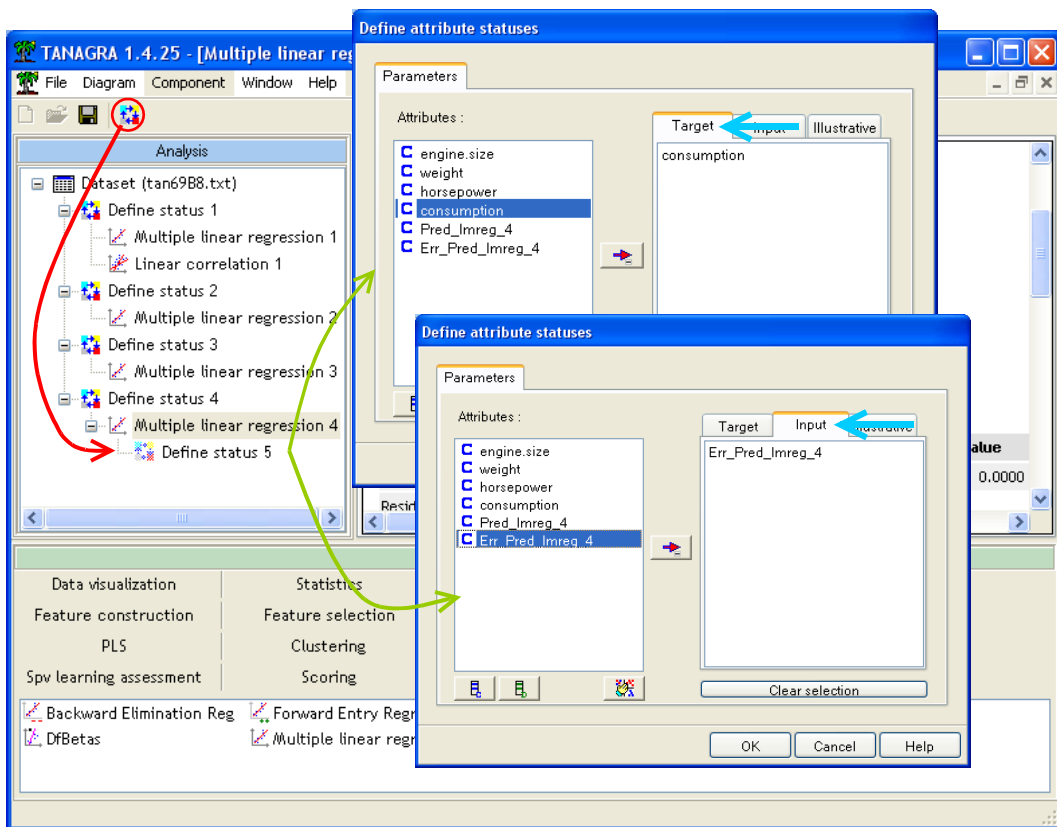


$R^2 = 90.07\%$  of the variance of HORSEPOWER is explained by the regression (A); essentially by the predictor ENGINE.SIZE if we consider the p-value of the test of significance (B). Definitely, HORSEPOWER is highly redundant with the control variables.

### 3.4.2 Correlation between residuals and the dependent variable Y

The component MULTIPLE LINEAR REGRESSION automatically produces two new variables that can be used in subsequent branches of the diagram: the prediction of the dependent variable and residuals of regression. We use the latter now.

We insert the DEFINE STATUS component into the diagram, behind the regression. The two new variables are visible. We place CONSUMPTION as TARGET, residuals ERR\_PRED\_LMREG\_4 as INPUT.



We calculate the correlation between these variables with the LINEAR CORRELATION component.

**Linear correlation 2**

**Parameters**

**Cross-tab parameters**

Sort results: non  
Input list: Target (Y) and input (X)

**Results**

Y	X	r	r <sup>2</sup>	t	Pr(>  t )
consumption	Err_Pred_lmreg_4	0.0831	0.0069	0.4252	0.6742

Computation time : 0 ms.  
Created at 14/06/2008 10:13:29

**Components**

Data visualization	Statistics	Nonparametric statistics	Instance selection
Feature construction	Feature selection	Regression	Factorial analysis
PLS	Clustering	Spv learning	Meta-spv learning
Spv learning assessment	Scoring	Association	

Bartlett's test   Fisher's test   Group exploration   **Linear correlation**  
Brown - Forsythe's test   Group characterization   Levene's test   More Univariate cont stat

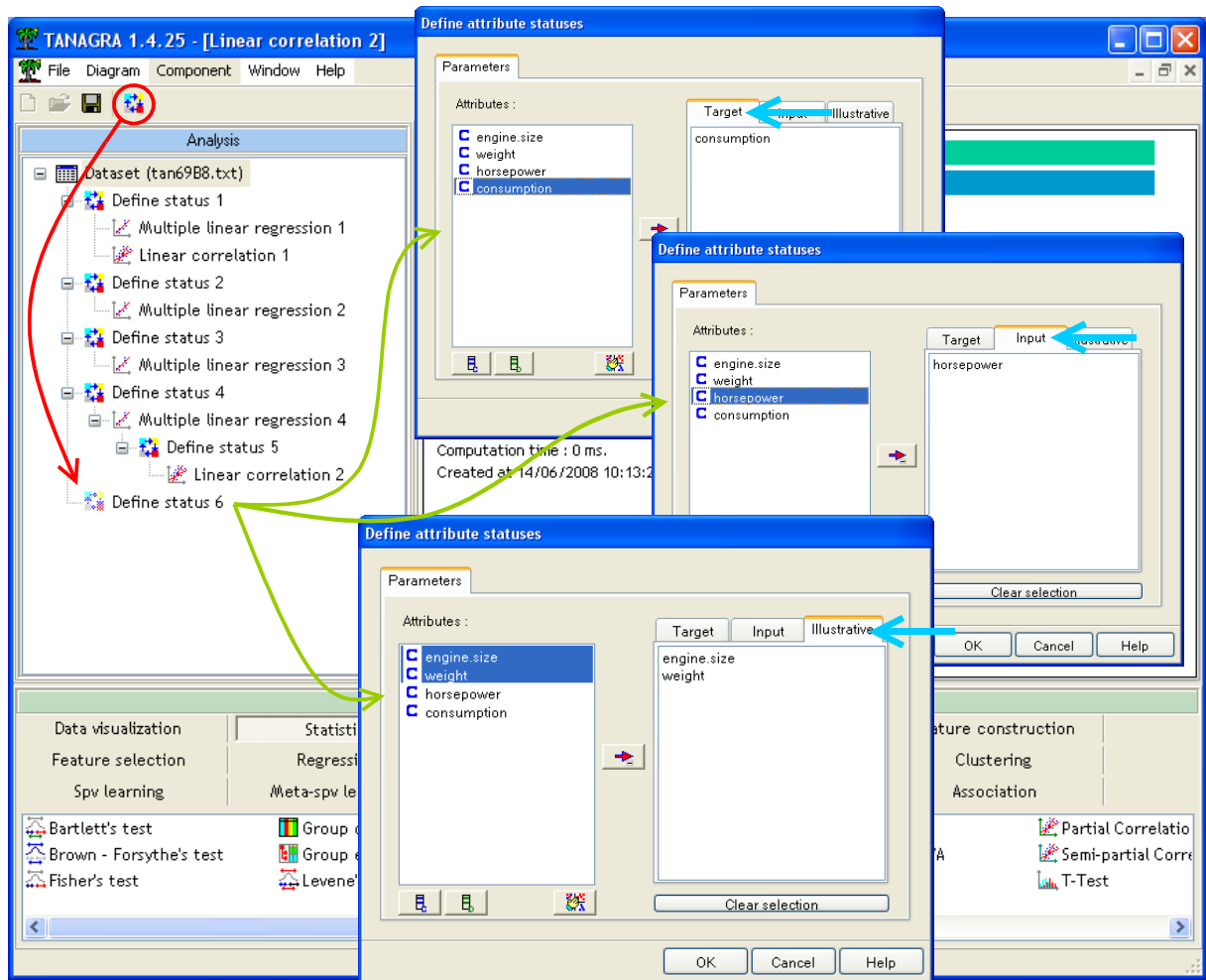
**The coefficient  $r = 0.0831$  is the semi-partial correlation.**

The t-test ( $t = 0.4252$ ) is wrong here. Because, the degrees of freedom is not well computed. The component does not know that one of the variables is a residual, obtained with a regression. It will be corrected in the following component, dedicated to semi-partial correlation calculation.

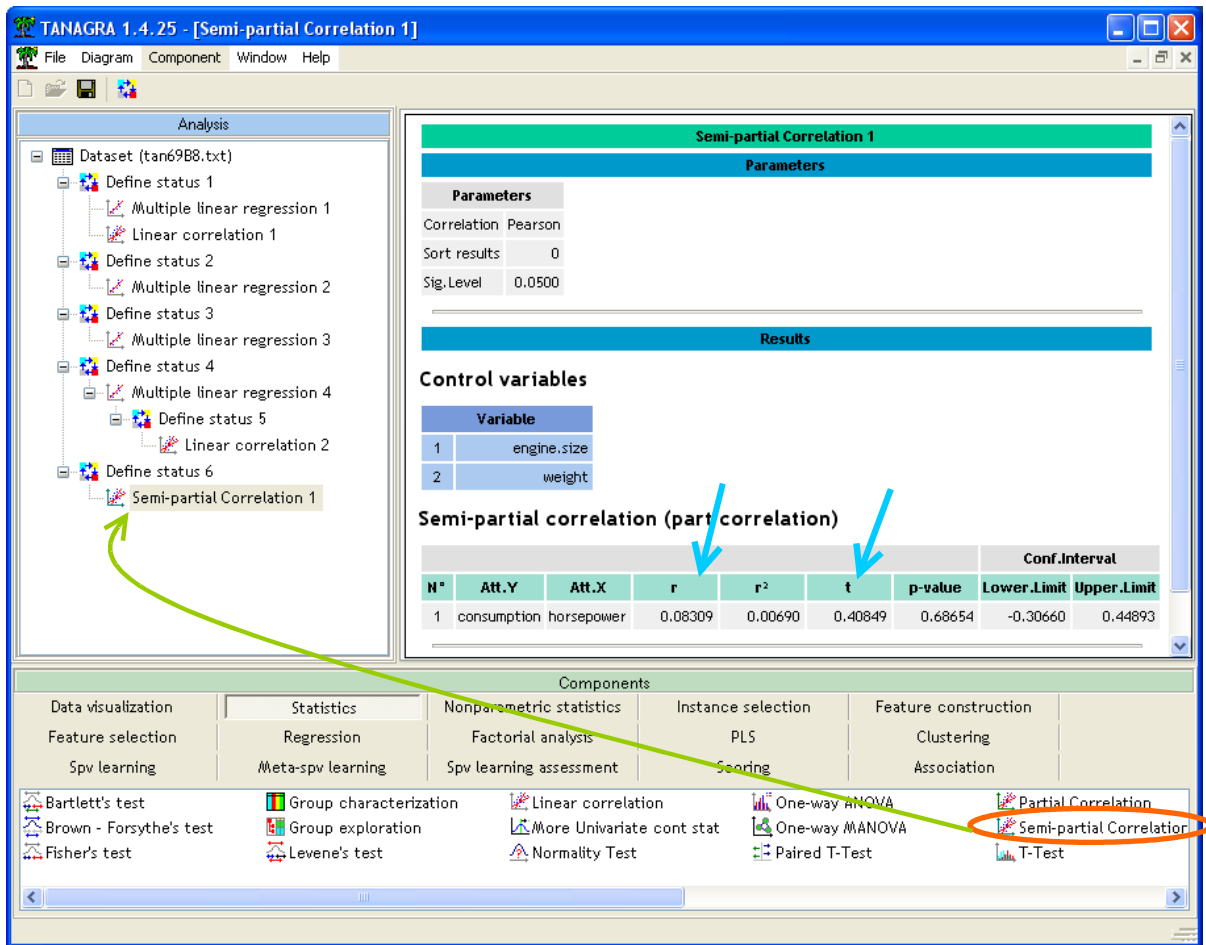
## 4 Using the SEMI-PARTIAL CORRELATION component

In this part of this tutorial, we present the component dedicated to calculating the semi-partial correlation. There are two advantages to use this component: (1) its implementation is facilitated; (2) calculations are explicitly defined, evaluation of degrees of freedom is correct this time.

We insert the DEFINE STATUS component. We set CONSUMPTION as TARGET, HORSEPOWER as INPUT, **ENGINE.SIZE** and **WEIGHT** as **ILLUSTRATIVE**.



We add now the SEMI-PARTIAL CORRELATION (STATISTICS tab) component into the diagram.



The semi-partial correlation is  $r = 0.0831$ . We find again the results above.

But now, the test of significance is correct. The t-statistic is 0.40849, with a degree of freedom equal to  $(28 - 2 - 2 = 24)$ . The p-value of the test is 0.68654. Compared to ENGINE.SIZE and WEIGHT, HORSEPOWER does not give useful information for the explanation of CONSUMPTION.

## 5 Conclusion

In this tutorial, we show the different ways of producing the semi-partial correlation with Tanagra. However, only the dedicated component (SEMI-PARTIAL CORRELATION) directly determines the adequate degrees of freedom for calculating the significance tests and confidence intervals.