1 Topic

Regression analysis with LazStats (OpenStat).

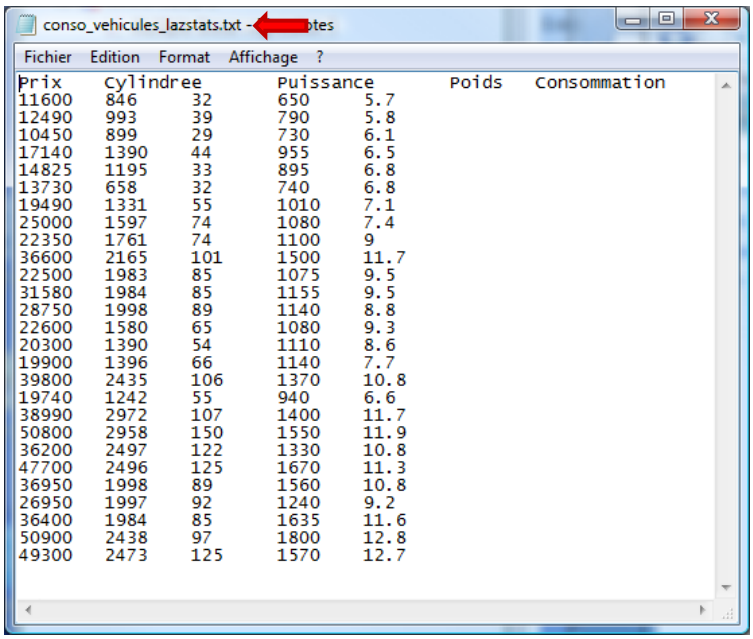
LazStat[1] is a statistical software which is developed by Bill Miller, the father of OpenStat, a well-know tool by statisticians since many years. These are tools of the highest quality. OpenStat is one of tools that I use when I want to validate my own implementations.

Several variants of OpenStat are available. In this tutorial, we study LazStat[2]. It is a version programmed in Lazarus, a development environment which is very similar to Delphi. It is based on the Pascal language. Projects developed in Lazarus benefit to the "write once, compile anywhere" principle i.e. we write our program on an OS (e.g. Windows), but we can compile it on any OS as long as Lazarus and the compiler are available (e.g. Linux). This idea has been proposed by Borland with Kylix[3] some years ago. We could program a project for both Windows and Linux. But, unfortunately, Kylix has been canceled. It seems that the Lazarus is more mature. In addition, it enables us also to compile the same project for the 32 bit and 64 bit versions of an OS.

In this tutorial, we present some functionality of LazStats about regression analysis.

2 Dataset

We use the tab separated text file « conso_vehicules_lazstats.txt ». We want to explain the consumption of cars from their price, engine size, horsepower and weight.



Prix	Cylindree		Puissance		Poids	Consommation
11600	846	32	650	5.7		
12490	993	39	790	5.8		
10450	899	29	730	6.1		
17140	1390	44	955	6.5		
14825	1195	33	895	6.8		
13730	658	32	740	6.8		
19490	1331	55	1010	7.1		
25000	1597	74	1080	7.4		
22350	1761	74	1100	9		
36600	2165	101	1500	11.7		
22500	1983	85	1075	9.5		
31580	1984	85	1155	9.5		
28750	1998	89	1140	8.8		
22600	1580	65	1080	9.3		
20300	1390	54	1110	8.6		
19900	1396	66	1140	7.7		
39800	2435	106	1370	10.8		
19740	1242	55	940	6.6		
38990	2972	107	1400	11.7		
50800	2958	150	1550	11.9		
36200	2497	122	1330	10.8		
47700	2496	125	1670	11.3		
36950	1998	89	1560	10.8		
26950	1997	92	1240	9.2		
36400	1984	85	1635	11.6		
50900	2438	97	1800	12.8		
49300	2473	125	1570	12.7		

This dataset is interesting because we have study it deeply in an e-book available on our website (**in French, "**Pratique de la Régression Linéaire Multiple – Diagnostic et Sélection de Variables**" -- Practical regression analysis - Diagnosis and variable selection**). We can refer to this book for the interpretation of the results of LazStats.

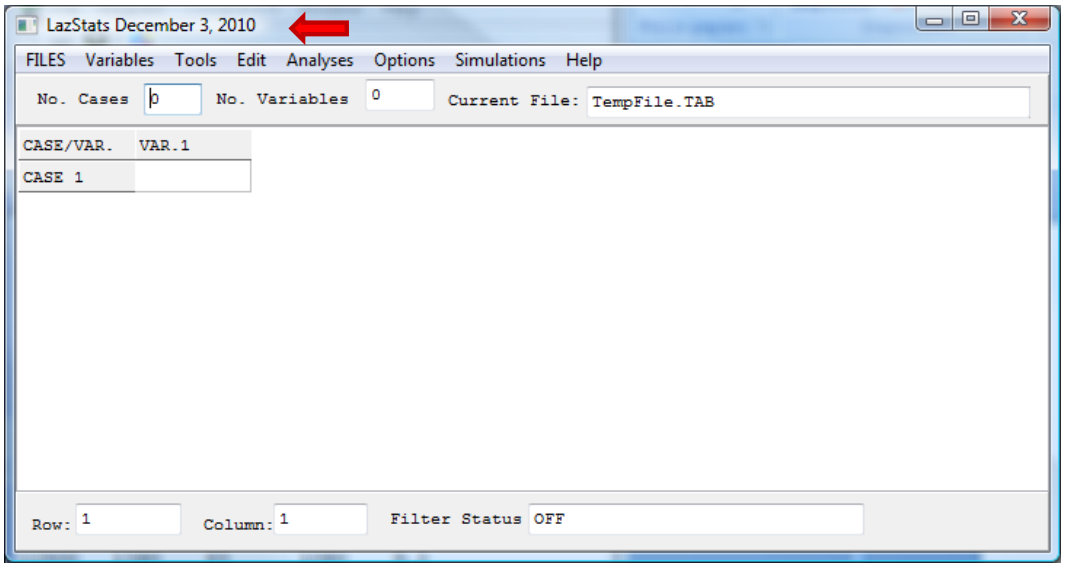
[1] http://www.statprograms4u.com/

[2] http://www.lazarus.freepascal.org/ ; it uses the Free Pascal Compiler - http://www.freepascal.org/

[3] http://en.wikipedia.org/wiki/Kylix_%28software%29

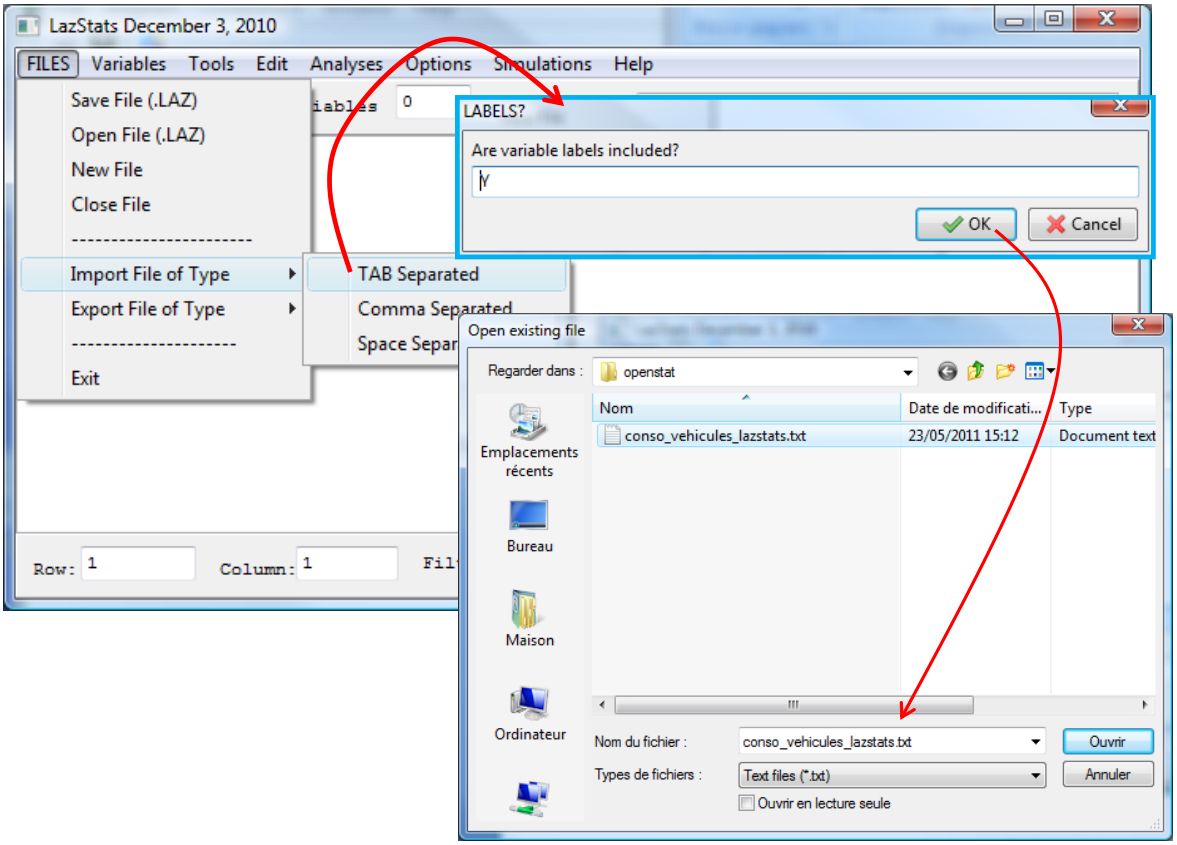
3 Linear regression with LazStats

First, we must download the setup file on the Bill Miller's website. We use the "December 3, 2010" version in this tutorial. After the installation process, we can launch the program.

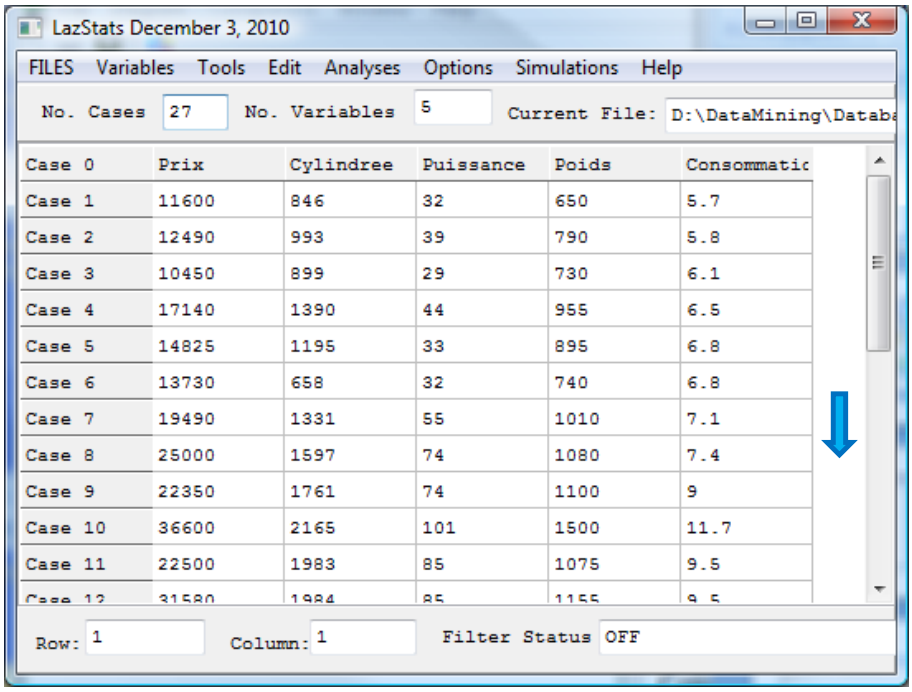


LazStats is driven by menu. The users of statistical tools such as SPSS or STATISTICA are accustomed to this mode of operation. A grid enables to visualize the dataset.

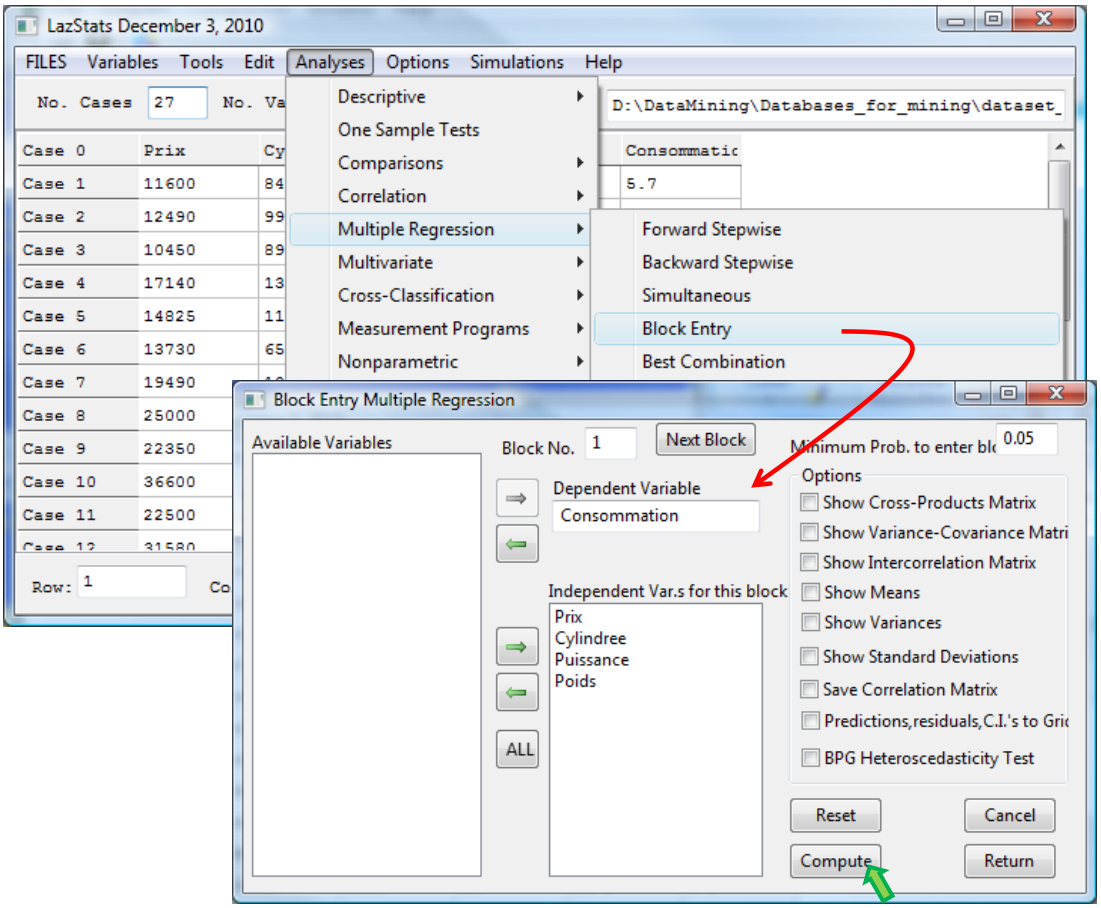
3.1 Importing the data file



We import the data file by clicking on the FILES / IMPORT FILE OF TYPE / TAB SEPARATED menu. LazStats asks if the first row corresponds to the variable name. We validate this choice. Then we set the file name « conso_vehicules_lazstats.txt ». The values are loaded into the grid.

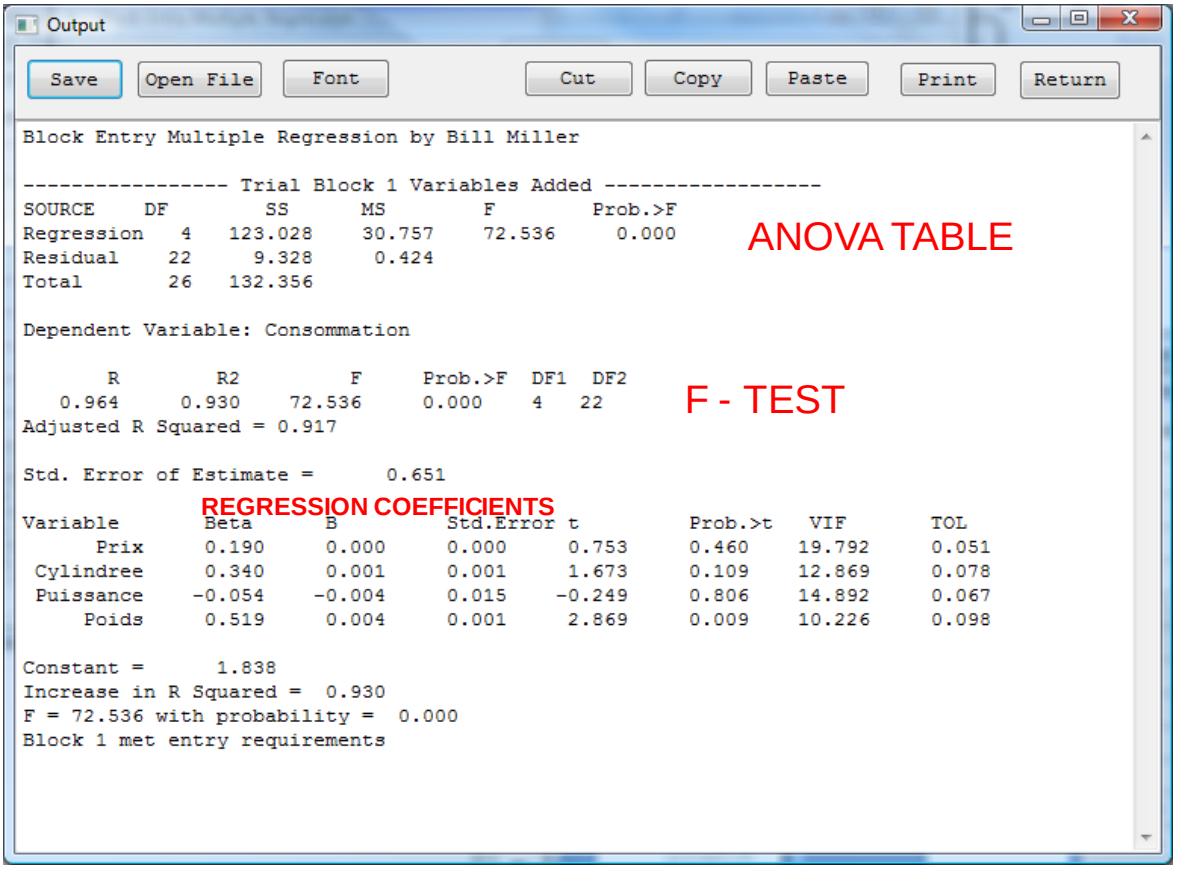


3.2 Regression analysis with all the available exogenous variables



We want to perform a regression with explains the consumption with all the available independent variables. We click on ANALYSES / MULTIPLE REGRESSION / BLOCK ENTRY menu. Into the dialog settings, we set CONSOMMATION as DEPENDENT VARIABLE, the others as INDEPENDENT VARS *FOR THIS BLOCK*. We click on the COMPUTE button.

A visualization window appears. It details the results of the regression[4].



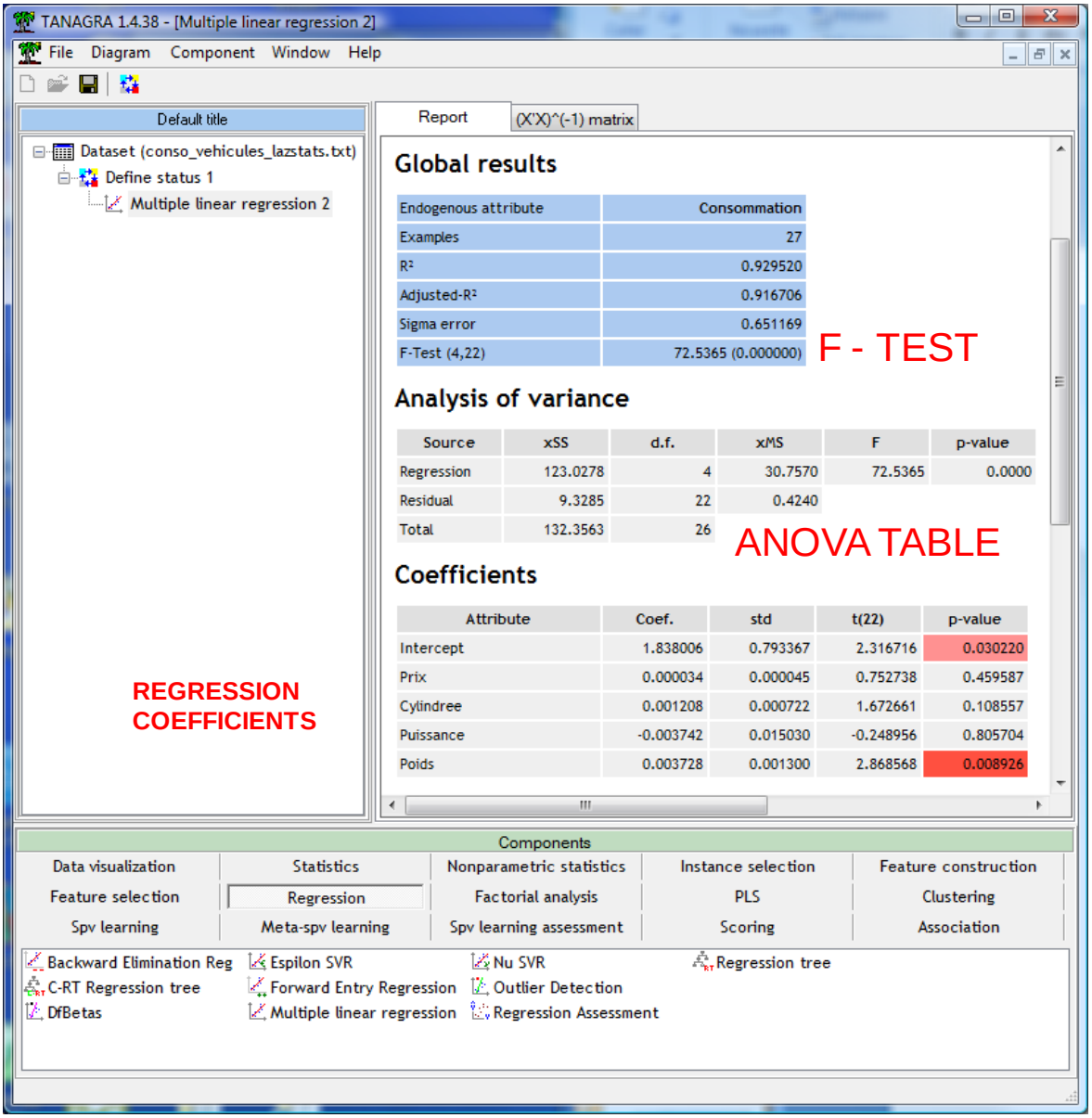
In the upper part of the window, we have the ANOVA table. It is followed by the test of significance of the regression. The R-squared is 0.930. We can say that the model is really good.

Then, we have the regression coefficients table. We have the test of significance for each coefficient. The beta coefficient allows to compare the influence of each factor on the consumption. For instance, we see here that the weight (poids) has the higher (positive) influence on the consumption i.e. when the weight increases, the consumption (liters by 100 km) increases also.

In the lower part, we have the significance of the block of the variables i.e. the adding this group of 4 independent variables increases significantly the R-squared. For our example, because we have only these 4 variables, the results are the same of the test for global significance of the regression.

For comparison, we give the results provided by TANAGRA on the same dataset. We obtain the same results of course.

[44] D. Garson details the results for SPSS. It is interesting to compare the outputs of LazStats with those of SPSS -- http://faculty.chass.ncsu.edu/garson/PA765/regress.htm.



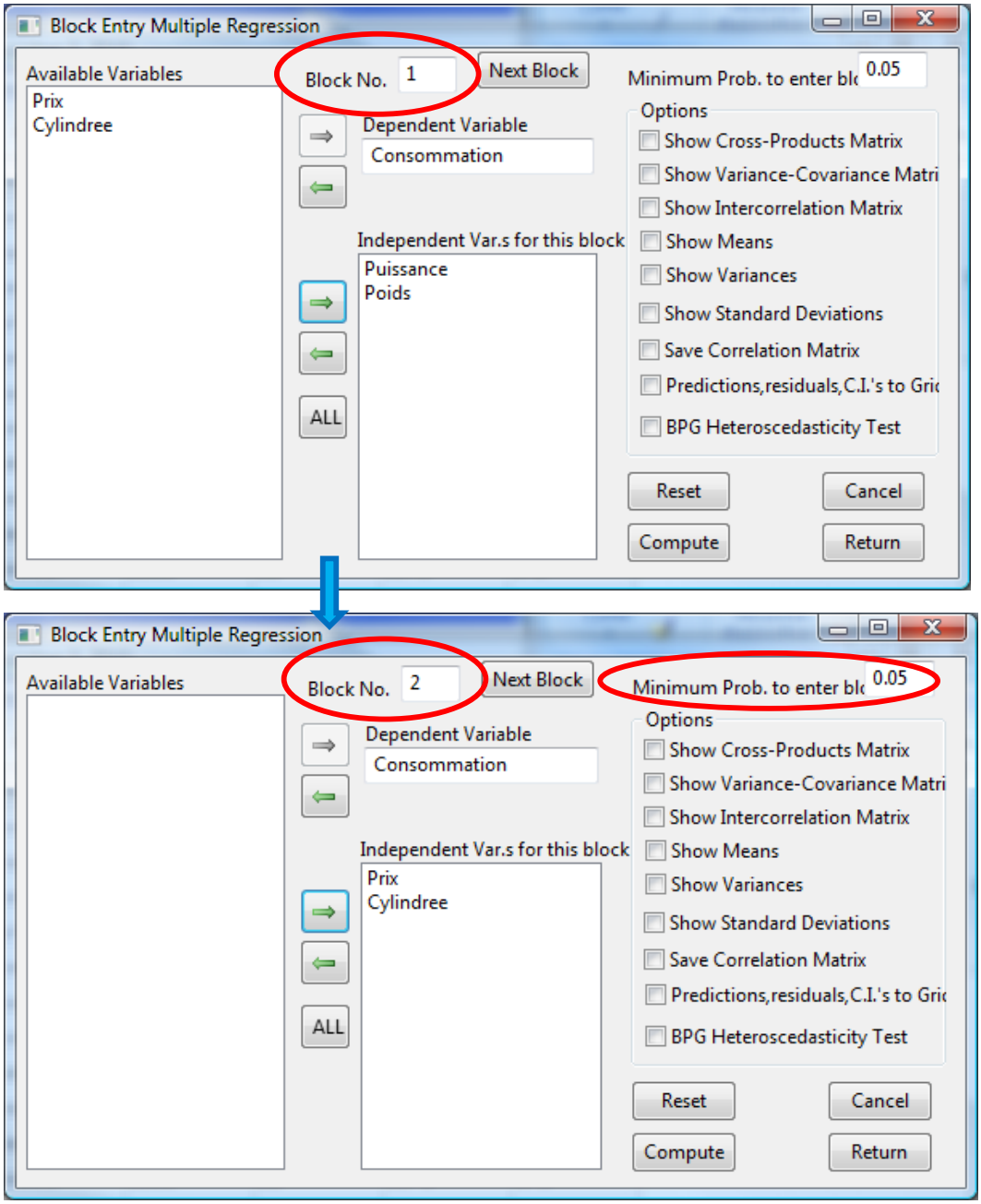
3.3 Block entry regression

"Regression block" is intended to evaluate the relevance of an additional group of variables in the regression. We specify the "blocks" during the selection of candidate independent variables.

We close all the visualization windows by clicking on the RETURN button. We click again on the ANALYSES / MULTIPLE REGRESSION / BLOCK ENTRY menu. We set CONSOMMATION as dependent variable. To evaluate simultaneously the relevance of PUISSANCE and POIDS, we set them as INDEPENDENT VARS FOR THIS BLOCK. Then, we click on the **NEXT BLOCK** button. We add PRIX and CYLINDREE into the "INDEPENDENT VARS FOR THIS BLOCK" list box.

Thus, we evaluate the simultaneous additional relevance of PRIX and CYLINDREE into the regression, knowing that PUISSANCE and POIDS are already selected.

The significance level for the test is 0.05.



We click on the COMPUTE button. The results are more detailed.

```
Block Entry Multiple Regression by Bill Miller

----------------- Trial Block 1 Variables Added ------------------
SOURCE     DF         SS       MS         F         Prob.>F
Regression    2    121.491    60.746    134.184      0.000
Residual     24     10.865     0.453
Total        26    132.356

Dependent Variable: Consommation
       R         R2          F      Prob.>F   DF1   DF2
   0.958      0.918    134.184      0.000     2    24
```

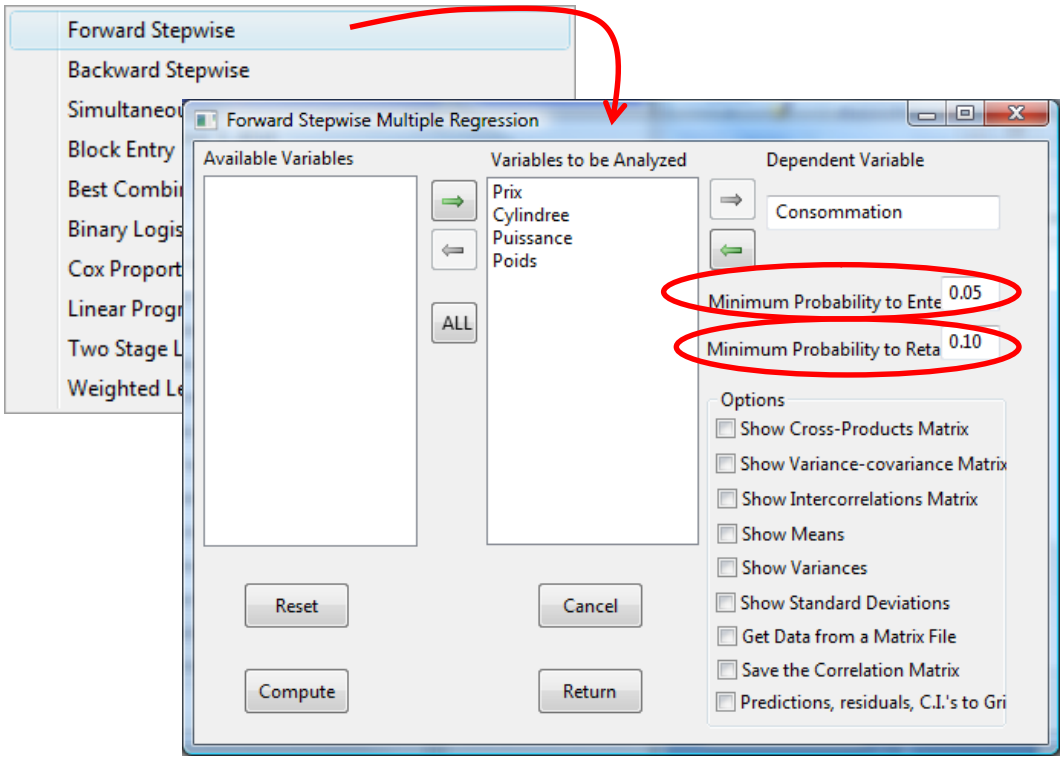
```
Adjusted R Squared = 0.911

Std. Error of Estimate =      0.673

Variable        Beta       B          Std.Error t          Prob.>t    VIF        TOL
 Puissance      0.304      0.021      0.008      2.724      0.012      3.648      0.274
     Poids      0.686      0.005      0.001      6.137      0.000      3.648      0.274

Constant =      1.620
Increase in R Squared =  0.918
F = 134.184 with probability =  0.000
Block 1 met entry requirements

----------------- Trial Block 2 Variables Added ------------------
SOURCE     DF         SS       MS         F         Prob.>F
Regression    4    123.028     30.757     72.536      0.000
Residual     22      9.328      0.424
Total        26    132.356

Dependent Variable: Consommation

       R         R2          F      Prob.>F   DF1   DF2
   0.964      0.930     72.536      0.000     4    22
Adjusted R Squared = 0.917

Std. Error of Estimate =      0.651

Variable        Beta       B          Std.Error t          Prob.>t    VIF        TOL
 Puissance     -0.054     -0.004      0.015     -0.249      0.806     14.892      0.067
     Poids      0.519      0.004      0.001      2.869      0.009     10.226      0.098
      Prix      0.190      0.000      0.000      0.753      0.460     19.792      0.051
 Cylindree      0.340      0.001      0.001      1.673      0.109     12.869      0.078

Constant =      1.838
Increase in R Squared =  0.012
F =  1.812 with probability =  0.187
Block 2 did not meet entry requirements
```

The coefficients of PUISSANCE and POIDS are significant at the 5% level, with F = 134.184 and a p-value = 0.000. At the next step, the regression is always globally significant. But, we note that the explanation provided by the two additional variables is not significant (F = 1.812, p-value = 0.187).

Although it seems a little complicated at first, this feature is particularly useful when we want to organize into a hierarchy the introduction of exogenous variables during the selection process. For instance, the introduction of a variable is only desirable if another variable is already present in the model to be consistent with the domain knowledge.

3.4 Stepwise regression – Forward variable selection

LazStats can perform a stepwise search for the selection of the most relevant variables in the regression[5]. We specify two significance levels: the first for the adding of a variable into the regression (0.05 for our dataset); the second for the removing (0.10). The process is stopped when we cannot either add or remove a variable into the model.



We click on the ANALYSES / MULTIPLE REGRESSION / FORWARD STEPWISE menu to perform a **forward selection process**. We specify the variables to use. We click on the COMPUTE button.

```
Stepwise Multiple Regression by Bill Miller
----------------- STEP 1 -----------------
SOURCE     DF         SS        MS         F         Prob.>F
Regression   1    118.133   118.133   207.632     0.000
Residual    25     14.224     0.569
Total       26    132.356

Dependent Variable: Consommation

       R         R2          F      Prob.>F   DF1   DF2
   0.945      0.893    207.632      0.000     1    25
Adjusted R Squared = 0.888

Std. Error of Estimate =      0.754
```

[5] http://faculty.chass.ncsu.edu/garson/PA765/regress.htm#stepwise1

Variable	Beta	B	Std.Error	t	Prob.>t	VIF	TOL
Poids	0.945	0.007	0.000	14.409	0.000	1.000	1.000

Constant = 1.035

Candidates for entry in next step.

Candidate	Partial	F Statistic	Prob.	DF1	DF2
Prix	0.4567	6.3243	0.0190	1	24

SOURCE	DF	SS	MS	F	Prob.>F
Regression	2	121.099	60.549	129.088	0.000
Residual	24	11.257	0.469		
Total	26	132.356			

Candidate	Partial	F Statistic	Prob.	DF1	DF2
Cylindree	0.5719	11.6631	0.0023	1	24

SOURCE	DF	SS	MS	F	Prob.>F
Regression	2	122.784	61.392	153.927	0.000
Residual	24	9.572	0.399		
Total	26	132.356			

Candidate	Partial	F Statistic	Prob.	DF1	DF2
Puissance	0.4859	7.4196	0.0118	1	24

Variable Cylindree will be added

---------- STEP 2 ----------

SOURCE	DF	SS	MS	F	Prob.>F
Regression	2	122.784	61.392	153.927	0.000
Residual	24	9.572	0.399		
Total	26	132.356			

Dependent Variable: Consommation

R	R2	F	Prob.>F	DF1	DF2
0.963	0.928	153.927	0.000	2	24

Adjusted R Squared = 0.922

Std. Error of Estimate = 0.632

Variable	Beta	B	Std.Error	t	Prob.>t	VIF	TOL
Poids	0.627	0.005	0.001	5.812	0.000	3.867	0.259
Cylindree	0.369	0.001	0.000	3.415	0.002	3.867	0.259

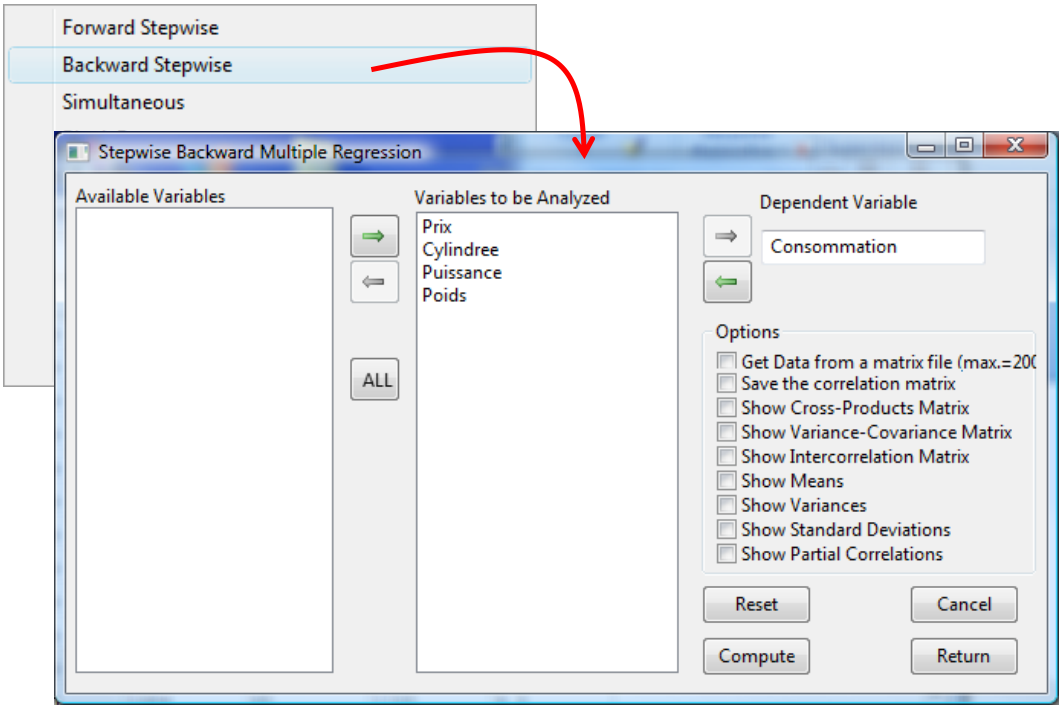
```
Constant =      1.392

Candidates for entry in next step.
Candidate  Partial  F Statistic  Prob.  DF1  DF2
SOURCE     DF         SS       MS         F         Prob.>F
Regression   3    123.002     41.001    100.806      0.000
Residual    23      9.355      0.407
Total       26    132.356

Prix         0.1507    0.5344    0.4721      1   23
SOURCE     DF         SS       MS         F         Prob.>F
Regression   3    122.788     40.929     98.380      0.000
Residual    23      9.569      0.416
Total       26    132.356

Puissance    0.0188    0.0082    0.9288      1   23
No further steps meet criterion for entry.

-------------FINAL STEP-----------
SOURCE     DF         SS       MS         F         Prob.>F
Regression   2    122.784     61.392    153.927      0.000
Residual    24      9.572      0.399
Total       26    132.356

Dependent Variable: Consommation

       R         R2          F      Prob.>F  DF1  DF2
   0.963      0.928    153.927      0.000    2   24
Adjusted R Squared = 0.922

Std. Error of Estimate =      0.632

Variable       Beta       B          Std.Error t          Prob.>t    VIF        TOL
     Poids     0.627      0.005      0.001     5.812      0.000      3.867      0.259
 Cylindree     0.369      0.001      0.000     3.415      0.002      3.867      0.259

Constant =      1.392
```

We observe that POIDS is the first selected variable, and then CYLINDREE is added. The next variable, PUISSANCE, is not significant (F = 0.0082, p-value = 0.9288).

3.5 Backward selection process

The backward regression begins with all the available independent variables. It removes sequentially the least relevant variables according to the test for significance of coefficients. It stops when all the remaining variables are relevant.

We click on the ANALYSES / MULTIPLE REGRESSION / BACKWARD STEPWISE menu to launch the backward regression.



We click on the COMPUTE button. We obtain the following results.

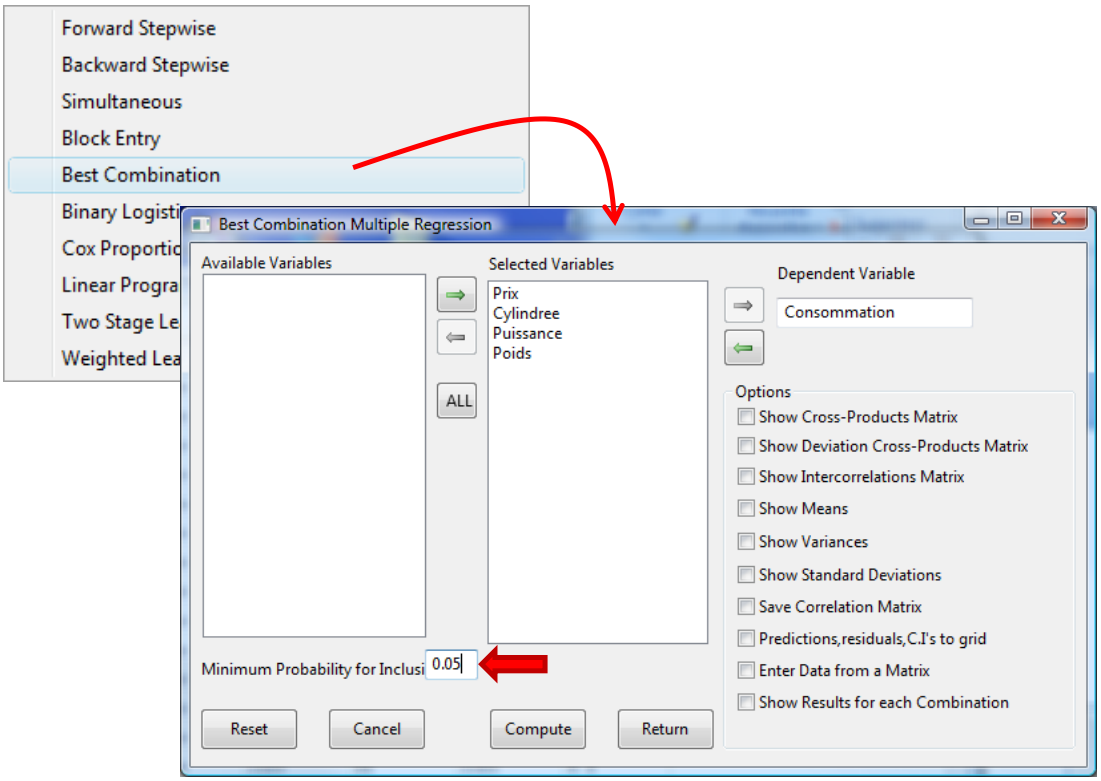
```
Step Backward Multiple Regression by Bill Miller

----------------- STEP   1 ------------------
Determinant of correlation matrix =   0.0001

SOURCE     DF         SS       MS         F         Prob.>F
Regression    4    123.028    30.757    72.536      0.000
Residual     22      9.328     0.424
Total        26    132.356

Dependent Variable: Consommation

       R         R2          F      Prob.>F   DF1   DF2
   0.964      0.930     72.536      0.000     4    22
Adjusted R Squared = 0.917

Std. Error of Estimate =      0.651

Variable        Beta       B          Std.Error t          Prob.>t   VIF        TOL
      Prix      0.190      0.000      0.000      0.753      0.460     19.792     0.051
 Cylindree      0.340      0.001      0.001      1.673      0.109     12.869     0.078
 Puissance     -0.054     -0.004      0.015     -0.249      0.806     14.892     0.067
     Poids      0.519      0.004      0.001      2.869      0.009     10.226     0.098
```

```
Constant =      1.838
Variable 3 (Puissance) eliminated

----------------- STEP   2 ------------------
Determinant of correlation matrix =   0.0011

SOURCE    DF        SS       MS        F        Prob.>F
Regression   3   123.002    41.001   100.806     0.000
Residual    23     9.355     0.407
Total       26   132.356

Dependent Variable: Consommation

       R        R2         F     Prob.>F  DF1  DF2
   0.964     0.929   100.806     0.000    3   23
Adjusted R Squared = 0.920

Std. Error of Estimate =      0.638

Variable       Beta      B         Std.Error t         Prob.>t   VIF       TOL
      Prix     0.162     0.000     0.000     0.731     0.472    16.001     0.062
 Cylindree     0.304     0.001     0.000     2.163     0.041     6.423     0.156
     Poids     0.530     0.004     0.001     3.072     0.005     9.676     0.103

Constant =      1.824
Variable 1 (Prix) eliminated

----------------- STEP   3 ------------------
Determinant of correlation matrix =   0.0187

SOURCE    DF        SS       MS        F        Prob.>F
Regression   2   122.784    61.392   153.927     0.000
Residual    24     9.572     0.399
Total       26   132.356

Dependent Variable: Consommation

       R        R2         F     Prob.>F  DF1  DF2
   0.963     0.928   153.927     0.000    2   24
Adjusted R Squared = 0.922

Std. Error of Estimate =      0.632

Variable       Beta      B         Std.Error t         Prob.>t   VIF       TOL
 Cylindree     0.369     0.001     0.000     3.415     0.002     3.867     0.259
     Poids     0.627     0.005     0.001     5.812     0.000     3.867     0.259
```

```
Constant =      1.392
Variable 1 (Cylindree) eliminated
----------------- STEP   4 ------------------
Determinant of correlation matrix =   0.1075
SOURCE     DF         SS       MS         F         Prob.>F
Regression   1    118.133   118.133   207.632      0.000
Residual    25     14.224     0.569
Total       26    132.356
Dependent Variable: Consommation
       R         R2          F      Prob.>F   DF1   DF2
   0.945      0.893    207.632      0.000     1    25
Adjusted R Squared = 0.888

Std. Error of Estimate =      0.754

Variable        Beta       B          Std.Error t          Prob.>t    VIF        TOL
     Poids      0.945      0.007      0.000     14.409     0.000      1.000      1.000

Constant =      1.035
```

3.6 « Best combination »

The best combination seems very similar to the forward process into LazStats. We have to specify the significance level for the inclusion of a variable (0.05).



We click on the COMPUTE button. We obtain the details of the process.

```
Best Combination Multiple Regression by Bill Miller
```

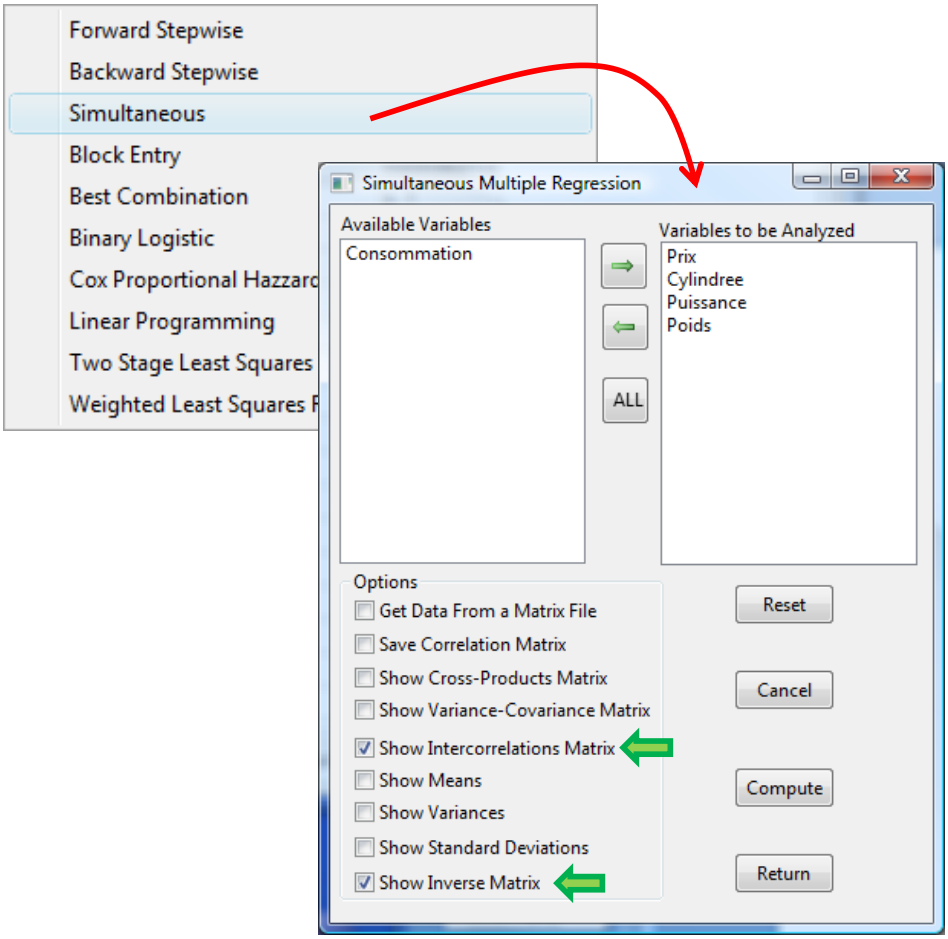
```
Variables entered in step  1
 4 Poids

Squared Multiple Correlation = 0.8925
Dependent variable = Consommation
ANOVA for Regression Effects :
SOURCE      df           SS           MS            F             Prob
Regression   1       118.1325       118.1325       207.6322         0.0000
Residual    25        14.2238         0.5690
Total       26       132.3563

Variables in the equation
VARIABLE            b        s.e. b    Beta     t    prob. t
           Poids    0.00678   0.0005   0.9447 14.409 0.0000
(Intercept)         1.03535

Increase in squared R for this step = 0.892534
F = 207.6322 with D.F. 1 and 25 with Probability = 0.0000
-------------------------------------------------------------


Variables entered in step  2
 2 Cylindree
 4 Poids

Squared Multiple Correlation = 0.9277
Dependent variable = Consommation
ANOVA for Regression Effects :
SOURCE      df           SS           MS            F             Prob
Regression   2       122.7842        61.3921       153.9275         0.0000
Residual    24         9.5721         0.3988
Total       26       132.3563

Variables in the equation
VARIABLE            b        s.e. b    Beta     t    prob. t
       Cylindree    0.00131   0.0004   0.3686  3.415 0.0023
           Poids    0.00450   0.0008   0.6273  5.812 0.0000
(Intercept)         1.39228

Increase in squared R for this step = 0.035145
F =  11.6631 with D.F. 1 and 24 with Probability = 0.0023
-------------------------------------------------------------


Variables entered in step  3
```

```
 1 Prix
 2 Cylindree
 4 Poids

Squared Multiple Correlation = 0.9293
Dependent variable = Consommation
ANOVA for Regression Effects :
SOURCE      df           SS           MS            F             Prob
Regression   3       123.0016      41.0005      100.8059          0.0000
Residual    23         9.3547       0.4067
Total       26       132.3563

Variables in the equation
VARIABLE            b        s.e. b    Beta     t    prob. t
            Prix    0.00003   0.0000   0.1621  0.731 0.4721
       Cylindree    0.00108   0.0005   0.3038  2.163 0.0412
           Poids    0.00380   0.0012   0.5297  3.072 0.0054
(Intercept)         1.82417

Increase in squared R for this step = 0.001642
F =    0.5344 with D.F. 1 and 23 with Probability = 0.4721
-------------------------------------------------------------

Last variable added failed entry test. Job ended.
```

3.7 Simultaneous regression

The redundancy between the independent variables is a difficulty in the regression. We can detect redundant variables by computing the correlation. But sometimes, a variable is highly correlated with a combination of the other variables. To detect this last configuration, we can calculate the regression of each independent variable with the set of the other independent variables. If "p" is the number of the independent variables, we should perform "p" regressions. In fact, this is not necessary. **We can obtain all these regression from the correlation matrix**.

We click on the ANALYSES / MULTIPLE REGRESSION / SIMULTANEOUS menu. We select the previous independent variables for the regression. We click on the COMPUTE button.



We obtain the inverse of the correlation matrix. On the main diagonal, we have the VIF criterion which is provided by the regression coefficients table (section 3.2) [e.g. VIF (PRIX) = 19.792].

```
Simultaneous Multiple Regression by Bill Miller

Product-Moment Correlations Matrix with   27 cases.

Variables
                    Prix     Cylindree     Puissance         Poids
       Prix        1.000         0.918         0.927         0.947
  Cylindree        0.918         1.000         0.956         0.861
  Puissance        0.927         0.956         1.000         0.852
      Poids        0.947         0.861         0.852         1.000

Determinant of correlation matrix =   0.0011

Inverse of correlation matrix with   27 cases.

Variables
                    Prix     Cylindree     Puissance         Poids
       Prix       19.792        -1.452        -7.513       -11.085
  Cylindree       -1.452        12.869        -9.798        -1.358
```

```
 Puissance      -7.513       -9.798       14.892        2.861
     Poids     -11.085       -1.358        2.861       10.226

Multiple Correlation Coefficients for Each Variable

  Variable       R         R2          F      Prob.>F  DF1  DF2
      Prix     0.974      0.949    144.072      0.000    3   23
 Cylindree     0.960      0.922     90.995      0.000    3   23
 Puissance     0.966      0.933    106.507      0.000    3   23
     Poids     0.950      0.902     70.732      0.000    3   23

Betas in Columns with   27 cases.

Variables
                   Prix    Cylindree    Puissance        Poids
      Prix       -1.000        0.113        0.505        1.084
 Cylindree        0.073       -1.000        0.658        0.133
 Puissance        0.380        0.761       -1.000       -0.280
     Poids        0.560        0.106       -0.192       -1.000

Standard Errors of Prediction
Variable     Std.Error
      Prix  3011.761
 Cylindree   188.031
 Puissance     9.034
     Poids   104.468

Raw Regression Coefficients with   27 cases.

Variables
                   Prix    Cylindree    Puissance        Poids
      Prix       -1.000        0.006        0.001        0.027
 Cylindree        1.457       -1.000        0.034        0.066
 Puissance      145.906       14.732       -1.000       -2.681
     Poids       22.464        0.213       -0.020       -1.000

Variable   Constant
      Prix-12570.317
 Cylindree   235.992
 Puissance     3.698
     Poids   520.312

Partial Correlations with   27 cases.


Variables
```

	Prix	Cylindree	Puissance	Poids
Prix	-1.000	0.091	0.438	0.779
Cylindree	0.091	-1.000	0.708	0.118
Puissance	0.438	0.708	-1.000	-0.232
Poids	0.779	0.118	-0.232	-1.000

For instance, for PRIX, we have the following equation.

PRIX = 1.457 * CYLINDREE + 145.906 * PUISSANCE + 22.464 * POIDS – 12570.317

The coefficient of determination for this regression is R^2 = 0.949; the F statistic for the global significance is F = 144.072 (p-value = 0.000). Undoubtedly, PRICE is strongly correlated with at least one of the other independent variables.

4 Conclusion

LazStats and OpenStat are simple tools. But they are really powerful and rely on a rigorous scientific work. Many tutorials and books are available on the author's website. They are interesting to learn how to use these tools, but they are also for the understanding of statistical methods.