

# 1 Objectif

## Régression avec le logiciel LazStats (OpenStat).

**LazStats**<sup>1</sup> est un logiciel de statistique programmé et diffusé par Bill Miller, le père du logiciel **OpenStat**<sup>2</sup>, très connu des statisticiens depuis un certain nombre d'années. Ce sont des outils de très grande qualité, avec une rigueur de calcul appréciable. OpenStat fait partie des logiciels de statistique que je privilégie lorsque je souhaite valider mes propres implémentations.

Les outils développés par Bill Miller ont beaucoup évolué avec le temps. Lui seul connaît véritablement leur historique. Pour me part, en me basant sur ce qui était perceptible sur le site web de diffusion, je distinguerai plusieurs périodes. La première version d'OpenStat qui a retenu mon attention était développée en Delphi. Le code source était en ligne (il l'est encore<sup>3</sup>). Je l'ai énormément étudié. Pour moi, il s'agissait d'un contrepoint très intéressant pour Tanagra, également écrit avec Delphi, mais dont les structures étaient avant tout optimisées pour les techniques de Data Mining. Par la suite, son auteur a traduit son code en C++ en passant au compilateur C++ Builder. A un moment donné, la dénomination du logiciel a été modifiée, elle est devenue Stat4U. Puis finalement, l'auteur est revenu à OpenStat, toujours développé en C++.

Aujourd'hui, une version compilable avec Lazarus<sup>4</sup> (langage Pascal Objet, une sorte de Delphi version libre) est également en ligne. Le principal intérêt de **LazStats** est de bénéficier du principe « write once, compile anywhere » (écrire une fois, compiler partout<sup>5</sup>). L'idée est simple : nous écrivons le logiciel avec l'EDI de Lazarus sous un système (ex. Windows) ; le code source peut être compilé sur n'importe quelle plate-forme (ex. Windows, Linux, etc.), pourvu que Lazarus accompagné du compilateur « Free Pascal Compiler » y soient disponibles. Cette idée avait déjà été mise en avant par Borland avec Kylix<sup>6</sup> il y a une dizaine d'année. Un des prototypes de Tanagra, lors de sa genèse, a d'ailleurs été implémenté sous Kylix. Mais le manque de fiabilité patent du compilateur – les plantages étaient fréquentissimes, je n'arrivais plus à distinguer mes propres erreurs des caprices du compilateur – m'avait dissuadé de poursuivre dans cette voie. Je suis sagement revenu à Delphi 6 pour Windows que j'utilise encore aujourd'hui.

Le projet Lazarus semble plus mature que Kylix (qui a été abandonné d'ailleurs). Et le logiciel LazStats, qui est certainement une émanation de la première version en Delphi de OpenStat, est de très bonne facture si j'en juge sa stabilité face aux multiples tests que j'ai pu effectuer. J'ai choisi de présenter la version Windows parce que j'ai l'habitude de travailler sous cet environnement. Une version Linux est accessible sur le site de diffusion pour ceux qui le désirent. Il est également possible de télécharger des versions pour Mac OSX et Linux 64 bits.

L'autre véritable évolution ces dernières années est la mise à disposition d'une documentation de plus en plus riche sur le site web d'OpenStat. Un ouvrage décrit les méthodes statistiques, des

---

<sup>1</sup> <http://www.statpages.org/miller/openstat/LazStatsPage.htm>

<sup>2</sup> <http://www.statpages.org/miller/openstat/OpenStatPage.htm>

<sup>3</sup> <http://www.statpages.org/miller/openstat/LegacyPage.htm>

<sup>4</sup> <http://www.lazarus.freepascal.org/>

<sup>5</sup> <http://fr.wikipedia.org/wiki/Lazarus>

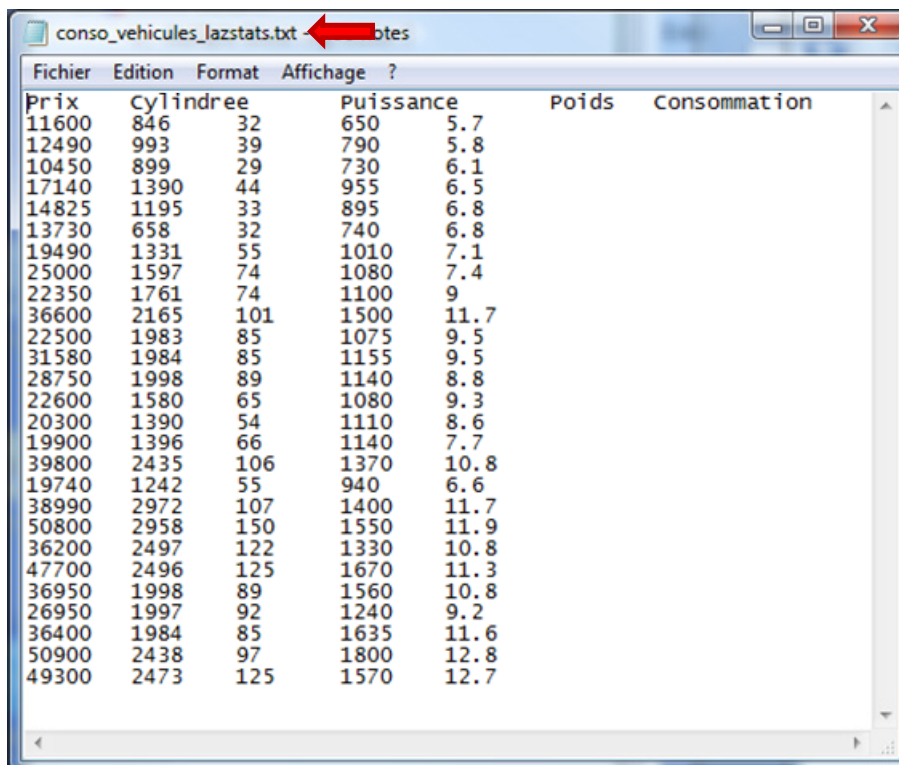
<sup>6</sup> [http://fr.wikipedia.org/wiki/Kylix\\_%28informatique%29](http://fr.wikipedia.org/wiki/Kylix_%28informatique%29)

tutoriels rédigés décrivent leur mise en œuvre et, pour enfoncer le clou, des tutoriels animés (fichiers `.wmv`) montrent les séquences de manipulations à réaliser pour mener les analyses. Le travail accompli est vraiment remarquable. Je m'y réfère souvent pour situer ce que je fais moi-même.

Enfin, ce tutoriel n'est pas complètement anodin. Dans un avenir plus ou moins lointain, lorsqu'il s'avèrera nécessaire de passer Tanagra en 64 bits, la solution Lazarus sera certainement la plus appropriée. Cerner les possibilités de cet outil de développement est une bonne manière d'évaluer l'ampleur de la tâche à venir.

## 2 Données

Nous utilisons le fichier « [conso\\_vehicules\\_lazstats.txt](#) » (format texte avec séparateur tabulation) pour décrire le logiciel. Il s'agit d'expliquer la consommation des automobiles à partir de leur prix, cylindrée, puissance et poids. Comme la très grande majorité des logiciels de statistique, LazStats sait lire ce type de format.

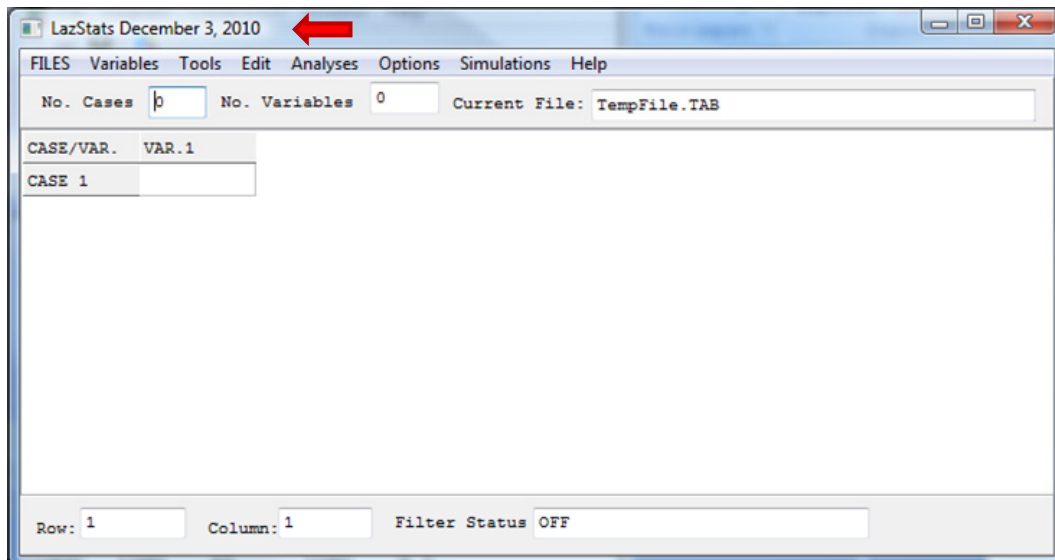


Prix	Cylindree	Puissance	Poids	Consommation
11600	846	32	650	5.7
12490	993	39	790	5.8
10450	899	29	730	6.1
17140	1390	44	955	6.5
14825	1195	33	895	6.8
13730	658	32	740	6.8
19490	1331	55	1010	7.1
25000	1597	74	1080	7.4
22350	1761	74	1100	9
36600	2165	101	1500	11.7
22500	1983	85	1075	9.5
31580	1984	85	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

Ce fichier présente une particularité très intéressante, nous l'avons étudié de manière détaillée dans le **chapitre 3** consacré à la colinéarité et la sélection de variables de notre support de cours accessible librement en ligne « **Pratique de la Régression Linéaire Multiple – Diagnostic et Sélection de Variables** » (<http://eric.univ-lyon2.fr/~ricco/publications.html>). Nous y ferons référence à plusieurs reprises pour comprendre les sorties de LazStats.

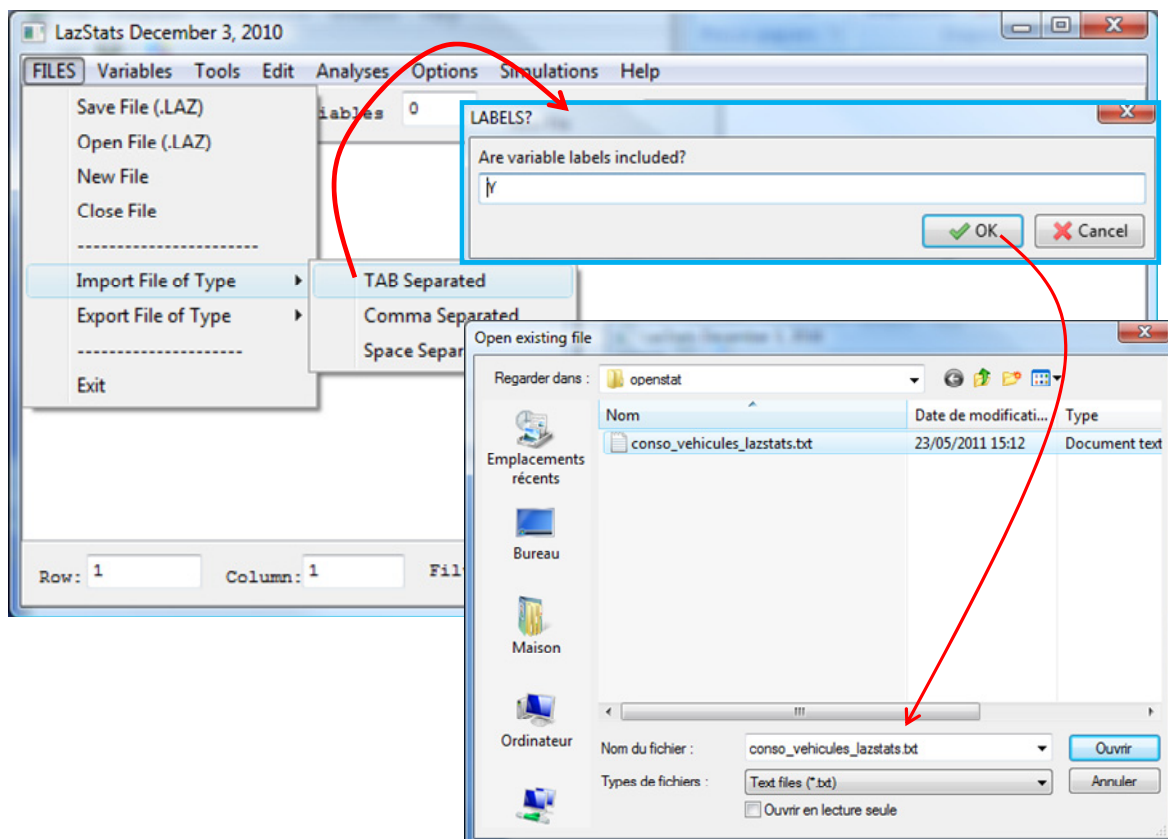
## 3 Régression linéaire avec LazStats

La première étape bien évidemment consiste à récupérer le SETUP de LazStats sur le site de Bill Miller (<http://statpages.org/miller/openstat/LazStatsPage.htm> - version du 3 décembre 2010 en ce qui nous concerne). Nous privilégions la version Windows dans ce tutoriel. Le logiciel s'installe très simplement. Nous le démarrons.



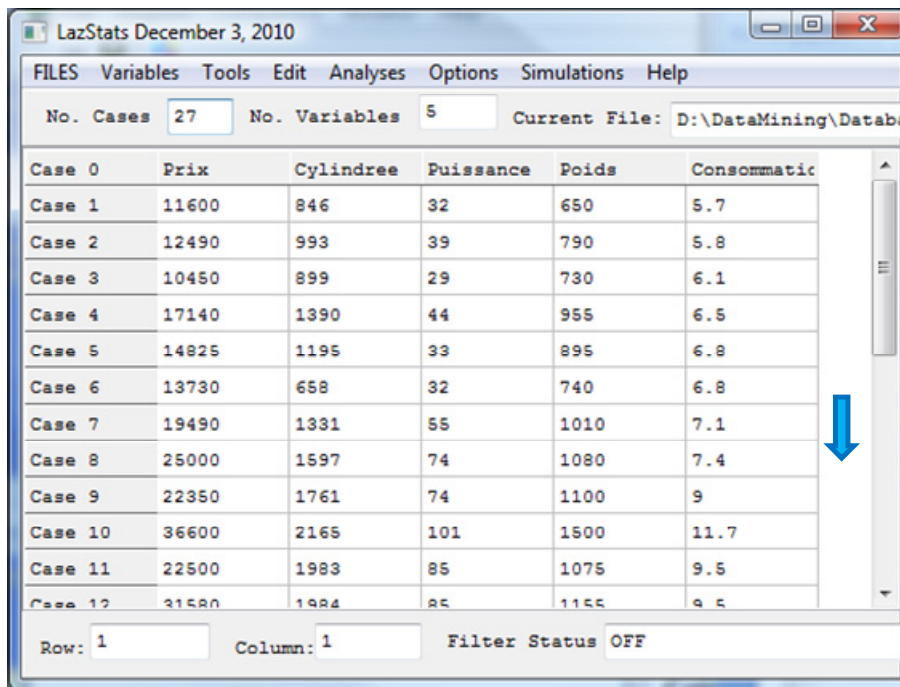
LazStats est entièrement piloté par menu. Une grille permet de visualiser les données. Les utilisateurs de SPSS, de STATISTICA,... ou de SIPINA, ne seront absolument pas dépaysés. Ainsi, une fois les données importées, les séquences de manipulations seront toujours les mêmes : choisir la technique statistique à mettre en œuvre ; définir les variables à traiter ; spécifier les éventuels paramètres de l'analyse ; lancer les calculs ; lire et interpréter les résultats. C'est ce que nous allons faire dans les sous-sections suivantes.

### 3.1 Importation des données

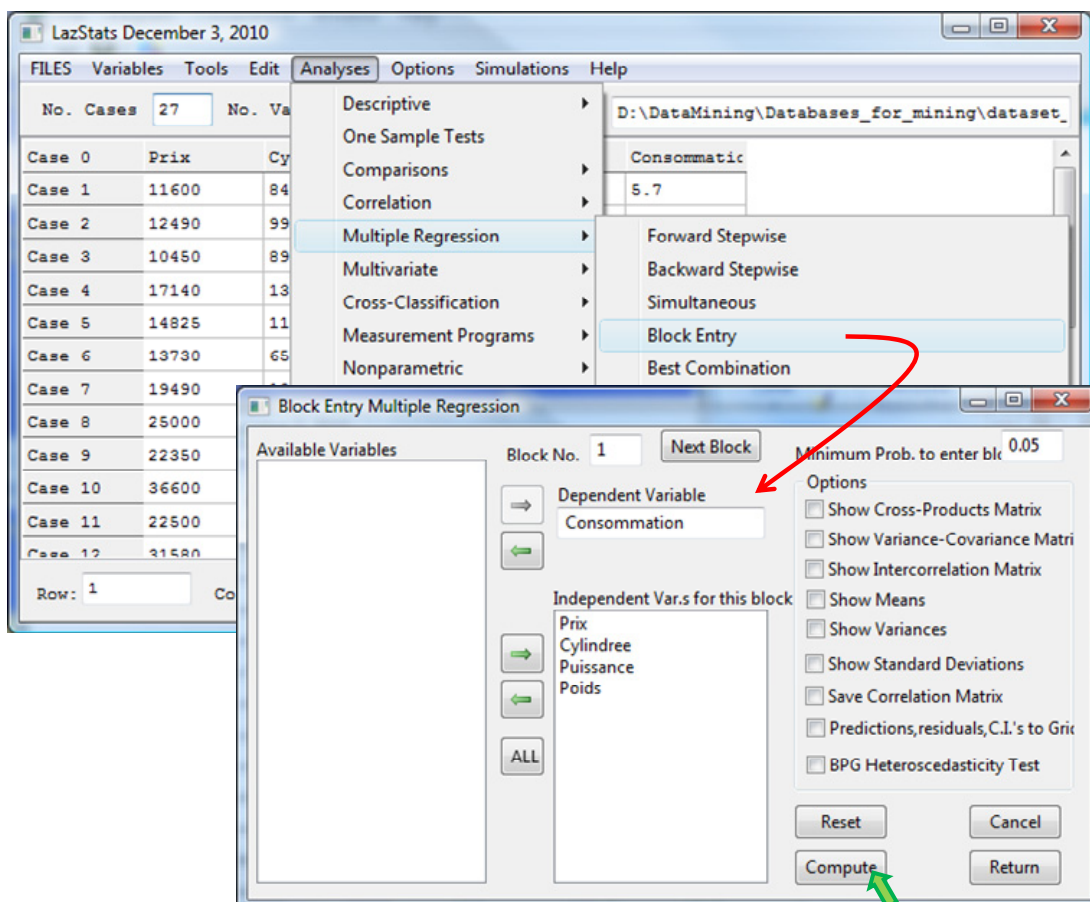


Avant toute chose, il nous faut importer les données. Pour ce faire, nous actionnons le menu FILES / IMPORT FILE OF TYPE / TAB SEPARATED pour charger le fichier au format texte avec le caractère

tabulation comme séparateur de colonnes. LazStats demande si la première ligne du fichier correspond aux noms des variables. Nous confirmons. Il ne nous reste plus qu'à désigner le fichier « conso\_vehicules\_lazstats.txt ». Les données sont affichées dans la grille.



### 3.2 Régression linéaire multiple



Nous souhaitons réaliser une régression avec la totalité des exogènes candidates. Nous actionnons le menu ANALYSES / MULTIPLE REGRESSION / BLOCK ENTRY. Dans la boîte de paramétrage, nous plaçons CONSOMMATION en DEPENDENT VARIABLE, les autres en INDEPENDENT VARS FOR THIS BLOCK (nous éclairerons le sens de cette précision dans la section suivante). Nous ne sélectionnons pas d'options.

Les résultats apparaissent dans une nouvelle fenêtre lorsque nous cliquons sur COMPUTE.

Block Entry Multiple Regression by Bill Miller

----- Trial Block 1 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			

ANOVA TABLE

Dependent Variable: Consommation

R	R2	F	Prob.>F	DF1	DF2
0.964	0.930	72.536	0.000	4	22

F - TEST

Adjusted R Squared = 0.917

Std. Error of Estimate = 0.651

REGRESSION COEFFICIENTS

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

Increase in R Squared = 0.930

F = 72.536 with probability = 0.000

Block 1 met entry requirements

Nous y voyons tour à tour<sup>7</sup> : le tableau d'analyse de variance ; le coefficient de corrélation multiple  $R = 0.964$  ; le coefficient de détermination  $R^2 = 0.930$  ; la statistique ( $F = 72.536$ ) du test de significativité globale de la régression ; avec sa probabilité critique ( $p$ -value = 0.000) et les degrés de liberté (4 ; 22) ; le coefficient de détermination ajusté 0.917 ; et l'estimation de l'écart type de l'erreur 0.651.

Vient ensuite la grille des coefficients. « Beta » est le coefficient standardisé, il rend comparable l'importance des variables dans la régression. « B » représente le coefficient de la droite de régression. « Std.error » est l'écart-type estimé des coefficients estimés. « t » est la statistique du test de significativité individuelle des coefficients ; avec sa probabilité critique « Prob.>t » (le test est bilatéral, c'est bien la valeur absolue de « t » qui est utilisée). « VIF » (variance inflation factor) est un indicateur permettant d'évaluer la colinéarité d'une variable avec l'ensemble des autres exogène, généralement, lorsque VIF est supérieur à 10, il faut s'inquiéter, ce qui est le cas pour toutes les variables de notre exemple. « TOL » est la tolérance, il est égal à  $(1/VIF)$ .

<sup>7</sup> D. Garson décrit en détail les résultats de SPSS, c'est aussi un élément de comparaison très intéressant pour la compréhension des sorties de LazStats -- <http://faculty.chass.ncsu.edu/garson/PA765/regress.htm>



Dans la partie basse de la fenêtre, le test de significativité globale est réitéré. Ici, on nous dit que le modèle composé des 4 variables est significatif. Le message qui accompagne le test est assez sibyllin « *Block 1 met entry requirements* », nous le comprendrons mieux dans la section suivante.

A titre de comparaison, nous donnons les résultats fournis par TANAGRA sur les mêmes données. Ils sont en tous points semblables bien évidemment.

**Global results**

Endogenous attribute	Consummation
Examples	27
R <sup>2</sup>	0.929520
Adjusted-R <sup>2</sup>	0.916706
Sigma error	0.651169
F-Test (4,22)	72.5365 (0.000000)

**Analysis of variance**

Source	xSS	d.f.	xMS	F	p-value
Regression	123.0278	4	30.7570	72.5365	0.0000
Residual	9.3285	22	0.4240		
Total	132.3563	26			

**Coefficients**

Attribute	Coef.	std	t(22)	p-value
Intercept	1.838006	0.793367	2.316716	0.030220
Prix	0.000034	0.000045	0.752738	0.459587
Cylindree	0.001208	0.000722	1.672661	0.108557
Puissance	-0.003742	0.015030	-0.248956	0.805704
Poids	0.003728	0.001300	2.868568	0.008926

**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

Backward Elimination Reg    Epsilon SVR    Nu SVR    Regression tree  
 C-RT Regression tree    Forward Entry Regression    Outlier Detection  
 DfBetas    Multiple linear regression    Regression Assessment

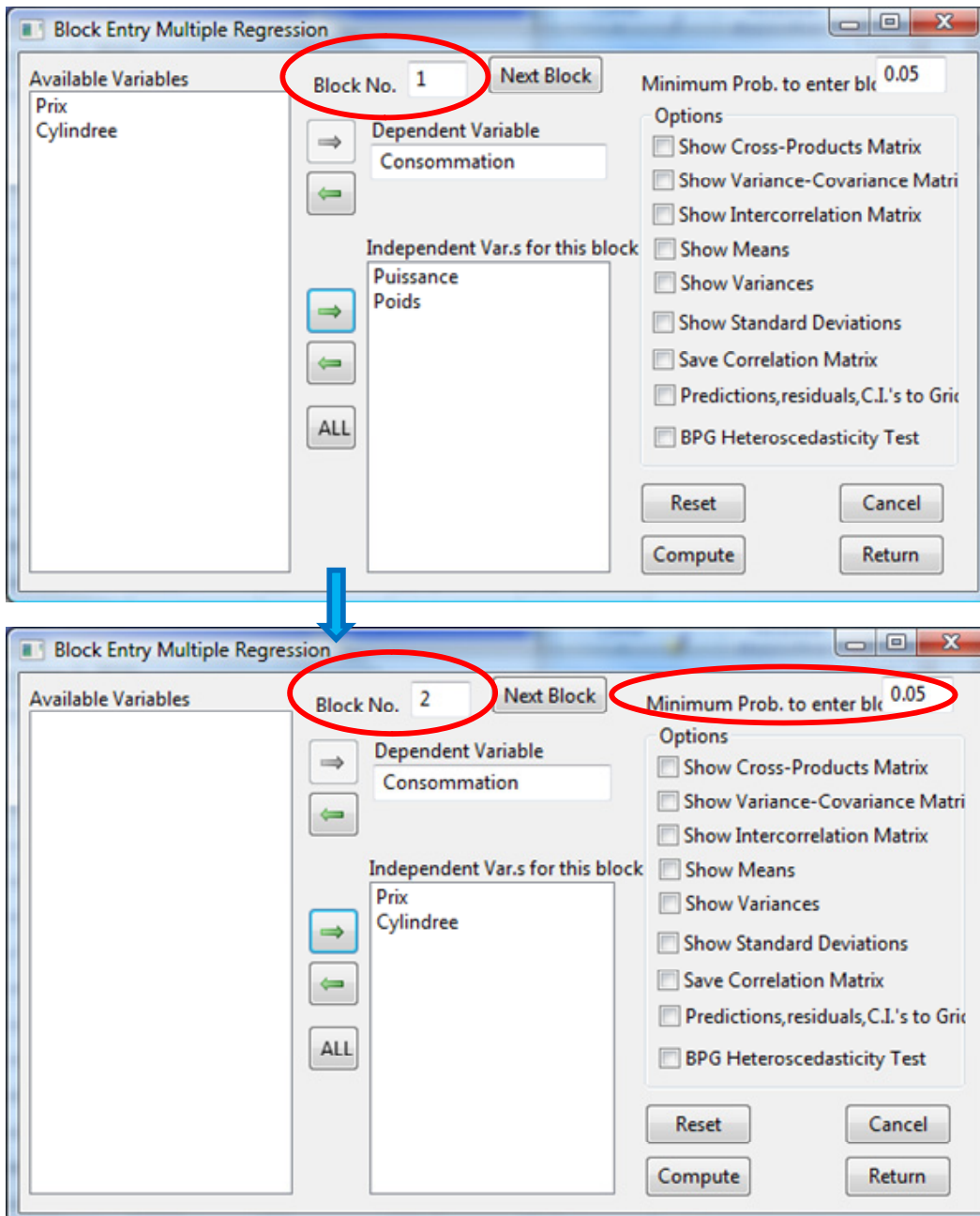
### 3.3 Régression par blocs

La « régression par blocs » vise à évaluer la pertinence d'un groupe additionnel de variables dans la régression. Les variables sont donc regroupées par blocs, spécifications que nous introduisons lors de la sélection des variables candidates.

Par sécurité, nous refermons les fenêtres en cliquant autant de fois que nécessaire sur les boutons RETURN. Puis, de nouveau, nous actionnons le menu ANALYSES / MULTIPLE REGRESSION / BLOCK ENTRY. Dans la boîte de paramétrage, nous plaçons CONSOMMATION en DEPENDENT

VARIABLE ; en INDEPENDENT VARS FOR THIS BLOKCK, nous insérons PUISSANCE et POIDS. Dans un premier temps, nous évaluerons la pertinence simultanée de ces deux variables.

Puis, nous cliquons sur le bouton **NEXT BLOCK**. Nous rajoutons en INDEPENDENT VARS FOR THIS BLOC, les variables PRIX et CYLINDREE. Dans un deuxième temps, **ce sera donc l'apport additionnel de ces deux variables, par rapport à PUISSANCE et POIDS, qui sera jugé**. Le risque du test est fixé à l'aide de l'option MINIMUM PROB. TO ENTER BLOCK, 0.05 (5%) en l'occurrence.



Nous cliquons sur COMPUTE. La fenêtre de résultats est plus fournie.

```
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
<b>Puissance</b>	0.304	0.021	0.008	2.724	0.012	3.648	0.274
<b>Poids</b>	0.686	0.005	0.001	6.137	0.000	3.648	0.274
Constant = 1.620							
Increase in R Squared = 0.918							
<b>F = 134.184 with probability = 0.000</b>							
<b>Block 1 met entry requirements</b>							
----- 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
<b>Puissance</b>	-0.054	-0.004	0.015	-0.249	0.806	14.892	0.067
<b>Poids</b>	0.519	0.004	0.001	2.869	0.009	10.226	0.098
<b>Prix</b>	0.190	0.000	0.000	0.753	0.460	19.792	0.051
<b>Cylindree</b>	0.340	0.001	0.001	1.673	0.109	12.869	0.078
Constant = 1.838							
<b>Increase in R Squared = 0.012</b>							
<b>F = 1.812 with probability = 0.187</b>							
<b>Block 2 did not meet entry requirements</b>							



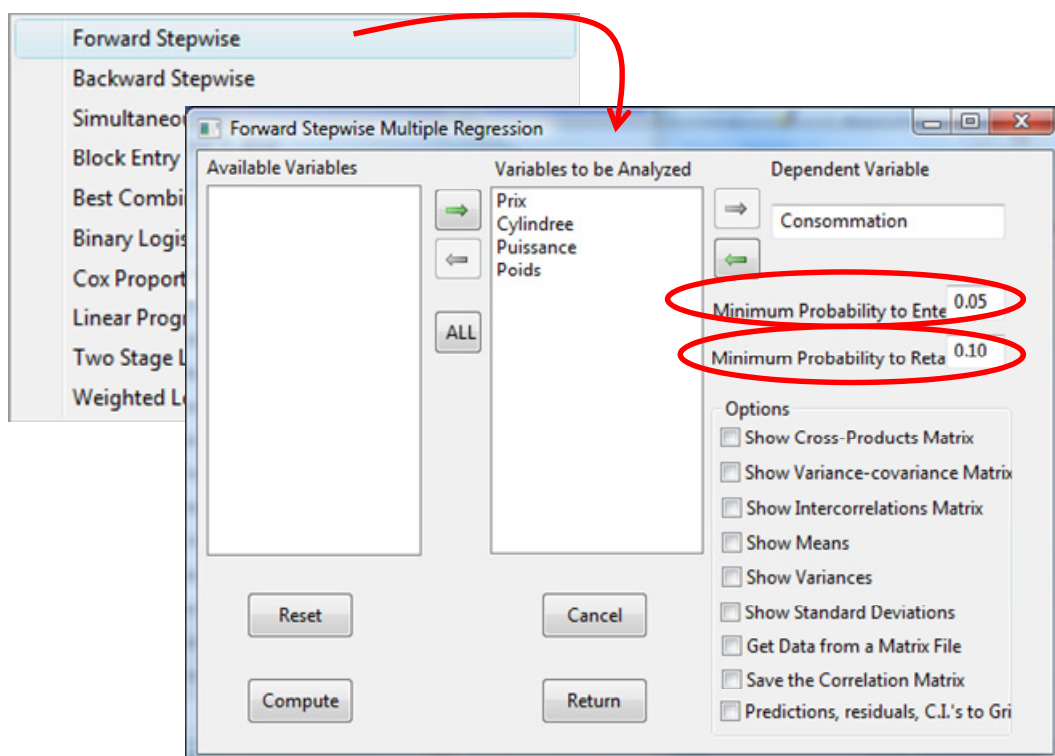
PUISSANCE et POIDS sont significatifs à 5%, avec  $F = 134.184$  et une probabilité critique 0.000.

A l'étape suivante, la régression avec 4 variables est globalement significative avec  $F = 72.536$  et  $p\text{-value} = 0.000$ . En revanche, lorsque l'on teste le pouvoir explicatif additionnel de PRIX et CYLINDREE par rapport au deux premières, on se rend compte qu'elles ne sont pas informatives, avec un  $F = 1.812$  et une probabilité critique 0.187. L'introduction de ces deux variables supplémentaires dans la régression n'est pas justifiée. Pour les férus d'économétrie, le calcul de la statistique F est basé sur l'écart entre les coefficients de détermination  $R^2$ , la formule est décrite dans le support « Econométrie – Régression linéaire simple et multiple » (section 10.4 -- <http://eric.univ-lyon2.fr/~ricco/publications.html>).

Même s'il paraît un peu compliqué au premier abord, ce dispositif est particulièrement utile lorsque nous voulons hiérarchiser l'introduction des exogènes dans un processus de sélection de variables. Par exemple, lorsque l'adjonction d'une variable particulière n'est souhaitable que si une autre variable est déjà présente dans le modèle pour être en accord avec les connaissances du domaine.

### 3.4 Régression stepwise – Sélection de variables

LazStats procède à une sélection de variables bi-directionnelle avec cette procédure. L'idée est la suivante : à l'étape  $t$ , il regarde s'il est possible d'ajouter une variable, si oui, il l'introduit, puis il teste si cette adjonction ne remet pas en cause une variable déjà présente dans le modèle. La procédure commence avec l'ajout d'une première variable. Deux niveaux de signification permettent de piloter la recherche: le premier intervient lors de l'ajout des variables (5%) ; le second lors de la tentative de retrait (10%). La première valeur est habituellement plus faible que la seconde, nous sommes plus exigeants lors de l'introduction que lors du retrait. Le processus s'arrête lorsque nous ne pouvons ni ajouter ni retirer de variables.



Nous activons le menu ANALYSES / MULTIPLE REGRESSION / FORWARD STEPWISE. Dans la boîte de paramétrage, nous sélectionnons toutes les variables, puis nous plaçons CONSOMMATION en DEPENDENT VARIABLE.

Nous cliquons sur COMPUTE.

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							
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	
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					
Prix	0.4567	6.3243	0.0190	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					
Cylindree	0.5719	11.6631	0.0023	1	24		
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					

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

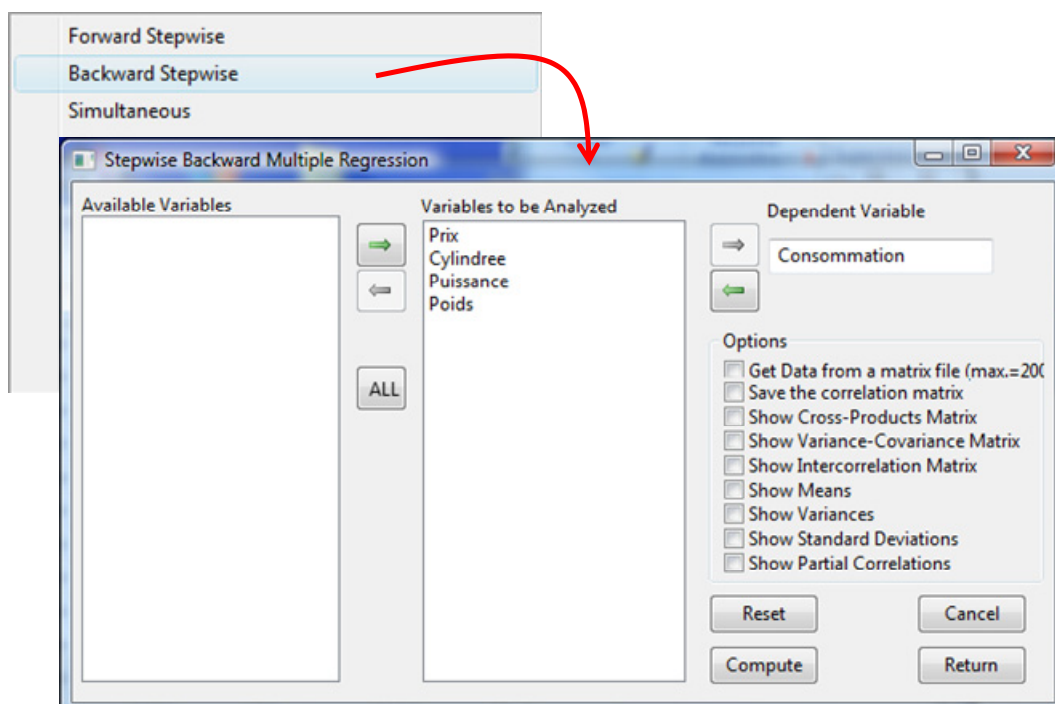
Variable	Beta	B	Std.Error	t	Prob.>t	VIF	TOL
Poids	0.627	0.005	0.001	5.812	0.000	3.867	0.259
Cylindree	0.369	0.001	0.000	3.415	0.002	3.867	0.259

Constant = 1.392

La première variable ajoutée est POIDS, puis CYLINDREE. Le processus s'arrête alors parce que la variable suivante, PUISSANCE, n'est pas pertinente avec : une corrélation partielle de 0.0188, un F partiel (statistique permettant de tester si la corrélation partielle est significative) de 0.0082, et une probabilité critique de p-value = 0.9288.

### 3.5 Régression backward – Sélection de variables

La sélection « backward » commence avec la totalité des variables candidates. A chaque étape, il retire la variable la moins significative, celle dont le t de Student est le plus petit *en valeur absolue*. Il s'arrête lorsqu'il n'y a plus qu'une variable. Nous fermons les fenêtres précédentes en cliquant autant de fois que nécessaire sur les boutons RETURN successifs. Arrivé à la fenêtre principale, nous actionnons le menu ANALYSES / MULTIPLE REGRESSION / BACKWARD STEPWISE.



Nous procédons de nouveau aux spécifications adéquates. Puis nous cliquons sur COMPUTE. Dans la fenêtre de résultats, il faut actionner plusieurs fois le bouton RETURN pour obtenir la succession de résultats.

```

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

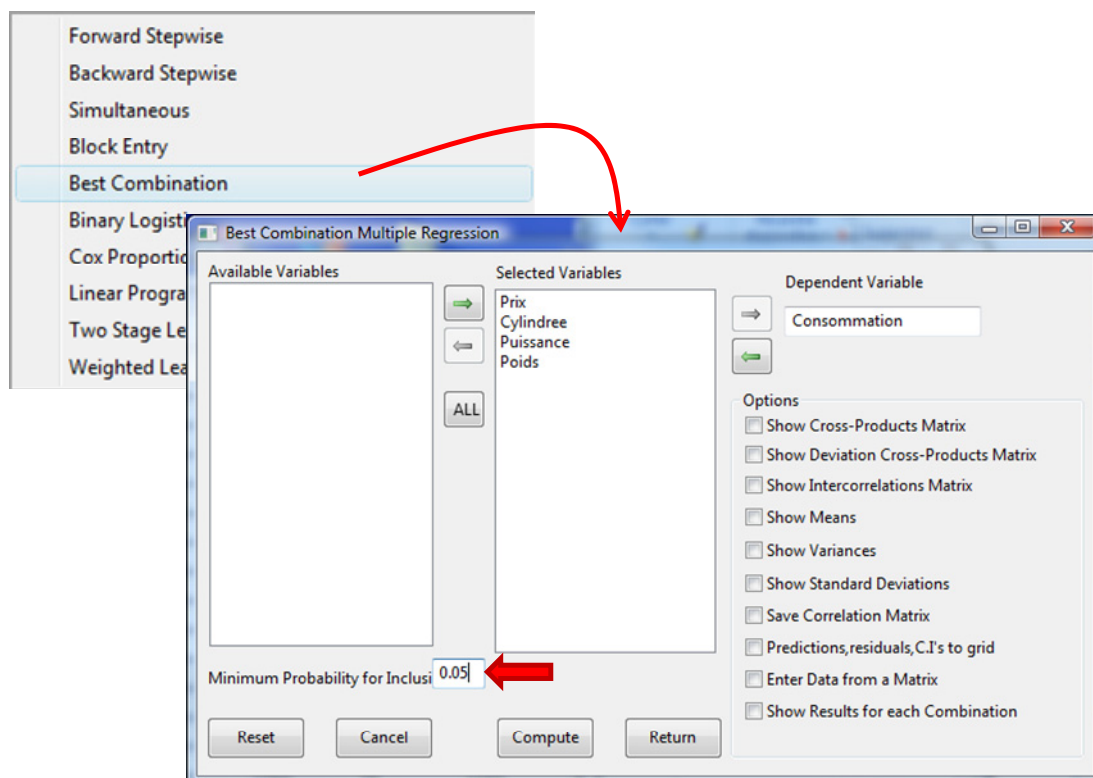


Std. Error of Estimate =		0.754					
Variable	Beta	B	Std. Error t		Prob.>t	VIF	TOL
Poids	0.945	0.007	0.000	14.409	0.000	1.000	1.000
Constant =		1.035					

### 3.6 Régression « best combination » – Sélection de variables

La sélection « best combination » correspond en réalité à une sélection pas à pas, par ajouts successifs. A chaque étape, la variable permettant le meilleur accroissement du coefficient de détermination  $R^2$  est mis en avant. Il est effectivement introduit dans le modèle si l'accroissement est significatif. Le processus est donc guidé par le seuil de significativité du test d'adjonction.

Nous actionnons le menu ANALYSES / MULTIPLE REGRESSION / BEST COMBINATION. Dans la boîte de dialogue, nous procédons aux spécifications idoines, nous n'oublions pas de spécifier le seuil des tests successifs : 0.05 pour 5%.



Nous cliquons sur COMPUTE. La fenêtre de résultats apparaît avec le détail de l'exploration des solutions.

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

-----

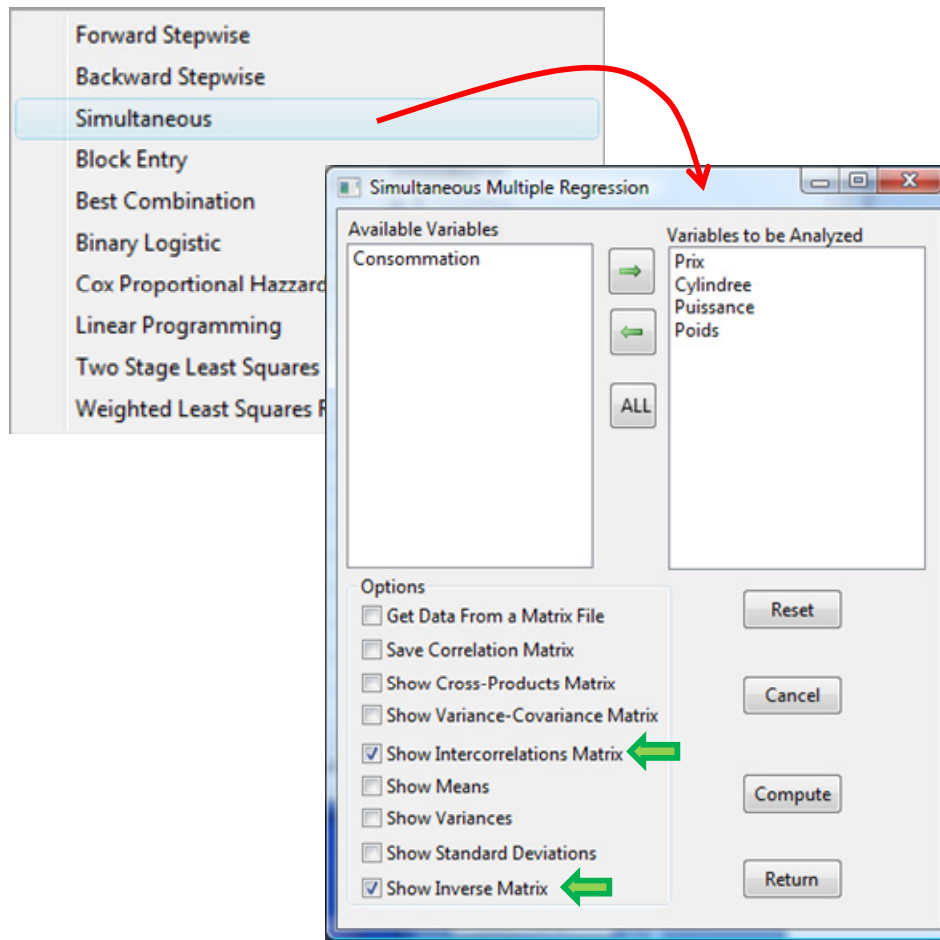
Last variable added failed entry test. Job ended.

### 3.7 Régressions croisées

Plus haut dans ce tutoriel, nous parlions du critère VIF. Il permet de rendre compte du degré de liaison de chaque exogène avec les autres explicatives du modèle. Dans son principe, il est basé sur la régression de chaque variable avec les autres. Dans les faits, il n'est pas nécessaire de former explicitement ces régressions, opérations coûteuses s'il en est, surtout lorsque nous avons à manipuler des grands fichiers. En réalité, le calcul est basé sur une inversion de la matrice de corrélations des exogènes. Nous obtenons sur la diagonale principale la valeur du VIF.

LazStats permet d'aller plus loin. Toujours en partant de la matrice des corrélations croisées, il sait reconstituer les régressions de chaque exogène avec les autres variables. Voyons ce qu'il en est.

Nous actionnons le menu ANALYSES / MULTIPLE REGRESSION / SIMULTANEOUS. Nous sélectionnons uniquement les exogènes de notre analyse, soit : PRIX, CYLINDREE, PUISSANCE et POIDS. Nous demandons à ce que la matrice de corrélation et son inverse soient affichés. Nous cliquons sur COMPUTE.



Le logiciel fournit dans un premier temps la matrice de corrélation et son inverse. Sur la diagonale principale de cette dernière, nous avons bien le critère VIF fourni par LazStats lors de la régression sur la totalité des variables (section 3.2) (ex. VIF de PRIX = 19.792 ; etc.).

```

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

Plus intéressant dans ce nouveau contexte, nous disposons des informations sur les régressions : le coefficient de détermination, la statistique de test de significativité globale, les coefficients de régression standardisés (betas), les coefficients de régression, et les corrélations partielles.

Pour clarifier les idées, voyons le cas de la variable prix. Le modèle de régression s'écrit :

$$\text{PRIX} = 1.457 * \text{CYLINDREE} + 145.906 * \text{PUISSANCE} + 22.464 * \text{POIDS} - 12570.317$$

Le coefficient de détermination de cette régression est égale à  $R^2 = 0.949$  ; la statistique de test de significativité globale est  $F = 144.072$ , avec une probabilité critique  $\text{Prob.}>F = 0.000$ . Manifestement, PRIX est fortement corrélée avec au moins une des variables explicatives candidates.

## 4 Conclusion

LazStats et OpenStat sont des outils très simples à manier. Ils bénéficient d'un travail de fond scientifique extrêmement rigoureux. Plusieurs fois, lorsque j'avais un doute sur mes propres implémentations, j'ai comparé mes résultats avec ceux de Bill Miller et, lorsque le doute persistait, j'allais directement voir le code source pour vérifier les ressemblances et les dissemblances.

Dans ce tutoriel, nous avons décrit leurs fonctionnalités en matière de régression linéaire multiple.