

1. Theme

The REGRESS tool included into the SIPINA package.

Few people know it. In fact, several tools are installed when we launch the SETUP file of SIPINA ([setup_stat_package.exe](#)) (Figure 1). This is the case of REGRESS which is intended to multiple linear regression.

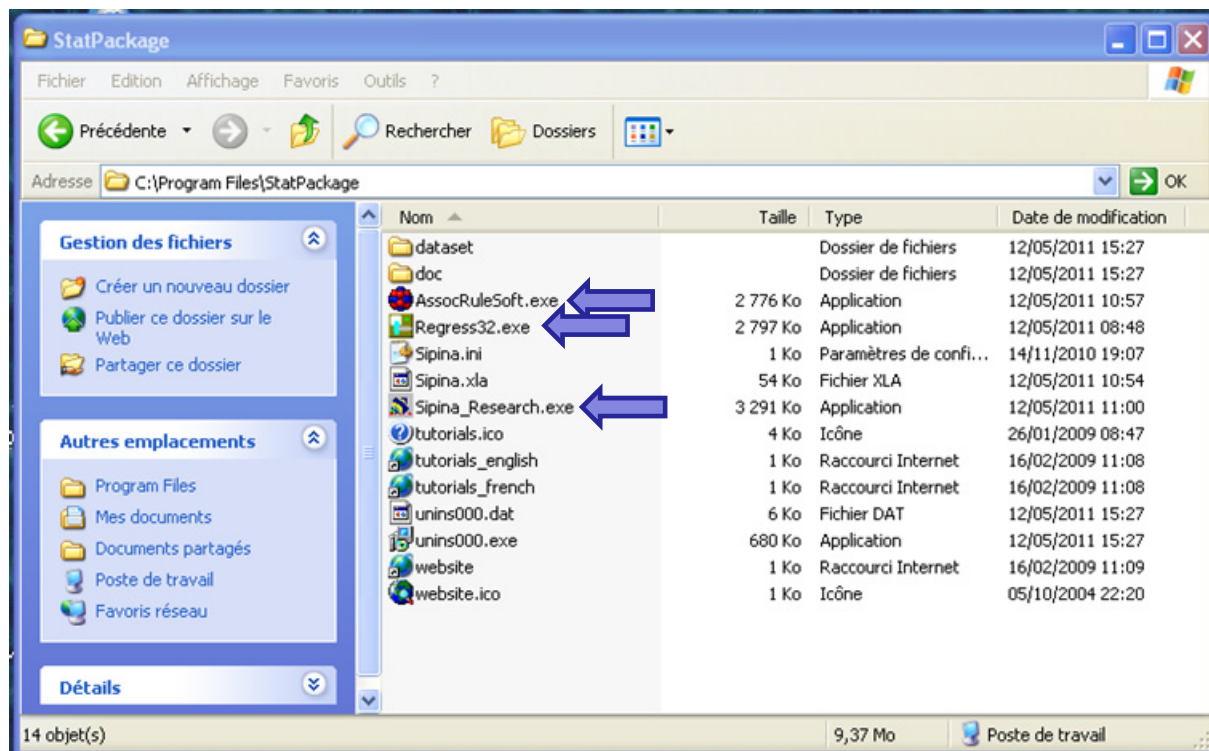


Figure 1 – Tools installed by the setup file of SIPINA

Even if a multiple linear regression procedure is incorporated to Tanagra, REGRESS can be useful essentially because it is very easy to use. The other reason to promote this tool is that it is directly in phase with the course about econometrics which I teach at the University (http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html).

2. Installing the SIPINA.XLA add-in for Excel

To install REGRESS, we must load the SIPINA Research version on the SIPINA website (from the version 3.7; may 2011), from the English URL

Data Mining Links
 French Data Mining Portal Tanagra Software Ricco's web site

SIPINA DOWNLOAD [French website](#)

Software	File
Most recent Sipina Research version - 32 bits. Implements several supervised learning methods (decision tree, neural network, linear discriminant analysis,...), model assessments (cross-validation, bootstrap,...) and association rule algorithm.	Sipina Research
Documentation	File
SIPINA Add-in for EXCEL spreadsheet (In english) An add-in for EXCEL(c) is incorporated in the SIPINA distribution. This add-in (SIPINA.XLA) enables to start a classification tree analysis, and more generally a data mining process, from your spreadsheet. This classification tree add-in appends a new menu in your spreadsheet. You select the cells range, activate the right menu: SIPINA is started and the selected dataset is automatically loaded.	1-Add-In Installation 2-How to use
Building decision tree interactively for the analysis of high blood pressure with SIPINA.	Tutorial
Using predefined learning (training) and test set for classifier performance evaluation. Definition of misclassification costs and the utilization of cost-sensitive decision tree classifier. Example in the classification of unsolicited e-mails (spams).	Tutorial
Computation of descriptive statistics on nodes during the interactive construction of the classification tree. Each node corresponds to a subpopulation, obtaining description of this subpopulation enables to better understand the significance of the rule. Both univariate and bivariate statistics are available.	Tutorial
Website about Tutorials for Sipina .	Website
Comparison of SIPINA with ORANGE -- Interactive construction of decision trees.	Tutorial
Comparison of SIPINA with TANAGRA and WEKA -- Training a neural network.	Tutorial
Other packages	File
XL-SIPINA is an attempt to embed the EXCEL(c) spreadsheet in a data mining software. It is mainly based on the Windows OLE technology. The ideas implemented here will be the starting point of wider project on association of a data mining software project and a free spreadsheet.	XL-Sipina French doc English doc
Old version of SIPINA (SIPINA v2.5) - 16 bits running under Windows 3.1. Implements	Setup Sipina v2.5 Documentation

Version française

Or from the French URL.

SIPINA - Un logiciel gratuit pour l'Induction des Arbres de Décision

Présentation
 Blog : Sipina - Arbres de décision
 Catégorie : data mining arbres de décision sipina Hi Tech
 Description : Sipina : fonctionnalités et références
 Partager ce blog
 Retour à la page d'accueil

Liens
[Sipina website en anglais](#)
[Télécharger Sipina](#)
[Blog des tutoriels](#)
[Le logiciel Tanagra](#)

Anciennes versions
[Sipina version 2.5](#)
[XL-Sipina](#)

Sipina
 SIPINA est un logiciel gratuit de Data Mining spécialisé dans l'induction des arbres de décision. Curieusement, c'est un des très rares outils en libre accès intégrant des fonctionnalités interactives lors de la construction d'un arbre de décision. Fonctionnalités qui, pourtant, font tout le sel de cette méthode dans une activité de fouille de données.

SIPINA implémente également d'autres méthodes supervisées. Mais son intérêt est moindre dans ce contexte. Depuis le développement et la diffusion de TANAGRA (Janvier 2004), je conseille systématiquement d'utiliser ce dernier. Il comporte non seulement les méthodes supervisées mais également une grande majorité des techniques de statistique et d'analyse de données telles que les analyses factorielles, la classification automatique, etc., et la possibilité de les faire coopérer entre elles.

Les différentes versions de SIPINA sont disponibles sur le web depuis 1995. La version actuelle n'a guère évolué depuis 2000. Elle est néanmoins distribuée car, comme je le disais plus haut, il y a très peu d'équivalents gratuits au monde. Le site de distribution en anglais est régulièrement consulté encore à ce jour, et le logiciel téléchargé. Il doit bien y avoir une raison à cela. J'ai donc décidé de la documenter un peu plus, aspect totalement négligé à l'époque de son développement. Je redécouvre d'ailleurs ainsi de très nombreuses fonctionnalités imaginées, expérimentées, et finalement connues de moi seul... autant que tout le monde en profite.

Configuré judicieusement, SIPINA peut traiter de très gros volumes (plusieurs millions d'observations - voir Sipina - Traitement des très grands fichiers) tout en conservant ses fonctionnalités interactives.

Ce site rassemble tout le matériel concernant SIPINA. Autre évolution notable, il est entièrement en français, le site initial ayant toujours été exclusivement en anglais. Le logiciel reste en anglais, mais les mots clés sont relativement simples à appréhender.

SIPINA est totalement gratuit, quel que soit le contexte d'utilisation.

Ricco Rakotomalala.

Documentation
 Fonctionnalités (6)
 Algos et méthodes (12)
 Doc. et tutoriels (17)
 Références en ligne (4)
 Bibliographie (6)
 Ils en parlent (2)

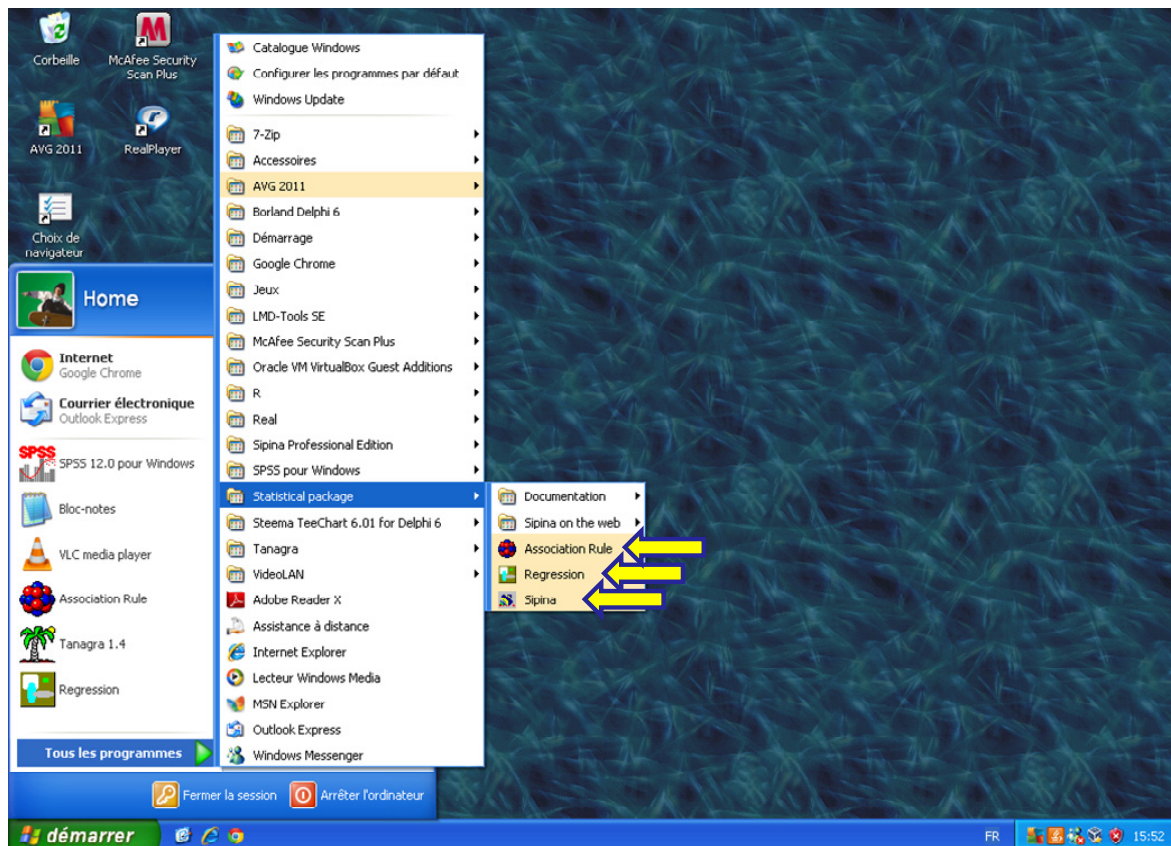
Principales doc.
 Intégrer Sipina dans Excel
 Prise en main de Sipina
 Apprentissage et test
 Méthodologie des arbres

Recherche

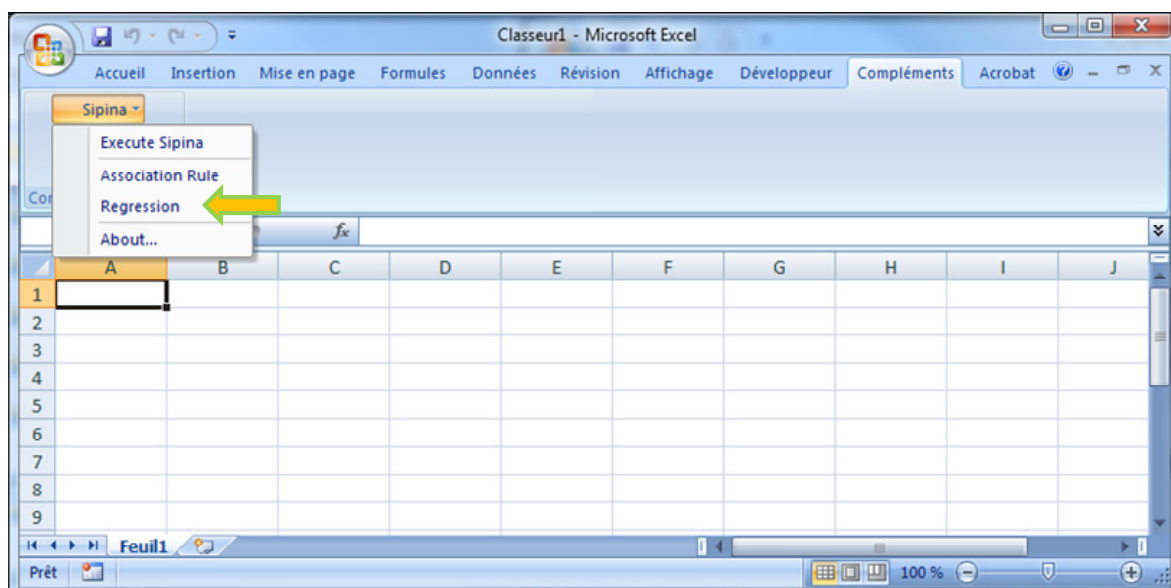
Articles récents
 Dessin "smart" de l'arbre dans la version 3.6
 Multithreading pour les arbres de décision
 Déploiement de modèles avec PMML
 Add-in Sipina pour Excel 2007 et 2010
 Sipina - Présentation de l'ancienne version 2.5
 Arbres de décision interactifs avec SPAD

Once we have loaded the setup file "[setup_stat_package.exe](#)", we launch this one. The tools are installed on our computer. We observe the "Statistical Package" folder into the START menu of Windows. Three tools are available: SIPINA which is intended to induction of

decision tree (supervised learning); a tool intended to the association rules construction; and REGRESS which is intended to the multiple linear regression.



We can now incorporate the **SIPINA.XLA** add-in into Excel spreadsheet application. I show how to do in previous tutorials: <http://data-mining-tutorials.blogspot.com/2010/08/sipina-add-in-for-excel.html> for up to Excel 2003; for Excel 2007 and 2010 - <http://data-mining-tutorials.blogspot.com/2010/08/tanagra-add-in-for-office-2007-and.html> (the description is related to Tanagra, but the adaptation to Sipina is easy).



Then, we launch Excel. Into the tab "**Compléments**" (in French version, probably "Add-ins" in English version), a menu SIPINA appears. When we activate the menu, we see three items, one for each tool associated to the Sipina package. All of them benefit to the same data management functionality: we can load the data file into the spreadsheet application; we can modify the values as we want; and last, we can send the dataset to the data mining tool.

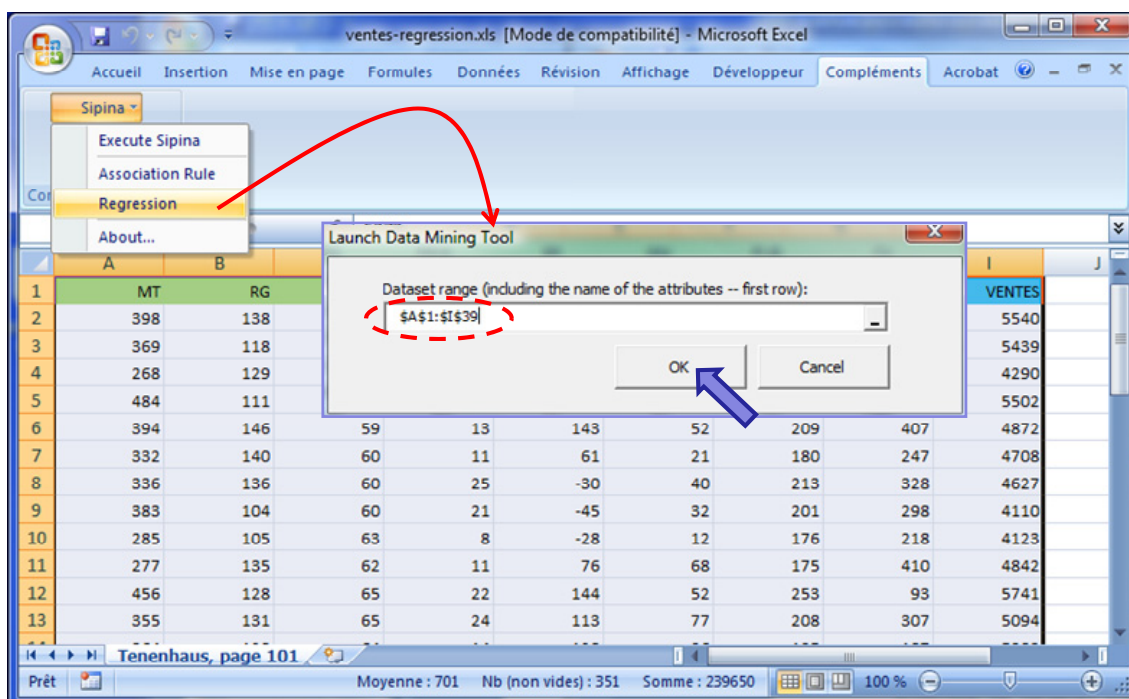
3. Dataset

To describe the use of REGRESS, we use the [ventes-regression.xls](#) data file, from Tenenhaus' book (page 101). We want to explain the volume of sales (VENTES) from a set of explanatory variables: MT (volume of sales into the branch), RG (put back to the wholesalers), PRIX (price), BR (budget of research), INV (investment), PUB (publicity), FV (cost of sale), TPUB (publicity for the branch). There are 38 instances into the data file.

Beyond the description of the actions to be made, our idea also is to check the results by comparing with those of our reference book. The author used the SPSS software.

4. Importing the dataset and launching REGRESS

We load the dataset into the Excel spreadsheet. After we select the data range, we click on the COMPLEMENTS (ADD-INS in English) / SIPINA / REGRESSION menu. A dialog box appears. We check the coordinates of the selection. Then, we click on OK.

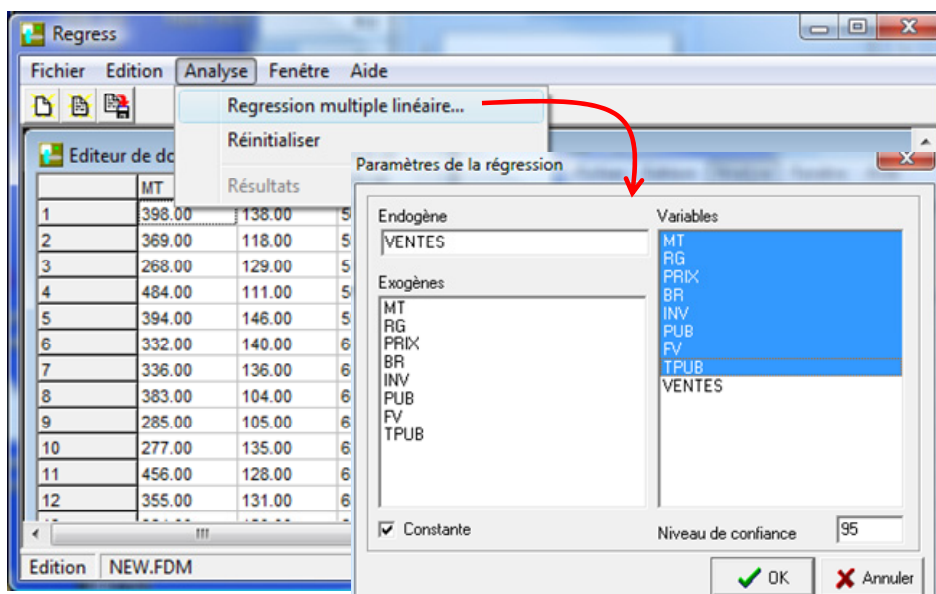


REGRESS is automatically launched. The dataset is loaded.

	MT	RG	PRIX	BR	INV	PUB	FV	TPUB
1	398.00	138.00	56.00	12.00	50.00	77.00	229.00	98.00
2	369.00	118.00	59.00	9.00	17.00	89.00	177.00	225.00
3	268.00	129.00	57.00	29.00	89.00	51.00	166.00	263.00
4	484.00	111.00	58.00	13.00	107.00	40.00	258.00	321.00
5	394.00	146.00	59.00	13.00	143.00	52.00	209.00	407.00
6	332.00	140.00	60.00	11.00	61.00	21.00	180.00	247.00
7	336.00	136.00	60.00	25.00	-30.00	40.00	213.00	328.00
8	383.00	104.00	60.00	21.00	-45.00	32.00	201.00	298.00
9	285.00	105.00	63.00	8.00	-28.00	12.00	176.00	218.00
10	277.00	135.00	62.00	11.00	76.00	68.00	175.00	410.00
11	456.00	128.00	65.00	22.00	144.00	52.00	253.00	93.00
12	355.00	131.00	65.00	24.00	113.00	77.00	208.00	307.00

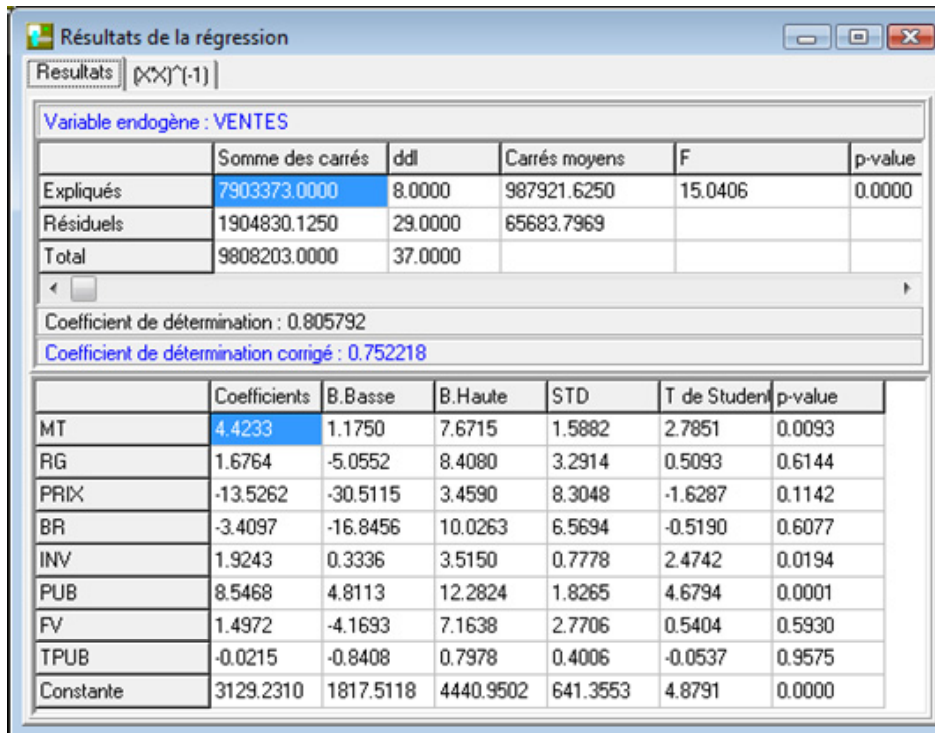
5. Multiple linear regression

VENTES is the dependent variable; the others are the independent ones. To define that, we click on ANALYSE / REGRESSION LINEAIRE MULTIPLE menu. Into the dialog box which appears, we set the appropriate selection by drag-and-drop.



By default, REGRESS performs a regression with constant (*Constante*). The confidence level for the calculation of the confidence intervals is 95% (*Niveau de confiance*). We validate our choices by clicking on OK. Several windows appear.

Résultats de la regression (Results of the regression). It provides the ANOVA table, the coefficient on determination $R^2 = 0.805792$, the adjusted $R^2 = 0.752218$, and the coefficients table (see <http://faculty.chass.ncsu.edu/garson/PA765/regress.htm> for the reading of these concepts). In this table, we observe: the estimated coefficients, their lower and upper limits of the confidence interval, their standard error, the t test, and the related p-value.



Résultats de la régression
 Résultats | $(X)^{-1}$

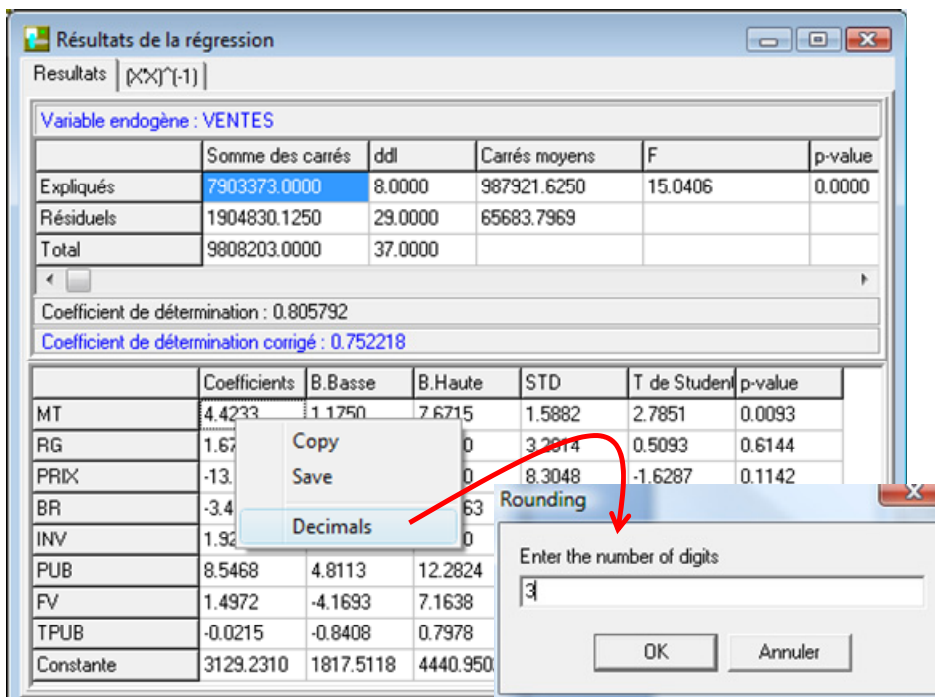
Variable endogène : VENTES

	Somme des carrés	ddl	Carrés moyens	F	p-value
Expliqués	7903373.0000	8.0000	987921.6250	15.0406	0.0000
Résiduels	1904830.1250	29.0000	65683.7969		
Total	9808203.0000	37.0000			

Coefficient de détermination : 0.805792
 Coefficient de détermination corrigé : 0.752218

	Coefficients	B.Basse	B.Haute	STD	T de Student	p-value
MT	4.4233	1.1750	7.6715	1.5882	2.7851	0.0093
RG	1.6764	-5.0552	8.4080	3.2914	0.5093	0.6144
PRIX	-13.5262	-30.5115	3.4590	8.3048	-1.6287	0.1142
BR	-3.4097	-16.8456	10.0263	6.5694	-0.5190	0.6077
INV	1.9243	0.3336	3.5150	0.7778	2.4742	0.0194
PUB	8.5468	4.8113	12.2824	1.8265	4.6794	0.0001
FV	1.4972	-4.1693	7.1638	2.7706	0.5404	0.5930
TPUB	-0.0215	-0.8408	0.7978	0.4006	-0.0537	0.9575
Constante	3129.2310	1817.5118	4440.9502	641.3553	4.8791	0.0000

We can modify the number of decimals for the display. We use the contextual menu and we click on the DECIMALS item.



Résultats de la régression
 Résultats | $(X)^{-1}$

Variable endogène : VENTES

	Somme des carrés	ddl	Carrés moyens	F	p-value
Expliqués	7903373.0000	8.0000	987921.6250	15.0406	0.0000
Résiduels	1904830.1250	29.0000	65683.7969		
Total	9808203.0000	37.0000			

Coefficient de détermination : 0.805792
 Coefficient de détermination corrigé : 0.752218

	Coefficients	B.Basse	B.Haute	STD	T de Student	p-value
MT	4.4233	1.1750	7.6715	1.5882	2.7851	0.0093
RG	1.6764	-5.0552	8.4080	3.2914	0.5093	0.6144
PRIX	-13.5262	-30.5115	3.4590	8.3048	-1.6287	0.1142
BR	-3.4097	-16.8456	10.0263	6.5694	-0.5190	0.6077
INV	1.9243	0.3336	3.5150	0.7778	2.4742	0.0194
PUB	8.5468	4.8113	12.2824	1.8265	4.6794	0.0001
FV	1.4972	-4.1693	7.1638	2.7706	0.5404	0.5930
TPUB	-0.0215	-0.8408	0.7978	0.4006	-0.0537	0.9575
Constante	3129.2310	1817.5118	4440.9502	641.3553	4.8791	0.0000

Context menu: Copy, Save, Decimals

Rounding dialog: Enter the number of digits: 3

Note that we can also copy the values to a spreadsheet application for subsequent calculations (COPY contextual menu).

	Somme des carrés	ddl	Carrés moyens	F	p-value
Expliqués	7903373.0000	8.0000	987921.6250	15.0406	0.0000
Résiduels	1904830.1250	29.0000	65683.7969		
Total	9808203.0000	37.0000			

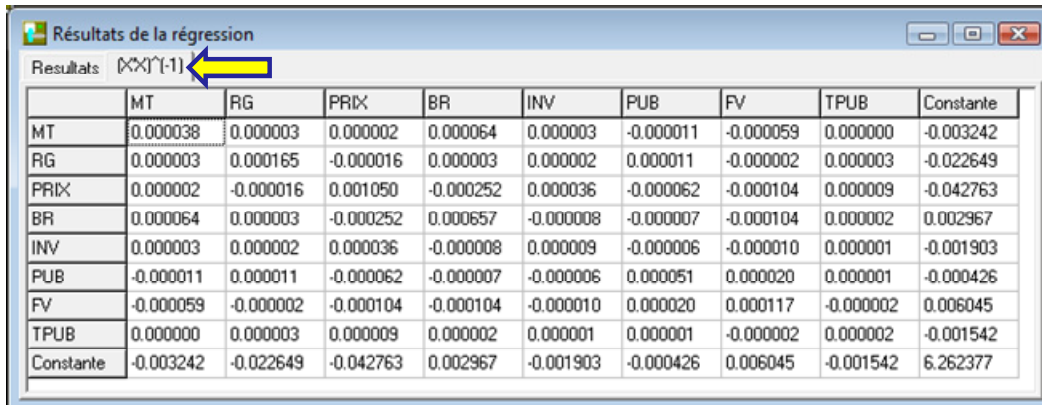
Coefficient de détermination : 0.805792
Coefficient de détermination corrigé : 0.752218

	Basse	B.Haute	STD	T de Student	p-value	
MT	4.42327261	1.17499816	7.672	1.588	2.785	0.009
RG	1.67640471	-5.05516768	8.408	3.291	0.509	0.614
PRIX	-13.5262318	-30.5115032	3.459	8.305	-1.629	0.114
BR	-3.40965939	-16.8456173	10.026	6.569	-0.519	0.608
INV	1.924	0.334	3.515	0.778	2.474	0.019
PUB	8.547	4.811	12.282	1.826	4.679	0.000
FV	1.497	-4.169	7.164	2.771	0.540	0.593
TPUB	-0.022	-0.841	0.798	0.401	-0.054	0.958
Constante	3129.231	1817.512	4440.950	641.355	4.879	0.000

The copy is performed with a maximum number of decimals for accurate calculations.

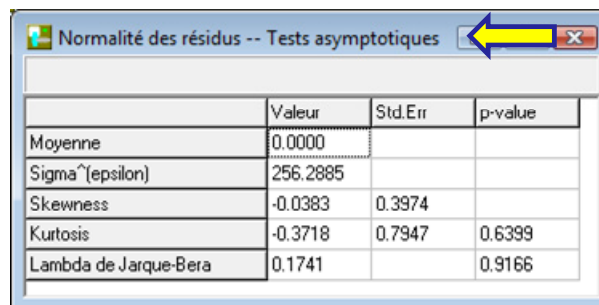
Label	Coefficients	B.Basse	B.Haute	STD	T de Student	p-value
MT	4.42327261	1.17499816	7.67154741	1.58821976	2.78505063	0.00933058
RG	1.67640471	-5.05516768	8.4079771	3.29135251	0.50933611	0.6143707
PRIX	-13.5262318	-30.5115032	3.4590404	8.30482292	-1.62872005	0.11418909
BR	-3.40965939	-16.8456173	10.0262985	6.56941223	-0.5190205	0.60768503
INV	1.92431557	0.33362997	3.51500106	0.77775395	2.47419572	0.01944845
PUB	8.54684162	4.81125164	12.2824316	1.82648897	4.6793828	6.18E-05
FV	1.49723995	-4.16929007	7.16376972	2.77060795	0.54040122	0.59304684
TPUB	-0.02151707	-0.84081632	0.79778218	0.4005903	-0.0537134	0.95753181
Constante	3129.23096	1817.51184	4440.9502	641.355347	4.87909079	3.5E-05

Covariance matrix between independent variables. Into the $(X'X)^{-1}$ tab, REGRESS provides the uncentered covariance matrix between the independent variables. This matrix is very useful for various calculations e.g. the test of significance of a set of parameters, the comparison between parameters, the computation of the leverage of one instance, the confidence interval for a prediction, etc.



	MT	RG	PRIX	BR	INV	PUB	FV	TPUB	Constante
MT	0.000038	0.000003	0.000002	0.000064	0.000003	-0.000011	-0.000059	0.000000	-0.003242
RG	0.000003	0.000165	-0.000016	0.000003	0.000002	0.000011	-0.000002	0.000003	-0.022649
PRIX	0.000002	-0.000016	0.001050	-0.000252	0.000036	-0.000062	-0.000104	0.000009	-0.042763
BR	0.000064	0.000003	-0.000252	0.000657	-0.000008	-0.000007	-0.000104	0.000002	0.002967
INV	0.000003	0.000002	0.000036	-0.000008	0.000009	-0.000006	-0.000010	0.000001	-0.001903
PUB	-0.000011	0.000011	-0.000062	-0.000007	-0.000006	0.000051	0.000020	0.000001	-0.000426
FV	-0.000059	-0.000002	-0.000104	-0.000104	-0.000010	0.000020	0.000117	-0.000002	0.006045
TPUB	0.000000	0.000003	0.000009	0.000002	0.000001	0.000001	-0.000002	0.000002	-0.001542
Constante	-0.003242	-0.022649	-0.042763	0.002967	-0.001903	-0.000426	0.006045	-0.001542	6.262377

Normalité des résidus (Tests for Normality of Residuals). Into the « Normalité des résidus – Tests asymptotiques » window, we have some information about the residuals. The mean (*Moyenne*) of the residuals is always equal to 0 for regression with intercept. The standard error of estimate [$\text{Sigma}^{\text{(epsilon)}}$] is 256.2885.

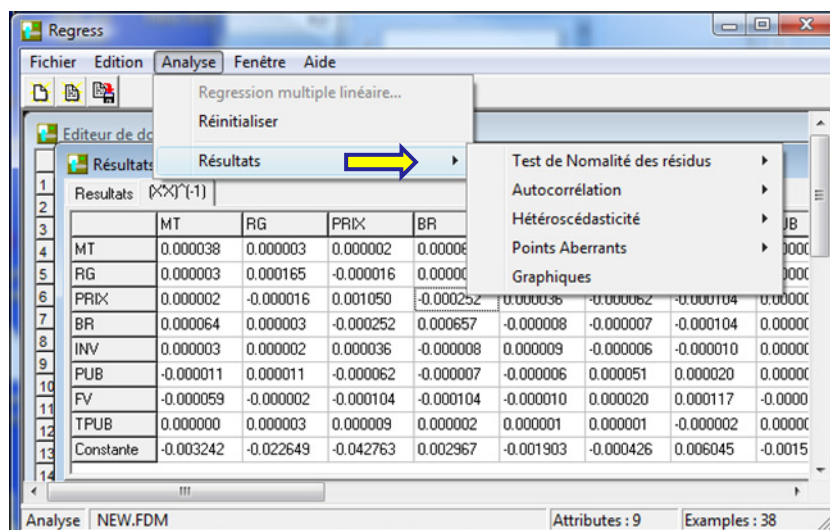


	Valeur	Std.Err	p-value
Moyenne	0.0000		
Sigma ^(epsilon)	256.2885		
Skewness	-0.0383	0.3974	
Kurtosis	-0.3718	0.7947	0.6399
Lambda de Jarque-Bera	0.1741		0.9166

Skewness is a measure of asymmetry of the residuals distribution. **Kurtosis** is a measure of its “peakedness”. **Lambda of Jarque Bera** enables to measure the departure from normality of the residuals probability distribution.

6. Additional results

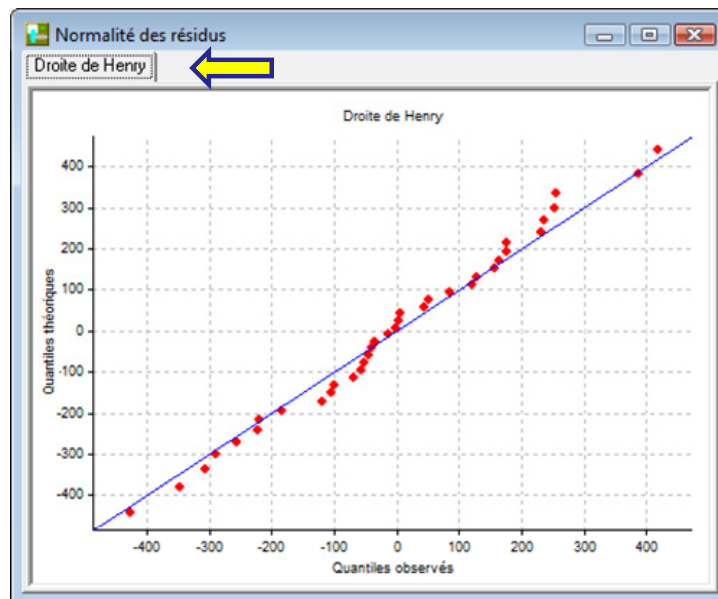
Beyond the windows generated automatically by REGRESS, we can obtain other results by activating the appropriate menus. To do that, we activate the ANALYSE / RESULTATS menu.



The screenshot shows the 'Regress' window with the 'Analyse' menu open. The 'Résultats' option is highlighted, and a yellow arrow points to it. The background shows the same regression results table as in the previous image.

We put aside the procedures dedicated to the detection and the treatment of the heteroscedasticity and the autocorrelation of residuals. They will be presented in a next tutorial.

Droite de Henry (Normal probability plot). By clicking on the ANALYSE / RESULTATS / TEST DE NORMALITE DES RESIDUS / DROITE DE HENRY menu, we obtain the [normal probability plot](#). If the points form an approximate straight line, we cannot reject the normality hypothesis.

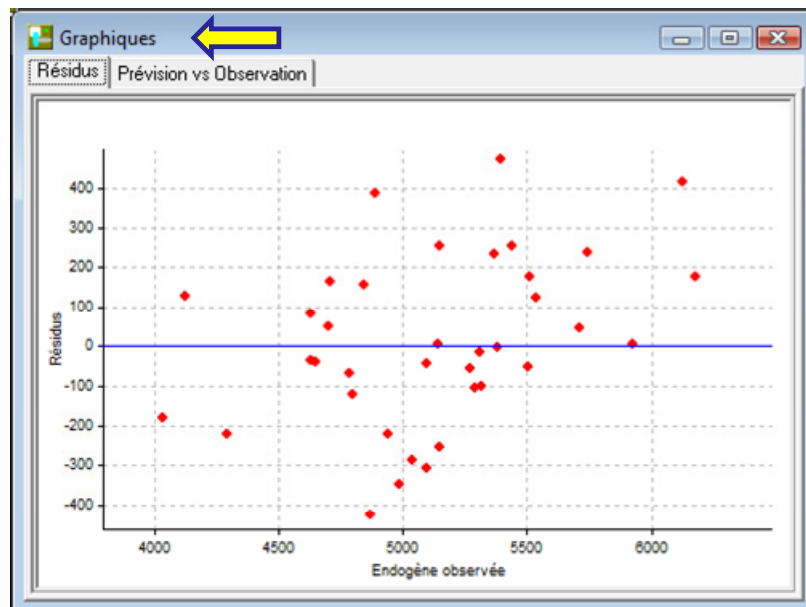


Détection des points atypiques (Detection of outliers). The ANALYSE / RESULTATS / POINTS ABERRANTS / AUTRES menu allows to obtain the indicators for the detection of outliers (Leverage [Hi], etc.; see <http://data-mining-tutorials.blogspot.com/2009/12/outliers-and-influential-points-in.html>).

The figure shows a window titled "Données Influentes et Points Aberrants" with a table of diagnostic statistics for 14 data points. The table includes columns for various indicators: Coupure, Hi, |DFFITS|, RSTUDENT, |COVRATIO|, Wilks, Mahalanobis, and Cook. Several values are highlighted in red boxes, indicating potential outliers or influential points.

	Coupure	Hi	DFFITS	RSTUDENT	COVRATIO	Wilks	Mahalanobis	Cook
	>0.4737	>0.9733	>2.0484	<[0.29,1.71]	<0.5405	>31.4270	>2.2229	
1		0.3354	0.4110	0.5784	1.8543	0.6825	17.1974	0.0192
2		0.2105	0.5807	1.1247	1.1671	0.8108	8.6254	0.0371
3		0.2531	0.5809	-0.9981	1.3404	0.7671	11.2231	0.0375
4		0.2801	0.1481	-0.2374	1.8707	0.7394	13.0324	0.0025
5		0.2872	1.3200	-2.0794	0.5284	0.7320	13.5343	0.1737
6		0.1457	0.2825	0.6841	1.3828	0.8774	5.1661	0.0090
7		0.2697	0.2307	0.3795	1.7932	0.7500	12.3239	0.0061
8		0.3379	1.8143	-2.5398	0.3205	0.6800	17.3982	0.3079
9		0.4284	0.5709	0.6594	2.0885	0.5870	26.0100	0.0369
10		0.3032	0.4792	0.7265	1.6636	0.7157	14.6902	0.0259
11		0.1916	0.4992	1.0253	1.2175	0.8302	7.5602	0.0276
12		0.1174	0.0662	-0.1815	1.5374	0.9065	3.8149	0.0005
13		0.1814	0.0010	-0.0022	1.6752	0.8407	7.0037	0.0000
14		0.2273	0.9668	1.7824	0.6747	0.7935	9.6197	0.0966

Graphiques (Some graphical representation). Last, we have two scatter plots (ANALYSIS / RESULTATS / GRAPHIQUES menu): (1) the first plots the residuals versus the values of the dependent variable; (2) the second plots the observed values of the dependent variable versus the predicted values.



7. Conclusion

REGRESS provides features fairly simple. It does not have all the advantages of more powerful tools such as TANAGRA (e.g. variable selection, etc.). But it has the advantage of being very easy to handle while being consistent with a degree course in Econometrics. As such, it may be useful for anyone wishing to learn about the regression without too much get involved in the learning of a new software.