# 1. Objectif

**Le logiciel REGRESS dans le package SIPINA**.

Peu de personnes le savent. En réalité, plusieurs logiciels sont installés lorsque l'on récupère et que l'on exécute le SETUP de SIPINA (setup_stat_package.exe) (Figure 1). Je n'en parle pas beaucoup parce que les autres techniques proposées (Régression Linéaire Multiple et Règles d'Association) sont déjà intégrées dans TANAGRA qui est très largement diffusé.

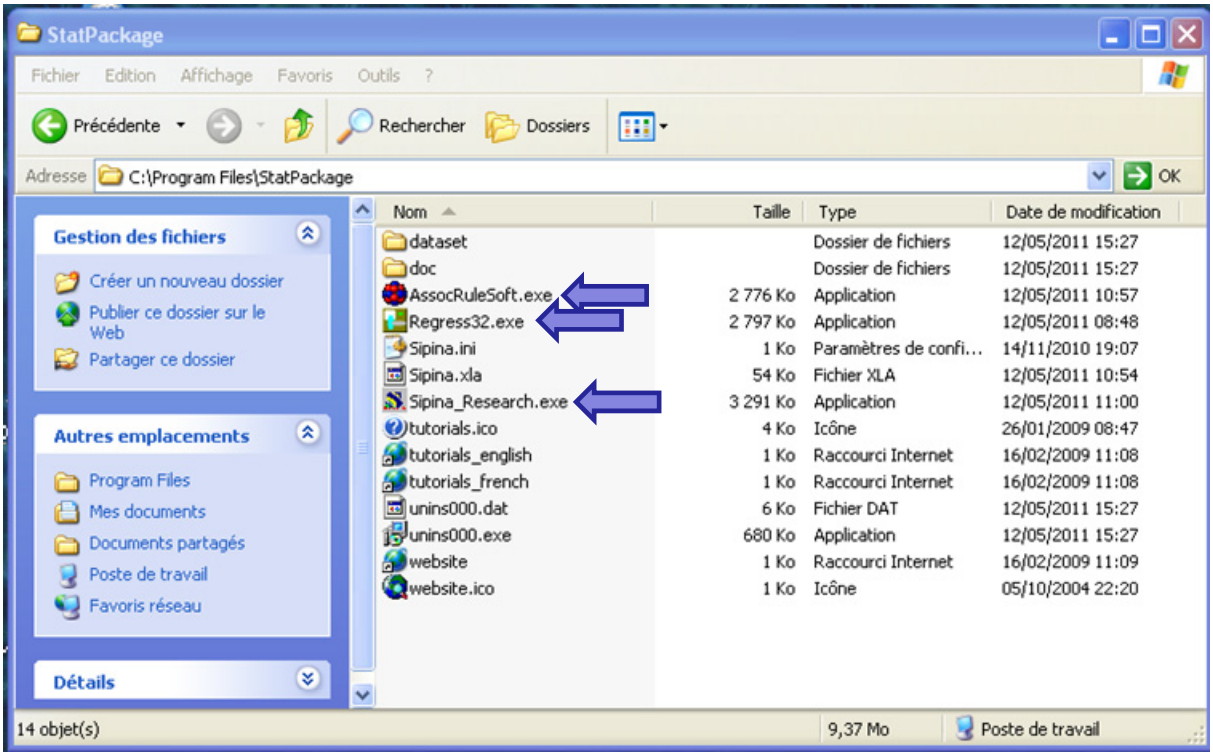



**Figure 1 - Applications installées par le setup de SIPINA**

Pourquoi en parler aujourd'hui alors ? Tout simplement parce que, concernant REGRESS en tous les cas, je me suis rendu compte en préparant le fascicule de cours consacré à la régression linéaire simple et multiple (Econométrie – Régression Linéaire Simple et Multiple), que le relatif manque de puissance du logiciel – par rapport à TANAGRA – est largement compensé par une grande facilité d'utilisation. Pour les utilisateurs qui souhaitent manipuler un outil simple, sans fioritures, REGRESS peut encore rendre de grands services.

L'autre raison est que j'ai souhaité proposé aux étudiants un outil qui cadre exactement avec le contenu de mon cours en Licence IDS (L3 Informatique Décisionnelle et Statistique - http://dis.univ-lyon2.fr/). Je m'acharne à les faire travailler sur tableur en TD. Mais une fois les connaissances acquises, disposer d'un outil qui permet de réaliser les calculs aisément est plutôt intéressant. A ce titre, tous les résultats de REGRESS sont en phase avec les

formules décrites dans mes fascicules de cours, celui cité ci-dessus, mais aussi celui consacré à la « Pratique de la Régression Linéaire Multiple ».

Enfin, la dernière raison est purement sentimentale. REGRESS est mon premier projet à vocation scientifique qui ait connu une certaine pérennité. Commencé en 1989 pour me faire la main sur Turbo Pascal, il a connu plusieurs mutations notables au fil des années : (1) mes connaissances en économétrie ont énormément évolué à partir du niveau Master (1992) ; (2) Patrick Sylvestre Baron, que j'ai eu comme enseignant, m'a demandé d'introduire des fonctionnalités supplémentaires afin de pouvoir l'utiliser pour les cours qu'il assurait à Lyon 2 (1994) ; (3) l'arrivée de Delphi m'a permis d'améliorer l'interface utilisateur à moindre effort (avec Turbo Pascal pour Windows, réaliser des grilles de saisie et d'affichage par exemple était vraiment compliqué) (1995) ; (4) enfin, le travail effectué pour la gestion de données de SIPINA (le fameux Data Manager de la version recherche) a été facilement porté dans REGRESS, résolvant en grande partie les problèmes de performances (traiter un fichier avec plusieurs milliers d'observations était mission impossible) qu'il pouvait y avoir sur les versions précédentes (1997).

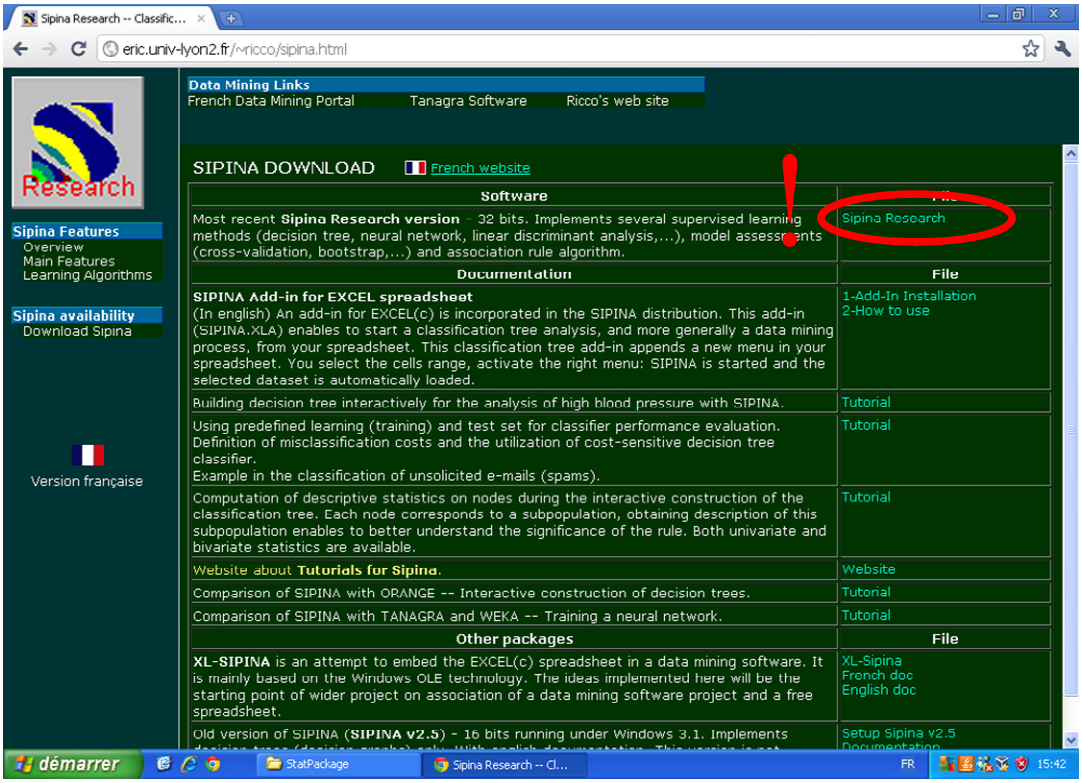
Depuis, REGRESS poursuivait sa petite vie tranquille sur mon site personnel sans que je ne m'en préoccupe réellement. Toute mon énergie étant tournée vers les autres outils d'obédience Data Mining que j'ai pu développer ensuite (SIPINA, puis TANAGRA).

Récemment, j'ai décidé de valoriser un peu plus SIPINA que j'avais plus ou moins abandonné depuis 2000, en documentant toute une série de fonctionnalités totalement méconnues et en lui dédiant un site web francophone (http://sipina.over-blog.fr/). A cette occasion, pris d'une frénésie nostalgique, j'avais décidé d'intégrer REGRESS dans le package, de même qu'un logiciel d'induction de règles d'association, première version de l'algorithme que j'ai ultérieurement perfectionné plusieurs fois dans TANAGRA (voir l'onglet ASSOCIATION dans la palette de composants). Mais il n'y avait pas vraiment une vraie volonté de valorisation à cette occasion. Il s'agissait surtout pour moi de ne pas abandonner des outils qui m'ont quand même mobilisé (surtout REGRESS) durant mes années d'études.

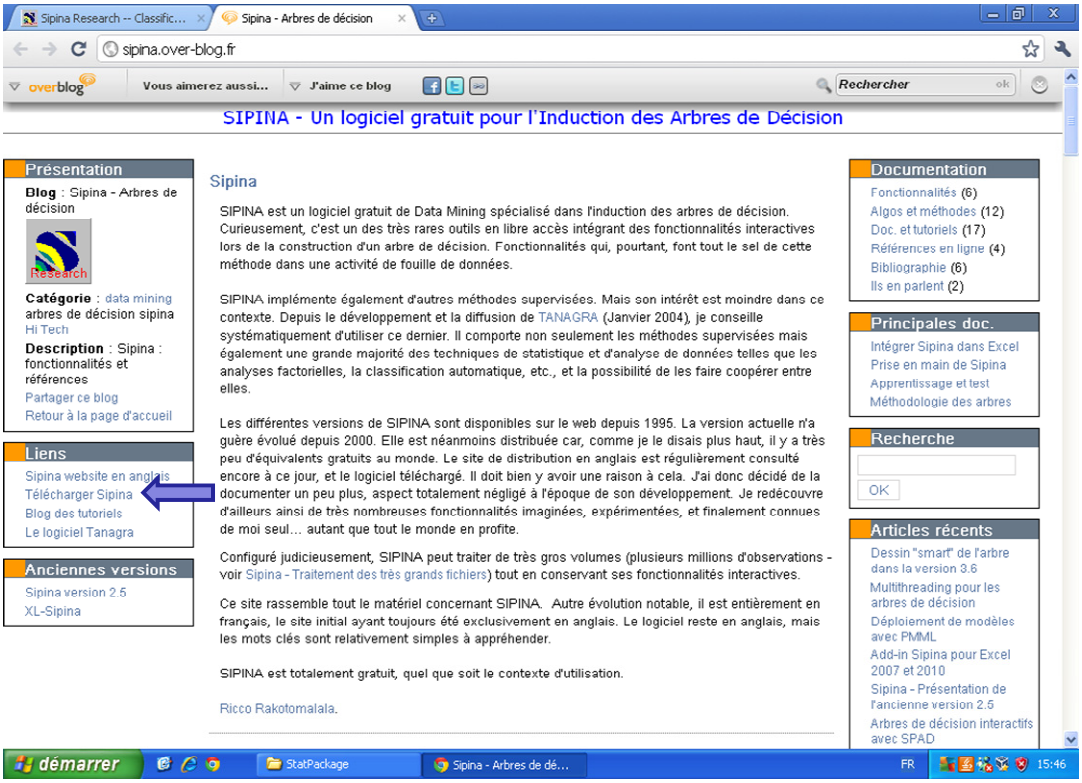
Aujourd'hui, comme je l'ai dit plus haut, j'ai réalisé de très nombreux calculs sur tableur en écrivant les supports pour mon cours d'Econométrie. J'avais besoin d'un outil qui me permette de croiser les résultats. REGRESS est revenu dans mon esprit. Je me suis dit qu'il fallait que je le sorte de sa léthargie. Je l'ai donc recompilé en introduisant deux améliorations : il peut s'intégrer dans le tableur Excel via une macro-complémentaire maintenant, la même que celle de SIPINA (**SIPINA.XLA**), cela accroît grandement sa facilité d'utilisation ; j'ai revérifié les formules pour qu'elles soient complètement cohérentes avec celles obtenues par tableur décrites dans mes fascicules de cours.

# 2. Installation de l'add-in SIPINA.XLA dans Excel

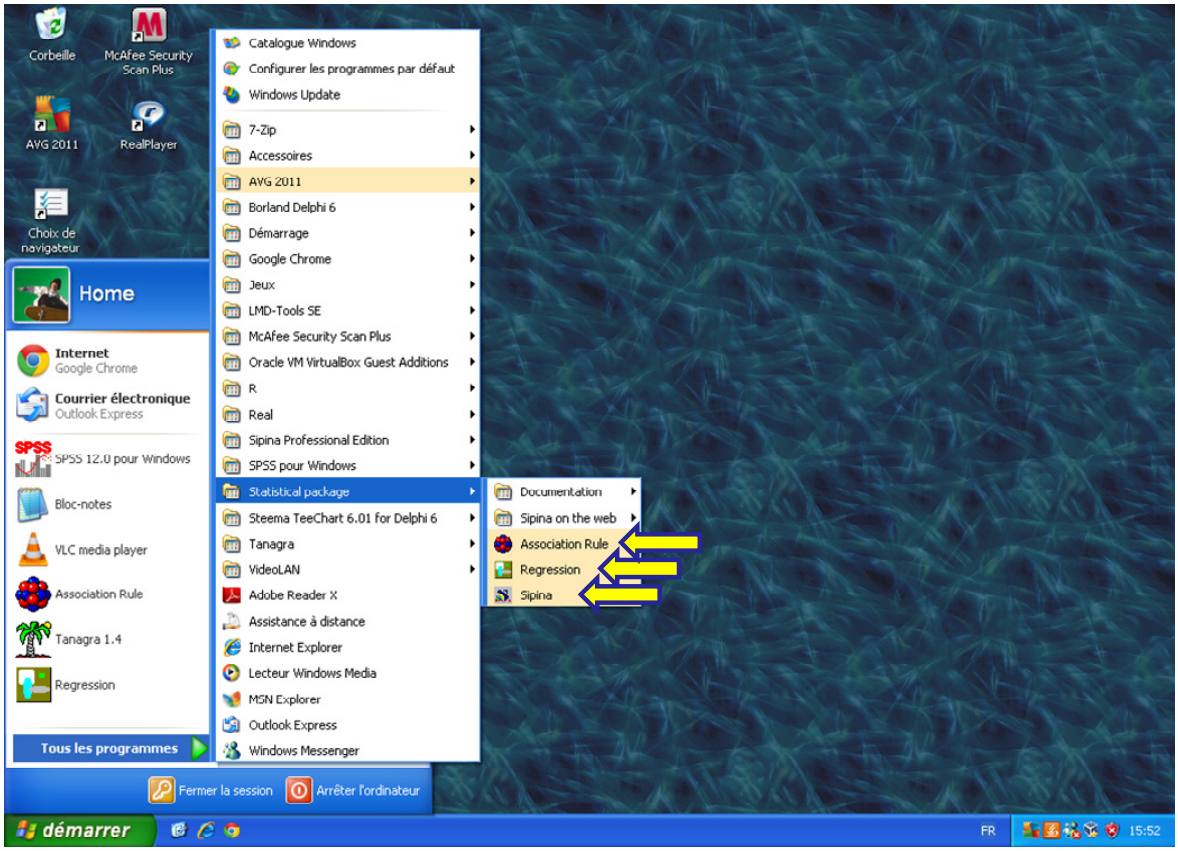
Pour installer REGRESS, il faut en réalité charger la SIPINA RESEARCH version sur le site web de SIPINA (***version 3.7*** du 12 mai 2011), sur la page anglophone




Ou sur la page francophone.



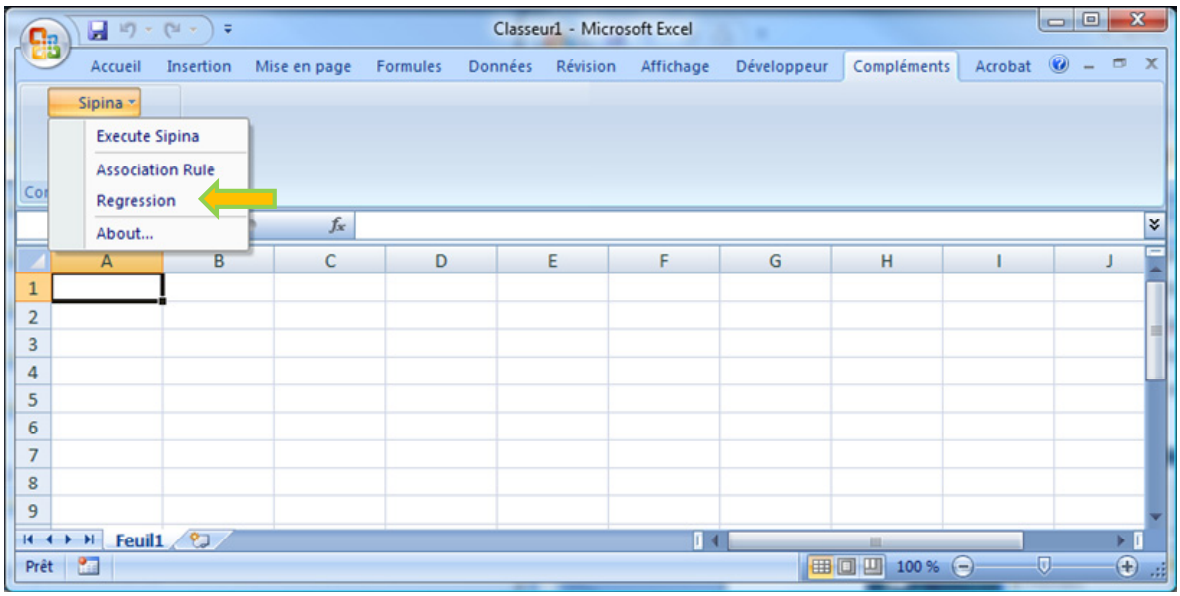
Une fois téléchargé, nous procédons à l'installation du logiciel de manière tout à fait classique en exécutant le fichier « **setup_stat_package.exe** ». Dans le menu « Démarrer » de Windows, sous le groupe « Statistical Package », nous avons effectivement les 3 logiciels : SIPINA proprement dit, spécialisé dans l'induction des arbres de décision ; un outil dédié à la génération des règles d'association ; et REGRESS, spécialisé dans la régression linéaire simple et multiple.




Nous pouvons installer la macro-complémentaire (add-in en anglais) **SIPINA.XLA** dans Excel. Nous y avons consacré plusieurs tutoriels par le passé : pour Excel 2003 et versions antérieures (http://sipina.over-blog.fr/article-17592277.html) ; pour Excel 2007 et 2010, nous décrivons la procédure pour Tanagra mais la transposition à SIPINA est triviale (http://tutoriels-data-mining.blogspot.com/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html).

Nous démarrons alors le tableur. Dans l'onglet « **Compléments** » d'Excel 2007 (nous avons un menu à part dans Excel 2003 et versions antérieures) apparaît le menu SIPINA. Lorsque nous le sélectionnons, nous voyons apparaître les items correspondants aux trois outils distribués avec le package SIPINA. Outre le logiciel SIPINA pour l'induction des arbres de décision, nous pouvons accéder au logiciel de génération de règle et au logiciel dédié à la régression REGRESS. Tous bénéficient de la même fonctionnalité de gestion des données :

nous pouvons charger notre fichier dans Excel, puis l'envoyer au logiciel spécialisé pour les traitements statistiques.




# 3. Données

Pour décrire la mise en œuvre de REGRESS, nous reprenons le fichier « ventes-regression.xls » décrit dans l'ouvrage de Michel Tenenhaus[1], page 101.

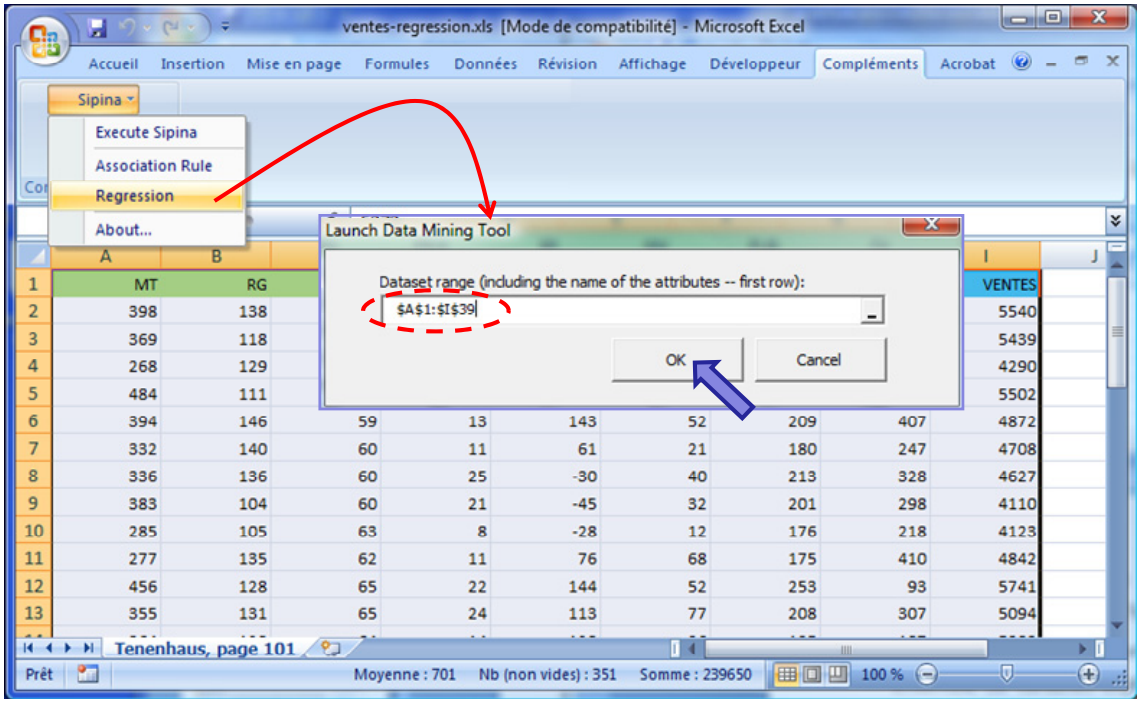
Il s'agit d'expliquer le volume de vente VENTES d'un produit quelconque à partir d'un ensemble de variables explicatives : MT (marché total de la branche), RG (remise aux grossistes), PRIX (prix du produit), BR (budget de recherche), INV (investissements), PUB (publicité), FV (frais de vente) et TPUB (total du budget de publicité dans la branche). Le fichier comporte 38 observations.

Au-delà de l'énumération des manipulations à effectuer, notre idée est également de suivre à la trace les résultats en les mettant en parallèle avec ceux de notre ouvrage de référence. L'auteur a utilisé le logiciel SPSS.

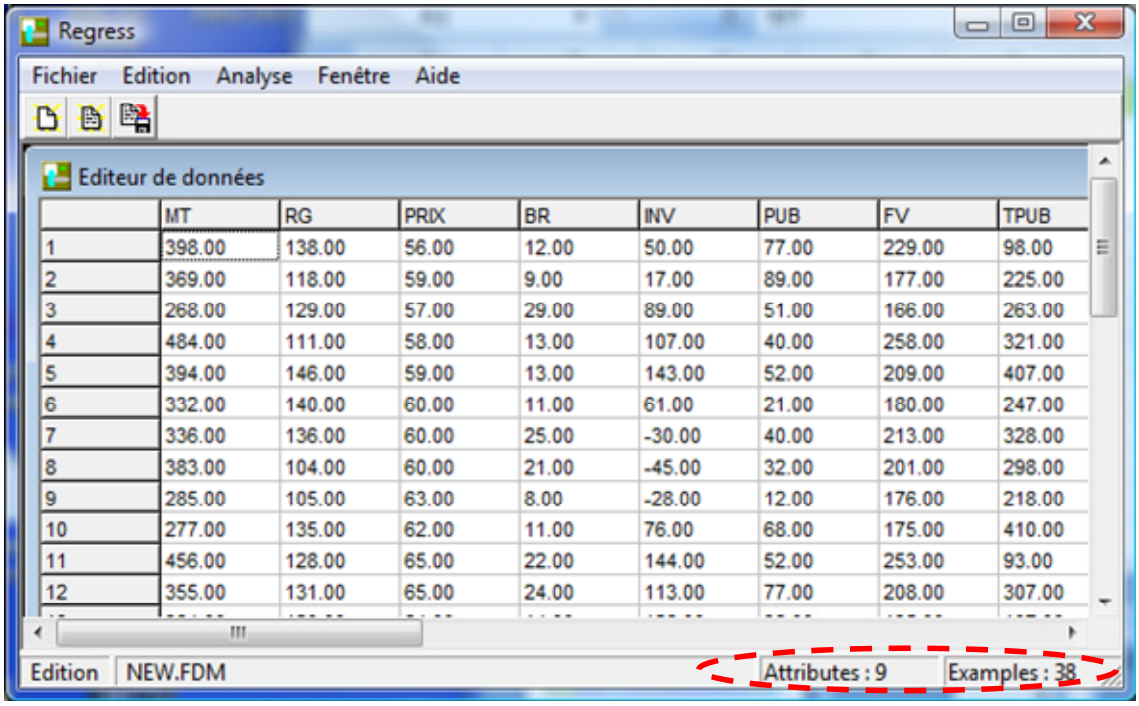
# 4. Importation des données et démarrage de REGRESS

Nous chargeons le fichier dans le tableur Excel. Avec avoir sélectionné la plage de données, nous actionnons le menu COMPLEMENTS / SIPINA / REGRESSION. Une boîte de dialogue apparaît, nous vérifions les coordonnées, puis nous cliquons sur OK.

[1] M. Tenenhaus, « Statistique – Méthodes pour décrire, expliquer et prévoir », Dunod, 2007 ; la source originelle est Wheelwright S.C. et Makridakis S., « Choix et valeur des méthodes de prévision », Editions d'Organisation, Paris, 1974.



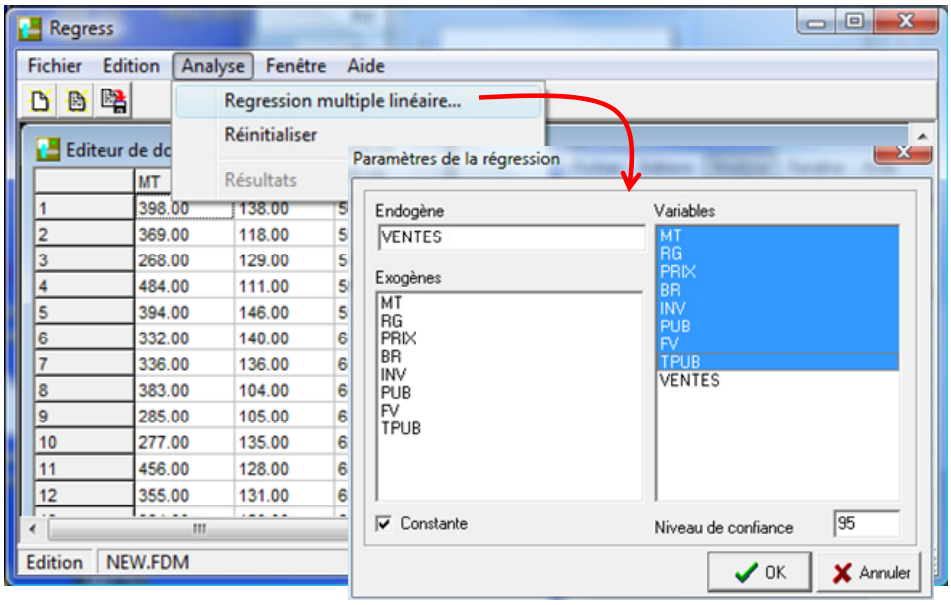

REGRESS est automatiquement démarré et les données chargées.




| | 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 |

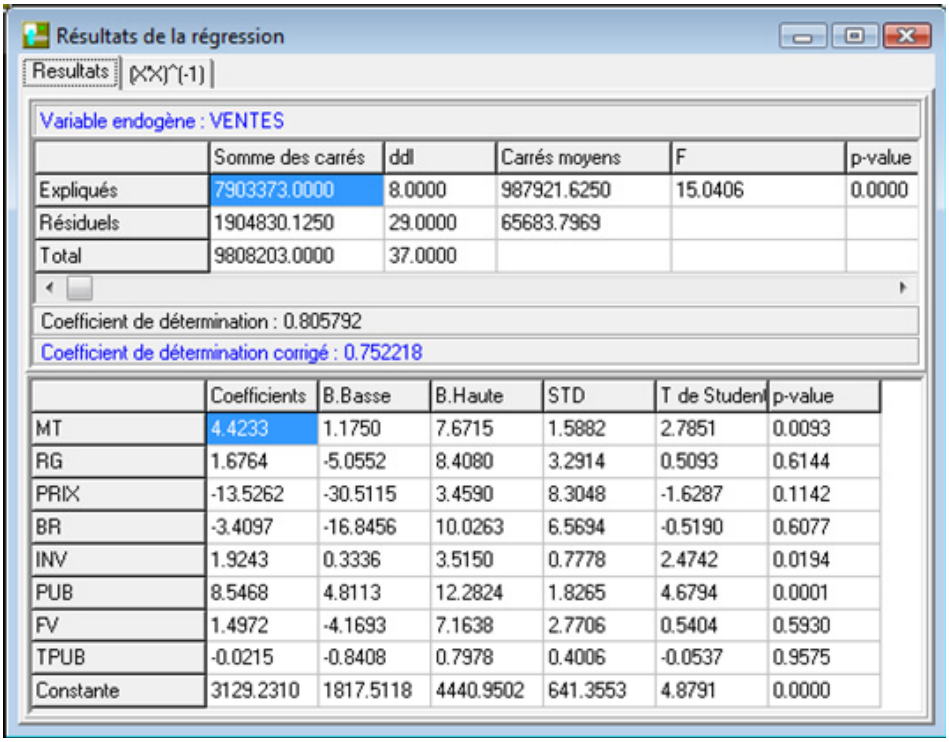
## 5. Régression linéaire multiple

Nous devons indiquer à REGRESS la variable cible VENTES et les explicatives (les autres). Nous actionnons le menu ANALYSE / REGRESSION LINEAIRE MULTIPLE. Dans la boîte de paramétrage qui apparaît, nous effectuons par glisser-déposer les sélections adéquates.




Par défaut, REGRESS effectue une régression avec constante. Le niveau de confiance pour le calcul des intervalles de confiance des coefficients est 95%. Nous validons. Les calculs sont automatiquement démarrés. Plusieurs fenêtres apparaissent.

**Résultats de la régression**. Elle fournit le tableau d'analyse de variance, le coefficient de détermination $R^2$ = 0.805792 (Tenenhaus, page 114), le $R^2$ ajusté = 0.752218 (Tenenhaus, page 116), et la grille des coefficients (Tenenhaus, page 107).

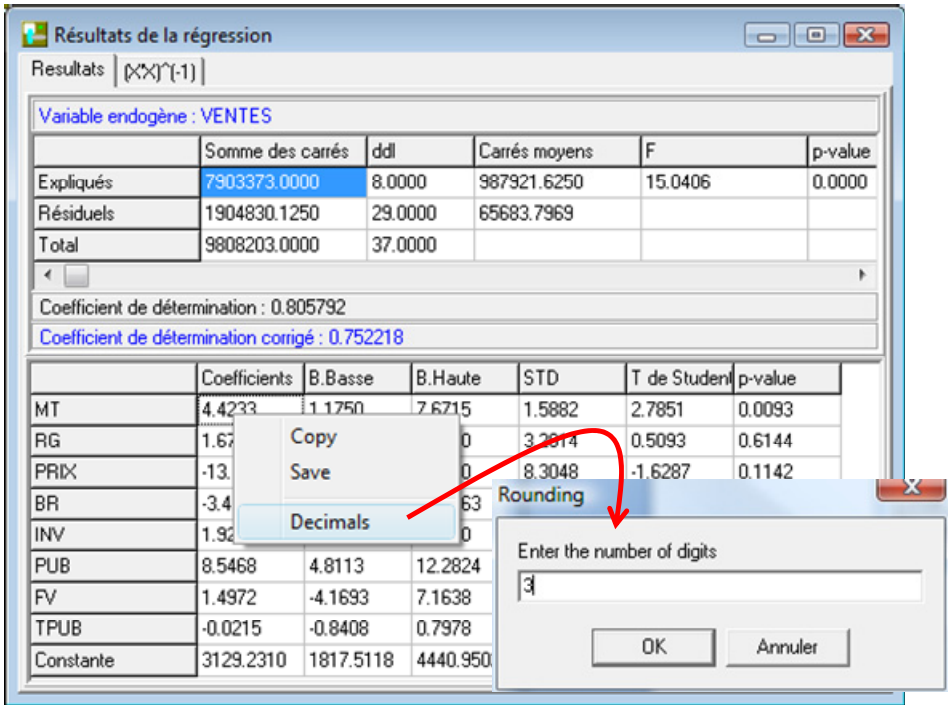



| | Somme des carrés | ddl | Carrés moyens | F | p-value |
|---|---|---|---|---|---|
| Expliqués | 7903373.0000 | 8.0000 | 987921.6250 | 15.0406 | 0.0000 |
| Résiduels | 1904830.1250 | 29.0000 | 65683.7969 | | |
| Total | 9808203.0000 | 37.0000 | | | |

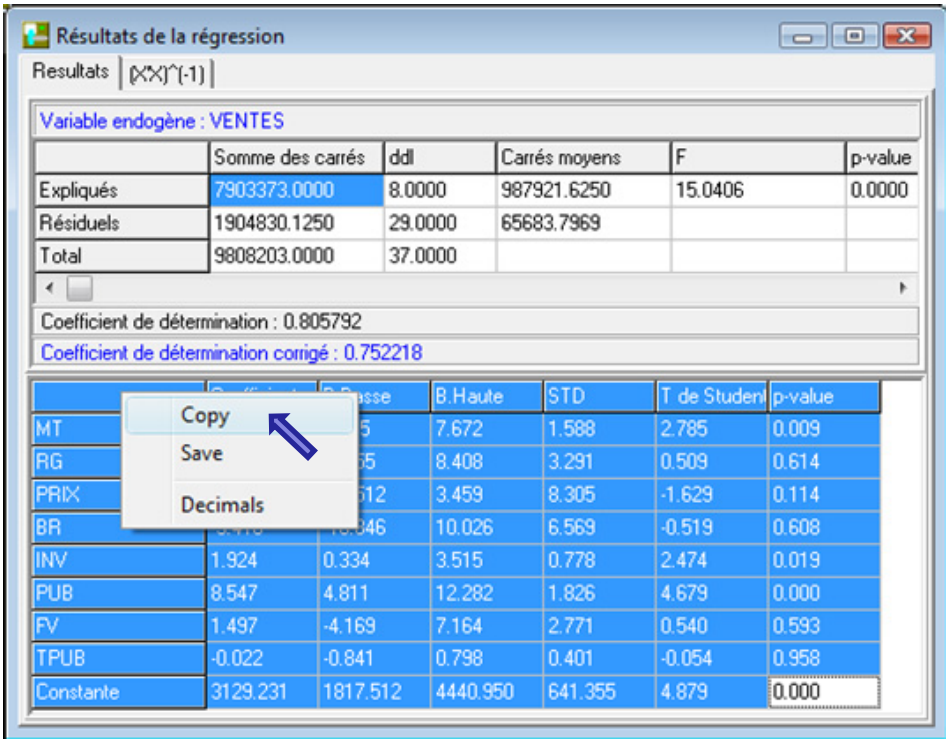
| | Coefficients | B.Basse | B.Haute | STD | T de Student | p-value |
|---|---|---|---|---|---|---|
| MT | 4.4233 | 1.1750 | 7.6715 | 1.5882 | 2.7851 | 0.0093 |
| RG | 1.6764 | -5.0552 | 8.4080 | 3.2914 | 0.5093 | 0.6144 |
| PRIX | -13.5262 | -30.5115 | 3.4590 | 8.3048 | -1.6287 | 0.1142 |
| BR | -3.4097 | -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 |

Dans cette dernière, nous y trouvons : les coefficients estimés, les bornes basses et hautes de leur intervalle de confiance, les écarts-type, le t de Student du test de significativité, et la probabilité critique (p-value) associée.

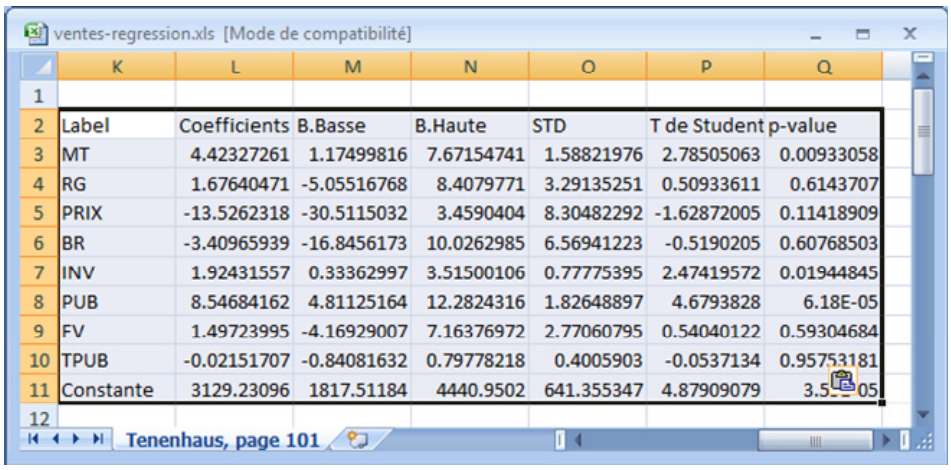
Il est possible de moduler la précision de l'affichage en faisant apparaître le menu contextuel (clic droit sur la grille) et en cliquant sur le menu DECIMALS. En choisissant 3 chiffres après la virgule, nous retrouvons exactement les valeurs (coefficients, p-value) du tableau 5.3 décrit dans notre ouvrage de référence.




Notons également qu'il est possible de copier les valeurs de la grille dans un tableur pour des calculs ultérieurs. Nous cliquons sur la première cellule en haut à gauche de la grille, le tableau est sélectionné, nous actionnons l'item COPY du menu contextuel.



La copie se fait avec la précision maximale. Libre à nous de fixer la précision qui nous sied.

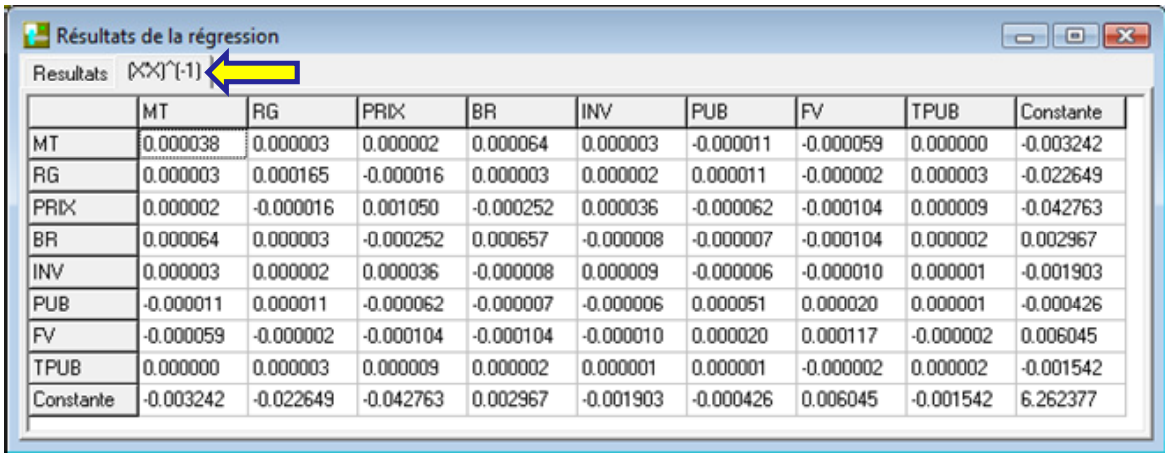


ventes-regression.xls [Mode de compatibilité]

| | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | Label | Coefficients | B.Basse | B.Haute | STD | T de Student | p-value |
| 3 | MT | 4.42327261 | 1.17499816 | 7.67154741 | 1.58821976 | 2.78505063 | 0.00933058 |
| 4 | RG | 1.67640471 | -5.05516768 | 8.4079771 | 3.29135251 | 0.50933611 | 0.6143707 |
| 5 | PRIX | -13.5262318 | -30.5115032 | 3.4590404 | 8.30482292 | -1.62872005 | 0.11418909 |
| 6 | BR | -3.40965939 | -16.8456173 | 10.0262985 | 6.56941223 | -0.5190205 | 0.60768503 |
| 7 | INV | 1.92431557 | 0.33362997 | 3.51500106 | 0.77775395 | 2.47419572 | 0.01944845 |
| 8 | PUB | 8.54684162 | 4.81125164 | 12.2824316 | 1.82648897 | 4.6793828 | 6.18E-05 |
| 9 | FV | 1.49723995 | -4.16929007 | 7.16376972 | 2.77060795 | 0.54040122 | 0.59304684 |
| 10 | TPUB | -0.02151707 | -0.84081632 | 0.79778218 | 0.4005903 | -0.0537134 | 0.95753181 |
| 11 | Constante | 3129.23096 | 1817.51184 | 4440.9502 | 641.355347 | 4.87909079 | 3.5[illegible]05 |
| 12 | | | | | | | |

Tenenhaus, page 101

**La matrice des covariances entre les variables**. Dans l'onglet (X'X)^(-1), REGRESS fournit la matrice des covariances non centrées entre les variables.



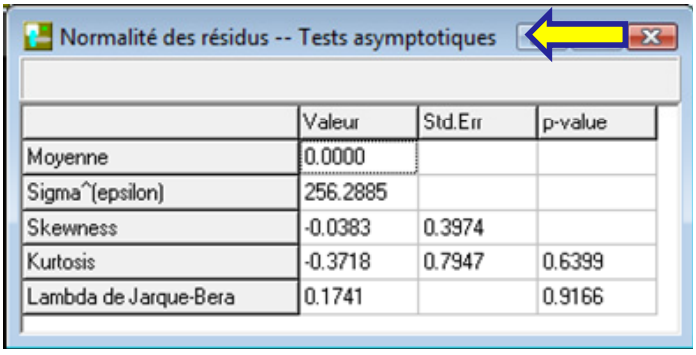
Résultats de la régression

Resultats (X'X)^(-1)

| | 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 |

Elle est très utile lors de l'inférence statistique : covariance des coefficients estimés, tests généralisés, calcul du levier des observations, intervalle de prédiction, etc.

**Normalité des résidus**. Dans la fenêtre « Normalité des résidus – Tests asymptotiques », nous disposons des informations sur les erreurs observées. La moyenne est forcément nulle dans une régression avec constante. Nous avons également l'estimation sans biais de l'écart-type de l'erreur (Tenenhaus, page 110).



Normalité des résidus -- Tests asymptotiques

| | 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 |

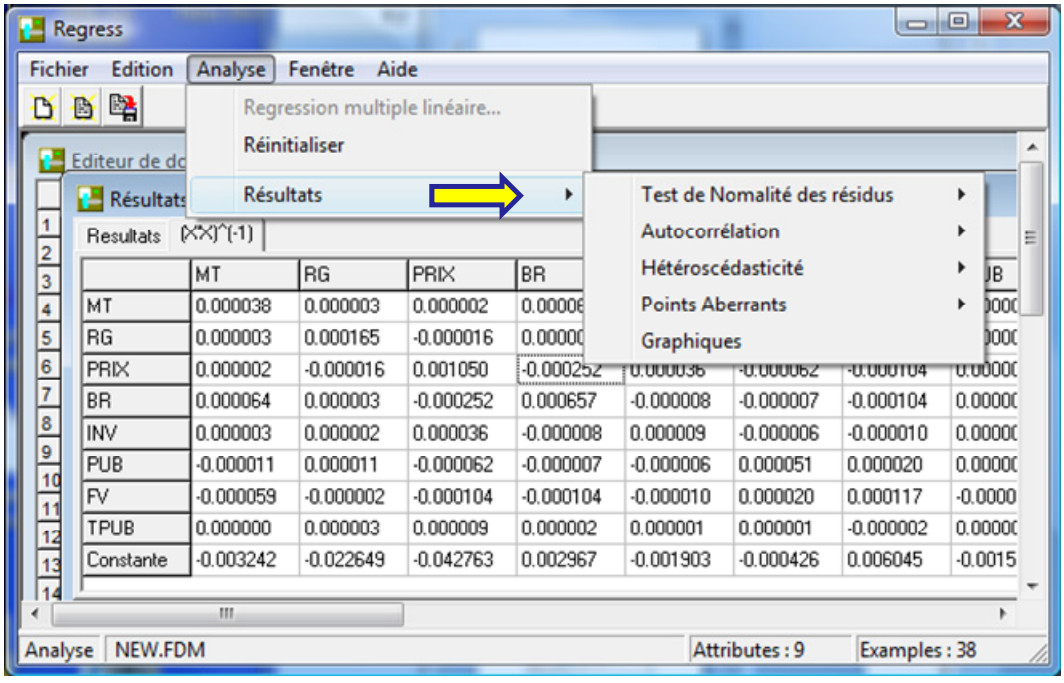
**Skewness**. Pour évaluer la non adéquation avec la loi normale, REGRESS fournit une estimation du coefficient d'asymétrie, de son écart-type et la probabilité critique (p-value) du test de conformité à la loi normale. Attention, il s'agit de tests asymptotiques. Ils sont décrits dans notre fascicule consacré à la « Pratique de la Régression »[2]. Dans notre exemple, la p-value étant élevée, plus grande que le risque alpha = 5% que nous nous sommes fixés, l'hypothèse de normalité des résidus ne peut pas être rejetée.

**Kurtosis**. Il s'agit du coefficient d'aplatissement, le schéma est identique au test précédent (Rakotomalala, section 1.3.3).

**Lambda de Jarque Bera**. C'est une combinaison des deux indicateurs précédents. Le test est plus puissant. La statistique est distribuée selon une loi du KHI-DEUX à 2 degrés de liberté (Rakotomalala, section 1.3.3). Ici également, on constate que l'hypothèse de normalité ne peut être rejetée.

# 6. Résultats additionnels

Au-delà des fenêtres de résultats générées automatiquement, il est possible d'accéder à d'autres informations relatives à la régression. Nous accédons au menu ANALYSE / RESULTATS pour ce faire.

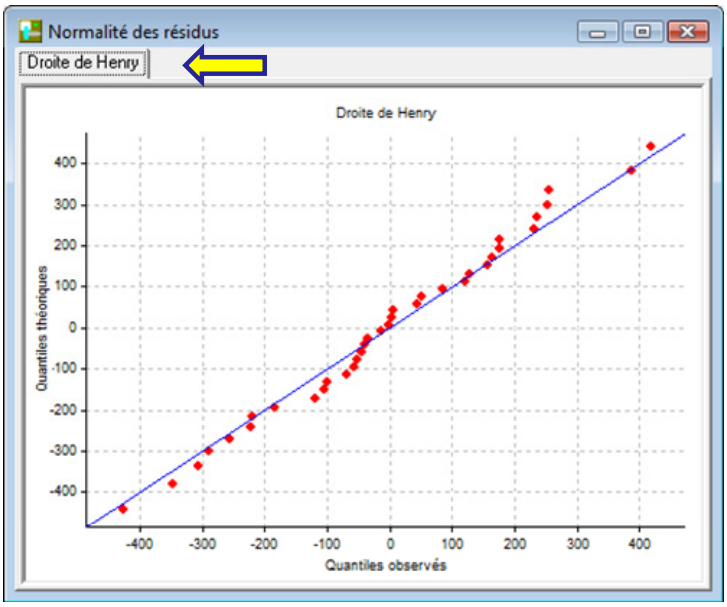



| | MT | RG | PRIX | BR | | | | |
|---|---|---|---|---|---|---|---|---|
| MT | 0.000038 | 0.000003 | 0.000002 | 0.00006 | | | | |
| RG | 0.000003 | 0.000165 | -0.000016 | 0.00000 | | | | |
| PRIX | 0.000002 | -0.000016 | 0.001050 | -0.000252 | 0.000036 | -0.000062 | -0.000104 | 0.00000 |
| BR | 0.000064 | 0.000003 | -0.000252 | 0.000657 | -0.000008 | -0.000007 | -0.000104 | 0.00000 |
| INV | 0.000003 | 0.000002 | 0.000036 | -0.000008 | 0.000009 | -0.000006 | -0.000010 | 0.00000 |
| PUB | -0.000011 | 0.000011 | -0.000062 | -0.000007 | -0.000006 | 0.000051 | 0.000020 | 0.00000 |
| FV | -0.000059 | -0.000002 | -0.000104 | -0.000104 | -0.000010 | 0.000020 | 0.000117 | -0.0000 |
| TPUB | 0.000000 | 0.000003 | 0.000009 | 0.000002 | 0.000001 | 0.000001 | -0.000002 | 0.00000 |
| Constante | -0.003242 | -0.022649 | -0.042763 | 0.002967 | -0.001903 | -0.000426 | 0.006045 | -0.0015 |

Nous laisserons de côté les tests et traitements de l'hétéroscédasticité et de l'auto corrélation des résidus. Ils feront l'objet de tutoriels spécifiques ultérieurement.

[2] R. Rakotomalala, « Pratique de la régression linéaire multiple – Diagnostic et sélection de variables », section 1.3.2, http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

**Droite de Henry**. Via le menu ANALYSE / RESULTATS / TEST DE NORMALITE DES RESIDUS / DROITE DE HENRY, nous accédons à un graphique permettant de vérifier l'adéquation de la distribution empirique des résidus avec la distribution gaussienne. Grosso modo, lorsque les points forment une droite, la compatibilité avec la loi normale ne peut pas être refutée. Ce qui est le cas de notre exemple.




**Détection des points atypiques**. Avec ANALYSE / RESULTATS / POINTS ABERRANTS / AUTRES, nous affichons les indicateurs d'évaluation du rôle de chaque observation dans la régression. L'objectif étant de détecter celles qui joueraient un rôle néfaste -- parce qu'exagéré -- sur les résultats obtenus (Tenenhaus, tableau 5.4, page 109).

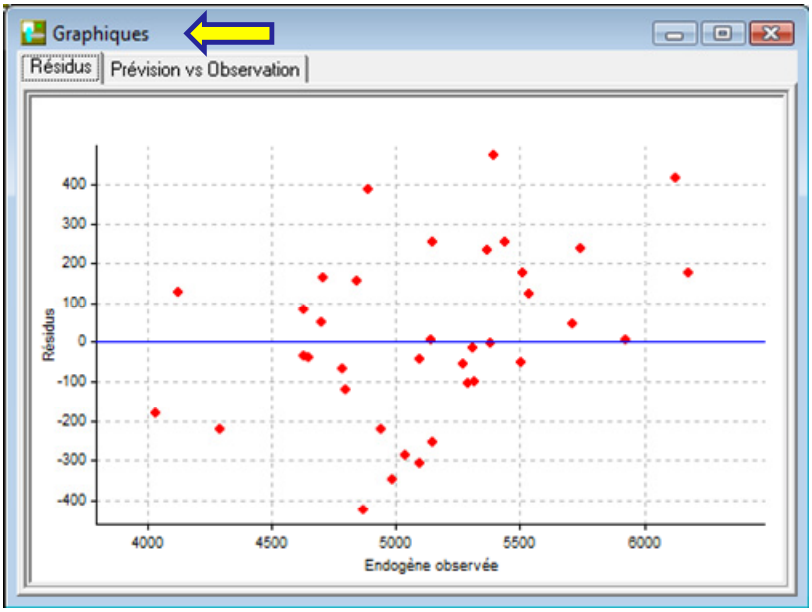
Données Influentes et Points Aberrants

| Coupure | >0.4737 | >0.9733 | >2.0484 | <>[0.29,1.71] | <0.5405 | >31.4270 | >2.2229 |
|---|---|---|---|---|---|---|---|
| | Hi | IDFFITSI | RSTUDENT | ICOVRATIO | Wilks | Mahalanobis | Cook |
| 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 |

Par rapport aux résultats de SPSS rapporté dans l'ouvrage, les résultats sont identiques, à l'exception du levier. La raison est que SPSS calcule le « levier centré ». Il faudrait ajouter la

valeur (1/n) où « n » est le nombre d'observations pour retrouver les valeurs de REGRESS, qui sont par ailleurs en adéquation avec les sorties de TANAGRA, SAS et R (http://tutoriels-data-mining.blogspot.com/2008/04/points-aberrants-et-influents-dans-la.html).

**Graphiques**. Dernière fenêtre que nous décrirons dans ce tutoriel, REGRESS fournit les graphiques « Résidus vs. Endogène » et « Endogène observée vs. Endogène calculée » via le menu ANALYSIS / RESULTATS / GRAPHIQUES.




# 7. Conclusion

REGRESS propose des fonctionnalités assez simples. Il ne possède pas tous les atouts d'outils plus puissants tels que TANAGRA par exemple (sélection de variables, etc.). Mais il a l'énorme avantage d'être très facile à manipuler tout en étant en phase avec un cours de Licence d'Econométrie. A ce titre, il peut être intéressant pour toute personne désireuse de s'initier à la régression sans avoir à s'investir outre mesure dans l'apprentissage d'un logiciel.