Objectif

Quelques stratégies pour combattre la colinéarité dans la régression linéaire multiple.

Un des plus gros écueils de la régression est la colinéarité c.-à-d. les variables exogènes sont excessivement corrélées. Les coefficients deviennent incohérents, en contradiction avec les connaissances du domaine. Des variables, a priori très importantes, paraissent non significatives, elles sont par conséquent éliminées à tort.

Il importe : (1) de déterminer s'il y a colinéarité dans la régression que nous menons ; (2) de proposer des solutions pour obtenir des résultats consistants.

Dans ce didacticiel, nous étudierons trois approches destinées à surmonter la colinéarité : la sélection de variables, la régression sur les composantes orthogonales, la régression PLS[1].

Données

Fichier de données

Les données décrivent **27** véhicules. L'objectif est de préciser les déterminants de la consommation d'une automobile à partir de certaines caractéristiques telles que le poids, la cylindrée (taille du moteur), etc. Nous avons beaucoup utilisé ce fichier car_consumption_colinearity_regression.xls dans nos didacticiels en raison de ses caractéristiques propices à de multiples analyses et commentaires.

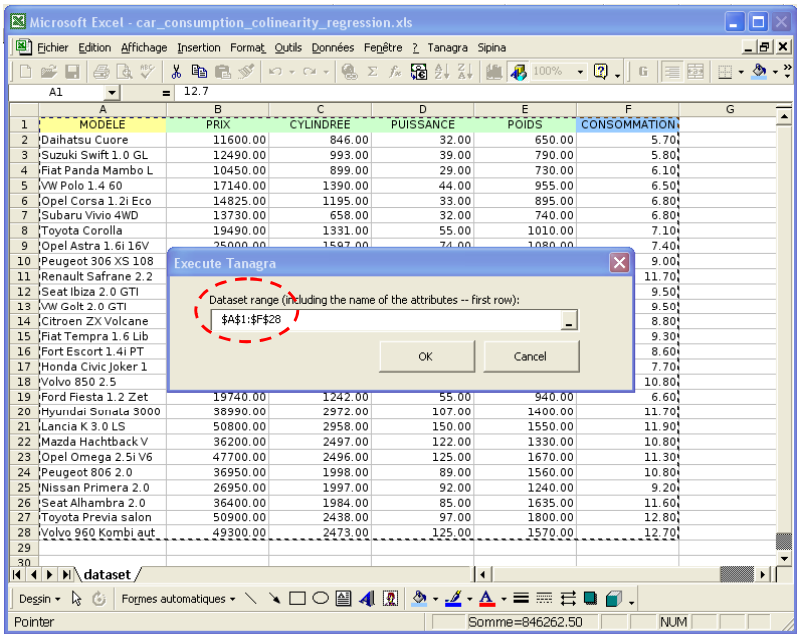
MODELE	PRIX	CYLINDREE	PUISSANCE	POIDS	CONSOMMATION
Daihatsu Cuore	11600.00	846.00	32.00	650.00	5.70
Suzuki Swift 1.0 GL	12490.00	993.00	39.00	790.00	5.80
Fiat Panda Mambo L	10450.00	899.00	29.00	730.00	6.10
VW Polo 1.4 60	17140.00	1390.00	44.00	955.00	6.50
Opel Corsa 1.2i Eco	14825.00	1195.00	33.00	895.00	6.80
Subaru Vivio 4WD	13730.00	658.00	32.00	740.00	6.80
Toyota Corolla	19490.00	1331.00	55.00	1010.00	7.10
Opel Astra 1.6i 16V	25000.00	1597.00	74.00	1080.00	7.40
Peugeot 306 XS 108	22350.00	1761.00	74.00	1100.00	9.00
Renault Safrane 2.2	36600.00	2165.00	101.00	1500.00	11.70
Seat Ibiza 2.0 GTI	22500.00	1983.00	85.00	1075.00	9.50
VW Golt 2.0 GTI	31580.00	1984.00	85.00	1155.00	9.50
Citroen ZX Volcane	28750.00	1998.00	89.00	1140.00	8.80
Fiat Tempra 1.6 Lib	22600.00	1580.00	65.00	1080.00	9.30
Fort Escort 1.4i PT	20300.00	1390.00	54.00	1110.00	8.60
Honda Civic Joker 1	19900.00	1396.00	66.00	1140.00	7.70
Volvo 850 2.5	39800.00	2435.00	106.00	1370.00	10.80
Ford Fiesta 1.2 Zet	19740.00	1242.00	55.00	940.00	6.60
Hyundai Sonata 3000	38990.00	2972.00	107.00	1400.00	11.70
Lancia K 3.0 LS	50800.00	2958.00	150.00	1550.00	11.90
Mazda Hachtback V	36200.00	2497.00	122.00	1330.00	10.80
Opel Omega 2.5i V6	47700.00	2496.00	125.00	1670.00	11.30
Peugeot 806 2.0	36950.00	1998.00	89.00	1560.00	10.80
Nissan Primera 2.0	26950.00	1997.00	92.00	1240.00	9.20
Seat Alhambra 2.0	36400.00	1984.00	85.00	1635.00	11.60
Toyota Previa salon	50900.00	2438.00	97.00	1800.00	12.80
Volvo 960 Kombi aut	49300.00	2473.00	125.00	1570.00	12.70

Figure 1 - Tableau de données

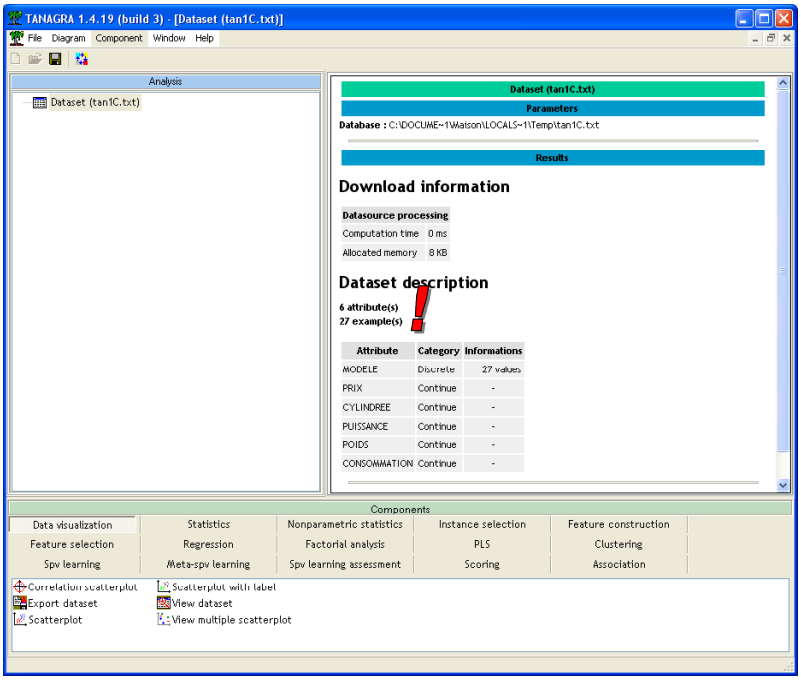
[1] Concernant le problème de la colinéarité dans la régression et les techniques destinées à y remédier, notamment la régression PLS, voir le document en ligne : S. Vancolen, « La régression PLS », Université de Neuchâtel ; http://doc.rero.ch/lm.php?url=1000,41,4,20070716085523-YM/mem_VancolenS.pdf

Créer un diagramme dans TANAGRA

Dans un premier temps, il faut initialiser un diagramme de traitements et charger les données dans le logiciel TANAGRA. Le plus simple est d'ouvrir le fichier car_consumption_colinearity_regression.xls dans le tableur EXCEL. Nous sélectionnons la plage de données et nous activons le menu TANAGRA/EXECUTE TANAGRA installée à l'aide de la macro complémentaire TANAGRA.XLA livrée avec le logiciel[2].



TANAGRA est automatiquement lancé, un nouveau diagramme de traitements est mis en place. Les données sont disponibles à la racine du diagramme.



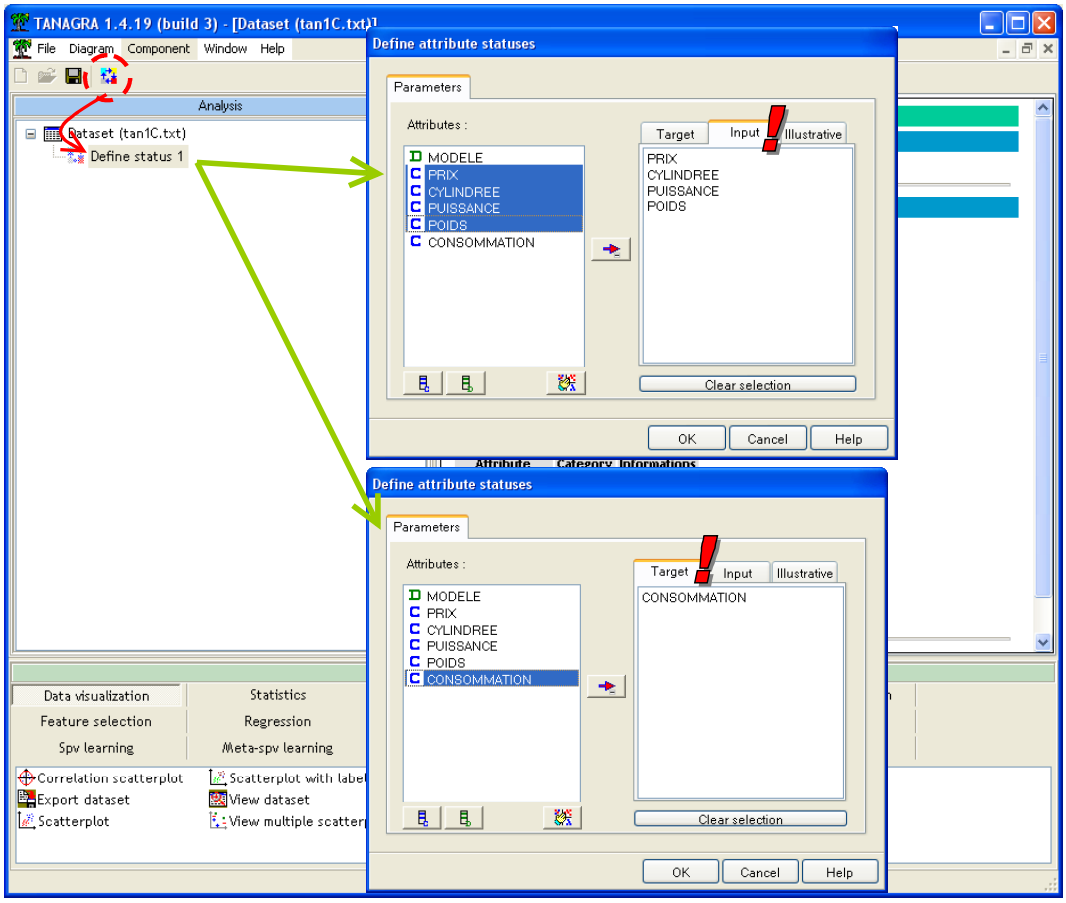
[2] Cette macro est disponible depuis la version 1.4.11 de TANAGRA. Un didacticiel décrit la procédure d'installation (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Excel_AddIn.pdf).

Régression linéaire multiple et colinéarité

Régression linéaire multiple

Dans un premier temps, sans connaissances préalables sur le sujet, nous décidons d'effectuer une régression linéaire multiple en intégrant toutes les variables explicatives candidates (les variables exogènes) : PRIX, CYLINDREE, PUISSANCE et POIDS. La variable expliquée (variable endogène) est CONSOMMATION.

Dans TANAGRA, nous devons définir avec un composant spécifique, DEFINE STATUS, le rôle des variables dans l'analyse. Pour ce faire, nous utilisons le raccourci dans la barre d'outils. La boîte de paramétrage apparaît automatiquement, nous plaçons en INPUT les variables explicatives, en TARGET, la variable à expliquer.



Nous introduisons le composant MULTIPLE LINEAR REGRESSION (onglet REGRESSION) dans le diagramme, à la suite de DEFINE STATUS 1. Nous activons le menu VIEW pour accéder aux résultats. Plusieurs éléments retiennent notre attention.

La régression semble excellente, le coefficient de détermination R^2 est égal à 0.9295 : près de 93% de la variance de CONSOMATION est expliquée par la régression. Nous sommes plutôt confiants par rapport aux résultats.

Pourtant, à la lecture des coefficients associés aux variables, plusieurs incohérences viennent perturber nos certitudes :

- Seul le poids est significatif pour expliquer la consommation. Le coefficient est positif, plus la voiture est lourde, plus elle consomme, cela paraît raisonnable.

- Ni la puissance, ni la cylindrée ne semblent peser sur la consommation ? C'est étrange. Sans être un grand expert, cela ne cadre pas du tout avec ce que l'on sait des automobiles. Ca laisse entendre qu'une Lada et une Ferrari de même poids auront une consommation identique, même si la seconde a un moteur 4 fois plus volumineux (ou plus puissant) que la première.

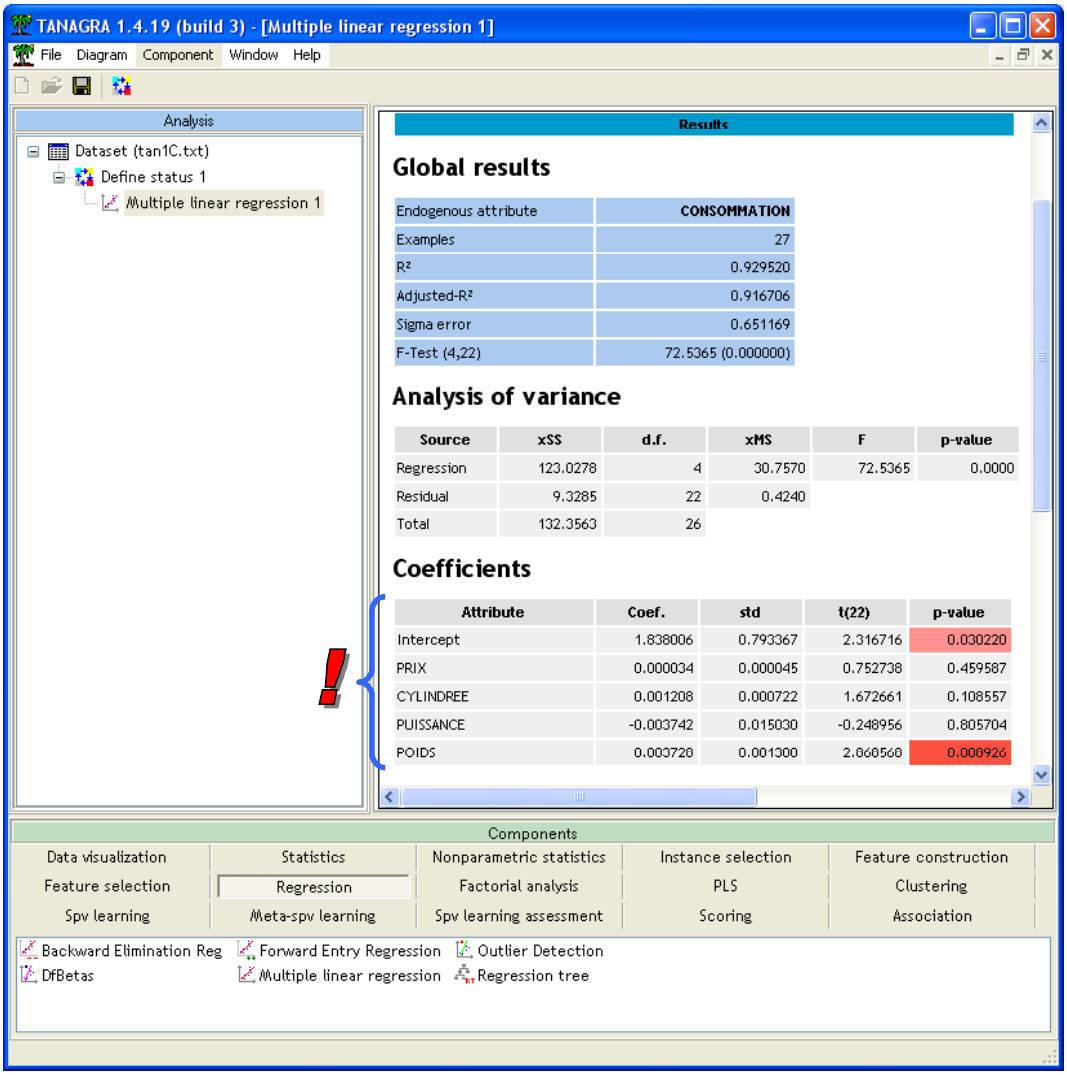


Figure 2 - Régression sur l'ensemble des variables

Détection de la colinéarité

Nous suspectons un problème de colinéarité. Certaines variables exogènes sont corrélées au point qu'elles se gênent lors des calculs. Cela se manifeste de différentes manières. Les résultats sont très instables, une petite modification des observations entraîne une forte modification des paramètres estimés. Les signes et les valeurs des paramètres sont incohérents, par rapport aux autres variables et par rapport aux connaissances du domaine. Dans notre cas, puissance aurait une influence négative sur la consommation ? Ce n'est pas très défendable. Enfin, des variables paraissent non pertinentes à cause d'une variance mal estimée. Le test de nullité des coefficients, le t de Student, renvoie des valeurs faussées.

Bref, nous avons une prétendue excellente régression mais inexploitable car nous ne pouvons pas en tirer une interprétation valable. Il est impossible de démonter le mécanisme de causalité du phénomène étudié.

Pour détecter la colinéarité, nous nous appuyons sur des calculs simples.

Cohérence des signes. Première stratégie, très rudimentaire, nous vérifions, pour chaque variable explicative, que le signe du coefficient dans la régression est identique au signe de la corrélation individuelle avec la variable expliquée. S'il y a contradiction, cela indiquerait que certaines explicatives interfèrent dans la liaison directe entre l'exogène incriminée et l'endogène.

Pour calculer ces corrélations, nous insérons (par un glisser-déposer) dans le diagramme TANAGRA, en dessous de DEFINE STATUS 1, le composant LINEAR CORRELATION (onglet STATISTICS). Son paramétrage par défaut lui permet de calculer la corrélation entre la variable TARGET et chaque variable INPUT.

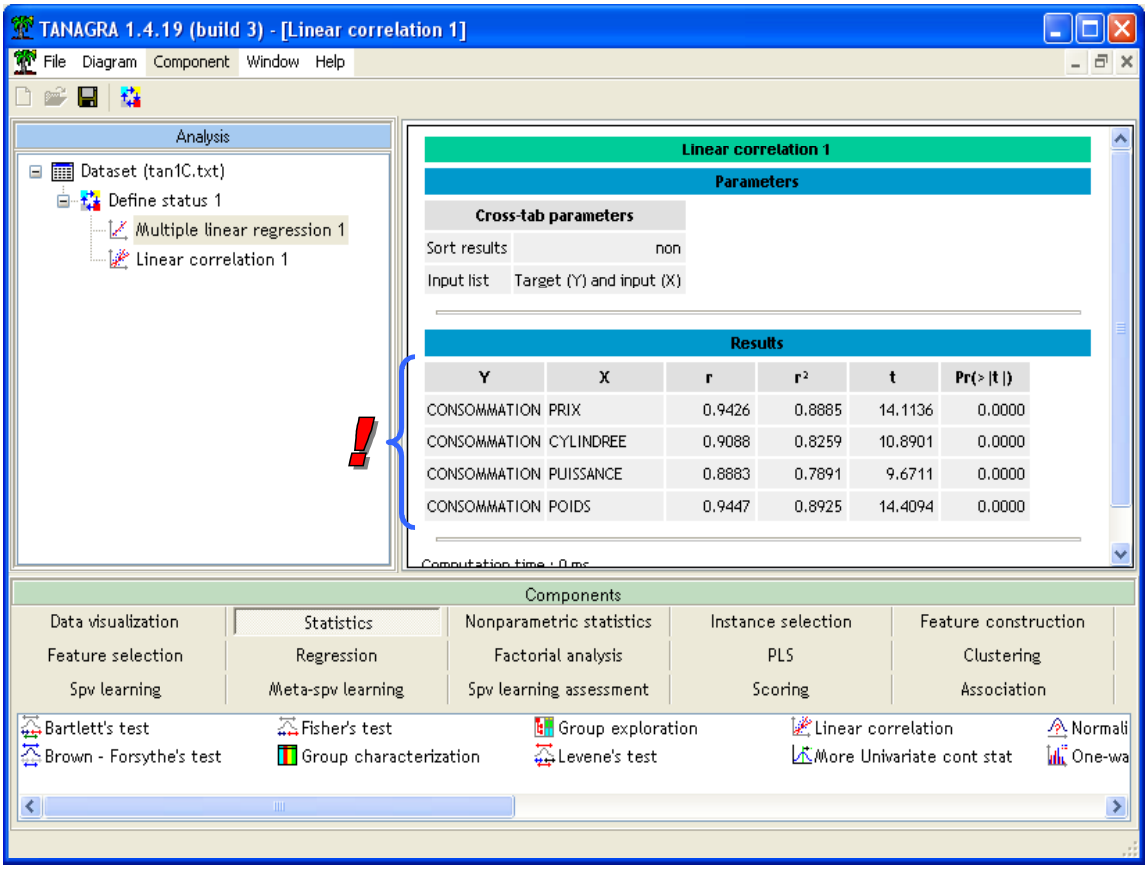


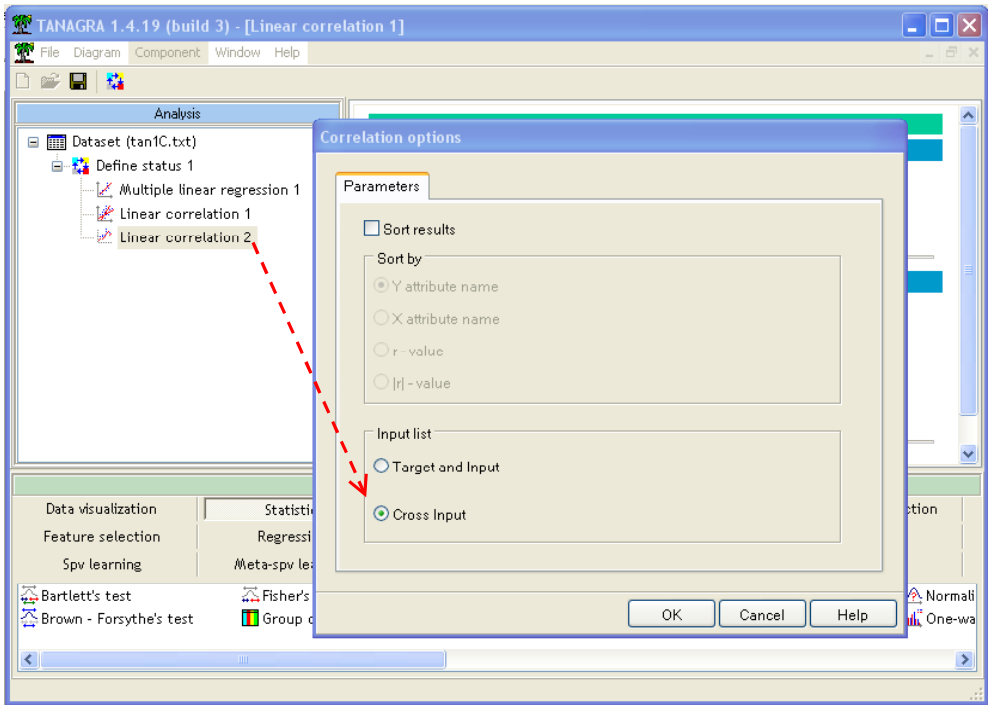
Figure 3 - Corrélation individuelle entre l'endogène et chaque explicative

Toutes les explicatives sont individuellement corrélées avec la CONSOMMATION (≥ 0.8883). Nous constatons également qu'il y a vraisemblablement un problème au niveau de la variable PUISSANCE. Elle est positivement corrélée avec l'endogène, pourtant le signe du coefficient de la régression est négatif. Une augmentation de la puissance entraînerait une baisse de la consommation ? Ce n'est pas très crédible. Il est vraisemblable qu'elle soit en (funeste) compétition avec une autre variable dans la régression.

Test de Klein. Il s'agit de calculer le carré de la corrélation entre chaque exogène. Si pour certains couples de variables, la valeur de l'indicateur est *proche du (ou supérieur au)* coefficient de détermination de la régression, il y a vraisemblablement un problème de colinéarité. L'intérêt, en

plus de détecter un éventuel problème dans la régression, est de pouvoir spécifier les variables redondantes qui peuvent mutuellement se nuire dans les calculs.

Nous plaçons de nouveau le composant LINEAR CORRELATION à la suite de DEFINE STATUS 1. Mais cette fois-ci, nous activons le menu PARAMETERS afin que les calculs portent sur le croisement des VARIABLES INPUT (option CROSS INPUT dans la section INPUT LIST).



Après avoir validé le paramétrage, nous cliquons sur le menu VIEW pour accéder aux résultats. Nous constatons que les variables exogènes sont toutes fortement corrélées entre elles.

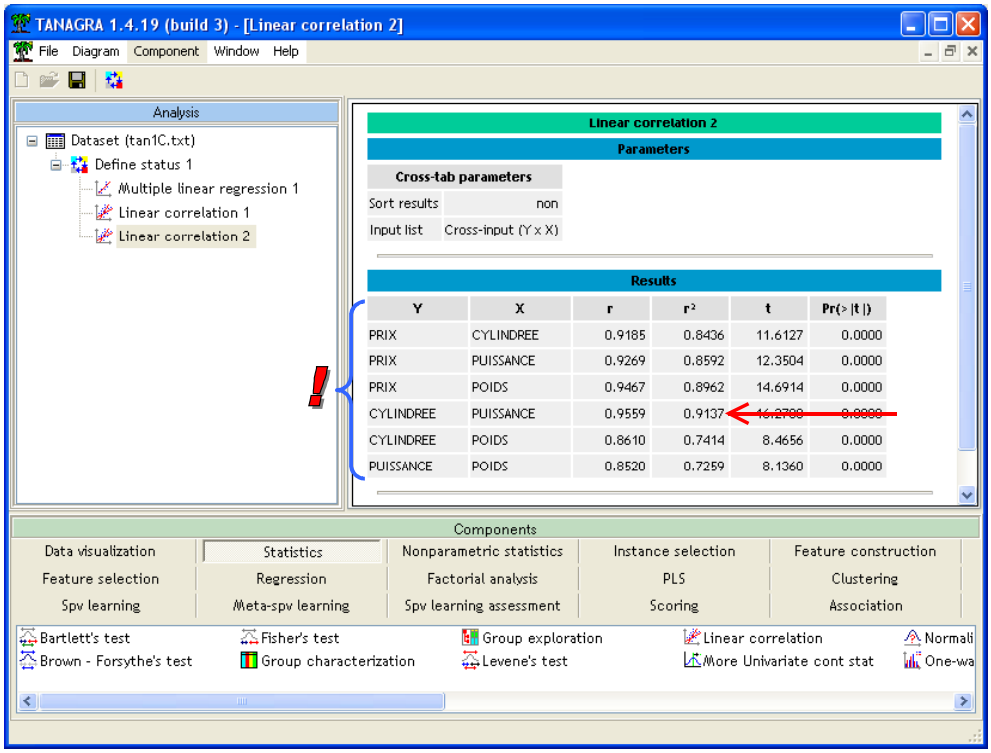


Figure 4 - Corrélations croisées entre les variables explicatives

C'est le cas en particulier des variables PUISSANCE et CYLINDREE où le carré de la corrélation r^2 (0.9137) se rapproche dangereusement du coefficient de détermination de la régression (0.9295).

Tous ces éléments sont autant d'indices qui laissent à penser qu'il y a un problème certain de colinéarité dans notre étude. Nous devons adopter une stratégie appropriée si nous voulons obtenir des résultats exploitables.

Sélection de variables

Même en l'absence de colinéarité des exogènes, réduire la dimensionnalité du problème étudié est toujours bénéfique. Discerner parmi les variables exogènes candidates, celles qui sont pertinentes pour prédire/expliquer les valeurs prises par l'endogène fait partie intégrante de la démarche de modélisation.

Nous allons mettre en œuvre une stratégie de recherche par avant (forward). Concrètement, l'ensemble de départ des variables sélectionnées est vide. A chaque étape, nous détectons l'exogène la plus corrélée (en valeur absolue) avec l'endogène, *compte tenu des variables déjà sélectionnées*. Nous nous basons pour cela sur la notion de corrélation partielle. Si elle est significativement non nulle, la variable est ajoutée dans le pool et nous essayons de détecter la meilleure variable suivante, etc. Dans le cas contraire, elle est refusée, ce qui constitue la règle d'arrêt du processus de sélection.

Nous insérons le composant FORWARD ENTRY REGRESSION à la suite de DEFINE STATUS 1 dans le diagramme TANAGRA. Nous cliquons sur le menu VIEW pour lancer les calculs et accéder aux résultats.

En plus des résultats habituels de la régression, nous disposons du détail du mécanisme de sélection. Plusieurs points sont à retenir :

- Les variables élues sont POIDS et CYLINDREE.
- Par rapport à la régression initiale, malgré la suppression de deux variables explicatives, la proportion de variance expliquée demeure très bonne avec un $R^2 = 0.9277$ (contre $R^2 = 0.9295$ pour le modèle à 4 variables).
- Introduites dans la régression, POIDS et CYLINDREE s'avèrent très significatives pour expliquer la consommation.
- Elles sont toutes deux liées positivement avec l'endogène c.-à-d. plus la voiture est lourde avec un gros moteur, plus sa consommation est élevée. Ce qui semble tout à fait acceptable.
- Les signes des coefficients sont en accord avec les signes des corrélations individuelles (Figure 3).
- Si nous nous référons au tableau des corrélations croisées (Figure 4), ces deux variables sont les moins corrélées du lot, avec un carré de la corrélation $r^2 = 0.7414$, bien en deçà du coefficient de détermination de la régression.
- Penchons nous maintenant sur les différentes étapes de la sélection (Figure 5) :
    - Au premier tour, la variable la plus corrélée avec l'endogène (en valeur absolue) est POIDS ($r = 0.9447$). Le F du test de nullité (F de Fisher est le carré du t de Student

lorsque le premier degré de liberté est 1) est égal à 207.63 avec un p-value < 0.0001. La liaison est très significative, la variable est introduite dans le modèle.

- Au tour suivant, la variable la plus corrélée avec la CONSOMMATION sachant les valeurs prises par POIDS (ou en enlevant l'effet POIDS) est CYLINDREE avec une corrélation, partielle cette fois-ci, de 0.5719, qui est également significative à 5% (p-value = 0.0118). Elle est également sélectionnée.

- Au 3ème tour, la variable la plus corrélée est PRIX (r-partiel = 0.1507), mais elle n'est pas significative à 5% (p-value = 0.4721). Nous ne la retenons pas, ce qui stoppe là le processus de sélection.

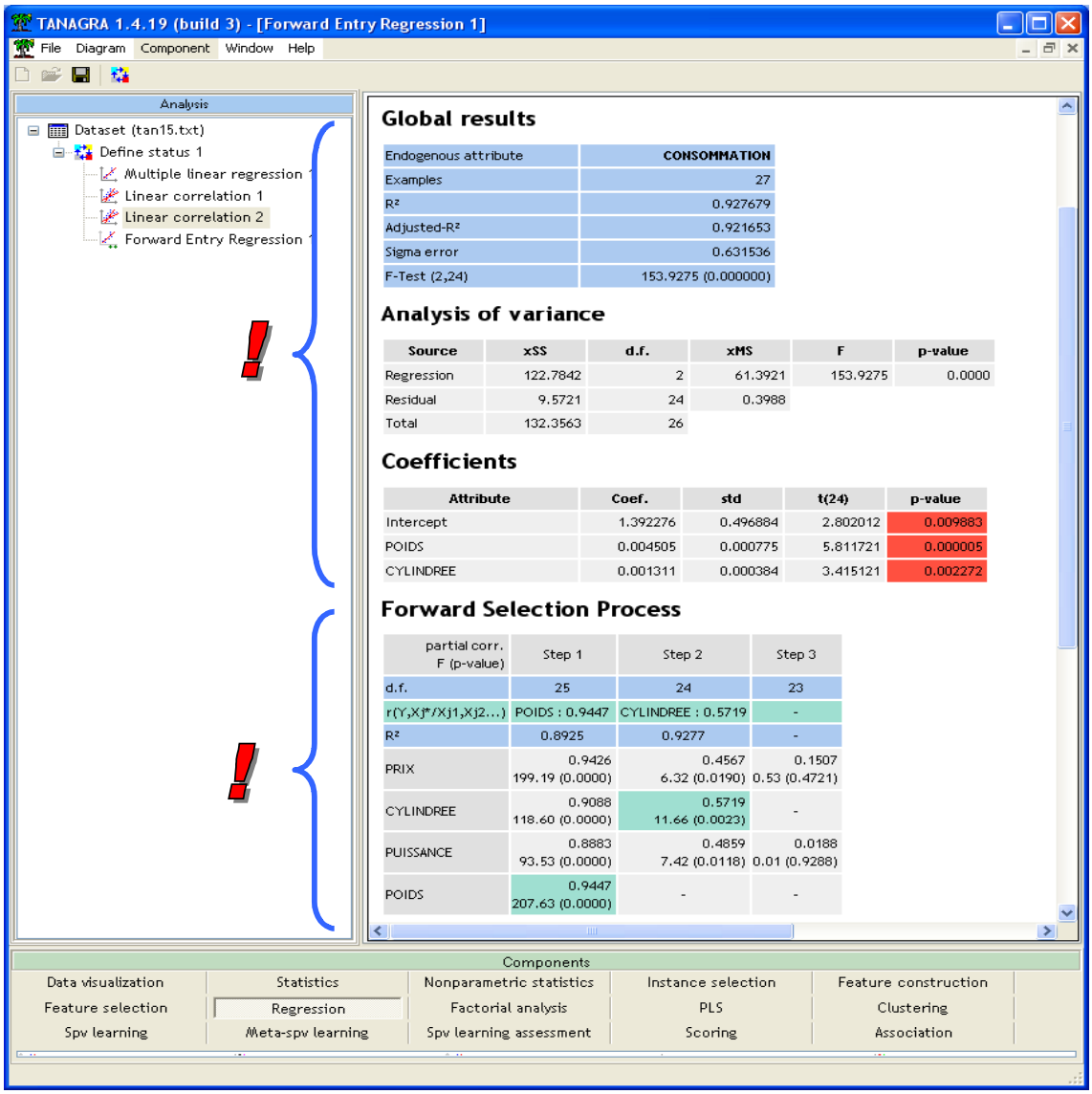


Figure 5 - Processus de sélection "forward"

Il existe d'autres techniques de sélection de variables (backward, stepwise, etc.). Elles sont plus ou moins performantes, selon les données et le contexte d'utilisation. En réalité, il ne faut pas trop se focaliser sur un ensemble *prétendument optimal* de variables explicatives pour la régression. Il est plus instructif de se pencher sur le rôle de chaque exogène.

Revenons sur la 2ème étape du processus, la variable CYLINDREE (r-partiel = 0.5719) est en compétition avec la variable PUISSANCE (r-partiel = 0.4859). Elle s'avère meilleure, elle est

sélectionnée, mais ce faisant nous excluons totalement la PUISSANCE dans l'explication de la consommation. Pourtant, son rôle n'est pas nul, loin de là. Si nous effectuons une régression avec les seules variables POIDS et PUISSANCE, nous obtiendrions un R^2 de 0.9179 (Figure 6), la variable PUISSANCE étant largement significative à 5% (p-value = 0.0118).

Global results

Endogenous attribute	**CONSOMMATION**
Examples	27
R^2	0.917912
Adjusted-R^2	0.911071
Sigma error	0.672833
F-Test (2,24)	134.1842 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	121.4914	2	60.7457	134.1842	0.0000
Residual	10.8649	24	0.4527		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	1.620097	0.560290	2.891532	0.008018
PUISSANCE	0.020937	0.007686	2.723896	0.011839
POIDS	0.004923	0.000802	6.137204	0.000002

Figure 6 - Régression avec les seules variables POIDS et PUISSANCE

Moralité de tout ceci : il faut utiliser avec précaution la sélection de variables. Certes, elle permet d'éliminer les variables inutiles, sans aucun lien avec la variable à prédire. Mais, elle élimine (masque) également les variables fortement liées avec l'endogène, qui peuvent être intéressantes pour l'interprétation, tout simplement parce qu'elles ont été numériquement dépassées par d'autres à une étape du processus de sélection.

Dans ce qui suit, nous nous penchons sur d'autres techniques qui, tout en luttant contre la colinéarité, essaient de restituer équitablement le rôle de chaque variable dans le processus d'explication de l'endogène.

Régression sur facteurs de l'ACP

L'ACP vise à produire des facteurs[3] (des variables synthétiques) combinaisons linéaires des variables originales en essayant de restituer au mieux les proximités entre les individus. L'intérêt est que les facteurs sont deux à deux indépendants et rangés par ordre décroissant d'importance : les plus informatifs sont placés en premier.

La régression sur facteurs consiste donc à : (1) Effectuer une ACP sur les exogènes ; (2) Utiliser certains (nous en discuterons plus loin) facteurs comme nouvelles variables explicatives de la régression. Ce faisant, nous éliminons totalement le problème de la colinéarité puisque les axes factoriels sont par définition orthogonaux ; (3) Reste alors à extrapoler les coefficients de la régression dans l'espace initial, celui des variables originelles.

[3] On parle également d'axes factoriels. Pour une description de l'ACP et de la lecture des résultats, nous conseillons le didacticiel TANAGRA - http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Acp.pdf

L'analyse en composantes principales (ACP)

Pour réaliser une ACP dans TANAGRA, nous insérons le composant PRINCIPAL COMPONENT ANALYSIS (onglet FACTORIAL ANALYSIS) en dessous de DEFINE STATUS 1. Nous lançons les calculs en cliquant sur le menu VIEW. Nous obtenons :

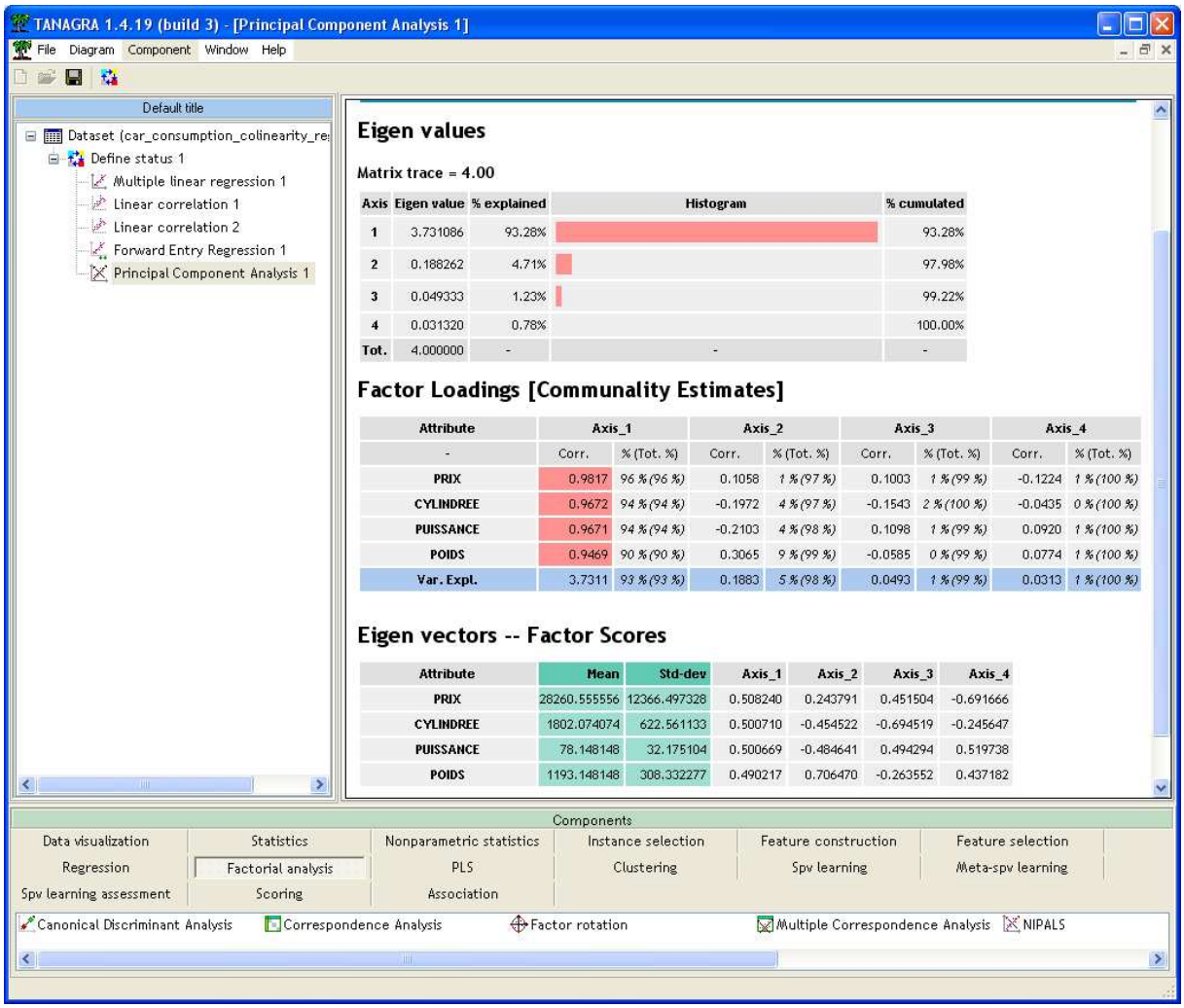


Figure 7 – Résultats de l'analyse en composantes principales

La fenêtre est subdivisée en 3 parties (nous simplifions à l'extrême la description des résultats dans ce didacticiel, notre objectif étant la régression) :

- EIGEN VALUES indique le pouvoir informationnel de chaque axe, on parle d'inertie expliquée. On note que 93.28% de la dispersion des points est restituée par le premier axe. La second n'en exprime que 4.71%. Les autres sont vraiment négligeables.

- FACTOR LOADINGS indique la corrélation de chaque variable avec les axes. On note par exemple que toutes les variables sont fortement corrélées avec le premier axe. C'est normal après coup, les véhicules avec un gros moteur sont, en général, puissants, lourds et chères. Sur le second axe, nous voyons s'opposer la puissance et le poids.

- EIGEN VECTORS - FACTOR SCORES fournissent les paramètres utilisés pour centrer et réduire les données lors des calculs (Mean et Std-dev), nous disposons alors des coefficients de projection sur chaque axe. Prenons le cas du premier véhicule de notre fichier (Daihatsu Core - Figure 1), sa coordonnée sur le premier axe est :

$$v_1 = 0.51 \times (\frac{11600 - 28260.6}{12366.5}) + 0.50 \times (\frac{846 - 1802.1}{622.6}) + 0.50 \times (\frac{32 - 78.1}{32.2}) + 0.49 \times (\frac{650 - 1193.1}{308.3})$$
$$= -3.035$$

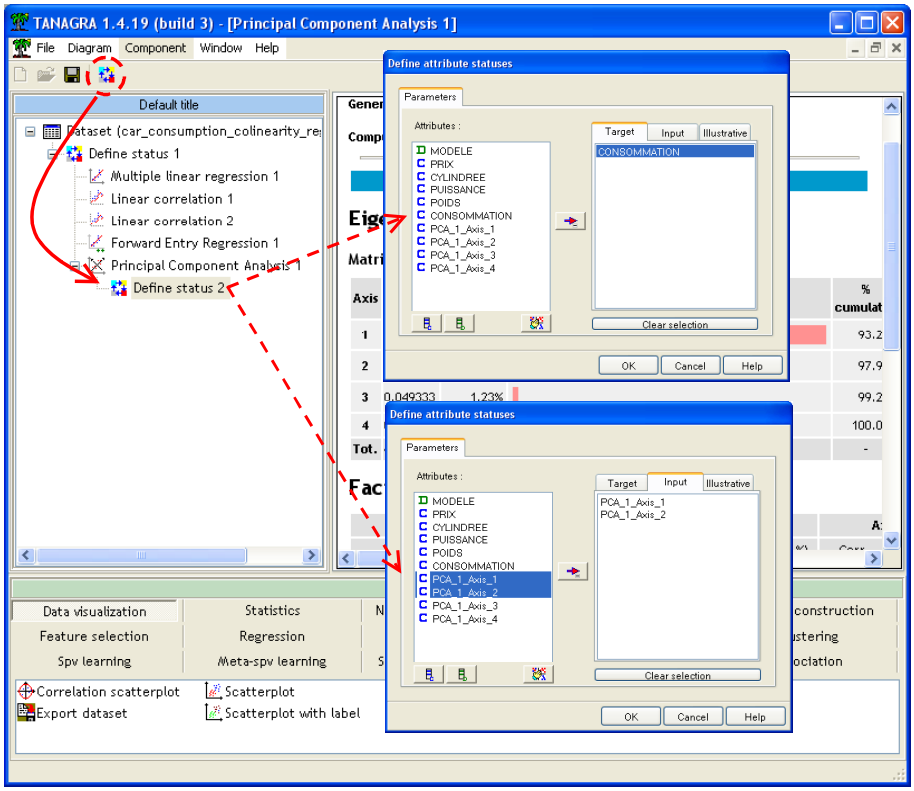
Il nous reste à effectuer la régression sur les axes factoriels.

Régression sur les axes factoriels

La question cruciale est « combien d'axes et lesquels doivent être utilisés dans la régression ? ». Le plus simple serait de tous les inclure et de laisser la régression choisir en observant la significativité des coefficients. Séduisante, cette solution présente une lacune forte. Il se peut que certains axes, très peu informatifs c.-à-d. avec un pourcentage d'inertie très faible, soient sélectionnés par la régression. Or ces facteurs sont par définition très instables, ils correspondent à des informations résiduelles dans la population. Positionner un individu par rapport à ces axes correspond à une disposition aléatoire dans un nouvel espace de description. La corrélation, trompeusement significative, constatée entre ces axes et la variable endogène a de fortes chances d'être simplement un artefact statistique.

Il nous faut donc préciser l'idée ci-dessus : nous sélectionnons les axes significatifs dans la régression, pour peu qu'ils soient porteurs d'informations, avec un pourcentage d'inertie *suffisamment* élevé. Notons un élément qui a son importance, les facteurs étant deux à deux orthogonaux, il suffit de lancer la régression sur les facteurs choisis puis d'éliminer ceux qui sont non significatifs.

Dans notre exemple, il paraît raisonnable de ne retenir que les deux premiers facteurs pour la régression. A cet effet, nous insérons un nouveau composant DEFINE STATUS après l'ACP, le plus simple toujours est de passer par le raccourci dans la barre d'outils. Nous plaçons en TARGET la variable CONSOMMATION, et en INPUT les facteurs PCA_1_Axis_1 et PCA_1_Axis_2.



Nous ajoutons à la suite le composant de régression (MULIPLE LINEAR REGRESSION). Nous observons les résultats.

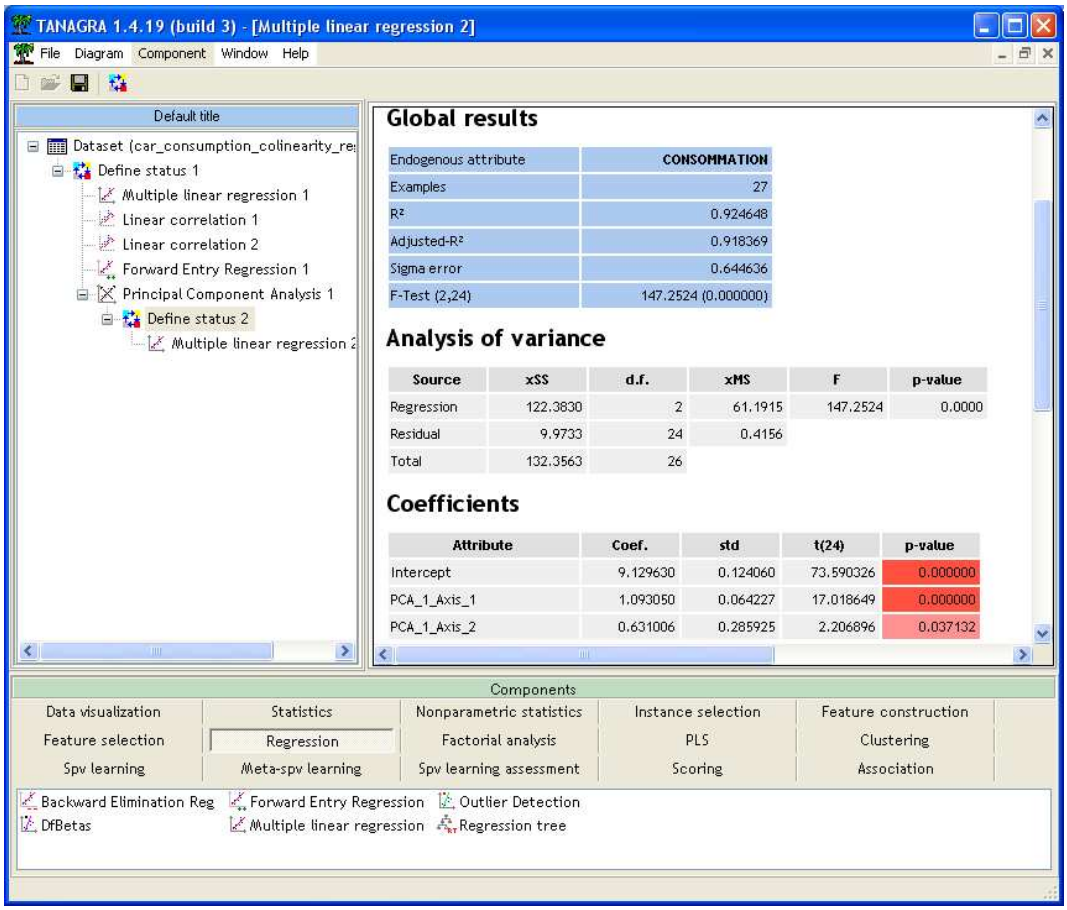


Figure 8 - Régression sur les 2 premiers axes factoriels

La régression est plutôt de bonne qualité puisque le R² est égal à 0.9246, tout à fait comparable à la régression après sélection des seules variables POIDS et CYLINDREE (Figure 5, R² = 0.9276). Les deux axes factoriels introduits sont significatifs à 5%. Si l'un des deux n'avait pas été significatif, nous l'aurions éliminé.

Notons que la régression avec les 4 facteurs a été tentée. Il s'est avéré que les facteurs 3 et 4 ne sont pas significatifs, ce qui conforte notre choix ci-dessus.

Calcul des coefficients de la régression dans l'espace initial

Si la procédure est satisfaisante numériquement, il reste une dernière étape. Les coefficients de la régression sur les facteurs, telles quelles, ne sont d'aucune utilité. Il nous faut impérativement re-calculer les coefficients sur les variables exogènes initiales pour les interpréter.

Pour ce faire, nous utilisons conjointement les résultats de l'ACP (Figure 7) et la régression sur les axes de l'ACP (Figure 8). Posons V1 et V2 les 2 premiers axes factoriels, m_x et σ_x la moyenne et l'écart type de la variable X. L'équation de régression sur facteurs s'écrit :

$$y = 9.13 + 1.09 \times V_1 + 0.63 \times V_2$$

$$= 9.13 + 1.09 \times \left[0.51 \times \left(\frac{prix - m_{prix}}{\sigma_{prix}} \right) + 0.50 \times \left(\frac{cylindree - m_{cylindree}}{\sigma_{cylindree}} \right) + \cdots \right]$$

$$+ 0.63 \left[0.24 \times \left(\frac{prix - m_{prix}}{\sigma_{prix}} \right) - 0.45 \times \left(\frac{cylindree - m_{cylindree}}{\sigma_{cylindree}} \right) + \cdots \right]$$

En développant cette expression, nous obtenons les coefficients suivants sur les variables centrées et réduites

$$\begin{aligned} y = & 9.13 \\ & + 0.7094 \times \left(\frac{prix - m_{prix}}{\sigma_{prix}} \right) \\ & + 0.2605 \times \left(\frac{cylindree - m_{cylindree}}{\sigma_{cylindree}} \right) \\ & + 0.2414 \times \left(\frac{puissance - m_{puissance}}{\sigma_{puissance}} \right) \\ & + 0.9816 \times \left(\frac{poids - m_{poids}}{\sigma_{poids}} \right) \end{aligned}$$

Restons un moment sur cette équation intermédiaire. Les coefficients de la régression sur les variables centrées et réduites, on parle de coefficients standardisés (standardized coefficient en anglais), sont comparables d'une variable à l'autre puisque nous nous affranchissons des unités. Il est possible d'évaluer l'importance relative des exogènes dans l'explication des valeurs de la consommation. **C'est à ce stade qu'intervient réellement l'interprétation des résultats dans la régression sur facteurs.** Nous constatons que le poids est le plus important, suivi du prix, cylindrée et puissance contribuent de la même manière pour expliquer la consommation. Nous comprenons mieux l'importance de ce résultat lorsque nous la comparons à la régression initiale (Figure 2) où POIDS éjectait les autres variables, CYLINDREE et PUISSANCE arborant des coefficients de signe opposé.

Un coefficient standardisé proche de zéro indique une variable peu pertinente dans la régression. Malheureusement, il n'existe pas de procédure simple pour tester la significativité d'un coefficient associé à une variable. Il faudrait passer par les procédures de ré-échantillonnage (ex. bootstrap) pour construire les intervalles de confiance.

En introduisant les valeurs de moyenne et écart type des variables affichées par l'ACP, il est possible de revenir à une équation de régression non-standardisée pour la prédiction.

$$y = 2.36954 + 0.00006 \times prix + 0.00042 \times cylindree + 0.00750 \times puissance + 0.00318 \times poids$$

TANAGRA ne fournit pas directement ces coefficients. Il faut effectuer une petite gymnastique qui consiste à copier les résultats de l'ACP et la régression sur facteurs dans un tableur ; puis, appliquer les formules ci-dessus pour obtenir les coefficients sur les variables initiales. Une feuille de calcul bien organisée permet de réaliser tout cela assez rapidement.

Régression PLS

La régression PLS vise également à élaborer des facteurs, résumés des variables initiales. A la différence que la méthode tient compte des informations de l'endogène pour élaborer les axes factoriels[4]. Ce faisant elle cumule un double avantage : (1) elle résiste bien à la colinéarité et fournit des coefficients qui permettent d'évaluer la contribution de chaque exogène dans l'explication des valeurs de la variable à prédire ; (2) les axes factoriels sont optimisés pour

[4] M. Tenenhaus, « La Régression PLS – Théorie et pratique », Technip, 1998.

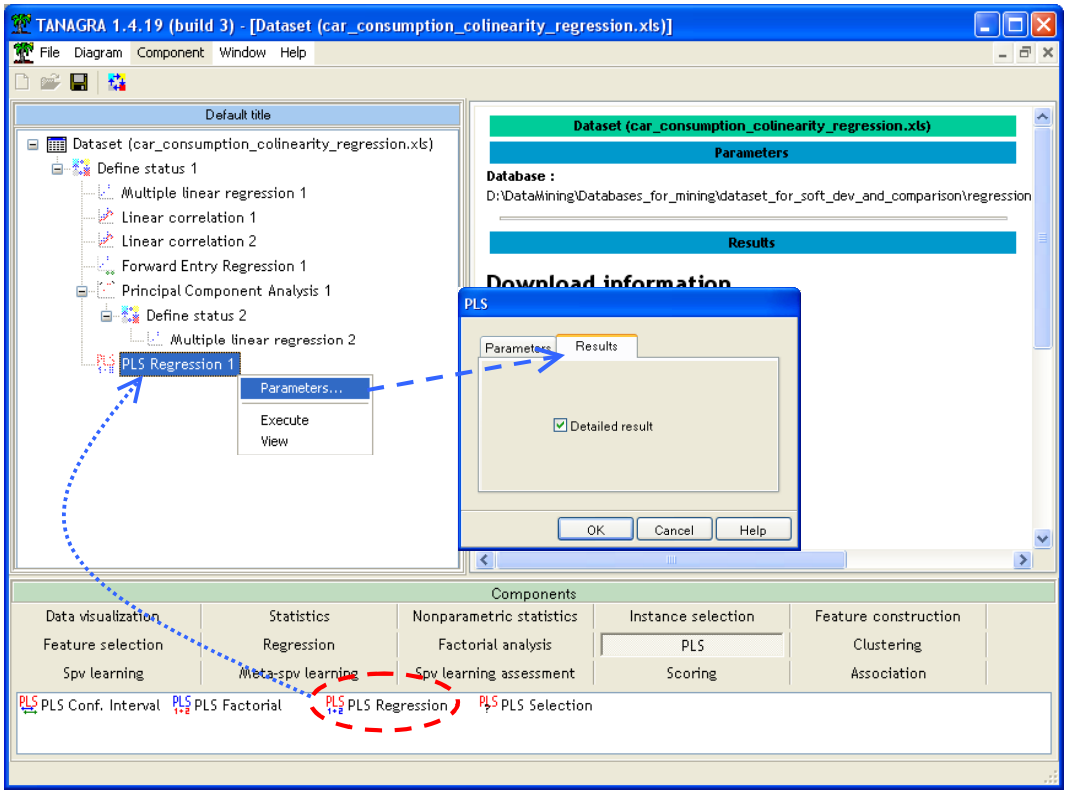
l'explication de Y, ils sont de significativité décroissante, nous n'avons plus besoin, comme dans la régression sur facteurs de l'ACP, de travailler en deux temps en choisissant d'abord les axes porteurs d'information, puis de vérifier leur efficacité dans la prédiction.

Nous conseillons l'ouvrage de M. Tenenhaus pour les détails de la régression PLS. L'article de Chavent et Patouille[5] paru dans la revue MODULAD est très intéressant également.

Lire les résultats de la régression PLS

Nous effectuer une régression PLS, nous insérons le composant PLS REGRESSION (onglet PLS) en dessous de DEFINE STATUS 1.

Au départ, la méthode produit L = MIN (5, nombre d'exogènes) axes factoriels. Dans notre exemple, nous produirons automatiquement 4 axes. Nous activons le menu PARAMETERS, puis l'option DETAILED RESULTS pour obtenir le détail des calculs.



Nous validons et nous cliquons sur VIEW, nous disposons de plusieurs niveaux de résultats.

Coefficients de la régression. Il s'agit des coefficients que nous utiliserons pour prédire la valeur d'une nouvelle observation (Figure 9). Nous constatons sans surprise que nous avons des coefficients identiques à ceux de la régression linéaire multiple (Figure 2), en effet nous avons utilisé toutes les informations linéairement exploitables de X pour expliquer les valeurs de Y (Nombre d'axes = nombre de variables).

[5] http://www-rocq.inria.fr/axis/modulad/archives/numero-30/chavent-30/chavent-30.pdf

Regression coefficients

X/Y	CONSOMMATION
PRIX	0.0000
CYLINDREE	0.0012
PUISSANCE	-0.0037
POIDS	0.0037
constant	1.8380

Figure 9 - Coefficients de la régression PLS - 4 axes

Part de variance expliquée des X (Redondance). Elle exprime la part de variance restituée par les axes factoriels concernant les variables exogènes. Lorsque le nombre d'axes est égal au nombre de variables, il est naturel que l'on obtienne $R^2 = 1$ (Figure 10, la dernière ligne du tableau). Nous constatons également que la part de variance expliquée par les axes diminue au fur et à mesure, elle est négligeable sur les 3ème et 4ème axes (resp. 0.0131 et 0.0078). Cela donne une indication : la bonne régression ne devrait retenir que les 2 premiers axes au maximum.

Si l'on devait faire un parallèle, il faudrait comparer ces valeurs avec ceux de l'ACP où on ne tenait pas compte de l'endogène pour élaborer les axes factoriels : nous restituons également 98% de l'information sur les deux premiers facteurs, à la différence que dans le cas de la régression PLS, les axes sont optimisés pour la prédiction.

R^2 coefficients and redundancy on inputs (X)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4
PRIX	0.9649 (0.9649)	0.0069 (0.9718)	0.0125 (0.9843)	0.0157 (1.0000)
CYLINDREE	0.9332 (0.9332)	0.0316 (0.9648)	0.0337 (0.9985)	0.0015 (1.0000)
PUISSANCE	0.9324 (0.9324)	0.0535 (0.9858)	0.0063 (0.9921)	0.0079 (1.0000)
POIDS	0.9005 (0.9005)	0.0932 (0.9937)	0.0001 (0.9938)	0.0062 (1.0000)
Redundancy	0.9327 (0.9327)	0.0463 (0.9790)	0.0131 (0.9922)	0.0078 (1.0000)

Figure 10 - Redondance sur les X - Régression PLS

Autre information très importante de ce tableau de résultats, nous disposons du détail pour chaque variable. Nous observons que, dès le 1er axe, les exogènes sont bien restituées ($R^2 \geq 0.9005$ quelle que soit la variable), aucune variable n'est négligée dans la régression.

Part de variance expliquée des Y (Redondance). L'objectif étant l'explication des valeurs de Y en fonction des X, ce tableau est crucial. Il indique la qualité de la prédiction (Figure 11). Encore une fois, si nous conservons tous les axes, nous aurions un $R^2 = 0.9295$, égal à celui de la régression multiple.

En inspectant en détail le tableau, nous constatons que la qualité de la régression s'améliore de manière anecdotique à partir du 2ème axe, le R^2 passe de 0.9270 à 0.9294 pour le 3ème axe. Le gain est nul si l'on passe au 4ème axe. Ces éléments, mis en relation avec le tableau des redondances sur les exogènes, laisse à penser que le choix des 2 premiers axes serait amplement suffisant.

R^2 coefficients and redundancy on targets (Y)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4
CONSOMMATION	0.9110 (0.9110)	0.0160 (0.9270)	0.0025 (0.9295)	0.0000 (0.9295)
Redundancy	0.9110 (0.9110)	0.0160 (0.9270)	0.0025 (0.9295)	0.0000 (0.9295)

Figure 11 - Redondance sur les Y - Régression PLS

Enfin, si nous avions voulu expliquer simultanément plusieurs endogènes, c'est possible avec la régression PLS, nous aurions le détail pour chaque variable dans le tableau des redondances en Y.

Tableau des VIP (Variable Importance in Projection). Dernier tableau très important pour l'interprétation, celui des VIP (Figure 12). Il indique la contribution d'une exogène pour l'explication de l'endogène à travers la composante n°k (Tenenhaus, page 139).

Variable Importance in the Projection

Attribute	Axis_1	Axis_2	Axis_3	Axis_4
PRIX	1.0230	1.0144	1.0143	1.0143
CYLINDREE	0.9863	0.9799	0.9822	0.9822
PUISSANCE	0.9641	0.9698	0.9691	0.9691
POIDS	1.0253	1.0345	1.0331	1.0331

Figure 12 - Tableau des VIP - Régression PLS

Contrairement à ce que laissait entendre la régression linéaire multiple, toutes les variables participent de manière plus ou moins équivalente pour l'explication de la consommation, quel que soit l'axe étudié. POIDS et PRIX semblent se démarquer très légèrement (VIP > 1), CYLINDREE et PUISSANCE sont d'égales contributions.

Nous retrouvons là des commentaires que nous avions déjà formulés concernant la régression sur les facteurs de l'ACP.

Choisir le nombre adéquat d'axes

Il reste un problème crucial, le choix du nombre d'axes. Nous avions commencé à émettre plusieurs remarques tendant à valider au maximum les 2 premiers facteurs. Mais ce choix repose avant tout sur des considérations plus ou moins subjectives, à la recherche d'un improbable « coude » dans l'évolution des redondances, tant sur les exogènes que sur l'endogène.

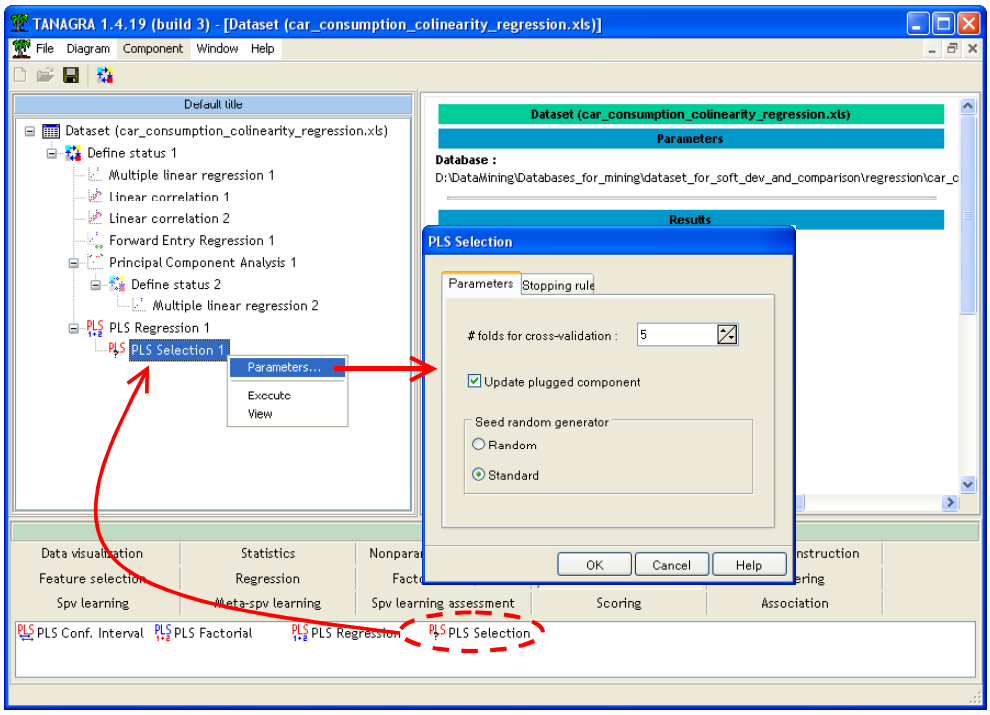
Cette première analyse peut être complétée par une procédure numérique basée uniquement sur les performances en prédiction.

Le composant PLS SELECTION de TANAGRA s'appuie sur le PRESS (Predicted Residual Sum of Squares) pour déterminer le nombre « optimal » d'axes de la PLS. Rappelons que le PRESS est formé à partir de la somme du carré des écarts, pour chaque observation **i**, entre la valeur de l'endogène et la valeur prédite de l'endogène lorsque cette observation n'a pas participé à la construction du modèle. Il s'agit bien d'une erreur de prédiction puisque l'observation **i** est utilisée comme donnée supplémentaire (Tenenhaus, page 77). Cet indicateur est autrement plus fiable que la somme des carrés des résidus classiques (RSS - Residual Sum of Squares) où l'observation est à la fois juge (on compare la prédiction de sa valeur) et partie (il a participé à la construction du modèle).

Un deuxième critère dérivé du PRESS et du RSS peut être utilisé pour déterminer le bon nombre de facteurs. Le Q^2 met en compétition le RSS de la solution comportant (h-1) axes avec le PRESS de

celle qui en comporte h. L'idée est de mesurer le *véritable* apport marginal de la h^{ème} composante au pouvoir prédictif du modèle (Tenenhaus, page 138).

Nous plaçons le composant PLS SELECTION (onglet PLS) dans le diagramme. Nous activons le paramétrage en cliquant sur le menu PARAMETERS.

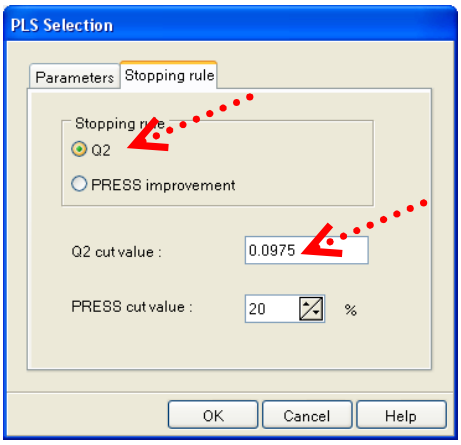


Dans le premier onglet PARAMETERS :

- Nous avons la possibilité de spécifier le nombre de subdivisions des données pour le calcul du PRESS. La valeur par défaut est 5 c.-à-d. les données sont divisées en 5 blocs, 4/5 pour la construction du modèle (en apprentissage), 1/5 est utilisé comme individus supplémentaires (en test) pour calculer les écarts de prédiction. Il s'agit de validation croisée, on opère une rotation des blocs de manière à ce qu'un bloc soit utilisé une fois comme données de test.

- « Update plugged component », si elle est cochée, indique que le composant précédent dans le diagramme, en l'occurrence PLS Regression 1, est automatiquement recalculé avec le nombre d'axes retenu par la procédure de sélection.

- Enfin, « Seed Random Generator » permet de spécifier le mode de fonctionnement du générateur de nombres aléatoires. Avec l'option par défaut « Standard », le composant reproduit exactement les mêmes séquences à chaque exécution.

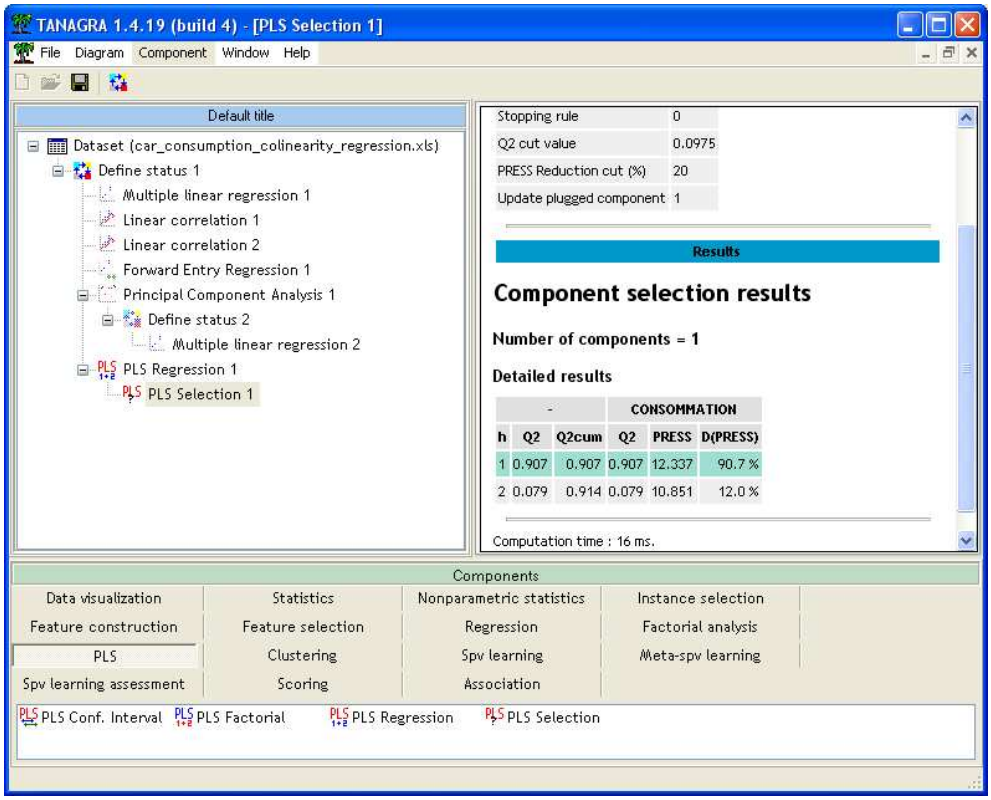


Dans le second onglet STOPPING RULE, nous pouvons spécifier le mode de détermination du nombre optimal d'axes :

- Nous pouvons utiliser le Q^2. Dans ce cas nous devons indiquer la valeur seuil qui permet de stopper la recherche. Par défaut, on rajoute un axe tant que Q2 > 0.05.
- Nous pouvons également utiliser directement le PRESS. Dans ce cas, nous indiquons la valeur limite de la décroissance du PRESS qui permet d'ajouter un axe. La valeur par défaut est 20% c.-à-d. si diminution du PRESS n'est pas meilleure que 20% lors du rajout d'un axe, ce dernier est refusé.

Dans notre exemple, nous choisissons le critère Q^2 et nous fixons comme valeur seuil 0.0975 conformément à ce qui est indiqué dans notre ouvrage de référence (Tenenhaus, page 138 ; qui cite la règle par défaut d'un logiciel commercial).

A l'exécution (menu VIEW), nous obtenons les résultats suivants.



Lorsque la régression PLS n'utilise qu'un seul axe, le Q^2 est égal à 0.907 avec un PRESS de 12.337 ; avec le second axe, le PRESS s'améliore et passe à 10.851, mais trop faiblement semble-t-il pour qu'il soit réellement significatif puisque le Q^2 lui est de 0.079, inférieur à 0.0975 que nous avons choisi comme valeur seuil.

Nous nous en tiendrons donc à une régression PLS avec un seul facteur, le composant PLS REGRESSION 1 a été automatiquement recalculé, nous accédons aux coefficients en activant le menu VIEW de PLS REGRESSION 1 (Figure 13). L'équation de régression s'écrit :

$$y = 2.835123 + 0.000045 \times prix + 0.000867 \times cylindree + 0.016397 \times puissance + 0.001820 \times poids$$

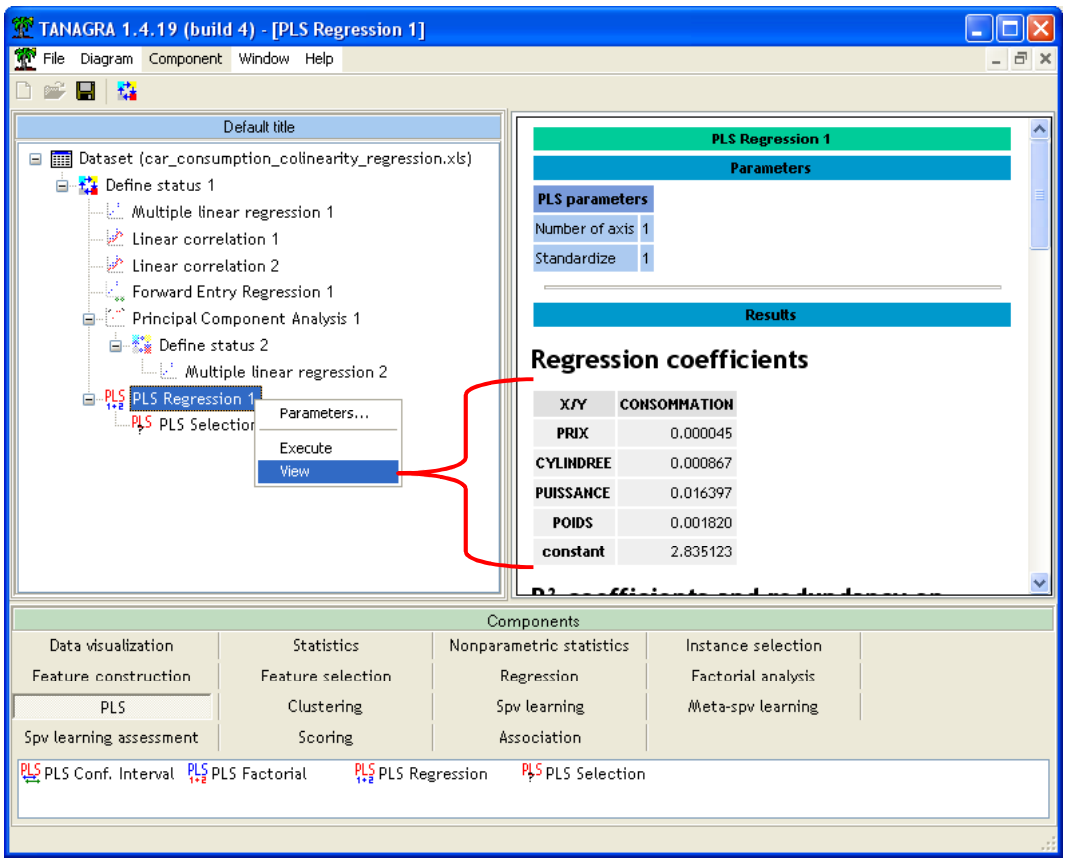


Figure 13 - Coefficients de l'équation de régression PLS

Pour bien situer les différences, nous récapitulons les paramètres estimés selon la méthode :

Variable	Régression linéaire multiple	Régression sur facteurs de l'ACP	Régression PLS
Constante	1.8380	2.36954	2.835123
Prix	0.0000	0.00006	0.000045
Cylindrée	0.0012	0.00042	0.000867
Puissance	-0.0037	0.00750	0.0016397
Poids	0.0037	0.00318	0.001820

La régression sur facteurs de l'ACP et la régression PLS proposent des équations cohérentes au niveau du signe des coefficients. La régression linéaire multiple, on le sait pourquoi maintenant, est sensiblement différente, la plus grosse carence étant le signe de PUISSANCE, en opposition avec les connaissances du domaine.

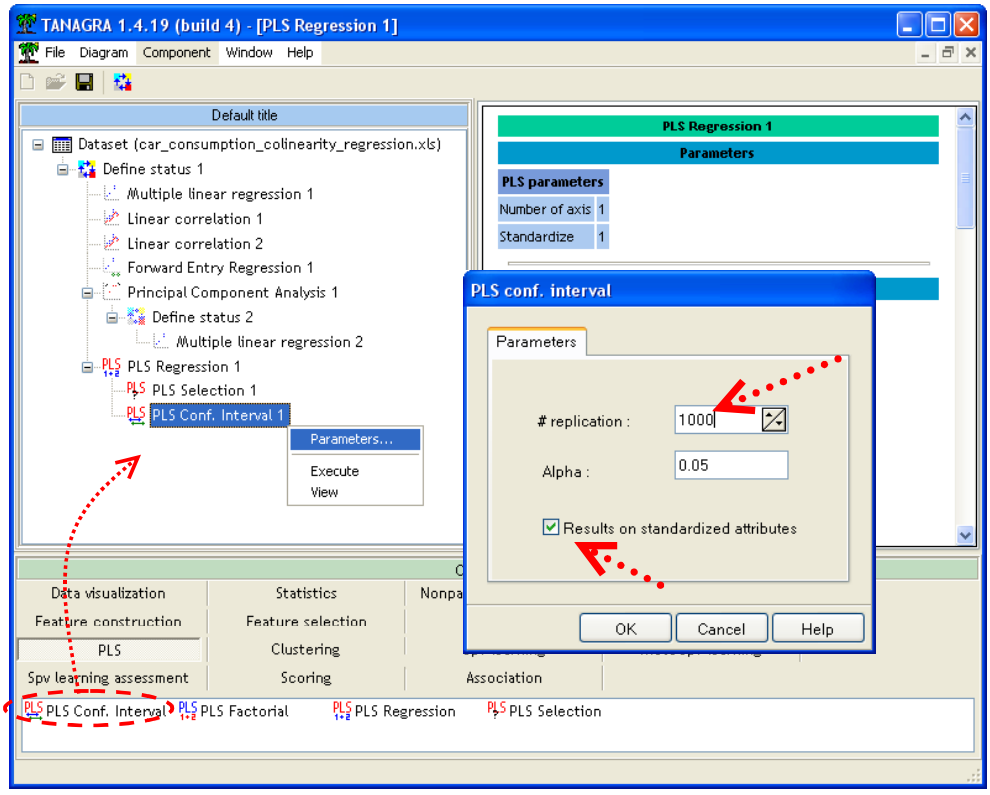
Expertiser les variables

Nous avons obtenu les coefficients de la régression PLS. Dans bien des cas, nous sommes également intéressés par leur intervalle de variation, ne serait-ce que pour apprécier leur stabilité. De même, nous pourrions en déduire des informations sur leur significativité : si l'intervalle contient la valeur 0 à un niveau de confiance donné, on peut légitimement se demander si la variable est réellement pertinente dans la régression.

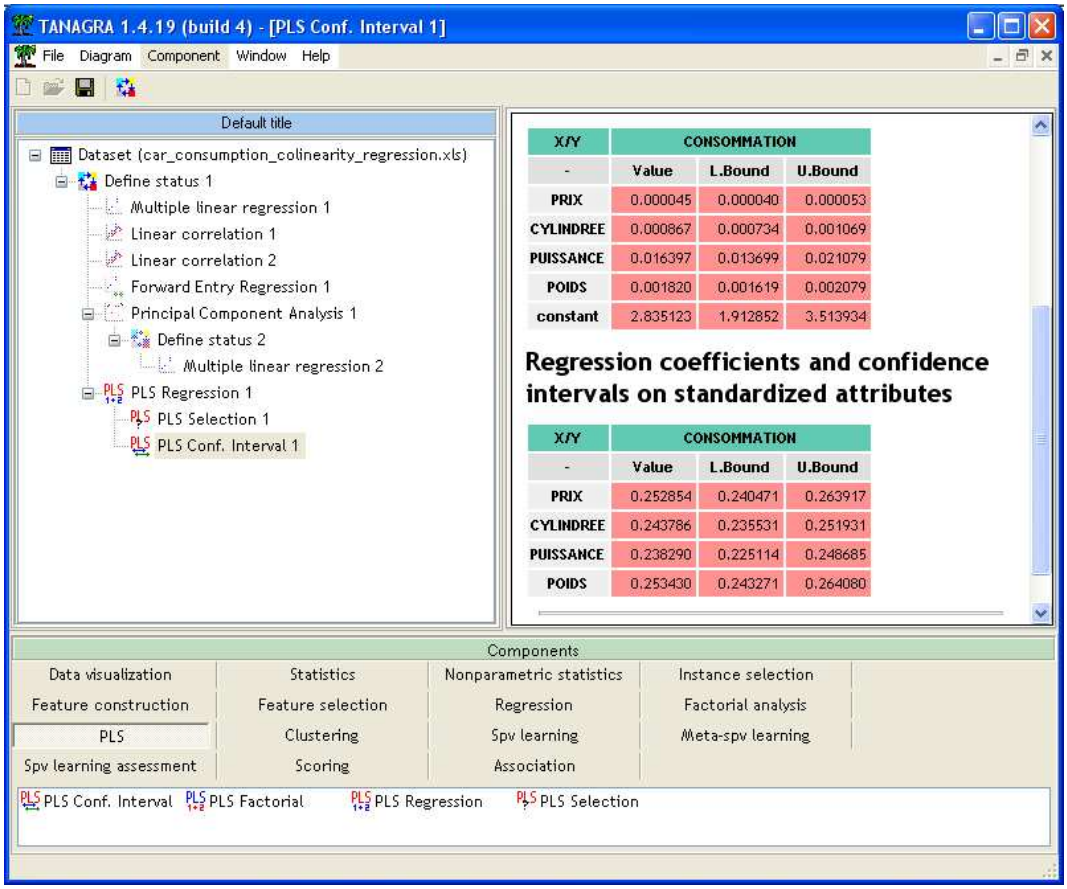
Le composant PLS CONF INTERVAL de TANAGRA calcule l'intervalle de variation des coefficients en utilisant une procédure bootstrap. La technique est simple, l'opération suivante est réitérée K fois : nous effectuons un tirage avec remise de n observations (n étant la taille du fichier), les coefficients de la régression sont calculés et collectés. A partir de ces données, nous pouvons construire la distribution empirique de chaque coefficient. Pour un niveau de confiance $1-\alpha$, nous en déduisons les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ qui correspondent aux bornes basses et hautes de l'intervalle de variation.

Nous plaçons le composant PLS CONF INTERVAL (onglet PLS) à la suite de PLS REGRESSION 1. Cela veut dire qu'il va utiliser les paramètres de ce dernier, le nombre d'axes notamment, pour ses calculs.

Nous activons le menu PARAMETERS, nous aimerions que les coefficients standardisés de la régression soient également affichés, que le nombre d'itérations bootstrap K=1000, le niveau de confiance étant 1-0.05 = 95%.



Nous lançons les calculs en cliquant sur le menu contextuel VIEW. Nous observons, si nous nous référons aux coefficients standardisés, que les variables ont toutes la même importance dans l'explication de l'endogène. Les coefficients présentent des valeurs très similaires.



De plus, ils sont éminemment stables, les intervalles de variation sont étroits, ce qui est remarquable compte tenu du faible effectif de notre fichier de données. Il faut voir une conséquence du nombre d'axes restreint que nous avons choisi, assurant une grande stabilité des résultats.

C'est là une des qualités décisives de la régression PLS. En jouant sur le nombre de facteurs, nous pouvons lisser/guider l'apprentissage et influer sur les caractéristiques du modèle. Nous n'intégrons que les informations *essentielles* des exogènes pour la prédiction de l'endogène.

Conclusion

Dans ce didacticiel, nous avons montré comment, avec TANAGRA, mettre en œuvre différentes stratégies de lutte contre la colinéarité dans une analyse de régression.

Les résultats peuvent diverger d'une technique à l'autre. Cela est normal dans la mesure où les approches ne sont pas les mêmes. Le plus important étant de lire convenablement les sorties des logiciels pour en tirer les conclusions adéquates.

Le cas de la variable PUISSANCE est édifiant quant à l'intérêt d'utiliser conjointement plusieurs techniques. Elle est fortement redondante avec la cylindrée, l'éjecter purement et simplement de

notre étude, comme nous y incitait la sélection de variables, masquerait une partie de la connaissance que nous pourrions extraire de ces données. En passant à la régression sur facteurs et à la régression PLS, le rôle important de cette variable dans l'explication de la consommation apparaît clairement.

Néanmoins, quelle que soit la qualité d'une technique statistique, rien ne remplace l'expertise humaine. Dans notre problème de consommation de véhicules, il est évident que le prix ne peut pas être une variable explicative de la consommation, même si nous comprenons aisément qu'elles soient *d'une certaine manière* liées. Aucune des approches n'a pu mettre en évidence ce non-sens.