1 Objectif

Comparer les résultats de la Régression PLS de Tanagra avec les autres logiciels, commerciaux (SIMCA-P, SPAD, SAS) ou gratuits (R).

Se comparer aux autres est toujours une bonne manière de faire avancer un logiciel.

Pour valider les implémentations. C'est un point essentiel. Bien que l'on s'appuie sur les mêmes références bibliographiques, que l'on met en place les mêmes algorithmes, les choix de programmation ne sont pas anodins (la gestion des conditions de convergence par exemple). Une manière simple de valider l'implémentation est, outre la documentation des algorithmes utilisés et la publication du code, de voir ce qui se passe avec les autres outils. Il est toujours perturbant de se rendre compte que deux logiciels scientifiques, censés réaliser les mêmes calculs, produisent des résultats plus ou moins différents. Il faut pouvoir (se) l'expliquer.

Pour améliorer la présentation de résultats. Il y a certains standards à respecter dans la production des rapports, consensus initié par les ouvrages de référence et/ou le(s) logiciel(s) leader(s) dans le domaine. Les utilisateurs ont besoin de repères. Certains ratios doivent être présentés d'une certaine manière. Il n'y a pas beaucoup de place pour la poésie là-dedans.

Notre implémentation de la Régression PLS repose essentiellement sur l'ouvrage de M. Tenenhaus (1998)[1] qui, lui même, fait beaucoup référence au logiciel SIMCA-P[2]. Profitant de l'accès à une version gratuite sur le site de l'éditeur (version 11 - limitée dans le temps), nous avons voulu comparer nos résultats sur un jeu de données. Nous avons étendu cette comparaison à d'autres logiciels.

Nous effectuons systématiquement ce travail de vérification lors de la programmation des composants. Il nous a semblé intéressant de l'intégrer dans un didacticiel. D'autant plus que nous constaterons, non sans surprises, que les résultats peuvent être (apparemment) très différents d'un logiciel à l'autre en ce qui concerne la Régression PLS.

Cette étude nous a emmené à introduire un nouveau composant dans Tanagra 1.4.24 (**PLSR**). Il intègre dans un cadre unique les composants PLS FACTORIAL, qui produit les scores factoriels, et PLS REGRESSION, qui effectue les prédictions. De plus, le mode de présentation des résultats, les noms des tableaux entre autres, est aligné sur les références anglo-saxonnes. Pour cela, nous nous sommes beaucoup appuyés sur les documents disponibles sur le site web de SIMCA-P (manuel de référence et tutoriel), et sur la description de D. Garson[3] qui, inlassablement, produit toujours des documents d'une qualité extraordinaire.

Enfin, ce didacticiel permet d'approfondir la lecture et l'interprétation des résultats de la régression PLS. Le précédent était peut-être un peu trop laconique (http://tutoriels-data-mining.blogspot.com/2008/04/rgression-pls.html).

[1] M. Tenenhaus, « La régression PLS – Théorie et Pratique », Technip, 1998.

[2] SIMCA-P for Multivariate Data Analysis. http://www.umetrics.com/default.asp/pagename/software_simcap/c/3

[3] D. Garson, « Partial Least Squares Regression », from *Statnotes: Topics in Multivariate Analysis.* Retrieved 05/18/2008 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

2 Données

Nous utilisons le fichier CARS_PLS_REGRESSION.XLS (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cars_pls_regression.xls).

Numéro	diesel	twodoors	sportsstyle	wheelbase	length	width	height	curbweight	enginesize	horsepower	horse_per_v	conscity	price	symboling
1	0	1	0	97	172	66	56	2209	109	85	0.0385	8.7	7975	2
2	0	0	0	100	177	66	54	2337	109	102	0.0436	9.8	13950	2
3	0	0	0	116	203	72	57	3740	234	155	0.0414	14.7	34184	-1
4	0	1	1	103	184	68	52	3016	171	161	0.0534	12.4	15998	3
5	0	0	0	101	177	65	54	2765	164	121	0.0438	11.2	21105	0
6	0	1	0	90	169	65	52	2756	194	207	0.0751	13.8	34028	3
7	1	0	0	105	175	66	54	2700	134	72	0.0267	7.6	18344	0
8	0	0	0	108	187	68	57	3020	120	97	0.0321	12.4	11900	0
9	0	0	1	94	157	64	51	1967	90	68	0.0346	7.6	6229	1
10	0	1	0	95	169	64	53	2265	98	112	0.0494	9.0	9298	1
11	1	0	0	96	166	64	53	2275	110	56	0.0246	6.9	7898	0
12	0	1	0	100	177	66	53	2507	136	110	0.0439	12.4	15250	2
13	0	1	1	94	157	64	51	1876	90	68	0.0362	6.4	5572	1
14	0	0	0	95	170	64	54	2024	97	69	0.0341	7.6	7349	1
15	0	1	1	95	171	66	52	2823	152	154	0.0546	12.4	16500	1
16	0	0	0	103	175	65	60	2535	122	88	0.0347	9.8	8921	-1
17	0	0	0	113	200	70	53	4066	258	176	0.0433	15.7	32250	0
18	0	0	0	95	165	64	55	1938	97	69	0.0356	7.6	6849	1
19	1	0	0	97	172	66	56	2319	97	68	0.0293	6.4	9495	2
20	0	0	0	97	172	66	56	2275	109	85	0.0374	8.7	8495	2

Figure 1 - Tableau de données : en vert les descripteurs, en bleu les variables cibles

L'objectif de l'étude est d'expliquer les indicateurs de coûts des véhicules (PRICE : coût à l'achat ; CONSCITY : consommation en ville ; SYMBOLING : étiquetage des assureurs qui détermine le montant de la prime d'assurance) à l'aide de leurs caractéristiques (motorisation, etc.).

3 La régression PLS

Rappelons succinctement le principe de la régression PLS. Elle a pour objectif d'expliquer un ensemble de variables cibles Y (variables à expliquer, variables à prédire) à partir d'un ensemble de variables explicatives X (les descripteurs, variables prédictives, variables explicatives). Pour ce faire, les descripteurs sont résumés en une série de facteurs t_h (axes factoriels, variables latentes, scores X) deux à deux orthogonaux. A la différence de l'analyse en composantes principales, ces facteurs sont construits de manière à expliquer le mieux Y. De la même manière, les variables cibles sont résumés dans une série de composantes u_h (scores Y). La contrainte d'orthogonalité ne s'applique pas à ces dernières. Elles seront surtout utiles pour l'interprétation.

Durant le processus de modélisation, la régression PLS va construire les séries de facteurs (u_h , t_h) de manière à ce que leur covariance soit maximum. Le nombre de facteurs ne peut pas excéder le nombre de variables explicatives.

4 Régression PLS avec TANAGRA et SIMCA-P

Dans cette partie, nous détaillons les rapports de **TANAGRA**. Nous mettrons systématiquement en miroir les sorties de **SIMCA-P**. Disons le tout de suite, **nous observons une correspondance exacte des résultats**, à croire que nous nous appuyons non seulement sur les mêmes formules mais également sur des choix techniques similaires (précision des calculs, etc.).

Concernant Tanagra, l'algorithme implémenté respecte scrupuleusement la description proposée dans l'ouvrage de Tenenhaus, page 128, « Algorithme de régression PLS2 classique ». Les initiés

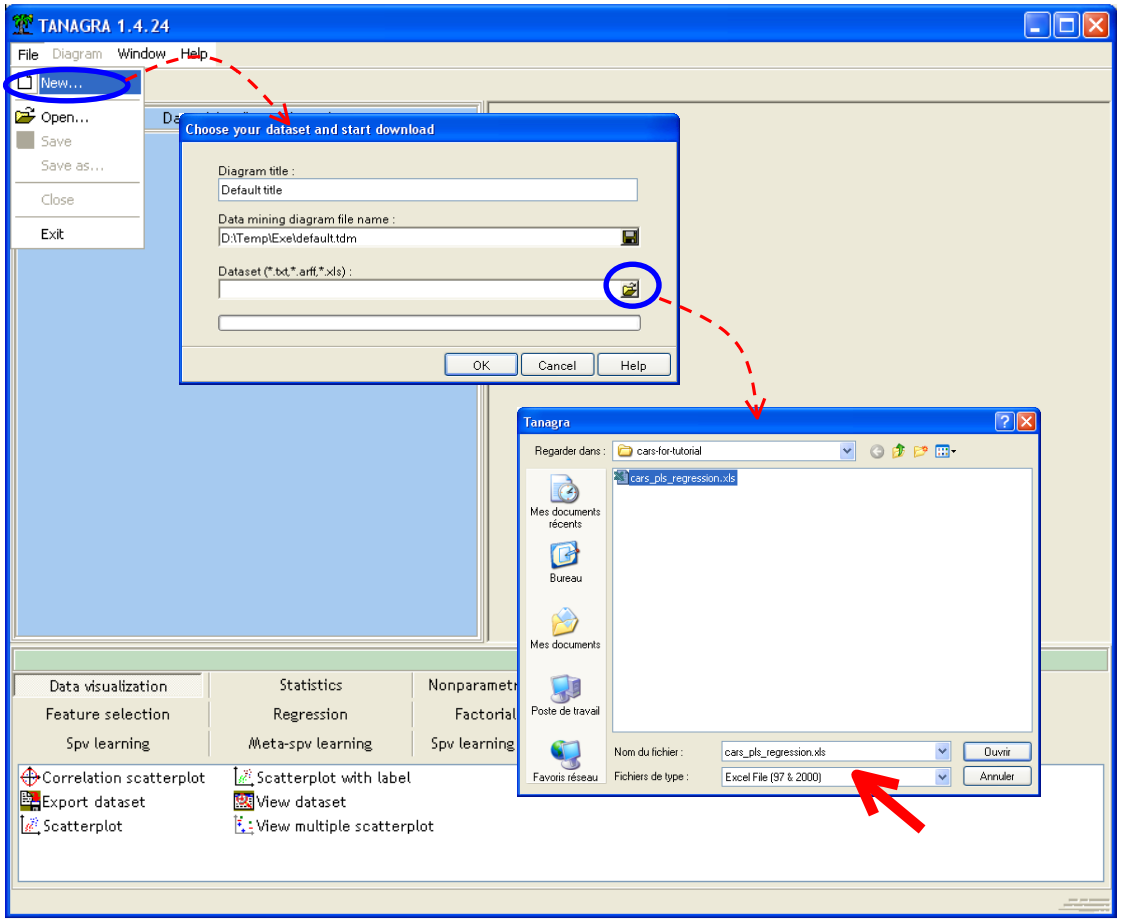
trouveront dans le code source de Tanagra les indications sur les formules (numéros de page, d'équation, …) utilisées pour les principales procédures. Attention, les données sont centrées et réduites d'office.

Dans ce didacticiel, nous nous attachons surtout à mettre en relation nos rapports avec ceux de SIMCA-P, le logiciel phare du domaine. Pour le lecteur désireux de s'initier à sa manipulation, une version gratuite, un tutoriel et un guide de l'utilisateur sont accessibles sur le site de l'éditeur.

4.1 Importation des données et création d'un diagramme

Il existe plusieurs manières de charger les données au format XLS dans TANAGRA. Nous choisissons l'importation directe[4]. Elle présente l'avantage de ne pas requérir la présence du logiciel EXCEL sur la machine (voir : http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-mode.html). Il faut en revanche que les données soient dans la première feuille de calcul, alignées en haut à gauche, la première ligne correspondant aux noms des variables. Notre configuration respecte ces spécifications. Attention, il ne faut pas que le fichier soit en cours d'édition lors de l'importation.

Après avoir démarré TANAGRA, nous activons le menu FILE / NEW pour créer un nouveau diagramme. Dans la boîte de sélection, nous spécifions le nom du fichier de données (CARS_PLS_REGRESSION.XLS) et le nom du fichier diagramme.

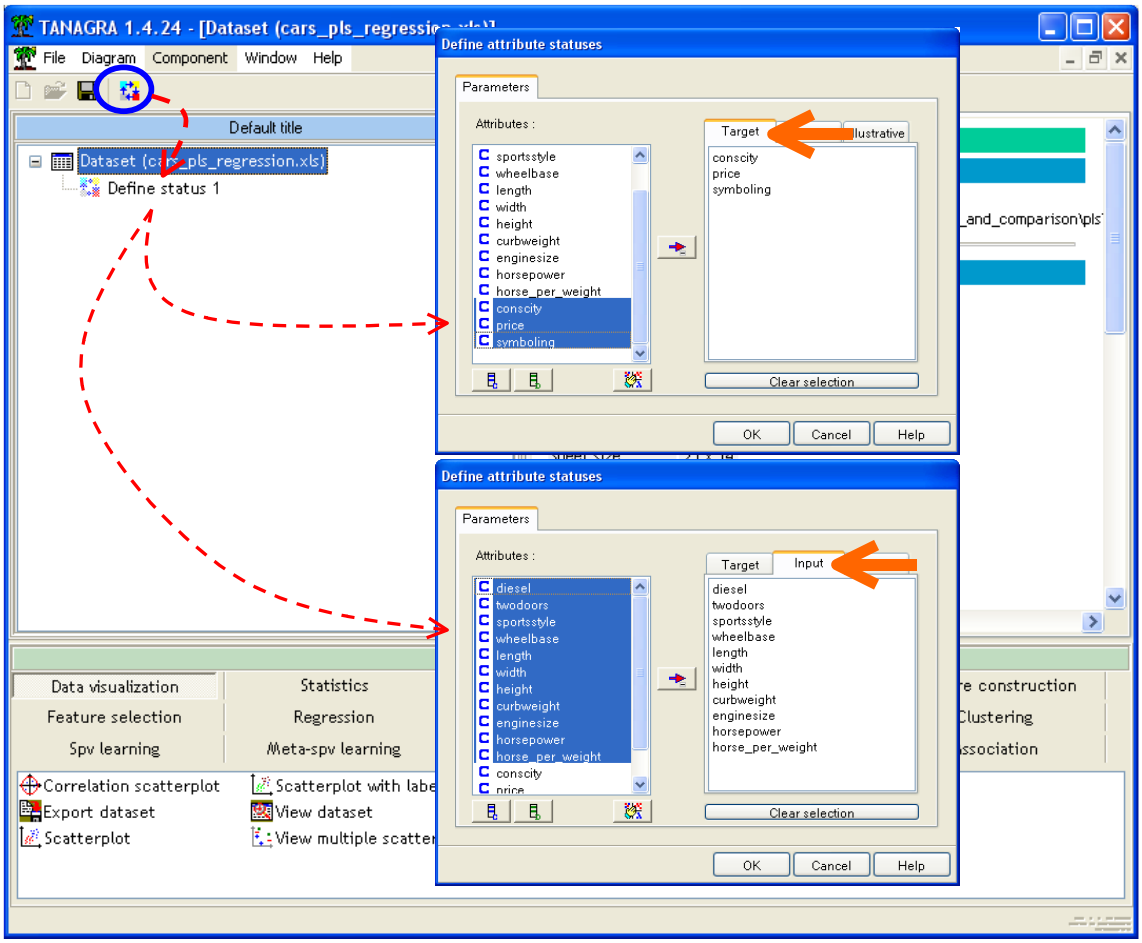


[4] L'autre possibilité d'importation est d'ouvrir le fichier dans le tableur. Puis à l'aide du nouveau menu TANAGRA dans EXCEL, inséré via la macro complémentaire TANAGRA.XLA, nous transférons les données. Voir : http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html

4.2 Variables cibles et variables prédictives

Nous voulons expliquer les variables CONSCITY (consommation en ville), PRICE (prix d'achat) et SYMBOLING (étiquetage des assureurs selon le risque) à l'aide des autres variables.

Nous insérons le composant DEFINE STATUS pour définir le rôle des variables. Nous utilisons le raccourci dans la barre d'outils.



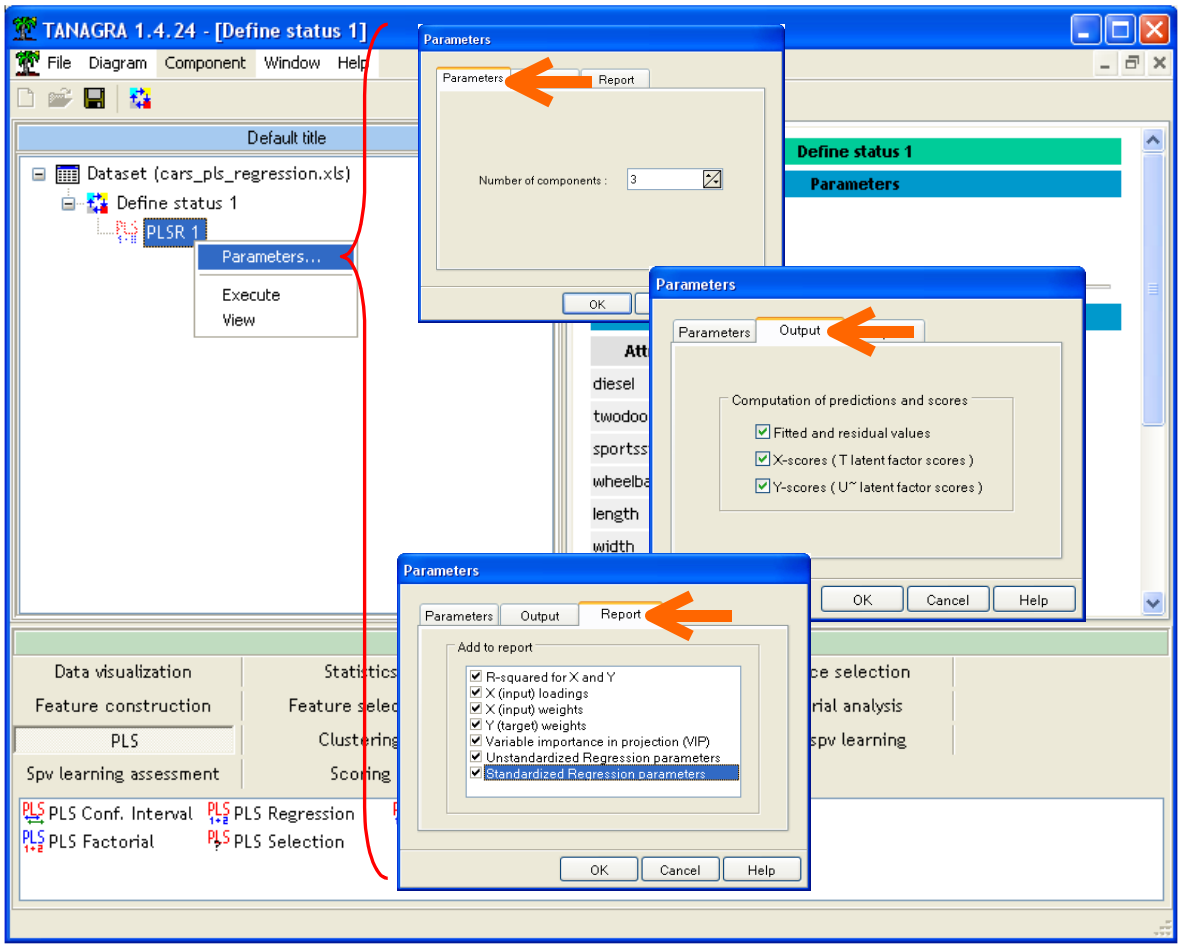
4.3 Paramétrage de la Régression PLS

Nous insérons le composant **PLSR** (onglet PLS) dans le diagramme. Nous activons le menu contextuel PARAMETERS… pour définir les paramètres. Dans le premier onglet PARAMETERS [1], nous spécifions le nombre de facteurs, nous choisissons la valeur « 3 ». Le nombre de facteurs est borné par le nombre de variables explicatives.

Dans le second onglet OUTPUT [2], nous spécifions les nouvelles variables qui sont produites par le composant. Elles sont disponibles pour les calculs dans les branches ultérieures du diagramme. Nous cochons toutes les options.

Enfin, dans le troisième onglet REPORT [3], nous indiquons les éléments qui seront intégrés au rapport. Ici également, nous sélectionnons tous les items.

Nous validons en cliquant sur OK. Puis, nous actionnons le menu contextuel VIEW du composant pour accéder aux résultats.



4.4 Description des résultats

Notre présentation et les commentaires associés reposent en grande partie sur les indications mis en avant dans les références suivantes : l'ouvrage de Tenenhaus (1998 ; pages 142 à 146) ; le guide de l'utilisateur de SIMCA-P (http://www.umetrics.com/pdfs/userguides/SIMCA-P_11_UG.pdf et http://www.umetrics.com/pdfs/userguides/SIMCA-P_11_Tutorial.pdf) ; le texte en ligne proposé par David Garson que nous utiliserons en priorité (http://www2.chass.ncsu.edu/garson/PA765/pls.htm)[5].

Précisons que les données correspondent à des véhicules en 1985 aux USA[6]. Les standards ne sont pas les mêmes qu'en Europe, surtout plus de 20 ans plus tard. Il ne faudra pas trop s'étonner de la teneur de certains résultats.

4.4.1 Qualité globale de la régression – Proportion de variance expliquée

Ce tableau indique la proportion de variance expliquée par les variables latentes, pour les descripteurs (X) et pour les variables cibles (Y), individuellement pour chaque facteurs et cumulativement pour les H premiers facteurs (Figure 2 et Figure 3).

Dans le premier cas, il traduit la qualité de représentation des variables sur les facteurs. Si nous retenons tous les facteurs (nombre de facteurs = nombre de variables explicatives), nous obtiendrons

[5] La référence de la page principale de son site web est http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

[6] http://archive.ics.uci.edu/ml/datasets/Automobile

forcément une proportion cumulée égale à 1. Dans notre exemple, les 3 premières variables latentes expliquent 81.927% de la variation des descripteurs. Une grande partie de l'information apportée par les descripteurs sont bien retranscrites.

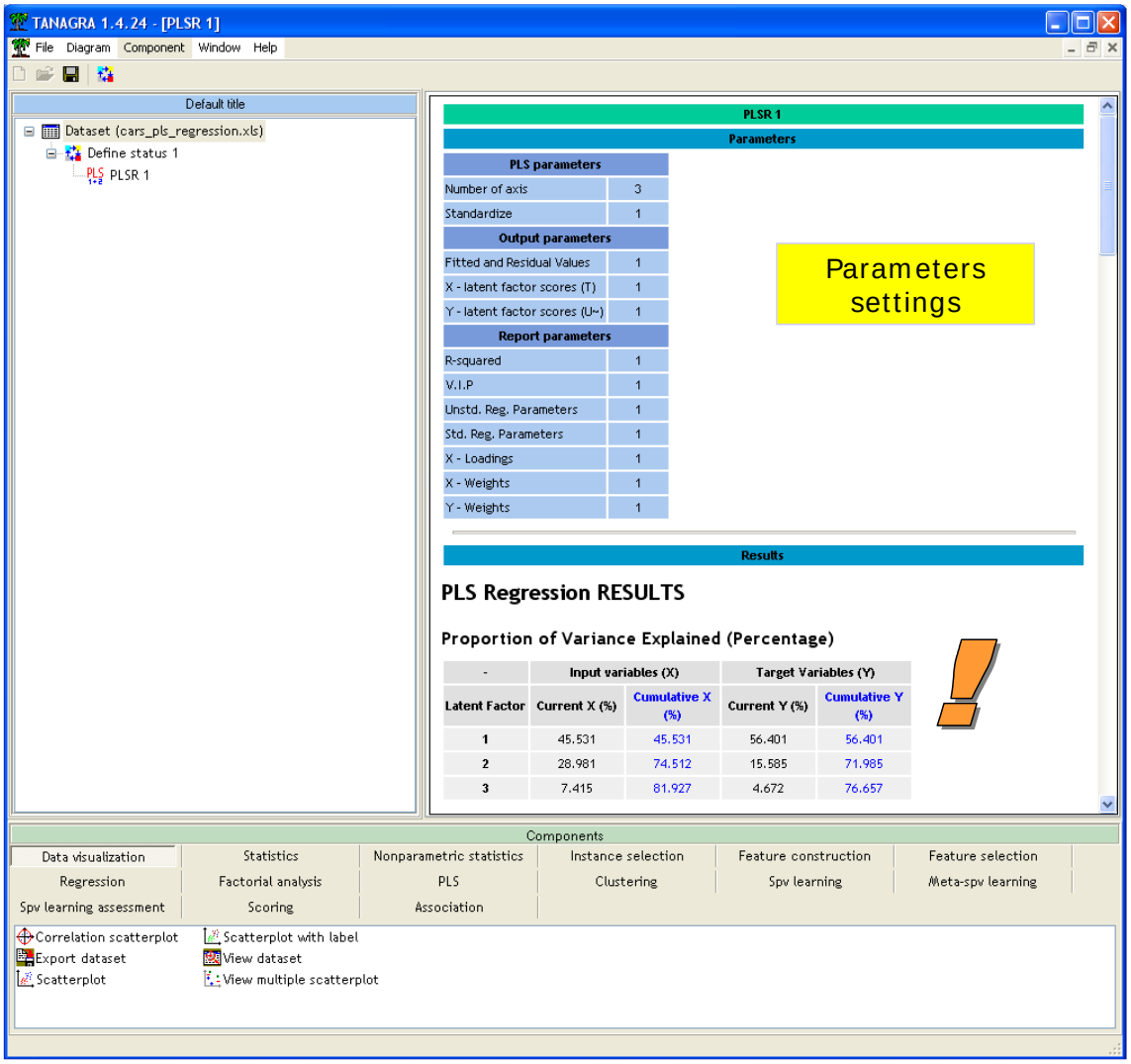


Figure 2 - Fenêtre principale de Tanagra - Proportion de variance expliquée

En ce qui concerne les variables cibles, nous observons le pouvoir explicatif du modèle. Généralement, mais pas toujours, la proportion de variance expliquée des variables cibles est plus faible que celle des variables expliquées.

Note : Si nous retenons tous les facteurs c.-à-d. nombre de facteurs = nombre de dscripteurs, nous obtiendrons le R^2 de la régression linéaire multiple.

Dans notre exemple, nous constatons que notre modèle est plutôt bon, 76.657% de la variabilité des Y sont expliquées par les 3 premières variables latentes.

Si la proportion cumulée est égale à 1, cela veut dire que la connaissance des valeurs de X permet de connaître avec certitudes les valeurs prises par les variables Y.

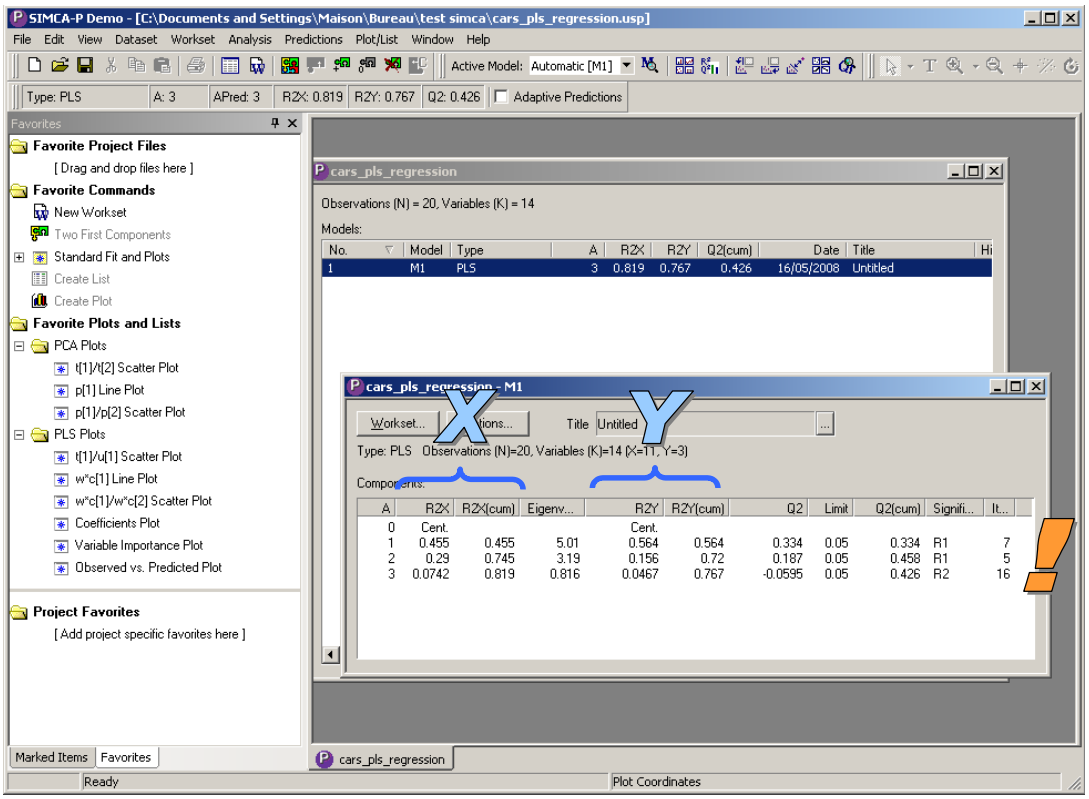


Figure 3 - Fenêtre principale de SIMCA-P - Proportion de variance expliquée

4.4.2 Variance expliquée des variables – R^2

Ce tableau (Figure 4) détaille le précédent. Il indique le carré de la corrélation de chaque variable (INPUT et TARGET) avec les variables latentes (Facteurs X) c.-à-d. la proportion de variance expliquée pour chaque variable.

Nous avons là un premier outil d'interprétation des variables latentes. Nous observons ainsi :

- Sur le 1er axe, les variables LENGTH (longueur), WIDTH (largeur), CURBWEIGHT (poids), ENGINESIZE (cylindrée du moteur) et, dans une moindre mesure HORSEPOWER (puissance) sont déterminantes. Nous n'avons pas le sens de la liaison à ce stade.

- Toujours sur ce premier facteur, si l'on se penche maintenant sur les variables cibles, nous constatons que les variables CONSCITY (consommation en ville) et PRICE (prix) sont bien expliquées. Il y a donc une certaine forme de liaison, qui reste à préciser, entre ces variables cibles et les variables explicatives énumérées précédemment.

- Sur le 2nd axe, représentant 28.98% de l'information disponible, les variables TWODOORS (2 portes), HORSE_PER_WEIGHT (rapport puissance poids) et hauteur (HEIGHT) sont importantes.

- Concernant les variables cibles, ce facteur explique 15.56% de la variation de Y, il est principalement en relation avec SYMBOLING (évaluation du risque affectée par les assureurs).

- Sur la dernière ligne des tableaux, nous retrouvons les proportions de variance expliquées (Figure 2).

- Le troisième axe est plus compliqué. La proportion de variance retranscrite est faible (7.42%). On y voit essentiellement une liaison entre le style sport (cabriolet, coupé ; un véhicule peut avoir un style sport sans être pour autant une pure sportive, style « fun » est peut être plus approprié ici) et la carburation diesel. Le sens de la liaison demande à être précisé.

R-squared

Input(s) vs. X-Scores						
-	R-squared			Cumulative R-squared		
Input	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
diesel	0.0871	0.1904	0.4243	0.0871	0.2775	0.7018
twodoors	0.0012	0.6790	0.0152	0.0012	0.6802	0.6954
sportsstyle	0.0189	0.2612	0.2483	0.0189	0.2801	0.5284
wheelbase	0.5527	0.3602	0.0393	0.5527	0.9129	0.9522
length	0.8236	0.1376	0.0027	0.8236	0.9613	0.9639
width	0.7692	0.0961	0.0118	0.7692	0.8652	0.8771
height	0.0157	0.5017	0.0173	0.0157	0.5174	0.5347
curbweight	0.9244	0.0258	0.0026	0.9244	0.9502	0.9528
enginesize	0.8967	0.0037	0.0240	0.8967	0.9005	0.9244
horsepower	0.7035	0.2546	0.0155	0.7035	0.9581	0.9737
horse_per_weight	0.2155	0.6775	0.0146	0.2155	0.8930	0.9076
Total Exp.	0.4553	0.2898	0.0742	0.4553	0.7451	0.8193

Target(s) vs. X-Scores						
-	R-squared			Cumulative R-squared		
Target	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
conscity	0.8836	0.0407	0.0030	0.8836	0.9243	0.9273
price	0.7790	0.0168	0.1129	0.7790	0.7958	0.9087
symboling	0.0294	0.4100	0.0242	0.0294	0.4394	0.4637
Total Exp.	0.5640	0.1558	0.0467	0.5640	0.7199	0.7666

Figure 4 - Tanagra – R² avec les facteurs X - Variables explicatives et expliquées

Summary [M1]

	1	2	3	4	5	6	7
1	Var ID (Primary)		R2VY	R2VY(cum)	Q2VY	Q2 limit	Q2VY(cum)
2	Total	Comp 1	0.5640	0.5640	0.3340	0.0500	0.3340
3		Comp 2	0.1558	0.7199	0.1868	0.0500	0.4584
4		Comp 3	0.0467	0.7666	-0.0595	0.0500	0.4262
5							
6	conscity	Comp 1	0.8836	0.8836	0.7999	0.0500	0.7999
7		Comp 2	0.0407	0.9243	0.2820	0.0500	0.8564
8		Comp 3	0.0030	0.9273	-0.0425	0.0500	0.8503
9							
10	price	Comp 1	0.7790	0.7790	0.5834	0.0500	0.5834
11		Comp 2	0.0168	0.7958	-0.3478	0.0500	0.5417
12		Comp 3	0.1129	0.9087	0.3452	0.0500	0.6999
13							
14	symboling	Comp 1	0.0294	0.0294	-0.3812	0.0500	-0.1000
15		Comp 2	0.4100	0.4394	0.2971	0.0500	0.2268
16		Comp 3	0.0242	0.4637	-0.2092	0.0500	0.1495
17							

Figure 5 – SIMCA-P – R² de chaque variable Y avec les facteurs X

SIMCA-P : Menu ANALYSIS / SUMMARY / MODEL OVERVIEW LIST (Figure 5)

4.4.3 LOADINGS des variables explicatives INPUT – Vecteur Ph

Les LOADINGS « reflètent » les corrélations entre les variables latentes et les variables explicatives. Ils complètent les R^2 (Figure 4) en précisant le sens de la liaison. Utilisés conjointement, ces tableaux permettent d'interpréter finement les facteurs t_h.

Dans la pratique, on considère qu'une valeur supérieure à **0.4** - en valeur absolue - est l'indication d'une liaison significative. Mais il faut être prudent là dessus. Mieux vaut se fier surtout à l'interprétation que l'on peut tirer des facteurs. On peut même être plus permissif encore et descendre le seuil à **0.25** (voir http://www2.chass.ncsu.edu/garson/PA765/pls.htm).

Attention, les LOADINGS ne correspondent pas aux coefficients de corrélation, d'où les guillemets ci-dessus. En revanche, ils permettent de positionner les variables de la même manière sur l'axe factoriel, et c'est bien ce qui importe dans l'interprétation. On se penchera avant tout sur les variables situées aux extrêmes.

Dans notre exemple, le 1er axe est formé par la conjonction des caractéristiques LENGTH, WIDTH, CURBWEIGHT, ENGINESIZE c.-à-d. les gros véhicules lourds, avec de gros moteurs, et relativement puissantes.

Sur le 2nd axe, les voitures basses à deux portes sont plutôt dynamiques (rapport puissance poids élevé). On voit maintenant sur le 3ème axe qu'être « fun » et carburer au diesel, c'est antinomique.

TANAGRA

X-loadings (Model Effect Loadings - Vector Ph)

Input	Factor1	Factor2	Factor3
diesel	-0.1326	-0.2496	-0.7368
twodoors	0.0154	0.4714	0.1393
sportsstyle	-0.0617	0.2924	0.5636
wheelbase	0.3342	-0.3433	0.2242
length	0.4080	-0.2122	0.0586
width	0.3942	-0.1773	0.1231
height	0.0563	-0.4052	0.1488
curbweight	0.4322	-0.0919	-0.0577
enginesize	0.4257	0.0349	-0.1751
horsepower	0.3770	0.2886	-0.1410
horse_per_weight	0.2087	0.4708	-0.1367

SIMCA-P

General List [M1]

	1	2	3	4
1	Var ID (Primary)	M1.p[1]	M1.p[2]	M1.p[3]
2	diesel	-0.1326	-0.2496	0.7368
3	twodoors	0.0154	0.4714	-0.1393
4	sportsstyle	-0.0617	0.2924	-0.5636
5	wheelbase	0.3342	-0.3433	-0.2242
6	length	0.4080	-0.2122	-0.0586
7	width	0.3942	-0.1773	-0.1231
8	height	0.0563	-0.4052	-0.1488
9	curbweight	0.4322	-0.0919	0.0577
10	enginesize	0.4257	0.0349	0.1751
11	horsepower	0.3770	0.2886	0.1410
12	horse_per_weigl	0.2087	0.4708	0.1367

Figure 6 - X-Loadings - Vecteur Ph (Lien INPUT - Facteurs X)

SIMCA-P : Menu ANALYSIS / LOADINGS / LINE PLOT, et choix de la série « p ».

4.4.4 WEIGHTS des variables cibles TARGET – Vecteur Ch

Les poids des variables cibles « reflètent » les corrélations entre ces variables et les scores u_h (Figure 7). Ils permettent de cerner ce qui est expliqué sur ces facteurs. Encore une fois ici, il ne s'agit pas du véritable coefficient de corrélation, mais le positionnement des variables est le même.

Pour notre étude, nous constatons que le premier axe explique essentiellement la conjonction entre le prix et la consommation en ville. Sur le second axe, on observe la pénalisation par les assureurs qui affectent un SYMBOLING élevé.

Il faut maintenant mettre en relation ces caractéristiques avec les variables explicatives.

TANAGRA

Y-Weights (Dependent Variable Weights - Vector Ch)

Target	Factor 1	Factor 2	Factor 3
conscity	0.4225	0.1154	0.0623
price	0.3968	0.0742	-0.3800
symboling	-0.0771	0.3663	-0.1760

SIMCA-P

General List [M1]

	1	2	3	4
1	Var ID (Primary)	M1.c[1]	M1.c[2]	M1.c[3]
2	conscity	0.4225	0.1154	-0.0623
3	price	0.3968	0.0742	0.3800
4	symboling	-0.0771	0.3663	0.1760

Figure 7 - Poids des variables cibles dans la définition des Facteurs Y

SIMCA-P : Menu ANALYSIS / LOADINGS / LINE PLOT, puis choix de la série « c ».

4.4.5 WEIGHTS des variables explicatives INPUT – Vecteurs Wh et Wh*

Les poids des variables explicatives « reflètent » la corrélation de ces variables avec les scores u_h (Figure 8). Ils indiquent le rôle de chaque descripteur dans l'explication globale de chaque axe.

On se rend compte dans la pratique que poids (WEIGHTS) et LOADINGS (section 4.4.3) positionnent les variables de manière similaire, les interprétations sont quasi identiques.

Dans notre exemple, on peut dire maintenant que les véhicules à prix et à consommation élevés sont associés aux lourds véhicules à gros moteurs. Sur le 2nd axe, il semble que les assureurs n'aiment pas les voitures performantes basses à 2 portes (*ben quoi, qu'est-ce qu'elle a ma 205 GTI 16 soupapes turbo avec un aileron de requin sur le capot avant ?*).

Les vecteurs Wh*, contrairement aux Wh, s'appliquent directement sur les variables explicatives (les Wh s'appliquent sur les résidus successifs des explicatives).

Nous observons que les signes sont inversés sur le 3ème axe (Figure 8). La direction de l'axe n'est pas la même, mais le positionnement relatif des variables est respecté. C'est ce qui importe.

TANAGRA

X-Weights (Model Effect Weights - Vector Wh and Wh*)

-	Wh			Wh*		
Input	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
diesel	-0.1396	-0.1605	-0.6992	-0.1396	-0.1731	-0.6691
twodoors	0.0262	0.4727	0.2439	0.0262	0.4750	0.1750
sportsstyle	-0.0723	0.1570	0.5264	-0.0723	0.1505	0.5077
wheelbase	0.2873	-0.3950	0.2310	0.2873	-0.3691	0.2727
length	0.3725	-0.1732	0.0704	0.3725	-0.1396	0.0761
width	0.3560	-0.1094	0.0001	0.3560	-0.0773	-0.0025
height	0.0142	-0.3782	0.1794	0.0142	-0.3769	0.2327
curbweight	0.4278	-0.1395	0.0468	0.4278	-0.1009	0.0448
enginesize	0.4490	-0.0117	-0.1792	0.4490	0.0288	-0.2005
horsepower	0.4144	0.3025	-0.1251	0.4144	0.3399	-0.1896
horse_per_weight	0.2645	0.5235	-0.1845	0.2645	0.5474	-0.2728

SIMCA-P

General List [M1]

	1	2	3	4	5	6	7
1	Var ID (Primary)	M1.w[1]	M1.w[2]	M1.w[3]	M1.w*[1]	M1.w*[2]	M1.w*[3]
2	diesel	-0.1396	-0.1605	0.6992	-0.1396	-0.1731	0.6691
3	twodoors	0.0262	0.4727	-0.2439	0.0262	0.4750	-0.1750
4	sportsstyle	-0.0723	0.1570	-0.5264	-0.0723	0.1505	-0.5077
5	wheelbase	0.2873	-0.3950	-0.2310	0.2873	-0.3691	-0.2727
6	length	0.3725	-0.1732	-0.0704	0.3725	-0.1396	-0.0761
7	width	0.3560	-0.1094	-0.0001	0.3560	-0.0773	0.0025
8	height	0.0142	-0.3782	-0.1794	0.0142	-0.3769	-0.2327
9	curbweight	0.4278	-0.1395	-0.0468	0.4278	-0.1009	-0.0448
10	enginesize	0.4490	-0.0117	0.1792	0.4490	0.0288	0.2005
11	horsepower	0.4144	0.3025	0.1251	0.4144	0.3399	0.1896
12	horse_per_weight	0.2645	0.5235	0.1845	0.2645	0.5474	0.2728

Figure 8 - Poids des descripteurs dans la définition des Facteurs Y

SIMCA-P : Menu ANALYSIS / LOADINGS / LINE PLOT, et choix des séries « w » et « w* ».

4.4.6 Importance des variables explicatives dans la régression (VIP)

Le tableau des VIP indique l'importance relative des variables dans l'explication des valeurs de Y, en tenant compte des H premiers facteurs (Figure 9). (VIP > 1) est une indication souvent mise en avant pour repérer les variables les plus intéressantes. A contrario, (VIP < 0.8) et valeur absolue du coefficient de la variable dans la régression proche de 0 inciterait à l'éliminer.

Dans notre exemple, on ne devrait lire que la dernière colonne puisque l'on retient 3 facteurs. Les colonnes précédentes rendent compte de l'évolution du rôle de la variable à mesure que l'on intègre de l'information. Prenons l'exemple de la variable TWODOORS (véhicules à deux portes). Sur le 1er axe, il ne joue absolument aucun rôle. Dès que l'on intègre le 2nd axe (en plus du 1er, il s'agit de l'influence cumulée des facteurs), il commence à se démarquer pour expliquer la pénalisation imposée par les assureurs. On constate l'évolution concomitante de la variable rapport puissance – poids (HORSE_PER_WEIGHT).

TANAGRA

VIP (Variable Importance in Projection)

Input	Factor 1	Factor 2	Factor 3
diesel	0.4630	0.4788	0.7369
twodoors	0.0870	0.7334	0.7383
sportsstyle	0.2399	0.3222	0.5322
wheelbase	0.9530	1.0408	1.0261
length	1.2354	1.1257	1.0924
width	1.1807	1.0586	1.0259
height	0.0472	0.5851	0.5857
curbweight	1.4187	1.2741	1.2353
enginesize	1.4893	1.3184	1.2860
horsepower	1.3746	1.3032	1.2670
horse_per_weight	0.8774	1.1206	1.0964

SIMCA-P

VIP List [M1]

	1	2
1	Var ID (Primary)	M1.VIP[3]
2	enginesize	1.2860
3	horsepower	1.2670
4	curbweight	1.2353
5	horse_per_weight	1.0964
6	length	1.0924
7	wheelbase	1.0261
8	width	1.0259
9	twodoors	0.7383
10	diesel	0.7369
11	height	0.5857
12	sportsstyle	0.5322

Figure 9 - Importance des descripteurs dans la projection

SIMCA-P : Menu ANALYSIS / VARIABLE IMPORTANCE / LIST. On ne peut afficher que les VIP pour un nombre choisi de facteurs. Les variables sont triées par ordre d'importance décroissante.

4.4.7 Coefficients non standardisés de la régression

Ce tableau retrace les coefficients de régression : une équation par variable cible (Figure 10).

TANAGRA

Unstandardized Regression Parameters (by target variable)

-	Target Variable(s)		
Input	conscity	price	symboling
diesel	-0.946269	4688.641651	0.208085
twodoors	0.450889	-393.099643	0.337465
sportsstyle	0.129202	-4734.098001	-0.081700
wheelbase	0.040765	-23.277199	-0.035576
length	0.036369	86.825747	-0.009458
width	0.192261	596.478952	-0.030636
height	-0.029355	-454.791393	-0.093726
curbweight	0.000863	2.341868	-0.000159
enginesize	0.010825	49.410891	0.000274
horsepower	0.013530	56.182058	0.003425
horse_per_weight	39.666768	201165.156382	23.329061
constant	-18.187248	-39378.332031	12.198230

SIMCA-P

Coefficient List [M1]

	1	2	3	4
1	Var ID (Primary)	M1.Coeff[3](conscity)	M1.Coeff[3](price)	M1.Coeff[3](symboling)
2	$constant	-18.187222	-39379.011719	12.198215
3	diesel	-0.946268	4688.631836	0.208090
4	twodoors	0.450888	-393.075836	0.337465
5	sportsstyl	0.129203	-4734.107422	-0.081705
6	wheelbase	0.040765	-23.277510	-0.035576
7	length	0.036369	86.826599	-0.009458
8	width	0.192261	596.485535	-0.030635
9	height	-0.029355	-454.787842	-0.093727
10	curbweight	0.000863	2.341853	-0.000159
11	enginesize	0.010825	49.410564	0.000274
12	horsepower	0.013530	56.182014	0.003425
13	horse_per_	39.666752	201165.500000	23.329159

Figure 10 - Coefficients non standardisés de la régression

Les coefficients non standardisés s'appliquent directement aux variables originelles, sans transformation préalable. Ils sont directement opérationnels pour la prédiction. Etant exprimés dans des unités différentes, ils ne permettent pas de discerner l'importance relative des variables.

On lit simplement ici :

- Lorsque la puissance (HORSEPOWER) augmente d'une unité, la consommation augmente de 0.013530 litres aux 100 km.
- Par rapport aux autres, les véhicules à 2 portes (TWODOORS = 1) ont une consommation supérieure de 0.45 l/100km en moyenne.
- Etc.

SIMCA-P : Menu ANALYSIS / COEFFICIENTS / LIST. Dans les propriétés, on peut définir les coefficients à afficher, on choisit UNSCALED.

4.4.8 Coefficients standardisés de la régression

Les coefficients standardisés (Figure 11) permettent une interprétation en termes d'écarts type. De fait, on peut positionner les variables les unes par rapport aux autres dans le processus d'explication :

- Lorsque ENGINESIZE augmente d'un écart type, la consommation augmentera de 0.18 écarts type.
- Lorsque HORSEPOWER augmente d'un écart type, la consommation augmentera 0.20 écarts type. Son influence est très similaire.

TANAGRA

Standardized Regression Parameters (by target variable)

-	Target Variable(s)				
-	Average		10.0517	14579.5000	1.0000
-	Std.Deviation		2.8730	9231.7001	1.1698
Input	Average	Std.Deviation	conscity	price	symboling
diesel	0.1500	0.3663	-0.120663	0.186062	0.065167
twodoors	0.3500	0.4894	0.076801	-0.020838	0.141172
sportsstyle	0.2000	0.4104	0.018456	-0.210452	-0.028662
wheelbase	99.5450	6.7530	0.095818	-0.017027	-0.205370
length	174.6500	11.5355	0.146026	0.108493	-0.093263
width	65.8000	2.1121	0.141344	0.136468	-0.055314
height	54.0000	2.2480	-0.022970	-0.110748	-0.180118
curbweight	2570.6500	572.3690	0.171900	0.145197	-0.077838
enginesize	134.5500	47.9215	0.180569	0.256491	0.011226
horsepower	106.1500	43.0034	0.202525	0.261709	0.125914
horse_per_weight	0.0406	0.0114	0.157935	0.249261	0.228124

SIMCA-P

Coefficient List [M1]

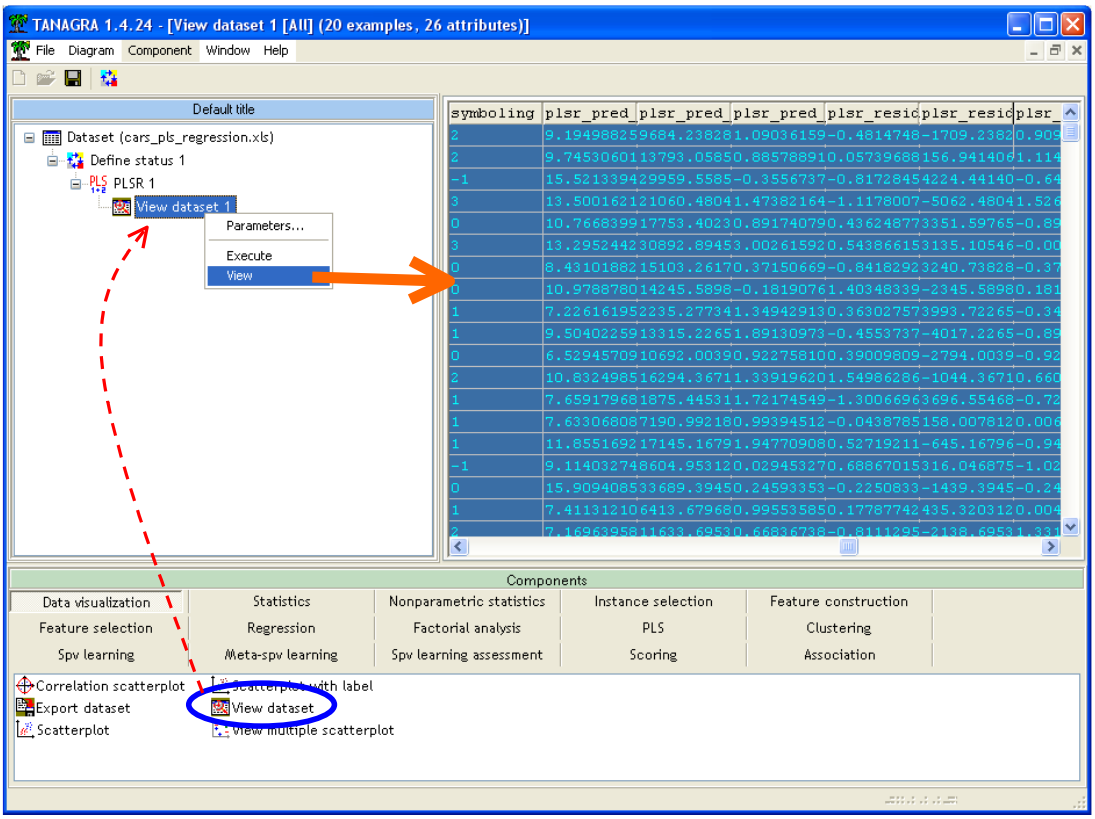
	1	2	3	4
1	Var ID (Primary)	M1.CoeffCS[3](conscity)	M1.CoeffCS[3](price)	M1.CoeffCS[3](symboling)
2	diesel	-0.120663	0.186062	0.065168
3	twodoors	0.076801	-0.020836	0.141172
4	sportsstyl	0.018456	-0.210453	-0.028664
5	wheelbase	0.095818	-0.017027	-0.205371
6	length	0.146026	0.108494	-0.093263
7	width	0.141344	0.136470	-0.055314
8	height	-0.022970	-0.110747	-0.180118
9	curbweight	0.171900	0.145196	-0.077838
10	enginesize	0.180569	0.256489	0.011226
11	horsepower	0.202525	0.261709	0.125914
12	horse_per_	0.157935	0.249262	0.228125

Figure 11 - Coefficients standardisés de la régression

SIMCA-P : Menu ANALYSIS / COEFFICIENTS / LIST. Dans les propriétés, nous définissons les coefficients à afficher, nous choisissons SCALED & CENTERED.

4.4.9 Prédiction et résidus de la régression

Lors du paramétrage de PLSR (Section 4.3), nous avions cochés toutes les options dans l'onglet OUTPUT. Dans ce cas, le composant produit des nouvelles variables disponibles pour les éventuels traitements en aval. Pour les visualiser, nous insérons le composant VIEW DATASET (onglet DATA VISUALIZATION) dans le diagramme. Nous cliquons sur le menu contextuel VIEW. Dans la grille qui s'affiche, nous retrouvons les données originelles et, en dernières positions, les nouvelles colonnes associées aux scores, aux prédictions et aux résidus.



Intéressons-nous dans un premier temps à la prédiction. Deux nouvelles variables sont créées pour chaque variable cible : la première correspond aux valeurs prédites, elle est automatiquement nommée PLSR_PRED_NOM_VARIABLE_CIBLE ; la seconde est l'erreur de prédiction c.-à-d. la différence entre les valeurs observées et prédites, nommée PLSR_RESIDUAL_NOM_VARIABLE_CIBLE.

Nous pouvons copier les données de la grille dans un tableur (EXCEL ou autres) pour les calculs ultérieurs ou tout simplement pour une meilleure mise en forme. Nous retrouvons les prévisions et résidus pour chaque variable cible (Figure 12). Nous pouvons rapprocher les valeurs avec la prévision (colonne 4) et le résidu (colonne 2) pour la variable CONSCITY de SIMCA-P (Figure 13) qui choisit une présentation individualisée des variables cibles.

	Prédiction			Résidus		
examples	ed_conscity_1	_pred_price_1	_symbolinq_1	al_conscity_1	sidual_price_1	_symbolinq_1
1	9.195	9684.238	1.090	-0.481	-1709.238	0.910
2	9.745	13793.059	0.886	0.057	156.941	1.114
3	15.521	29959.559	-0.356	-0.817	4224.441	-0.644
4	13.500	21060.480	1.474	-1.118	-5062.480	1.526
5	10.767	17753.402	0.892	0.436	3351.598	-0.892
6	13.295	30892.895	3.003	0.544	3135.105	-0.003
7	8.431	15103.262	0.372	-0.842	3240.738	-0.372
8	10.979	14245.590	-0.182	1.403	-2345.590	0.182
9	7.226	2235.277	1.349	0.363	3993.723	-0.349
10	9.504	13315.227	1.891	-0.455	-4017.227	-0.891
11	6.529	10692.004	0.923	0.390	-2794.004	-0.923
12	10.832	16294.367	1.339	1.550	-1044.367	0.661
13	7.659	1875.445	1.722	-1.301	3696.555	-0.722
14	7.633	7190.992	0.994	-0.044	158.008	0.006
15	11.855	17145.168	1.948	0.527	-645.168	-0.948
16	9.114	8604.953	0.029	0.689	316.047	-1.029
17	15.909	33689.395	0.246	-0.225	-1439.395	-0.246
18	7.411	6413.680	0.996	0.178	435.320	0.004
19	7.170	11633.695	0.668	-0.811	-2138.695	1.332
20	8.757	10007.340	0.716	-0.043	-1512.340	1.284

Figure 12 - Prévision et résidus de Tanagra

Residuals List [M1]

	1	2	3	4
1	Obs ID (Primary)	Observed-Pred.	M1.YVar(conscity)	M1.YPred[3](conscity)
2	1	-0.481	8.714	9.195
3	2	0.057	9.803	9.745
4	3	-0.817	14.704	15.521
5	4	-1.118	12.382	13.500
6	5	0.436	11.203	10.767
7	6	0.544	13.839	13.295
8	7	-0.842	7.589	8.431
9	8	1.403	12.382	10.979
10	9	0.363	7.589	7.226
11	10	-0.455	9.049	9.504
12	11	0.390	6.920	6.529
13	12	1.550	12.382	10.832
14	13	-1.301	6.359	7.659
15	14	-0.044	7.589	7.633
16	15	0.527	12.382	11.855
17	16	0.689	9.803	9.114
18	17	-0.225	15.684	15.909
19	18	0.178	7.589	7.411
20	19	-0.811	6.359	7.170
21	20	-0.043	8.714	8.757
22		RMSEE	0.844	

Figure 13 - SIMCA-P - Résidus, Variance des résidus et Prévision pour CONSCITY

Note : Petit avantage pour SIMCA-P qui propose une estimation de la variance de l'erreur (Figure 13, colonne 3). C'est indispensable si l'on souhaite construire des intervalles de prévision.

4.4.10 Composantes PLS sur les variables explicatives (Scores X, Vecteur « t_h »)

TANAGRA

examples	plsr_t_1_1	plsr_t_2_1	plsr_t_3_1
1	-0.8663	0.2919	0.5479
2	-0.1672	-0.3067	-0.0102
3	4.9653	-1.9137	0.4263
4	2.1854	1.9583	0.8166
5	0.6839	-0.2210	-0.2338
6	1.8294	4.1295	-1.9338
7	-0.5753	-2.1501	-1.1698
8	1.1884	-2.0594	0.9338
9	-2.6462	0.6861	0.8899
10	-0.9021	1.7776	-0.2343
11	-2.2872	-1.4091	-1.5550
12	0.3808	0.9137	0.0872
13	-2.6206	1.7195	1.2210
14	-1.8943	-0.3876	0.0527
15	0.6998	2.6030	0.5076
16	-0.3893	-1.9024	0.9250
17	5.0939	-0.8270	-0.2904
18	-2.0754	-0.4084	0.0810
19	-1.7067	-1.7503	-1.2840
20	-0.8963	-0.7439	0.2222

SIMCA-P

General List [M1]

	1	2	3	4
1	Obs ID (Primary)	M1.t[1]	M1.t[2]	M1.t[3]
2	1	-0.8663	0.2919	-0.5479
3	2	-0.1672	-0.3067	0.0102
4	3	4.9653	-1.9137	-0.4263
5	4	2.1854	1.9583	-0.8166
6	5	0.6839	-0.2210	0.2338
7	6	1.8294	4.1295	1.9338
8	7	-0.5753	-2.1501	1.1698
9	8	1.1884	-2.0594	-0.9338
10	9	-2.6462	0.6861	-0.8899
11	10	-0.9021	1.7776	0.2343
12	11	-2.2872	-1.4091	1.5550
13	12	0.3808	0.9137	-0.0872
14	13	-2.6206	1.7195	-1.2210
15	14	-1.8943	-0.3876	-0.0527
16	15	0.6998	2.6030	-0.5076
17	16	-0.3893	-1.9024	-0.9250
18	17	5.0939	-0.8270	0.2904
19	18	-2.0754	-0.4084	-0.0810
20	19	-1.7067	-1.7503	1.2840
21	20	-0.8963	-0.7439	-0.2222

Figure 14 - Vecteurs "t_h" pour Tanagra et SIMCA-P

SIMCA-P : Menu ANALYSIS / SCORES / LINE PLOT, puis choix de la série « t ».

Communément appelés SCORES X dans les logiciels de langue anglaise, il s'agit des composantes latentes définies sur les variables explicatives. La formule adéquate est

$$t_h = X \times w_h^*$$

La projection des individus dans ces nouveaux repères renforce l'interprétation des résultats. Elle permet aussi de mieux situer les cas individuels, voire de distinguer des groupes comme cela se fait usuellement en analyse en composantes principales.

4.4.11 Composantes PLS sur les variables expliquées (Scores Y, Vecteur « $\tilde{u}_h$ »)

Communément appelés SCORES Y dans les références anglaises, il s'agit cette fois ci de la projection des individus dans l'espace des facteurs calculés sur les variables cibles.

$$\tilde{u}_h = Y \times c_h$$

TANAGRA – Vecteur « $\tilde{u}$ »

examples	plsr_utilde_1_1	plsr_utilde_2_1	plsr_utilde_3_1
1	-1.5986	1.3483	0.5151
2	-0.3790	1.9482	-0.7250
3	4.8511	-1.8419	-2.2595
4	0.7953	4.7797	-1.7224
5	1.5083	-1.4015	-0.5196
6	3.6883	6.1095	-5.6858
7	-0.3933	-2.4954	-0.3230
8	0.8585	-1.5756	1.7366
9	-2.1089	-1.0852	1.6193
10	-1.0954	-0.5408	1.0912
11	-1.9944	-3.2200	1.9944
12	0.8940	2.6938	-0.7113
13	-2.7209	-1.4428	1.6212
14	-1.9682	-1.0264	1.3621
15	1.2440	0.7128	-0.1589
16	-0.4328	-4.4560	2.9478
17	4.8369	0.3605	-2.5362
18	-2.0310	-1.0527	1.4769
19	-2.4206	0.8099	-0.1189
20	-1.5333	1.3757	0.3957

SIMCA-P – Vecteur « u »

General List [M1]

	1	2	3	4
1	Obs ID (Primary)	M1.u[1]	M1.u[2]	M1.u[3]
2	1	-1.5986	1.6312	-0.1186
3	2	-0.3790	2.0028	0.9746
4	3	4.8511	-3.4632	0.1017
5	4	0.7953	4.0662	-0.5627
6	5	1.5083	-1.6249	0.2020
7	6	3.6883	5.5121	2.5856
8	7	-0.3933	-2.3075	1.7039
9	8	0.8585	-1.9636	-1.4895
10	9	-2.1089	-0.2212	-0.3102
11	10	-1.0954	-0.2463	-1.3811
12	11	-1.9944	-2.4732	0.0918
13	12	0.8940	2.5694	0.0402
14	13	-2.7209	-0.5872	-0.8207
15	14	-1.9682	-0.4079	-0.0060
16	15	1.2440	0.4843	-1.5150
17	16	-0.4328	-4.3289	-1.7999
18	17	4.8369	-1.3028	-0.2193
19	18	-2.0310	-0.3750	0.0011
20	19	-2.4206	1.3672	2.0088
21	20	-1.5333	1.6683	0.5133

Figure 15 - Vecteur "$\tilde{u}$" pour Tanagra, Vecteur "u" pour SIMCA-P

SIMCA-P : Menu ANALYSIS / SCORES / LINE PLOT, puis choix de la série « u ».

Horreur et damnation, mis à part la première colonne, les résultats sont divergents. Une erreur se serait-elle glissée dans les calculs ?

Heureusement, il n'en est rien (*ouf !*). Conformément aux indications de notre principale référence (Tenenhaus, page 136), à la différence de SIMCA-P, nous préférons proposer les composantes $\tilde{u}_h$ qui sont d'interprétation plus aisée. En effet, elles s'expriment comme une combinaison linéaire des variables cibles.

4.4.12 Quelques graphiques

La Régression est aussi (avant tout ?) une méthode factorielle. Certains auteurs associent d'ailleurs l'acronyme PLS à « Projection to Latent Structures ». A ce titre, on en tire vraiment la quintessence avec les différents graphiques (cartes) qui permettent de positionner les individus entre eux, les variables entre elles.

Cartes des variables. Concernant les cartes des variables basées sur les LOADINGS et les WEIGHTS, il n'est pas possible de les réaliser directement dans Tanagra. Le plus simple dans ce cas est de copier les résultats dans un tableur (menu COMPONENT / COPY RESULTS). Nous avons ainsi toute latitude pour élaborer les graphiques adéquats.

Ci dessous nous reprenons le graphique de SIMCA-P croisant les poids Wh* et Ch dans le premier plan factoriel. Nous retrouvons les commentaires relatifs aux concomitances et oppositions mis en lumière lors de l'analyse des tableaux associés. La représentation graphique devient réellement décisive lorsque le nombre de variables est important.

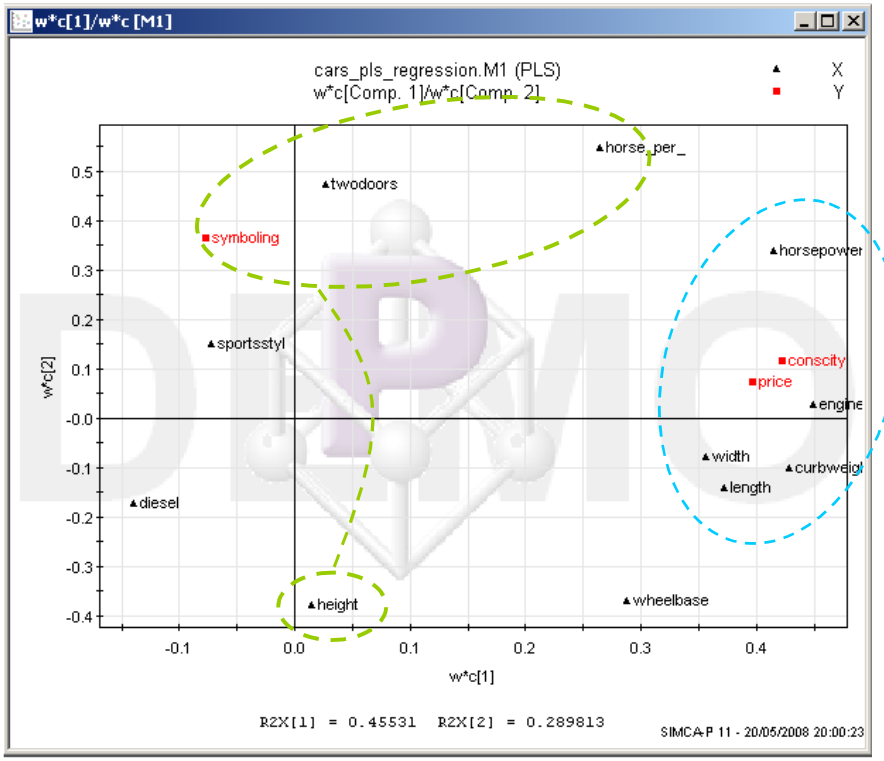
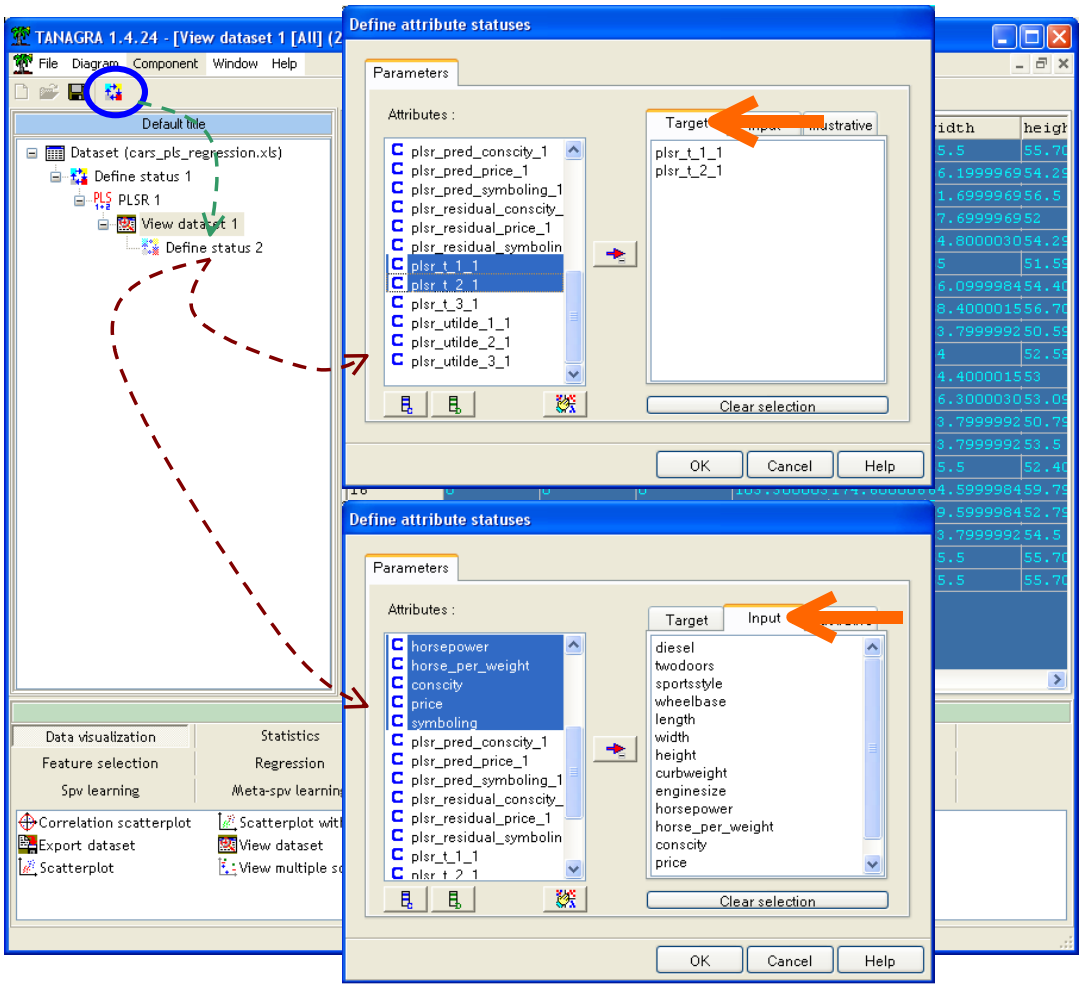


Figure 16 - Cartes des variables de SIMCA-P - w*c[1] vs. w*c[2]

Corrélations variables – facteurs : cercle des corrélations. Nous présentions les WEIGHTS et LOADINGS comme des grandeurs proportionnelles aux corrélations des variables sur les axes, tout du moins permettant de positionner les variables de la même manière sans que l'on ait besoin de produire explicitement ces corrélations. Mais puisque les facteurs sont automatiquement calculés, nous pouvons positionner les variables sur les axes factoriels en construisant le véritable cercle des corrélations. Voyons ce qu'il en est et comparons le résultat avec la carte ci-dessus (Figure 16).

Pour construire le cercle de corrélation dans Tanagra, nous devons tout d'abord spécifier les facteurs à placer en abscisse et en ordonnée, puis les variables que nous voulons intégrer dans le graphique. Nous utilisons pour cela le composant DEFINE STATUS que nous insérons à la suite de PLSR1. Nous plaçons en TARGET les deux premières variables latentes, en INPUT les variables de l'étude (DIESEL…SYMBOLING).

Note : Le composant étant assez générique, nous aurions pu intégrer dans le graphique des variables illustratives n'ayant pas initialement participé aux calculs.



Il ne nous reste plus qu'à insérer à la suite le composant CORRELATION SCATTERPLOT (onglet DATA VISUALIZATION).

En termes de positionnement relatif des variables, le cercle des corrélations (Figure 17) rappelle singulièrement la carte des variables vue précédemment (Figure 16). Ce qui confirme l'intérêt des tableaux des Ph, Ch et Wh* pour l'interprétation des résultats.

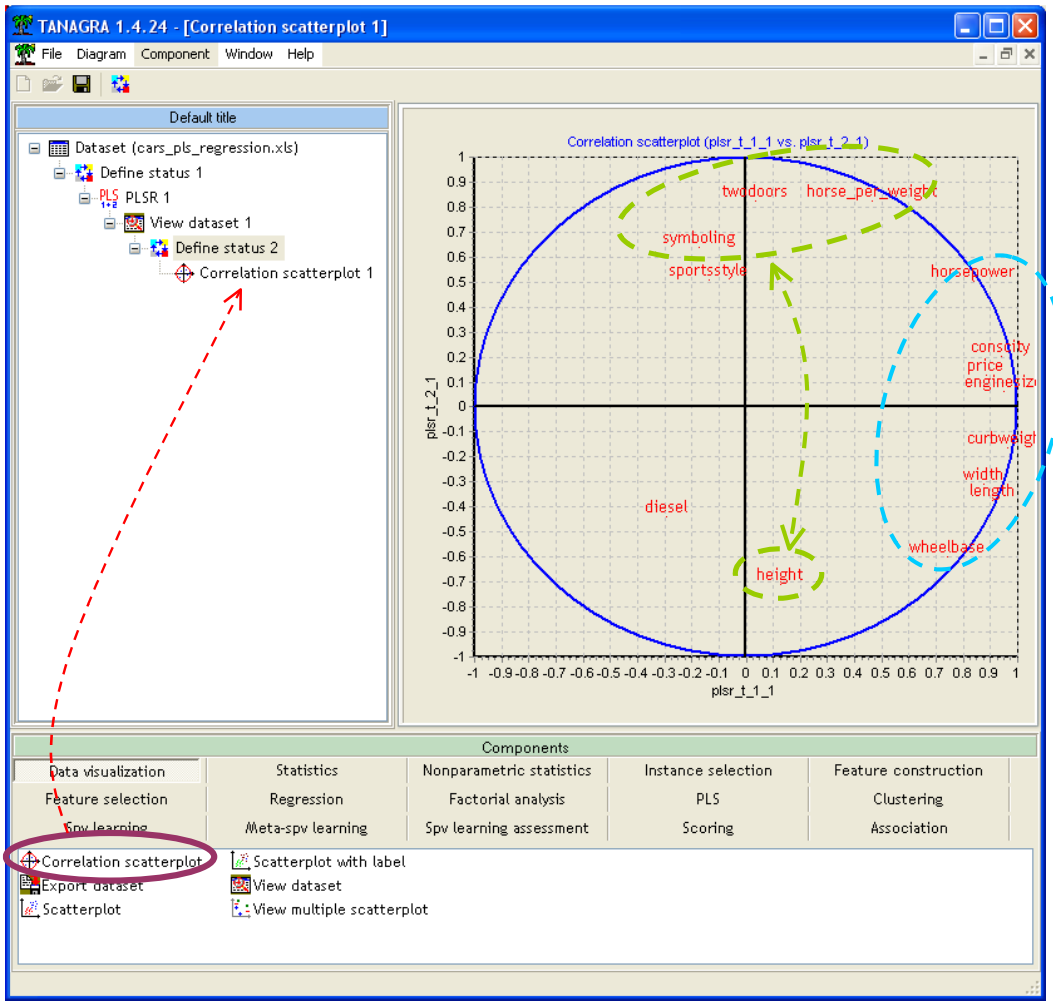


Figure 17 - Cercle des corrélations - 1er plan factoriel

Représentation des individus dans les plans factoriels. Un aspect très instructif des méthodes factorielles est la possibilité de positionner les individus, d'analyser des proximités, voire de délimiter des regroupements. Essayons de reproduire quelques graphiques présentés dans l'ouvrage de Tenenhaus (pages 150 à 156).

Pour construire un graphique nuage de points, nous insérons le composant SCATTERPLOT WITH LABEL (onglet DATA VISUALIZATION) dans le diagramme. Nous pouvons choisir interactivement les facteurs positionnés en abscisse et en ordonnée. Les données sont étiquetées par les numéros d'observations.

Nous pouvons ainsi produire successivement :

- Le plan des composantes (t1, t2) (Figure 18). On observe le rôle important des véhicules n°3 et n°17 sur le premier axe, des véhicules puissants et imposants ; de la voiture n°6 sur le second axe, une sacrée teigneuse si on se réfère à ses caractéristiques.

- Le plan des composantes (t1, $\widetilde{u}_1$) (Figure 19). On note principalement que la première variable latente explique plutôt bien le premier facteur Y. Ce que nous avions déjà constaté dans le tableau des proportions de variances expliquées. Nous remarquerons cependant que la

relation est dominée par la position extrême des observations n°3 et n°17. Sur un aussi petit effectif, ce type de situation est inévitable. Quelques points pèsent fortement sur les résultats.

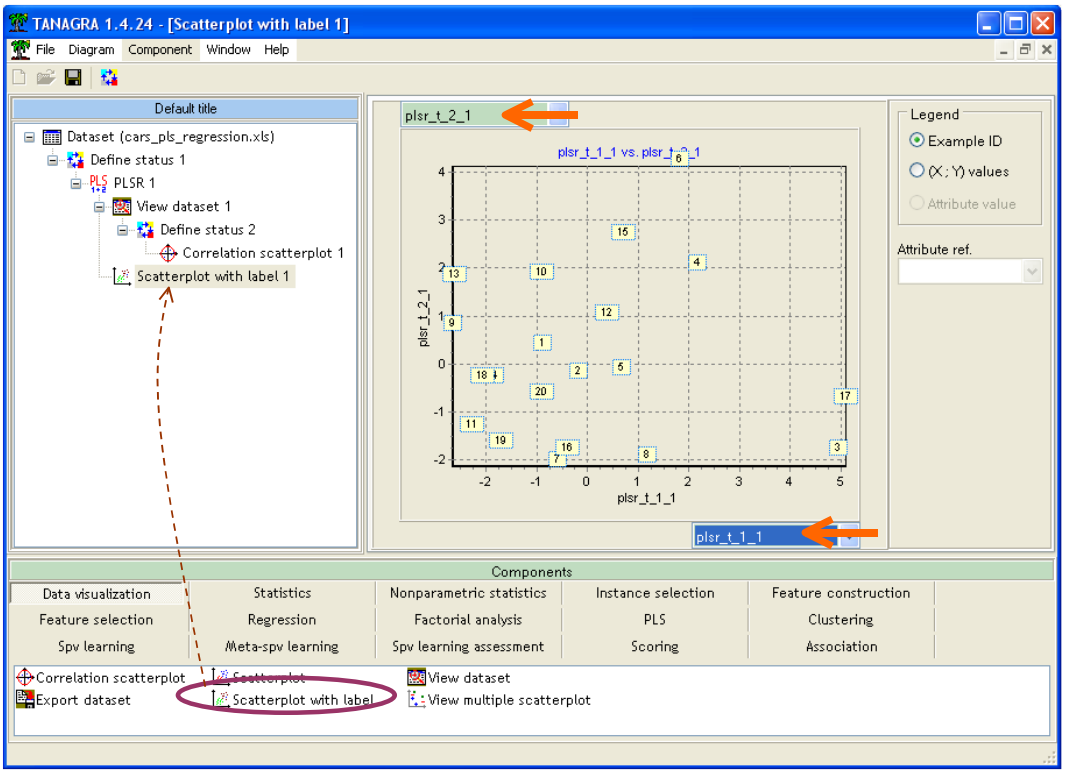


Figure 18 - Graphique des composantes (t1, t2)

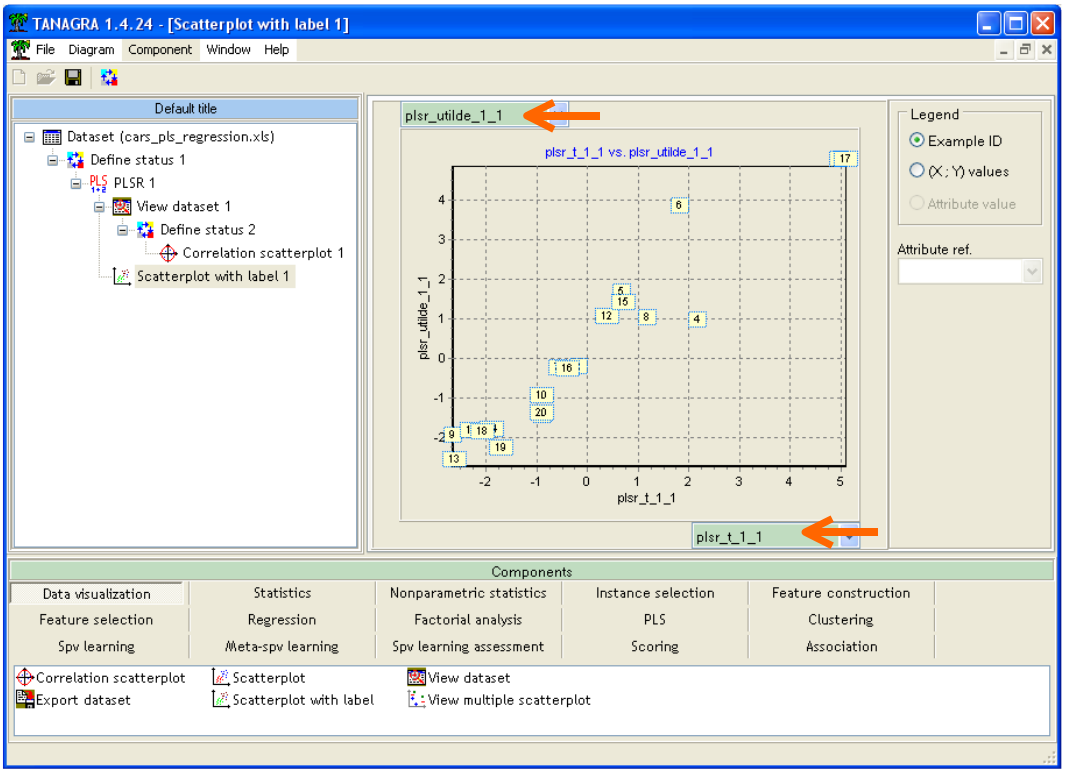
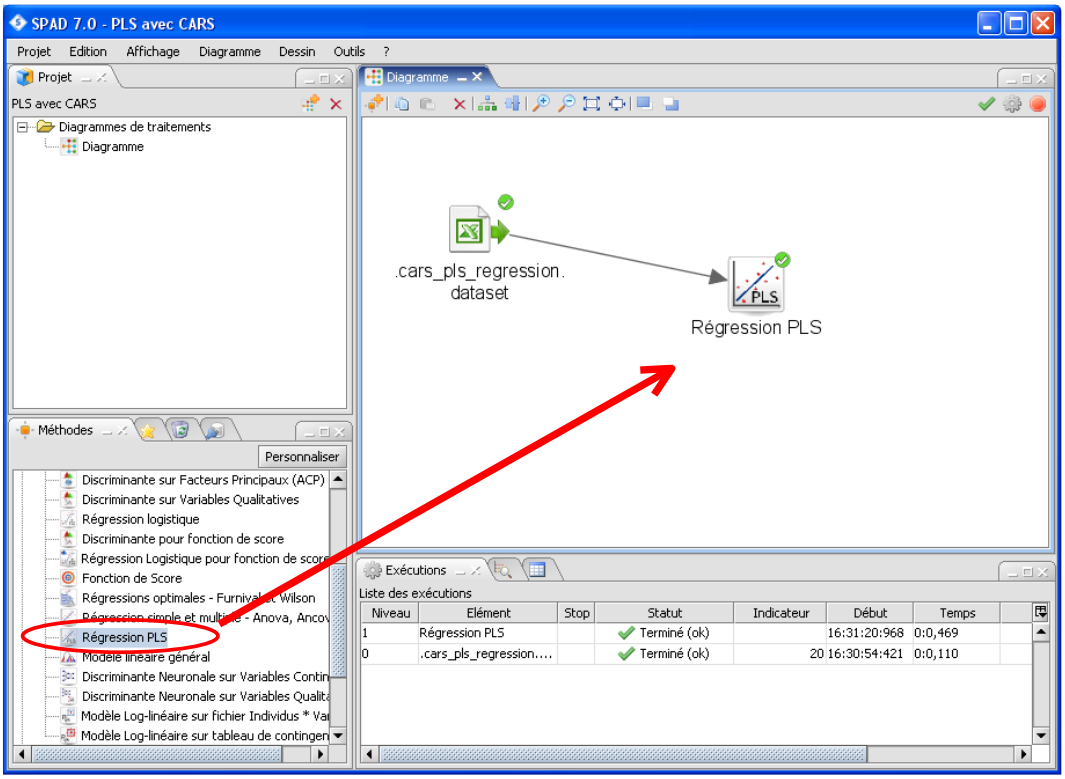


Figure 19 - Graphique des composantes (t1, $\widetilde{u}_1$)

5 Régression PLS avec les autres logiciels

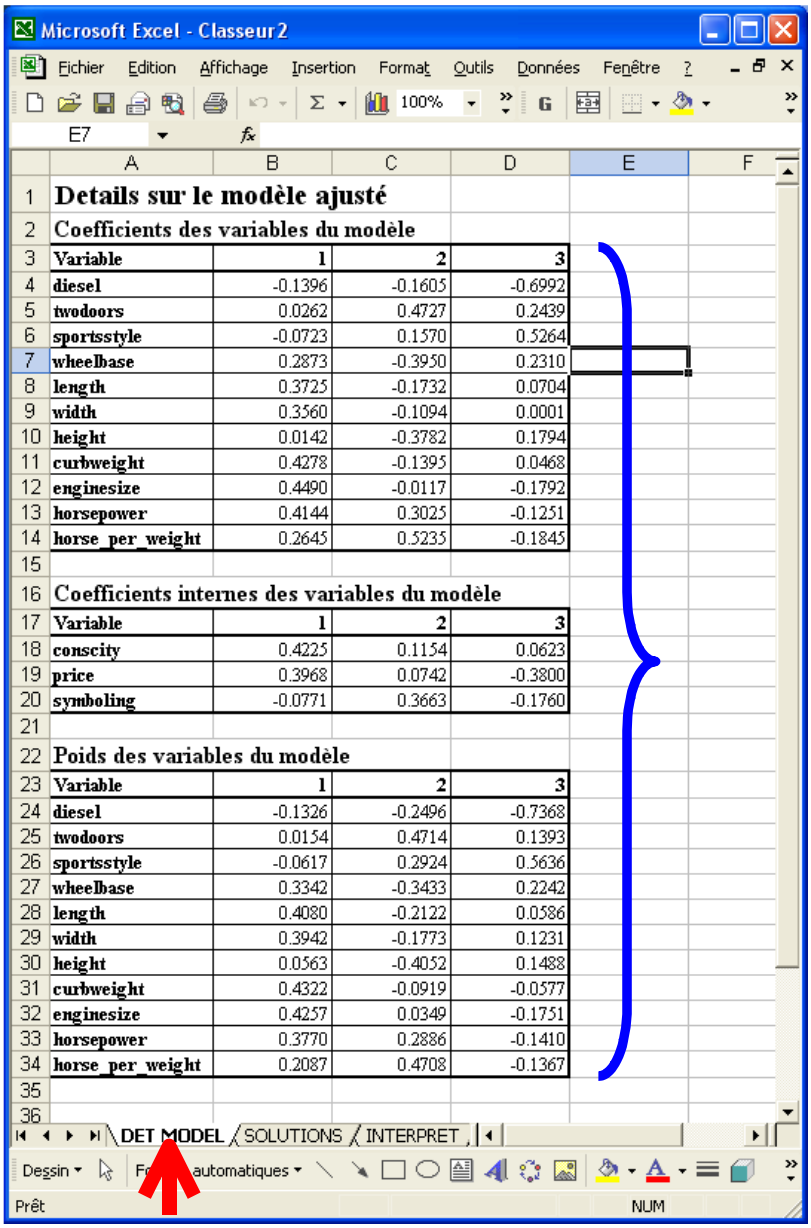
5.1 Régression PLS avec SPAD

SPAD, la version 7.0 que nous avons testée en tous les cas, intègre un composant « Régression PLS ». Nous avons élaboré la filière : elle importe directement le classeur EXCEL puis implémente la régression. Ci-dessous la fenêtre principale de l'application.



Les tableaux de résultats peuvent être consultés automatiquement dans un nouveau classeur EXCEL généré automatiquement. Nous retrouvons les résultats usuels de la régression PLS répartis dans différentes feuilles. **Les valeurs proposées sont en adéquation avec les résultats fournis par Tanagra et SIMCA-P**. Il est vraisemblable que ces logiciels développent de manière similaire les mêmes algorithmes.

Dans la feuille **DET MODEL**, nous observons : « Coefficients des variables du modèle », ils correspondent au vecteur Wh c.-à-d. les WEIGTHS des variables explicatives (à rapprocher avec la Figure 8) ; « Coefficients internes des variables du modèle » sont les vecteurs Ch c.-à-d. les WEIGHTS des variables cibles (cf. Figure 7); et enfin, « Poids des variables du modèles » sont les LOADINGS des explicatives c.-à-d. les vecteurs Ph (cf. Figure 6).



Microsoft Excel - Classeur2

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

E7

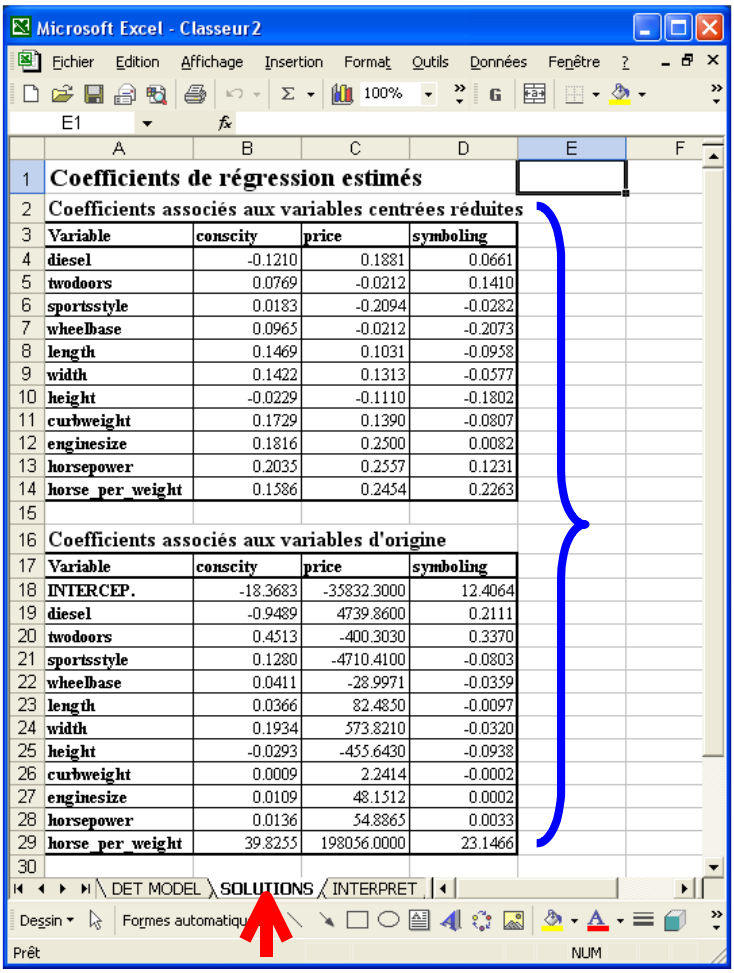
	A	B	C	D	E	F
1	**Details sur le modèle ajusté**					
2	**Coefficients des variables du modèle**					
3	**Variable**	**1**	**2**	**3**		
4	**diesel**	-0.1396	-0.1605	-0.6992		
5	**twodoors**	0.0262	0.4727	0.2439		
6	**sportsstyle**	-0.0723	0.1570	0.5264		
7	**wheelbase**	0.2873	-0.3950	0.2310		
8	**length**	0.3725	-0.1732	0.0704		
9	**width**	0.3560	-0.1094	0.0001		
10	**height**	0.0142	-0.3782	0.1794		
11	**curbweight**	0.4278	-0.1395	0.0468		
12	**enginesize**	0.4490	-0.0117	-0.1792		
13	**horsepower**	0.4144	0.3025	-0.1251		
14	**horse_per_weight**	0.2645	0.5235	-0.1845		
15						
16	**Coefficients internes des variables du modèle**					
17	**Variable**	**1**	**2**	**3**		
18	**conscity**	0.4225	0.1154	0.0623		
19	**price**	0.3968	0.0742	-0.3800		
20	**symboling**	-0.0771	0.3663	-0.1760		
21						
22	**Poids des variables du modèle**					
23	**Variable**	**1**	**2**	**3**		
24	**diesel**	-0.1326	-0.2496	-0.7368		
25	**twodoors**	0.0154	0.4714	0.1393		
26	**sportsstyle**	-0.0617	0.2924	0.5636		
27	**wheelbase**	0.3342	-0.3433	0.2242		
28	**length**	0.4080	-0.2122	0.0586		
29	**width**	0.3942	-0.1773	0.1231		
30	**height**	0.0563	-0.4052	0.1488		
31	**curbweight**	0.4322	-0.0919	-0.0577		
32	**enginesize**	0.4257	0.0349	-0.1751		
33	**horsepower**	0.3770	0.2886	-0.1410		
34	**horse_per_weight**	0.2087	0.4708	-0.1367		
35						
36						

DET MODEL | SOLUTIONS | INTERPRET

Dessin Formes automatiques

Prêt NUM

Dans la feuille **SOLUTIONS**, nous obtenons les paramètres estimés de la régression : « Coefficients associés aux variables centrées réduites » sont les coefficients standardisés (cf. Figure 11) ; « Coefficients associés aux variables d'origine » correspondent aux coefficients non standardisés (cf. Figure 10).



Microsoft Excel - Classeur2

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

E1

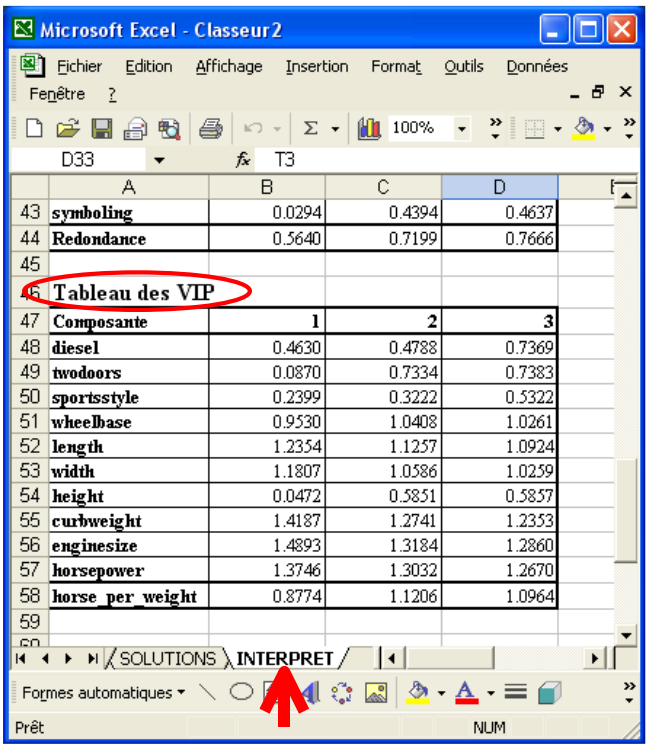
	A	B	C	D	E	F
1	**Coefficients de régression estimés**					
2	**Coefficients associés aux variables centrées réduites**					
3	**Variable**	**conscity**	**price**	**symboling**		
4	**diesel**	-0.1210	0.1881	0.0661		
5	**twodoors**	0.0769	-0.0212	0.1410		
6	**sportsstyle**	0.0183	-0.2094	-0.0282		
7	**wheelbase**	0.0965	-0.0212	-0.2073		
8	**length**	0.1469	0.1031	-0.0958		
9	**width**	0.1422	0.1313	-0.0577		
10	**height**	-0.0229	-0.1110	-0.1802		
11	**curbweight**	0.1729	0.1390	-0.0807		
12	**enginesize**	0.1816	0.2500	0.0082		
13	**horsepower**	0.2035	0.2557	0.1231		
14	**horse_per_weight**	0.1586	0.2454	0.2263		
15						
16	**Coefficients associés aux variables d'origine**					
17	**Variable**	**conscity**	**price**	**symboling**		
18	**INTERCEP.**	-18.3683	-35832.3000	12.4064		
19	**diesel**	-0.9489	4739.8600	0.2111		
20	**twodoors**	0.4513	-400.3030	0.3370		
21	**sportsstyle**	0.1280	-4710.4100	-0.0803		
22	**wheelbase**	0.0411	-28.9971	-0.0359		
23	**length**	0.0366	82.4850	-0.0097		
24	**width**	0.1934	573.8210	-0.0320		
25	**height**	-0.0293	-455.6430	-0.0938		
26	**curbweight**	0.0009	2.2414	-0.0002		
27	**enginesize**	0.0109	48.1512	0.0002		
28	**horsepower**	0.0136	54.8865	0.0033		
29	**horse_per_weight**	39.8255	198056.0000	23.1466		
30						

DET MODEL SOLUTIONS INTERPRET

Dessin Formes automatiques

Prêt NUM

Enfin, la feuille **INTERPRET** correspond aux tableaux d'aides à l'interprétation. Nous retrouvons la séquence proposée dans Tenenhaus (pages 145 à 146), ci-dessous les VIP (cf. Figure 9).



Microsoft Excel - Classeur2

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

D33 T3

	A	B	C	D
43	**symboling**	0.0294	0.4394	0.4637
44	**Redondance**	0.5640	0.7199	0.7666
45				
46	**Tableau des VIP**			
47	**Composante**	**1**	**2**	**3**
48	**diesel**	0.4630	0.4788	0.7369
49	**twodoors**	0.0870	0.7334	0.7383
50	**sportsstyle**	0.2399	0.3222	0.5322
51	**wheelbase**	0.9530	1.0408	1.0261
52	**length**	1.2354	1.1257	1.0924
53	**width**	1.1807	1.0586	1.0259
54	**height**	0.0472	0.5851	0.5857
55	**curbweight**	1.4187	1.2741	1.2353
56	**enginesize**	1.4893	1.3184	1.2860
57	**horsepower**	1.3746	1.3032	1.2670
58	**horse_per_weight**	0.8774	1.1206	1.0964
59				

SOLUTIONS INTERPRET

Formes automatiques

Prêt NUM

5.2 Régression PLS avec SAS

Nous avons lancé la PROC PLS de SAS, version 9.1. La documentation du logiciel est accessible en ligne (http://support.sas.com/91doc/docMainpage.jsp, faire Rechercher : PROC PLS).

La PROC PLS est très riche. Pour obtenir des résultats comparables aux nôtres, il faut faire attention aux options de calcul, à savoir METHOD = PLS et ALGORITHM = NIPALS. Ci-dessous la syntaxe de notre commande. L'option DETAILS permet de produire les affichages détaillés, OUTPUT est destinée à créer les nouvelles variables (prévisions, résidus et scores), NFAC correspond au nombre de facteurs demandés.

```
%let cost = conscity price symboling;
%let charac = diesel twodoors sportsstyle wheelbase length width height curbweight
enginesize horsepower horse_per_weight;
proc pls data=cars method=pls (algorithm=nipals) details nfac=3;
model &cost=&charac;
output out=pls        predicted=predY1-predY3
                      yresidual=resY1-resY3
                      xscore=xsc1-xsc3
                      yscore=ysc1-ysc3;
run;
```

Premier affichage, la proportion de variance expliquée par les facteurs, individuellement et cumulativement, pour les variables explicatives et expliquées (à rapprocher avec Figure 2 et Figure 3). **Les valeurs sont identiques à ceux de Tanagra et SIMCA-P**.

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors

Number of Extracted Factors	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	45.5310	45.5310	56.4009	56.4009
2	28.9813	74.5123	15.5845	71.9855
3	7.4150	81.9273	4.6715	76.6570

Puis viennent les LOADINGS des variables explicatives. Contrairement à Tanagra et SIMCA-P (Figure 6), les facteurs sont en ligne, les variables en colonnes. Le tableau est subdivisé en sous blocs.

Number of Extracted Factors	diesel	twodoors	sportsstyle	wheelbase	length	width
1	-0.131855	0.015262	-0.061349	0.332197	0.405519	0.391885
2	-0.244389	0.461515	0.286265	-0.336150	-0.207791	-0.173595
3	-0.721275	0.136361	0.551743	0.219469	0.057399	0.120507

Model Effect Loadings

Number of Extracted Factors	height	curbweight	enginesize	horsepower	horse_per_ weight
1	0.055946	0.429607	0.423140	0.374797	0.207439
2	-0.396712	-0.090016	0.034193	0.282591	0.460990
3	0.145641	-0.056496	-0.171430	-0.138054	-0.133844

Si ce n'était qu'un problème de représentation, on pourrait s'en accommoder facilement. Plus ennuyeux ici est que l'on constate des différences assez sensibles avec les valeurs proposées par Tanagra et SIMCA-P. Comment comprendre ces divergences ? Les autres indicateurs (WEIGHTS) sont à l'avenant. Les poids des variables cibles ci-dessous posent réellement problème (cf. Figure 7), les écarts ne peuvent pas être dus à des disparités de précision des calculs.

Dependent Variable Weights

Number of Extracted Factors	conscity	price	symboling
1	0.722639	0.678529	-0.131878
2	0.295016	0.189718	0.936468
3	0.147212	-0.897485	-0.415751

Passé le moment de perplexité, on se rend compte que **SAS propose en réalité des poids (et loadings) normalisés**. En effectuant les corrections, les résultats sont totalement cohérents.

Prenons le premier facteur, on constate après vérifications que $0.722369^2 + 0.678529^2 + (-0.131878)^2 = 1$. Si nous normalisons nos propres coefficients (cf. Figure 7), nous retrouvons les coefficients produits par SAS (**ouf !**).

Tous ces logiciels réalisent les mêmes calculs. Mais les choix de présentation sont différents. Il faut en être conscient simplement. De toute manière, à partir du moment où les proportions de variance expliquée par les facteurs étaient les mêmes d'un logiciel à l'autre, les disparités ne pouvaient être imputables qu'à des choix de mise en forme.

Note : Assez étrangement, il n'est pas possible d'obtenir les coefficients des droites de régression avec les options que nous avions utilisées. Il faut spécifier METHOD = SIMPLS pour que SAS produise les coefficients standardisés. Les valeurs proposées sont proches des coefficients obtenus avec la méthode NIPALS des autres logiciels.

5.3 Régression PLS avec R – Le package PLS

Nous avons utilisé la version 2.6.0 de R (http://www.r-project.org/). Nous avons fait appel au package PLS (http://cran.r-project.org/web/packages/pls/index.html). La documentation disponible est un peu élusive, la bonne référence est certainement l'article publié par les auteurs de la librairie dans « Journal of Statistical Software » (http://www.jstatsoft.org/v18/i02).

Nous avons introduit les commandes suivantes pour lancer la régression PLS dans R.

```
#clear all
rm(list=ls())

#************************
#some packages
#************************
#xls data file handling
library(xlsReadWrite)
#pls regression
library(pls)

#downloading the dataset
setwd("directory of the dataset")
cars.data                                                    <-
read.xls(file="cars_pls_regression.xls",rowNames=FALSE,sheet=1)
#checking the variables
```

```
summary(cars.data)
#subdivide the dataset into matrix Y and X
Y <- as.matrix(cars.data[,12:14])
X <- as.matrix(cars.data[,1:11])

#pls regression : 3 axes, nipals
cars.pls <- mvr(Y ~ X, ncomp = 3, method = "oscorespls", scale = TRUE)
summary(cars.pls)
```

Un premier affichage vient résumer les résultats de la régression PLS. Ici nous avons des vraies divergences par rapport aux autres logiciels. En effet, les proportions de variance expliquées par les facteurs ne sont pas les mêmes. Détallons cela.

```
> summary(cars.pls)
Data:   X dimension: 20 11
        Y dimension: 20 3
Fit method: oscorespls
Number of components considered: 3
TRAINING: % variance explained
           1 comps  2 comps  3 comps
X           45.333    66.30    81.73   [1]
conscity    87.988    89.17    91.94
price       80.482    88.47    92.72   [2]
symboling    2.546    33.95    41.78
```

Les proportions cumulées de variance des variables explicatives retranscrites par les facteurs [1] sont (45.33%, 66.30% et 81.73%), assez différents des proportions produites par les autres logiciels (cf. par exemple Figure 3) où l'on a plutôt (45.53%, 74.51%, 81.93%). L'écart est surtout problématique pour le second axe factoriel.

Si l'on s'intéresse maintenant à la part de variance expliquée cumulativement par les 3 premiers facteurs pour chaque variable cible (3ème colonne de [2]), les valeurs (CONSCITY : 91.94%, PRICE : 92.72%, SYMBOLING : 41.78%) sont assez discordantes par rapport aux autres logiciels où nous avons respectivement (92.73%, 90.87%, 46.37%) (cf. Figure 5).

D'ores et déjà on peut imaginer que les ratios (LOADINGS, WEIGHT) et coefficients de régression seront également différents. Listons quand même les commandes adéquates.

```
#loadings for X
loadings(cars.pls)

#weights for X
loading.weights(cars.pls)

#weights for Y
Yloadings(cars.pls)

#regression coefficients
coef(cars.pls)

#prediction
fitted(cars.pls)
```

```
#residuals
residuals(cars.pls)
```

Observons les paramètres des équations de régression (à rapprocher avec Figure 11).

```
> coef(cars.pls)
, , 3 comps

                    conscity       price    symboling
diesel           -0.26017692  1557.55068  0.002923431
twodoors          0.24626479  -889.76672  0.149642873
sportsstyle       0.06821250 -1820.00863  0.057453989
wheelbase         0.27648511   -34.11222 -0.186698496
length            0.44138223   462.08244 -0.119974109
width             0.41972747   578.93918 -0.115984611
height           -0.09896129 -1246.64032 -0.203604184
curbweight        0.50771044  1634.46156 -0.045387443
enginesize        0.51005416  3122.95037  0.029842667
horsepower        0.57798583  2488.51319  0.121556434
horse_per_weight  0.43161789  2200.81123  0.198713642
```

Au regard des coefficients de PRICE, on imagine bien qu'il s'agit d'un problème de normalisation. Les coefficients sont manifestement exprimés dans des unités différentes. Nous avons un début de réponse dans l'article de présentation du package (http://www.jstatsoft.org/v18/i02/paper ; dernier paragraphe, page 3), les auteurs indiquent qu'il existe une multitude de normalisations possibles durant les calculs. Ce qui rend les comparaisons difficiles. Malgré plusieurs tentatives, les passerelles vers les autres logiciels n'ont pas été trouvées[7].

6 Conclusion

Dans ce didacticiel, nous introduisons le composant PLSR (version 1.4.24 de Tanagra) qui implémente la Régression PLS. Dérivé des outils déjà existants, nous avons surtout fait un effort de synthèse en fusionnant des fonctionnalités dispersées, puis en concevant une structure de rapports en accord avec les sorties des logiciels reconnus dans le domaine.

Le second aspect important de ce document est la validation des calculs. Elle passe (entre autres) par une comparaison approfondie de nos résultats avec ceux fournis par les autres logiciels. Il m'a semblé intéressant de montrer comment cela est organisé, quelles en sont les principales étapes. Dans la pratique, ce processus est réitéré sur plusieurs fichiers avant que la version du logiciel avec le nouveau le composant ne soit mise en ligne.

7 Note de mise à jour (23.11.2018)

Un internaute avait donné suite à mon interrogation ci-dessous (cf. note de bas de page). La correction portait bien sur les normalisations à effectuer pour orienter les coefficients de R (package PLS) vers ceux de SIMCA-P et TANAGRA (Figure 11). Merci Grégori pour ton aide, je sais bien que tu

[7] R.R. : J'avoue avoir un peu beaucoup cherché/tenté, en vain. La documentation ne m'a pas permis de comprendre les divergences. Si un lecteur a une idée, je suis preneur.

m'as envoyé tes suggestions en 2015 et que je ne corrige le tutoriel qu'en 2018 mais bon, mieux vaut tard que jamais a-t-on coutume de dire.

Solution. En réalité, l'option scale de la fonction mvr() du package « pls » ne porte que sur les explicatives X. Pour retrouver nos résultats de référence (Figure 11), nous devons explicitement standardiser les colonnes constituant la matrice de variables expliquées Y. Voici le code R corrigé avec le résultat associé.

Importation des données et préparation des matrices

```
#la librairie "xlsReadWrite"" est obsolète aujourd'hui (en 2018)
#nous utilisons "xlsx""
library(xlsx)

#chargement
setwd("… votre dossier …")
cars <- read.xlsx("cars_pls_regression.xls",sheetIndex=1,header=TRUE)

#vérification
print(str(cars))
```

```
## 'data.frame':    20 obs. of  14 variables:
##  $ diesel          : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ twodoors        : num  1 0 0 1 0 1 0 0 0 1 ...
##  $ sportsstyle     : num  0 0 0 1 0 0 0 0 1 0 ...
##  $ wheelbase       : num  97.3 99.8 115.6 102.9 101.2 ...
##  $ length          : num  172 177 203 184 177 ...
##  $ width           : num  65.5 66.2 71.7 67.7 64.8 65 66.1 68.4 63.8 64 ...
##  $ height          : num  55.7 54.3 56.5 52 54.3 51.6 54.4 56.7 50.6 52.6 ...
##  $ curbweight      : num  2209 2337 3740 3016 2765 ...
##  $ enginesize      : num  109 109 234 171 164 194 134 120 90 98 ...
##  $ horsepower      : num  85 102 155 161 121 207 72 97 68 112 ...
##  $ horse_per_weight: num  0.0385 0.0436 0.0414 0.0534 0.0438 ...
##  $ conscity        : num  8.71 9.8 14.7 12.38 11.2 ...
##  $ price           : num  7975 13950 34184 15998 21105 ...
##  $ symboling       : num  2 2 -1 3 0 3 0 0 1 1 ...
## NULL
```

Préparation des matrices.

```
#isoler les variables expliquées (Y) et explicatives
Y <- as.matrix(cars[,12:14])
X <- as.matrix(cars[,1:11])
```

Le cœur de la correction est ici : nous standardisons explicitement les variables de Y via la fonction scale()

```
#standardisation de Y
Z <- scale(Y,scale=TRUE,center=TRUE)
print(head(Z))
```

```
##         conscity       price  symboling
## [1,] -0.46579135 -0.71541535  0.8548504
## [2,] -0.08667628 -0.06818896  0.8548504
## [3,]  1.61934155  2.12360668 -1.7097008
## [4,]  0.81122784  0.15365534  1.7097008
## [5,]  0.40075739  0.70685788 -0.8548504
## [6,]  1.31827958  2.10670838  1.7097008
```

Les 3 variables sont centrées et réduites.

Régression PLS

Nous pouvons lancer la régression maintenant. Il fallait lire attentivement la documentation : l'option scale ne porte que sur les X, pas sur les Y ! Le hic était là.

```
#librarie PLS
library(pls)

##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings

#régression
modele <- mvr(Z ~ X, method="oscorespls",ncomp=3,scale=TRUE)

#affichage des coefficients
print(coef(modele))

## , , 3 comps
##
##                     conscity       price   symboling
## diesel           -0.12066342  0.18606243  0.06516670
## twodoors          0.07680084 -0.02083768  0.14117163
## sportsstyle       0.01845594 -0.21045234 -0.02866247
## wheelbase         0.09581766 -0.01702724 -0.20537021
## length            0.14602576  0.10849312 -0.09326336
## width             0.14134433  0.13646841 -0.05531428
## height           -0.02296951 -0.11074767 -0.18011779
## curbweight        0.17190020  0.14519668 -0.07783768
## enginesize        0.18056845  0.25649056  0.01122586
## horsepower        0.20252515  0.26170901  0.12591390
## horse_per_weight  0.15793490  0.24926123  0.22812411
```

Nous obtenons des coefficients standardisés de la régression identiques à ceux de SIMCA-P et TANAGRA (Figure 11) !