1 Objectif

Calculer les corrélations semi-partielles avec Tanagra.

La régression linéaire multiple[1] vise à expliquer les valeurs d'une variable dépendante (Y) à l'aide d'une série de variables indépendantes ou explicatives (Z1, …, Zp). La corrélation semi-partielle quantifie le pouvoir explicatif additionnel d'une variable supplémentaire (X), une fois que nous lui avons retranché les informations déjà portées par les variables (Z1,…,Zp). Une manière simple de la calculer est de réaliser les 2 régressions, avec et sans la présence de X, l'écart entre les deux coefficients de détermination des régressions correspond au carré de la corrélation semi-partielle.

Une autre manière de la produire est de calculer les résidus de la régression de X sur (Z1, …, Zp). Ils correspondent à la fraction de X non expliquée par les variables indépendantes. La corrélation semi-partielle est obtenue en calculant le coefficient de corrélation de Pearson entre Y et la variable résiduelle. La nature asymétrique du processus apparaît clairement, l'appellation « corrélation semi-partielle » est pertinente de ce point de vue. On peut faire le parallèle avec la corrélation partielle qui, elle, est symétrique. En effet la corrélation est calculée sur les résidus de X/Z1,…,Zp et Y/Z1,…, Zp dans ce cas.

Dans ce didacticiel, nous montrons les différentes manières de produire la corrélation semi-partielle. Nous comparons les résultats avec le composant dédié de TANAGRA (SEMI-PARTIAL CORRELATION).

Les aspects théoriques en relation avec ce didacticiel sont disponibles dans un support de cours accessible en ligne http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf (chapitre 5). Nous reprenons d'ailleurs l'exemple illustratif qui y est développé.

2 Données

Nous cherchons à prédire la consommation de véhicules (Y : CONSUMPTION) à partir de la puissance (X : HORSEPOWER), la cylindrée (Z1 : ENGINE.SIZE) et le poids (Z2 : WEIGHT). L'objectif est de déterminer l'apport d'information de HORESPOWER par rapport aux autres variables explicatives.

3 Obtenir la corrélation semi-partielle de différentes manières

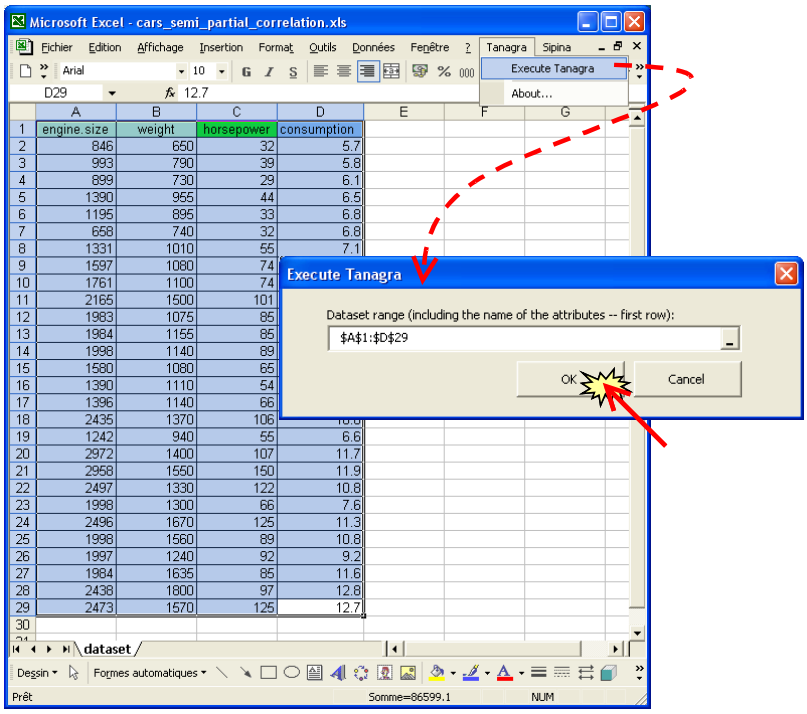
3.1 Importer les données

Le plus simple pour lancer Tanagra et charger les données est d'ouvrir le fichier XLS[2] dans le tableur EXCEL. Nous sélectionnons la plage de données. La première ligne doit correspondre au nom des variables. Puis nous activons le menu TANAGRA / EXECUTE TANAGRA qui a été installé avec la macro complémentaire TANAGRA.XLA[3]. Une boîte de dialogue apparaît. Nous vérifions la sélection. Si tout est en règle, nous validons en cliquant sur le bouton OK.

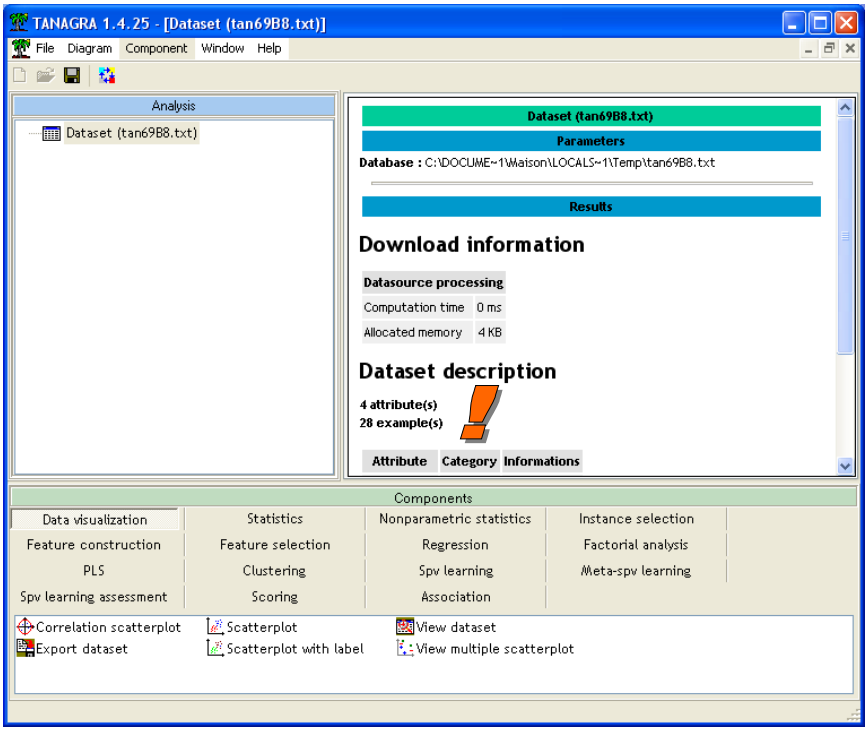
[1] http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

[2] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cars_semi_partial_correlation.xls

[3] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.



Les données importées comporte 28 observations et 4 variables.

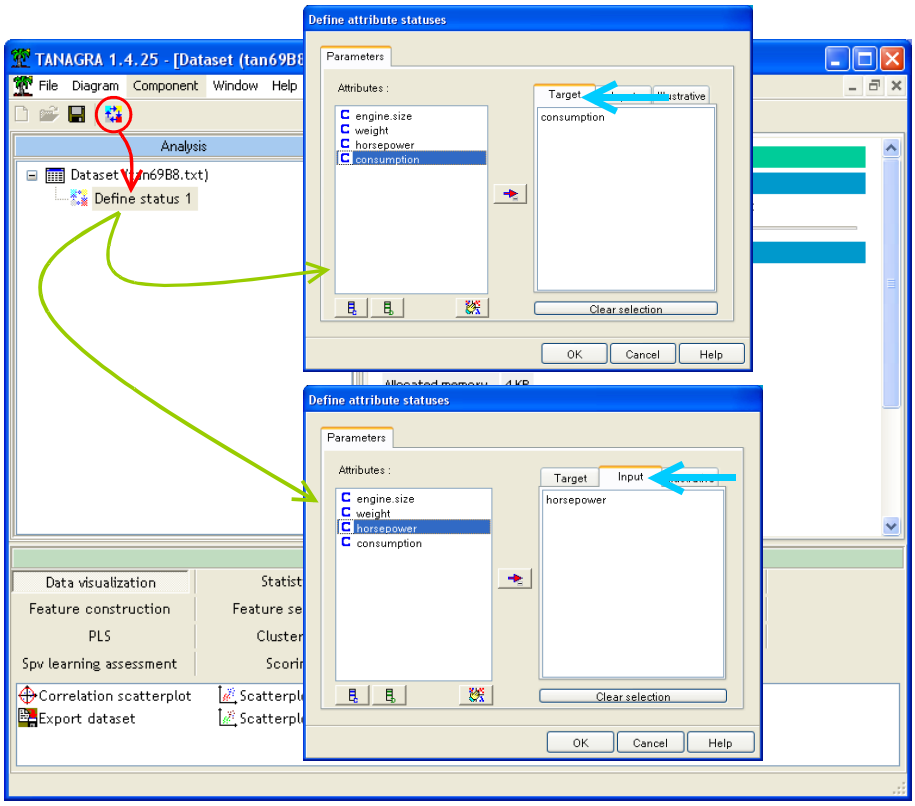


3.2 Régression simple et corrélation

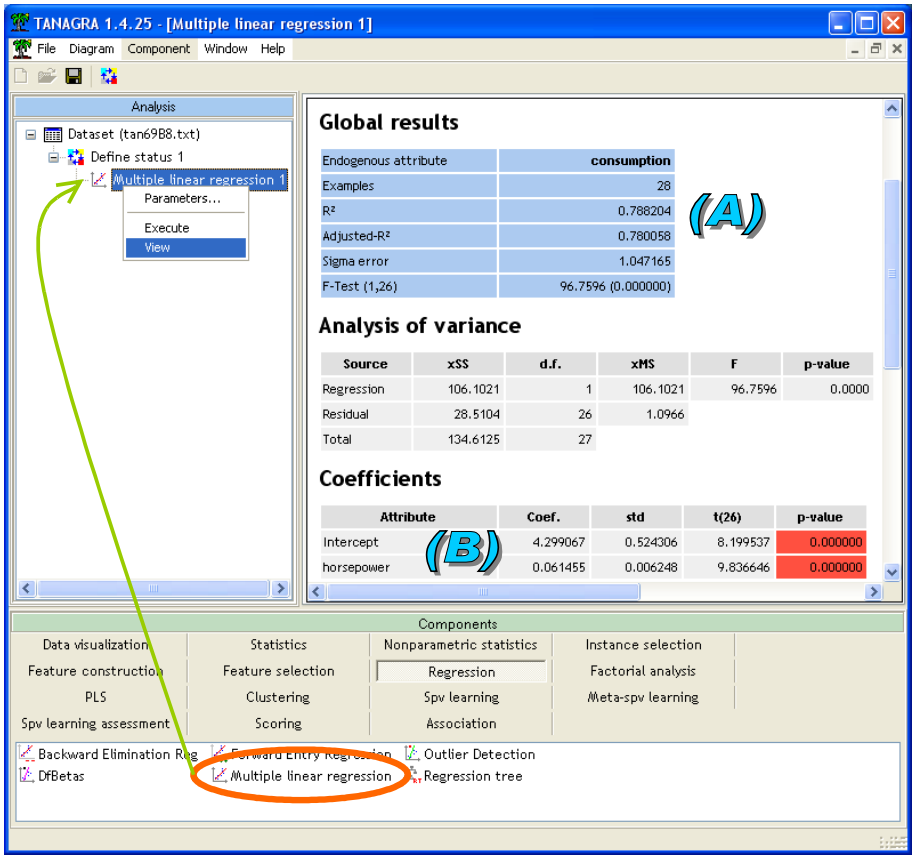
3.2.1 Régression simple et coefficient de détermination

Dans un premier temps, essayons de cerner le rôle de HORSEPOWER seul dans l'explication de CONSUMPTION. Pour cela, nous allons réaliser une régression simple.

Nous insérons le composant DEFINE STATUS dans le diagramme. Passer par le raccourci est le plus simple. Nous plaçons en TARGET la variable CONSUMPTION, en INPUT la variable HORSEPOWER.



Nous introduisons alors le composant MULTIPLE LINEAR REGRESSION (onglet REGRESSION). Nous actionnons le menu VIEW pour obtenir les résultats.



Deux résultats doivent retenir notre attention : (A) le coefficient de détermination $R^2 = 0.7882$ c.-à-d. 78.82% de la variance de Y est expliquée par la régression, c'est plutôt un bon résultat, la régression est globalement très significative avec un F = 96.7596 ; (B) la variable HORSEPOWER est bien entendu très significative avec un t de Student de 9.8366 et une p-value < 0.0001, le coefficient est positif, les voitures puissantes consomment plus.

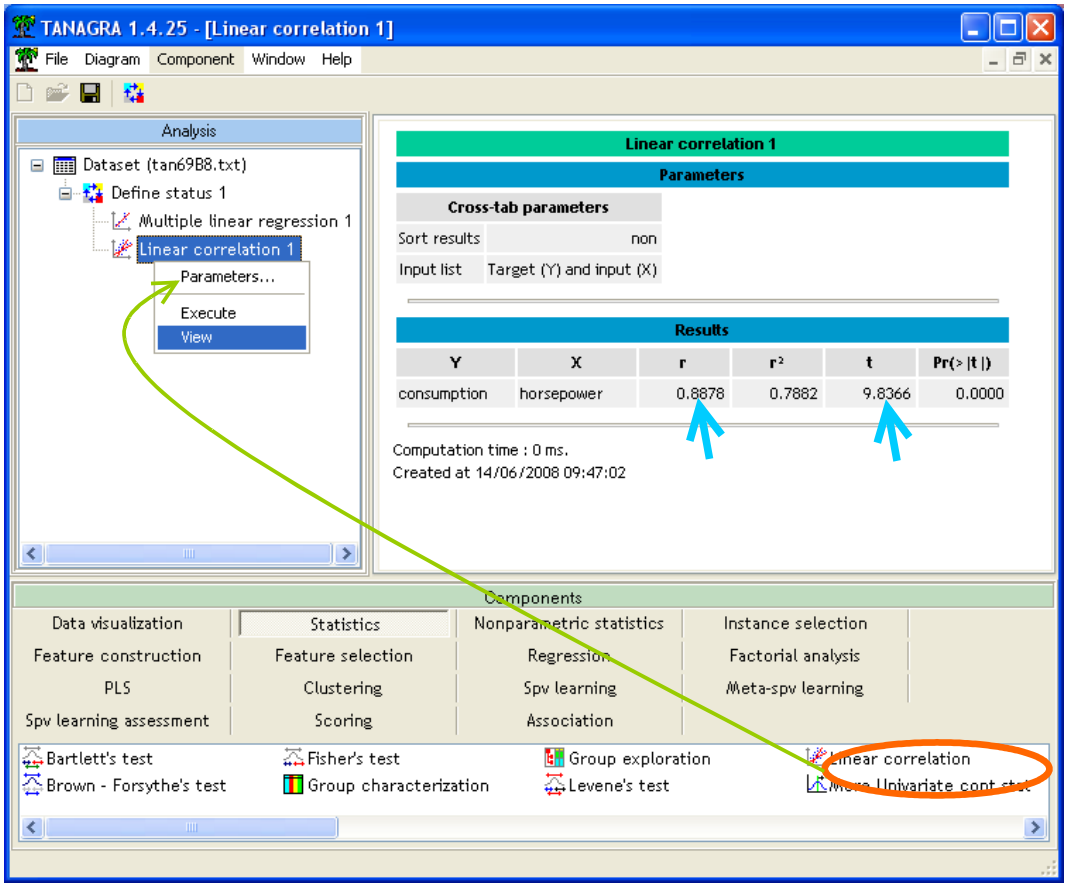
Dans la régression linéaire simple, le carré du t de Student, évaluation de la pente, est égal au F de Fisher, évaluation globale c.-à-d. $t^2 = (9.8366)^2 = 96.7596$.

C'est un résultat plutôt encourageant. On pourrait s'en tenir là, sauf qu'il y a deux autres variables dans le fichier. Il serait intéressant d'approfondir l'étude en analysant leur rôle éventuel dans l'explication de la consommation. Nous y viendrons plus loin (section 3.2.2).

3.2.2 Corrélation

Une autre manière d'analyser la liaison entre CONSUMPTION et HORSEPOWER est de calculer le coefficient de corrélation de Pearson. Son carré s'interprète également comme la proportion de variance expliquée.

Nous insérons le composant LINEAR CORRELATION (onglet STATISTICS) dans le diagramme. Nous actionnons le menu VIEW.



La corrélation est positive r = 0.8878, confirmant le signe du coefficient de la régression. Son carré est égal au coefficient de détermination de la régression simple $r^2 = 0.7882$. La corrélation est significative, on retrouve le t de Student du test de la pente de la régression t = 9.8366. Tester la pente de la régression simple et tester le coefficient de corrélation sont équivalents.

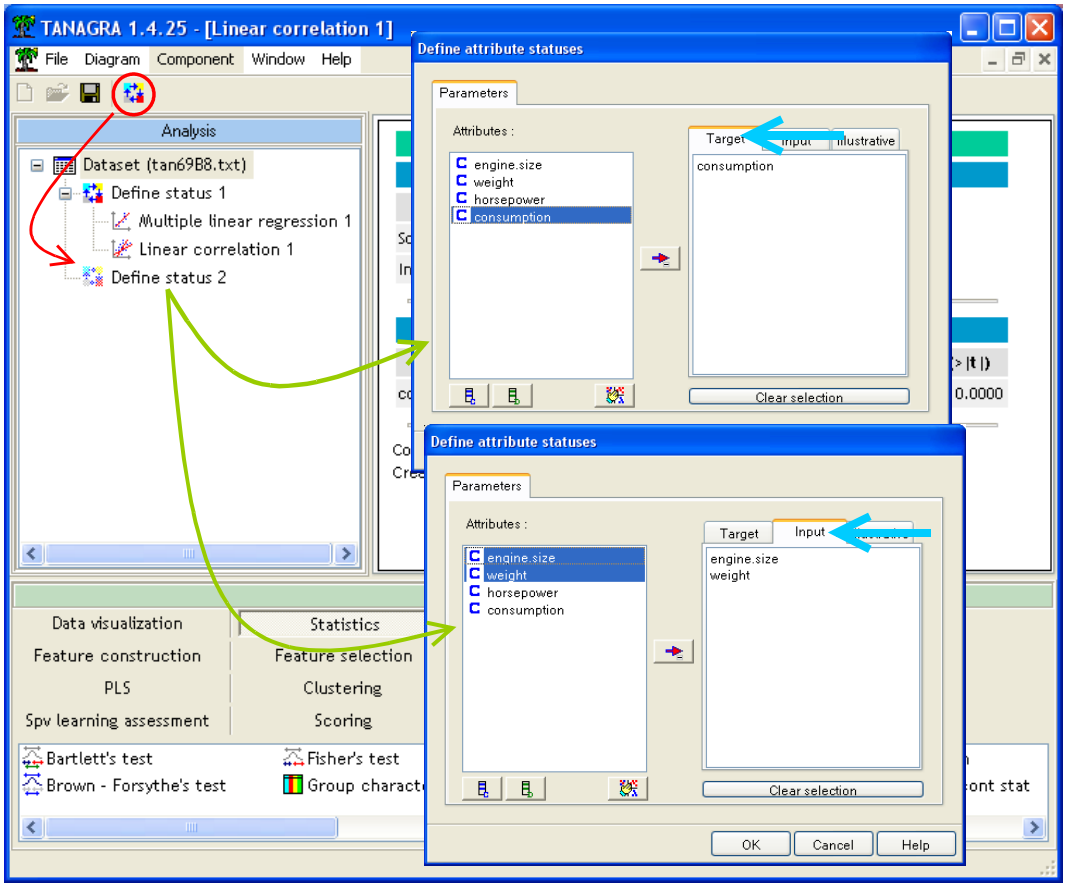
3.3 Comparaison de régressions

Nous ne pouvons pas en rester là. En effet, connaissant un peu les voitures, on se rend compte que des informations importantes sont mises de côté dans notre première analyse : la puissance est bien souvent déterminée par la cylindrée (taille du moteur), de même les puissantes berlines sont aussi des véhicules imposants, lourds. La question que l'on peut se poser est donc la suivante : par rapport à ENGINE.SIZE (Z1) et WEIGHT (Z2), est-ce qu'il y a dans HORSEPOWER (X) des informations supplémentaires qui pourraient expliquer la consommation ?

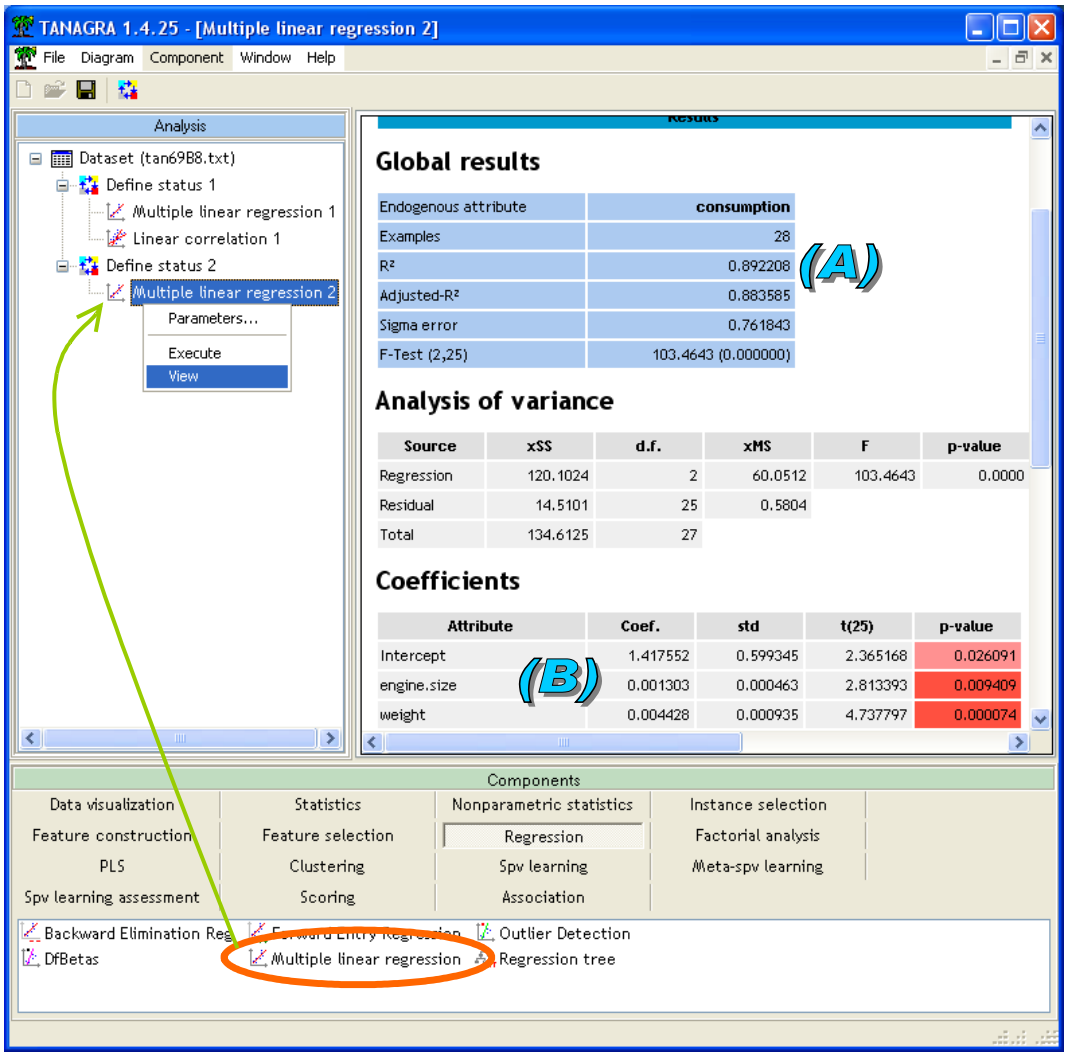
3.3.1 Régression Y / Z1, Z2

Voyons dans un premier temps dans quelle mesure les variables Z1 et Z2, que l'on appelle aussi variables de contrôle en référence à la corrélation partielle, expliquent la consommation des véhicules. Pour ce faire, nous allons mettre en place la régression multiple Y / Z1, Z2.

Nous introduisons un nouveau composant DEFINE STATUS dans le diagramme, nous plaçons CONSUMPTION en TARGET, ENGINE.SIZE et WEIGHT en INPUT.



Nous introduisons alors le composant de régression multiple.



Le coefficient de détermination $R^2 = 0.8922$ est élevé, la régression explique 89.22% de la variance de Y (A) ; les deux variables Z1 et Z2 sont significatives avec des p-value < 0.01 (B).

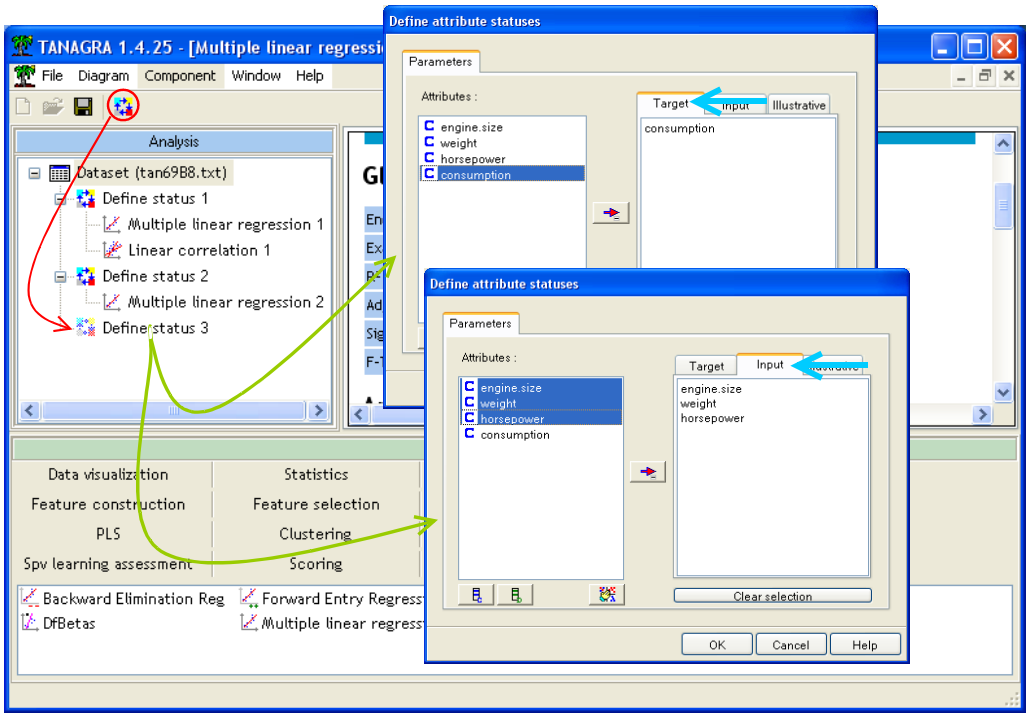
Ces deux variables conjuguées expliquent mieux la consommation ($R^2 = 0.8922$) que la seule puissance ($R^2 = 0.7882$)[4].

3.3.2 Régression Y / X, Z1, Z2

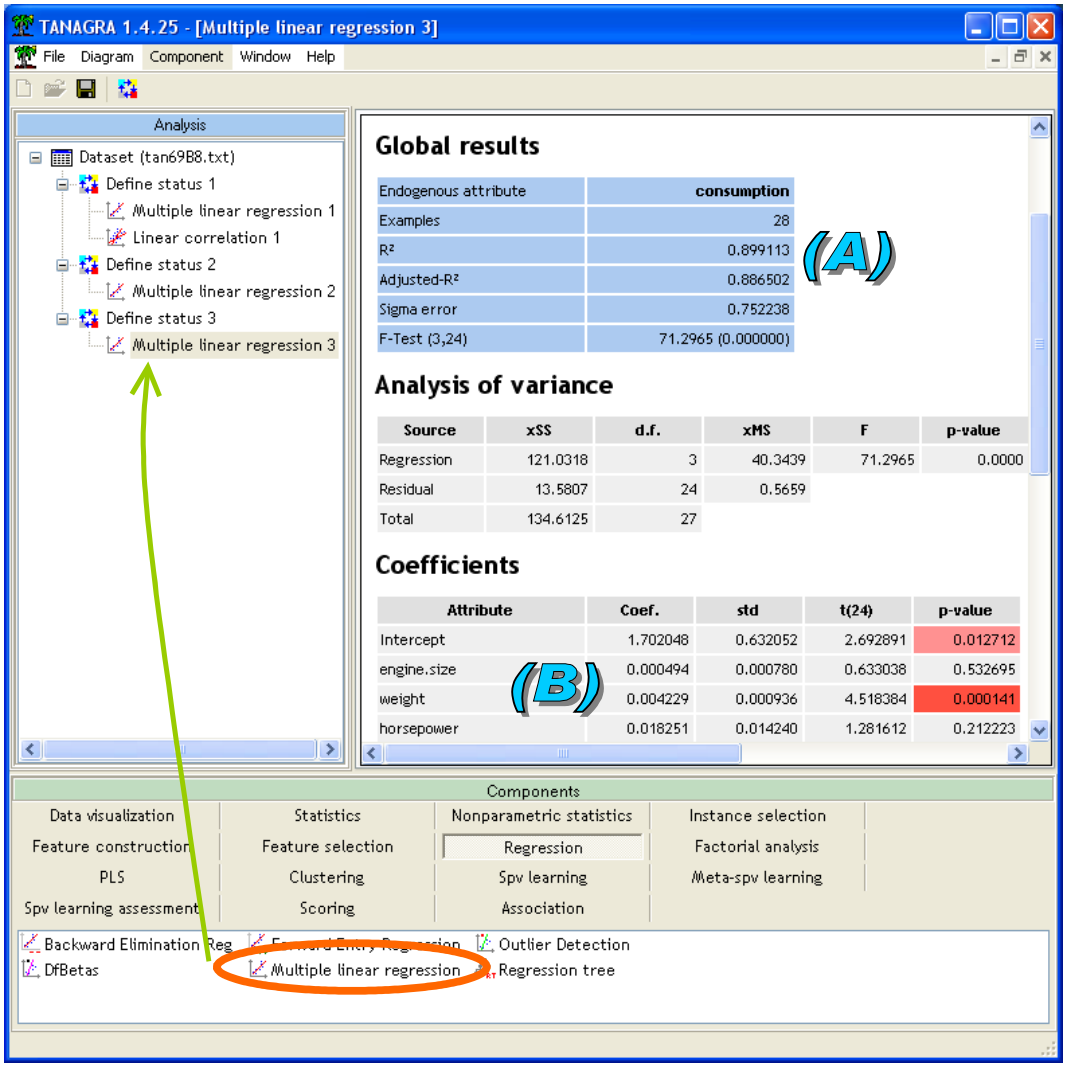
Etant entendu que l'on utilise les variables Z1 et Z2 dans la régression, est-ce l'introduction de X emmène de l'information supplémentaire ? Pour cela nous allons réaliser la régression Y / X, Z1, Z2 et nous allons comparer les coefficients de détermination. On sait par ailleurs que le R^2 de la régression à 3 variables sera supérieur ou égal à celui à 2 variables, les modèles étant imbriqués, ce qui importe donc, c'est d'évaluer l'ampleur de l'écart.

Insérons à nouveau le composant DEFINE STATUS dans le diagramme, nous plaçons en TARGET la variable CONSUMPTION, en INPUT les 3 autres variables.

[4] Restons prudent quand même par rapport à de telles comparaisons, les degrés de liberté ne sont pas les mêmes. Utiliser le R^2 ajusté serait déjà plus judicieux dans ce contexte.



Puis, nous introduisons la régression linéaire multiple.



Le coefficient de détermination est $R^2 = 0.899113$ (A). L'écart avec la régression précédente, $d^2 = 0.899113 - 0.892208 = 0.006905$, exprime l'information supplémentaire, non redondante, introduite par la variable X, par rapport à Z1 et Z2. **La racine carrée d = 0.0831 est le coefficient de corrélation semi-partielle**.

A première vue, la valeur semble assez faible. Par rapport à ENGINE.SIZE et WEIGHT, HORSEPOWER amène peu d'information intrinsèque pour expliquer la consommation. Nous verrons plus loin comment tester la significativité de la corrélation semi-partielle (section 4).

Assez curieusement, mis à part WEIGHT, les autres variables ne semblent pas significatives dans cette régression (B). Il faut surtout y voir l'effet nuisible de la colinéarité entre les variables explicatives.

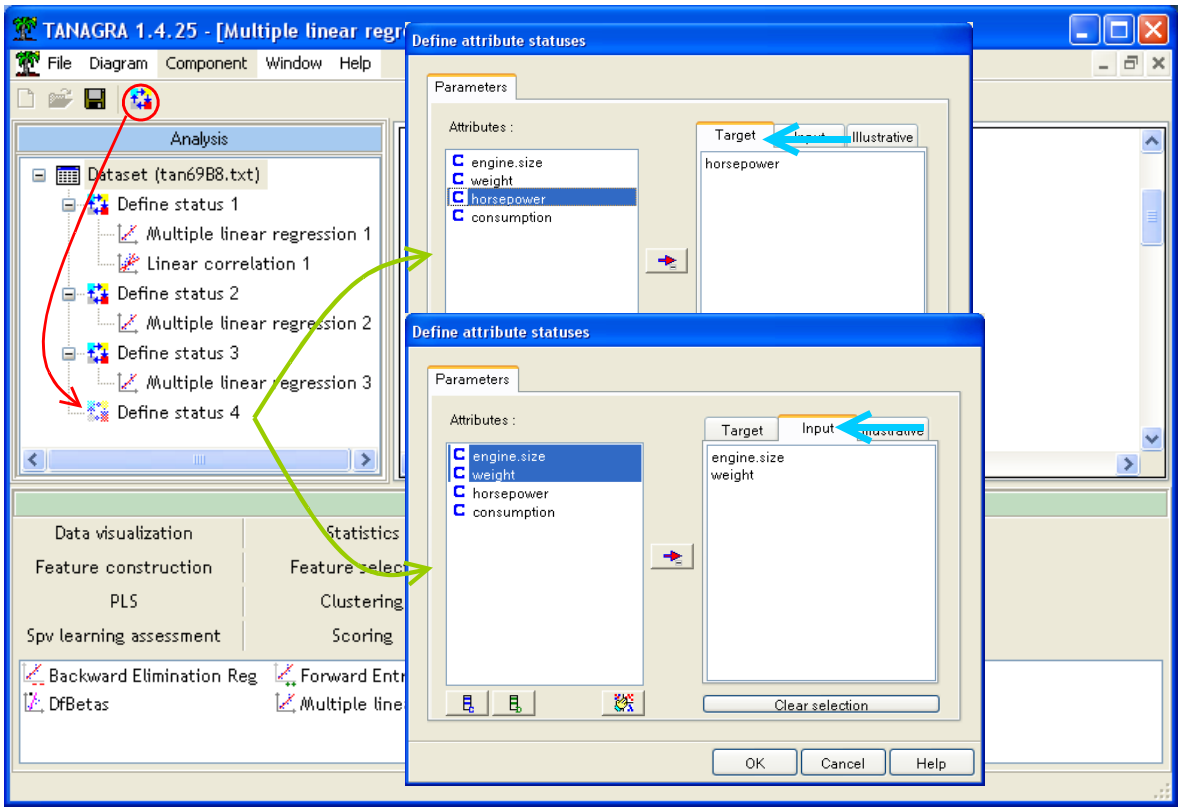
3.4 Résidus de X/Z1,…,Zp et corrélation de Pearson

Nous avons la possibilité de calculer la corrélation semi-partielle d'une autre manière.

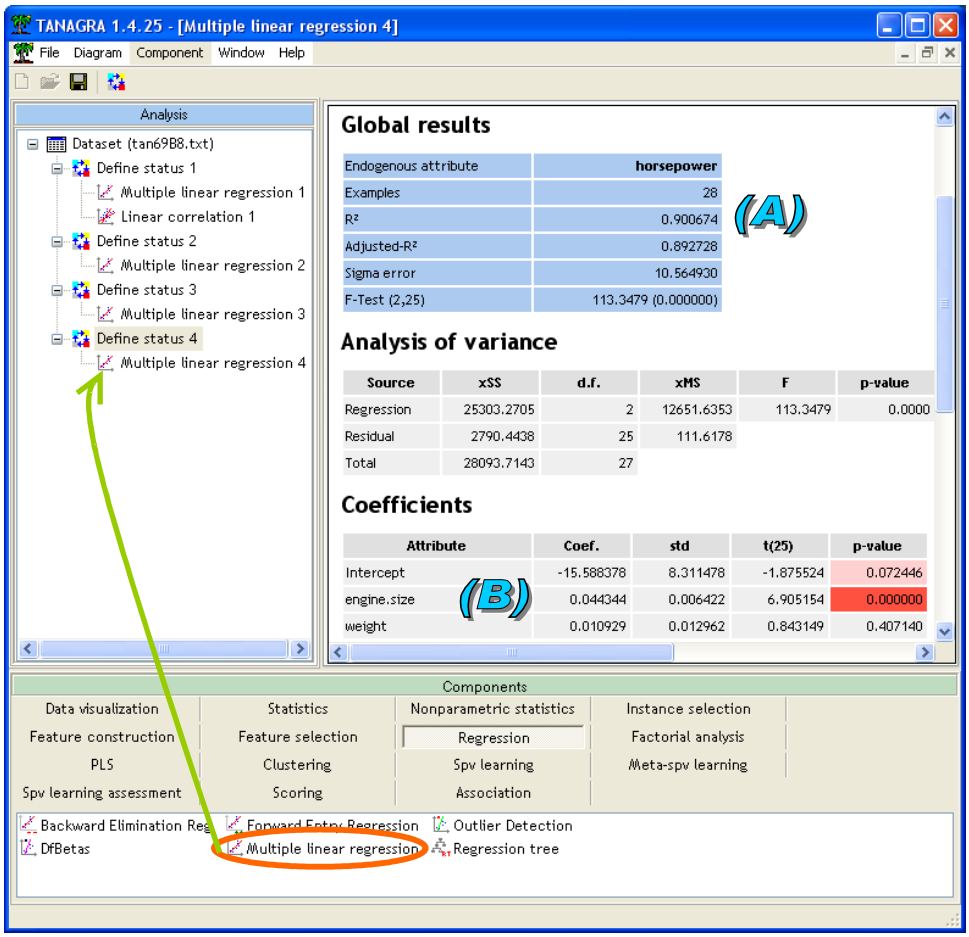
Dans un premier temps, nous calculons le résidu de la régression de X en Z1 et Z2, nous retranchons ainsi de X l'information portée par les variables de contrôle. Puis nous calculons la corrélation entre les résidus et la variable dépendante Y. Le coefficient obtenu est la corrélation semi-partielle.

3.4.1 Résidus de X/Z1,Z2

Pour obtenir ce résidu, nous devons tout d'abord réaliser la régression de X/Z1,Z2. Nous insérons le composant DEFINE STATUS dans le diagramme, nous plaçons HORSEPOWER en TARGET, ENGINE.SIZE et WEIGHT en INPUT.



Puis nous réalisons la régression.

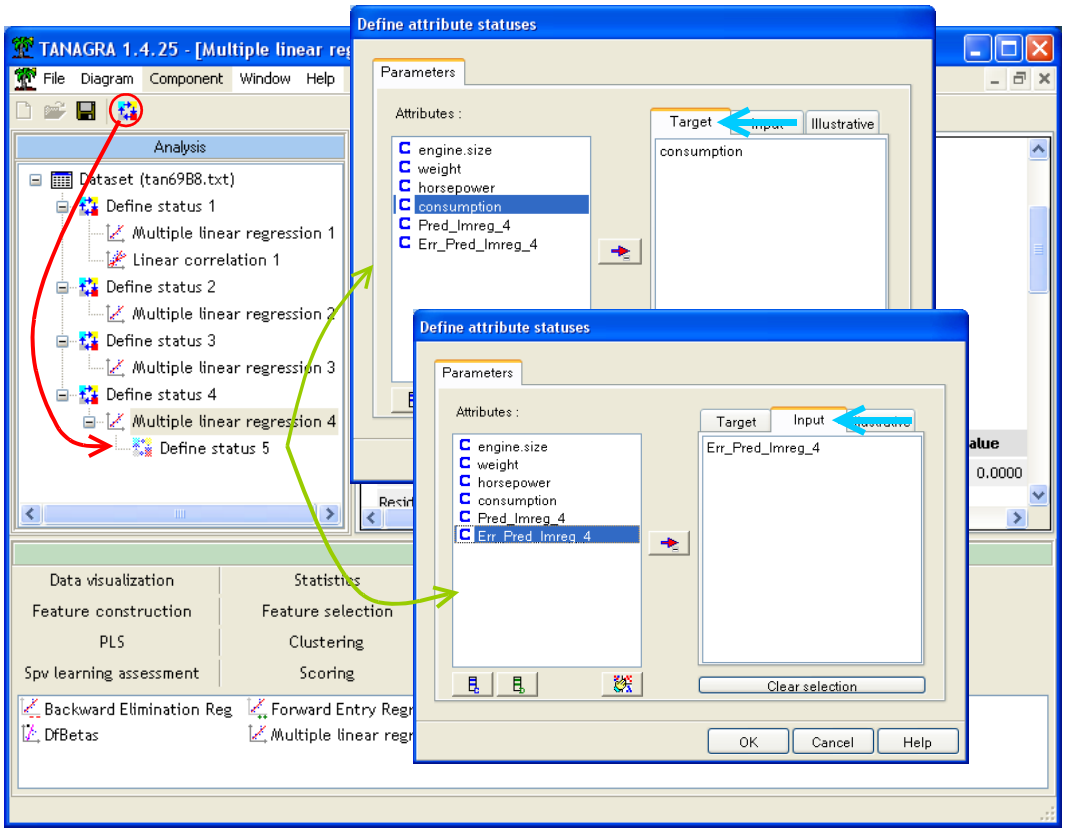


R^2 = 90.07% de la variance de HORSEPWER est expliquée par la régression (A), principalement par ENGINE.SIZE d'ailleurs si on en juge la significativité des coefficients (B). Manifestement HORSEPOWER est fortement redondante par rapport aux variables de contrôle.

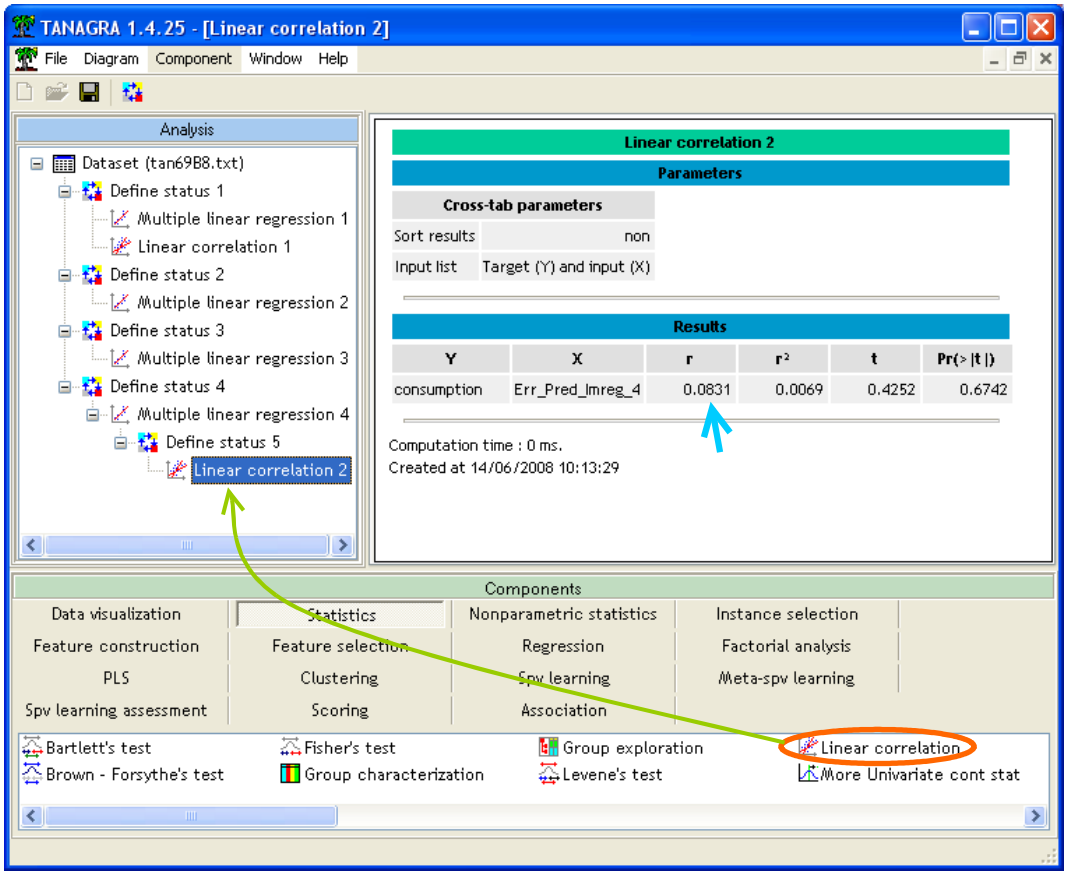
3.4.2 Corrélation résidus et variable dépendante Y

Le composant MULTIPLE LINEAR REGRESSION produit automatiquement deux nouvelles variables que l'on peut utiliser dans les branches subséquentes du diagramme : la prédiction de la variable dépendante et les résidus de la régression. Utilisons cette dernière maintenant.

Nous introduisons le composant DEFINE STATUS dans le diagramme, à la suite de la régression. Les deux nouvelles variables sont visibles. Nous plaçons en TARGET la variable CONSUMPTION, en INPUT le résidu ERR_PRED_LMREG_4.



Nous calculons la corrélation entre ces variables via le composant LINEAR CORRELATION.



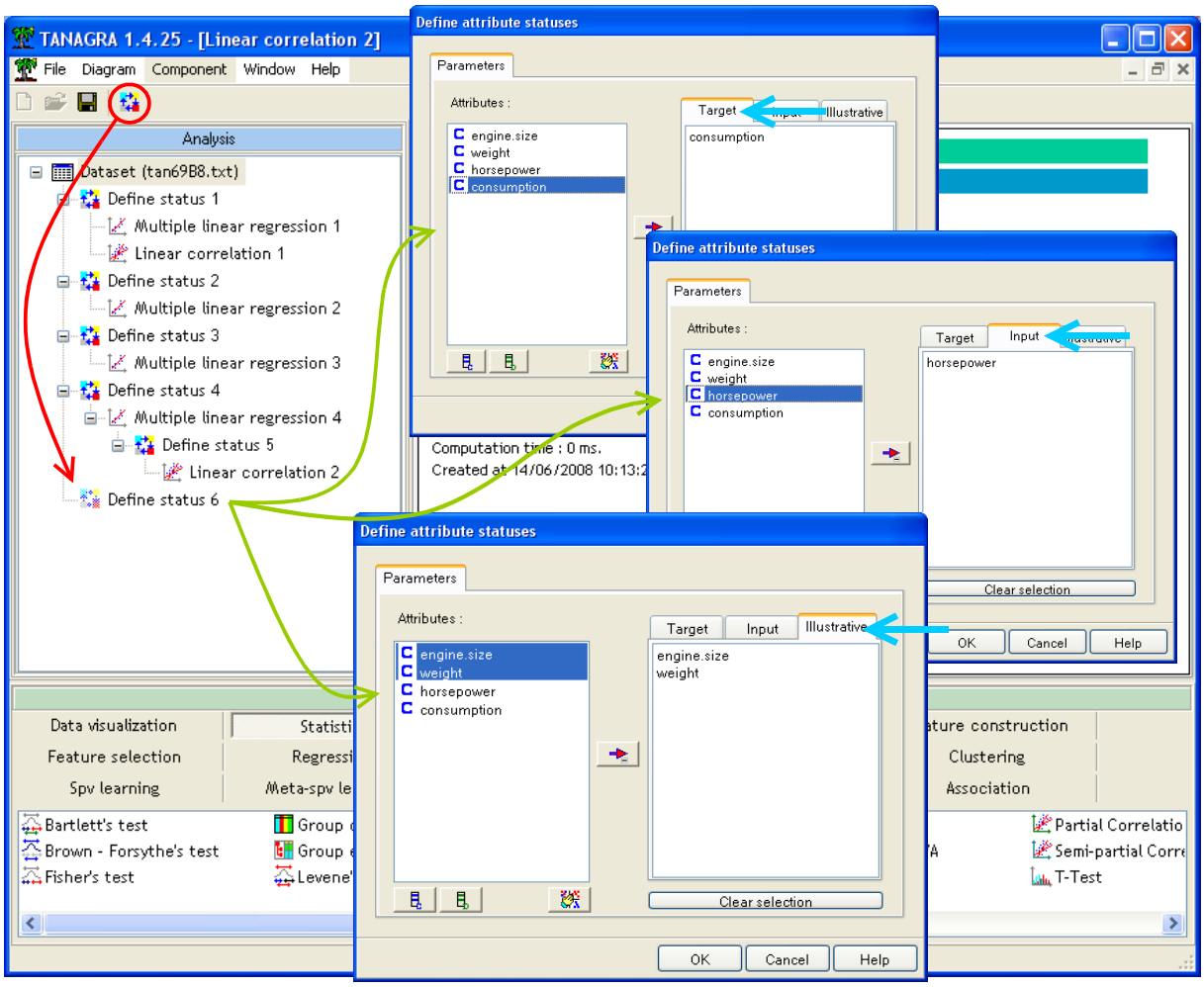
Le coefficient r = 0.0831 correspond à la corrélation semi-partielle. La valeur est cohérente avec l'approche précédente.

Notons qu'un test de significativité avec un **t de Student = 0.4252** est proposé. **Mais nous ne devons pas en tenir compte ici**. En effet, le composant ne sait pas qu'une des variables introduites dans la corrélation est le fruit d'un calcul ayant mis à contribution d'autres variables, les degrés de libertés calculés sont erronés, la valeur proposée est donc faussée.

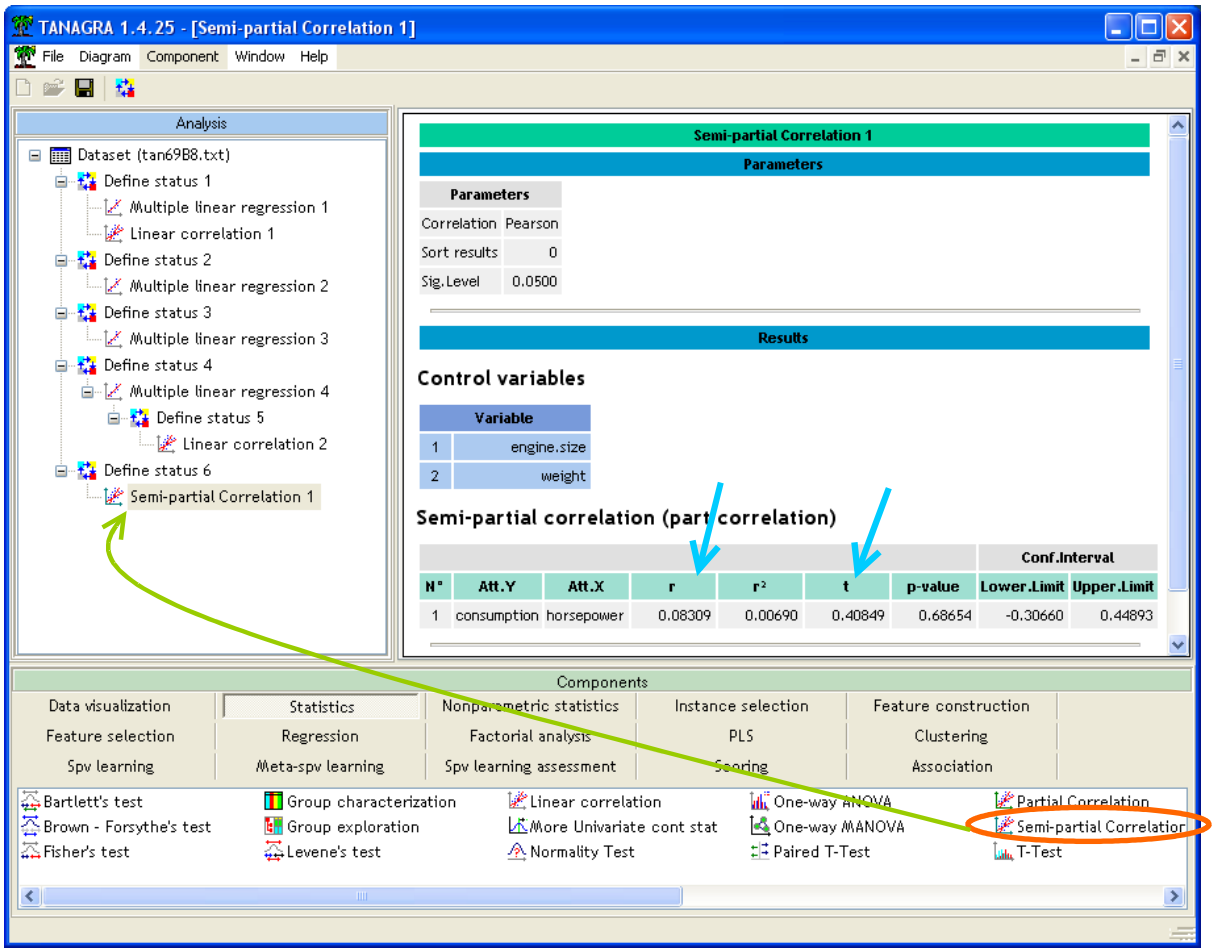
4 Utiliser le composant SEMI-PARTIAL CORRELATION

Dans cette partie de ce didacticiel, nous introduisons le composant dédié au calcul de la corrélation semi-partielle. Son avantage est double : (1) sa mise en œuvre est facilitée ; (2) les calculs étant explicitement définis, l'évaluation des degrés de libertés est correcte cette fois-ci.

Introduisons pour la dernière fois le composant DEFINE STATUS. Nous plaçons en TARGET la variable dépendante CONSUMPTION, en INPUT la variable à évaluer HORSEPOWER, **et en ILLUSTRATIVE, les variables de contrôle ENGINE.SIZE et WEIGHT**.



Nous insérons alors le composant SEMI-PARTIAL CORRELATION (onglet STATISTICS) dans le diagramme.



La corrélation semi-partielle est r = 0.0831. Les différentes approches sont cohérentes.

A la différence maintenant que nous pouvons tester la significativité du coefficient, le t de Student = 0.40849. Sous H0, il suit une loi de Student à (n – p – 2 = 28 – 2 – 2 = 24) degrés de liberté. Au risque 5%, la conclusion est que le coefficient n'est pas significatif du tout, avec une probabilité critique égale à 0.68654.

Par rapport à ENGINE.SIZE et WEIGHT, HORSEPOWER ne porte pas d'informations supplémentaires qui pourrait être utile à l'explication de la consommation.

5 Conclusion

Dans ce didacticiel, nous montrons les différentes manières de produire la corrélation semi-partielle dans Tanagra. Toutefois, seul le composant dédié (SEMI-PARTIAL CORRELATION) détermine directement les degrés de liberté adéquats pour le calcul des tests de significativité et des intervalles de confiance.