Objectif

Calculer le coefficient de corrélation linéaire pour un grand nombre de variables.

Calculer un indicateur et trier selon cet indicateur fait partie des tâches récurrentes de la fouille de données. Dans ce didacticiel, nous montrons comment mettre en place rapidement le calcul du coefficient de corrélation linéaire (1) d'une variable de référence avec une liste de variables, dans le cadre de la sélection de variables explicatives d'une régression par exemple ; (2) croisé entre une série de variables, cela peut être utilisé pour détecter les colinéarités entre variables explicatives.

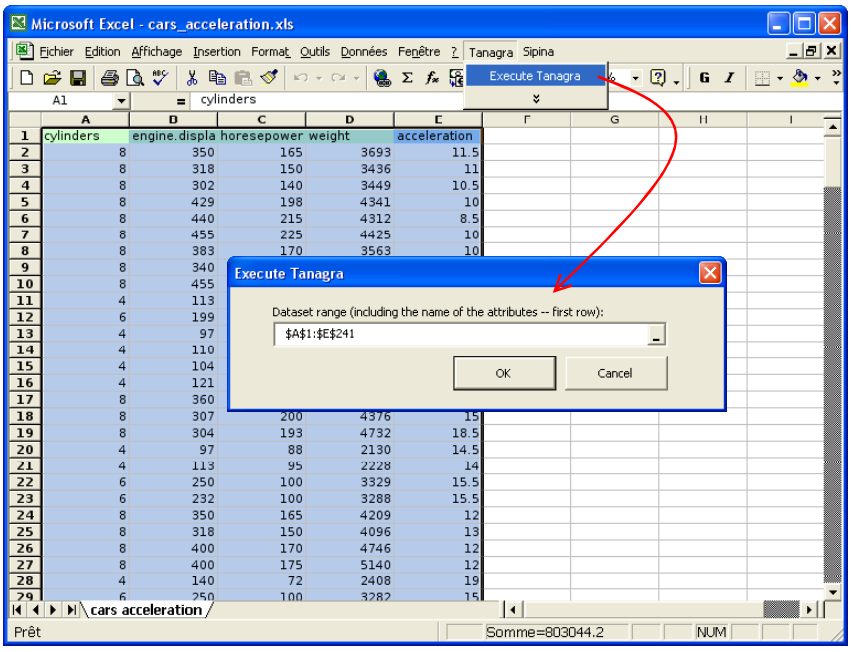
Données

Nous utilisons le fichier CARS_ACCELERATION.XLS. La variable d'intérêt est l'accélération qui indique le temps nécessaire pour parcourir une certaine distance, départ arrêté. Les variables, éventuellement explicatives, sont des descripteurs caractérisant les véhicules (poids, nombre de cylindres, etc.).

Coefficient de corrélation linéaire

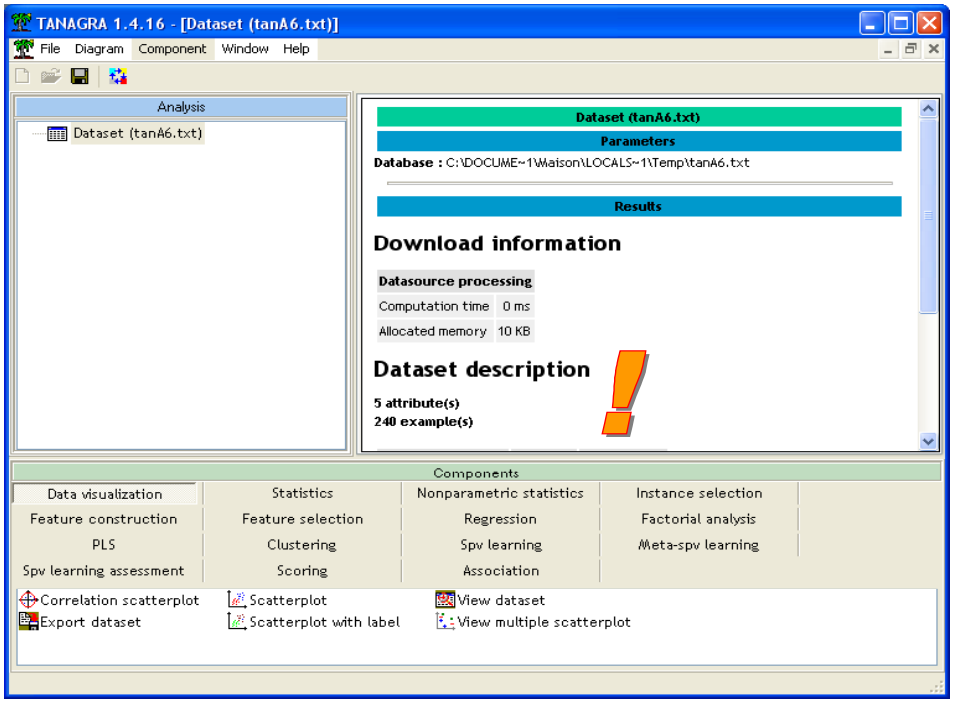
Créer un diagramme

Le plus simple maintenant[1] est de charger le fichier dans le tableur EXCEL puis d'utiliser la macro complémentaire TANAGRA.XLA. Pour ce faire, sélectionner les données et activer le nouveau menu TANAGRA/EXECUTE TANAGRA.



Valider la sélection dans la boîte de dialogue qui apparaît si elle est correcte. Les données sont maintenant chargées dans une nouvelle session de travail de TANAGRA, un nouveau diagramme a été automatiquement créé.

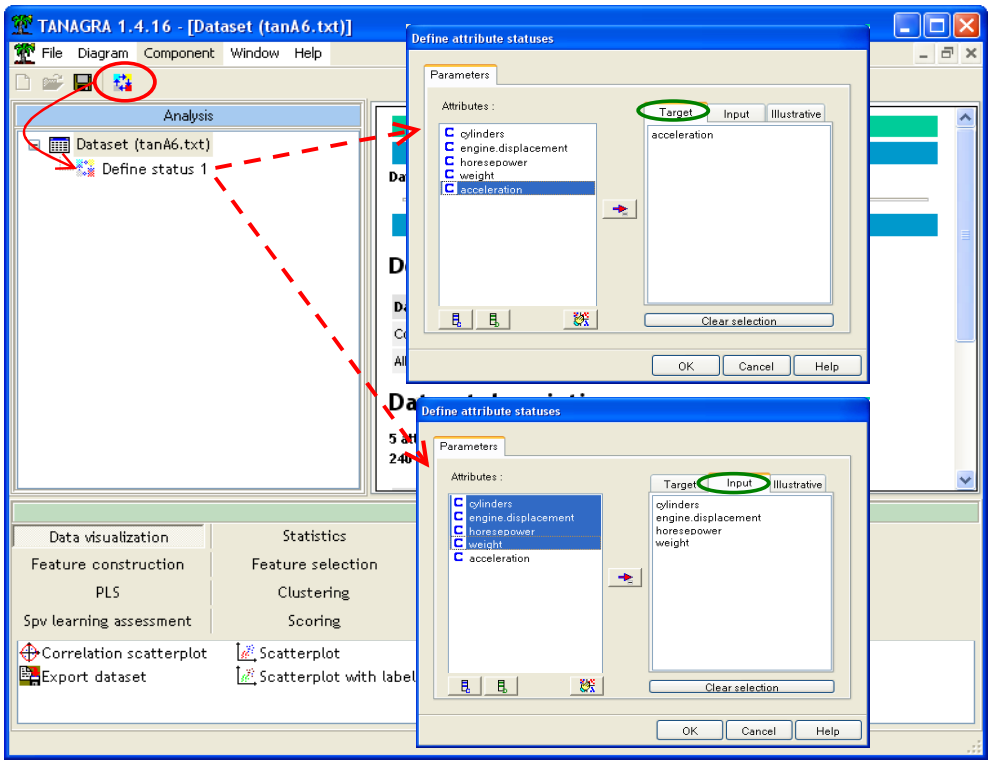
[1] Depuis la version 1.4.11 de TANAGRA, voir le didacticiel en ligne http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Excel_AddIn.pdf si la macro n'a pas été installée.



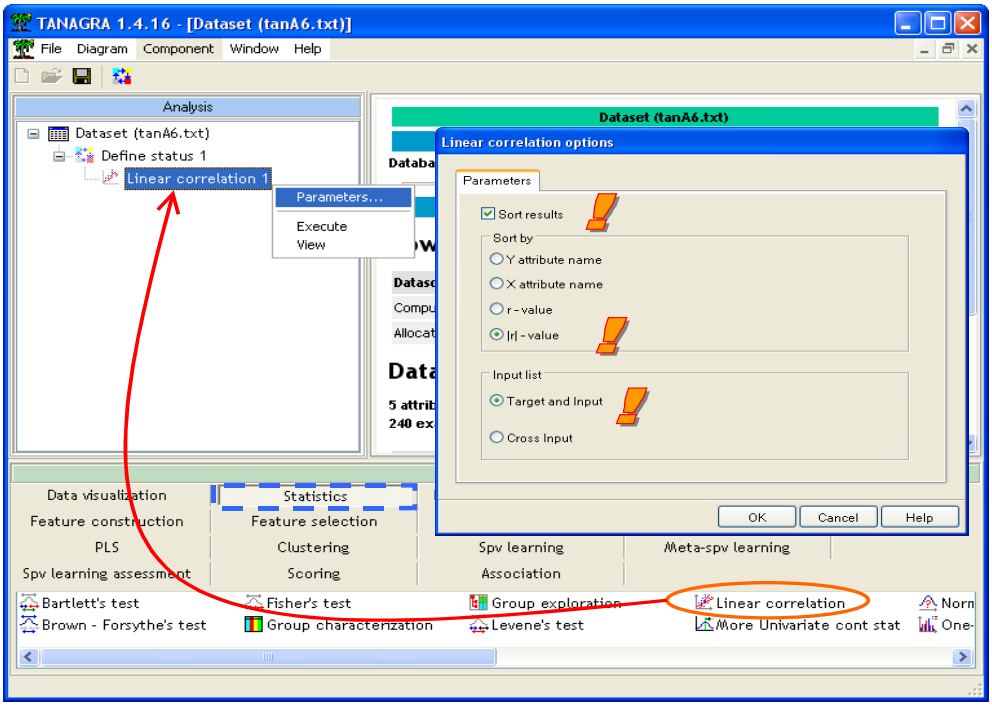
Corrélation variable d'intérêt et autres variables

Notre variable d'intérêt est ACCELERATION. Nous voulons connaître les variables qui lui sont le plus corrélées.

Nous plaçons le composant DEFINE STATUS dans le diagramme, le plus simple est de passer par le raccourci dans la barre d'outils. Dans la boîte de paramétrage qui apparaît automatiquement, nous mettons en TARGET la variable ACCELERATION, en INPUT toutes les autres variables. Elles sont toutes quantitatives dans ce fichier.

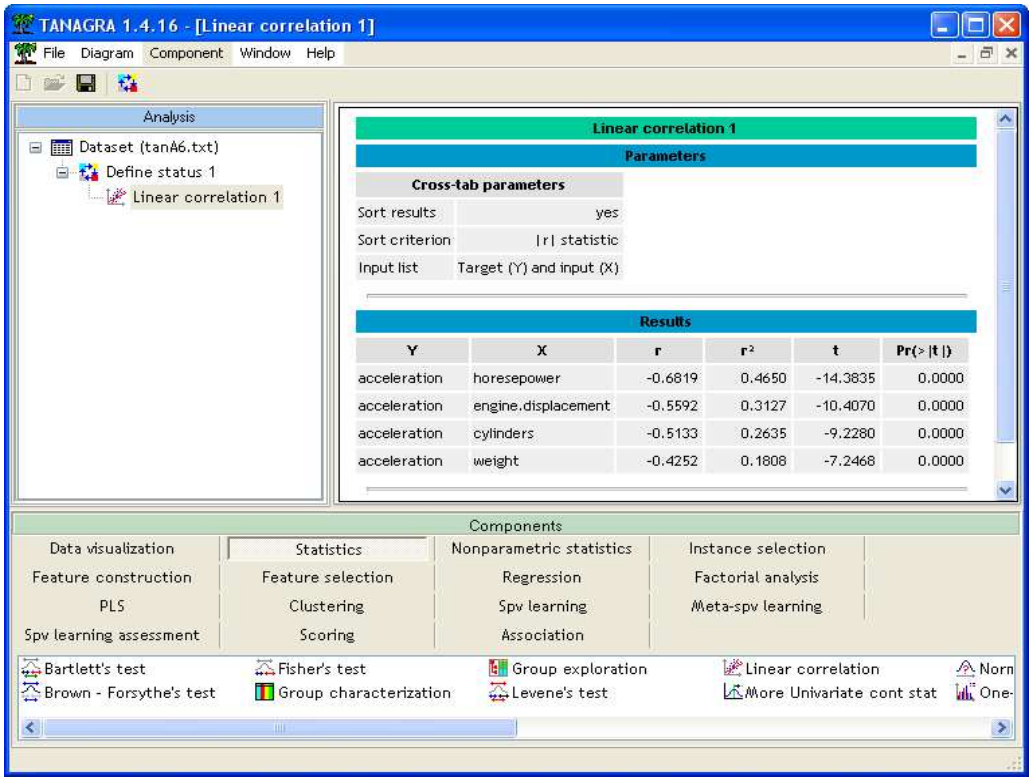


Puis, nous insérons le composant LINEAR CORRELATION (Onglet STATISTICS) dans le diagramme. Nous activons le menu PARAMETERS, nous indiquons que les résultats doivent être triés selon la valeur absolue du coefficient de corrélation décroissant, de manière à ce que les liaisons les plus fortes apparaissent en premier.



Nous ne modifions pas le paramètre INPUT LIST. Nous verrons son utilité plus loin, tout au plus remarquerons-nous que nous avons bien des variables à la fois en TARGET et en INPUT. L'option par défaut nous convient.

Le paramétrage validé, nous cliquons sur le menu VIEW pour accéder aux résultats.

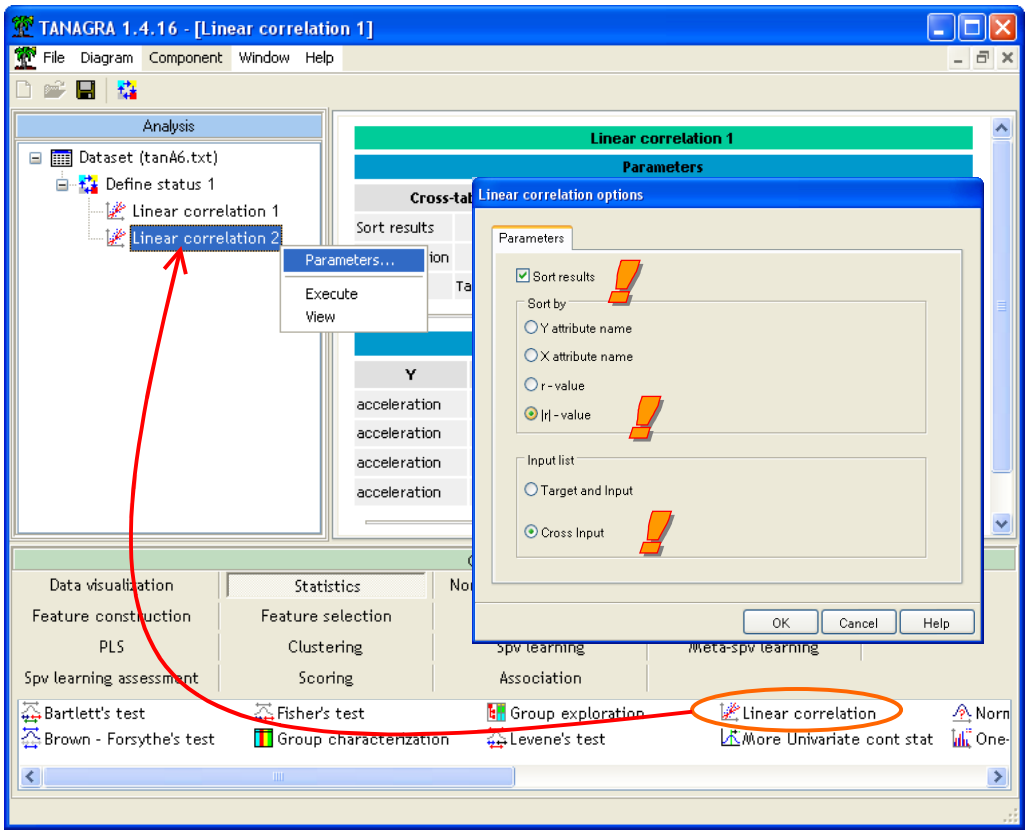


La variable la plus corrélée avec ACCELERATION est HORSEPOWER. Ce qui semble assez logique. La variable suivante est ENGINE.DISPLACEMENT (qui indique le volume du moteur[2]). Nous observons que toutes les corrélations sont significatives, même à un seuil très bas.

Corrélations croisées

Sans être un grand spécialiste, nous nous doutons bien que certaines variables INPUT doivent porter le même type d'informations. Par exemple, plus un moteur est volumineux, plus il sera puissant, vraisemblablement. La première étude ci-dessus doit être complétée avec une seconde analyse qui cherche à détecter les éventuelles redondances. Dans notre cas, il s'agit de calculer les corrélations croisées entre les variables INPUT.

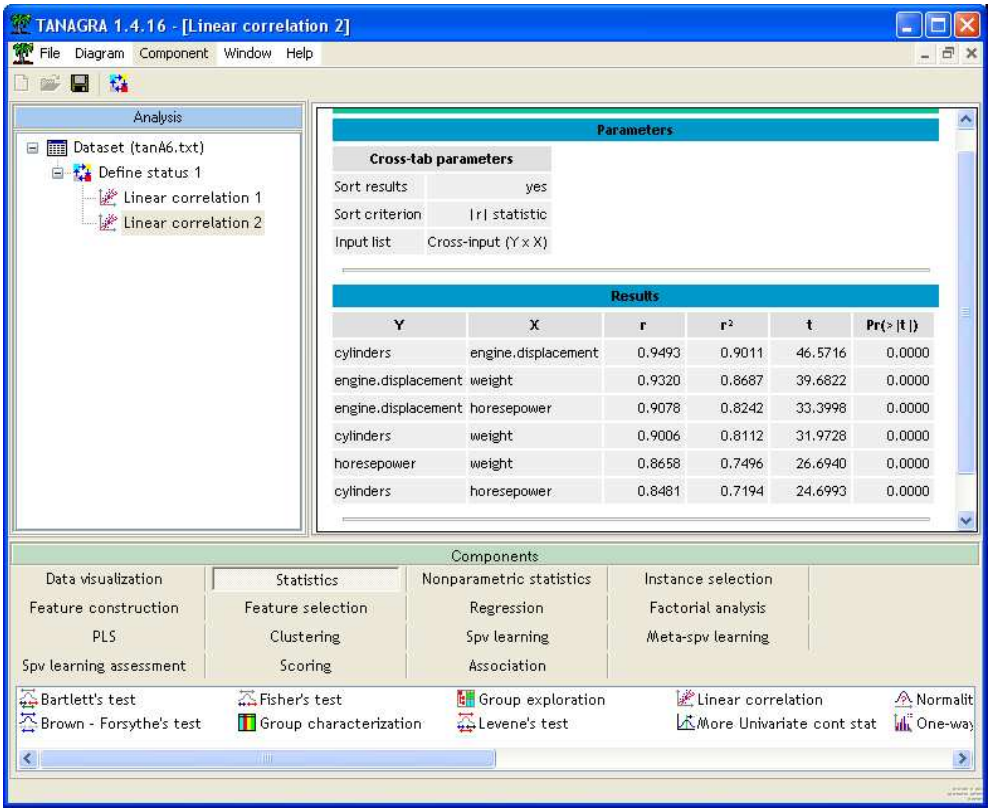
Pour ce faire, nous insérons de nouveau le composant LINEAR CORRELATION en dessous du composant DEFINE STATUS 1 dans le diagramme. Nous activons le menu PARAMETERS, nous voulons toujours que les résultats soient triés, mais de surcroît, nous indiquons maintenant que le calcul doit porter sur les variables indiquées en INPUT.



TANAGRA va donc calculer toutes les corrélations entre ces variables, si le nombre de variable est important, nous imaginons bien la masse de calcul que ça représente, tout l'intérêt de l'outil apparaît ici, les calculs sont réalisés rapidement et les résultats sont triés par intérêt décroissant.

Nous validons ce paramétrage et nous cliquons sur le menu VIEW pour accéder aux résultats.

[2] Le véritable terme technique est « cylindrée du moteur ». Mais pour les non-spécialistes, cette terminologie entraîne souvent une confusion avec le nombre de cylindres, qui est une autre donnée, présente également dans ce fichier (CYLINDERS).



ACCELERATION n'apparaît plus dans ce calcul, c'est tout à fait normal. Nous constatons qu'effectivement les variables CYLINDERS (nombre de cylindres) et ENGINE.DISPLACEMENT (volume total du moteur) sont fortement et positivement liés. Ce qui n'est pas une surprise.

De même, la puissance est liée positivement au volume du moteur, ce qui semble logique également connaissant un tant soit peu les voitures. Ce dernier résultat est très important car il éclaire différemment les corrélations entre ACCELERATION et les variables INPUT, nous comprenons qu'en réalité les deux premières corrélations (ACCELERATION vs. HORSEPOWER et ACCELERATION vs. ENGINE.DISPLACEMENT) traduisent en réalité le même phénomène.

Corrélation partielle

Revenons sur les premiers résultats, la corrélation négative (-0.4252) entre l'accélération (ACCELERATION) et le poids (WEIGHT) pose problème car il ne correspond pas du tout à ce que l'on sait des voitures. Il semble dire qu'un camion accélérerait plus vite qu'une formule 1, ce qui est absurde.

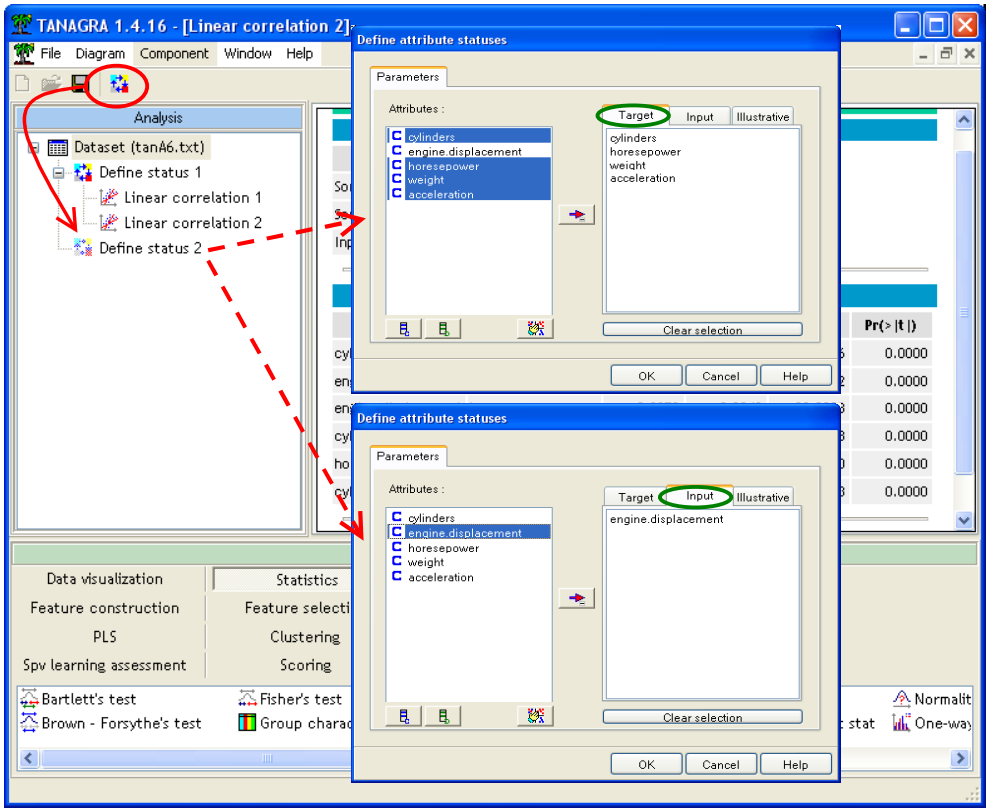
Ce résultat est éclairé d'un jour nouveau par les corrélations croisées. Nous constatons que le poids est lui-même fortement corrélé positivement (0.9320) avec le volume du moteur (ENGINE.DISPLACEMENT). C'est conforme à ce que l'on sait, les automobiles sont organisées en gamme, les grosses cylindrées sont également souvent des limousines luxueuses et volumineuses, par conséquent assez lourdes.

Nous constatons par ailleurs que la variable ENGINE.DISPLACEMENT est fortement corrélée avec toutes les autres variables (0.9493 avec CYLINDERS, 0.9078 avec HORSEPOWER). Pour compléter notre analyse, nous pourrions annihiler le rôle de cette variable pour évaluer correctement le rôle joué par les autres variables sur l'accélération : c'est la notion de corrélation partielle.

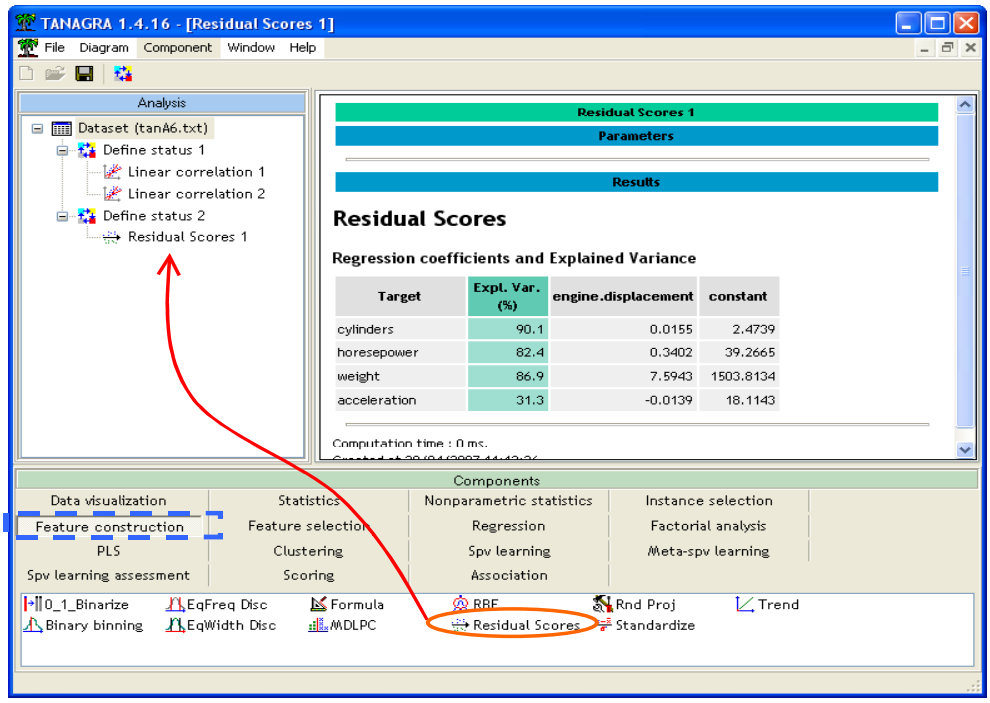
Nous effectuerons le calcul en deux temps : (1) retirer de l'ensemble des variables de l'étude l'effet induit par ENGINE.DISPLACEMENT ; (2) étudier alors les corrélations à partir des variables résultantes c.-à-d. les variables où l'influence de la variable conditionnelle a été déduite.

Construire les variables résiduelles

Insérons de nouveau le composant DEFINE STATUS à la racine de notre diagramme. Nous plaçons en TARGET les variables CYLINDERS, HORSEPOWER, WEIGHT et ACCELERATION ; en INPUT la variable ENGINE.DISPLACEMENT.



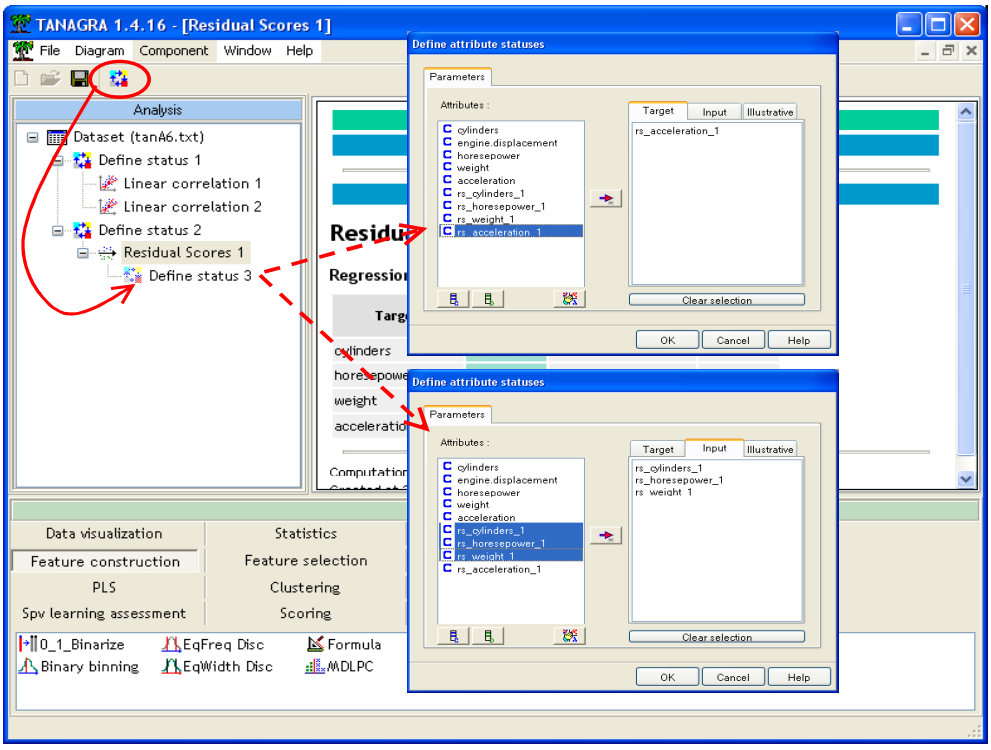
L'idée est de retirer des variables TARGET l'effet induit par la (les) variables INPUT(s) en calculant les résidus de la régression linéaire (multiple). Nous plaçons à cet effet le composant RESIDUAL SCORES (onglet FEATURE CONSTRUCTION). Nous activons le menu VIEW pour accéder aux résultats.



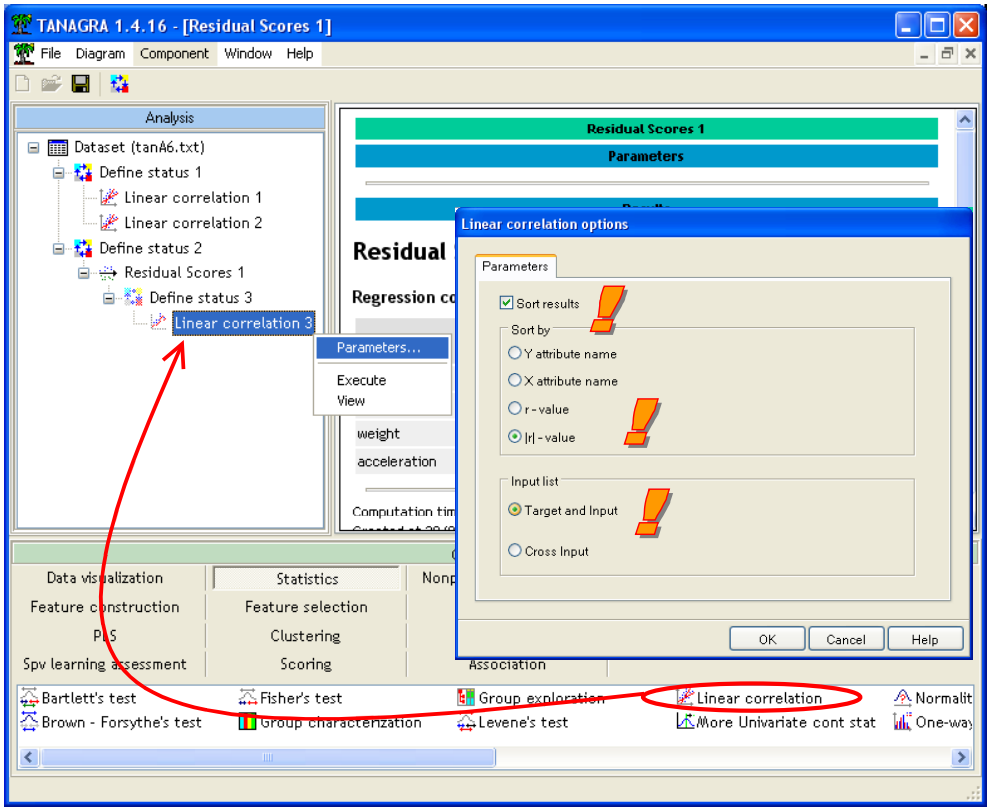
Nous constatons qu'ENGINE.DISPLACEMENT explique plus de 80% de la variabilité des variables CYLINDERS, HORSEPOWER et WEIGHT. Le résultat est moins tranché en ce qui concerne la variable ACCELERATION, où néanmoins 31% a été expliquée, ce qui n'est pas négligeable.

Calculer les corrélations partielles

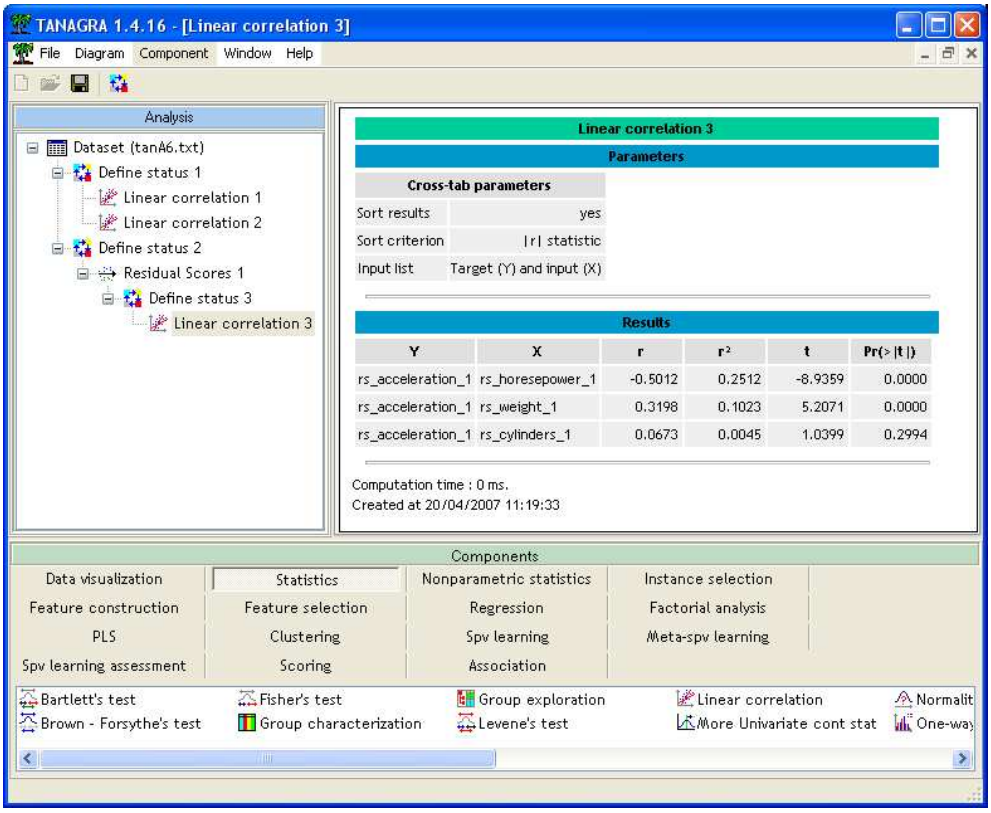
Nous voulons maintenant calculer les corrélations entre ACCELERATION d'une part et les autres variables d'autre part, après avoir retiré l'effet induit par ENGINE.DISPLACEMENT. Nous insérons de nouveau le composant DEFINE STATUS dans le diagramme, nous plaçons en TARGET la variable résiduelle d'ACCELERATION (RS_ACCELERATION_1), et en INPUT les variables résiduelles de CYLINDERS, HORSEPOWER et WEIGHT (RS_CYLINDERS_1, RS_HORSEPOWER_1 et RS-WEIGHT_1).



Puis nous insérons le composant LINEAR CORRELATION en veillant à ce que les résultats soient triés selon un coefficient de corrélation décroissant en valeur absolue.



Nous cliquons sur le menu VIEW pour accéder aux résultats.



La puissance (HORSEPOWER) conserve son rôle, plus les voitures sont puissantes, moins il leur faudra du temps pour parcourir une distance. Ca reste logique. La corrélation partielle reste significative[3] notamment parce qu'à volume de moteur égal, la puissance n'est pas la même d'une voiture à l'autre.

Plus intéressant maintenant est la corrélation avec le poids (WEIGHT). Nous constatons qu'elle est positive maintenant (0.3198). Moins la voiture est lourde, moins il lui faudra du temps pour atteindre une certaine vitesse. Ouf, ce n'est pas demain que l'on verra des tanks sur les circuits de formule 1 !!! Nous constatons effectivement que le rôle du poids a été totalement brouillé par l'influence d'ENGINE.DISPLACEMENT dans notre première analyse.

Conclusion

Très souvent, pour défricher un fichier, nous devons effectuer des séries de calculs (statistiques descriptives, corrélations, etc.) qui sont fastidieuses. Les outils que nous avons présentés dans ce didacticiel essaient de réduire ce travail répétitif de manière à ce que le statisticien se concentre plus sur une autre partie autrement plus ardue, la lecture et l'interprétation des résultats.

[3] Pour un vrai test, il faudrait en réalité corriger les degrés de libertés. TANAGRA ne tient pas compte de cette information. Néanmoins, il serait étonnant que cette correction modifie le sens des résultats ici.