Présentation du didacticiel

Dans ce didacticiel, nous montrons les étapes à suivre pour effectuer une analyse de type « régression ».

Le fichier que nous utilisons contient les caractéristiques de moteurs d'automobiles, et leur consommation. C'est la liaison entre cette donnée « consommation » et les autres variables du fichier que nous allons modéliser par régression.

Vous apprendrez à utiliser les opérateurs suivants :

Onglet	**Opérateur**	**Commentaire**
Data visualization	View dataset	Visualisation du fichier de données
Feature selection	Define status	Précision des variables à utiliser
Regression	Multiple linear regression	

Charger les données dans Tanagra

➢ Ouvrir un diagramme existant

1 – Choisissez *File/Open…* dans le menu de Tanagra.

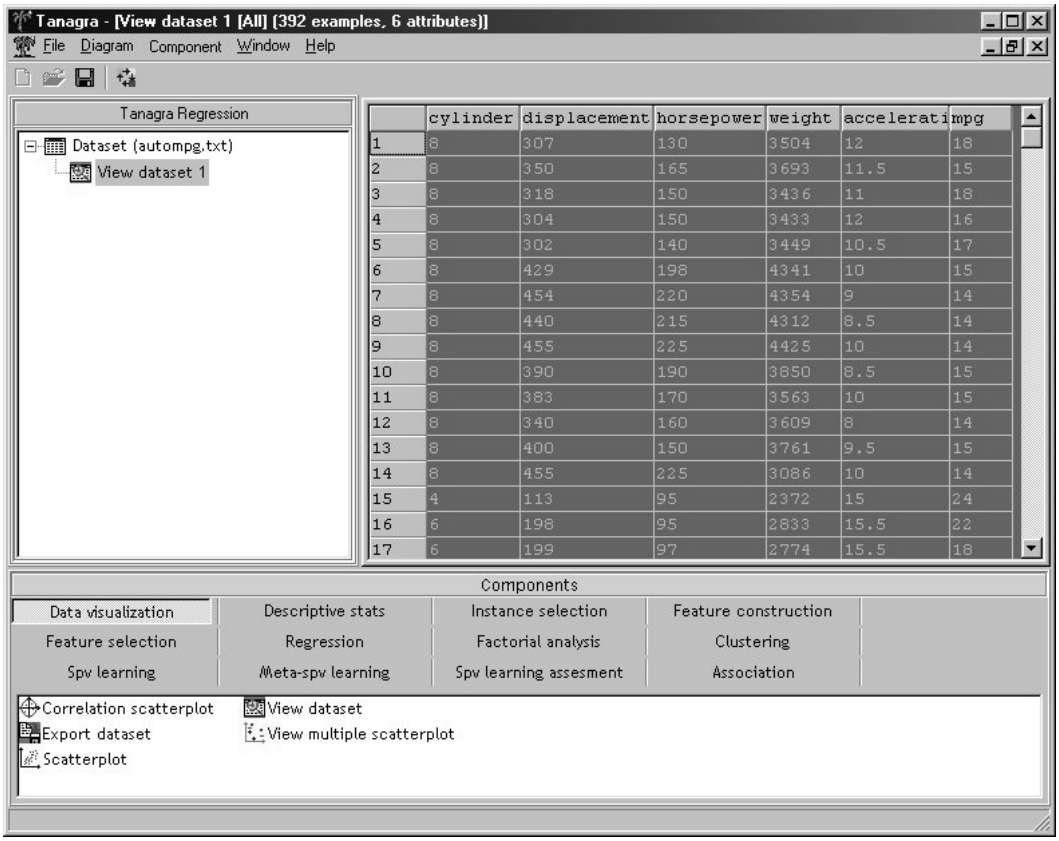
2 – Allez chercher le fichier « autompg.bdm » situé dans le sous-répertoire « Dataset » de Tanagra.

➢ Ajouter un opérateur au diagramme pour visualiser les données

1 – Ajoutez un opérateur **View dataset** au diagramme. Pour cela, cliquez sur l'onglet DATA VISUALIZATION de la palette des opérateurs.

Positionnez la souris sur l'opérateur **View dataset** et, en maintenant le bouton gauche de la souris enfoncé, amenez-le sur le diagramme. Relâchez quand vous êtes au-dessus du nœud "Dataset" (il doit apparaître sélectionné comme ci-dessous).

2 – Cliquez ensuite sur le nœud "View dataset" pour le sélectionner (s'il ne l'est pas déjà), et faites apparaître son menu contextuel par clic droit : choisissez la commande *View*. Les données apparaissent dans le cadre de droite.

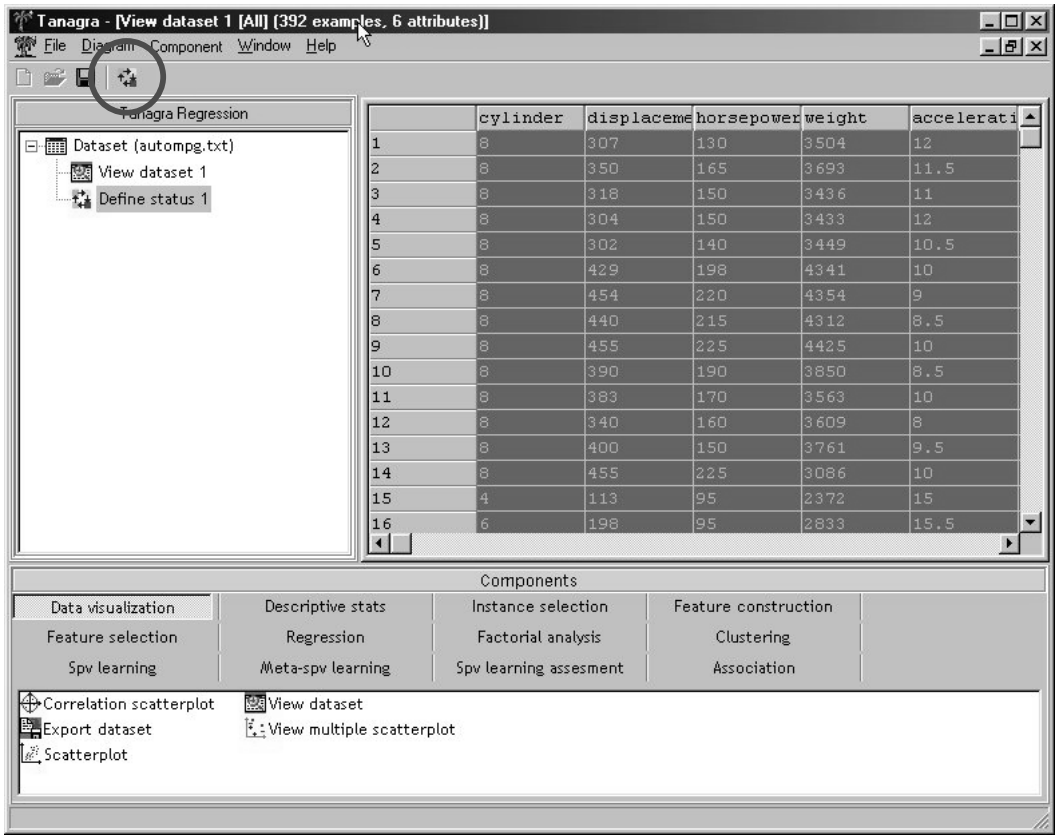


	cylinder	displacement	horsepower	weight	acceleratı	mpg
1	8	307	130	3504	12	18
2	8	350	165	3693	11.5	15
3	8	318	150	3436	11	18
4	8	304	150	3433	12	16
5	8	302	140	3449	10.5	17
6	8	429	198	4341	10	15
7	8	454	220	4354	9	14
8	8	440	215	4312	8.5	14
9	8	455	225	4425	10	14
10	8	390	190	3850	8.5	15
11	8	383	170	3563	10	15
12	8	340	160	3609	8	14
13	8	400	150	3761	9.5	15
14	8	455	225	3086	10	14
15	4	113	95	2372	15	24
16	6	198	95	2833	15.5	22
17	6	199	97	2774	15.5	18

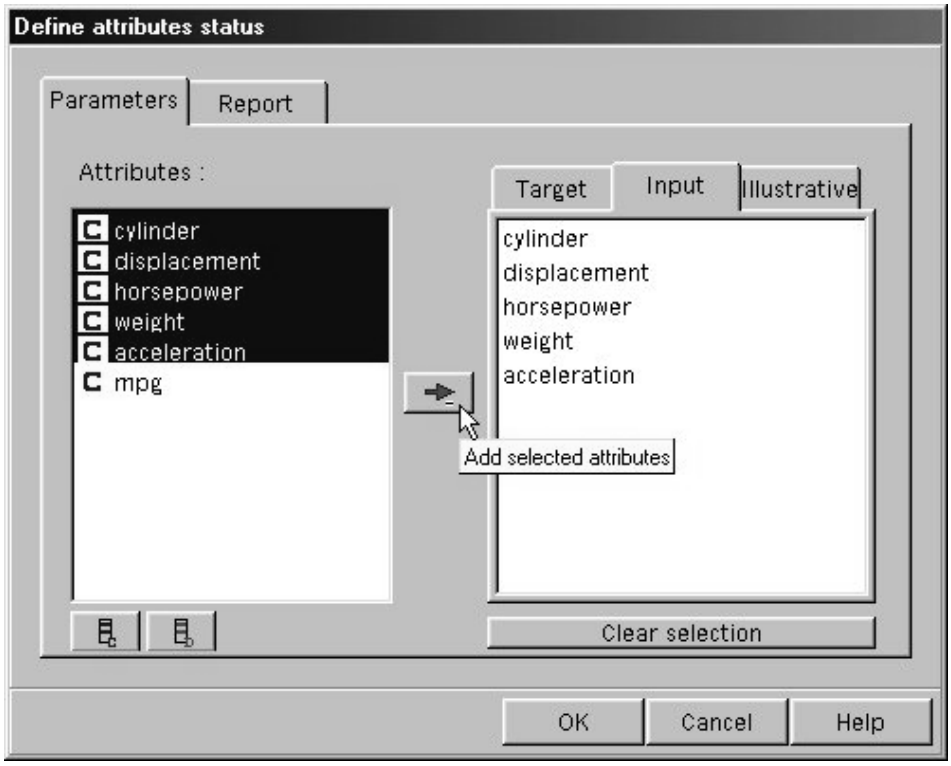
Les véhicules, au nombre de 392, sont caractérisés par 6 variables : nombre de cylindres, déplacement, puissance, poids, accélération et consommation (miles per gallon).

Définir les paramètres de la régression

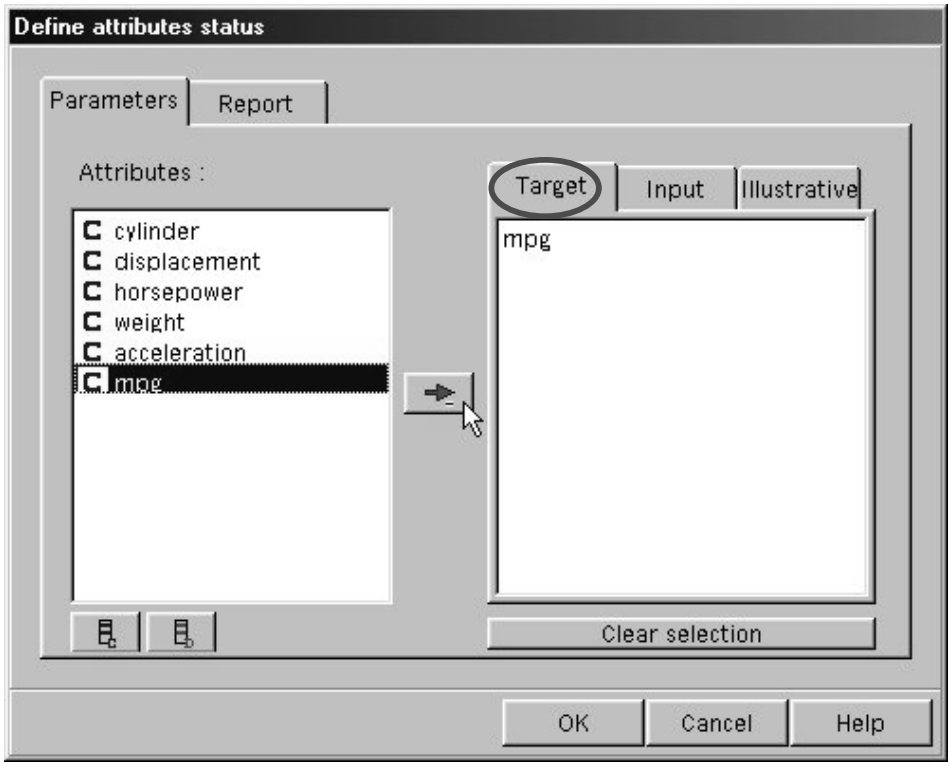
1 – Placez-vous sur le nœud « Dataset » et ajoutez un opérateur **Define Status** en cliquant sur son icône dans la barre des raccourcis. La fenêtre de dialogue permettant de définir le statut des variables apparaît automatiquement.



2 – Assurez-vous que c'est l'onglet « Input » qui est actif. Sélectionnez les cinq premières variables continues de la liste, et cliquez enfin sur le bouton flèche pour les passer dans la liste des Input.

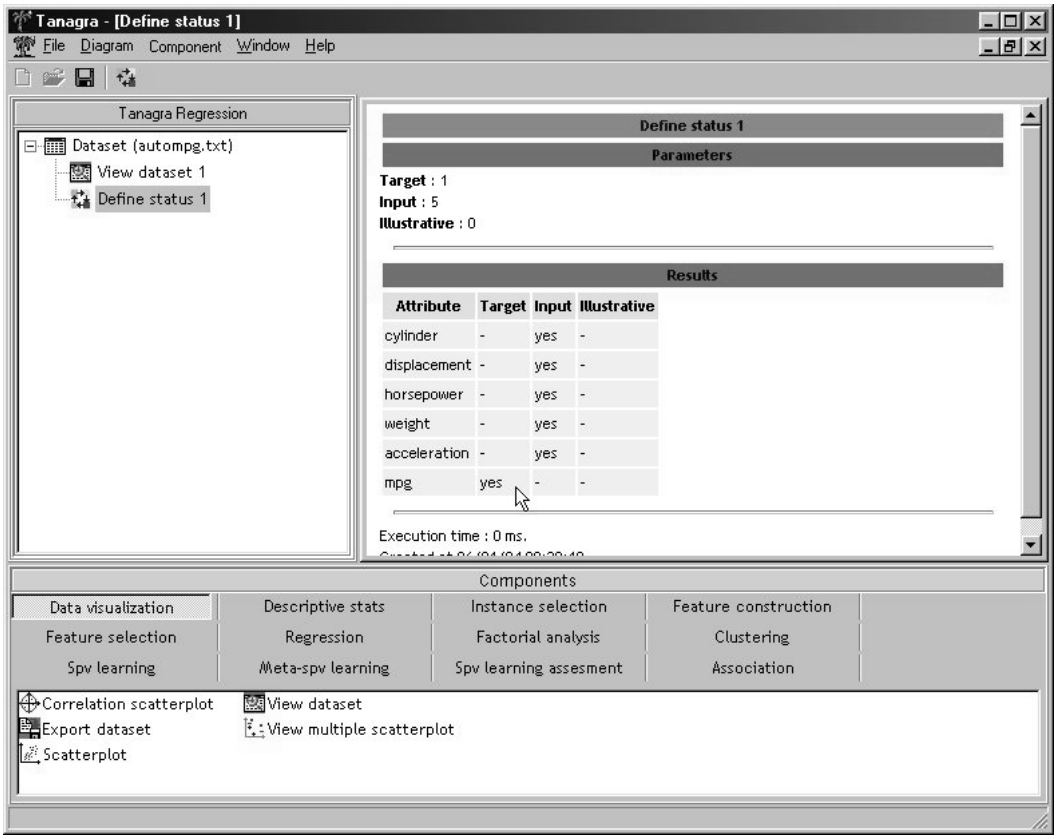


3 – Toujours en restant dans la fenêtre de dialogue, activez l'onglet Target. Cliquez sur la variable « mpg » pour la sélectionner, puis sur le bouton flèche.



4 – Vous venez de définir la variable à prédire (« mpg » = Target) et les variables explicatives (les autres = Input). Appuyez sur OK pour valider et fermer cette fenêtre.

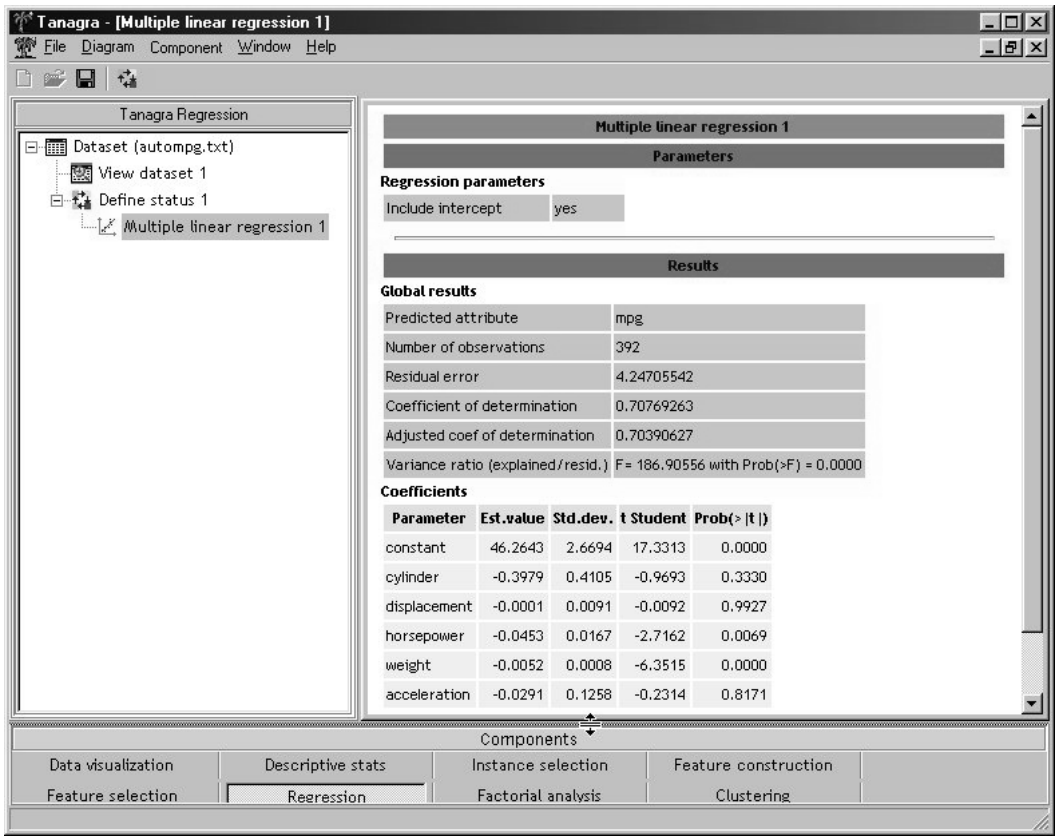
5 – Double-cliquez sur l'opérateur **Define status** que vous venez d'ajouter, vous pouvez ainsi vérifier le statut de vos variables dans le cadre de droite.



Effectuer la régression

1 – Ajoutez un opérateur **Multiple linear regression** (onglet REGRESSION) au diagramme, sous le nœud «Define status 1».

2 – Comme il n'y a rien à paramétrer de plus au niveau de cet opérateur, vous pouvez activer directement la commande *View* du menu contextuel. Les résultats de la régression apparaissent dans le cadre de droite.



On constate que la régression est d'assez bonne qualité (le coefficient de détermination est de 0.70). Les variables les plus significatives sont la puissance (horse power) et le poids (weight). Il faudrait cependant tenir compte des problèmes de colinéarité dans cet exemple.