Activités du grain 07

Robert Sabatier
Université de Montpellier, Montpellier, France

Christelle Reynes
Université de Montpellier, Montpellier, France

Myrtille Vivien
Université de Montpellier, Montpellier, France

V19.02

Table des matières

Cette activité se compose de deux parties : un exercice de reproduction, destiné à refaire un exercice présenté dans le grain ; et un exercice de compréhension du grain.

1 Exercice de reproduction du grain 07

A partir du jeu de données d'analyses des terephtalates utilisé dans le cours, vous allez réaliser une régression simple et une régression multiple.

Dans chemFlow, créez un nouvel historique : *CheMoocs-exercice-grain07* puis chargez les fichiers *x_tereph.tab* et *y_tereph.tab* depuis **chemflow/shared data/data libraries/chemoocs/grain07**.

— 1. Dans **x_tereph.tab** sélectionnez la quarantième longueur d'onde du tableau, qu'on notera NIR40, et calculez le coefficient de corrélation et les coefficients du modèle de régression linéaire de la variable *densité* en fonction de NIR40.

Deux solutions sont proposées, au choix : une solution simple qui passe par l'édition d'un graphique, le coefficient de corrélation fait partie des sorties ; et une solution plus compliquée mais aussi plus générale, qui passe par la concaténation de matrices.

Solution 1 : graphique.

Nous utiliserons la fonction **scatter plot** avec l'option **plot type** → *points*. Sélectionnez les deux fichiers : *y_tereph.tab* avec la variable *c2 :density* puis *x_tereph.tab* avec la variable *c41 : nir.40*. Il faut aussi cocher l'option **add bissectrice and statistic parameters** qui conduit à l'impression du R^2 et de la droite de régression sur la figure. R est déduit en prenant la racine carrée de R^2 et en lui donnant un signe positif si la droite de régression monte de gauche à droite de la figure, négatif si elle descend.

Solution 2 : concaténation.

Le coefficient de régression simple entre une ou plusieurs variables (R) est calculé par la fonction **statistics/correlation matrix**. Cette fonction ne s'applique que sur une matrice, or nos informations sont ici dans deux matrices. Il va donc falloir les réunir dans une seule matrice, grâce à la fonction **utils/merge files**. Cliquer deux fois sur **insert dataset** puis renseignez les options suivantes :

— **1-dataset/select dataset** → *y_tereph.tab*
— **1-dataset/select column(s) of dataset** → *c2 :density*
— **2-dataset/select dataset** → *x_tereph.tab*
— **2-dataset/select column(s) of dataset** → *c41 : nir.40*

— **merging by** → *columns*

On obtient le fichier *merge files on...* avec lequel il devient possible de calculer les coefficients de corrélation, avec la fonction **correlation matrix** et les options :

— **dataset** → *merge files on...*

— **column for correlation coefficients** → *c2 :density c3 :nir.40*

La fonction **regressions/slr** réalise une régression simple. Les paramètres à choisir sont les suivants :

— **select x data** → *x_tereph.tab*

— **column of x data chosen for the calculation** → *nir.40*

— **select y data** → *y_tereph.tab*

— **column of y data chosen for the calculation** → *density*

— 2. Sélectionner les variables nir.40 et nir.219 et réaliser la régression de la variable densité en fonction de ces deux variables. Donnez le coefficient de corrélation multiple au carré R^2.

Note : La validation croisée est un critère que nous utiliserons pour régler la dimension de certaines méthodes vues plus tard, comme la PCR ou la PLSR. La MLR, quand elle est calculable, donne un résultat unique : pas de dimension à régler. C'est pourquoi les paramètres de validation croisée seront ignorés. S'ils sont présents dans la méthode de calcul de la MLR, ils ne seront pas renseignés ou ils seront laissés à leurs valeurs par défaut.

Les deux variables sont extraites de *x_tereph.tab* avec la fonction **edit files** et les options :

— **select dataset** → *x_tereph.tab*

— **select operation** → *extract*

— **select from** → *columns*

— **enter column number(s)** → *41,220*

Notez que la virgule a été utilisée pour séparer plusieurs valeurs individuelles, alors que les deux points utlisés dans un grain précédent indiquaient un intervalle de valeurs.

Le fichier obtenu est *new x_tereph.tab*. La régression multiple peut maintenant être appliquée avec la fonction **regressions/mlr** et les options :

— **select x data** → *new x_tereph.tab*

— **select y data** → *y_tereph.tab*

— **column of y chosen for the calculation** → *c2 :density*

— **centering option** → *yes*

Les sorties sont : le modèle de régression (non lisible directement dans ChemFlow) ; les valeurs

prédites en validation croisée ; les b-coefficients du modèle de régression, les valeurs de RMSEC-RMSECV et les résidus. Les valeurs prédites par le modèle ne font pas partie des sorties, il faut les calculer, c'est à dire appliquer le modèle sur les mêmes données spectrales. On utilisera **regressions/applies a regression model...** avec :

— **select the regression model** → *mlr on (new x_tereph.tab ; y_tereph.tab) :model*

— **select x data** → *new x_tereph*

— **have you reference data of the new dataset ?** → *yes*

— **dataset** → *y_tereph.tab*

— **column of dataset chosen for the calculation** → *c2 :density*

Remarque : le nombre de variables latentes n'est pas à renseigner pour la MLR qui ne l'utilise pas ; laisser la valeur par défaut.

Le fichier de sortie **rmsep of new x_tereph.tab from mlr on (new x_tereph.tab ;y_tereph.tab) :model** contient 2 valeurs : le RMSEP (dont on verra la signification plus tard) et le R^2.

2 Exercice de compréhension du grain 07

Les données utilisées ici concernent 187 huiles d'olives dont l'origine géographique est connue. L'objectif de cet exercice est de prédire l'acide linolénique C18-3ω3 en fonction des spectres PIR. Nous utiliserons les fichiers *pir.tab* et *ags.tab*.

— 1. Charger les fichiers *pir.tab* et *ags.tab*.

— 2. Sélectionner 10 longueurs d'ondes à partir de la première, réparties uniformément toutes les 62 longueurs d'ondes avec la fonctionnalité séquence (*seq*).

On utilisera la fonction **tools/utils/edit files**. Deux options : soit indiquer nominativement les longueurs d'onde, soit les calculer en en prenant une toutes les 62.

Option 1 : indication nominative. Attention, dans ChemFlow, il faut rajouter 1 pour tenir compte de la première colonne qui contient les labels des lignes, et donc mettre : *2,64, 126, 188, 250, 312, 374, 436, 498, 560* dans le champ **enter column number(s)**. Les longueurs d'onde correspondantes sont : 1000, 1124, 1248, 1372, 1496, 1620, 1744, 1868, 1992 et 2116nm.

Option 2 : calcul d'une séquence.

Remplir le champ **enter column number(s)** avec *seq(2,613,62)*.

— 3. Quel est le nombre d'observations ? de variables explicatives ?

— 4. Réalisez la régression multiple de la variable linoléique en fonction des 10 longueurs d'ondes

sélectionnées. Donnez le coefficient de corrélation multiple.

La fonction **regressions/mlr** est utilisée avec les options :

- **select x data** → *new pir.tab*
- **select y data** → *ags.tab*
- **column of y chosen for the calculation** → *c11 :c18-3ω3*
- **centering option** → *yes*

Les valeurs prédites ainsi que le R^2 sont ensuite obtenues en appliquant le modèle obtenu sur les mêmes données : fonction **regressions/applies a regression model...** avec :

- **select the regression model** → *mlr on ags.tab :model*
- **select x data** → *new pir.tab*
- **have you reference data of the new dataset ?** → *yes*
- **dataset** → *ags.tab*
- **column of dataset chosen for the calculation** → *c11 :c18-3ω3*

Le R^2 est dans la sortie : *rmsep of new pir.tab from mlr on (new pir.tab x ags.tab) :model.*

— 5. Réalisez un premier graphique représentant les résidus sur les y contre les valeurs prédites, et un second graphique représentant le qq-plot des résidus.

Le qq-plot des résidus en fonction des valeurs prédites est obtenu avec la fonction **plots/qqplot** avec :

- **dataset containing residuals from a regression model** → *mlr on (new pir.tab ; ags.tab) : yresiduals*
- **column for x axis** → *c2 :residuals*

— 6. Rajouter les labels de l'origine géographique sur le graphique des résidus - valeurs prédites (le premier graphique de la figure précédente). Que pensez-vous de ce modèle ?