Cette étude s'inspire d'un cas réel.

Le fichier Excel 'FW_groupexx', où xx est le n° du groupe contient la description de 100 'objets' à l'aide de 50 descripteurs.

Le fichier Excel 'FW_groupexx_obs' contient 25 objets décrits à l'aide des mêmes descripteurs. Mais sur ces 25 objets on a mesuré une réponse, qui se trouve en colonne 1 de ce fichier. Aucun de ces 25 éléments ne figure dans 'FW_groupexx'.

Le problème que l'on se pose est de faire le lien entre les descripteurs et la réponse à l'aide d'un modèle linéaire, construit avec un nombre limité de descripteurs (4 ou 5 au plus), à choisir parmi les 50 initiaux.

Le 1er descripteur correspond à la variable x_1, le deuxième à x_2, …, le 50^{ième} à x_{50}. Dans le modèle, on s'autorise des termes de degré 1 et des interactions. Par exemple, si i,j,k sont les indices des variables choisies, le modèle peut contenir les variables :

- x_i, x_j, x_k
- $x_i\,x_j$, $x_i\,x_k$, $x_j\,x_k$

Un modèle contenant 3 descripteurs comprendra donc au plus 7 coefficients, correspondant aux 6 variables précédentes et au terme constant. Mais il ne les contient pas nécessairement tous, car on recherche un modèle 'satisfaisant' et contenant le moins possible de variables. Un modèle sera dit satisfaisant si l'erreur de prédiction sur les 25 'objets' est inférieure à ou d'un ordre de grandeur égale à 1,5.

Plusieurs problèmes sont à résoudre :

1) Certains objets peuvent être mal renseignés.

2) 50 descripteurs conduisent à 1225 interactions. Donc le nombre de variables potentielles pour ce problème est de 1275 (interactions, plus 50). C'est beaucoup trop. Ce serait bien de pouvoir en éliminer.

   On remarque que les descripteurs sont très corrélés entre eux. Cela signifie que certains peuvent être très bien prédits à l'aide de plusieurs autres grâce à une relation linéaire (cette fois-ci sans interactions). On suggère donc de commencer par repérer ces descripteurs redondants en utilisant le fichier 'FW_groupexx'. Un descripteur sera redondant si il peut être prédit à l'aide des autres avec un coefficient de détermination supérieur à 0,95.

3) Certains objets peuvent être suspects ou 'outliers'. Ils sont éventuellement difficiles à détecter. Cela doit se faire en cours de modélisation.

4) Le risque de sur-ajustement est très élevé : on a peu d'observations et beaucoup de régresseurs possibles.

A. Pour récapituler, on propose d'écrire successivement 3 algorithmes pour résoudre ce problème.

1) Elimination des descripteurs inutiles.

L'algorithme procède de la façon suivante :

- à partir de 'FW_groupexx', on essaie de trouver les relations linéaires à 1 variable qui lient un régresseur à un autre. On élimine ceux qui peuvent être prédits avec un $R^2 > 0.95$.

- même chose avec les relations linéaires à 2 variables.

- même chose avec les relations linéaires à 3 variables.

On espère éliminer ainsi 10, 15 variables explicatives redondantes.

2) régression 'stepwise'.

On considère tous les sous-ensembles de 3 variables parmi les prédicteurs restants.

- Chacun de ces sous-ensembles permet de construire une relation linéaire avec interactions qui comprend au plus ces 3 variables x_i, x_j, x_k et leurs interactions. On cherche à l'aide d'une régression 'stepwise' le modèle plus pertinent pour ce sous-ensemble de variables.
- A l'aide du modèle obtenu, on calcule le PRESS

Le modèle final sera celui pour lequel le PRESS est le plus faible.

3) Régression 'Least Absolute Deviation'.

La régression 'stepwise' de la question précédente est dans le cas de notre problème questionnable : d'une part les régressions de type L2 sont réputées peu résistantes aux 'outliers', d'autre part le faible nombre d'échantillons de la base de calibration risque d'induire des problèmes de stabilité. On propose d'effectuer une régression de type 'L1' à la place de la régression 'stepwise'. On utilisera avec profit le package 'L1Pack' ou 'Blossom'.

4) Comment comparer les résultats des questions 2 et 3 ?