TP n°3 – Régression Linéaire et challenge TOTAL –

Pour réaliser ce TP **en binôme**, on utilisera le fichier de commandes R '**ScriptTP3.R'** (disponible sur **Campus**) ainsi que le fichier de données « **TrainSample.csv »**.

On s'intéresse ici aux techniques de régression linéaire en vue du challenge TOTAL. Pour pouvoir tester « librement » les différentes méthodes, on redécoupe le fichier de données '**TrainSample.csv** (correspondant à 460 individus) en un sous-ensemble d'apprentissage de taille 360 et un sous-ensemble test de taille 100.

Partie 1.

1. On commence par déterminer la variable la plus corrélée avec la réponse à prédire « **GasCum360** » que l'on utilise pour construire un modèle de régression linéaire simple (p = 1 prédicteur). De même pour l'autre variable à prédire « **OilCum360** ». En déroulant le script fourni, obtenir le meilleur score en jouant sur le niveau de confiance des intervalles de prédiction calculés.

2. Les hypothèses du modèle linéaire simple n'étant clairement pas satisfaites dans les deux cas, on propose de corriger le modèle linéaire en considérant une transformation simple des variables à prédire (passage au log après translation des valeurs) et en ajoutant un terme quadratique pour la réponse prédite (régression linéaire multiple). A nouveau, obtenir le meilleur score possible. Peut-on valider les modèles obtenus ?

Partie 2.

Reprendre la partie 1 en ajoutant d'autres prédicteurs (bien choisis) et en considérant d'autres transformations simples des variables.