

TP final – Régression Linéaire et challenge TOTAL –

On s'intéresse uniquement aux techniques de régression linéaire pour le challenge TOTAL. Pour pouvoir tester « librement » les différents modèles, on rappelle que le fichier de données '**TrainSample.csv**' (correspondant à 460 individus) est partitionné en un sous-ensemble d'apprentissage de taille 360 et un sous-ensemble test de taille 100. On pourra étudier l'influence de ce découpage aléatoire sur les résultats.

On fournit le fichier de commandes '**ScriptTPfinal.R**' qui reprend et complète le script utilisé pour répondre aux questions posées lors du TP3.

On pourra s'aider au début des indications suivantes compte tenu des résultats déjà obtenus.

1. On a déterminé au TP n°3 la variable la plus corrélée avec la réponse à prédire « **GasCum360** » que l'on a utilisée ensuite pour construire un modèle de régression linéaire simple (donc avec $p = 1$ prédicteur). De même pour l'autre variable à prédire « **OilCum360** ». Améliorer le score obtenu en se basant sur des intervalles de prédiction non symétriques contrairement à ceux calculés par défaut avec le modèle linéaire.
2. On a également considéré 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 en jouant sur la forme des intervalles de prédiction.
3. Considérer un modèle linéaire complet utilisant les réponses transformées et tous les prédicteurs.
4. Essayer d'améliorer le résultat obtenu en essayant de construire un modèle réduit (on pourra utiliser le test de Fisher de comparaison de modèles emboîtés, utiliser la fonction **anova** de R).
5. Divers : régression sur composantes principales (ACP sur les prédicteurs), régression quantile, régression logistique, régression L^1 , etc.

Travail à réaliser **en binôme** sous la forme d'un **compte-rendu synthétique** (quelques pages, pas de code demandé)

sous le format **Nom1Nom2TPReg.pdf**,

A rendre au plus tard le **lundi 13/novembre** (début du cours de « Séries Temporelles »)