TP2 - Régression Linéaire Multiple

pour la prédiction de ventes

On dispose d'un échantillon statistique de **n = 300** données de **ventes** globales journalières (en euros) d'une station-service ouverte tous les jours de la semaine. Conjointement, on fournit les températures moyennes journalières (en °C) ainsi que les jours de la semaine correspondants (lundi, …, dimanche). Ces données sont dans le fichier **'Data_app.txt'** (données d'apprentissage).

Le but est de construire le meilleur modèle prédictif des ventes en fonction des variables **'température'** et **'type de jour'**.

Le critère utilisé pour évaluer un modèle est la moyenne des carrés des erreurs de prédiction (en anglais, Root Mean Square Error) :

$$\textbf{RMSE} = \sqrt{\textbf{MSE}} \quad \textbf{où MSE} = \frac{1}{\textbf{n.test}} \sum_{\textbf{i=1}}^{\textbf{n.test}} (y_i - \hat{y}_i)^2$$

où n.test = 100 est le nombre de données 'test', y_i les valeurs de vente réellement observées pour ces données 'test' (fichier **Ventes_test.txt**) et $\hat{y}_i$ les prédictions obtenues à partir du modèle en question pour les données de test (fichier **Data_test.txt**).

1. Commencer par une **régression linéaire simple** des ventes sur la température journalière. Obtenir le graphique montrant les données ainsi que la droite de régression. Qu'observe-t-on ? Pouvez-vous expliquer ? Tracer les résidus contre la réponse prédite. Conclusion ? Quel score obtient-on ?

2. Comment pourrait-on améliorer le modèle à partir de simples régressions linéaires ? Score obtenu ?

3. Envisager une **régression linéaire multiple** faisant intervenir la variable température, la variable binaire associée aux jours de week-end (= 1 si jour de we et = 0 sinon) et enfin celle associée au fait que le jour soit un dimanche (=1 si dimanche et = 0 sinon). Calculer le score et comparer avec 2.

4. Dans le modèle RLM précédent, il peut être pertinent d'envisager une interaction entre les prédicteurs. Quel(s) nouveau(x) prédicteur(s) faut-il introduire ? Faire l'étude et retrouver le score obtenu en 2.

5. Réfléchir encore à d'autres améliorations possibles…