TP2 HPC-BigData 2023 - Partie 1 : Modèle Linéaire Gaussien

Adaptation statistique des prévisions d'Ozone du modèle MOCAGE

Description des données :

Le but de ce TP est l'emploi de techniques de modélisation linéaire pour adapter localement les prévisions de concentration d'ozone du modèle MOCAGE de chimie-transport de Météo-France à l'échéance 24H. Ce post-traitement statistique nommé adaptation statistique (AS) a pour objectifs de débiaiser les prévisions de MOCAGE et réduire l'amplitude de ses erreurs.

La stratégie adoptée dans ce TP est de proposer au final un unique modèle linéaire gaussien opérant la régression de l'ozone observé (**prédictand = variable O3o**) sur cinq sites en France (Aix, Cadarache-StPaul, Plan d'Aups, Rambouillet, et Munchhausen en Alsace ; ***voir carte en annexe***) par exploitation des 7 prédicteurs potentiels suivants :

O3p :	[O_3] prévue par MOCAGE à l'échéance considérée ($\mu g/m^3$)
TEMPE :	Température prévue par MOCAGE pour l'échéance considérée (°C)
RMH2O :	Rapport de mélange prévu par MOCAGE pour l'échéance considérée (g/kg)
FF :	Force du vent prévue par MOCAGE pour l'échéance considérée (m/s)
NO2 :	[NO_2] prévue par MOCAGE à l'échéance considérée ($\mu g/m^3$)
JJ :	Jour de la semaine, facteur à deux modalités codées : S pour jours ouvrés, F pour fins de semaine et jours fériés
STATION :	Nom de la station, facteur à cinq modalités codées Aix, Cad, Pla, Ram et Als.

Le modèle statistique devra donc exploiter le nom du site comme prédicteur (une autre stratégie aurait été d'élaborer un modèle statistique par site) et sera ainsi capable d'adapter sa prévision au site d'intérêt. On dispose dans le fichier **DataTP.txt** de mesures de concentration d'ozone **O3o** réalisées lors de quatre étés et des prévisions associées issues du modèle MOCAGE, classées par station et par ordre chronologique.

1. Chargement des données :

Charger les données dans une data.frame :
data=read.table("DataTP.txt", header=TRUE)

2. **<u>Etude préliminaire et régression simple sur la station d'Aix</u> :**

- Créer une data.frame contenant les données relatives à la station d'Aix.

- Calculer les statistiques de base sur l'ozone observé **O3o** et prévu par MOCAGE **O3p** à Aix, comparer leurs distributions. Peut-on considérer que les moyennes de ces deux variables sont différentes (*t.test*) ? (comparer d'abord les variances avec *var.test*)

- Estimer le coefficient de corrélation entre ozone prévu et observé à Aix, interpréter. Faire la régression de l'ozone observé **O3o** par l'ozone prévu **O3p** (*lm*) et interpréter les sorties de la fonction *summary*.
  (→ estimation du modèle **O3o** = $\beta_0 + \beta_1$.**O3p** + **e**)

- Représenter le nuage de points (*plot*). Ajouter sur ce graphe la droite de régression (*abline*). Commenter.

- Représenter la série chronologique relative à l'ozone observé **O3o** (*plot*) et ajouter sur ce graphe des croix bleues pour les prévisions de MOCAGE **O3p** et des croix rouges pour les prévisions issues du modèle linéaire simple (*points*). Commenter le graphe.

3. **<u>Régression multiple</u> : prédicteurs quantitatifs** (à partir de la data.frame 'data')

- Comparer, dans une même fenêtre graphique, les histogrammes des différents prédicteurs quantitatifs. Quel pré-traitement des données préconiseriez-vous ?

- Etudier le lien entre les différentes variables quantitatives : ***pairs(data[,c(-1,-7)])***

- Faire la régression de l'ozone observé par les prédicteurs quantitatifs disponibles :

  regmult=lm(O3o~O3p+TEMPE+RMH2O+log(NO2)+FF,data)

  Analyser les sorties de la fonction *summary* ainsi que la 'design matrix' (*model.matrix*). Quels prédicteurs semblent pertinents ? Justifier.

- Vérifier les hypothèses du modèle linéaire gaussien:
  - homoscédasticité :
  plot(fitted(regmult),residuals(regmult))
  - normalité :
  hist(residuals(regmult)) ; qqnorm(residuals(regmult))
  - indépendance :
  acf(residuals(regmult))

  Vérifier l'hypothèse de linéarité de la réponse :
  plot(fitted(regmult),data$O3o)

Annexe : carte des stations