TP2 HPC-BigData 2022 – Partie 2 : Modèle Linéaire Gaussien

Adaptation statistique des prévisions d'Ozone du modèle MOCAGE

4. ANOVA: prédicteurs qualitatifs et contraintes d'identification

• Faire la régression de l'ozone mesuré **O3o** par le facteur **JJ** : anova1=lm(O3o ~ JJ,data)

Analyser les sorties de la fonction *summary* et la 'design matrix'. Quelle contrainte d'identification R a-t-il imposée par défaut ? Interpréter alors les paramètres estimés du modèle.

 Imposer la contrainte 'somme des contributions de chaque modalité = 0' : anova2=lm(O3o ~ C(JJ,sum),data)
Refaire l'analyse précédente.
Le modèle dépend-il de la contrainte imposée ?

5. ANCOVA: modèle complet et sélection automatique des prédicteurs

• Estimer le modèle complet d'analyse de covariance en considérant l'ensemble des prédicteurs potentiels, et afficher le bilan de l'estimation.

 $regcomplet=lm(O3o\sim O3p+TEMPE+RMH2O+log(NO2)+FF+STATION+JJ,data)$

- Quels prédicteurs vous semblent pertinents ? Comparer avec le modèle regmult.
- **Sélection automatique** : le critère d'Akaike (**AIC** Akaike Information Criterion)

On cherche le modèle qui minimise l'indice AIC, n étant la dimension de l'archive, j étant la dimension du modèle $Y=T\beta+e$:

$$AIC = ln \left(\frac{\left\| Y - T\hat{\beta} \right\|^2}{n} \right) + \frac{2j}{n}$$

Quel est l'intérêt d'exploiter un tel critère ?

Effectuer une sélection descendante (charger le package MASS : library(MASS)):

regaic=stepAIC(regcomplet)

Commenter le modèle sélectionné.

• Utilisation du critère de Schwartz (**BIC** – Bayesian Information Criterion) :

On cherche le modèle qui minimise l'indice **BIC** :

$$BIC = \ln \left(\frac{\left\| \mathbf{Y} - \mathbf{T} \hat{\boldsymbol{\beta}} \right\|^2}{\mathbf{n}} \right) + \frac{\mathbf{j} \ln(\mathbf{n})}{\mathbf{n}}$$

BIC est une variante de **AIC** qui exploite une pénalisation en jln(n) au lieu de 2j. Quel va être l'impact sur la sélection par rapport à la sélection **AIC** ? Les variables sélectionnées sont-elles identiques au modèle **regaic** précédent ? regbic=stepAIC(regcomplet,k=log(nrow(data)))

Estimer un modèle **BIC** avec interactions d'ordre 2 : regbicint=stepAIC(lm(03o~.*.,data),k=log(nrow(data)))

6. **Evaluation des modèles :**

- Comparer sur un même graphe les observations d'ozone **O30**, les prévisions brutes de MOCAGE **O3p** et les prévisions obtenues après post-traitements statistiques (modèle **BIC** avec interactions et régression simple exploitant le prédicteur **O3p**). Commenter.
- Coder une fonction R calculant le biais et le RMSE d'une série de prévisions. Estimer ces scores pour les prévisions **O3p** du modèle MOCAGE et pour les différentes prévisions statistiques. Commenter.
- Ce mode d'évaluation vous semble-t-il satisfaisant? Quels problèmes pose-t-il? Proposer une stratégie d'évaluation plus rigoureuse.
- A l'aide des fonctions R *sample* et *setdiff*, créer un fichier d'apprentissage *datapp* contenant 80% des données et un fichier de test *datatest* contenant les données restantes.
- Réestimer les modèles précédents sur les données d'apprentissage. Estimer les scores sur apprentissage puis sur test (*predict*). Quelle analyse pouvez-vous faire ? Illustrer le phénomène de sur-apprentissage en estimant un modèle très complexe.
- Analyser le script CV.R fourni puis l'exécuter. Interpréter le graphe obtenu et conclure sur le modèle à finalement proposer.