Modélisation statistique - TD5
Régression linéaire multiple avec R

Un ingénieur s'intéresse à la consommation d'eau dans une usine. Il cherche à relier la consommation d'eau de l'usine avec différentes variables mesurées chaque mois depuis 17 mois. Il dispose des données suivantes :

— Temperature : la température moyenne en degrés Celcius
— Production : la production de l'usine en kg.
— Days : le nombre de jours de production dans le mois.
— Persons : le nombre de personnes ayant travaillé dans le mois.
— Water : la consommation d'eau mensuelle en m^3.

Partie 1 : Estimation

Il modélise la consommation d'eau en utilisant une régression linéaire multiple avec les variables explicatives suivantes :

— Production
— Temperature
— Days
— Persons

1. Charger les données sur **R**.
2. Quel modèle est estimé ici ? Ecrire le modèle de façon matricielle et non matricielle. Pour l'écriture matricielle, donner avec précision tous les vecteurs et matrices intervenants.
3. Implémenter le modèle sur **R** à l'aide de l'instruction suivante :

```
reg=lm(Water~Production + Temperature + Days + Persons ,data=data)
summary(reg)
```

4. Rappeler l'expression et la loi du vecteur $\hat{\boldsymbol{\beta}}$.
5. La table `Coefficients` donne l'estimation de β_j, de l'écart-type de $\hat{\beta}_j$, ainsi que la statistique de test et la probabilité critique, pour le test $\mathcal{H}_0 = \{\beta_j = 0\}$ contre $\mathcal{H}_1 = \{\beta_j \neq 0\}$. Donner l'expression de cette statistique de test et sa loi sous $\mathcal{H}_0$.
6. Donner la définition de R^2 et rappeler son lien avec σ^2.
7. Donner le degré de liberté du `RSE`, ainsi que son expression. En déduire une estimation de σ^2.
8. Donner la statistique de Fisher du test global du modèle. Le modèle est-il globalement significatif au risque $\alpha = 5\%$?

Partie 2 : Choix de modèle

D'après les conclusions de la première partie, le modèle est globalement significatif. Cependant, l'ingénieur se demande si toutes les variables sont nécessaires.

1. Dans un premier temps il décide d'étudier la corrélation entre les différentes variables. Calculer la corrélation entre chaque variable que pouvez-vous en conclure ?
2. L'ingénieur décide de supprimer une à une les variables du modèle et de comparer les critères AIC. Recopier l'instruction ci-après et interpréter les résultats ci-dessous.

   ```
   step(reg)
   ```

3. L'ingénieur décide de supprimer la variable jour du modèle. Dans cette situation, il s'interroge sur le meilleur modèle entre les deux suivants

   ```
   mod1=lm(Water~Production + Temperature + Persons ,data=data)
   mod2=lm(Water~Production + Persons, data=data)
   ```

   Il décide de réaliser un test de comparaison de deux modèles emboités. Les deux modèles décrits ci-dessus sont-ils emboités ? Rappeler les hypothèses de test. Obtenez les résultats du test avec l'instruction `anova(mod1,mod2)` et conclure.

Partie 3 : Validation

L'ingénieur décide de modéliser la consommation d'eau avec le modèle `mod2`.

1. Rappeler les hypothèses qu'il doit vérifier pour valider ce modèle.
2. Dans ce but il produit les graphiques suivants :

   ```
   plot(mod2,which=1)
   plot(mod2,which=2)
   plot(mod2,which=4)
   ```

   Les hypothèses sont-elles validées selon vous ? Que conseilleriez-vous à l'ingénieur ? Des données vous paraissent-elles aberrantes ?
3. Rappeler la définition de levier pour l'observation i. Que mesure-t-il ?
4. Rappeler la définition de la Distance de Cook. Que mesure-t-elle ? Quel lien existe-t-il entre le levier et la distance de Cook ?

Partie 4 : Prédiction

L'ingénieur retient le modèle `mod2`. Il veut désormais prédire la consommation d'eau de l'usine le mois prochain, pour lequel il sait que la production envisagée est de 5000 kg avec 180 employés.

1. Donner une estimation de la quantité d'eau consommée le mois prochain, ainsi que l'expression de son intervalle de confiance de niveau 5%.
2. Donner l'expression de l'intervalle de prévision pour la quantité d'eau consommée le mois prochain.
3. La société envisage d'acheter une nouvelle machine et de produire 5000 kg avec 100 employés. Le modèle permet-il de prévoir la future consommation d'eau ?