TP 3 - séance en salle

Aurélien Garivier

24 février 2011

Le contenu du TP, les données nécessaires et les hyperliens sont accessibles en ligne à l'adresse suivante :

`http://perso.telecom-paristech.fr/~garivier/centrale/`

1 Utilisation de R et modèle linéaire

Si vous n'avez pas encore pu tester certaines des expériences numériques vues en cours, profitez-en pour le faire, et pour poser vos questions. En particulier, vous pourrez reprendre le travail effectué sur le fichier `ozone.txt`. Vous trouverez dans ces fichiers tous les éléments de syntaxe sur le langage R pour traiter la suite du TP.

2 Analyse de la covariance

Le fichier `eucalyptus.txt` contient des mesures portant sur 1429 arbres. Pour chacun, sont reportés sa hauteur (variable `ht`), sa circonférence (variable `circ`) et la partie du champ dont il provient (variable `bloc`).

- Représentez la hauteur des arbres en fonction de la circonférence. Une régression linéaire simple, du type

$$ht = \alpha_0 + \alpha_1 circ\,, \tag{1}$$

  vous semble-t-elle adaptée ? Réalisez-là, et notez le coefficient de détermination obtenu.
- On propose le modèle suivants :

$$ht = \alpha_0 + \alpha_1 circ + \alpha_2 \sqrt{circ} + \epsilon$$

  Réaliser la régression, puis déterminer s'il est préférable à la régression linéaire simple.
- Ce modèle est-il préférable aux modèles alternatifs suivants :

$$ht = \exp(\alpha_0 + \alpha_1 circ) + \epsilon$$

$$ht = \alpha_0 + \alpha_1 circ + \alpha_2 \sqrt{circ} + \alpha_3 circ^2 + \epsilon$$

– On garde désormais le modèle linéaire simple (1), mais on souhaite savoir si le coefficient α_1 dépend du bloc dont provient l'arbre (pour α_0, on suppose que ça n'est pas le cas). Ecrire le modèle linéaire correspondant. Il peut être résolu avec R grâce aux commandes suivantes :

```
> arbres <- read.table("eucalyptus.txt", header=T, row.names=1)
> arbres[,"bloc"]<- as.factor(arbres[,"bloc"])
> mod <- lm(ht ~ bloc:circ, data=arbres)
```

Etudier ce que font ces commandes, puis regarder ce qu'on obtient. Grâce à la commande `anova`, déterminer si ce modèle est préférable à (1).

3 Validation du modèle

Après avoir effectué une régression linéaire avec la commande `lm`, et sauvé le résultat dans la variable `modlin`, essayez la commande suivante :

```
> plot(modlin)
```

Elle produit des graphes qui permettent une validation graphique du modèle linéaire gaussien :

– Résidus studentisés
– Test visuel de normalité : q-q plot
– Influence des points sur l'estimation : distance de Cook

Essayez de comprendre ces graphes, avec l'aide de R et les explications de l'enseignant, et construisez des cas particuliers permettant de mettre en lumière l'intérêt de chacun d'entre eux.