

Modèles à effets mixtes : un exemple simple

Des courbes de croissance figurent en annexe de tous les carnets de santé. Nous allons dans cet exercice construire des courbes de croissance pour des enfants âgés de 3 à 14 ans. On dispose des données obtenues au cours d'un suivi longitudinal effectué sur un échantillon de 200 enfants : 100 garçons et 100 filles. La taille (en cm) de chaque enfant a été mesurée tous les ans à une date proche de son anniversaire. Il n'existe aucun lien de parenté entre les enfants de cette enquête. Le fichier *taille.txt* est un fichier ASCII structuré de la façon suivante :

- la première colonne (intitulée *Ident*) contient un identificateur de l'enfant. Il s'agit d'un nombre entier compris entre 1 et 200,
- la seconde colonne (intitulée *age*) contient l'âge exprimé en année,
- la troisième colonne (intitulée *taille*) contient la taille en cm,
- enfin, la dernière colonne (intitulée *sexe*) contient une codification du sexe de l'enfant : 1 pour les garçons et 2 pour les filles.

Ce fichier ne contient pas d'erreur de saisie.

Cet exercice est très directif : les différentes étapes de l'analyse vous seront suggérées et des questions intermédiaires vous permettront de répondre aux questions d'intérêt. Malheureusement, les analyses statistiques ne sont pas aussi balisées dans la vraie vie...

Certaines commandes R dont vous aurez besoin sont indiquées dans le texte en *italique*.

1. Charger le fichier de données dans R.

Vous pouvez utiliser les commandes suivantes :

```
taille1 <- read.table("C:\\chemin\\taille.txt", header=TRUE)
```

chemin doit contenir l'adresse du dossier contenant le fichier *taille.txt*. Les noms de dossier doivent être séparés par des \\

Cette commande construit un tableau de données intitulé *taille1* avec les données du fichier *taille.txt*. L'option *header = TRUE* signifie au logiciel que la première ligne contient les intitulés des colonnes correspondantes.

```
attach(taille1)
```

est une commande qui rend directement accessibles les variables du tableau de données.

2. Représentation graphique des données

Représentez graphiquement les courbes de taille en fonction de l'âge : *plot(age, taille, type="b")*

Vous pouvez aussi faire les graphiques pour chaque sexe. Pour obtenir la taille en fonction du poids pour les garçons (ie *sexe=1*) tapez : *plot(age[sexe==1], taille[sexe==1], type="b")*

Voir la doc de R pour les autres types de graphique

1. Quelle est l'allure générale des courbes ?
2. Essayez de trouver une courbe simple qui pourrait résumer cette allure générale pour tous les individus.
3. Les données vous semblent-elles homoscedastiques ?
4. Proposez une explication ?
5. Pour les deux sexes effectuer un "stem and leaf" de la taille des enfants de 3 ans et 14 ans. Par exemple, pour les garçons de 3 ans tapez : *stem(taille[sexe==1 & age == 3])*
6. Que suggèrent ces stem & leaf ?

3. Modélisation (dans cette partie on raisonnera conditionnellement à l'âge)

1. A partir de votre analyse graphique, proposez, pour chaque individu, un modèle qui décrit l'évolution de la taille en fonction de l'âge. Vous vous attacherez à préciser les notations que vous utilisez (ce que représente chaque indice, chaque lettre, chaque effet) ainsi que les hypothèses que vous effectuez :
 - liste des paramètres (fixes) du modèle,
 - liste des variables aléatoires, hypothèses sur leur distribution, hypothèses sur leur indépendance etc...
2. En utilisant ce modèle, calculez la variance de la taille d'un individu particulier (conditionnellement au vecteur de ses "paramètres individuels"). Ce modèle conditionnel est-il homoscedastique ? Expliquez ce que cela signifie pratiquement ? Cette homo(hétéro)-scédasticité vous semble-t-elle raisonnable ?
3. Calculez la covariance entre des tailles obtenues à des âges différents sur un même individu (conditionnellement au vecteur de ses "paramètres individuels"). Quelle est la conséquence de ce résultat ? (regardez les hypothèses que vous avez effectuées en 3.1)
4. En utilisant ce modèle, calculez la variance de la taille. Ce modèle est-il homoscedastique ? Permet-il d'expliquer ce que vous avez vu au cours de l'analyse graphique des données (question 2.3) ?

5. Calculez la covariance entre des tailles obtenues à des âges différents. Quelle est la conséquence de ce résultat ? (regardez les hypothèses que vous avez effectuées en 3.1)
6. Chargez la bibliothèque NLME de R en utilisant la commande : `library(nlme)`.

Le sexe est codé avec des 1 et des 2. Pour un logiciel il s'agit de nombres avec lesquels il peut faire des calculs. Il est donc nécessaire dans un premier temps de déclarer sexe comme un facteur (ie une variable explicative qualitative). On crée une nouvelle variable intitulée `sex` qui sera cette fois ci déclarée comme un facteur avec la commande : `sex<-factor(sexe)`

Rq : si votre ordinateur ne dispose pas de beaucoup de mémoire vous pouvez taper `sexe<-factor(sexe)` à la place de la commande précédente.

On dispose de plusieurs tailles variant avec l'âge pour chaque individu. Afin de déclarer cette structuration particulière des données il faut créer un nouveau tableau de données, que nous nommerons `taille2`, en utilisant la commande

```
taille2<-groupedData(taille~age|Ident)
```

Tapez `plot(taille2)` commentez le graphique obtenu.

Dans un premier temps, on omettra les possibles différences entre les courbes de croissance des garçons et des filles. A l'aide de la commande `lme`, dont la syntaxe a été détaillée en cours, estimez les paramètres du modèle que vous avez écrit en 3.1. Vous devez taper une commande de la forme

```
lme1<-lme(taille~modele1,data=taille2,random=~modele2)
```

où `modele1` est la représentation du modèle que vous avez écrit en 3.1 et `modele2` est la déclaration des facteurs à effets aléatoires.

Tapez ensuite `plot(lme1)`

```
qqnorm(lme1, ~ resid(., type = "p"), abline = c(0, 1))
```

```
qqnorm(lme1, ~ranef(.))
```

En quoi ces graphiques sont ils indispensables ?

Interpréter ces graphiques.

4 Utilisation de la modélisation (dans cette partie on raisonnera conditionnellement à l'âge)

1. La commande `summary(lme1)` permet d'afficher l'estimation des paramètres du modèle défini en 3.6. Interprétez chaque terme de la sortie.
2. A l'aide des estimations des paramètres, une prédiction de la taille pour un enfant de 5 ans. Proposez des intervalles de prédiction de sécurités 0.50 et 0.95 de la taille des enfants âgés de 5 ans.
3. Un enfant a 4 ans et il mesure 89.7 cm. Sa taille à 3 ans était de 86.2 cm. On souhaite calculer sa taille la plus probable lorsqu'il aura 5 ans. Expliquez pourquoi la taille prédite dans cette question est différente de celle obtenue en 4.2. Quel calcul faut il effectuer pour déterminer cette taille ? Le calcul précédent donne (69.42, 5.33) , déduisez-en la taille de cet enfant à 5 ans.
4. Ecrivez un modèle dans lequel l'effet du facteur sexe intervient. Vous devez taper une commande de la forme

```
lme2<-lme(taille~modele3,data=taille2,random=~modele3).
```

Utilisez les graphiques de diagnostic que vous connaissez,

Après avoir montré l'emboîtement des modèles que vous avez écrit en 3.6 et 4.6, faites un test de rapport de vraisemblance en tapant : `anova(lme1,lme2)`

La croissance des garçons est elle différente de celle des filles entre 3 et 14 ans ?