**Exercice SQUID**

Cette étude vise à explorer la variation saisonnière de la reproduction chez des organismes marins notamment chez une espèce de calmar (*Loligo forbesi*). Si cette étude permet d’explorer plusieurs variables sur des femelles et males, nous analyserons comment la longueur du manteau dorsal (DML) et le mois (MONTH) affectent le poids des gonades (Testisweight).

* Réaliser le modèle suivant :

Testisweighti = intercept + DMLi + Monthi + DMLi x Monthi + residualsi

* Que constatez-vous et quelles solutions proposez-vous ?

On veut lire la table

dataSquid<-read.table("C:\\Users\\olivier.pays\\Documents\\enseignement\\M2 Analyses de donn??es 3\\Exercices Ateliers\\Atelier 3\\Squid.txt", header=T, dec=".")

head(dataSquid)

NA 🡪 lignes vides à supprimer

Pour supprimer :

Nom=nom[-nrow(nom),]

View nom

on teste la normalité

hist(nom$variable que je veux)

puis

Dth<-rnorm(1000, mean(Donnees),sd(Donnees))

On met les données en catégorielle : corrélation entre l’âge et le poids

dataSquid$fMONTH <- factor(dataSquid$MONTH)

on regarde la distribution de la variable

hist(dataSquid$Testisweight)

*les résidus = plus la distance entre le point et la droite est faible, mieux c’est*

modèle demandé :

M1 <- lm(Testisweight ~ DML\*fMONTH,data=dataSquid)

OU

m1<-lm(Testisweight~DML+fMONTH+DML\*fMONTH,data=Donnees)

cela implique que les résidus suivent une loi normale avec une moyenne de 0 et une variance de 0²

on souhaite un nuage de points

plot(Donnees$DML,Donnees$Testisweight)

boxplot(Testisweight~fMONTH, data=Donnees)

ici, la variable n’est pas constante, il y a une variation saisonnière des gonades.

De même, en regardant les résidus, il y a un soucis dans l’homogénéité de la variance, on cherche donc la source du problème

par(mfrow=c(2,2))

on divise l’endroit des graphiques en 4 (2 lignes et 2 colonnes)

plot(m1,which = c(1),col=1,add.smooth = FALSE,caption = "")

#which = c(1) = choisi le 1er graph sur les 4

#col=1 = sur la case 1

#add.smooth = FALSE pas de ligne rouge sur la moyenne

Le caption = “” signifie que l’on ne veut pas de titre pour gagner de la place

plot(Donnees$fMONTH,resid(m1),xlab="Month",ylab="Residuals")

on veut voir les résidus de fMONTH par rapport à la moyenne m1

plot(Donnees$DML,resid(m1),xlab="DML",ylab="Residuals")

on veut voir les résidus de DML par rapport à la moyenne m1

##On voit que DML a cet effet car la distrib DML correspond bcp au 1er graph

**OPTION 1 : The fixed variance structure**

##On va dire à la formule des résidus que la variance dépend de DML, plus DML est grand plus la variance est grande (proportionnelle, linéaire)

#va permettre de corriger l'homogénéité

On utiliser la library

library(nlme)

mlm<-gls(Testisweight~DML\*fMONTH,data=Donnees)

vf1fixed<-varFixed(~DML)

##Pour que la variance dépend de DML = que la variance soit proportionnelle

mgls1<-gls(Testisweight~DML\*fMONTH, weights = vf1fixed, data=Donnees)

on ajoute weight pour que cela soit pris en compte malgré la variance fixée

summary(mgls1)

le summary permet de voir si R a bien compris ce que je lui demande

par(mfrow=c(1,1))

pour ouvrir une fenêtre de graph

plot(mgls1,which=c(1))

pour avoir le graph

hist(resid(mgls1))

pour avoir l’histogramme

en regardant l’histogramme, on observe que cela suit une loi normale

**OPTION 2 : The VarIdent variance structure**

On fait la même chose mais avec month (qui est une variable catégorielle)

vf2<-varIdent(form=~1|fMONTH)

mgls2<-gls(Testisweight~DML\*fMONTH,weights = vf2,data=Donnees)

« weights » : on n’oublie pas de prendre en compte le poids

summary(mgls2)

plot(mgls2,which=c(1))

hist(resid(mgls2))

C’est un peu différent de l’option 1

**OPTION 3 : The varPower Variance structure**

Cela permet d’observer une loi de puissance 🡪 + DML augmente, + la variance augmente et ce, sans proportionnalité

vf3<-varPower(form=~DML)

##varPower(form=~DML) formule de puissance

mgls3<-gls(Testisweight~DML\*fMONTH,weights = vf3, data=Donnees)

summary(mgls3)

plot(mgls3)

##rend la variance encoure plus homgène

hist(resid(mgls3))

**OPTION 5 : on essai avec la fonction exponentielle aka The VarExp Variance structure**

vf5<-varExp(form=~DML)

mgls5<-gls(Testisweight~DML\*fMONTH,weights = vf5, data=Donnees)

summary(mgls5)

plot(mgls5)

hist(resid(mgls5))

On observe que l’option de puissance est mieux car elle prend plus de valeurs.

**OPTION 4 : The varPower Variance structure**

On fait l’option puissance sur DML et sur MOIS (on mixte l’option 2 et 3)

vf4<-varPower(form=~DML|fMONTH)

mgls4<-gls(Testisweight~DML\*fMONTH,weights = vf4, data=Donnees)

summary(mgls4)

plot(mgls4)

hist(resid(mgls4))

##Pour sélectionner le meilleur modèle => sélection sur critères

##approximation sur le modèle qui rpz la réalité => vraissemblance => calcul de la distance en la réalité et le modèle

## il faut un AIC le plus bas et le comparer au modèle nul avec aucune correction (les AIC avec une différence plus petite que 2 sont équivalents ex : 210 et 212 seront équivalents)

AIC(mlm,mgls1,mgls2,mgls3,mgls4,mgls5)

##le modèle 4 est le mieux

On valise graphiquement le modèle avec :

E2<-resid(mgls4,type="normalized")

coplot(E2~DML|fMONTH,data=Donnees,ylab = "norm",xlab="month")