Travaux Pratiques
La Régression Linéaire

1 Introduction

On commence par charger les données. Il s'agit de mesures de concetration de minerais sur 359 sites dans le jura. On choisit de modéliser la variable Nickel (Ni) à l'aide d'autres variables. Les unités des différentes variables sont les mêmes.

```
jura=read.table("http://informatique-mia.inra.fr/biosp/sites/
informatique-mia.inra.fr.biosp-d7/files/jura.txt",header=T)
```

2 Analyses descriptives

Une analyse préliminaire à la modélisation est de réaliser quelques analyses descriptives à l'aide des commandes pairs(), boxplot et cor().

```
dim(jura)
boxplot(jura)
pairs(jura)
round(cor(jura),2)
```

Question : Quelle variable peut-on proposer pour faire une régression simple pour la variable Ni ?

3 Le test de corrélation

On veut tester la significativité de la corrélation entre Ni et Co. Pour cela, on utlise la commande `cor.test` en exécutant :

```
cor.test(jura$Ni,jura$Co)
```

Question : Retrouver la corrélation avec la formule page 6 du cours et la commande `cor(.,.)`
Question : Retrouver les résultats de cette commande en utilisant les formules données dans le cours (page 6 et 7)

4 Un modèle de régression simple

4.1 Choix du modèle

On choisit de s'intéresser aux données brutes (pas de transformation de la variable Ni) et au modèle

$$\mathcal{M}_1 : Ni = a + b \times Co.$$

Voici les commandes pour réaliser la régression et obtenir différentes informations.

```
res1=lm(Ni~Co,jura)
summary(res1)
```

Question : Retrouver l'estimation de a et b (page 13) Sur un même graphique, représenter les données, l'estimation obtenue pour le modèle et les valeurs prédites par le modèle
Question : Retrouver la valeur de la statistique pour le paramétre b **Question :** Retrouver la p-valeur associée au test sur le paramètre de pente de la régression.
Question : Retrouver le coefficient de détermination R^2 (page 24) à partir de somme de carrés et vérifier que $R^2 = \rho^2(Zn, Ni)$.

4.2 Analyse des résidus

Faire une analyse des résidus de la régression $Ni \sim Co$: représenter les résidus en fonction de la variable explicative Co, en fonction de la variable modélisée Ni et utiliser la fonction *qqnorm()*.
Question : Commenter ces résultats.

5 La régression multiple

On souhaite passer à la régression multiple. On se demande donc quelle variable ajouter au modèle de régression simple.

5.1 Le coefficient de régression partielle

On donne ci-dessous le code pour calculer la corrélation partielle entre y et x en supposant z fixé.

```
corp<-function(y,x,z)
{
  n<-cor(y,x)-cor(y,z)*cor(x,z)
  d<-sqrt((1-cor(y,z)^2)*(1-cor(x,z)^2))
  r<-n/d
  print(r)
}
```

5.2 Choix d'une nouvelle covariable

Calculer les corrélations partielles entre la variable Ni et les variables Cd, Cr, Cu, Pb et Zn en supposant Co fixé.
Question : Quelle covaribale peut-on choisir pour réaliser une régression multiple ? Quel serait ce choix si le critère était la corrélation simple ?

5.3 Un premier modèle de régression multiple

On choisit le modèle de régression $\mathcal{M}_2 : Ni = a + a_1 \times Co + a_2 \times Cr$

```
res2=lm(Ni~Co+Cr,jura)
summary(res2)
```

Question : Faire l'étude des résidus et conclure.
Question : Comparer le R^2, le R^2 ajusté et les variances estimées entre $\mathcal{M}_1$ et $\mathcal{M}_2$.

5.4 Interprétation du coefficient de corrélation partielle

On souhaite un peu mieux comprendre le principe du coefficient de corrélation partielle. Pour cela, on demande d'exécuter les commandes suivantes.

```
tmp=lm(Cr~Co,jura)
cor(res1$res,tmp$res)
corp(jura$Ni,jura$Cr,jura$Co)
```

Question : Commenter ces résultats.

5.5 Supplément : le coefficient de corrélation partielle d'ordre supérieur à 1

On souhaite à nouveau ajouter une nouvelle variable au précédent modèle. On définit le coefficient de corrélation partielle entre x et y conditionnellement à z_1 et z_2. On le note $\rho_{y,x|z_1,z_2}$. Voici la démarche pour le calculer :

1. On retire de y l'information de z_1 et z_2

$$e_1 = y - (\hat{a}_0 + \hat{a}_1 z_1 + \hat{a}_2 z_2)$$

2. On retire de x l'information de z_1 et z_2

$$e_2 = x - (\hat{b}_0 + \hat{b}_1 z_1 + \hat{b}_2 z_2)$$

3. On calcule la corrélation simple entre les 2 résidus

$$\rho_{y,x|z_1,z_2} = \rho_{e_1,e_2} = cor(e_1, e_2)$$

Question : Calculer $\rho_{N_i,Cd|Co,Cr}$ et $\rho_{N_i,Zn|Co,Cr}$ et commenter.

On réalise les modèles de régression suivants :

```
res3=lm(Ni~Co+Cr+Zn,jura)
summary(res3)
res4=lm(Ni~Co+Cr+Zn+Cd,jura)
summary(res4)
```

Question : Commenter.

6 Comparaison/choix de modèles

On s'intéresse aux modèles suivants :

```
res3=lm(Ni~Co+Cr+Zn,jura)
summary(res3)
res4=lm(Ni~Co+Cr+Zn+Cd,jura)
summary(res4)
res5=lm(Ni~Co+Cr+Cd+Zn+Pb,jura)
summary(res5)
```

Question : Que remarque-t-on ?

6.1 test de comparaison de 2 modèles.

On veut donc comparer le modèle de régression de res5 et res3

```
anova(res3,res5)
```

Question : Que peut-on conclure ?

6.2 Le critère AIC

On compare toujours les modèles, mais à l'aide du critère AIC

```
AIC(res1)
AIC(res2)
AIC(res3)
AIC(res4)
AIC(res5)
```

Question : Que concluez-vous ?
Question : Refaire la démarche avec le critère BIC.

Les formules de l'AIC et BIC sont :

```
AIC=-2*logLik+2*k
BIC=-2*logLik+n*k
```

avec logLik la log-vraisemblance des paramètres (voir la fonction logLik() de R), n est le nombre d'individus observés et k le nombre de paramètres.

Question : Retrouver les valeurs AIC et BIC pour res3

6.3 Le package MuMin

Le package MuMIn (MultiModelInference) permet d'aborder le problème de sélèctions de modèles en utlisant des critères de vraisemblance pénalisés tels que le AIC, le BIC ou le AICc.

Installer le package MUMIn : `install.packages("MuMIn")` puis le charger `library(MuMIn)`

```
options(na.action = "na.fail")
toto <- lm(y ~ ., data = jura)
ddAIC <- dredge(toto,rank=AIC)
ddAIC
subset(ddAIC, delta < 4)
```

Question : Que peut-on conclure ?
Question : Refaire la même analyse en utilisant le critère BIC. Que remaque t'on par rapport à une sélèction basée sur le AIC ? par rapport à l'approche basée sur la corrélation partielle ?