TP d'Econométrie II avec le logiciel R

Moindres carrés non-linéaires

L'objectif de ce TP est de se familiariser avec l'étude des modèles semi-paramétriques non-linéaires et qui ne peuvent pas se transformer en un modèle linéaire.

Un exemple simulé

On commence par étudier un modèle souvent utilisé pour modéliser les croissances de populations. Ainsi, tape-t-on les commandes suivantes:

```
library(car)
n=100
x=1:n; eps=rnorm(n,0,1)
Y=4/(1+exp(-(-5+0.3*x)))+eps
ts.plot(Y)
dataY=data.frame(x,Y)
Y.nls=nls(Y ~ theta1/(1+exp(-(theta2+theta3*x))),
start=list(theta1 = 10, theta2 = 0, theta3 = 0.1),data=dataY,trace=TRUE)
summary(Y.nls); plot(Y.nls)
```

Quel est le modèle simulé? Comment sont calculées les p-values? Que concluez-vous? Changer l'initialisation de l'algorithme. Changez également le bruit du modèle. Que constatez-vous? Augmentez la taille de l'échantillon. Que constatez-vous? Faire une boucle de manière à tracer des histogrammes des estimateurs. Quel est le coefficient de détermination R^2 pour une telle estimation?

Un exemple sur données réelles

On travaille sur une base de données précisant l'évolution du poids de rats (l'étude statistique qui suit est dûe à Marc Lavielle, DR INRIA Polytechnique).

```
data.rat <- read.csv("ratWeight.csv")
data.select <- subset(data.rat, gender=="Female" & regime=="Control")
library(ggplot2);  theme_set(theme_bw())
ggplot(data=data.select, aes(x=week, y=weight)) + geom_point() + geom_line() + facet_wrap(~id)
```

Qu'a-t-on fait? Que déduire?
On s'intéresse à un rat particulier:

```
di <- subset(data.select, id=="B38837")
pl <- ggplot(data=di) +  geom_point(aes(week,weight), size=3, colour="#993399"); pl
poly2 <- lm(weight ~poly(week,2,raw=T), data=di)
poly3 <- lm(weight ~poly(week,3,raw=T), data=di)
poly4 <- lm(weight ~poly(week,4,raw=T), data=di)
poly5 <- lm(weight ~poly(week,5,raw=T), data=di)
poly.df <- data.frame(week=seq(0,18,by=0.5))
poly.df$p2 <- predict(poly2,poly.df)
poly.df$p3 <- predict(poly3,poly.df)
poly.df$p4 <- predict(poly4,poly.df)
poly.df$p5 <- predict(poly5,poly.df)
```

```
library(reshape2)
poly.df <- melt(poly.df, id="week", variable.name = "model", value.name = "prediction")
print(ggplot() +
        geom_line(data=poly.df, aes(x=week,y=prediction,colour=model), size=0.75) +
        geom_point(data=di, aes(x=week,y=weight), colour="black", size=3) +
        geom_point(data=di, aes(x=week,y=weight), colour="white", size=2) )
merge(AIC(poly2,poly3,poly4,poly5),BIC(poly2,poly3,poly4,poly5))
```

Qu'a-t-on fait? Que pensez vous de cette modélisation?
On utilise désormais un modèle non-linéaire:

```
nls1 <- nls(weight ~ A*exp(-b*exp(-k*week)), data=di, start=c(A=500, b=0.8, k=0.1))
summary(nls1)
nlreg.df <- data.frame(week=seq(0,25,by=0.5))
nlreg.df$f1 <- predict(nls1,nlreg.df)
pl +  geom_line(data=nlreg.df, aes(x=week,y=f1), colour="#339900", size=1)
```

Quel est le modèle? Qu'en pensez-vous?
On propose deux autres modèles non linéaires:

```
nls2 <- nls(weight ~ A*(1-b*exp(-k*week)), data=di, start=c(A=500, b=0.8, k=0.1))
nls3 <- nls(weight ~ A/(1+ exp(-gamma*(week-tau))), data=di, start=c(A=500, gamma=0.8, tau=0))
nlreg.df$f2 <- predict(nls2,nlreg.df)
nlreg.df$f3 <- predict(nls3,nlreg.df)
nlreg.df <- melt(nlreg.df, id="week", variable.name = "model", value.name = "prediction")
pl + geom_line(data=nlreg.df, aes(x=week,y=prediction,colour=model), size=0.75)
cbind(AIC(nls1,nls2,nls3),BIC(nls1,nls2,nls3))
par(mfrow=c(1,2))
plot(predict(nls2), di$weight)
abline(a=0, b=1, lty=1, col="magenta")
residual.nls2 <- resid(nls2)/sd(resid(nls2))
plot(predict(nls2), residual.nls2)
abline(a=0, b=0, lty=2, col="magenta")
```

Quelle conclusion?

Exercice

Recommencer avec une simulation du modèle $Y_i = \theta_1 + \theta_2 * x_i^{\theta_3} + \varepsilon_i$. On commencera par le cas $\theta_1 = 2$, $\theta_2 = 0.1$, $\theta_3 = 0.2$ et $\varepsilon_i \sim \mathcal{U}[-5, 5]$. Estimer les paramètres et interpréter les tests et graphique. Faire ensuite une transformation de Box-Cox pour retrouver θ_3 et ensuite effectuer une régression linéaire pour obtenir θ_1 et θ_2. Conclusions? Recommencer plusieurs fois.