TP d'Econométrie II avec le logiciel R

Sélection de variables pour le modèle linéaire

L'objectif de ce TP est de mettre en application les méthodes permettant de sélectionner des variables potentiellement explicatives. Ceci peut se faire par ses suites descendantes de tests de Fisher imbriqués, ou bien par des critères de sélection de modèles prédictifs (AIC, BIC et Cp).

Tests de Ficher imbriqués

Supposons que nous disposons des valeurs prises par n (par exemple $n = 100$) réalisations de six variables $Z^{(j)}$, $j = 1, \cdots, 6$, précisées ci-dessous. On simule les réalisations Y_i de la variable à expliquer Y, en ayant décidé arbitrairement que Y ne dépendrait linéairement que de $Z^{(2)}$, $Z^{(4)}$ et $Z^{(6)}$, c'est-à-dire que l'on simule les réalisations:

$$Y_i = \beta_0^* + \beta_2^* Z_i^{(2)} + \beta_4^* Z_i^{(4)} + \beta_6^* Z_i^{(6)} + \varepsilon_i^* \quad \text{pour} \quad i = 1, \cdots, n, \tag{1}$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma_*^2)$ pour tout $i = 1, \cdots, n$, les ε_i^* étant indépendants les uns les autres. Les différentes valeurs choisies des coefficients sont: $\beta_0^* = 5$, $\beta_2^* = -0.003$, $\beta_4^* = 8$, $\beta_6^* = 0.03$ et $\sigma_*^2 = 5^2$. Ceci constituera ce que nous appellerons dorénavant *le vrai modèle* (nous réserverons au vrai modèle la signalisation * pour le distinguer). On supposera donc maintenant que sont connues les différentes valeurs Y_i et $Z_i^{(1)}, \cdots, Z_i^{(6)}$ pour $i = 1, \cdots, n$. En revanche, le "vrai" modèle lui-même, c'est-à-dire les variables potentiellement explicatives intervenant vraiment (soit $Z^{(2)}$, $Z^{(4)}$ et ,$Z^{(6)}$) n'est pas connu. Comme toujours, les ε_i, la variance σ_*^2, les valeurs des différents coefficients β_j^*, sont inconnus. On va donc essayer de "retrouver" ce modèle par une suite de test de Fisher, et on pourra alors disposer d'estimation des différents paramètres :

```
set.seed(2022)
n=100; i=1:n;
Z1=i;Z2=i^2;Z3=runif(n,-10,10);Z4=cos(i/10);Z5=1/sqrt(i);Z6=i*rchisq(n,5);
epsilon=rnorm(n,0,5);
Y=5-0.002*Z2+8*Z4+0.03*Z6+epsilon;
j=1:(n+50);
ZZ6=c(Z6,(n+c(1:50))*rchisq(50,5))
YY=5-0.002*(j^2)+8*cos(j/10)+0.03*ZZ6
plot(j,YY,"l",xlim=c(-10,160),ylim=c(-40,40))
points(i,Y)
```

Commenter les différentes commandes exécutées. On effectue alors une première régression, dont on peut éliminer la variable la moins significative, c'est-à-dire celle possédant la p-value au test de Student la plus forte (et si celle-ci est supérieure à 0.05). On recommence alors une régression avec toutes les variables sauf celle qui vient d'être éliminée. Puis, si besoin, on élimine la variable la moins significative... et on renouvelle les mêmes étapes jusqu'à ce qu'aucune p-value ne soit supérieure à 0.05. Cela peut se faire par les commandes suivantes:

```
model1=lm(Y~Z1+Z2+Z3+Z4+Z5+Z6)
summary(model1)
model2=update(model1,~. -Z1)
summary(model2)
model3=update(model2,~. -Z5)
summary(model3)
```

On aboutit alors à un modèle sélectionné, dont on peut également tracé les prédictions sur le graphe précédent. Conclusions?

Premier exercice

Comparer les graphiques de contrôle de la régression complète et de la régression avec le modèle sélectionné. Comparer également les paramètres estimés et les vrais paramètres. Aboutit-on (parfois? toujours?) au vrai modèle? Pour répondre à cette question, on peut recommencer la simulation plusieurs fois, ainsi que les étapes de sélection de modèle. Conclusion?

Utilisation de critères de sélection de modèles prédictifs

On commence par utiliser le critère du Cp de Mallows. Cela peut se faire de la manière suivante:

```
library(leaps)
Z=matrix(c(Z1,Z2,Z3,Z4,Z5,Z6),ncol=6);
colnames(Z)=c("Z1","Z2","Z3","Z4","Z5","Z6");
r=leaps(Z,Y);
names(r)
r$which;
r$Cp;as.data.frame(ZZ)
t=(r$Cp==min(r$Cp));
cc=colnames(Z)[r$whi[t]]
cc
j=(n+1):(n+50)
new=data.frame(Z1=j,Z2=j^2,Z3=runif(50,-10,10),Z4=cos(j/10),Z5=1/sqrt(j),Z6=j*rchisq(50,5))
y.pred.cp=predict(regCp,new)
MSE=mean((YY-y.pred.cp)^2)
points(j,y.pred.cp,pch=19)
```

Quel modèle a été sélectionné? Expliquer les différentes variables associée au data.frame r. Comparer les prédictions avec celles obtenues par tests de Fisher imbriqués.
On utilise maintenant les critères BIC et AIC. Voici les commandes nécessaires:

```
library(MASS)
ZZ=as.data.frame(Z);
y.lm=lm(Y~.,data=ZZ);
y.bic=stepAIC(y.lm,k=log(n),direction = c("backward"))
y.bic2=stepAIC(y.lm,k=log(n),direction = c("both"))
y.aic=stepAIC(y.lm,k=2,trace=FALSE)
library(glmulti)
modBIC=glmulti(Y~Z1+Z2+Z3+Z4+Z5+Z6,crit=bic,level=1)
```

Quels modèles ont été sélectionnés? Expliquer les différentes résultats numériques affichés. Comparer les prédictions avec les précédentes.

Deuxième exercice

Recommencer la simulation plusieurs fois, et comparer les résultats obtenus. Comment pourrait-on faire sur cet exemple pour savoir le critère le plus efficace?

Troisième exercice

Recommencer la simulation en remplaçant cette fois-ci $Z^{(3)}$ par une réalisation d'une variable uniforme sur $[-4, 8]$? A quelle conclusions numériques aboutissez vous? Pourquoi?