Polytech'Lille/GIS4

Fiche TP Régression linéaire : sélection de modèle

C. Preda / A. Ehrhardt

L'objectif de ce TP est de mettre en œuvre sous R la sélection de modèles de régression linéaire. Le jeux de données considéré est décrit ci-dessous.

Description des jeux de données

donnees_cornell.xls

On trouve dans le livre de Michel Tenenhaus ([10] page 78) l'exemple suivant tiré de Cornell (1990). On cherche à connaître l'influence des proportions de sept composants sur l'indice d'octane moteur de douze différents mélanges d'essences. Les variables sont les suivantes :

- y : indice d'octane moteur
- x_1 : distillation directe (entre 0 et 0.21)
- x_2 : reformat (entre 0 et 0.62)
- x_3 : naphta de craquage thermique (entre 0 et 0.12)
- x_4 : naphta de craquage catalytique (entre 0 et 0.62)
- x_5 : polymère (entre 0 et 0.12)
- x_6 : alkylat (entre 0 et 0.74)
- x_7 : essence naturelle (entre 0 et 0.08)

TABLE 4.1 – *Données Cornell*

x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
0	0,23	0	0	0	0,74	0,03	98,7
0	0,1	0	0	0,12	0,74	0,04	97,8
0	0	0	0,1	0,12	0,74	0, 04	96,6
0	0,49	0	0	0,12	0,37	0,02	92
0	0	0	0,62	0,12	0,18	0,08	86,6
0	0,62	0	0	0	0,37	0,01	91,2
0,17	0,27	0,1	0,38	0	0	0,08	81,9
0,17	0,19	0,1	0,38	0,02	0,06	0,08	83,1
0,17	0,21	0,1	0,38	0	0,06	0,08	82,4
0,17	0,15	0,1	0,38	0,02	0,1	0,08	83,2
0,21	0,36	0,12	0,25	0	0	0,06	81,4
0	0	0	0,55	0	0,37	0,08	88,1

On demande :

1. Réaliser les statistiques descriptives univariées et bivariées (y versus les autres variables)

2. Réaliser le modèle de régression linéaire entre y et toutes les autres variables (fonction R : lm). Que constatez vous ?

3. Puisque n =12 > p=7, il ne reste qu'à vérifier qu'il n'y a pas une relation entre les variables explicatives (multi-collinéarité). En effet, les variables X's représentent les taux de chaque composante dans l'essence. Du coup, la somme sur ligne doit faire 100%. Vérifier (fonction *apply*). Donc on n'a pas besoin de toutes les 7 variables puisque 6 suffisent ! On calculera aussi le déterminant de la matrice X^TX (voir cours). Utiliser la foction *det* en R.

4. One ne peut pas donc faire un modèle avec toutes les variables. Mais lesquelles éliminer ? On procédera à une sélection des variables. Explorez la fonction *regsubsets* du package « leaps »

```
m = lm(Y~., data = d)
summary(m)

library(leaps)
choix = regsubsets(Y~., int= T, nbest = 1, nvmax=7, method="exh", data = d)
res = summary(choix)
print(res)

#choix du meilleur modèle selon le critère BIC
plot(choix, scale="bic")
```

4-a) Quel est le meilleur modèle ? Combien de variables fait-il rentrer dans la régression ?
Estimer ce modèle, analyser la validité et les performances du modèle complet (R^2, significativité coefficients).

4-b) quel est le meilleur modèle avec deux variables ?

5. Remplacer précédemment le critère BIC par le critère Cp, R^2 ajusté (adjr2) ou encore R^2 . (évidemment AIC donne le mêmes résultats que BIC à cause du lien entre les deux critères). Préciser pour chaque critère le meilleur modèle.

6. Les recherches précédentes étaient exhaustives. Cela pose un problème lorsque le nombre de variables est grand. Faisons une sélection de variables pas-à-pas.

```
library(MASS)
m_0 <- lm(Y ~ 1, data=d) # modele sans aucune variable explicative
m_all <- lm(Y ~ ., data=d) # modele avec toutes les variables
m_back = stepAIC(m_all,direction="backward")
m_forw = stepAIC(m_0,direction="forward",scope=list(upper=m_all,lower=m_0))
m_stepwise = stepAIC(m_0,direction="both",scope=list(upper=m_all,lower=m_0))
```

Comparer les 3 modèles obtenues selon leur pouvoir prédictif : PRESS (à le calculer grace au leviers h_ii) :

```
press=function (fit)
{
h=lm.influence (fit)$ h
return (sqrt (mean ((residuals (fit)/ (1- h))^2 )))  # voir cours
}
```