Régression Multiple et Sélection de Modèles

support Applications

Objectif de la session: Régression linéaire multiple. Sélection de modèles.
Rappel: `help(fonc)` pour obtenir de l'aide sur la fonction nommée "fonc".

Les différents Critères de sélection d'un modèle de régression:

Le tableau suivant présent une liste de critères habituellement utilisés pour caractériser un modèle de régression.

$RSS = \sum_i (y_i - \hat{y}_i)^2$;

Notation	Définition	Critère	Objectif	R
R^2	$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$	Détermination	-	Fonction `lm()`
R^2_{adj}	$R^2_{adj} = 1 - \frac{n-1}{n-p}(1-R^2)$	Détermination ajusté	Max. R^2_{adj}	`lm()`
$\hat{\sigma}^2_p$	$\hat{\sigma}^2_p = \frac{RSS(p)}{n-p}$	Estim. non biaisée des rés.	Min. $\hat{\sigma}_p$	Fonction `lm()`
AIC	$\simeq nlog(\frac{RSS}{n}) + 2p$	Information d'Aikaike	Min. AIC	`extractAIC()`
BIC	$\simeq nlog(\frac{RSS}{n}) + log(n)p$	Information Bayésien	Min. BIC	`extractAIC(,k=log(n))`
C_p	$= \frac{RSS(p)}{\sigma^2} - (n-2p)$	C_p de Mallows	Min. BIC	`regsubsets()`

La fonction `step()` de R est utilisée pour comparer et sélectionner un modèle parcimonieux (avec peu de variables) en optimisant le critère AIC. La fonction débute avec le modèle complet, puis élimine à chaque étape la variable pour lequel le critère AIC du modèle partiel est minimal (modèle sans la variable) et pour lequel le coefficient β est testé non significatif (seuil $\alpha = 0.1$). Le procédure s'arrête quand le coefficient de la variable candidate au retrait est significatif.

Illustration: Etude sur les données immobilières

Les critères ci-dessous sont calculés à partir des données de transaction immobilières du TD précédent.

Modèle	R^2	R^2_{adj}	$\hat{\sigma}_p$	`AIC()`	`extractAIC()`
$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$	0.8345	0.8149	34.73	203.40	144.6
$Y = \beta_1 X_1 + \epsilon$	0.8344	**0.8249**	**33.77**	**201.43**	**142.67**
$Y = \beta_2 X_2 + \epsilon$	0.72	0.7096	43.5	211.56	152.80

```
 step(res);
**Start:  AIC=144.65 Y ~ X1 + X2    Df Sum of Sq   RSS   AIC
                                    - X2    1        30 20530   143
                                    <none>              20499   145
                                    - X1    1     13560 34059   153
**Step:  AIC=142.68 Y ~ X1          Df Sum of Sq    RSS    AIC
                                    <none>              20530   143
                                    - X1    1    103257 123787   177
**Call: lm(formula = Y ~ X1, data = tab) Coefficients: (Intercept)->319.470   X1-> 2.750
```

Calcul automatique des critères de sélection emboitée

Etudier les fichiers "USCrimeinfo.txt" et "UsCrime.txt". La variable cible (Y) est la première variable colonne du fichier.

1. Charger le fichier dans l'environnement R en utilisant la fonction `tab=read.table()`. Quel est le nombre d'observations disponibles? Visualiser les nuages de points entre les variables. Que constate-t-on?
2. Calculer la matrice de corrélations. Interpréter le résultat. Utiliser les instructions graphiques proposées (section library `corrplot`) pour mettre en valeur les corrélations entre les variables.

3. **Modèle de régression multiple:** On souhaite étudier le modèle linéaire permettant d'expliquer la variable cible Y en fonction des autres variables disponibles (X). Expliciter formellement le modèle attendu.

   - Utiliser vos connaissances pour analyser (rapidement) le modèle de régression linéaire multiple où Y est la variable cible et (X) les variables explicatives au complet ($p = 14$). `reg=lm("R~.",data=tab)`.
   - Ce modèle est-il significatif globalement? (justifier)
   - Tester (rapidement) la significativité de chacun des coefficients du modèle (justifier).
   - Calculer la valeur RSS (somme quadratique des résidus) pour le modèle complet à $p = 14$ paramètres.
   - Calculer le critère AIC pour le modèle complet à l'aide de l'instruction `extractAIC(reg)`. Retrouver cette valeur en utilisant la définition du critère $AIC = -2L + 2p$, où L est la log vraisemblance du modèle et p le nombre de paramètres inconnus.
   - Pour indication, la fonction R `extractAIC(reg)` propose également le calcul du critère BIC. Consulter l'aide de la fonction pour calculer le critère BIC. Noter la principale différence entre les deux crtères.

     -Pour information il existe également sous R une fonction `AIC()` qui propose un calcul lègèrement différent de la fonction `extractAIC()`.

4. **Sélection de modèles:** Le but est ici de trouver un modèle parcimonieux (utilisant un nombre restreint p_0 de variables $p_0 < p$) tout en proposant un ajustement linéaire acceptable.

   (a) **Régression Backward.** Exécuter les instructions suivantes:

   ```
   regbackward=step(reg,direction='backward')
   summary(regbackward)
   ```

   Commenter les variables successivement éliminées. Décriver le modèle réduit sélectionné puis comparer le au modèle initial complet.

   (b) **Régression Forward.** Etudier la fonction `step()` de R. Puis exécuter les instructions suivantes:

   ```
   regforward=step(lm(R~1,data=tab),list(upper=reg),direction='forward');
   summary(regforward);
   ```

   Commenter les variables successivement sélectionnées. Vérifier le critère AIC proposé. Décriver le modèle réduit sélectionné puis comparer le au modèle initial complet, et au modèle sélectionné par la régression backward. Que constatez-vous? Quelles sont les limites de cette approche ?

   (c) **Régression Stepwise.** Exécuter les instructions suivantes:

   ```
   regboth=step(reg,direction='both')
   summary(regboth)
   ```

   Commenter les variables successivement selectionnées puis éliminées. Comparer les trois modèles de sélection.

   (d) Exécuter l'instruction `formula(s0)` où $s0$ est un objet retourné par la fonction step. Noter que l'instruction `reg0=lm(formula(s0),data=tab);` vous permet automatiquement de réappliquer et d'étudier, `summary(reg0)`, le modèle sélectionné.

   (e) Etudier l'aide de la fonction step() pour mettre en place une pénalisation de type BIC. Quel est le modèle obtenu? Conclusions.

   (f) **Sélection automatique et gestion des formules R** *-niveau avancé R, à réaliser à la fin du TP-***:** Analyser puis exécuter les instructions suivantes. Aidez-vous de la fonction `print()` si besoin.

   ```
   xnames=names(tab)[2:ncol(tab)];
   for (k in 1:length(xnames)) {
   print(sprintf('---> %s',xnames[k]));
   fk=paste("R ~", paste(xnames[-k], collapse= "+"));
   regk=lm(as.formula(fk),data=tab); print(summary(regk));
   mes=sprintf("%5s --> %5.2f",xnames[k],AIC(regk)); print(mes)}
   ```

   Ces instructions décrivent la première étape d'une méthode de sélection de modèles. Décriver la méthode proposée. Quelle est la valeur du critère AIC pour le modèle complet initial?

Recherche exhaustive et sélection de modèles

La librairie `leaps` de R permet d'effectuer une recherche exhaustive du "meilleur" modèle parmi tous les sous-modèles. Les deux fonctions principales de la librarie sont `leaps()` et `regsubsets()`. Une aide et disponible pour chacune de ces deux fonctions. **Précautions:** Les arguments de ces fonctions diffèrent des fonctions précedemment utilisées.

1. Charger la librairie dans l'environnement de travail et préparer les données:

```
library(leaps);
X=tab[,2:ncol(tab)]; xnames=names(X); Y=tab[,"R"];
```

2. **Recherche exhaustive de modèles. Fonction `leaps()`:**

   (a) **Critère** R^2. Etudier puis exécuter les instructions suivantes. Etudier l'objet `reg`. Modifier la valeur du paramètre `nbest=2`.

```
reg=leaps(X,Y,method="r2",nbest=1);reg
x11();plot(reg$r2,main='R2',xlab='nb modèle',type="both");
```

   • En étudiant le contenu de l'objet reg, indiquer le meilleur modèle (ici au sens du R^2) comportant 1 puis 3 variables.

   (b) **Critère** R^2 **ajusté**. Etudier puis exécuter les instructions suivantes.

```
reg=leaps(X,Y,method="adjr2",nbest=1);reg
x11();plot(reg$adjr2,main='adjR2',xlab='nb modèle',type="both");
```

   • A l'aide de l'évolution de R^2_{adj}, proposer un modèle réduit adapté.

   • En vous aidant des instructions suivantes, récupérer automatiquement les variables du modèle sélectionné, et estimer, puis tester les coefficients. Comparer ce modèle au modèle proposé par sélection emboitée utilisant le critère AIC puis BIC.

```
m0=6; ind0=reg$which[m0,];
fk=paste("R~",paste(xnames[ind0],collapse= "+"))
mes=sprintf('Modèle réduit (Cp) -> %s ',fk); mes
summary(lm(fk,data=tab));
```

   (c) In est également possible d'étudier le **Critère** C_p de Mallows. Etudier puis exécuter les instructions suivantes.

```
reg=leaps(X,Y,method="Cp",nbest=1);reg
x11();plot(reg$Cp,main='Cp',type="both");
```

   • En vous aidant des instructions suivantes, récupérer automatiquement les variables du modèle sélectionné, et estimer, puis tester les coefficients. Comparer ce modèle aux modèles proposés précédemment par sélection.

```
m0=which.min(reg$Cp);
ind0=reg$which[m0,]; fk=paste("R~",paste(xnames[ind0],collapse="+"))
mes=sprintf('Modèle réduit (Cp) -> %s ',fk); mes
summary(lm(fk,data=tab));
```

   • Recalculer le critère de Mallows pour le modèle à une variable et à p variables et comparer aux résultats proposés par R.

3. **Recherche exhaustive de modèles. Fonction `regsubsets()`:**

   (a) **Sélection Forward**. Etudier puis exécuter progressivement les instructions suivantes. Etudier l'objet `reg`.

   ```
   reg=regsubsets(X,Y,nbest=1,method="forward",nvmax=ncol(tab));
   summary(reg); mod=summary(reg); attributes(mod);
   ```

   • Comparer les critères de sélection à l'aide des instructions suivantes (attention, l'évolution du R^2 est donné à l'aide de la fonction leaps()):

   ```
   x11(); par(mfrow=c(2,2)); plot(mod$adjr2);
   plot(mod$cp); plot(mod$bic); plot(regleaps$r2);
   ```

   • Puis, comparer les modèles sélectionnés:

   ```
   x11();par(mfrow=c(2,2)); plot(reg,scale="r2",main='Backward');
   plot(reg,scale="adjr2"); plot(reg,scale="Cp");
   plot(reg,scale="bic");
   ```

   (b) Modifier la valeur du paramètre `nbest=2` et `nvmax=8`, et étudier de nouveaux modèles.

   (c) **Autres sélections**. Etudier puis comparer les autres méthodes de sélection: forward, seqrep, et exaustive.

library corrplot

La library `corrplot` de R offre des outils de visualisation intéressants de la matrice de corrélation.
Exécuter les instructions suivantes et observer le résultat. Pour rappel, vous pouvez obtenir des informations complémentaires en utilisant l'aide de la fonction `help(corrplot)`.

```
library("corrplot")
#visu avec des cercles de couleur
mycorr=cor(tab);

### Visualisation de type I
corrplot(mycorr)

#sauvegarde de l'image dans un fichier au format jpg
x11(); corrplot(mycorr)
savePlot("mygraph.jpg",type="jpg"); #sauvegarde sur le DD

### Visualisation de type II
corrplot(mycorr,method="shade",shade.col=NA,tl.col="black",tl.srt=45)

#sauvegarde de l'image dans un fichier au format png
x11(); corrplot(mycorr,method="shade",shade.col=NA,tl.col="black",tl.srt=45)
savePlot("mygraph.png",type="png")

### Visualisation de type III

corrplot(mycorr,method="shade",shade.col=NA,tl.col="black",tl.srt=45,
addCoef.col="black",addcolorlabel="no",order="AOE")

#sauvegarde de l'image dans un fichier au format eps
x11(); corrplot(mycorr,method="shade",shade.col=NA,tl.col="black",tl.srt=45,
addCoef.col="black",addcolorlabel="no",order="AOE")
savePlot("mygraph.eps",type="eps")
```