Économétrie 1, M1 MAEF, Partiel
14/01/2016, Durée : deux heures
Documents et téléphones portables ne sont pas autorisés.
L'usage de la calculatrice est autorisé.

Exercice 1. Régression linéaire

Description : Les données sont un échantillon aléatoire de dossiers de reventes de maisons à partir des fichiers utilisés par des agents immobiliers comme une base d'information.

Nombre de cas : 20
Les noms des variables :
PRICE (P) : prix de vente
SQFT (S) : feet carrés de surface habitable
AGE (A) : âge de la maison
FEATS (F) : nombre des 11 fonctions (lave-vaisselle, réfrigérateur, micro-ondes, broyeur, machine à laver, interphone, lucarne(s), compacteur, sèche-linge, matériel pour handicapé, accès à la télévision par câble)
TAX (T) : taxe annuelle

Vous êtes chargé(e) d'une étude visant à fournir un modèle de la taxe annuelle en fonction des variables sur la propriété de la maison (P, S, A et F).

Dans un premier temps, nous considérons le modèle à quatre variables explicatives. Les résultats sont présentés dans le tableau suivant.

```
Call:
lm(formula = TAX ~ PRICE + SQFT + AGE + FEATS)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 165.77265  182.41948   0.909 0.377855
PRICE         0.29756    0.06582   4.521 0.000406
SQFT          0.38491    0.06295   6.115 1.98e-05
AGE          -7.44801    1.99673  -3.730 0.002011
FEATS        25.03941   14.30090   1.751 0.100381

Residual standard error: 133.7 on 15 degrees of freedom
Multiple R-squared:  0.833,Adjusted R-squared:  0.7885
F-statistic:  18.7 on 4 and 15 DF,  p-value: 1.073e-05
```

1. Réécrivez le modèle en utilisant les résultats de régression.
2. Le modèle est il globalement significatif ?
3. Les coefficient ont-ils un signe conforme à vos attentes?
4. Testez la position du coefficient de PRICE par rapport à 0.3 avec le niveau de test 5%.
5. Y a t il une variable qui doit être éliminée ? Laquelle ?

Pour déterminer le vrai modèle, nous allons appliquer quatre méthodes : régression pas-à-pas descendante, les critères Cp de Mallows, AIC et BIC. Un extrait des résultats obtenus est fourni après.

6. Listez les variables qui sont retenues par chaque méthode.

```
> ############################
> # Régression descendante #
> ############################
```

```
> drop1(lm(TAX~PRICE+SQFT+AGE+FEATS),test="F")
Model:
TAX ~ PRICE + SQFT + AGE + FEATS
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              268001 200.06
PRICE   1    365177 633177 215.25 20.4389  0.000406 ***
SQFT    1    668079 936079 223.07 37.3924  1.978e-05 ***
AGE     1    248591 516591 211.19 13.9136  0.002011 **
FEATS   1     54773 322774 201.78  3.0656  0.100381

> drop1(lm(TAX~PRICE+SQFT+AGE),test="F")
Model:
TAX ~ PRICE + SQFT + AGE
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              322774 201.78
PRICE   1    354954 677728 214.62  17.595 0.0006861 ***
SQFT    1    670292 993066 222.26  33.227 2.903e-05 ***
AGE     1    267398 590171 211.85  13.255 0.0022013 **

> drop1(lm(TAX~PRICE+SQFT),test="F")
Model:
TAX ~ PRICE + SQFT
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>               590171 211.85
PRICE   1    314564  904735 218.39  9.0611 0.0078820 **
SQFT    1    680025 1270196 225.18 19.5883 0.0003701 ***

> drop1(lm(TAX~SQFT),test="F")
Model:
TAX ~ SQFT
       Df Sum of Sq     RSS    AIC F value   Pr(>F)
<none>               904735 218.39
SQFT    1    699976 1604711 227.85  13.926 0.001527 **
> ###############
> # Mallows Cp #
> ###############
> X=matrix(c(PRICE,SQFT,AGE,FEATS),ncol=4);
> r=leaps(X,TAX,nbest=1);
> r$whi;
      1    2     3     4
1 FALSE TRUE FALSE FALSE
2  TRUE TRUE FALSE FALSE
3  TRUE TRUE  TRUE FALSE
4  TRUE TRUE  TRUE  TRUE
> r$Cp;
[1] 34.638031 19.031901  6.065641  5.000000
> #############
> # step AIC #
> #############
> step(reg, k=2)
Start:  AIC=200.06
TAX ~ PRICE + SQFT + AGE + FEATS

        Df Sum of Sq    RSS    AIC
```

```
<none>               268001 200.06
- FEATS  1     54773 322774 201.78
- AGE    1    248591 516591 211.19
- PRICE  1    365177 633177 215.25
- SQFT   1    668079 936079 223.07

Call:
lm(formula = TAX ~ PRICE + SQFT + AGE + FEATS)

Coefficients:
(Intercept)        PRICE         SQFT          AGE
   165.7727       0.2976       0.3849      -7.4480
      FEATS
    25.0394

> ############
> # step BIC #
> ############
> step(reg, k=log(100))
Start:  AIC=213.09
TAX ~ PRICE + SQFT + AGE + FEATS

        Df Sum of Sq    RSS    AIC
- FEATS  1     54773 322774 212.20
<none>               268001 213.09
- AGE    1    248591 516591 221.61
- PRICE  1    365177 633177 225.68
- SQFT   1    668079 936079 233.50

Step:  AIC=212.2
TAX ~ PRICE + SQFT + AGE

        Df Sum of Sq    RSS    AIC
<none>               322774 212.20
- AGE    1    267398 590171 219.66
- PRICE  1    354954 677728 222.43
- SQFT   1    670292 993066 230.07
```

Exercice 2. Théorème de Gauss-Markov

Rappelons que pour le modèle linéaire, l'estimateur des moindres carrés $\hat{\theta}$ est optimal parmi les estimateurs linéaires sans biais. C'est-à-dire, si $\tilde{\theta}$ est un autre estimateur linéaire sans biais, alors pour C un vecteur réel quelconque de même taille que θ on a

$$\mathrm{Var}\,(C'\tilde{\theta}) \geq \mathrm{Var}\,(C'\hat{\theta}).$$

Soient θ_1 et θ_2 deux paramètres réels inconnus et soit:

Y_1 un estimateur sans biais de $\theta_1 + \theta_2$ et de variance σ^2;

Y_2 un estimateur sans biais de $2\theta_1 - \theta_2$ et de variance σ^2;

Y_3 un estimateur sans biais de $6\theta_1 + 3\theta_2$ et de variance $9\sigma^2$,

les estimateurs Y_1, Y_2 et Y_3 étant indépendants. Quels estimateurs de θ_1 et θ_2 proposeriez vous ?

Questions de cours

Une variable Y constituée de n observations Y_i suit un modèle linéaire statistique si on peut écrire que

$$Y = X \cdot \theta + \varepsilon,$$

où X est une matrice à n lignes et k colonnes, $k < n$, θ est un vecteur inconnu constitué de k réels qui sont des paramètres du modèle, et ε est un vecteur aléatoire appelé erreur/résidu du modèle. Les 4 postulats du modèle sont vérifiés. La méthode des moindres carrés ordinaires (MMCO) consiste à déterminer θ minimisant la somme des carrés des résidus (SCR) suivante

$$\mathrm{SCR}(\theta) = \|Y - X \cdot \theta\|^2.$$

1. Montrer que l'estimateur $\hat{\theta}$ de θ par la MMCO vérifie l'égalité suivante

$$\hat{\theta} = (X' \cdot X)^{-1} \cdot X' \cdot Y.$$

2. Montrer que la variable aléatoire $\mathrm{SCR}(\hat{\theta}) = \|Y - X \cdot \hat{\theta}\|^2$ suit une loi du χ^2 à $n - k$ degrés de liberté multipliée par la variance de ε_i.