Đặng Hữu Phước Vinh

1752052

Fin exercice

I/

Mon modèle est sur une liste de nombre de livre pour savoir quand des gens achètent des livres. L'édition peut savoir quand le temp pour peut publier les livres.

Mon modèle a :

num_of_book : Le moyen nombre des livres que cette personne achète dans un mois.

genre : La type du livre que cette personne est la plus aime.

age : L'âge de cette personne.

place : La place où cette personne vit.

num_of_place : C'est le nombre de la place que cette personne vienne pour acheter des livres

form : c'est la forme que cette personne toujours achète des livres (0 est online et 1 et directement à librairie, rue des livres, la fête des livres. . .)

month : C'est le mois que cette personne achète la plus des livres

duration : La moyenne durée que cette personne utilise quand il/elle achète des livres.

time : la temp que cette personne achète des livres

total : C'est la somme des argents pour achète des livres dans une fois

salary : C'est le salaire de cette personne

J'utilise le modèle num_of_book = β0_chapeau + β1_chapeau.age + β2_chapeau.num_of_place + β3_chapeau.form + β4_chapeau.month + β5_chapeau.duration + β6_chapeau.total + β7_chapeau.salary

Pour teste quels éléments effectuent à la nombre des livres que cette personne achète dans un mois pour choisir le bon moment à publier des livres.

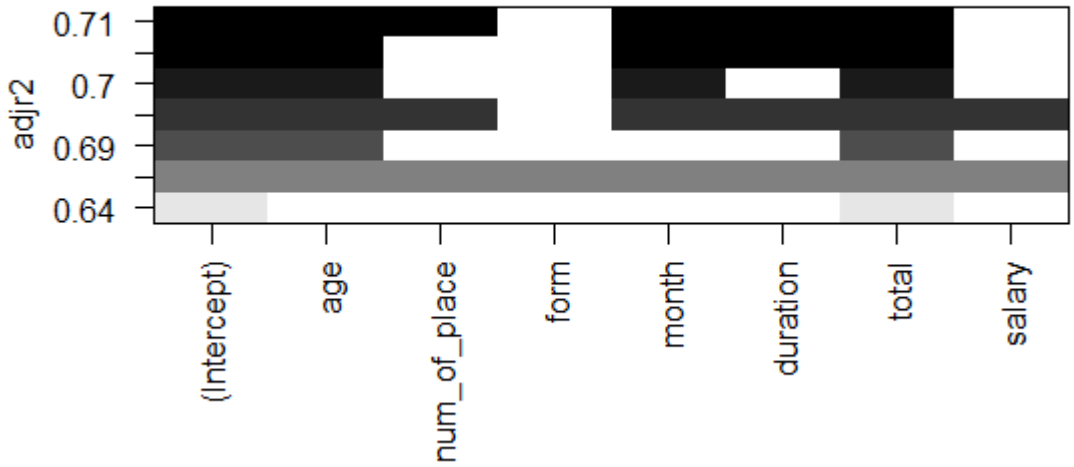
J'utilise la méthode exhaustive avec R carrés ajouté pour gérer mon modèle

```
I_Ex <- read.csv(file.choose())
I_Ex
library(leaps)

recherche.ex<-regsubsets(num_of_book~age+num_of_place+form+month+duration+total+s
                         nvmax=10,method="exhaustive",really.big=T,data=I_Ex)
res_summary <-summary(recherche.ex)

plot(recherche.ex,scale="adjr2")

res_summary$adjr2
```




```
> max.adjr2 <- which.max(res_summary$adjr2)
> max.adjr2
[1] 5
> names(which(res_summary$which[max.adjr2,]==TRUE))
[1] "(Intercept)"  "age"          "num_of_place" "month"        "duration"
[6] "total"
>
```

Le modèle avec age, num_of_place, month, duration et total est le meilleur parce qu'il a le plus grand R carré ajouté

Le modèle : num_of_book = β0_chapeau + β1_chapeau.age + β2_chapeau.num_of_place + + β4_chapeau.month + β5_chapeau.duration + β6_chapeau.total

```
> coef(result1)
  (Intercept)           age  num_of_place         month      duration         total
-4.584588e+00  4.461885e-01  1.005323e+00  6.583592e-01 -2.317636e+00  3.766656e-07
>
```

C'est le coefficient du modèle

Le summary du modèle :

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.585e+00  4.407e+00  -1.040   0.3172
age           4.462e-01  1.793e-01   2.489   0.0272 *
num_of_place  1.005e+00  9.533e-01   1.055   0.3108
month         6.584e-01  4.364e-01   1.509   0.1553
duration     -2.318e+00  1.480e+00  -1.566   0.1413
total         3.767e-07  1.961e-07   1.921   0.0770 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.658 on 13 degrees of freedom
Multiple R-squared:  0.792,     Adjusted R-squared:  0.7121
F-statistic: 9.903 on 5 and 13 DF,  p-value: 0.0004439
```

Le tableau ANOVA :

```
Analysis of Variance Table

Response: num_of_book
             Df Sum Sq Mean Sq F value    Pr(>F)
age           1 541.06  541.06 40.4280 2.503e-05 ***
num_of_place  1  23.32   23.32  1.7426   0.20958
month         1  16.20   16.20  1.2107   0.29114
duration      1  32.70   32.70  2.4430   0.14206
total         1  49.37   49.37  3.6890   0.07698 .
Residuals    13 173.98   13.38
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Alors, on voir la personne qui est plus âgé, des livres et des coûts pour acheter sont le plus. L’édition peut publier des livres sont le type adapté à ces âges. En plus, le temp que des gens achète plusieurs livres est des mois qui sont dans l’été et l’automne. Non seulement ça, les gens qui achètent plusieurs livres toujours aller à plusieurs places pour chercher et voir des livres. Alors, l’édition peut distribuer des livres à autres places.

II/

Ex1/

a/stepwise

Avec par à par ascendant

Étape 1/

C’est des tobs du chaque modèle avec 1 variable mais je juste chois 10 variables et ne chois pas 2 variables « vent » et « pluie » parce que ces variables n’ont pas des numéros.

T9

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.6076    11.4551  -2.235   0.0274 *
T9            6.3130     0.6151  10.263   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196     9.0335  -3.035    0.003 **
T12           5.4687     0.4125  13.258   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ne12

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.0201     4.9807  26.105  < 2e-16 ***
Ne12         -7.9150     0.9042  -8.753 2.77e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ne15

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  118.226      5.424  21.799  < 2e-16 ***
Ne15          -5.781      1.012  -5.712 9.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vx12

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  97.3009     2.7898  34.877  < 2e-16 ***
Vx12          4.3435     0.8675   5.007 2.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

maxO3v

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.50249    6.57153   4.337 3.21e-05 ***
maxO3v       0.68235    0.06929   9.848  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

J’ai choisi la variable « T12 » parce qu’elle a le plus grand tobs.

Puis, j’ajoute chaque variable avec la variable « T12 »

T12 + T9

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.4603     9.9976  -2.847  0.00528 **
T12           5.2757     0.8825   5.978  2.9e-08 ***
T9            0.2830     1.1424   0.248  0.80482
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + Ne12

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7077    15.0884   0.511  0.61050
T12           4.4649     0.5321   8.392 1.92e-13 ***
Ne12         -2.6940     0.9426  -2.858  0.00511 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + Ne15

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.2661    12.1408  -0.681   0.4974
T12           4.9875     0.4552  10.957   <2e-16 ***
Ne15         -1.8207     0.7889  -2.308   0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + Vx12

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4242     9.3943  -1.535  0.12758
T12           5.0202     0.4140  12.125  < 2e-16 ***
Vx12          2.0742     0.5987   3.465  0.00076 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T2 + maxO3v

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.43810    8.00289  -3.678 0.000366 ***
T12           4.07197    0.44195   9.214 2.66e-15 ***
maxO3v        0.35425    0.06318   5.607 1.57e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

J’ai choisi le modèle T2 + maxO3v parce qu’il a le plus grand tobs

Après ça, je faire comme étape 2

T12 + maxO3v + T9

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.85474    8.82497  -2.930  0.00414 **
T12           4.69732    0.78455   5.987 2.84e-08 ***
maxO3v        0.36788    0.06476   5.681 1.15e-07 ***
T9           -0.99557    1.03184  -0.965  0.33678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + maxO3v + Ne12

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.70739   13.17105   0.509  0.61161
T12          3.02540    0.52418   5.772 7.61e-08 ***
maxO3v       0.35757    0.06038   5.922 3.83e-08 ***
Ne12        -2.77351    0.82282  -3.371  0.00104 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + maxO3v + Ne15

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.02913   10.79660  -1.207   0.2302
T12           3.69832    0.46576   7.940 2.04e-12 ***
maxO3v        0.34480    0.06222   5.542 2.14e-07 ***
Ne15         -1.55467    0.70100  -2.218   0.0287 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + maxO3v + Vx12

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.50621    8.30071  -2.109 0.037256 *
T12           3.71410    0.43154   8.607 6.69e-14 ***
maxO3v        0.34124    0.06012   5.676 1.17e-07 ***
Vx12          1.89257    0.52882   3.579 0.000518 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

J’ai choisi le modèle T12 + maxO3v + Vx12 parce qu’il a le plus grand tobs.

Je faire comme étape 2

T12 + maxO3v + Vx12 + T9

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.86214    8.89000  -1.784 0.077214 .
T12           4.05106    0.77045   5.258 7.51e-07 ***
maxO3v        0.34869    0.06195   5.629 1.47e-07 ***
Vx12          1.85375    0.53565   3.461 0.000775 ***
T9           -0.52478    0.99246  -0.529 0.598064
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + maxO3v + Vx12 + Ne12

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.51363   12.96158   0.271   0.7869
T12          3.10885    0.51429   6.045 2.22e-08 ***
maxO3v       0.34701    0.05927   5.855 5.31e-08 ***
Vx12         1.37749    0.57615   2.391   0.0186 *
Ne12        -1.86207    0.89109  -2.090   0.0390 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T12 + maxO3v + Vx12 + Ne15

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.15502   10.46702  -0.970  0.33414
T12           3.55667    0.45212   7.867 3.12e-12 ***
maxO3v        0.33771    0.06011   5.618 1.54e-07 ***
Vx12          1.67366    0.56130   2.982  0.00355 **
Ne15         -0.82725    0.71935  -1.150  0.25271
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

J’ai choisi le modèle T12 + maxO3v + Vx12 + Ne12 parce qu’il a le plus grand tobs

Après, je faire comme l’étape 2

T12 + maxO3v + Vx12 + Ne12 + T9

```
(Intercept)  3.50351   13.02312   0.269  0.78844
T12          3.05772    0.90622   3.374  0.00104 **
maxO3v       0.34607    0.06110   5.664 1.28e-07 ***
Vx12         1.37755    0.57885   2.380  0.01911 *
Ne12        -1.88066    0.93529  -2.011  0.04689 *
T9           0.07023    1.02240   0.069  0.94537
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tobs = 0.069 mais p-value > 0.05 donc T9 n’est pas significative => supprimer

T12 + maxO3v + Vx12 + Ne12 + Ne15

```
(Intercept)  3.5318    13.0555   0.271   0.7873
T12          3.1094     0.5175   6.008 2.68e-08 ***
maxO3v       0.3469     0.0598   5.801 6.90e-08 ***
Vx12         1.3763     0.5823   2.364   0.0199 *
Ne12        -1.8505     1.0721  -1.726   0.0873 .
Ne15        -0.0167     0.8536  -0.020   0.9844
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tosb = -0.020 mais p-value > 0.05 donc Ne15 n’est pas significative => supprimer

Alors, le modèle que j’ai choisi est :

maxO3 = β0_chapeau + β2_chapeau.T12 + β6_chapeau.maxO3v + β5_chapeau.Vx12 + β3_chapeau.Ne12

```
> coef(reg2653)
(Intercept)         T12       maxO3v         Vx12         Ne12
  3.5136259   3.1088538   0.3470085   1.3774933  -1.8620687
>
```

maxO3 = 3.5136259 + 3.1088538*T12 + 0.3470085*maxO3v + 1.3774933*Vx12 - 1.8620687*Ne12

Le résultat quand j'utilise la fonction step() dans R

```
II_Ex1
recherche.ex1_2 <- lm(maxO3 ~ T9 + T12 + Ne12 + Ne15 + Vx12 + maxO3v, data = II_E
```

```
step(recherche.ex1_2)
```

```
Call:
lm(formula = maxO3 ~ T12 + Ne12 + Vx12 + maxO3v, data = II_Ex1)

Coefficients:
(Intercept)          T12         Ne12         Vx12       maxO3v
      3.514        3.109       -1.862        1.377        0.347
```

C'est le même résultat.

b/stagewise

D'abord, je teste quel variable X se corréler avec variable Y

Cor entre maxO3 et maxO3v

```
> cor(II_Ex1$maxO3,II_Ex1$maxO3v)
[1] 0.684516
>
```

Cor entre maxO3 et T9

```
> cor(II_Ex1$maxO3,II_Ex1$T9)
[1] 0.6993865
```

Cor entre maxO3 et T12

```
> cor(II_Ex1$maxO3,II_Ex1$T12)
[1] 0.7842623
>
```

Cor entre maxO3 et Vx12

```
> cor(II_Ex1$maxO3,II_Ex1$Vx12)
[1] 0.4307959
```

Cor entre maxO3 et Ne12

```
> cor(II_Ex1$maxO3,II_Ex1$Ne12)
[1] -0.6407513
```

Cor entre maxO3 et Ne15

```
> cor(II_Ex1$maxO3,II_Ex1$Ne15)
[1] -0.4783021
```

J'ai choisi la variable T12. J'ajoute à le modèle.

Après, je teste chaque variable avec le modèle qui est trouvé.

Cor entre le modèle avec maxO3v

```
> cor(res1,II_Ex1$maxO3v)
[1] 0.3908311
>
```

Cor entre le modèle avec T9

```
> cor(res1,II_Ex1$T9)
[1] 0.01113529
>
```

Cor entre le modèle avec Vx12

```
> cor(res1,II_Ex1$Vx12)
[1] 0.2991681
>
```

Cor entre le modèle avec Ne12

```
> cor(res1,II_Ex1$Ne12)
[1] -0.1983459
>
```

Cor entre le modèle avec Ne15

```
> cor(res1,II_Ex1$Ne15)
[1] -0.1918546
```

J'ai choisi le modèle avec maxO3v

Puis, je faire comme l'étape 2

Cor entre le modèle et T9

```
> cor(res2,II_Ex1$T9)
[1] -0.2351255
>
```

Cor entre le modèle et Vx12

```
> cor(res2,II_Ex1$Vx12)
[1] 0.2300461
>
```

Cor entre le modèle et Ne12

```
> cor(res2,II_Ex1$Ne12)
[1] -0.06181155
>
```

Cor entre le modèle et Ne15

```
> cor(res2,II_Ex1$Ne15)
[1] -0.07745322
```

J'ai choisi le modèle avec T9

Puis, je faire comme l'étape 2

Cor entre le modèle et Vx12

```
> cor(res3,II_Ex1$Vx12)
[1] 0.2904782
>
```

Cor entre le modèle et Ne12

```
> cor(res3,II_Ex1$Ne12)
[1] -0.1778346
```

Cor entre le modèle et Ne15

```
> cor(res3,II_Ex1$Ne15)
[1] -0.1583407
>
```

J'ai choisi le modèle avec Vx12

Puis, je faire comme l'étape 2

Cor entre le modèle et Ne12

```
> cor(res4,II_Ex1$Ne12)
[1] -0.03093136
```

Cor entre le modèle et Ne15

```
> cor(res4,II_Ex1$Ne15)
[1] -0.03437606
>
```

J'ai supprimé tous les deux parce que le coefficient de la corrélation < 0.15 => pas de corrélation.

Alors, le meilleur modèle est :

maxO3 = β0_chapeau + β1_chapeau.T12 + β2_chapeau.maxO3v + β3_chapeau.T9 + β4_chapeau.Vx12

```
(Intercept)         T12       maxO3v          T9        Vx12
-15.8621374   4.0510650   0.3486859  -0.5247776   1.8537461
>
```

maxO3 = -15. 8621374 + 4.0510650*T12 + 0.3486859*maxO3v - 0.5247776*T9 + 1.8537461*Vx12

Ex2/

J'ai utilisé exhaustive méthode avec R adjusté pour choisir la modèle.

J'installer des informations du la fichier tauxaccidents.csv

```r
#Ex2
setwd("C:\Users\Vinh\Desktop\bt\donnees\tauxaccidents.csv")
II_Ex2 <- read.csv(file.choose())
II_Ex2
library(leaps)
#J'ai change des variables du format x_i,numero au format x_inumero parce que
#le R ne reconnais pas le format dernier dans le fichier tauxaccidents.csv
#Exemple: x_i,1 -> x_i1
recherche.ex2<-regsubsets(y_i~x_i1+x_i2+x_i3+x_i4+x_i5+x_i6+x_i7+x_i8+x_i9+x_i10+x_i
                          nvmax=10,method="exhaustive",really.big=T,data=II_Ex2)
res_summary <-summary(recherche.ex2)
plot(recherche.ex2,scale="adjr2")
```

C'est le graphique du modèle.

On peut voir le R carré ajusté de chaque modèle. Mais la modèle qui a le plus grand R carré ajusté est y_i = β0_chapeau + β1_chapeau.x_i1 + β3_chapeau.x_i3 + β4_chapeau.x_i4 + β8_chapeau.x_i8 + β9_chapeau.x_i9 + β12_chapeau.x_i12

```
> res_summary <-summary(recherche.ex2)
> plot(recherche.ex2,scale="adjr2")
> res_summary$adjr2
 [1] 0.5538002 0.6239635 0.6728057 0.6855571 0.7010699 0.7039853 0.6989896 0.6918046
 [9] 0.6826197 0.6722698
> max.adjr2 <- which.max(res_summary$adjr2)
> max.adjr2
[1] 6
> names(which(res_summary$which[max.adjr2,]==TRUE))
[1] "(Intercept)" "x_i1"        "x_i3"        "x_i4"        "x_i8"
[6] "x_i9"        "x_i12"
>
```

C’est le résultat que R compte. R choit 6 variables qui aident le modèle a un sens basé sur R carré ajusté.

```
Call:
lm(formula = y_i ~ x_i1 + x_i4 + x_i8 + x_i9 + x_i12, data = II_Ex2)

Coefficients:
(Intercept)         x_i1         x_i4         x_i8         x_i9        x_i12
   10.10349     -0.06871     -0.11023      0.79395      0.06545     -0.72254

> |
```

Mais quand j’essaye avec la méthode stepwise avec « forward », le meilleur modèle est y_i = β0_chapeau + β1_chapeau.x_i1 + β4_chapeau.x_i4 + β8_chapeau.x_i8 + β9_chapeau.x_i9 + β12_chapeau.x_i12, juste 5 variables (supprimer le x_i3).