Contrôle de Modèle linéaire

STID2 - IUT VANNES

Année 2021/2022

03/01/22

SUJET

- Durée : **1h30** - Calculatrice et feuille de synthèse A4 recto-verso autorisées.
- **Rendre uniquement la feuille de réponses, sans oublier d'indiquer votre nom et prénom.**
- La qualité de la rédaction, les explications fournies et la précision du vocabulaire utilisé seront prises en compte dans la notation.

Exercice 1

Vous êtes consultant pour un client qui souhaite savoir comment accroître les ventes d'un produit. Vous disposez d'une base de données qui fournit les ventes de ce produit dans 200 entreprises différentes ainsi que les budgets alloués par chacune d'elles à la publicité pour ce produit sur les 3 médias publicitaires que sont la télévision, la radio et les journaux. Les variables à votre disposition sont donc les suivantes :

- "Sales" : la quantité de produits vendus (en milliers d'unités) ;
- "TV" : le budget consacré à la publicité à la télévision (en milliers de dollars) ;
- "Radio" : le budget consacré à la publicité à la radio (en milliers de dollars) ;
- "Newspaper" : le budget consacré à la publicité dans les journaux (en milliers de dollars).

L'objectif est d'expliquer/prédire la quantité de produits vendus en fonction de l'investissement publicitaire. On liste ci-dessous les sorties des 7 modèles possibles pouvant être ajustés à partir des 3 variables explicatives à disposition :

- **Modèle 1** :

```
lm(formula = Sales ~ TV, data = data_market)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

- **Modèle 2** :

```
lm(formula = Sales ~ Radio, data = data_market)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7305  -2.1324   0.7707   2.7775   8.1810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
Radio        0.20250    0.02041   9.921   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

- **Modèle 3** :

```
lm(formula = Sales ~ Newspaper, data = data_market)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2272  -3.3873  -0.8392   3.5059  12.7751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

- **Modèle 4** :

```
lm(formula = Sales ~ TV + Radio, data = data_market)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
Radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

- **Modèle 5** :

```
lm(formula = Sales ~ TV + Newspaper, data = data_market)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6231 -1.7346 -0.0948  1.8926  8.4512

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.774948   0.525338  10.993  < 2e-16 ***
TV          0.046901   0.002581  18.173  < 2e-16 ***
Newspaper   0.044219   0.010174   4.346 2.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.121 on 197 degrees of freedom
Multiple R-squared:  0.6458,Adjusted R-squared:  0.6422
F-statistic: 179.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

- **Modèle 6** :

```
lm(formula = Sales ~ Radio + Newspaper, data = data_market)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5289  -2.1449   0.7315   2.7657   7.9751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.188920   0.627672  14.640   <2e-16 ***
Radio       0.199045   0.021870   9.101   <2e-16 ***
Newspaper   0.006644   0.014909   0.446    0.656
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.284 on 197 degrees of freedom
Multiple R-squared:  0.3327,Adjusted R-squared:  0.3259
F-statistic: 49.11 on 2 and 197 DF,  p-value: < 2.2e-16
```

- **Modèle 7** :

```
lm(formula = Sales ~ TV + Radio + Newspaper, data = data_market)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

1. Afin de choisir le meilleur modèle parmi les 7 modèles possibles, on met en place une sélection de modèle avec une recherche forward basée sur le critère du R^2 ajusté.

1.1 Quelle variable ajoute-t-on en premier et pourquoi ?

1.2 Ensuite, quelle(s) variable(s) ajoute-t-on dans l'ordre ?

1.3 Quel est le critère d'arrêt de la procédure ?

1.4 Quel modèle parmi les 7 modèles est finalement retenu ?

2. On met ensuite en place une sélection de modèle avec une recherche de type backward basée sur les tests de Student.

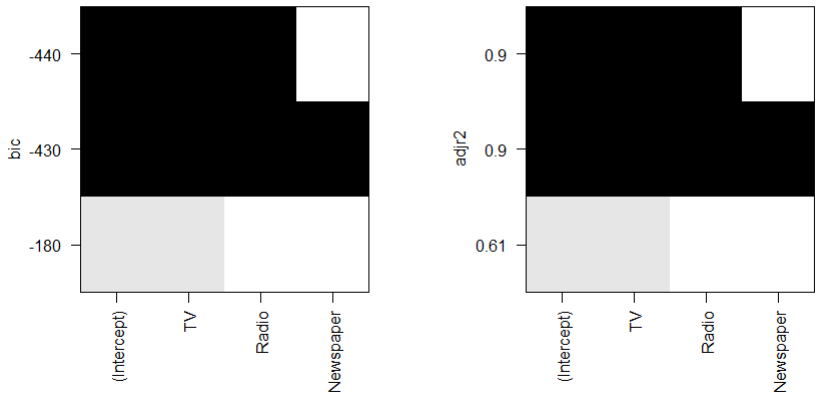
2.1 De quel modèle démarre-t-on ?

2.2 Quelle variable retire-t-on en premier et pourquoi ?

2.3 Quel est le critère d'arrêt de la procédure ?

2.4 Quel modèle parmi les 7 modèles est finalement retenu ?

3. On met enfin en place deux sélections de modèle utilisant une recherche exhaustive : une basée sur le critère du R^2 ajusté, une autre sur le critère BIC. On obtient la sortie graphique suivante avec R :



3.1 On constate que les deux critères conduisent au même choix de modèle : est-ce le cas sur tous les jeux de données ?

3.2 Donner l'équation du modèle ajusté retenu.

3.3 D'après le modèle retenu, combien de milliers de produits peut-on espérer vendre si on ne fait aucun investissement publicitaire ?

4. Est-il possible de mettre en place une procédure de sélection de modèle (forward, backward ou exhaustive) basée sur le critère du R^2 standard ?

5. Le test de Fisher emboité peut être utilisé pour comparer certains modèles entre eux. Combien de tests de Fisher emboités différents peuvent être mis en place à partir de nos variables ?

6. On choisit de mettre en place un test de Fisher emboité pour comparer les modèles 4 et 7.

6.1 Écrire précisément les hypothèses de test H_0 et H_1 utilisées.

6.2 Calculer la valeur prise par la statistique de test F à partir de la sortie ci-dessous :

```
Model 1: Sales ~ TV + Radio
Model 2: Sales ~ TV + Radio + Newspaper
  Res.Df    RSS     Df Sum of Sq   F   Pr(>F)
1    197    556.91
2    196    556.83  1  0.088717    ?    ?
```

6.3 Le quantile de niveau $1-\alpha$ de la loi de Fisher de paramètres $(1, 196)$ vaut 3,89 pour $\alpha = 0,05$. Quelle est la conclusion du test ?

6.4 La p-value du test est-elle inférieure ou supérieure à 5% ?

6.5 Cette p-value apparait également dans une des sorties des 7 modèles : laquelle et pourquoi ?

6.6 Quel modèle préfère-t-on finalement conserver entre les deux ?

7. Dans le modèle 4 expliquant les ventes de produits par les variables "TV" et "Radio" :

7.1 Que vaut la variance estimée $\hat{V}(\hat{\beta}_0)$ de l'estimateur $\hat{\beta}_0$ donnée dans la sortie R ?

7.2 Comment peut-on retrouver ce résultat à partir de la matrice $(X'X)^{-1}$ ci-dessous, où X la matrice des variables explicatives du modèle ?

$$(X'X)^{-1} = \begin{pmatrix} 0.03069062 & -9.550600 \times 10^{-5} & -5.000833 \times 10^{-4} \\ -0.0000955060 & 6.838015 \times 10^{-7} & -2.167244 \times 10^{-7} \\ -0.0005000833 & -2.167244 \times 10^{-7} & 2.286584 \times 10^{-5} \end{pmatrix}$$

7.3 Est-ce que le test de Fisher global donné sur la dernière ligne de la sortie nous permet de confirmer que tous les coefficients du modèle sont significativement différents de 0 ?

Exercice 2

On dispose des variables suivantes, pour 64 stations de ski françaises :

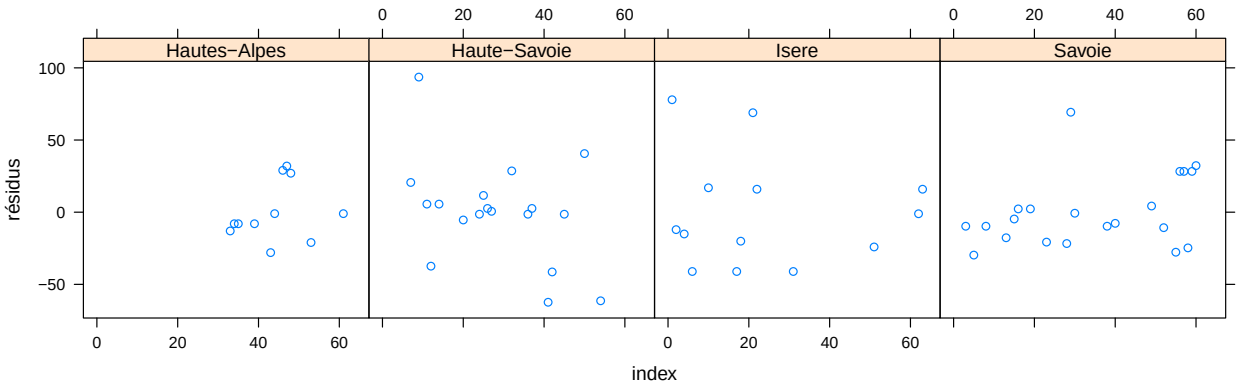
- le prix du forfait semaine ("Forfait") ;
- le département dans lequel se situe la station ("Departement"), ayant 4 modalités :
  — "Hautes-Alpes" ;
  — "Haute-Savoie" ;
  — "Isere" ;
  — "Savoie".

On se demande si le prix du forfait semaine est influencé par le département dans lequel se situe la station. On décide de mettre en place une analyse de la variance et on obtient la sortie suivante :

```
Analysis of Variance Table

Response: Forfait
            Df Sum Sq Mean Sq F value  Pr(>F)
Departement  3  14985  4995.0   ?      0.00285 **
Residuals   60  57350   955.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Donner la valeur de " ?" dans la sortie ci-dessus.
2. Rappeler les hypothèses du test mis en place et donner la conclusion du test.
3. Que peut-on conclure à partir du graphique et de la sortie R ci-dessous ?




```
Bartlett test of homogeneity of variances

data:  residus by Departement
Bartlett's K-squared = 6.6577, df = 3, p-value = 0.08365
```

4. On met en place une analyse de la covariance pour expliquer le prix du Forfait par une variable quantitative : le nombre de remontées dans la station ("NbRemontees") et la variable qualitative "Departement". On obtient l'ajustement ci-dessous avec R :

```
lm(formula = Forfait ~ -1 + Departement + Departement:NbRemontees,
    data = Ski)

Residuals:
    Min      1Q  Median      3Q     Max
-59.268  -9.047   1.797   9.432  94.342

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
DepartementHautes-Alpes             110.24866   18.40730   5.989 1.58e-07 ***
DepartementHaute-Savoie             142.15416    7.15670  19.863  < 2e-16 ***
DepartementIsere                     73.57683   10.88964   6.757 8.73e-09 ***
DepartementSavoie                    87.84187   10.84195   8.102 5.25e-11 ***
DepartementHautes-Alpes:NbRemontees   0.89065    0.35356   2.519  0.01465 *
DepartementHaute-Savoie:NbRemontees   0.09191    0.07085   1.297  0.19987
DepartementIsere:NbRemontees          1.67091    0.31927   5.234 2.58e-06 ***
DepartementSavoie:NbRemontees         1.31497    0.39341   3.342  0.00148 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 23.64 on 56 degrees of freedom
Multiple R-squared:  0.9743,Adjusted R-squared:  0.9706
F-statistic: 265.2 on 8 and 56 DF,  p-value: < 2.2e-16
```

A partir de ce modèle, quel prix peut-on prédire pour un forfait dans une station située en Isère et ayant 30 remontées ?