Régression multiple : exercice

Enoncé

Le propriétaire de la société Showtime Movie Theaters voudrait estimer le chiffre d'affaires hebdomadaire en fonction des dépenses publicitaires. Les données historiques d'un échantillon de huit semaines sont présentées dans le tableau ci-dessous (cf. fichier en ligne Showtime).

Chiffre d'affaires hebdomadaire (milliers de dollars)	Publicité télévisée (milliers de dollars)	Publicité dans les journaux (milliers de dollars)
96	5,0	1,5
90	2,0	2,0
95	4,0	1,5
92	2,5	2,5
95	3,0	3,3
94	3,5	2,3
94	2,5	4,2
94	3,0	2,5

a) Estimer l'équation de la régression en considérant le montant des dépenses publicitaires télévisées comme variable indépendante.

b) Estimer l'équation de la régression en considérant les dépenses publicitaires télévisées et dans les journaux comme variables indépendantes.

c) Est-ce que le coefficient de l'équation estimée de la régression associé aux dépenses publicitaires télévisées est le même dans les questions (a) et (b) ? Interpréter le coefficient dans chaque cas.

d) Quelle est l'estimation du revenu brut d'une semaine lorsque 3 500 dollars sont dépensés en publicité télévisée et 1 800 dollars en publicité dans les journaux.

e. A partir des paramètres SCT et SCR donnés par R, calculez R^2 et R^2 ajusté ;

f. Lorsque, seules les dépenses publicitaires télévisées sont considérées comme variables indépendantes, R^2=0,653 et R^2ajusté=0,595. Les résultats de la régression sont-ils préférables ? Justifiez.

g. Testez les hypothèses suivantes avec $\alpha = 0{,}01$:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : \beta_1 \text{ et/ou } \beta_2 \text{ n'est pas égal à zéro}$$

pour le modèle $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ où x_1 correspond aux dépenses publicitaires télévisées (en milliers de dollars) et x_2 aux dépenses publicitaires dans les journaux (en milliers de dollars).

h. Utilisez $\alpha = 0{,}05$ pour tester la significativité de β_1. Qu'en concluez-vous ?
Idem pour β_2.

Résolution

A

Dans R concernant la variable demandée voici la réponse obtenue :

```
Call:
lm(formula = Chiffre.d.affaires.hebdomadaire..milliers.de.dollars. ~
   Publicité.télévisée..milliers.de.dollars., data = Dataset)

Residuals:
   Min     1Q  Median     3Q    Max
-1.8454 -0.6498 -0.1522  0.7512  1.5507

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    88.6377     1.5824  56.016 2.17e-09 ***
Publicité.télévisée..           1.6039     0.4778   3.357   0.0153 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.215 on 6 degrees of freedom
Multiple R-squared:  0.6526,    Adjusted R-squared:  0.5946
F-statistic: 11.27 on 1 and 6 DF,  p-value: 0.01529
```

Y = 1.639*X+ 88.6377. En regardant nos P-Valeur, on voit que l'bêta0 (88) est largement dessous de 5% donc significatif MAIS le bêta1 (1.6) lui est limite limite. On peut l'interprêter en disant que si on augmente le chiffre de pub à la télé de 1 on augmente le chiffre d'affaire de 1.6. Pour interprêter le bêta0 ; on peut dire que si on a 0% de pub à la télé on aurait 88 6377 dollars.
Pour rappel, le bêta0 est l'ordonné à l'origine. Ayant un schéma de ligne droite, la ligne commence au point (0, 886377).

B

En premier lieu, je vais devoir observer quelle est la colonne à choisir en premier dans mes colonnes explicatives. Pour savoir cela, je regarde celle qui possède le plus grand coefficiant de corrélation par rapport à la variable à expliquer. Dans R -> Statistique > résumé > matrice de corrélation.

Voici la réponse obtenue dans notre cas :

```
                                                Chiffre.d.affaires.hebdomadaire..milliers.de.dollars. Publicité.dans.les.journaux..milliers.de.dollars.
Chiffre.d.affaires.hebdomadaire..milliers.de.dollars.                                      1.00000000                                      -0.02053029
Publicité.dans.les.journaux..milliers.de.dollars.                                         -0.02053029                                       1.00000000
Publicité.télévisée..milliers.de.dollars.                                                  0.80780741                                      -0.55640063
                                                Publicité.télévisée..milliers.de.dollars.
Chiffre.d.affaires.hebdomadaire..milliers.de.dollars.                                 0.8078074
Publicité.dans.les.journaux..milliers.de.dollars.                                    -0.5564006
Publicité.télévisée..milliers.de.dollars.                                             1.0000000
```

On remarque dans la matrice du bas, une corrélation de 0.80…. qui sera la plus grande. On se charge de prendre cette colonne donc en premier !
Pour ce faire j'ai rajouté un A devant le nom de la colonne car R prends par ordre alphabétique.

```
Call:
lm(formula = Chiffre.d.affaires.hebdomadaire..milliers.de.dollars. ~
   APublicité.dans.les.journaux..milliers.de.dollars.
+Publicité.télévisée..milliers.de.dollars.,
   data = Dataset)

Residuals:
    1     2     3     4     5     6     7     8
-0.6325 -0.4124  0.6577 -0.2080  0.6061 -0.2380 -0.4197  0.6469

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      83.2301  1.5739  52.882 4.57e-08 ***
APublicité.dans.les.journaux..    1.3010    0.3207   4.057 0.009761 **
Publicité.télévisée..milliers.de  2.2902    0.3041   7.532 0.000653 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6426 on 5 degrees of freedom
Multiple R-squared:  0.919,      Adjusted R-squared:  0.8866
F-statistic: 28.38 on 2 and 5 DF,  p-value: 0.001865
```

Voici le résultat donné par R pour la régression multilinéaire. Ici on à bêta1 = 1.3010 et bêta2 = 2.2902. Son équation sera : Y = bêta1*X1 + Bêta2*X2 + Bêta0

C

A dépense de publicité tv fixe, si j'augmente de la pub journeaux de 1000 dollars, j'augmente mon chiffre d'affaire de 1301 dollars.

A dépense de pub journeaux fixe, si j'augmente de la pub de la tv de 1000 dollars, j'augmente mon chiffre d'affaire de 22902 dollars.

D Y = 1.3010* 1.8 + 3.5*2.2902 + 83.2301 = 93.5872

E

Pour trouver dans R les SCE et SCR on tape en ligne de commande dans le R partie du dessus, « anova(RegModel.3) « (R^2 = SCR/SCT)

Analysis of Variance Table

Response: Chiffre.d.affaires.hebdomadaire..milliers.de.dollars.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
APublicité.dans.les.journ.	1	0.0107	0.0107	0.026	0.8781448	
Publicité.télévisée..millier	1	23.4247	23.4247	56.730	0.0006532	***
Residuals	5	2.0646	0.4129			

SCR est la somme des carrés de la régression divisé par le total. DF degré de liberté. Pas d'explication. Sum of Square (somme des carrés de la regression). 2.0646 on a la somme des carrés résiduels.

En gros, SCR = 0.0107 + 23.4247 // SCE = 2.0646 // SCT = SCR + SCE.

F

Pour R^2 et R^2Ajusté lequel est meilleur ? Les deux.

Le modèle ici n'est pas robuste car on remarque que la valeur pour la publicité dans les journeaux 0.87 (Pr(>F)) démontre ça n'a pas de sens(Test variance). Du coup on peut dire que le test de regression linéaire est plus correct. Attention vaut mieux comparer le R carré ajusté que le R carré normal. Le r Carré ajusté tiens compte qu'on a notre R qui augmente par nos différentes variables a rajouter. Il faut donc comparer le R carré normal en linéaire et le R carré ajusté en multiple.

G

Notre Test F est non significatif donc ca veut dire que la régression n'est pas significative dans son ensemble. (façon jolie de dire que nos variable n'explique pas significativement notre chiffre d'affaire).

H

Au vue des p-valeur 0.009 et 000.6 on remarque que les tests sont correct séparément.