DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

MTH2302D - PROBABILITÉS ET STATISTIQUE

Devoir - Hiver 2023

Date de remise : 18 avril avant 23h59 (dans Moodle)

Veuillez remplir le tableau suivant et joindre cette page à votre rapport.

|  |  |
| --- | --- |
| **Identification de l’étudiant(e)** | |
| **Nom :** Benzekri | **Prénom :** Omar |
| **Groupe :** 01 | **Matricule :** 2244082 |

|  |  |
| --- | --- |
| Question | Note |
| a) | /4 |
| b) | /7 |
| c) | /12 |
| d) | /5 |
| Présentation | /2 |
| TOTAL | /30 |

Mardi le 18 avril 2023

**Contexte**

Le devoir est une étude de cas qui consiste en une analyse de données (recueillies durant une période de temps) relatives aux ventes de sièges d’automobile (pour enfants) d’un fabricant.

**Les données**

Table

Description automatically generatedLes données à analyser sont constituées d’un échantillon de 195 observations (points de vente) avec quatre variables mesurant un certain nombre de caractéristiques socio-économiques à différents points de vente répartis dans plusieurs villes. Le Tableau 1 ci-dessous présente les variables de l’étude (numéro de colonne dans le fichier, symbole, nom, et description).

**Logiciel**

Pour ce devoir, j’ai utilisé RStudio afin d’étudier et analyser les données.

Text

Description automatically generated**Données**

Sales Price Advertising Region

43 5.05 117 0 1

9 4.97 160 0 1

5 6.42 126 5 1

91 10.71 79 10 0

36 13.39 134 20 1

6 5.61 154 9 0

27 5.56 146 0 0

177 8.19 155 0 0

29 4.90 144 13 0

70 10.77 103 17 0

61 9.43 129 11 0

267 7.57 99 2 1

209 6.98 97 0 0

15 8.68 104 10 0

42 6.85 154 5 1

86 8.68 86 0 0

254 6.37 132 15 1

213 14.90 82 0 0

109 8.78 100 0 1

115 8.77 114 7 0

245 6.03 129 10 1

138 4.90 128 0 0

45 6.43 107 0 1

73 13.28 96 7 1

22 6.71 93 0 1

259 8.33 131 11 0

101 4.68 135 0 0

228 5.31 129 10 1

94 4.42 108 0 1

4 4.34 111 0 0

289 3.02 90 11 0

287 7.54 122 0 1

90 6.52 118 0 1

75 9.58 129 23 1

62 4.83 107 3 1

39 5.30 97 0 0

251 8.73 121 16 1

275 10.51 119 9 0

34 10.96 86 11 1

283 8.85 91 0 1

270 11.07 96 11 0

107 6.38 128 5 1

225 11.67 87 10 1

178 2.34 144 15 1

83 8.41 77 13 1

110 6.50 150 16 0

155 11.85 120 15 1

89 12.11 104 18 0

196 11.17 118 11 1

199 8.79 101 13 0

114 7.40 97 4 1

226 9.01 100 9 0

163 6.92 119 13 1

238 4.36 123 2 0

125 8.23 139 5 1

112 7.36 133 0 1

103 6.20 118 0 1

51 6.56 111 7 1

253 7.81 118 13 0

166 4.43 145 1 1

136 6.67 125 5 1

74 11.62 139 4 1

81 7.38 93 0 0

162 6.62 151 12 1

44 9.34 49 0 0

242 7.23 128 18 1

239 7.30 117 0 1

171 9.31 106 9 1

229 6.63 108 0 1

50 3.63 149 0 1

23 4.47 147 7 0

197 6.97 129 19 0

188 11.48 77 15 1

185 8.67 112 14 1

191 7.52 128 0 1

160 5.53 132 8 1

285 8.86 104 0 1

19 1.82 133 0 1

55 5.73 144 0 1

14 4.95 110 5 1

161 5.36 101 0 1

35 10.61 149 0 1

100 2.07 126 0 0

256 5.83 112 7 0

184 6.50 94 3 1

48 4.62 138 0 1

142 9.14 90 0 1

132 6.52 116 3 1

255 6.81 125 0 0

202 9.48 132 10 0

216 2.93 160 5 0

63 9.32 70 0 0

195 8.77 128 13 1

220 8.71 144 5 0

211 13.14 105 10 1

37 9.03 110 13 1

221 8.74 124 0 1

154 5.71 118 4 1

57 6.01 127 11 1

58 4.56 135 0 1

108 7.77 115 6 1

66 12.30 94 10 0

149 6.89 110 10 0

261 8.55 92 23 0

274 8.31 117 0 1

104 7.71 69 0 1

282 11.48 87 13 1

21 9.50 120 11 1

186 12.29 131 13 1

33 6.54 124 0 0

214 5.52 116 0 1

18 5.32 102 6 1

47 7.22 151 2 0

212 4.69 124 0 0

181 12.61 104 10 0

52 7.37 128 8 1

134 8.39 84 5 1

12 8.98 90 0 0

257 10.07 107 11 1

159 10.43 24 0 1

241 8.80 119 0 1

279 5.36 117 0 0

122 4.10 133 6 1

93 0.53 159 7 1

82 4.88 107 3 0

268 3.24 138 0 0

262 12.49 127 24 0

69 9.45 92 12 0

46 8.97 125 0 0

31 6.71 137 17 1

167 6.88 108 5 1

247 10.31 121 0 1

10 5.16 114 0 0

206 5.42 103 15 1

128 10.00 88 0 0

233 6.90 90 20 1

88 11.70 126 7 0

269 4.53 125 0 1

219 8.47 101 10 1

120 6.59 102 0 1

98 8.77 117 11 1

56 7.96 124 0 1

130 4.38 108 0 1

133 4.20 144 0 1

164 4.74 140 4 1

222 5.07 96 0 1

92 7.95 119 3 0

105 2.05 157 0 1

77 11.19 105 7 0

193 3.67 131 0 1

187 8.09 122 0 0

157 8.67 115 14 0

263 7.45 129 5 1

117 5.12 100 10 0

227 8.22 141 0 0

249 6.41 131 2 1

1 5.40 163 13 0

59 10.08 130 10 0

217 7.80 104 12 0

96 7.78 116 3 1

265 7.56 93 0 0

78 7.91 129 3 1

240 0.37 191 7 1

124 9.01 115 14 1

3 4.21 137 14 0

232 7.80 98 0 0

243 12.13 109 12 0

169 7.49 157 0 1

203 10.27 109 12 1

237 3.47 81 0 0

230 5.55 97 8 1

139 1.00 185 0 0

198 8.01 118 12 1

147 12.44 70 14 0

65 6.41 136 0 0

258 6.88 112 0 0

246 12.49 55 12 1

53 6.67 173 13 1

119 9.39 120 14 1

72 9.32 108 16 1

150 4.67 111 0 0

85 3.90 131 0 1

244 7.78 64 0 1

205 4.15 128 3 1

201 10.49 114 8 0

192 10.62 116 19 1

156 4.42 94 7 1

170 8.65 120 18 0

146 11.91 84 0 1

7 3.47 103 2 1

135 13.91 68 0 0

280 7.53 113 11 0

41 7.81 102 15 1

174 6.20 137 12 0

40 7.70 89 12 0

On initialise les variables:

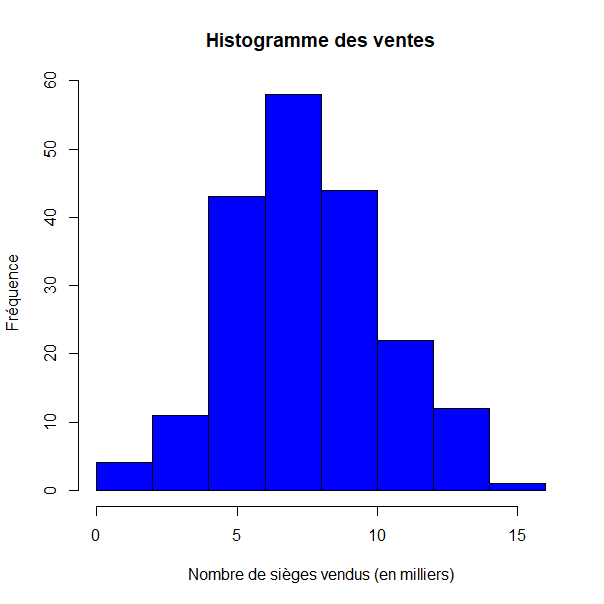
Text

Description automatically generated

**Phase I : Analyse statistique descriptive et inférence.**

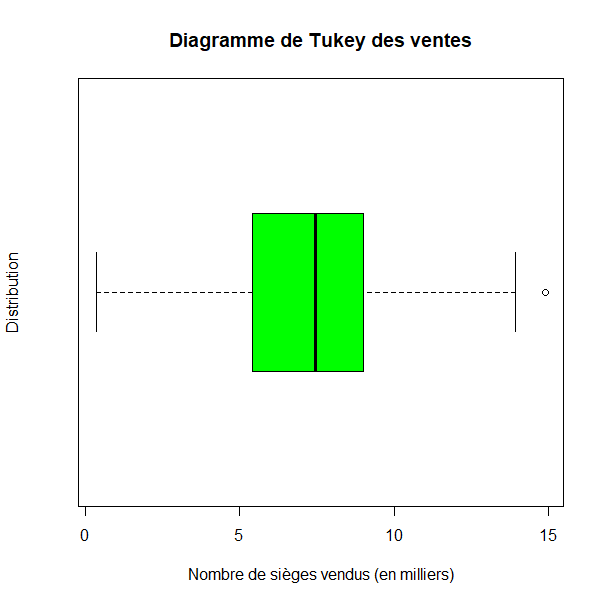
**Text

Description automatically generateda)**

****

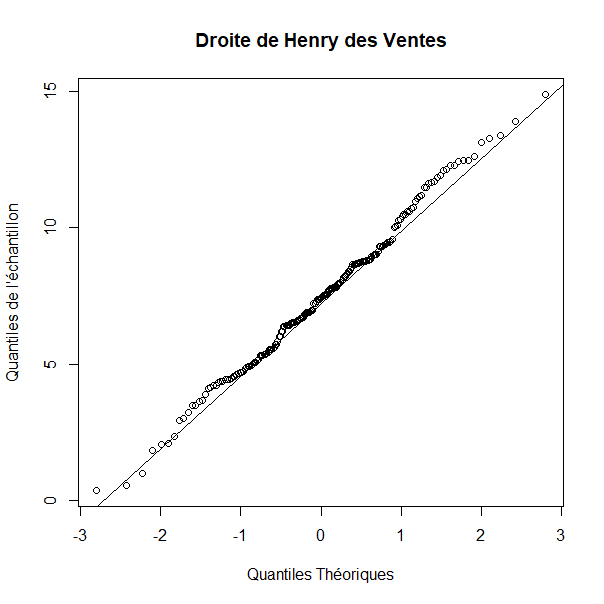
**Figure 1.** Histogramme des ventes

Dans cet histogramme, on peut observer la plage des ventes, qui varie de 0 à environ 15000. La hauteur des barres de l'histogramme représente l'effectif, ce qui permet de constater que l'effectif le plus fréquent se situe entre 6000 et 8000 ventes, avec environ 50 occurrences. On peut donc supposer que la médiane se situe également entre ces deux valeurs. Par ailleurs, on note un faible effectif entre 14000 et 15000 ventes. La forme de l'histogramme ressemble à celle de la distribution normale, ce qui permettrait d'envisager cette hypothèse. Toutefois, il n'est pas possible de confirmer cette hypothèse car la symétrie par rapport à la moyenne ne peut pas être mesurée à partir de cet histogramme.



**Figure 2.** Boîte de Tukey pour les ventes

Le diagramme de Tukey des ventes ci-dessus permet de relever plusieurs données intéressantes lors de l'analyse statistique descriptive des ventes. Tout d'abord, ce diagramme représente les quartiles de la variable. Dans ce cas, le premier quartile se situe autour de 5500 ventes, la médiane se situe autour de 7500 ventes et le troisième quartile se situe autour de 9000 ventes. Par ailleurs, il est possible d'estimer les valeurs extrêmes à partir de ce diagramme en boîte à moustaches : la valeur minimale est de 400 ventes et la valeur maximale est de 13500 ventes. Ainsi, on peut conclure que la moitié des données se situent entre 5500 ventes et 9000 ventes car la boîte en mauve représente l'écart interquartile contenant 50% des observations par définition. En outre, on peut remarquer que la médiane est plus proche du deuxième quartile que du troisième.

****

**Figure 3.** Droite de Henry pour les ventes

Sur la droite de Henry des ventes représentée ci-dessus, les données de ventes sont représentées par des petits cercles et la droite représente la droite de Henry qui suit une probabilité normale. Étant donné que la majorité des observations se situent sur cette droite, on peut supposer que les ventes suivent une loi normale. Seules les observations situées aux extrémités sont nettement en dehors du tracé de la droite normale. Pour vérifier cette hypothèse de normalité, il est nécessaire d'effectuer un test de normalité Shapiro-Wilk. Ce test permet d'obtenir le carré d'un coefficient de corrélation W et la valeur P, deux indicateurs qui sont nécessaires pour confirmer l'hypothèse. L'hypothèse nulle (H0) stipule que Y (Ventes) suit une loi normale, tandis que l'hypothèse alternative (H1) soutient que Y (Ventes) ne suit pas une loi normale.

Shapiro-Wilk normality test

data: sales

W = 0.99348, p-value = 0.5458

**Figure 4.** Test de normalité (Shapiro-Wilk)

On a obtenu W=0.99348 et p-value=0.5458. Premièrement, on est poussé à accepter H0, car notre valeur de W n’est pas proche de 0.70 (limite inférieure de cette statistique) et est en fait très proche de 1; la limite supérieure de cette statistique. On doit rejeter H0 lorsque la valeur de W est petite et proche de 0.70. De plus, H0 est acceptée lorsque la P-value, limitée par 0 et 1, est grande. Ayant obtenu une p-value de 0.9872 très grande et proche de sa limite supérieure, on peut donc accepter H0. Puisque cette hypothèse nulle est acceptée dans les deux cas, on peut donc dire que la variable Y représentant les ventes suit effectivement une loi normale.

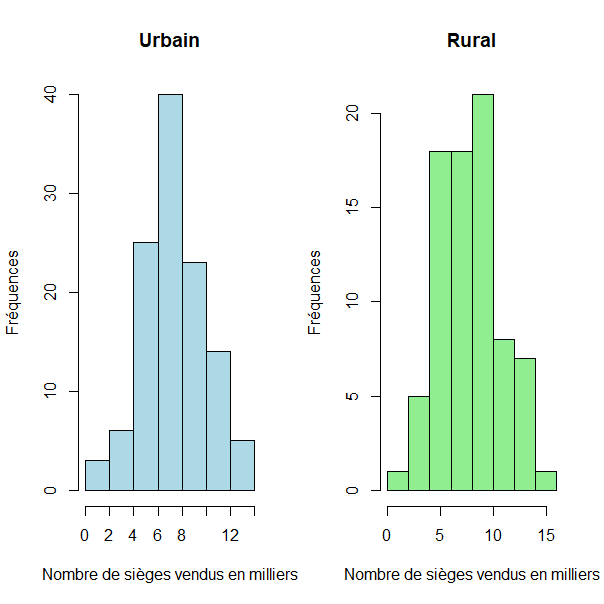
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Moyenne | Q1 | Q2 (Mediane) | Q3 | Écart type | Intervalle de confiance |
| 7.467 | 5.410 | 7.450 | 9.010 | 2.743 | [7.079,7.854] |

**Figure 5.** Tableau de statistiques descriptives

Le tableau de statistiques descriptives présenté ci-dessus présente la moyenne, les quantiles et l'intervalle de confiance des ventes.

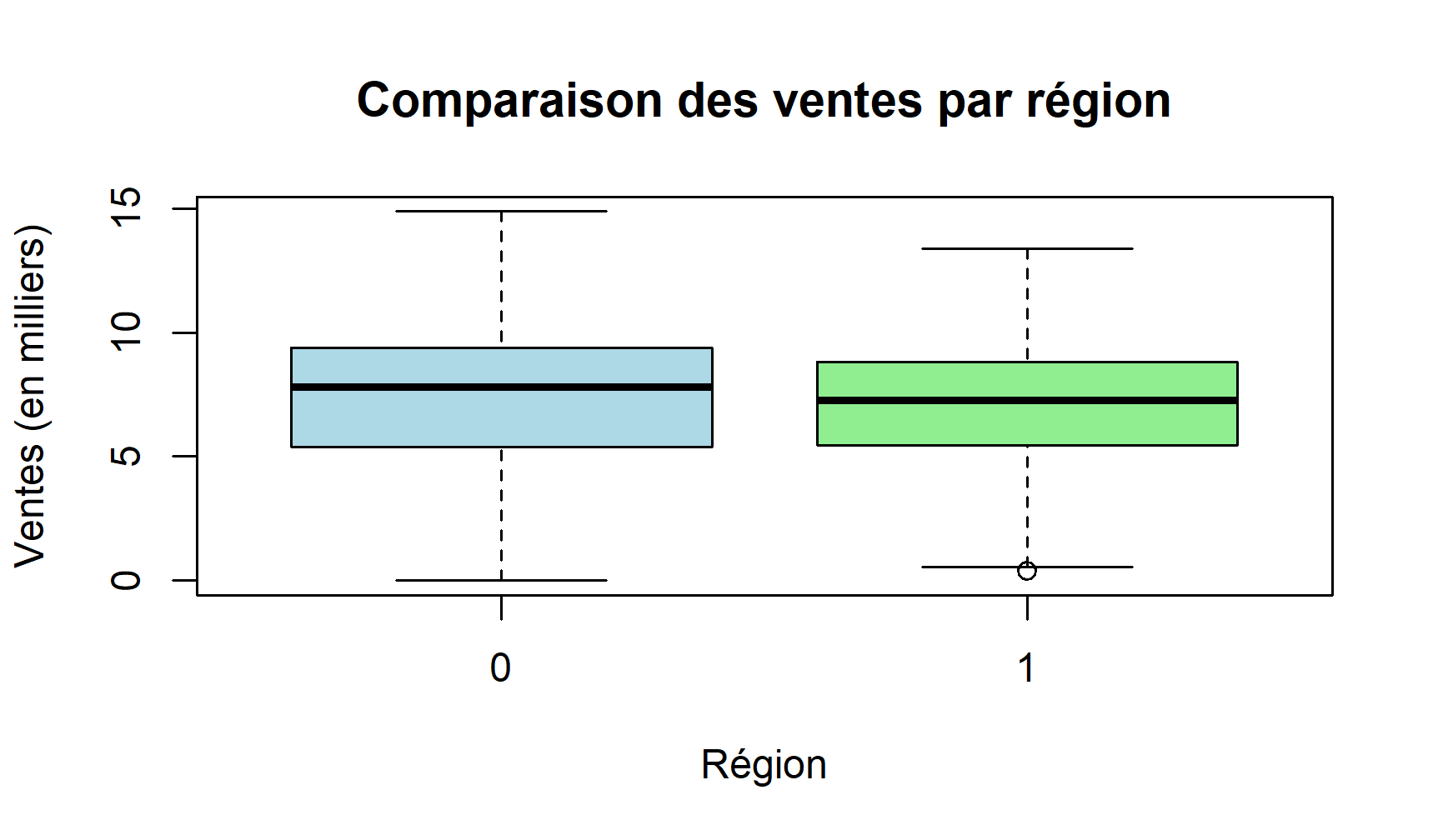
**Text

Description automatically generatedb)**

****

**Figure 6.** Histogramme des ventes pour les points de vente urbains et ruraux

L'histogramme correspondant à la région 0 représente les ventes effectuées dans les points de vente situés en milieu rural, tandis que l'autre représente celles en milieu urbain. Premièrement, il est à noter que les ventes de la région 1 semblent davantage suivre la distribution en cloche de la loi normale que celles de la région 0. Deuxièmement, la classe avec l'effectif le plus important diffère entre les deux groupes. Dans la région 0, il s'agit de la classe située entre 8000 et 10000 ventes, tandis que dans la région 1, il s'agit de la classe située entre 6000 et 8000 ventes.

****

**Figure 7.** Boîte de Tukey pour les ventes par région

On peut observer une disparité de l’étendue des données par région. En milieu rural, les données s’étendent de 2000 ventes à proche de 15000 ventes, tandis qu’en milieu urbain, les observations s’étendent de 1000 ventes à 13500 ventes. Une certaine disparité est aussi présente dans l’écart interquartile, celui de la région 0 est plus grand que dans la région 1. De plus, la médiane des ventes dans la région rurale est autour de 7750 ventes, tandis qu’en région urbaine, la médiane semble être plus proche de 7250 ventes.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Moyenne | Q1 | Q2 | Q3 | Écart type | Variance | Intervalle |
| Urbain | 7.295 | 5.495 | 7.265 | 8.812 | 2.656 | 7.052 | [6.806,7.783] |
| Rural | 7.707 | 5.380 | 7.800 | 9.385 | 2.898 | 8.398 | [7.057,8.355] |

**Figure 8.** Tableau de statistiques descriptives par groupe

Le tableau de statistiques descriptives présenté ci-dessus présente la moyenne, les quantiles et les intervalles de confiance des ventes dans les région urbaine et rurale.

F test to compare two variances

data: sales[region == 0] and sales[region == 1]

F = 1.1647, num df = 78, denom df = 115, p-value = 0.4535

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.7799153 1.7677926

sample estimates:

ratio of variances

1.164701

**Figure 9.** Test d’hypothèses sur l’égalité des variances pour les deux groupes

L'hypothèse nulle est formulée en termes d'égalité des variances (H0 : σ02 = σ12) tandis que l'hypothèse alternative suppose leur non-égalité (H1 : σ02 ≠ σ12). Pour effectuer ce test, il est nécessaire de connaître la valeur de F0 et celle de la loi de Fisher F(α/2, n0-1, n1-1). Si F0 > F (α/2, n0-1, n1-1), alors l'hypothèse nulle est rejetée. Dans ce cas-ci, on obtient F0 = 1.1647 et F = 5.102369, ce qui nous permet d'accepter l'hypothèse nulle et de confirmer l'égalité des variances entre les deux groupes. La p-value étant supérieure à 0,05, on peut en conclure que les variances ne présentent pas de différence significative entre les deux groupes.

Welch Two Sample t-test

data: sales[region == 0] and sales[region == 1]

t = 1.0453, df = 159.01, p-value = 0.2975

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.3773656 1.2260191

sample estimates:

mean of x mean of y

7.719241 7.294914

**Figure 10.** Test d’hypothèses sur l’égalité des moyennes pour les deux groupes

On pose l’hypothèse nulle comme l’égalité des moyennes pour les deux groupes (H0 : μ0=μ1) et l’hypothèse alternative comme la non-égalité des moyennes pour les deux groupes (H1 : μ0≠μ1). Pour ce test, on a besoin de la valeur de t0 et la valeur de la loi de Student t(α/2, v). On rejette l’hypothèse nulle si abs(t0) > t. Calculée ci-dessus, on obtient t0 = 1.0453, t=1.974438 et v=df=159.01, on peut donc accepter l’hypothèse nulle dans ce cas-ci et confirmer l’hypothèse de l’égalité des moyennes pour les deux groupes.

**Phase II : Recherche du meilleur modèle.**

**c)**

Text

Description automatically generatedModèle 1 : Y = β0 + β1X1 + ε :

Call:

lm(formula = sales ~ price)

Residuals:

Min 1Q Median 3Q Max

-5.8192 -1.8188 -0.1694 1.5349 6.8463

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.48516 0.87060 15.489 < 2e-16 \*\*\*

price -0.05180 0.00734 -7.058 2.96e-11 \*\*\*

---

Residual standard error: 2.452 on 193 degrees of freedom

Multiple R-squared: 0.2052, Adjusted R-squared: 0.201

F-statistic: 49.81 on 1 and 193 DF, p-value: 2.963e-11

**Figure 11.** Tableau des coefficients de régression du modèle 1.

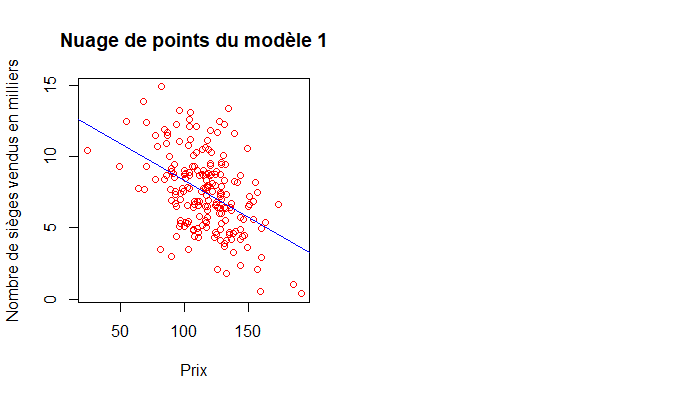
Response: sales

Df Sum Sq Mean Sq F value Pr(>F)

price 1 299.56 299.562 49.814 2.963e-11 \*\*\*

Residuals 193 1160.63 6.014

**Figure 12.** Tableau d’analyse de la variance du modèle 1.



**Figure 13.** Test de signification du modèle 1.

Chart, scatter chart

Description automatically generated**Figure 14.** Analyse des résidus du modèle 1.

Pour évaluer la normalité des résidus, nous devons nous fier au « Normal Probability Plot » représenté ci-dessus. On peut observer que les résidus en mauve suivent une ligne droite et se situe majoritairement sur la ligne de normalité en rouge indiquant que l’hypothèse de normalité est respectée. De plus, les résidus sont dispersés de façon homogène autour de la droite horizontale dans le graphique « Residuals vs Fitted » indiquant l’homoscédasticité. Cependant, dans ce même graphique, les résidus se situant en dehors de l’intervalle -2 à 2 doivent être rejetées, car ce sont des données atypiques.

2.5 % 97.5 %

(Intercept) 11.7680393 15.20228356

price -0.0662782 -0.03732599

**Figure 15.** Intervalle de confiance pour les paramètres du modèle 1.

Calculée ci-dessus, avec un niveau de confiance 95%, le paramètre β0 est compris entre 11.7680393 et 15.20228356 et le paramètre β1 est compris entre -0.0662782 et -0.03732599.

Modèle 2 : Y = β0\*X1^(β1)\*e^(ε) :

Graphical user interface, text, application

Description automatically generatedÉquation transformée : ln(Y) = ln(β0) + β1ln(X1) + ε :

Call:

lm(formula = log(sales) ~ log(price))

Residuals:

Min 1Q Median 3Q Max

-2.46161 -0.21102 0.06951 0.27545 0.82030

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.0144 0.6445 9.332 < 2e-16 \*\*\*

log(price) -0.8657 0.1361 -6.362 1.41e-09 \*\*\*

---

Residual standard error: 0.4495 on 193 degrees of freedom

Multiple R-squared: 0.1734, Adjusted R-squared: 0.1691

F-statistic: 40.48 on 1 and 193 DF, p-value: 1.411e-09

**Figure 16.** Tableau des coefficients de régression du modèle 2.

Analysis of Variance Table

Response: log(sales)

Df Sum Sq Mean Sq F value

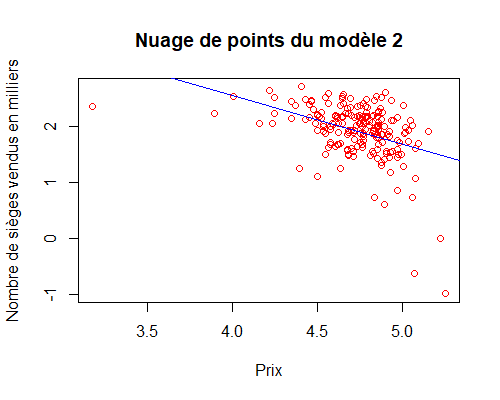
log(price) 1 8.180 8.1802 40.479

Residuals 193 39.003 0.2021

Pr(>F)

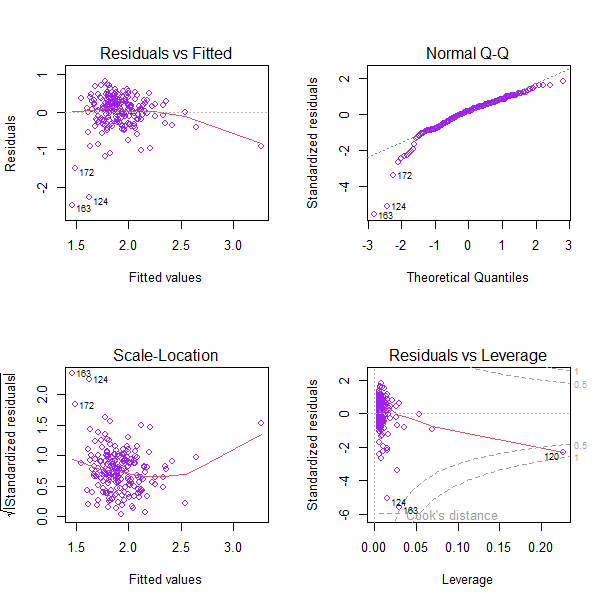
log(price) 1.411e-09 \*\*\*

Residuals **Figure 17.** Tableau d’analyse de la variance du modèle 2.



**Figure 18.** Test de signification du modèle 2.

Les tableaux ci-dessous me permettent de faire le test de signification des coefficients. Pour ces tests, le critère de rejet de l’hypothèse est que la p-value soit plus petite que α. On pose comme hypothèse nulle que les coefficients soient égaux à zéro (H0 : β0 = 0 β1= 0) ainsi que l’hypothèse alternative que les coefficients ne soient pas égaux à zéro (H1 : β0 ≠ 0 β1≠0). Dans ce modèle, α=0.05, la p-value (β0) = 2e-16 et la p-value (β1) = 1.41e-9. Ces valeurs sont significativement plus petites que 0.05, on peut donc rejeter l’hypothèse nulle dans les deux tests. On peut donc conclure que la relation linéaire entre le prix (X1) et les ventes (Y) est significative. De plus, la valeur de R2 trouvée dans le tableau d’analyse de la variance est de 0.1361, étant loin de 1 ceci indique que le modèle n’est pas une bonne représentation de la relation entre les ventes et le prix. Aussi, sur le nuage de points ci-dessus on peut observer que les observations suivent majoritairement la ligne bleue représentant la relation entre le prix et les ventes, ce qui confirme le rejet de l’hypothèse nulle de β1.



**Figure 19.** Analyse des résidus du modèle 2.

Pour vérifier la normalité des résidus, il convient de se référer au « Normal Probability Plot » représenté ci-dessus. On peut constater que près de 90% des résidus en mauve suivent une ligne droite et se situent en grande partie sur la ligne de normalité en rouge, ce qui indique que l'hypothèse de normalité est respectée pour la plupart des observations. De plus, les résidus sont distribués de manière homogène autour de la droite horizontale dans le graphique « Residuals vs Fitted », ce qui témoigne de l'homoscédasticité. Toutefois, les résidus situés en dehors de l'intervalle -2 à 2 doivent être considérés comme des données atypiques et rejetés.

2.5 % 97.5 %

(Intercept) 4.743289 7.285531

log(price) -1.134109 -0.597350

**Figure 20.** Intervalle de confiance pour les paramètres du modèle 1.

Calculée ci-dessus, avec un niveau de confiance 95%, le paramètre β0 est compris entre 4.743289 et 7.285531 et le paramètre β1 est compris entre 1.134109 et -0.597350.

Modèle 3 : Y = β0\*e^(β1\*X1+ε) :

Graphical user interface, text, application

Description automatically generatedÉquation transformée : ln(Y) = ln(β0) + β1\*X1 + ε :

Call:

lm(formula = log(sales) ~ price)

Residuals:

Min 1Q Median 3Q Max

-2.17532 -0.21112 0.07678 0.26388 0.85112

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.065320 0.154007 19.904 < 2e-16 \*\*\*

price -0.009865 0.001298 -7.598 1.27e-12 \*\*\*

---

Residual standard error: 0.4338 on 193 degrees of freedom

Multiple R-squared: 0.2303, Adjusted R-squared: 0.2263

F-statistic: 57.73 on 1 and 193 DF, p-value: 1.267e-12

**Figure 21.** Tableau des coefficients de régression du modèle 3.

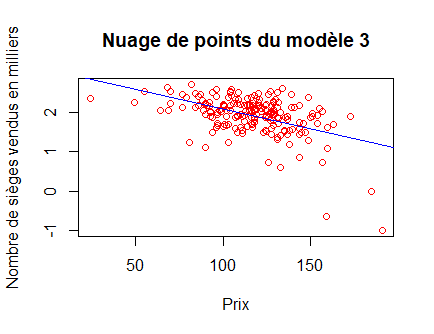
Analysis of Variance Table

Response: log(sales)

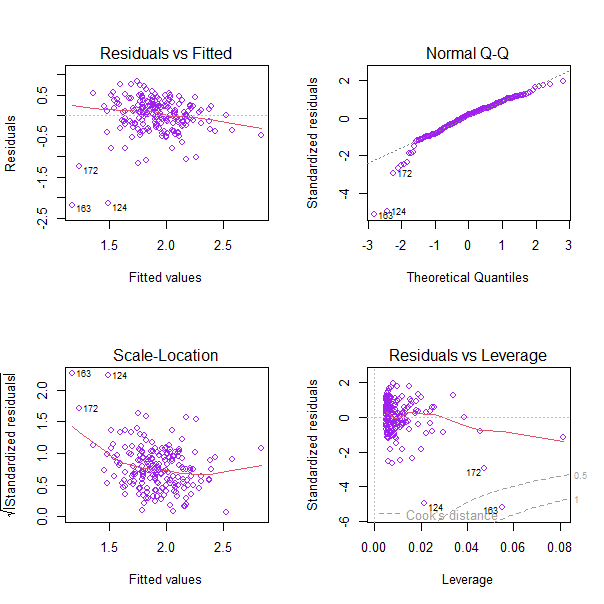
Df Sum Sq Mean Sq F value Pr(>F)

price 1 10.864 10.8643 57.733 1.267e-12 \*\*\*

Residuals 193 36.319 0.1882 **Figure 22.** Tableau d’analyse de la variance du modèle 3.



**Figure 23.** Test de signification du modèle 3.

Les tableaux ci-dessous me permettent de faire le test de signification des coefficients. Pour ces tests, le critère de rejet de l’hypothèse est que la p-value soit plus petite que α. On pose comme hypothèse nulle que les coefficients soient égaux à zéro (H0 : β0 = 0 β1= 0) ainsi que l’hypothèse alternative que les coefficients ne soient pas égaux à zéro (H1 : β0 ≠ 0 β1≠0). Dans ce modèle, α=0.05, la p-value (β0) < 2e-16 et la p-value (β1) = 1.27e-12. Ces valeurs sont significativement plus petites que 0.05, on peut donc rejeter l’hypothèse nulle dans les deux tests. On peut donc conclure que la relation linéaire entre le prix (X1) et les ventes (Y) est significative. De plus, la valeur de R2 trouvée dans le tableau d’analyse de la variance est de 0.23, étant loin de 1 ceci indique que le modèle n’est pas une bonne représentation de la relation entre les ventes et le prix. Aussi, sur le nuage de points ci-dessus on peut observer que les observations suivent majoritairement la ligne bleue représentant la relation entre le prix et les ventes, ce qui confirme le rejet de l’hypothèse nulle de β1.

**Figure 24.** Analyse des résidus du modèle 3.

Pour évaluer la normalité des résidus, nous devons nous fier au « Normal Probability Plot » représenté ci-dessus. On peut observer qu’environ 85% des résidus en mauve suivent une ligne droite et se situent en bonne partie sur la ligne de normalité en rouge indiquant que l’hypothèse de normalité est respectée pour la plupart des observations. De plus, les résidus sont dispersés de façon homogène autour de la droite horizontal

dans le graphique « Residuals vs Fitted » indiquant l’homoscédasticité. Cependant, dans ce même graphique, les résidus se situant en dehors de l’intervalle -2 à 2 doivent être rejetées, car ce sont des données

atypiques.

2.5 % 97.5 %

(Intercept) 2.76156702 3.369073355

price -0.01242595 -0.007304401

**Figure 25.** Intervalle de confiance pour les paramètres du modèle 3.

Calculée ci-dessus, avec un niveau de confiance 95%, le paramètre β0 est compris entre 2.76156702 et 3.369073355 et le paramètre β1 est compris entre -0.01242595 et -0.007304401.

Graphical user interface, text, application

Description automatically generatedModèle 4 : Y = β0+β1\*X2+ε :

Call:

lm(formula = sales ~ advertising)

Residuals:

Min 1Q Median 3Q Max

-7.2065 -1.9808 0.0461 1.6108 8.3708

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.52919 0.26084 25.032 < 2e-16 \*\*\*

advertising 0.14962 0.02936 5.097 8.2e-07 \*\*\*

---

Residual standard error: 2.582 on 193 degrees of freedom

Multiple R-squared: 0.1186, Adjusted R-squared: 0.1141

F-statistic: 25.98 on 1 and 193 DF, p-value: 8.198e-07

**Figure 26.** Tableau des coefficients de régression du modèle 4.

Analysis of Variance Table

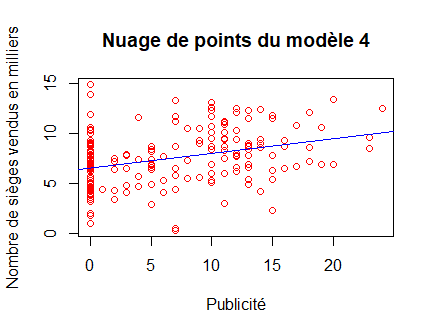
Response: sales

Df Sum Sq Mean Sq F value Pr(>F)

advertising 1 173.23 173.232 25.979 8.198e-07 \*\*\*

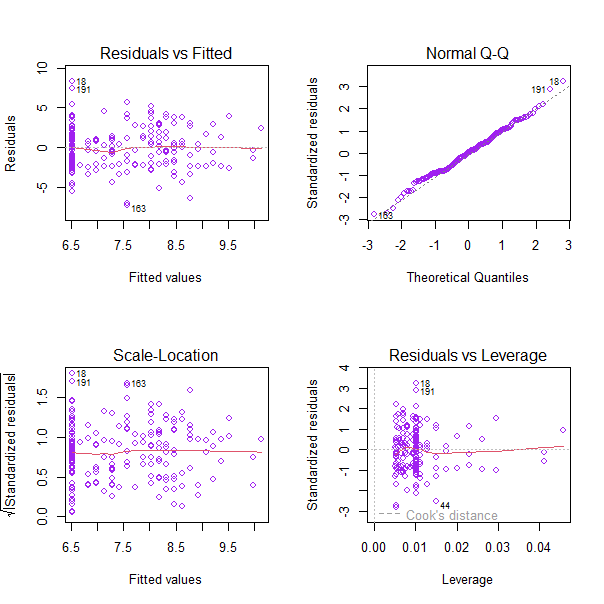
Residuals 193 1286.96 6.668

**Figure 27.** Tableau d’analyse de la variance du modèle 4.



**Figure 28.** Test de signification du modèle 4.

Pour ces tests, le critère de rejet de l’hypothèse est que la p-value soit plus petite que α. On pose comme hypothèse nulle que les coefficients soient égaux à zéro (H0 : β0 = 0 β1=0) ainsi que l’hypothèse alternative que les coefficients ne soient pas égaux à zéro (H1 : β0 ≠ 0 β1≠0). Dans ce modèle, α=0.05, la p-value (β0) < 2e-16 et la p-value (β1) = 8.2e-07. Ces valeurs sont significativement plus petites que 0.05, on peut donc rejeter l’hypothèse nulle dans les deux tests. On peut donc conclure que la relation linéaire entre la publicité (X2) et les ventes (Y) est significative. De plus, la valeur de R2 trouvée dans le tableau d’analyse de la variance est de 0.1186, étant loin de 1 ceci indique que le modèle n’est pas une bonne représentation de la relation entre les ventes et la publicité. Aussi, sur le nuage de points ci-dessus on peut observer que les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre la publicité et les ventes, ce qui confirme le rejet de l’hypothèse nulle de β1.

**Figure 29.** Analyse des résidus du modèle 4.

Pour évaluer la normalité des résidus, nous devons nous fier au « Normal Probability Plot » représenté ci-dessus. On peut observer que les résidus en mauve suivent majoritairement une ligne droite et se situent en bonne partie sur la ligne de normalité en rouge indiquant que l’hypothèse de normalité est respectée. De plus, les résidus sont dispersés de façon homogène autour de la droite horizontale dans le graphique « Residuals vs Fitted » indiquant l’homoscédasticité. Cependant, dans ce même graphique, les résidus se situant en dehors de l’intervalle -2 à 2 doivent être rejetées, car ce sont des données atypiques.

2.5 % 97.5 %

(Intercept) 6.01472880 7.0436493

advertising 0.09172383 0.2075203

**Figure 30.** Intervalle de confiance pour les paramètres du modèle 4.

Calculée ci-dessus, avec un niveau de confiance 95%, le paramètre β0 est compris entre 6.01472880 et 7.0436493 et le paramètre β1 est compris entre 0.09172383 et 0.2075203.

Modèle 5 : Y = β0\* (8+X2)^(β1)\*e^(ε) :

Graphical user interface, text, application, email

Description automatically generatedÉquation transformée : ln(Y) = ln(β0) + β1\*ln(8+X2) + ε :

Call:

lm(formula = log(sales) ~ log(8 + advertising))

Residuals:

Min 1Q Median 3Q Max

-2.95670 -0.22109 0.08772 0.27969 0.92424

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.16405 0.20211 5.760 3.28e-08 \*\*\*

log(8 + advertising) 0.29483 0.07777 3.791 0.000201 \*\*\*

---

Residual standard error: 0.477 on 193 degrees of freedom

Multiple R-squared: 0.0693, Adjusted R-squared: 0.06448

F-statistic: 14.37 on 1 and 193 DF, p-value: 0.0002006

**Figure 31.** Tableau des coefficients de régression du modèle 5.

Analysis of Variance Table

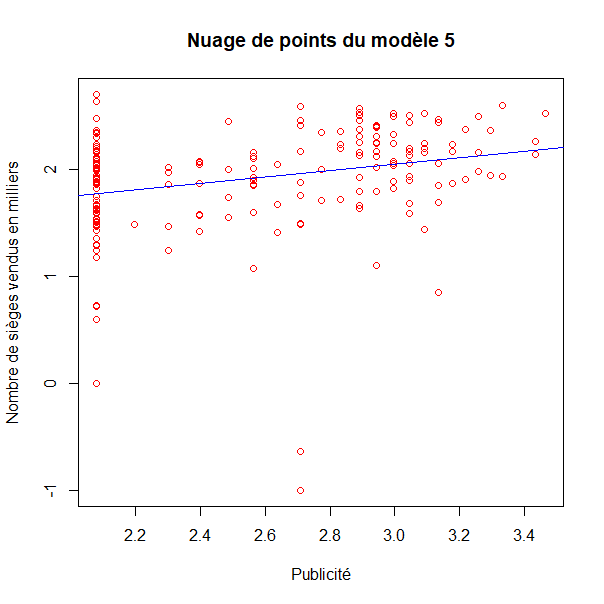
Response: log(sales)

Df Sum Sq Mean Sq F value Pr(>F)

log(8 + advertising) 1 3.270 3.2698 14.371 0.0002006 \*\*\*

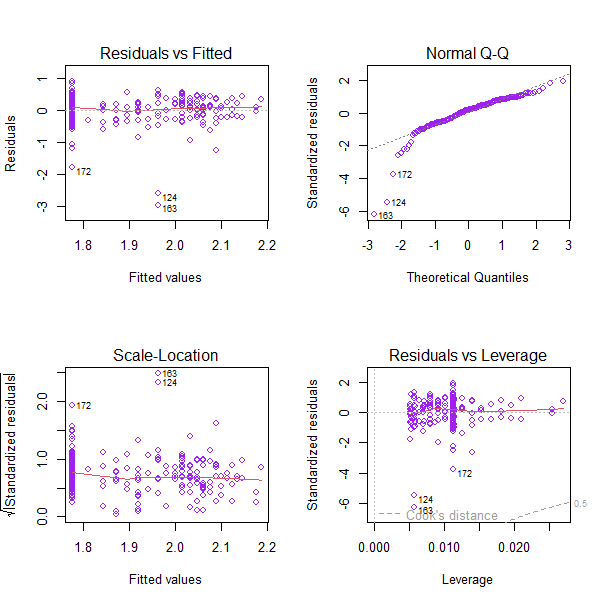
Residuals 193 43.913 0.2275

**Figure 32.** Tableau d’analyse de la variance du modèle 5.



**Figure 33.** Test de signification du modèle 5.

Pour ces tests, le critère de rejet de l’hypothèse est que la p-value soit plus petite que α. On pose comme hypothèse nulle que les coefficients soient égaux à zéro (H0 : β0 = 0 β1=0) ainsi que l’hypothèse alternative que les coefficients ne soient pas égaux à zéro (H1 : β0 ≠ 0 β1≠0). Dans ce modèle, α=0.05, la p-value (β0) < 2e-16 et la p-value (β1) = 0.000201. Ces valeurs sont significativement plus petites que 0.05, on peut donc rejeter l’hypothèse nulle dans les deux tests. On peut donc conclure que la relation linéaire entre la publicité (X2) et les ventes (Y) est significative. De plus, la valeur de R2 trouvée dans le tableau d’analyse de la variance est de 0.0693, étant loin de 1 ceci indique que le modèle n’est pas une bonne représentation de la relation entre les ventes et la publicité. Aussi, sur le nuage de points ci-dessus on peut observer que les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre la publicité et les ventes, ce qui confirme le rejet de l’hypothèse nulle de β1.



**Figure 34.** Analyse des résidus du modèle 5.

Pour évaluer la normalité des résidus, nous devons nous fier au « Normal Probability Plot » représenté ci-dessus. On peut observer que 85% des résidus en mauve suivent une ligne droite et se situent en bonne partie sur la ligne de normalité en rouge indiquant que l’hypothèse de normalité est respectée pour la majorité des observations. De plus, les résidus sont dispersés de façon homogène autour de la droite horizontale dans le graphique « Residuals vs Fitted » indiquant l’homoscédasticité. Cependant, dans ce même graphique, les résidus se situant en dehors de l’intervalle -2 à 2 doivent être rejetées, car ce sont des données atypiques.

2.5 % 97.5 %

(Intercept) 0.7654201 1.56267

log(8 + advertising) 0.1414325 0.44822

**Figure 35.** Intervalle de confiance pour les paramètres du modèle 5.

Calculée ci-dessus, avec un niveau de confiance 95%, le paramètre β0 est compris entre 0.7654201 et 1.56267 et le paramètre β1 est compris entre 0.1414325 et 0.44822.

Modèle 6 : Y = β0\*e^(β1\*X2+ε) :

Graphical user interface, text, application, email

Description automatically generatedÉquation transformée : ln(Y) = ln(β0) + β1\*X2+ ε :

Call:

lm(formula = log(sales) ~ advertising)

Residuals:

Min 1Q Median 3Q Max

-2.92923 -0.23988 0.09063 0.27957 0.91712

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.784246 0.048007 37.166 < 2e-16 \*\*\*

advertising 0.021533 0.005403 3.986 9.54e-05 \*\*\*

---

Residual standard error: 0.4753 on 193 degrees of freedom

Multiple R-squared: 0.07605, Adjusted R-squared: 0.07126

F-statistic: 15.88 on 1 and 193 DF, p-value: 9.539e-05

**Figure 36.** Tableau des coefficients de régression du modèle 6.

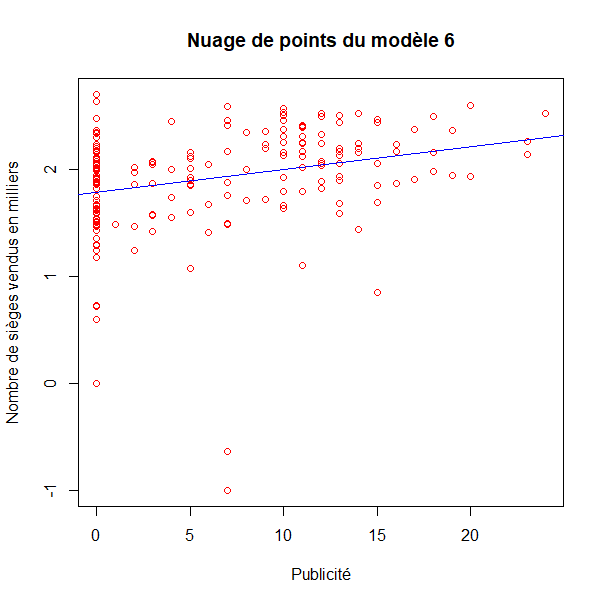
Analysis of Variance Table

Response: log(sales)

Df Sum Sq Mean Sq F value Pr(>F)

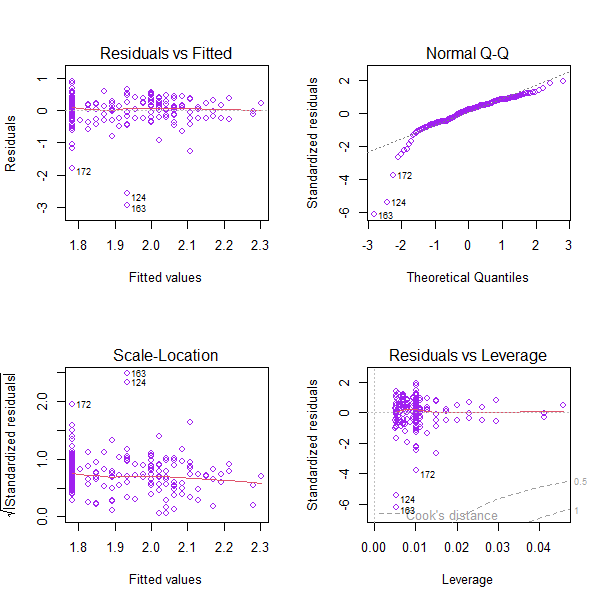
advertising 1 3.588 3.5881 15.885 9.539e-05 \*\*\*

Residuals 193 43.595 0.2259

**Figure 37.** Tableau d’analyse de la variance du modèle 6.

**Figure 38.** Test de signification du modèle 6.

Pour ces tests, le critère de rejet de l’hypothèse est que la p-value soit plus petite que α. On pose comme hypothèse nulle que les coefficients soient égaux à zéro (H0 : β0 = 0 β1=0) ainsi que l’hypothèse alternative que les coefficients ne soient pas égaux à zéro (H1 : β0 ≠ 0 β1≠0). Dans ce modèle, α=0.05, la p-value (β0) < 2e-16 et la p-value (β1) = 0.00111. Ces valeurs sont significativement plus petites que 0.05, on peut donc rejeter l’hypothèse nulle dans les deux tests. On peut donc conclure que la relation linéaire entre la publicité (X2) et les ventes (Y) est significative. De plus, la valeur de R2 trouvée dans le tableau d’analyse de la variance est de 0.05515, étant loin de 1 ceci indique que le modèle n’est pas une bonne représentation de la relation entre les ventes et la publicité. Aussi, sur le nuage de points ci-dessus on peut observer que les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre la publicité et les ventes, ce qui confirme le rejet de l’hypothèse nulle de β1.



**Figure 39.** Analyse des résidus du modèle 6.

Pour évaluer la normalité des résidus, nous devons nous fier au « Normal Probability Plot » représenté ci-dessus. On peut observer que 75% des résidus en mauve suivent une ligne droite et se situent en bonne partie sur la ligne de normalité en rouge indiquant que l’hypothèse de normalité est respectée pour la majorité des observations. De plus, les résidus sont dispersés de façon homogène autour de la droite horizontale dans le graphique « Residuals vs Fitted » indiquant l’homoscédasticité. Cependant, dans ce même graphique, les résidus se situant en dehors de l’intervalle -2 à 2 doivent être rejetées, car ce sont des données atypiques.

2.5 % 97.5 %

(Intercept) 1.68955949 1.87893259

advertising 0.01087728 0.03218965

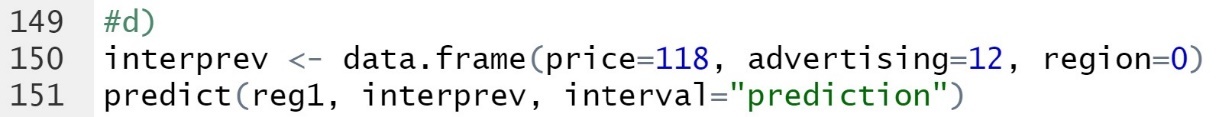
**Figure 40.** Intervalle de confiance pour les paramètres du modèle 6.

Calculée ci-dessus, avec un niveau de confiance 95%, le paramètre β0 est compris entre 1.68955949 et 1.87893259 et le paramètre β1 est compris entre 0.01087728 et 0.03218965.

Comparaison et choix du modèle le plus préférable :

En me basant sur le test de signification, R2, et la normalité des résidus, je choisis le modèle 1. Le modèle 1 démontre bien la relation significative entre le prix et les ventes. De plus, son R2 est celui qui se rapproche le plus de 1. Les résidus de ce modèle suivent mieux la droite normale que n’importe quel autre modèle. Bref, ce modèle représente le mieux la relation linéaire entre le prix et les ventes.

**d)**



fit lwr upr

1 7.372514 2.523374 12.22165

**Figure 41.** Intervalle de prévision des ventes

Avec un niveau de confiance de 95% ainsi que X1=102, X2=14, X3=0, on peut affirmer que les ventes seraient entre 2.523374 et 12.22165 milliers de ventes. Dans ce modèle, en utilisant les valeurs des coefficients trouvées dans le tableau des coefficients de régression, les ventes serait de 7.37251 milliers de ventes.