

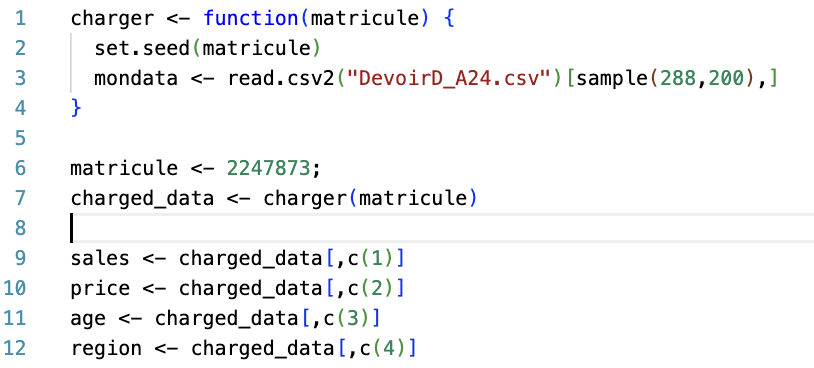
**Contexte :**

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

**Les données :**

A screenshot of a computer

Description automatically generatedLes données à analyser sont constituées d’un échantillon de 200 observations (points de vente) formées de 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).

****

**Figure 1 : Code pour générer les données personnalisées**

**Logiciel :**

Nous avons utilisé l’éditeur de texte Visual Studio code en langage R pour la première phase et RSudio pour la deuxième phase.

**Phase 1 : Analyse statistique descriptive et inférence**

**a)**

A screenshot of a computer code

Description automatically generated

**A blue graph with black text

Description automatically generated**

**Figure 2 : Histogramme des ventes**

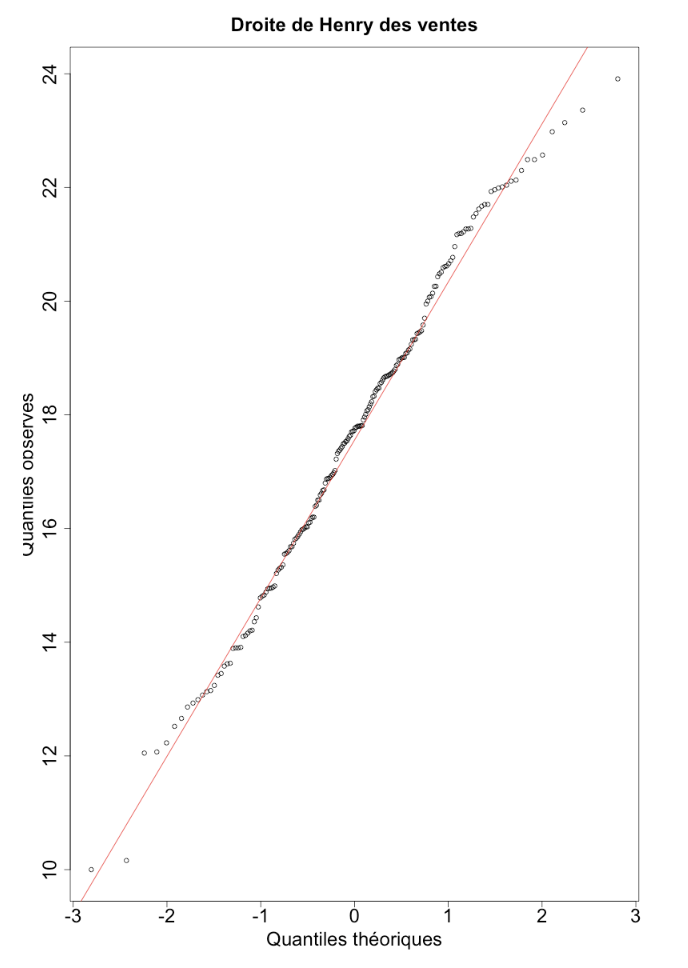
Dans cet histogramme, on peut observer la plage des ventes, qui varie de 0 à environ 24 000. La hauteur des barres de l'histogramme représente l'effectif, ce qui permet de constater que l'effectif le plus fréquent se situe entre 16 000 et 18 000 ventes, avec environ 50 occurrences. On peut donc porter comme hypothèse que la médiane se situe également entre ces deux valeurs. Par ailleurs, on note un très faible effectif entre 10 000 et 12 000 ventes. La forme de l'histogramme ressemble à celle de la distribution normale, ce qui permettrait d'envisager cette hypothèse. Toutefois, il est impossible de confirmer cette hypothèse, car on ne peut mesurer la symétrie de la distribution par rapport à la médiane seulement avec cet histogramme.

**A green rectangle with a line

Description automatically generated**

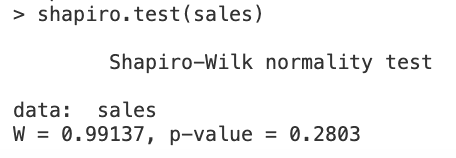
**Figure 3 : Boite de Turkey des ventes**

Tout d'abord, cette boite de Turkey représente les quartiles de la variable. Dans ce cas, le premier quartile se situe autour de 16 000 ventes, la médiane se situe autour de 17 800-18 000 ventes et le troisième quartile se situe autour de 19 500 ventes. Par ailleurs, il est possible d'estimer les valeurs extrêmes à partir de ce diagramme: la valeur minimale est d’environ 10 000 ventes et la valeur maximale est d’environ 24 000 ventes. Ainsi, on peut conclure que la moitié des données se situent entre 16 000 et 19 500 ventes, car la boîte de couleur verte représente l'écart interquartile contenant 50% des observations par définition (<https://fr.wikipedia.org/wiki/Boîte_à_moustaches> ). De plus, on peut remarquer que la médiane est un peu plus proche du deuxième quartile que du troisième. Il est aussi important de noter qu’il y a une donnée aberrante (N’est situé dans aucun des quartiles) à environ 10 000 ventes.



**Figure 4 : Droite de Henry des ventes**

Sur la droite de Henry des ventes représentée sur la figure 4, on voit que la droite suit une série de petits points représentant les données de vente. Étant donné que la majorité des données de vente 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 rouge 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 de normalité. 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.



**Figure 5 : Résultat du test Shapiro-Wilk**

D’une part, on voit que le résultat présent sur la figure 5 pour la valeur de W est très élevé et se rapproche de 1. Dans ce contexte, une valeur de W proche de 1 indique que les données sont proches d’une distribution normale. D’autre part, dans le test de Shapiro-Wilk, lorsque la valeur de *p* est inférieure au seuil critique de 0.05, alors les données ne suivent pas une distribution normale et on doit rejeter H0. Dans notre cas, la valeur de *p* est supérieure au seuil critique et donc on ne rejette pas H0. Ainsi, il n’y a pas de preuve suffisante pour conclure que les données de sales ne suivent pas une distribution normale. Par conséquent, nous pouvons supposer que sales (La variable Y) est normalement distribué.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Moyenne** | **Q1** | **Q2(médian)** | **Q3** | **Écart type** | **Intervalle de confiance** |
| 17.62 | 15.68 | 17.75 | 19.43 | 2.802118 | [17.23028, 18.01172] |

A screenshot of a computer code

Description automatically generated**Tableau 2 : Statistiques descriptives des ventes (en milliers)**

**b)**

A graph of a number of different sizes

Description automatically generated with medium confidence

**Figure 6 : Histogrammes des ventes pour les points de vente ruraux et urbain**

Dans la région rurale, les ventes sont concentrées entre 12 000 et 22 000 sièges, avec des pics de fréquence autour de 14 000, 16 000 et 18 000 sièges. En milieu urbain, les ventes s’étendent de 10 000 à 22 000 sièges, avec un pic de fréquence plus élevé autour de 18 000 sièges, indiquant des ventes plus fréquentes et globalement plus élevées. De plus, la région urbaine présente une dispersion plus large, montrant une variation plus importante des ventes. En résumé, les ventes semblent globalement plus élevées et plus concentrées en milieu urbain, bien qu’un test statistique soit nécessaire pour confirmer cette différence, comme le test de *Student* vu en cours, qui permet de comparer les moyennes des deux groupes (rural et urbain).

A graph showing a bar chart

Description automatically generated with medium confidence

**Figure 7 : Boites de Turkey pour les ventes par région**

Les deux boîtes de Tukey présent dans la figure 7 comparent les ventes de sièges d’automobile en milliers dans les régions rurale (0) et urbaine (1). La médiane, représentée par la ligne noire au centre de chaque boîte, est proche de 18 000 sièges pour la région urbaine et 17 500 pour la région rurale. La hauteur des boîtes, correspondant à l’étendue interquartile (IQR), montre une dispersion des valeurs centrales comparable entre les ventes en milieu rural et urbain à une centaine de ventes près. Les “moustaches” des boîtes, qui représentent l’étendue des valeurs non aberrantes, montrent également une variabilité similaire des ventes dans les deux régions. Sans oublier, deux valeurs aberrantes pour la région urbaine qui sont proches de 10 000 ventes. En conclusion, malgré de légère différence, les deux distributions de ventes sont globalement similaire en milieu rural et urbain en termes de médiane et de dispersion

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Moyenne** | **Q1** | **Q2(médian)** | **Q3** | **Écart type** | **Variance** | **Intervalle de confiance** |
| Rural | 17.64 | 15.76 | 17.50 | 19.47 | 2.661843 | 7.085407 | [17.04707, 18.23951] |
| Urbain | 17.61 | 15.68 | 17.78 | 19.32 | 2.900855 | 8.414958 | [17.08431, 18.12858] |

**Tableau 3 : Tableau de statistiques descriptives pour chacun des groupes**

A white background with black text

Description automatically generated

**Figure 8 : Résultat du test d’hypothèse sur l’égalité des variances**

Comme la figure 8 le montre, le test F effectué compare les variances des ventes de sièges dans les régions rurale (0) et urbaine (1). L’hypothèse nulle (H\_0) de ce test suppose que les variances des ventes dans les deux régions sont égales, tandis que l’hypothèse alternative (H\_1) stipule que les variances sont différentes. La statistique F calculée est de 0.842, avec une p-valeur de 0.4157. Étant supérieure au seuil de signification de 0.05, cette p-valeur indique qu’il n’y a pas de différence significative entre les variances des deux groups. L’intervalle de confiance à 95 % pour le ratio des variances est [0.5662, 1.2745], et comme il inclut la valeur 1, cela signifie que le ratio des variances pourrait bien être de 1, renforçant l’idée qu’il n’y a pas de différence significative entre les deux variances.

A screenshot of a computer code

Description automatically generated

**Figure 9 : Résultat du test d’hypothèse sur l’égalité des moyennes**

Comme on le voit dans la figure 9, le test de Welch compare les moyennes des ventes de sièges entre les régions rurale (0) et urbaine (1) pour vérifier s’il existe une différence significative. L’hypothèse nulle suppose que les moyennes sont égales, tandis que l’hypothèse alternative indique une différence. Dans le résultat on obtient une valeur de *t* = 0.092333, une valeur proche de 0. Ce qui signifie qu’il n’y a quasiment pas de différence entre la moyenne de chacun des groupes. Ensuite, la p-valeur de 0.9265, bien supérieure au seuil critique de 0.05, ce qui signifie qu’il n’y a pas de différence statistiquement significative entre les moyennes des deux régions. L’intervalle de confiance [-0.7506526, 0.8243423] inclut 0, ce qui confirme que la différence réelle pourrait être nulle. En conclusion, les moyennes des ventes dans les régions rurale et urbaine sont statistiquement similaires

**Phase 2 : Recherche du meilleur modèle**

**c)**

**A screenshot of a computer program

Description automatically generatedModèle 1 :**

*Modèle 1 : Y = β0 + β1X1 + ε ;*

A screenshot of a computer

Description automatically generated

**Figure 10 : Résultat du tableau des coefficients de régression du modèle 1**

A close-up of numbers

Description automatically generated

**Figure 11 : Résultat du tableau d’analyse de la variance du modèle 1**

A graph of red dots

Description automatically generated

**Figure 12 : Régression linéaire de la variable (Test de signification du modèle 1)**

**A group of graphs showing different values

Description automatically generated**

**Figure 13 : Analyse des résidus du modèle 1**

**Test significativité́ du modèle 1 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 1.872e-11 < seuil critique = 0.05, donc on rejette H0
* Ainsi, le modèle 1 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 1 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* A number and text on a white background

  Description automatically generatedDans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (23, 26, 112) pouvant être rejeté pour une meilleure représentation du modèle

**Figure 14 : Intervalle de confiance pour les paramètres du modèle 1**

Comme on le voit sur la figure 14, avec un niveau de confiance 95%, le paramètre

β0 ∈  [21.89412686 25.26490213]

et

β1 ∈  [-0.06602391 -0.03740456]

**Modèle 2 :**

A screenshot of a computer

Description automatically generated

**Figure 15 : Résultat du tableau des coefficient de régression du modèle 2**

A table with numbers and letters

Description automatically generated

**Figure 16 : Résultat d’analyse du tableau de variance du modèle 2**

A graph with red dots

Description automatically generated

**Figure 17 : Nuage de points du modèle 2**

A graph of a graph

Description automatically generated with medium confidence

**Figure 18 : Analyse des résidus du modèle 2**

**Test significativité́ du modèle 2 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 3.117e-6 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre l’âge et les ventes.
* Ainsi, le modèle 2 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 2 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (25, 99, 150) pouvant être rejeté pour une meilleure représentation du modèle

A number and percentage symbol

Description automatically generated with medium confidence

**Figure 19 : Intervalle de confiance pour les paramètres du modèle 2**

Comme on le voit sur la figure 19, avec un niveau de confiance 95%, le paramètre

β0 ∈  [19.2706291 21.76261108]

et

β1 ∈  [-0.0773122 -0.03229735]

**Modèle 3 :A screenshot of a computer program

Description automatically generated**

**Figure 20 : Résultat du tableau des coefficient de régression du modèle 3**

**A close-up of a table

Description automatically generated**

**Figure 21 : Résultat du tableau des variances du modèle 3**

**A graph with red dots

Description automatically generated**

**Figure 22 : Nuage de points du modèle 3**

A collage of graphs

Description automatically generated

**Figure 23 : Analyse des résidus du modèle 3**

**Test significativité́ du modèle 3 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 1.264e-11 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre le prix^2 et les ventes.
* Ainsi, le modèle 3 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 3 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (23, 26, 112) pouvant être rejeté pour une meilleure représentation du modèle

A number and a number

Description automatically generated with medium confidence

**Figure 24 : Intervalle de confiance pour les paramètres du modèle 3**

Comme on le voit sur la figure 24, avec un niveau de confiance 95%, le paramètre

β0 ∈  [19.856145801 21.7296740605]

et

β1 ∈  [-0.000291329 -0.0001659811]

**Modèle 4 :A screenshot of a computer error

Description automatically generated**

**Figure 25 : Résultat du tableau des coefficient de régression du modèle 4**

A table with numbers and a table

Description automatically generated

**Figure 26 : Résultat du tableau des variances du modèle 4**

A graph with red dots

Description automatically generated

**Figure 27 : Nuage de points du modèle 4**

A group of graphs showing different values

Description automatically generated

**Figure 28 : Analyse des résidus du modèle 4**

**Test significativité du modèle 4 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 3.339e-06 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre l’âge^2 et les ventes.
* Ainsi, le modèle 4 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 4 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (26, 156, 164) pouvant être rejeté pour une meilleure représentation du modèle

A number and numbers on a white background

Description automatically generated

**Figure 29 : Intervalle de confiance pour les paramètres du modèle 4**

Comme on le voit sur la figure 24, avec un niveau de confiance 95%, le paramètre

β0 ∈  [18.4555770124 19.9587494618]

et

β1 ∈  [-0.00007312044 -0.0003043969]

**Modèle 5 :**A screenshot of a computer program

Description automatically generated

**Figure 30 : Résultat du tableau des coefficient de régression du modèle 5**

A close-up of a table

Description automatically generated

**Figure 31 : Résultat du tableau des variances du modèle 5**

A graph with red dots

Description automatically generated

**Figure 32 : Nuage de points du modèle 5**

A collage of graphs

Description automatically generated

**Figure 33 : Analyse des résidus du modèle 5**

**Test significativité du modèle 5 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 1.405e-11 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre log(Prix) et les ventes.
* Ainsi, le modèle 5 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 5 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (23, 26, 99) pouvant être rejeté pour une meilleure représentation du modèle

A number and text on a white background

Description automatically generated

**Figure 34 : Intervalle de confiance pour les paramètres du modèle 5**

Comme on le voit sur la figure 34, avec un niveau de confiance 95%, le paramètre

β0 ∈  [3.109845728 3.308235024]

et

β1 ∈  [-0.003907058 -0.002222646]

**Modèle 6 :**

A screenshot of a computer

Description automatically generated

**Figure 35 : Résultat du tableau des coefficient de régression du modèle 6**

A close-up of a table

Description automatically generated

**Figure 36 : Résultat du tableau des variances du modèle 6**

A graph of red dots

Description automatically generated

**Figure 37 : Nuage de points du modèle 6**

A graph of a graph

Description automatically generated with medium confidence

**Figure 38: Analyse des résidus du modèle 6**

**Test significativité du modèle 6 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 3.583e-06 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre l’âge et le logarithme des ventes.
* Ainsi, le modèle 6 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 6 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (26, 99, 155) pouvant être rejeté pour une meilleure représentation du modèle

A number and text on a white background

Description automatically generated

**Figure 39 : Intervalle de confiance pour les paramètres du modèle 6**

Comme on le voit sur la figure 39, avec un niveau de confiance 95%, le paramètre

β0 ∈  [2.952091306 3.099066879]

et

β1 ∈  [-0.004538801 -0.001883263]

**Modèle 7 :**

A screenshot of a computer

Description automatically generated

**Figure 40 : Résultat du tableau des coefficient de régression du modèle 7**

A close-up of a table

Description automatically generated

**Figure 41 : Résultat du tableau des variances du modèle 7**

A graph with red dots

Description automatically generated

**Figure 42 : Nuage de points du modèle 7**

A graph of a graph

Description automatically generated with medium confidence

**Figure 43: Analyse des résidus du modèle 7**

**Test significativité du modèle 7 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 3.84 -10 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre le logarithme du prix et le logarithme des ventes.
* Ainsi, le modèle 7 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 7 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (23, 26, 99) pouvant être rejeté pour une meilleure représentation du modèle

A number and percentages on a white background

Description automatically generated

**Figure 44 : Intervalle de confiance pour les paramètres du modèle 7**

Comme on le voit sur la figure 44, avec un niveau de confiance 95%, le paramètre

β0 ∈  [3.8114780 4.6286974]

et

β1 ∈  [-0.3754187 -0.2025329]

**Modèle 8 :**

A screenshot of a computer

Description automatically generated

**Figure 45 : Résultat du tableau des coefficient de régression du modèle 8**

A close-up of a table

Description automatically generated

**Figure 46 : Résultat du tableau des variances du modèle 8**

A graph with red dots

Description automatically generated

**Figure 47 : Nuage de points du modèle 8**

A graph of a graph

Description automatically generated with medium confidence

**Figure 48: Analyse des résidus du modèle 8**

**Test significativité du modèle 8 :**

* Hypothèse H0 : *β1 =* 0 et H1 : *β1* ≠ 0
* On a p-value = 6.398e-06 < seuil critique = 0.05, donc on rejette H0
* Dans le nuage de points, les observations sont dispersées majoritairement autour de la ligne bleue représentant la relation entre le logarithme de l’âge et le logarithme des ventes.
* Ainsi, le modèle 8 est significatif au seuil critique de 0.05

**Analyse des résidus du modèle 8 :**

* On peut voir que dans le graphique «  Q-Q residuals » les résidus suivent une certaine ligne droite de normalité, l'hypothèse de normalité est donc vraie.
* Dans le graphique « Residuals vs Fitted » les résidus sont répartis de façon homogène autour de la ligne rouge horizontale indiquant une certaine homoscédasticité.
* Dans le graphique « Residuals vs Fitted » on peut aussi voir 3 points atypiques (26, 99, 155) pouvant être rejeté pour une meilleure représentation du modèle

A number and percentage symbol

Description automatically generated with medium confidence

**Figure 49 : Intervalle de confiance pour les paramètres du modèle 7**

Comme on le voit sur la figure 49, avec un niveau de confiance 95%, le paramètre

β0 ∈  [3.1943836 3.69854464]

et

β1 ∈  [-0.2151176 -0.08675208]

**Comparaison des modèles et choix du meilleur modèle :**

Après analyse des différents modèles, on a les modèles 1, 3 et 5 qui se destinguent des autres, car ils sont tous significatifs et ont une valeur de R2 assez élevé par rapport aux autres.  
Le modèle 3 reste plus significatif que le modèle 1 et le modèle 5, car il a une valeur de R2 plus grande (0.2032). C'est donc le modèle 3 qui représente le mieux la relation linéaire entre le prix et les ventes. Ainsi bien que le modèle 1 soit le plus représentatif et valide, il ne permet pas de bien expliquer la variabilité des valeurs du modèle.

**d) Choix du meilleur modèle :**

A black text on a white background

Description automatically generated

Ainsi avec un niveau de confiance de 95%, on peut prédire que le nombre de ventes sera compris entre **12.82404 et 22.71386** milliers pour un prix de 115$ et l’âge de 35 ans. Dans ce modèle (Modèle 3), en utilisant les valeurs des coefficients trouvées dans le tableau des coefficients de régression (β0 et β1), les ventes serait de 17.76895 milliers.