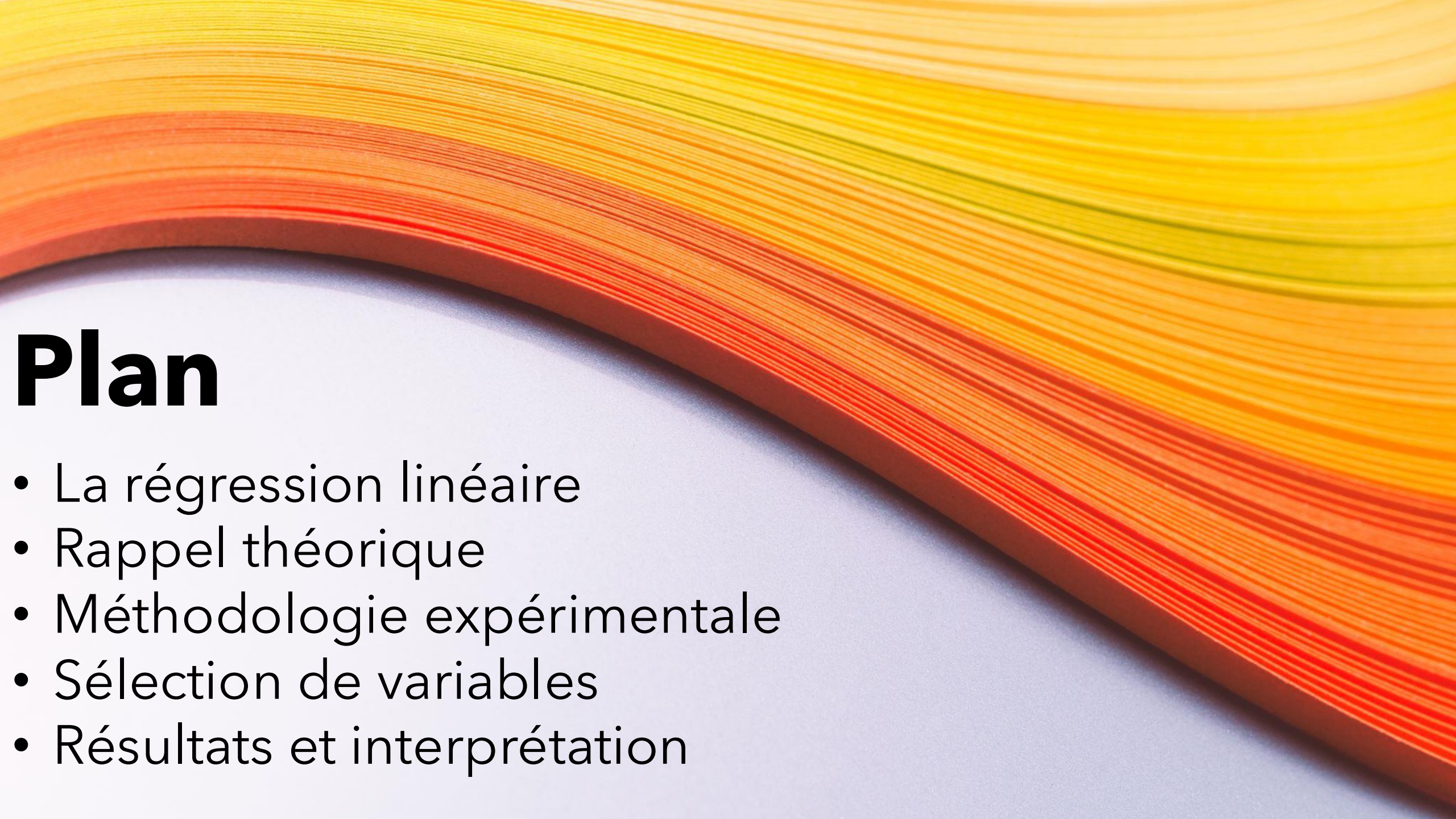
Régression linéaire

Martelli Gino – Senis Tahitoa – Vieville Sébastien

Introduction

On va étudier un jeu de données réel issu de R, le célèbre `mtcars`.
Objectif : comprendre, modéliser et prédire des variables d'intérêts à partir d'autres variables explicatives. Plus concrètement, il s'agit de voir si certaines caractéristiques d'un véhicule permettent de prédire, par exemple, sa consommation.



Plan

- La régression linéaire
- Rappel théorique
- Méthodologie expérimentale
- Sélection de variables
- Résultats et interprétation

Pourquoi la régression linéaire ?

Objectif :

- **Modéliser** une relation entre :
  - Y : variable dépendante (ex. : consommation *mpg*)
  - X : variable(s) explicative(s) (ex. : poids, cylindres)

Applications :

- **Prédiction** de valeurs futures
- **Interprétation** des relations entre variables
- **Aide à la décision :** En ingénierie, marketing, etc.

Rappel Théorique Régression Linéaire Simple

L'objectif est de modéliser la relation linéaire entre une variable explicative X et une variable réponse Y :

$$Y_i = ax_i + b + \varepsilon_i$$

- a : pente
- b : ordonnée à l'origine
- $\varepsilon i \sim N(0,\sigma 2)$ i.i.d.

On minimise la somme des carrés des résidus pour trouver les estimateurs :

$$\min_{a,b} \sum_{i=1}^{n} (Y_i - ax_i - b)^2$$

- Estimateurs :

$$\hat{a}_n = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{Y}_n}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

$$\hat{b}_n = \bar{Y}_n - \hat{a}_n \bar{x}_n$$

Rappel Théorique Régression Linéaire Simple

Propriétés des estimateurs :

- Non biaisés : $\mathbb{E}[\hat{a}_n] = a, \quad \mathbb{E}[\hat{b}_n] = b$
- Variances : $\mathbb{V}[\hat{b}_n] = \frac{\sigma^2 . \sum_{i=1}^{n} x_i^2}{n . \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}, \; \mathbb{V}[\hat{a}_n] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$
- Covariances : $\mathrm{Cov}(\hat{a}_n, \hat{b}_n) = -\frac{\bar{x}\sigma^2}{\sum (x_i - \bar{x})^2}$

Rappel Théorique Régression Linéaire Multiple

L'objectif est de modéliser la relation entre une variable réponse Y et plusieurs variables explicatives X1,X2,...,Xp .

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad \text{avec } \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d.}$$

Forme matricielle : $Y = X\beta + \varepsilon$

$Y \in \mathbb{R}^n$: vecteur des observations
$X \in \mathbb{R}^{n \times (p+1)}$: matrice des variables explicatives (1ère colonne de 1)
$\beta \in \mathbb{R}^{p+1}$: vecteur des coefficients
$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$: bruit gaussien centré, variance σ^2

Méthodologie expérimentale

Préparation des données

- Jeu de données utilisé : `mtcars`
- Variables sélectionnées :

`mpg` : Miles per gallon (consommation)	**`vs`** : Forme du moteur (V-shape)
`qsec` : Temps sur 1/4 mile (sec)	**`hp`** : Horsepower (puissance)
`cyl` : Nombre de cylindres	**`gear`** : Nombre de vitesses
`drat` : Rear axle ratio (rapport pont arrière)	**`wt`** : Poids du véhicule

Méthodologie expérimentale

```
A = mtcars
F1 = as.factor(A[,8])
A[,8] = F1
set.seed(3)
set.seed(3*floor(100*runif(1,0,3)))
set1 = sample(1:32,1)
B = A[-set1,]
Y = B[,1]
u = 1:11
v = u[-c(1,8,9)]
set2 = c(8, sample(v,6,replace=FALSE))
X = B[,set2]
```

Création d'un **sous-échantillon** de `mtcars` :

- La **variable cible mpg**
- **7 variables explicatives** : vs + 6 autres tirées au hasard
- Une voiture est retirée aléatoirement pour rendre l'échantillon unique

Sélection de variables

Objectif : Garder les variables qui **expliquent le mieux** `mpg` sans surcharger le modèle

Critère utilisé :

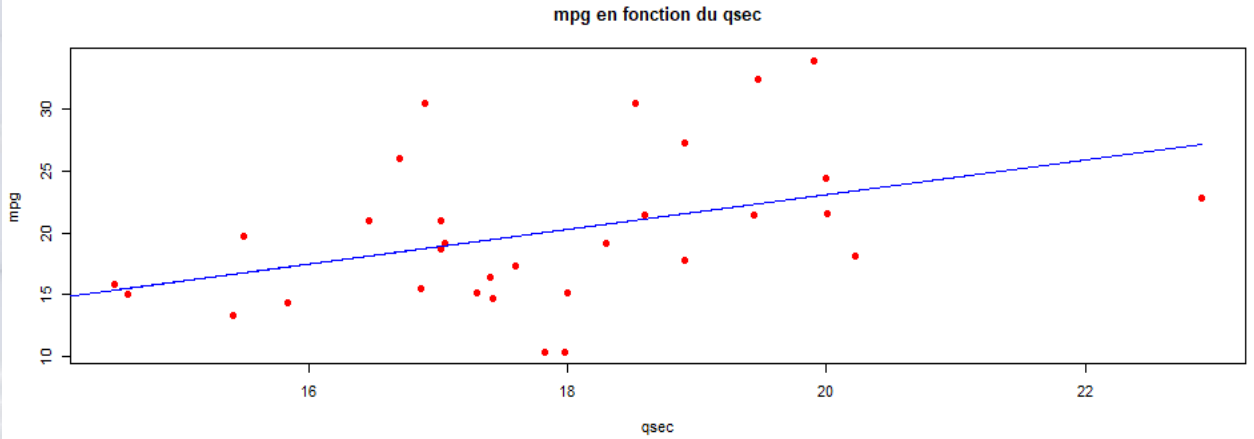
- **R² ajusté** :
  - Corrige le R² classique qui **augmente toujours** quand on ajoute des variables
  - Le R² ajusté **n'augmente que si la variable ajoutée apporte réellement de l'information**

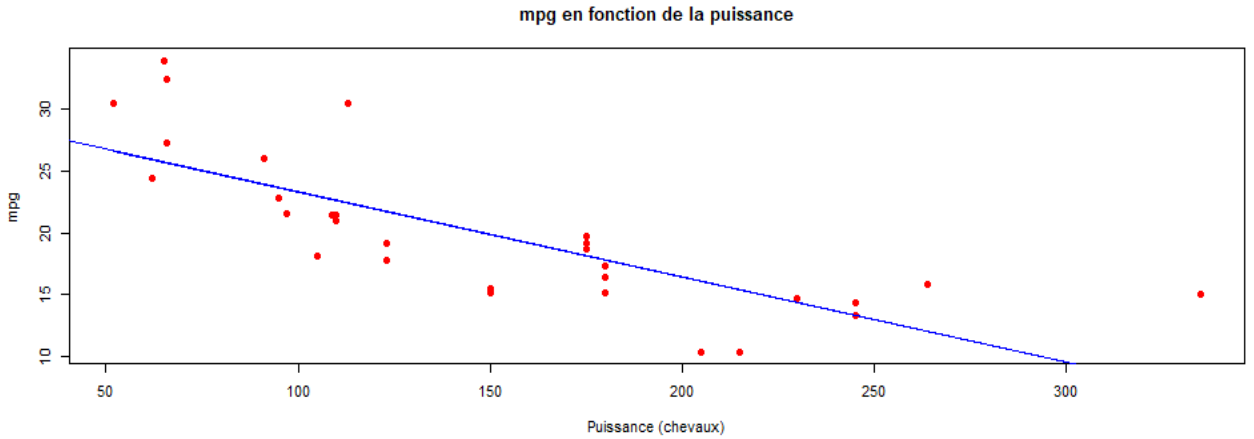
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

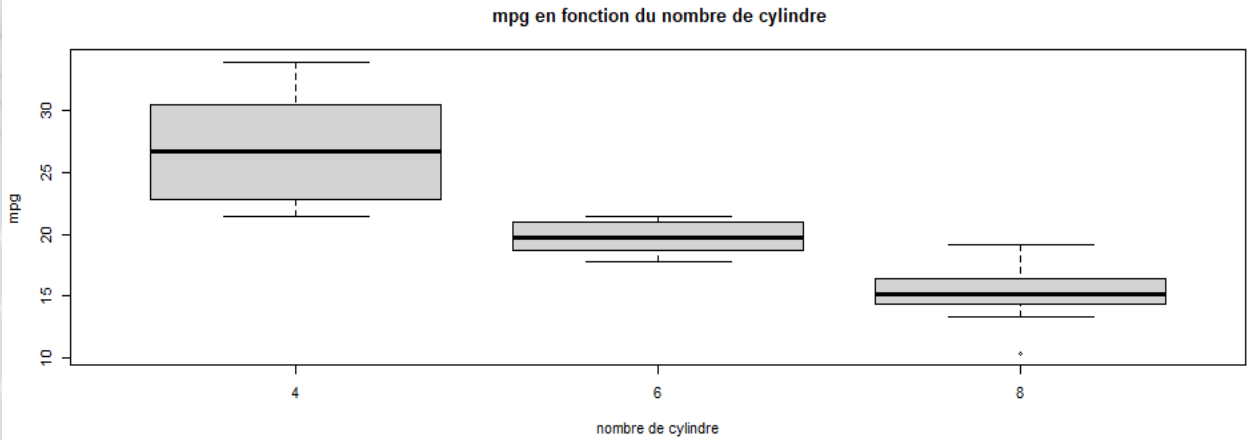
$$R^2_{\text{ajusté}} = 1 - \frac{(1-R^2)(n-1)}{n - rang(X)}$$

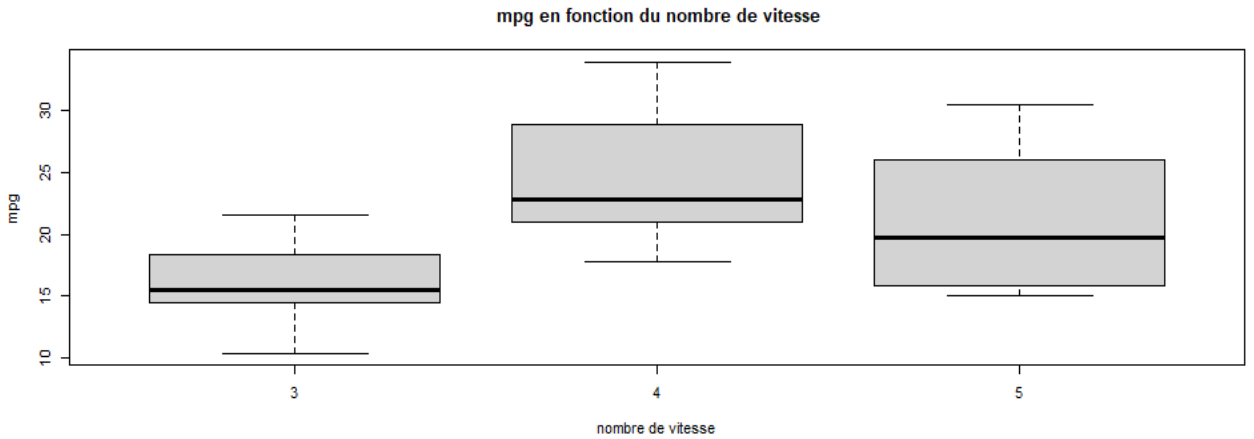
Résultats et interprétation : simple

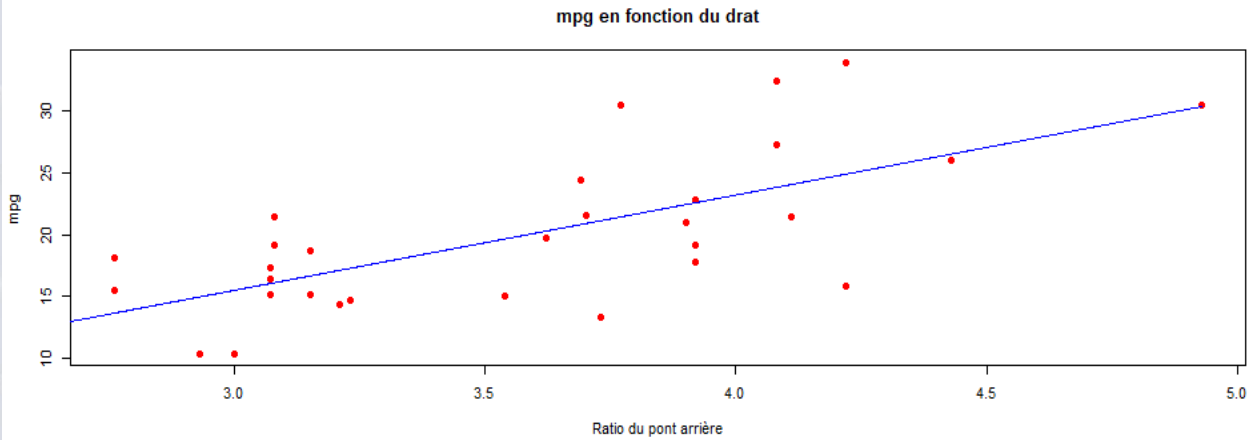
- Modélisation de la droite ax + b
- Prédiction
- Vérification de R^2
- Vérification de l'hypothèse de bruit
- Test d'adéquation de Kolmogorov-Smirnov

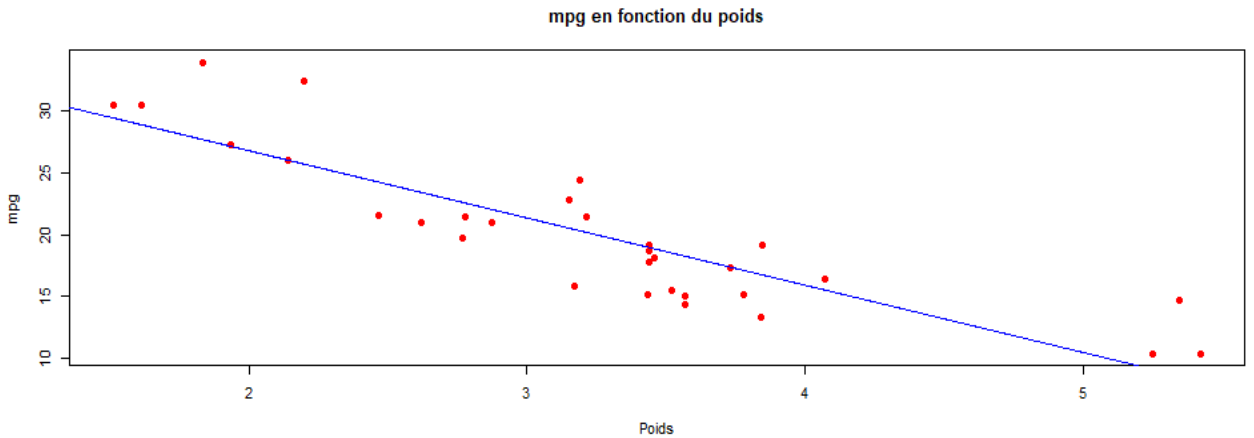

mpg en fonction du qsec
mpg
qsec
10
15
20
25
30
16
18
20
22

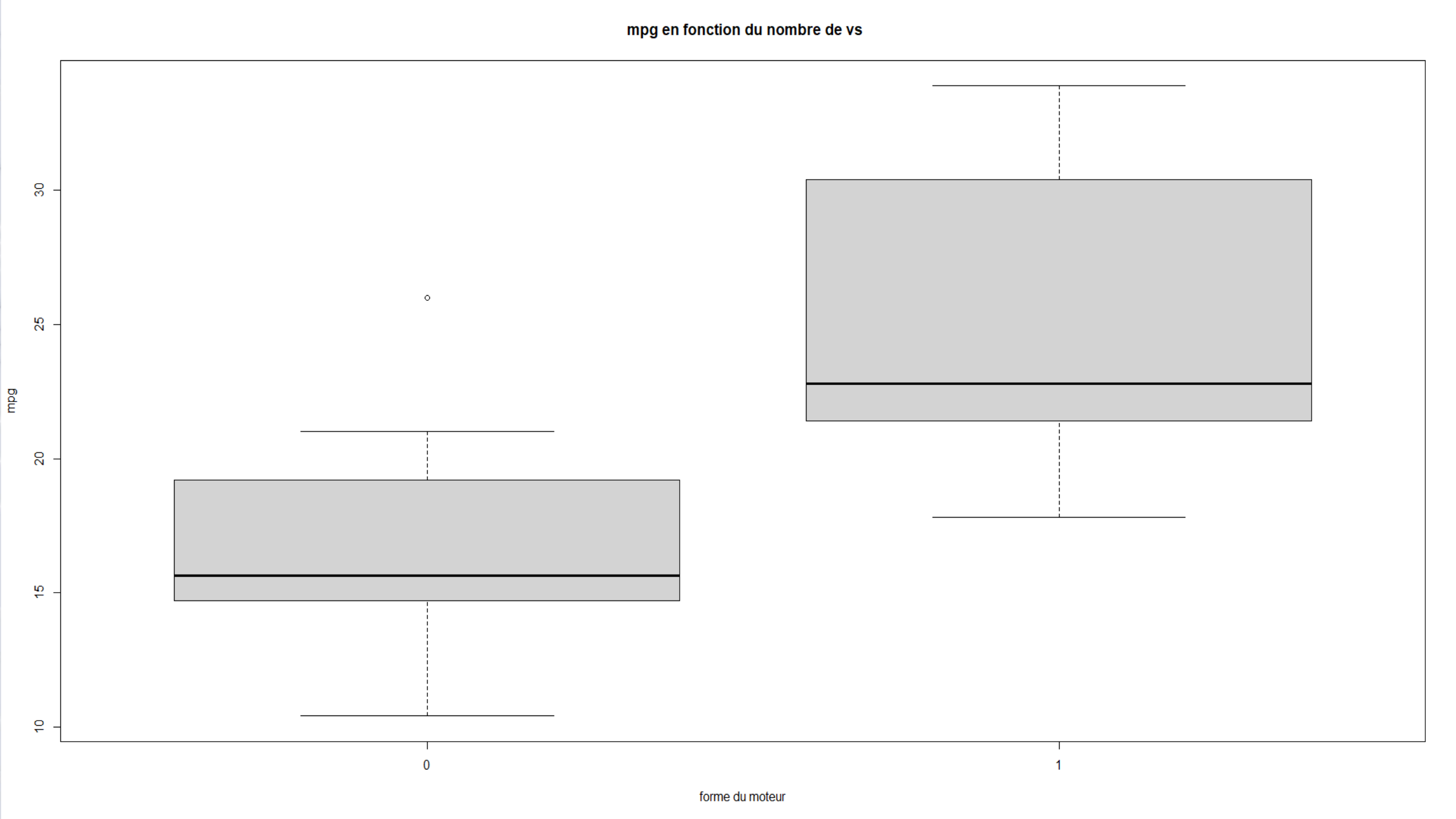

mpg en fonction de la puissance
mpg
Puissance (chevaux)
10
15
20
25
30
50
100
150
200
250
300


mpg en fonction du nombre de cylindre
mpg
nombre de cylindre
10
15
20
25
30
4
6
8

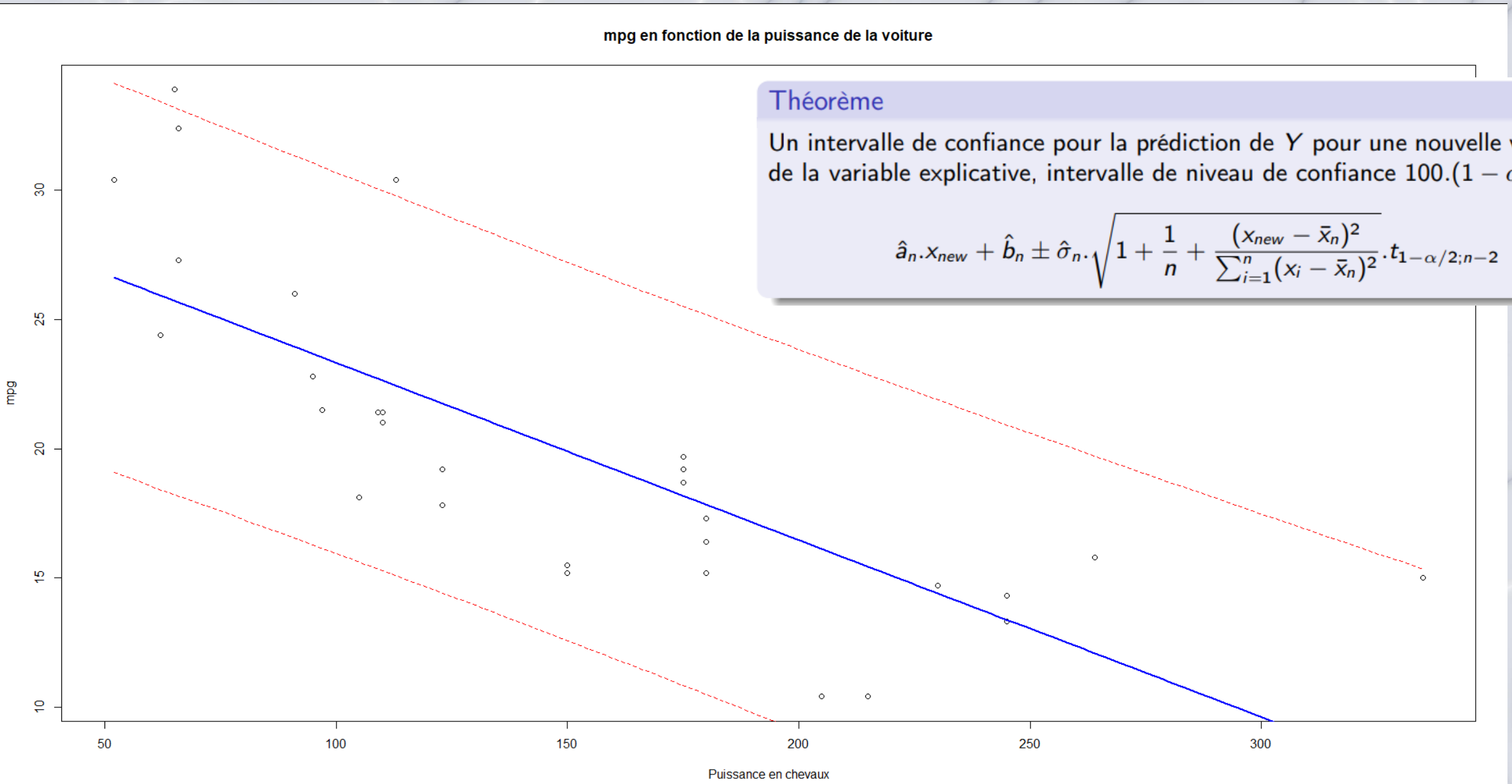

mpg en fonction du nombre de vitesse
mpg
nombre de vitesse
10
15
20
25
30
3
4
5


mpg en fonction du drat
mpg
Ratio du pont arrière
10
15
20
25
30
3.0
3.5
4.0
4.5
5.0


mpg en fonction du poids
mpg
Poids
10
15
20
25
30
2
3
4
5


mpg en fonction du nombre de vs
mpg
10
15
20
25
30
0
1
forme du moteur

Prédiction



Théorème

Un intervalle de confiance pour la prédiction de Y pour une nouvelle valeur x_{new} de la variable explicative, intervalle de niveau de confiance $100.(1-\alpha)\%$, est :

$$\hat{a}_n.x_{new} + \hat{b}_n \pm \hat{\sigma}_n.\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}.t_{1-\alpha/2;n-2}$$

Prédiction

Y - Y_prédiction

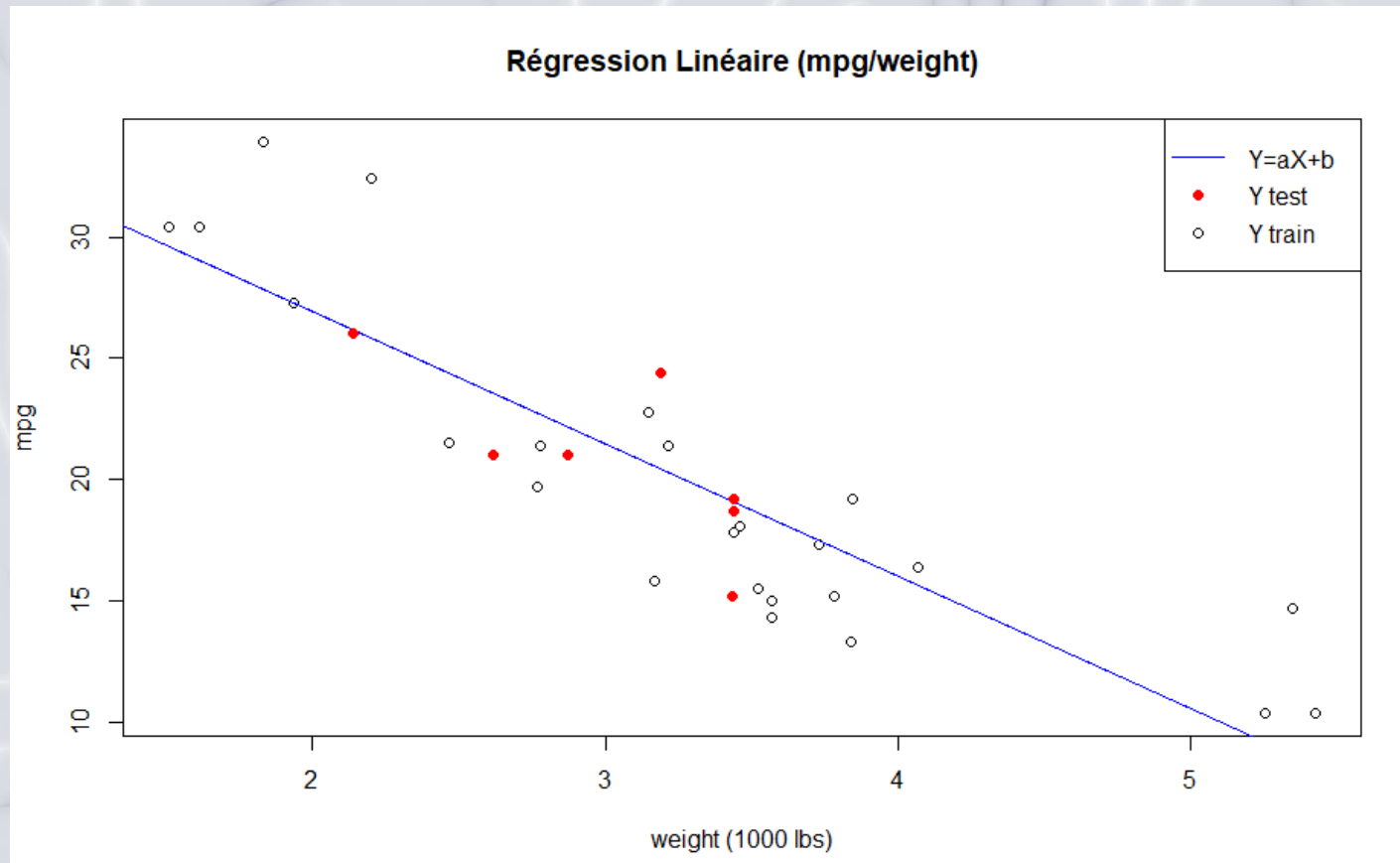
```
-2.5494606 -1.1589177 -0.3779108  3.9588118  0.1220892 -3.9051764 -0.1669531
```

Moyenne de l'erreur en valeur absolue.

$$\frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{T}} |y_i - (\beta_0 + \beta_1 x_i)| = \boxed{1.974084}$$

Moyenne de l'erreur au carré.

$$\frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{T}} (y_i - (\beta_0 + \beta_1 x_i))^2 = \boxed{6.282424}$$



Prédiction

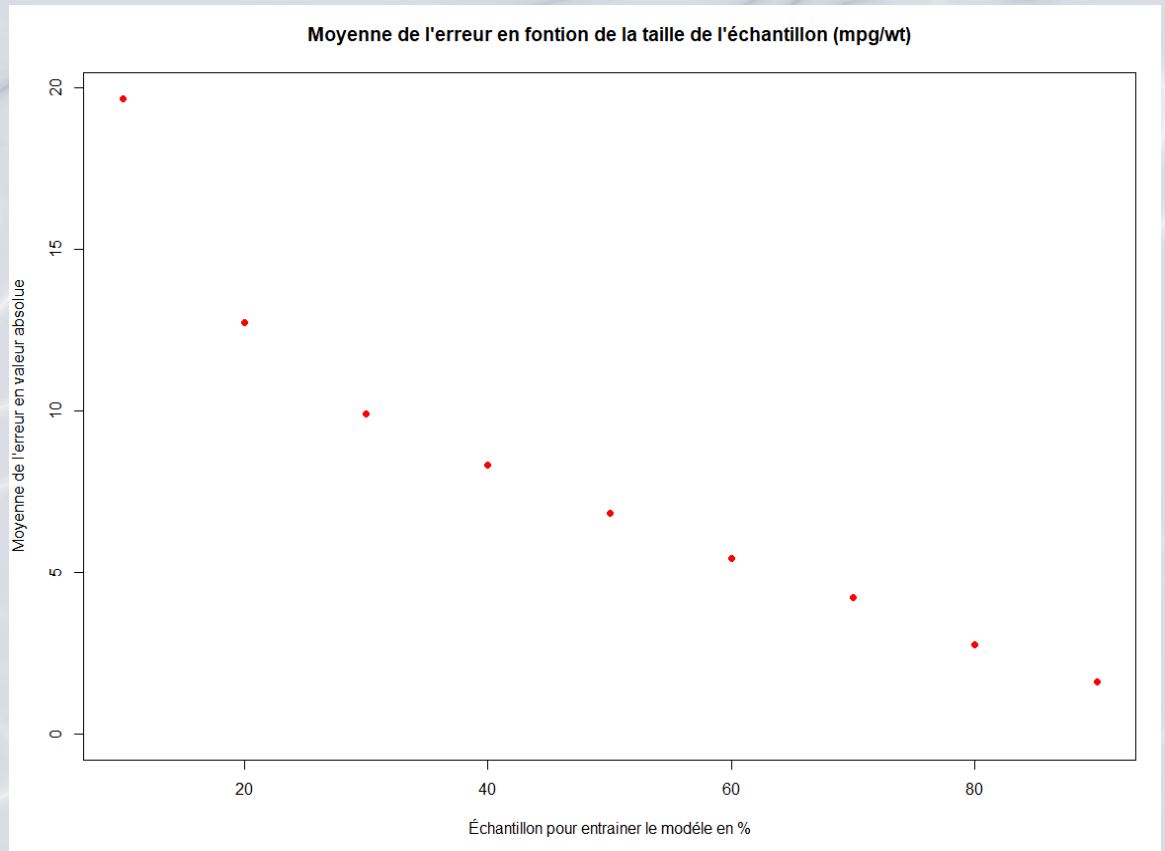
Moyenne de l'erreur en valeur absolue.

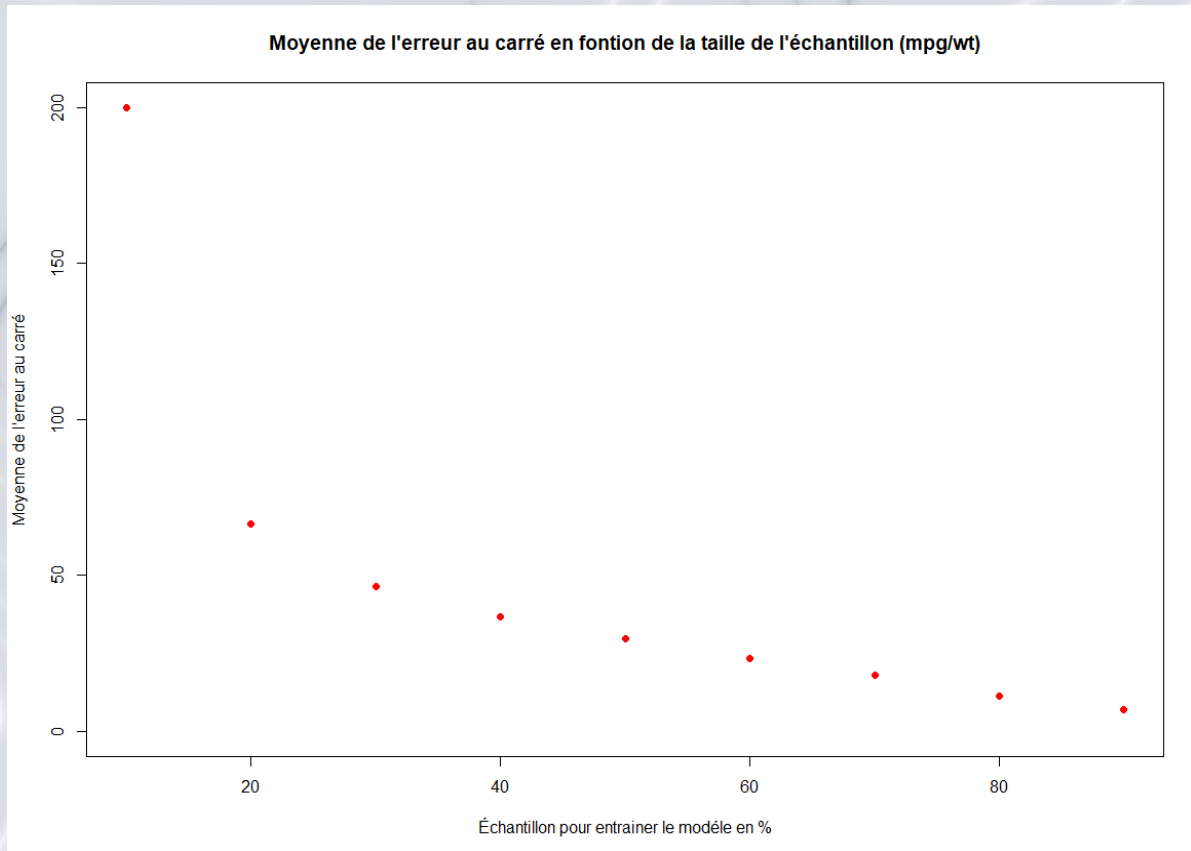
$$\frac{1}{100}\sum_{i=1}^{100}\left[\frac{1}{n_{\text{test}}}\sum_{j\in\mathcal{T}_i}\left|y_j-(\beta_0^{(i)}+\beta_1^{(i)}x_j)\right|\right]=\boxed{2.814278}$$

Moyenne de l'erreur au carré.

$$\frac{1}{100}\sum_{i=1}^{100}\left[\frac{1}{n_{\text{test}}}\sum_{j\in\mathcal{T}_i}\left(y_j-(\beta_0^{(i)}+\beta_1^{(i)}x_j)\right)^2\right]=\boxed{11.6429}$$

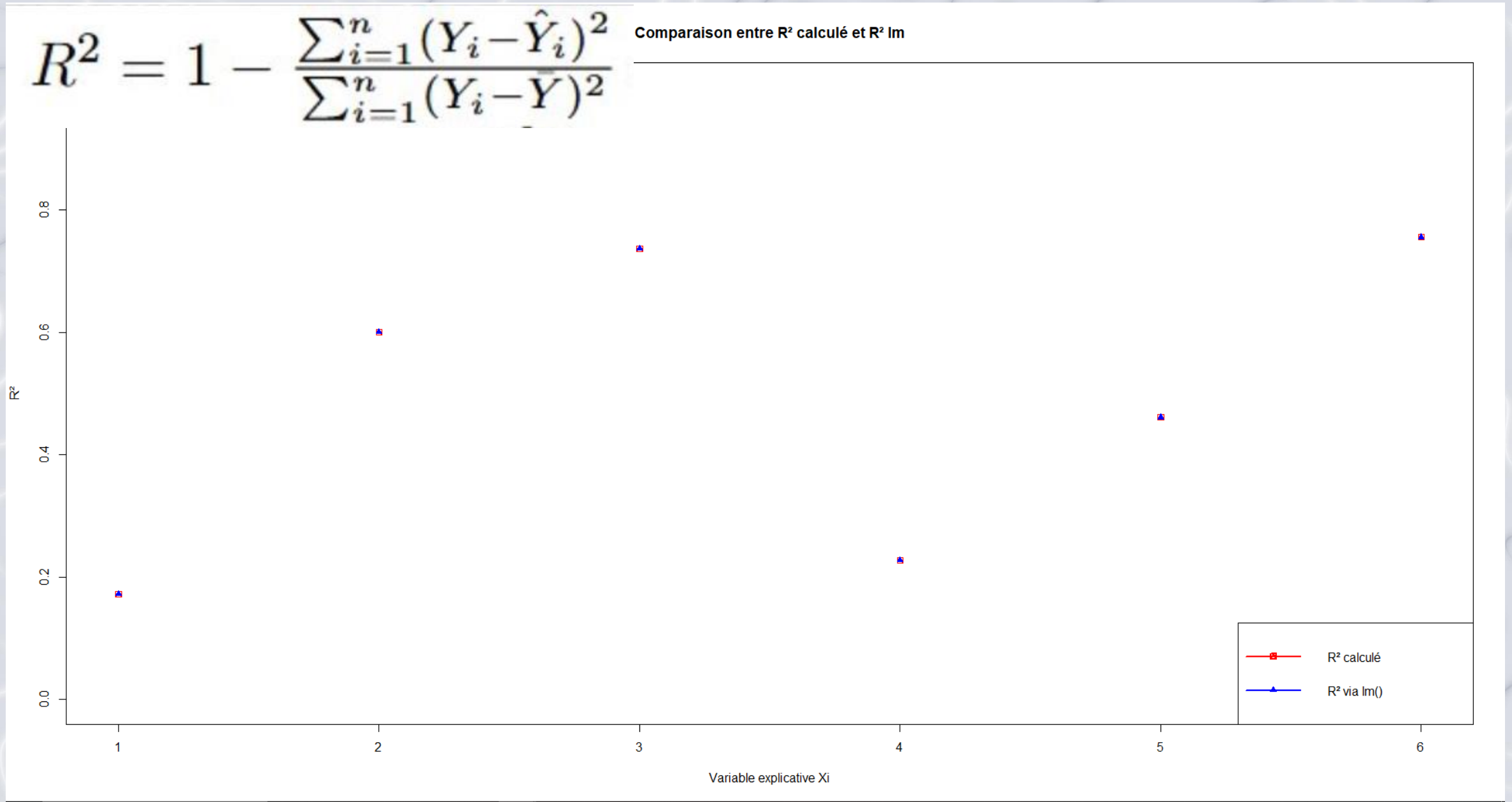
Prédiction


Moyenne de l'erreur en fontion de la taille de l'échantillon (mpg/wt)
Moyenne de l'erreur en valeur absolue
0
5
10
15
20
20
40
60
80
Échantillon pour entrainer le modéle en %


Moyenne de l'erreur au carré en fontion de la taille de l'échantillon (mpg/wt)
Moyenne de l'erreur au carré
0
50
100
150
200
20
40
60
80
Échantillon pour entrainer le modéle en %

Vérification de R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$



P-value

Régression linéaire avec les sept variables

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.35699   18.19650   1.448  0.16098
Xmatvs      -0.11380    2.05901  -0.055  0.95640
Xmatqsec     0.37576    0.68080   0.552  0.58631
Xmathp      -0.01626    0.01788  -0.910  0.37239
Xmatcyl     -0.56616    0.98961  -0.572  0.57280
Xmatgear     0.37213    1.22369   0.304  0.76378
Xmatdrat     0.71457    1.59828   0.447  0.65899
Xmatwt      -3.34725    1.09780  -3.049  0.00569 **
```

Régression linéaire sans la variable "weight"

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.92726   20.46757   1.951   0.0629
Xmatvs       1.50090    2.30818   0.650   0.5217
Xmatqsec    -0.78193    0.65552  -1.193   0.2446
Xmathp      -0.04099    0.01848  -2.218   0.0363
Xmatcyl     -1.38083    1.10535  -1.249   0.2236
Xmatgear     0.65031    1.41558   0.459   0.6501
Xmatdrat     1.17507    1.84578   0.637   0.5304
```

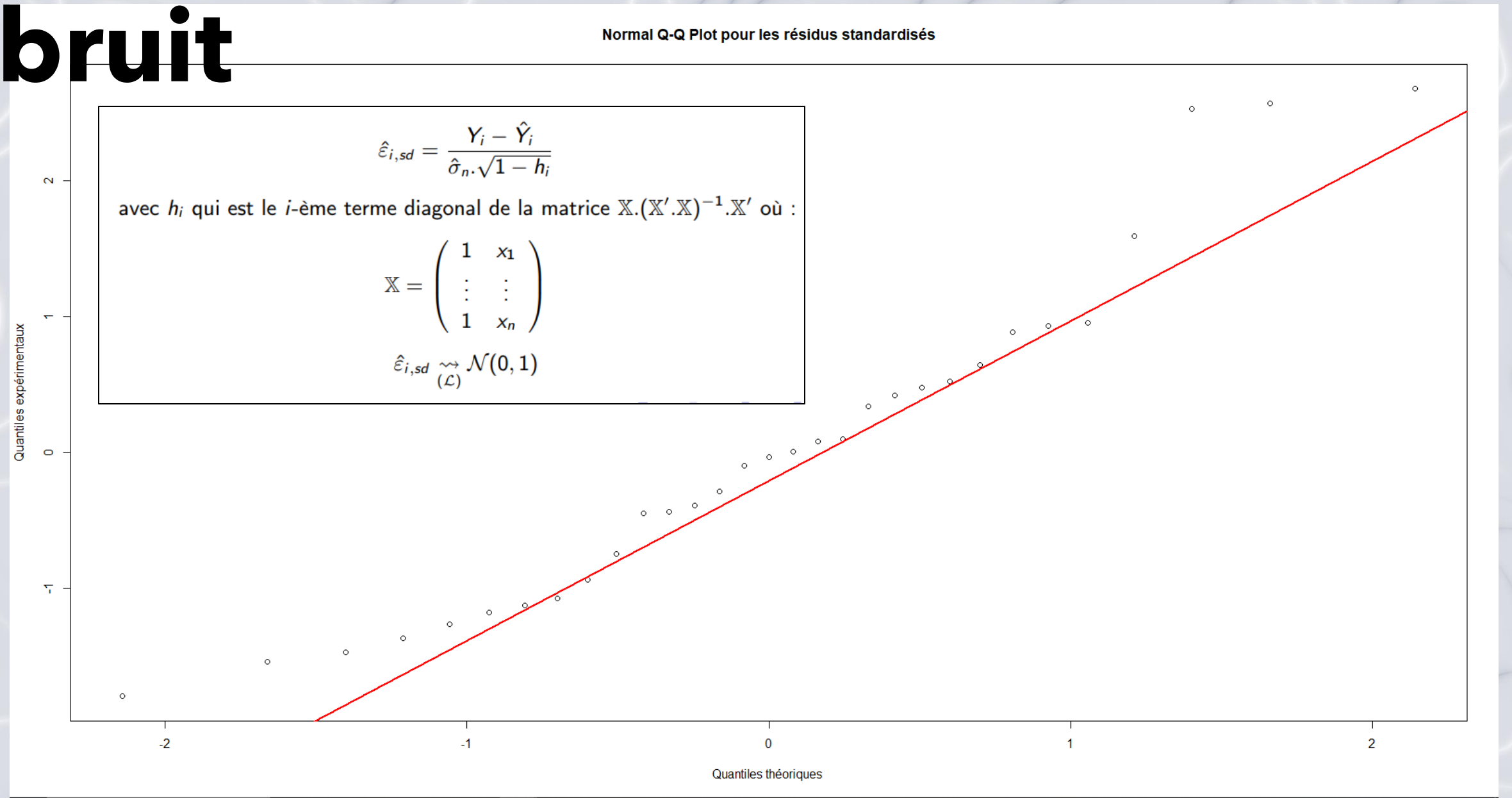
Validité des hypothèses sur le bruit

$$\hat{\varepsilon}_{i,sd} = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_n . \sqrt{1 - h_i}}$$

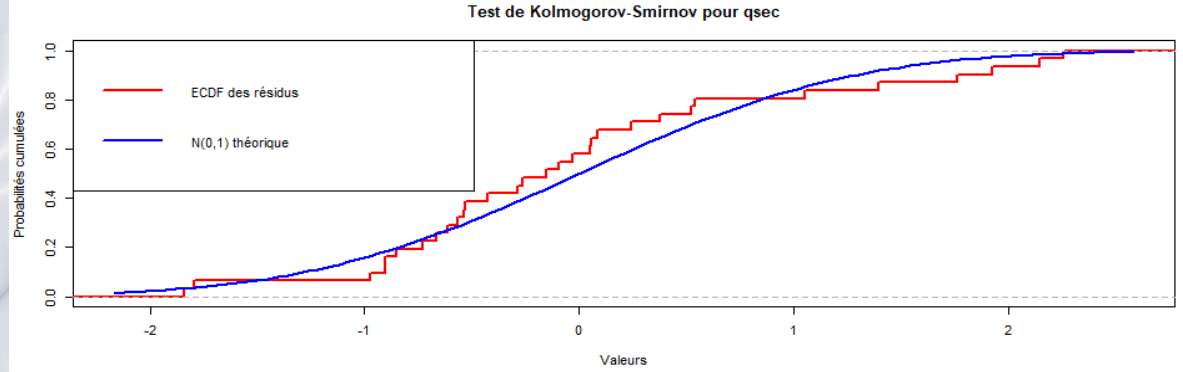
avec h_i qui est le i-ème terme diagonal de la matrice $\mathbb{X}.(\mathbb{X}'.\mathbb{X})^{-1}.\mathbb{X}'$ où :

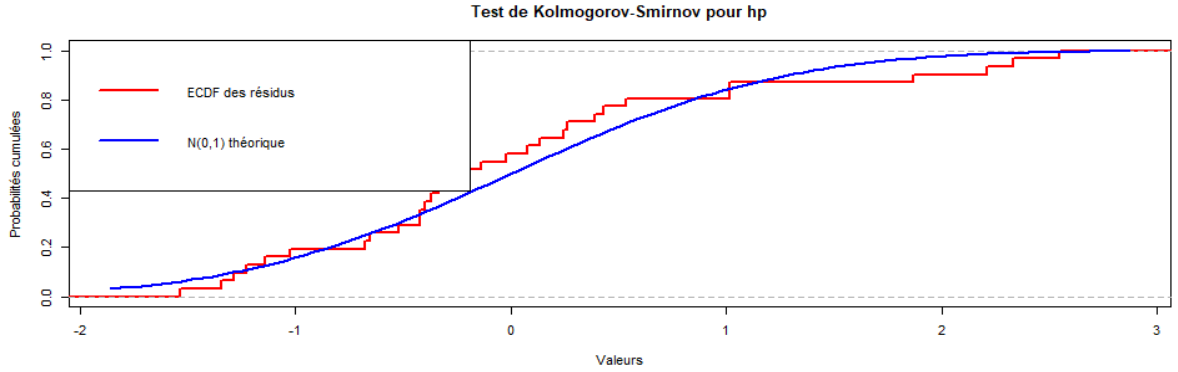
$$\mathbb{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

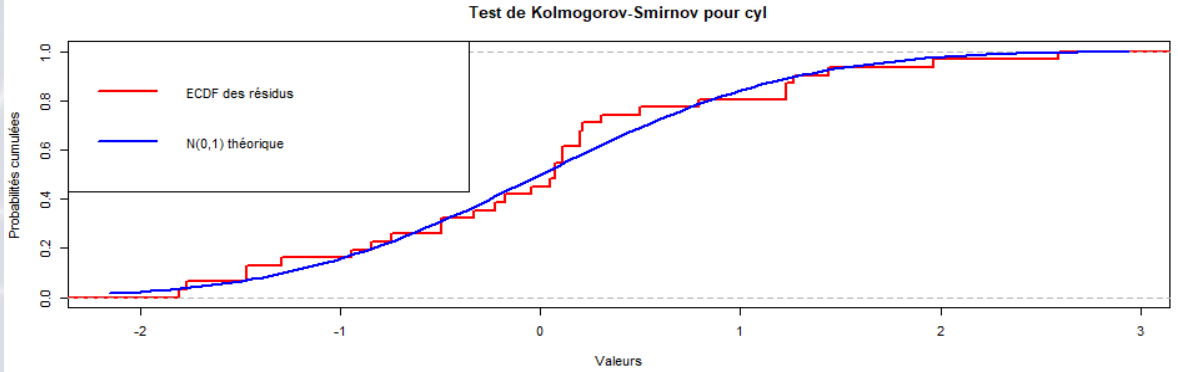
$$\hat{\varepsilon}_{i,sd} \underset{(\mathcal{L})}{\rightsquigarrow} \mathcal{N}(0,1)$$

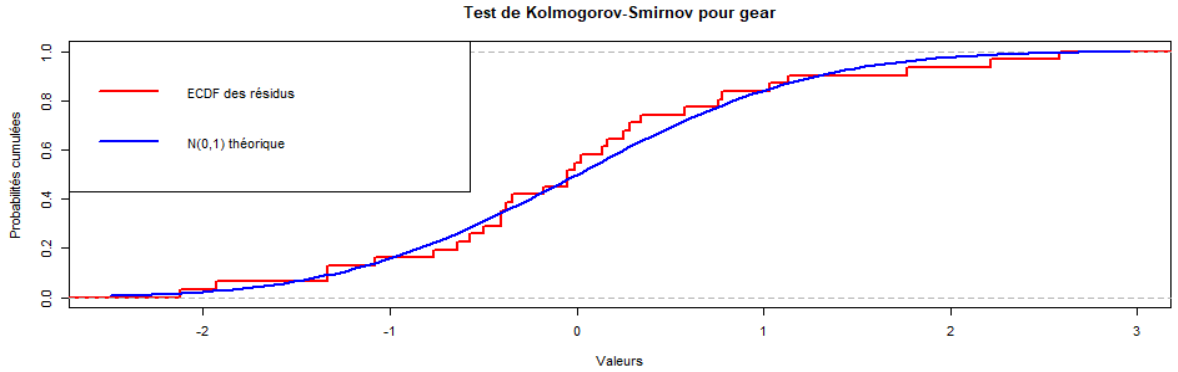


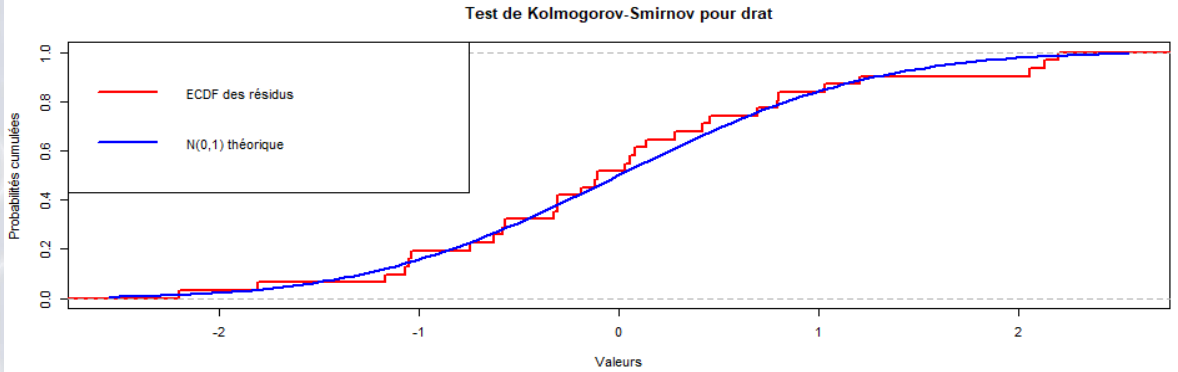
Kolmogorov-Smirnov

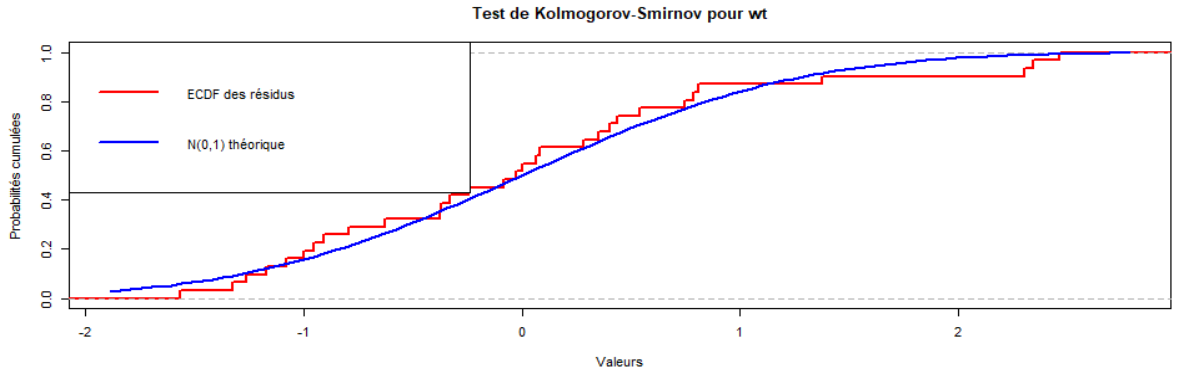








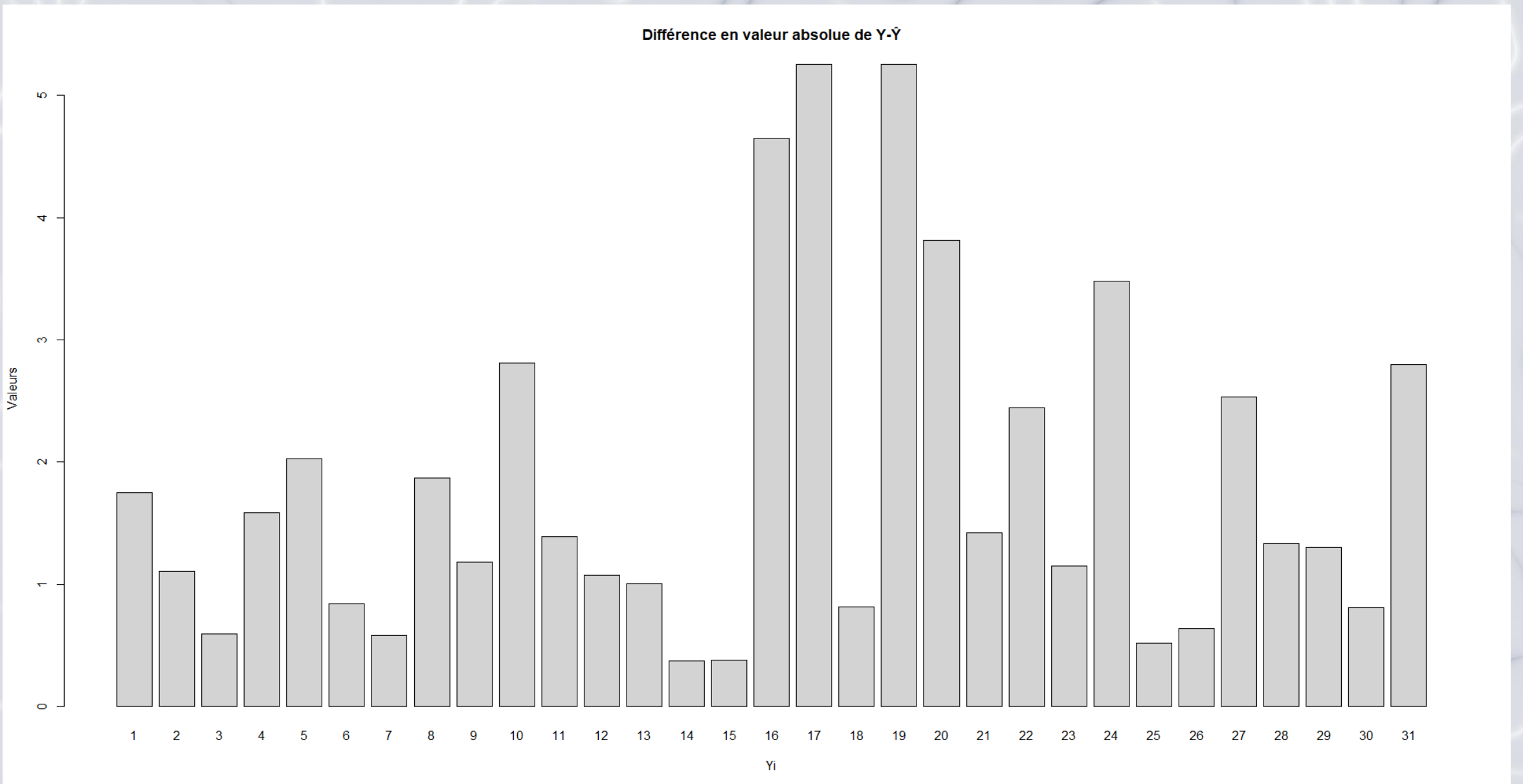




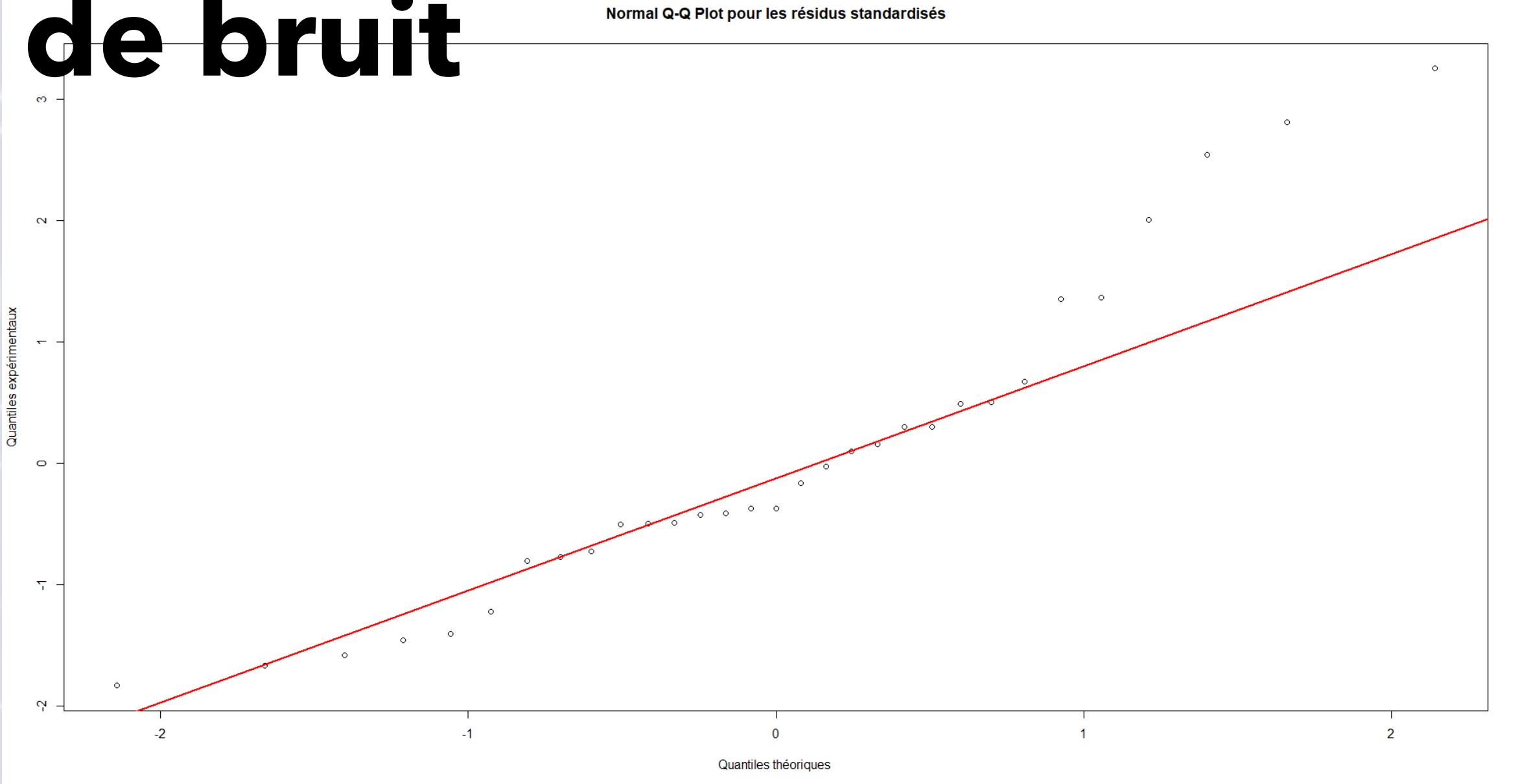
Résultats et interprétation : multiple

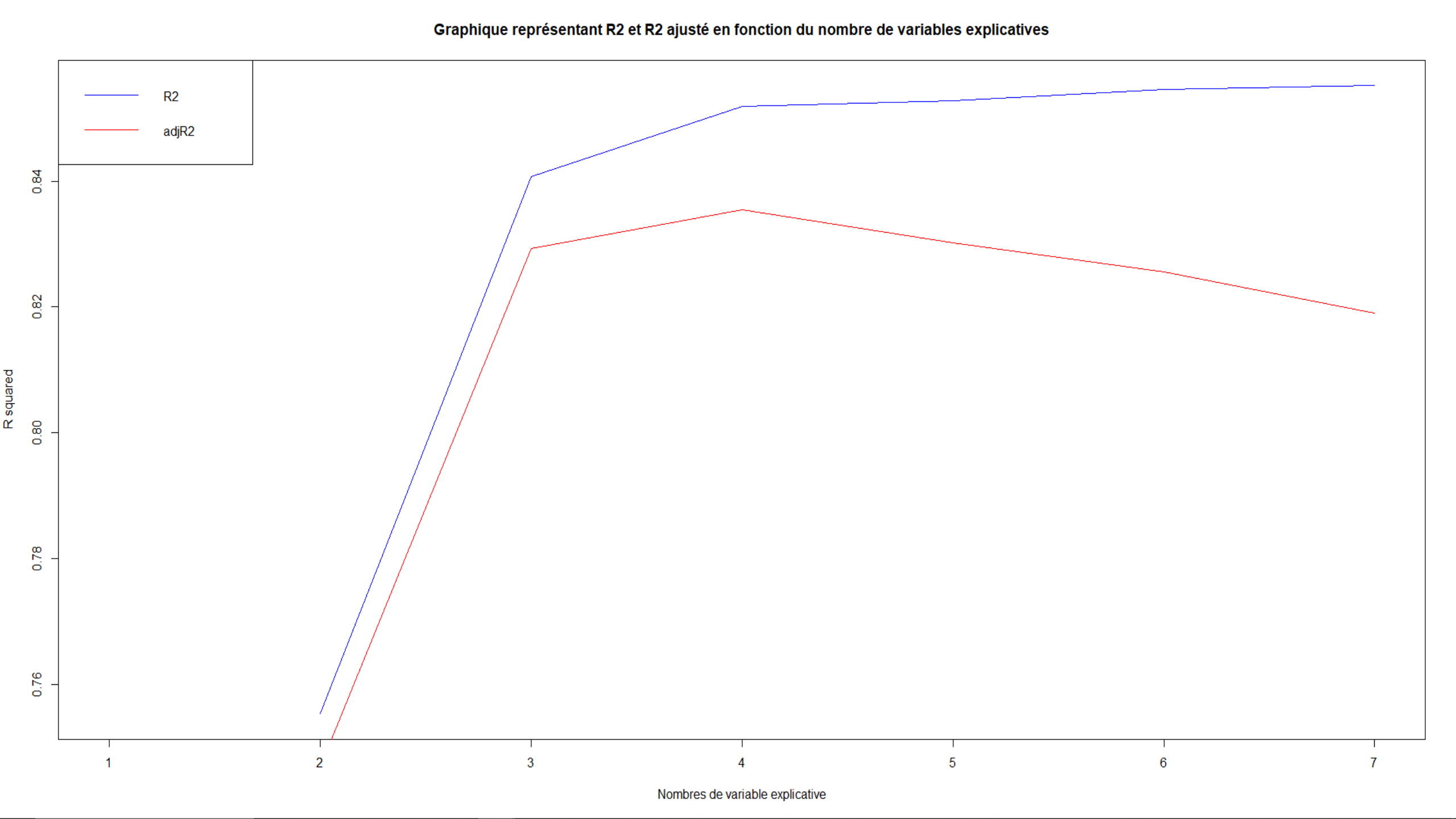
- Calcul de l'estimateur de Y et vérification
- Vérification de l'hypothèse de bruit
- Sélection de variables

Estimateur Ŷ



Vérification de l'hypothèse de bruit




Graphique représentant R2 et R2 ajusté en fonction du nombre de variables explicatives
R2
adjR2
R squared
0.76
0.78
0.80
0.82
0.84
1
2
3
4
5
6
7
Nombres de variable explicative

Sélection de variable

leaps(XmatR,Y,method=c("adjr2"))

```
$adjr2
 [1] 0.7468515 0.7277089 0.5868283 0.4423163 0.2008664 0.1436442 0.8292735 0.8194037 0.8192713 0.7459503 0.7390146
[12] 0.7346708 0.7318932 0.7315151 0.7305607 0.7225687 0.8353932 0.8314675 0.8264371 0.8236717 0.8232604 0.8231137
[23] 0.8220467 0.8220277 0.8199197 0.7623246 0.8301654 0.8300895 0.8293728 0.8270361 0.8267107 0.8262427 0.8250043
[34] 0.8201593 0.8190102 0.8182333 0.8255074 0.8247047 0.8236582 0.8234100 0.8193299 0.7517068 0.8189761
```

```
 2 FALSE FALSE  TRUE FALSE  TRUE FALSE
 2 FALSE FALSE  TRUE  TRUE FALSE FALSE
 3 FALSE  TRUE  TRUE FALSE FALSE  TRUE
 3  TRUE FALSE  TRUE FALSE FALSE  TRUE
 3 FALSE FALSE  TRUE  TRUE FALSE  TRUE
 3  TRUE FALSE FALSE FALSE  TRUE  TRUE
 3 FALSE  TRUE FALSE FALSE  TRUE  TRUE
 3 FALSE FALSE  TRUE FALSE  TRUE  TRUE
 3 FALSE  TRUE FALSE  TRUE FALSE  TRUE
 3  TRUE  TRUE FALSE FALSE FALSE  TRUE
 3  TRUE FALSE FALSE  TRUE FALSE  TRUE
 3  TRUE  TRUE  TRUE FALSE FALSE FALSE
 4 FALSE  TRUE  TRUE FALSE  TRUE  TRUE
 4  TRUE  TRUE  TRUE FALSE FALSE  TRUE
 4 FALSE  TRUE  TRUE  TRUE FALSE  TRUE
```

- Horse-power (hp)
- Cylinder (cyl)
- Weight (wt)

Conclusion

Ce projet nous a permis de :

- Découvrir le modèle de régression linéaire.
- Consolider les fondements mathématiques en probabilité.
- Appliquer ces outils sur un jeu de données réel.
- Apprendre à construire des modèles utiles pour mieux comprendre des données.