Soutenance de projet

Régression linéaire

NIO Katell PROUZET Charlotte

Sommaire

I. Introduction

Data frame mtcars :

- 32 lignes → 32 modèles de voitures
- 11 colonnes → 11 variables techniques

Nom de la colonne de mtcars	vs	disp	qsec	wt	hp	drat	carb
Signification	Type de moteur (0 en V, 1 en ligne)	Cylindrée	Temps au ¼ mile	Poids de la voiture	Puissance	Rapport de pont arrière	Nb de carburateurs
Unité	Catégorielle	pouces cubes	sec	1000 livres	chevaux		

II - Modèle de régression linéaire simple

Objectifs

Mathématiquement, le modèle s'écrit :

$$Y = aX + b + \varepsilon$$

où :

- Y est la variable à prédire,
- X est la variable explicative,
- a est le coefficient directeur (la pente),
- b est l'ordonnée à l'origine (interception),
- ε représente le bruit.

> Par la méthode des moindres carrés :

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

Validité de notre modèle

1) Le coefficient de détermination R^2

$$0 \leqslant R^2 \leqslant 1$$

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

> coefficient de corrélation de Pearson r

$$R^2 = r^2$$

> fonction **cor()** en R

Indice	1	2	3	4	5	6	7
R^2	0	0.717	0.194	0.756	0.600	0.460	0.300

Tableau répertoriant les valeurs de R^2 pour chaque colonne de X

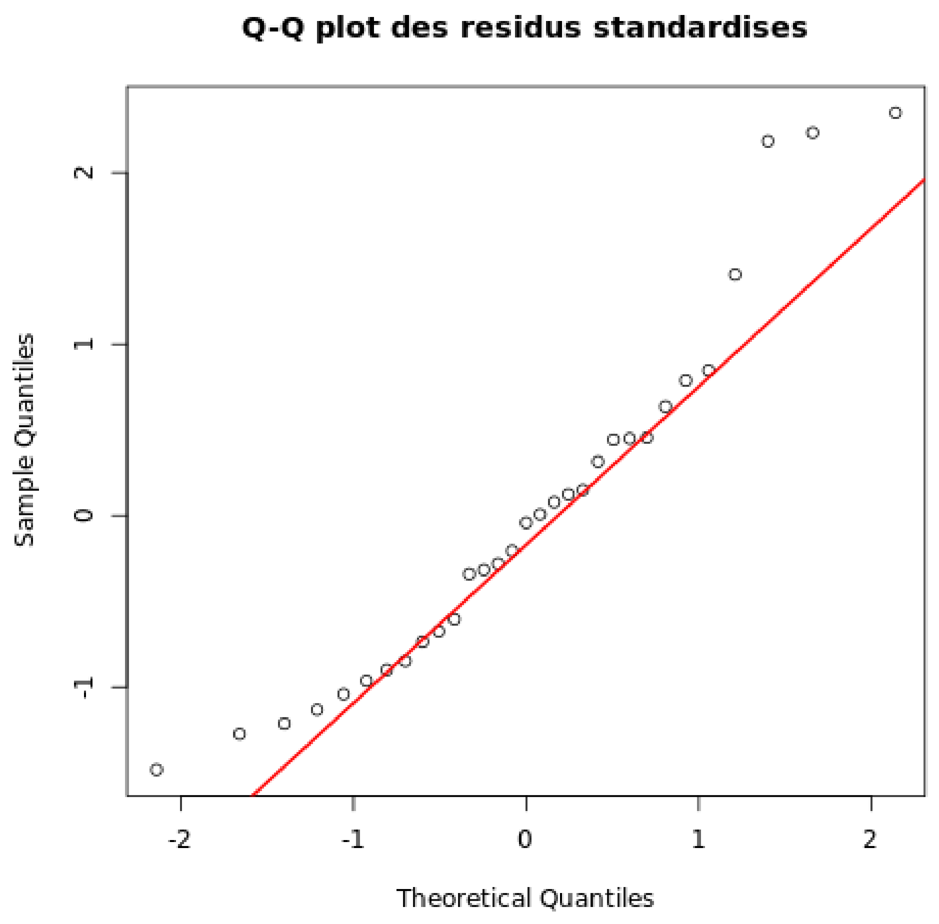
Validité de notre modèle

2) Test du paramètre a

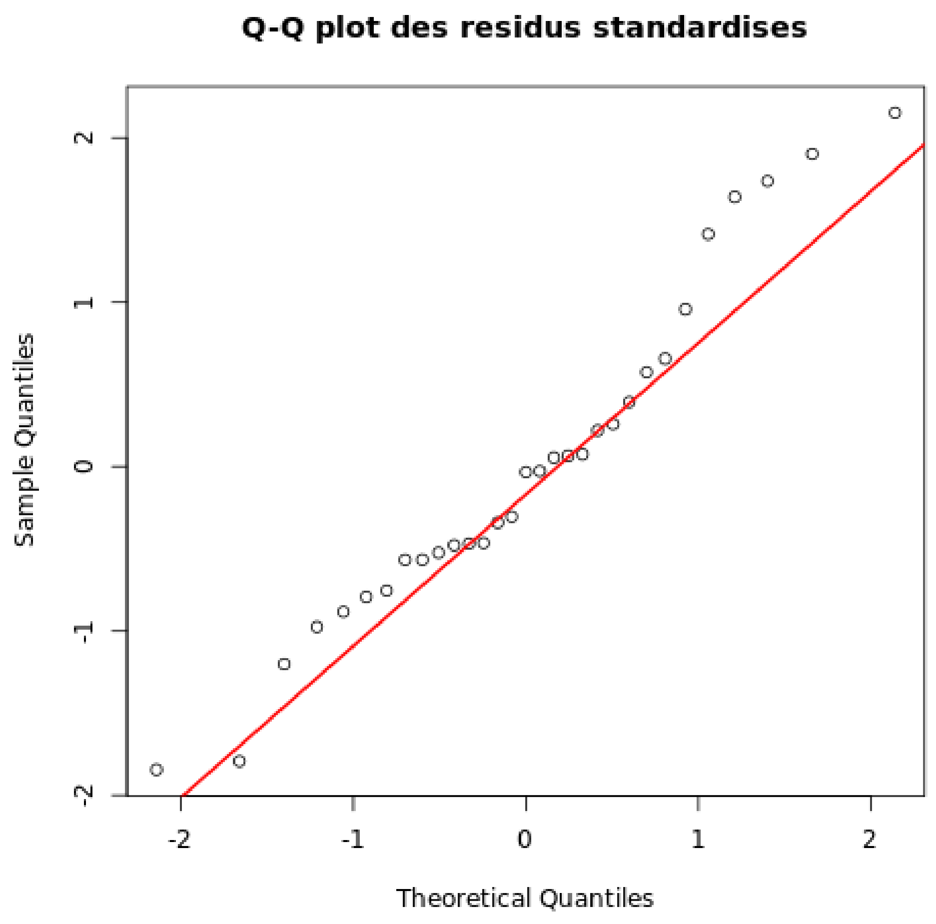
>> On vérifie en amont l'**hypothèse de gaussianité du bruit**

a) Résidus standardisés

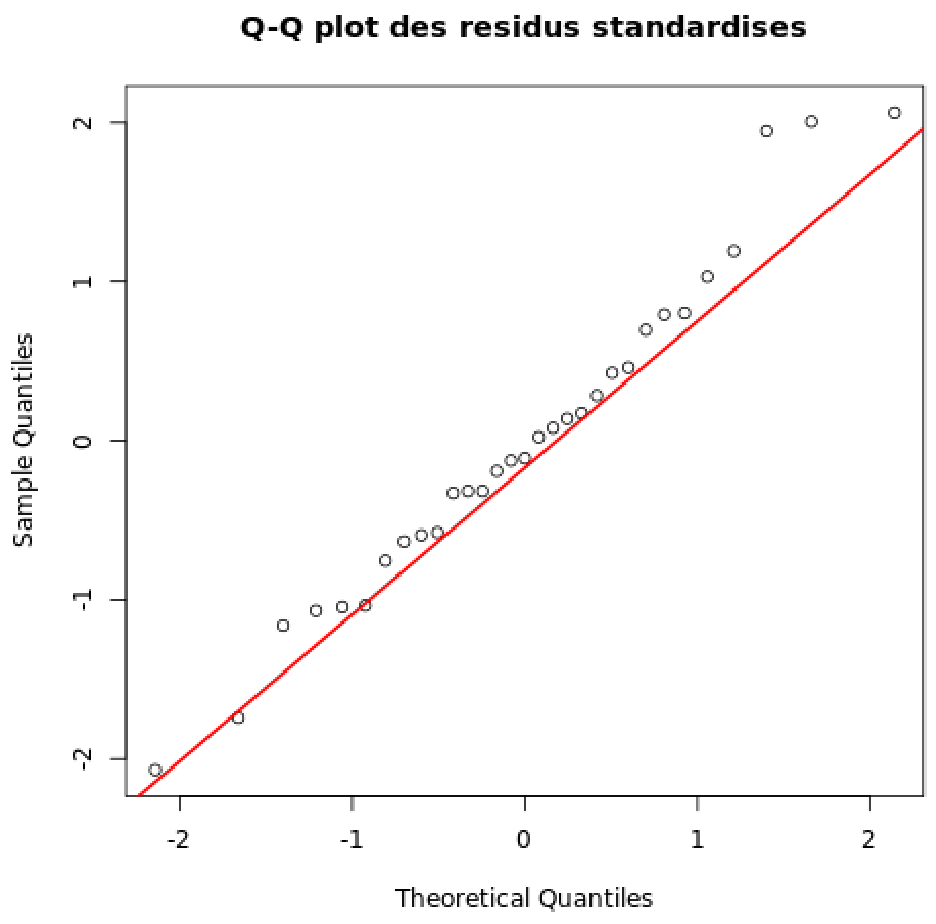
> Loi normale standard



Q-Q plot pour X[,4]



Q-Q plot pour X[,3]



Q-Q plot pour X[,6]

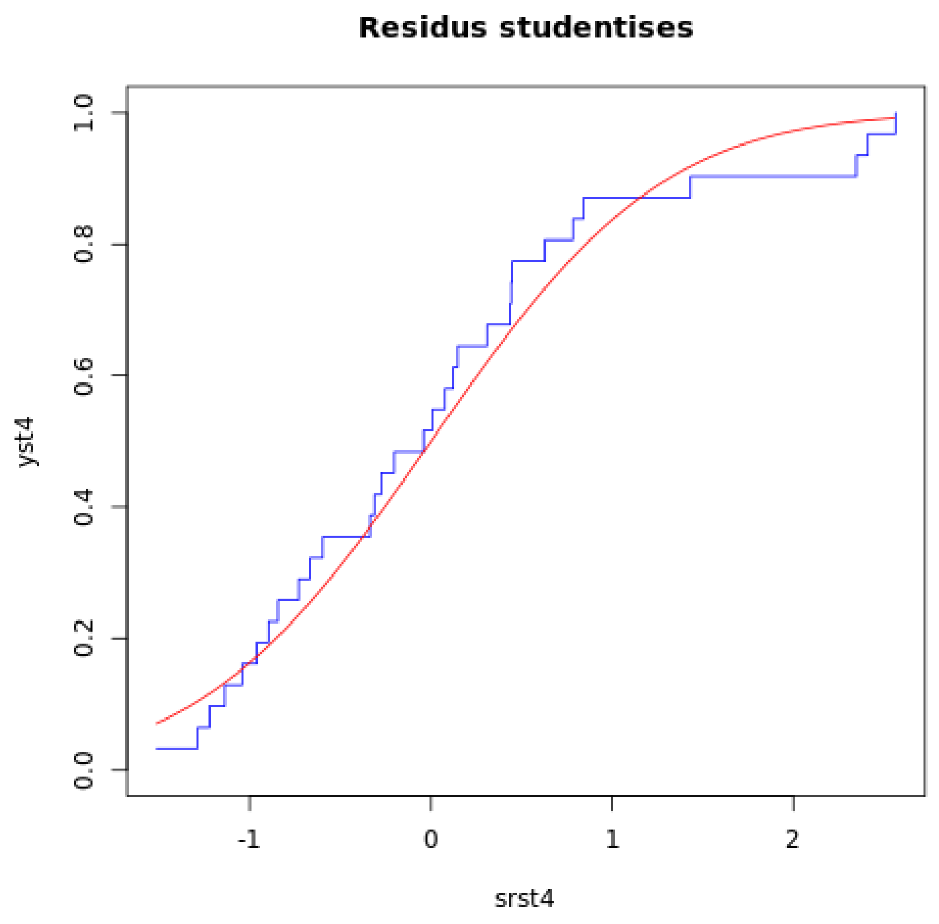
Validité de notre modèle

2) Test du paramètre a

>> On vérifie en amont l'**hypothèse de gaussianité du bruit**

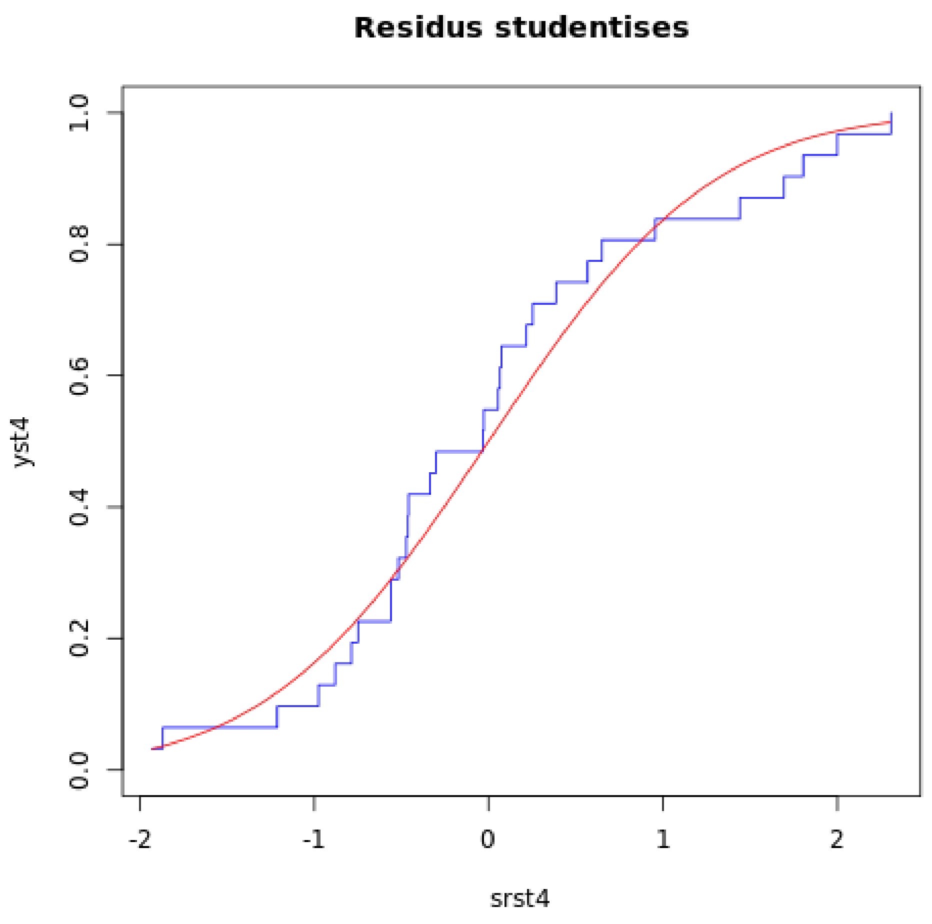
a) Résidus studentisés

> Loi de Student à (n-3) degré de liberté



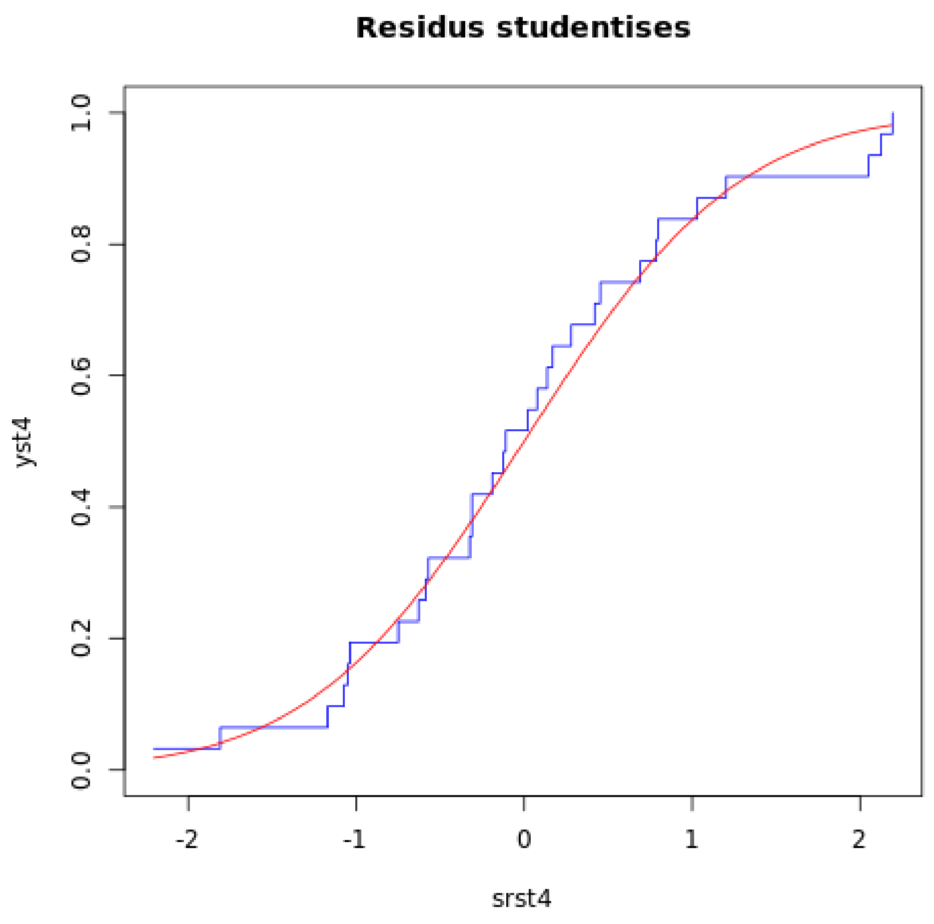
Plot pour X[,4]

p-valeur : 0.8667



Plot pour X[,3]

p-valeur : 0.7544



Plot pour X[,6]

p-valeur : 0.9835

Validité de notre modèle

2) Test du paramètre a

>> l'hypothèse de gaussianité du bruit est vérifiée

> Hypothèse du test :

- H0 : a=0 (pas d'effet de X sur Y)
- H1 : a≠0

On utilise la statistique : $T_a = \dfrac{\widehat{a}_n}{\widehat{\sigma}_n \cdot \sqrt{\dfrac{1}{\sum(x_i - \bar{x}_n)^2}}}$

avec sous H0 :

$$T_a \sim \mathcal{T}(n-2)$$

> Proposition :

- On rejette H0 au seuil alpha si : $|T_a| > t_{1-\alpha/2,\, n-2}$

> Règle de décision :

- Si p-valeur < alpha : on décide H1
- Si p-valeur > alpha : on ne rejette pas H0

> Résultats :

Indice	X[,4]	X[,3]	X[,6]
p-valeur	1.078e-10	0.00653	1.362e-05

Tableau répertoriant les p-valeur obtenues pour différentes colonnes de X

> Décision : **On décide H1**

Prédiction

> Une prévision de la variable réponse Y est donnée par :

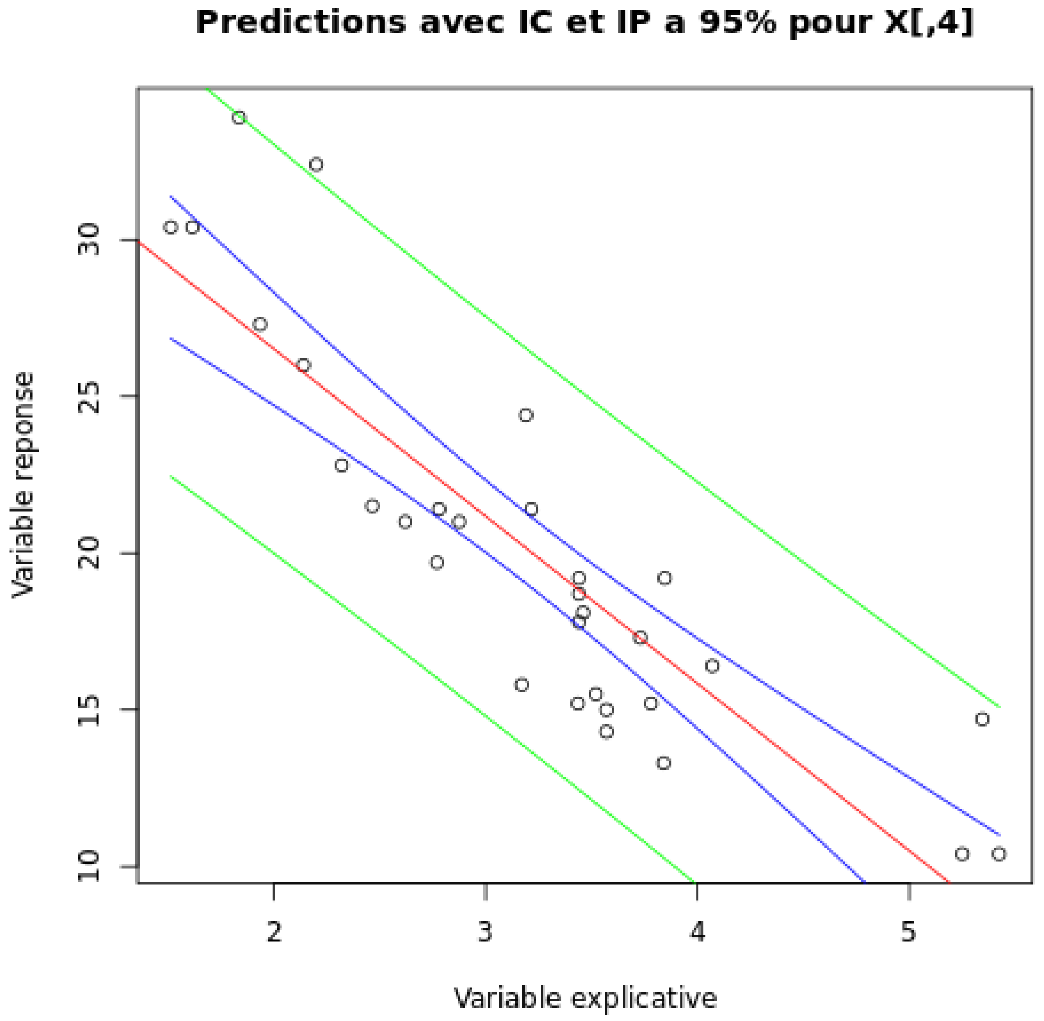
$$\widehat{y}_{\text{new}} = \widehat{a}_n x_{\text{new}} + \widehat{b}_n$$

> Utilisation de la fonction **predict()** en R

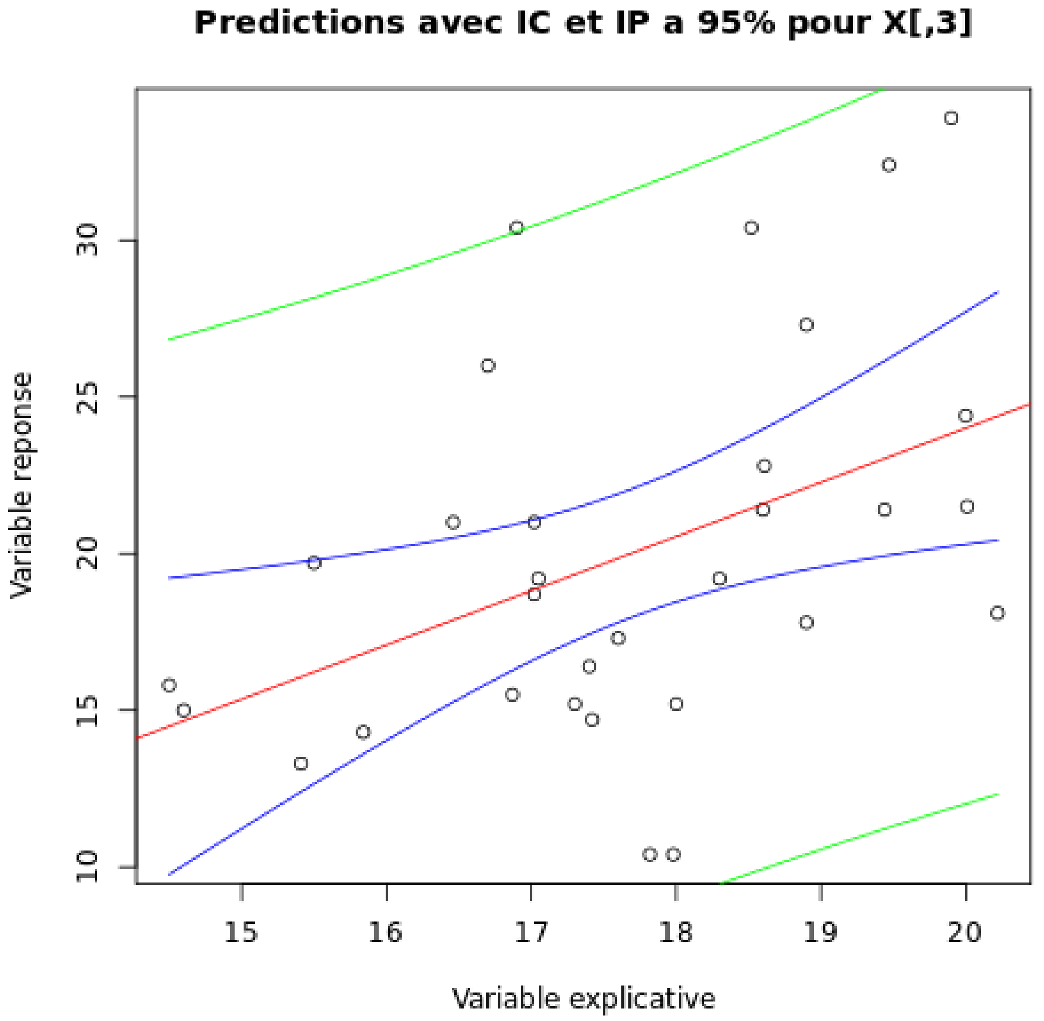
> Graphes obtenus :

IC : $\widehat{y}_{\text{new}} \pm \widehat{\sigma}_n \cdot \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{\sum(x_i - \bar{x}_n)^2}} \cdot t_{1-\alpha/2,\, n-2}.$

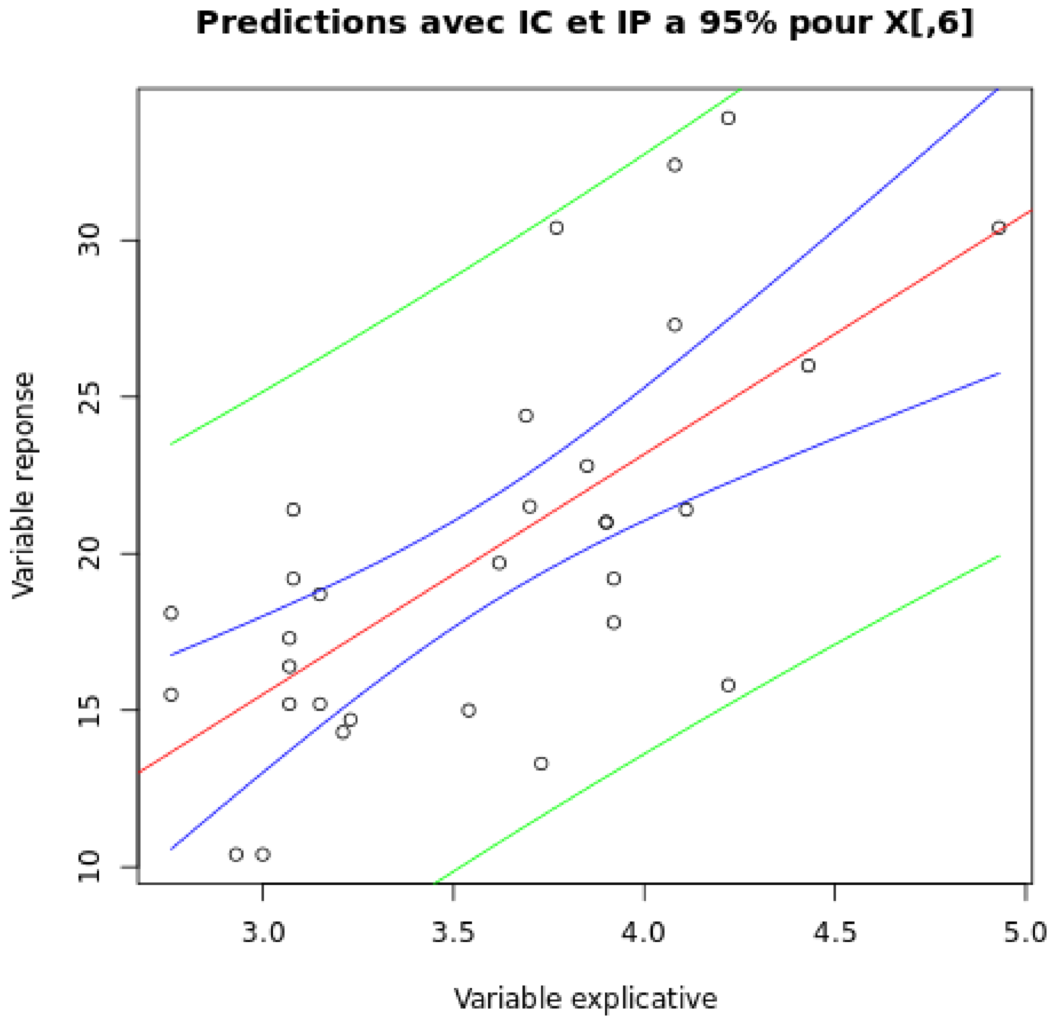
IP : $\widehat{y}_{\text{new}} \pm \widehat{\sigma}_n \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{\sum(x_i - \bar{x}_n)^2}} \cdot t_{1-\alpha/2,\, n-2}.$



Prédiction avec IC et IP à 95% pour X[,4]



Prédiction avec IC et IP à 95% pour X[,3]



Prédiction avec IC et IP à 95% pour X[,6]

III - Modèle de régression linéaire multiple

Objectif

Principe :

- 1 variable réponse quantitative : Y
- p variables explicatives qui sont quantitatives : X (matrice)

Objectif :

Modéliser au mieux une variable dépendante Y à partir de plusieurs variables explicatives X1, X2, ..., Xp.

Modèle Mathématique :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad \Leftrightarrow \qquad Y = X\beta + \varepsilon$$

On détermine bêta chapeau avec la méthode des moindres carrés :

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Validité de notre modèle

1) X'X inversible :

- det(X) ≠ 0
- rang(X) = nb de colonnes de X

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

$$\downarrow$$

$$\hat{Y} = X\hat{\beta}$$

2) R² ajusté :

- R^2 ajusté < 1
- R^2 ajusté < R^2

$$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

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

3) Tests sur les coefs bêta

- Les résidus standardisés doivent suivre une loi normale centrée réduite $\hat{\varepsilon}_{i,sd} \underset{(\mathcal{L})}{\rightsquigarrow} \mathcal{N}(0,1)$
- Les résidus studentisés doivent suivre une loi de Student à n - rang(X) -1 degrés de liberté. $\hat{\varepsilon}_{i,st} \sim T(n - rang(\mathbb{X}) - 1)$

Validité de notre modèle

2) R^2 ajusté :

Sélection de variable : a) Méthode pas à pas

X	vs	disp	qsec	wt	hp	drat	carb

On retire les variables qualitatives : vs et carb

X1	disp	qsec	wt	hp	drat

R^2 ajusté : 0.823

X2	disp	qsec	wt	drat

R^2 ajusté : 0.827

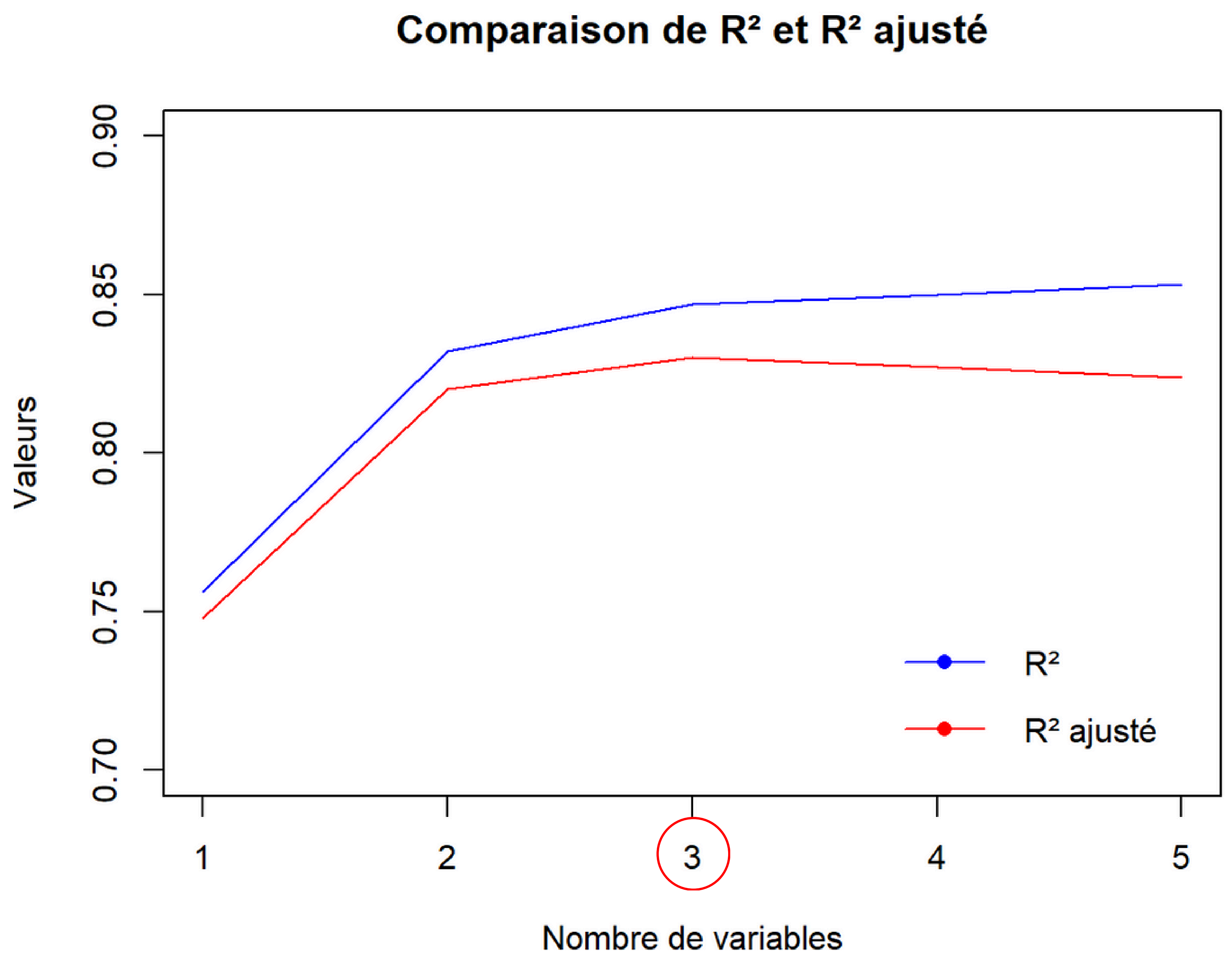
X3	qsec	wt	drat

R^2 ajusté : 0.830

Validité de notre modèle

2) R^2 ajusté :

Sélection de variable : b) Méthode exhaustive (comparaison R^2 et R^2 ajusté)



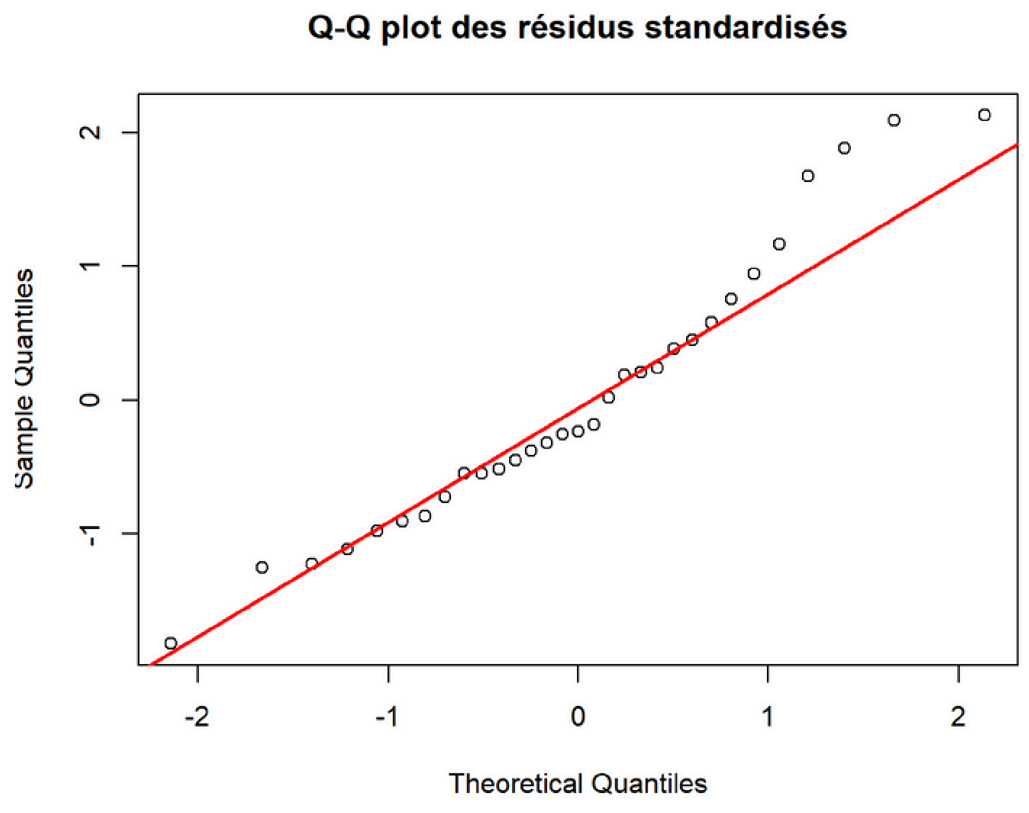
Validité de notre modèle

3) Tests :

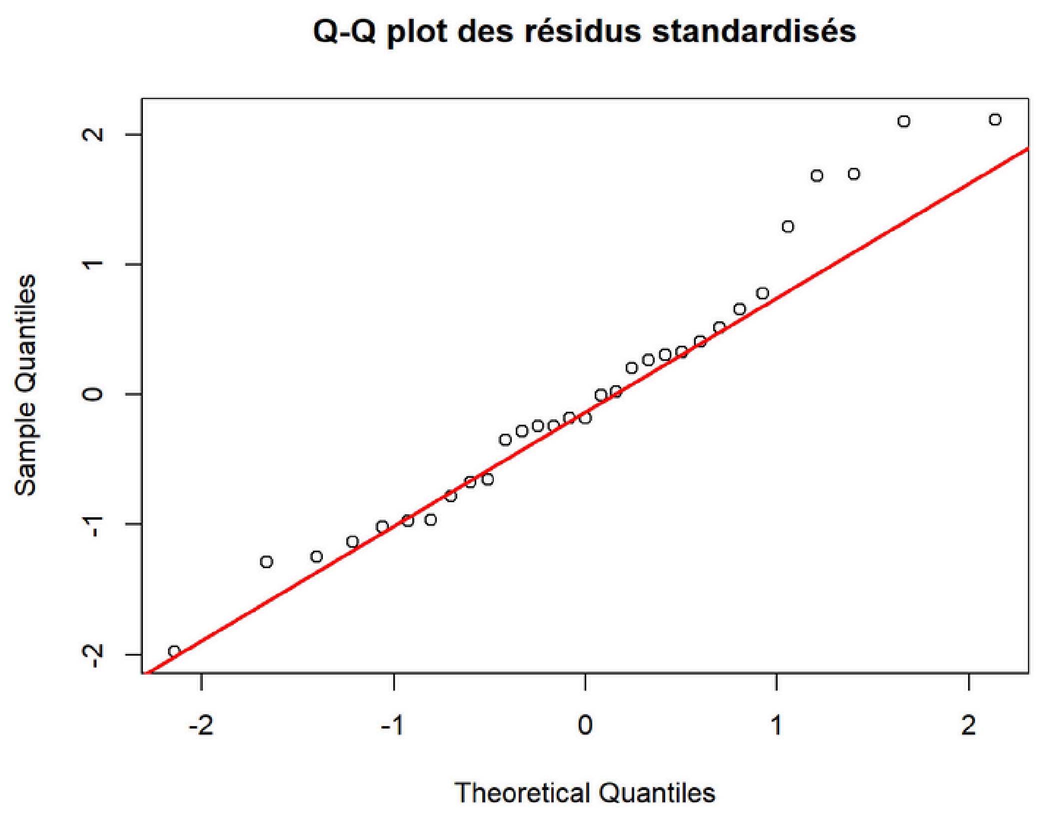
>> On vérifie en amont l'**hypothèse de gaussianité du bruit**

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

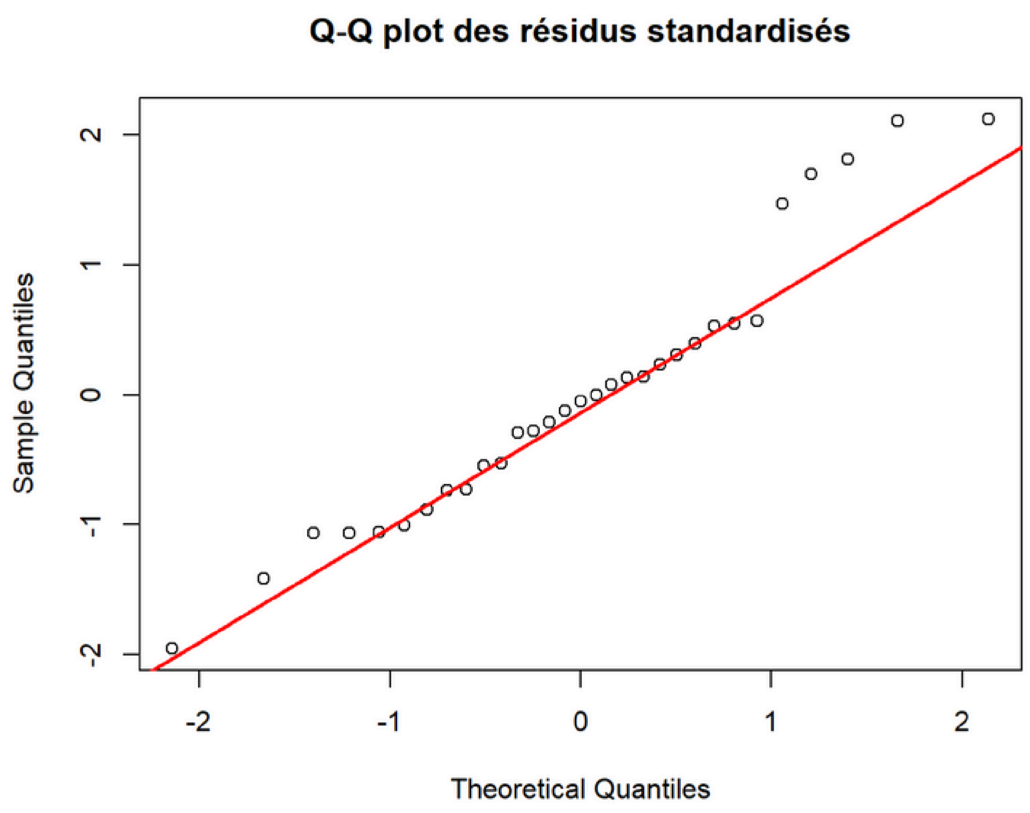
X2



X3



X4



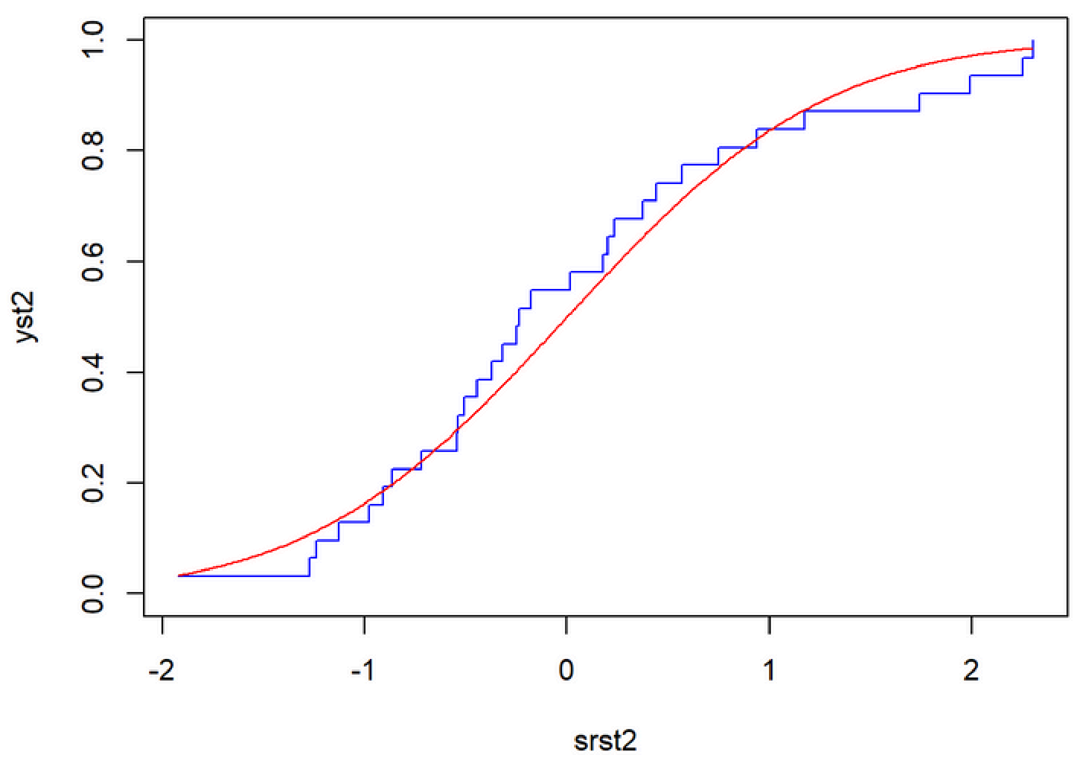
Validité de notre modèle

3) Tests :

>> On vérifie en amont l'**hypothèse de gaussianité du bruit**

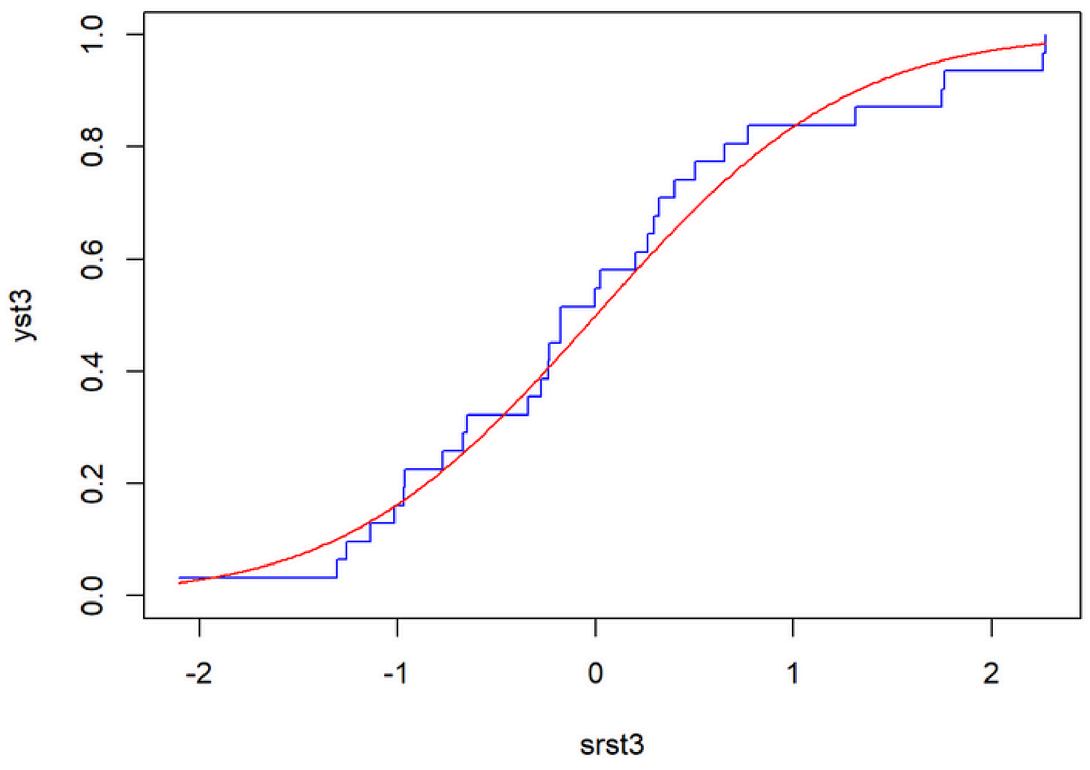
b) Résidus studentisés $\hat{\varepsilon}_{i,st} \sim T(n - rang(\mathbb{X}) - 1)$

X2



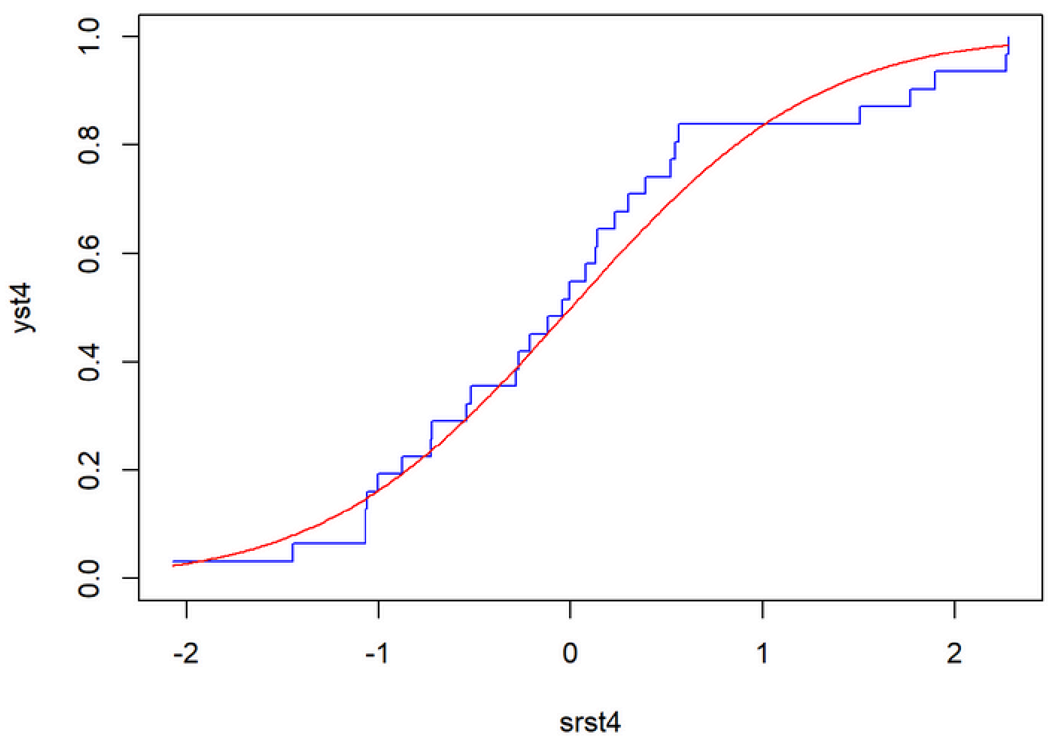
p-valeur : 0.7283 >> 0.05

X3



p-valeur : 0.9485 >> 0.05

X4



p-valeur : 0.6424 >> 0.05

Validité de notre modèle

3) Tests :

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (aucune variable explicative n'a d'effet),
- $H_1 : \exists j \in \{1, \ldots, p\}$ tel que $\beta_j \neq 0$.

On utilise la statistique de test de Fisher suivante :

$$F = \frac{\|\widehat{Y} - \bar{Y}\|^2/(rang(X) - 1)}{\|Y - \widehat{Y}\|^2/(n - rang(X))}$$

Sous H_0, la statistique F suit une loi de Fisher $\mathcal{F}(rang(X)-1, n-rang(X))$. On rejette H_0 si la p-valeur est inférieure à $\alpha = 5\%$.

Modèle	X2	X3	X4
p-valeur	≈ 1.77e-09	≈ 2.33e-10	≈ 3.87e-11

→ On choisit H1

IV - Analyse de la variance à un facteur

Objectif

Principe :

- 1 variable réponse quantitative : Y
- 1 variable explicative qualitative : facteur

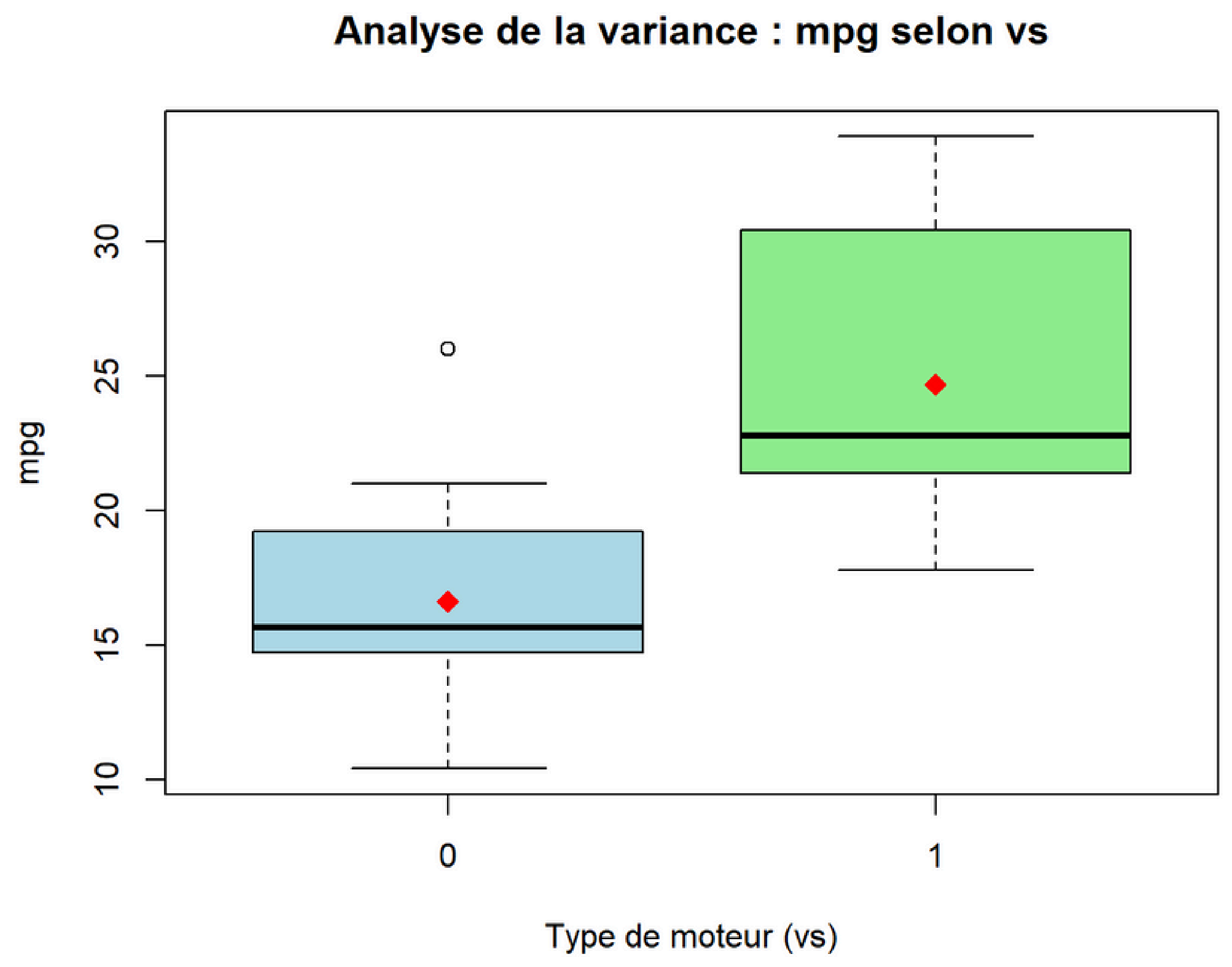
Objectif :

Comparer les moyennes de plusieurs modalités pour déterminer si elles se comportent de la même façon, ou si elles diffèrent de manière significative. Cela permet de répondre à des questions du type :

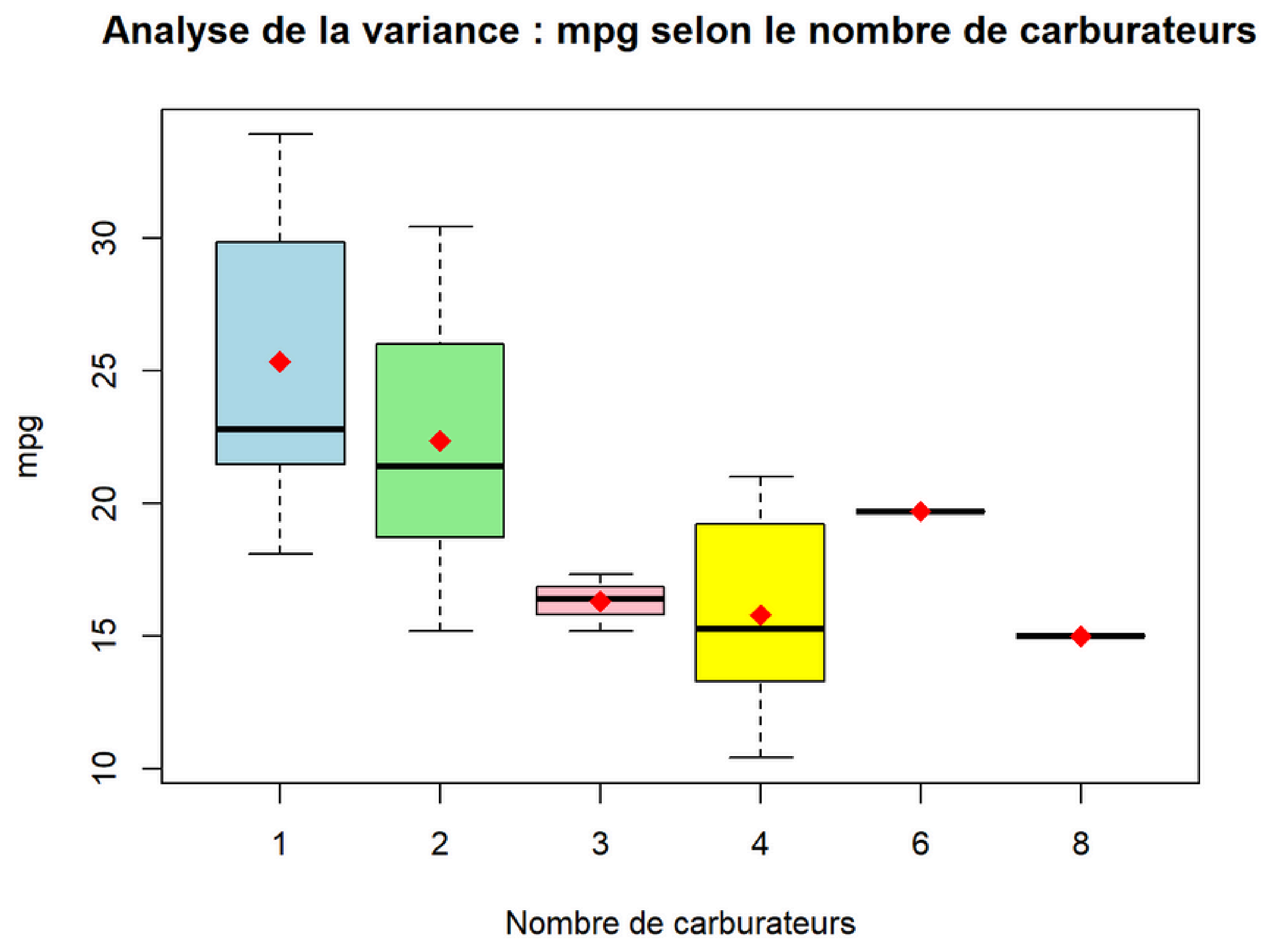
- La consommation d'essence (mpg) depend-elle du type de moteur (vs) ?
- Le nombre de carburateurs (carb) influence-t-il la consommation d'essence (mpg)?

Visualisation par boîtes à moustaches

Pour vs :



Pour carb :



Tests

1) Test sur les moyennes

On considère une variable réponse Y (quantitative) et un facteur A (qualitatif) ayant k modalités, notées $A_1, A_2, \ldots, A_k$. À chaque modalité A_j, on observe un certain nombre n_j de valeurs de Y.

Hypothèses du test :

- H_0 : les moyennes des modalités sont égales. Autrement dit, le facteur A n'a pas d'effet :

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- H_1 : au moins une moyenne est différente :

$$H_1 : \exists(i,j) \text{ tel que } \mu_i \neq \mu_j$$

Pour vs :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
vs	1	492.3	492.3	22.8	4.74e-05	***
Residuals	29	626.2	21.6			

Pour carb :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
carb	5	493.2	98.63	3.943	0.00898	**
Residuals	25	625.3	25.01			

Tests

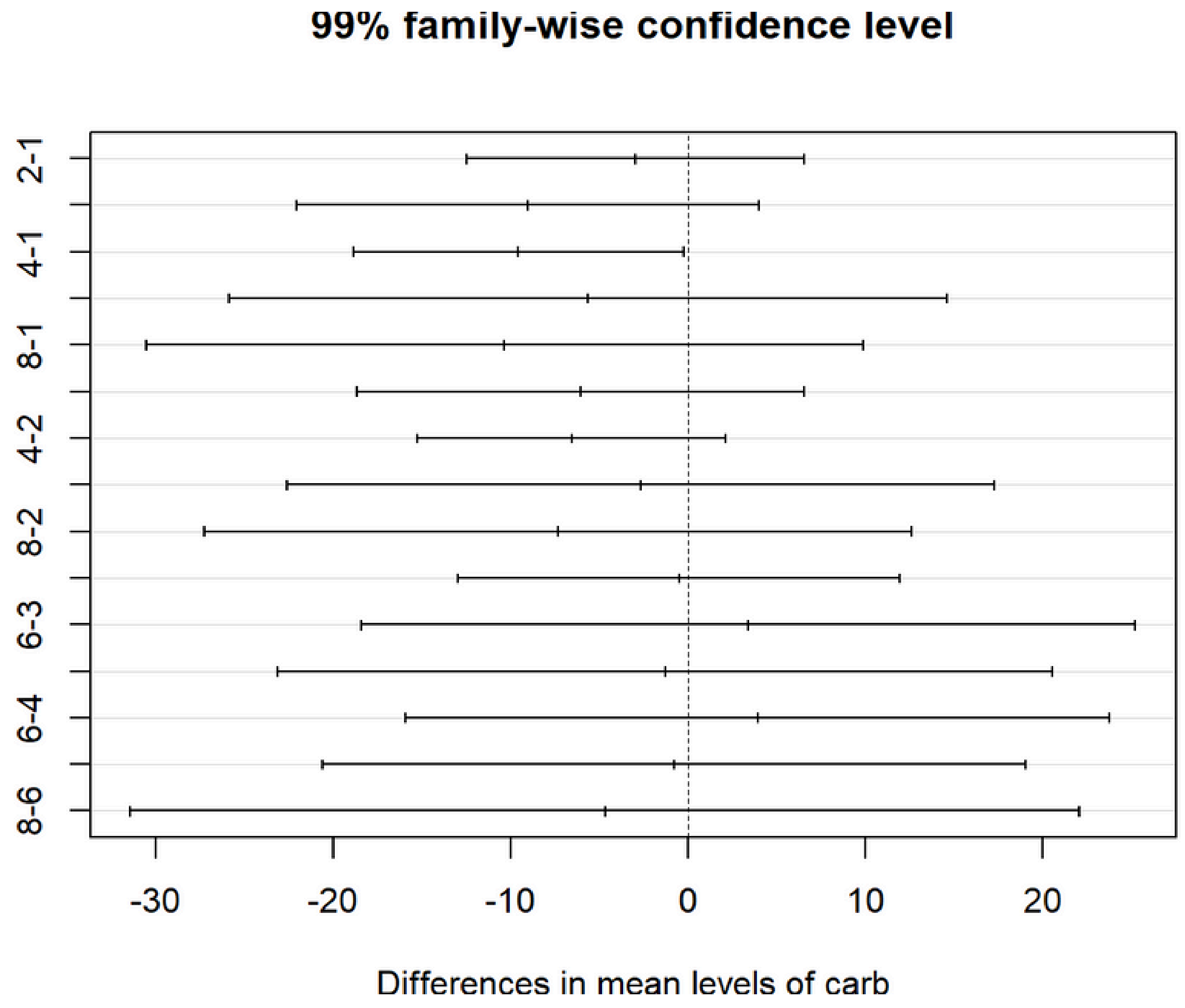
2) Test de Tukey

Objectif : Déterminer quelles modalités sont significativement différentes entre elles. Ce test permet de comparer toutes les paires de modalités deux à deux.

Interprétation des intervalles de confiance : Si l'intervalle de confiance contient la valeur 0, cela signifie qu'il n'y a pas de différence significative entre les deux modalités comparées. Les modalités peuvent être regroupées.

vs

carb

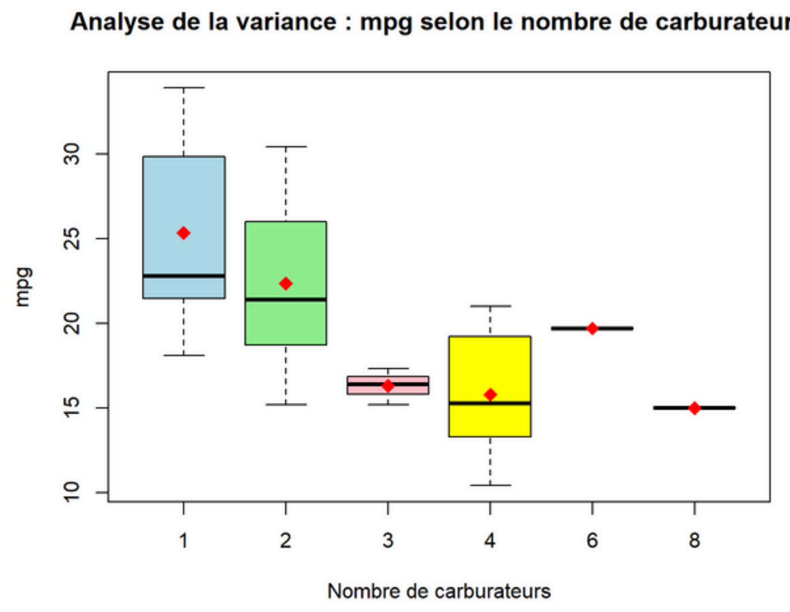


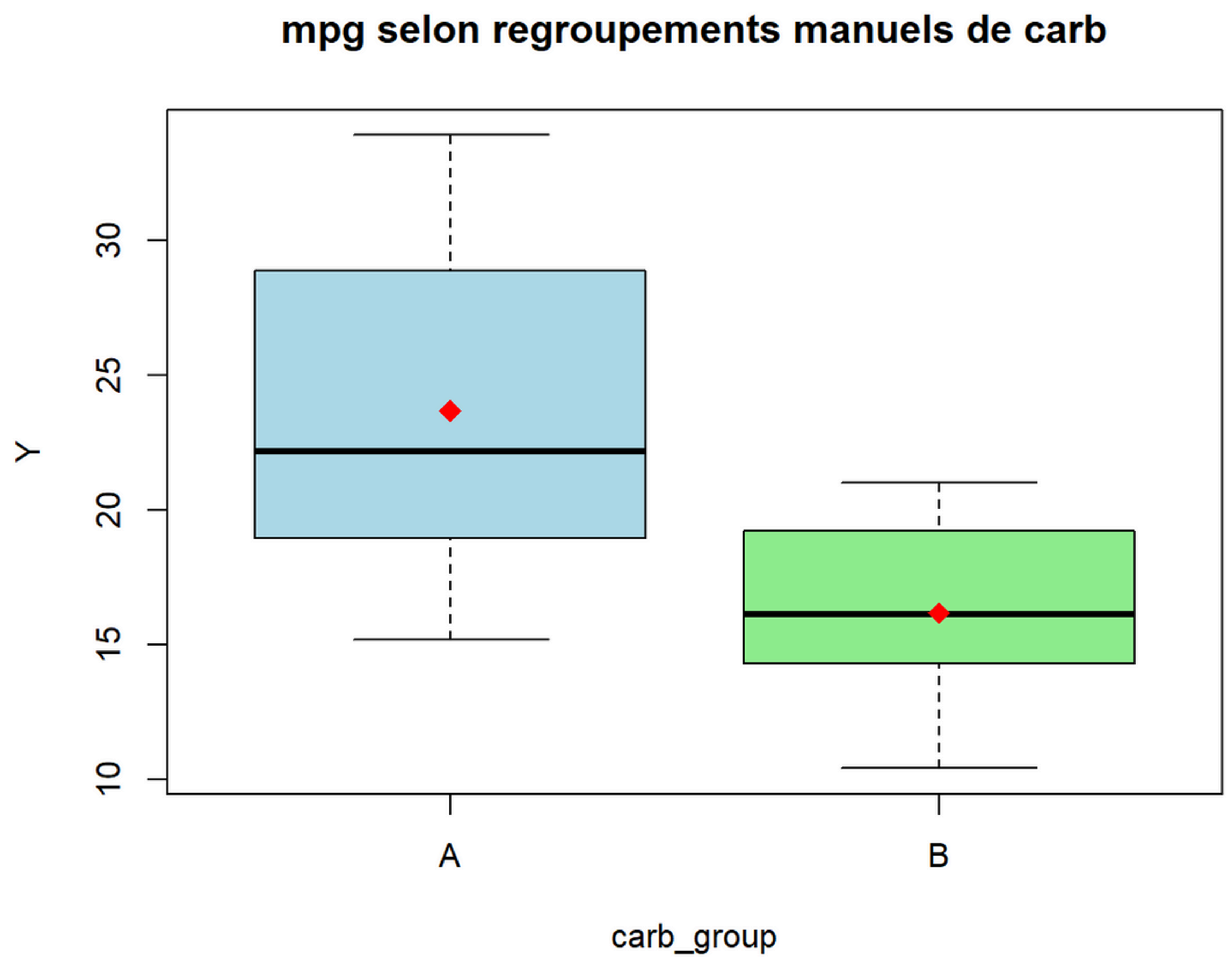
Regroupement des modalités

1) Visualisation par boîtes à moustaches

Choix des 2 groupes :

- **Groupe A** = carb 1, 2 et 6
- **Groupe B** = carb 3, 4 et 8





Regroupement des modalités

2) Tests

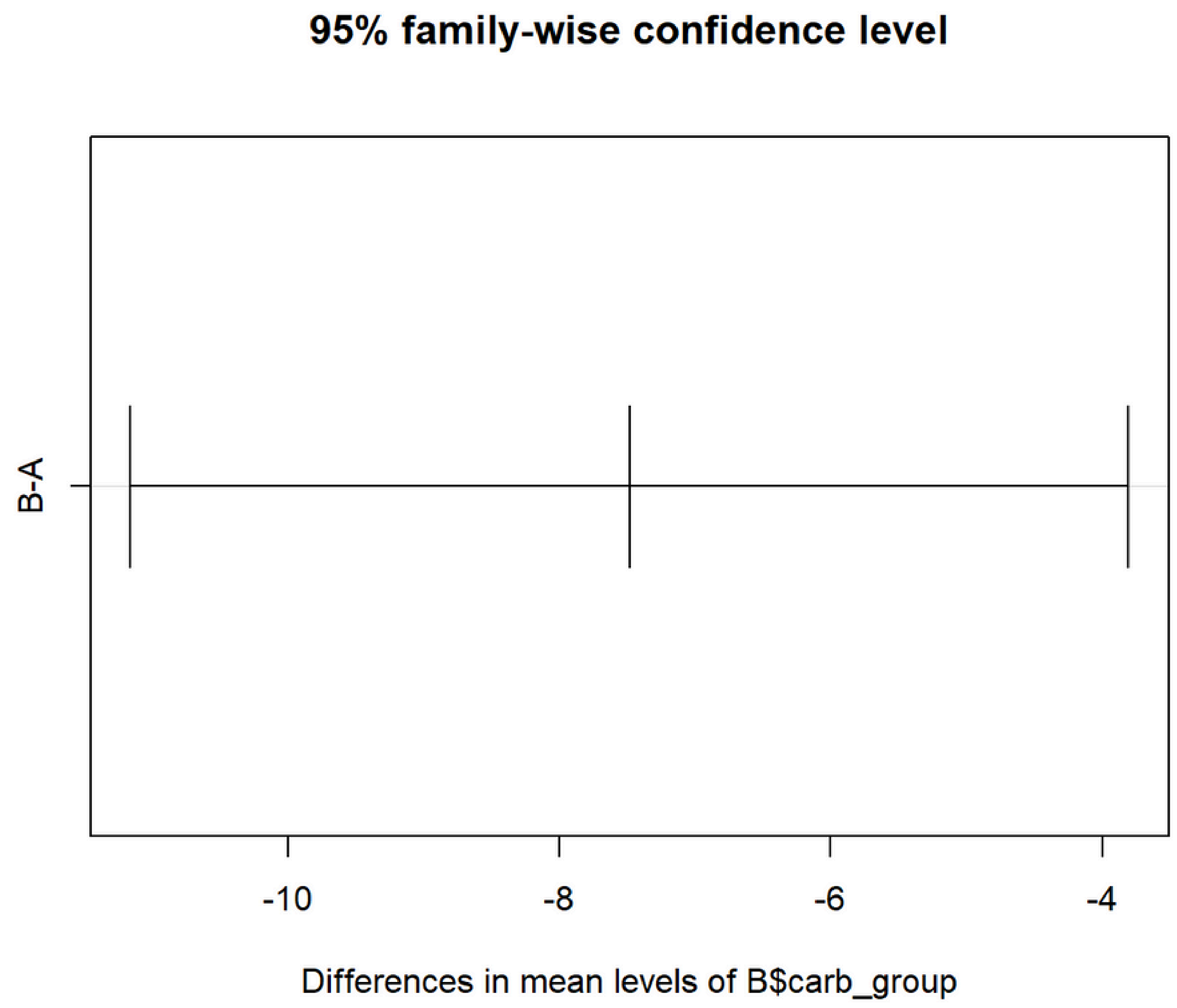
a) Test sur les moyennes

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
B$carb_group	1	418.2	418.2	17.36	0.000268	***
Residuals	28	674.4	24.1			

p-valeur = 2.68e-05 << 0.05

→ on choisit H1

b) Test de Tukey



Conclusion

Merci pour votre attention !