MOOC Statistique pour ingénieur
Thème 4 : Régression linéaire
Vidéo 4 : Validation du modèle

Anca Badea

Institut Mines-Télécom
Mines Saint-Étienne

Sommaire

Sommaire

Comment valider le modèle ?

outil principal de vérification des hypothèses : analyse des résidus

$\widehat{\varepsilon}_i = y_i - \widehat{y}_i \quad / \quad \widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$

- normalité
- homoscédasticité
- indépendance
- forme linéaire de la relation entre y et x

La forme des graphiques des résidus donne des informations importantes sur la linéarité et sur la homoscédasticité.

Si l'étude des graphiques montre que les hypothèses ne sont pas validées, alors la partie concernant l'estimation du modèle, les tests, les IC, ... n'est pas utilisable non plus.

Quelques propriétés

- la somme des résidus est nulle

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0$$

- la somme des résidus pondérée par la valeur correspondante de la variable explicative est nulle

$$\sum_{i=1}^{n} x_i \, \widehat{\varepsilon}_i = 0$$

- la somme des résidus pondérée par la valeur correspondante de la prédiction de la variable expliquée est nulle

$$\sum_{i=1}^{n} \widehat{y}_i \, \widehat{\varepsilon}_i = 0$$

Les résidus standardisés, les résidus studentisés

on admet $\mathbb{E}(\widehat{\varepsilon}_i) = 0, \quad \mathbb{V}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$
avec h_{ii} qui dépend uniquement des valeurs de x
variance des résidus bruts n'est pas constante !!!

les résidus standardisés

$$R_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}^* \sqrt{1 - h_{ii}}}$$

$\widehat{\varepsilon}_i$ et $\widehat{\sigma}^*$ sont corrélés !!!

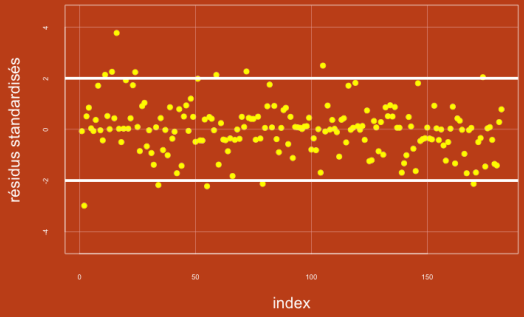
$\widehat{\sigma}^* \longrightarrow \widehat{\sigma}^*_{(i)}$

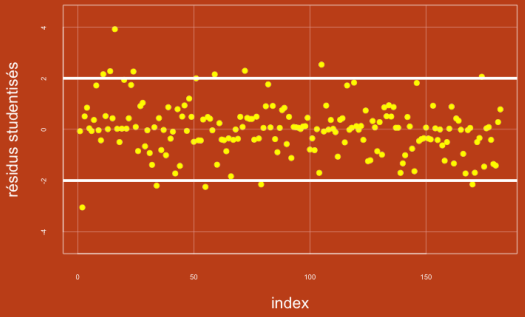
les résidus studentisés

$$T_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}^*_{(i)} \sqrt{1 - h_{ii}}} \sim \mathcal{T}(n - 3)$$

Les résidus en pratique

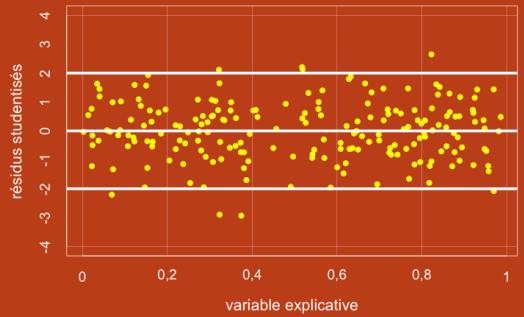
- comparer les résidus standardisés / studentisés aux bornes ± 2





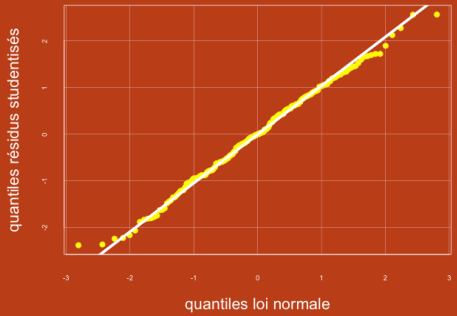
Les résidus en pratique

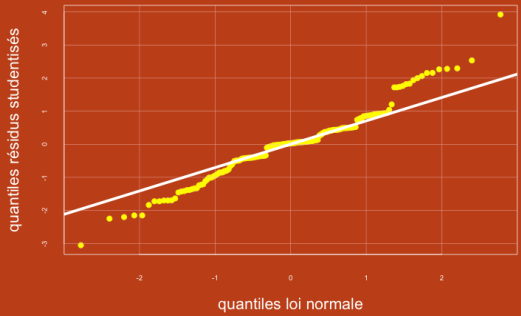
- tracer les couples (x_i, t_i) et vérifier qu'ils sont
  - *normalement* distribués autour de la droite d'équation $y = 0$
  - sans forme particulière



Les résidus en pratique

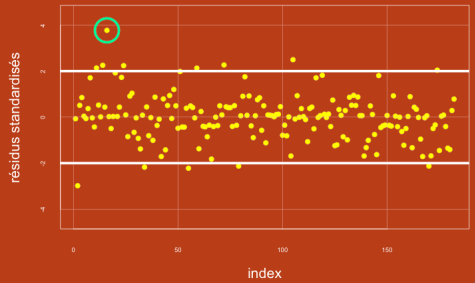
- tracer un graphique de type quantile-quantile et la droite de Henri pour vérifier la *normalité* des résidus





Les résidus en pratique

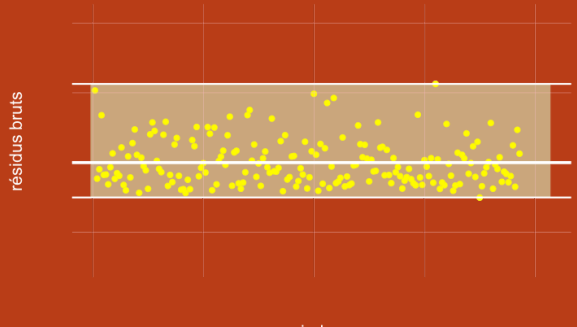
- étudier attentivement les grandes valeurs de r_i ($|r_i| > 3$) : potentiellement outliers



- r_i et t_i diffèrent peu

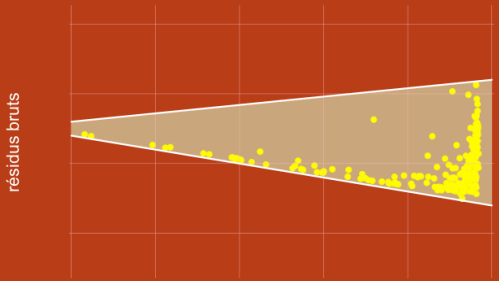
Les graphiques de base

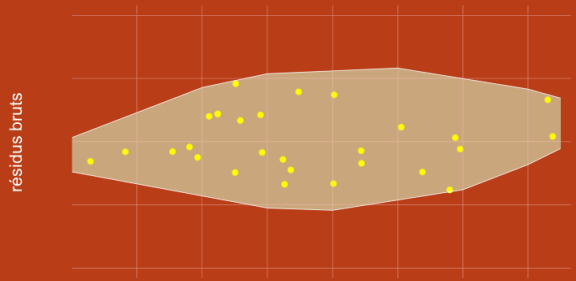
- graphique des résidus $(\widehat{\varepsilon}_i, r_i, t_i)$ de façon séquentielle
- graphique des résidus $(\widehat{\varepsilon}_i, r_i, t_i)$ en fonction de $\widehat{y}_i$ ou bien de x_i
  - si les résidus sont dans une bande horizontale et ils varient de manière aléatoire à l'intérieur, il n'y a pas de défaut évident



Les graphiques de base

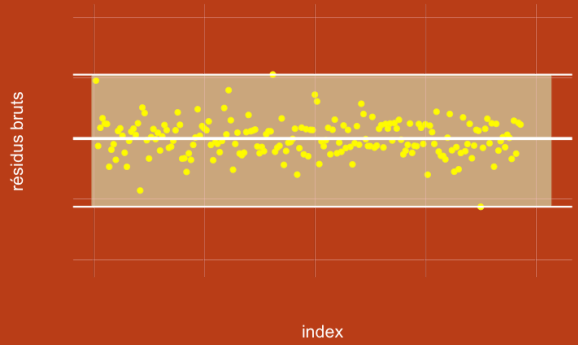
- graphique des résidus $(\widehat{\varepsilon}_i, r_i, t_i)$ en fonction de $\widehat{y}_i$ ou bien de x_i
  - si des formes particulières apparaissent
    - entonnoir ou double arc $\Longrightarrow$ la variance des erreurs pas constante





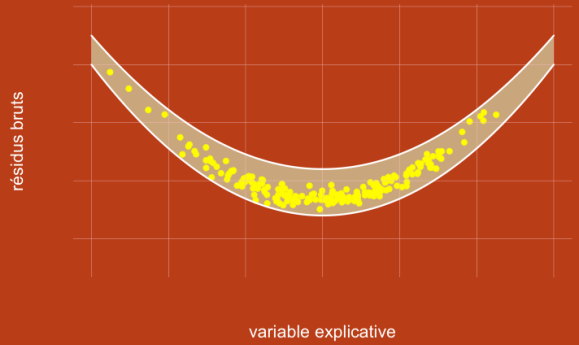
Les graphiques de base

- graphique des résidus ($\widehat{\varepsilon}_i$, r_i, t_i) en fonction de $\widehat{y}_i$ ou bien de x_i
  - si des formes particulières apparaissent
    - comment on peut réparer : transformations de la variable explicative / expliquée (ici log)



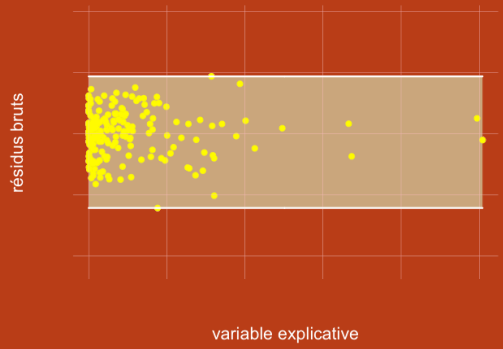
Les graphiques de base

- graphique des résidus $(\widehat{\varepsilon}_i, r_i, t_i)$ en fonction de $\widehat{y}_i$ ou bien de x_i
  - si des formes particulières apparaissent
    - forme incurvée $\Longrightarrow$ non-linéarité



Les graphiques de base

- graphique des résidus ($\widehat{\varepsilon}_i$, r_i, t_i) en fonction de $\widehat{y}_i$ ou bien de x_i
  - si des formes particulières apparaissent
    - comment on peut réparer : transformations de la variable explicative / expliquée (ici x^2)



Les graphiques de base

- résidus anormalement grands détectés
  - d'outliers potentiels
  - si ces résidus apparaissent pour les valeurs extrêmes des $\widehat{y}_i$, il est possible que
    - soit la variance n'est pas constante,
    - soit la vraie relation entre x et y n'est pas linéaire

Sommaire

Les données

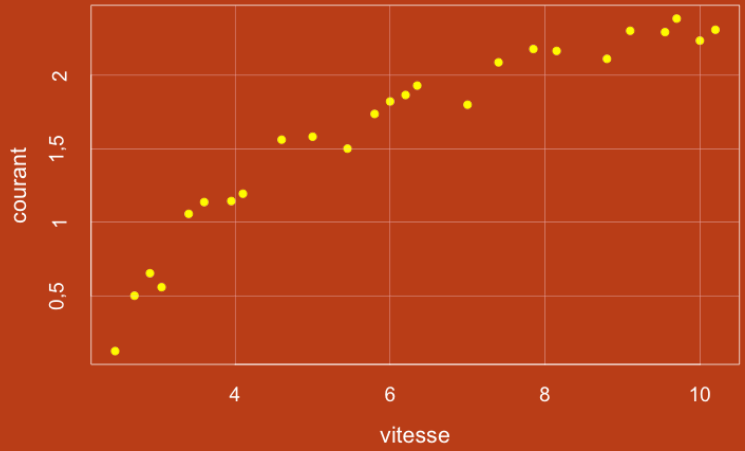
production d'électricité à partir d'énergie éolienne [1]

- vitesse du vent (miles/heure) : *x*
- sortie de courant continu (volts) : *y*

	x vitesse	y courant		x vitesse	y courant
1	2,45	0,123	14	6,20	1,866
2	2,70	0,500	15	6,35	1,930
3	2,90	0,653	16	7,00	1,800
4	3,05	0,558	17	7,40	2,088
5	3,40	1,057	18	7,85	2,179
6	3,60	1,137	19	8,15	2,166
7	3,95	1,144	20	8,80	2,112
8	4,10	1,194	21	9,10	2,303
9	4,60	1,562	22	9,55	2,294
10	5,00	1,582	23	9,70	2,386
11	5,45	1,501	24	10,00	2,236
12	5,80	1,737	25	10,20	2,310
13	6,00	1,822	14		

1. Montgomery, Peck, Vining *Introduction to Linear Regression Analysis* (2006)

Les données



Premier modèle

$$Y = \beta_0 + \beta_1 \times x + \varepsilon$$

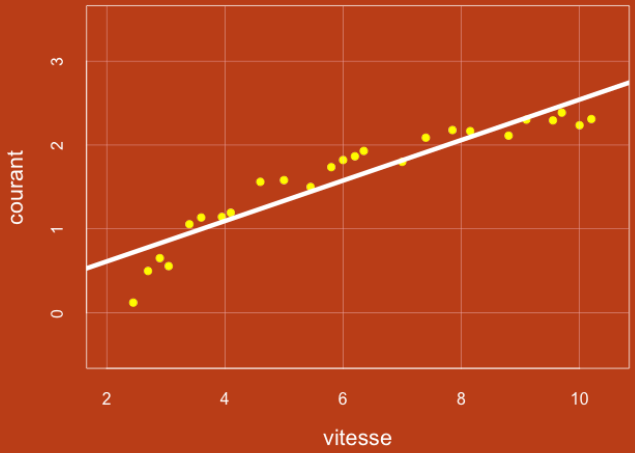
Variable	Coefficient	Ecart-type	t	p_{val}
Intercept	$0,13$	$0,13$	$1,04$	$0,31$
vitesse	$0,24$	$0,02$	$12,66$	8e-12

$\widehat{\sigma}^*$	ddl	R^2	R^2_{adj}
$0,24$	23	$0,875$	$0,869$

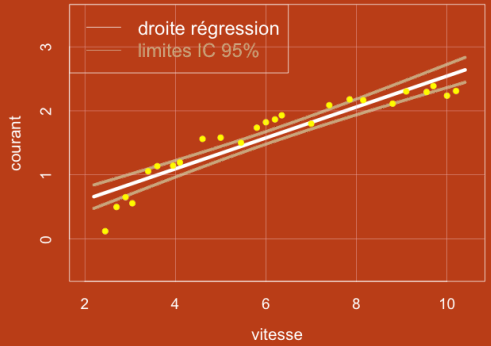
F-statistique : $160,3$	sur 1 et 23 ddl	p-val : 8 e-12

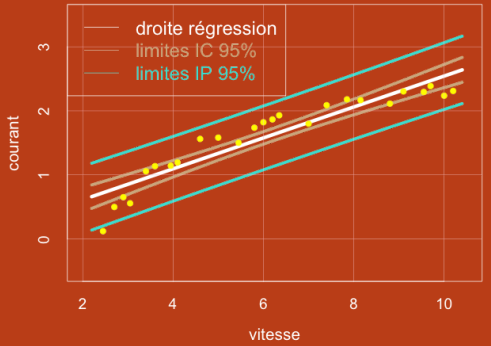
$$\widehat{courant} = 0,13 + 0,24 \times vitesse$$

Premier modèle

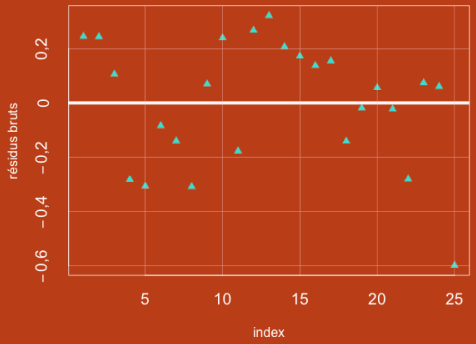


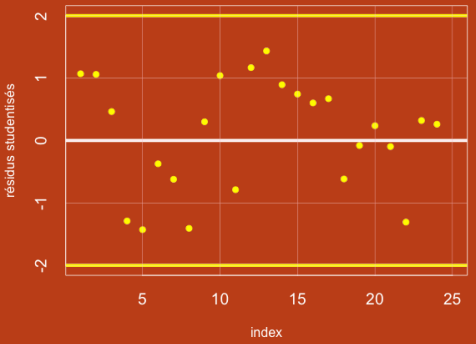
Premier modèle



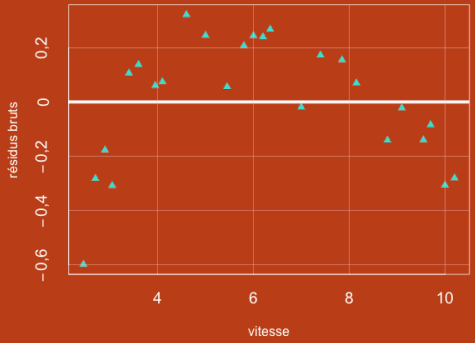


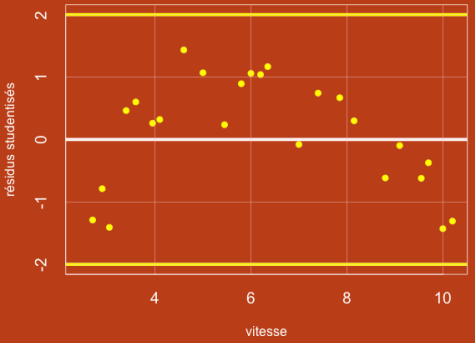
Premier modèle



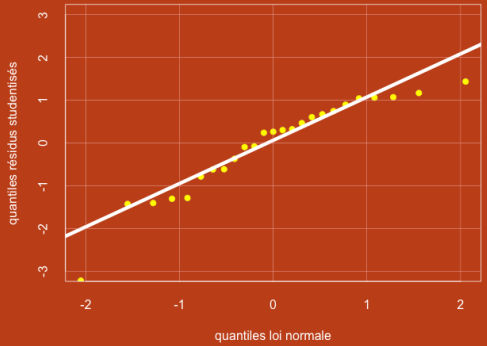


Premier modèle



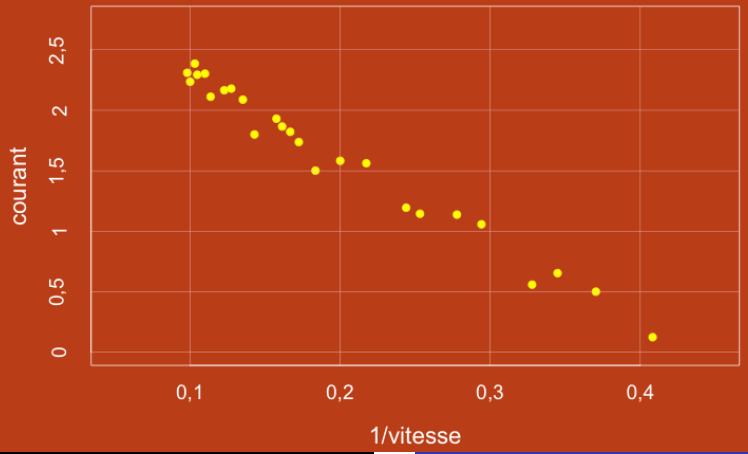


Premier modèle



Les données transformées

vitesse $\mapsto 1/$*vitesse*



Deuxième modèle

$Y = \beta_0 + \beta_1 \times x' + \varepsilon$, avec $x' = 1/x$

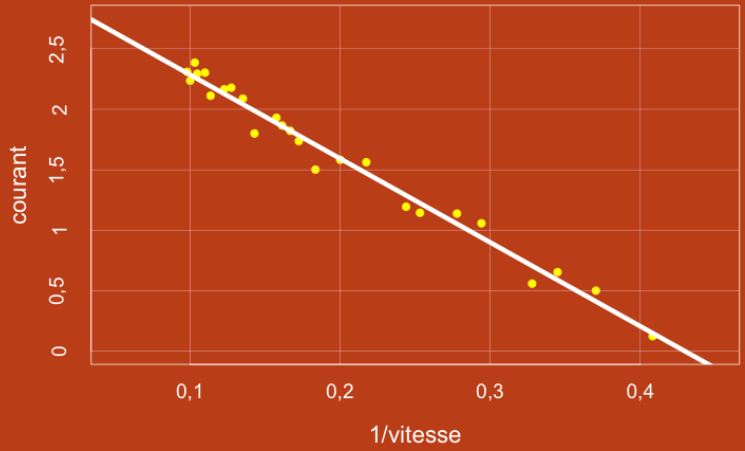
Variable	Coefficient	Ecart-type	t	p_{val}
Intercept	$2,98$	$0,05$	$66,34$	<2e-16
1/vitesse	$-6,93$	$0,21$	$-33,59$	<2e-16

$\widehat{\sigma}^*$	ddl	R^2	R^2_{adj}
$0,094$	23	$0,98$	$0,98$

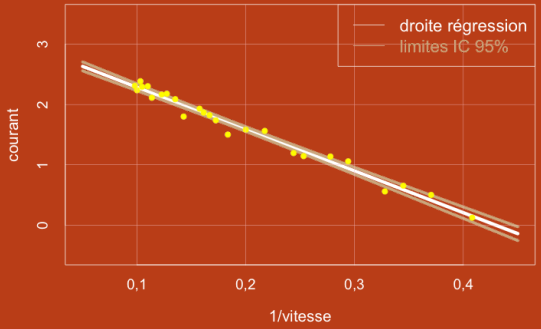
F-statistique : 1128	sur 1 et 23 ddl	p-val : <2e-16

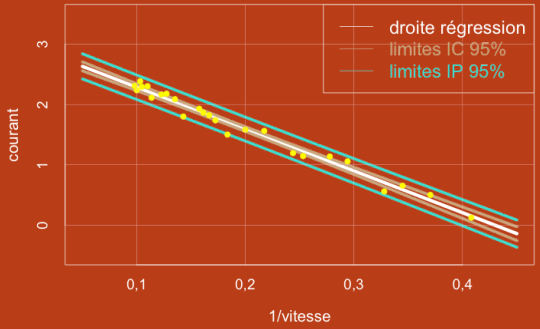
$$\widehat{courant} = 2,98 - 6,93 \times \frac{1}{vitesse}$$

Deuxième modèle

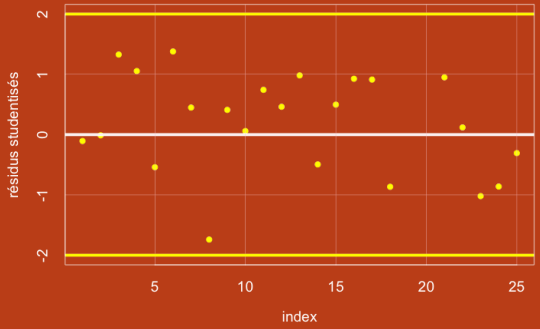


Deuxième modèle

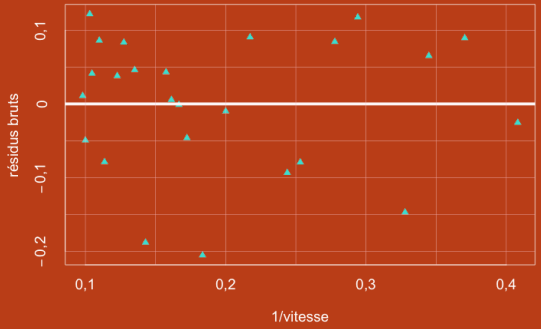


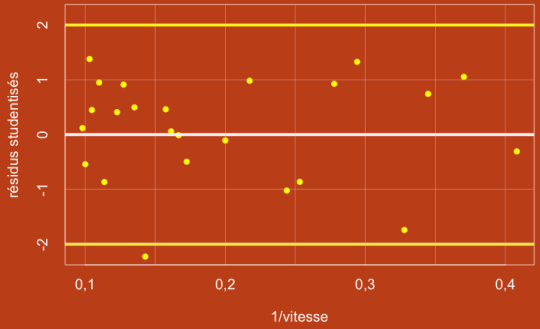


Deuxième modèle

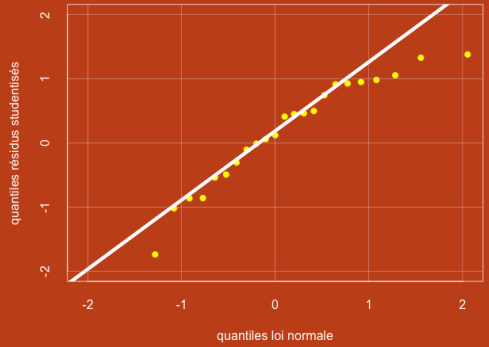


Deuxième modèle





Deuxième modèle



Conclusion

Hypothèses principales :

- la relation entre la variable expliquée et la variable explicative est approximativement linéaire
- l'erreur du modèle est d'espérance nulle et de variance constante σ^2
- les erreurs ne sont pas corrélées
- les erreurs sont de loi normale

Ce qu'il faut retenir :

- toujours douter de la validité de ces hypothèses
- faire des analyses pour valider le modèle estimé
- la non validité des hypothèses peut conduire à un modèle instable
- on ne peut pas détecter les écarts de ces hypothèses en examinant les statistiques standard : R^2, statistique de Student ou de Fisher, car se sont des propriétés globales du modèle

Conclusion générale

- estimer les paramètres d'un modèle de régression linéaire simple par la méthode des MC
- réaliser des tests d'hypothèses et construire des IC pour ces paramètres
- analyser les résidus pour déterminer si le modèle est adéquat et pour vérifier que les hypothèses de départ sont satisfaites
- transformer les données si nécessaire
- prévoir la valeur d'une observation future et construire un intervalle de prévision


INSTITUT
Mines-Télécom