Introduction à la régression
cours n°3

ENSM.SE – axe MSA

L. Carraro

Ecole Nationale
Supérieure des Mines
SAINT-ETIENNE

Retour sur TP tailles

- Compréhension du problème
  - Rôle des variables sexe et poids
- Analyses graphiques unidimensionnelles
  - Valeurs à corriger
  - Enfants trop jeunes à supprimer (≤16 ans ?)
  - Adultes plus âgés

Autres analyses graphiques

- Corrélation des prédicteurs
  - poids/taille → utilisation de l'IMC
  - taille père/taille mère → (tp+tm)/2, tp-tm, tp/tm…
- Analyses bidimensionnelles :
  - Effet apparent de sexe, taille père, taille mère.
  - Interaction sexe*taille père
-  Corrélations et termes d'interaction

Modèle linéaire

- Un modèle par sexe ou un modèle global ?
- Attention aux tables d'ANOVA de R
- Ne pas rejeter **toutes** les variables dont les p-valeurs sont fortes.
- Penser à utiliser R^2 (ou R^2 adj) et RMSE
- Interpréter les matrices de corrélation et non de covariance.
- Validation : attention aux outils de séries temporelles.

Exemple de table d'ANOVA

```
summary(lm(formula=data$taille~data$IMCm+data$IMCd+data$td+data$tm+data
   $sexe+data$tm*data$sexe))

Call:
lm(formula = data$taille ~ data$IMCm + data$IMCd + data$td +
    data$tm + data$sexe + data$tm * data$sexe)

Residuals:
       Min          1Q      Median          3Q         Max
-0.0811588 -0.0272833  0.0004906  0.0320491  0.1254566

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.6829240  0.4709861   1.450   0.1522
data$IMCm          -0.0046741  0.0029860  -1.565   0.1227
data$IMCd          -0.0009283  0.0034064  -0.273   0.7861
data$td            -0.0680851  0.1011091  -0.673   0.5032
data$tm             0.6494623  0.2833605   2.292   0.0254 *
data$sexeG          0.0585179  0.5089905   0.115   0.9088
data$tm:data$sexeG  0.0378336  0.3024440   0.125   0.9009

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04522 on 61 degrees of freedom
Multiple R-Squared: 0.6946,     Adjusted R-squared: 0.6646
F-statistic: 23.12 on 6 and 61 DF,  p-value: 4.927e-14
```

Exemple de résultat simple et de bonne qualité statistique

```
Call:
lm(formula = taille ~ sex3 + I((tm + tp)/2), data = data.def)

Residuals:
      Min        1Q    Median        3Q       Max
-0.083743 -0.029934  0.003118  0.024200  0.134862

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.750417   0.005288 331.039  < 2e-16 ***
sex3           0.115513   0.012230   9.445 4.65e-14 ***
I((tm + tp)/2) 0.627895   0.097913   6.413 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04487 on 69 degrees of freedom
Multiple R-Squared: 0.6657,   Adjusted R-squared: 0.656
F-statistic:  68.7 on 2 and 69 DF,  p-value: < 2.2e-16
```

Matrice de covariance des coefficients estimés

```
> cov2cor(vcov(mod))
                  (Intercept)          sex3 I((tm + tp)/2)
(Intercept)     1.000000e+00 -5.983812e-17   1.098277e-15
sex3           -5.983812e-17  1.000000e+00  -5.485673e-02
I((tm + tp)/2)  1.098277e-15 -5.485673e-02   1.000000e+00
```

Le même modèle présentable

```
Call:
lm(formula = taille ~ sexe + I((tm + tp)/2), data = data.def)

Residuals:
      Min        1Q    Median        3Q       Max
-0.083743 -0.029934  0.003118  0.024200  0.134862

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.66378    0.01059 157.149  < 2e-16 ***
sexeG           0.11551    0.01223   9.445 4.65e-14 ***
I((tm + tp)/2)  0.62789    0.09791   6.413 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04487 on 69 degrees of freedom
Multiple R-Squared: 0.6657,   Adjusted R-squared: 0.656
F-statistic:  68.7 on 2 and 69 DF,  p-value: < 2.2e-16
```

Suite du cours : rappel démarche

- Observations graphiques
  - prédicteurs, réponse, prédicteurs entre eux et contre réponse (dont interactions)
- Modélisation et inférence
  - estimation paramètres + corrélation, ANOVA, résidus
- Observations à problème ou influentes
- Prévisions

Exemple 3
distributeurs de boisson

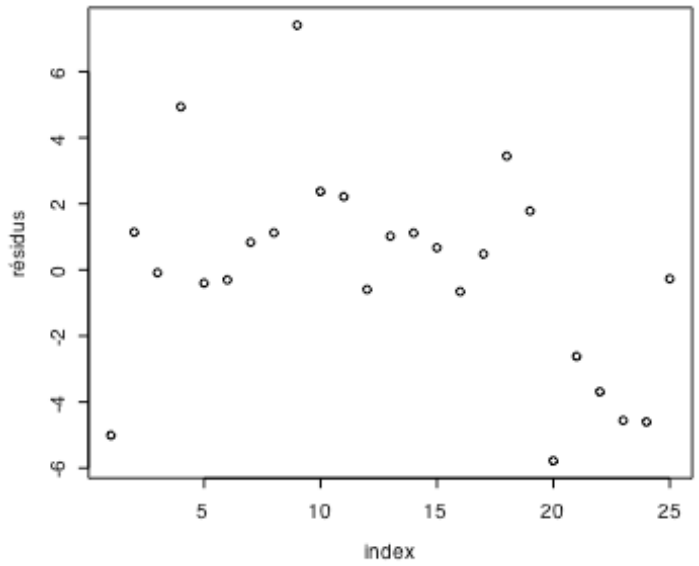
- réponse = temps
- prédicteurs = distance, nombre de caisses
- après analyses graphiques préliminaires, modèle candidat :

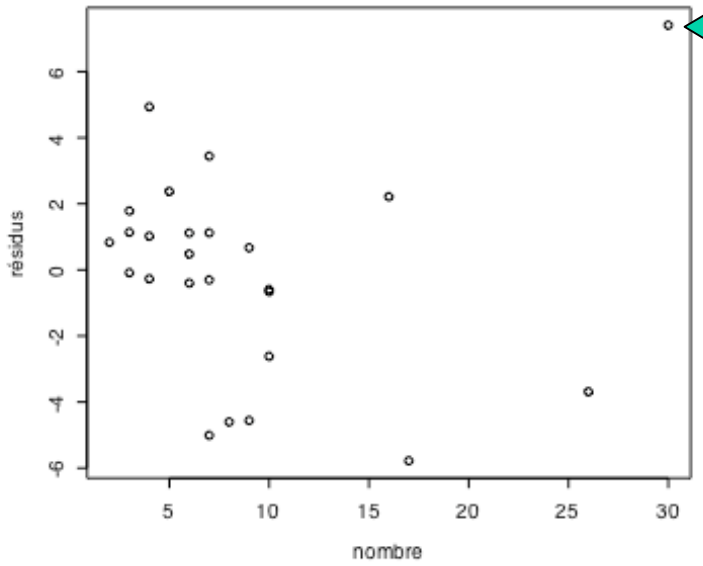
$$temps = \beta_0 + \beta_{nb}\, nb + \beta_{dist}\, dist$$

Exemple 3 - résumé modèle

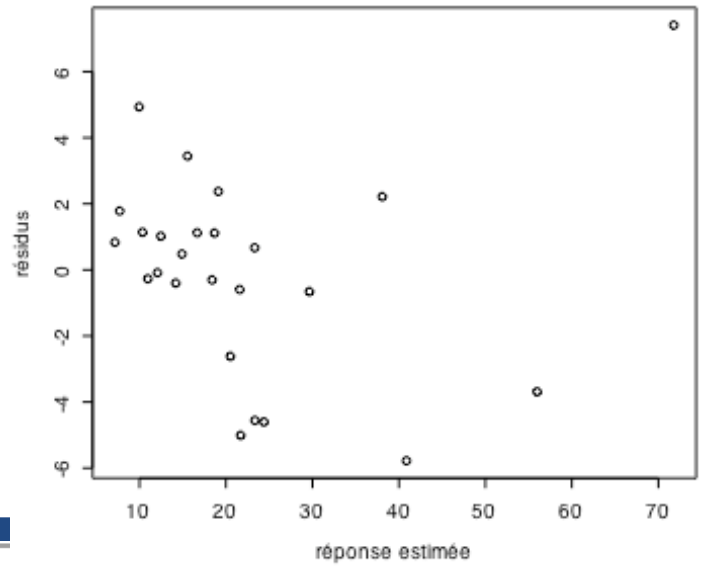
```
Call:
lm(formula = temps ~ nb + distance, data = boissons)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7771 -0.6576  0.4817  1.1395  7.4093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.353134   1.095117   2.149 0.042918 *
nb          1.615100   0.170484   9.474  3.2e-09 ***
distance    0.014373   0.003608   3.984 0.000627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.255 on 22 degrees of freedom
Multiple R-Squared: 0.9597,      Adjusted R-squared: 0.956
F-statistic: 261.7 on 2 and 22 DF,  p-value: 4.601e-16
```

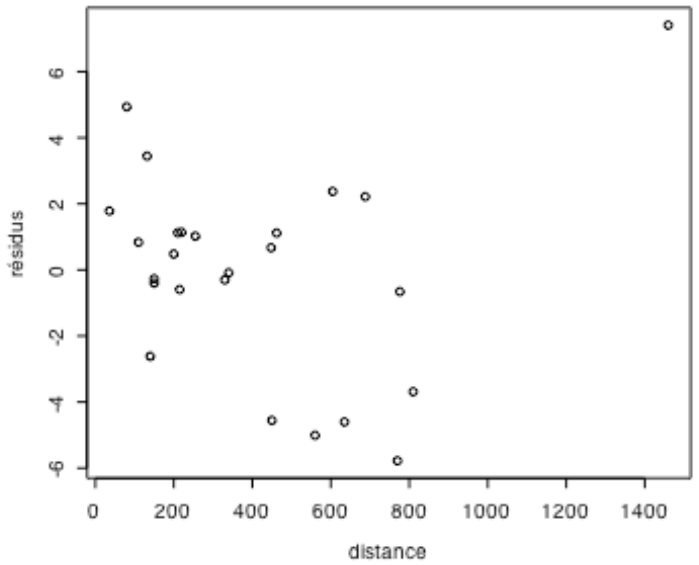
Exemple 3 - résidus bruts





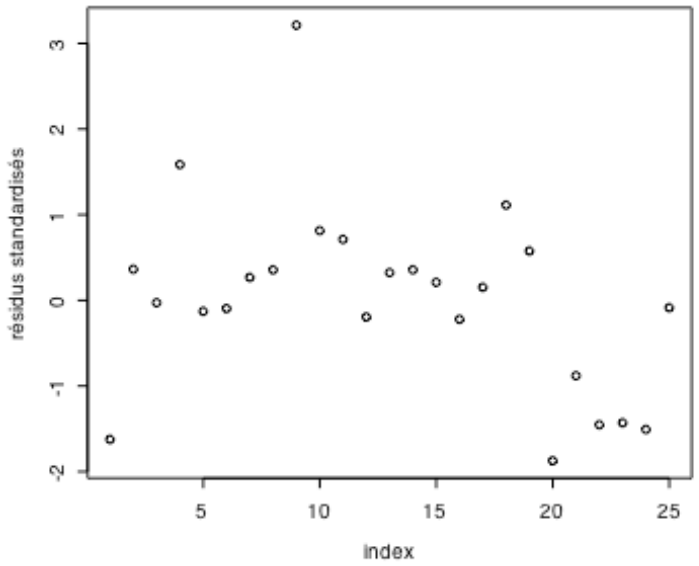
Problème ??

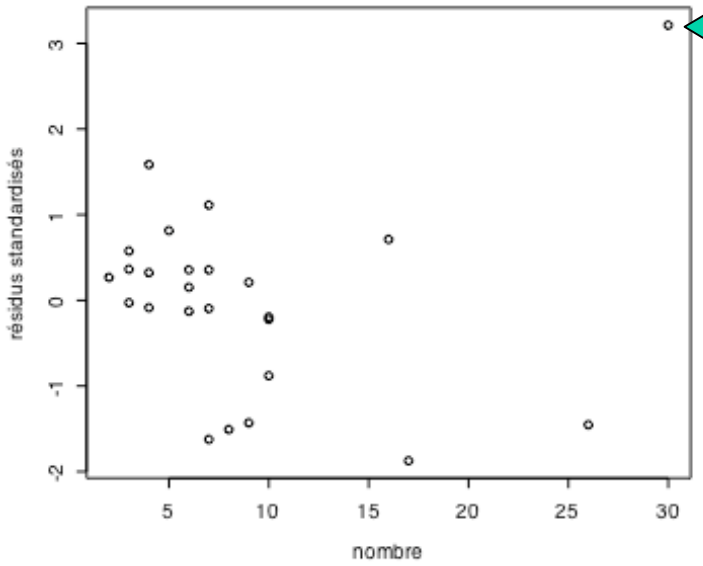




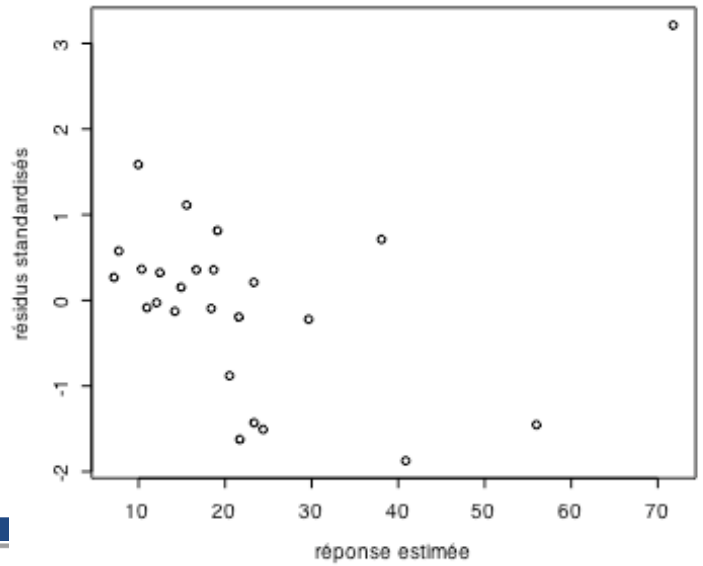
Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

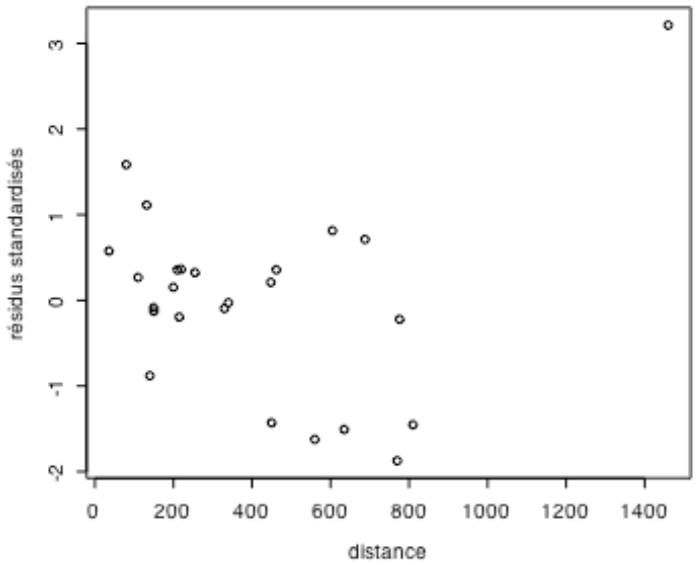
Exemple 3 - résidus standardisés



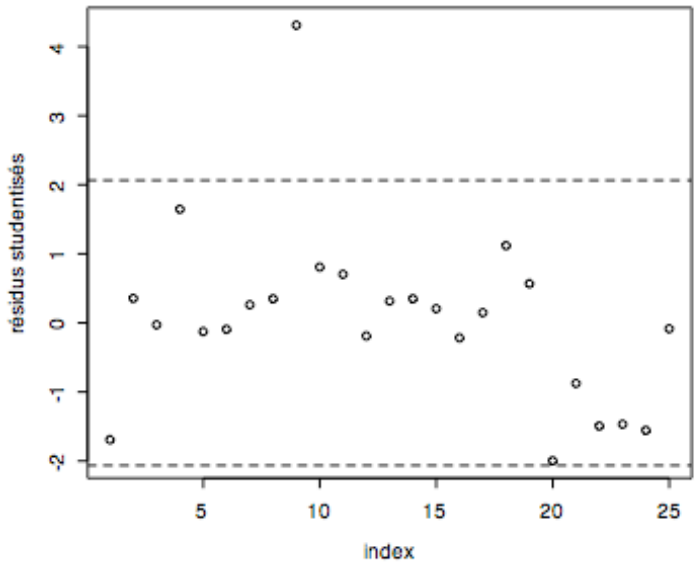


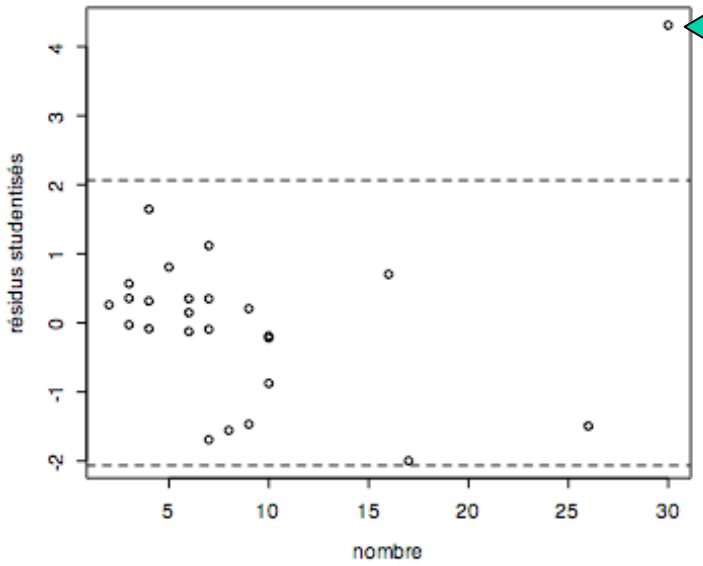
Problème ??



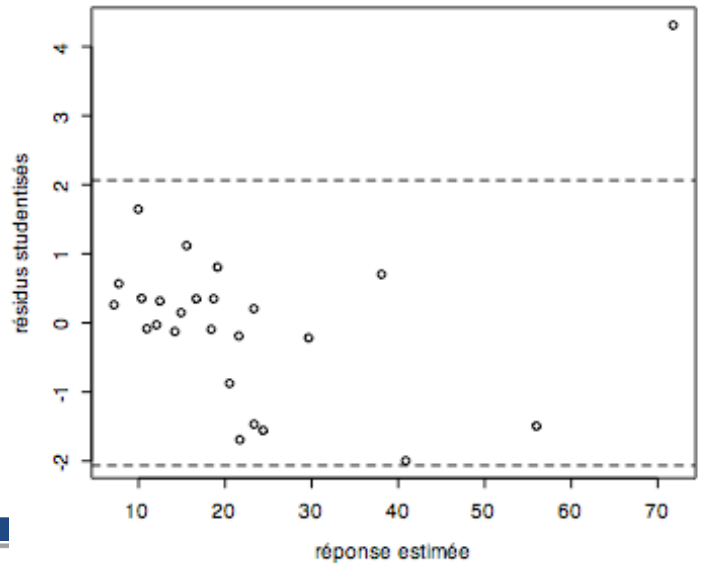


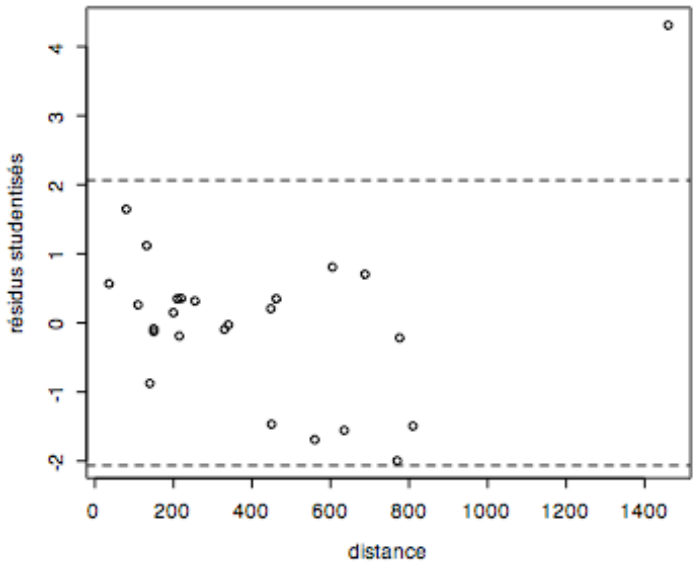
Exemple 3 - résidus studentisés





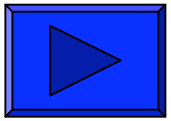
Problème !!





Observations aberrantes
Observations influentes

- Retour sur l'exemple 3
  - Observation n°9 aberrante (résidus studentisés)
  - Résidus bruts, standardisés, studentisés très différents pour cette observation

- Simulation Excel 

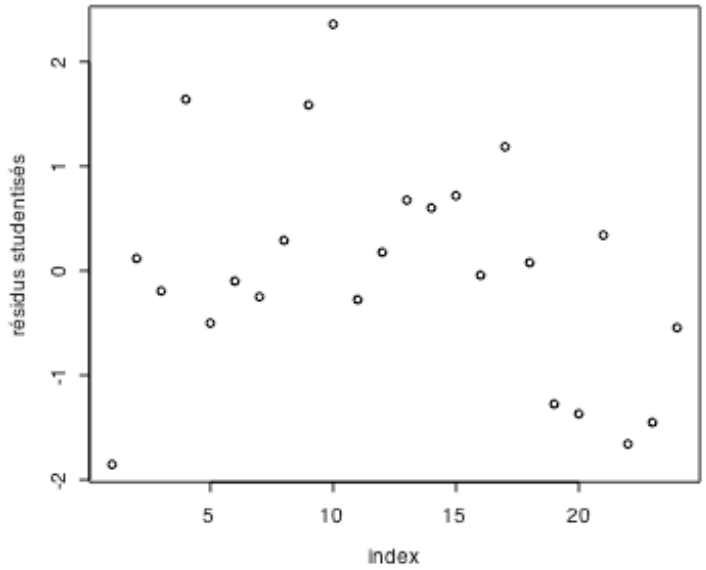
Exemple 3 - rappel modèle 25 données

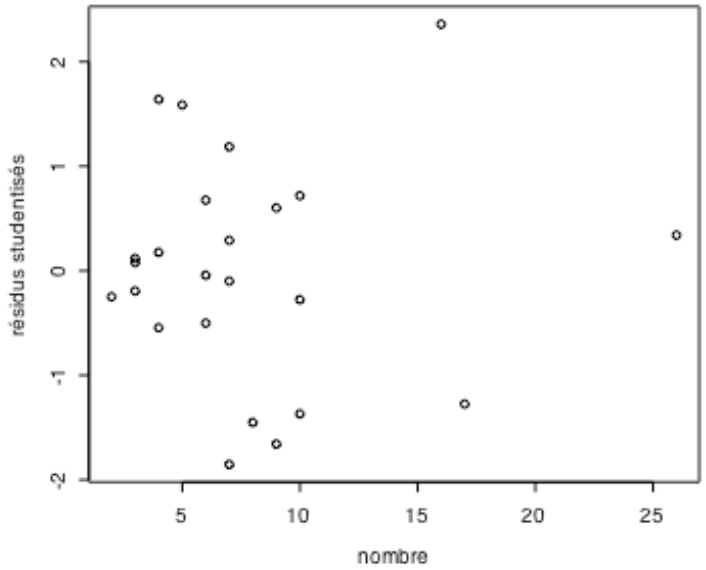
```
Call:
lm(formula = temps ~ nb + distance, data = boissons)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7771 -0.6576  0.4817  1.1395  7.4093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.353134   1.095117   2.149 0.042918 *
nb          1.615100   0.170484   9.474  3.2e-09 ***
distance    0.014373   0.003608   3.984 0.000627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.255 on 22 degrees of freedom
Multiple R-Squared: 0.9597,      Adjusted R-squared: 0.956
F-statistic: 261.7 on 2 and 22 DF,  p-value: 4.601e-16
```

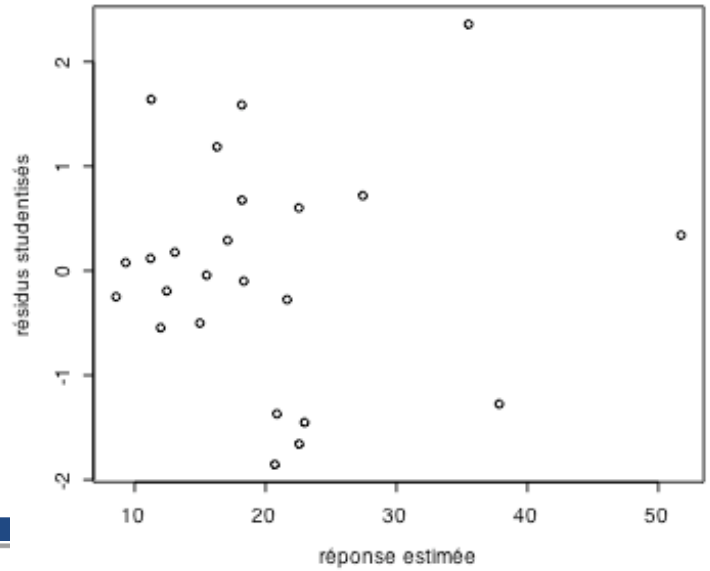
Exemple 3 - modèle sans obs. n° 9

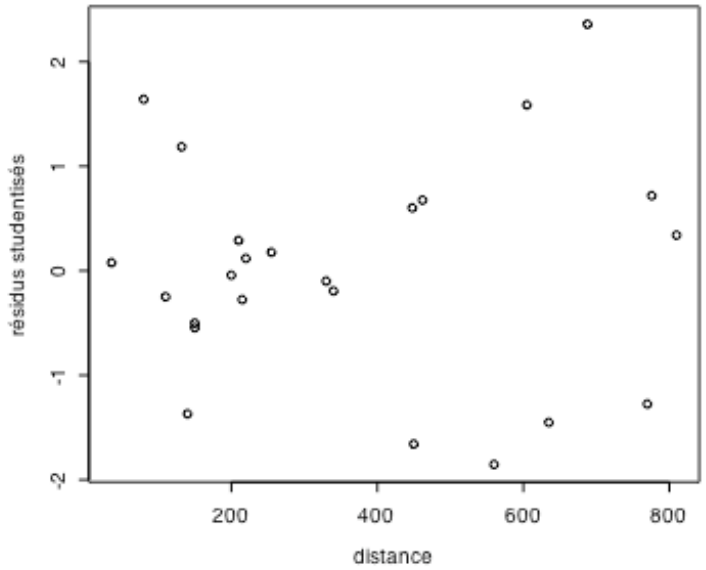
```
Call:
lm(formula = temps ~ nb + distance, data = boissons2)

Residuals:
     Min       1Q   Median       3Q      Max
-4.01359 -1.21265  0.03958  1.47758  4.79225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.456173   0.951015   4.686 0.000126 ***
nb          1.497050   0.130008  11.515 1.55e-10 ***
distance    0.010318   0.002849   3.621 0.001601 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.426 on 21 degrees of freedom
Multiple R-Squared: 0.9488,      Adjusted R-squared: 0.9439
F-statistic: 194.6 on 2 and 21 DF,  p-value: 2.798e-14
```

sans obs n°9 - résidus studentisés

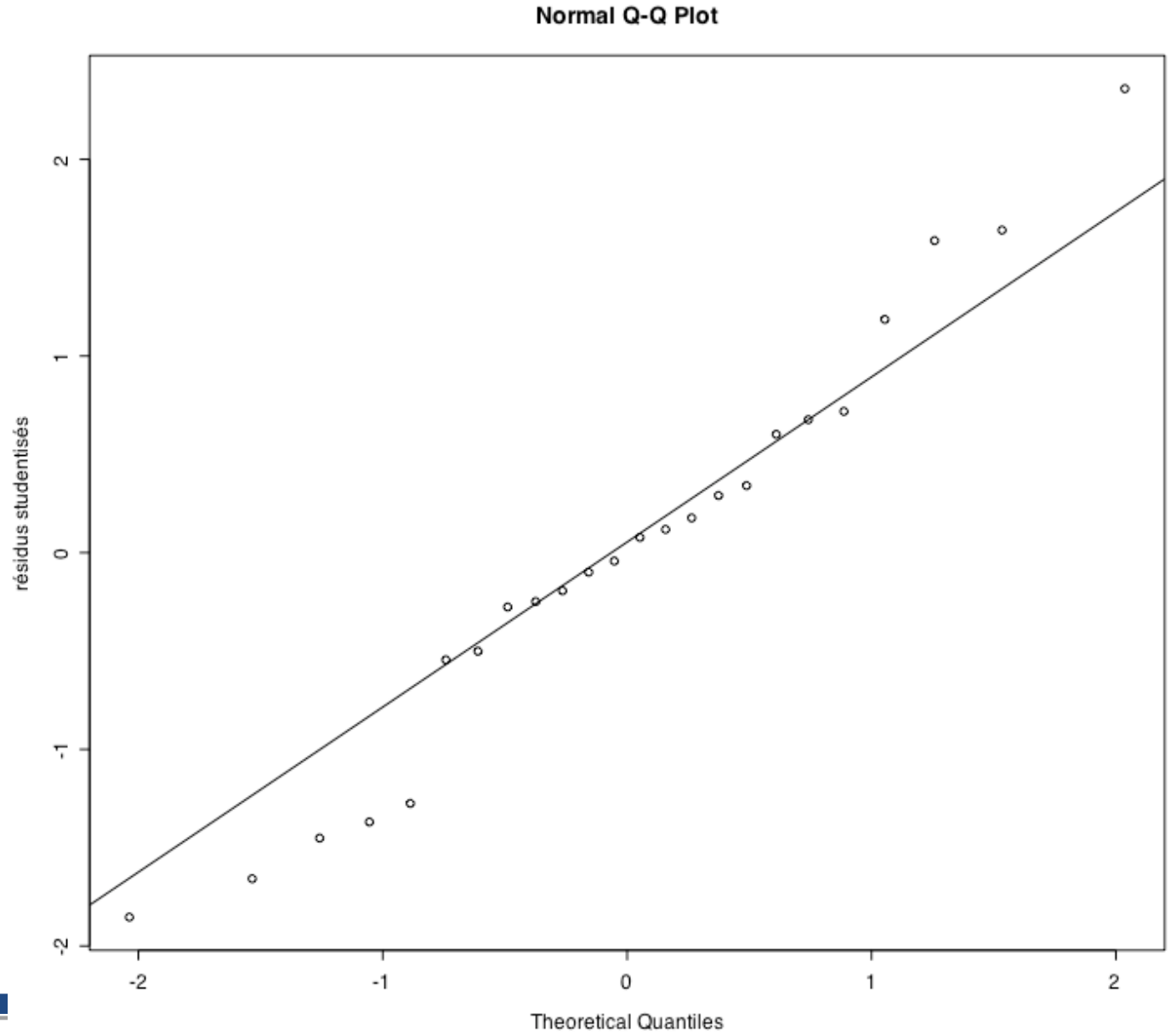








Sans obs n°9 - droite de Henri



Tests d'adéquation

	Kolmogorov Student	Kolmogorov Gaussienne	Shapiro-Wilk Gaussienne
commandes	`ks.test(res, "pt",df)`	`ks.test(res, "pnorm",0,1)`	`shapiro.test (res)`
25 données	D = 0.18	D = 0.17	W = 0.87
	p-value = 0.38 OK	p-value = 0.39 OK	p-value = 0.004 NON
sans obs. 9	D = 0.10	D = 0.11	W = 0.97
	p-value = 0.95 OK	p-value = 0.92 OK	p-value = 0.69 OK

Détection de valeurs influentes

- Levier h_{ii} :

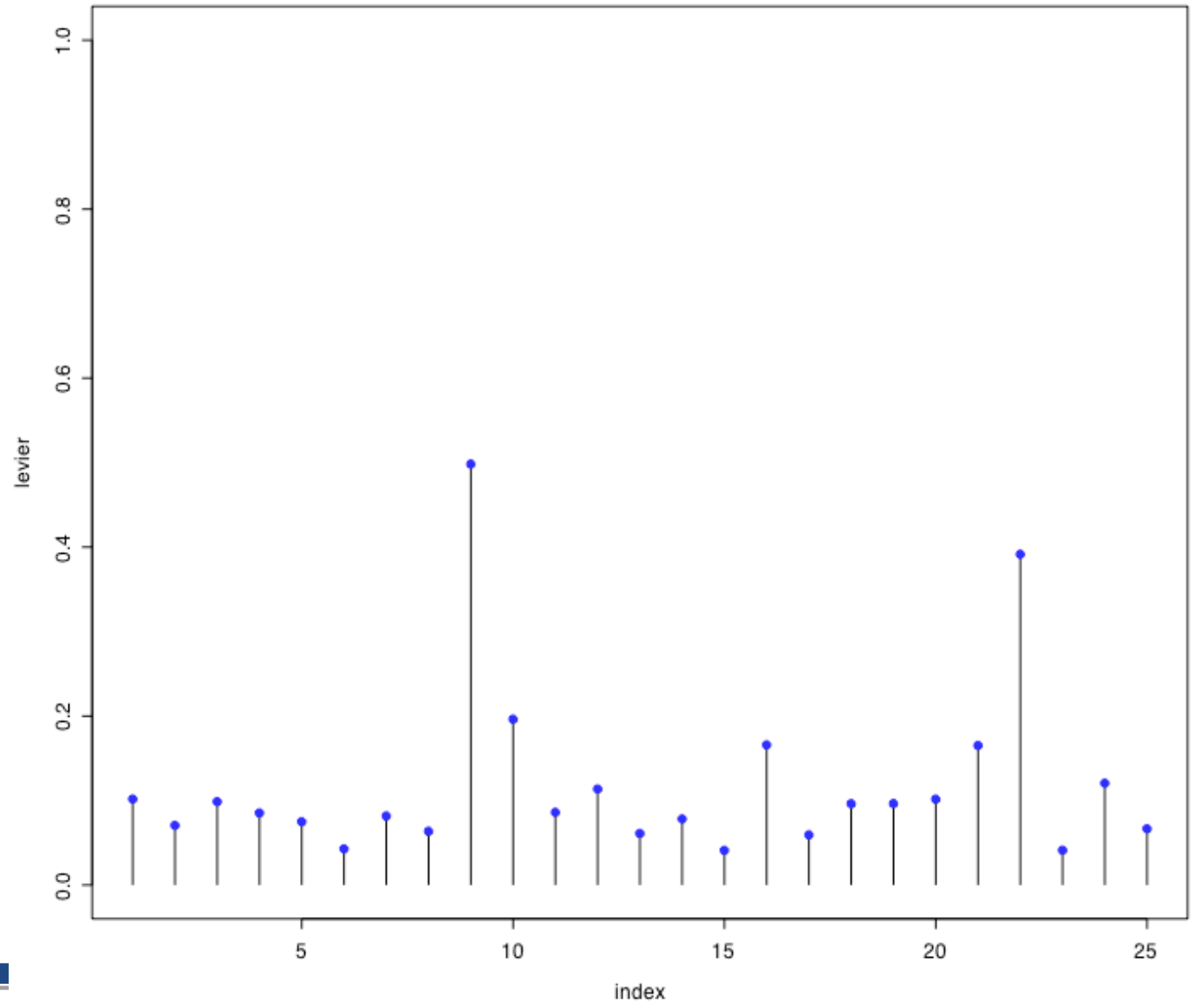
$$\hat{y}_i = \sum_{i=1}^{n} h_{ij}\, y_j \text{ et } \mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2 (1-h_{ii})$$

- Distance Cook D_i :

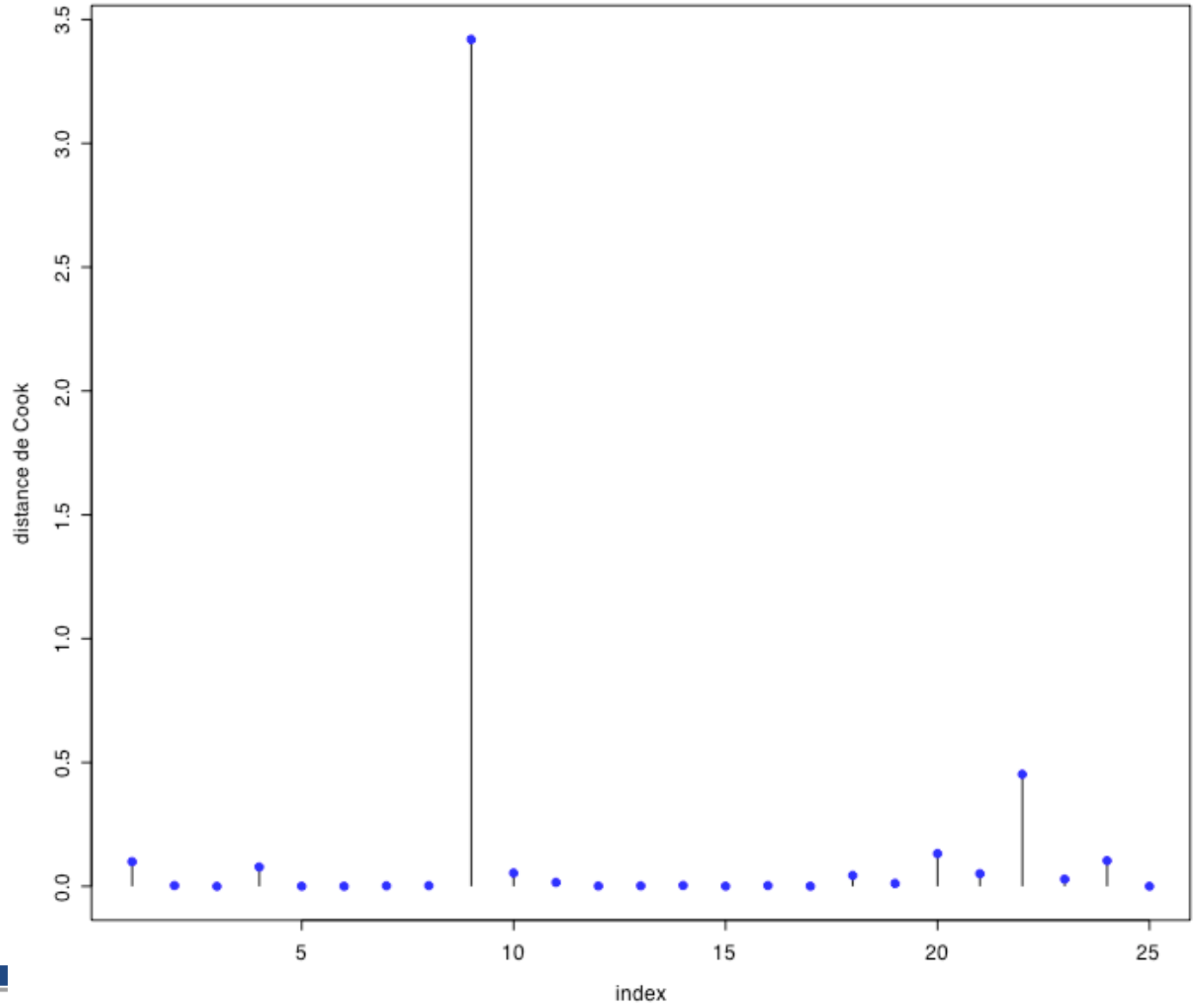
Distance a-dimensionnelle entre les réponses estimées avec ou sans l'observation n° i.

$$D_i = \frac{T_i^2}{p+1} \frac{h_{ii}}{1-h_{ii}}$$

Exemple 3 - leviers



Exemple 3 – distances de Cook



Prévisions

- Nouvelle valeur des prédicteurs x_{new}
- Prévision pour y :

$$\hat{y}_{new} = x_{new}\,\hat{\beta}$$

  - Intervalle de confiance pour $x_{new}\,\beta$ (réponse espérée) ?
  - Intervalle de prévision pour la réponse y_{new} ?

Intervalles de confiance/prévision

- De la forme :

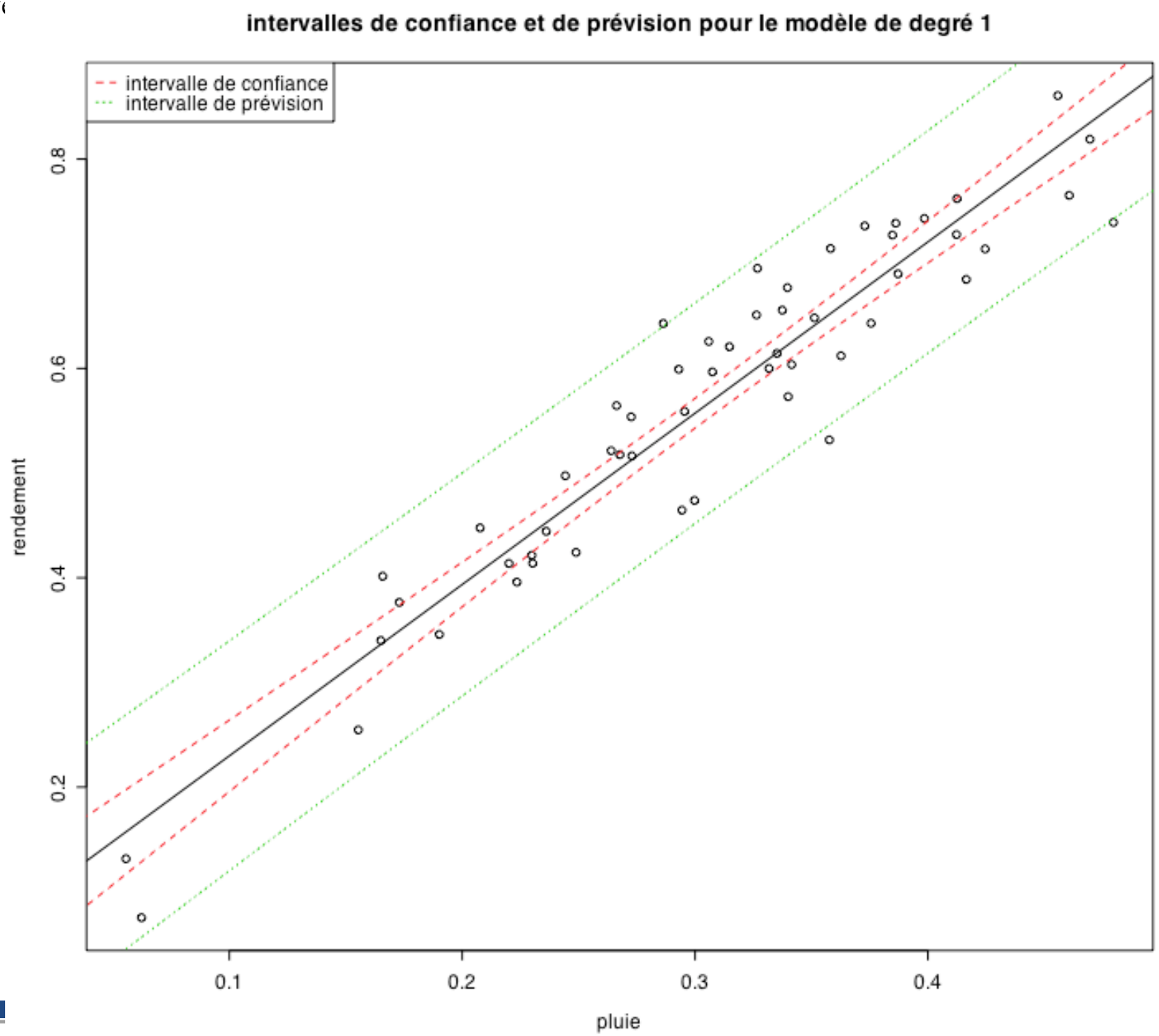
$$[x_{new}\hat{\beta} - s(x_{new})\, t^{-1}_{n-p-1}(1-\alpha/2),\ x_{new}\hat{\beta} + s(x_{new})\, t^{-1}_{n-p-1}(1-\alpha/2)]$$

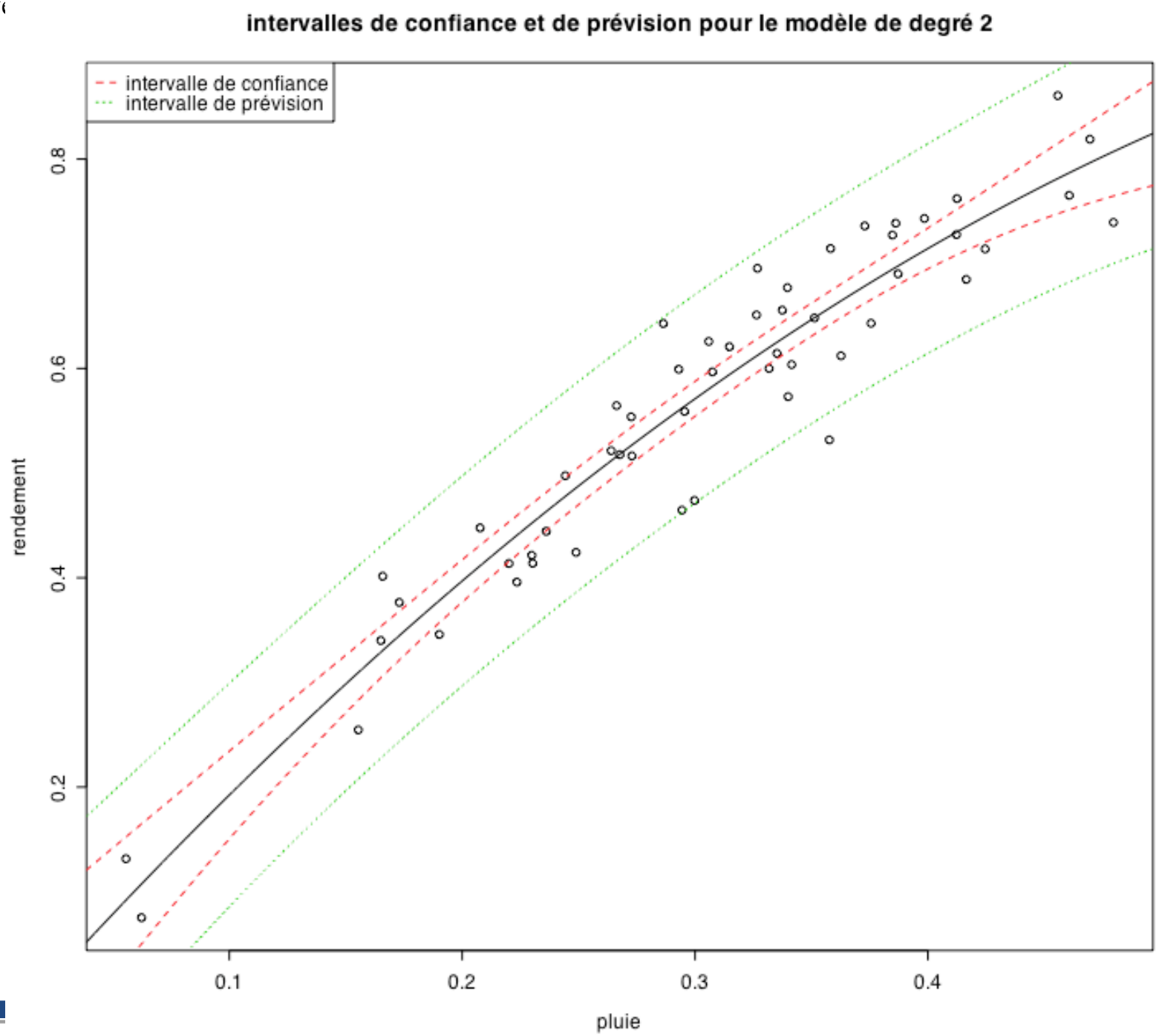
  - Confiance :

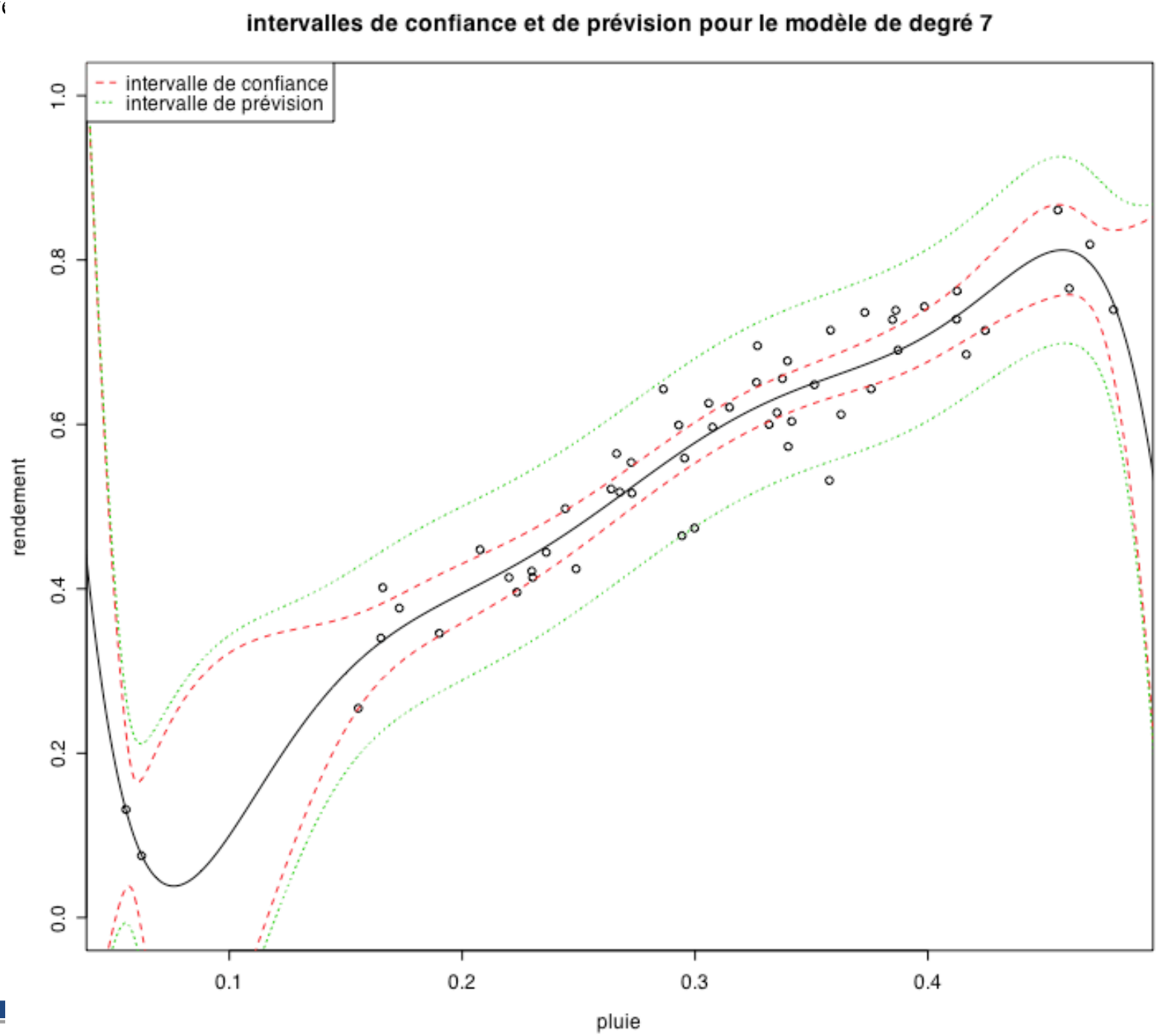
    La pente est-elle >1 ? la droite passe-t-elle par 0 ?

  - Prévision :

    A quel rendement s'attendre pour 20 mm de pluie ?


intervalles de confiance et de prévision pour le modèle de degré 1
intervalle de confiance
intervalle de prévision
rendement
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4
pluie


intervalles de confiance et de prévision pour le modèle de degré 2
intervalle de confiance
intervalle de prévision
rendement
pluie
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4


intervalles de confiance et de prévision pour le modèle de degré 7
intervalle de confiance
intervalle de prévision
rendement
pluie
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4