Introduction à la régression
cours n°4

ENSM.SE – axe MSA

L. Carraro

Ecole Nationale
Supérieure des Mines
SAINT-ETIENNE

Résidus et validation
ex 3 des distributeurs de boisson

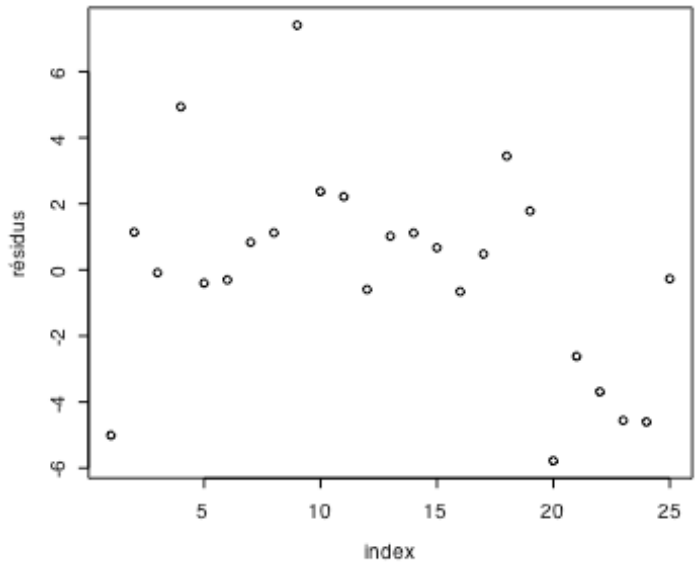
- réponse = temps
- prédicteurs = distance, nombre de caisses
- 25 observations
- Après analyses graphiques préliminaires, modèle candidat :

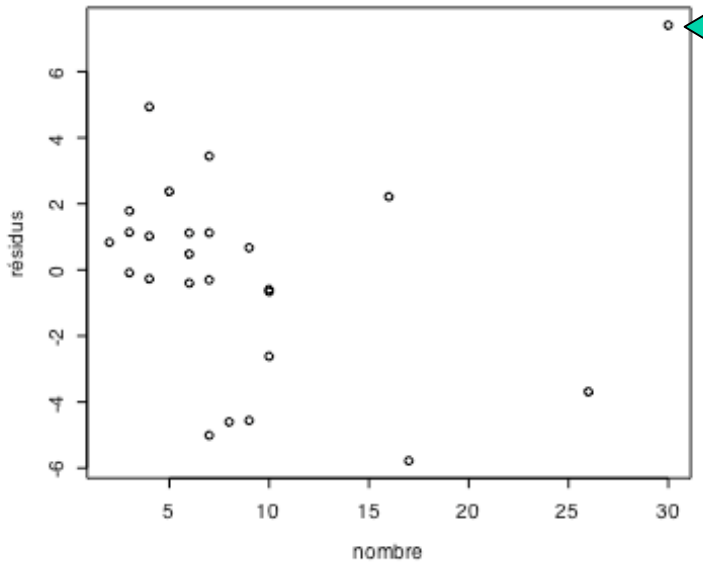
$$temps = \beta_0 + \beta_{nb}\, nb + \beta_{dist}\, dist$$

Exemple 3 - résumé modèle

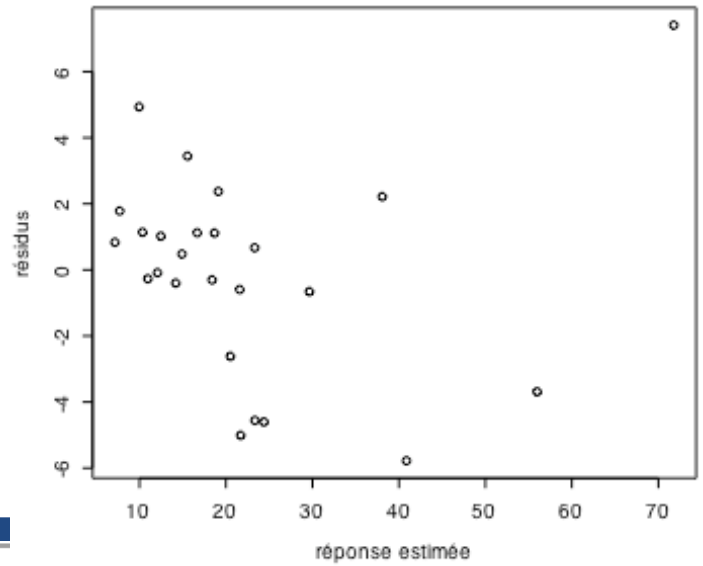
```
Call:
lm(formula = temps ~ nb + distance, data = boissons)

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
```

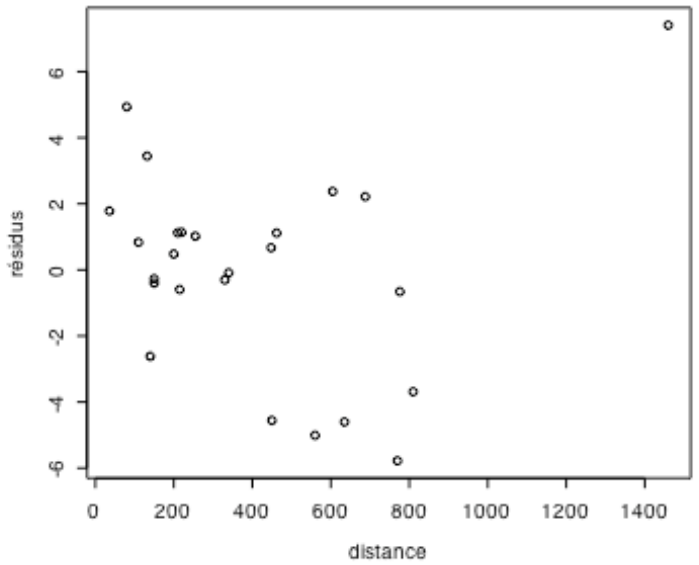
Exemple 3 - résidus bruts





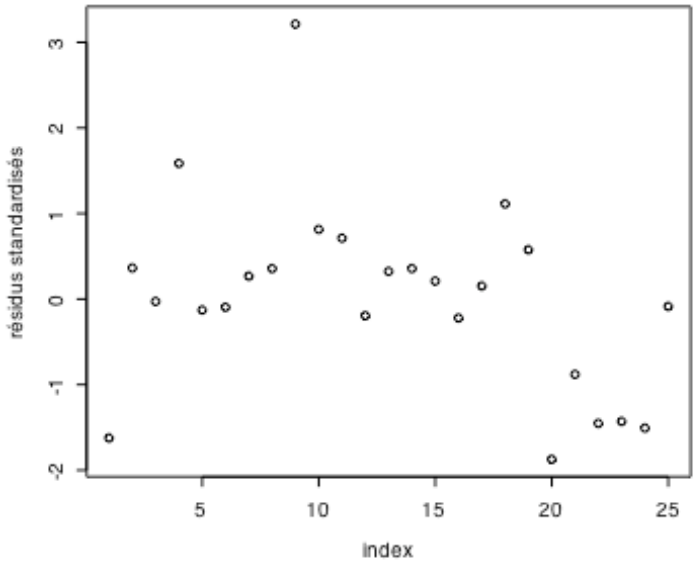
Problème ??

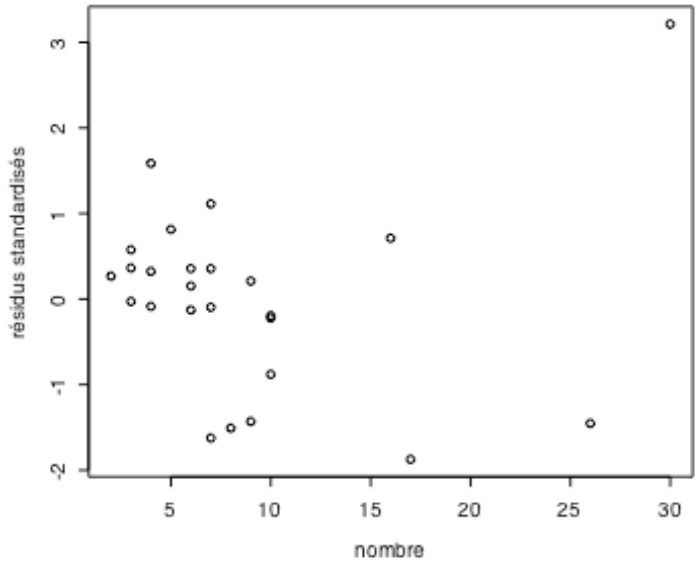


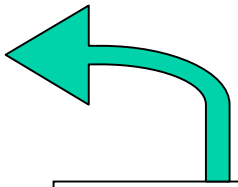


Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

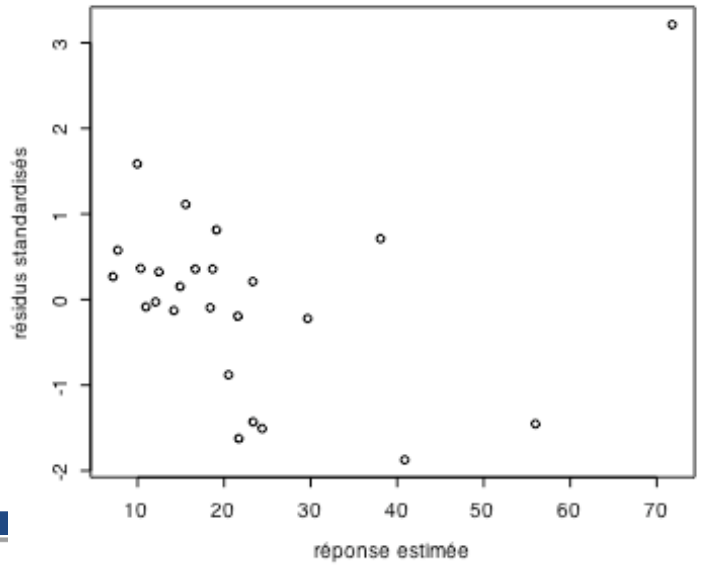
Exemple 3 - résidus standardisés

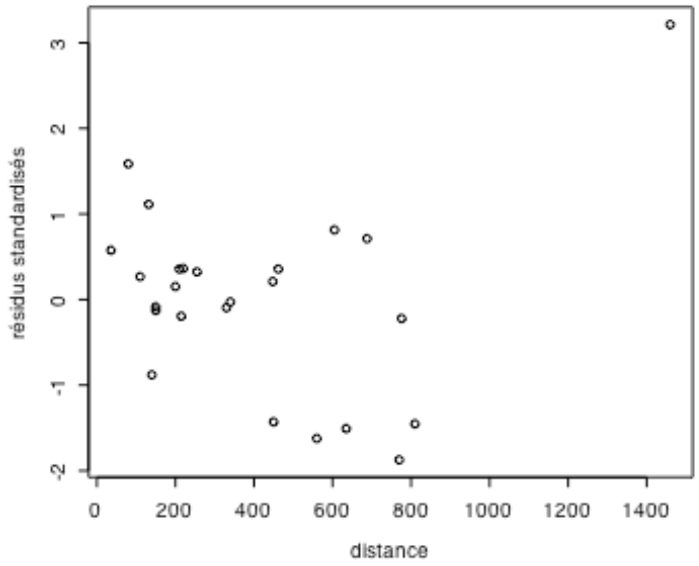






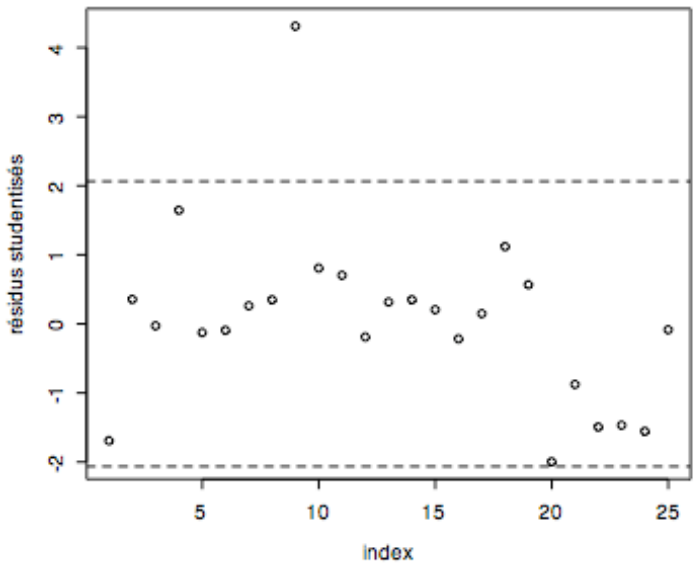
Problème ??

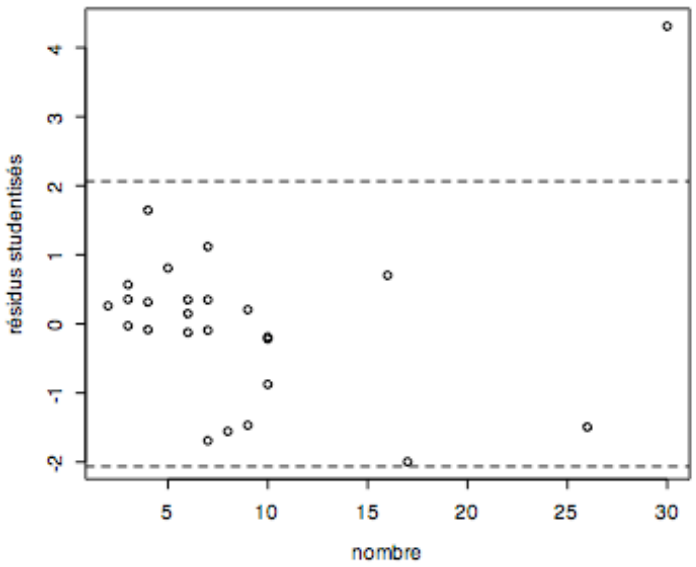




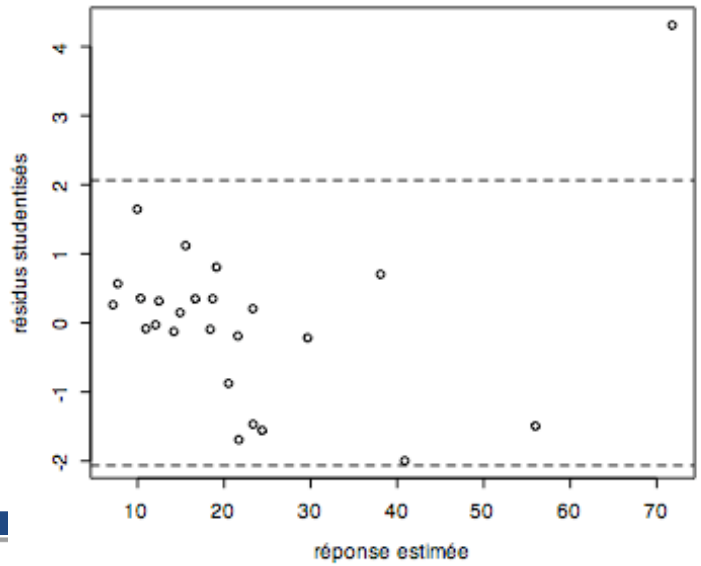
Ecole Nationale Supérieure des Mines SAINT-ETIENNE

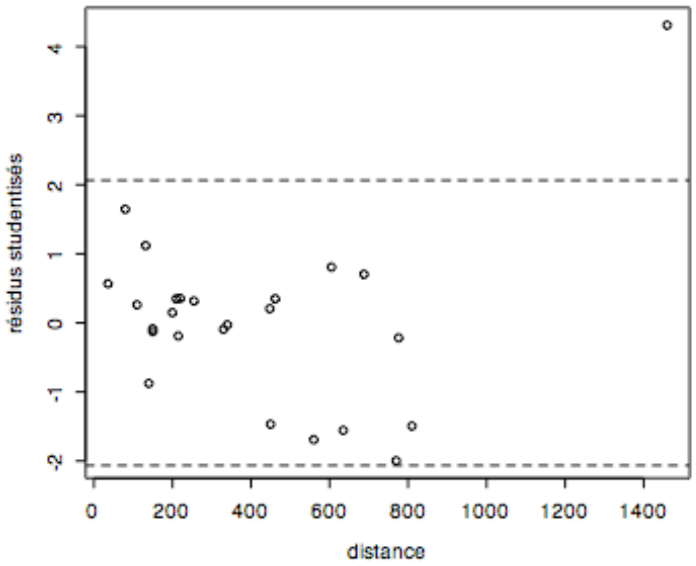
Exemple 3 - résidus studentisés





Problème !!





Ecole Nationale Supérieure des Mines SAINT-ETIENNE

Observations aberrantes
Observations influentes

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

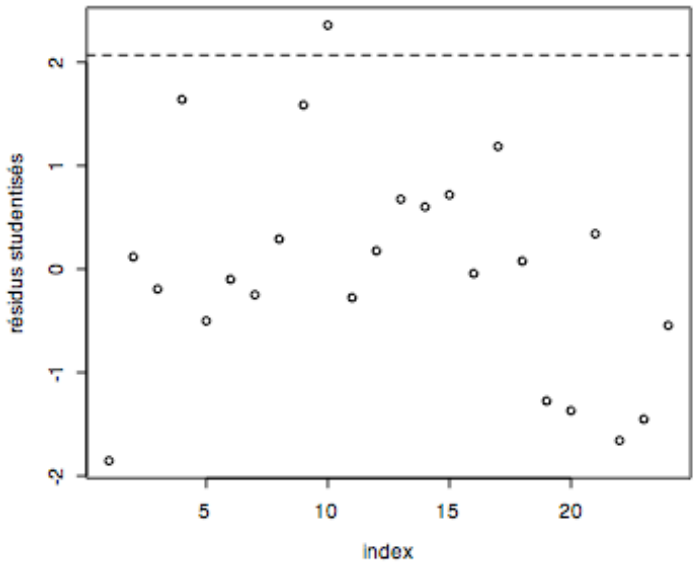
Exemple 3 - rappel modèle 25 données

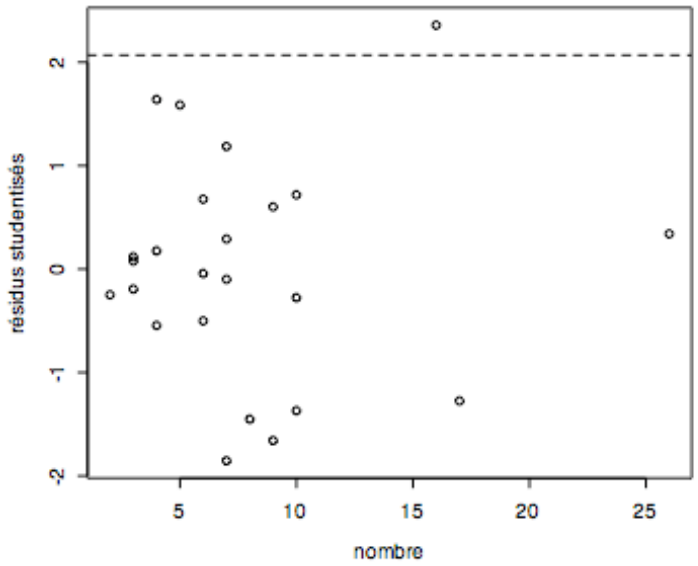
```
Call:
lm(formula = temps ~ nb + distance, data = boissons)

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
```

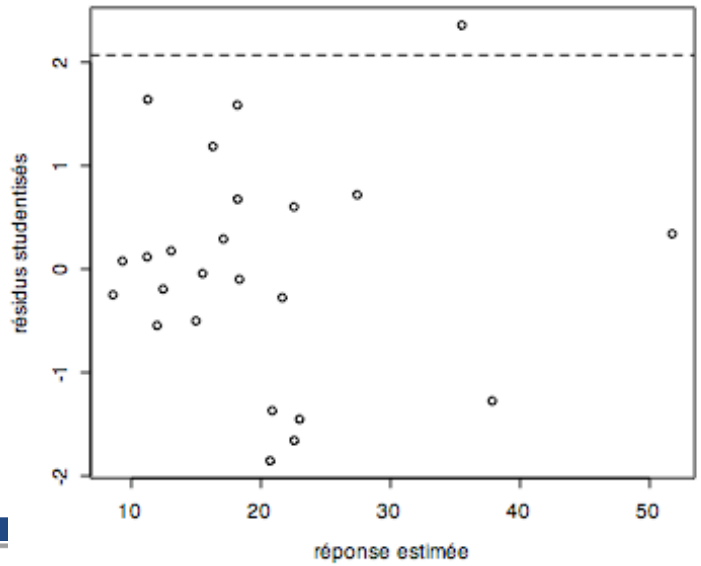
Exemple 3 - modèle sans obs. n° 9

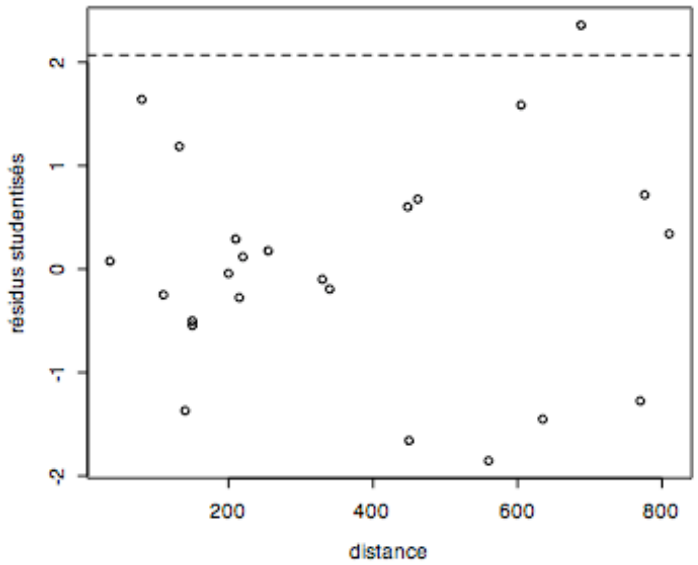
```
Call:
lm(formula = temps ~ nb + distance, data = boissons2)

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

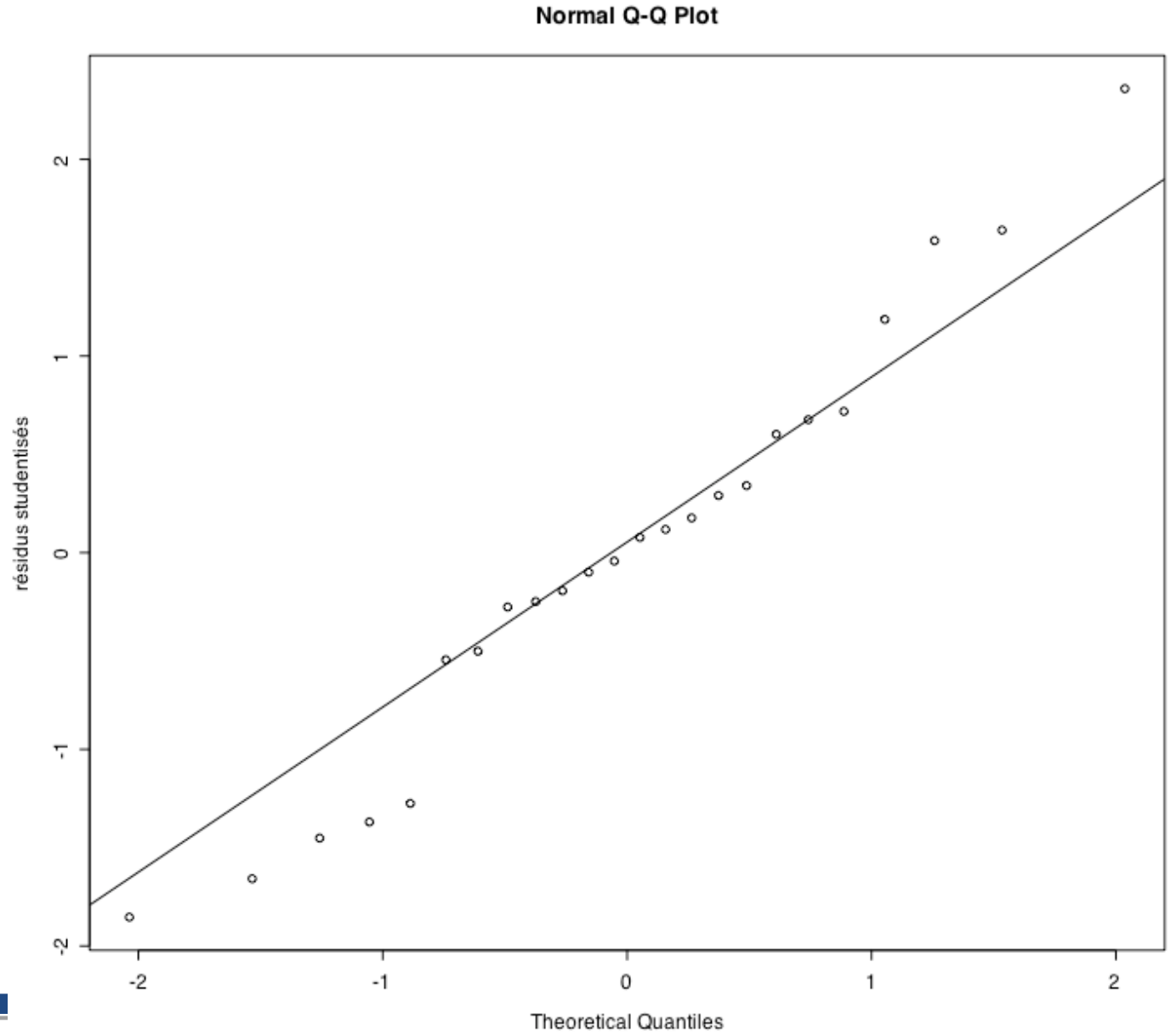








Sans obs n°9 - droite de Henri



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

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 :

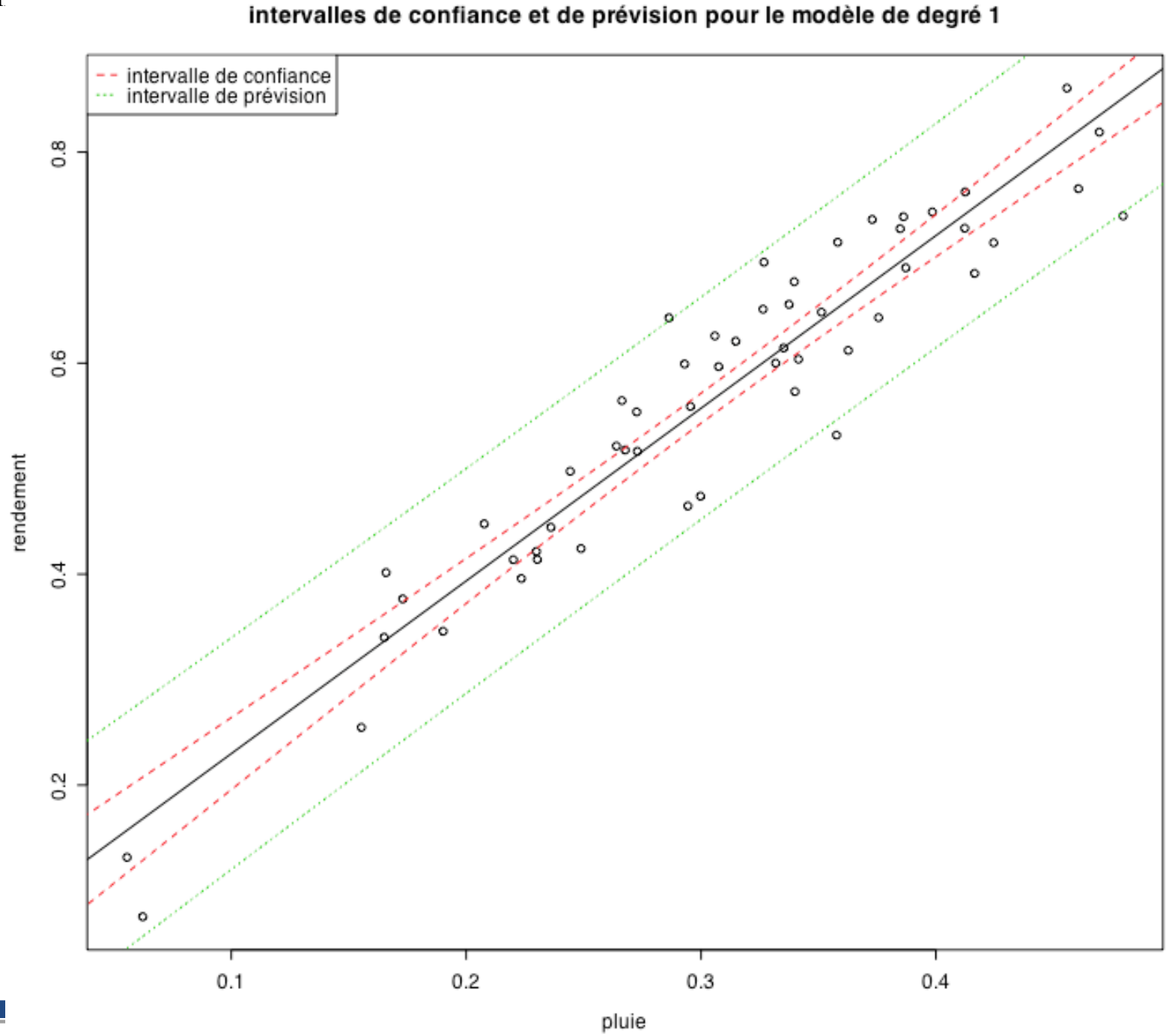
$$[x_{new}\hat{\beta} - s(x_{new})\, t^{-1}_{n-p-1}(1-\alpha/2),\ x_{new}\hat{\beta} + s(x_{new})\, t^{-1}_{n-p-1}(1-\alpha/2)]$$

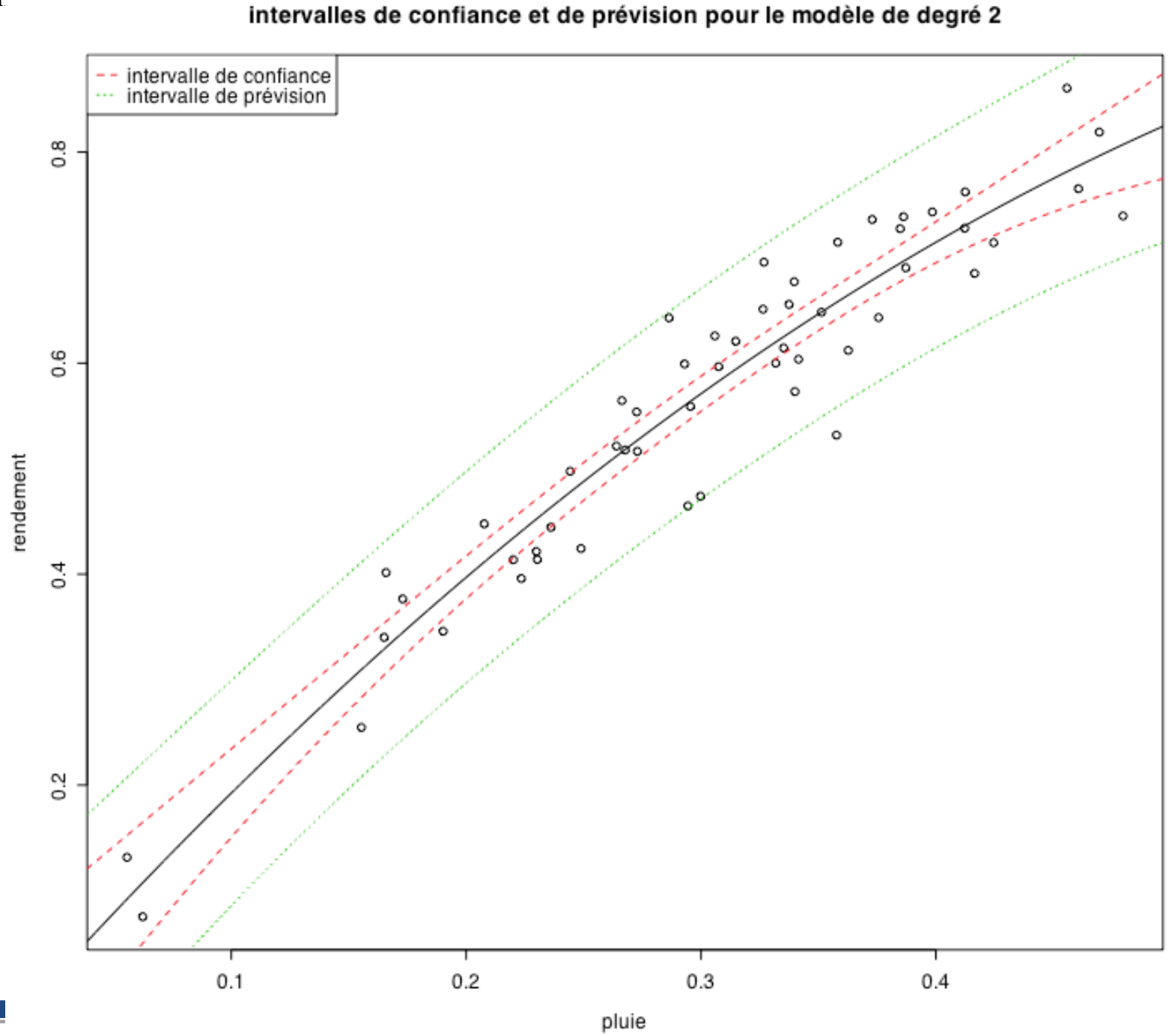
- Confiance :

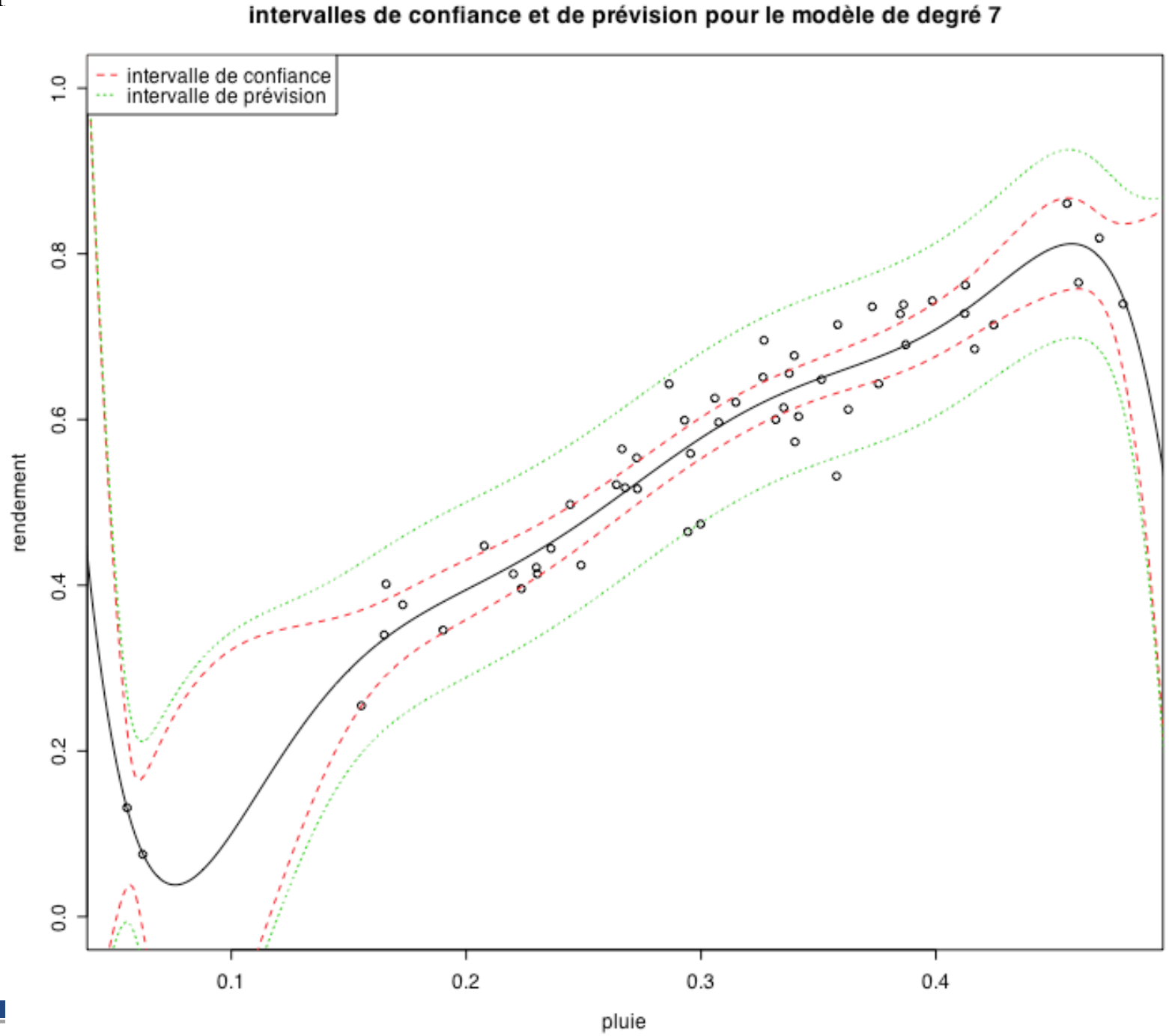
  La pente est-elle >1 ? la droite passe-t-elle par 0 ?

- 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
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 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

Retour TP exploration pétrolière

- Compréhension du problème
  - Monte Carlo
  - Approximation du simulateur : plan d'expériences et régression
- Qualité des plans
  - Bonne répartitition a priori
  - Critères (D, A…)
  - Aspects qualitatifs éventuels à l'issue

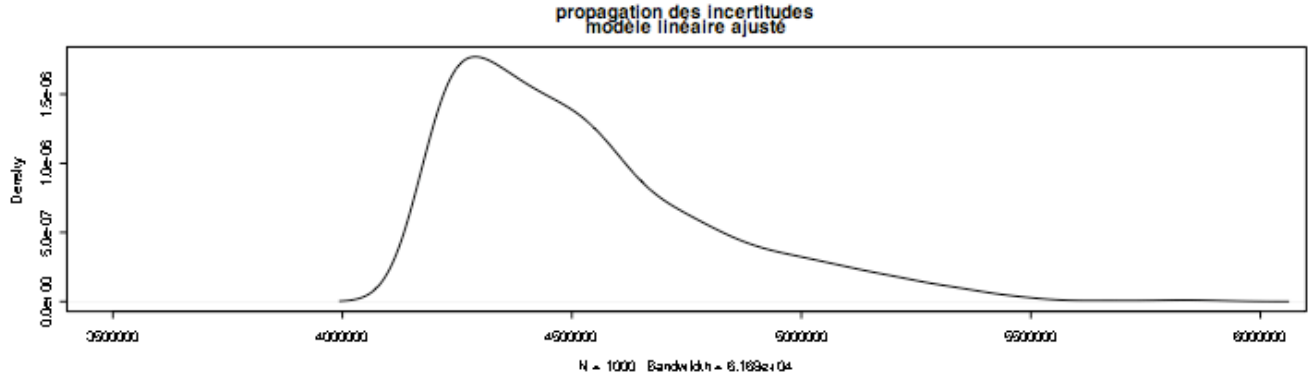
Propagation des incertitudes

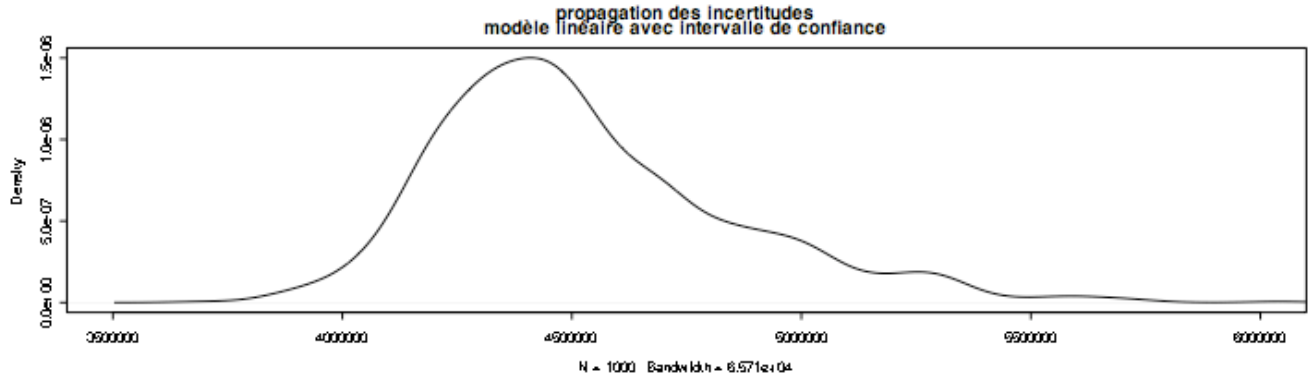
- Production à 10 ans : CP
- Facteurs : PORO, MK3, KR
- Simulateur :

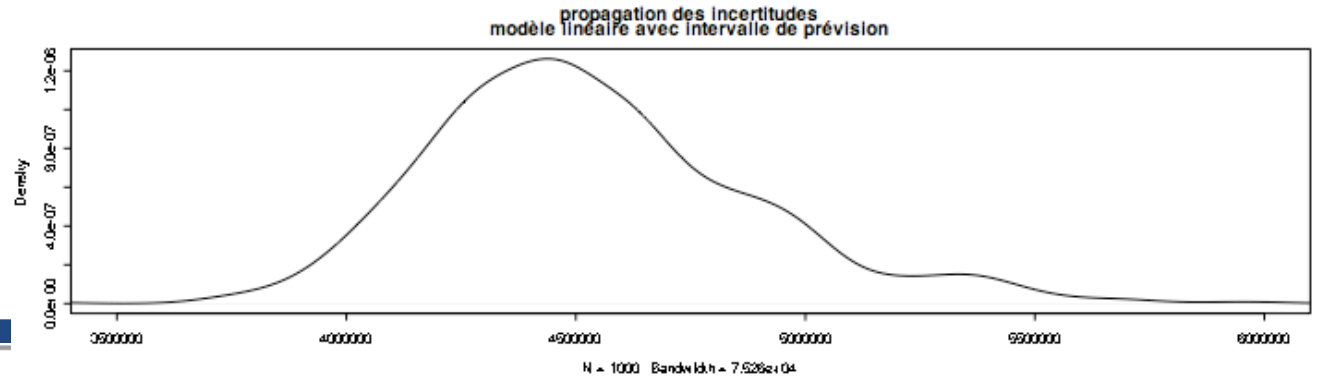
$$CP = simu(PORO, MK3, KR) = simu(x)$$

- Approximation de simu par un polynome de degré 2 : $P(x) = X(x)\,\beta$, β de taille 10.
- Estimation de la régression à partir d'un plan d'expérience X de taille 15*10

Propagation des incertitudes sous les hypothèses du modèle linéaire

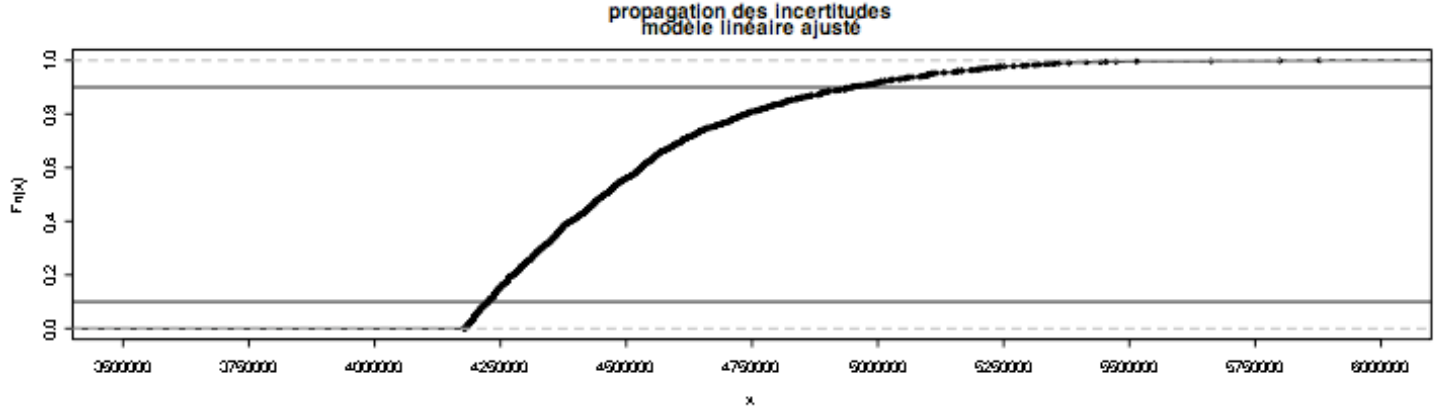


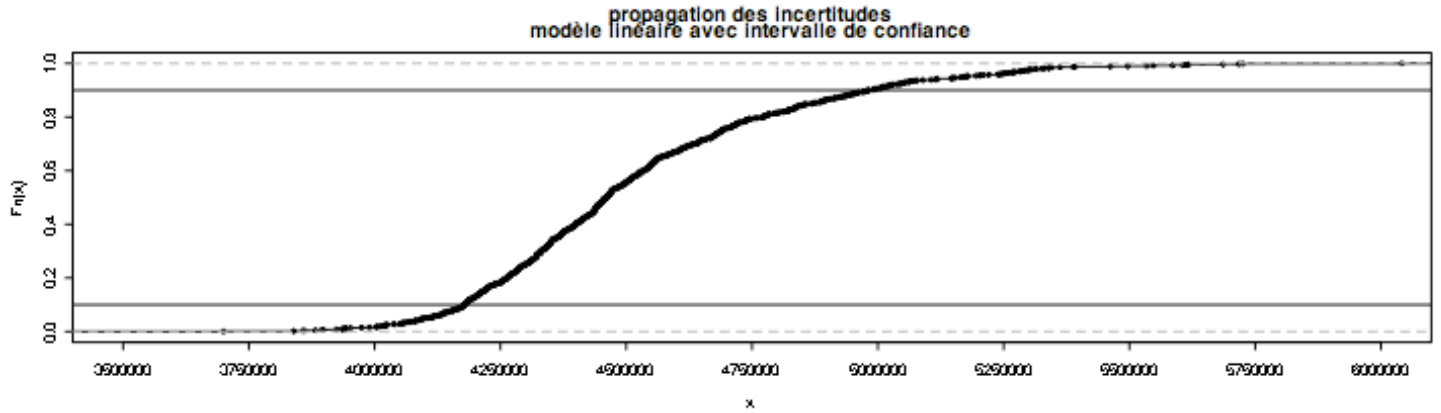


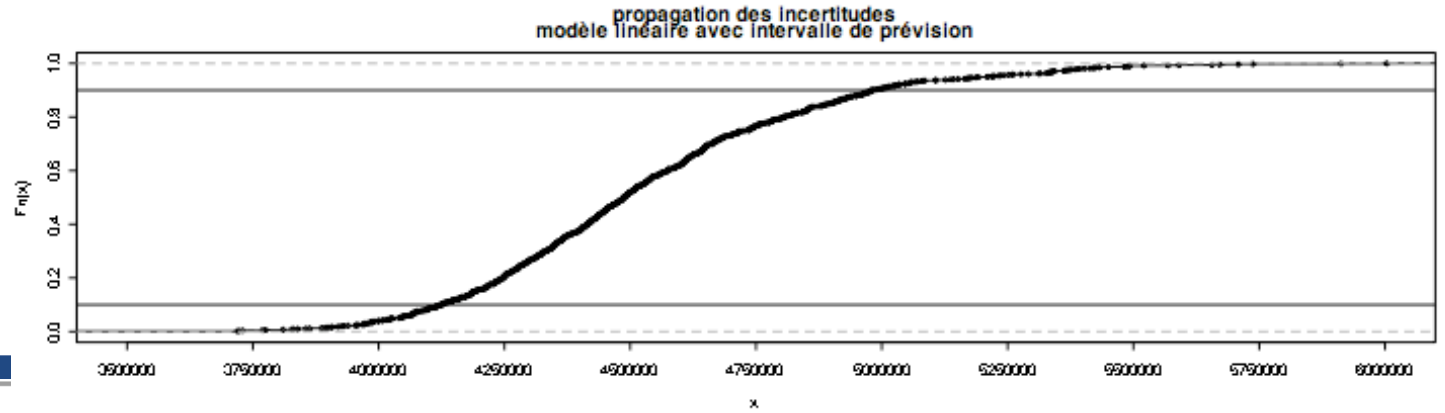


Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

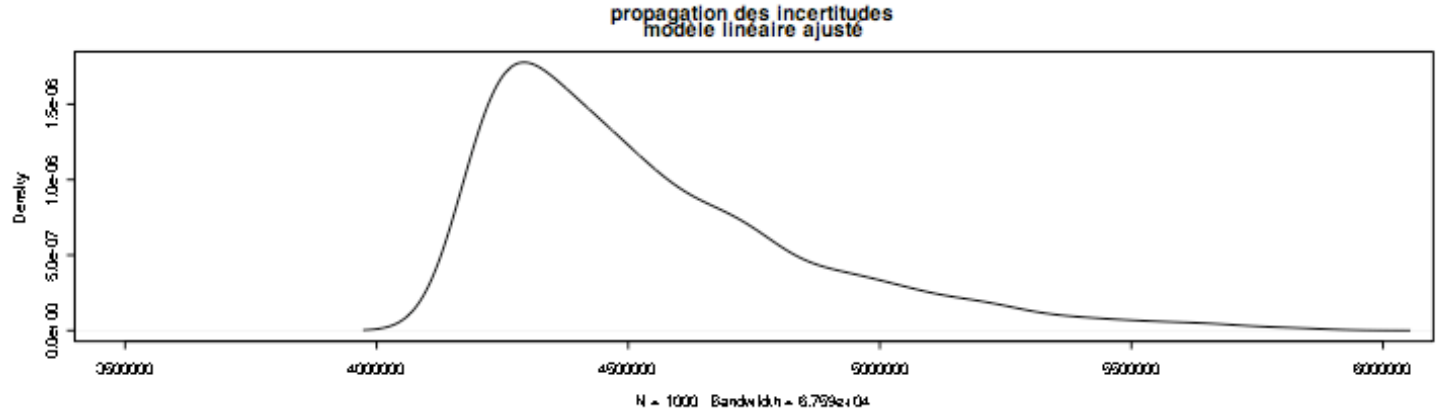
Fonctions de répartition

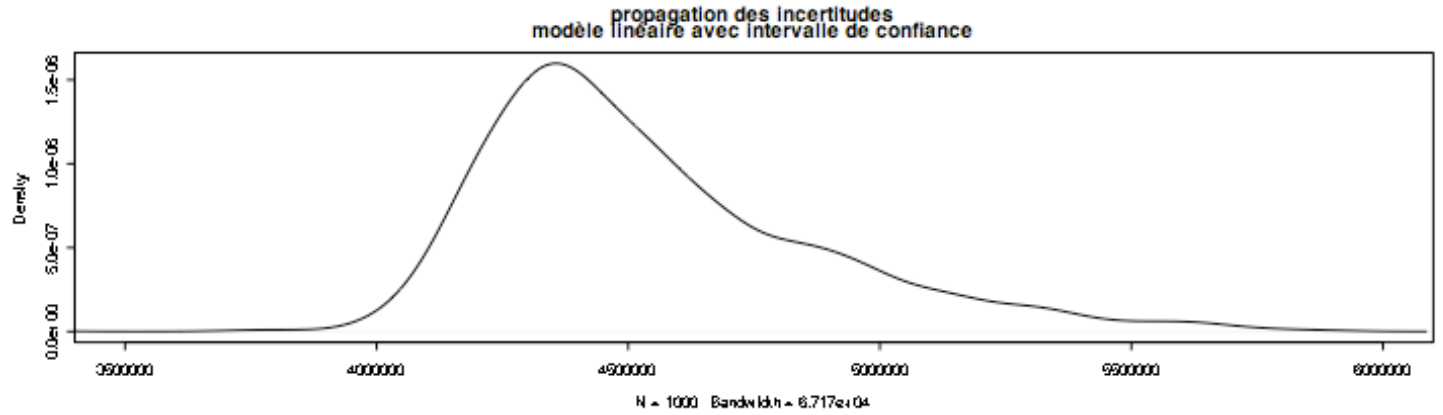


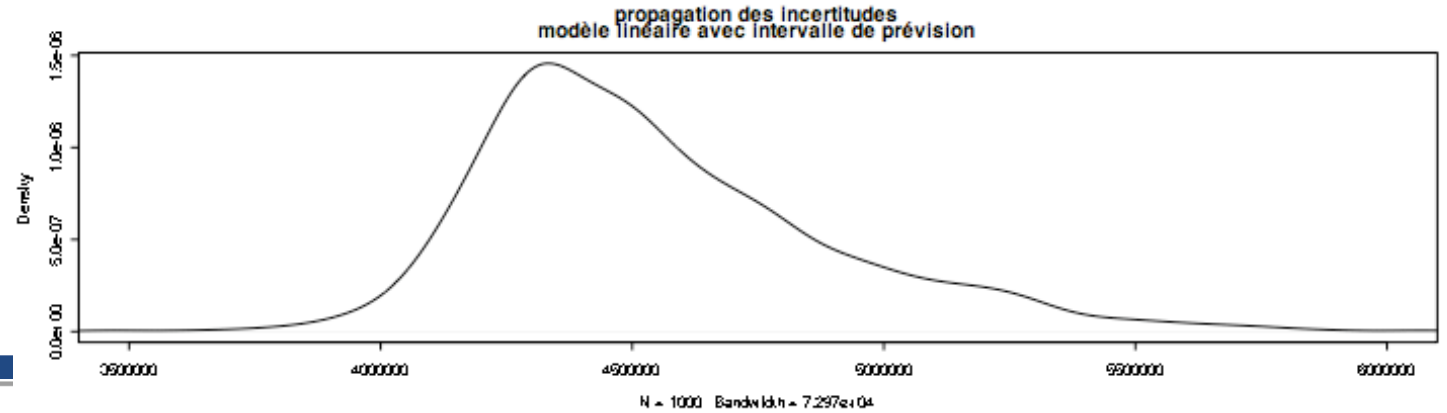




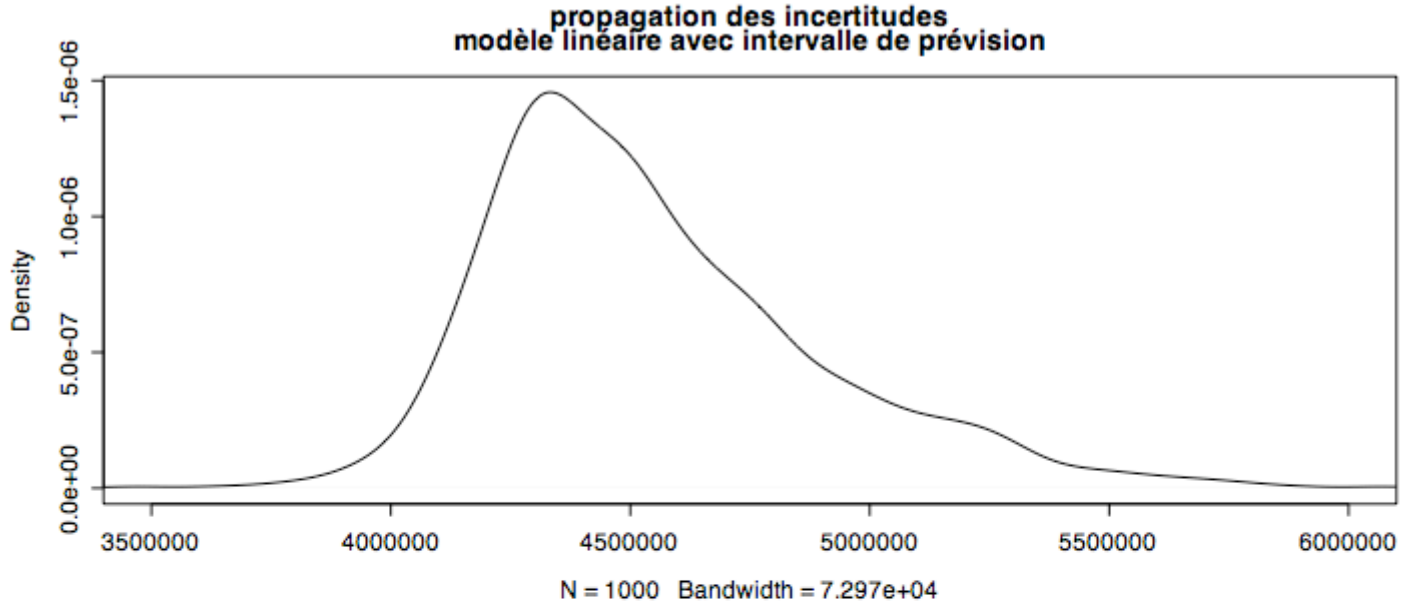
Par simulation

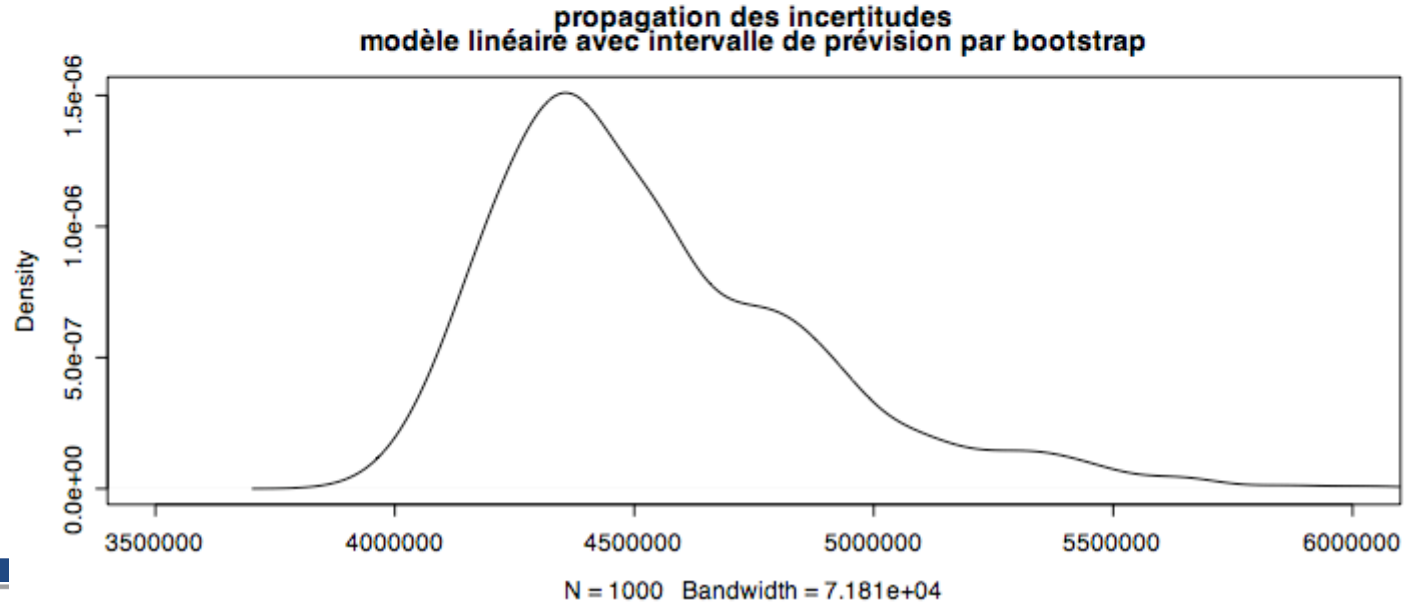






Simulation gaussienne et bootstrap





Impact du plan d'expériences