telecom Saint-Etienne

Probabilités et Statistiques

Année 2009/2010

laurent.carraro@telecom-st-etienne.fr

olivier.roustant@emse.fr

telecom Saint-Etienne

Cours n°11

Régression linéaire

1/3 Modèle - Validation

telecom Saint-Etienne

Problématique de la régression

- Expliquer une **réponse** y
  grâce à des **prédicteurs** $x_1, \ldots, x_p$

... dans un contexte incertain

... à partir d'expériences

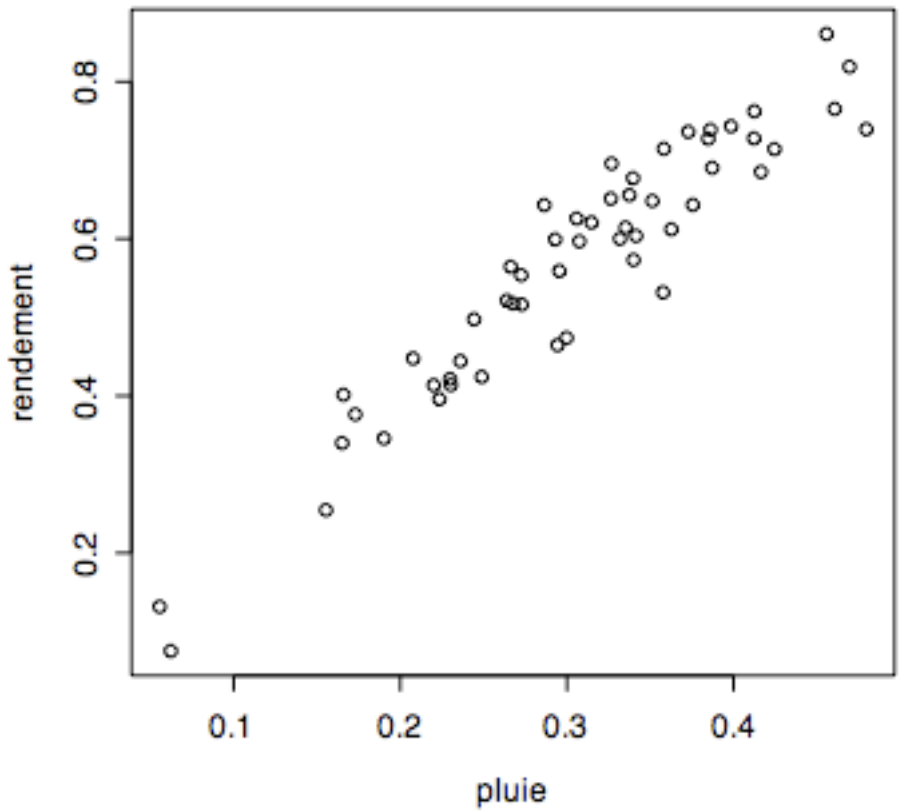
... dans un but **prédictif**

Exemple 1

- Réponse : rendement d'une culture de blé (fraction d'un rendement maximum observé)
- Prédicteur : quantité de pluies printanières (m)
- 54 observations

- Objectif : prévoir le rendement grâce à la hauteur de pluie observée

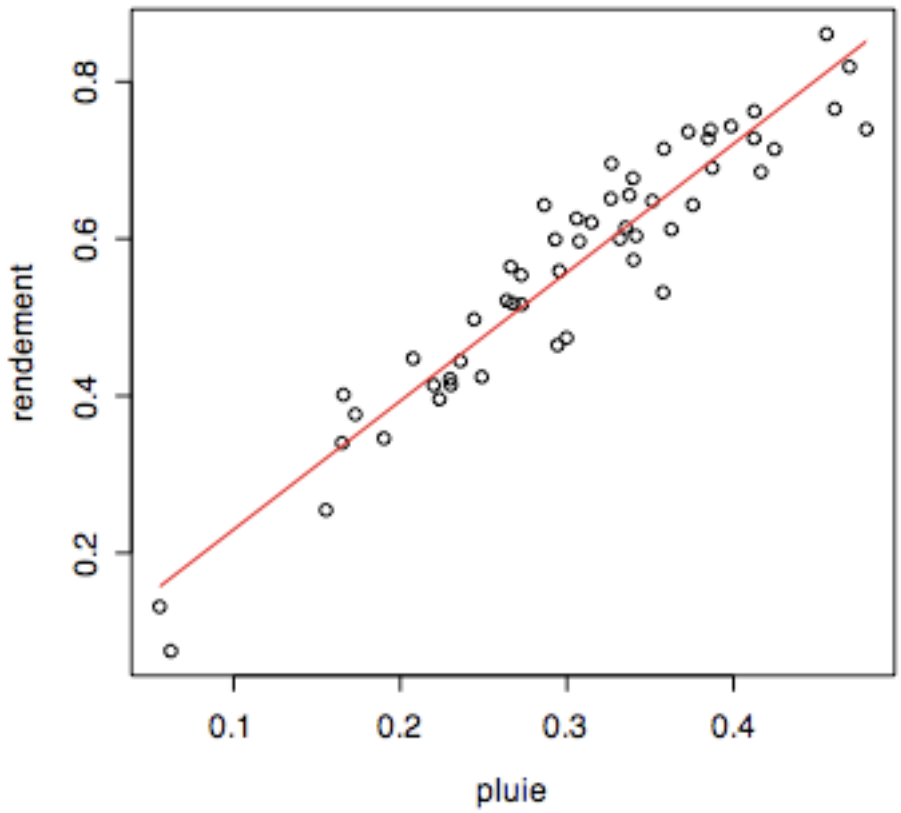
Observation des données




```
donnees <- read.table("reg_pluie.txt", dec=",", sep="\t", header=TRUE)
plot(donnees)
```

1er modèle (formulation incomplète)

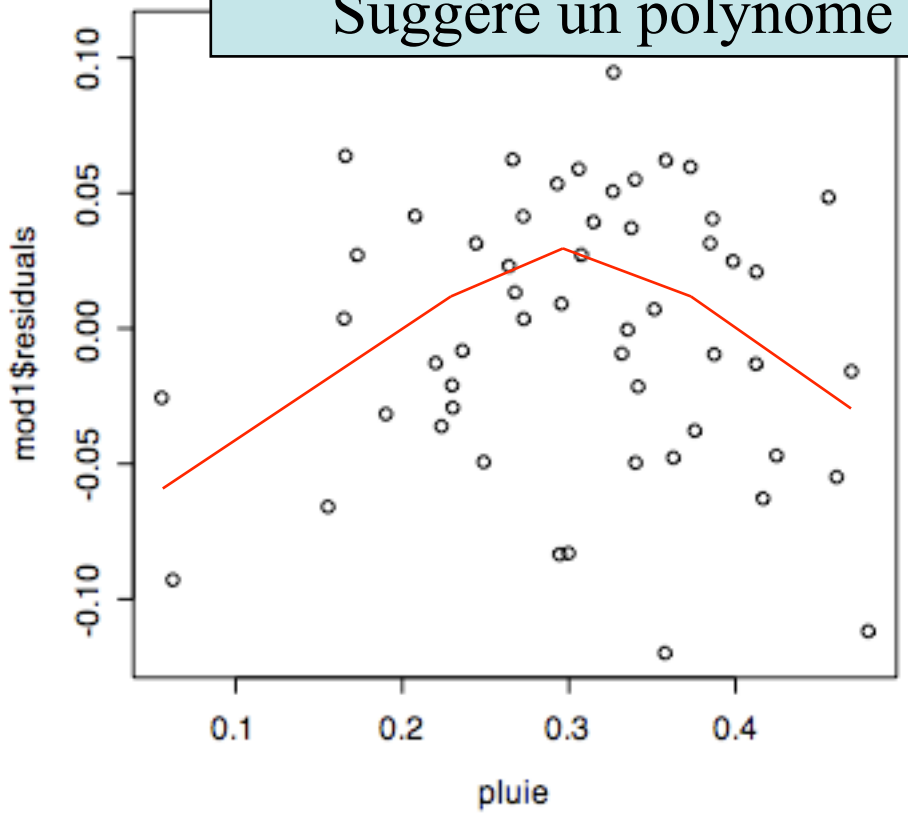
$$y_i = \beta_0 + \beta_1 x_i + e_i$$




```
mod1 <- lm(rendement~pluie, data=donnees)
plot(rendement~pluie, data=donnees)
lines(donnees$pluie, mod1$fitted.values, col="red")
```

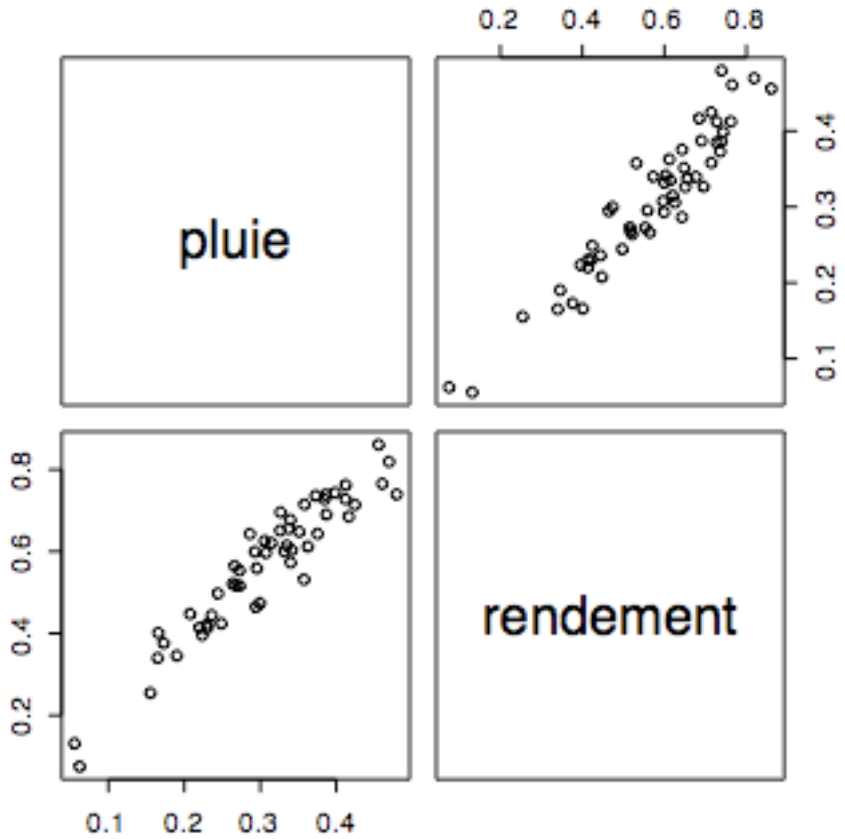
Regardons les erreurs (résidus)

Courbure quadratique des résidus :
Suggère un polynôme de degré 2




```
plot(mod1$residuals~pluie,data=donnees)
```

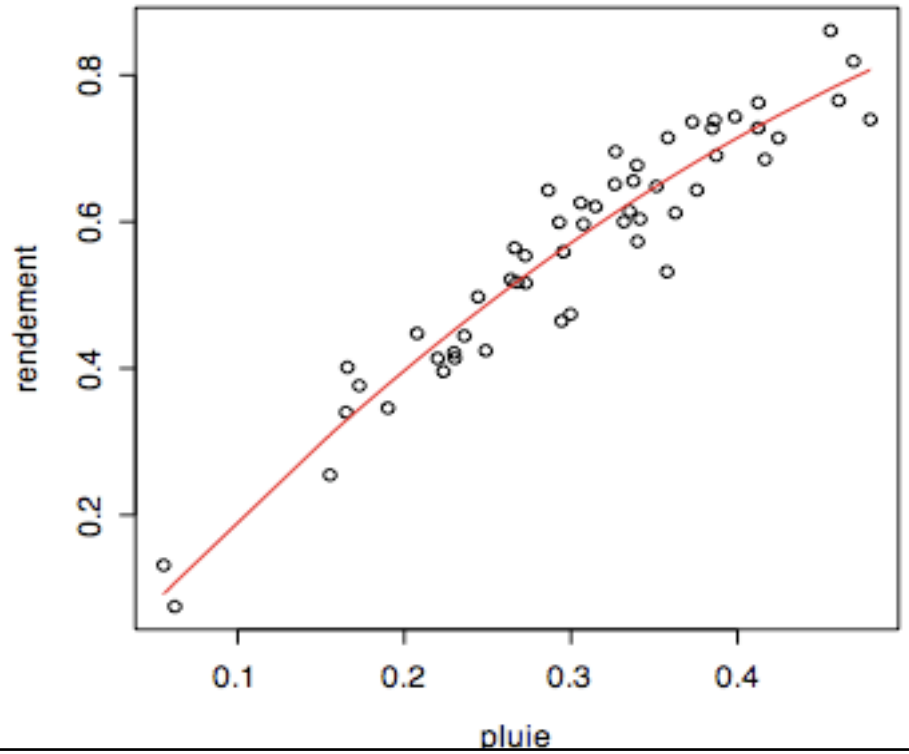
Observation des données (suite)




```
pairs(donnees)
```

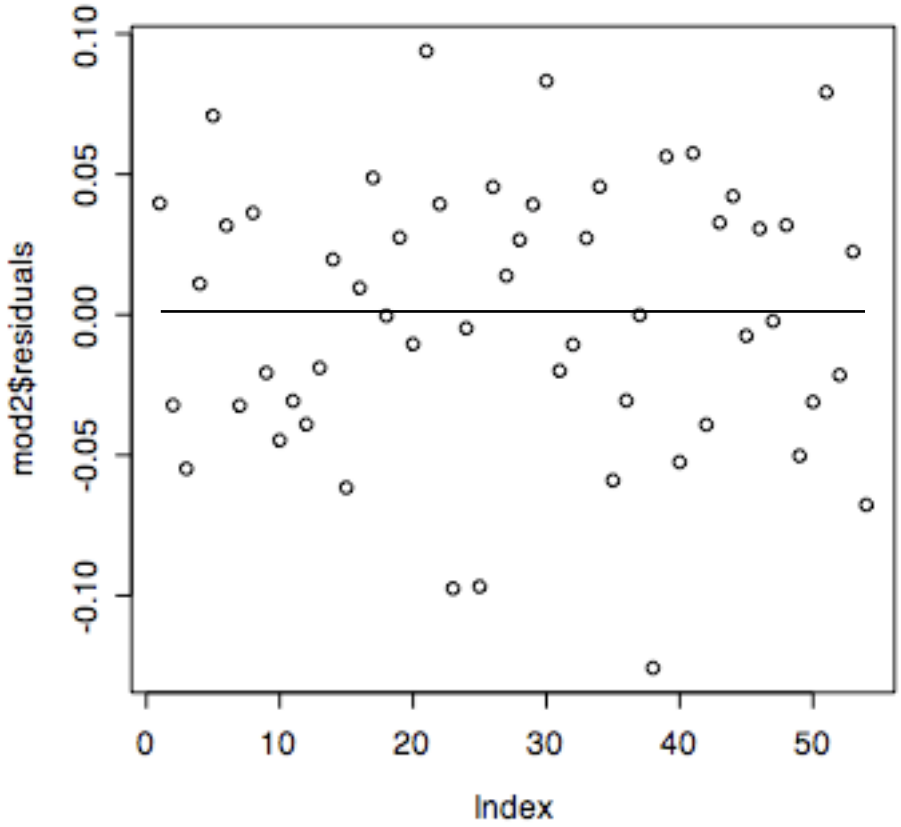
2nd modèle (form. Incomplète)

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + e_i$$



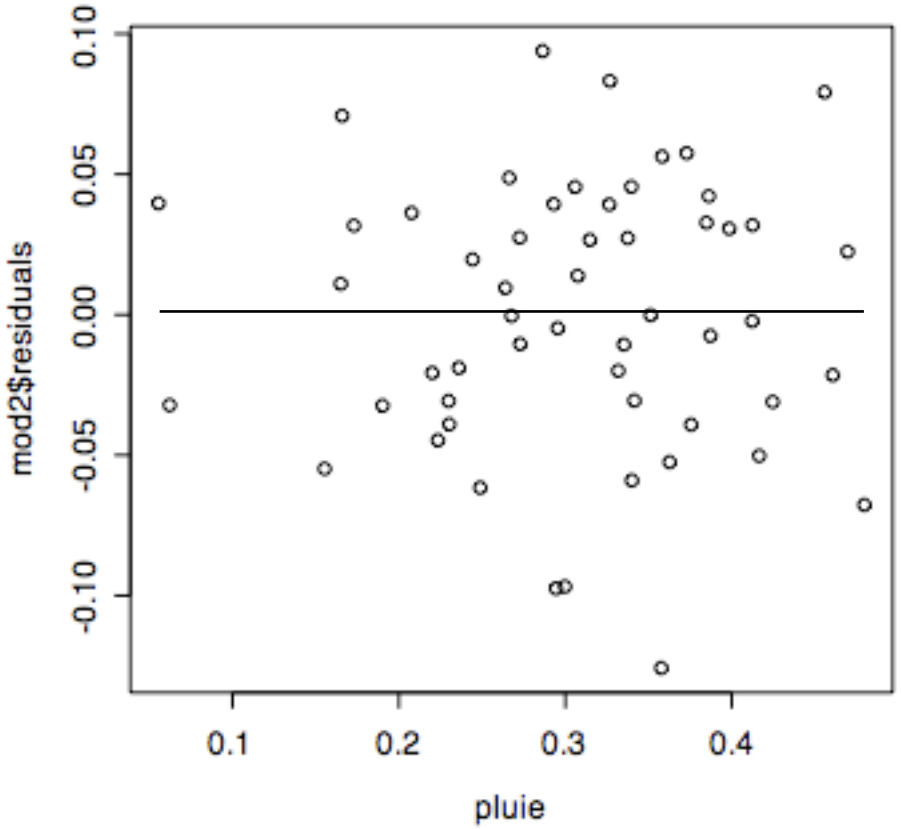

```
mod2 <- lm(rendement~pluie+I(pluie^2), data=donnees)
plot(rendement~pluie, data=donnees)
lines(donnees$pluie, mod2$fitted.values, col="red")
```

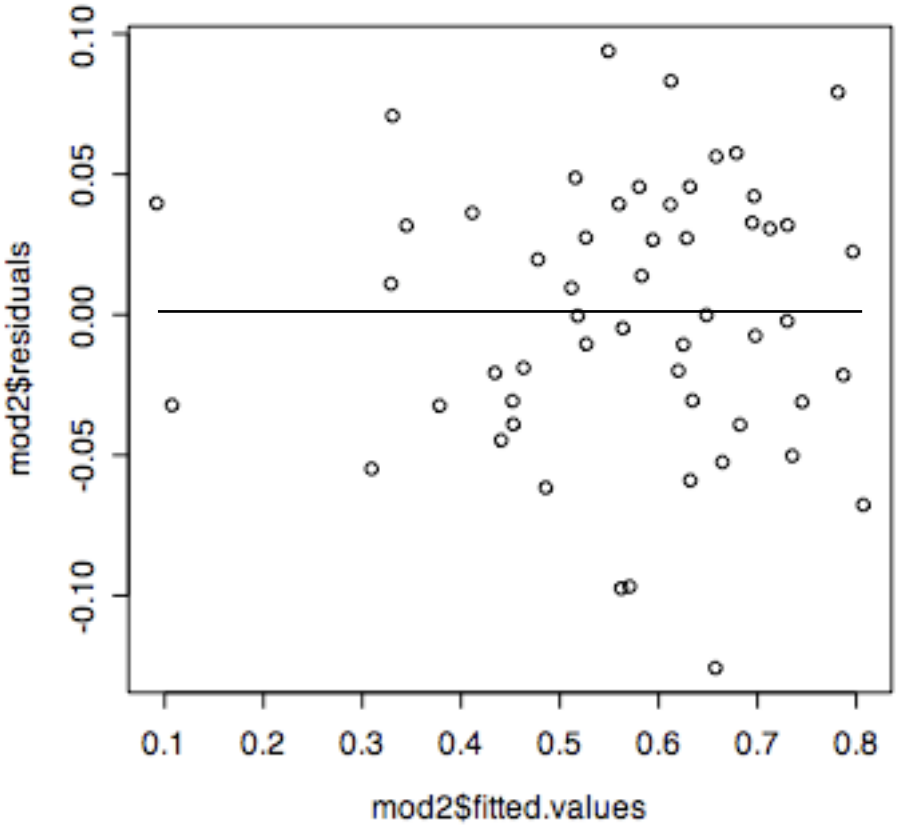
Résidus 2nd modèle




```
plot(mod2$residuals)
```

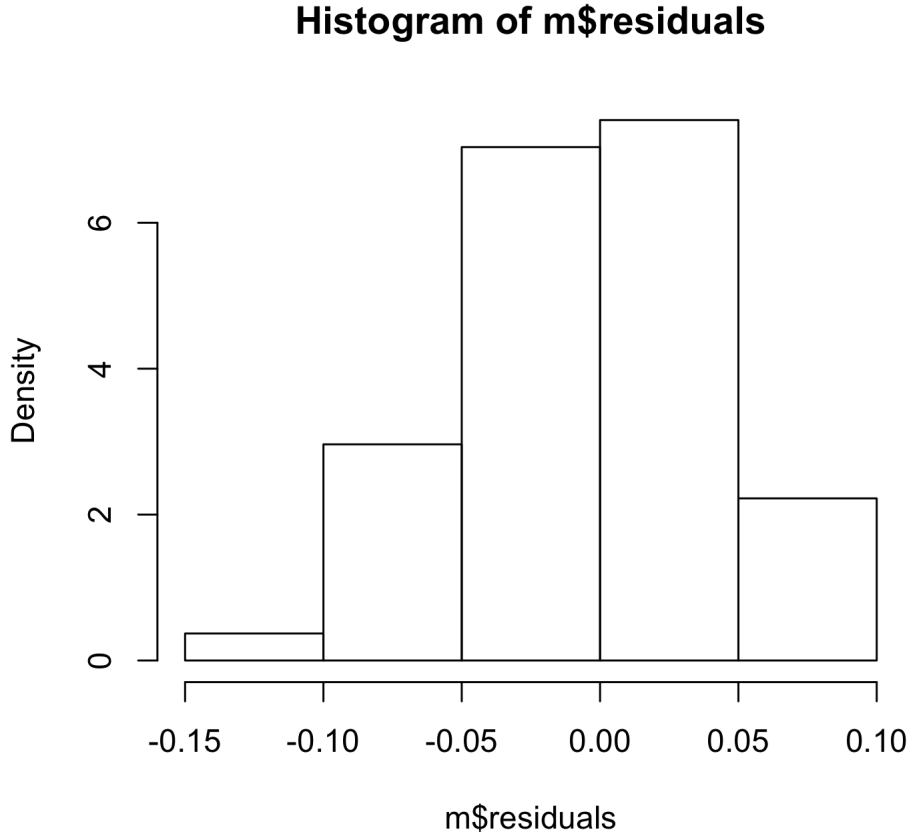
Résidus 2nd modèle (suite)

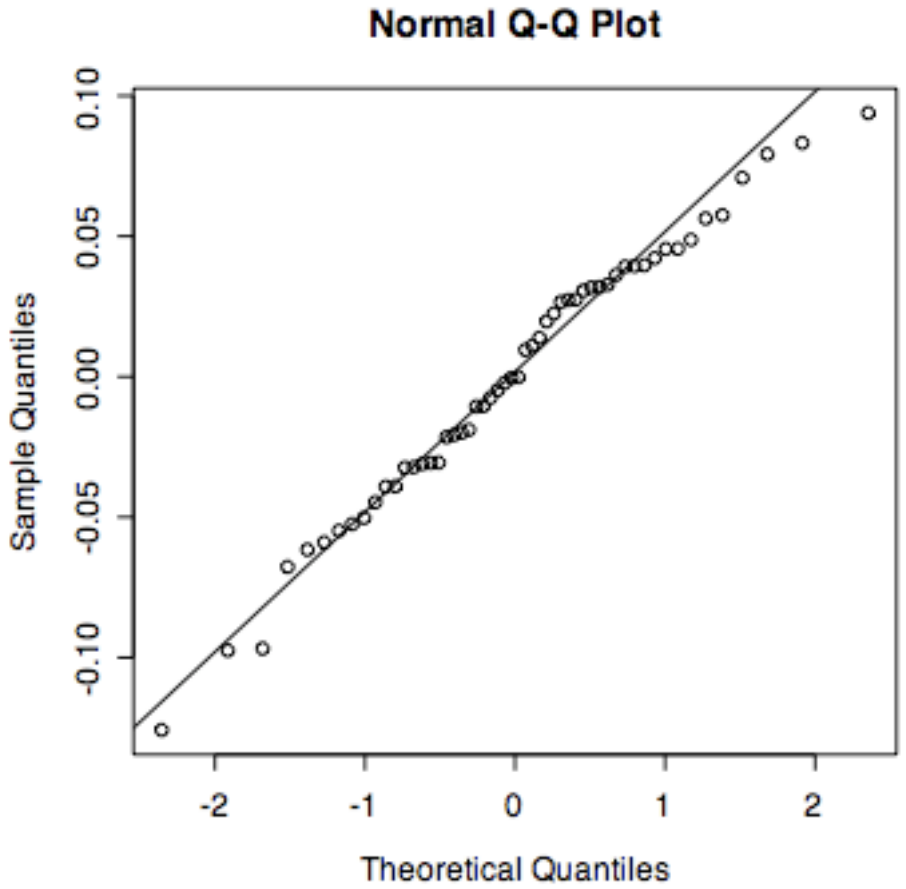





```
plot(mod2$residuals~pluie,data=donnees)
plot(mod2$residuals~mod2$fitted.values, data=donnees)
```

Loi des résidus ?



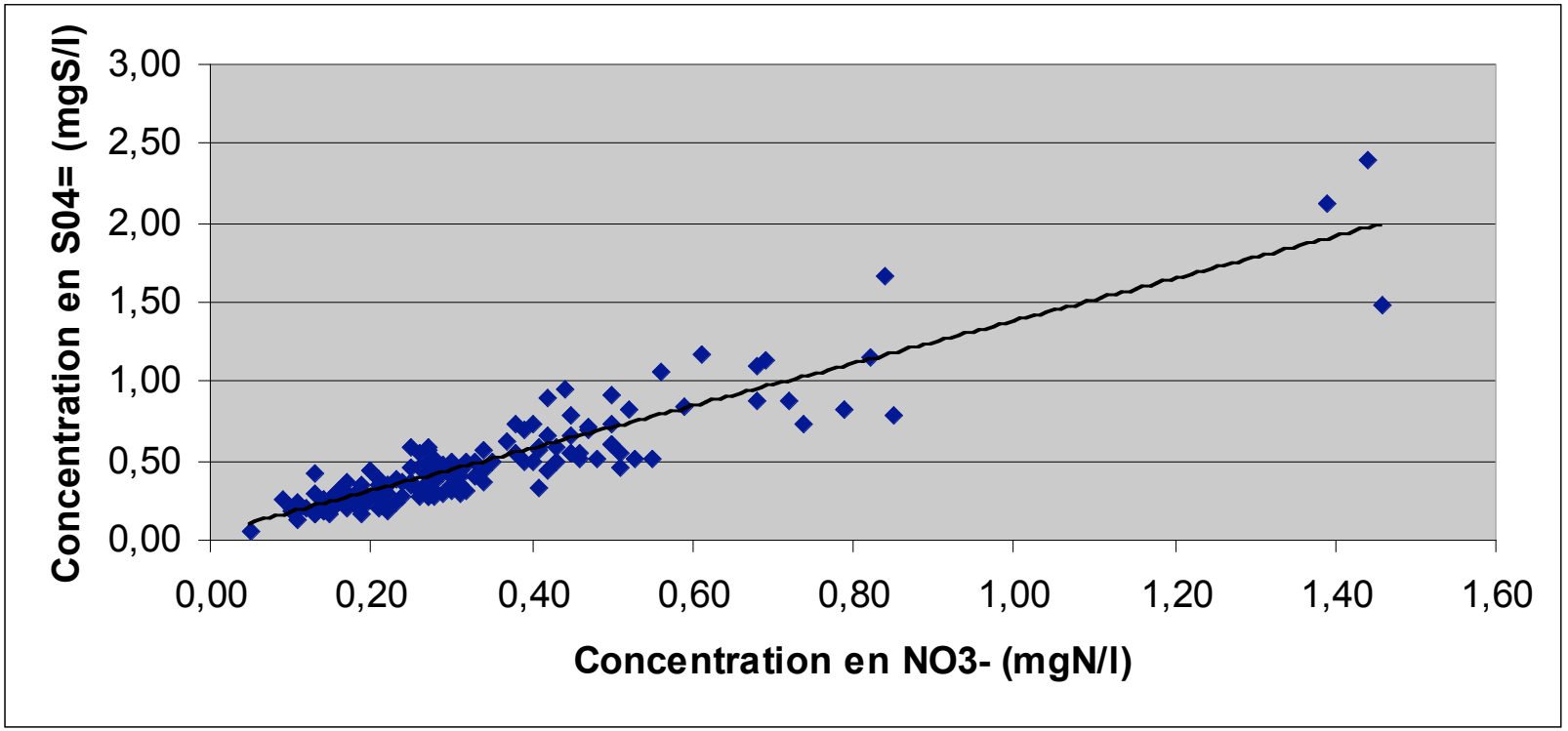



```
qqnorm(mod2$residuals); qqline(mod2$residuals)
```

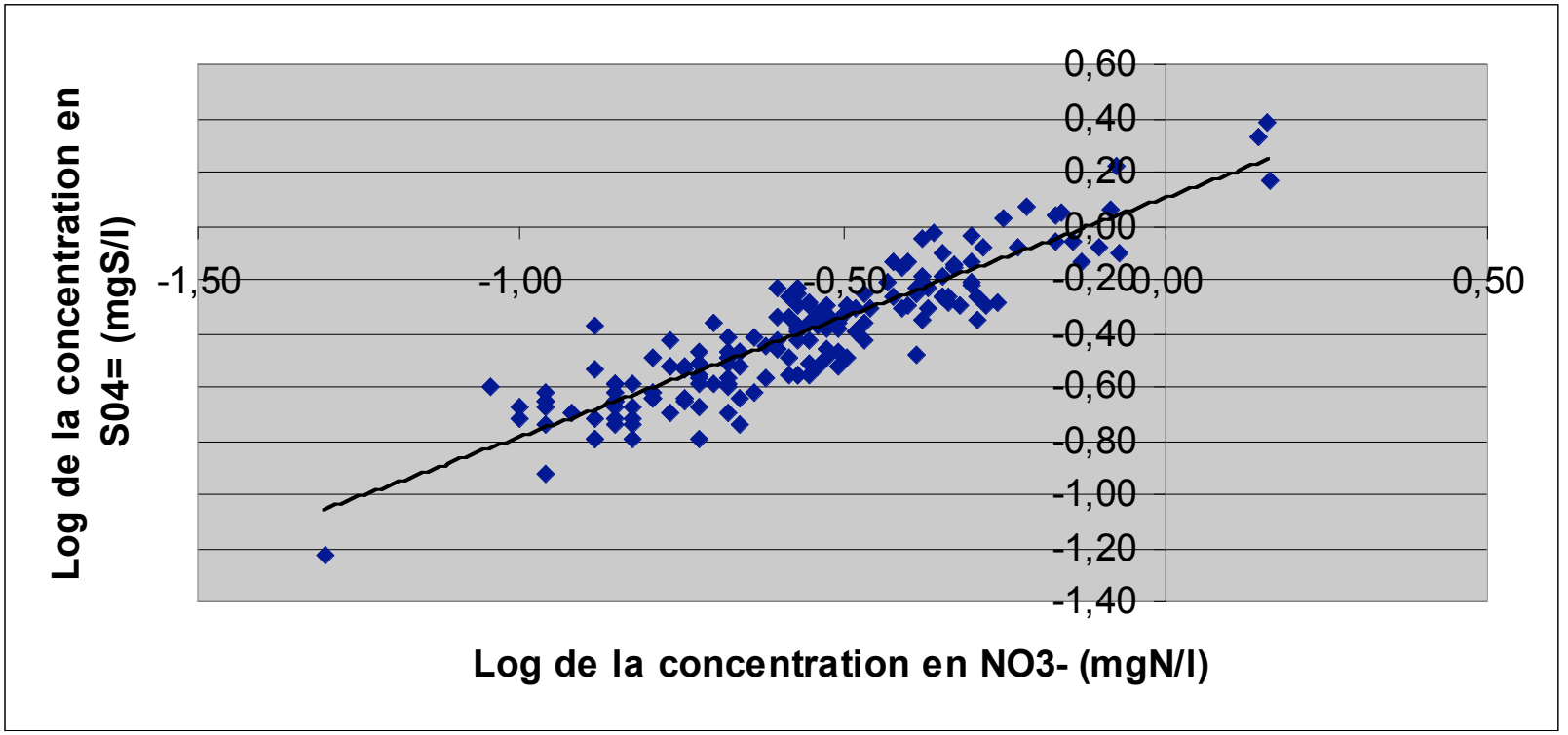
Exemple 2

- Réponse : concentration en ions SO4 --
- Prédicteur : concentration en ions NO3-

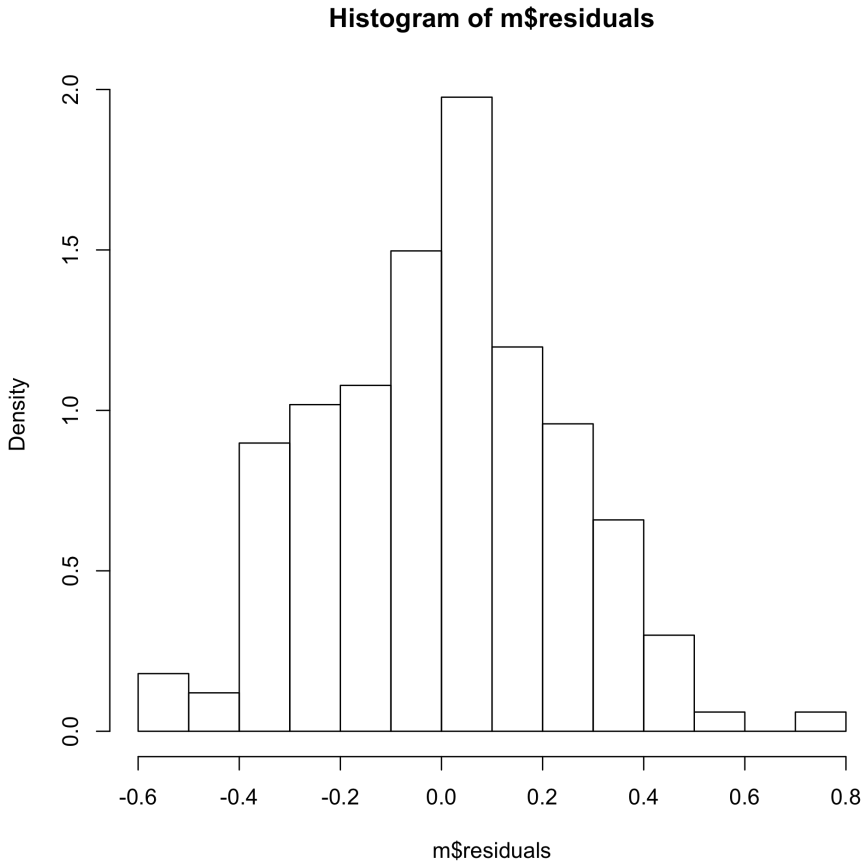
- Objectif : prévoir la concentration en ions SO4-- connaissant celle en ions NO3-

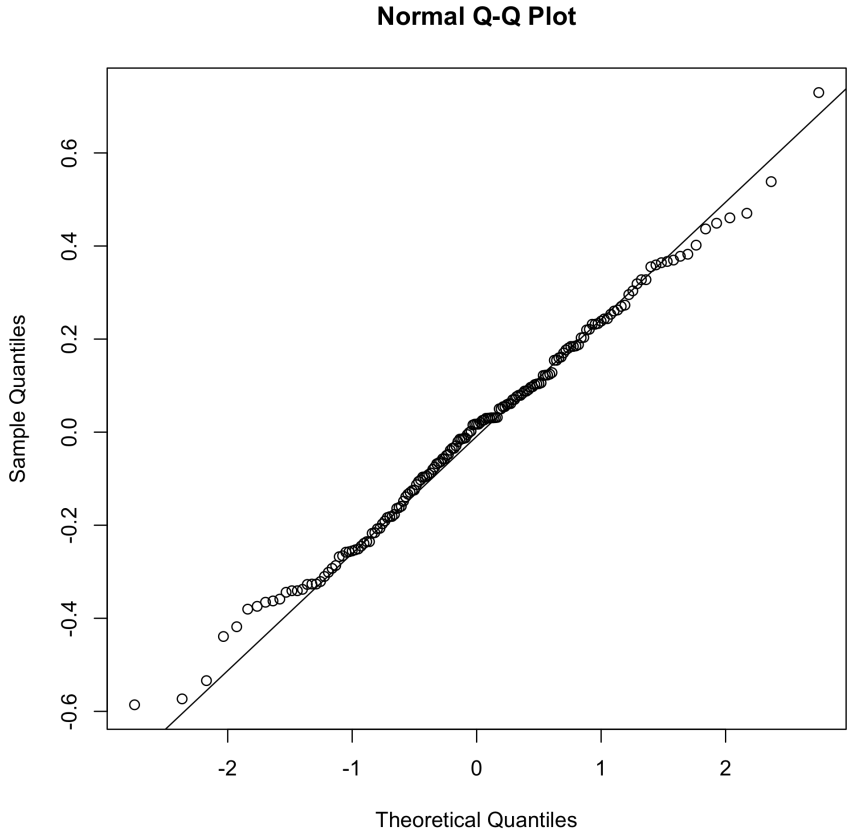

Concentration en S04= (mgS/l)
3,00
2,50
2,00
1,50
1,00
0,50
0,00
0,00
0,20
0,40
0,60
0,80
1,00
1,20
1,40
1,60
Concentration en NO3- (mgN/l)

Résolution du problème d'hétéroscédasticité



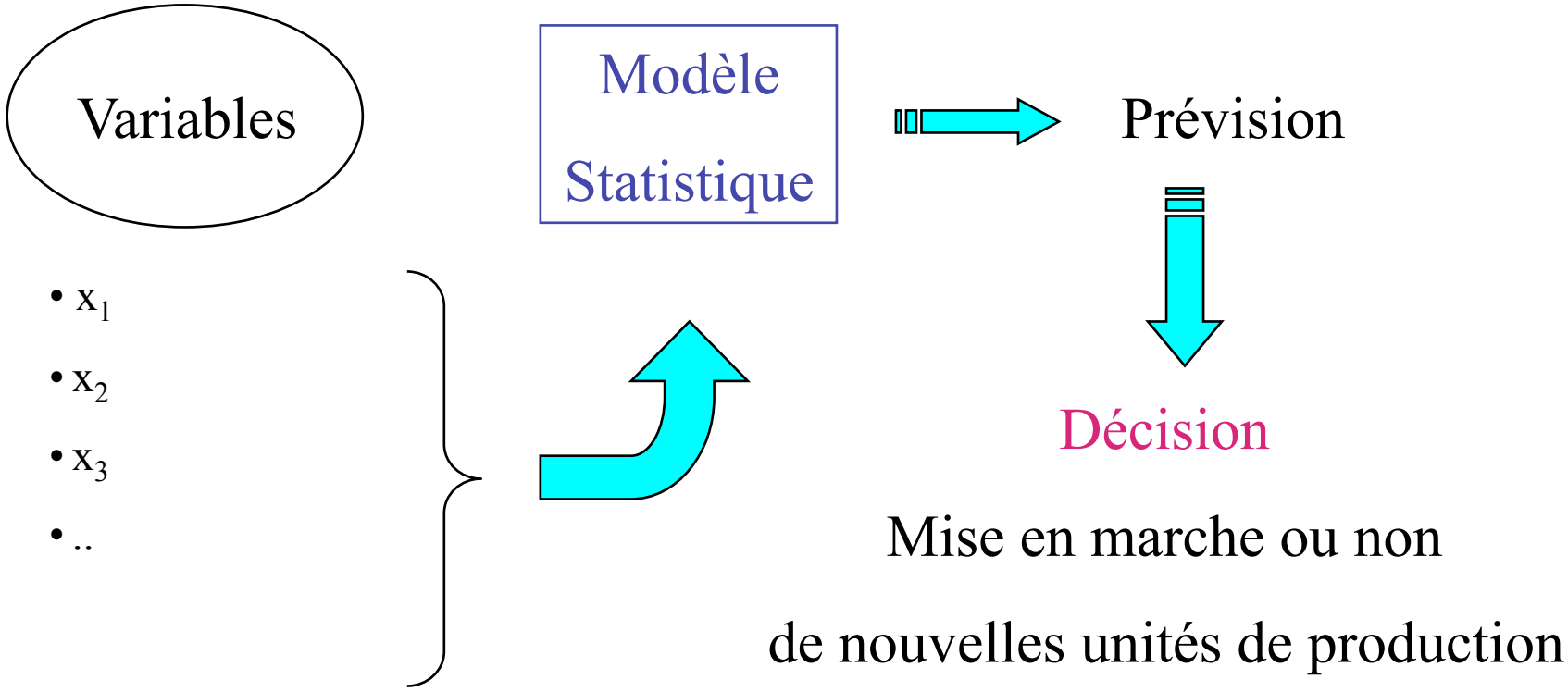
Loi des résidus ?



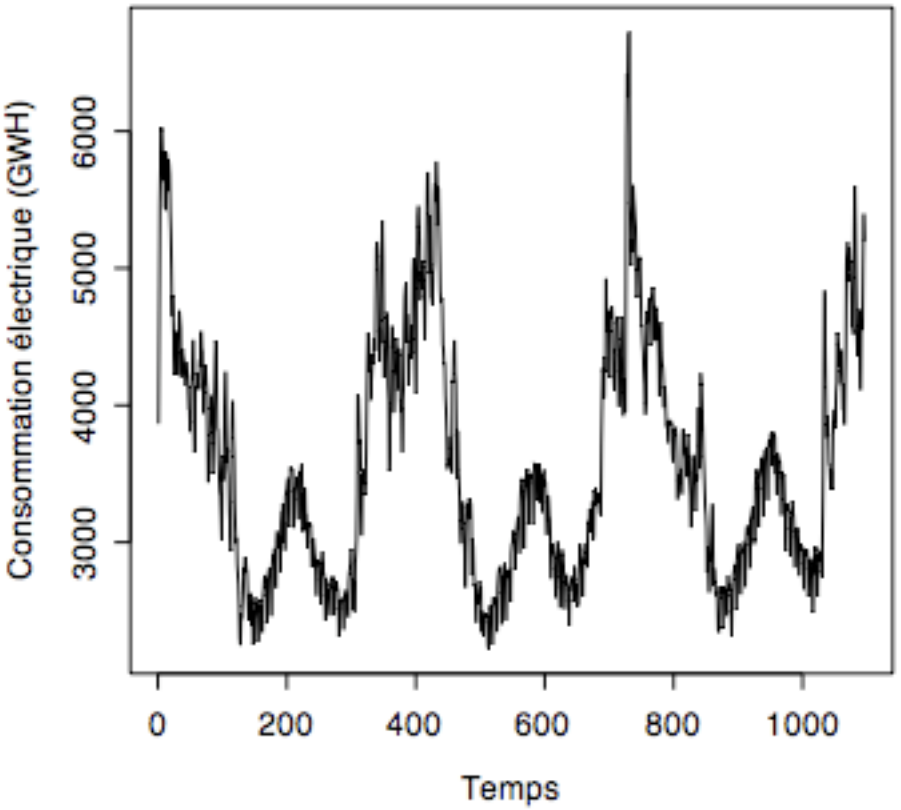


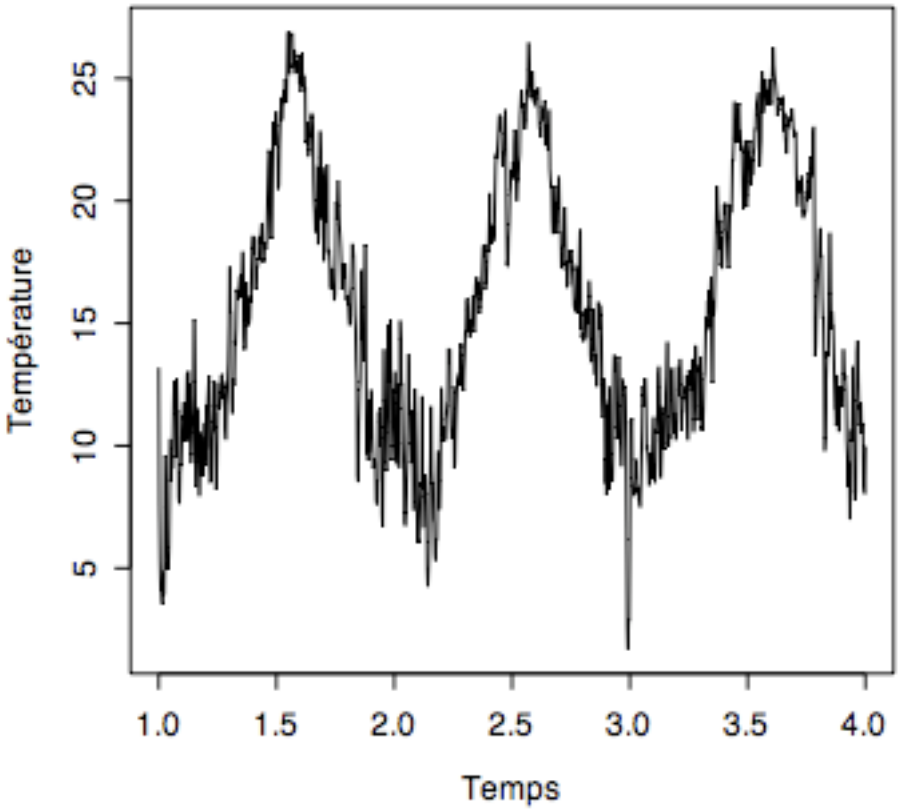
Exemple 3 : entreprise d'énergie

- But : prévoir la consommation électrique

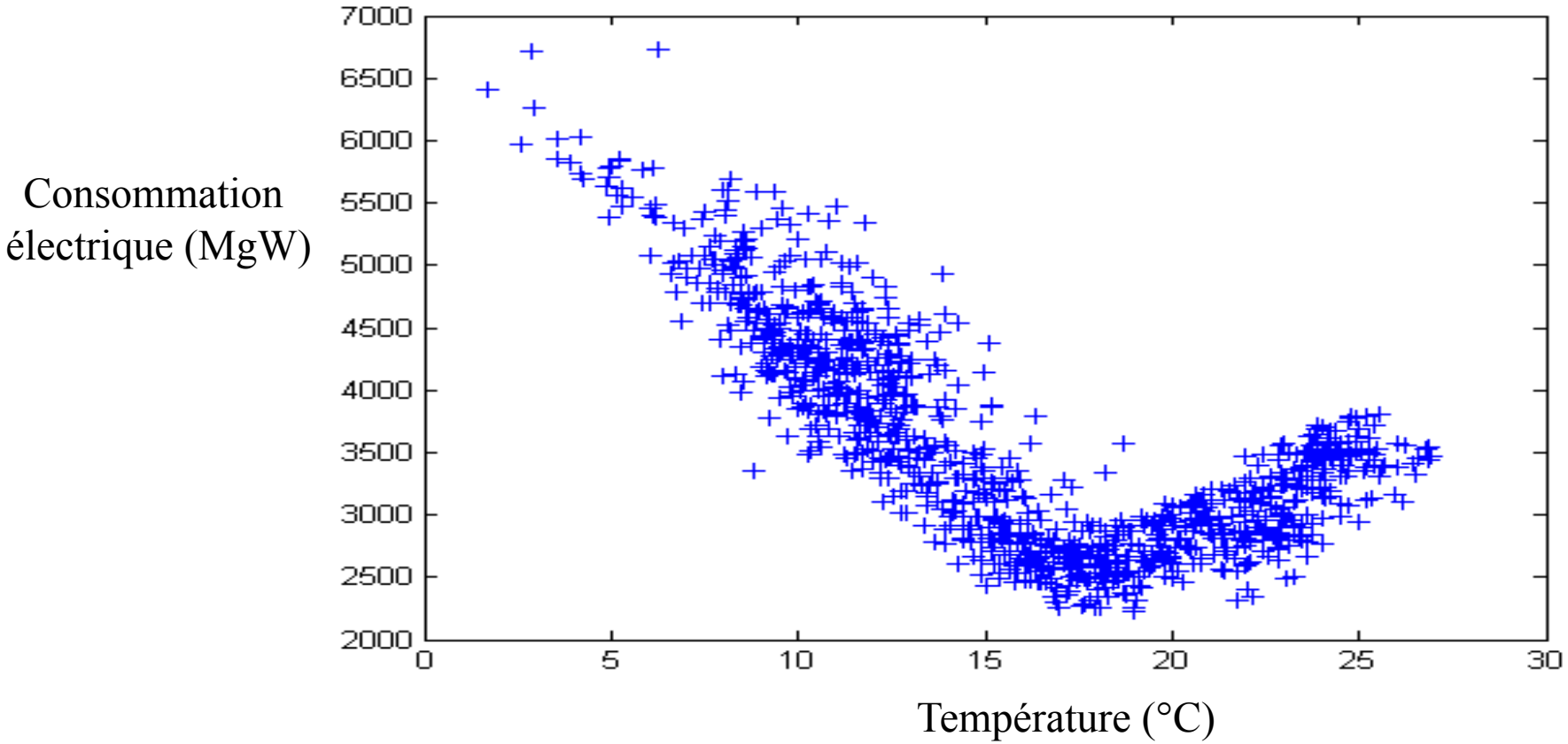


Modélisation de la consommation électrique
Un prédicteur privilégié : la température





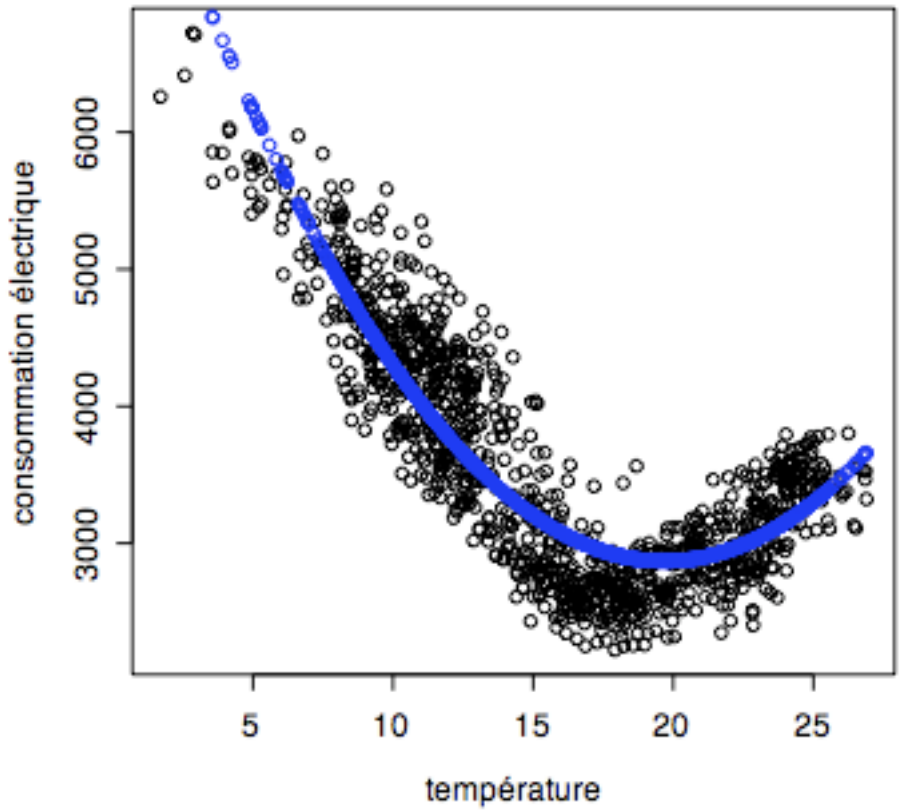
Modélisation de la consommation électrique
Dépendance non linéaire

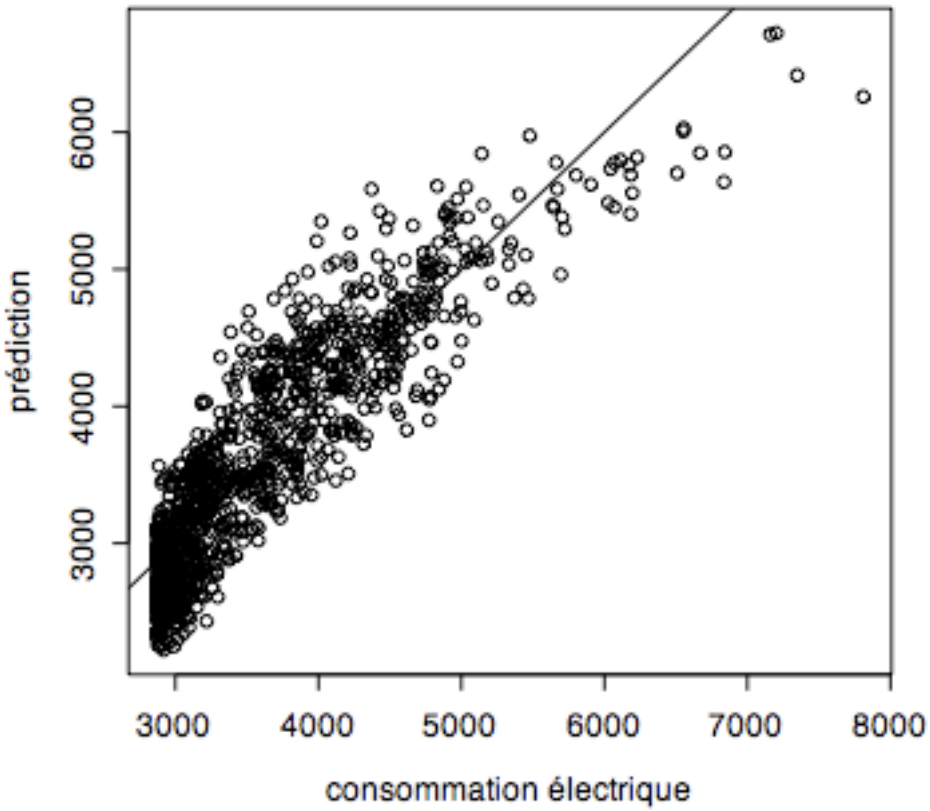


Modélisation de la consommation électrique

Dépendance non linéaire, mais modèle linéaire !

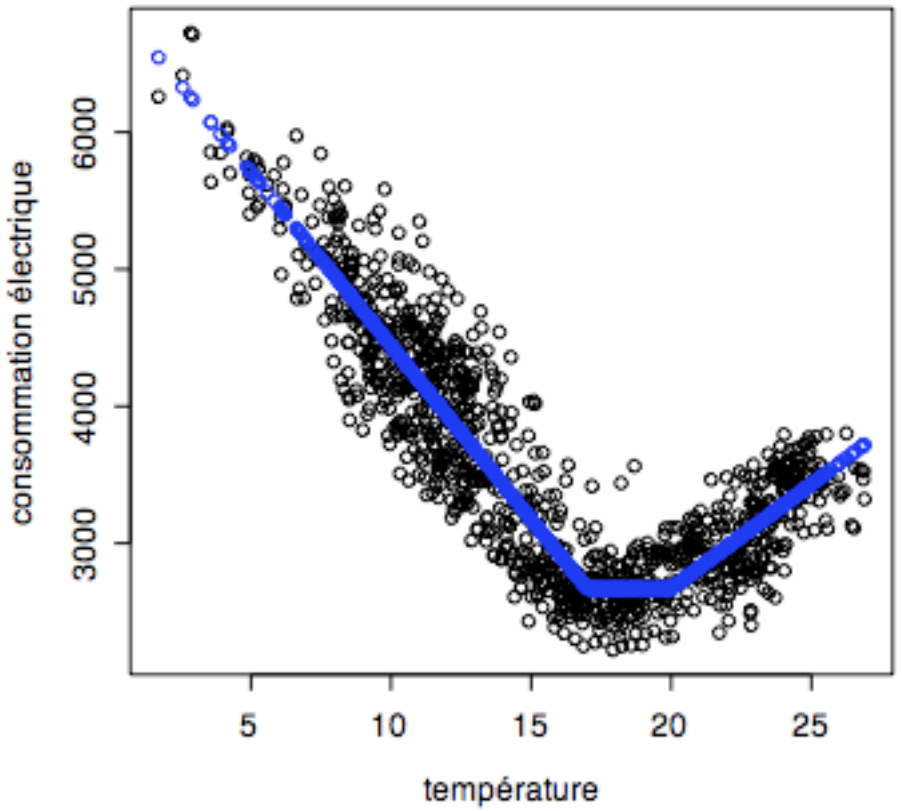
1er modèle : $C(t) = b_0 + b_1 T(t) + b_2 T(t)^2 + e(t)$

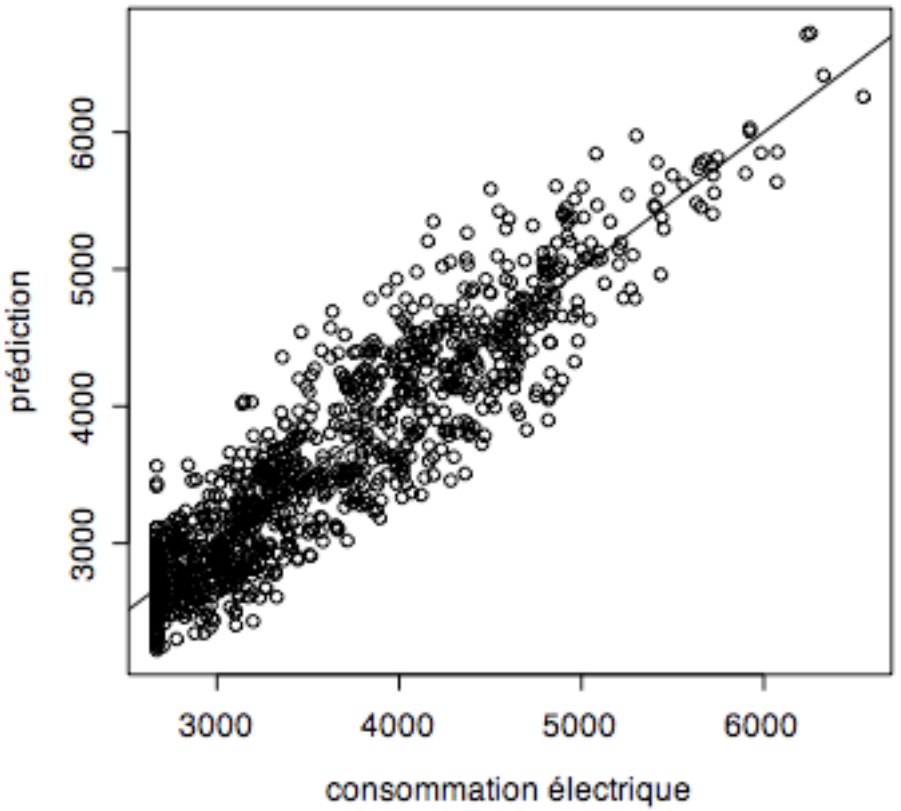




Modélisation de la consommation électrique

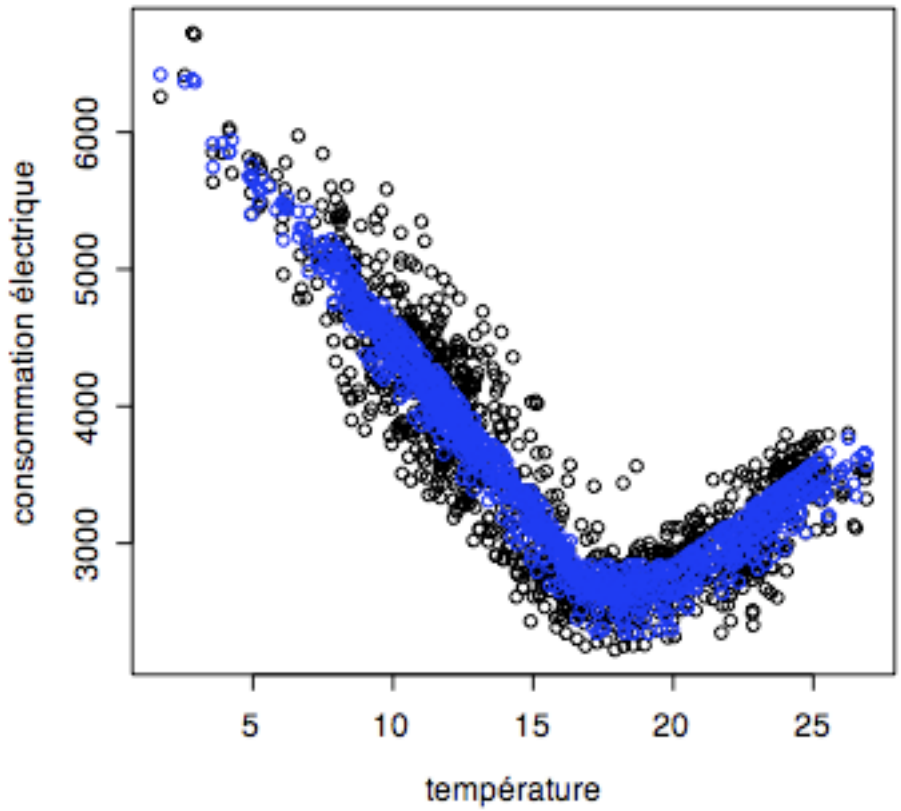
2ème modèle : $C(t) = b_0 + b_1(17\text{-}T(t))_+ + b_2(T(t)\text{-}20)_+ + e(t)$

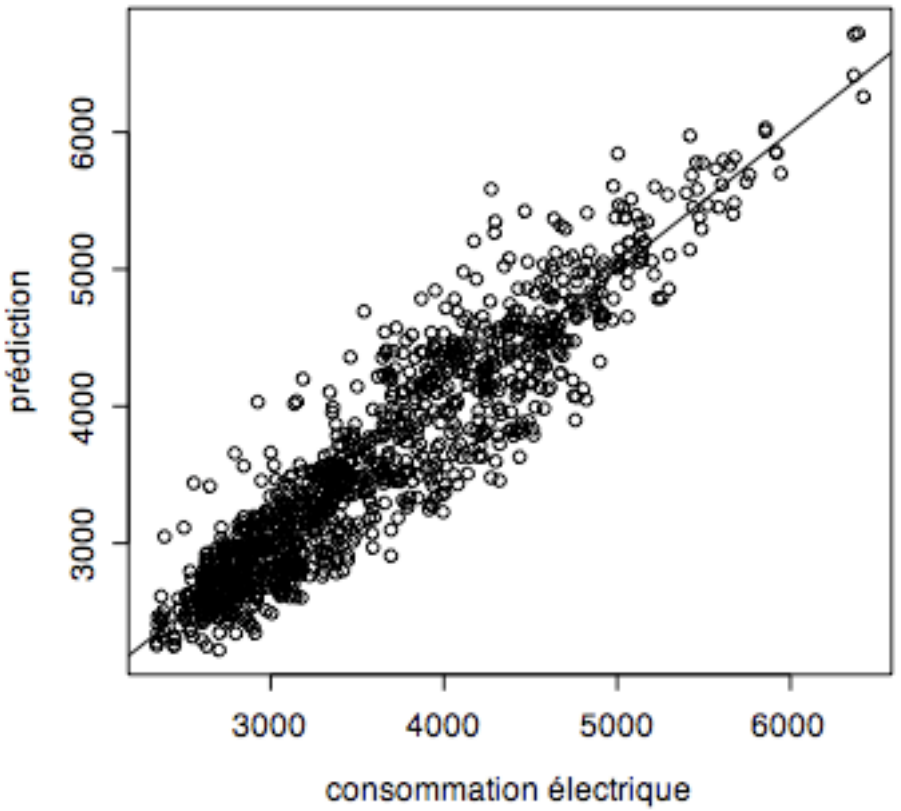




Modélisation de la consommation électrique

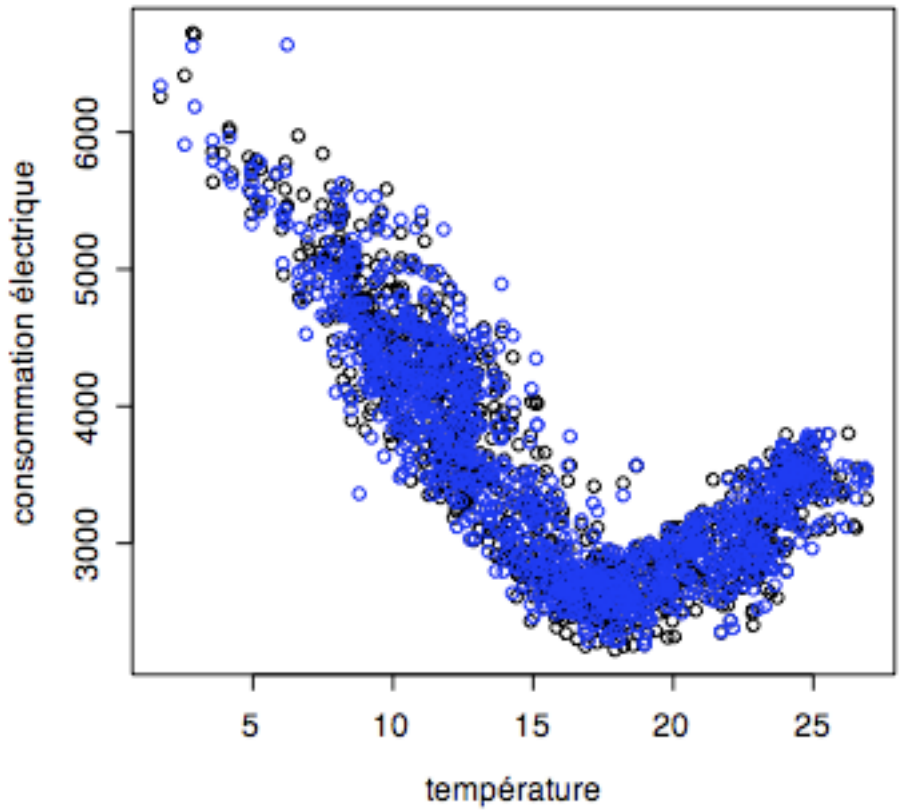
Modèle linéaire incluant t, T(t), jour de la semaine

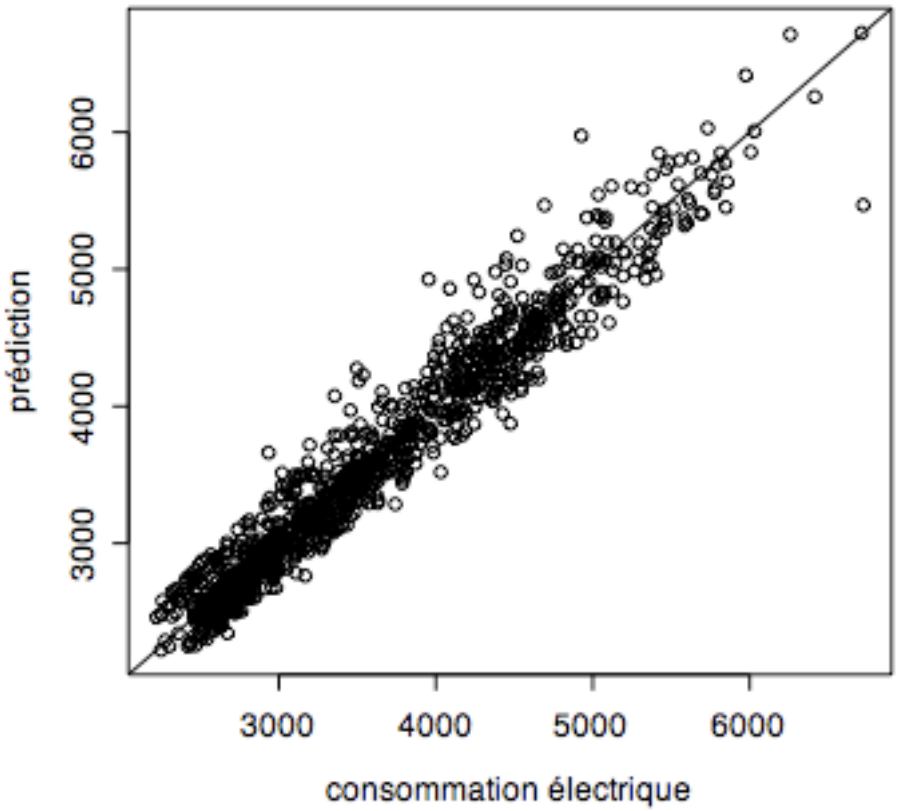




Modélisation de la consommation électrique

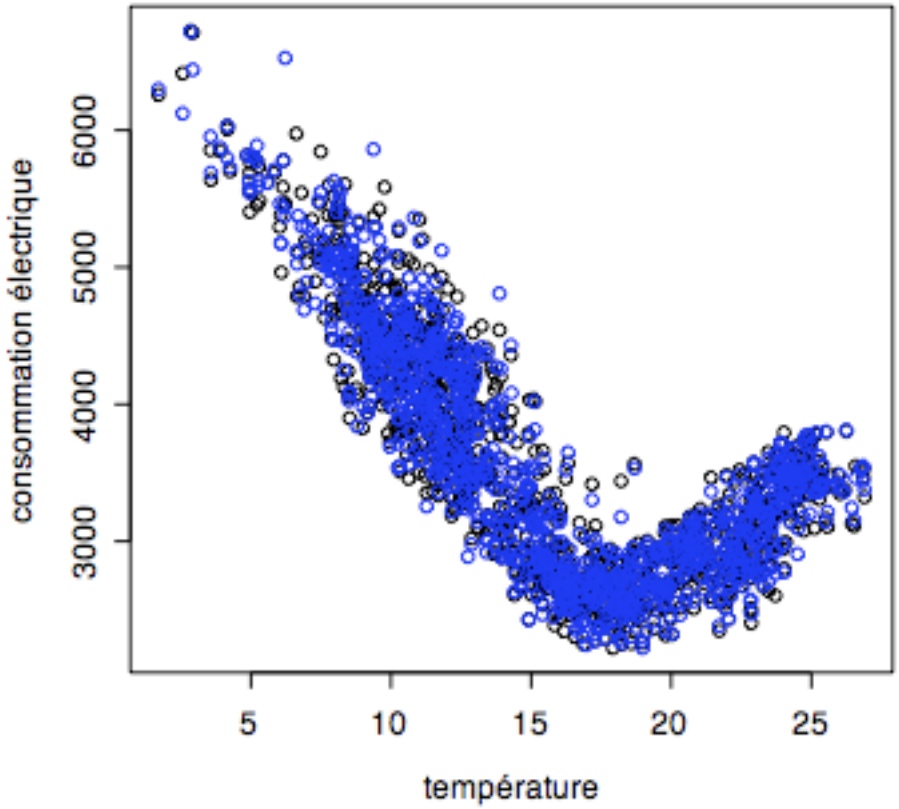
Modèle temporel $C(t) = b_0 + b_1 C(t-1) + e(t)$

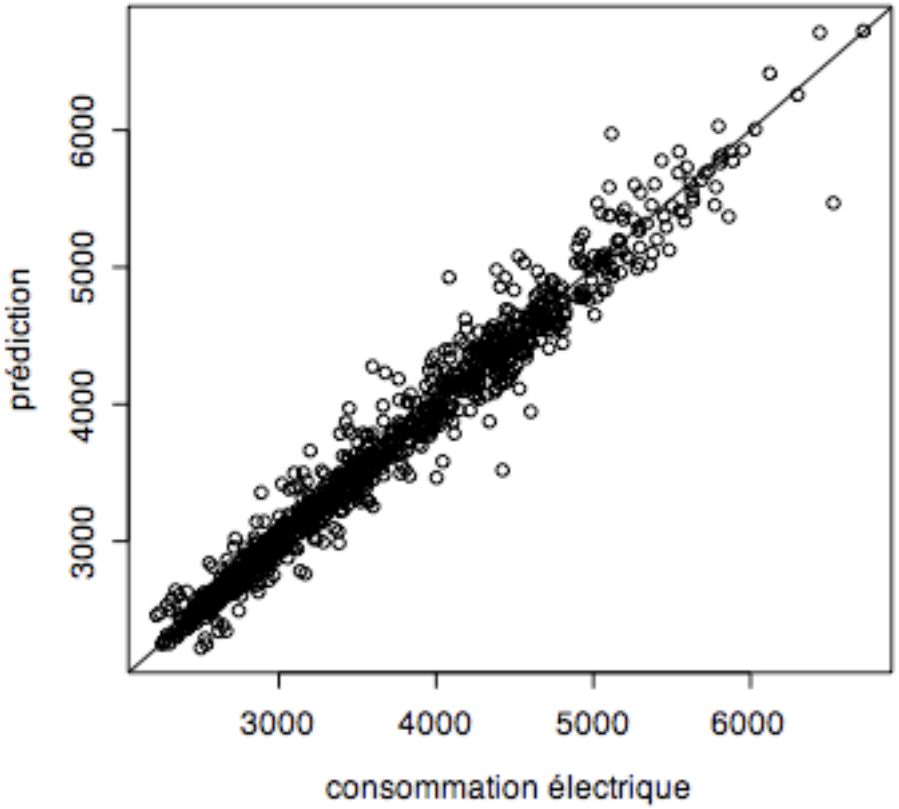




Modélisation de la consommation électrique

1 modèle mixte régression / temporel (mais pas le meilleur !)





Modélisation statistique : démarche

- Travail préliminaire
  - Etude des données
  - Choix des prédicteurs

  ⇒ proposition d'un modèle de régression
  - Plusieurs modèles possibles

- Etude statistique
  - Estimation, analyse des résultats
  - Validation du modèle : vérification des hypothèses
  - Si non validé, retour à la 1ère étape !

Dans ce cours…

- On se limite à 1 ou 2 prédicteurs
- On n'étudiera pas de données temporelles sauf si elles sont indépendantes dans le temps
  - Nécessite des connaissances en séries temporelles

Définition du modèle linéaire

- Y vecteur des réponses
- X matrice du plan d'expériences
- β vecteurs des paramètres
- ε vecteurs des écarts au modèle :

$$Y = X\beta + \varepsilon$$

$\varepsilon_1, \ldots, \varepsilon_n$ indépendants et de même loi $N(0, \sigma^2)$

Ecriture développée :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$$

telecom Saint-Etienne

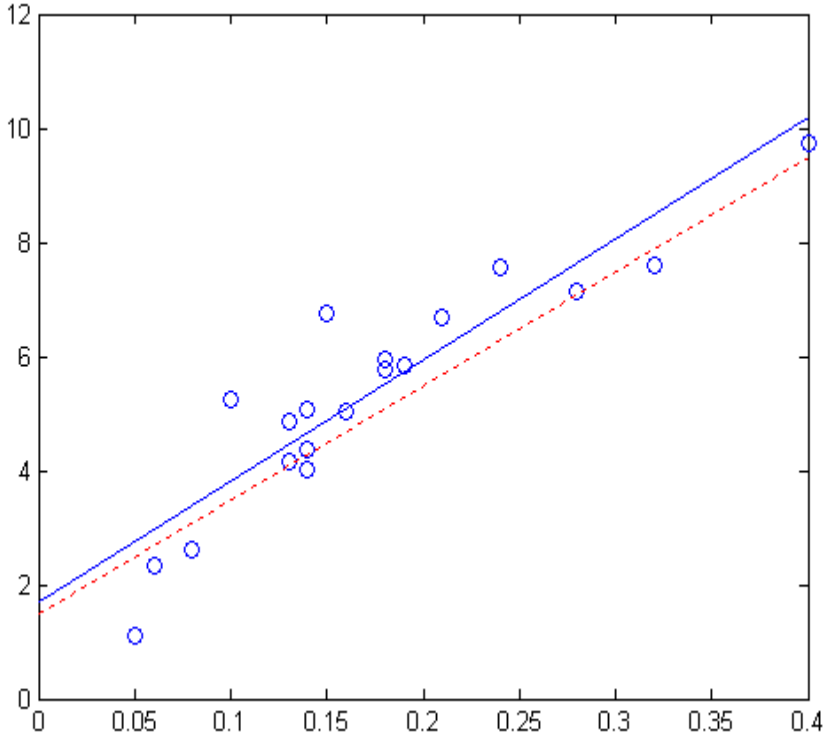
Exercice

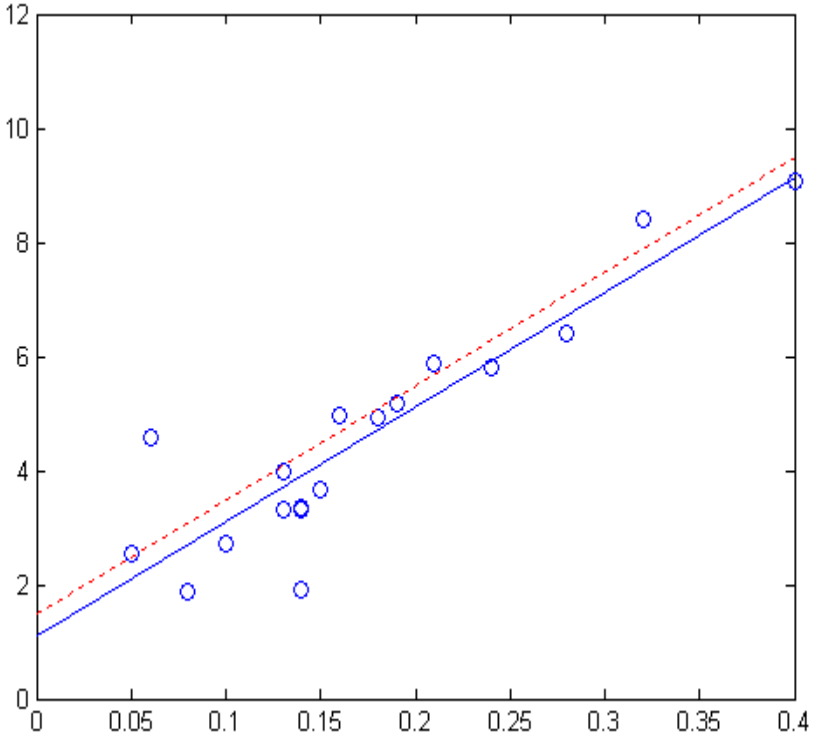
On suppose que $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ sont des v.a. i.i.d de loi $N(0, \sigma^2)$. Dans quels cas peut-on se ramener à un modèle de régression linéaire ?

1. $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$
2. $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$
3. $y_i = \beta_0 \exp(\beta_1 x_i) \times |\varepsilon_i|$
4. $y_i = \beta_0 / (1 + \beta_1 x_i) + \varepsilon_i$

Interprétation

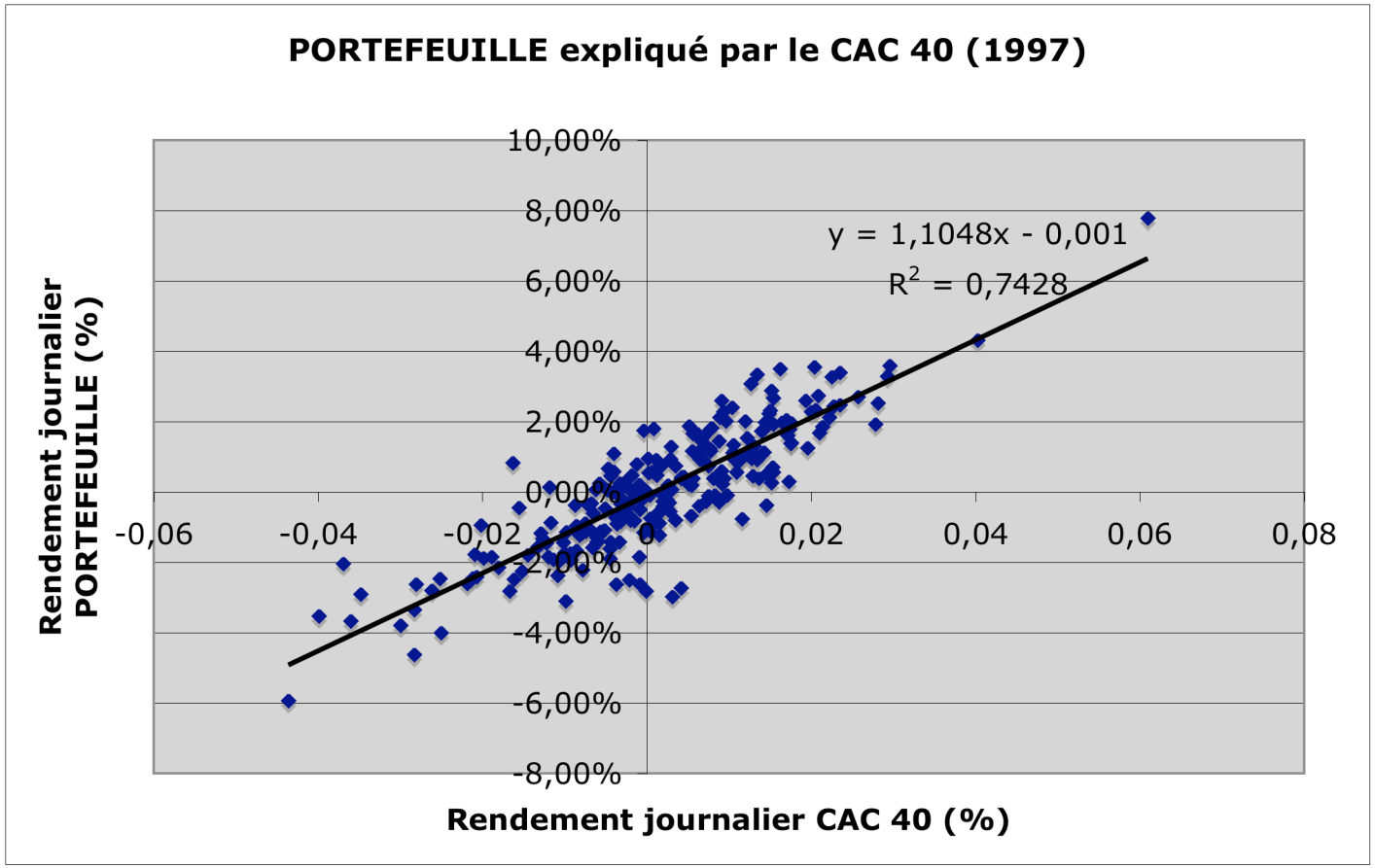
Générateur de jeux de *n* réponses i.i.d et de loi normale





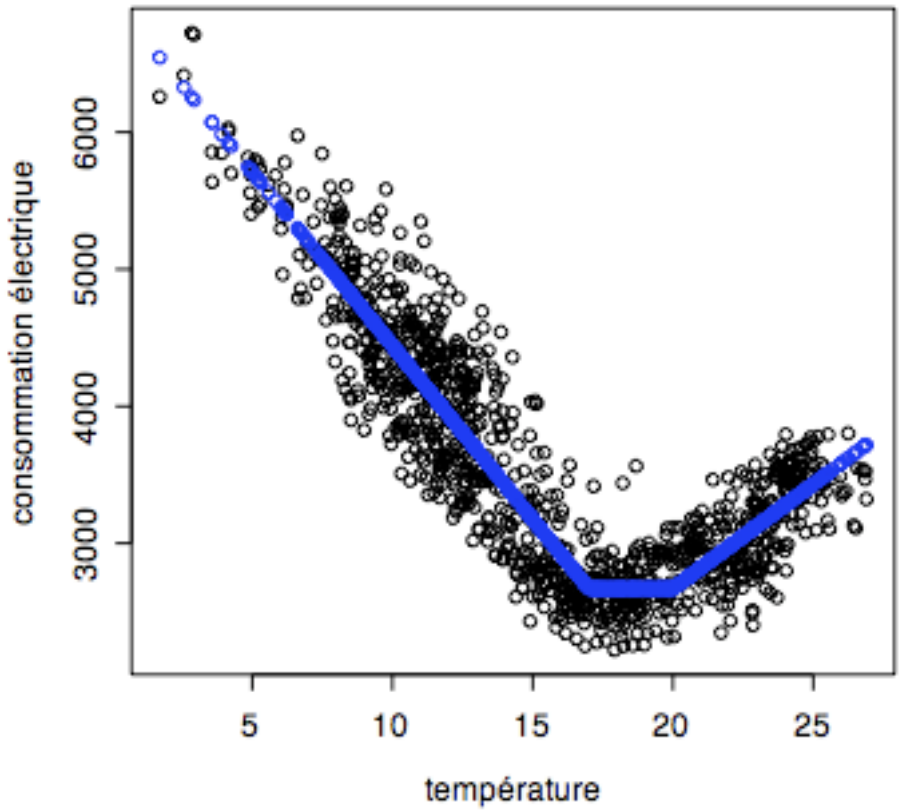
Exemples (1/2)

Que penser du modèle sur ces données ?



Exemples (2/2)

Et là ?



Validation du modèle

- On cherche à savoir si l'hypothèse du modèle linéaire est validée :
  - Avait t-on le droit de supposer que $e_1, \ldots, e_n$ sont indépendants et de même loi normale ?

- On va étudier les « résidus » :

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \ldots + \hat{\beta}_p x_{p,i})$$

Que regarder sur les résidus ?

- Les résidus ne devraient pas montrer de régularité (indépendance + même loi), et être centrés sur 0
  - Tracé des résidus
  - Tracé des résidus contre chaque prédicteur
  - Tracé des résidus contre la réponse estimée

Voir transparents 10 et 11

Que regarder sur les résidus (2)

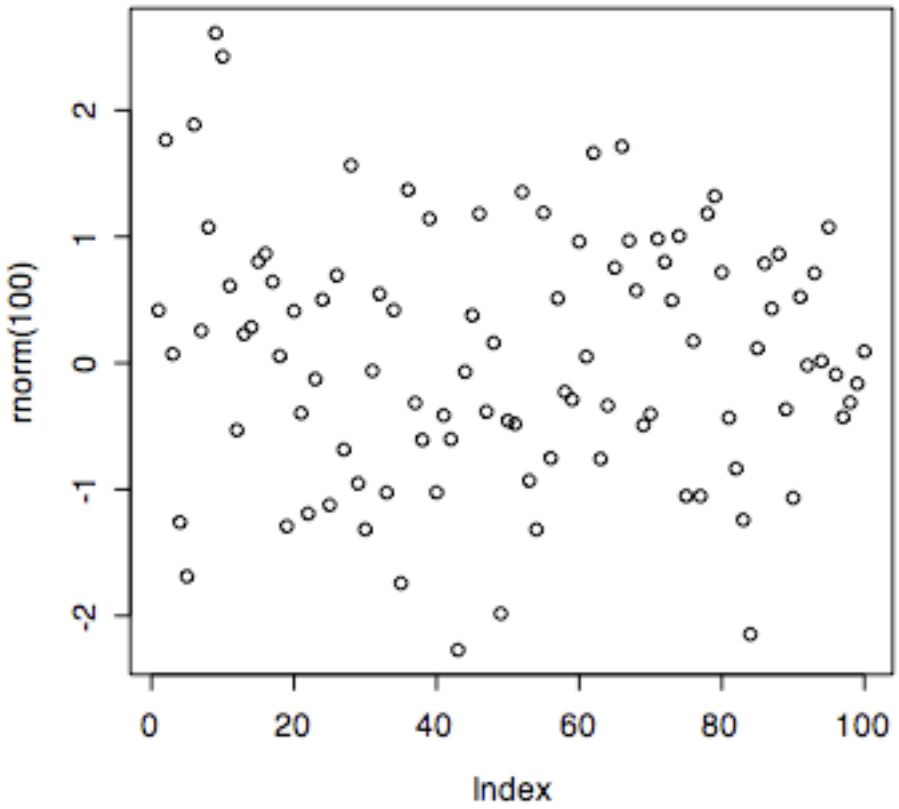
- Les résidus devraient être normaux
  - Tracé de la droite de Henri
  - Test statistique de normalité (Kolmogorov-Smirnov par ex.)

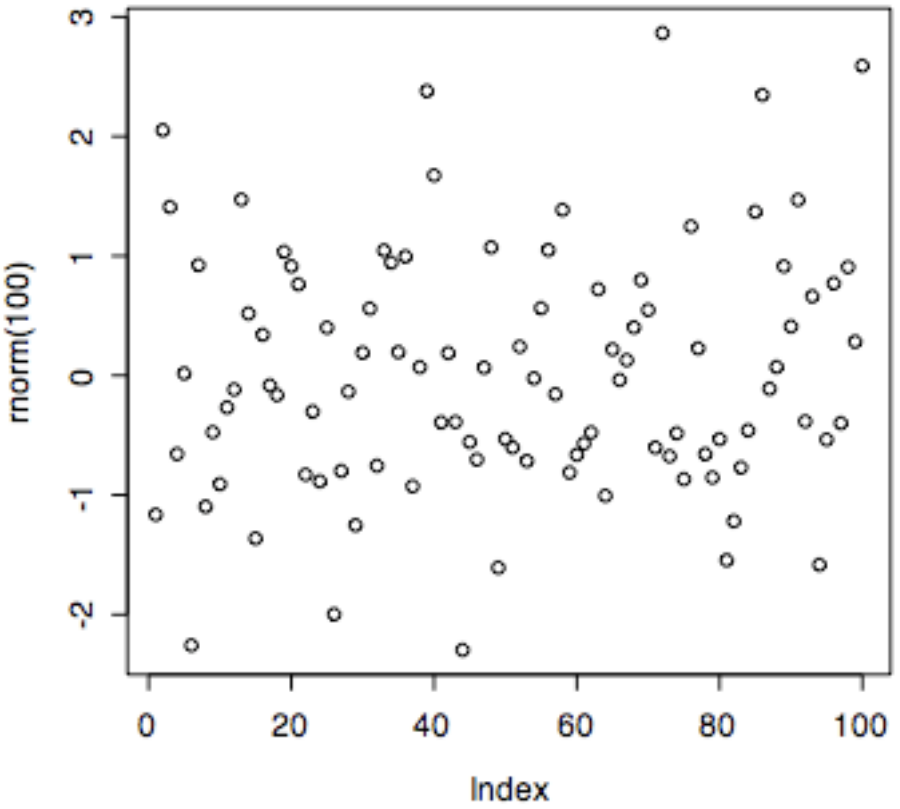
Voir transparent 12

- Il conviendrait d'étudier les résidus standardisés et même studentisés
  - Pas au programme cette année !

Ex. de résidus corrects

- Simulations -





Difficulté de la dimension > 1

Exemple :

- y : coût d'un produit
- x_1 : somme des salaires versés
- x_2 : nombre d'heures travaillées

- x_1 et x_2 sont fortement liées ! Ajuster un modèle de régression revient à essayer de faire passer un plan sur des données réparties autour d'un axe !
- Exercice : quels prédicteurs considérer ?

⇒ problème des prédicteurs corrélés