Introduction à la régression
cours n°4
Interprétation géométrique
Précision et validation du modèle

ENSM.SE – 1A

Olivier Roustant - Laurent Carraro

Objectifs du cours

- Connaître l'interprétation géométrique de la régression linéaire
- Savoir utiliser le coefficient de détermination R^2 pour apprécier la précision d'une régression
- Savoir valider ou invalider un modèle linéaire

Prévisions ponctuelles

- Considérons le modèle linéaire

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

avec $e_1, \ldots, e_n$ i.i.d $N(0,\sigma^2)$

- Prévision sans la régression : $\bar{y} := \sum_{i=1}^{n} y_i / n$

- Prévision avec la régression : $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \ldots + \hat{\beta}_p x_{p,i}$

Intérêt du modèle de régression

- Le modèle de régression a de l'intérêt si les erreurs $y_i - \hat{y}_i$ sont petites relativement aux erreurs $y_i - \bar{y}$ que l'on ferait sans avoir de prédicteurs
  - Donne envie de regarder $\dfrac{\|Y - \hat{Y}\|}{\|Y - \bar{Y}\|}$ avec ||.|| la norme usuelle de R^n et :

$$\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)' \qquad \bar{Y} = (\bar{y}, \ldots, \bar{y})' = \bar{y}(1, \ldots, 1)' =: \bar{y}1$$

Des moindres carrés à la géométrie

- Estimation par moindres carrés

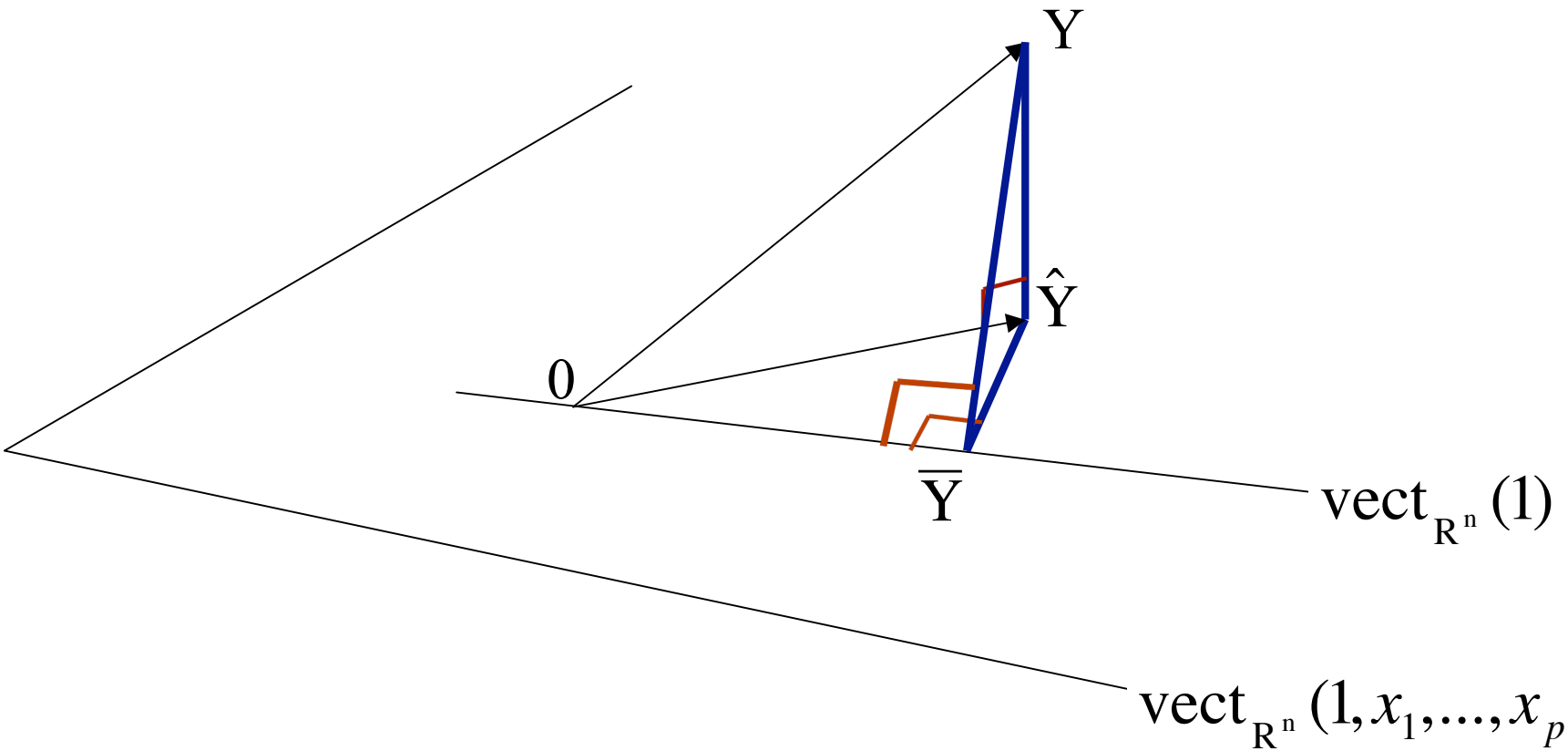
$$\hat{\beta} = Arg\min_{\beta} \|Y - X\beta\|^2$$

- Interprétation géométrique
  - $\hat{Y} = X\hat{\beta}$ est la projection orthogonale dans R^n de Y sur le plan engendré par les prédicteurs (1 inclus)

Des angles droits partout !

- Exercice
  - Montrer que $\overline{Y}$ est la projection orthogonale de Y sur la droite de R^n engendrée par $1 = (1,\ldots,1)$
  - En déduire que $\overline{Y}$ est aussi la projection orthogonale de $\hat{Y}$ sur cette même droite

Interprétation géométrique

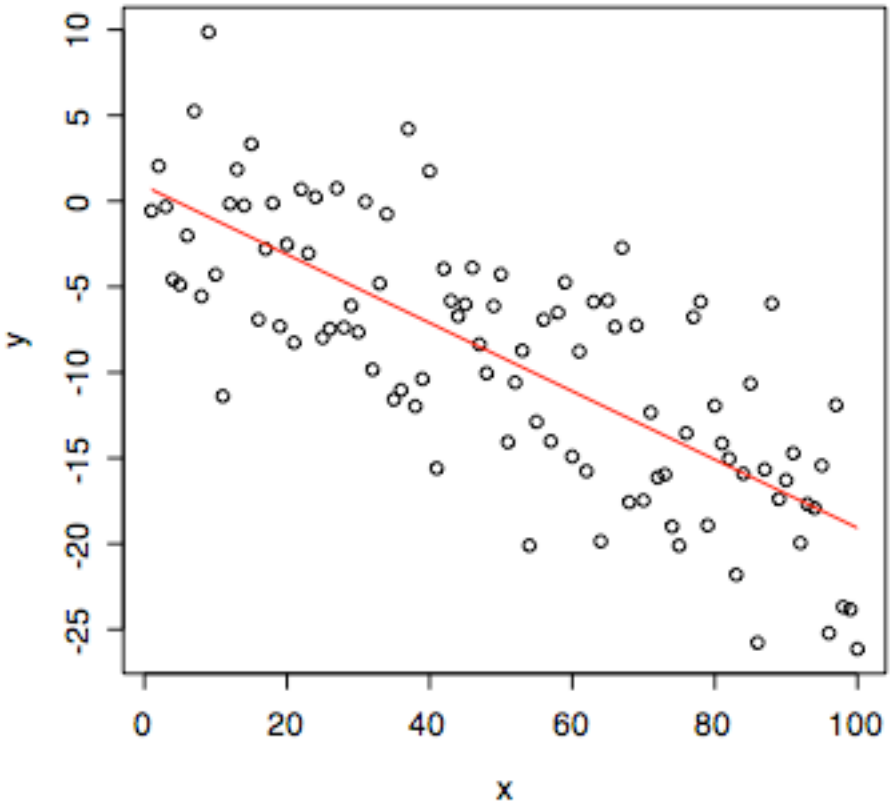


Coefficient de détermination R^2

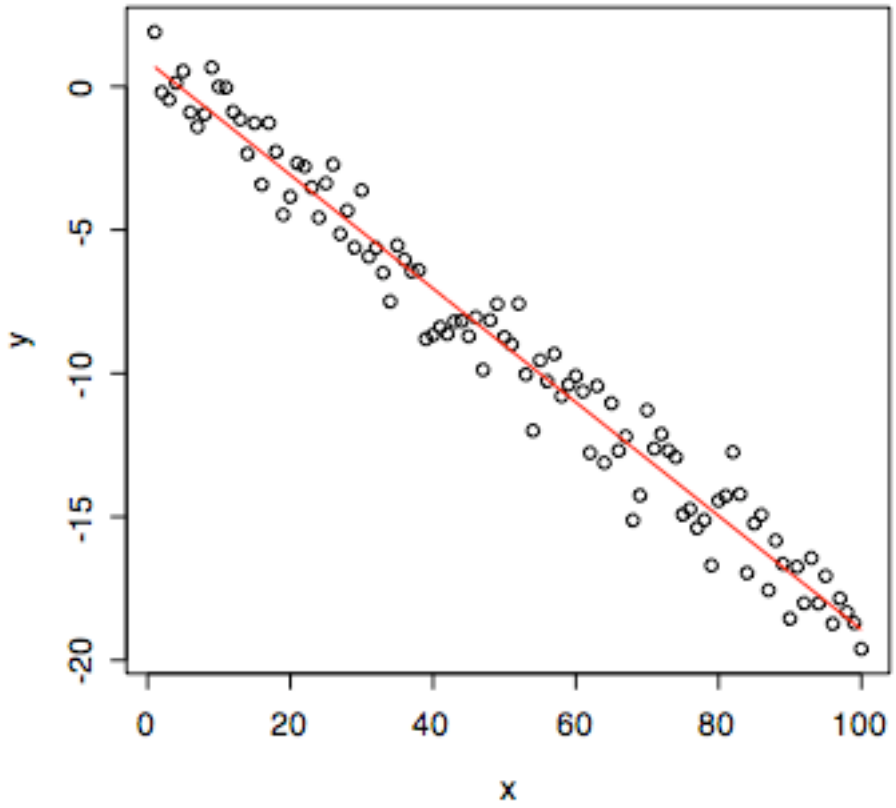
- Définition
  - $R^2 = 1 - \frac{\|Y - \hat{Y}\|^2}{\|Y - \bar{Y}\|^2} = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} \in [0,1]$
- Interprétation :
  - Pourcentage de variance expliquée par la régression
    - Numérateur et dénominateur s'interprètent effectivement comme la variance empirique des quantités considérées
  - La réponse est bien expliquée par la régression lorsque R^2 est proche de 1

Exemple

$y_i = 1 - 0.2\, i + e_i$ avec $e_1, \ldots, e_{100}$ i.i.d $N(0, s^2)$

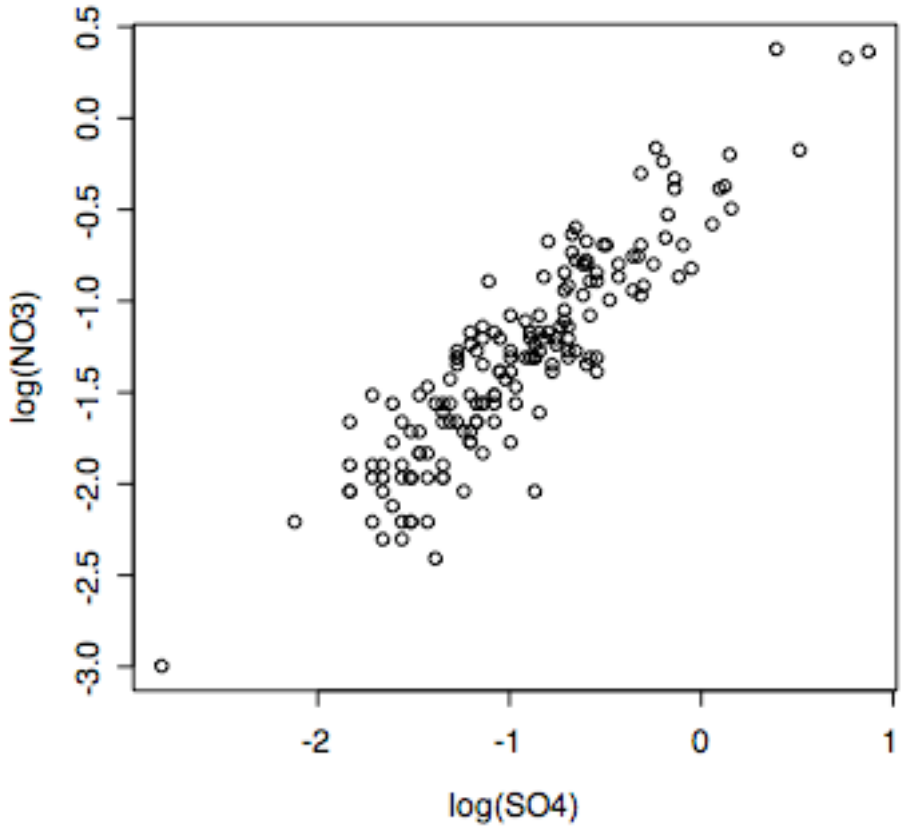


s=5, R^2 =0.583



s=1, R^2 =0.969

Données de pollution (cf cours 1)



```
Call:
lm(formula = log(NO3) ~ log(SO4), data = pollution)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80424 -0.14485 -0.01087  0.16564  0.56666

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -0.43642    0.03679  -11.86   <2e-16 ***
log(SO4)     0.92168    0.03356   27.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'

Residual standard error: 0.2417 on 163 de
Multiple R-Squared: 0.8205,          Adjusted R-squared: 0.8195
F-statistic: 754.4 on 1 and 165 DF,  p-value: < 2.2e-16
```

R^2 de 82% : le prédicteur explique plutôt bien la réponse

Quelques dangers du R^2

- Le coefficient de détermination peut être égal à 1, et le modèle inexploitable
  - C'est ce qui arrive avec un modèle linéaire polynômial de degré n-1 avec n points…

- Attention : la précision du modèle n'est valable qu'aux points d'observation x_i
  - En dehors, on ne peut rien dire

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

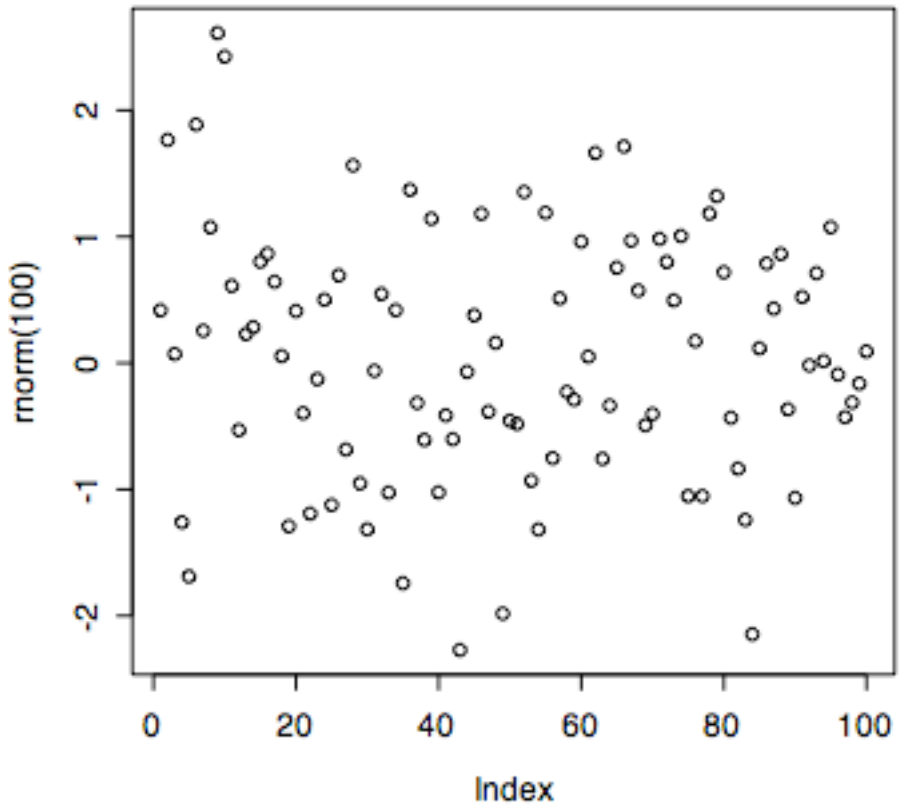
Que faut-il regarder sur les résidus (2)

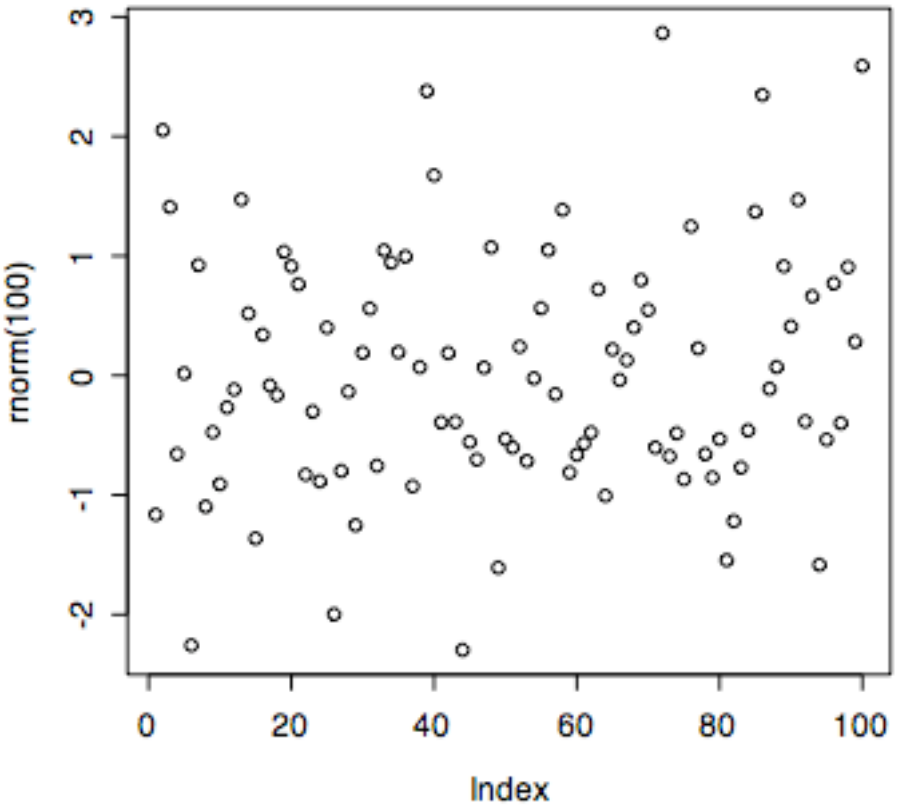
- Les résidus devraient être normaux
  - Tracé de la droite de Henri

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

Exemple de résidus corrects

- Simulations -

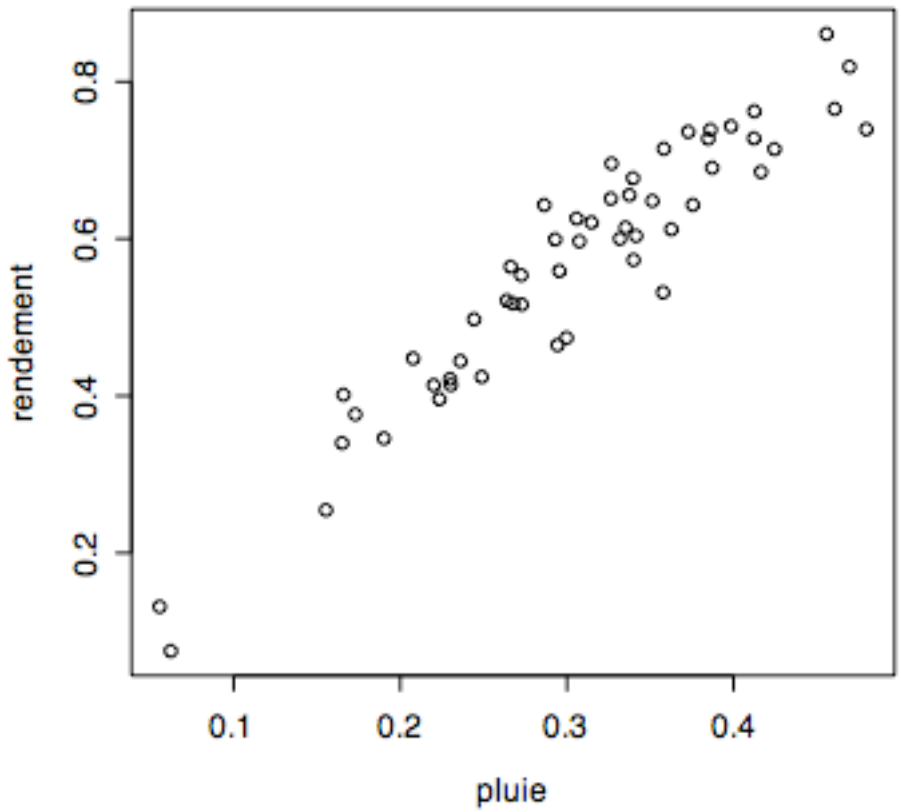




Exemple d'étude

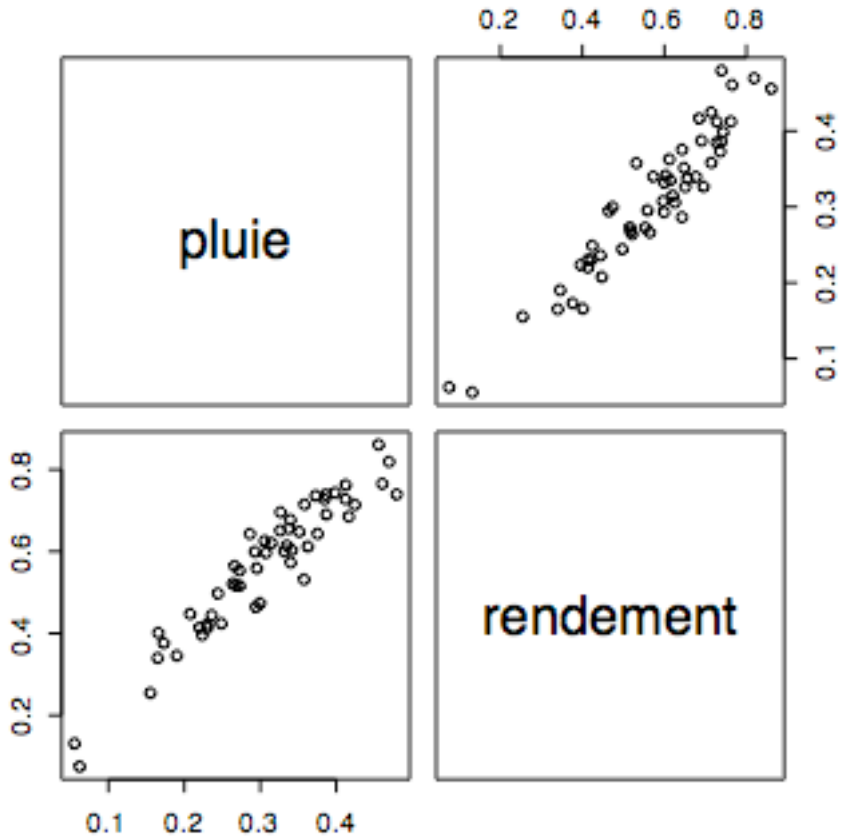
- Réponse : pourcentage d'un rendement maximal de blé
- Prédicteur : quantité de pluies printanières (en m)
- 54 observations

Observation des données




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

Observation des données (suite)

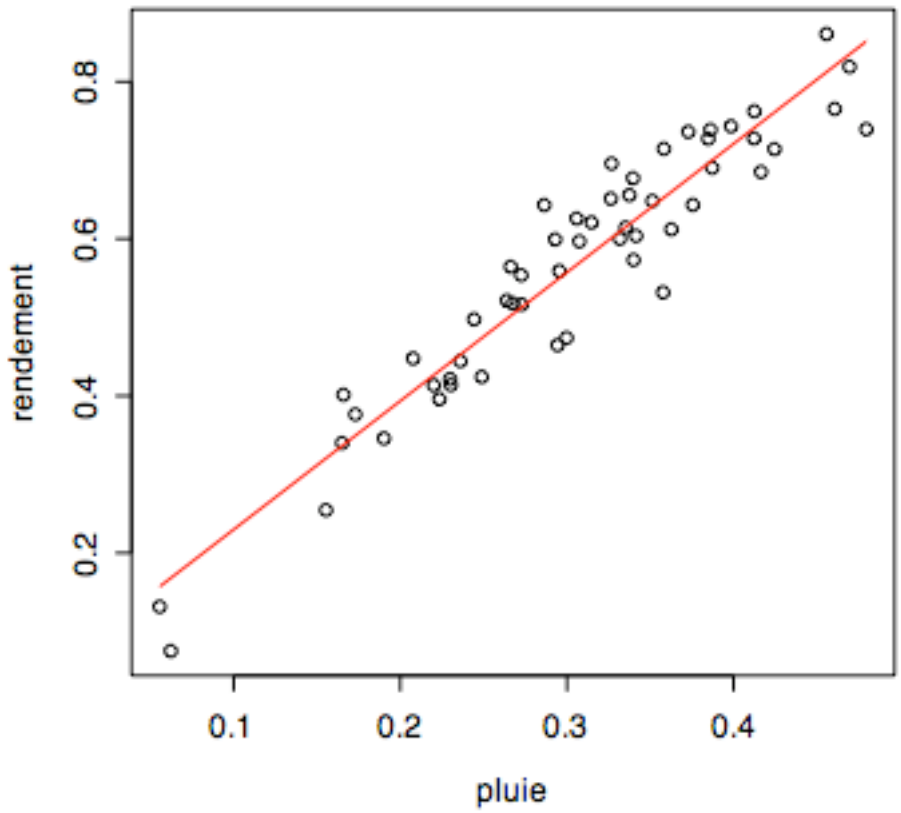


Ecole Nationale Supérieure des Mines

```
pairs(donnees)
```

1er modèle

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{avec} \quad e_1, \ldots, e_{54} \quad \text{i.i.d } N(0, \sigma^2)$$




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

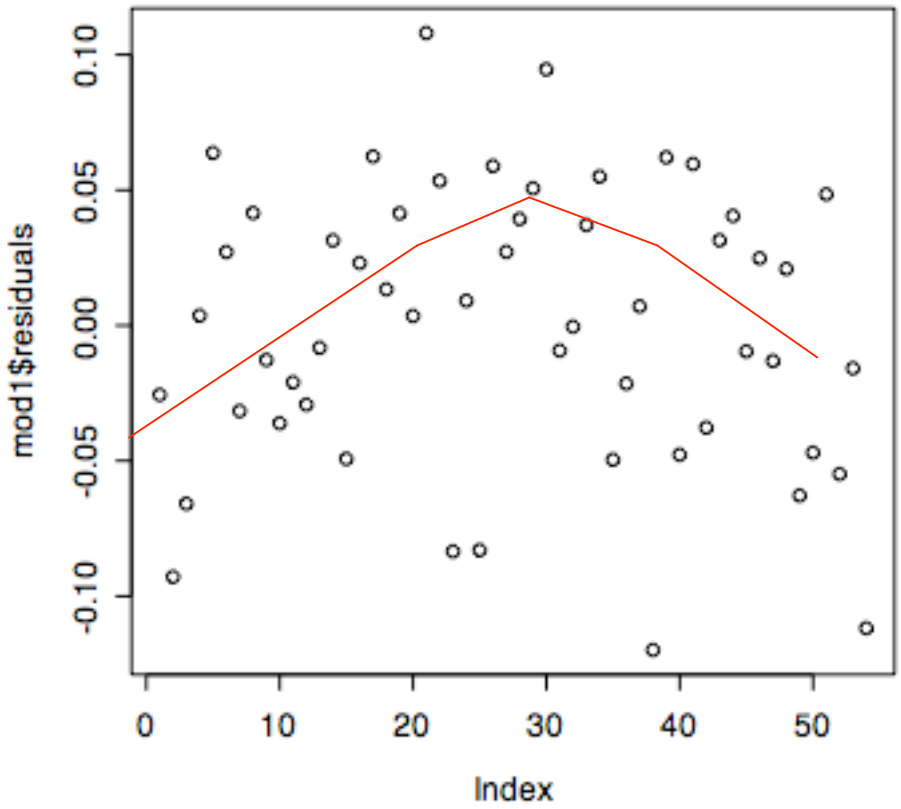
Table d'ANOVA

```
Call:
lm(formula = rendement ~ pluie, data = donnees)

Residuals:
      Min         1Q    Median         3Q        Max
-0.119861 -0.034987  0.003603  0.040208  0.108037

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06620    0.02405   2.752  0.00813 **
pluie        1.63673    0.07526  21.747  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05201 on 52 degrees of freedom
Multiple R-Squared: 0.9009,   Adjusted R-squared: 0.899
F-statistic: 472.9 on 1 and 52 DF,  p-value: < 2.2e-16
```



```
mod1 <- lm(rendement~pluie, data=donnees)
summary(mod1)
```

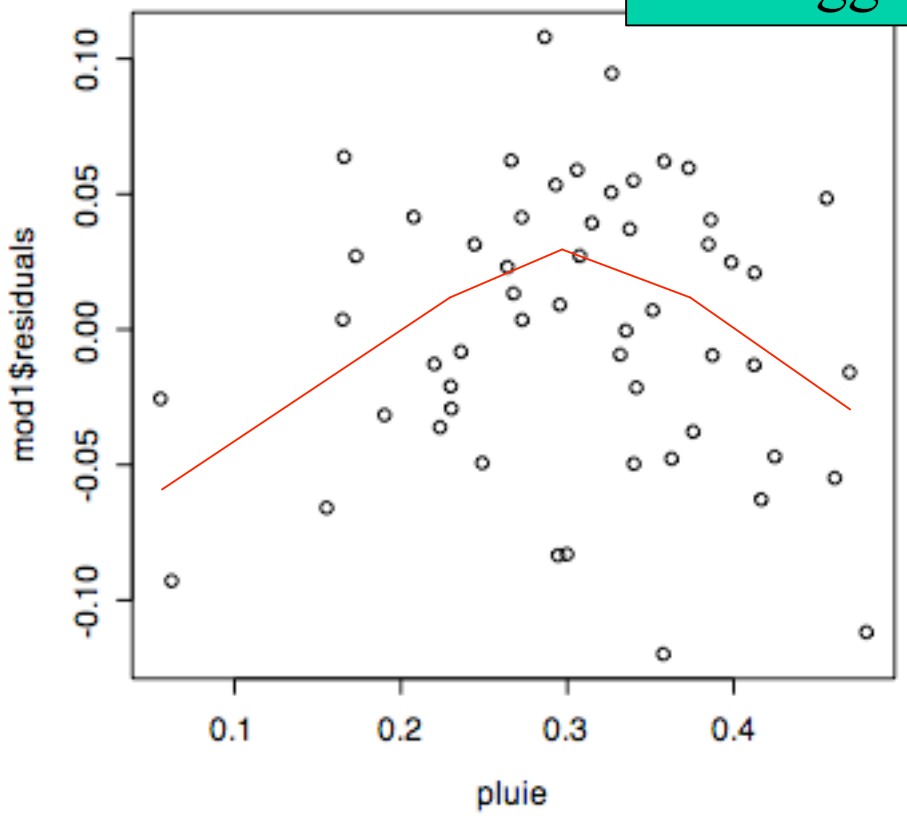
Examen des résidus

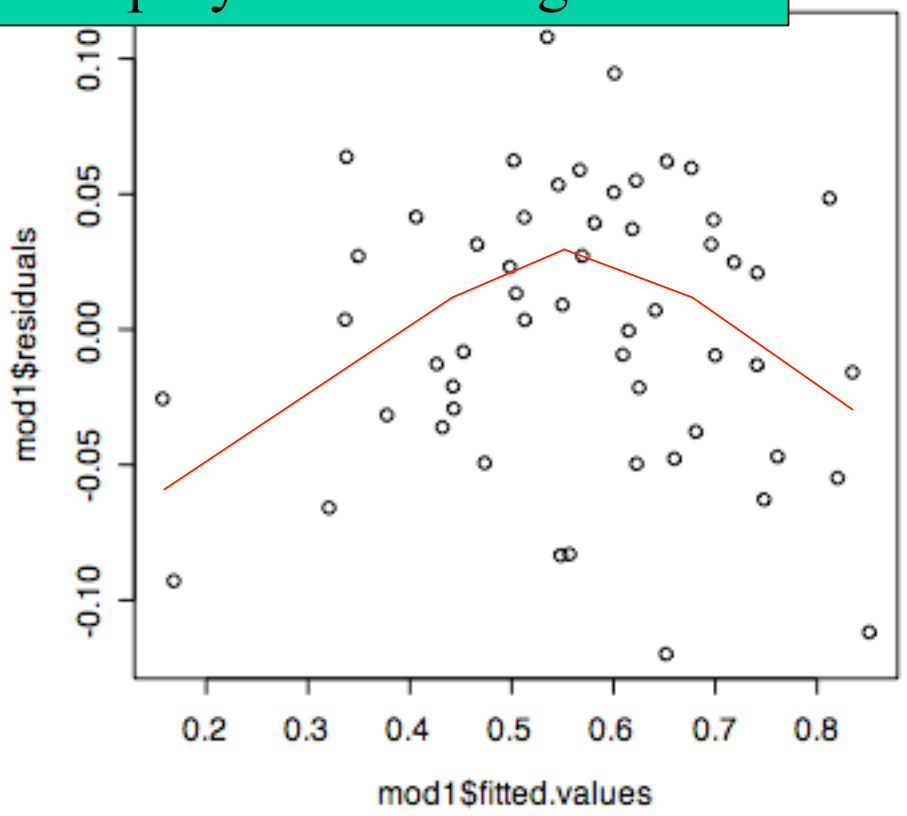



```
plot(mod1$residuals)
```

Examen des résidus (suite)

Courbure quadratique des résidus :
Suggère un polynôme de degré 2

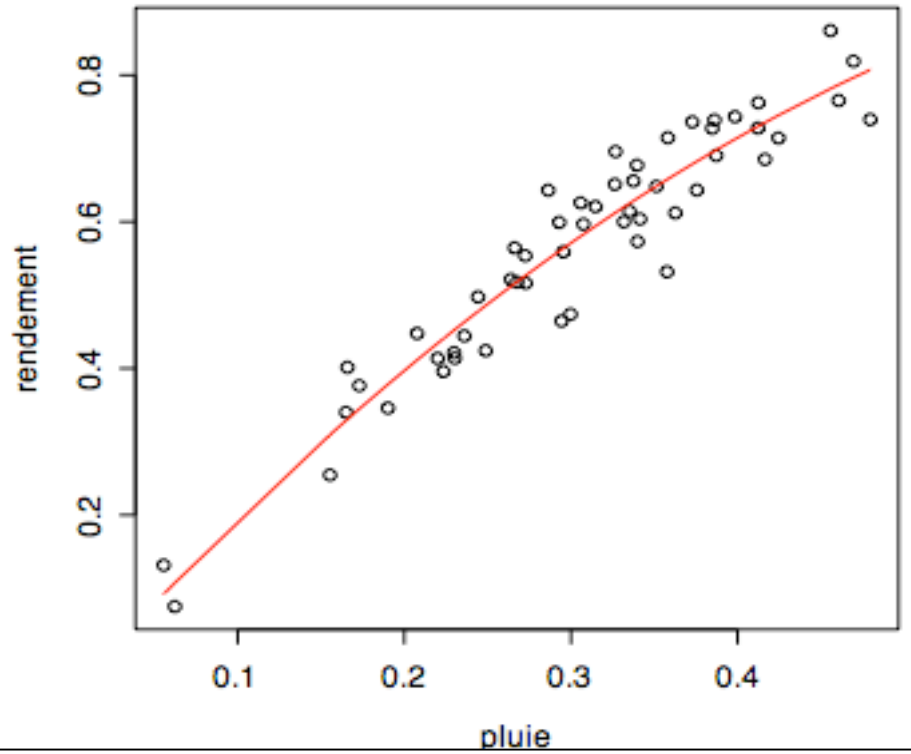





```
plot(mod1$residuals~pluie,data=donnees)
plot(mod1$residuals~mod1$fitted.values, data=donnees)
```

2ème modèle

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + e_i \quad \text{avec} \quad e_1, \ldots, e_{54} \quad \text{i.i.d } N(0, \sigma^2)$$




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

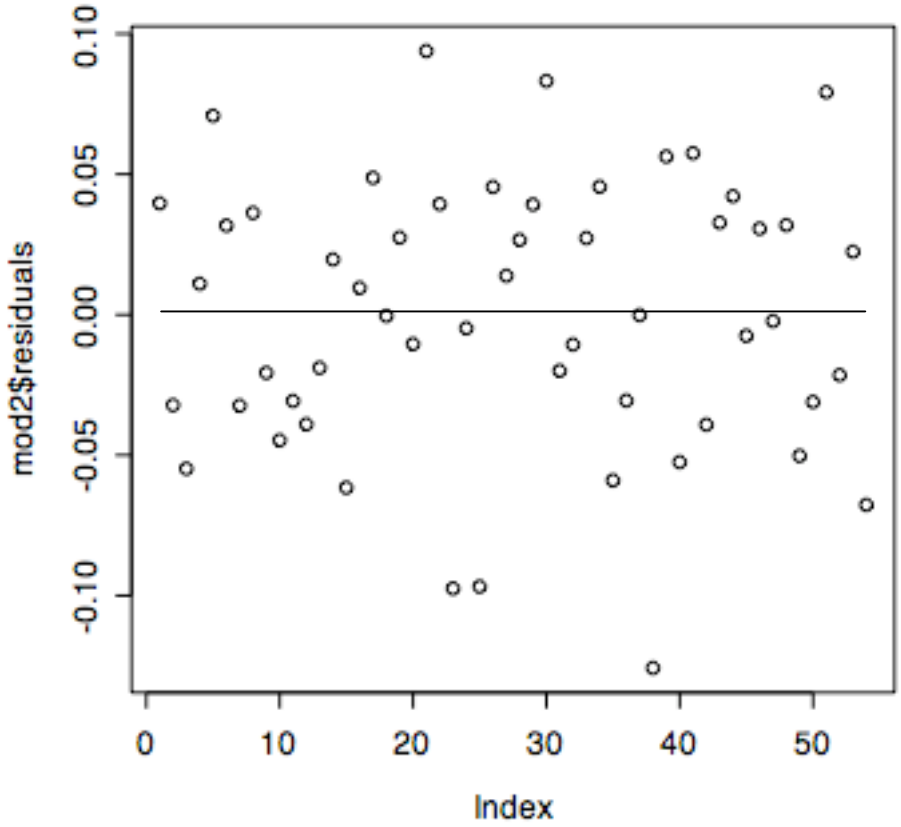
Table d'ANOVA

```
Call:
lm(formula = rendement ~ pluie + I(pluie^2), data = donnees)

Residuals:
      Min         1Q     Median         3Q        Max
-0.125759 -0.031894 -0.000287  0.035384  0.093880

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04246    0.04512  -0.941  0.35109
pluie        2.50171    0.31868   7.850 2.49e-10 ***
I(pluie^2)  -1.52278    0.54701  -2.784  0.00752 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04893 on 51 degrees of freedom
Multiple R-Squared: 0.914,    Adjusted R-squared: 0.9106
F-statistic:   271 on 2 and 51 DF,  p-value: < 2.2e-16
```



```
mod2 <- lm(rendement~pluie+I(pluie^2), data=donnees)
summary(mod2)
```

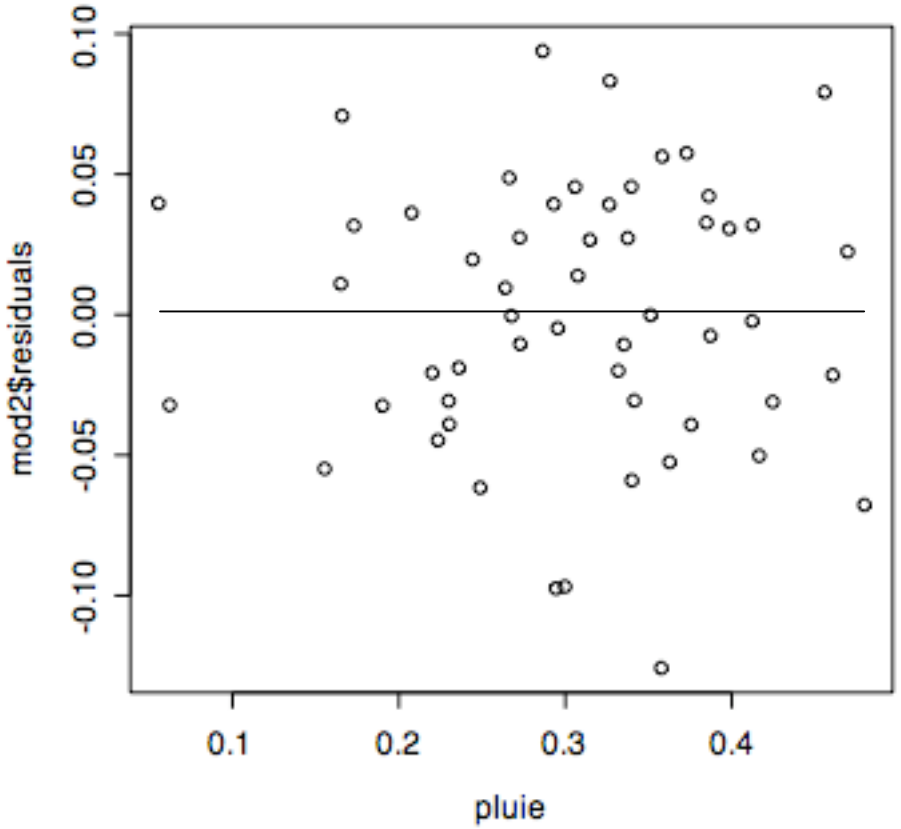
Examen des résidus du 2nd modèle

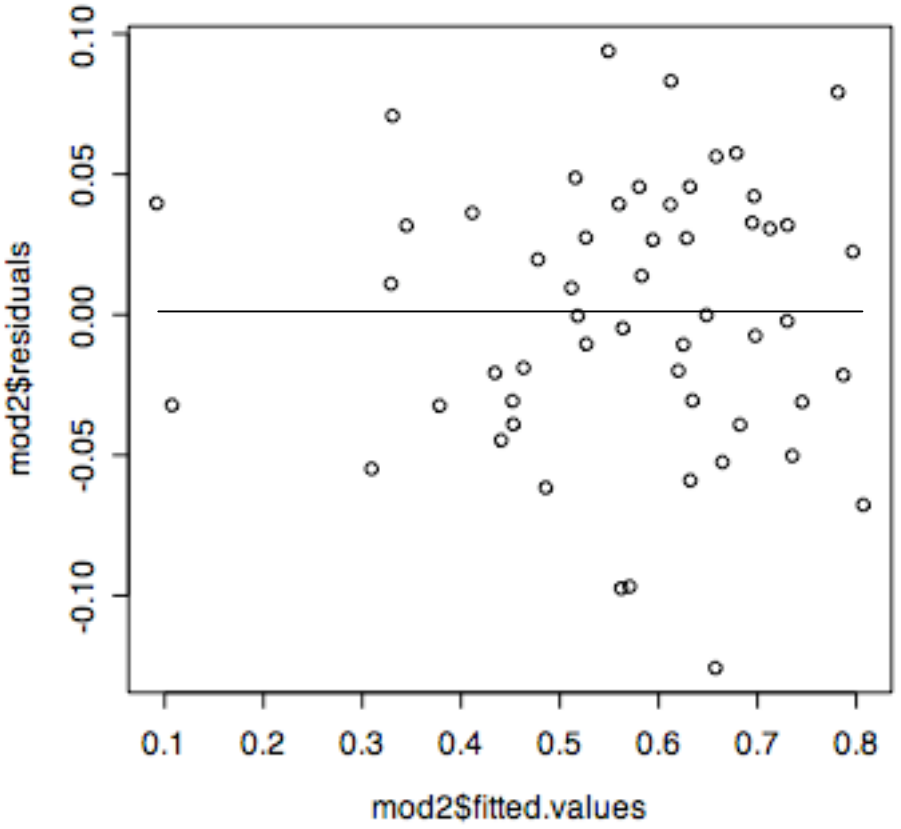


Ecole Nationale Supérieure des Mines

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

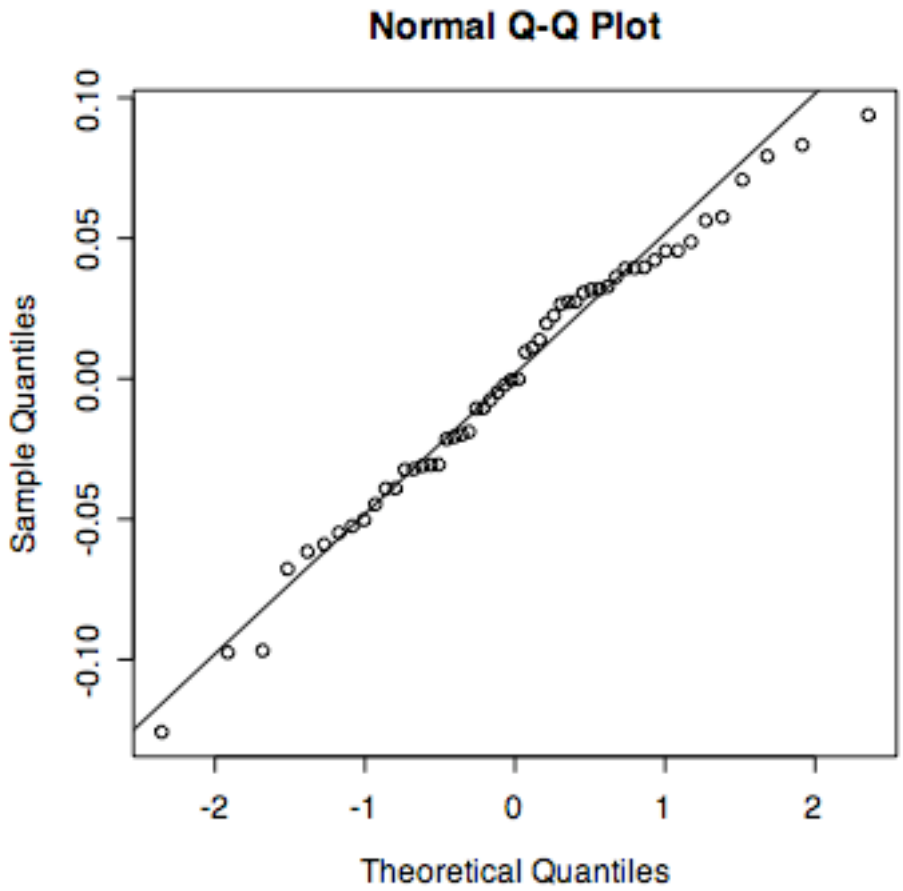
Examen des résidus (suite)






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

Dernière étape : normalité des résidus



Ecole Nationale Supérieure des Mines

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

Conclusion

- Le modèle de régression linéaire polynomial de degré 2 est validé
- On pourra l'utiliser pour faire des prévisions