4. Régression

4.1 Introduction

Motivation

La **régression** concerne la relation entre une variable d'intérêt et d'autres variables. On note

- la variable d'intérêt, la **variable de réponse**, y, et on la considère comme variable aléatoire
- les autres variables, les **covariables** (variables explanatoires) sont notées $x^{(1)}, \ldots, x^{(p)}$, on les considère comme fixées

On peut s'interesser à

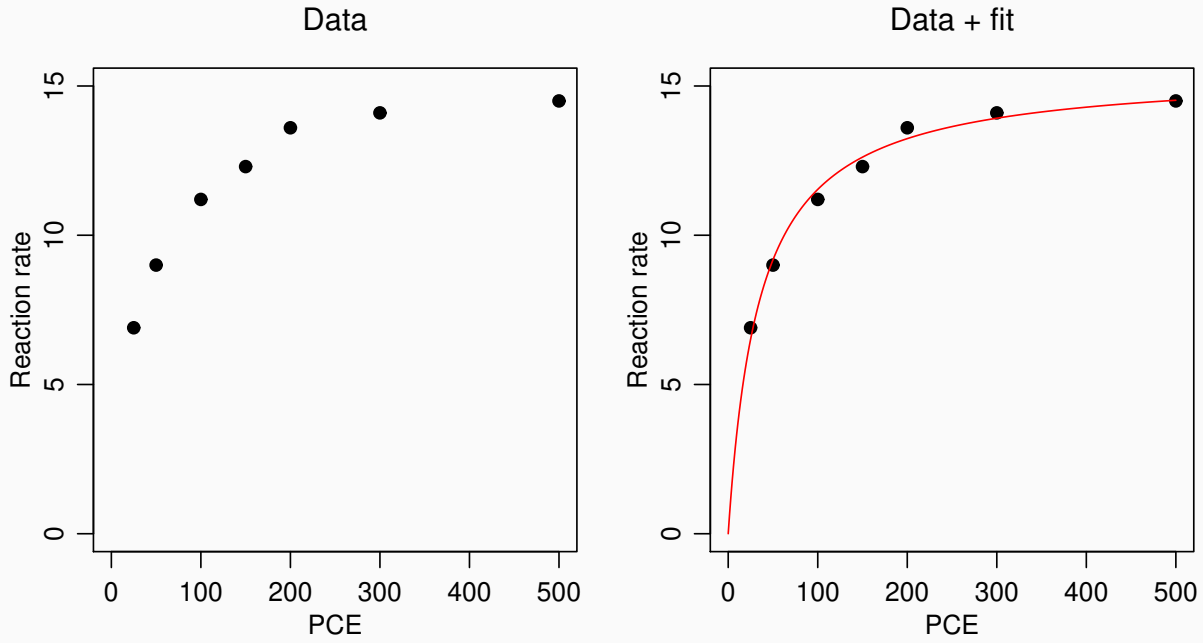
- l'**estimation** d'une relation eventuelle entre y et les $x^{(j)}$, ou
- la **prévision** des valeurs futures/manquantes de y sur la base des $x^{(j)}$ correspondantes

Réaction chimique

Professeur Christophe Holliger (SIE) : on essaye de déterminer les paramètres kinétiques d'une 'reductive dehalogenase dechlorinating tetrachloroethene (PCE)'. Ceci dépend de la concentration du substrat, et la vitesse de la réaction peut être exprimé par l'équation de Michaelis–Menten

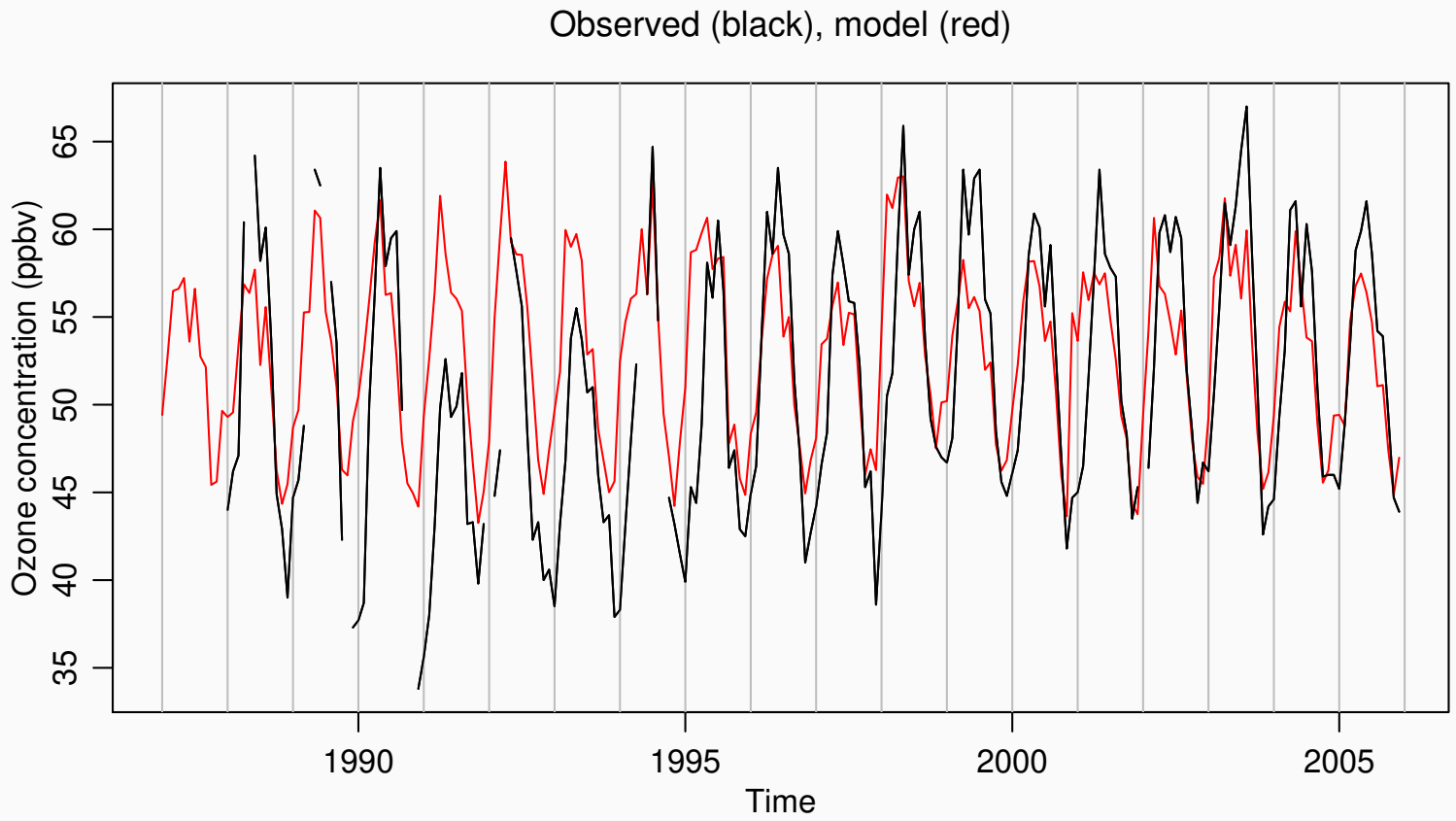
$$y = \frac{\gamma_0 x}{\gamma_1 + x},$$

où x est la concentration de PCE, γ_0 est la vitesse maximale, et γ_1 est la concentration quand $y = \gamma_0/2$. Comment estimer γ_0 et γ_1 ? Quelles sont leurs incertitudes ?



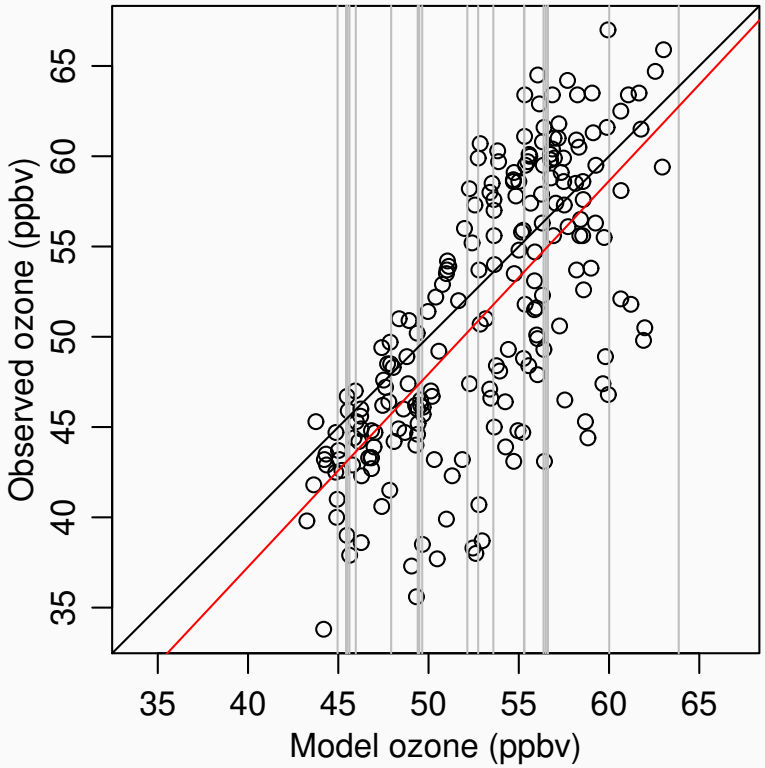
Ozone atmosphérique

Observations de la concentration de l'ozone au Jungfraujoch, de janvier 1987 à décembre 2005 (quelques valeurs manquantes), et resultats d'une modélisation



Soient y les données réeles et x les resultats du modèle

Relation linéaire ?



Les lignes verticales grises montrent des x dont les y sont manquants. La ligne noire montre la relation $y = x$, et la ligne rouge montre la meilleure estimation d'une relation linéaire entre y et x.

Comment utiliser la relation entre les resultats du modèle x et les données observées y pour estimer les y manquants ?

Problème d'ajustement

- On considère une variable de réponse y que l'on cherche à expliquer par une covariable x
- On dispose d'un ensemble de points

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

  qu'on peut représenter par un nuage de points (scatterplot) comme ceux d'auparavant
- D'une manière générale, le **problème d'ajustement** consiste à trouver une courbe $y = \mu(x)$ qui résume "le mieux possible" le nuage de points. La fonction $\mu(x)$ dépend de paramètres qu'il faut estimer
- S'il y a une **relation linéaire**, on peut utiliser la corrélation pour mesurer la dépendance linéaire entre les y et x. La régression linéaire permet de résumer cette dépendance par une droite

Moindres carrés

- Les écarts verticaux entre les données y_j et la courbe $\mu(x_j)$ sont

$$y_j - \mu(x_j), \quad j = 1, \ldots, n$$

- On cherche les paramètres de la fonction $\mu(x)$ pour minimiser la **somme des carrés** des écarts verticaux

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2$$

- L'ajustement est dit **linéaire** si $\mu(x) = a + \beta x$. Dans ce cas, il faut minimiser

$$S(a, \beta) = \sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 = \sum_{j=1}^{n} \{y_j - (a + \beta\, x_j)\}^2$$

Estimateurs de moindres carrés

Théorème Soient $(x_1, y_1), \ldots, (x_n, y_n)$ issues d'un relation $y = a + \beta x$ et telles que pas tous les x_j sont égaux. Alors les **estimateurs de moindres carrés** de a et β sont

$$\widehat{a}_n = \overline{y}_n - \widehat{\beta}_n \overline{x}_n, \quad \widehat{\beta}_n = \frac{\sum_{j=1}^n (x_j - \overline{x}_n) y_j}{\sum_{j=1}^n (x_j - \overline{x}_n)^2}$$

Définition: La droite

$$\widehat{a}_n + \widehat{\beta}_n x$$

s'appelle la **droite des moindres carrés**, la **valeur ajustée** qui correspond à (x_j, y_j) est

$$\widehat{y}_j = \widehat{a}_n + \widehat{\beta}_n x_j,$$

et la différence

$$r_j = y_j - \widehat{y}_j = y_j - (\widehat{a}_n + \widehat{\beta}_n x_j)$$

s'appelle un **résidu**

Preuve

Il faut minimiser

$$S(a,\beta) = \sum_{i=1}^{n}(y_i - a - \beta x_i)^2$$

en a et β. On calcule

$$\frac{dS}{da}(a,\beta) = -2\sum_{i=1}^{n}(y_i - a - \beta x_i) = 2na + 2n\beta\overline{x}_n - 2n\overline{y}_i$$

$$\frac{dS}{d\beta}(a,\beta) = -2\sum_{i=1}^{n} x_i(y_i - a - \beta x_i) = 2na\overline{x}_n + 2\beta\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i$$

$$\frac{d^2S}{da^2}(a,\beta) = 2n > 0 \qquad \frac{d^2S}{d\beta^2}(a,\beta) = 2\sum_{i=1}^{n} x_i^2 \qquad \frac{d^2S}{d\beta da}(a,\beta) = 2n\overline{x}_n$$

La matrice hessienne est donc

$$H = \begin{pmatrix} 2n & 2n\overline{x}_n \\ 2n\overline{x}_n & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix} \qquad \text{définie positive}$$

car $2n > 0$ et $\det(H) = 4n[(\sum_{i=1}^{n} x_i^2) - n\overline{x}_n] = 4n\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 > 0$

Propriétés

- La droite de moindres carrés passe par $(\overline{x}_n, \overline{y}_n)$
- $\sum_{j=1}^{n} r_j = 0$
- $\sum_{j=1}^{n} x_j\, r_j = \sum_{j=1}^{n} x_j(y_j - \widehat{y}_j) = 0$
- $\sum_{j=1}^{n} \widehat{y}_j\, r_j = 0$

(voir série d'exercices). Donc

$$\sum_{j=1}^{n}(y_j - \bar{y}_n)^2 = \sum_{j=1}^{n}\Big(\underbrace{y_j - \widehat{y}_j}_{r_j} + \widehat{y}_j - \bar{y}_n\Big)^2 = \cdots = \sum_{j=1}^{n}(\widehat{y}_j - \bar{y}_n)^2 + \sum_{j=1}^{n} r_j^2,$$

nous donnant la **décomposition de la somme des carrés** total

$$\mathrm{SC}_{\mathrm{Total}} = \mathrm{SC}_{\mathrm{R}} + \mathrm{SC}_{\mathrm{E}}$$

en une partie due à la régression (variation expliquée par le modèle) et une partie due à l'erreur (variation non-expliquée par le modèle)

Ozone atmosphérique

- Il y a $n = 207$ paires (Observée, Modèle) $= (y_j, x_j)$, et en plus 21 valeurs de x sans valeur observée
- Avec les n paires complètes on trouve comme droite des moindres carrés

$$\widehat{y} = \widehat{a}_n + \widehat{\beta}_n x = -5.511 + 1.069x$$

avec décomposition de la somme des carrés

$$\mathrm{SC}_{\mathrm{Total}} = \mathrm{SC}_{\mathrm{R}} + \mathrm{SC}_{\mathrm{E}} = 5813 + 5832.$$

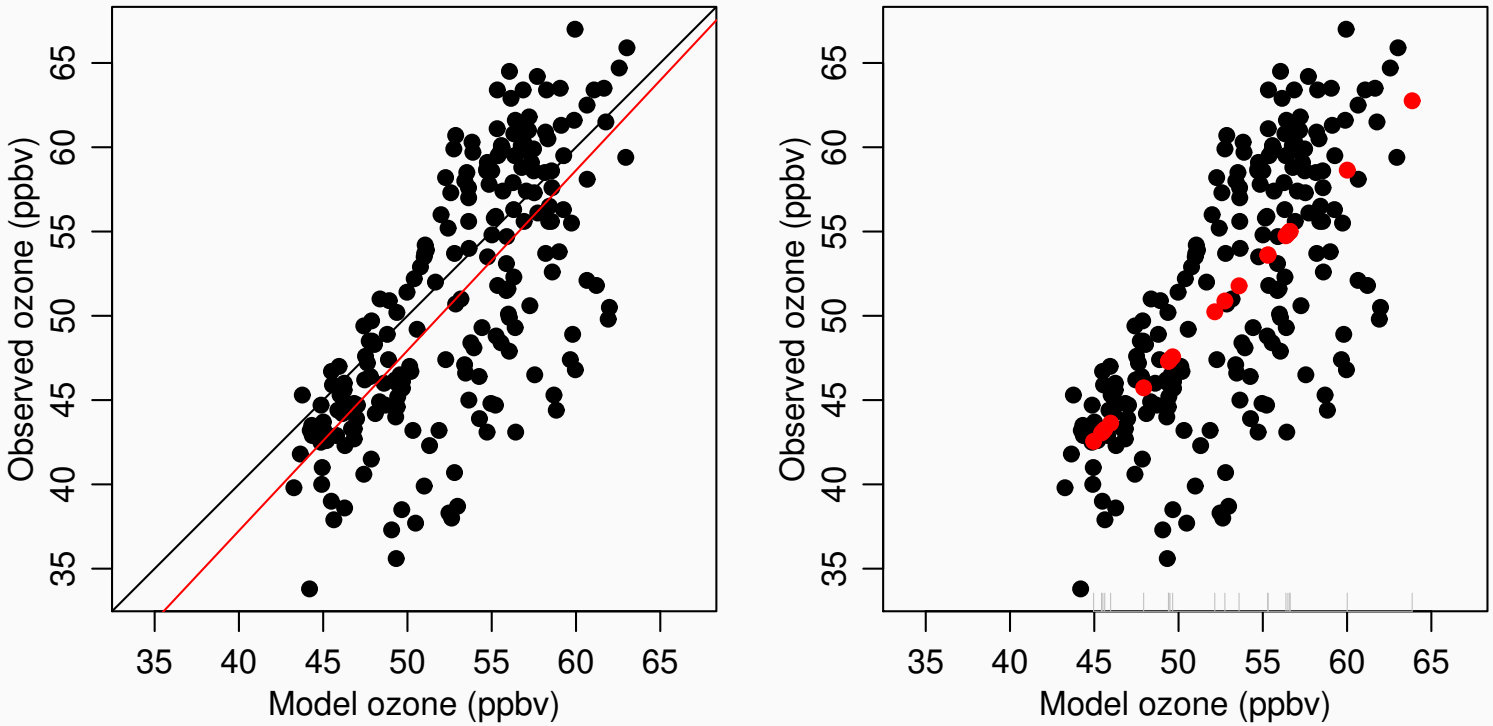
La régression explique donc une moitié de la somme des carrés totale

- Pour un paire (Observée, Modèle) $= (\,?,\, x_+)$ dont la valeur observée manque, on peut la remplacer par la valeur ajustée correspondante,

$$\widehat{y}_+ = \widehat{a}_n + \widehat{\beta}_n x_+.$$

On parle d'**imputation** de donnée

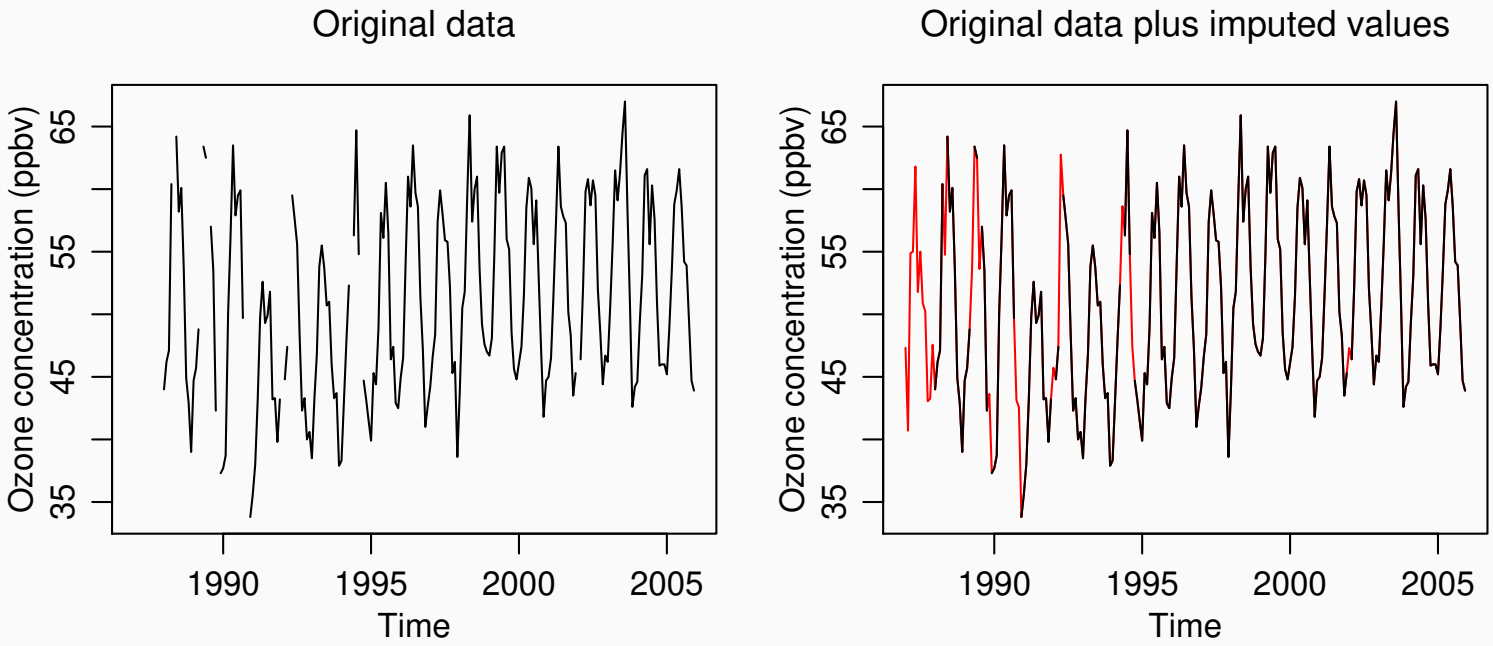
Modèle ajusté



Gauche : droite $y = x$ et droite ajustée $\widehat{y} = \widehat{a}_n + \widehat{\beta}_n x = -5.511 + 1.069x$

Droite : valeurs ajustées pour des valeurs manquantes de x

Données imputées



Gauche : données originales

Droite : données originales (noir) avec valeurs imputées (rouge). Comparer avec la diapositive 226.

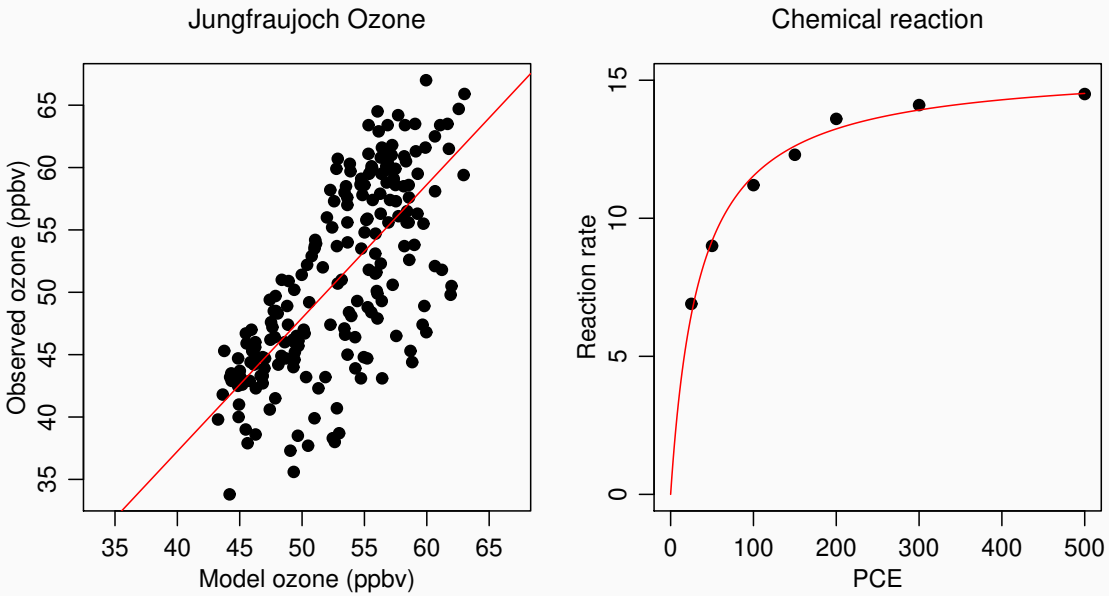
4.2 Modèle statistique

Modèle normale

- On observe une version perturbée d'une relation $y = \mu(x)$
- Pour modéliser ceci, on peut souvent supposer que

$$y_j \overset{\text{ind}}{\sim} \mathcal{N}\left\{\mu(x_j), \sigma^2\right\} \quad \text{ou bien} \quad y_j = \mu(x_j) + \epsilon_j, \quad \epsilon_j \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Ainsi la dépendance entre la réponse y et la variable explicative x est donnée par $\mathbb{E}(y) = \mu(x)$, alors que le bruit dépend de σ^2
- À gauche : $\mu(x)$ linéaire, σ^2 grand, donc beaucoup de bruit
- À droite : $\mu(x)$ non-linéaire, σ^2 petite, donc peu de bruit



Linéarité

- La linéarité du modèle concerne les paramètres :

$$y = a + \beta x + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$ est la différence entre y et la droite $a + \beta x$

- Le modèle

$$y = a + \beta x + \gamma x^2 + \delta x^3 + \epsilon$$

est linéaire en $(a, \beta, \gamma, \delta)$.

- Le modèle

$$y = \gamma_0 x^{\gamma_1} \eta, \quad \eta \sim \exp(1),$$

devient linéaire après transformation logarithmique :

$$\log y = \log \gamma_0 + \gamma_1 \log x + \log \eta = a + \beta x' + \log \eta$$

- Le modèle

$$y = \frac{\gamma_0 x}{\gamma_1 + x} + \epsilon$$

n'est pas linéaire en les paramètres γ_0, γ_1

Estimation des paramètres

- Dans le cas $\mu(x) = a + \beta x$ il y a trois paramètres inconnus : (intercepte, pente, bruit), $\theta = (a, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$
- Nous utilisons la méthode de maximum de vraisemblance pour les estimer
- La log vraisemblance est

$$\ell(a, \beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \overbrace{\sum_{j=1}^{n} \{y_j - (a + \beta x_j)\}^2}^{S(a,\beta)} - \frac{n}{2} \log(2\pi),$$

et en maximisant celle-ci par rapport à θ nous trouvons

$$\widehat{a}_n = \overline{y}_n - \widehat{\beta}_n \overline{x}_n, \quad \widehat{\beta}_n = \frac{\sum_{j=1}^{n} (x_j - \bar{x}_n) y_j}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2}, \quad \widehat{\sigma}_n^2 = n^{-1} \sum_{j=1}^{n} r_j^2$$

avec $r_j = y_j - \widehat{y}_j$ les **résidus** et $\widehat{y}_j = \widehat{a}_n + \widehat{\beta}_n x_j$ les **valeurs ajustées**

- Les estimateurs $\widehat{a}_n$ et $\widehat{\beta}_n$ sont les estimateurs de moindres carrés et sont sans biais, mais $\mathbb{E}(\widehat{\sigma}_n^2) < \sigma^2$, et on utilise souvent l'estimateur non-biaisé (comparer avec 174)

$$S_n^2 = \frac{1}{n-2} \sum_{j=1}^{n} r_j^2 = \frac{1}{n-2} \sum_{j=1}^{n} (y_j - \widehat{y}_j)^2$$

Inférence pour les paramètres du modèle linéaire simple

- Le coefficient β (pente) est plus intéressant que a (ordonnée à l'origine). On se concentre donc ici sur le premier
- On peut montrer que

$$\mathrm{Var}(\widehat{\beta}_n) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$$

- On estime σ^2 par S^2 pour estimer cette variance. En prenant la racine carrée on obtient **l'erreur type** (standard error)

$$\widehat{\mathrm{sd}}(\widehat{\beta}_n) = \frac{S_n}{\sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}$$

- On peut montrer que

$$\frac{\widehat{\beta}_n - \beta}{\widehat{\mathrm{sd}}(\widehat{\beta}_n)} \sim t_{n-2}$$

On a donc un pivot. On peut construire des intervalles de confiance et tester des hypothèses

Intervalles de confiance pour β

On en déduit des intervalles de confiance pour β au niveau de confiance $1-\alpha$, pour $\alpha \in (0,1)$:

- Intervalle de confiance bilatéral symétrique :
$$\left[\widehat{\beta}_n - t_{n-2,1-\alpha/2}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}, \widehat{\beta}_n + t_{n-2,1-\alpha/2}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right].$$
- Intervalle de confiance unilatéral à gauche :
$$\left(-\infty, \widehat{\beta}_n + t_{n-2,1-\alpha}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right].$$
- Intervalle de confiance unilatéral à droite :
$$\left[\widehat{\beta}_n - t_{n-2,1-\alpha}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}, \infty\right).$$

Comparer avec diapositive 184 : $[\widehat{\theta} \pm t_{k,1-\alpha/2}\widehat{sd}(\widehat{\theta})]$: en 184, $k = n-1, \widehat{\theta} = \overline{Y}_n$, ici $k = n-2, \widehat{\theta} = \widehat{\beta}_n$

Tests pour β

On peut effectuer les tests statistiques classiques au niveau de significativité α, pour $\alpha \in (0, 1)$:

- Test bilatéral $H_0 : \beta = \beta_0$ contre $H_1 : \beta \neq \beta_0$. On rejette H_0 si et seulement si $|t_{\text{obs}}| > t_{n-2,1-\alpha/2}$.
- Test unilatéral à gauche $H_0 : \beta = \beta_0$ contre $H_1 : \beta < \beta_0$. On rejette H_0 si et seulement si $t_{\text{obs}} < t_{n-2,1-\alpha}$.
- Test unilatéral à droite $H_0 : \beta = \beta_0$ contre $H_1 : \beta > \beta_0$. On rejette H_0 si et seulement si $t_{\text{obs}} > t_{n-2,1-\alpha}$.

La statistique de test est

$$T = \frac{\widehat{\beta_n} - \beta_0}{\widehat{\text{sd}}(\widehat{\beta_n})} = \frac{\widehat{\beta_n} - \beta_0}{S_n / \sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}$$

qui suit la loi t_{n-2} quand H_0 est vraie

Nos données

```
> JungOzone
   Observed Model
1        NA 49.42
2      40.7 52.79
3        NA 56.49
4        NA 56.61
5      61.8 57.22
6        NA 53.59
7        NA 56.61
8        NA 52.75
9        NA 52.15
10       NA 45.43
...
> MM <- data.frame(
+   x=c(25, 50, 100, 150, 200, 300, 500),
+   y=c(6.9, 9.0, 11.2, 12.3, 13.6, 14.1, 14.5))
> MM
    x    y
1  25  6.9
2  50  9.0
3 100 11.2
4 150 12.3
5 200 13.6
6 300 14.1
7 500 14.5
```

Inférence

Voici le résultat de l'ajustement du modèle linéaire aux données d'ozone :

```
> fit <- lm(Observed~Model,data=JungOzone)
> summary(fit)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.51072    3.98014  -1.385    0.168
Model        1.06903    0.07479  14.294   <2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 5.334 on 205 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-Squared: 0.4992,Adjusted R-squared: 0.4967
F-statistic: 204.3 on 1 and 205 DF,  p-value: < 2.2e-16
```

Exemple : données d'ozone (inférence)

- On sait d'après les slides précédentes que l'intervalle de confiance bilatéral symétrique pour β au niveau de confiance $1-\alpha$ est
$$\left[\hat{\beta}_n - t_{n-2,1-\alpha/2}\hat{\mathrm{sd}}(\hat{\beta}_n), \hat{\beta}_n + t_{n-2,1-\alpha/2}\hat{\mathrm{sd}}(\hat{\beta}_n)\right].$$
- Ainsi, en lisant les sorties du logiciel, on obtient qu'une réalisation de l'IC précédent pour β au niveau de confiance 95% est donnée par
$$1.06903 \pm t_{205,0.975} \times 0.07479 \approx 1.07 \pm 1.97 \times 0.07 = [0.93, 1.21].$$
- Souvent, on veut tester si le terme impliquant la covariable est significatif. Cela revient à tester $H_0 : \beta = 0$.
- Ici, le scatter plot semble clairement indiquer que β est différent de 0 et on effectue donc plutôt le test $H_0 : \beta = 1$. On choisit comme niveau de significativité $\alpha = 0.05$. On rejette H_0 si et seulement si la valeur absolue de la réalisation t_{obs} de
$$T = \frac{\hat{\beta}_n - 1}{\hat{\mathrm{sd}}(\hat{\beta}_n)}$$
est strictement supérieure à $t_{n-2,1-\alpha/2} = t_{205,0.975} \approx 1.97$. On a $t_{\mathrm{obs}} \approx 0.92$ et on ne rejette donc pas H_0.

Modèle nonlinéaire (non-examinable)

- Les mêmes idées s'appliquent aux modèles nonlinéaires, mais comme approximations
- Il faut donner des valeurs initiales pour γ_0 et γ_1, en principe il faut en essayer plusieurs, car il est possible que la vraisemblance ait des maximas locaux
- Pour ajuster le modèle $\mu(x) = \gamma_0 x/(\gamma_1 + x)$ aux données chimiques :

```
> fit <- nls(y~g0*x/(g1+x),data=MM, start=c(g0=1,g1=1))
> summary(fit)

Formula: y ~ g0 * x/(g1 + x)

   Estimate Std. Error t value Pr(>|t|)
g0  15.5269     0.2876   53.99 4.12e-08 ***
g1  34.5990     2.8777   12.02 7.02e-05 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.3341 on 5 degrees of freedom
```

Coefficient de détermination

- Nous avons dêjà vu la **décomposition de la somme des carrés** total

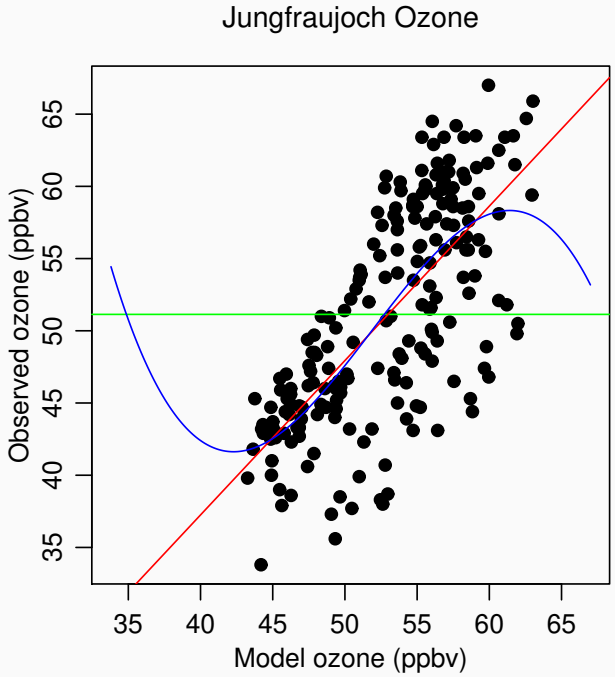
$$\sum_{j=1}^{n}(y_j-\bar{y})^2=\sum_{j=1}^{n}(\widehat{y}_j-\bar{y})^2+\sum_{j=1}^{n}r_j^2,\quad \text{soit}\quad \mathrm{SC}_{\mathrm{Total}}=\mathrm{SC}_{\mathrm{R}}+\mathrm{SC}_{\mathrm{E}},$$

  en une partie SC_{R} due à la régression et une partie SC_{E} due à l'erreur
- Le proportion (ou pourcentage) de la variation totale expliquée par le modèle

$$R^2=\frac{\mathrm{SC}_{\mathrm{R}}}{\mathrm{SC}_{\mathrm{Total}}}=\frac{\mathrm{SC}_{\mathrm{Total}}-\mathrm{SC}_{\mathrm{E}}}{\mathrm{SC}_{\mathrm{Total}}}$$

  est appelé **coefficient de détermination** ; $0\leq R^2\leq 1$
- Si $R^2\approx 1$, alors $y_j\approx\widehat{y}_j$ pour tout j et donc tous les $r_j\approx 0$, et donc le modèle explique les données presque parfaitement
- Si $R^2\approx 0$, alors l'inclusion de x n'explique presque rien de la variation totale
- Pour les données d'ozone, $R^2=0.5$, donc la moitié de la variance est expliquée par le modèle
- Pour les données chimiques, $R^2=0.99$, donc le modèle explique presque toute la variation

Comparaison des modèles



- Voici trois modèles :

$$\begin{aligned} \text{constant (vert)} : \quad y &= a + \epsilon, \\ \text{linéaire (rouge)} : \quad y &= a + \beta x + \epsilon, \\ \text{cubique (bleu)} : \quad y &= a + \beta x + \gamma x^2 + \delta x^3 + \epsilon? \end{aligned}$$

- Le rouge semble être bien meilleur que le vert, mais que le rouge et le bleu semblent être semblables. Comment tester ces constats ?

Décomposition de la variance

- Comparons le modèle constante $y = a + \epsilon$ et le modèle linéaire $y = a + \beta x + \epsilon$
- Pour tester s'il vaut la peine d'ajouter βx, on calcule

$$F = \frac{\mathrm{SC_R}/1}{\mathrm{SC_E}/(n-2)} \quad \sim \quad F_{1,n-2}$$

  si l'hypothèse nulle $H_0 : \beta = 0$ que le modèle est constant est vraie
- F_{d_1,d_2} est la **loi de Fisher(–Snedecor)** avec d_1 et d_2 degrés de liberté
- Pour un niveau de significativité $\alpha \in (0,1)$ donné, il faut comparer la valeur observée de F avec le $1-\alpha$ quantile $F_{1,n-2,1-\alpha}$ (rejet pour grandes valeurs de F)
- Pour les données d'ozone, on trouve $f_{obs} = 204.32$, à comparer avec $F_{1,205,0.95} = 3.887$
- Ce test est équivalent au t-test pour $H_0 : \beta = 0$ vu précédemmant car : $T \sim t_\nu \Longrightarrow T^2 \sim F_{1,\nu}$

Statistique F

- Pour tester $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$ dans le modèle linéaire

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_q x_i^{(q)} + \beta_{q+1} x_i^{(q+1)} + \cdots \beta_p x_i^{(p)} + \epsilon,$$

on a deux sommes des carrés, l'un $\mathrm{SC}_{E,p}$ qui correspond au modèle avec $x^{(1)}, \ldots, x^{(p)}$ et l'autre $\mathrm{SC}_{E,q}$ qui correspond au modèle réduit avec $x^{(1)}, \ldots, x^{(q)}$, $q < p$. On a $\mathrm{SC}_{E,p} \leq \mathrm{SC}_{E,q}$, et pour tester H_0 on calcule

$$F = \frac{(\mathrm{SC}_{E,q} - \mathrm{SC}_{E,p})/(p-q)}{\mathrm{SC}_{E,p}/(n-p-1)} \sim F_{p-q,n-p-1}$$

si $H_0 : \beta = 0$ est vraie

- On rejette H_0 au niveau α si $f_{obs} > F_{p-q,n-p-1,1-\alpha}$
- Pour les données d'ozone, pour tester $\gamma = \delta = 0$ dans le modèle cubique

$$y = a + \beta x + \gamma x^2 + \delta x^3 + \epsilon,$$

on a $n = 207$, $p = 3$, $q = 1$, et

$$F = \frac{(5831.9 - 5712.2)/(3-1)}{5712/(207-3-1)} = 2.13 \sim F_{3-1,207-3-1} = F_{2,203},$$

dont le 0.95 quantile est $F_{2,203,0.95} = 3.04$.

Validation du modèle de régression linéaire (non-examinable)

- Le modèle normale $y \sim \mathcal{N}\left\{\mu(x), \sigma^2\right\}$ implique que

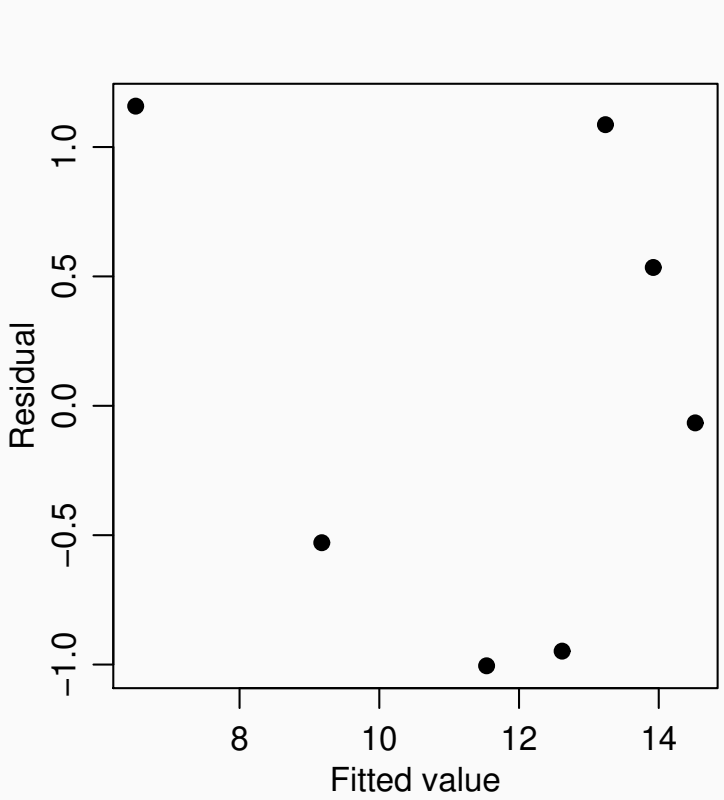
$$\frac{y - \mu(x)}{\sigma} \sim \mathcal{N}(0, 1),$$

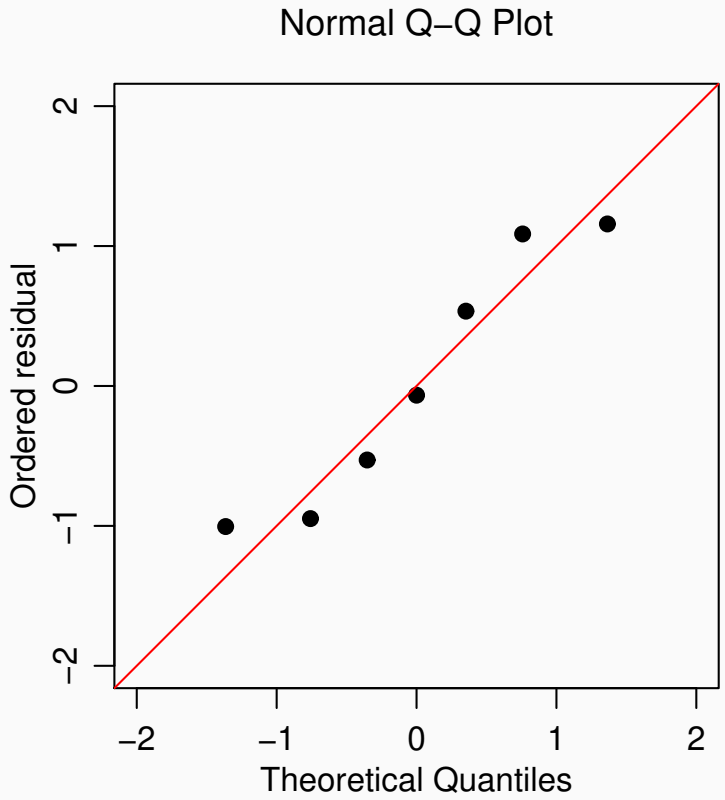
et donc que le **résidu standardisé**

$$r_j^S = \frac{r_j}{s_n} = \frac{y_j - \widehat{y}_j}{s_n} = \frac{r_j}{s_n} = \frac{y_j - (\widehat{a}_n + \widehat{\beta}_n x_j)}{s_n} \overset{\text{app}}{\sim} \mathcal{N}(0, 1)$$

- On teste cela graphiquement avec un quantile-quantile plot (Q-Q plot) normal. C'est un graphique des quantiles empiriques des données (ici les résidus standardisés) contre les quantiles théoriques d'une loi $\mathcal{N}(0, 1)$. Si les r_i^S suivent effectivement la loi $\mathcal{N}(0, 1)$, alors les points du Q-Q plot doivent se trouver (plus ou moins) sur la diagonale $y = x$. Des écarts trop importants par rapport à la diagonale indiquent une violation de l'hypothèse de normalité des erreurs.
- Par ailleurs, il faut qu'il n'y ait pas de relation entre les r_j^S et les valeurs ajustées $\widehat{y}_j$

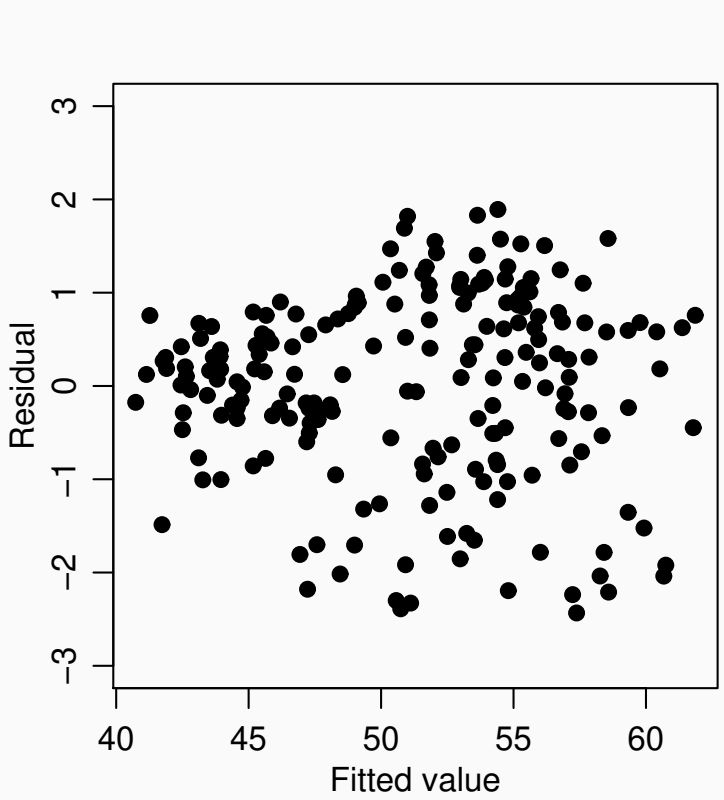
Données chimiques (non-examinable)

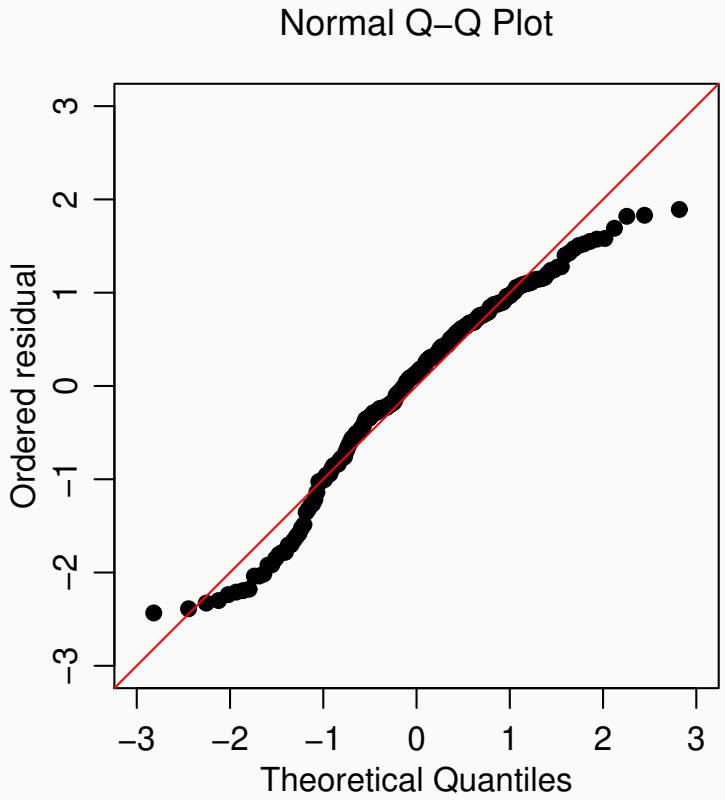




- À gauche : r_j^S contre $\widehat{y}_j$
- À droite : QQplot des r_j^S
- Avec $n = 7$, il est presque impossible de contradire le modèle

Données d'ozone (non-examinable)





- À gauche : r_j^S contre $\widehat{y}_j$
- À droite : QQplot des r_j^S
- La loi des erreurs n'est pas normale, mais asymétrique, et la variance semble changer avec $\mathbb{E}(y)$