3.3 Tests statistiques

Démarche scientifique

Toute **démarche scientifique** s'effectue selon le même schéma. Afin d'analyser la plausibilité d'une théorie, on itère les étapes suivantes :

- Enoncé d'une hypothèse (théorie) pouvant être contredite par des données.
- Récolte de données
- Comparaison des données avec les prédictions/implications de l'hypothèse.
- Non-rejet, rejet ou modification éventuelle de l'hypothèse.

Dans un cadre statistique, en supposant que l'on dispose d'un modèle pour le phénomène étudié, on itère les étapes suivantes :

- Enoncé d'une hypothèse (typiquement sur les paramètres du **modèle statistique**). Cette hypothèse peut être contredite par des données (via une statistique, appelée **statistique de test**).
- Récolte de données
- **Rejet (ou non) de l'hypothèse** à partir de la comparaison entre les données et les implications de l'hypothèse. En cas d'écart, à partir de quel seuil juge-t-on cet écart **significatif**, i.e., suffisamment important pour justifier le rejet de l'hypothèse ?

Exemple

Exemple Question : L'alcool ralentit-il les réflexes ?

Afin d'étudier l'effet de l'alcool sur les réflexes, on fait passer à 14 sujets un test de dextérité avant et après qu'ils aient consommé 100 ml de vin. Leurs temps de réaction (en ms) avant et après sont donnés dans le tableau suivant :

Sujet	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Avant	57	54	62	64	71	65	70	75	68	70	77	74	80	83
Après	55	60	68	69	70	73	74	74	75	76	76	78	81	90

Cadre statistique : [1] Hypothèse nulle et alternative

Etant donné un modèle statistique (de densité $f(x;\theta)$), nous voulons choisir entre deux théories concurrentes à propos du paramètre θ. Ces dernières forment une paire d'hypothèses :

$$H_0 : \text{l'hypothèse } \underline{\text{nulle}} \qquad \text{vs} \qquad H_1 : \text{l'hypothèse } \underline{\text{alternative}}.$$

Exemple. Dans une population décrite par la loi $\mathcal{N}(\mu, \sigma^2)$, nous pouvons former des hypothèses sur μ comme suit :

$$\underbrace{\left\{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}\right\}}_{\text{paire bilatérale}} \quad \text{ou} \quad \underbrace{\left\{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{array}\right\} \quad \text{ou} \quad \left\{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array}\right\}}_{\text{paires unilatérales}}.$$

Cadre statistique : [2] Statistique de test

Comment choisir entre les deux hypothèses ?

- Nous tirons un échantillon $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ tiré de la population. Comment l'utiliser pour prendre notre décision ?
- Nous choisissons une statistique $T = T_n = g(X_1, ..., X_n)$ qui a tendance à prendre des valeurs "typiques" sous l'hypothèse nulle H_0 (i.e., si H_0 est vraie) et "extrêmes" (dans la direction de l'hypothèse alternative H_1) sous H_1
- Ainsi, si on observe une valeur plutôt "extrême" ("extrême" dans la direction de l'hypothèse alternative H_1) de T, nous avons de l'évidence contre H_0.

Notre règle de décision est donc :

- Rejeter H_0 si la valeur observée de T est assez extrême (au-delà d'une valeur critique à déterminer).
- Ne pas rejeter H_0 si la valeur observée de T n'est pas assez extrême.

Exemple : paire bilatérale

Soient $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, où σ^2 est inconnue, et considérons la paire d'hypothèses :

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} \right\}.$$

On parle de paire bilatérale car $\mu \neq \mu_0$ est équivalent à $\mu < \mu_0$ ou $\mu > \mu_0$.

Considérons la statistique de test $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

- Si H_0 est vraie, alors $T \sim t_{n-1}$ (donc si H_0 est vraie, T prend typiquement des valeurs proches de 0).
- Compte tenu de H_1, nous considérons donc les valeurs de T comme "extrêmes" si elles sont "éloignées" de 0. Notons qu'ici, la notion d'"extrême" dans la direction de l'hypothèse alternative H_1 signifie une valeur "extrême" de la valeur absolue de T.
- Nous allons rejeter H_0 si $|T|$ est suffisamment élevée, i.e., $|T| > v^*$, où $v^* > 0$ est une valeur critique à déterminer.

Exemple : paire unilatérale

Soient $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, où σ^2 est inconnu, et considérons la paire d'hypothèses :

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right\}.$$

Considérons la statistique de test $T_n = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

- Si H_0 est vraie, alors $T \sim t_{n-1}$.
- Compte tenu de H_1, nous considérons donc les valeurs de T comme "extrêmes" si elles sont fortement négatives. Donc ici, la notion d'"extrême" dans la direction de l'hypothèse alternative H_1 signifie une valeur "extrême" de $|\min(T, 0)|$ et non de $|T|$.
- Nous allons donc rejeter H_0 si T est suffisamment négative, i.e., $T < v_*$, où $v_* < 0$ est la valeur critique à déterminer.

Cadre statistique : [3] Significativité statistique

Choix de la valeur critique (par exemple v^* et v_*) : Comment définir suffisamment élevée ou suffisamment négative. En d'autres termes, quelle ampleur est considérée comme significative ?

Pour répondre à cette question, il faut considérer les deux types d'erreurs que l'on peut commettre lorsque l'on se décide en faveur de l'une des hypothèses :

Décision \ Vérité	H_0 vraie	H_0 fausse
Non-rejet de H_0		Erreur de type II
Rejet de H_0	Erreur de type I	

Erreur de type I (**faux positif**) considérée plus grave que l'erreur de type II (faux négatif) — filtre de spam

Cadre statistique : [3] Significativité statistique

- On ne peut pas contrôler les deux erreurs à la fois :

$$\Pr(\text{erreur type I}) \text{ petite}$$
$$\Updownarrow$$
$$\text{rejet uniquement pour des valeurs très extrêmes}$$
$$\Updownarrow$$
$$\text{difficile de rejetter}$$
$$\Updownarrow$$
$$\Pr(\text{erreur type II}) \text{ grande}$$

- L'asymétrie entre la gravité des erreurs nous aide à choisir H_0 et H_1
- On va contrôler l'erreur de type I, qui est plus grave
- Parfois on choisit H_0 par convenience mathématique (de sorte que mathématiquement il serait plus facile de contrôler l'erreur de type I)

Cadre statistique : [3] Significativité statistique

- Nous choisissons la valeur maximale que l'on tolère pour la probabilité d'erreur de type I (éventuellement en tenant compte de l'avis d'un spécialiste). Cette quantité est notée α et appelée **niveau/seuil de significativité du test** ; $\alpha \in (0, 1)$. On choisit généralement une valeur faible pour α. Typiquement, $\alpha = 0.1, 0.05, 0.01, 0.001$; le plus souvent, $\alpha = 0.05$.

- La valeur critique est déterminée de manière à ce que

$$\Pr_{H_0}[\text{Rejet de } H_0] = \alpha.$$

- Ainsi, la **valeur critique** est telle que

$$\Pr_{H_0}[|T| > \text{ valeur critique}] = \alpha \quad \text{(cas bilatéral)},$$

$$\Pr_{H_0}[T < \text{ valeur critique}] = \alpha \quad \text{ou}$$

$$\Pr_{H_0}[T > \text{ valeur critique}] = \alpha \quad \text{(cas unilatéral)}.$$

- Les probabilités sont sous l'hypothèse que H_0 **est vraie** !

Cadre statistique : [3] Significativité statistique

Exemple, paire bilatérale : Soient $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, où σ^2 est inconnu, et considérons la paire $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

Nous allons rejeter H_0 si $|T| = \left| \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \right|$ est assez large, c'est à dire $|T| > v^*$.

Soit α le niveau de significativité. La valeur critique v^* satisfait

$$\Pr_{H_0}[|T| > v^*] = \alpha,$$

i.e.,

$$\Pr_{H_0}[T < -v^* \text{ ou } T > v^*] = \alpha.$$

Quand H_0 est vraie $T \sim t_{n-1}$. On doit donc choisir

$$v^* = t_{n-1,1-\alpha/2},$$

où $t_{n-1,1-\alpha/2}$ est le $(1 - \alpha/2)$ quantile de la loi de Student t_{n-1}.

Cadre statistique : [4] La p-valeur

Au lieu d'utiliser des valeurs critiques pour choisir entre H_0 et H_1, nous pouvons utiliser une autre approche, basée sur la notion de **p-valeur**.

- La p-valeur (notée p_{obs}) est la probabilité d'obtenir une valeur de la statistique de test au moins aussi élevée (élevée dans la direction de H_1) que celle que nous avons observée si H_0 était vraie.
- Supposons que la réalisation de la statistique de test sur nos données est $T = t_{\text{obs}}$. Alors :
  - Cas bilatéral : $p_{\text{obs}} = \Pr_{H_0}[|T| > |t_{\text{obs}}|]$,
  - Cas unilatéral à gauche : $p_{\text{obs}} = \Pr_{H_0}[T < t_{\text{obs}}]$,
  - Cas unilatéral à droite : $p_{\text{obs}} = \Pr_{H_0}[T > t_{\text{obs}}]$.
- Petite valeurs de p_{obs} s'opposent à H_0 car elles démontrent que la realité observée serait très improbable si l'hypothèse nulle H_0 était vraie.
- Cas bilatéral (203) : $p_{obs} = 2(1 - F_{t_{n-1}}(|t_{obs}|))$, où $F_{t_{n-1}}$ est la fonction de répartition de la loi de Student t_{n-1}.

$$p_{obs} = P_{H_0}(|T| > |t_{obs}|)$$

$$= P_{H_0}(T > |t_{obs}|) + P_{H_0}(T < -|t_{obs}|)$$

Sous H_0, $T \sim t_{n-1}$ qui est symétrique

$$= 2 P_{H_0}(T > |t_{obs}|) = 2(1 - P_{H_0}(T \leq |t_{obs}|))$$

$$= 2(1 - F_{t_{n-1}}(|t_{obs}|))$$

On peut utiliser la p-valuer pour faire un test d'hypothèse :

$$\boxed{\text{rejetter } H_0 \qquad \Longleftrightarrow \qquad p_{obs} < \alpha}$$

Exemple bilatérale (203) $p_{obs} = 2(1 - F_{t_{n-1}}(t_{obs}))$ donc

$$p_{obs} < \alpha \iff F_{t_{n-1}}(|t_{obs}|) > 1 - \alpha/2 \iff |t_{obs}| > t_{n-1,1-\alpha/2}$$

De manière générale, l'approche de la p-valeur est équivalente à l'approche des valeurs critiques. Cependant, la p-valeur p_{obs} fournit une information plus facilement interprétable que la valeur t_{obs}. Il s'agit d'une mesure de l'évidence contre H_0 contenue dans les données.

Attention : la p-valeur **n'est pas** la probabilité que H_0 soit vraie.

Résumé : les éléments d'un test

A Une hypothèse nulle H_0 à tester contre une hypothèse alternative H_1.

B Une statistique de test T, choisie de telle sorte que des valeurs "extrêmes" de T (en direction de H_1) suggèrent que H_0 est fausse. La valeur observée de T est t_{obs}.

C Un niveau de significativité α, qui la probabilité d'erreur de type I (rejet de H_0 quand H_0 est vraie) maximale que nous allons tolérer.

D1 Des valeurs critiques, telles que quand T tombe au-delà de ces valeurs, nous rejetons H_0 en faveur de H_1. Les valeurs critiques sont choisies pour respecter le niveau de significativité α.

Au lieu de D1, nous pouvons utiliser l'approche équivalente D2 :

D2 Une valeur p_{obs} donnant la probabilité d'observer une valeur de T aussi élevée que t_{obs} sous H_0. On rejette alors H_0 en faveur de H_1 quand $p_{\text{obs}} \leq \alpha$.

Exemple

Exemple On a contrôlé 10 compteurs d'électricité nouvellement fabriqués et obtenu les valeurs suivantes (en MW) :
983 1002 998 996 1002 983 994 991 1005 986.
On suppose qu'il s'agit de réalisation d'un échantillon iid d'une loi normale. On aimerait savoir s'il y a un écart entre la moyenne attendue de 1000 MW et la moyenne réelle des compteurs qui sortent de la fabrication. Nous avons obtenu $\bar{x} = 994 < 1000$. S'agit-il d'un hasard ou une faute de production ?

On va prendre $\alpha = 5\%$.

Solution Exemple 208

Supposons que nos observations $x_1, \ldots, x_n$ soient des réalisations de variables aléatoires $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, avec σ^2 inconnu. On veut tester : $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, où $\mu_0 = 1000$. On prend comme statistique de test

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{ sous } H_0 : \mu = \mu_0.$$

Dans notre cas $n = 10$, $\mu_0 = 1000$, $\bar{x} = 994$, et

$$s^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 64.88,$$

donc $t_{\text{obs}} = -2.35$. $= \dfrac{994 - 1000}{\sqrt{64.88/10}}$

On rejette H_0 si et seulement si $|t_{\text{obs}}| > t_{n-1,1-\alpha/2}$ (cas bilatéral). , $t_{n-1,1-\alpha/2} = 2.262$ (voir les tables), et comme $t_{\text{obs}} = -2.35 < -2.262$, on rejette l'hypothèse H_0.

$p_{\text{obs}} \simeq 0.046 < 0.05 \Rightarrow$ rejet

Intervalles de confiance et tests

- Soient $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ avec μ, σ inconnus
- Considérons $H_0 : \mu = 0$ et $H_1 : \mu \neq 0$
- On rejette H_0 au niveau α si et seulement si

$$|T_n| = \left| \sqrt{n} \frac{\overline{Y}_n}{S_n} \right| > t_{n-1,1-\alpha/2}$$

- Intervalle de confiance (IC) de niveau $1 - \alpha$

$$\left[\overline{Y}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n, \overline{Y}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n \right]$$

- Cet intervalle contient zéro si et seulement si $|\overline{Y}_n| \leq t_{n-1,1-\alpha/2} S_n / \sqrt{n}$ et donc

$$\text{on rejette } H_0 \iff 0 \text{ n'est pas dans l'intervalle de confiance}$$

- De manière générale, rejet de $H_0^{\theta_0} : \theta = \theta_0$ est équivalent à $\{$l'IC ne contient pas $\theta_0\}$ si on se base sur la même statistique de test
- α petit $\iff$ difficile à rejetter $\iff$ IC de niveau $1 - \alpha$ large

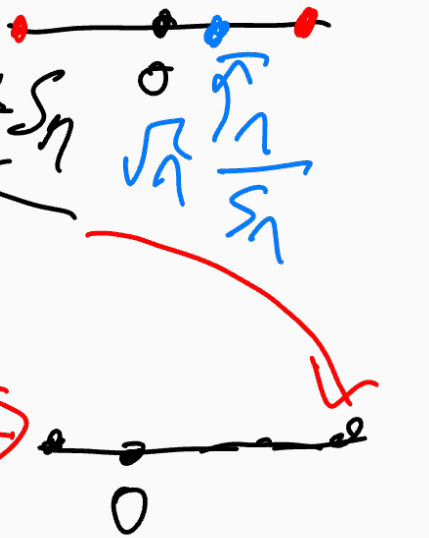
Intervalles de confiance (IC) et tests

On rejette pas $\Leftrightarrow \sqrt{n}\left|\frac{\bar{Y}_n}{S_n}\right| \leq t_{n-1,1-\frac{\alpha}{2}}$

$\Leftrightarrow -t_{n-1,1-\frac{\alpha}{2}} \leq -\sqrt{n}\frac{\bar{Y}_n}{S_n} \leq t_{n-1,1-\frac{\alpha}{2}}$

$\Leftrightarrow \bar{Y}_n - \frac{t_{n-1,1-\frac{\alpha}{2}} S_n}{\sqrt{n}} \leq 0 \leq \bar{Y}_n + \frac{t_{n-1,1-\frac{\alpha}{2}} S_n}{\sqrt{n}}$

$\Leftrightarrow 0 \in IC(1-\alpha)$



- Rejet $\iff$ $p_{obs} \leq \alpha$ $\iff$ $\theta_0 \notin$IC de niveau $1-\alpha$
- **IC** : α fixé, pour quels $\theta_0 \in \mathbb{R}$ $H_0^{\theta_0}$ n'est pas rejetée ?
- ***p*-valeur** : θ_0 fixé, pour quels $\alpha \in (0,1)$ $H_0^{\theta_0}$ est rejetée ?

$\theta = \theta_0$

3.4 Tests khi-deux

Le test khi-deux

- On se pose la question de l'adéquation d'une distribution théorique à des données
- Supposons que dans une expérience on observe n résultats différents avec des
  - **fréquences observées** dans k classes disjointes $o_1, \ldots, o_k$, alors que les
  - **fréquences théoriques** correspondantes sont $e_1, \ldots, e_k$,
  - où $\sum_{i=1}^{k} o_i = \sum_{i=1}^{k} e_i = n$
- On a H_0 : "les observations proviennent de la loi théorique spécifiée"
- Une mesure de l'écart entre les o_j et les e_j est donnée par la **statistique khi-deux**

$$T_n = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Si n est grand et les e_i ne sont pas trop petites (règle de pouce : $e_i \geq 5$ pour la plupart), alors $T_n \overset{\text{app}}{\sim} \chi^2_\nu$ sous H_0, où
  - χ^2_ν est la loi **khi-carré**, la loi de $\sum_{i=1}^{\nu} Z_i^2$ où $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
  - $\nu = k - 1$ si les e_i peuvent être calculés sans devoir estimer des paramètres inconnus
  - $\nu = k - 1 - c$ si les e_i sont calculés après avoir estimé c paramètres

Exemple

P valeur $W \sim \chi^2_5$ $p_{obs} = \Pr(W > t_{obs}) \approx 0{,}121$ $\Rightarrow$ Pas de rejet

rejet de H_0 au niveau α $\iff$ $t_{obs} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} > \chi^2_{\nu,1-\alpha}$ $8{,}5 \approx 11{,}1$

$\alpha = 0{,}05$

Exemple $n = 60$ jets d'un dé ont donné la répartition suivante :

Valeur x_i	1	2	3	4	5	6	
Valeur o_i	8	10	9	16	13	4	60
e_i	10	10	10	10	10	10	

$\nu = k - 1 = 5$

$\chi^2_{5,0{,}95}$

Ici $k = 6$ et H_0 : "équilibre du dé" est équivalente au modèle même si $\alpha = 0{,}1$

$$\Pr(X = x) = 1/6, \quad x \in \{1, \ldots, 6\}.$$

$t_{obs} = \frac{(8-10)^2}{10} + \frac{(10-10)^2}{10} + \cdots + \frac{(4-10)^2}{10} = 8{,}5$

$t_{obs} < \chi^2_{5,0{,}95} \Rightarrow$ Pas de rejet

Exemple

Exemple L'intelligence (QI) X de $n = 1000$ personnes est testée :

$a_i = 0$ $b_i = 70$

Score X	$[0, 70)$	$[70, 85)$	$[85, 100)$	$[100, 115)$	$[115, 130)$	$[130, \infty)$
Nombre o_i	34	114	360	344	120	28

e_i 22.75 135.91 341.34 341.34 135.91 22.75

Est-il plausible que $X \sim \mathcal{N}(100, 15^2)$?

On a

$$e_i = n\Pr_{\mathcal{N}(100,15^2)}(a_i \leq X \leq b_i) = n\Pr\left(\frac{a_i - 100}{15} \leq \frac{X - 100}{15} \leq \frac{b_i - 100}{15}\right)$$

$$= n\Phi\left(\frac{b_i - 100}{15}\right) - n\Phi\left(\frac{a_i - 100}{15}\right)$$

$\nu = 6 - 1 = 5$

$\alpha = 0{,}05$

$t_{obs} = \frac{(34 - 22{,}75)^2}{22{,}75} + \cdots + \approx 13{,}21 > \approx \chi^2_{5;0{,}95}$

on rejette H_0 , $p_{obs} \approx 0{,}021$

Tableaux de contingence

Un **tableau de contingence** est une classification de n objets ou individus selon plusieurs critères

- Une question fondamentale concerne **l'indépendance** des critères
- Supposons qu'on observe deux caractères A (h classes) et B (k classes) sur chacun de n individus, donnant le **tableau de contingence** suivant :

	B						
A	1	2	$\cdots$	j	$\ldots$	k	Σ
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	n_{h1}	n_{h2}	$\cdots$	n_{hj}	$\cdots$	n_{hk}	$n_{h\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	$n_{\cdot\cdot} = n$

Sommes des lignes

Sommes de colonnes

Indépendance

- Soit n_{ij} le nombre de personnes tombant dans la classe i du caractère A et dans la classe j du caractère B, et soit

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{h} n_{ij}, \quad i \in \{1, \ldots, h\}, j \in \{1, \ldots, k\}$$

- On désire tester H_0 : A **et** B **sont indépendants**. Dans ce cas

$$\Pr(A = i, B = j) = \Pr(A = i) \times \Pr(B = j), \quad i \in \{1, \ldots, h\}, j \in \{1, \ldots, k\},$$

et les probabilités empiriques sont

$$\widehat{\Pr(A = i)} = \frac{n_{i\cdot}}{n}, \quad \widehat{\Pr(B = j)} = \frac{n_{\cdot j}}{n}$$

Ainsi, sous H_0, l'(i, j)ième ~~élément du tableau de contingence est~~ valeur attendue

$$E_{ij} = n \times \Pr(A = i, B = j) = n \times \Pr(A = i) \times \Pr(B = j)$$

qu'on va estimer par

$$e_{ij} = n \frac{n_{i\cdot}}{n} \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

Calcul sous H_0

- Sous H_0 et pour n grand, la statistique T_n suit une distribution χ^2_ν avec $\nu = (h-1)(k-1)$, car on a dû estimer $c = (h-1)+(k-1)$ probabilités, et

$$hk - 1 - c = kh - 1 - (h-1) - (k-1) = (h-1)(k-1)$$

- Pour tester à un niveau de significativité α fixé, on rejette H_0 si $t_{\text{obs}} > \chi^2_{(h-1)(k-1),1-\alpha}$, sinon on ne rejette pas H_0

Exemple On a relevé parmi 95 personnes la couleur de leurs yeux (caractère A) et celle de leurs cheveux (caractère B) et on a obtenu les résultats suivants :

	B		
A	Cheveux clairs	Cheveux sombres	
Yeux bleus	$n_{11} = 32$	$n_{12} = 12$	$n_{1.} = 44$
Yeux bruns	$n_{21} = 14$	$n_{22} = 22$	$n_{2.} = 36$
Autres	$n_{31} = 6$	$n_{32} = 9$	$n_{3.} = 15$
Σ	$n_{.1} = 52$	$n_{.2} = 43$	$n = 95$

On désire tester si la couleur des cheveux est indépendante de celle des yeux

Donc on a H_0 : indépendance entre couleur des cheveux et couleur des yeux

$p_{obs} \approx 0{,}0048$

O_{ij}			
	32 $\frac{44\times52}{95}=e_{11}$	12	44
l_{ij}	14	22 $\frac{36\times43}{95}$	36
	6 $\frac{15\times52}{95}$	9	15
	52	43	95

$$t_{obs} = \frac{\left(32-\frac{44\times52}{95}\right)^2}{\frac{44\times52}{95}} + \dots + \frac{\left(9-\frac{15\times43}{95}\right)^2}{\frac{15\times43}{95}}$$

$\approx 10{,}67$ $\quad r = (k-1)(l-1) = 2$ $\quad \chi^2_{2;0,95} \approx 5{,}99 < t_{obs}$

$\leadsto$ rejet

Régularité (non-examinable)

Les conditions de régularité sont compliquées. Elles sont fausses le plus souvent quand

- un des paramètres est discret
- le support de $f(y;\theta)$ dépend de θ
- le vrai θ se trouve sur une borne des valeurs possibles

Elles sont satisfaites dans la grande majorité des cas rencontrés en pratique

Exemple Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0,\theta)$, trouver la vraisemblance $L(\theta)$ et le $\widehat{\theta}_{\text{ML}}$. Montrer que la loi limite de $n(\theta - \widehat{\theta}_{\text{ML}})/\theta$ quand $n \to \infty$ est exp(1). Discuter.

Preuve (non-examinable)

- Les fonctions de densité et de répartition de y_j sont

$$f(y;\theta) = \theta^{-1} I(0 \le y \le \theta), \quad F(y) = y/\theta, \quad 0 \le y \le \theta.$$

- L'indépendance donne

$$L(\theta) = \prod \theta^{-1} I(Y_j < \theta) = \theta^{-n} I(\max Y_j \le \theta), \quad \theta > 0,$$

qui est maximisée au point $\widehat{\theta}_{\mathrm{ML}} = M_n = \max Y_j$

- On a

$$\Pr(M_n \le x) = \prod_{j=1}^{n} \Pr(Y_j \le x) = (x/\theta)^n, \quad 0 \le x \le \theta$$

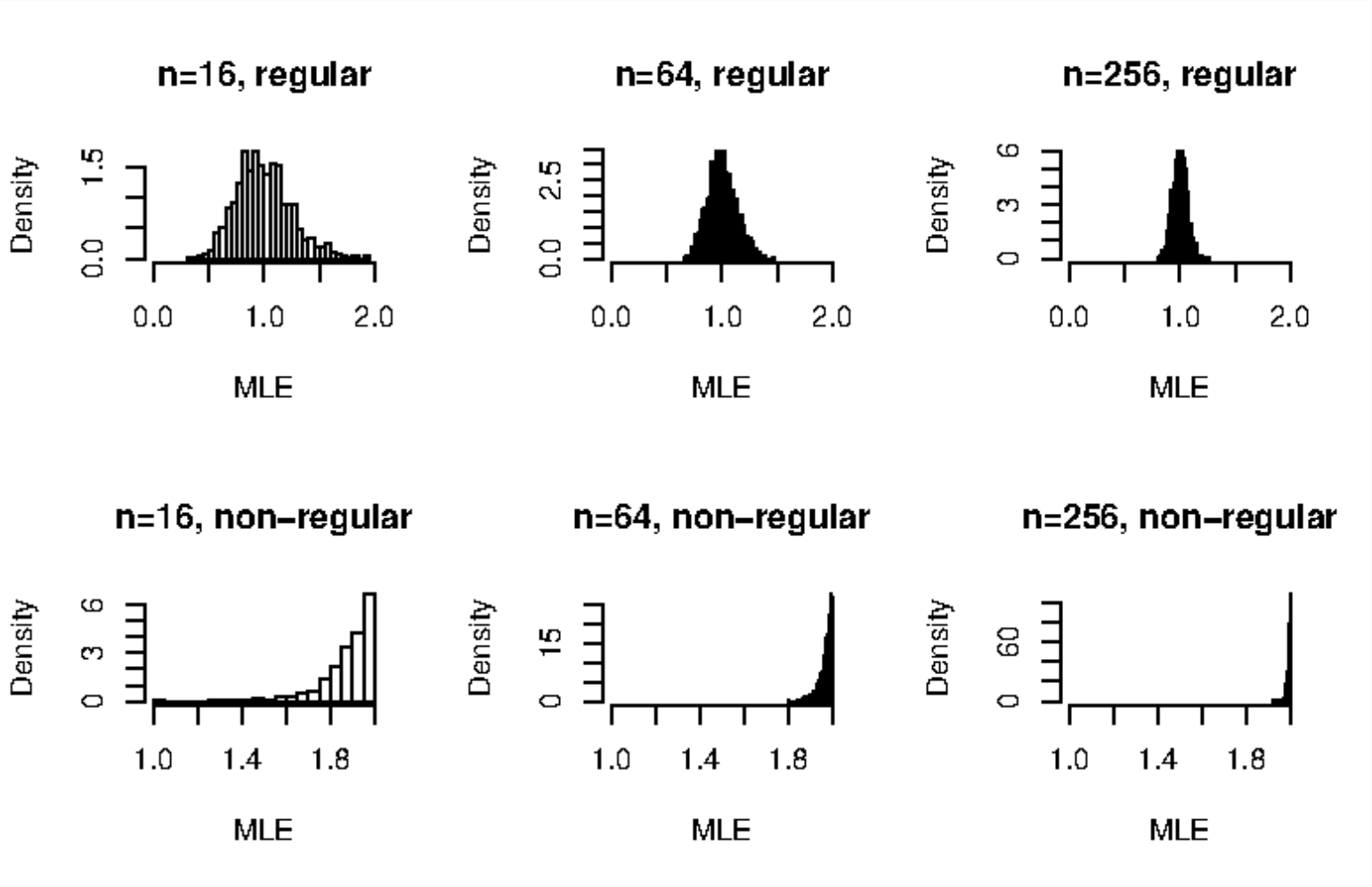
- Donc pour $x \ge 1$ et $n \ge x$,

$$\Pr\left\{n(\theta - \widehat{\theta}_{\mathrm{ML}})/\theta \le x\right\} = \Pr(\widehat{\theta}_{\mathrm{ML}} \ge \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n$$
$$\to 1 - \exp(-x),$$

comme requis

Exemple (non examinable)

Comparaison des lois de $\widehat{\theta}$ dans un cas régulier (en haut, avec écart-type $\propto n^{-1/2}$ et loi limite normale) et dans un cas non-régulier (en bas, avec écart-type $\propto n^{-1}$ et loi limite non-normale)



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

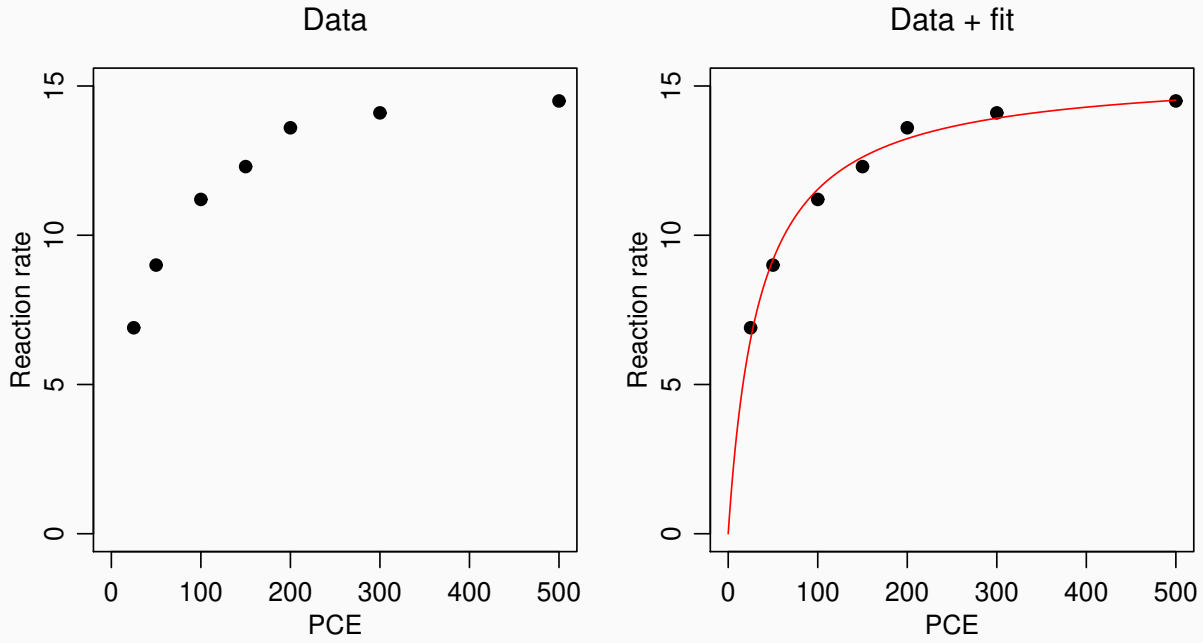
- l'**estimation** d'une relation eventuelle entre y et les $x^{(j)}$, ou
- la **prévision** des valeurs futures/manquantes de y sur la base des $x^{(j)}$ correspondantes

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

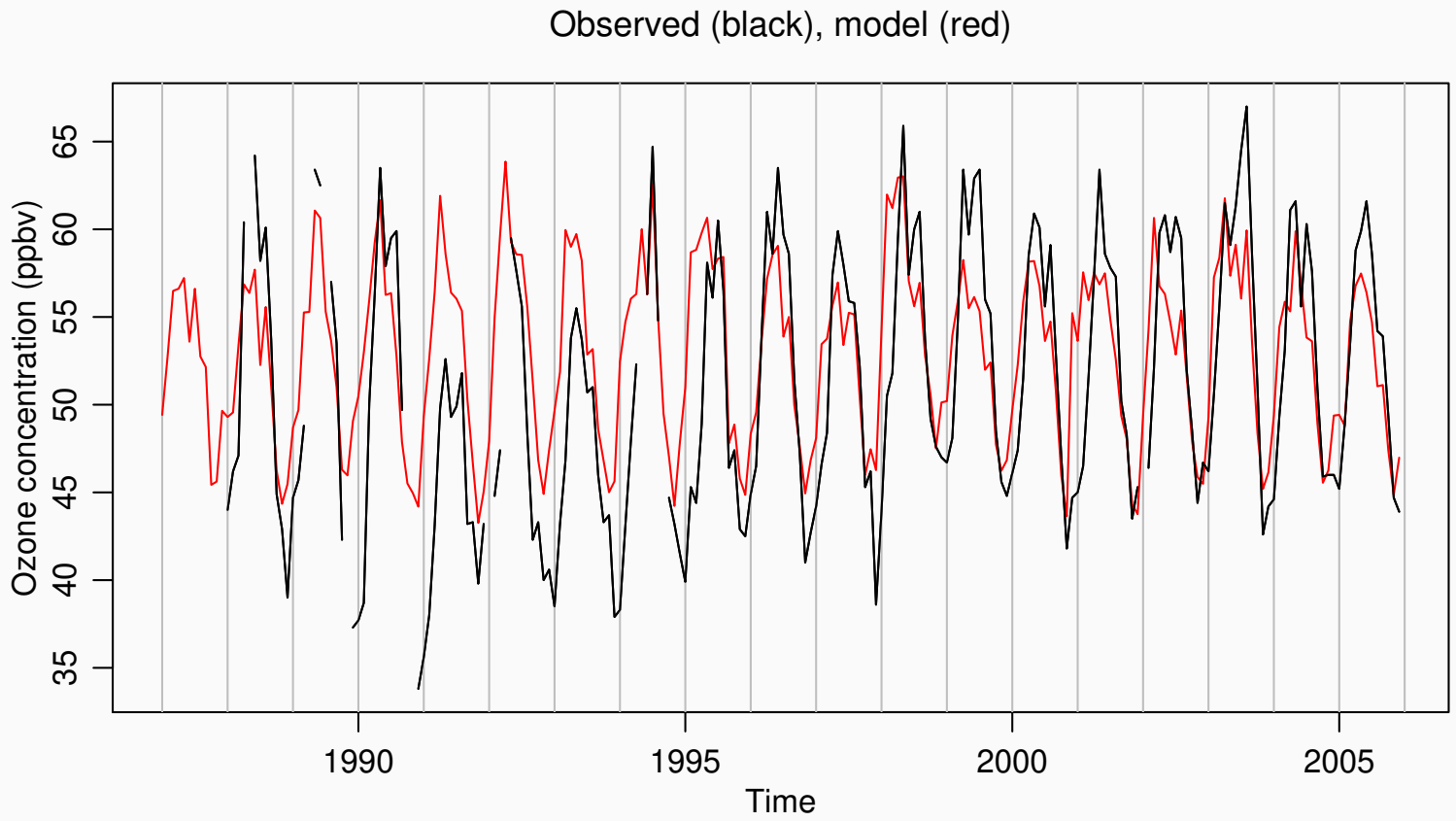
$$y = \frac{\gamma_0 x}{\gamma_1 + x},$$

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 ?



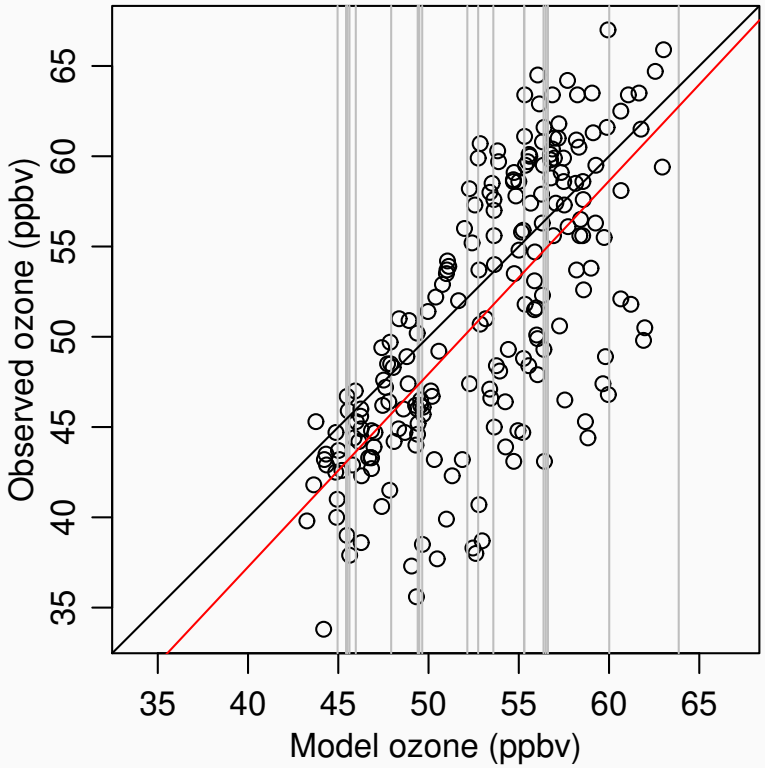
Ozone atmosphérique

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

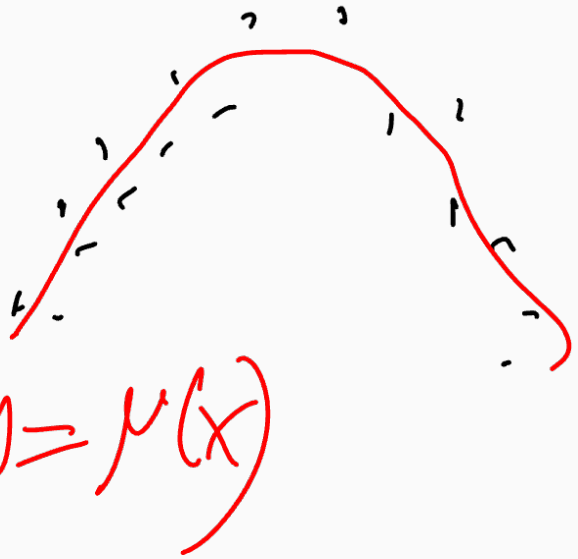
une seule covariable

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

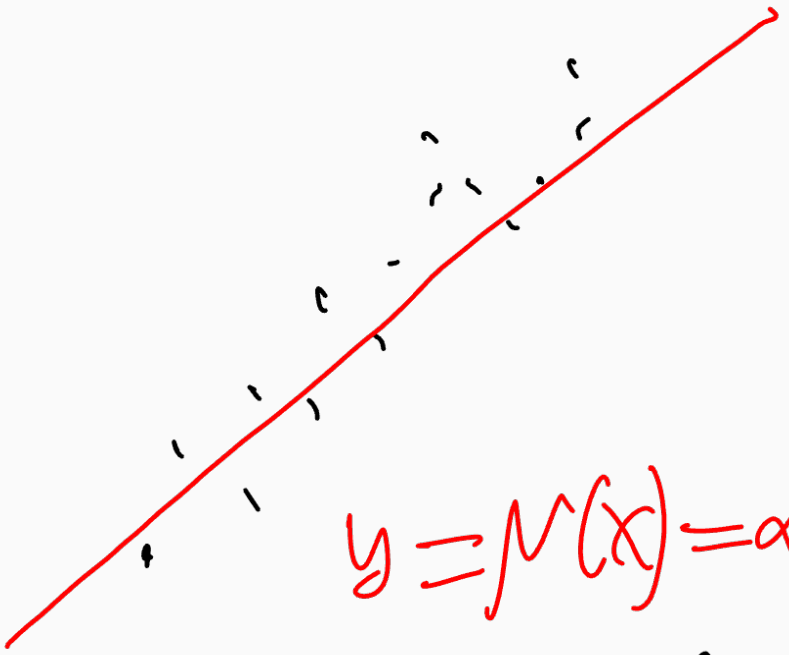
$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

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



$y = \mu(x)$



$y = \mu(x) = \alpha + \beta x$

relation linéaire

Moindres carrés

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

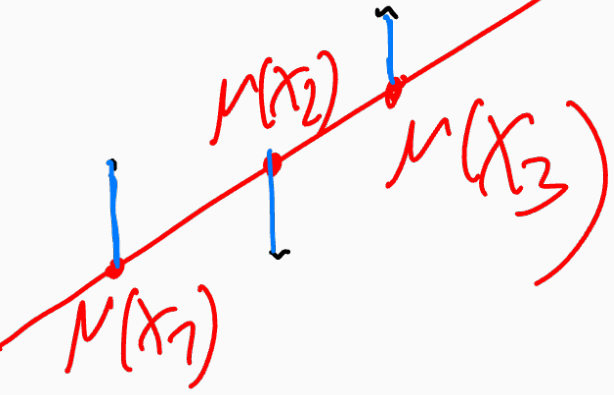
$$y_j - \mu(x_j), \quad j = 1, \ldots, n$$

- 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

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



$S(a,\beta) = \sum (\text{lignes bleues})^2$

Estimateurs de moindres carrés

Théorème Soient $(x_1, y_1), \ldots, (x_n, y_n)$ issues d'une 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

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

$$\sum (x_j - \overline{x}_n) y_j = \sum (x_j - \overline{x}_n)(y_j - \overline{y}_n) = \sum x_j (y_j - \overline{y}_n)$$

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

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

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

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

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

$$\frac{\partial}{\partial a} = 0 \longleftrightarrow 2na + 2n\beta\bar{x}_n - 2n\bar{y}_n = 0$$

$$\longleftrightarrow \boxed{a = \bar{y}_n - \beta\bar{x}_n}$$

$$\frac{\partial}{\partial \beta} = 0 \longleftrightarrow 0 = 2na\bar{x}_n + 2\beta\sum x_i^2 - 2\sum x_i y_i$$

$$= 2n\bar{x}_n(\bar{y}_n - \beta\bar{x}_n) + 2\beta\sum x_i^2 - 2\sum x_i y_i$$

$$\Rightarrow \beta\left(2\sum x_i^2 - 2n\bar{x}_n^2\right) = 2\sum x_i y_i - 2n\bar{x}_n\bar{y}_n$$

Preuve

Il faut minimiser

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

en a et β. On calcule

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

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

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

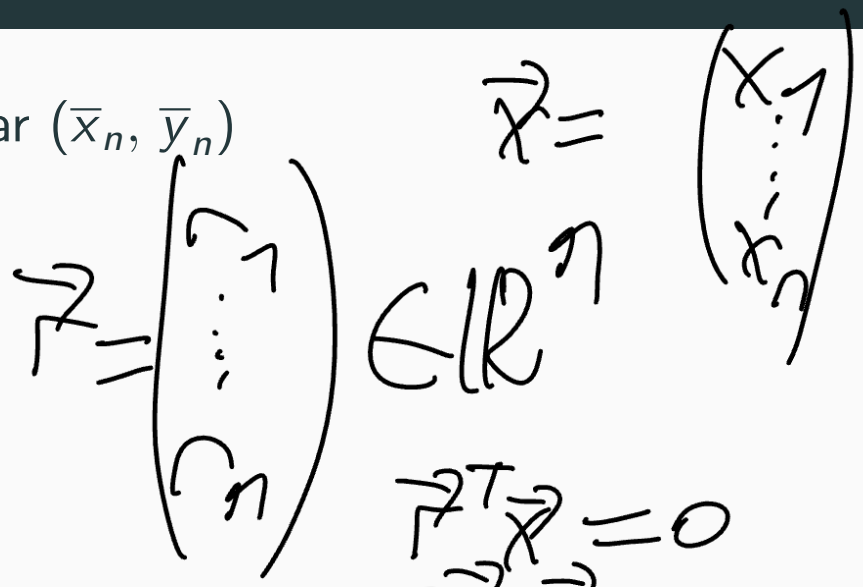
La matrice hessienne est donc

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

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 $(\overline{x}_n, \overline{y}_n)$
- $\sum_{j=1}^{n} r_j = 0$
- $\sum_{j=1}^{n} x_j\, r_j = \sum_{j=1}^{n} x_j(y_j - \widehat{y}_j) = 0$
- $\sum_{j=1}^{n} \widehat{y}_j\, r_j = 0$

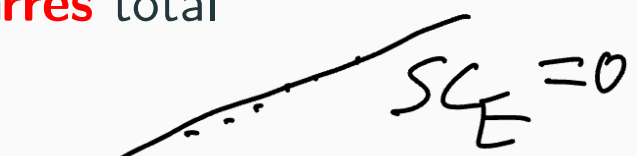


(voir série d'exercices). Donc

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

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

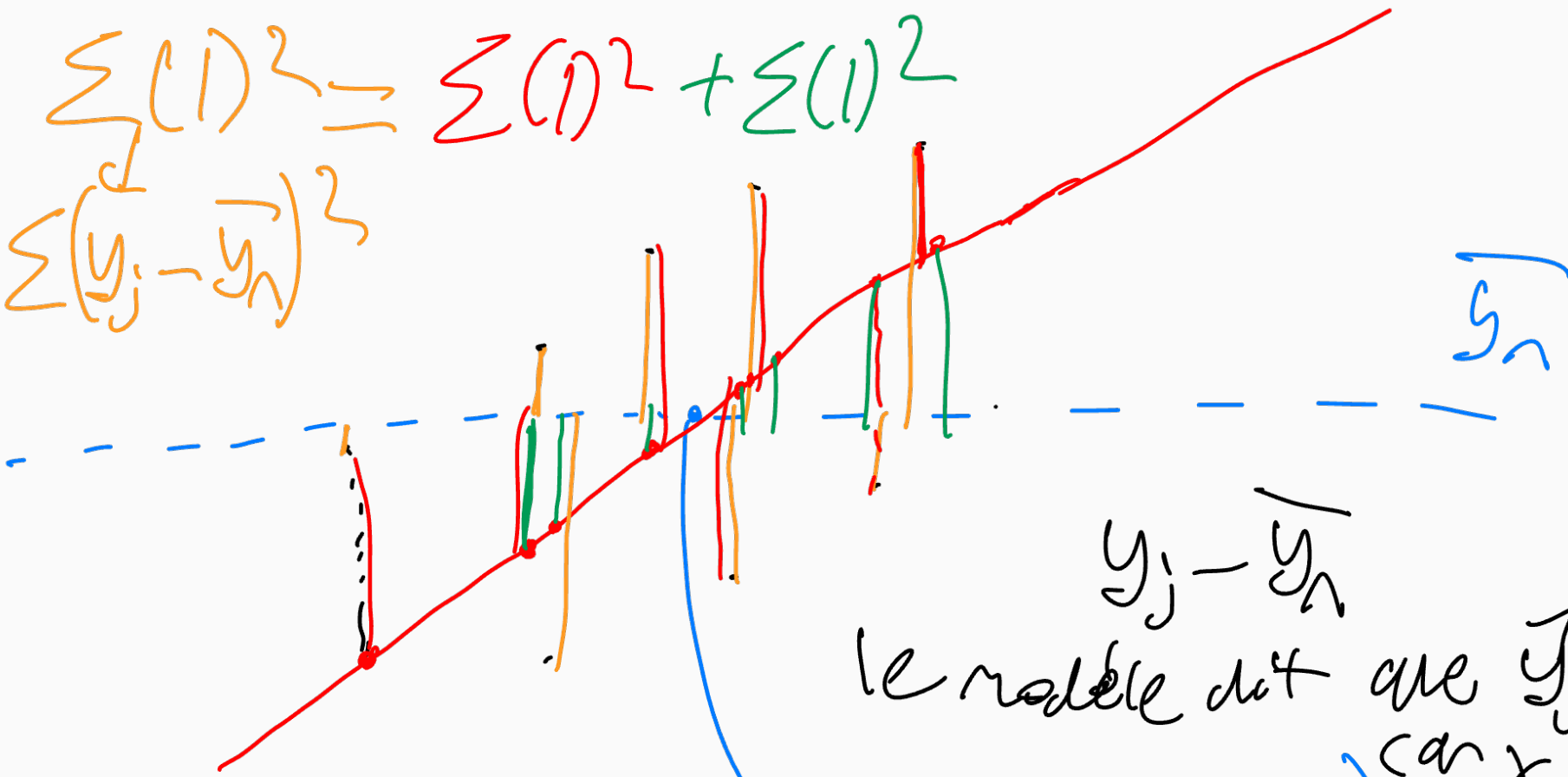
$$\mathrm{SC}_{\mathrm{Total}} = \mathrm{SC}_{\mathrm{R}} + \mathrm{SC}_{\mathrm{E}}$$



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)

$\sum (\cdot)^2 = \sum (\cdot)^2 + \sum (\cdot)^2$

$\sum (y_j - \bar{y}_n)^2$



$\bar{y}_n$

$y_j - \bar{y}_n$

le modèle dit que $y_j \neq \bar{y}_n$ car $x_j \neq \bar{x}_n$

$(\bar{x}_n, \bar{y}_n)$

$\sum (\cdot)^2 = \sum n_j^2 = SC_E$

$\sum (\cdot)^2 = \sum (y_j - \bar{y}_n) = SC_R$

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

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

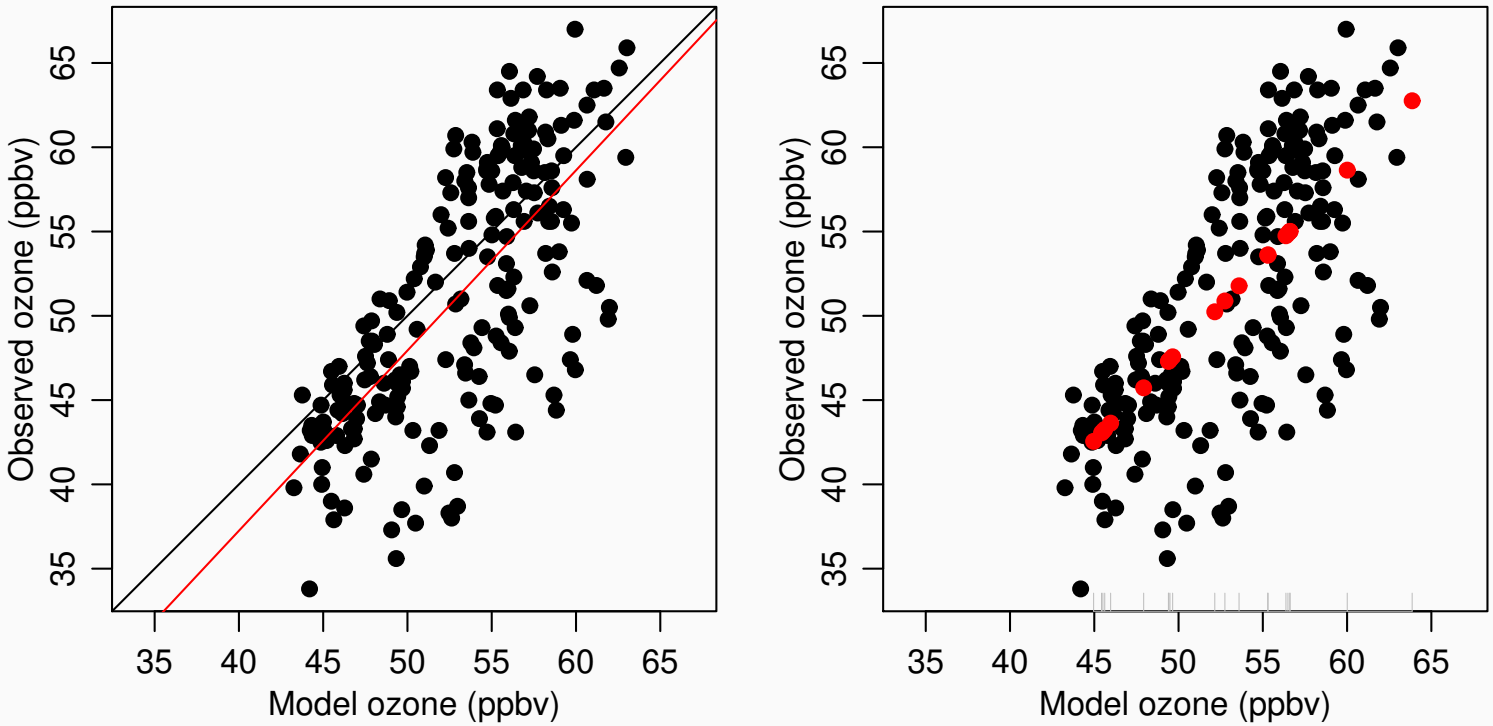
La régression explique donc une moitié de la somme des carrés totale

- 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

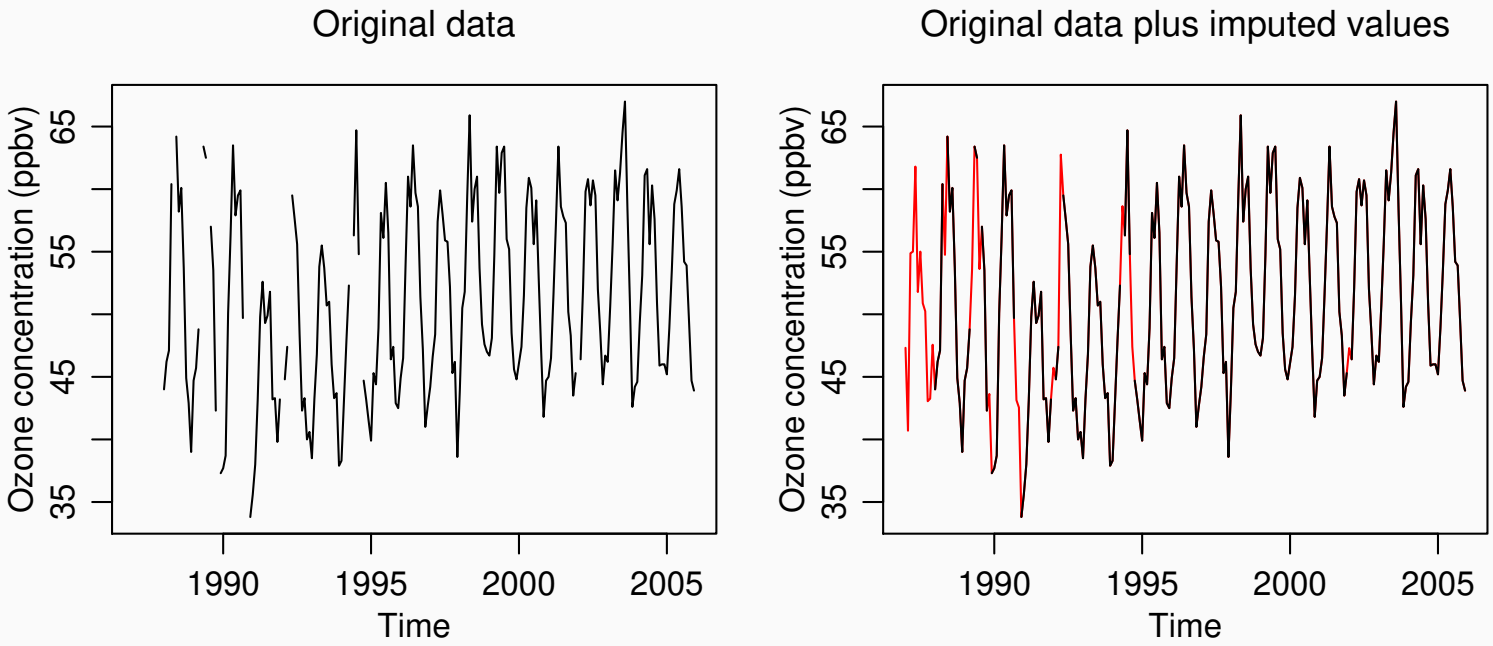
Modèle ajusté



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.

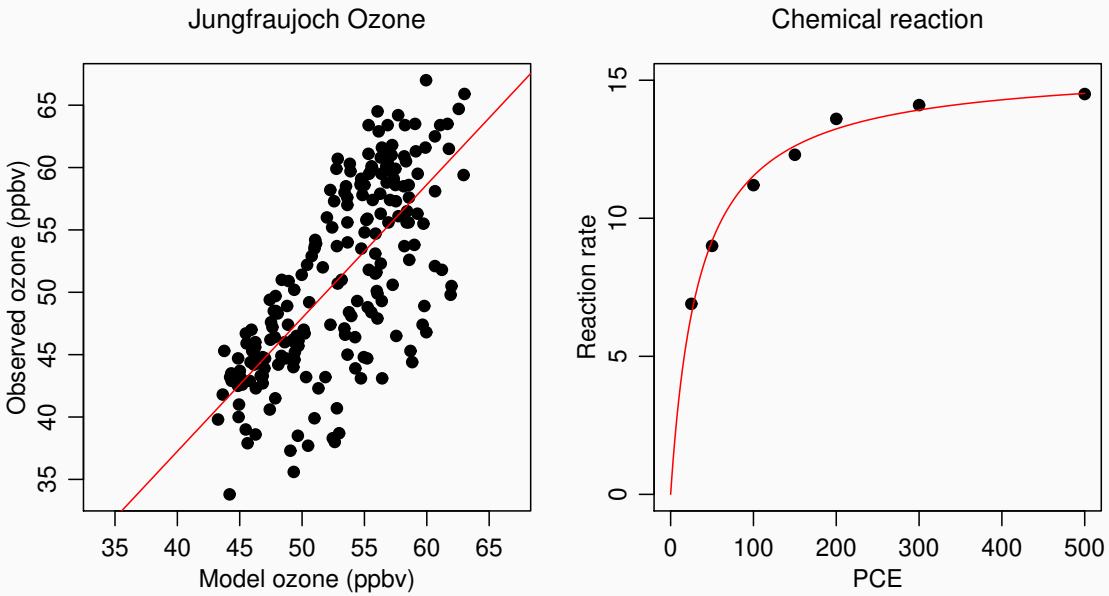
4.2 Modèle statistique

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 $(a, \beta, \gamma, \delta)$.

$y = a + \beta \sin(x) + \epsilon$,

- Le modèle

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

devient linéaire après transformation logarithmique :

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

y x

- 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 1/n)

182

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

$w = \sigma^2$ $\quad h(w) = l(\alpha_n, \beta_n, w)$

$$h(w) = -\frac{n}{2}\log w - \frac{1}{2w} S(\alpha_n, \beta_n) - \frac{n}{2}\log(2\pi)$$

$$S(\alpha_n, \beta_n) = \sum \varepsilon_j^2$$

$$\frac{\partial}{\partial w} = \frac{-n}{2w} + \frac{1}{2w^2}\sum \varepsilon_j^2 = 0 \iff w = \frac{1}{n}\sum \varepsilon_j^2$$

$$\| \\ \sigma^2$$

$$\frac{\partial^2}{\partial w^2} = \frac{n}{2w^2} - \frac{\sum \varepsilon_j^2}{w^3}$$

$\frac{\partial^2}{\partial w^2}(w) < 0 \rightarrow$ w est un maximum

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
$$1.06903 \pm t_{205,0.975} \times 0.07479 \approx 1.07 \pm 1.97 \times 0.07 \approx [0.93, 1.21].$$
- 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

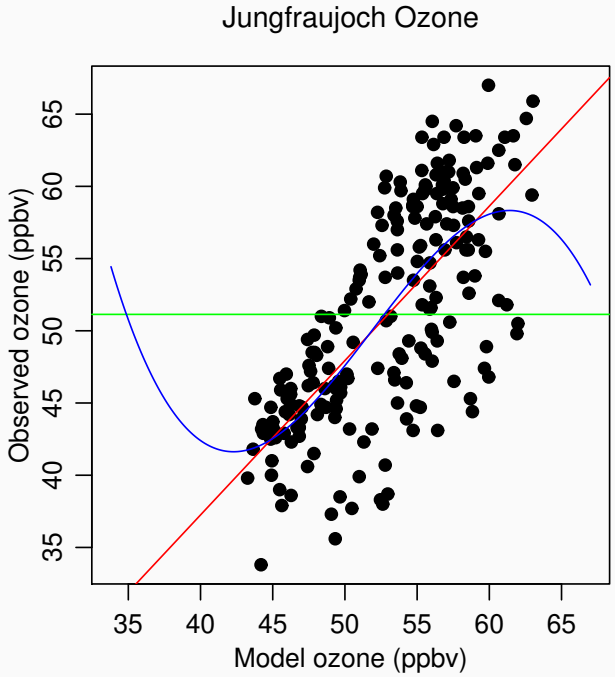
$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\widehat{y}_j - \bar{y})^2 + \sum_{j=1}^{n} r_j^2, \quad \text{soit} \quad \mathrm{SC}_{\mathrm{Total}} = \mathrm{SC}_{\mathrm{R}} + \mathrm{SC}_{\mathrm{E}},$$

  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

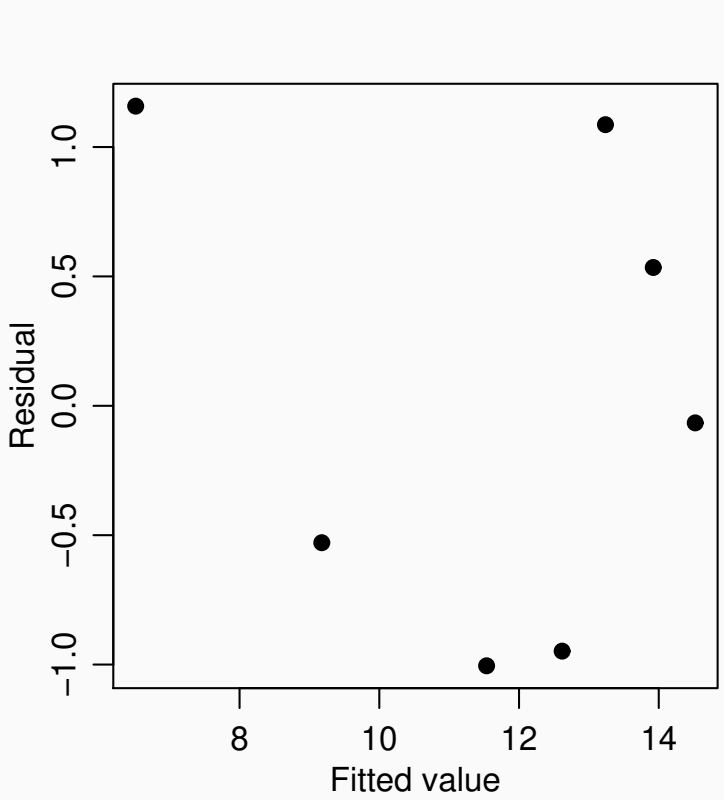
$$\frac{y - \mu(x)}{\sigma} \sim \mathcal{N}(0, 1),$$

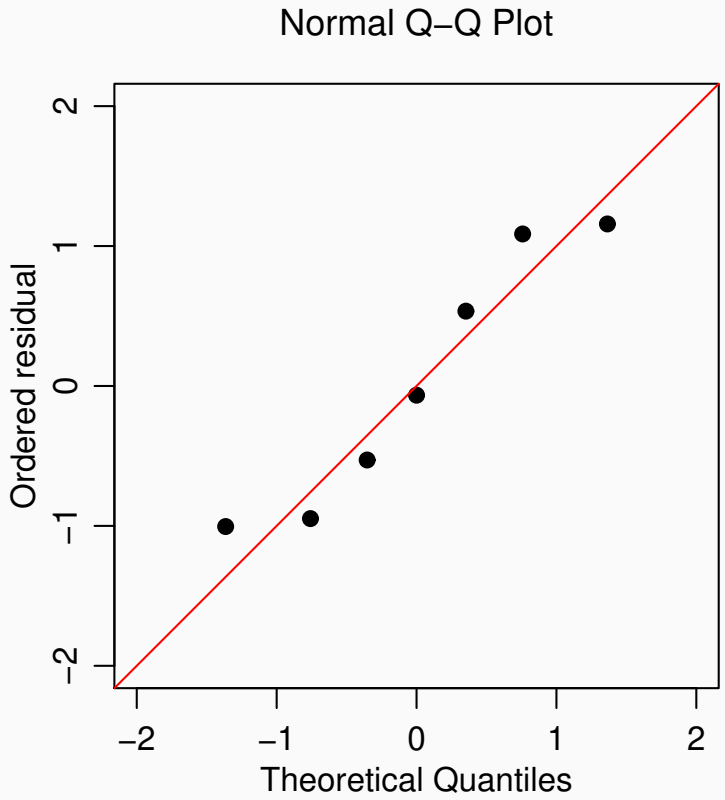
et donc que le **résidu standardisé**

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

- 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 $\widehat{y}_j$

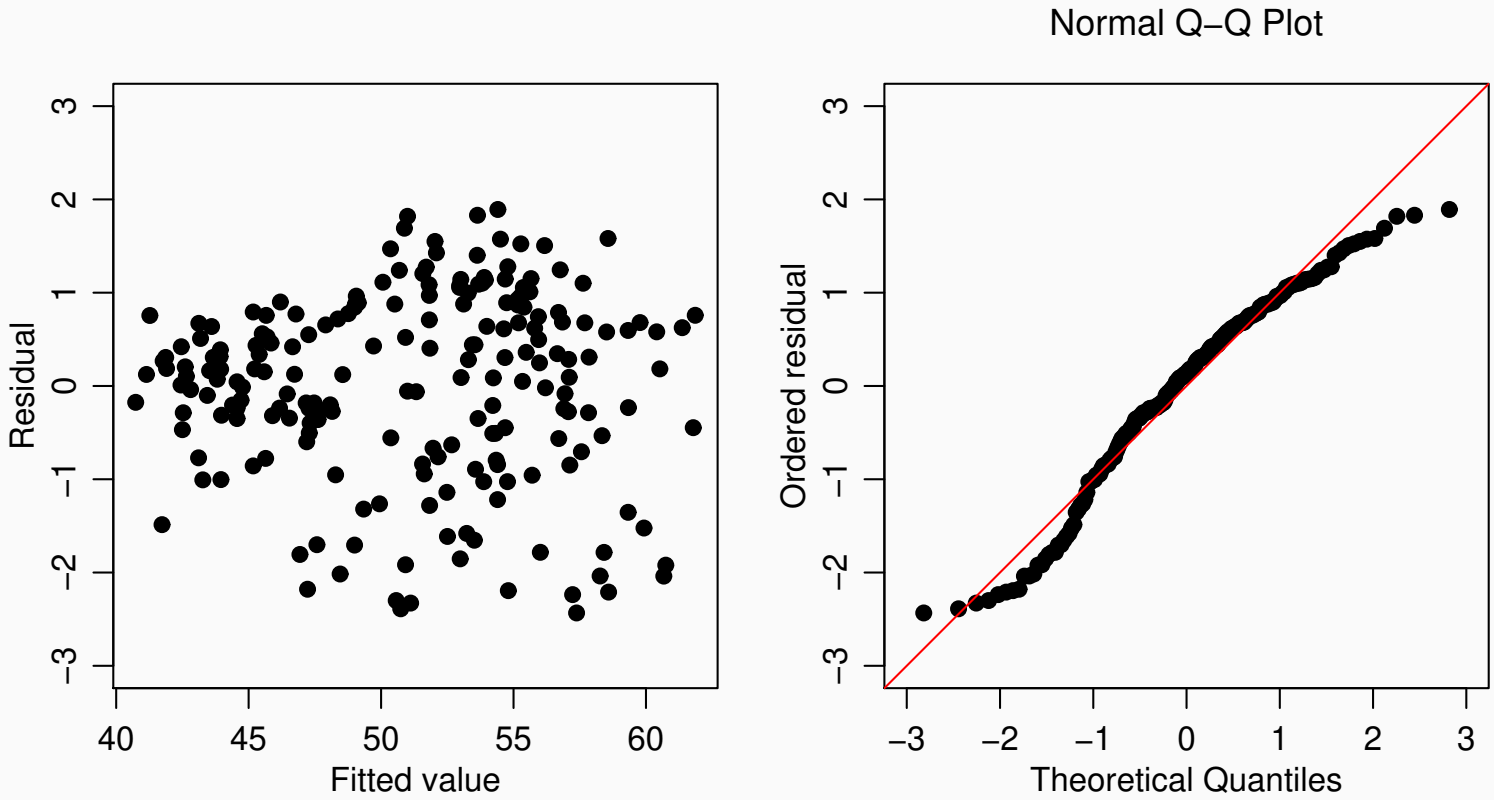
Données chimiques (non-examinable)





- 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 $\mathbb{E}(y)$