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

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

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

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

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

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

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

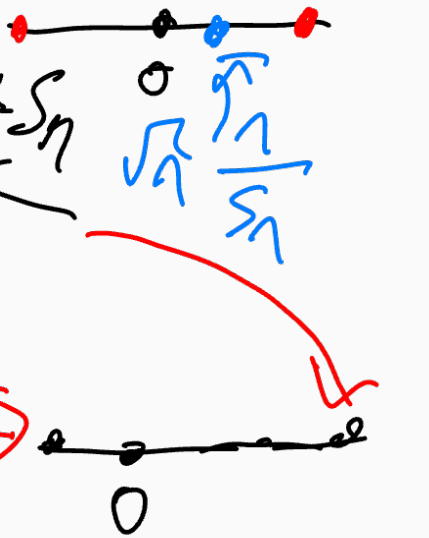
Intervalles de confiance (IC) et tests

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	

$r = 6 - 1 = 5$

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

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	12	44
	14	22	36
	6	9	15
	52	43	95

$e_{11} = \frac{44 \times 52}{95}$

$e_{12} = \frac{36 \times 43}{95}$

$e_{31} = \frac{15 \times 52}{95}$

$$t_{obs} = \frac{\left(32 - \frac{44 \times 52}{95}\right)^2}{\frac{44 \times 52}{95}} + \dots + \frac{\left(9 - \frac{15 \times 43}{95}\right)^2}{\frac{15 \times 43}{95}}$$

$\approx 10{,}67$

$r = (h-1)(k-1) = 2$

$\chi^2_{2;0,95} \approx 5{,}99 < t_{obs}$

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

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

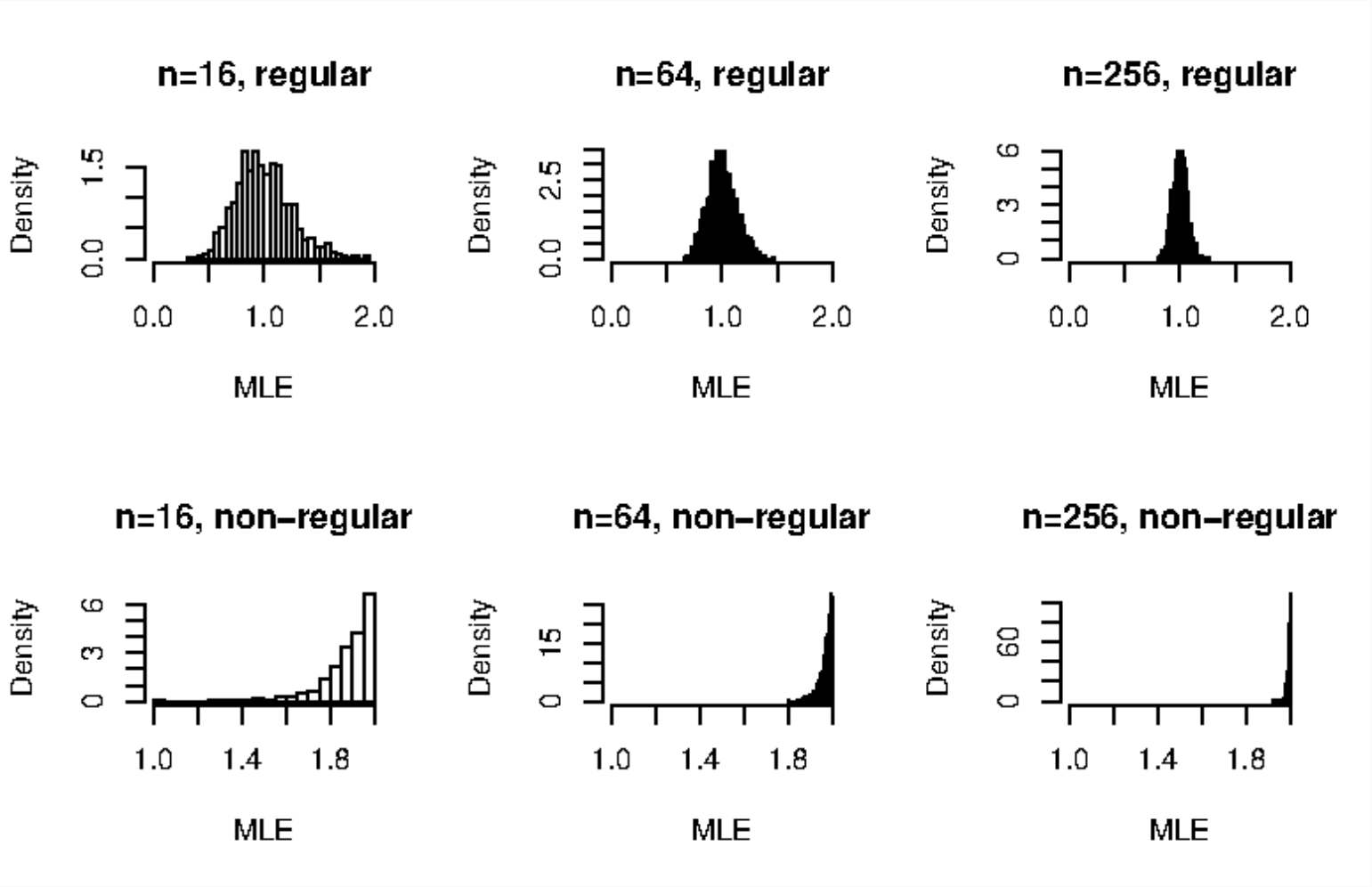
- Donc pour $x \ge 1$ et $n \ge x$,

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

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

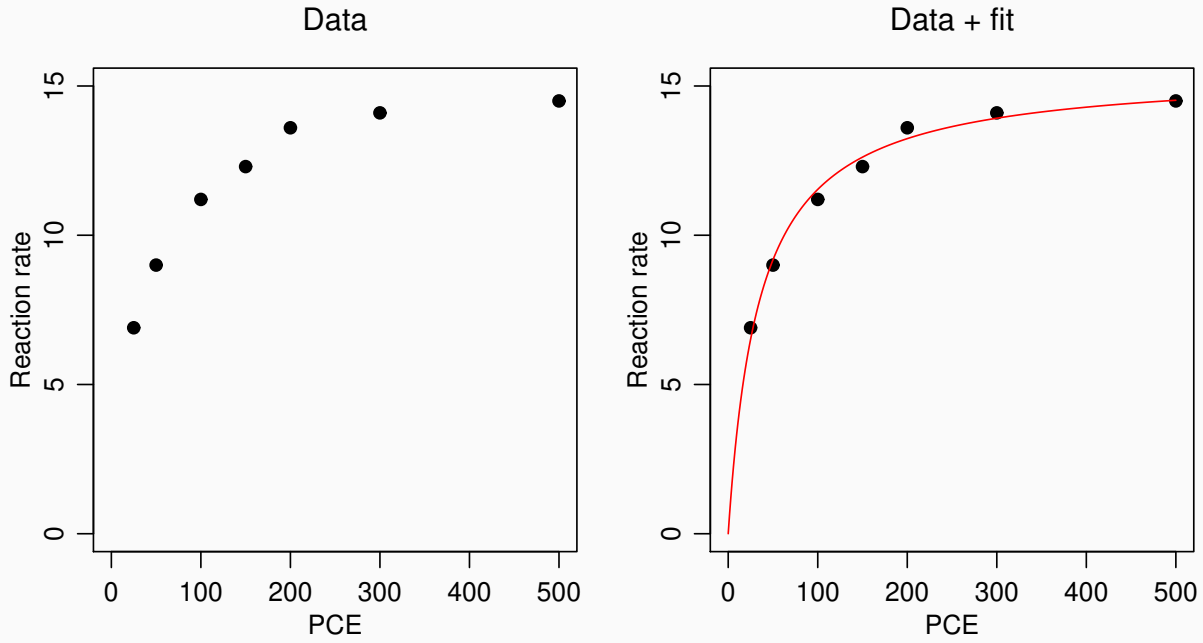
- l'**estimation** d'une relation eventuelle entre y et les $x^{(j)}$, ou
- la **prévision** des valeurs futures/manquantes de y sur la base des $x^{(j)}$ correspondantes

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

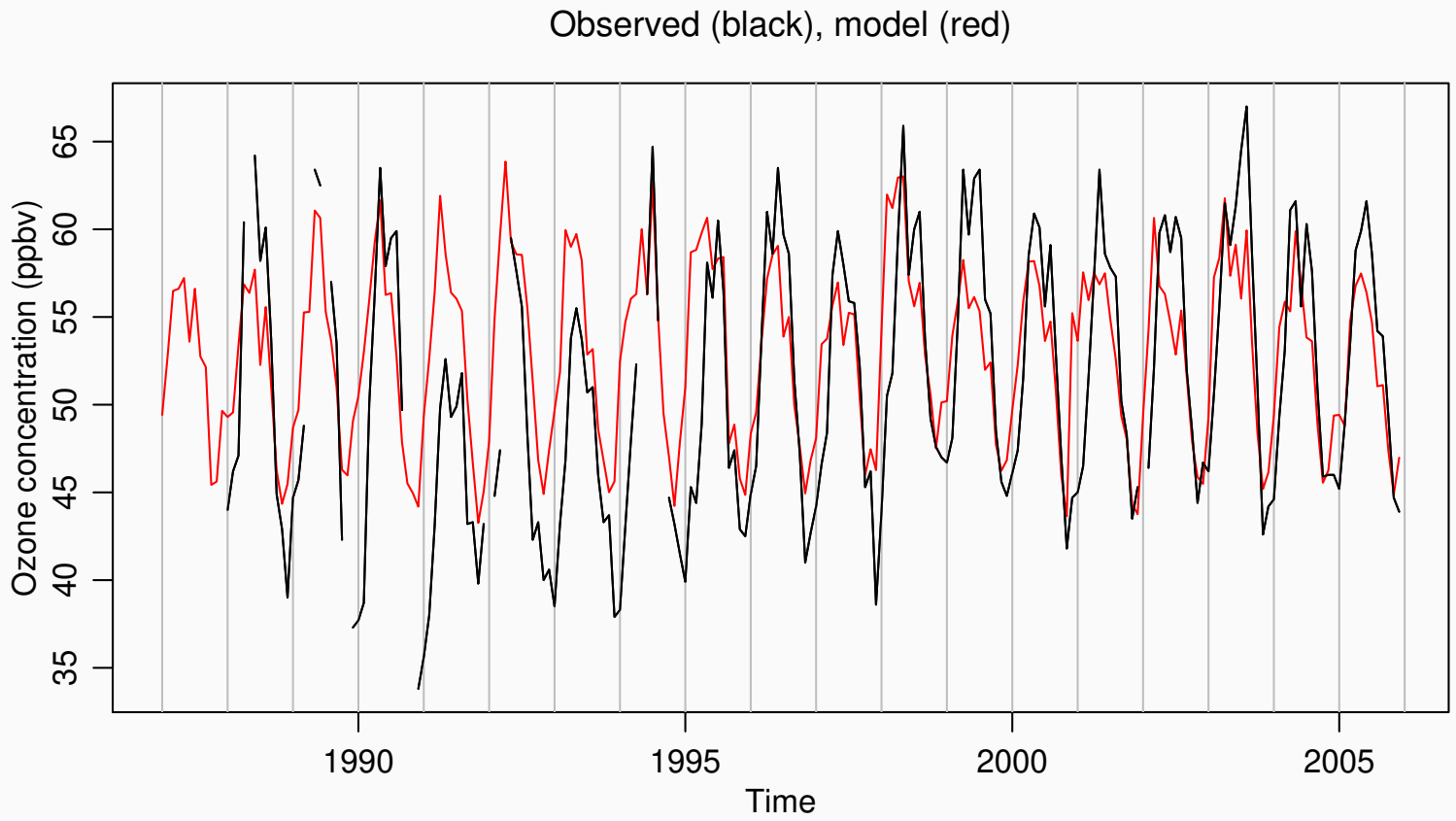
$$y = \frac{\gamma_0 x}{\gamma_1 + x},$$

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 ?



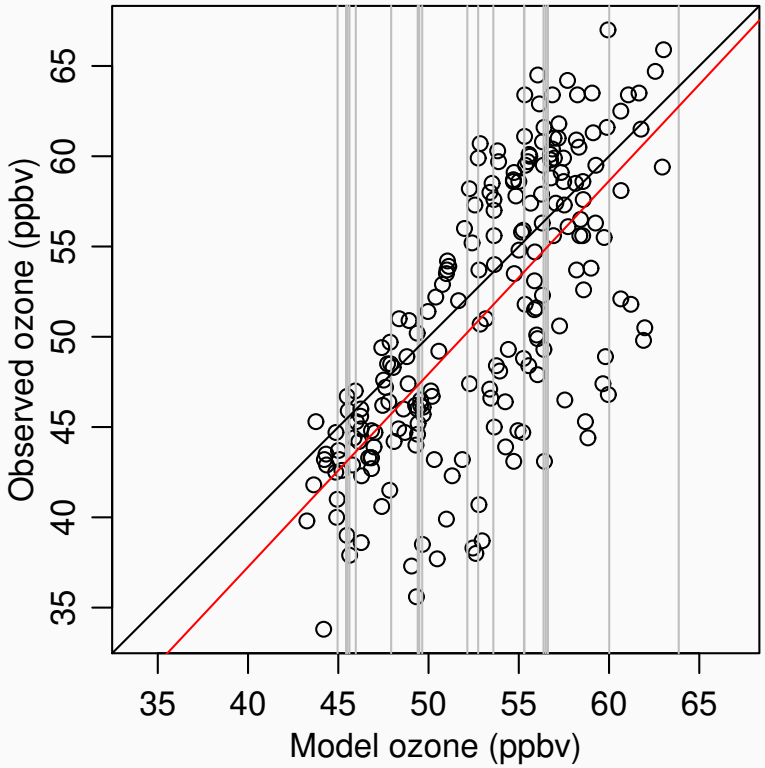
Ozone atmosphérique

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

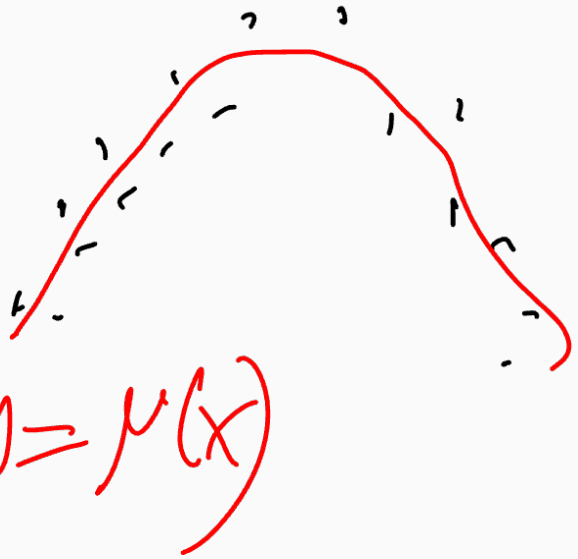
une seule covariable

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

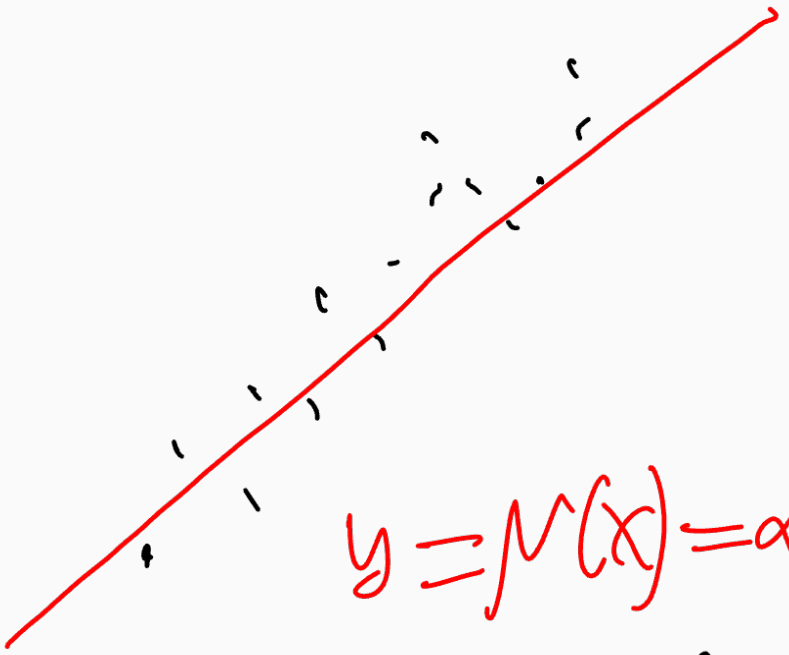
$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

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