ECOLE DES MINES DE NANTES

Modélisation stochastique et analyse de données

Formation FIL - Année 1

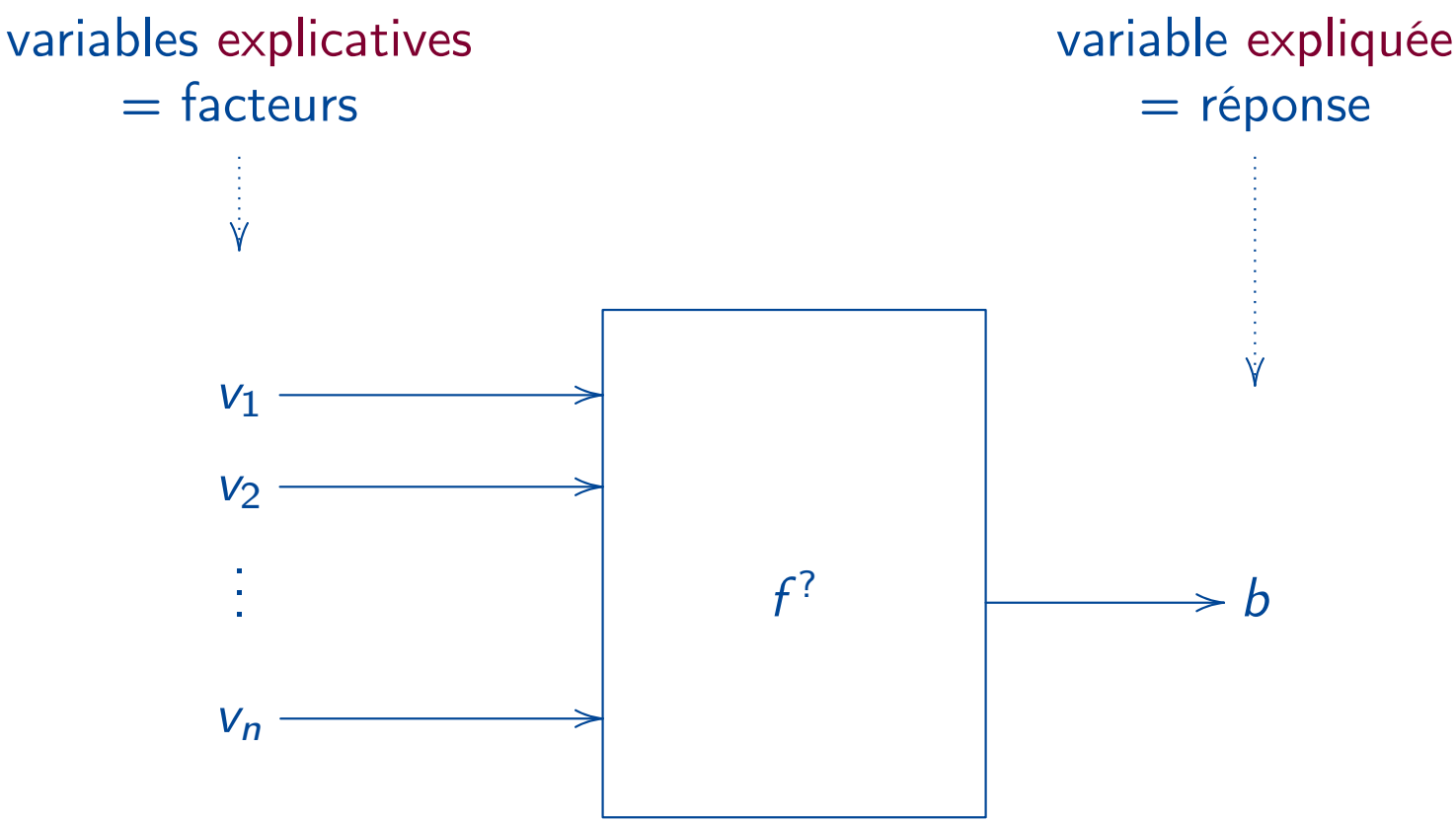
Régression par la méthode des moindres carrés – 2011/2012

Tony Bourdier

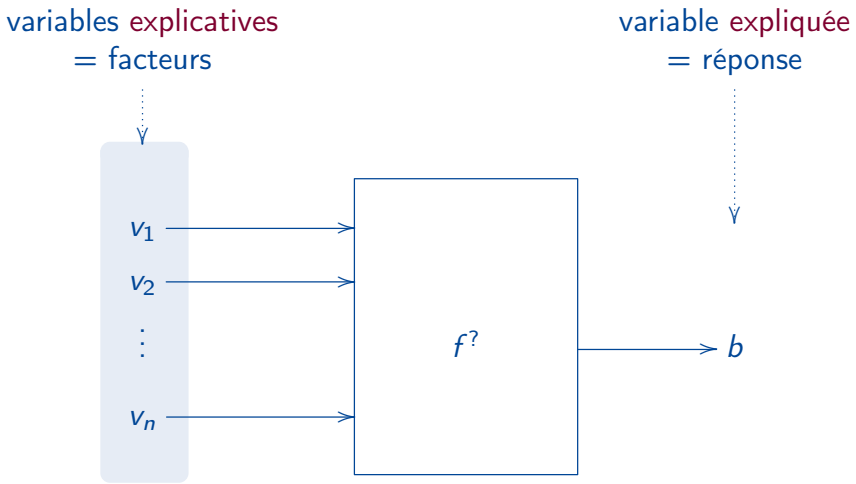
Plan

ECOLE DES MINES DE NANTES

Le principe de la régression



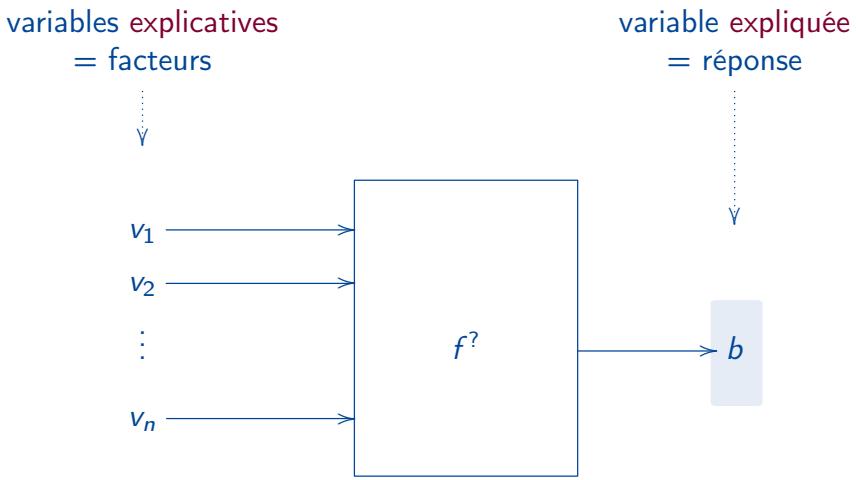
Le principe de la régression



Variables explicatives : données que l'on peut toujours mesurer / obtenir

âge, taille, masse, temps, pression, abscisse, ...

Le principe de la régression



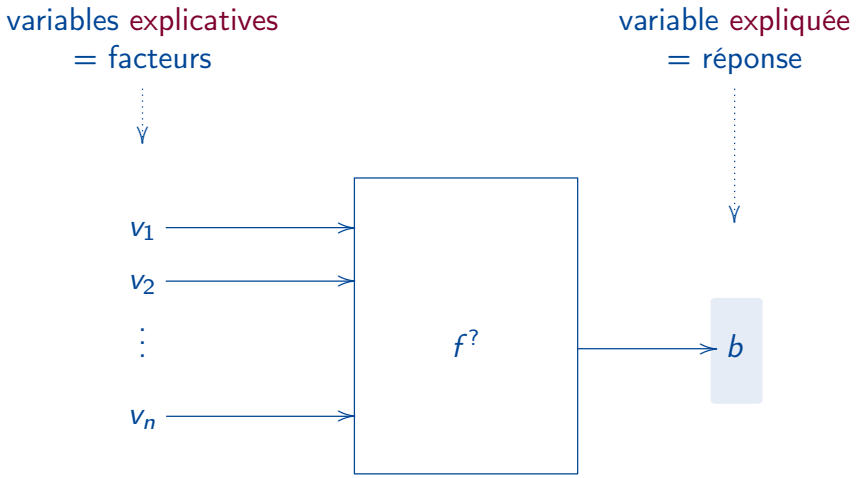
Variable expliquée : donnée que l'on souhaite calculer à partir des variables explicatives parce que . . .

. . . l'on ne peut pas toujours ou difficilement la mesurer

$p = 2.\pi.r$

- rayon r facilement mesurable
- périmètre plus difficilement !

Le principe de la régression



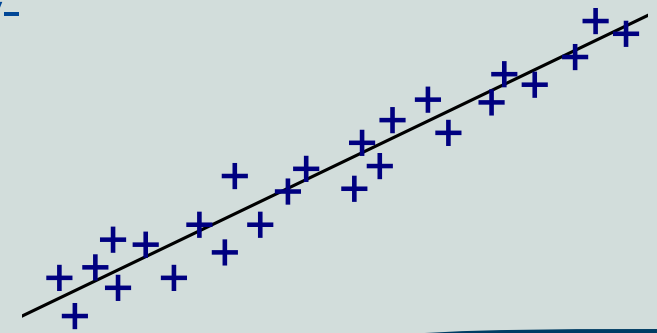
Variable expliquée : donnée que l'on souhaite calculer à partir des variables explicatives parce que . . .

. . . l'on ne peut en obtenir que des mesures imprécises / « bruitées »

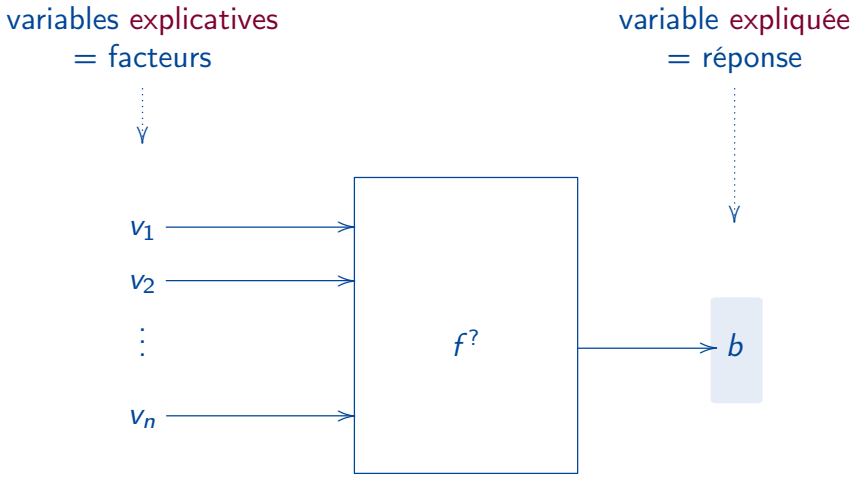
On a obtenu des données par mesure « physique »

- entachées de « bruit »
- on souhaite « corriger » les données

On parle d'« ajustement »



Le principe de la régression



Variable expliquée : donnée que l'on souhaite calculer à partir des variables explicatives parce que . . .

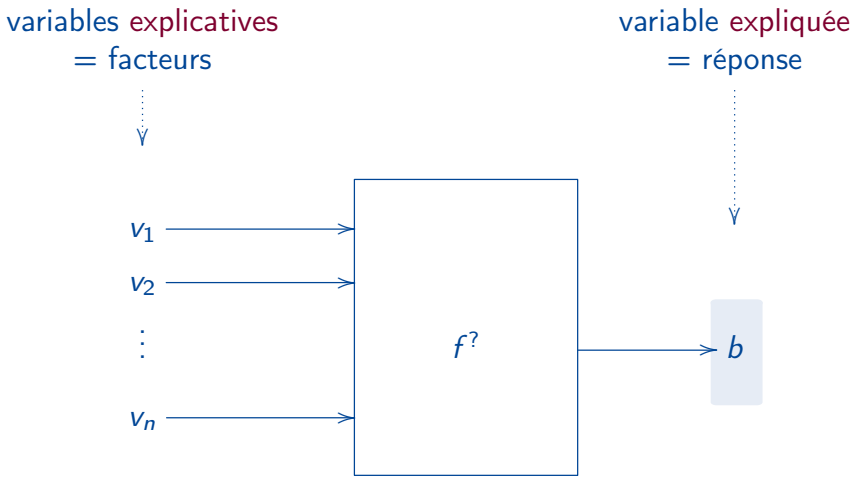
. . . l'on peut la mesurer dans certaines conditions et l'on souhaiterait pouvoir l'estimer dans d'autres conditions

Une donnée évoluant au cours du temps t

- on peut la mesurer si $t \leq current\_time$ (à moins d'être devin)
- on veut connaître sa valeur pour $t' > current\_time$

On parle de « prédiction »

Le principe de la régression



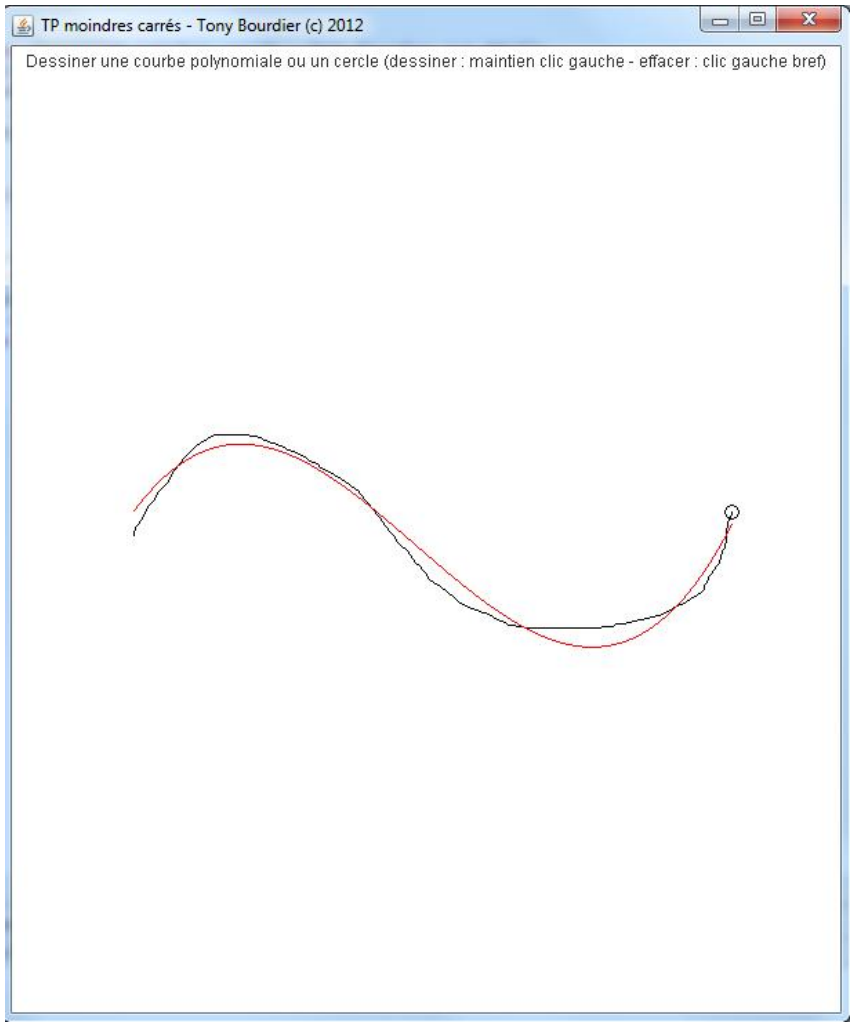
Variable expliquée : donnée que l'on souhaite calculer à partir des variables explicatives parce que . . .

. . . l'on veut établir sa dépendance avec les variables explicatives

Production, travail et capital : Cobb et Douglas affirment que la production P dépend du capital K (valeur des usines, . . .) et du travail fourni T selon la relation suivante :

$$P = f(K, T) = a_1 \times K^{a_2} \times T^{a_3}$$

Applications



- Compression de données

Applications

ECOLE DES MINES DE NANTES

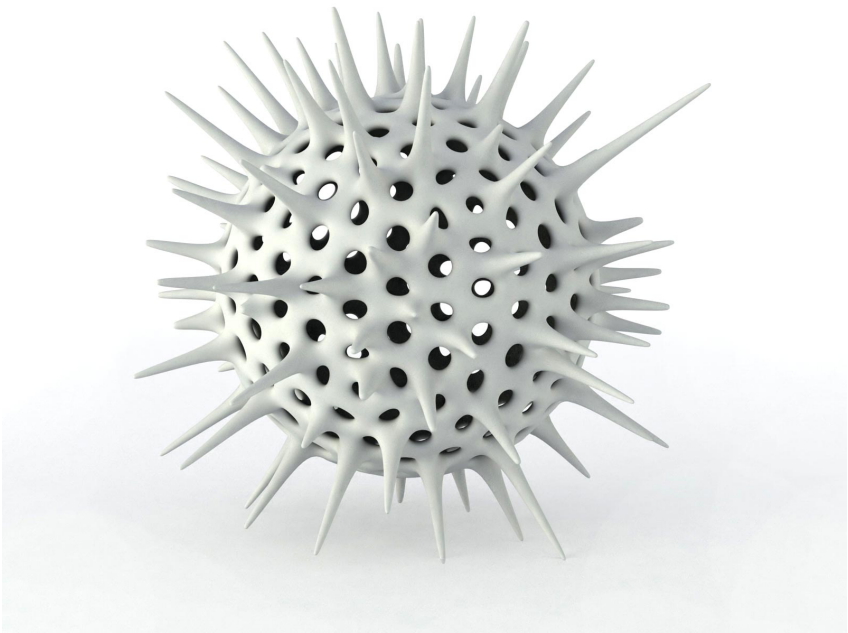
- Compression de données
- Données manquantes dans les B.D.D.

Applications

- Compression de données
- Données manquantes dans les B.D.D.
- Estimation des préférences des utilisateurs

Applications

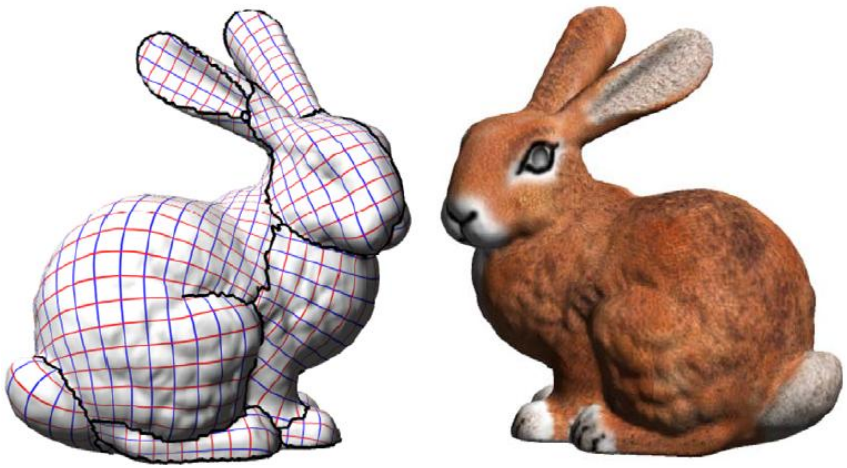
ECOLE DES MINES DE NANTES



- Compression de données
- Données manquantes dans les B.D.D.
- Estimation des préférences des utilisateurs
- Détection d'intrusion / de virus

Applications

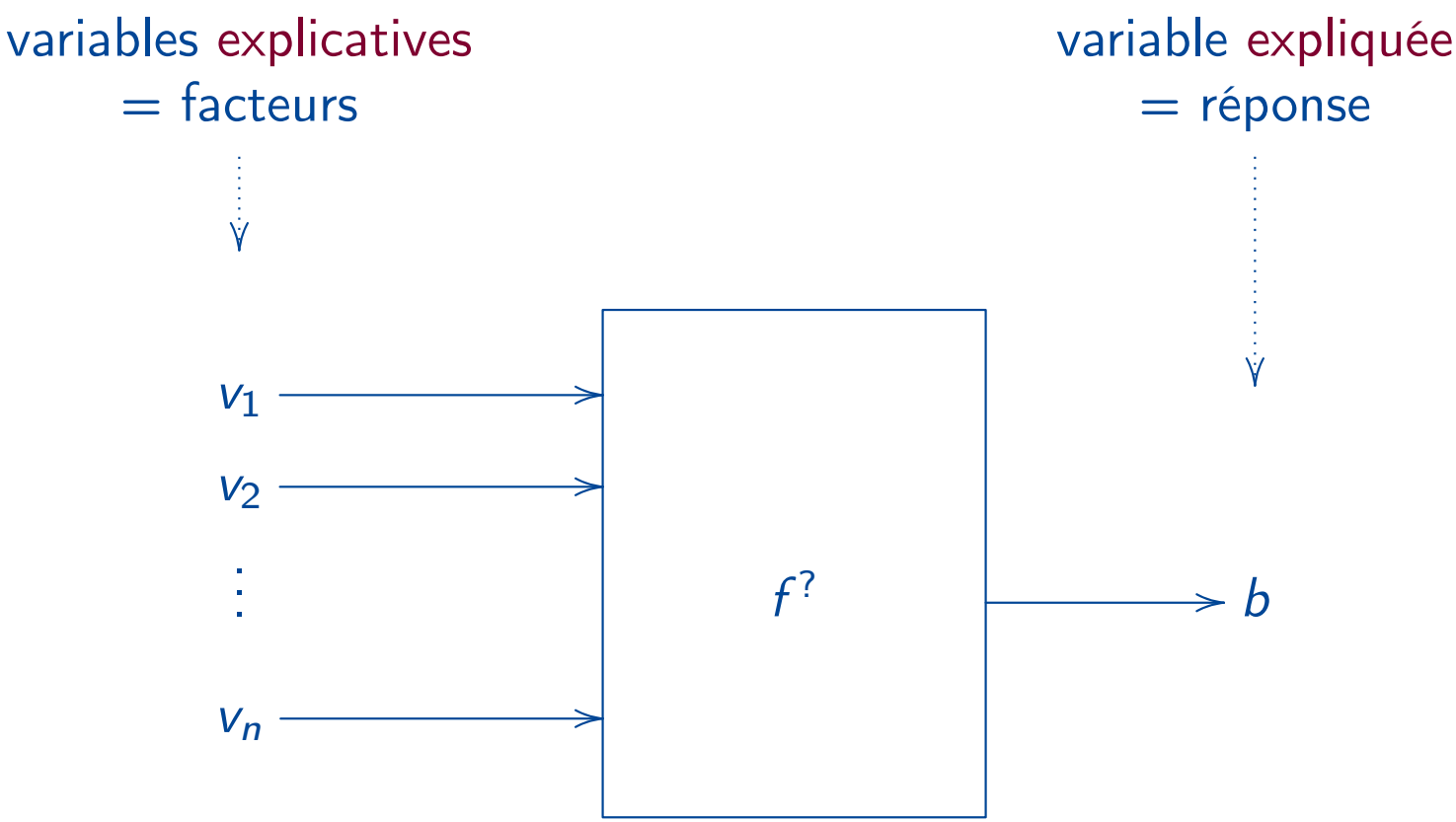
ECOLE DES MINES DE NANTES



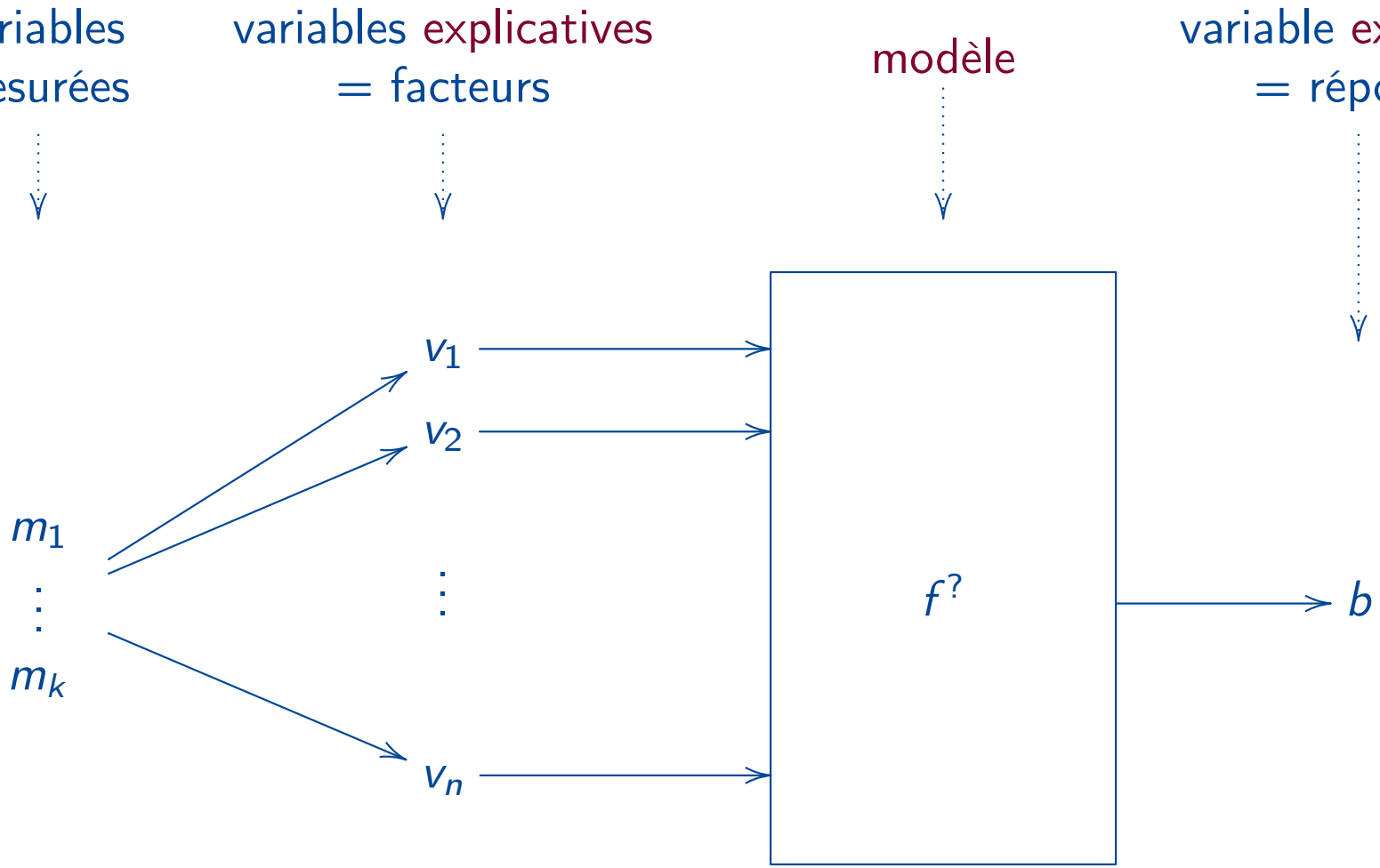
©Bruno Levy (2007)

- Compression de données
- Données manquantes dans les B.D.D.
- Estimation des préférences des utilisateurs
- Détection d'intrusion / de virus
- Plaquage de texture
- ...

Le principe de la régression



Le principe de la régression



Modèle

ECOLE DES MINES DE NANTES

Modèle = fonction caractérisée par :

- des variables explicatives
- des paramètres
- faisant apparaître une variable d'erreur

- $y = f(x) = \sum_{i=1}^{3} \alpha_i.x^i + \varepsilon = \underbrace{\alpha_1.x + \alpha_2.x^2 + \alpha_3.x^3}_{\hat{y}} + \varepsilon$
- $p = f(k,t) = \underbrace{m.k^{\alpha}.t^{\beta}}_{\hat{p}} + \varepsilon$
- $q = f(t) = \underbrace{\lambda.e^{-t/\tau}}_{\hat{q}} + \varepsilon$
- $y = f(t) = \underbrace{a.cos\left(2\pi.t/T\right)}_{\hat{y}} + \varepsilon$

Modèle

ECOLE DES MINES DE NANTES

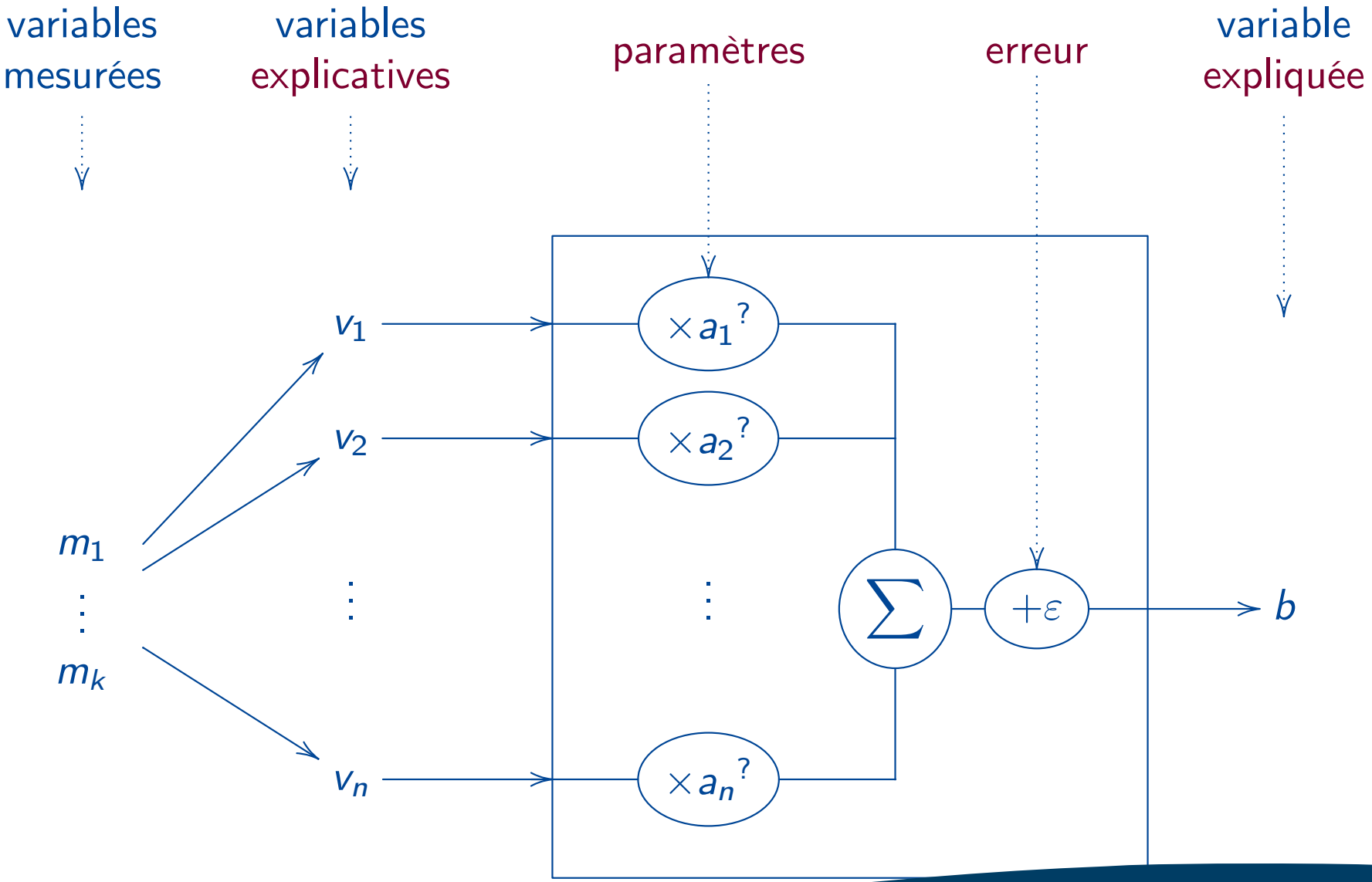
Modèle = fonction caractérisée par :

- des variables explicatives
- des paramètres
- faisant apparaître une variable d'erreur

- $y = f(x) = \sum_{i=1}^{3} \alpha_i.x^i + \varepsilon = \underbrace{\alpha_1.x + \alpha_2.x^2 + \alpha_3.x^3}_{\hat{y}} + \varepsilon$
- $p = f(k,t) = \underbrace{m.k^{\alpha}.t^{\beta}}_{\hat{p}} + \varepsilon$
- $q = f(t) = \underbrace{\lambda.e^{-t/\tau}}_{\hat{q}} + \varepsilon$
- $y = f(t) = \underbrace{a.cos\left(2\pi.t/T\right)}_{\hat{y}} + \varepsilon$

Pour chaque modèle, identifiez les variables mesurées, explicatives et la variable expliquée.

Régression linéaire



Modèle linéaire

ECOLE DES MINES DE NANTES

Modèle linéaire = fonction caractérisée par :

- des variables explicatives
- des paramètres
- faisant apparaître une variable d'erreur
- linéaire en ses paramètres

- $y = f(x) = \sum_{i=0}^{3} \alpha_i . x^i + \varepsilon = \underbrace{\alpha_0 + \alpha_1 . x + \alpha_2 . x^2}_{\hat{y}} + \varepsilon$
- $y = f(t) = \underbrace{a + b.cos\left(\frac{t}{2}\right) + c.sin\left(\frac{t}{2}\right)}_{\hat{y}} + \varepsilon$

Modèle linéaire

ECOLE DES MINES DE NANTES

Modèle linéaire = fonction caractérisée par :

- des variables explicatives
- des paramètres
- faisant apparaître une variable d'erreur
- linéaire en ses paramètres

- $y = f(x) = \sum_{i=0}^{3} \alpha_i . x^i + \varepsilon = \underbrace{\alpha_0 + \alpha_1 . x + \alpha_2 . x^2}_{\hat{y}} + \varepsilon$
- $y = f(t) = \underbrace{a + b.cos\left(\frac{t}{2}\right) + c.sin\left(\frac{t}{2}\right)}_{\hat{y}} + \varepsilon$

Pour chaque modèle, identifiez les variables mesurées, explicatives et la variable expliquée.

Régression linéaire

ECOLE DES MINES DE NANTES

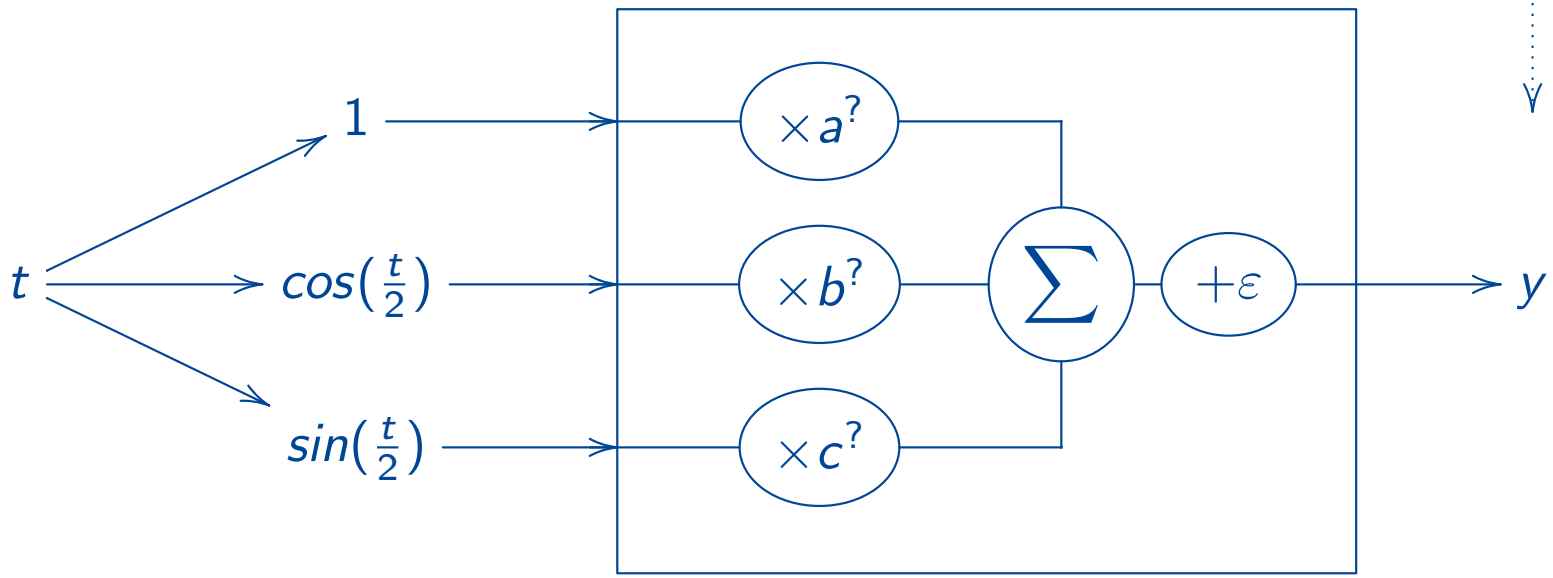
Modèle : $y = f(t) = \underbrace{a + b.cos\left(\frac{t}{2}\right) + c.sin\left(\frac{t}{2}\right)}_{\hat{y}} + \varepsilon$

variable mesurée

variables explicatives

paramètres

variable expliquée



Régression

ECOLE DES MINES DE NANTES

Objectif : On souhaite établir une relation entre une variable (dite expliquée) et d'autres variables (dites explicatives).

Le principe :

1. On propose une forme de relation (le modèle) qui comporte des inconnues (les paramètres)
2. On cherche les valeurs optimales [1] des paramètres :
   - On prend un échantillon (des variables explicatives et expliquée)
   - On estime les paramètres à partir de cet échantillon [2]

[1] selon un certain critère

[2] la méthode d'estimation dépend du critère d'optimalité choisi

Régression

ECOLE DES MINES DE NANTES

1. On propose une forme de relation (le modèle) qui comporte des inconnues (les paramètres)

Exemple : $y = f(t) = a + b.cos\left(\frac{t}{2}\right) + c.sin\left(\frac{t}{2}\right) + \varepsilon$

2. On prend un échantillon (des variables explicatives et expliquée)

	mesures	calcul des var. explicatives			calcul de la var. expliquée
$\text{mesure}_1 \rightarrow$	(t_1, y_1)	1	$cos\,(t_1/2)$	$sin\,(t_1/2)$	y_1
$\text{mesure}_2 \rightarrow$	(t_2, y_2)	1	$cos\,(t_2/2)$	$sin\,(t_2/2)$	y_2
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\text{mesure}_m \rightarrow$	(t_m, y_m)	1	$cos\,(t_m/2)$	$sin\,(t_m/2)$	y_m

Forme matricielle

ECOLE DES MINES DE NANTES

Dans le cas linéaire, on exprime le modèle sous forme matricielle à partir de l'échantillon :

$$\underbrace{\begin{pmatrix} \text{mesure}_1 \rightarrow \\ \text{mesure}_2 \rightarrow \\ \vdots \\ \vdots \\ \text{mesure}_m \rightarrow \end{pmatrix}}_{\hat{b}} \underbrace{\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix}}_{A} \times \underbrace{\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}}_{x} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_m \end{pmatrix}}_{\varepsilon} = \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_m \end{pmatrix}}_{b}$$

- A : matrice des données (variables explicatives)
- x : vecteur des paramètres
- $\hat{b}$: vecteur réponse estimé par le modèle
- ε : vecteur des erreurs
- b : vecteur réponse (variable expliquée)

Forme matricielle

ECOLE DES MINES DE NANTES

Dans le cas linéaire, on exprime le modèle sous forme matricielle à partir de l'échantillon :

$$\underbrace{\begin{pmatrix} \text{mesure}_1 \rightarrow \\ \text{mesure}_2 \rightarrow \\ \vdots \\ \vdots \\ \text{mesure}_m \rightarrow \end{pmatrix}}_{\hat{b}} \underbrace{\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix}}_{A} \times \underbrace{\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}}_{x} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_m \end{pmatrix}}_{\varepsilon} = \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_m \end{pmatrix}}_{b}$$

- A : valeurs données par l'échantillon
- x : valeurs recherchées
- $\hat{b}$: valeurs déduites une fois x déterminé ($\hat{b} = A.x$)
- ε : valeurs déduites une fois x déterminé ($\varepsilon = b - \hat{b}$)
- b : valeurs données par l'échantillon

Forme matricielle

ECOLE DES MINES DE NANTES

Exemple : $y = f(t) = a + b.cos\left(\frac{t}{2}\right) + c.sin\left(\frac{t}{2}\right) + \varepsilon$

$$\begin{pmatrix} 1 & cos\,(t_1/2) & sin\,(t_1/2) \\ 1 & cos\,(t_2/2) & sin\,(t_2/2) \\ \vdots & \vdots & \vdots \\ 1 & cos\,(t_m/2) & sin\,(t_m/2) \end{pmatrix} \times \begin{pmatrix} a \\ b \\ c \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

Forme matricielle

ECOLE DES MINES DE NANTES

Donnez la forme matricielle du problème de regression défini par :

- le modèle : $y = f(x) = a + b.x + c.x^2$
- l'échantillon suivant :

i	$(x_i\ ,\ y_i)$
1	(3 , 8.1)
2	(2 , 4.25)
3	(4 , 14.15)
4	(5 , 22.05)
5	(9 , 73.5)
6	(7 , 43.85)
7	(1 , 1.9)

Critères

ECOLE DES MINES DE NANTES

Quel critère pour estimer les paramètres ?

Critères

ECOLE DES MINES DE NANTES

Quel critère pour estimer les paramètres ?

Solution 1 (Maximum de vraisemblance) :

Hypothèses : $\begin{cases} \varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ donc } b \sim \mathcal{N}\left(\hat{b}, \sigma^2\right) \\ \forall k \neq j, \varepsilon_k \text{ independant de } \varepsilon_j \end{cases}$

Trouver les a_i qui maximisent la probabilité d'avoir obtenu l'échantillon :

$$\max\left(\mathbb{P}(b = b_1) \times \mathbb{P}(b = b_2) \times \ldots \times \mathbb{P}(b = b_m)\right)$$

Abus de notation : $\mathbb{P}$ doit être remplacé par la fonction de densité.

Critères

ECOLE DES MINES DE NANTES

Solution 2 :

Trouver les a_i qui minimisent l'erreur pour l'échantillon considéré :

$$\min \|\varepsilon\|^2 \quad = \quad \min \|b - \hat{b}\|^2 \quad = \quad \min \left(\sum_{k=1}^{m} \left(b_k - \hat{b}_k \right)^2 \right)$$

Critères

ECOLE DES MINES DE NANTES

Solution 2 (méthode des moindres carrés) :

Trouver les a_i qui minimisent l'erreur pour l'échantillon considéré :

$$\min \|\varepsilon\|^2 \quad = \quad \min \|b - \hat{b}\|^2 \quad = \quad \min \left(\sum_{k=1}^{m} \left(b_k - \hat{b}_k \right)^2 \right)$$

Critères

ECOLE DES MINES DE NANTES

Solution 2 (méthode des moindres carrés) :

Trouver les a_i qui minimisent l'erreur pour l'échantillon considéré :

$$\min \|\varepsilon\|^2 \quad = \quad \min \|b - \hat{b}\|^2 \quad = \quad \min \left(\sum_{k=1}^{m} \left(b_k - \hat{b}_k \right)^2 \right)$$

solution 1 = solution 2 ! ! !

Solution

ECOLE DES MINES DE NANTES

On suppose l'existence[1] de la fonction suivante :

$$\begin{array}{rccl} pmc & : & \mathcal{M}_{m,n} \times \mathbb{R}^m & \rightarrow & \mathbb{R}^n \\ & & (\quad A \quad , \; b \quad) & \mapsto & \hat{x} \; t.q. \; \|A.\hat{x} - b\|^2 = \min\limits_{x \in \mathbb{R}^{\boldsymbol{n}}} \|A.x - b\|^2 \end{array}$$

ou (informatiquement) :

```
double[]  pmc (Matrix A, double[] b);
//@ requires A.getRowDimension()    == b.length
//@ ensures  A.getColumnDimension() == \result.length
```

[1] On établira cette fonction lors de la séance suivante.

...

ECOLE DES MINES DE NANTES

Si on dispose de la fonction *pmc*, alors on n'a plus rien à faire ...

...

Si on dispose de la fonction *pmc*, alors on n'a plus rien à faire ...

Lourdement tu te trompes.
Bien modéliser le problème tu dois !

...

Si on dispose de la fonction *pmc*, alors on n'a plus rien à faire ...

Lourdement tu te trompes.
Bien modéliser le problème tu dois !

Modéliser ?

...

ECOLE DES MINES DE NANTES

Si on dispose de la fonction *pmc*, alors on n'a plus rien à faire ...

Lourdement tu te trompes.
Bien modéliser le problème tu dois !

Modéliser ?

Bien identifier la matrice A et le vecteur b il faut.

Ben *a priori* si on a toutes les infos, j'vois pas l'problème.

...

ECOLE DES MINES DE NANTES

Ben *a priori* si on a toutes les infos, j'vois pas l'problème.

Non nécessairement linéaire, le problème d'origine est.

Ben *a priori* si on a toutes les infos, j'vois pas l'problème.

Non nécessairement linéaire, le problème d'origine est.

S'il n'est pas linéaire, alors on n'peut rien faire, si ?

...

ECOLE DES MINES DE NANTES

Ben *a priori* si on a toutes les infos, j'vois pas l'problème.

Non nécessairement linéaire, le problème d'origine est.

S'il n'est pas linéaire, alors on n'peut rien faire, si ?

Changer de variable(s) pour linéariser le problème souvent tu peux.

...

ECOLE DES MINES DE NANTES

OK, et c'est tout ?

ECOLE DES MINES DE NANTES

OK, et c'est tout ?

Non. De dimension supérieure à 1 la variable expliquée peut être.

...

OK, et c'est tout ?

Non. De dimension supérieure à 1 la variable expliquée peut être.

Qu'est-ce qu'on fait dans ce cas là ?

...

OK, et c'est tout ?

Non. De dimension supérieure à 1 la variable expliquée peut être.

Qu'est-ce qu'on fait dans ce cas là ?

Compositionnellement ton problème tu définiras.

Exemples

ECOLE DES MINES DE NANTES

Formulez chacun des problèmes suivants, lorsque cela est possible, sous la forme d'un problème de moindre carrés :

- c.-à.-d. identifier les matrices A et b et
- explicitez le résultat du problème (en invoquant la fonction *pmc*).

Exemples

ECOLE DES MINES DE NANTES

#	Modèle	Mesures
1	$z = a + b.x + c.y + \varepsilon$	$(x_i, y_i, z_i)_{i=1,\dots,m}$
2	$z = a + b.x + c.y + d.x^2 + e.x.y + f.y^2 + \varepsilon$	$(x_i, y_i, z_i)_{i=1,\dots,m}$
3	$p = m.k^{\alpha}.t^{\beta} + \varepsilon$	$(p_i, k_i, t_i)_{i=1,\dots,m}$
4	$z = \lambda.e^{\lambda.t} + \varepsilon$	$(z_i, t_i)_{i=1,\dots,m}$
5	$y = a.cos\,(2\pi.t/T) + \varepsilon$	$(y_i, t_i)_{i=1,\dots,m}$
6	$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a + r.cos(\theta) \\ b + r.sin(\theta) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$	$(x_i, y_i, \theta_i)_{i=1,\dots,m}$
7	$v = \frac{2}{a.t^2} + \varepsilon$	$(v_i, t_i)_{i=1,\dots,m}$
8	$v = \frac{2}{1 + a.t^2} + \varepsilon$	$(v_i, t_i)_{i=1,\dots,m}$
9	$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a + b.t \\ b + c.t^2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$	$(x_i, y_i, t_i)_{i=1,\dots,m}$
10	$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a + b.t \\ c.e^{-b.t} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$	$(x_i, y_i, t_i)_{i=1,\dots,m}$
11	$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a + b.t \\ b.e^{-c.t} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$	$(x_i, y_i, t_i)_{i=1,\dots,m}$

ECOLE DES MINES DE NANTES

Une question ? Un doute ? Une incertitude ?