DM/ML

Régression linéaire

Jamal Atif
jamal.atif@dauphine.fr

Université Paris-Dauphine, M2ID

2015-2016

Copyright

Ce cours est adapté librement des ressources suivantes :

- Livre : Régression avec R, Pierre-André Cornillon et Eric Matzner- Lober, Springer 2010.

Exemple introductif

Un analyste en webmarketing s'intéresse à la relation de causalité entre les frais de publicité et les ventes d'un produit donné. En particulier il cherche à savoir s'il est possible d'expliquer le nombre de ventes par les frais de publicité. Il collecte l'échantillon des données suivant :

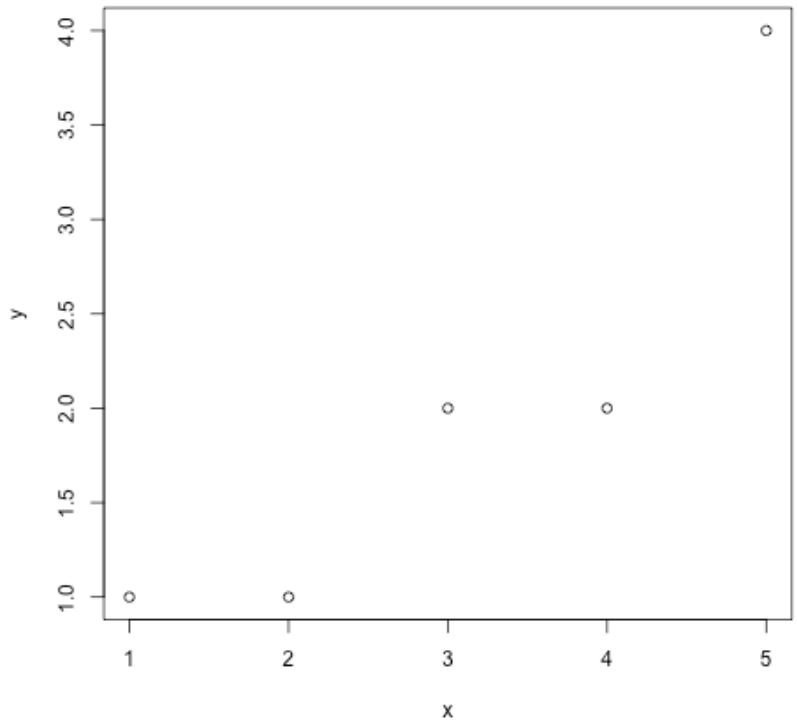
Pub. (€)	Ventes (unité)
1	1
2	1
3	2
4	2
5	4

Il s'agit alors :

- de trouver un *modèle* permettant d'expliquer *Ventes* en fonction de *Pub*,
- de *prédire* les valeurs de *Ventes* par de nouvelles valeurs de *Pub*.

$$\Rightarrow \text{Régression}$$

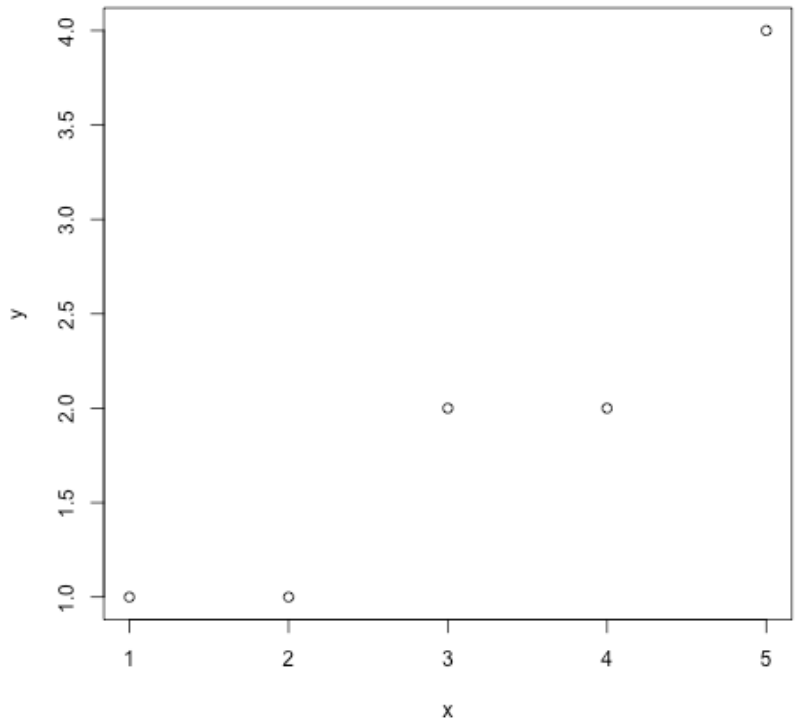
Nuage de points



- Pub (x) est une variable indépendante, dite variable explicative ou encore de régression
- Ventes (y) est dépendante dite réponse.

Première démarche

Examiner le nuage de points



Trouver une relation entre la variable x et la variable y, telle que :

$$y_i \approx f(x_i), i = 1, \cdots, n$$

Modèle

$$y_i \approx f(x_i)$$

$\rightsquigarrow$ Se fixer une famille de fonction $\mathcal{F}$ (ex : fonctions linéaires) et une fonction de coût L telle que :

$\sum_{i=1}^{n} L(y - f(x))$ est minimale pour une fonction $f \in \mathcal{F}$ donnée,

où n représente le nombre de données disponibles (taille de l'échantillon) et L une fonction de coût ou de perte (Loss).

$$\Updownarrow$$

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} L(y - f(x)),\ i = 1, \cdots, n$$

Exemples de L :

- $L(x) = \mid x \mid$,
- $L(x) = x^2$,
- etc.

Modèle de régression linéaire simple

- $\mathcal{F}$ est une famille de fonctions linéaires (affines) de $\mathbb{R}$ dans $\mathbb{R}$.
- On suppose disposer d'un échantillon de n points (x_i, y_i).

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \; \forall i = 1, \cdots, n$$

- ε_i modélisent le bruit et sont supposés aléatoires (les points n'étant jamais parfaitement alignées sur une droite).
- β_1 et β_2 sont les paramètres inconnues du modèles.

Modèle de régression linéaire simple

Hypothèses :

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \ \forall i = 1, \cdots, n$$

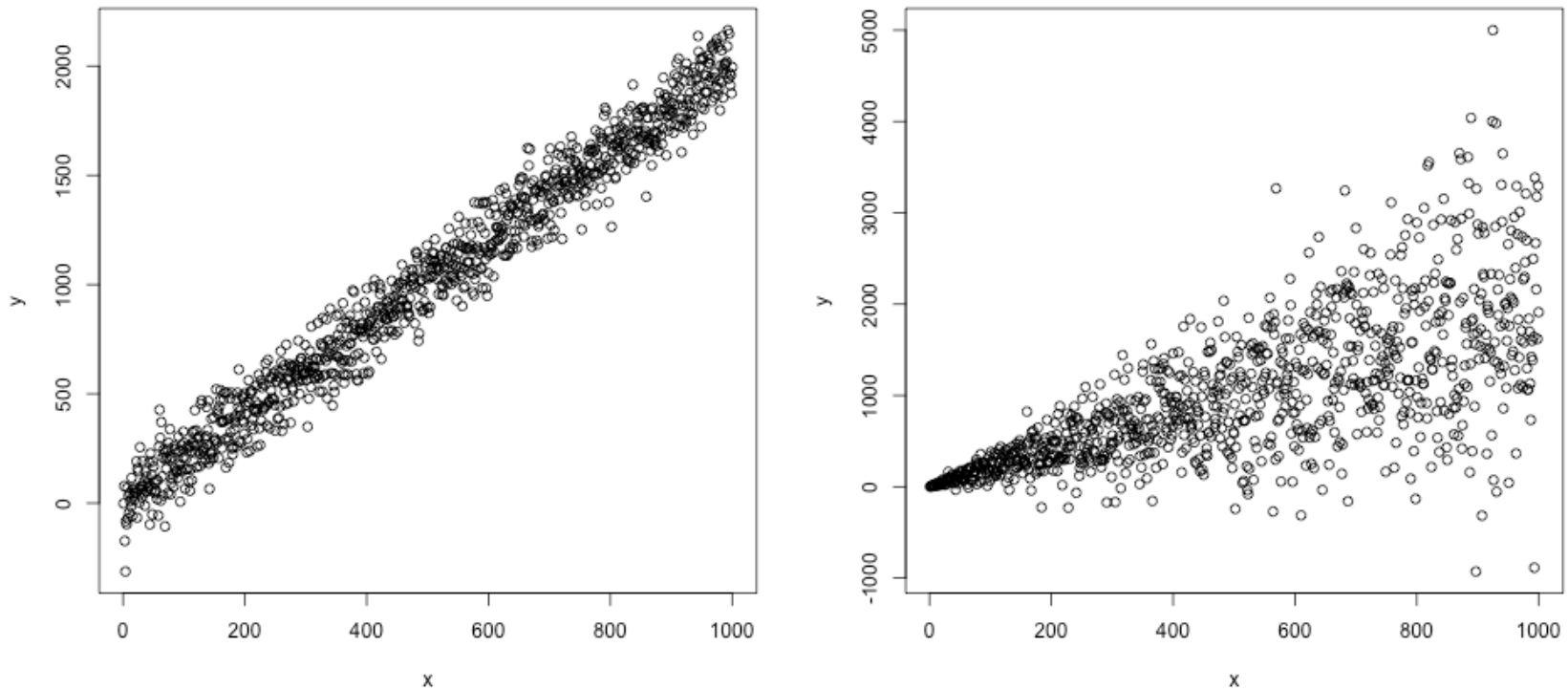
1. le bruit est une variable aléatoire d'espérance nulle et de variance inconnue fixe (homoscédacité) : $\mathbb{E}(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$,
2. ε_i et ε_j sont décorrélés pour tout $i \neq j$: $cov(\varepsilon_i, \varepsilon_j) = 0$
3. ε_i une v.a distribuée selon une loi normale de moyenne nulle et de variance σ^2 : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Modèle de régression linéaire simple

Hypothèses :

Homoscédacité vs Hétéroscédacité

Ex : $y_i = 2x_i + 1 + \varepsilon_i$

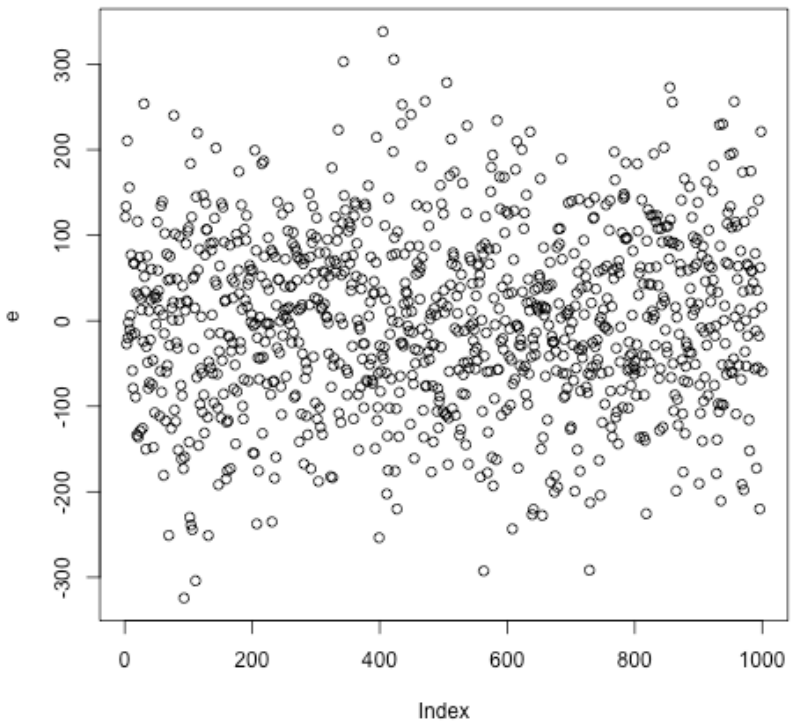


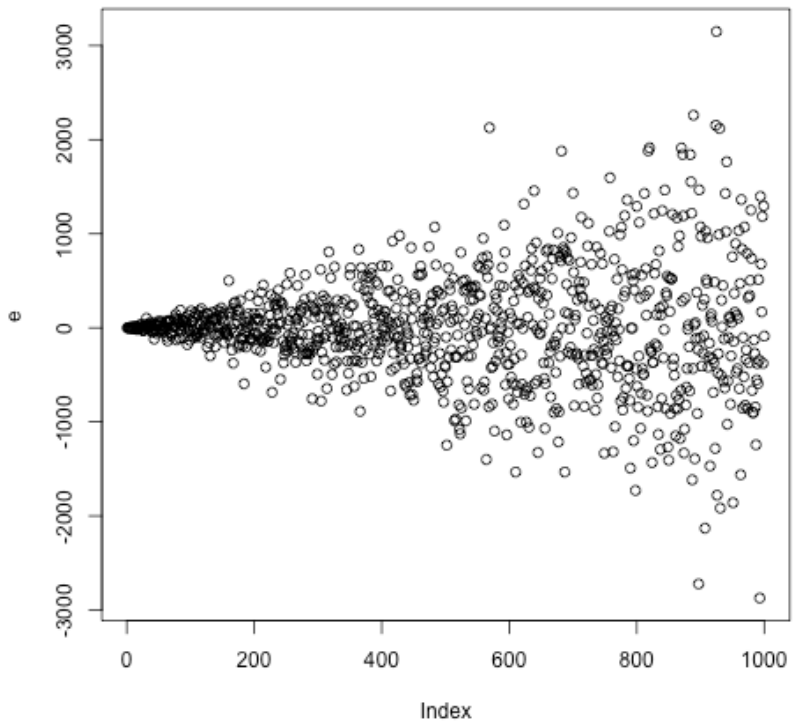
Nuage de points (x_i, y_i).

Modèle de régression linéaire simple

Hypothèses :

Homoscédacité vs Hétéroscédacité





Nuage de points (x_i, ε_i).

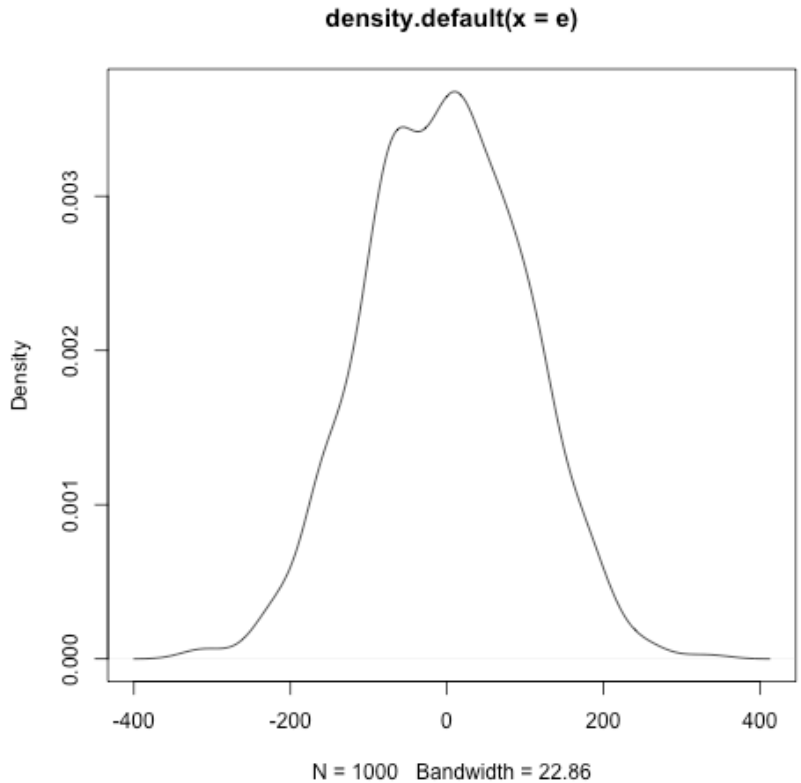
$\varepsilon_i \sim \mathcal{N}(0, x_i)$

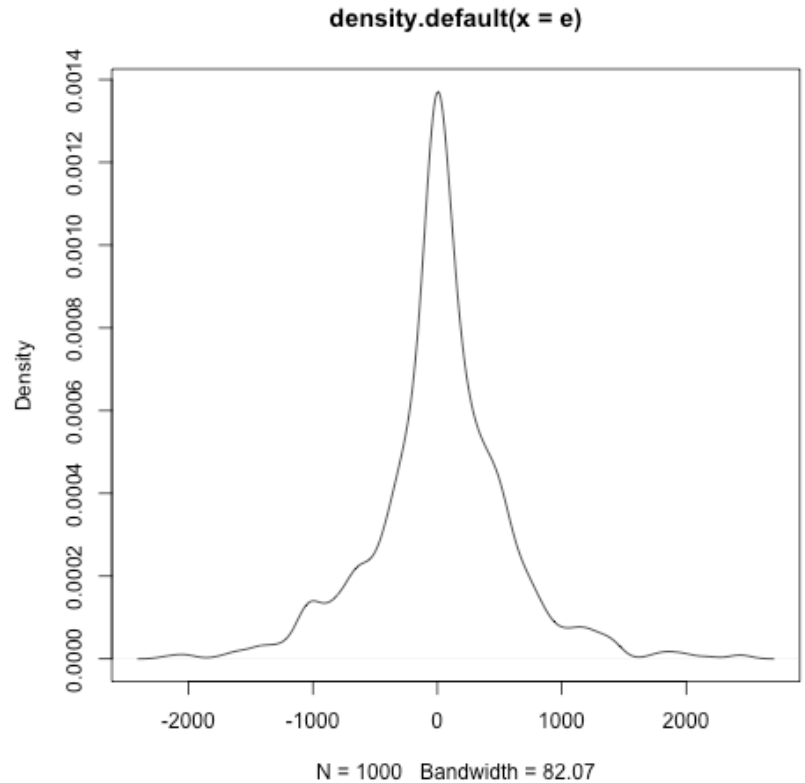
Modèle de régression linéaire simple

Hypothèses :

Homoscédacité vs Hétéroscédacité

Ex : $y_i = 2x_i + 1 + \varepsilon_i$



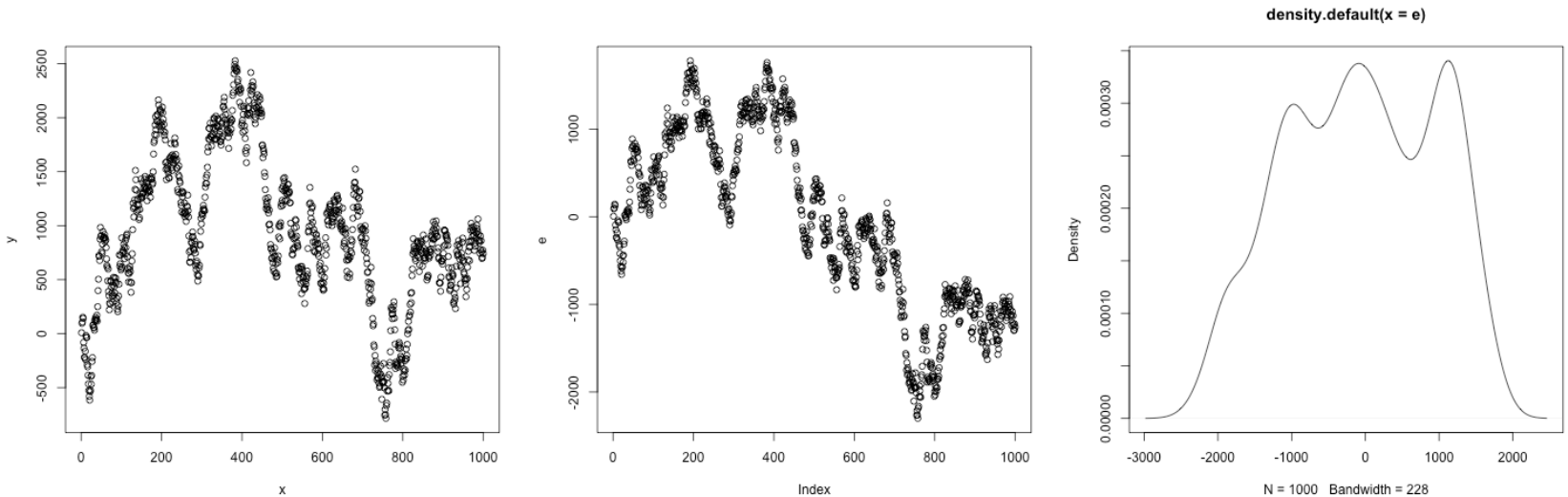


Distribution des erreurs.

Modèle de régression linéaire simple

Hypothèses :

Dépendance vs indépendance du bruit Ex : $y_i = 2x_i + 1 + \varepsilon_i$



A gauche (x_i, y_i), au milieu (x_i, ε_i) et à droite la distribution de ε.

$$\varepsilon_{i+1} \sim \mathcal{N}(\varepsilon_i, 100)$$

Estimateur des Moindres carrés Ordinaires

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i - f(x_i)),\ i = 1, \cdots, n$$

On fixe $L(x) = x^2$ et $f(x) = \beta_1 + \beta_2 x$

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Estimateur des Moindres carrés Ordinaires

Cela revient à minimiser le carré du bruit ε_i pour chaque i :

$$\varepsilon_i = y_i - \beta_1 + \beta_2 x_i = y_i - \tilde{y}_i$$

y_i : le point observé, et $\tilde{y}_i$ le point de la droite *théorique*.

$$\begin{aligned}
(\hat{\beta}_1, \hat{\beta}_2) &= \underset{\beta_1,\beta_2}{\text{argmin}} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2, \\
&= \underset{\beta_1,\beta_2}{\text{argmin}} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2, \\
&= \underset{\beta_1,\beta_2}{\text{argmin}} \sum_{i=1}^{n} \varepsilon_i^2, \\
&= \underset{\beta_1,\beta_2}{\text{argmin}} \, ||\varepsilon||^2
\end{aligned}$$

Illustration

Au tableau !

Calcul des estimateurs de β_1 et β_2

- On notera $S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$
- $S(\beta_1, \beta_2)$ est quadratique donc convexe et différentiable $\Rightarrow$ admet un minimum unique en $(\hat{\beta}_1, \hat{\beta}_2)$.
- On calcule les points pour lesquelles les dérivées partielles de S en β_1 et β_2 s'annulent. On obtient les équations normales suivantes :

$$\begin{cases} \dfrac{\partial S}{\partial \hat{\beta}_1} = -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \\ \dfrac{\partial S}{\partial \hat{\beta}_2} = -2 \displaystyle\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \end{cases}$$

Calcul des estimateurs de β_1 et β_2

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

$$\Rightarrow \hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\Rightarrow \hat{\beta}_1 = \overline{\mathrm{y}} - \hat{\beta}_2 \overline{\mathrm{x}}.$$

Calcul des estimateurs de β_1 et β_2

$$\frac{\partial S}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

$$\Rightarrow \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\Rightarrow \hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \overline{\mathrm{y}}}{\sum x_i^2 - \sum x_i \overline{\mathrm{x}}} = \frac{\sum x_i (y_i - \overline{\mathrm{y}})}{\sum x_i (x_i - \overline{\mathrm{x}})} = \frac{\sum (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{\sum (x_i - \overline{\mathrm{x}})(x_i - \overline{\mathrm{x}})}$$

Quelques remarques

- La relation $\hat{\beta}_1 = \overline{\mathrm{y}} - \hat{\beta}_2 \overline{\mathrm{x}}$ montre que la droite des moindres carrés passe par le centre de gravité des nuages $(\overline{x}, \overline{y})$.
- Les expressions obtenues pour $\hat{\beta}_1$ et $\hat{\beta}_2$ montrent que ces deux estimateurs sont linéaires par rapport au vecteur y.
- L'estimateur $\hat{\beta}_2$ peut s'écrire (exercice de TD) :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \overline{\mathrm{x}})\varepsilon_i}{\sum(x_i - \overline{\mathrm{x}})^2}$$

$\rightarrow$ La variation de $\hat{\beta}_1$ et $\hat{\beta}_2$ vient seulement de ε

Quelques propriétés de $\hat{\beta}_1$ et $\hat{\beta}_2$

Sous les hypothèses 1 et 2 (centrage, décorrélation et homoscédacité) β_1 et β_2 sont des estimateurs sans biais de β_1 et β_2.

Nous savons que :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \overline{\mathrm{x}})\varepsilon_i}{\sum(x_i - \overline{\mathrm{x}})^2}$$

Dans cette expression, seuls les bruits ε_i sont aléatoires d'espérance nulle. Nous avons donc :

$$\mathbb{E}(\hat{\beta}_2) = \beta_2$$

Pour $\hat{\beta}_1$, on part de l'expression :

$$\hat{\beta}_1 = \overline{\mathrm{y}} - \hat{\beta}_2\overline{\mathrm{x}},$$

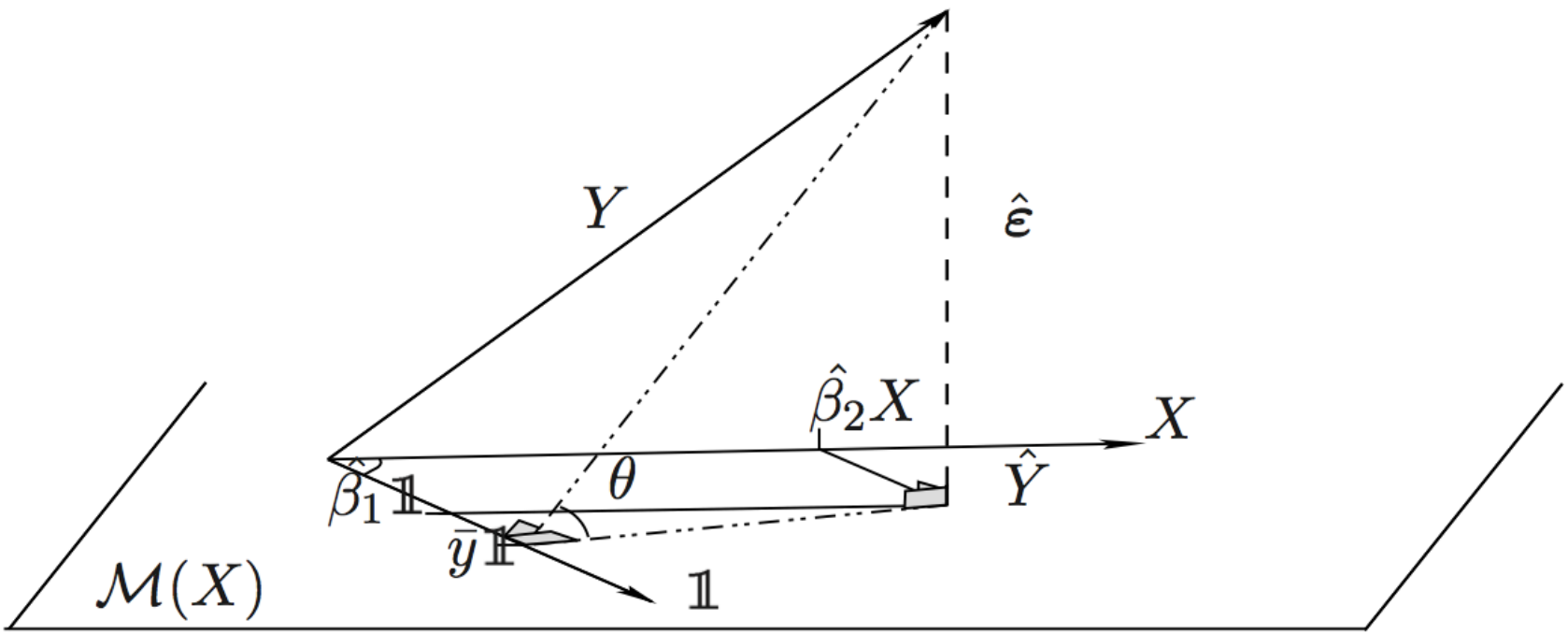
d'où l'on tire :

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}(\overline{\mathrm{y}}) - \overline{\mathrm{x}}\mathbb{E}(\hat{\beta}_2) = \beta_1 + \overline{\mathrm{x}}\beta_2 - \overline{\mathrm{x}}\beta_2 = \beta_1$$

Interprétation géométrique

Le problème de régression peut prendre la forme matricielle : $\mathrm{y} = A\mathrm{b} + \varepsilon$, avec

$$A = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \mathrm{b} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$



Cas d'erreurs gaussiennes

Hypothèses supplémentaires sur le modèle :

1. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
2. ε_i sont mutuellement indépendants

Il s'en suit :

$$\forall i \in \{1, \cdots, n\} \;\; y_i \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)$$

Estimateurs du maximum de vraisemblance

La vraisemblance vaut

$$\begin{aligned}\mathcal{L}(\mathrm{y};\beta_1,\beta_2,\sigma^2) &= \prod_{i=1}^{n}\mathcal{N}(\beta_1+\beta_2 x_i,\sigma^2)\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1-\beta_2 x_i)^2\right]\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}S(\beta_1,\beta_2)\right]\end{aligned}$$

$\Rightarrow$ Trouver $(\hat{\beta}_1,\hat{\beta}_2,\hat{\sigma}^2)$ qui maximisent la vraisemblance.
$\Rightarrow$ Pour simplifier le calcul on prend la log-vraisemblance : $\log(\mathcal{L})$.

$$\log\mathcal{L}(\mathrm{y};\beta_1,\beta_2,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}S(\beta_1,\beta_2)$$

Estimateurs du maximum de vraisemblance

$\hat{\beta}_{1mv}, \hat{\beta}_{2mv}$

$$\log \mathcal{L}(\mathrm{y}; \beta_1, \beta_2, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} S(\beta_1, \beta_2)$$

- Maximiser par rapport à (β_1, β_2) revient à minimiser $-S(\beta_1, \beta_2)$.

$\Rightarrow$ Les estimateurs du maximum de vraisemblance de β_1 et β_2 sont égaux aux estimateurs des moindres carrés.

$$\hat{\beta}_{1mv} = \hat{\beta}_1, \quad \hat{\beta}_{2mv} = \hat{\beta}_2$$

Estimateurs du maximum de vraisemblance

$\hat{\sigma}^2_{mv}$

Il reste à maximiser $\log \mathcal{L}(\mathrm{y}; \hat{\beta}_1, \hat{\beta}_2, \sigma^2)$ par rapport à σ^2.

$$\frac{\partial \mathcal{L}(\mathrm{y}; \hat{\beta}_1, \hat{\beta}_2, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} S(\hat{\beta}_1, \hat{\beta}_2)$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

$$\frac{\partial \mathcal{L}(\mathrm{y}; \hat{\beta}_1, \hat{\beta}_2, \sigma^2)}{\partial \hat{\sigma}^2_{mv}} = 0 \Rightarrow \hat{\sigma}^2_{mv} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$$

- L'estimateur du maximum de vraisemblance de σ^2 est différent de l'estimateur MCO $\hat{\sigma}^2$.
- L'estimateur du maximum de vraisemblance de σ^2 est donc biaisé. En effet $\mathbb{E}(\hat{\sigma}^2_{mv}) = \frac{1}{n} \sum \mathbb{E}(\hat{\varepsilon}_i^2) = \frac{n-2}{n} \sigma^2$

$\Rightarrow$ Ce biais est d'autant plus négligeable que le nombre d'observations est grand.

Régression linéaire multiple

Exemple

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
V	9.25	-6.15	-4.92	11.57	-6.23	2.76	10.15	13.5	21.27	13.79
N_{12}	5	7	6	5	2	7	4	6	1	4
O_3	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

TABLE : 10 données journalières de température, vent, nébulosité et ozone.

Modèle :

$$O_{3i} \approx f(T_i, V_i, N_i)$$

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{i}^{n} L(y_i - f(\mathrm{x}_i))$$

$$\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}, f(\mathrm{x}) = f(\mathrm{x}^1, \mathrm{x}^2, \cdots, \mathrm{x}^d) = \sum_{j=1}^{d} \beta_j \mathrm{x}^j\}$$

Modélisation, suite

$$
\begin{aligned}
y_i &= \beta_i \mathrm{x}_i^1 + \beta_2 \mathrm{x}_i^2 + \cdots + \beta_d \mathrm{x}_i^d + \epsilon_i, i = 1, \cdots, n \\
&= \boldsymbol{\beta}^T \mathrm{x}_i + \epsilon_i
\end{aligned}
$$

Ecriture matricielle :

$$
\mathrm{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}
$$

$$
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{x}_1^1 & \cdots & \mathrm{x}_1^d \\ 1 & \mathrm{x}_2^1 & \cdots & \mathrm{x}_2^d \\ \vdots & \cdots & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots \\ 1 & \mathrm{x}_n^1 & \cdots & \mathrm{x}_n^d \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}
$$

Solution MCO

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T \mathrm{x}_i)^2 \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} ||\mathrm{y} - X\boldsymbol{\beta}||_2^2\end{aligned}$$

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathrm{y}$$

Matrice de Projection :

$$P_X = (X^T X)^{-1} X^T$$

Estimateurs du Maximum de vraisemblance

Modèle

$$y_i = \boldsymbol{\beta}^T \mathrm{x}_i + \epsilon_i \sim \mathcal{N}(\boldsymbol{\beta}^T \mathrm{x}_i, \sigma^2)$$

Vraisemblance :

$$\mathcal{L}(\mathrm{y}, \beta, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^T \mathrm{x}_i)^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}||\mathrm{y} - X\beta||_2^2\right]$$

Log-vraisemblance

$$\mathcal{LL}(\mathrm{y}, \beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - -\frac{1}{2\sigma^2}||\mathrm{y} - X\beta||_2^2$$

$$\Rightarrow \hat{\beta}_{mv} = \hat{\beta} = (X^T X)^{-1} X^T \mathrm{y}$$

Régression non-linéaire

Modèle :

$$f_\beta(\mathrm{x}) = \sum_{i=0}^{M-1} \beta_j \phi_j(\mathrm{x})$$

$$= (\beta_0, \cdots, \beta_{M-1}) \begin{pmatrix} \phi_0(\mathrm{x}) \\ \vdots \\ \phi_{M-1}(\mathrm{x}) \end{pmatrix} = \beta^T \boldsymbol{\phi}(\mathrm{x})$$

oİ les $\phi_j(\mathrm{x})$ sont dites **fonctions de base**

Exemples de fonctions de base

Cas linéaire

$$\phi_j(\mathrm{x}) = \mathrm{x}_j, M = d + 1$$

Régression polynomiale

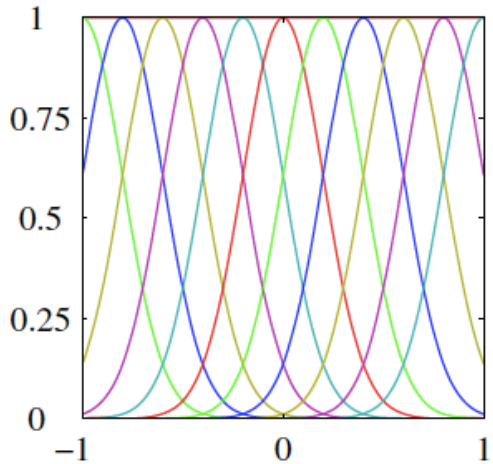
Cas unidimensionnel

$$\phi_j(\mathrm{x}) = \mathrm{x}^j$$

Régression noyau

Fonctions de base gaussiennes

$$\phi_j(\mathrm{x}) = \exp\left(-\frac{(\mathrm{x}-\mu_j)^2}{2\sigma^2}\right)$$



Estimateurs du maximum de vraisemblance

Modèle

$$y_i = \boldsymbol{\beta}^T \boldsymbol{\phi}(\mathrm{x}_i) + \epsilon_i \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{\phi}(\mathrm{x}_i), \sigma^2)$$

Vraisemblance (iid) :

$$\mathcal{L}(\mathrm{y}, \beta, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^T \boldsymbol{\phi}(\mathrm{x}_i))^2\right]$$

Log-vraisemblance

$$\mathcal{LL}(\mathrm{y}, \beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^T \boldsymbol{\phi}(\mathrm{x}_i))^2$$

Posons

$$S(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta^T \boldsymbol{\phi}(\mathrm{x}_i))^2$$

Estimateurs du maximum de vraisemblance

$$\nabla S(\beta) = -\sum_{i=1}^{n}(y_i - \beta^T \boldsymbol{\phi}(\mathrm{x}_i))\boldsymbol{\phi}(\mathrm{x}_i)^T$$

D'oİ

$$\sum_{i=1}^{n} y_i \boldsymbol{\phi}(\mathrm{x}_i)^T - \hat{\beta}^T \left(\sum_{i=1}^{n} \boldsymbol{\phi}(\mathrm{x}_i)\boldsymbol{\phi}(\mathrm{x}_i)^T \right) = 0$$

$$\Longrightarrow \ \hat{\beta}_{mv} = (\Phi^T \Phi)^{-1} \Phi^T \mathrm{y}$$

Avec Φ la matrice dite du *plan dexpérience ou* design matrix

$$\Phi = \begin{pmatrix} \phi_0(\mathrm{x}_1) & \phi_1(\mathrm{x}_1) & \cdots & \phi_{M-1}(\mathrm{x}_1) \\ \phi_0(\mathrm{x}_2) & \phi_1(\mathrm{x}_2) & \cdots & \phi_{M-1}(\mathrm{x}_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(\mathrm{x}_n) & \phi_1(\mathrm{x}_n) & \cdots & \phi_{M-1}(\mathrm{x}_n) \end{pmatrix}$$

Régression de Ridge

Régularisation L_2

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta^T \boldsymbol{\phi}(\mathrm{x}_i))^2 + \frac{\lambda}{2}||\beta||_2^2$$

Solution

$$\hat{\beta} = \left(\lambda I + \Phi^T \Phi\right)^{-1} \Phi^T \mathrm{y}$$

Régression Lasso

Régularisation L_1

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta^T \boldsymbol{\phi}(\mathrm{x}_i))^2 + \frac{\lambda}{2}||\beta||_1$$