

Cours

Modèle de régression

K. BELATTAR,

Département Informatique - Université d'Alger 1

Régression linéaire

Régression linéaire : le modèle le plus simple de la régression en apprentissage supervisé.

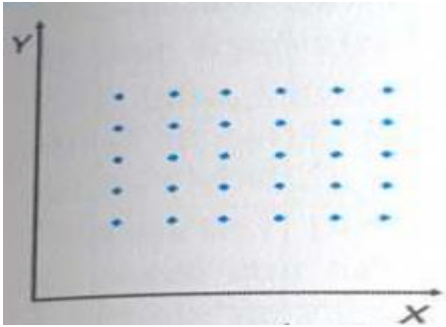
Chercher une **relation linéaire** entre la variable explicative (indépendante) X et la variable expliquée (dépendante) Y .

L'idée est de pouvoir ensuite faire des prévisions sur Y lorsque X est mesurée.

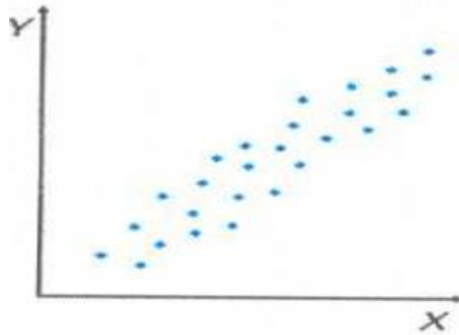
Régression linéaire

- Lorsqu'il existe une **seule variable** d'entrée (x), ➔ la méthode est appelée **régression linéaire simple**.
- Lorsqu'il y a **plusieurs variables d'entrée**, ➔ la méthode est appelée **régression linéaire multiple**.
- Pour déterminer la **nature de la relation**, on utilise **le nuage de point**.

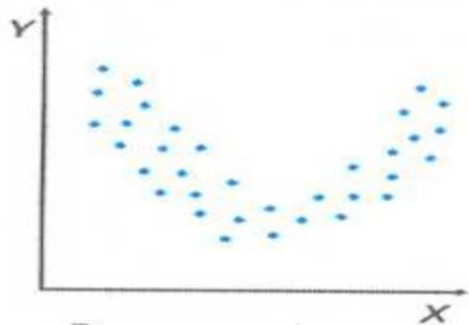
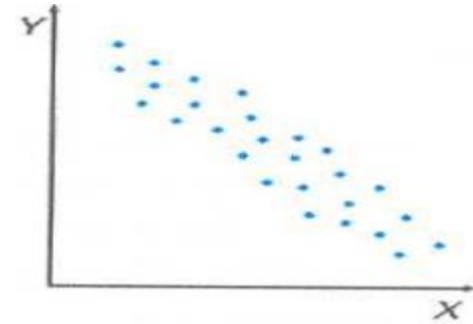
Formes des nuages de points



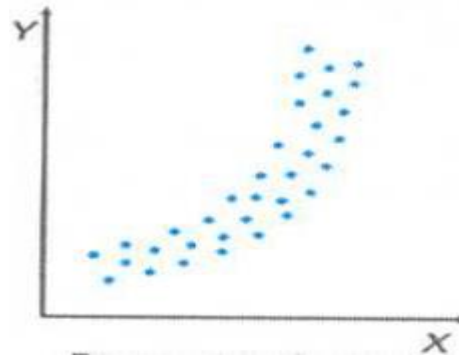
Indépendance



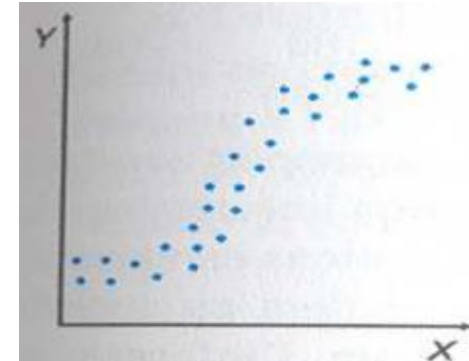
Ajustement linéaire (affine)



Ajustement parabolique
(polynôme de degrés 2)



Ajustement exponentiel



Ajustement logistique

Choix d'ajustement du modèle

- Ajustement linéaire: $r(x; y)$
- Ajustement exponentiel: $r(x; \ln y)$
- Ajustement puissance: $r(\ln x; \ln y)$
- Ajustement logistique (logarithmique): $r(\ln x; y)$

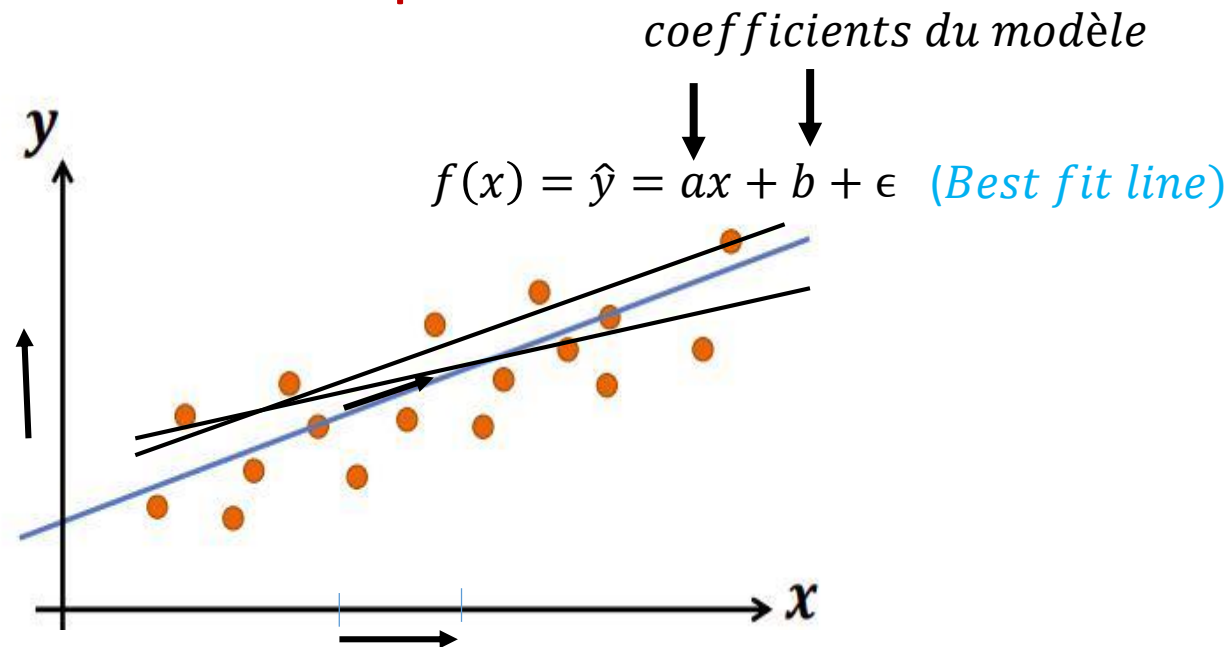
$$r(X; Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad -1 \leq r_{X,Y} \leq 1$$

→ Retenir l'ajustement correspondant au plus grand coefficient de corrélation r

Régression linéaire

Modèle de régression

(1) Régression linéaire simple



Exemple: relation entre la surface (x) et le prix (y) des maisons

Régression linéaire

(2) Régression linéaire multiple

$$\hat{y}_i = \theta_0 + \sum_{j=1}^d x_{ij}\theta_j = \theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + \dots + x_{id}\theta_d + \epsilon$$

θ_0 : *intercept (bais)*

$x = (x_1, x_2, \dots, x_d)$ avec $x_0 = 1$

θ_j : *Coefficients du modèle (poids) relatifs à x_1, x_2, \dots, x_d*

Exemple: Relation entre surface, nombre de pièces, localisation et prix des maisons.

Régression linéaire

Chercher à estimer la qualité de l'ajustement de la fonction f aux données « Best fit line » consiste à:

- Trouver les **meilleurs paramètres (biais et coefficients)** du modèle de la régression
- Estimer les valeurs de θ_j telles que **la droite passe au mieux par tous les exemples d'apprentissage**
- **Les valeurs prédites $f(x)$ sont très proches des valeurs réelles (y) dans tous les exemples d'apprentissage.**

Estimation des paramètres du modèle linéaire

(1) Régression linéaire simple

$$\hat{y} = ax + b$$

$$a = \frac{cov(x, y)}{var(x)}, b = \bar{Y} - a \bar{x}$$

$$cov(x, y) = \overline{xy} - \bar{x} \bar{y}$$

$$var(x) = \overline{x^2} - (\bar{x})^2$$

$$\text{Résidu } e_i = y_i - \hat{y}$$

Si résidu est anormalement élevé alors l'individu i est atypique.

Estimation des paramètres du modèle linéaire

La qualité de l'ajustement (*coefficient de détermination* R^2)

$$\text{Variance résiduelle (SCR)} = \sum_{i=1}^n e_i^2$$

$$\text{Variance expliquée par le modèle (SCE)} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Variation totale des observations (SCT) = SCR + SCE ou

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$R^2 = \frac{SCE}{SCT}$$

Statistique $\sigma^2 = \frac{SCR}{n-2}$ / n est le nombre de individus

Cette statistique est utilisée pour calculer l'erreur type, construire les intervalles de confiance et effectuer des tests d'hypothèse dans le cadre de l'analyse de régression.

Estimation des paramètres du modèle linéaire

(2) Régression linéaire multiple

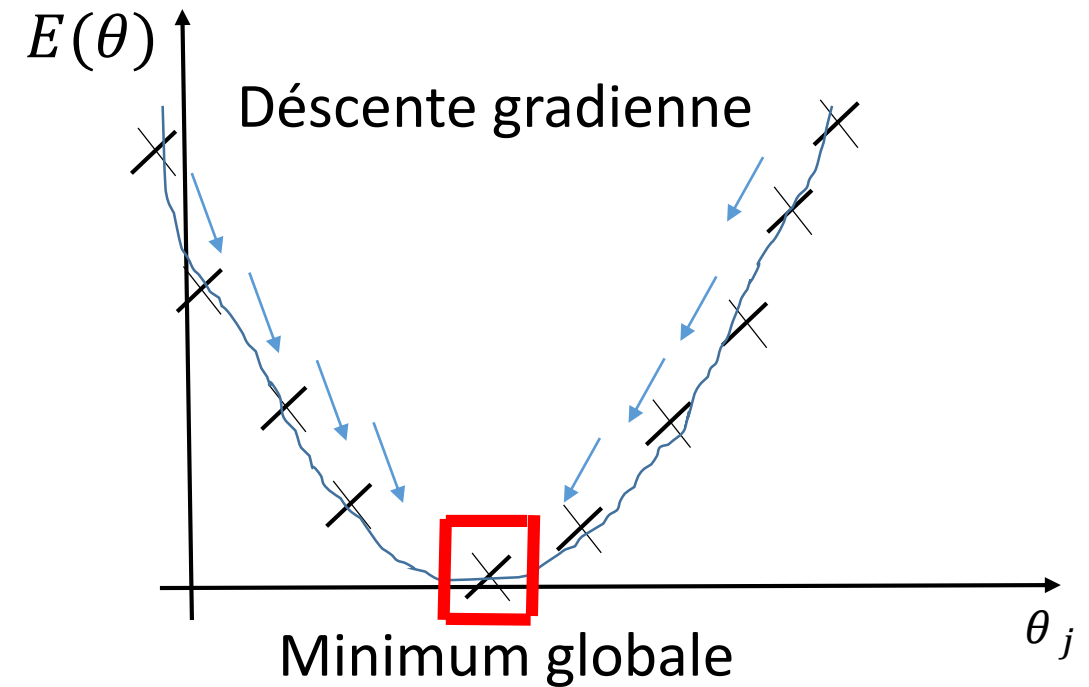
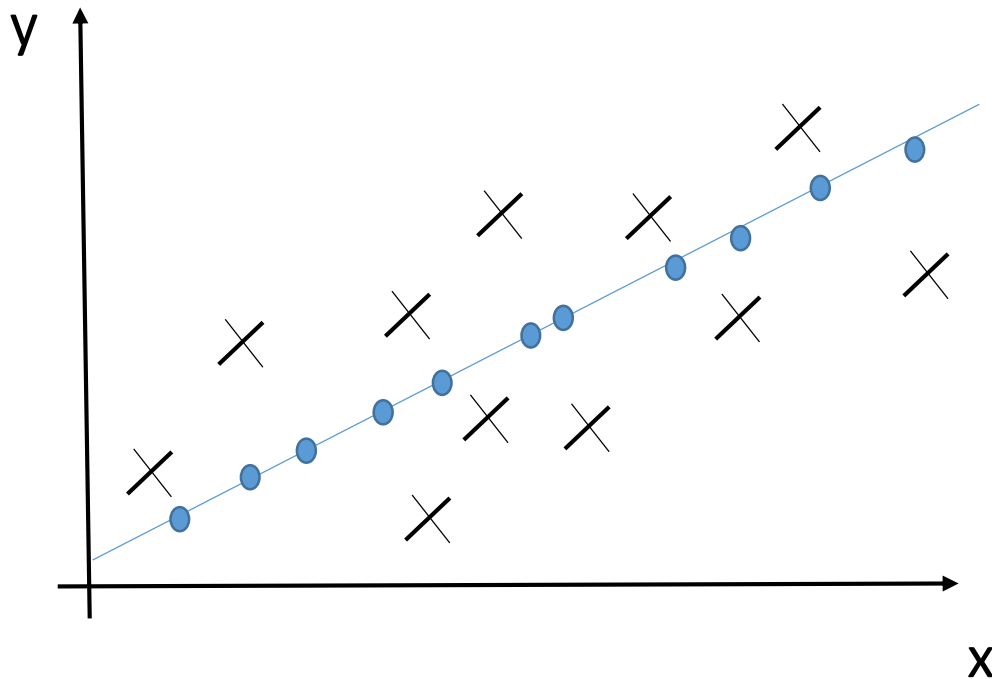
- Utiliser la méthode des moindres carrés ordinaires qui consiste à minimiser la somme des résidus au carré (fonction de coût E) sur l'ensemble des exemples N .

$$E(\tilde{\theta}) = \frac{1}{2N} \sum_{i=1}^N (y_i - \widehat{y}_i)^2$$

- La fonction d'erreur permet de mesurer combien la prédiction du modèle est loin des «vraies» valeurs
- Choisir $\tilde{\theta}$ qui **minimise la fonction de coût**

Estimation des paramètres du modèle linéaire

(2) Régression linéaire multiple



Estimation des paramètres du modèle linéaire

(2) Régression linéaire multiple

Notation matricielle

$$\tilde{\theta} = (X^T X)^{-1} X^T y$$

L'inverse de $(X^T X)$ existe si les colonnes de **X** sont linéairement indépendantes.

Estimation des paramètres du modèle linéaire

Notation matricielle

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \theta_0 + \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

$$\tilde{\theta} = (X^T X)^{-1} X^T y$$

Estimation des paramètres du modèle linéaire

Prediction



Input features
(one sample has d features)

Model parameters
(d weights and 1 bias)

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

Estimation des paramètres du modèle linéaire

Temperature of the building



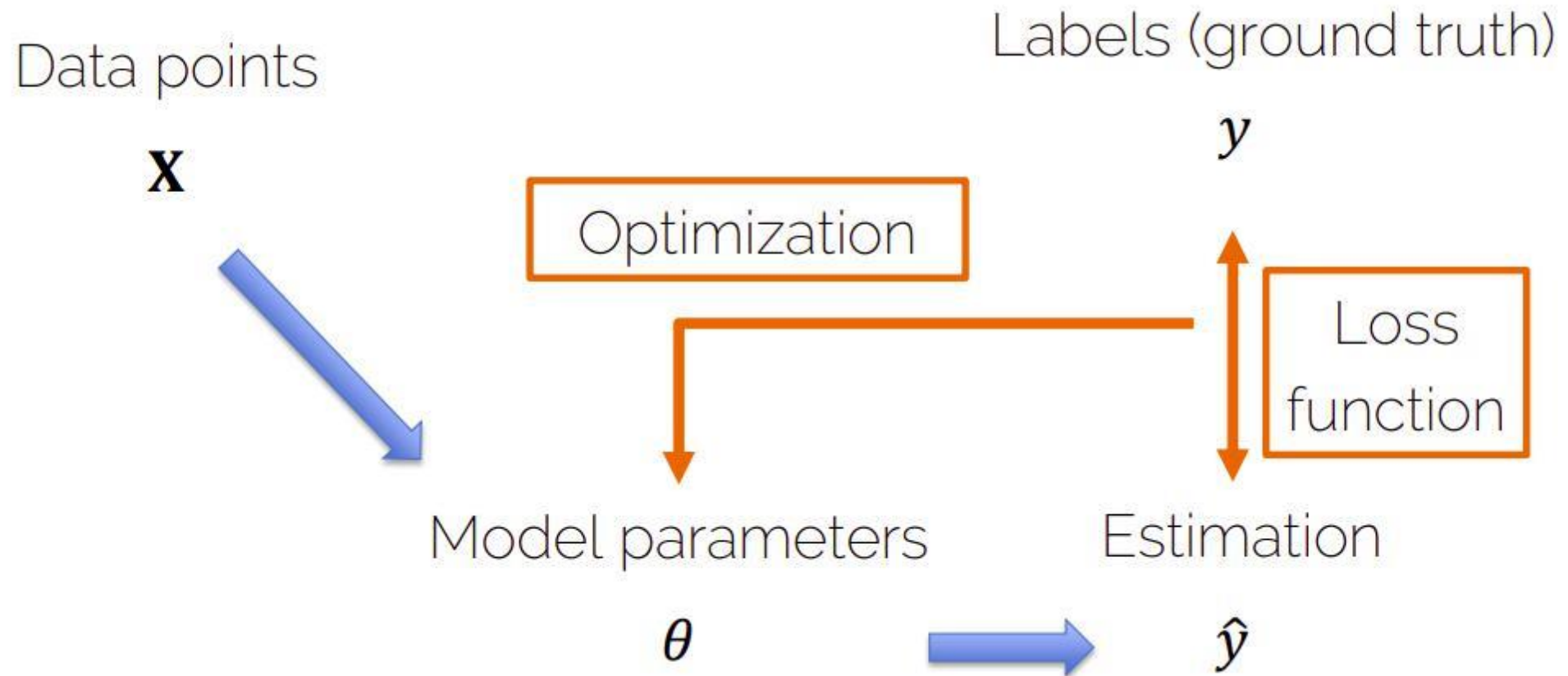
Outside temperature
Humidity
Number people
Sun exposure (%)

MODEL

How do we obtain the model?

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} 1 & 25 & 50 & 2 & 50 \\ 1 & -10 & 50 & 0 & 10 \end{bmatrix} \cdot \begin{bmatrix} 0.2 \\ 0.64 \\ 0 \\ 1 \\ 0.14 \end{bmatrix}$$

Estimation des paramètres du modèle linéaire



Estimation des paramètres du modèle linéaire

Exemple:

x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43



$$X = \begin{bmatrix} 0.86 & 1 \\ 0.09 & 1 \\ -0.85 & 1 \\ 0.87 & 1 \\ -0.44 & 1 \\ -0.43 & 1 \\ -1.10 & 1 \\ 0.40 & 1 \\ -0.96 & 1 \\ 0.17 & 1 \end{bmatrix} \quad Y = \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

$x_1 \quad x_0$

Estimation des paramètres du modèle linéaire

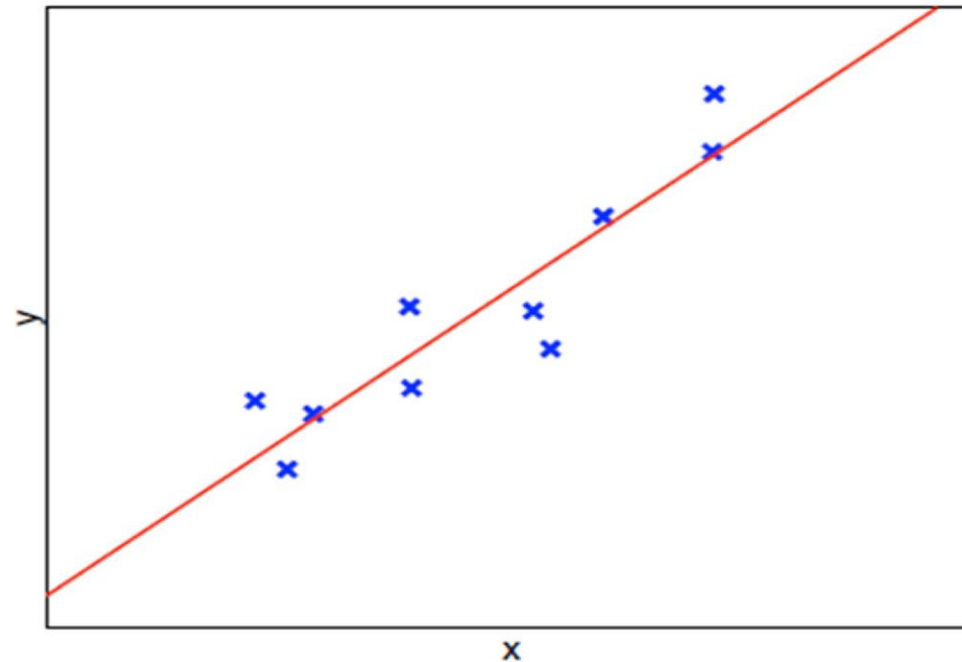
$$X^T X = \begin{bmatrix} 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.86 & 1 \\ 0.09 & 1 \\ -0.85 & 1 \\ 0.87 & 1 \\ -0.44 & 1 \\ -0.43 & 1 \\ -1.10 & 1 \\ 0.40 & 1 \\ -0.96 & 1 \\ 0.17 & 1 \end{bmatrix} = \begin{bmatrix} 4.95 & -1.39 \\ -1.39 & 10 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix} = \begin{bmatrix} 6.49 \\ 8.34 \end{bmatrix}$$

Estimation des paramètres du modèle linéaire

$$\tilde{\theta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 4.95 & -1.39 \\ -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 1.60 \\ 1.05 \end{bmatrix}$$

La meilleure ligne est égale alors à: $y = 1.6x + 1.05$



Fonctions d'ordre supérieur

- Si $(X^T X)$ n'est pas réversible, alors ajouter des termes d'ordre supérieur. Appliquer une transformation des entrées dans un autre espace puis faire la régression dans le nouveau espace.

- Soit x une variable d'entrée unidimensionnelle. Si nous voulons appliquer un polynôme d'ordre supérieur aux données d'apprentissage, on aura:

$$\hat{y} = \theta_0 x_0 + \theta_1 x + \theta_2 x^2 + \epsilon$$

- Pour un polynôme d'ordre m , on aura:

$$\mathbf{X} = \begin{bmatrix} x^{(1)m} & \dots & x^{(1)2} & x^{(1)} & 1 \\ x^{(2)m} & \dots & x^{(2)2} & x^{(2)} & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x^{(N)m} & \dots & x^{(N)2} & x^{(N)} & 1 \end{bmatrix}$$

Si $m=2$, c'est une régression quadratique.

Régression linéaire

- La **complexité** d'un modèle de régression
- **Complexité = Nombre de coefficients** utilisés dans le modèle.
- Les coefficients sont estimés lors de l'apprentissage du modèle

Régularisation des paramètres de la régression

- Le modèle de régression peut être non stable et sensible aux variables d'entrée (les coefficients estimés de la régression sont larges).

1) Introduire une pénalité supplémentaire pour le modèle (qui a des coefficients large) → La régularisation des paramètres du modèle.

2) Imposer une distribution à priori des paramètres du modèle.

- La régularisation des paramètres permet de:
 - Pénaliser les valeurs extrêmes des paramètres.
 - Minimiser l'erreur (variance) et réduire la complexité du modèle.

Régularisation des paramètres de la régression

- Lasso Regression: les moindres carrés ordinaires sont modifiés pour minimiser également la somme absolue des coefficients (appelée régularisation L1).

Fonction de coût pour « Lasso regression »

$$\sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - \sum_{j=0}^d \theta_j * x_{ij})^2 + \lambda \sum_{j=0}^d |\theta_j|$$

Indépendant des données d'apprentissage

pour $t > 0, \sum_{j=0}^d |\theta_j| < t$

λ : terme de pénalité qui permet de régulariser les coefficients.

Régularisation des paramètres de la régression

- **Ridge Regression** : les moindres carrés ordinaires sont modifiés pour minimiser également la somme absolue au carré des coefficients (appelée régularisation L2).

Fonction de coût pour « Ridge regression »

$$\sum_{i=1}^N (y_i - \widehat{y}_i)^2 = \sum_{i=1}^N (y_i - \sum_{j=0}^d \theta_j * x_{ij})^2 + \lambda \sum_{j=0}^d \theta_j^2, \text{ pour } c > 0, \sum_{j=0}^d \theta_j^2 < c$$

Indépendant des données d'apprentissage

λ : terme de pénalité qui permet de régulariser les coefficients.

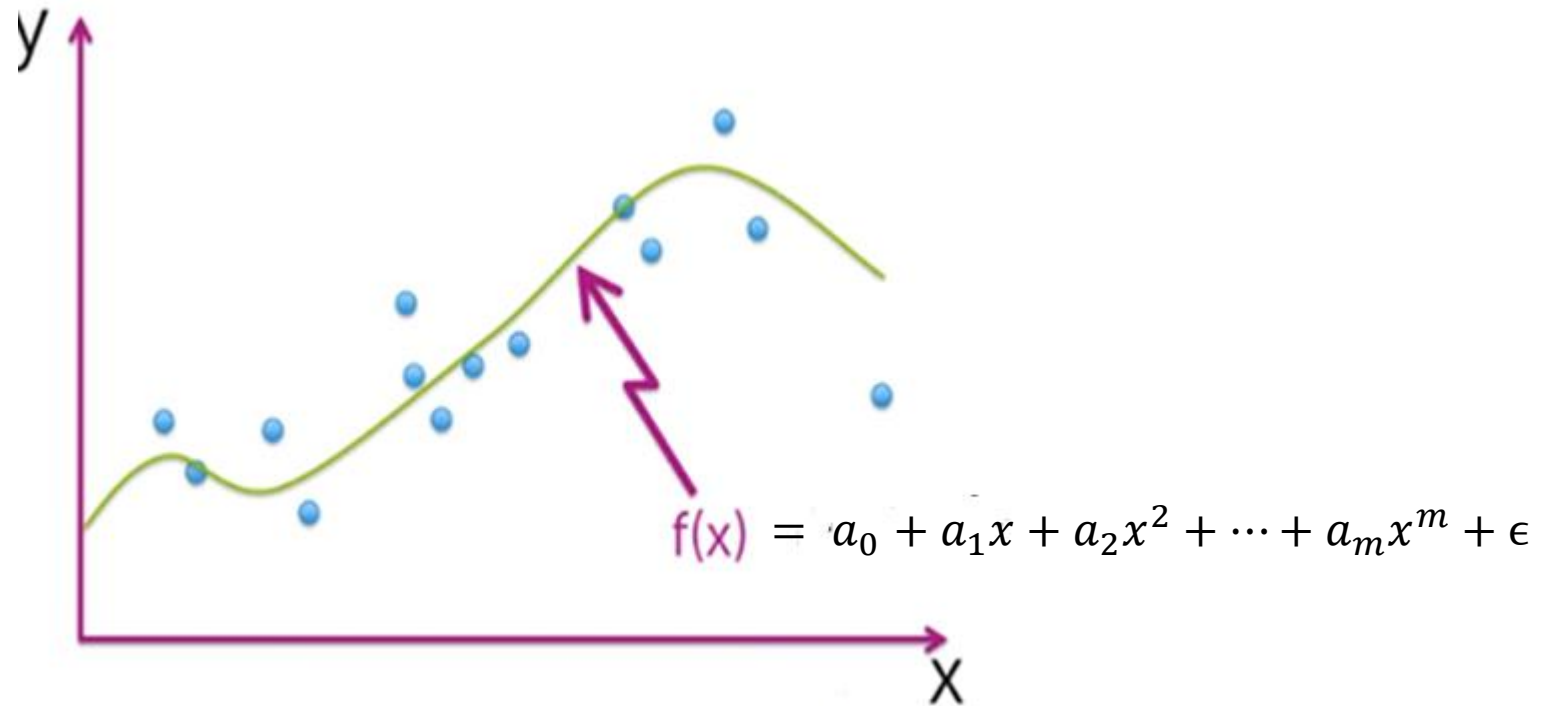
Régression polynomiale non linéaire

- La régression polynomiale est un modèle non linéaire simple où nous avons deux variables Y et X :
 - Y est la variable expliquée et
 - X est la variable explicative , seulement ici X est donnée sous forme d'une fonction plus complexe de la seule entrée X.
- Ajuster le nuage de points par une courbe d'équation:

$$f(x) = \hat{y}_i = a_0 + a_1x + a_2x^2 + \dots + a_mx^m + \epsilon$$

$$f(x) = \sum_{j=0}^m a_j x^j$$

Régression polynomiale non linéaire



Estimation des paramètres du modèle polynomiale

Pour un ensemble données d'entraînement $(x_i, y_i), i = 1 \dots N$

Chercher les meilleurs paramètres $a_j / j = 1 \dots m$:

$$\text{Min } E(a_j) = \frac{1}{2N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + \dots + a_m x_i^m + \epsilon$$

- *Calculer la dérivée partielle de l'erreur pour chaque coefficient: $\frac{\partial E}{\partial a_j} = 0$*
→ Résoudre un système linéaire de plusieurs équations avec des paramètres inconnus.

Estimation des paramètres du modèle polynomiale

- On utilise des algorithmes itératifs pour résoudre ce système linéaire. Parmi ces algorithmes :
 - [Algorithme de Gauss-Newton](#) ;
 - [Algorithme de Levenberg-Marquardt](#) ;
 - [Algorithme de descente du gradient](#).

Estimation des paramètres du modèle polynomiale

Notation matricielle:

$$a_j = (X^T X)^{-1} X^T y$$

Si on a polynôme d'ordre m :

$$\mathbf{X} = \begin{bmatrix} x^{(1)m} & \dots & x^{(1)2} & x^{(1)} & 1 \\ x^{(2)m} & \dots & x^{(2)2} & x^{(2)} & 1 \\ \vdots & \ddots & & \vdots & \\ x^{(N)m} & \dots & x^{(N)2} & x^{(N)} & 1 \end{bmatrix}$$

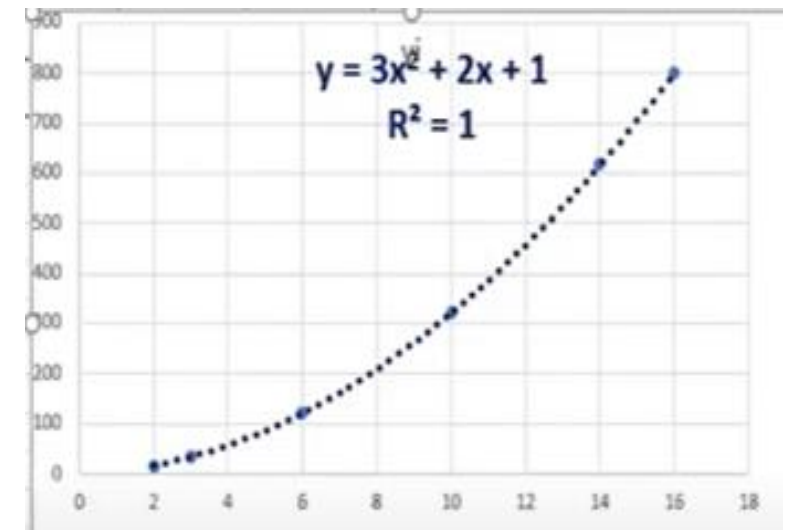
Estimation des paramètres du modèle polynomiale

Exemple:

On suppose qu'on a régression quadratique ($m=2$) avec une seule variable x , chercher le meilleur ajustement de la fonction f par rapport aux données d'apprentissage suivantes:

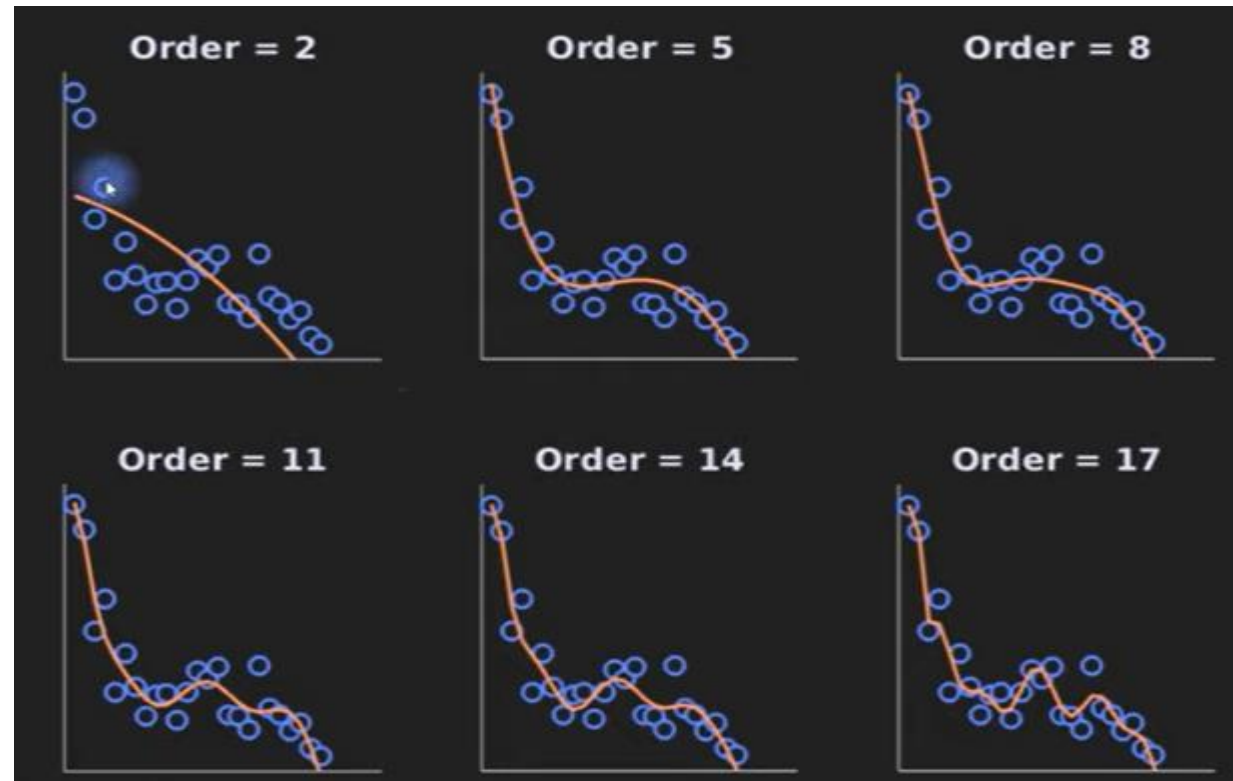
x_i	y_i
2	17
3	34
6	121
10	321
14	617
16	801

$$\longrightarrow f(x) = 3x^2 + 2x + 1$$



Estimation des paramètres du modèle polynomiale

Quel est le degrés optimal du polynôme qui s'ajuste au mieux avec les données d'apprentissage?



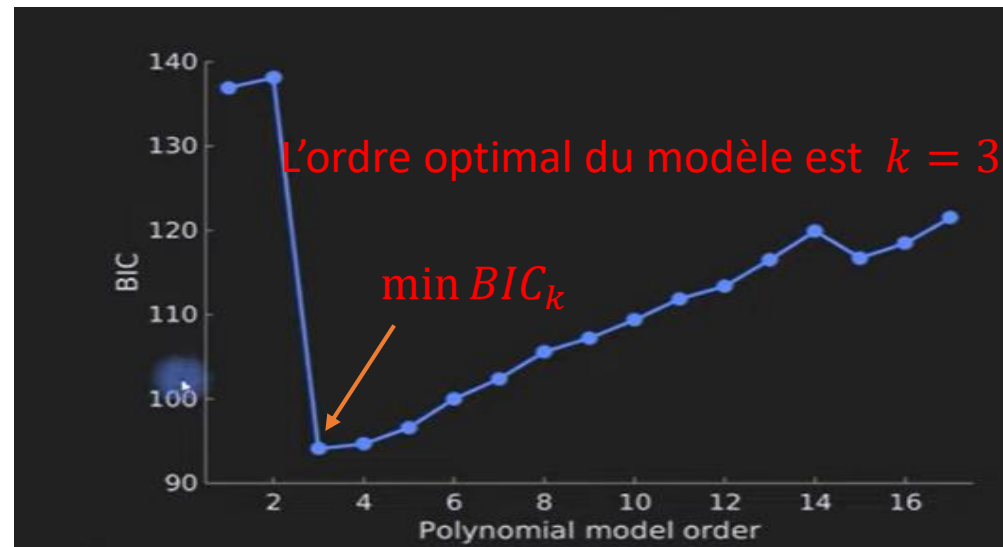
Estimation des paramètres du modèle polynomiale

Sélection d'ordre du modèle polynomiale

Critère d'information de Bayes:

$$k_{\text{BIC}} = \operatorname{argmin} \text{BIC}(k), \quad \text{BIC}_k = n \log(SS_\epsilon) + NP \log(n)$$

n : est le nombre de données, SS_ϵ : $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ et NP : est le nombre de paramètres du modèle



Régularisation des paramètres de la régression polynomiale

Pénaliser la fonction de coût :

$$E(\theta) = \frac{1}{2N} \sum_{i=1}^N (y_i - \widehat{y}_i)^2 + \frac{\lambda}{2} \sum_{j=0}^d \|\theta_j\|^2$$

$$\|\theta_j\|^2 = \theta_0^2 + \theta_1^2 + \dots + \theta_D^2$$

λ : terme de pénalité qui permet de régulariser les coefficients.

Régularisation des paramètres de la régression polynomiale

Valeurs des paramètres optimaux w_j pour les différents degrés du polynôme (M)

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

Plus λ est grand, plus la régularisation est forte

Modèle polynomiale et multi-colinéarité

Problème de multi-colinéarité

x_i, x_i^2, \dots, x_i^m sont des variables hautement corrélés, de fait qu'elles dépendent tous de x .

→ Standardisations des valeurs de x .

$$x'_i = \frac{x_i - \bar{x}}{std_x}$$

Régression non linéaire exponentielle

Ajustement exponentiel

$y = b e^{ax}$, a et b sont des nombres réels

$$\ln y = \ln(b e^{ax})$$

$$\ln y = \ln b + \ln e^{ax}$$

$$\ln y = ax + \ln b$$

Posons $Y = \ln y$ et $B = \ln b$

$Y = ax + B$ (relation de type linéaire ou affine)

Régression non linéaire exponentielle

Nuage de points $(x_i, y_i) \rightarrow$ ajustement exponentiel

x_i	y_i	y_{i+1}/y_i	$\ln y_i$
1	55		4.01
2	120	2.18	4.79
3	275	2.29	5.62
4	610	2.22	6.41
5	1365	2.24	7.22
6	3050	2.23	8.02

Nuage de points $(x_i, \ln y_i) \rightarrow$ ajustement linéaire

Régression non linéaire de puissance

Ajustement puissance

$y = b x^a$, a et b sont des nombres réels

$$\ln y = \ln(b x^a)$$

$$\ln y = \ln b + \ln x^a$$

$$\ln y = a \ln x + \ln b$$

Posons $Y = \ln y$, $X = \ln x$ et $B = \ln b$

$Y = aX + B$ (relation de type linéaire ou affine)

Régression non linéaire puissance

Nuage de points $(x_i, y_i) \rightarrow$ ajustement puissance

x_i	y_i	$\ln x_i$	$\ln y_i$
8	55	2.07	4.01
2.9	120	2.19	4.79
10.7	275	2.37	5.62
12.8	610	2.55	6.41
15	1365	2.71	7.22
18	3050	2.89	8.02

Nuage de points $(\ln x_i, \ln y_i) \rightarrow$ ajustement linéaire

Régression non linéaire logarithmique

Ajustement logarithmique

$y = a \ln x + b$ a et b sont des nombres réels

Posons $X = \ln x$

$y = aX + b$ (relation de type linéaire ou affine)

Résumé

1-Définir la problématique et l' **ensemble de données**

➔ **Hypothèse** (modèle de la **régression adéquat**)

2-**Minimiser la fonction du coût ➔ calcul des coefficients du modèle**

3- À l'aide du modèle, effectuer des prédictions

4- Évaluer performances