Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 2 : La régression lineaire

Ouadfel Salima
Faculté NTIC/IFA
salima.ouadfel@univ-constantine2.dz

Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 2 : La régression linéaire

Faculté NTIC/IFA

salima.ouadfel@univ-constantine2.dz

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
Nouvelles technologies	IFA	Master1	STIC

Apprentissage supervisé

L'apprentissage supervisé (*supervised learning*) consiste à générer **un modèle de prédiction** à partir de données annotées. L'annotation consiste à affecter à chaque donnée une étiquette qui est la réponse à prédire.

C'est une analyse prédictive

Apprentissage non supervisé

En apprentissage non supervisé (*unsupervised learning*), l'algorithme prend en entrée des données non annotées (sans leur label) et découvre par lui-même des similarités ou des différences entre les données à partir des features qui les décrivent.

C'est une analyse descriptive

La régression

Le modèle de régression est un **modèle d'apprentissage supervisé** qui permet de **prédire** une variable expliquée Y continue (dépendante) à partir de variables explicatives (indépendantes) X.

Exemples:

- Prédire le prix d'une maison Y en se basant sur sa surface (X_1), nombre de chambres (X_2) et son âge (X_3).
- Prédire le poids d'une personne adulte Y en fonction de sa taille (X).

La régression

Prédire le prix d’une maison

test

area	bedrooms	balcony	age
4600	5	3	1
2050	2	2	2
1450	2	2	3

area	bedrooms	balcony	age
1200	2	0	2
2300	3	2	5
2500	4	2	1
3650	5	3	3
1800	3	1	5
3000	3	1	4
1222	1	0	2

price
500000
620000
122500
6000000
2122000
120000
450000

Apprentissage

Modèle

Prédiction

price
6500000
1530000
1563330

Données d’apprentissage

Labels

label

La régression linéaire

La régression

Prédire le poids d'une personne

$Taille$
160
165
170
175
180
185
190

Données d'apprentissage

$Poids$
66
68.23
70.44
71.29
73.86
74.68
76.28

Labels

Apprentissage

test

$Taille$
155
195

Modèle

Prédiction

$Poids$
68
80

label

La régression lineaire

Le modèle de régression est linéaire quand la variable expliquée Y change linéairement en fonction des variables explicatives indépendantes X.

On parle de modèle de régression linéaire :

- simple si le modèle permet de prédire **la variable expliquée Y** à partir **d'une variable explicative X**
- multiple si elle permet de prédire la **variable expliquée** Y à partir de **plusieurs variables explicatives X_i.**

La régression linéaire simple

Prédire le poids d'une personne adulte en fonction de sa taille.

	X:Variable explicative	Y:Variable expliquée
	Taille	*Poids*
1	160	66
2	165	68.23
3	170	70.44
4	175	71.29
5	180	73.86
6	185	74.68
7	190	76.28

La régression linéaire multiple

Exemple: Prédire le prix d'une maison en se basant sur sa surface, nombre de chambres et son âge.

X:Variables explicatives

Y:Variable expliquée

Surface	Chambres	Age	Prix
2600	2	20	550000
3000	3	15	585000
3200	4	18	610000
3600	4	10	595000
4000	5	8	760000

Le modèle de régression linéaire simple:

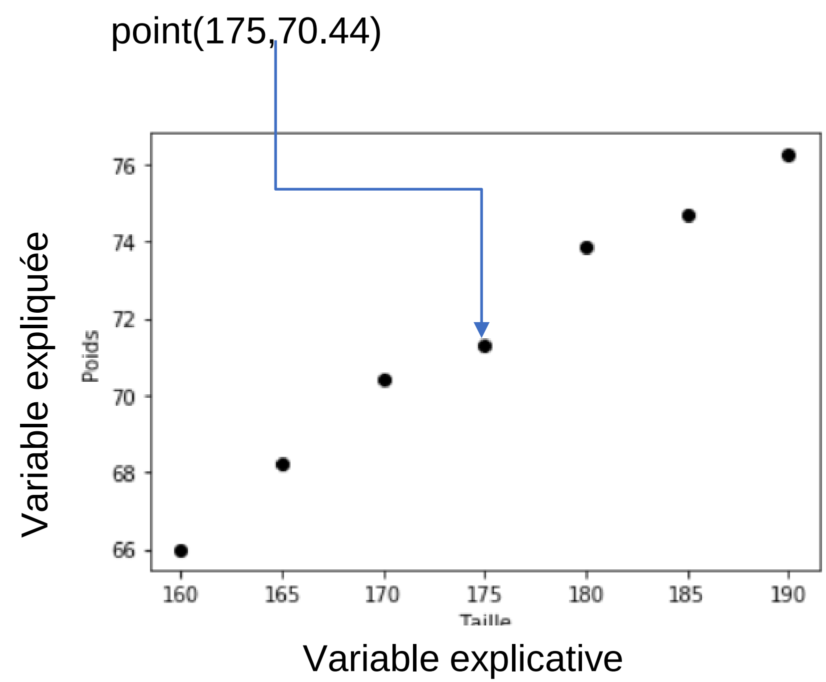
Représentation graphique Nuage de points

Soit n données (n samples) $(x_i, y_i)\ i = 1..n$ tel que:

x_i est une variable explicative et y_i est une variable expliquée

Un nuage de points est une représentation graphique dans un plan des paires $(x_i, y_i)\ i = 1..n$ tel que la variable expliquée est sur l'axe Y et la variable explicative est sur l'axe X.

	X:Variable explicative	Y:Variable expliquée
	Taille	*Poids*
1	160	66
2	165	68.23
3	170	70.44
4	175	71.29
5	180	73.86
6	185	74.68
7	190	76.28

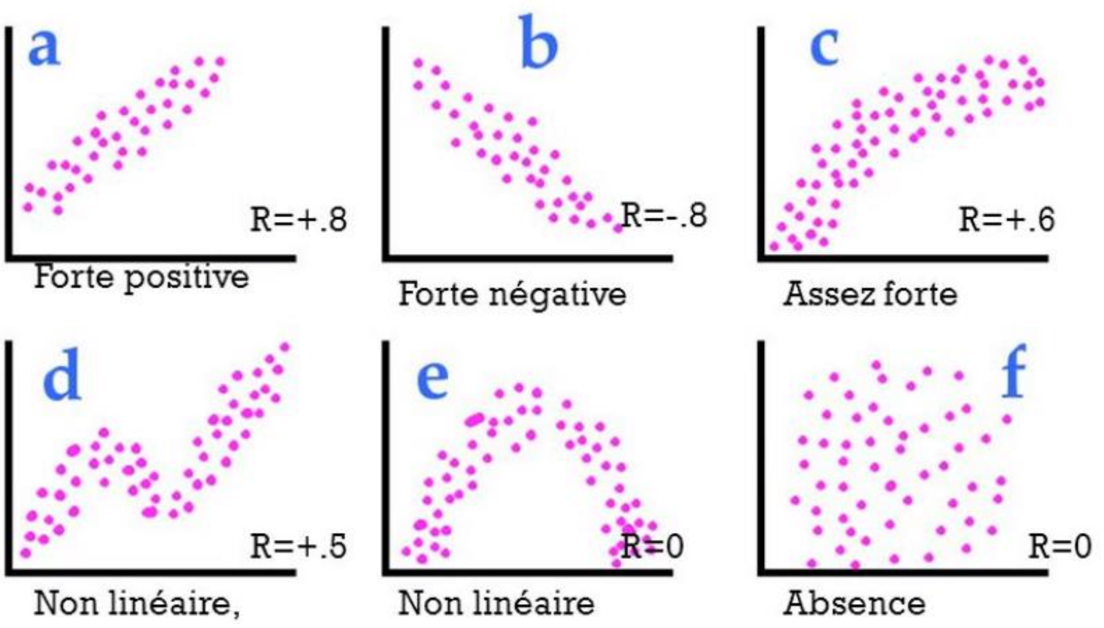


Le modèle de régression linéaire simple:

Représentation graphique Nuage de points

Les nuages de points montrent le type de relation entre les deux variables continues X et Y.

R est le coefficient de corrélation.
R est entre -1 et +1

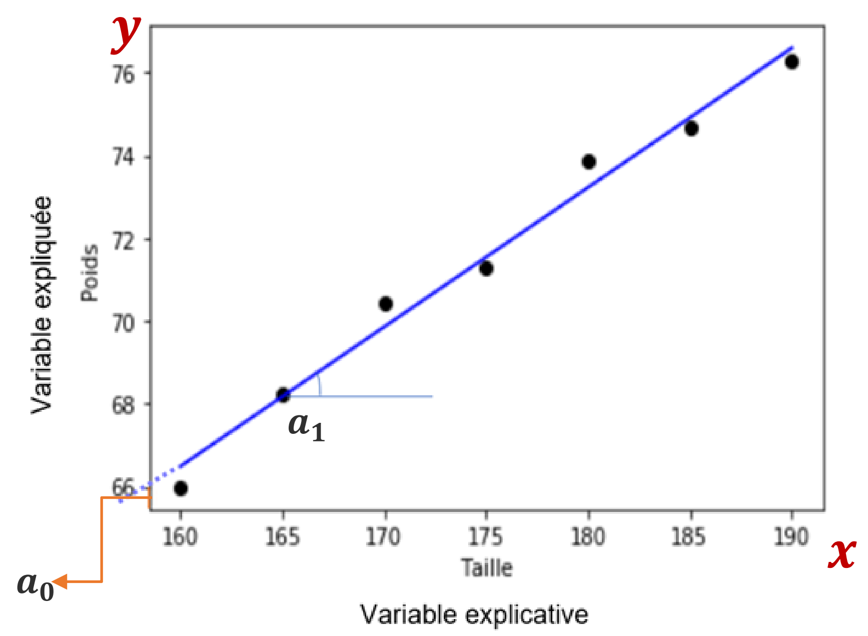


Le modèle de régression linéaire simple:

Le modèle de régression linéaire qui exprime une relation linéaire entre une variable explicative $\boldsymbol{x}$ et une variable explique $\boldsymbol{y}$, est donné par l'équation suivante:

$$\boldsymbol{y = a_0 + a_1 x + \varepsilon}$$

$\boldsymbol{a_0}$: point d'intersection ($\boldsymbol{x}$=0)
$\boldsymbol{a_1}$: pente de la droite
ε : erreur de prédiction



L'équation de la régression linéaire simple

$$y_i = \boldsymbol{a_0} + \boldsymbol{a_1} x_i + \varepsilon_i$$

$i^{\text{ème}}$ donnée expliquée de Y

Terme intersection

Influence de la $i^{\text{ème}}$ donnée explicative de x

Erreur sur la $i^{\text{ème}}$ donnée

Le modèle de régression linéaire simple

Valeur observée

$$y_i \quad = a_0 + a_1 x_i \; + \varepsilon_i$$

Valeur prédite

$$\widehat{y_i} = \hat{a}_0 + \hat{a}_1 x_i$$

Erreur de prédiction

$$\varepsilon_i = y_i \; - \widehat{y_i}$$

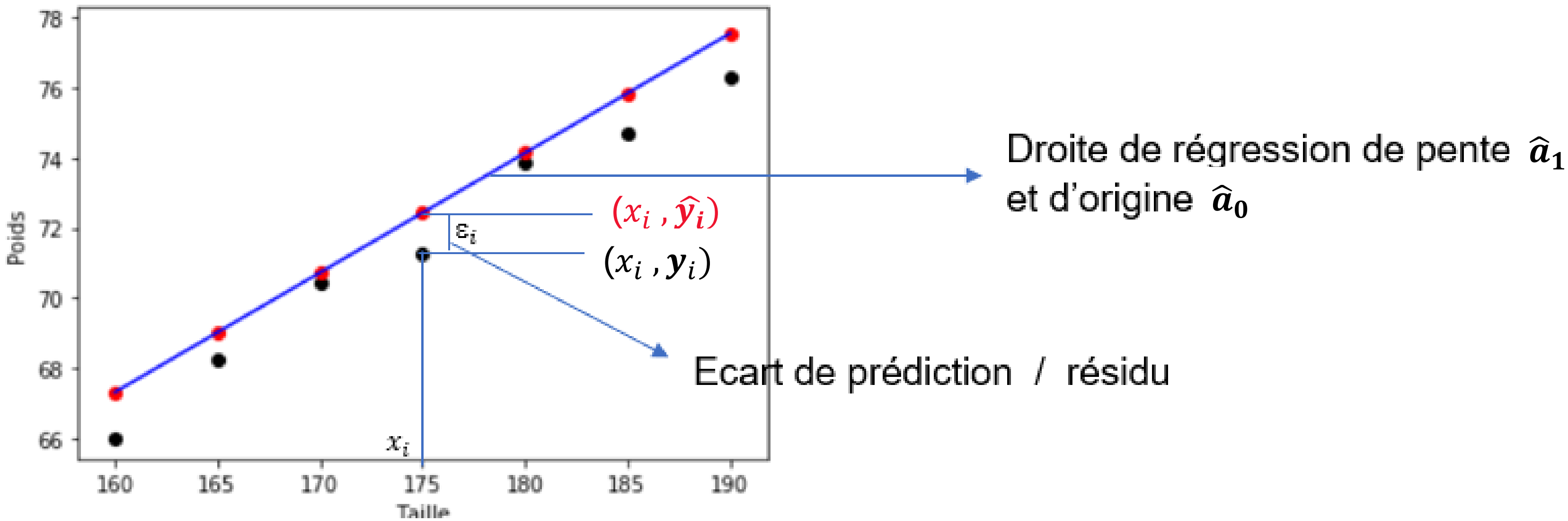
Le modèle de régression linéaire simple

Exemple

x_i :valeur de la variable taille

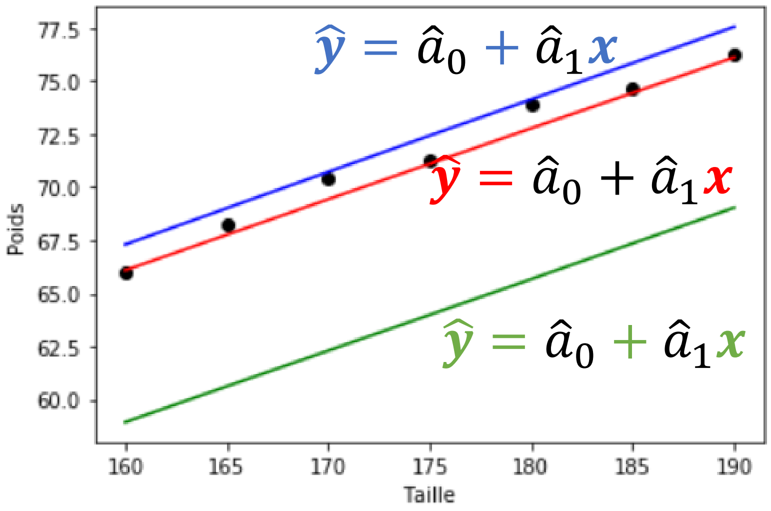
$\boldsymbol{y}_i$: valeur réelle de la variable poids

$\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$: valeur prédite de la variable poids



Représentation graphique de la droite de la régression linéaire simple

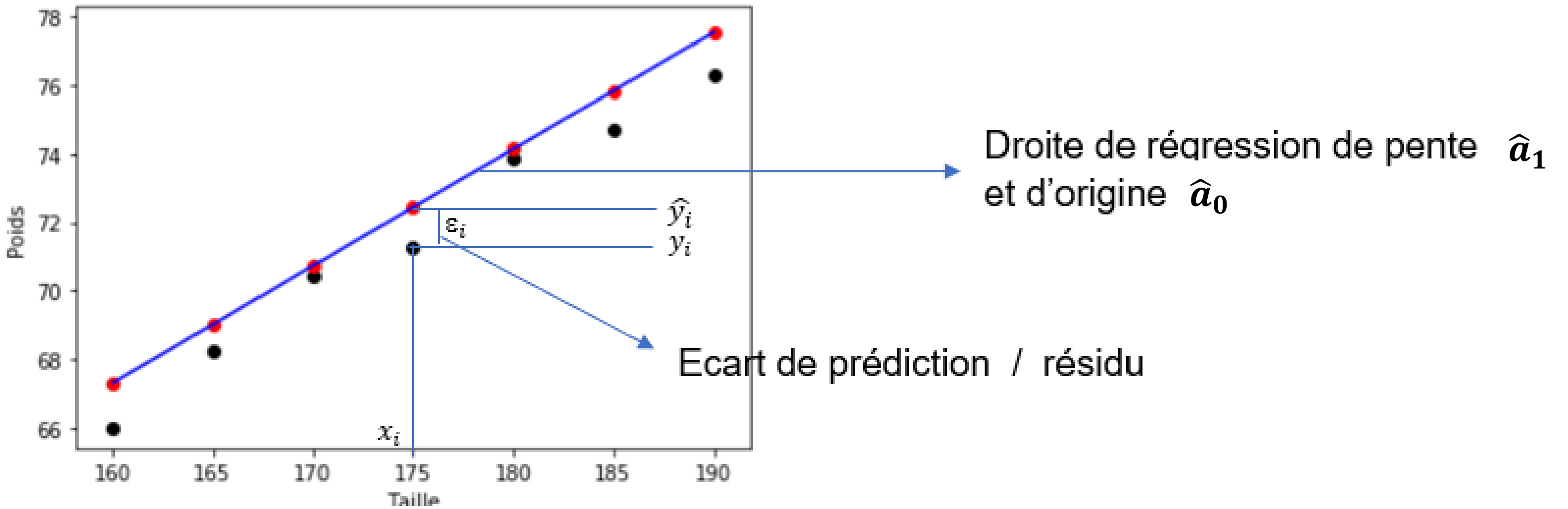
Le modèle de régression linéaire simple



On peut trouver plusieurs droites de régression selon les valeurs de $\hat{a}_0$ et $\hat{a}_1$

Comment choisir la bonne droite?

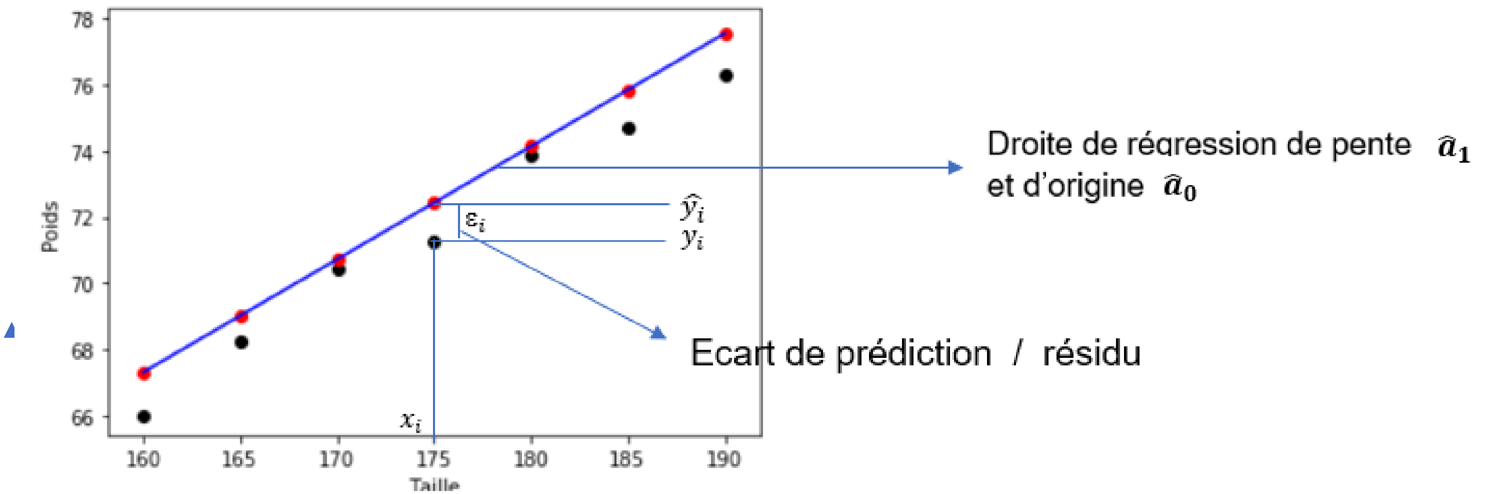
Le modèle de régression linéaire simple



La bonne droite est celle **qui s'ajuste le mieux** aux couples (x_i, y_i) $i = 1..n$

La bonne droite est celle qui minimise $\frac{1}{n}\sum_i^n \varepsilon_i^2$

Le modèle de régression linéaire simple



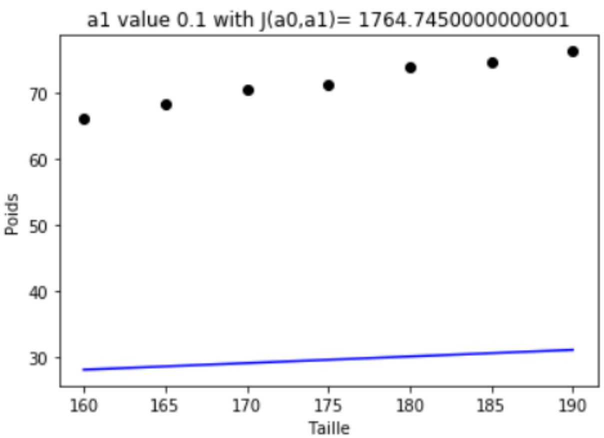
On cherche les paramètres $\boldsymbol{a_0}, \boldsymbol{a_1}$ telle que :

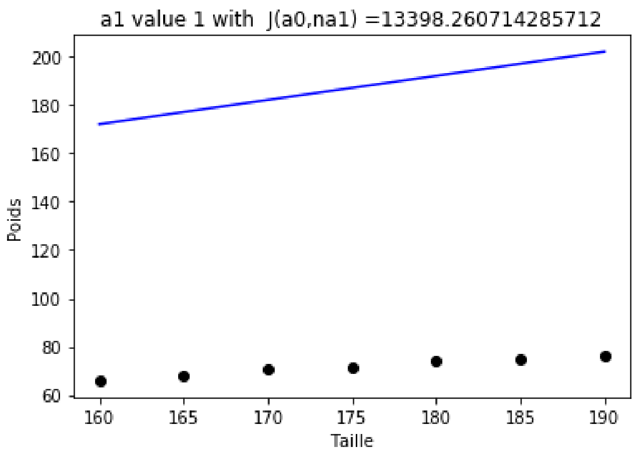
$$\frac{1}{n}\sum_i^n \varepsilon_i^2 = \frac{1}{n}\left[\left(y_1 - (\boldsymbol{a_0} + \boldsymbol{a_1}\boldsymbol{x_1})\right)^2 + \ldots\ldots \left(y_n - (\boldsymbol{a_0} + \boldsymbol{a_1}\boldsymbol{x_n})\right)^2\right]$$ est minimum

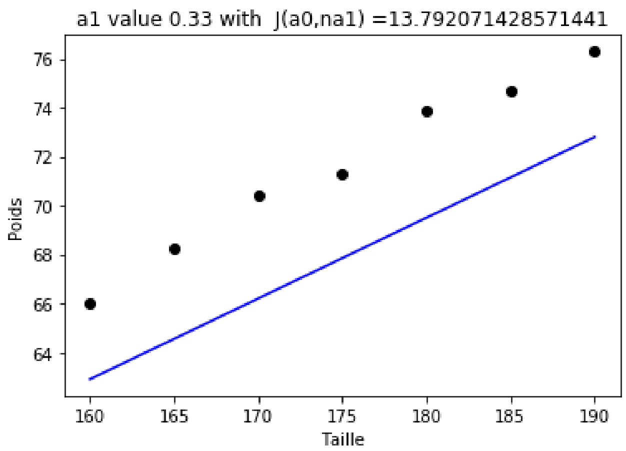
$J(a_0, a_1) = \frac{1}{n}\sum_i^n \varepsilon_i^2$ est la fonction coût du modèle de régression simple.

Le modèle de régression linéaire simple

Si on fixe a_0 et on fait varier a_1







La bonne droite est celle qui minimise $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$

Le modèle de régression linéaire simple:

La méthode des moindre carrées

Chercher les valeurs $\hat{a}, \hat{b}$ qui minimisent la somme des carrés.

$$\boldsymbol{J(\hat{a}_0, \hat{a}_1)} = argmin \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2$$

$$\boldsymbol{J}(\hat{a}_0, \hat{a}_1) = argmin \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = argmin \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{a}_0 + \hat{a}_1 x_i))^2$$

$$\hat{a}_0 = \frac{cov(x, y)}{var(x)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{a}_1 = \bar{y} - \hat{a}_0 \bar{x}$$

avec $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Le modèle de régression linéaire simple

Modèle : $y = a_0 + a_1 x + \varepsilon$

Parametres: $a_0 \;\; et \;\; a_1$

Fonction coût: $J(a_0, a_1) = \frac{1}{n} \sum_i^n \varepsilon_i^2$

But: minimiser $J(a_0, a_1)$

$$\hat{a}_1 = \frac{cov(x, y)}{var(x)} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

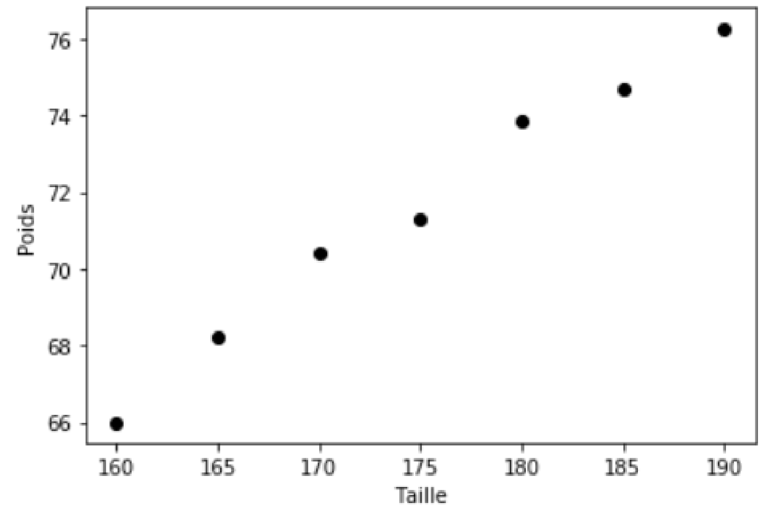
$$\hat{a}_0 = \overline{y} - \hat{a}_1 \overline{x}$$

Le modèle de régression linéaire simple:

La méthode des moindre carrées

Exemple

$Taille$	$Poids$
160	66
165	68.23
170	70.44
175	71.29
180	73.86
185	74.68
190	76.28



Le modèle de régression linéaire simple:

La méthode des moindre carrées

Exemple

$x = Taille$	$y = Poids$	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
160	66	-15	-5.54	225	83.1
165	68.23	-10	-3.31	100	33.1
170	70.44	-5	-1.1	25	5.5
175	71.29	0	-0.25	0	0
180	73.86	5	2.32	25	11.6
185	74.68	10	3.14	100	31.4
190	76.28	15	4.74	225	71.1

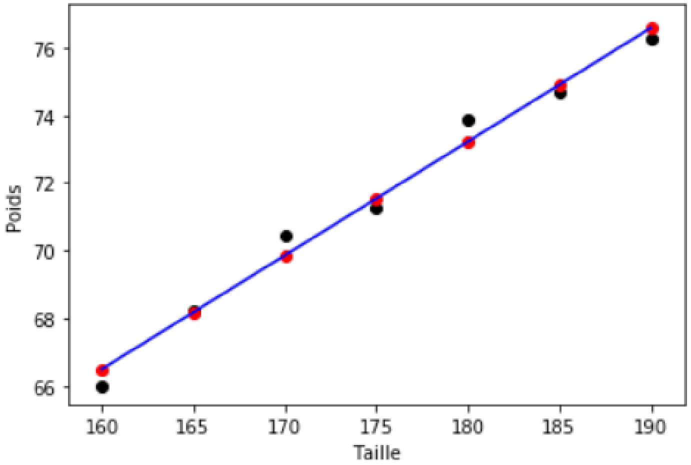
$\bar{x} = 175 \quad \bar{y} = 71.54$

$\sum_{i=1}^{7}(x_i - \bar{x})^2 = 700 \qquad \sum_{i=1}^{7}(x_i - \bar{x})(y_i - \bar{y}) = 235.80$

Le modèle de régression linéaire simple:

La méthode des moindre carrées

Exemple



$x = Taille$	$y = Poids$	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
160	66	-15	-5.54	225	83.1
165	68.23	-10	-3.31	100	33.1
170	70.44	-5	-1.1	25	5.5
175	71.29	0	-0.25	0	0
180	73.86	5	2.32	25	11.6
185	74.68	10	3.14	100	31.4
190	76.28	15	4.74	225	71.1

$\bar{x} = 175 \quad \bar{y} = 71.54$

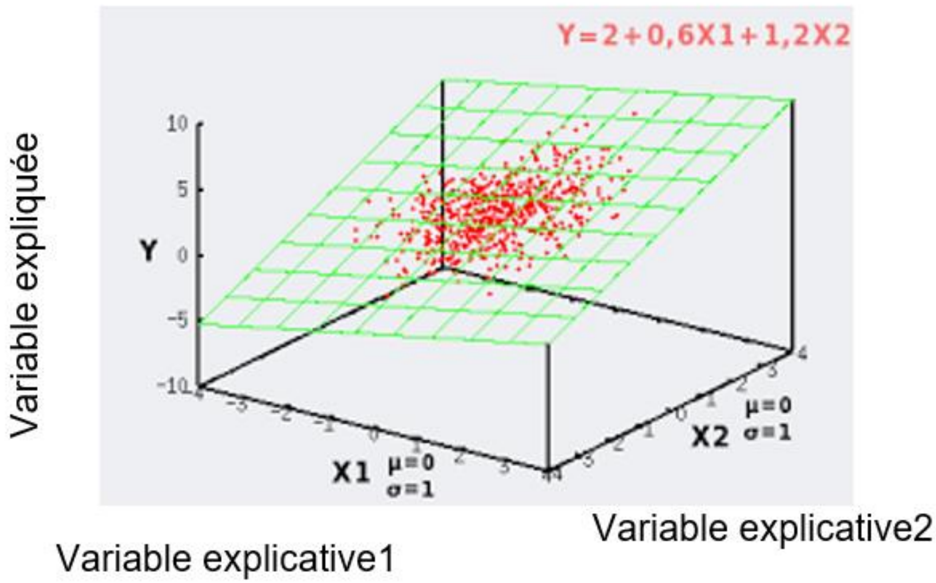
$\sum_{i=1}^{7}(x_i-\bar{x})^2 = 700 \qquad \sum_{i=1}^{7}(x_i-\bar{x})(y_i-\bar{y}) = 235.80$

$$\hat{a}_1 = \frac{\sum_{i=1}^{7}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{7}(x_i-\bar{x})^2} = 0.33686$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x} = 12.58$$

Le modèle de régression linéaire multiple

Le modèle de régression multiple est une généralisation du modèle de régression simple. On cherche à prédire une variable expliquée en fonction de deux ou plusieurs variables explicatives.



Le modèle de régression linéaire multiple

- **L'équation de regression multiple**

Le modèle théorique de la régression lineaire multiple est décrit par l'equation suivante:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots . + a_px_p + \varepsilon$$

où

$a_0, a_1, a_2, \ldots . . a_p$ **sont les parametres du modèle**

ε **représente le terme d'erreur**

Le modèle de régression linéaire multiple

L'équation de régression linéaire multiple

$$y_i \quad = a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \cdots . + a_p x_{ip} + \varepsilon_i$$

I^{eme} donnée de y

Terme d'intersection

Influence de la variable x_1

Influence de la variable x_2

Influence de la variable x_p

erreur de la I^{eme} donnée De x

Le bon hyperplan est celui **qui s'ajuste le mieux** aux couples $(x_i, \boldsymbol{y_i})$ $\boldsymbol{i = 1..n}$

Le modèle de régression linéaire multiple

La fonction coût est :

$$J(a_0, a_1,\ a_2, \ldots..a_p) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$$

Telle que

$$\boldsymbol{\varepsilon_i} = \left(\boldsymbol{y_i} - \left(a_0 + \boldsymbol{a_1 x_{i1}} + \boldsymbol{a_2 x_{i2}} + \boldsymbol{a_3 x_{i3}} + \cdots. + \boldsymbol{a_p x_{ip}}\right)\right)$$

Le modèle de régression linéaire multiple

Chercher les valeurs $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3 \ldots., \hat{a}_p$ qui minimisent la somme des carrés.

$$J(\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3 \ldots., \hat{a}_p) = argmin \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2$$

$$J(\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3 \ldots., \hat{a}_p) = argmin \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Avec

$$\hat{\boldsymbol{y}} = \hat{a}_0 + \hat{a}_1 \boldsymbol{x_1} + \hat{a}_2 \boldsymbol{x_2} + \hat{a}_3 \boldsymbol{x_3} + \cdots + \hat{a}_p \boldsymbol{x_p}$$

Le modèle de régression linéaire multiple

- **Estimation des parametres par la méthodes des moindres carrées**

Ecriture matricielle

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} \color{red}{1} & x_{11} \; .... & x_{1p} \\ \color{red}{1} & x_{21} \; .... & x_{2p} \\ & . & \\ & . & \\ \color{red}{1} & x_{n1} \; .... & x_{np} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ a_p \end{pmatrix}}_{a} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

$$Y = Xa + \varepsilon$$

La regression linéaire

Le modèle de régression linéaire multiple

- **Estimation des parametres par la méthodes des moindres carrées**

Ecriture matricielle

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ & & . & \\ & & . & \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ a_p \end{pmatrix}}_{a} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

$$Y = Xa + \varepsilon$$

$$\hat{a} = (X^T X)^{-1} X^T Y$$

Le modèle de régression linéaire multiple

Modèle : $$\boldsymbol{y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots . + a_px_p + \varepsilon}$$

Parametres: $\boldsymbol{a_0, a_1,\ a_2, \ldots\ldots a_p}$

Fonction coût: $$\boldsymbol{J(a_0, a_1,\ a_2, \ldots\ldots a_p) = \frac{1}{n}\sum_i^n \varepsilon_i^2}$$

But: minimiser: $\boldsymbol{J(a_0, a_1,\ a_2, \ldots\ldots a_p)}$

Transformation:
$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ & & \cdot & \\ & & \cdot & \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ a_p \end{pmatrix}}_{\boldsymbol{a}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

$$\boldsymbol{Y = Xa} + \varepsilon$$

$$\hat{a} = (X^TX)^{-1}X^TY$$

Le modèle de régression linéaire multiple

retour au modèle de régression simple

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}. \qquad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \qquad \mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} = n \begin{pmatrix} \bar{y} \\ s_{xy} + \bar{x}\bar{y} \end{pmatrix},$$

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\
&= \frac{1}{n^2\left\{\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2\right\}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\
&= \frac{1}{n^2 s_x^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\
&= \frac{1}{n^2 s_x^2} \begin{pmatrix} n s_x^2 + n\bar{x}^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \\
&= \frac{1}{n s_x^2} \begin{pmatrix} s_x^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix},
\end{aligned}$$

Le modèle de régression linéaire multiple

retour au modèle de régression simple

$$\hat{a} = (X^T X)^{-1} X^T Y$$

$$\hat{a} = \frac{1}{s_x^2}\begin{pmatrix}(s_x^2+\bar{x}^2)\bar{y}-\bar{x}(s_{xy}+\bar{x}\bar{y})\\ -\bar{x}\bar{y}+(s_{xy}+\bar{x}\bar{y})\end{pmatrix} = \begin{pmatrix}\bar{y}-\bar{x}\frac{s_{xy}}{s_x^2}\\ \frac{s_{xy}}{s_x^2}\end{pmatrix}$$

avec

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

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

Le modèle de régression linéaire multiple

Exemple

On veut prédire le prix d'une maison à partir de sa surface, du nombre de chambres et de son age.

On a l'ensemble d'apprentissage suivant:

Surface	Chambres	Age	Prix
2600	2	20	550000
3000	3	15	585000
3200	4	18	610000
3600	4	10	595000
4000	5	8	760000

Modèle de régression linéaire multiple

Le modèle de régression linéaire multiple

x_1 Surface	x_2 Chambres	x_3 Age	y Prix
2600	2	20	550000
3000	3	15	585000
3200	4	18	610000
3600	4	10	595000
4000	5	8	760000

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \varepsilon$$

Ecriture matricielle

$$y = aX + \varepsilon$$

X=

1	2600	2	20
1	3000	3	15
1	3200	4	18
1	3600	4	10
1	4000	5	8

y=

550000
585000
610000
595000
760000

$$\hat{a} = (X^T X)^{-1} X^T Y$$

La régression linéaire

La méthode des moindres carrés est une méthode analytique qui permet de trouver une solution exacte. Mais elle devient difficile avec un très grand nombre de données et de caractéritisques à cause de l'inversion de la matrice.

Une alternative à la méthode des moindres carrés est une méthode approximative : l'algorithme ***Descente de Gradient*** pour trouver une solution approximative.