Réseaux de neurones artificiels

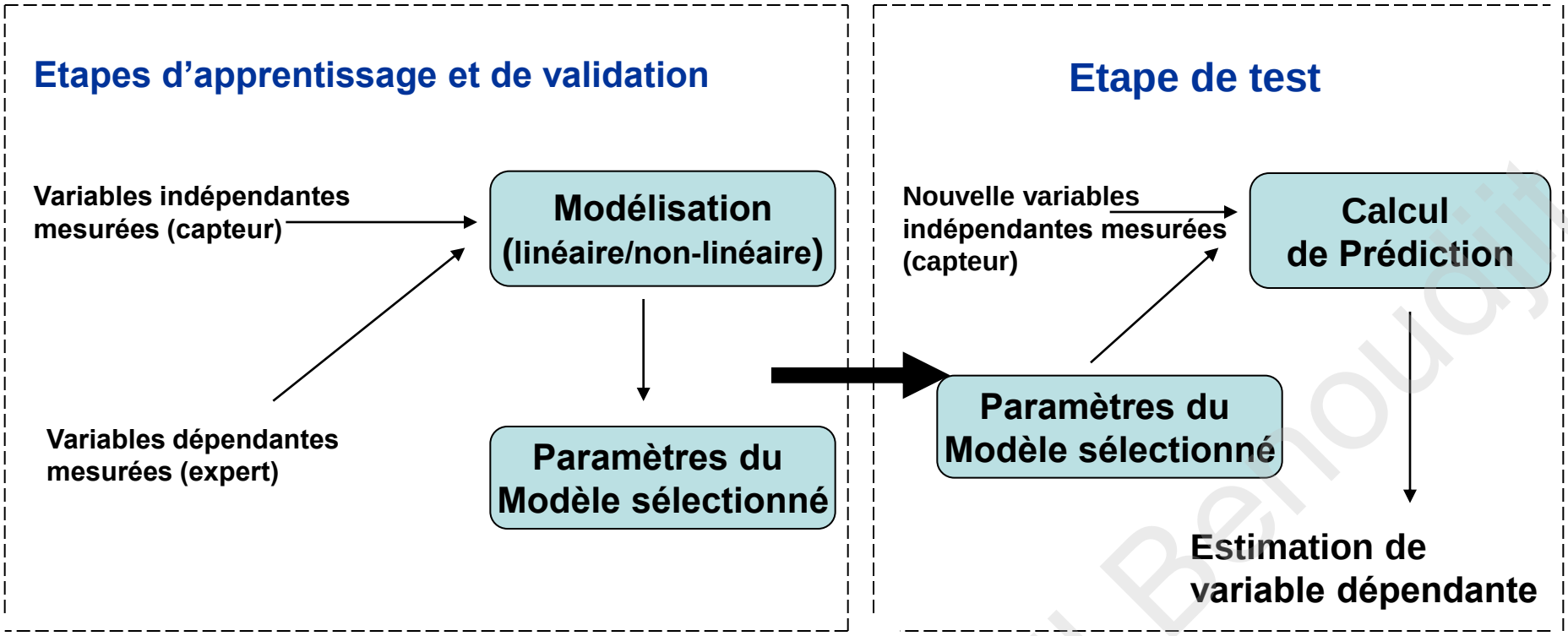
Pr. Nabil Benoudjit

Université de Batna -2-

Email: n.benoudjit@univ-batna2.dz

Modèle linéaire (Régression linéaire simple et Multiple)

Comment construire et évaluer un modèle?



Remarque: Pour le modèle linéaire on a besoin seulement de deux étapes

- Apprentissage (Estimation des paramètres du modèle)
- Test (Tester les performances du modèle)

Base de données (Datasets) (1)

- Avant de construire un modèle, les échantillons sont souvent subdivisés en ensembles de données d'apprentissage, de validation et de test. Les distinctions entre ces ensembles sont cruciales, mais les termes d'ensemble de données de validation et d'ensemble de données de test sont souvent confondus dans la littérature.
- **Ensemble d'apprentissage (Training set):**
  - L'ensemble de données d'apprentissage est utilisé pour apprendre ou construire un modèle. Par exemple, dans la régression linéaire, l'ensemble de données d'apprentissage est utilisée pour ajuster le modèle de régression linéaire, à savoir pour calculer les coefficients de régression. Dans un modèle non linéaire tel que un réseau neuronal, l'ensemble de données d'apprentissage est utilisée pour estimer les poids de réseau.

Base de données (Datasets) (2)

- **Validation dataset:**
  - Une fois qu'un modèle est construit sur des données d'apprentissage, nous devons trouver la précision du modèle sur les données inconnues. A cet effet, le modèle doit être utilisé sur un ensemble de données qui n'a pas été utilisé dans le processus d'apprentissage. Si nous devions utiliser les mêmes données d'apprentissage pour calculer la précision de l'ajustement du modèle, nous obtenons une estimation trop optimiste de la précision du modèle. En effet, le processus d'apprentissage garantit que la précision du modèle sur les données d'apprentissage est aussi élevée que possible - le modèle est spécifiquement adapté aux données d'apprentissage. Pour obtenir une estimation plus réaliste de la façon dont le modèle se comporte avec des données inconnues, nous avons besoin de mettre de côté une partie des données d'origine et de ne pas l'utiliser dans le processus d'apprentissage. Cette base de données est connu sous le nom d'ensemble de données de validation. Après ajustement du modèle sur l'ensemble de données d'apprentissage, nous devrions tester ses performances sur cet ensemble.

Base de données (Datasets) (3)

- **Test dataset:**

  - L'ensemble de validation est souvent utilisé pour trouver le meilleur modèle non linéaire. Par exemple, nous pourrions essayer différents modèles de réseaux neuronaux avec diverses architectures (par exemple avec différent nombre de neurones dans la couche cachée du RBF) et de tester la précision de chacun sur l'ensemble de données de validation pour choisir la meilleure architecture (les meilleurs paramètres).
  - Ainsi, nous avons besoin de mettre de côté une autre partie des données, qui n'est pas utilisé ni en apprentissage, ni dans la validation. Cet ensemble est connu sous le nom d'ensemble de données de test. La précision du modèle sur les données de test donne une estimation réaliste de la performance du modèle sur des données complètement inconnus.

Régression Linéaire simple (1)

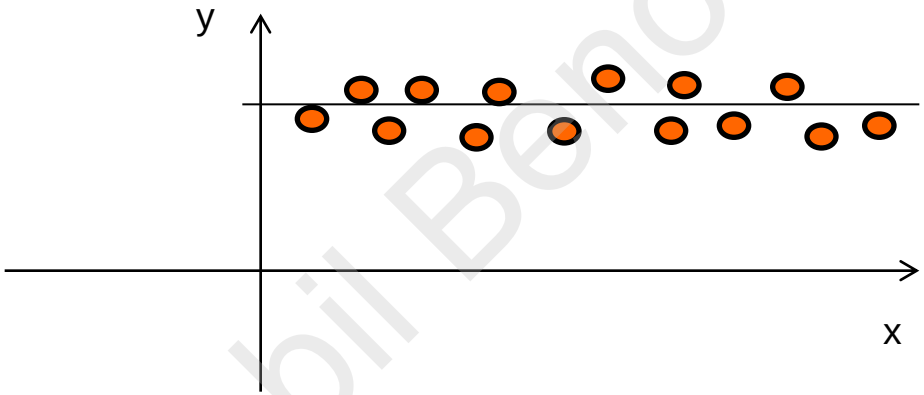
Liaison linéaire entre X et Y

- Avant d'estimer la droite de régression, il faut vérifier:
  - empiriquement (graphiquement) que la liaison entre les 2 variables est de nature linéaire.
- A défaut, l'interprétation du test de la pente de la droite de régression peut être erronée.

Cas 1:

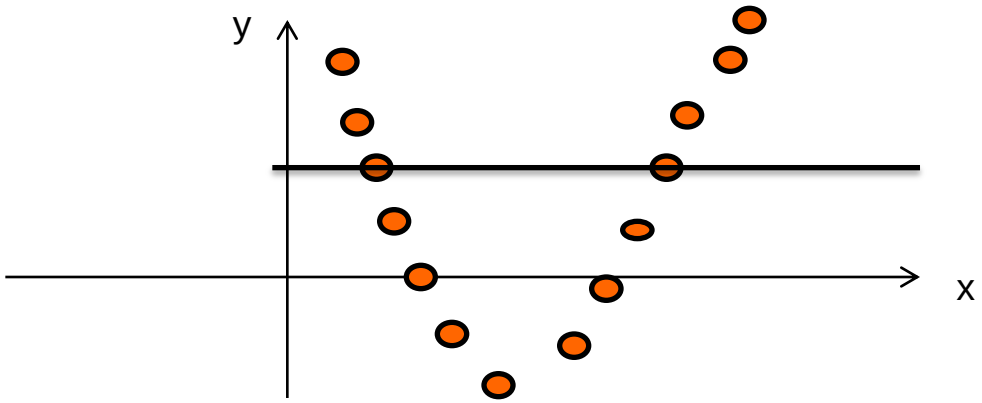
* La nature de la liaison est linéaire (le nuage de points est résumé au mieux par une droite horizontale d'équation y = b)
* La condition d'application est vérifiée
* Il est possible d'utiliser la régression linéaire simple pour quantifier la liaison entre les 2 variables.

(Conclusion : X et Y sont indépendants [Y constant quelle que soit la valeur de X]).



Régression Linéaire simple (2)

Cas 2:



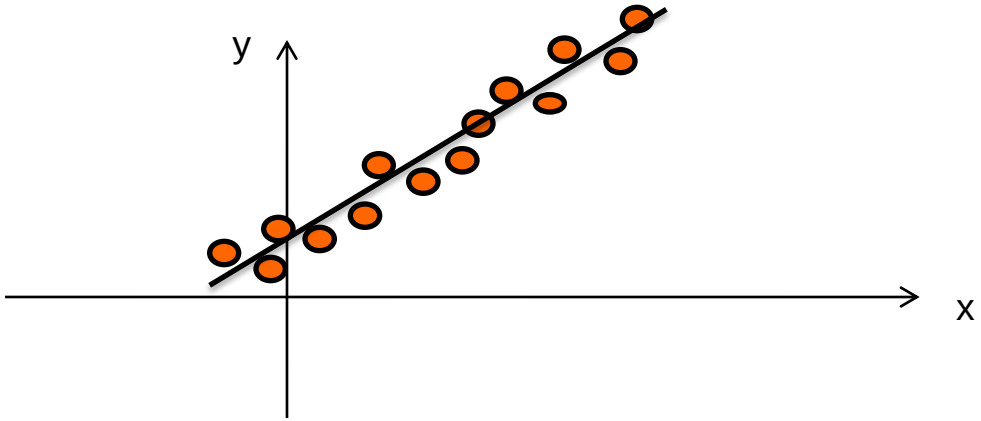
*** Le nuage de points n'est pas résumé au mieux par une droite mais plutôt par une fonction quadratique.**

La condition d'application n'est pas vérifiée.

(Conclusion: Il ne faut pas utiliser la régression linéaire simple pour quantifier la liaison entre les 2 variables x et y)

Régression Linéaire simple (3)

Cas 3:

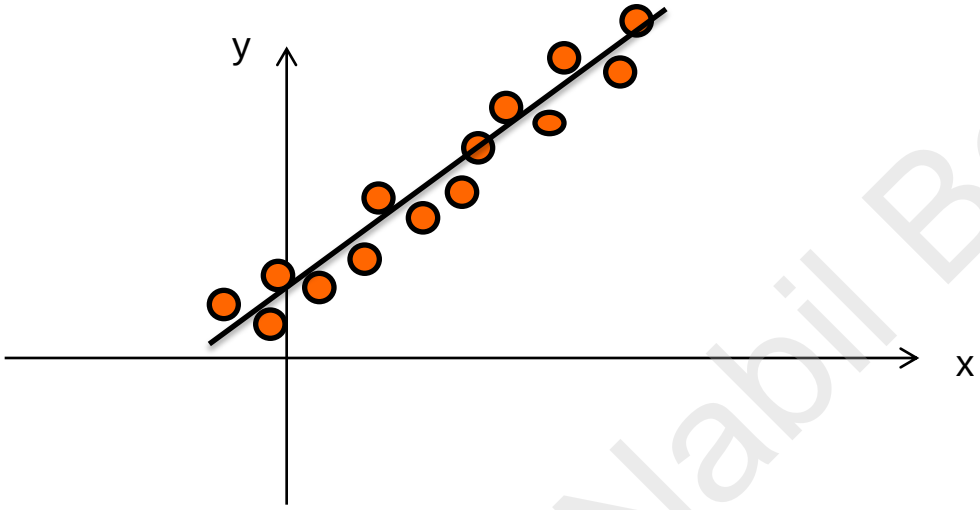


- **La nature de la liaison est linéaire (le nuage de points est résumé au mieux par une droite d'équation y = a*x + b).**
- **La condition d'application est vérifiée.**
- **Il est possible d'utiliser la régression linéaire simple pour quantifier la liaison entre les 2 variables.**

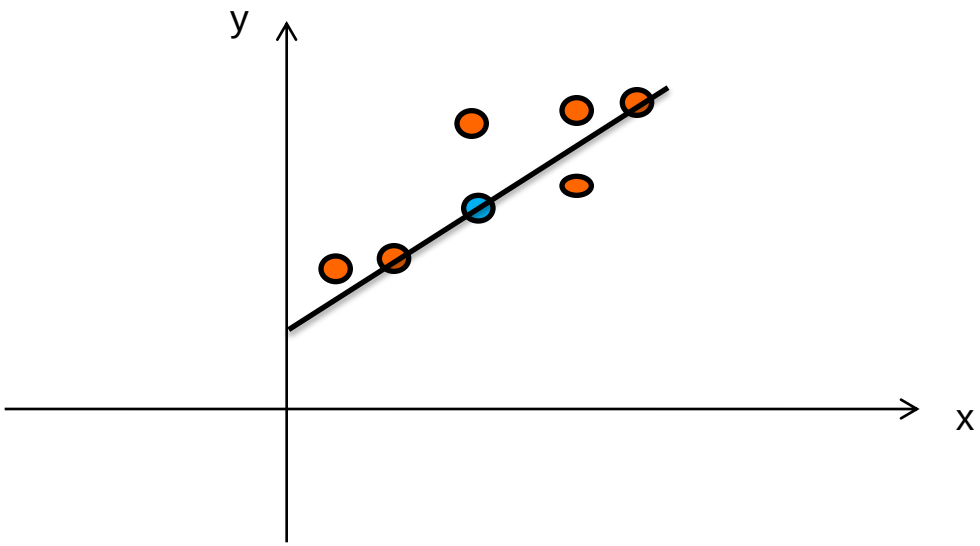
(Conclusion : il existe une liaison linéaire entre X et Y)

Régression Linéaire simple (4)

- La régression s'adresse à un type de problème où les 2 variables quantitatives continues **x** et **y** ont un rôle asymétrique : la variable **y** dépend de la variable **x**.
- La liaison entre la variable **y** dépendante (**dite expliquée**) et la variable **x** indépendante (**dite explicative**) peut être modélisée par une fonction de type **y = a*x + b**, représentée graphiquement par une droite.
- La variable **x**, peut être soit aléatoire, soit contrôlée c'est-à-dire qu'elles est connue **sans erreur**.



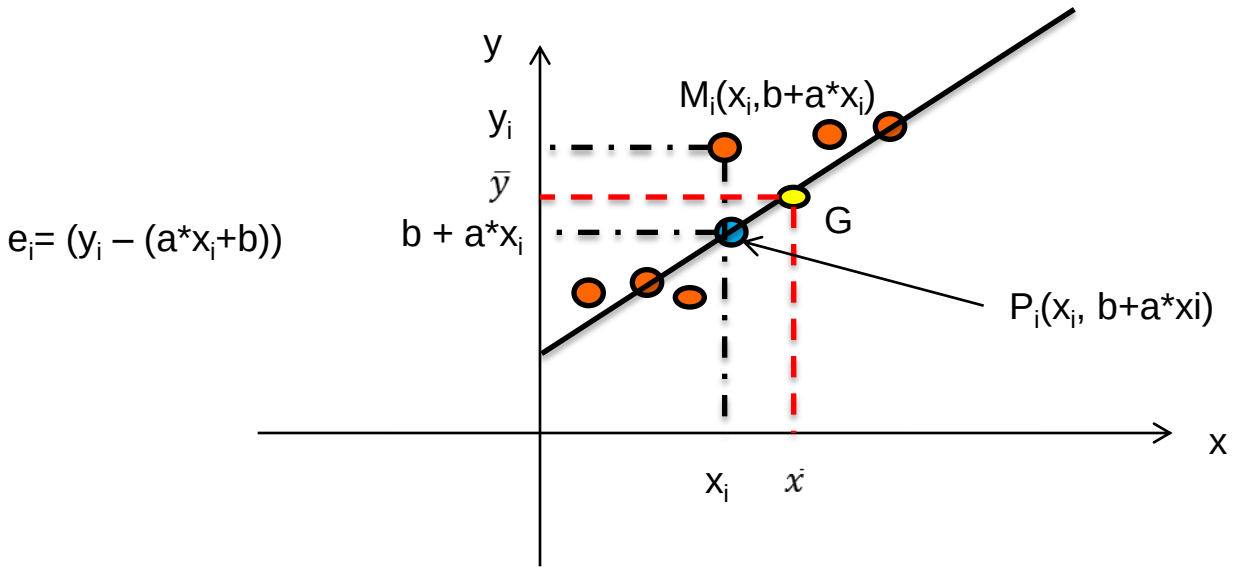
Estimation des paramètres par la méthode des moindres carrés (1)



* Chaque échantillon (individu) i est caractérisé par un couple de coordonnées (x_i, y_i) et est représenté par un point sur le graphique.

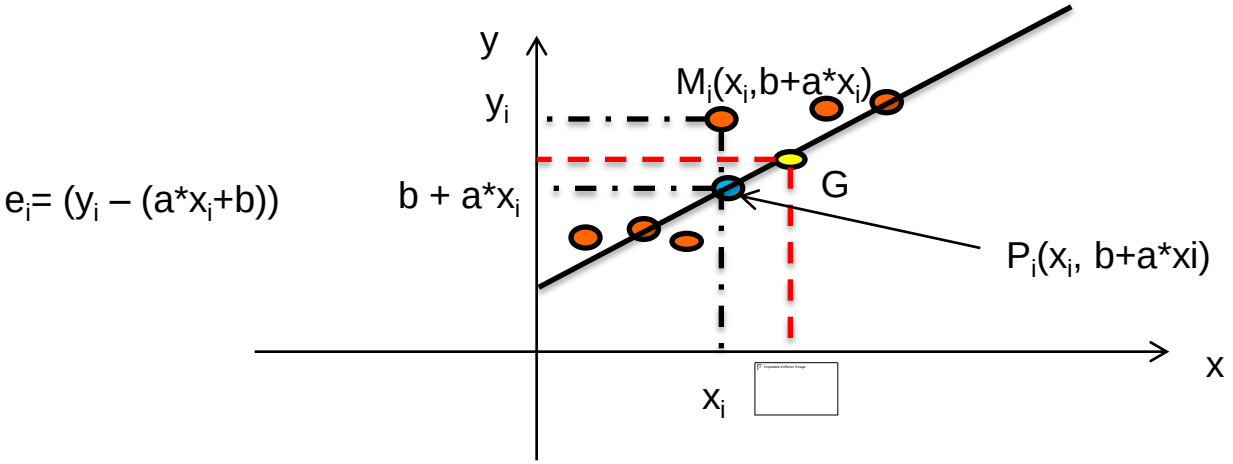
- L'ensemble des individus forme un nuage de points.
- (x_1,y_1), (x_2,y_2), ……..(x_n,y_n).
- L'objectif de trouver y = f(x) où f fonction linéaire donc y = a*x + b

Estimation des paramètres par la méthode des moindres carrés (2)



- L'objectif de trouver y = f(x) où f fonction linéaire donc y = b + a*x
- Les écarts sur la droite sont notés e_i, peuvent être positifs ou négatifs.
- Donc déterminer la droite $D_{y/x}$ pour laquelle: $\sum_{i=1}^{n} e_i^2$ soit minimale

Estimation des paramètres par la méthode des moindres carrés (3)



- L'objectif de trouver y = f(x) où f fonction linéaire donc y = b + a*x
- Les écarts sur la droite sont notés e_i, peuvent être positifs ou négatifs.
- Donc déterminer la droite $D_{y/x}$ pour laquelle: $\sum_{i=1}^{n} e_i^2$ soit minimale

Estimation des paramètres par la méthode des moindres carrés (4)

❑

$E = \sum_{y=1}^{n} e_i^2$

On a: $y = ax + b \quad (1)$

Au point G on a: $\bar{y} = a * \bar{x} + b \quad (2)$

On fait la soustraction de (1) - (2) on aura $(y - \overline{y}) = a * (x - \bar{x})$

sachant que $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ et $\bar{y} = \sum_{i=n}^{n} \frac{y_i}{n}$

Donc $E = \sum (y_i - (a * x_i + b))^2$

E minimale implique $\frac{\partial E}{\partial a} = 0$ et $\frac{\partial E}{\partial b} = 0$

$\frac{\partial E}{\partial b} = 2(-1) \sum (y_i - (a * x_i + b))$ et $\frac{\partial E}{\partial a} = -2 \sum \big(y_i - (a * x_i + b)\big) x_i$

Estimation des paramètres par la méthode des moindres carrés (5)

$$\frac{\partial E}{\partial b} = 0 \Longrightarrow -2\sum(y_i - (a * x_i + b)) = 0 \Longrightarrow \sum\left(y_i - (a * x_i + b)\right) = 0$$

$$\sum\left(y_i\right) - \sum(a * x_i + b) = 0$$

$$\sum y_i = \sum(a * x_i + b)$$

$$\sum y_i = a\sum x_i + \sum b$$

$$\sum y_i = a\sum x_i + n * b$$, on divise les deux membres par *n*

$$\sum_{i=1}^{n} \frac{y_i}{n} = a\sum_{i=1}^{n} \frac{x_i}{n} + b$$

$$\bar{y} = a * \bar{x} + b$$

Ce résultat traduit bien que le point $G(\bar{x}, \bar{y})$ appartient à la droite $D_{y/x}$ ce point est appelé le centre du gravité du nuage de points.

Alors $b = \bar{y} - a * \bar{x}$

Estimation des paramètres par la méthode des moindres carrés (6)

$$\frac{\partial E}{\partial a} = -2\sum_{i=1}^{n}(y_i - (a * x_i + b))x_i$$

$$-2\sum\left(y_i - (a * x_i + b)\right)x_i = 0$$

$$\sum\left(x_i y_i - \left(a * x_i^2 + b * x_i\right)\right) = 0$$

$$\sum\left(x_i y_i\right) = \sum a * x_i^2 + \sum b * x_i$$

Où: $b = \bar{y} - a * \bar{x}$

Alors : $\sum x_i y_i = \sum a * x_i^2 + \sum(\bar{y} - a * \bar{x})x_i$

$\sum x_i y_i = \sum a * x_i^2 + \sum \bar{y} x_i - \sum a * \bar{x} x_i$,

$\sum x_i\, y_i = a\sum x_i^2 + \bar{\bar{y}}\sum x_i - a * \bar{x}\sum x_i$, on divise les deux membres par *n*

$$\sum \frac{x_i y_i}{n} = a\sum \frac{x_i^2}{n} + \bar{y}\sum \frac{x_i}{n} - a * \bar{x}\sum \frac{x_i}{n}$$

$$\sum \frac{x_i y_i}{n} = a(\sum \frac{x_i^2}{n} - \bar{x}\bar{x}) + \bar{x}\bar{y}$$

$$a\left[\sum_{i=1}^{n} \frac{x_i^2}{n} - \bar{x}^2\right] = \sum_{i=1}^{n} \frac{x_i y_i}{n} - \bar{x}\bar{y}$$

Estimation des paramètres par la méthode des moindres carrés (7)

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

Estimation des paramètres par la méthode des moindres carrés (8)

- En statistique on peut écrire ceci:

Où $\sigma_x^2 = var(x) = \sum_{i=1}^{n} \frac{x_i^2}{n} - \bar{x}^2$

Et $\sigma_{xy} = cov(x,y) = \sum_{i=1}^{n} \frac{x_i y_i}{n} - \bar{x}\bar{y}$

Donc $a * \sigma_x^2 = \sigma_{xy}$ ====➔ $a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{cov(x,y)}{var(x)}$

Sous forme Matricielle on a:

y = x*A

Sous la forme matricielle

- Trouver le couple (a, b) de telle sorte que

$$\begin{pmatrix} ax_1 + b \\ : \\ ax_n + b \end{pmatrix} \text{ est proche de } \begin{pmatrix} y_1 \\ : \\ y_n \end{pmatrix}$$

Trouver la solution au sens des moindres carrés sous forme matricielle

$$\begin{pmatrix} x_1 & 1 \\ : & : \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ : \\ y_n \end{pmatrix}$$

Donc $X * A = Y \Longrightarrow A = (X^T X)^{-1} X^T Y$

Critères de performances:

- Coefficient de détermination R^2:

$$R^2 = 1 - \frac{SCE_D}{SCE_{\bar{y}}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- Erreur quadratique moyenne (MSE):

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

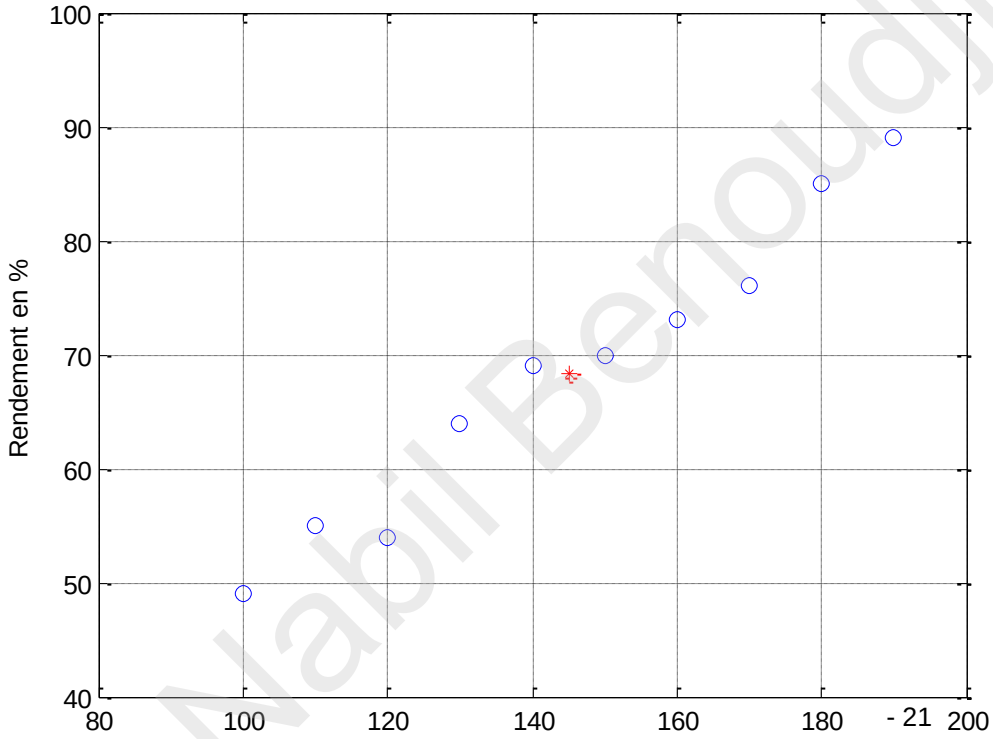
Exemple 1 Régression linéaire (1)

- L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit a donné les valeurs suivante pour la température x_i et le rendement correspondant y_i.

Température °C [100 110 120 130 140 150 160 170 180 190]

Rendement % [49 55 54 64 69 70 73 76 85 89]

La figure suivante représente le nuage de points pour ces données et suggère une relation linéaire

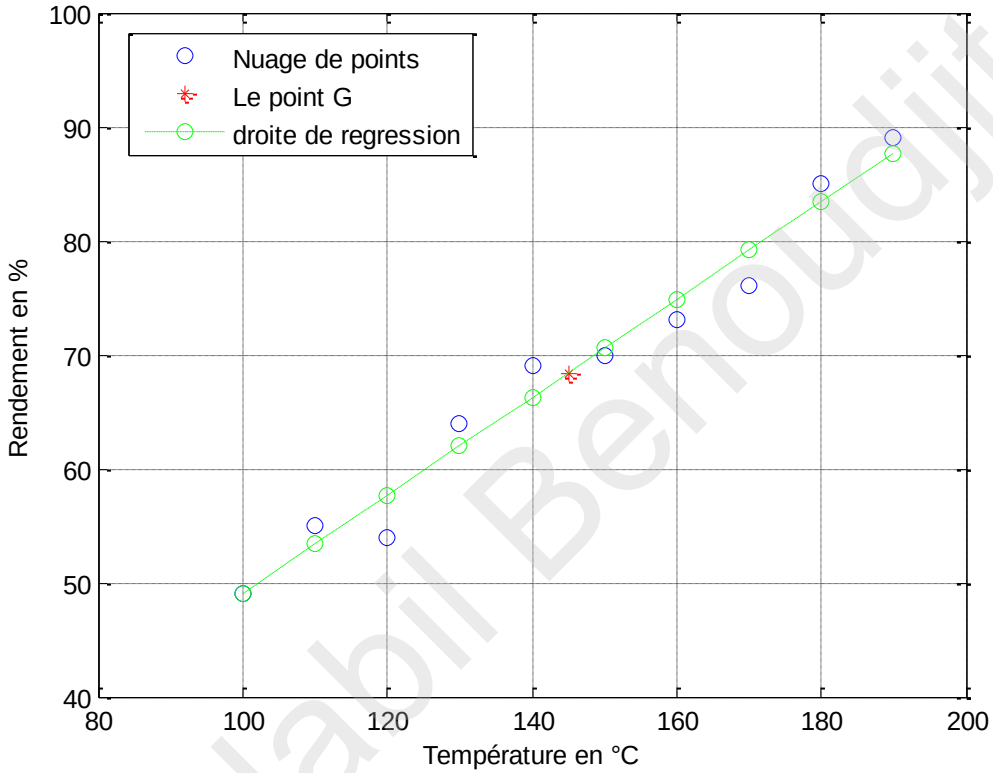


Exemple 1 Régression linéaire (2)

- Donc le modèle de la droite est : y = a *x + b
- En utilisant les deux équations ci-dessous on estime la meilleure droite de régression

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

$$b = \bar{y} - a * \bar{x}$$



Après calcul:
Y = a*x + b = 0,4291*x + 6,1818

Avec : R2 = 0,9710 et MSE = 4,5418

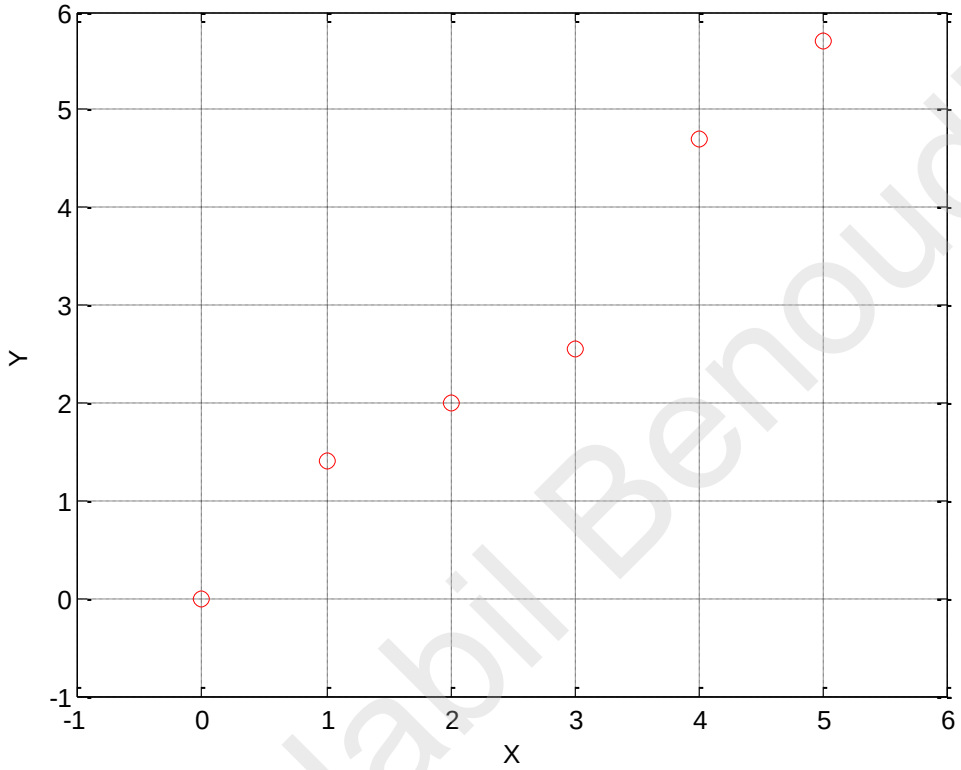
Exemple 2 Régression linéaire (1)

- Trouver la meilleure droite à travers les couples de points suivants:

X = [0 1 2 3 4 5]

Y= [0 1.4 2 2.55 4.7 5.7]

La figure suivante représente le nuage de points pour ces données et suggère une relation linéaire

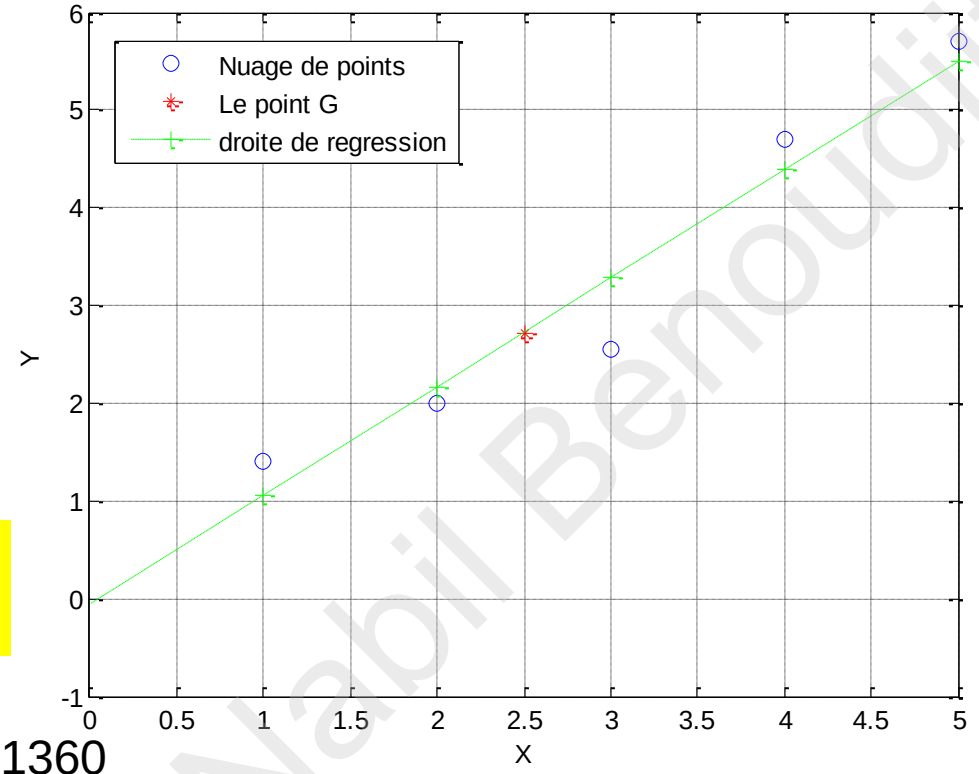


Exemple 2 Régression linéaire (2)

- Donc le modèle de la droite sous la forme matricielle est : y = A*X

$$A = (X^T X)^{-1} X^T Y$$

$$X^T X = \begin{matrix} 55 & 15 \\ 15 & 6 \end{matrix} \quad \text{et} \quad (X^T X)^{-1} = \begin{matrix} 0.0571 & -0.1429 \\ -0.1429 & 0.5238 \end{matrix} \quad \text{et} \quad X^T Y = \begin{matrix} 60.35 \\ 16.35 \end{matrix}$$



Après calcul:
Y = a*x + b = 1,1129*x – 0,0571

Avec : R2 = 0,96 et MSE = 0,1360

Régression Linéaire Multiple (MLR) (1)

- La régression linéaire multiple est une méthode d'analyse de données quantitatives. Elle a pour but de mettre en évidence la liaison pouvant exister entre une variable dite **expliquée**, que l'on notera **Y** et plusieurs autres variables dites **explicatives** que l'on notera X_1, X_2, ... , X_k.

- Les k variables X_i, i = 1, ... , k peuvent être soit aléatoires, soit contrôlées c'est-à-dire qu'elles sont connues **sans erreur**. Nous supposerons dans la suite que **les variables X_i, i = 1, ... , k sont contrôlées**. Nous nous intéressons aux modèles dit **linéaires**, c'est-à-dire aux modèles du type :

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_kX_k$$

Régression Linéaire Multiple (MLR) (2)

- dans lequel a_0, a_1, ... , a_k sont des réels appelés **coefficients du modèle.**

- Le modèle de la Régression Linéaire Multiple (MLR) sous la forme matricielle est sous la forme suivante:

$$y = X b + e$$

- L'estimation des coefficients du modèle inconnus constitué par le vecteur **b** est réalisée en minimisant une fonction coût, par exemple la somme des carrés résiduels

$$SS_{res} = \sum_{i=1}^{m} \left(\hat{y}_i - y_i \right)^2$$

Régression Linéaire Multiple (MLR) (3)

- Quand m = n+1, (m est le nombre d'échantillons (observations) et n est le nombre de variables)

$$\mathbf{X}^{-1}\mathbf{y} = \mathbf{X}^{-1}\mathbf{X}\mathbf{b},$$

$$\mathbf{b} = \mathbf{X}^{-1}\mathbf{y}.$$

- quand m > n+1

$$\mathbf{X}^{T}\mathbf{y} = \mathbf{X}^{T}\mathbf{X}\mathbf{b},$$

$$(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{X}\mathbf{b},$$

$$\mathbf{b} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y}.$$

Régression Linéaire Multiple (MLR) (4)

- Quand m < n+1

$$\mathbf{b} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}.$$

- NB: Souvent dans la pratique la matrice X comprend plus de variables que d'échantillons, alors la colinéarité est garantie.

Exemple 1: Régression linéaire Multiple (1)

- Trouver la régression linéaire multiple liant les variables explicatives x_1 et x_2 avec la variable expliquée y.

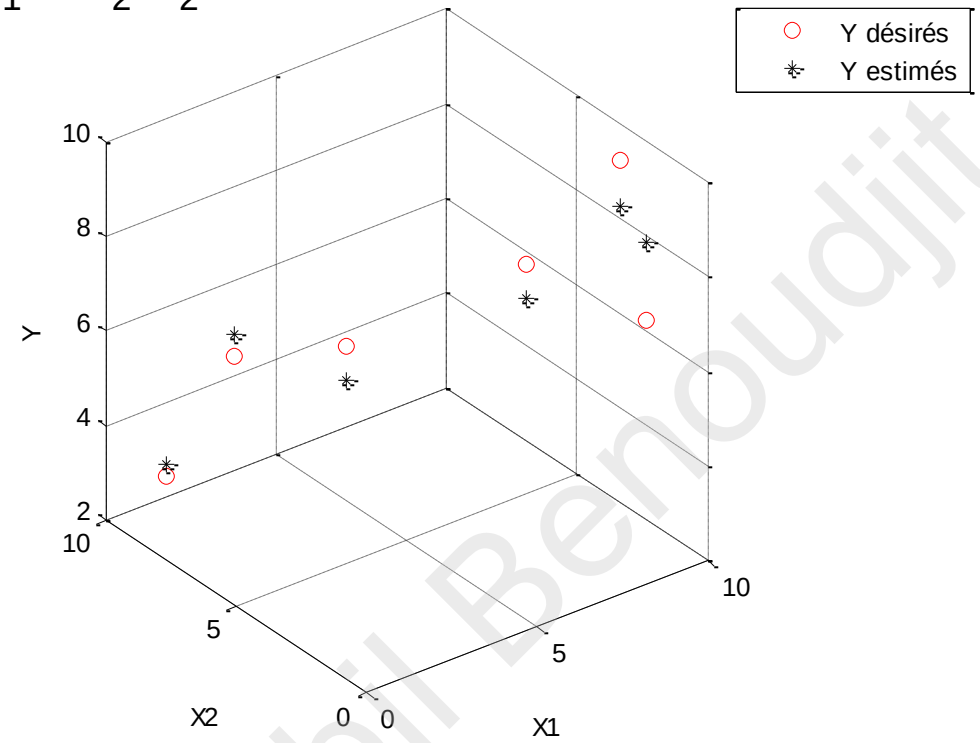
$$y = a_1*x_1 + a_2*x_2 + b$$

x_1 = [1 3 4 7 9 9]

x_2 = [9 9 6 3 1 2]

Y = [3 5 6 8 7 10]

[y-désiré	y-estimé	erreur]
3	3,2734	0,0748
5	5,4714	0,2222
6	5,2812	0,5166
8	7,2891	0,5054
7	8,6276	2,6491
10	9,0573	0,8887



Après calcul: R^2 = 0,8354

$Y = a_1*x_1 + a_2*x_2 + b = 1,0990*x_1 + 0,4297*x_2 – 1,6227$

Exemple 1: Régression linéaire Multiple (2)

- Les données sont centrés ===➔ b = 0
- Le modèle de la régression linéaire multiple liant les variables explicatives x_1 et x_2 avec la variable expliquée y.

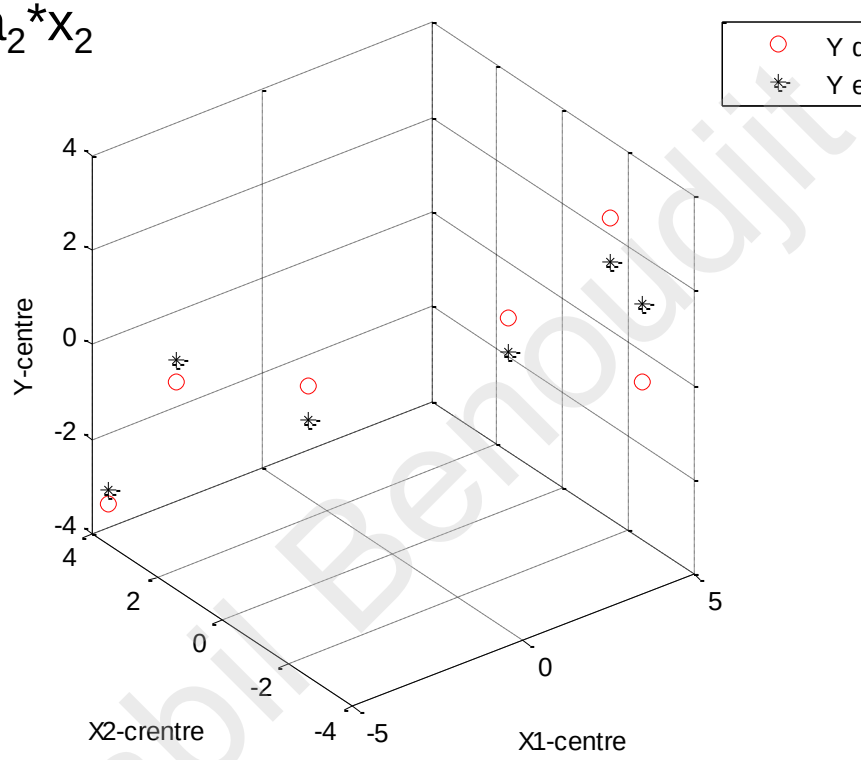
$$y = a_1*x_1 + a_2*x_2$$

x_1 = [-4,5 -2,5 -1,5 1,5 3,5 3,5]

x_2 = [4 4 1 -2 -4 -3]

y = [-3,5 -1,5 -0,5 1,5 0,5 3,5]

[y-désiré	y-estimé	erreur]
-3,5	-3,2266	0,0748
-1,5	-1,0286	0,2222
-0,5	-1,2187	0,5166
1,5	0,7891	0,5054
0,5	2,1276	2,6491
3,5	2,5573	0,8887



Après calcul: R^2 = 0,8354

$Y = a_1*x_1 + a_2*x_2 = 1,0990*x_1 + 0.4297*x_2$

Comparaison entre modèle linéaire et modèle non linéaire

- Nombre de paramètres fixe
- Nombre de paramètres Variable

- Petit nombre de paramètres
- Grand nombre de paramètres

- Apprentissage direct
- Apprentissage adaptatif

- Pas de minima locaux
- Présence de minima locaux

- Réservé aux problème linéaire
- Valide pour n'importe quel problème

Merci pour votre attention