

# Prétraitement et Exploration de données “Preprocessing”

**Onesime Mbulayi**

PhD Student  
University of Padova | Italy  
Data Scientist & Bioinformatics



# Machine learning Pipeline review



# Analyse exploratoire et pretraitement des données

Comprendre les données



Les données statistiques peuvent être divisées en **deux grandes catégories** : données **quantitatives** et données **qualitatives**.

Chacune de ces catégories a ses propres sous-types, caractéristiques, et contextes d'utilisation.

Les données quantitatives sont des données numériques qui mesurent la quantité ou l'ampleur d'une variable.

Les données quantitatives peuvent être subdivisées en deux types principaux :

**Données Continues:** Les données continues sont des valeurs numériques qui peuvent prendre une infinité de valeurs dans un intervalle donné.

Exemples :

- Température : 22.5°C, 36.7°C
- Taille : 165.5 cm, 180.2 cm
- Poids : 70.3 kg, 54.8 kg
- Temps : 4.76 secondes, 12.34 minutes

**Données Qualitatives:** Les données qualitatives, aussi appelées données catégorielles, décrivent des attributs ou des caractéristiques non numériques.

Les données qualitatives sont subdivisées en deux types principaux :

**Données Nominales :** Les données nominales sont des données qualitatives qui représentent des catégories sans ordre intrinsèque.

◦ **Exemples :**

- Couleur des cheveux : brun, blond, roux
- Genre : homme, femme, autre
- Type de produit : alimentaire, électronique, vestimentaire

- **Données Ordinales** : Les données ordinales sont des données qualitatives avec un ordre ou un rang significatif entre les catégories. Cependant, les écarts entre les catégories ne sont pas nécessairement égaux.

Exemples :

- Niveaux de satisfaction : satisfait, neutre, insatisfait
- Niveau d'éducation : primaire, secondaire, universitaire
- Classement de courses : premier, deuxième, troisième

# Autre types de données

Images & videos, audios ...



# Format de données

Les données peuvent être classifiées en trois grandes catégories selon leur organisation et leur format :

- les données structurées,
- semi-structurées
- non structurées.

Chacune de ces catégories présente des caractéristiques distinctes qui influencent la manière dont les données sont stockées, traitées et analysées.

**Données Structurées:** Les données structurées sont des données organisées de manière rigide et prévisible, généralement dans des formats tabulaires avec des lignes et des colonnes.

Exemple:

Nom	Age	Sexe	Taille /m	Poid /kg
KOMBA	25	M	1.80	72
Nzulu	54	F	1.52	65

## Données Semi-Structurées :

Les données semi-structurées ne suivent pas strictement un modèle de données fixe, mais elles contiennent des balises ou des marqueurs pour séparer les éléments et faciliter leur organisation.

Ces données ne sont pas aussi rigides que les données structurées mais possèdent néanmoins une certaine forme de structure.

Les formats comme XML et JSON sont couramment utilisés pour stocker et échanger des données semi-structurées.

## Données Semi-Structurées :

Exemple fichier JSON:

```
{ "building": "1007",  
  "coord": { "type": "Point",  
              "coordinates" : [-73.856077, 40.848447]  
            },  
  "street": "Morris Park Ave",  
  "zipcode": "10462"  
}
```

**Données Non Structurées** : Les données non structurées sont des données qui ne suivent aucun modèle fixe ou prévisible. Elles sont généralement stockées dans leur format brut et ne sont pas organisées selon un schéma ou une structure prédéfinie.

Exemples :

- Documents texte : rapports, articles, essais.
- Contenus multimédias : photos, vidéos, enregistrements audio.
- Publications sur les réseaux sociaux : tweets, publications Facebook, commentaires.

# Exploration et analyse univariée

## Traitement des valeurs manquantes

### 1. Suppression des données :

- Suppression des cas : Supprimer toutes les observations avec des valeurs manquantes.
  - Avantages : Simple à implémenter.
  - Inconvénients : Peut entraîner une perte importante de données

## Traitement des valeurs manquantes

**2. Suppression des variables :** Supprimer les variables avec un grand nombre de valeurs manquantes.

- Avantages : Évite les biais introduits par des valeurs manquantes systématiques.
- Inconvénients : Perte d'informations potentiellement importantes.

## Traitement des valeurs manquantes

### Imputation des données :

- Imputation par la moyenne/médiane/mode :
  - Remplacer les valeurs manquantes par la moyenne ( pour les variables quantitatives), la médiane ou la mode ( pour les variables qualitatives).
  - Avantages : Simple à mettre en œuvre et maintient la taille de l'ensemble de données.
  - Inconvénients : Réduit la variance des données et peut introduire un biais.



## Traitement des valeurs manquantes

### ◦ Imputation par régression :

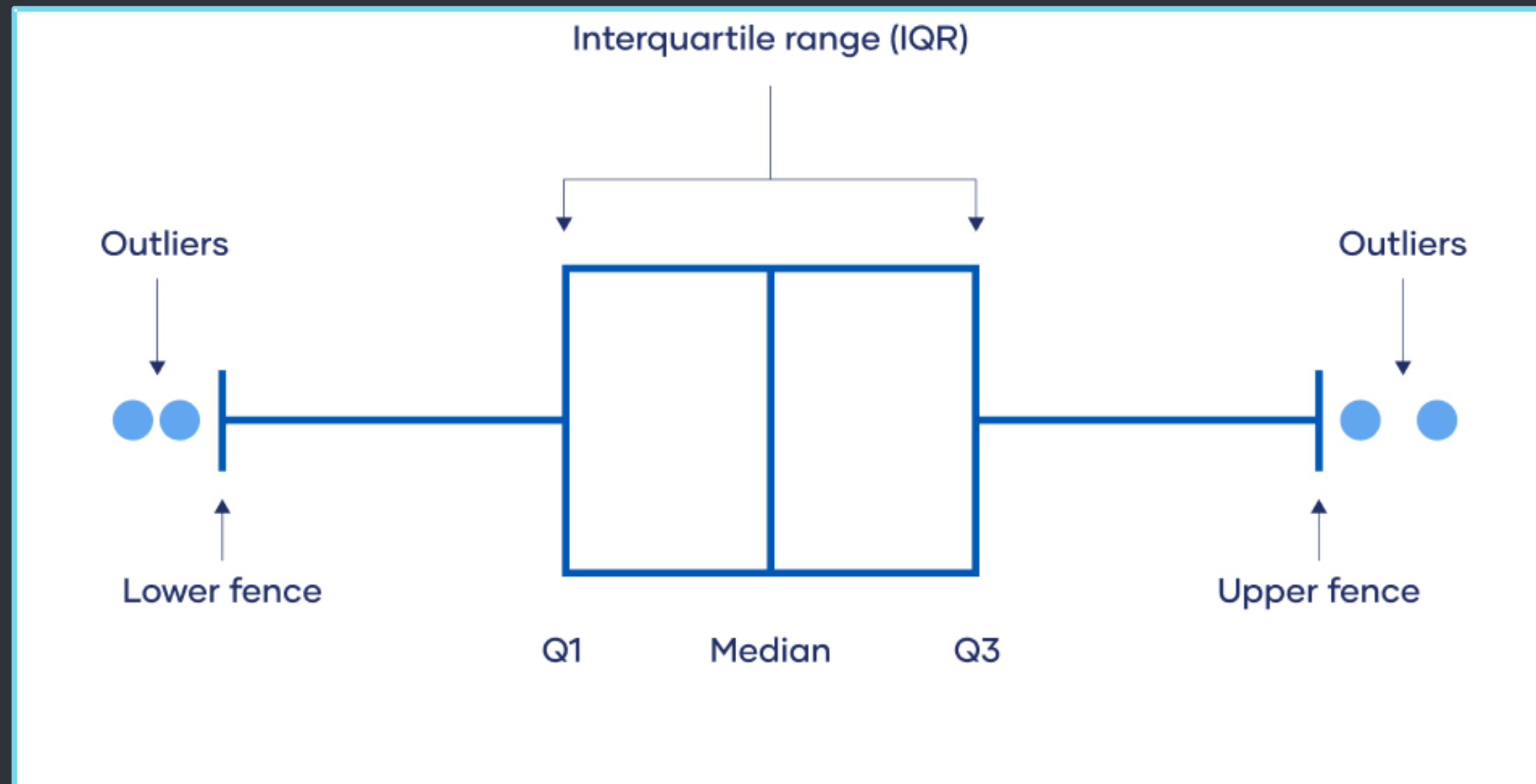
- Utiliser des modèles de régression pour prédire les valeurs manquantes en fonction des autres variables.
- Avantages : Peut capturer les relations entre les variables et fournir des imputations plus précises.
- Inconvénients : Complexe à implémenter et peut sous-estimer la variance des imputations.

## Les valeurs aberrantes:

Les valeurs aberrantes, également appelées outliers, sont des observations dans un ensemble de données qui s'écartent significativement des autres observations.

Ces valeurs peuvent indiquer des variations inhabituelles, des erreurs de mesure ou des cas exceptionnels, et elles peuvent avoir un impact important sur les analyses statistiques et les modèles de données.

## Les valeurs aberrantes:



## Mesure de dispersion:

Les mesures de dispersion sont des statistiques qui décrivent l'étendue et la variabilité d'un ensemble de données.

Elles sont essentielles pour comprendre comment les données se distribuent autour de la mesure de tendance centrale (comme la moyenne ou la médiane).

## Variance:

La variance mesure de la dispersion des données par rapport à la moyenne. Elle représente la moyenne des carrés des écarts par rapport à la moyenne.

Pour un ensemble de données  $\{x_1, x_2, \dots, x_n\}$  avec une moyenne  $\bar{x}$  tel que

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

la variance  $\sigma^2$  est donnée par :

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

**Variance:**

Exemple :

Index	Valeurs
1	3
2	5
3	7
4	8
5	10

La moyenne des données est calculée comme suit :

$$\bar{x} = \frac{3 + 5 + 7 + 8 + 10}{5} = \frac{33}{5} = 6.6$$

Calcul de la variance :

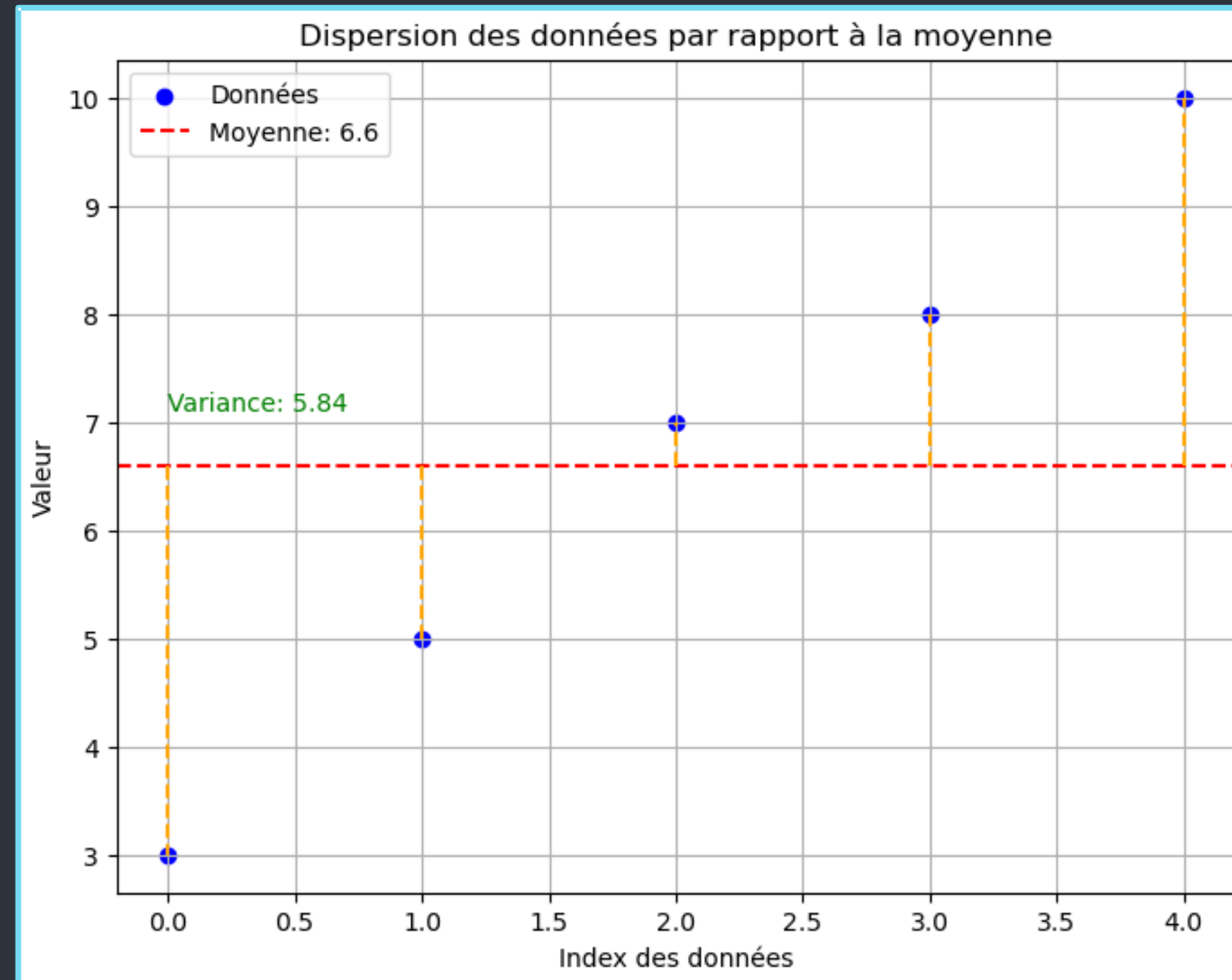
$$\begin{aligned} \sigma^2 &= \frac{(3 - 6.6)^2 + (5 - 6.6)^2 + (7 - 6.6)^2 + (8 - 6.6)^2 + (10 - 6.6)^2}{5} \\ &= \frac{12.96 + 2.56 + 0.16 + 1.96 + 11.56}{5} = \frac{29.2}{5} = 5.84 \end{aligned}$$

$$\sigma^2 = 5.84$$

## Variance:

Exemple :

Index	Valeurs
1	3
2	5
3	7
4	8
5	10



## Écart-type (Standard Deviation):

Racine carrée de la variance, l'écart-type ramène la mesure de la dispersion aux mêmes unités que les données initiales.

Il est obtenu par :

$$\sigma = \sqrt{\sigma^2} \quad (3)$$

$$\sigma = \sqrt{5.84} = 2.41$$



## Mesure de tendance centrales

Les mesures de tendance centrale sont des valeurs qui représentent ou résument un ensemble de données en identifiant le point central ou typique de la distribution.

Elles fournissent des informations sur la valeur autour de laquelle les autres valeurs de l'ensemble de données ont tendance à se regrouper.

## Moyenne (ou moyenne arithmétique)

La moyenne est la somme de toutes les valeurs divisée par le nombre de valeurs.

Elle est la mesure de tendance centrale la plus utilisée, surtout lorsqu'il n'y a pas de valeurs aberrantes qui pourraient la fausser.

Voir l'équation (1)

## Médiane

La médiane est la valeur qui sépare la moitié supérieure des valeurs de la moitié inférieure dans un ensemble de données ordonné.

C'est une mesure particulièrement utile lorsqu'il y a des valeurs aberrantes, car elle n'est pas influencée par celles-ci comme la moyenne.

◦ Comment la trouver :

- Pour un nombre impair de valeurs : La médiane est la valeur du milieu une fois les données triées.
- Pour un nombre pair de valeurs : La médiane est la moyenne des deux valeurs du milieu.

## Médiane

◦ Exemple :

- Pour un ensemble de données impair : [3,5,7,8,10][3, 5, 7, 8, 10][3,5,7,8,10], les données triées sont déjà ordonnées, donc la médiane est 7.
- Pour un ensemble de données pair : [3,5,7,8,10,12][3, 5, 7, 8, 10, 12][3,5,7,8,10,12], les valeurs centrales sont 7 et 8, donc la médiane est :

$$\text{Médiane} = \frac{7 + 8}{2} = 7.5$$

## Mode

Le mode est la valeur qui apparaît le plus fréquemment dans un ensemble de données.

Un ensemble de données peut avoir un mode, plusieurs modes (bimodal, multimodal) ou aucun mode si toutes les valeurs sont uniques.

- Exemple :

- Pour un ensemble de données  $[3, 5, 7, 8, 10, 8, 8]$ , le mode est 8, car il apparaît trois fois, plus souvent que les autres valeurs.

# Analyse multivariée

L'analyse multivariée est une branche des statistiques qui examine et interprète des données impliquant plusieurs variables simultanément.

Elle englobe diverses techniques pour comprendre les relations entre plusieurs variables et comment ces variables interagissent entre elles.

## La corrélation linéaire

La corrélation linéaire est une mesure statistique qui indique la force et la direction d'une relation linéaire entre deux variables.

Elle évalue dans quelle mesure deux variables quantitatives sont linéairement liées.

En d'autres termes, elle mesure à quel point les points sur un diagramme de dispersion se rapprochent d'une droite.

## Coefficient de Corrélation de Pearson (r)

Le coefficient de corrélation linéaire le plus courant est le coefficient de corrélation de Pearson.

Il varie entre **-1** et **1** et est calculé à l'aide de la formule suivante :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (4)$$



## Coefficient de Corrélation de Pearson ( $r$ )

### Interprétation du Coefficient de Corrélation

- $r=1$ : Corrélation linéaire positive parfaite (lorsque  $X$  augmente,  $Y$  augmente de manière proportionnelle).
- $r=-1$ : Corrélation linéaire négative parfaite (lorsque  $X$  augmente,  $Y$  diminue de manière proportionnelle).
- $r=0$ : Aucune corrélation linéaire
- $0 < r < 1$ : Corrélation positive (les deux variables augmentent ensemble).
- $-1 < r < 0$ : Corrélation négative (une variable augmente tandis que l'autre diminue).

# Coefficient de Corrélation de Pearson (r)

Exemple : Tableau de Données Météorologiques

Jour	Humidité (%)	Température (°C)	Pluie (mm)	Vent (km/h)	Chaleur (°C)	Interprétation de r
1	80	25	10	5	28	r=1 : Corrélation linéaire positive parfaite (Température ↔ Chaleur)
2	70	20	0	15	25	r=0.8 : Corrélation positive (Humidité ↔ Pluie)
3	60	22	5	20	27	r=0 : Aucune corrélation linéaire (Vent ↔ Température)
4	40	30	0	25	32	r=-0.7 : Corrélation négative (Température ↔ Humidité)
5	20	35	0	30	35	r=-1 : Corrélation linéaire négative parfaite (Vent ↔ Pluie)

## Matrice de corrélation

Une matrice de corrélation est un tableau rectangulaire qui affiche les coefficients de corrélation entre plusieurs variables.

Chaque cellule de la matrice représente le coefficient de corrélation entre deux variables spécifiques.

Cette matrice est utile pour comprendre la force et la direction des relations linéaires entre plusieurs variables en un seul coup d'œil.

## Matrice de corrélation

Si on a trois variables  $X_1, X_2$ , et  $X_3$ , la matrice de corrélation  $r$  serait :

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

Ici,  $r_{ij}$  représente le coefficient de corrélation entre  $X_i$  et  $X_j$ .

# Normalisation et Standardisation

La normalisation et la standardisation sont deux techniques couramment utilisées en statistique et en apprentissage automatique pour préparer les données avant de les utiliser dans des modèles.

Elles sont particulièrement utiles lorsque les caractéristiques (features) des données ont des échelles différentes.

## 1. Normalisation

La normalisation, aussi appelée mise à l'échelle min-max (Min-Max Scaling), consiste à redimensionner les valeurs des caractéristiques pour qu'elles se situent dans une plage spécifique, généralement  $[0,1]$  ou  $[-1,1]$ .

Pour une valeur  $x_i$  dans un ensemble de données la normalisation est donnée par :

$$\hat{x} = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (5)$$

$x_i$  : La valeur initiale.

$\hat{x}$  : La valeur normalisée.

$\min(X)$  : La valeur minimale de la caractéristique.

$\max(X)$  : La valeur maximale de la caractéristique.

## Quand utiliser la normalisation ?

Lorsque les caractéristiques des données ont des plages différentes et que vous voulez les ramener à une échelle commune.

Souvent utilisée dans les algorithmes d'apprentissage automatique qui ne font pas d'hypothèses sur la distribution des données, comme les réseaux de neurones ou les méthodes de plus proches voisins (KNN).

## 2. Standardisation

La standardisation consiste à recentrer les données pour qu'elles aient une moyenne de 0 et une variance (ou écart-type) de 1. Cela transforme les données en une distribution normale centrée.

Pour une valeur  $x_i$ , la standardisation s'obtient par:

$$z = \frac{x_i - \mu}{\sigma} \quad (6)$$

$x_i$  : La valeur initiale.

$\mu$  : La moyenne de la caractéristique.

$\sigma$  : L'écart-type de la caractéristique.

$z$  : La valeur standardisée.



## Quand utiliser la standardisation ?

Lorsque les données suivent une distribution normale (ou sont supposées le faire).

Utile dans les algorithmes qui supposent ou sont sensibles à la distribution des données, comme la régression linéaire, les modèles de machine à vecteurs de support (SVM), et les algorithmes de clustering comme K-means.

## Création et la sélection de caractéristiques (features)

La création et la sélection de caractéristiques (features) sont des étapes cruciales dans le pipeline de machine learning.

Elles influencent directement la performance et la précision des modèles.

## 1. Création de Caractéristiques

La création de caractéristiques implique la transformation des données brutes en un format plus utile pour les modèles de machine learning.

### MÉTHODES COURANTES :

- **Extraction de Caractéristiques** : Créez des caractéristiques à partir de données existantes en utilisant des méthodes statistiques ou des
- **Encodage des Variables Catégorielles** : Transformez les variables catégorielles en numériques. (One-hot encoding, Label encoding, Ordinal encoding)
- **Création de Nouvelles Variables** : Combinez ou transformez les variables existantes. Exemples : Interaction entre variables, polynômes, logarithmes, etc.

## 2. Sélection de Caractéristiques

La sélection de caractéristiques consiste à choisir les variables les plus pertinentes pour le modèle, afin de réduire la complexité et améliorer la performance.

- **Régression Lasso** : Utilise la régularisation L1 pour effectuer la sélection de caractéristiques en réduisant les poids des caractéristiques non importantes à zéro.
- **Tests Statistiques** : Utilisez des tests statistiques (comme le test de chi-carré pour les variables catégorielles) pour évaluer l'importance des caractéristiques.
- **Corrélation** : Sélectionnez les caractéristiques en fonction de leur corrélation avec la variable cible.
- **Analyse en Composantes Principales (PCA)** : Réduit la dimensionnalité des données tout en préservant la variance.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)** : Réduit la dimensionnalité pour les visualisations en préservant les relations locales.

# **‘Apprentissage Supervisé : Regression**

# Apprentissage Supervisé

- Est une branche de l' apprentissage automatique (machine learning) qui consiste à **entraîner un modèle** à partir d'un **ensemble de données étiquetées**. Ces données contiennent à la fois des entrées (les caractéristiques) et des sorties (les résultats attendus).
- Cela signifie que chaque donnée d'entrée est associée à une sortie correcte (**appelée étiquette ou label**). Le modèle apprend à associer les entrées aux sorties correspondantes, afin de pouvoir ensuite faire des prédictions sur de nouvelles données.

# Branches de l'apprentissage supervisé

**1. Classification** : Le but est de prédire une catégorie ou une classe à partir de données d'entrée.

Exemples :

- Prédire si un e-mail est un spam ou non.
- Classifier des images d'animaux en chien, chat, etc.

**2. Régression** : Ici, l'objectif est de prédire une valeur continue en fonction des données d'entrée.

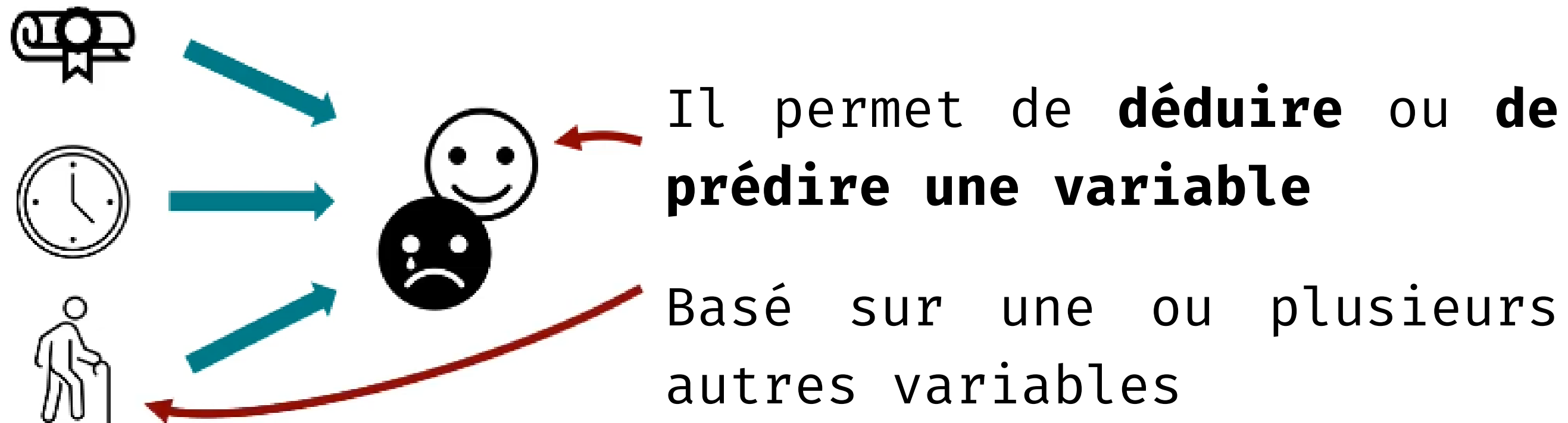
Exemples :

- Prédire le prix d'une maison en fonction de ses caractéristiques (taille, emplacement, etc.).
- Estimer les ventes futures d'un produit.

Dans ces deux cas, le modèle apprend en ajustant ses paramètres pour minimiser l'écart entre ses prédictions et les étiquettes réelles fournies dans les données d'entraînement.

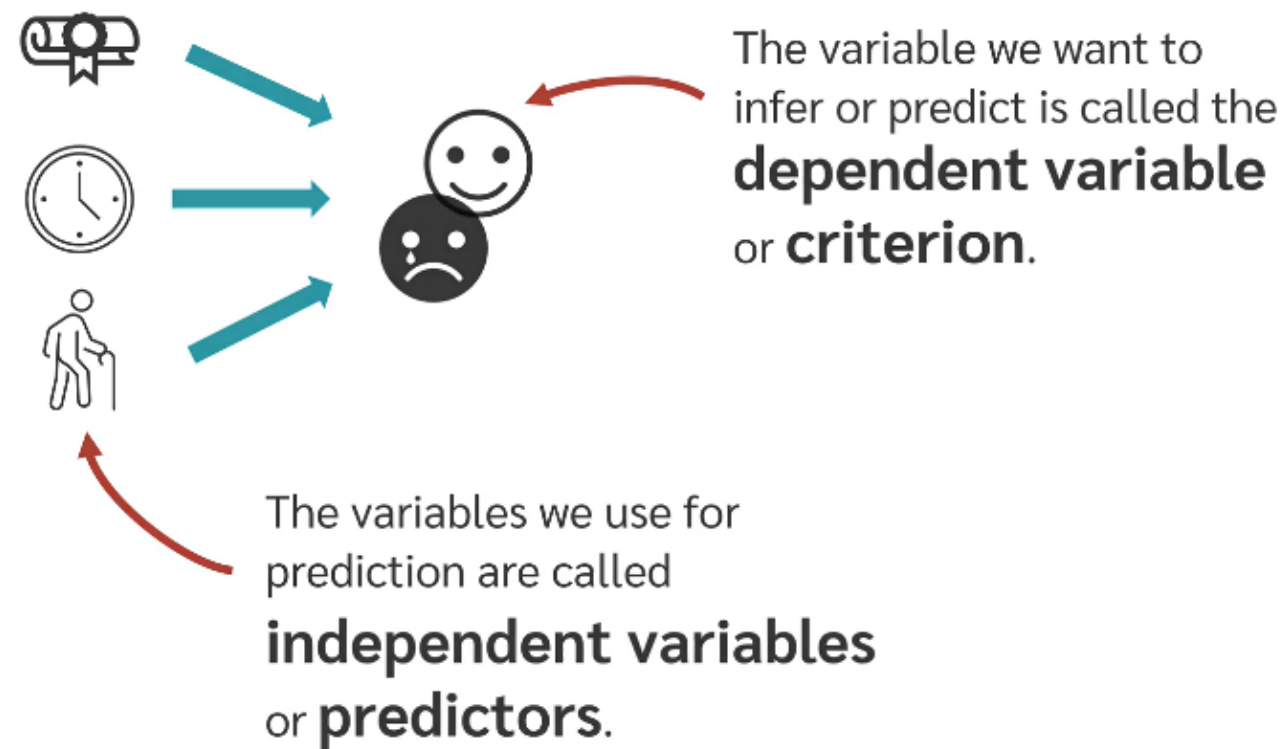
# Qu'est ce que la Regression ?

Une analyse de régression est une méthode de modélisation des relations entre les variables.

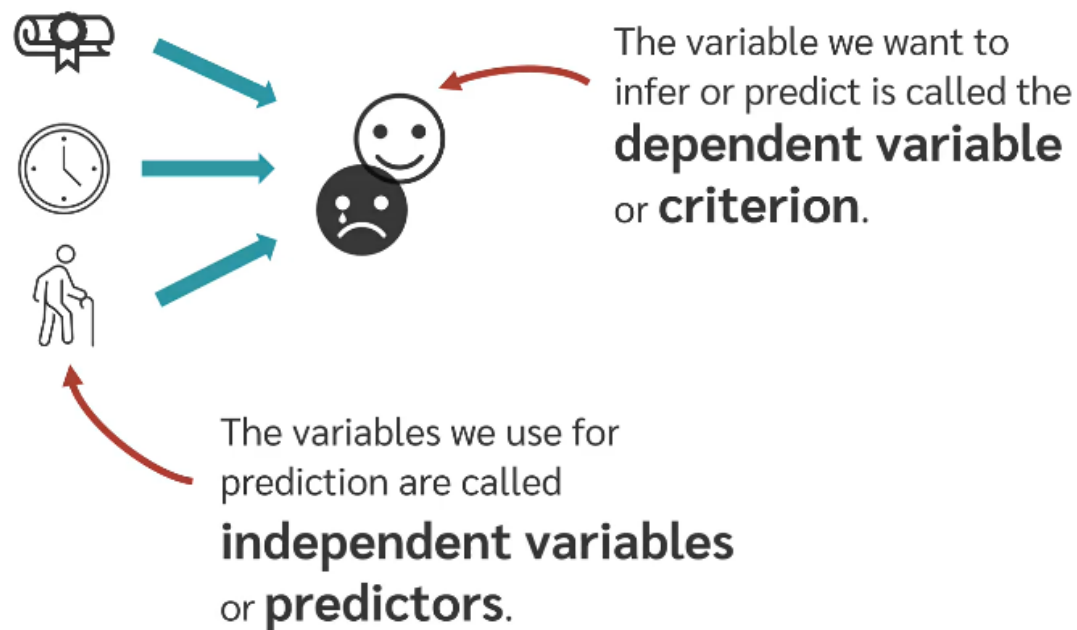




# Qu'est ce que Regression ?



La variable que nous voulons déduire ou prédire est appelée **variable dépendante** ou **critère**.



Les variables que nous utilisons pour la prédiction sont appelées **variables indépendantes** et **prédicteurs**.

# Régression linéaire (Prédiction manuelle)

Considérons un exemple où les données de ventes sur cinq semaines (en milliers) sont présentées sous forme de tableaux.

$x_i$ ( Semaine)	$y_i$ (Vente en dollard)
1	1200
2	1800
3	2600
4	3200
5	3800

Appliquer la technique de régression linéaire pour prédire les ventes des 7e et 12e semaines

# Régression linéaire (Prédiction manuelle)

L'équation de régression linéaire est donnée par

$$y = a_0 + a_1 * x + e$$

où

$$a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$$

$$a_0 = \bar{y} - a_1 * \bar{x}$$

	<b>Xi (Semaine)</b>	<b>yi( Ventes)</b>	<b>xi ^2</b>	<b>Xi * Yi</b>
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
<b>Somme</b>	<b>15</b>	<b>12.6</b>	<b>55</b>	<b>44.4</b>
<b>Moyenne</b>	<b>moy (x) =3</b>	<b>moy(y) = 2.52</b>	<b>moy(x^2) = 11</b>	<b>moyenne (x*y) = 8.88</b>

- $\bar{x} = 3$        $\bar{y} = 2.52$        $\overline{x^2} = 11$        $\overline{xy} = 8.88$

- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2} = \frac{8.88 - 3 * 2.52}{11 - 3^2} = 0.66$

- $a_0 = \bar{y} - a_1 * \bar{x} = 2.52 - 0.66 * 3 = 0.54$

L'équation de régression est

- $y = a_0 + a_1 * x$

- $y = 0.54 + 0.66 * x$

La vente prévue pour la 7e semaine (quand  $x = 7$ ) est

- **$y = 0.54 + 0.66 \times 7 = 5.16$**

La vente prévue pour la 12e semaine (quand  $x = 12$ ) est

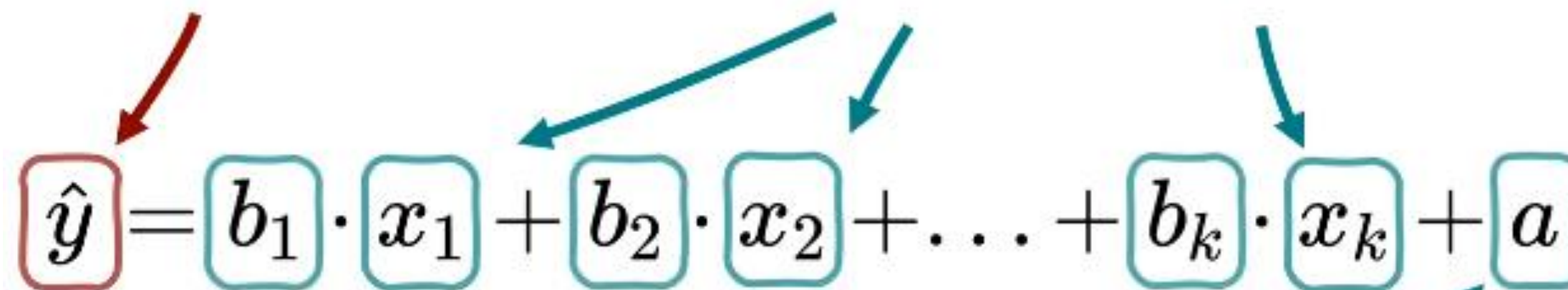
**$$y = 0.54 + 0.66 \times 12 = 8.46$$**

# Régression linéaire Multiple

Dans la régression linéaire, voici notre équation de régression

Nous avons la **variable dépendante**

Les **variables indépendante**



The diagram shows the multiple linear regression equation:  $\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_k \cdot x_k + a$ . The terms are enclosed in rounded rectangular boxes. A red arrow points from the text 'variable dépendante' to the  $\hat{y}$  box. A teal arrow points from the text 'Les variables indépendante' to the  $x_1$  box. Another teal arrow points from the same text to the  $x_k$  box. A teal arrow points from the text 'Et les coefficients de régression' to the  $b_1$  box. A final teal arrow points from the same text to the  $a$  box.

Et les **coefficients de régression**