Chapitre II : Analyse Prédictive : (Prédire une quantité : Régression)

II- 1- Régression Linéaire: Simple et Multiple

PLAQUETTE COMMERCIALE

Régression Linéaire

Définition :

La **Régression linéaire** : est une méthode statistique utilisée pour :

- modéliser la relation entre une variable dépendante expliquée continue et une ou plusieurs variables indépendantes explicatives.
- L'objectif est de déterminer l'équation d'une droite (ou d'un hyperplan, dans le cas de plusieurs variables) qui représente au mieux la relation entre les variables.

Régression Linéaire

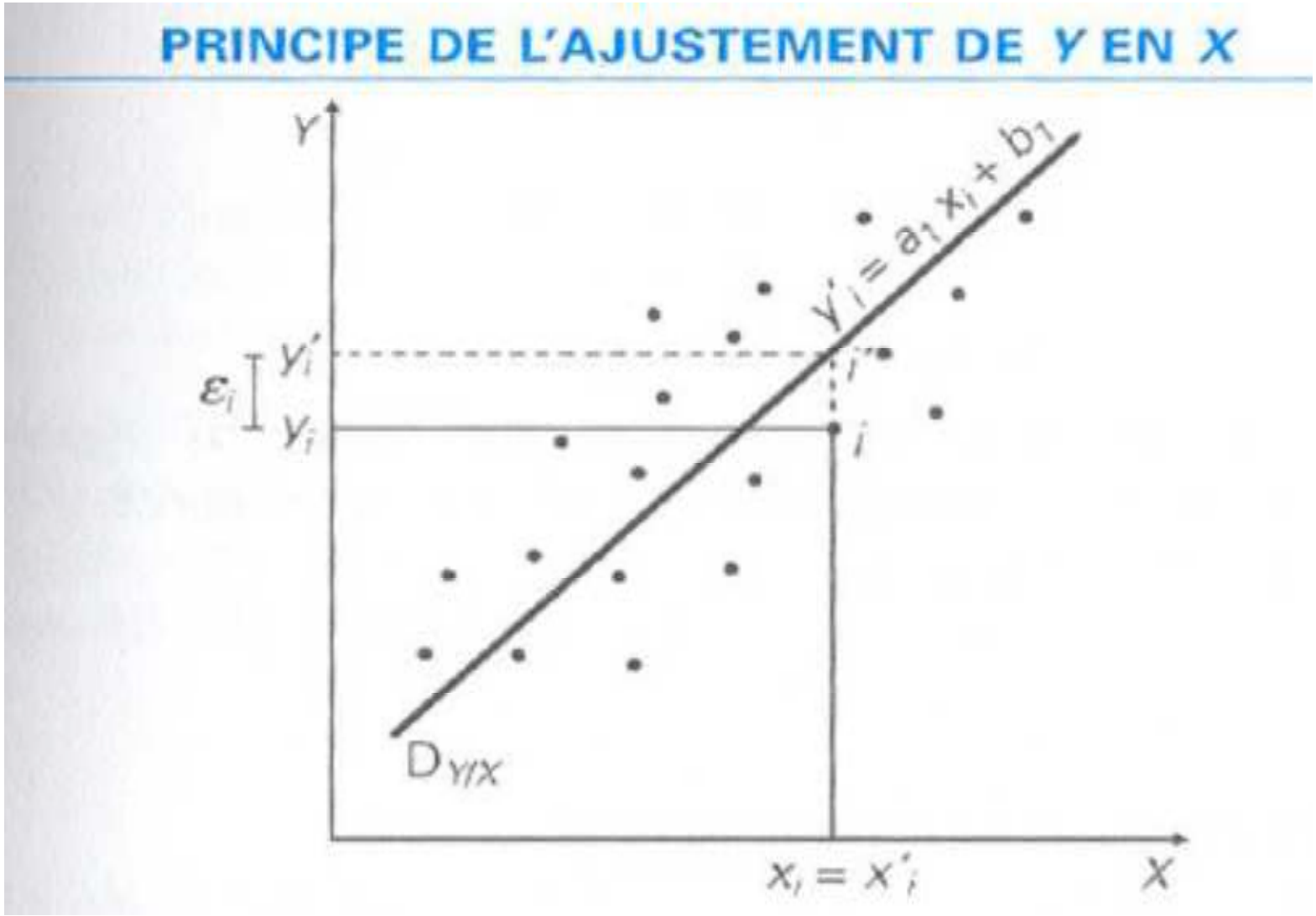
* Lorsqu'il existe une seule variable d'entrée (x), ➔ la méthode est appelée régression linéaire simple.

* Lorsqu'il y a plusieurs variables d'entrée, ➔ la méthode est appelée régression linéaire multiple.

I- La R**égression Linéaire Simple d'un Point de vue Statistique:**

-Il ya deux variables quantitatives X,Y.
-X variable explicative; Y Expliquée
➔**hypothèse** = les données proviennent d'un phénomène qui à la forme d'une **droite**.
-Il existe une relation linéaire entre l' entrée X et la sortie Y

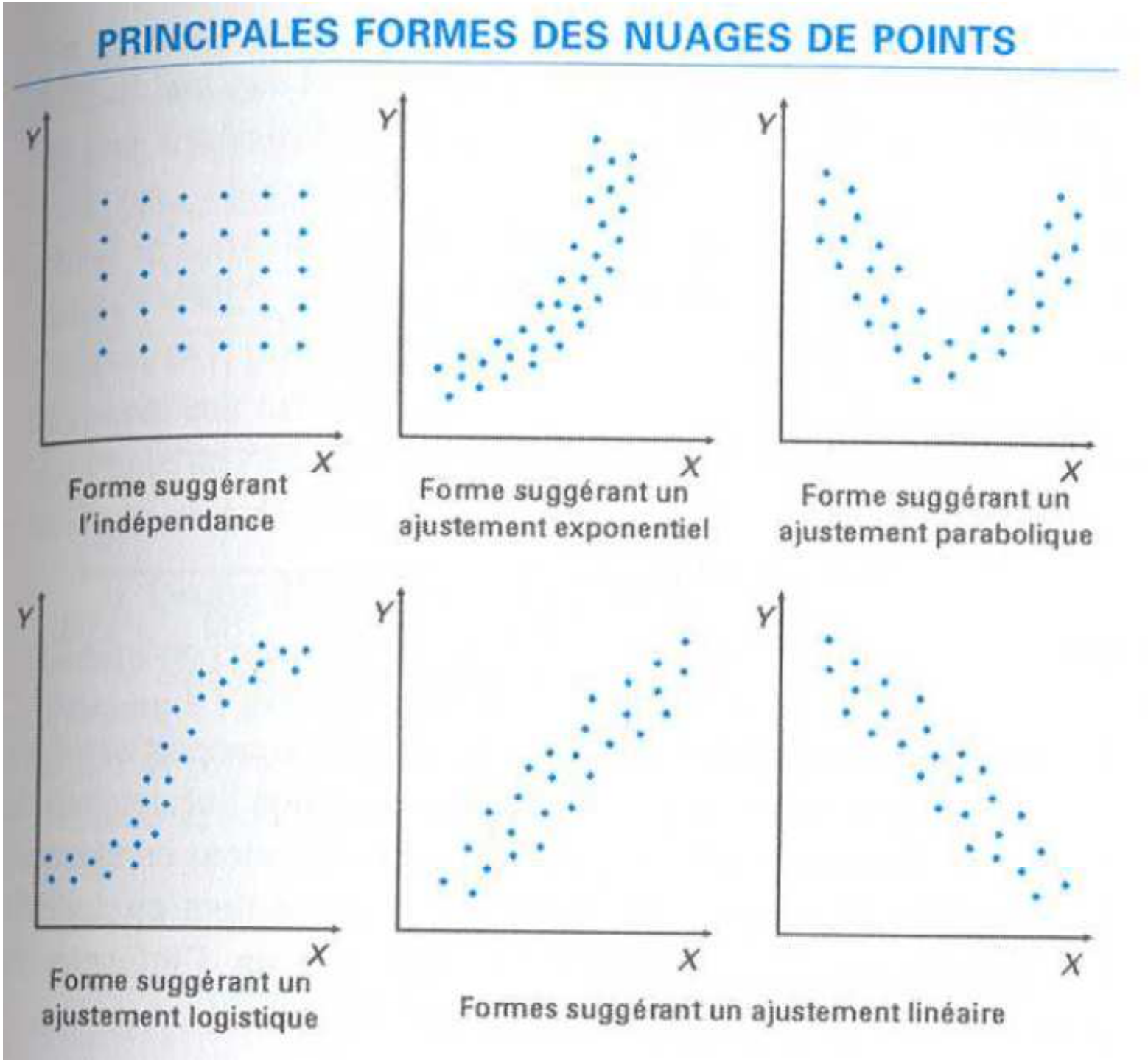
Régression Linéaire Simple



➔Un **graphique de corrélation permet de vérifier rapidement l'existence d'un lien.**

➔La forme du **nuage de points** obtenus détermine la nature de la Liaison statistique entre deux variables.

PRINCIPALES FORMES DES NUAGES DE POINTS



Objectifs de la régression linéaire:

✓**Prédiction :** Elle permet de prédire la valeur de la variable dépendante en fonction des valeurs des variables indépendantes.

Ex. prédire les ventes en fonction des dépenses publicitaires.

✓**Identification des variables importantes :** Elle permet d'identifier les variables qui sont les plus importantes pour expliquer le phénomène étudié.

✓……

Méthodologie:

Il faut:

- Estimer les coefficients a et b de la droite $Y=ax + b$
- de manière à minimiser la somme des carrés des résidus $\sum(ei)^2$ (c'est-à-dire les différences entre les valeurs prédites et les valeurs réelles), ce qui est souvent appelé la Méthode des Moindres Carrés Ordinaire: **MMCO**.

Principe de MMCO:

➢On calcule la distance entre chaque point des données et la ligne de régression.

-Cette opération engendre **des résidus ei par rapport à Y ,** Les valeurs de x restent inchangées ;

➢Nous calculons la distance et la somme de toutes les erreurs au carré: $\sum(ei)^2$

➢ Minimiser la somme des erreurs :

$$\textbf{Min } (\sum(ei)^2)$$

Régression linéaire Apprentissage du modèle

1. **Régression linéaire simple**

Avec une régression linéaire simple lorsque nous avons une seule entrée, nous pouvons utiliser des statistiques pour estimer les coefficients.

(moyenne, écart type, corrélations et covariance.

C'est un exercice amusant, mais pas vraiment utile dans la pratique.

Régression linéaire Apprentissage du modèle

$$a_1 = \frac{COV_{xy}}{V_x} = \frac{1/n \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{1/n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\bar{y} = a_1 \bar{x} + b_1 \Leftrightarrow b_1 = \bar{y} - a_1 \bar{x}$$

Régression linéaire Apprentissage du modèle

2. **Régression Linéaire Multiple**:

-Lorsque nous avons plus d'une entrée, nous pouvons utiliser les moindres carrés ordinaire MMCO pour estimer les valeurs des coefficients.

-Minimiser la somme des résidus au carré. $\sum(ei)^2$

Problématique :

-Expliquer la variable y à l'aide d'une combinaison de plusieurs variables xi soient : x1 ,x2 ,x3 ,…..xk. On parle alors de régression linéaire multiple tel que :

$$Y = \sum_{i=1}^{k} X_i a_i + b + e$$

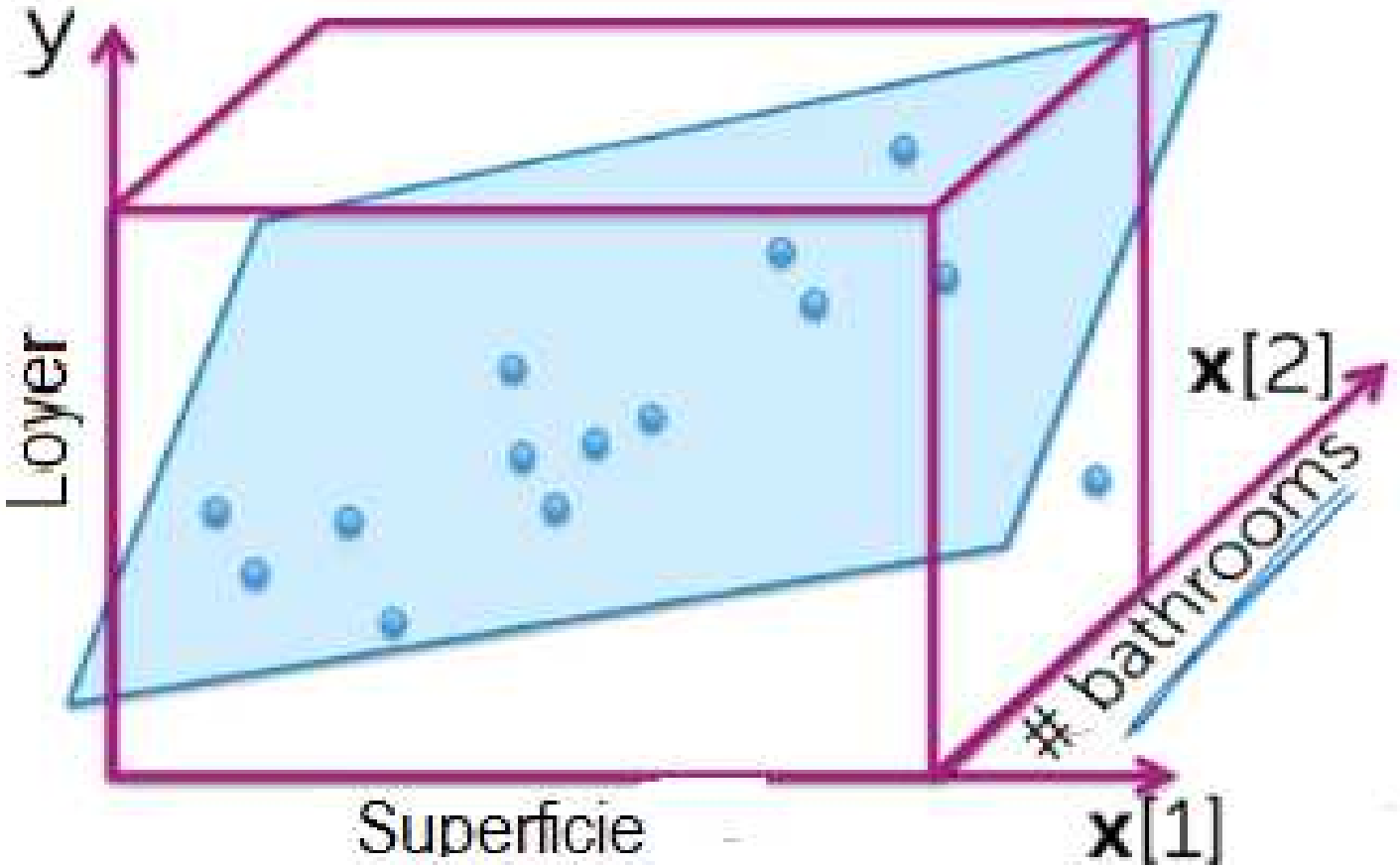
e : erreur résiduelle.

ai : coefficient de regression à estimer à partir des n observations.

b: terme constant.

Exemple :

Si on est en présence de trois variables y, x1 , x2 expliquer y par x1 , x2 c'est essayer de trouver les coefficients de l'équation : y= a1x1 + a2x2 + b …….équation d'un plan de degré 2 .



Formulation du problème :

-y = f(x1,x2,…..xk)

-y : variable expliquée

-x1,x2,x3,……xk ,

xi : variables explicatives

- $Y = a_1 x_{1+} a_2 x_{2+} a_3 x_{3+} \ldots . a_k x_k + b + e$

forme linéaire

-Nous avons n observations de la variable y

-Il s'agit de résoudre le système suivant :

$$y_1=a_1x_{11} + a_2x_{12} + a_3x_{13} +\ldots\ldots a_kx_{1k} + b + e_1$$
$$y_2=a_1x_{21} + a_2x_{22} + a_3x_{23} +\ldots\ldots a_kx_{2k} + b + e_2$$
$$..$$
$$..$$
$$y_n=a_1x_{n1} + a_2x_{n2} + a_3x_{n3} +\ldots\ldots a_kx_{nk} + b + e_n$$

FORME (1)

-On dispose alors de n équations et (k+1+n) inconnus,

-il existe une infinité de solutions pour ce système(car pas de contraintes sur les inconnus).

-La solution qui nous intéresse est celle qui minimise la somme des erreurs au carrés.

-On choisit la méthode des moindres carrés **MMCO**

Représentation du modèle de régression linéaire

Information:

Il est courant de parler de la **complexité** d'un modèle de régression comme la régression linéaire .

Elle est égale :

Complexité = **Nombre de coefficients** utilisés dans le modèle.

3. **Estimation des coefficients «sans isolation du terme constant»**

La forme (1) peut être écrite sous une forme matricielle comme suit :

$$Y = Xa + e$$

$$\rightarrow e = Y - Xa$$

3. **Estimation des coefficients «sans isolation du terme constant»**

Min $\sum ei^2 = ?$

$\varphi = \sum ei^2 = e^t e.$

On détermine ‘a’ vecteur de dimensions (k+1,1) qui minimise $\varphi = e^t e$

$$e = y - xa$$

$$\varphi = (y^t - a^t x^t)(y - xa)$$

Définition :

un Minimum d'une fonction de plusieurs variables ne peut se produire qu'en un point ou les dérivées partielles par rapport a ses inconnus s'annulent.

φ est une <u>fonction à plusieurs inconnus,</u> une condition nécessaire d'extremum est l'annulation des dérivées partielles (Gradients).

Min(φ) = ?

$$\frac{\partial \varphi}{\partial a} = 0 = -2x^t y + 2$$

-Résultat Exact
-L'inverse d'une matrice prend bcp de temps

➔ $x^t y = x^t x a$ ➔ $a = (x^t x)^{-1} x^t y$

Trouver une Solution Approximative

Prédiction en utilisant le modèle trouvé :

Ce modèle mathématique peut alors être utilisé pour prédire une valeur y en fonction des valeurs xi.

Exemple: estimer le loyer d'un logement à condition de ne pas trop s'éloigner des valeurs déjà observées.

-Régression linéaire Développée dans le domaine de la statistique,

-Mais Empruntée par l'apprentissage machine.

-C'est à la fois :

- un algorithme statistique et
- un algorithme d'apprentissage par machine (Machine Learning.)

- **Remarques** :

- SI on a beaucoup de données alors,
  Calcul solution Exacte très long
(Inverse matrice, n'est pas gratuit en temps calcul !).

ou

  Pas de solution pour dérivé=0

➔ Calcul d'une Approximation de la solution

➔ Algorithme ***Descente de Gradient***

3. Descente de Gradient

Lorsqu'il y a une ou plusieurs entrées, vous pouvez utiliser un processus d'optimisation des valeurs des coefficients en réduisant de manière itérative l'erreur du modèle.

Cette opération s'appelle Gradient Descent et commence par des valeurs aléatoires pour chaque coefficient.

Régression linéaire Apprentissage du modèle

-Un taux d'apprentissage (le pas) est utilisé comme facteur d'échelle et les coefficients sont mis à jour dans le sens d'une minimisation de l'erreur.

-Le processus est répété jusqu'à ce qu'une erreur de somme au carré soit atteinte ou qu'aucune amélioration supplémentaire ne soit possible.

Algorithme descent de Gradient

Début

Iteration t=1

η : le pas (stepsize)

W(t)=0 (ou Rand)

While not converged (ou t<=100)

W(t+1)🡐w(t) - η(dg(w)/dw)

T🡐t+1

Fin

Problème du Surajustement (overfitting)

Ce problème se produit lorsqu'un modèle est trop complexe par rapport à la quantité (petite) de données disponible pour son entraînement.

✓Cela signifie que le modèle peut bien fonctionner sur les données d'entraînement,

✓ mais il ne généralise pas bien sur de nouvelles données qu'il n'a pas encore vu.

➢Solution on utilise la **régularisation**.

Régularisation

Il existe des extensions de la formation du modèle linéaire appelées méthodes de régularisation.

Celles-ci cherchent à la fois à

- **minimiser** la somme de l'erreur au carré
- + à **réduire** la complexité du modèle .

Régularisation

Deux exemples populaires de procédures de régularisation pour la régression linéaire:

- **Lasso Regression** L1 : les moindres carrés ordinaires sont modifiés pour minimiser également la somme absolue des coefficients:

$$Min\left(\sum_{i}^{n} {e_i}^2 + \lambda \sum_{j}^{k} \|a_j\|\right)$$

- λ tend vers l'infini ($+\infty+\infty$), les coefficients vont tendre vers 0 mais pourront atteindre 0.
- "Lasso" permet de "sélectionner" les covariables en vue de rendre le modèle plus simple (vu que les coefficients peuvent devenir nuls).

Régularisation

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

$$Min(\sum_{i}^{n} {e_i}^2 + \lambda \sum_{j}^{k} \|a_j\|^2$$

Ce paramètre λ est utilisé afin de pondérer plus ou moins l'importance de la minimisation du modèle en lui-même ou celle des coefficients.

- λ =0 même résultat MMC.
- λ tend vers l'infini ($+\infty+\infty$), les coefficients vont tendre vers 0 (mais sans jamais l'atteindre).
- Valeurs intermédiaires, les coefficients vont être plus petits que dans le modèle non-régularisé.

Préparation des données pour la régression linéaire

Préparation des Données:

En pratique, il faut **préparer les données** avant d'appliquer MMCO afin que les prédictions soient corrects.

Essayez différentes préparations de données en utilisant des heuristiques et voyez ce qui convient le mieux à votre problème.

contoso

Faire des prédictions avec la régression linéaire

- **Supprimer la colinéarité.** La régression linéaire surchargera vos données lorsque vous avez des variables d'entrée hautement corrélées.

Pensez à calculer des corrélations par paires pour vos données d'entrée et à supprimer les plus corrélées.

Préparation des données pour la régression linéaire

- **Supprimer le bruit**: La régression linéaire suppose que vos variables d'entrée et de sortie ne sont pas bruyantes.

Pensez à utiliser des opérations de nettoyage, ceci est très important pour la variable de sortie vous devez supprimer les valeurs aberrantes dans la variable de sortie (y) si possible.

Faire des prédictions avec la régression linéaire

- Il existe différentes techniques pour nettoyer les données avant de trouver un modèle. Ses méthodes ne font pas l'objet de ce cours.

Nous supposons donc que toutes les données utilisées dans ce cours sont déjà nettoyées.