

## 6. LA RÉGRESSION NON PARAMÉTRIQUE

# RÉGRESSION NON PARAMÉTRIQUE

Vous avez déjà parlé de régression dans la première partie de ce cours : l'objectif est d'étudier le comportement d'une variable aléatoire, disons  $Y$ , qui est liée à une deuxième variable aléatoire,  $X$ , selon une fonction  $g(X) : Y = g(X)$ .

L'approche paramétrique consiste à avancer des hypothèses à propos de la fonction  $g$ , et ensuite estimer les paramètres inconnus de cette fonction, à partir d'un échantillon observé.

E.g. régression linéaire :

$$Y = \beta X + \varepsilon$$

Paramètre inconnu      Erreur ou résidu

# RÉGRESSION NON PARAMÉTRIQUE

Vous avez déjà parlé de régression dans la première partie de ce cours : l'objectif est d'étudier le comportement d'une variable aléatoire, disons  $Y$ , qui est liée à une deuxième variable aléatoire,  $X$ , selon une fonction  $g(X) : Y = g(X)$ .

L'approche paramétrique consiste à avancer des hypothèses à propos de la fonction  $g$ , et ensuite estimer les paramètres inconnus de cette fonction, à partir d'un échantillon observé.

E.g. régression linéaire :

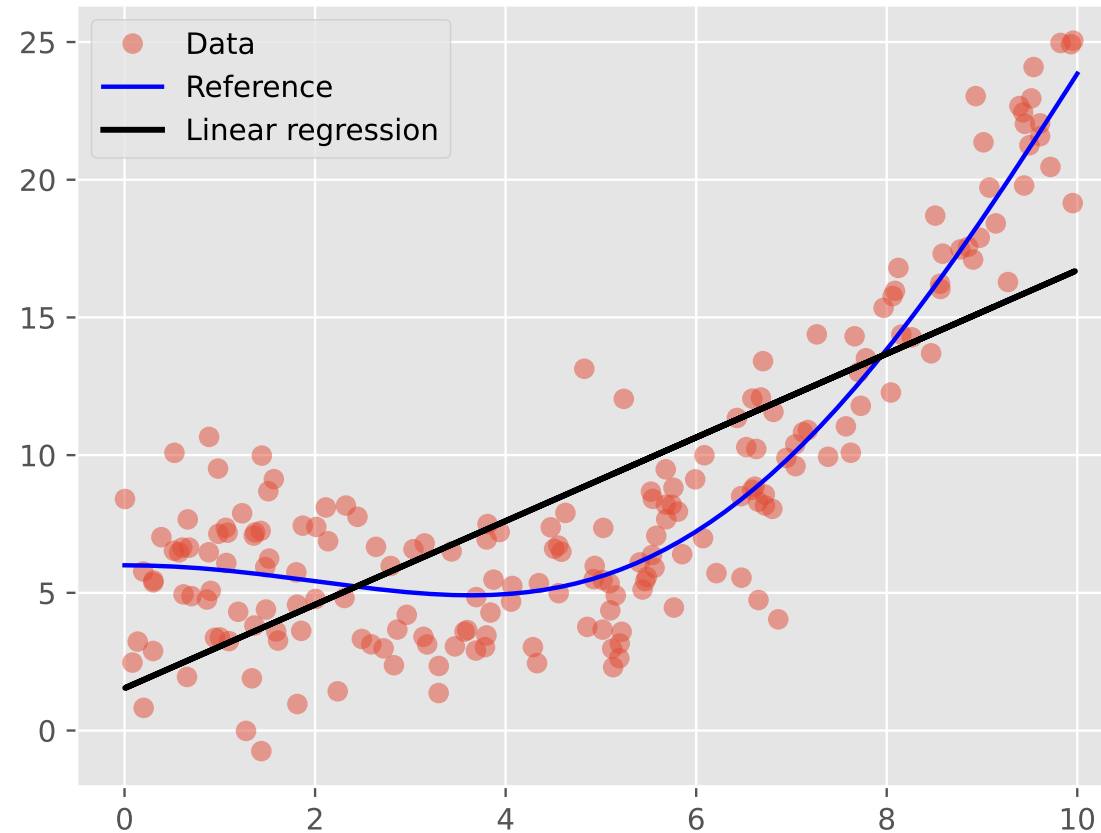
$$Y = \beta X + \varepsilon$$

Paramètre inconnu      Erreur ou résidu



# RÉGRESSION NON PARAMÉTRIQUE

Comme nous l'avons commenté pour l'estimation non paramétrique de la densité, malgré l'approche paramétrique aie des avantages (parcimonie, prédiction, calcul, poids de l'échantillon), les même problèmes liés au choix parfois erronée du modèle se posent ici également, d'où l'intérêt à explorer des méthodes non paramétriques.



# RÉGRESSION NON PARAMÉTRIQUE

Le problème peut être formaliser comme suit :

Soit  $\mathcal{D}_N = \{(x_1, y_1), \dots, (x_N, y_N)\}$  un  $N$ -échantillon, réalisation des variables  $X$  et  $Y$ , nous souhaitons modéliser le comportement de  $X$  comme fonction de  $Y$ , sans avancer des hypothèse sur la forme de la fonction  $g$  :

$$Y = g(X) + \varepsilon$$

# RÉGRESSION NON PARAMÉTRIQUE

Le problème peut être formaliser comme suit :

Soit  $\mathcal{D}_N = \{(x_1, y_1), \dots, (x_N, y_N)\}$  un  $N$ -échantillon, réalisation des variables  $X$  et  $Y$ , nous souhaitons modéliser le comportement de  $X$  comme fonction de  $Y$ , sans avancer des hypothèse sur la forme de la fonction  $g$  :

$$Y = g(X) + \varepsilon$$

Centré, de variance  $\sigma^2$  : bruit

# RÉGRESSION NON PARAMÉTRIQUE

Le problème peut être formaliser comme suit :

Soit  $\mathcal{D}_N = \{(x_1, y_1), \dots, (x_N, y_N)\}$  un  $N$ -échantillon, réalisation des variables  $X$  et  $Y$ , nous souhaitons modéliser le comportement de  $X$  comme fonction de  $Y$ , sans avancer des hypothèse sur la forme de la fonction  $g$  :

$$Y = g(X) + \varepsilon \leftarrow \text{Centré, de variance } \sigma^2$$

Pour résoudre ce problème, nous cherchons à déterminer  $g$  comme solution du problème de minimisation suivant :

$$\min [Y - g(X)]^2$$

# RÉGRESSION NON PARAMÉTRIQUE

Le problème peut être formaliser comme suit :

Soit  $\mathcal{D}_N = \{(x_1, y_1), \dots, (x_N, y_N)\}$  un  $N$ -échantillon, réalisation des variables  $X$  et  $Y$ , nous souhaitons modéliser le comportement de  $X$  comme fonction de  $Y$ , sans avancer des hypothèse sur la forme de la fonction  $g$  :

$$Y = g(X) + \varepsilon \leftarrow \text{Centré, de variance } \sigma^2$$

Pour résoudre ce problème, nous cherchons à déterminer  $g$  comme solution du problème de minimisation suivant :

$$\min [Y - g(X)]^2$$

D'où on obtiens (sous des conditions de régularité) :

$$g(X) = \mathbb{E}(Y|X)$$



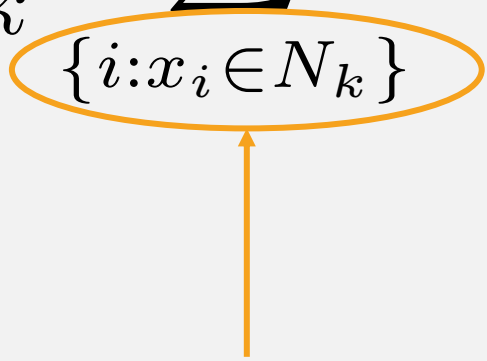
## 7. RÉGRESSION PAR K-PLUS- PROCHES-VOISINS

Un des approches plus intuitifs consiste à utiliser les k-plus-proches-voisins pour déterminer la valeur de  $g$  dans un point  $x$  donnée.

- Soit  $k$  le nombre des voisins que l'on souhaite considérer, choisi a priori.
- Pour un  $x$  donnée, soit  $N_k(x) := \{x_i \in \mathcal{D}_N \mid x_i \in \text{Dist}^\nearrow(x)[:k]\}$ , l'ensemble des premiers  $k$  plus proches voisins de  $x$ . Dans la suite on omettra  $x$  de la notation et écrira simplement  $N_k$ .
- Pourriez-vous imaginer comment peut on définir  $\hat{g}_k^{KNN}(x)$  à l'aide de cette notation ?

Idée : pensez à l'exemple de l'utilisation de k-plus-proches-voisins pour la classification, le même type de raisonnement s'applique ici pour la régression.

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$


On considère toutes les observations dans notre échantillon d'entraînement qui sont dans la boule de rayon  $k$  centrée sur  $x$

# RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

Pour chacune des observations  $x_i$  choisies, on va récupérer l'observation  $y_i$  correspondante et on en fait la moyenne

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

On répète ce procédé  
pour tout x dans  
l'intervalle considéré

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

Télécharger le fichier TP5\_Regression\_partiel.ipynb :  
[ibalelli.github.io](https://ibalelli.github.io) → Teaching → Modélisation statistique avancée

## 8. RÉGRESSION PAR RÉGRESSOGRAMME



## RÉGRESSION NON PARAMÉTRIQUE : RÉGRESSOGRAMME

Le régressogramme, et en quelque sorte l'équivalent de l'histogramme vu pour l'estimation non paramétrique d'une densité de probabilité.

- I. Partitionner l'espace  $[m, m + l]$  contenant les observations de la variable  $X$  en  $b$  intervalles de taille  $\nu$ .

# RÉGRESSION NON PARAMÉTRIQUE : RÉGRESSOGRAMME

Le régressogramme, et en quelque sorte l'équivalent de l'histogramme vu pour l'estimation non paramétrique d'une densité de probabilité.

1. Partitionner l'espace  $[m, m + l]$  contenant les observations de la variable  $X$  en  $b$  intervalles de taille  $\nu$ .
2. Pour tout intervalle  $[m + (i - 1)\nu, m + i\nu], i = 1, \dots, b$ , procéder au comptage des  $x_j$  :

$$\forall i = 1, \dots, b, C_i := \sum_{j=1}^N \mathbb{I}_{[m+(i-1)\nu, m+i\nu]}(x_j)$$

# RÉGRESSION NON PARAMÉTRIQUE : RÉGRESSOGRAMME

Le régressogramme, et en quelque sorte l'équivalent de l'histogramme vu pour l'estimation non paramétrique d'une densité de probabilité.

1. Partitionner l'espace  $[m, m + l]$  contenant les observations de la variable  $X$  en  $b$  intervalles de taille  $\nu$ .
2. Pour tout intervalle  $[m + (i - 1)\nu, m + i\nu], i = 1, \dots, b$ , procéder au comptage des  $x_j$  :

$$\forall i = 1, \dots, b, C_i := \sum_{j=1}^N \mathbb{I}_{[m+(i-1)\nu, m+i\nu]}(x_j)$$

3. Enfin, pour chaque  $x \in [m + (i - 1)\nu, m + i\nu]$ , la  $i$ -ème boîte, et pour chaque  $i = 1, \dots, b$ , nous estimons la fonction de régression en prenant la moyenne des  $y_i$  correspondantes à la classe considérée. Qu'est-ce qu'on obtiens ?

# RÉGRESSION NON PARAMÉTRIQUE : RÉGRESSOGRAMME

Le régressogramme, et en quelque sorte l'équivalent de l'histogramme vu pour l'estimation non paramétrique d'une densité de probabilité.

1. Partitionner l'espace  $[m, m + l]$  contenant les observations de la variable  $X$  en  $b$  intervalles de taille  $\nu$ .
2. Pour tout intervalle  $[m + (i - 1)\nu, m + i\nu], i = 1, \dots, b$ , procéder au comptage des  $x_j$  :

$$\forall i = 1, \dots, b, C_i := \sum_{j=1}^N \mathbb{I}_{[m+(i-1)\nu, m+i\nu]}(x_j)$$

3. Enfin, pour chaque  $x \in [m + (i - 1)\nu, m + i\nu]$ , la  $i$ -ème boîte, et pour chaque  $i = 1, \dots, b$ , nous estimons la fonction de régression en prenant la moyenne des  $y_i$  correspondantes à la classe considérée :

$$\hat{g}_{\nu}^{\text{Reg}}(x) = \frac{\sum_{j=1}^N \mathbb{I}_{[m+(i-1)\nu, m+i\nu]}(x_j) y_j}{C_i}$$

# RÉGRESSION NON PARAMÉTRIQUE : RÉGRESSOGRAMME

Les mêmes observations faites à propos des limites de cette approche s'appliquent ici également : il s'agit d'une approximation constante par morceaux. Pour l'obtenir, nous avons du choisir a priori certains paramètres :

- La valeur (ou coordonnée)  $m$
- Le nombre total d'intervalles (ou boîtes)  $b$
- La longueur (ou volume) de chaque intervalle/boîte  $v$

## 9. RÉGRESSION PAR NOYAUX

## RÉGRESSION NON PARAMÉTRIQUE : NOYAUX

Comme vu pour le cas de l'estimation de la densité, dans ce cas également une manière naturelle pour rendre l'estimation plus lisse est celle d'utiliser des noyaux  $K$  (un paramètre de lissage  $\nu$  est à définir ici également).

Sur la base des éléments introduits précédemment, écrire l'estimateur  $\hat{g}_\nu^K(x)$  pour un noyau générique  $K$ .

## RÉGRESSION NON PARAMÉTRIQUE : NOYAUX

Comme vu pour le cas de l'estimation de la densité, dans ce cas également une manière naturelle pour rendre l'estimation plus lisse est celle d'utiliser des noyaux  $K$  (un paramètre de lissage  $\nu$  est à définir ici également).

$$\hat{g}_{\nu}^K(x) = \frac{\sum_{j=1}^N K\left(\frac{x_j - x}{\nu}\right) y_j}{\sum_{j=1}^N K\left(\frac{x_j - x}{\nu}\right)}$$



# RÉGRESSION NON PARAMÉTRIQUE : NOYAUX

Les mêmes noyaux peuvent être reproposés ici :

- Le noyau gaussien :  $K(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
- Le noyau d'Epanechnikov :  $K(z) := \frac{3}{4}(1 - z^2)\mathbb{I}_{[-1,1]}(z)$
- Le noyau triangulaire :  $K(z) := (1 - |z|)\mathbb{I}_{[-1,1]}(z)$
- Le noyau uniforme :  $K(z) := \frac{1}{2}\mathbb{I}_{[-1,1]}(z)$

Télécharger le fichier TP5\_Regression\_partie2.ipynb : [ibalelli.github.io](https://ibalelli.github.io) → Teaching → Modélisation statistique avancée

## 10. PRÉDICTION NON PARAMÉTRIQUE

## PRÉDICTION

À quoi pouvez vous vous attendre lorsque vous souhaitez effectuer une prédiction pour un nouveau point  $x$  ? Si  $x$  est à l'intérieur de l'intervalle d'entraînement ? Et si  $x$  est à l'extérieur ?

## PRÉDICTION

