CHAPITRE 2.
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 : $\mathrm{Y} = g(X)$.

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 : $\mathrm{Y} = g(X)$.

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.

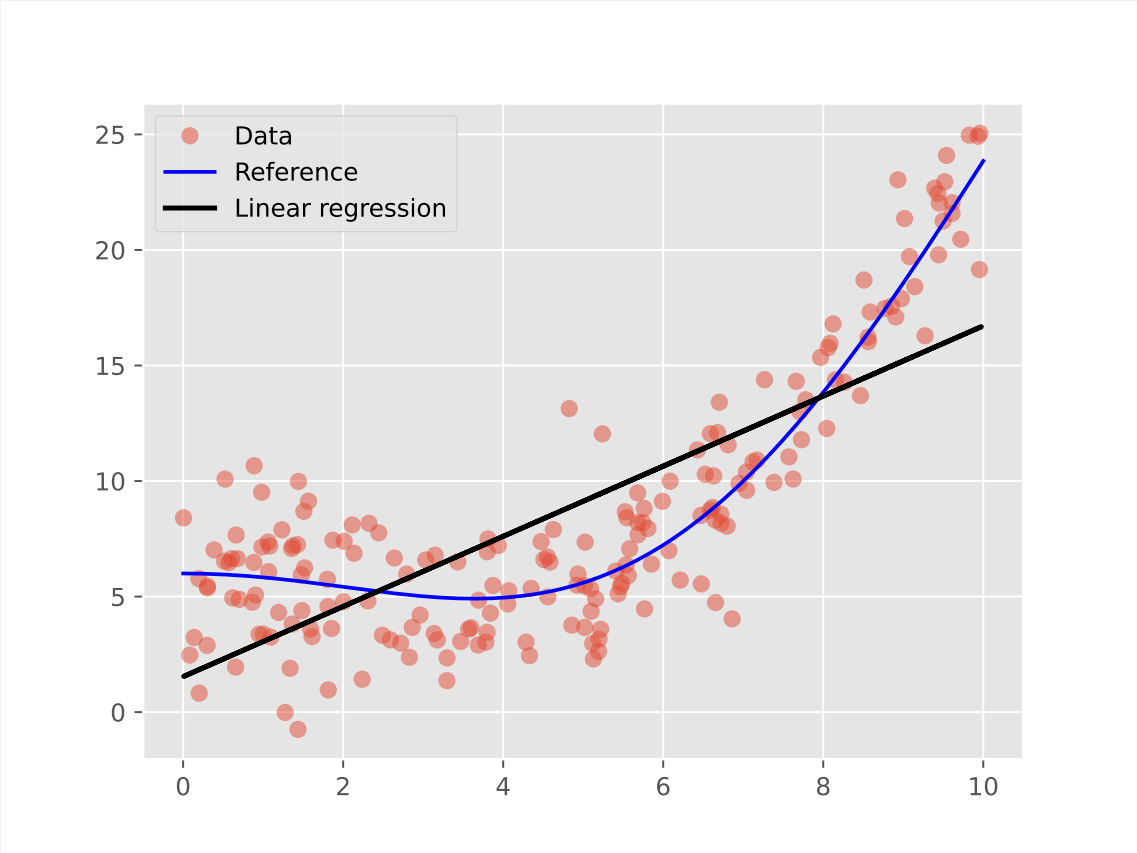
E.g. régression linéaire :

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

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 :

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

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

Centré, de variance σ^2 : bruit

2.1. RÉGRESSION PAR K-PLUS-PROCHES-VOISINS

RÉGRESSION NON PARAMÉTRIQUE : 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 $N_k(x) := \{x_i \in \mathcal{D}_N \mid x_i \in Dist^{\nearrow}\ (x)[:k]\}$, 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.

RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

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

RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

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

On considère toutes les observations dans notre échantillon d'entraînement qui sont dans la boulle 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

RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

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

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

RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

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

Télécharger le fichier TP5_Regression_partie1.ipynb :
ibalelli.github.io → Teaching → Modélisation statistique avancée