

TP n° 5 : Modèle de régression non-paramétrique

Exercice 1. Le fichier "NP.txt" disponible ici :

<https://chesneau.users.lmno.cnrs.fr/NP.txt>

contient un tableau de deux colonnes : la première donne 1000 valeurs (générées indépendamment) d'une variable explicative X et la seconde celles de la variable à expliquer Y . Nous considérons que les 800 premières lignes sont celles dont nous disposons pour faire des estimations alors que les 200 dernières lignes contiennent les abscisses pour lesquelles nous voudrions avoir une estimation de la précision de la fonction de régression (les valeurs de Y correspondantes ne seront donc pas utilisées dans les estimations, et ne sont pas censées être connues dans un premier temps).

1. Créer deux tables, **App** et **Val**, contenant respectivement les 800 premières et les 200 dernières lignes du tableau.
2. Représenter Y en fonction de X . Un modèle de régression linéaire semble-t-il adapté ?
3. En utilisant `npregbw` de la librairie `np`, déterminer la taille optimale de la fenêtre si l'on travaille avec un noyau d'Epanechnikov.
4. En utilisant un tel noyau ainsi que la fenêtre obtenue, donner les valeurs estimées de l'espérance conditionnelle de Y pour les 200 valeurs de X contenues dans la table **Val**, sans utiliser `npreg` (on fera donc des boucles).
5. Représenter les valeurs obtenues sur un graphique, avec les valeurs de X en abscisse. Inclure dans ce graphique les vraies valeurs de Y . Le résultat semble-t-il satisfaisant ? Même si la prévision est une application importante, l'objectif était d'estimer l'espérance conditionnelle. Ici, la vraie espérance conditionnelle est donnée par

$$\mathbb{E}(Y \mid \{X = x\}) = x \cos\left(\frac{x}{150}\right).$$

(On ne dispose évidemment pas de cette information en pratique). Représenter sur un même graphique, avec les valeurs de X en abscisse, la vraie valeur de l'espérance conditionnelle et son estimation.

6. Faire un graphique identique au précédent, mais présentant en plus les estimations obtenues en utilisant `npreg` avec un noyau gaussien. Avec quel noyau a-t-on visuellement un meilleur résultat ? Comparer numériquement ces deux estimateurs : on pourra comparer à partir de la connaissance que nous avons de la vraie fonction de régression, mais aussi faire une comparaison plus réaliste (par rapport aux conditions normales d'évaluation) qui n'utilise que les données de la table initiale.