Université Paris-Dauphine – Année 2019-2020
Executive Master : Régression non-paramétrique

Attention ! *Il n'est pas nécessaire de traiter toutes les questions pour obtenir une bonne évaluation. Le devoir est conçu pour un travail personnel de 2 - 3 heures. Certaines questions sont exploratoires et n'admettent pas nécessairement de solution unique.*

Modalités : A envoyer par mail (convertir au format pdf) avant le 16 septembre 2019 à l'adresse `celine.duval@parisdescartes.fr`

Situation

On dispose de deux jeux de données $(X_i, Y_i)_{1\leq i\leq 2000}$ où les X_i et les Y_i sont des réalisations de variables aléatoires réelles admettant la représentation

$$Y_i = r(X_i) + \sigma(X_i)\xi_i, \quad i = 1, \ldots 2000,$$

où

- Les ξ_i sont indépendantes et identiquement distribuées, avec $\mathbb{E}[\xi_1] = 0$ et $\mathbb{E}[\xi_1^2] = 1$, et ont une densité μ.
- La fonction $x \mapsto \sigma(x)$ est strictement positive. Si σ est constante on parle d'un modèle homoscédastique, sinon le modèle est dit hétéroscédastique.
- Les X_i sont indépendantes et identiquement distribuées de densité $g : [0, 10] \to \mathbb{R}_+$, et indépendantes des ξ_i.
- La fonction $r : \mathbb{R} \to \mathbb{R}$ vérifie $|r(x)| \leq 1$ pour tout $x \in [0, 10]$.

Les objectifs sont :

a. Reconstruire $x \mapsto g(x)$ graphiquement et étudier si g est la densité uniforme ou non.

b. Reconstruire $x \mapsto r(x)$ graphiquement.

c. Explorer les propriétés de $x \mapsto \mu(x)$ et $x \mapsto \sigma(x)$.

On dispose de deux jeux de données,

- `Data1` : dont la première colonne correspond aux X_i et la seconde colonne correspond aux Y_i. Dans ce jeu de données la variance des erreurs ne dépend pas de X.
- `Data2` : dont la première colonne correspond aux X_i et la seconde colonne correspond aux Y_i. Les différences avec les données `Data1` sont la loi μ des erreurs ξ et le fait que σ non constante.

Ainsi on a les mêmes valeurs pour les X_i et la même fonction de régression r dans `Data1` et `Data2`.

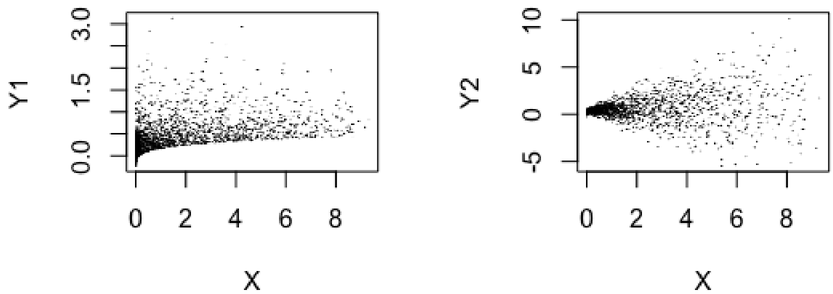

FIGURE 1 – Les deux jeux de données, `Data1` à gauche et `Data2` à droite, représentent les observations X en abscisse et Y en ordonnée.

1 Étude de la densité g des X

Pour cette partie on utilisera la première colonne des données `Data1`*.*

1.1. Construire un estimateur non-paramétrique $\widehat{g}_{n,h}(x)$ de $g(x)$ pour une fenêtre de lissage $h > 0$ donnée et représenter graphiquement $x \mapsto \widehat{g}_{n,h}(x)$ pour différentes valeurs de h que vous choisirez. On discutera la raison pour laquelle ce choix est important et ce qui se produit si h est mal choisi.

1.2. Représenter graphiquement $x \mapsto \widehat{g}_{n,\widehat{h}_n}(x)$, où $\widehat{h}_n$ est la fenêtre donnée par validation croisée ou par une autre méthode que l'on précisera.

1.3. Implémenter un QQ-plot pour vérifier empiriquement l'hypothèse $g(x) = 1/10$ pour tout $x \in [0, 10]$. L'hypothèse selon laquelle g est uniforme semble-t-elle raisonnable ?

1.4. Dans quelle zone de l'espace l'estimation de r sera plus précise ? Pourquoi ?

2 Reconstruction de $r(x)$

Pour cette partie on utilisera les données `Data1`*.*

2.1. Est-il plausible de penser que la fonction r est linéaire ? Tracer $Y1$ en fonction de $\log(X)$, que remarque-t-on ?

2.2. Construire un estimateur non-paramétrique $\widehat{r}_{n,h}(x)$ de $r(x)$ pour une fenêtre de lissage $h > 0$ bien choisie et le représenter graphiquement.

2.3. On se propose maintenant d'estimer r en régressant $Y1$ sur $\log(X)$. Construire un estimateur non-paramétrique $\widetilde{r}_{n,h}(x)$ de $\tilde{r}(x)$ dans le modèle $Y_1 = \tilde{r}(\log(X)) + \epsilon$, pour une fenêtre de lissage $h > 0$ bien choisie. Superposer sur le graphe de la question précédente l'estimateur $\widetilde{r}_{n,h}$ (sur le même graphe $x \mapsto \widehat{r}_{n,h}(x)$ et $x \mapsto \widetilde{r}_{n,h}(\log(x))$).

2.4. Que remarque-t-on ? Comment peut-on l'expliquer ?

3 Étude de la densité μ des ξ_i

3.1 A partir du jeu de données `Data1`

3.1.1. On cherche à estimer $x \mapsto \mu(x)$. Pour cela, on coupe l'échantillon en deux, selon que $i \in \mathcal{J}_- = \{1, \ldots, 1000\}$ ou que $i \in \mathcal{J}_+ = \{1001, \ldots, 2000\}$. On note $\widehat{r}_{n,h}^{(-)}(x)$ (pour un choix de h établi à la question 2.2) l'estimateur construit à l'aide de $(X_i, Y_i)_{1 \leq i \leq 1000}$ et on pose

$$\widetilde{\xi}_i = Y_i - \widehat{r}_{n,h}^{(-)}(X_i), \;\; i \in \mathcal{J}_+.$$

Quelle est la distribution approximative de $\widetilde{\xi}_i$?

3.1.2. En déduire un estimateur de $x \mapsto \mu(x)$ et l'implémenter graphiquement.

3.1.3. *(Facultatif.)* Quel est l'intérêt d'avoir découpé le jeu de données selon $\mathcal{J}_+$ et $\mathcal{J}_-$?

3.1.4. La densité $x \mapsto \mu(x)$ peut-elle être gaussienne ? Proposer un protocole pour le vérifier empiriquement et l'implémenter.

3.1.5. *(Facultatif.)* Comment peut-on tester si le modèle est bien homoscédastique ?

3.2 A partir du jeu de données `Data2`

On cherche à estimer $x \mapsto \mu(x)$ et $x \mapsto \sigma(x)$. Pour cela, on coupe à nouveau l'échantillon en deux et on considère à nouveau $\widetilde{\xi}_i$.

3.2.1. Justifier qu'en régressant $\widetilde{\xi}_i^2$ sur X_i on obtient un estimateur de $x \mapsto \sigma^2(x)$. L'implémenter et le visualiser graphiquement. En comparant avec le jeu de données (Figure 1 à droite), retrouve-t-on un résultat attendu ?

3.2.2. La densité $x \mapsto \mu(x)$ peut-elle être gaussienne ? Proposer un protocole pour le vérifier empiriquement et l'implémenter. *On pourra penser à renormaliser $\widetilde{\xi}_i$ par la fonction estimée à la question précédente et s'aider des questions de la Section 3.1.*