Devoir de statistique non paramétrique

Le travail sera fait individuellement ou en binôme, rédigé en utilisant LaTeX. Les fichiers (.tex, .pdf, .Rmd, .R, etc.) sont à déposer sur la page Moodle du cours, avant le 19 mars.

A Estimation de l'espérance d'une loi symétrique

Le fichier *devA.txt* contient $n = 200$ observations issues d'une loi symétrique, de centre de symétrie (ou paramètre de position) $\theta_0 \in \mathbb{R}$ inconnu. La densité f de X, par rapport à la mesure de Lebesgue sur R, peut donc s'écrire sous la forme :

$$f(x) = f_0(x - \theta_0) = f_0(\theta_0 - x), \ \forall x \in \mathbb{R},$$

où f_0 est est une densité symétrique sur R : $f_0(x) = f_0(-x)$ pour tout $x \in \mathbb{R}$.

Ecrire le programme, avec le logiciel de votre choix,

pour obtenir le "meilleur" estimateur $\widehat{f}_n$ au sens de l'erreur quadratique intégrée moyenne :

$$d\left(\widehat{f}_n, f\right) = E\int \left(\widehat{f}_n(x) - f(x)\right)^2 dx\,.$$

Expliquer le choix de la méthode et des paramètres de lissage. Déterminer graphiquement une approximation de θ_0. Proposer un estimateur asymptotiquement efficace $\widehat{\theta}_n$ de θ_0 et écrire le programme pour l'obtenir. Expliquer le choix des paramètres de lissage pour les estimateurs non paramétriques utilisés.

B Estimation de l'espérance sous l'hypothèse MAR

Le fichier *devB.txt* contient des données $(Y_i, X_i) \in \mathbb{R}^2$, $1 \le i \le 200$, issues d'un modèle de régression

$$Y = m(X) + \varepsilon, \; E(\varepsilon|X) = 0.$$

Les valeurs de la réponse Y ne sont pas toutes observées, mais l'hypothèse de données manquantes au hasard (MAR) est satisfaite : si la variable binaire D est telle que $D = 1$ si Y est observé et $D = 0$ si Y n'est pas observé, alors

$$Y \perp\!\!\!\perp D \mid X.$$

Autrement dit, si on note avec $p(X) = P(D = 1 \mid X)$ le score de propension, alors

$$P(D = 1 \mid Y, X) = p(X).$$

Ecrire le programme, avec le logiciel de votre choix, pour obtenir un estimateur consistant (et asymptotiquement efficace) $\widehat{\alpha}_n$ de $\alpha_0 = E(Y)$.

Expliquer le choix de la méthode et des paramètres de lissage.