Statistiques avec SAS : Sujet du TD N°6

Jean-Sébastien Roy (`js@jeannot.org`)

1 Estimation non paramétrique d'une densité

Créer une table, que l'on appellera `deuxn`, contenant quelques centaines de tirages d'observations `xi` suivant une loi N(0,1) pour 50% d'entre eux, et N(4,1.5) pour les 50% restants.

Créer une table, que l'on appellera `grid`, contenant une seule variable `x`, et dont les valeurs parcourent l'intervalle [-4,9] avec un pas tel que la table contienne une centaine de valeurs.

Calculer dans une table que l'on appellera `exact`, la valeur exacte de la densité pour l'ensemble des valeurs de la table `grid`.

Créer une table, que l'on appellera `stats`, contenant le nombre d'observations n, l'écart type $\widehat{\sigma}$, et l'intervalle inter-quartiles Q_{range} de la variable `xi` de la table `deuxn`.

Soit :

$$SNR = \widehat{\sigma}\left(\frac{4}{3n}\right)^{\frac{1}{5}}$$

Créer une table, que l'on appellera `h`, contenant une dizaine de valeur de $h = SNR \cdot e^f$ où f varie dans l'intervalle $[\ln(0.05), \ln(3)]$. La valeur h est appelée fenêtre.

Créer une table, que l'on appellera `kesti`, contenant pour l'ensemble des valeurs de `h` (dans la table `h`) et de `x` (dans la table `grid`) la variable `y` définie ainsi :

Soit le noyau gaussien :

$$K(d) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2}$$

Alors on appelle estimateur de Rozenblatt-Parzen la fonction :

$$y_h(x) = \frac{\sum_i K\left(\frac{x-x_i}{h}\right)}{nh}$$

Tracer une un même graphe les fonctions y_h et la densité précédemment calculée.

Comparer en fonction de h la moyenne des carrées des écarts sur les points de la grille, entre y_h et la densité précédemment calculée.

Recalculer et tracer y_h pour les valeurs de h suivantes :

$$h = SNR$$

$$h = SROT = 0.9 \cdot \min\left(\widehat{\sigma}, \frac{Q_{range}}{1.34}\right) \cdot n^{-\frac{1}{5}}$$

$$h = OS = 3\widehat{\sigma}\left(\frac{1}{70\sqrt{\pi}n}\right)^{\frac{1}{5}}$$

Ainsi que pour la valeur de h obtenue en réalisant une estimation non paramétrique de la densité avec SAS/Insight.

2 Estimation fonctionnelle non paramétrique

Créer une table `points`, contenant pour une centaine de valeurs de x_i suivant une loi uniforme sur $[0, 1]$, la valeur $y_i = f(x_i)$ de la fonction f définie par : $f(x) = \frac{1}{2} - 2x$ si $x < \frac{1}{5}$, et $f(x) = x - \frac{1}{10}$sinon ; fonction que l'on perturbera par l'ajout d'un bruit suivant une loin normale d'écart type $\frac{1}{10}$.

Calculer de manière similaire à la question précédente les tables `exact`, `stats`, `h` et `grid`.

Calculer de même les valeurs de l'estimateur de Nadaraya-Watson (1964) :

$$y_h(x) = \frac{\sum_i K\left(\frac{x-x_i}{h}\right) y_i}{\sum_i K\left(\frac{x-x_i}{h}\right)}$$

Tracer les courbes simultanément et les comparer avec la courbe exacte. On calculera de même la moyenne des carrés des écarts en fonction de h.

Calculer et tracer en fonction de h la validation croisée (Clark, 1975) qui correspond la moyenne des carrés des écarts entre y_i et la valeur $y_h(x)$ calculée sans prendre en compte le point i, ce qui correspond à la valeur du PRESS en régression linéaire.