Devoir 1

STT6415 - Régression

Échéance: vendredi 29 janvier.

Dans les problèmes énoncés ici, les données correspondent aux données du diabète sucré insulino-dépendant. Les problèmes concernent la régression (non paramétrique) de log(C-peptide) en fonction de l'âge.

1. Considérez la fonction de lissage donnée par la ligne-mobile.

   (a) Trouvez la matrice de lissage $S(x)$. C'est-à-dire, trouvez une expression générale pour la composante $S_{ij}(x)$. Justifiez votre réponse.

   (b) Effectuez la régression de log(C-peptide) comme fonction de l'âge en utilisant la ligne-mobile avec (i) $K = 2$ plus-proches-voisins symétriques (c'est-à-dire, deux voisins à gauche, et deux voisins à droite), et (ii) $K = 4$ plus-proches-voisins (pas nécessairement symétriques). Faites les graphiques des courbes et comparez les résultats.

2. Maintenant, effectuez la régression en utilisant une fonction de lissage par noyaux (par exemple, Epanechnikov).

   (a) Effectuez la régression pour une grille de valeurs du paramètre de lissage (la bande passante), par exemple, $\lambda \in [0, 10]$. Choisissez la valeur de la banda passante par validation croisée et par validation croisée généralisée. Aussi, effectuez la régression en utilisant une estimation de la valeur optimale asymptotique de la bande passante. Comparez les résultats. Quelles sont les degrés de liberté associés aux trois régressions?

   (b) Effectuez le même exercice, mais cette fois considérez $\sqrt{\text{âge}}$ comme la variable explicative. Comparez les résultats. Commentez.

3. Considérez les fonctions suivantes avec r noeuds intérieures $\{\xi_1, \ldots \xi_r\}$

$$\mathcal{S}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{r} \theta_j (x - \xi_j)_+^3.$$

   Ceci est la base de splines cubiques tronquées. Montrez que ces fonctions sont des splines cubiques, c'est-à-dire, montrez qu'elles satisfassent:

   (a) $\mathcal{S}$ est un polynôme cubique dans chacun des intervalles $[\xi_j, \xi_{j+1})$, $j = 1, \ldots, r-1$

   (b) $\mathcal{S}'$ et $\mathcal{S}''$ son continues.

   (c) $\mathcal{S}'''$ a des discontinuités de saut aux noeuds ξ_j, $j + 1, \ldots, r$.

4. L'effet de bord. Soit (x_i, y_i), $i = 1, ..., n$, un échantillon. Considérez la fonction de lissage par noyaux.

(a) Quelle est l'ordre de taille de la bande passante à x_1 ? Quelles valeurs de x_1 ont besoin de grandes largeurs de bande?

(b) Montrez que le "MSE locale" au bord de gauche x_1 est d'ordre $O(n^{-2/3})$.

5. Le noyau d'Epanechnikov. Supposez que le noyau les plus efficace est un noyau de la forme

$$K(t) = \begin{cases} at^2 + b, & -c < t < c \\ 0, & |t| >= c. \end{cases}$$

Montrez que les valeurs optimales de a, b et c sont

$$a = -3/(4 \times 5 \times \sqrt{5}),$$
$$b = 3/(4 \times \sqrt{5}), et$$
$$c = \sqrt{5}.$$