Оценивание плотностей. Непараметрическая регрессия.

1 (2 балл) Рассмотрим задачу непараметрической регрессии $Y = m(X) + \varepsilon$, где $X = (X_1, \ldots, X_n)$ – вектор признаков. В рамках метода локальной линейной регрессии

$$\sum_{i=1}^{n} q_{h(x)}(X_i - x)(Y_i - a(x) - b(x)(X_i - x))^2 \longrightarrow \min_{a(x), b(x)}$$

получить оценку $\widehat{m}(x)$ регрессионной функции $m(x)$ в явном виде.

2 (2 балла) По данным $\{X_i\}_{i=1}^{n}$ выбрать оптимальную ширину окна пропускания, построить ядерную оценку плотности и её 95%-ый доверительный интервал (т.е. доверительную полосу).

3 (2 балла) Выданы данные $\{(X_i, Y_i), i = 1, \ldots, n\}$. Рассмотрим задачу непараметрической регрессии $Y_i = m(X_i) + \varepsilon_i$. Построить непараметрическую регрессию с помощью оценки Надарая-Ватсона и методом сглаживающего сплайна (функция smooth.spline в R):

$$SS(h) = \sum_{i=1}^{n} (Y_i - m(X_i))^2 + h \int_{X_{(1)}}^{X_{(n)}} [m''(x)]^2 dx \longrightarrow \min_{m}$$

и вывести графики получившихся приближений. Зачем, на ваш взгляд, добавлять в выражение суммы квадратов интеграл от квадрата второй производной функции m? Какой выбор h, на ваш взгляд, является оптимальным? Где отличаются графики оценки Надарая-Ватсона и сглаживающего сплайна и как вы это объясните?

4 (3 балла) Выданы данные $\{(y_i, x_{ij}), i = 1, \ldots, n,\ j = 1, \ldots, k\}$. Построить по ним оценку функции $m(x)$ в модели непараметрической регрессии $Y = m(X) + \varepsilon$ методом Надарая-Ватсона, а также линейной регрессией и алгоритмом настройки с возвращением. Сравнить методы с помощью кросс-валидации.

5 (3 балла) Выданы данные $\{(y_i, x_{ij}), i = 1, \ldots, n+q,\ j = 1, \ldots, k\}$, причем $y_{n+1}, \ldots, y_{n+q}$ неизвестны. Используя пройденные методы регрессионного анализа, в рамках линейной регрессионной модели и модели непараметрической регрессии предсказать значения откликов объектов с номерами $n + 1, \ldots, n + q$. Описать и объяснить проделанные процедуры.