

Tutorial 4

1. The air pollutants include nitrogen dioxide (NO_2), Carbon dioxide (CO), sulphur dioxide (SO_2), respirable particulates (PM), Ozone (O_3) and others. Pollutants can be classified as either primary or secondary. Primary pollutants are substances directly produced by a process, such as ash from a volcanic eruption or the carbon monoxide gas from a motor vehicle exhaust. The primary can be controlled by reducing the emission of harmful gas. Secondary pollutants are not emitted. Rather, they form in the air when primary pollutants react or interact. An important example of a secondary pollutant is ozone. For the air pollution **data** in Hong Kong, fit the following partially linear regression model

$$\text{Ozone} = \beta_1 * \text{NO}_2 + \beta_2 * \text{SO}_2 + \beta_3 * \text{Particulate} + \beta_4 * \text{Humidity} + g(\text{Temperature}) + \varepsilon$$

- (a) What is your findings from the modelling?
 - (b) suppose a day's NO_2 , SO_2 , Particulate, Humidity and Temperature are respectively 58, 21, 76, 18 and 80. Predict the Ozone level in that day?
2. Suppose $(X_i, Y_i), i = 1, \dots, n$ are samples from (X, Y) . Write the details for the CV bandwidth selection in local linear kernel estimation of $E(Y|X = x)$ or the regression function in model $Y = m(X) + \varepsilon$.
 3. Suppose the true function is a k -th order polynomial, i.e.

$$m(x) = a_0 + a_1x + \dots + a_kx^k$$

Suppose $(X_i, Y_i), i = 1, \dots, n$ are sample from

$$Y = m(X) + \varepsilon$$

i.e.

$$Y_1 = m(X_1) + \varepsilon_1$$

$$Y_2 = m(X_2) + \varepsilon_2$$

...

$$Y_n = m(X_n) + \varepsilon_n$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent from each other and $E\varepsilon_i = 0$ for all i ; X_i are non-random values. Then it is claimed that the k -th order local polynomial kernel smooth

estimator is unbiased, i.e. $E\hat{m}(x) = m(x)$. Prove the claim for a special case where $k = 2$.

4. For the local linear kernel estimator $\hat{m}(x)$

$$\begin{pmatrix} \hat{m}(x) \\ \hat{m}'(x) \end{pmatrix} = \left(\sum_{i=1}^n K_h(X_i - x) \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix}^\top \right)^{-1} \sum_{i=1}^n K_h(X_i - x) \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} Y_i$$

(a) prove that

$$\hat{m}(x) = \frac{\sum_{i=1}^n \{s_{n,2}(x)K_h(X_i - x) - s_{n,1}(x)K_h(X_i - x)((X_i - x)/h)\}Y_i}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)}$$

where $s_{n,k}(x) = \sum_{i=1}^n K_h(X_i - x)\{(X_i - x)/h\}^k$ for $k = 0, 1, 2$

(b) By approximating

$$Y_i \approx m(x) + m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2 + \varepsilon_i,$$

Prove that the coefficient of $m'(x)$ in the expansion of $\hat{m}(x)$ is zero. i.e.

$$\begin{aligned} \hat{m}(x) \approx & m(x) + \frac{1}{2}m''(x)h^2 \frac{s_{n,2}^2(x) - s_{n,1}(x)s_{n,3}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)} \\ & + \frac{\sum_{i=1}^n \{s_{n,2}(x)K_h(X_i - x) - s_{n,1}(x)K_h(X_i - x)((X_i - x)/h)\}\varepsilon_i}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)} \end{aligned}$$