Non-parametric regression

Motivation

- Linear regression fits a linear line, which might be a poor fit for general datasets.
- Need a powerful estimator of E(Y|X) without making any assumption about the functional form.
- $\mathrm{E}(Y|\, x_1, \ldots, x_k) = \mathrm{m}(\mathrm{x})$ where **x = ($x_1, \ldots, x_k$)**

- The function m(x) we will derive under the assumption that f(X,Y) and f(X) are both estimated using kernels.

$$m(x) = E(Y|x) = \int_y y \underbrace{f(Y|x)} \, dy$$

X is also a random variable

$$D = \{(x_1, y_1) \cdots (x_n, y_n)\}$$

$$= \int_y y \frac{f(x\ y)}{f(x)} \, dy$$

$\hat{f}(x\ y)$

$\hat{f}(x)$

estimate using KDE on D

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

$$\hat{f}(x, y) = K_{xy}([X_i, Y_i], (x\ y)) = K(X_i, x) K(Y_i, y)$$

$$= K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right)$$

$$E(Y|X) = \int_y y \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \quad \hat{f}(x\, y)$$

$$\int_y y\, K\left(\frac{Y_i - y}{h}\right) dy$$

change of variable $= Y_i$

$m(Y)$

$$= \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} = \hat{Y}$$

$$\text{Test } (x) \longrightarrow \hat{Y} = \sum_{i=1}^{n} w_i(x) Y_i$$

$$w_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

Non-parametric estimate of E[Y|x]

Starting with the definition of conditional expectation,

$$\operatorname{E}(Y \mid X=x) = \int y f(y \mid x)\, dy = \int y \frac{f(x,y)}{f(x)}\, dy$$

we estimate the joint distributions *f*(*x*,*y*) and *f*(*x*) using kernel density estimation with a kernel ***K***:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)K_h(y-y_i),$$

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i),$$

We get:

$$\begin{aligned}
\hat{\operatorname{E}}(Y \mid X=x) &= \int y \frac{\hat{f}(x,y)}{\hat{f}(x)}\, dy,\\
&= \int y \frac{\sum_{i=1}^{n} K_h(x-x_i)K_h(y-y_i)}{\sum_{j=1}^{n} K_h(x-x_j)}\, dy,\\
&= \frac{\sum_{i=1}^{n} K_h(x-x_i)\int y\, K_h(y-y_i)\, dy}{\sum_{j=1}^{n} K_h(x-x_j)},\\
&= \frac{\sum_{i=1}^{n} K_h(x-x_i) y_i}{\sum_{j=1}^{n} K_h(x-x_j)},
\end{aligned}$$

which is the Nadaraya–Watson estimator.

Nadaraya–Watson kernel regression

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}$$

Demo: https://colab.research.google.com/github/tufts-ml-courses/cs135-23f-assignments/blob/main/labs/day20-KernelRegression.ipynb

Choosing bin-width is again a problem

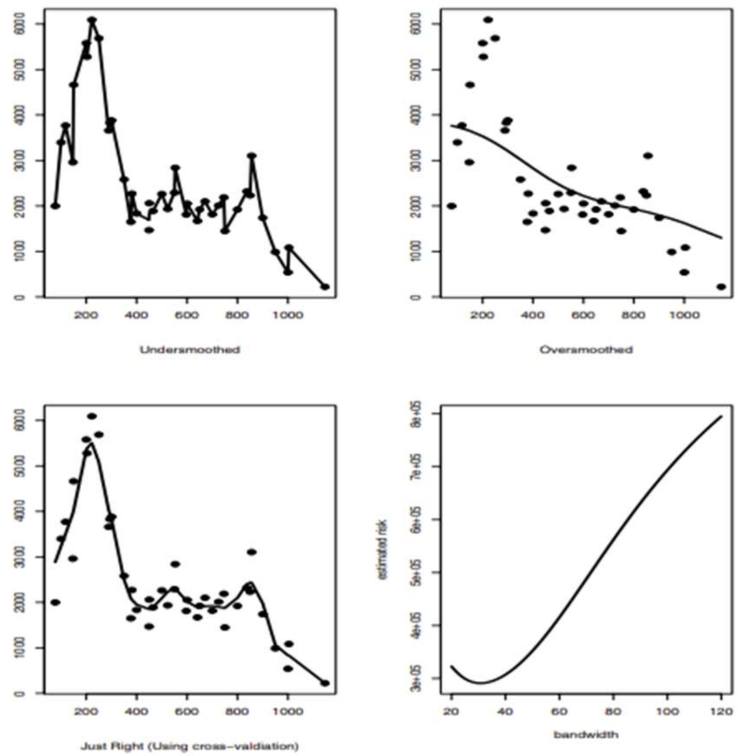

FIGURE 20.8. Regression analysis of the CMB data. The first fit is undersmoothed, the second is oversmoothed, and the third is based on cross-validation. The last panel shows the estimated risk versus the bandwidth of the smoother. The data are from BOOMERaNG, Maxima, and DASI.

$n+m$

$m<n$

D

train split — D_n

Validation split — D_m

$\hat{f}_{h_1}(x) \rightarrow E_1$

Error of $\hat{f}_{h_j}$ on D_m

$\vdots$

$\hat{f}_{h_{10}}(x)$ — E_{10}

Choose h with minimum error.

Multidimensional extension.

$$K\left(\frac{x_i - x}{h}\right) = K\left(\frac{\|x_i - x\|^2}{h}\right) \qquad x \in R^k$$

$$x \equiv [x_1 \ldots x_k]$$

$$L_p = \|x_i - x\|_p$$

$$= \left(\sum_{j=1}^{k} |x_{ij} - x_j|^p\right)^{1/p}$$

Summary of regression

- Linear regression (1-D data)
  - MLE estimates of slope and intercept
  - Unbiased chi-squared distribution based estimate of σ^2
- Distribution of parameters
- Linear regression (Arbitrary k)
  - Just the derivation of MLE estimate
- Kernel regression
  - Just the final estimate.