Kernel Regression on synthetic and real data

Mirko Bicchierai, Chiara Peppicelli, Jay Senoner

Università di Firenze
Statistical Learning

April 19, 2024

Introduction

Suppose that we observed n independent pair of points $\{(x_i, y_i)\}_{i=1}^n$, and that the variables **X** and **Y** have a functional relationship of the form:

$$y_i = m(x_i) + \epsilon_i$$

Objective: Estimate the regression function $m : \mathbb{R}^p \to \mathbb{R}$.
We will use the **Nadaraya–Watson** estimator:

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

where $K_h(x) = \frac{1}{h} K(\frac{x}{h})$ is a **kernel** with bandwith h.

Observation

The Nadaraya–Watson estimator can be seen as a **weighted average** of $Y_1, \ldots, Y_n$ by means of the set of weights $\{W_i(x)\}_{i=1}^n$ (they add to one). The set of varying weights depends on the evaluation point x. That means that the Nadaraya–Watson estimator is a **local mean** of $Y_1, \ldots, Y_n$ about $X = x$.

Bandwith tuning using Leave-One-Out CV

Following an analogy with the fit of the linear model, we could look for the bandwidth h such that it minimizes an RSS of the form:

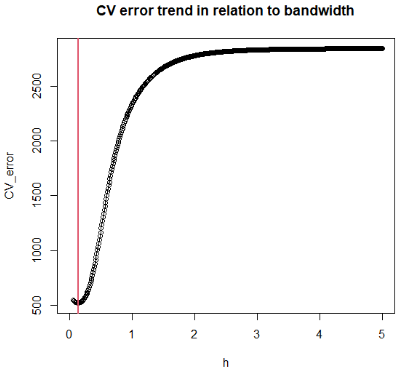
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i; p, h))^2$$

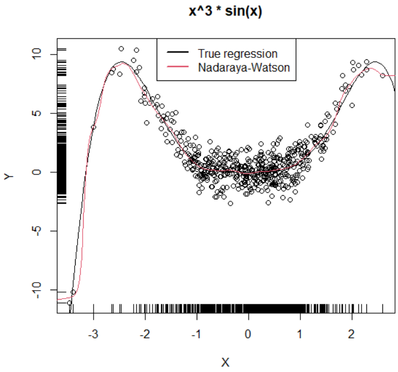
Attempting to minimize the **RSS** always leads to values of $h \approx 0$. To overcome this problem we compare Y_i with the **leave-one-out** estimate of m, yielding the **least-square CV error**:

$$CV(h) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{-i}(X_i; p, h))^2, \qquad h_{CV} := \arg\min_{h>0} CV(h)$$

Synthetic data 1

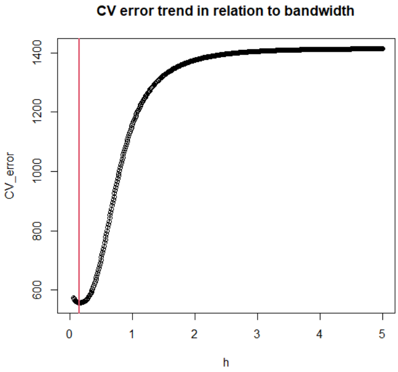
h_{CV}: **0.13**

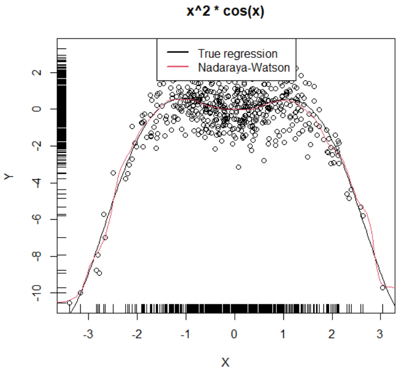




Synthetic data 2

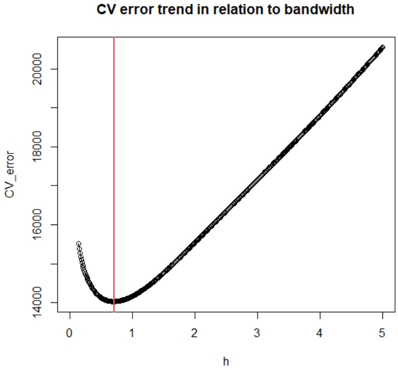
h_{CV}: **0.15**

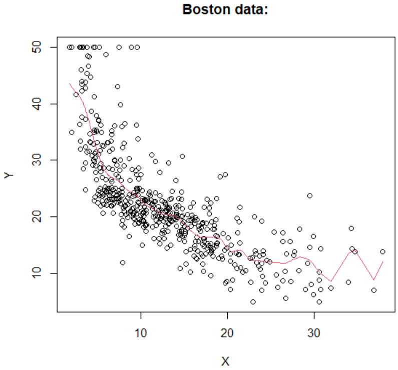




Boston dataset

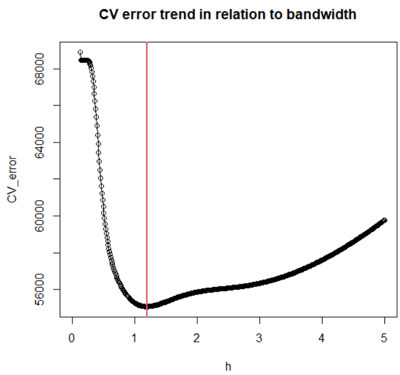
Response variable: **medv** , Predictor variable: **lstat**, h_{CV}: **0.7**

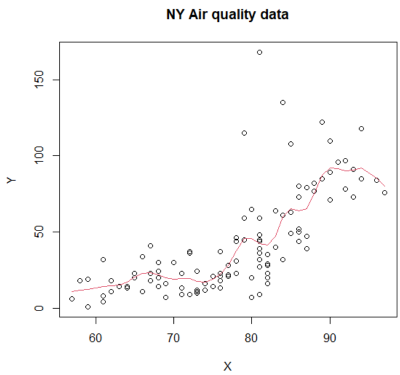




Air Quality dataset

Response variable: **Ozone** , Predictor variable: **Temp**, h_{CV}: **1.19**





References

Eduardo García-Portugués (2024).
Predictive modelling course notes.
https://bookdown.org/egarpor/PM-UC3M/npreg-kre.html#npreg-kre-bwd.
[MSc in Big Data Analytics, Carlos III University of Madrid].

Nathaniel E. Helwig (2021).
Nonparametric regression(smoothers) in r.
http://users.stat.umn.edu/~helwig/notes/smooth-notes.html.
[Department of Psychology and School of Statistics, University of Minnesota, January 04].