3.5 Comparison of Linear Regression with K-Nearest Neighbors

Zongyi Liu

2023-05-18

3.5 Comparison of Linear Regression with K-Nearest Neighbors

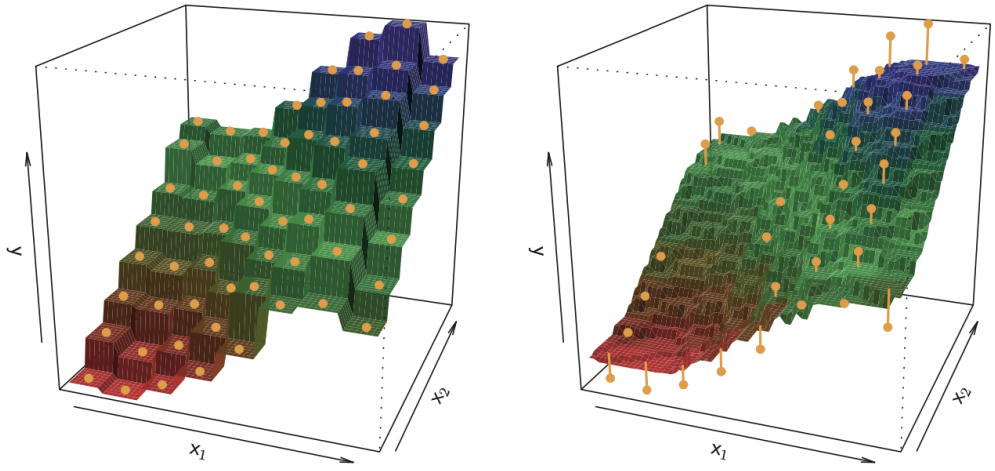
As discussed before, linear regression is an example of parametric regression, which has advantages, but also has disadvantage: by construction, they make strong assumptions about the form of $f(X)$.

In contrast, non-parametric methods do not explicitly assume a parametric form for $f(X)$, therefore provides an alternative and more flexible approach for performing regression.

Here we consider the K-nearest neighbors regression **(KNN regression)**. KNN regression is closely related to the KNN classifier (in Chapter 2). Here $f(x_0)$ can be estimated as:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

In plots, we can see as

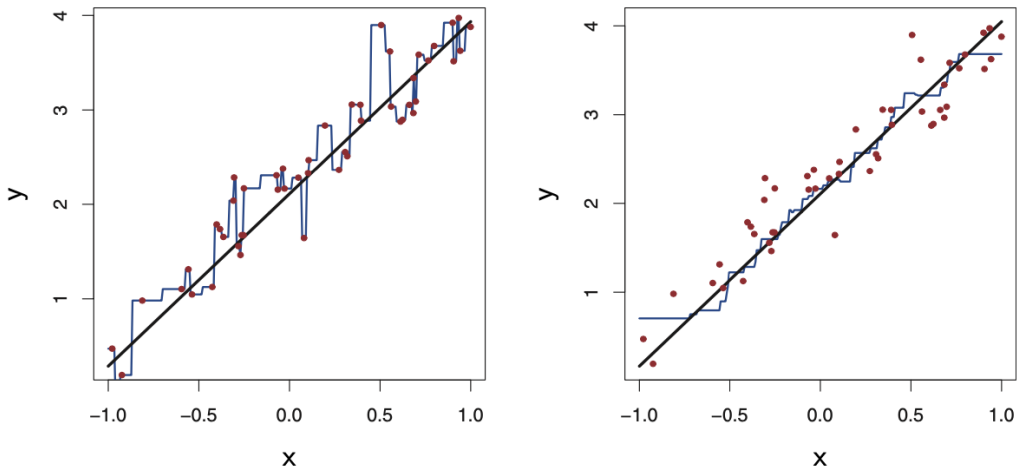


In the left panel, K=1, and in the right panel, K=9, which is much smoother and has smaller regions than the left one. We can conclude that the optimal K value depends on the bias-variance trade-off. A small value for K provides the most flexible fit, which will have low bias but high variance; a large K value provides a smoother and less variable fit.

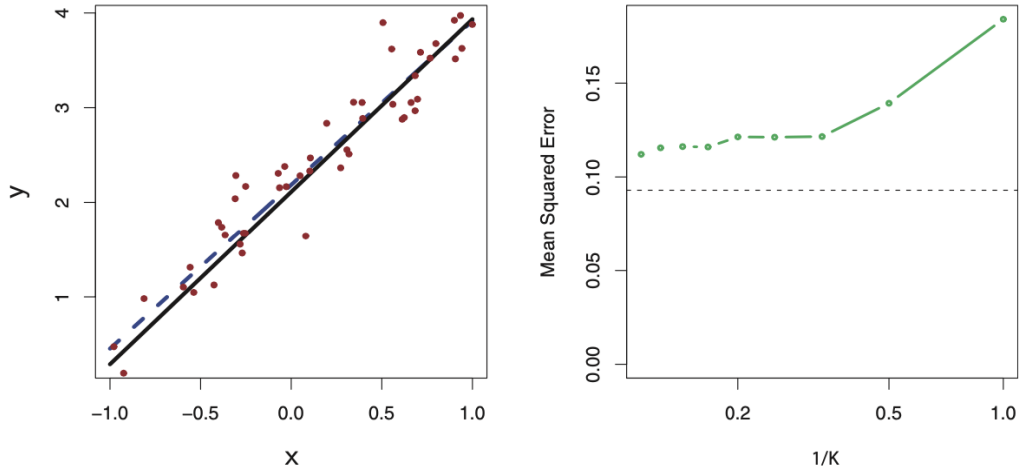
3.5.1 In what condition will the parametric method be better?

The answer is if the parametric form that has been selected is close to the true form of f.

When Real Function is Linear

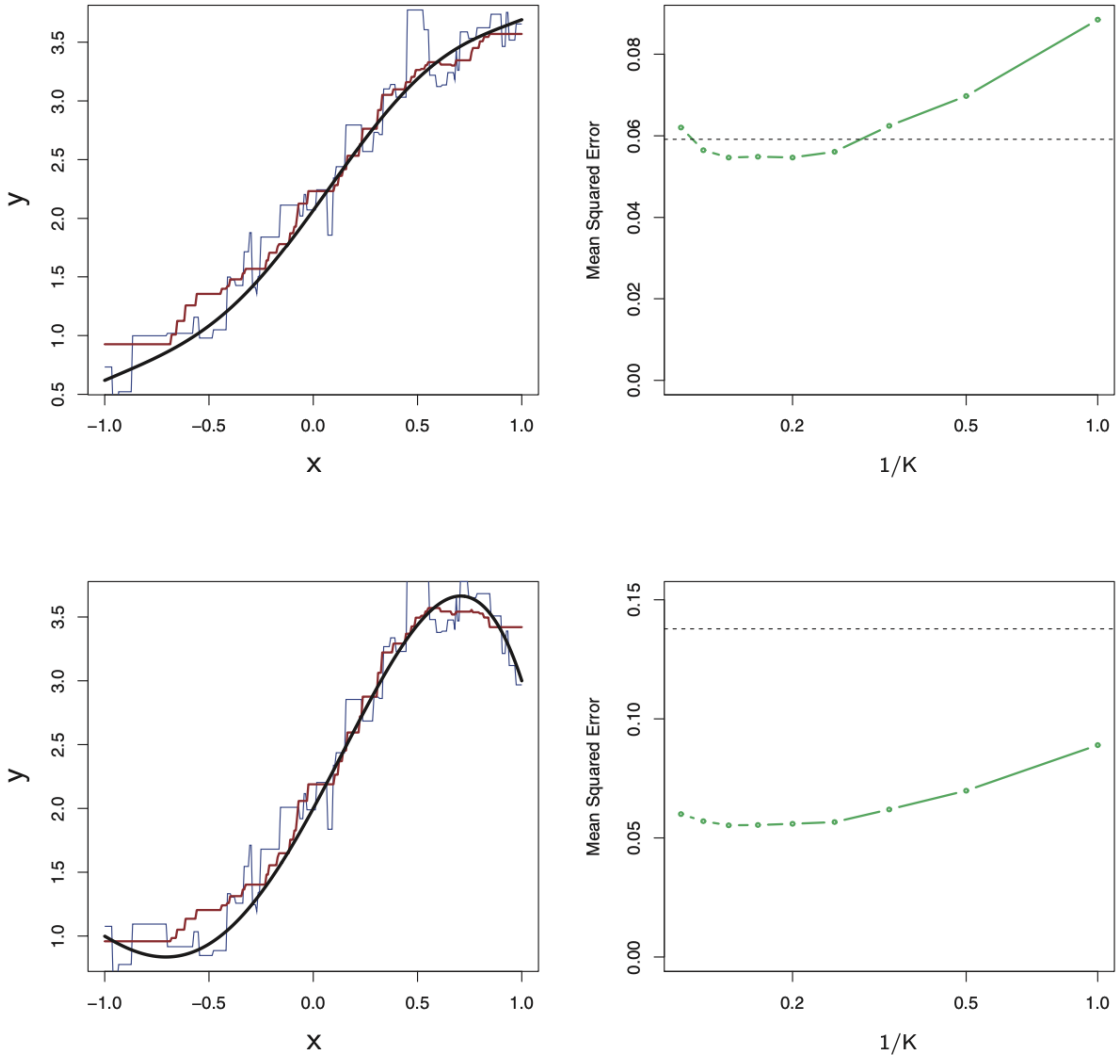


- Left: K=1
- Right: K=9



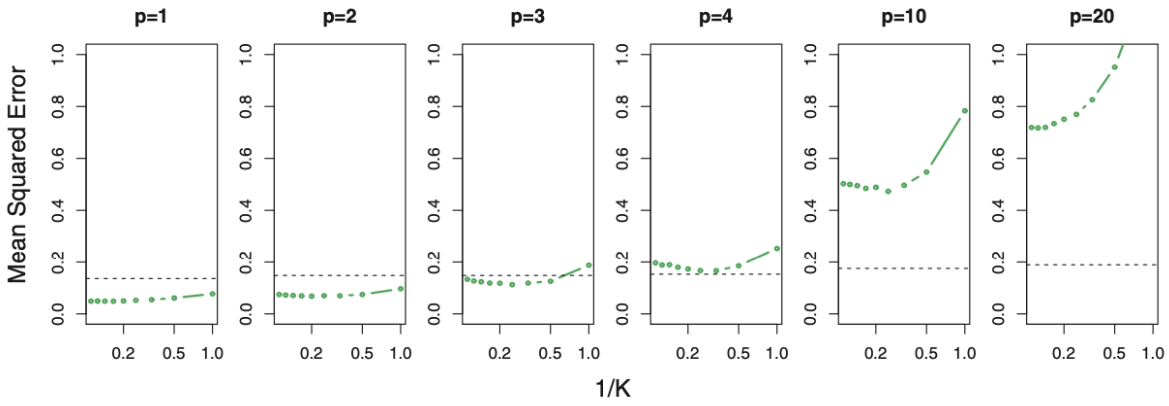
- Left: same data set doing linear regression
- Right: dashed horizontal line: the least squares test set MSE; the green solid line: MSE for KNN as a function of 1/K (the MSE of linear regression is much smaller)

When Real Function is Non-Linear



- Above: K=1, here the MSE is increasing as 1/K increases
- Below: K=9, here MSE of non-parametric regression is always smaller than the linear one

Dimension



The test MSE for linear regression and KNN. As p increases, the MSE of KNN increases much faster than linear regression, meaning its performance degrades much more faster as p increases.

However, even in problems in which the dimension is small, we might prefer linear regression to KNN from an interpretability standpoint.