Comparison of Linear Regression with K-NN

- As stated in Basics-of-Statistical-Learning
- Linear Regression is a parametric approach
- KNN is a non-parametric approach

Parametric approaches assumes the form of $f(X)$ which can lead to some problems

- The real relationship between the Response and Predictors isn't as we assumed
- Overfitting problems
  Non Parametric is more of a flexible approach to perform regression

The KNN Regression

Its close to the K-Nearest Neighbors that uses Byes' Classifier which deals with Classification problems.

Given value for K and a prediction point x_0

1. Identifies the K training observation that are closest to x_0 the point we wanna predict
2. then it estimates $f(X)$ using average of all the training the responses in $\mathcal{N}_0$
   - **Simple Average** :

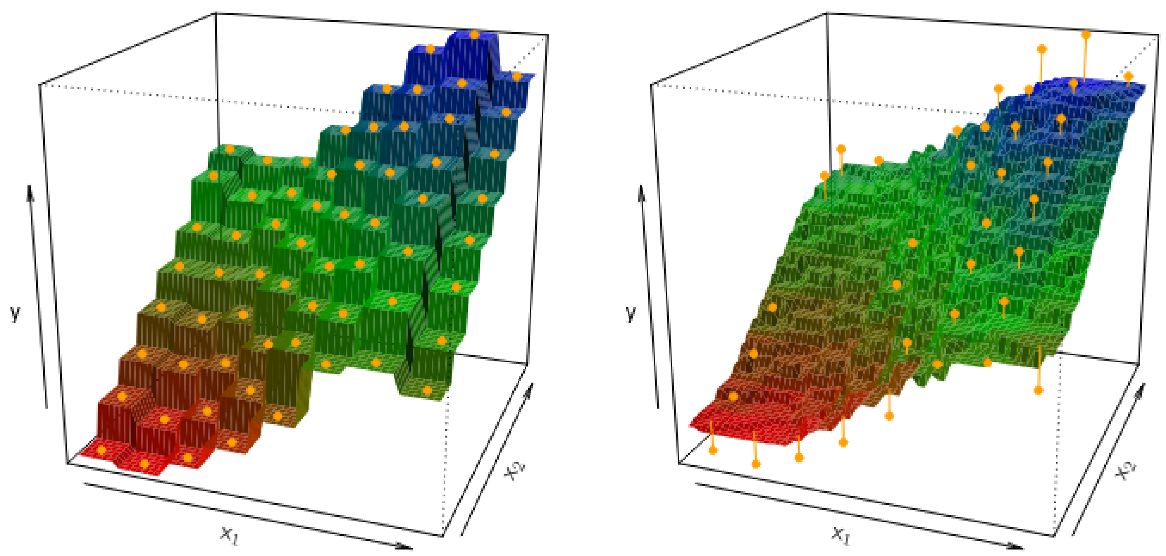
$$\hat{f}(X) = \frac{1}{K} \sum_{x \in \mathcal{N}_o} y_i$$

   - It averages out the values for all the nearest points to x_0
   - **Weighted Average** : Calculate the distance so the closest neighbors contribute more

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$
$$\hat{y} = \frac{\sum w_i y_i}{w}$$

   With :

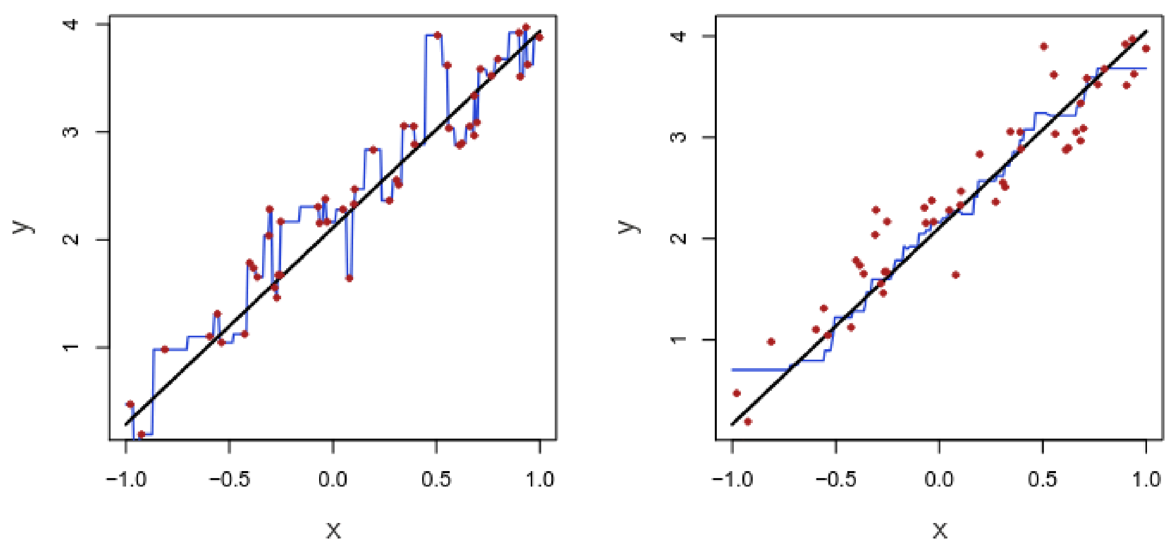
$$w_i = \frac{1}{d(x, y)}$$



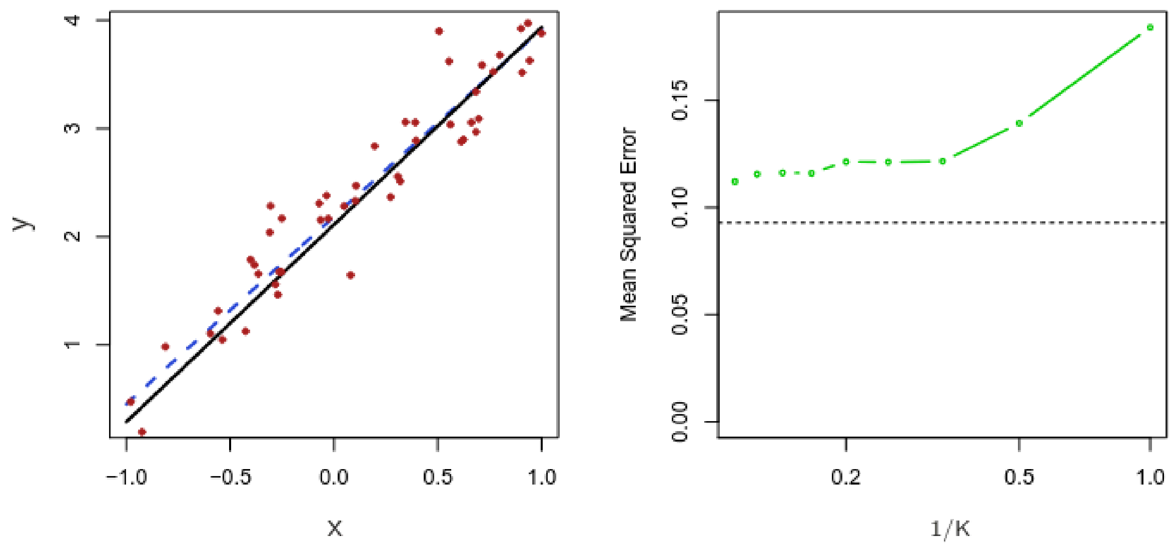
- When $K = 1$ The KNN fits perfectly the training observations (data) Left Figure
- When $K = 9$ The KNN smoother fit averaging out between 9 data points
- The optimal value for K depends on the $Bais - Variance$ Trade off, small K provides the most flexible fit which corresponds for a low bias and a high variance
- Large K values provides smoother and less variable fit Low Variance, But may cause for a higher bias hiding some patterns in the true form of $f(X)$

K-NN Regression Vs Linear Regression

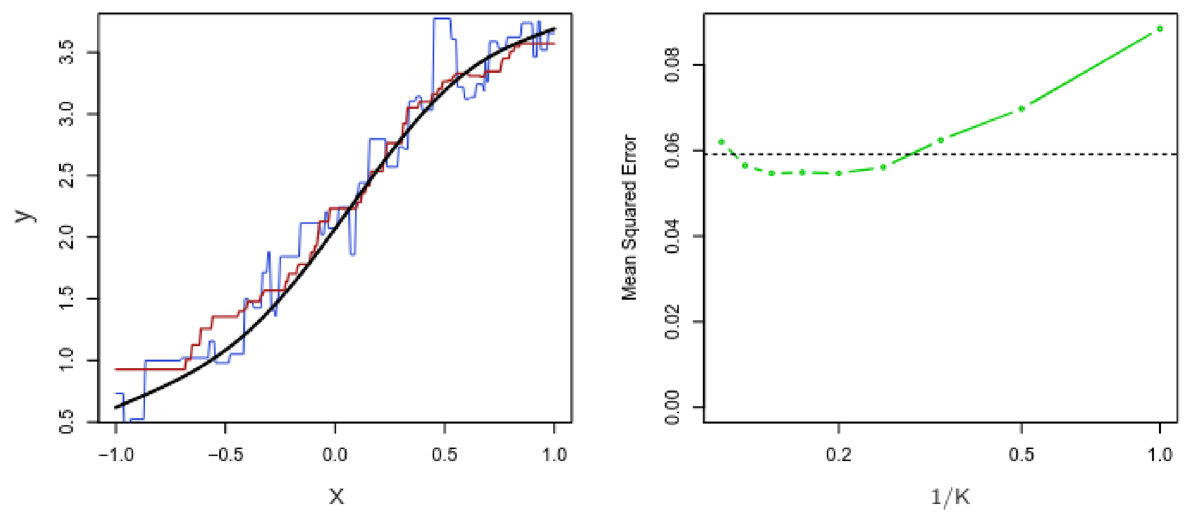
- The **Parametric** approach will out perform the **non- parametric** one when the true relationship is Linear (Or the assumed form of $\hat{f}(X)$ is close to the real shape of $f(X)$)
- The non-parametric approach incurs cost in variance without reducing the bias

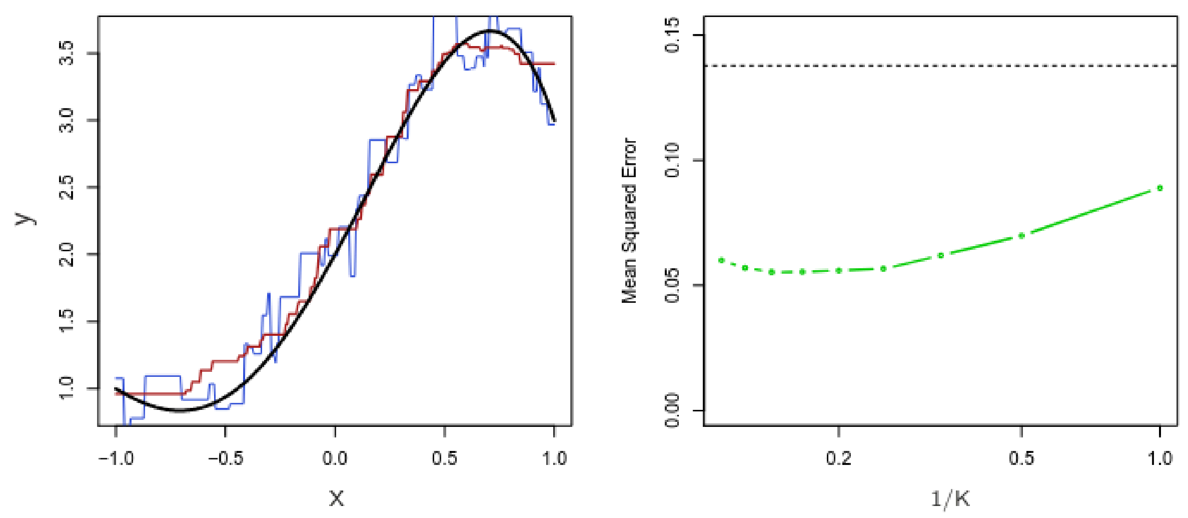


- The True relationship is Linear represented in a black line
- The **Left Figure** is K-NN Regression with $K = 1$
- The **Right Figure** is $K = 9$ with gives a closer and smoother fit

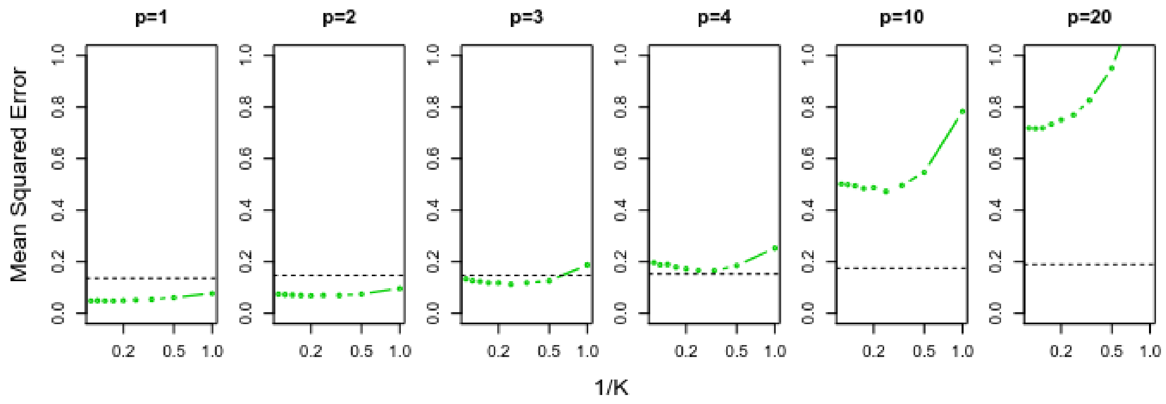


- The Blue dashed Line is a Linear Regression Fit
- The MSE of the Linear Regression is better than the K-NN Regression





- Here we notice that the more the true form of $f(X)$ is not linear the K-NN regression outperform **the parametric Linear Regression** shown in the MSE graph
- $K = 1$ is the blue line fit
- $K = 9$ is the Red line fit



- In higher dimensions The KNN Regression tend to perform poorly compared to the Parametric Regression
- Cause in higher dimensions the distance start to mean less and less $Dimensionality\ Curse$
- The non-parametric approach needs a large observation
- And the Parametric Regression is more interperable

Conclusion

Key Points	Parametric Regression	Non-Parametric Regression
Low $p\ Variables$	May under-fit if the form is too simple	Often performs better due to flexibility
Higher $p\ Variables$	Handles better if form of $f(X)$ is correct	Perform worse, suffers from $demensionality\ Curse$
Non-Linear $f(X)$	High bias	Performs better if the observations are large enough
Linear $f(X)$	Performs well with low variance	May overfit/noise , higher variance
Overfitting Risk	Moderate - depends on the model complexity	High if the K isn't optimal
Data Efficiency	Works well with small data sets	Requires Large amount of observations to generalize well
Interpretability	High - Easy to explain coefficients - Works great for Inference	Low - Harder to explain the results
Computation	Faster	Slower in higher dimensions