# K-Nearest Neighbors regression

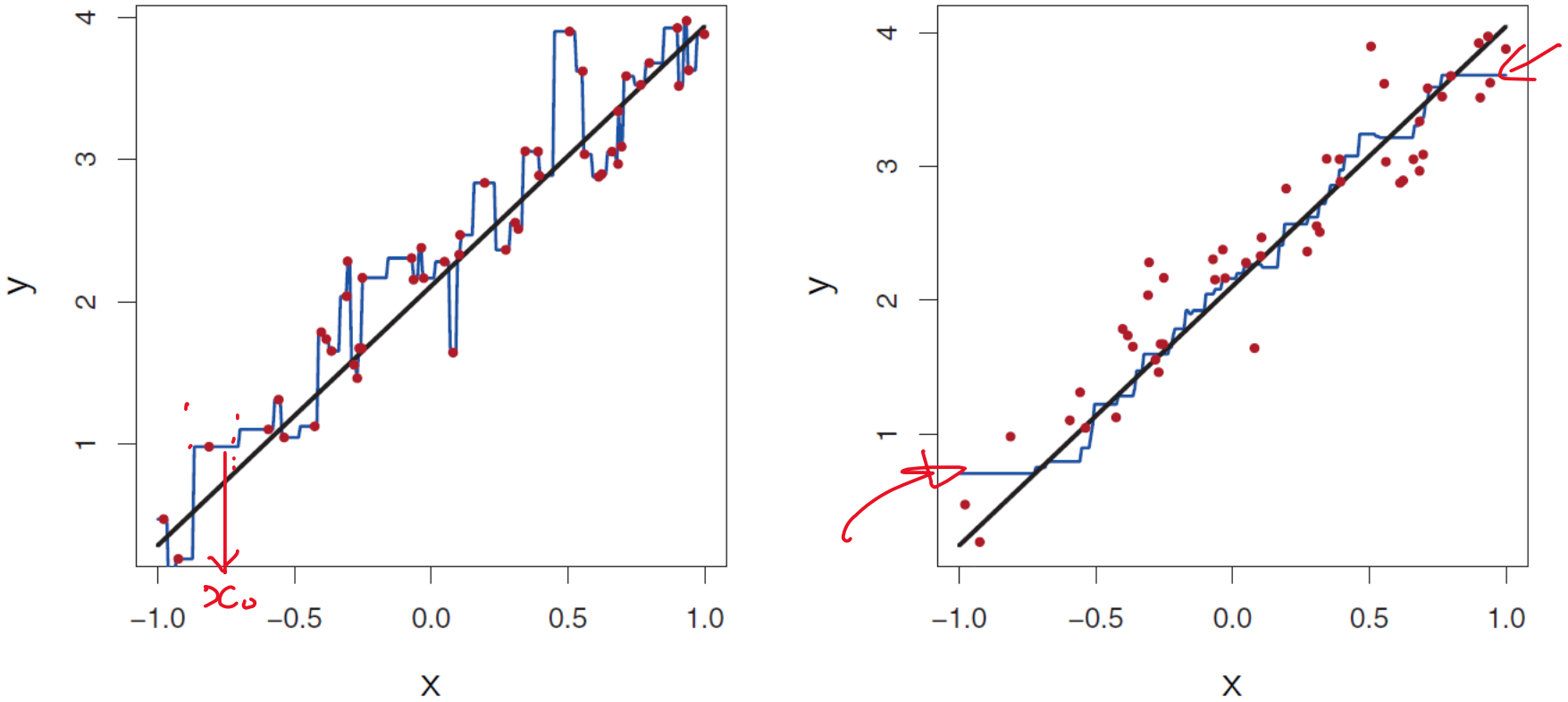
Prof. Asim Tewari
IIT Bombay

# *K*-Nearest Neighbors regression

Given a value for K and a prediction point $x_0$, KNN regression first identifies the K training observations that are closest to $x_0$, represented by $N_0$. It then estimates $f(x_0)$ using the average of all the training responses in $N_0$. In other words,

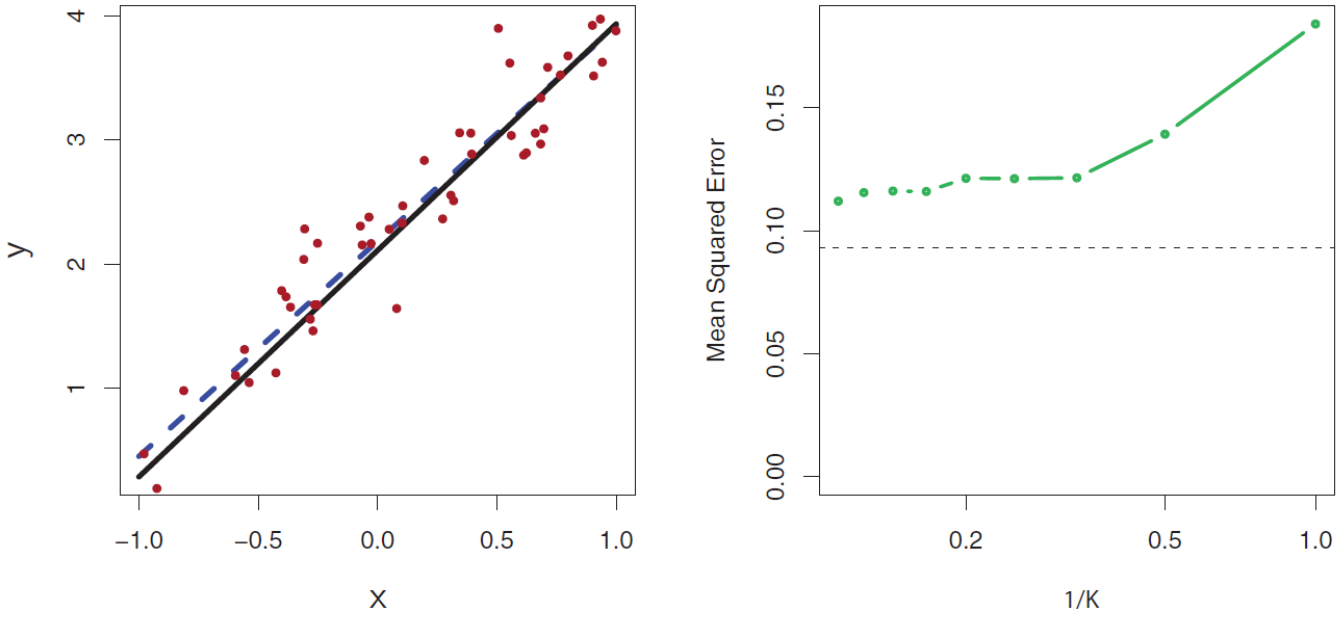
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

# Comparison of Linear Regression with *K*-Nearest Neighbors



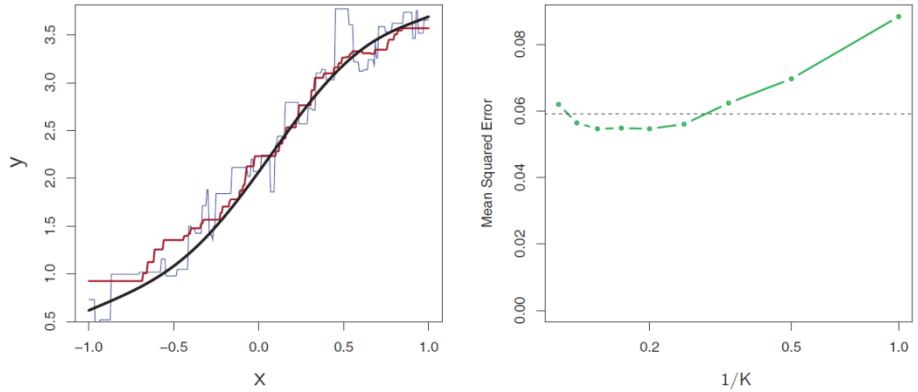

Plots of f(X) using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line. Left: The blue curve corresponds to K = 1 and interpolates (i.e. passes directly through) the training data. Right: The blue curve corresponds to K = 9, and represents a smoother fit.

# Comparison of Linear Regression with *K*-Nearest Neighbors

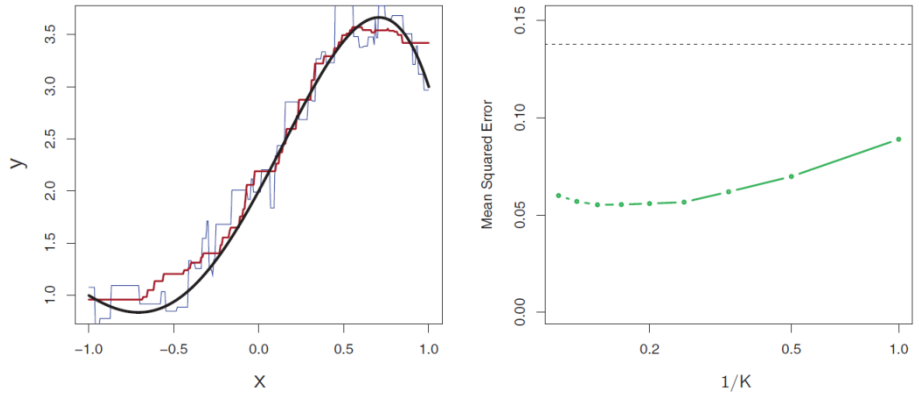



*The same data set as in the previous slide is investigated further.* Left: *The blue dashed line is the least squares fit to the data. Since f(X) is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of f(X).* Right: *The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of 1/K (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since f(X) is in fact linear. For KNN regression, the best results occur with a very large value of K, corresponding to a small value of 1/K.*

# Comparison of Linear Regression with *K*-Nearest Neighbors

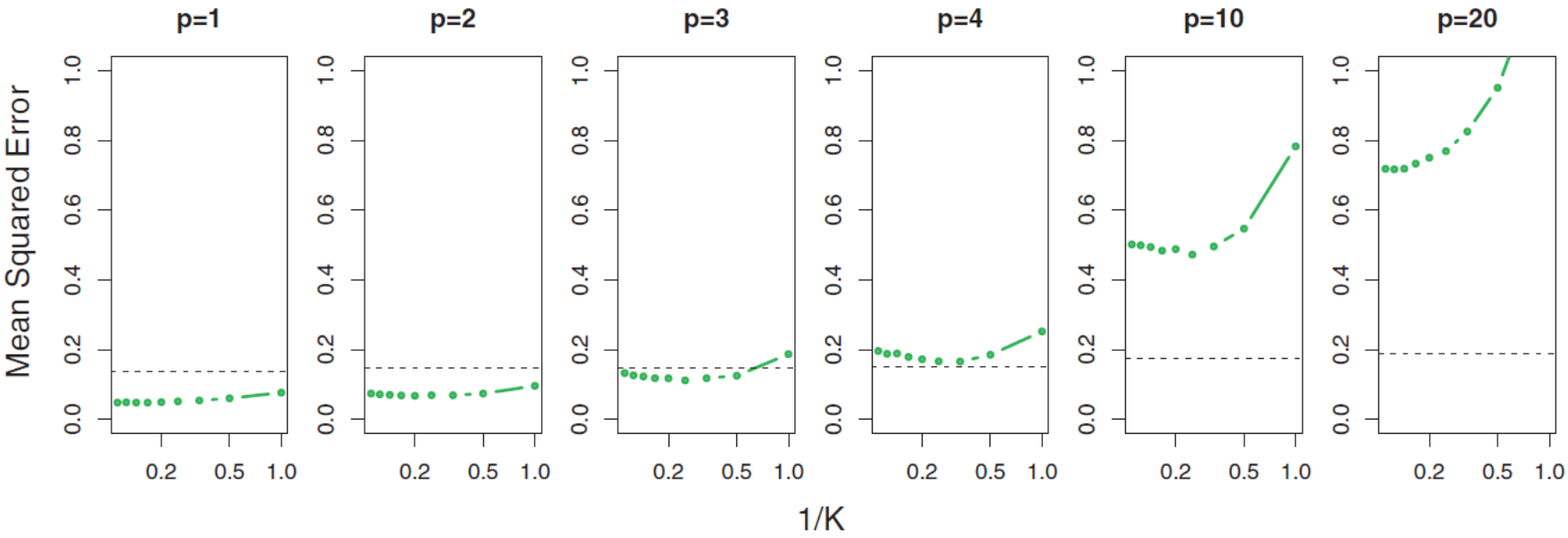



Left: *In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with K* = 1 *(blue) and K* = 9 *(red) are displayed.* Right: *For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of* 1*/K (green) are displayed.*




Left and Right: *As in the top panel, but with a strongly non-linear relationship between X and Y .*

# Comparison of Linear Regression with *K*-Nearest Neighbors




Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non–linear in the first variable, as in the lower panel in, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.