DATASCI 347 Machine Learning

Lecture 2: KNN and Bias-variance decomposition

Ruoxuan Xiong

Suggested reading: ISL Chapter 2

Lecture plan

- K-nearest neighbors (KNN) regression
- Bias-variance decomposition for regression problems (MSE)

K-nearest neighbors regression

- A non-parametric approach
- K is a user-defined constant
  - K is an integer, e.g., $1,2,3,\cdots$
- Given a value for K and a prediction point x_0, $\hat{f}(x_0)$ is the average of the responses of K nearest neighbors

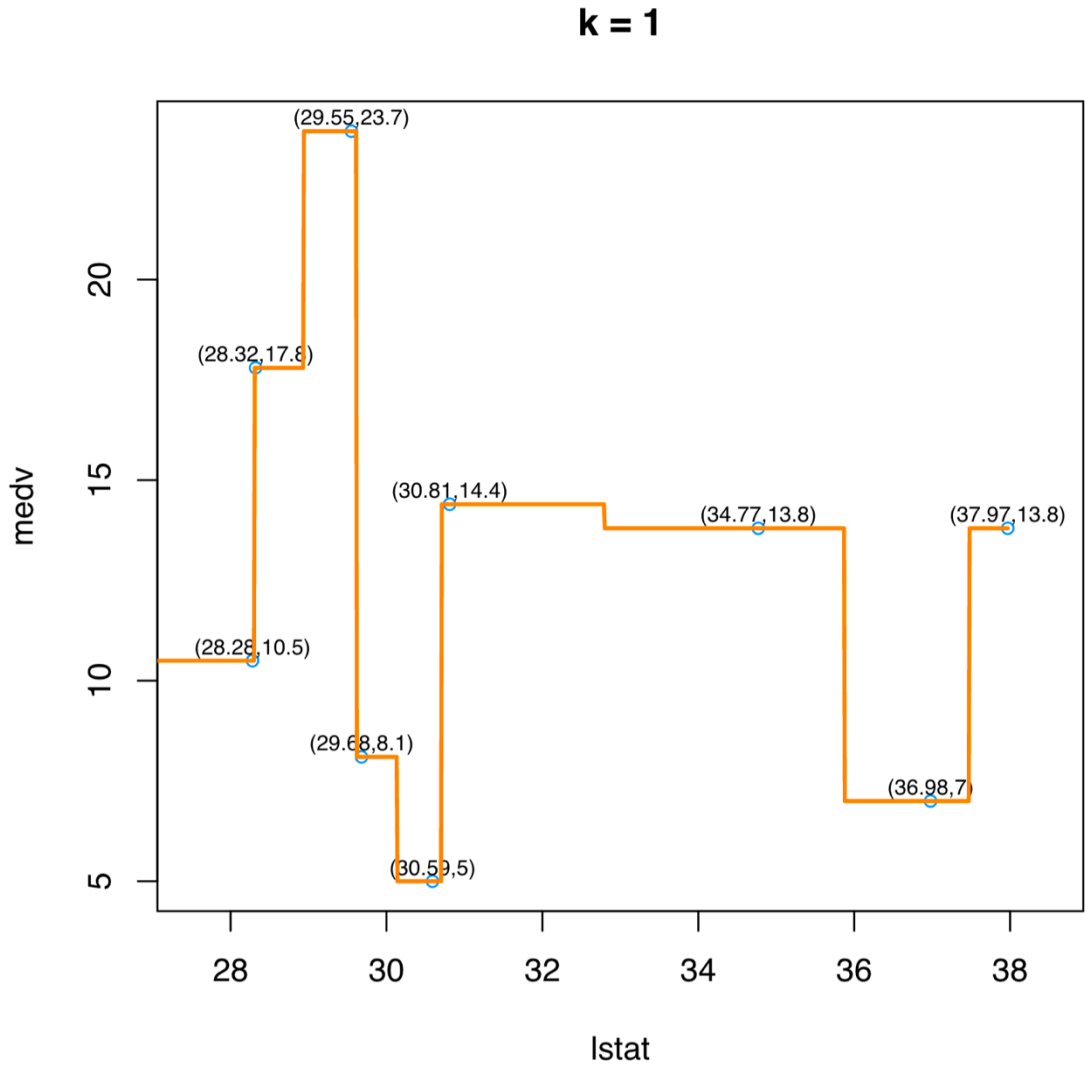
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_K(x_0)} y_i$$

  - $N_K(x_0)$ is the set of K training observations that are closest to x_0

Example: 1-nearest neighbor regression

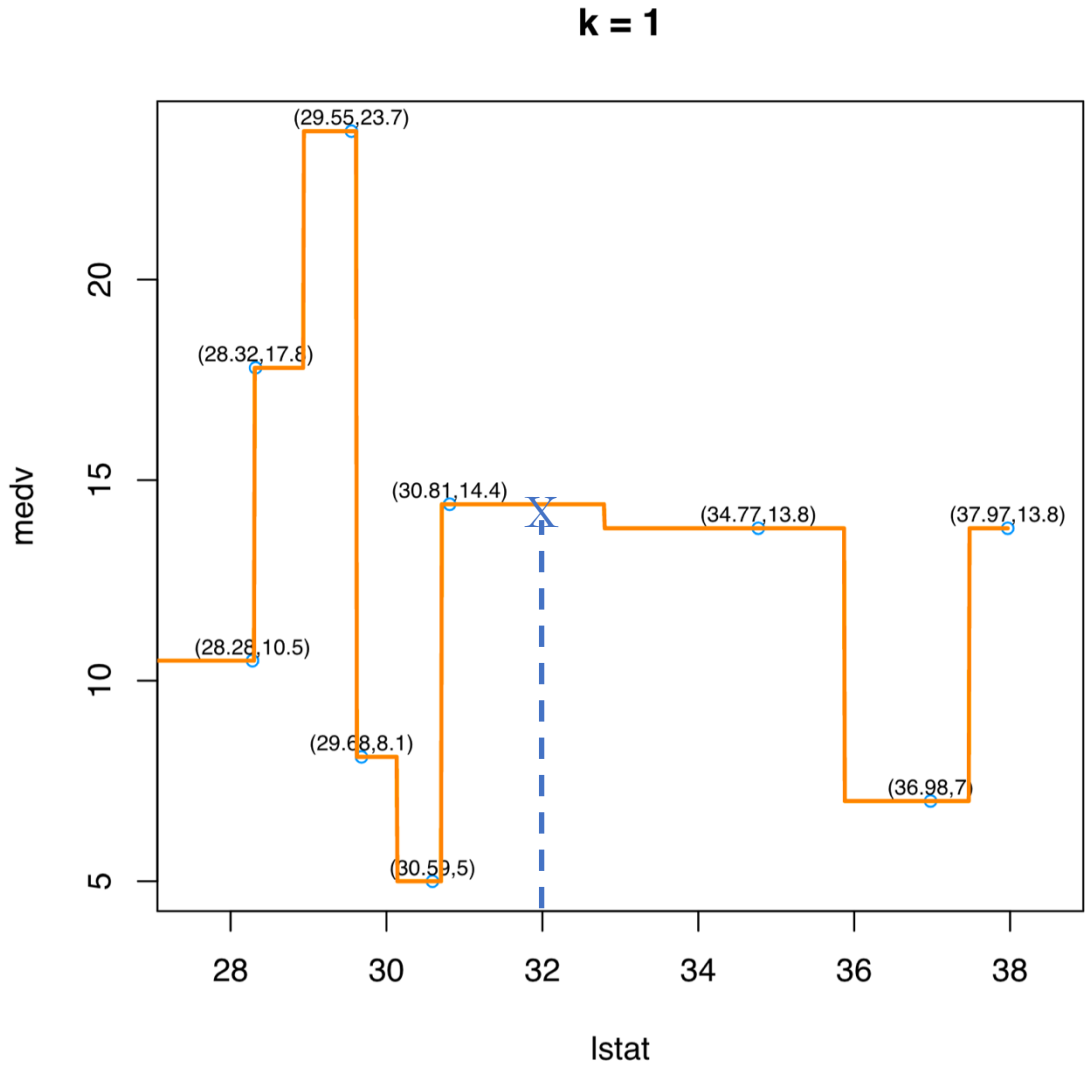
- Prediction of the median house value of a neighbor given the percentage of households with low socioeconomic status (lstat)

- Orange curve: $\hat{f}(x_0)$
  - $\hat{f}(x_0)$ equals to the response of x_0's nearest neighbor
  - $\hat{f}(x_0)$ is a step function



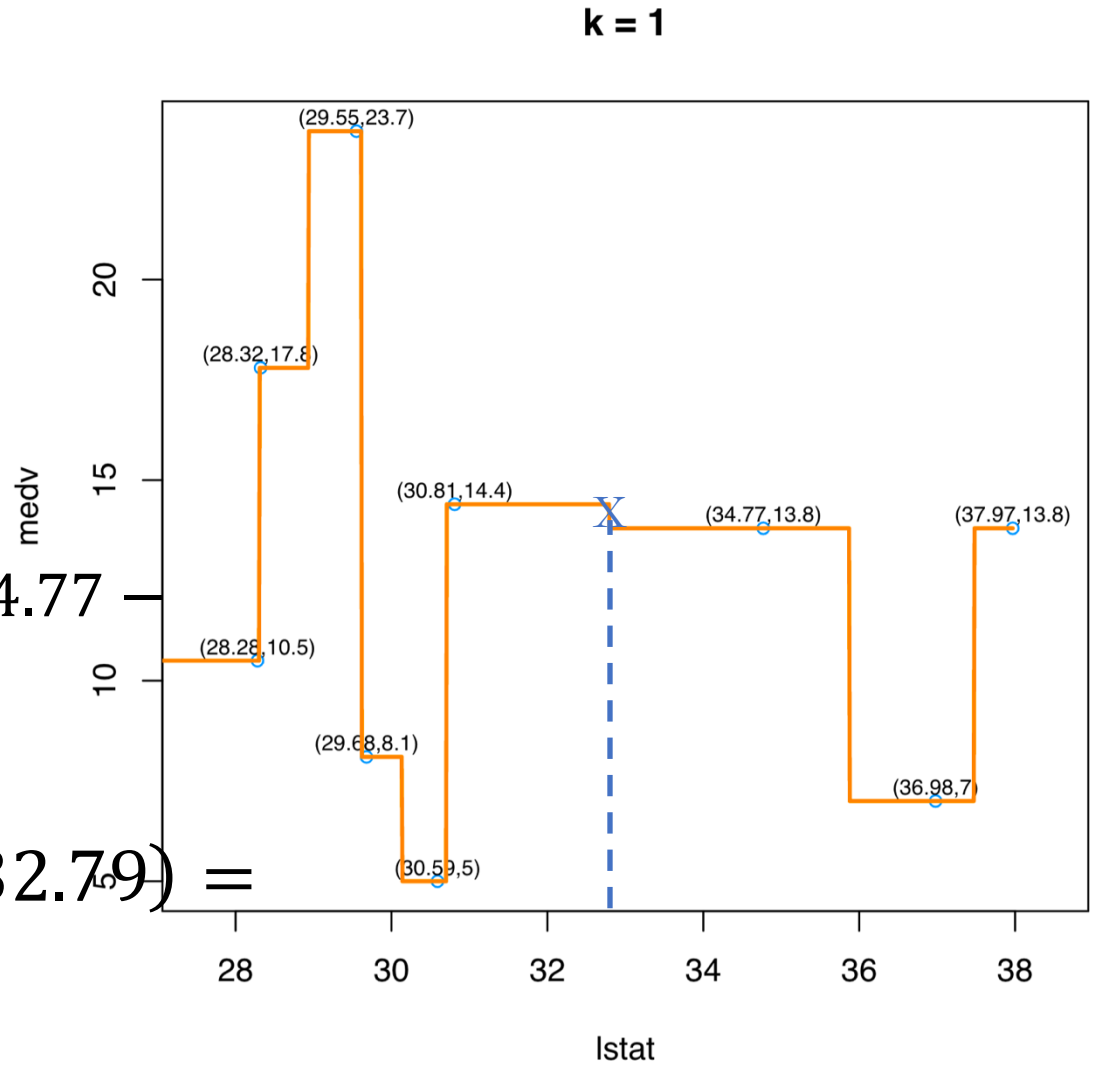
Prediction at $x_0 = 32$ $(K = 1)$

- $x_0 = 32$
- $N_K(x_0) = \{30.81\}$
- $\hat{f}(x_0 = 32) = 14.4$



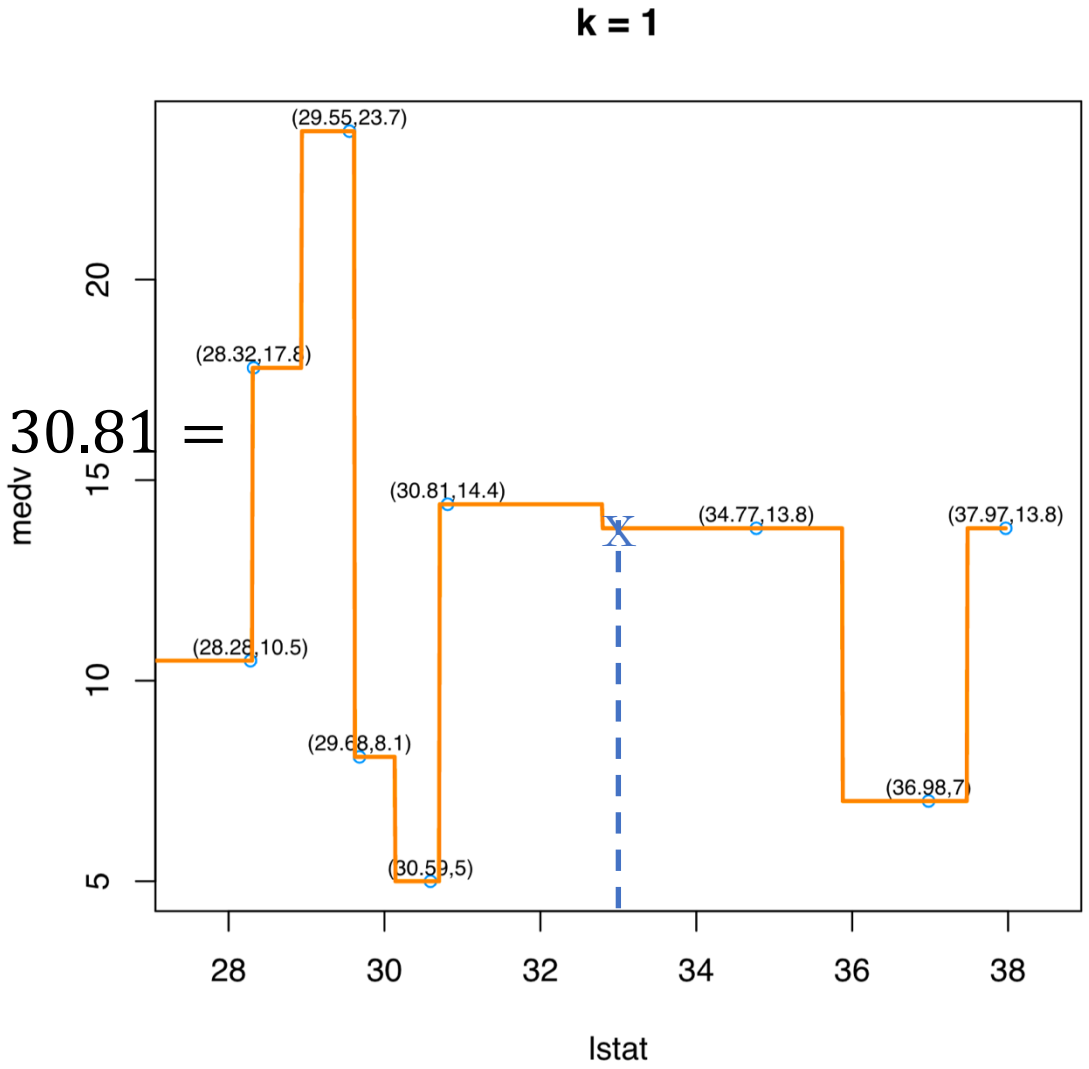
Now let us understand why $\hat{f}(x_0)$ is a step function

- $x_0 = 32.79$
  - It is a switching point!
- $N_K(x_0) = \{30.81\}$ or $N_K(x_0) = \{34.77\}$
  - Note that $32.79 - 30.81 = 1.98 = 34.77 - 32.79$
- $\hat{f}(x_0 = 32.79) = 14.4$ or $\hat{f}(x_0 = 32.79) = 13.8$



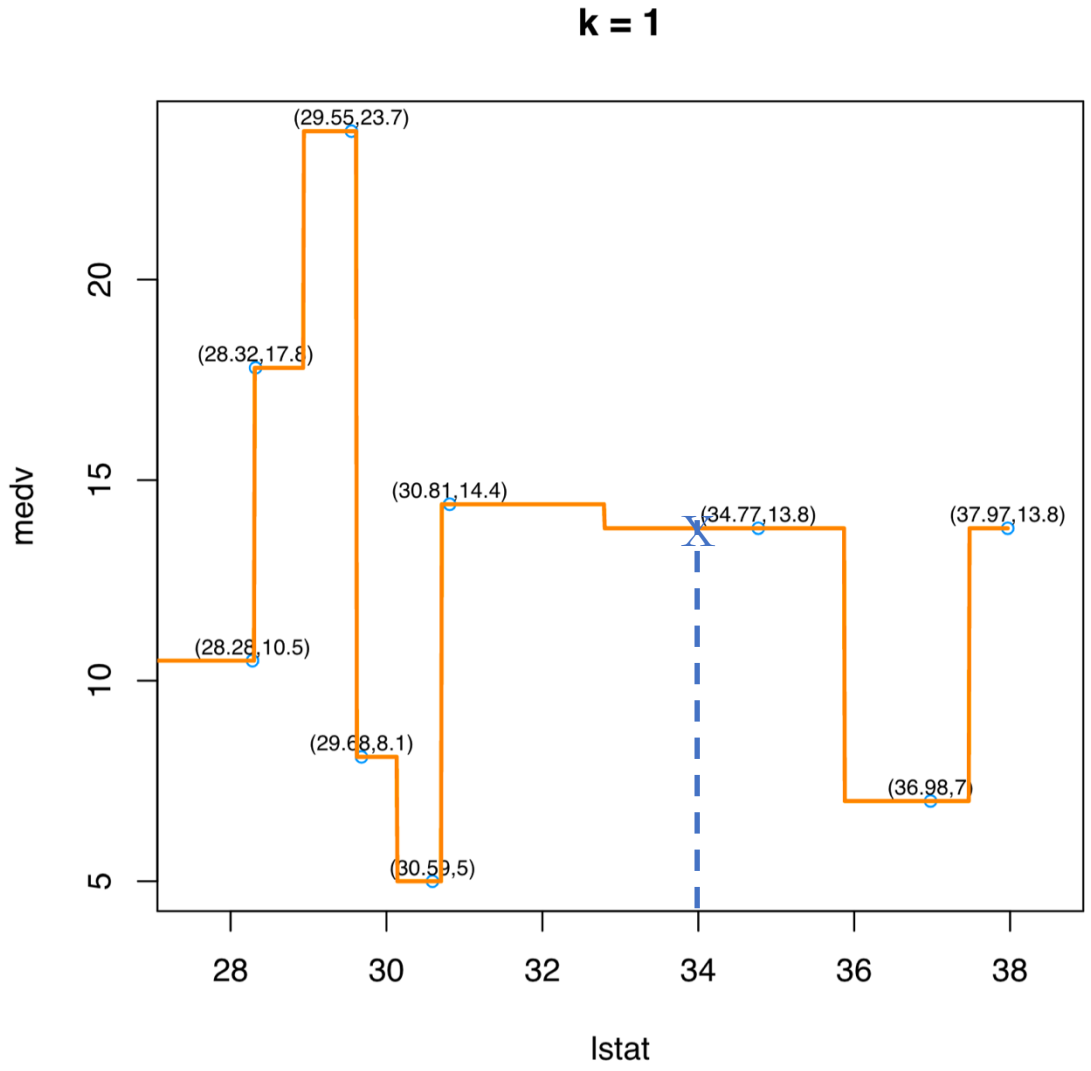
Prediction at $x_0 = 33$ ($K = 1$)

- $x_0 = 33$
- $N_K(x_0) = \{34.77\}$
  - Note that $34.77 - 33 = 1.77 < 33 - 30.81 = 2.19$
- $\hat{f}(x_0 = 33) = 13.8$



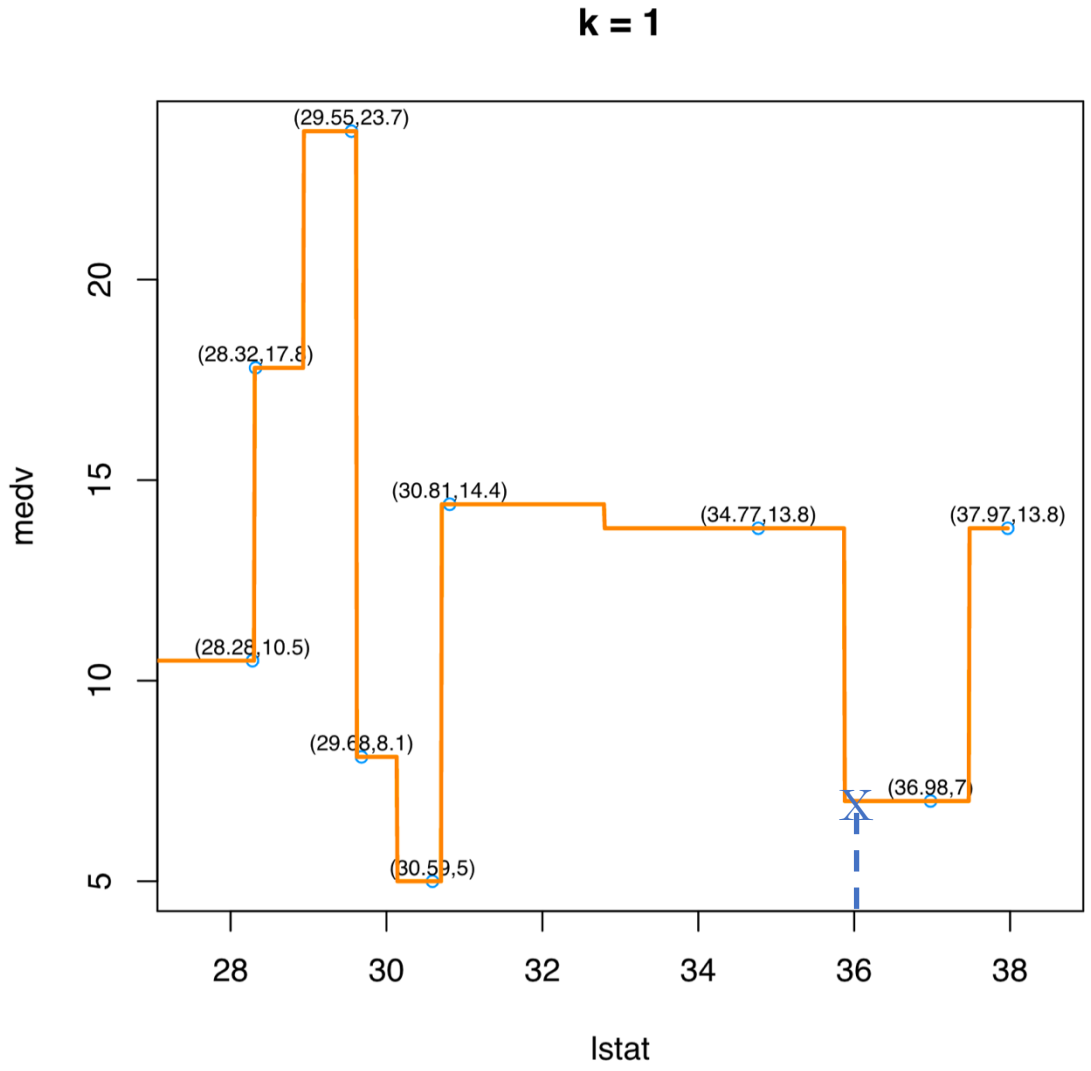
Prediction at $x_0 = 34$ ($K = 1$)

- $x_0 = 34$
- $N_K(x_0) = \{34.77\}$
- $\hat{f}(x_0 = 34) = 13.8$



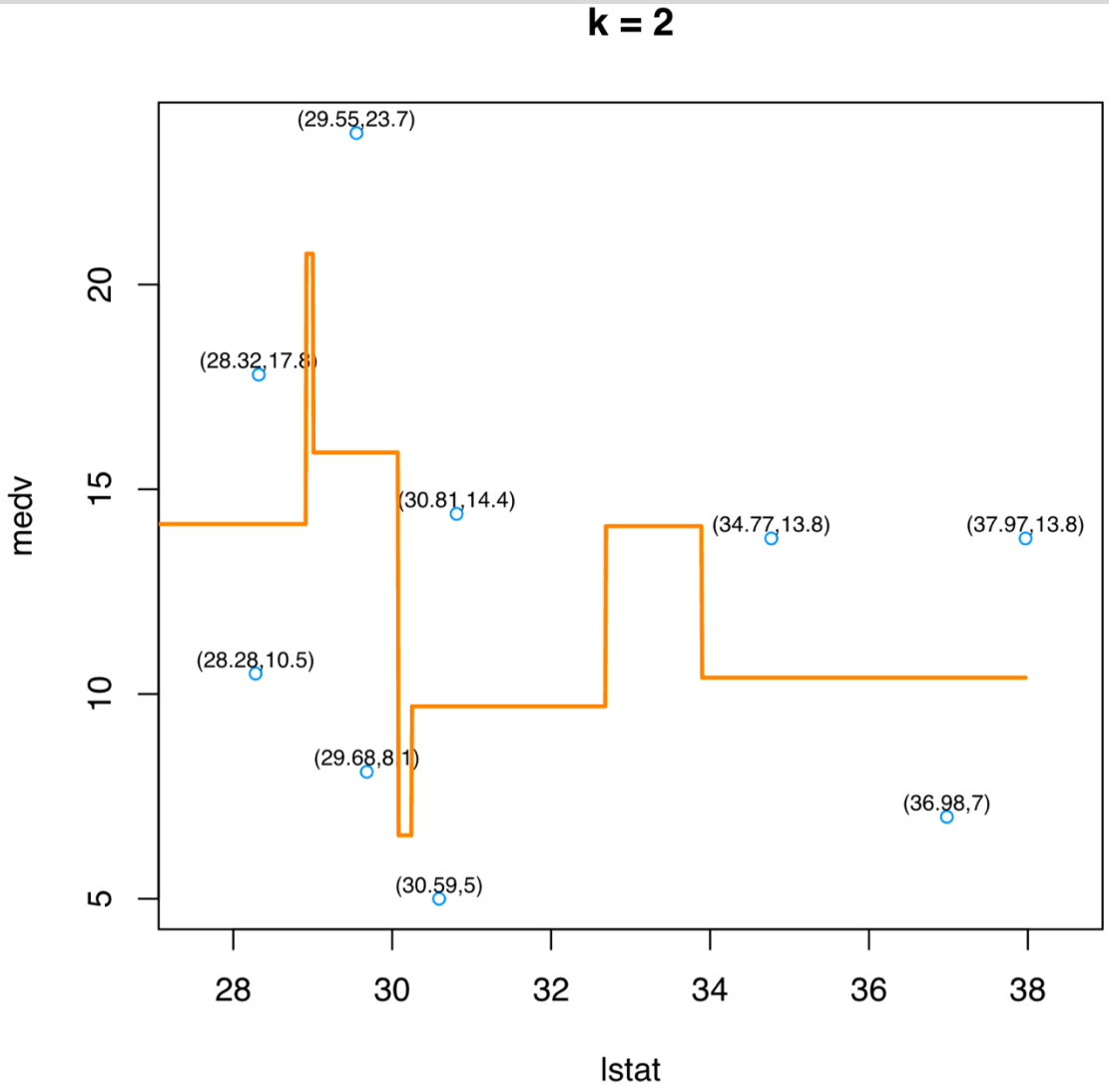
Prediction at $x_0 = 36$ $(K = 1)$

- $x_0 = 36$
- $N_K(x_0) = \{36.98\}$
- $\hat{f}(x_0 = 36) = 7$



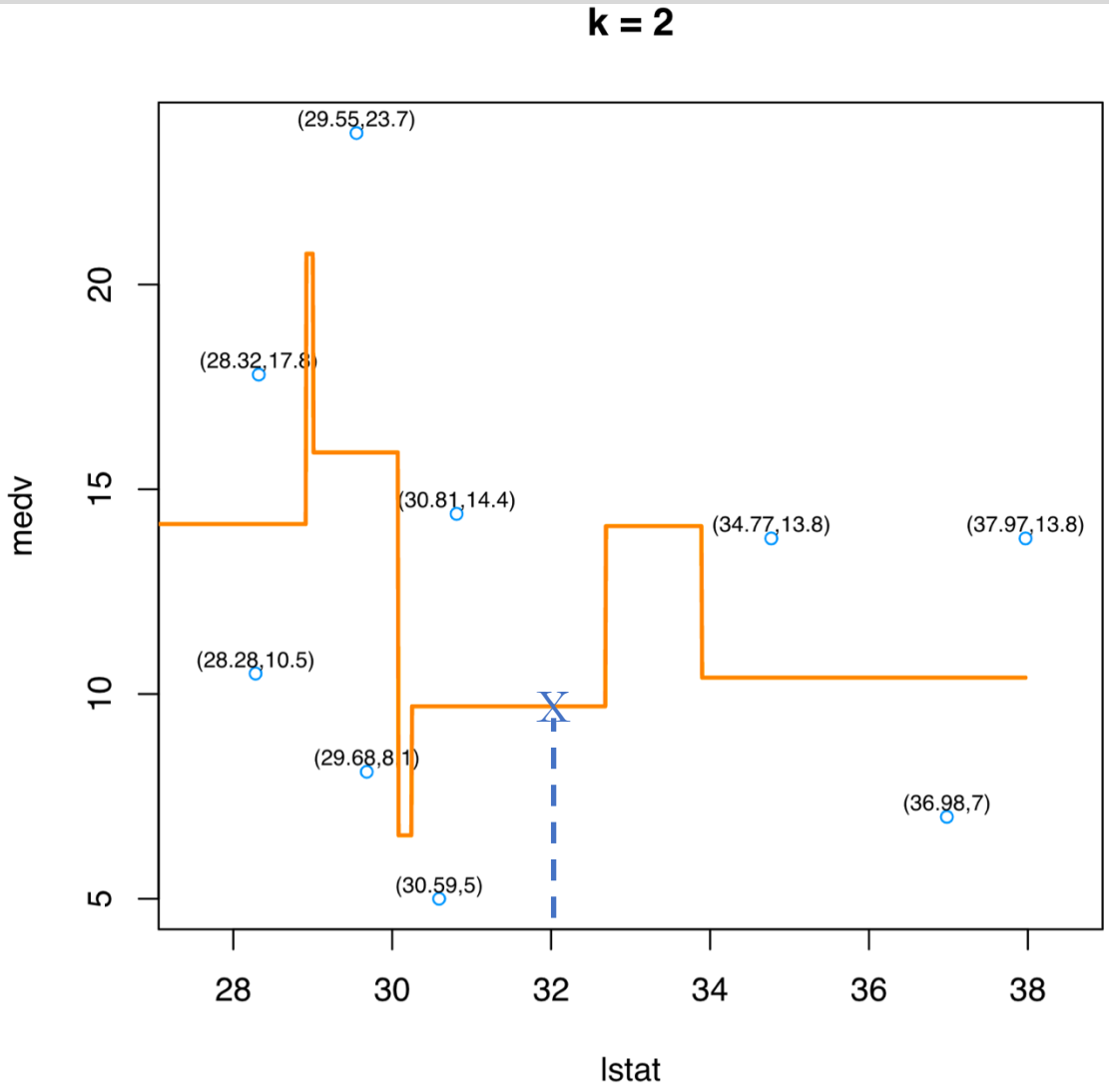
Example: 2-nearest neighbor regression

- $\hat{f}(x_0)$ equals to the average of responses of x_0's 2 nearest neighbors



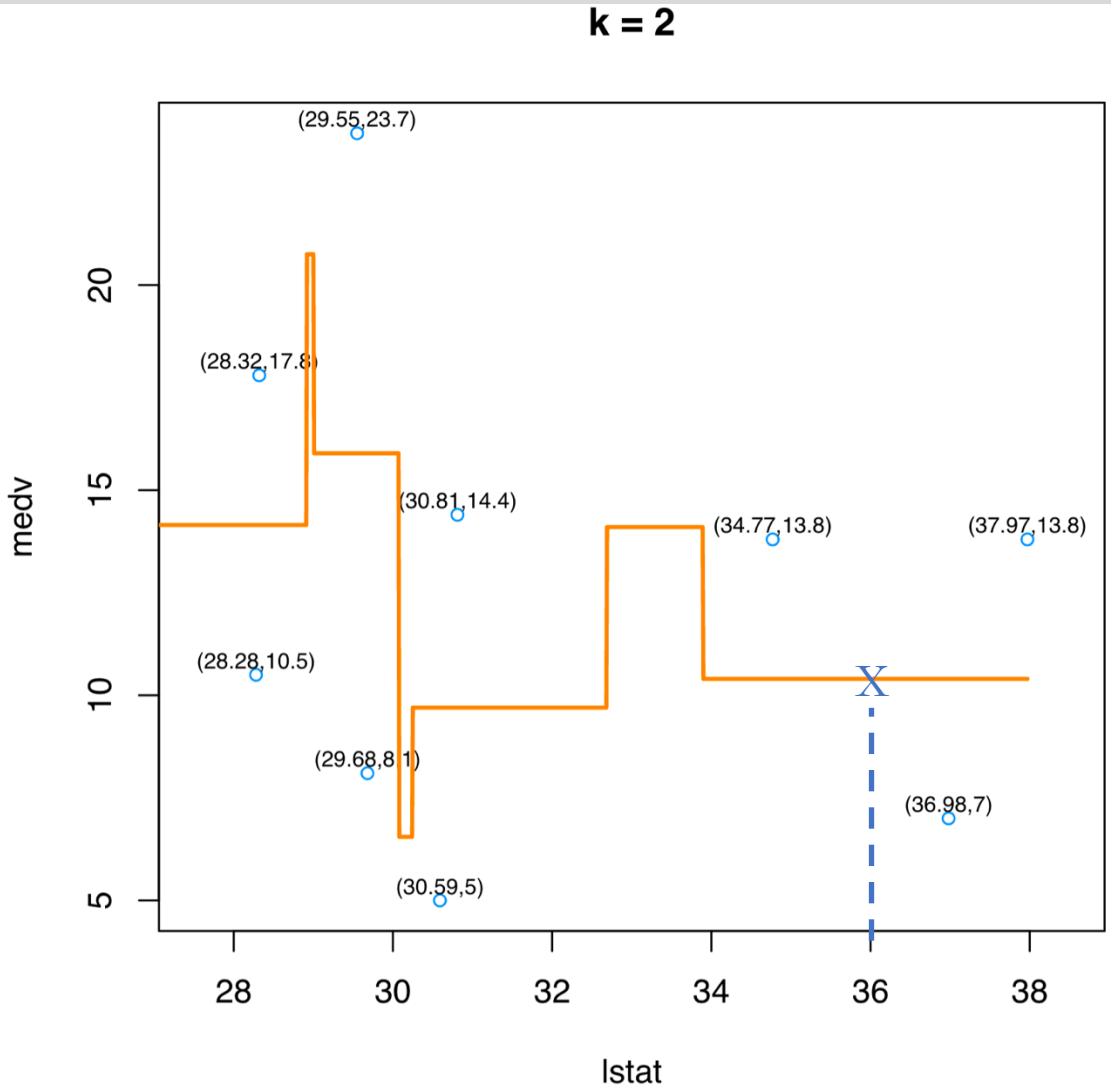
Prediction at $x_0 = 32$ $(K = 2)$

- $x_0 = 32$
- $N_K(x_0) = \{30.59, 30.81\}$
- $\hat{f}(x_0 = 32) = \frac{5+14.4}{2}$



Prediction at $x_0 = 36$ $(K = 2)$

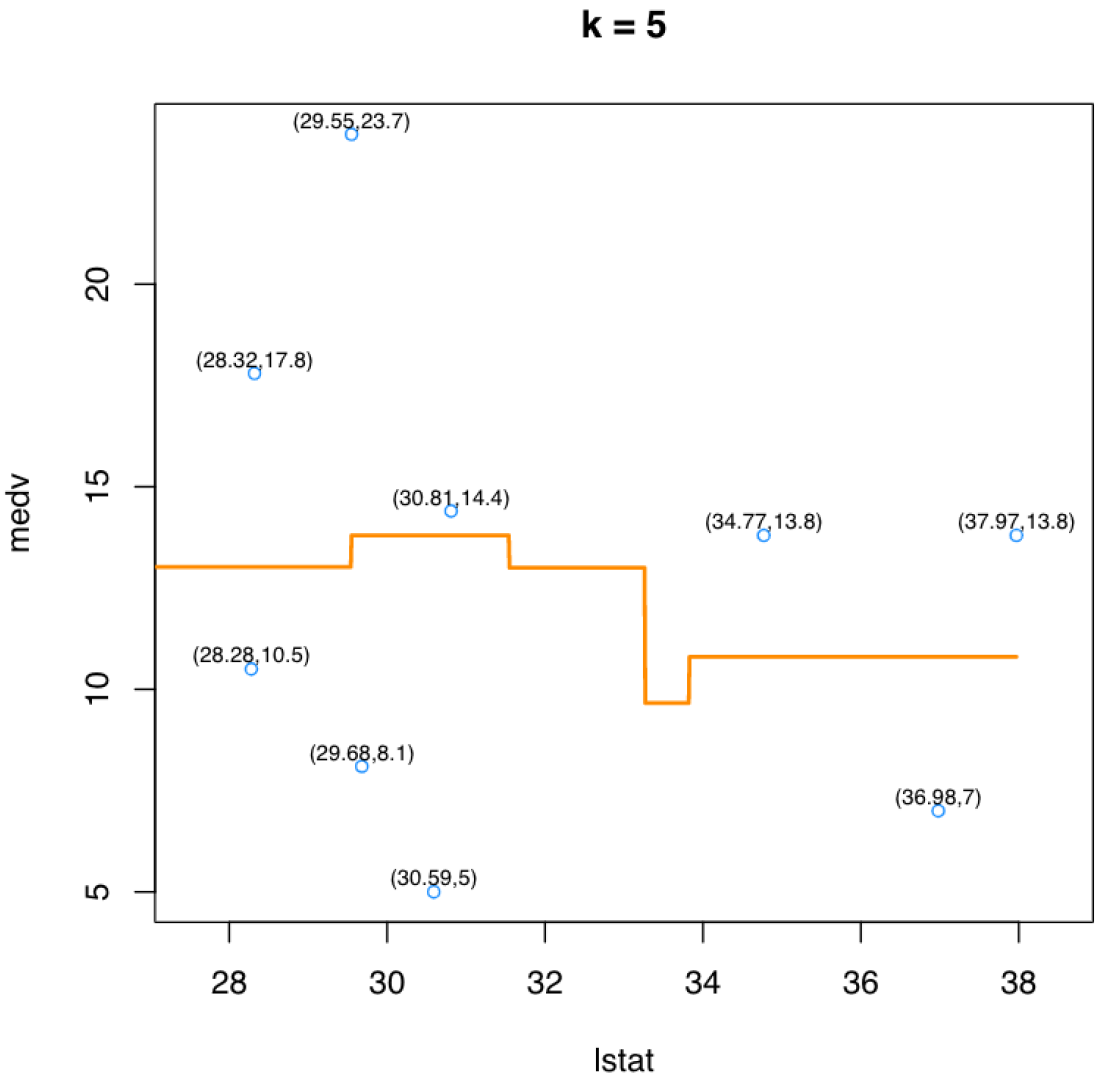
- $x_0 = 36$
- $N_K(x_0) = \{34.77, 36.98\}$
- $\hat{f}(x_0 = 36) = \frac{13.8+7}{2}$



Question

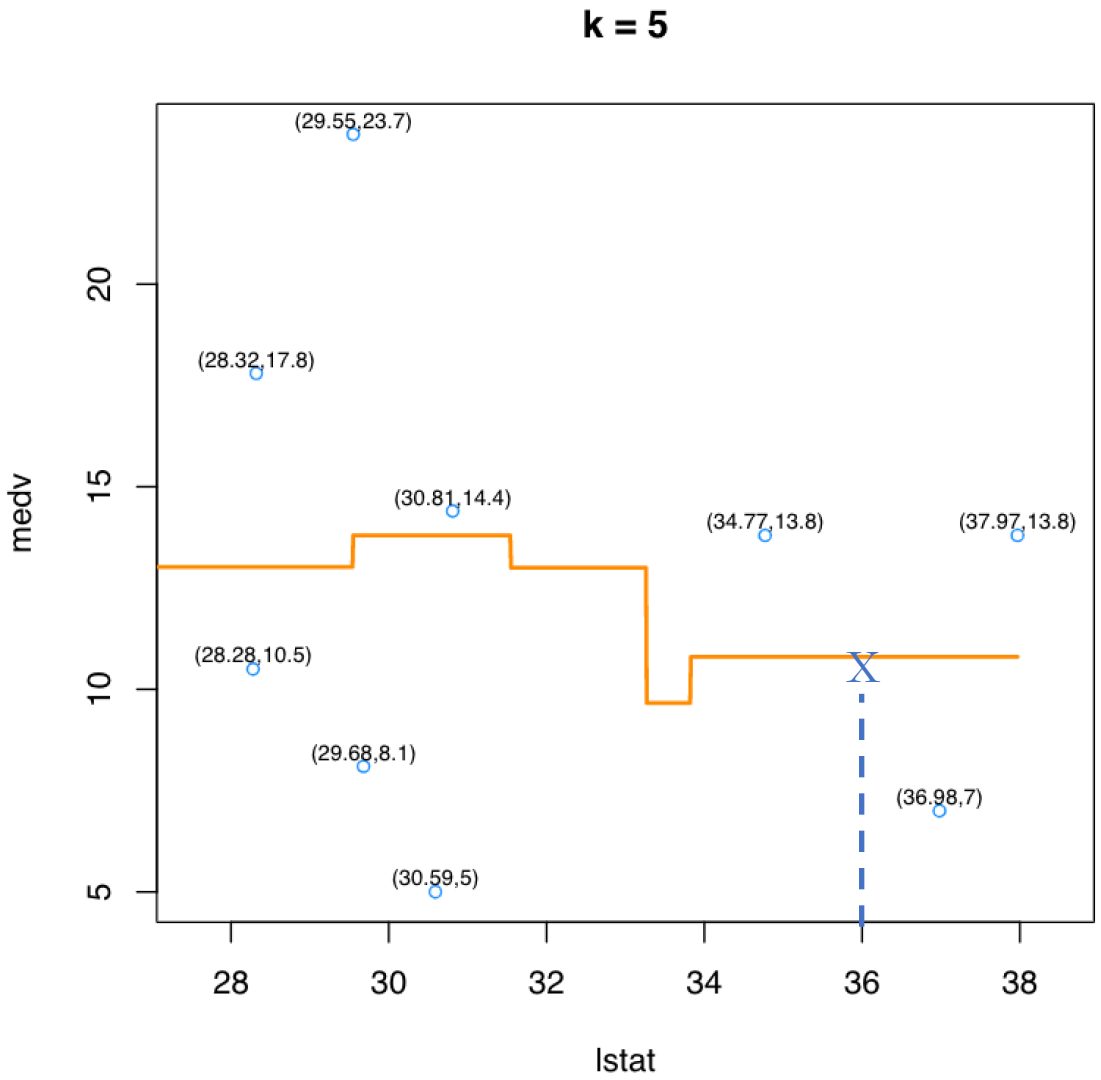
- For $K = 5$, what is $\hat{f}(x_0)$ for $x_0 = 36$?

- $\hat{f}(x_0)$ equals to the average of responses of x_0's 5 nearest neighbors

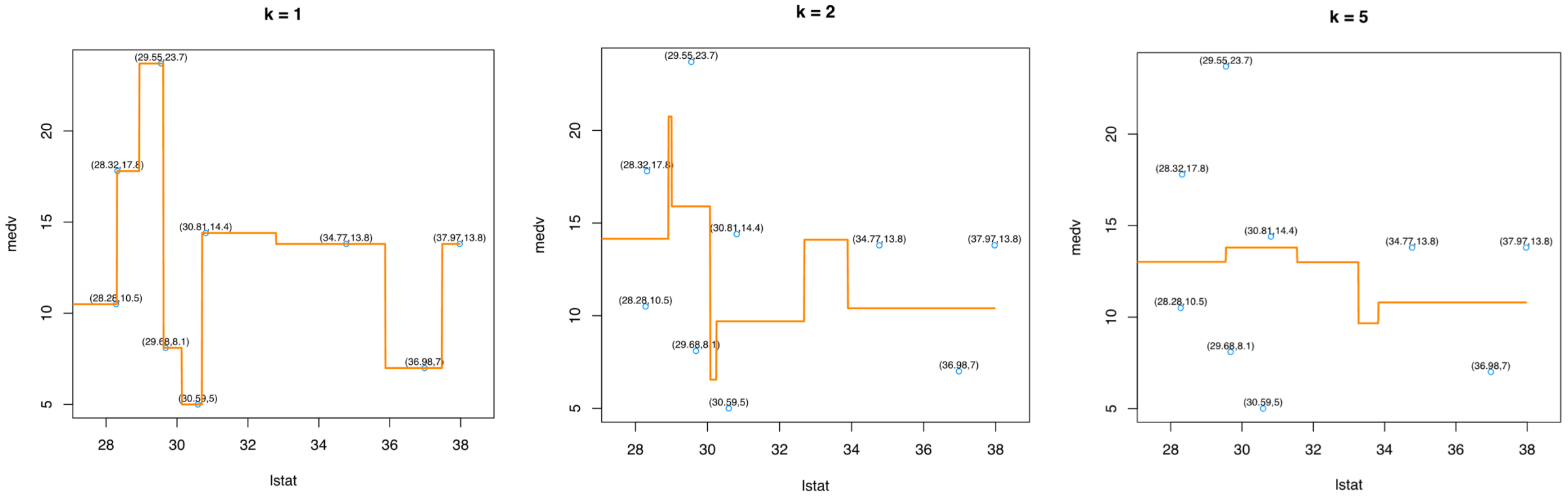


Prediction at $x_0 = 36$ ($K = 5$)

- $x_0 = 36$
- $\hat{f}(x_0)$ equals to the average of responses of x_0's 5 nearest neighbors
  $N_K(x_0)$
  $= \{30.59, 30.81, 34.77, 36.98, 37.97\}$
- $\hat{f}(x_0 = 36) = \frac{5+14.4+13.8+7+13.8}{5}$
- $\hat{f}(x_0)$ is smoother as K increases



$\hat{f}(x_0)$ is smoother for a larger K

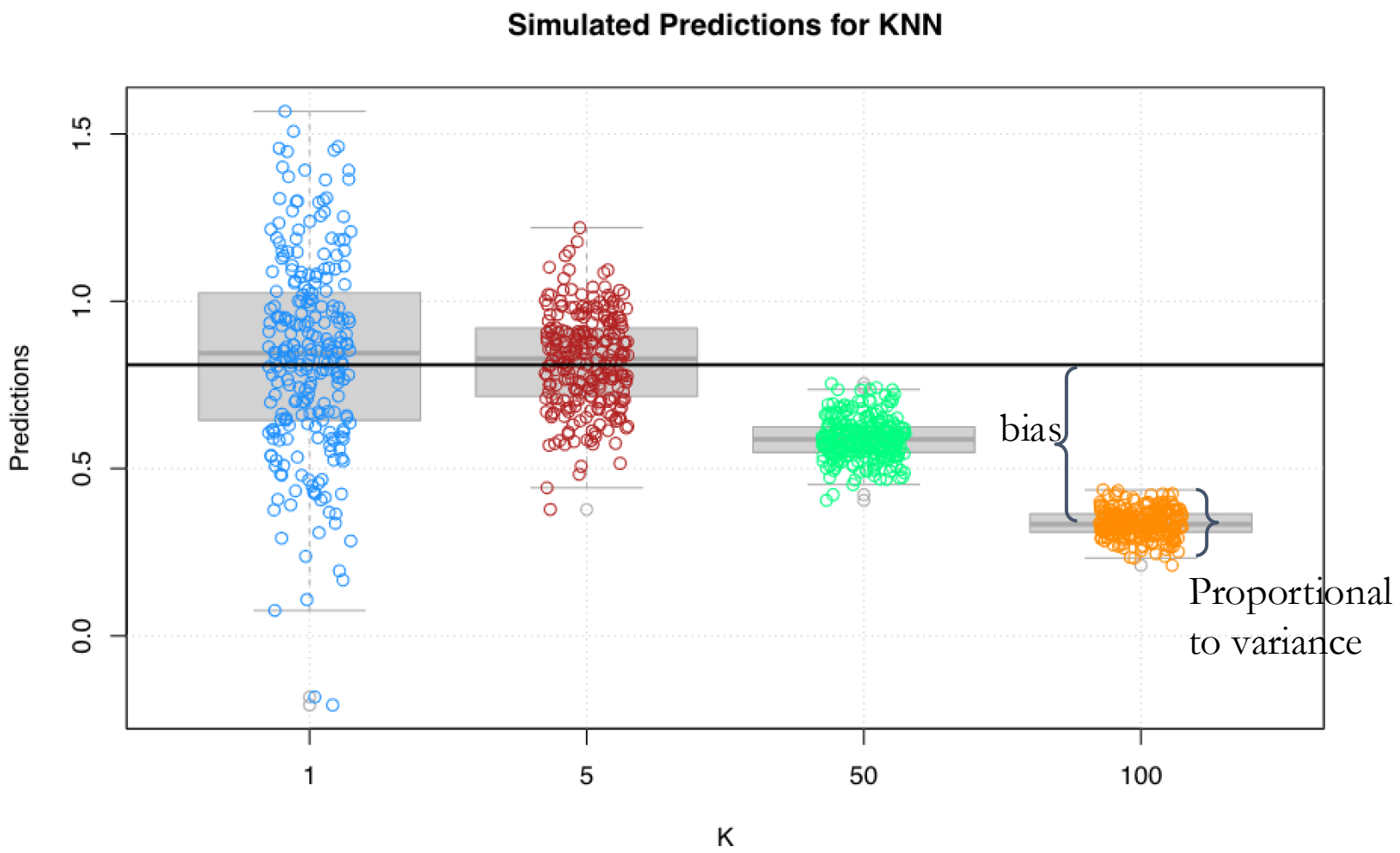


- *Question: Is the model more flexible or less flexible for a larger K?*

Using KNN to approximate a quadratic model

- Train a KNN model to learn the unknown true function $f(x) = x^2$ (x is a scalar). $Y = X^2 + \epsilon, X \in U[0,1], E[\epsilon \mid X] = 0$

- $x_0 = 0.9$
- The truth, $f(x_0 = 0.9) = x_0^2 = 0.81$
- We have 250 datasets
- For each dataset, we fit KNN with $K = 1, 5, 50, 100$, and plot $\hat{f}(x_0 0.9)$
- $\hat{f}$ *depends on training data* $\mathcal{D}$



Each dot = one prediction $\hat{f}(x_0)$ from a different dataset
The **horizontal line** = true value $f(x_0) = 0.81$
The **center of the cloud** = average prediction
The **spread of the cloud** $\propto$ variance of the model

Bias-variance tradeoff

- The square of bias

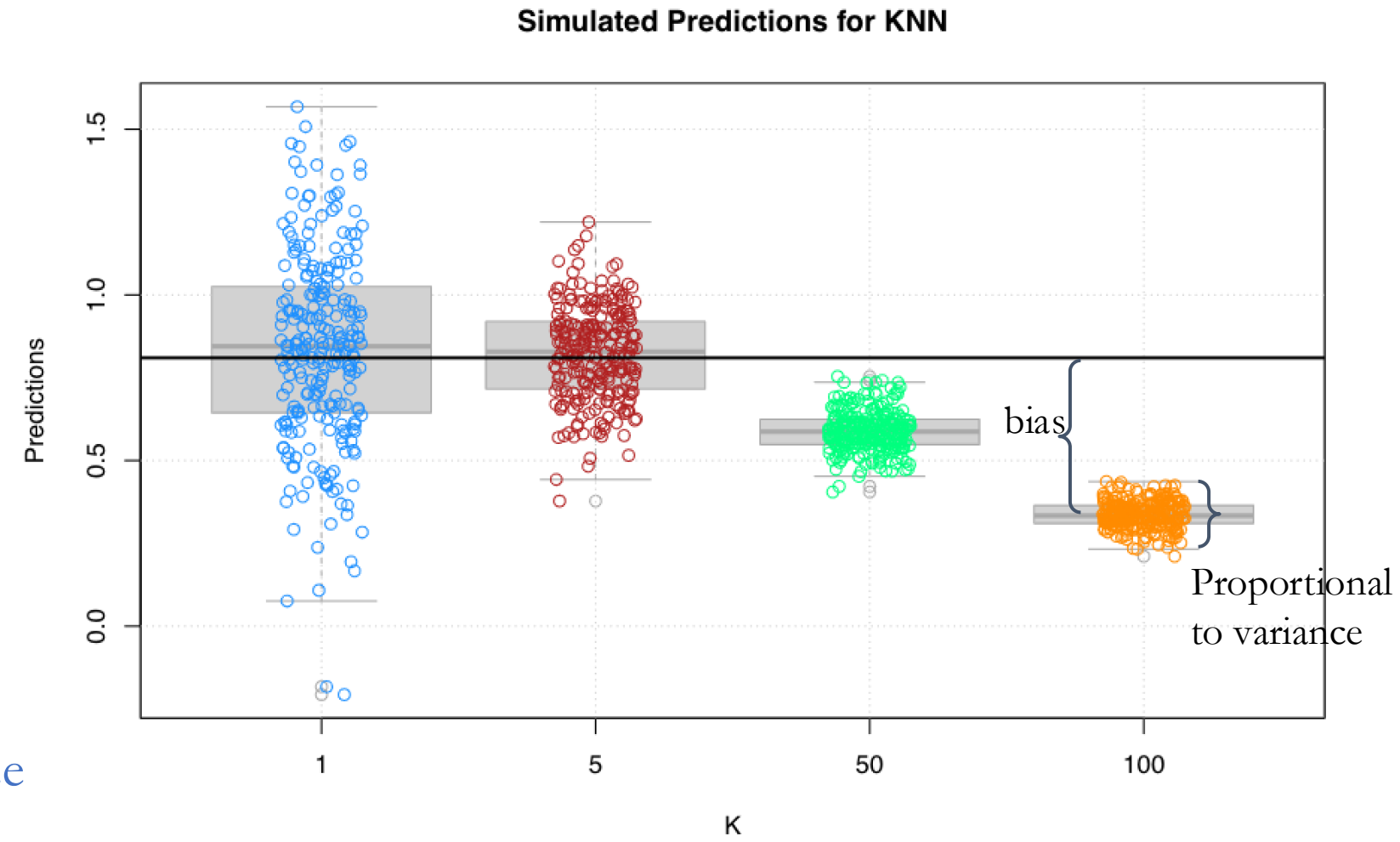
  $\hat{f}_5(x) \approx \hat{f}_1(x) < \hat{f}_{50}(x) < \hat{f}_{100}(x)$

  ➢Increasing K increases bias

- Variance

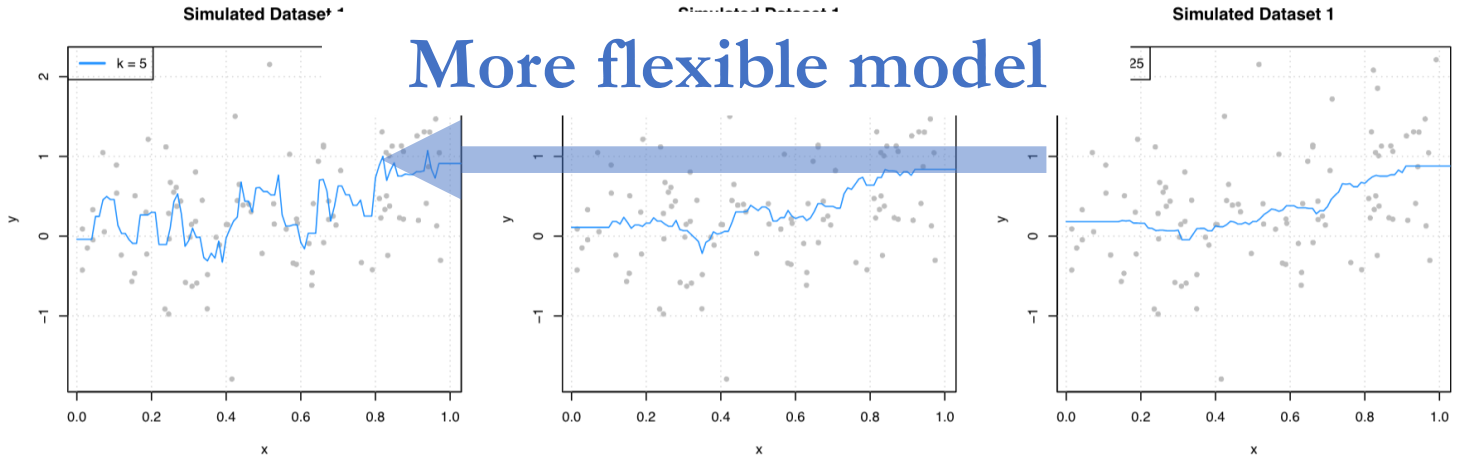
  $\hat{f}_{100}(x) < \hat{f}_{50}(x) < \hat{f}_5(x) < \hat{f}_1(x)$

  ➢Increasing K reduces variance

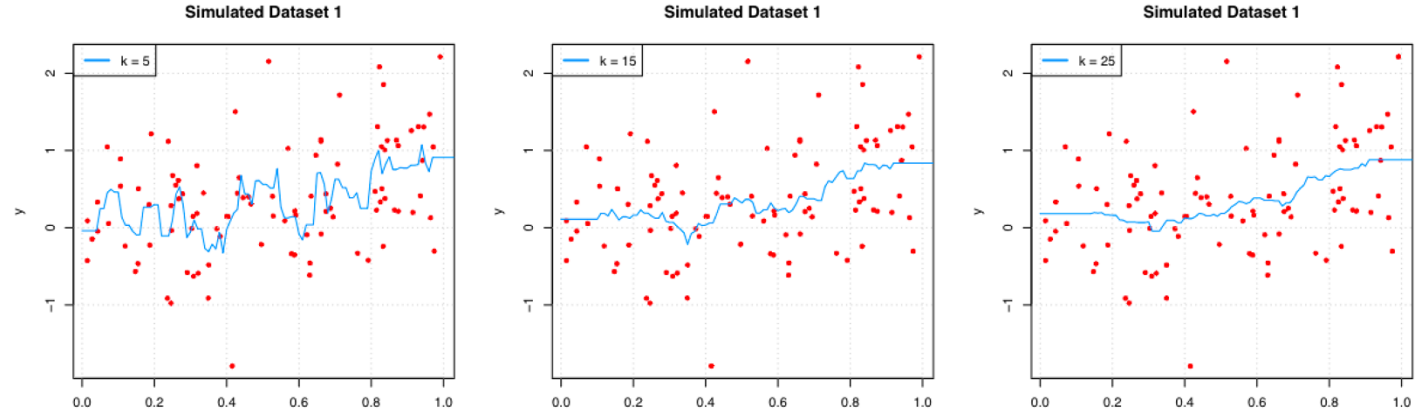


At test point x_0, bias is $f(x_0) - E\left[\hat{f}(x_0)\right]$ and variance is $\mathrm{Var}\left(\hat{f}(x_0)\right)$

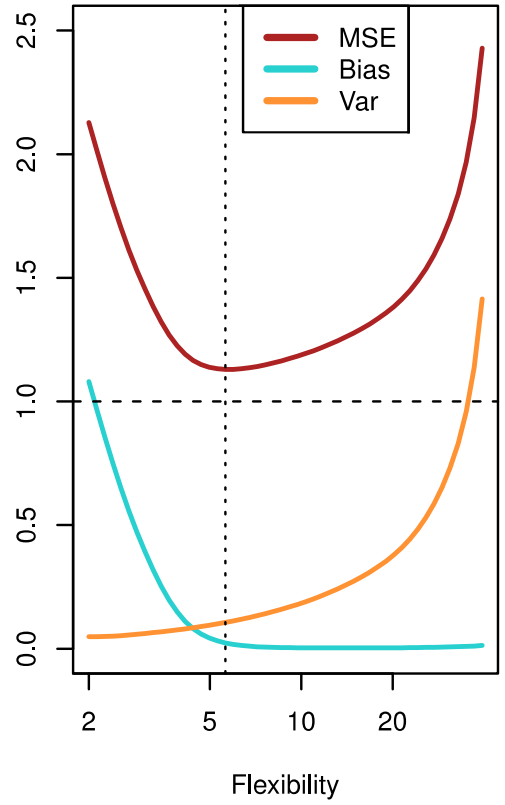
Bias and variance vary with model flexibility



Training MSE = 0.354 Training MSE = 0.412 Training MSE = 0.429



Test MSE = 0.564 Test MSE = 0.507 Test MSE = 0.524



Flexibility proportional to $1/K$

Bias-variance decomposition of test MSE

- Suppose the true regression function is f
- The response satisfies $Y = f(X) + \varepsilon$, with $\mathrm{E}[\varepsilon \mid X] = 0$
- Let the training dataset be $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$
- The fitted model $\hat{f}$ *depends on training data* $\mathcal{D}$
  - If we change the training data $\mathcal{D}$, we get a different estimate $\hat{f}$
- Prediction at a fixed test point x_0
  - The MSE at x_0 can be decomposed as

$$\mathrm{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

Expected value over $\mathcal{D}$ and $Y|X$

$$\mathrm{V}[\varepsilon \mid X] = \sigma_\varepsilon^2$$

Decomposition of MSE

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]} + \underbrace{V_{Y|X}[Y \mid X = x_0]}$$

Expected value over $Y|X$ and $\mathcal{D}$ | Reducible error | Irreducible error

- The MSE at x_0 is the **expected prediction error** for
  - a **new response** Y_0
  - at a **fixed input** $X_0 = x_0$
  - using a **model** $\hat{f}$ trained on random data $\mathcal{D}$

- Two sources of *randomness*
  - Noise in the outcome: Y_0 is random because $Y = f(X) + \varepsilon$ and ε is random
  - Randomness in fitted model $\hat{f}$: $\hat{f}$ is random because it depends on randomly sampled $\mathcal{D}$

Irreducible error

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

  - Irreducible error: the variance of Y given that $X = x_0$
    - If $Y = f(X) + \varepsilon$ with $\text{E}[\varepsilon \mid X] = 0$ and $\text{V}[\varepsilon \mid X] = \sigma_\varepsilon^2$, then $V_{Y|X}[Y \mid X = x_0] = \sigma_\varepsilon^2$
    - Irreducible error can not be reduced for any $\hat{f}$

Reducible error

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

Expected value over $Y|X$ and $\mathcal{D}$

  - Reducible error: the expected squared error by using $\hat{f}(x_0)$ to estimate $f(x_0)$
    - The randomness comes from training data $\mathcal{D}$ used to obtain $\hat{f}$ (both f and x_0 are fixed)
    - Reducible error can be reduced by using a better $\hat{f}$

Bias-variance decomposition of reducible error

- *Reducible error* can be decomposed as the *squared bias* and *variance*

$$E_{\mathcal{D}}\left[\left(f(x_0)-\hat{f}(x_0)\right)^2\right] = \underbrace{\left(f(x_0)-E_{\mathcal{D}}[\hat{f}(x_0)]\right)^2}_{\text{bias}^2(\hat{f}(x_0))} + \underbrace{E_{\mathcal{D}}\left[\left(\hat{f}(x_0)-E_{\mathcal{D}}[\hat{f}(x_0)]\right)^2\right]}_{\text{var}(\hat{f}(x_0))}$$

- *Take home exercise: Prove this decomposition*
- Hint: Use the property

$$E_{\mathcal{D}}\left[\left(f(x_0)-\hat{f}(x_0)\right)^2\right] = E_{\mathcal{D}}\left[\left(\underbrace{f(x_0)-E_{\mathcal{D}}[\hat{f}(x_0)]}_{A} + \underbrace{E_{\mathcal{D}}[\hat{f}(x_0)]-\hat{f}(x_0)}_{B}\right)^2\right]$$

$$E_{\mathcal{D}}[(A+B)^2] = A^2 + 2A\underbrace{E_{\mathcal{D}}[B]}_{=0} + E_{\mathcal{D}}[B^2] = A^2 + E_{\mathcal{D}}[B^2]$$

(A is nonrandom and $E_{\mathcal{D}}[A^2] = A^2$)

Summing up the decomposition

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{\text{bias}^2\left(\hat{f}(x_0)\right) + \text{var}\left(\hat{f}(x_0)\right)}_{\text{Reducible error}} + \underbrace{\sigma_\varepsilon^2}_{\text{Irreducible error}}$$

Linear regression vs K-nearest neighbors

- KNN is only better when the function f is far from linear (in which case linear model is *misspecified*)

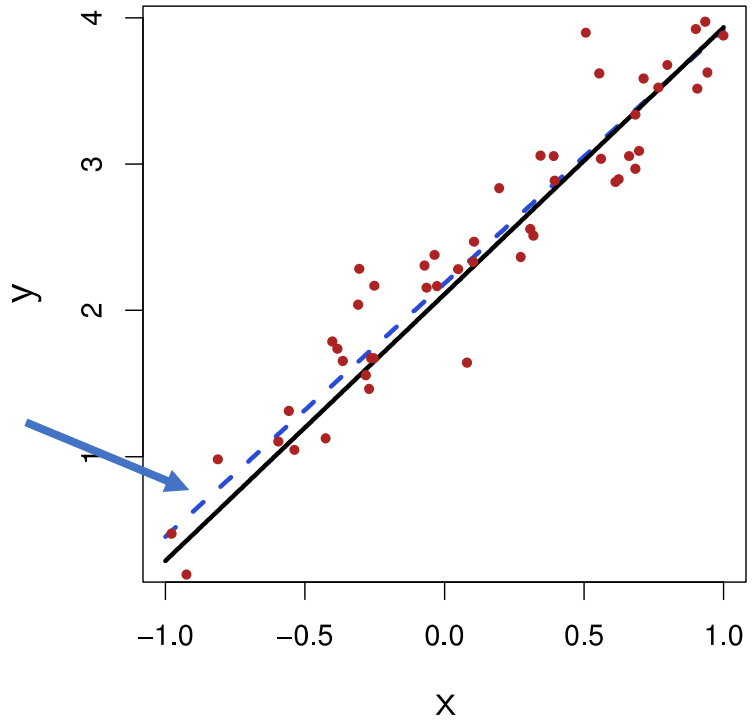
- When n is not much larger than p, even if f is nonlinear, linear regression can outperform KNN

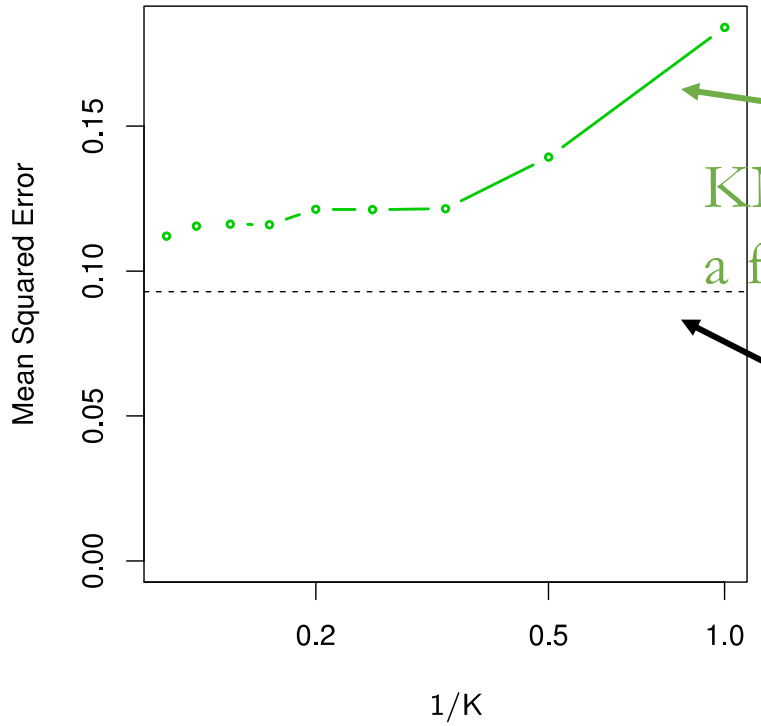
- KNN has smaller bias, but this comes at a price of higher variance

Linear models can dominate KNN

- Truth is linear, plot of test MSE vs. $1/K$ shows KNN worse than linear regression.

linear regression





KNN test MSE as a function of $1/K$

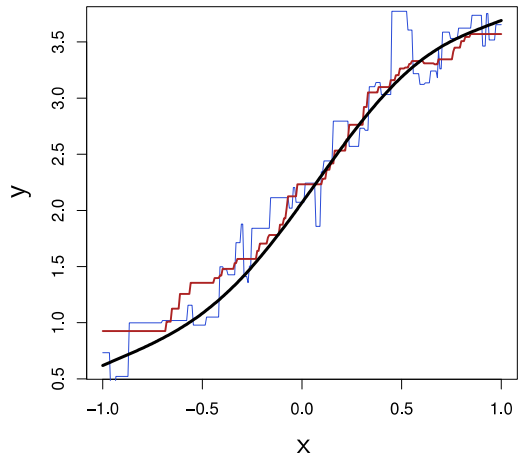
linear regression test MSE

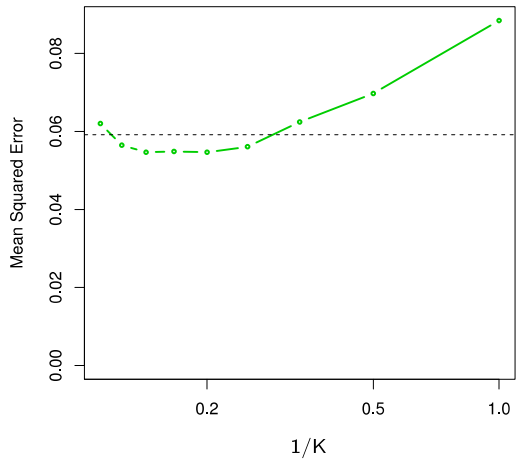
Increasing deviations from linearity

Slightly nonlinear relationship between X and Y

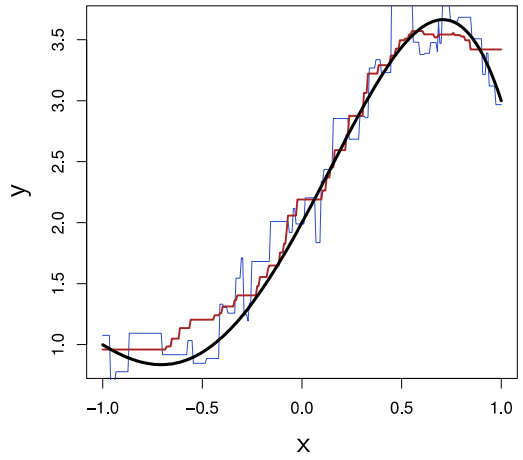
Blue: $K = 1$

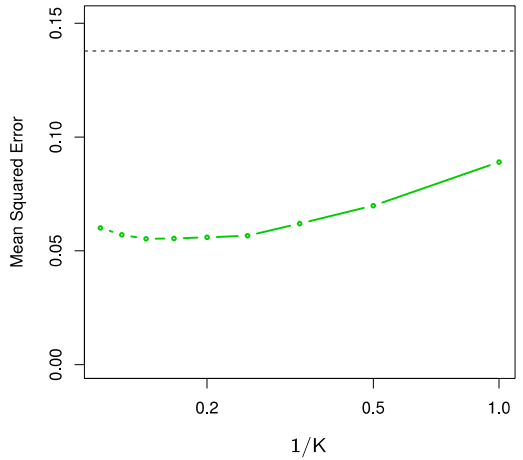
Red: $K = 9$





Strongly nonlinear relationship between X and Y





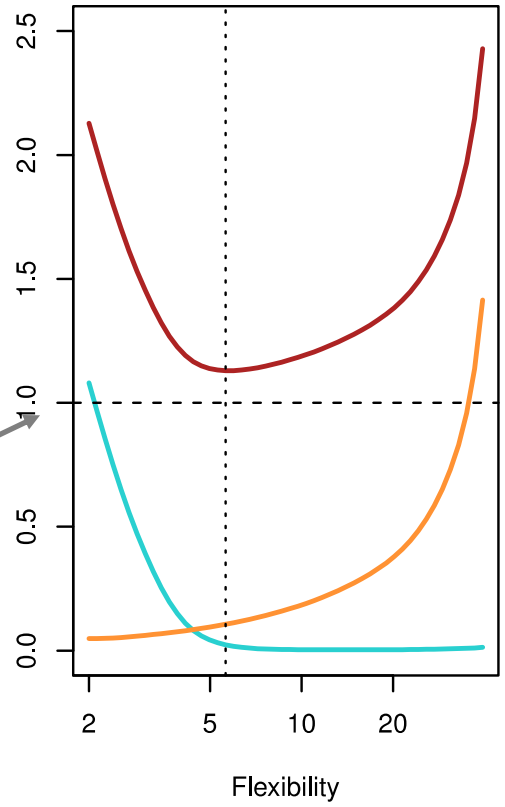
Bias-variance tradeoff under various SNR regimes

$Y = f(X) + \varepsilon$
$\mathrm{Var}(\varepsilon) = 1$

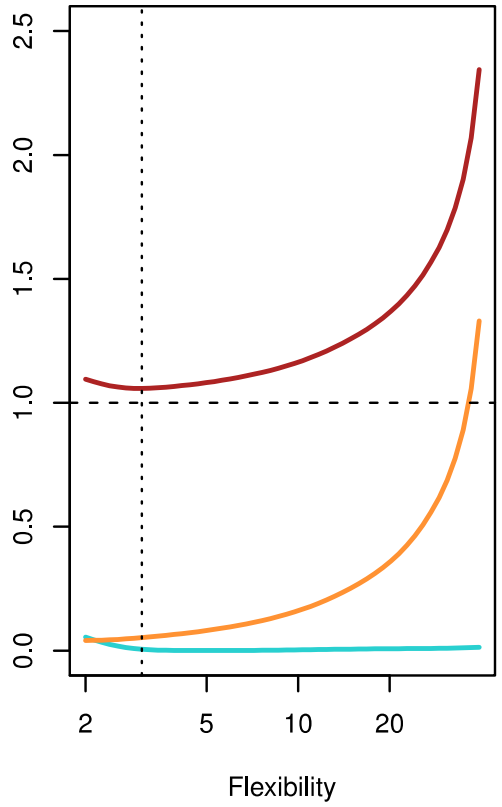
SNR: Signal-to-noise ratio
Signal measured by $\mathrm{Var}(f(X))$
Noise measured by $\mathrm{Var}(\varepsilon)$

Irreducible error

f is complicated, low SNR



f is simple, low SNR



f is complicated, high SNR