Ch3 Regression to Machine Learning

Contents

Textbook: James et al. ISLR 2ed.

3 Subsection

[ToC]

A.1 Statistical Learning

- General Model

$$Y = f(X) + \epsilon$$

- We don't want to assume that $f(X)$ is linear function.
- Two types of motivation:
  - Model Estimation
  - Prediction
- Pattern recognition

A.2 How do we find 'overall pattern'? - Inference

- Want to understand the relationship between X and Y
- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

A.3 How do we find 'overall pattern'? - Prediction

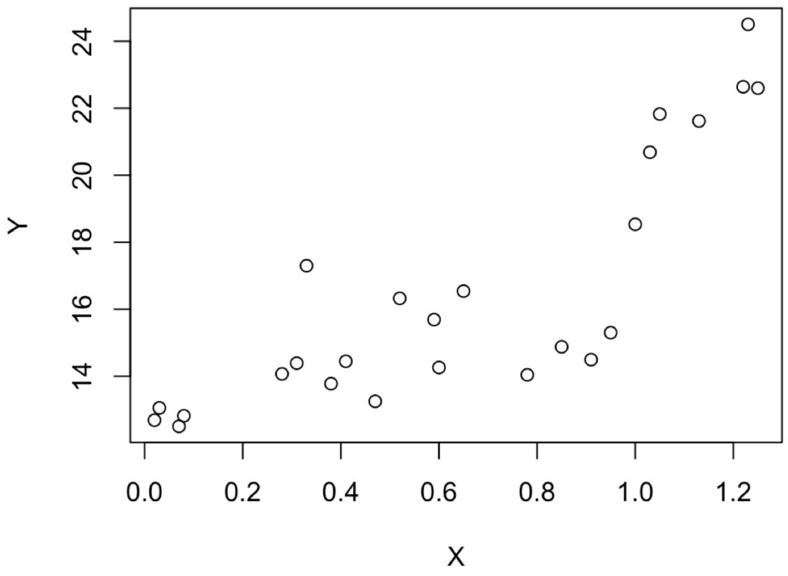
- Want to guess the next Y as accurate as possible

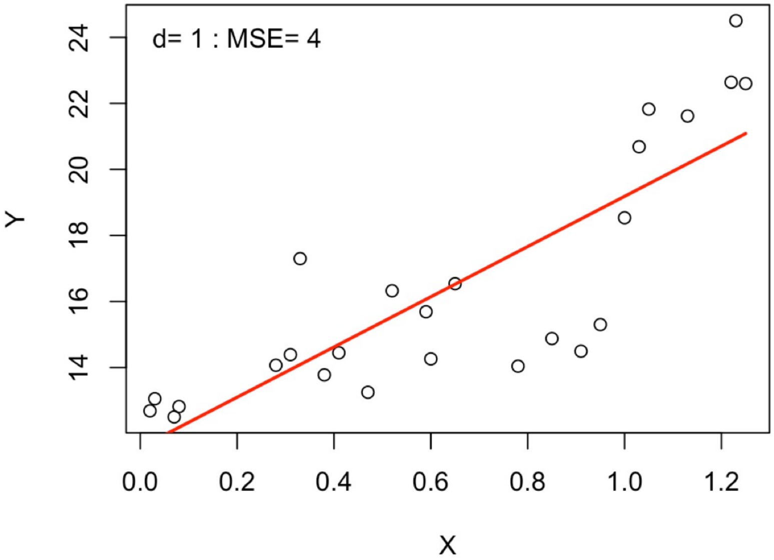
$$\hat{Y} = \hat{f}(X)$$

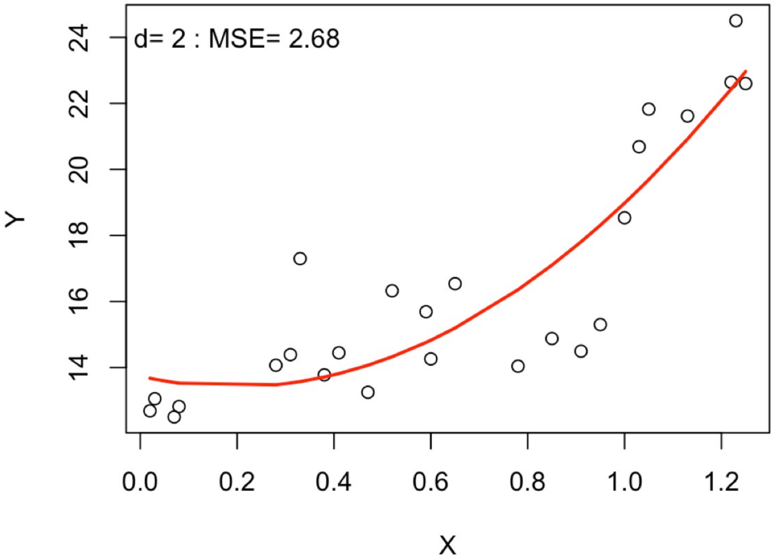
- f can be a black box
- reducible error and irreducible error in prediction
- Want to reduce prediction Mean Squared Error:

$$MSE = E(Y - \hat{Y})^2 = E(Y - \hat{f}(X))^2$$

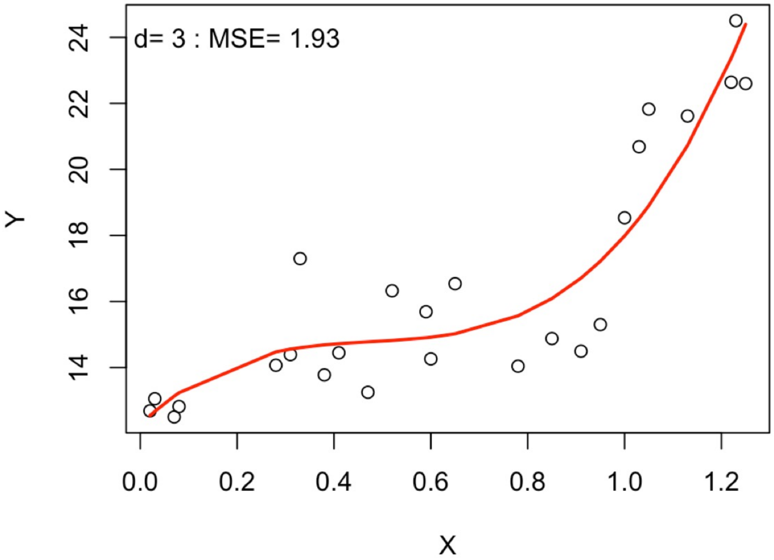
A.4 Polynomial Regression 1

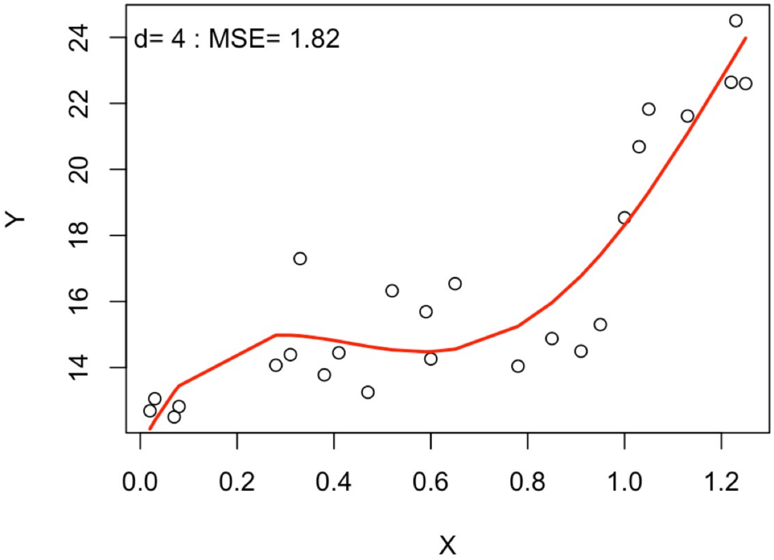


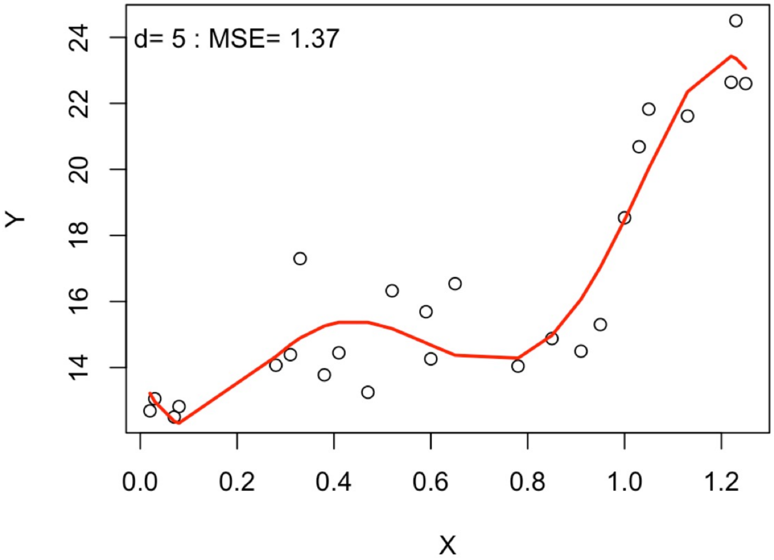

d= 1 : MSE= 4
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

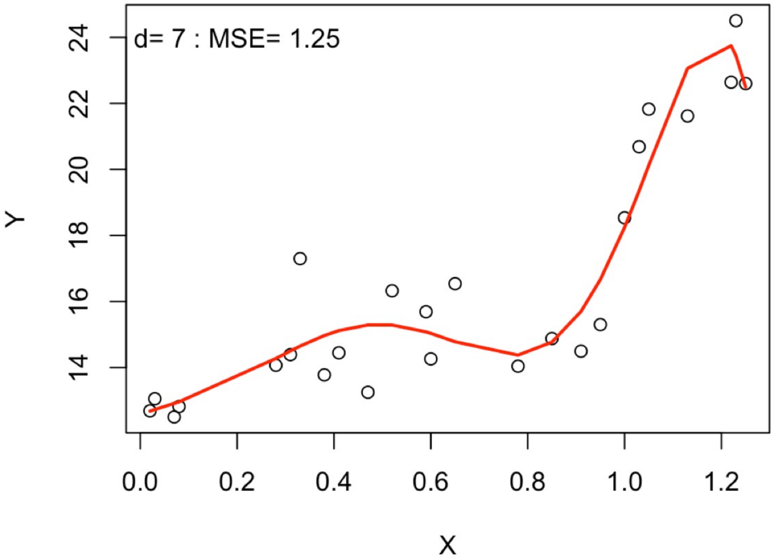


d= 2 : MSE= 2.68
Y
14
16
18
20
22
24
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2

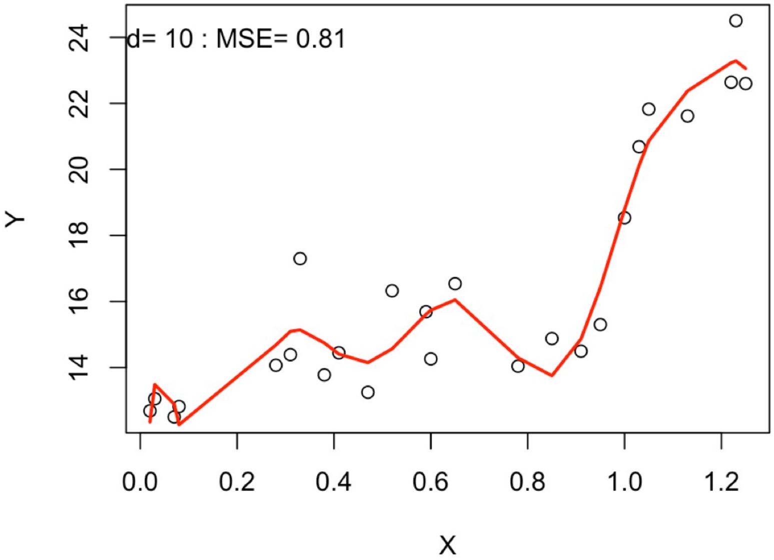

d= 3 : MSE= 1.93
Y
X
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2

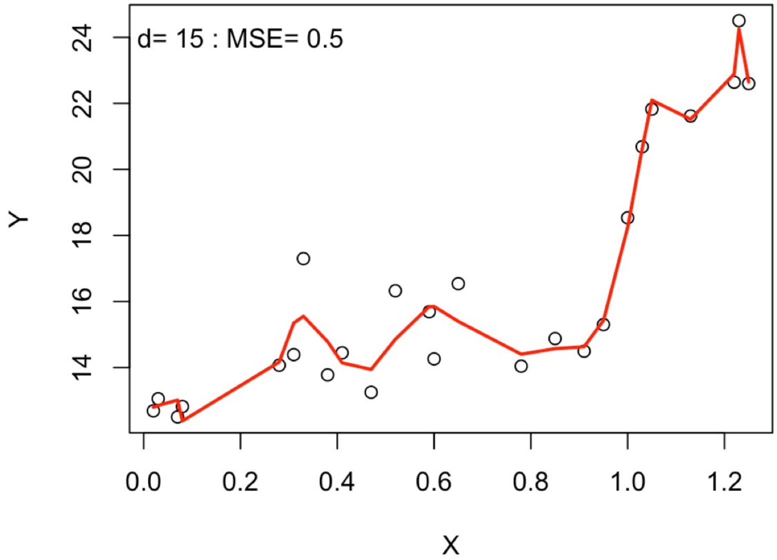

d= 4 : MSE= 1.82
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X


d= 5 : MSE= 1.37
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

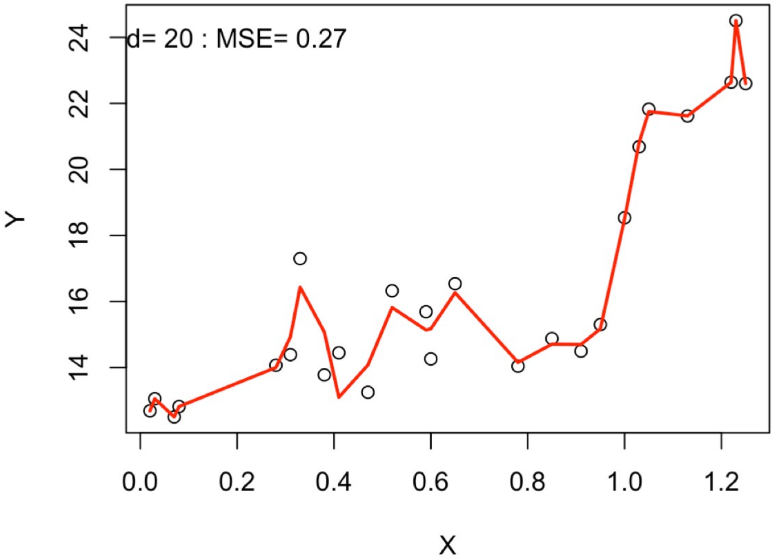


d= 7 : MSE= 1.25
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X


d= 10 : MSE= 0.81
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X



d= 15 : MSE= 0.5
Y
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X


d= 20 : MSE= 0.27
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

A.5 Problem

- More flexibility in the model is always going to result in better fit to the data.
- Better fitting model is not always inferential.
- Better fitting Leave some out and use it for 'validation' and 'testing'.
- Underlying mechanism:

$$Y = f(X) + \epsilon$$

A.6 Measure of Quality of Fit

- Training MSE (sample)

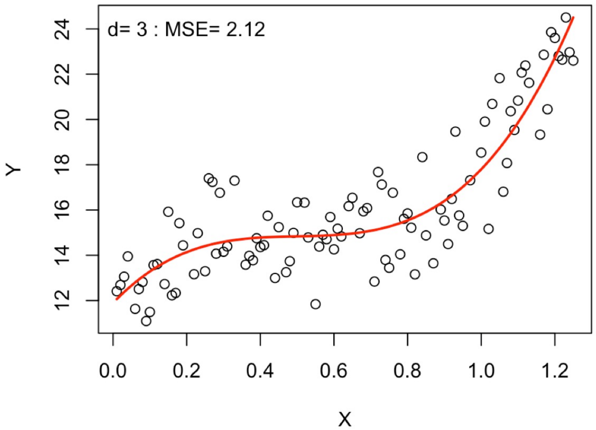
$$\text{MSE}_{tr} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

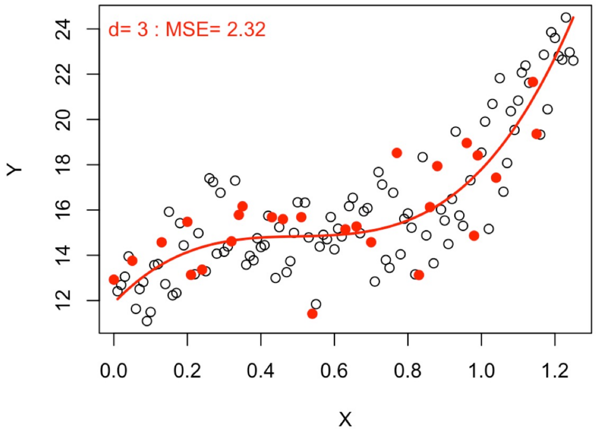
- But we want minimum Prediction MSE

$$\text{MSE} = E\left(Y - \hat{f}(X)\right)^2$$

- Solution: look at Test MSE (sample) as estimator

$$\mathrm{MSE}_{test} = \frac{1}{m} \sum_{j=1}^{m} \left(y_j - \hat{f}(x_j)\right)^2$$

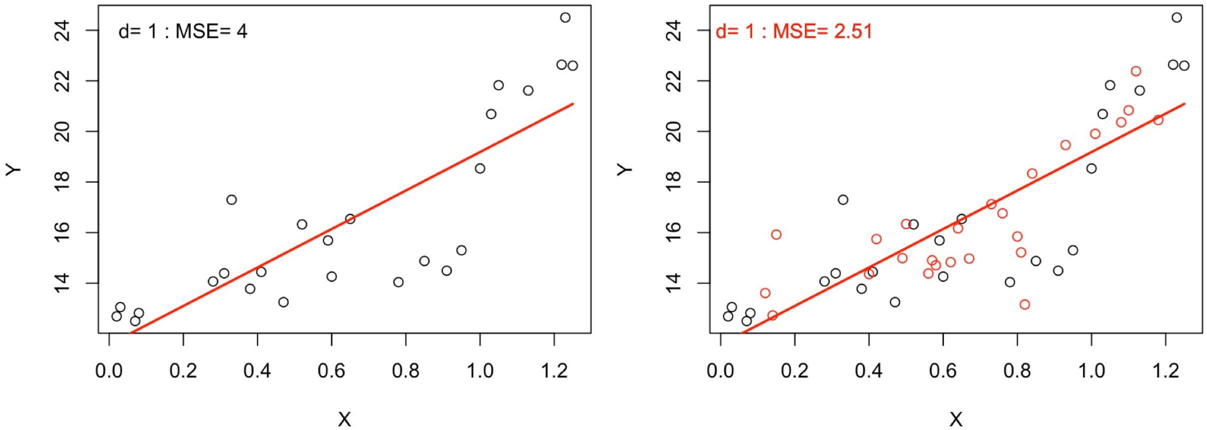


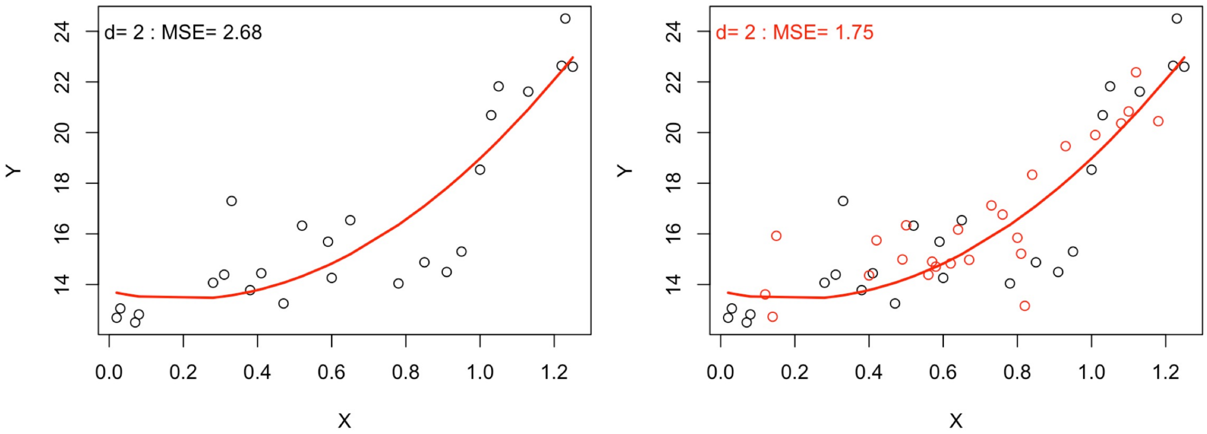


A.7 KEY CONCEPT

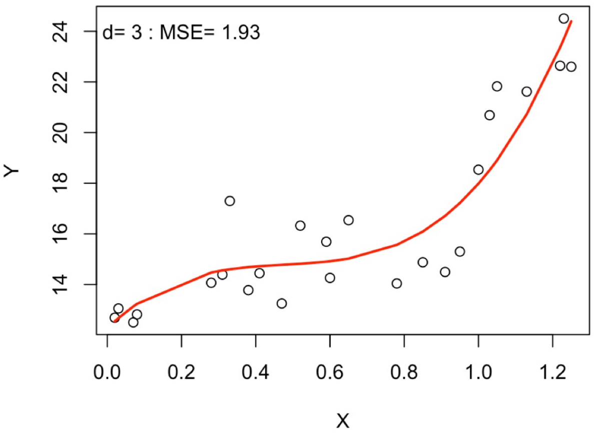
- Cross Validation
- Don't use all data when you are fitting a model
- Leave some out and use it for 'validation' and 'testing'.

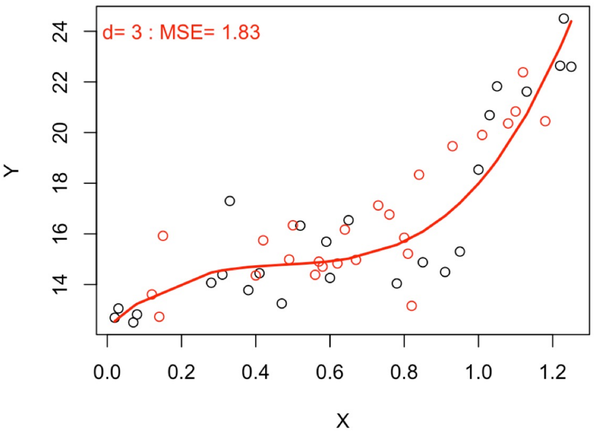
A.8 Leave-some-out Fitting Procedure 1

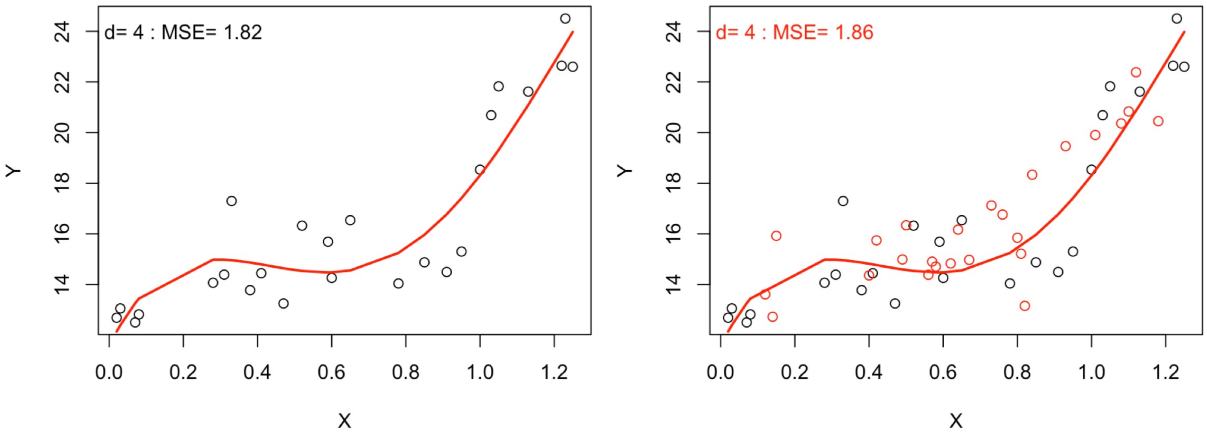


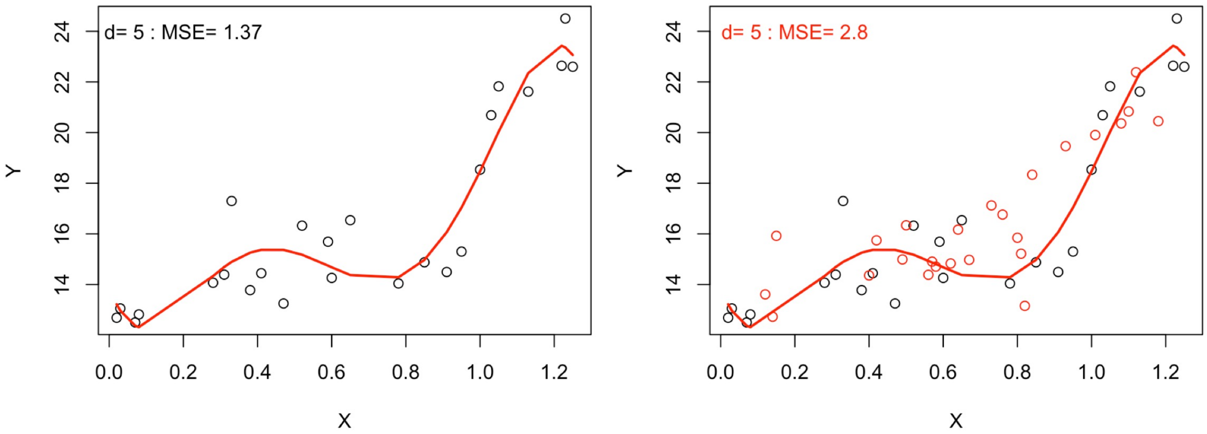


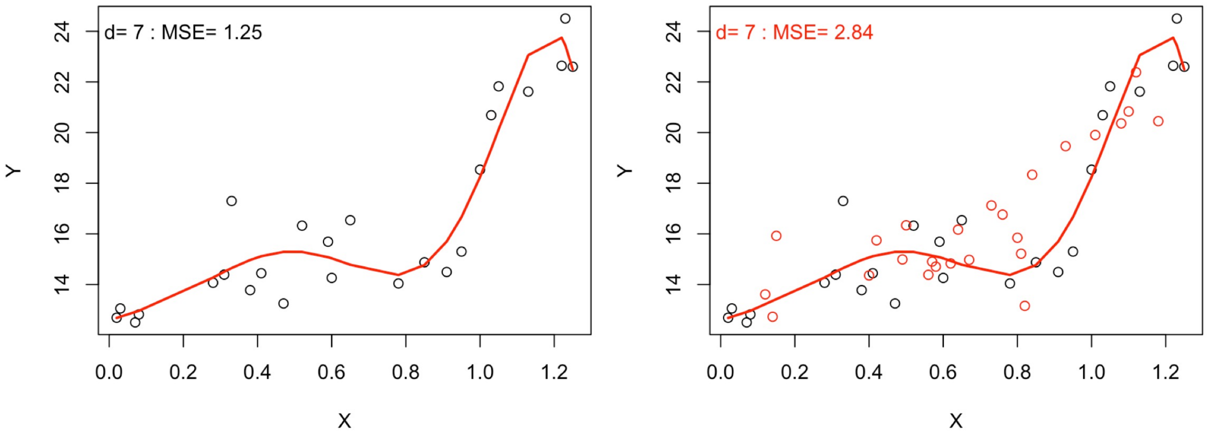
d= 2 : MSE= 2.68
d= 2 : MSE= 1.75
Y
X
0.0 0.2 0.4 0.6 0.8 1.0 1.2
14 16 18 20 22 24


d= 3 : MSE= 1.93
Y
X
0.0 0.2 0.4 0.6 0.8 1.0 1.2
14 16 18 20 22 24

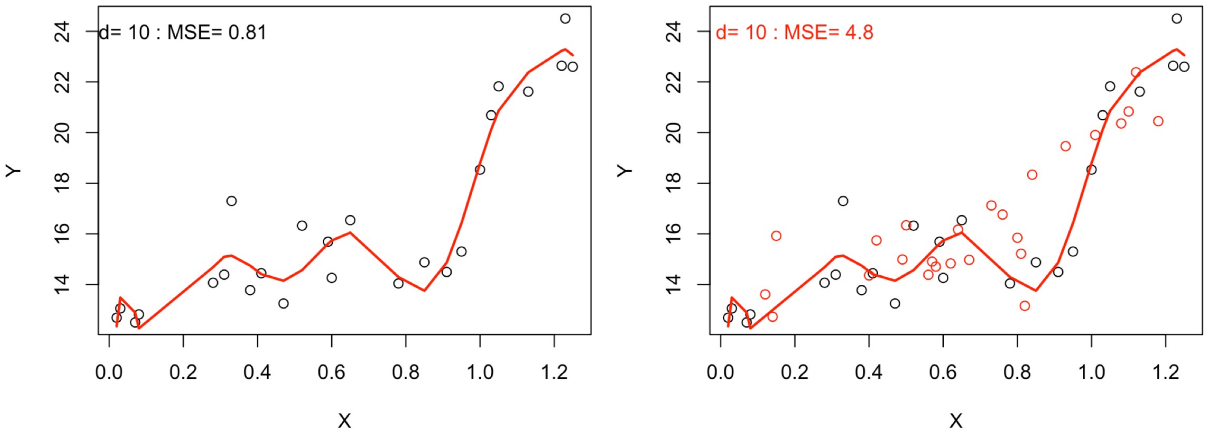

d= 3 : MSE= 1.83
Y
X
0.0 0.2 0.4 0.6 0.8 1.0 1.2
14 16 18 20 22 24

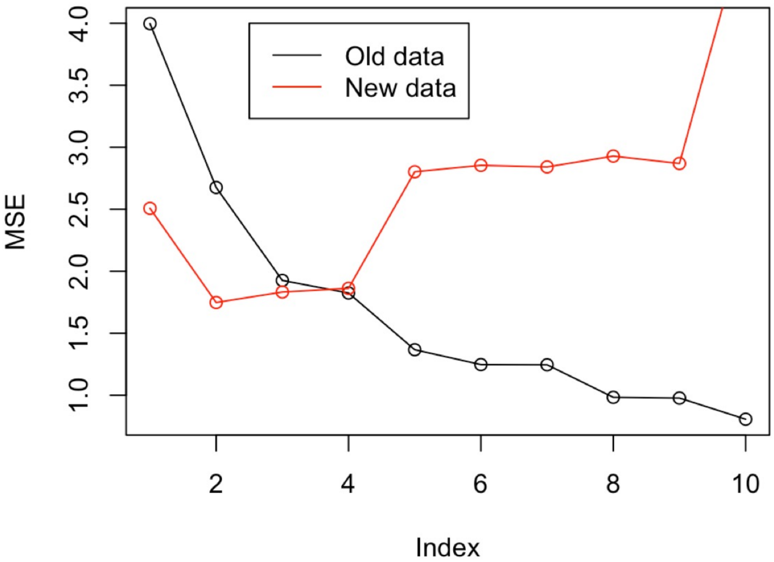

d= 4 : MSE= 1.82
d= 4 : MSE= 1.86
Y
X
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2


d= 5 : MSE= 1.37
d= 5 : MSE= 2.8
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
14
16
18
20
22
24



d= 7 : MSE= 1.25
d= 7 : MSE= 2.84
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
14
16
18
20
22
24

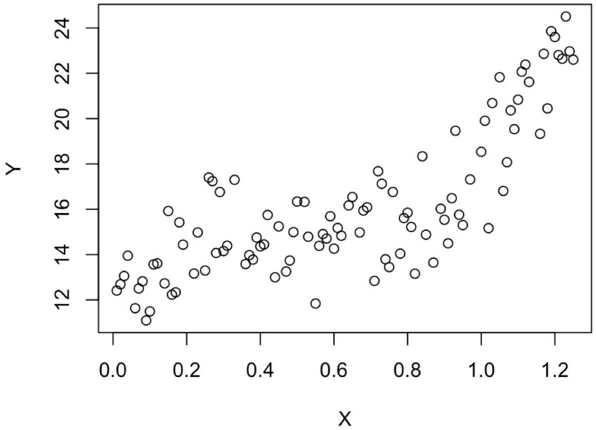

d= 10 : MSE= 0.81
d= 10 : MSE= 4.8
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
14
16
18
20
22
24

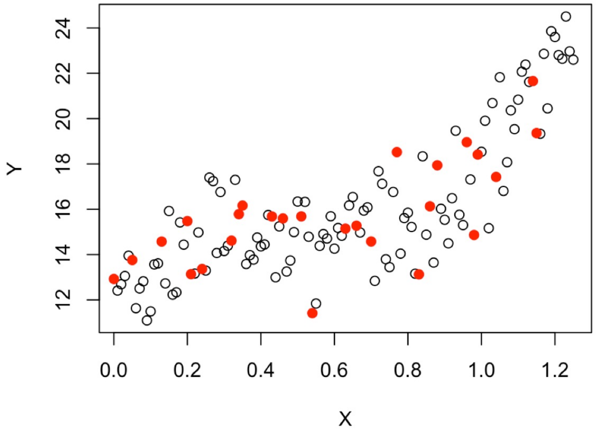


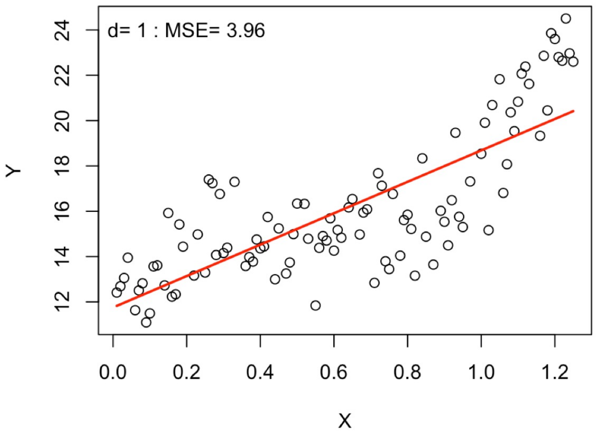
Old data
New data
MSE
Index
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2
4
6
8
10

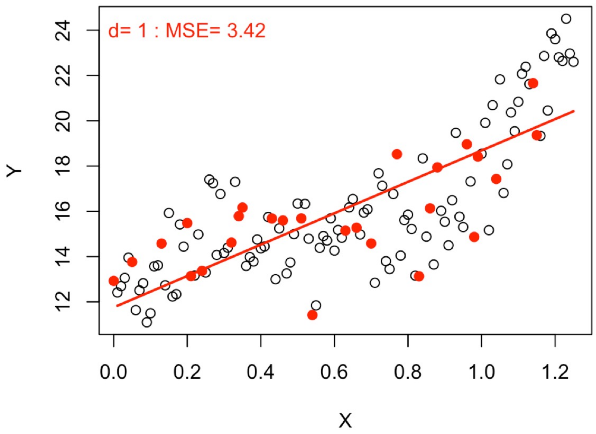
A.9 Leave-some-out Fitting 2

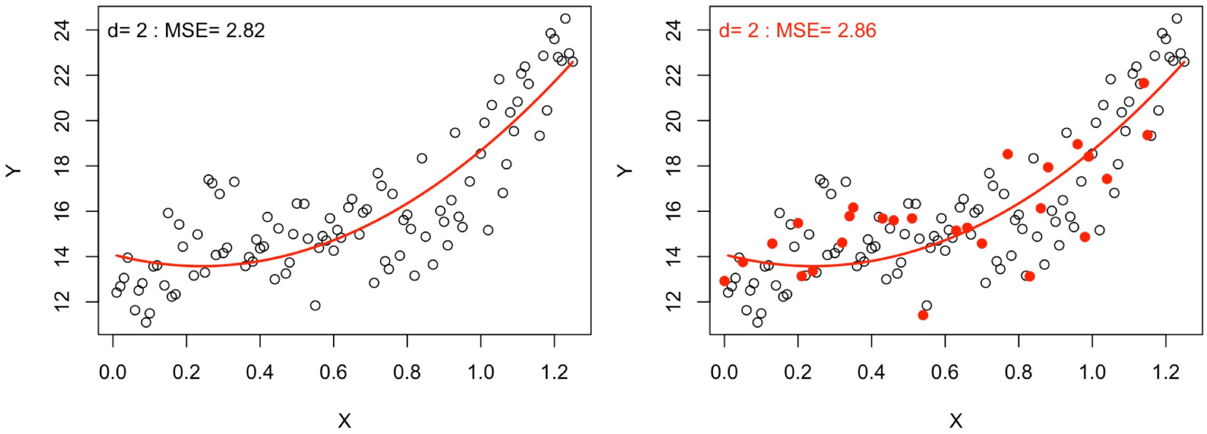
Larger dataset. $n = 100$ and $m = 26$.



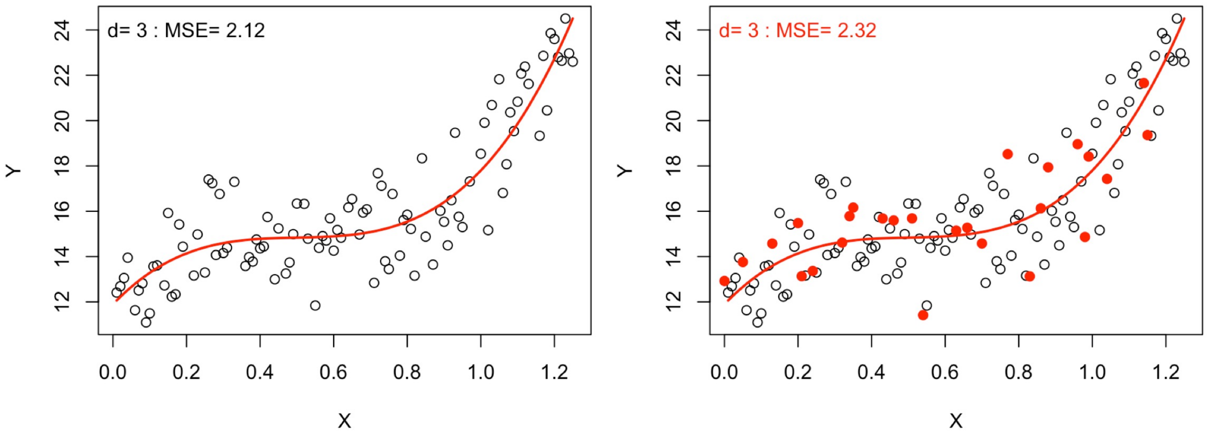


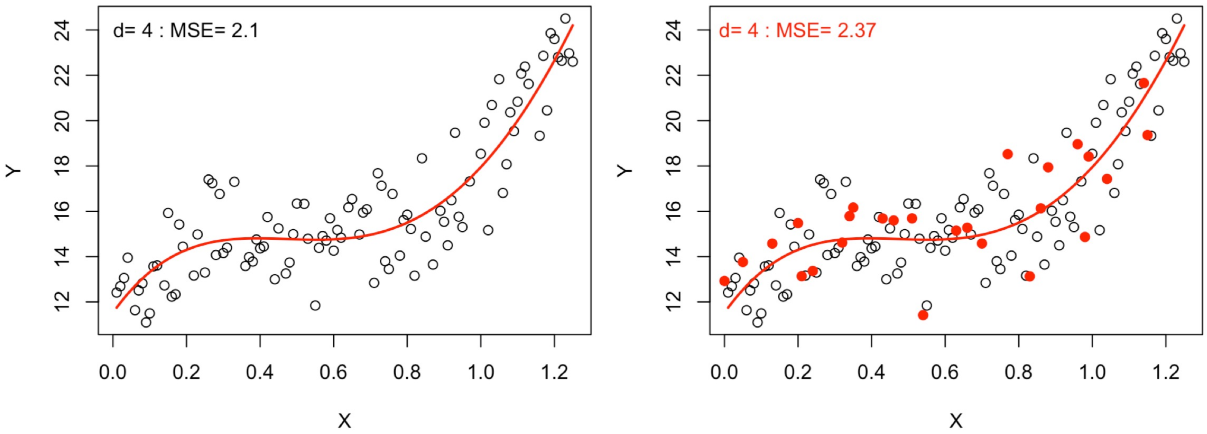

d= 1 : MSE= 3.96
Y
X
24
22
20
18
16
14
12
0.0
0.2
0.4
0.6
0.8
1.0
1.2

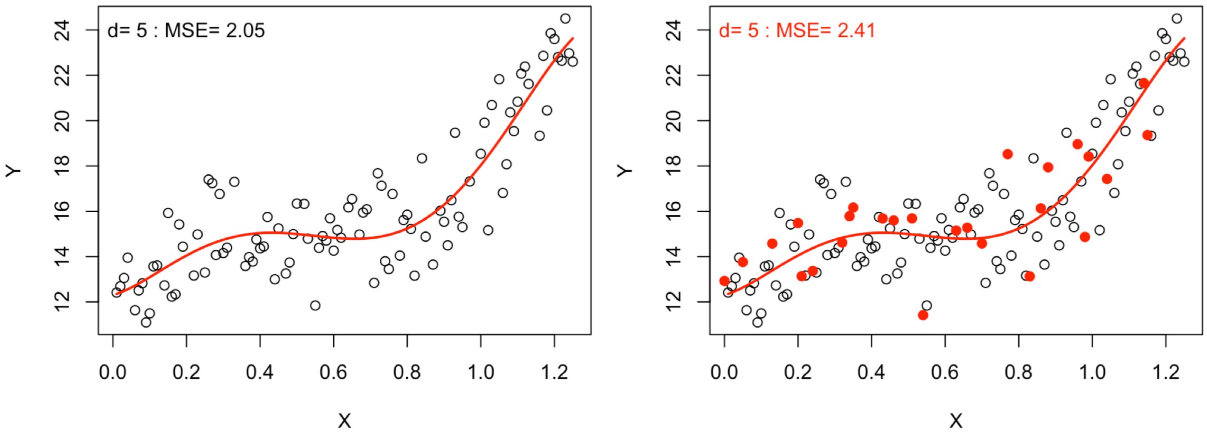

d= 1 : MSE= 3.42
Y
X
24
22
20
18
16
14
12
0.0
0.2
0.4
0.6
0.8
1.0
1.2

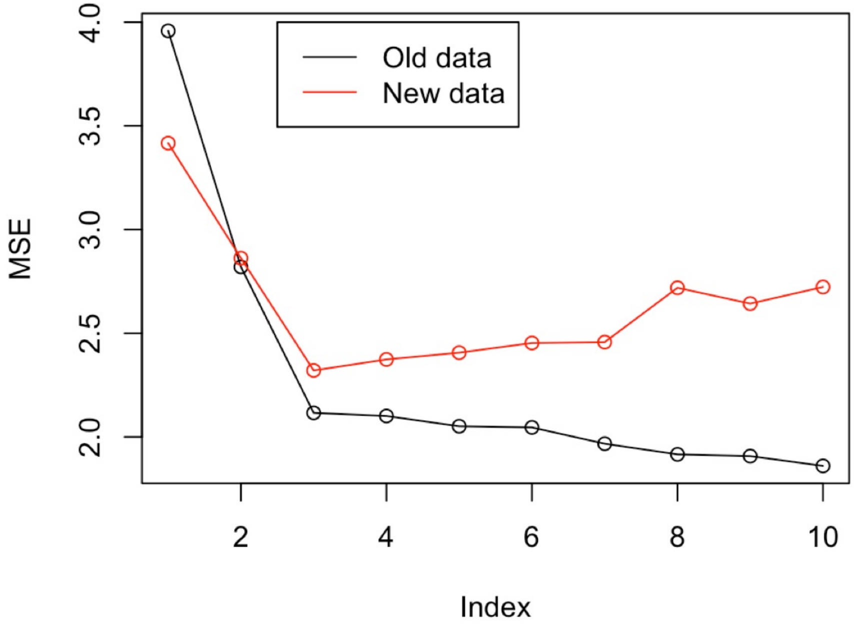


d= 2 : MSE= 2.82
d= 2 : MSE= 2.86
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
12
14
16
18
20
22
24


d= 3 : MSE= 2.12
d= 3 : MSE= 2.32
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
12
14
16
18
20
22
24


d= 4 : MSE= 2.1
Y
X
d= 4 : MSE= 2.37
Y
X


d= 5 : MSE= 2.05
d= 5 : MSE= 2.41
Y
X
12
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Old data
New data
MSE
4.0
3.5
3.0
2.5
2.0
2
4
6
8
10
Index

A.10 k-fold Cross Validation

- Training, Validation, and Testing Set
- Usually $k = 5$ or $k = 10$. We use $k = 5$ in this class.
- Randomly divide n into 6 groups.
- For example, if $n = 150$ and $k = 5$,

```
n=150
[----------Training Set 125------------]   [Test Set]
[fold 1][fold 2][fold 3][fold 4][fold 5]
[  25  ][  25  ][  25  ][  25  ][  25  ]   [   25   ]
```

A.11 k-fold Cross Validation

- Hyperparameter - parameter in the model that controls flexibility.
- e.g. Polynomial Regression $\rightarrow d$.
- Use Cross-Validation within the training set to tune the hyperparameter.
- Once you pick your hyperparameter, test the final model using [Training Set] to fit the model, and use [Test Set] to find testing MSE.
- Cross-Validation of the training set can be used repeatedly.
- Test Set should be used only once per method at the end.

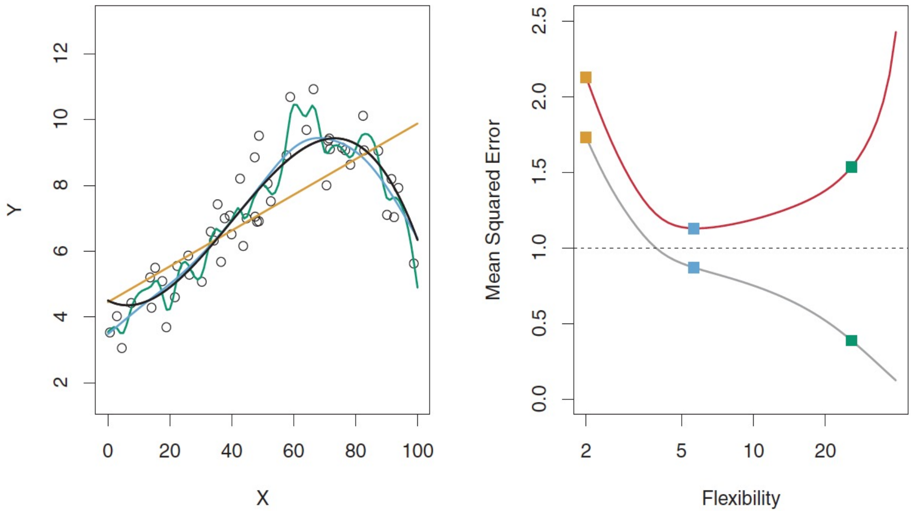
A.12 CV within the Training Set

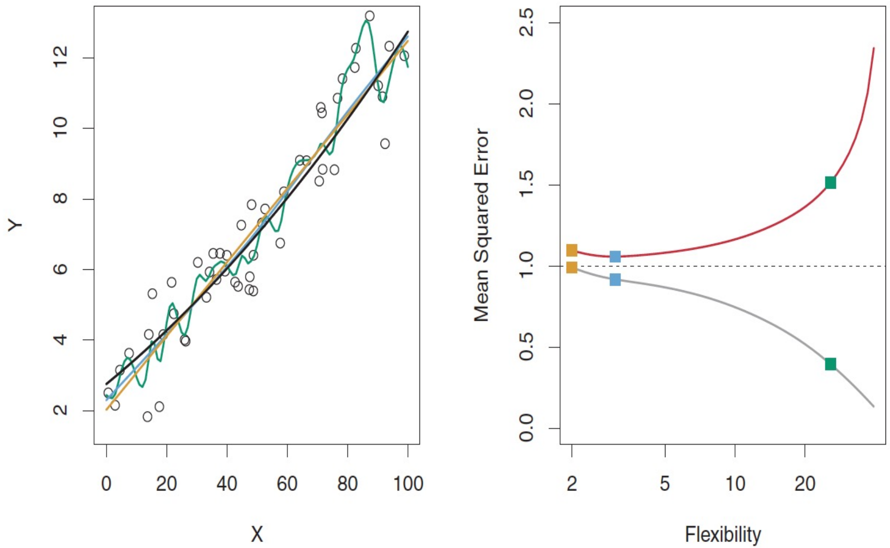
```
Let k=4
Data = [A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z]
n=26
Training = 5*4 = 20 obs.
Test = 6 obs.

[  fold 1 ][  fold 2 ][  fold 3 ][  fold 4 ]    [     Test    ]
[J G M T V][L B Q E I][N U X A P][Z K S D F]    [O C W Y H R]    # random assignment
```

```
CV rd 1 Training:               [fold 2][fold 3][fold 4]  ->  Training MSE
CV rd 1 Validation: [fold 1]                              ->  Validation MSE

CV rd 2 Training:   [fold 1]        [fold 3][fold 4]      ->  Training MSE
CV rd 2 Validation:         [fold 2]                      ->  Validation MSE

CV rd 3 Training:   [fold 1][fold 2]        [fold 4]      ->  Training MSE
CV rd 3 Validation:                 [fold 3]              ->  Validation MSE

CV rd 4 Training:   [fold 1][fold 2][fold 3]              ->  Training MSE
CV rd 4 Validation:                         [fold 4]      ->  Validation MSE

#--- Pick hyperparameter based on Average Validation MSE ---

Final Fit Training: [   -- All Training Set ---    ]              ->  Training MSE
Final Fit Test:                                        [  Test  ] ->  Test MSE

#--- Method should be compared to other Medhod using Test MSE
```

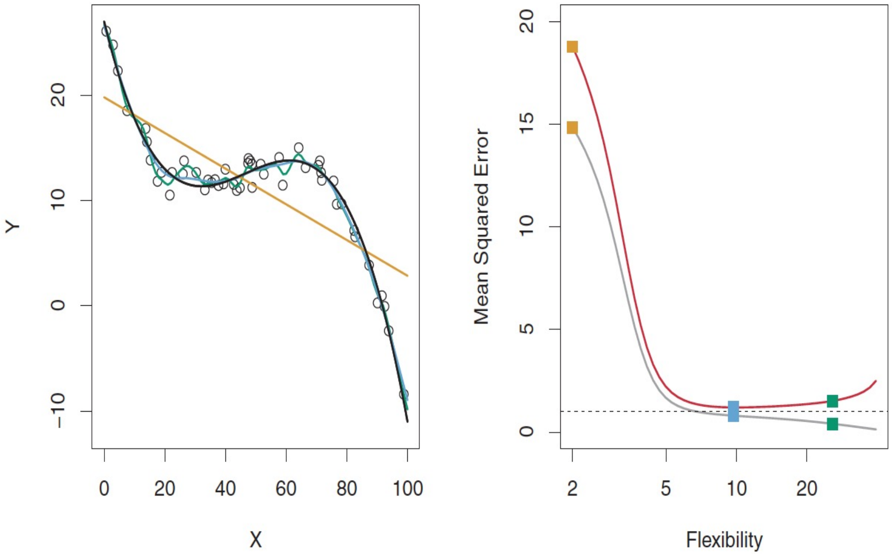
A.13 Training MSE vs Validation MSE

Also called in-sample vs out-sample





Y
2
4
6
8
10
12
0
20
40
60
80
100
X
Mean Squared Error
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility


Y
X
Mean Squared Error
Flexibility

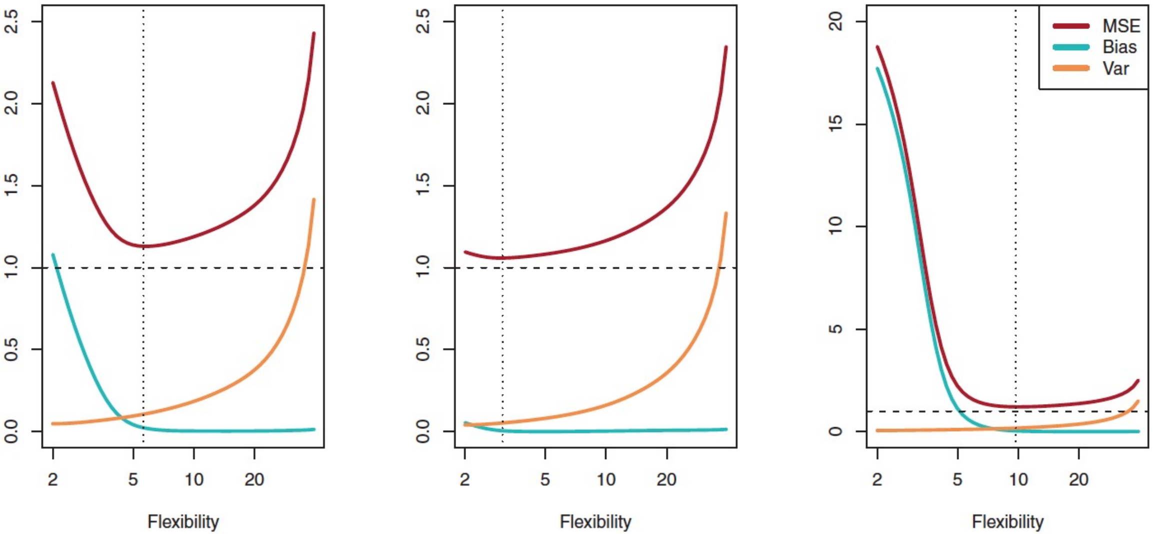
A.14 Assessing Model Prediction Accuracy

Assessing Model Prediction Accuracy

A.15 Bias-Variance Trade-Off

Prediction MSE can be decomposed as

$$
\begin{aligned}
E\Big(Y - \hat{f}(X)\Big)^2 &= E\Big(f(X) + \epsilon - \hat{f}(X)\Big)^2 \\
&= E\Big(f(X) - E(\hat{f}(X)) + E(\hat{f}(X)) - \hat{f}(X) + \epsilon\Big)^2 \\
&= E\Big(f(X) - E(\hat{f}(X))\Big)^2 + E\Big(E(\hat{f}(X)) - \hat{f}(X)\Big)^2 + E\big(\epsilon^2\big) \\
&= Var(\hat{f}(X)) \; + \; Bias(\hat{f}(X))^2 \; + \; Var(\epsilon)
\end{aligned}
$$



0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility
0
5
10
15
20
2
5
10
20
Flexibility
MSE
Bias
Var

A.16 Prediction MSE

- $$E\big(Y - \hat{f}(X)\big)^2 = Var(\hat{f}(X)) + Bias(\hat{f}(X))^2 + Var(\epsilon)$$
- can't have low variance and low bias
- has lower bound

A.17 In the Classification Setting

- Instead of MSE, work with Errror Rate:

$$ER = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

A.18 Trade-off in the new approach

- Classical Statistics (Probabilistic Model)

$$Y = f(X) + \epsilon$$

  - Assume parametric model for $f(\cdot)$ and ϵ.
  - Sampling Probability of $(y_1, \cdots, y_n)$, which are realizations of r.v. Y.
  - Esimate parameters for $f(\cdot)$ and ϵ.
  - Because the model distinguish the mechanism $f(\cdot)$ vs noise ϵ, looking at in-sample fit was enough (if the assumption is correct).
  - Predict future Y using the estimated model.

- Pros and Cons
  - Model is interpretable.
  - Future effect of the model is easier to calculate.
  - No need for out-sample validation (test set), if assumption is correct.
  - Popular models are mathematically optimized already, to save the computational task.
  - Theory on prediction interval. Based on the assumption, often distribution on the prediction error is available.

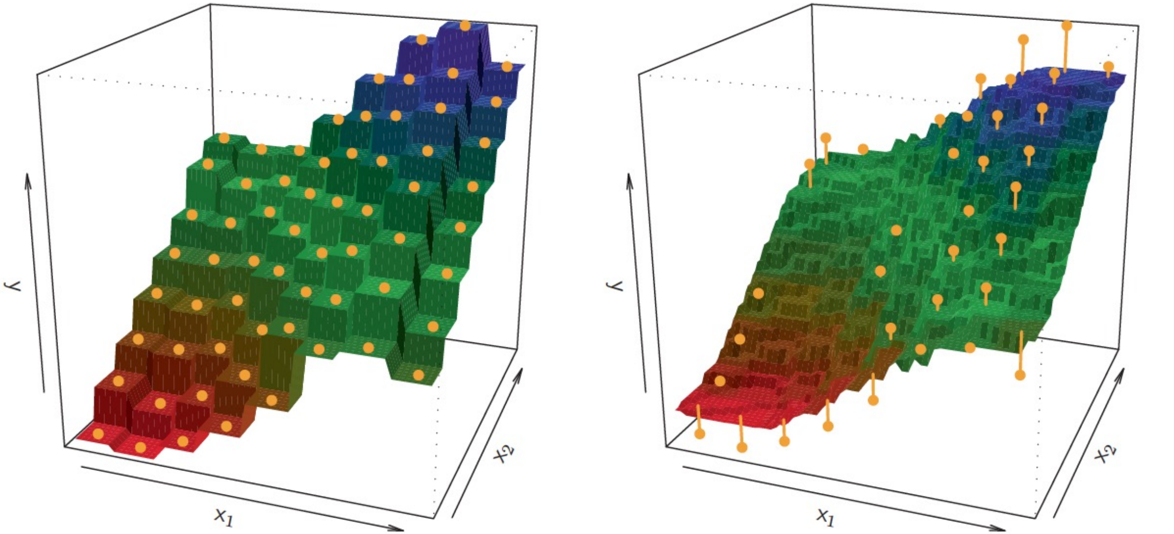
A.19 K-Nearest Neighbor

- One of elementary supervised learning model.
- Pick a point x_0, find K nearest observations.
- $f(x_0)$ is estimated by the average of all K neighbors:

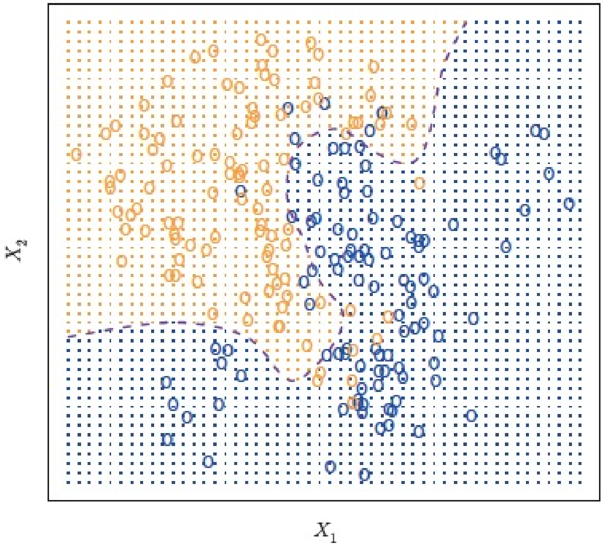
$$\hat{f}(x_0) = \frac{1}{K} \sum y_i.$$

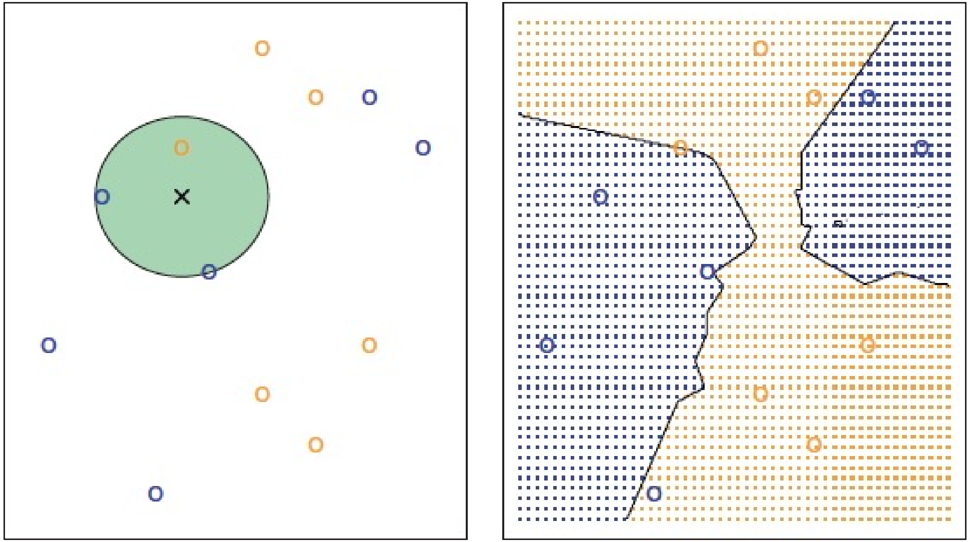
- K is the hyperparameter.
- Can be used for Regression or Classification

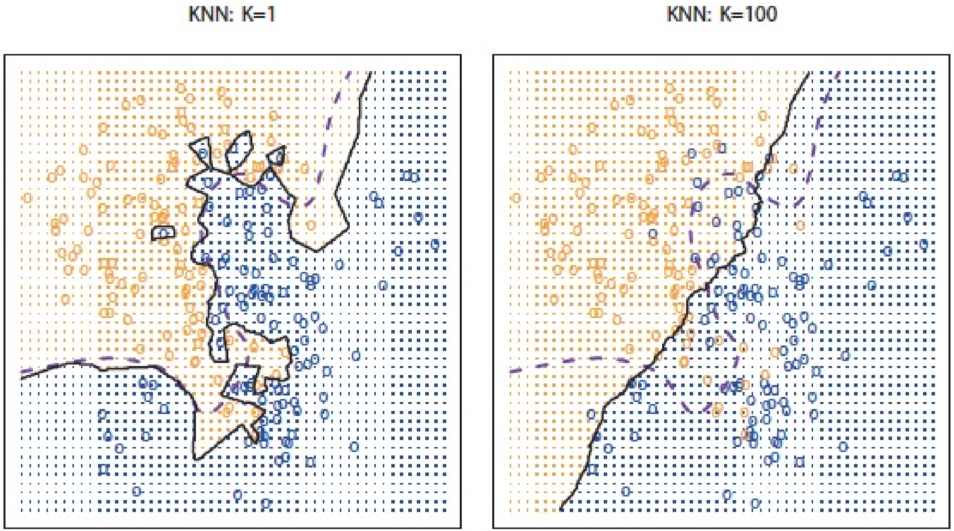
K=1 (left) and K=9 (right)

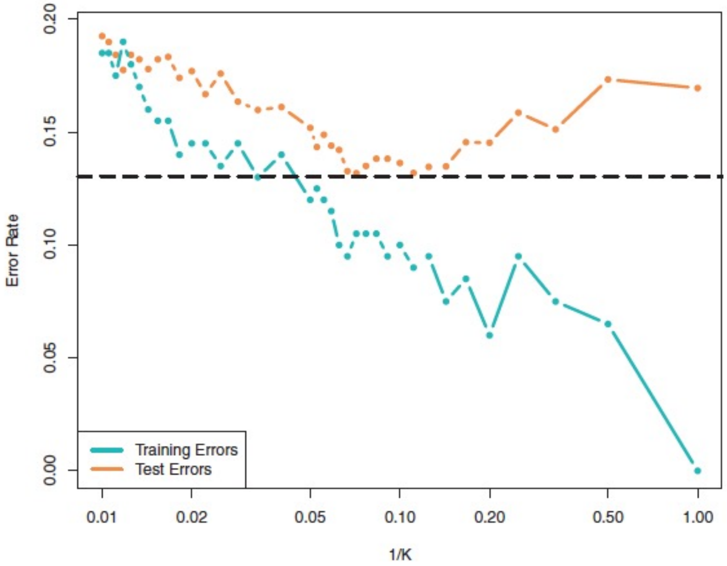


A.20 K-NN examples






KNN: K=1
KNN: K=100



Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors