Module II: Machine Learning: Foundations and Algorithms

Nirav P Bhatt
Department of Biotechnology
Robert Bosch Centre for Data Science and Artificial Intelligence
Indian Institute of Technology Madras, Chennai – 600036, India

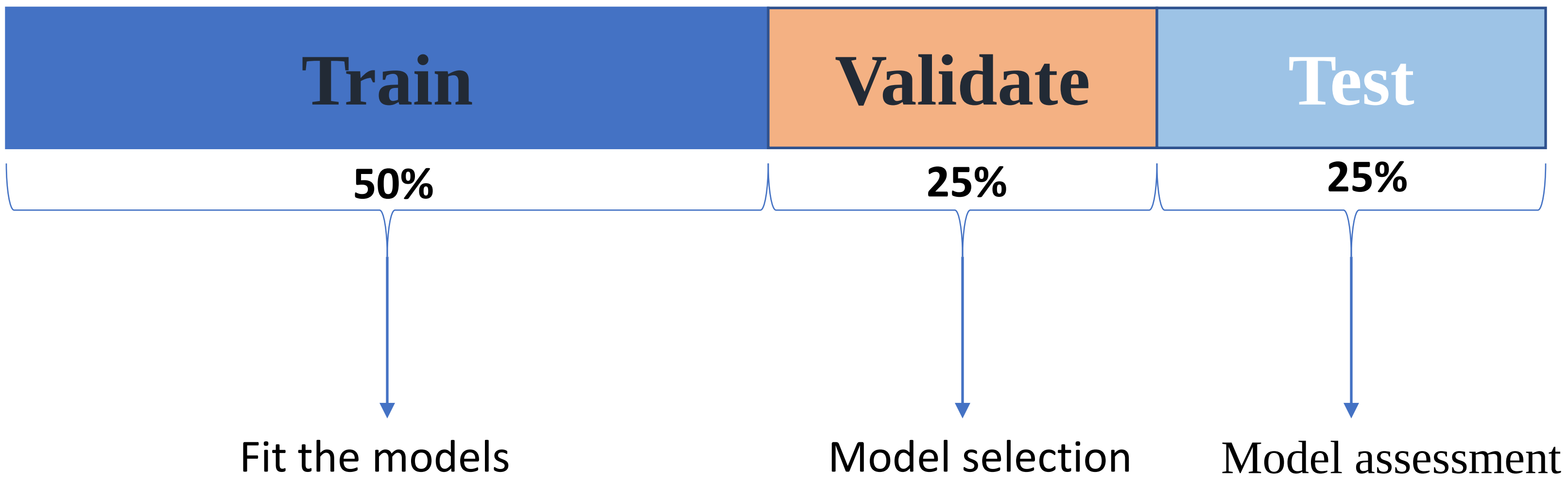
Model Selection and Assessment

- Development of model
  - **Building a model**(s): Verifying all the assumptions
  - Validation of model: Predictive ability of the model
  - Testing the model on new data
- Two goals: Validation of model
  - **Model selection:** Comparing the performance of several models to find the best one
  - **Model assessment:** Assessing the predictive ability of the chosen final model on new data

Model Selection and Assessment

- Model selection is important in multiple linear and nonlinear models
- Data-rich situation: Randomly divide the data in three parts

Ideal Scenario: Data-rich situation

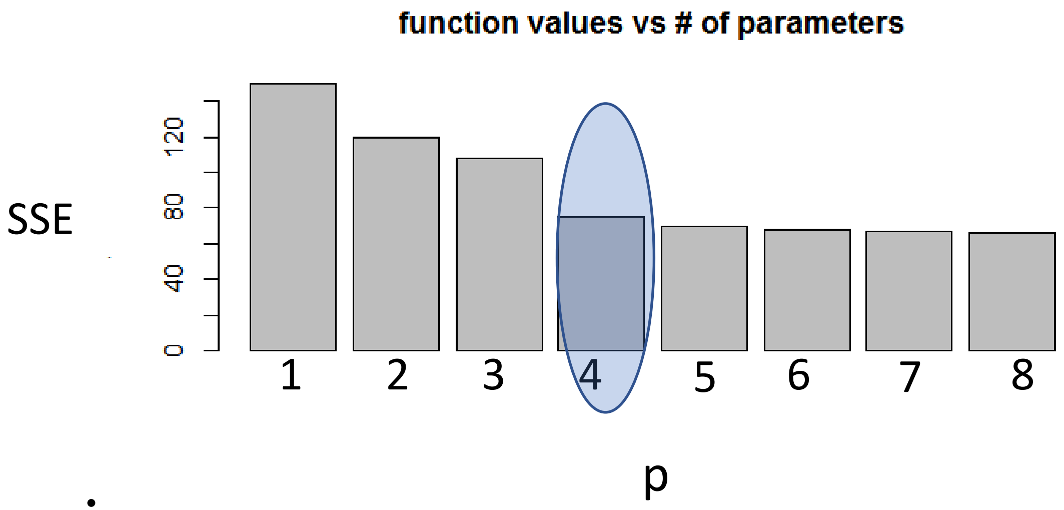


Practice: Limited Amount of Data

Best Model in Practice? Need a Criterion

Model Complexity

- SSE: $S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
- *SSE* values decreases with increase in parameters (or variables)



- Over fitting as *p* increases
- Under fitting not including parameters
- How to determine # of parameters?

Model Complexity

“Give me three parameters, I can fit an elephant
Give me five and I can include his tail!”
- Fogler and Gurmen (ECRE, 2006)

Principle of Parsimony:
Simplest model which fits the data well should be chosen

Model Complexity

- Trade off: *SSE* values and model parameters

- Model selection criterion?

  SSE values + model complexity

- SSE values: Assess the quality of the model

- Model complexity: Principle of Parsimony

  penalize complex models = number of parameters increases

K Nearest Neighbors Classifier

- Data: $\{(x_1, y_1), (x_2, y_2), \ldots.., (x_n, y_n)\}$
  - Features: $(x^1, x^2, \ldots. x^p)$: x_i
  - Label: y_i
- New test data x_o
  - What is the corresponding label?

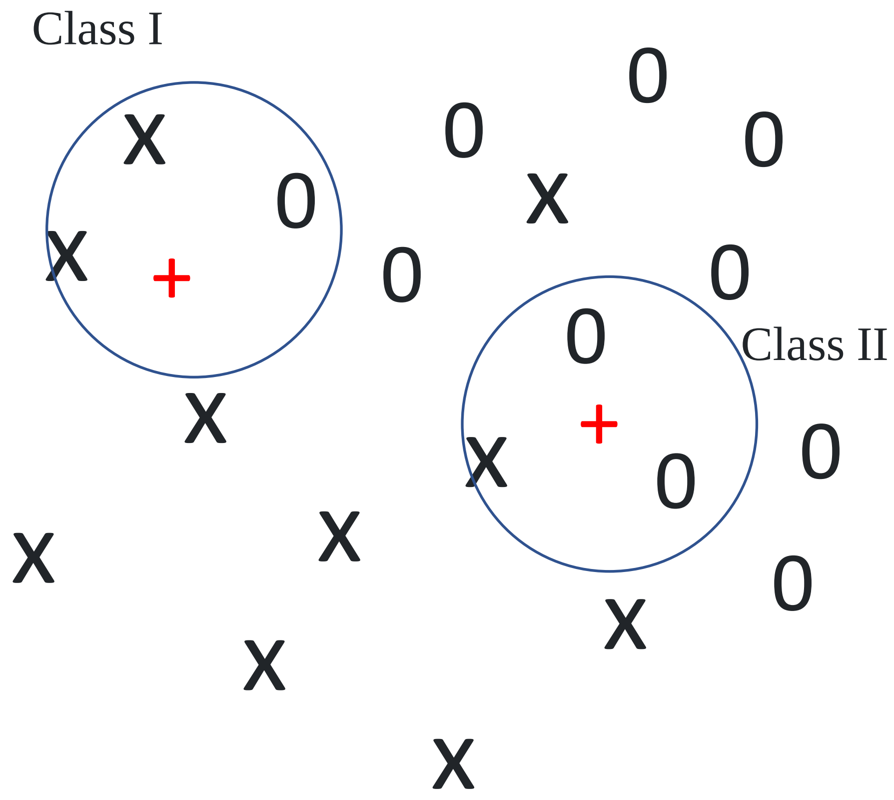
- Instant based Classifier
  - Use the data (or training data) for classification (no models)
  - Non-parametric method

K Nearest Neighbors Classifier

- How can we find the new Label?
- Old adage: Something walks and talks like peacock beware of statistics it may be hen
- kNN Idea: Something walks and talks like peacock it is high likely to be peacock not hen

K Nearest Neighbors Classifier

x: Class I and 0: Class II



- kNN classifier
  - Training Data: $\{(x_1, y_1), (x_2, y_2), \ldots.., (x_n, y_n)\}$
  - A distance Metric
  - Number of neighbors: K

K Nearest Neighbors Classifier

Algorithm

1. Data $\{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)\}$
2. For new data point, x_o
3. Find the nearest point(s)

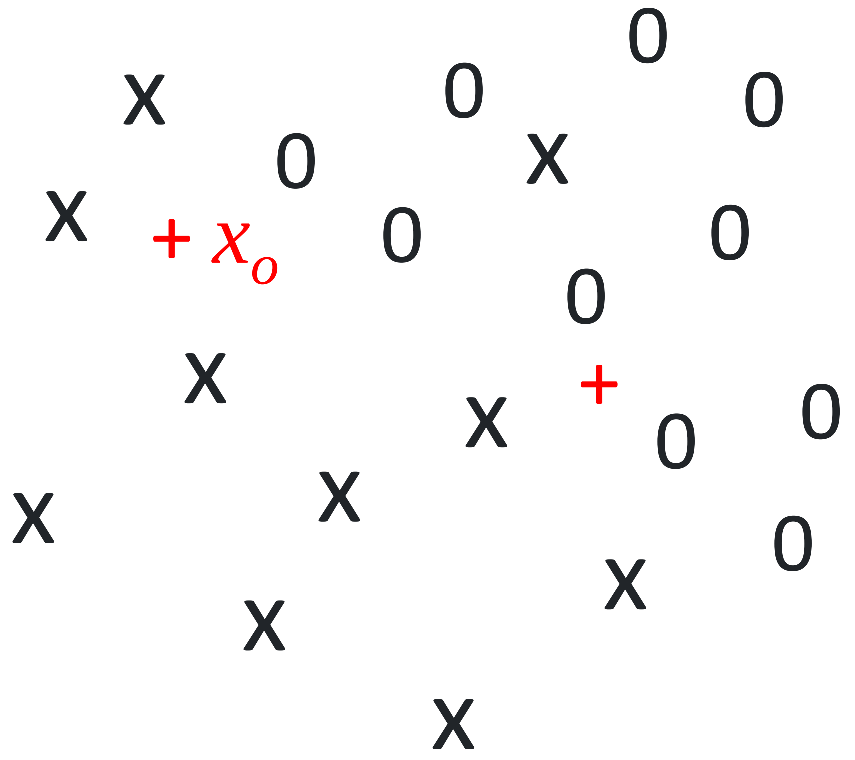
$$n^* = \underset{n=1,\ldots,n}{\operatorname{argmax}} ||x_0 - x_n||^2$$

4. Label $y_o = y_{n*}$ based on majority votes

K Nearest Neighbors Classifier

Example:

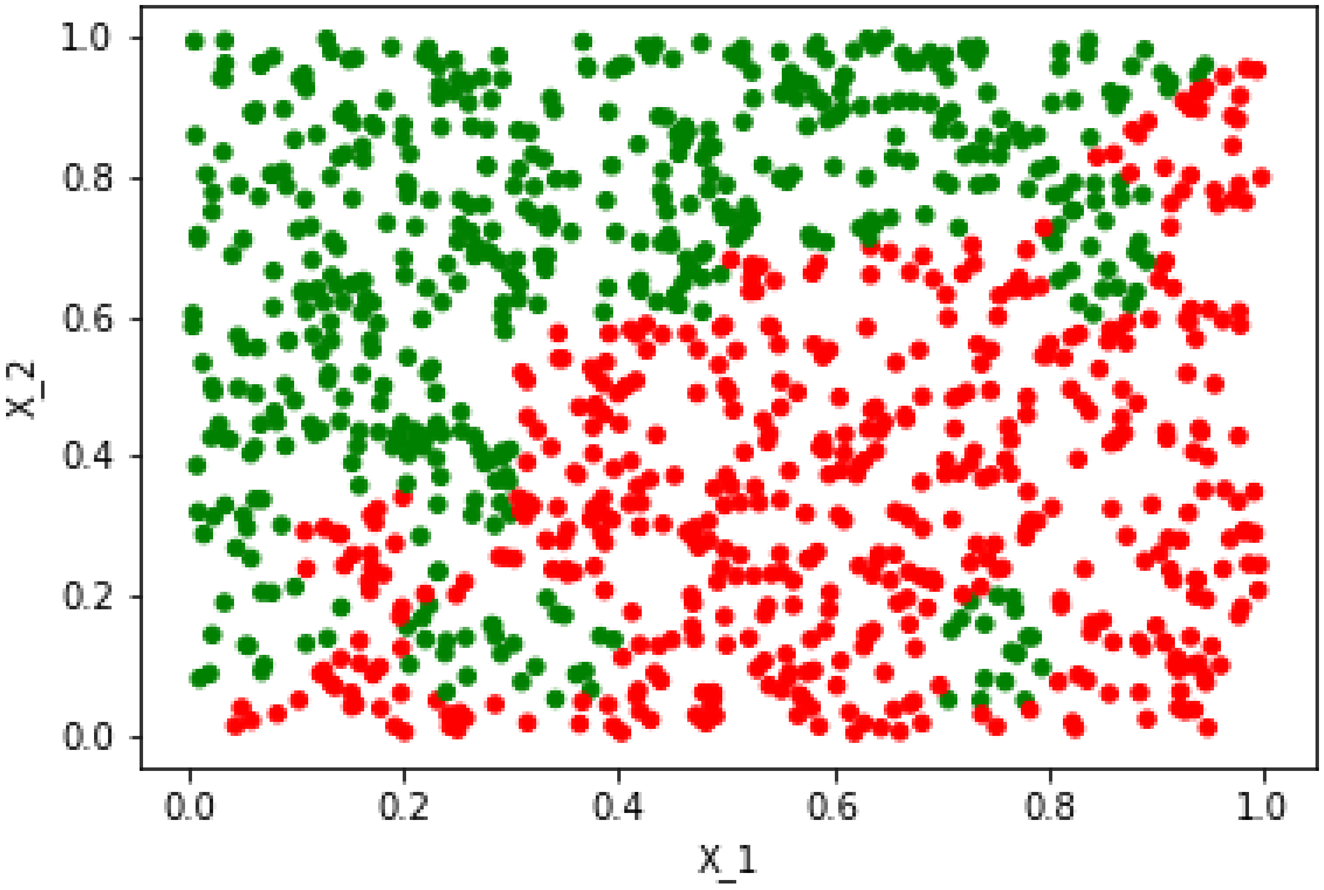
x: Class I and 0: Class II



- K=3
- Compute conditional probability
  - P(Y=Class I | x=x_o)= 0.67
  - P (Y= Class II| x=x_o)=0.33

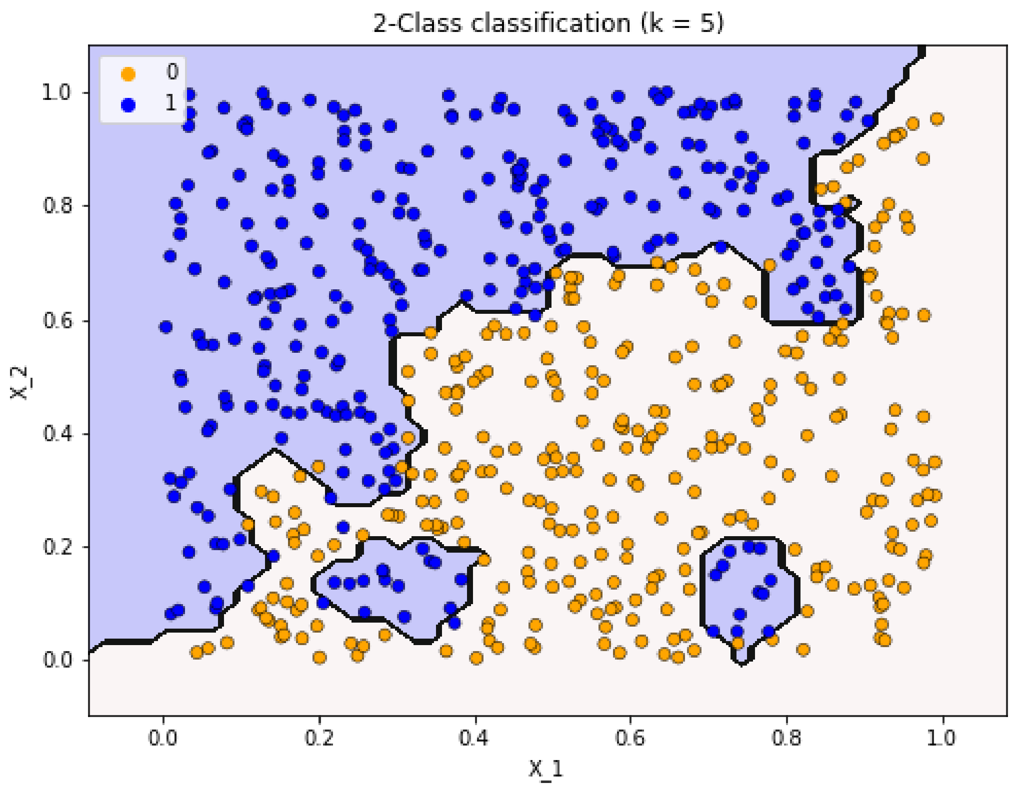
K Nearest Neighbors Classifier

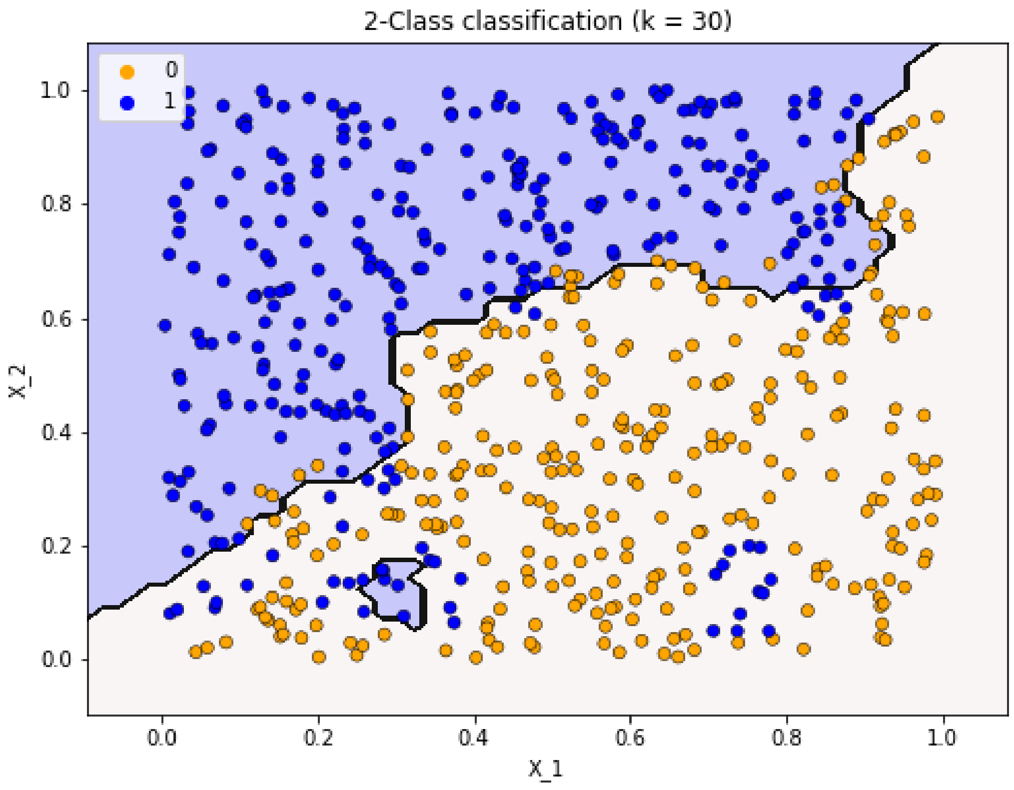
2-class classification problem with 2 features



K Nearest Neighbors Classifier

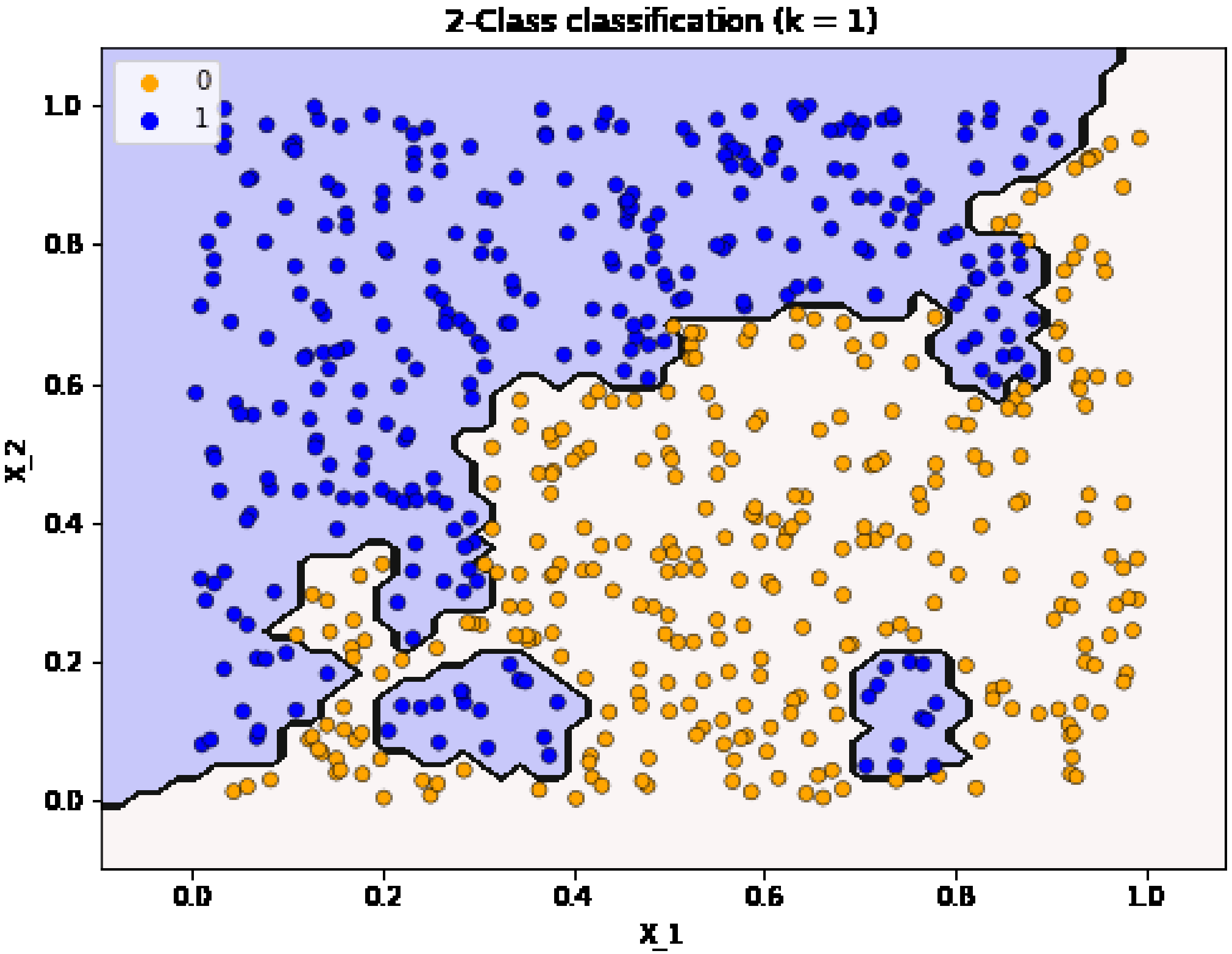
2-class classification problem with 2 features





K Nearest Neighbors Classifier

2-class classification problem with 2 features

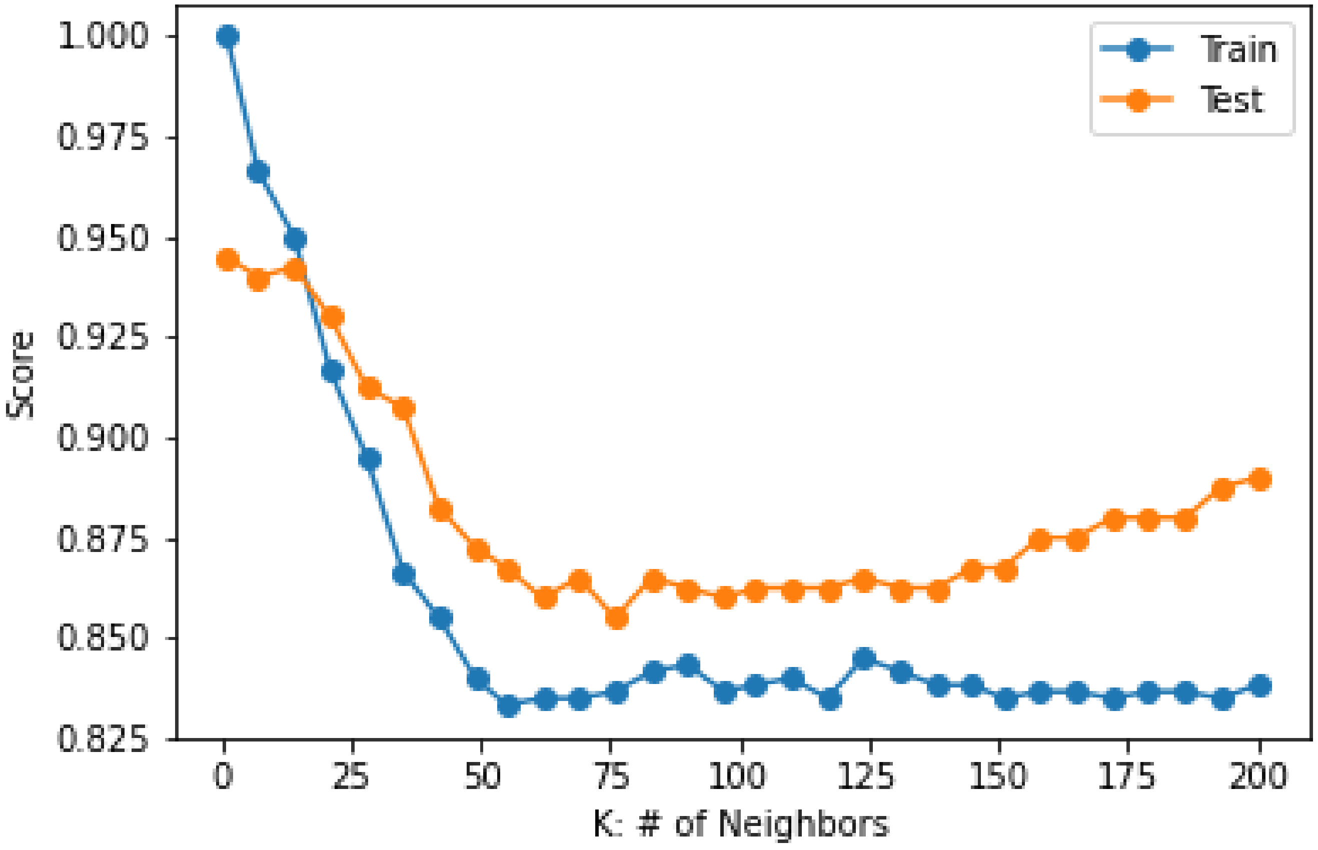


K Nearest Neighbors Classifier

- Choice of K
- Large K value
  - Less flexible model
- Small K value
  - Flexible model
  - But sensitive to noisy data point

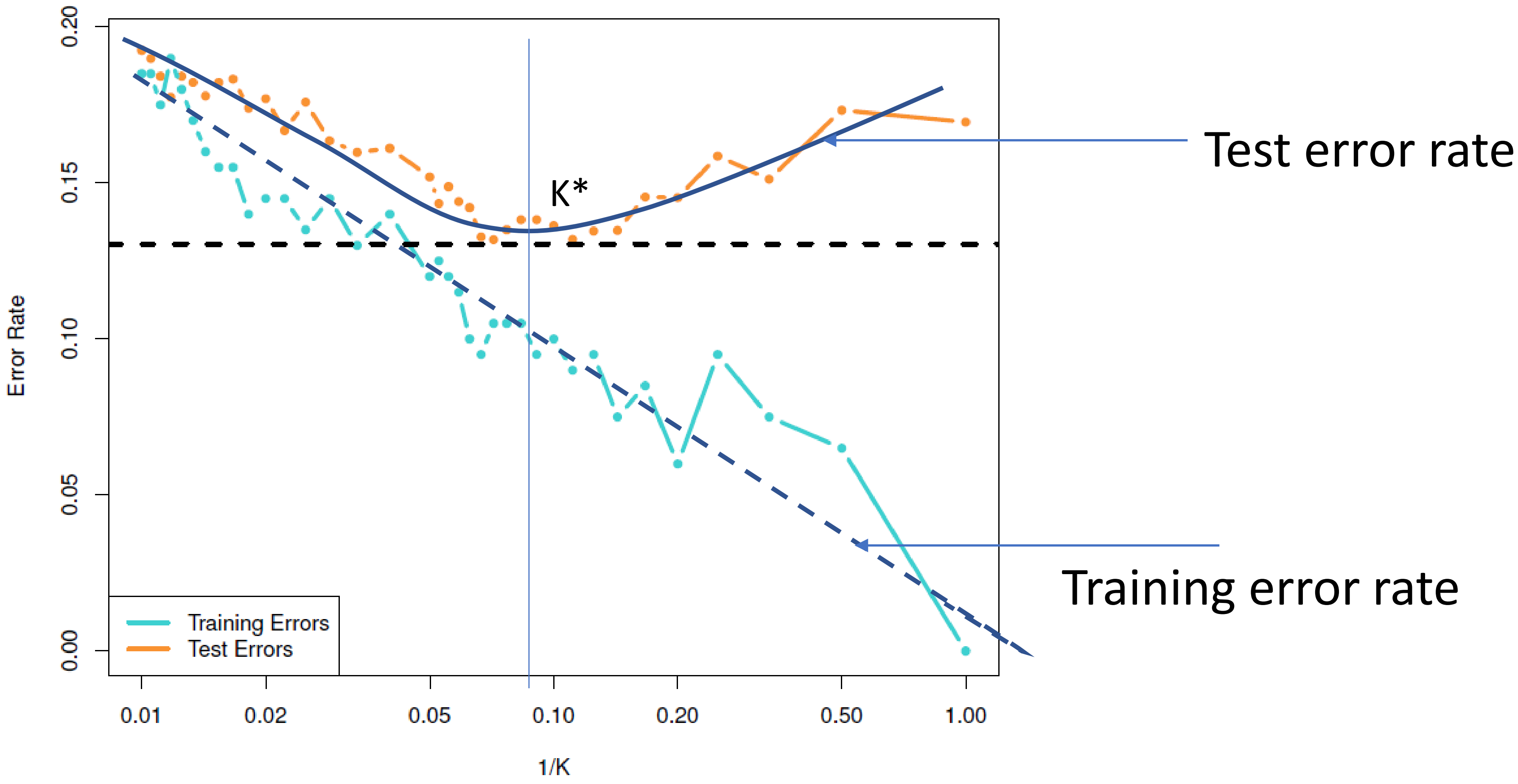
K Nearest Neighbors Classifier

How do we decide the “K”?



K Nearest Neighbors Classifier

How do we decide the “K”?



[1] James, G., Witten, D., Hastie, T., and Tibshirani, R. An Introduction to statistical learning, 2021

Irreducible and Reducible Errors

Mean Square Error between the actual and predicted y *using* the fit $\hat{f}(x, \hat{p})$

$$E[(y - \hat{y})^2] = [f(x,p) - \hat{f}(x,\hat{p})]^2 + Var(\epsilon)$$

Irreducible Error $\quad Var(\epsilon)$

Reducible Error $\quad [f(x,p) - \hat{f}(x,\hat{p})]^2$

Definition: Bias

θ : Unknown True value of parameter or function

$\hat{\theta}$: Estimated θ

$\bar{\theta}$: $E[\hat{\theta}]$

Bias $E[(\bar{\theta} - \theta)^2]$

- Measure of accuracy
- Low bias: Accurately fitted function or estimated parameters

Definition: Variance

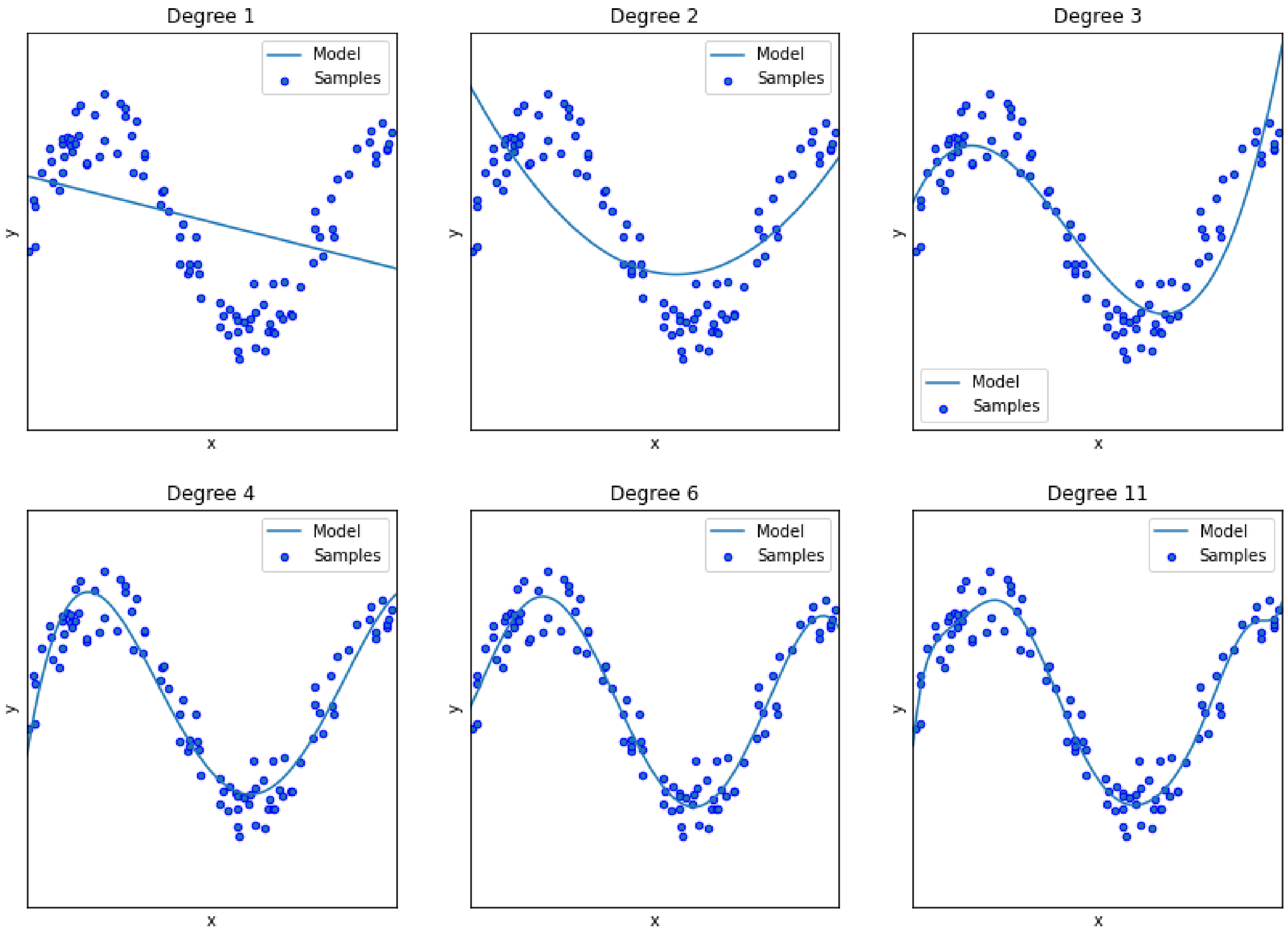
Variance $E[(\hat{\theta} - \bar{\theta})^2]$

- Measure of precision
- Low variance: Model parameters or function approximation does not change with training data significantly

Mean Square error of estimator

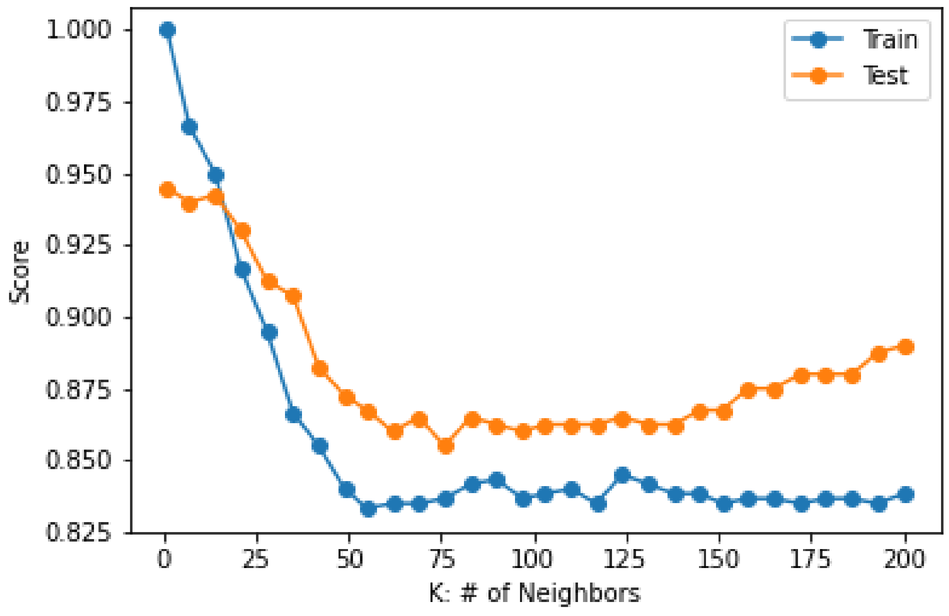
$$MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

Bias-Variance Trade-off

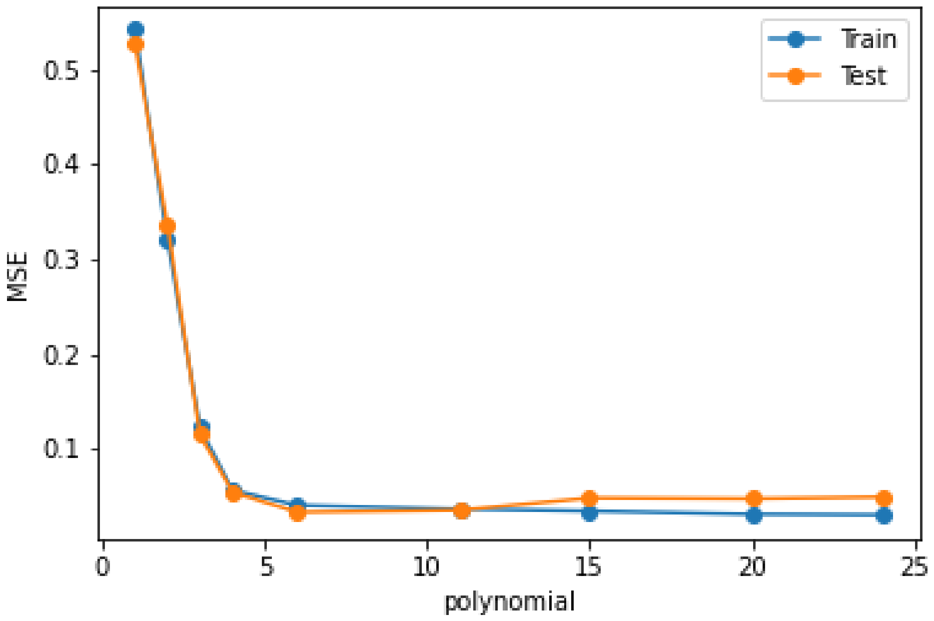


Bias-Variance Trade-off

kNN Classifier



Linear Regression



Linear Regression and Bias-Variance Trade-off

- Multiple linear regression: Revisiting

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} & x_{p+1,1} & \cdots & x_{q,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} & x_{p+1,2} & \cdots & x_{q,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots & & & \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} & x_{p+1,n} & \cdots & x_{q,n} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_p & \beta_{p+1} \cdots & \beta_q \end{bmatrix}^T,$$

- The linear model in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0},\ Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

- Solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Linear Regression and Bias-Variance Trade-off

- The full model can be partitioned into *p* variables and *r*=*q*-*p* variables

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}$$

- The least-squares estimates

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\boldsymbol{\beta}}_p^* \\ \hat{\boldsymbol{\beta}}_r^* \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Linear Regression and Bias-Variance Trade-off

MLR with p variables

- Multiple linear regression: *p (<q) subset of variables*

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

- The linear model in matrix form

$$\mathbf{y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\epsilon}$$

- Solution

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T \mathbf{y}$$

MLR with p variables: Properties of Estimates

- Expected value and variance of $\hat{\beta}_p$

$$E[\hat{\boldsymbol{\beta}}_p] = \boldsymbol{\beta}_p + \mathbf{A}\boldsymbol{\beta}_r, \quad \mathbf{A} = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{X}_r$$

$$Var[\hat{\boldsymbol{\beta}}_p] = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\sigma^2$$

- Properties of $\hat{\beta}_p$ and $\hat{\beta}_p^*$
  - $\hat{\beta}_p$: Biased Estimates of β_p unless β_r=0 or $\mathbf{X}_p^T\mathbf{X}_r = 0$
  - $Var[\hat{\boldsymbol{\beta}}^*] > Var[\hat{\boldsymbol{\beta}}_p]$
  - $\hat{\sigma}_p^2$: Biased estimate of σ^2

Model Complexity: Prediction on *y*

- Effect on model miss specification on prediction?
- Prediction of $\hat{y}^*$ corresponding to **x=[$\mathbf{x_p}$: $\mathbf{x_r}$]**^{T}
  - $\hat{y}^* = \boldsymbol{x^T}\widehat{\boldsymbol{\beta}}^*$

$$E[\hat{y}^*] = \mathbf{x}^T\boldsymbol{\beta}$$

  - $\hat{y} = \boldsymbol{x}_p^T\widehat{\boldsymbol{\beta}}_p$with

$$Var[\hat{y}^*] = \sigma^2(1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x})$$

$$E[\hat{y}] = \mathbf{x}_p^T\,\boldsymbol{\beta}_p + \mathbf{x}_p^T\,\mathbf{A}\boldsymbol{\beta}_r$$

$$Var[\hat{y}] = \sigma^2(1 + \mathbf{x}_p^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_p)$$

Model Complexity: Decomposition of Prediction error

- Properties of $\hat{y}^* = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}^*$ and $\hat{y} = \boldsymbol{x}_p^T\widehat{\boldsymbol{\beta}}_p$
  - $\hat{y}$ is biased unless $\mathbf{X}_p^T\mathbf{X}_r\boldsymbol{\beta}_r = 0$
  - $Var[\hat{y}^*] \geq Var[\hat{y}]$

- Expected prediction error

$$E[(\hat{y}-y)^2] = \underbrace{\sigma^2}_{\text{Irreducible error}} + \underbrace{(\mathbf{x}_p^T Var[\hat{\boldsymbol{\beta}}_p]\mathbf{x}_p)}_{\text{Variance}} + \underbrace{(\mathbf{x}_p^T \mathbf{A}\boldsymbol{\beta}_r - \mathbf{x}_r\beta_r)^2}_{\text{Bias}}$$

Irreducible error

Variance:
Expected s.d. of $\boldsymbol{x}_p^T\widehat{\boldsymbol{\beta}}_p$
around the true mean

Bias

Bias-Variance Trade-off and Prediction error

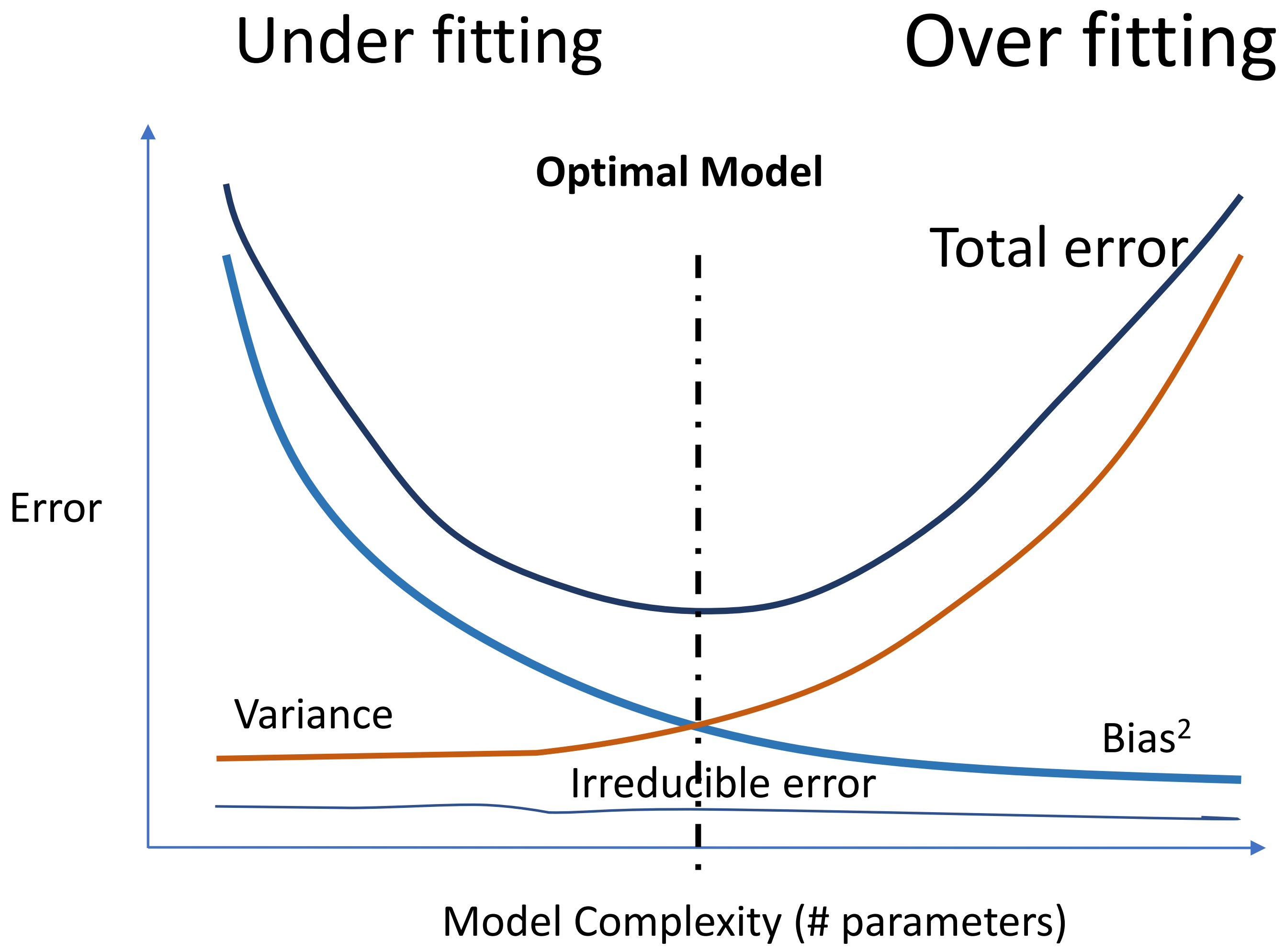
kNN MSE

$$E[(\hat{y}_{x_o} - y)^2] = Var(\epsilon) + \frac{1}{K}\sigma^2 + (f(x_o) - \frac{1}{K}\sum_{i \in \mathcal{A}} f(x_i))^2$$

Linear Regression MSE

$$E[(\hat{y} - y)^2] = \sigma^2 + (\mathbf{x}_p^T Var[\hat{\boldsymbol{\beta}}_p]\mathbf{x}_p) + (\mathbf{x}_p^T \mathbf{A}\boldsymbol{\beta}_r - \mathbf{x}_r\beta_r)^2$$

Bias-Variance Trade-off



Model Validation: Methods

Model Validation

Analytical Methods
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Mallows Cp

Resampling Methods
- Validation set approach
- K-fold cross validation
- Bootstrap

Subset selection
- Best subsets
- Backward elimination
- Forward Selection
- Stepwise regression

Analytical Methods: AIC

- AIC defined as:

$$AIC_p = \frac{SSE_p}{n} + 2\frac{p}{n}\hat{\sigma}^2$$

Penalty on
# of parameters

- Balance between demands of accuracy (fit, first term) and simplicity of model (second term)
- Penalty on the model with the large number of variables
- Smaller the value of AIC, the better the model

Analytical Methods: BIC

- BIC defined as:

$$BIC_p = \frac{SSE_p}{\sigma^2} + log(n)p$$

Penalty

- The number of observations also plays a role
- A good model: A small value of BIC
- $n > e^2$, BIC penalize model having large p
- Selects simpler models

Analytical Methods: AIC vs BIC

- For given a set of models (including the true model), BIC is asymptotically consistent
  - The probability that BIC chooses the correct model tends to 1 as *n* tends to infinity
  - AIC chooses relatively complex model as *n* tends to infinity
- Finite sets, BIC chooses models that are too simple in comparison of AIC

Example 1

- Linear regression model,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \text{error}_i$$

$$\beta_0 = -560,\ \beta_1 = 0.08,\ \beta_2 = 1.56$$

- Given data for *52* observations:

$$y,\ x_1,\ x_2,\ x_3$$

- Objective:
  - Find a relationship between y and xs

Example 1

- Build a full regression model

```
Call:
lm(formula = Y_noisy ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4246  -3.2720   0.3424   2.6490  13.0606

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.754e+02  1.669e+01 -34.466   <2e-16 ***
X1           8.116e-02  1.568e-03  51.766   <2e-16 ***
X2           1.586e+00  4.240e-02  37.413   <2e-16 ***
X3           2.777e-03  6.926e-03   0.401     0.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.441 on 45 degrees of freedom
Multiple R-squared:  0.9899, Adjusted R-squared:  0.9892
F-statistic:  1473 on 3 and 45 DF,  p-value: < 2.2e-16
```

Example 1

- Build a reduced regression model,

```
Call:
lm(formula = Y_noisy ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.9277  -3.6406   0.3646   3.1255  13.1142 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.744e+02  1.639e+01  -35.06   <2e-16 ***
X1           8.152e-02  1.271e-03   64.16   <2e-16 ***
X2           1.584e+00  4.159e-02   38.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.391 on 46 degrees of freedom
Multiple R-squared:  0.9899,	Adjusted R-squared:  0.9894 
F-statistic:  2251 on 2 and 46 DF,  p-value: < 2.2e-16
```

Example 1: Cont…

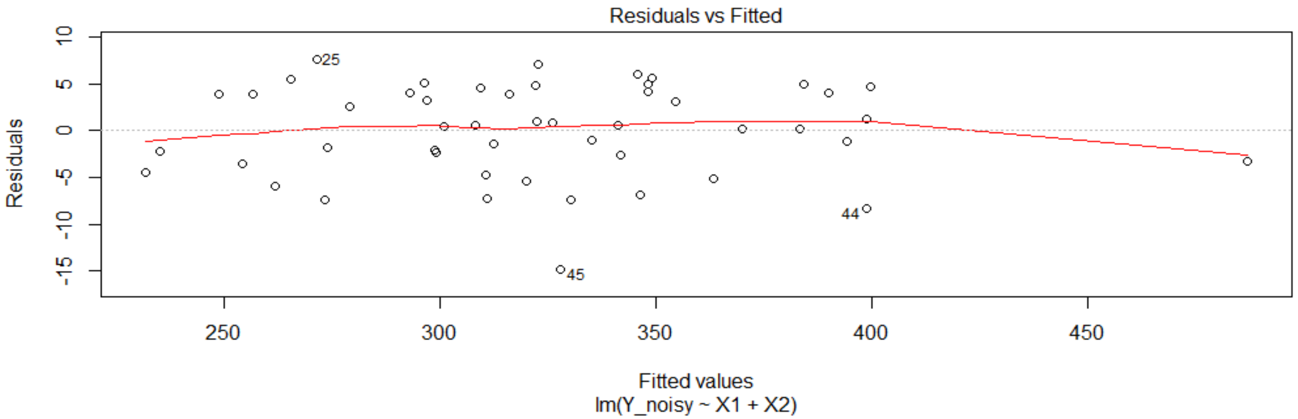
- Fit a linear model with x_1, x_2, x_3
- Fitting objective function: SSE

Models $f(x_i, \Theta)$	**AIC**	**BIC**
$B_0 + B_2x_1 + B_3x_3$	302.0091	311.4682
$B_0 + B_1x_1 + B_2x_2 + B_3x_3$	303.6010	314.9519
$B_0 + B_1x_1 + B_2x_2$	301.3854	308.9527
$B_0 + B_1x_1$	474.3469	480.0223

Minimum AIC & BIC

Example 1:Cont....

- Residual plot analysis: No patterns, linearity assumption holds



- Parameter estimates for the model $\beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$\beta_{0,e} = -543.43(\mathit{-560}),$$
$$\beta_{1,e} = 0.078(\mathit{0.08}),$$
$$\beta_{2,e} = 1.53(\mathit{1.56})$$

Example 2

- Nonlinear regression model,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1i}^{\ 2} + \text{error}_i$$

$$\beta_0 = 560,\ \beta_1 = 3,\ \beta_2 = 1.56,\ \beta_3 = 0.08$$

- Given data:

$$y,\ x_1,\ x_2,\ x_3$$

- Objective:
  - Find a relationship and parameters

Example 2

- Build a full linear regression,

```
Call:
lm(formula = Y_noisy2 ~ ., data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-36067 -25670 -10913  19164  85646

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.896e+06  1.011e+05 -18.744   <2e-16 ***
X1           7.574e+02  9.526e+00  79.512   <2e-16 ***
X2           4.567e+02  2.584e+02   1.768   0.0837 .
X3           3.327e+00  4.220e+01   0.079   0.9375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33230 on 46 degrees of freedom
Multiple R-squared:  0.9956, Adjusted R-squared:  0.9953
F-statistic:  3493 on 3 and 46 DF,  p-value: < 2.2e-16
```

Example 2: Cont…

- Fit linear model with x_1, x_2, x_3

Models $f(x_i, \Theta)$	AIC	BIC
$B_0 + B_1x_1+B_2x_2+B_3x_3$	1188.846	1198.406
$B_0 + B_1x_1+B_3x_3$	1190.131	1197.779
$B_0 + B_1x_1+B_2x_2$	1186.853	1194.501
$B_0 + B_1x_1$	1188.158	1193.894
$B_0+B_2x_2+B_3x_3$	1433.363	1441.011

Minimum AIC & BIC

Example 2

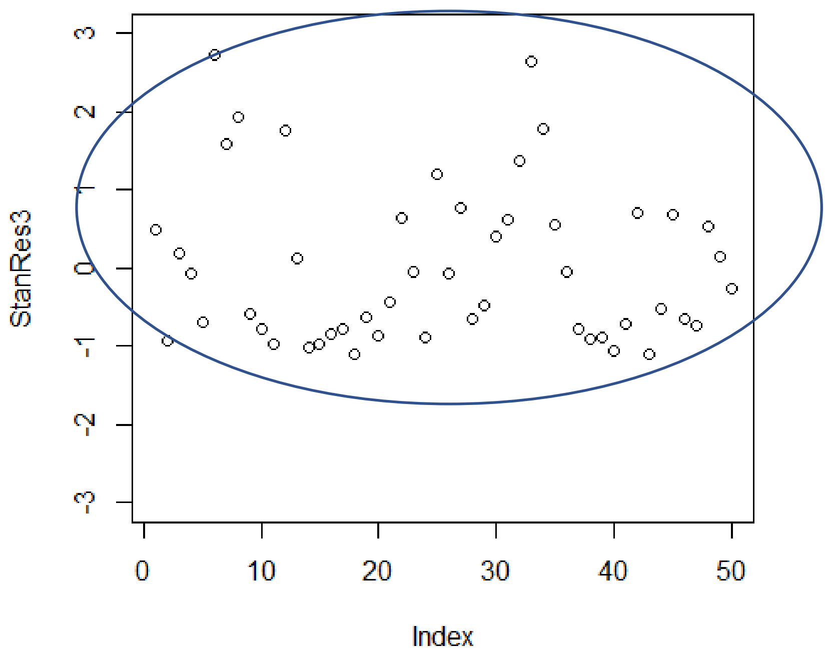
- Build a linear regression model R output,

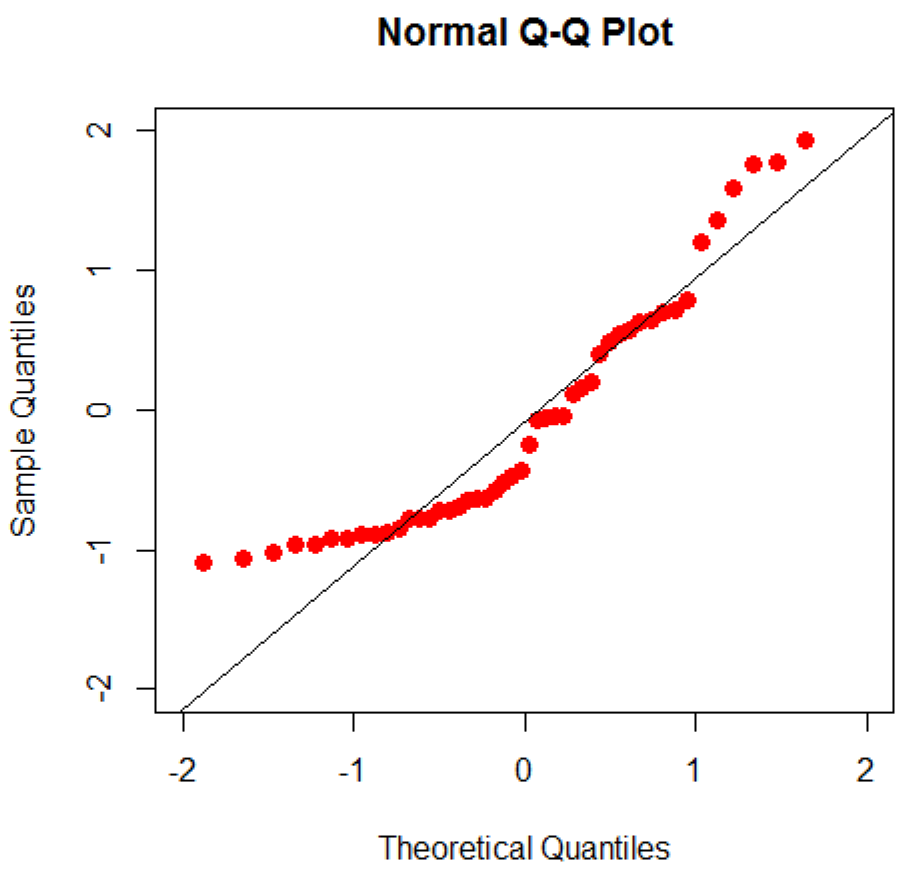
```
Call:
lm(formula = Y_noisy2 ~ X1 + X2)

Residuals:
   Min     1Q Median     3Q    Max 
-35675 -25150 -10864  19325  85446 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.895e+06  9.923e+04 -19.095   <2e-16 ***
X1           7.579e+02  7.632e+00  99.309   <2e-16 ***
X2           4.540e+02  2.532e+02   1.793   0.0795 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32880 on 47 degrees of freedom
Multiple R-squared:  0.9956,	Adjusted R-squared:  0.9954 
F-statistic:  5352 on 2 and 47 DF,  p-value: < 2.2e-16
```

Improved model but x_2 coefficient is not significant

Example 2

- Diagnostics Plot





No outliers

Pattern: Yes, Shifting of error

Normally distributed errors: Not sure

Nonlinear model?

Example 2: Cont…

- Fit nonlinear model with x_1, x_2 and nonlinear terms

Models $f(x_i, \Theta)$	**AIC**	**BIC**
$B_0 + B_1x_1+B_2x_2+B_3x_1x_2$	1183.4581	1193.018
$B_0 + B_1x_1+B_2x_2+B_3x_1^2$	420.1689	429.729
$B_0 + B_1x_1+B_2x_2+B_3x_2^2$	1188.2036	1197.764
$B_0 + B_1x_1+B_2x_2+B_3x_1^2+B_4x_2^2$	422.0468	433.519

Minimum AIC & BIC

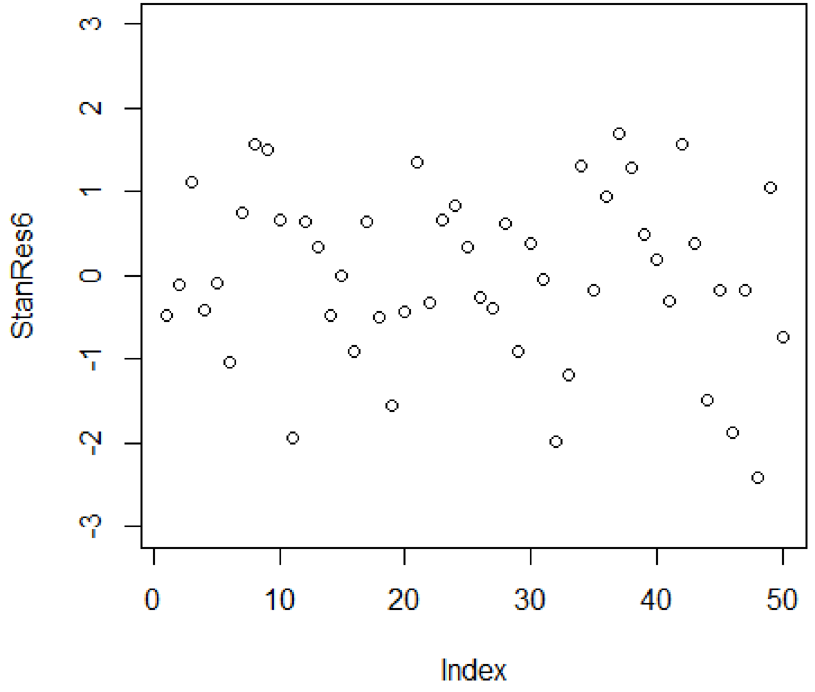
Example 2

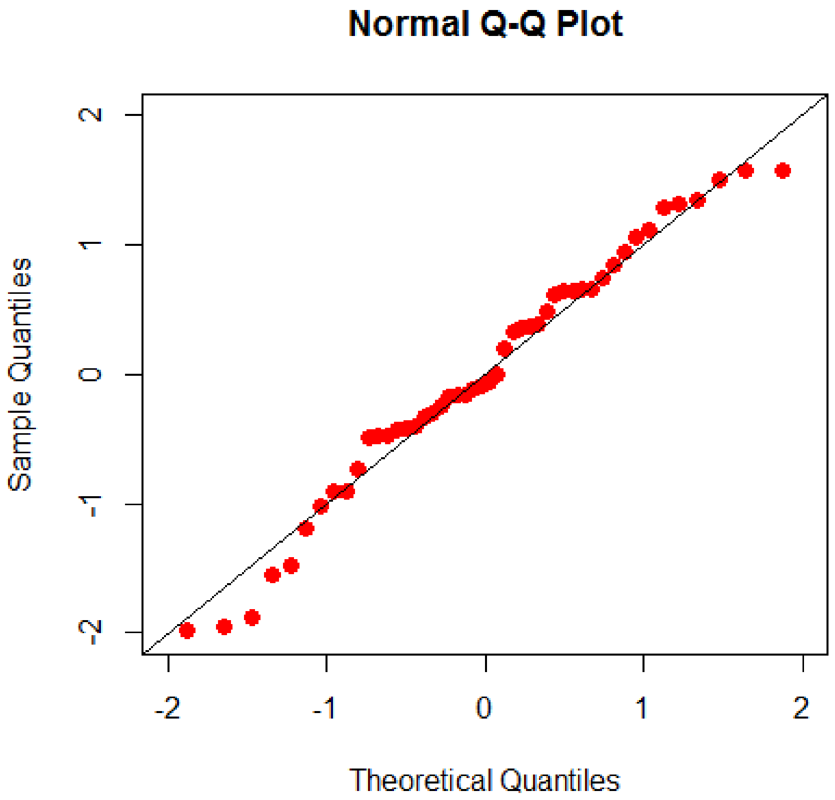
- Build a nonlinear regression model,

```
Call:
lm(formula = Y_noisy2 ~ X1 + X2 + X1sr)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.334  -7.403  -1.110  10.375  26.702 

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)    
(Intercept) 4.062e+02  1.430e+02     2.841  0.00667 ** 
X1          3.049e+00  5.372e-02    56.762  < 2e-16 ***
X2          1.669e+00  1.274e-01    13.106  < 2e-16 ***
X1sr        7.999e-02  5.680e-06 14084.807  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16 on 46 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 1.513e+10 on 3 and 46 DF,  p-value: < 2.2e-16
```

Example 2

- Diagnostics





No outliers
No Pattern

Normally distributed errors

Example 2

- Nonlinear regression model,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i}^2 + \text{error}_i$$

$$\beta_0 = 560,\ \beta_1 = 3,\ \beta_2 = 1.56,\ \beta_3 = 0.08$$

- Given data:

$$y,\ x_1,\ x_2,\ x_3$$

- Objective:
  - Find a relationship and parameters

```
Call:
lm(formula = Y_noisy2 ~ X1 + X2 + X1sr)

Residuals:
    Min      1Q  Median      3Q     Max
-37.334  -7.403  -1.110  10.375  26.702

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 4.062e+02  1.430e+02     2.841  0.00667 **
X1          3.049e+00  5.372e-02    56.762  < 2e-16 ***
X2          1.669e+00  1.274e-01    13.106  < 2e-16 ***
X1sr        7.999e-02  5.680e-06 14084.807  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16 on 46 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 1.513e+10 on 3 and 46 DF,  p-value: < 2.2e-16
```

Subset selection methods

- *Automatically picking variables from a global model?*
- Methods: (i) Best subsets, and (ii) Stepwise
- Best subsets:
  - Builds 2^p where p: # of parameters
  - Best according to some statistics such as AIC, BIC, C_p
- Stepwise:
  - consider a smaller set of models
  - Three approach
    - Backward elimination
    - Forward selection
    - Stepwise regression

Subset selection: Backward Elimination

- Procedure to select variables
  1. Fit the full model with all p independent variables (IV) and
  2. Fit the model by removing each IV and compute AIC values
  3. Find the IV whose removal leads to minimum AIC
  4. Remove the corresponding IV
  5. Fit a new model without this variable and Go to Step 2
  6. If the removal of any variable does not reduce AIC, stop and keep current model

Subset selection: Forward Selection

- Procedure to select variables
  1. Fit all simple regression models
  2. Compute AIC
  3. Consider the IV with the lowest AIC
  4. Fit all two variable models including this variable and Go to Step 2
  5. If Current AIC > previous AIC, stop and keep previous model

Subset selection: Stepwise Regression

- Procedure to select variables
  1. Start like Backward Selection
  2. Find the IV whose removal leads to minimum AIC and remove the IV, and $AIC_{removal}$= minimum AIC for the next iteration
  3. New IV must have $AIC_{enter} \leq AIC_{removal}$ to enter
  4. Re-test all "*old variables*" that have already been entered,
  5. Old variables must have $AIC_{enter} > AIC_{with\ IV}$ to stay in model
  6. Continue until no new variables can be entered and no old variables need to be removed

Example 2:Cont…

Term removal(-) or addition(+)	AIC
- x_2^2	278.41
- x_2	278.71
- x_3	278.74
- x_1x_2	278.78
None	280.35
- x_1	431.86
- x_1^2	1040.28

Model: $B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_1^2 + B_5x_2^2 + B_6x_1x_2$

Remove x_2^2

$AIC_{removal} = 278.41$

Term removal(-) or addition(+)	AIC
- x_1x_2	276.79
- x_3	276.93
None	278.41
+ x_2^2	280.35
- x_2	281.74
- x_1	432.12
- x_1^2	1039.88

Model: $B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_1^2 + B_6x_1x_2$

Remove x_1x_2

$AIC_{enter} =< AIC_{removal} = 278.41$

Add x_2^2, $AIC_{with\ IV} > AIC_{enter}$

$AIC_{removal} = 276.79$

Example 2:Cont…

Term removal(-) or addition(+)	AIC
- x_3	276.28
None	276.79
+ x_1x_2	278.41
+ x_2^2	278.78
- x_2	349.09
- x_1	493.16
- x_1^2	1044.95

Model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2$

Remove x_3

$AIC_{enter} =< AIC_{removal} = 276.79$

Add x_1x_2 , $AIC_{with\ IV} > AIC_{enter}$

Add x_2^2 , $AIC_{with\ IV} > AIC_{enter}$

Term removal(-) or addition(+)	AIC
None	276.28
+ x_3	276.79
+ x_1x_2	276.93
+ x_2^2	278.15
- x_2	347.16
- x_1	491.32
- x_1^2	1042.96

$AIC_{removal} = 276.28$

Model: $B_0 + B_1x_1 + B_2x_2 + B_4x_1^2$

Add x_3 , $AIC_{with\ IV} > AIC_{enter}$

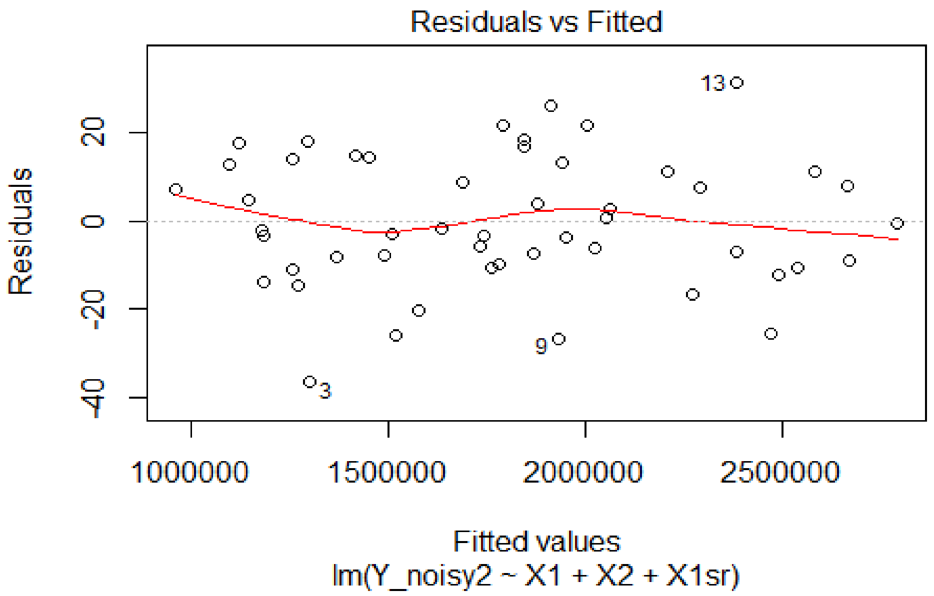
Add x_1x_2 , $AIC_{with\ IV} > AIC_{enter}$

Add x_2^2 , $AIC_{with\ IV} > AIC_{enter}$

No improvement in AIC by adding or removing IVs

Example 2: Cont…

- Identified model: $B_0 + B_1x_1 + B_2x_2 + B_4x_1^2$
- Residual plot analysis



Parameter estimates

$B_{0,e}$= 541.6(*560*),
$B_{1,e}$= 3.02(*3*),
$B_{2,e}$= 1.49(*1.56*)
$B_{3,e}$= 0.08(*0.08*)

Subset selection: Best Subset Selection

- Best model from among the 2^q possibilities
- Algorithm to select best model in two stages
- Procedure to select variables
  1. Let P_0 denote the null model (only intercept)
  2. For $l=1,2,\ldots,q$
     a. Fit all ${}^{q}c_{l}$ models that contain exactly l predictors
     b. Pick the best among these ${}^{q}c_{l}$ models based on AIC or BIC, and call it P_k
  3. Select a single best model from among $P_0, \ldots, P_q$

Algorithm reduces the problem to one of q+1 possible models

Resampling Methods

- Validation of models by repeatedly drawing random samples from a training set
  - K-fold cross validation
  - Bootstrap
- **Objective**:
  - Predict the performance of model(s) on the test sets using the training sets
- Resampling methods useful for data scarce situations

Resampling Methods

- Consider the following data set
  - Training set: $\{(\boldsymbol{x}_1, y_1);(\boldsymbol{x}_2, y_2);\ldots;(\boldsymbol{x}_n, y_n)\}$
  - Test point: $(\boldsymbol{x}_0, y_0)$ such n_t observations
- Training error rate

$$MSE_{Training} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2$$

Not of our interest for predictive ability of the model

- Test error rates

$$\boxed{MSE_{Test} = \frac{1}{n_t}\sum_{i=1}^{n}(y_{0,i} - \mathbf{x}_{0,i}^T\hat{\boldsymbol{\beta}})^2}$$

Of our interest

Data scarcity: Test data are not available

Resampling Methods

- Expected test error

$$E[(\hat{y}_0 - y_0)^2] = \underbrace{\sigma^2}_{\text{Irreducible error}} + \underbrace{(\mathbf{x_0}_p^T Var[\hat{\boldsymbol{\beta}}_p]\mathbf{x_0}_p)}_{\text{Variance}} + (\underbrace{\mathbf{x_0}_p^T \mathbf{A}\boldsymbol{\beta}_r - \mathbf{x_0}_r \beta_r}_{\text{Bias}})^2$$

- Interpretation of variance: The amount by which $\widehat{\boldsymbol{\beta}}_p$ would change if we estimate it using different training sets
- Interpretation of Bias: The amount of error introduced by approximating a problem with a simpler model
- *Select the model that achieves low variance and low bias*

Resampling Methods

- Random Sampling
  - Select a subset of data with equal probability of being chosen
  - Large data set
  - Small data set or imbalance classification dataset: introduce sampling bias

Resampling Methods

- Stratified Sampling
  - Select a subset of data from each stratum with equal probability of being chosen
  - Provides the representation of data set in training and test phases
  - Reduces sampling bias

Validation Set Approach

- Enough data: (1) Training set, (2) Validation set, and (3) Test set
- Not enough data: Generate validation sets from a training set
- Validation set approach: Divides (often randomly) the training set into two parts

1 2 3 4 n

  - A training set

1 2 3 4 n_t

  - A validation set (or hold-out set)

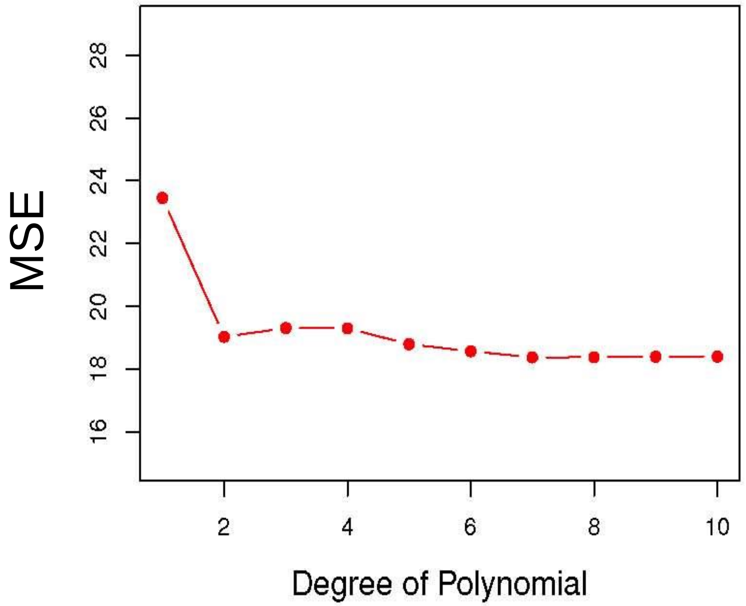
1 2 3 4 n_v

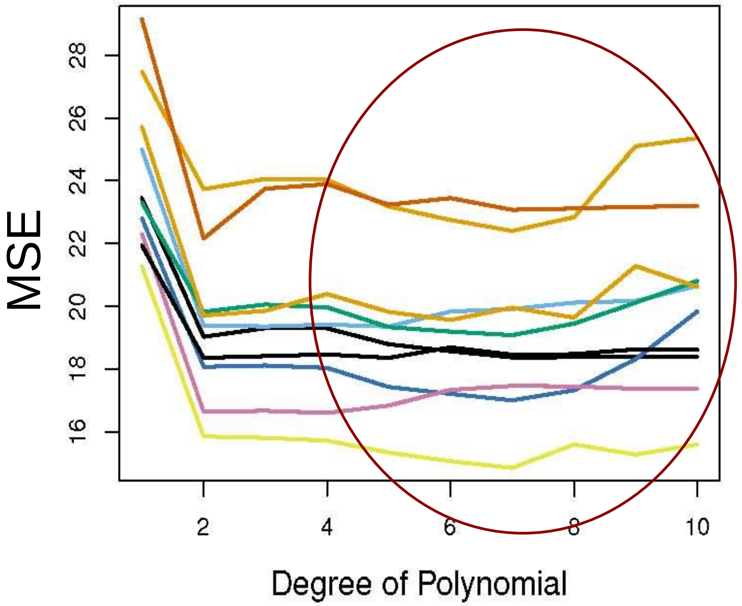
- Use training set, to fit the model
- Use validation set, to predict validation set errors

  Provides an estimate of test error rates

Validation Set Approach: Example

- Example: mileage~ horsepower[1]
- Nonlinear Model: mileage~*f*(horsepower)



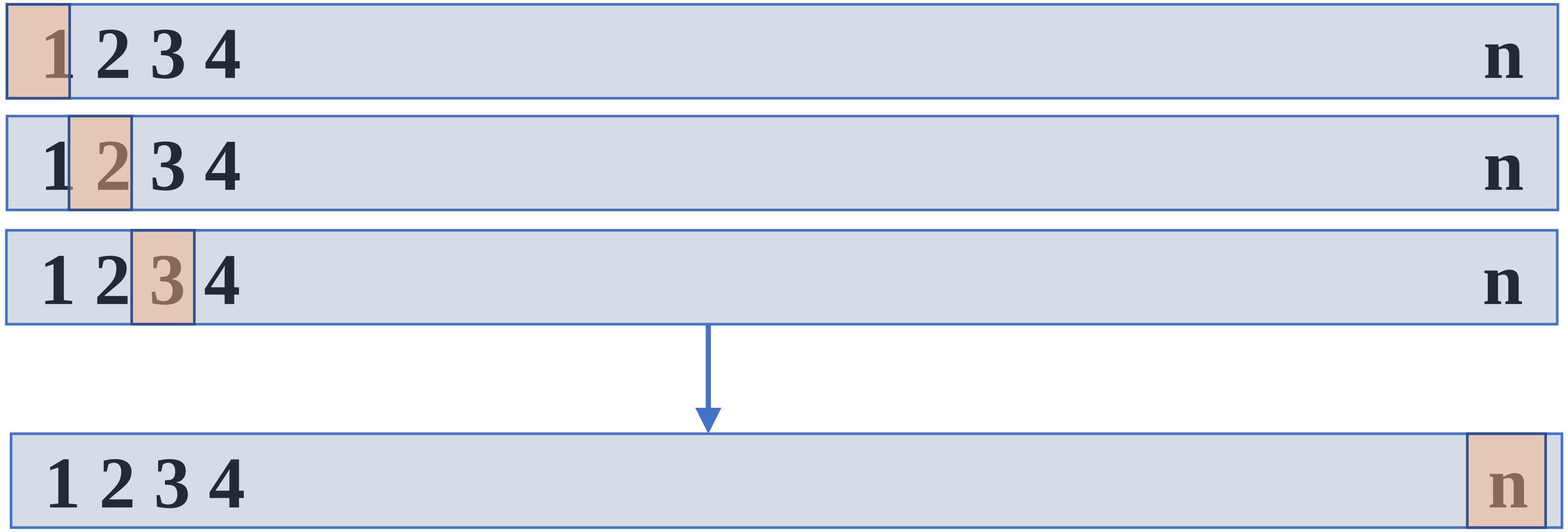


High variability in estimates of test error

[1]Tibshirani et al (2013)

Leave-one-out-cross-validation (LOOCV)

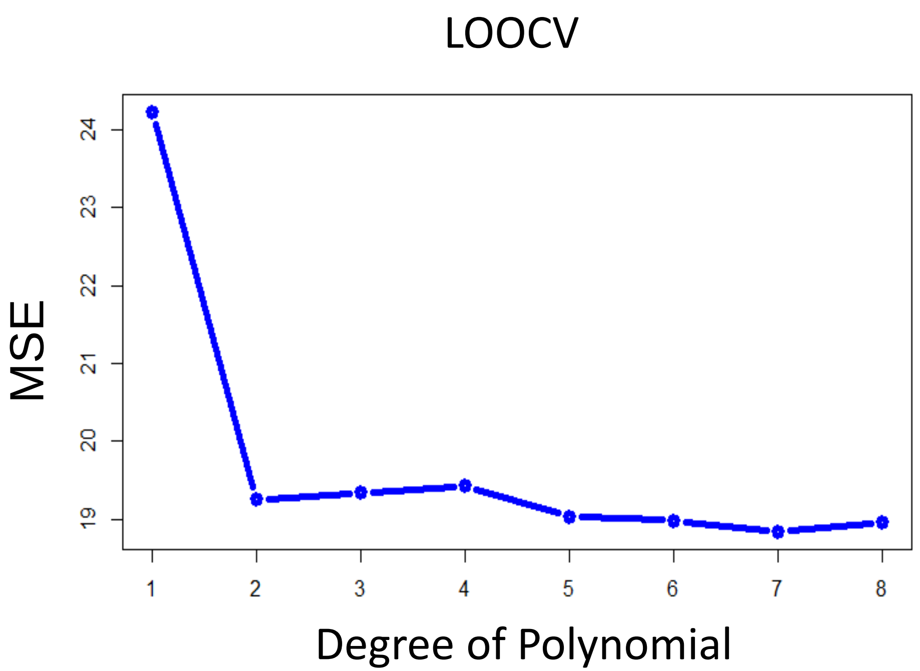
- Build model using *(n-1)* samples and predict the response (y_i) for *the remaining sample*

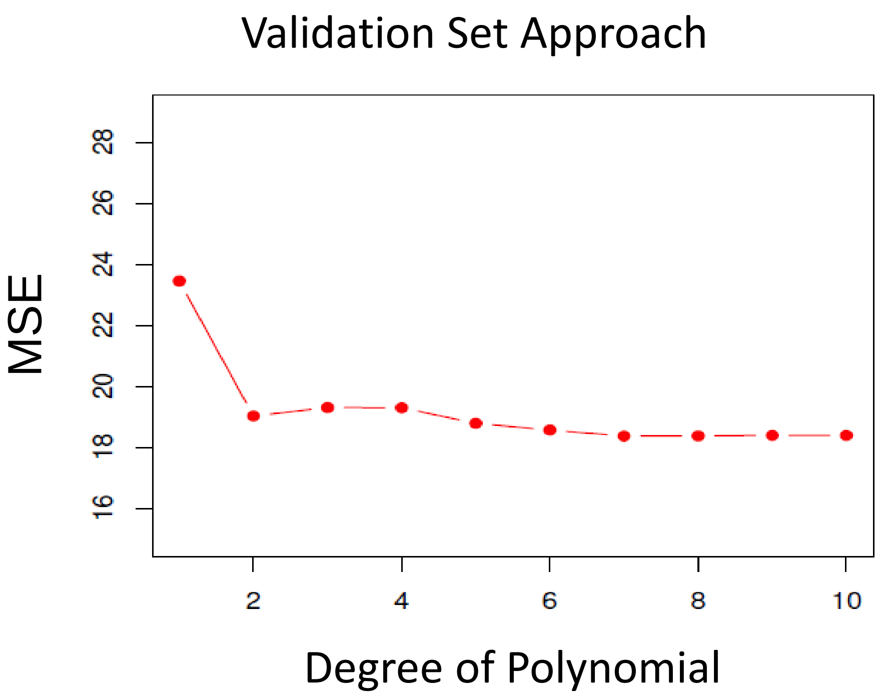


$$CV_1 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(1)})^2$$

LOOCV: Example

- Example: mileage~ horsepower[1]
- Nonlinear Model: mileage~*f*(horsepower)





[1]Tibshirani et al (2013)

Leave-one-out-cross-validation (LOOCV)

- Advantages
  - Far less bias comparison to the validation set approach

    Training set contains *(n-1)* observations each iteration
  - Yield the same results

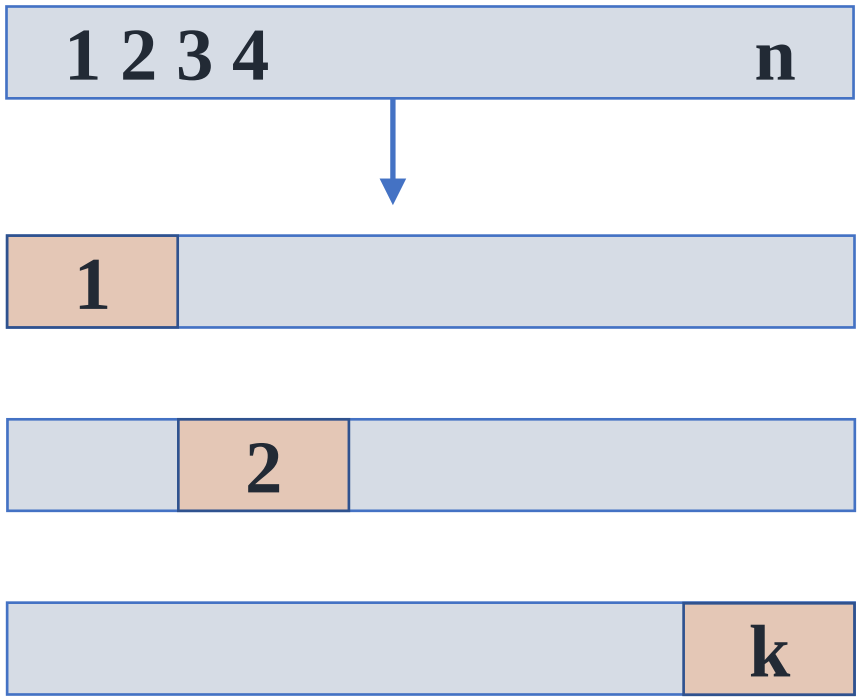
    No randomness in the training/validation set splits
  - Does not overestimate the test error rate as much as the validation set approach
- Disadvantages
  - Expensive to implement due to fitting happens *n* times
  - Asymptotical incorrect (n tends to infinity) it does not choose correct model
  - It may select a model of excessive size (more variables) than the optimal model

k-Fold Cross Validation

- Training data into k disjoint samples of equal size,

  $Z_1, Z_2 \ldots, Z_k$
- For each validation sample Z_i
  - Use remaining data to fit the model
  - Predict the response for the validation sample Z_i and compute mean square error (MSE_i),
  - Repeat for all k samples
  - The k-fold CV

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$



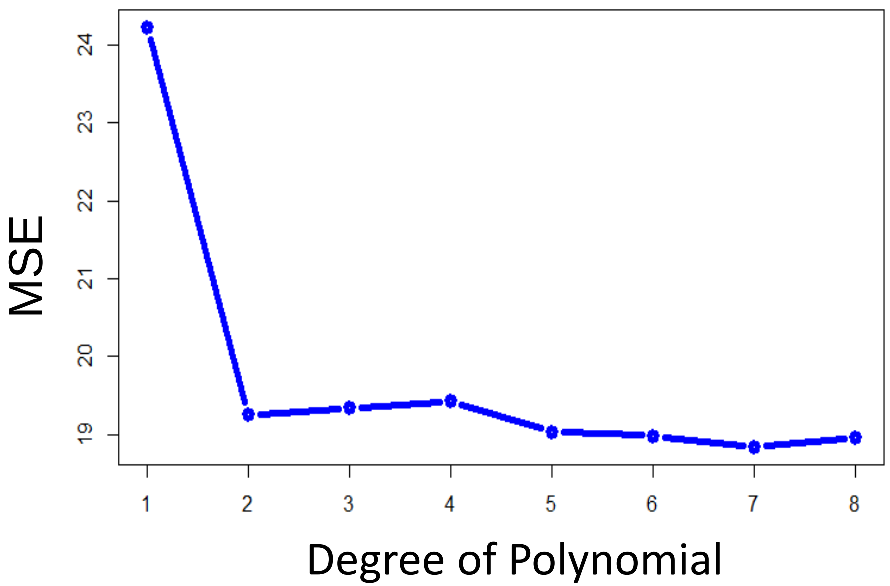
k-fold Validation

- For $k=n$, Leave-one-out-cross-validation (LOOCV)
- In practice, k=5 or 10 is taken,
- Less computation cost
- For computationally intensive learning methods
  - LOOCV fits the model n times
  - k-fold CV fits the model k times

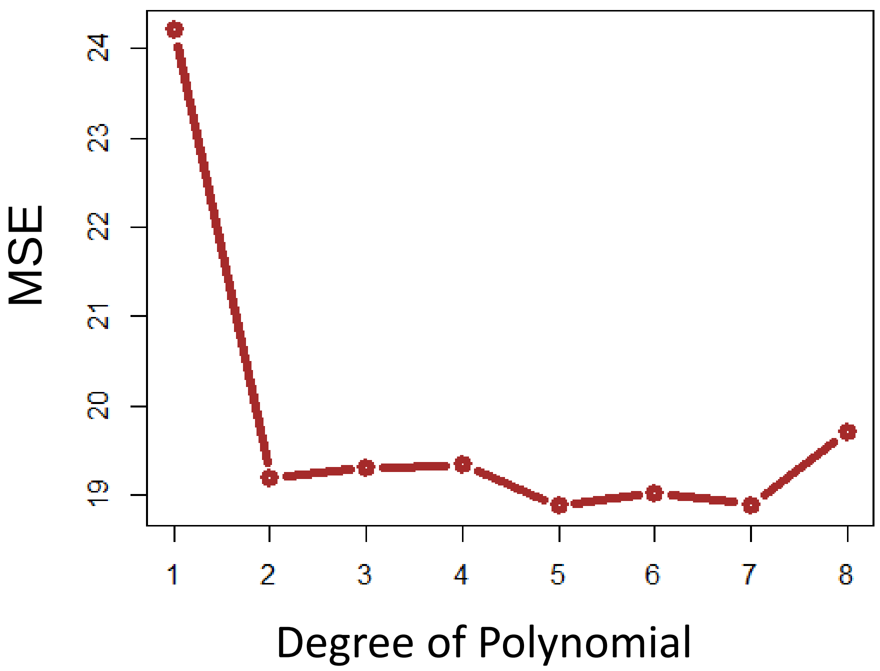
k-fold CV: Example

- Example: mileage~ horsepower[1]
- Nonlinear Model: mileage~*f*(horsepower)

LOOCV



10-fold CV

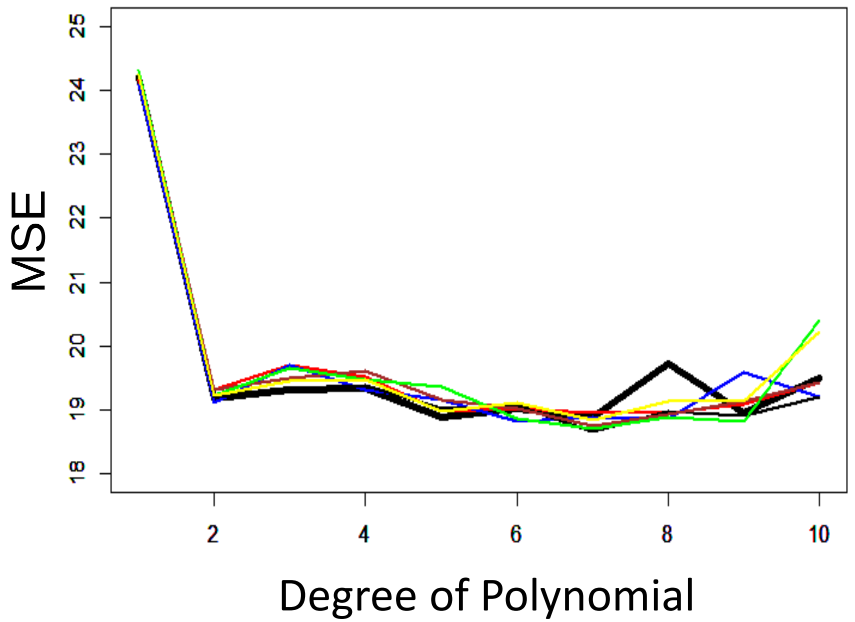


[1]Tibshirani et al (2013)

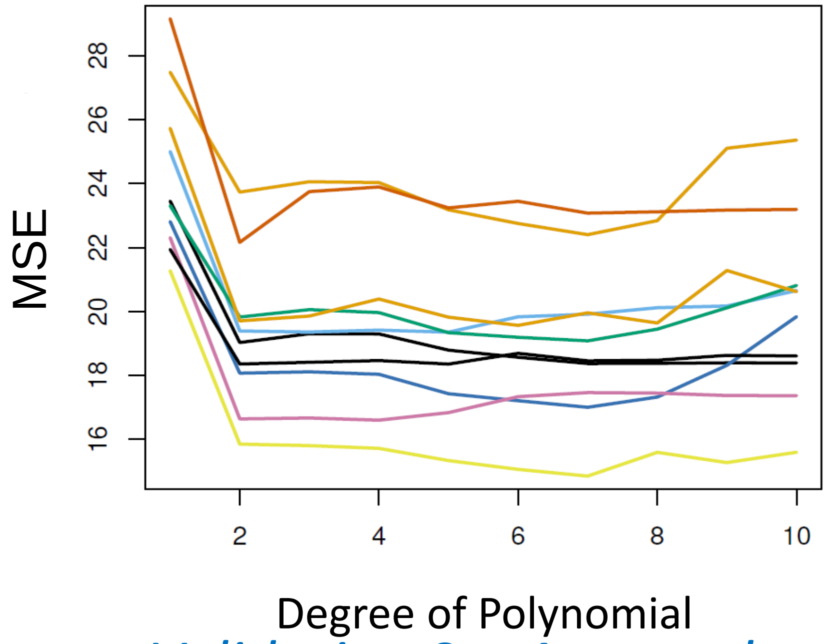
k-fold CV: Example

- Example: mileage~ horsepower[1]
- Nonlinear Model: mileage~*f*(horsepower)

10-fold CV



Validation Set Approach



k-fold CV has lower variability in comparison to Validation Set Approach

[1]Tibshirani et al (2013)

k-fold CV: Bias-Variance Trade-off

- Bias reduction in test error: LOOCV is preferred
  - LOOCV provides nearly unbiased estimates: $(n\text{-}1)$ observations in training set
  - k-fold CV provides intermediate level of biased estimates: $(k\text{-}1)n/k$ observations in training set
- Variance reduction in test error: k-fold CV
  - LOOCV leads to higher variance: Training on almost identical $(n\text{-}1)$ observations
  - k-fold CV ($k<n$) leads to lower variance: Training on $(k\text{-}1)n/k$ observations having overlap between the training sets in each model is smaller

5- or 10-fold CV yields test error rate estimates having moderate bias and variances

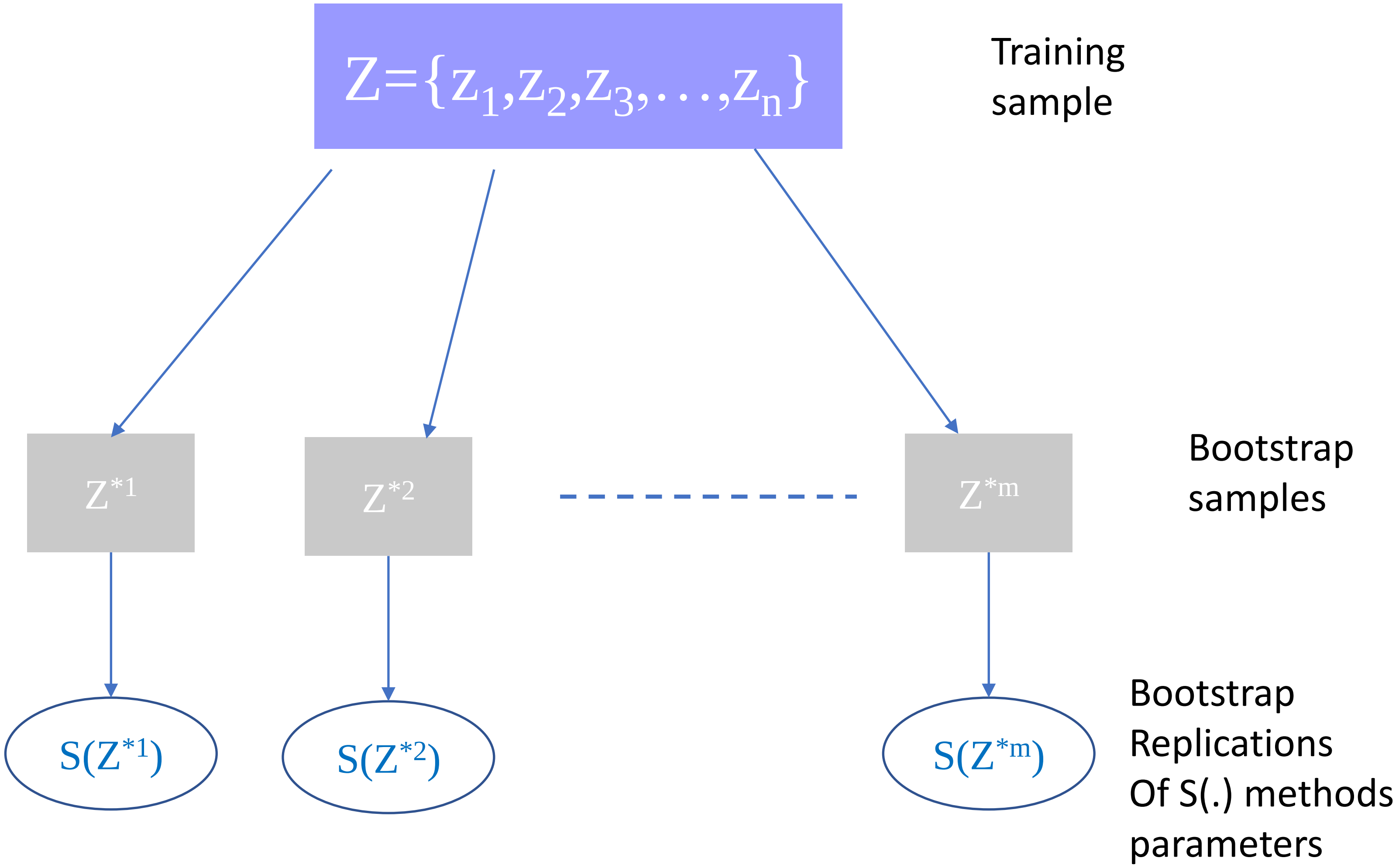
Cross-validation: Classification Problems

- Quantitative outcome y_i of Regression problems
- In CV, MSE is used to quantify test error
- Classification problem: y_i is qualitative
- CV?
- Use the number of misclassified observations
- LOOCV error rate

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} Err_i$$

with $Err_i = I(y_i \neq \hat{y}_i)$, I is an indicator function

Bootstrap



$Z=\{z_1,z_2,z_3,\ldots,z_n\}$
Training sample
Z^{*1}
Z^{*2}
Z^{*m}
Bootstrap samples
$S(Z^{*1})$
$S(Z^{*2})$
$S(Z^{*m})$
Bootstrap Replications Of S(.) methods parameters

Bootstrap

- Normally used for quantifying the uncertainty associated with a given estimator
- Training set: $Z=\{z_1,z_2,\ldots,z_n\}$ where $z_i=(x_i,y_i)$
- Draw samples with replacement from the training set such that each sample size = original training size
- Repeat the sampling for *m* times: m data sets Z^{*m}
- Compute the quantity of interest (ex. Regression parameters) from the each data set
- Estimation of prediction errors

$$MSE_{boot} = \frac{1}{m}\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T\hat{\boldsymbol{\beta}}^{*m})^2$$

Bootstrap

- Estimation of prediction error

$$MSE_{boot} = \frac{1}{m}\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T\hat{\boldsymbol{\beta}}^{*m})^2$$

- MSE_{boot} does not provide a good estimate, why?
- The original training set is acting as test set
- Boot strap sets are near to the training set
- A better bootstrap estimate of prediction error is

$$\boxed{MSE_{boot} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{m\in C^{-i}}(y_i - \mathbf{x_i}^T\hat{\boldsymbol{\beta}}^{*m})^2}$$

where C^{-i} the set of indices of the sample m that not having i^{th} observation

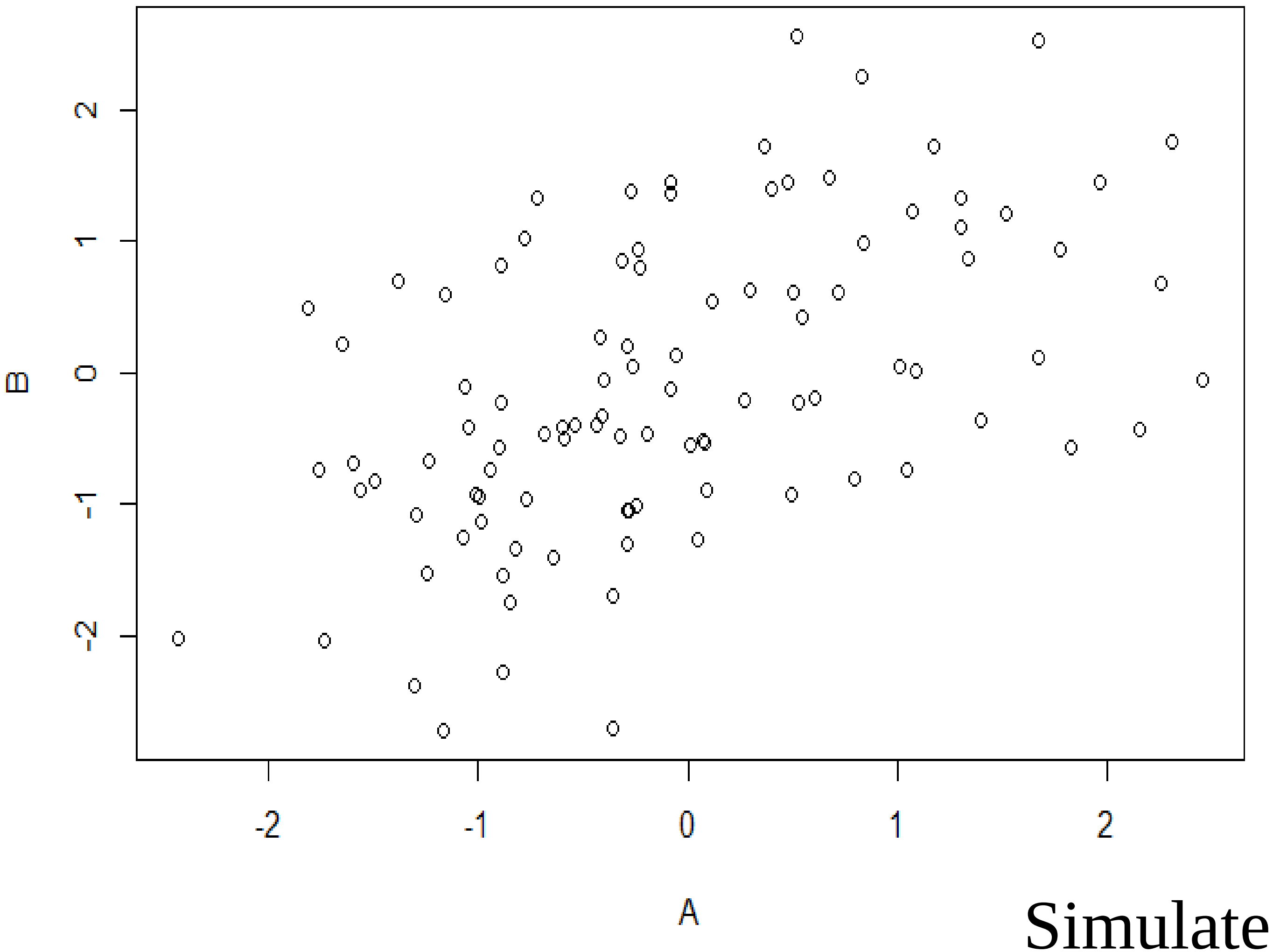
Bootstrap: Example

- Two instruments: A and B
- Property C= αA+(1- α)B, α is a parameter
- Variability associated with each instrument
- Objective : Choose α such that variance of C is minimized
- α value at minimum var(C) can be given by

$$\alpha = \frac{\sigma_B^2 - \sigma_{AB}}{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}}$$

- $\sigma_A^2, \sigma_B^2, \sigma_{AB}^2$: Unknown
- Compute them using past data sets

Bootstrap: Example



Simulated α=0.6

Bootstrap: Example

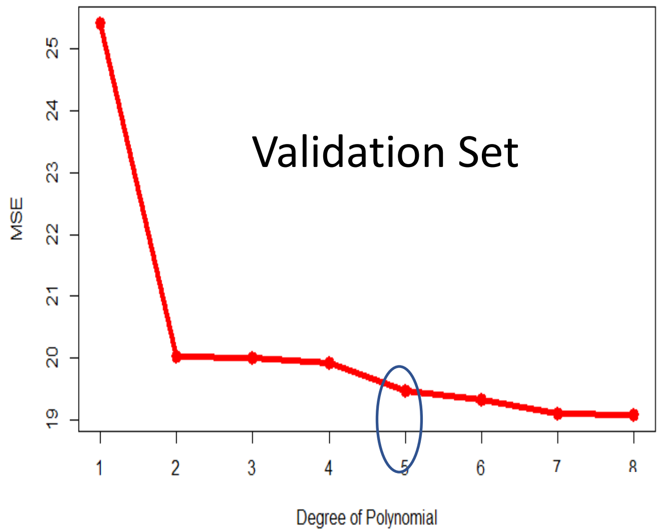
- n=100 observations
- m= n Bootstrap samples
- Compute unknown estimates of Quantities $\hat{\sigma}_A^2$,$\hat{\sigma}_B^2$, $\hat{\sigma}_{AB}^2$ and
- $\hat{\alpha}$ for each bootstrap sample using

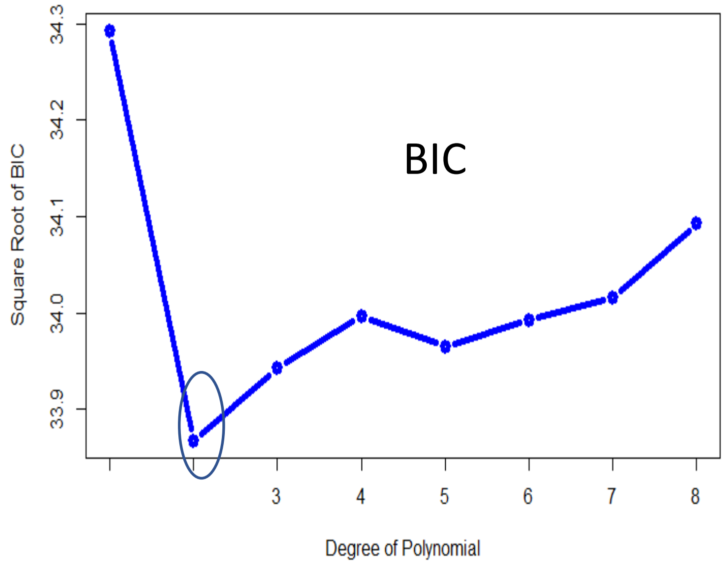
$$\alpha = \frac{\sigma_B^2 - \sigma_{AB}}{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}}$$

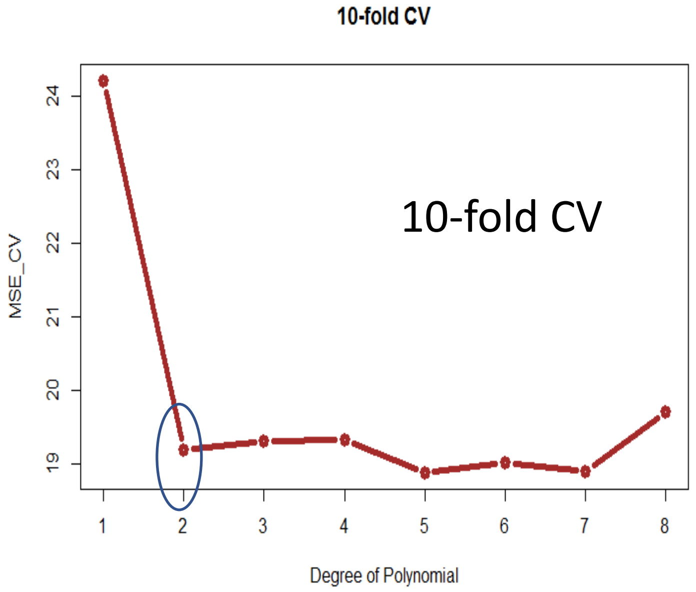
$$\hat{\alpha}=0.5964$$

Conclusion:

Choosing the Optimal Model?

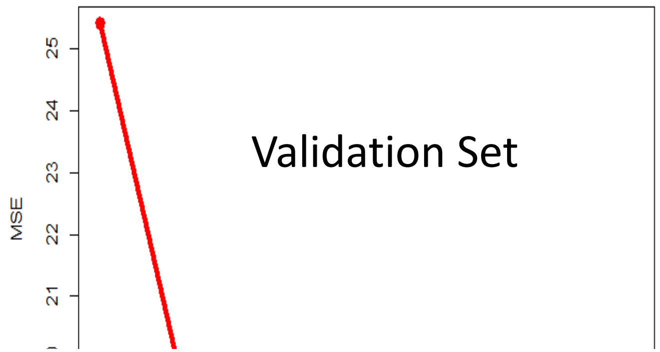


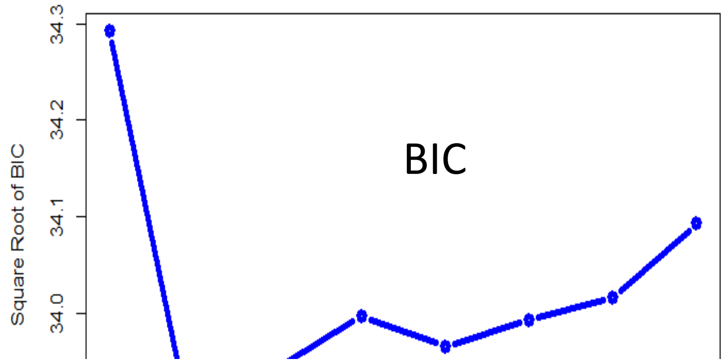




Conclusion:

Choosing the Optimal Model?

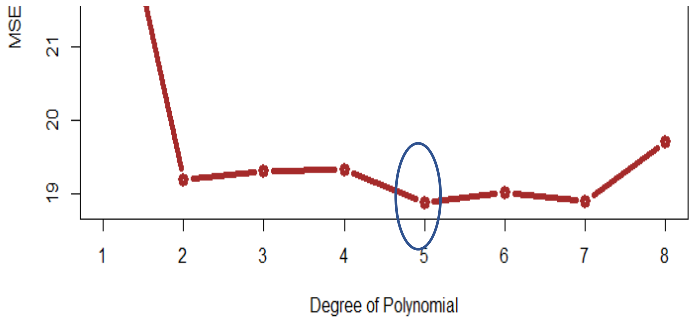




One-standard-error rule

Compute standard error of the test MSE for each model

Select the smallest model for which the test error is within one standard error of the lowest point on the curve



References

- Gareth J, Daniela W, Trevor H, Robert T. An introduction to statistical learning: with applications in R. Spinger; 2021 (Chapter 2 and Chapter 5)
- Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics; 2001. (Chapter 7)