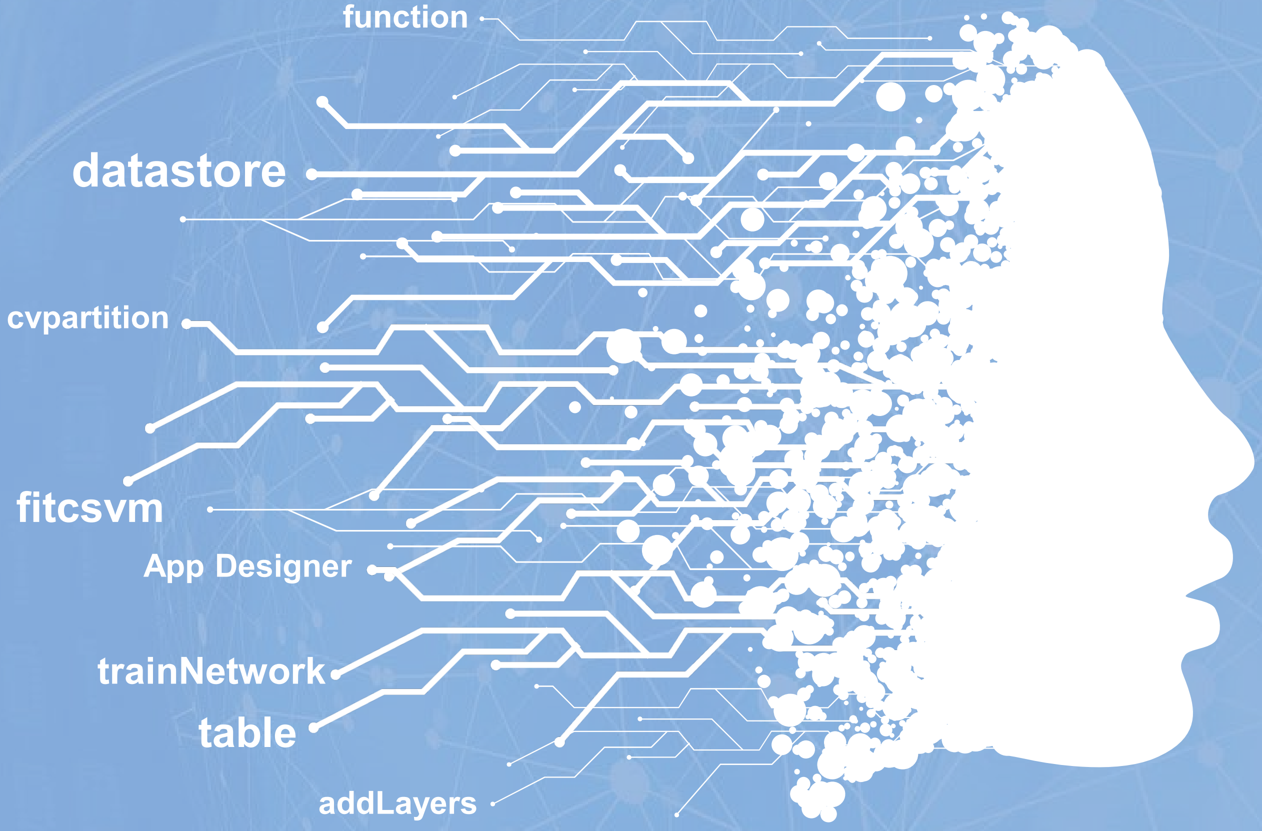
Team Project Report

- **Record a 10-minute video (40%, before 6/5) & participate in Q&A (30%)**
- **Please include the following information in your report**
  - Objectives and division of work
  - Description and citation of dataset
  - Machine learning method and main content of the code
  - Demonstration of App use (data input/presentation, machine learning parameters setting/training, performance evaluation)
  - Discussion: Problems encountered and solutions, follow-up improvement and application
- **Email to alvin4016@nycu.edu.tw before 6/7 (30%)**
  - Please name the email title as “Matlab project: Group #”
  - Presentation slides, App files, and usage data set



Model Validation

MATLAB進階程式語言與實作

盧家鋒 Chia-Feng Lu, Ph.D.
Department of Biomedical Imaging and Radiological Sciences, NYCU
alvin4016@nycu.edu.tw

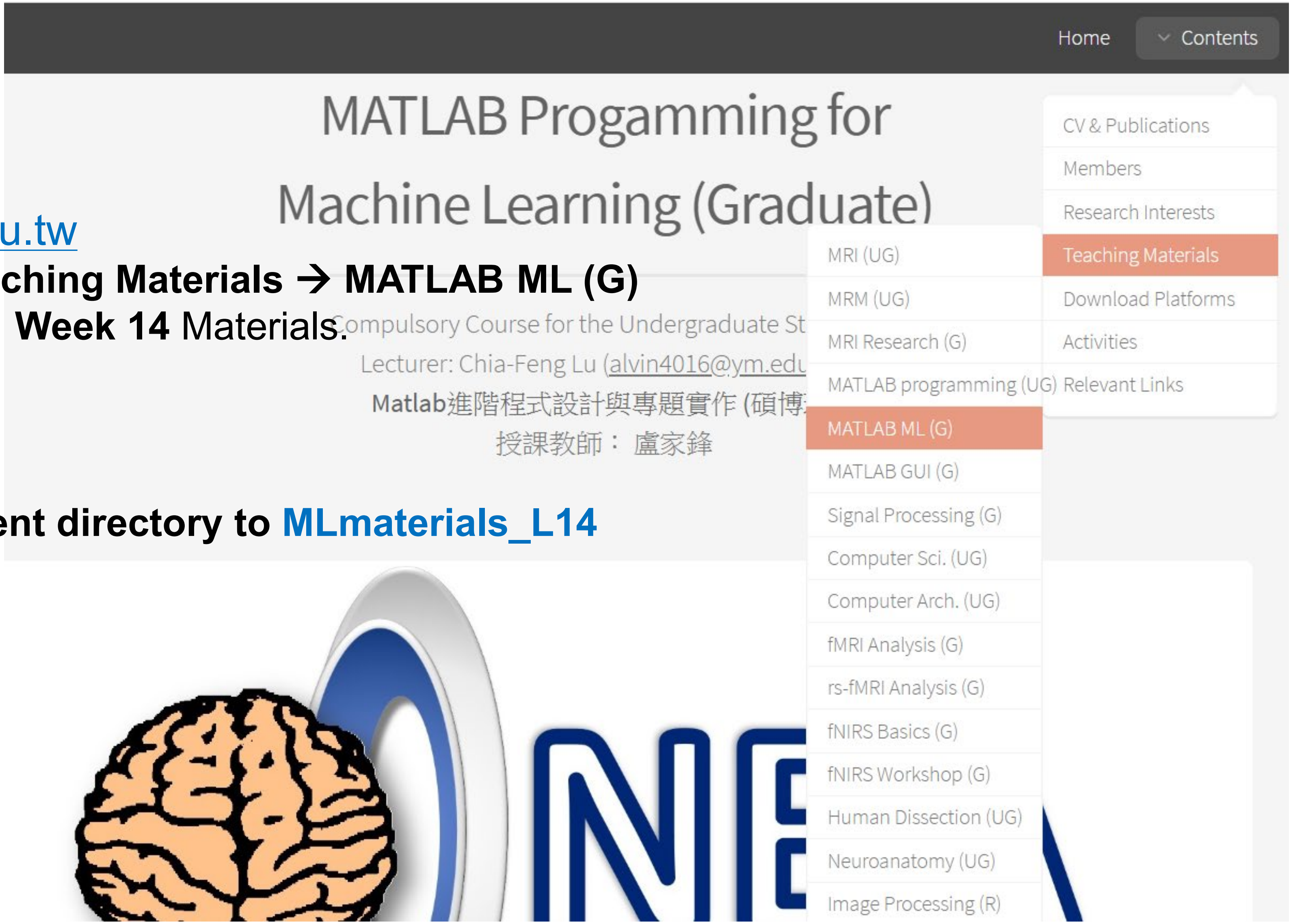
Teaching Materials

cflu.lab.nycu.edu.tw

Contents → Teaching Materials → MATLAB ML (G)

Please download **Week 14** Materials.

Please set current directory to MLmaterials_L14



MATLAB Progamming for Machine Learning (Graduate)

Compulsory Course for the Undergraduate St

Lecturer: Chia-Feng Lu (alvin4016@ym.edu

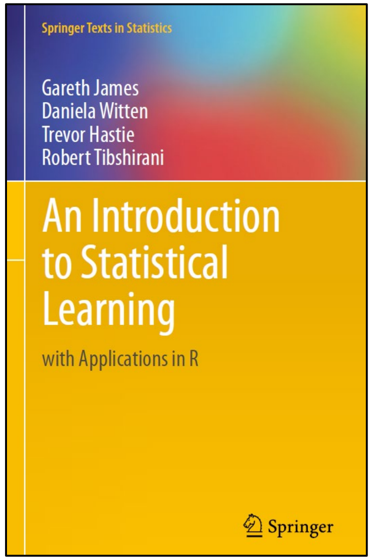
Matlab進階程式設計與專題實作 (碩博

授課教師： 盧家鋒

References

[Textbook 3]

- **An Introduction to Statistical Learning, 2nd edition, 2013**
  Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
- **Online resources:** https://github.com/rghan/ISLR
- **Online resources:** https://github.com/JWarmenhoven/ISLR-python
- **Resampling Methods (Ch.5)**



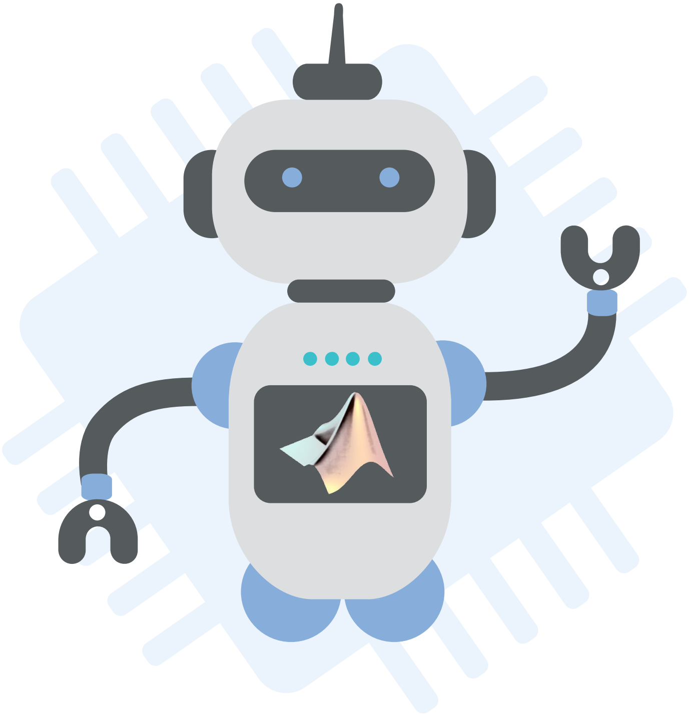
Contents in this Week

01 Resampling Methods – Cross validation

Classification & Regression validation

02 Resampling Methods – Bootstrap

Variance estimation but not for prediction error estimation



Resampling Methods

Cross Validation

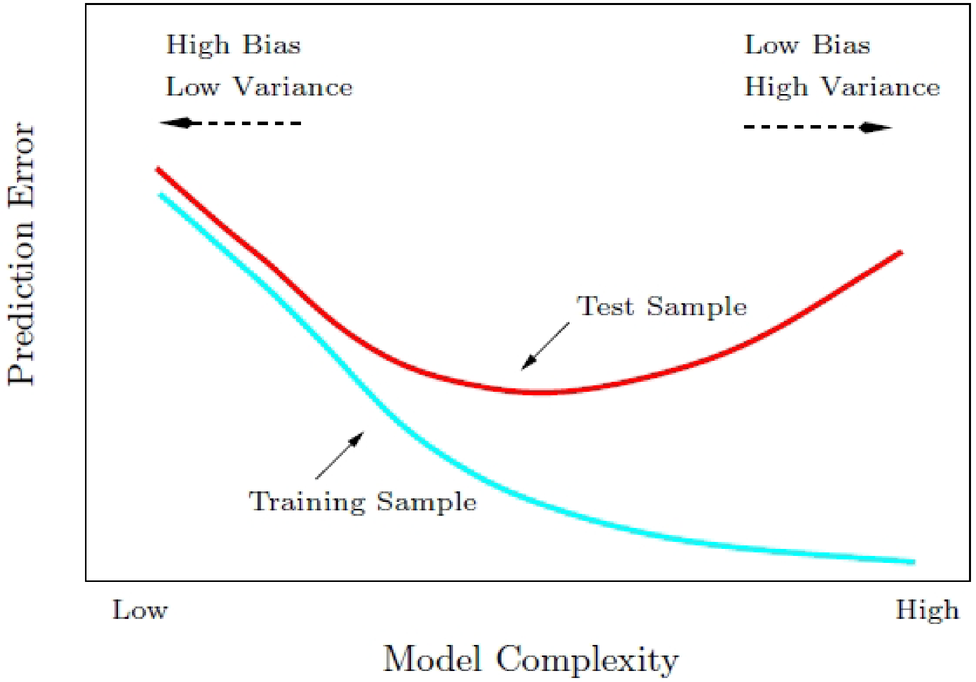
Model Performance

- How can we ensure the machine/statistical learning model is good enough?

沒見過的招式
(unpredictable)

Training Error vs Test error

- **Training error:**
  - Calculated based on the observations used in its training.
- **Test error:**
  - The average error for predicting the response on a new observation, one that was not used in training the method.
- The training error rate often **dramatically underestimates** the test error.



Prediction-Error Estimates

- **Best solution**: a large designated test set. (Often not available…)
- Instead, considering methods that estimate the test error by **holding out** a subset of the training observations from the fitting process.

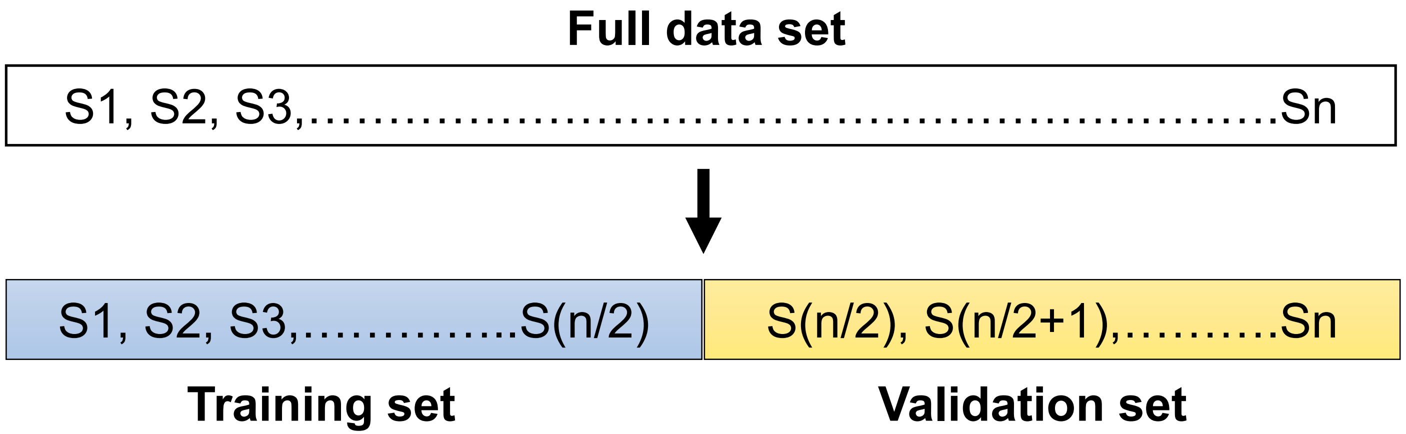
- **Validation-set approach**
- **Leave-one-out cross validation**
- **K-fold cross validation**

Resampling Methods

- **Repeatedly drawing samples** from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.
- Can be computationally expensive, because they involve fitting the same statistical method multiple times using different subsets of the training data.

- Very useful for the **model assessment** and **model selection.**

Validation-Set Approach



Full data set

S1, S2, S3,……………………………………………………….Sn

S1, S2, S3,…………..S(n/2)

S(n/2), S(n/2+1),……….Sn

Training set

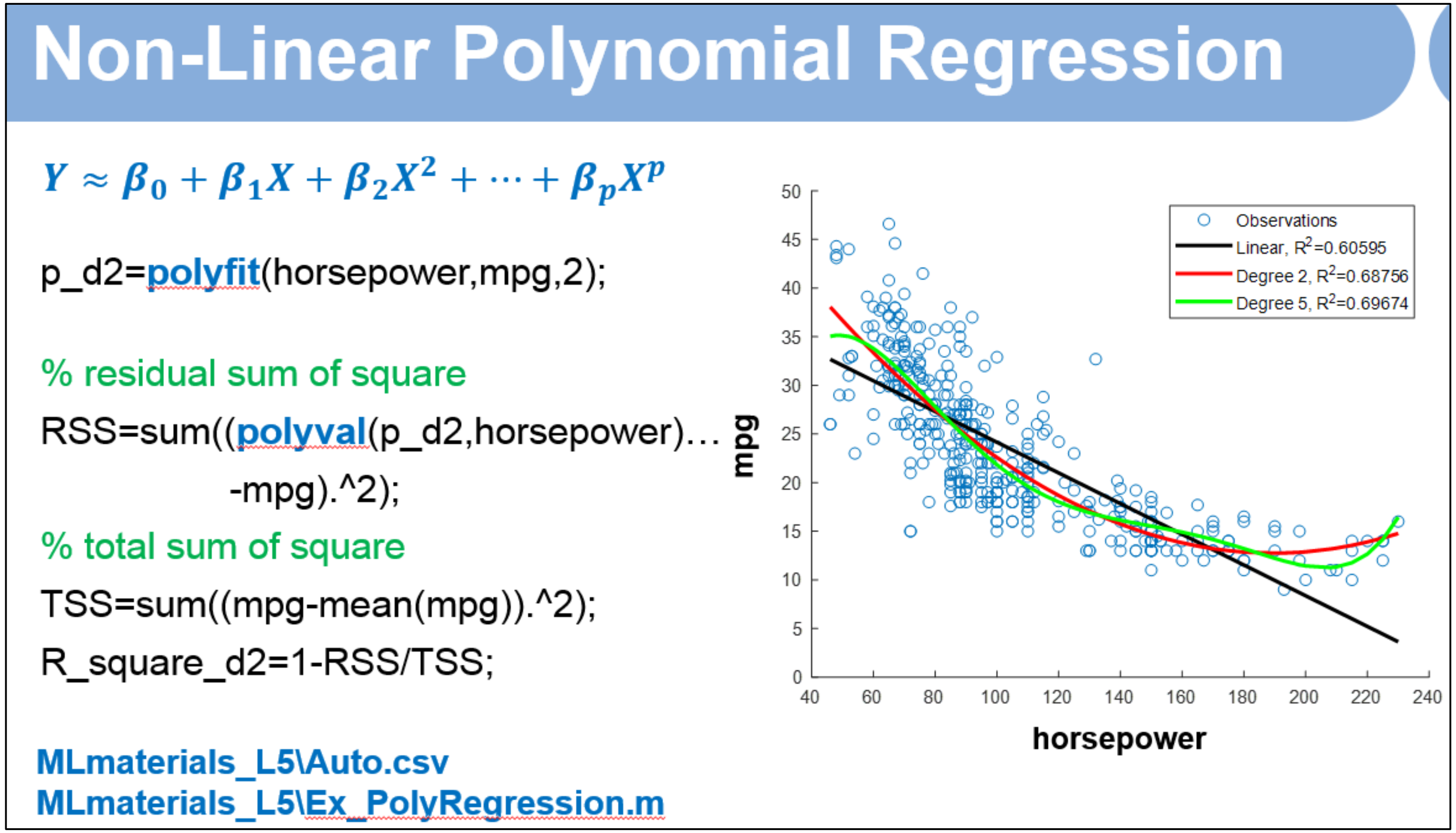
Validation set

- A random splitting into two halves: left part is training set, right part is validation set.
- The machine/statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

Exercise: Automobile data

- Recall the exercise of polynomial regression (in Lesson 5)

- It is natural to wonder whether a cubic or higher-order fit might provide even better results.



Non-Linear Polynomial Regression

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$$

```
p_d2=polyfit(horsepower,mpg,2);

% residual sum of square
RSS=sum((polyval(p_d2,horsepower)...
                -mpg).^2);
% total sum of square
TSS=sum((mpg-mean(mpg)).^2);
R_square_d2=1-RSS/TSS;
```

MLmaterials_L5\Auto.csv
MLmaterials_L5\Ex_PolyRegression.m

Exercise: Automobile data

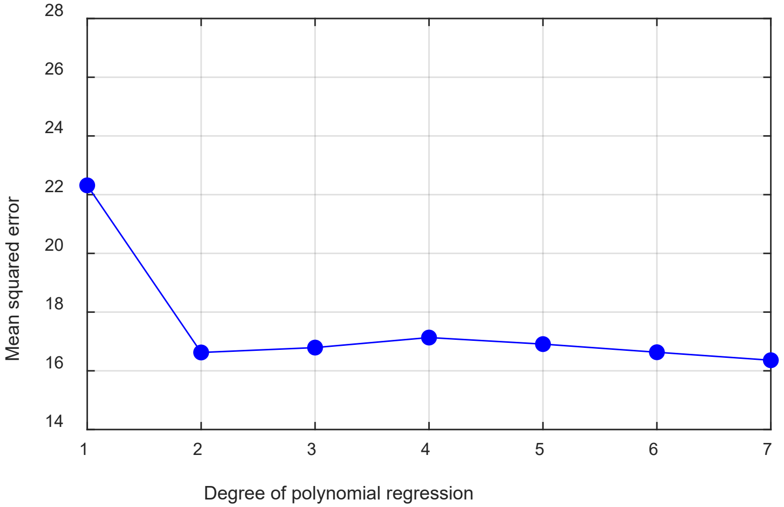
- We answer this question in Chapter 3 by looking at the p-values or R-square associated with a cubic term and higher-order polynomial terms in a linear regression.

- But we could also answer this question using the validation method.
- We randomly split the 392 observations into two sets
  - a training set containing 196 observations;
  - a validation set containing the remaining 196 observations.

Exercise: Automobile data

```
% randomly hold out 50% data for validation
C = cvpartition(length(mpg),'HoldOut',0.5);

for dg=1:7 % the degree of polynomial regression
    % Build the regression model only use the training set
    p{dg}=polyfit(horsepower(C.training),mpg(C.training),dg);
    % Calculate mean squared error on the validation set
    MSE(dg)=mean((polyval(p{dg},horsepower(C.test))-mpg(C.test)).^2);
end

figure, plot(1:7,MSE,'b.-','markersize',16)
xlabel('Degree of polynomial regression')
ylabel('Mean squared error')
grid on
```

>> cvpartition



MLmaterials_L14\Ex_PolyRegression.m

Drawbacks of Validation Set Approach

- The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations – those that are included in the training set rather than in the validation set – are used to fit the model.
- This suggests that the validation set error may tend to overestimate the test error compared to the model fit on the entire data set.

K-fold Cross Validation

- Widely used approach for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.

- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other K – 1 parts(combined), and then obtain predictions for the left-out kth part.

	1	2	3	4	5
When K = 5	**Validation**	**Train**	**Train**	**Train**	**Train**

Five-fold Cross Validation

Dataset includes 99 observations. **Shuffle data** and then…

$$CV_{(K)} = \sum_{i=1}^{K} \frac{n_i}{N} MSE_i$$

	19 observations	20 observations	20 observations	20 observations	20 observations	
Run 1	**Validation**	**Train**	**Train**	**Train**	**Train**	→ MSE_1
Run 2	**Train**	**Validation**	**Train**	**Train**	**Train**	→ MSE_2
Run 3	**Train**	**Train**	**Validation**	**Train**	**Train**	→ MSE_3
Run 4	**Train**	**Train**	**Train**	**Validation**	**Train**	→ MSE_4
Run 5	**Train**	**Train**	**Train**	**Train**	**Validation**	→ MSE_5

$$CV_{(5)} = (19/99)MSE_1 + (20/99)MSE_2 + \ldots. + (20/99)MSE_5$$

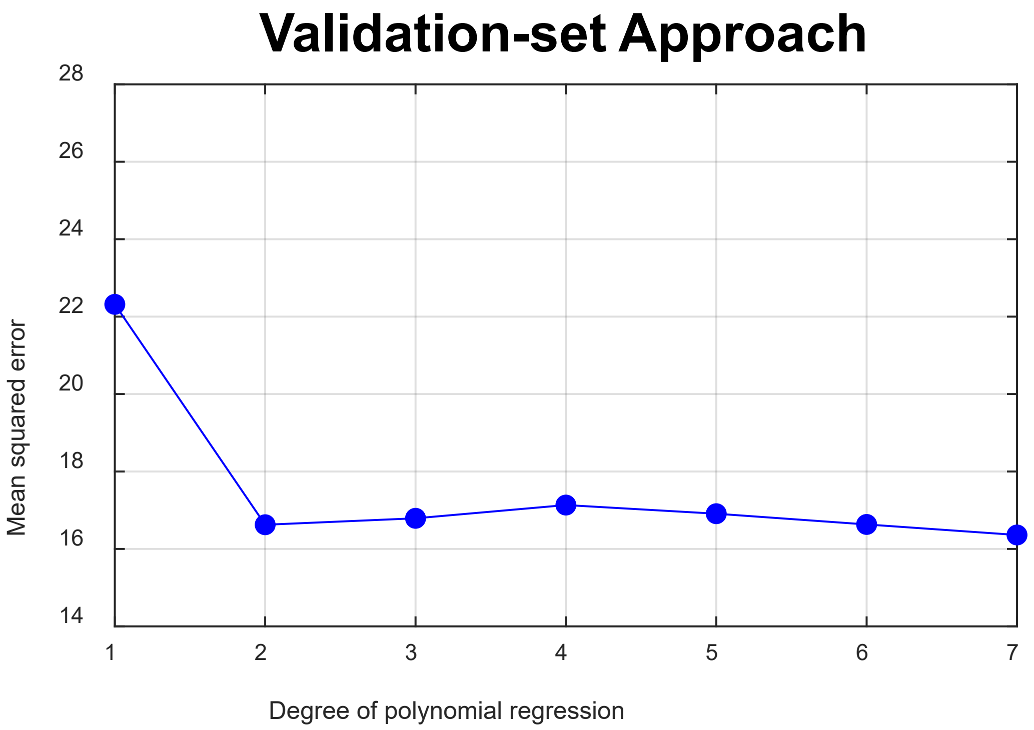
Exercise: Automobile data

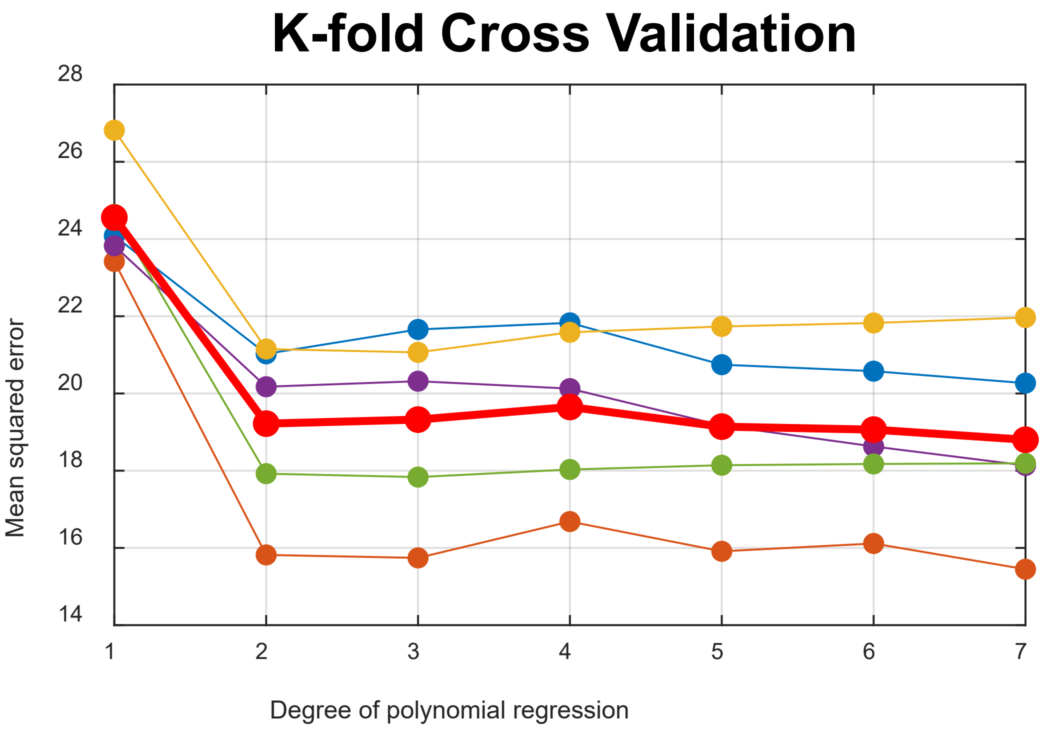
>> cvpartition

```
% randomly partition data to K-fold for cross validation
K=5;
C = cvpartition(length(mpg),'Kfold',K);
figure, hold on
for i=1:K
    for dg=1:7 % the degree of polynomial regression
        % Build the regression model only use the training set
        p{dg,i}=polyfit(horsepower(C.training(i)),mpg(C.training(i)),dg);
        % Calculate mean squared error on the validation set
        MSE(dg,i)=mean((polyval(p{dg,i},horsepower(C.test(i)))-mpg(C.test(i))).^2);
    end
    plot(1:7,MSE(:,i),'.-','markersize',16)
end
plot(1:7,mean(MSE,2),'r.-','markersize',16,'linewidth',2)
```

MLmaterials_L14\Ex_PolyRegression.m

Comparison

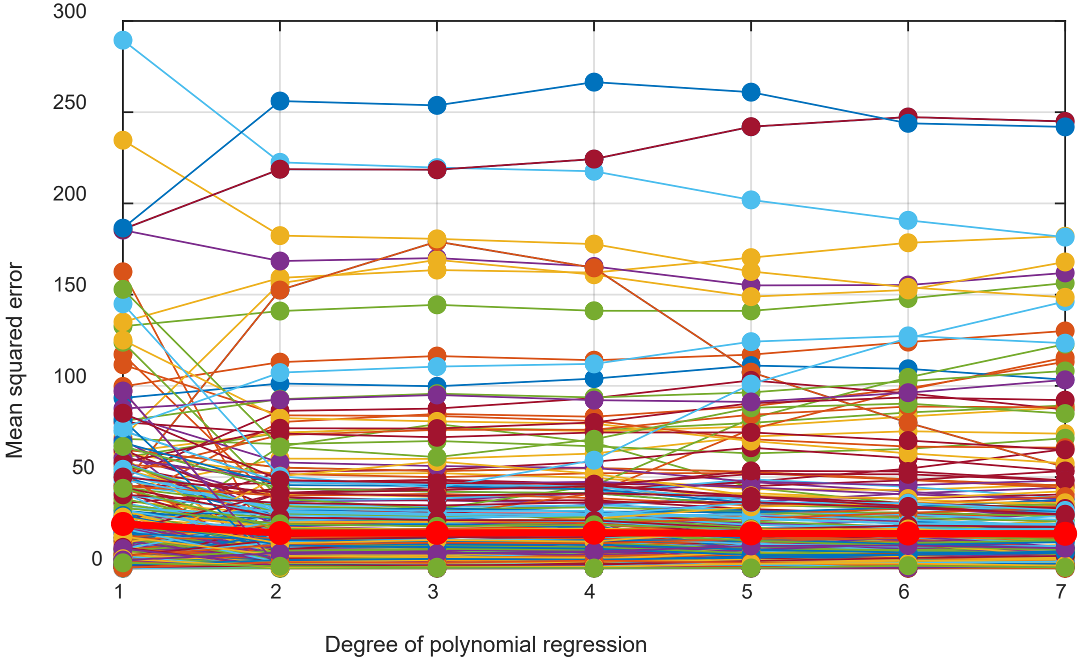

Validation-set Approach
Mean squared error
28
26
24
22
20
18
16
14
1
2
3
4
5
6
7
Degree of polynomial regression


K-fold Cross Validation
Mean squared error
28
26
24
22
20
18
16
14
1
2
3
4
5
6
7
Degree of polynomial regression

More details for K-fold CV

- Setting K = n yields n-fold or leave-one out cross-validation (LOOCV).
- LOOCV sometimes useful, but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.

- For *K*-fold cross-validation,
  - $\boldsymbol{K \uparrow \rightarrow \quad \text{Variance} \uparrow \ Bias \downarrow}$
- A common choice is K = 5 or 10.



Cross Validation on Classification

- Rather than using MSE to quantify test error, we instead use the number of misclassified observations.

$$\mathbf{CV_{(K)}} = \sum_{i=1}^{K} \frac{n_i}{N} Err_i$$

- Where $Err_i = I(y_i \neq \hat{y}_i)$

Cross Validation on Classification

Dataset includes 99 observations. **Shuffle data** and then…

$\mathbf{CV_{(K)}} = \sum_{i=1}^{K} \frac{n_i}{N} Err_i$

	19 observations	20 observations	20 observations	20 observations	20 observations	
Run 1	**Validation**	**Train**	**Train**	**Train**	**Train**	→ 2 **X**, 17 **O**
Run 2	**Train**	**Validation**	**Train**	**Train**	**Train**	→ 3 **X**, 17 **O**
Run 3	**Train**	**Train**	**Validation**	**Train**	**Train**	→ 5 **X**, 15 **O**
Run 4	**Train**	**Train**	**Train**	**Validation**	**Train**	→ 4 **X**, 16 **O**
Run 5	**Train**	**Train**	**Train**	**Train**	**Validation**	→ 1 **X**, 19 **O**

$$\begin{aligned} CV_{(5)} &= (19/99)*(2/19) + (20/99)*(3/20) + \ldots. + (20/99)*(1/20) \\ &= (2+3+5+4+1)/99 = 15/99 \end{aligned}$$

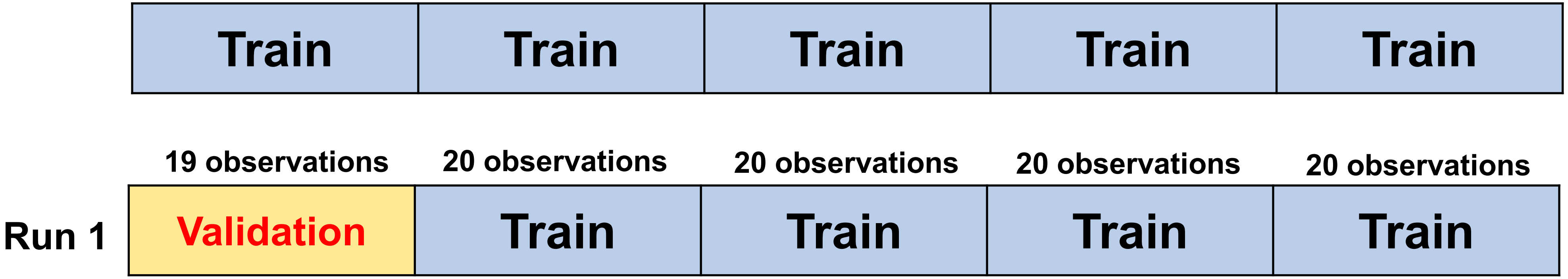
Cross Validation: Right or Wrong?

- Consider a simple classifier applied to some two-class data:
  1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
  2. We then apply a classifier using only these 100 predictors.

How do we estimate the test set performance of this classifier?

Wrong!

- This would ignore the fact that in Step 1, the procedure has already seen the labels of all the data, and made use of them (not just the training dataset).
- This is a form of training and must be conducted by the cross validation process.



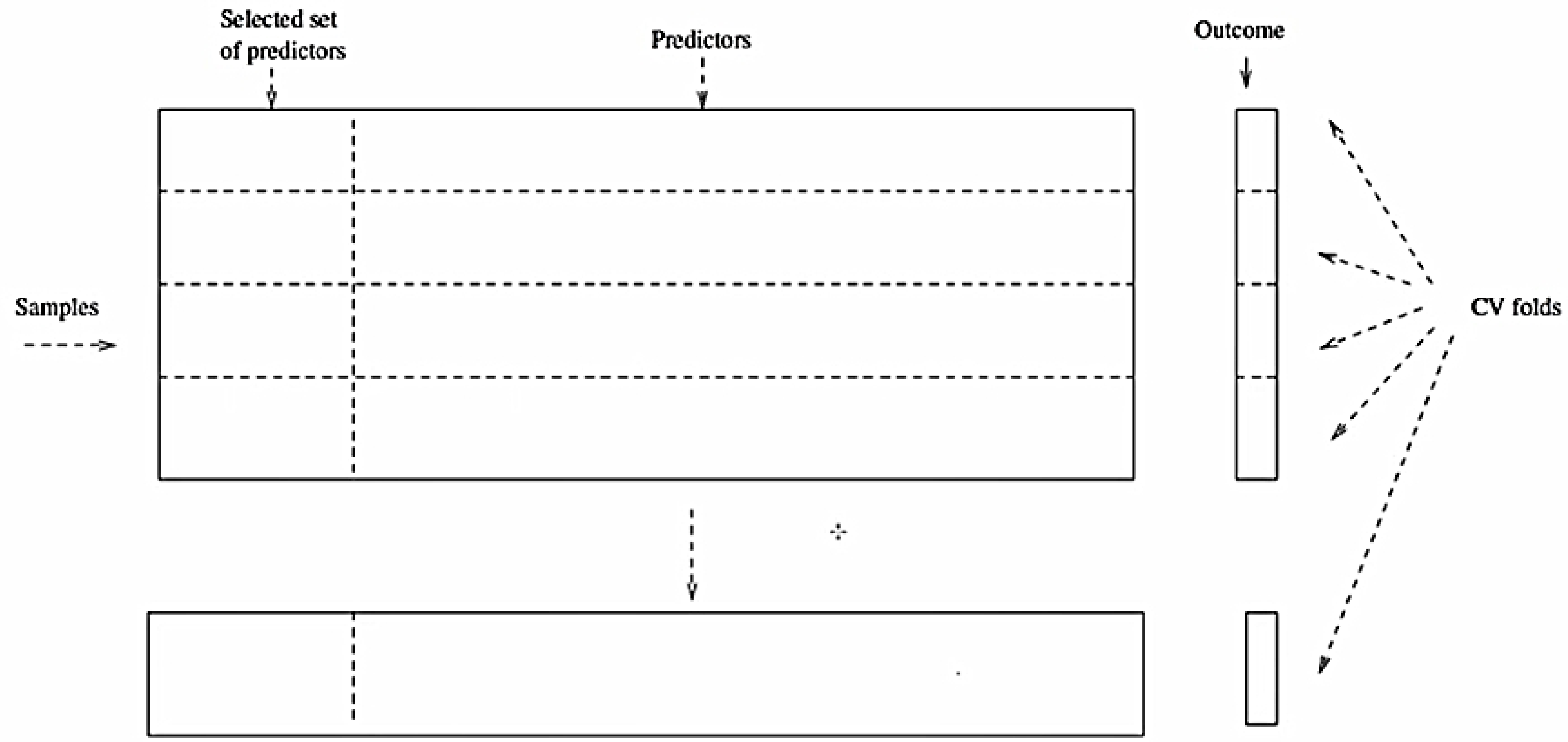
Train	Train	Train	Train	Train

	19 observations	20 observations	20 observations	20 observations	20 observations
Run 1	Validation	Train	Train	Train	Train

- This error was made in many high profile papers.

The Wrong and Right Way

- Wrong: Apply cross-validation only in step 2.
- Right: Apply cross-validation to step 1 and 2.



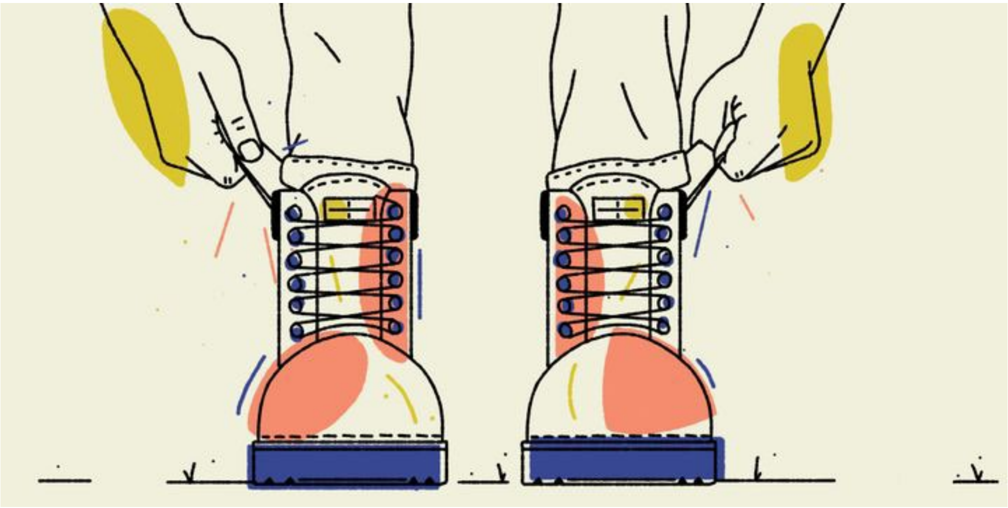
Resampling Methods

Bootstrap

Where does the name came from?

- The use of the term bootstrap derives from the phrase to pull oneself up by one’s bootstraps, widely thought to be based on one of the eighteenth century “The Surprising Adventures of Baron Munchausen” by Rudolph Erich Raspe:

The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.



ISABELLA CARAPELLA/HUFFPOST

The Bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainly associated with a given estimator or statistical learning method.

- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

An Investment Example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively.
- We will invest a fraction α of our money in X, and will invest the remaining 1 - α in Y.
- Choose α to minimize the total risk, or variance, of our investment.

  → minimize Var(αX + (1 - α)Y).
- The value that minimizes the risk is given by

$$\boldsymbol{\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}}$$

Where $\sigma^2{}_X$ = Var(X), $\sigma^2{}_Y$ = Var(Y), and σ_{XY} = Cov(X,Y)

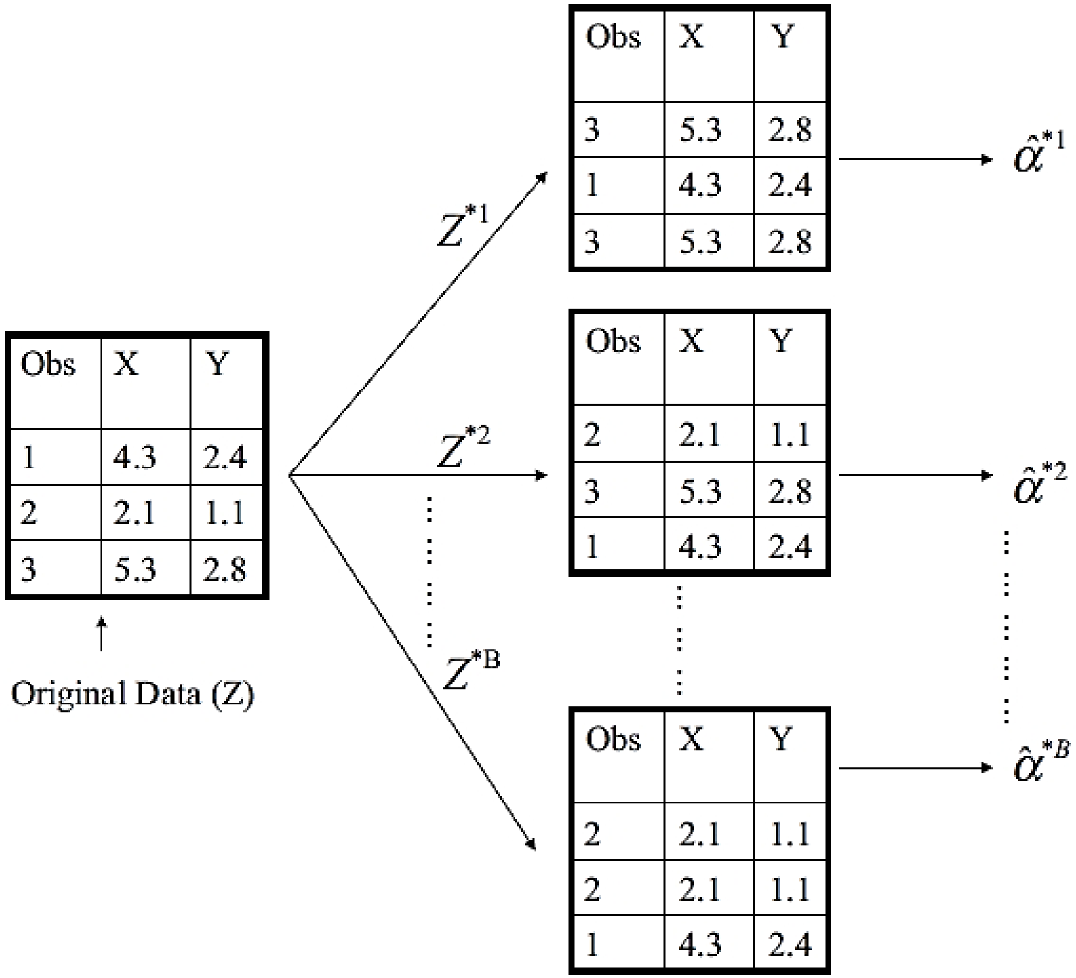
An Investment Example

- But the values of ${\sigma^2}_X$, ${\sigma^2}_Y$, and σ_{XY} are unknown.
- We can compute estimates for these quantities, $\hat{\sigma}_X^2$,$\hat{\sigma}_Y^2$, and $\hat{\sigma}_{XY}$, using a data set that contains measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using

$$\boldsymbol{\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}}$$

In the real world…

- For real data, we cannot always generate new samples from the original population.
- Rather than repeatedly obtaining independent data sets from the population, bootstrap method instead obtains distinct data sets by repeatedly sampling observations from the original data set with replacement.



Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

Exercise: Bootstrap sampling

```
%% load Portfolio data:
[num,txt,raw]=xlsread('Portfolio.csv');
X=num(:,1);  Y=num(:,2);
 %% Calculate the estimated alpha from the measurement
temp=cov(X,Y);
alpha_est = (var(Y)-temp(2))/(var(X)+var(Y)-2*temp(2));
 %% Perform bootstrap sampling
alpha_bs = bootstrp(1000,@alpha_funct,X,Y);

mean(alpha_bs)
std(alpha_bs)
```

>> bootstrp

MLmaterials_L14\Ex_Bootstrap.m

alpha_funct

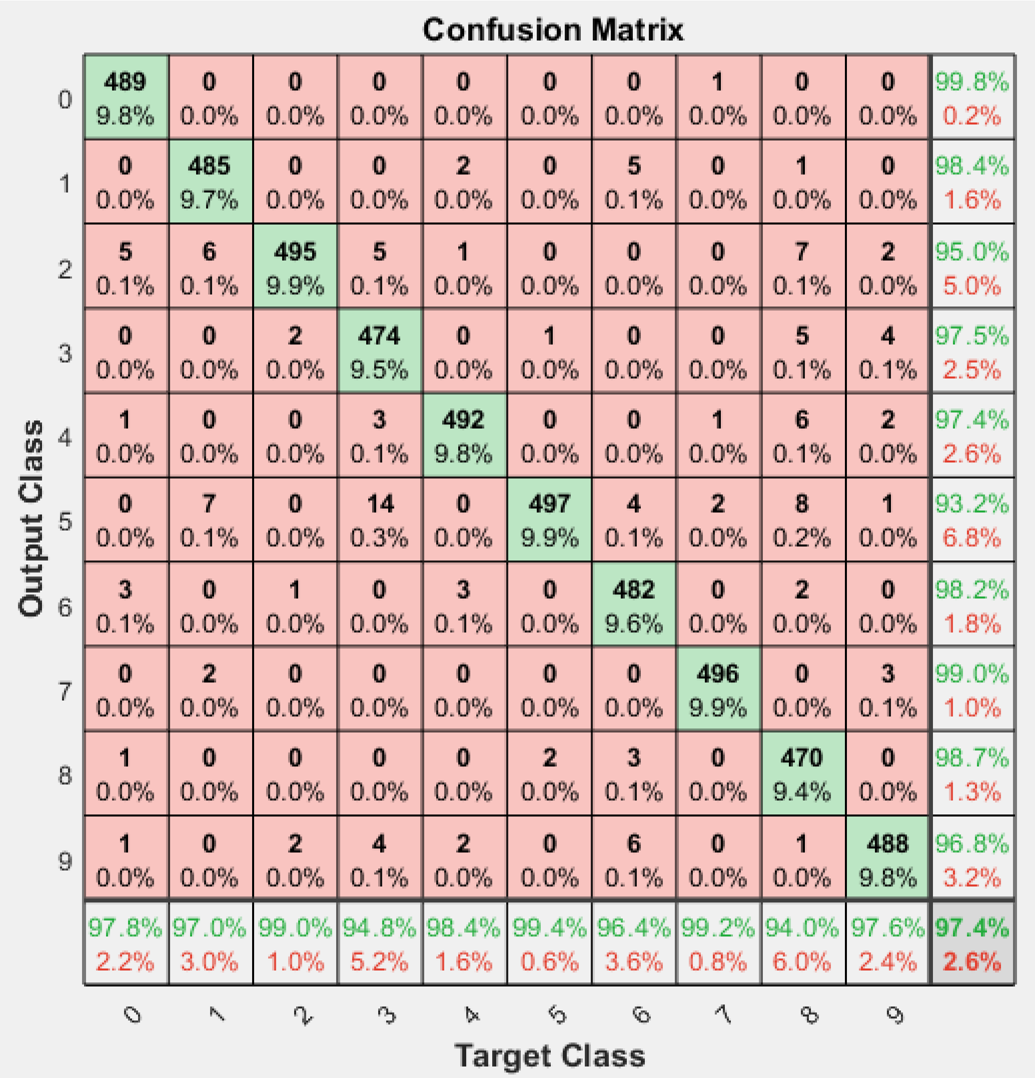
```
function alpha=alpha_funct(X,Y)

temp=cov(X,Y);
alpha=(var(Y)-temp(2))/(var(X)+var(Y)-2*temp(2));
```

Can the bootstrap estimate prediction error?

- In cross-validation, each of the K validation folds is distinct from the other K – 1 folds used for training : there is no overlap. This is crucial for its success.
- If we want to estimate prediction error using the bootstrap, one may think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.
- But each bootstrap sample has significant overlap with the original data. About two-thirds of the original data points appear in each bootstrap sample.
- Because of the above procedure produces overlap between training and test data, **DO NOT use bootstrap to estimate prediction error.**

Confusion Matrix		**True status**			
		Yes	No		
Predicted status	Yes	**True Positive** (TP)	False Positive (FP) Type I error	Positive Predictive Rate, Precision TP/(TP+FP)	False Discovery Rate FP/(TP+FP)
	No	False Negative (FN) Type II error	**True Negative** (TN)	False Omission Rate FN/(FN+TN)	Negative Predictive Rate TN/(FN+TN)
		True positive Rate Sensitivity or Recall TP/(TP+FN)	False positive Rate FP/(FP+TN)	F1 score =2*precision*Recall/(precision+Recall)	
Accuracy **(TP+TN)/T**		False Negative Rate FN/(TP+FN)	True Negative Rate Specificity TN/(FP+TN)		

Confusion Matrix

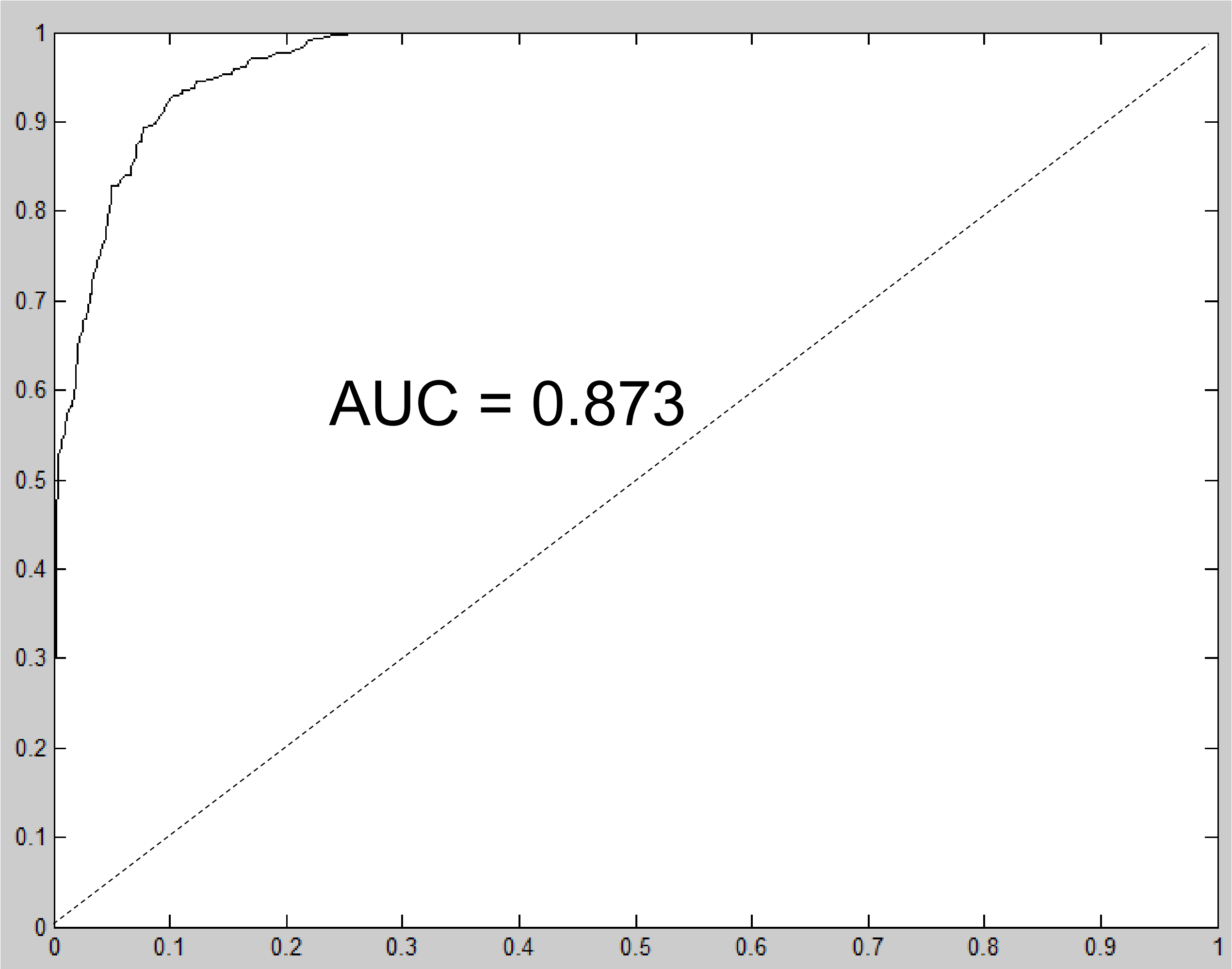


	0	1	2	3	4	5	6	7	8	9	
0	489 9.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	99.8% 0.2%
1	0 0.0%	485 9.7%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	5 0.1%	0 0.0%	1 0.0%	0 0.0%	98.4% 1.6%
2	5 0.1%	6 0.1%	495 9.9%	5 0.1%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	7 0.1%	2 0.0%	95.0% 5.0%
3	0 0.0%	0 0.0%	2 0.0%	474 9.5%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	5 0.1%	4 0.1%	97.5% 2.5%
4	1 0.0%	0 0.0%	0 0.0%	3 0.1%	492 9.8%	0 0.0%	0 0.0%	1 0.0%	6 0.1%	2 0.0%	97.4% 2.6%
5	0 0.0%	7 0.1%	0 0.0%	14 0.3%	0 0.0%	497 9.9%	4 0.1%	2 0.0%	8 0.2%	1 0.0%	93.2% 6.8%
6	3 0.1%	0 0.0%	1 0.0%	0 0.0%	3 0.1%	0 0.0%	482 9.6%	0 0.0%	2 0.0%	0 0.0%	98.2% 1.8%
7	0 0.0%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	496 9.9%	0 0.0%	3 0.1%	99.0% 1.0%
8	1 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	3 0.1%	0 0.0%	470 9.4%	0 0.0%	98.7% 1.3%
9	1 0.0%	0 0.0%	2 0.0%	4 0.1%	2 0.0%	0 0.0%	6 0.1%	0 0.0%	1 0.0%	488 9.8%	96.8% 3.2%
	97.8% 2.2%	97.0% 3.0%	99.0% 1.0%	94.8% 5.2%	98.4% 1.6%	99.4% 0.6%	96.4% 3.6%	99.2% 0.8%	94.0% 6.0%	97.6% 2.4%	**97.4%** **2.6%**

>> confusion

>> plotconfusion

Receiver Operating Characteristics (ROC) Curve



True positive rate (sensitivity)

False positive rate (1-specificity)

>> perfcurve

AUC: Arear under the ROC curve

THE END

Contact:

盧家鋒 alvin4016@nycu.edu.tw