Data Analysis 2021 Spring





**Lecture 08: Resampling Methods**

April 28 & May 3, 2021

**Taesoo Kwon**

[tskwon@seoultech.ac.kr](mailto:tskwon@seoultech.ac.kr)

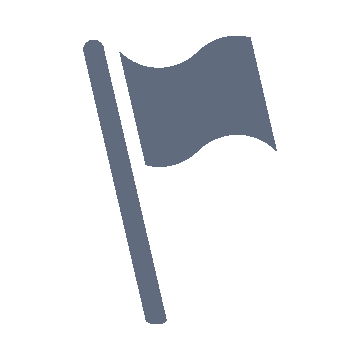
©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

# Course Schedule (Tentative)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Note** | **Date (W)** | **Date (M)** |
| 1 | Orientation, Statistical Learning (Ch2) | Online | 03/03 | 03/08 |
| 2 | Statistical Learning (Ch2), Python Programming | Online | 03/10 | 03/15 |
| 3 | Probability & Statistics | Online | 03/17 | 03/22 |
| 4 | Probability & Statistics | Online | 03/24 | 03/29 |
| 5 | Linear Regression (Ch3) | Online | 03/31 | 04/05 |
| 6 | Linear Regression (Ch3) | Online | 04/07 | 04/12 |
| 7 | Classification (Ch4) | Online | 04/14 | 04/19 |
| 8 | **Midterm exam** | **Class hours (W1-W7)** | **04/21** | **04/26** |
| **9** | Resampling Methods (Ch5) | Online | 04/28 | 05/03 |
| 10 | Linear Model Selection and Regularization (Ch6) | Online | 05/05 | 05/10 |
| 11 | Moving Beyond Linearity (Ch7) | Online | 05/12 | 05/17 |
| 12 | Tree-Based Methods (Ch8) | Online | 05/19 | 05/24 |
| 13 | Support Vector Machines (Ch9) | Online | 05/26 | 05/31 |
| 14 | Unsupervised Learning (Ch10) | Online | 06/02 | 06/07 |
| 15 | **Final exam** | **7pm or Class hours (W9-W14)** | **06/09or14** | **06/09or14** |

©copyright CIEL 2015 **2**





* Resampling methods

**OUTLINES**

* + Cross-validation
  + Bootstrap
* Python lab
* Summary & Next class

©copyright CIEL 2015



**Resampling Methods: Ch 5**



## Resampling methods

* Python lab
* Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

# Cross-Validation and Bootstrap

* In the section we discuss two resampling methods: cross-validation and bootstrap.
* These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
* For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

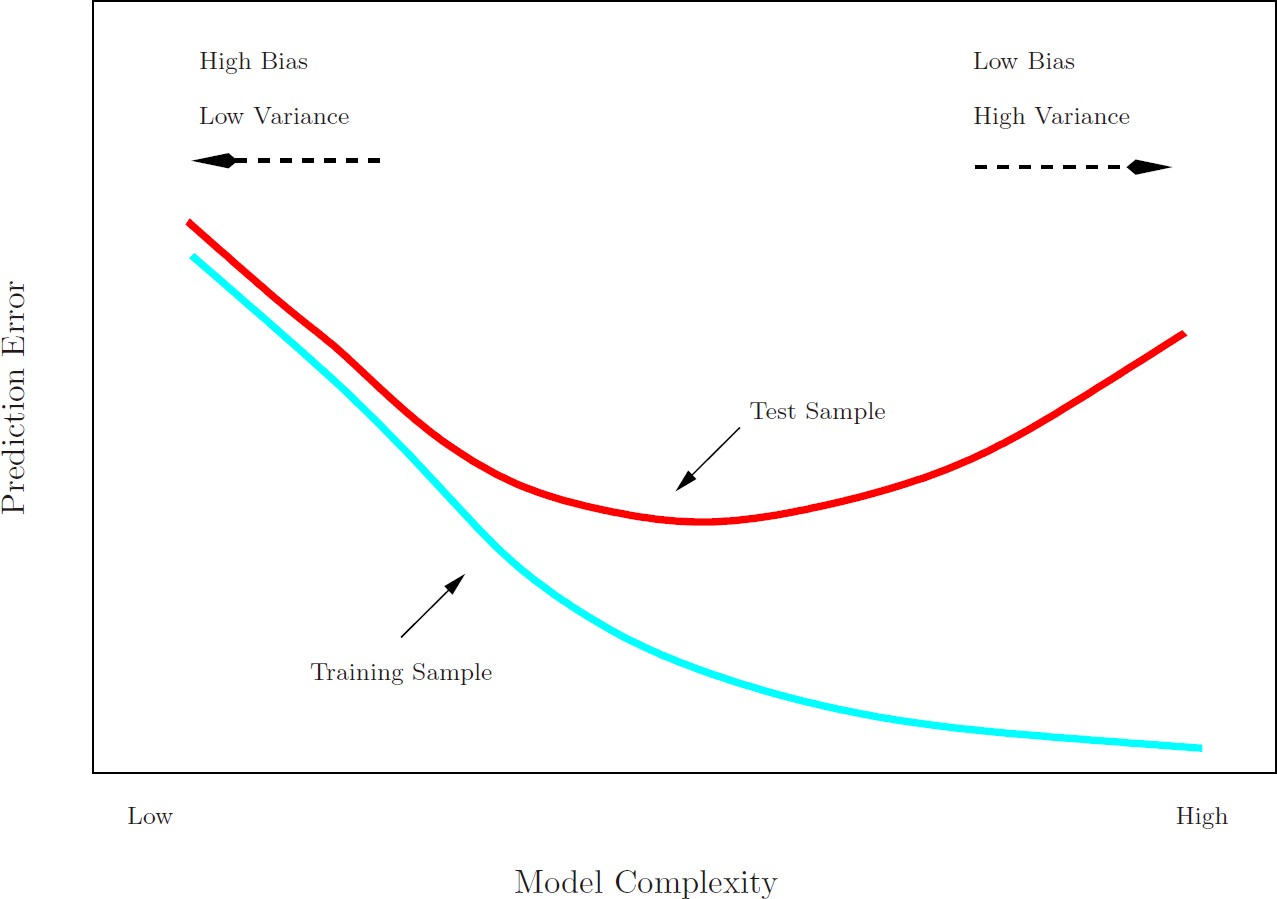
©copyright CIEL 2015 **5**

# Training Error versus Test Error

* Recall the distinction between the test error and the training error:
* The test error is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
* In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training.
* But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

©copyright CIEL 2015 **6**

# Training- versus Test-Set Performance



©copyright CIEL 2015 **7**

# More on Prediction-Error Estimates

* Best solution: a large designated test set
  + Often not available
* Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate.
  + These include the 𝐶𝐶𝑃𝑃 statistic, AIC and BIC. They are discussed elsewhere in this course
* Here we instead consider a class of methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations

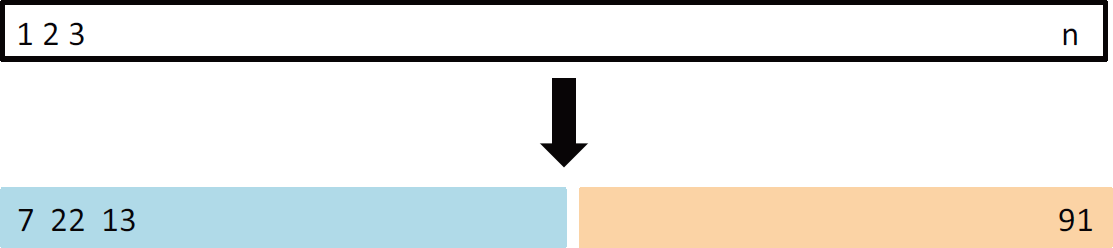
©copyright CIEL 2015 **8**

# Validation-Set Approach

* Here we randomly divide the available set of samples into two parts: a training set and a validation or hold-out set
* The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
* The resulting validation-set error provides an estimate of the test error.
  + This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

©copyright CIEL 2015 **9**

# Validation Process



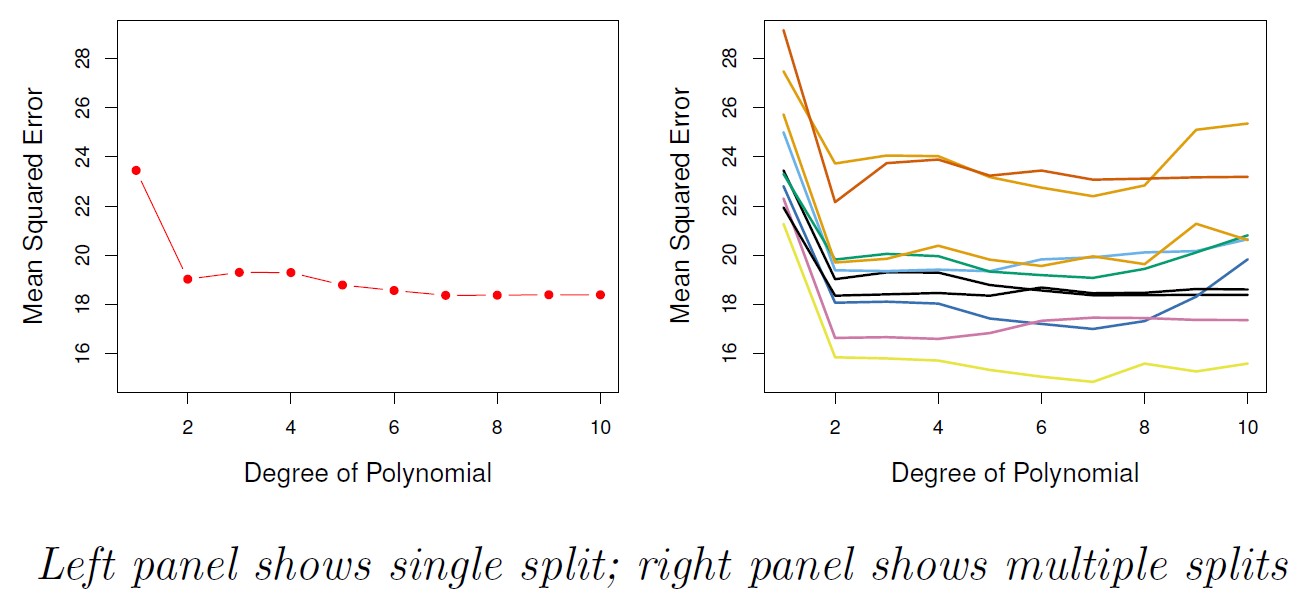
* A random splitting into two halves: left part is training set, right part is validation set

©copyright CIEL 2015 **10**

# Example: Automobile Data

* Want to compare linear vs higher-order polynomial terms in a linear regression
* We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations

**11**



©copyright CIEL 2015

# 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 for the model fit on the entire data set. Why?

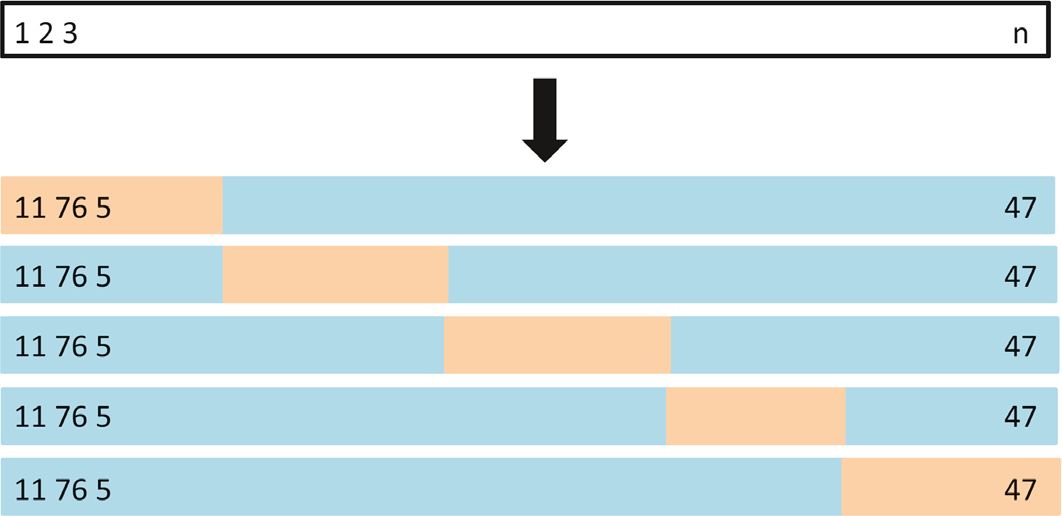
©copyright CIEL 2015 **12**

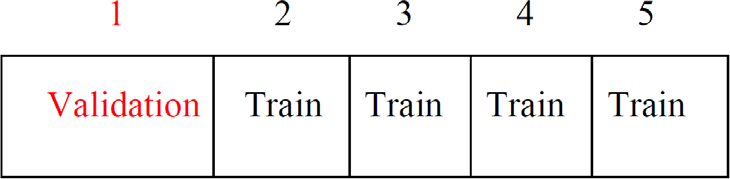
# 𝒌𝒌-fold Cross-validation

* 𝑘𝑘-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 𝑘𝑘 equal-sized parts
  + We leave out part 𝑖𝑖, fit the model to the other 𝑘𝑘 − 1 parts (combined), and then obtain predictions for the left-out 𝑖𝑖th part
* This is done in turn for each part 𝑖𝑖 = 1,2, ⋯ , 𝑘𝑘, and then the results are combined

©copyright CIEL 2015 **13**

# 𝒌𝒌-fold Cross-Validation in Detail

* Divide data into 𝑘𝑘 roughly equal-sized parts (𝑘𝑘 = 5 here)

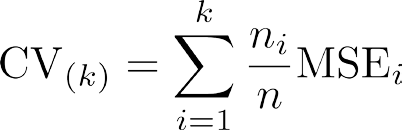


* + Textbook

©copyright CIEL 2015 **14**

# The Details

* Let the 𝑘𝑘 parts be 𝐶𝐶1, 𝐶𝐶2, ⋯ 𝐶𝐶𝐾𝐾, where 𝐶𝐶𝑘𝑘 denotes the indices of the observations in part 𝑘𝑘



where MSE𝑖𝑖 = ∑𝑗𝑗∈𝐶𝐶

removed

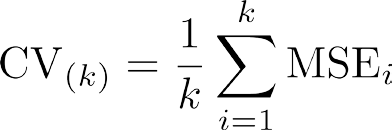
𝑦𝑦𝑗𝑗 − 𝑦𝑦�𝑗𝑗

𝑖𝑖

/𝑛𝑛𝑖𝑖 , and

𝑦𝑦�𝑖𝑖 is the fit for observation 𝑖𝑖, obtained from the data with part 𝑘𝑘

* + If 𝑛𝑛𝑖𝑖 is the same for all 𝑖𝑖,

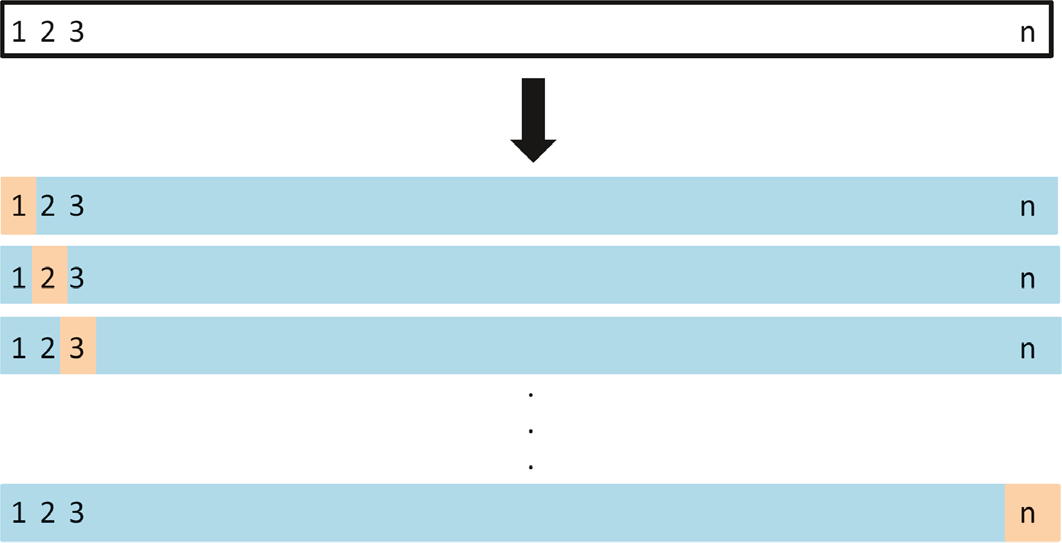


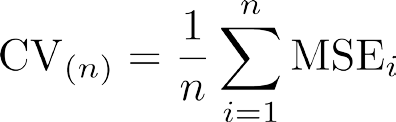
* Setting 𝑘𝑘 = 𝑛𝑛 yields 𝑛𝑛-fold or leave-one out cross-validation (LOOCV)

©copyright CIEL 2015 **15**

# LOOCV

* Far less bias from 𝑛𝑛 − 1 observations: almost as many as in the entire data set
* Performing LOOCV multiple times always yields the same result
* Drawback: potential to be expensive to implement

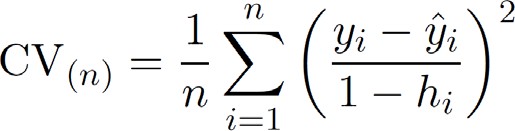




©copyright CIEL 2015 **16**

# A Nice Special Case for LOOCV

* With least-squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit!
  + The following formula holds



* + 𝑦𝑦�𝑖𝑖 is the 𝑖𝑖th fitted value from the original least squares fit, and ℎ𝑖𝑖 is the leverage
  + This is like the ordinary MSE, except the 𝑖𝑖th residual is divided by 1 − ℎ𝑖𝑖
* 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
* A better choice is 𝑘𝑘 = 5 or 10

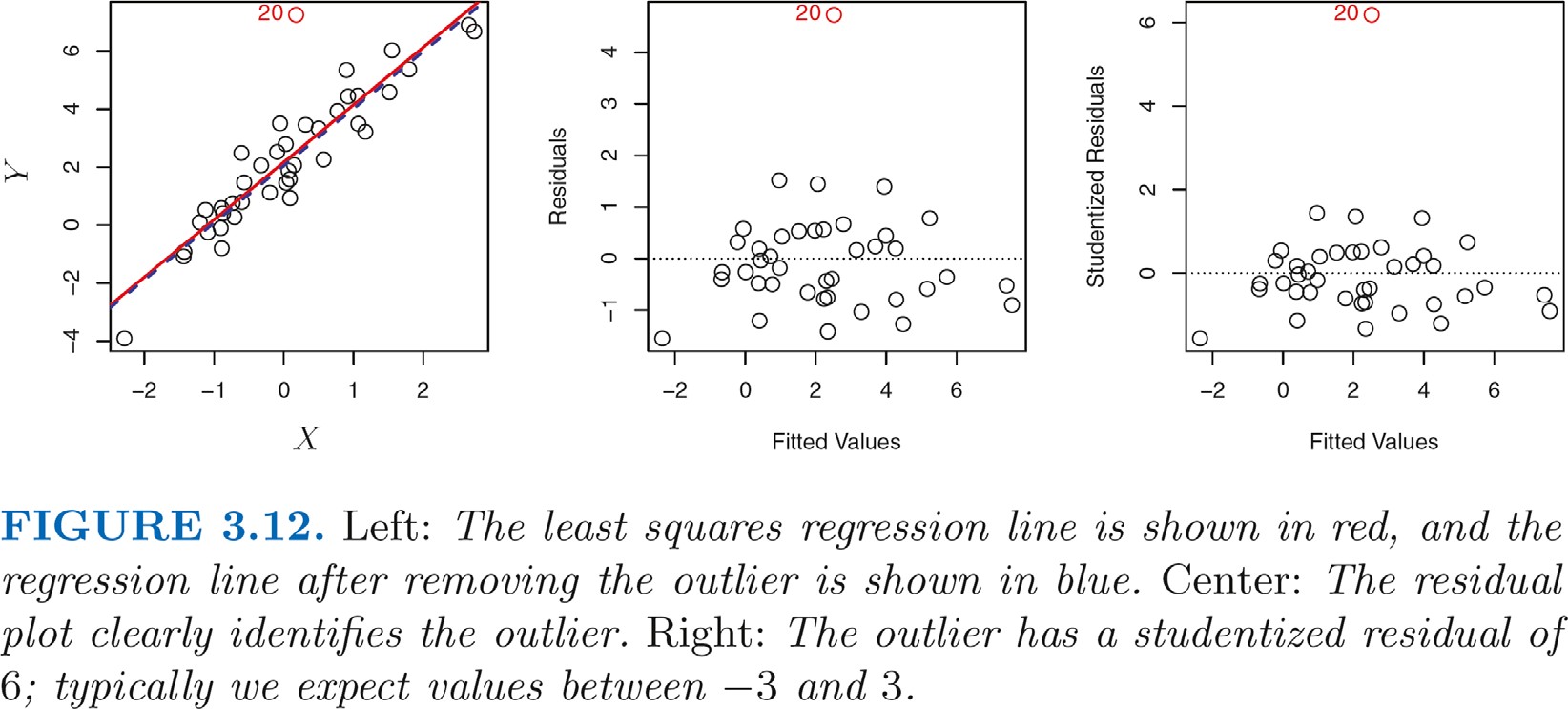
©copyright CIEL 2015 **17**

# Outliers: Back to Ch 3.3-4

* Outlier
  + A point for which 𝑦𝑦𝑖𝑖 is far from the value predicted by the model
* Even if an outlier does not have much effect on the least squares fit
  + E.g., RSE 1.09 w/ outlier while being 0.77 w/o outlier
  + E.g., 𝑅𝑅2 0.892 w/ outlier while being 0.805 w/o outlier
* Residual plots
  + How large a residual?
* Studentized residuals
  + Computed by dividing each residual by its estimated standard error
  + Greater than 3: possible outlier

©copyright CIEL 2015 **18**

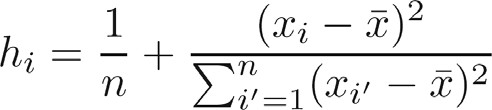
# Outliers: Back to Ch 3.3-4 [cont.]



©copyright CIEL 2015 **19**

# High Leverage Points: : Back to Ch 3.3-5

* Observations with high leverage have an unusual value for 𝑥𝑥𝑖𝑖
  + C.f. outliers are observations for which response 𝑦𝑦𝑖𝑖 is unusual given predictor 𝑥𝑥𝑖𝑖
  + Much more substantial impact on the least squares line
* Identifying high leverage observations
  + In simple linear regression: fairly easy
  + In multiple linear regression: difficult to plotting all dimensions of data simultaneously
* Leverage statistic: large  high leverage



* + In simple linear regression

Standardized distance

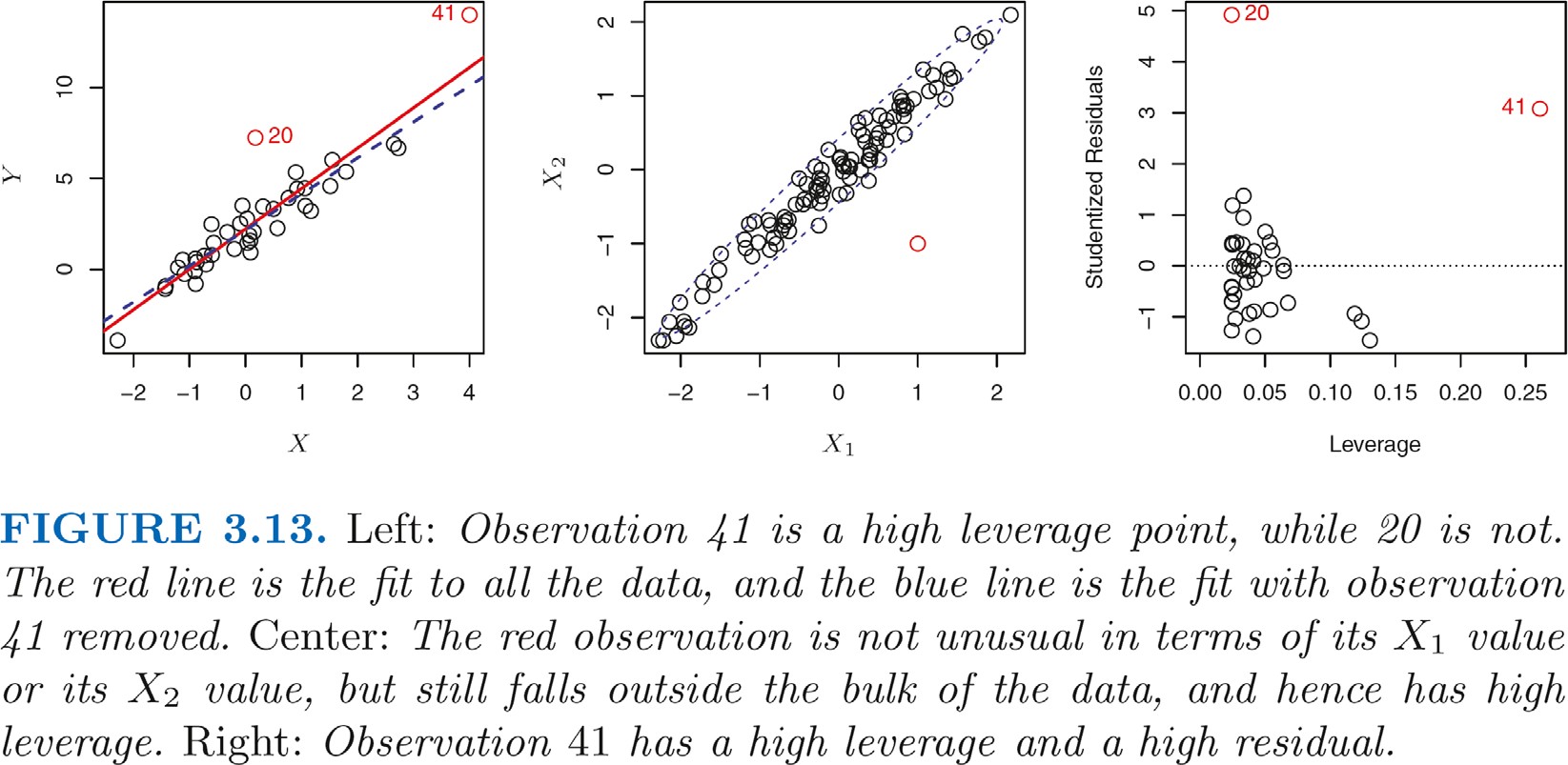


* + Multiple linear regression: similarly standardized distance, i.e., relative distance of each from center

o FYI: from hat matrix

©copyright CIEL 2015 **20**

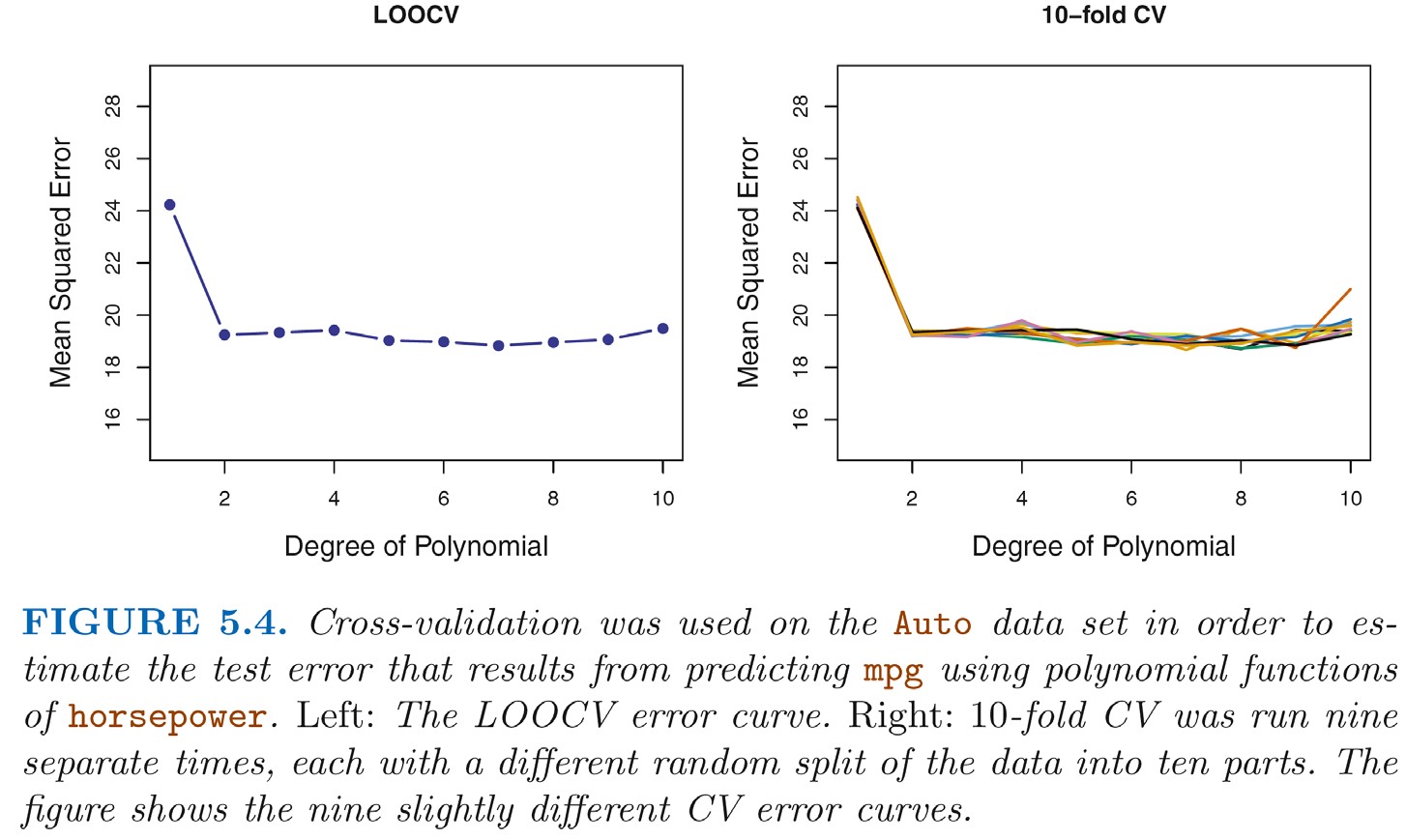
# High Leverage Points: Back to Ch 3.3-5 [cont.]



©copyright CIEL 2015 **21**

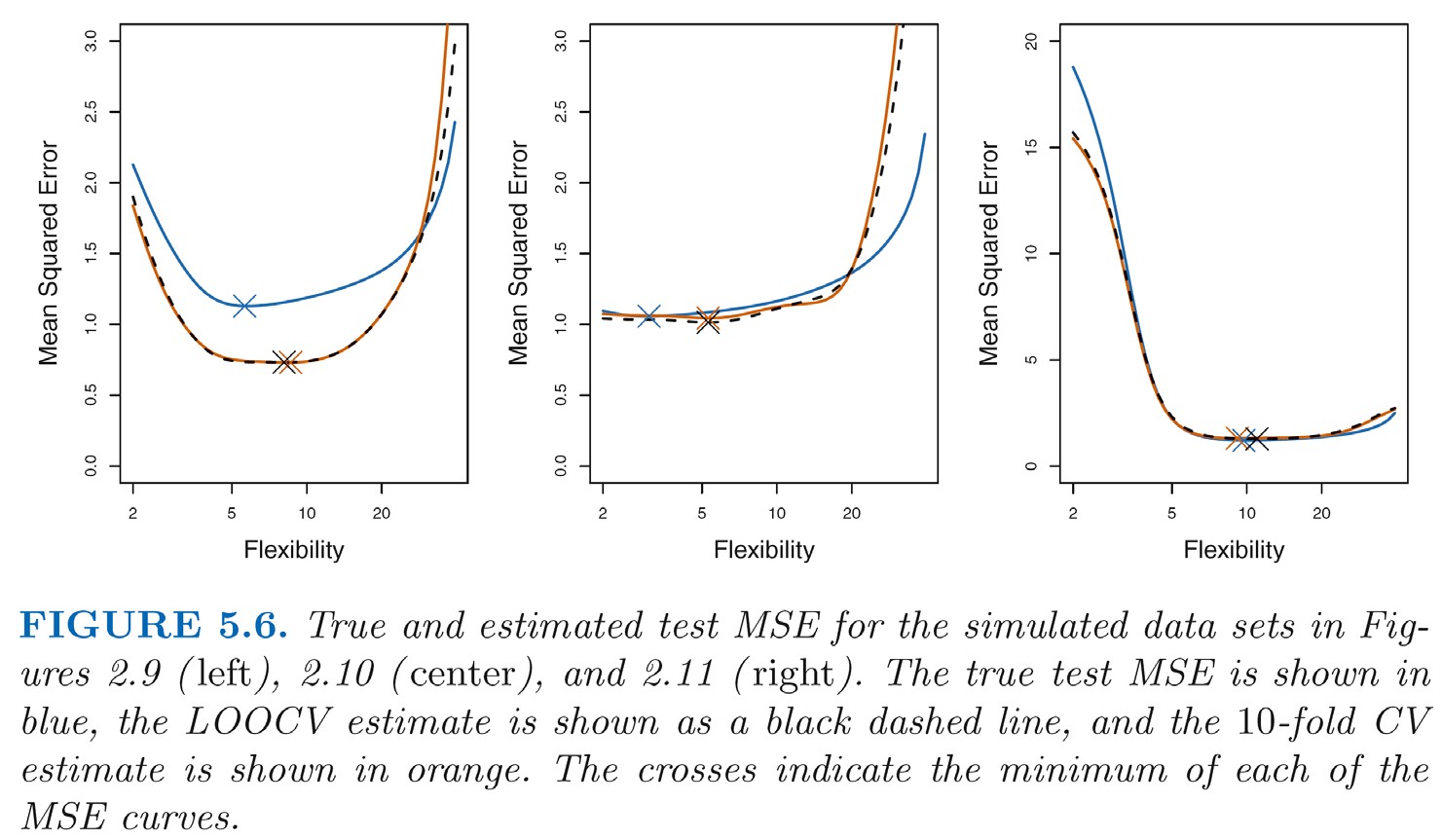
# Auto Data Revisited

©copyright CIEL 2015 **22**



# True and Estimated Test MSE for Simulated data

©copyright CIEL 2015 **23**

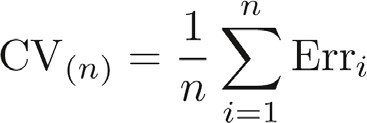


# Other Issues with Cross-Validation

* Since each training set is only (𝑘𝑘 − 1)/𝑘𝑘 as big as the original training set, the estimates of prediction error will typically be biased upward
* This bias is minimized when 𝑘𝑘 = 𝑛𝑛 (LOOCV), but this estimate has high variance, as noted earlier
* 𝑘𝑘 = 5 or 10 provides a good compromise for this bias-variance tradeoff

©copyright CIEL 2015 **24**

# Cross-Validation for Classification Problems

* LOOCV error rate

 

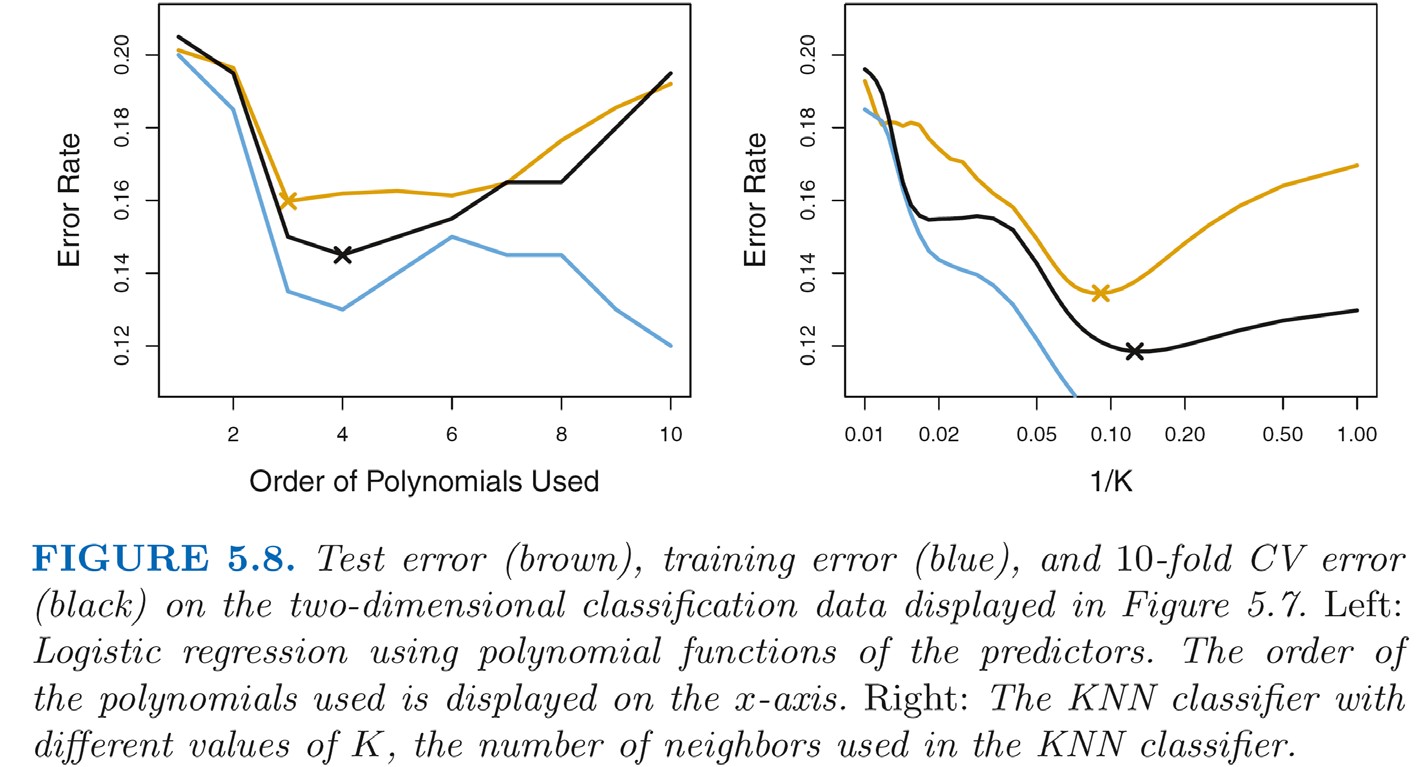
* 𝑘𝑘-fold CV error rate and validation set error rate are defined analogously

©copyright CIEL 2015 **25**

# Cross-Validation for Classification Problems [cont.]

* Minimum very close to the best order of polynomials or 𝐾𝐾

©copyright CIEL 2015 **26**



# Bootstrap

* The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty 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.

©copyright CIEL 2015 **27**

# A Simple Example

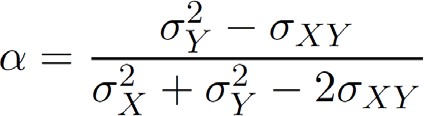
* Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of

𝑋𝑋 and 𝑌𝑌, respectively, where 𝑋𝑋 and 𝑌𝑌 are random quantities

* We will invest a fraction 𝛼𝛼 of our money in 𝑋𝑋, and will invest the remaining 1 − 𝛼𝛼 in 𝑌𝑌
* We wish to choose 𝛼𝛼 to minimize the total risk, or variance, of our investment. In other words, we want to minimize Var

𝛼𝛼𝑋𝑋 + 1 − 𝛼𝛼 𝑌𝑌

* One can show that the value that minimizes the risk is given by



* + 𝜎𝜎2 = Var

𝑋𝑋

𝑋𝑋

𝑌𝑌

, 𝜎𝜎2 = Var

, and 𝜎𝜎𝑋𝑋𝑌𝑌 = Cov

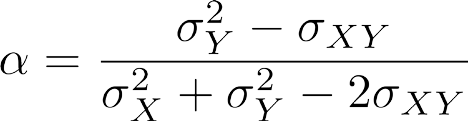


©copyright CIEL 2015 **28**

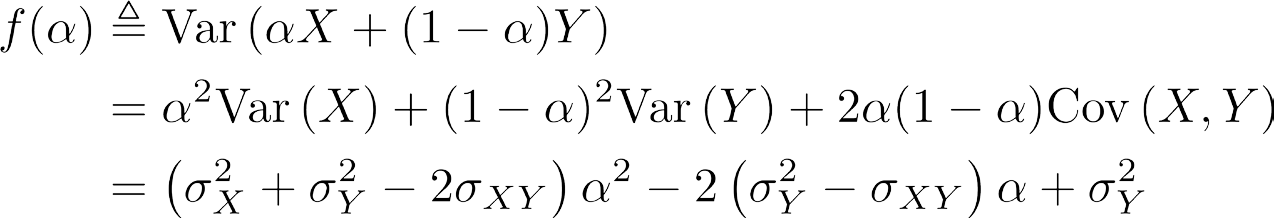
𝑌𝑌

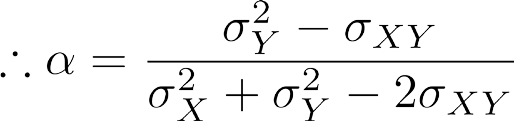
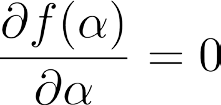
𝑋𝑋, 𝑌𝑌

**[FYI] Optimal** 𝜶𝜶



* Proof

Convex w.r.t. 𝛼𝛼



©copyright CIEL 2015 **29**

# A Simple Example [cont.]

* But the values of 𝜎𝜎2, 𝜎𝜎2, and 𝜎𝜎𝑋𝑋𝑌𝑌 are unknown

𝑋𝑋 𝑌𝑌

* We can compute estimates for these quantities, measurements for 𝑋𝑋 and 𝑌𝑌

𝑋𝑋

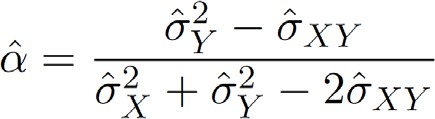
𝑌𝑌

𝜎𝜎�2,

𝜎𝜎�2, and

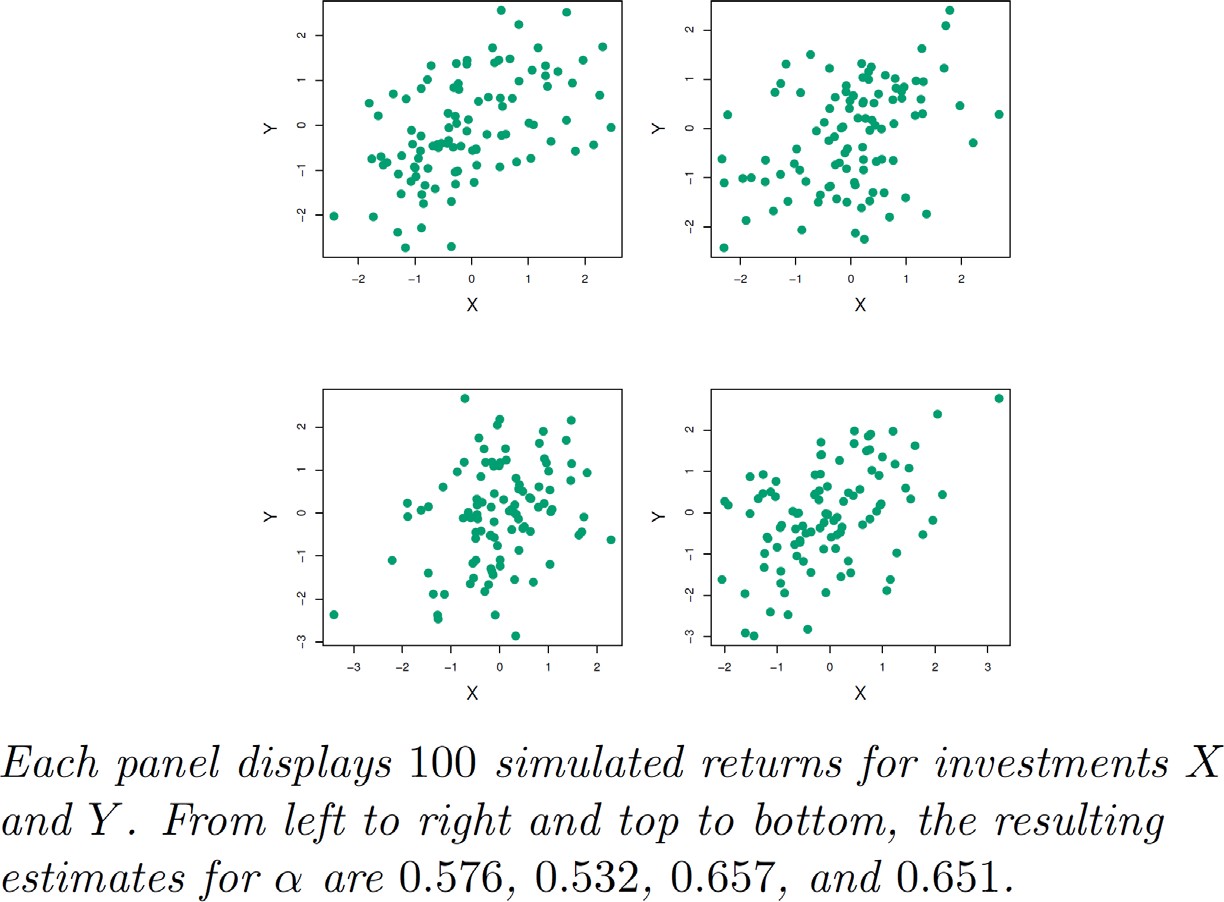
𝜎𝜎�𝑋𝑋𝑌𝑌, using a data set that contains

* We can then estimate the value of 𝛼𝛼 that minimizes the variance of our investment using



©copyright CIEL 2015 **30**

# A Simple Example [cont.]



©copyright CIEL 2015 **31**

# A Simple Example [cont.]

* To estimate the standard deviation of 𝛼𝛼�, we repeated the process of simulating 100 paired

observations of 𝑋𝑋 and 𝑌𝑌, and estimating 𝛼𝛼 1,000 times

* We thereby obtained 1,000 estimates for 𝛼𝛼, which we can call

𝛼𝛼�1,

𝛼𝛼�2, ⋯,

𝛼𝛼�1000

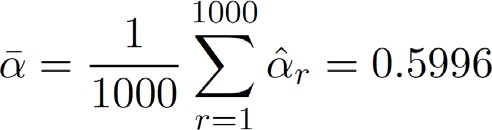
* For these simulations the parameters were set to 𝜎𝜎2 = 1, 𝜎𝜎2 = 1.25, and 𝜎𝜎𝑋𝑋𝑌𝑌 = 0.5, and so we

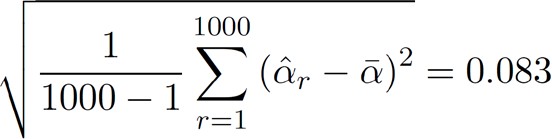
𝑋𝑋 𝑌𝑌

know that the true value of 𝛼𝛼 is 0.6 (indicated by the red line)

©copyright CIEL 2015 **32**

# A Simple Example [cont.]

* The mean over all 1,000 estimates for 𝛼𝛼 is
  + Very close to 𝛼𝛼 = 0.6, and the standard deviation of the estimates is



* This gives us a very good idea of the accuracy of

𝛼𝛼�: SE

≈ 0.083

* So roughly speaking, for a random sample from the population, we would expect from 𝛼𝛼 by approximately 0.08, on average

𝛼𝛼�

𝛼𝛼�

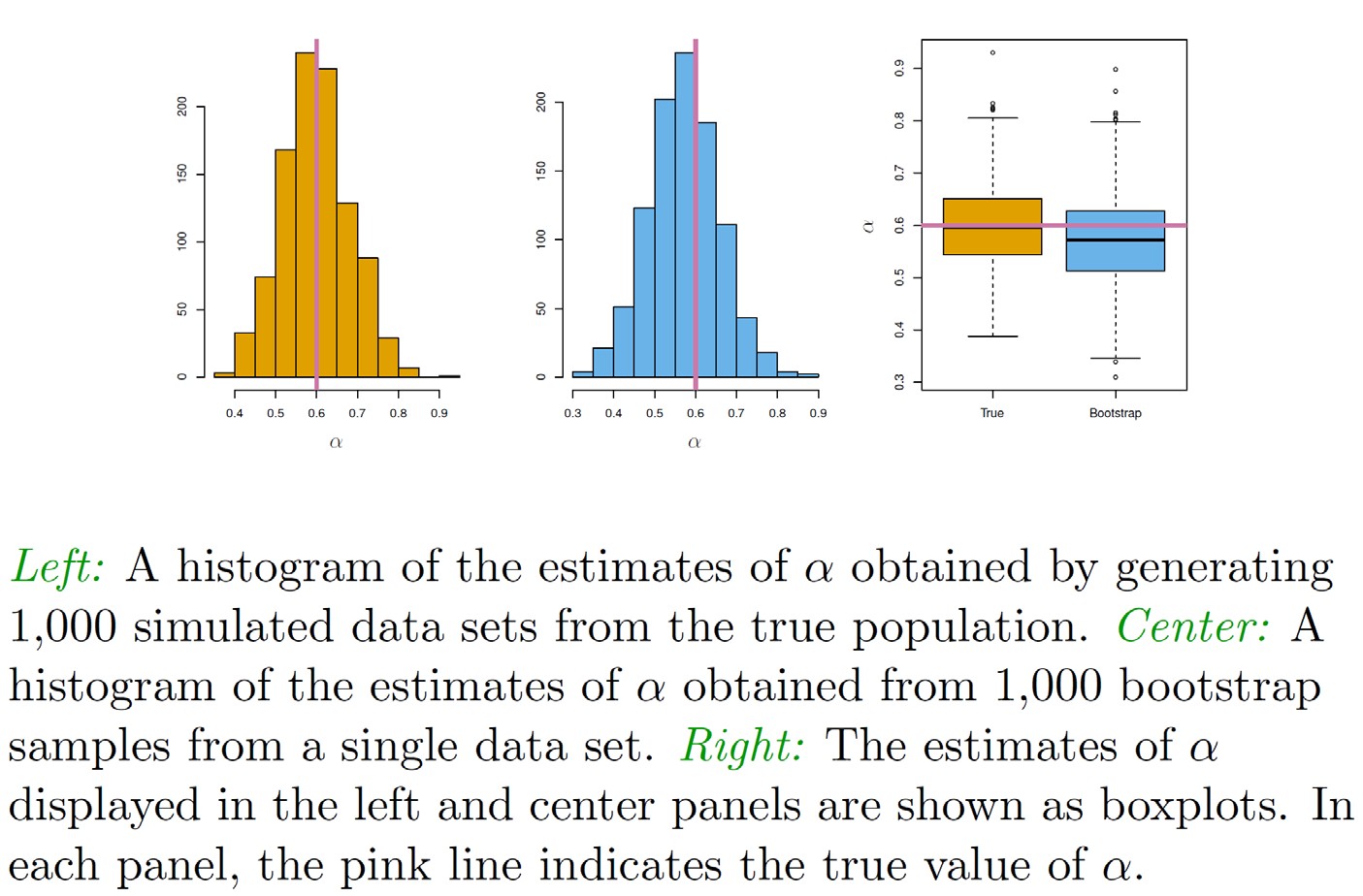
to differ



©copyright CIEL 2015 **33**

# A Simple Example: Results

©copyright CIEL 2015 **34**

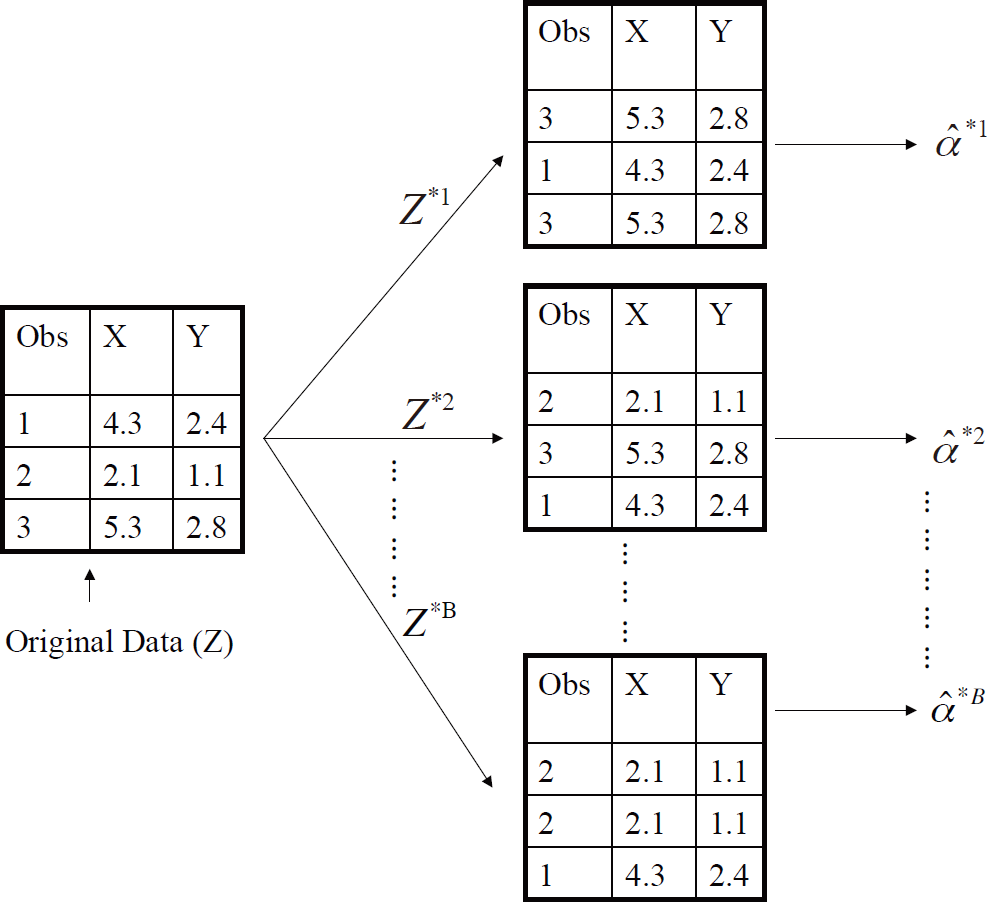


# Now Back to the Real World

* The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population
* However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples
* Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement
*  Each of these “bootstrap data sets" is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all

©copyright CIEL 2015 **35**

# Example with just 3 Observations

* A graphical illustration of the bootstrap approach on a small sample containing 𝑛𝑛 = 3 observations.
* Each bootstrap data set contains 𝑛𝑛 observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of 𝛼𝛼

©copyright CIEL 2015 **36**

# Bootstrap Procedure

* Denoting the first bootstrap data set by 𝑍𝑍∗1, we use 𝑍𝑍∗1 to produce a new bootstrap estimate

for 𝛼𝛼, which we call

𝛼𝛼�∗1

* This procedure is repeated 𝐵𝐵 times for some large value of 𝐵𝐵 (say 100 or 1000), in order to

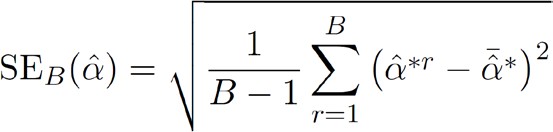
produce 𝐵𝐵 different bootstrap data sets, 𝑍𝑍∗1, 𝑍𝑍∗2, ⋯ , 𝑍𝑍∗𝐵𝐵, and 𝐵𝐵 corresponding 𝛼𝛼 estimates,

𝛼𝛼�∗1,

𝛼𝛼�∗2, ⋯ ,

𝛼𝛼�∗𝐵𝐵

* We estimate the standard error of these bootstrap estimates using the formula



* This serves as an estimate of the standard error of

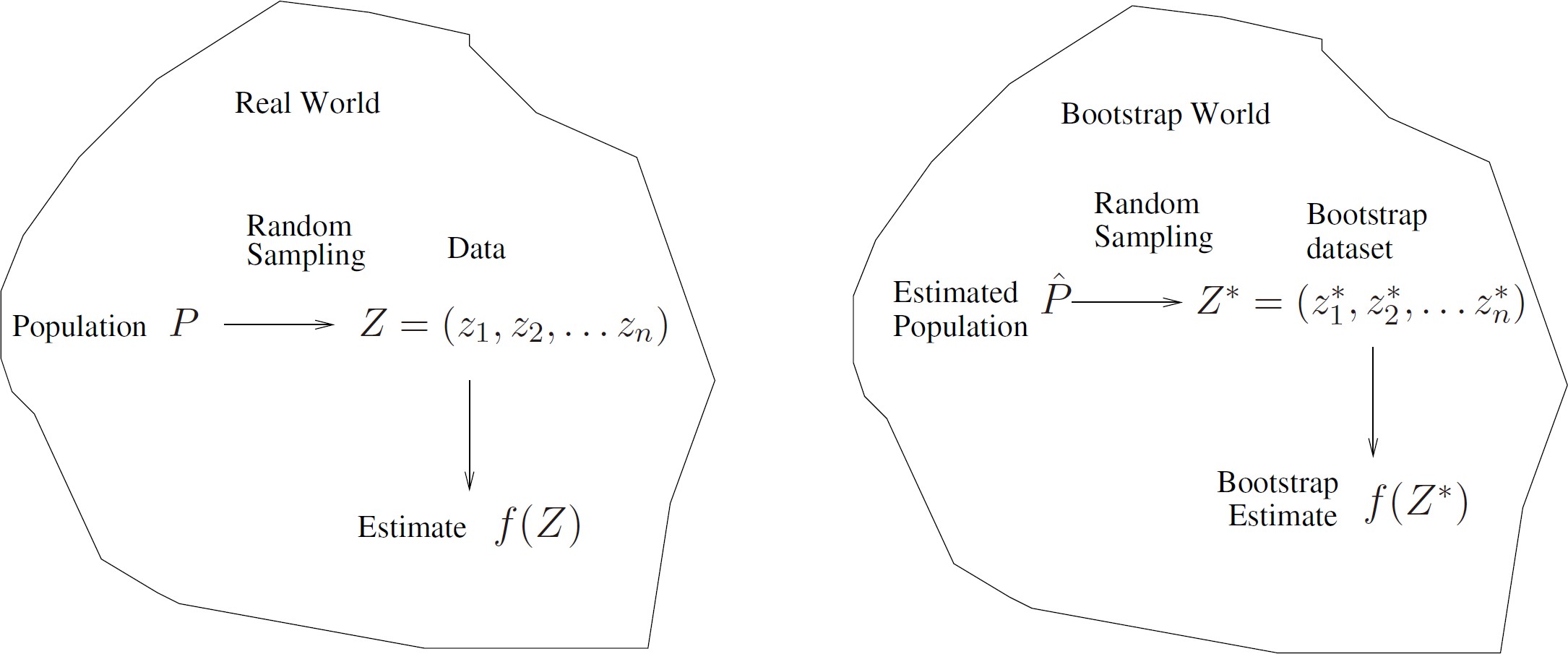
𝛼𝛼�

estimated from the original data set



©copyright CIEL 2015 **37**

# A General Picture for Bootstrap



©copyright CIEL 2015 **38**



**Python Lab**

* Resampling methods

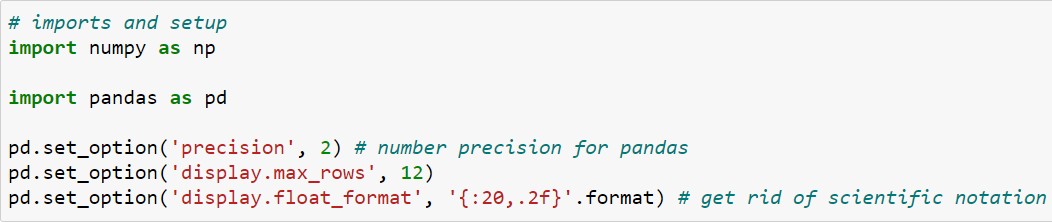
## Python lab

* Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

# 5.3.1 The Validation Set Approach

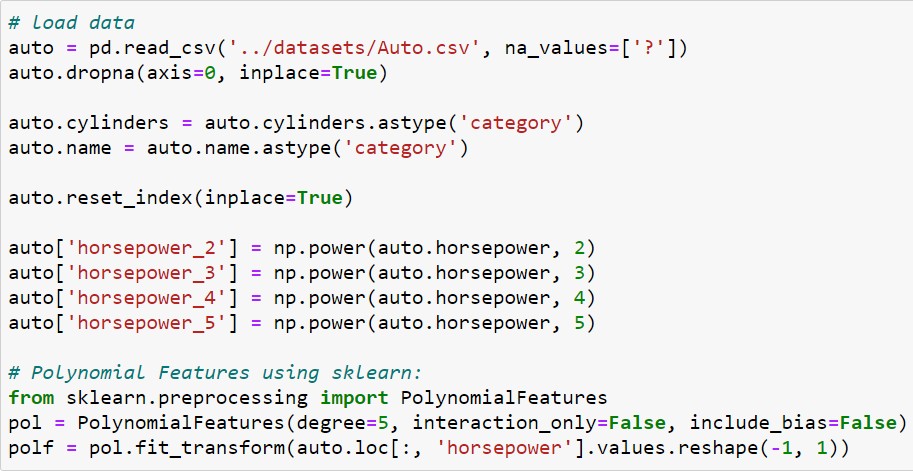
* + Using Python Libraries
    - Import the libraries that are often used for data analysis



©copyright CIEL 2015 **40**

# 5.3.1 The Validation Set Approach

* + Load data: Auto data set



Filter axis labels based on whether values for each label have missing data

Sequence of values to replace with NA



Use the default index not creating a new object

Generate a new feature matrix consisting of all polynomial combinations of the features with degree less than or equal to the specified degree.

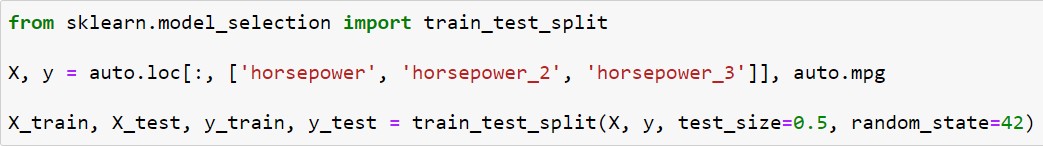
For example, if an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2].

Fit to data, then transform it Selects single column or subset of columns by label

©copyright CIEL 2015 **41**

# 5.3.1 The Validation Set Approach

* + Splitting training & test sets



Split arrays or matrices into random train and test subsets

Proportion of dataset to include in train split

Controls the shuffling applied to the data before applying split

©copyright CIEL 2015 **42**

# 5.3.1 The Validation Set Approach

* + Training, predicting, and evaluating

Whether to calculate the intercept for this mode



Training

Predicting

Evaluating: MSE

Degree 2: lowest MSE

©copyright CIEL 2015 **43**

# 5.3.2 Leave-One-Out Cross-Validation

* + Training & test data set for LOOCV



Equivalent to KFold(n\_splits=n) and LeavePOut(p=1) where n is the number of samples.

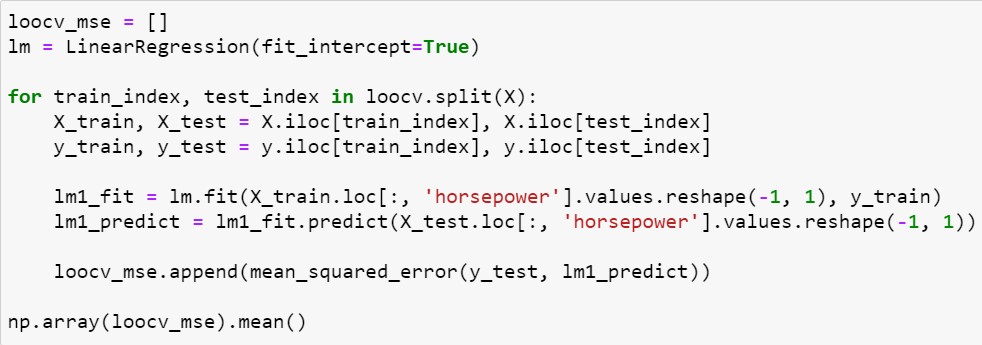
Return the number of splitting iterations in the cross-validator

Now, each column of auto has 392 elements

©copyright CIEL 2015 **44**

# 5.3.2 Leave-One-Out Cross-Validation

* + Mean of MSE for LOOCV



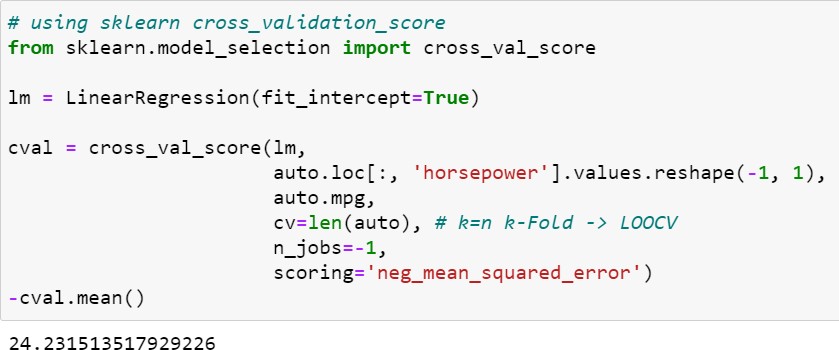
Split results of LOOCV

Average of MSE for each observation

©copyright CIEL 2015 **45**

# 5.3.2 Leave-One-Out Cross-Validation

* + Mean of MSE for LOOCV using cross\_validation\_score function
    - None, to use the default 5-fold cross validation,



* + - int, to specify the number of folds in a (Stratified)KFold,
    - …

Number of jobs to run in parallel

-1 means using all processors

Strategy to evaluate the performance of the cross- validated model on the test set

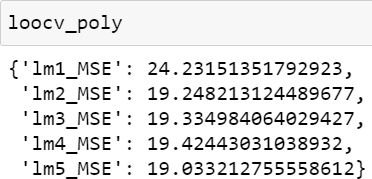
neg\_mean\_squared\_error: minus value of MSE



©copyright CIEL 2015 **46**

# 5.3.2 Leave-One-Out Cross-Validation

* + Loop for 5 degree polynomial linear regressions with LOOCV



Degree 12: sharp drop

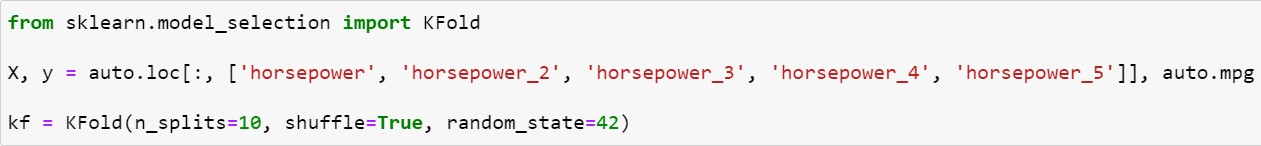
But, other increasing: no clear improvement

©copyright CIEL 2015 **47**



# 5.3.3 𝒌𝒌-Fold Cross-Validation

* + 𝑘𝑘-Folds cross-validator



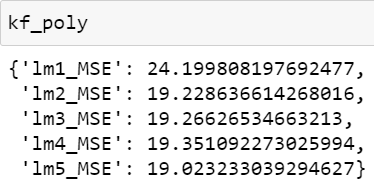
Number of folds

Whether to shuffle the data before splitting into batches

©copyright CIEL 2015 **48**

# 5.3.3 𝒌𝒌-Fold Cross-Validation

* + Loop for 5 degree polynomial linear regressions with k-Fold CV



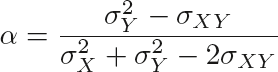
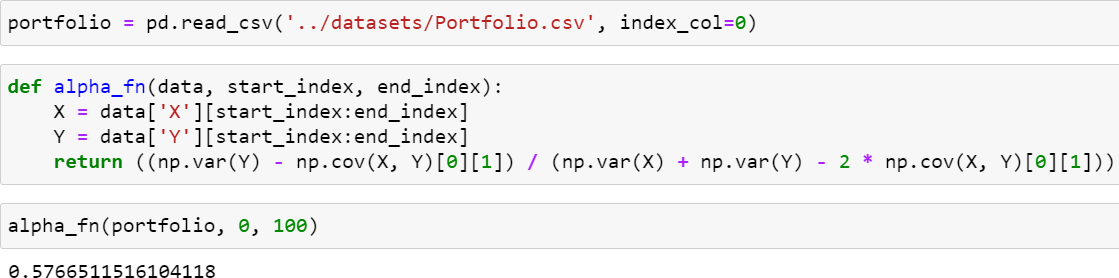
Similar to LOOCV results

©copyright CIEL 2015 **49**



# 5.3.4 The Bootstrap

* + Load data & optimal 𝛼𝛼: Portfolio data set in Section 5.2



Return 𝛼𝛼 that minimizes risk

Value of optimal 𝛼𝛼 that minimizes risk

©copyright CIEL 2015 **50**

# 5.3.4 The Bootstrap

* + Resampling with replacement and bootstrap



Implements resampling with replacement

Default strategy: one step of the bootstrapping procedure

Far from optimal 𝛼𝛼

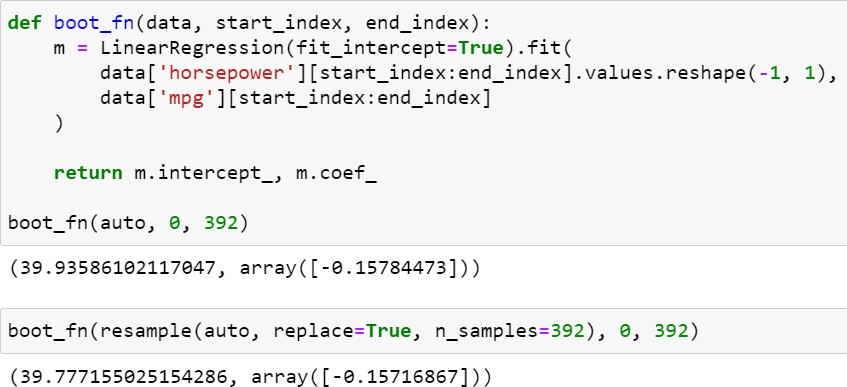
Bootstrap using 1000 samplings

Closer to optimal 𝛼𝛼

©copyright CIEL 2015 **51**

# 5.3.4 The Bootstrap

* + Bootstrap for assessing variability of coefficient estimates



Linear regression

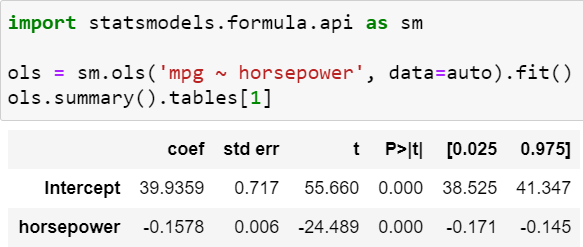
Obtaining coefficient estimates

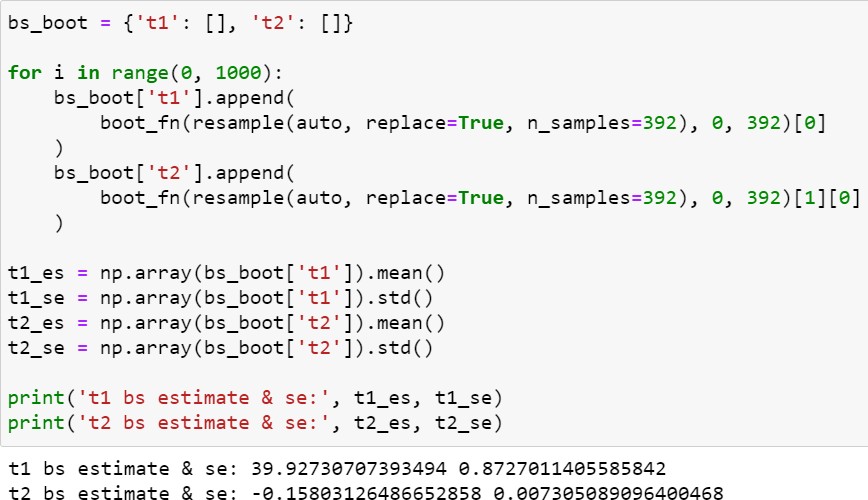
Bootstrap for obtaining coefficient estimates

©copyright CIEL 2015 **52**

# 5.3.4 The Bootstrap

Coefficient estimates using entire data set

* + Bootstrap for assessing variability of coefficient estimates [cont.]



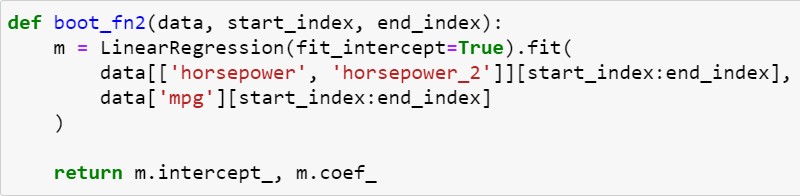
Bootstrap using 1000 samplings

Maybe closer to real coefficients t1: intercept, t2: slope

©copyright CIEL 2015 **53**

# 5.3.4 The Bootstrap

* + Bootstrap for coefficient estimates of polynomial regression

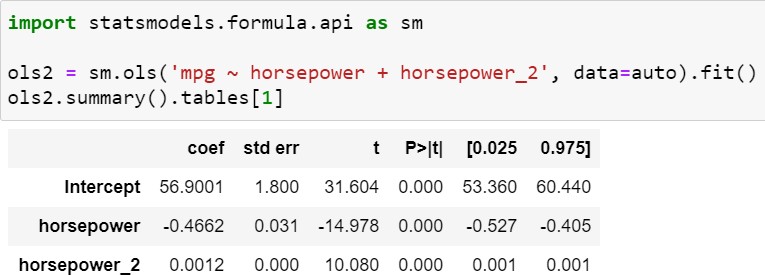


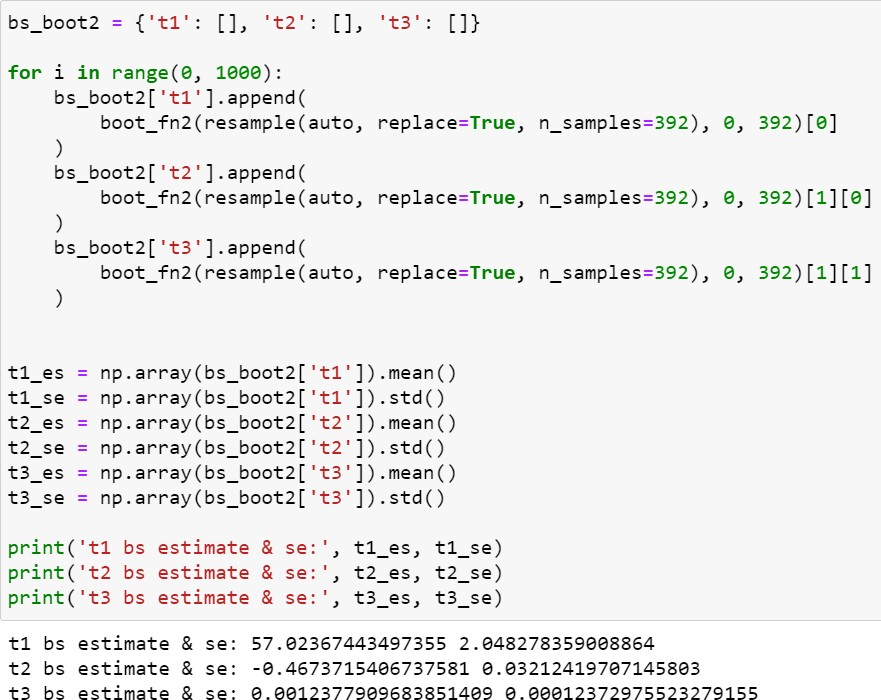
Polynomial regression

©copyright CIEL 2015 **54**

# 5.3.4 The Bootstrap

Coefficient estimates using entire data set

* + Bootstrap for coefficient estimates of polynomial regression



Bootstrap using 1000 samplings

Maybe closer to real coefficients t1: intercept, t2: degree 1,

t3: degree 2

©copyright CIEL 2015 **55**



**Summary & Next Class**

* + - Resampling methods
    - Python lab

## Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

# Summary

* + Resampling methods
* Cross-validation
  + Validation set approach, 𝑘𝑘-fold, LOOCV
  + 𝑘𝑘-fold: typically, 𝑘𝑘 = 5 or 10
* Bootstrap
  + For estimating parameters such as regression coefficients
  + Python lab
* scikit-learn

o model\_slection.train\_test\_split, metrics.mean\_squared\_error

o model\_slection.LeaveOneOut, model\_slection.cross\_val\_score

o model\_slection.KFold

o utils.resample

©copyright CIEL 2015 **57**

# Assignments

* + eClass > Assignments
* Upload 2 or 3 files (do not compress them)
  + Python practices in today’s lecture
* Upload a single ipynb file
* Referring to the lecture slides marked with [P]
* File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_1.ipynb”, e.g., **20211234\_02\_1.ipynb**
  + Textbook exercise problems for today’s lecture
* Conceptual
  + Solving the given problems, then upload your own solution (only docx/hwp formats, not pdf/jpg/bmp etc.)
  + Only include your answers (not need to describe problems)
  + File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_2.ipynb”, e.g., **20211234\_02\_2.docx**
* Applied
  + Implement your Python code for the given problems, then upload another single ipynb file
  + File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_1.ipynb”, e.g., **20211234\_02\_3.ipynb**
  + If not complying with the above policies, some penalty on assignment scores may be given.

©copyright CIEL 2015 **58**

# Course Schedule (Tentative)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Note** | **Date (W)** | **Date (M)** |
| 1 | Orientation, Statistical Learning (Ch2) | Online | 03/03 | 03/08 |
| 2 | Statistical Learning (Ch2), Python Programming | Online | 03/10 | 03/15 |
| 3 | Probability & Statistics | Online | 03/17 | 03/22 |
| 4 | Probability & Statistics | Online | 03/24 | 03/29 |
| 5 | Linear Regression (Ch3) | Online | 03/31 | 04/05 |
| 6 | Linear Regression (Ch3) | Online | 04/07 | 04/12 |
| 7 | Classification (Ch4) | Online | 04/14 | 04/19 |
| 8 | **Midterm exam** | **Class hours (W1-W7)** | **04/21** | **04/26** |
| 9 | Resampling Methods (Ch5) | Online | 04/28 | 05/03 |
| **10** | Linear Model Selection and Regularization (Ch6) | Online | 05/05 | 05/10 |
| 11 | Moving Beyond Linearity (Ch7) | Online | 05/12 | 05/17 |
| 12 | Tree-Based Methods (Ch8) | Online | 05/19 | 05/24 |
| 13 | Support Vector Machines (Ch9) | Online | 05/26 | 05/31 |
| 14 | Unsupervised Learning (Ch10) | Online | 06/02 | 06/07 |
| 15 | **Final exam** | **7pm or Class hours (W9-W14)** | **06/09or14** | **06/09or14** |

©copyright CIEL 2015 **59**