MSBA7002 Business Statistics Tutorial 1

Yutao DENG

The University of Hong Kong

November 7, 2023

Outline

Model Selection

Two goals of fitting a model

- Prediction Accuracy: for the new data
- Model Interpretability: better understanding of the relation and causality

Two types of error measurements involved in statistical learning

- “Wellness of fit”: training error
  - R^2
  - Training RMSE
  - ...
- “Prediction accuracy”: testing error
  - Cross validation RMSE
  - ...

Model Comparison Criteria

Quiz 1

Which of Criteria could be used in model parameters selection ?

Select all that apply:

A. R^2
B. adjusted R^2
C. Akaike Information Criterion (AIC)
D. Mallow's C_p
E. Bayesian Information Criterion (BIC)
F. Training RMSE
G. Cross-Validation RMSE

Three Types of ANOVA[1]

ANOVA

- Short model v.s. Long model

$$Y \sim \sum_{i=1}^{p} X_i \tag{1}$$

$$Y \sim \sum_{i=1}^{p} X_i + \sum_{j=p+1}^{p+q} X_j \tag{2}$$

- Check if the ”additional” SSR is significant compared to that of original model (F-value).

$$\frac{(SSR_2 - SSR_1)/(df_2 - df_1)}{SSR_2/df_2} \sim F(df_2 - df_1, df_2)$$

[1] https://mcfromnz.wordpress.com/2011/03/02/anova-type-iiiiii-ss-explained/

Three Types of ANOVA

Type I ANOVA

$$Y \sim X_1 + X_2 + X_3$$

- anova(fit1, fit2) $\Leftarrow$ basic R function
- Specific the order. e.g. $X_2 \rightarrow X_3 \rightarrow X_1$
- fit them steps by steps (from "short" model to "longer" model)

$$Y \sim X_2 \tag{3}$$
$$Y \sim X_2 + X_3 \tag{4}$$
$$Y \sim X_2 + X_3 + X_1 \tag{5}$$

- Check if the "additional" SSR is significant compared to that of original model (F-value).

Three Types of ANOVA

Type I ANOVA

Remarks:

- q in equation (2) could be larger than 1, for each comparison.
- Order will drive the final result.
- Stop once insignificant p-value ($p > 0.05$) appears.

Three Types of ANOVA

Type II ANOVA

$$Y \sim X_1 + X_2 + X_3$$

- Anova(fit) $\Leftarrow$ R function in "car" package
- Start from the longest model with all independent variables
- Try to delete one independent variables

$$Y \sim X_1 + X_2 + X_3 \tag{6}$$
$$Y \sim X_2 + X_3 \tag{7}$$
$$Y \sim X_1 + X_3 \tag{8}$$
$$Y \sim X_1 + X_2 \tag{9}$$

- Check if the "reduced" SSR is significant compared to that of original model (F-value).

Three Types of ANOVA

Type II ANOVA

Remarks:

- q in equation (2) must be 1.
- Delete the most insignificant independent variable first, and the do Type-II anova iteratively until all $\widehat{\beta}$ is significant.
- Other way to delete variable?

Bias and Variance

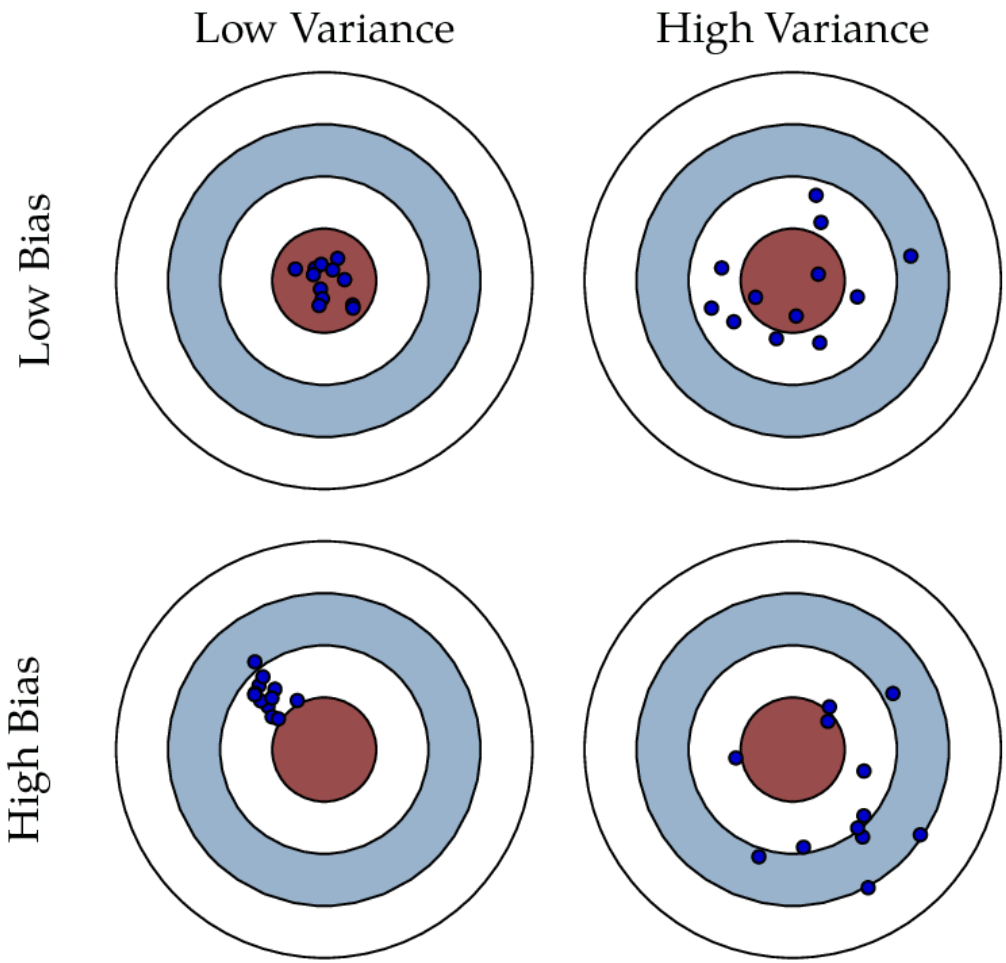


Figure: Bias: how much far off on average the model is from the truth.
Variance: how much that estimate varies around its average

Derive the Bias-Variance Decomposition

Preliminary Knowledge

X, Y are random variables; α is a constant; f is the real model; $\hat{f}$ is the estimation of the model; y and $\hat{y}$ are the response and the predicted value respectively.

1. $E(X+Y) = E(X) + E(Y)$
2. $E(\alpha X) = \alpha E(X)$
3. $\mathrm{Var}(X) = E(X^2) - [E(X)]^2 \Rightarrow E(X^2) = \mathrm{Var}(X) + [E(X)]^2$
4. $E(y) = f$
5. $E(\hat{y}) = E(\hat{f})$
6. $\mathrm{Var}(y) = \sigma_\epsilon$
7. ϵ and $\hat{f}$ are independent

Derive the Bias-Variance Decomposition

Derivation

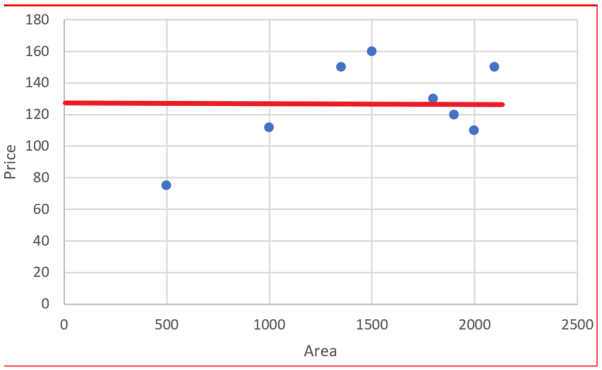
Proof.

Starting point: MSE loss

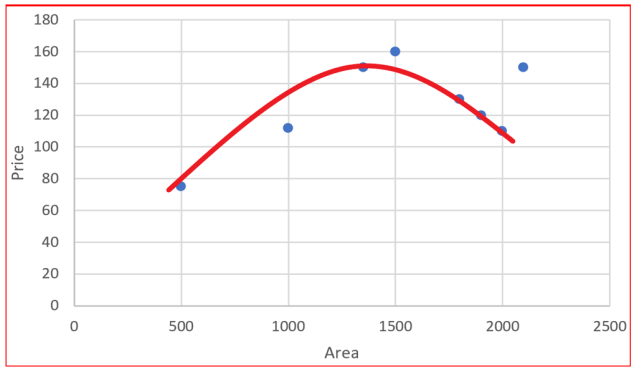
$$
\begin{aligned}
E\left(y-\hat{y}\right)^2 &= E(y^2 - 2y\hat{y} + \hat{y}^2), \\
&= E(y^2) - 2fE(\hat{f}) + E(\hat{y}^2), \\
&= \text{Var}(y) + [E(y)]^2 - 2fE(\hat{f}) + \text{Var}(\hat{y}) + [E(\hat{y})]^2, \\
&= \sigma_\epsilon + f^2 - 2fE(\hat{f}) + \text{Var}(\hat{y}) + [E(\hat{f})]^2, \\
&= \sigma_\epsilon + [f - E(\hat{f})]^2 + \text{Var}(\hat{y}), \\
&= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}.
\end{aligned}
$$

Bias Variance Trade-off

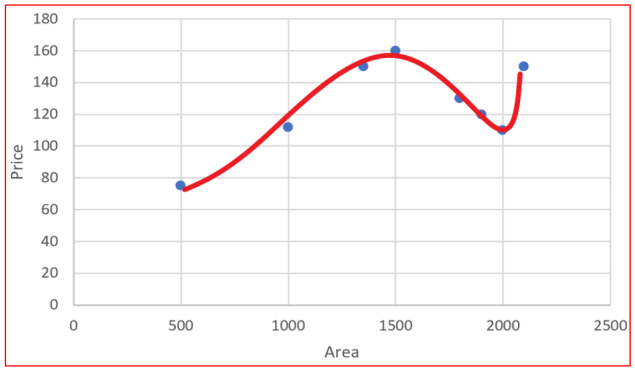
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$



High Bias - underfit



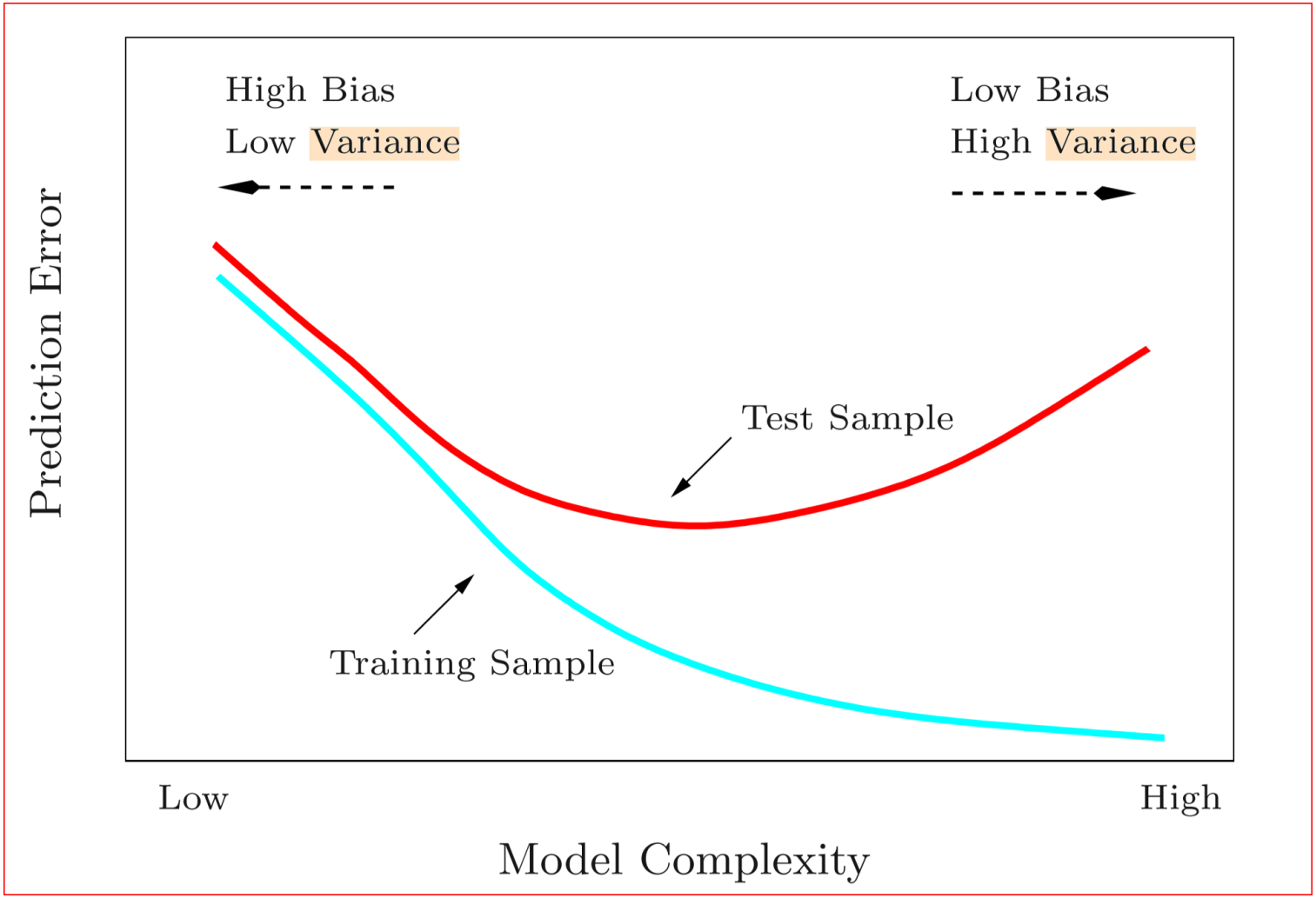
Just Fit



High Variance – overfit

$$y = \alpha, \qquad y = \beta x^2 + \alpha, \qquad y = \sum_{i=1}^{4} \beta_i x^i + \alpha$$

Bias Variance Trade-off



Double descent*

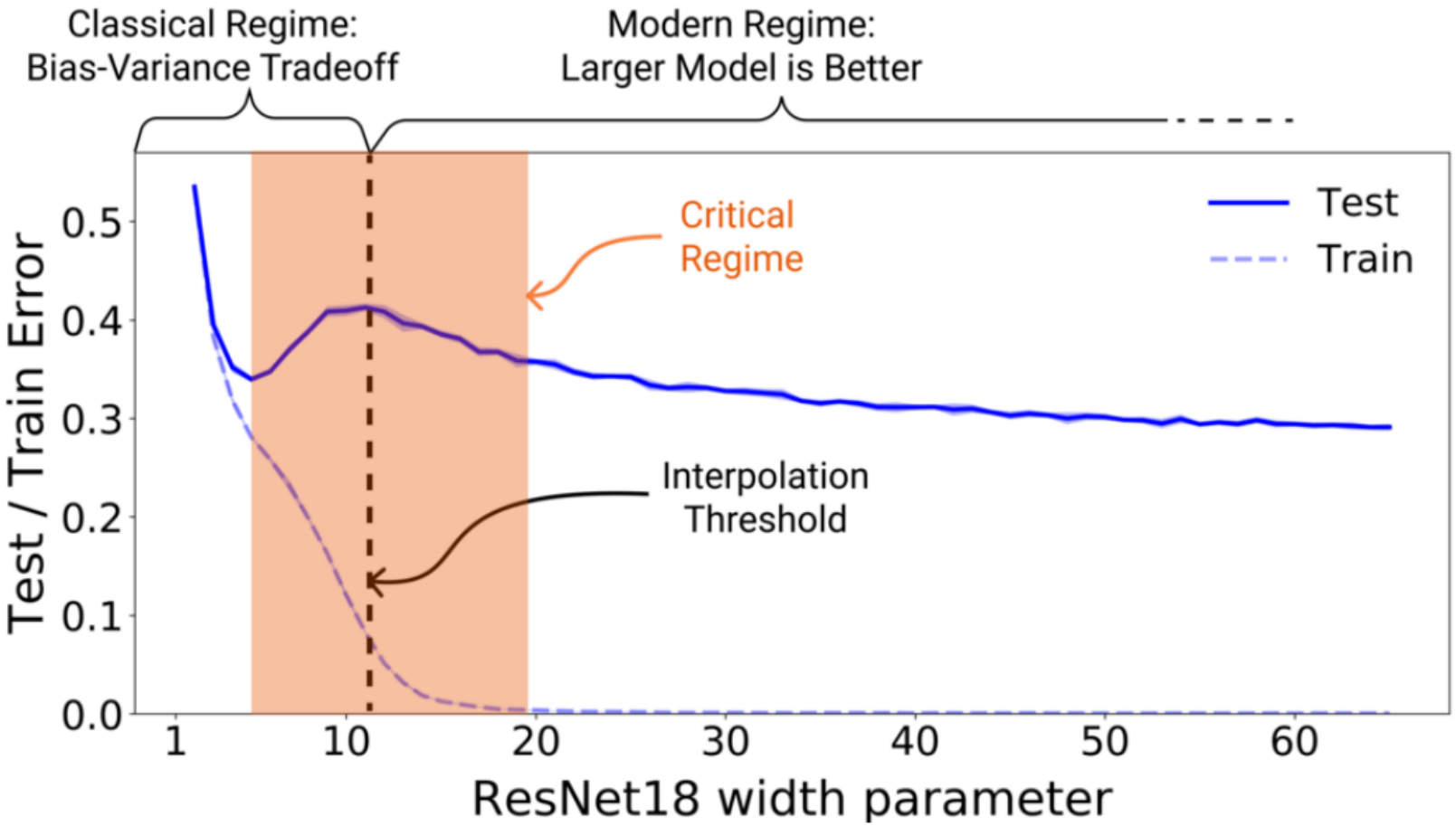


Figure: Nakkiran P, Kaplun G, Bansal Y, et al. Deep double descent: Where bigger models and more data hurt[J]. Journal of Statistical Mechanics: Theory and Experiment, 2021, 2021(12): 124003.

Regularization

Two Regularization Methods

Lasso regression (L1 penalty)

$$\min_{\beta} \left(\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

$$\iff \min_{\beta} \left(\text{RSS} \right) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

Ridge regression (L2 penalty)

$$\min_{\beta} \left(\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

$$\iff \min_{\beta} \left(\text{RSS} \right) \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

Regularization

Two Regularization Methods

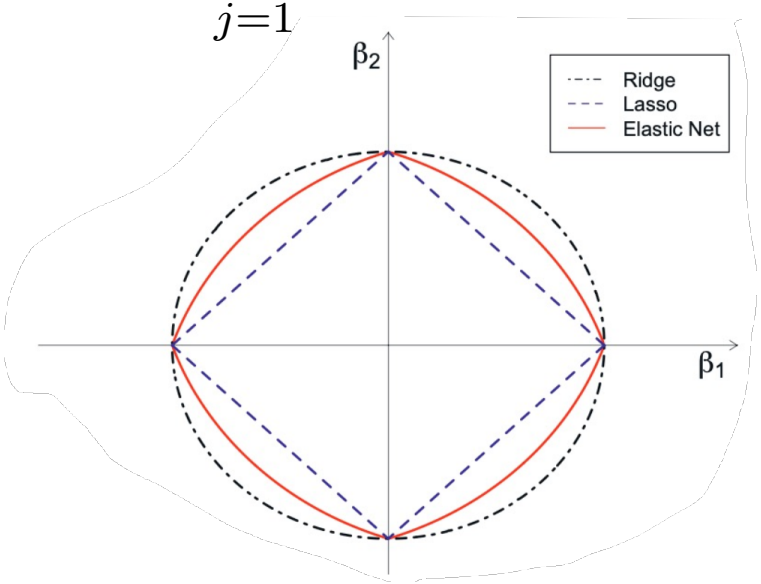
Elastic net (L1 penalty + L2 penalty)

$$\min_{\beta} \left(\text{RSS} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right)$$

$$\iff \min_{\beta} \left(\lambda \text{RSS} + \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

Remarks:

- Selecting a good weight for regularization is critical; cross-validation is used for this.
- Standardizing the predictors before adding regularization.



Validation Set and Cross-Validation

- Why do we need validation?

Validation Set and Cross-Validation

- Why do we need validation?
  - No test data is available when fitting the model.

Validation Set and Cross-Validation

- Why do we need validation?
  - No test data is available when fitting the model.
- Validation procedure:
  - Divide the whole sample into two parts:
  - The model is fit on the training set, accessed on the validation set.

Validation Set and Cross-Validation

- Why do we need validation?
  - No test data is available when fitting the model.
- Validation procedure:
  - Divide the whole sample into two parts:
  - The model is fit on the training set, accessed on the validation set.
- How can we use the whole sample for validation?

Validation Set and Cross-Validation

- Why do we need validation?
  - No test data is available when fitting the model.
- Validation procedure:
  - Divide the whole sample into two parts:
  - The model is fit on the training set, accessed on the validation set.
- How can we use the whole sample for validation?
  - Use Cross-Validation
  - Randomly divide the data into K equal-sized parts.

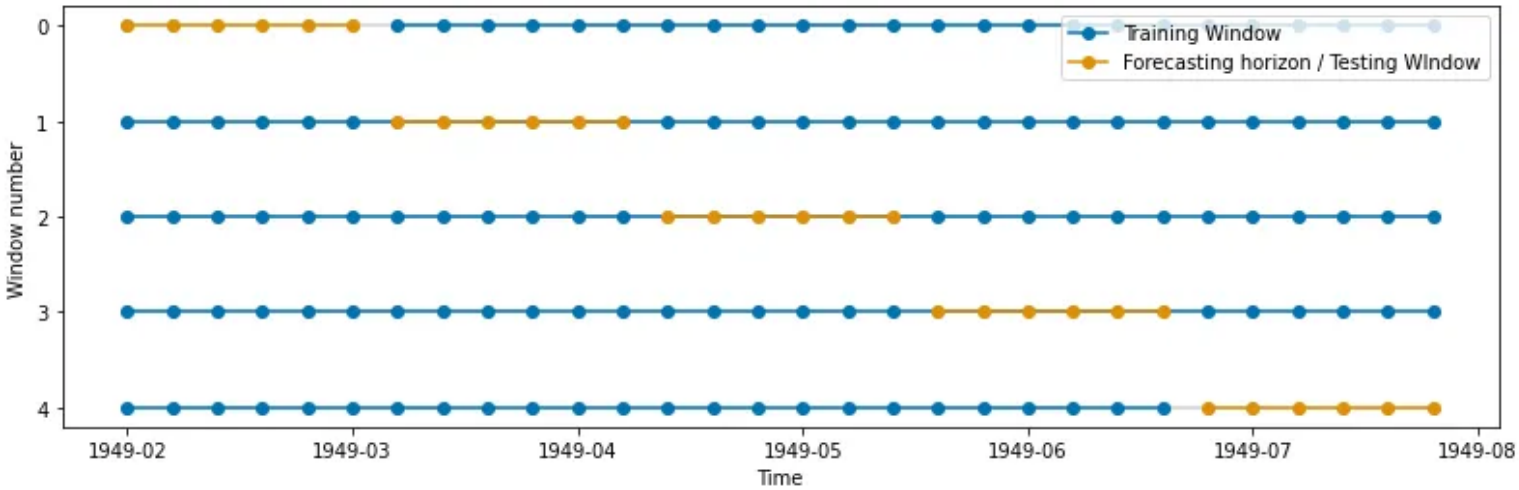
Validation Set and Cross-Validation

- Why do we need validation?
  - No test data is available when fitting the model.
- Validation procedure:
  - Divide the whole sample into two parts:
  - The model is fit on the training set, accessed on the validation set.
- How can we use the whole sample for validation?
  - Use Cross-Validation
  - Randomly divide the data into K equal-sized parts.
- Key point of Cross-Validation:
  - Model fitting and validation are two independent procedure.
  - Can we use Cross-Validation on time-series data?

Validation Set and Cross-Validation

Adjustment for time-series*

- Cross-Validation

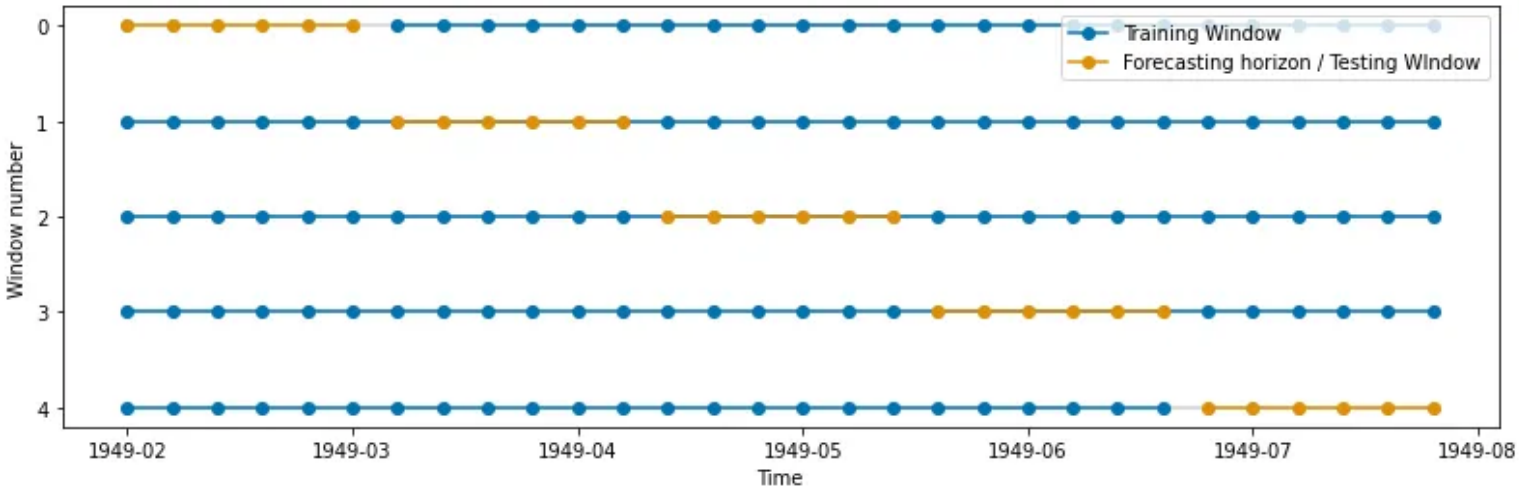


- Do we have any problems?

Validation Set and Cross-Validation

Adjustment for time-series*

- Cross-Validation

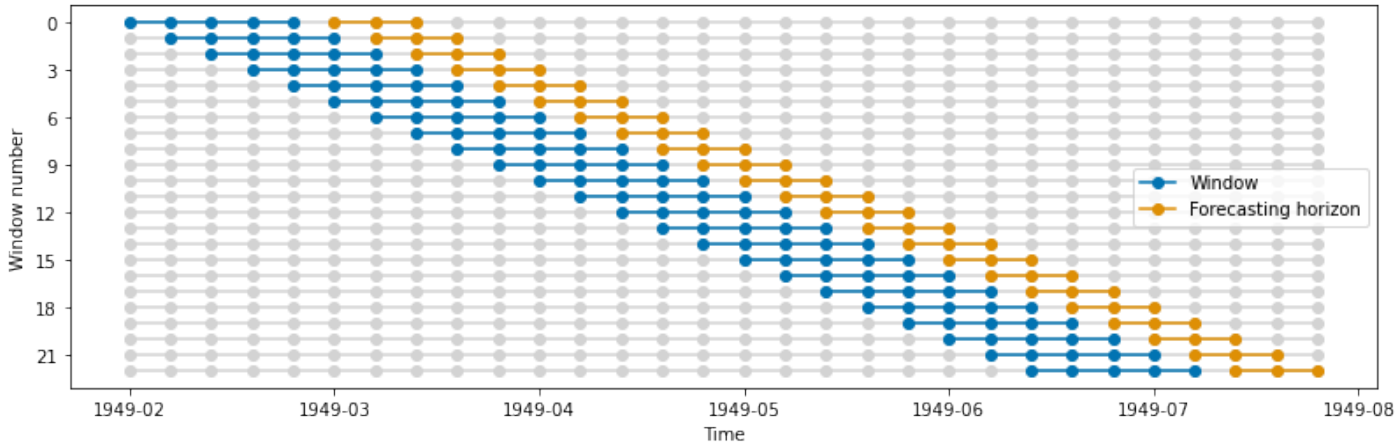


- Do we have any problems?
  - Forecast / test data occurs before the training data.
  - Data leakage.

Validation Set and Cross-Validation

Adjustment for time-series*

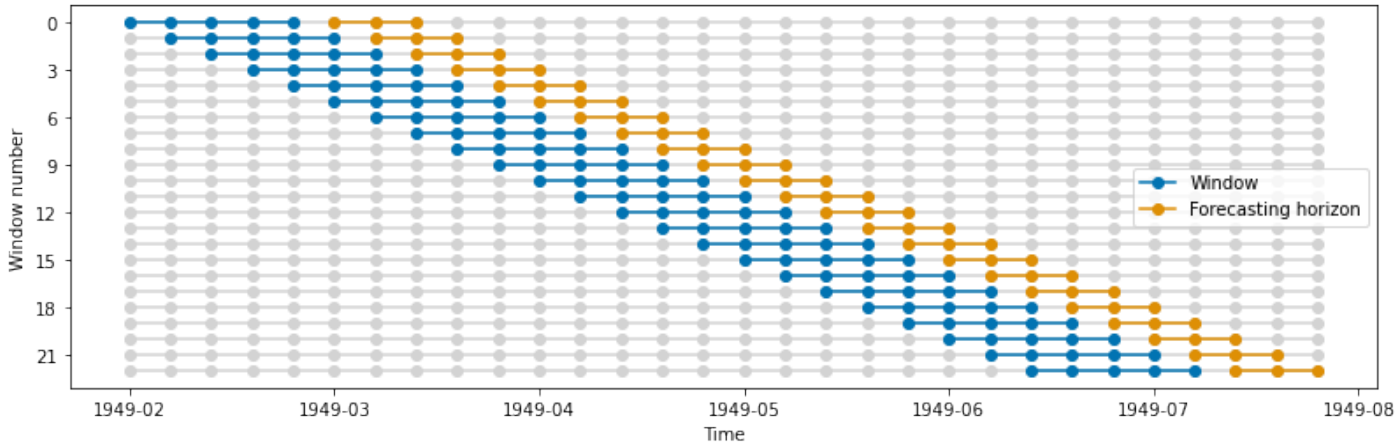
- Rolling windows



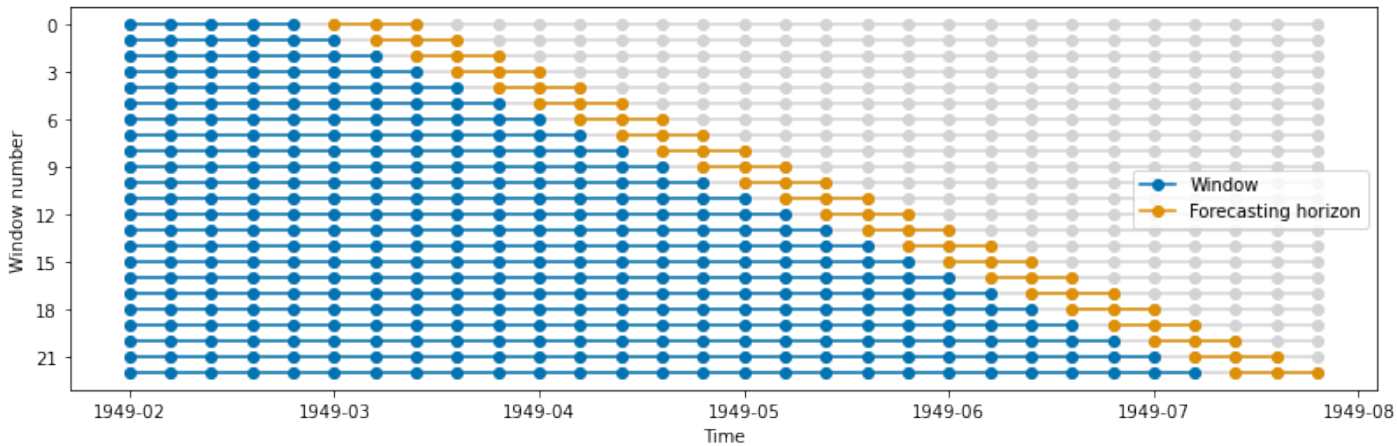
Validation Set and Cross-Validation

Adjustment for time-series*

- Rolling windows



- Expanding windows



Outline

Introduction

- R Markdown is a file format for making dynamic documents with R (similar to iPython Notebook!).
- Structure if R Markdown
  - Headers for information and settings
  - R Chunks: small block to implement R codes
  - Text and math equations
  - Table and plots(more details in the next tutorial)

```
install.packages("rmarkdown")
```

YAML Header

```
---
title: "Tutorial 1"
author: "Your name"
date: "Oct 30, 2023"
output:
  pdf_document:
    toc: yes
    toc_depth: "4"
  html_document:
    code_folding: show
    highlight: haddock
    theme: lumen
    toc: yes
    toc_depth: 4
    toc_float: yes
 ---
```

YAML Header

```
---
title: "Tutorial 1"
author: "Your name"
date: "Oct 30, 2023"
output:
  pdf_document:
    toc: yes
    toc_depth: "4"
  html_document:
    code_folding: show
    highlight: haddock
    theme: lumen
    toc: yes
    toc_depth: 4
    toc_float: yes
 ---
```

- Some information required in the title page

YAML Header

```
---
title: "Tutorial 1"
author: "Your name"
date: "Oct 30, 2023"
output:
  pdf_document:
    toc: yes
    toc_depth: "4"
  html_document:
    code_folding: show
    highlight: haddock
    theme: lumen
    toc: yes
    toc_depth: 4
    toc_float: yes
 ---
```

- Formatting the output PDF file. Please click *here* for a more detailed introduction.

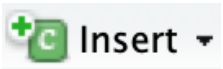
YAML Header

```
---
title: "Tutorial 1"
author: "Your name"
date: "Oct 30, 2023"
output:
  pdf_document:
    toc: yes
    toc_depth: "4"
  html_document:
    code_folding: show
    highlight: haddock
    theme: lumen
    toc: yes
    toc_depth: 4
    toc_float: yes
 ---
```

- Formatting the output HTML file. Please click *here* for a more detailed introduction.

R Chunks

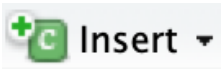
Unlike regular texts, a chunk is where the code will be executed in an R Markdown file. Two ways to quickly add a regular R chunk

- the keyboard shortcut Ctrl+Alt+I (OS X: Cmd+Option+I)
- the Add Chunk command 

in the editor toolbar

R Chunks

Unlike regular texts, a chunk is where the code will be executed in an R Markdown file. Two ways to quickly add a regular R chunk

- the keyboard shortcut Ctrl+Alt+I (OS X: Cmd+Option+I)
- the Add Chunk command 

in the editor toolbar

Regular R chunk

```
```{r}
```
```

R Chunks

```
```{r, include = FALSE}
# code: not shown
# results: not shown
```
```

```
```{r, fig.show='hide'}
# code: shown
# fig results: not shown
```
```

```
```{r, echo = FALSE}
# code: not shown
# results: shown
```
```

```
```{r, warning = FALSE}
# Not printing warnings
```
```

```
```{r, results = 'hide'}
# code: shown
# not showing text results
```
```

```
```{r, fig.cap = "..."}
# Add caption to figures
```
```

Working Directory

Claim global setting in ”R setup” chunk.

For example, we could set working directory in this chunk.

````
```{r, setup}
setwd(some_dir) # set the working dir to some_dir
```
````

Caveat

Duplicated ”r setup” chuck is not allowed,

or RMarkdown will report error :(

Useful links

- Formatting the text!
- Formatting the output PDF file
- Formatting the output HTML file
- R Markdown documentation
- (*advance level*) R Markdown gallery

Useful links

Installing R, RStudio, and useful links

- Install R:

  `https://cran.rstudio.com/`
- Install RStudio:

  `www.rstudio.com/products/rstudio/download/`
- Data Visualization - A Practical Introduction

  `https://socviz.co/`
- RStudio Cheat Sheets

  `www.rstudio.com/resources/cheatsheets/`

R Implementation

Use the rmd file

- Introduction of R
  - Numeric and string objects
  - Vectors, Matrices and Dataframes
  - Defining functions and Control flows
- R Implementation
  - EDA
  - Linear Model
  - Model Selection
    - ANOVA
    - Regularization
    - Subset selection