Statistical Learning for Data Science 1

229351

Dates

- Sec 01 : Tu 14:30-16:30 at SCB4202
  Lab : F 14:30-16:30 at STB107

- Sec 02 : M 11:00-13:00 at SCB4202
  Lab : Th 11:00-13:00 at STB207

Instructors

- Donlapark Ponnoprat
  Email: `donlapark.p--a--cmu.ac.th`
  Office: STB304

- Phimphaka Taninpong
  Email: `phimphaka.t--a--cmu.ac.th`
  Office: STB201

Lectures

- Mainly focuses on predictions/forecasts

- Covers four main topics:
    - Principal component analysis
    - Linear regression
    - Time series analysis
    - Logistic regression

- Prerequisites: comfort with basic algebra and probability

Main principle

- Predictors $X = (X_1, X_2, \ldots, X_p)$

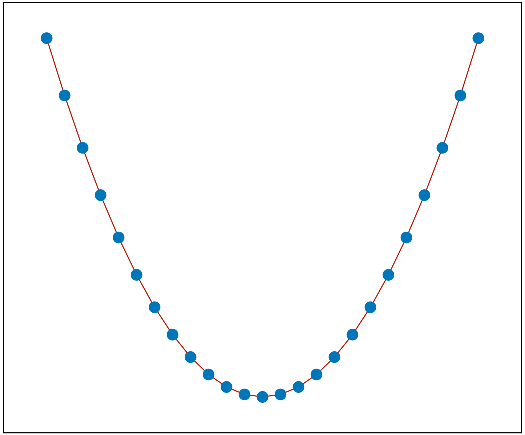
- Response Y

Assumption: There's some function f and error ϵ such that

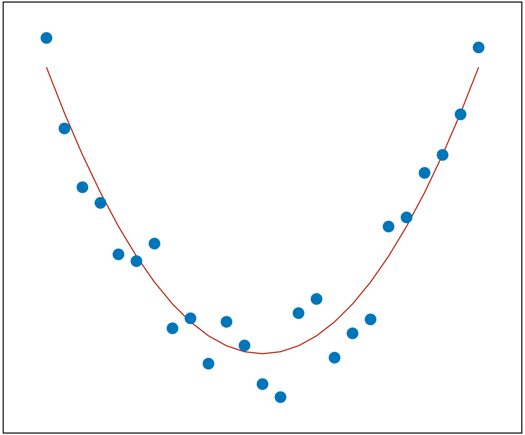
$$Y = f(X) + \epsilon$$

Here, ϵ is a **random noise**.

Without noises



With noises



Our goals

- Prediction/forecast
    - Learn f from noisy data
    - Make predictions.

Our goals

- Prediction/forecast
  - Learn f from noisy data
  - Make predictions.

- Make decisions
  - Is there any relationship between X and Y?
  - Can we remove some variables?

  We will use technique from statistics: the **hypothesis testing**.

Labs

- 10-12 labs
- Mainly in google Colab (python)
- Recommended to work in groups, but write your own solutions!
- Turn in your Colab file on Microsoft teams

Homework

- 4 homework, due once a month
- conceptual problems & coding problems
- turn in solutions via Microsoft Teams

Kaggle competition

- Try to build a model that is as accurate as possible!
- We give you training data → build a model → evaluate on test data
- The competition will take place on kaggle (www.kaggle.com); a kaggle account is required
- Compete as a group of $1 - 3$ people

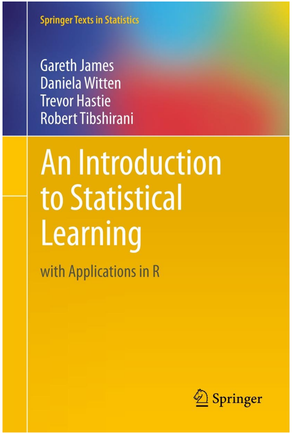
kaggle report

- After the competition, you will have to write a kaggle report
- 4-10 pages, one- or two-column format
- Examples of kaggle reports are given on the course website
- Due on Mar 28

Grading scheme

4 homework	10%
10-12 labs	15%
Kaggle	15%
Midterm (TBD)	30%
Final (Mar 28)	30%

Textbooks



Springer Texts in Statistics

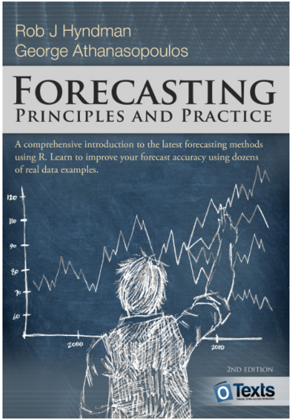
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

James et al. An Introduction to Statistical Learning



Rob J Hyndman
George Athanasopoulos

FORECASTING
PRINCIPLES AND PRACTICE

A comprehensive introduction to the latest forecasting methods using R. Learn to improve your forecast accuracy using dozens of real data examples.

2ND EDITION

OTexts

Hyndman et al. Forecasting: Principles and Practice

Course website

Syllabus, homework, slides can be found at

`donlapark.github.io/ds351`

We appreciate your comments on these materials!