1. Introduction

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Statistical Modelling

- Statistical modelling: A simplified, mathematically-formalized way to approximate reality (i.e., what generates data).

- Main purposes of statistical modelling:
  - Prediction: Regression, Classification.
  - Information: Patterns, Associations, etc.

- Two cultures of statistical modelling (Breiman, L., 2001):
  - Data modelling culture.
  - Algorithmic modelling culture.

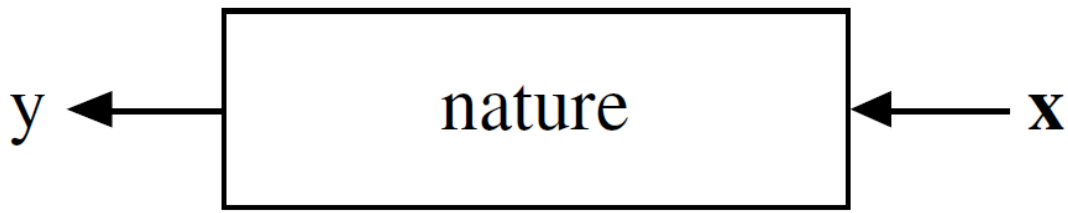

Figure: Prediction problem; black box between y and $\boldsymbol{x}$.

Data Modelling Culture

- Traditional statistical models.
- This culture starts with assuming a stochastic data model for the inside of the black box.
- Common data models: $y = f(\boldsymbol{x}, \epsilon, \boldsymbol{\theta})$,
  - ϵ: Error; $\boldsymbol{\theta}$: Model parameters.
- Mainly for both prediction and information.
- Model validation: Goodness-of-fit tests and residual diagnosis.

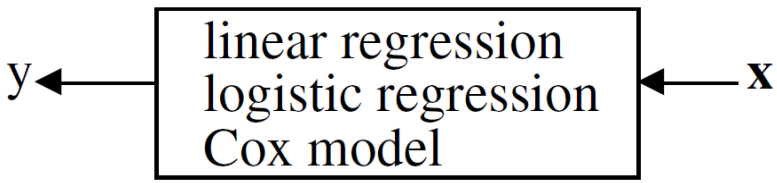

Figure: Data modelling culture in the prediction problem

Algorithmic Modelling Culture

- This culture considers that inside of the black box is complex and unknown.
- Finding a function $f(\boldsymbol{x})$ by an algorithm operating on $\boldsymbol{x}$.
- Mainly for prediction.
- Model validation: Prediction accuracy (e.g., test error).

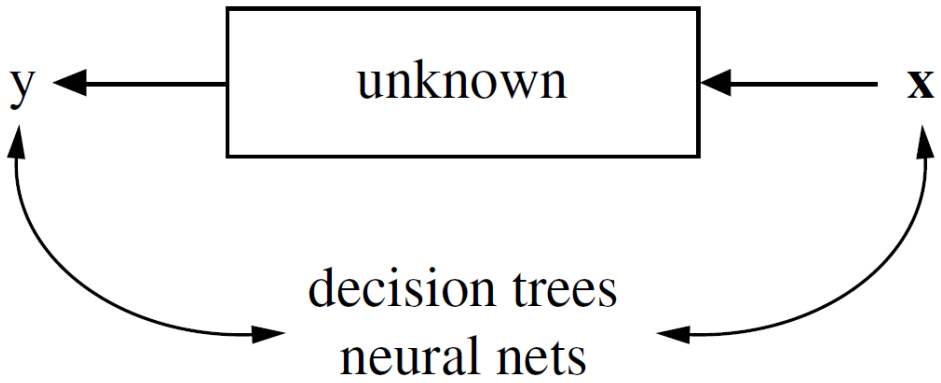

Figure: Algorithmic modelling culture in the prediction problem

Advantages of Data Modelling

- Interpretability: Association, variable selection, mechanism of data generation.
- It can consider uncertainties (e.g., confidence interval, standard error of estimates, etc.)
- It is easy to consider subject-matters for data (e.g., data quality, experimental design, background of data, etc.)

Disadvantages of Data Modelling

- Distributional assumptions are required.
- Conclusions are about the model's mechanism, not about data nature's mechanism.
- If models are a poor emulation of data nature, conclusions could be wrong.
- Model validation problem: Goodness-of-fit tests and residual analysis are often fail to validate models (e.g., linearity tests are usually hard to reject in high-dimension).
- Different models with similar goodness-of-fits might lead to different conclusions.

Advantages of Algorithmic Modelling

- Better prediction accuracy.
- Model specification is not required.
- No distributional assumptions (only iid sample is assumed).
- Different models with similar goodness-of-fits) $\Rightarrow$ Model averaging $\Rightarrow$ Better prediction.

Disadvantages of Algorithmic Modelling

- Poor interpretability due to complex nature of models.
- Intensive computing power.
- Results from CV or test sample are not very stable.
- Results of tuning parameter from CV or test sample have some bias.

Modelling for Data Analysis

- As a statistician or data analyst, we do not have to distinguish data and algorithmic modellings.
- Both modelling techniques are useful tools for data analysis.
- Big data present great opportunities. However, it is not the size that matters.
- Asking the right question, using the right model, applying the right statistics make all the difference.