Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for CE Engineers

In-Class Material: Class 01

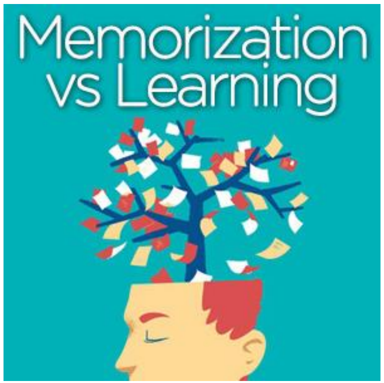
Introduction (ISL Chapter 1)

1. Statistical Learning

(a) Definition: a vast set of tools and processes for U__________ Data

“(Machine) learning is NOT m_____________”
– Yaser Abu-Mostafa

U_________ data means (1) identifying the relationship between input(s) and output, or similarity between inputs, and (2) being able to predict/cluster for cases that have not been experienced



©gakkomom.com

(b) Types of statistical learning

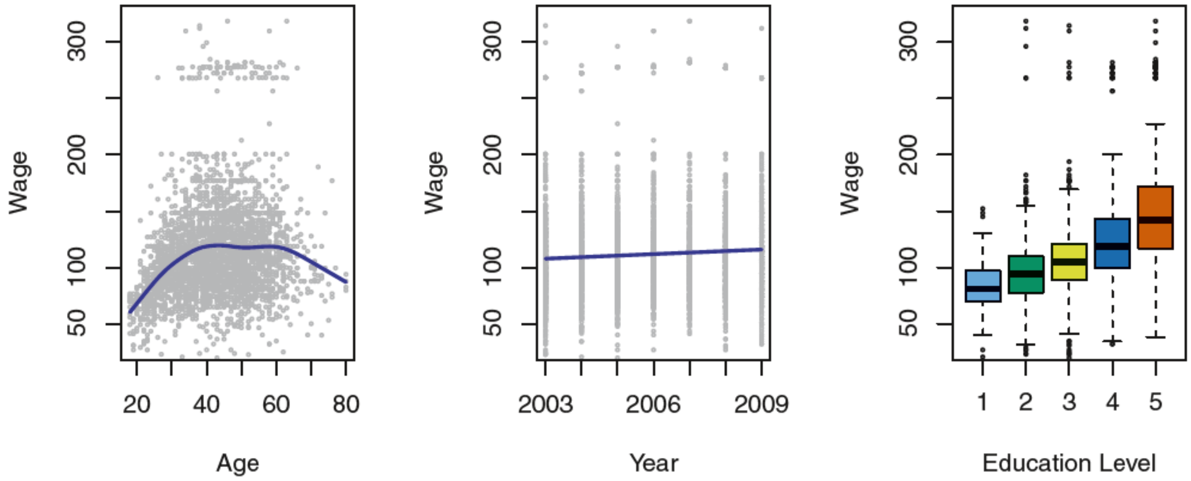
- **Supervised** – predicting an output based on one or more inputs
  - Regression (Ch3): c__________ or q___________ output
  - Classification (Ch4): c__________ or q___________ output

- **Unsupervised** – inputs but no supervising output
  - Clustering (Ch10): structuring or grouping

2. Real-world Data Sets (Considered throughout the “ISL” Textbook)

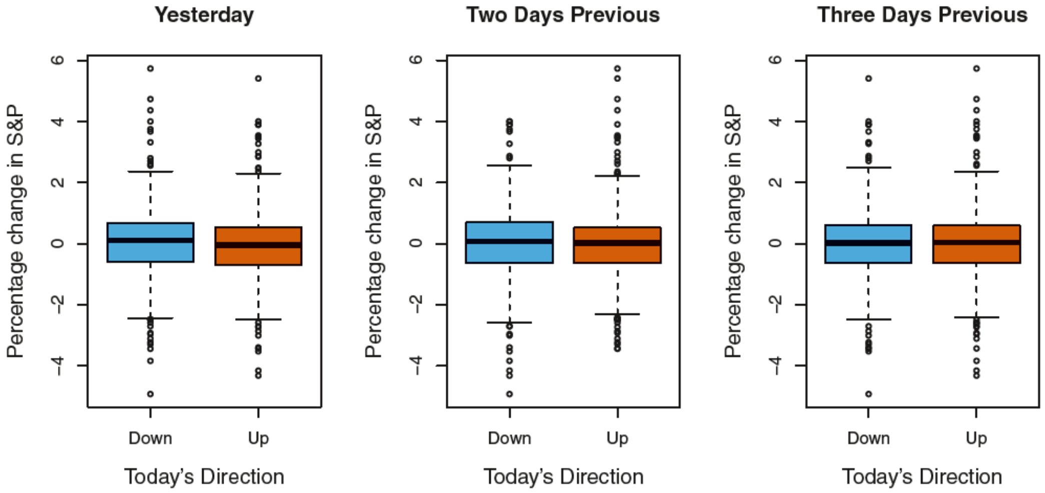
Notes: You can obtain these data sets by installing “ISLR” package to your R environment

(a) Wage data: wages for a group of males from the Atlantic region of the United States



One can understand the relationship between the output (wage) and important inputs (age, year, education, etc.) and predict the wage for given input values by linear r_________ (Ch3) and nonlinear r_________ (Ch7)

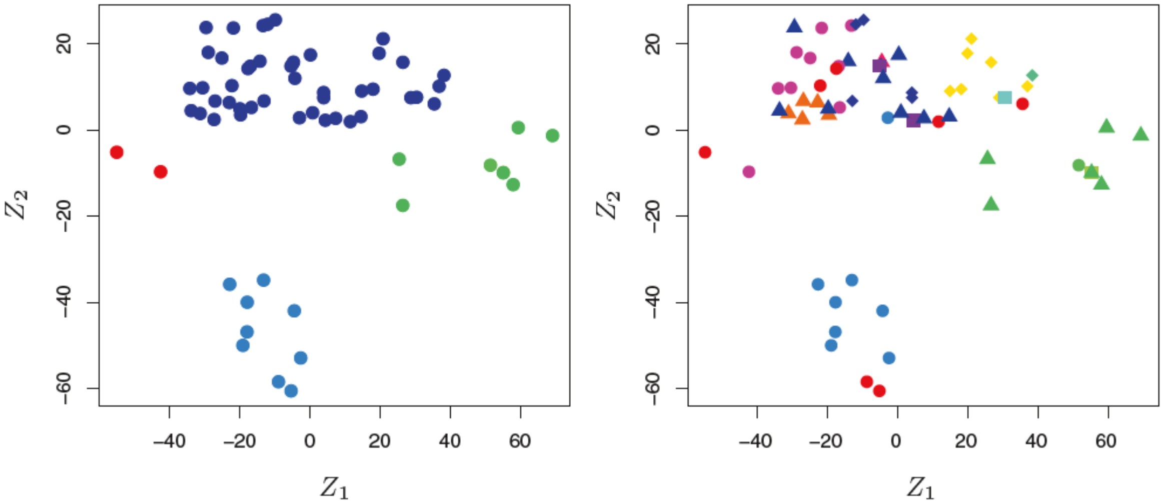
(b) Smarket data: daily movements in the S&P 500 stock index from 2001 to 2005



The goal is to predict whether the index will *increase* or *decrease* using the past 5 days' percentage change (result: around 60% accuracy) → C__________ problem (Ch4)

(c) NCI60 data: 6,830 gene expression measurements for each of 64 cancer cell lines

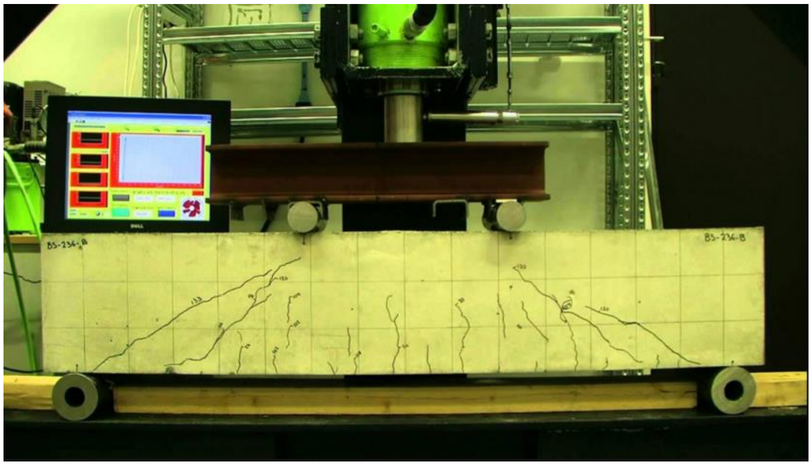
6,830 input variables → approximately represented by the first two *principal components* Z_1 and Z_2 (Details of this dimension reduction method available in Ch6)



Using a C________ method (Ch10), four groups are identified for the 64 cancer cell lines (Left). Actually, it is known that there exist 14 different types of cancer (Right). It seems that the cell lines with the same type of cancer belong to the same cluster

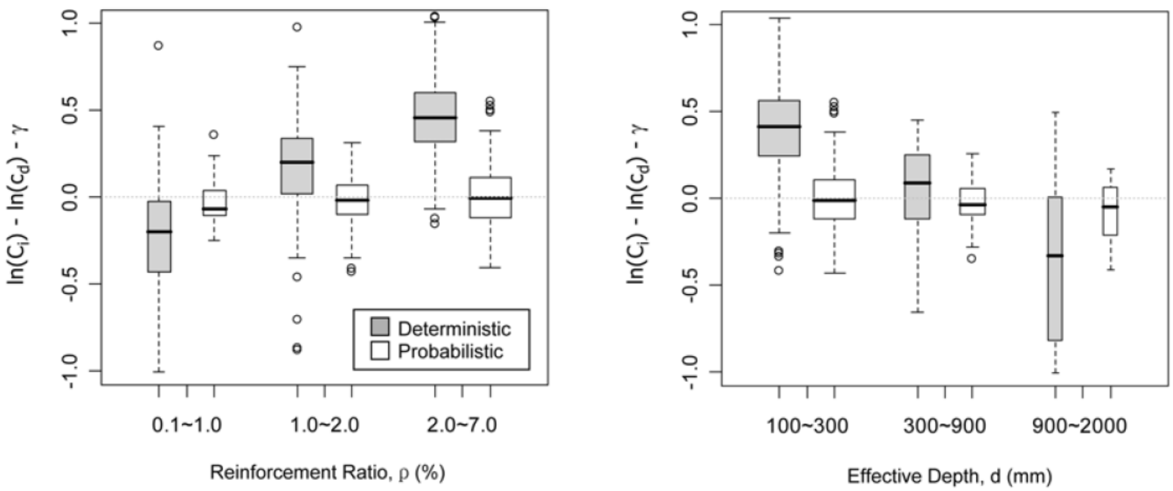
3. Statistical Learning in CEE Research

(a) **Regression:** Predict the shear strengths of reinforced concrete beams for given characteristics (geometry, material, etc.) of a beam



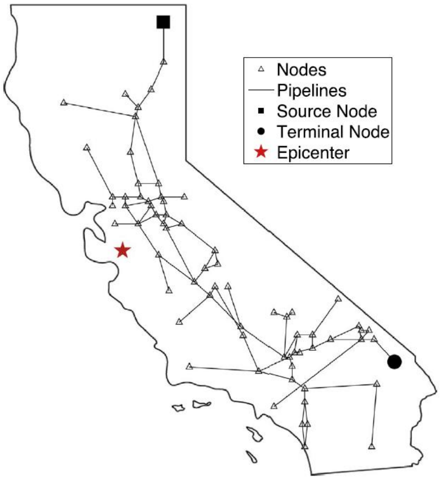
Source: https://www.youtube.com/watch?v=DPQIpT1ZvXY

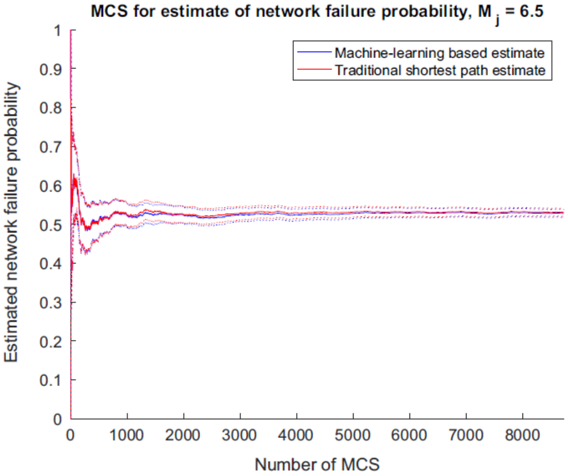
Using nonlinear regression (based on Bayesian parameter estimation), Song et al. (2010) were able to predict the shear strength accurately while quantifying the uncertainty in the predictions.



Song, J., W.-H. Kang, K.S. Kim, and S. Jung (2010). Probabilistic shear strength models for reinforced concrete beams without shear reinforcement. *Structural Engineering & Mechanics*, Vol. 34(1), 15-38.

(b) **Classification**: Predict whether two points in an infrastructure network, e.g. gas pipeline network would be still connected if a certain group of network components are damaged by a scenario earthquake





Using logistic regression (Ch4) and support vector machine (Ch9), Stern et al. (2017) were able to approximately predict the connectivity (without using network analysis algorithm) and thus improved the efficiency of Monte Carlo simulation.

Stern, R.E., J. Song, and D.B. Work (2017). Accelerated Monte Carlo system reliability analysis through machine-learning-based surrogate models of network connectivity. *Reliability Engineering & System Safety*. Vol. 164, 1-9.

(c) **Clustering**: Group similar earthquake ground motion records in terms of their important features

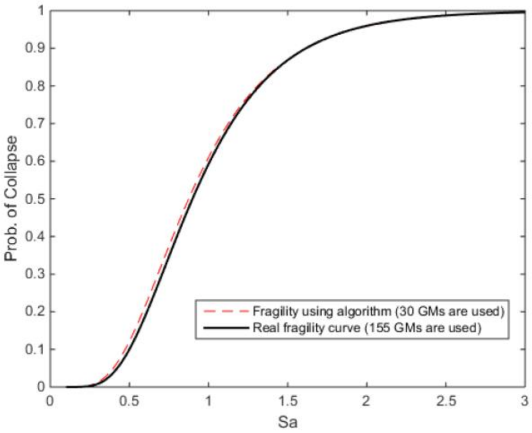


Image (left) Source: https://www.youtube.com/watch?v=8zyyPLNMSfw

Using K-means clustering method (Ch10), Kim (2017) was able to obtain similar fragility curves using only 30 ground motions each of which represents a cluster, compared to those constructed by 155 ground motions.

T. Kim (2017). *Enhancing seismic fragility analysis of structural system: developing intensity measure and ground motion selection algorithm*. MS Thesis, Seoul National University.

4. Brief History of Statistical Learning

(a) Underlying concepts developed:

- **Method of Least Squares** (linear regression): Legendre and Gauss (beginning of 19th century)
- **Linear discriminant analysis** (classification): Fisher (1936)
- **Logistic regression** (classification and regression): Various authors (1940s)
- **Generalized linear models** (entire class of statistical learning methods including linear and logistic regression): the term coined by Nelder and Wedderburn (early 1970s)

(b) Advanced methods developed:

- Many more techniques for linear methods (by the end of 1970s)
- Computing technology → enables fitting nonlinear relationship (by the 1980s)
- **Classification and regression trees**: Breiman, Friedman, Olshen and Stone (mid 1980s; demonstrated detailed practical implementation, e.g. cross-validation for model selection)
- **Generalized additive models**: nonlinear extension of generalized linear model; the term coined by Hastie and Tibshirani (1986)

(c) Emerge as a new subfield in statistics:

- Inspired by the advent of machine *learning* and other disciplines
- Increasing availability of powerful and relatively user-friendly software, such as R system (a brief introduction to "R" will be provided in Class 04)

5. Organization of Course and ISL Textbook

Class No	***Contents***
1	Introduction (Ch1)
2-4	Statistical Learning (Ch2)
5-8	Linear Regression (Ch3)
9-11	Classification (Ch4)
12-13	Resampling Methods (Ch5)
14-17	Linear Model Selection and Regularization (Ch6)
18-19	Moving Beyond Linearity (Ch7)
20-21	Tree-Based Methods (Ch8)
22-23	Support Vector Machines (Ch9)
24-26	Unsupervised Learning (Ch10)

Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for CE Engineers

In-Class Material: Class 02

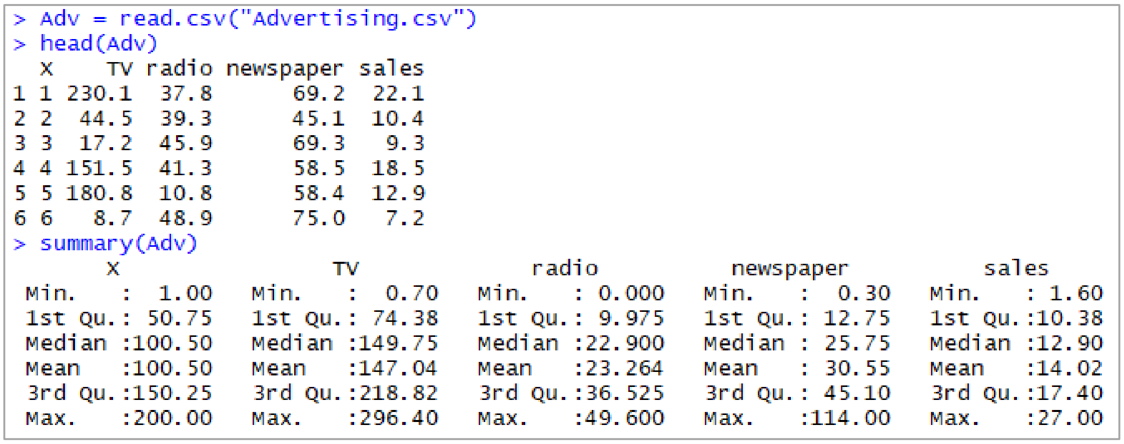
Statistical Learning (ISL Chapter 2)

What is “Statistical Learning”?

1. Mathematical description and key terminologies
2. Two main reasons for statistical learning: P________ and I__________
3. How to estimate? P________ vs N___________ methods
4. Prediction Accuracy vs Model Interpretability
5. Supervised vs Unsupervised Learning
6. Regression vs Classification

1. Mathematical Description and Key Terminologies

(a) Advertising data (“Advertising.csv” at the ISLR textbook website): budgets (in $k) spent for advertisement through TV, radio and newspaper, and corresponding sales (in thousands of units) for 200 different markets



```
> Adv = read.csv("Advertising.csv")
> head(Adv)
  X    TV radio newspaper sales
1 1 230.1  37.8      69.2  22.1
2 2  44.5  39.3      45.1  10.4
3 3  17.2  45.9      69.3   9.3
4 4 151.5  41.3      58.5  18.5
5 5 180.8  10.8      58.4  12.9
6 6   8.7  48.9      75.0   7.2
> summary(Adv)
       X                TV              radio          newspaper           sales      
 Min.   :  1.00   Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60  
 1st Qu.: 50.75   1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38  
 Median :100.50   Median :149.75   Median :22.900   Median : 25.75   Median :12.90  
 Mean   :100.50   Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02  
 3rd Qu.:150.25   3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40  
 Max.   :200.00   Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00  
```

(b) General mathematical model describing the true relationship

The relationship $\text{sales} = f(\text{TV}, \text{radio}, \text{newspaper}) + \text{error/noise}$ can be described by a general mathematical model

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

- $X = (X_1, X_2, \ldots, X_p)$: **input variables**, predictors, independent variables, features, or variables
- Y: **output variable**, response, or dependent variable
- $f(\cdot)$: fixed but unknown **function** of $X_1, X_2, \ldots, X_p$, representing s________ information that X provides about Y
- ϵ: r_______ error term, which is independent of X and has mean z_______

“Statistical learning” refers to a set of approaches for e__________ $f(\cdot)$

2. Main Reasons for Statistical Learning: Prediction & Inference

(a) **Prediction**: want to predict Y using a set of inputs X

$$\hat{Y} = \hat{f}(X)$$

- $\hat{f}(\cdot)$: e___________ for $f(\cdot)$
- $\hat{Y}$: prediction of Y resulting from the estimated model, i.e. _____
- If prediction is the only reason for statistical learning, it is fine to treat $\hat{f}(\cdot)$ as a "b______ box" (as long as it yields accurate predictions)

Errors of prediction by $\hat{Y} = \hat{f}(X)$, i.e.

$$Y = f(X) + \epsilon \quad \text{vs} \quad \hat{Y} = \hat{f}(X)$$

- **Reducible Error**: inaccuracy of the imperfect estimate $\hat{f}(\cdot)$
  → can be reduced by using the most appropriate statistical learning technique, etc.
- **Irreducible Error**: Even if $\hat{f}(\cdot) = f$, i.e. perfect estimation, $\hat{Y} = f(X) \neq Y$
  - Because the true response Y is also a function of ___
  - Reasons for non-zero irreducible error ϵ
    a. <u>Unmeasured variables</u>: $f(\cdot)$ may miss important but unmeasured variables
    b. <u>Unmeasurable variation</u>: $f(\cdot)$ itself may vary

These errors can be described mathematically as follows:

$$\begin{aligned} \mathrm{E}\left(Y - \hat{Y}\right)^2 &= \mathrm{E}\left[f(X) + \epsilon - \hat{f}(X)\right]^2 \\ &= \\ &= \left[f(X) - \hat{f}(X)\right]^2 + \mathrm{Var}(\epsilon) \end{aligned}$$

This course focuses on methods (Ch3, 4, 8, 9 and 10) and techniques (Ch5, 6 and 7) for estimating $f(\cdot)$ with the aim of minimizing the r___________ error.

(b) **Inference**: want to u_________ the way that Y is a_____________ as $X_1, X_2, \ldots, X_p$ change, i.e. the r__________ between X and Y → the estimate cannot be treated as a b________ box

Questions to answer by inference:

- **Which predictors** are associated with the response?
  - Identifying the few important ones and using them
- **What is the relationship** between the response and each predictor?
  - Positive, negative, depending on the others? e.g. $Y = 2X_1 - 3X_2 + X_3X_4$
- Is the relationship **linear or more complicated**?

You might be interested in "Prediction only", "Inference only" or "Both" (See Page 20).

3. How to Estimate?

Statistical learning aims to find a function $\hat{f}(\cdot)$ such that $Y \approx f(X)$ for *any* observation (X, Y) based on t________ data:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \text{ where } x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$$

(a) **Parametric method**: reduces the problem of estimating $f(\cdot)$ down to one of estimating a set of model p__________; a two-step m_______-based approach

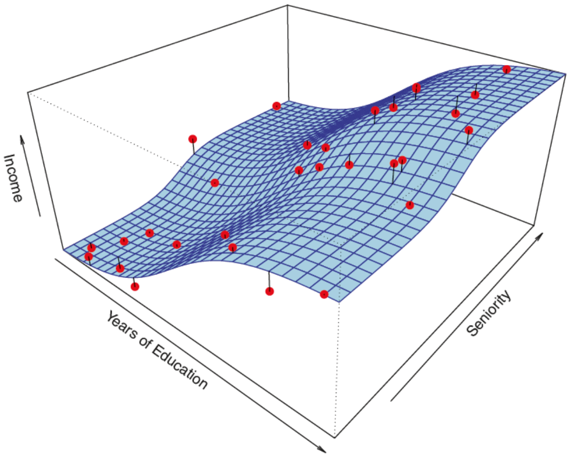
- **First step**: make an assumption about the f________ f______, or shape of f
  If a linear model is assumed:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

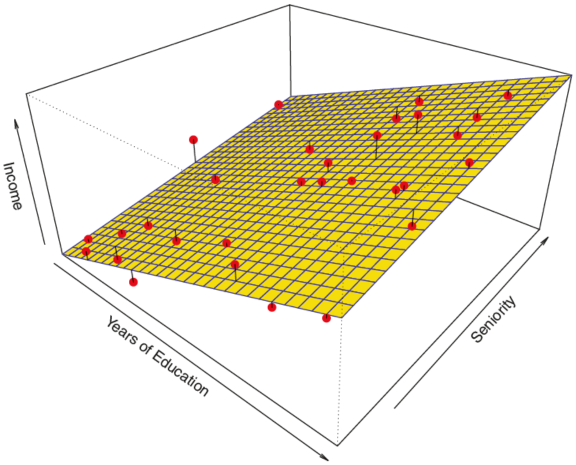
- **Second step**: Estimate the p__________ (instead of arbitrary function) such that the model with the estimated parameters can approximate the response well, e.g.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Example: linear regression of “simulated” Income data (“Income2.csv” at ISLR site)



$Y = f(X) + \epsilon$



Parametric estimation by linear model (Ch3)

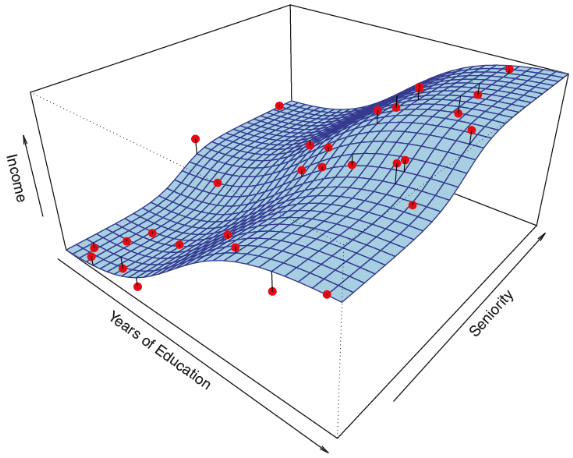
- **Advantage**: much easier to estimate
- **Disadvantage**: the model we choose will usually NOT match the true unknown form of $f(\cdot)$ → may produce a poor estimate

Then, why not using more flexible models?

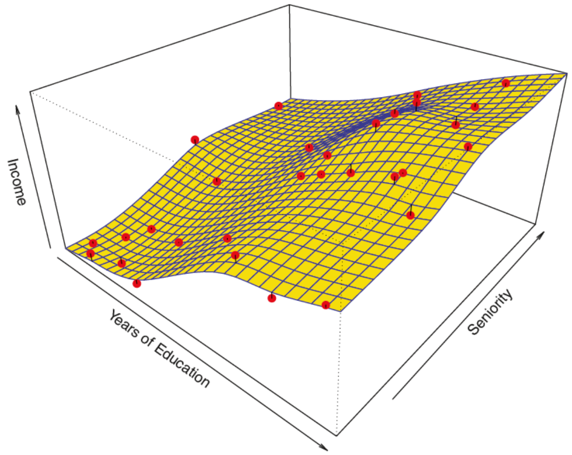
- **Greater number of parameters** required
- **“Overfitting” the data**: may follow the errors or noise too closely

(b) **Non-parametric method**: does not make explicit assumptions about the functional form of $f(\cdot)$

- **Advantage**: potential to accurately fit a wider range of possible shapes
- **Disadvantage**: a very l_______ number of observations is required

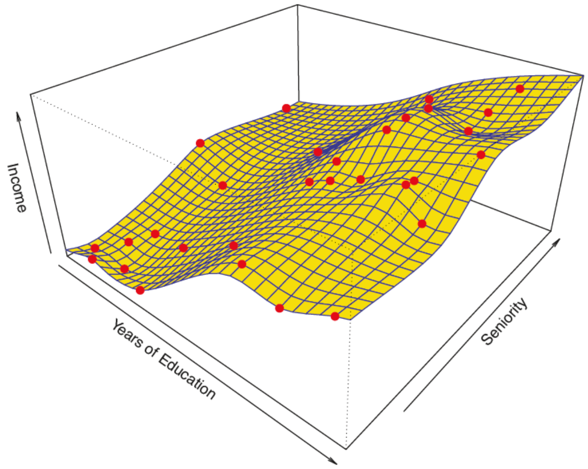


$Y = f(X) + \epsilon$



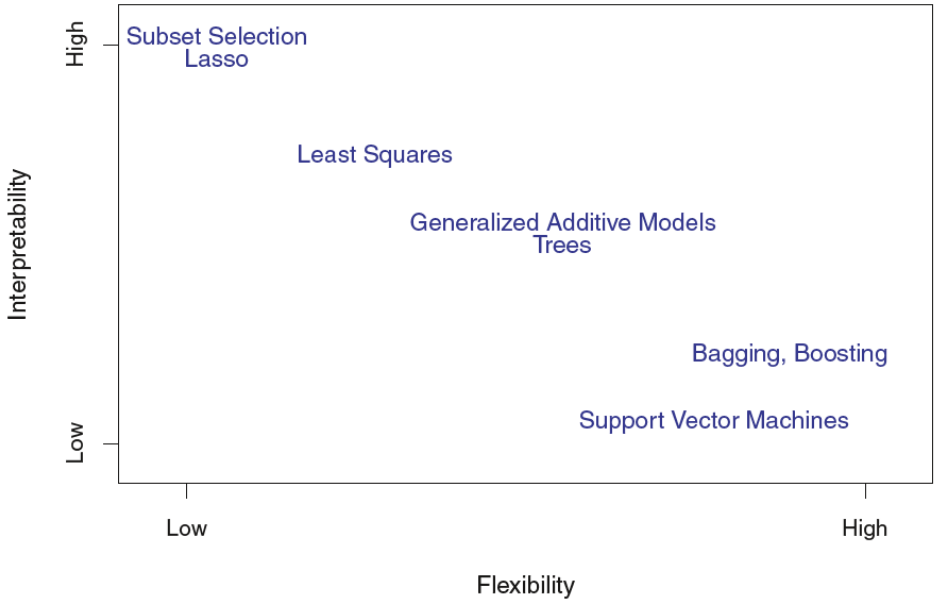
Non-parametric estimation by spline (Ch7)

… and **"overfitting" issue**: the estimate shown below fits the observed data perfectly, and is far more variable than the true function (blue surface above) → Ch5 discusses this issue in detail



Non-parametric estimation by spline with a lower level of smoothness

4. **Prediction Accuracy vs Model Interpretability**



Why would we ever choose to use a more restrictive method instead of a very flexible approach?

- **Much more i____________**
- **Less prone to the o__________ issue**

5. **Supervised vs Unsupervised Learning**

Discussed in Class 01; Read Section 2.1.4. for details.

6. **Regression vs Classification**

Discussed in Class 01; Read Section 2.1.5. for details.

Important note: “We tend to select statistical learning methods on the basis of whether the r__________ is quantitative or qualitative; i.e. we might use linear regression when quantitative and logistic regression when qualitative. However, whether p________ are qualitative or quantitative is generally considered less important. Most of the statistical learning methods discussed in this book can be applied regardless of the predictor variable type, provided that any qualitative predictors are properly *coded* before the analysis is performed. This is discussed in Chapter 3.”