STK2100

Riccardo De Bin

debin@math.uio.no

- what is statistical learning (data mining)
- linear models
- variable transformations

Data Mining

Machine Learning

Data Science

Artificial Intelligence

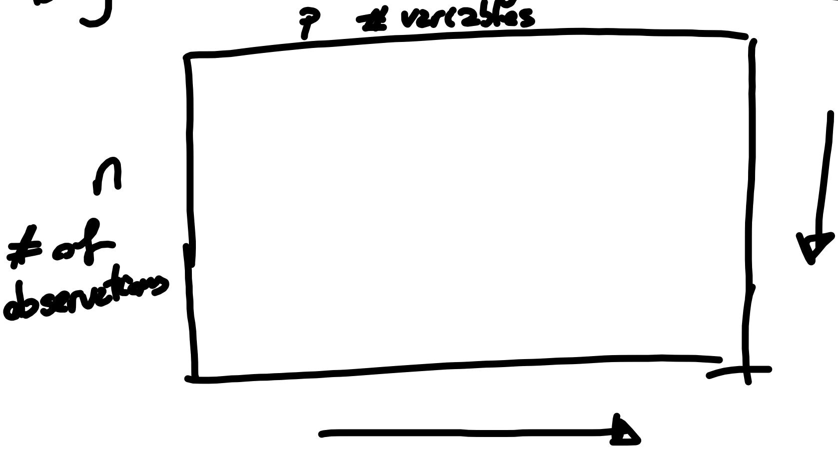
Statistics

Statistical Learning

- black-box vs interpretable model
- prediction vs explanation

Huge amount of data: examples
- receipts
- credit card
- telephone companies
- web
- genetic data
- academic
- fraud (class imbalance)

Big Data vs High-dimensional data



Big Data

- statistics
- programming skills
- database
  - memory issues
  - slow computation

$p >> n$ ⇒ high-dimensional data

- complex situation (traditional tools stop working)
- memory issues
- methodological issues
  - curse of dimensionality
  - $p>>n$ problems

What is a model?

"All models are wrong, but some are useful" (G.E.P. Box)

$$y = \underbrace{f(x)}_{\text{model}} + \varepsilon \rightarrow \text{error (everything we cannot explain)}$$

systematic part

Azzalini & Scarpa (2012)'s definition:

'A model is a simplified representation of the phenomenon of interest, functional for a specific objective'

- simplified representation
  - portability / usability
    - focus on the important aspect
    - eliminate everything not essential
  - degree of complexity
    - trade off (e.g. bias/variance trade-off)

simplicity

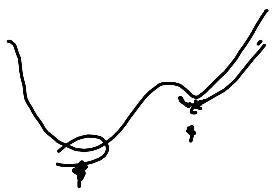
- functional for a specific objective
  - same data, we can have different models based on the goal (description/explanation vs prediction)

Is there a true model?

Experimental studies vs observational studies

Pressing a button / Against

- understand the theory behind → what are strenghts and weaknesses of possible approaches
- interpret the results → good analyst behind
- reliability → find a global minimum and not a local minimum / to be sure that we find



Software

R

→ CRAN → Manuals

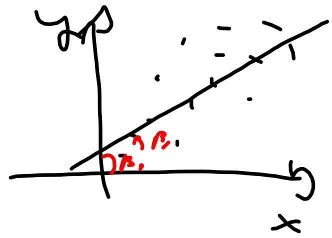
→ RStudio

→ Google's R Style Guide

Basic: linear model

simplest way to relate two variables

simple linear regression model



$$y = \underbrace{f(x;\beta)} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$$

y = response, dependent variable, outcome

x = independent variable, input, predictor, explanatory variable

β_0, β_1 = regression parameters / coeffients

\# regr. coeff = 2 → β_0 = intercept (always)

β_1 = slope

ε = error $\quad E[\varepsilon_i] = 0$

$Var(\varepsilon_i) = \sigma^2$ homoskedasticity

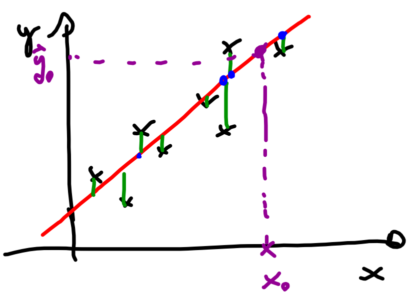
$Cor(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

GOAL: estimate the values of β_0, β_1 using the information in the data (x_i, y_i), $i = 1, \dots, n$ → sample size

HOW: by minimizing a loss function (objective function)

most often the sum of squares (square loss)

$$\hat\beta = \underset{\beta}{\operatorname{argmin}} \Big\{ \underbrace{\sum_{i=1}^{n} (y_i - f(x_i, \beta))^2}_{RSS(\beta) \,\circ\, D(\beta)} \Big\} \qquad \| y - f(x;\beta) \|^2$$



$\hat y_i = f(x_i; \hat\beta) = \hat\beta_0 + \hat\beta_1 x_i$ fitted value

$\hat y_0 = f(x_0; \hat\beta) = \hat\beta_0 + \hat\beta_1 x_0$ predicted value

$f(x;\beta)$ is not a line $\longrightarrow$ polynomial form

$$f(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_{p-1} x^{p-1}$$

β is a p-dimensional vector

$f(x,\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

the model is linear in the parameters:
- conceptually and mathematically simple
- easy to compute

$$f(x;\beta) = \underset{(n\times p)}{X}\,\underset{(p\times 1)}{\beta}$$

X is a matrix $X = (1, x, \dots, x^{p-1})$ X is also called design matrix

In broad generality, the linear model has form

$$y = X\beta + \varepsilon$$

Minimizing the sum of squares

$$D(\beta) = \sum_{i=1}^{n} (y_i - x_i^T\beta) = (y - X\beta)^T(y - X\beta)$$

$$\frac{\partial D(\beta)}{\partial \beta} = X^T(y - X\beta)$$

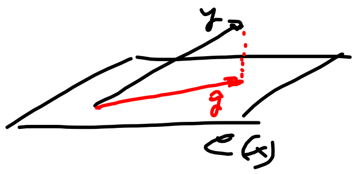
$$\frac{\partial D(\beta)}{\partial \beta} = 0 \qquad X^Ty - X^TX\beta = 0$$

$$X^TX\beta = X^Ty$$

$$\hat\beta = (X^TX)^{-1}X^Ty$$ LEAST SQUARES ESTIMATOR

$$\hat y = X\hat\beta = \underbrace{X(X^TX)^{-1}X^T}_{\text{HAT MATRIX}}y$$

HAT MATRIX $\leftarrow$ PROJECTION MATRIX

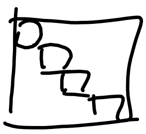


space spanned by the columns of X

$$D(\hat\beta) = \|y - \hat y\|^2$$

We can use the deviance to estimate σ^2: $s^2 = \frac{D(\hat\beta)}{n-p}$ ($\hat\sigma^2$)

and the variance of $\hat\beta$ $\widehat{Var}(\hat\beta) = s^2(X^TX)^{-1}$ $\to$ $p\times p$ matrix in which the diagonal terms (j,j) are the variance of $\hat\beta_j$



Include information about fuel type
(a new variable that can help explaining the variability of y)

create a dummy variable $I_A = \begin{cases} 0 & \text{if fuel type = "gas"} \\ 1 & \text{" " " = "diesel"} \end{cases}$

simplest way to include the variable $\rightarrow$ additive effect

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta I_A + \varepsilon$$

in matrix form

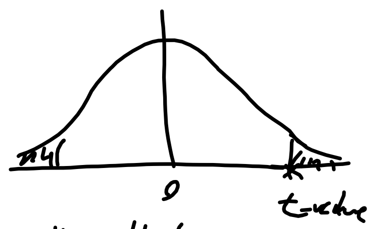
$$y = X\beta + \varepsilon$$

where

$$X = (1, x, x^2, x^3, I_A)$$
$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$$

Assuming $\varepsilon \sim N(0; \sigma^2) \rightarrow$ linear Gaussian regression

t-values: $\dfrac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} = \dfrac{\hat{\beta}_j}{\sqrt{s^2 (X^T X)_{[j,j]}}}$



p-values = Pr(obtaining a t-value larger in absolute value than that we actually obtained under the null hypothesis $\beta_j = 0$)