

STK 2100

Riccardo De Bin

office 815 NHA
debin@math.uio.no

Tuesday exercises

Wednesday 13:15 - 14:00

Thursday 19:45 - 12:00

Content of the course:

- Introduction
 - what is a model and linear regression
- ~~Linear~~ regression
- overoptimism, overfitting, bias-variance trade-off
- data split training/validation/test
 - cross-validation, bootstrapping, ...
- penalised regression (lasso, ridge)
- other regression techniques

Central in this course: data

- increasingly amount of recorded data
- health
- economy
- image recognition
- genetic data

} Big Data
n of observations is large n
very high

high-dimensional data

curse of dimensionality

$p \gg n$

Normally in this course we deal with observational study, the data are collected from the real world, are not collected after a planned experiment.

What is a model?

$$\underline{y} = \underline{f(x)} + \varepsilon$$

y : response variable / outcome / ~~independent~~ variable
 x : inputs / independent variables / predictors / covariates
 ε : error term

"A model is a simplified representation of the phenomenon of interest, functional for a specific objective"
 (Azzolini & Scorre)

- simplified representation
 - usability / portability
 - remove what is not necessary
- functional for a specific objective

"All models are wrong but some are useful" (Box)

→ same data, interested in the same phenomenon, but different models

(x, y)

$f(x)$

$\hat{f}(x)$ $\hat{f}(x)$ $\hat{f}(x)$

E.g. models for prediction vs. models for description

$$y_e \leftrightarrow \hat{f}(x_e)$$

$$\hat{f}(x) \leftrightarrow f(x)$$

$$\underline{E(\hat{f}(x) - f(x))}$$

$$\textcircled{y} = f(x) + \varepsilon$$

$$E[Y] = f(x)$$

continuous $\rightarrow$ regression

discrete/categorical $\rightarrow$ classification

} supervised learning

Linear regression

$$y = f(x) + \varepsilon$$

$\hookrightarrow$ parametric models $f(x) = f(x; \beta)$

$$f(x; \beta) = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

• car example

203 cars in circulation in US in 1985

27 characteristics

y = distance with 2 liter of fuel

$$\text{distance} = \beta_0 + \beta_1 \text{engine size} + \beta_2 \text{type of fuel} + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$\beta_0, \beta_1, \beta_2$ = regression coefficients

$$\text{distance} = \beta_0 + \beta_1 \text{engine size} + \varepsilon$$

β_0 = intercept

β_1 = slope

ε = error term

$$E[\varepsilon_i] = 0$$

$$\text{Var}(\varepsilon_i) = \sigma^2$$

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

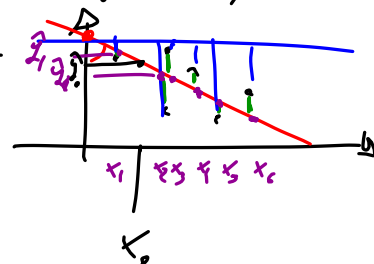
$$i, j = 1, \dots, n$$

CPAL: estimate the values of β_0 and β_1 using the information in the data

How: minimize a loss function (objective function)

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - f(x_i; \beta))^2$$

$RSS(\beta)$ or $J(\beta)$



Once we have found $\hat{\beta}$, we define:

fitted values: $\hat{y}_i = f(x_i; \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

predicted values: $\hat{y}_0 = f(x_0; \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 x_0$ $\leftarrow$ for a new value of x

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

