UE DS52 : Foundations of Data Science
Introduction

Serge Iovleff

UTBM

3 mars 2022

« All models are wrong, but some are useful »

Georges Box

« Prediction can be very difficult, especially about the future. »

Niels Bohre

Plan

1. Objectives and content
2. Examples (from “The elements of statistical learning”)
3. What are Data ?
4. Fundamentals of Statistics
   - Maximum Likelihood
   - The Multivariate Gaussian Model
   - Linear Regression

Course Language

... will be in English

Why?

- Essentially *all* machine learning literature is in English.
- Knowing the proper *terminology* is essential!
- Good to improve your English skills!
- Questions and answers in emails/homework/exams may be answered in French (However, this is not encouraged...).

Issues addressed, Part 1: Fundamentals

- Examine and understand the (characteristics of the) data $\Rightarrow$: Transformations, correlation analysis, visualizations (transversal)
- Statistical inference: Maximum likelihood, Estimators, Bias-Variance, Fisher Information [+ Consistency, Efficiency and TCL]
- Optimization techniques in the context of statistical/Machine learning $\Rightarrow$ Gradient, Newton, Quasi-Newton optimization methods, EM (?) with applications

Issues addressed, Part 2: Techniques

- Unsupervised learning ⇒ automatic classification, density estimation, dimension reduction, vector quantization
- Supervised learning ⇒ Linear Regression, Ridge, Lasso, decision trees, forest trees, Kernel methods, Boosting, Bagging, K-NN, Support Vector Machines, logistic regression, Optimal Margin Classifier...
- Machine Learning ⇒ multi-layer perceptrons, GAN, other methods (including) [20h with M. Séna APEKE]

Prerequisites

- Basic probability theory
- Mathematics: basic knowledge of linear algebra, derivation, integration
- Computer science: basic programming skills
- Basic knowledge of python programming

Background reading

- Standard background reading:
  - C.M. Bishop, *Pattern Recognition and Machine Learning* (2006), Springer
  - T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (2015), Springer Verlag, `https://web.stanford.edu/~hastie/Papers/ESLII.pdf`
  - K.P. Murphy, *Machine Learning: a Probabilistic Perspective* (2012), MIT Press
  - S. Rogers, M. Girolami, *A First Course in Machine Learning* (2016), CRC Press
- Mathematics for machine learning background:
  - Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong, *Mathematics for Machine Learning* (2020), Cambridge University Press, `https://mml-book.github.io/`
- Other resources:
  - D. Barber, *Bayesian Reasoning and Machine Learning* (2012), Cambridge University Press, `http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf`
  - R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification* (2nd ed. 2001), Willey-Interscience

Evaluations

- About 1-2 projects/homeworks during the training sessions
- Two computer-based assessments (MCQs) during Tutorial sessions
- Two hours “Final” exam

Plan

1. Objectives and content
2. Examples (from “The elements of statistical learning”)
3. What are Data ?
4. Fundamentals of Statistics
   - Maximum Likelihood
   - The Multivariate Gaussian Model
   - Linear Regression

Example 1: Email Spam

- The data for this example consists of information from 4601 email messages, in a study to try to predict whether the email was `junk` email, or `spam`.

	george	spam	email	you	your	hp	free	hpl	!
`spam`	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.51
`email`	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.11	0.18

Table: Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

- This is a supervised learning problem, with the outcome the class variable `email/spam`. It is also called a **classification** problem.

Example 2: Prostate Cancer

- The data for this example come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures, in 97 men who were about to receive a radical prostatectomy.
- The goal is to predict the log of PSA (`lpsa`) from a number of measure- ments including log cancer volume (`lcavol`), log prostate weight `lweight`, `age`, log of benign prostatic hyperplasia amount `lbph`, seminal vesicle invasion `svi`, log of capsular penetration `lcp`, Gleason score `gleason`, and percent of Gleason scores 4 or 5 `pgg45`.
- Figure in the Notebook is a scatterplot matrix of the variables. Some correlations with `lpsa` are evident, but a good predictive model is difficult to construct by eye.
- This is a supervised learning problem, known as a **regression** problem, because the outcome measurement is quantitative.

Example 3: Handwritten Digit Recognition

- The data from this example come from the handwritten ZIP codes on envelopes from U.S. postal mail. Each image is a segment from a five digit ZIP code, isolating a single digit. The images are 16×16 eight-bit grayscale maps, with each pixel ranging in intensity from 0 to 255.
- Some sample images are shown in the Notebook
- This is a **classification** problem for which the **error rate** needs to be kept very low to avoid misdirection of mail. In order to achieve this low error rate, some objects can be assigned to a “don’t know” category, and sorted instead by hand.

Example 4: DNA Expression Microarrays

- DNA stands for deoxyribonucleic acid, and is the basic material that makes up human chromosomes.
- DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present for that gene.
- Microarrays are considered a breakthrough technology in biology, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells.
- The challenge here is to understand how the genes and samples are organized.
- Typical questions include the following:
  - (a) which samples are most similar to each other, in terms of their expression profiles across genes?
  - (b) which genes are most similar to each other, in terms of their expression profiles across samples?
  - (c) do certain genes show very high (or low) expression for certain cancer samples?
- The problem (a) is an unsupervised learning problem called **clustering**.

Plan

1. Objectives and content
2. Examples (from “The elements of statistical learning”)
3. What are Data ?
4. Fundamentals of Statistics
   - Maximum Likelihood
   - The Multivariate Gaussian Model
   - Linear Regression

According to The Oxford Learner's Dictionary (2021) data are "*facts or information, especially when examined and used to find out things or to make decisions.*"

Data, Information, Knowledge In common usage, **data** is a collection of discrete or continuous values that convey **information**, describing the quantity, quality, fact, statistics, other basic units of **meaning**, or simply sequences of symbols that may be further **interpreted** formally.

A **datum** is an individual value in a collection of data.

All Is Data Data are collected, and thus those data that are collected are limited by what the scientist apprehends. For this reason, one must remember that the data are **incomplete**.

Data Are Socially Constructed for a Purpose Social construction is about how relationships and frames of reference give rise to the meanings of things that happen in daily life. The social construction of data means that it also includes the scientist's reflections on the data.

One individual, many characteristics

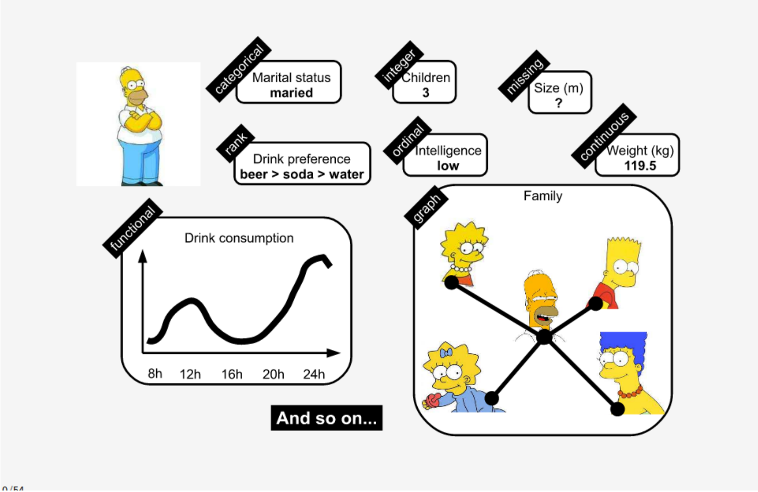

Figure: A representative heterogeneous datum about a representative individual

Tables

- Data refer to a *set* of observations, with each **observation** (sometimes referred to as an **individual** or **entity**) groups together the values taken by the **variables** (or **features**, or **characteristics**) involved in the study.
- Sets of observations are generally represented by data **tables**, with each row corresponding to an observation (an individual, a datum) and each column (or group of columns, for a nominal variable) to a variable:

	X_1	X_2	$\dots$	X_p
I_1	x_{11}	x_{12}	$\dots$	x_{1p}
I_2	x_{21}	x_{22}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_n	x_{n1}	x_{n2}	$\dots$	x_{np}

Table: Standard table for a data set. Rows are referred as **Observations**, *n* is the number of observations. Columns are referred as **Variables** (or **Series**).

Coding ordinal variables

- Representing the values of qualitative variables is not as straightforward.
- Indeed, while some modeling methods can directly employ a symbolic representation of these values, others require numerical representations.
- Let's consider an ordinal variable, such as a Lickert scale. Is it possible to represent its different values by numbers that respect the same order relationship, as in the following table?

Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly agree
1	2	3	4	5

- This representation introduces identical distances between successive modalities, while only the order relationship is associated with the meaning of the variable.
- Is disagreement at equal distance from strongly disagree and neither disagree nor agree?

Coding ordinal variables

- An often-preferred representation is to associate as many binary numerical variables with this ordinal variable as there are different modalities, using the following binary coding:

Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly agree
00001	00011	00111	01111	11111

- This representation preserves the This representation preserves the order relationship between modalities, while avoiding imposing too strong a constraint on the representation of successive modalities, since a subsequent modality is associated with the passage of a new binary variable to 1.
- The disadvantage of this representation is that it increases the size of the observations.

Coding categorical variables

- For a categorical variable, its direct representation by a quantitative variable using ordinal coding, as in the following table, is strongly discouraged.

Teacher	Doctor	Technician	...
1	2	3	...

Table: **Wrong** way to encode categorical variables

- Such a representation introduces an order between modalities, an order that is **absent** from the meaning of the variable.
- For example, a prediction that is uncertain between `Teacher` and `Technician` will be assimilated to `Doctor`.
- The preferred representation for a categorical variable uses disjunctive (or one-hot) coding

Teacher	Doctor	Technician	...
$100\dots000$	$010\dots000$	$001\dots000$	...

- This representation allows the different modalities to be separated, but also has the disadvantage of increasing the size of the observations.

Plan

1. Objectives and content
2. Examples (from “The elements of statistical learning”)
3. What are Data ?
4. Fundamentals of Statistics
   - Maximum Likelihood
   - The Multivariate Gaussian Model
   - Linear Regression

Plan

1. Objectives and content
2. Examples (from “The elements of statistical learning”)
3. What are Data ?
4. Fundamentals of Statistics
   - Maximum Likelihood
   - The Multivariate Gaussian Model
   - Linear Regression

Reminders on Maximum Likelihood (ML)

The main elements of a maximum likelihood estimation problem are the following:

- A sample
$$\mathcal{D}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$$
that we use to make statements about the probability distribution that generated the sample;
- The sample $\mathcal{D}_n$ is regarded as the **realization** of a random vector $\mathbf{X}_n$, whose distribution is unknown and needs to be estimated;
- there is a set $\Theta \subset \mathbb{R}^d$ of real vectors (called the parameter space) whose elements (called parameters) are put into correspondence with the possible distributions of $\mathbf{X}_n$;
- The joint probability mass function (if $\mathbf{X}$ discrete) or the joint probability density (if $\mathbf{X}$ continuous) is a function of $\boldsymbol{\theta}$ for fixed $\mathcal{D}_n$. It is called likelihood and it is denoted by
$$L(\mathcal{D}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta})$$

Reminders on Maximum Likelihood (2)

- The estimator of $\boldsymbol{\theta}$'s maximum likelihood is the one that **maximize** (minimize) the (negative) log-likelihood of the observations

$$l(\boldsymbol{\theta}|\mathbf{x}_1,\dots,\mathbf{x}_n) \stackrel{\text{def}}{=} l_n(\boldsymbol{\theta}) = (-)\sum_{i=1}^{n} \log\left(f(\mathbf{x}_i;\boldsymbol{\theta})\right).$$

- The estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is a solution of the equation (multidimensional)

$$\frac{\partial}{\partial\boldsymbol{\theta}} l_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}) = \mathbf{0}.$$

- Under conditions of regularity of the likelihood, the ML estimator $\hat{\boldsymbol{\theta}}_n$ has good properties (asymptotically unbiased, Gaussian, convergent, etc.). More on this later.

Exercices

1 For each of the distributions below, what is the parameter $\boldsymbol{\theta}$?

1. Binomial $\mathcal{B}(n; p)$?
2. Poisson $\mathcal{P}(\lambda)$?
3. Uniform $\mathcal{U}(a, b)$?
4. Exponential $\mathcal{E}(\lambda)$?
5. Normal $\mathcal{N}(\mu; \sigma^2)$?

2 Consider a sample of n iid random variables $X_1, \dots, X_n$.

- X_i was drawn from distribution $F = \mathrm{Ber}(\theta)$ with unknown parameter θ.
- Observed sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- How likely is this sample if, say, $\theta = 0.4$?
- Answer

$$\mathbb{P}\left(\text{sample} \,|\theta = 0.4\right) = (0.4)^8 (0.6)^2 = 0.000236$$

- Is there a better choice for θ?

Plan

1 Objectives and content

2 Examples (from “The elements of statistical learning”)

3 What are Data ?

4 Fundamentals of Statistics
- Maximum Likelihood
- The Multivariate Gaussian Model
- Linear Regression

Example: The Gaussian multivariate distribution

Definition

The density probability distribution (pdf) of a multivariate Gaussian random vector is

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Theorem (ML of the multivariate Gaussian)

Let $\mathbf{X}$ *be a random variable with probability law* $\mathcal{N}(\boldsymbol{\mu};\boldsymbol{\Sigma})$. *And let* $(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n)$ *be a sample drawn from this distribution. The maximum likelihood estimators for* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *are*

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\boldsymbol{\mu}})(\mathbf{x}_i-\hat{\mu})^T.\end{aligned}$$

Writing the log-likelihood

By the independence of the random vectors, the joint density of the data $\mathbf{x}_i,\ i = 1, 2, \ldots, n$ is the product of the individual densities, that is

$$
\begin{aligned}
l_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log \prod_{i=1}^{n} f(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) \\
&= \sum_{i=1}^{n} \left(-\frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_i - \mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)
\end{aligned}
$$

Deriving $\hat{\boldsymbol{\mu}}$

To take the derivative with respect to $\boldsymbol{\mu}$ and equate to zero we will make use of the following matrix calculus identity:

$$\frac{\partial}{\partial \mathbf{w}} \mathbf{w}^{\mathbf{T}} \mathbf{A} \mathbf{w} = \mathbf{2} \mathbf{A} \mathbf{w}, \quad \text{if } \mathbf{w} \text{ does not depend on } \mathbf{A} \text{ and } \mathbf{A} \text{ is } \textit{symmetric}.$$

$$\frac{\partial}{\partial \mu} l_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = 0$$

Since $\boldsymbol{\Sigma}$ is positive definite

$$0 = n\boldsymbol{\mu} - \sum_{i=1}^{n} \mathbf{x}_i$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \bar{\mathbf{x}}$$

which is (obviously) the sample mean vector.

Deriving $\hat{\mathbf{\Sigma}}$ (1)

Deriving the MLE for the covariance matrix requires more work and the use of the following linear algebra and calculus properties:

- The trace is invariant under cyclic permutations of matrix products: $\operatorname{tr}[ABC] = \operatorname{tr}[CAB] = \operatorname{tr}[BCA]$
- Since $x^T Ax$ is scalar, we can take its trace and obtain the same value: $x^T Ax = \operatorname{tr}\left[x^T Ax\right] = \operatorname{tr}\left[xx^T A\right]$
- $\frac{\partial}{\partial A}\operatorname{tr}[AB] = B^T$
- $\frac{\partial}{\partial A}\log|A| = (A^{-1})^T = (A^T)^{-1}$
- The determinant of the inverse of an invertible matrix is the inverse of the determinant: $|A| = \frac{1}{|A^{-1}|}$

Combining these properties allows us to calculate

$$\frac{\partial}{\partial A}x^T Ax = \frac{\partial}{\partial A}\operatorname{tr}\left[xx^T A\right] = [xx^T]^T = \left(x^T\right)^T x^T = xx^T$$

Which is the **outer produc**t of the vector x with itself.

Deriving $\hat{\boldsymbol{\Sigma}}$ (2)

We can now re-write the log-likelihood function and compute the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$ (note C is constant)

$$
\begin{aligned}
l_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathrm{C} - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \\
&= \mathrm{C} + \frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\right]
\end{aligned}
$$

and thus

$$
\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{n}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \text{ Since } \boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}
$$

Equating to zero and solving for $\boldsymbol{\Sigma}$

$$
\begin{aligned}
0 &= n\boldsymbol{\Sigma} - \sum_{i=1}^{m}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \\
\hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T
\end{aligned}
$$

Plan

1. Objectives and content

2. Examples (from “The elements of statistical learning”)

3. What are Data ?

4. Fundamentals of Statistics
   - Maximum Likelihood
   - The Multivariate Gaussian Model
   - Linear Regression

The Regression Model

- The objective is to estimate the parameters of the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

  where y_i is the dependent variable, x_i is a $1 \times K$ vector of **regressors**, $\boldsymbol{\beta}$ is the $K \times 1$ vector of regression coefficients to be estimated and $\boldsymbol{\epsilon}_i$ is an unobservable error term.

- The sample is made up of n IID observations

$$(y_i, \mathbf{x}_i), \quad i = 1, \dots, n.$$

- The regression equations can be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where the $n \times 1$ vector of observations of the dependent variable is denoted by $\mathbf{Y}$, the $n \times K$ matrix of **regressors** is denoted by $\mathbf{X}$, and the $N \times 1$ vector of error terms is denoted by $\boldsymbol{\epsilon}$.

Assumptions

- We assume that the vector of errors $\boldsymbol{\epsilon}$ has a multivariate normal distribution conditional on $\mathbf{X}$, with mean equal to 0 and covariance matrix equal to

$$\sigma^2 I_n$$

where I_n is the $n \times n$ identity matrix and

$$\sigma^2 = \mathrm{Var}\left(\epsilon_i \,|\mathbf{X}\right)$$

is the second parameter to be estimated.

- Furthermore, it is assumed that the matrix of regressors X has **full-rank**.

Implications of the assumptions

- The assumption that the covariance matrix of epsilon is diagonal implies that the entries of $\boldsymbol{\epsilon}$ are mutually independent (i.e., $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\epsilon}_j$ for $i \neq j$).
- Moreover, they all have a normal distribution with mean 0 and variance σ^2.
- By the properties of **linear transformations of normal random variables**, the dependent variable y_i is conditionally normal, with mean $\mathbf{x}_i^T\boldsymbol{\beta}$ and variance σ^2.
- Therefore, its conditional probability density function is

$$f(y_i|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{\sigma^2}\right)$$

- The likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{Y}, \mathbf{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right)$$

The ML estimators

- The log-likelihood function is

$$l_n(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

- The ML estimators of the regression coefficients and of the variance of the error terms are

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2 \end{aligned}$$

- Note that $\hat{\boldsymbol{\beta}}$ does not depend on $\hat{\sigma}^2$, so that this is an explicit solution.