CentraleSupélec

Statistics and Learning

Lecturers (alphabetic order):

Julien Bect, Gilles Faÿ, Ziad Kobeissi, Laurent Le Brusquet, Vincent Lescarret, Arshak Minasyan, Arthur Tenenhaus† & Xujia Zhu

† Course coordinator

Lecture 6/9

Introduction to supervised learning
Linear models for regression

Course objectives

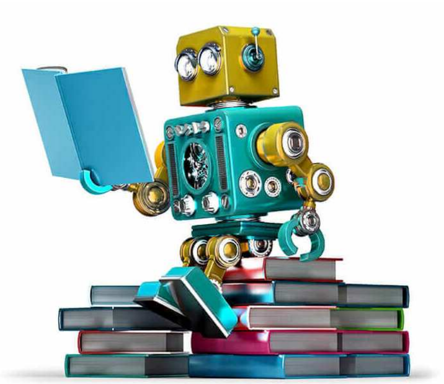
- Introduce the basic concepts of statistical learning
- Establish the mathematical framework for regression and classification problems
- Learn how to build and use linear regression models

Lecture outline

Lecture outline

Lecture outline

Machine learning (*apprentissage automatique*)



One possible definition...

"*Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience*." → data !

(P. Langley and H. A. Simon (1995). Comm. of the ACM, 38(11):54–64)

Image: J. Walsh (2016). Machine Learning: The Speed-of-Light Evolution of AI and Design. https://www.autodesk.com/redshift/machine-learning/

Statistical learning: a "disciplinary" point of view

Computer science → Artificial intelligence (AI) → Machine learning → **Statistical learning** (*apprentissage statistique*)

Mathematics ("mathematical sciences") → Statistics → **Statistical learning** (*apprentissage statistique*)

Remark: in practice, "machine learning" (*apprentissage automatique*) and "statistical learning" (*apprentissage statistique*) are often used interchangeably.

Example: handwritten character recognition

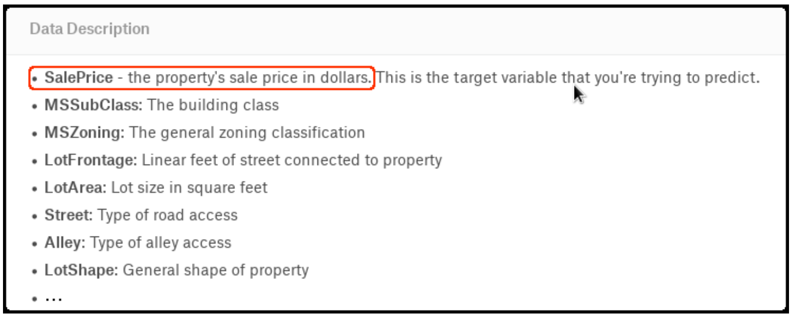
A subset of the MNIST database
containing 70 000 b&w images† of size 28 × 28 pixels

Supervised learning problems: examples are provided with a label.

➠ Learn to classify a new image in one of the 10 classes.

† 60 000 training examples and 10 000 test examples
Source: `https://www.openml.org/search?type=data&id=554`

Example: real estate pricing in Ames (Iowa)



Data Description

- **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
- **MSSubClass:** The building class
- **MSZoning:** The general zoning classification
- **LotFrontage:** Linear feet of street connected to property
- **LotArea:** Lot size in square feet
- **Street:** Type of road access
- **Alley:** Type of alley access
- **LotShape:** General shape of property
- ...

Database of real estate transactions data
(sales price + 79 attributes; 1460 transactions)

Supervised learning problem: here, the price plays the role of a label.

➠ Learn to predict the price of a house from its 79 attributes.

Source: Kaggle competition "House Prices: Advanced Regression Techniques"
(https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

Several forms of learning

- **Supervised** learning: examples with labels.
  - analogy: learning with a teacher.

➠ Lectures 6 to 8

- Unsupervised learning: examples without labels
  - analogy: learning without a teacher, pattern discovery

➠ Lecture 9

and also... (not covered in this course)

- Active learning
  - the labels are queried sequentially;
  - example: detection of bank frauds
    → in-depth analysis of "suspicious" cases only.
- Reinforcement learning
- Transfer learning
- ...

Numerous fields of application

- Computer vision
- Speech recognition
- Natural Language Processing (NLP)
- Fraud detection
- Personalized medicine
- Recommender systems & targeted marketing
- ...

Lecture outline

ML vocabulary: instance space and label space

Instance space: $\mathcal{X}$

- instances $x_1, \ldots, x_n \in \mathcal{X}$

Label space: $\mathcal{Y}$

- labels $y_1, \ldots, y_n \in \mathcal{Y}$

MNIST example:

Class: zero, one, ... nine

$\mathcal{X} = [0,1]^{28\times 28}$ $\quad\quad$ $\mathcal{Y} = \{\text{“zero”}, \ldots, \text{“nine”}\}$

In this and the following lectures, we will always assume:

$\mathcal{X} = \mathbb{R}^p$

$\mathcal{Y} = \mathbb{R} \to$ regression, or
$\mathcal{Y} = \{0,1\} \to$ classification†.

† more precisely: *binary* classification. However, binary classification methods can also be useful for "multi-class" problems (such as MNIST)...

Statistical model

The statistical model of supervised learning

i) In supervised learning, we consider an iid n-sample:

$$(X_1, Y_1), \ \ldots, \ (X_n, Y_n) \overset{\text{iid}}{\sim} \mathrm{P}^{X,Y}$$

where $\mathrm{P}^{X,Y}$ is an unknown probability measure on $\mathcal{X} \times \mathcal{Y}$.

ii) Unless explicitely mentioned, we make no assumption on the distribution: $\theta = \mathrm{P}^{X,Y}$ and $\Theta = \{\text{probability measures on } \mathcal{X} \times \mathcal{Y}\}$.

Notation. We denote by (X, Y) another pair of RVs, which follows the same distribution $\mathrm{P}^{X,Y}$ but is not observed.

⚠ change of notation (wrt previous lectures)

➠ observations: $\quad X_i \in \mathcal{X} \quad \to \quad (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

Goal

Goal of supervised learning (informally)

We want to "learn" from data† a prediction function‡

$$\begin{array}{rccc} \hat{h}: & \mathcal{X} & \to & \mathcal{Y} \\ & x & \mapsto & y = \hat{h}(x) \end{array}$$

such that the RVs Y and $\hat{h}(X)$ are as "close" as possible.

† We should write $\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n))\dots$
‡ If $\mathcal{Y}$ is finite, it is also called classification function or "classifier".

To this end, let us consider a loss function:

$$\begin{array}{rccc} L: & \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ & (y, \tilde{y}) & \mapsto & L(y, \tilde{y}). \end{array}$$

➠ $L(y, \hat{h}(x))$ quantifies the loss when y is predicted by $\hat{h}(x)$.

Goal (cont'd)

Definition: risk (generalization error)

Given a loss function L and a prediction function h, the risk, or generalization error, is defined as :

$$R(h) = \mathbb{E}\left(L(Y, h(X))\right),$$

where the expectation is with respect to (X, Y).

(NB: the concept of "risk" in this context differs from that in the previous lectures)

⚠ This risk depends on the unknown distribution $\theta = \mathrm{P}^{X,Y}$:

$$R_\theta(h) = \iint_{\mathcal{X}\times\mathcal{Y}} L(y, h(x)) \, \mathrm{P}^{X,Y}(\mathrm{d}x, \mathrm{d}y).$$

➠ From now on, we will simply write $R(h)$.

Goal (cont'd)

The optimal prediction function depends on the unknown distribution $\mathrm{P}^{X,Y}$:

$$h^* \;=\; h^*(\mathrm{P}^{X,Y}) \;=\; \mathrm{argmin}_h \; R(h).$$

(existence/uniqueness not guaranteed)

Goal of supervised learning

We want to construct, from the data $(X_1, Y_1), \dots, (X_n, Y_n)$, a prediction function

$$\begin{array}{rccc} \hat{h}: & \mathcal{X} & \to & \mathcal{Y} \\ & x & \mapsto & y = \hat{h}(x) \end{array}$$

such that the risk $R(\hat{h})$ is as close as possible to the optimal risk

$$R^* = \inf_h R(h)$$

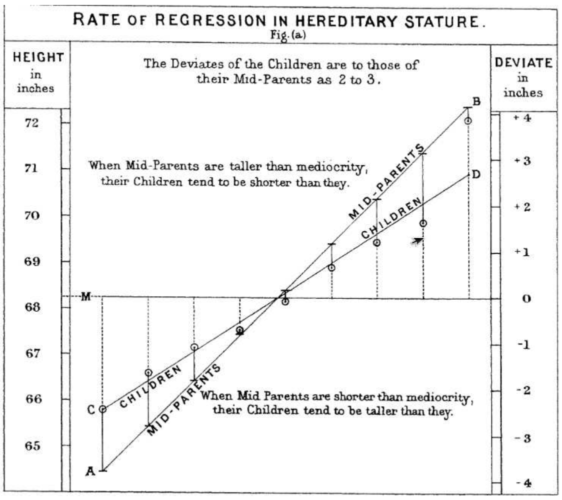
(also called "Bayes risk").

Lecture outline

Lecture outline

Regression

We consider in the rest of this lecture the regression case: $\mathcal{Y} = \mathbb{R}$.



Francis Galton (1886). "Regression Towards Mediocrity in Hereditary Stature", *Journal of the Anthropological Institute*, 15:246–263.

Stat. vocab.: Y = response variable / X = explanatory variables.

Quadratic loss

Consider for a start the quadratic loss:

$$L(y, \tilde{y}) \;=\; (y - \tilde{y})^2.$$

(this is the most commonly used in regression settings)

Proposition

For the quadratic loss, the optimal prediction function is

$$\forall x \in \mathcal{X}, \quad h^*(x) = \mathbb{E}(Y|X = x).$$

Vocabulary : $x \mapsto \mathbb{E}(Y|X = x)$ is sometimes called "regression function".

We will consider this loss function until further notice.

Quadratic loss (cont'd)

Proof. By the law of total expectation, we get:

$$R(h) \;=\; \mathbb{E}\left(\underbrace{\mathbb{E}\left((Y - h(X))^2 \mid X \right)}_{\circledast} \right).$$

Le term $\circledast$ can be decomposed as :

$$\begin{aligned}
&\mathbb{E}\left((Y - h(X))^2 \mid X \right) \\
&\quad = \mathbb{E}\left((Y - \mathbb{E}(Y \mid X) + \mathbb{E}(Y \mid X) - h(X))^2 \mid X \right) \\
&\quad = \mathrm{var}(Y \mid X) + (\mathbb{E}(Y \mid X) - h(X))^2 .
\end{aligned}$$

The first term does not depend on h, and the second one is minimal when $h(X) = \mathbb{E}(Y \mid X)$ a.s.. $\square$

Empirical risk

Recall that the joint distribution $\mathrm{P}^{X,Y}$ is unknown

➠ the risk $R(h)$ cannot be computed.

Definition: empirical risk

We call empirical risk the risk

$$\hat{R}_n(h) \;=\; \iint_{\mathcal{X}\times\mathcal{Y}} L(y,\, h(x))\, \hat{\mathrm{P}}_n(\mathrm{d}x, \mathrm{d}y) \;=\; \frac{1}{n}\sum_{i=1}^{n} L(Y_i,\, h(X_i))$$

associated to the empirical measure $\hat{\mathrm{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$.

With the quadratic loss :

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(X_i))^2 \,.$$

Empirical risk minimization

A general learning method:

1. Choose a family $\mathscr{H}$ of prediction functions.
2. Select the function h which minimizes the empirical risk:

$$\hat{h}^{\mathrm{ERM}} \;=\; \operatorname{argmin}_{h\in\mathscr{H}}\, \hat{R}_n(h).$$

Example: "linear" (affine) prediction functions

$$\mathscr{H} \;=\; \Big\{ h : \mathbb{R}^p \to \mathbb{R} \;\Big|\; \exists \beta \in \mathbb{R}^{p+1},\ \forall x \in \mathcal{X}, \\ h(x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} \Big\}$$

⚠ the ERM method is reasonable if $\mathscr{H}$ is "not too large"

➠ otherwise, complex models must be *penalized* (more on this in Lecture 8)

Other examples of families of prediction functions

- linear models with general basis functions

$$h(x) = \beta_1 h_1(x) + \ldots + \beta_K h_K(x),$$

where the functions $h_k : \mathcal{X} \to \mathbb{R}$ are known;

- additive models

$$h(x) = h_1(x^{(1)}) + \ldots + h_p(x^{(p)}),$$

where the h_k's belong to a given family of $\mathbb{R} \to \mathbb{R}$ functions;

- neural networks,
- decision trees,
- generalized linear/additive models
- ...

Lecture outline

Residual sum of squares

We consider prediction functions h of the form :

$$h(x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} = \beta^\top x$$

$$\text{with } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ and } x = \begin{pmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(p)} \end{pmatrix}.$$

Definition: RSS / least squares criterion

Empirical risk: $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \beta^\top X_i\right)^2$.

We define the Residual Sum of Squares (RSS):

$$\mathrm{RSS}(\beta) = \sum_{i=1}^n \left(Y_i - \beta^\top X_i\right)^2$$

or least squares criterion.

Matrix-vector notations

$$\text{Let } \underline{X} = \begin{pmatrix} 1 & X_1^{(1)} & \ldots & X_1^{(p)} \\ 1 & X_2^{(1)} & \ldots & X_2^{(p)} \\ \vdots & & & \vdots \\ 1 & X_n^{(1)} & \ldots & X_n^{(p)} \end{pmatrix} \text{ and } \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

➠ $\underline{X}$ has size $n \times (p+1)$ and $\underline{Y}$ has length n.

Matrix form of the criterion

$$\begin{aligned} \mathrm{RSS}(\beta) &= \|\underline{Y} - \underline{X}\beta\|^2 \\ &= (\underline{Y} - \underline{X}\beta)^\top (\underline{Y} - \underline{X}\beta) \\ &= \beta^\top \underline{X}^\top \underline{X}\beta - 2\underline{Y}^\top \underline{X}\beta + \underline{Y}^\top \underline{Y} \end{aligned}$$

Minimization of the least squares criterion

Assumption

We assume $\underline{X}^\top \underline{X}$ invertible

➠ implies $p+1 \leq n$.

Let $\tilde{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}$. Then:

$$\begin{aligned} \mathrm{RSS}(\beta) &= \beta^\top \underline{X}^\top \underline{X}\beta - 2\underline{Y}^\top \underline{X}\beta + \underline{Y}^\top \underline{Y} \\ &= (\beta - \tilde{\beta})^\top \underline{X}^\top \underline{X}(\beta - \tilde{\beta}) + \mathrm{c} \end{aligned}$$

where c is a constant (i.e., does not depend on β).

Indeed: $\tilde{\beta}^\top \underline{X}^\top \underline{X}\beta = \underline{Y}^\top \underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{X}\beta = \underline{Y}^\top \underline{X}\beta$.

Minimization of the least squares criterion

Reminder : $\mathrm{RSS}(\beta) = (\beta - \tilde{\beta})^\top \underline{X}^\top \underline{X}(\beta - \tilde{\beta}) + \mathrm{c}$.

We have:

- (i) $\forall a \in \mathbb{R}^{p+1}$, $a^\top \underline{X}^\top \underline{X} a = \|\underline{X} a\|^2 \geq 0$,
- (ii) $\underline{X}^\top \underline{X}$ is invertible, hence positive definite.

(i) implies that $\mathrm{RSS}(\beta)$ is minimal at $\tilde{\beta}$;

(ii) implies that the minimizer is unique ($a^\top \underline{X}^\top \underline{X} a = 0 \implies a = 0$).

Proposition: least squares estimator

When $\underline{X}^\top \underline{X}$ is invertible,

$$\hat{\beta} = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}$$

is the unique minimizer of the RSS function.

➠ complement: matrix calculus

Goodness of fit

Without explanatory variables, we would have

$$\hat{h}(x) = \hat{\beta}_0, \quad \text{with} \quad \hat{\beta}_0 = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Let us set $\text{TSS} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ $\rightarrow$ Total Sum of Squares.

Definition: coefficient of determination R^2

Reminder : $\text{RSS}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i\right)^2$. We set :

$$R^2 = 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}}.$$

Properties.

proof: see exercise 1

- $0 \leq R^2 \leq 1$,
- $R^2 = 1 \iff \forall i,\ Y_i = \hat{\beta} X_i$.

"Ozone" example: presentation of the data

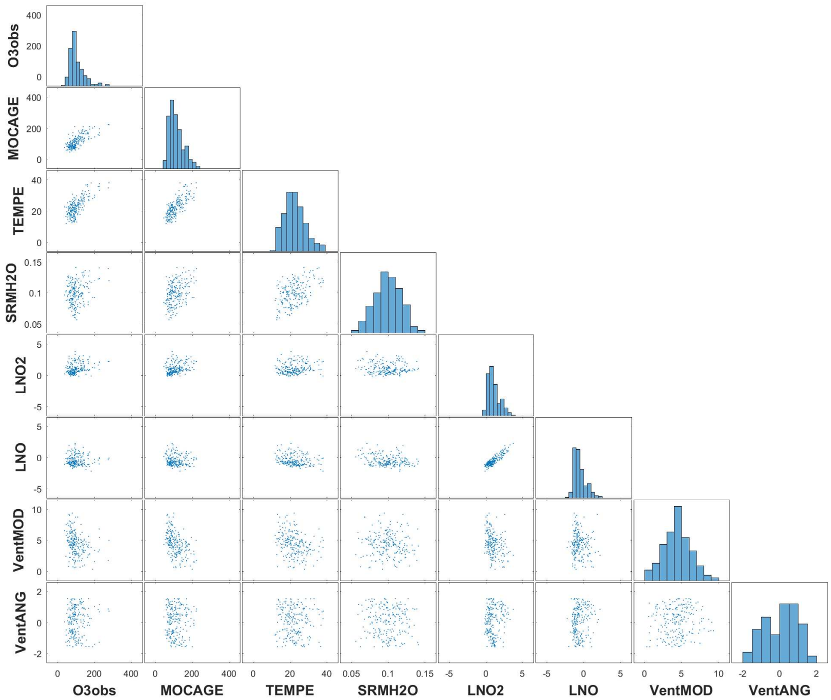
variable	description
O3obs	concentration of ozone on day $t+1$
MOCAGE	pollution prediction obtained by a deterministic computation fluid dynamics (CFD) model
TEMPE	MétéoFrance temperature forecast for day $t+1$
RMH2O	humidity ratio at day t
NO2	nitrogen dioxide concentration on day t
NO	nitrogen monoxide concentration on day t
VentMOD	wind strength on day t
VentANG	wind orientation of day t

Learning task

- predict the ozone concentation on day $t+1$ from data available on day t
- predict if the concentration will exceed 150 μg/m^3 (classification task, cf. lecture #7).

Application and data obtained from https://github.com/wikistat/Apprentissage/tree/master/Pic-ozone

"Ozone" example: data visualization



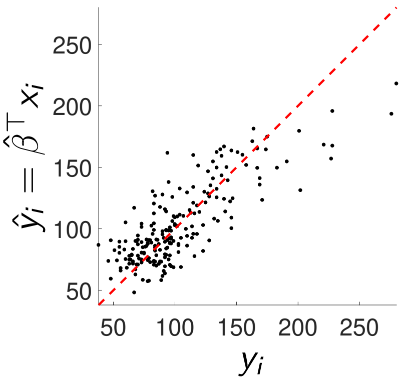
"Ozone" example: linear regression

Linear regression using $n = 210$ days of data.

Remark. All variables standardized for the sake of interpretability.

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

Coefficient of determination. $R^2 = 65.7\%$



Observations:

- the negative coefficient associated to NO2 is surprising (but NO2 is correlated with NO);
- RMH2O, VentMOD and VentANG appear to be of lesser importance;
- the model explains partly the data.

Lecture outline

Properties of the least squares estimator

Recall that, until now: $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathrm{P}^{X,Y}$.

➠ in the section, we assume instead deterministic X_i's

(equivalently, we work "conditionally on the X_i's").

Assume moreover that there exists $\beta \in \mathbb{R}^{p+1}$ such that

(i) $\forall i, \quad Y_i = \beta^\top X_i + \epsilon_i$

where the errors ϵ_i are

(ii) centered: $\mathbb{E}(\epsilon_i) = 0$,

(iii) uncorrelated: $i \neq j \; \Rightarrow \; \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$,

(iv) homoscedastic: $\mathrm{var}(\epsilon_i) = \sigma^2$ for some $\sigma^2 > 0$.

Properties of the least squares estimator

Proposition

Under these assumptions, $\hat{\beta}$ is an unbiased estimator:

$$\mathbb{E}\left(\hat{\beta}\right) \; = \; \beta,$$

and its covariance matrix is:

$$\text{var}\left(\hat{\beta}\right) \; = \; \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1}.$$

Properties of the least squares estimator

Proof.

Recall that the X_i's are assumed deterministic.

Let $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Then:

$$\text{(i)} \quad \Rightarrow \quad \begin{cases} \underline{Y} & = \; \underline{X}\beta + \underline{\epsilon} \\ \hat{\beta} & = \; (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y} = \beta + (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{\epsilon} \end{cases}$$

$$\text{(ii)} \quad \Rightarrow \quad \mathbb{E}\left(\hat{\beta}\right) = \beta + (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \mathbb{E}\left(\underline{\epsilon}\right) = \beta$$

$$\text{(iii)+(iv)} \quad \Rightarrow \quad \begin{aligned} \text{var}\left(\hat{\beta}\right) & = \; (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \; \text{var}\left(\underline{\epsilon}\right) \; \underline{X} \, (\underline{X}^\top \underline{X})^{-1} \\ & = \; \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1} \end{aligned}$$

$\square$

Distribution of $(\hat\beta, \hat\sigma^2)$ under a normality assumption

Assume furthermore that (v) $\underline{\epsilon}$ is Gaussian:

$$\log \mathcal{L}(\beta, \sigma^2; \underline{Y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta^\top X_i\right)^2 .$$

Proposition: MLE of (β, σ^2) (see PC 6)

The MLE is
$$\begin{cases} \hat\beta = \arg\min_\beta \sum_{i=1}^n \left(Y_i - \beta^\top X_i\right)^2, \\ \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat\beta^\top X_i\right)^2 . \end{cases}$$

➠ We recover the least square estimator of β

Student's theorem: distribution of $(\hat\beta, \hat\sigma^2)$ (see PC 6)

- $\hat\beta \sim \mathcal{N}\left(\beta, \sigma^2 \left(\underline{X}^\top \underline{X}\right)^{-1}\right)$,
- $\hat\sigma^2 \sim \frac{\sigma^2}{n}\chi^2(n-p-1)$,
- $\hat\beta$ et $\hat\sigma^2$ are independent.

Tests / CI on the value of a component of β

We know that $\hat\beta_j \sim \mathcal{N}\left(\beta_j, \sigma^2 v_j\right)$ with $v_j = \left[\left(\underline{X}^\top \underline{X}\right)^{-1}\right]_{j,j}$.

Pivotal function

$$T = \frac{\hat\beta_j - \beta_j}{\sqrt{\frac{n\hat\sigma^2 v_j}{n-p-1}}} \sim \mathcal{T}(n-p-1)$$

with $\mathcal{T}(n-p-1)$: Student's t distrib. with $n-p-1$ degrees of freedom

➠ Student's t distribution

Remark:

$$\frac{n\hat\sigma^2}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i - \hat\beta^\top X_i\right)^2$$

is an unbiased estimator of σ^2 (see PC 6).

Proof

It follows from Student's theorem that

- $U = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \sim \mathcal{N}(0,1)$
- $V = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$,
- and U and V are independent.

Thus

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{n\hat{\sigma}^2 v_j}{n-p-1}}} = \frac{U}{\sqrt{\frac{V}{n-p-1}}} \sim \mathcal{T}(n-p-1),$$

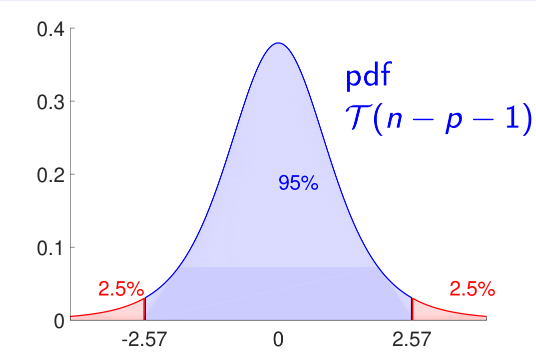
by definition of the Student's t distribution with $k = n-p-1$ degrees of freedom. □

Test for $H_0 : \beta_j = 0$ / $H_1 : \beta_j \neq 0$

Let $0 < \alpha < 1$.

Take $\beta_j = 0$ in the def. of T (i.e., assume H_0) and

$$\delta = \mathbb{1}_{|T| > q_{1-\frac{\alpha}{2}}}$$



Exact confidence interval for β_j

$$\left[\hat{\beta}_j - \sqrt{\frac{n\hat{\sigma}^2 v_j}{n-p-1}}\, q_{1-\frac{\alpha}{2}},\ \hat{\beta}_j + \sqrt{\frac{n\hat{\sigma}^2 v_j}{n-p-1}}\, q_{1-\frac{\alpha}{2}}\right]$$

q_r: quantile of order r of $\mathcal{T}(n-p-1)$

"Ozone" example: CIs and p-values

	$\mathrm{CI}_{95\%}$	t	pval
β_0	$[100.1, 106.7]$	62.9	$< 10^{-6}$
MOCAGE	$[21.1, 36.8]$	7.4	$< 10^{-6}$
TEMPE	$[16.5, 28.5]$	7.6	$< 10^{-6}$
RMH2O	$[-7.0, 0.6]$	-1.7	0.095
NO2	$[-53.0, -15.7]$	-3.7	$< 10^{-3}$
NO	$[19.8, 55.4]$	4.2	$< 10^{-3}$
VentMOD	$[-2.7, 5.4]$	0.7	0.49
VentANG	$[-0.8, 6.0]$	1.6	0.12

with t: realization of T for the corresponding coefficient

Remark: regression without RMH2O, VentMOD et VentANG

➟ the coefficient of determination drops from 65.7% to 64.5%.

Lecture outline

“Ozone” example: data corruption

Assume that 5 out of n measurements of ozone concentration ($n = 210$) are corrupted, i.e., approx. 2% of the sample.

Estimated coefficients without and with corrupted data:

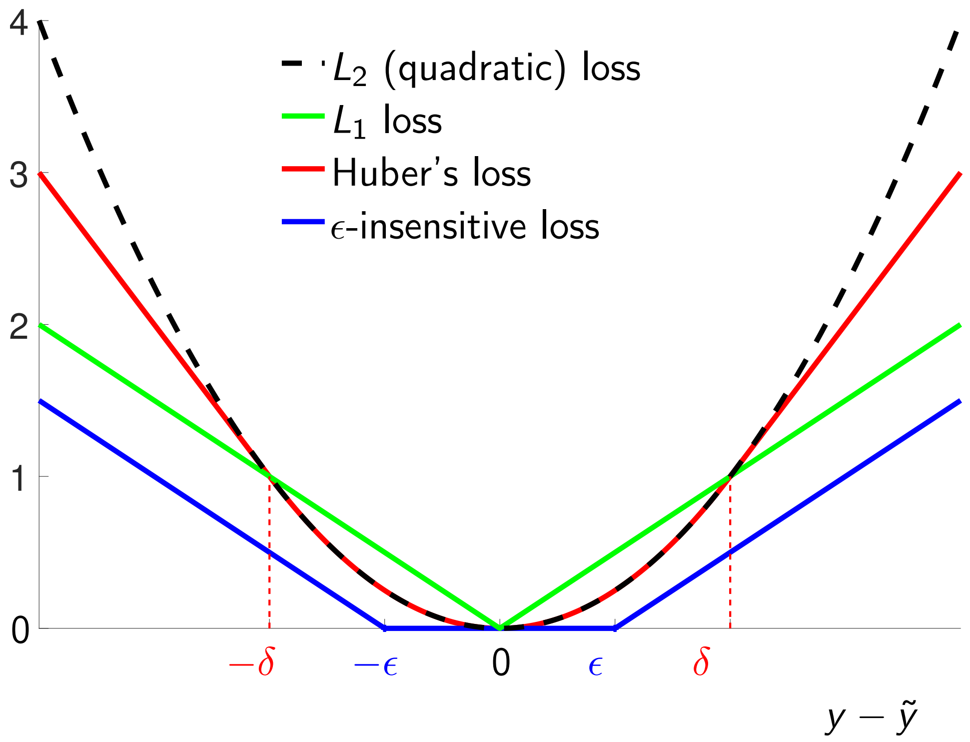
	β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
w/o	103.4	28.9	22.6	-3.2	-34.4	37.6	1.4	2.6
with	125.2	79.2	-15.6	24.2	-155.1	141.4	4.7	24.9

➠ Strong sensitivity of the coefficients to “outliers”.

Solution

Use a loss function that leads to a prediction function with better robustness properties than the quadratic loss.

Usual loss functions



L_1 loss

Loss function : $L(y, \tilde{y}) = |y - \tilde{y}|$.

Proposition (see PC 6)

For the L_1 loss, the optimal prediction function is

$$\forall x \in \mathcal{X}, \quad h^*(x) = \mathrm{med}(Y|X = x)$$

"Ozone" example

	β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
w/o	100.8	27.5	19.2	-3.3	-32.2	33.9	-1.0	3.9
with	101.4	28.3	18.6	-1.6	-35.1	37.5	0.5	3.2

➠ better stability with respect to outliers.

Lecture outline

Limitations of "ordinary least squares"

Recall that $\underline{X}$ has size #individuals $\times$ #variables ($n \times (p+1)$).

Critical cases for "ordinary least squares"

- when $\underline{X}^\top \underline{X}$ not invertible,
- or poorly conditioned.

Typical cases:

- when the number of variables is large
  ($p + 1 > n$, sometimes $p \gg n$)

  Example: genomics.

- when there are strong correlations between explanatory variables

  Example: "ozone" data (cf. variables NO and NO2)
  ➠ lack of interpretability of the coefficients

One possible solution: penalized regression

A penalty term is added to the empirical risk:

$$\hat{\beta} = \arg\min_{\beta} \underbrace{\mathrm{RSS}(\beta)}_{\text{data "fidelity"}} + \underbrace{\lambda}_{\text{hyperparameter}} \underbrace{\Omega(\beta)}_{\text{penalty}} .$$

➠ see Lecture 8

Lecture outline

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

3 – Standard exercices (with solutions)

3.1 – Questions

3.2 – Solutions

4 – Appendices

Exercise 1 (Regression seen as a projection)

solution

Let $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $1 \leq i \leq n$, denote an n-sample of observations.

Consider the linear regression model from slide 21:

$$h(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x^{(j)} = \beta^\top x, \quad x \in \mathbb{R}^{p+1},$$

and the corresponding least squares estimator:

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i \right)^2 .$$

As in slide 22, we denote by

- $\underline{X} \in \mathbb{R}^{n \times (p+1)}$ the matrix of regressors,
- $\underline{Y} \in \mathbb{R}^n$ the vector of responses.

Exercise 1 (Regression seen as a projection)

solution

Questions

1. Set $\underline{\hat{Y}} = \underline{X}\hat{\beta}$. Prove that $\underline{\hat{Y}}$ is the projection of $\underline{Y}$ onto the image of $\underline{X}$.
2. Give the expression of the projection matrix, assuming that $\underline{X}^\top \underline{X}$ is invertible.
3. Prove that the coefficient of determination, defined in slide 25, satisfies the property $0 \leq R^2 \leq 1$, with $R^2 = 1$ iff $\forall i,\ Y_i = \hat{\beta}^\top X_i$.

Lecture outline

Solution of exercise 1

back to questions

❶ Reminders:

- The projection of $y \in \mathbb{R}^n$ onto a closed convex set $C \subset \mathbb{R}^n$ is the unique $y^* \in C$ such that $\|y - y^*\| = \min_{v \in C} \|y - v\|$.
- The image of $\underline{X}$, which we will denote by $\mathrm{Im}(\underline{X})$, is the linear subspace of $\mathbb{R}^n$ generated by the columns of $\underline{X}$:

$$\mathrm{Im}(\underline{X}) = \left\{ v \in \mathbb{R}^n \mid \exists \beta \in \mathbb{R}^{(p+1)},\ v = \underline{X}\beta \right\}.$$

To begin with, note that

- $\mathrm{Im}(\underline{X})$ is indeed a closed convex set (since all linear subspaces are closed in finite dimension),
- $\underline{\hat{Y}} = \underline{X}\hat{\beta}$ belongs to $\mathrm{Im}(\underline{X})$.

Solution of exercise 1

back to questions

Furthermore, for all $v = \underline{X}\beta \in \text{Im}(\underline{X})$, using the fact that

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \ \|\underline{Y} - \underline{X}\beta\|^2,$$

we find that

$$\begin{aligned} \|\underline{Y} - \hat{\underline{Y}}\| &= \|\underline{Y} - \underline{X}\hat{\beta}\| \\ &\leq \|\underline{Y} - \underline{X}\beta\| = \|\underline{Y} - v\|, \end{aligned}$$

therefore $\hat{\underline{Y}}$ is indeed the projection of $\underline{Y}$ onto $\text{Im}(\underline{X})$.

❷ Using the expression of $\hat{\beta}$ established in class, we can write the projection of $\underline{Y}$ onto $\text{Im}(\underline{X})$ as

$$\hat{\underline{Y}} = \underline{X}\hat{\beta} = \underline{X}\left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}.$$

Solution of exercise 1

back to questions

This being true for all $\underline{Y} \in \mathbb{R}^n$, we conclude that the matrix of the projection operator is:

$$P = \underline{X}\left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top.$$

❸ Recall the characterization of the projection onto a linear subspace:

Theorem

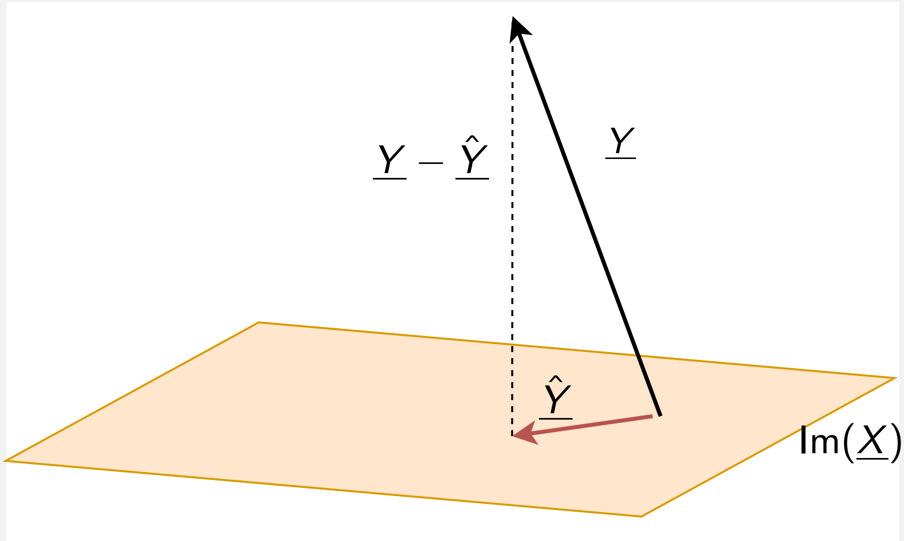
Let $y \in \mathbb{R}^n$ and let F be a linear subspace of $\mathbb{R}^n$. Then, y^* is the projection of y onto F if, and only if,

- $y^* \in F$,
- $y - y^* \in F^\perp$.

Solution of exercise 1

back to questions

We apply the theorem with $F = \text{Im}(\underline{X})$ and $y = \underline{Y}$.



Solution of exercise 1

back to questions

Consider now the coefficient of determination:

$$R^2 = 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}}, \quad \text{where} \begin{cases} \text{TSS} & = \left\| \underline{Y} - \bar{Y} 1_{n\times 1} \right\|^2 \\ \text{RSS}(\beta) & = \| \underline{Y} - \underline{X}\beta \|^2 \end{cases}$$

Let us decompose the TSS:

$$\begin{aligned} \text{TSS} &= \left\| \underline{Y} - \hat{\underline{Y}} + \hat{\underline{Y}} - \bar{Y} 1_{n\times 1} \right\|^2 & (1) \\ &= \left\| \underline{Y} - \hat{\underline{Y}} \right\|^2 + \left\| \hat{\underline{Y}} - \bar{Y} 1_{n\times 1} \right\|^2 & (2) \\ &= \text{RSS}(\hat{\beta}) + \left\| \hat{\underline{Y}} - \bar{Y} 1_{n\times 1} \right\|^2 . \end{aligned}$$

The transition from (1) to (2) follows from Pythgora's theorem.

Solution of exercise 1

back to questions

Indeed,

- $\underline{\hat{Y}} \in \text{Im}(\underline{X})$ and $\underline{Y} - \underline{\hat{Y}} \in \text{Im}(\underline{X})^{\perp}$ since $\underline{\hat{Y}}$ is the projection of $\underline{Y}$ onto the linear subspace $\text{Im}(\underline{X})$.
- $\underline{\hat{Y}} - \bar{Y} 1_{n\times 1} \in \text{Im}(\underline{X})$ since $1_{n\times 1} \in \text{Im}(\underline{X})$.

Thus:

(i) $0 \leq \text{RSS}(\hat{\beta}) \leq SCT$, therefore $0 \leq R^2 \leq 1$,

(ii) $R^2 = 1$ iff $SCR(\hat{\beta}) = 0$ iff $\underline{Y} = \underline{X}\hat{\beta}$.

$\square$

Lecture outline

Matrix calculus

The result can also be found using matrix calculus.

Let $v \in \mathbb{R}^q$, $z \in \mathbb{R}^q$ and $M \in \mathbb{R}^{q\times q}$.

1) differentiation of $h(z) = v^\top z = \sum_{j=1}^q v_j z_j$

$$\nabla_z h(z) = \begin{pmatrix} \frac{\partial h}{\partial z_1} \\ \vdots \\ \frac{\partial h}{\partial z_q} \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_q \end{pmatrix} = v \quad \text{therefore} \quad \nabla_z \left(v^\top z\right) = v.$$

2) differentiation of $h(z) = z^\top M z = \sum_{i,j=1}^p z_i M_{i,j} z_j$

$$\nabla_z h(z) = \begin{pmatrix} \frac{\partial h}{\partial z_1} \\ \vdots \\ \frac{\partial h}{\partial z_q} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^q M_{1,j} z_j + \sum_{i=1}^q M_{i,1} z_i \\ \vdots \\ \sum_{j=1}^q M_{1,j} z_j + \sum_{i=1}^q M_{i,1} z_i \end{pmatrix}$$

$$\text{therefore} \quad \nabla_z \left(z^\top M z\right) = (M + M^\top) z.$$

Matrix calculus (cont'd)

Application to the minimization of the least squares criterion.

Recall that

$$\mathrm{RSS}(\beta) \;=\; \beta^\top \underline{X}^\top \underline{X}\beta \;-\; 2\underline{Y}^\top \underline{X}\beta \;+\; \underline{Y}^\top \underline{Y}$$

Thus we have

$$\nabla_\beta \mathrm{RSS}(\beta) \;=\; 2\underline{X}^\top \underline{X}\beta - 2\underline{X}^\top \underline{Y} \;=\; 2\left(\underline{X}^\top \underline{X}\beta - \underline{X}^\top \underline{Y}\right),$$

and finally:

$$\nabla_\beta \mathrm{RSS}(\hat\beta) = 0 \quad \Longrightarrow \quad \hat\beta = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{Y}.$$

$\square$

back to slide ??

Data standardization

Let $\underline{X} = (X_1, \ldots, X_n)$ be an n-sample taking values in $\mathbb{R}^p$.

Data standardization consists in transforming $\underline{X}$ to $\underline{\widetilde{X}}$ as follows:

$$\widetilde{X}_i^{(j)} = \frac{X_i^{(j)} - \overline{X}_n^{(j)}}{S_n^{(j)}}, \qquad 1 \leq i \leq n, \quad 1 \leq j \leq p,$$

where $\overline{X}_n^{(j)}$ and $S_n^{(j)}$ are the sample average and standard deviation of the j-th variable, respectively:

$$\overline{X}_n^{(j)} = \frac{1}{n} \sum_{i=1}^{n} X_i^{(j)},$$

$$\left(S_n^{(j)}\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i^{(j)} - \overline{X}_n^{(j)}\right)^2.$$

back to slide 28

Student's t distribution $\mathcal{T}(k)$

Definition of $\mathcal{T}(k)$, k integer ≥ 1

Let U and V be two RVs such that

- $U \sim \mathcal{N}(0, 1)$
- $V \sim \chi^2(k)$
- U and V are independent

then $T = \frac{U}{\sqrt{\frac{V}{k}}}$ follows a Student's t distribution with k degrees of freedom.

Properties

$\mathcal{T}(k) \xrightarrow[k \to \infty]{d} \mathcal{N}(0, 1)$

Exercise : prove it.

Probability density function

$$f(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

Mean

- for $k \geq 2$, $\mathbb{E}_k(T) = 0$

Variance

- for $k \geq 3$, $\text{var}_k(T) = \frac{k}{k-2}$

back to slide ??