

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

1/84

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

2/84

Lecture outline

- 1 – Introduction to (supervised) statistical learning
- 2 – Linear regression
- 3 – Standard exercises (with solutions)
- 4 – Appendices

3/84

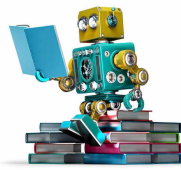
Lecture outline

- 1 – Introduction to (supervised) statistical learning
 - 1.1 – Statistical learning
 - 1.2 – The mathematical framework of supervised learning
- 2 – Linear regression
- 3 – Standard exercises (with solutions)
- 4 – Appendices

Lecture outline

- 1 – Introduction to (supervised) statistical learning
 - 1.1 – Statistical learning
 - 1.2 – The mathematical framework of supervised learning
- 2 – Linear regression
- 3 – Standard exercises (with solutions)
- 4 – Appendices

Machine learning (*apprentissage automatique*)



One possible definition...

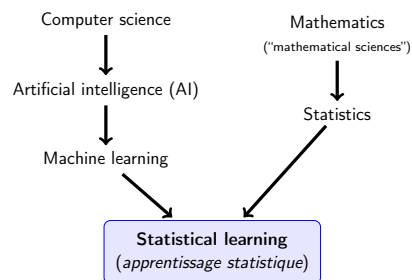
"Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge *from experience*." → data !

(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/>

4/54

Statistical learning: a "disciplinary" point of view



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

5/54

Example: handwritten character recognition



A subset of the MNIST database
containing 70 000 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>

6/54

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>)

7/84

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

▶ ...

8/84

Numerous fields of application

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

9/84

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

3 – Standard exercises (with solutions)

4 – Appendices

ML vocabulary: instance space and label space

Instance space: \mathcal{X}

► instances $x_1, \dots, x_n \in \mathcal{X}$

Label space: \mathcal{Y}

► labels $y_1, \dots, y_n \in \mathcal{Y}$

MNIST example:



Class: zero, one, ... nine

$$\mathcal{X} = [0, 1]^{28 \times 28}$$

$$\mathcal{Y} = \{\text{"zero"}, \dots, \text{"nine"}\}$$

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

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

$$\mathcal{Y} = \mathbb{R} \rightarrow \text{regression, or}$$

$$\mathcal{Y} = \{0, 1\} \rightarrow \text{classification}^\dagger.$$

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

10/84

Statistical model

The statistical model of supervised learning

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

$$(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X, Y}$$

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

ii) Unless explicitly mentioned, we make **no assumption on the distribution**: $\theta = P^{X, Y}$ 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 $P^{X, Y}$** but is **not observed**.

⚠ change of notation (wrt previous lectures)

$$\Rightarrow \text{observations: } X_i \in \mathcal{X} \rightarrow (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$$

11/84

Goal

Goal of supervised learning (informally)

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

$$\begin{aligned} \hat{h} : \mathcal{X} &\rightarrow \mathcal{Y} \\ x &\mapsto y = \hat{h}(x) \end{aligned}$$

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

† We should write $\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n))$.

‡ If \mathcal{Y} is finite, it is also called **classification function** or "classifier".

To this end, let us consider a **loss function**:

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

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

12/84

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 :

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

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 $\theta = P^{X, Y}$** :

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

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

13/84

Goal (cont'd)

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

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

(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**

$$\begin{aligned} \hat{h}: \mathcal{X} &\rightarrow \mathcal{Y} \\ x &\mapsto y = \hat{h}(x) \end{aligned}$$

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").

14/84

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

- 2.1 – Introduction to regression models
- 2.2 – Linear model / quadratic loss
- 2.3 – Back to statistical inference
- 2.4 – Other loss functions
- 2.5 – Limitations of "ordinary least squares"

3 – Standard exercises (with solutions)

4 – Appendices

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

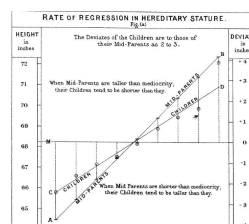
- 2.1 – Introduction to regression models
- 2.2 – Linear model / quadratic loss
- 2.3 – Back to statistical inference
- 2.4 – Other loss functions
- 2.5 – Limitations of "ordinary least squares"

3 – Standard exercises (with solutions)

4 – Appendices

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.

15/84

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**.

16/24

Quadratic loss (cont'd)

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

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

Le term \circledast can be decomposed as :

$$\begin{aligned} & \mathbb{E}((Y - h(X))^2 | X) \\ &= \mathbb{E}((Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - h(X))^2 | X) \\ &= \text{var}(Y | X) + (\mathbb{E}(Y | X) - h(X))^2. \end{aligned}$$

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

17/24

Empirical risk

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

\Rightarrow the risk $R(h)$ cannot be computed.

Definition: empirical risk

We call **empirical risk** the risk

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

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

With the quadratic loss :

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

18/24

Empirical risk minimization


A general learning method:

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

$$\hat{h}^{\text{ERM}} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_n(h).$$

Example: "linear" (affine) prediction functions

$$\mathcal{H} = \left\{ h : \mathbb{R}^p \rightarrow \mathbb{R} \mid \exists \beta \in \mathbb{R}^{p+1}, \forall x \in \mathcal{X}, \right. \\ \left. h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)} \right\}$$

 the ERM method is reasonable if \mathcal{H} is "not too large"
 \Rightarrow otherwise, complex models must be *penalized* (more on this in Lecture 8)

19/24

Other examples of families of prediction functions

- ▶ **linear models** with general basis functions

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

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

- ▶ **additive models**

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

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

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

20/84

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of "ordinary least squares"

3 – Standard exercises (with solutions)

4 – Appendices

Residual sum of squares

We consider prediction functions h of the form :

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

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

Definition: RSS / least squares criterion

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

We define the **Residual Sum of Squares (RSS)**:

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

or **least squares criterion**.

21/84

Matrix-vector notations

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

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

Matrix form of the criterion

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

22/84

Minimization of the least squares criterion

Assumption

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

⇒ implies $p + 1 \leq n$.

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

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

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

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

23/54

Minimization of the least squares criterion

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

We have:

❶ $\forall a \in \mathbb{R}^{p+1}, a^T \underline{X}^T \underline{X} a = \|\underline{X}a\|^2 \geq 0$,

❷ $\underline{X}^T \underline{X}$ is invertible, hence positive definite.

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

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

Proposition: least squares estimator

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

$$\hat{\beta} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$$

is the unique minimizer of the RSS function.

→ complement: matrix calculus

24/54

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

Definition: coefficient of determination R^2

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

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

Properties.

→ proof: see exercise 1

► $0 \leq R^2 \leq 1$,

► $R^2 = 1 \iff \forall i, Y_i = \hat{\beta}^T X_i$.

25/54

"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

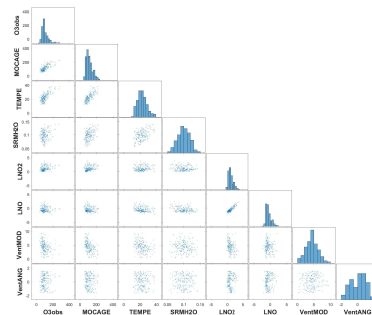
Learning task

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

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

26/54

"Ozone" example: data visualization



27/84

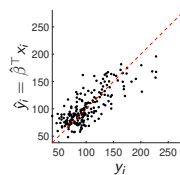
"Ozone" example: linear regression

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

Remark. All variables [standardized](#) for the sake of interpretability.

β_0	MOCAQE	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 partially the response variable (O3obs).

28/84

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of "ordinary least squares"

3 – Standard exercises (with solutions)

4 – Appendices

Properties of the least squares estimator

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

⇒ 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

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

where the errors ϵ_i are

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

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

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

29/84

Properties of the least squares estimator

Proposition

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

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

and its **covariance matrix** is:

$$\text{var}(\hat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

30/34

Properties of the least squares estimator

Proof.

Recall that the X_i 's are assumed deterministic.

Let $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$. Then:

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

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

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

□

31/34

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

Assume furthermore that **(v) ϵ is Gaussian**:

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

Proposition: MLE of (β, σ^2)

(see PC 6)

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

⇒ We recover the least squares estimator of β

Student's theorem: distribution of $(\hat{\beta}, \hat{\sigma}^2)$

(see PC 6)

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

32/34

Tests / CI on the value of a component β_j of β

We know that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j)$ with $v_j = [(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}$.

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:

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

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

33/34

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. \square

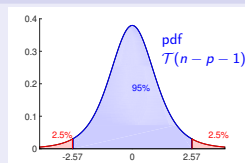
34/54

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 $\mathcal{T}(n - p - 1)$

35/54

"Ozone" example: CIs and p-values

	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%.

36/54

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

- 2.1 – Introduction to regression models
- 2.2 – Linear model / quadratic loss
- 2.3 – Back to statistical inference
- 2.4 – Other loss functions
- 2.5 – Limitations of "ordinary least squares"

3 – Standard exercises (with solutions)

4 – Appendices

“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

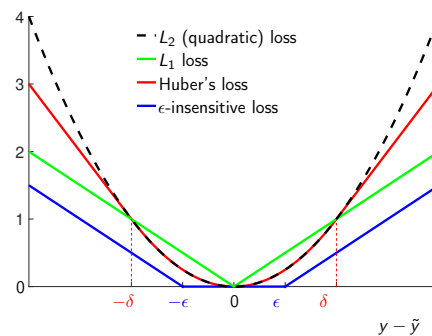
⇒ Strong sensitivity of the coefficients to “outliers”.

Solution

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

37/54

Usual loss functions



38/54

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

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

“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.

39/54

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

3 – Standard exercises (with solutions)

4 – Appendices

Limitations of “ordinary least squares”

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

Critical cases for “ordinary least squares”

- ▶ when $\underline{X}^T \underline{X}$ 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

40/54

One possible solution: penalized regression

A **penalty** term is added to the empirical risk:

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

⇒ see Lecture 8

41/54

Summary and preview

We have seen and will practice in PC 6:

- ▶ the mathematical framework for regression (and classification),
- ▶ the development and application of linear regression models.
- ▶ the properties of the least squares estimator.

We will cover in Lecture 7:

- ▶ performance metrics for classifiers,
- ▶ the development and application of logistic regression,
- ▶ tree-based models and neural networks.

42/54

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

3 – Standard exercises (with solutions)

3.1 – Questions

3.2 – Solutions

4 – Appendices

Lecture outline

- 1 – Introduction to (supervised) statistical learning
- 2 – Linear regression
- 3 – Standard exercises (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:

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

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.

43/54

Exercise 1 (Regression seen as a projection)

[→ solution](#)

Questions

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

44/54

Lecture outline

- 1 – Introduction to (supervised) statistical learning
- 2 – Linear regression
- 3 – Standard exercises (with solutions)
 - 3.1 – Questions
 - 3.2 – Solutions
- 4 – Appendices

Solution of exercise 1

[back to questions](#)

1 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 $\text{Im}(\underline{X})$, is the linear subspace of \mathbb{R}^n generated by the columns of \underline{X} :

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

To begin with, note that

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

46/54

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

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

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

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

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

46/54

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:

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

3 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,

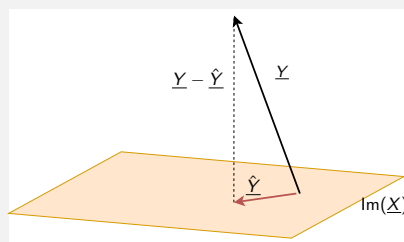
- ▶ $y^* \in F$,
- ▶ $y - y^* \in F^\perp$.

47/54

Solution of exercise 1

[back to questions](#)

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



48/54

Solution of exercise 1

[back to questions](#)

Consider now the coefficient of determination:

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

Let us decompose the TSS:

$$\text{TSS} = \|\underline{Y} - \hat{\underline{Y}} + \hat{\underline{Y}} - \bar{Y}\mathbf{1}_{n \times 1}\|^2 \quad (1)$$

$$= \|\underline{Y} - \hat{\underline{Y}}\|^2 + \|\hat{\underline{Y}} - \bar{Y}\mathbf{1}_{n \times 1}\|^2 \quad (2)$$

$$= \text{RSS}(\hat{\beta}) + \|\hat{\underline{Y}} - \bar{Y}\mathbf{1}_{n \times 1}\|^2.$$

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

49/54

Solution of exercise 1

[back to questions](#)

Indeed,

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

► $\hat{\underline{Y}} - \bar{Y}\mathbf{1}_{n \times 1} \in \text{Im}(\underline{X})$ since $\mathbf{1}_{n \times 1} \in \text{Im}(\underline{X})$.

Thus:

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

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

□

50/54

Lecture outline

1 – Introduction to (supervised) statistical learning

2 – Linear regression

3 – Standard exercises (with solutions)

4 – Appendices

Matrix calculus

The result can also be found using matrix calculus.

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

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

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

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

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

therefore $\nabla_{\mathbf{z}} (\mathbf{z}^\top \mathbf{M} \mathbf{z}) = (\mathbf{M} + \mathbf{M}^\top) \mathbf{z}$.

51/54

Matrix calculus (cont'd)

Application to the minimization of the least squares criterion.

Recall that

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

Thus we have

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

and finally:

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

□

➡ back to slide ??

82/84

Data standardization

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

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

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

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

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

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

➡ back to slide 20

83/84

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

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

Exercise : prove it.

Mean

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

Probability density function

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

Variance

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

➡ back to slide ??

84/84