CentraleSupélec

# Statistique et apprentissage


Chargés de cours (ordre alphabétique) :

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

[†] Coordinateur du cours

Cours 6/9

# Introduction à l'apprentissage supervisé
# Modèles linéaires pour la régression

## Objectifs du cours 6

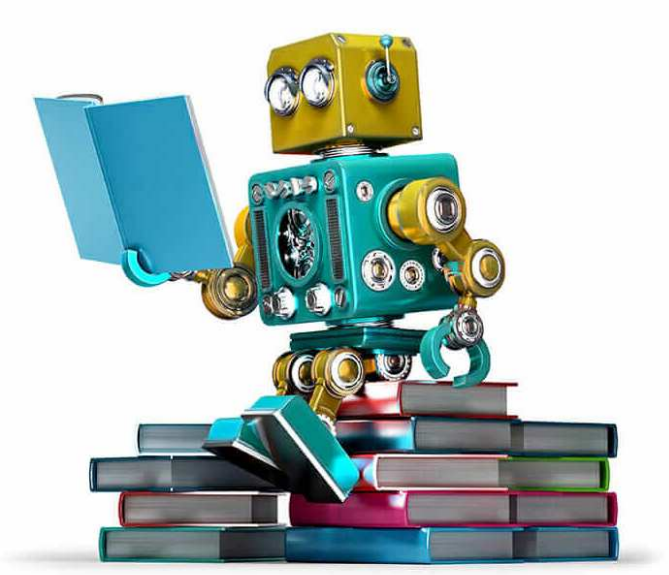
- Présenter les principes de base de l'apprentissage statistique
- Poser le cadre mathématique de la régression et de la classification
- Savoir construire et utiliser des modèles de régression linéaire

# Plan du cours

# Plan du cours

# Plan du cours

# Apprentissage automatique (*machine learning*)



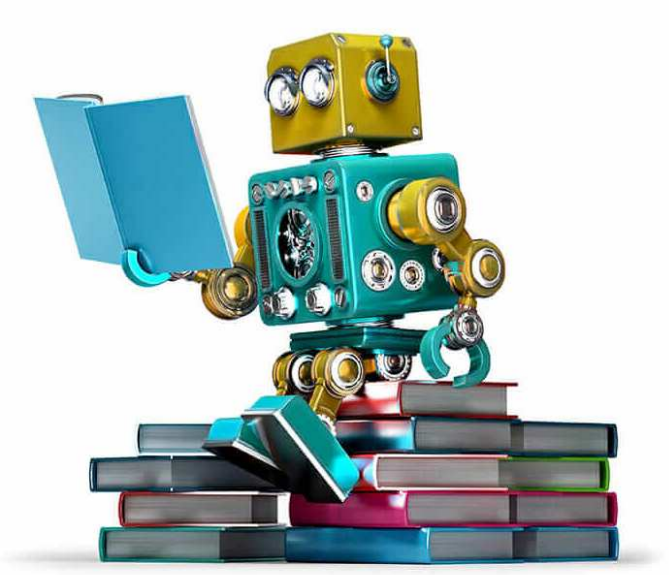
## Une définition possible...

*« Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience. »*

(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/

# Apprentissage automatique (*machine learning*)



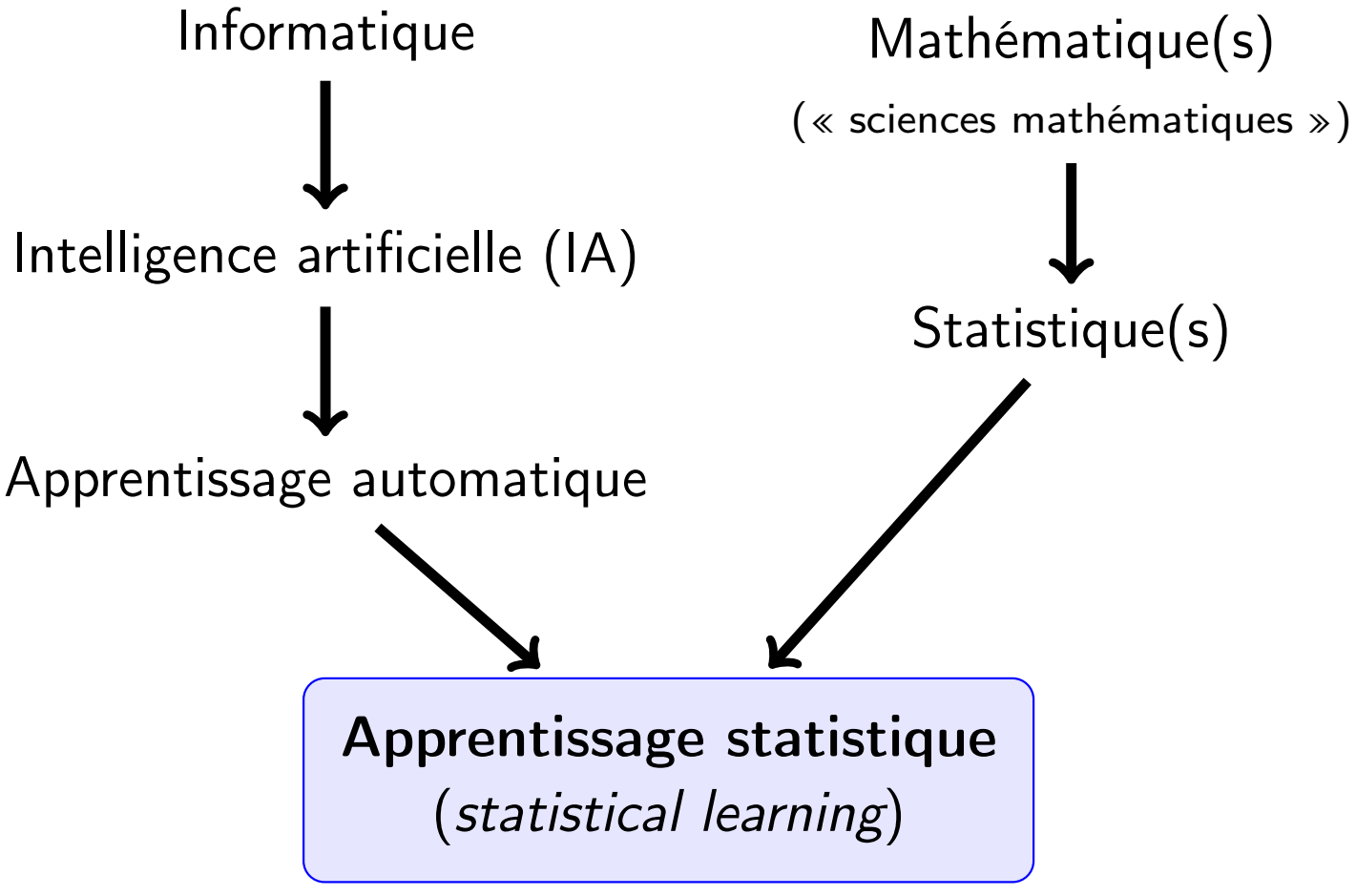
## Une définition possible...

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

(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/

# Apprentissage statistique : une vision « disciplinaire »




Remarque : en pratique, « apprentissage automatique » (*machine learning*) et « apprentissage statistique » (*statistical learning*) sont très souvent utilisés de manière interchangeable.

# Exemple : reconnaissance de caractères manuscrits

Sous-ensemble de la base de données MNIST
contenant 70 000 images[†] de $28 \times 28$ pixels

Problème d'apprentissage supervisé : exemples fournis avec une étiquette.

➠ Apprendre à classer une nouvelle image dans l'une des 10 classes.

[†] 60 000 pour l'apprentissage et 10 000 pour les tests
**Source** : `https://www.openml.org/search?type=data&id=554`

# Exemple : reconnaissance de caractères manuscrits

Sous-ensemble de la base de données MNIST
contenant 70 000 images[†] de $28 \times 28$ pixels

Problème d'apprentissage supervisé : exemples fournis avec une étiquette.

➠ Apprendre à classer une nouvelle image dans l'une des 10 classes.

[†] 60 000 pour l'apprentissage et 10 000 pour les tests
Source : `https://www.openml.org/search?type=data&id=554`

# Exemple : évaluation de biens immobiliers à 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
- ...

Base de données de transactions immobilières
(prix de vente + 79 attributs ; 1460 transactions)

Prob. d'apprentissage supervisé : ici le prix joue le rôle d'une étiquette.

➠ Apprendre à prédire le prix d'un bien à partir des 79 attributs.

Source : compétition Kaggle « House Prices : Advanced Regression Techniques »
(`https://www.kaggle.com/c/house-prices-advanced-regression-techniques`)

# Exemple : évaluation de biens immobiliers à 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
- ...

Base de données de transactions immobilières
(prix de vente + 79 attributs ; 1460 transactions)

Prob. d'apprentissage supervisé : ici le prix joue le rôle d'une étiquette.

➠ Apprendre à prédire le prix d'un bien à partir des 79 attributs.

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

# Différentes formes d'apprentissage

- Apprentissage supervisé : exemples avec des étiquettes
  - analogie : apprentissage avec l'aide d'un enseignant.

➠ cours 6 à 8

- Apprentissage non supervisé : exemples sans étiquette
  - analogie : apprentissage sans enseignant, découverte.

➠ cours 9

et aussi. . . (hors programme)

- Apprentissage actif
  - les étiquettes sont obtenues séquentiellement, sur demande ;
  - exemple : détection de fraudes bancaires → analyse en profondeur des dossiers « suspects » uniquement.
- Apprentissage par renforcement
- Apprentissage par transfert
- . . .

# Différentes formes d'apprentissage

- Apprentissage supervisé : exemples avec des étiquettes
  - analogie : apprentissage avec l'aide d'un enseignant.

➠ cours 6 à 8

- Apprentissage non supervisé : exemples sans étiquette
  - analogie : apprentissage sans enseignant, découverte.

➠ cours 9

et aussi. . . (hors programme)

- Apprentissage actif
  - les étiquettes sont obtenues séquentiellement, sur demande ;
  - exemple : détection de fraudes bancaires → analyse en profondeur des dossiers « suspects » uniquement.
- Apprentissage par renforcement
- Apprentissage par transfert
- . . .

# Différentes formes d'apprentissage

- Apprentissage supervisé : exemples avec des étiquettes
  - analogie : apprentissage avec l'aide d'un enseignant.

➠ cours 6 à 8

- Apprentissage non supervisé : exemples sans étiquette
  - analogie : apprentissage sans enseignant, découverte.

➠ cours 9

et aussi... (hors programme)

- Apprentissage actif
  - les étiquettes sont obtenues séquentiellement, sur demande ;
  - exemple : détection de fraudes bancaires $\rightarrow$ analyse en profondeur des dossiers « suspects » uniquement.
- Apprentissage par renforcement
- Apprentissage par transfert
- ...

# Nombreux domaines d'application

- Vision par ordinateur
- Reconnaissance de la parole
- Traitement des langues naturelles
- Détection de fraude
- Médecine personnalisée
- Systèmes de recommandation & marketing ciblé
- ...

# Plan du cours

# Espace des exemples et espace des étiquettes

Espace des exemples : $\mathcal{X}$

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

Espace des étiquettes : $\mathcal{Y}$

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

Exemple MNIST :

Classe : zéro, un, ... neuf

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

$\mathcal{Y} = \{\text{« zéro »}, \ldots, \text{« neuf »}\}$

Dans la suite du cours on supposera toujours :

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

$\mathcal{Y} = \mathbb{R} \rightarrow$ régression, ou
$\mathcal{Y} = \{0,1\} \rightarrow$ classification[†].

[†] plus précisément : classification *binaire*. Cependant, les méthodes de classification binaire peuvent également être utiles dans le cadre de problèmes « multi-classes » (par ex. MNIST)...

# Espace des exemples et espace des étiquettes

Espace des exemples : $\mathcal{X}$

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

Espace des étiquettes : $\mathcal{Y}$

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

Exemple MNIST :

Classe : zéro, un, ... neuf

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

$\mathcal{Y} = \{\text{« zéro »}, \ldots, \text{« neuf »}\}$

Dans la suite du cours on supposera toujours :

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

$\mathcal{Y} = \mathbb{R} \rightarrow$ régression, ou
$\mathcal{Y} = \{0,1\} \rightarrow$ classification[†].

[†] plus précisément : classification *binaire*. Cependant, les méthodes de classification binaire peuvent également être utiles dans le cadre de problèmes « multi-classes » (par ex. MNIST)...

# Espace des exemples et espace des étiquettes

Espace des exemples : $\mathcal{X}$

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

Espace des étiquettes : $\mathcal{Y}$

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

Exemple MNIST :

Classe : zéro, un, . . . neuf

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

$\mathcal{Y} = \{\text{« zéro »}, \ldots, \text{« neuf »}\}$

Dans la suite du cours on supposera toujours :

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

$\mathcal{Y} = \mathbb{R} \to$ régression, ou
$\mathcal{Y} = \{0,1\} \to$ classification[†].

[†] plus précisément : classification *binaire*. Cependant, les méthodes de classification binaire peuvent également être utiles dans le cadre de problèmes « multi-classes » (par ex. MNIST). . .

# Espace des exemples et espace des étiquettes

Espace des exemples : $\mathcal{X}$

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

Espace des étiquettes : $\mathcal{Y}$

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

Exemple MNIST :

Classe : zéro, un, . . . neuf

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

$$\mathcal{Y} = \{\text{« zéro »}, \ldots, \text{« neuf »}\}$$

Dans la suite du cours on supposera toujours :

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

$\mathcal{Y} = \mathbb{R} \to$ régression, ou
$\mathcal{Y} = \{0,1\} \to$ classification[†].

[†] plus précisément : classification *binaire*. Cependant, les méthodes de classification binaire peuvent également être utiles dans le cadre de problèmes « multi-classes » (par ex. MNIST). . .

# Modèle statistique

## Modèle statistique de l'apprentissage supervisé

i) En apprentissage supervisé on considère un *n*-échantillon iid :

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

avec $\mathrm{P}^{X,Y}$ une mesure de probabilité sur $\mathcal{X} \times \mathcal{Y}$, inconnue.

ii) Sauf mention explicite du contraire, on ne fera pas d'hypothèse sur la loi : $\theta = \mathrm{P}^{X,Y}$ et $\Theta = \{\text{mesures de probabilité sur } \mathcal{X} \times \mathcal{Y}\}$.

**Notation.** On notera $(X, Y)$ un autre couple de loi de VA, suivant la même loi $\mathrm{P}^{X,Y}$ mais non observé.

⚠ au changement de notation (par rapport aux cours précédents)

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

# Modèle statistique

## Modèle statistique de l'apprentissage supervisé

i) En apprentissage supervisé on considère un $n$-échantillon iid :

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

avec $\mathrm{P}^{X,Y}$ une mesure de probabilité sur $\mathcal{X} \times \mathcal{Y}$, inconnue.

ii) Sauf mention explicite du contraire, on ne fera pas d'hypothèse sur la loi : $\theta = \mathrm{P}^{X,Y}$ et $\Theta = \{\text{mesures de probabilité sur } \mathcal{X} \times \mathcal{Y}\}$.

**Notation.** On notera $(X, Y)$ un autre couple de loi de VA, suivant la même loi $\mathrm{P}^{X,Y}$ mais non observé.

⚠ au changement de notation (par rapport aux cours précédents)

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

# Modèle statistique

## Modèle statistique de l'apprentissage supervisé

i) En apprentissage supervisé on considère un $n$-échantillon iid :

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

avec $\mathrm{P}^{X,Y}$ une mesure de probabilité sur $\mathcal{X} \times \mathcal{Y}$, inconnue.

ii) Sauf mention explicite du contraire, on ne fera pas d'hypothèse sur la loi : $\theta = \mathrm{P}^{X,Y}$ et $\Theta = \{\text{mesures de probabilité sur } \mathcal{X} \times \mathcal{Y}\}$.

**Notation.** On notera $(X, Y)$ un autre couple de loi de VA, suivant la même loi $\mathrm{P}^{X,Y}$ mais non observé.

⚠ au changement de notation (par rapport aux cours précédents)

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

# Objectif

## Objectif (informel) de l'apprentissage supervisé

On veut "apprendre" des données[†] une fonction de prédiction[‡]

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

telle que les VA $Y$ et $\hat{h}(X)$ soient aussi « proches » que possible.

[†] On devrait donc écrire $\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \ldots, (X_n, Y_n))\ldots$

[‡] Si $\mathcal{Y}$ est fini, on parle plutôt de fonction de classification ou « classifieur ».

Pour cela, on se donne une fonction de perte :

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

- $L(y, \hat{h}(x))$ traduit la perte lorsque l'on prédit $y$ par $\hat{h}(x)$.

# Objectif

## Objectif (informel) de l'apprentissage supervisé

On veut "apprendre" des données[†] une fonction de prédiction[‡]

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

telle que les VA $Y$ et $\hat{h}(X)$ soient aussi « proches » que possible.

[†] On devrait donc écrire $\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \ldots, (X_n, Y_n))\ldots$

[‡] Si $\mathcal{Y}$ est fini, on parle plutôt de fonction de classification ou « classifieur ».

Pour cela, on se donne une fonction de perte :

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

➠ $L(y, \hat{h}(x))$ traduit la perte lorsque l'on prédit $y$ par $\hat{h}(x)$.

# Objectif (suite)

## Définition : risque (erreur de généralisation)

Etant donnée une fonction de perte $L$ et une fonction de prédiction $h$, on définit le risque, ou erreur de généralisation :

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

l'espérance portant sur le couple $(X, Y)$.

(NB : le terme « risque » a ici un sens différent des cours précédents)

⚠ Ce risque dépend de la loi $\theta = \mathrm{P}^{X,Y}$ inconnue :

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

➠ Dans la suite on notera simplement $R(h)$.

# Objectif (suite)

## Définition : risque (erreur de généralisation)

Etant donnée une fonction de perte $L$ et une fonction de prédiction $h$, on définit le risque, ou erreur de généralisation :

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

l'espérance portant sur le couple $(X, Y)$.

(NB : le terme « risque » a ici un sens différent des cours précédents)

⚠ Ce risque dépend de la loi $\theta = \mathrm{P}^{X,Y}$ inconnue :

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

➠ Dans la suite on notera simplement $R(h)$.

# Objectif (suite)

La fonction de prédiction optimale dépend de la loi $\mathrm{P}^{X,Y}$ inconnue :

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

(existence/unicité non garanties)

## Objectif de l'apprentissage supervisé

On veut construire, à partir des données $(X_1, Y_1), \ldots, (X_n, Y_n)$ une fonction de prédiction

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

telle que le risque $R(\hat{h})$ soit aussi proche que possible du risque optimal

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

(aussi appelé « risque bayésien »).

# Objectif (suite)

La fonction de prédiction optimale dépend de la loi $\mathrm{P}^{X,Y}$ inconnue :

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

(existence/unicité non garanties)

## Objectif de l'apprentissage supervisé

On veut construire, à partir des données $(X_1, Y_1), \ldots, (X_n, Y_n)$ une fonction de prédiction

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

telle que le risque $R(\hat{h})$ soit aussi proche que possible du risque optimal

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

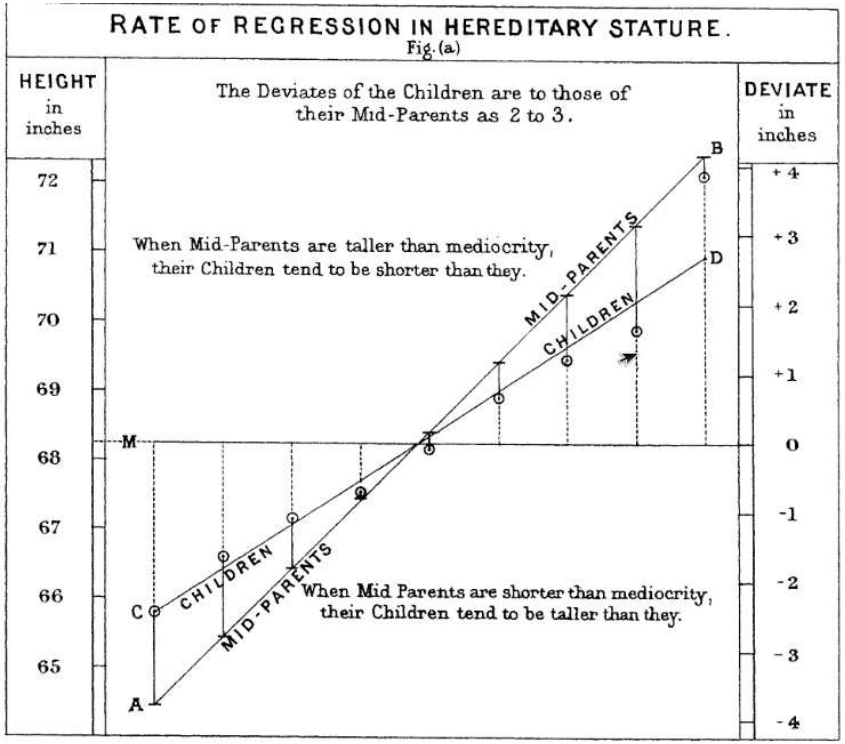
(aussi appelé « risque bayésien »).

# Plan du cours

# Plan du cours

# Régression

On s'intéresse dans le suite de ce cours à la regression : $\mathcal{Y} = \mathbb{R}$.

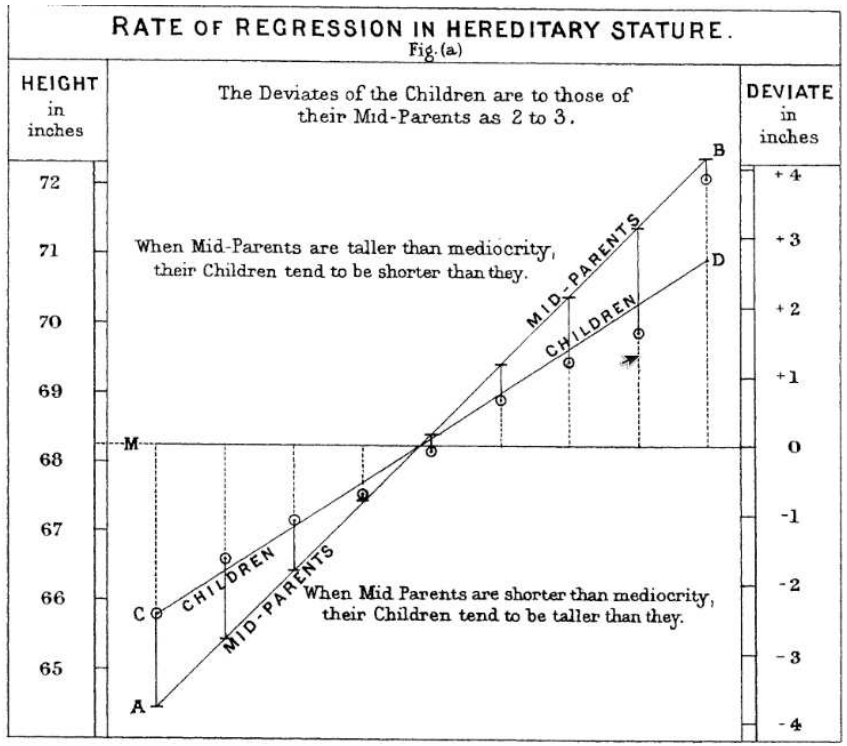



Francis Galton (1886). "Regression Towards Mediocrity in Hereditary Stature", *Journal of the Anthropological Institute*, 15:246–263.

Vocab. stat. : $Y$ = variable expliquée / $X$ = variables explicatives.

# Régression

On s'intéresse dans le suite de ce cours à la regression : $\mathcal{Y} = \mathbb{R}$.




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

Vocab. stat. : $Y$ = variable expliquée / $X$ = variables explicatives.

# Perte quadratique

Considérons pour commencer la perte quadratique :

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

(la plus couramment utilisée en regression)

**Proposition**

Pour la perte quadratique, la fonction de prédiction optimale est

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

Vocab. : $x \mapsto \mathbb{E}(Y|X = x)$ est parfois appelée « fonction de regression ».

On considérera cette fonction de perte jusqu'à nouvel ordre.

# Perte quadratique

Considérons pour commencer la perte quadratique :

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

(la plus couramment utilisée en regression)

## Proposition

Pour la perte quadratique, la fonction de prédiction optimale est

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

Vocab. : $x \mapsto \mathbb{E}(Y|X = x)$ est parfois appelée « fonction de regression ».

On considérera cette fonction de perte jusqu'à nouvel ordre.

# Perte quadratique

Considérons pour commencer la perte quadratique :

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

(la plus couramment utilisée en regression)

## Proposition

Pour la perte quadratique, la fonction de prédiction optimale est

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

Vocab. : $x \mapsto \mathbb{E}(Y|X = x)$ est parfois appelée « fonction de regression ».

On considérera cette fonction de perte jusqu'à nouvel ordre.

# Perte quadratique (suite)

**Démonstration.** Par la formule de l'espérance totale on a :

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

Le terme $\circledast$ se décompose :

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

Le premier terme ne dépend pas de $h$, et le second est minimal lorsque $h(X)=\mathbb{E}(Y \mid X)$ p.s.. □

# Perte quadratique (suite)

**Démonstration.** Par la formule de l'espérance totale on a :

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

Le terme $\circledast$ se décompose :

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

Le premier terme ne dépend pas de $h$, et le second est minimal lorsque $h(X) = \mathbb{E}(Y \mid X)$ p.s.. □

# Perte quadratique (suite)

**Démonstration.** Par la formule de l'espérance totale on a :

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

Le terme $\circledast$ se décompose :

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

Le premier terme ne dépend pas de $h$, et le second est minimal lorsque $h(X) = \mathbb{E}(Y \mid X)$ p.s.. □

# Risque empirique

Rappel : la loi jointe $\mathrm{P}^{X,Y}$ est inconnue

➠ le risque $R(h)$ ne peut être calculé.

**Définition : risque empirique**

On appelle risque empirique le risque

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

calculé avec $\mathrm{P}^{X,Y}$ égal à la mesure empirique $\hat{\mathrm{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$.

Avec la perte quadratique :

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

# Risque empirique

Rappel : la loi jointe $\mathrm{P}^{X,Y}$ est inconnue

➠ le risque $R(h)$ ne peut être calculé.

## Définition : risque empirique

On appelle risque empirique le risque

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

calculé avec $\mathrm{P}^{X,Y}$ égal à la mesure empirique $\hat{\mathrm{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$.

Avec la perte quadratique :

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

# Minimisation du risque empirique

Une méthode générale d'apprentissage :

1. Choisir une famille $\mathscr{H}$ de fonctions de prédiction.
2. Sélectionner la fonction $h$ qui minimise le risque empirique :

$$\hat{h}^{\text{MRE}} \;=\; \text{argmin}_{h \in \mathscr{H}} \; \hat{R}_n(h).$$

**Exemple : fonctions de prédiction « linéaires » (affines)**

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

⚠ la méthode MRE est raisonnable si $\mathscr{H}$ n'est « pas trop grande »

➠ sinon, il faut *pénaliser* les modèles trop complexes (voir plus loin)

# Minimisation du risque empirique

Une méthode générale d'apprentissage :

1. Choisir une famille $\mathscr{H}$ de fonctions de prédiction.
2. Sélectionner la fonction $h$ qui minimise le risque empirique :

$$\hat{h}^{\text{MRE}} \;=\; \text{argmin}_{h \in \mathscr{H}} \; \hat{R}_n(h).$$

## Exemple : fonctions de prédiction « linéaires » (affines)

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

⚠ la méthode MRE est raisonnable si $\mathscr{H}$ n'est « pas trop grande »

➠ sinon, il faut *pénaliser* les modèles trop complexes (voir plus loin)

# Autres exemples de familles de fonctions de prédiction

- modèle linéaire avec fonctions de base quelconques

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

les $h_k : \mathcal{X} \to \mathbb{R}$ étant connues ;

- modèle additif

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

avec les $h_k$ dans une famille de fonctions $\mathbb{R} \to \mathbb{R}$ à préciser ;

- réseaux de neurones,
- arbres de décisions,
- modèles linéaires/additifs généralisés (GLM/GAM)
- ...

# Autres exemples de familles de fonctions de prédiction

- modèle linéaire avec fonctions de base quelconques

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

  les $h_k : \mathcal{X} \to \mathbb{R}$ étant connues ;

- modèle additif

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

  avec les $h_k$ dans une famille de fonctions $\mathbb{R} \to \mathbb{R}$ à préciser ;

- réseaux de neurones,
- arbres de décisions,
- modèles linéaires/additifs généralisés (GLM/GAM)
- ...

# Autres exemples de familles de fonctions de prédiction

- modèle linéaire avec fonctions de base quelconques

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

  les $h_k : \mathcal{X} \to \mathbb{R}$ étant connues ;

- modèle additif

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

  avec les $h_k$ dans une famille de fonctions $\mathbb{R} \to \mathbb{R}$ à préciser ;

- réseaux de neurones,
- arbres de décisions,
- modèles linéaires/additifs généralisés (GLM/GAM)
- ...

# Plan du cours

# Somme des carrés des résidus

On considère des fonctions $h$ de la forme :

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

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

**Définition : SCR / critère des moindres carrés**

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

On définit la Somme des Carrés des Résidus (SCR) :

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

ou critère des moindres carrés.

# Somme des carrés des résidus

On considère des fonctions $h$ de la forme :

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

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

## Définition : SCR / critère des moindres carrés

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

On définit la Somme des Carrés des Résidus (SCR) :

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

ou critère des moindres carrés.

# Notations matricielles

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

➠ $\underline{X}$ de taille $n \times (p+1)$ et $\underline{Y}$ de longueur $n$.

**Ecriture matricielle du critère**

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

# Notations matricielles

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

➠ $\underline{X}$ de taille $n \times (p+1)$ et $\underline{Y}$ de longueur $n$.

## Ecriture matricielle du critère

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

# Minimisation du SCR

## Hypothèse

On suppose : $\underline{X}^\top \underline{X}$ inversible

➠ implique $p + 1 \leq n$.

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

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

où $c$ est une grandeur indépendante de $\beta$.

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

# Minimisation du SCR

## Hypothèse

On suppose : $\underline{X}^\top \underline{X}$ inversible

➠ implique $p+1 \leq n$.

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

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

où $c$ est une grandeur indépendante de $\beta$.

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

# Minimisation du SCR

## Hypothèse

On suppose : $\underline{X}^\top \underline{X}$ inversible

➠ implique $p + 1 \leq n$.

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

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

où $c$ est une grandeur indépendante de $\beta$.

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

# Minimisation du SCR

## Hypothèse

On suppose : $\underline{X}^\top \underline{X}$ inversible

➠ implique $p + 1 \leq n$.

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

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

où $c$ est une grandeur indépendante de $\beta$.

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

# Minimisation du SCR

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

On a :

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

ii. $\underline{X}^\top \underline{X}$ est inversible, donc définie positive.

(i) implique que $\mathrm{SCR}(\beta)$ est minimum en $\tilde{\beta}$ ;

(ii) implique que le minimum est unique ($a^\top \underline{X}^\top \underline{X} a = 0 \implies a = 0$).

**Proposition : estimateur des moindres carrés**

Lorsque $\underline{X}^\top \underline{X}$ est inversible,

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

est l'unique vecteur qui minimise la fonction SCR.

complément : différentiation vectorielle

# Minimisation du SCR

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

On a :

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

ii. $\underline{X}^\top \underline{X}$ est inversible, donc définie positive.

(i) implique que $\mathrm{SCR}(\beta)$ est minimum en $\tilde{\beta}$ ;

(ii) implique que le minimum est unique ($a^\top \underline{X}^\top \underline{X} a = 0 \implies a = 0$).

**Proposition : estimateur des moindres carrés**

Lorsque $\underline{X}^\top \underline{X}$ est inversible,

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

est l'unique vecteur qui minimise la fonction SCR.

complément : différentiation vectorielle

# Minimisation du SCR

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

On a :

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

ii. $\underline{X}^\top \underline{X}$ est inversible, donc définie positive.

(i) implique que $\mathrm{SCR}(\beta)$ est minimum en $\tilde{\beta}$ ;

(ii) implique que le minimum est unique ($a^\top \underline{X}^\top \underline{X} a = 0 \implies a = 0$).

**Proposition : estimateur des moindres carrés**

Lorsque $\underline{X}^\top \underline{X}$ est inversible,

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

est l'unique vecteur qui minimise la fonction SCR.

complément : différentiation vectorielle

# Minimisation du SCR

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

On a :

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

ii. $\underline{X}^\top \underline{X}$ est inversible, donc définie positive.

(i) implique que $\mathrm{SCR}(\beta)$ est minimum en $\tilde{\beta}$ ;

(ii) implique que le minimum est unique ($a^\top \underline{X}^\top \underline{X} a = 0 \implies a = 0$).

## Proposition : estimateur des moindres carrés

Lorsque $\underline{X}^\top \underline{X}$ est inversible,

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

est l'unique vecteur qui minimise la fonction SCR.

complément : différentiation vectorielle

# Quantifier l'apport des variables explicatives

Pour une régression sans variables explicatives, on aurait

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

On définit : $\mathrm{SCT} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ $\rightarrow$ Somme des Carrés Totale.

**Définition : coefficient de détermination $R^2$**

Rappel : $\mathrm{SCR}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i\right)^2$. On définit :

$$R^2 = 1 - \frac{\mathrm{SCR}(\hat{\beta})}{\mathrm{SCT}}.$$

**Propriétés.**

démonstration : voir exercice 1

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

# Quantifier l'apport des variables explicatives

Pour une régression sans variables explicatives, on aurait

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

On définit : $\text{SCT} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \rightarrow$ Somme des Carrés Totale.

## Définition : coefficient de détermination $R^2$

Rappel : $\text{SCR}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i\right)^2$. On définit :

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

**Propriétés.**

démonstration : voir exercice 1

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

# Quantifier l'apport des variables explicatives

Pour une régression sans variables explicatives, on aurait

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

On définit : $\text{SCT} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \rightarrow$ Somme des Carrés Totale.

## Définition : coefficient de détermination $R^2$

Rappel : $\text{SCR}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i\right)^2$. On définit :

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

**Propriétés.**

démonstration : voir exercice 1

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

# Exemple « Ozone » : présentation des données

| variable | description |
|---|---|
| O3obs | concentration d'ozone observée au jour $t+1$ |
| MOCAGE | prévision de pollution obtenue par un modèle déterministe de mécanique des fluides |
| TEMPE | température prévue par MétéoFrance pour le jour $t+1$ |
| RMH2O | rapport d'humidité au jour $t$ |
| NO2 | concentration en dioxyde d'azote au jour $t$ |
| NO | concentration en monoxyde d'azote au jour $t$ |
| VentMOD | force du vent au jour $t$ |
| VentANG | orientation du vent au jour $t$ |

## Questions posées

- prédire la concentration d'ozone du jour $t+1$ à partir des données disponible au jour $t$
- prédire si la concentration dépassera le seuil de $150\,\mu g/m^3$ (problème de classification traité au cours 7/9).

Application et données issues de https://github.com/wikistat/Apprentissage/tree/master/Pic-ozone

# Exemple « Ozone » : présentation des données

| variable | description |
|---|---|
| O3obs | concentration d'ozone observée au jour $t+1$ |
| MOCAGE | prévision de pollution obtenue par un modèle déterministe de mécanique des fluides |
| TEMPE | température prévue par MétéoFrance pour le jour $t+1$ |
| RMH2O | rapport d'humidité au jour $t$ |
| NO2 | concentration en dioxyde d'azote au jour $t$ |
| NO | concentration en monoxyde d'azote au jour $t$ |
| VentMOD | force du vent au jour $t$ |
| VentANG | orientation du vent au jour $t$ |

## Questions posées

- prédire la concentration d'ozone du jour $t+1$ à partir des données disponible au jour $t$
- prédire si la concentration dépassera le seuil de 150 $\mu g/m^3$ (problème de classification traité au cours 7/9).

Application et données issues de https://github.com/wikistat/Apprentissage/tree/master/Pic-ozone

# Exemple « Ozone » : présentation des données

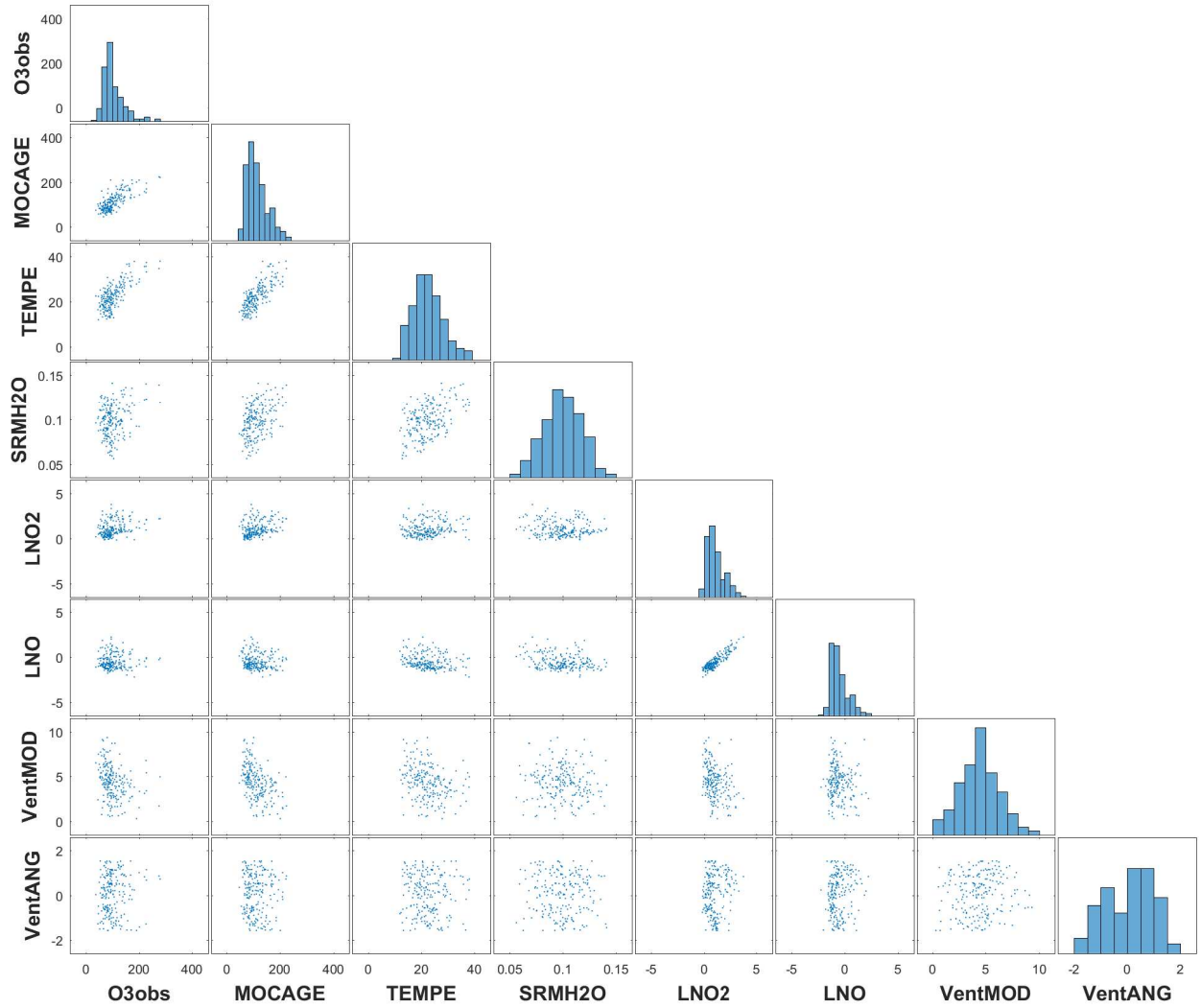
| variable | description |
|---|---|
| O3obs | concentration d'ozone observée au jour $t+1$ |
| MOCAGE | prévision de pollution obtenue par un modèle déterministe de mécanique des fluides |
| TEMPE | température prévue par MétéoFrance pour le jour $t+1$ |
| RMH2O | rapport d'humidité au jour $t$ |
| NO2 | concentration en dioxyde d'azote au jour $t$ |
| NO | concentration en monoxyde d'azote au jour $t$ |
| VentMOD | force du vent au jour $t$ |
| VentANG | orientation du vent au jour $t$ |

## Questions posées

- prédire la concentration d'ozone du jour $t+1$ à partir des données disponible au jour $t$
- prédire si la concentration dépassera le seuil de $150\,\mu g/m^3$ (problème de classification traité au cours 7/9).

Application et données issues de https://github.com/wikistat/Apprentissage/tree/master/Pic-ozone

# Exemple « Ozone » : exploration des données



# Exemple « Ozone » : coefficients de régression

Régression linéaire effectuée sur $n = 210$ jours.

**Remarque.** Les variables ont été centrées et réduites pour faciliter l'interprétation.

| $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|
| 103.4 | 28.9 | 22.5 | -3.2 | -34.4 | 37.9 | 1.4 | 2.6 |

**Coefficient de détermination.** $R^2 = 65.7\%$

Observations :

- le coefficient négatif associé à NO2 est surprenant
  (mais NO2 est corrélée avec NO)
- RMH2O, VentMOD et VentANG semblent secondaires
- Le modèle explique une partie des données.

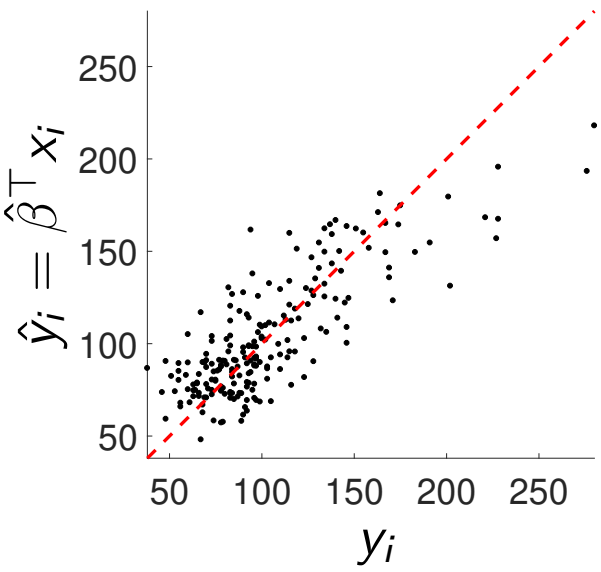
# Exemple « Ozone » : coefficients de régression

Régression linéaire effectuée sur $n = 210$ jours.

**Remarque.** Les variables ont été centrées et réduites pour faciliter l'interprétation.

| $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|
| 103.4 | 28.9 | 22.5 | -3.2 | -34.4 | 37.9 | 1.4 | 2.6 |

**Coefficient de détermination.** $R^2 = 65.7\%$




Observations :

- le coefficient négatif associé à NO2 est surprenant
  (mais NO2 est corrélée avec NO)
- RMH2O, VentMOD et VentANG semblent secondaires
- Le modèle explique une partie des données.

# Plan du cours

# Propriétés de l'estimateur des moindres carrés

Rappel : jusqu'ici $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathrm{P}^{X,Y}$.

➠ dans cette partie, on suppose plutôt les $X_i$ déterministes

(ou, de façon équivalente : on travaille « conditionnellement aux $X_i$ »).

Supposons désormais qu'il existe $\beta \in \mathbb{R}^{p+1}$ tel que

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

avec des erreurs $\epsilon_i$

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

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

(iv) homoscédastiques : $\mathrm{var}(\epsilon_i) = \sigma^2$ pour un certain $\sigma^2 > 0$.

# Propriétés de l'estimateur des moindres carrés

Rappel : jusqu'ici $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathrm{P}^{X,Y}$.

➠ dans cette partie, on suppose plutôt les $X_i$ déterministes

(ou, de façon équivalente : on travaille « conditionnellement aux $X_i$ »).

Supposons désormais qu'il existe $\beta \in \mathbb{R}^{p+1}$ tel que

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

avec des erreurs $\epsilon_i$

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

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

(iv) homoscédastiques : $\mathrm{var}(\epsilon_i) = \sigma^2$ pour un certain $\sigma^2 > 0$.

# Propriétés de l'estimateur des moindres carrés

## Proposition

Sous ces hypothèses, l'estimateur $\hat{\beta}$ est sans biais :

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

et sa matrice de covariance s'écrit :

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

# Propriétés de l'estimateur des moindres carrés

**Démonstration.**

Rappel : les $X_i$ sont supposés déterministes.

On note $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Alors :

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

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

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

□

# Propriétés de l'estimateur des moindres carrés

**Démonstration.**

Rappel : les $X_i$ sont supposés déterministes.

On note $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Alors :

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

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

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

□

# Propriétés de l'estimateur des moindres carrés

**Démonstration.**

Rappel : les $X_i$ sont supposés déterministes.

On note $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Alors :

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

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

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

□

# Loi de l'estimateur $(\hat{\beta}, \hat{\sigma}^2)$ sous hypothèse gaussienne

On suppose de plus que (v) $\underline{\epsilon}$ est gaussien :

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

**Proposition : EMV de $(\beta, \sigma^2)$** (voir TD 6)

L'EMV s'écrit : $\begin{cases} \hat{\beta} = \arg\min_\beta \sum_{i=1}^n \left(Y_i - \beta^\top X_i\right)^2, \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat{\beta}^\top X_i\right)^2. \end{cases}$

➠ On retrouve pour $\beta$ l'estimateur des moindres carrés

**Théorème de Student : loi de $(\hat{\beta}, \hat{\sigma}^2)$** (voir TD 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 $\hat{\sigma}^2$ sont indépendants.

# Loi de l'estimateur $(\hat{\beta}, \hat{\sigma}^2)$ sous hypothèse gaussienne

On suppose de plus que (v) $\underline{\epsilon}$ est gaussien :

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

**Proposition : EMV de $(\beta, \sigma^2)$** (voir TD 6)

L'EMV s'écrit : $\begin{cases} \hat{\beta} = \arg\min_\beta \sum_{i=1}^n \left(Y_i - \beta^\top X_i\right)^2, \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat{\beta}^\top X_i\right)^2. \end{cases}$

➠ On retrouve pour $\beta$ l'estimateur des moindres carrés

**Théorème de Student : loi de $(\hat{\beta}, \hat{\sigma}^2)$** (voir TD 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 $\hat{\sigma}^2$ sont indépendants.

# Loi de l'estimateur $(\hat{\beta}, \hat{\sigma}^2)$ sous hypothèse gaussienne

On suppose de plus que (v) $\underline{\epsilon}$ est gaussien :

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

## Proposition : EMV de $(\beta, \sigma^2)$ (voir TD 6)

L'EMV s'écrit : $\begin{cases} \hat{\beta} = \arg\min_\beta \sum_{i=1}^n \left(Y_i - \beta^\top X_i\right)^2, \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat{\beta}^\top X_i\right)^2. \end{cases}$

➠ On retrouve pour $\beta$ l'estimateur des moindres carrés

## Théorème de Student : loi de $(\hat{\beta}, \hat{\sigma}^2)$ (voir TD 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 $\hat{\sigma}^2$ sont indépendants.

# Tests / IC sur la valeur d'un élément de $\beta$

On sait que $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 v_j\right)$ avec $v_j = \left[\left(\underline{X}^\top \underline{X}\right)^{-1}\right]_{j,j}$.

## Fonction pivotale

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

avec $\mathcal{T}(n-p-1)$ : loi de Student à $n-p-1$ degrés de liberté

Loi de Student

**Remarque** :

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

est un estimateur sans biais de $\sigma^2$ (voir TD 6).

# Tests / IC sur la valeur d'un élément de $\beta$

On sait que $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 v_j\right)$ avec $v_j = \left[\left(\underline{X}^\top \underline{X}\right)^{-1}\right]_{j,j}$.

## Fonction pivotale

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

avec $\mathcal{T}(n-p-1)$ : loi de Student à $n-p-1$ degrés de liberté

Loi de Student

**Remarque** :

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

est un estimateur sans biais de $\sigma^2$ (voir TD 6).

# Tests / IC sur la valeur d'un élément de $\beta$

On sait que $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 v_j\right)$ avec $v_j = \left[\left(\underline{X}^\top \underline{X}\right)^{-1}\right]_{j,j}$.

## Fonction pivotale

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

avec $\mathcal{T}(n-p-1)$ : loi de Student à $n-p-1$ degrés de liberté

Loi de Student

**Remarque** :

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

est un estimateur sans biais de $\sigma^2$ (voir TD 6).

# Démonstration

D'après le théorème de Student,

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

Ainsi

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

par définition de la loi de Student à $k = n - p - 1$ degrés de libertés. □

# Démonstration

D'après le théorème de Student,

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

Ainsi

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

par définition de la loi de Student à $k = n - p - 1$ degrés de libertés. □

## Test pour $H_0 : \beta_j = 0$ / $H_1 : \beta_j \neq 0$

Soit $0 < \alpha < 1$.

On prend $\beta_j = 0$ dans la déf. de $T$ (on se place sous $H_0$) et

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

## Intervalle de confiance exact pour $\beta_j$

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

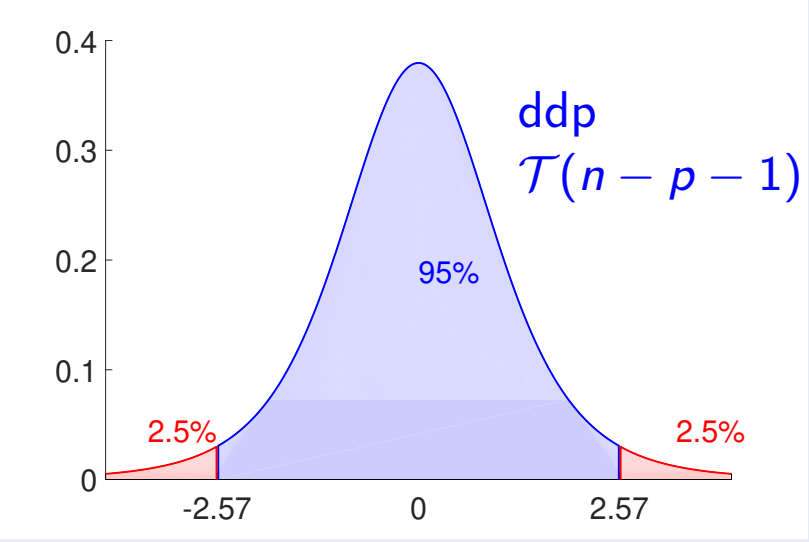
$q_r$ : quantile d'ordre $r$ de la loi $\mathcal{T}(n-p-1)$

## Test pour $H_0 : \beta_j = 0$ / $H_1 : \beta_j \neq 0$

Soit $0 < \alpha < 1$.

On prend $\beta_j = 0$ dans la déf. de $T$ (on se place sous $H_0$) et

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




## Intervalle de confiance exact pour $\beta_j$

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

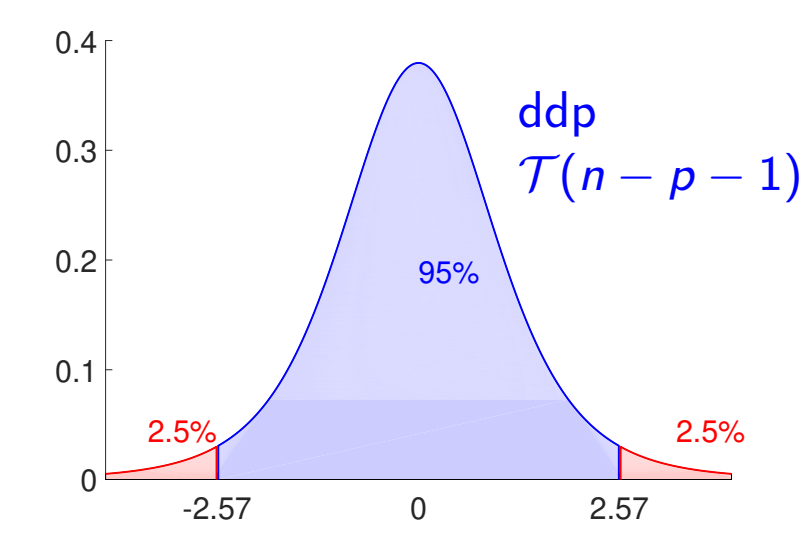
$q_r$ : quantile d'ordre $r$ de la loi $\mathcal{T}(n-p-1)$

## Test pour $H_0 : \beta_j = 0$ / $H_1 : \beta_j \neq 0$

Soit $0 < \alpha < 1$.

On prend $\beta_j = 0$ dans la déf. de $T$ (on se place sous $H_0$) et

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




## Intervalle de confiance exact pour $\beta_j$

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

$q_r$ : quantile d'ordre $r$ de la loi $\mathcal{T}(n-p-1)$

# Exemple « Ozone » : test des coefficient $\beta_j$

| | $IC_{95\%}$ | $t$ | pval |
|---|---|---|---|
| $\beta_0$ | [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 |

avec $t$ : valeur réalisée de $T$ pour le coefficient concerné

Remarque : régression sans les variables RMH2O, VentMOD et VentANG

➠ le coefficient de détermination passe de 65.7% à 64.5%.

# Exemple « Ozone » : test des coefficient $\beta_j$

| | $IC_{95\%}$ | $t$ | pval |
|---|---|---|---|
| $\beta_0$ | [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 |

avec $t$ : valeur réalisée de $T$ pour le coefficient concerné

Remarque : régression sans les variables RMH2O, VentMOD et VentANG

➠ le coefficient de détermination passe de 65.7% à 64.5%.

# Plan du cours

# Exemple « Ozone » : cas d'un échantillon corrompu

Considérons que 5 mesures de concentration d'ozone sur $n$ ($n = 210$) sont corrompues, soit environ 2% de l'échantillon.

Coefficients estimés avec/sans données corrompues :

| | $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|---|
| sans | 103.4 | 28.9 | 22.6 | -3.2 | -34.4 | 37.6 | 1.4 | 2.6 |
| avec | 125.2 | 79.2 | -15.6 | 24.2 | -155.1 | 141.4 | 4.7 | 24.9 |

➠ Grande sensibilité des coefficients de régression aux « outliers ».

**Solution**

Utiliser une fonction de perte qui conduit à une fonction de prédiction plus robuste que la fonction de perte quadratique.

# Exemple « Ozone » : cas d'un échantillon corrompu

Considérons que 5 mesures de concentration d'ozone sur $n$ ($n = 210$) sont corrompues, soit environ 2% de l'échantillon.

Coefficients estimés avec/sans données corrompues :

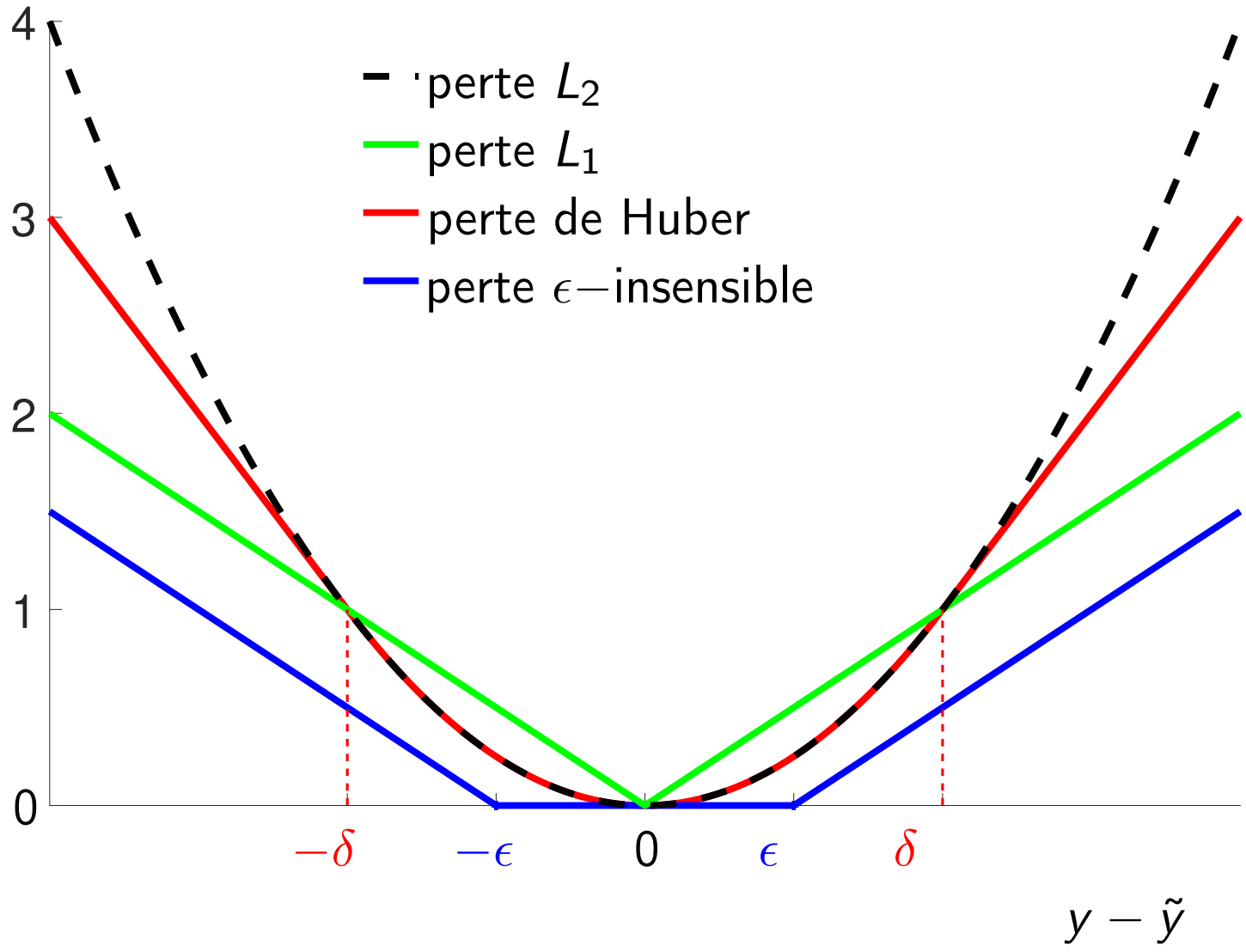
| | $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|---|
| sans | 103.4 | 28.9 | 22.6 | -3.2 | -34.4 | 37.6 | 1.4 | 2.6 |
| avec | 125.2 | 79.2 | -15.6 | 24.2 | -155.1 | 141.4 | 4.7 | 24.9 |

➠ Grande sensibilité des coefficients de régression aux « outliers ».

**Solution**

Utiliser une fonction de perte qui conduit à une fonction de prédiction plus robuste que la fonction de perte quadratique.

# Fonctions de perte usuelles



# Perte $L_1$

Fonction de perte : $L(y, \tilde{y}) = |y - \tilde{y}|$.

**Proposition** (voir TD 6)

Pour la perte $L_1$, la fonction de prédiction optimale est

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

**Exemple « Ozone »**

| | $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|---|
| sans | 100.8 | 27.5 | 19.2 | -3.3 | -32.2 | 33.9 | -1.0 | 3.9 |
| avec | 101.4 | 28.3 | 18.6 | -1.6 | -35.1 | 37.5 | 0.5 | 3.2 |

➠ meilleure stabilité par rapport à la présence d'outliers.

# Perte $L_1$

Fonction de perte : $L(y, \tilde{y}) = |y - \tilde{y}|$.

## Proposition (voir TD 6)

Pour la perte $L_1$, la fonction de prédiction optimale est

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

### Exemple « Ozone »

| | $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|---|
| sans | 100.8 | 27.5 | 19.2 | -3.3 | -32.2 | 33.9 | -1.0 | 3.9 |
| avec | 101.4 | 28.3 | 18.6 | -1.6 | -35.1 | 37.5 | 0.5 | 3.2 |

➠ meilleure stabilité par rapport à la présence d'outliers.

# Perte $L_1$

Fonction de perte : $L(y, \tilde{y}) = |y - \tilde{y}|$.

## Proposition (voir TD 6)

Pour la perte $L_1$, la fonction de prédiction optimale est

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

**Exemple « Ozone »**

| | $\beta_0$ | MOCAGE | TEMPE | RMH2O | NO2 | NO | VentMOD | VentANG |
|---|---|---|---|---|---|---|---|---|
| sans | 100.8 | 27.5 | 19.2 | -3.3 | -32.2 | 33.9 | -1.0 | 3.9 |
| avec | 101.4 | 28.3 | 18.6 | -1.6 | -35.1 | 37.5 | 0.5 | 3.2 |

➠ meilleure stabilité par rapport à la présence d'outliers.

# Plan du cours

# Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables ($n \times (p+1)$).

**Situations critiques les « moindres carrés ordinaires »**

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible,
- ou mal conditionnée.

Cas typiques :

- lorsque le nombre de variables est important
  ($p + 1 > n$, voire $p \gg n$)

  Exemple avec $\underline{X}$ de taille #patients $\times$ #gènes

- lorsque il y a de fortes corrélations entre les variables explicatives

  Exemple « Ozone » avec les variables NO et NO2

  ➠ interprétation hasardeuse des coefficients de régression

# Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables $(n \times (p+1))$.

**Situations critiques les « moindres carrés ordinaires »**

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible,
- ou mal conditionnée.

Cas typiques :

- lorsque le nombre de variables est important
  $(p+1 > n$, voire $p \gg n)$

  Exemple avec $\underline{X}$ de taille #patients $\times$ #gènes

- lorsque il y a de fortes corrélations entre les variables explicatives

  Exemple « Ozone » avec les variables NO et NO2
  ➠ interprétation hasardeuse des coefficients de régression

# Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables $(n \times (p+1))$.

**Situations critiques les « moindres carrés ordinaires »**

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible,
- ou mal conditionnée.

Cas typiques :

- lorsque le nombre de variables est important
  $(p+1 > n$, voire $p \gg n)$

  Exemple avec $\underline{X}$ de taille #patients $\times$ #gènes

- lorsque il y a de fortes corrélations entre les variables explicatives

  Exemple « Ozone » avec les variables NO et NO2

  ➠ interprétation hasardeuse des coefficients de régression

# Limites des « moindres carrés ordinaires »

Rappel : matrice $\underline{X}$ de taille #individus $\times$ #variables ($n \times (p+1)$).

**Situations critiques les « moindres carrés ordinaires »**

- lorsque la matrice $\underline{X}^\top \underline{X}$ n'est pas inversible,
- ou mal conditionnée.

Cas typiques :

- lorsque le nombre de variables est important
  ($p + 1 > n$, voire $p \gg n$)

  Exemple avec $\underline{X}$ de taille #patients $\times$ #gènes

- lorsque il y a de fortes corrélations entre les variables explicatives

  Exemple « Ozone » avec les variables NO et NO2
  ➠ interprétation hasardeuse des coefficients de régression

# Solution possible : régression pénalisée

Au critère précédent (SCR), on ajoute une pénalité :

$$\hat{\beta} = \arg\min_{\beta} \underbrace{\text{SCR}(\beta)}_{\text{« attache » aux données}} + \underbrace{\lambda}_{\text{hyperparamètre}} \underbrace{\Omega(\beta)}_{\text{pénalité}} .$$

➠ voir cours 8

# Solution possible : régression pénalisée

Au critère précédent (SCR), on ajoute une pénalité :

$$\hat{\beta} = \arg\min_{\beta} \underbrace{\text{SCR}(\beta)}_{\text{« attache » aux données}} + \underbrace{\lambda}_{\text{hyperparamètre}} \underbrace{\Omega(\beta)}_{\text{pénalité}} .$$

➠ voir cours 8

# Plan du cours

# Plan du cours

# Exercice 1 (La régression vue comme une projection) corrigé

Soient, pour $1 \leq i \leq n$, des observations $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$.

On considère le modèle de régression linéaire du slide 21 :

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

et l'estimateur des moindres carrés associé :

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

Comme au slide 22, on note

- $\underline{X} \in \mathbb{R}^{n \times (p+1)}$ la matrice des régresseurs,
- $\underline{Y} \in \mathbb{R}^n$ le vecteur des réponses.

# Exercice 1 (La régression vue comme une projection) corrigé

## Questions

1. On note $\underline{\hat{Y}} = \underline{X}\hat{\beta}$. Montrer que $\underline{\hat{Y}}$ est le projeté de $\underline{Y}$ sur l'image de $\underline{X}$.
2. Donner l'expression de la matrice de projection dans le cas où $\underline{X}^\top \underline{X}$ est inversible.
3. Montrer que le coefficient de détermination, défini au slide 25, vérifie $0 \leq R^2 \leq 1$, avec $R^2 = 1$ ssi $\forall i,\ Y_i = \hat{\beta}^\top X_i$.

# Plan du cours

# Corrigé de l'exercice 1

retour au sujet

❶ Rappels :

- Le projeté de $y \in \mathbb{R}^n$ sur un convexe fermé $C \subset \mathbb{R}^n$ est l'unique $y^* \in C$ tel que $\|y - y^*\| = \min_{v \in C} \|y - v\|$.
- L'image de $\underline{X}$, que nous noterons $\mathrm{Im}(\underline{X})$, est le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\underline{X}$ :

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

On remarque pour commencer que

- $\mathrm{Im}(\underline{X})$ est bien un convexe fermé (puisqu'en dimension finie tous les sev sont fermés),
- $\underline{\hat{Y}} = \underline{X}\hat{\beta}$ est dans $\mathrm{Im}(\underline{X})$.

retour au sujet

De plus, pour tout $v = \underline{X}\beta \in \mathrm{Im}(\underline{X})$, en utilisant que

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

on trouve que

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

donc $\hat{\underline{Y}}$ est bien le projeté de $\underline{Y}$ sur $\mathrm{Im}(\underline{X})$.

❷ En utilisant l'expression de $\hat{\beta}$ établie en cours, on peut exprimer le projeté de $\underline{Y}$ sur $\mathrm{Im}(\underline{X})$ comme

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

Ceci étant vrai pour tout $\underline{Y} \in \mathbb{R}^n$, on en déduit l'expression de la matrice de projection :

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

❸ On rappelle la caractérisation du projeté sur un sev :

**Théorème**

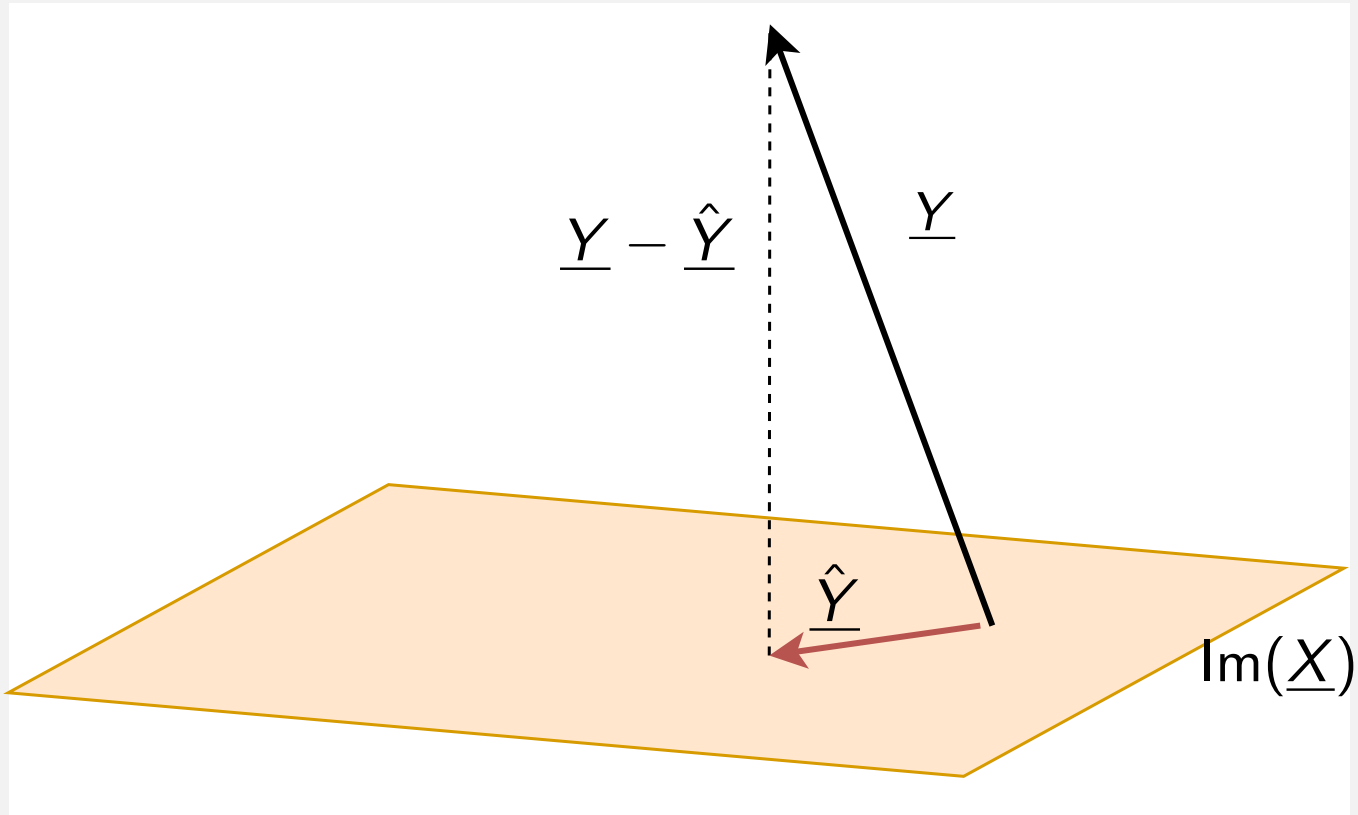
Soit $y \in \mathbb{R}^n$ et soit $F$ un sous-espace vectoriel de $\mathbb{R}^n$. Alors $y^*$ est le projeté de $y$ sur $F$ si, et seulement si,

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

# Corrigé de l'exercice 1

retour au sujet

On applique le théorème avec $F = \mathrm{Im}(\underline{X})$ et $y = \underline{Y}$.



# Corrigé de l'exercice 1

retour au sujet

Considérons maintenant le coefficient de détermination :

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

Décomposons le terme SCT :

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

Le passage de (1) à (2) découle du théorème Pythagore.

# Corrigé de l'exercice 1

retour au sujet

En effet,

- $\underline{\hat{Y}} \in \mathrm{Im}(\underline{X})$ et $\underline{Y} - \underline{\hat{Y}} \in \mathrm{Im}(\underline{X})^{\perp}$ puisque $\underline{\hat{Y}}$ est le projeté de $\underline{Y}$ sur le sous-espace vectoriel $\mathrm{Im}(\underline{X})$.
- $\underline{\hat{Y}} - \bar{Y} 1_{n\times 1} \in \mathrm{Im}(\underline{X})$ puisque $1_{n\times 1} \in \mathrm{Im}(\underline{X})$.

Ainsi :

i. $0 \leq \mathrm{SCR}(\hat{\beta}) \leq SCT$, donc $0 \leq R^2 \leq 1$,

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

□

# Plan du cours

# Différentiation par rapport à un vecteur

Le résultat peut également être obtenu par « différentiation vectorielle ».

Soient $v \in \mathbb{R}^q$, $z \in \mathbb{R}^q$ et $M \in \mathbb{R}^{q \times q}$.

## 1) différentiation de $h(z) = v^\top z = \sum_{j=1}^{q} v_j z_j$

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

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

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

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

# Différentiation par rapport à un vecteur (suite)

Application à la minimisation du critère des moindres carrés.

Rappelons que

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

On a donc

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

et finalement :

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

□

retour au slide 24

# Centrer et réduire les variables

Soit $\underline{X} = (X_1, \ldots, X_n)$ un $n$-échantillon à valeurs dans $\mathbb{R}^p$.

Centrer et réduire les variables consiste à considérer les observations $\widetilde{X}_i^{(j)}$ définies par

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

où $\overline{X}_n^{(j)}$ et $S_n^{(j)}$ sont respectivement la moyenne et l'écart-type empiriques de la $j$-ème variable :

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

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

retour au slide 28

# Quelques mots sur la loi de Student $\mathcal{T}(k)$

## Définition de $\mathcal{T}(k)$, $k$ entier$\geq 1$

Soient deux VA $U$ et $V$ tel que :

- $U \sim \mathcal{N}(0,1)$
- $V \sim \chi^2(k)$
- $U$ et $V$ indépendantes

alors $T = \frac{U}{\sqrt{\frac{V}{k}}}$ suit une loi de Student à $k$ degrés de liberté.

## Propriété

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

Exercice : le prouver.

## Densité de probabilité

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

## Moyenne

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

## Variance

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

retour au slide 33