

# **SD-TSIA204**

## **Statistics : linear models**

**Joseph Salmon**

<http://josephsalmon.eu>

Télécom ParisTech, Institut Mines-Télécom

# Outline

Introduction : OLS with two features

Multivariate least square

- Matrix model

- Least squares definition

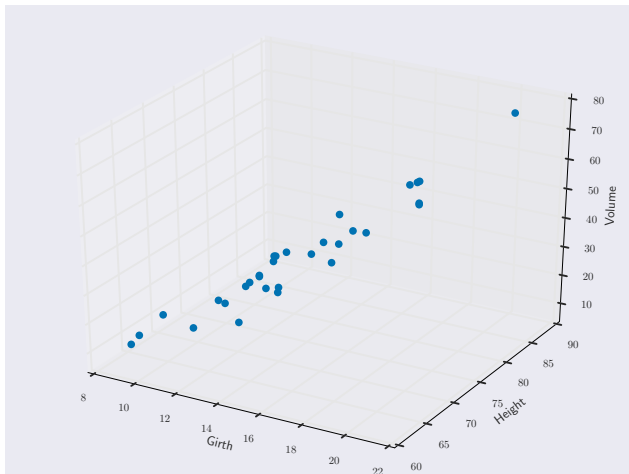
- Optimization

- Uniqueness issues

- Closed-form solution, prediction and residual

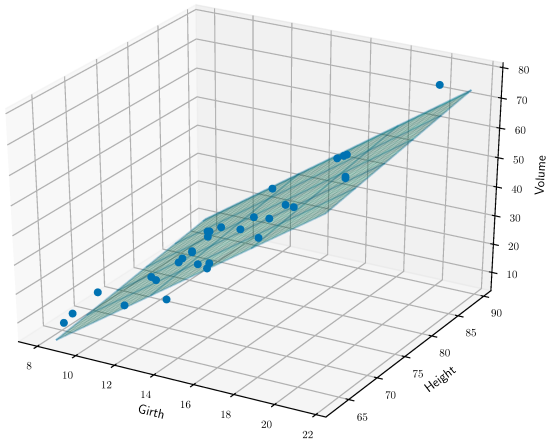
# Toward multivariate models

Tree volume as a function of height / girth ( ■ ■ : *circonférence* )



# Toward multivariate models

Tree volume as a function of height / girth (■ : *circonférence*)



## Python commands

```
from matplotlib.mplot3d import Axes3D
# Load data
url = 'http://vincentarelbundock.github.io/
      Rdatasets/csv/datasets/trees.csv'
dat3 = pd.read_csv(url)
# Fit regression model
X = dat3[['Girth', 'Height']]
X = sm.add_constant(X)
y = dat3['Volume']
results = sm.OLS(y, X).fit().params
XX = np.arange(8, 22, 0.5)
YY = np.arange(64, 90, 0.5)
xx, yy = np.meshgrid(XX, YY)
zz = results[0] + results[1]*xx + results[2]*yy
fig = plt.figure()
ax = Axes3D(fig)
ax.plot(X['Girth'], X['Height'], y, 'o')
ax.plot_wireframe(xx, yy, zz, rstride=10, cstride=10)
plt.show()
```

results output const:-57.98, Girth: 4.70, Height: 0.33

# Outline

Introduction : OLS with two features

## Multivariate least square

- Matrix model

- Least squares definition

- Optimization

- Uniqueness issues

- Closed-form solution, prediction and residual

# Model

One observes  $p$  features  $(\mathbf{x}_1, \dots, \mathbf{x}_p)$

Model in dimension  $p$

$$y_i = \theta_0^\star + \sum_{j=1}^p \theta_j^\star x_{i,j} + \varepsilon_i$$

$$\varepsilon_i \stackrel{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \dots, n$$

$$\mathbb{E}(\varepsilon) = 0$$

Rem: we assume (frequentist point of view) there exists a “true” parameter  $\boldsymbol{\theta}^\star = (\theta_0^\star, \dots, \theta_p^\star)^\top \in \mathbb{R}^{p+1}$

# Dimension $p$

## Matrix model

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}}_X \underbrace{\begin{pmatrix} \theta_0^* \\ \vdots \\ \theta_p^* \end{pmatrix}}_{\boldsymbol{\theta}^*} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

Equivalently :  $\boxed{\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}}$

Column notation :  $X = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p)$  with  
 $\mathbf{x}_0 = \mathbf{1}_n = (1, \dots, 1)^\top$

Line notation :  $X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = (x_1, \dots, x_n)^\top$

Rem: often  $\mathbf{x}_0$  will be omitted by simplicity, e.g., center  $\mathbf{y}$  first



# Vocabulary

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$$

- ▶  $\mathbf{y} \in \mathbb{R}^n$  : observations vector
- ▶  $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$  : **design** matrix (with features as columns)
- ▶  $\boldsymbol{\theta}^* \in \mathbb{R}^{p+1}$  : (unknown) **true** parameter to be estimated
- ▶  $\boldsymbol{\varepsilon} \in \mathbb{R}^n$  : noise vector

“Observations” point of view :  $y_i = \langle x_i, \boldsymbol{\theta}^* \rangle + \varepsilon_i$  for  $i = 1, \dots, n$   
 $\langle \cdot, \cdot \rangle$  stands for standard inner product (■ ■ : *produit scalaire*)

“Features” point of view :  $\mathbf{y} = \sum_{j=0}^p \theta_j^* \mathbf{x}_j + \boldsymbol{\varepsilon}$

# Outline

Introduction : OLS with two features

## Multivariate least square

- Matrix model

- Least squares definition

- Optimization

- Uniqueness issues

- Closed-form solution, prediction and residual

## (Ordinary) Least squares

A least square estimator is any solution of the following problem :

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \left( \frac{1}{2} \|\mathbf{y} - X\theta\|_2^2 \right)$$

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \frac{1}{2} \sum_{i=1}^n \left[ y_i - \left( \theta_0 + \sum_{j=1}^p \theta_j x_{i,j} \right) \right]^2$$

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \frac{1}{2} \sum_{i=1}^n [y_i - \langle x_i, \theta \rangle]^2$$

Rem: a solution always exists, as we are minimizing a coercive continuous function (**coercive** :  $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ )

Rem: uniqueness is not guaranteed

Rem: the  $\frac{1}{2}$  term does not change the optimization problem, but simplifies gradient computation

# Outline

Introduction : OLS with two features

## Multivariate least square

- Matrix model

- Least squares definition

### Optimization

- Uniqueness issues

- Closed-form solution, prediction and residual

# First order condition / Fermat's rule

## Theorem : Fermat's rule

If  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is differentiable at a local minimum  $\theta^*$  then the gradient of  $f$  vanishes at  $\theta^*$ , i.e.,  $\nabla f(\theta^*) = 0$ .

Rem: sufficient condition when  $f$  is convex !

For least squares  $f : \theta \mapsto \frac{1}{2} \|\mathbf{y} - X\theta\|_2^2$  or

$$\begin{aligned} f(\theta) &= \frac{1}{2} \|\mathbf{y} - X\theta\|_2^2 \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle X\theta, \mathbf{y} \rangle + \frac{1}{2} \theta^\top X^\top X \theta \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \theta, X^\top \mathbf{y} \rangle + \frac{1}{2} \theta^\top X^\top X \theta \end{aligned}$$

## Gradient computation

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\boldsymbol{\theta}$  as the vector satisfying :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{for any } h$$

For the  $f$  of interest here, this reads

$$f(\boldsymbol{\theta} + h) = \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2} (\boldsymbol{\theta} + h)^\top X^\top X (\boldsymbol{\theta} + h)$$

## Gradient computation

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\boldsymbol{\theta}$  as the vector satisfying :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{for any } h$$

For the  $f$  of interest here, this reads

$$\begin{aligned} f(\boldsymbol{\theta} + h) &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2} (\boldsymbol{\theta} + h)^\top X^\top X (\boldsymbol{\theta} + h) \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\ &\quad + \frac{1}{2} \boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \end{aligned}$$

## Gradient computation

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\boldsymbol{\theta}$  as the vector satisfying :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{for any } h$$

For the  $f$  of interest here, this reads

$$\begin{aligned} f(\boldsymbol{\theta} + h) &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2} (\boldsymbol{\theta} + h)^\top X^\top X (\boldsymbol{\theta} + h) \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\ &\quad + \frac{1}{2} \boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\ &= f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \end{aligned}$$



## Gradient computation

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\boldsymbol{\theta}$  as the vector satisfying :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{for any } h$$

For the  $f$  of interest here, this reads

$$\begin{aligned} f(\boldsymbol{\theta} + h) &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2} (\boldsymbol{\theta} + h)^\top X^\top X (\boldsymbol{\theta} + h) \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\ &\quad + \frac{1}{2} \boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\ &= f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\ &= f(\boldsymbol{\theta}) + \underbrace{\langle h, X^\top X \boldsymbol{\theta} - X^\top \mathbf{y} \rangle}_{\nabla f(\boldsymbol{\theta})} + \underbrace{\frac{1}{2} h^\top X^\top X h}_{o(h)} \end{aligned}$$

## Gradient computation

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\boldsymbol{\theta}$  as the vector satisfying :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{for any } h$$

For the  $f$  of interest here, this reads

$$\begin{aligned} f(\boldsymbol{\theta} + h) &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta} + h, X^\top \mathbf{y} \rangle + \frac{1}{2} (\boldsymbol{\theta} + h)^\top X^\top X (\boldsymbol{\theta} + h) \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \boldsymbol{\theta}, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\ &\quad + \frac{1}{2} \boldsymbol{\theta}^\top X^\top X \boldsymbol{\theta} + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\ &= f(\boldsymbol{\theta}) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \boldsymbol{\theta}^\top X^\top X h \\ &= f(\boldsymbol{\theta}) + \underbrace{\langle h, X^\top X \boldsymbol{\theta} - X^\top \mathbf{y} \rangle}_{\nabla f(\boldsymbol{\theta})} + \underbrace{\frac{1}{2} h^\top X^\top X h}_{o(h)} \end{aligned}$$

Hence,

$$\nabla f(\boldsymbol{\theta}) = X^\top X \boldsymbol{\theta} - X^\top \mathbf{y} = X^\top (X \boldsymbol{\theta} - \mathbf{y})$$

## Gradient computation

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\theta$  as the vector satisfying :

$$f(\theta + h) = f(\theta) + \langle h, \nabla f(\theta) \rangle + o(h) \quad \text{for any } h$$

For the  $f$  of interest here, this reads

$$\begin{aligned} f(\theta + h) &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \theta + h, X^\top \mathbf{y} \rangle + \frac{1}{2} (\theta + h)^\top X^\top X (\theta + h) \\ &= \frac{1}{2} \|\mathbf{y}\|^2 - \langle \theta, X^\top \mathbf{y} \rangle - \langle h, X^\top \mathbf{y} \rangle \\ &\quad + \frac{1}{2} \theta^\top X^\top X \theta + \frac{1}{2} h^\top X^\top X h + \theta^\top X^\top X h \\ &= f(\theta) - \langle h, X^\top \mathbf{y} \rangle + \frac{1}{2} h^\top X^\top X h + \theta^\top X^\top X h \\ &= f(\theta) + \underbrace{\langle h, X^\top X \theta - X^\top \mathbf{y} \rangle}_{\nabla f(\theta)} + \underbrace{\frac{1}{2} h^\top X^\top X h}_{o(h)} \end{aligned}$$

Hence,

$$\nabla f(\theta) = X^\top X \theta - X^\top \mathbf{y} = X^\top (X \theta - \mathbf{y})$$

## Alternative gradient formulation in finite dimension

The gradient of  $f$ ,  $\nabla f$  is defined for any  $\boldsymbol{\theta}$  as the vector satisfying :

$$f(\boldsymbol{\theta} + h) = f(\boldsymbol{\theta}) + \langle h, \nabla f(\boldsymbol{\theta}) \rangle + o(h) \quad \text{for any } h$$

Property : the gradient can be formulated as the vector of partial derivatives

$$\nabla f(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_0} \\ \vdots \\ \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix}$$

## Least squares - normal equation

$$\nabla f(\boldsymbol{\theta}) = 0 \Leftrightarrow X^\top X \boldsymbol{\theta} - X^\top \mathbf{y} = X^\top (X \boldsymbol{\theta} - \mathbf{y}) = 0$$

### Theorem

Fermat's rule ensures that any solution  $\hat{\boldsymbol{\theta}}$  satisfies :

**Normal equation :**

$$X^\top X \hat{\boldsymbol{\theta}} = X^\top \mathbf{y}$$

$\hat{\boldsymbol{\theta}}$  is solution of the linear system " $A\boldsymbol{\theta} = b$ " for a matrix  $A = X^\top X$  and right hand side  $b = X^\top \mathbf{y}$

Rem: uniqueness does not hold when features are **co-linear**, and then there are an infinite number of solutions

---

**Exo:** code (in Python) gradient descent for least squares

---

# Vocabulary (and abuse of terms)

## Definition

We call **Gramian matrix** (■ ■ : *matrice de Gram*) the matrix

$$X^{\top} X$$

whose general term is  $[X^{\top} X]_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$

Rem:  $X^{\top} X$  is often referred to as the feature correlation matrix (true for standardized columns)

Rem: when columns are scaled such that  $\forall j \in \llbracket 0, p \rrbracket, \|\mathbf{x}_j\|^2 = n$ , the Gramian diagonal is  $(n, \dots, n)$

The vector  $X^{\top} \mathbf{y} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix}$  represents the correlation

between the observations and the features

# Outline

Introduction : OLS with two features

## Multivariate least square

- Matrix model

- Least squares definition

- Optimization

- Uniqueness issues**

- Closed-form solution, prediction and residual

## Least squares and uniqueness

Let  $\hat{\boldsymbol{\theta}}$  be a solution of  $\boxed{X^\top X \hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$

**Non uniqueness** : happens for non trivial kernel, *i.e.*, when  $\text{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$

Assume  $\boldsymbol{\theta}_K \in \text{Ker}(X)$  with  $\boldsymbol{\theta}_K \neq 0$ , then

$$X(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X\hat{\boldsymbol{\theta}}$$

$$\text{and then } (X^\top X)(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) = X^\top \mathbf{y}$$

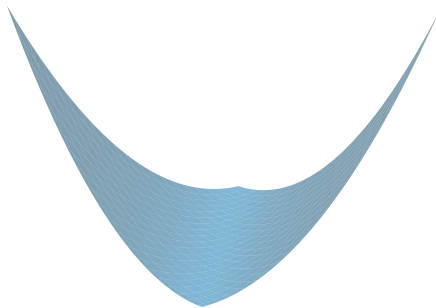
Conclusion : the set of least squares solutions is an affine sub-space

$$\boxed{\hat{\boldsymbol{\theta}} + \text{Ker}(X)}$$



## Optimization in $\mathbb{R}^d$

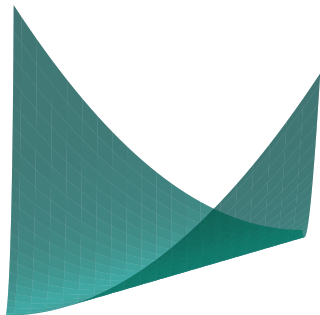
Convex case,  $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$ , where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

## Optimization in $\mathbb{R}^d$

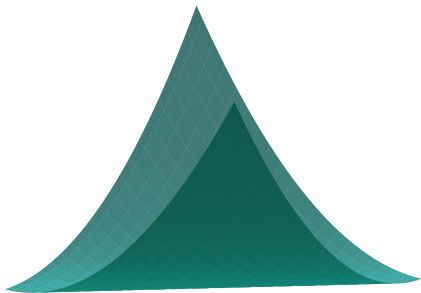
Convex case,  $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$ , where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

## Optimization in $\mathbb{R}^d$

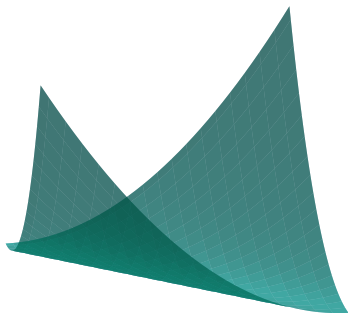
Convex case,  $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$ , where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

## Optimization in $\mathbb{R}^d$

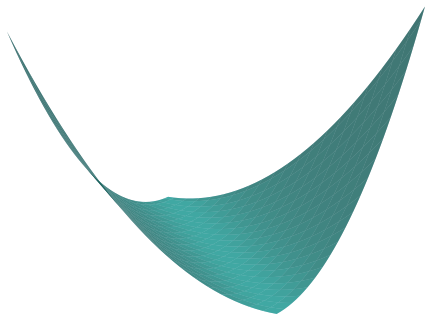
Convex case,  $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$ , where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

## Optimization in $\mathbb{R}^d$

Convex case,  $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$ , where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

## Non uniqueness : single feature case

Reminder :

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

If  $\text{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^2 : X\boldsymbol{\theta} = 0\} \neq \{0\}$  there exists  $(\theta_0, \theta_1) \neq (0, 0)$  :

$$\begin{cases} \theta_0 + \theta_1 x_1 & = 0 \\ \vdots & \vdots & = \vdots \\ \theta_0 + \theta_1 x_n & = 0 \end{cases} \quad (\star)$$

1. If  $\theta_1 = 0$  :  $(\star) \Rightarrow \theta_0 = 0$ , so  $(\theta_0, \theta_1) = (0, 0)$ , **contradiction**
2. If  $\theta_1 \neq 0$  :
  - 2.1 If  $\forall i, x_i = 0$  then  $X = (\mathbf{1}_n, 0)$  and  $\theta_0 = 0$
  - 2.2 Otherwise there exists  $x_{i_0} \neq 0$  and  $\forall i, x_i = -\theta_0/\theta_1 = x_{i_0}$ ,  
i.e.,  $X = [\mathbf{1}_n \quad x_{i_0} \cdot \mathbf{1}_n]$

Interpretation :  $\mathbf{x}_1 \propto \mathbf{1}_n$ , i.e.,  $\mathbf{x}_1$  is constant

# Interpretation for multivariate cases

Reminder : we write  $X = (\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p)$ , the features being column-wise (each are of length  $n$ )

The property  $\text{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$  means that there exists a linear dependence between the features

$\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p$ ,

Reformulation :  $\exists \boldsymbol{\theta} = (\theta_0, \dots, \theta_p)^\top \in \mathbb{R}^{p+1} \setminus \{0\}$  s.t.

$$\theta_0 \mathbf{1}_n + \sum_{j=1}^p \theta_j \mathbf{x}_j = 0$$

# Algebra reminder

## Definition

**Rank of a matrix :**  $\text{rank}(X) = \dim(\text{Span}(\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p))$  ;  
 $\text{Span}(\cdot)$  : the space generated by .

Property :  $\text{rank}(X) = \text{rank}(X^\top)$

## Rank-nullity theorem

$$\text{rank}(X) + \dim(\text{Ker}(X)) = p + 1$$

$$\text{rank}(X^\top) + \dim(\text{Ker}(X^\top)) = n$$

Rem:

$$\text{rank}(X) \leq \min(n, p + 1)$$

See **Golub and Van Loan (1996)** for details

---

**Exo:**  $\text{Ker}(X) = \text{Ker}(X^\top X)$

---



# Algebra reminder (continued)

## Matrix inversion

A square matrix  $A \in \mathbb{R}^{m \times m}$  is invertible

- ▶ if and only if its kernel is trivial :  $\text{Ker}(A) = \{0\}$
- ▶ if and only if it is full rank  $\text{rank}(A) = m$

---

**Exo:** Show that  $\text{Ker}(A) = \{0\}$  is equivalent to  $A^\top A$  invertible

---

# Outline

Introduction : OLS with two features

## Multivariate least square

- Matrix model

- Least squares definition

- Optimization

- Uniqueness issues

- Closed-form solution, prediction and residual

# Closed-form solution for least squares

## Closed-form solution for full rank matrix

If  $X$  is full (column) rank (i.e., if  $X^\top X$  is non-singular) then

$$\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Rem: recover the empirical mean if  $X = \mathbf{1}_n$  :  $\hat{\boldsymbol{\theta}} = \frac{\langle \mathbf{1}_n, \mathbf{y} \rangle}{\langle \mathbf{1}_n, \mathbf{1}_n \rangle} = \bar{y}_n$

Rem: for a single feature  $X = \mathbf{x} = (x_1, \dots, x_n)^\top$  :  $\hat{\boldsymbol{\theta}} = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|^2}, \mathbf{y} \right\rangle$

**Beware** : in practice **avoid** inverting the matrix  $X^\top X$  :

- ▶ this is numerically time consuming
- ▶ the matrix  $X^\top X$  might be big if “ $p \gg n$ ”, e.g., in biology  $n$  patients ( $\approx 100$ ),  $p$  genes ( $\approx 50000$ )

---

**Exo:** recover formula for 1D case with intercept

# Prediction

## Definition

**Prediction vector :**  $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}}$

Rem:  $\hat{\mathbf{y}}$  depends linearly on the observation vector  $\mathbf{y}$

Reminder : an **orthogonal projector** is a matrix  $H$  such that

1.  $H$  is symmetric :  $H^\top = H$
2.  $H$  is idempotent :  $H^2 = H$

## Proposition

Writing  $H_X$  the orthogonal projector onto the space span by the columns of  $X$ , one gets  $\hat{\mathbf{y}} = H_X \mathbf{y}$

Rem: if  $X$  is full (column) rank, then  $H_X = X(X^\top X)^{-1}X^\top$  is called the **hat matrix**

## Prediction (continued)

If a new observation  $x_{n+1} = (x_{n+1,1}, \dots, x_{n+1,p})$  is provided, the associated prediction is :

$$\hat{y}_{n+1} = \langle \hat{\boldsymbol{\theta}}, (1, x_{n+1,1}, \dots, x_{n+1,p})^\top \rangle$$

$$\hat{y}_{n+1} = \hat{\theta}_0 + \sum_{j=1}^p \hat{\theta}_j x_{n+1,j}$$

Rem: the normal equation ensures **equi-correlation** between observations and features :

$$\begin{aligned} (X^\top X) \hat{\boldsymbol{\theta}} &= X^\top \mathbf{y} \Leftrightarrow X^\top \hat{\mathbf{y}} = X^\top \mathbf{y} \\ &\Leftrightarrow \begin{pmatrix} \langle \mathbf{x}_0, \hat{\mathbf{y}} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \hat{\mathbf{y}} \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix} \end{aligned}$$

---

**Exo:** Let  $P = \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \in \mathbb{R}^{n \times n}.$

1. Check that  $P$  is an orthogonal projection matrix
2. Determine  $\text{Im}(P)$ , the range of  $P$
3. For  $\mathbf{x} = (x_1, \dots, x_n)^\top$ ,  $\bar{x}_n$  is the empirical mean and  $\sigma_{\mathbf{x}}$  is the standard deviation :

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \qquad \sigma_{\mathbf{x}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Show that  $\sigma_{\mathbf{x}} = \|(\text{Id}_n - P)\mathbf{x}\|/\sqrt{n}.$

---

# Residuals and normal equation

## Definition

$$\text{Residual(s)} : \quad \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\theta}} = (\text{Id}_n - H_X)\mathbf{y}$$

Reminder :

$$\text{Normal Equation : } \boxed{(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$$

Thanks to the residual definition, the later yields :

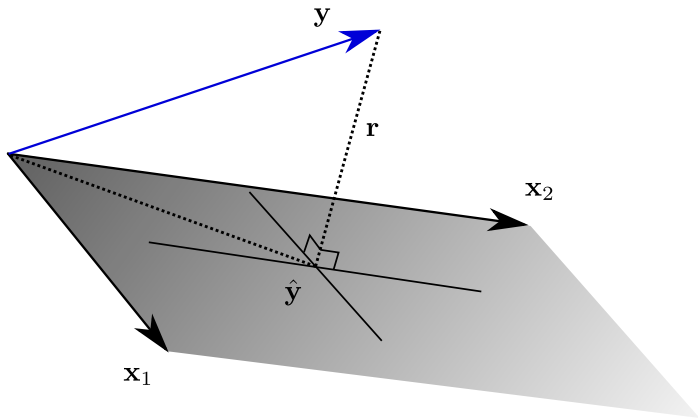
$$X^\top (X\hat{\boldsymbol{\theta}} - \mathbf{y}) = 0 \Leftrightarrow X^\top \mathbf{r} = 0 \Leftrightarrow \mathbf{r}^\top X = 0$$

With  $X = (\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p)$ , this can be rewritten

$$\forall j = 1, \dots, p : \langle \mathbf{r}, \mathbf{x}_j \rangle = 0 \text{ and } \bar{r}_n = 0$$

Interpretation : residuals are orthogonal to features

## Visualization : predictors and residuals ( $p = 2$ )





# References I

- ▶ G. H. Golub and C. F. van Loan.

*Matrix computations.*

Johns Hopkins University Press, Baltimore, MD, third edition, 1996.