# SD 204 : Linear model
# Properties of Ordinary Least Squares

**François Portier, Joseph Salmon**
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

# Outline

The fixed and random design models

Coefficient estimation

Noise level

Random design model

Miscellaneous

# Table of Contents

# The fixed design model

## Model I

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

$$x_i^\top = (1, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1}$$

$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ for } i = 1, \ldots, n$$

$$\mathbb{E}(\varepsilon) = 0, \; \mathrm{Var}(\epsilon) = \sigma^2$$

- $x_i$ is deterministic
- $\sigma^2$ is called the noise level

## Examples

- Physical experiment when the analyst is choosing the design *e.g.*,temperature of the experiment
- Some features are not random *e.g.*,time, location.

# The fixed design Gaussian model

## Model I with Gaussian noise

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

$$x_i^\top = (1, x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^{p+1}$$

$$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \text{ for } i = 1, \dots, n$$

## Examples

- Parametric model : specified by the two parameters $(\boldsymbol{\theta}, \sigma)$
- Strong assumption

# The random design model

**Model II**

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^{p+1} \\
(\varepsilon_i, x_i) &\overset{i.i.d}{\sim} (\varepsilon, x), \text{ for } i = 1, \dots, n \\
\mathbb{E}(\varepsilon|x) &= 0, \ \mathrm{Var}(\varepsilon|x) = \sigma^2
\end{aligned}
$$

Rem: here, the features are modelled as random (they might also suffer from some noise)

# The ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left( y_i - \theta_0 - \sum_{k=1}^{p} \theta_k x_{i,k} \right)^2$$

## How to deal with these two models ?

- The estimator is the same for both models
- The mathematics involved are different for each case
- The study of the fixed design case is easier as many closed formulas are available
- The two models lead to the same estimators of the variance $\sigma^2$

## Important formula

In both models, whenever $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times (p+1)}$ has full rank,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}$$

# Table of Contents

# Bias

**Proposition**

Under model I, whenever the matrix $X$ has full rank, the least squares estimator is unbiased, i.e.,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^\star$$

Proof :

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

# Bias

## Proposition

Under model I, whenever the matrix $X$ has full rank, the least squares estimator is unbiased, i.e.,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^\star$$

<u>Proof</u> :

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

$$B = \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon)) - \boldsymbol{\theta}^\star$$

# Bias

**Proposition**

Under model I, whenever the matrix $X$ has full rank, the least squares estimator is unbiased, i.e.,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^\star$$

Proof :

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

$$B = \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon})) - \boldsymbol{\theta}^\star$$

$$B = (X^\top X)^{-1} X^\top X \boldsymbol{\theta}^\star + (X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon) - \boldsymbol{\theta}^\star = 0$$

# Bias

**Proposition**

Under model I, whenever the matrix $X$ has full rank, the least squares estimator is unbiased, i.e.,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^\star$$

Proof :

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

$$B = \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon})) - \boldsymbol{\theta}^\star$$

$$B = (X^\top X)^{-1} X^\top X \boldsymbol{\theta}^\star + (X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star = 0$$

# Quadratic risk

## Definition

The **quadratic** risk is given by

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$$

where $\|\cdot\|$ is the Euclidean norm

## Bias/Variance decomposition

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

Proof :

$$\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2$$

# Quadratic risk

## Definition

The **quadratic** risk is given by

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$$

where $\|\cdot\|$ is the Euclidean norm

## Bias/Variance decomposition

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

Proof :

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
=& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
& + 2\mathbb{E}\langle \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\rangle
\end{aligned}$$

# Quadratic risk

## Definition

The **quadratic** risk is given by

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$$

where $\|\cdot\|$ is the Euclidean norm

## Bias/Variance decomposition

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

Proof :

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\ =& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\ &+ 2\mathbb{E}\langle \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\rangle \\ =& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2\end{aligned}$$

# Quadratic risk

## Definition

The **quadratic** risk is given by

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$$

where $\|\cdot\|$ is the Euclidean norm

## Bias/Variance decomposition

$$\boxed{\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2}$$

Proof :

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2 =& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
=& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2 \\
&+ 2\mathbb{E}\langle \mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\rangle \\
=& \mathbb{E}\|\boldsymbol{\theta}^\star - \mathbb{E}(\hat{\boldsymbol{\theta}})\|^2 + \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2
\end{aligned}$$

# Bias/Variance decomposition

Reminder : as the bias vanishes when $X$ has full rank,

$$\mathbb{E}\|\boldsymbol{\theta}^{\star} - \hat{\boldsymbol{\theta}}\|^2 = \mathbb{E}\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}\|^2$$

# The trace of a matrix

**Definition**

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of $A$ is the sum of the diagonal elements of $A$ and is denoted by $\operatorname{tr}(A)$ :

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n \times n}$, and $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linearity)

# The trace of a matrix

**Definition**

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of $A$ is the sum of the diagonal elements of $A$ and is denoted by $\operatorname{tr}(A)$ :

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n \times n}$, and $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linearity)
- $\operatorname{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$

# The trace of a matrix

## Definition

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of $A$ is the sum of the diagonal elements of $A$ and is denoted by $\text{tr}(A)$ :

$$\text{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\text{tr}(A) = \text{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n \times n}$, and $\alpha \in \mathbb{R}$, $\text{tr}(\alpha A + B) = \alpha\,\text{tr}(A) + \text{tr}(B)$ (linearity)
- $\text{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- For any $A, B \in \mathbb{R}^{n \times n}$, $\text{tr}(AB) = \text{tr}(BA)$

# The trace of a matrix

**Definition**

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of $A$ is the sum of the diagonal elements of $A$ and is denoted by $\operatorname{tr}(A)$ :

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n \times n}$, and $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linearity)
- $\operatorname{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- For any $A, B \in \mathbb{R}^{n \times n}$, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$
- $\operatorname{tr}(PAP^{-1}) = \operatorname{tr}(A)$, hence if $A$ is diagonalisable, the trace is the sum of the eigenvalues

# The trace of a matrix

## Definition

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of $A$ is the sum of the diagonal elements of $A$ and is denoted by $\mathrm{tr}(A)$ :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n \times n}$, and $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linearity)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- For any $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, hence if $A$ is diagonalisable, the trace is the sum of the eigenvalues
- If $H$ is an orthogonal projector $\mathrm{tr}(H) = \mathrm{rank}(H)$

# The trace of a matrix

**Definition**

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of $A$ is the sum of the diagonal elements of $A$ and is denoted by $\operatorname{tr}(A)$ :

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n \times n}$, and $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linearity)
- $\operatorname{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- For any $A, B \in \mathbb{R}^{n \times n}$, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$
- $\operatorname{tr}(PAP^{-1}) = \operatorname{tr}(A)$, hence if $A$ is diagonalisable, the trace is the sum of the eigenvalues
- If $H$ is an orthogonal projector $\operatorname{tr}(H) = \operatorname{rank}(H)$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon) - \boldsymbol{\theta}^\star)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon) - \boldsymbol{\theta}^\star)\right]$$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top \varepsilon)^\top ((X^\top X)^{-1} X^\top \varepsilon)\right] = \mathbb{E}(\varepsilon^\top X (X^\top X)^{-2} X^\top \varepsilon)$$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})^\top((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-2}X^\top\boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\varepsilon^\top X(X^\top X)^{-1}(X^\top X)^{-1}X^\top\varepsilon)] \text{ (thx to } \operatorname{tr}(u^\top u) = u\top u)$$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})^\top((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-2}X^\top\boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})] \text{ (thx to } \operatorname{tr}(u^\top u) = u\top u)$$

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1}X^\top\varepsilon\varepsilon^\top X(X^\top X)^{-1}\right]\right)$$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top ((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-2} X^\top \boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})] \text{ (thx to } \operatorname{tr}(u^\top u) = u \top u)$$

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1}\right]\right)$$

$$= \operatorname{tr}[(X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) X (X^\top X)^{-1}]$$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})^\top((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-2}X^\top\boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})]$$ (thx to $\operatorname{tr}(u^\top u) = u\top u$)

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}\right]\right)$$

$$= \operatorname{tr}[(X^\top X)^{-1}X^\top\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)X(X^\top X)^{-1}]$$

$$= \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

# Estimation risk

**Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

Proof :

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top((X^\top X)^{-1}X^\top(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}\left[((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})^\top((X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-2}X^\top\boldsymbol{\varepsilon})$$

$$= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon})] \text{ (thx to } \operatorname{tr}(u^\top u) = u \top u)$$

$$= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1}X^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}\right]\right)$$

$$= \operatorname{tr}[(X^\top X)^{-1}X^\top\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)X(X^\top X)^{-1}]$$

$$= \sigma^2 \operatorname{tr}((X^\top X)^{-1})$$

# Prediction risk

Prediction risk (normalized) $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Under model I, whenever the matrix $X$ has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\text{rank}(X)}{n}$$

Because $X$ has full rank, $\text{rank}(X) = p + 1$.

Proof : As before

$$n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right]$$

$$= \mathbb{E}(\varepsilon^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \varepsilon)$$

# Prediction risk

**Prediction risk (normalized)** $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Under model I, whenever the matrix $X$ has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\text{rank}(X)}{n}$$

Because $X$ has full rank, $\text{rank}(X) = p + 1$.

Proof : As before

$$\begin{aligned} n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \end{aligned}$$

# Prediction risk

**Prediction risk (normalized) $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\operatorname{rank}(X)}{n}$$

Because $X$ has full rank, $\operatorname{rank}(X) = p + 1$.

Proof : As before

$$\begin{aligned} n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \end{aligned}$$

# Prediction risk

**Prediction risk (normalized)** $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Under model I, whenever the matrix $X$ has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\operatorname{rank}(X)}{n}$$

Because $X$ has full rank, $\operatorname{rank}(X) = p + 1$.

Proof : As before

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \\
&= \operatorname{tr}[\mathbb{E}(H_X \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H_X^\top)] = \operatorname{tr}\left(H_X \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) H_X^\top\right)
\end{aligned}$$

# Prediction risk

**Prediction risk (normalized)** $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Under model I, whenever the matrix $X$ has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\text{rank}(X)}{n}$$

Because $X$ has full rank, $\text{rank}(X) = p + 1$.

Proof : As before

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \text{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \text{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \\
&= \text{tr}[\mathbb{E}(H_X \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H_X^\top)] = \text{tr}\left(H_X \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) H_X^\top\right) \\
&= \sigma^2 \,\text{tr}(H_X) = \sigma^2 \,\text{rank}(H_X) = \sigma^2 \,\text{rank}(X)
\end{aligned}$$

# Prediction risk

**Prediction risk (normalized) $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$**

Under model I, whenever the matrix $X$ has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\operatorname{rank}(X)}{n}$$

Because $X$ has full rank, $\operatorname{rank}(X) = p + 1$.

Proof : As before

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H_X^\top H_X \boldsymbol{\varepsilon})] \\
&= \operatorname{tr}[\mathbb{E}(H_X \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H_X^\top)] = \operatorname{tr}\left(H_X \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) H_X^\top\right) \\
&= \sigma^2 \operatorname{tr}(H_X) = \sigma^2 \operatorname{rank}(H_X) = \sigma^2 \operatorname{rank}(X)
\end{aligned}$$

# Covariance matrix

## Covariance of $\hat{\boldsymbol{\theta}}$

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Proof :

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}})$$
$$= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\right]$$
$$= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon) - \boldsymbol{\theta}^\star)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon) - \boldsymbol{\theta}^\star)^\top\right]$$

# Covariance matrix

**Covariance of $\hat{\boldsymbol{\theta}}$**

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Proof :

$$\begin{aligned}
&\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \varepsilon)((X^\top X)^{-1} X^\top \varepsilon)^\top\right]
\end{aligned}$$

# Covariance matrix

## Covariance of $\hat{\boldsymbol{\theta}}$

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Proof :

$$\begin{aligned}
&\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1}
\end{aligned}$$

# Covariance matrix

## Covariance of $\hat{\boldsymbol{\theta}}$

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Proof :

$$\begin{aligned}
&\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1} \\
&= (X^\top X)^{-1} X^\top (\sigma^2 \mathrm{Id}_n) X (X^\top X)^{-1}
\end{aligned}$$

# Covariance matrix

## Covariance of $\hat{\boldsymbol{\theta}}$

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

<u>Proof</u> :

$$\begin{aligned}
&\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1} \\
&= (X^\top X)^{-1} X^\top (\sigma^2 \, \mathrm{Id}_n) X (X^\top X)^{-1} \\
&= \sigma^2 (X^\top X)^{-1}
\end{aligned}$$

# Covariance matrix

## Covariance of $\hat{\boldsymbol{\theta}}$

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Proof :

$$\begin{aligned}
&\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top\right] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] X (X^\top X)^{-1} \\
&= (X^\top X)^{-1} X^\top (\sigma^2 \,\mathrm{Id}_n) X (X^\top X)^{-1} \\
&= \sigma^2 (X^\top X)^{-1}
\end{aligned}$$

# Table of Contents

# Estimation of the noise level

- An estimator of the noise level $\sigma^2$ is given by
$$\boxed{\frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

- Another estimator which is unbiased is defined by
$$\boxed{\hat{\sigma}^2 = \frac{1}{n - \mathrm{rank}(X)}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

# Estimation of the noise level

**$\hat{\sigma}^2$ is unbiased**

Under model I, whenever the matrix $X$ has full rank, we have

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2$$

Proof :

$$\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2 = \mathbf{y}^\top(\mathrm{Id}_n - H_X)\mathbf{y} = \boldsymbol{\varepsilon}^\top(\mathrm{Id}_n - H_X)\boldsymbol{\varepsilon} = \mathrm{tr}((\mathrm{Id}_n - H_X)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)$$

# Heteroscedasticity

Model I and Model II are homoscedastic models, *i.e.*,we assume that the noise level $\sigma^2$ does not depend on $x_i$

Heteroscedastic Model : we allow $\sigma^2$ to change with the observation $i$, we denote by $\sigma_i^2 > 0$ the associated variance

$$\hat{\boldsymbol{\theta}} \in \mathop{\arg\min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left( \frac{y_i - \langle \boldsymbol{\theta}, x_i \rangle}{\sigma_i} \right)^2 = \mathop{\arg\min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} (y - X\boldsymbol{\theta})^\top \Omega (y - X\boldsymbol{\theta})$$

with $\Omega = \operatorname{diag}(\frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_n^2})$

---

**Exo**: give a closed formula for $\hat{\boldsymbol{\theta}}$ when $X^\top \Omega X$ has full rank

---

---

**Exo**: give a necessary and sufficient condition for $X^\top \Omega X$ to be invertible

---

# Gaussian model

## Proposition

Under model I with Gaussian noise, whenever the matrix $X$ has full rank, we have

(i) $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$ are independent random variables

(ii) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim \mathcal{N}(0, \sigma^2 (X^\top X/n)^{-1})$ for every $n$

(iii) $(n - \mathrm{rank}(X))\frac{\hat{\sigma}^2}{\sigma^{*2}} \sim \chi^2_{n-\mathrm{rank}(X)}$ for every $n$

(iv) Let $\hat{s}_k = (X^\top X/n)^{-1}_{k,k}$,

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta^*}{\sqrt{\hat{s}_k \hat{\sigma}^2}}\right) \sim \mathcal{T}_{n-\mathrm{rank}(X)}$$

where $\mathcal{T}_{n-\mathrm{rank}(X)}$ stands for a student distribution with $n - \mathrm{rank}(X)$ degrees of freedom

# Table of Contents

# Bias and variance

**Proposition**

Under model II, whenever the matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ has full rank, we have

$$\mathbb{E}(\hat{\boldsymbol{\theta}} \mid X) = \boldsymbol{\theta}^\star$$
$$\mathrm{Var}(\hat{\boldsymbol{\theta}} \mid X) = (X^\top X)^{-1}\sigma^2$$

Proof : The same as in the case of fixed design with the conditional expectation

Rem:We cannot compute the $\mathbb{E}(\hat{\boldsymbol{\theta}})$ nor $\mathrm{Var}(\hat{\theta})$ because the matrix $X$ has full rank is now random !
Rem:One solution is to rely on asymptotic convergence

# Asymptotics

## Asymptotics of $\hat{\boldsymbol{\theta}}$

Under model II, whenever the covariance matrix $\operatorname{cov}(X)$ has full rank, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2 S^{-1})$$

with $S = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$

<u>Outline of the proof</u> : It could happen that $\hat{\boldsymbol{\theta}}$ is not uniquely defined, so we put

$$\hat{\boldsymbol{\theta}} = \left(X^\top X\right)^+ X^\top Y$$

where $A^+$ is the generalized inverse of $A$

- With high probability, we have that $X^\top X$ is invertible because $\frac{X^\top X}{n} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ goes to $S$

# Asymptotics

Outline of the proof :

- As a consequence, in the asymptotics we can replace $\left(X^\top X\right)^+$ by $\left(X^\top X\right)^{-1}$ (that we shall admit)

Then we use that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) = \left(\frac{X^\top X}{n}\right)^{-1} \left(\frac{X^\top \epsilon}{\sqrt{n}}\right)$$

- The term on the right $\frac{X^\top \varepsilon}{\sqrt{n}}$ converges to $\mathcal{N}(0, \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\sigma^2)$ in distribution
- The term on the left $\left(\frac{X^\top X}{n}\right)^{-1}$ goes to $S^{-1}$ in probability

# Asymptotics

- In the random design model, since closed formulas for the bias and variance of $\boldsymbol{\theta}$ are lacking ; Asymptotics is used to validate the procedure and to build-up the variance estimator

**Variance estimation**

By the previous Proposition, the variance to estimate is

$$\sigma^2 S^{-1}$$

a natural "Plug-in" estimator is

$$\hat{\sigma}^2 \hat{S}_n^{+}$$

with $\hat{\sigma}^2 = \frac{1}{n - \mathrm{rank}(X)} \|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2$

Rem:It coincides with the estimator in the case of fixed design

# Variance estimation

## Noise level is conditionally unbiased

Under model $\textsc{ii}$, whenever the matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ has full rank, we have

$$\mathbb{E}(\hat{\sigma}^2 \mid X) = \sigma^2$$

---

**Exo**: Write the proof

---

## Convergence of the variance estimator

Under model $\textsc{ii}$, if the covariance matrix $\operatorname{cov}(X)$ has full rank, we have

$$\hat{\sigma}^2 \hat{S}_n^+ \to \sigma^2 S^{-1}$$

in probability

# Table of Contents

# Qualitative variables

A variable is qualitative, when its state space is discrete (non-necessarily numeric)

Exemple : colors, gender, cities, etc.

Classically : "One-hot encoder" consists in representing a qualitative variable with several dummy variables (valued in $\{0, 1\}$)

If each $x_i$ is valued in $a_1, \ldots, a_K$, we define the following $K$ explanatory variables : $\forall k \in [\![1, K]\!], \mathbb{1}_{a_k} \in \mathbb{R}^n$ is given by

$$\forall i \in [\![1, n]\!], \quad (\mathbb{1}_{a_k})_i = \begin{cases} 1, & \text{if } x_i = a_k \\ 0, & \text{else} \end{cases}$$

# Examples

Binary case : M/F, yes/no, I like it/I don't.

| Client | Gender |
|---|---|
| 1 | H |
| 2 | F |
| 3 | H |
| 4 | F |
| 5 | F |

$$\longrightarrow \begin{pmatrix} F & H \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

General case : colors, cities, etc.

| Client | Colors |
|---|---|
| 1 | Blue |
| 2 | Blanc |
| 3 | Red |
| 4 | Red |
| 5 | Blue |

$$\longrightarrow \begin{pmatrix} \text{Blue} & \text{Blanc} & \text{Red} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

# Somme difficulties

Correlations : $\sum_{k=1}^{K} \mathbb{1}_{a_k} = \mathbf{1}_n$ ! We can drop-off one modality (*e.g.*,`drop_first=True` dans `get_dummies` de `pandas`)

Without intercept, with all modalities : $X = [\mathbb{1}_{a_1}, \ldots, \mathbb{1}_{a_K}]$. If $x_{n+1} = a_k$ then $\hat{y}_{n+1} = \hat{\boldsymbol{\theta}}_k$

With intercept, with one less modality : $X = [\mathbf{1}_n, \mathbb{1}_{a_2}, \ldots, \mathbb{1}_{a_K}]$, dropping-off the first modality

If $x_{n+1} = a_k$ then $\hat{y}_{n+1} = \begin{cases} \hat{\boldsymbol{\theta}}_0, & \text{if } k = 1 \\ \hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_k, & \text{else} \end{cases}$

Rem: might give null column in Cross-Validation (if a modality is not present in a CV-fold)

Rem: penalization might help (*e.g.*,Lasso, Ridge)

---

**Exo**: Compute the OLS for $X = [\mathbb{1}_{a_1}, \ldots, \mathbb{1}_{a_K}] \in \mathbb{R}^{n \times K}$

---

# What if $n < p$?

Many of the things presented before need to be adapted

For instance : if $\text{rank}(X) = n$, then $H_X = \text{Id}_n$ and $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}} = \mathbf{y}$ !
The vector space generated by the columns $[\mathbf{x}_0, \ldots, \mathbf{x}_p]$ is $\mathbb{R}^n$, making the observed signal and predicted signal are **identical**

Rem: typical kind of problem in large dimension (when $p$ is large)

Possible solution : variable selection, *cf.*Lasso and greedy methods (coming soon)

# Web sites and books

- Python Packages for OLS :
  `statsmodels`
  `sklearn.linear_model.LinearRegression`
- McKinney (2012) about `python` for statistics
- Lejeune (2010) about the Linear Model
- Delyon (2015) Advanced course on regression
  https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf

# References I


- B. Delyon.
  Régression, 2015.
  https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf.
- M. Lejeune.
  *Statistiques, la théorie et ses applications*.
  Springer, 2010.
- W. McKinney.
  *Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython*.
  O'Reilly Media, 2012.