SD 204 : Linear model
Properties of Ordinary Least Squares

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

Plan

The fixed and random design models

Estimation of $\boldsymbol{\theta}$

- Bias
- Estimation and prediction risk
- Variance

Noise level

- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model

- Bias and variance

Miscellaneous

- Qualitative variables
- Large dimension $p > n$

The fixed design model

Model I

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

- x_i is deterministic
- σ^2 is called the noise level

Examples

- Physical experiment when the annalist 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

- Its a parametric model because specified by the two parameters $(\boldsymbol{\theta}, \sigma)$
- Strong assumption

The random design model

Model II

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

$$\mathbf{x}_i^\top = (1, \mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,p}) \in \mathbb{R}^{p+1}$$

$$(\varepsilon_i, \mathbf{x}_i) \overset{i.i.d}{\sim} (\varepsilon, \mathbf{x}), \text{ for } i = 1, \dots, n$$

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

Examples

- As soon as the variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ are unpredictable, the random modelling is justified, e.g., wind speed, random survey

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 this two models ?

- The estimator is the same for both models
- The mathematics involved are different in either cases
- 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 σ^2

Important formula

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

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

Sommaire

The fixed and random design models

Estimation of $\boldsymbol{\theta}$

- Bias
- Estimation and prediction risk
- Variance

Noise level

- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model

- Bias and variance

Miscellaneous

- Qualitative variables
- Large dimension $p > n$

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.,

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

Proof :

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

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

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 :

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

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,

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

Sommaire

The fixed and random design models

Estimation of $\boldsymbol{\theta}$

- Bias
- Estimation and prediction risk
- Variance

Noise level

- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model

- Bias and variance

Miscellaneous

- Qualitative variables
- Large dimension $p > n$

The trace of a matrix

Définition

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

Définition

Let $A \in \mathbb{R}^{n\times n}$ denote a matrix. The **trace** of A, denoted $\mathrm{tr}(A)$ is the sum of the diagonal element 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)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$

The trace of a matrix

Définition

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix. The **trace** of A, denoted $\mathrm{tr}(A)$ is the sum of the diagonal element 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)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- For any $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

The trace of a matrix

Définition

Let $A \in \mathbb{R}^{n\times n}$ denote a matrix. The **trace** of A, denoted $\mathrm{tr}(A)$ is the sum of the diagonal element 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)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- 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

The trace of a matrix

Définition

Let $A \in \mathbb{R}^{n\times n}$ denote a matrix. The **trace** of A, denoted $\mathrm{tr}(A)$ is the sum of the diagonal element 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)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- 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 $\mathrm{tr}(H) = \mathrm{rang}(H)$

The trace of a matrix

Définition

Let $A \in \mathbb{R}^{n\times n}$ denote a matrix. The **trace** of A, denoted $\mathrm{tr}(A)$ is the sum of the diagonal element 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)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} A_{i,j}^2$
- 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 $\mathrm{tr}(H) = \mathrm{rang}(H)$

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

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

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

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

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

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

Proof :

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

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

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

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

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

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

Proof :

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

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

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

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

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

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

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

Proof :

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

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

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

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

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

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

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

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

Proof :

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

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

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

Proof :

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

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

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

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

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

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

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

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

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

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

Proof : As before

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

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

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

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

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

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

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

Proof : As before

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

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

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

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

<u>Proof</u> : As before

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

Sommaire

The fixed and random design models

Estimation of θ

- Bias
- Estimation and prediction risk
- Variance

Noise level

- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model

- Bias and variance

Miscellaneous

- Qualitative variables
- Large dimension $p > n$

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 :

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

Covariance matrix

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

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

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

Proof :

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

Covariance matrix

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

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

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

Proof :

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

Covariance matrix

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

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

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

<u>Proof</u> :

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

Sommaire

The fixed and random design models

Estimation of $\boldsymbol{\theta}$
- Bias
- Estimation and prediction risk
- Variance

Noise level
- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model
- Bias and variance

Miscellaneous
- Qualitative variables
- Large dimension $p > n$

Estimation of the noise level

- An estimator of the noise level σ^2 is given by

$$\boxed{\frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

- Another estimator which is unbiased is defined by

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

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 :

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

Sommaire

The fixed and random design models

Estimation of $\boldsymbol{\theta}$
- Bias
- Estimation and prediction risk
- Variance

Noise level
- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model
- Bias and variance

Miscellaneous
- Qualitative variables
- Large dimension $p > n$

Heteroscedasticity

Model I and Model II are homoscedastic model, i.e., we assume that the noise level σ^2 does not depend on x_i

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

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

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

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

---

Sommaire

The fixed and random design models

Estimation of $\boldsymbol{\theta}$
- Bias
- Estimation and prediction risk
- Variance

Noise level
- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model
- Bias and variance

Miscellaneous
- Qualitative variables
- Large dimension $p > n$

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

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

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

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

Sommaire

The fixed and random design models

Estimation of θ

- Bias
- Estimation and prediction risk
- Variance

Noise level

- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model

- Bias and variance

Miscellaneous

- Qualitative variables
- Large dimension $p > n$

Bias and variance

Proposition

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

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

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

Outline of the proof : 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 $X^\top X / n = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ 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

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

- The term on the right $X^\top \epsilon/\sqrt{n}$ converges to $\mathcal{N}(0, \mathbb{E}[\mathbf{x}\mathbf{x}^T]\sigma^2)$ in distribution
- The term on the left $\left(X^\top X/n\right)^{-1}$ 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 $\hat{\sigma}^2 = \frac{1}{n - \mathrm{rg}(X)} \|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2$

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

Variance estimation

Noise level is conditionally unbiased

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

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

in probability

Sommaire

The fixed and random design models

Estimation of $\boldsymbol{\theta}$

- Bias
- Estimation and prediction risk
- Variance

Noise level

- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model

- Bias and variance

Miscellaneous

- Qualitative variables
- Large dimension $p > n$

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 explicative 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

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

General case : colors, cities, etc.

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

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

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 gives 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 $X = [\mathbb{1}_{a_1}, \ldots, \mathbb{1}_{a_K}] \in \mathbb{R}^{n \times K}$

Sommaire

The fixed and random design models

Estimation of θ
- Bias
- Estimation and prediction risk
- Variance

Noise level
- Estimation of the noise level
- Heteroscedasticity
- Gaussian noise

Random design model
- Bias and variance

Miscellaneous
- Qualitative variables
- Large dimension $p > n$

What if $n < p$?

Many of the things presented before needs to be adapted

For instance : if $\mathrm{rg}(X) = n$, then $H_X = \mathrm{Id}_n$ 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: This is the 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`