MS BGD
MDI 720 : Statistiques

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

Plan

Analyse de performance
- Biais
- Variance

Impact du niveau de bruit
- Estimation du niveau de bruit
- Cas hétéroscédastique

Aparté
- Variables qualitatives
- Grande dimension $p > n$

Sommaire

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

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

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

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

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

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

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

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

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

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

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

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

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

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

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

$$B = 0 \quad \text{(bruit centré)}$$

Biais

Rappel : $\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}$

Proposition

Sous l'hypothèse que $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ et que la matrice X est de plein rang, alors l'estimateur des moindres carrés est sans biais :

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

Rem: l'hypothèse $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ signifie que $\forall i \in [\![1, n]\!], \mathbb{E}(\varepsilon_i) = 0$

Démonstration :

$$\begin{aligned}
B =& \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^\star = \mathbb{E}((X^\top X)^{-1} X^\top \mathbf{y}) - \boldsymbol{\theta}^\star \text{(plein rang)} \\
B =& \mathbb{E}((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon})) - \boldsymbol{\theta}^\star \\
B =& (X^\top X)^{-1} X^\top X\boldsymbol{\theta}^\star + (X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star \\
B =& 0 \quad \text{(bruit centré)}
\end{aligned}$$

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

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

Décomposition biais/variance

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

<u>Démonstration</u> :

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

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

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

Décomposition biais/variance

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

Démonstration :

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

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

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

Décomposition biais/variance

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

<u>Démonstration</u> :

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

Risque quadratique

Définition

Le **risque quadratique** est la quantité suivante :

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

Décomposition biais/variance

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

<u>Démonstration</u> :

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

Décomposition biais/variance

Rappel : pour les moindres carrés le biais est nul sous l'hypothèse que X est de plein rang et que le bruit est centré.

Ainsi

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^\star = 0$$

et

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

Sommaire

Intermède sur la trace

Définition

Soit $A \in \mathbb{R}^{n \times n}$ une matrice carrée. La **trace** de A, notée $\operatorname{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linéarité)

Intermède sur la trace

Définition

Soit $A \in \mathbb{R}^{n \times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$

Intermède sur la trace

Définition

Soit $A \in \mathbb{R}^{n \times n}$ une matrice carrée. La **trace** de A, notée $\operatorname{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

- $\operatorname{tr}(A) = \operatorname{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$, $\operatorname{tr}(\alpha A + B) = \alpha \operatorname{tr}(A) + \operatorname{tr}(B)$ (linéarité)
- $\operatorname{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$

Intermède sur la trace

Définition

Soit $A \in \mathbb{R}^{n \times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\, \mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, donc si A est diagonalisable, sa trace est la somme de ses valeurs propres

Intermède sur la trace

Définition

Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, donc si A est diagonalisable, sa trace est la somme de ses valeurs propres
- Si H est un projecteur orthogonal $\mathrm{tr}(H) = \mathrm{rang}(H)$

Intermède sur la trace

Définition

Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée. La **trace** de A, notée $\mathrm{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, et tout $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linéarité)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^n \sum_{j=1}^n A_{i,j}^2 := \|A\|_F^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, donc si A est diagonalisable, sa trace est la somme de ses valeurs propres
- Si H est un projecteur orthogonal $\mathrm{tr}(H) = \mathrm{rang}(H)$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration :

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

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

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration :

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

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

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

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration :

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

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

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

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

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration :

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

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

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

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

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

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \operatorname{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration :

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

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

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

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

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

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

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \,\mathrm{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration :

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

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

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

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

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

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

$$= \sigma^2 \,\mathrm{tr}\left[(X^\top X)^{-1}\right]$$

Risque d'estimation

Hypothèse de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \,\mathrm{Id}_n$

Risque d'estimation $\mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Si le modèle est homoscédastique et que X est de plein rang

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

<u>Démonstration</u> :

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

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

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

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

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

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

$$= \sigma^2 \,\mathrm{tr}\left[(X^\top X)^{-1}\right]$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

Démonstration : début identique

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

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

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

Démonstration : début identique

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

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

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

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

Démonstration : début identique

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

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

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

$$= \mathbb{E}(\varepsilon^\top H_X^2 \varepsilon) = \mathbb{E}(\varepsilon^\top H_X \varepsilon)$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

Démonstration : début identique

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

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

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

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

Démonstration : début identique

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

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

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

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

<u>Démonstration</u> : début identique

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

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

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

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque de prédiction (normalisé) $\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Si le modèle est homoscédastique

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \frac{\mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2}{n} = \frac{\operatorname{rang}(X)}{n}\sigma^2$$

- l'erreur est proportionnelle au niveau de bruit σ^2
- l'erreur est proportionnelle à $1/n$ (n : taille échantillon)
- l'erreur est proportionnelle à $\operatorname{rang}(X)$ ($\operatorname{rang}(X)$: nombre de variables explicatives indépendantes) ;

Attention si $\operatorname{rang}(X) \approx n$, l'erreur n'est pas petite...

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Si le modèle est homoscédastique et que X est de plein rang

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

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Si le modèle est homoscédastique et que X est de plein rang

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

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

Si le modèle est homoscédastique et que X est de plein rang

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

Démonstration : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

Sommaire

Estimateur du niveau de bruit

- On peut construire un estimateur de le variance σ^2 du bruit :

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

ou si l'on souhaite un estimateur sans biais :

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

- Motivation "débiaisage" : théorie des tests

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \mathbf{y}^\top(\mathrm{Id}_n - H_X)\mathbf{y} = \boldsymbol{\varepsilon}^\top(\mathrm{Id}_n - H_X)\boldsymbol{\varepsilon} = \sum_{i=1}^{n-\mathrm{rg}(X)} \tilde{\varepsilon}_i^2$$

Cas gaussien : si $\varepsilon_i \overset{\textit{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, alors $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$ suit une loi du χ^2 à $n - \mathrm{rg}(X)$ degrés de liberté

<u>Rem</u>: implicitement on fait donc encore l'hypothèse $n > p$

Estimateur du niveau de bruit (II)

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

<u>Preuve</u> :

$$\begin{aligned}
\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2 =& \mathbf{y}^\top (\mathrm{Id}_n - H_X)\mathbf{y} \\
\mathbb{E}(\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2) =& \mathbb{E}\left[(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon})^\top (\mathrm{Id}_n - H_X)(X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon})\right]
\end{aligned}$$

Comme $(\mathrm{Id}_n - H_X)X = 0$ et $\mathbb{E}(\boldsymbol{\varepsilon}^\top X\boldsymbol{\theta}^\star) = 0$ on obtient :

$$\begin{aligned}
\mathbb{E}(\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2) =& \mathbb{E}\left[\boldsymbol{\varepsilon}^\top (\mathrm{Id}_n - H_X)\boldsymbol{\varepsilon}\right] \\
=& \sigma^2 \operatorname{tr}(\mathrm{Id}_n - H_X) \\
=& \sigma^2 (n - \mathrm{rang}(X))
\end{aligned}$$

Sommaire

Cas hétéroscédastique

L'estimateur MCO $\hat{\boldsymbol{\theta}}$ postule implicitement que les variables $y_1, \dots, y_n$ ont même niveau de bruit

Rem: pour cela reprendre le calcul du maximum de vraisemblance d'un modèle gaussien avec variance σ^2 fixée / connue

Modèle hétéroscédastique : on suppose que le niveau de bruit diffère pour chaque y_i et on note σ_i^2 la variance associée

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

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

Exo: donner une formule explicite si $X^\top \Omega X$ est de plein rang

---

Sommaire

Variables qualitatives

On parle de variable **qualitative**, quand une variable ne prend que des modalités discrètes et/ou non-numériques.

Exemple : couleurs, genre, ville, etc.

Encodage classique : variables fictives/indicatrices
(: *dummy variables*)="encodage à chaud"
(: *one-hot encoder*).

Si la variable $\mathbf{x} = (x_1, \ldots, x_n)^\top$ peut prendre K modalités $a_1, \ldots, a_K$ on crée K variables : $\forall k \in [\![1, K]\!], \mathbb{1}_{a_k} \in \mathbb{R}^n$ définies par

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

et on remplace $\mathbf{x} \in \mathbb{R}^n$ par $[\mathbb{1}_{a_1}, \ldots, \mathbb{1}_{a_K}] \in \mathbb{R}^{n \times K}$

Exemple d'encodage

Cas binaire : M/F, oui/non, j'aime/j'aime pas.

Client	Genre
1	H
2	F
3	H
4	F
5	F

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

Cas général : couleur, villes, etc.

Client	Couleurs
1	Bleu
2	Blanc
3	Rouge
4	Rouge
5	Bleu

$$\longrightarrow \quad \begin{pmatrix} \text{Bleu} & \text{Blanc} & \text{Rouge} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Quelques difficultés

Corrélations : $\sum_{k=1}^{K} \mathbb{1}_{a_k} = \mathbf{1}_n$! On peut enlever une des modalités (*e.g.*, `drop_first=True` dans `get_dummies` de `pandas`)

Interprétation sans constante et avec toutes les modalités :
$X = [\mathbb{1}_{a_1}, \ldots, \mathbb{1}_{a_K}]$. Si $x_{n+1} = a_k$ alors $\hat{y}_{n+1} = \hat{\boldsymbol{\theta}}_k$

Interprétation avec constante et avec une modalité en moins :
$X = [\mathbf{1}_n, \mathbb{1}_{a_2}, \ldots, \mathbb{1}_{a_K}]$, en enlevant la première modalité

Si $x_{n+1} = a_k$ alors $\hat{y}_{n+1} = \begin{cases} \hat{\boldsymbol{\theta}}_0, & \text{si } k = 1 \\ \hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_k, & \text{sinon} \end{cases}$

Rem: création possible d'une colonne nulle par validation croisée (CV), difficultés limitées par régularisation (*e.g.*, Lasso, Ridge)

Exo: Calculer l'estimateur des moindres carrés avec $X = [\mathbb{1}_{a_1}, \ldots, \mathbb{1}_{a_K}]$ obtenu par des *dummy variables* avec une seule variable explicative ayant K modalités

---

Sommaire

Et si $n < p$?

Beaucoup des choses vues avant ont besoin d'être révisées :

Par exemple : si $\mathrm{rg}(X) = n$, alors $H_X = \mathrm{Id}_n$ et $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}} = \mathbf{y}$!
En effet, l'espace engendré par les colonnes $[\mathbf{x}_0, \ldots, \mathbf{x}_p]$ est $\mathbb{R}^n$, et donc le signal observé et le signal prédit sont **identiques**

Rem: c'est un problème inhérent à la grande dimension (grand nombre de variables explicatives p)

Solutions possibles : sélection de variables, *cf.* cours sur le Lasso et méthodes gloutonnes (à venir), régularisation, etc.

Sites web et livres pour aller plus loin

- Éléments de pré-traitement en manipulation de données : "Feature Engineering", HJ van Veen
- Packages Python pour les moindres carrés :
  `statsmodels`
  `sklearn.linear_model.LinearRegression`
- McKinney (2012) concernant `python` pour les statistiques
- Lejeune (2010) concernant le modèle linéaire (notamment)
- Delyon (2015) cours plus avancé sur la régression :
  `https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf`

Références 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.