MS BGD: MDI720
Modèle linéaire: propriétés

François Portier
Telecom ParisTech

Septembre 2018

Biais

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

Proposition

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

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

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

Démonstration :

(plein rang)

Biais

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

Proposition

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

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

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

Démonstration :

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

Biais

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

Proposition

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

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

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

Démonstration :

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

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

Biais

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

Proposition

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

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

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

Démonstration :

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

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

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

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

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

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

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

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

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

et

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

Intermède sur la trace

Définition

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

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

Quelques propriétés utiles :

- $\mathsf{tr}(A) = \mathsf{tr}(A^\top)$

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)

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 $\mathrm{tr}(A)$ 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 $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

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 $\mathsf{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

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

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(X\widehat{\boldsymbol{\theta}} - X\boldsymbol{\theta}^\star)^\top (X\widehat{\boldsymbol{\theta}} - X\boldsymbol{\theta}^\star)\right] \\
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(H_X\mathbf{y} - X\boldsymbol{\theta}^\star)^\top (H_X\mathbf{y} - X\boldsymbol{\theta}^\star)\right] \\
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(H_X\varepsilon)^\top (H_X\varepsilon)\right] \\
&= \mathbb{E}(\boldsymbol{\epsilon}^\top H_X^2 \boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}^\top H_X \boldsymbol{\epsilon})
\end{aligned}$$

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

$$\begin{aligned}
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(X\widehat{\boldsymbol{\theta}} - X\boldsymbol{\theta}^\star)^\top (X\widehat{\boldsymbol{\theta}} - X\boldsymbol{\theta}^\star)\right] \\
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(H_X\mathbf{y} - X\boldsymbol{\theta}^\star)^\top (H_X\mathbf{y} - X\boldsymbol{\theta}^\star)\right] \\
n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \widehat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(H_X\varepsilon)^\top (H_X\varepsilon)\right] \\
&= \mathbb{E}(\boldsymbol{\epsilon}^\top H_X^2 \boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}^\top H_X \boldsymbol{\epsilon}) \\
&= \mathsf{tr}[\mathbb{E}(H_X \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top H_X^\top)] = \mathsf{tr}\left(H_X \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) H_X^\top\right)
\end{aligned}$$

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

- 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 $\mathsf{rang}(X)$ ($\mathsf{rang}(X)$: nombre de variables explicatives indépendantes) ;

Attention si $\mathsf{rang}(X) \approx n$, 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

$$\mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

$$\mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

$$\mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

$$\mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

$$\mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}})$

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

$$\mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$$

Démonstration : notons $V = \mathbb{C}\mathrm{ov}(\widehat{\boldsymbol{\theta}})$

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

Estimateur du niveau de bruit

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

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

ou si l'on souhaite un estimateur sans biais :

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

- Motivation "débiaisage" : théorie des tests

$$\|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 = \mathbf{y}^\top(\mathsf{Id}_n - H_X)\mathbf{y} = \boldsymbol{\epsilon}^\top(\mathsf{Id}_n - H_X)\boldsymbol{\epsilon} = \sum_{i=1}^{n-\mathsf{rang}(X)} \tilde{\varepsilon}_i^2$$

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

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

Estimateur du niveau de bruit (II)

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

Preuve :

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

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

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

Cas hétéroscédastique

L'estimateur MCO $\widehat{\boldsymbol{\theta}}$ 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

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(\frac{y_i - \langle \boldsymbol{\theta}, x_i \rangle}{\sigma_i} \right)^2 = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} (\mathbf{y} - X\boldsymbol{\theta})^\top \Omega (\mathbf{y} - X\boldsymbol{\theta})$$

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

---

1. Analyse de performance
   - Biais
   - Variance

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

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

1. Analyse de performance

2. Impact du niveau de bruit

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

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

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

et on remplace $\mathbf{x} \in \mathbb{R}^n$ par $[, \dots,] \in \mathbb{R}^{n \times K}$

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

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

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

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

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 $X = [, \ldots,]$ obtenu par des *dummy variables* avec une seule variable explicative ayant K modalités

---

Et si $n < p$?

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

Par exemple : si $\mathsf{rang}(X) = n$, alors $H_X = \mathsf{Id}_n$ et $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\theta}} = \mathbf{y}$!
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

[Del15] B. Delyon. Régression, 2015. `https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf`.

[Lej10] M. Lejeune. *Statistiques, la théorie et ses applications*. Springer, 2010.

[McK12] W. McKinney. *Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2012.