

MS IA : MDI721
Modèle linéaire: propriétés

Pavlo Mozharovskyi François Portier
Télécom Paris

Septembre 2019

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

Biais

Variance

2. Impact du niveau de bruit

3. Aparté

Biais

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

Proposition

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

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

Rem : l'hypothèse $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ signifie que $\forall i \in \llbracket 1, n \rrbracket, \mathbb{E}(\epsilon_i) = 0$

Démonstration :

(plein rang)

Biais

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

Proposition

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

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

Rem : l'hypothèse $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ signifie que $\forall i \in \llbracket 1, n \rrbracket, \mathbb{E}(\epsilon_i) = 0$

Démonstration :

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

Biais

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

Proposition

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

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

Rem : l'hypothèse $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ signifie que $\forall i \in \llbracket 1, n \rrbracket, \mathbb{E}(\epsilon_i) = 0$

Démonstration :

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

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

Biais

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

Proposition

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

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

Rem : l'hypothèse $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ signifie que $\forall i \in \llbracket 1, n \rrbracket, \mathbb{E}(\epsilon_i) = 0$

Démonstration :

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

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

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

$$B = \mathbf{0} \quad \text{(bruit centré)}$$

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

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

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

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

Risque quadratique

Définition

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

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

Décomposition biais/variance

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

Démonstration :

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

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

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

et

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

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

1. Analyse de performance

Biais

Variance

2. Impact du niveau de bruit

3. Aparté

Intermède sur la trace

Définition

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

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

Quelques propriétés utiles :

- $\text{tr}(A) = \text{tr}(A^T)$

Intermède sur la trace

Définition

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

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

Quelques propriétés utiles :

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

Intermède sur la trace

Définition

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

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

Quelques propriétés utiles :

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

Intermède sur la trace

Définition

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

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

Quelques propriétés utiles :

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

Intermède sur la trace

Définition

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

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

Quelques propriétés utiles :

- $\text{tr}(A) = \text{tr}(A^\top)$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, et tout $\alpha \in \mathbb{R}$,
 $\text{tr}(\alpha A + B) = \alpha \text{tr}(A) + \text{tr}(B)$ (linéarité)
- $\text{tr}(A^\top A) = \sum_{i=1}^n \sum_{j=1}^n A_{i,j}^2 := \|A\|_F^2$
- Pour toutes matrices $A, B \in \mathbb{R}^{n \times n}$, $\text{tr}(AB) = \text{tr}(BA)$
- $\text{tr}(PAP^{-1}) = \text{tr}(A)$, 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 $\text{tr}(A)$ vaut la somme des éléments diagonaux de A :

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

Quelques propriétés utiles :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque d'estimation

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

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

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

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

Démonstration :

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

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

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

$$\begin{aligned}n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E} \left[(\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}^*)^\top (\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}^*) \right] \\n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E} \left[(\mathbf{H}_X \mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*)^\top (\mathbf{H}_X \mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*) \right] \\n \cdot R_{\text{pred}}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) &= \mathbb{E} \left[(\mathbf{H}_X \boldsymbol{\epsilon})^\top (\mathbf{H}_X \boldsymbol{\epsilon}) \right] \\&= \mathbb{E}(\boldsymbol{\epsilon}^\top \mathbf{H}_X^2 \boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}^\top \mathbf{H}_X \boldsymbol{\epsilon}) \\&= \text{tr}[\mathbb{E}(\mathbf{H}_X \boldsymbol{\epsilon} \boldsymbol{\epsilon}^\top \mathbf{H}_X^\top)] = \text{tr}(\mathbf{H}_X \mathbb{E}(\boldsymbol{\epsilon} \boldsymbol{\epsilon}^\top) \mathbf{H}_X^\top)\end{aligned}$$

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

Démonstration : début identique

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

Risque de prédiction

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

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

Si le modèle est homoscédastique

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

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

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

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

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

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

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

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

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

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

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

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

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

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

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

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

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

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

Terme de variance/covariance

Matrice de variance/covariance des moindres carrés

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

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

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

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

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

Estimation du niveau de bruit

Cas hétéroscédastique

3. Aparté

Estimateur du niveau de bruit

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

Preuve :

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

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

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

1. Analyse de performance

2. Impact du niveau de bruit

Estimation du niveau de bruit

Cas hétéroscédastique

3. Aparté

Cas hétéroscédastique

L'estimateur MCO $\hat{\theta}$ 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

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

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

La formule explicite :

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

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 $\mathbf{a}_1, \dots, \mathbf{a}_K$ on crée K variables : $\forall k \in \llbracket 1, K \rrbracket, \in \mathbb{R}^n$ définies par

$$\forall i \in \llbracket 1, n \rrbracket, \quad ()_i = \begin{cases} 1, & \text{if } x_i = \mathbf{a}_k \\ 0, & \text{sinon} \end{cases}$$

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

→

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

Cas général : couleur, villes, etc.

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

→

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

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 : $X = [, \dots,]$.

Si $x_{n+1} = a_k$ alors $\hat{y}_{n+1} = \hat{\theta}_k$

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

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

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

1. Analyse de performance

2. Impact du niveau de bruit

3. Aparté

Variables qualitatives

Grande dimension $p > n$

Et si $n < p$?

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

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

En effet, l'espace engendré par les colonnes $[\mathbf{x}_0, \dots, \mathbf{x}_p]$ 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.