Modélisation statistique

#2.c Géométrie des moindres carrés

Dr. Léo Belzile
HEC Montréal

Rappels d'algèbre linéaire

Si $\mathbf{X}$ est une matrice $n \times p$, l'espace engendré par les colonnes de $\mathbf{X}$ est

$$\mathrm{S}(\mathbf{X}) = \{\mathbf{X}\boldsymbol{a}, \boldsymbol{a} \in \mathrm{R}^p\}$$

L'équation du modèle linéaire

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

correspond donc à un élément (inconnu) du sous-espace vectoriel engendré $\mathbf{X}$ plus un aléa.

Moindres carrés ordinaires

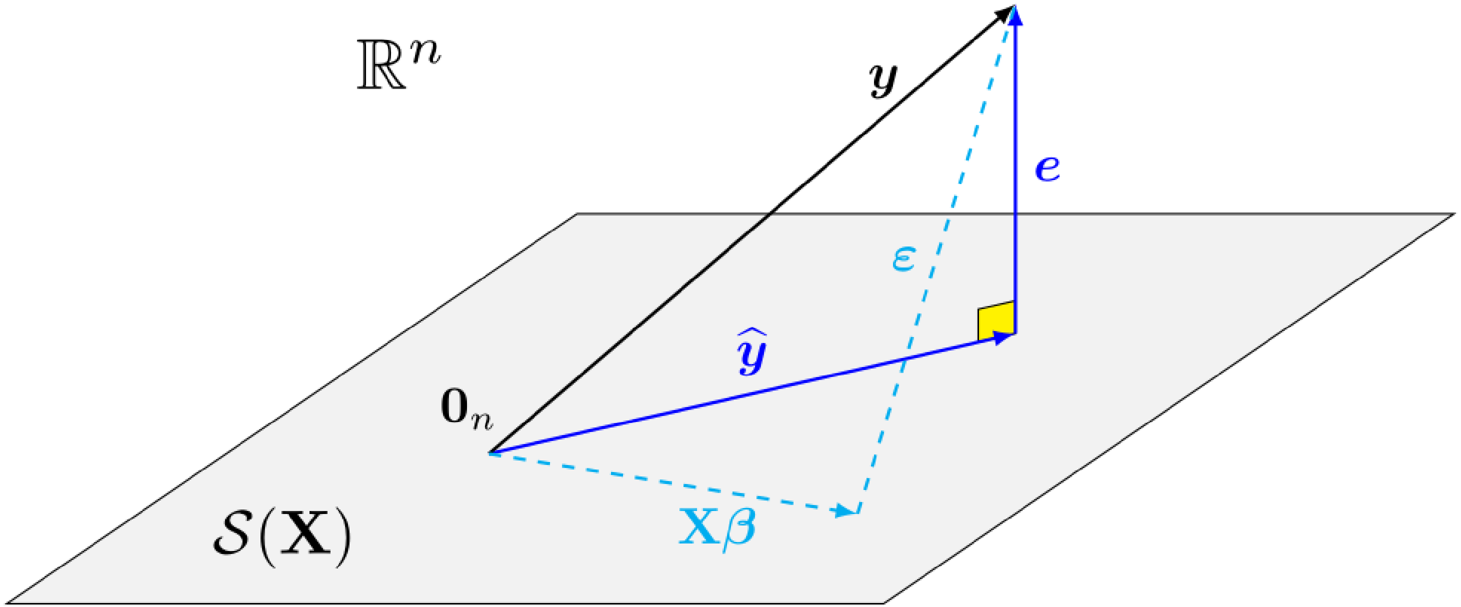
Trouver l'élément de $\mathrm{S}(\mathbf{X})$ le plus près (distance minimale) de $\boldsymbol{y}$, soit

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathrm{R}^p} ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||^2$$

Intuition: $\varepsilon_1, \ldots, \varepsilon_n$ et $\beta_0, \ldots, \beta_{p-1}$ sont inconnus, on ne peut les recouvrer (n observations, $n + p$ inconnues).

Géométrie des colonnes

On cherche à trouver la meilleur approximation p-dimensionnelle dans $S(X)$.



La solution du problème des moindres carrés est la projection de y sur $S(X)$, soit Hy, où $H = X(X^\top X)^{-1}X^\top$ est la matrice de projection.

Décomposition orthogonale

On écrit

$$\begin{aligned} \boldsymbol{y} &= \mathbf{H}\boldsymbol{y} + (\mathbf{I}_n - \mathbf{H})\boldsymbol{y} \\ &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e} \end{aligned}$$

Les résidus $\boldsymbol{e}$ sont orthogonaux aux colonnes de $\mathbf{X}$ et donc aux valeurs ajustées $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

- Par le théorème de Pythagore, $\|\boldsymbol{y}\|^2 = \|\widehat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$.

Si $\mathbf{1}_n \in \mathcal{S}(\mathbf{X})$ (le modèle inclut une ordonnée à l'origine)

- La moyenne des résidus $\boldsymbol{e}$ est nulle.
- La régression linéaire de $\hat{\boldsymbol{y}}$ sur $\boldsymbol{e}$ a une ordonnée à l'origine et une pente nulle (aucune corrélation linéaire).