TD 6 : modèle linéaire

Les questions marquées d'un astérisque **(*)** *sont facultatives.*

Exercice 1. (Régression polynomiale) Soit $n \in \mathbb{N}^*$. On considère le modèle de régression polynômiale : pour tout $i \in \{1, \dots, n\}$,

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_{m-1} X_i^{m-1} + \varepsilon_i.$$

1. Écrire ce modèle linéaire sous forme matricielle $Y = \mathbf{X}\beta + \varepsilon$.
2. **(*)** Montrer que si $m = n$, $\det(\mathbf{X}) = \prod_{1 \le i < j \le n}(X_j - X_i)$ (déterminant de matrices de Vandermonde).
   Indication : montrer que ce déterminant est un polynome de degré $n(n-1)/2$ *en les* X_i*, regarder quand il s'annule, et calculer le coefficient de son monôme* $X_2 X_3^2 \dots X_n^{n-1}$.

On suppose les X_i distincts et $m \le n$.

3. Montrer que $\mathbf{X}^\top \mathbf{X}$ est inversible et écrire l'estimateur des moindres carrés.

Exercice 2. (Coefficient de détermination) Soit $n \in \mathbb{N}^*$. On considère le modèle linéaire $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ pour $i \in \{0, \dots, n\}$, avec $Y_i, X_i \in \mathbb{R}$. On note $\hat\beta$ l'estimateur des moindres carrés de β, $\hat Y_i = \hat\beta_0 + \hat\beta_1 X_i$ et $\bar Y_n$ la moyenne empirique des Y_i.

1. Montrer que $\sum_{i=1}^n (Y_i - X_i \hat\beta) = 0$ et que $\sum_{i=1}^n (Y_i - X_i\hat\beta) X_i = 0$.
2. En déduire $\sum_{i=1}^n (Y_i - \hat Y_i)(\hat Y_i - \bar Y_n) = 0$.
3. Prouver que $\sum_{i=1}^n (Y_i - \bar Y_n)(\hat Y_i - \bar Y_n) = \sum_{i=1}^n (\hat Y_i - \bar Y_n)^2$.
4. Démontrer que $\sum_{i=1}^n (\hat Y_i - Y_i)^2 + \sum_{i=1}^n (\hat Y_i - \bar Y_n)^2 = \sum_{i=1}^n (Y_i - \bar Y_n)^2$.

Le coefficient de détermination est défini comme le rapport

$$R^2 = \frac{\left(\sum_{i=1}^n (Y_i - \bar Y_n)(\hat Y_i - \bar Y_n)\right)^2}{\sum_{i=1}^n (Y_i - \bar Y_n)^2 \sum_{i=1}^n (\hat Y_i - \bar Y_n)^2}.$$

5. En utilisant les questions précédentes, montrer que

$$R^2 = \frac{\sum_{i=1}^n (\hat Y_i - \bar Y_n)^2}{\sum_{i=1}^n (Y_i - \bar Y_n)^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat Y_i)^2}{\sum_{i=1}^n (Y_i - \bar Y_n)^2}$$

Exercice 3. (Régression ridge) Dans un modèle linéaire $Y = \mathbf{X}\beta + \epsilon$, lorsque la matrice $\mathbf{X}^\top\mathbf{X}$ n'est pas ou presque pas inversible, l'estimateur des moindres carrés peut ne pas être défini ou être instable et prendre des valeurs anormalement grandes. L'estimateur ridge a été introduit pour résoudre ce problème : c'est le minimiseur du critère des moindres carrés pénalisé

$$\|Y - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2,$$

où $\lambda > 0$. La différence par rapport à l'estimateur des moindres carrés habituel est le terme $\lambda\|\beta\|^2$, qui assure que β ne prend pas de trop grandes valeurs.

1. Montrer que le gradient en β du critère des moindres carrés pénalisé s'écrit $-2(Y - \mathbf{X}\beta)^\top \mathbf{X} + 2\lambda\beta^\top$.
2. En déduire une matrice M telle que le minimiseur de ce critère s'écrit $\hat\beta = M^{-1} X^\top Y$.
3. Commenter : la matrice M est-elle inversible ? Est-elle sujette à instabilité comme $\mathbf{X}^\top\mathbf{X}$?