TD 6 : modèle linéaire

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

Avertissement : les applications et valeurs numériques des exercices ci-dessous sont pure invention.

Exercice 1. On s'intéresse au lien entre le taux d'arbres touchés par des malformations et le taux de polluants dans le sol. Pour cela, on effectue des mesures dans différents environnements. Les résultats sont reportés dans le tableau ci-dessous. X_i est le taux de polluants mesuré, Y_i le taux d'arbres malformés.

X_i	4,5	5,6	5,2	9,3	6,0	5,3	8,7	6,9
Y_i	9,6	9,5	8,7	13,6	10,4	7,7	13,1	10,9

$$\sum_i X_i = 51,5 \qquad \sum_i Y_i = 83,5$$

$$\sum_i X_i^2 = 352,53 \qquad \sum_i Y_i^2 = 900,93 \qquad \sum_i X_i Y_i = 560,51$$

1. Quelle est la variable à expliquer et la variable explicative ?
2. On cherche une dépendance affine entre X et Y. Écrire le modèle linéaire correspondant.
3. Estimer les paramètres du modèle linéaire.
4. Quelle valeur peut-on prédire pour le taux d'arbres malformés si le taux de pollution est de 8 ?

Exercice 2. On cherche à prédire la solidité de poutres métalliques sans les soumettre à des tests coûteux sous presse hydraulique. Pour cela, on observe l'intensité de la résonance de ces poutres à une fréquence donnée, bien plus simple à mesurer. On note S_i la solidité de la poutre i, I_i l'intensité de sa résonance. On suppose la dépendance entre ces variables affine, et on calibre la méthode avec l'échantillon suivant.

S_i	4 655	3 429	4 745	4 340	4 381	4 675	4 732	4 338	3 898	4 347	3 259
I_i	384	386	268	260	305	362	273	384	282	348	362

$$\sum_i S_i = 46\,799 \qquad \sum_i I_i = 3\,614$$

$$\sum_i S_i^2 = 201\,748\,439 \qquad \sum_i I_i^2 = 1\,213\,602 \qquad \sum_i S_i I_i = 15\,289\,107$$

1. Quelle est la variable à expliquer et la variable explicative ?
2. Écrire le modèle linéaire correspondant.
3. Estimer les paramètres du modèle linéaire.

Dans les questions suivantes, on se place dans le cadre du modèle linéaire gaussien homoscédastique de variance inconnue.

4. Rappeler ce que cela signifie.
5. Rappeler la formule de l'estimateur non biaisé de la variance du bruit. On admettra qu'il vaut ici $\hat{\sigma}^2 = 262\,124$.
6. Peut-on conclure avec un test de niveau 5% qu'il y a une dépendance entre la solidité d'une poutre et l'intensité de sa résonance ? On écrira soigneusement le test utilisé. On pourra utiliser la table de quantile ci-dessous.

ordre :	95%	97,5%
$\mathcal{T}(9)$	1,833	2,262
$\mathcal{T}(10)$	1,812	2,228
$\mathcal{T}(11)$	1,796	2,201
$\mathcal{N}(0,1)$	1,644	1,960

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. L'enjeu de cet exercice est de mettre en évidence ce comportement et d'étudier un estimateur qui ne souffre pas de ce problème : l'estimateur ridge.

Supposons $\mathbf{X} \in \mathbb{R}^{n\times d}$ de rang d et soit $\mathbf{X} = U\Sigma V^\top$ sa décomposition en valeurs singulières. Pour tout $i \in \{1,\dots,d\}$, on note $\sigma_i = \Sigma_{ii}$ la i-ème valeur singulière de $\mathbf{X}$, et on suppose $\sigma_1 \geqslant \cdots \geqslant \sigma_d > 0$.

1. Soit $\delta > 0$ et soit $e_d = (0,0,\dots,0,1)^\top \in \mathbb{R}^d$. Soit $Y^{\text{perturb.}} = Y + \delta U e_d$.
   (a) Calculer $\|Y - Y^{\text{perturb.}}\|$.
   Indication : si U est une matrice orthogonale, alors $\|Ux\| = \|x\|$ pour tout x.

   Notons $\hat\beta$ l'estimateur des moindres carrés de β obtenu à partir de $\mathbf{X}$ et Y, et $\hat\beta^{\text{perturb.}}$ l'estimateur des moindres carrés de β obtenu à partir de $\mathbf{X}$ et $Y^{\text{perturb.}}$.
   (b) Calculer $\|\hat\beta - \hat\beta^{\text{perturb.}}\|$.
   (c) Comment se comporte le ratio $\frac{\|\hat\beta - \hat\beta^{\text{perturb.}}\|}{\|Y - Y^{\text{perturb.}}\|}$ lorsque $\mathbf{X}^\top\mathbf{X}$ est presque pas inversible, c'est-à-dire dans la limite $\sigma_d \to 0$? Peut-on dire que l'estimateur des moindres carrés est instable ?

L'estimateur ridge est le minimiseur du critère des moindres carrés pénalisé

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

où $\lambda > 0$ est fixé à l'avance.

2. Montrer que le gradient en β du critère des moindres carrés pénalisé s'écrit $-2\mathbf{X}^\top(Y - \mathbf{X}\beta) + 2\lambda\beta$.
   Soit $f : \mathbb{R}^d \longrightarrow \mathbb{R}^p$ une fonction $\mathcal{C}^1$. Son gradient $\nabla_x f$ en x est la matrice de taille $d \times p$ tel que $(\nabla_x f)_{ij} = (\frac{\partial}{\partial x_i} f_j)(x)$.
3. En déduire une matrice M telle que le minimiseur de ce critère s'écrit $\hat\beta = M^{-1}\mathbf{X}^\top Y$.
4. Commenter : la matrice M est-elle inversible ? Est-elle sujette à instabilité comme $\mathbf{X}^\top\mathbf{X}$?

Exercice 4. (*) (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 $m = n$, $\det(\mathbf{X}) = \prod_{1\leqslant i<j\leqslant n}(X_j - X_i)$ (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 $X_2X_3^2\dots X_n^{n-1}$.

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

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