Modélisation statistique – partiel

M1 Mathématiques et Intelligence Artificielle, Université Paris-Saclay
2025-2026

Mardi 4 novembre 2025
13h30 – 16h30 (tiers-temps 13h30 – 17h30)

Une feuille A4 recto manuscrite est autorisée en tant que support. La calculatrice est autorisée.

Avant de commencer :

- *Le sujet comporte trois exercices. Il est demandé de les traiter dans l'ordre sur la copie.*
- *Des résultats pourront être admis d'une question sur l'autre, à la condition de l'écrire clairement.*
- *Une attention particulière sera portée aux questions d'interprétation, à la rigueur et la concision de la rédaction.*
- *Un barème indicatif est donné :*

  Exercice 1 (questions de cours) : 4 points,

  Exercice 2 : 8 points,

  Exercice 3 : 8 points.

Exercice 1 – Des questions de cours

1. Montrer que si X est un vecteur aléatoire de $\mathbb{R}^d$, de matrice de covariance finie Σ, alors pour tous $A \in \mathbb{R}^{m\times d}$ et $b \in \mathbb{R}^m$, $Y = AX + b$ a une matrice de covariance finie donnée par $A\Sigma A^T$. *Après avoir montré qu'il est fini, on pourra calculer* $\mathrm{Cov}(Y_i, Y_j)$ *pour tout* $1 \le i, j \le m$ *en fonction des coefficients de* Σ *et de* A.

2. Dans le modèle linéaire gaussien $Y = X\theta + \varepsilon$. Rappeler sans justification l'expression de l'estimateur des moindres carrés de θ, noté $\hat{\theta}$, et de l'estimateur non biaisé de la variance, noté $\hat{\sigma}^2$. Rappeler sans justification de quelle projection orthogonale $X\hat{\theta}$ est le résultat. Puis, montrer que $\hat{\theta}$ et $\hat{\sigma}^2$ sont indépendants.

Exercice 2 – Consommation d'eau annuelle d'un bâtiment

Une société d'environnement souhaite prédire la consommation annuelle d'eau (`conso_eau`, en m^3) de bâtiments de bureau en fonction de plusieurs caractéristiques physiques. On dispose d'un échantillon de $n = 32$ bâtiments, où les variables mesurées sont :

- `conso_eau` : consommation d'eau annuelle du bâtiment (en m^3) ;
- `surface` : surface utile du bâtiment (en m^2) ;
- `recup` : indicateur binaire égal à 1 si le bâtiment dispose d'un système de récupération d'eau de pluie, 0 sinon ;
- `occupants` : nombre moyen d'occupants dans le bâtiment ;
- `hauteur` : hauteur moyenne du bâtiment (en m).

On se place sous les hypothèses du modèle linéaire gaussien :

$$\texttt{conso\_eau} = \theta_0 + \theta_1 \cdot \texttt{surface} + \theta_2 \cdot \texttt{recup} + \theta_3 \cdot \texttt{occupants} + \theta_4 \cdot \texttt{hauteur} + \varepsilon,$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. On notera ce modèle sous sa forme vectorielle classique $Y = X\theta + \varepsilon$. Une régression linéaire a été effectuée avec l'estimateur des moindres carrés. Les résultats sont partiellement donnés ci-dessous :

coordonnée i	$\hat{\theta}_i$	$\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{i,i}}$
`(Intercept)`	240	51.4
`surface`	4.10	1.20
`recup`	-85	29.1
`occupants`	4.50	1.10
`hauteur`	2.8	3.1

On dispose également des informations suivantes :

$$SCR := \sum_{i=1}^{n} (Y_i - X_i\hat{\theta})^2 = 26\,600, \qquad SCT := \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 158\,000.$$

On donne également les quantiles suivants :

	$\beta = 0.95$	$\beta = 0.975$	$\beta = 0.98$	$\beta = 0.99$	$\beta = 0.995$
$d = 26$	1.706	2.056	2.479	2.779	3.435
$d = 27$	1.703	2.052	2.473	2.771	3.421
$d = 28$	1.701	2.048	2.467	2.763	3.408

Quantiles de la loi de Student $\mathcal{T}(d)$ de niveau β, pour β et d variables.

	$d_2 = 26$	$d_2 = 27$	$d_2 = 28$
$d_1 = 3$	4.637	4.601	4.568
$d_1 = 4$	4.140	4.106	4.074
$d_1 = 5$	3.818	3.785	3.754

Quantiles de la loi de Fisher $\mathcal{F}(d_1, d_2)$ de niveau $\beta = 0.99$.

1. D'après les résultats de la régression, quelles variables expliquent significativement la consommation d'eau au seuil de 2% ? Justifier à l'aide de tests de Student.
2. Ecrire le coefficient de détermination R^2 en fonction de SCT et de SCR. Faire l'application numérique et interpréter sa valeur.
3. Donner l'estimateur sans biais de σ^2, noté $\widehat{\sigma}^2$, et préciser sa loi sous les hypothèses du modèle (avec les valeurs numériques des paramètres).
4. Donner un intervalle de confiance pour le coefficient θ_1 associé à la variable `surface` de niveau $1 - \alpha$. Faire l'application numérique pour $\alpha = 0.05$.
5. Ecrire la statistique de Fisher associée au test global de nullité des coefficients (hors constante) en fonction de SCR et de SCT. Faire l'application numérique, et conclure grâce à la table des quantiles. Préciser le niveau d'erreur.
6. Le coefficient négatif associé à `recup` permet-il de conclure que le système de récupération d'eau cause une baisse de consommation ?

Exercice 3 – Modèle linéaire avec bruits corrélés

On se place dans un modèle linéaire *non supposé gaussien*, où Y, vecteur aléatoire de $\mathbb{R}^n$, s'écrit

$$Y = X\theta + \varepsilon, \tag{1}$$

avec $X \in \mathbb{R}^{n \times p}$ une matrice déterministe, $\theta \in \mathbb{R}^p$ un vecteur de paramètres inconnus, et ε un vecteur aléatoire de $\mathbb{R}^n$ centré. *Contrairement au cas standard, on suppose ici que le bruit ε a une matrice de covariance $\Sigma \in \mathbb{R}^{n \times n}$ supposée symétrique définie positive, qui n'est pas forcément de la forme $\sigma^2 I_n$.* On suppose que Σ est connue. On admet que ces nouvelles hypothèses, on a encore (la démonstration de ce résultat est la même que dans le cours) :

$$\text{le modèle (1) est identifiable en } \theta \iff X^T X \text{ inversible}\,.$$

Dans toute la suite de cet exercice, nous nous plaçons sous les hypothèses d'identifiabilité ci-dessus. On note :

$$\hat{\theta} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2,$$

avec $\|\cdot\|$ norme euclidienne usuelle sur $\mathbb{R}^n$. On rappelle qu'il est unique par hypothèse, et dans cet exercice il est appelé *estimateur des moindres carrés ordinaire.*

1. Rappeler la forme close de $\hat{\theta}$, et déterminer l'espérance puis la matrice de covariance de $\hat{\theta}$ en fonction de θ, Σ et X. *Attention, le modèle a changé par rapport à celui du cours : les résultats de cette question peuvent donc différer de ceux du modèle classique.*

Notons $\|\cdot\|_\Sigma$ la norme euclidienne sur $\mathbb{R}^n$ définie pour tout $u \in \mathbb{R}^n$ par $\|u\|_\Sigma^2 = u^T \Sigma^{-1} u$. On considère le nouveau problème d'optimisation :

$$\operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} \|Y - X\theta\|_\Sigma^2\,. \tag{2}$$

2. On admet l'existence de $S \in \mathbb{R}^{n\times n}$, symétrique définie positive telle que $S^2 = \Sigma^{-1}$. En utilisant S, montrer que $X^T\Sigma^{-1}X$ est symétrique définie positive.

3. Prouver que le problème (2) admet une unique solution appelée *estimateur des moindres carrés généralisé*, notée $\hat{\theta}_g$, qui vaut

$$\hat{\theta}_g = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y\,.$$

4. Déterminer l'espérance puis la matrice de covariance de $\hat{\theta}_g$ en fonction de θ, Σ et X.

On note $\preceq$ l'ordre partiel sur les matrices symétriques de taille $m \times m$ défini par $A \preceq B \iff \forall u \in \mathbb{R}^m, u^TAu \leq u^TBu$.

5. En utilisant la matrice S définie à la question 1, montrer que

$$X(X^T\Sigma^{-1}X)^{-1}X^T \preceq \Sigma$$

On pourra partir de $v^TX(X^T\Sigma^{-1}X)^{-1}X^Tv$ pour $v \in \mathbb{R}^n$, écrire $v = SS^{-1}v$ et observer que $\Pi := SX((SX)^TSX)^{-1}(SX)^T$ est une certaine projection orthogonale.

6. En déduire que $\mathrm{Var}(\hat{\theta}_g) \preceq \mathrm{Var}(\hat{\theta})$. Pourquoi cela vous semble-t-il logique ? Commenter.

7. (Question bonus.) Y a t-il une façon de modifier nos données (X, Y) (préprocessing) dans le modèle de cet exercice pour se ramener au modèle linéaire classique du cours ? Interpréter.

Fin du sujet.