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. *Solution. L'opération faite sur X étant linéaire, l'espace des v.a. de carré intégrable étant un espace vectoriel, les coordonnées de $Y = AX + b$ sont toujours de carré intégrable. De plus, pour $1 \le i, j \le m$, en utilisant la bilinéarité de la covariance, on a*

$$\begin{aligned}\mathrm{Cov}(Y_i, Y_j) &= \mathrm{Cov}\left(\sum_{k=1}^{d} A_{ik}X_k + b_i, \sum_{\ell=1}^{d} A_{j\ell}X_\ell + b_j\right)\\ &= \sum_{1\le k,\ell\le d} A_{ik}A_{j\ell}\mathrm{Cov}(X_k, X_\ell)\\ &= \sum_{1\le k,\ell\le d} A_{ik}\Sigma_{k\ell}A_{j\ell}\,.\end{aligned}$$

Par ailleurs,

$$(A\Sigma A^T)_{ij} = \sum_{k=1}^{d} A_{ik}(\Sigma A^T)_{kj} = \sum_{k=1}^{d}\sum_{\ell=1}^{d} A_{ik}\Sigma_{k\ell}A_{j\ell},$$

ce qui prouve le résultat voulu.

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. *Solution. On a d'après le cours*

$$\hat\theta = (X^TX)^{-1}X^TY \quad \textit{et} \quad \hat\sigma^2 = \frac{\|Y - X\hat\theta\|^2}{n-p}\,.$$

De plus, d'après le cours, $X\hat\theta$ est le résultat de la projection orthogonale de Y sur $\mathrm{Im}(X)$. Si l'on note Π cette projection, on a donc $X\hat\theta = \Pi Y$. De même, $Y - X\hat\theta = (I_n - \Pi)Y$ est la projection orthogonale de Y sur $\mathrm{Im}(X)^\perp$. Comme $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$ d'après les hypothèses, le théorèem de Cochran implique que les vecteurs gaussiens $X\hat\theta$ et $Y - X\hat\theta$ sont indépendants. Puis, comme $\hat\sigma^2$ est une fonction de $Y - X\hat\theta$, et $\hat\theta = ((X^TX)^{-1}X^T)X\hat\theta$, donc une fonction de $X\hat\theta$, ils sont à leur tour 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.

   Solution. *La valeur critique de tous ces tests de Student est le quantile pour $d = n-p = 32-5 = 27$ et $\beta = 0.99$, soit 2.771. Les statistiques de tests sont obtenues via la formule*

   $$t_i = \frac{\hat{\theta}_i}{\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{i,i}}}.$$

   On fait les calculs et on compare les valeurs absolues :

   - $|t_{surface}| = 3.42 > 2.771 \Rightarrow$ *significatif ;*
   - $|t_{recup}| = 2.92 > 2.771 \Rightarrow$ *significatif ;*
   - $|t_{occupants}| = 4.09 > 2.771 \Rightarrow$ *significatif ;*
   - $|t_{hauteur}| = 0.90 < 2.771 \Rightarrow$ *non significatif.*

   Les variables **`surface`**, **`recup`** *et* **`occupants`** *expliquent significativement la consommation d'eau, mais* **`hauteur`** *ne semble pas utile dans ce modèle.*

2. Ecrire le coefficient de détermination R^2 en fonction de SCT et de SCR. Faire l'application numérique et interpréter sa valeur.

   Solution.

   $$R^2 = 1 - \frac{SCR}{variance\ totale} = 1 - \frac{26600}{158000} = 1 - 0.168 = 0.832.$$

   Le modèle explique environ 83.2% *de la variabilité de la consommation d'eau : c'est un pouvoir explicatif très correct.*

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).

   Solution.

   $$\widehat{\sigma}^2 = \frac{SCR}{n-p} = \frac{26600}{27} \approx 985.2.$$

   Sous les hypothèses du modèle linéaire gaussien :

   $$\widehat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p) = \frac{\sigma^2}{27}\chi^2(27).$$

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 $\alpha = 0.05$. *Solution. L'estimateur de θ_1 est $\widehat{\theta}_1$, et l'on sait d'après le cours que* $\frac{\widehat{\theta}_1 - \theta_1}{\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{1,1}}} \sim \mathcal{T}(n-p)$. *Un intervalle de confiance pour θ_1 de niveau $1-\alpha$ est donc*

   $$[\widehat{\theta}_1 \pm t_{1-\alpha/2}\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{1,1}}],$$

   avec t_β le quantile d'ordre β de la loi $\mathcal{T}(n-p)$. Application numérique :

   $$[4.10 \pm 2.052 \times 1.20] = [1.64, 6.56].$$

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.

   Solution. En appliquant la formule du cours, la statistique de Fisher pour le test de nullité des coefficients est donnée par

   $$F = \frac{\|X\hat{\theta} - \bar{Y}\|^2/(p-1)}{\|Y - X\hat{\theta}\|^2/(n-p)} = \frac{[\|Y - \bar{Y}\|^2 - \|Y - X\hat{\theta}\|^2]/(p-1)}{\|Y - X\hat{\theta}\|^2/(n-p)} = \frac{(SCT - SCR)/(p-1)}{SCR/(n-p)},$$

   d'après Pythagore. L'application numérique donne :

   $$F = \frac{(158000 - 26600)/4}{26600/27} \approx 33.35.$$

   Le quantile de $\mathcal{F}(4, 24)$ au seuil 0.99 vaut 4.106, *donc on rejette clairement H_0 avec niveau d'erreur* 1%. *Le modèle est globalement significatif : au moins une variable explique la consommation d'eau, ce qui n'est pas une conclusion très forte.*

6. Le coefficient négatif associé à `recup` permet-il de conclure que le système de récupération d'eau cause une baisse de consommation ?

   Solution. Non : la régression établit une association statistique, pas une causalité. D'autres variables non observées (âge du bâtiment, comportement des usagers, climat, etc.) peuvent expliquer simultanément la présence d'un système de récupération et une moindre consommation. Une étude expérimentale (ou quasi-expérimentale) serait nécessaire pour conclure à la causalité.

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 :

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

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.* *Solution. On a d'après le cours que*

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

ce qui entraîne

$$\mathbb{E}[\hat{\theta}] = (X^T X)^{-1} X^T \mathbb{E}[Y] = (X^T X)^{-1} X^T X\theta = \theta$$

et

$$\mathrm{Var}[\hat{\theta}] = (X^T X)^{-1} X^T \mathrm{Var}[Y] X (X^T X)^{-1} = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

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 :

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

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. *Solution. La symétrie est évidente. Soit $u \in \mathbb{R}^p$. On a $u^T X^T \Sigma^{-1} X u = u^T X^T S^2 X u = \|SXu\|^2 \geq 0$, et qui vaut 0 si et seulement si $SXu = 0$ si et seulement si $Xu = 0$ (car S est inversible), si et seulement si $u = 0$ par identifiabilité (X est injective).*

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

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

Solution. Notons, pour $\theta \in \mathbb{R}^p$, $f(\theta) = \|Y - X\theta\|_\Sigma^2 = (Y - X\theta)^T \Sigma^{-1} (Y - X\theta)$. On a que f est de classe C^∞, faisons comme dans le cours et calculons son gradient

$$\begin{aligned}\nabla_\theta f(\theta) &= \nabla_\theta (Y - X\theta)^T \Sigma^{-1} (Y - X\theta)\\ &= \nabla_\theta (Y^T\Sigma^{-1}Y - 2(X^T\Sigma^{-1}Y)^T\theta + \theta^T X^T \Sigma^{-1} X\theta)\\ &= -2X^T\Sigma^{-1}Y + 2X^T\Sigma^{-1}X\theta,\end{aligned}$$

qui s'annule en $\hat{\theta}_g = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$, et l'inverse a bien un sens grâce à la question précédente. Justifions que ce point critique est un minimum global. Cela est établi en prenant la hessienne de f qui n'est autre que

$$Hf(\theta) = 2X^T\Sigma^{-1}X,$$

qui est symétrique définie positive d'après la question précédente. La fonction f est donc convexe : elle admet un minimum global en $\hat{\theta}_g = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$.

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

$$\mathbb{E}[\hat{\theta}_g] = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbb{E}[Y] = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\theta = \theta$$

et

$$\mathrm{Var}[\hat{\theta}_g] = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathrm{Var}[Y]\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} = (X^T\Sigma^{-1}X)^{-1}.$$

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.

Solution. *Pour tout $v \in \mathbb{R}^n$,*

$$\begin{aligned} v^TX(X^T\Sigma^{-1}X)^{-1}X^Tv &= v^TX(X^TS^TSX)^{-1}X^Tv \\ &= (S^{-1}v)^T\underbrace{SX((SX)^TSX)^{-1}(SX)^T}_{=\Pi}S^{-1}v, \end{aligned}$$

où Π est la projection orthogonale sur $\mathrm{Im}(SX)$. En particulier, on a $\|\Pi w\| \leq \|w\|$ pour tout $w \in \mathbb{R}^n$, ce qui implique, par Cauchy-Schwarz,

$$(S^{-1}v)^T\Pi S^{-1}v = \langle S^{-1}v, \Pi S^{-1}v\rangle \leq \|S^{-1}v\|\|\Pi S^{-1}v\| \leq \|S^{-1}v\|^2 = v^T(S^{-1})^2v = v^T\Sigma v.$$

D'où $X(X^T\Sigma^{-1}X)^{-1}X^T \preceq \Sigma$.

6. En déduire que $\mathrm{Var}(\hat{\theta}_g) \preceq \mathrm{Var}(\hat{\theta})$. Pourquoi cela vous semble-t-il logique ? Commenter. *Solution.* *On applique l'inégalité des formes quadratiques à $X(X^TX)^{-1}w \in \mathbb{R}^n$ avec $w \in \mathbb{R}^p$. Cela donne*

$$\begin{aligned} &w^T(X^TX)^{-1}X^TX(X^T\Sigma^{-1}X)^{-1}X^TX(X^TX)^{-1}w \leq w^T(X^TX)^{-1}X^T\Sigma X^TX(X^TX)^{-1}w \\ \iff\ & w^T(X^T\Sigma^{-1}X)^{-1}w \leq w^T(X^TX)^{-1}X^T\Sigma X^TX(X^TX)^{-1}w \\ \iff\ & w^T\mathrm{Var}(\hat{\theta}_g)w \leq w^T\mathrm{Var}(\hat{\theta})w, \end{aligned}$$

Ce qui conclut la question. On a vu que $\hat{\theta}_g$ et $\hat{\theta}$ sont tous les deux sans biais mais que $\hat{\theta}_g$ est meilleur que $\hat{\theta}$ au sens de l'ordre précédent. C'est logique, car $\hat{\theta}_g$ utilise Σ sans sa définition, et pas $\hat{\theta}$, qui a donc de l'information en moins.

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. *Solution.* *Oui, il suffit de remarquer que si $Y = X\theta + \varepsilon$ avec $\mathrm{Var}(\varepsilon) = \Sigma$, $SY = SX + \varepsilon'$ avec $Var(\varepsilon') = I_n$, où S est définie à la question 2. On peut donc voir le modèle généralisé tout simplement comme le modèle linéaire classique dans lequel on a créé de la dépendance entre les données en multipliant par une matrice S.*

Fin du sujet.