

DM 1 – Régression linéaire multiple

APM_41033_EP – Statistiques

Promotion X2025

Les informations importantes concernant les DM sont sur Moodle, dans la rubrique *Recherches maison notées*.

On observe n vecteurs dans $\mathbb{R}^{p+1}$, $(Z_i^1, \dots, Z_i^p, Y_i)_{i=1}^n$.

Notre objectif est d'expliquer la variable quantitative Y_i (réponse) par les p variables déterministes $Z_i^1, \dots, Z_i^p$ dites *explicatives* (ou encore *régresseurs*).

Le *modèle linéaire* consiste à supposer que

$$Y_i = \beta_1 Z_i^1 + \dots + \beta_p Z_i^p + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées, et où

$$\theta := (\beta_1, \dots, \beta_p, \sigma^2) \in \Theta := \mathbb{R}^p \times \mathbb{R}_+^*.$$

Il est plus pratique de représenter le modèle sous la forme matricielle

$$Y = Z\beta + \sigma\varepsilon,$$

avec

$$Y := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta := \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad Z := \begin{bmatrix} Z_1^1 & \dots & Z_1^p \\ Z_2^1 & \dots & Z_2^p \\ \vdots & & \vdots \\ Z_n^1 & \dots & Z_n^p \end{bmatrix}.$$

Nous supposons par ailleurs que :

(H1) $n > p$ et Z est de rang p .

(H2) $\mathbb{E}[\varepsilon] = 0$ et $\text{Var}(\varepsilon) = I_n$.

1 Estimation par moindres carrés du vecteur β

Pour estimer le paramètre $\beta \in \mathbb{R}^p$, la *méthode des moindres carrés* consiste à chercher un estimateur $\hat{\beta} \in \mathbb{R}^p$ qui minimise la fonction

$$J_n : \beta \mapsto \sum_{i=1}^n (Y_i - \beta_1 Z_i^1 - \dots - \beta_p Z_i^p)^2 = \|Y - Z\beta\|^2.$$

1. Soit H la matrice de projection orthogonale sur l'image de Z . Déterminer l'expression de H de deux manières différentes, par exemple :

- En écrivant que, pour $u \in \mathbb{R}^n$, Hu est dans l'image de Z et $u - Hu$ est orthogonal à l'image de Z .
- En cherchant les points critiques de la fonction $x \mapsto \|u - Zx\|^2$.

2. En déduire l'existence et l'unicité de l'estimateur des moindres carrés, ainsi que son expression :

$$\hat{\beta} = (Z^\top Z)^{-1} Z^\top Y.$$

L'estimateur $\hat{\beta}$ est dit *linéaire*, car il est obtenu comme combinaison linéaire des observations $Y_1, \dots, Y_n$.

3. Montrer que l'estimateur des moindres carrés est un estimateur sans biais de β .
 4. Montrer que, pour tout $\theta \in \Theta$,

$$\text{Var}_\theta(\hat{\beta}) = \sigma^2 (Z^\top Z)^{-1}.$$

5. En observant que, pour tout $w \in \mathbb{R}^\ell$, $\text{Tr}(ww^\top) = \|w\|^2$, montrer que

$$\hat{\sigma}^2 := \frac{1}{n-p} \|Y - Z\hat{\beta}\|^2$$

est un estimateur sans biais de σ^2 .

2 Optimalité de l'estimateur des moindres carrés $\hat{\beta}$

Soit B une matrice déterministe de taille $p \times n$; on pose $\tilde{\beta} := BY$.

6. À quelle condition sur B l'estimateur $\tilde{\beta}$ est-il sans biais ?
 7. On suppose la condition de la question précédente vérifiée.
 Démontrer que pour tout $\theta \in \Theta$, $\text{Cov}_\theta(\tilde{\beta}, \tilde{\beta} - \hat{\beta}) = 0$, et en déduire que $\text{Var}_\theta(\tilde{\beta}) \succeq \text{Var}_\theta(\hat{\beta})$ ¹.
 Interpréter.

3 Cas de la régression linéaire gaussienne

On suppose maintenant que Y est une observation du modèle gaussien

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathcal{N}_n(Z\beta, \sigma^2 I_n) : (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*\}).$$

8. Pour tout $\theta \in \Theta$, déterminer la loi de $\hat{\beta}$ sous $\mathbb{P}_\theta$.
 9. Pour tout $\theta \in \Theta$, déterminer la loi de $\hat{\sigma}^2$ sous $\mathbb{P}_\theta$.
 10. Pour tout $\theta \in \Theta$, montrer que $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants sous $\mathbb{P}_\theta$.
 11. Déterminer l'estimateur du maximum de vraisemblance de $\theta = (\beta, \sigma^2)$.
 Ses différentes composantes sont-elles biaisées ?

On admet par ailleurs, dans ce contexte, que si l'on dispose d'une nouvelle observation Y_0 indépendante des précédentes, associée au vecteur $z_0 = (Z_0^1, \dots, Z_0^p)^\top$, alors un intervalle de prédiction de Y_0 au niveau $1 - \alpha$ est donné par :

$$\left[z_0^\top \hat{\beta} - q_{n-p}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + z_0^\top (Z^\top Z)^{-1} z_0}, z_0^\top \hat{\beta} + q_{n-p}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + z_0^\top (Z^\top Z)^{-1} z_0} \right],$$

avec q_{n-p} les quantiles de la loi de Student à $n - p$ degrés de liberté.

1. On rappelle que pour $A, B \in \mathbb{R}^{p \times p}$ symétriques, $A \succeq B \iff \forall x \in \mathbb{R}^p, x^\top A x \geq x^\top B x$.

4 Tests statistiques dans le modèle gaussien

Remarque : les tests devraient être vus au cours de la quatrième séance.

12. Soit $x \in \mathbb{R}^p$ non nul. Montrer que, sous $\mathbb{P}_\theta$,

$$\frac{x^\top \hat{\beta} - x^\top \beta}{\hat{\sigma} \sqrt{x^\top (Z^\top Z)^{-1} x}}$$

suit une loi de Student à $n - p$ degrés de liberté.

13. Soit $\alpha \in]0, 1[$. Déterminer un intervalle de confiance bilatéral de niveau de couverture $1 - \alpha$ pour $\beta^\top x$.

14. Soient $\alpha \in]0, 1[$, $c \in \mathbb{R}$. Construire un test de niveau α de

$$H_0 : \beta^\top x = c \quad \text{contre} \quad H_1 : \beta^\top x \neq c.$$

Déterminer la p -valeur de ce test.

15. Soit A une matrice de taille $r \times p$ et de rang $r \leq p$. Montrer que, sous $\mathbb{P}_\theta$,

$$\frac{1}{r\hat{\sigma}^2} \left\{ A(\hat{\beta} - \beta) \right\}^\top \left[A(Z^\top Z)^{-1} A^\top \right]^{-1} \left\{ A(\hat{\beta} - \beta) \right\}$$

suit une loi de Fisher à $(r, n - p)$ degrés de liberté.

16. On suppose $p \geq 2$ et on fixe $\alpha \in]0, 1[$.

Déterminer une région de confiance de niveau $1 - \alpha$ pour le vecteur (β_1, β_2) , où $\beta = (\beta_1, \dots, \beta_p)^\top$.
Quelle est sa nature géométrique ?

5 Applications numériques

On observe, pour $t = 1, \dots, n$,

$$X_t = \alpha_0 + \alpha_1 t + s_{r(t)} + \sigma \varepsilon_t,$$

où $(\alpha_0, \alpha_1) \in \mathbb{R}^2$, $(s_1, \dots, s_{T-1}) \in \mathbb{R}^{T-1}$, $s_T = -\sum_{k=1}^{T-1} s_k$ et $r(t) = 1 + ((t-1) \bmod T)$.

De plus, les variables $\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes et identiquement distribuées, avec $\varepsilon_i \sim \mathcal{N}(0, 1)$.

On définit les matrices $V_n \in \mathbb{R}^{n \times 2}$ et $C_n \in \mathbb{R}^{n \times (T-1)}$ par :

$$V_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & n \end{pmatrix}^\top, \quad (C_n)_{t,j} = \mathbf{1}_{\{r(t)=j\}} - \mathbf{1}_{\{r(t)=T\}}.$$

On pose également

$$Z_n = \begin{bmatrix} V_n & C_n \end{bmatrix} \in \mathbb{R}^{n \times (T+1)}, \quad \beta = (\alpha_0, \alpha_1, s_1, \dots, s_{T-1})^\top \in \mathbb{R}^{T+1}.$$

17. Vérifier que le modèle s'écrit

$$X = Z_n \beta + \sigma \varepsilon,$$

et montrer, en supposant $n \geq T + 1$, que Z_n est de rang $T + 1$.

On utilise les données fournies dans <https://www.insee.fr/fr/statistiques/serie/000436391> (téléchargeables au format `csv` par exemple), qui donnent le nombre de naissances en France métropolitaine chaque mois depuis janvier 1946 jusqu'à décembre 2025.

On prend $T = 12$ pour les 12 mois de l'année.

18. Visualiser les données sur la période janvier 1946 - décembre 2025.

On découpe désormais le temps en intervalles de 20 ans en se décalant de 10 ans : de janvier 1946 à décembre 1965 (où t varie de 1 à $n = 240$), de janvier 1956 à décembre 1975 (où t varie également de 1 à $n = 240$), etc.

19. Pour chaque intervalle de temps de 20 ans, calculer l'estimateur des moindres carrés $\hat{\beta}$ ainsi que $\hat{\sigma}^2$.

Tracer les courbes associées au modèle par-dessus les observations.

20. Pour chaque paramètre β_i et pour σ^2 , et pour chaque intervalle de temps, déterminer un intervalle de confiance de niveau de couverture 0,95.

21. Pour chaque intervalle de 20 ans pour lequel les 10 années suivantes sont observées dans le jeu de données, tracer les intervalles de prédiction pour ces 20 années suivantes, et les superposer aux observations.

Commenter.

Pour quelle(s) raison(s) a-t-on découpé en blocs de 20 ans, plutôt que d'utiliser directement toutes les valeurs ?

22. Dans cette question, on s'intéresse uniquement à la période comprise entre janvier 2006 et décembre 2025.

Pour chacune des questions suivantes, formuler la question comme un test d'hypothèse, construire une procédure de test, et calculer numériquement pour conclure.

(i) Y a-t-il en général autant de naissances en avril qu'en juin ?

(ii) Y a-t-il réellement un « effet saisonnier » (correspondant au vecteur s) ?

Réflexivité sur le travail en groupe

Pour chaque DM et EN, la **dernière question du devoir** vous demande d'expliquer brièvement l'organisation adoptée au sein du groupe : stratégie de travail, répartition des tâches, relectures, etc. Cette question a avant tout pour but d'induire un bref travail réflexif sur votre travail de groupe, pour, éventuellement, vous améliorer dans l'avenir – voir plus d'explication dans la feuille de consignes.

Cette question sera notée, mais elle n'a pas de "bonne réponse". Nous vous demandons simplement une réponse honnête et sérieuse, et vous obtiendrez les points dans ce cas.

Si l'un des membres du groupe ne participe pas au travail comme prévu, aucun membre ne sera pénalisé pour cela ; il conviendra simplement de décrire honnêtement la manière dont le travail a été réalisé. Si par hasard des désaccords forts apparaissent quant à votre perception du travail effectué (ce qui n'est pas souhaitable, mais peut survenir dans de rares cas), vous pouvez en faire état.

23. Répondre aux questions ci-dessous.

1. Quelle a été la répartition du travail au sein du groupe ? (qui a fait quoi ?)

Si vous avez procédé à une répartition des exercices : avez-vous pris un temps d'échange pour exposer votre travail aux autres ?

2. Une fois le DM terminé, quel membre du groupe maîtrise quelle partie du DM ?

3. Qu'est-ce qui a marché dans le travail en groupe ? Le fait de travailler en groupe vous a-t-il apporté ou appris quelque chose ?

4. Qu'est-ce qui *n'a pas* marché dans le travail en groupe ?

Que pourriez vous améliorer si c'était à refaire (ou pour les prochains devoirs) ?

5. Y a-t-il eu de l'entraide ? Si oui, qui a aidé qui, et en quoi ?