

HEC Paris – ESCP Europe

Mathématiques II

Mathématiques approfondies

Sujet adapté – 2012

Les parties du sujet contenant des éléments n'étant plus au programme depuis 2022 ont été adaptées pour convenir aux nouveaux programmes de la filière Mathématiques approfondies en vigueur depuis 2023.

Table des matières

1	Quelques résultats statistiques et algébriques	2
	<i>Thèmes abordés : estimations ; espaces vectoriels ; applications linéaires ; algèbre bilinéaire.</i>	2
2	Le modèle de régression linéaire	3
	<i>Thèmes abordés : variables à densité ; convergences de variables ; estimations ; applications linéaires.</i>	3
3	Hypothèse de normalité et prévision	4
	<i>Thèmes abordés : variables à densité ; estimations ; réduction ; applications linéaires. .</i>	4

Toutes les variables aléatoires qui interviennent dans ce problème sont réelles et définies sur un même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, où $\mathbb{P}$ peut dépendre de paramètres réels inconnus a, b, σ , etc. Elles admettent toutes une espérance et une variance ; si J désigne l'une de ces variables aléatoires, on note $\mathbb{E}(J)$ son espérance et $\mathbb{V}(J)$ sa variance.

Si J_1, J_2 et $J_1 + J_2$ sont des variables aléatoires à densité, on admet alors l'existence de la covariance de J_1 et J_2 , notée $\text{Cov}(J_1, J_2)$, qui est définie par

$$\text{Cov}(J_1, J_2) = \frac{1}{2}(\mathbb{V}(J_1 + J_2) - \mathbb{V}(J_1) - \mathbb{V}(J_2)).$$

On admet que les covariances de variables aléatoires à densité vérifient les mêmes règles de calcul que celles des variables aléatoires discrètes.

Pour tout (k, ℓ) de $(\mathbb{N}^*)^2$, on note $M_{k, \ell}(\mathbb{R})$ l'ensemble des matrices à k lignes et ℓ colonnes à coefficients réels ; on note $M_k(\mathbb{R})$ l'ensemble des matrices carrées d'ordre k . On note tQ la transposée d'une matrice Q . Dans tout le problème, n désigne un entier supérieur ou égal à 3.

L'objet du problème est l'étude de quelques propriétés du modèle de régression linéaire élémentaire.

1 Quelques résultats statistiques et algébriques

Thèmes abordés : estimations ; espaces vectoriels ; applications linéaires ; algèbre bilinéaire.

On considère une population d'individus statistiques dans laquelle on étudie deux caractères quantitatifs $\mathcal{X}$ et $\mathcal{Y}$. On extrait de cette population un échantillon de n individus sélectionnés selon des valeurs choisies du caractère $\mathcal{X}$ et numérotés de 1 à n .

Pour tout $i \in [1, n]$, les réels x_i et y_i sont les observations respectives de $\mathcal{X}$ et de $\mathcal{Y}$ pour l'individu i de l'échantillon. On suppose que les réels $x_1, x_2, \dots, x_n$ ne sont pas tous égaux.

Soient a et b deux paramètres réels. On pose, pour tout $i \in [1, n]$,

$$u_i = y_i - (ax_i + b). \quad (*)$$

1. On note $\bar{X}$ (resp. $\bar{Y}$) et s_X^2 (resp. s_Y^2) la moyenne empirique et la variance empirique de la série statistique $(x_i)_{1 \leq i \leq n}$ (resp. $(y_i)_{1 \leq i \leq n}$) ; on rappelle que

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2.$$

- a) Montrer que $s_X^2 > 0$.
- b) Établir les formules

$$\sum_{i=1}^n (x_i - \bar{X})y_i = \sum_{i=1}^n x_i y_i - n\bar{X}\bar{Y} \quad \text{et} \quad ns_X^2 = \sum_{i=1}^n x_i^2 - n\bar{X}^2.$$

- c) On pose, pour tout $i \in [1, n]$, $\alpha_i = \frac{x_i - \bar{X}}{ns_X^2}$. Montrer que

$$\sum_{i=1}^n \alpha_i = 0, \quad \sum_{i=1}^n \alpha_i x_i = 1, \quad \sum_{i=1}^n \alpha_i^2 = \frac{1}{ns_X^2}.$$

2. On pose

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in M_{n,1}(\mathbb{R}), \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \in M_{n,1}(\mathbb{R}),$$

$$\theta = \begin{pmatrix} a \\ b \end{pmatrix} \in M_{2,1}(\mathbb{R}), \quad M = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \in M_{n,2}(\mathbb{R}).$$

Les n relations $(*)$ s'écrivent sous la forme matricielle suivante : $y = M\theta + u$.

- a) Quel est le rang de la matrice M ?
 - b) Calculer la matrice tMM et justifier son inversibilité.
3. L'espace vectoriel $\mathbb{R}^n$ est muni de sa structure euclidienne canonique. Soit F le sous-espace vectoriel engendré par les vecteurs $(x_1, x_2, \dots, x_n)$ et $(1, 1, \dots, 1)$ de $\mathbb{R}^n$. On note K la matrice du projecteur orthogonal de $\mathbb{R}^n$ sur F dans la base canonique de $\mathbb{R}^n$ et $G = I - K$, où I désigne la matrice identité de $M_n(\mathbb{R})$.

- a) On cherche les matrices $\hat{\theta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ de $M_{2,1}(\mathbb{R})$ qui minimisent $\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - (\hat{a}x_i + \hat{b}))^2$. Montrer que ce problème admet une unique solution $\hat{\theta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ et qu'elle vérifie la relation ${}^tMM\hat{\theta} = {}^tMy$.
- b) Montrer que $\hat{a} = \sum_{i=1}^n \alpha_i y_i$ et $\hat{b} = \bar{Y} - \hat{a}\bar{X}$.
- c) Exprimer K en fonction de M et tM .
- d) Soit $\hat{u}$ la matrice-colonne de $M_{n,1}(\mathbb{R})$ de composantes $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_n$, définie par $\hat{u} = y - M\hat{\theta}$. Montrer que $\hat{u} = Gy = {}^tGu$.
- e) En déduire les égalités

$$\sum_{i=1}^n \hat{u}_i = 0 \quad \text{et} \quad \sum_{i=1}^n \hat{u}_i^2 = {}^t_y Gy = {}^t_u Gu.$$

2 Le modèle de régression linéaire

Thèmes abordés : variables à densité ; convergences de variables ; estimations ; applications linéaires.

Le contexte et les notations sont ceux de la partie I. Dans cette partie, on cherche à modéliser les fluctuations aléatoires du caractère $\mathcal{Y}$ sur l'échantillon.

Les hypothèses du modèle de régression linéaire élémentaire sont les suivantes :

- les réels a et b sont des paramètres inconnus ;
- pour tout $i \in [1, n]$, la valeur x_i du caractère $\mathcal{X}$ est connue et la valeur y_i du caractère $\mathcal{Y}$ est la réalisation d'une variable aléatoire Y_i ;
- pour tout $i \in [1, n]$, Y_i est la somme d'une composante déterministe $ax_i + b$, fonction affine de la valeur choisie x_i , et d'une composante aléatoire U_i ;
- les variables aléatoires $U_1, U_2, \dots, U_n$ sont mutuellement indépendantes, de même loi, possèdent

une densité, et, pour tout $i \in [1, n]$, $\mathbb{E}(U_i) = 0$ et $\mathbb{V}(U_i) = \sigma^2$, où le paramètre inconnu σ est strictement positif.

Le modèle de régression linéaire s'écrit alors : pour tout $i \in [1, n]$,

$$Y_i = ax_i + b + U_i. \quad (1)$$

L'objectif consiste à estimer les paramètres inconnus a , b et σ^2 du modèle (1).

On pose, pour tout $n \geq 3$, $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ et $\bar{U}_n = \frac{1}{n} \sum_{i=1}^n U_i$.

4. On note A_n et B_n les deux variables aléatoires définies par

$$A_n = \sum_{i=1}^n \alpha_i Y_i \quad \text{et} \quad B_n = \bar{Y}_n - A_n \bar{X},$$

où le réel α_i a été défini à la question 1.c.

- a) Montrer que A_n et B_n sont des estimateurs sans biais de a et b respectivement.
- b) Établir les formules suivantes :

$$\mathbb{V}(A_n) = \frac{\sigma^2}{ns_X^2} \quad \text{et} \quad \mathbb{V}(B_n) = \left(1 + \frac{\bar{X}^2}{s_X^2}\right) \frac{\sigma^2}{n}.$$

- c) Calculer $\text{Cov}(A_n, B_n)$.

5. Dans cette question uniquement, l'entier n n'est plus fixé. On suppose l'existence de

$$\lambda = \lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad \mu^2 = \lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2,$$

avec $(\lambda, \mu) \in \mathbb{R} \times \mathbb{R}_+^*$. Montrer que les deux suites $(A_n)_{n \geq 3}$ et $(B_n)_{n \geq 3}$ convergent en probabilité vers a et b respectivement.

6. a) On pose, pour tout $i \in [1, n]$, $\hat{U}_i = Y_i - A_n x_i - B_n$. Calculer $\mathbb{E}(\hat{U}_i)$.
- b) Établir l'égalité

$$\sum_{i=1}^n \hat{U}_i^2 = \sum_{i=1}^n (U_i - \bar{U}_n)^2 - ns_X^2 (A_n - a)^2.$$

- c) Calculer $\mathbb{E} \left(\sum_{i=1}^n \hat{U}_i^2 \right)$. En déduire un estimateur sans biais de σ^2 .

3 Hypothèse de normalité et prévision

Thèmes abordés : variables à densité ; estimations ; réduction ; applications linéaires.

Le contexte et les notations de cette partie sont ceux des parties I et II. De plus, on suppose dans cette partie que, pour tout $i \in [1, n]$, la variable aléatoire U_i suit une loi normale $\mathcal{N}(0, \sigma^2)$.

Pour $\theta > 0$ et $k > 0$, on dit qu'une variable aléatoire suit la loi Gamma à deux paramètres $\Gamma(\theta, k)$ lorsqu'elle admet pour densité

$$f(x) = \begin{cases} \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)} & \text{si } x > 0, \\ 0 & \text{si } x \leq 0. \end{cases}$$

Équivalamment, si X suit la loi gamma à un paramètre de forme k , notée $\gamma(k)$, alors θX suit la loi $\Gamma(\theta, k)$.

On pose

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}.$$

Le modèle (1) de la partie II s'écrit alors matriciellement : $Y = M\theta + U$.

Soient $W_1, W_2, \dots, W_q$ ($q \in \mathbb{N}^*$) q variables aléatoires réelles définies sur $(\Omega, \mathcal{A}, \mathbb{P})$. On définit le vecteur aléatoire $(W_1, W_2, \dots, W_q)$ à valeurs dans $\mathbb{R}^q$, en associant à tout $\omega \in \Omega$ le vecteur $(W_1(\omega), W_2(\omega), \dots, W_q(\omega))$ de $\mathbb{R}^q$.

On dit que le vecteur aléatoire $(W_1, W_2, \dots, W_q)$ est normal si, pour tout q -uplet $(p_1, p_2, \dots, p_q)$ de nombres réels, différent de $(0, 0, \dots, 0)$, la variable aléatoire $\sum_{i=1}^q p_i W_i$ suit une loi normale de variance non nulle.

Dans le cas où le vecteur $(W_1, W_2, \dots, W_q)$ est normal, on admet que les variables aléatoires $W_1, W_2, \dots, W_q$ sont mutuellement indépendantes si et seulement si, pour tout $(i, j) \in [1, q]^2$ avec $i \neq j$, $\text{Cov}(W_i, W_j) = 0$.

7. a) Montrer que le vecteur aléatoire $(Y_1, Y_2, \dots, Y_n)$ est normal mais que le vecteur $(Y_1 - \bar{Y}_n, Y_2 - \bar{Y}_n, \dots, Y_n - \bar{Y}_n)$ ne l'est pas.
 b) Déterminer la loi de chacune des variables aléatoires A_n et B_n . Le vecteur aléatoire (A_n, B_n) est-il normal ?
8. Soit S une matrice inversible de $M_n(\mathbb{R})$. On note T la matrice-colonne des composantes du vecteur aléatoire $(T_1, T_2, \dots, T_n)$ telle que $T = SU$.
 a) Montrer que le vecteur $(T_1, T_2, \dots, T_n)$ est normal.
 b) On suppose que la matrice S est orthogonale. Montrer que $T_1, T_2, \dots, T_n$ sont mutuellement indépendantes.
9. Soient $\hat{U}_1, \hat{U}_2, \dots, \hat{U}_n$ les variables aléatoires qui ont été définies dans la question 6. On note $\hat{U}$ la matrice-colonne de composantes $\hat{U}_1, \hat{U}_2, \dots, \hat{U}_n$, définie par $\hat{U} = Y - M\hat{\theta}$.
 a) Montrer que $\hat{U} = GU$, où la matrice G a été définie dans la question 3.
 b) Justifier l'existence d'une matrice orthogonale R de $M_n(\mathbb{R})$ et d'une matrice diagonale D de $M_n(\mathbb{R})$, telles que $G = RD^t R$. Quels sont les éléments diagonaux de D ?
 c) Soit Z la matrice-colonne de composantes $Z_1, Z_2, \dots, Z_n$, définie par $Z = {}^t R U$. Quelle est la loi de $\sum_{i=1}^{n-2} Z_i^2$?
 d) En déduire que la variable aléatoire $\sum_{i=1}^n \hat{U}_i^2$ suit la loi $\Gamma\left(2\sigma^2, \frac{n-2}{2}\right)$.
 e) Soit p un réel donné vérifiant $0 < p < 1$. Établir l'existence d'un réel c_n , ne dépendant pas des paramètres inconnus a, b et σ^2 , tel que

$$\mathbb{P}\left(\left[\sum_{i=1}^n \hat{U}_i^2 \geq c_n \sigma^2\right]\right) = p.$$

Dans les questions 10 et 11, on suppose qu'une $(n+1)$ -ième valeur de $\mathcal{X}$, notée x_{n+1} , est choisie mais que la valeur correspondante y_{n+1} de $\mathcal{Y}$ est inconnue. On suppose que y_{n+1} est la réalisation d'une variable aléatoire Y_{n+1} qui vérifie $Y_{n+1} = ax_{n+1} + b + U_{n+1}$, où les variables aléatoires $U_1, U_2, \dots, U_{n+1}$ sont mutuellement indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

10. On pose, pour tout n -uplet $r = (r_1, r_2, \dots, r_n)$ de $\mathbb{R}^n$,

$$\widehat{Y}_{n+1}^{(r)} = \sum_{i=1}^n r_i Y_i.$$

L'ensemble $\{\widehat{Y}_{n+1}^{(r)} ; r \in \mathbb{R}^n\}$ est l'ensemble des « prédictors linéaires » de Y_{n+1} .

- a) Soit g la fonction définie sur $\mathbb{R}^n$ à valeurs réelles, telle que, pour tout $r = (r_1, r_2, \dots, r_n)$ de $\mathbb{R}^n$,

$$g(r_1, r_2, \dots, r_n) = \sum_{i=1}^n r_i^2.$$

On rappelle que, pour tout $i \in [1, n]$, $\alpha_i = \frac{x_i - \bar{X}}{ns_X^2}$. Montrer que la fonction g admet un minimum absolu sous les contraintes $\sum_{i=1}^n r_i = 1$ et $\sum_{i=1}^n x_i r_i = x_{n+1}$, atteint en l'unique point $r^* = (r_1^*, r_2^*, \dots, r_n^*)$, où, pour tout $i \in [1, n]$,

$$r_i^* = \frac{1}{n} + (x_{n+1} - \bar{X})\alpha_i.$$

- b) Montrer que parmi les prédictors linéaires $\widehat{Y}_{n+1}^{(r)}$ de Y_{n+1} qui vérifient $\mathbb{E}(\widehat{Y}_{n+1}^{(r)}) = \mathbb{E}(Y_{n+1})$ pour tout $(a, b) \in \mathbb{R}^2$, $\widehat{Y}_{n+1}^{(r^*)}$ est celui qui a la plus petite variance. Vérifier que $\widehat{Y}_{n+1}^{(r^*)} = A_n x_{n+1} + B_n$.
11. a) Déterminer la loi de la variable aléatoire $Y_{n+1} - (A_n x_{n+1} + B_n)$.
- b) On note Φ la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Soit p un réel donné vérifiant $\frac{1}{2} < p < 1$. Justifier l'existence d'un réel d_n , que l'on exprimera à l'aide de Φ^{-1} , ne dépendant pas de a , b et σ^2 , tel que

$$\mathbb{P}(|Y_{n+1} - (A_n x_{n+1} + B_n)| \leq d_n \sigma) = p.$$

- c) En déduire, à l'aide de la question 9.e, un intervalle dont les bornes ne dépendent que des $(Y_i)_{1 \leq i \leq n}$, des $(x_i)_{1 \leq i \leq n+1}$, de c_n et d_n , qui contienne Y_{n+1} avec une probabilité supérieure ou égale à $2p - 1$. S'agit-il d'un intervalle de confiance au sens usuel du terme ?