Université de Caen — *Correction de l'*Examen du 20 avril 2018. Durée 2h00 — **Analyse de données**

UFR des Sciences — Par F. Chamroukhi — **L3 Maths/Miashs**

Consignes :
- Sont interdits : Documents, calculettes, téléphones, écouteurs, ordinateurs, tablettes.
- Il est interdit de composer avec un crayon.
- Votre feuille double d'examen doit porter, à l'emplacement réservé, vos nom, prénom, et signature.
- Cette zone réservée doit être cachée par collage.
- Vos feuilles intercalaires doivent être toutes numérotées.
- Le barème est donné à titre indicatif.

Exercice 1 *(4 pts)* Soit (X, Y) un coupe de variables aléatoires réelles et soit $((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon de n observations. Chacune des situations présentées dans la Figure 1 représente le nuage de données d'un échantillon de taille $n = 500$. Pour chaque situation, donner une valeur approchée du cœfficient de corrélation linéaire empirique r et justifier votre réponse.

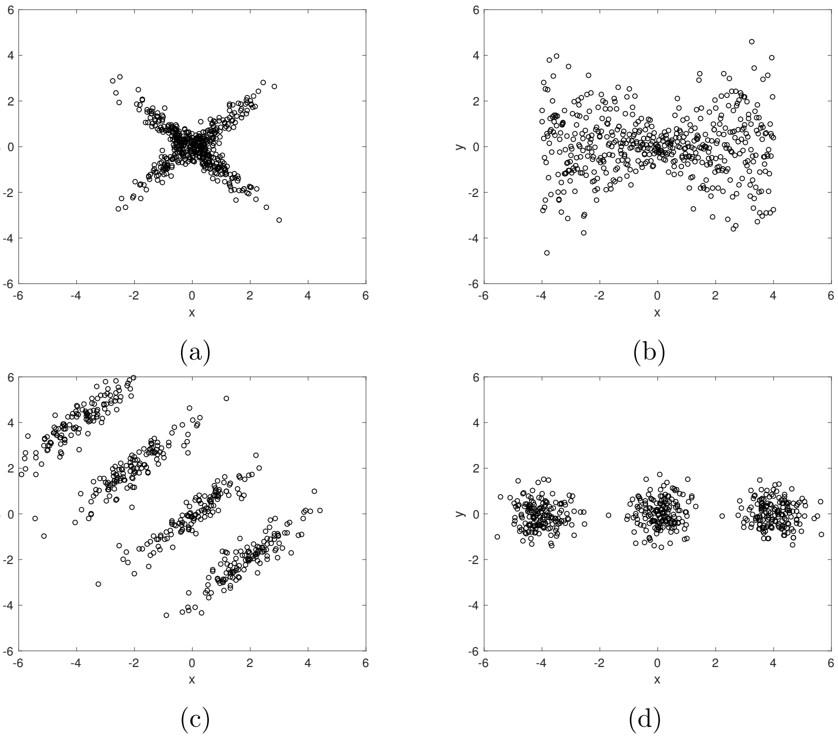

Figure 1 – Nuages de données (○)

Solution 1 Voir TD ((a,b,b) r très proche de 0 car absence de lien linéaire entre les deux variables ; (c) corrélation négative (anti-corrélation), paradoxe de Simpson, r proche de -0.7/8)

Exercice 2 *(6 pts)* On considère un jeu de données multivariées $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où $\boldsymbol{x}_i \in \mathbb{R}^p$ est un individu décrit par p variables réelles. On suppose que les variables sont potentiellement corrélées et que p est potentiellement grand, et on cherche à réduire la dimension en projetant ce jeu de données dans un espace de dimension M plus réduite et ce au sens de l'ACP. Soient $\boldsymbol{x} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ le vecteur de la moyenne empirique des individus et $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ la matrice de variances-covariances empirique.

1. On suppose que l'on projette les données dans un espace de dimension 1, i.e., une droite de vecteur directeur $\boldsymbol{u}_1$, que l'on suppose normé (i.e.,$\boldsymbol{u}^\top \boldsymbol{u} = 1$). Donner l'expression de la variance des données projetées dans cet espace et exprimer là en fonction de $\mathbf{S}$ et $\boldsymbol{u}_1$. On la notera v_1.

2. En sachant que l'ACP consiste à maximiser la variance dans l'espace projeté v_1, montrer que le vecteur $\boldsymbol{u}_1$ qui définissant cet espace correspond au vecteur propre associé à la plus grande valeur propre de $\mathbf{S}$.
3. On suppose maintenant que l'on projette les données dans un espace de dimension $M \leq p$ déterminé par le sous espace vectoriel de base orthonormée $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$, i.e., $\boldsymbol{u}_j^\top \boldsymbol{u}_j = 1 \ \forall j$ et $\boldsymbol{u}_j^\top \boldsymbol{u}_k = 0 \ \forall j \neq k$. Montrer que les vecteurs de cette base sont les vecteurs propres de $\mathbf{S}$ organisés selon l'ordre décroissant des valeurs propres associées.
4. En interprétant le lien entre les valeurs propres et les variances expliquées par chaque axe de l'espace projeté, donner une idée pour choisir la dimension M.

Solution 2

1. La coordonnée de la projection linéaire du vecteur $\boldsymbol{x}_i$ dans l'espace défini par le vecteur $\boldsymbol{u}_1$ est $\boldsymbol{x}_i^\top \boldsymbol{u}_1$. Il en est de même pour $\bar{\boldsymbol{x}}$. La variance empirique des données projetées sur dans espace est donc donnée par :

$$
\begin{aligned}
v_1 = v(\boldsymbol{u}_1) &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top \boldsymbol{u}_1 - \bar{\boldsymbol{x}}^\top \boldsymbol{u}_1)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top \boldsymbol{u}_1 - \bar{\boldsymbol{x}}^\top \boldsymbol{u}_1)^\top(\boldsymbol{x}_i^\top \boldsymbol{u}_1 - \bar{\boldsymbol{x}}^\top \boldsymbol{u}_1) \\
&= \boldsymbol{u}_1^\top \Big(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top\Big)\boldsymbol{u}_1 \\
&= \boldsymbol{u}_1^\top \mathbf{S} \boldsymbol{u}_1 \cdot
\end{aligned}
$$

2. L'ACP consiste à maximiser la variance dans l'espace projeté v_1. Le vecteur $\boldsymbol{u}_1$ définissant cet espace est donc déterminé en résolvant le problème suivant :

$$
\begin{aligned}
\boldsymbol{u}_1 &= \arg\max_{\boldsymbol{u}_1 \in \mathbb{R}^d} v(\boldsymbol{u}_1) \text{ sous la contrainte } \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1 \\
&= \arg\max_{\boldsymbol{u}_1 \in \mathbb{R}^d} \boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1 \text{ sous la contrainte } \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1.
\end{aligned}
$$

Pour résoudre ce problème d'optimisation sous contrainte, on passe par le Lagrangien définissant le critère à maximiser tenant compte de la contrainte et défini par

$$
\ell(\boldsymbol{u}_1) = \boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^T \boldsymbol{u}_1) \tag{1}
$$

où$\lambda \in \mathbb{R}$ est le multiplicateur inconnu de Lagrange et on a donc

$$
\begin{aligned}
\boldsymbol{u}_1 &= \arg\max_{\boldsymbol{u}_1 \in \mathbb{R}^d} \ell(\boldsymbol{u}_1) \\
&= \arg\max_{\boldsymbol{u}_1 \in \mathbb{R}^d} \boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^T \boldsymbol{u}_1) \cdot
\end{aligned}
$$

En dérivant et en annulant la fonction $\ell(\boldsymbol{u}_1)$ par rapport à $\boldsymbol{u}_1$, on obtient :

$$
\begin{aligned}
\frac{\partial\left\{\boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1 + \lambda(1 - \boldsymbol{u}_1^T \boldsymbol{u}_1)\right\}}{\partial \boldsymbol{u}_1} = 0 \quad &\overset{\frac{\partial \boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1}{\partial \boldsymbol{u}_1} = (\mathbf{S}+\mathbf{S}^T)\boldsymbol{u}}{\Longrightarrow} \quad 2\mathbf{S}\boldsymbol{u}_1 - 2\lambda_1 \boldsymbol{u}_1 = 0 \\
&\Rightarrow \quad \mathbf{S}\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1.
\end{aligned}
$$

D'après cette dernière équation, on voit donc que le vecteur $\boldsymbol{u}_1$ correspond au vecteur propre associé à la valeur propre λ_1 de $\mathbf{S}$. Comme on cherche à maximiser la variance, la valeur propre λ_1 doit donc être la plus grande des valeurs propres de $\mathbf{S}$.

3. On suppose que l'on projette les données dans un espace de dimension $M \leq p$ déterminé par le sous espace vectoriel de base orthonormée $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$, i.e., $\boldsymbol{u}_j^\top \boldsymbol{u}_j = 1 \, \forall j$ et $\boldsymbol{u}_j^\top \boldsymbol{u}_k = 0 \, \forall j \neq k$. Montrons que les vecteurs de cette base sont les vecteurs propres de $\mathbf{S}$ organisés selon l'ordre décroissant des valeurs propres associées $(\lambda_1, \ldots, \lambda_M)$. Nous venons de montrer que ce résultat est vrai pour $M = 1$. Supposant qu'il est aussi vrai pour M. Alors pour le $(M+1)-$ème vecteur de la base de l'espace projetée $\boldsymbol{u}_{M+1}$ qui est donc

   (i) normé, i.e : $\boldsymbol{u}_{M+1}^T \boldsymbol{u}_{M+1} = 1$,

   (ii) orthogonal aux autres vecteurs $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ de la base, i.e : $\boldsymbol{u}_{M+1}^T \boldsymbol{u}_k = 0 \; \forall \; k \neq M+1$,

   doit vérifier que la variance expliquée par l'axe dont il définit, notée $v(\boldsymbol{u}_{M+1})$ et donnée par $v(\boldsymbol{u}_{M+1}) = \boldsymbol{u}_{M+1}^T \mathbf{S} \boldsymbol{u}_{M+1}$, est maximale. En introduisant le multiplicateur de Lagrange λ_{M+1} pour intégrer la contrainte (i) et les M multiplicateurs $\eta_k, k = 1, \ldots, M$ pour les M contraintes (ii), cela correspond donc à résoudre le problème d'optimisation suivant :

$$\boldsymbol{u}_{M+1} = \arg \max_{\boldsymbol{u}_{M+1}} \ell(\boldsymbol{u}_{M+1})$$

   où

$$\ell(\boldsymbol{u}_{M+1}) = \boldsymbol{u}_{M+1}^T \mathbf{S} \boldsymbol{u}_{M+1} + \lambda_{M+1}(1 - \boldsymbol{u}_{M+1}^T \boldsymbol{u}_{M+1}) + \sum_{k=1}^{M} \eta_k \boldsymbol{u}_{M+1}^T \boldsymbol{u}_k$$

   est le nouveau Lagrangien. En dérivant et en annulant la dérivée on trouve

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{u}_{M+1})}{\partial \boldsymbol{u}_{M+1}} = 0 \quad &\Rightarrow \quad 2\mathbf{S}\boldsymbol{u}_{M+1} - 2\lambda_{M+1}\boldsymbol{u}_{M+1} + \sum_{k=1}^{M} \eta_k \boldsymbol{u}_k = 0 \\
&\Rightarrow \quad \underbrace{\boldsymbol{u}_j^T \mathbf{S} \boldsymbol{u}_{M+1}}_{0} - \lambda_{M+1} \underbrace{\boldsymbol{u}_j^T \boldsymbol{u}_{M+1}}_{0} + \underbrace{\sum_{k=1}^{M} \eta_k \boldsymbol{u}_j^T \boldsymbol{u}_k}_{\eta_j \boldsymbol{u}_j^T \boldsymbol{u}_j = \eta_j} = 0 \\
&\Rightarrow \quad \eta_j = 0 \quad \text{pour tout} \quad j = 1, \ldots, M
\end{aligned} \tag{2}$$

   D'après (2), nous obtenons donc :

$$\mathbf{S}\boldsymbol{u}_{M+1} = \lambda_{M+1} \boldsymbol{u}_{M+1}.$$

   Alors $\boldsymbol{u}_{M+1}$ est nécessairement le $(M+1)$-ème vecteur propre de $\mathbf{S}$ associé à la valeur propre λ_{M+1}. La variance projetée dans la direction $\boldsymbol{u}_{M+1}$ est donnée par

$$\begin{aligned}
\boldsymbol{u}_{M+1}^T \mathbf{S} \boldsymbol{u}_{M+1} &= \boldsymbol{u}_{M+1}^T \lambda_{M+1} \boldsymbol{u}_{M+1} \\
&= \lambda_{M+1}.
\end{aligned}$$

   Elle est donc maximale si on prend $\boldsymbol{u}_{M+1}$ le vecteur propore associée à la plus grande valeur propre parmi celles qui n'ont pas été déjà sélectionnées.
   Le résultat que l'on cherchait à montrer est donc vrai pour la dimension $M+1$. Par récurrence, il est donc vrai pour tout $M \leq p$.

4. On vient de voir que la variance expliquée par chaque axe de l'espace projeté correspond à la valeur propre associée au vecteur propre définissant cet axe. Comme l'objectif de l'ACP est de maximiser la variance dans l'espace projetée, on peut donc retenir les axes qui explique un certain pourcentage cumulé des variances suffisamment grand (e.g 85%, 90%, ou 95%) de la variance totale de façon à ne pas perdre trop d'information. Cela correspond à prendre la valeur maximale de M vérifiant $100 \left(\frac{\sum_{m=1}^{M} \lambda_m}{\sum_{m=1}^{p} \lambda_m} \right) \% \leq \alpha$ où α vaut par exemple 85%, 90% ou 95%.

Exercice 3 *(10 pts)* On considère un échantillon aléatoire indépendant $((X_1, Y_1), \ldots, (X_n, Y_n))$ du couple (X, Y) où $X \in \mathbb{R}$ est une variable explicative et $Y \in \mathbb{R}$ une variable expliquée à prédire. On considère le modèle de régression linéaire pondérée suivant

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3}$$

où les ε_i sont des v.a indépendantes Gaussiennes centrées de variance $\frac{\sigma^2}{w_i}$, $w_i > 0$ est le poids connu de la $i$ème observation, et $(\beta_0, \beta_1) \in \mathbb{R}^2$ sont les deux cœfficients de régression inconnus. On dispose d'un échantillon indépendant d'observations $((x_1, y_1), \ldots, (x_n, y_n))$ et les poids associées $(w_1, \ldots, w_n)$, à partir desquels on cherche à estimer les paramètres (β_0, β_1) par maximum de vraisemblance.

On sait facilement montrer que selon le modèle (3) la loi de Y_i conditionnellement à $X_i = x_i$, est Gaussienne d'espérance $\beta_0 + \beta_1 x_i$ et de variance $\frac{\sigma^2}{w_i}$ de densité notée $f(y_i|X_i = x_i; \beta_0, \beta_1)$.

1. On rappelle que si Z est une v.a Gaussienne d'espérance μ et de variance v^2, sa densité est alors donnée par $f(z; \mu, v^2) = \frac{1}{\sqrt{2\pi v^2}} \exp\left(-\frac{1}{2v^2}(z-\mu)^2\right)$. En déduire la densité $f(y_i|X_i = x_i; \beta_0, \beta_1)$.
2. Donner l'expression de la fonction de log-vraisemblance
$$\ln L(\beta_0, \beta_1) = \ln f(y_1, \ldots, y_n|X_1 = x_1, \ldots, X_n = x_n; \beta_0, \beta_1). \tag{4}$$
3. Montrer en maximisant (6) que l'estimateur du maximum de vraisemblance (EMV) $\widehat{\beta}_0$ de β_0 et l'EMV $\widehat{\beta}_1$ de β_1 sont donnés par :
$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y}_w - \widehat{\beta}_1 \overline{X}_w, \\ \widehat{\beta}_1 &= \sum_{i=1}^n w_i (X_i - \overline{X}_w)(Y_i - \overline{Y}_w) \Big/ \sum_{i=1}^n (X_i - \overline{X}_w)^2,\end{aligned}$$
où $\overline{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$ est la moyenne empirique pondérée des prédicteurs X_i et $\overline{Y}_w = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$ celle des réponses Y_i.
4. On considère maintenant la formulation matricielle du problème. Soit $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$ le vecteur paramètre du modèle. En remarquant que maximiser (6) revient à minimiser la fonction
$$R_w(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \mathbf{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$
où $\mathbf{W} \in \mathbb{R}^{n \times n}$ est la matrice diagonale de termes les poids $(w_1, \ldots, w_n)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ est le vecteur des réponses, et $\boldsymbol{X} = ((1, X_1)^\top, \ldots, (1, X_n)^\top)^\top \in \mathbb{R}^{n \times 2}$ est la matrice de design, montrer que l'EMV $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est donné par
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \mathbf{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{W} \boldsymbol{Y}.$$
5. A quoi correspond l'EMV dans ce cas ?

Solution 3

1. La loi de Y_i conditionnellement à $X_i = x_i$ est Gaussienne d'espérance $\beta_0 + \beta_1 x_i$ et de variance $\frac{\sigma^2}{w_i}$. Sa densité $f(y_i|x_i; \beta_0, \beta_1)$ est donc donnée par
$$f(y_i|x_i; \beta_0, \beta_1) = \frac{1}{\sqrt{2\pi \frac{\sigma^2}{w_i}}} \exp\left(-\frac{1}{2\frac{\sigma^2}{w_i}}\big(y_i - (\beta_0 + \beta_1 x_i)\big)^2\right) = \frac{\sqrt{w_i}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} w_i \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2\right). \tag{5}$$

2. L'expression de la fonction de log-vraisemblance pour cet échantillon i.i.d est donnée par :

$$\begin{aligned}\ln L(\beta_0,\beta_1) &= \ln f(y_1,\ldots,y_n|X_1=x_1,\ldots,X_n=x_n;\beta_0,\beta_1)\\ &= \ln\prod_{i=1}^n f(y_i|x_i;\beta_0,\beta_1)\\ &= \ln\prod_{i=1}^n \frac{\sqrt{w_i}}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{1}{2\sigma^2}w_i\big(y_i-(\beta_0+\beta_1x_i)\big)^2\Big)\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1x_i)\big)^2+c. \qquad (6)\end{aligned}$$

où c est une constante qui ne dépend ni de β_0 ni de β_1.

3. Montrer en maximisant (6) que l'estimateur du maximum de vraisemblance (EMV) $\widehat{\beta}_0$ de β_0 et l'EMV $\widehat{\beta}_1$ de β_1 sont donnés par :

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y}_w-\widehat{\beta}_1\overline{X}_w,\\ \widehat{\beta}_1 &= \sum_{i=1}^n w_i(X_i-\overline{X}_w)(Y_i-\overline{Y}_w)\Big/\sum_{i=1}^n w_i(X_i-\overline{X}_w)^2,\end{aligned}$$

où $\overline{X}_w=\frac{\sum_{i=1}^n w_iX_i}{\sum_{i=1}^n w_i}$ est la moyenne empirique pondérée des prédicteurs X_i et $\overline{Y}_w=\frac{\sum_{i=1}^n w_iY_i}{\sum_{i=1}^n w_i}$ celle des réponses Y_i.

On voit que maximiser (6) par rapport aux coefficients β' est équivalent à minimiser la somme pondérée des carrés des résidus $R_w(\beta_0,\beta_1)=\sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1x_i)\big)^2$ et on peut donc écrire

$$\begin{aligned}(\hat{\beta}_0,\hat{\beta}_1) &= \arg\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} R_w(\beta_0,\hat{\beta}_1)\\ &= \arg\min_{(\beta_0,\beta_1)\in\mathbb{R}^2}\sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1x_i)\big)^2. \qquad (7)\end{aligned}$$

Résoudre cette équation revient à prendre les paramètres (β_0,β_1) qui annulent ses dérivées partielles premières et pour lesquels les dérivées secondes sont positives. Ainsi, pour β_0 nous avons

$$\frac{\partial R_w}{\partial\beta_0} = \frac{\partial\sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1x_i)\big)^2}{\partial\beta_0} = -2\sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1x_i)\big)$$

En annulant cette dérivée nous avons :

$$\sum_{i=1}^n w_iy_i-\hat{\beta}_0\sum_{i=1}^n w_i-\hat{\beta}_1\sum_{i=1}^n w_ix_i = 0,$$

ce qui donne

$$\begin{aligned}\beta_0 &= \frac{\sum_{i=1}^n w_iy_i}{\sum_{i=1}^n w_i}-\beta_1\frac{\sum_{i=1}^n w_ix_i}{\sum_{i=1}^n w_i}\\ &= \bar{y}_w-\beta_1\bar{x}_w. \qquad (8)\end{aligned}$$

Pour β_1, en remplaçant l'expression de β_0 (qui est fonction de β_1) dans l'expression de R_w, nous avons

$$\begin{aligned}
\frac{\partial R_w(\beta_0,\beta_1)}{\partial \beta_1} &= \frac{\partial \sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1 x_i)\big)^2}{\partial \beta_1}\\
&= \frac{\partial \sum_{i=1}^n w_i\big(y_i-(\bar{y}_w-\beta_1\bar{x}_w+\beta_1 x_i)\big)^2}{\partial \beta_1}\\
&= \frac{\partial \sum_{i=1}^n w_i\big(y_i-\bar{y}_w-\beta_1(x_i-\bar{x}_w)\big)^2}{\partial \beta_1}\\
&= 2\sum_{i=1}^n w_i(x_i-\bar{x}_w)\big(y_i-\bar{y}_w-\beta_1(x_i-\bar{x}_w)\big)\\
&= 2\sum_{i=1}^n w_i(x_i-\bar{x}_w)(y_i-\bar{y}_w)-2\beta_1\sum_{i=1}^n w_i(x_i-\bar{x}_w)^2.
\end{aligned}$$

En annulant cette dérivée on trouve :

$$\hat{\beta}_1=\frac{\sum_{i=1}^n w_i(x_i-\bar{x}_w)(y_i-\bar{y}_w)}{\sum_{i=1}^n w_i(x_i-\bar{x}_w)^2}\,. \tag{9}$$

D'après (8) et (9), es estimateurs sont donc donnés par :

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y}_w-\widehat{\beta}_1\overline{X}_w,\\
\widehat{\beta}_1 &= \sum_{i=1}^n w_i(X_i-\overline{X}_w)(Y_i-\overline{Y}_w)\Big/\sum_{i=1}^n (X_i-\overline{X}_w)^2.
\end{aligned}$$

4. On considère maintenant la formulation matricielle du problème. Soit $\boldsymbol{\beta}=(\beta_0,\beta_1)^\top\in\mathbb{R}^2$ le vecteur paramètre du modèle. On voit que maximiser (6) par rapport au vecteur $\boldsymbol{\beta}$ est équivalent à minimiser la somme pondérée des carrés des résidus $R_w(\boldsymbol{\beta})=\sum_{i=1}^n w_i\big(y_i-(\beta_0+\beta_1 x_i)\big)^2$ qui s'écrit sous la forme matricielle suivante :

$$R_w(\boldsymbol{\beta})=(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top\mathbf{W}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$$

L'EMV de $\boldsymbol{\beta}$ est donc donné par :

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^2} R_w(\boldsymbol{\beta})\\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top\mathbf{W}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\boldsymbol{Y}^\top\mathbf{W}\boldsymbol{Y}-(\boldsymbol{X}\boldsymbol{\beta})^\top\mathbf{W}\boldsymbol{Y}-\boldsymbol{Y}^\top\mathbf{W}(\boldsymbol{X}\boldsymbol{\beta})+(\boldsymbol{X}\boldsymbol{\beta})^\top\mathbf{W}(\boldsymbol{X}\boldsymbol{\beta})\\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\boldsymbol{Y}^\top\mathbf{W}\boldsymbol{Y}-2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\mathbf{W}\boldsymbol{Y}+\boldsymbol{\beta}^\top\boldsymbol{X}^T\mathbf{W}\boldsymbol{X}\boldsymbol{\beta}\cdot
\end{aligned}$$

En annulant la dérivée, qui est donnée par

$$\frac{\partial R_w(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\boldsymbol{X}^\top\mathbf{W}\boldsymbol{Y}+2\boldsymbol{X}^T\mathbf{W}\boldsymbol{X}\boldsymbol{\beta}$$

on trouve

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{W}\boldsymbol{Y}.$$

5. L'EMV dans ce cas est celui des moindres carrés pondérés