Ajustement affine d'une série statistique double : droite des moindres carrés

1 Série statistique pondérée et série statistique double pondérée

Définition 1 (série statistique pondérée)**.** *Une série statistique pondérée X est la donnée de n réels pondérés $(x_1, p_1), \ldots, (x_n, p_n)$ où les poids sont des réels strictement positifs vérifiant $p_1 + \cdots + p_n = 1$.*

Par défaut, lorsque les poids ne sont pas précisés, on prendra $p_1 = \cdots = p_n = 1/n$.

Définition 2 (moyenne et variance d'une série statistique pondérée)**.** *Soit $X = \{(x_1, p_1), \ldots, (x_n, p_n)\}$ une série statistique pondérée. La moyenne pondérée et la variance pondérée de la série X valent :*

$$\overline{X} = \sum_{i=1}^{n} p_i x_i \text{ et } v(X) = \sum_{i=1}^{n} p_i (x_i - \overline{X})^2$$

Par définition de la variance $v(X) \geq 0$ de plus, $v(X) = 0$ si et seulement si $x_1 = \cdots = x_n$.

Définition 3 (série statistique double pondérée)**.** *Une série statistique double pondérée est la donnée de n couples de $\mathbb{R}^2$ pondérés $((x_1, y_1), p_1), \ldots, ((x_n, y_n), p_n)$ où les poids sont des réels strictement positifs vérifiant $p_1 + \cdots + p_n = 1$.*

Définition 4 (Covariance d'une série statistique double pondérée)**.** *Soit $(X, Y) = \{((x_1, y_1), p_1), \ldots, ((x_n, y_n), p_n)\}$ une série statistique double pondérée. La covariance pondérée entre les séries X et Y vaut*

$$cov(X, Y) = \sum_{i=1}^{n} p_i (x_i - \overline{X})(y_i - \overline{Y})$$

Proposition 1. *Soit $X = \{(x_1, p_1), \ldots, (x_n, p_n)\}$ une série statistique. On pose $\overline{X^2} = \sum_{i=1}^{n} p_i x_i^2$. Alors, la variance pondérée de X vérifie*

$$v(X) = \overline{X^2} - \overline{X}^2.$$

Soit $(X, Y) = ((x_1, y_1), p_1), \ldots, ((x_n, y_n), p_n)$ une série statistique double. On pose $\overline{XY} = \sum_{i=1}^{n} p_i x_i y_i$. Alors, la covariance pondérée de (X, Y) vérifie

$$cov(X, Y) = \overline{XY} - \overline{X} \times \overline{Y}$$

Démonstration. Comme, par définition, $v(X) = cov(X, X)$, la première propriété est une conséquence immédiate de la seconde propriété dont la preuve est donnée ci-dessous.

$$\begin{aligned} cov(X, Y) &= \sum_{i=1}^{n} p_i (x_i - \overline{X})(y_i - \overline{Y}) \\ &= \sum_{i=1}^{n} (p_i x_i y_i) - \overline{Y} \sum_{i=1}^{n} p_i x_i - \overline{X} \sum_{i=1}^{n} p_i y_i + \overline{X} \times \overline{Y}, \\ &= \overline{XY} - \overline{Y} \times \overline{X} - \overline{X} \times \overline{Y} + \overline{X} \times \overline{Y}, \\ &= \overline{XY} - \overline{X} \times \overline{Y}. \end{aligned}$$

□

Exemple 1. *Sept groupes d'étudiants ont réalisé un projet scientifique. Pour l'évaluation X est la note sur quinze du rapport écrit et Y est la note sur cinq de la soutenance orale. Les notes sont recueillis dans une série statistique double pondérée (X, Y) donnée ci-dessous où le poids est la taille du groupe sur l'effectif total :*

$$(X,Y) = \{((12,4),0.1),((9,3),0.1),((6,3),0.15),((5,2),0.15),((13,4),0.2),((8,2),0.1),((7,1),0.2)\}.$$

Les moyennes, variances et la covariance de cette série statistique double pondérée sont données ci-dessous :

— *La moyenne pondérée des notes des rapports vaut :*

$$\overline{X} = 12 \times 0.1 + 9 \times 0.1 + 6 \times 0.15 + 5 \times 0.15 + 13 \times 0.2 + 8 \times 0.1 + 7 \times 0.2 = 8.55.$$

— *La moyenne pondérée des notes des soutenances orales vaut :*

$$\overline{Y} = 4 \times 0.1 + 3 \times 0.1 + 3 \times 0.15 + 2 \times 0.15 + 4 \times 0.2 + 2 \times 0.1 + 1 \times 0.2 = 2.65.$$

— *Le carré pondéré moyen pour la série X est égal à :*

$$\overline{X^2} = 144 \times 0.1 + 81 \times 0.1 + 36 \times 0.15 + 25 \times 0.15 + 169 \times 0.2 + 64 \times 0.1 + 49 \times 0.2 = 81.65.$$

Ainsi, la variance des notes des rapports écrits vaut $v(X) = \overline{X^2} - \overline{X}^2 = 8.5475$.

— *Le carré pondéré moyen pour la série Y est égal à :*

$$\overline{Y^2} = 16 \times 0.1 + 9 \times 0.1 + 9 \times 0.15 + 4 \times 0.15 + 16 \times 0.2 + 4 \times 0.1 + 1 \times 0.2 = 8.25.$$

Ainsi la variance des notes des soutenances orales vaut $v(Y) = \overline{Y^2} - \overline{Y}^2 = 1.2275$.

— *Le produit pondéré moyen de la série double (X, Y) est égal à :*

$$\overline{XY} = 48 \times 0.1 + 27 \times 0.1 + 18 \times 0.15 + 10 \times 0.15 + 52 \times 0.2 + 16 \times 0.1 + 7 \times 0.2 = 25.1.$$

Ainsi, la covariance entre les notes des rapports écrits et des notes des soutenances orales vaut $cov(X, Y) = \overline{XY} - \overline{X} \times \overline{Y} = 2.4425$.

2 Ajustement affine : droite des moindres carrés pondérés

La droite des moindres carrés pondérés minimise la somme pondérée des différences entre y_i et $ax_i + b$ au carré ; c'est-à-dire que les coefficients a et b de cette droite minimisent l'expression : $\sum_{i=1}^{n} p_i(y_i - (ax_i + b))^2)$. L'équation de cette droite est explicite et est donnée au théorème suivant.

Théorème 1 (Droite des moindres carrés pondérés)**.** *Soit $(X, Y) = \{((x_1, y_1), p_1), \ldots, ((x_n, y_n), p_n)\}$ une série statistique double pondérée avec $v(X) > 0$ (c'est-à-dire que $n \geq 2$ et que les réels $x_1, \ldots, x_n$ ne sont pas tous égaux). La fonction*

$$\forall a \in \mathbb{R}\ \forall b \in \mathbb{R} \quad \phi(a,b) = \sum_{i=1}^{n} p_i(y_i - (ax_i + b))^2$$

atteint sont minimum en un unique point (a^*, b^*) donné par

$$a^* = \frac{cov(X,Y)}{v(X)} \text{ et } b^* = \overline{Y} - a^*\overline{X}.$$

On appelle droite des moindres carrés pondérés la droite d'équation $y = a^*x + b^*$.

La démonstration suivante ne nécessite aucune connaissance en calcul différentiel.

Démonstration. Les égalités suivantes permettent de décomposer ϕ comme une somme de deux carrés et d'un

reste ne dépendant ni de a ni de b :

$$\begin{aligned}
\phi(a,b) &= \sum_{i=1}^{n} p_i(y_i - ax_i - b)^2 \\
&= b^2 \sum_{i=1}^{n} p_i - 2b \sum_{i=1}^{n} p_i(y_i - ax_i) + \sum_{i=1}^{n} p_i(y_i - ax_i)^2 \\
&= b^2 - 2b(\overline{Y} - a\overline{X}) + \sum_{i=1}^{n} p_i(y_i - ax_i)^2 \\
&= b^2 - 2b(\overline{Y} - a\overline{X}) + \sum_{i=1}^{n} p_i y_i^2 - 2a \sum_{i=1}^{n} x_i y_i + a^2 \sum_{i=1}^{n} x_i^2 \\
&= b^2 - 2b(\overline{Y} - a\overline{X}) + \overline{Y^2} - 2a\overline{XY} + a^2\overline{X^2} \\
&= (b - (\overline{Y} - a\overline{X}))^2 - (\overline{Y} - a\overline{X})^2 + \overline{Y^2} - 2a\overline{XY} + a^2\overline{X^2} \\
&= (b - (\overline{Y} - a\overline{X}))^2 + v(Y) + a^2 v(X) - 2a\,cov(X,Y) \\
&= (b - (\overline{Y} - a\overline{X}))^2 + v(X)\left(a - \frac{cov(X,Y)}{v(X)}\right)^2 - \frac{cov(X,Y)^2}{v(X)} + v(Y)
\end{aligned}$$

On remarque que l'expression précédente est minimale lorsque les deux carrés s'annulent c'est-à-dire lorsque $a = \frac{cov(X,Y)}{v(X)} = a^*$ et $b = \overline{Y} - a^*\overline{X} = b^*$. □

Quelques remarques sur ce théorème sont données ci-dessous :

— La droite des moindres carrés pondérés passe par le point moyen de la série statistique double $(\overline{X}, \overline{Y})$.

— Lorsque $v(Y) > 0$ (c'est-à-dire que $n \geq 2$ et que les réels $y_1, \ldots, y_n$ ne sont pas tous égaux) le minimum de la fonction ϕ vaut

$$\phi(a^*, b^*) = v(Y)\left(1 - \frac{cov(X,Y)^2}{v(X)v(Y)}\right).$$

Ce minimum est appelé somme des carrés résiduels pondérés.

Exemple 2. *On considère la série statistique double pondérée (X, Y) donnée à l'exemple 1. Comme $a^* = cov(X,Y)/v(X) = 0.2858$ et $b^* = \overline{Y} - a^*\overline{X} = 0.2068$, la droite des moindres carrés pondérés, représentée ci-dessous, a pour équation $y = 0.2858x + 0.2068$. Par exemple, pour une note de 11 au rapport écrit on prédit, via la droite des moindres carrés pondérés, une note moyenne à la soutenance orale de $0.2858 \times 11 + 0.2068 = 3.3506$.*

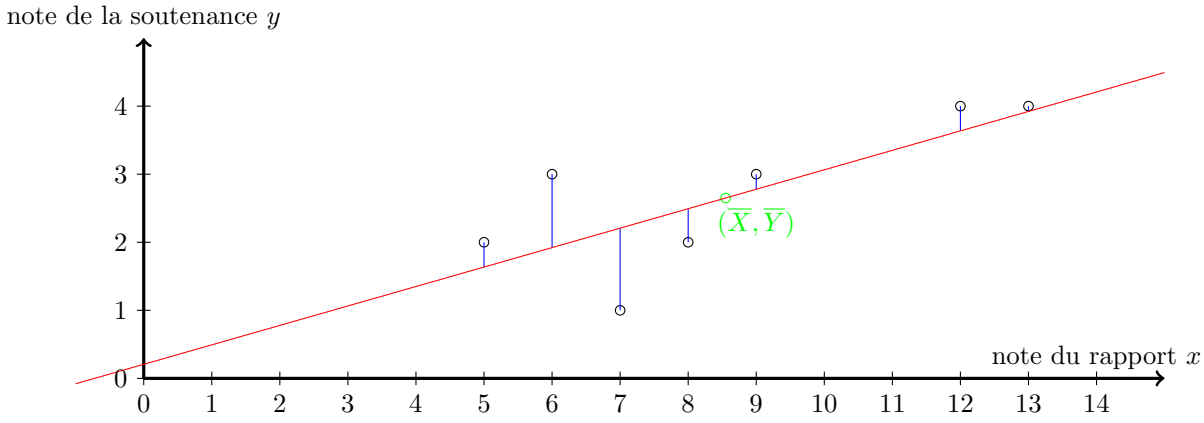

FIGURE 1 – La droite rouge est la droite des moindres carrés pondérés de la série statistique double (X, Y) décrit à l'exemple 1. Géométriquement, la somme des carrés résiduels pondérés $\phi(a^*, b^*) = v(Y)\left(1 - cov(X,Y)^2/(v(X)v(Y))\right)$ est égale à la somme pondérée des carrés des longueurs en bleues.

3 Coefficient de corrélation linéaire et décomposition de la variance

Le coefficient de corrélation linéaire, que l'on définit juste après, est lié à la somme des carrés résiduels pondérés.

Définition 5 (Coefficient de corrélation linéaire)**.** *Soit* $(X,Y) = (((x_1,y_1),p_1),\ldots,((x_n,y_n),p_n))$ *une série statistique double pondérée avec* $v(X) > 0$ *et* $v(Y) > 0$*. On appelle coefficient de corrélation linéaire le réel suivant :*

$$r(X,Y) = \frac{cov(X,Y)}{\sqrt{v(X)}\sqrt{v(Y)}}.$$

Proposition 2. *Soit* $(X,Y) = (((x_1,y_1),p_1),\ldots,((x_n,y_n),p_n))$ *une série statistique double pondérée avec* $v(X) > 0$ *et* $v(Y) > 0$*. Le coefficient de corrélation linéaire vérifie* $|r(X,Y)| \leq 1$*. De plus,* $|r(X,Y)| = 1$ *si et seulement si les points* $(x_1,y_1),\ldots,(x_n,y_n)$ *sont alignés.*

Démonstration. Clairement la somme des carrés résiduels pondérés est positive, c'est-à-dire que $\phi(a^*,b^*) \geq 0$. Ainsi,

$$v(Y)\left(1 - \frac{cov(X,Y)^2}{v(X)v(Y)}\right) \geq 0 \Rightarrow -1 \leq \frac{cov(X,Y)}{\sqrt{v(X)v(Y)}} \leq 1.$$

Enfin, $|r(X,Y)| = 1$ si et seulement si $\phi(a^*,b^*) = 0$ si et seulement si $p_i(y_i - (a^*x_i + b^*))^2 = 0$ pour tout $i \in \{1,\ldots,n\}$ si et seulement si les points $(x_1,y_1),\ldots,(x_n,y_n)$ sont alignés sur la droite d'équation $y = a^*x + b^*$. □

Proposition 3 (Décomposition de la variance)**.** *Soit* $(X,Y) = \{((x_1,y_1),p_1),\ldots,((x_n,y_n),p_n)\}$ *une série statistique double pondérée,* a^* *et* b^* *le coefficient directeur et l'ordonnée à l'origine de la droite des moindres carrés pondérés. Pour* $i \in \{1,\ldots,n\}$ *on pose* $\widehat{y}_i = a^*x_i + b^*$ *la* $i^{\text{ème}}$ *valeur ajustée ,* $\widehat{e}_i = y_i - \widehat{y}_i$ *le* $i^{\text{ème}}$ *résidu et on définit* $\widehat{Y}$ *la série pondérée des valeurs ajustées et* $\widehat{E}$ *la série pondérés des résidus :*

$$\widehat{Y} = \{(\widehat{y}_1,p_1),\ldots,(\widehat{y}_n,p_n)\} \text{ et } \widehat{E} = \{(\widehat{e}_1,p_1),\ldots,(\widehat{e}_n,p_n)\}.$$

Les propriétés suivantes sont satisfaites :

1. *La moyenne pondérée de la série statistique des valeurs ajustées est égale à la moyenne pondérée de la série* Y *:* $\overline{\widehat{Y}} = \overline{Y}$*.*
2. *La moyenne pondérée de la série statistique des résidus est nulle :* $\overline{\widehat{E}} = 0$*.*
3. *La variance de* Y *est la somme de la variance des valeurs ajustées et de la variance des résidus :* $v(Y) = v(\widehat{Y}) + v(\widehat{E})$*.*
4. *Le coefficient de corrélation linéaire vérifie* $r(X,Y)^2 = v(\widehat{Y})/v(Y)$*.*

Démonstration. 1) La moyenne pondérée de la série statistique des valeurs ajustées est égale à la moyenne pondérée de la série Y, en effet :

$$\begin{aligned}
\overline{\widehat{Y}} &= \sum_{i=1}^{n} p_i(a^*x_i + b^*) \\
&= \sum_{i=1}^{n} p_i(a^*x_i + \overline{Y} - a^*\overline{X}) \\
&= \underbrace{\sum_{i=1}^{n} p_i\overline{Y}}_{=\overline{Y}} + a^* \underbrace{\sum_{i=1}^{n} p_i(x_i - \overline{X})}_{=0} = \overline{Y}
\end{aligned}$$

2) La moyenne pondérée de la série statistique des résidus est nulle, en effet :

$$\begin{aligned}
\overline{\widehat{E}} &= \sum_{i=1}^{n} p_i(y_i - a^*x_i - b^*) \\
&= \sum_{i=1}^{n} p_i(y_i - a^*x_i - \overline{Y} + a^*\overline{X}) \\
&= \underbrace{\sum_{i=1}^{n} p_i(y_i - \overline{Y})}_{=0} + a^* \underbrace{\sum_{i=1}^{n} p_i(\overline{X} - x_i)}_{=0} = 0
\end{aligned}$$

3) Montrons que $v(Y) = v(\widehat{Y}) + v(\widehat{E})$:

$$\begin{aligned} v(Y) &= \sum_{i=1}^{n} p_i (y_i - \overline{Y})^2, \\ &= \sum_{i=1}^{n} p_i (y_i - \widehat{y}_i + \widehat{y}_i - \overline{Y})^2, \\ &= \underbrace{\sum_{i=1}^{n} p_i (y_i - \widehat{y}_i)^2}_{=\overline{\widehat{E}^2}} + \underbrace{\sum_{i=1}^{n} p_i (\widehat{y}_i - \overline{Y})^2}_{=v(\widehat{Y})} + 2 \sum_{i=1}^{n} p_i (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{Y}), \\ &= v(\widehat{E}) + v(\widehat{Y}) + 2 \sum_{i=1}^{n} p_i (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{Y}). \end{aligned}$$

Ainsi, il suffit de montrer que $\sum_{i=1}^{n} p_i (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{Y}) = 0$:

$$\begin{aligned} \sum_{i=1}^{n} p_i (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{Y}) &= \sum_{i=1}^{n} p_i (y_i - a^* x_i - b^*)(a^* x_i + b^* - \overline{Y}), \\ &= \sum_{i=1}^{n} p_i (y_i - a^* x_i - \overline{Y} + a^* \overline{X})(a^* x_i + \overline{Y} - a^* \overline{X} - \overline{Y}), \\ &= \sum_{i=1}^{n} p_i (y_i - \overline{Y})(a^* x_i - a^* \overline{X}) + \sum_{i=1}^{n} p_i (a^* \overline{X} - a^* x_i)(a^* x_i - a^* \overline{X}), \\ &= a^* cov(X,Y) - (a^*)^2 v(X) = \frac{cov(X,Y)^2}{v(X)} - \frac{cov(X,Y)^2}{v(X)} = 0 \end{aligned}$$

4) La somme des carrés résiduels pondérés est égale à $\phi(a^*, b^*) = \overline{\widehat{E}^2} = v(\widehat{E})$. Ainsi,

$$\frac{v(\widehat{Y})}{v(Y)} = \frac{v(Y) - v(\widehat{E})}{v(Y)} = 1 - \frac{\phi(a^*, b^*)}{v(Y)} = 1 - \frac{v(Y)(1 - r(X,Y)^2)}{v(Y)} = r(X,Y)^2.$$

□

Une interprétation géométrique de la décomposition $v(Y) = v(\widehat{Y}) + v(\widehat{E})$: est donnée à la figure 2.

Exemple 3. *On considère la série statistique double pondérée (X, Y) donnée à l'exemple 1.*

— *La série pondérée des valeurs ajustées $\widehat{Y}$ et la série pondérés des résidus $\widehat{E}$ sont données ci-dessous :*

$$\widehat{Y} = \{(3.6359, 0.1), (2.7786, 0.1), (1.9213, 0.15), (1.6356, 0.15), (3.9216, 0.2), (2.4928, 0.1), (2.2071, 0.2)\}.$$

$$\widehat{E} = \{(0.3641, 0.1), (0.2214, 0.1), (1.0787, 0.15), (0.3644, 0.15), (0.0784, 0.2), (-0.4928, 0.1), (-1.2071, 0.2)\}.$$

— *Les variances pondérées des séries statistiques $Y, \widehat{Y}$ et $\widehat{E}$ valent $v(Y) = 1.2275$, $v(\widehat{Y}) = 0.6980$ et $v(\widehat{E}) = 0.5295$. On vérifie que $v(Y) = v(\widehat{Y}) + v(\widehat{E})$.*

— *Le coefficient de corrélation linéaire vaut $r(X,Y) = cov(X,Y)/\sqrt{v(X)v(Y)} = 0.7541$. On vérifie que $r(X,Y)^2 = v(\widehat{Y})/v(Y)$.*

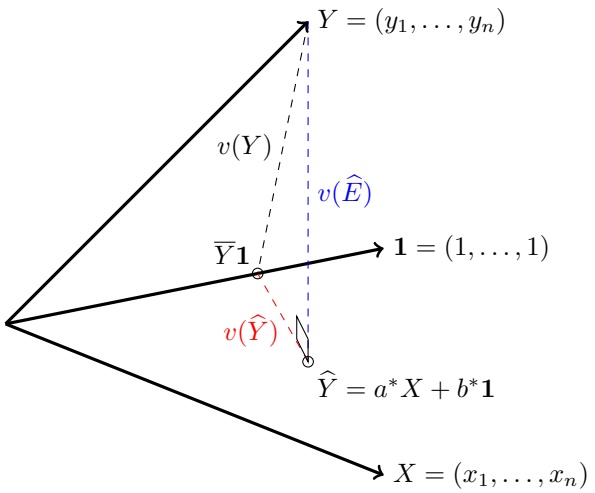

FIGURE 2 – Cette figure illustre géométriquement la décomposition de la variance $v(Y) = v(\widehat{Y}) + v(\widehat{E})$. On considère le produit scalaire $\langle u, v\rangle = \sum_{i=1}^n p_i u_i v_i$ et la norme euclidienne $\|u\| = \sqrt{\langle u, u\rangle}$. La projection orthogonale de $Y = (y_1, \ldots, y_n)$ sur l'espace vectoriel engendré par $\mathbf{1} = (1, \ldots, 1)$ et $X = (x_1, \ldots, x_n)$ est égal à $\widehat{Y} = a^*Y + b^*\mathbf{1}$. Ainsi, d'après le théorème de Pythagore on a $\underbrace{\|Y - \overline{Y}\mathbf{1}\|^2}_{v(Y)} = \underbrace{\|Y - \widehat{Y}\|^2}_{v(\widehat{E})} + \underbrace{\|\widehat{Y} - \overline{Y}\mathbf{1}\|^2}_{v(\widehat{Y})}$.