6. Moindres carrés et statistiques

Exercice 1 (Un peu de statistiques)
On se donne à nouveau n points du plan de coordonnées (x_i, y_i), $1 \leq i \leq n$, et on cherche l'équation $y = ax + b$ de la droite qui minimise la somme des carrés des distances verticales des points aux droites. En statistiques, cette droite est appelée la *droite de régression linéaire* des points (x_i, y_i).

(a) Montrer qu'on peut réécrire le système de l'exercice 4 de la feuille 5 sous la forme :

$$\left\{ \begin{array}{rcrcl} a\overline{x} & + & b & = & \overline{y} \\ a\frac{\sum x_i^2}{n} & + & b\overline{x} & = & \frac{\sum x_i y_i}{n} \end{array} \right. ,$$

où :
- $\overline{x} = \dfrac{x_1 + \cdots + x_n}{n}$ est la *moyenne observée* des abscisses ;
- $\overline{y} = \dfrac{y_1 + \cdots + y_n}{n}$ est la *moyenne observée* des ordonnées ;
- $\sum x_i^2 = \displaystyle\sum_{i=1}^{n} x_i^2 = x_1^2 + \cdots + x_n^2$ et $\sum x_i y_i = \displaystyle\sum_{i=1}^{n} x_i y_i = x_1 y_1 + \cdots + x_n y_n$.

(b) La *variance observée* des abscisses est la moyenne des carrés des écarts à la moyenne $\overline{x}$:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} .$$

Montrer la formule de König :

$$Var(x) = \frac{\sum x_i^2}{n} - \overline{x}^2 .$$

L'*écart-type observé* est : $\sigma(x) = \sqrt{Var(x)}$.

(c) La *covariance observée* est la moyenne des produits des écarts aux moyennes $\overline{x}$ et $\overline{y}$:

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})}{n} .$$

Montrer que $Cov(x, y) = \dfrac{\sum x_i y_i}{n} - \overline{x}\,\overline{y}$.

(d) En déduire que l'équation générale de la droite de régression linéaire est :

$$y = \frac{Cov(x, y)}{Var(x)}x + \overline{y} - \frac{Cov(x, y)}{Var(x)}\overline{x} .$$

Noter que la droite de régression linéaire passe par le *point moyen* $(\overline{x}, \overline{y})$.

(e) Retrouver les résultats de l'exercice 5 de la feuille 5 à l'aide de cette formule.

Exercice 2 (Corrélation)
Etant donnés n points (x_i, y_i), $1 \leq i \leq n$, le *cœfficient de corrélation linéaire* r est défini par :

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}} = \frac{Cov(x,y)}{\sigma(x)\sigma(y)} \ .$$

(a) Montrer que l'équation de la droite de régression linéaire peut s'écrire : $\dfrac{y-\overline{y}}{\sigma(y)} = r\dfrac{x-\overline{x}}{\sigma(x)}$.

(b) On note S_{min} la somme des carrés des distances verticales des points (x_i, y_i) à la droite de régression linéaire : c'est le minimum de la fonction $S(a,b)$ de la feuille 5. Etablir la formule :

$$S_{min} = n\,Var(y)(1-r^2) \ .$$

Ceci montre d'une part que $r^2 \leq 1$, donc $-1 \leq r \leq 1$, et d'autre part que plus r est proche de 1 ou -1, plus S_{min} est petit, et plus les points donnés sont proches de la droite de régression linéaire. On dira alors que la relation entre x et y est *proche d'une relation linéaire*. Inversement, si r est proche de 0, la relation n'est pas proche d'une relation linéaire.

(c) Calculer le cœfficient de corrélation linéaire pour chacun des ensembles de points des exercices 2 et 5 de la feuille 5. Lesquels sont proches d'une relation linéaire ?

Exercice 3 (Parabole)
(a) Quelle courbe les points (0 ; 3), (1 ; 1), (2 ; 0), (4 ; 1) et (6 ; 4) semblent-ils dessiner ?

(b) Montrer que la distance verticale d'un point (x_1, y_1) à la *parabole* d'équation $y = ax^2 + bx + c$ est $|y_1 - ax_1^2 - bx_1 - c|$. En déduire que la fonction donnant la somme des carrés des distances verticales de n points (x_i, y_i), $1 \leq i \leq n$, à la parabole est :

$$P(a,b,c) = (y_1 - ax_1^2 - bx_1 - c)^2 + \cdots + (y_n - ax_n^2 - bx_n - c)^2 \ .$$

(c) Calculer les trois dérivées partielles : $\frac{\partial P}{\partial a}(a,b,c)$, $\frac{\partial P}{\partial b}(a,b,c)$, $\frac{\partial P}{\partial c}(a,b,c)$.

(d) Montrer que les trois dérivées partielles s'annulent en (a,b,c) si et seulement si :

$$\left\{ \begin{array}{rcrcrcl} a\sum x_i^2 & + & b\sum x_i & + & cn & = & \sum y_i \\ a\sum x_i^3 & + & b\sum x_i^2 & + & c\sum x_i & = & \sum x_i y_i \\ a\sum x_i^4 & + & b\sum x_i^3 & + & c\sum x_i^2 & = & \sum x_i^2 y_i \end{array} \right.$$

(e) Ecrire le système correspondant aux cinq points donnés dans la question 1 et le résoudre. La solution (a,b,c) fournit l'équation de la parabole <u>la plus proche des cinq points au sens des moindres carrés</u>, que l'on tracera.

Exercice 4 (Exponentielle)
Supposons savoir *a priori* (par exemple parce que la théorie le prévoit) que la loi régissant la dépendance d'un paramètre y par rapport à un paramètre x est de la forme :

$$y = k\mathrm{e}^{ax} \ , \quad \text{avec } k > 0 \text{ et } a \in \mathbb{R} \ .$$

(a) Trouver deux nouvelles variables X et Y dépendant respectivement de x et de y, qui suivent une loi linéaire, c'est-à-dire correspondant à l'équation d'une droite : $Y = AX + B$.

(b) Des mesures donnent les couples : (−1 ; 1,2), (0 ; 2), (1 ; 3,3), (2 ; 5,4), (3 ; 9) (la première coordonnée correspond au paramètre x, la seconde à y). Déterminer la droite de régression linéaire pour les points correspondant aux nouvelles variables X et Y. L'approximation est-elle satisfaisante ? En déduire les valeurs de k et a.

Exercice 5 (Comparaison)
On donne les points $(-1\,;2)$, $(0\,;0)$, $(0\,;1)$ et $(1\,;2)$. Les tracer sur une figure.

(a) Déterminer l'équation de la droite la plus proche des quatre points au sens des moindres carrés, puis la somme S_{min} des carrés des distances verticales des points à cette droite.

(b) Faire de même avec la parabole la plus proche.

(c) Laquelle des deux courbes approxime le mieux les quatre points ?