Feuille 7 : Régression

Exercice 1 (Premier exemple de régression)
On considère la série statistique à deux variables suivante :

x_i	1	5	6	1	4
y_i	-1	4	6	0	3

1. *Calculer les moyennes empiriques de x et de y.*
2. *Calculer les variances empiriques de x et de y, la covariance empirique de x et y.*
3. *Donner une équation de la droite de régression de y sur x.*
4. *Tracer le nuage de points et la droite de régression.*

Exercice 2 (Température d'incubation)
On se propose d'étudier l'influence de la température sur la durée d'incubation des oeufs de grenouilles. On choisit 6 échantillons de 200 oeufs chacun. Le nombre x d'éclosions au 22-ème jour est le suivant

température t_i d'incubation en degré Celsius	6	6,4	6,8	7,2	7,6	8
nombre x_i d'éclosions à la température t_i	131	144	157	170	190	189

1. *Dessiner le nuage des données et equisser une droite D approchant ce nuage.*
2. *Calculer le coefficient de corrélation observé et écrire l'équation de la droite de régression de x en t. Étudier la qualité de l'ajustement.*
3. *Calculer le nombre d'éclosions prédit pour un échantillon de 200 oeufs au 22-ème jour pour une température de 7,5 degrés.*

Exercice 3 (Coïncidence)
Soient $(x_i, y_i)_{1\leq i\leq n}$ une série statistique à deux variables. À quelle condition les deux droites de régression, de x sur y et de y sur x, coïncident-elles ?

Exercice 4 (Régression logarithmique)
On a mesuré les variables x et y sur 10 individus et obtenu les résultats suivants :

individu i	1	2	3	4	5	6	7	8	9	10
x_i	13	16	23	29	35	43	49	55	58	63
y_i	16,5	17,9	20,3	22	23,5	25,3	26,5	27,6	28,2	29,1

1. *Calculer la droite de régression linéaire de y en x.*
2. *On pose $z_i = \log x_i$ (logarithme en base 10). Chercher les valeurs de α, β, γ qui minimisent la somme $\sum_{i=1}^{10}(y_i - \alpha - \beta x_i - \gamma z_i)^2$.*
3. *Représenter sur le même graphique le nuage de points (x_i, y_i), la droite de régression de la question 1), et la courbe de la fonction $y = \alpha + \beta x + \gamma \log(x)$ avec les coefficients trouvés dans la question 2).*

Exercice 5 (Données brutes)
On recueille des données $(x_i, y_i)_{1\leq i\leq 20}$ telles que :

$$\frac{1}{20}\sum_{i=1}^{20} x_i = 34.9, \quad \frac{1}{20}\sum_{i=1}^{20} x_i^2 = 1246.3, \quad \frac{1}{20}\sum_{i=1}^{20} y_i = 18.34, \quad \frac{1}{20}\sum_{i=1}^{20} y_i^2 = 339.2, \quad \frac{1}{20}\sum_{i=1}^{20} y_i x_i = 646.32.$$

1. *Calculer l'équation de la droite de régression linéaire de y en x.*
2. *Calculer le coefficient de corrélation observé et écrire l'équation de la droite de régression de x en y.*
3. *Calculer le coefficient de détermination.*

Exercice 6 (Régression et région de confiance)
On considère le modèle de régression linéaire simple suivant, où les ε_i sont i.i.d del oi $\mathcal{N}(0, \sigma^2)$.

$$\forall i \in \{1, ..., n\} \quad y_i = \beta_1 + \beta_2 x_i + \epsilon_i.$$

Soit un échantillon $(x_i, y_i)_{1\leq i\leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0, \ \sum_{i=1}^{100} x_i^2 = 400, \ \sum_{i=1}^{100} x_i y_i = 100, \ \sum_{i=1}^{100} y_i = 100, \ , \ \hat{\sigma}^2 = 1.$$

1. *Exprimer les intervalles de confiance à 95% pour β_1 et β_2.*
2. *Donner l'équation de la région de confiance à 95% de (β_1, β_2).*

Rappel : l'ensemble des points (x, y) tels que $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'intérieur d'une ellipse centrée en (x_0, y_0), dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, 0)$ et $(0, y_0 \pm b)$.

Exercice 7 (Moindres carrés)
Nous considérons le modèle statistique suivante :

$$\forall i \in \{1, ..., n\} \quad y_i = \beta x_i + \epsilon_i.$$

On suppose que $\mathbb{E}(\epsilon_i) = 0$, pour tout $i \in \{1, ..., n\}$ et $Cov(\epsilon_i, \epsilon_j) = \mathbb{1}_{\{i=j\}}\, \sigma^2$, pour $(i, j) \in \{1, ..., n\}^2$.

1. *En revenant à la définition des moindres carrés, montrer que l'estimateur des moindres carrés de β vaut*
$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$
2. *Montrer que la droite passant par l'origine et le centre de gravité du nuage de points est $y = \beta^* x$, avec*
$$\beta^* = \frac{\sum y_i}{\sum x_i}.$$
3. *Montrer que $\hat{\beta}$ et β^* sont tous deux des estimateurs sans biais de β.*
4. *En utilisant l'inégalité de Cauchy–Schwarz, montrer que $Var(\beta^*) > Var(\hat{\beta})$ sauf dans le cas où tous les x_i sont égaux. Ce résultat était-il prévisible ?*