Série 13

Exercice 1. Pour le modèle statistique de régression linéaire $Y = a + b\,X + \eta$, quand on observe les points $(x_i,\, y_i)$ $i = 1, 2, \ldots, n$, les réalisations des estimateurs de a et b, $\hat{a}$ et $\hat{b}$, peuvent s'écrire :

$$\hat{b} = \sum_{j=1}^{n} w_j\, y_j \quad \text{et} \quad \hat{a} = \sum_{j=1}^{n} v_j\, y_j \quad \text{avec} \quad w_j = \frac{x_j - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad v_j = \frac{1}{n} - \bar{x}\, w_j\,.$$

Montrer les propriétés suivantes :

a) $\displaystyle\sum_{j=1}^{n} w_j = 0$ et $\displaystyle\sum_{j=1}^{n} v_j = 1$.

b) $\displaystyle\sum w_j^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ et $\displaystyle\sum v_j^2 = \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

c) $\sum w_j\, x_j = 1$ et $\sum v_j\, x_j = 0$.

Exercice 2. On cherche à calculer la relation entre la moyenne $\bar{y}$ d'un examen et celle $\bar{x}$ d'un test bonus. On sait que leur coefficient de corrélation empirique vaut $\hat{r} = 0.6$. En utilisant les données suivantes :

$$\bar{y} \;=\; 55, \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 20, \quad \bar{x} = 70, \quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 10,$$

trouver l'équation de la droite de régression $x \to \hat{a} + \hat{b}\bar{x}$ de la moyenne à l'examen en fonction de celle au test bonus.

Exercice 3. Au problème de la régression simple classique peut s'ajouter l'obligation de devoir passer par un point (x_0, y_0) spécifique. Ce type de régression est dite régression forcée. On considère un ensemble de points (x_i, y_i) pour $i = 1, \ldots, n$, et le modèle $y_i = \beta(x_i - x_0) + y_0 + \eta_i$ où les η_i sont des Gaussiennes indépendantes de moyenne 0 et de variance 1.

a) Trouver la valeur $\hat{\beta}$ de β qui minimise l'erreur de prédiction $\sum_i \{y_i - y_0 - \beta(x_i - x_0)\}^2$. On va utiliser cette valeur comme estimation (réalisation de l'estimateur) de β, la pente de la droite de régression forcée par un point (x_0, y_0).

b) Appliquer le résultat de la question a) au cas $(x_0, y_0) = (\bar{x}, \bar{y})$ et comparer avec l'estimateur de la pente de la droite de régression simple classique.

c) Déterminer le modèle de régression forcée par le point de coordonnées $(3, 6)$ pour les données ci-dessous :

x_i	1	2	3	4	5	6	7
y_i	2.5	5.0	3.5	7	6.5	8.25	8

Exercice 4. Un échantillon aléatoire de 10 hommes âgés de 30 ans a été choisi. On a relevé les données de leur salaire annuel (Y) et le nombre d'années d'école qu'ils ont suivies (X). On a trouvé les valeurs suivantes : $\Sigma\, x = 151 \quad \Sigma\, y = 96.3 \quad \Sigma\, xy = 1526.30 \quad \Sigma\, x^2 = 2357 \quad \Sigma\, y^2 = 1036.13$.

En supposant que la moyenne de la loi des salaires est une fonction linéaire du nombre d'années d'école :

a) Déterminer les estimations des paramètres a et b du modèle $Y = a + b\,X + \eta$, ainsi que l'estimation de σ^2 (variance de η).

b) Tester l'hypothèse $H_0 : b = 0$ contre l'alternative $H_1 : b \neq 0$ au seuil de signification $\alpha = 0.01$.

Exercice 5. Transformations des données. Pour certains jeux de données, il peut être judicieux de calculer une transformation des données avant d'effectuer une régression linéaire.

Les données suivantes montrent les valeurs expérimentales de la pression P d'une masse donnée de gaz pour différentes valeurs du volume V.

Volume V (cm^3)	54.3	61.8	72.4	88.7	118.6	194.0
Pression P (kg/cm^2)	61.2	49.5	37.6	28.4	19.2	10.1

D'après les principes de la thermodynamique, on a la relation $P\,V^\gamma = C$, où γ et C sont des constantes dépendant des conditions d'expérience.

a) Transformer $P\,V^\gamma = C$ en modèle linéaire avec $Y = \log(P)$. Quels sont les paramètres dont on cherche la valeur ?

b) En utilisant a), trouver les estimateurs de γ et $\log(C)$.

c) Estimer P quand $V = 100$ cm^3.

d) Calculer l'intervalle de confiance à 95% pour la pente de la droite du modèle de régression trouvé en a).