Travaux Dirigés

Exercice 1. Dans le modèle de régression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon = \beta_0 + \mathbf{X}^T \beta + \epsilon,$$

(a) Déterminer les éléments de la matrice de design $\mathbf{X}$.

(b) Calculer le prédicteur linéaire $\mathbf{X}^T \beta$.

Exercice 2.

(a) Utiliser la méthode des mondres carrées pour montrer que l'estimateure du vecteur $\widehat{\beta}$ est donné par

$$\widehat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}\mathbf{y}$$

(b) Calculer σ^2 la variance de variable aléatoire Gaussienne ϵ.

Exercice 3.

(a) Soit $X \rightsquigarrow \mathcal{B}(p)$ "la loi de Bernoulli". Déterminer la loi de X.

(b) Calculer la moyenne et la variance de X.

Exercice 4.

(a) Soit $X \rightsquigarrow \mathcal{B}(n, p)$ "la loi binomiale (ou multinomiale),". Déterminer la loi de X.

(b) Calculer la moyenne et la variance de X.

Exercice 5.

(a) Soit $X \rightsquigarrow \mathcal{P}(\lambda)$ "la loi de Poisson". Déterminer la loi de X.

(b) Calculer la moyenne et la variance de X.

Exercice 6.

- Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O_3 dans l'air (en microgrammes par millilitre). En particulier, on cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone de la journée par la température T_{12} à midi. Les données sont :

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3	115.4	76.8,	113.8	81.6,	115.4	125	83.6	75.2	136.8	102.8

∘ On cherche à expliquer la variable $Y = O_3$ à partir de la variable $X_1 = T_{12}$.

• Ecrire le modèle de régression linéaire correspondant et déterminer la matrice de design.

• Représentez, à l'aide d'une graphe adapté, la relation entre la variable X_1 et Y.

• Donner une estimation des différents paramètres du modèle.

• Calculer le coefficient de détermination $r^2 = \rho_{2(X_1,Y)}$, avec

$$\rho_{(X_1,Y)} = \frac{cov\,(Y, X_1)}{\sqrt{Var\,\{Y\}}\sqrt{Var\,\{X_1\}}}$$

• Importer les données sous R et étudier l'effet de la variable X_1 sur la variable Y.

Exercice 7. Les données utilisées sont le taux de décès par attaque cardiaque chez les hommes de 55 à 59 ans dans différents pays.

$Y = (124, 49, 181, 4, 22, 152, 75, 54, 43, 41, 17, 22, 16, 10, 63, 170, 125, 15, 221, 171, 97, 254)$

$X_1 = (33, 31, 38, 17, 20, 39, 30, 29, 35, 31, 23, 21, 8, 23, 37, 40, 38, 25, 39, 33, 38, 39)$

$X_2 = (8, 6, 8, 2, 4, 6, 7, 7, 6, 5, 4, 3, 3, 3, 6, 8, 6, 4, 7, 7, 6, 8)$

$X_3 = (81, 55, 80, 24, 78, 52, 52, 45, 50, 69, 66, 45, 24, 43, 38, 72, 41, 38, 52, 52, 66, 89)$

Les variables sont les suivantes :

• Y : $100 * log$(nombre de décès par crise cardiaque pour 100000 hommes de 55 à 59 ans).

• X_1 : nombre de téléphones pour 1000 habitants.

• X_2 : calories grasses en pourcentage du total de calories.

• X_3 : calories protéines animales en pourcentage du total de calories.

a) On cherche tout d'abord à expliquer la variable Y à partir de la variable X_1, X_2 et X_3.

• Ecrire le modèle de régression linéaire correspondant et déterminer la matrice de design.

• Représentez, à l'aide d'une graphe adapté, la relation entre les variables X_1 et Y, X_2 et Y, X_3 et Y.

• Donner une estimation des différents paramètres du modèle.

• Importer les données sous R et étudier l'effet des variables X_1, X_2 et X_3 sur la variable Y.