Planche de TP 7
Regression et modèle linéaire

Exercice 1 (Regression linéaire simple)

Soient X et ϵ deux variables aléatoires telles que $X \sim \mathcal{N}(5,7)$ et $\epsilon \sim \mathcal{N}(0,1)$. Générer 30 réalisations de la variable Y définies par $Y = 2 + 0.5 \times X + \epsilon$.

1. Tracer le nuage de points correspondant aux réalisations du couple (X, Y).
2. Rappeler la formule des estimateurs des coefficients de la droite de regression au sens des moindres carrés. Quelles valeurs prennent ces derniers sur l'échantillon précédant.
3. Tracer sur le même graphique la droite de regression.
4. Utiliser la commande `lm` et commenter les résultats obtenus.

Exercice 2 (Télévision et agressivité)

Un psychologue mène une étude auprès de 4 enfants et s'intéresse au lien entre le nombres d'heures consacrées à regarder la télévision chaque semaine et le taux d'agressivité. Les résultats obtenus sont les suivants :

Heures de télé	Score d'agressivité
5	8
20	12
8	6
2	4

1. Tracer le nuage de points correspondant.
2. Tracer la droite de regression.
3. Peut-on affirmer au vu des observations que le nombre d'heures passées devant la télé et le taux d'agressivité sont corrélées ?

Exercice 3 (Prédiction)

Cet exercice utilise le fichier `tdf.dat` en partie étudié dans la planche de TP 4. On s'intéresse cette fois-ci à l'étude de la vitesse moyenne sur l'épreuve en fonction de la distance totale du parcours.

1. Tracer la droite de regression correspondante.
2. Tester l'hypothèse selon laquelle il existe une corrélation linéaire entre la longueur du parcours et le vitesse moyenne.
3. Donner une prédiction de la vitesse moyenne sur l'épreuve 2008.
4. Tracer sur le même graphique les contours à 95% pour respectivement la valeur moyenne de $\mathbb{E}[Y/X = x]$ et l'intervalle de prédiction pour une valeur unique (on prendra soins d'utiliser des couleurs différentes pour chaque type de courbe).

Exercice 4 (Télévisions et espérance de vie)

Cet exercice utilise les fichiers television.txt et television.dat.

1. Créer deux graphiques décrivant la valeur de l'espérance de vie en fonction de respectivement le nombre de personnes par télé et le nombre de personnes par physiciens. Commentez le résultat obtenu. Quelle type de relation lie ces différentes variables ?
2. Superposer aux graphiques précédants les deux droites de regression.
3. Tester l'hypothèse selon laquelle la pente des deux droites et positive. Que peut-on conclure du résultat obtenu ?
4. Reprendre les questions 1-3 en étudiant cette fois-ci le logarithme du nombre d'habitants par télé en fonction de l'espérance de vie.

Exercice 5 (Regression multiples à deux composantes)

Soient $X1$, $X2$ et ϵ trois variables aléatoires telles que $X1 \sim \mathcal{N}(4,1)$, $X1 \sim \mathcal{N}(1,1)$ et $\epsilon \sim \mathcal{N}(0,0.5)$. Générer 30 réalisations de la variable Y définies par $Y = 2 \times X1 - X2 + \epsilon$.

1. Créer un data.frame *obs* contenant les différentes réalisations du triplet $(Y, X1, X2)$. Taper plot(obs) et commenter le graphique obtenu.
2. Rappeler la formule des estimateurs des coefficients de la droite de regression au sens des moindres carrés. En utilisant seulement les fonctionnalités matricielles de **R**, précisez les valeurs de ces derniers sur l'échantillon précédant.
3. Utiliser maintenant la commande lm et commenter les résultats obtenus.

Exercice 6 (Précipitations aux Mali)

Cet exercice utilise le fichier Mali.dat. Ce dernier contient le relevé de précipitation en divers points du Mali (spécifiés par leur lattitude et leut longitude). Préciser les coefficents de regression linéaire expliquant l'intensité des précipitation en fonctions des coordonnées géographiques.

Exercice 7 (Hauteur de neige et variables environnementales)

Cet exercice utilise le fichier neige.dat contenant le relevé de la hauteur de neige en différents endroits en fonction de la valeur de certaines variables environnementales (pente, altitude, orientation, etc,...).

1. Construire les graphiques de la hauteur de neige en fonction des différentes variables et commenter les résultats obtenus.
2. Tracer les droites de regression de la hauteur de neige en fonction de la pente, puis en fonction de l'orientation et de l'altitude. Pour chacune d'entre elles, tester l'hypothèse de nullité du coefficient de corrélation linéaire.
3. Effectuer une regression multiple de la hauteur de neige en fonction des différentes variables environnementales et conclure.