RÉGRESSIONS.

Exercice 1 Les données concernant le poids et la taille des athlètes médaillés sont contenus dans le fichier taille_poids_olympics.txt.

1. Télécharger ce fichier depuis UMTICE et charger les données dans R.
2. Faire la régression de la taille par le poids, puis du poids par la taille. Que remarquez-vous ?
3. Pour chacune des deux régressions, tracer sur un même graphique le nuage de points, la droite de régression.

Exercice 2 Les données concernant la vitesse des coureurs et leur taux de lactate dans le sang sont contenues dans le fichier lactate.txt.

1. Télécharger ce fichier depuis UMTICE et charger les données dans R.
2. Faire la régression du taux de lactate par rapport à la vitesse. Qu'en concluez-vous ?
3. Tracer sur un même graphique le nuage de points, la droite de régression et les intervalles de confiance et de prédiction. Est-ce que ce graphique confirme la conclusion de la question précédente ?

Exercice 3 Lors d'un TP de biochimie, les données suivantes ont été obtenues :

X	10	20	30	40	50	60
Y	0.058	0.117	0.204	0.26	0.292	0.348

1. Effectuer la régression linéaire de Y par X et tracer sur un même graphique le nuage de points et la droite de régression.
2. Techniquement on sait que si X=0, alors Y=0 (c'est la seule chose dont on soit certain). Faire la régression de Y par X en forçant le passage par l'origine. Tracer sur le graphique précédent la nouvelle droite de régression obtenue.
3. Pour les deux régressions précédentes, calculer la somme des erreurs au carré. Comparer.

Exercice 4 Après avoir extrait l'ADN de différentes cultures cellulaires, un technicien vérifie la qualité de ces extractions en dosant l'ADN contenu dans chaque tube par spectrophotométrie. Il mesure la densité optique à 260 nm et à 280 nm.

Partie 1 : statistiques descriptives.

1. Téléchargez le fichier « densité optique_1.txt » sur UMTICE et enregistrez-le dans H.
2. Chargez les données dans R. Vous appellerez le jeu de données DO1.
3. Par la méthode de votre choix, déterminez s'il existe des valeurs aberrantes dans chacune des deux variables. Indiquez la méthode utilisée et le nombre de valeurs aberrantes trouvées pour chaque variable.
4. Que feriez-vous de ces valeurs ? Expliquez pourquoi.

Le technicien pense avoir corrigé toutes ses erreurs.

1. Téléchargez le fichier « densité optique_2.txt » sur UMTICE et enregistrez-le dans H.
5. Chargez les données dans R. Vous appellerez le jeu de données DO2.
6. Recherchez d'éventuelles données aberrantes dans ce nouveau fichier.
7. À partir du jeu de données DO2, et après avoir arrondi les données à un chiffre après la virgule en utilisant les commandes suivantes :

   ```
   DO2$DO260b<-round(DO2$DO260,digits=1)
   DO2$DO280b<-round(DO2$DO280,digits=1)
   ```

   calculez la moyenne, la médiane, la variance et l'écart-type pour chacune des deux variables.

Partie 2 : statistiques descriptives.

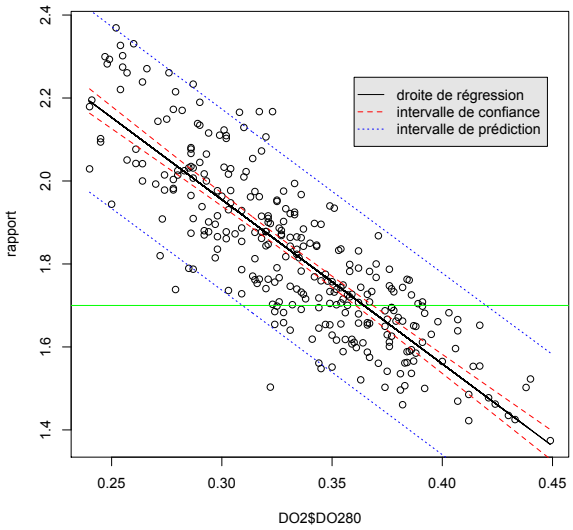
Dans la suite on n'utilisera plus que cette variable D02 contenant les valeurs non arrondies. Le rapport entre la densité optique à 260nm et la densité optique à 280 nm renseigne sur la contamination de l'extrait d'ADN par des protéines.

1. En utilisant la fonction `cov.test` de R, calculez le coefficient de corrélation entre DO260 et DO280. Existe-t-il une relation linéaire entre ces deux variables ?
2. Vérifiez que l'équation de la droite de régression linéaire de la DO à 260 nm en fonction de la DO à 280 nm est donnée par :

$$\mathrm{DO260} = 0,52\mathrm{DO280} + 0,43.$$

3. Calculez la DO à 260 nm prédite par la régression pour les valeurs suivantes de D0 à 280 nm : 0.250, 0.300, 0.350, 0.400, 0.450.
4. On estime qu'un extrait d'ADN est contaminé par des protéines pour un rapport DO260/DO280 <1,7. Calculez ce rapport pour les 5 couples de valeurs de la question précédente. Que pouvez-vous en conclure ?
5. Calculer le rapport r=DO260/DO280 pour les 300 tubes.

6. On utilise un modèle de régression linéaire pour déterminer si on peut conclure quant à la contamination des extraits d'ADN par des protéines d'après la valeur de DO à 280 nm sans calculer le rapport. Calculer les coefficients de la régression de r par rapport à DO280 et tracer sur un même graphique le nuage de points correspondant, les intervalles de prédiction et de confiance, et la droite constante r=1,7.
   Voici la représentation graphique des résultats obtenus :



   Le technicien peut-il déterminer si un extrait d'ADN est contaminé par des protéines d'après la valeur de DO à 280 nm sans calculer le rapport DO260/DO280 ? Vous pouvez distinguer différents cas.
