TP n°2
Régression et intervalles de confiance, le retour

1 Régression linéaire

Le but de cet exercice est d'étudier le lien entre la taille des pieds et le quotient intellectuel. Pour cela, une étude scientifique a été réalisée sur un échantillon représentatif de 100 hommes. On pourra trouver les résultats ici sous la forme d'une matrice

$$[P, Q] = \begin{bmatrix} P_1 & Q_1 \\ P_2 & Q_2 \\ \vdots & \vdots \\ P_{100} & Q_{100} \end{bmatrix},$$

où P_i et Q_i sont respectivement la pointure (française) et le QI de la $i^{\text{ème}}$ personne.

1. Télécharger la matrice $[P, Q]$, et tracer le nuage de points à l'aide de Scilab.
2. Quel est le coefficient de corrélation de P et Q ? Cela vaut-il la peine de continuer l'exercice ?
3. Calculer les coefficients de la droite de régression de Q sur P puis la tracer. Conclure.
4. Estimer l'intelligence de Jeison Orlando Rodriguez Hernandez, l'homme ayant les plus grands pieds du monde (40,1 centimètres, soit une pointure de 57).
5. Le QI de Stephen Hawking est estimé à 160. Comment estimer la taille de ses pieds ?

2 Un autre modèle de régression

Monsieur B. souhaite vendre sa maison de 280m^2 située à St Tropez, et aimerait savoir combien il pourrait en tirer. Pour cela, il fait une étude des prix de vente de maisons réalisées dans son voisinage. Cette étude est disponible ici, sous la forme d'une matrice

$$[S, P] = \begin{bmatrix} S_1 & P_1 \\ S_2 & P_2 \\ \vdots & \vdots \\ S_{200} & P_{200} \end{bmatrix},$$

où S_i et P_i sont respectivement la surface (en mètres carrés) et le prix de vente (en milliers d'euros) de la $i^{\text{ème}}$ maison.

1. Télécharger la matrice $[S, P]$ et afficher le nuage de points. Quel type de régression (i.e. dépendance de P en S) suggère le nuage de points ?

2. Dans la suite, on va effectuer une régression quadratique de P sur S, i.e. supposer que les variables dépendent l'une de l'autre suivant la relation

$$P = uS^2 + vS + w.$$

   a. Montrer que les coefficients u, v et w satisfont le système suivant

$$\left\{ \begin{array}{rcl} u\overline{S^2} + v\overline{S} + w & = & \overline{P} \\ u\overline{S^4} + v\overline{S^3} + w\overline{S^2} & = & \overline{PS^2} \\ u\overline{S^3} + v\overline{S^2} + w\overline{S} & = & \overline{PS} \end{array} \right. .$$

   b. Créer une fonction `regquad(s,p)` retournant les coefficients u, v et w définis à la question précédente. Afficher le nuage de points et la courbe de régression quadratique de P sur S.

   c. Estimer le prix de vente de la maison de Monsieur B.

3. Monsieur B. s'aperçoit qu'il a oublié de prendre en compte si les maisons vendues disposaient ou non d'une piscine (argument de poids sur la Côte d'Azur). Il raffine alors son étude précédente, désormais disponible ici, sous la forme

$$[S, P, J] = \begin{bmatrix} S_1 & P_1 & J_1 \\ S_2 & P_2 & J_2 \\ \vdots & \vdots & \vdots \\ S_{200} & P_{200} & J_{200} \end{bmatrix},$$

   où J_i vaut 1 si la $i^{\text{ème}}$ maison disposait d'une piscine, et 0 sinon. Dans la suite, on dénotera par S_0 (resp. P_0) la sous-matrice de S (resp. P) correspondant aux maisons ne disposant pas de piscine (et, de la même manière, S_1 et P_1).

   a. Tracer le nuage de points, en différenciant les maisons disposant d'une piscine des autres.

   b. A l'aide de la fonction `regquad` définie précédemment, tracer la régression quadratique de P sur S en noir, celle de P_0 sur S_0 en rouge, celle de P_1 sur S_1 en bleu.

   c. Malheureusement pour lui, Monsieur B. ne dispose pas (encore) de piscine dans sa maison. Estimer le prix maximum auquel Monsieur B. pourrait faire construire une piscine pour que cela soit rentable lors de la vente de la maison.

3 Estimation de la variance quand la moyenne est connue

On considère $X_1, \ldots, X_n$ un échantillon de taille n d'une loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 à estimer.

1. a. Calculer la loi de la variable aléatoire

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} X_k^2 = \frac{n}{\sigma^2} \overline{X_n^2}.$$

   b. Comment en déduire un intervalle de confiance exact à 95% du paramètre σ^2 ?

2. Utiliser cette méthode pour obtenir un intervalle de confiance au niveau 95% pour l'échantillon enregistré ici.