TP N°2 : Régression polynomiale

1 Régression linéaire

L'objectif de cet exercice est d'établir une relation linéaire entre différentes données issues d'un groupe de 100 élèves. Pour cela, on part d'une table de données de 100 élèves comprenant leur tailles, leur pointures et leur résultats au contrôle terminal. On pourra trouver les données ici sous la forme d'une matrice

$$(T, P, N) = \begin{pmatrix} T_0 & P_0 & N_0 \\ T_1 & P_1 & N_1 \\ \vdots & \vdots & \vdots \\ T_{99} & P_{99} & N_{99} \end{pmatrix}$$

où T_i, P_i et N_i sont respectivement la taille (en cm), la pointure (française) et la note au contrôle terminal (sur 20) de la $i^{ème}$ personne.

(i) Télécharger la matrice (T, P, N) à l'aide de la fonction *loadtxt* de *numpy* et afficher les nuages de points des paires (T, P), (T, N) et (P, N).

(ii) A votre avis, quelle paire est la meilleure candidate pour faire de la régression linéaire? Calculer les coefficients de corrélation des différentes paires grace à *np.corrcoef* de *numpy* et conclure.

On suppose que le couple (P, T) suit un modèle de régression linéaire, c'est à dire que pour tout i,

$$P_i = aT_i + b + \varepsilon_i$$

où les ε_i sont des variables aléatoires i.i.d centrées de variance σ^2 représentant l'erreur et a, b sont des paramètres réels à déterminer.

(iii) Utiliser la fonction *polyfit* de *numpy* afin de déterminer deux estimateurs $\hat{a}$ et $\hat{b}$ de a et b.

(iv) Calculer l'erreur quadratique moyenne du modèle,

$$EQM = \frac{1}{100} \sum_{i=0}^{99} (P_i - (\hat{a}T_i + \hat{b}))^2,$$

c'est à dire l'écart quadratique moyen entre les données réelles et celles estimées par le modèle linéaire.

On cherche à améliorer les résultats en exploitant le fait que les 20 premières données correspondent à des élèves masculins et les 80 dernières à des élèves féminins.

(v) Extraire la sous-matrice (P', T', N') correspondant aux hommes si vous êtes de sexe masculin et aux femmes si vous êtes de sexe féminin. Estimer les nouveaux paramètres a', b' et calculer la nouvelle erreur quadratique moyenne EQM'.

(vi) Comparer EQM et EQM' et conclure. Utiliser ce modèle pour estimer votre pointure.

2 Régression quadratique

Dans cet exercice on considère les durées moyennes mensuelles d'ensoleillement à Rennes entre 1981 et 2010 enregistrées ici(source: meteo-bretagne.fr). On cherche alors à estimer la durée moyenne d'ensoleillement en septembre en fonction des autres données. Ici, les données sont de la forme

$$(M, E) = \begin{pmatrix} M_0 & E_0 \\ M_1 & E_1 \\ \vdots & \vdots \\ M_{11} & E_{11} \end{pmatrix}$$

où $M_i = i + 1$ correspond au numéro du mois et E_i correspond à la durée moyenne d'ensoleillement du mois $i + 1$.

(i) Télécharger la matrice, retirer le couple (M_8, E_8) (c'est lui que l'on cherchera à estimer à partir des autres) et afficher le nuage de point. Quel type de régression polynomiale semble la plus adaptée?

On suppose que les E_i suivent un modèle de régression quadratique, c'est à dire que pour tout i,

$$E_i = aM_i^2 + bM_i + c + \varepsilon_i$$

où les ε_i sont des variables aléatoires i.i.d centrées représentant l'erreur et a, b, c sont des paramètres réels à déterminer. On cherche à estimer les paramètres du modèle par la méthode des moindres carrés, c'est à dire que l'on cherche le triplet $(\hat{a}, \hat{b}, \hat{c})$ qui minimise l'écart entre les données observées et la courbe quadratique estimée, c'est à dire

$$(\hat{a}, \hat{b}, \hat{c}) = \underset{a,b,c}{\arg\min} \sum_{i=0, i\neq 8}^{11} (E_i - aM_i^2 - bM_i - c)^2$$

(ii) (Optionnel) Montrer que le triplet $(\hat{a}, \hat{b}, \hat{c})$ est solution du système linéaire suivant (Penser aux dérivées partielles)

$$\begin{cases} \hat{a}\sum M_i^2 + \hat{b}\sum M_i + \hat{c}\sum 1 &= \sum E_i \\ \hat{a}\sum M_i^3 + \hat{b}\sum M_i^2 + \hat{c}\sum M_i &= \sum M_i E_i \\ \hat{a}\sum M_i^4 + \hat{b}\sum M_i^3 + \hat{c}\sum M_i^2 &= \sum M_i^2 E_i \end{cases}$$

et montrer que sous l'hypothèse que les ε_i suivent une loi $\mathcal{N}(0, \sigma^2)$ alors le triplet $(\hat{a}, \hat{b}, \hat{c})$ correspond au maximum de vraisemblance du modèle.

(iii) Créer une fonction $regquad(M, E)$ retournant les coefficients de la régression quadratique de E sur M. Vérifier que les résultats coïncident avec la fonction *polyfit*.

(iv) Afficher le nuage de points et la courbe de régression quadratique de E sur M. Estimer la quantité d'ensoleillement pour le mois de septembre et comparer avec la vrai valeur.

(v) Refaire les mêmes questions mais cette fois-ci en cherchant à estimer E_{11} à partir des autres données. Que pensez-vous du résultat? Ce modèle est-il vraiment pertinent pour ce type de données? Proposer un meilleur modèle.

3 Régression cubique

Pour finir, on s'intéresse au nombre de mariages à Rennes entre les années 1986 et 2017 que l'on trouve ici (source: data.rennesmetropole.fr) et on cherche à estimer le nombre de mariages qu'il y aura en 2018. Ici, les données sont de la forme

$$(A, M) = \begin{pmatrix} A_0 & M_0 \\ A_1 & M_1 \\ \vdots & \vdots \\ A_{30} & M_{30} \end{pmatrix}$$

où les A_i correspondent à l'année et les M_i au nombre de mariage durant l'année A_i.

(i) Télécharger la matrice (A, M) et afficher le nuage de points correspondant.

(ii) On cherche à modéliser la relation entre M et A par un modèle polynomiale de degré 30. Que dire du polynôme P qui minimise $\sum_{i=1}^{31}(M_i - P(A_i))^2$ parmi les polynômes de degré $\leqslant 30$?

(iii) Le polynôme qui atteint ce minimum colle parfaitement à nos données. Si on le trace on trouve ceci. Pensez-vous que le modèle est pertinent? Quel est le problème?

On supposera plutôt que les M_i suivent un modèle de régression cubique, c'est à dire que pour tout i,

$$M_i = aA_i^3 + bA_i^2 + cA_i + d + \varepsilon_i$$

où les ε_i sont des variables aléatoires i.i.d centrées représentant l'erreur et a, b, c, d sont des paramètres réels à déterminer. On cherche à les estimer par le 4-uplet $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$ qui minimise la fonction

$$f : (a, b, c, d) \mapsto \sum_i (M_i - (aA_i^3 + bA_i^2 + cA_i + d))^2.$$

(iv) Créer une fonction qui prend en argument la liste [a,b,c,d] et renvoie $f(a, b, c, d)$. On remplacera les A_i par $A_i - 1986$ afin d'éviter que la fonction ne prenne de valeurs trop grandes.

(v) Déterminer $\hat{a}, \hat{b}, \hat{c}$ et $\hat{d}$ à l'aide de la fonction *scipy.optimize.minimize*. On prendra $[0, 0, 0, 0]$ comme valeur d'initialisation. Comparer les résultats avec *polyfit*.

(vi) Tracer la courbe de régression cubique et estimer le nombre de mariage à Rennes en 2018.

4 Régression sinusoïdale

A l'aide de la fonction *scipy.optimize.minimize*, trouver la courbe de la forme

$$y = a\cos(\pi x/6 + b) + c$$

qui approche au mieux les données de l'exercice 2 au sens des moindres carrés. Comparer avec celle obtenue par régression quadratique. Pensez-vous que ce modèle est plus adapté?