S2 PeiP — Orsay 2014-2015

Mathématiques

Travaux pratiques de math sur Scilab. Feuille 3.

Moindres carrés, régression linéaire et « data fitting »

Exercice 1 - *Droite de régression : un exemple médical.*

On donne la consommation de viande (en grammes par jour et par personne), et l'incidence du cancer du colon (pour 100000 femmes et par an) dans différents pays industrialisés :

pays	consommation de viande	taux de cancer
Japon	26	7.5
Finlande	101	9.8
Israël	124	16.4
Grande Bretagne	205	23.3
États Unis	284	34

1. Représenter graphiquement les données. Qu'observe-t-on ? Nous allons chercher la droite des moindres carrés, ou droite de régression linéaire, et évaluer la qualité de la corrélation.

2. Rentrer les données sous forme de vecteurs colonnes X et Y. À l'aide du cours (paragraphe 3.5), déterminer les coefficients a et b de la droite de régression $\widetilde{Y} = aX + b$, puis le vecteur $\widetilde{Y}$. Représenter graphiquement cette droite en plus des données initiales.

3. Évaluer la qualité de la régression en calculant le (carré du) coefficient de corrélation $r^2 = \cos^2\theta = \dfrac{\|\widetilde{Y}\|^2}{\|Y\|^2}$ (cf. paragraphe 3.6 du cours).

Exercice 2 - *Un célèbre exemple de loi puissance.*

On donne dans le tableau ci-dessous la période de révolution T des planètes (en année terrestre) et leur distance moyenne a au Soleil (en unité astronomique).

planète	distance a	période T
Mercure	0.387	0.241
Terre	1	1
Jupiter	5.20	11.86
Uranus	19.18	84.0
Pluton	39.53	248.5

1. Représenter les données. Calculer et tracer la droite de régression correspondante. Déterminer le coefficient de corrélation. Est-il mauvais en soi ? Et pourtant, les données n'ont pas l'air « alignées » !

De fait, un certain Johannes Kepler, eu l'idée en 1618 de chercher à lier ces données par une loi puissance : $T = Ca^m$.

2. Trouver les coefficients de la droite de régression linéaire entre les données $\log T$ et $\log a$. En déduire C et a. Interpréter les résultats (comparer avec les valeurs théoriques de C et m, connues en fait).

Exercice 3 - *Fitting d'une loi sinusoïdale.*

On donne ci-dessous la durée d du jour (en heures) le n-ème jour de l'année 2008 à Rome :

jour	n	d
1er Février	32	10
17 Mars	77	12
30 Avril	121	14
31 Mai	152	15

On cherche à exprimer d sous la forme

$$d = a + b\sin\Big(\frac{2\pi}{366}n\Big) + c\cos\Big(\frac{2\pi}{366}n\Big).$$

1. En s'aidant du paragraphe 3.6 du cours, déterminer les « meilleurs » constantes a, b et c au sens des moindres carrés.

2. Représenter les données, et vérifier que le coefficient de corrélation est excellent.

3. Quelle a été la durée du jour le plus long à Rome en 2008 ? (La valeur réelle est de 15h, 13 min, 39 secondes).

Exercice 4 - *Relation linéaire à trois paramètres.*

Dans le tableau suivant, issu d'un livre américain [1] on donne la taille H, le sexe S et le poids P de quelques jeunes adultes (H est donnée en pieds au dessus de 5 pieds, $S = 0$ pour un homme et 1 pour une femme, le poids est donné en Pounds) :

H	S	P
2	1	110
12	0	180
5	1	120
11	1	160
6	0	160

On cherche a mettre P sous la forme

$$P \simeq c_0 + c_1 H + c_2 S$$

au sens des moindres carrés.

1. Avant de faire les calculs, à quels signes vous attendez-vous pour c_0, c_1 et c_2, si les données sont représentatives de la population générale ? Que représentent pratiquement ces coefficients (si la relation cherchée est raisonnable) ?

2. Calculer les meilleurs coefficients c_i aux moindres carrés. Vérifier la qualité de l'approximation sur l'échantillon :

i) en comparant les valeurs réelles de P à celles « calculées » $\widetilde{P} = c_0 + c_1 H + c_2 S$,

ii) en calculant le coefficient de corrélation de cette projection.

3. Comment dépendent ces calculs des unités de mesure ?

1. "Linear Algebra with Applications" de Otto Bretscher. Un livre fabuleux, dont je vous recommande chaudement la lecture, car vous pourrez y parfaire votre connaissance de l'algèbre linéaire tout en pratiquant votre anglais !