Régression linéaire et analyse de la variance

Licence MIASHS - L3

Année 2015 - 2016

Simplice Dossou-Gbété - Jonathan Jung

Feuille de TD n° 1 : Régression linéaire

Exercice 1. Les données ci-dessous décrivent un échantillon de 34 puits par le pH et la concentration en bicarbonate (ppm) dans l'eau souterraine.

Observation	pH	Bicarbonate
1	7.60	157
2	7.10	174
3	8.20	175
4	7.50	188
5	7.40	171
6	7.80	143
7	7.30	217
8	8.00	190
9	7.10	142
10	7.50	190
11	8.10	215
12	7.00	199
13	7.30	262
14	7.80	105
15	7.30	121
16	8.00	81
17	8.50	82
18	7.10	210
19	8.20	202
20	7.90	155
21	7.60	157
22	8.80	147
23	7.20	133
24	7.90	53
25	8.10	56
26	7.70	113
27	8.40	35
28	7.40	125
29	7.30	76
30	8.50	48
31	7.80	147
32	6.70	117
33	7.10	182
34	7.30	87

(a) Effectuer un diagramme en tige et feuille en considérant les 30 premières observations de la concentration en bicarbonate du tableau de données ci-dessus.

(b) Montrer, à l'aide d'un diagramme de dispersion, qu'il n'y a pas d'évidence contre l'hypothèse que la variation moyenne du pH en fonction de la concentration en bicarbonate peut être décrite par un modèle linéaire. Peut-on considérer qu'il n'y a pas d'évidence contre l'hypothèse d'homoscédasticité ?

(c) Calculer les estimations des paramètres de ce modèle sur les 30 premières observations du tableau de données ci-dessus.

(d) Discuter la qualité globale de l'ajustement du modèle aux données.

(e) Peut-on considérer un test d'hypothèse de la nullité de la pente basé sur la loi de Student ?

(f) Peut-on considérer un test d'hypothèse de la nullité de la pente basé sur la loi de Fischer ?

(g) Peut-on considérer qu'il existe des observations influentes ?

(h) Peut-on considérer qu'il existe des observations atypiques ?

(i) Calculer les prévisions des 4 dernières observations ainsi que leurs intervalles de confiance.

(j) Peut-on considérer qu'il n'y a pas d'évidence contre l'hypothèse d'homoscédasticité et de linéarité ?

(k) Peut-on considérer que l'échantillon est compatible avec l'hypothèse de normalité ?

(l) Peut-on considérer que l'échantillon est compatible avec l'hypothèse d'indépendance des observations ?

(m) Reprenez cet exercice en étudiant le diagramme de dispersion de la variation moyenne du pH en fonction du logarithme de la concentration en bicarbonate.

Exercice 2. La taille d'un athlète peut être un bon indicateur de ses résultats en saut en hauteur. Les données utilisées ici présentent la taille et la performance de 20 champions du monde.

Nom	Taille	Performance
Jacobs (EU)	1.73	2.32
Noji (EU)	1.73	2.31
Conway (EU)	1.83	2.40
Matei (Roumanie)	1.84	2.40
Austin (EU)	1.84	2.40
Ottey (Jamaique)	1.78	2.33
Smith (GB)	1.84	2.37
Carter (EU)	1.85	2.37
McCants (EU)	1.85	2.37
Sereda (URSS)	1.86	2.37
Grant (GB)	1.85	2.36
Paklin (URSS)	1.91	2.41
Annys (Belgique)	1.87	2.36
Sotomayor (Cuba)	1.96	2.45
Sassimovitch (URSS)	1.88	2.36
Zhu Jianhua (Chine)	1.94	2.39
Brumel (URSS)	1.85	2.28
Sjoeberg (Suède)	2.00	2.42
Yatchenko (URSS)	1.94	2.35
Povarnitsine (URSS)	2.01	2.40

(a) Représenter la variabilité de cet échantillon à l'aide d'un diagramme de dispersion.

(b) Peut-on faire l'hypothèse d'une relation fonctionnelle entre la performance moyenne des athlètes en saut en hauteur et leur taille ?

(c) Quel est le modèle de cette relation fonctionnelle s'il y en a un ?

(d) En fonction de la réponse à la précédente question, réaliser un ajustement de ce modèle aux données.

(e) Peut-on considérer qu'il existe des observations influentes ?

(f) Peut-on considérer qu'il existe des observations atypiques ?

(g) Peut-on considérer qu'il n'y a pas d'évidence contre l'hypothèse d'homoscédasticité et de linéarité ?

(h) Peut-on considérer que l'échantillon est compatible avec l'hypothèse de normalité ?

(i) Peut-on considérer que l'échantillon est compatible avec l'hypothèse d'indépendance des observations ?

Exercice 3. Le tableau suivant contient la liste de 14 pays d'Amérique du Nord et d'Amérique Centrale dont la population dépassait le milion d'habitants en 1985. Pour chaque pays, on mesure le taux de natalité (nombre de naissances annuel pour 1000 habitants ainsi que le taux d'urbanisation (pourcentage de la population vivant dans des villes de plus de 100 000 habitants). On fait l'hypothèse de régression linéaire simple du type $y_i = a + b\,x_i$, c'est-à-dire que le taux de natalité dépend linéairement du taux d'urbanisation.

Pays	Urbanisation	Natalité
Canada	55.0	16.2
Costa-Rica	27.3	30.5
Cuba	33.3	16.9
USA	56.5	16.0
El Salvador	11.5	40.2
Guatemala	14.2	38.4
Haïti	13.9	41.3
Honduras	19.0	43.9
Jamaïque	33.1	28.3
Mexique	43.2	33.9
Nicaragua	28.5	44.2
Trinidade/Tobago	6.8	24.6
Panama	37.7	28.0
Rep. Dominicaine	37.1	33.1

(a) Représenter la variabilité de cet échantillon à l'aide d'un diagramme de dispersion.

(b) Peut-on faire l'hypothèse d'une relation fonctionnelle entre le taux de natalité moyen et le taux d'urbanisation ?

(c) Si oui, quel est le modèle de cette relation fonctionnelle s'il y en a un ?

(d) En fonction de la réponse à la précédente question, réaliser un ajustement de ce modèle aux données.

(e) Peut-on considérer qu'il existe des observations influentes ?

(f) Peut-on considérer qu'il existe des observations atypiques ?

(g) Peut-on considérer qu'il n'y a pas d'évidence contre l'hypothèse d'homoscédasticité et de linéarité ?

(h) Peut-on considérer que l'échantillon est compatible avec l'hypothèse de normalité ?

(i) Peut-on considérer que l'échantillon est compatible avec l'hypothèse d'indépendance des observations ?