T. D. n° 9
Régression linéaire simple

Les deux premiers exercices s'inspirent du livre de Y. Dodge, *Analyse de régression appliquée*, aux éditions Dunod, 1999. Le dernier exercice provient du livre de G. Baillargeon, *Probabilités, Statistique et Techniques de régression*, aux éditions SMG, 1989.

Exercice 1. Les athlètes de saut en hauteur.

La taille d'un athlète peut jouer un rôle important dans ses résultats en saut en hauteur. Les données utilisées ci-dessous présentent donc la taille et la performance de 30 champions du monde.

1. À partir de l'échantillon proposé, utiliser la méthode des moindres carrés ordinaires pour estimer les paramètres de la droite de régression linéaire :

$$(\text{Performance}) = \beta_0 + \beta_1 \times (\text{Taille}) + \varepsilon.$$

2. Compléter le tableau d'analyse de la variance (dit aussi tableau d'ANOVA) :

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens	F_{obs}	F_c
expliquée					
résiduelle					
totale					

   Puis réaliser le test de Fisher au seuil de significativité $\alpha = 5\%$. Que concluez-vous ?

3. Quel pourcentage de la variation totale des performances est expliqué par la variable taille ? Que pensez-vous de ce résultat ? Que faudrait-il faire en tant que chargé de cette étude ?

Observation	Nom	Taille	Performance
1	Jacobs (EU, 1978)	1,73	2,32
2	Noji (EU, 1936)	1,73	2,31
3	Conway (EU, 1989)	1,83	2,40
4	Matei (Roumanie, 1990)	1,84	2,40
5	Austin (EU, 1996)	1,84	2,39
6	Ottey (Jamaique)	1,78	2,33
7	Smith (GB, 1992 et 1993)	1,85	2,37
8	Carter (EU)	1,85	2,37
9	McCants (EU)	1,85	2,37
10	Sereda (URSS)	1,86	2,37
11	Grant (GB)	1,85	2,36
12	Paklin (URSS, 1985)	1,91	2,41
13	Annys (Belgique, 1985)	1,87	2,36
14	Sotomayor (Cuba, 1993)	1,95	2,45
15	Sassimovitch (URSS)	1,88	2,36
16	Zhu Jianhua (Chine, 1984)	1,94	2,39
17	Brumel (URSS, 1963)	1,85	2,28
18	Sjöberg (Suède, 1987)	2,00	2,42
19	Yatchenko (URSS, 1978)	1,94	2,35
20	Povarnitsyn (URSS, 1985)	2,01	2,40
21	Voronin (Russie, 2000)	1,91	2,40
22	Ukhov (Russie, 2012)	1,92	2,39
23	Essa Barshim (Qatar, 2012)	1,89	2,39
24	Holm (Suède, 2005)	1,81	2,40
25	Sjöberg (Suède, 1987)	2,00	2,41
26	Prezelj (Slovénie, 2012)	1,94	2,32
27	Forsyth (Australie, 1997)	2,00	2,36
28	Kemp (Bahamas, 1995)	1,87	2,38
29	Buss (Allemand, 2001)	1,95	2,36
30	Freitag (Sud-Africain, 2005)	2,04	2,38

Exercice 2. Calories.

Soient les données présentées dans le tableau ci-dessous. Il s'agit du nombre de calories consommées par jour et par personne et du pourcentage de la superficie agricole dans 30 pays.

1. Représenter graphiquement Y en fonction de X.
2. Estimer, à l'aide du logiciel `R`, les paramètres β_0 et β_1 du modèle linéaire simple :
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$
3. Calculer, à l'aide du logiciel `R`, le tableau d'analyse de la variance correspondant à cette régression.
4. Calculer, à l'aide du logiciel `R`, un intervalle de confiance à 95% autour de la droite des moindres carrés ordinaires.

5. Représenter sur le graphique de la question 1 la droite des moindres carrés ordinaires et l'intervalle de confiance calculé à la question 4.

Observation	Pays	% Superficie agricole (1000Ha)	Calories par jour et par personne
1	Luxembourg	131,0	4 713
2	Autriche	2 869,0	4 023
3	Israël	520,5	3 831
4	États-Unis	411 262,5	3 754
5	Italie	9 227,0	3 754
6	Grèce	8 152,0	3 706
7	Portugal	3 636,0	3 635
8	Uruguay	14 378,0	3 576
9	Cuba	6 570,0	3 547
10	France	29 090,0	3 541
11	Lituanie	2 805,9	3 530
12	Arabie Saoudite	173 355,0	3 527
13	Algérie	41 383,0	3 510
14	Irlande	4 555,0	3 503
15	Pologne	14 779,0	3 503
16	Danemark	2 690,0	3 494
17	Maroc	30 103,8	3 492
18	Canada	62 597,0	3 486
19	Pays-Bas	1 894,8	3 479
20	République tchèque	4 229,0	3 303
21	Roumanie	13 982,0	3 263
22	Suède	3 066,0	3 114
23	Suisse	1 522,9	3 085
24	Brésil	275 030,0	2 926
25	Bulgarie	62 597,0	2 839
26	Chine	519 148,2	2 740
27	Pérou	21 500,0	2 583
28	Inde	179 799,0	2 529
29	Sénégal	9 505,0	2 513
30	Pakistan	26 550,0	2 422

Exercice 3. Le composant électronique.

Un certain composant électronique est fabriqué une fois par mois par l'entreprise Micro-Systèmes. La quantité fabriquée varie avec la demande du marché. Dans le but de planifier la production et d'établir certaines normes sur le nombre d'hommes-minutes exigés pour la production de différents lots de ce composant électronique, le responsable de la production a relevé l'information suivante pour 30 cédules de production. Le nombre d'hommes-minutes est identifié par Y et la quantité fabriquée par X.

1. Quelle serait la première étape à franchir avant d'aborder tout calcul préliminaire ?
2. Le responsable de la production envisage d'utiliser le modèle linéaire simple comme modèle prévisionnel. Spécifier ce modèle et bien identifier chacune des composantes du modèle dans le contexte de ce problème.
3. Déterminer l'équation de la droite des moindres carrés ordinaires.
4. D'après l'équation de la droites des moindres carrés ordinaires, si le nombre d'unités à fabriquer augmente de 10, quelle sera l'augmentation correspondante du nombre moyen d'hommes-minutes requis ?
5. En l'absence de l'information que nous donne la quantité à fabriquer, quelle serait une bonne estimation du nombre d'hommes-minutes requis ?
6. Quelle correction peut-il apporter à son estimation du nombre moyen d'hommes-minutes requis en introduisant la connaissance de X dans son analyse ?
7. Donner la valeur de $s(\widehat{\beta}_1)$ et tester, au seuil de significativté $\alpha = 5\%$ les deux hypothèses suivantes :
$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{contre} \qquad \mathcal{H}_1 : \beta_1 \neq 0.$$
8. Donner la variation expliquée par la droite des moindres carrés ordinaires et la variation inexpliquée par la droite.
9. Déterminer le pourcentage de variation expliqué par la droite de régression.
10. Donner une estimation du nombre moyen d'hommes-minutes requis pour : $x_h = 42; x_h = 57; x_h = 72$.
11. Pour quelle quantité X_n, l'estimation du nombre moyen d'hommes-minutes requis serait-elle la plus précise ?
12. Entre quelles valeurs peut se situer le vrai nombre moyen d'hommes-minutes requis pour les lots dont la quantité a été déterminée à la question 11. ? Utiliser un niveau de confiance de 95%.
13. Quelle est la marge d'erreur dans l'estimation effectuée en 12. ?

y_i	150	192	264	371	300	358	192	134	242	238
x_i	35	42	64	88	70	85	40	30	55	60
y_i	226	302	340	182	169	149	183	273	362	311
x_i	38	40	68	85	70	51	72	80	44	39
y_i	347	183	123	213	217	237	313	331	193	158
x_i	75	37	28	55	52	53	74	76	48	37