Feuille de Travaux Dirigés n° 1
Régression linéaire simple

Exercice I.1. D'après Baillargeon, *Probabilités, Statistiques et techniques de régression*, Les éditions SMG, 1995.
On donne les couples d'observations suivants :

x_i	18	7	14	31	21	5	11	16	26	29
y_i	55	17	36	85	62	18	33	41	63	87

a) Tracer à l'aide de **R** le diagramme de dispersion des couples $(x_i; y_i)$. À partir de ce diagramme, observez-vous une liaison linéaire entre ces deux variables ? Comment procédez-vous pour confirmer ce que vous observez ?

b) À l'aide de **R**, donner pour ces observations la droite des moindres carrés. Demander à **R** de vous calculer les valeurs prédites pour les différentes valeurs des x_i. Tracer à l'aide de **R** la droite des moindres carrés.

c) À partir de cette question et ce jusqu'à la fin de l'exercice, vous ne devez plus utiliser **R**, si ce n'est que la calculatrice de **R** !
Quelle est une estimation plausible de Y à $x_i = 21$?

d) Quel est l'écart entre la valeur observée de Y à $x_i = 21$ et la valeur estimée avec la droite des moindres carrés ? Comment appelle-t-on cet écart ?

e) Est-ce que la droite des moindres carrés obtenue en b) passe par le point $(\overline{x}; \overline{y})$? Peut-on généraliser cette conclusion à n'importe laquelle droite de régression ?

Remarque : Voici quelques lignes de commande qui pourront par exemple vous aider à répondre aux questions. À vous de savoir à quoi elles répondent.

```
> Chemin <- "C :\\Documents and Settings\\Bertrand\\Bureau\\"
> Exo1<-read.table(paste(Chemin,"Exo1-TD5-Estimation.csv"
+,sep=""),sep=" ;",header=T)
> str(Exo1)
'data.frame' : 10 obs. of 2 variables :
$ x_i : int 18 7 14 31 21 5 11 16 26 29
$ y_i : int 55 17 36 85 62 18 33 41 63 87
> plot(Exo1)
> Droite<-lm(y_i ~ x_i,data=Exo1)
> coef(Droite)
(Intercept) x_i
1.021341 2.734756
> fitted(Droite)
1 2 3 4 5 6 7 8 9 10
50.24695 20.16463 39.30793 85.79878 58.45122 14.69512 31.10366
+ 44.77744 72.12500 80.32927
> abline(coef(Droite),col="red")
```

```
>fitted(Droite)[5]
> residuals(Droite)
1 2 3 4 5 6 7 8 9 10
4.7530488 -3.1646341 -3.3079268 -0.7987805 3.5487805 3.3048780
+ 1.8963415 -3.7774390 -9.1250000 6.6707317
> residuals(Droite)[5]
5
3.548780
>mean(Exo1 $ x_i)
17.8
> mean(Exo1 $ y_i)
49.7
> 1.021341+2.734756*17.8
49.7
```

Exercice I.2. L'étudiant et le professeur, qui a raison entre les deux ?
Un étudiant de l'ENGREF veut utiliser la régression linéaire pour estimer le volume en bois utilisable d'un arbre debout en fonction de l'aire du tronc mesuré à 25 cm du sol. Il a choisi au hasard 10 arbres et a mesuré, à la base, l'aire correspondante (en cm^2). Il a par la suite enregistré, une fois l'arbre coupé, le volume correspondant en m^3.

Vol.	0,152	0,284	0,187	0,350	0,416	0,230	0,242	0,276	0,383	0,140
Aire	297	595	372	687	790	520	473	585	762	232

a) À l'aide de **R**, calculer le coefficient de corrélation linéaire entre ces deux variables. Que concluez-vous ?

b) Tester l'hypothèse suivante :

$$(H_0) : \rho(X, Y) = 0.$$

Utiliser $\alpha = 0,05$.

c) Son professseur lui mentionne qu'il peut, à l'oeil, évaluer avec une assez bonne précision le volume d'un arbre. L'étudiant un peu perplexe lui lance un défi : "je gage 1 euro que je fais mieux que vous avec le modèle des moindres carrés". "D'accord."
Ayant justement un arbre tout près, le professeur lui dit, après une expertise de quelques minutes que cet arbre a un volume de 0,22 m^3. Sans plus tarder ; l'étudiant mesure l'aire de la base de l'arbre et obtient 465 cm^2. Calculer avec la droite des moindres carrés obtenue avec Mintab, l'estimation la plus plausible du volume de l'arbre.

d) On s'acharne par la suite à couper l'arbre et le volume correspondant est 0,24 m^3. Celui qui a le plus faible écart de prévision empoche le pari. Lequel s'est enrichi de 1 euro ?

e) Est ce que le volume moyen des arbres échantillonnés aurait donné une estimation aussi bonne que la droite des moindres carrés pour cet arbre ?

Exercice I.3. D'après Frontier, Davoult, Gentilhomme, Lagadeuc *Statistique pour les sciences de la vie et de l'environnement*, Dunod, 2001. On étudie l'influence d'un antibiotique sur une culture bactérienne. On répartit dans 10 tubes des volumes égaux de culture additionnés d'une quantité X d'antibiotique, et on mesure, après incubation, la densité optique D. Les résultats sont les suivants.

X	0,2	0,2	0,4	0,4	0,6	0,6	0,8	0,8	1,0	1,0
D	19	21	35	38	64	66	115	130	200	210

a) Un ajustement linéaire semble-t-il justifié ? Pour répondre à cette question, utiliser **R**. Que devez-vous calculer comme coefficient avec **R** ?

b) En transformant une des deux variables avec une fonction adaptée, déterminer une équation de régression en précisant quelles sont la variable explicative et la variable expliquée ?

c) Donner à l'aide de **R**, une prévision de D pour une quantité d'antibiotique $X = 0,5$. Calculer, toujours à l'aide de **R**, l'intervalle de sécurité à 95% de cette prévision.

Exercice I.4. D'après Frontier, Davoult, Gentilhomme, Lagadeuc *Statistique pour les sciences de la vie et de l'environnement*, Dunod, 2001. On mesure le poids frais et le poids sec de 20 prélèvements de plancton. Les résultats sont les suivants (exprimés en g par 10 m^3 d'eau de mer)

poids frais	20,4	28,4	48,7	28,8	32,9	85,2	32,2	27,8	27,0	36,7
poids sec	3,6	3,4	5,6	4,1	3,3	9,3	3,7	3,2	2,9	4,5
poids frais	20,4	24,3	24,3	18,0	31,7	25,7	41,2	53,0	61,0	61,2
sec	2,6	2,8	3,1	2,6	4,4	2,8	4,6	6,0	7,2	6,3

a) Calculer à l'aide de **R** le coefficient de corrélation linéaire entre le poids frais et le poids sec. Est-il significatif et à quel seuil ? Repérer un *outsider* parmi les couples de valeurs ; l'éliminer et reprendre la question. Pour cette dernière partie de question, vous devez appliquer la procédure étudiée en cours.

b) La teneur en eau de chaque prélèvement planctonique est estimée par la différence entre poids frais et poids sec. Estimer sa variance à l'aide de **R**.

c) Y a-t-il un sens à calculer le coefficient de corrélation entre le poids frais et la teneur en eau ainsi estimée, et pourquoi ?

d) Donner, à l'aide de **R**, la droite permettant de connaître approximativement le poids sec après une mesure de poids frais. Quelle est, dans ces conditions, la proportion de variance du poids sec expliquée par la régression ?

e) Soit un poids frais de 40 grammes. Calculer, à l'aide de **R**, la valeur la plus probable du poids sec, et son intervalle de sécurité à 95%.