T. D. n^{o} 7
Régression linéaire simple avec R

Exercice 1. D'après Baillargeon, Probabilités, Statistiques et techniques de régression, Les éditions SMG, 1995. Cet exercice doit se traiter en grande partie avec R.
Nous donnons 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

1. La première étape est d'obtenir les données. Pour cela, vous pouvez les télécharger sur mon site, puis les enregistrer sur le bureau du poste. Par exemple, depuis le bureau de mon ordinateur portable, les lignes de commande à taper sous R sont les suivantes :
```
> Chemin <- "C :\\Documents and Settings\\Bertrand\\Bureau\\"
> Exo1<-read.table(paste(Chemin,"Exo1-TD7-Estimation.csv",
+sep=""),+sep=" ;",header=T)
```
2. Tracer le diagramme de dispersion des couples $(x_i; y_i)$. À la vue de ce diagramme, pouvons-nous soupçonner une liaison linéaire entre ces deux variables ?
3. Déterminer pour ces observations la droite des moindres carrés, c'est-à-dire donner les coefficients de la droite des MC.
4. Donner les ordonnées des y_i calculés par la droite des moindres carrés correspondant aux différentes valeurs des x_i.
5. Tracer ensuite la droite sur le même graphique.
6. Quelle est une estimation plausible de Y à $x_i = 21$?
7. 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 appelons-nous cet écart ?
8. Est-ce que la droite des moindres carrés obtenue en b) passe par le point $(\overline{x}; \overline{y})$? Pouvons-nous généraliser cette conclusion à n'importe laquelle droite de régression ?

   Remarque : Voici quelques lignes de commande qui pourront 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-TD7-Estimation.csv"
+,sep=""),sep=" ;",header=T)
> Exo1
```
x_i y_i
```
1 18 55
2 7 17
3 14 36
4 31 85
5 21 62
```

```
6 5 18
7 11 33
8 16 41
9 26 63
10 29 87
> 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")
> 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
```

Exercice 2. D'après Baillargeon, Probabilités, Statistiques et techniques de régression, Les éditions SMG, 1995. Cet exercice doit se traiter en grande partie avec R.

La société de Transport Bertrand veut établir une politique d'entretien des camions de sa flotte. Tous sont de même modèle et utilisés à des transports semblables. La direction de la société est d'avis qu'une liaison statistique entre le coût direct de déplacements (*cents* par *km*) et l'espace de temps écoulé depuis la dernière inspection de ce camion serait utile. Nous avons donc recueilli un certain nombre de données sur ces deux variables. Nous souhaitons utiliser la régression linéaire comme modélisation statistique.

Coût direct	10	18	24	22	27	13	10	24	25	8	16
Nombre de mois	3	7	10	9	11	6	5	8	7	4	6
Coût direct	20	28	22	19	18	26	14	20	26	30	12
Nombre de mois	9	12	8	10	9	11	6	8	10	12	5

1. Quelle variable devrions-nous identifier variable dépendante (Y) et laquelle devrions-nous identifier variable explicative (X) ?
2. Tracer le diagramme de dispersion de ces observations. Est-ce que le nuage de points suggère une forme de liaison particulière ?

3. Calculer l'équation de la droite des moindres carrés.
4. Avec l'équation de la droite des moindres carrés, quelle est l'estimation la plus plausible du coût direct de déplacement pour des camions dont la dernière inspection remonte à 6 mois ?
5. D'après les résultats de cette étude, un délai supplémentaire d'un mois pour l'inspection d'un camion occasionnera-t-il une augmentation ou une diminution du coût direct ? Quelle sera vraisemblablement la valeur de cette variation de coût ?
6. Déterminer la variation totale dans le coût direct de déplacement.
7. L'équation de la droite des moindres carrés pour les données de la société est : $\hat{y}_i = 1,54941 + 2,26087\, x_i$. Calculer la variation qui est expliquée par la droite des moindres carrés.
8. Quelle est la variation résiduelle ?
9. Calculer le coefficient R^2 et interpréter le résultat.

Exercice 3. Cet exercice doit se traiter en grande partie avec `R`.
Une étudiante en sociologie veut analyser, dans le cadre d'un projet de fin de session, s'il existe une relation linéaire entre la densité de population dans les régions métropolitaines et le taux de criminalité correspondant dans ces régions.
Le taux de criminalité (Y) est indiqué en nombre de crimes par 10 000 habitants et la densité de population (X) est mesurée en milliers d'habitants par km^2.

Région	1	2	3	4	5	6	7	8	9	10	11	12
x_i	7,7	5,8	11,5	2,1	3,7	3,6	7,5	4,2	3,8	10,3	8,6	7,2
y_i	12	9	15	4	4	2	10	3	5	11	10	11

1. Tracer le diagramme de dispersion de ces observations.
2. Calculer les coefficients de la droite des moindres carrés.
3. À quelle augmentation du taux de criminalité pouvons-nous nous attendre pour une variation unitaire (ici 1 000 habitants par km^2) de la densité de population ?
4. Estimer le taux de criminalité le plus plausible pour une densité de population de 75 000 habitants par km^2.
5. À l'aide des calculs préliminaires, calculer la variation totale du taux de criminalité.
6. Calculer la variation qui est expliquée par la droite des moindres carrés.
7. Quelle proportion de la variation totale est expliquée par la droite des moindres carrés ?

Exercice 4. Cet exercice doit se traiter en grande partie avec `R`.
Un étudiant en techniques forestières 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

1. Déterminer les coefficients de la droite des moindres carrés.
2. 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'étudaint mesure l'aire de la base de l'arbre et obtient $465\, cm^2$. Calculer avec la droite des moindres carrés, l'estimation la plus plausible du volume de l'arbre.
3. L'étudiant 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 ?
4. 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 5. Cet exercice doit se traiter en grande partie avec `R`.
L'entreprise INFORMATEX se spécialise dans l'analyse de systèmes et la programmation sur ordinateur de problèmes techniques et de gestion. Elle veut utiliser la régression dans une étude sur le temps requis, par ses analystes-programmeurs, pour programmer des projets complexes.

Cette étude pourrait permettre à la firme d'établir des normes quant au temps requis pour programmer certains projets et d'assurer éventuellement une meilleure planification des ressources humaines. Les données du tableau suivant représentent le temps total en heures requis pour programmer différents projets en fonction du nombre d'instructions dans chaque programme.

Temps total en heures	40	55	62	58	82	94	120
Nombre d'instructions	60	82	100	142	190	220	285
Temps total en heures	134	128	140	152	174	167	218
Nombre d'instructions	354	400	425	440	500	530	640

1. Si nous voulons expliquer les fluctuations dans le temps requis pour programmer les projets quelle variable devons-nous identifier comme variable dépendante ? Comme variable explicative ?
2. Qu'est-ce qui peut renseigner l'entreprise sur la forme de liaison statistique qui peut exister entre ces deux variables ?
3. Quelle méthode d'ajustement linéaire devons-nous utiliser pour obtenir les estimateurs des coefficients de la droite de régression ?

4. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
5. Quelle est l'équation de la droite des moindres carrés ?
6. Si nous ne tenons pas compte du nombre d'instructions, quelle valeur pourrions-nous utiliser comme estimation du temps moyen de programmation des projets ?
7. Quelle correction pouvons-nous apporter à l'estimation obtenue en 6., en tenant compte du nombre d'instructions par l'entremise de la droite des moindres carrés ?
8. D'après la droite des moindres carrés, à quelle augmentation du temps de programmation pouvons-nous nous attendre lorsque le nombre d'instructions augmente de 50 ?
9. Pour chaque nombre d'instructions suivant, estimer le temps de programmation à l'aide de la droite des moindres carrés

Nombre d'instructions	100	220	440
Estimation du temps de programmation			

10. Selon les résultats observés, quels sont les écarts de prévision de l'équation des moindres carrés pour le nombre d'instruction en 9. ?
11. Si nous avions utilisé l'estimation obtenue en 6. au lieu de celles déduites de l'équation des moindres carrés pour effectuer les prévisions selon le nombre d'instructions spécifié en 9., quels auraient été alors, dans chaque cas, les écarts de prévision ?
12. Pour chaque valeur x_i spécifié en k), vérifier la relation $y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$.
13. Calculer la variation totale, la variation expliquée par la droite des moindres carrés et la variation résiduelle.
14. Quelle proportion de la variation totale dans le temps de programmation est expliquée par la droite des moindres carrés ? Quelle proportion demeure inexpliquée par la droite ?
15. Nous avions fixé le R^2 à $0,90$ comme valeur minimale pour considérer la droite des moindres carrés d'utilité pratique. D'après les résultats obtenus, devrions-nous utiliser la droite des moindres carrés comme outil de prévision ?