Feuille de Travaux Dirigés n° 10
Régression linéaire simple avec R

Exercice X.1. Étude de la pollution de l'air. Cet exercice est issu du livre « Statistiques avec R », Pierre-André Cornillon et autres, aux éditions PUR, 2008.

1. Récupérer les données dans R en exécutant les instructions suivantes :
   `>ozone<-read.table("...ozone.txt",header=T)`
2. Quelles sont les différentes variables ? Quelle est leur nature ? Obtenir les statistiques descriptives du jeu de données.
3. Nous allons nous intéresser plus particulièrement à deux variables du jeu de données : la variable `"maxO3"` et la variable `"T12"`. Pour cela, nous allons calculer les statistiques élémentaires sur ces deux variables, en utilisant la commande `summary`
   `>summary(ozone[,c("maxO3","T12")])`
4. Nous allons maintenant représenter le nuage de points (x_i, y_i). Pour cela, exécuter la ligne de commande suivante :
   `>plot(maxO3~T12,data=ozone,pch=15,cex=.5)`
5. Nous allons maintenant estimer les paramètres du modèle de régression linéaire. Pour cela, nous allons utiliser la fonction `lm` (lm pour linear model). Cette fonction permet d'ajuster un modèle linéaire. Exécuter les lignes de commande suivantes :
   `>reg.simple<-lm(maxO3~T12,data=ozone)`
   `>summary(reg.simple)`
6. Nous pouvons consulter la liste des différents résultats de l'objet `reg.simple` avec :
   `>names(reg.simple)`
7. Nous pouvons récupérer les coefficients avec :
   `>reg.simple$coef`
   ou en utilisant la fonction `coef`, ce qui est la méthode recommandée :
   `>coef(reg.simple)`
8. **Remarque :** Nous avons auparavant construit un modèle avec constante. Si jamais la constante n'était pas significative ou si encore le modèle ne doit pas contenir de constante, alors nous devons procéder de la manière suivante pour le construire :
   `>reg.ss.constante<-lm(maxO3~T12-1,data=ozone) >reg.ss.constante`
9. Nous allons maintenant tracer la droite de régression. Nous pouvons simplement appliquer la commande `abline(reg.simple)` mais nous préférons nous restreindre au domaine d'observation de la variable explicative. Nous créons donc une grille de points sur les abscisses et appliquons le modèle trouvé sur cette grille :

```
>plot(max03~T12,data=ozone,pch=15,cex=.5)
>grillex<-seq(min(ozone[,"T12"]),max(ozone[,"T12"]),
+length=100)
>grilley<-reg.simple$coef[1]+reg.simple$coef[2]*grillex
>lines(grillex,grilley,col=2)
```

10. Nous allons maintenant analyser les résidus. Les résidus sont obtenus par la fonction `residuals`, cependant les résidus obtenus ne sont pas de même variance (hétéroscédastiques). Nous allons utiliser alors les résidus studentisés, qui eux sont de même variance.

```
>res.simple<-rstudent(reg.simple)
>plot(res.simple,pch=15,cex=.5,ylab="Résidus",ylim=c(-3,3))
>abline(h=c(-2,0,2),lty=c(2,1,2))
```

11. Pour terminer nous allons donner la procédure pour prévoir une nouvelle valeur. Ayant une nouvelle observation `xnew`, il suffit d'utiliser les estimations pour prévoir la valeur de Y correspondante. Cependant, la valeur prédite est de peu d'intérêt sans l'intervalle de confiance associé. Voyons cela sur un exemple. Nous disposons d'une nouvelle observation de la température `T12` égale à 19 degrés pour le 1er octobre 2001.

```
>xnew=19
>xnew<-as.data.frame(xnew)
>colnames(xnew)<-"T12"
>predict(reg.simple,xnew,interval="pred")
```

Il faut noter que l'argument `xnew` de la fonction `predict` doit être un data-frame avec les mêmes noms de variables explicatives (ici `T12`). La valeur prévueest `76.5` et l'intervalle de prévision à 95% est `[41.5,111.5]`. Pour représenter sur un même graphique l'intervalle de confiance d'une valeur lissée et l'intervalle de confiance d'une prévision, nous calculons ces intervalles pour l'ensemble des points ayant servi à dessiner la droite de régression. Nous faisons figurer les deux sur le même graphique.

```
>grillex.df<-data.frame(grillex)
>dimnames(grillex.df)[[2]]<-"T12"
>ICdte<-predict(reg.simple,new=grillex.df,interval="conf",
+level=0.95)
>ICprev<-predict(reg.simple,new=grillex.df,interval="pred",
+level=0.95)
>plot(max03~T12,data=ozone,pch=15,cex=.5)
>matlines(grillex,cbind(ICdte,ICprev[,-1]),lty=c(1,2,2,3,3),
+col=1)
>legend("topleft",lty=2 :3,c("prev","conf"))
```

Exercice X.2. 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 le site qui vous est consacré, 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,"MASTER2TD10_EX2.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 à la question 3. passe par le point $(\overline{x}; \overline{y})$? Pouvons-nous généraliser cette conclusion à n'importe laquelle droite de régression ?

Remarque : Voici les lignes de commande qui pourront vous aider à répondre aux questions ainsi que quelques sorties.

1. ```
   > Chemin<-"C :\\Documents and Settings\\Bertrand\\
   + Bureau\\"
   > Exo1<-read.table(paste(Chemin,"Exo1-TD5-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
```

2. `> plot(Exo1)`
3. `> Droite<-lm(y_i~x_i,data=Exo1)`

```
> coef(Droite)
(Intercept)         x_i
   1.021341    2.734756
```

4. `> fitted(Droite)`

```
       1        2        3        4        5        6
50.24695 20.16463 39.30793 85.79878 58.45122 14.69512
       7        8        9       10
31.10366 44.77744 72.12500 80.32927
```

5. `> abline(coef(Droite),col="red")`
6.
7. `> residuals(Droite)`

```
          1          2          3          4          5
 4.7530488 -3.1646341 -3.3079268 -0.7987805  3.5487805
          6          7          8          9         10
3.3048780   1.8963415 -3.7774390 -9.1250000  6.6707317
```

```
> residuals(Droite)[5]
       5
3.548780
```

Exercice X.3. 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 X.4. 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 X.5. 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'étudiant 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 X.6. 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 ?

Les quatre exercices suivant proviennent du livre de Y. Dodge, *Analyse de régression appliquée*, aux éditions Dunod, 1999.

Exercice X.7. Les athlètes.
La taille d'un athlète peut jouer un rôle important dans ses résultats en saut en hauteur. Les données utilisées ici présentent donc la taille et la performance de 20

champions du monde.

Observation	Nom	Taille	Performance
1	Jacobs (EU)	1,73	2,32
2	Noji (EU)	1,73	2,31
3	Conway (EU)	1,83	2,40
4	Matei (Roumanie)	1,84	2,40
5	Austin (EU)	1,84	2,40
6	Ottey (Jamaique)	1,78	2,33
7	Smith (GB)	1,84	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)	1,91	2,41
13	Annys (Belgique)	1,87	2,36
14	Sotomayor (Cuba)	1,96	2,45
15	Sassimovitch (URSS)	1,88	2,36
16	Zhu Jianhua (Chine)	1,94	2,39
17	Brumel (URSS)	1,85	2,28
18	Sjoeberg (Suède)	2,00	2,42
19	Yatchenko (URSS)	1,94	2,35
20	Povarnitsine (URSS)	2,01	2,40

1. À partir de l'échantillon proposé, utiliser la méthode des moindres carrés pour estimer les paramètres de la 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	Degrés de liberté	Somme des carrés	Moyenne des carrés	F_{obs}
Régression				
Résiduelle				
Totale				

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 ?

Exercice X.8. Un exercice pour pratiquer.
Nous disposons des données suivantes au sujet de deux variables d'intérêt X et Y :

x_i	7	9	9	10	13	17	19	20	21	25
y_i	5	4	6	4	1	2	0	1	1	0

Nous nous référons au modèle linéaire :
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

1. Estimer les paramètres β_0 et β_1 par la méthode des moindres carrés.
2. Pour chacun de ces deux paramètres, trouver un intervalle de confiance avec un niveau de confiance de 99%.
3. Soit $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i \, (i = 1, \ldots, n)$ où $\widehat{\beta_0}$ et $\widehat{\beta_1}$ sont les estimateurs de β_0 et β_1 obtenus en 1). Démontrer que nous avons $\sum \widehat{Y_i} = \sum Y_i$, par deux méthodes (mathématique, et avec `R`).
4. Donner les intervalles de confiance pour les $\mu_Y(X)$.
5. Représenter graphiquement les points (x_i, y_i), la droite de régression et l'ensemble des intervalles de confiance pour les $\mu_Y(X)$.

Exercice X.9. Trois exemples.

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F_{obs}	F_t	R^2
Régression	1	501,76	501,76	7,575	4,75	0,387
résiduelle	12	794,90	66,24			
Totale	13	1 296,66				

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F_{obs}	F_t	R^2
Régression	1	34,186	34,186	43,44	7,71	0,916
résiduelle	4	3,148	0,787			
Totale	5	37,333				

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F_{obs}	F_t	R^2
Régression	1	179,76	179,76	27,87	4,08	0,367
résiduelle	48	309,56	6,45			
Totale	49	489,32				

1. En comparant dans les exemples ci-dessus, les liens entre les valeurs F_{obs}, F_t (F_t est la valeur lue dans la table de Fisher) et R^2, quelles sont, selon vous, les meilleures régressions ?
2. Avant de calculer le coefficient de détermination R^2, en n'utilisant que les valeurs F_{obs} et F_t, quelle règle pourrions-nous énoncer pour réperer rapidement une bonne analyse de régression ?

Exercice X.10. Calories.
Soient les données présentées dans le tableau ci-dessous. Il s'agit du nombre de calories consommées par jour et du pourcentage de population agricole dans 11

pays.

Observation i	Pays	% Population agricole	Calories par jour et par personne
1	Suisse	4, 0	3 432
2	France	5, 7	3 273
3	Suède	4, 9	3 049
4	USA	3, 0	3 642
5	Ex-URSS	14, 8	3 394
6	Chine	69, 6	2 628
7	Inde	63, 8	2 204
8	Brésil	26, 2	2 643
9	Pérou	38, 3	2 192
10	Algérie	24, 7	2 687
11	Ex-Zaire	65, 7	2 159

1. Représenter graphiquement Y en fonction de X.
2. Estimer les paramètres β_0 et β_1 du modèle :
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$
3. Construire le tableau d'analyse de la variance correspondant à cette régression.
4. Construire un intervalle de confiance à 95% autour de la droite de régression.
5. Représenter sur le graphique de la question (1) la droite de régression et l'intervalle de confiance calculé à la question (4).

Le dernier exercice provient du livre de G. Baillargeon, *Probabilités, Statistique et Techniques de régression*, aux éditions SMG, 1989.

Exercice X.11. 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 productoin a relevé l'information suivante pour 15 cédules de production. Le nombre d'hommes-minutes est identifié par Y et la quantité fabriquée par X.

y_i	150	192	264	371	300	358	192	134	242	238	226	302	340	182	169
x_i	35	42	64	88	70	85	40	30	55	60	51	72	80	44	39

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 régression.

4. D'après l'équation de régression, 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 les deux hypothèses suivantes avec la loi de Student.
$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{contre} \qquad \mathcal{H}_1 : \beta_1 \neq 0.$$
8. Donner la variation qui est expliquée par la droite de régression et la variation qui est inexpliquée par la droite.
9. Déterminer le pourcentage de variation qui est 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. ?