Devoir 1 - Solutions

Patrick Boily

2023-01-28

Q1

a) Soient $U_i \sim \chi^2(r_i)$ des variables aléatoires indépendantes avec $r_1 = 5$, $r_2 = 10$. Posons

$$X = \frac{U_1/r_1}{U_2/r_2}.$$

En utilisant `R`, trouvez s et t tels que

$$P(X \leq s) = .95 \quad \text{et} \quad P(X \leq t) = .99.$$

Solution: la variable aléatoire X suit donc une loi de Fisher avec 5 et 10 degrés de liberté. On peut obtenir les valeurs recherchées à l'aide du code suivant.

```
s = qf(0.95,5,10)
t = qf(0.99,5,10)
```

On s'attend à ce que $s < t$, ce qui est effectivement le cas:

```
s
```

```
## [1] 3.325835
```

```
t
```

```
## [1] 5.636326
```

b) Soient $Z \sim N(0,1)$ et $U \sim \chi^2(10)$ deux variables aléatoires indépendantes. Posons

$$V = \frac{Z}{\sqrt{U/10}}.$$

En utilisant `R`, trouvez w tel que $P(V \leq w) = 0.95$.

Solution: la variable aléatoire Z suit ainsi une loi T de Student avec 10 degrés de liberté. On peut obtenir la valeur recherchée à l'aide du code suivant.

```
w = qt(0.95,10)
w
```

```
## [1] 1.812461
```

Q2

Soient $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$, et $a \in \mathbb{R}$. Définissons $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v} + a$. Trouvez le gradient de f par rapport à $\mathbf{Y}$. Écrivez une fonction `R` qui calcule $f(\mathbf{Y})$ étant donnés $\mathbf{v}, a$. Évaluez la fonction en $\mathbf{Y} = (1, 0, -1)$, pour $\mathbf{v} = (1, 2, -3)$ et $a = -2$.

Note: nous écrirons les vecteurs soit dans un format colonne ou dans un format ligne, de façon plus ou moins arbitraire. À vous de déterminer le format qui fait en sorte que les dimensions soient compatibles (c'est vrai pour l'ensemble du cours).

Solution: le gradient de f par rapport à est

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \nabla_{\mathbf{Y}}(\mathbf{Y}^\top \mathbf{v} + a) = \nabla_{\mathbf{Y}}(\mathbf{Y}^\top \mathbf{v}) + \nabla_{\mathbf{Y}}(a) = \mathbf{v} + \mathbf{0} = \mathbf{v}.$$

Voici un bloc de code qui évalue la fonction f.

```
ma.fonction <- function(Y,v,a){
  sum(Y*v)+a
}
```

On l'essaie:

```
ma.fonction(Y=c(1,0,-1),v=c(1,2,-3),a=-2)
```

```
## [1] 2
```

C'est effectivement la valeur de $\mathbf{Y}^\top \mathbf{v} + a = (1, 0, -1) \cdot (1, 2, -3) - 2 = 1 \cdot 1 + 0 \cdot 2 + (-1) \cdot (-3) - 2 = 2$.

Q3

Soient $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$, $\boldsymbol{\mu} = (1, 0, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, et $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Posons $\mathbf{W} = A\mathbf{Y}$. Quelle loi le vecteur aléatoire $\mathbf{W}$ suit-il? Prélevez 100 observations de ce vecteur aléatoire avec `R` et placez les dans un graphique.

Indice: vous pouvez utiliser la fonction `mvrnorm()` provenant de la librairie `MASS`.

Solution: si $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, alors

$$\mathbf{W} = A\mathbf{Y} \sim \mathcal{N}(A\boldsymbol{\mu}, A\boldsymbol{\Sigma}A^\top).$$

Mais

$$A\boldsymbol{\mu} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

et

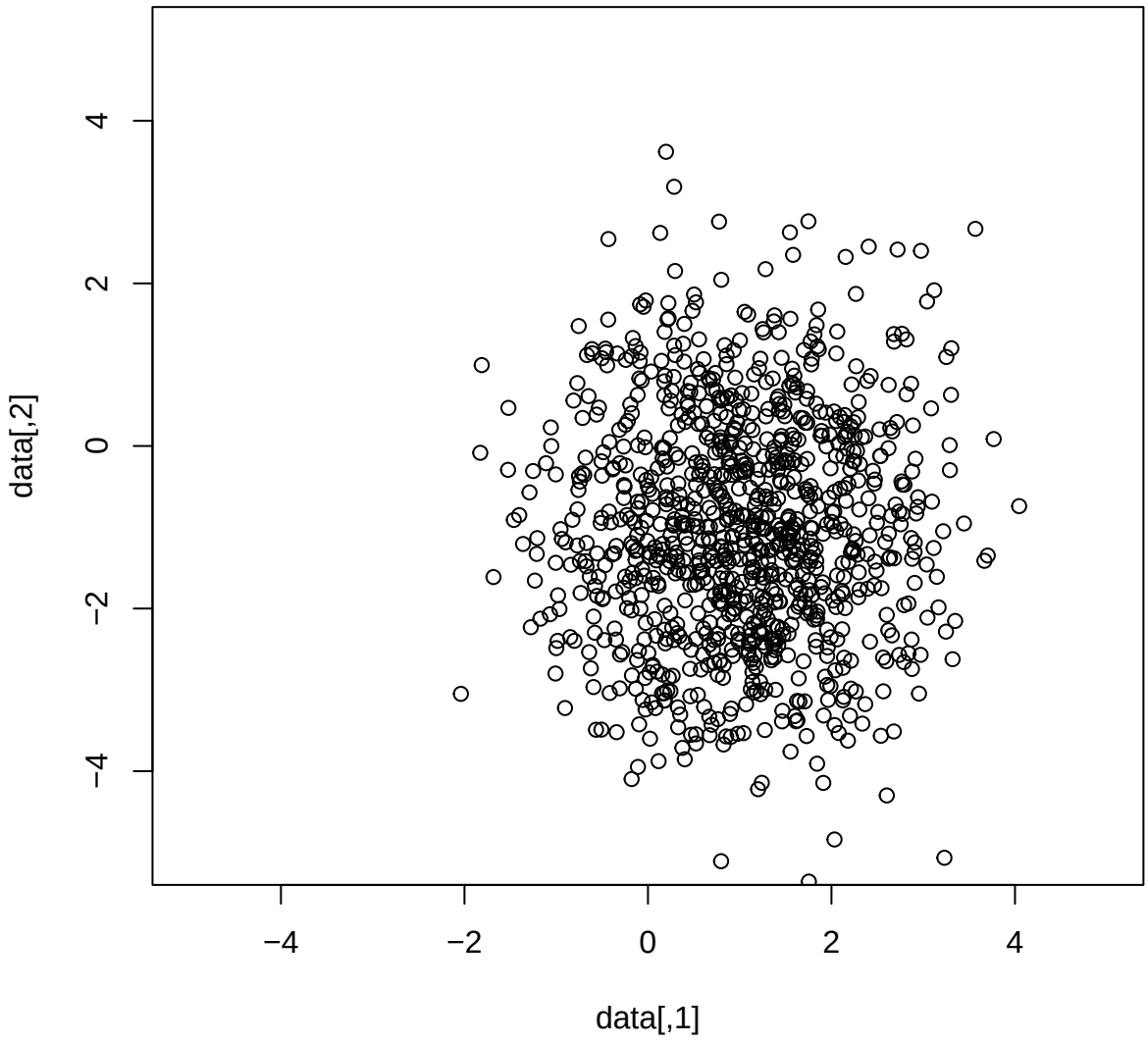
$$A\boldsymbol{\Sigma}A^\top = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

On store ces matrices dans `R`:

```r
mu.W <- matrix(c(1,-1),2,1)
Sigma.W <- matrix(c(1,0,0,2),2,2)
```

On peut utiliser la fonction `mvnorm()` de la librairie `MASS` afin de prélever un échantillon de vecteurs aléatoires $\mathbf{W}$ de taille $n = 1000$ (je sais que j'ai dit $n = 100$, mais cela fonctionne pour n'importe quelle taille n). Vos échantillons peuvent être différents, bien sûr.

```r
set.seed(0)
data = MASS::mvrnorm(n = 1000, mu.W, Sigma.W)
plot(data, xlim=c(-5,5), ylim=c(-5,5))
```



On vérifie que l'échantillon a bien les caractéristiques escomptées:

```
mean(data[,1])
```

```
## [1] 1.024786
```

```
mean(data[,2])
```

```
## [1] -1.022386
```

```
var(data[,1])
```

```
## [1] 1.068649
```

```
var(data[,2])
```

```
## [1] 1.992027
```

```
cov(data[,1],data[,2])
```

```
## [1] 0.01841607
```

Q4

Soit $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_4)$. Posons $\overline{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$. En utilisant `R`, prélevez 1000 observations des variables aléatoires suivantes:

a) $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$

b) $4\overline{Y}^2$

c) $(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2$

Dans chacun des cas, tracez un histogramme des observations.

Solution: nous avons $n = 4$ et $\sigma^2 = 9$. Par hypothèse, les variables aléatoires Y_1, Y_2, Y_3, Y_4 sont indépendantes, mais ce n'est pas la même chose que de dire que $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$, $4\overline{Y}^2$, et $(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2$ le sont également.

Cependant, on observe que a) correspond à $Q_A(\mathbf{Y})$, b) à $Q_B(\mathbf{Y})$, et c) à $Q_C(\mathbf{Y})$. Selon le théorème de Cochran, a), b), et c) sont donc indépendantes, et

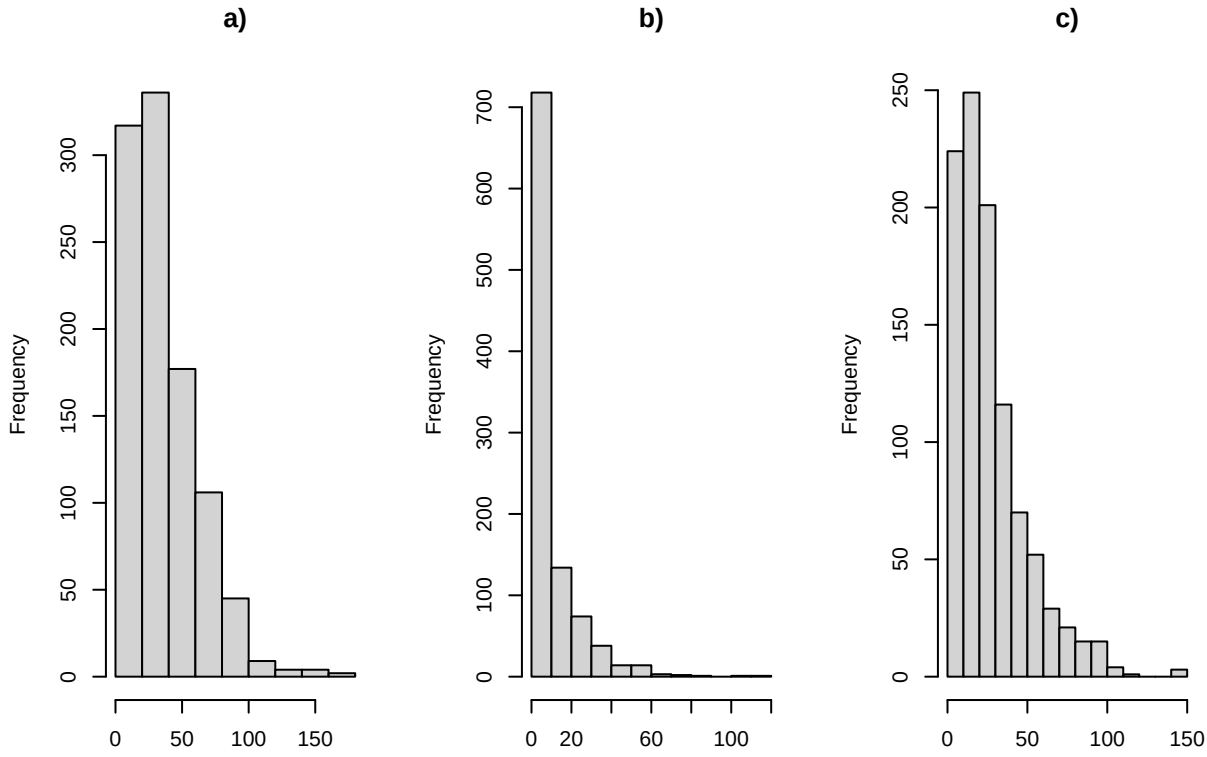
$$\frac{Q_A(\mathbf{Y})}{\sigma^2} = \frac{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2}{9} \sim \chi^2(4), \quad \frac{Q_B(\mathbf{Y})}{\sigma^2} = \frac{4\overline{Y}^2}{9} \sim \chi^2(1),$$

et

$$\frac{Q_C(\mathbf{Y})}{\sigma^2} = \frac{(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2}{9} \sim \chi^2(4 - 1 = 3)$$

Nous pouvons ainsi prélever 1000 observations chacunes à partir des lois $\chi^2(4), \chi^2(1), \chi^2(3)$, multiplier les éechantillons obtenus par $\sigma^2 = 9$, et tracer les histogrammes.

```r
set.seed(0)
par(mfrow = c(1,3))
hist(9*rchisq(1000,4),main="a)", xlab="")
hist(9*rchisq(1000,1),main="b)", xlab="")
hist(9*rchisq(1000,3),main="c)", xlab="")
```



Q5

Considérons la fonction $f : \mathbb{R}^3 \to \mathbb{R}$ définie par

$$f(\mathbf{Y}) = Y_1^2 + \tfrac{1}{2}Y_2^2 + \tfrac{1}{2}Y_3^2 - Y_1Y_2 + Y_1 + 2Y_2 - 3Y_3 - 2.$$

En utilisant `R`, trouvez le(s) point(s) critique(s) de f. Si c'est un point critique unique, donne-t-il naissance à un maximum global de f? Un minimum global? Un col?

Solution: on ré-écrit

$$f(\mathbf{Y}) = \frac{1}{2}\begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} - \begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix}\begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix} - 2.$$

Les points critiques de f sont ceux pour lesquels $\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \mathbf{0}$. Mais

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} - \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix},$$

d'où le point critique recherché résoud

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix} \Longrightarrow \mathbf{Y}^* = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix}.$$

C'est effectivement une matrice inversible, puisque son déterminant est non-nul:

```
A=matrix(c(2,-1,0,-1,1,0,0,0,1), nrow=3, ncol=3)
det(A)
```

```
## [1] 1
```

On calcule l'inverse (la fonction `inv()` se retrouve dans la librairie `matlib`) et le produit matriciel à l'aide de `R`.

```
v=matrix(c(-1,-2,3), nrow=3, ncol=1)
Y0 = matlib::inv(A)%*%v
Y0
```

```
##      [,1]
## [1,]   -3
## [2,]   -5
## [3,]    3
```

On détermine la nature du point critique en calculant les valeurs propres de la matrice.

```
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 2.618034 1.000000 0.381966
##
## $vectors
##            [,1] [,2]      [,3]
## [1,]  0.8506508    0 0.5257311
## [2,] -0.5257311    0 0.8506508
## [3,]  0.0000000    1 0.0000000
```

Puisqu'elles sont toutes positives, $\mathbf{Y}^*$ correspond donc à un **minimum global**.

On peut essayer de se convaincre que c'est bien le cas en évaluant la fonction f à un paquet de points $\mathbf{Y}$ et en confirmant que les valeurs de f sont toute plus élevées que $f(\mathbf{Y}^*)$.

Voici un bloc de code qui implémente f en `R`, ainsi que la valeur de $f(\mathbf{Y}^*)$

```
ma.func <- function(Y1,Y2,Y3){
  Y1^2+1/2*Y2^2+1/2*Y3^2-Y1*Y2+Y1+2*Y2-3*Y3-2
}
ma.func(Y0[1],Y0[2],Y0[3])
```

```
## [1] -13
```

On choisit $n = 1000$ vecteurs $\mathbf{Z} = (Z_1, Z_2, Z_3)$ au hasard dans le cube $[-10, 10]^3$, et on constate que la plus petite valeur de $f(\mathbf{Z})$ dans l'ensemble est effectivement plus grande que $f(\mathbf{Y}^*) = -13$.

```
set.seed(0)     # replication
X1 = runif(1000,-10,10)
X2 = runif(1000,-10,10)
X3 = runif(1000,-10,10)

x=c()

for(j in 1:1000){
  x[j]=ma.func(X1[j],X2[j],X3[j])
}

min(x)
```

```
## [1] -12.98493
```

Ce n'est pas une preuve, bien sûr (la démonstration, c'est ce qui se retrouve un peu plus haut), mais c'est au moins compatible avec notre résultat.

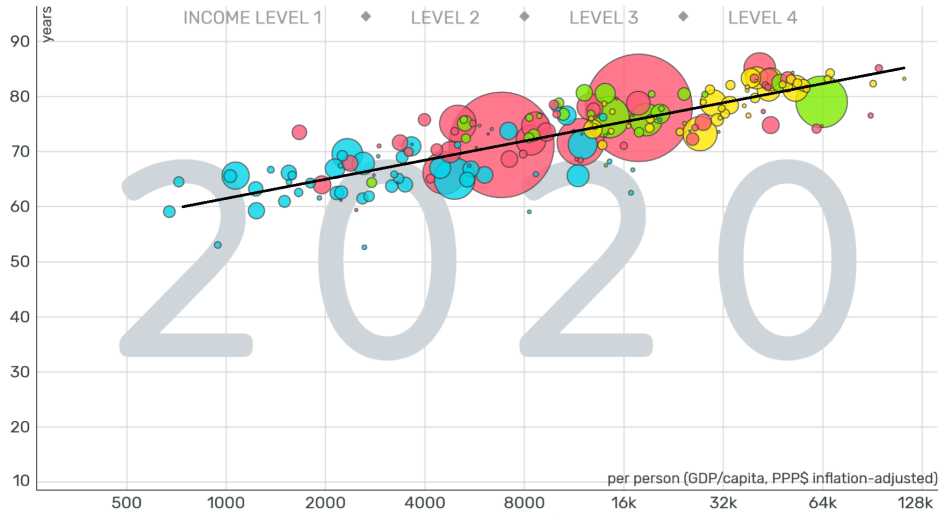
Q6

a) Identifiez la variable réponse Y et la variable prédicteur X dans chacuns des exemples présentés aux pages 4 et 5 des notes de cours (chapitre 2). Y a-t-il une relation linéaire entre X et Y. À l'oeil, tracez la droite d'ajustement linéaire approximative (et donnez son équation).

Indice: servez vous de captures d'écran et d'un logiciel tel que Paint, PowerPoint, ou GIMP pour superposer la droite.

Solution: dans le premier cas, la variable réponse Y est l'espérance de vie des pays de la planète en 2020, tandis que la variable prédicteur X est le revenu par personne (en dollars ajustés pour l'inflation) de ces mêmes pays.

La relation semble linéaire, mais attention! ... l'échelle du prédicteur est logarithmique: il y a donc une relation linéaire entre Y et $\log_2(X)$.



Les points $(\log_2(2000), 65)$ et $(\log_2(32000), 79)$ se retrouvent sur la droite de pente et d'ordonnée à l'origine

$$m = \frac{79 - 65}{\log_2(32000) - \log_2(2000)} \quad \text{et} \quad b = 79 - m \log_2(32000) :$$

```
m = (79-65)/(log2(32000)-log2(2000))
b = 79-m*log2(32000)
m
```

```
## [1] 3.5
```

```
b
```

```
## [1] 26.61976
```

L'équation de la "droite" est ainsi

$$Y = 3.5 \log_2(X) + 26.62.$$

On vérifie que c'est raisonnable: si $X = 8000$, nous avons

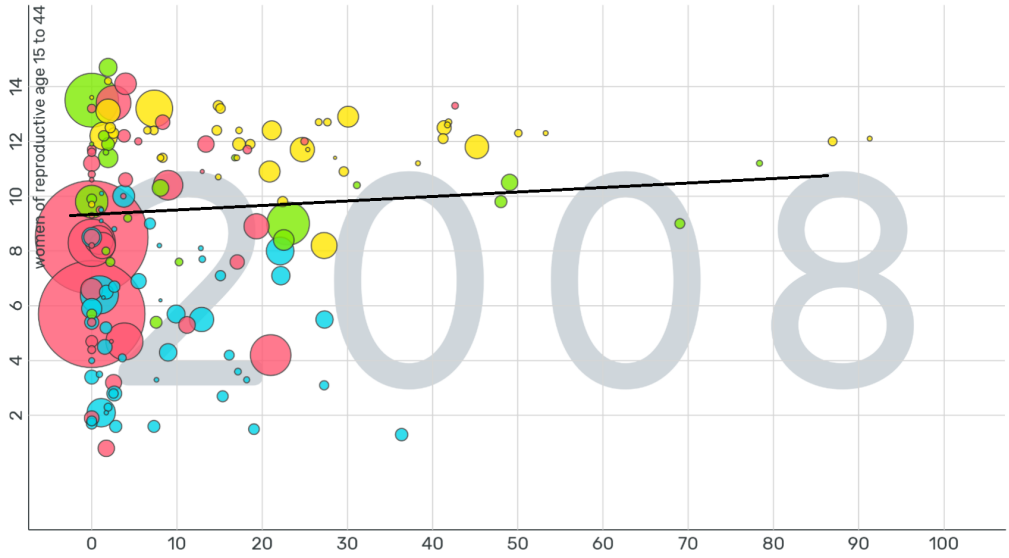
$$Y = 3.5 \log_2(8000) + 26.62 :$$

```
m*log2(8000)+26.62
```

```
## [1] 72.00024
```

ce qui concorde ma foi assez bien avec le graphique.

Dans le deuxième cas, la réponse Y est la longueur moyenne de la scolarité par pays en 2008, et le prédicteur X est l'index de démocratie directe par pays en 2008. Il n'y a pas vraiment de relation entre X et Y (linéaire ou non).



J'ai tracé une droite, mais je ne peux pas juger de sa qualité ... je ne sais même pas si la pente devrait être positive ou négative; c'est un exercice en futilité d'essayer de calculer l'équation dans ce cas et on laisse tout simplement tomber (on pourrait le faire exactement en se servant de `R`... si on avait l'ensemble de données à notre disposition).

b) Considérez les 4 exemples présentés à la page 9 des notes de cours (chapitre 2). La variance de l'erreur est-elle constante? Les termes d'erreurs sont-ils indépendants les uns des autres?

Solution: la variance de l'**erreur** ε_i (en supposant un modèle linéaire) est constante en haut à gauche, plus ou moins constante en haut à droite, mais pas constante en bas. Les termes d'erreurs semblent indépendants en haut, mais non en bas.

Q7

Considérons l'ensemble de données `Autos.xlsx` se retrouvant sur Brightspace. Le prédicteur est `VKM.q` (X, distance quotidienne moyenne, en km); la réponse est `CC.q` (Y, consommation de carburant quotidienne moyenne, en L). Utilisez `R` afin de:

a) tracer le nuage de points de Y en fonction de X;

b) déterminer le nombre d'observations n;

c) calculer les quantités $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_i Y_i$, $\sum Y_i^2$;

d) déterminer les équations normales de la droite d'ajustement;

e) déterminer les coefficients de la droite d'ajustement (sans utiliser la fonction `lm()`), et

f) superposer la droite d'ajustement sur le nuage de point.

Solution: on doit commencer par charger l'ensemble de données. On peut soit convertir le fichier `.xlsx` en fichier `.csv`, ou utiliser la fonction `read_excel()` de la librairie `readxl`, ou utiliser une quelconque autre méthode.

```
Autos <- readxl::read_excel("Autos.xlsx")
str(Autos)
```

```
## tibble [996 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Type : chr [1:996] "PUPC" "PUPC" "PUPC" "PUPC" ...
##  $ Age  : num [1:996] 0 1 10 1 3 5 9 6 3 9 ...
##  $ Rural: num [1:996] 0 0 0 1 1 1 0 0 0 0 ...
##  $ VKM.q: num [1:996] 330 264 251 235 230 230 215 208 203 196 ...
##  $ CC.q : num [1:996] 49 33 44 22 38 31 28 19 31 19 ...
```

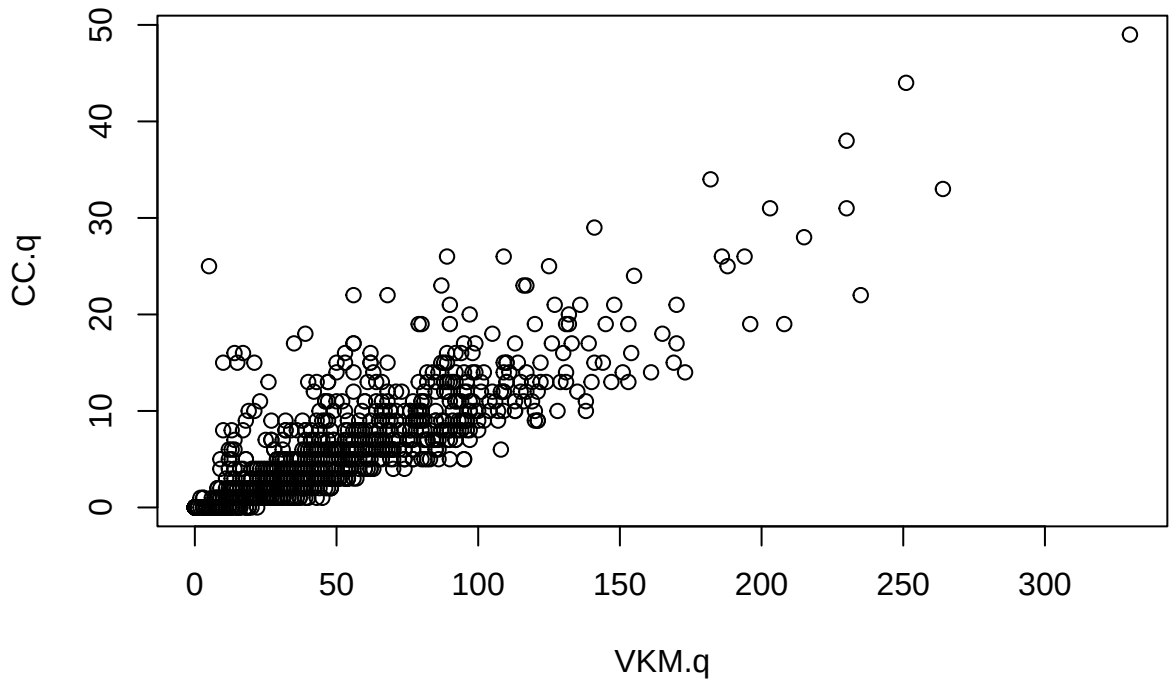
Ensuite, on ne garde que le prédicteur X et la réponse Y.

```
library(tidyverse)  # pour avoir acces a select() et |>
Autos = Autos |> select(VKM.q,CC.q)
str(Autos)
```

```
## tibble [996 x 2] (S3: tbl_df/tbl/data.frame)
##  $ VKM.q: num [1:996] 330 264 251 235 230 230 215 208 203 196 ...
##  $ CC.q : num [1:996] 49 33 44 22 38 31 28 19 31 19 ...
```

a) On trace le nuage de point avec le code suivant.

```
plot(Autos)
```



La relation semble bien au moins un peu linéaire.

b) On peut déterminer le nombre d'observations n de plusieurs façons, comme, par exemple:

```
n = nrow(Autos)
n
```

```
## [1] 996
```

c) On calcule les quantités demandées:

```
X = Autos$VKM.q
Y = Autos$CC.q

(somme.X = sum(X))
```

```
## [1] 48173
```

```
(somme.Y = sum(Y))
```

```
## [1] 5766
```

```
(somme.X2 = sum(X^2))
```

```
## [1] 4100349
```

```
(somme.XY = sum(X*Y))
```

```
## [1] 495119
```

```
(somme.Y2 = sum(Y^2))
```

```
## [1] 70208
```

d) Il y a plusieurs façons d'exprimer les équations normales. Sous forme matricielle, par exemple, on peut retrouver

$$\begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}.$$

Avec les valeurs calculées au préalable, nous obtenons ainsi

$$\begin{pmatrix} 996 & 48173 \\ 48173 & 4100349 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 5766 \\ 495119 \end{pmatrix}.$$

e) On obtient les estimateurs b_0, b_1 des coefficients en résolvant les équations normales (sans utiliser `lm()`, comme la question le demande):

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum X_i Y_i - n\overline{XY}}{\sum X_i^2 - n\overline{X}^2} \quad \text{et} \quad b_0 = \overline{Y} - b_1 \overline{X}.$$

```
Sxy = somme.XY-n*mean(X)*mean(Y)
Sxx = somme.X2-n*(mean(X))^2
(b1 = Sxy/Sxx)
```

```
## [1] 0.1221413
```

```
(b0 = mean(Y)-b1*mean(X))
```

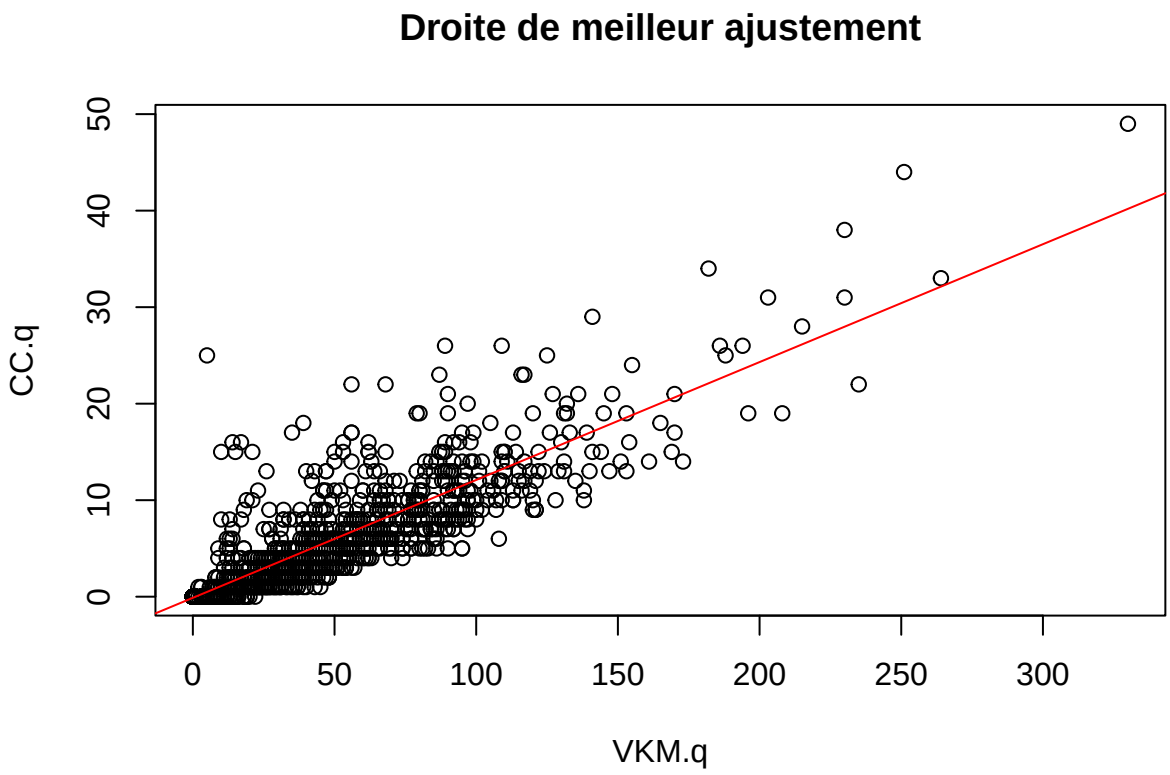
```
## [1] -0.1183883
```

L'équation de la droite de meilleur ajustement est ainsi

$$Y = -0.1183883 + 0.1221314X.$$

f) Voici le tracé de la droite superposé sur le nuage de points.

```
plot(Autos, main="Droite de meilleur ajustement")
abline(c(b0,b1), col="red")
```



La droite passe le test du "pif", mais on ne peut pas nécessairement interpréter les coefficients comme on l'aimerait: si $X = 0$ (aucune distance parcourue, quotidiennement), nous obtenons $Y = -0.1184$ (une quantité **négative** de carburant consommé).

Q8

Utilisez la fonction `lm()` de `R` afin d'obtenir les coefficients de la droite d'ajustement et les résidus. Montrez (en calculant les quantités requises directement) que les 5 premières propriétés des résidus (p.25 dans les notes de cours du chapitre 2) sont satisfaites.

Solution: on voit facilement que la droite obtenue au problème précédent est la bonne.

```
mod = lm(Y ~ X)
mod$coefficients
```

```
## (Intercept)           X
##  -0.1183883   0.1221413
```

On peut aussi aller chercher les résidus et les valeurs ajustées:

```
e = mod$residuals
Y.hat = mod$fitted.values
```

On se sert de X_i, Y_i, $\hat{Y}_i$ et e_i pour montrer que les 5 propriétés des résidus sont valides pour l'ajustement:

a) $\overline{e} = 0$

```
mean(e)
```

```
## [1] 2.215574e-15
```

b) $\overline{Y} = \overline{\hat{Y}}$

```
mean(Y)
```

```
## [1] 5.789157
```

```
mean(Y.hat)
```

```
## [1] 5.789157
```

c) $\sum X_i e_i = 0$

```
sum(X*e)
```

```
## [1] 1.024109e-10
```

d) $\sum \hat{Y}_i e_i = 0$

```
sum(Y.hat*e)
```

```
## [1] 3.852065e-11
```

e) $(\overline{X}, \overline{Y})$ se retrouve sur la droite d'ajustement

```
mean(Y)
```

```
## [1] 5.789157
```

```
b0+b1*mean(X)
```

```
## [1] 5.789157
```

Q9

À l'aide de `R`, calculez les coefficients de corrélation de Pearson et de Spearman entre le prédicteur et la réponse. Y a-t-il une association linéaire forte ou faible entre ces deux variables? Servez-vous des corrélations afin de justifier votre réponse.

Solution: on peut calculer la corrélation de Pearson directement, ou encore utiliser la fonction `cor()`.

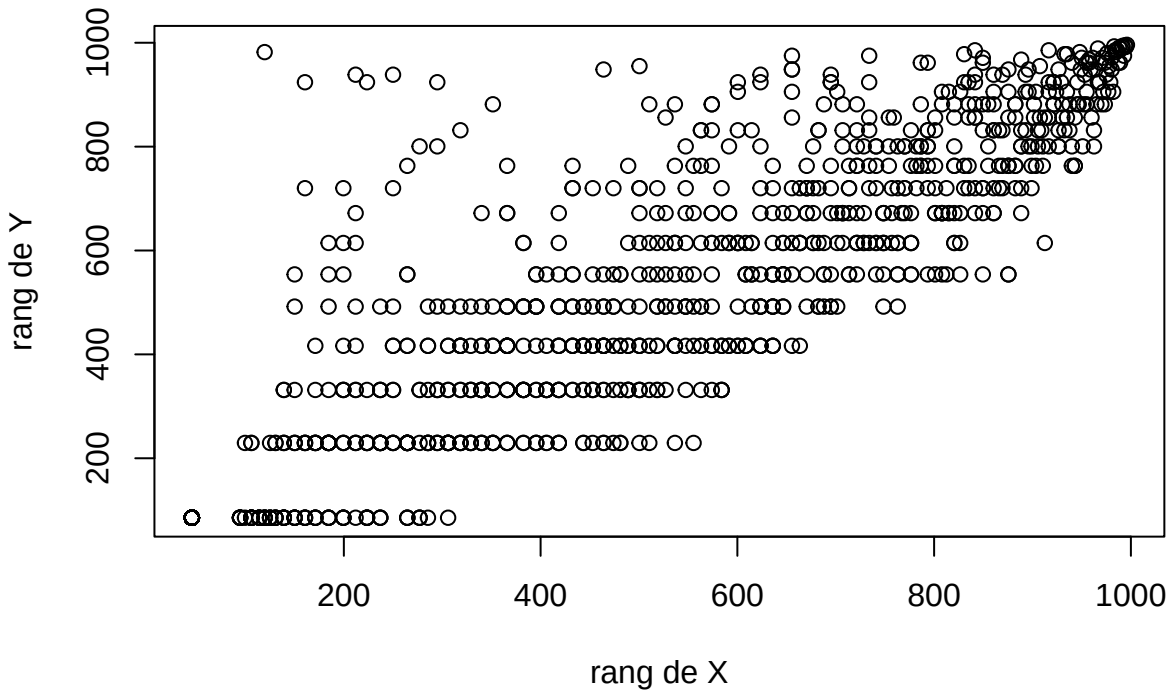
```
(r = cor(X,Y))
```

```
## [1] 0.8468566
```

Pour calculer la corrélation de Spearman, on s'y prend comme suit. On obtient les rangs de X et Y à l'aide de:

```
rX = rank(X)
rY = rank(Y)

plot(rX,rY, xlab="rang de X", ylab="rang de Y")
```



La corrélation de Spearman est la corrélation de Pearson des rangs:

```
(r.S = cor(rX,rY))
```

```
## [1] 0.8713188
```

On peut aussi l'obtenir avec:

```
cor.test(X,Y, method="spearman")
```

```
## Warning in cor.test.default(X, Y, method = "spearman"): Cannot compute exact p-
## value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  X and Y
## S = 21190504, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8713188
```

On ne peut pas facilement juger de la force de la relation et de la linéarité entre X et Y seulement par l'entremise de ces corrélations, quoique les valeurs $r_S = 0.87$ et $r = 0.85$ semblent toutes deux suggérer qu'une relation linéaire n'est pas hors de question; c'est le nuage de point qui vient clore le débat en faveur d'une linéarité presque certaine.

On peut aussi aborder le problème sous un autre angle: toutes les voitures n'ont pas la même facteur de conversion entre la distance parcourue et la consommation de carburant (surtout que la vitesse et autres habitudes de conduite peuvent venir influencer les données), mais en général, on pourrait s'attendre à ce que la relation soit linéaire.

Q10

À l'aide de `R`, déterminez la décomposition en sommes de carrés de la régression.

Solution: la décomposition en somme de carrés est

$$\text{SST} = \text{SSR} + \text{SSE},$$

où $\text{SST} = S_y y$, $\text{SSR} = b_1^2 S_x x$ et $\text{SSE} = \sum e_i^2$.

On a alors:

```
(SST = somme.Y2-n*(mean(Y))^2)
```

```
## [1] 36827.72
```

```
(SSR = b1^2*Sxx)
```

```
## [1] 26411.59
```

```
(SSE = sum(e^2))
```

```
## [1] 10416.13
```

On voit ainsi que

$$36827.72 = 26411.59 + 10416.13 :$$

```
SSR+SSE
```

```
## [1] 36827.72
```