Homework 1 - Solutions

Patrick Boily

2023-05-25

Q1

a) Let $U_i \sim \chi^2(r_i)$ be independent random variables with $r_1 = 5$, $r_2 = 10$. Set

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

Using `R`, find s and t such that

$$P(X \leq s) = 0.95 \quad \text{and} \quad P(X \leq t) = 0.99.$$

Solution: the random variable X follows a Fisher distribution with 5 and 10 degrees of freedom. We can obtain the values we are looking for using the following code:

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

We expect $s < t$, which is indeed the case:

```
s
```

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

```
t
```

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

b) Let $Z \sim N(0,1)$ and $U \sim \chi^2(10)$ be two independent random variables. Let

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

Using `R`, find w such that $P(V \leq w) = 0.95$.

Solution: the random variable Z thus follows a Student's T distribution with 10 degrees of freedom. We can obtain the value we are looking for with the following code:

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

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

Q2

Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$, and $a \in \mathbb{R}$. Define $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v} + a$. Find the gradient of f with respect to $\mathbf{Y}$. Write a function in `R` that computes $f(\mathbf{Y})$ given $\mathbf{v}, a$. Evaluate the function at $\mathbf{Y} = (1, 0, -1)$, for $\mathbf{v} = (1, 2, -3)$ and $a = -2$. **Note:** in the course, we will write vectors either as columns format or as rows, in a more or less arbitrary way. It is up to you to determine which one makes the dimensions compatible.

Solution: the gradient of f with respect to $\mathbf{Y}$ is

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

Here is a code block that evaluates the function f function.

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

Let's try it:

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

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

This is indeed the value of $\mathbf{Y}^\top \mathbf{v} + a = (1, 0, -1) \cdot (1, 2, -3) - 2 = 1 \cdot 1 + 0 \cdot 2 + (-1) \cdot (-3) - 2 = 2$.

Q3

Let $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$, $\boldsymbol{\mu} = (1, 0, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Let $\mathbf{W} = A\mathbf{Y}$. What distribution does the random vector $\mathbf{W}$ follow? Draw a sample of size 100 for this random vector with `R` and plot them in a graph.

Note: you may use the function `mvrnorm()` from the `MASS` package to help along (but you do not have to).

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

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

But

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

and

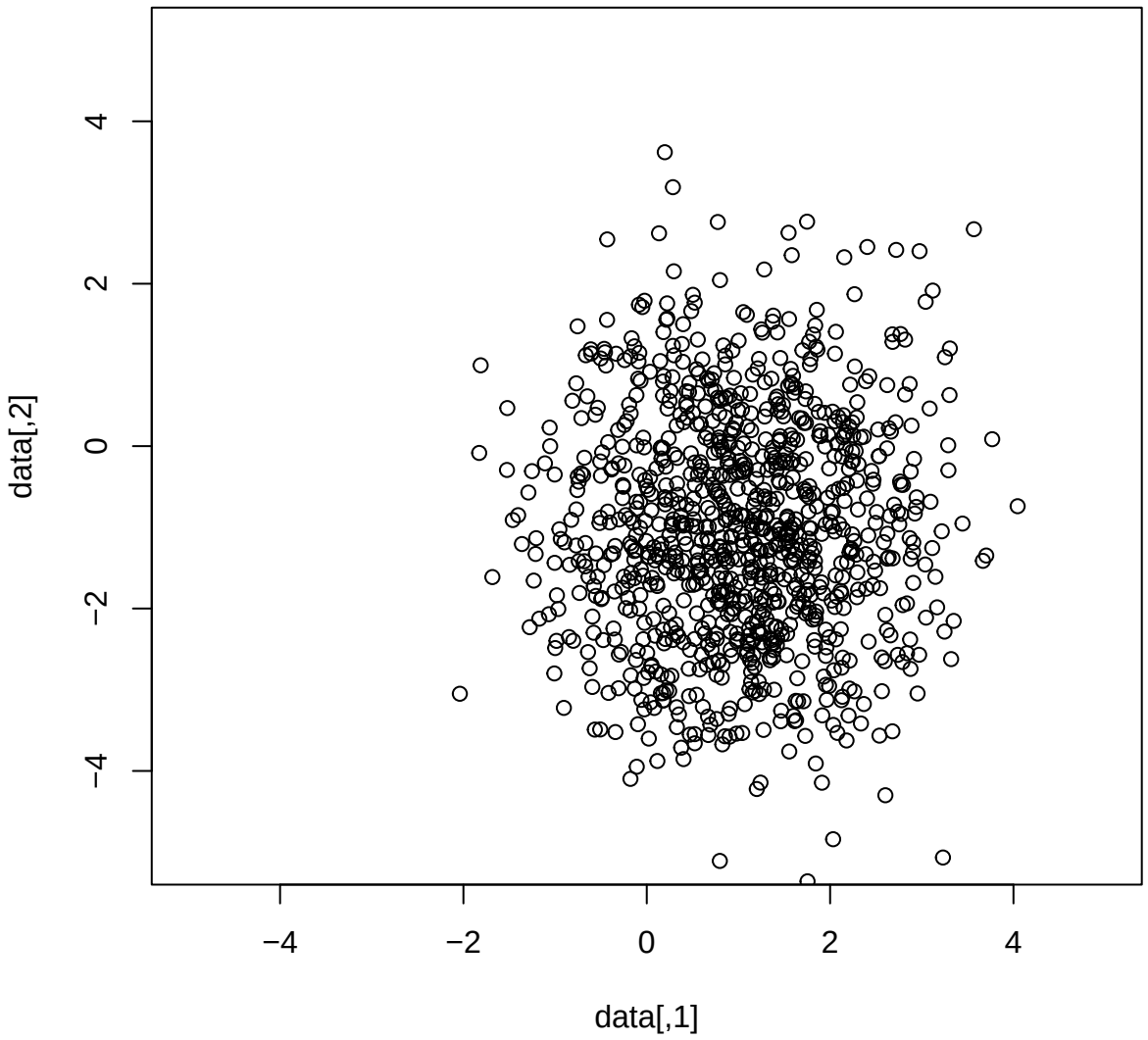
$$A\boldsymbol{\Sigma}A^\top = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0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

We store these matrices in `R`:

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

We can use the `MASS` package's `mvnorm()` function to draw a sample of random vectors $\mathbf{W}$ of size $n = 1000$ (I know I said $n = 100$, but it works for any size n). Your samples can be different depending on the seed, of course.

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



We verify that the sample has the expected characteristics:

```
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

Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_4)$. and set $\overline{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$. Using `R`, draw 1000 observations from the following random variables:

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

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

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

In each case, plot a histogram of the observations.

Solution: we have $n = 4$ and $sigma^2 = 9$. By assumption, the random variables Y_1, Y_2, Y_3, Y_4 are independent, but that is not the same thing as saying that $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$, $4\overline{Y}^2$, and $(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2$ are independent.

However, it can be seen that a) matches with $Q_A(\mathbf{Y})$, b) with $Q_B(\mathbf{Y})$, and c) with $Q_C(\mathbf{Y})$. According to Cochran's theorem, a), b), and c) are therefore independent, and

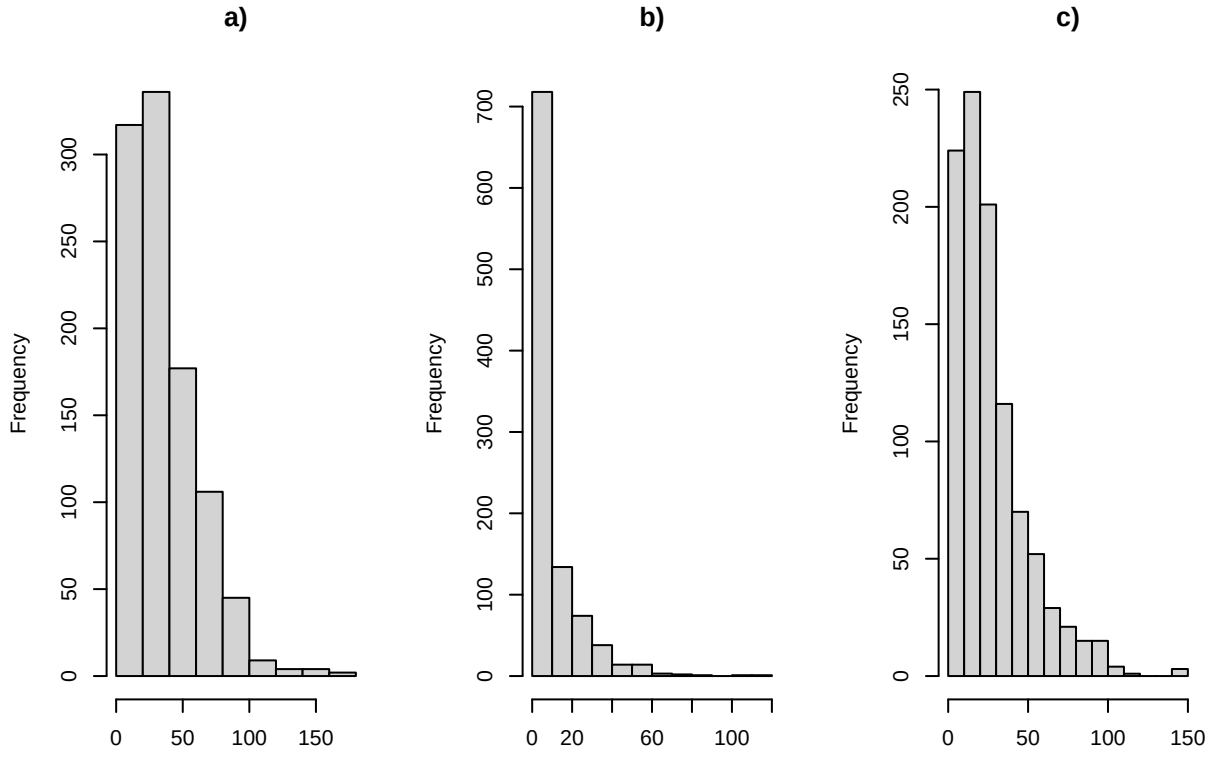
$$\frac{Q_A(\mathbf{Y})}{\sigma^2} = \frac{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2}{9} \sim \chi^2(4), \quad \frac{Q_B(\mathbf{Y})}{\sigma^2} = \frac{4\overline{Y}^2}{9} \sim \chi^2(1),$$

and

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

We can thus draw 1000 observations each from the distributions $\chi^2(4), \chi^2(1), \chi^2(3)$, multiply the samples obtained by $\sigma^2 = 9$, and plot the histograms.

```
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

Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

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

Using R, find the critical point(s) of f. If it is unique, does it give rise to a global maximum of f? A global minimum? A saddle point?

Solution: re-write

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

The critical points of f are those for which $\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \mathbf{0}$. But

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

from which we conclude that the critical point sought solves

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

The matrix is indeed invertible, since its determinant is non-zero:

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

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

We compute the inverse (the function `inv()` can be found in the `matlib` library) and the matrix product using R (that is not the only way to do this, however... `solve()` works as well).

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

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

The nature of the critical point is determined by computing the eigenvalues of the matrix.

```
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
```

Since they are all positive, $\mathbf{Y}^*$ corresponds to a **global minimum**.

We can convince ourselves that this likely the case by evaluating the function f at a bunch of points $\mathbf{Y}$ and confirming that the values of f are all higher than $f(\mathbf{Y}^*)$.

Here is a block of code that implements f in `R`, along with the value of $f(\mathbf{Y}^*)$.

```
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
```

We choose $n = 1000$ vectors $\mathbf{Z} = (Z_1, Z_2, Z_3)$ at random in the cube $[-10, 10]^3$, and we find that the smallest value of $f(\mathbf{Z})$ in the set is indeed larger than $f(\mathbf{Y}^*) = -13$.

```
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
```

This is not a proof, of course (the demonstration is found above), but it is at least compatible with our result.

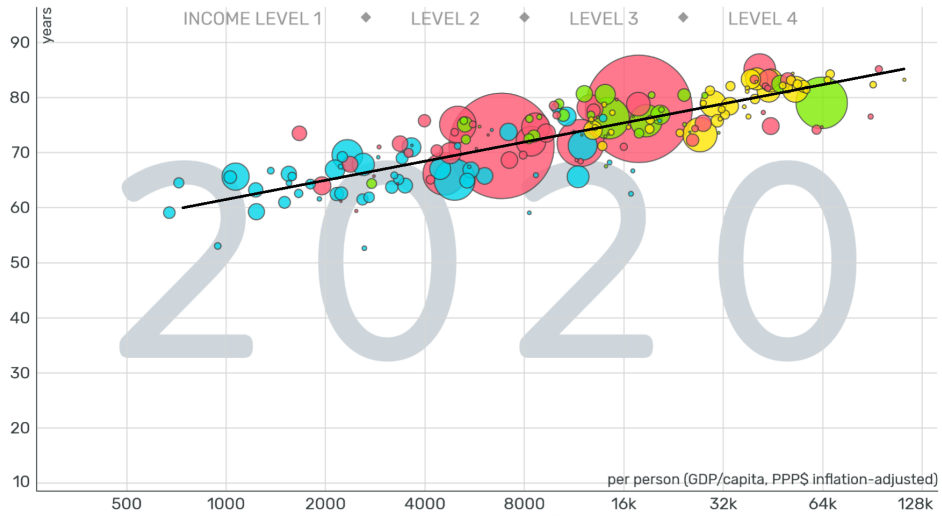
Q6

a) Identify the response variable Y and the predictor variable X in each of the examples shown on slides 4 and 5 of the course notes (Chapter 2). Is there a linear relationship between X and Y. Draw the approximate line of linear fit (and give its equation).

Hint: use screenshots and software (Paint, PowerPoint, GIMP, etc.) to overlay the line.

Solution: in the first case, the response variable Y is the life expectancy of the world's countries in 2020, while the predictor variable X is the per capita income (adjusted for inflation) of these same countries.

The relationship seems linear, but be careful! ... the scale of the predictor is logarithmic, so the linear relationship is between Y and $\log_2(X)$.



The observations $(\log_2(2000), 65)$ and $(\log_2(32000), 79)$ are on the line with slope and intercept

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

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

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

```
b
```

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

The equation of the "line" is thus

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

We verify that this is reasonable: if $X = 8000$, we have

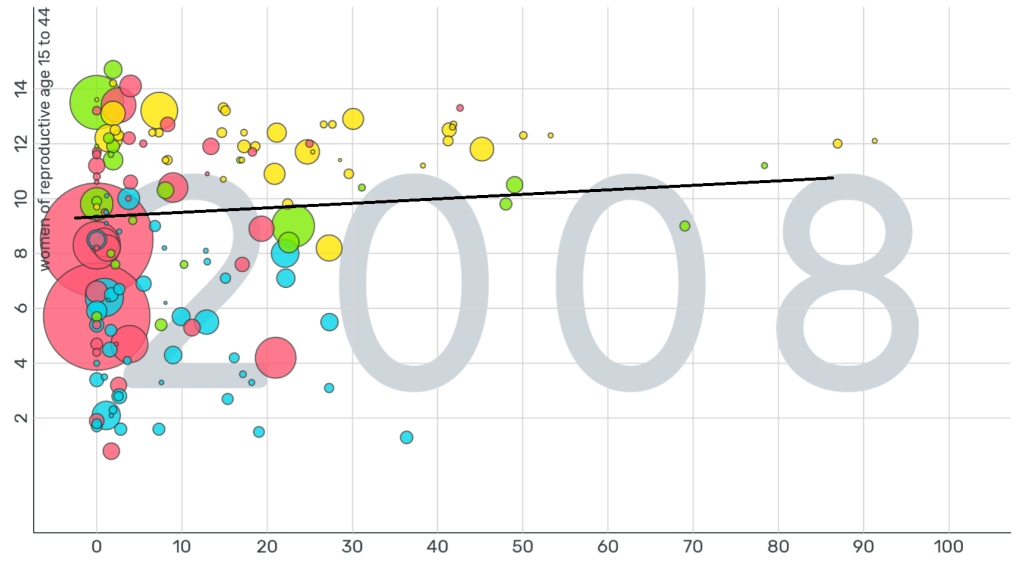
$$Y = 3.5 \log_2(8000) + 26.62 :$$

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

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

which is quite consistent with the graph.

In the second case, the response Y is the average length of schooling by education by country in 2008, and the predictor X is the direct democracy index by country in 2008. There does not appear to be a relationship between X and Y (linear or not).



I drew a line, but I can't assess its quality... I don't even know if the slope should be positive or negative; it is an exercise in futility to try to calculate the equation in this case and we might as well drop it altogether (we could do it using 'R'... if we had the data set at our disposal).

b) Consider the 4 examples shown on page 9 of the course notes (chapter 2). Is the variance of the error terms constant? Are the error terms independent of each other?

Solution: the variance of the **error** ε_i (assuming a linear assuming a linear model) is constant in the top left image, more or less constant in the top right image, but not constant in the bottom images. The error terms seem to be independent at the top, but not at the bottom.

Q7

Consider the dataset `Autos.xlsx` found on Brightspace. The predictor variable is `VKM.q` (X, the average daily distance driven, in km); the response variable is `CC.q` (Y, the average daily fuel consumption, in L). Use `R` to:

a) display the scatterplot of Y versus X;

b) determine the number of observations n in the dataset;

c) compute the quantities $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_i Y_i$, $\sum Y_i^2$;

d) find the normal equations of the line of best fit;

e) find the coefficients of the line of best fit (without using `lm()`), and

f) overlay the line of best fit onto the scatterplot.

Solution: the first step is to load the data set. You can either convert the `.xlsx` file into a `.csv` file, or use the `read_excel()` function from the `readxl` library, or use any other method that does the trick.

```
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 ...
```

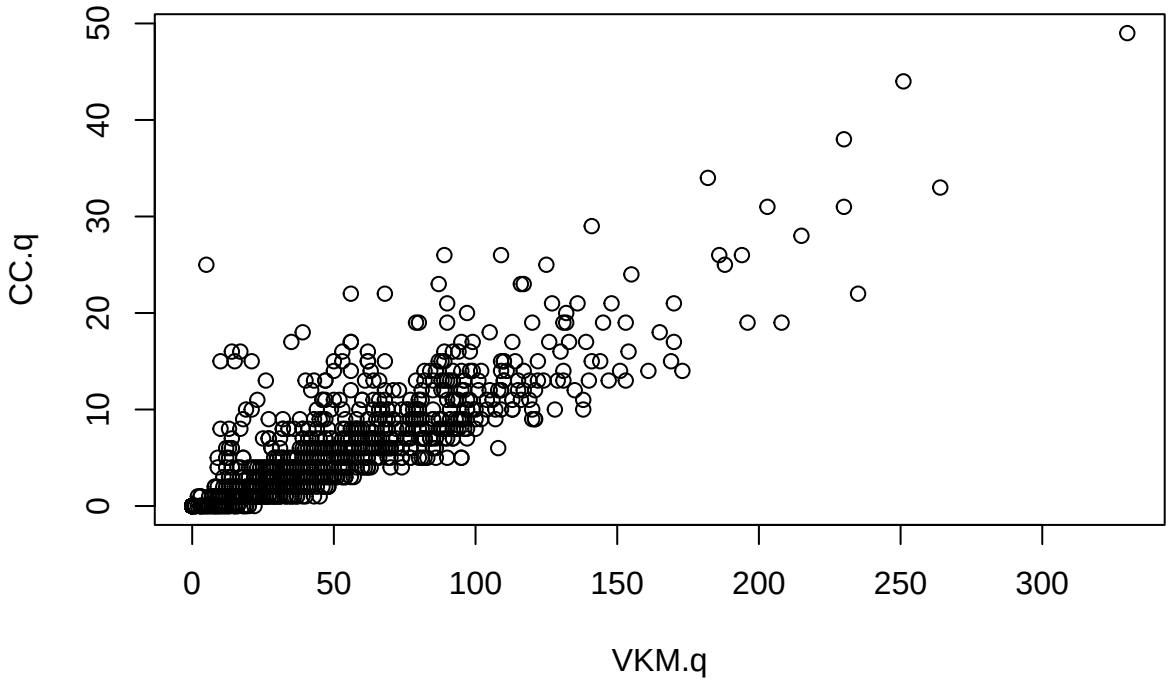
Next, we retain only the predictor X and the response Y.

```
library(tidyverse)  # to access select() and |>
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) We display the scatterplot with the following code.

```
plot(Autos)
```



The relationship seems linear (at least a little).

b) We can find the number of observations n in several ways, such as:

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

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

c) We compute the required quantities:

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

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

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

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

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

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

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

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

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

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

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

d) There are several ways of expressing normal equations. In matrix form, for example, we have

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

Using the previously calculated values, we obtain

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

e) The coefficient estimators b_0, b_1 are obtained by solving the normal equations (without using `lm()`, as required in the question):

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

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

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

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

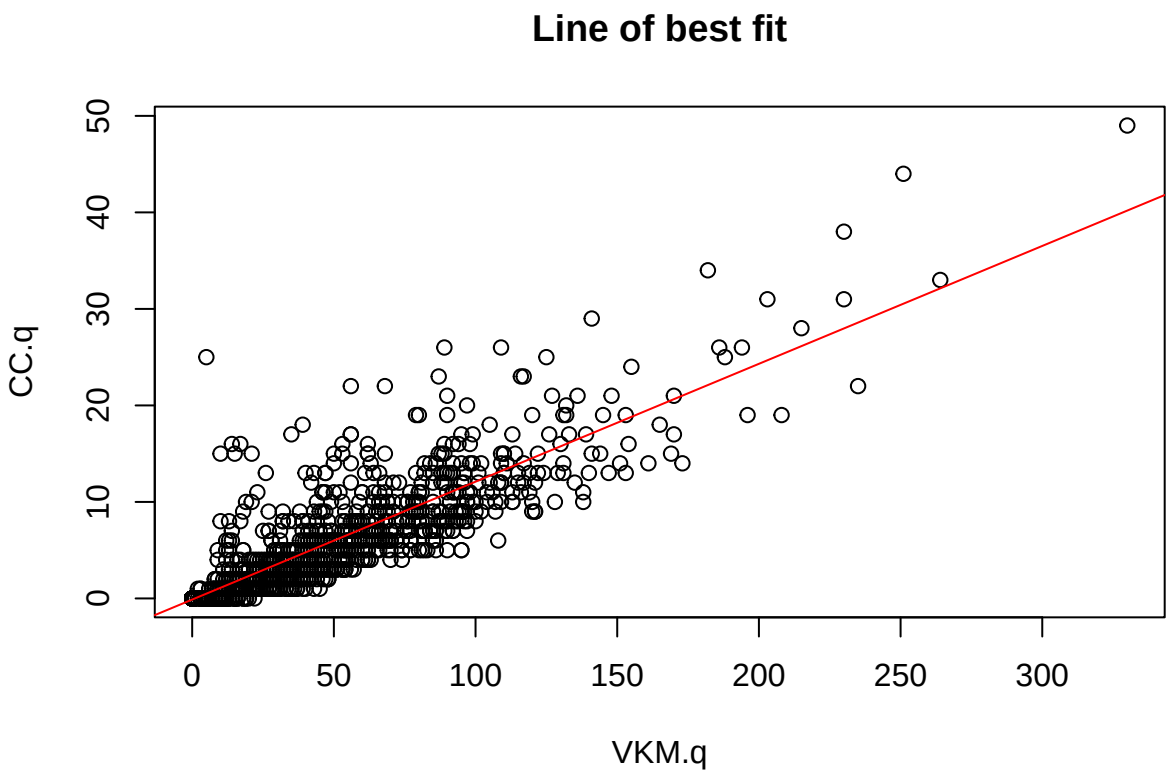
```
## [1] -0.1183883
```

The equation of the OLS line of best fit is thus

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

f) Here's the line superimposed on the scatterplot.

```
plot(Autos, main="Line of best fit")
abline(c(b0,b1), col="red")
```



The line passes the "smell" test, but we can't necessarily interpret the coefficients as we'd like: if $X = 0$ (no daily distance travelled), we get $Y = -0.1184$ (a **negative** amount of fuel consumed), which is patently impossible.

Q8

Use the `R` function `lm()` to obtain the coefficients of the line of best fit and the residuals. Show (by calculating the required quantities directly) that the first 5 properties of residuals (p.@ 25 in the course notes, Chapter 2) are satisfied.

Solution: it's easy to see that the straight line obtained in the previous problem is the right one.

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

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

Residuals and fitted values can also be retrieved:

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

We use X_i, Y_i, $\hat{Y}_i$ and e_i to show that the 5 properties of the residuals are valid for the fit:

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) $(\overline{X}, \overline{Y})$ is on the regression line

```
mean(Y)
```

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

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

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

Q9

Using R, compute the Pearson and Spearman correlation coefficients between the predictor and the response. Is there a strong or weak linear association between these two variables? Use the correlation values and diagrams to justify your answer.

Solution: you can calculate the Pearson correlation directly, or use the `cor()` function.

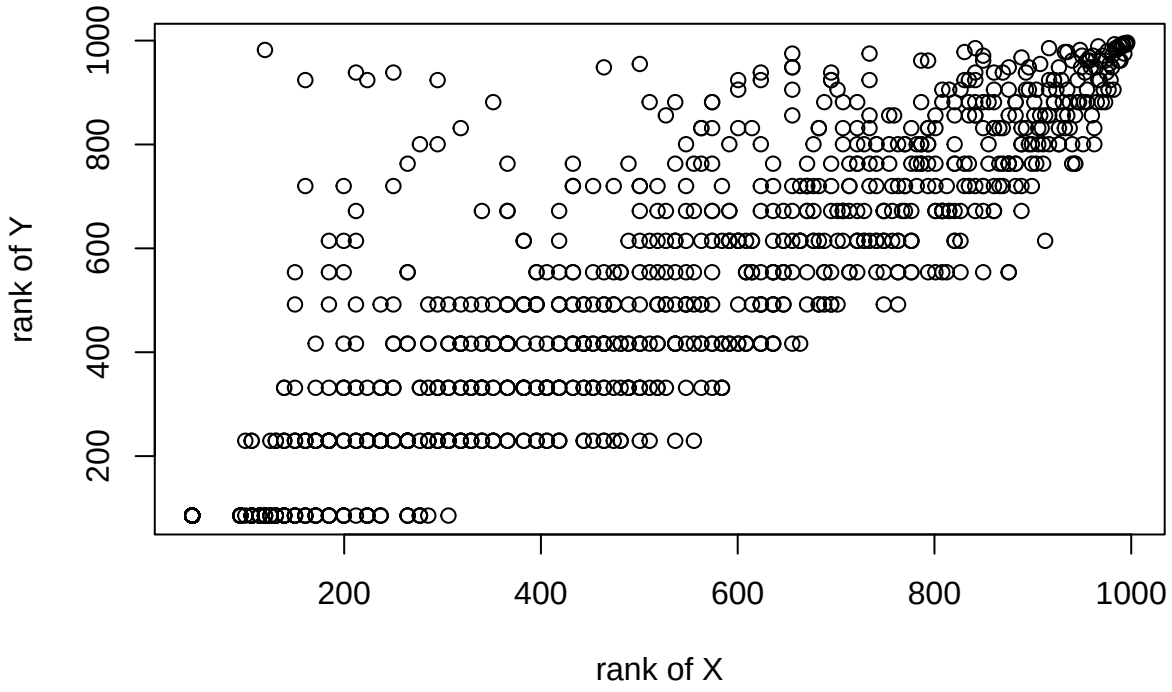
```
(r = cor(X,Y))
```

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

To calculate the Spearman correlation, proceed as follows. The ranks of X and Y are obtained using:

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

plot(rX,rY, xlab="rank of X", ylab="rank of Y")
```



Spearman's correlation is Pearson's correlation for the ranks:

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

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

It can also be obtained *via*:

```
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
```

We can't easily gauge the strength of the relationship and the linearity between X and Y through these correlations alone, although the values $r_S = 0.87$ and $r = 0.85$ both seem to suggest that a linear relationship isn't out of the question; it's the scatterplot that ends the debate in favour of almost certain linearity.

We can also look at the problem from another angle: not all cars have the same conversion factor between distance travelled and fuel consumption (especially as speed and other driving habits can influence the data), but in general, we might expect the relationship to be linear.

Q10

Using `R`, find the decomposition into sums of squares for the regression.

Solution: the sum-of-squares decomposition is

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

where $\text{SST} = S_{yy}$, $\text{SSR} = b_1^2 S_{xx}$, and $\text{SSE} = \sum e_i^2$.

We thus have:

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

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

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

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

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

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

We see that

$$36827.72 = 26411.59 + 10416.13 :$$

```
SSR+SSE
```

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