PEC3

Angel Hugo Montes Hernández, Fernando Moral Algaba

30 de diciembre de 2018

Sección 1 (8 puntos)

1. (1 punto) Buscad un conjunto de datos relacionados con la Bioestadística o Bioinformática. Para ello, podéis utilizar recursos conocidos de la PEC1, por ejemplo, como es el caso de http://www.bioinformatics.org/sms2/index.html o de http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets. También podéis utilizar otros recursos propios que conozcáis o que sean de vuestro interés, y siempre teniendo en cuenta que sean datos públicos que podéis utilizar. Tenéis que explicar la procedencia de los datos así como incluir las referencias que correspon dan y justificar porqué habéis elegido estos datos.

Duchenne Muscular Dystrophy Dataset

This dataset is from M. Percy, listed in Table 38 of DF Andrews and AM Herzberg: Data, New York: Springer-Verlag, 1985 and also available on StatLib. The 209 observations correspond to blood samples on 192 patients (17 patients have two samples in the dataset) collected in a project to develop a screening program for female relatives of boys with DMD. The program's goal was to inform a woman of her chances of being a carrier based on serum markers as well as her family pedigree. Another question of interest is whether age and season should be taken into account. Enzyme levels were measured in known carriers (75 samples) and in a group of non-carriers (134 samples). Note that the original observation numbers (within subject) on this dataset do not agree with replicates of hospital IDs, so they have been recomputed here. Another anomaly of the dataset is that 16 out of 17 subjects having two blood samples drawn had differing carrier status for the two observations.

The first two serum markers, creatine kinase and hemopexin (ck,h), are inexpensive to obtain, while the last two, pyruvate kinase and lactate dehydroginase (pk,ld), are more expensive. It is of interest to measure how much pk and ld add toward predicting the carrier status. The importance of age and sample date is also of interest. Percy noted that the water supply for the lab changed during the study.

Frank E Harrell Jr

Last modified: Fri Dec 27 16:55:37 EST 2002

Table 1: Resumen Variables (desde URL)

Name	Labels	Class	Storage	NAs
hospid	Hospital ID		double	0
age	Age in Years		double	0
sdate	Date of Study	date	double	0
ck	Creatine Kinase		double	0
h	Hemopexin		double	0
pk	Pyruvate Kinase		double	8
ld	Lactate Dehydroginase		double	7
carrier	Carrier of DMD		double	0
obsno	Observation Number within Patient		double	0

2. (1 punto) Utilizando R, mostrad y explicad qué tipo de fichero habéis importado y las variables que forman parte de él (tipo, clasificación,...), así como todo aquello que creáis relevante. Incluir capturas de pantalla y las instrucciones en R que habéis utilizado para importar y mostrar los datos.

```
## head(mydata)
```

```
##   X hospid age sdate ck         h       pk ld carrier obsno
## 1 1    657  27  6497 22  99.00000 10.79883 NA       0     1
## 2 2    667  31  6528 29  94.00000 11.79883 NA       0     1
## 3 3    669  22  6558 22  85.50000 15.00000 NA       0     1
## 4 4    671  25  6497 41  87.29688 15.00000 NA       0     1
## 5 5    673  26  6558 28  93.50000  7.00000 NA       0     1
## 6 6    675  38  6558 45 108.00000 13.69922 NA       0     1
```

```
## tail(mydata)
```

```
##       X hospid age sdate  ck          h pk  ld carrier obsno
## 204 204   1493  40  7288 123  25.398438 NA 275       1     1
## 205 205   1496  32  7288 610 111.687500 NA 593       1     1
## 206 206   1513  30  7288 510  60.195312 NA 272       1     1
## 207 207   1531  36  7319  55  20.699219 NA 262       1     1
## 208 208   1536  31  7319  45  13.798828 NA 217       1     1
## 209 209   1538  59  7319  25   9.199219 NA 316       1     1
```

```
## colnames(mydata)
```

```
##  [1] "X"       "hospid"  "age"     "sdate"   "ck"      "h"       "pk"
##  [8] "ld"      "carrier" "obsno"
```

```
## str(mydata)
```

```
## 'data.frame':    209 obs. of  10 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ hospid : int  657 667 669 671 673 675 682 762 763 764 ...
##  $ age    : int  27 31 22 25 26 38 24 22 22 25 ...
##  $ sdate  : int  6497 6528 6558 6497 6558 6558 6497 6740 6740 6740 ...
##  $ ck     : num  22 29 22 41 28 45 26 34 51 37 ...
##  $ h      : num  99 94 85.5 87.3 93.5 ...
##  $ pk     : num  10.8 11.8 15 15 7 ...
##  $ ld     : int  NA NA NA NA NA NA NA 144 149 167 ...
##  $ carrier: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ obsno  : int  1 1 1 1 1 1 1 1 1 1 ...
```

3. (2 puntos) Realizad un mínimo de seis preguntas objetivo que den una idea de la información contenida en el conjunto de datos escogido. Para ello, podéis basaros en el tipo de consultas realizadas a la Sección 2 de la PEC1 y también utilizando, en alguno de los casos, la definición de funciones tal como se trabaja en el LAB3.

4. 1. ¿Cual es el numero total de pacientes del estudio?

```
## [1] "total pacientes: 192"
```

3. 2. ¿Cuantas observaciones se han realizado?

```
## [1] "total observaciones: 209"
```

3. 3. Haz una tabla que indique cuantos pacientes hay de cada edad

```
## Frecuencia edad
## 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
##  7  1 13  3  4 14 15 18  6  4 15 14 13  9  6 12 10  3  4 12  2  2  3  2  2
```

```
## 45 48 52 53 54 58 59 61
##  1  1  2  3  1  2  3  2
```

3. 4. Indica el valor medio de la Hemopexina y la desviación para el conjunto de observaciones

```
## [1] "media Hemopexina: 84.28"
```

```
## [1] "sd: 17.06"
```

3. 5. Muestra la ID de los pacientes portadores que tengan un valor de Creatine Kinase inferior al valor medio

```
## Portadores con ck inferior a la media
```

```
##   [1]  657  667  669  671  673  675  682  762  763  764  765  766  767  768
##  [15]  769  770  771  773  774  776  777  778  779  781  782  785  786  789
##  [29]  789  790  791  798  801  802  804  810  813  818  819  824  825  829
##  [43]  831  895  896  899  902  903  904  907  908  909  910  911  911  913
##  [57]  913  914  916  917  918  919  920  921  924  926  927  929  929  933
##  [71]  934  936  938  940  941  942  943  947  947  948  949  949  951  956
##  [85]  966  970  987  989  990 1001 1003 1007 1009 1009 1010 1011 1012 1014
##  [99] 1015 1016 1017 1019 1021 1022 1024 1050 1066 1141 1153 1155 1168 1217
## [113] 1218 1219 1220 1223 1236 1244 1245 1246 1247 1248 1249 1249 1250 1252
## [127] 1253 1253 1254 1255 1258 1259 1260 1261 1262 1281 1285 1287 1289 1290
## [141] 1292 1294 1295 1296 1296 1298 1300 1301 1302 1303 1305 1306 1307 1310
## [155] 1311 1323 1324 1325 1327 1328 1332 1354 1395 1531 1536 1538
```

3. 6. Indica el número de pacientes cuyo valor de Creatine Kinase, Hemopexin y Pyruvate Kinase sean superiores a la media.Indica si son o no portadores.

```
## [1] "Pacientes con h, ck y pk superiores a la media: 24"
```

```
## [1] "Pacientes portadores con h, ck y pk superiores a la media: 23"
```

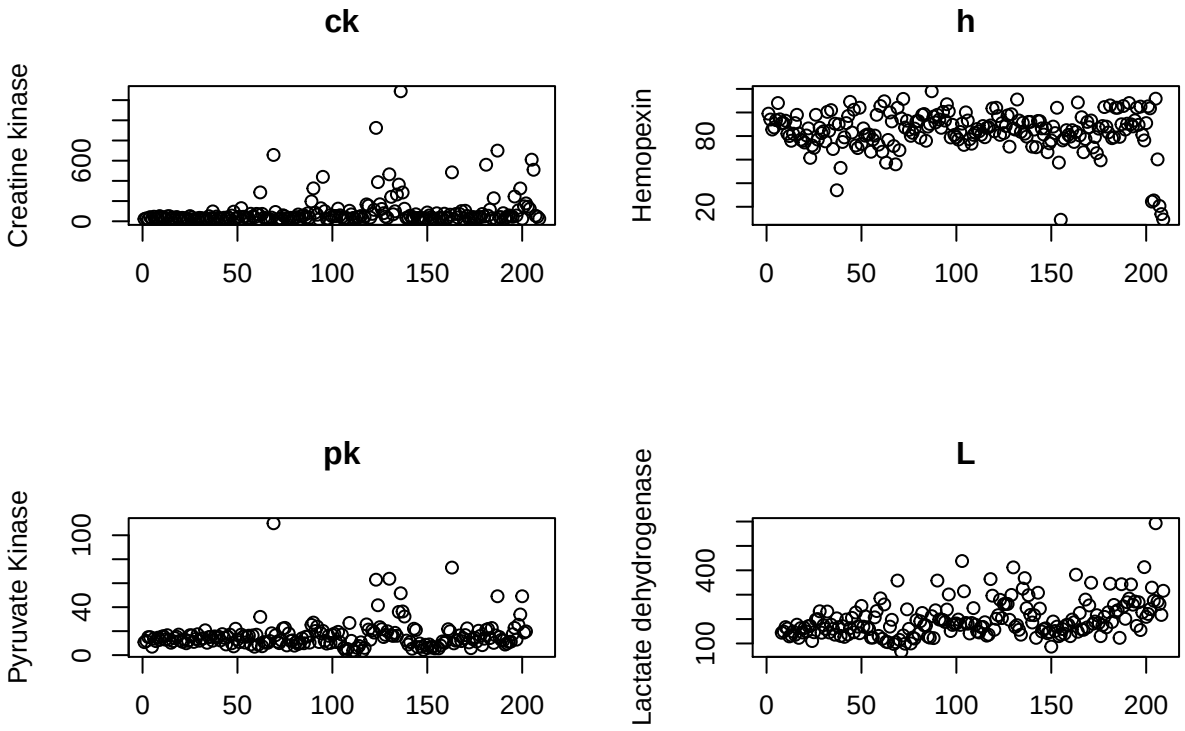
4. (1 punto) Realizad un análisis descriptivo de los datos. Este estudio debe incluir, como se vio en la Sección 3 de la PEC1, un resumen paramétrico de los datos y diversas representaciones gráficas de los mismos basadas en determinados criterios. Dejamos a vuestra elección el tipo de gráficos y los criterios utilizados.

```
summary(mydata)
```

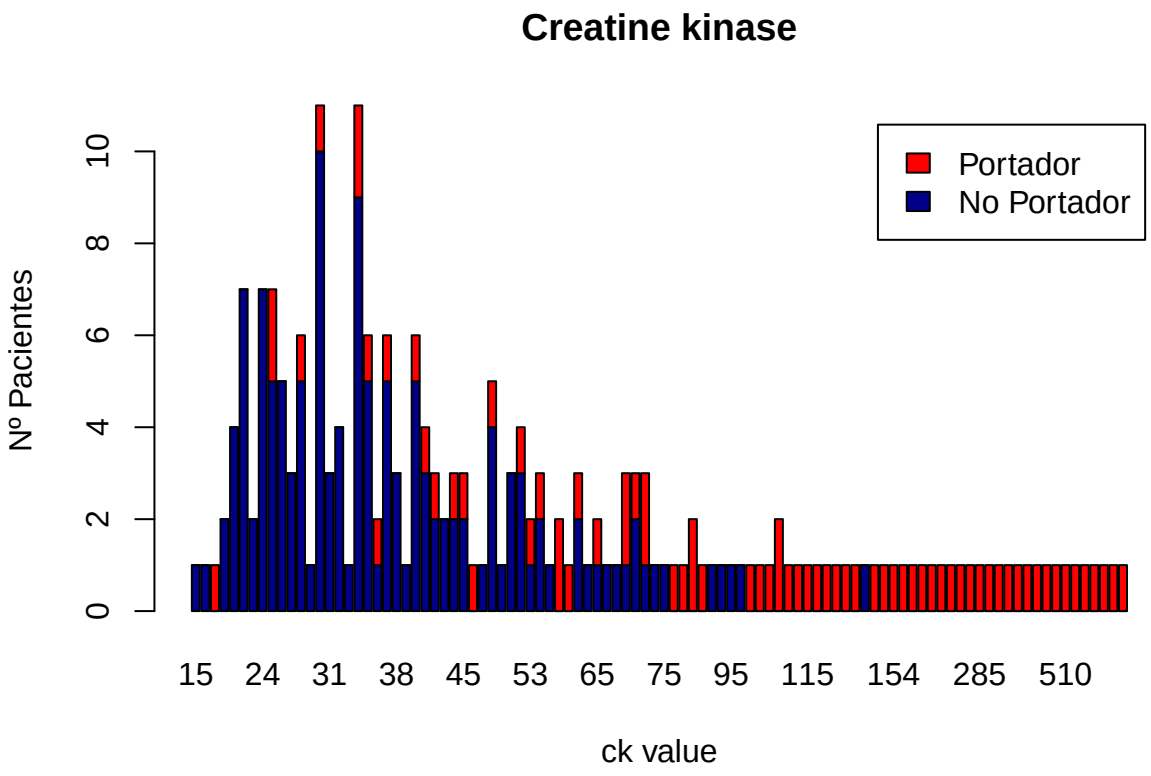
```
##        X           hospid          age            sdate
##  Min.   :  1   Min.   : 657   Min.   :20.00   Min.   :6497
##  1st Qu.: 53   1st Qu.: 907   1st Qu.:26.00   1st Qu.:6832
##  Median :105   Median :1009   Median :31.00   Median :7013
##  Mean   :105   Mean   :1055   Mean   :32.16   Mean   :6992
##  3rd Qu.:157   3rd Qu.:1255   3rd Qu.:36.00   3rd Qu.:7135
##  Max.   :209   Max.   :1538   Max.   :61.00   Max.   :7319
##
##        ck                h                pk               ld
##  Min.   :  15.00   Min.   :  9.00   Min.   :  2.80   Min.   : 66.0
##  1st Qu.:  30.00   1st Qu.: 78.00   1st Qu.: 10.30   1st Qu.:148.2
##  Median :  41.00   Median : 86.00   Median : 13.70   Median :177.0
##  Mean   :  92.26   Mean   : 84.28   Mean   : 16.07   Mean   :198.6
##  3rd Qu.:  73.00   3rd Qu.: 93.19   3rd Qu.: 17.40   3rd Qu.:231.8
##  Max.   :1288.00   Max.   :118.00   Max.   :110.00   Max.   :593.0
##                                     NA's   :8        NA's   :7
##     carrier           obsno
##  Min.   :0.0000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:1.000
```

```
##  Median :0.0000   Median :1.000
##  Mean   :0.3589   Mean   :1.081
##  3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :2.000
##
```

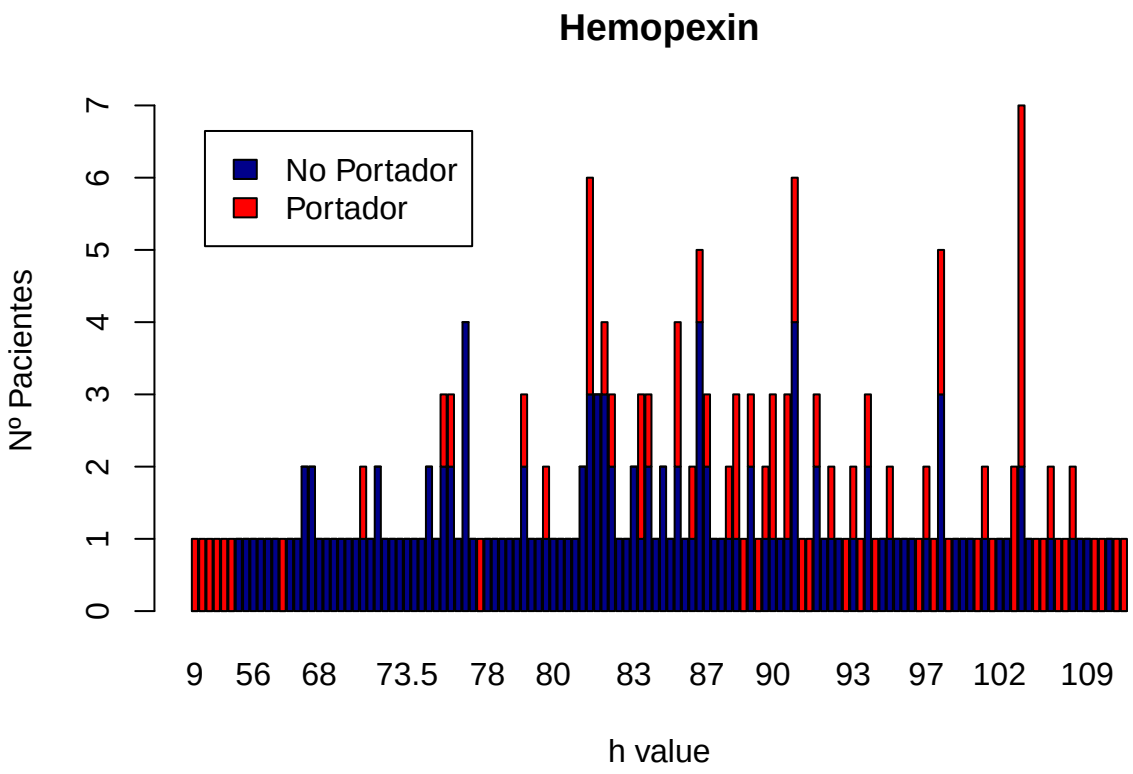
```
par(mfrow=c(2,2))
plot(mydata$ck, xlab ="", ylab = "Creatine kinase", main = "ck" )
plot(mydata$h, xlab = "", ylab = "Hemopexin", main = "h")
plot(mydata$pk, xlab = "", ylab = "Pyruvate Kinase", main= "pk")
plot(mydata$ld, xlab = "", ylab = "Lactate dehydrogenase", main = "L")
```



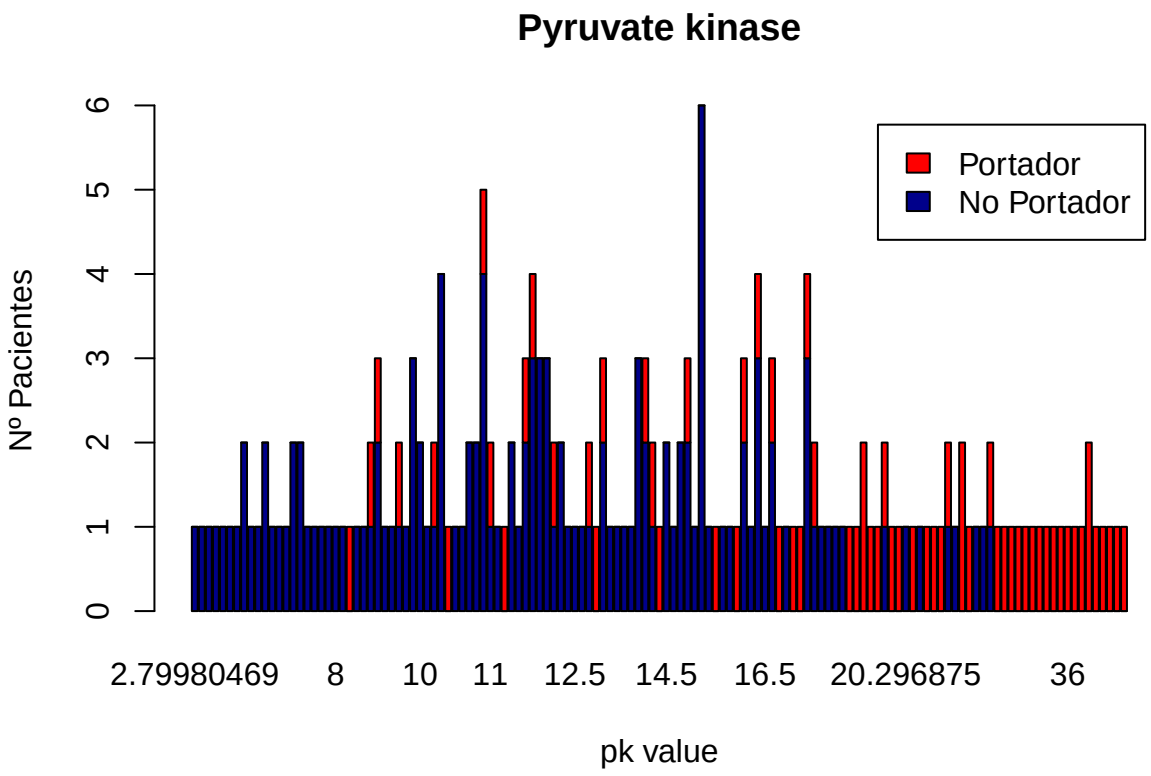

```
par(mfrow=c(1,1))
counts_ck <- table(mydata$carrier, mydata$ck)
barplot(counts_ck, main="Creatine kinase",
  xlab="ck value",ylab="Nº Pacientes", col=c("darkblue","red"),
  legend = c("No Portador", "Portador") )
```



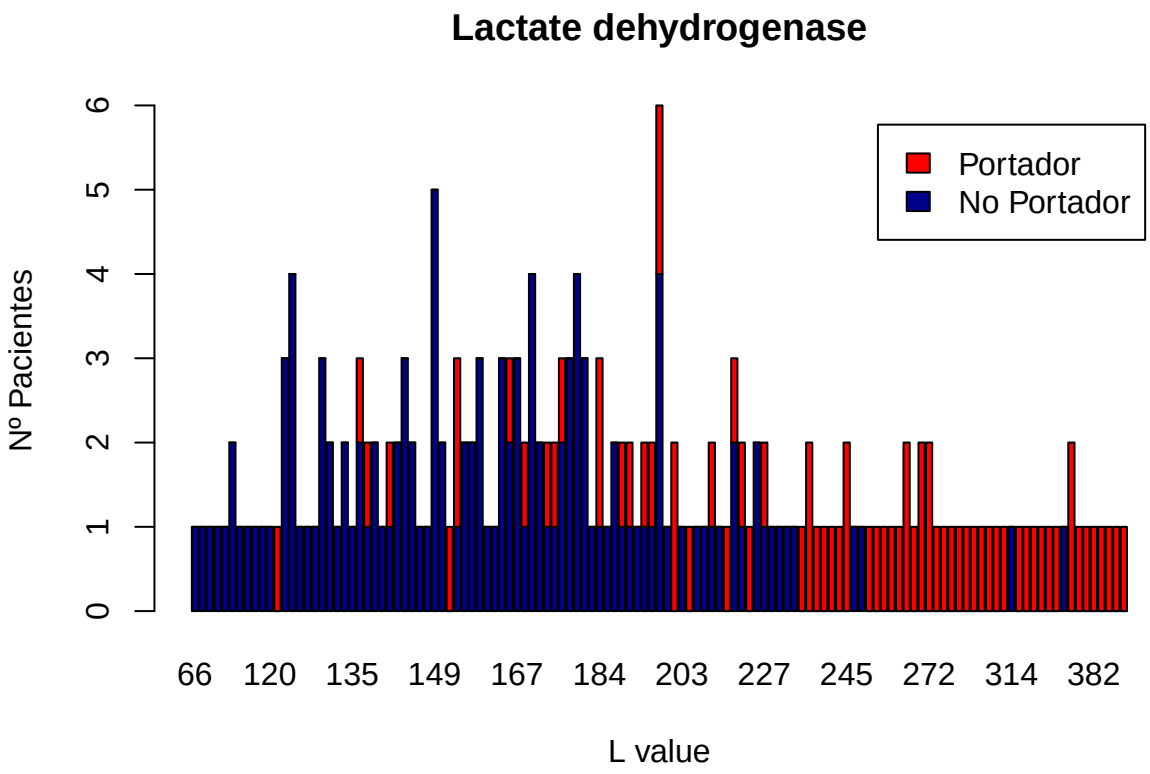

```
counts_h <- table(mydata$carrier, mydata$h)
barplot(counts_h, main="Hemopexin",
  xlab="h value",ylab="Nº Pacientes", col=c("darkblue","red"))
legend("topleft",inset=0.05, legend = c("No Portador", "Portador"), fill =  c("darkblue","red"), col=c(
```




```
counts_pk <- table(mydata$carrier, mydata$pk)
barplot(counts_pk, main="Pyruvate kinase",
  xlab="pk value",ylab="Nº Pacientes", col=c("darkblue","red"),
  legend = c("No Portador", "Portador"))
```




```
counts_l <- table(mydata$carrier, mydata$l)
barplot(counts_l, main="Lactate dehydrogenase",
  xlab="L value",ylab="Nº Pacientes", col=c("darkblue","red"),
  legend = c("No Portador", "Portador"))
```




```
# Según los gráficos apreciamos que niveles altos de los biomarcadores Creatine kinase,
# Pyruvate kinase y Lactate dehydrogenase, son más claramente indicativos que los niveles
# de kemopexin para determinar si se es portador.
```

5. (1 punto) Realizad, basándoos en los conceptos trabajados en el LAB4 y PEC2, un mínimo de tres cuestiones que respondan a una cuestión de probabilidad y un mínimo de una cuestión que corresponda a un breve modelo de simulación.

6. 1. Como hemos visto en el apartado 3.1 el total de pacientes es de 192. Calcular el porcentaje de portadores. Sabiendo dicho porcentaje determinar que probabilidad existe de que al tomar 10 pacientes al azar más de 6 sean portadores.

```
# Pacientes
pacientes <- length(unique(mydata$hospid))
sprintf("total pacientes: %d", pacientes)
```

```
## [1] "total pacientes: 192"
```

```
# portadores
portadores <- length(unique(mydata$hospid[mydata$carrier==0]))
sprintf("Portadores: %d", portadores )
```

```
## [1] "Portadores: 133"
```

```
# % portadores
porcentaje <- (portadores/pacientes)*100
sprintf("Porcentaje: %.2f", porcentaje )
```

```
## [1] "Porcentaje: 69.27"
```

```r
# Probabilidad más de 6 portadores
probabilidad <- 1 - pbinom(6, 10, porcentaje*0.01, lower.tail=TRUE, log.p= FALSE)
sprintf("Probabilidad: %.2f", probabilidad)
```

```
## [1] "Probabilidad: 0.63"
```

5. 2. Supongamos que el número de pacientes diagnosticados erroneamente de DMD siguen una distribución de poisson con una media de 2.4 pacientes por cada 1000 casos. ¿ Cual es la probabilidad de que se erre el diagnostico 2 veces entre mil pacientes?

```r
# E(x) = 2.4 = lambda

probabilidad2 <- dpois(2,2.4)
sprintf("Probabilidad: %.2f", probabilidad2)
```

```
## [1] "Probabilidad: 0.26"
```

5. 3. Sabiendo que el 64 % de las muestras corresponde a portadores y el 36% a no portadores y que entre las muestras de portadores el 35 % corresponde a mayores de 25 años mientras que en las de no portadores son el 47 % los mayores de 25 años, ¿ que probabilidad hay de que tomemos una muestra al azar perteneciente a una persona mayor de 25 años y esta sea portadora?

```r
# p = portador
# np = no portador
# M = mayor de 25 años

# Aplicando el teorema de bayes

# P(p|M)= (P(p)*P(M|p)) / (P(p)*P(M|p) + P(np)*p(M|np))
probabilidad3 = (0.64*0.35) / ((0.64*0.35)+ 0.36*0.47)
sprintf("La probabilidad es de: %.2f", probabilidad3)
```

```
## [1] "La probabilidad es de: 0.57"
```

5. 4. Para el biomarcador “Lactate dehydrogenase” La media y desviación entre las muestras de portadores es m_port = 256.2 y de sd_port = 81.1 respectivamente, siendo de m_no_port = 164.6 y sd_no_port = 41.4 para los no portadores. Supongamos que ambas distribuciones corresponden a distribuciones normales con sus respectivas medias y desviaciones. Genera 500 valores de cada distribución, unelos en un solo conjunto y compara el histograma del conjunto así generado con el histograma de los valores experimentales.Compara tambien la media y desviacion con las del estudio experimental.

```r
set.seed(6509)
port<-rnorm(500, 256.2, 81.1)
no_port<-rnorm(500, 164.6, 41.4)
simulado <- c(port,no_port)

sprintf("media simulacion: %f",mean(simulado))
```

```
## [1] "media simulacion: 206.797648"
```

```r
sprintf("desviacion simulacion: %f",sd(simulado))
```

```
## [1] "desviacion simulacion: 78.176530"
```

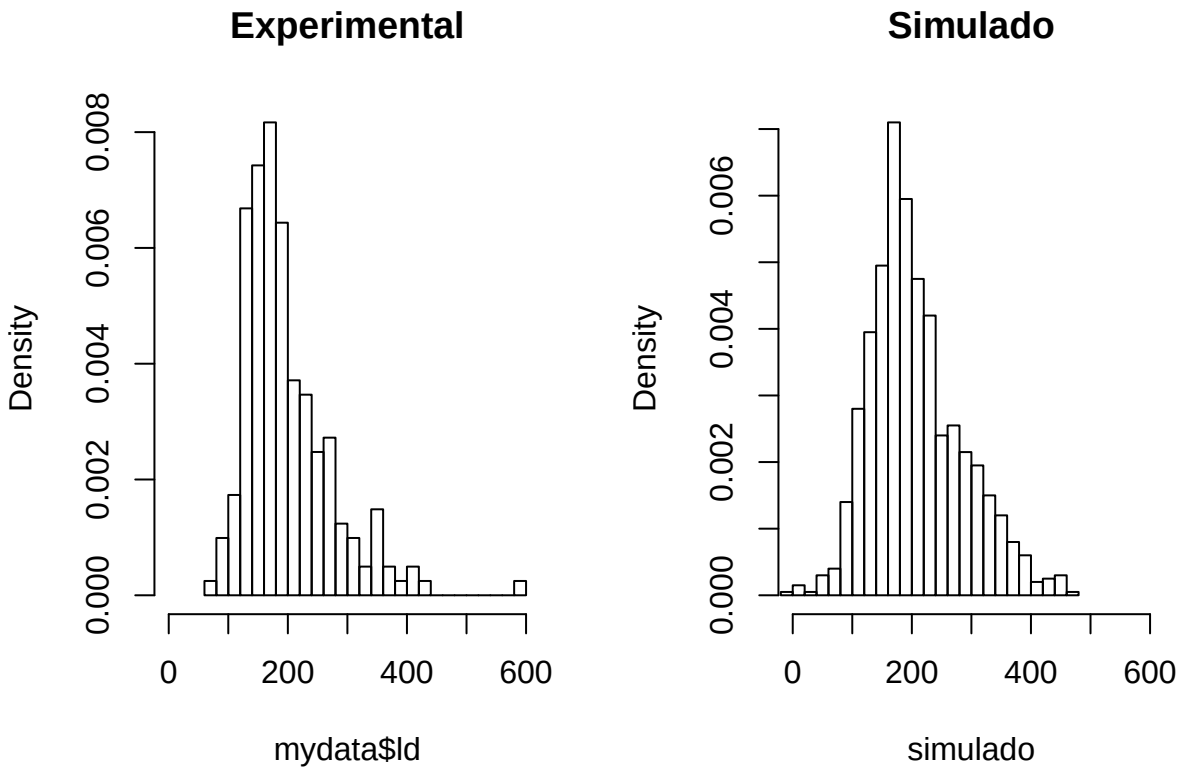
```r
m_estudio <- mean(mydata$ld, na.rm=TRUE)
sd_estudio <- sd(mydata$ld, na.rm=TRUE)

sprintf("media experimental: %.2f",m_estudio)
```

```
## [1] "media experimental: 198.59"
```

```
sprintf("desviacion experimental: %.2f", sd_estudio)
```

```
## [1] "desviacion experimental: 73.92"
```

```
par(mfrow=c(1,2))
hist(mydata$ld, xlim = c(0,600), freq=FALSE, breaks= 20, main = "Experimental")
hist(simulado, xlim = c(0,600), freq = FALSE,breaks= 20, main = "Simulado")
```



6. (1 punto) Realizad un breve análisis de regresión a partir de las variables que disponéis y utilizando el criterio que responda a alguna pregunta de interés que os hayáis planteado.

```
# Pregunta 1: ¿Hay relación entre la edad de los pacientes y los niveles de ck?

# Para dilucidad esta cuestión realizamos un análisis de regresión lineal de la
# variable independiente age respecto a la variable dependiente ck

reg_lin <- lm(ck~age, mydata)
summary(reg_lin)
```

```
##
## Call:
## lm(formula = ck ~ age, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -145.95  -56.56  -40.75  -13.13 1187.74
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.801     40.734   0.044   0.9648
## age            2.813      1.224   2.298   0.0226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.3 on 207 degrees of freedom
## Multiple R-squared:  0.02488,    Adjusted R-squared:  0.02016
## F-statistic: 5.281 on 1 and 207 DF,  p-value: 0.02256
```

```
# Obtenemos un p-valor<0.05. Es estadísticamente significativo. Hay relación
# lineal entre age y los niveles de ck

# La recta de regresión es ck = 1.801 + 2.813*age
# Por cada año que se suma esperamos que los niveles de ck aumenten en 2.813

# Confirmamos los resultados anteriores con las siguientes ecuaciones

rto.estimates <- function(x, y) {
  b1 <- sum(x * y) / sum(x^2)
  ssr <- b1^2 * sum(x^2)
  sse <- sum(y^2) - ssr
  mse <- sse / (length(x) - 1)
  msr <- ssr / 1
  res.std.err <- sqrt(mse)
  f.stat <- msr / mse
  std.error <- sqrt(mse / sum(x^2))

  r2 <- ssr / (sse + ssr)

  adj.r2 <- r2 - (1 - r2) * (2 - 1) / (length(x) - 1)

  res <- data.frame(rbind(b1, res.std.err, f.stat, std.error, r2, adj.r2))
  rownames(res) <- c('b1', 'Residual Standard Error', 'F-Statistic', 'b1 Standard Error',
                     'r-squared', 'Adjusted r-squared')
  colnames(res) <- 'Estimates'

  print(format(res, scientific = FALSE, digits = 3))
}

rto.estimates(mydata$age, mydata$ck)
```

```
##                          Estimates
## b1                           2.865
## Residual Standard Error    150.983
## F-Statistic                 83.346
## b1 Standard Error            0.314
## r-squared                    0.286
## Adjusted r-squared           0.283
```

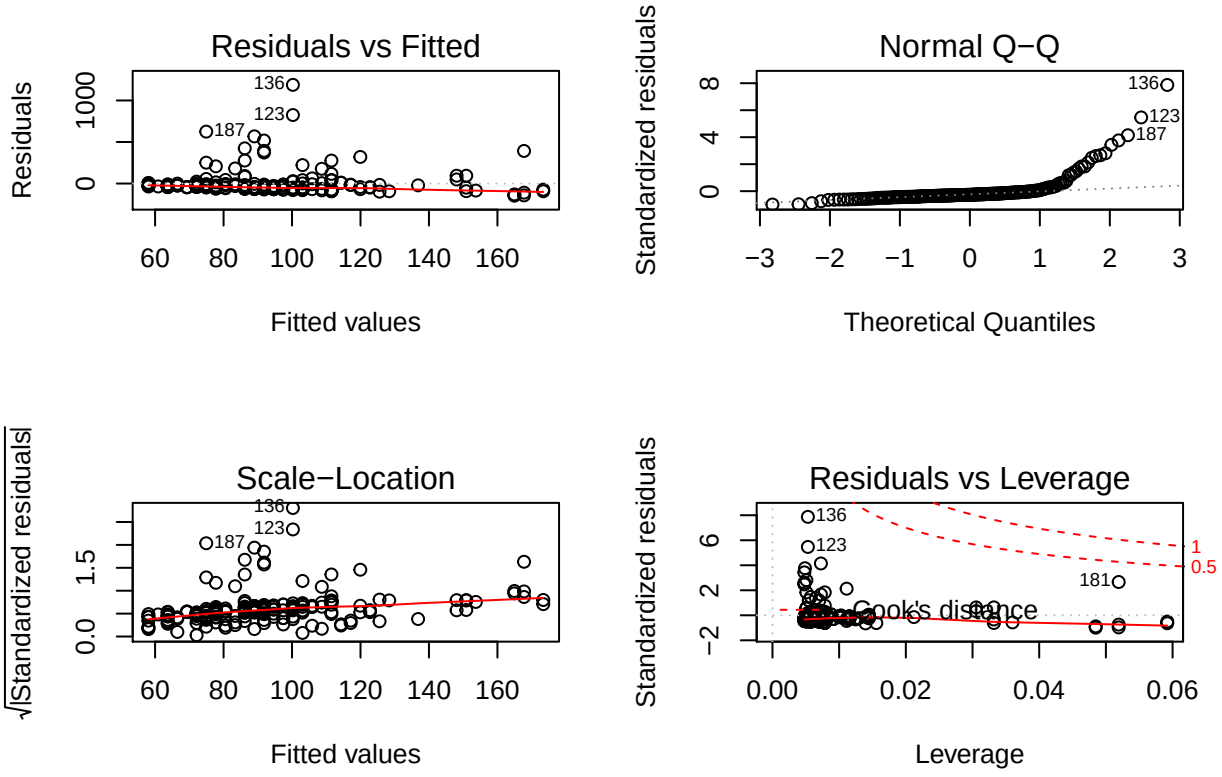
```
# Calculamos los intervalos de confianza al 95%
```

```r
confint(reg_lin, level = 0.95)
```

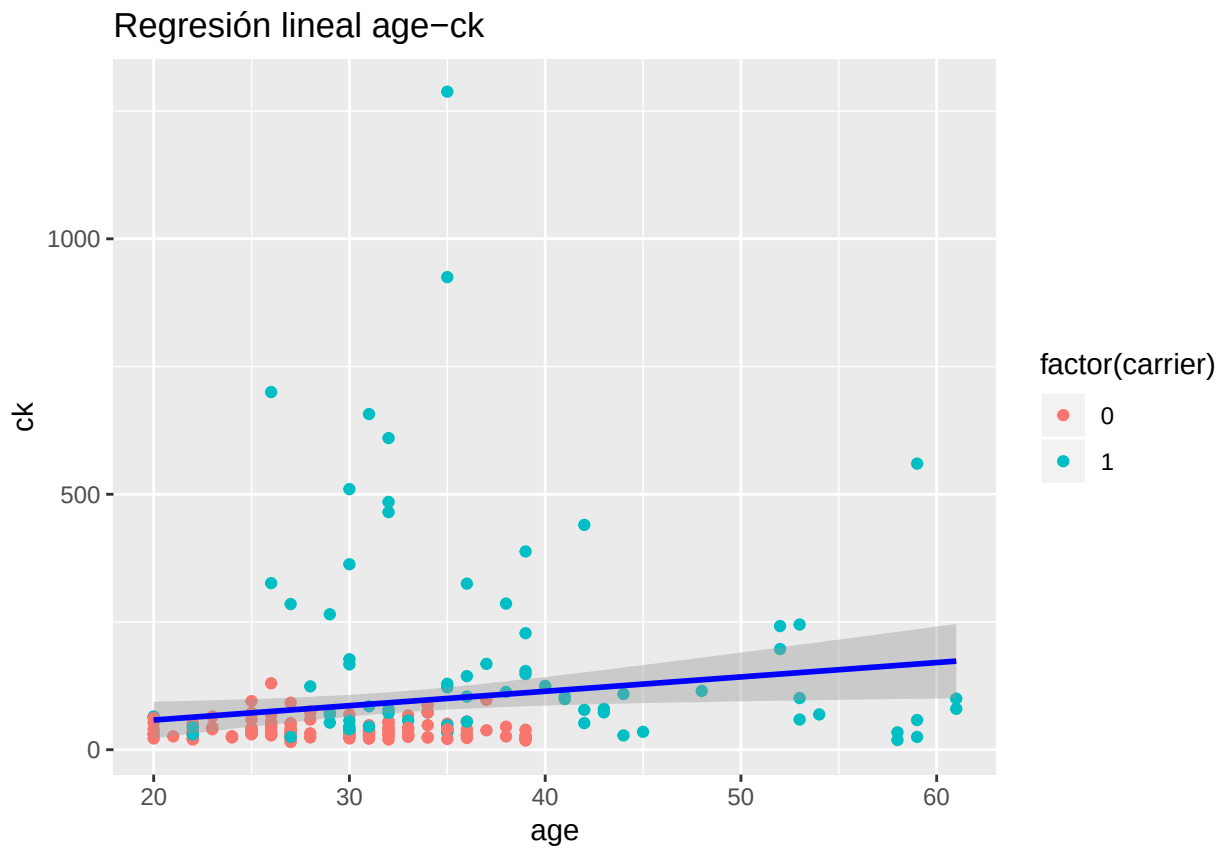
```
##                   2.5 %    97.5 %
## (Intercept) -78.5053996 82.106746
## age           0.3996384  5.226357
```

```r
# Hacemos un diagnóstico y evaluamos la calidad del modelo de regresión

par(mfrow = c(2, 2))
plot(reg_lin)

# El modelo parece bastante aceptable desde el punto de vista de la normalidad de los datos.



# Representamos la nube de puntos y la recta de regresión lineal

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```




```r
ggplot(mydata, aes(age,ck))+geom_point(aes(color=factor(carrier)))+
  stat_smooth(method="lm", col="blue")+labs(title="Regresión lineal age-ck")
```




```
# Calculamos la correlación que puede existir entre las variables age y ck

cor(mydata$age, mydata$ck)
```

```
## [1] 0.15772
```

```
# Existe una correlación lineal positiva entre age y ck
```

```
# Pregunta 2: ¿Hay relación entre la edad de los pacientes y los dias de hospitalización?

# Para responder a esta pregunta realizamos un análisis de Regresión lineal de age frente
# a los días de hospitalización

reg_lin2 <- lm(hospid~age, mydata)
summary(reg_lin2)
```

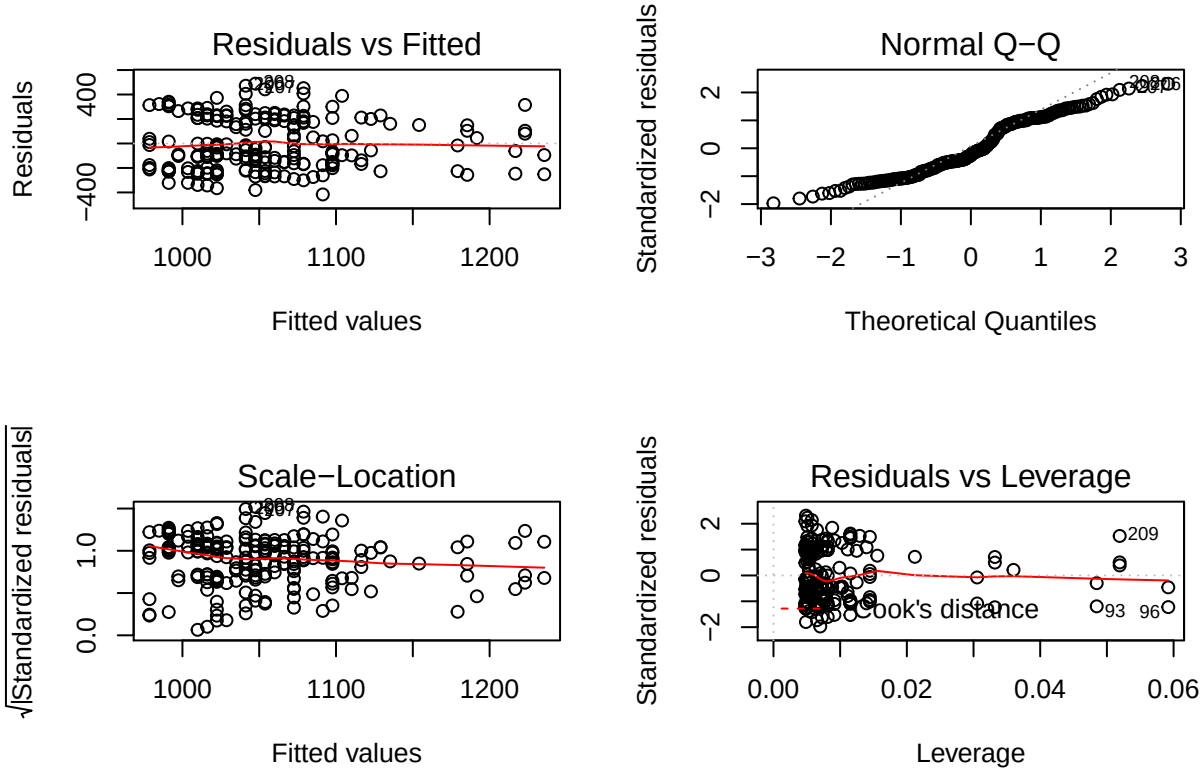
```
##
## Call:
## lm(formula = hospid ~ age, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -416.39 -171.66  -54.24  202.49  488.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  853.144     57.018  14.963  < 2e-16 ***
## age            6.270      1.714   3.659 0.000321 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211.9 on 207 degrees of freedom
## Multiple R-squared:  0.06075,    Adjusted R-squared:  0.05621
## F-statistic: 13.39 on 1 and 207 DF,  p-value: 0.0003213
```

```
# El p-valor < 0.05. Existe significación. Hay relación entre la edad y los días de
# hospitalización


# La recta de regresión lineal es hospid = 853.144 + 6.270*age
# Por cada año que se sume esperamos que los días de hospitalización aumenten en 6.270

# Confirmamos los datos de la regresión usando las siguientes ecuaciones

rto.estimates <- function(x, y) {
  b1 <- sum(x * y) / sum(x^2)
  ssr <- b1^2 * sum(x^2)
  sse <- sum(y^2) - ssr
  mse <- sse / (length(x) - 1)
  msr <- ssr / 1
  res.std.err <- sqrt(mse)
  f.stat <- msr / mse
  std.error <- sqrt(mse / sum(x^2))

  r2 <- ssr / (sse + ssr)

  adj.r2 <- r2 - (1 - r2) * (2 - 1) / (length(x) - 1)

  res <- data.frame(rbind(b1, res.std.err, f.stat, std.error, r2, adj.r2))
  rownames(res) <- c('b1', 'Residual Standard Error', 'F-Statistic', 'b1 Standard Error',
                     'r-squared', 'Adjusted r-squared')
  colnames(res) <- 'Estimates'

  print(format(res, scientific = FALSE, digits = 3))
}

rto.estimates(mydata$age, mydata$hospid)
```

```
##                          Estimates
## b1                          31.047
## Residual Standard Error    304.915
## F-Statistic               2399.315
## b1 Standard Error            0.634
## r-squared                    0.920
## Adjusted r-squared           0.920
```

```
# Hacemos un diagnóstico y evaluamos la calidad del modelo de regresión

par(mfrow = c(2, 2))
plot(reg_lin2)
```




```
# Basándonos en estos gráficos anteriores podemos decir que el modelo parece aceptable
# desde el punto de vista de la normalidad


# Calculamos los intervalos de confianza al 95%

confint(reg_lin2, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 740.733270 965.555347
## age           2.891549   9.647906
```

```
# Representamos la nube de puntos y la recta de regresión lineal

ggplot(mydata, aes(age,hospid))+geom_point(aes(color=factor(carrier)))+
  stat_smooth(method="lm", col="blue")+labs(title="Regresión lineal age-hospid")
```


```
# Determinamos la correlación entre la edad y los días de hospitalización

cor(mydata$age, mydata$hospid)
```

```
## [1] 0.2464714
```

```
# Existe una correlación lineal positiva entre la edad y los días de hospitalización
```

7. (1 punto) A partir de los datos de origen y el estudio realizado, haced una valoración final. Para ello, podéis basaros en las siguientes preguntas: "disponemos de conclusiones finales?", "sería necesario hacer un análisis más avanzado?", "faltan datos para obtener otro tipo de información como...?",....

```
# La distrofia muscular de Duchenne (DMD) es una miopatía progresiva que produce
# degeneración y debilidad muscular y que tiene un tipo de herencia ligado al
# cromosoma X recesiva. Es una enfermedad grave y relativamente frecuente que se
# asocia con un deterioro clínico progresivo e imparable.

# El gen DMD codifica la proteína distrofina que, en el músculo, une la matriz
# extracelular (laminina) al citoesqueleto de actina.

# Existen una serie de biomarcadores clínicos usados en el diagnóstico de la DMD
# como son los niveles séricos de enzimas como ck y h. Incrementos patológicos
# combinados de ambos enzimas debido a la rabdomiolisis o destrucción de células
# musculares como consecuencia de la DMD son buenos parámetros para estudiar la
# gravedad y el grado de desarrollo de la DMD.


#En las fases preclínica y temprana de la enfermedad, la concentración sérica de
```

```
# creatincinasa (ck) es muy elevada (50-100 veces superior al límite alto de la
# normalidad) debido a que esta enzima es liberada a partir del músculo afectado
# en aquellos pacientes que sufren la distrofia.

# La inmensa mayoría de las mujeres portadoras no presenta manifestaciones clínicas,
# aunque alrededor del 70% muestra concentraciones séricas de creatina quinasa
# ligeramente elevadas que van aumentando progresivamente con la edad.

# Los dos hechos anteriores encajan con el resultado obtenido en la regresión lineal
# que nos permitía determinar si existía una relación entre la edad y los niveles de
# ck, un hecho que, como hemos visto por los resultados obtenidos, así ocurre. Esto
# lo podemos ver en la nube de puntos de la gráfica de regresión lineal age-ck donde
# las mujeres portadoras (carrier=1) tienen valores más elevados del enzima.

# Dado que, conforme se va envejeciendo, se va sufriendo un empeoramiento progresivo
# como consecuencia de la patología, y que podría requerir un mayor tratamiento
# hospitalario, parece razonable confirmar los resultados obtenidos en el segundo
# modelo de regresión lineal que considera que aumentan los días de hospitalización
# conforme se va aumentando la edad de los pacientes.

# Para completar el estudio se necesitarian más variables y datos. Por ejemplo, sería
# interesante obtener datos acerca del tratamiento. El tratamiento con glucocorticoides
# puede retrasar la progresión de la enfermedad durante varios años, por lo que sería
# necesario disponer de esa información y ver si podría afectar de alguna manera a los
# niveles de ck o a los días de hospitalización.

# También serían interesantes estudios genéticos como el análisis de polimorfismos o
# mutuaciones presentes en la región promotora o en el gen de la distrofina y en genes
# que codifican para las diferentes enzimas como la ck. También el estudio de diferentes
# isoformas del gen en una serie de tejidos celulares podría ser una buena fuente de
# información adicional.
```

Sección 2 (2 puntos)

A lo largo del curso se ha trabajado con datos cuyo origen era diverso pero, básicamente, correspondían a archivos de tipo texto o hojas de cálculo. En este ejercicio se os pide que realicéis un breve estudio acerca de cómo gestionar la información a partir de una base de datos. En particular, se pide:

- Seleccionar una base de datos de libre acceso y importad, desde Rstudio, estos datos. Mostrad el código utilizado y el resultado obtenido por pantalla.

Respuesta:

Importaremos una base de datos hecha en SQLite que corresponde a la calidad de las aguas en la unión europea

Waterbase is the generic name given to the EEA's databases on the status and quality of Europe's rivers, lakes, groundwater bodies and transitional, coastal and marine waters, on the quantity of Europe's water resources, and on the emissions to surface waters from point and diffuse sources of pollution.

Esta base de datos puede encontrarse en la página web de la **European Environment Agency** en la URL: https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-1

La base de datos no está disponible para conectar remotamente pero puede ser descargada en el siguiente enlace a fecha 12/01/2019 : http://ftp.eea.europa.eu/www/waterbase/WISE4/v2016_1/Waterbase_v2016_1_WISE4_sqlite.zip

Una vez descompromida la base de datos ocupa aprox. 8 GB por lo que no se adjuntará el archivo a este ejercicio.

```
library(RSQLite)
```

```
## Warning: package 'RSQLite' was built under R version 3.5.2
```

```
# Establecemos una conexión con la base de datos
con <- dbConnect(SQLite(), "S:/Descargas/Waterbase_v2016_1_WISE4.sqlite")

# Obtenemos un listado de las tablas que forman la base de datos
dbListTables(con)
```

```
## [1] "MonitoringSite_DerivedData"
## [2] "T_WISE4_AggregatedData"
## [3] "T_WISE4_AggregatedDataByWaterBody"
## [4] "T_WISE4_BiologyEQRClassificationProcedure"
## [5] "T_WISE4_BiologyEQRData"
## [6] "T_WISE4_DisaggregatedData"
```

```
# Importamos una tabla
mydf <- dbReadTable(con, "T_WISE4_BiologyEQRClassificationProcedure")

# La tabla se importa como un data frame
class(mydf)
```

```
## [1] "data.frame"
```

```
# Obtenemos de la base de datos el nombre de los campos que componen la tabla
dbListFields(con, "T_WISE4_BiologyEQRClassificationProcedure")
```

```
##  [1] "CountryCode"
##  [2] "observedPropertyDeterminandBiologyEQRCode"
##  [3] "parameterWaterBodyCategory"
##  [4] "parameterNCSWaterBodyType"
##  [5] "parameterWFDIntercalibrationWaterBodyType"
##  [6] "parameterNaturalAWBHMWB"
##  [7] "parameterICStatusOfDeterminandBiologyEQR"
##  [8] "parameterBoundaryValueClasses12"
##  [9] "parameterBoundaryValueClasses23"
## [10] "parameterBoundaryValueClasses34"
## [11] "parameterBoundaryValueClasses45"
## [12] "procedureBiologicalAnalyticalMethodDescription"
## [13] "resultObservationStatus"
## [14] "Remarks"
## [15] "metadata_versionId"
## [16] "metadata_beginLifeSpanVersion"
## [17] "metadata_statusCode"
## [18] "metadata_observationStatus"
## [19] "metadata_statements"
## [20] "UID"
```

```
# Obsérvese que se obtienen el mismo resultado si se utiliza la instrucción names()
# de R sobre el data frame de la tabla importada
```

```
names(mydf)
```

```
##  [1] "CountryCode"
##  [2] "observedPropertyDeterminandBiologyEQRCode"
##  [3] "parameterWaterBodyCategory"
##  [4] "parameterNCSWaterBodyType"
##  [5] "parameterWFDIntercalibrationWaterBodyType"
##  [6] "parameterNaturalAWBHMWB"
##  [7] "parameterICStatusOfDeterminandBiologyEQR"
##  [8] "parameterBoundaryValueClasses12"
##  [9] "parameterBoundaryValueClasses23"
## [10] "parameterBoundaryValueClasses34"
## [11] "parameterBoundaryValueClasses45"
## [12] "procedureBiologicalAnalyticalMethodDescription"
## [13] "resultObservationStatus"
## [14] "Remarks"
## [15] "metadata_versionId"
## [16] "metadata_beginLifeSpanVersion"
## [17] "metadata_statusCode"
## [18] "metadata_observationStatus"
## [19] "metadata_statements"
## [20] "UID"
```

- Realizad un par de consultas, desde Rstudio, a partir de estos datos y mostrad el código utilizado y resultado obtenido por pantalla.

```
# Podemos hacer una query SQL
q <- dbSendQuery(con, "SELECT parameterBoundaryValueClasses23 FROM
                 T_WISE4_BiologyEQRClassificationProcedure WHERE CountryCode = 'ES'" )

# Hacemos el fetch para obtener el resultado de la query
res1<- dbFetch(q)

# consultamos los últimos registros resultado de nuestra query
tail(res1)
```

```
##     parameterBoundaryValueClasses23
## 169                            0.59
## 170                            0.59
## 171                            0.68
## 172                            0.59
## 173                            0.72
## 174                            0.59
```

```
# Limpiamos el fetch
dbClearResult(q)

# Desconectamos de la base de datos
dbDisconnect(con)
# Dado que hemos importado una tabla a R podemos trabajar directamente con ella sin
# estar conectados a la base de datos. Realizaremos la misma consulta pero en formato
# R sobre el data frame importado. Los resultados obtenidos son equivalentes.
res1 <- mydf$parameterBoundaryValueClasses23[mydf$CountryCode=='ES']
tail(res1)
```

```
## [1] 0.59 0.59 0.68 0.59 0.72 0.59
```