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.
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.

```
# Descomentar para usar uno u otro dataset

# Diabetes data

#mydata <- read.csv("http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.csv",header=T, se

# Duchenne muscular dystrophy dataset
mydata <- read.csv("http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/dmd.csv", header= T, sep=",

head(mydata, n=5)
```

```
##   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
```

```
names(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.

```
# Total pacientes

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

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

```
# total observaciones

sprintf("total observaciones: %d", length(mydata$obsno))
```

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

```
# Frecuencia edades

table(ordered(mydata$age), dnn = "age frequency")
```

```
## age frequency
## 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
```

```
# Valor medio Hemopexina

sprintf("media Hemopexina: %f", mean(mydata$h))
```

```
## [1] "media Hemopexina: 84.283278"
```

```
# id portadores con Hemopexin inferior a la media de Hemopexin

id_port <- function(df){

  m <- mean(mydata$h)
  port_hp_inf_media <- df[df$h < m,]
  return(port_hp_inf_media$hospid)
  }
id <- id_port(mydata)
id
```

```
##  [1]  765  766  767  768  771  773  774  776  777  779  781  782  786  789
## [15]  790  791  804  811  818  819  824  895  899  901  903  906  907  908
## [29]  909  910  911  913  914  916  917  920  921  924  934  936  942  947
## [43]  987  990 1001 1003 1007 1009 1011 1012 1013 1014 1016 1017 1021 1050
## [57] 1115 1135 1153 1193 1203 1208 1218 1220 1245 1247 1248 1249 1250 1253
## [71] 1253 1254 1255 1259 1260 1262 1266 1285 1289 1289 1294 1295 1296 1298
## [85] 1303 1305 1305 1307 1358 1381 1487 1493 1513 1531 1536 1538
```

```
# número pacientes con ck, pk y h superiores a la media

age_high_val <- function(df){

  # medias
  m1 <- mean(mydata$h)
  m2 <- mean(mydata$ck)
```

```
  m3 <- mean(mydata$pk, na.rm=TRUE)

  sup_medias <- df[(df$h > m1)&(df$ck > m2)&(df$pk>m3),1]

  return(length(sup_medias))
  }
l <- age_high_val(mydata)
sprintf("Pacientes con h, ck pk superiores a la media: %d", l)
```

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

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.
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 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.
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...?",....

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