PEC 3 - TIPOLOGÍA DE DATOS - ANÁLISIS Y LIMPIEZA DE DATOS

Javier Botija y Nelson Salinas

25/05/2021

Table of Contents

Tema principal: Limpieza y análisis de datos en pacientes con/sin enfermedades cardiacas

Introducción

Según la **Organización Mundial de Salud**, una de las principales causas de defunciones a nivel mundial, son directamente causadas por enfermedades cardiacas, en el año 2012 murieron aproximadamente 17.5 millones de personas en relación a alguna patologia presentada en el corazón, esto representa mas del 30% porciento de muertes a nivel mundial OMS. En los próximos años se pronostica que para el año 2030, 23.6 millones de personas moriran a causa de alguna enfermedad cardiaca. Es por ello importante dar analisis a las variables que participan en el estudio para conocer las relaciones que existen entre las mismas y su efectos en cuanto a pacientes que sufren alguna patologia cardiaca segun el dataset en estudio. Incluso la **Organización Panamericana de Salud** en su página web detalla la importancia de los cuidados cardiovasculares que deben de tener los pacientes y publicaciones cientificas en relacion a dichas enfermedades OPS.

Este trabajo será útil para poder analizar la cantidad de pacientes que se detectan con enfermedades cardiacas y que relación tienen con otras variables, por lo que se puede crear un plan de acción para generar recomendaciones a estos tipos de pacientes por la parte médica correspondiente.

Descripción

El desarrollo de esta actividad abarcará la resolución de los puntos determinados para cada actividad, se tomará una base de datos de pacientes que se han realizado exámenes de corazón y tienen como referencia si tienen o no alguna patologia, esto permitirá resolver cual es la relación que hay entre ciertas variables de análisis y los cuidados respectivos en cuanto a las patologias detectadas.

Objetivos de la actividad

- Aplicar los conocimientos obtenidos para poder tratar un Dataset de información.
- Poder analizar las variables y su relación.
- Concluir y resolver el problema planteado.

Problemática

La problemática que se desea resolver es verificar cual es la incidencia de patologia cardiacas entre los dos sexos participantes (Masculino Femenino), y a ello considerar cuales son los pacientes que deben ser mayormente vigilados en relación a lo ya indicado por la OMS, para la resolución del problema se tomará en cuenta la relación que tienen las variables y con ello poder concluir con un resultado final.

Dataset

El dataset fue obtenido desde uno de los repositorios con mayor acogida a nivel mundial denominado UCI Machine Learning Repository. Link de acceso UCI, en el mismo existen gran cantidad de datasets para analizar. Para el desarrollo de esta actividad se toma en consideración un dataset que contiene información de pacientes con o sin enfermedades cardiacas donde se exponen varios parámetros para su análisis y relación entre variables.

Esta base de datos se compone de 4 datasets separados, de 4 ubicaciones diferentes (Cleveland, Hungría, Suiza y VA Long Beach). Se compone originalmente de 76 atributos, aunque están reducidos y procesados en subconjuntos de 14 características principales.

Esta base de datos contiene 76 atributos, pero todos los experimentos publicados se refieren al uso de un subconjunto de 14 de ellos. En particular, la base de datos de Cleveland es la única que han utilizado los investigadores hasta la fecha.

Nombre del dataset: Heart Disease Data Set

Formato: DATA

Link de acceso: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Cantidad de registros: 920 registros global, 303 registros para análisis cleveland

Variables: 14

Las 14 variables que se usan son las siguientes (# número original de los 76):

```
  -- 1. #3  (age)   age in years
  -- 2. #4  (sex)   sex (1 = male; 0 = female)
  -- 3. #9  (cp)    chest pain type
    -- Value 1: typical angina
    -- Value 2: atypical angina
    -- Value 3: non-anginal pain
    -- Value 4: asymptomatic
  -- 4. #10 (trestbps)    resting blood pressure (in mm Hg on admission
to the
    hospital)
  -- 5. #12 (chol)    serum cholestoral in mg/dl
  -- 6. #16 (fbs)   (fasting blood sugar > 120 mg/dl)  (1 = true; 0 =
false)
  -- 7. #19 (restecg)   resting electrocardiographic results
    -- Value 0: normal
    -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST
                elevation or depression of > 0.05 mV)
    -- Value 2: showing probable or definite left ventricular hypertrophy
                by Estes' criteria
  -- 8. #32 (thalach)   maximum heart rate achieved
```

```
  -- 9. #38 (exang)   exercise induced angina (1 = yes; 0 = no)
  -- 10. #40 (oldpeak)    ST depression induced by exercise relative to
rest
  -- 11. #41 (slope)    the slope of the peak exercise ST segment
    -- Value 1: upsloping
    -- Value 2: flat
    -- Value 3: downsloping
  -- 12. #44 (ca)   number of major vessels (0-3) colored by flourosopy
  -- 13. #51 (thal)   3 = normal; 6 = fixed defect; 7 = reversable defect
  -- 14. #58 (num)    (the predicted attribute) diagnosis of heart
disease (angiographic disease status)
    -- Value 0: < 50% diameter narrowing
    -- Value 1: > 50% diameter narrowing
```

Es una base de datos supervisada, y por tanto tiene una variable llamada 'num' con valores entre 0 y 4, que representa la presencia de una enfermedad cardíaca en el sujeto observado, siendo 0 la no enfermedad a 4 la más grave. Este estudio es importante, dado que ayudará a discernir, al menos, entre la presencia de una enfermedad (del 1 al 4) con respecto a la ausencia (0).

Hay que tener en cuenta que aunque todas las variables son numéricas, muchas de ellas son categóricas: sex, cp, fbs, restecg, exang, slope, ca, thal y la propia 'num'.

Hoy por hoy se potencia mucho este tipo de estudios, dado que conseguir predecir las enfermedades mediante recogida de variables es una de las mayores utilidades del análisis de datos. Puede ser una manera de diagnosticar precozmente enfermedades y actuar así preventinemente a ellas.

Lo que pretendemos realizar es un modelo que nos indique, mediante algunas variables,la posibilidad de contraer una enfermedad, y en qué grado.

Para la correcta resolución de este documento, cargamos las librerías que vamos a utilizar a lo largo del mismo.

```
# cargamos las librería necesrias
library(ggplot2)
library(dplyr)
library(VIM)
library(nortest)
```

Integración y selección de los datos de interés a analizar.

Aunque la base de datos completa es de 4 subdatasets, la propia web nos advierte que los investigadores sólo han usado la base de datos de Cleveland. Vamos a observar por qué motivo es.

```
# leemos los ficheros
clev <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-
databases/heart-disease/processed.cleveland.data',
```

```
                      stringsAsFactors = FALSE,
                      sep = ',',
                      header = FALSE)
hung <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-
databases/heart-disease/processed.hungarian.data',
                      stringsAsFactors = FALSE,
                      sep = ',',
                      header = FALSE)
swit <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-
databases/heart-disease/processed.switzerland.data',
                      stringsAsFactors = FALSE,
                      sep = ',',
                      header = FALSE)
LBva <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-
databases/heart-disease/processed.va.data',
                      stringsAsFactors = FALSE,
                      sep = ',',
                      header = FALSE)
# unimos todos los sub-dataset en uno único para poder trabajar con el
heart = rbind(clev,hung, swit, LBva)
# Dado que es un fichero sin cabeceras, cargamos los nombres de las
columnas
names(heart) <-
c("age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","o
ldpeak","slope","ca","thal","num")
# Vamos a mostrar la estructura del conjunto de datos
str(heart)
```

```
## 'data.frame':    920 obs. of  14 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: chr  "145" "160" "120" "130" ...
##  $ chol    : chr  "233" "286" "229" "250" ...
##  $ fbs     : chr  "1" "0" "0" "0" ...
##  $ restecg : chr  "2" "2" "2" "0" ...
##  $ thalach : chr  "150" "108" "129" "187" ...
##  $ exang   : chr  "0" "1" "1" "0" ...
##  $ oldpeak : chr  "2.3" "1.5" "2.6" "3.5" ...
##  $ slope   : chr  "3" "2" "2" "3" ...
##  $ ca      : chr  "0.0" "3.0" "2.0" "0.0" ...
##  $ thal    : chr  "6.0" "3.0" "7.0" "3.0" ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
```

```
# valores básicos de cada variable
summary(heart)
```

```
##       age             sex               cp          trestbps
##  Min.   :28.00   Min.   :0.0000   Min.   :1.00   Length:920
##  1st Qu.:47.00   1st Qu.:1.0000   1st Qu.:3.00   Class :character
##  Median :54.00   Median :1.0000   Median :4.00   Mode  :character
```

```
##  Mean   :53.51   Mean   :0.7891   Mean   :3.25
##  3rd Qu.:60.00   3rd Qu.:1.0000   3rd Qu.:4.00
##  Max.   :77.00   Max.   :1.0000   Max.   :4.00
##      chol               fbs              restecg           thalach
##  Length:920         Length:920         Length:920         Length:920
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##     exang             oldpeak             slope                ca
##  Length:920         Length:920         Length:920         Length:920
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##      thal                num
##  Length:920         Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :1.0000
##                     Mean   :0.9957
##                     3rd Qu.:2.0000
##                     Max.   :4.0000
```

La base de datos tiene valores nulos. Estos no son representados con un valor vacío, sino que se rellenan con un valor '?', por eso muchas de las columnas se han cargado como chr.

```
# Sí hay valores ?, que no los trataremos dado que entendemos que son
valores desconocidos
colSums(heart=="?")
```

```
##      age      sex       cp trestbps     chol      fbs  restecg
thalach
##        0        0        0       59       30       90        2
55
##    exang  oldpeak    slope       ca     thal      num
##       55       62      309      611      486        0
```

Por este motivo, sabemos que no podemos unir el resto de subsets, dado que muchos de ellos carecen de un número de variables significativas. **Por tanto nos quedaremos con los 303 datos del subset de Cleveland, tal como se anticipaba**.

```
# nos quedamos sólo con los datos de Cleveland
heart = clev
# Dado que es un fichero sin cabeceras, cargamos los nombres de las
```

```
columnas
names(heart) <-
c("age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","o
ldpeak","slope","ca","thal","num")
# Vamos a mostrar la estructura del conjunto de datos
str(heart)

## 'data.frame':    303 obs. of  14 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ restecg : num  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : num  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ ca      : chr  "0.0" "3.0" "2.0" "0.0" ...
##  $ thal    : chr  "6.0" "3.0" "7.0" "3.0" ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
```

Limpieza de los datos.

Con la finalidad de obtener datos integros para el analisis de relacion y la comprensión del valor de los datos, se procederá a realizar la limpieza de los datos. En este punto vamos a convertir todos los valores chr a numéricos:

```
# pasamos a doble los campos con ?
heart$ca <- as.double(heart$ca)

## Warning: NAs introducidos por coerción

heart$thal <- as.double(heart$thal)

## Warning: NAs introducidos por coerción

# valores básicos de cada variable
summary(heart)

##       age             sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
##  Median :56.00   Median :1.0000   Median :3.000   Median :130.0
##  Mean   :54.44   Mean   :0.6799   Mean   :3.158   Mean   :131.7
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##
##       chol            fbs            restecg          thalach
```

```
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :241.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.7   Mean   :0.1485   Mean   :0.9901   Mean   :149.6
##  3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##
##      exang           oldpeak         slope             ca
##  Min.   :0.0000   Min.   :0.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :2.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.601   Mean   :0.6722
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :3.000   Max.   :3.0000
##                                                  NA's   :4
##       thal            num
##  Min.   :3.000   Min.   :0.0000
##  1st Qu.:3.000   1st Qu.:0.0000
##  Median :3.000   Median :0.0000
##  Mean   :4.734   Mean   :0.9373
##  3rd Qu.:7.000   3rd Qu.:2.0000
##  Max.   :7.000   Max.   :4.0000
##  NA's   :2
```

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Aunque la función summary nos ha adelantado los valores NA, vamos a sacar sólo estos datos, asegurándonos que así es.

```
 # Números de valores desconocidos por campo
sapply(heart, function(x) sum(is.na(x)))
```

```
##      age      sex       cp trestbps     chol      fbs  restecg
thalach
##        0        0        0        0        0        0        0
0
##    exang  oldpeak    slope       ca     thal      num
##        0        0        0        4        2        0
```

Dado que tenemos muy pocos registros con datos nulos, podríamos optar por eliminar dichas filas, pero dado que nuestra base de datos no es muy numerosa en sujetos, optaremos por la imputación de valores. Podríamos optar por asignarles la media, pero en este caso utilizaremos el método de los KNN (k nearest neighborhoods) o k vecinos próximos, dado que se supone que tienen una relación entre ellos.

```
# Aplicamos la imputación
heart$ca <- kNN(heart)$ca
heart$thal <- kNN(heart)$thal
```

```
# validamos que no quedan valores NA
sapply(heart, function(x) sum(is.na(x)))
```

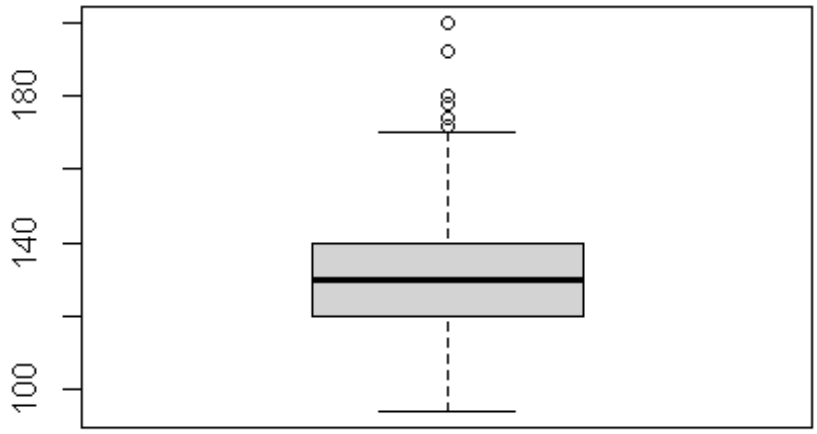
```
##      age      sex       cp trestbps     chol      fbs  restecg
thalach
##        0        0        0        0        0        0        0
0
##    exang  oldpeak    slope       ca     thal      num
##        0        0        0        0        0        0
```

Por lo consecuente se verifica que ya no existen valores nulos.

Identificación y tratamiento de valores extremos.

Para la identificación de valores extremos o outliers, los obtendremos gráficamente de las variables que son continuas y no categóricas, y que los tienen (previo estudio de dichas variables) y también los obtendremos por la función de R que los lista.

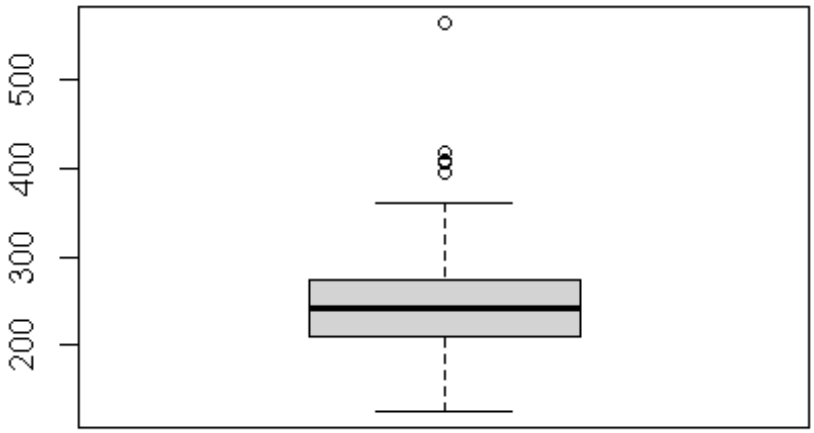
```
# gráfico boxplot para obtener outliers
boxplot(heart$trestbps)
```



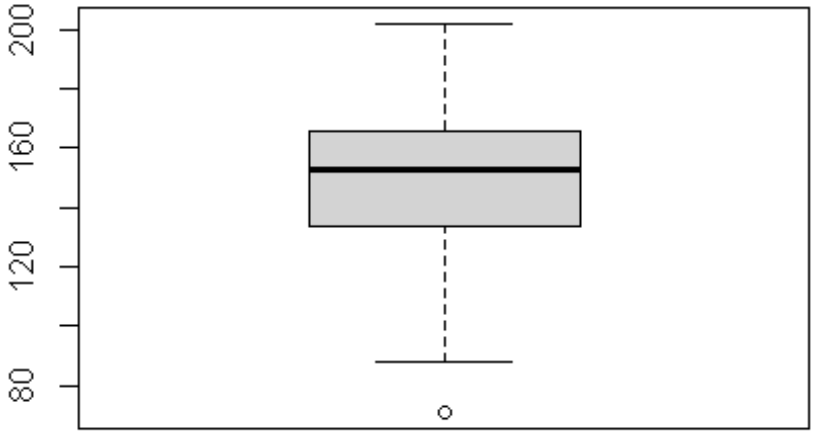

```
# obtenemos los valores
boxplot.stats(heart$trestbps)$out
```

```
## [1] 172 180 200 174 178 192 180 178 180
```

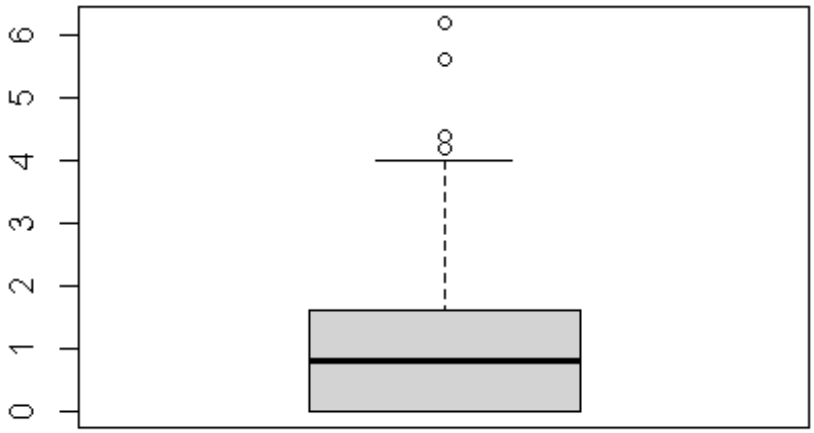
```
boxplot(heart$chol)
```




```
boxplot.stats(heart$chol)$out

## [1] 417 407 564 394 409

boxplot(heart$thalach)
```




```
boxplot.stats(heart$thalach)$out

## [1] 71

boxplot(heart$oldpeak)
```




```
boxplot.stats(heart$oldpeak)$out
```

```
## [1] 6.2 5.6 4.2 4.2 4.4
```

En este caso, aunque los valores superan los XXX, ni los vamos a eliminar, ni los vamos a ponderar, dado que representan precisamente valores posibles y que pueden marcar la diferencia para indicar la enfermedad.

Exportación de los datos procesados

```
 # Exportación de los datos limpios en .csv
write.csv(heart, "heart_clean.csv")
```

Análisis de los datos.

Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para proceder al análisis, vamos a realizar diferentes grupos que permitirían comparar resultados.

```
# Agrupación por sexo
heart.male <- heart[heart$sex == 1,]
heart.female <- heart[heart$sex == 0,]
# Agrupación por fbs
heart.fbs <- heart[heart$fbs == 1,]
```

```
heart.nofbs <- heart[heart$fbs == 0,]
# Agrupación por grupos de edad
heart.30 <- heart[heart$age < 40,]
heart.40 <- heart[heart$age >= 40 & heart$age < 50,]
heart.50 <- heart[heart$age >= 50 & heart$age < 60,]
heart.60 <- heart[heart$age >= 60 & heart$age < 70,]
heart.70 <- heart[heart$age >= 70,]
```

Podríamos realizar más segmentaciones de las mostradas, pero en este estudio, sólo vamos a utilizar la hecha por sexo.

Comprobación de la normalidad y homogeneidad de la varianza.

Para verificar que los valores se aproximan a una población distribuida, realizaremos la prueba de normalidad de Anderson-Darling, que consiste en obtener un p-valor superior a un nivel de significación (alfa) de 0,05.

```
alpha = 0.05
col.names = colnames(heart)
for (i in 1:ncol(heart)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(heart[,i]) | is.numeric(heart[,i])) {
    p_val = ad.test(heart[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(heart) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## age, sex, cp,
## trestbps, chol, fbs,
## restecg, thalach, exang,
## oldpeak, slope, ca,
## thalnum
```

Seguidamente, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos por sexo y por fbs (glucemia en ayunas).

En el siguiente test, considera que la hipótesis nula es que ambas varianzas son iguales.

```
fligner.test(num ~ sex, data = heart)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
```

```
## data:  num by sex
## Fligner-Killeen:med chi-squared = 18.335, df = 1, p-value = 1.853e-05
```

Al obtener un p-valor inferior a 0,05, rechazamos la hipótesis nula y diremos que las varianzas no son igules.

Ahora lo haremos con el fbs:

```
fligner.test(num ~ fbs, data = heart)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  num by fbs
## Fligner-Killeen:med chi-squared = 0.63363, df = 1, p-value = 0.426
```

En este caso sí, al tener un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Contraste de hipótesis

En segundo lugar, vamos a realizar un contraste de hipótesis sobre uno de los segmentos antes realizados (por sexo), para determinar si la posibilidad de tener una enfermedad coronaria es diferente dependiendo del sexo del sujeto. Se suele decir que los hombres padecen más este tipo de enfermedades que las mujeres.

Lo primero que tendremos que validar es que nuestras muestras sean superiores a 30 (preferiblemente más que 50) para poder realizar el contraste.

```
dim(heart.female)

## [1] 97 14

dim(heart.male)

## [1] 206  14
```

Dado que sí lo son, el contraste de hipótesis siguiente es válido.

La hipótesis que nos plantemos consistiría en:

• Hipótesis nula: para la variable que refleja el grado de enfermedad (num), el valor medio de los hombres es igual al de las mujeres

• Hipótesis alternativa: para dicha variable, el valor medio de los hombres es mayor que el de las mujeres.

O lo que es lo mismo, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es bilateral atendiendo a la formulación de la hipótesis alternativa:

$H0 : \mu1 - \mu2 = 0$ $H1 : \mu1 - \mu2 > 0$

donde $\mu1$ es la media de la enfermedad de las mujeres (población de la que se extrae la primera muestra) y $\mu2$ es la media de la enfermedad de los hombres (población de la que extrae la segunda). Consideremos un nivel de confianza de 0.95 ($\alpha = 0, 05$).

```
t.test( x = heart.male$num,
        y = heart.female$num,
        alternative = "greater",
        conf.level  = 0.95)

##
##  Welch Two Sample t-test
##
## data:  heart.male$num and heart.female$num
## t = 4.2851, df = 225.1, p-value = 1.354e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3626725       Inf
## sample estimates:
## mean of x mean of y
## 1.1262136 0.5360825
```

El p-valor obtenido es menor que el valor de significación fijado, por tanto rechazamos la hipótesis nula. En este caso podemos concluir que, efectivamente, **la posibilidad de tener mayor enfermedad en los hombres es mayor que en las mujeres.**

Los valores son próximos, dado que son valores entre 0 y 4, pero la media es superior en hombres que en mujeres.

Correlación

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre el valor que determina la enfermedad.

Se recorren cada culumna del dataset realizando el coeficiente de correlación de Spearman, puesto que ya hemos visto que nuestros datos que no siguen una distribución normal.

```
# creamos la matriz vacía
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
# en nuestro caso todas, con respecto al valor que determina la
enfermedad
for (i in 1:(ncol(heart) - 1)) {
```

```
  if (is.integer(heart[,i]) | is.numeric(heart[,i])) {
    spearman_test = cor.test(heart[,i],
                             heart[,length(heart)],
                             method = "spearman")
    # creamos matriz de resultado
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = spearman_test$estimate
    pair[2][1] = spearman_test$p.value
    # la añadimos a la de resultados
    corr_matrix <- rbind(corr_matrix, pair)
    # nombramos la fila con el nombre de la columna
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(heart)[i]
  }
}
# imprimimos la matriz
print(corr_matrix)
```

```
##             estimate      p-value
## age       0.24169585 2.109189e-05
## sex       0.25786634 5.435562e-06
## cp        0.48693348 1.911373e-19
## trestbps  0.13716366 1.689306e-02
## chol      0.10929344 5.739478e-02
## fbs       0.05164306 3.703448e-01
## restecg   0.18124799 1.533833e-03
## thalach  -0.44394442 4.591425e-16
## exang     0.43525478 1.946468e-15
## oldpeak   0.46303189 1.657777e-17
## slope     0.40030234 4.344070e-13
## ca        0.53363042 1.062652e-23
## thal      0.53969585 2.654246e-24
```

Podemos ver que no existe una variable que correlacione por sí misma con respecto al valor esperado. Hay algunas que superan el 0.5, siendo este un valor bajo.

Si alguna variable, como por ejemplo el fbs o glucemia en ayunas, es porque aunque es un valor numérico, realmente es un valor categórico (1 > 120 mg/dl, 0 <= 120 mg/dl).

Modelo de regresión lineal

Como planteábamos en el inicio de este documento, sería muy interesante poder predecir la posibilidad de tener enfermedades coronarias en base a algunas variables. Vamos a calcular varios modelos de regresión lineal utilizando diferentes variables, cuantitativas (continuas) y/o cualitativas (categóricas), obteniendo de cada uno de ellos el coeficiente de determinación (R2), para así saber cual se ajusta mejor.

Aunque partiremos de los datos de correlación de las variables obtenidos en el apartado anterior, añadiremos otros para ver si mejoran o no los modelos.

```
# modelo con variables más correlacionadas
modelo1 <- lm(num ~ thal + ca + cp + oldpeak + thalach, data = heart)
```

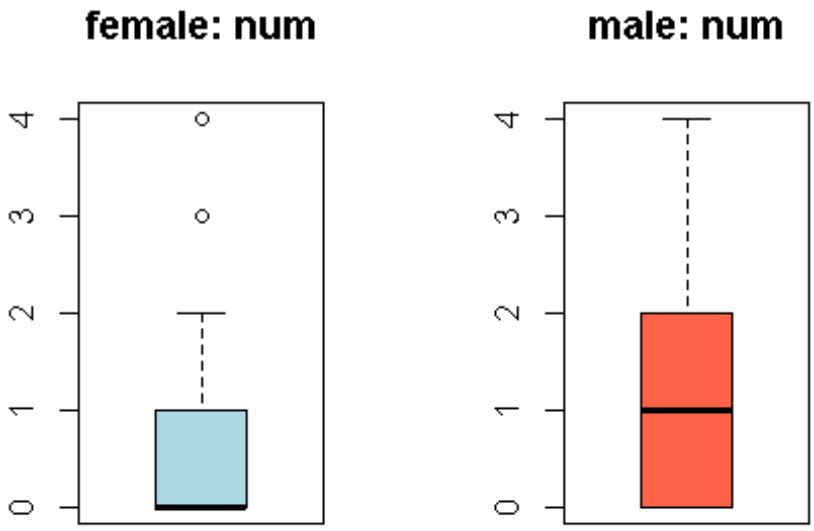
```
# modelo con variables con outliers
modelo2 <- lm(num ~ trestbps + chol + thalach + oldpeak, data = heart)
# modelo con variables más correlacionadas sin outliers
modelo3 <- lm(num ~ thal + ca + cp, data = heart)
# modelo de todas las variables
modelo4 <- lm(num ~ age + sex + cp + trestbps + chol + fbs + restecg +
thalach + exang + oldpeak + slope + ca + thal, data = heart)
tabla.r2 <- matrix(c(1, summary(modelo1)$r.squared,
                     2, summary(modelo2)$r.squared,
                     3, summary(modelo3)$r.squared,
                     4, summary(modelo4)$r.squared), ncol = 2, byrow =
TRUE)
colnames(tabla.r2) <- c("Modelo", "R2")
tabla.r2
```

```
##      Modelo        R2
## [1,]      1 0.5413242
## [2,]      2 0.3271283
## [3,]      3 0.4697399
## [4,]      4 0.5685422
```

Vemos que el último modelo es el que mejor coeficiente de determinación tiene, aunque está claro que el valor sigue siendo bajo.

Representación de los resultados a partir de tablas y gráficas.

```
Conf = matrix(c(1:2), nrow=1, byrow=TRUE)
layout(Conf)
boxplot(heart.female$num, main='female: num', col = 'lightblue' )
boxplot(heart.male$num, main='male: num', col = 'tomato' )
```

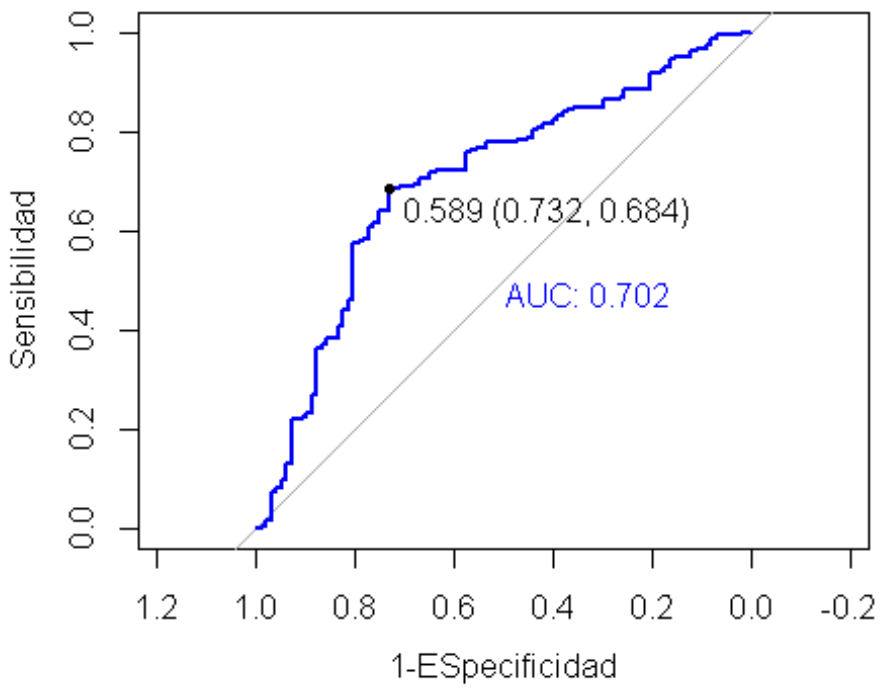


En las gráficas por sexo se puede notar la inferencia que tiene la variable num para hombres como para mujeres, lo cual detalla, que hay mayor incidencia en el sexo masculino en padecer una enfermedad cardiaca.

Otro análisis válido es poder demostrar mediante la curva ROC que tan factible es nuestro modelo, para este analisis se tomara en consideración el modelo4 que es el mejor coeficiente de determinación tuvo.

```
library(pROC)

## Warning: package 'pROC' was built under R version 4.0.5

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:colorspace':
##
##     coords

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

prob=predict(modelo4, heart, type="response")
r=roc(heart$sex,prob, data=heart)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: Sin retraso = 0, Con retraso = 1
## Setting direction: sinretraso > conretraso
plot (r,print.auc=T,print.thres = "best",
          col="blue",xlab="1-ESpecificidad",ylab="Sensibilidad")
```



Evidenciando los resultados, podemos ver que el AUC tiene un valor de 70%, lo cual detalla que el modelo4 es medianamente óptimo para los análisis de predicción aunque la recomendación es que se acerque al 100% para ser considerado óptimo.

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

En este ejercicio se ha realizado un análisis de una base de datos con valores numéricos que hemos tenido primero que completar, y también se ha realizado un análisis de datos outliers. Es importante tener claro cómo se cumplimentan los valores vacíos y cómo tratar los outliers si fuera el caso, dado que influyen mucho en los análisis posteriores.

Posteriormente hemos realizado un análisis de las variables más significativas. El análisis nos ha llevado a concluir que no existe una relación lineal ni univariante ni

multivariante para concluir los datos. Es posible que para poder predecir correctamente las enfermedades tengamos que aplicar otros modelos u obtener mayor cantidad de datos que pueda ajustar el modelo para obtener mejor evidencia al caso de resolver.

En relación a la problematica planteada, se puede concluir que se pudo obtener los resultados requeridos, y es que **los pacientes masculinos tienen mayor influencia en enfermedades cardiacas que las mujeres**, para ello es de vital importancia realizar algun plan de acción comunitario sobre los pacientes del dataset de cleveland, con el fin de evitar desarrollos de patologias cardiacas a futuro en el supuesto caso, ya que el dataset fue recopilado en el año 1988.

Para futuros estudios. Segun la OMS, es importante que dichos pacientes tomen en consideración la reducción de malos hábitos como fumar, beber alcohol, no hacer ejercicio, comer grasas saturadas, entre otros. Mas info.. Es importante que se recogan nuevos datos para poder determinar la posición actual en relacion a enfermedades cardiovasculares.

Bibliografía

- Calvo M, Subirats L, Pérez D (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- OMS (2021). Enfermedades cardiovasculares. Link: https://www.who.int/cardiovascular_diseases/about_cvd/es/
- UCI (2021). Machine Learning Repository. Heart Disease. Link: https://archive.ics.uci.edu/ml/datasets/Heart+Disease