M2.851 – TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS: PRA 2

Morten Nyborg

Enero 2023

Contents

Introducción

Esta práctica ha sido realizada de forma individual por Morten Nyborg

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Descripción del dataset.

El dataset con el que trabajamos, https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset, proviene de un conjunto de datos más grande http://archive.ics.uci.edu/ml/datasets/Heart+Disease que contiene datos de riesgo de enfermedad cardiaca. Los datos contienen 303 registros y originalmente 75 atributos, aunque los datos más usados por los investigadores y por nosotros en esta práctica son los 14 atributos que describiremos más abajo.

Además esta tabla de datos, el dataset propuesto en kaggle tiene otro conjunto que tiene una sola columna con 3586 registros. Estos datos no encajan con los 303 registros que tenemos, así que no tomamos ninguna acción para integrarlos.

Mas abajo veremos los atributos y sus características con más detalle. Ahora proporcionamos una descripción básica de los atributos.

- **Age**, numérica, continua - Edad del paciente.
- **Sex**, categórica - Sexo del paciente. (1 = hombre; 0 = mujer).
- **cp**, categórica - Tipo de dolor de pecho:
  1: dolor de pecho típico angina. 2: dolor de pecho atípico. 3: otro dolor, no de pecho. 0: asintomático
- **trtbps**, numérica, continua - tensión arterial en reposo (mm Hg).
- **chol**, numérica, continua - colesterol en mg/dl, grabado con un sensor de IMC.
- **fbs**, numérica, continua - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false).
- **restecg**, categórica - resultados electrocardiograficos en reposo - 1 = normal. 2 = anomalías en las ondas ST-T. 0 = hypertrophy
- **thalach** - numérica, continua - valor máximo de puslasciones
- **exng** - categórica, el ejercicio causa dolor de pecho, 1 = sí 0 = no.
- **old peak**, numércia, continua - depresión ST (cuando el segmento ST está por debajo de lo normal) inducida por el ejercicio en relación al reposo.
- **slp** - categórica - pendiente en el ejercicio máximo del segmento ST. 2 = pendiente ascendente 1 = penndiente lisa 0 = pendiente descendetne
- **caa** - numérica, discreta - número de vasos sangúineos importantes (0-3).
- **thall** - thalassemia (1-3) - categórica - 0 = null. 1 = defecto fijo. 2 = normal. 3 = defecto reversible.
- **output** - categórica - diagnóstico de enfermedad cardiovascular 0: $< 50\%$ estrechamiento en el diámetro, menos posibilidad de enfermedad 1: $> 50\%$ estrechamiento del diámetro, más posibilidad de enfermedad

¿Por qué es importante el dataset?.

Con una búsqueda rápida en internet determinamos que la enfermedad cardíaca mata cada año a 697,000 personas en Estados Unidos, lo que supone 1 de cada 5 muertes. En España 15.6% de las muertes están causadas por enfermedades cardiacas. Por este motivo es importante conocer los factores que influyen en el desarrollo de enfermedades cardiovasculares. Además, los datos tienen información sobre el sexo de los pacientes, y es común pensar que los hombres sufren más que las mujeres de enfermedades cardiovasculares. Sin embargo, las mujeres sufren de enfermedades cardiovasculares casi tanto como los hombres, aunque los síntomas de la enfermedad pueden variar https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease/art-20046167.

Por este motivo, al tener datos de ambos sexos, podemos realizar varios análisis estadísticos que nos permitirán responder a varias preguntas relacionadas con las diferencias en la enfermedad cardiovascular para los dos sexos.

¿Qué problema pretende responder el dataset?.

El dataset dispone de información para responder a varias preguntas sobre los factores que influyen en la enfermedad cardiaca, por ejemplo, podríamos preguntarnos:
* ¿El colesterol alto causa enfermedades cardiovasculares?
* ¿La tensión arterial tiene alguna relación con enfermedades cardiovasculares?
* ¿Ser hombre es factor de riesgo para sufrir enfermedades cardiovasculares?
* ¿Cuales de los atributos del dataset tiene mayor influencia en el riesgo de sufrir enfermedades cardiovasculares?
Según el artículo de mayoclinic, la enfermedad cardiovascular en las mujeres se manifiesta de forma distinta que en los hombres, por ejemplo, la web afirma que en mujeres el dolor en el pecho no es tan común. De aquí podemos formular la pregunta: * ¿La proporción de mujeres que sufren enfermedad cardiaca y son asintomáticas es la misma que la proporción de hombres con enfermedad cardiaca y asintomáticos?

Podemos usar todas las variables del dataset para responder a preguntas, pero nos vamos a centrar en las preguntas siguientes:

- ¿Las mujeres tienen el mismo riesgo de padecer enfermedad cardiovascular que los hombres?
- ¿Quien tiene más riesgo de padecer enfermedades cardiovasculares: los hombres o las mujeres?
- ¿La edad influye en el riesgo de padecer enfermedad cardiovascular?
- ¿Las mujeres asintomáticas en cuanto al dolor de pecho tienen el mismo riesgo de padecer enfermedad cardiovascular que los hombres asintomáticos?
- ¿La edad, el sexo, la existencia o no de angina de pecho son factores que influyen en el diagnóstico de enfermedad cardiovascular?

Con estas preguntas podemos quedarnos solamente con las variables del dataset que nos interesan. Al principio analizamos y visualizaremos todas las variables y comentaremos brevemente las visualizaciones. Despues seleccionaremos un subconjunto de datos.

Carga del dataset:

Leemos los datos y mostramos las primeras filas.

```
# Carga de el archivo que contiene los datos.
heart_attack <- read.csv("../data/heart.csv")
```

Integración y selección de los datos

Integración de los Datos

La integración de datos forma parte de la limpieza de datos y tiene como objetivo integrar datos de varias fuentes para crear una estructura única y coherente que será usada en los análisis posteriores.

La integración puede ser horizontal o vertical. En el primer caso se añaden atributos al juego de datos, en el segundo caso se añaden más instancias al juego de datos.

Los datos que tenemos para esta práctica consisten de dos partes: el dataset descrito más arriba, y un fichero csv con un valor por fila con saturaciones de oxígeno. Los números de registros en los dos datasets son muy diferentes y el propietario de los datos no proporciona suficiente información para encontrar la relación que exsite entre los dos conjuntos.

Selección de los Datos.

Screening de los datos. En este apartado hacemos la selección de datos. Una de las tareas de la selección es screening: ver la estructura de los datos, ver las características de cada atributo y identificar correlaciones si las hay, eliminar información redundante.

Los datos que usamos tienen 303 filas y 14 columnas.

```
dim(heart_attack)
```

```
## [1] 303  14
```

Vemos las columnas que tiene el dataset.

Usamos ´sapply´ para ver los tipos de variable de cada columna. Todas las variables son numércas, la mayoría enteros, excepto oldpeak.

```
options(width = 80)
#list types for each variable
sapply(heart_attack, class)
```

```
##       age       sex        cp    trtbps      chol       fbs   restecg  thalachh
## "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
##      exng   oldpeak       slp       caa     thall    output
## "integer" "numeric" "integer" "integer" "integer" "integer"
```

Aunque las variables son interpretadas por R como enteros, no estamos de acuerdo con esto, las variables que consideramos realmente numéricas son age,trtbps, chol, thalachh, oldpeak. Las variables sex, cp, fbs, restecg, exng, slp, caa son variables categóricas, por esto las transformamos en factores.

```
num.var <-heart_attack %>%
  dplyr::select("age","trtbps","chol","thalachh","oldpeak") # numerical

cat.var<- heart_attack %>%
  dplyr::select("sex","cp","fbs","restecg","exng","slp","caa",
         "thall","output")%>%
  mutate_if(is.numeric, as.factor) #categorical variables.

#Put together categorical  and numerical variables again.
heart_attack1 = cbind(cat.var,num.var)
```

Después de haber transformado algunas variables en factores, mostramos el tipo de datos que corresponden a cada variable. Sexo es un factor de dos niveles, cp un factor de 4 niveles, etc. También podemos ver las variables numércicas.

```
str(heart_attack1)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
##  $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
##  $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
##  $ exng    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
##  $ slp     : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
##  $ caa     : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ thall   : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
##  $ output  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
```

Usamos el comando ´summary´para mostrar las estadísticas de cada atributo. R calcula el mínimo, máximo, media, mediana, cuantiles para las variables numéricas y un recuento de instancias que pertenecen a cada nivel del factor para los atributos cualitativos.

Algunas interpretaciones: las edades de las personas están entre 29 y 77 años, la media es cercana a la mediana.

Los datos contienen muchas menos mujeres que hombres y más casos de enfermedad que de no enfermedad.

```
options(width = 80)
summary(heart_attack1)
```

```
##  sex     cp      fbs     restecg exng    slp     caa     thall   output
##  0: 96   0:143   0:258   0:147   0:204   0: 21   0:175   0:  2   0:138
##  1:207   1: 50   1: 45   1:152   1: 99   1:140   1: 65   1: 18   1:165
##          2: 87           2:  4           2:142   2: 38   2:166
##          3: 23                                   3: 20   3:117
##                                                  4:  5
##
##       age             trtbps           chol         thalachh        oldpeak
##  Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0   Min.   :0.00
##  1st Qu.:47.50   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.5   1st Qu.:0.00
##  Median :55.00   Median :130.0   Median :240.0   Median :153.0   Median :0.80
##  Mean   :54.37   Mean   :131.6   Mean   :246.3   Mean   :149.6   Mean   :1.04
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:166.0   3rd Qu.:1.60
##  Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0   Max.   :6.20
```

Duplicados

Vemos si hay alguna línea duplicada en los datos y detectamos una. Consideraremos esta entrada como duplicada, aunque podría ser una coincidencia que dos pacientes tengan exactamente las mismas lecturas. Originalmente los datos tenían información del paciente, como el número de la seguridad social, estos datos han sido eliminados y ya no tenemos forma de saber si tenemos un duplicado verdadero.

Tratare,os el duplicado más adelante.

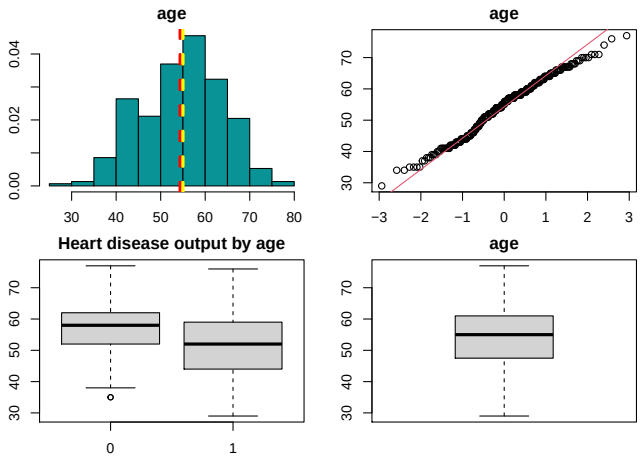
```
heart_attack1[duplicated(heart_attack1),]
```

```
##     sex cp fbs restecg exng slp caa thall output age trtbps chol thalachh
## 165   1  2   0       1    0   2   4     2      1  38    138  175      173
##     oldpeak
## 165       0
```

Vamos a crear una función para hacer visualizaciones para distintas variables. La función mostrará un histograma, un qq plot, dos box plots: uno separado por el resultado (enfermo o no) y otro de la variable en general.

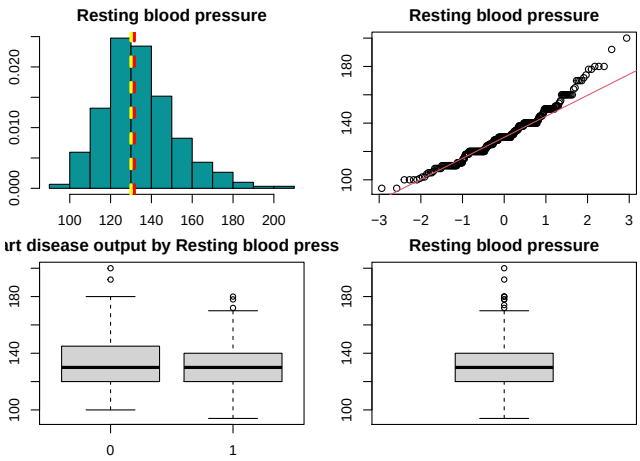
En primer lugar vemos el box plot de la variable edad. Vemos que la media está cerca de la mediana, en el qq plot la distribución está bastante ajustaa a la línea diagonal, parece que la edad tiene una distribución normal, o se parece. Lo comprobaremos con un test más tarde.

En los box plots vemos que las personas que tienen el resultado (output) de poca posibilidad de enfermar tienen una mediana mayor. La mediana de edad en las personas que enferman es menor que la de los que no enferman.



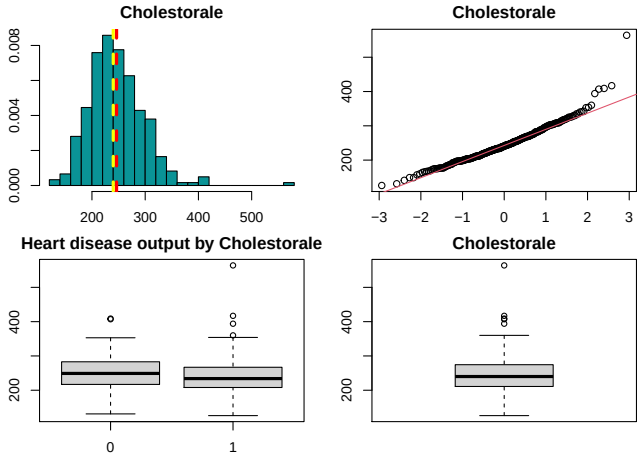
Otra variable numércia es la presión sanguínea. La distribución tiene una cola a la derecha, hay algunas personas que tienen la presión muy alta y esto debe ser lo que vemos en los outlierrs del boxplot. en el qq plot los puntos se ajustan bastante a la mayor parte de la línea diagonal, algo menos por los extremos. Estos puntos también podrían pertenecer a los outliers.

En los box plots, el caso de enfermedad: sí tiene una presión de sangre mediana igual que en enfermedad: no. Las personas con riesgo bajo de enfermedad tienen la presión sanguínea en rangos más amplios, los valores entre los cuantiles 25 y 75 están entre 120 y 145 aproximadamente. Las personas con más riesgo de enfermar tienen la presión entre 120 y 140.



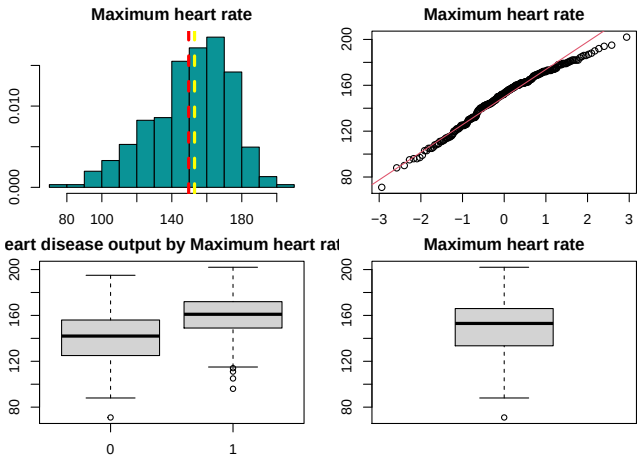
El colesterol tiene una distribución también parecida a la normal, por el histograma y el qq plot, con la media y la mediana con valores parecidos, también hay una cola a la derecha, que significa que hay personas con colesterol muy alto. Tendremos que investigar si el valor tan alto es posible y encontramos que un valor de más de 240 es un riesgo, aquí tenemos algún valor por encima de 500 que consideramos no posible.

En el boxplot vemos algo interesante, es que las personas que tienen riesgo de enfermar tienen un colesterol mediano algo más bajo que el de las personas con menos riesgo de enfermar.

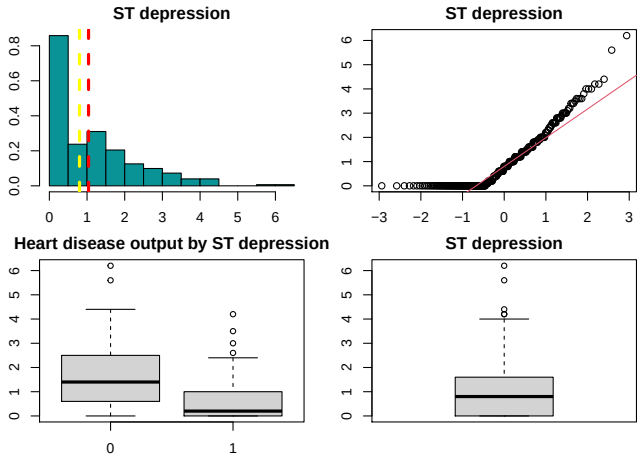


Los datos de frecuencia cardiaca tienen algo más de variación entre la media y la mediana, la distribución está algo inclinada a la derecha. En el qqplot también se ajusta bien en la mayoría de la línea.

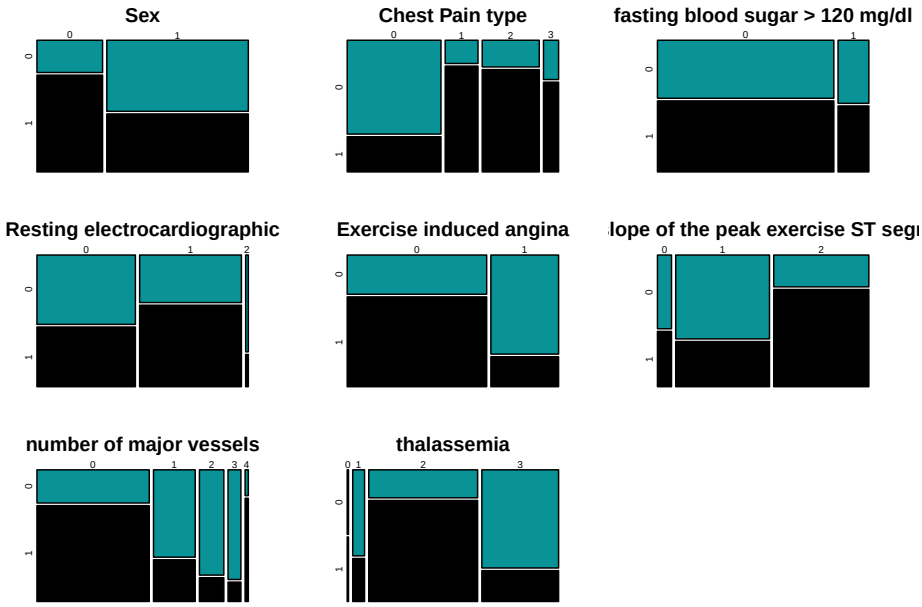
Aquí hay una diferencia sustancial entre los boxplots de riesgo de enfermar y poco riesgo de enfermar. Las personas sin riesgo de enfermar tienen una frecuencia cardíaca mediana de 140, las personas con mas riesgo tienen la frecuencia mediana de 160.



La variable ST depression se distribuye principalmente al inicio del eje x, la mayoría de personas tiene st depression cercana a cero, muy pocos llegan a valores altos de esta variable. Esta distribución no podemos considerarla como normal, el qq plot tiene muchos valores alejados de la línea y el histograma no se parece nada a una distribución normal. El valor mediano de esta variable esm uy bajo para personas que tienen riesgo de enfrermar.



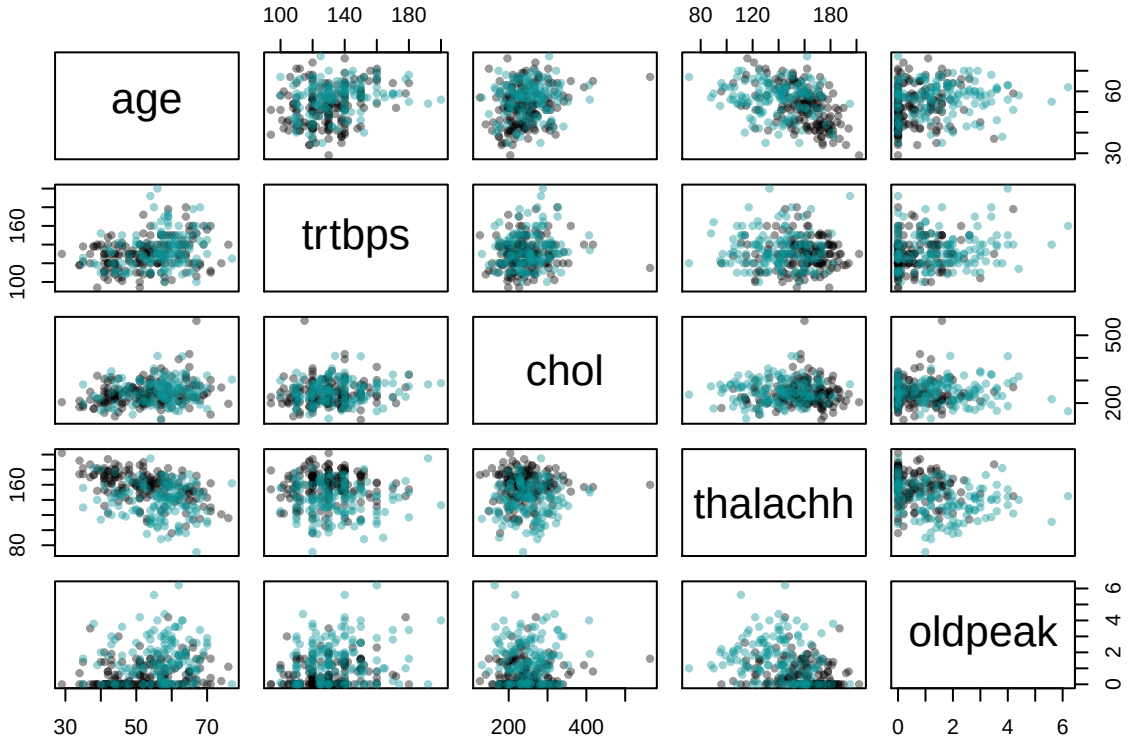
Ahora visualizamos los datos categóricos.



Algunas interpretaciones:

- Usamos el color negro para riesgo de enfermedad. El sexo mujer parece tener más riesgo de enfermedad cardiovascular que el sexo masculino.
- Los síntomas de dolor en el pecho muestran riesgo de enfermedad, también hay casos asintomáticos con riesgo.
- El azúcar en sangre en ayuno es bastante parecido para casos de enfermedad y de no enfermedad.
- El electrocardiograma es más alto en casos de posible enfermedad.
- Las personas que no tienen dolor de pecho inducido por ejercicio tienen riesgo de enfermedad.
- Las personas con el valor de thalassemia de 2 tienen más riesgo de enfermedad que las otras personas.

También vemos las nubes de puntos por pares de variables, para ver si detectamos algunos patrones. Usamos pairplot.



En color negro mostramos el riesgo de enfermedad.

Edad:
* con thalachh - las personas con edades bajas, con talachh altos tienen riesgo de enfermedad.
* colesterol - personas de edad baja y colesterol bajo tienen riesgo.

trtbps:
* con thalachh - personas con trtbps medio o bajo y thalachh alto tienen riesgo.

thalachh:
* oldpeak - personas con thalachh medio alto y oldpeak bajo tienen riesgo.

Eliminación del duplicado. Finalmente, ahora que hemos analizado todas las variables, nos quedamos con las mencionadas al inicio, pero antes comprobamos si hay un duplicado, lo encontramos y lo eliminamos.

```
heart_attack1[duplicated(heart_attack1),]
```

```
##     sex cp fbs restecg exng slp caa thall output age trtbps chol thalachh
## 165   1  2   0       1    0   2   4     2      1  38    138  175      173
##     oldpeak
```

```
## 165       0
```

```
heart_attack1 <- heart_attack1[!duplicated(heart_attack1), ]
```

```
heart_attack1[duplicated(heart_attack1),]
```

```
##  [1] sex      cp       fbs      restecg  exng     slp      caa      thall
##  [9] output   age      trtbps   chol     thalachh oldpeak
## <0 rows> (or 0-length row.names)
```

Aquí vamos a seleccionar las variables que usaremos para los tests y modelos. Estas variables son age, sex, cp, output. Además creamos una vraible dicotómica que tendrá valor 0 para observaciones con dolor en el pecho y valor 1 para observaciones sin dolor en el pecho.

```
final_vars <- c('age', 'sex', 'cp', 'output')
heart_attack2 <- heart_attack1[,final_vars]
```

Creación de nuevas variables

Creamos una variable para dividir los casos de dorol de pecho: usamos 0 para los asintomáticos y en 1 recogemos los otros tres tipos de dolor de pecho.

```
#add categorical variable named 'asymptomatic' using values from 'cp' column
heart_attack2$asymptomatic <- as.factor(ifelse(heart_attack2$cp == 0, 1, 0))
head(heart_attack2,3)
```

```
##   age sex cp output asymptomatic
## 1  63   1  3      1            0
## 2  37   1  2      1            0
## 3  41   0  1      1            0
```

```
head(heart_attack2, 3)
```

```
##   age sex cp output asymptomatic
## 1  63   1  3      1            0
## 2  37   1  2      1            0
## 3  41   0  1      1            0
```

Limpieza de datos

¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Antes de seleccionar los datos habíamos encontrado un duplicado y lo eliminamos en el apartado anterior. Ahora si buscamos duplicados vamos a encontrar muchos porque la mayoría de variables, excepto edad son discretas y tienen pocos posibles valores.

Vemos cuales son los valores únicos en cada variable. Encontramos edades entre 29 y 77 años, 2 niveles para sexo, 4 niveles (de 0 a 3 para cp), más detalles abajo.

```
sapply(heart_attack2, function(x) unique(sort(x)))
```

```
## $age
##  [1] 29 34 35 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
## [26] 59 60 61 62 63 64 65 66 67 68 69 70 71 74 76 77
##
## $sex
## [1] 0 1
## Levels: 0 1
##
## $cp
## [1] 0 1 2 3
```

```
## Levels: 0 1 2 3
##
## $output
## [1] 0 1
## Levels: 0 1
##
## $asymptomatic
## [1] 0 1
## Levels: 0 1
```

Los datos que hemos seleccionado para responder a nuestras preguntas no tienen ceros y elementos vacíos. En el caso de que hubiera, podríamos decidir entre eliminar los registros que tengan elementos vacíos o sino imputar los valores vacíos.En el ejemplo del dataset antes de seleccionar las variables que vamos a usar tenemos dos casos de nulos: la variable thall según la descripción de los datos no puede tomar el valor cero, el cero corresponde a null, la variable caa no puede tomar valor 4 porque según la descripción puede tomar valores de 0 a 3.

Aunque no vamos a usar los datos, podemos hacer una demostración del paquete mice para imputar datos categóricos. En primer lugar sustituimos los valores que no tienen sentido por un NA, después comprobamos que hay valores NA. Iniciamos la imputaciòn y usamos el método polyreg, regresión polinomial.

```
heart_attack1[heart_attack1$caa == 4 | heart_attack1$thall == 0,]
```

```
##     sex cp fbs restecg exng slp caa thall output age trtbps chol thalachh
## 49    0  2   0       0    0   2   0     0      1  53    128  216      115
## 93    1  2   0       1    0   2   4     2      1  52    138  223      169
## 159   1  1   0       1    0   1   4     3      1  58    125  220      144
## 164   1  2   0       1    0   2   4     2      1  38    138  175      173
## 252   1  0   1       0    1   1   4     3      0  43    132  247      143
## 282   1  0   1       1    1   1   0     0      0  52    128  204      156
##     oldpeak
## 49      0.0
## 93      0.0
## 159     0.4
## 164     0.0
## 252     0.1
## 282     1.0
```

```
#df["pages"][df["pages"] == 0] <- NA
```

Como resultado de imputar los valores tenemos un dataframe sin ningún valor NA. Mice ha ajustado una regresión para imputar los valores a partir del resto de los datos.

```
library(mice)

heart_attack1["thall"][heart_attack1["thall"] == 0] <- NA
heart_attack1["caa"][heart_attack1["caa"] == 4] <- NA
sapply(heart_attack1, function(x) sum(is.na(x)))
```

```
##      sex       cp      fbs  restecg     exng      slp      caa    thall
##        0        0        0        0        0        0        4        2
##   output      age   trtbps     chol thalachh  oldpeak
##        0        0        0        0        0        0
```

```
init <- mice(heart_attack1)
```

```
##
##  iter imp variable
##   1   1  caa  thall
##   1   2  caa  thall
##   1   3  caa  thall
```

```
##   1   4  caa  thall
##   1   5  caa  thall
##   2   1  caa  thall
##   2   2  caa  thall
##   2   3  caa  thall
##   2   4  caa  thall
##   2   5  caa  thall
##   3   1  caa  thall
##   3   2  caa  thall
##   3   3  caa  thall
##   3   4  caa  thall
##   3   5  caa  thall
##   4   1  caa  thall
##   4   2  caa  thall
##   4   3  caa  thall
##   4   4  caa  thall
##   4   5  caa  thall
##   5   1  caa  thall
##   5   2  caa  thall
##   5   3  caa  thall
##   5   4  caa  thall
##   5   5  caa  thall
```

```
meth <- init$method
predM <- init$predictorMatrix


meth[c("thall")]="polyreg"
meth[c("caa")]="polyreg"
imputed <- mice(heart_attack1, method=meth, predictorMatrix=predM, m=5)
```

```
##
##  iter imp variable
##   1   1  caa  thall
##   1   2  caa  thall
##   1   3  caa  thall
##   1   4  caa  thall
##   1   5  caa  thall
##   2   1  caa  thall
##   2   2  caa  thall
##   2   3  caa  thall
##   2   4  caa  thall
##   2   5  caa  thall
##   3   1  caa  thall
##   3   2  caa  thall
##   3   3  caa  thall
##   3   4  caa  thall
##   3   5  caa  thall
##   4   1  caa  thall
##   4   2  caa  thall
##   4   3  caa  thall
##   4   4  caa  thall
##   4   5  caa  thall
##   5   1  caa  thall
##   5   2  caa  thall
##   5   3  caa  thall
##   5   4  caa  thall
```

```
##   5   5  caa  thall
```

```
imputed <- complete(imputed)
sapply(imputed, function(x) sum(is.na(x)))
```

```
##      sex       cp      fbs  restecg     exng      slp      caa    thall
##        0        0        0        0        0        0        0        0
##   output      age   trtbps     chol thalachh  oldpeak
##        0        0        0        0        0        0
```

Elementos vacíos en el dataset, elementos iguales a cero

El dataset no tiene elementos vacíos con esta selección de datos. En el apartado anterior hemos introducido elementos vacíos en lugar de algunos valores atípicos, a modo de ejemplo, para poder imputar estos valores con el paquete mice. Una vez imputados los valores, no han quedado elementos vacíos.

En cuanto a los valores iguales a cero, ya hemos visto que alguns variables toman el valor cero, pero estos valores son correctos. No hemos visto otra variable que tome valores fuera de lo normal para tener que tomar medidas sobre ellas.

```
# Estructura de los datos - Type = 2

kbl(ExpData(data=heart_attack2,type=2),booktabs =T)%>%
  kable_styling(latex_options =c("striped","scale_down", "hold_position"))
```

Index	Variable_Name	Variable_Type	Sample_n	Missing_Count	Per_of_Missing	No_of_distinct_values
1	age	integer	302	0	0	41
2	sex	factor	302	0	0	2
3	cp	factor	302	0	0	4
4	output	factor	302	0	0	2
5	asymptomatic	factor	302	0	0	2

Identifica y gestiona los valores extremos.

La función que usamos antes para ver las distribuciones devuelve las estadísticas del boxplot, con esto podemos encontrar los outliers de cada una de las variables numéricas con las que nos hemos quedado. Podemos ver que algunas de las variables tienen valores fuera de lo común, pero como no vamos a usar estas variables, no vamos a tomar ninguna acción para imputar valores.
La única variable numérica que vamos a usar es age, y no tiene ningún duplicado ni valores que parezcan extremos. No necesitamos tomar ninguna acción.

```
boxplot.age$out
```

```
## numeric(0)
```

```
boxplot.chol$out
```

```
## [1] 417 564 394 407 409
```

```
boxplot.bp$out
```

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

```
boxplot.heartrate$out
```

```
## [1] 71
```

```
boxplot.stdep$out
```

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

Análisis de los datos.

Como hemos dicho en la introducción queremos responder a las preguntas: * ¿Las mujeres tienen el mismo riesgo de padecer enfermedad cardiovascular que los hombres? * ¿Quien tiene más riesgo de padecer enfermedades cardiovasculares: los hombres o las mujeres? * ¿La edad influye en el riesgo de padecer enfermedad cardiovascular? * ¿Las mujeres asintomáticas en cuanto al dolor de pecho tienen el mismo riesgo de padecer enfermedad cardiovascular que los hombres asintomáticos? * ¿La edad, el sexo, la existencia o no de angina de pecho son factores que influyen en el diagnóstico de enfermedad cardiovascular?

Como tenemos límite de espacio, nos limitamos a esto.

Selección de los grupos de datos que se quieren analizar/comparar (p. ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

Creamos un subconjuntos de datos para las pruebas. En primer lugar, vamos a testear si las mujeres tienen tanto riesgo de enfermedad cardiovascular como los hombres. Es habitual pensar que las mujeres tienen menos riesgo, pero los gráficos que hemos hecho muestran que las mujeres tienen más riesgo.

Para esta prueba seleccionamos unos datos que tengan solo personas con riesgo de enfermedad, a partir de ahí sacamos un dataset para hombres y uno para mujeres.

```
heart.disease <- heart_attack2[heart_attack2$output==1,]

heart.disease.male <- heart.disease[heart.disease$sex==1,]
heart.disease.female <- heart.disease[heart.disease$sex==0,]
```

Para aplicar la regresión logística guardamos particiones de entrenamiento y test. Aquí usamos datos de personas con y sin riesgo. Aunque la validación cruzada que vamos a usar con la regresión logística ya hace la partición en entrenamiento y test, reservamos un 10% de los datos para hacer una predicción y calcular la matriz de confusión.

```
library(tidyverse)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(42)
# guarda las líneas que son para train, hace un subset con estas líneas
training.samples <- heart_attack2$cp %>%
  createDataPartition(p = 0.9, list = FALSE)
train.data  <- heart_attack2[training.samples, ]
```

```
# subset del os datos con las líneas que no son de train
test.data <- heart_attack2[-training.samples, ]
```

```
head(train.data, 3)
```

```
##   age sex cp output asymptomatic
## 1  63   1  3      1            0
## 2  37   1  2      1            0
## 3  41   0  1      1            0
```

```
head(test.data,3)
```

```
##    age sex cp output asymptomatic
```

```
## 26   71    0   1       1             0
## 31   41    0   1       1             0
## 32   65    1   0       1             1
```

Comprobación de la normalidad y homogeneidad de la varianza.

Un requisito previo para aplicar un contrastee de hipótesis paramétrico es que la media muestral siga una distribución normal y la varias sea homocenastica. Haremos las comprobaciones a continuación.

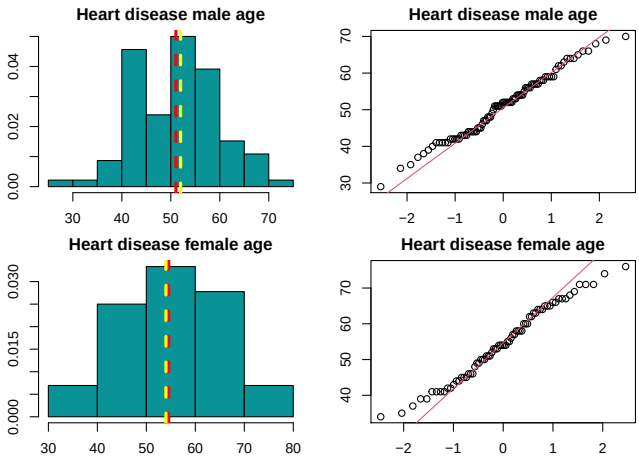
Creamos una función que va a hacer un grupo de plots para ayudarnos a testear la normalidad.

```
plot.data.norm <- function(df1, df2, name.df1, name.df2, name.var){
oldpar = par(mfrow = c(2,2), mar=c(2,2,2,2))
truehist(df1, main = paste(name.df1, name.var), col = "#0a9396")
abline(v = mean(df1), col="red", lwd=3, lty=2);
abline(v = median(df1), lwd=3, lty=2, col="yellow");
qqnorm(df1, main = paste(name.df1, name.var));qqline(df1, col = 2 )

truehist(df2, main = paste(name.df2, name.var), col = "#0a9396")
abline(v = mean(df2), col="red", lwd=3, lty=2);
abline(v = median(df2), lwd=3, lty=2, col="yellow");
qqnorm(df2, main = paste(name.df2, name.var));qqline(df2, col = 2 )
}
```

Inspección visual de normalidad

Hacemos histogramas parecidos a los anteriores, esta vez mostramos nada más las personas que sí tienen riesgo de enfermedad, por sexo. Los hombres son muchos más que las mujeres en el dataset, la distribución de las edades con alguna excepción se parece a la normal. Tenemos una muestra bastante grande de hombres y de mujeres, podríamos aplicar el teorema central del límite que dice que la media muestral sigue una distribución normal si la muestra es lo bastante grande, aunque la población de origen no siga esta distribución. Para asegurarnos aplicamos un test de normalidad.



Contraste de normalidad de Lilliefors El test de Kolmogorov-Smirnov plantea la hipótesis nula de normalidad y la hipótesis alternativa de no normalidad. Tenemos un p value mayor que 0.05 en ambos casos(mujeres p=0.3939 y hombres p=0.08653) y esto permite que aceptemos

$$H_0$$

por la que la variable sigue una distribución normal.

```
lillie.test(heart.disease.female$age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```
## data:  heart.disease.female$age
## D = 0.07548, p-value = 0.3939
```

```
lillie.test(heart.disease.male$age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  heart.disease.male$age
## D = 0.086439, p-value = 0.08653
```

Una opción más conservadora para testear la normalidad es la prueba de Shapiro-Wilk que se considera más robusta. Hacemos esta prueba y también tenemos un pvalue que nos permite afirmar con un nivel de confianza del 95% que tanto la media de edad de los hombres como la media de edad de las mujeres siguen una distribución normal.

```
shapiro.test(heart.disease.female$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heart.disease.female$age
## W = 0.97977, p-value = 0.2991
```

```
shapiro.test(heart.disease.male$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heart.disease.male$age
## W = 0.98619, p-value = 0.4455
```

El test de la homocedasticidad también es un contraste de hipótesis que indica si las varianzas entre dos grupos son iguales o no. En la hipótesis 0 se asume que las varianzas son iguales y en la H1 que son diferentes. Un p value menor que el nivel de significancia 0.05 permite rechazar H0.

-Las varianzas en la edad de los hombres y las mujeres son iguales y podemos hacer esta afirmación con un nivel de confianza del 95%.

```
var.test(heart.disease.male$age, heart.disease.female$age)
```

```
##
##  F test to compare two variances
##
## data:  heart.disease.male$age and heart.disease.female$age
## F = 0.70577, num df = 91, denom df = 71, p-value = 0.1169
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4508303 1.0914650
## sample estimates:
## ratio of variances
##           0.705767
```

Para hacer el test de Levene que nos proporcionan los materiales de clase tenemos que usar los datos sin partir, comprobamos si la varianza de la edad es igual para los dos sexos, un valor pvalue alto indicaría que estamos cometiendo un error al descartar

$$H_0$$

que indica la igualdad de varianzas (homocedasticidad). En este caso, igual que con el var.test tenemos el resultado de pvalue alto, por lo tanto podemos afirmar con un nivel de confianza del 95% que las varianzas son homocedasticas.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
leveneTest(heart_attack2$age ~ heart_attack2$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.4435 0.5059
##       300
```

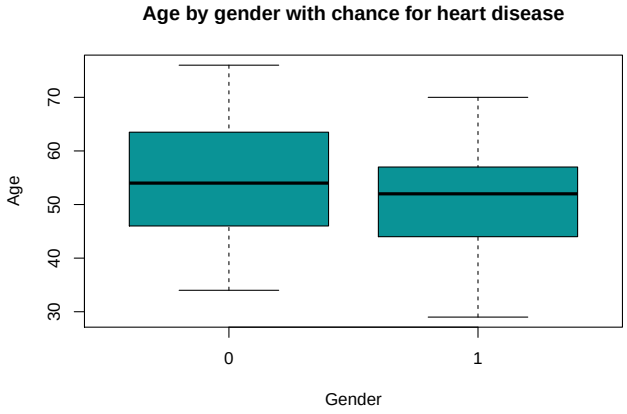
Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Prueba 1. Contrastes de hipótesis

Contraste sobre la media "Nos preguntamos si las mujeres con riesgo de enfermedad cardiovascular tienen la misma edad media como los hombres"

Análisis visual Visualizamos los datos de la edad de hombres y mujeres. El valor de la mediana para mujeres con riesgo de enfermedad cardiovascular es un poco más alta que la de los hombres. Las edades que se encuentran en las cajas (entre cuantiles 25 y 75) tienen mayor rango para las mujeres.

Las mujeres con riesgo que entran entre los cuantiles 25 y 75 tienen edades de entre 45 y 65 aproximadamente. Las edades para los hombres entre estos cuantiles están entre algo más de 40 y 55 aproximadamente.



La hipótesis nula y la alternativa

¿Podemos aceptar que los hombres con riesgo cardiovascular tienen una edad media mayor que las mujeres con riesgo de cardiovascular?

$$H_0$$

: La edad media de los hombres = la edad media de las mujeres

$$H_1$$

: La edad media de los hombres > la edad media de las mujeres

Justificación del test a aplicar * Tenemos dos muestras independientes: hombres y mujeres y realizamos un contraste de hipótesis de dos muestras. * Un contraste sobre la media. * Las varianzas poblacionales son desconocidas * Test unilateral por la derecha. * Según el Teorema del límite central asumimos la distribución normal de la media muestral (muestra mayor que 30) * Las varianzas muestrales son iguales, esto lo sabemos del test de homocedasticidad aplicado antes * Con todo lo dicho, aplicamos un Test paramétrico.

```
t.test(heart.disease.male$age, heart.disease.female$age, alternative = "greater", var.equal = TRUE)
```

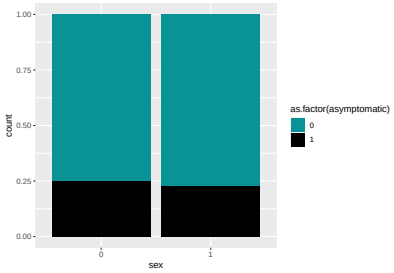
```
##
##  Two Sample t-test
##
## data:  heart.disease.male$age and heart.disease.female$age
## t = -2.3799, df = 162, p-value = 0.9908
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -5.953391       Inf
## sample estimates:
## mean of x mean of y
##  51.04348  54.55556
```

El test nos da un pvalue cercano a 1, esto quiere decir que no podemos rechazar H0, que dice que la edad media de los hombres y de las mujeres con riesgo de enfermedad cardiovascular es igual. Podemos afirmar esto con un nivel de confianza del 95 porciento.

Contraste sobre la proporción Proporciones de hombres y mujeres con riesgo de enfermedad cardiovascular que no tienen ningún dolor en el pecho.

Uno de los artículos que usamos para introducir el tema mencionaba que los síntomas de la enfermedad cardiovascular en hombres y mujeres son diferentes, las mujeres no tienen tanto de dolor en el pecho como los hombres, según el artículo. Por este motivo, vamos a comprobar si hay alguna diferencia, usando un contraste sobre la proporción.

Visualizamos la variable Asymtomatic según género, si(=1) o no(=0). La visualización muestra que los asintomáticos con riesgo de enfermedad tienen una proporción parecida entre los hombres y las mujeres.



Hipótesis nula y alternativa ¿La proporción de personas que no tienen ningún síntoma de dolor en el pecho es diferente para hombres que para mujeres? La hipótesis nula dice que la proporción de asintomáticos con enfermedad cardiovascular es la misma entre hombres y mujeres, mientras que la H1 dice que la proporción no es igual entre los sexos.

$$H_0$$

:

$$p_{Hombres} = p_{Mujeres}$$

La proporción de personas sin síntoma de dolor en el pecho es la misma para hombres que para mujeres

$$H_1$$

:

$$p_{Hombres} \neq p_{Mujeres}$$

La proporción de personas con síntoma de dolor en el pecho es diferente para hombres que para mujeres.

Tipo de test * Tenemos dos muestras independientes: la proporción de hombres sin dolor en el pecho y con enfermedad y la proporción de mujeres sin dolor en el pecho y con enfermedad y realizamos un contraste de hipótesis de dos muestras. * Un contraste sobre la proporción. * test bilateral la proporción es igual o no. * Asumimos que las proporciones muestrales presentan una distribución aproximadamente normal (según el teorema central del límite) ya que las muestras son lo suficientemente grandes. * Por el tamaño de la muestra podemos asumir que la desviación estándar para cada sexo se calcula por `sqrt(prop_sexo(1-prop_sexo) / observ_sexo)`

```
# Calculate number of observations for men.
obs.male <- nrow(heart.disease.male)
# Calculate number of observations for women.
obs.female <- nrow(heart.disease.female)

# Calculate the proportion asymptomatic men.
p.male <- sum(heart.disease.male$asymptomatic==1) / obs.male
# Calculate the proportion asymptomatic women
p.female <- sum(heart.disease.female$asymptomatic==1) / obs.female
```

Usamos el test sobre la proporción de R con un nivel de significancia del 95 porciento. El estadístico pvalue es mayor que el nivel de significancia y no podemos rechazar la hipótesis nula, por esto podemos afirmar que la proporción de hombres que no tienen dolor en el pecho y tienen riesgo de enfermedad cardiovascular es la misma que la proporción de mujeres sin dolor en el pecho con riesgo de enfermedad cardiovascular. Afirmamos esto con un nivel de confianza del 95 porciento.

```
success <- c(p.male*obs.male, p.female*obs.female)
nn <- c(obs.male, obs.female)

prop.test(x=success, n=nn, alternative="two.sided", correct=FALSE, conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 0.10531, df = 1, p-value = 0.7455
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1534937  0.1100154
## sample estimates:
##    prop 1    prop 2
## 0.2282609 0.2500000
```

Prueba 2. chisq.test.

Aplicamos el test chisq.test para comprobar si existe asociación entre la variable output y las variables sex y asymptomatic. El test tiene como hipótesis nula que no existe asociación entre las variables testeadas y para concluir que sí existe asociación tendríamos que rechazar

$$H_0$$

.

Creamos las tablas de contingencia entre la variable output y las variables sex y asymptomatic.

```
# Contingency table output vs sex
table.1 <- table(heart_attack2$output, heart_attack2$sex)
table.1
```

```
##
##       0   1
##   0  24 114
```

```
##   1  72  92
```

```
# Contingency table output vs RS_cat2.
table.2 <- table(heart_attack2$output, heart_attack2$asymptomatic)
table.2
```

```
##
##       0   1
##   0  34 104
##   1 125  39
```

```
chisq.test(table.1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table.1
## X-squared = 23.084, df = 1, p-value = 1.551e-06
```

```
chisq.test(table.2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table.2
## X-squared = 77.926, df = 1, p-value < 2.2e-16
```

El test chi al cuadrado comprueba si hay asociación entre variables categóricas. Las variables que testeamos son sex y asyntomatic, y los dos tests nos dan un pvalue por debajo de 0.05, esto significa que sí hay relación entre estas varaibles y la variable objetivo.
El resultado debajo toma como referencia el sexo femenino y el hombre tiene 0.27% de posibilidades de enfermar de las que tiene una mujer, tiene menos posibilidades de enfermedad.

De la misma forma, en el siguiente odds ratio vemos que ser asintomático es factor de protección frente a la posibilidad de enfermar.

```
oddsratio(table.1)
```

```
## $data
##
##          0   1 Total
##   0     24 114   138
##   1     72  92   164
##   Total 96 206   302
##
## $measure
##    odds ratio with 95% C.I.
##      estimate     lower     upper
##   0 1.0000000        NA        NA
##   1 0.2710974 0.1556892 0.4593234
##
## $p.value
##    two-sided
##       midp.exact fisher.exact   chi.square
##   0           NA           NA           NA
##   1 6.384415e-07 1.009872e-06 8.281875e-07
##
## $correction
## [1] FALSE
##
## attr(,"method")
```

```
## [1] "median-unbiased estimate & mid-p exact CI"
```

```
oddsratio(table.2)
```

```
## $data
##
##           0   1 Total
##   0      34 104   138
##   1     125  39   164
##   Total 159 143   302
##
## $measure
##    odds ratio with 95% C.I.
##      estimate      lower     upper
##   0 1.0000000         NA        NA
##   1 0.1032723 0.06000774 0.1734141
##
## $p.value
##    two-sided
##     midp.exact fisher.exact   chi.square
##   0         NA           NA           NA
##   1          0 1.007752e-19 3.779078e-19
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "median-unbiased estimate & mid-p exact CI"
```

Prueba 3. Regresión logísitica

Vamos a realizar una serie de regresiones logísticas con las variables que hemos usado en los contrastes anteriores para confirmar lo que ya mehos visto. En primer lugar vemos en qué medida el riesgo de enfermedad está relacionado con el sexo.

```
# Build the model.
model1 <- glm(formula = output ~ sex, family = binomial(link = logit), data = train.data)
# Summary of model.
summary(model1)
```

Explicar la enfermedad a partir del sexo.

```
##
## Call:
## glm(formula = output ~ sex, family = binomial(link = logit),
##     data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6259  -1.0903   0.7876   1.2671   1.2671
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0116     0.2384   4.244 2.20e-05 ***
## sex1         -1.2200     0.2809  -4.343 1.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 376.52  on 272  degrees of freedom
## Residual deviance: 356.10  on 271  degrees of freedom
## AIC: 360.1
##
## Number of Fisher Scoring iterations: 4
```

El resultado muestra asteriscos *** en la parte de los coeficientes, esto indica que la variable sex1 es significativa. En presencia de sex1 (hombre), el riesgo de tener enfermedad cardiovascular disminuye. El estadístico Pr(>|z|) con un valor por debajo de 0.05 confirma esto.

Notamos que el valor de AIC es 360, que no tiene mucho valor en este momento, pero será útil para comparar los modelos .

Debajo calculamos el ODDS ratio. Para el sexo 1 (hombre) está por debajo de cero, esto significa que ser hombre es un factor de protección de la enfermedad cardiovascular.

Con el intervalo de confianza que muestra el resultado podemos decir que el 95% de las OR estarán en el intervalo de 0.17 y 0.51.

```
# Odds ratio
exp(coefficients(model1))
```

```
## (Intercept)        sex1
##   2.7500000   0.2952295
```

```
exp(cbind(coef(model1), confint(model1)))
```

```
##                            2.5 %    97.5 %
## (Intercept) 2.7500000 1.7492876 4.4733966
## sex1        0.2952295 0.1677977 0.5063592
```

Explicar la enfermedad a partir del sexo y la edad. Hacemos otra regresión logística añadiendo a la variable anterior la variable que representa la edad.

```
# Build the model.
model2 <- glm(formula = output ~ sex + age, family = binomial(link = logit), data = train.data)
# Summary of model.
summary(model2)
```

```
##
## Call:
## glm(formula = output ~ sex + age, family = binomial(link = logit),
##     data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0532  -1.0211   0.5421   1.0017   1.7653
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.87283    0.97023   5.022 5.10e-07 ***
## sex1        -1.48244    0.30210  -4.907 9.24e-07 ***
## age         -0.06732    0.01604  -4.196 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 376.52  on 272  degrees of freedom
## Residual deviance: 336.61  on 270  degrees of freedom
## AIC: 342.61
##
## Number of Fisher Scoring iterations: 4
```

Las variables explicativas son estadísticamente significativas para explicar el riesgo de enfermedad. Para el sexo masculino el riesgo de enfermedad disminuye. También el aumento de la edad en una unidad hace que disminuya el riesgo de enfermedad en 0.067.

Antes de llegar a una conclusión sobre la influencia de la edad y el sexo conjuntamente sobre el resultado de riesgo de enfermedad o no, comprobamos si la edad influye al resultado y si la edad tiene alguna influencia sobre el sexo. Hacemos esto para descartar introducir confusión con la variable edad. Los resultados los interpretamos igual que antes: la edad es una variable estadísticamente significativa para predecir el reisgo de la enfermedad, a mayor edad, menor riesgo, esto lo indica el coeficiente de la variable edad.

Encontramos un AIC de 342.6, este valor mejora el AIC en el modelo anterior, entonces, añadir la edad aumenta la capacidad predictiva del modelo.

Por otra parte, la edad no tiene influencia en el sexo, la edad no es estadisticamente significativa para predecir el sexo.

Con esto podemos descartar el riesgo de introducir confusión al añadir la variable edad.

```
# Verify age related to output
related1 <- glm(formula = output ~ age, family = binomial(link = logit), data = train.data)
# Summary of model.
summary(related1)
```

```
##
## Call:
## glm(formula = output ~ age, family = binomial(link = logit),
##     data = train.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.698  -1.177   0.838   1.072   1.584
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.95649    0.81427   3.631 0.000282 ***
## age         -0.05099    0.01466  -3.477 0.000507 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 376.52  on 272  degrees of freedom
## Residual deviance: 363.65  on 271  degrees of freedom
## AIC: 367.65
##
## Number of Fisher Scoring iterations: 4
```

```
# Verify age related to sex
related2 <- glm(formula = sex ~ age, family = binomial(link = logit), data = train.data)

# Summary of model
summary(related2)
```

```
##
## Call:
```

```
## glm(formula = sex ~ age, family = binomial(link = logit), data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7925  -1.3732   0.8037   0.9194   1.1780
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.47692    0.84344   2.937  0.00332 **
## age         -0.03218    0.01506  -2.137  0.03259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 346.13  on 272  degrees of freedom
## Residual deviance: 341.43  on 271  degrees of freedom
## AIC: 345.43
##
## Number of Fisher Scoring iterations: 4
```

Explicar la enfermedad a partir del sexo, la edad y el síntoma de angina de pecho. Añadimos la variable asyntomatic para ver si hay algún cambio en la regresión. Las 3 variables usadas en esta regresión logística son significativas. No tener síntomas reduce la probabilidad de padecer enfermedad.

```
# Build the model.
model3 <- glm(formula = output ~ sex + age + asymptomatic, family = binomial(link = logit), data = train.d
# Summary of model.
summary(model3)
```

```
##
## Call:
## glm(formula = output ~ sex + age + asymptomatic, family = binomial(link = logit),
##     data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2309  -0.6776   0.2908   0.7565   2.2152
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    6.23645    1.19370   5.224 1.75e-07 ***
## sex1          -1.70081    0.36009  -4.723 2.32e-06 ***
## age           -0.06727    0.01900  -3.540    4e-04 ***
## asymptomatic1 -2.45921    0.31708  -7.756 8.77e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 376.52  on 272  degrees of freedom
## Residual deviance: 262.93  on 269  degrees of freedom
## AIC: 270.93
##
## Number of Fisher Scoring iterations: 4
```

En el tercer modelo tuvimos un AIC de 271, que mejora la capacidad predictiva de los dos modelos anteriores.

Validación cruzada Finalmente, usamos cross validation con 5 folds para estimar los valores de accuracy del modelo. Usamos 5 folds porque el tamaño de los datos no es muy grande y dividirlos en 10 folds sería tener el conjunto de test de 30 instancias, puede no ser suficiente.

```
# Define training control
set.seed(42)
train.control <- trainControl(method = "cv", number = 5)
# Train the model
model <- train(output ~ sex + age + asymptomatic , data = train.data, method = "glmnet",
               trControl = train.control)
# Summarize the results
print(model)
```

```
## glmnet
##
## 273 samples
##   3 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 218, 219, 219, 218, 218
## Resampling results across tuning parameters:
##
##   alpha  lambda       Accuracy   Kappa
##   0.10   0.000527303  0.7694276  0.5331906
##   0.10   0.005273030  0.7694276  0.5331906
##   0.10   0.052730302  0.7729293  0.5408676
##   0.55   0.000527303  0.7694276  0.5331906
##   0.55   0.005273030  0.7694276  0.5331906
##   0.55   0.052730302  0.7764983  0.5491894
##   1.00   0.000527303  0.7694276  0.5331906
##   1.00   0.005273030  0.7730640  0.5405526
##   1.00   0.052730302  0.7692256  0.5351093
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0.55 and lambda = 0.0527303.
```

Ahora podemos acceder a los valores de accuracy del modelo y calcular el valor medio, que nos dará una idea del rendimiento del modelo.

```
accuracies <- model$results$Accuracy
mean(accuracies)
```

```
## [1] 0.7709839
```

Predicción.

Hacemos la predicción con el modelo entrenado con validación cruzada, en la predicción podemos tener las probabilidades o el resultado final. Con el resultado final que son ceros y unos y los datos del conjunto de test conseguimos una matriz de confusión que muestra que predice el no riesgo de enfermedad correctamente 9 veces, predice el riesgo de enfermedad correctamente 14 veces, 4 veces predice que no hay enfermedad y sí la hay, 4 veces predice que hay enfemedad cuando no la hay.

```
probabilities <- model %>% predict(test.data, type = "prob")
head(probabilities, 3)
```

```
##           0         1
## 1 0.2418399 0.7581601
## 2 0.1074928 0.8925072
```

```
## 3 0.7978097 0.2021903
```

```
predicted.classes <- model %>% predict(test.data, type = "raw")
predicted.classes
```

```
##  [1] 1 1 0 0 1 1 1 0 1 1 1 1 0 1 1 1 0 0 0 0 0 0 1 1 1 0 0 0 1
## Levels: 0 1
```

```
confusionMatrix(predicted.classes, test.data$output)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0  9  4
##          1  4 12
##
##                Accuracy : 0.7241
##                  95% CI : (0.5276, 0.8727)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 0.04437
##
##                   Kappa : 0.4423
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.6923
##             Specificity : 0.7500
##          Pos Pred Value : 0.6923
##          Neg Pred Value : 0.7500
##              Prevalence : 0.4483
##          Detection Rate : 0.3103
##    Detection Prevalence : 0.4483
##       Balanced Accuracy : 0.7212
##
##        'Positive' Class : 0
##
```

Representación de los resultados a partir de tablas y gráficas..

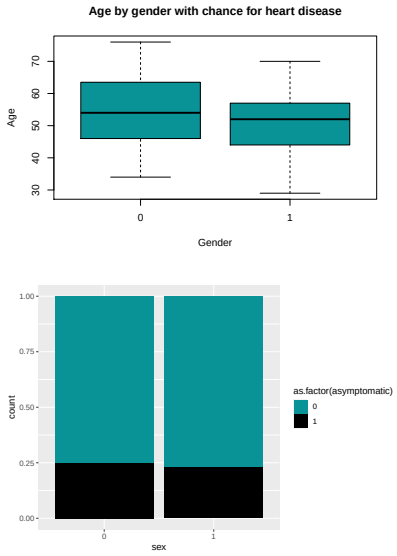
En esta práctica hemos usado el juego de datos de enfermedad cardiovascular y nos hemos enfocado en responder a unas preguntas muy concretas, que nos han requerido usar una parte reducida de los datos. En general, partimos de la idea extendida de que las mujeres no sufren enfermedad cardiovascular y buscamos pruebas para confirmarla o refutarla.

Para analizar las diferencias entre los hombres y las mujeres en el riesgo de padecer enfermedad cardiovascular, planteamos las preguntas de investigación.
* ¿Las mujeres tienen el mismo riesgo de padecer enfermedad cardiovascular que los hombres?
* ¿Quien tiene más riesgo de padecer enfermedades cardiovasculares: los hombres o las mujeres?
* ¿La edad influye en el riesgo de padecer enfermedad cardiovascular?
* ¿Las mujeres asintomáticas en cuanto al dolor de pecho tienen el mismo riesgo de padecer enfermedad cardiovascular que los hombres asintomáticos?
* ¿La edad, el sexo, la existencia o no de angina de pecho son factores que influyen en el diagnóstico de enfermedad cardiovascular?

Representación de los resultados de contrastes de hipótesis.

Pregunta	Variable	Variable	Tipo Contraste	P-value	Resultado
La edad media de los hombres con riesgo de enfermedad cardiovascular es igual que la edad media de las mujeres con riesgo.	age	output	media	0.99084	H0: no hay diferencia entre las medias.
La proporción de hombres con enfermedad cardiovascular y sin síntoma de angina es la misma que l proporción de mujeres con enfermedad cardiovascular y con síntoma de angina.	asymptomatic	output	proporción	0.7455	H0: no hay diferencia entre las proporciones

Los gráficos que ayudan a mostrar el resultado de los contrastes de forma visual son el boxplot por edad y sexo, un plot que muestra la proporción de las personas con enfermedad entre cada sexo.



Presentación del os resultados de chisq.test

Pregunta	Variable	Variable	P-value	Resultado
¿Existe relación entre la variable sex y la variable output?	sex	output	1.551e-06	H1: sí, existe asociación entre las variables
¿Existe relación entre la variable asyntomatic y la variable output?	asymptomatic	output	$<$ 2.2e-16	H1: sí, existe asociación entre las variables

Presentación de resultados de regresiones logísticas

Pregunta	Variable	Variable	Resultado
La variable sexo ayuda a explicar la enfermedad cardiovascular	sex	output	Sí, y el sexo masculino contribuye negativamente a la probabilidad de enfermedad

Pregunta	Variable	Variable	Resultado
Añadir la edad a la regresión ayuda a explicar la enfermedad	sex, age	output	Sí, la edad es relevante para explicar la enfermedad, además cuando aumenta l edad, el riesgo de enfermedad disminuye.
La existencia de angina de pecho ayuda a explicar la probabilidad de enfermedad.	sex, age, asymptomatic	output	Sí, no tener angina de pecho disminuye el riesgo de padecer enfermedad cardiovascular.

El modelo de regresión logística que explica el riesgo de enfermedad a r del sexo, edad y síntoma de angina, tiene un valor de accuracy de 0.77, como resultado de la validación cruzada.

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

Hemos realizado contrastes de hipótesis, un test chi al cuadrado para ver si las variables están relacionadas, y unas regresiones logísticas. En estos casos nos limitamos a usar las variables edad, sexo, asintomático. Las pruebas que hemos realizado ayudan a responder a las preguntas planteadas en los apartados anteriores.

En primer lugar, hemos descubierto que la edad media de las mujeres que tienen riesgo de sufrir enfermedad cardiaca es igual a la de los hombres, podemos afirmarlo con un nivel de confianza del 95%. En el boxplot del apartado anterior parecía que la edad mediana de las mujeres con enfermedad era algo más alta que la de los hombres, con el contraste hemos podido desmentir la idea.

Por otra parte, comprobamos si es verdad que la proporción de mujeres que tienen enfermedad cardiovascular la pasan sin tener el síntoma de angina es igual que para los hombres. Resulta que en proporción hay tantos hombres como mujeres que tienen la enfermedad y no tienen angina.

También aplicamos unas pruebas chisq.test para ver si las variables categóricas que representan el sexo y el síntoma de angina están relacionadas con la variable objetivo, el resultado es que sí están relacionadas, y así en la siguientea regresiones logísticas que aplicamos usamos estas variables para predecir el riesgo de enfermedad.

En las regresiones logísticas podemos ver estas mismas variables con un enfoque un poco diferente. La primera regresión nos muestra que el sexo masculino reduce la probabilidad de enfermedad, esto parece que contradice a la opinión general de que las mujeres sufren menos enfermedad cardiovascular. Otra conclusión desde la regresión logística es que a mayor edad parece haber menos riesgo de enfermedad. La última conclusión de la regresión logística es que ser asintomático reduce el riesgo de enfermedad, y esto sí podríamos esperarlo. Al experimentar con regresiones logísticas añadiendo variables, cada vez hemos visto que el valor de AIC mejoraba, esto se debe a que las variables que añadimos ayudaron a explicar algo mejor la variable output.

En conclusión: las mujeres que sufren enfermedad cardiovascular tienen de media una edad similar a la de los hombres que tienen enfermedad, cuando sufren enfermedad cardiovscular, hay tanta proporción de mujeres que tienen angina como de hombres. Y según las regresiones logísticas y el odds ratio, los hombres tienen algo menos de riesgo de enfermar.

Código.

https://github.com/mnyborg77/tcvd_PRA2.

Vídeo.

https://drive.google.com/file/d/1UjK8VkLZBKqF-K6BR1T2JRi8rYQkdfH9/view?usp=sharing