PEC 3

Andrés Sánchez Ruiz y Marc Bañuls Tornero

26/12/2019

Contents

al lector a comprender el ámbito de estudio. Además, se valorará positivamente la coherencia de resultados y las justificaciones presentadas. **42**

1) (1 punto) Buscar un conjunto de datos relacionados con la Bioestadística o la Bioinformática. Posibles recursos bibliográficos son aquellos presentados en la PEC1, por ejemplo: http://www.bioinformatics.org/sms2/index.html; o bien http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets. También es posible utilizar fuentes propias o de interés, siempre teniendo en cuenta que sean datos públicos para los usuarios. Será necesario especificar la procedencia y la justificación de la elección de los datos.

Clicando sobre el segundo de los enlaces que se nos ofrece en el enunciado accedemos al repositorio del cual descargamos el archivo relacionado con la diabetes en formato csv. Hemos escogido estos datos porque la diabetes es una grave enfermedad a tratar en la actualidad y estos datos, aunque obtenidos hace tiempo, pueden ser útiles para tener un mejor entendimiento sobre esta enfermedad. Además, hay una suficiente cantidad de datos con las que realizar distintos estudios y comparaciones de variables para esta PEC.

2) (1 punto) Utilizando R, mostrar y explicar qué tipo de ficheros se han importado, las variables de estudio (clasificación, . . .), así como todo aquello relevante para el estudio. Incluir capturas de pantall a y las instrucciones en R utilizadas para importar y mostrar los datos.

Una vez descargado el archivo hemos accedido a RStudio, determinado la carpeta en la que trabajaremos (PEC3) y cargado nuestro data frame en la variable diabetes. Todo el conjunto de datos puede observarse mediante la funcion View().

```
diabetes <- read.csv("diabetes.csv", header=TRUE)
View(diabetes)
```

Las variables del estudio se encuentran subdivididas en columnas en el el dataframe `diabetes`. Podemos observar los nombres de todas las variables del estudio y el número de éstas:

```
names(diabetes)
```

```
##  [1] "id"       "chol"     "stab.glu" "hdl"      "ratio"    "glyhb"
##  [7] "location" "age"      "gender"   "height"   "weight"   "frame"
## [13] "bp.1s"    "bp.1d"    "bp.2s"    "bp.2d"    "waist"    "hip"
## [19] "time.ppn"
```

```
length(names(diabetes))
```

```
## [1] 19
```

Observamos que tenemos 19 variables dentro de este dataframe. La variable “id” sirve como identificador de los pacientes que han participado al estudio. La variable “chol” determina los niveles de colesterol total. La variable “stab.glu” indica el nivel de estabilidad de la glucosa en sangre. Cuanto mayor el valor de estabilidad más fiables son los datos de ese registro. La variable “hdl” indica la concentración de “High

Density Lipoprotein" o llamado vulgarmente el colesterol bueno, debido a que este tipo de colesterol es transportado al hígado para ser posteriormente eliminado o metabolizado (no se acumula en las arterias). La variable "ratio" indica el ratio entre colesterol y hdl, para así observar el ratio de colesterol malo respecto al bueno en el paciente. La variable "glyhb" determina la concentración de hemoglobina glicosilada. Esta concentración se utiliza principalmente para diagnosticar diabetes en los pacientes. En este estudio, si el paciente tiene una concentración de hemoglobina glicosilada mayor a 7, se considera que dicho paciente tiene diabetes. La variable "time.ppn" mide el "postprandial time", que consiste en el tiempo requerido por el paciente en llegar al pico de glucosa en sangre justo después de una toma de alimento. A continuación tenemos diferentes variables para tener en cuenta la fisiología del paciente junto con su procedencia: localización ("location"), edad ("age" medida en años), sexo ("gender"), altura ("height" medida en pulgadas), peso ("weight" medido en libras), constitución ("frame"), cintura y cadera("waist" y "hip" respectivamente, ambas medidas en pulgadas), primera y segunda presión arterial sistólica y primera y segunda presión arterial diastólica ("bp.1s","bp.2s","bp.1d" y "bp.2d" respectivamente).

Utilizando la función `str()` podemos observar los tipos de variables con los que vamos a realizar el estudio:

```
str(diabetes)
```

```
## 'data.frame':    403 obs. of  19 variables:
##  $ id      : int  1000 1001 1002 1003 1005 1008 1011 1015 1016 1022 ...
##  $ chol    : int  203 165 228 78 249 248 195 227 177 263 ...
##  $ stab.glu: int  82 97 92 93 90 94 92 75 87 89 ...
##  $ hdl     : int  56 24 37 12 28 69 41 44 49 40 ...
##  $ ratio   : num  3.6 6.9 6.2 6.5 8.9 ...
##  $ glyhb   : num  4.31 4.44 4.64 4.63 7.72 ...
##  $ location: Factor w/ 2 levels "Buckingham","Louisa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  46 29 58 67 64 34 30 37 45 55 ...
##  $ gender  : Factor w/ 2 levels "female","male": 1 1 1 2 2 2 2 2 2 1 ...
##  $ height  : int  62 64 61 67 68 71 69 59 69 63 ...
##  $ weight  : int  121 218 256 119 183 190 191 170 166 202 ...
##  $ frame   : Factor w/ 4 levels "","large","medium",..: 3 2 2 2 3 2 3 3 2 4 ...
##  $ bp.1s   : int  118 112 190 110 138 132 161 NA 160 108 ...
##  $ bp.1d   : int  59 68 92 50 80 86 112 NA 80 72 ...
##  $ bp.2s   : int  NA NA 185 NA NA NA 161 NA 128 NA ...
##  $ bp.2d   : int  NA NA 92 NA NA NA 112 NA 86 NA ...
##  $ waist   : int  29 46 49 33 44 36 46 34 34 45 ...
##  $ hip     : int  38 48 57 38 41 42 49 39 40 50 ...
##  $ time.ppn: int  720 360 180 480 300 195 720 1020 300 240 ...
```

Principalmente vamos a trabajar con variables numéricas, excepto con las variables de localización, sexo, y constitución, los cuales son factores.

3) (2 puntos) Con la Sección 2 de la PEC1 como base, elaborar y analizar una serie de cuestiones, que ayuden a explorar y a familiarizarse mejor con los datos de estudio. Además, en algunos casos, puede utilizarse la definición de funciones y el lenguaje SQL estudiado en el LAB3.

Al haber definido y comprendido las variables dentro de este dataset, podemos obtener información variada para sentar unas posibles bases con las que trabajar o realizar hipótesis posteriormente. Esta información puede ser obtenida mediante subsets y otros tratamientos y cálculos del dataset `variable`.

Primero observamos la cantidad de pacientes con los que vamos a trabajar (utilizando para ello la variable "id"):

```
length(diabetes$id)
```

```
## [1] 403
```

Esto nos indica que el dataset está comprendido por 403 pacientes. Otro dato importante que debemos saber previo estudio es saber si hay datos perdidos o nulos en el dataset. Para ello podemos realizar una tabla general de este dataset.

```
table(is.na(diabetes))
```

```
##
## FALSE  TRUE
##  7094   563
```

Observamos que hay 569 "missing values", una significativa cantidad que puede afectar a posteriores estudios de los datos. Por ello, es recomendable observar específicamente donde se encuentran estos "missing values" y ver en que o donde pueden afectar. Utilizamos la función `colSums()` con este fin.

```
colSums(is.na(diabetes))
```

```
##       id     chol stab.glu      hdl    ratio    glyhb location      age
##        0        1        0        1        1       13        0        0
##   gender   height   weight    frame    bp.1s    bp.1d    bp.2s    bp.2d
##        0        5        1        0        5        5      262      262
##    waist      hip time.ppn
##        2        2        3
```

Esto nos indica que hay 13 "missing values" en los registros de hemoglobina glicosilada, una variable de cierta importancia para posteriores estudios (con esta variable determinamos si el paciente tiene diabetes o no), y algún que otro missing value en otras variables del dataset. La gran mayoría de los otros "missing values" (524) corresponden a la segunda presión arterial sistólica y diastólica. Estas dos variables tienen cierta importancia médica pero en los estudios a realizar en esta PEC no va a ser determinante, por lo que la ausencia de estos valores no resulta significativa.

Para realizar un análisis superficial de los datos, podemos realizar varios filtros de los datos:

Primero podemos obtener el número total de pacientes con diabetes y pacientes sanos:

```
p_diabetes<- length(subset(diabetes$id, diabetes$glyhb >= 7))
p_sano<- length(subset(diabetes$id, diabetes$glyhb < 7))
p_diabetes
```

```
## [1] 60
```

```
p_sano
```

```
## [1] 330
```

Observamos que un bajo porcentaje de los pacientes del estudio tiene diabetes (60 de los 403 totales).

Posteriormente analizamos los datos de la persona con los niveles de colesterol más bajo y más alto.

```
max_chol<-subset(diabetes, diabetes$chol == (max(na.omit(diabetes$chol))))
min_chol<-subset(diabetes, diabetes$chol == (min(na.omit(diabetes$chol))))
max_chol
```

```
##      id chol stab.glu hdl ratio glyhb   location age gender height weight
## 63 2778  443      185  23  19.3 14.31 Buckingham  51 female     70    235
##     frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
## 63 medium   158    98   148    88    43  48      420
```

```
min_chol
```

```
##     id chol stab.glu hdl ratio glyhb   location age gender height weight
## 4 1003   78       93  12   6.5  4.63 Buckingham  67   male     67    119
##   frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
## 4 large   110    50    NA    NA    33  38      480
```

Con estos dos registros vemos claras diferencias en el ratio de colesterol/hdl y sobre todo la concentración de hemoglobina glicosilada.

El siguiente conjunto de datos se ha obtenido de una muestra de 403 afroamericanos residentes en el estado de Virginia, Estados Unidos. Originalmente la intención de este estudio fue determinar la prevalencia de la obesidad, la diabetes y otros riesgos cardiovasculares dentro de este grupo. No obstante, en el siguiente análisis principalmente intentaremos determinar si existía una mayor propensión a padecer diabetes según si los habitantes fuesen de Louisa o Buckingham, aunque también se tendrán en cuenta aquellas relacionadas con la salud como el HDL, colesterol, IMC... etc. Nuestra hipótesis es que aunque no se trataba de poblaciones especialmente grandes, la de la Buckingham era 10 veces superior (15.000 frente a 1.400) a la de Louisa, lo cual probablemente influenciaba en cierta forma en su estilo de vida (alimentación, ejercicio diario... etc). De entre todas la variables tomaremos la de Glycosilated hemoglobin (glyhb) como indicador de la presencia de diabetes (segun se puede leer en el enlace http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html): aquellos valores que sean mayores que 7 serán tomados como pacientes que sufren esta enfermedad, mientras que aquellos que tengan un valor inferior seran tomados como grupo control.

Otra cosa que podemos añadir a esta base de datos son los valores de índice de masa corporal (bmi en ingés) de cada paciente, que dará información más detallada sobre sus posibles alteraciones en su relación entre peso y altura. Para ello utilizamos la fórmula:

$$BMI = \frac{mass_{kg}}{height_m^2} = \frac{mass_{lb}}{height_{lb}^2} * 703$$

Por lo tanto debemos realizar esta operación para cada registro en la base de datos:

```
diabetes$imc <- (diabetes$weight  /diabetes$height**2) * 703
```

```
Lcontrol <- sqldf("SELECT * FROM diabetes WHERE glyhb<7 AND location='Louisa'")
Ldiabeticos <- sqldf("SELECT * FROM diabetes WHERE glyhb>7 AND location='Louisa'")

Bcontrol <- sqldf("SELECT * FROM diabetes WHERE glyhb<7 AND location='Buckingham'")
Bdiabeticos <- sqldf("SELECT * FROM diabetes WHERE glyhb>7 AND location='Buckingham'")
```

4) (1 punto) Realizar un análisis descriptivo de los datos. El análisis debe incluir (tal y como aparece en la Sección 3 de la PEC1) un resumen paramétrico de los datos y su representación gráfica, que mejor defina y complemente cada una de dichas variables.

Primero realizamos un resumen paramétrico de todas las variables:

```
# Resumen paramétrico de variables
summary(diabetes)
```

```
##        id             chol          stab.glu          hdl        
##  Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00  
##  1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00  
##  Median :15766   Median :204.0   Median : 89.0   Median : 46.00  
##  Mean   :15978   Mean   :207.8   Mean   :106.7   Mean   : 50.45  
```

```
##  3rd Qu.:20336   3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00
##  Max.   :41756   Max.   :443.0   Max.   :385.0   Max.   :120.00
##                  NA's   :1                       NA's   :1
##      ratio            glyhb             location       age
##  Min.   : 1.500   Min.   : 2.68   Buckingham:200   Min.   :19.00
##  1st Qu.: 3.200   1st Qu.: 4.38   Louisa    :203   1st Qu.:34.00
##  Median : 4.200   Median : 4.84                    Median :45.00
##  Mean   : 4.522   Mean   : 5.59                    Mean   :46.85
##  3rd Qu.: 5.400   3rd Qu.: 5.60                    3rd Qu.:60.00
##  Max.   :19.300   Max.   :16.11                    Max.   :92.00
##  NA's   :1        NA's   :13
##     gender        height          weight         frame         bp.1s
##  female:234   Min.   :52.00   Min.   : 99.0         : 12   Min.   : 90.0
##  male  :169   1st Qu.:63.00   1st Qu.:151.0   large :103   1st Qu.:121.2
##               Median :66.00   Median :172.5   medium:184   Median :136.0
##               Mean   :66.02   Mean   :177.6   small :104   Mean   :136.9
##               3rd Qu.:69.00   3rd Qu.:200.0                3rd Qu.:146.8
##               Max.   :76.00   Max.   :325.0                Max.   :250.0
##               NA's   :5       NA's   :1                    NA's   :5
##      bp.1d            bp.2s           bp.2d            waist
##  Min.   : 48.00   Min.   :110.0   Min.   : 60.00   Min.   :26.0
##  1st Qu.: 75.00   1st Qu.:138.0   1st Qu.: 84.00   1st Qu.:33.0
##  Median : 82.00   Median :149.0   Median : 92.00   Median :37.0
##  Mean   : 83.32   Mean   :152.4   Mean   : 92.52   Mean   :37.9
##  3rd Qu.: 90.00   3rd Qu.:161.0   3rd Qu.:100.00   3rd Qu.:41.0
##  Max.   :124.00   Max.   :238.0   Max.   :124.00   Max.   :56.0
##  NA's   :5        NA's   :262     NA's   :262      NA's   :2
##       hip           time.ppn           imc
##  Min.   :30.00   Min.   :   5.0   Min.   :15.20
##  1st Qu.:39.00   1st Qu.:  90.0   1st Qu.:24.13
##  Median :42.00   Median : 240.0   Median :27.80
##  Mean   :43.04   Mean   : 341.2   Mean   :28.78
##  3rd Qu.:46.00   3rd Qu.: 517.5   3rd Qu.:32.24
##  Max.   :64.00   Max.   :1560.0   Max.   :55.78
##  NA's   :2       NA's   :3        NA's   :6
```

En los datos del resumen paramétrico cabe destacar que en el tercer cuartil (es decir, 75% de los pacientes) aún no tenemos pacientes con diabetes. Observando el IMC de la población del estudio, sabemos que para el estudio se han escogido personas principalmente con cierto sobrepeso, ya que en el primer cuartil ya tenemos un ibm de 24 y la media de la población está cercana a 29, que se traduce a un sobrepeso de grado II (llamado también preobesidad). Un último dato a mencionar es que los pacientes elegidos para el estudio tienen todos una edad adulta (el paciente más joven tiene 19 años) y la mayoría de éstos tienen entre 30 y 50 años, cosa que puede influir en posteriores investigaciones en la PEC.

Para representar de manera útil los datos, podemos realizar un resumen paramétrico basado en las concentraciones de hemoglobina glicosilada entre los pacientes procedentes de Louisa con los de Buckingham. Además podemos dividir estos pacientes en un grupo control (donde su concentración de hemoglobina glicosilada es menor a 7) y grupo diabético (cuando la concentración de hemoglobina glicosilada es mayor a 7). A continuacion se mostrará un resumen de las variables que se tendrán en cuenta más adelante:

```
#Resumen parametrico
Louisa <- c(mean(Lcontrol$glyhb), mean(Ldiabeticos$glyhb))
Buckingham <- c(mean(Bcontrol$glyhb), mean(Bdiabeticos$glyhb))
Louisasd <- c(sd(Lcontrol$glyhb), sd(Ldiabeticos$glyhb))
Buckinghamsd <- c(sd(Bcontrol$glyhb), sd(Bdiabeticos$glyhb))
```

```r
grupos <- c("Control", "Diabéticos")
resumenparametrico <- data.frame("Media Glyhb" = grupos, Louisa, Buckingham)
resumenparametrico
```

```
##   Media.Glyhb    Louisa Buckingham
## 1     Control  4.686725   4.843585
## 2  Diabéticos 10.345172   9.949677
```

```r
Louisa <- c(mean(Lcontrol$age), mean(Ldiabeticos$age))
Buckingham <- c(mean(Bcontrol$age), mean(Bdiabeticos$age))
resumenparametrico1 <- data.frame("Media edad" = grupos, Louisa, Buckingham)
resumenparametrico1
```

```
##   Media.edad   Louisa Buckingham
## 1    Control 44.84211   44.46541
## 2 Diabéticos 55.82759   60.80645
```

```r
Louisa <- c(mean(Lcontrol$hdl), mean(Ldiabeticos$hdl))
Buckingham <- c(mean(na.omit(Bcontrol$hdl)), mean(Bdiabeticos$hdl))
resumenparametrico2 <- data.frame("Media HDL" = grupos, Louisa, Buckingham)
resumenparametrico2
```

```
##    Media.HDL  Louisa Buckingham
## 1    Control 51.7076   50.60127
## 2 Diabéticos 44.2069   46.29032
```

```r
Louisa <- c(mean(na.omit(Lcontrol$waist)), mean(Ldiabeticos$waist))
Buckingham <- c(mean(Bcontrol$waist), mean(Bdiabeticos$waist))
resumenparametrico3 <- data.frame("Media cintura" = grupos, Louisa, Buckingham)
resumenparametrico3
```

```
##   Media.cintura   Louisa Buckingham
## 1       Control 37.36095   37.33962
## 2    Diabéticos 39.03448   42.61290
```

```r
Louisa <- c(mean(na.omit(Lcontrol$hip)), mean(Ldiabeticos$hip))
Buckingham <- c(mean(Bcontrol$hip), mean(Bdiabeticos$hip))
resumenparametrico4 <- data.frame("Media cadera" = grupos, Louisa, Buckingham)
resumenparametrico4
```

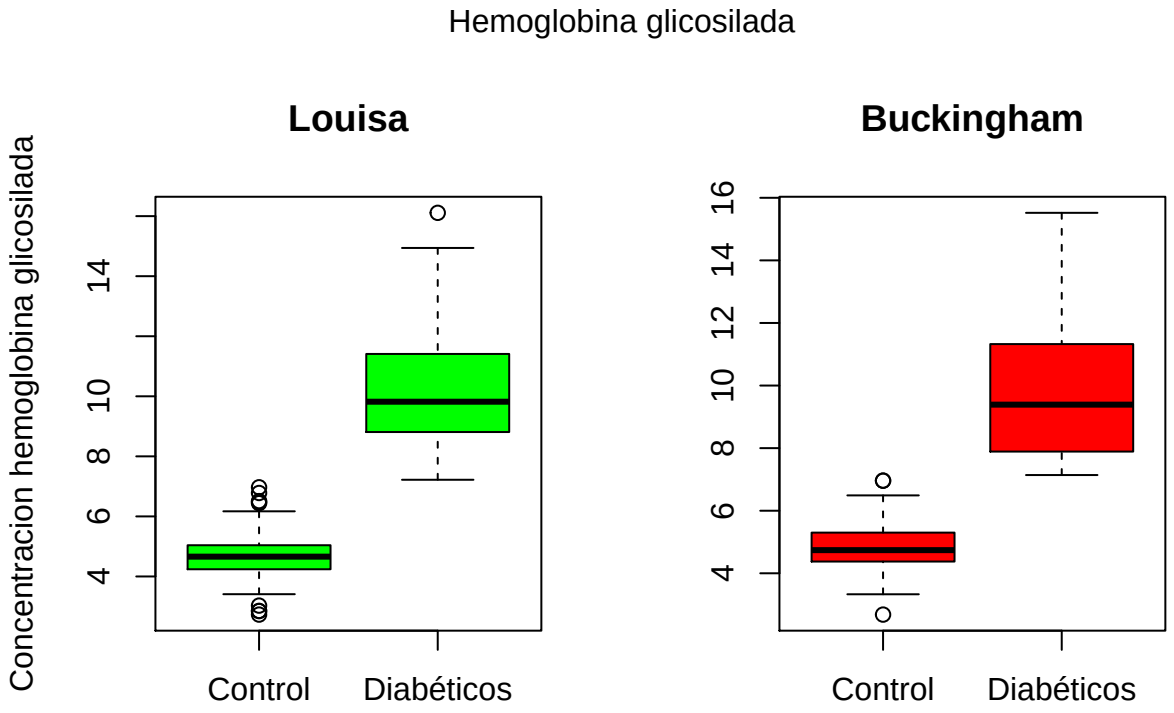
```
##   Media.cadera  Louisa Buckingham
## 1      Control 43.2071   42.14465
## 2   Diabéticos 44.0000   45.74194
```

```r
Louisa <- c(mean(Lcontrol$chol), mean(Ldiabeticos$chol))
Buckingham <- c(mean(na.omit(Bcontrol$chol)), mean(Bdiabeticos$chol))
resumenparametrico5 <- data.frame("Media colesterol" = grupos, Louisa, Buckingham)
resumenparametrico5
```

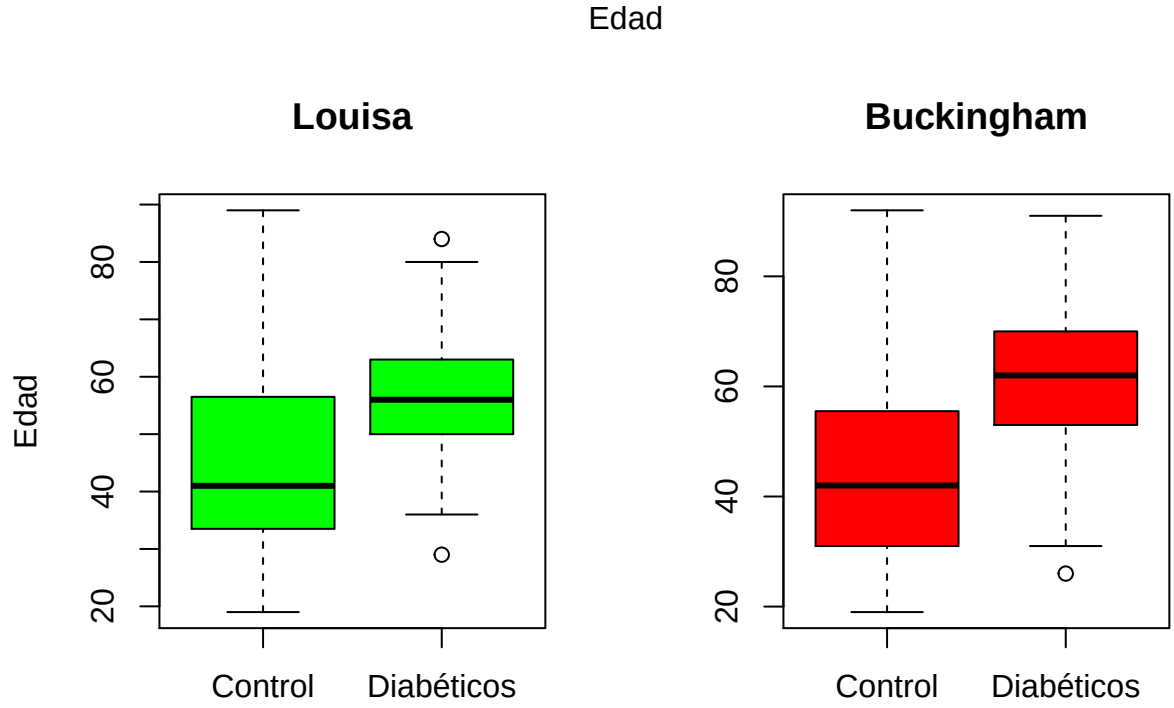
```
##   Media.colesterol   Louisa Buckingham
## 1          Control 207.9181   198.4810
## 2       Diabéticos 227.9310   229.2258
```

```r
#Representación gráfica de los datos del estudio
par(mfrow=c(1,2), oma = c(0,0,2,0))
boxplot(Lcontrol$glyhb, Ldiabeticos$glyhb, names=grupos, main = "Louisa",
        ylab = "Concentracion hemoglobina glicosilada",col = "green")
boxplot(Bcontrol$glyhb, Bdiabeticos$glyhb, names=grupos, main = "Buckingham", col = "red")
```

```
mtext("Hemoglobina glicosilada", outer = TRUE, cex=1, line=0)
```

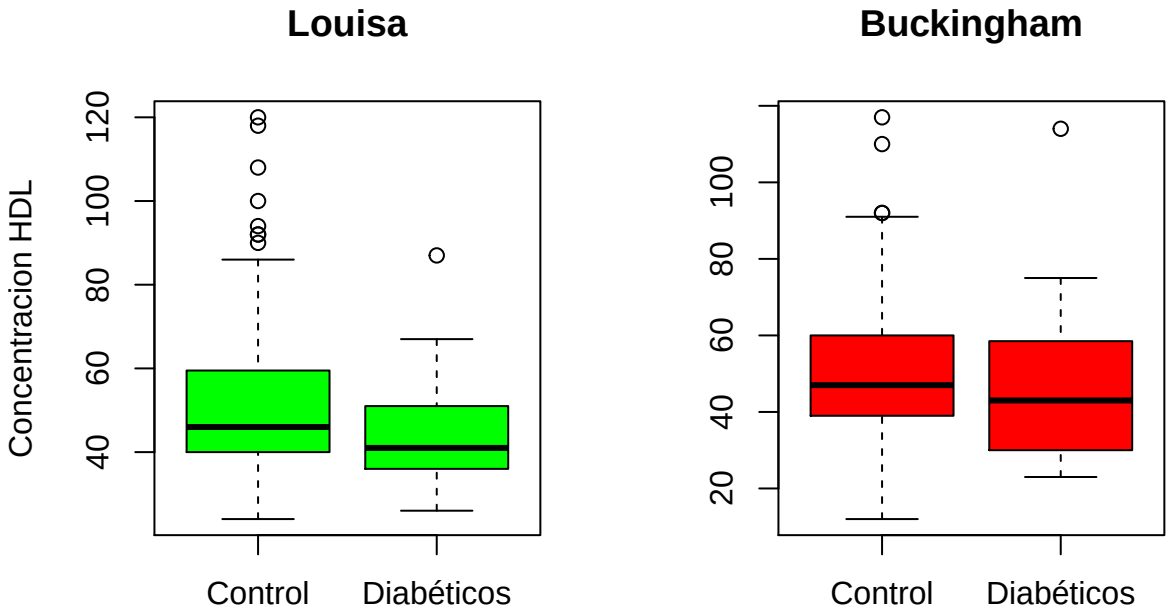



```
par(mfrow=c(1,2), oma = c(0,0,2,0))
boxplot(Lcontrol$age, Ldiabeticos$age, names=grupos, main = "Louisa",
        ylab = "Edad",col = "green")
boxplot(Bcontrol$age, Bdiabeticos$age, names=grupos, main = "Buckingham", col = "red")
mtext("Edad", outer = TRUE, cex=1, line=0)
```



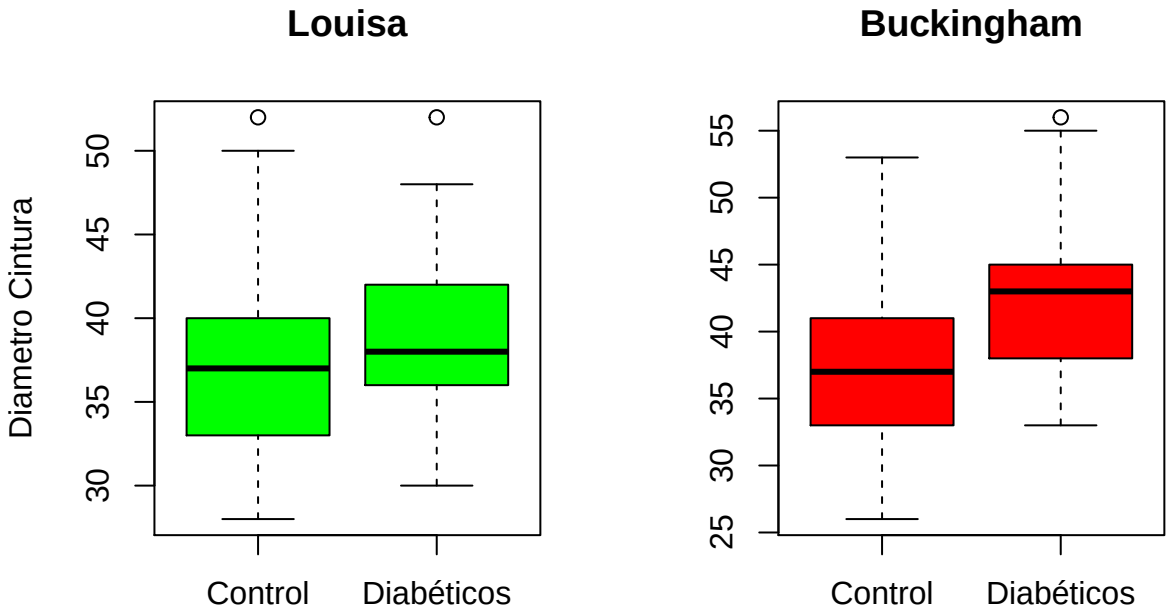

```
par(mfrow=c(1,2), oma = c(0,0,2,0))
boxplot(Lcontrol$hdl, Ldiabeticos$hdl,  names=grupos, main = "Louisa",
        ylab = "Concentracion HDL",col = "green")
boxplot(Bcontrol$hdl, Bdiabeticos$hdl, names=grupos, main = "Buckingham", col = "red")
mtext("HDL", outer = TRUE, cex=1, line=0)
```

HDL



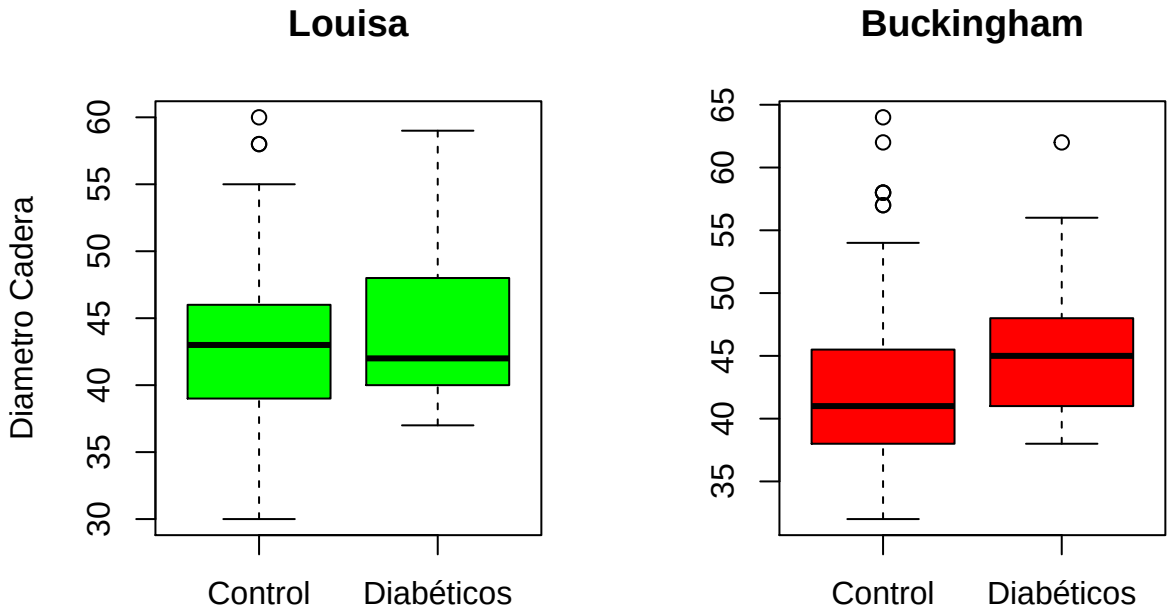

```
par(mfrow=c(1,2), oma = c(0,0,2,0))
boxplot(Lcontrol$waist, Ldiabeticos$waist, names=grupos, main = "Louisa",
        ylab = "Diametro Cintura",col = "green")
boxplot(Bcontrol$waist, Bdiabeticos$waist, names=grupos, main = "Buckingham", col = "red")
mtext("Cintura", outer = TRUE, cex=1, line=0)
```

Cintura

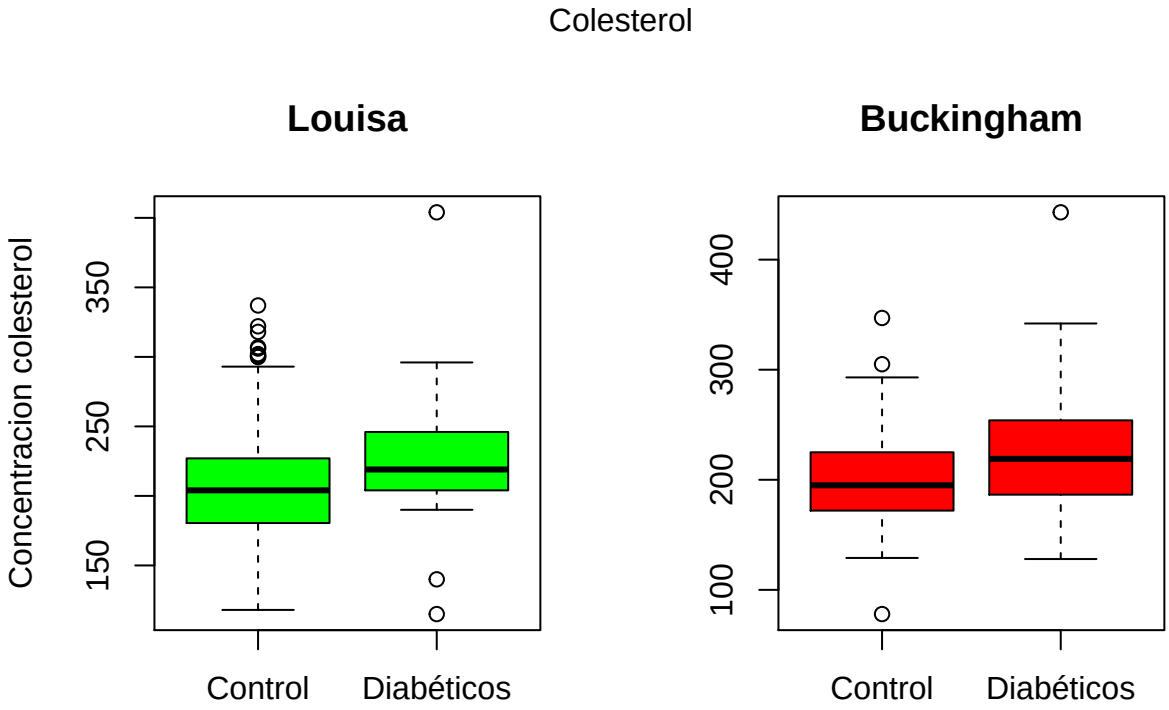



```
par(mfrow=c(1,2), oma = c(0,0,2,0))
boxplot(Lcontrol$hip, Ldiabeticos$hip, names=grupos, main = "Louisa",
        ylab = "Diametro Cadera",col = "green")
boxplot(Bcontrol$hip, Bdiabeticos$hip, names=grupos, main = "Buckingham", col = "red")
mtext("Cadera", outer = TRUE, cex=1, line=0)
```

Cadera




```
par(mfrow=c(1,2), oma = c(0,0,2,0))
boxplot(Lcontrol$chol, Ldiabeticos$chol, names=grupos, main = "Louisa",
        ylab = "Concentracion colesterol",col = "green")
boxplot(Bcontrol$chol, Bdiabeticos$chol, names=grupos, main = "Buckingham", col = "red")
mtext("Colesterol", outer = TRUE, cex=1, line=0)
```



Como puede observarse las medias obtenidas en ambas poblaciones para cada uno de los grupos son similares, a excepcion de la variable cadera que parece estar ligaremente distribuida de distinta forma en el grupo de diabéticos de Louisa. La significancia de estas diferencias se comprobará a lo largo del siguiente estudio.

5) (1,5 punto) Complementando el apartado anterior, elaborar un análisis de regresión de dos conjuntos de variables (LAB2 y Ejercicio 6 de la PEC1). La elección de las variables, los resultados, así como su relación deben de estar correctamente justificada.

Para observar posibles correlaciones entre las variables del estudio, utilizamos la función `cor()` omitiendo los "missing values" de la base de datos. Para eliminar de la correlación valores factoriales que impiden el procesamiento de la tabla entera, realizamos un query sql para obtener una nueva base de datos sin estos valores:

```
# Eliminamos además las variables 'stab.glu' y 'time.ppn', ya que estas variables
# son informativas pero no afectan a los otros valores
diabetes_2 <- sqldf("SELECT chol,hdl,ratio,glyhb,age,height,weight,waist,hip,imc FROM diabetes")
```

```
cor(na.omit(diabetes_2))
```

```
##              chol         hdl       ratio       glyhb          age
## chol    1.00000000  0.18797855  0.47688710  0.27103964  0.2547360345
## hdl     0.18797855  1.00000000 -0.68421379 -0.14665730  0.0293171722
## ratio   0.47688710 -0.68421379  1.00000000  0.34448729  0.1654136248
## glyhb   0.27103964 -0.14665730  0.34448729  1.00000000  0.3361330235
## age     0.25473603  0.02931717  0.16541362  0.33613302  1.0000000000
```

```
## height -0.06968623 -0.09180354  0.08332396  0.05412979 -0.0910550527
## weight  0.05641402 -0.29839036  0.27974250  0.15863127 -0.0646278594
## waist   0.11749415 -0.28569377  0.30975212  0.23816614  0.1520176885
## hip     0.07106823 -0.23335910  0.20351732  0.14089484  0.0009374293
## imc     0.08755875 -0.24533283  0.22692674  0.12474171 -0.0124085177
##             height      weight       waist           hip         imc
## chol   -0.06968623  0.05641402  0.11749415  0.0710682253  0.08755875
## hdl    -0.09180354 -0.29839036 -0.28569377 -0.2333591044 -0.24533283
## ratio   0.08332396  0.27974250  0.30975212  0.2035173204  0.22692674
## glyhb   0.05412979  0.15863127  0.23816614  0.1408948352  0.12474171
## age    -0.09105505 -0.06462786  0.15201769  0.0009374293 -0.01240852
## height  1.00000000  0.25089840  0.05569341 -0.1063940703 -0.26703016
## weight  0.25089840  1.00000000  0.85128530  0.8286568690  0.85856513
## waist   0.05569341  0.85128530  1.00000000  0.8331178139  0.81215895
## hip    -0.10639407  0.82865687  0.83311781  1.0000000000  0.88571221
## imc    -0.26703016  0.85856513  0.81215895  0.8857122065  1.00000000
```

En esta tabla de correlaciones podemos observar algunas variables que tienen una correlación significativa entre sí:
Existe una alta relación entre la variable "ratio", la concentración de colesterol y la concentración de hdl. Esto se debe a que la variable "ratio", como ya hemos dicho anteriormente, es la relación entre la concentración de colesterol y la concentración de hdl.

Observamos además una leve correlación entre estas concentraciones de hemoglobina glicosilada y la concentración de colesterol, junto con una leve correlación negativa con la concentración de hdl. Esto podría indicar que existe una pequeña correlación entre la aparición de diabetes y altos niveles de colesterol y/o bajos niveles de hdl en sangre. Asimismo, se observa una correlación positiva significativa entre la edad y la concentración de hemoglobina glicosilada, indicando que conforme se llegan a edades más avanzadas aumenta la concentración de este tipo de hemoglobina. Por úlimo, también encontramos relacion entre el peso y altura con el diámetro de la cintura y cadera de los pacientes.

Sabiendo estas correlaciones podemos realizar varios modelos de regresión y su representación en gráficos para observar si la correlación es significativa.

5.1. Regresión ratio con hdl y colesterol

5.1.1- Ratio/Chol

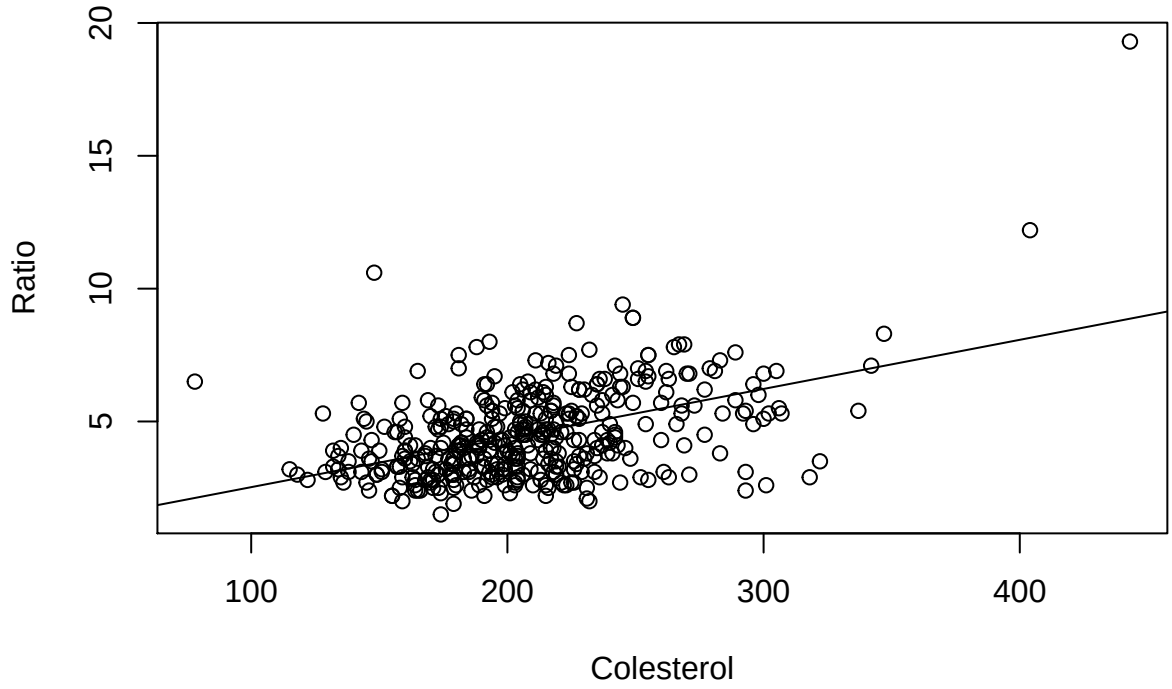
Realizamos el modelo de regresión lineal entre el ratio y la concentración de colesterol y otro modelo entre el ratio y la concentración de hdl. En ambos modelos tomamos por variable explicada la del ratio.

```
modelo_r_chol<- lm(diabetes$ratio ~ diabetes$chol)
summary(modelo_r_chol)
```

```
##
## Call:
## lm(formula = diabetes$ratio ~ diabetes$chol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6959 -0.9966 -0.0959  0.9339 10.4312
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.67928    0.36343   1.869   0.0623 .
## diabetes$chol  0.01849    0.00171  10.811   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.522 on 400 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2261, Adjusted R-squared:  0.2242
## F-statistic: 116.9 on 1 and 400 DF,  p-value: < 2.2e-16
```

```
plot(diabetes$chol,diabetes$ratio, xlab="Colesterol", ylab="Ratio")
abline(modelo_r_chol)
```



Según el valor de R cuadrado ajustado sabemos que el modelo no sigue una tendencia definida. Aun así, el p-valor es significativamente menor a 0.05 y por lo tanto podemos rechazar la hipótesis nula, aceptando que existe una correlación entre estas dos variables. Por otro lado, p-valor del intercepto no es significativo (mayor a 0.05), probablemente debido a que solo tiene en cuenta los valores que toman los predictores cuando la variable explicada equivale a 0. En casos como este donde el valor 0 no se alcanza el interceptor no tiene significancia. En la gráfica de las variables representadas podemos observar por qué el valor de R ajustado es bajo (la mayoría de valores se encuentran en una zona del gráfico excepto unos pocos). Aun así, se observa que a mayores concentraciones de colesterol (aunque disminuye el número de pacientes) aumenta paulatinamente el ratio.

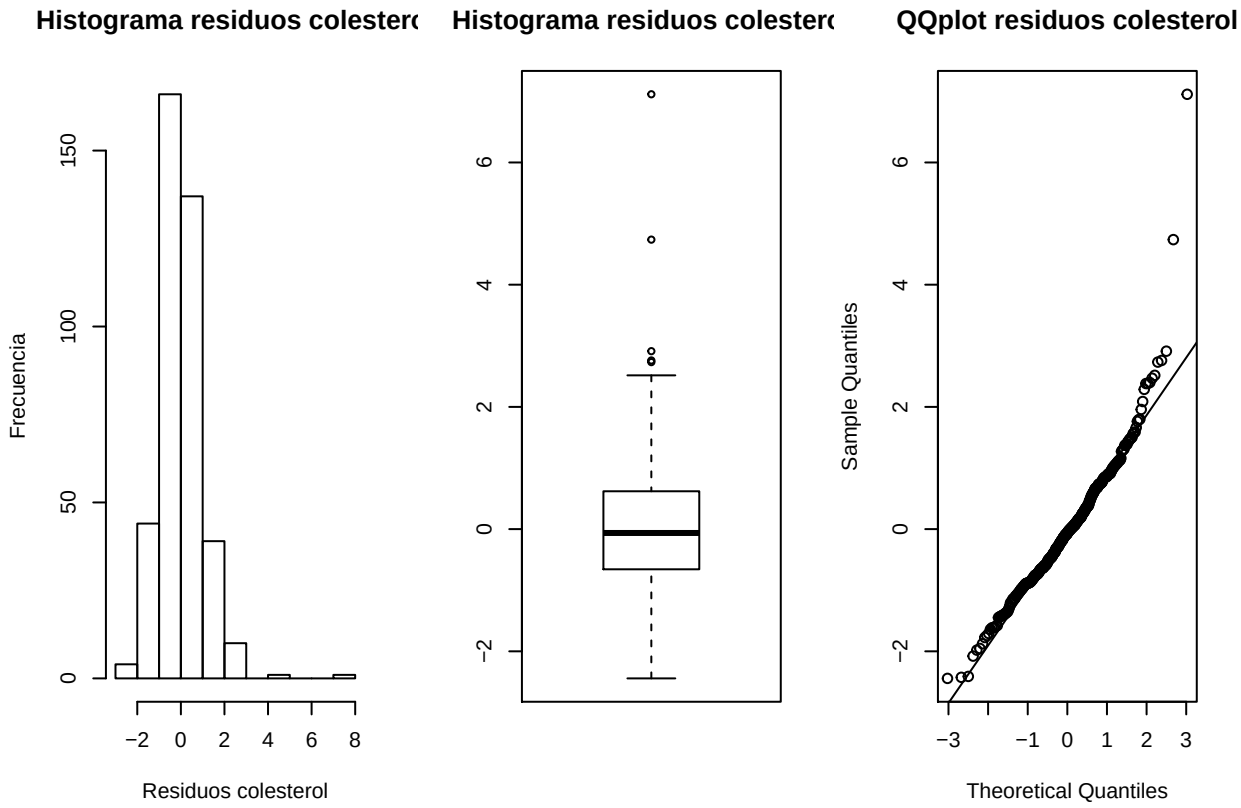
Ahora calculamos los residuos del modelo y representamos los valores en un gráfico de normalidad:

```
residuos_r_chol<-rstandard(modelo_r_chol)
par(mfrow=c(1,3))

hist(residuos_r_chol, xlab="Residuos colesterol", ylab = "Frecuencia",
     main = "Histograma residuos colesterol")
```

```r
boxplot(residuos_r_chol, main="Histograma residuos colesterol")

qqnorm(residuos_r_chol, main="QQplot residuos colesterol")
qqline(residuos_r_chol)
```



En estas gráficas observamos que el modelo parece seguir una distribución normal y que sus residuos se encuentran dispersos (observable en los outliers y la amplitud de los bigotes del boxplot). Además, en el diagrama de puntos se puede observar que los residuos siguen la línea del modelo de manera bastante exacta (exceptuando los primeros y últimos valores, algo usual).

5.1.2- Ratio/HDL

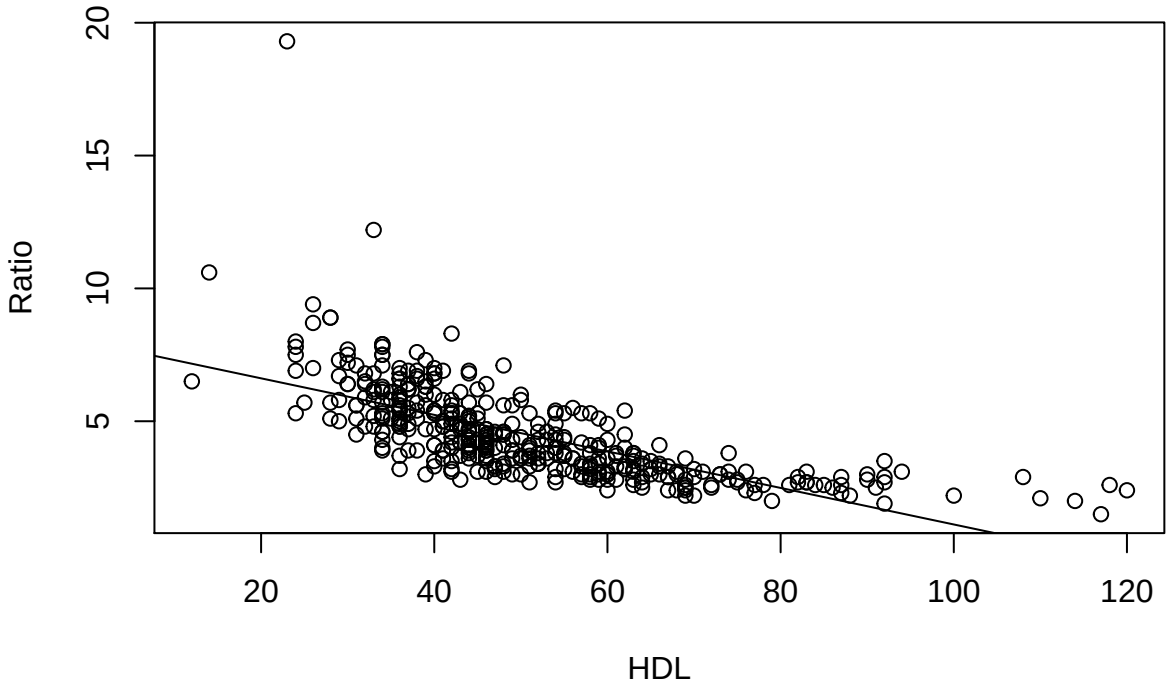
Respecto al modelo de regresión entre el ratio y la concentración de hdl, podemos realizar los mismos pasos:

```r
modelo_r_hdl<- lm(diabetes$ratio ~ diabetes$hdl)
summary(modelo_r_hdl)
```

```
##
## Call:
## lm(formula = diabetes$ratio ~ diabetes$hdl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3148 -0.7586 -0.1897  0.5571 12.8913
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.990022   0.193897   41.21   <2e-16 ***
```

```
## diabetes$hdl -0.068755   0.003637  -18.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.257 on 400 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4718, Adjusted R-squared:  0.4705
## F-statistic: 357.3 on 1 and 400 DF,  p-value: < 2.2e-16
```

```
plot(diabetes$hdl,diabetes$ratio, xlab="HDL", ylab="Ratio")
abline(modelo_r_hdl)
```



Este modelo tiene un mejor ajuste a los datos que el anterior, aunque sigue siendo un ajuste pobre. Aun así obtenemos que hay una correlación negativa entre estas dos variables. En la gráfica observamos con cierta facilidad que conforme aumenta la concentración de hdl, disminuye el valor del ratio entre el colesterol y el hdl (tal y como era de esperar).

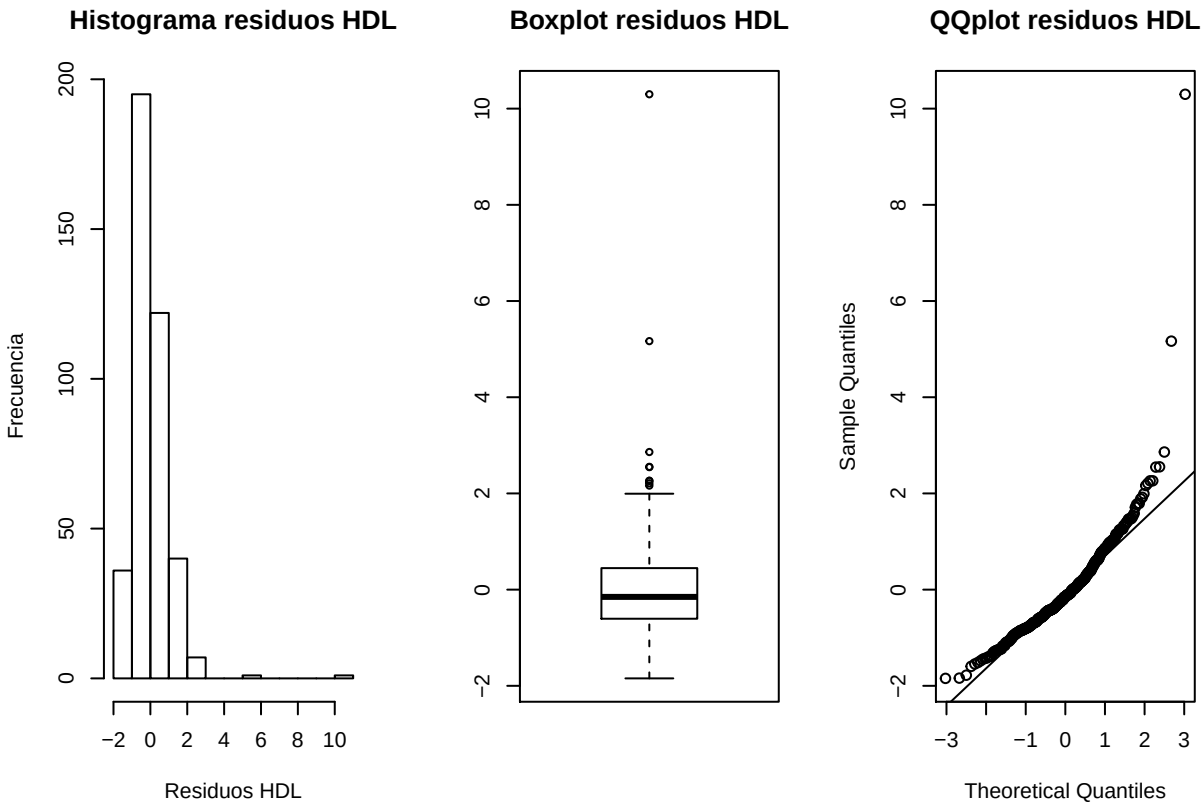
Ahora calculamos los residuos del modelo y representamos los valores en un gráfico de normalidad:

```
residuos_r_hdl<-rstandard(modelo_r_hdl)
par(mfrow=c(1,3))

hist(residuos_r_hdl, main="Histograma residuos HDL", xlab="Residuos HDL",ylab="Frecuencia")

boxplot(residuos_r_hdl, main="Boxplot residuos HDL")

qqnorm(residuos_r_hdl, main="QQplot residuos HDL")
qqline(residuos_r_hdl)
```



En este caso también observamos que los valores se comportan de manera normal y el diagrama de cajas es similar al anterior modelo, sólo que parece tener una mayor precisión (desviación estándar menor). En el diagrama de puntos de los residuos observamos que estos valores siguen de igual manera la línea del modelo aunque de manera menos exacta que el modelo anterior.

5.2. Regresión glyhb con chol y HDL

5.2.1 Glyhb/Chol

Mediante la correlación de pearson podemos ver que existe cierta correlación entre la concentración de hemoglobina glicosilada y la concentración de colesterol y de dhl. En este apartado vamos a realizar dos modelos de regresión para comprobar su significancia.

Utilizamos la variable glyhb como variable explicada y las variables chol y HDL como explicativas para cada modelo.

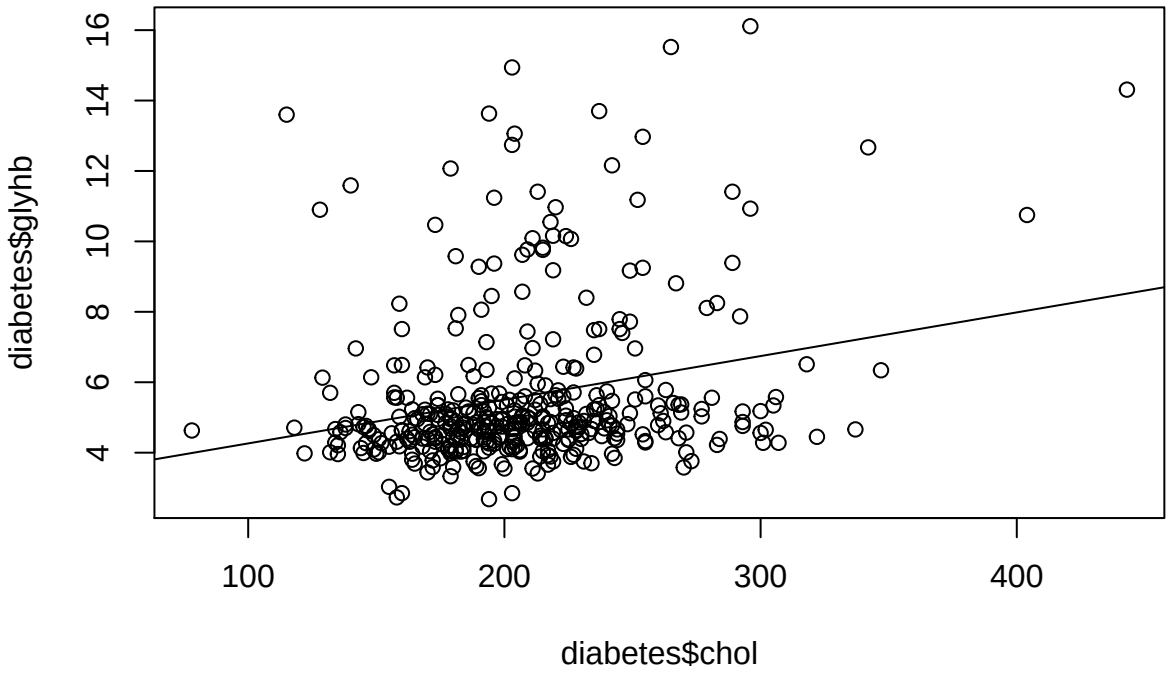
```
modelo_glyhb_chol<- lm(diabetes$glyhb ~ diabetes$chol)
summary(modelo_glyhb_chol)
```

```
##
## Call:
## lm(formula = diabetes$glyhb ~ diabetes$chol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7911 -1.2187 -0.6495  0.1696  9.4164
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.022246   0.524240   5.765 1.67e-08 ***
## diabetes$chol 0.012403   0.002472   5.017 8.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.178 on 387 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.06106,    Adjusted R-squared:  0.05863
## F-statistic: 25.17 on 1 and 387 DF,  p-value: 8.032e-07
```

A pesar de que el valor r es muy bajo el p-valor sigue siendo menor a 0.05, lo que nos indica que estos datos están correlacionados. Podremos observarlo con mayor detenimiento en la representación de los datos:

```
plot(diabetes$chol,diabetes$glyhb)
abline(modelo_glyhb_chol)
```



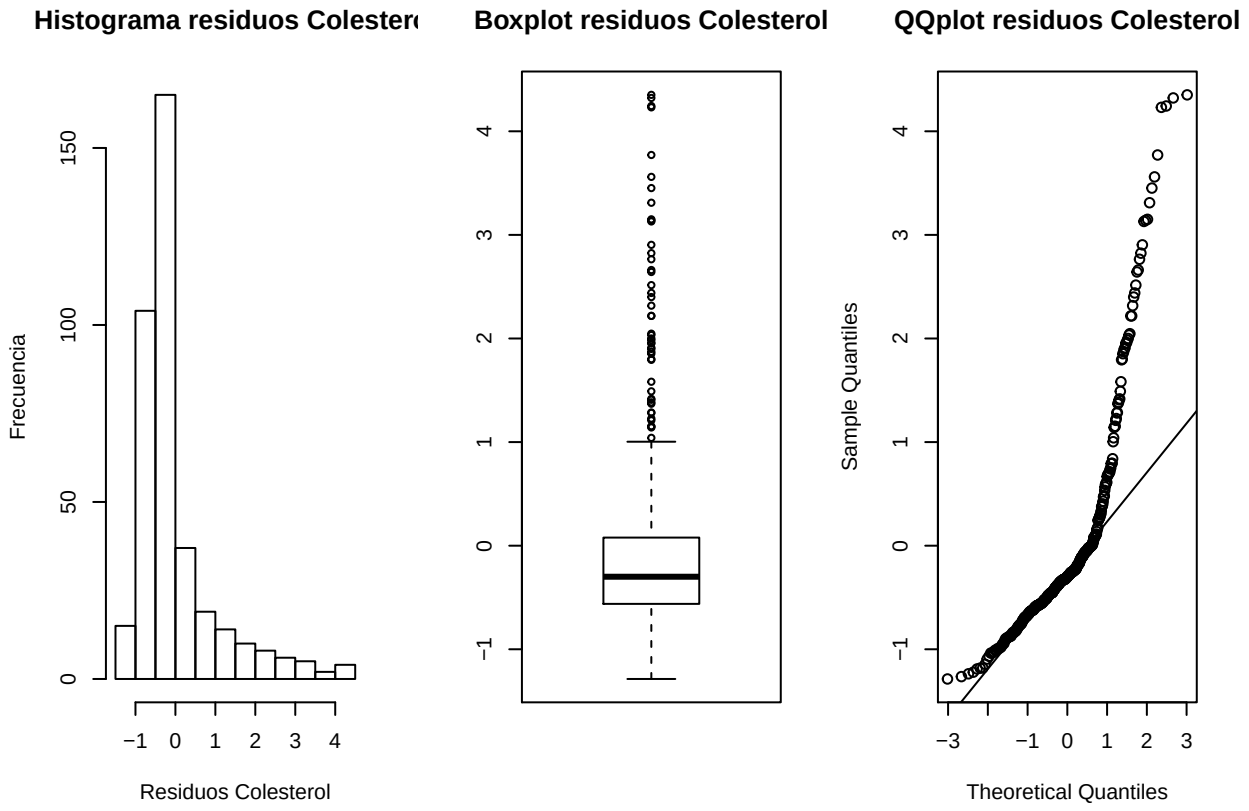
Debido al problema de tener pocos individuos con diabetes, encontramos el por qué de la dificultad del modelo en ajustar los datos. En este caso se puede intuir una sutil correlación entre estas variables con un aumento de probabilidad de tener diabetes conforme aumenta la concentración de colesterol. Para poder confirmar esta correlación serían necesarios más pacientes. Debido a que la representación de los datos es pobre, no se puede confirmar la correlación entre estas dos variables.

```
residuos_glyhb_chol<-rstandard(modelo_glyhb_chol)
par(mfrow=c(1,3))

hist(residuos_glyhb_chol, main="Histograma residuos Colesterol", xlab="Residuos Colesterol",
     ylab="Frecuencia")
```

```r
boxplot(residuos_glyhb_chol, main="Boxplot residuos Colesterol")

qqnorm(residuos_glyhb_chol, main="QQplot residuos Colesterol")
qqline(residuos_glyhb_chol)
```



En estas representaciones observamos como existe una normalidad de estos datos pero poco significativa. De igual manera, en el diagrama de cajas se observa una significativa cantidad de outliers y la media del diagrama no se encuentra centrada en 0. Cabe destacar también que los residuos no siguen correctamente el modelo normal (visible en el diagrama de puntos) en gran parte de éste.

5.2.2 glyhb/HDL

Ahora realizamos el mismo proceso para el modelo de regresión entre la hemoglobina glicosilada (explicada) y el hdl (explicativa):

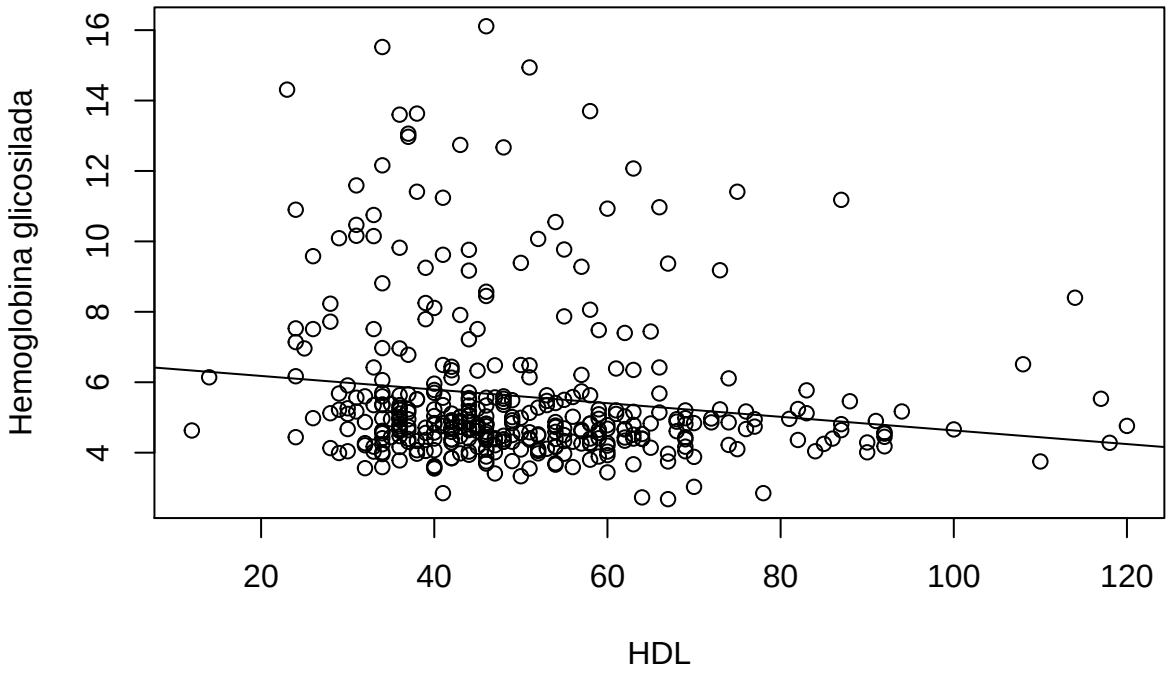
```r
modelo_glyhb_hdl<- lm(diabetes$glyhb ~ diabetes$hdl)
summary(modelo_glyhb_hdl)
```

```
##
## Call:
## lm(formula = diabetes$glyhb ~ diabetes$hdl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9224 -1.2092 -0.7061  0.0686 10.4343
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.565734   0.346607  18.943  < 2e-16 ***
## diabetes$hdl -0.019348   0.006521  -2.967  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.222 on 387 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.02224,    Adjusted R-squared:  0.01972
## F-statistic: 8.804 on 1 and 387 DF,  p-value: 0.003192
```

Obtenemos un ajuste pobre, y esta vez aunque sigue pareciendo haber una correlación significativa, esta significancia es menor que la anterior.

```
plot(diabetes$hdl,diabetes$glyhb, xlab="HDL", ylab="Hemoglobina glicosilada")
abline(modelo_glyhb_hdl)
```



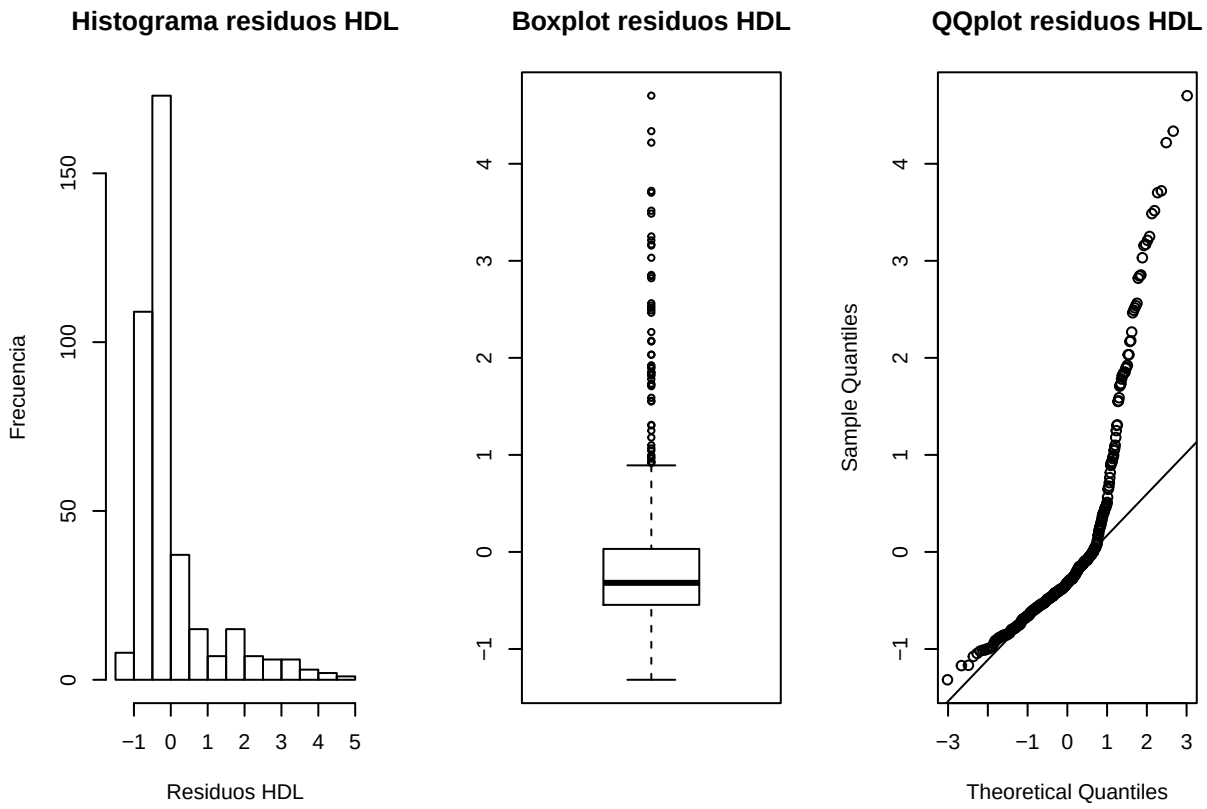
En este caso también tenemos el problema de poca representación de datos, aunque se intuye de manera muy sutil una correlación negativa entre la hemoglobina glicosilada y el HDL.

```
residuos_glyhb_hdl<-rstandard(modelo_glyhb_hdl)
par(mfrow=c(1,3))

hist(residuos_glyhb_hdl, main="Histograma residuos HDL", xlab="Residuos HDL",
     ylab="Frecuencia")

boxplot(residuos_glyhb_hdl, main="Boxplot residuos HDL")
```

```r
qqnorm(residuos_glyhb_hdl, main="QQplot residuos HDL")
qqline(residuos_glyhb_hdl)
```



Ocurre en este modelo el mismo problema que el modelo anterior (normalidad dudosa con una elevada cantidad de outliers y un pobre ajuste del modelo de normalidad en los residuos), por lo que podemos decir que la correlación no parece ser significativa debido a la gran cantidad de residuos que dificultan la creación de un modelo correctamente ajustado.

5.3. Regresión entre edad y colesterol

5.3.1 Edad/colesterol

Hemos encontrado una significativa relación entre la edad de los pacientes y la concentración de hemoglobina glicosilada de éstos, por lo que hemos creído conveniente realizar un modelo de regresión lineal para observar si esta correlación es significativa. Utilizamos el modelo de regresión lineal utilizando como variable explicada la edad y como variable explicativa.

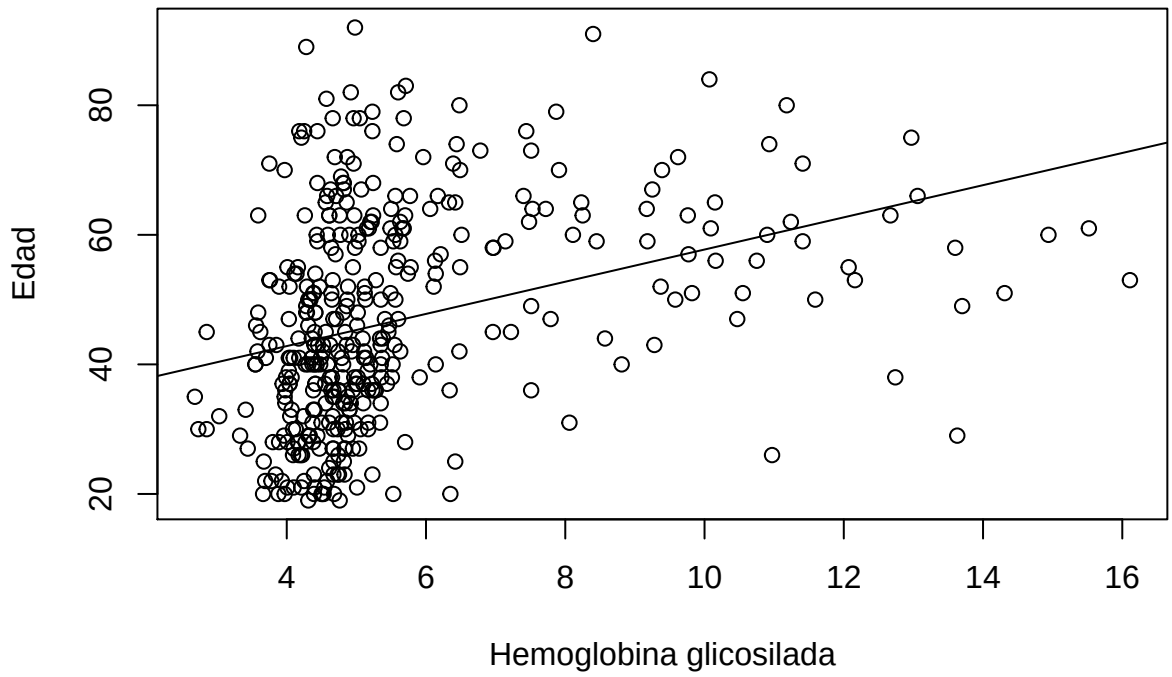
```r
modelo_age_glyhb<- lm(diabetes$age ~ diabetes$glyhb)
summary(modelo_age_glyhb)
```

```
##
## Call:
## lm(formula = diabetes$age ~ diabetes$glyhb)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37.75 -11.07  -2.17  10.79  46.74
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      32.889      2.108  15.603  < 2e-16 ***
## diabetes$glyhb    2.484      0.350   7.096 6.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.48 on 388 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.1149, Adjusted R-squared:  0.1126
## F-statistic: 50.36 on 1 and 388 DF,  p-value: 6.138e-12
```

Observamos que el valor ajustado de R al cuadrado es bajo, indicando un bajo ajuste del modelo con los datos de las variables. Por otro lado, los p-valores para el intercepto y la variable explicada indican que estos resultados son significativos, pues permiten rechazar la hipótesis nula (p valor menor a 0.05). Con estos datos se indica con cierta seguridad que estas dos variables tienen una correlación consistente. De la misma manera que en anteriores modelos, podemos obtener una gráfica de estos valores para observar visualmente la correlación de las variables y la línea del modelo de regresión

```
plot(diabetes$glyhb,diabetes$age, xlab="Hemoglobina glicosilada", ylab="Edad")
abline(modelo_age_glyhb)
```



Aquí podemos identificar la correlación entre la edad y la concentración de hemoglobina glicosilada. También entendemos el pobre ajuste del modelo respecto a los datos, ya que la mayoría de los pacientes de distintas edades tienen una concentración de hemoglobina baja. Esto provoca que la pendiente del modelo no pueda ser determinada de manera eficiente. Observamos que hay una elevada cantidad de pacientes entre los

20 y los 60 años que tienen una concentración de hemoglobina glicosilada entre 4 y 6, indicando que a la mayoría de los pacientes no se les ha diagnosticado diabetes (como habíamos contado anteriormente en las variables `p_diabetes` y `p_sano` con 60 y 330 pacientes respectivamente). Aun así, se puede observar una ligera correlación en que a mayor edad, mayor probabilidad de tener diabetes. Esta hipótesis surje de que a pesar de haber un bajo porcentaje de los pacientes con diabetes, se observa que estos tienen mayoritariamente una edad comprendida entre los 50 y 70 años, mientras que en edades entre 20 y 50 años la cantidad de pacientes con diabetes es negligible.

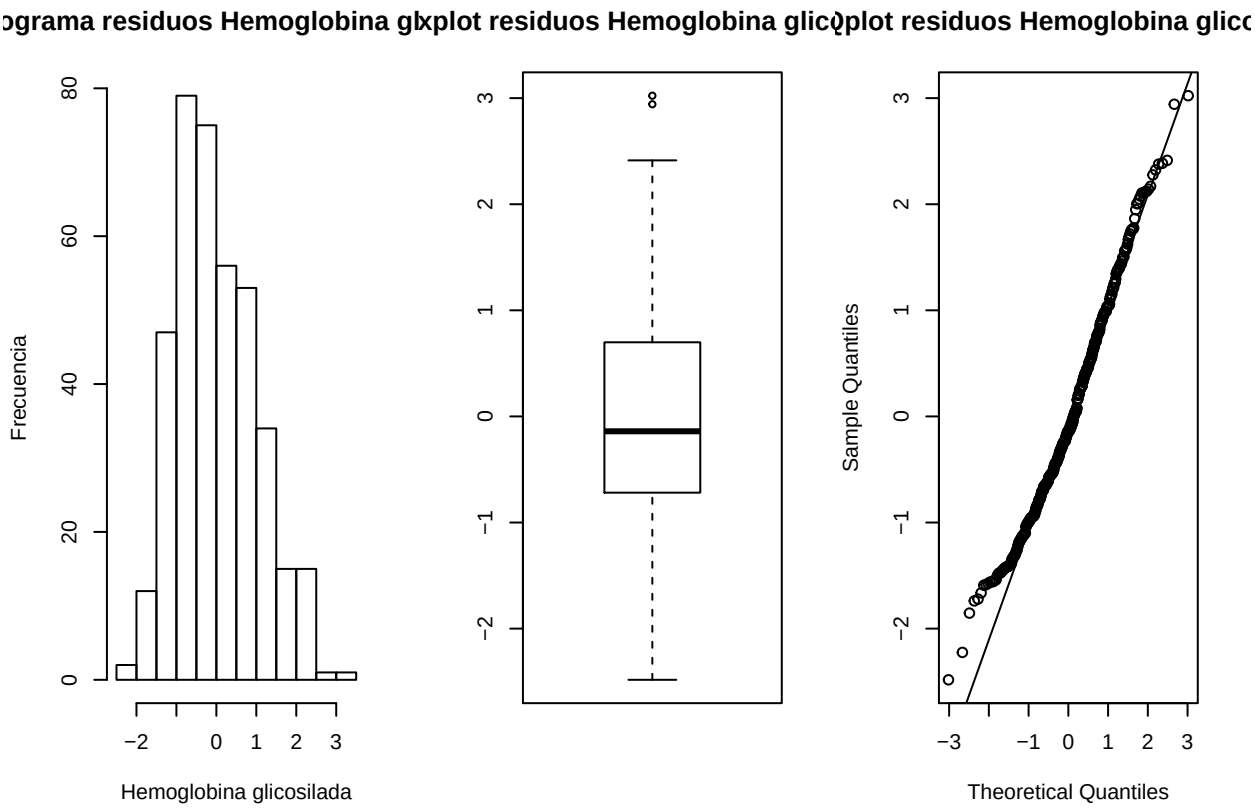
Ahora calculamos los residuos del modelo y representamos los valores en un gráfico de normalidad:

```
residuos_age_glyhb<-rstandard(modelo_age_glyhb)
par(mfrow=c(1,3))

hist(residuos_age_glyhb, main="Histrograma residuos Hemoglobina glicosilada",
     xlab="Hemoglobina glicosilada", ylab="Frecuencia")

boxplot(residuos_age_glyhb, main="Boxplot residuos Hemoglobina glicosilada")

qqnorm(residuos_age_glyhb, main="QQplot residuos Hemoglobina glicosilada")
qqline(residuos_age_glyhb)
```



En estas gráficas observamos que los residuos se comportan de manera normal y que los valores tienen una media cercana a 0 y una desviación estándar entre 0.5 y 1, aunque sus outliers llegan a valores significativamente alejados de esta media y desviación estándar. En el diagrama de cajas observamos que los residuos ajustan casi perfectamente con el modelo normal.

5.4 Regresión imc con tamaño de cadera y cintura

5.4.1 IMC/cadera

Aunque el motivo sea lógico, hemos observado una alta correlación entre el imc de los pacientes con sus diámetros de cintura y cadera, por lo que también podemos realizar dos modelos de regresión entre estas variables. Cabe mencionar que el peso también tiene una alta correlación con los diámetros de cintura y cadera de los pacientes, pero hemos pensado en realizar sólo modelos de regresión utilizando los valores de IMC de los pacientes para no generar datos redundantes.

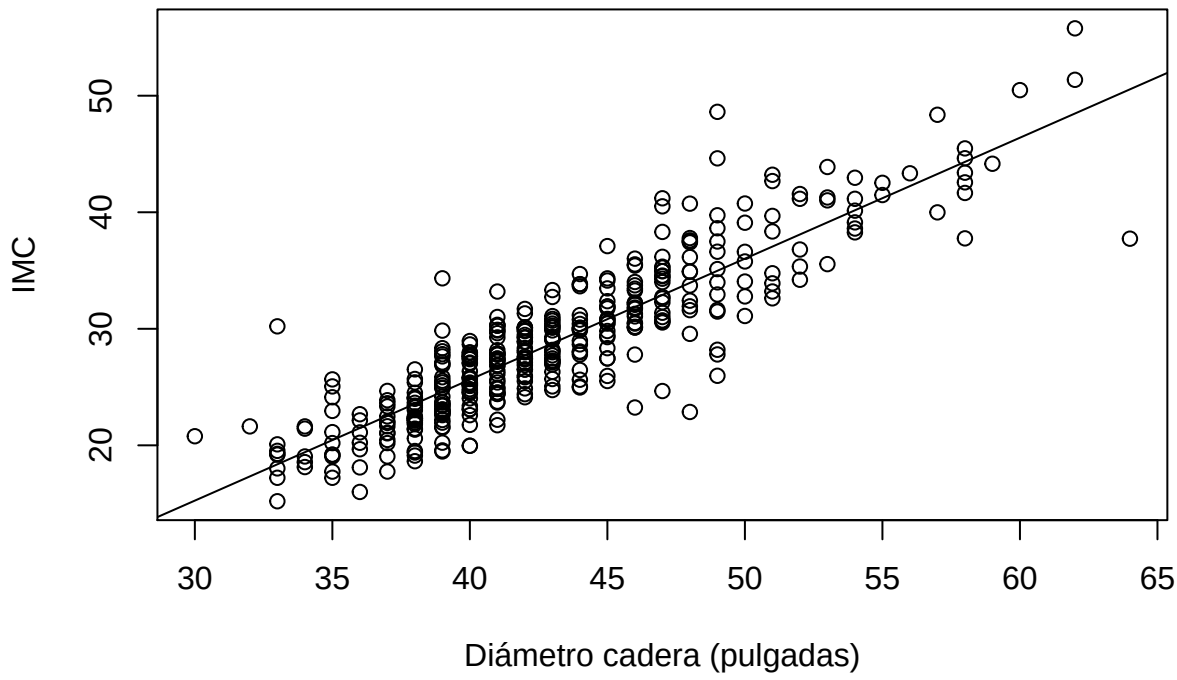
Primero realizamos el modelo de regresión tomando la variable IMC como explicada y el tamaño de cadera como variable explicativa:

```
modelo_imc_hip<- lm(diabetes$imc ~ diabetes$hip)
summary(modelo_imc_hip)
```

```
##
## Call:
## lm(formula = diabetes$imc ~ diabetes$hip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8145  -1.7490  -0.1593   1.8197  13.6392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15.88037    1.19284  -13.31   <2e-16 ***
## diabetes$hip   1.03793    0.02747   37.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.079 on 393 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.7841, Adjusted R-squared:  0.7836
## F-statistic:  1427 on 1 and 393 DF,  p-value: < 2.2e-16
```

El valor de R ajustado elevado indica que el modelo se ajusta a los datos de manera correcta (aunque no perfecta) y gracias al p valor menor a 0.05 sabemos que las dos variables tienen una correlación significativa. Podemos confirmar esto mediante la representación de los datos junto con la línea del modelo:

```
plot(diabetes$hip,diabetes$imc, xlab="Diámetro cadera (pulgadas)", ylab="IMC")
abline(modelo_imc_hip)
```



Se puede observar que existe una tendencia de correlación positiva en estos datos. Cuanto mayor el imc, los diámetros de cadera de los pacientes aumentan de manera clara

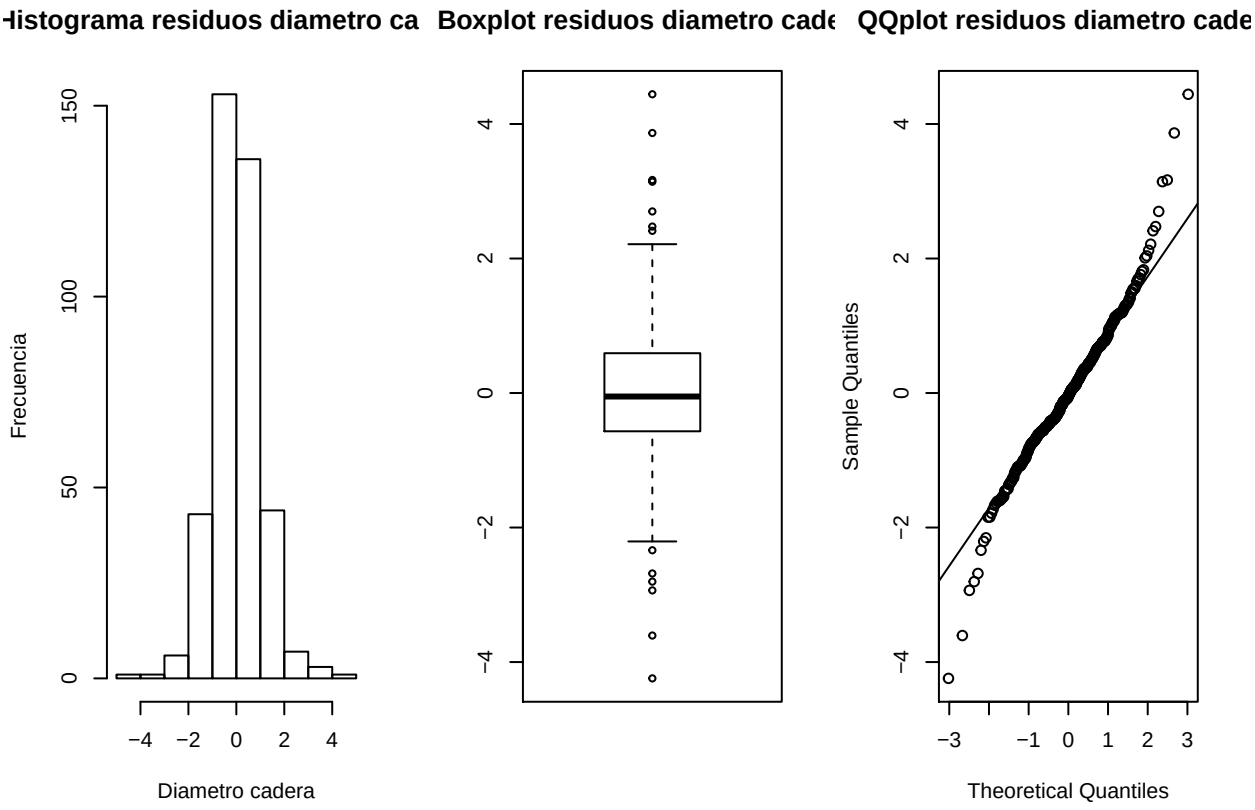
Ahora calculamos los residuos del modelo y representamos los valores en un gráfico de normalidad:

```
residuos_imc_hip<-rstandard(modelo_imc_hip)
par(mfrow=c(1,3))

hist(residuos_imc_hip, main="Histograma residuos diametro cadera", xlab="Diametro cadera",
     ylab="Frecuencia")

boxplot(residuos_imc_hip, main="Boxplot residuos diametro cadera")

qqnorm(residuos_imc_hip, main="QQplot residuos diametro cadera")
qqline(residuos_imc_hip)
```



En el histograma observamos que los datos se comportan de manera normal. En el diagrama de cajas obtenemos una media en 0 y una pequeña desviación estándar, aunque existe la presencia de outliers significativamente alejados de estos valores. En el diagrama de puntos observamos que los residuos siguen de manera ajustada la línea del modelo de residuos (excepto al inicio y final del modelo). Con estos datos podemos confirmar con seguridad que existe una correlación positiva entre el imc y el diámetro de la cadera de los pacientes (el tamaño de la cadera de los pacientes aumenta conforme aumenta el imc de éstos).

5,4.2 IMC/cintura

Ahora realizamos el modelo de regresión con la variable imc como explicada y el tamaño de cintura como variable explicativa:

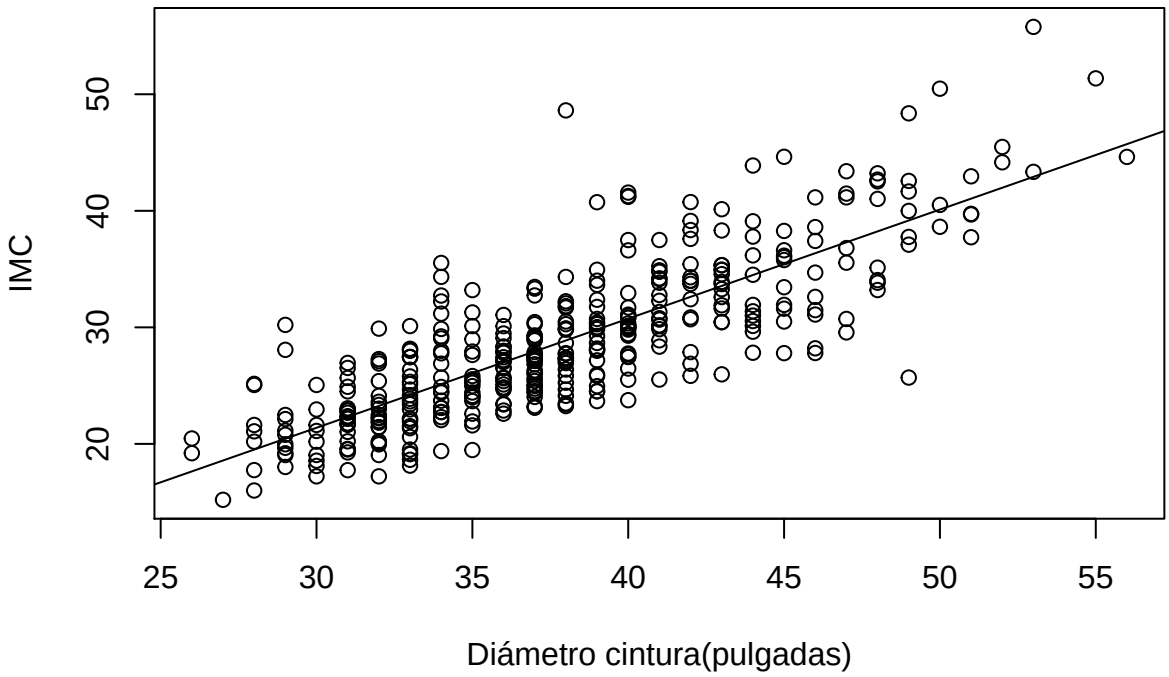
```
modelo_imc_waist<- lm(diabetes$imc ~ diabetes$waist)
summary(modelo_imc_waist)
```

```
##
## Call:
## lm(formula = diabetes$imc ~ diabetes$waist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4797  -2.3956  -0.4817   2.1211  19.7431
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.70080    1.29918  -5.158 3.97e-07 ***
## diabetes$waist  0.93618    0.03387  27.639  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.862 on 393 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6603, Adjusted R-squared:  0.6594
## F-statistic: 763.9 on 1 and 393 DF,  p-value: < 2.2e-16
```

El valor de R ajustado elevado indica que el modelo se ajusta a los datos, aunque con menor precisión que el modelo anterior. El p valor es menor a 0.05 y por lo tanto sabemos que las dos variables tienen una correlación significativa. Realizamos las representaciones de las variables junto con el modelo para observar esta correlación:

```
plot(diabetes$waist,diabetes$imc, xlab="Diámetro cintura(pulgadas)", ylab="IMC")
abline(modelo_imc_waist)
```



De igual manera que en el modelo anterior, se puede observar que existe una tendencia de correlación positiva en estos datos: Cuanto mayor el IMC, mayores los diámetros de cintura.

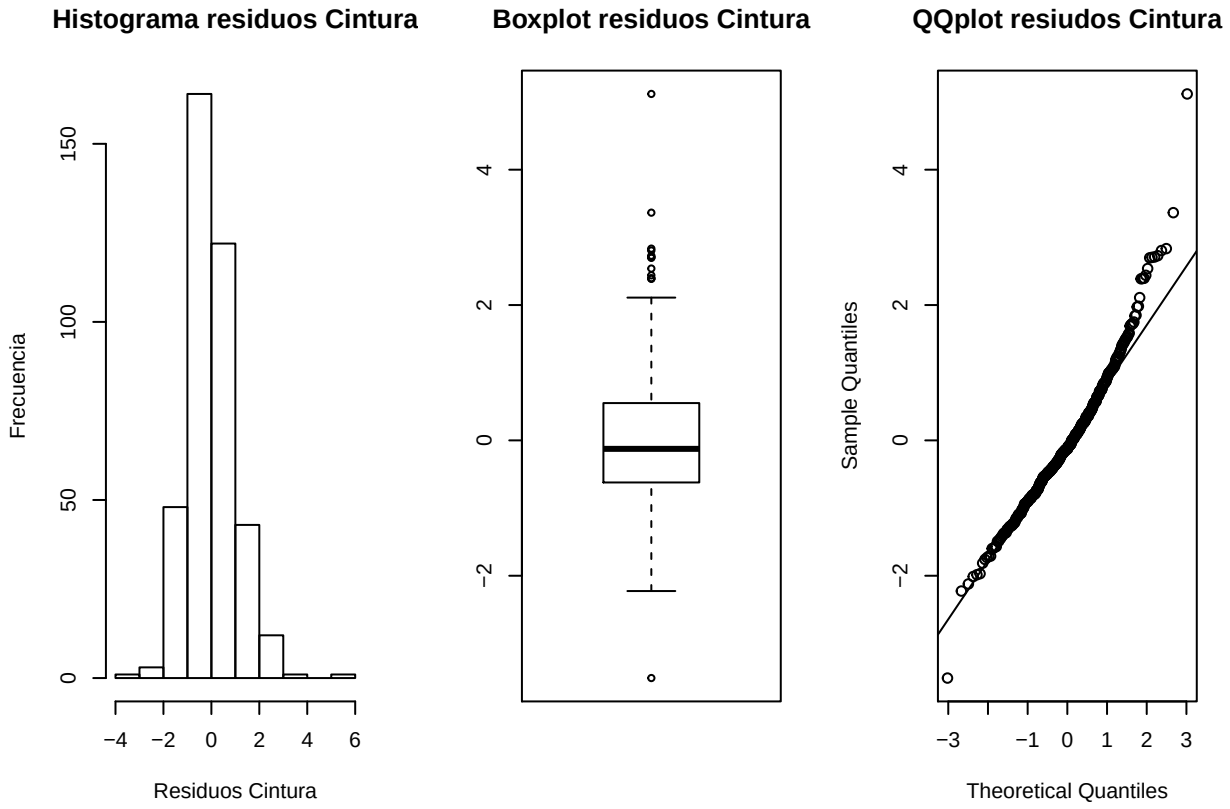
Ahora calculamos los residuos del modelo y representamos los valores en un gráfico de normalidad:

```
residuos_imc_waist<-rstandard(modelo_imc_waist)
par(mfrow=c(1,3))

hist(residuos_imc_waist, main="Histograma residuos Cintura", xlab="Residuos Cintura",
     ylab="Frecuencia")

boxplot(residuos_imc_waist, main="Boxplot residuos Cintura")
```

```
qqnorm(residuos_imc_waist, main="QQplot resiudos Cintura")
qqline(residuos_imc_waist)
```



En estas gráficas se puede observar que residuos siguen una distribución normal, con una baja cantidad de outliers, una media cercana a cero y desviación estándard baja. En el diagrama de puntos de los residuos observamos que los datos se ajustan al modelo. Por lo tanto, afirmamos también que existe una correlación positiva entre el imc y el diámetro de la cintura de los pacientes (el tamaño de la cintura de los pacientes aumenta conforme aumenta su imc).

6) Realizar, a partir de los conceptos trabajados en el LAB4 y la PEC2, un estudio probabilístico (a elección propia) de al menos 3 de las variables, que ayude a esclarecer cuestiones de relevancia que se plantean en los ámbitos de acción estudiados.

Distribución binomial de la variable glyhb

A partir de las variables que muestran los pacientes control y diabéticos en Louisiana y Buckingham, podemos obtener un probabilidad de adquirir diabetes según el porcentaje de afectados.

```
total_L<- nrow(Ldiabeticos) + nrow(Lcontrol)
total_B<- nrow(Bdiabeticos) + nrow(Bcontrol)
porcentaje_diab_L<-nrow(Ldiabeticos) / total_L
porcentaje_diab_B<-nrow(Bdiabeticos) / total_B
resultado1 <- data.frame("Louisa" = porcentaje_diab_L, "Buckingham"= porcentaje_diab_B)
```

```
resultado1
```

```
##   Louisa Buckingham
## 1  0.145  0.1631579
```

Obtenemos que un 14.5% de la población del estudio procedente de Louisiana es diabética, mientras que de la muestra de Buckingham el 16.3% padece esta enfermedad. A partir de estos valores podemos hacer varios estudios de probabilidad. Como el porcentaje indica la probabilidad de ser o no diabético, realizaremos los estudios siguiendo una distribución binomial. Hemos observado que los datos del estudio son antiguos (1997) y las dos poblaciones son pequeñas, ya que hay 1663 y 133 personas (en el censo de 2010, por lo que debe haber variado) para Louisa y Buckingham respectivamente. Utilizaremos entonces estos datos para realizar estudios de probabilidad con su tamaño poblacional. Cabe mencionar que al tener una tabla de registros de Buckingham tan pequeña, la probabilidad de tener diabetes en esta población no está fuertemente validada. Es decir, haría falta una mayor población para disminuir el posible margen de error en esta probabilidad (podría incluso tener una probabilidad menor a Louisa, pero la falta de registros nos impide poder realizar de manera fiable el porcentaje). Por ello, aunque trabajemos y comparemos estas dos localizaciones, es recomendable dudar de los resultados finales y a ser posible generar en un futuro estudios con un mayor número de registros.

Probabilidad de que más la mitad de la población sea diabética (hemoglobina glicosilada mayor a 7)

```
censo_L <- 1663
censo_B <- 133
mitad_L_diab<- pbinom(censo_L/2, size = censo_L, prob = porcentaje_diab_L, lower.tail = FALSE)
mitad_B_diab<- pbinom(censo_B/2, size = censo_B, prob = porcentaje_diab_B, lower.tail = FALSE)
mitad_L_diab
```

```
## [1] 5.089227e-256
```

```
mitad_B_diab
```

```
## [1] 1.271177e-19
```

Podemos concluir que en ambas poblaciones es muy improbable que más de la mitad de la población sea diabética.

Probabilidad de que 20 personas o menos sean diabéticas

```
veinte_L_diab <- pbinom(20, size = censo_L, prob = porcentaje_diab_L)
veinte_B_diab <- pbinom(20, size = censo_B, prob = porcentaje_diab_B)
veinte_L_diab
```

```
## [1] 2.895948e-83
```

```
veinte_B_diab
```

```
## [1] 0.3986536
```

Observamos que en este caso la población de Louisa al ser de más de 1000 habitantes también resulta improbable que haya tan solo 20 personas o menos con diabetes en la población total, mientras que en la población de Buckingham ya tiene un 40% de probabilidades de que 20 o menos personas sean diabéticas de toda la población. Esto es debido a que 20 personas en Buckingham es ya un alto porcentaje de la población total (un 15%) y además la probabilidad de aparición de diabetes era un 2% mayor a la de la población de Louisa.

Probabilidad de que exactamente 50 personas sean diabéticas

```
cincuenta_L_diab <- dbinom(50, size = censo_L, prob = porcentaje_diab_L)
cincuenta_B_diab <- dbinom(50, size = censo_B, prob = porcentaje_diab_B)
cincuenta_L_diab
```

```
## [1] 3.699545e-56
```

```
cincuenta_B_diab
```

```
## [1] 2.00922e-09
```

En este caso ambas probabilidades son muy bajas, ya que en un número exacto de personas con diabetes afecta proporcionalmente la población de la muestra total. En esta caso como hablamos de una población de 1000 y 100 habitantes sería muy difícil saber con certeza el número exacto de "aciertos", siendo menos probable acertar en una población de más de 1000 habitantes (Louisa) que en una de más de 100 (Buckingham), aunque igualmente difícil de acertar.

Distribución normal de la variable chol

Unos valores interesantes con los que trabajar y realizar un estudio probabilístico son los niveles de colesterol de la población de nuestra base de datos. Estos valores indican de manera precisa el colesterol acumulado en sangre, y resulta de gran importancia saber esto debido a que niveles de colesterol altos están relacionados a varias enfermedades. Además, en caso de personas con diabetes estos niveles de colesterol son de media más elevados.

En este caso trabajaremos con todos los valores de la base de datos sin diferenciar por localización. Suponemos que los valores de colesterol de los pacientes siguen una distribución normal, por lo que buscamos los valores de la media y desviación estándar de esta variable.

```
media<-mean(diabetes$chol, na.rm = TRUE)
desvest<-sd(diabetes$chol, na.rm = TRUE)
media
```

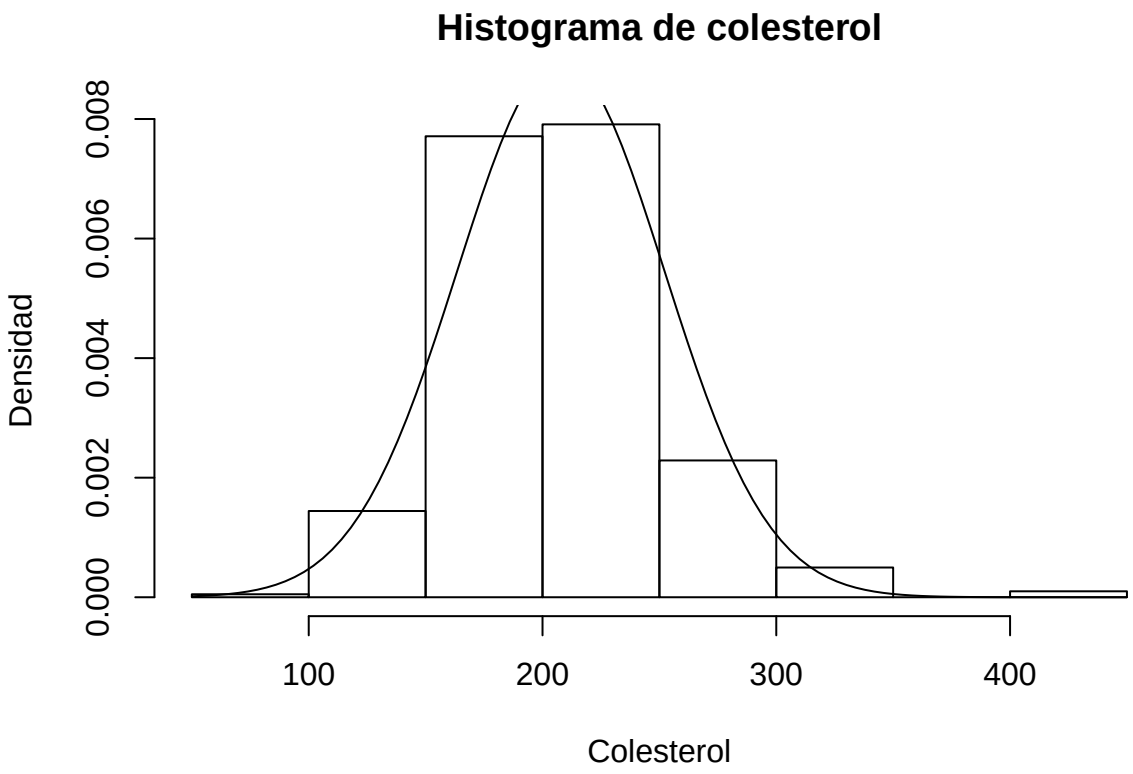
```
## [1] 207.8458
```

```
desvest
```

```
## [1] 44.44556
```

A partir de estos datos podemos superponer una curva de densidad de la distribución normal en el histograma de la variable de hemoglobina glicosilada:

```
hist(diabetes$chol, freq = FALSE, main = "Histograma de colesterol", xlab = "Colesterol",
     ylab = "Densidad")
curve(dnorm(x,mean = media,sd = desvest), add = TRUE)
```



Observamos que los datos de colesterol de nuestra base de datos si parecen seguir una distribución normal, y obtenemos que la media de colesterol es 207.85 y su desviación estándar es 44.45. A partir de estos datos podemos realizar diferentes estudios estadísticos. Para realizar las distintas preguntas nos basaremos en los rangos de colesterol total en los que se subdivide a los pacientes (fuente: https://medlineplus.gov/spanish/pruebas-de-laboratorio/niveles-de-colesterol/)

Probabilidad de que un paciente tenga colesterol por debajo de 200 mg/dL (niveles de colesterol considerados deseables)

```
chol_200<- pnorm(200, mean = media, sd = desvest, lower.tail = TRUE)
chol_200
```

```
## [1] 0.4299406
```

Obtenemos que sólo un 43% de la población del estudio tiene unos niveles deseables de colesterol, indicando entonces que más de la mitad de la población del estudio no tiene unos niveles de colesterol adecuados.

Probabilidad de que un paciente tenga el colesterol entre 200 y 239 mg/dL (niveles de colesterol por encima del rango normal)

```
# Encontramos la diferencia de probabilidades entre 239 mg/dL y 200 mg/dL, por lo que podemos utilizar
# la probabilidad encontrada en el apartado anterior
chol_239<- pnorm(239, mean = media, sd = desvest, lower.tail = TRUE)
chol_200_239 <- chol_239 - chol_200
chol_200_239
```

```
## [1] 0.3283931
```

Esto nos indica que un 33% de la población del estudio tiene un colesterol por encima del rango normal, pero aún no se considera colesterol alto

Probabilidad de que el paciente tenga el colesterol mayor a 240 mg/dL (considerado colesterol alto)

```
chol_mayor_240 <- pnorm(240, mean = media, sd = desvest, lower.tail = FALSE)
chol_mayor_240
```

```
## [1] 0.2347011
```

Encontramos que el 23% de la población del estudio tiene colesterol alto.

Distribución normal de la variable imc

La variable del índice de masa corporal tiene una gran importancia cuando se habla de enfermedades como diabetes (sobretodo la diabetes tipo 2, provocada principalmente por una mala alimentación o falta de ejercicio, promoviendo la obesidad y así un IMC elevado). Podemos realizar un estudio probabilístico de esta variable para observar el porcentaje de gente que se encuentra en los distintos rangos de IMC. Cabe comentar primero que valores tienen estos rangos de IMC(fuente: http://www.calculoimc.com/):

IMC	Clasificación
<16.00	Delgadez severa
16.00-16.99	Delgadez moderada
17.00-18.49	Delgadez aceptable
18.50-24.99	Peso normal
25.00-29.99	Sobrepeso
30.00-34.99	Obeso Tipo I
35.00-40.00	Obeso Tipo II
>40.00	Obeso Tipo III

En nuestro caso dividiremos la clasificación del IMC en tres rangos: delgadez (de 16.00 a 18.49), Peso normal (de 18.50 a 24.99), y sobrepeso (25.00 o superior).
Antes de realizar los estudios de probabilidad debemos considerar cómo se distribuyen los datos de IMC. Creemos que estos datos siguen una distribución normal, por lo que hacemos la misma comprobación que en la variable anterior (chol). Calcularemos la media y desviación estándar de esta variable y realizaremos un histograma con los valores de frecuencia de la variable junto con una curva que indique la curva de una distribución normal.

```
media_imc<-mean(diabetes$imc, na.rm = TRUE)
desvest_imc<-sd(diabetes$imc, na.rm = TRUE)
media_imc
```

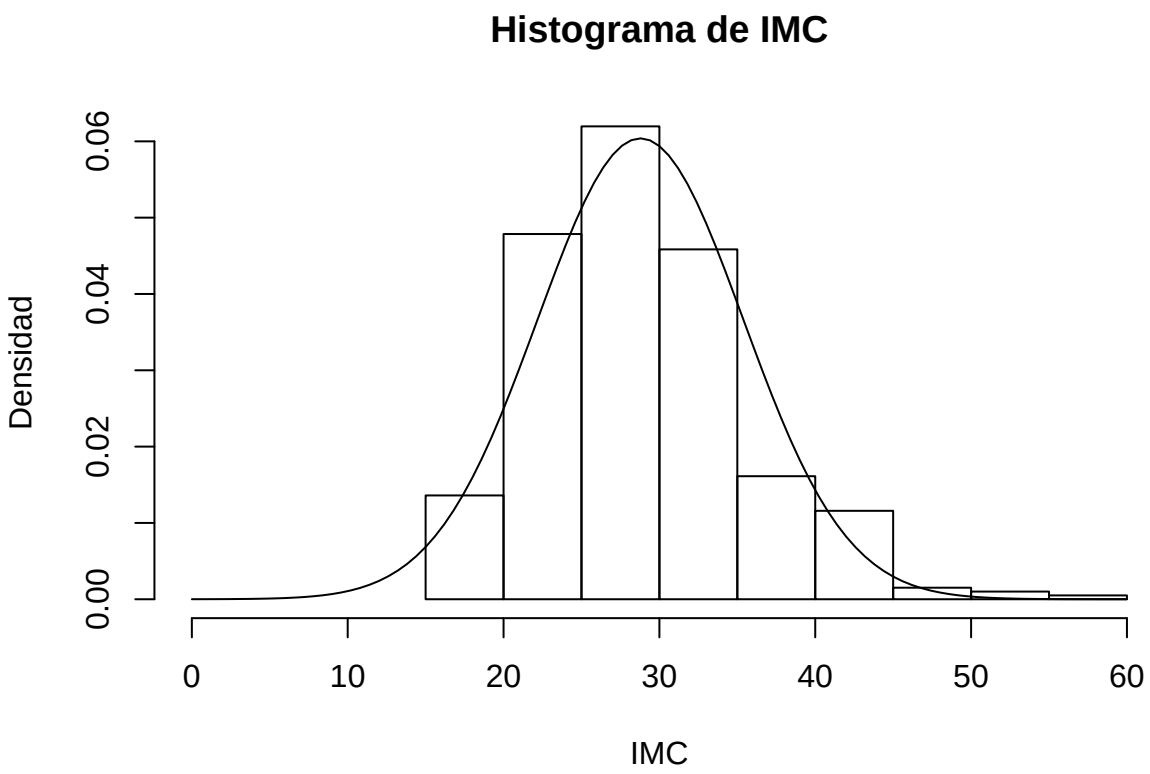
```
## [1] 28.78457
```

```
desvest_imc
```

```
## [1] 6.60589
```

A partir de estos datos podemos superponer una curva de densidad de la distribución normal en el histograma de la variable de hemoglobina glicosilada:

```
hist(diabetes$imc, freq = FALSE, main = "Histograma de IMC", xlab = "IMC", ylab = "Densidad",
     xlim = range(0,60))
curve(dnorm(x,mean = media_imc,sd = desvest_imc), add = TRUE)
```



Observamos en este histograma que como habíamos pensado el IMC parece seguir una distribución normal, por lo que podemos seguir con el estudio probabilístico.

Probabilidad de que un paciente tenga delgadez

```
delgadez <- pnorm(18.49, mean =media_imc, sd= desvest_imc )
delgadez
```

```
## [1] 0.0595701
```

Tan solo hay un 6% de probabilidad de que el paciente tenga un IMC menor a 18.50. Esto indica que en esta población es poco probable que haya pacientes delgados.

Probabilidad de que un paciente tenga peso normal

```
normal <- pnorm(24.99, mean =media_imc, sd= desvest_imc ) - pnorm(18.50, mean = media_imc,
                                                                  sd= desvest_imc )
normal
```

```
## [1] 0.2230914
```

Existe un 22.3% de probabilidad de que el paciente tenga un IMC considerado normal en esta población. Si no tuvieramos en cuenta que el estudio proviene de pacientes en los que se estudia la aparición de diabetes tipo 2 según el ratio de cadera y cintura (por lo que se supone que buscaron pacientes con tendencia a tener sobrepeso), sería una probabilidad preocupante. Por lo tanto esta probabilidad demuestra el alto porcentaje de personas con sobrepeso en la población, porcentaje que vamos a obtener en el siguiente apartado.

Probabilidad de que un paciente tenga sobrepeso

```
sobrepeso <- pnorm(25, mean =media_imc, sd= desvest_imc, lower.tail = FALSE )
sobrepeso
```

```
## [1] 0.7166467
```

Obtenemos que la probabilidad de tener sobrepeso en esta población de estudio es de un 71.6%, indicando consecuentemente que la mayoría de pacientes del estudio tienen sobrepeso de algún tipo (de leve a obesidad).

Como dato a destacar que vale la pena repetir, decir que los pacientes de este estudio fueron seleccionados de manera específica para realizar el estudio por su fisiología y características relacionadas con una posible relación a padecer diabetes tipo 2. Por lo tanto, los valores y porcentajes obtenidos en esta PEC están basados solo en datos muestrales (la base de datos) que no representan de manera fiable las poblaciones de donde proceden los pacientes (Louisa y Buckingham). Igualmente, los resultados obtenidos en este estudio probabilístico de las tres variables elegidas nos ayudan a entender mejor los valores de estas variables en el conjunto de la base de datos.

7) Complementando el apartado anterior, elaborar un análisis ANOVA de dos conjuntos de variables (LAB5 y Ejercicio 6 de la PEC2). La elección de las variables, los resultados, así como su relación deben de estar correctamente. Además, realizar un test cluster de las variables, y si existe un fuerte agrupamiento, elaborar un dendograma (LAB5).

En el siguiente análisis ANOVA nuestra intención será comprobar si existe alguna diferencia en las medias de la variable "hemoglobina glicosilada" entre las dos poblaciones. Ello podría estar relacionado con una mayor propension a contraer esta enfermedad según si un individuo perteneciese a una población u otra. Para ello utilizaremos el factor 'Location'.

No obstante, antes de llevarlo a cabo debemos comprobar que ambos grupos cumplen las condiciones para completar un test ANOVA: -Ambos grupos han de seguir distribuciones normales. -Para ambos grupos debe cumplirse que sus varianzas sean homocedásticas, es decir, que tengan las mismas varianzas.

Para ello en primer lugar pasamos a comprobar que ambas distribuciones son normales mediante un test shaphiro.wilk ya que tenemos muestras lo suficientemente grandes.

```
library(nortest)
by(data = diabetes, INDICES = diabetes$location, FUN = function(x){shapiro.test(x$glyhb)})
```

```
## diabetes$location: Buckingham
##
##  Shapiro-Wilk normality test
##
## data:  x$glyhb
## W = 0.74153, p-value < 2.2e-16
##
## ------------------------------------------------------------
## diabetes$location: Louisa
##
##  Shapiro-Wilk normality test
##
## data:  x$glyhb
## W = 0.69874, p-value < 2.2e-16
```

En este tipo de test la hipotesis nula es que el conjunto de datos sigue una distribución normal y como podemos ver en ambos casos obtenemos un p-valor extremadamente pequeño, lo cual nos hace imposible rechazarla.

Aunque no es necesario, comprobamos mediante un test bartlett la homocedasticidad de ambos grupos.

```
bartlett.test(diabetes$glyhb~diabetes$location)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  diabetes$glyhb by diabetes$location
## Bartlett's K-squared = 0.054488, df = 1, p-value = 0.8154
```

Para este tipo de test la hipotesis nula es que las varianzas son iguales y como podemos ver obtenemos un p-valor elevado lo cual no nos permite rechazar la hipótesis nula.

Este tipo de análisis de normalidad ha sido también llevado sobre el resto de variables. No obstante, no se ha obtenido ninguna en la cual se cumpla una distribución normal, lo cual resulta extraño ya que se trata de muestras suficientemente grandes como para cumplir la regla de los grandes números. Esto nos hace pensar que quizá la recogida de datos quizá no se haya dado de forma aleatoria.

```
by(data = diabetes, INDICES = diabetes$location, FUN = function(x){shapiro.test(x$chol)})
```

```
## diabetes$location: Buckingham
##
##  Shapiro-Wilk normality test
##
## data:  x$chol
## W = 0.94741, p-value = 1.132e-06
##
## ------------------------------------------------------------
## diabetes$location: Louisa
##
##  Shapiro-Wilk normality test
##
## data:  x$chol
## W = 0.96326, p-value = 3.956e-05
```

```
by(data = diabetes, INDICES = diabetes$location, FUN = function(x){shapiro.test(x$imc)})
```

```
## diabetes$location: Buckingham
##
##  Shapiro-Wilk normality test
##
## data:  x$imc
## W = 0.95806, p-value = 1.422e-05
##
## ------------------------------------------------------------
## diabetes$location: Louisa
##
##  Shapiro-Wilk normality test
##
## data:  x$imc
## W = 0.96186, p-value = 3.161e-05
```

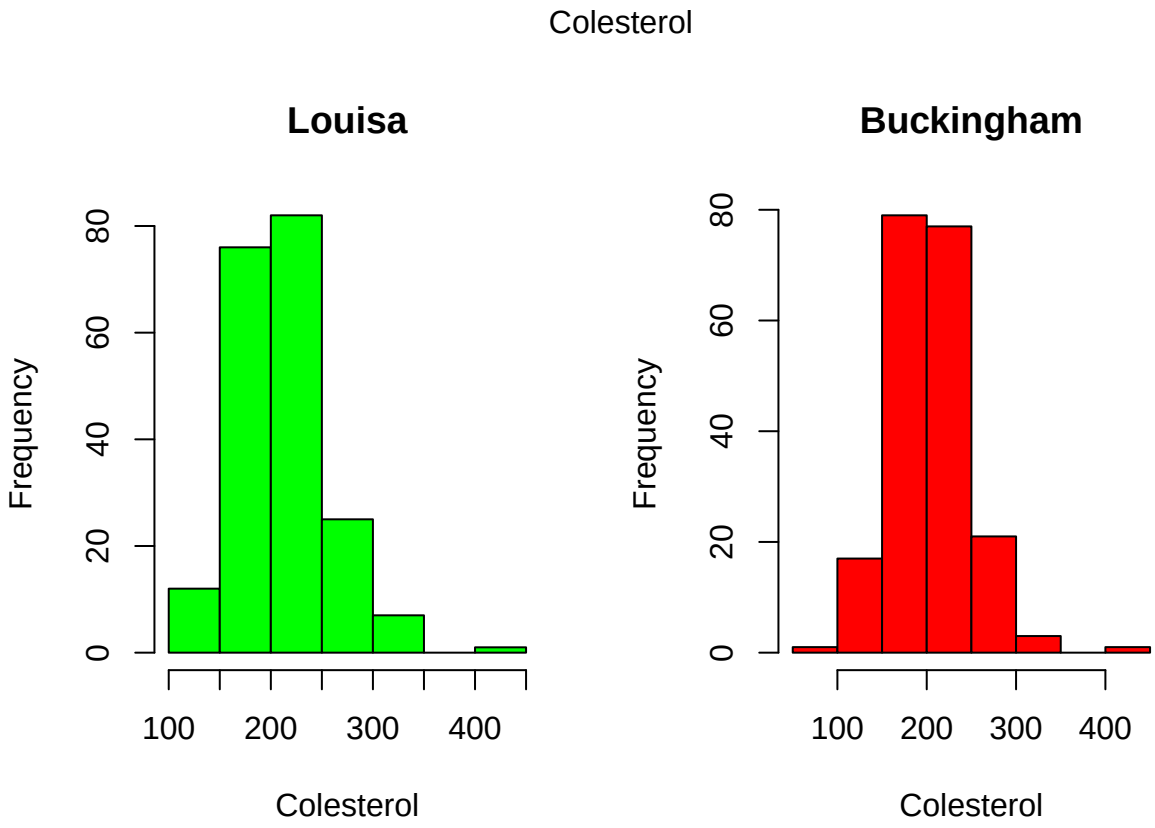
```
by(data = diabetes, INDICES = diabetes$location, FUN = function(x){shapiro.test(x$hdl)})
```

```
## diabetes$location: Buckingham
##
##  Shapiro-Wilk normality test
##
## data:  x$hdl
## W = 0.9425, p-value = 4.032e-07
##
## ------------------------------------------------------------
## diabetes$location: Louisa
##
##  Shapiro-Wilk normality test
##
## data:  x$hdl
## W = 0.89643, p-value = 1.198e-10
```

Dado que no se cumple la normalidad en ninguna variable pasamos a buscar test no paramétricos que sí nos puedan ofrecer alguna conclusión. Dado que estamos llevando a cabo un análisis de las diferencias que encontramos entre ambas poblaciones (nuestro factor) el cual consta de dos niveles o grupos (Louisa y Buckingham), creemos que el test estadistico más adecuado podría ser el de U-Mann-Whitney. Esto es debido a que cumple las siguientes condiciones: -La variable dependiente (Chol, hdl, glyhb... etc) se trata de una variable continua. -La variable independiente (location) se trata de una variable categórica con dos niveles (Louisa y Buckingham). -Las datos recogidos son independientes. -Aunque no siguen normalidad, algunas de ellas si que se tienen cierta tendencia a seguir una distribucion normal, como demostramos a continuación en los histogramas.

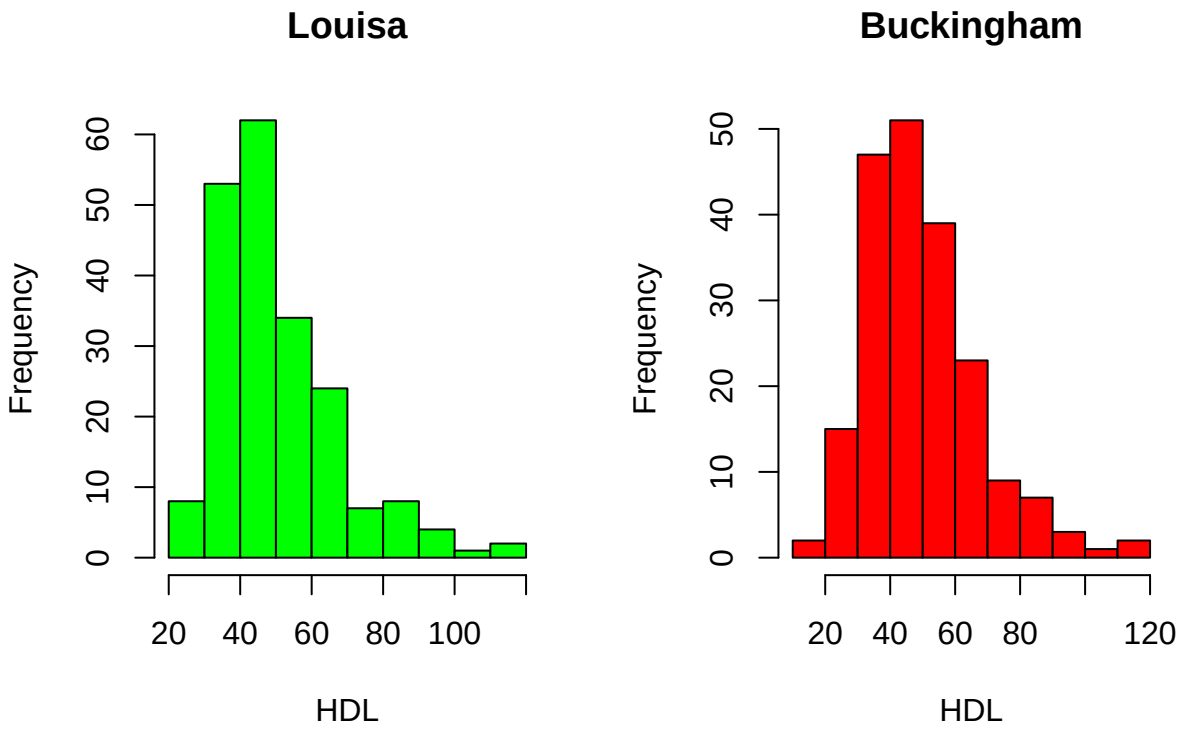
```
LOUISAFINAL<- sqldf("SELECT * FROM diabetes WHERE location='Louisa'")
BUCKINGHAMFINAL<- sqldf("SELECT * FROM diabetes WHERE location='Buckingham'")

par(mfrow=c(1,2), oma = c(0,0,1,0))
hist(LOUISAFINAL$chol, col="green", main="Louisa", xlab="Colesterol")
hist(BUCKINGHAMFINAL$chol, col="red", main="Buckingham", xlab="Colesterol")
mtext("Colesterol", outer = TRUE, cex=1, line=0)
```



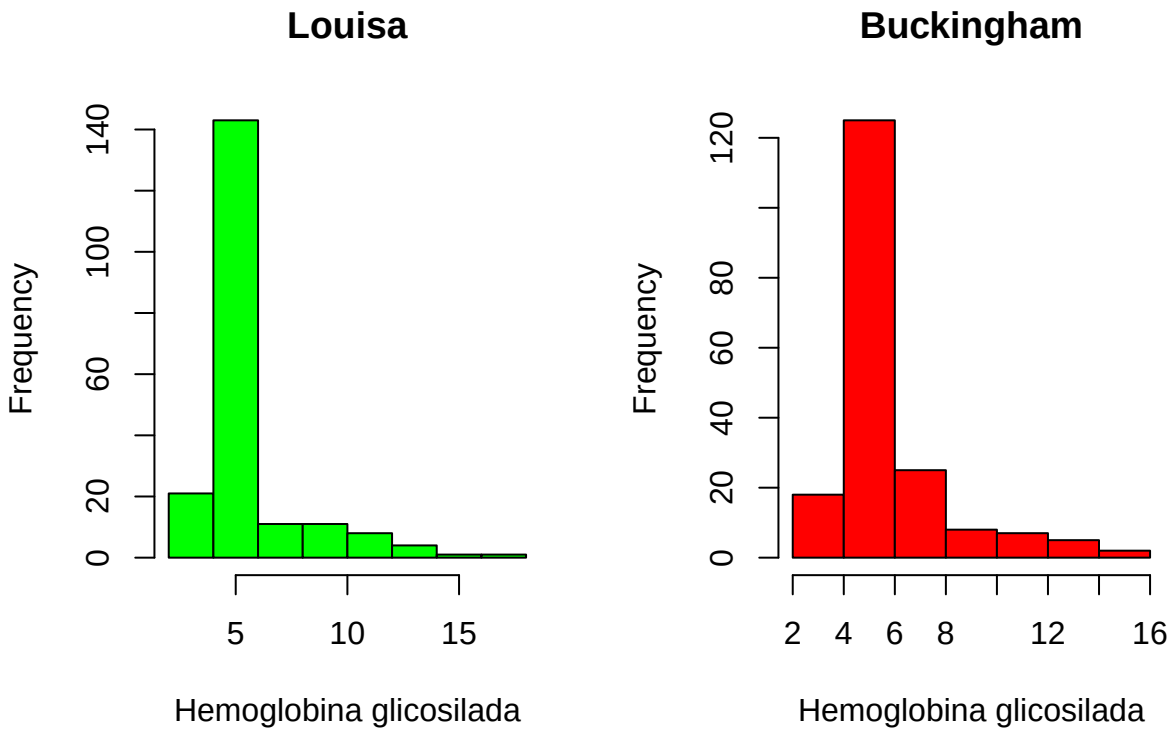

```
par(mfrow=c(1,2), oma = c(0,0,1,0))
hist(LOUISAFINAL$hdl, col="green", main="Louisa", xlab="HDL")
hist(BUCKINGHAMFINAL$hdl, col="red", main="Buckingham", xlab="HDL")
mtext("HDL", outer = TRUE, cex=1, line=0)
```

HDL

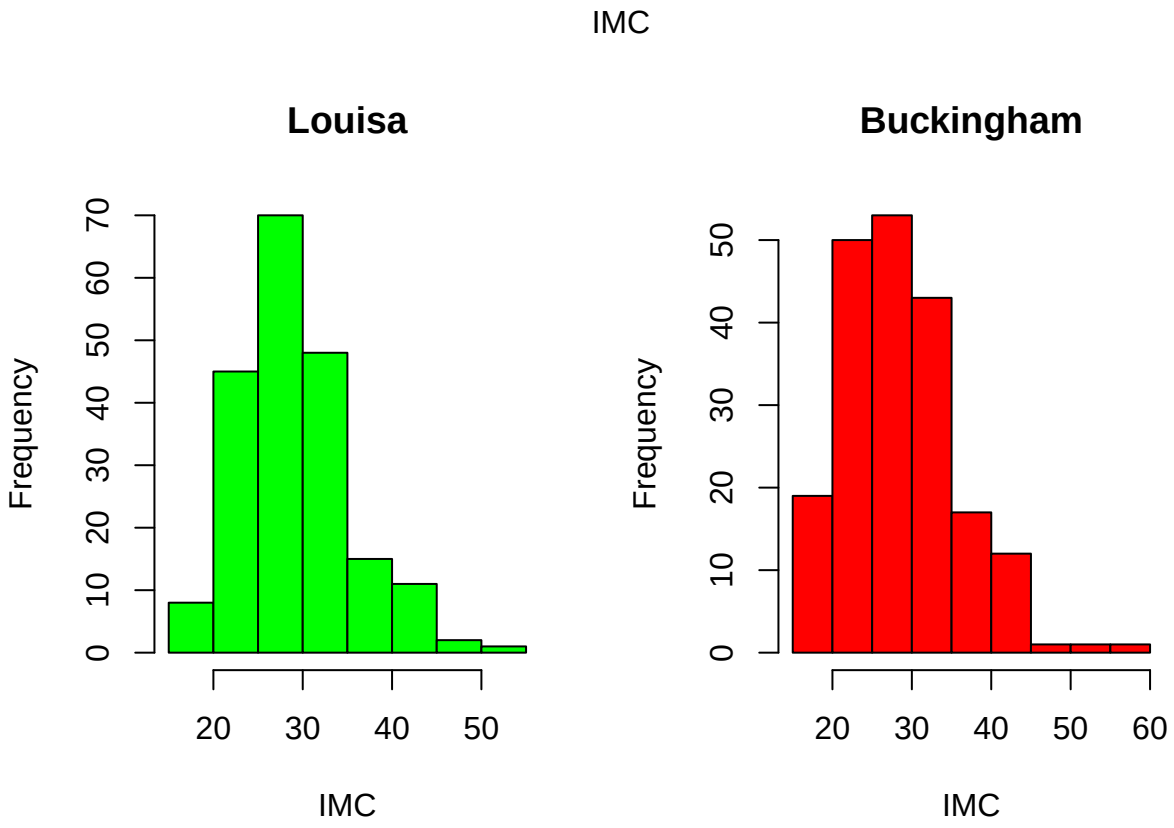



```
par(mfrow=c(1,2), oma = c(0,0,1,0))
hist(LOUISAFINAL$glyhb, col="green", main="Louisa", xlab="Hemoglobina glicosilada")
hist(BUCKINGHAMFINAL$glyhb, col="red", main="Buckingham", xlab="Hemoglobina glicosilada")
mtext("Hemoglobina Glicosilada", outer = TRUE, cex=1, line=0)
```

Hemoglobina Glicosilada




```
par(mfrow=c(1,2), oma = c(0,0,1,0))
hist(LOUISAFINAL$imc, col="green", main="Louisa", xlab="IMC")
hist(BUCKINGHAMFINAL$imc, col="red", main="Buckingham", xlab="IMC")
mtext("IMC", outer = TRUE, cex=1, line=0)
```



Como puede observarse, la variable que menos se asemeja a una distribución normal es la de la hemoglobina glicosilada. Por ello, seguiremos adelante con el resto de ellas. En los test Mann-Whitney U la hipótesis nula es que no existe ninguna diferencia en las variables por el hecho de pertencer a una ciudad u otra:

```
wilcox.test(diabetes$chol ~ diabetes$location, data=diabetes, na.rm=TRUE, paired=FALSE,
            exact=FALSE, conf.int=TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  diabetes$chol by diabetes$location
## W = 18712, p-value = 0.2018
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -13.00001   2.99997
## sample estimates:
## difference in location
##              -5.000035
```

```
wilcox.test(diabetes$hdl ~ diabetes$location, data=diabetes, na.rm=TRUE, paired=FALSE,
            exact=FALSE, conf.int=TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  diabetes$hdl by diabetes$location
## W = 19871, p-value = 0.7788
## alternative hypothesis: true location shift is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.000036  2.000039
## sample estimates:
## difference in location
##          -3.245369e-05
```

```
wilcox.test(diabetes$imc ~ diabetes$location, data=diabetes, na.rm=TRUE, paired=FALSE,
            exact=FALSE, conf.int=TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  diabetes$imc by diabetes$location
## W = 18673, p-value = 0.3692
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -1.8288147  0.6445607
## sample estimates:
## difference in location
##             -0.5585889
```

Dado que el p-valor es en todos los casos mayor que el nivel de significación no nos es posible refutar la hipótesis nula, de forma que podemos decir que pertencer a una ciudad o a otra no tiene influencia sobre estas variables.

8) (1 punto) A partir de los datos de origen y el estudio realizado (incluyendo todos los puntos anteriores), presentar un apartado de conclusiones. Esta sección debe incluir un resumen de los principales resultados obtenidos en apartados anteriores, que ayuden al lector a comprender el ámbito de estudio. Además, se valorará positivamente la coherencia de resultados y las justificaciones presentadas.

La idea principal cuando escogimos esta base de datos fue observar como podían cambiar ciertas variables segun si los individuos perteneciesen a una poblacion u otra. En primer lugar, llevamos a cabo una serie de regresiones lineales con el fin de observar que relación existia entre las variables, de forma que los resultados pudiesen sugerir aquellas que pudiesen ser más interesantes. Realizando varias regresiones hemos visto que existía una correlación en todas las regresiones hechas, pero en unas se ajustaban a los datos de la correlación mejor que otras. En concreto, nos gustaría destacar la correlación entre el Ratio y la concentración de HDL y las correlaciones de IMC con el diámetro de cadera y cintura, ya que estas son las que mejor se han ajustado al modelo de regresión lineal. En la regresión del Ratio con el HDL hemos observado una visible correlación negativa, y como era de esperar, hemos encontrado una clara correlación positiva con un buen ajuste de los datos tanto en la regresión entre IMC y el diámetro de cadera como con la regresión del IMC con el diámetro de la cintura.

Respecto análisis probabilístico ha sido muy interesante observar que 1/4 de las personas que han participado poseian niveles altos de colesterol o que casi 3/4 de los pacientes poseían algún grado de sobrepeso (mientras que los pacientes con delgadez apenas representaban el 6%). Esta información nos ha resultado útil ya que nos ha permitido hacernos una idea del tipo de personas encuestadas en las dos localizaciones en el condado de Virginia, teniendo en cuenta la forma y calidad de vida de las personas en el año de realización del estudio.

Por último, en el ejercicio 7 se ha intentado comprobar si por el hecho de pertencer a una ciudad u otra los habitantes eran mas propensos a tener diferencias en los indicadores de salud, como pueder ser la hemoglobina

glicosilada, el colesterol (estrechamente relacionado con el HDL ya que es su transportador) o el IMC. Como se muestra, se ha intentado llevar a cabo un analisis ANOVA para observar diferencias, no obstante, no ha sido posible su aplicacion ya que incumplía una de las dos condiciones más importantes de este test: ninguna variable seguía una distribución normal. Por ello se ha optado por un test Mann Whitney U ya para éste si que se cumplían las condiciones necesarias. Como era de esperar, el hecho de pertenecer a Louisa o Buckingham no tiene influencia significativa en la salud de los individuos que han participado en este estudio.