Tipología de datos: Práctica 2

Autor: Maria Victoria Diaz

Junio 2021

Contents

Análisis de sobrevivencia por insuficiencias cardiacas: Caso estudio Pakistán.

El dataset que se analizará en este trabajo, es producto de un estudio realizado por Ahman et.al (2017),el cual se centró en el análisis de la supervivencia de los pacientes con insuficiencia cardíaca que fueron admitidos en el Instituto de Cardiología y el hospital aliado de Faisalabad-Pakistán durante el período comprendido entre abril y diciembre (2015). Todos los pacientes tenían 40 años o más, tenían disfunción sistólica ventricular izquierda y pertenecían a la clase III y IV de la NYHA. Se determinaron: la edad, la fracción de eyección, la creatinina sérica, el sodio sérico, la anemia, las plaquetas, la creatinina fosfocinasa, la presión arterial, el sexo, la diabetes y el tabaquismo como factores que pueden contribuir a la mortalidad.

En este trabajo, quiero determinar cuáles de todos estos factores son los que más influyen en la mortalidad por insuficiencias cardíacas, cómo se relacionan, y si existen factores de riesgo diferentes a los aquí considerados para ayudar en un futuro a enfocar los distintos programas de prevención de enfermedades del corazón en la sociedad.

Descripción de las variables

```
## 'data.frame':    299 obs. of  13 variables:
##  $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...
##  $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
##  $ diabetes                : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
##  $ high_blood_pressure     : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...
##  $ platelets               : num  265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...
##  $ smoking                 : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...
##  $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

El estudio se realizó en 299 pacientes con insuficiencia cardíaca, y las variables presentes en el dataset fueron:

- Age: Edad del paciente.
- Anaemia: Si el paciente tiene o no un decenso en los glóbulos rojos.
- High blood pressure: Si el paciente tiene o no hipertensión.
- Creatinine phosphokinase (mcg/L): Nivel de creatinina fosfoquinasa en la sangre.
- Diabetes: Si el paciente tiene o no diabetes.
- Ejection fraction Percentage: Porcentaje de sangre saliendo del corazón en cada contracción.
- Sex: 0 - Mujer; 1- Hombre.
- Plateletes (kiloplatelets/mL): Cantidad de plaquetas en la sangre.
- Serum creatinine: Nivel de creatinina presente en la sangre.
- Serum sodium: Nivel de sodio presente en la sangre.
- Smoking: Si el paciente fuma o no.
- Time: Tiempo de seguimiento al paciente.
- Death event: Si el paciente murió o no durante el período de seguimiento.

Análisis exploratorio de variables cuantitativas

En primer lugar, analizaré las variables cuantitativas una a una y miraré si hay presencia de valores perdidos o también de valores extremos:

```
heart_c<- heart %>% select(age, creatinine_phosphokinase,
                           ejection_fraction, platelets, serum_creatinine,
                           serum_sodium, time, DEATH_EVENT)

heart_c %>% split(.$DEATH_EVENT) %>% map(summary)
```

```
## $`0`
##       age        creatinine_phosphokinase ejection_fraction
##  Min.   :40.00   Min.   :  30.0           Min.   :17.00
##  1st Qu.:50.00   1st Qu.: 109.0           1st Qu.:35.00
##  Median :60.00   Median : 245.0           Median :38.00
##  Mean   :58.76   Mean   : 540.1           Mean   :40.27
##  3rd Qu.:65.00   3rd Qu.: 582.0           3rd Qu.:45.00
##  Max.   :90.00   Max.   :5209.0           Max.   :80.00
##    platelets      serum_creatinine  serum_sodium        time
##  Min.   : 25100   Min.   :0.500    Min.   :113.0   Min.   : 12.0
##  1st Qu.:219500   1st Qu.:0.900    1st Qu.:135.5   1st Qu.: 95.0
```

```
##  Median :263000   Median :1.000    Median :137.0   Median :172.0
##  Mean   :266658   Mean   :1.185    Mean   :137.2   Mean   :158.3
##  3rd Qu.:302000   3rd Qu.:1.200    3rd Qu.:140.0   3rd Qu.:213.0
##  Max.   :850000   Max.   :6.100    Max.   :148.0   Max.   :285.0
##  DEATH_EVENT
##  0:203
##  1:  0
##
##
##
##
##
## $`1`
##       age        creatinine_phosphokinase ejection_fraction
##  Min.   :42.00   Min.   :  23.0           Min.   :14.00
##  1st Qu.:55.00   1st Qu.: 128.8           1st Qu.:25.00
##  Median :65.00   Median : 259.0           Median :30.00
##  Mean   :65.22   Mean   : 670.2           Mean   :33.47
##  3rd Qu.:75.00   3rd Qu.: 582.0           3rd Qu.:38.00
##  Max.   :95.00   Max.   :7861.0           Max.   :70.00
##    platelets      serum_creatinine  serum_sodium        time
##  Min.   : 47000   Min.   :0.600    Min.   :116.0   Min.   :  4.00
##  1st Qu.:197500   1st Qu.:1.075    1st Qu.:133.0   1st Qu.: 25.50
##  Median :258500   Median :1.300    Median :135.5   Median : 44.50
##  Mean   :256381   Mean   :1.836    Mean   :135.4   Mean   : 70.89
##  3rd Qu.:311000   3rd Qu.:1.900    3rd Qu.:138.2   3rd Qu.:102.25
##  Max.   :621000   Max.   :9.400    Max.   :146.0   Max.   :241.00
##  DEATH_EVENT
##  0: 0
##  1:96
##
##
##
##
```

Los pacientes que sobrevivieron tenían entre 40 y 90 años. Los niveles medios de creatina fosfoquinasa en sangre fueron de 245 mcg/L. El porcentaje de sangre saliendo del corazón en cada contracción iba de 17 a 80% y la cantidad de plaquetas en la sangre iba de 25100 a 850000 kiloplatelets/mL. Los niveles medios de creatinina en sangre eran 1, y de creatinina en sodio eran de 137. Finalmente, a estos pacientes se les hizo seguimiento entre 12 y 285 días.

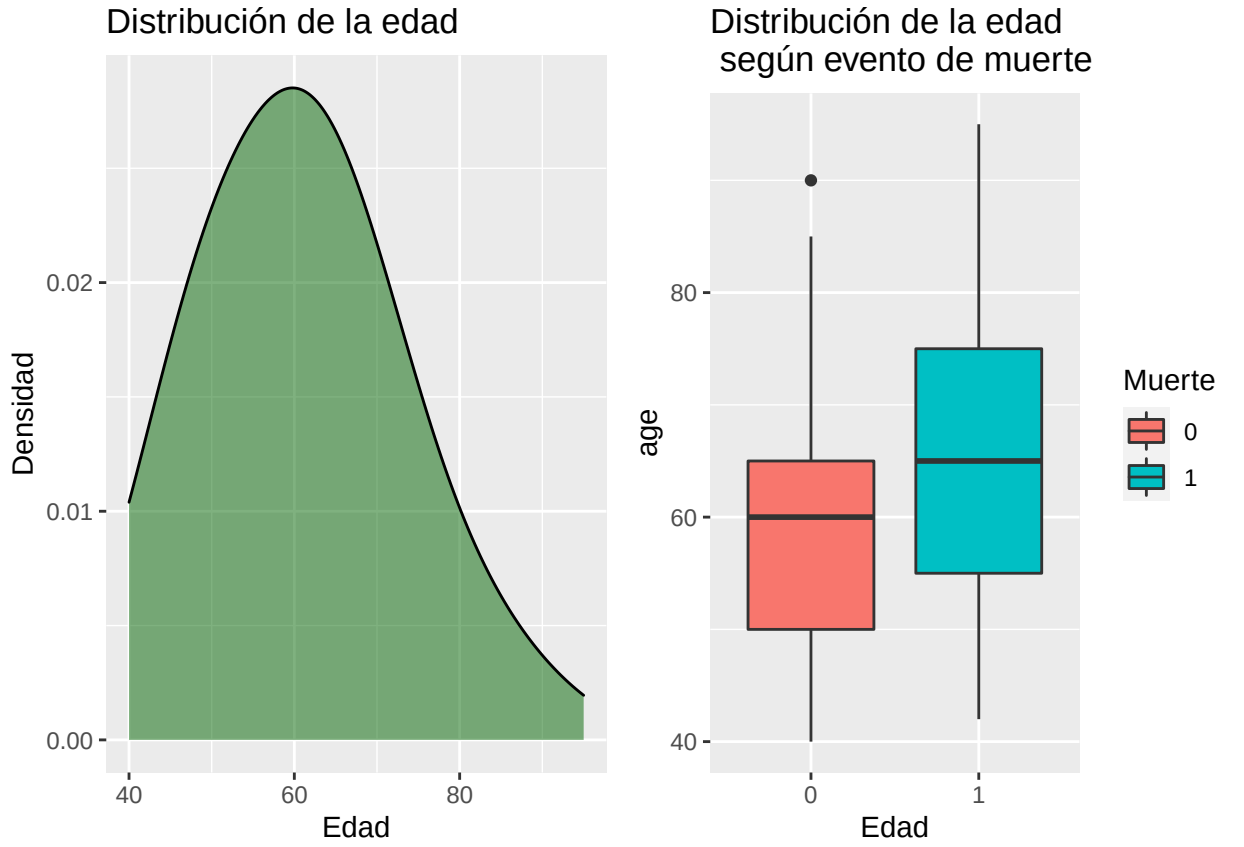
Ahora, los pacientes que fallecieron tenían entre 42 y 95 año. Los niveles medios de creatina fosfoquinasa en sangre fueron de 259 mcg/L. El porcentaje de sangre saliendo del corazón en cada contracción iba de 14 a 70% y la cantidad de plaquetas en la sangre iba de 47000 a 621000 kiloplatelets/mL. Los niveles medios de creatinina en sangre eran 1.3, y de creatinina en sodio eran de 135.5. Finalmente, estos pacientes murieron entre el día 4 y el 241.

Nótese que he reportado los valores de la mediana, ya que con estos resultados se puede anticipar que la distribución de las variables es asimétrica.

Ahora, analizaré la distribución de cada variable y los supuestos de normalidad y homocedasticidad mediante las pruebas de Anderson Darling y Levene, la cual es menos sensible a la no normalidad de los datos:

Edad

```
p1 <- ggplot(heart)+geom_density(aes(age), fill = "darkgreen", alpha=0.5, adjust=2) +
  labs(x="Edad", y="Densidad")+ guides(fill=guide_legend(title=""))+
  ggtitle("Distribución de la edad")

p2 <- ggplot(heart, aes(x = DEATH_EVENT, y = age, fill = DEATH_EVENT))+geom_boxplot() +
  labs(x="Edad")+ guides(fill=guide_legend(title="Muerte")) +
ggtitle("Distribución de la edad \n según evento de muerte")

grid.arrange(p1,p2,ncol = 2)
```

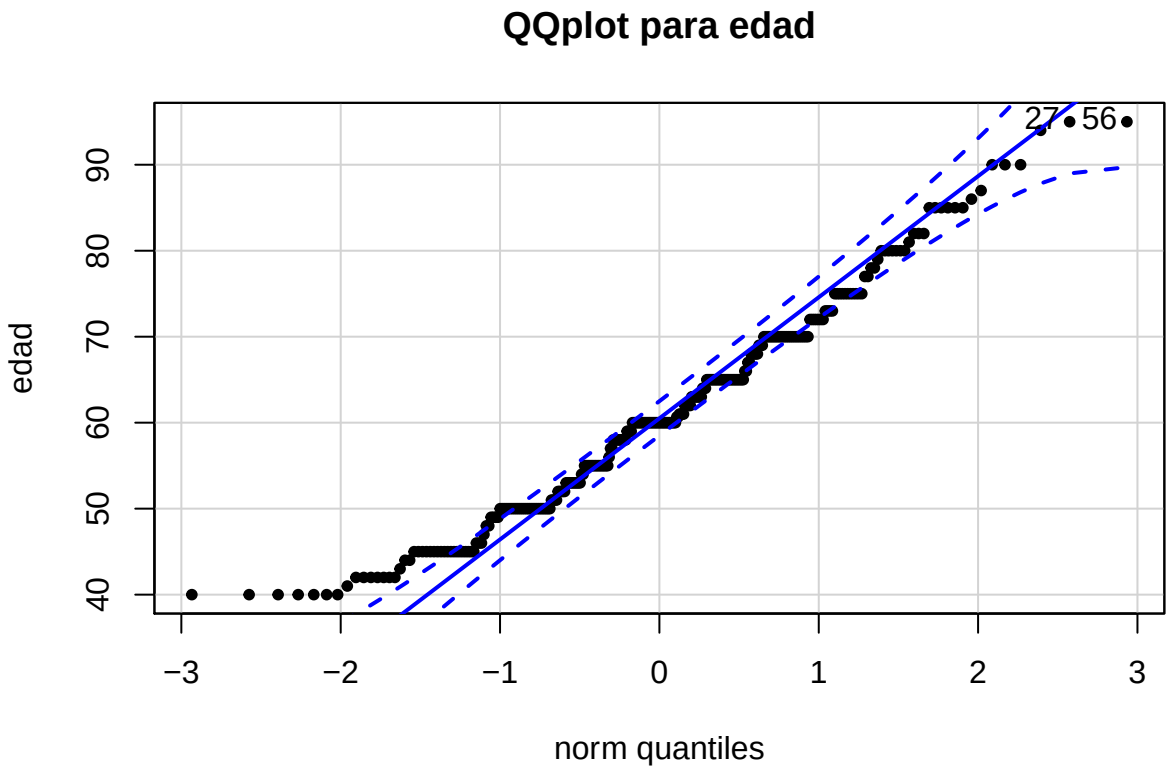


Se observa una distribución asimétrica y una gran cantidad de pacientes entre los 50 y 70 años de eda. También, que la edad media de las personas que murieron es mayor que la de las personas que no murieron. Por último, las edades de los que sobrevivieron son menos variables que de los que no.

```
ad.test(heart$age)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$age
## A = 1.6424, p-value = 0.0003171
```

```
qqPlot(heart$age, pch=20, ylab='edad',
       main='QQplot para edad')
```




```
## [1] 27 56
```

```
leveneTest(y = heart$age, group = heart$DEATH_EVENT, center = "median")
```

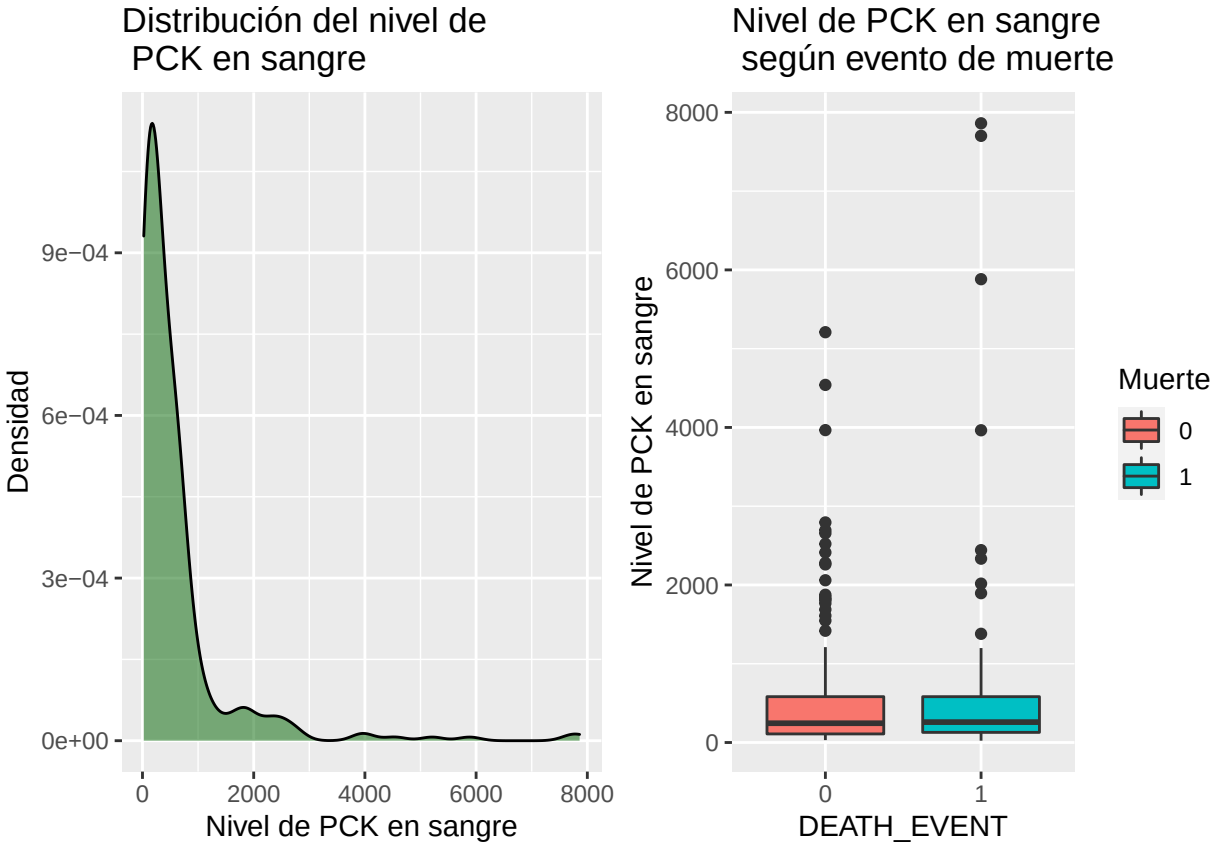
```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value   Pr(>F)
## group   1  7.1338 0.007981 **
##       297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Según el test de Levene para homogeneidad de varianza, existe suficiente evidencia en los datos para decir que las varianzas de las edades no son iguales entre los que murieron y los que no, también, que la variable no sigue una distribución normal. Y como se puede ver en el QQPlot, las observaciones 27 y 56 están mucho más desviadas de la distribución de la variable, y estas observaciones corresponden a personas con 95 años.

Creatina fosfoquinasa

```
p1<- ggplot(heart)+geom_density(aes(creatinine_phosphokinase),
                                fill = "darkgreen", alpha=0.5, adjust=2) +
  labs(x="Nivel de PCK en sangre", y="Densidad")+
guides(fill=guide_legend(title="")) +
  ggtitle("Distribución del nivel de \n PCK en sangre")
```

```
p2<- ggplot(heart, aes(x = DEATH_EVENT, y = creatinine_phosphokinase,
                       fill = DEATH_EVENT ))+geom_boxplot() +
  labs(y="Nivel de PCK en sangre")+
guides(fill=guide_legend(title="Muerte")) +
  ggtitle("Nivel de PCK en sangre \n según evento de muerte")


grid.arrange(p1,p2,ncol = 2)
```

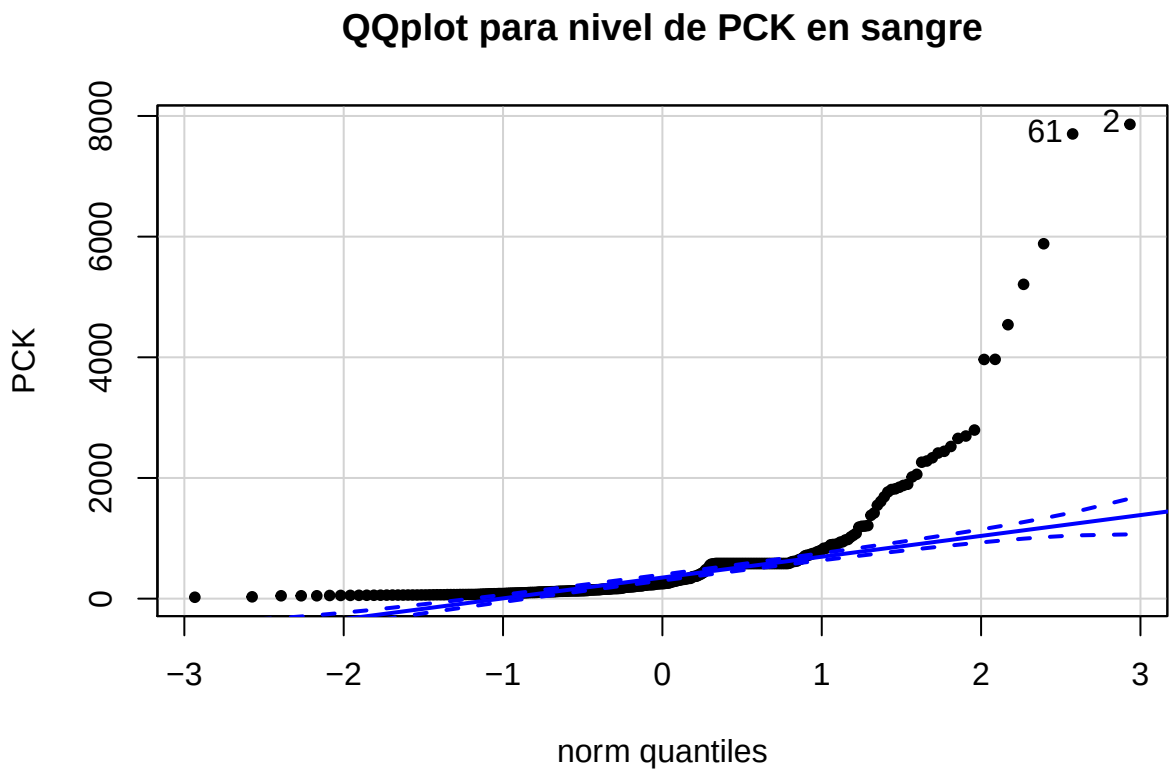


Se observa una distribución asimétrica y al parecer multimodal y una gran cantidad de pacientes con niveles de PCK menores a 2000. También, que estos niveles no son muy diferentes entre los que murieron y los que no. Se destaca la poca presencia de valores bastante altos, por encima de los 5000 mcg/L, estos valores sin duda son extremos, sin embargo, más adelante analizaré si al ser atípicos influyen o no de manera significativa en los análisis posteriores. De momento, estos resultados indican una alta prevalencia de accidentes cerebro-vasculares,infartos, inflamación cardíaca, cirugías previas al corazón, pero también pueden indicar la realización de una actividad física intensa o incluso consumo de drogas ilícitas.

```
ad.test(heart$creatinine_phosphokinase)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$creatinine_phosphokinase
## A = 41.906, p-value < 2.2e-16
```

```
qqPlot(heart$creatinine_phosphokinase, pch=20, ylab='PCK',
       main='QQplot para nivel de PCK en sangre')
```




```
## [1]  2 61
```

```
leveneTest(y = heart$creatinine_phosphokinase, group = heart$DEATH_EVENT, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1  1.0303 0.3109
##       297
```

El test de normalidad, indica que existe suficiente evidencia en los datos para decir que no provienen de una distribución normal, pero que sí hay homogeneidad de varianza entre los que murieron y los que no. El QQPlot, indica que las observaciones 61 y 2 esstán especialmente más alejadas de la distribución de la variable, y efectivamente pertenecen a personas cuyos niveles exceden los 7000 mcg/L.

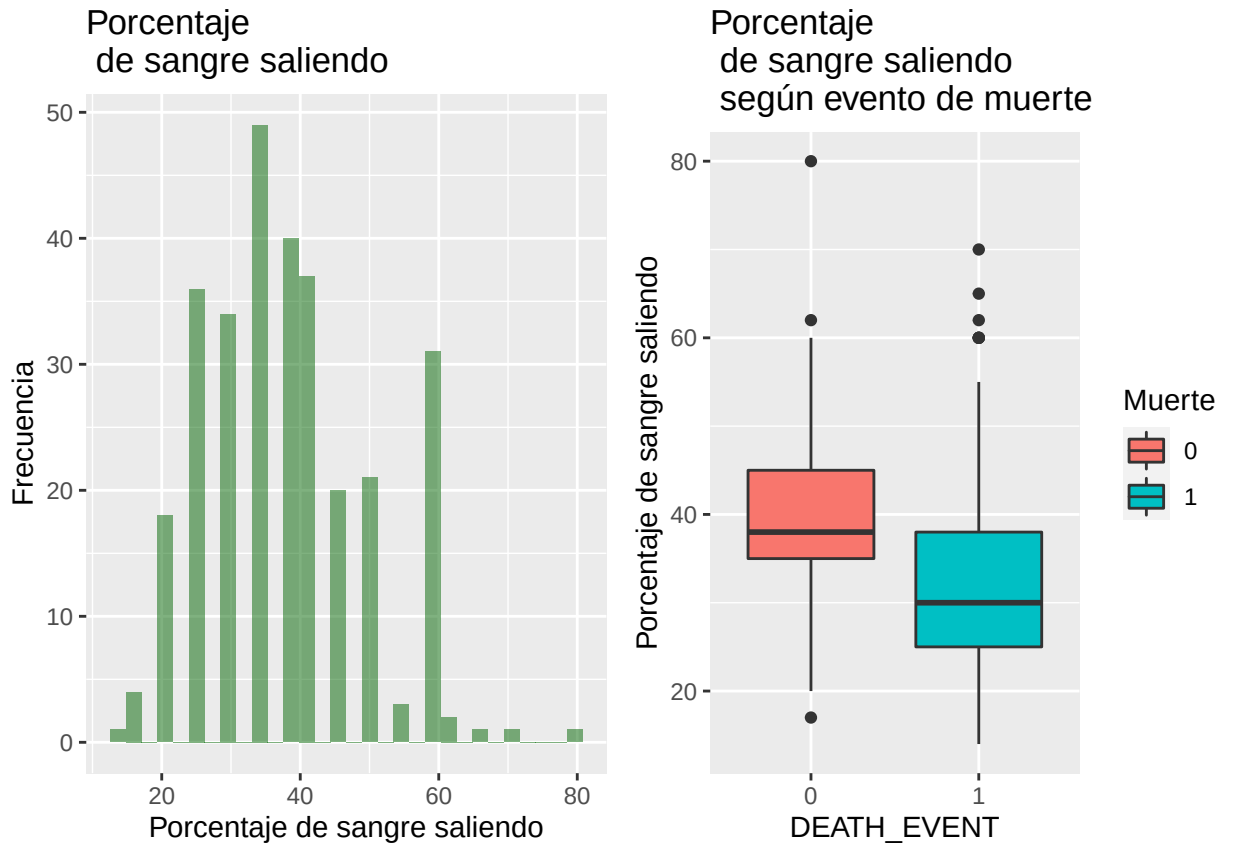
Fracción de eyección

```
p1<- ggplot(heart)+geom_histogram(aes(ejection_fraction),
                                  fill = "darkgreen", alpha=0.5, adjust=2) +
  labs(x="Porcentaje de sangre saliendo", y="Frecuencia")+
  guides(fill=guide_legend(title="")) +
ggtitle("Porcentaje \n de sangre saliendo")
```

```
## Warning: Ignoring unknown parameters: adjust
```

```
p2<- ggplot(heart, aes(x = DEATH_EVENT, y = ejection_fraction,
                       fill = DEATH_EVENT ))+geom_boxplot() +
  labs(y="Porcentaje de sangre saliendo")+
  guides(fill=guide_legend(title="Muerte"))+
```

```
  ggtitle("Porcentaje \n de sangre saliendo \n según evento de muerte")

grid.arrange(p1,p2,ncol = 2)
```

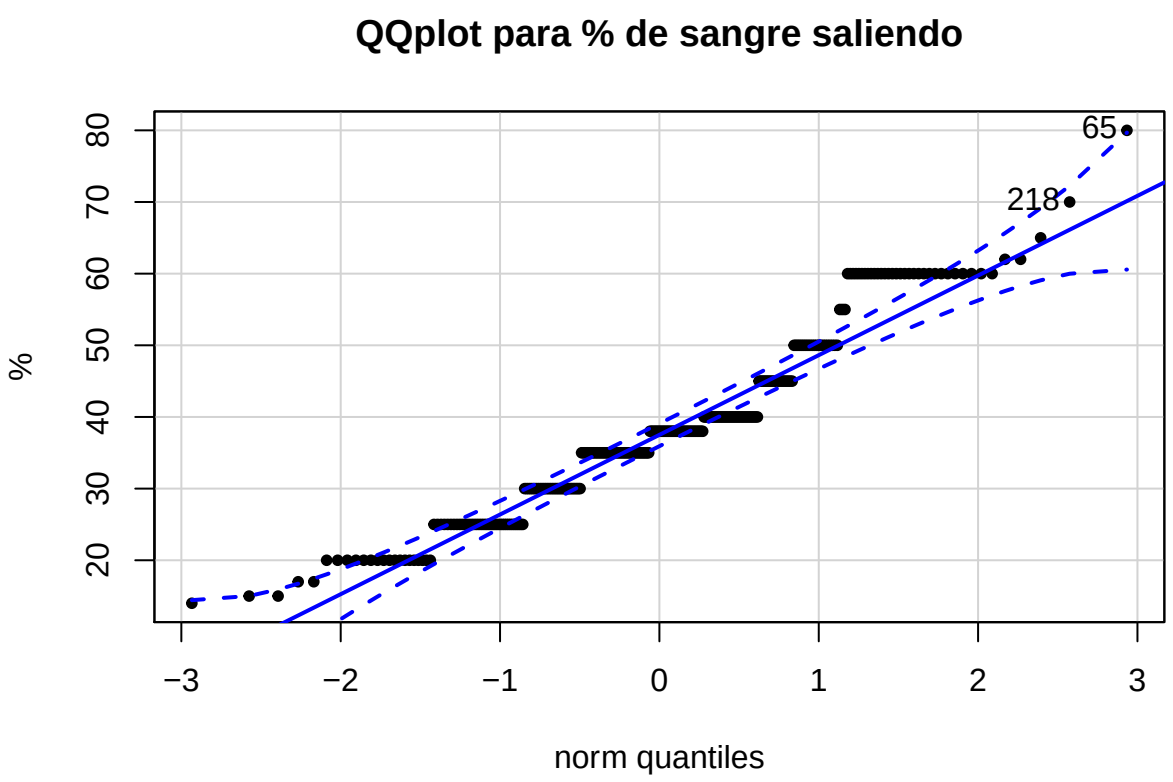


Se observa que la fracción de eyección es mayor en las personas que sobrevivieron que en las que no. Una fracción de eyección en personas sanas, en su mayoría va de 50-70%. Sin embargo, si el músculo cardíaco es grueso y rígido,el ventrículo contiene un volumen de sangre más bajo que lo normal,por lo que la cantidad total de sangre bombeada no sería la suficiente para satisfacer las necesidades del cuerpo, por esta razón, esta fracción se ve menor en los fallecidos.

```
ad.test(heart$ejection_fraction)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$ejection_fraction
## A = 5.802, p-value = 2.59e-14
```

```
qqPlot(heart$ejection_fraction, pch=20, ylab='%',
       main='QQplot para % de sangre saliendo')
```




```
## [1]  65 218
```

```
leveneTest(y = heart$ejection_fraction, group = heart$DEATH_EVENT, center = "median")
```

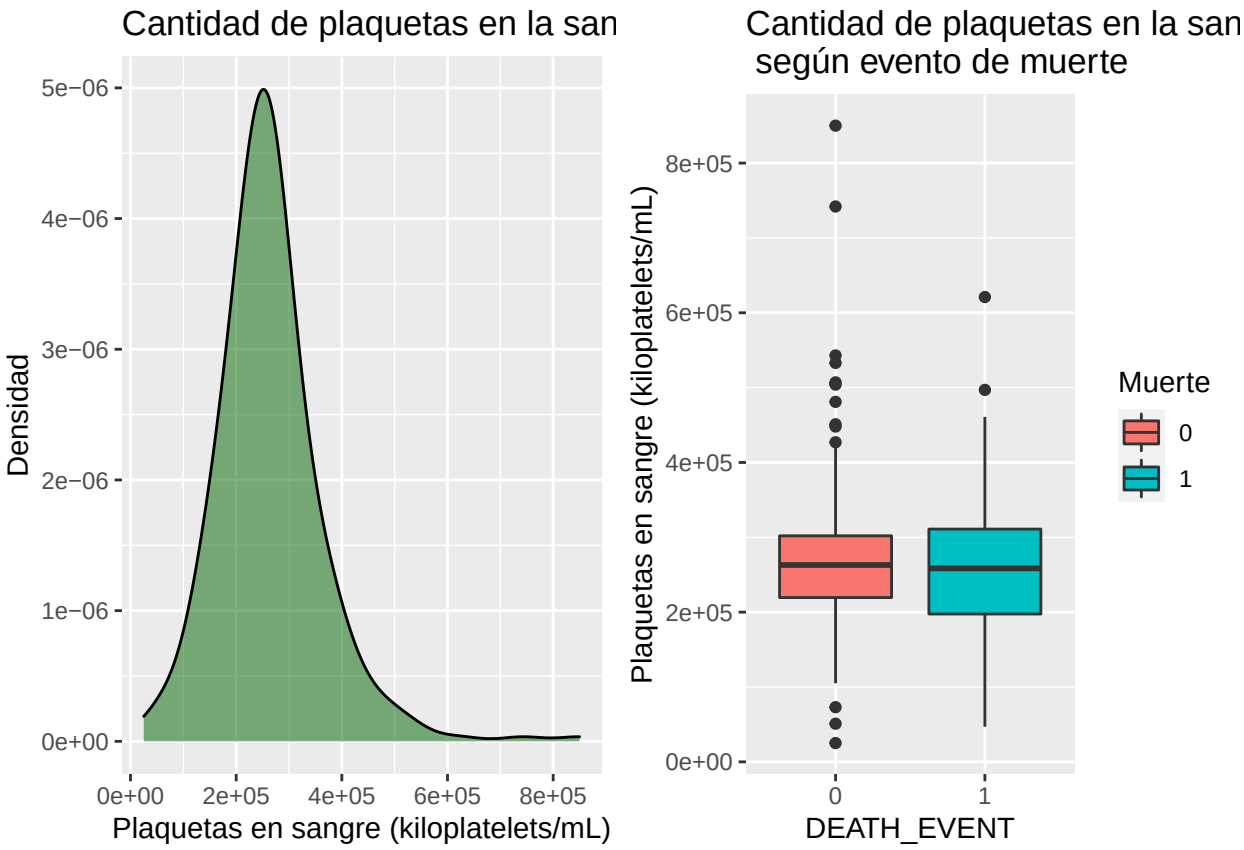
```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1  3.7021 0.0553 .
##       297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El test de normalidad indica que existe suficiente evidencia en los datos para decir que los datos no provienen de una distribución normal, sin embargo, si existe evidencia (a un nivel de significancia del 5%) para decir que las varianzas son iguales entre los que murieron y los que no. Las observaciones 65 y 218 corresponden a fracciones de eyección entre 70 y 80%, lo que puede indicar en estos pacientes una enfermedad cardíaca.

Plaquetas

```
p1<-ggplot(heart)+geom_density(aes(platelets),
                               fill = "darkgreen", alpha=0.5, adjust=2)+
labs(x="Plaquetas en sangre (kiloplatelets/mL)", y="Densidad")+
guides(fill=guide_legend(title="")) +
  ggtitle("Cantidad de plaquetas en la sangre")

p2<-ggplot(heart, aes(x = DEATH_EVENT, y = platelets,
                      fill = DEATH_EVENT ))+geom_boxplot() +
labs(y="Plaquetas en sangre (kiloplatelets/mL)")+
```

```
guides(fill=guide_legend(title="Muerte")) +
  ggtitle("Cantidad de plaquetas en la sangre \n según evento de muerte")


grid.arrange(p1,p2,ncol = 2)
```

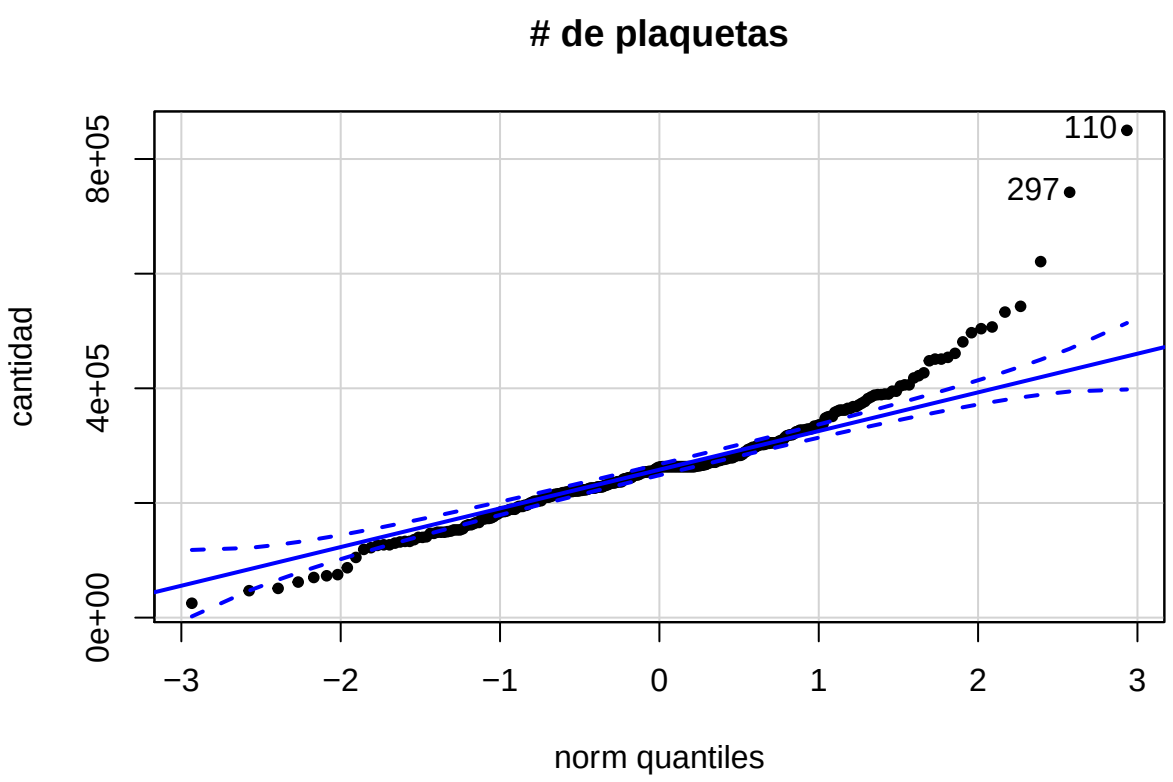


Se observa que la cantidad de plaquetas en la sangre es similar entre las personas que murieron y las que no. Existen valores superiores a los 600000 kiloplatelets/mL pero son muy pocos.

```
ad.test(heart$platelets)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$platelets
## A = 4.989, p-value = 2.306e-12
```

```
qqPlot(heart$platelets, pch=20, ylab='cantidad',
       main='# de plaquetas')
```




```
## [1] 110 297
```

```
leveneTest(y = heart$platelets, group = heart$DEATH_EVENT, center = "median")
```

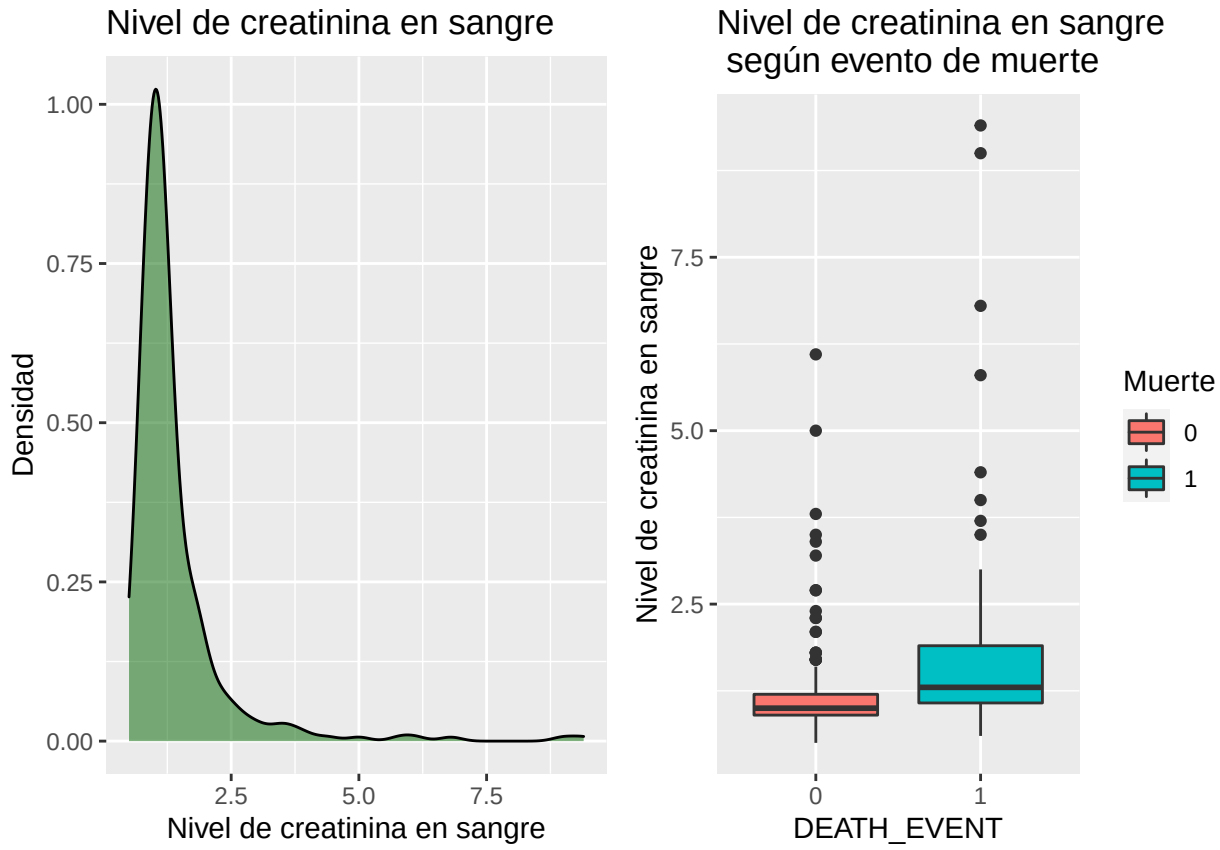
```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1   1.085 0.2984
##       297
```

Existe evidencia suficiente en los datos para decir que no provienen de una distribución normal, sin embargo hay evidencia de homogeneidad de varianza entre los que murieron y los que no. Las observaciones 297 y 110 que corresponden a una cantidad de plaquetas superior a los 740000 son las que están más alejadas de la distribución de la variable, sin embargo, eliminarlas no afecta significativamente dicha distribución.

Creatinina en sangre

```
p1<-ggplot(heart)+geom_density(aes(serum_creatinine),
                               fill = "darkgreen", alpha=0.5, adjust=2) +
labs(x="Nivel de creatinina en sangre", y="Densidad")+
guides(fill=guide_legend(title="")) +
  ggtitle("Nivel de creatinina en sangre")

p2<-ggplot(heart, aes(x = DEATH_EVENT, y = serum_creatinine,
                      fill = DEATH_EVENT ))+geom_boxplot() +
labs(y="Nivel de creatinina en sangre")+
  guides(fill=guide_legend(title="Muerte"))+
  ggtitle("Nivel de creatinina en sangre \n según evento de muerte")
```

```
grid.arrange(p1,p2,ncol = 2)
```

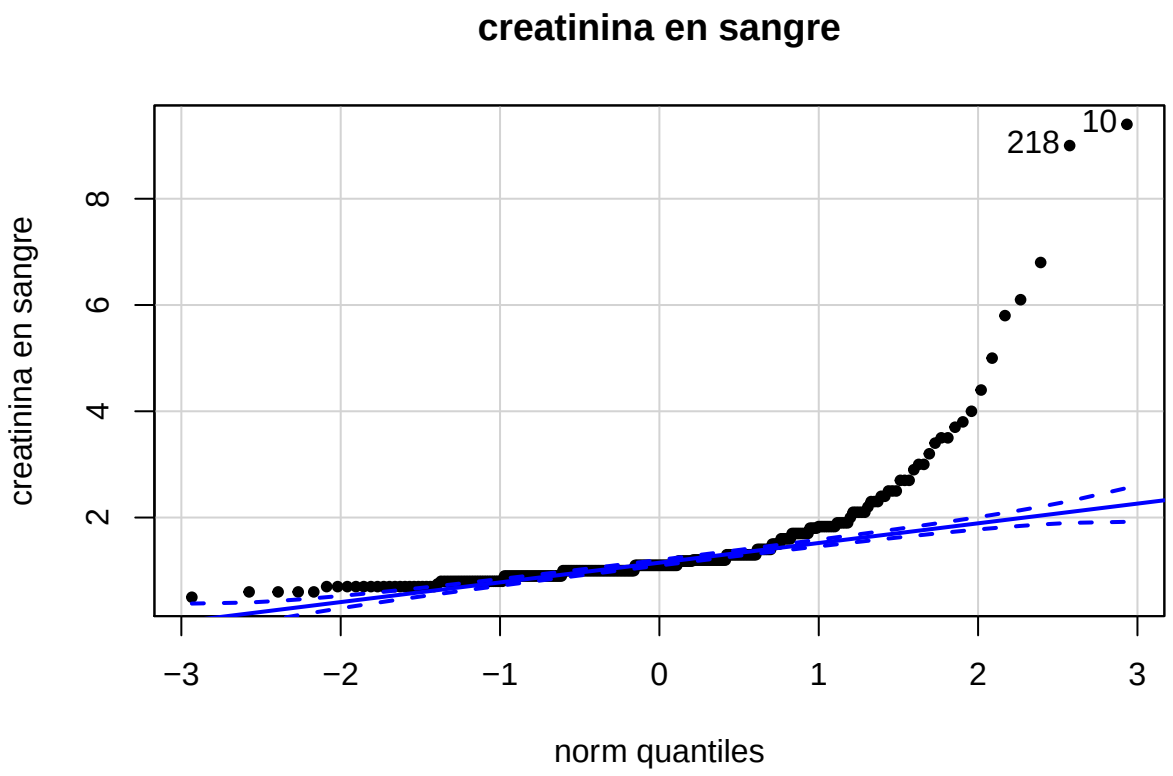


Se observa los niveles de creatinina en la sangre son mucho menores en las personas que no murieron. La distribución de la variable es asimétrica y con presencia de pocos valores superiores a 5 mg/dL. Esto indicaría la presencia de enfermedades renales tanto en los pacientes que sobrevivieron como en los que no, sin embargo, los que murieron tenían mucha más probabilidad de desarrollar insuficiencia renal debido a sus altísimos niveles de creatinina.

```
ad.test(heart$serum_creatinine)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$serum_creatinine
## A = 36.451, p-value < 2.2e-16
```

```
qqPlot(heart$serum_creatinine, pch=20, ylab='creatinina en sangre',
       main='creatinina en sangre')
```




```
## [1]  10 218
```

```
leveneTest(y = heart$serum_creatinine, group = heart$DEATH_EVENT, center = "median")
```

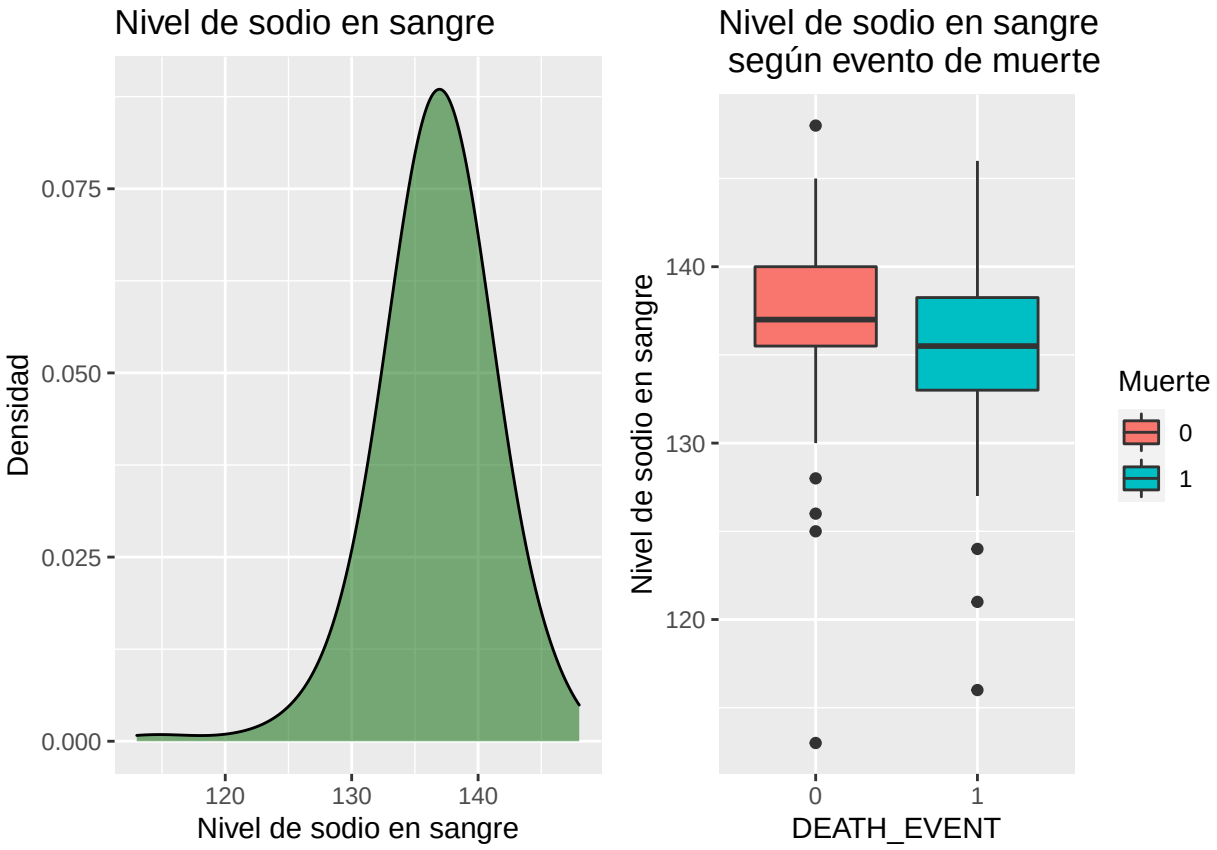
```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value    Pr(>F)
## group   1  16.242 7.087e-05 ***
##       297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No existe evidencia suficiente en los datos que indique la normalidad y la homogeneidad de varianza.

Sodio en sangre

```
p1<-ggplot(heart)+geom_density(aes(serum_sodium),
                               fill = "darkgreen", alpha=0.5, adjust=2) +
labs(x="Nivel de sodio en sangre", y="Densidad")+
guides(fill=guide_legend(title="")) +
  ggtitle("Nivel de sodio en sangre")

p2<-ggplot(heart, aes(x = DEATH_EVENT, y = serum_sodium,
                      fill = DEATH_EVENT ))+geom_boxplot() +
labs(y="Nivel de sodio en sangre")+
guides(fill=guide_legend(title="Muerte")) +
  ggtitle("Nivel de sodio en sangre \n según evento de muerte")
```

```
grid.arrange(p1,p2,ncol = 2)
```

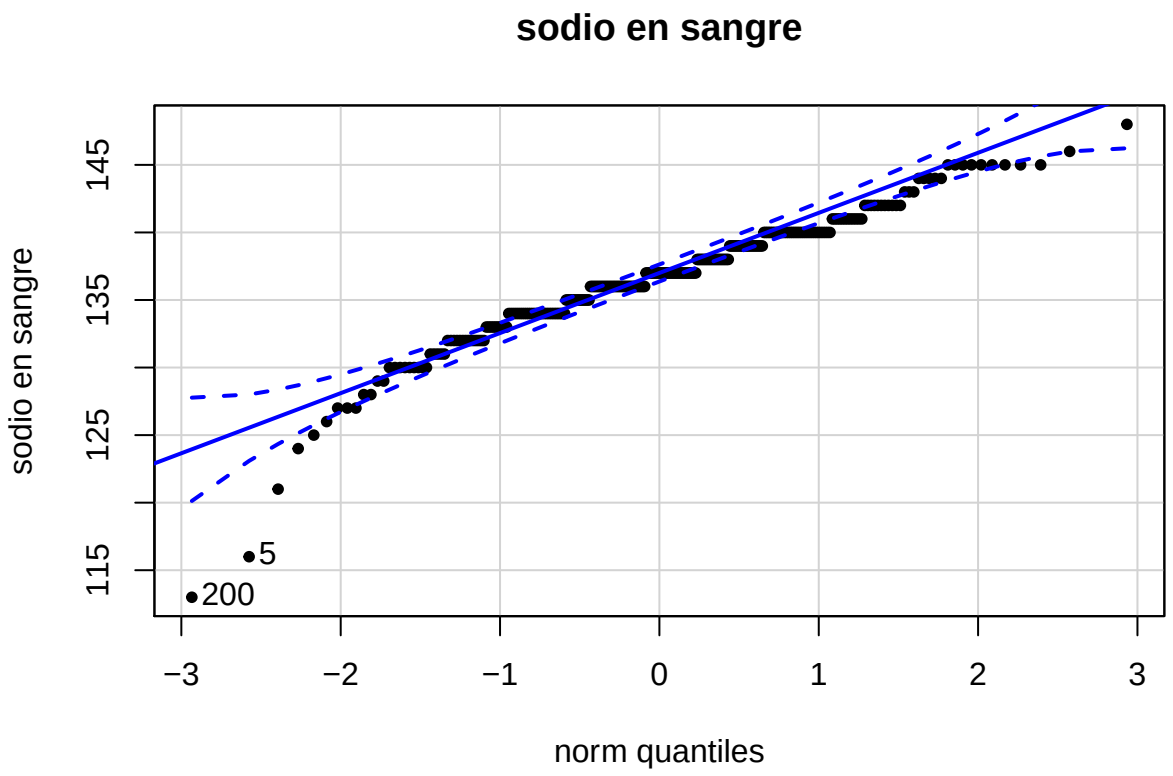


Se observa que los niveles de sodio en la sangre son altos pero menores en las personas que murieron. La distribución de la variable es asimétrica y con presencia de pocos valores inferiores a 125. La mayoría de las personas tenían niveles normales de sodio. Los que tenían niveles bajos, probablemente presentaban altos niveles de azúcar en sangre, o acumulación en la orina de productos de desecho de la descomposición de la grasa, o problemas hormonales.

```
ad.test(heart$serum_sodium)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$serum_sodium
## A = 3.0938, p-value = 8.902e-08
```

```
qqPlot(heart$serum_sodium, pch=20, ylab='sodio en sangre',
       main='sodio en sangre')
```




```
## [1] 200   5
```

```
leveneTest(y = heart$serum_sodium, group = heart$DEATH_EVENT, center = "median")
```

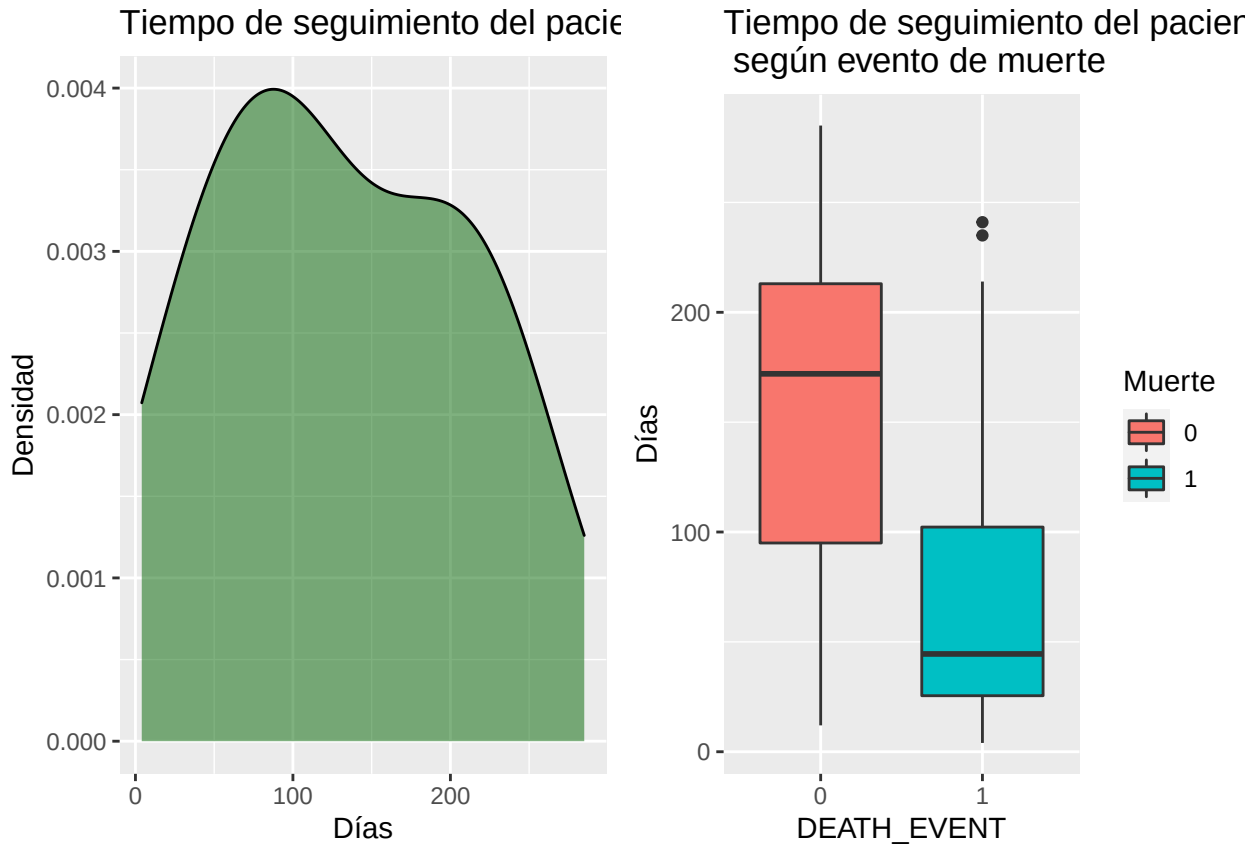
```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   1   5.274 0.02234 *
##       297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No existe evidencia suficiente en los datos para decir que los datos siguen una distribución normal, sin embargo, a un nivel de significancia del 5% los datos no muestran varianzas homogéneas entre los que murieron y los que no. Ahora, a un nivel de significancia del 1% si se evidencian varianzas homogéneas. Desde mi experiencia, considero que si existen varianzas homogéneas.

Tiempo

```
p1<-ggplot(heart)+geom_density(aes(time),
                               fill = "darkgreen", alpha=0.5, adjust=2)+
labs(x="Días", y="Densidad")+
  guides(fill=guide_legend(title="")) +
  ggtitle("Tiempo de seguimiento del paciente")

p2<- ggplot(heart, aes(x = DEATH_EVENT, y = time,
                       fill = DEATH_EVENT ))+geom_boxplot()+
labs(y="Días")+
```

```
  guides(fill=guide_legend(title="Muerte")) +
ggtitle("Tiempo de seguimiento del paciente \n según evento de muerte")



grid.arrange(p1,p2,ncol = 2)
```

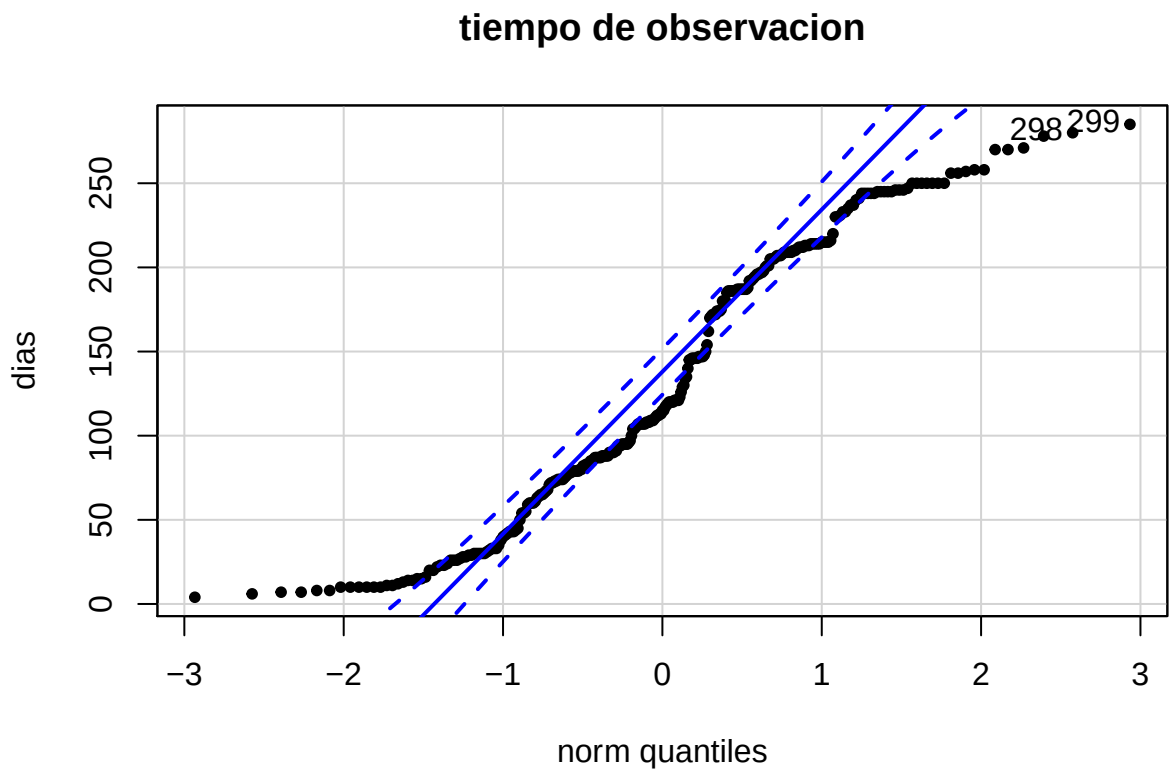


La cantidad de tiempo en la que se monitoreó a cada paciente, fue mayor en los que sobrevivieron que en los que no.

```
ad.test(heart$time)
```

```
##
##  Anderson-Darling normality test
##
## data:  heart$time
## A = 4.9702, p-value = 2.559e-12
```

```
qqPlot(heart$time, pch=20, ylab='dias',
       main='tiempo de observacion')
```




```
## [1] 299 298
```

```
leveneTest(y = heart$time, group = heart$DEATH_EVENT, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value   Pr(>F)
## group   1  7.9512 0.005129 **
##       297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No existe evidencia en los datos para decir que provienen de una distribución normal con varianzas iguales.

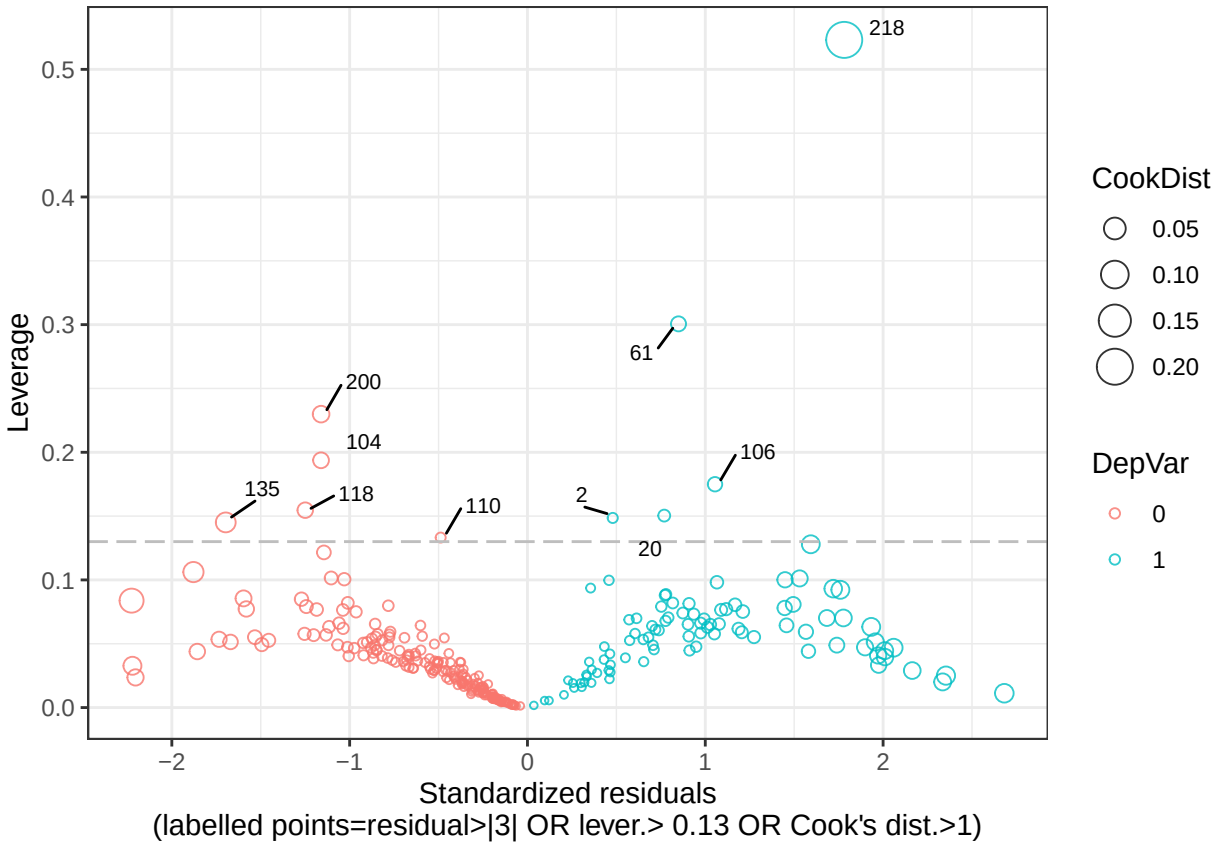
Identificación y tratamiento de valores atípicos

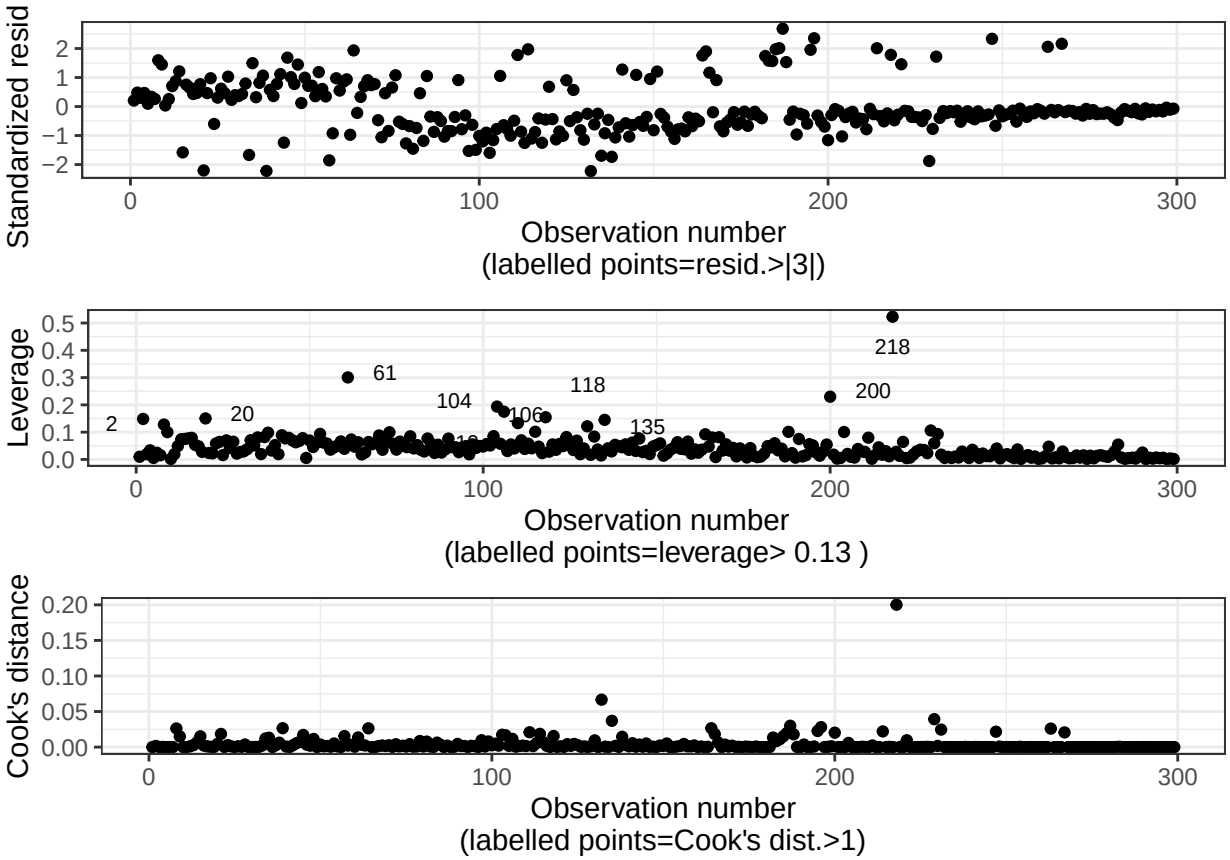
```
influyentes <- function(data){

  library('ggplot2')
  library('ggrepel')
  library('gridExtra')

  options(scipen=999)
  group <-names(data[1])
  independent <- names(data)[names(data) != group]
  variables <- paste(independent, collapse="+")
  frm <<- as.formula(paste(group, "~", variables, sep = ""))
```

```
  fit <<- glm(frm, data=data, family="binomial")
  print(summary(fit))

  df.diagn <- data.frame(points=as.numeric(rownames(data)), y=fit$y,
                         pred.prob=fit$fitted.values, res=rstandard(fit),
                         CookDist=cooks.distance(fit), DepVar=as.factor(fit$y),
                         leverage=hatvalues(fit))

  n.of.predictors <- sum(hatvalues(fit))-1
  lev.thresh <- round(3*((n.of.predictors+1)/nrow(data)),3)
  df.diagn$lever.check <- ifelse(df.diagn$leverage>lev.thresh,
                                 "lever. not ok","lever. ok")

  obs_per_factor <- plyr::count(df.diagn$DepVar)
  U <- wilcox.test(df.diagn$pred.prob ~ df.diagn$DepVar)$statistic
  auc <- round(1-U/(obs_per_factor$freq[1]*obs_per_factor$freq[2]), 3)

  p1 <- ggplot(df.diagn, aes(x=res, y=leverage, color=DepVar)) +
    geom_point(aes(size = CookDist), shape=1, alpha=.80) +
    geom_hline(yintercept = lev.thresh, colour="grey", linetype = "longdash") +
    theme_bw() +
    geom_text_repel(data = subset(df.diagn, abs(res) > 3 |
                                    leverage > lev.thresh | CookDist > 1),
aes(label = points), size = 2.7, colour="black",
box.padding = unit(0.35, "lines"), point.padding = unit(0.3, "lines")) +
labs(x = paste("Standardized residuals\n(labelled points=residual>|3| OR lever.>", lev.thresh,"OR Cook's

  p2 <- ggplot(df.diagn, aes(x=points, y=res, label=points)) +
    geom_point() + theme_bw() +
    geom_text_repel(data = subset(df.diagn, abs(res) > 3),
                    aes(label = points), size = 2.7, colour="black",
                    box.padding = unit(0.35, "lines"), point.padding = unit(0.3, "lines")) +
    labs(x = "Observation number\n(labelled points=resid.>|3|)", y="Standardized residuals")

  p3 <- ggplot(df.diagn, aes(x=points, y=leverage, label=points)) +
    geom_point() + theme_bw() +
    geom_text_repel(data = subset(df.diagn, leverage > lev.thresh),
                    aes(label = points), size = 2.7, colour="black",
                    box.padding = unit(0.35, "lines"), point.padding = unit(0.3, "lines")) +
    labs(x = paste("Observation number\n(labelled points=leverage>",lev.thresh,")"), y="Leverage")

  p4 <- ggplot(df.diagn, aes(x=points, y=CookDist, label=points)) +
    geom_point() + theme_bw() +
    geom_text_repel(data = subset(df.diagn, CookDist > 1),
                    aes(label = points), size = 2.7, colour="black",
                    box.padding = unit(0.35, "lines"), point.padding = unit(0.3, "lines")) +
    labs(x = "Observation number\n(labelled points=Cook's dist.>1)", y="Cook's distance")

  print(p1)
  grid.arrange(p2, p3, p4, ncol=1)


}
```

```r
heart1<- data.frame(group = heart$DEATH_EVENT, heart[,-13])
influyentes(data = heart1)
```

```
##
## Call:
## glm(formula = frm, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1848  -0.5706  -0.2401   0.4466   2.6668
##
## Coefficients:
##                              Estimate   Std. Error z value
## (Intercept)              10.184929023  5.656570325   1.801
## age                       0.047419074  0.015800632   3.001
## anaemia1                 -0.007470452  0.360489086  -0.021
## creatinine_phosphokinase  0.000222229  0.000177933   1.249
## diabetes1                 0.145149775  0.351188640   0.413
## ejection_fraction        -0.076662501  0.016329130  -4.695
## high_blood_pressure1     -0.102679427  0.358706893  -0.286
## platelets                -0.000001200  0.000001889  -0.635
## serum_creatinine          0.666093340  0.181492576   3.670
## serum_sodium             -0.066981072  0.039735098  -1.686
## sex1                     -0.533658016  0.413918039  -1.289
## smoking1                 -0.013492224  0.412617798  -0.033
## time                     -0.021044626  0.003014394  -6.981
##                                 Pr(>|z|)
## (Intercept)                     0.071774 .
## age                             0.002690 **
## anaemia1                        0.983467
## creatinine_phosphokinase        0.211684
## diabetes1                       0.679380
## ejection_fraction        0.00000266827448 ***
## high_blood_pressure1            0.774688
## platelets                       0.525404
## serum_creatinine                0.000242 ***
## serum_sodium                    0.091855 .
## sex1                            0.197299
## smoking1                        0.973915
## time                     0.00000000000292 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 219.55  on 286  degrees of freedom
## AIC: 245.55
##
## Number of Fisher Scoring iterations: 6
```

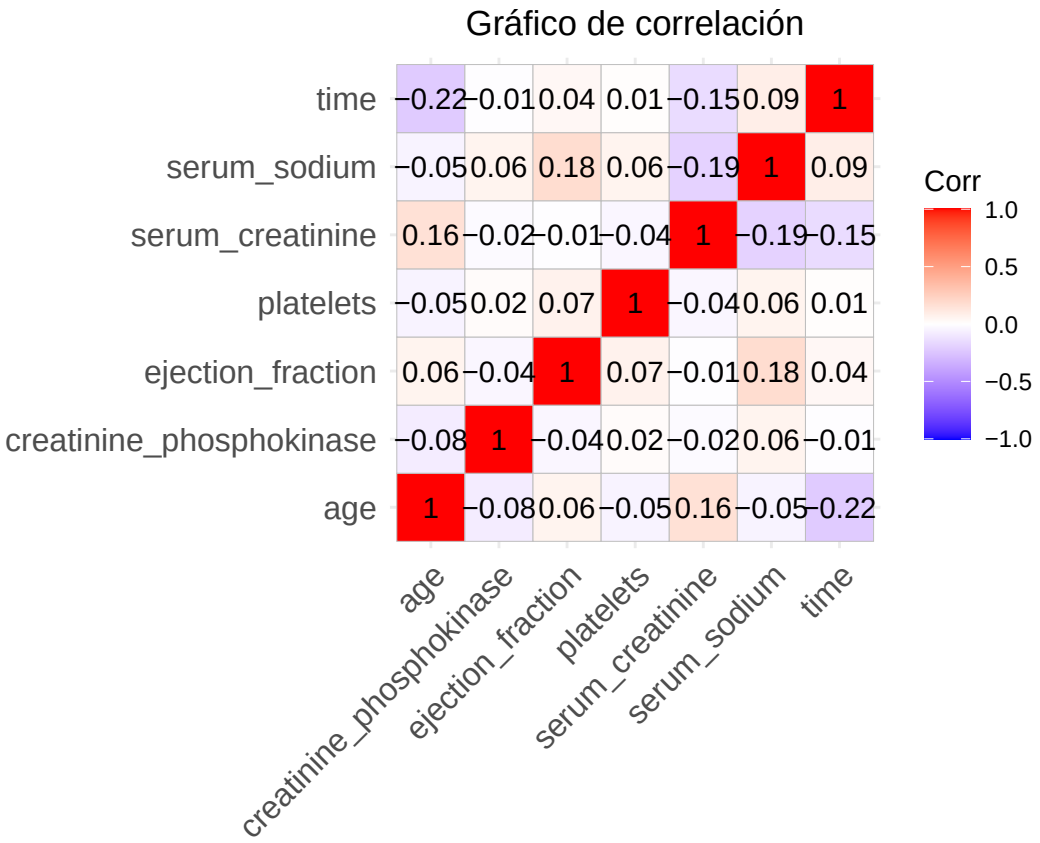

218
61
200
104
106
135
118
110
2
20
0.5
0.4
0.3
0.2
0.1
0.0
Leverage
−2
−1
0
1
2
Standardized residuals
(labelled points=residual>|3| OR lever.> 0.13 OR Cook's dist.>1)
CookDist
0.05
0.10
0.15
0.20
DepVar
0
1



Según los gráficos de leverage y residuos estudentizados, existen 9 valores extremos en todo el dataset, sin embargo, la distancia de Cook indica que estos valores no son influyentes, y según lo que se vió la presencia de los distintos valores extremos en las variables no modifican su distribución significativamente. Por lo tanto, no los eliminaré. Ahora, con respecto al modelo logístico ajustado, se puede ver a priori que las variables edad, fracción de eyección, tiempo y creatinina en suero son significativamente estadísticas y pueden ayudar mucho a la clasificación

Correlación

```
corr <- cor(heart_c[,-8])

ggcorrplot(corr, lab = TRUE) +theme(legend.position="right") +
  ggtitle('Gráfico de correlación') +
  theme(plot.title = element_text(hjust = 0.5))
```



No hay correlaciones muy fuertes entre las variables, esto anticipa que no es muy probable que existan problemas de multicolinealidad en los modelos que se vayan a ajustar.

Diferencia de medias

Dado que las variables que se manejan aquí no siguen una distribución normal, realizaré una prueba no paramétrica para comparar las medias entre los que murieron y los que no:

```
library(tidyverse)


sds<-heart_c%>%group_by(DEATH_EVENT)%>%summarise_all(list(sd), na.rm=TRUE)%>%
  gather("Variable", "Sd", -DEATH_EVENT)%>%
  spread(DEATH_EVENT, Sd)%>%rename("Sd_0"='0', "Sd_1"="1")

medians<-heart_c%>%group_by(DEATH_EVENT)%>%summarise_all(list(median), na.rm=TRUE )%>%
  gather("Variable", "Median", -DEATH_EVENT)%>%
  spread(DEATH_EVENT, Median)%>%rename("Median_0"='0', "Median_1"="1")

summary_report<-medians%>%
  inner_join(sds, "Variable")%>%
  mutate(DiffInMedians=Median_1-Median_0)

variables<-colnames(heart_c)[-8]
pvals<-{}
vars<-{}
```

```
for (i in variables) {

  xxx<-heart_c%>%select(c("DEATH_EVENT", i))

  x1<-xxx%>%filter(DEATH_EVENT=="0")%>%dplyr::select(c(2))%>%na.omit()%>%pull()
  x2<-xxx%>%filter(DEATH_EVENT=="1")%>%dplyr::select(c(2))%>%na.omit()%>%pull()
  wc<-wilcox.test(x1,x2)


  pvals<-c(pvals,round(wc$p.value,4) )
  vars<-c(vars,i)

}
wc_df<-data.frame(Variable=vars,pvalues=pvals)
wc_df$Variable<-as.character(wc_df$Variable)
summary_report<-summary_report%>%inner_join(wc_df, by="Variable")

library(kableExtra)
summary_report %>% kable() %>% kable_styling()
```

Variable	Median_0	Median_1	Sd_0	Sd_1	DiffInMedians	pvalues
age	60	65.0	10.6378902	13.214556	5.0	0.0002
creatinine_phosphokinase	245	259.0	753.7995716	1316.580640	14.0	0.6840
ejection_fraction	38	30.0	10.8599627	12.525303	-8.0	0.0000
platelets	263000	258500.0	97531.2022835	98525.682857	-4500.0	0.4256
serum_creatinine	1	1.3	0.6540827	1.468562	0.3	0.0000
serum_sodium	137	135.5	3.9829234	5.001579	-1.5	0.0003
time	172	44.5	67.7428724	62.378281	-127.5	0.0000

Según la tabla, la fracción de eyección, la creatinina en suero, la creatinina en sodio, el tiempo de seguimiento y la edad provienen de distribuciones con medias estadística y significativamente diferentes, ya que los test sugieren que los datos ofrecen evidencia suficiente para rechazar la hipótesis nula de igualdad de media.

Análisis exploratorio de variables cualitativas

Analizaré la cantidad de personas fallecidas y no fallecidas dentro de cada categoría de las variables categóricas:

```
summary_categories <- data.frame(No_muerte = length(which(heart$anaemia == 0 &
                                                    heart$DEATH_EVENT == 0)),
                                 Muerte = length(which(heart$anaemia == 0 &
                                                    heart$DEATH_EVENT == 1)))

summary_categories <- rbind(summary_categories, c(length(which(heart$anaemia == 1 &
                                                    heart$DEATH_EVENT == 0)),
                                                  length(which(heart$anaemia == 1 &
                                                    heart$DEATH_EVENT ==1))))

summary_categories <- rbind(summary_categories, c(length(which(heart$diabetes == 0 &
                                                    heart$DEATH_EVENT == 0)),
                                                  length(which(heart$diabetes == 0 &
                                                    heart$DEATH_EVENT == 1))))
```

```
summary_categories <- rbind(summary_categories, c(length(which(heart$diabetes == 1 &
                                                                heart$DEATH_EVENT == 0)),
                                                   length(which(heart$diabetes == 1 &
                                                                heart$DEATH_EVENT == 1))))

summary_categories <- rbind(summary_categories,
                            c(length(which(heart$high_blood_pressure == 0 &
                                           heart$DEATH_EVENT == 0)),
                              length(which(heart$high_blood_pressure == 0 &
                                           heart$DEATH_EVENT == 1))))

summary_categories <- rbind(summary_categories,
                            c(length(which(heart$high_blood_pressure == 1 &
                                           heart$DEATH_EVENT == 0)),
                              length(which(heart$high_blood_pressure == 1 &
                                           heart$DEATH_EVENT == 1))))

summary_categories <- rbind(summary_categories, c(length(which(heart$sex == 0 &
                                                                heart$DEATH_EVENT == 0)),
                                                   length(which(heart$sex == 0 &
                                                                heart$DEATH_EVENT == 1))))

summary_categories <- rbind(summary_categories, c(length(which(heart$sex == 1 &
                                                                heart$DEATH_EVENT == 0)),
                                                   length(which(heart$sex == 1 &
                                                                heart$DEATH_EVENT == 1))))

summary_categories <- rbind(summary_categories, c(length(which(heart$smoking == 0 &
                                                                heart$DEATH_EVENT == 0)),
                                                   length(which(heart$smoking == 0 &
                                                                heart$DEATH_EVENT == 1))))

summary_categories <- rbind(summary_categories, c(length(which(heart$smoking == 1 &
                                                                heart$DEATH_EVENT == 0)),
                                                   length(which(heart$smoking == 1 &
                                                                heart$DEATH_EVENT == 1))))


summary_categories <- summary_categories %>%
  mutate(muestra_total = summary_categories$Muerte + summary_categories$No_muerte)

summary_categories<- summary_categories %>%
  mutate(porcentaje_no_muerte = round((summary_categories$No_muerte/summary_categories$muestra_total)*10
  mutate(porcentaje_muerte = round((summary_categories$Muerte/summary_categories$muestra_total)*100, 4))

rownames(summary_categories)<- c("no anemia","anemia", "no diabetes",
                                 "diabetes", "no hipertension", "hipertension",
                                 "mujer", "hombre", "no fuma", "fuma")


summary_categories %>% kable()
```

	No_muerte	Muerte	muestra_total	porcentaje_no_muerte	porcentaje_muerte
no anemia	120	50	170	70.5882	29.4118
anemia	83	46	129	64.3411	35.6589
no diabetes	118	56	174	67.8161	32.1839
diabetes	85	40	125	68.0000	32.0000
no hipertension	137	57	194	70.6186	29.3814
hipertension	66	39	105	62.8571	37.1429
mujer	71	34	105	67.6190	32.3810
hombre	132	62	194	68.0412	31.9588
no fuma	137	66	203	67.4877	32.5123
fuma	66	30	96	68.7500	31.2500

La tabla muestra que solo el 35% de los pacientes que tenían anemia murieron, y el 70.6% de los que no tenían anemia se salvaron. También, que no existe mucha diferencia entre el porcentaje de personas que murieron teniendo o no diabetes. El 37% de los que tenían hipertensión fallecieron, y el 31% de los que fumaban también murieron. No existe mucha diferencia entre el porcentaje de hombres y mujeres que murieron. Sin embargo, el 86.4% de las personas que murieron tenían diabetes, o hipertensión o fumaban, pero la consideración de estos factores como factores potenciales de riesgo para muerte por insuficiencia cardíaca, en especial la diabetes y la hipertensión, estaría más asociada con otras características que muestran la afectación al funcionamiento normal del corazón.

Lo anterior lo comprobamos realizando un test de asociación Chi-cuadrado para saber si la ditribución de las muertes está asociada con la presencia de alguna de las variables anteriores:

Los resultados, muestran que efectivamente existe una evidencia en los datos para decir que no existe una asociación fuerte entre la presencia de diabetes, hipertensión, fumar o el sexo y la muerte o no del paciente.

Modelación

Dado que el dataset presenta imbalance importante en los grupos de fallecidos (32.11%) y sobrevivientes (67.89%), no ajustaré una regresión logística para saber la probabilidad de fallecer o no según estas variables, sino que ajustaré una serie de modelos de ML para clasificar las observaciones, conocer el poder de clasificación que juntas tienen estas variables y ver cuáles son las más importantes.

```
df <- heart_c[,-8] %>%  scale() %>% as.data.frame()
df<- data.frame(Y = heart$DEATH_EVENT, df,sex = heart$sex,
                smoking = heart$smoking, diabetes = heart$diabetes,
                high_blood_pressure = heart$high_blood_pressure,
                anaemia = heart$anaemia)

library(caret)
library(caretEnsemble)
```

Previamente, estandaricé las variables cuantitativas y seleccioné los datos de entrenamiento (70%) y de testeo (30%) mediante la función createDataPartition del paquete Caret.

```
training<- read.csv("C:/Users/mvdiaz/Downloads/training_dataset.csv")
training$Y<-as.factor(training$Y)
training$sex<-as.factor(training$sex)
training$smoking<-as.factor(training$smoking)
training$diabetes<-as.factor(training$diabetes)
training$high_blood_pressure<-as.factor(training$high_blood_pressure)
training$anaemia<-as.factor(training$anaemia)


testing <- read.csv("C:/Users/mvdiaz/Downloads/testing_dataset.csv")
```

```
testing$Y<-as.factor(testing$Y)
testing$sex<-as.factor(testing$sex)
testing$smoking<-as.factor(testing$smoking)
testing$diabetes<-as.factor(testing$diabetes)
testing$high_blood_pressure<-as.factor(testing$high_blood_pressure)
testing$anaemia<-as.factor(testing$anaemia)
```

```
control_prmt <- trainControl(method          = "LGOCV",
                             p               = 0.7,
                             number          = 10,
                             savePredictions = "final",
                             verboseIter     = FALSE)

cat(">>> Fitting 5 models: random forest, SVM, knn, FDA, avNNet ...\n")
capture.output(model_list <- caretEnsemble::caretList(
  Y ~ .,
  data       = training,
  trControl  = control_prmt,
  tuneList   = list(ranger = caretModelSpec(method = "ranger", importance = "impurity")),
  methodList = c("svmRadial", "knn", "avNNet") #"bagFDA",
  ))


cat(">>> Calculating correlation and accuracy metrics over resamples ...\n")
# It is expected high accuracy and un-correlation between them
results <- list(model_correlations = modelCor(resamples(model_list)),
                model_accuracies   = summary(resamples(model_list)),
                final_model        = model_list)

cat(">>> Calculating contingency table and metrics for training data ...\n")
training_preds <- predict(model_list, newdata = training) %>% data.frame
training_preds$ensemble <- apply(training_preds, 1, function(x){tt <- table(x);
return(names(tt[which.max(tt)]))}) %>% factor
# results$Training_predictions <- training_preds

results$Training_CM <- training_preds %>%
  purrr::map(function(x) suppressWarnings(caret::confusionMatrix(data = x,
                                                 reference = training$Y)))

results$Training_MCC <- training_preds %>%
  purrr::map(function(x) suppressWarnings(mltools::mcc(preds = x,
                                                       actuals = training$Y))) %>%
  unlist

cat(">>> Calculating contingency table and metrics for testing data ...\n")
testing_preds <- predict(model_list, newdata = testing) %>% data.frame
testing_preds$ensemble <- apply(testing_preds, 1, function(x){tt <- table(x);
return(names(tt[which.max(tt)]))}) %>% factor
# results$Testing_predictions <- testing_preds

results$Testing_CM <- testing_preds %>%
  purrr::map(function(x) suppressWarnings(caret::confusionMatrix(data = x,
                                                                 reference = testing$Y)))
```

```
results$Testing_MCC <- testing_preds %>%
  purrr::map(function(x) suppressWarnings(mltools::mcc(preds = x,
                                                       actuals = testing$Y))) %>%
  unlist


impVar_list <- lapply(1:length(model_list), function(i){
  vImportance <- caret::varImp(object = model_list[[i]])
  impVar <- data.frame(impVar = rownames(vImportance$importance)[1:3])
  return(impVar)
})


impVar_list <- do.call(cbind, impVar_list)
colnames(impVar_list) <- names(model_list)
results$Important_variables <- impVar_list
```

Ajustados los modelos, veré el MCC (Coeficiente de correlación de Mathews), para saber el accuracy de los modelos. Esto, debido al imbalance de los datos.

```
results$Testing_MCC
```

```
##    ranger svmRadial       knn    avNNet  ensemble
## 0.5677336 0.5129523 0.4370415 0.4820789 0.5027933
```

Se observa que el Random Forest obtuvo un mayor accuracy, es decir, que clasificó mejor los datos ya que logró acertar en un 55.9% de ellos. El modelo que peor los clasificó fue el K-Nearest Neighborhood.

```
results$Testing_CM$ensemble$table
```

```
##            Reference
## Prediction  0  1
##          0 55 13
##          1  5 15
```

También, se ve que en general los modelos clasificaron mejor a los sobrevivientes, y tiene sentido porque representan la mayor cantidad de los datos.

```
results$Important_variables
```

```
##                     ranger                svmRadial
## 1                      age                      age
## 2 creatinine_phosphokinase creatinine_phosphokinase
## 3        ejection_fraction        ejection_fraction
##                          knn                   avNNet
## 1                        age                      age
## 2 creatinine_phosphokinase creatinine_phosphokinase
## 3        ejection_fraction        ejection_fraction
```

Adicionalmente, las variables más importantes para la clasificación fueron la edad, los niveles de creatinina fosfoquinasa y la fracción de eyección.

Conclusiones

Durante el desarrollo de este trabajo, se pudo analizar los diversos factores de riesgo para mortalidad por fallas cardíacas en personas mayores de 40 años en una ciudad de Pakistán. Se encontró que en estos pacientes hay niveles de creatinina fosfoquinasa más altos de lo normal, lo que podría indicar desde ya un alta prevalencia

de accidentes cerebro-vasculares,infartos, inflamación cardíaca y cirugías previas al corazón, sin embargo, esto también puede evidenciar la realización de una actividad física intensa o incluso consumo de drogas ilícitas, por esta razón, considero importante incluir en futuros análisis variables que den cuenta del consumo de sustancias psicoactivas, frecuencia e intensidad de la actividad física, o incluso de los niveles de estrés a los que esté sometido el paciente, ya que esto sin duda afecta el funcionamiento normal del corazón iniciando desde su ritmo. Ahora, las fracciones de eyección se ven normales, sin embargo, son menores entre los fallecidos, y tiene sentido ya que si el músculo cardíaco es grueso y rígido,el ventrículo contiene un volumen de sangre más bajo que lo normal,por lo que la cantidad total de sangre bombeada no sería la suficiente para satisfacer las necesidades del cuerpo; esto podría mostrar que los eventos de muerte fueron en parte por insuficiencias cardíacas, sin embargo, existen pacientes que aún teniendo esta patología, presentan una fracción de eyección normal, en este sentido, parece ser que esta variable por sí sola no determina una causal de muerte segura. La cantidad de plaquetas es normal. Los niveles de creatinina en sangre tan altos como los que se ven, muestran la presencia de enfermedades renales tanto en los pacientes que sobrevivieron como en los que no, sin embargo, los que murieron tenían mucha más probabilidad de desarrollar insuficiencia renal debido a eso niveles. Los que tenían niveles bajos de sodio en sangre, probablemente presentaban altos niveles de azúcar en sangre, o acumulación en la orina de productos de desecho de la descomposición de la grasa, lo que sería también un indicio de daños renales. Adicionalmente, se observó que tener diabetes no necesariamente contribuye a una muerte por fallas cardiacas, tener anemia contribuye un poco, sin embargo, como se vio, los niveles de plaquetas son en general buenos en estos pacientes.

Finalmente, las variables que más ayudan a determinar una muerte por posible falla cardíaca, son los niveles de creatinina tanto en sangre como fosfoquinasa, acompañados de la fracción de eyección y la edad, ya que probablemente a mayor edad, más deterioro del organismo y su funcionamiento en general.

Con respecto a los modelos ajustados, se logró un accuracy apenas superior al 55%, lo que soportaría la necesidad de inclusión de más variables como las mencionadas anteriormente y que den cuenta de los hábitos tanto físicos como nutricionales de las personas.

Contribuciones

```
Contribuciones<-data.frame(Investigacion_previa = "Maria Victoria Diaz",
                           Redaccion_de_las_respuestas =  "Maria Victoria Diaz",
                           Desarrollo_codigo = "Maria Victoria Diaz")

Contribuciones %>% kable()
```

Investigacion_previa	Redaccion_de_las_respuestas	Desarrollo_codigo
Maria Victoria Diaz	Maria Victoria Diaz	Maria Victoria Diaz