Examen parcial

Alejo Nicolás Romano

30/5/2022

Instituto Tecnológico de Buenos Aires

16.43 - Bioestadística

Profesoras:

- SANTA MARÍA, Victoria
- BERRINO, Eugenia
- CIRIGNOLI, Camila

Alumno:

- ROMANO, Alejo Nicolás

Legajo: 59351

Fecha de Entrega: 30 / 05 / 2022

Antes de empezar, se configura el Working-Directory.

```
setwd('E:/ITBA/Bioestadistica/parcial')
```

y se importan las librerías a utilizar en el trabajo.

```
library(ggplot2)
library(magrittr)
library(knitr)
library(dplyr)
library(ggpubr)
library(nortest)
library(cowplot)  # subplots
library(car)
library(ggExtra)  # Graficos afuera de otros graficos
library(epitools)
library(gmodels)
library(readxl)
library(survival)
library(survminer)
library(lubridate)
library(fastDummies)
library(ResourceSelection)
library(lmtest)
```

Para empezar, se carga el archivo **parcial1Q22.xls** en el objeto **datos**. Para verificar que el objeto es efectivamente un dataframe, se utiliza la función **class()**.

```
datos <- read_xls('parcial1Q22.xls')
class(datos)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

A continuación, se observan un resumen de la estructura del dataframe.

```
str(datos)
```

```
## tibble [1,363 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ID            : num [1:1363] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Edad          : num [1:1363] 56 48 48 52 50 50 45 58 54 46 ...
##  $ PAS           : num [1:1363] 132 170 108 110 125 120 126 125 140 114 ...
##  $ PAD           : num [1:1363] 78 84 70 80 85 74 90 65 84 76 ...
##  $ Colesterolemia: num [1:1363] 204 187 340 232 187 217 225 199 217 192 ...
##  $ EC            : num [1:1363] 1 0 0 0 0 1 0 0 1 0 ...
##  $ Hombre        : num [1:1363] 1 0 1 0 1 1 1 1 1 1 ...
##  $ cig           : num [1:1363] 0 0 0 10 0 30 0 20 5 20 ...
##  $ packs/y       : num [1:1363] 0 0 0 1 0 2 0 1 1 1 ...
```

Se puede observar que el dataframe contiene 9 variables de tipo numérico, con 1363 observaciones cada una. La primera variable corresponde simplemente a un índice de los pacientes, por lo que no tiene relevancia en el estudio.

Para obtener un resumen de los datos de cada variable, se realiza el siguiente **summary**.

```
summary(datos)
```

```
##        ID              Edad            PAS             PAD       
##  Min.   :   1.0   Min.   :45.00   Min.   : 90.0   Min.   : 50.00  
##  1st Qu.: 341.5   1st Qu.:48.00   1st Qu.:130.0   1st Qu.: 80.00  
##  Median : 682.0   Median :52.00   Median :142.0   Median : 90.00  
##  Mean   : 682.0   Mean   :52.38   Mean   :147.8   Mean   : 90.02  
##  3rd Qu.:1022.5   3rd Qu.:56.00   3rd Qu.:160.0   3rd Qu.: 98.00  
##  Max.   :1363.0   Max.   :62.00   Max.   :300.0   Max.   :160.00  
##                                                                   
##  Colesterolemia        EC             Hombre            cig        
##  Min.   : 96.0   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000  
##  1st Qu.:200.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000  
##  Median :230.0   Median :0.0000   Median :0.0000   Median : 0.000  
##  Mean   :234.6   Mean   :0.1966   Mean   :0.4718   Mean   : 7.988  
##  3rd Qu.:264.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:20.000  
##  Max.   :430.0   Max.   :1.0000   Max.   :1.0000   Max.   :60.000  
##                                                    NA's   :1       
##     packs/y      
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.5441  
##  3rd Qu.:1.0000  
##  Max.   :2.0000  
##  NA's   :1
```

Se puede observar que las variables **Hombre** y **EC** son dicotómicas, por lo que se tiene que redefinir el tipo de variable a factor. Por otra parte, se puede observar que las variables **cig** y **packs/y** presentan una observación con valores NA.

Se pasa las variables **Hombre** y **EC** a tipo factor.

```
datos$Hombre <- as.factor(datos$Hombre)
datos$EC <- as.factor(datos$EC)

data.frame(Variable = names(datos), Tipo = sapply(datos, class),
           row.names = NULL) %>%
kable()
```

Variable	Tipo
ID	numeric
Edad	numeric
PAS	numeric
PAD	numeric
Colesterolemia	numeric
EC	factor
Hombre	factor
cig	numeric
packs/y	numeric

Luego, se decide eliminar las filas que contienen datos faltantes (NA's), dado que son pocas observaciones y el tamaño de la muestra es grande. Esto se hace con el siguiente código.

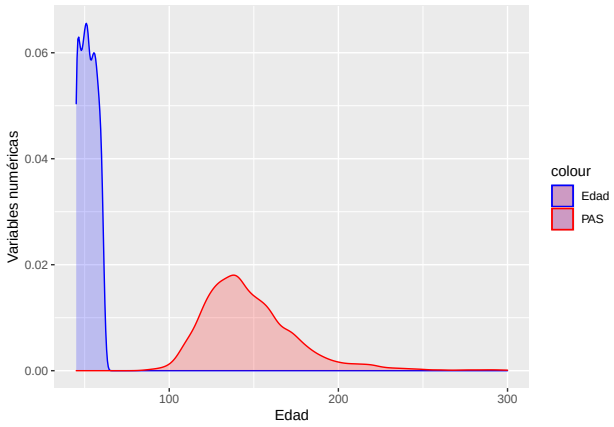
```
datos <- datos[complete.cases(datos), ]
```

Una vez analizada la estructura del dataframe, se obtiene la estadística básica de las variables mediante la función **summary()**.

```
summary(datos)
```

```
##        ID              Edad            PAS              PAD      Colesterolemia
##  Min.   :   1.0   Min.   :45.00   Min.   : 90.0   Min.   : 50   Min.   : 96.0
##  1st Qu.: 341.2   1st Qu.:48.00   1st Qu.:130.0   1st Qu.: 80   1st Qu.:200.0
##  Median : 681.5   Median :52.00   Median :142.0   Median : 90   Median :230.0
##  Mean   : 681.6   Mean   :52.38   Mean   :147.7   Mean   : 90   Mean   :234.6
##  3rd Qu.:1021.8   3rd Qu.:56.00   3rd Qu.:160.0   3rd Qu.: 98   3rd Qu.:264.0
##  Max.   :1363.0   Max.   :62.00   Max.   :300.0   Max.   :160   Max.   :430.0
##  EC       Hombre       cig            packs/y
##  0:1094   0:719   Min.   : 0.000   Min.   :0.0000
##  1: 268   1:643   1st Qu.: 0.000   1st Qu.:0.0000
##                   Median : 0.000   Median :0.0000
##                   Mean   : 7.988   Mean   :0.5441
##                   3rd Qu.:20.000   3rd Qu.:1.0000
##                   Max.   :60.000   Max.   :2.0000
```

```
plt = ggplot(data = datos)
plt + geom_density(aes(x = Edad, color = 'Edad') , fill = 'blue', alpha = 0.2) +
  geom_density(aes(x = PAS, color = 'PAS'), fill = 'red', alpha = 0.2) +
  theme(legend.position = 'right') +
  scale_color_manual(values = c('Edad' = 'blue', 'PAS' = 'red')) +
  labs(y = "Variables numéricas")
```



Ejercicio 1

El trabajo planteado corresponde al diseño de un estudio de cohortes, dado que el estudio es de tipo observacional y consiste en seguir un grupo de muestras en un intervalo de tiempo. Esto tiene sentido dado que este diseño de estudio se utiliza para evaluar la relación entre una exposición y la ocurrencia de un evento de interés, lo que permite identificar factores de riesgo.

La pregunta PICO para esta investigación es la siguiente:

- **Población (P).** Pacientes que consultaron por primera vez en el año 2010 y 2011 al servicio de Cardiología del Hospital de Clínicas.
- **Intervención (I).** Las covariables del estudio, es decir, la exposición que tienen los sujetos (sexo, colesterol, cigarrillos, presiones, etc.).
- **Comparación (C).** No hay comparación, dado que no se evalúan poblaciones diferentes entre sí.
- **Resultado (O).** Enfermedad coronaria (EC).

Ejercicio 2

El diseño experimental puede presentar los siguientes sesgos:

- **De selección.** Dado que los pacientes asisten al servicio de cardiología de un hospital, estos presentan mayor prevalencia del evento de interés. Además, puede ser que presenten otros factores de riesgo o enfermedades graves, los cuales pueden interferir en el estudio o incluso hacer que el paciente abandone el estudio (p. ej. sujeto fallece por otra enfermedad cardíaca).
- **De clasificación.** Es probable que los sujetos ya presenten el evento de interés al inicio del estudio, dado que asisten al servicio de cardiología, pero que aún no se haya detectado.
- **De información.** Es posible que haya errores en la medición o tomas de datos en la etapa inicial del estudio, lo cual afecta la exposición de los sujetos y, por ende, afecta la relación con el outcome.

Ejercicio 3

Las variables **Edad**, **PAS**, **PAD**, **Colesterolemia**, **Cig** y **Packs/y** son variables numéricas, por lo que se estudia la distribución de estas variables mediante la siguiente función de normalidad:

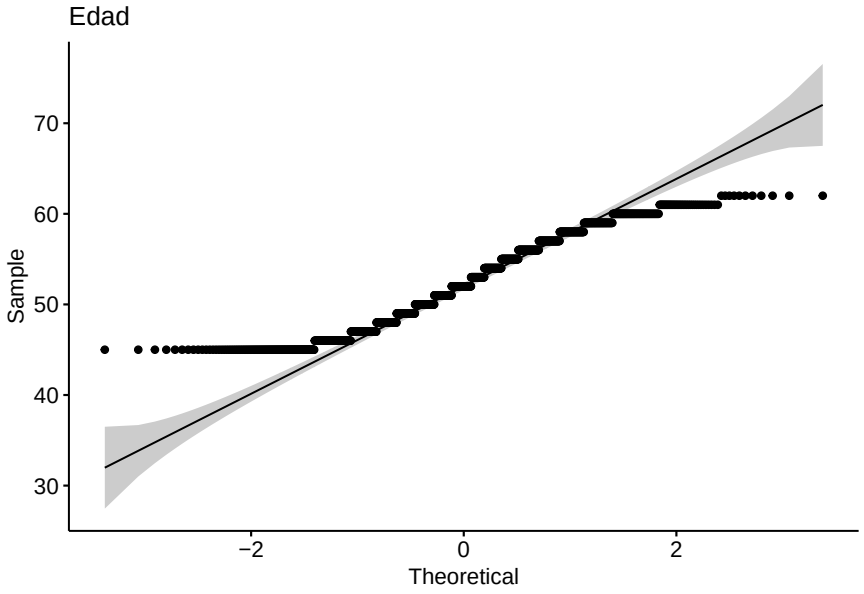
```
pruebas_normalidad = function(dataframe, xlab = "X", ylab = "Densidad", main = "Título")
{
  # Tests de Normalidad
  print(lillie.test(dataframe))
  print(shapiro.test(dataframe))

  # Q-Q plot
  print(ggqqplot(dataframe, main=main))

  # Estadísticos básicos
  min <- min(dataframe)
  max <- max(dataframe)
  media <- mean(dataframe)
  desvio <- sd(dataframe)

  # Histograma
  hist(dataframe, freq = F, main = main, xlab = xlab, ylab = ylab, border = "salmon",
       col = "khaki")
  curve(dnorm(x, media, desvio), from = min, to = max, add = TRUE, col = "black")
}
```

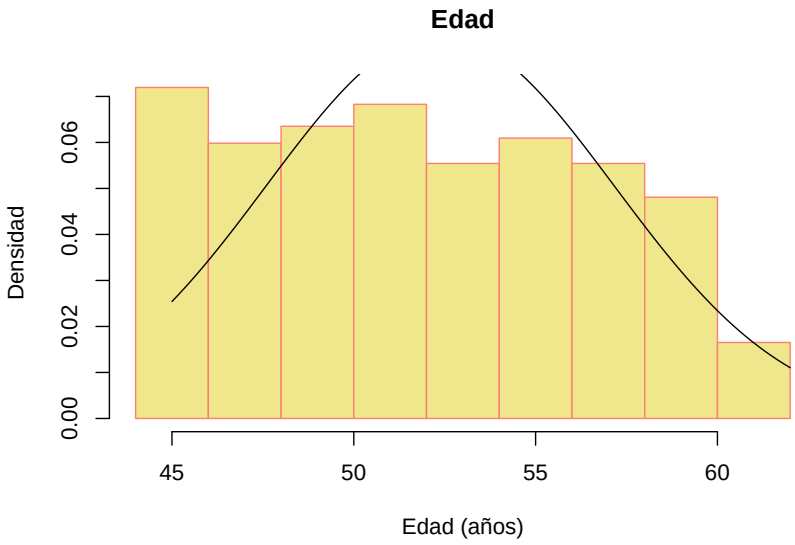
Esta función realiza los Test Shapiro-Wilk y el Test Kolgomorov-Smirnov, los cuales toman como hipótesis nula que la variable analizada presenta una distribución normal, mientras que toman como hipótesis alternativa que la variable no cumple con la normalidad.

Para la variable **Edad**, se obtienen los siguientes resultados.

```
pruebas_normalidad(datos$Edad, "Edad (años)", "Densidad", "Edad")
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataframe
## D = 0.08332, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  dataframe
## W = 0.95401, p-value < 2.2e-16
```





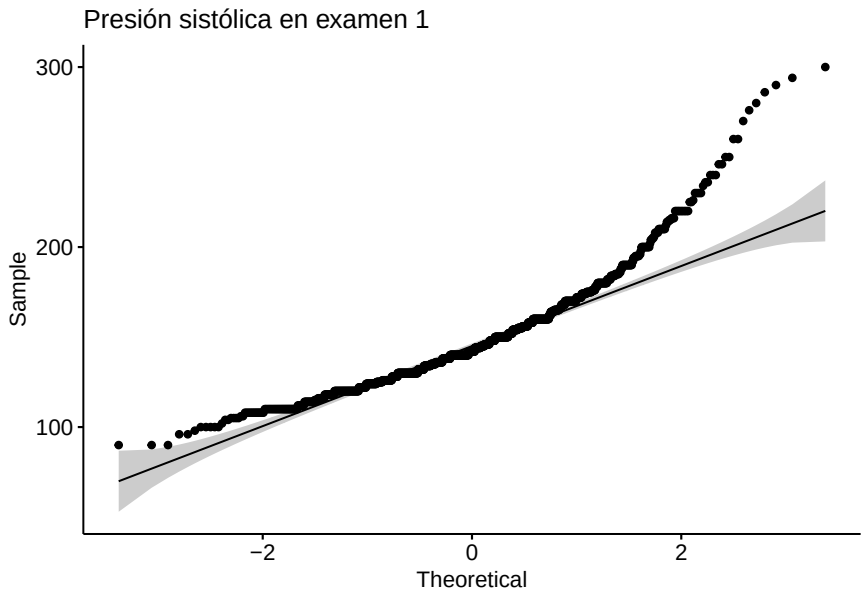
Se puede observar que ambos p-values ($< 2,2.10^{-16}$) dan menores a 0.05, por lo que se rechaza la hipótesis nula mencionada y se acepta la hipótesis alternativa. Esto quiere decir que la variable **Edad** NO tiene distribución normal.

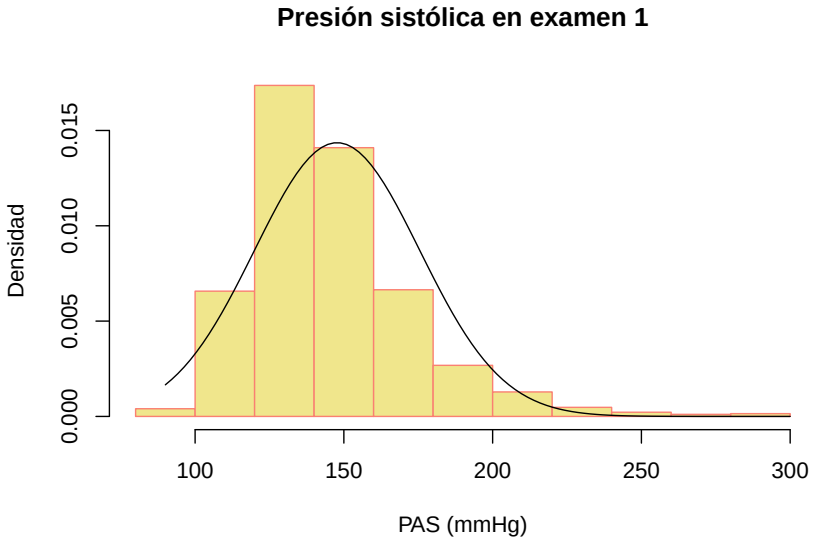
Con respecto a los métodos gráficos, en el Q-Q plot se puede observar que la distribución se aleja de la recta de distribución teórica normal. Los puntos salen de la región gris tanto en el medio como en el extremo de la derecha, lo que indica que estos puntos caen por fuera del intervalo de confianza tomado (95%). Por lo tanto, este gráfico muestra que la distribución de la variable en cuestión no es normal, lo que es coherente con los test de normalidad. Además, se puede observar en el histograma que la distribución no se asemeja a la campana simétrica teórica de una distribución normal, lo cual indica que la variable no presenta distribución normal.

Para la variable **PAS**, se obtienen los siguientes resultados.

```
pruebas_normalidad(datos$PAS, "PAS (mmHg)", "Densidad", "Presión sistólica en examen 1")
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataframe
## D = 0.10243, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  dataframe
## W = 0.91433, p-value < 2.2e-16
```





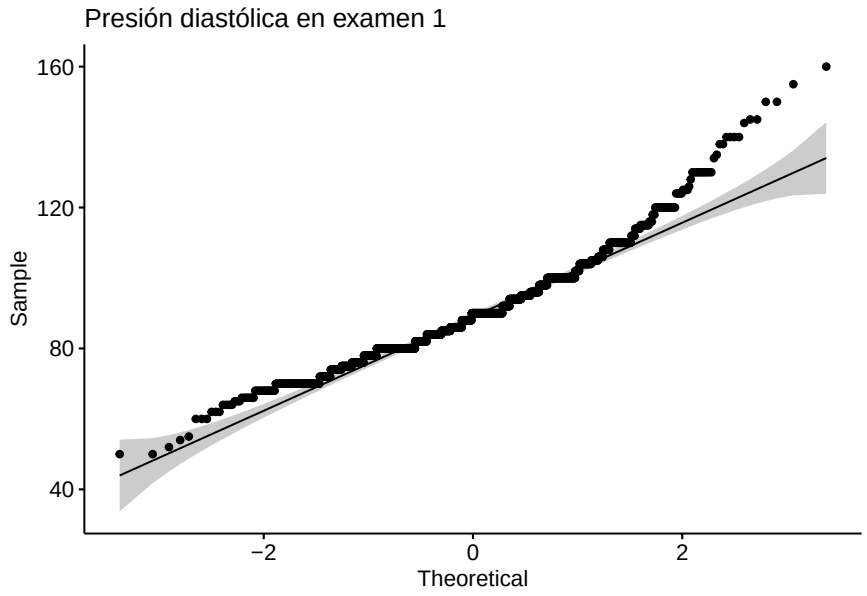
Se puede observar que ambos p-values ($< 2,2.10^{-16}$) dan menores a 0.05, por lo que se rechaza la hipótesis nula mencionada y se acepta la hipótesis alternativa. Esto quiere decir que la variable **PAS** NO tiene distribución normal.

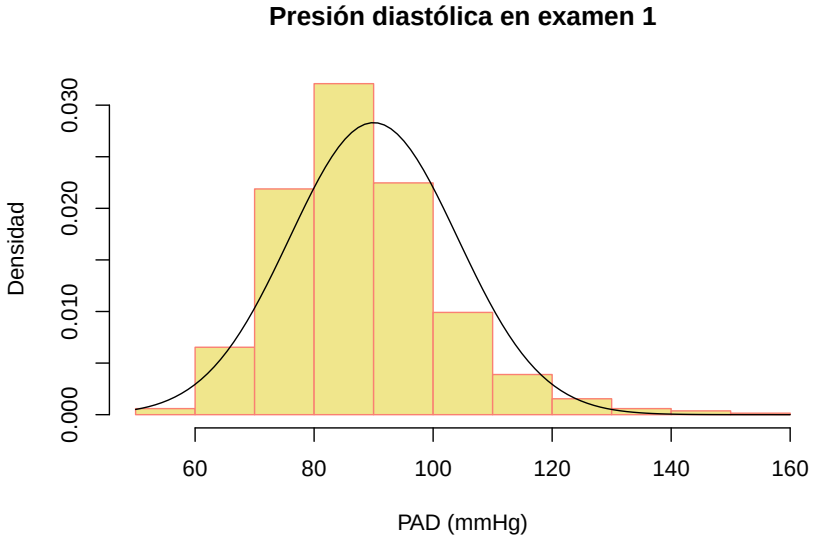
Con respecto a los métodos gráficos, en el Q-Q plot se puede observar que la distribución se aleja de la recta de distribución teórica normal. Los puntos salen de la región gris tanto en el medio como en el extremo de la derecha, lo que indica que estos puntos caen por fuera del intervalo de confianza tomado (95%). Por lo tanto, este gráfico muestra que la distribución de la variable en cuestión no es normal, lo que es coherente con los test de normalidad. Además, se puede observar en el histograma que la distribución no se asemeja a la campana simétrica teórica de una distribución normal, sino que los datos tienden hacia la izquierda. Esto indica que la variable no presenta distribución normal.

Para la variable **PAD**, se obtienen los siguientes resultados.

```
pruebas_normalidad(datos$PAD, "PAD (mmHg)", "Densidad", "Presión diastólica en examen 1")
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataframe
## D = 0.11093, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  dataframe
## W = 0.95592, p-value < 2.2e-16
```





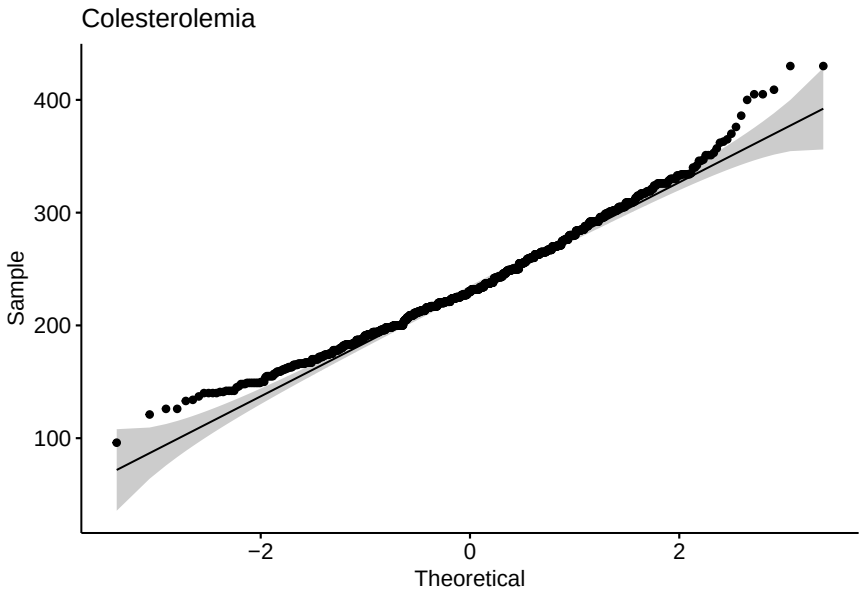
Se puede observar que ambos p-values ($< 2, 2.10^{-16}$) dan menores a 0.05, por lo que se rechaza la hipótesis nula mencionada y se acepta la hipótesis alternativa. Esto quiere decir que la variable **PAD** NO tiene distribución normal.

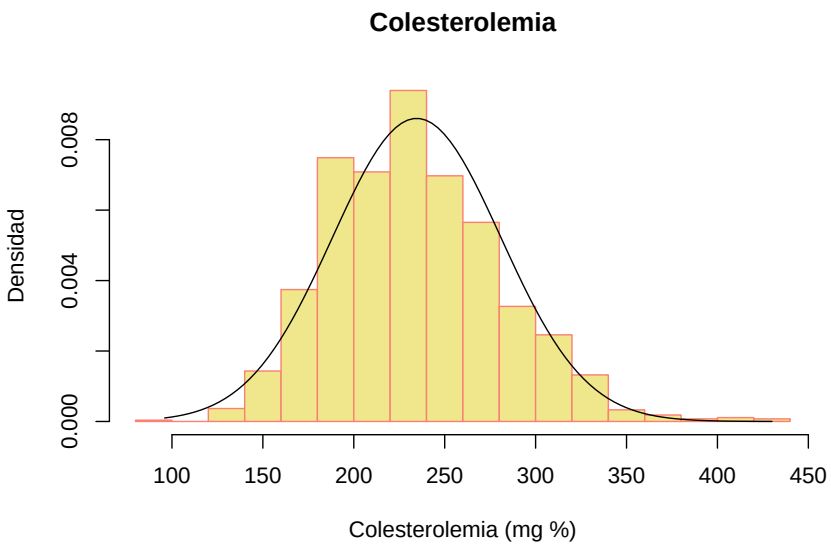
Con respecto a los métodos gráficos, en el Q-Q plot se puede observar que la distribución se aleja de la recta de distribución teórica normal. Los puntos salen de la región gris tanto en el medio como en el extremo de la derecha, lo que indica que estos puntos caen por fuera del intervalo de confianza tomado (95%). Por lo tanto, este gráfico muestra que la distribución de la variable en cuestión no es normal, lo que es coherente con los test de normalidad. Además, se puede observar en el histograma que la distribución no se asemeja a la campana simétrica teórica de una distribución normal, sino que los datos tienden hacia la izquierda. Esto indica que la variable no presenta distribución normal.

Para la variable **Colesterolemia**, se obtienen los siguientes resultados.

```
pruebas_normalidad(datos$Colesterolemia, "Colesterolemia (mg %)", "Densidad", "Colesterolemia")
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataframe
## D = 0.058874, p-value = 5.798e-12
##
##
##  Shapiro-Wilk normality test
##
## data:  dataframe
## W = 0.98606, p-value = 3.742e-10
```





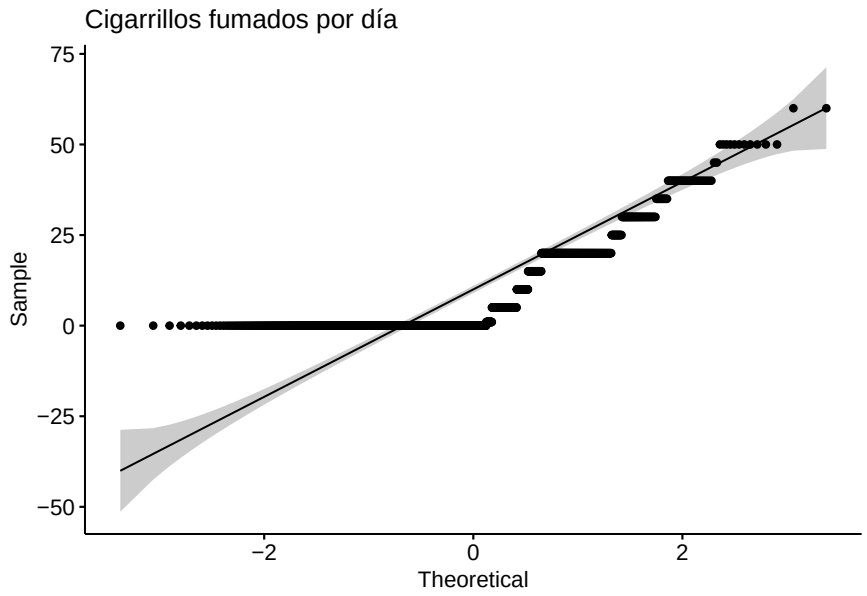
Se puede observar que ambos p-values ($5,798.10^{-12}$ para Kolmogorov y $3,742.10^{-10}$ para Shapiro) dan menores a 0.05, por lo que se rechaza la hipótesis nula mencionada y se acepta la hipótesis alternativa. Esto quiere decir que la variable **Colesterolemia** NO tiene distribución normal.

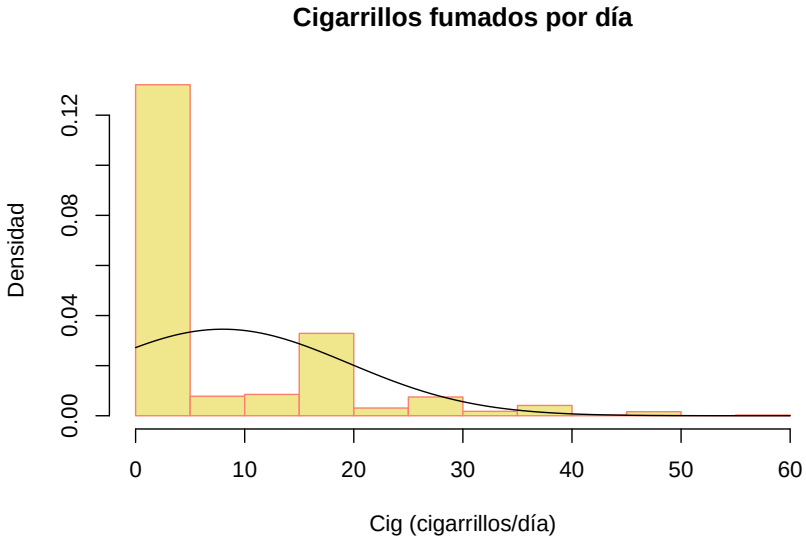
Con respecto a los métodos gráficos, en el Q-Q plot se puede observar que la distribución se aleja de la recta de distribución teórica normal. Los puntos salen de la región gris tanto en el medio como en el extremo de la derecha, lo que indica que estos puntos caen por fuera del intervalo de confianza tomado (95%). Por lo tanto, este gráfico muestra que la distribución de la variable en cuestión no es normal, lo que es coherente con los test de normalidad. Además, se puede observar en el histograma que la distribución se asemeja un poco a la campana simétrica teórica de una distribución normal, pero los datos tienden levemente hacia la izquierda. Esto indica que la variable no presenta distribución normal.

Para la variable **cig**, se obtienen los siguientes resultados.

```
pruebas_normalidad(datos$cig, "Cig (cigarrillos/día)", "Densidad", "Cigarrillos fumados por día")
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataframe
## D = 0.30466, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  dataframe
## W = 0.72426, p-value < 2.2e-16
```





Se puede observar que ambos p-values ($< 2,2.10^{-16}$) dan menores a 0.05, por lo que se rechaza la hipótesis nula mencionada y se acepta la hipótesis alternativa. Esto quiere decir que la variable **cig** NO tiene distribución normal.

Con respecto a los métodos gráficos, en el Q-Q plot se puede observar que la distribución se aleja de la recta de distribución teórica normal. Los puntos salen de la región gris tanto en el medio como en el extremo de la derecha, lo que indica que estos puntos caen por fuera del intervalo de confianza tomado (95%). Por lo tanto, este gráfico muestra que la distribución de la variable en cuestión no es normal, lo que es coherente con los test de normalidad. Además, se puede observar en el histograma que la distribución no se asemeja a la campana simétrica teórica de una distribución normal, lo cual indica que la variable no presenta distribución normal.

Se puede observar que las variables analizadas no presentan distribución normal. Además, dado que son observaciones independientes (por el diseño del estudio), se utiliza el Test Wilcoxon rank-sum para el análisis univariado de cada variable numérica continua (comparar el outcome para cada variable por separado).

La variable **Hombre** es dicotómica, no presenta unidades y tiene una distribución binomial. Por lo tanto, se debe usar el Test Chi cuadrado (o Fisher) para comparar el outcome para esta variable.

La varibale **Packs** es categórica ordinal, dado que se utiliza una escala que permite medir la cantidad de paquetes de cigarrillos fumados por año. No presenta unidad y su distribución es xxxxxxx. Por lo tanto, se debe utilizar el Test Chi cuadrado para comparar el outcome para esta variable.

Por último, la variable **EC** es la variable respuesta (el resto de las variables son explicatorias) y es dicotómica. No presenta unidad y su distribución es binomial.

Ejercicio 4

En primer lugar, se analizan las variables numéricas continuas, que son las siguientes: **Edad**, **PAS**, **PAD**, **Colesterolemia** y **cig**. En primer lugar, se calculan las proporciones o estadísticos de los dos grupos según la variable **EC**. Dado que son variables numéricas continuas que no presentan distribución normal, se calcula la mediana y el rango intercuartil para cada una.

```
mediana_edad <- aggregate(datos$Edad, list(datos$EC), median)
mediana_pas <- aggregate(datos$PAS, list(datos$EC), median)
mediana_pad <- aggregate(datos$PAD, list(datos$EC), median)
mediana_colest <- aggregate(datos$Colesterolemia, list(datos$EC), median)
mediana_cig <- aggregate(datos$cig, list(datos$EC), median)

paste('Mediana - (No EC | Si EC)')
```

```
## [1] "Mediana - (No EC | Si EC)"
```

```
mediana_edad[,2]
```

```
## [1] 52 54
```

```
mediana_pas[,2]
```

```
## [1] 140.0 154.5
```

```
mediana_pad[,2]
```

```
## [1] 88 92
```

```
mediana_colest[,2]
```

```
## [1] 230 232
```

```
mediana_cig[,2]
```

```
## [1] 0.0 0.5
```

```r
iqr_edad <- aggregate(datos$Edad, list(datos$EC), IQR)
iqr_pas <- aggregate(datos$PAS, list(datos$EC), IQR)
iqr_pad <- aggregate(datos$PAD, list(datos$EC), IQR)
iqr_colest <- aggregate(datos$Colesterolemia, list(datos$EC), IQR)
iqr_cig <- aggregate(datos$cig, list(datos$EC), IQR)

paste('IQR - (No EC | Si EC)')
```

```
## [1] "IQR - (No EC | Si EC)"
```

```r
iqr_edad[,2]
```

```
## [1] 8 8
```

```r
iqr_pas[,2]
```

```
## [1] 30 36
```

```r
iqr_pad[,2]
```

```
## [1] 16 20
```

```r
iqr_colest[,2]
```

```
## [1] 63.00 70.25
```

```r
iqr_cig[,2]
```

```
## [1] 15 20
```

Luego, se realizan los correspondientes test para evaluar si los dos grupos para cada variable son estadísticamente diferentes o no. Para estas variables, dado que no presentan distribución normal y los dos grupos son independientes entre sí, se realiza el Test Wilcoxon rank-sum. Este Test tiene como hipótesis nula que las medianas de los dos grupos analizados (para cada variable) tienen medianas iguales, mientras que tiene como hipótesis alternativa que las medianas son significativamente diferentes.

Para la variable **edad**, el test se realiza con el siguiente código.

```r
wilcox.test(datos$Edad[datos$EC==0], datos$Edad[datos$EC==1],
                           exact = F, conf.int = T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  datos$Edad[datos$EC == 0] and datos$Edad[datos$EC == 1]
## W = 120048, p-value = 4.046e-06
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -2.0000497 -0.9999591
## sample estimates:
## difference in location
##              -1.999964
```

El Test da como resultado un p-value ($4,046.10^{-06}$) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, hay diferencias significativas entre las edades de los grupos con y sin EC.

Para la variable **PAS**, el test se realiza con el siguiente código.

```
wilcox.test(datos$PAS[datos$EC==0], datos$PAS[datos$EC==1],
                           exact = F, conf.int = T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  datos$PAS[datos$EC == 0] and datos$PAS[datos$EC == 1]
## W = 106197, p-value = 2.475e-12
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -15.999958  -9.000039
## sample estimates:
## difference in location
##              -12.00001
```

El Test da como resultado un p-value ($2,475.10^{-12}$) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, hay diferencias significativas entre las presiones sistólicas en el primer examen de los grupos con y sin EC.

Para la variable **PAD**, el test se realiza con el siguiente código.

```
wilcox.test(datos$PAD[datos$EC==0], datos$PAD[datos$EC==1],
                           exact = F, conf.int = T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  datos$PAD[datos$EC == 0] and datos$PAD[datos$EC == 1]
## W = 113964, p-value = 1.46e-08
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -7.000041 -3.999981
## sample estimates:
## difference in location
##              -5.000015
```

El Test da como resultado un p-value ($1,460.10^{-08}$) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, hay diferencias significativas entre las presiones diastólicas en el primer examen de los grupos con y sin EC.

Para la variable **Colesterolemia**, el test se realiza con el siguiente código.

```
wilcox.test(datos$Colesterolemia[datos$EC==0], datos$Colesterolemia[datos$EC==1],
                           exact = F, conf.int = T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  datos$Colesterolemia[datos$EC == 0] and datos$Colesterolemia[datos$EC == 1]
## W = 135018, p-value = 0.04481
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -1.300000e+01 -5.248038e-06
## sample estimates:
## difference in location
##              -6.000036
```

El Test da como resultado un p-value (0.04481) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, hay diferencias significativas entre las colesterolemias en el primer examen de los grupos con y sin EC. Hay que destacar que el p-value obtenido se encuentra muy cerca del valor 0.05.

Para la variable **cig**, el test se realiza con el siguiente código.

```
wilcox.test(datos$cig[datos$EC==0], datos$cig[datos$EC==1],
                           exact = F, conf.int = T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  datos$cig[datos$EC == 0] and datos$cig[datos$EC == 1]
## W = 133661, p-value = 0.01382
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -5.569379e-05 -4.646154e-05
## sample estimates:
## difference in location
##          -1.314778e-05
```

El Test da como resultado un p-value (0.01382) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, hay diferencias significativas entre los cigarrillos fumados por día de los grupos con y sin EC.

Con respecto a la variable **Hombre**, esta es dicotómica con distribución binomial. Primero, se calculan las proporciones de sujetos entre los valores de esta variable y los de la variable resultado, **EC**.

```
summary(datos$Hombre[datos$EC == 0])
```

```
##   0   1
## 615 479
```

```
summary(datos$Hombre[datos$EC == 1])
```

```
##   0   1
## 104 164
```

Se puede observar que hay 615 sujetos no hombres sin EC, 479 sujetos hombres sin EC, 104 sujetos no hombres con EC y 164 sujetos hombres con EC. La proporción en porcentaje de estos valores, utilizando el total de las observaciones (1362), son 45.15% para sujetos no hombres sin EC, 35.17% para sujetos hombres sin EC, 7.64% para sujetos no hombres con EC y 12.04% para sujetos hombres con EC.

Dado que los grupos a evaluar (con y sin EC) son independientes entre sí, se utiliza el Test Chi Cuadrado para evaluar la diferencia entre dichos grupos. Este Test toma como hipótesis nula que todas las probabilidades de ocurrencia de las celdas de la tabla de contingencia son iguales entre sí, es decir, que las variables son independientes. La hipótesis alternativa establece que al menos una de esas probabilidades es distinta al resto, es decir, que las variables no son independientes. A continuación, se realiza el test indicando que se utilice la corrección de Yates, ya que la tabla de contingencia es de 2x2.

ver si poner la tabla de contingencia

```
chisq_1 <- chisq.test(datos$Hombre, datos$EC, correct = TRUE)
chisq_1
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$Hombre and datos$EC
## X-squared = 25.486, df = 1, p-value = 4.455e-07
```

El Test da como resultado un p-value ($4,455.10^{-07}$) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, las variables **Hombre** y **EC** son independientes entre sí (no hay relación).

Por último, con respecto a la variable **packs**, esta es categórica ordinal, dado que utiliza una escala de tres valores. Primero, se calculan las proporciones de sujetos entre los valores de esta variable y los de la variable resultado, **EC**.

```
sum(datos$EC == 0 & datos$`packs/y` == 0)
```

```
## [1] 614
```

```
sum(datos$EC == 0 & datos$`packs/y` == 1)
```

```
## [1] 388
```

```
sum(datos$EC == 0 & datos$`packs/y` == 2)
```

```
## [1] 92
```

```
sum(datos$EC == 1 & datos$`packs/y` == 0)
```

```
## [1] 134
```

```
sum(datos$EC == 1 & datos$`packs/y` == 1)
```

```
## [1] 99
```

```
sum(datos$EC == 1 & datos$`packs/y` == 2)
```

```
## [1] 35
```

Se puede observar que hay 614 sujetos sin EC y con consumo bajo de packs (categoría 0), 388 sujetos sin EC y con consumo intermedio de packs (categoría 1), 92 sujetos sin EC y con consumo alto de packs (categoría 2), 134 sujetos con EC y con consumo bajo de packs, 99 sujetos con EC y con consumo intermedio de packs y 35 sujetos con EC y con consumo alto de packs. La proporción en porcentaje de estos valores, utilizando el total de las observaciones (1362), son 45.08% para sujetos sin EC y con consumo bajo de packs, 28.49% para sujetos sin EC y con consumo intermedio de packs, 6.75% para sujetos sin EC y con consumo alto de packs, 9.84% para sujetos con EC y con consumo bajo de packs, 7.27% para sujetos con EC y con consumo intermedio de packs y 2.57% para sujetos con EC y con consumo alto de packs.

Dado que los grupos a evaluar (con y sin EC) son independientes entre sí, se utiliza el Test Chi Cuadrado para evaluar la diferencia entre dichos grupos. Este Test toma como hipótesis nula que todas las probabilidades de ocurrencia de las celdas de la tabla de contingencia son iguales entre sí, es decir, que las variables son independientes. La hipótesis alternativa establece que al menos una de esas probabilidades es distinta al resto, es decir, que las variables no son independientes.

ver si poner la tabla de contingencia

```
chisq_1 <- chisq.test(datos$`packs/y`, datos$EC)
chisq_1
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$`packs/y` and datos$EC
## X-squared = 6.5932, df = 2, p-value = 0.03701
```

El Test da como resultado un p-value (0.03701) menor a 0.05. Por lo tanto, se rechaza la hipótesis nula y se acepta la alternativa, es decir, las variables **packs/y** y **EC** son independientes entre sí (no hay relación).

Ejercicio 5

Dado que la variable resultado es categórica y dicotómica, se debe usar un modelo de regresión logística para poder obtener la probabilidad de ocurrencia de dicha variable a partir de las variables explicatorias. De esta manera, se puede realizar un análisis multivariado de la base de datos que se está estudiando.

Antes de realizar el modelo completo, se realiza las variables dummies para la variable packs, es decir, se generan tres nuevas variables, una para cada valor posible de packs. Así, se obtienen variables dicotómicas (con dos valores: 0 y 1).

```
packs_dummy <- dummy_cols(datos$`packs/y`)
colnames(packs_dummy)<-c('packs','packs0', 'packs1', 'packs2')
datos_dummy<-cbind(datos,packs_dummy)
str(datos_dummy)
```

```
## 'data.frame':    1362 obs. of  13 variables:
##  $ ID            : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Edad          : num  56 48 48 52 50 50 45 58 54 46 ...
##  $ PAS           : num  132 170 108 110 125 120 126 125 140 114 ...
##  $ PAD           : num  78 84 70 80 85 74 90 65 84 76 ...
##  $ Colesterolemia: num  204 187 340 232 187 217 225 199 217 192 ...
##  $ EC            : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 2 1 ...
##  $ Hombre        : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 2 2 ...
```

```
##  $ cig           : num  0 0 0 10 0 30 0 20 5 20 ...
##  $ packs/y       : num  0 0 0 1 0 2 0 1 1 1 ...
##  $ packs         : num  0 0 0 1 0 2 0 1 1 1 ...
##  $ packs0        : int  1 1 1 0 1 0 1 0 0 0 ...
##  $ packs1        : int  0 0 0 1 0 0 0 1 1 1 ...
##  $ packs2        : int  0 0 0 0 0 1 0 0 0 0 ...
```

```
datos_dummy$packs0 <- as.factor(datos_dummy$packs0)
datos_dummy$packs1 <- as.factor(datos_dummy$packs1)
datos_dummy$packs2 <- as.factor(datos_dummy$packs2)
str(datos_dummy)
```

```
## 'data.frame':    1362 obs. of  13 variables:
##  $ ID            : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Edad          : num  56 48 48 52 50 50 45 58 54 46 ...
##  $ PAS           : num  132 170 108 110 125 120 126 125 140 114 ...
##  $ PAD           : num  78 84 70 80 85 74 90 65 84 76 ...
##  $ Colesterolemia: num  204 187 340 232 187 217 225 199 217 192 ...
##  $ EC            : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 2 1 ...
##  $ Hombre        : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 2 2 ...
##  $ cig           : num  0 0 0 10 0 30 0 20 5 20 ...
##  $ packs/y       : num  0 0 0 1 0 2 0 1 1 1 ...
##  $ packs         : num  0 0 0 1 0 2 0 1 1 1 ...
##  $ packs0        : Factor w/ 2 levels "0","1": 2 2 2 1 2 1 2 1 1 1 ...
##  $ packs1        : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 2 2 ...
##  $ packs2        : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
```

En primer lugar, se realiza el modelo general linealizado completo, es decir, utilizando todas las variables del dataframe, con excepción de la variable ID y la variable packs sin modificar (**Modelo saturado**). Este modelo de regresión logística queda definida entonces por la siguiente ecuación:

$$logit(p) = \beta_0 + \beta_1 Edad + \beta_2 PAS + \beta_3 PAD + \beta_4 Colesterolemia + \beta_5 Hombre + \beta_6 cig + \beta_7 packs0 + \beta_8 packs1 + \beta_9 packs2 \quad (1)$$

Se genera el modelo y se imprime los resultados.

```
logistica_1 <- glm(EC ~ Edad + PAS + PAD + Colesterolemia + Hombre + cig +
                     packs0 + packs1 + packs2, data = datos_dummy, family = "binomial")
summary(logistica_1)
```

```
##
## Call:
## glm(formula = EC ~ Edad + PAS + PAD + Colesterolemia + Hombre +
##     cig + packs0 + packs1 + packs2, family = "binomial", data = datos_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6654  -0.6918  -0.5282  -0.3499   2.4849
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.768825   1.147716  -7.640 2.17e-14 ***
```

```
## Edad            0.057720   0.015689   3.679 0.000234 ***
## PAS             0.015013   0.004018   3.737 0.000187 ***
## PAD             0.004891   0.007857   0.623 0.533586
## Colesterolemia  0.004614   0.001580   2.920 0.003495 **
## Hombre1         0.906132   0.164425   5.511 3.57e-08 ***
## cig             0.009888   0.014490   0.682 0.495009
## packs01        -0.070526   0.546254  -0.129 0.897272
## packs11        -0.079069   0.380820  -0.208 0.835520
## packs21               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.8  on 1361  degrees of freedom
## Residual deviance: 1239.3  on 1353  degrees of freedom
## AIC: 1257.3
##
## Number of Fisher Scoring iterations: 4
```

El test realizado para los coeficientes tiene como hipótesis nula que el coeficiente β_i es igual a cero, y tiene como hipótesis alternativa que el coeficiente es distinto de cero. Se puede observar que las variables **PAD**, **cig**, **packs0** y **packs1** presentan un p-value mayor a 0.05, por lo que se acepta la hipótesis nula y, por ende, estos coeficientes son iguales a 0 (no significativos para el modelo). Por otro lado, el resto de las variables, incluida la ordenada al origen (β_0) presentan p-values menores a 0.05, por lo que se rechaza la hipótesis nula y se acepta la alternativa. Esto quiere decir que los coeficientes asociados a estas variables son significativos y no nulos: la ordenada al origen es -8.768825, el coeficiente para la edad es 0.057720, el coeficiente para la PAS es 0.015013, el coeficiente para la colesterolemia es 0.004614 y el coeficiente para hombre es 0.906132. Estos coeficientes indican la relación que existe entre sus respectivas variables y la probabilidad de ocurrencia de EC.

Cabe destacar que la variable **packs2** tiene valores NaNs dado que, al hacer variables dummy, se comparan las variables contra una de ellas. En este caso, la variable que se utiliza para comparar es **packs2**.

Por último, el modelo queda definida por la siguiente ecuación:

$$logit(p) = -8.768825 + 0.057720Edad + 0.015013PAS + 0.004614Colesterolemia + 0.906132Hombre \quad (2)$$

Luego, se obtienen los Odds Ratio y sus respectivos intervalos de confianza.

```
OR1 = exp(coefficients(logistica_1))
b = exp(confint(logistica_1))
OR = cbind(OR1,b)
OR
```

```
##                         OR1        2.5 %      97.5 %
## (Intercept)    0.0001555063 1.586333e-05 0.001432487
## Edad           1.0594180896 1.027455e+00 1.092686429
## PAS            1.0151264415 1.007179e+00 1.023201997
## PAD            1.0049030196 9.895332e-01 1.020531367
## Colesterolemia 1.0046243616 1.001513e+00 1.007740937
## Hombre1        2.4747308567 1.796847e+00 3.425357220
## cig            1.0099366699 9.817073e-01 1.039190472
```

```
## packs01        0.9319034785 3.228206e-01 2.757004101
## packs11        0.9239765522 4.394462e-01 1.960567902
## packs21                  NA           NA          NA
```

Se puede observar en los resultados de OR que las variables **PAD**, **cig**, **packs0** y **packs1** tienen intervalos de confianza que incluyen al 1. Esto quiere decir que dichas variables son independientes del outcome (**EC**). Por otro lado, el resto de las variables no son independientes del outcome y son significativas. Para la **edad**, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que por cada año que aumenta la edad, el riesgo de sufrir una EC aumenta en un 6% (o aumenta 1.06 veces). Para la **PAS**, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que por cada aumento unitario de PAS, el riesgo de sufrir una EC aumenta un 1.5% (o aumenta 1.015 veces). Para la variable **colesterolemia**, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que por cada aumento unitario de colesterolemia, el riesgo de sufrir una EC aumenta en un 0.46% (o 1.0046 veces). Para la variable **Hombre**, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que ser hombre aumenta el riesgo de sufrir una EC en un 147% (o aumenta 2.4747 veces).

omnibus para ver si los coeficientes son globalmente significativos.

```
datos_dummy$EC <- as.numeric(datos_dummy$EC)
omnibus <- aov(logistica_1, data = datos_dummy)
summary(omnibus)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## Edad              1   3.33   3.331  22.776 2.02e-06 ***
## PAS               1   6.53   6.533  44.663 3.41e-11 ***
## PAD               1   0.16   0.159   1.088   0.2970
## Colesterolemia    1   0.36   0.360   2.462   0.1169
## Hombre            1   6.50   6.499  44.434 3.82e-11 ***
## cig               1   0.46   0.459   3.138   0.0767 .
## packs0            1   0.01   0.010   0.066   0.7974
## packs1            1   0.01   0.008   0.051   0.8205
## Residuals      1353 197.91   0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

conclusion

Para saber que tan bien ajusta el modelo generado al conjunto de datos, se realiza la prueba de bondad de ajuste de Hosmer-Lemeshow. En función de que tan bien esta ajustado el modelo a los datos, se puede inferir que tan certero o acertado va a ser el modelo. Este test compara las frecuencias esperadas con las frecuencias observadas del modelo. Si las frecuencias observadas son similares a las frecuencias esperadas, entonces el modelo presenta un buen ajuste. La hipótesis nula que se plantea para esta prueba es que la diferencia entre frecuencias observadas y las esperadas es igual a cero (observado igual a esperado), mientras que la hipótesis alternativa indica que esta diferencia es distinta de cero.

```
prediccion <- predict(logistica_1, type="response")
hosmer <- hoslem.test(logistica_1$y, logistica_1$fitted.values) # Test Hosmwe-Lemeshow
hosmer
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  logistica_1$y, logistica_1$fitted.values
## X-squared = 8.3939, df = 8, p-value = 0.396
```

Como el Test de bondad de ajuste de Hosmer-Lemeshow da un p-value (0.396) mayor a 0.05, entonces la diferencia entre lo observado y lo esperado es igual a cero y, por ende, el modelo ajusta bien al conjunto de datos que se tiene.

Ejercicio 6

A continuación, se decide utilizar el método **backward elimination** para ir sacando aquellas variables que no son significativas para el modelo, es decir, que no tengan una contribución parcial. Para esto, se elimina la variable que tiene menor contribución, es decir, la variable que obtenga el p-value más grande. Estos p-values se obtienen de los test de hipótesis aplicados a los coeficientes, donde las hipótesis son análogas a las mencionadas en el modelo saturado.

Dado que en el modelo saturado la variable **packs0** es la que presenta un p-value mayor, es la que se elimina primero.

```
datos_dummy$EC <- as.factor(datos_dummy$EC)
logistica_2_1 <- glm(EC ~ Edad + PAS + PAD + Colesterolemia + Hombre + cig +
                     packs1 + packs2, data = datos_dummy, family = "binomial")
summary(logistica_2_1)
```

```
##
## Call:
## glm(formula = EC ~ Edad + PAS + PAD + Colesterolemia + Hombre +
##     cig + packs1 + packs2, family = "binomial", data = datos_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6654  -0.6918  -0.5282  -0.3499   2.4849
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.839351   1.035192  -8.539  < 2e-16 ***
## Edad            0.057720   0.015689   3.679 0.000234 ***
## PAS             0.015013   0.004018   3.737 0.000187 ***
## PAD             0.004891   0.007857   0.623 0.533586
## Colesterolemia  0.004614   0.001580   2.920 0.003495 **
## Hombre1         0.906132   0.164425   5.511 3.57e-08 ***
## cig             0.009888   0.014490   0.682 0.495009
## packs11        -0.008543   0.251505  -0.034 0.972904
## packs21         0.070526   0.546254   0.129 0.897272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.8  on 1361  degrees of freedom
## Residual deviance: 1239.3  on 1353  degrees of freedom
## AIC: 1257.3
##
## Number of Fisher Scoring iterations: 4
```

Se observa que, al eliminar la variable **packs0**, todas las variables se mantuvieron igual en términos de

significancia, salvo **colesterolemia**, la cual pasó a ser significativa. Ahora se elimina **packs1** dado que presenta el p-value más grande.

```
datos_dummy$EC <- as.factor(datos_dummy$EC)
logistica_2_2 <- glm(EC ~ Edad + PAS + PAD + Colesterolemia + Hombre + cig +
                     packs2, data = datos_dummy, family = "binomial")
summary(logistica_2_2)
```

```
##
## Call:
## glm(formula = EC ~ Edad + PAS + PAD + Colesterolemia + Hombre +
##     cig + packs2, family = "binomial", data = datos_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6639  -0.6924  -0.5283  -0.3497   2.4837
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.842607   1.030816  -8.578  < 2e-16 ***
## Edad            0.057750   0.015663   3.687 0.000227 ***
## PAS             0.015013   0.004018   3.736 0.000187 ***
## PAD             0.004893   0.007856   0.623 0.533421
## Colesterolemia  0.004614   0.001580   2.921 0.003492 **
## Hombre1         0.906043   0.164397   5.511 3.56e-08 ***
## cig             0.009510   0.009280   1.025 0.305501
## packs21         0.085153   0.336005   0.253 0.799938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.8  on 1361  degrees of freedom
## Residual deviance: 1239.3  on 1354  degrees of freedom
## AIC: 1255.3
##
## Number of Fisher Scoring iterations: 4
```

Las significancias se mantuvieron para las variables y sus coeficientes, por lo que se procede a eliminar la variable **packs2**, la cual presenta el p-value más grande.

```
datos_dummy$EC <- as.factor(datos_dummy$EC)
logistica_2_3 <- glm(EC ~ Edad + PAS + PAD + Colesterolemia + Hombre + cig,
                     data = datos_dummy, family = "binomial")
summary(logistica_2_3)
```

```
##
## Call:
## glm(formula = EC ~ Edad + PAS + PAD + Colesterolemia + Hombre +
##     cig, family = "binomial", data = datos_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.6594  -0.6934  -0.5286  -0.3501   2.4847
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.854237   1.030009  -8.596  < 2e-16 ***
## Edad            0.057782   0.015663   3.689 0.000225 ***
## PAS             0.015006   0.004018   3.735 0.000188 ***
## PAD             0.004955   0.007853   0.631 0.528020
## Colesterolemia  0.004618   0.001579   2.924 0.003454 **
## Hombre1         0.903372   0.164023   5.508 3.64e-08 ***
## cig             0.011229   0.006319   1.777 0.075563 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.8  on 1361  degrees of freedom
## Residual deviance: 1239.4  on 1355  degrees of freedom
## AIC: 1253.4
##
## Number of Fisher Scoring iterations: 4
```

Nuevamente se mantuvieron las significancias de las variables y sus coeficientes, por lo que se elimina la variable **PAD**.

```
datos_dummy$EC <- as.factor(datos_dummy$EC)
logistica_2_4 <- glm(EC ~ Edad + PAS + Colesterolemia + Hombre + cig,
                     data = datos_dummy, family = "binomial")
summary(logistica_2_4)
```

```
##
## Call:
## glm(formula = EC ~ Edad + PAS + Colesterolemia + Hombre + cig,
##     family = "binomial", data = datos_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6639  -0.6910  -0.5285  -0.3507   2.4924
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.637776   0.969724  -8.907  < 2e-16 ***
## Edad            0.056192   0.015455   3.636 0.000277 ***
## PAS             0.017009   0.002469   6.888 5.64e-12 ***
## Colesterolemia  0.004683   0.001577   2.971 0.002972 **
## Hombre1         0.910863   0.163688   5.565 2.63e-08 ***
## cig             0.011103   0.006318   1.757 0.078868 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.8  on 1361  degrees of freedom
```

```
## Residual deviance: 1239.8  on 1356  degrees of freedom
## AIC: 1251.8
##
## Number of Fisher Scoring iterations: 4
```

Se puede observar que todos los coeficientes para las variables del modelo dan significativos, salvo para la variable **cig**, por lo que se procede a eliminarla.

```
datos_dummy$EC <- as.factor(datos_dummy$EC)
logistica_2_5 <- glm(EC ~ Edad + PAS + Colesterolemia + Hombre,
                     data = datos_dummy, family = "binomial")
summary(logistica_2_5)
```

```
##
## Call:
## glm(formula = EC ~ Edad + PAS + Colesterolemia + Hombre, family = "binomial",
##     data = datos_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5400  -0.6947  -0.5271  -0.3492   2.4779
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.347166   0.950481  -8.782  < 2e-16 ***
## Edad            0.051574   0.015195   3.394 0.000689 ***
## PAS             0.016868   0.002459   6.859 6.93e-12 ***
## Colesterolemia  0.004762   0.001573   3.028 0.002461 **
## Hombre1         1.010859   0.152870   6.613 3.78e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1350.8  on 1361  degrees of freedom
## Residual deviance: 1242.9  on 1357  degrees of freedom
## AIC: 1252.9
##
## Number of Fisher Scoring iterations: 4
```

Finalmente, se llegó a un modelo donde todos los coeficientes asociados a las variables **Edad**, **PAS**, **Colesterolemia** y **Hombre** dan significativos (p-value menor a 0.05), con coeficientes iguales a 0.051574, 0.016868, 0004762 y 1.010859, respectivamente. Además, se obtiene una ordenada al origen significativa con un valor de -8.347166. El modelo reducido queda entonces definido por la siguiente ecuación:

$$logit(p) = \beta_0 + \beta_1 Edad + \beta_2 PAS + \beta_4 Colesterolemia + \beta_5 Hombre \tag{3}$$

$$logit(p) = -8.347166 + 0.051574 Edad + 0.016868 PAS + 0004762 Colesterolemia + 1.010859 Hombre \tag{4}$$

Luego, se obtienen los Odds Ratio y sus respectivos intervalos de confianza.

```
OR2 = exp(coefficients(logistica_2_5))
b = exp(confint(logistica_2_5))
OR2 = cbind(OR2,b)
OR2
```

```
##                         OR2        2.5 %      97.5 %
## (Intercept)    0.0002370675 3.569668e-05 0.001486371
## Edad           1.0529271105 1.022126e+00 1.084914391
## PAS            1.0170110804 1.012156e+00 1.021977984
## Colesterolemia 1.0047733370 1.001676e+00 1.007876632
## Hombre1        2.7479605402 2.042667e+00 3.721337052
```

Se puede observar en los resultados de OR que ninguna de las variables tienen OR iguales a 1 o intervalos que incluyan al 1, por lo que todas las variables no son independientes del outcome y son significativas. Para la **edad**, se puede decir que por cada año que aumenta la edad, el riesgo de sufrir una EC aumenta en un 5.29% (o aumenta 1.0529 veces). Para la **PAS**, se puede decir que por cada aumento unitario de PAS, el riesgo de sufrir una EC aumenta un 1.7% (o aumenta 1.017 veces). Para la variable **colesterolemia**, se puede decir que por cada aumento unitario de colesterolemia, el riesgo de sufrir una EC aumenta en un 0.477% (o 1.00477 veces). Para la variable **Hombre**, se puede decir que ser hombre aumenta el riesgo de sufrir una EC en un 174796% (o aumenta 2.74796 veces).

omnibus para ver si los coeficientes son globalmente significativos.

```
datos_dummy$EC <- as.numeric(datos_dummy$EC)
omnibus2 <- aov(logistica_2_5, data = datos_dummy)
summary(omnibus2)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## Edad               1   3.33   3.331  22.785 2.01e-06 ***
## PAS                1   6.53   6.533  44.681 3.38e-11 ***
## Colesterolemia     1   0.38   0.382   2.616    0.106
## Hombre             1   6.60   6.604  45.166 2.66e-11 ***
## Residuals       1357 198.41   0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

conclusion

Luego, se realiza la prueba de bondad de ajuste de Hosmer-Lemeshow, mencionada y utilizada en el modelo saturado (ver hipótesis del test)

```
prediccion2 <- predict(logistica_2_5, type="response")
hosmer2 <- hoslem.test(logistica_2_5$y, logistica_2_5$fitted.values) # Test Hosmwe-Lemeshow
hosmer2
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  logistica_2_5$y, logistica_2_5$fitted.values
## X-squared = 10.022, df = 8, p-value = 0.2635
```

Como el Test de bondad de ajuste de Hosmer-Lemeshow da un p-value (0.2635) mayor a 0.05, entonces la diferencia entre lo observado y lo esperado es igual a cero y, por ende, el modelo ajusta bien al conjunto de datos que se tiene.

Por último, para comparar ambos modelos, se utiliza el Test Likelihood ratio. Este test compara los valores de LogLik y analiza si la diferencia en estos valores de ajuste es significativa o no. La hipótesis nula indica que la diferencia es nula (no significativa), mientras que la alternativa indica que la diferencia no es nula y que es significativa.

```
lrtest(logistica_1, logistica_2_5)
```

```
## Likelihood ratio test
##
## Model 1: EC ~ Edad + PAS + PAD + Colesterolemia + Hombre + cig + packs0 +
##     packs1 + packs2
## Model 2: EC ~ Edad + PAS + Colesterolemia + Hombre
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -619.67
## 2   5 -621.43 -4 3.5102     0.4763
```

Como el Test Likelihood ratio da un p-value (0.4763) mayor a 0.05, se acepta la hipótesis nula. Por lo tanto, la diferencia entre los dos valores LogLik es nula y no significativa, lo cual indica que los dos modelos ajusten de igual manera (o de manera similar).

Sin embargo, se decide elegir el segundo modelo dado que todas las variables son significativas y presenta un LogLik más chico.

Ejercicio 7

En el análisis univariado se observó que la edad, la PAS, la PAD la colesterolemia y la cantidad de cigarrillos fumados por día presentan diferencias significativas al comparar los sujetos con y sin EC. Por otro lado, se observó que el sexo (ser hombre) y la categoría packs influyen en las ECs.

Luego, en el análisis multivariado se realizó un modelo de regresión logística saturado y se observó que las variables PAD, cig, packs0 y packs1 (variables dummy de packs) no son significativas para dicho modelo (p-values > 0.05), mientras que el resto de las variables sí lo es (p-values < 0.05). Además, se investigó las magnitudes de las relaciones existentes entre las variables significativas y el outcome mediante el cálculo de los Odds Ratio. Se observó que para la edad, por cada año que aumenta, el riesgo de sufrir una EC aumenta en un 6% (o aumenta 1.06 veces). Para la PAS, por cada aumento unitario de PAS, el riesgo de sufrir una EC aumenta un 1.5% (o aumenta 1.015 veces). Para la variable colesterolemia, por cada aumento unitario de colesterolemia, el riesgo de sufrir una EC aumenta en un 0.46% (o 1.0046 veces). Para la variable Hombre, el hecho de ser hombre aumenta el riesgo de sufrir una EC en un 147% (o aumenta 2.4747 veces). Además, gracias al Test de bondad de ajuste de Hosmer-Lemeshow, se probó que el modelo generado presenta un buen ajuste al conjunto de datos (p-value = 0.396).

Como segundo modelo, se propuso uno de regresión logística que contenga únicamente variables significativas. Para obtener este modelo, se realizó el backward elimination, lo que terminó en un modelo con las variables significativas edad, PAS, colesterolemia y Hombre (p-values < 0.05). Además, mediante el cálculo de los OR, se observó que las cuatro variables presentan una relación con el outcome (EC). Para la edad, por cada año que aumenta, el riesgo de sufrir una EC aumenta en un 5.29% (o aumenta 1.0529 veces). Para la PAS, por cada aumento unitario, el riesgo de sufrir una EC aumenta un 1.7% (o aumenta 1.017 veces). Para la variable colesterolemia, por cada aumento unitario, el riesgo de sufrir una EC aumenta en un 0.477% (o 1.00477 veces). Para la variable Hombre, el hecho de ser hombre aumenta el riesgo de sufrir una EC en un 175% (o aumenta 2.74796 veces). También se probó, mediante el Test de Hosmer-Lemeshow que el modelo generado presenta un buen ajuste a los datos (p-value = 0.2635).

Por último, mediante el Test Likelihood Ratio, se comprobó que no hay diferencia significativa entre los dos modelos en cuanto al ajuste a los datos (p-value = 0.4763). Sin embargo, se decidió elegir el segundo modelo

ya que todas las variables son significativas, tiene menor cantidad de variables y presenta un LogLik más chico (-621.43 contra -619.67), es decir, un mejor ajuste.