Memoria TFM UOC

Esteban Hernández Maldonado

22 de abril de 2021

Contents

Memoria (PEC 4).

1. Fase exploratoria

1.1 Lectura y preparación de los datos

El fichero de datos contiene 145 variables. Muchas de ellas se pueden omitir porque son las respuestas a cada de una de las preguntas de tres cuestionarios diferentes y en este estudio se utilizarán sólo las puntuaciones estandarizadas del cuestionario DMSES completo y el de cada uno de su cuatro apartados.

También se descartarán las variables que son la categorización de otras variables del fichero. Si alguna categorización o agrupación de valores fuera necesaria se creará en su debido momento.

Por otra lado se crearán variables nuevas:

- BMI: la variable continua BMI puede estar más relacionada con la diabetes que no las variables peso y altura (estas dos variables se descartarán una vez se haya calculado el BMI).
- sobrepeso: la variable categórica sobrepeso (1:Sí, 0:No) se calculará a partir del BMI (BMI>25: sobrepeso)
- obesidad: la variable categórica obesidad (1:Sí, 0:No) se calculará a partir del BMI (BMI>30: obesidad)
- alerta: la variable categórica alerta (1:Sí, 0:No) se calculará a partir de la variable hba1c (hba1c>7%). Se considera que un valor de hba1c > 7% indica que la diabetes no está siendo bien contralado por el paciente.

```
library(MASS)
library(dplyr)
library(ggplot2)
library(mgcv)
library(gnm)
library(GGally)
library(gmodels)
library(Hmisc)
library(caret)
library(ROCR)
library(gridExtra)
library(corrplot)
library(kableExtra)
library(faraway)
library(ggpubr)
library(klaR)


setwd("C:/Users/ehm24/Documents/master/TFM")

datos.raw <- read.csv("./datos/MontiFinal.csv")

# eliminamos las preguntas de los cuestionarios, categorizaciones de variables y
# otras variables no relevantes para el estudio
datos <- datos.raw[,-c(1,2,15,22,26:84,113:131,137:142,144:145)]

# definimos los factores

# Patient identifier
datos$id.code <- as.factor(datos$id.code)
# Hospital identifier
datos$hospid <- factor(datos$hospid,levels=c(1,2,3,4),
                       labels=c("phupaman","srinakarin","wegaronrat","chula"))
# Sex
datos$gender <- factor(datos$gender,levels=c(1,2),
                       labels=c("Hombre","Mujer"))
# Marital status
```

```r
datos$mstatus <- factor(datos$mstatus,levels=c(1,2,3,4,5),
                        labels=c("Soltero","´Casado",
                                 "Viudo","Divorciado","Separado"))
# Education level
datos$edu <- factor(datos$edu,
                    levels=c(1,2,3,4,5,6),
                    labels=c("Sin estudios" , "Primarios", "Secundarios",
                             "Grado", "Master", "Grado Superior"))
# Religion
datos$religion <- factor(datos$religion,levels=c(1,2,3,4),
                         labels=c("Budismo", "Islam", "Cristianismo","Otra religión"))
# Income
datos$income <- factor(datos$income, levels=c(1,2,3,4,5,6),
                       labels=c("menos de 4,999 bath", "5,000-9,999  bath",
                                "10,000-14,999 bath", "15,000-19,999 bath",
                                "20,000-24,999 bath", "más de 25,000 bath"))
# Family history of T2DM
datos$famhx <- factor(datos$famhx,levels=c(0,1),labels=c("No", "Sí"))
# Comorbid disease
datos$comob <- factor(datos$comob,levels=c(0,1),labels=c("No", "Sí"))
# Comorbidity lipidemia
datos$comlip <- factor(datos$comlip,levels=c(0,1),labels=c("No", "Sí"))
# Comorbidity hypertension
datos$comht <- factor(datos$comht,levels=c(0,1),labels=c("No", "Sí"))
# Comorbidity coronary heart disease
datos$comchd <- factor(datos$comchd,levels=c(0,1),labels=c("No", "Sí"))
# Comorbidity kidney disease
datos$comkid <- factor(datos$comkid,levels=c(0,1),labels=c("No", "Sí"))
# Other comorbidity
datos$comoth <- factor(datos$comoth,levels=c(0,1),labels=c("No", "Sí"))
# Type of Treatment of DM
datos$dmrx <- factor(datos$dmrx,
                     levels=c(1,2,3,4),
                     labels=c("Sin medicación","Hipoglicémico oral",
                              "Insulina","Ambos"))
# Smoking
datos$smk <- factor(datos$smk,levels=c(1,2,3),
                    labels=c("Sí","Exfumador","No"))
# Alcohol
datos$alcohol <- factor(datos$alcohol,levels=c(1,2,3),
                        labels=c("Sí","Exbebedor","No"))
# Diabetes Complication
datos$compli <- factor(datos$compli,levels=c(0,1),labels=c("No", "Sí"))
# Cerebrovascular Accident
datos$cva <- factor(datos$cva,levels=c(0,1),labels=c("No", "Sí"))
# Cerebral Infraction
datos$cereinfrac <- factor(datos$cereinfrac,levels=c(0,1),labels=c("No", "Sí"))
# Ischemic Stroke
datos$ishemic <- factor(datos$ishemic,levels=c(0,1),labels=c("No", "Sí"))
# Stroke, Not specify
datos$stroke <- factor(datos$stroke,levels=c(0,1),labels=c("No", "Sí"))
# Cerebral Hemorrhage
datos$cerebhem <- factor(datos$cerebhem,levels=c(0,1),labels=c("No", "Sí"))
```

```r
# Transient Ischemic Attack
datos$tia <- factor(datos$tia,levels=c(0,1),labels=c("No", "Sí"))
# Angina pectoris
datos$angia <- factor(datos$angia,levels=c(0,1),labels=c("No", "Sí"))
# Congestive Heart Failure
datos$chf <- factor(datos$chf,levels=c(0,1),labels=c("No", "Sí"))
# Myocardial Infraction
datos$mi <- factor(datos$mi,levels=c(0,1),labels=c("No", "Sí"))
# Coronary Revascularization
datos$cororevas <- factor(datos$cororevas,levels=c(0,1),labels=c("No", "Sí"))
# Peripheral Arterial Disease
datos$pad <- factor(datos$pad,levels=c(0,1),labels=c("No", "Sí"))
# Neuropathy
datos$neuropath <- factor(datos$neuropath,levels=c(0,1),labels=c("No", "Sí"))
# Renal Insufficiency
datos$renal <- factor(datos$renal,levels=c(0,1),labels=c("No", "Sí"))
# Diabetes Nephropathy
datos$dn <- factor(datos$dn,levels=c(0,1),labels=c("No", "Sí"))
# Diabetes Retinopathy
datos$dr <- factor(datos$dr,levels=c(0,1),labels=c("No", "Sí"))
# Other complication
datos$othcomp <- factor(datos$othcomp,levels=c(0,1),labels=c("No", "Sí"))


# creación de variables derivadas

# alerta: 1-Sí si hba1c > 7% (diabetes no controlada)
datos$alerta <- factor(ifelse(datos$hba1c > 7,1,0),
                       levels=c(0,1),
                       labels=c("No", "Sí"))
# BMI
datos$bmi <- datos$weight/(datos$height/100)**2
# sobrepeso, BMI >25
datos$sobrepeso <- factor(ifelse(datos$bmi > 25,1,0),
                          levels=c(0,1),labels=c("No", "Sí"))
# obesidad, BMI > 30
datos$obesidad <- factor(ifelse(datos$bmi > 30,1,0),
                         levels=c(0,1),labels=c("No", "Sí"))

# transformación a variable tipo date de todas las variables fecha
# estas variables son útiles para poder determinar que datos tienen
# información actualizada
datos$fecha.hba1c <- as.Date(datos$hba1cdate,"%d/%m/%y")
datos$fecha.ldl <- as.Date(datos$ldldate,"%d/%m/%y")
datos$fecha.hdl <- as.Date(datos$hdldate,"%d/%m/%y")
datos$fecha.trig <- as.Date(datos$trigdate,"%d/%m/%y")
datos$fecha.bp <- as.Date(datos$bpdate,"%d/%m/%y")

# eliminación de variables no necesarias
drop.cols <- c("weight","height","hba1cdate","ldldate","hdldate","trigdate","bpdate")
datos <- datos %>% select(-all_of(drop.cols))
```

1.1.1 Lectura y creación de variables derivadas

1.1.2 Datos faltantes Se tiene sólo datos faltantes en dos variables tipo fecha (fecha.ldl y fecha.bp) y en sólo dos pacientes.

```
# datos faltantes
datos[!complete.cases(datos),]
```

```
##     id.code     hospid gender mstatus age         edu religion
## 225     225 wegaronrat Hombre ´Casado  56 Secundarios    Islam
## 352     352 wegaronrat Hombre   Viudo  70   Primarios  Budismo
##                  income dmdura famhx comob comlip comht comchd comkid comoth
## 225 menos de 4,999 bath      4    No    Sí     Sí    Sí     No     No     No
## 352 menos de 4,999 bath      5    No    Sí     Sí    Sí     No     No     No
##                    dmrx       smk   alcohol hba1c ldl hdl trig sbp dbp compli
## 225 Hipoglicémico oral        No        No   6.0  99  68  111 137  95     No
## 352              Ambos Exfumador Exbebedor   6.7  76  46   72 133  22     Sí
##     cva cereinfrac ishemic stroke cerebhem tia angia chf mi cororevas pad
## 225  No         No      No     No       No  No    No  No No        No  No
## 352  No         No      Sí     No       No  No    No  No No        No  No
##     neuropath renal dn dr othcomp DMSES.Diet DMSES.Monitor DMSES.Physical
## 225        No    No No No      No  0.3336945     0.1225384     -0.6551139
## 352        No    No No No      No  0.1123347     0.3966739      0.3133142
##     DMSES.Regimen DMSES.Total    DK.10 alerta      bmi sobrepeso obesidad
## 225    -0.5854147  -0.7842956 4.685086     No 33.62209        Sí       Sí
## 352     0.1183689   0.9406917 4.550101     No 22.65625        No       No
##     fecha.hba1c  fecha.ldl  fecha.hdl fecha.trig   fecha.bp
## 225  2016-02-15       <NA> 2015-06-18 2015-06-18 2016-05-16
## 352  2016-01-15 2016-05-13 2016-05-23 2016-05-23       <NA>
```

1.1.3 Outliers Las variables continuas son:
age
dmdura
hba1c
ldl
hdl
trig
sbp
dbp
DMSES.Diet
DMSES.Monitor
DMSES.Physical
DMSES.Regimen
DMSES.Total
DK.10

```
flag_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  z <-as.factor(as.integer(is.na(y)))
  return(z)
}
```

```r
datos$age_out <- flag_outliers(datos$age)
datos$dmdura_out <- flag_outliers(datos$dmdura)
datos$hba1c_out <- flag_outliers(datos$hba1c)
datos$ldl_out <- flag_outliers(datos$ldl)
datos$hdl_out <- flag_outliers(datos$hdl)
datos$trig_out <- flag_outliers(datos$trig)
datos$sbp_out <- flag_outliers(datos$sbp)
datos$dbp_out <- flag_outliers(datos$dbp)
datos$DMSES.Diet_out <- flag_outliers(datos$DMSES.Diet)
datos$DMSES.Monitor_out <- flag_outliers(datos$DMSES.Monitor)
datos$DMSES.Physical_out <- flag_outliers(datos$DMSES.Physical)
datos$DMSES.Regimen_out <- flag_outliers(datos$DMSES.Regimen)
datos$DMSES.Total_out <- flag_outliers(datos$DMSES.Total)
datos$DK.10_out <- flag_outliers(datos$DK.10)
datos$bmi_out <- flag_outliers(datos$bmi)

summary(datos)
```

```
##     id.code           hospid       gender          mstatus         age
##  1      :  1   phupaman  : 60   Hombre:208   Soltero   : 57   Min.   :26.00
##  2      :  1   srinakarin: 77   Mujer :492   ´Casado   :462   1st Qu.:59.00
##  3      :  1   wegaronrat:241                Viudo     :165   Median :65.00
##  4      :  1   chula     :322                Divorciado: 13   Mean   :65.16
##  5      :  1                                 Separado  :  3   3rd Qu.:73.00
##  6      :  1                                                  Max.   :95.00
##  (Other):694
##              edu               religion                   income
##  Sin estudios  : 47   Budismo      :543   menos de 4,999 bath:318
##  Primarios     :381   Islam        :152   5,000-9,999  bath  : 95
##  Secundarios   :146   Cristianismo :  5   10,000-14,999 bath : 86
##  Grado         : 99   Otra religión:  0   15,000-19,999 bath : 48
##  Master        : 25                       20,000-24,999 bath : 48
##  Grado Superior:  2                       más de 25,000 bath :105
##
##      dmdura      famhx    comob    comlip   comht    comchd   comkid   comoth
##  Min.   : 4.00   No:370   No: 42   No: 95   No: 97   No:662   No:512   No:699
##  1st Qu.: 7.00   Sí:330   Sí:658   Sí:605   Sí:603   Sí: 38   Sí:188   Sí:  1
##  Median :10.00
##  Mean   :13.53
##  3rd Qu.:20.00
##  Max.   :45.00
##
##                  dmrx            smk           alcohol        hba1c
##  Sin medicación    : 12   Sí       : 23   Sí       : 42   Min.   : 2.000
##  Hipoglicémico oral:409   Exfumador: 88   Exbebedor: 89   1st Qu.: 6.400
##  Insulina          : 94   No       :589   No       :569   Median : 7.100
##  Ambos             :185                                   Mean   : 7.579
##                                                           3rd Qu.: 8.300
##                                                           Max.   :15.000
##
##       ldl              hdl              trig             sbp
##  Min.   :  8.7   Min.   : 17.00   Min.   : 29.30   Min.   : 93.0
##  1st Qu.: 81.0   1st Qu.: 41.88   1st Qu.: 95.75   1st Qu.:123.0
```

```
##  Median : 97.0   Median : 50.00   Median :127.00   Median :133.0
##  Mean   :100.8   Mean   : 50.75   Mean   :149.02   Mean   :134.8
##  3rd Qu.:116.0   3rd Qu.: 58.00   3rd Qu.:175.00   3rd Qu.:145.0
##  Max.   :221.0   Max.   :116.00   Max.   :999.00   Max.   :217.0
##
##       dbp          compli    cva       cereinfrac ishemic  stroke   cerebhem
##  Min.   : 22.00   No:429   No:699   No:700     No:688   No:698   No:700
##  1st Qu.: 65.00   Sí:271   Sí:  1   Sí:  0     Sí: 12   Sí:  2   Sí:  0
##  Median : 73.00
##  Mean   : 73.02
##  3rd Qu.: 80.00
##  Max.   :112.00
##
##  tia       angia     chf       mi       cororevas pad      neuropath renal
##  No:699   No:700   No:692   No:675   No:699    No:696   No:695    No:582
##  Sí:  1   Sí:  0   Sí:  8   Sí: 25   Sí:  1    Sí:  4   Sí:  5    Sí:118
##
##
##
##
##
##   dn        dr       othcomp    DMSES.Diet        DMSES.Monitor
##  No:619   No:561   No:699   Min.   :-1.45946   Min.   :-0.84225
##  Sí: 81   Sí:139   Sí:  1   1st Qu.:-0.53693   1st Qu.:-0.23065
##                             Median :-0.06537   Median :-0.03718
##                             Mean   : 0.00000   Mean   : 0.00000
##                             3rd Qu.: 0.52029   3rd Qu.: 0.24502
##                             Max.   : 1.03434   Max.   : 0.55056
##
##  DMSES.Physical     DMSES.Regimen       DMSES.Total          DK.10
##  Min.   :-1.28709   Min.   :-2.86027   Min.   :-5.6289   Min.   : 0.000
##  1st Qu.:-0.28906   1st Qu.: 0.07856   1st Qu.:-1.0268   1st Qu.: 4.224
##  Median : 0.01757   Median : 0.09450   Median :-0.0258   Median : 5.486
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 5.535
##  3rd Qu.: 0.29446   3rd Qu.: 0.10978   3rd Qu.: 1.0416   3rd Qu.: 6.919
##  Max.   : 0.79336   Max.   : 0.51344   Max.   : 2.5059   Max.   :10.000
##
##  alerta        bmi        sobrepeso obesidad  fecha.hba1c
##  No:333   Min.   :14.31   No:301    No:518   Min.   :2006-02-01
##  Sí:367   1st Qu.:22.91   Sí:399    Sí:182   1st Qu.:2016-03-01
##           Median :25.96                      Median :2016-05-04
##           Mean   :27.08                      Mean   :2016-03-07
##           3rd Qu.:30.14                      3rd Qu.:2016-06-07
##           Max.   :76.03                      Max.   :2016-07-04
##
##    fecha.ldl            fecha.hdl            fecha.trig
##  Min.   :2012-11-08   Min.   :2006-04-23   Min.   :2012-11-08
##  1st Qu.:2016-02-04   1st Qu.:2016-01-13   1st Qu.:2016-02-02
##  Median :2016-04-21   Median :2016-03-29   Median :2016-04-21
##  Mean   :2016-02-19   Mean   :2016-01-26   Mean   :2016-02-27
##  3rd Qu.:2016-05-31   3rd Qu.:2016-05-31   3rd Qu.:2016-05-31
##  Max.   :2016-09-03   Max.   :2016-12-09   Max.   :2016-12-21
##  NA's   :1
##     fecha.bp          age_out dmdura_out hba1c_out ldl_out hdl_out trig_out
```

```
##  Min.   :2006-04-27   0:692   0:692      0:666     0:674   0:683   0:665
##  1st Qu.:2016-04-22   1:  8   1:  8      1: 34     1: 26   1: 17   1: 35
##  Median :2016-05-19
##  Mean   :2016-05-14
##  3rd Qu.:2016-06-16
##  Max.   :2035-03-27
##  NA's   :1
##  sbp_out dbp_out DMSES.Diet_out DMSES.Monitor_out DMSES.Physical_out
##  0:684   0:687   0:700          0:700             0:696
##  1: 16   1: 13                                    1:  4
##
##
##
##
##
##  DMSES.Regimen_out DMSES.Total_out DK.10_out bmi_out
##  0:578             0:698           0:694     0:686
##  1:122             1:  2           1:  6     1: 14
##
##
##
##
##
##
```

```
# flags para outliers
vars_out <- c("age_out","dmdura_out","hba1c_out","ldl_out","hdl_out","trig_out","sbp_out","dbp_out",
              "DMSES.Diet_out","DMSES.Monitor_out","DMSES.Physical_out","DMSES.Regimen_out",
              "DMSES.Total_out","DK.10_out")

# la variable datos$out valdrá uno si alguna de las variables continuas es outlier, 0 en caso contrario
datos$out <- apply(datos[,vars_out],1,max)


# variables continuas
vars_cont <- c("age","dmdura","hba1c","ldl","hdl","trig","sbp","dbp",
              "DMSES.Diet","DMSES.Monitor","DMSES.Physical","DMSES.Regimen",
              "DMSES.Total","DK.10")

# variables categóricas
vars_cat <- c("id.code","hospid","gender","mstatus","edu","religion","income",
              "famhx","comob","comlip","comht","comchd","comkid","comoth","dmrx",
              "smk","alcohol","compli","cva","cereinfrac","ishemic","stroke","cerebhem",
              "tia","angia","chf","mi","cororevas","pad","neuropath","renal","dn",
              "dr","othcomp","alerta","sobrepeso","obesidad",vars_out,"out")

# otras variables
vars_otras <- c("fecha.hba1c","fecha.ldl","fecha.hdl","fecha.trig","fecha.bp")

# todas las variables
vars <- c(vars_cat,vars_cont,vars_otras)

# pacientes con alguna variable continua outlier
pat_out <- datos %>% filter(out==1)
```

```r
dim(pat_out)
```

```
## [1] 234  73
```

```r
dim(datos)
```

```
## [1] 700  73
```

```r
attach(datos)
```

La estructura del fichero inicial, ya con todas las variables derivadas creadas es la que sigue:

```r
# descripcion del fichero con el que se trabajará
str(datos)
```

```
## 'data.frame':    700 obs. of  73 variables:
##  $ id.code          : Factor w/ 700 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ hospid           : Factor w/ 4 levels "phupaman","srinakarin",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ gender           : Factor w/ 2 levels "Hombre","Mujer": 1 2 1 2 1 2 1 2 2 2 ...
##  $ mstatus          : Factor w/ 5 levels "Soltero","´Casado",..: 2 2 2 3 3 2 2 2 2 2 ...
##  $ age              : int  51 55 47 54 71 55 59 64 58 69 ...
##  $ edu              : Factor w/ 6 levels "Sin estudios",..: 4 2 2 2 2 2 2 2 2 1 ...
##  $ religion         : Factor w/ 4 levels "Budismo","Islam",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ income           : Factor w/ 6 levels "menos de 4,999 bath",..: 5 2 1 1 1 1 6 1 1 1 ...
##  $ dmdura           : int  4 4 5 20 10 4 4 22 4 14 ...
##  $ famhx            : Factor w/ 2 levels "No","Sí": 2 2 1 1 1 1 1 1 2 1 ...
##  $ comob            : Factor w/ 2 levels "No","Sí": 2 2 2 2 2 2 2 2 2 1 2 ...
##  $ comlip           : Factor w/ 2 levels "No","Sí": 2 2 2 2 1 2 2 2 1 2 ...
##  $ comht            : Factor w/ 2 levels "No","Sí": 2 1 2 2 2 1 2 2 1 2 ...
##  $ comchd           : Factor w/ 2 levels "No","Sí": 1 1 1 1 2 1 1 1 1 1 ...
##  $ comkid           : Factor w/ 2 levels "No","Sí": 1 1 2 1 2 1 1 1 1 1 ...
##  $ comoth           : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ dmrx             : Factor w/ 4 levels "Sin medicación",..: 2 2 2 4 3 2 2 2 2 4 ...
##  $ smk              : Factor w/ 3 levels "Sí","Exfumador",..: 2 3 2 3 2 3 3 3 3 3 ...
##  $ alcohol          : Factor w/ 3 levels "Sí","Exbebedor",..: 2 3 2 3 1 3 2 3 3 3 ...
##  $ hba1c            : num  9.7 7.1 8 7.2 8.8 8 9.8 6.3 6.7 7.7 ...
##  $ ldl              : num  123.7 89.8 119 202.1 58.6 ...
##  $ hdl              : num  45.2 49.5 50.7 47 35.7 35.3 42 44.9 46 35.6 ...
##  $ trig             : num  248.9 276.9 165.6 84.7 149.5 ...
##  $ sbp              : int  130 120 168 121 118 119 119 150 124 169 ...
##  $ dbp              : int  80 70 94 56 65 63 83 76 76 93 ...
##  $ compli           : Factor w/ 2 levels "No","Sí": 1 1 2 1 2 1 1 1 1 1 ...
##  $ cva              : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cereinfrac       : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ishemic          : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ stroke           : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cerebhem         : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tia              : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ angia            : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ chf              : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mi               : Factor w/ 2 levels "No","Sí": 1 1 1 1 2 1 1 1 1 1 ...
```

```
##  $ cororevas         : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pad               : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ neuropath         : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ renal             : Factor w/ 2 levels "No","Sí": 1 1 2 1 2 1 1 1 1 1 ...
##  $ dn                : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ dr                : Factor w/ 2 levels "No","Sí": 1 1 2 1 1 1 1 1 1 1 ...
##  $ othcomp           : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DMSES.Diet        : num  -0.132233 -0.000414 1.031174 -1.102578 -1.064826 ...
##  $ DMSES.Monitor     : num  -0.133 0.158 0.527 -0.168 -0.37 ...
##  $ DMSES.Physical    : num  -0.1775 0.09218 0.79015 -0.00571 0.02158 ...
##  $ DMSES.Regimen     : num  0.0909 0.106 -0.0232 0.0916 0.0824 ...
##  $ DMSES.Total       : num  -0.352 0.355 2.325 -1.184 -1.331 ...
##  $ DK.10             : num  5.77 6.66 4.27 5.19 5.65 ...
##  $ alerta            : Factor w/ 2 levels "No","Sí": 2 2 2 2 2 2 2 1 1 2 ...
##  $ bmi               : num  23 25.6 22.6 25.7 21.1 ...
##  $ sobrepeso         : Factor w/ 2 levels "No","Sí": 1 2 1 2 1 1 2 1 2 1 ...
##  $ obesidad          : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fecha.hba1c       : Date, format: "2014-07-10" "2014-07-26" ...
##  $ fecha.ldl         : Date, format: "2015-11-06" "2015-02-16" ...
##  $ fecha.hdl         : Date, format: "2015-11-06" "2015-02-16" ...
##  $ fecha.trig        : Date, format: "2015-11-06" "2015-02-16" ...
##  $ fecha.bp          : Date, format: "2016-02-19" "2016-02-24" ...
##  $ age_out           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ dmdura_out        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ hba1c_out         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ldl_out           : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ hdl_out           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ trig_out          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sbp_out           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ dbp_out           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DMSES.Diet_out    : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DMSES.Monitor_out : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DMSES.Physical_out: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DMSES.Regimen_out : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...
##  $ DMSES.Total_out   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DK.10_out         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ bmi_out           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ out               : chr  "0" "0" "1" "1" ...
```

1.2 Exploración unidimensional

Una primera vista a las variables de nuestro fichero:

```
summary(datos[,-c(1)])
```

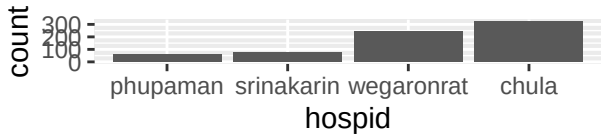
```
##         hospid       gender          mstatus         age        
##  phupaman  : 60   Hombre:208   Soltero    : 57   Min.   :26.00  
##  srinakarin: 77   Mujer :492   ´Casado    :462   1st Qu.:59.00  
##  wegaronrat:241                Viudo      :165   Median :65.00  
##  chula     :322                Divorciado: 13   Mean   :65.16  
##                                Separado   :  3   3rd Qu.:73.00  
##                                                  Max.   :95.00  
##
```

```
##              edu               religion                   income
##  Sin estudios  : 47   Budismo      :543   menos de 4,999 bath:318
##  Primarios     :381   Islam        :152   5,000-9,999  bath  : 95
##  Secundarios   :146   Cristianismo :  5   10,000-14,999 bath : 86
##  Grado         : 99   Otra religión:  0   15,000-19,999 bath : 48
##  Master        : 25                       20,000-24,999 bath : 48
##  Grado Superior:  2                       más de 25,000 bath :105
##
##      dmdura      famhx    comob    comlip   comht    comchd   comkid   comoth
##  Min.   : 4.00   No:370   No: 42   No: 95   No: 97   No:662   No:512   No:699
##  1st Qu.: 7.00   Sí:330   Sí:658   Sí:605   Sí:603   Sí: 38   Sí:188   Sí:  1
##  Median :10.00
##  Mean   :13.53
##  3rd Qu.:20.00
##  Max.   :45.00
##
##                  dmrx            smk           alcohol        hba1c
##  Sin medicación    : 12   Sí       : 23   Sí       : 42   Min.   : 2.000
##  Hipoglicémico oral:409   Exfumador: 88   Exbebedor: 89   1st Qu.: 6.400
##  Insulina          : 94   No       :589   No       :569   Median : 7.100
##  Ambos             :185                                   Mean   : 7.579
##                                                           3rd Qu.: 8.300
##                                                           Max.   :15.000
##
##       ldl              hdl              trig             sbp
##  Min.   :  8.7   Min.   : 17.00   Min.   : 29.30   Min.   : 93.0
##  1st Qu.: 81.0   1st Qu.: 41.88   1st Qu.: 95.75   1st Qu.:123.0
##  Median : 97.0   Median : 50.00   Median :127.00   Median :133.0
##  Mean   :100.8   Mean   : 50.75   Mean   :149.02   Mean   :134.8
##  3rd Qu.:116.0   3rd Qu.: 58.00   3rd Qu.:175.00   3rd Qu.:145.0
##  Max.   :221.0   Max.   :116.00   Max.   :999.00   Max.   :217.0
##
##       dbp          compli   cva      cereinfrac ishemic  stroke   cerebhem
##  Min.   : 22.00   No:429   No:699   No:700     No:688   No:698   No:700
##  1st Qu.: 65.00   Sí:271   Sí:  1   Sí:  0     Sí: 12   Sí:  2   Sí:  0
##  Median : 73.00
##  Mean   : 73.02
##  3rd Qu.: 80.00
##  Max.   :112.00
##
##  tia      angia    chf       mi       cororevas pad      neuropath renal
##  No:699   No:700   No:692   No:675   No:699    No:696   No:695    No:582
##  Sí:  1   Sí:  0   Sí:  8   Sí: 25   Sí:  1    Sí:  4   Sí:  5    Sí:118
##
##
##
##
##
##   dn       dr      othcomp    DMSES.Diet       DMSES.Monitor
##  No:619   No:561   No:699   Min.   :-1.45946   Min.   :-0.84225
##  Sí: 81   Sí:139   Sí:  1   1st Qu.:-0.53693   1st Qu.:-0.23065
##                             Median :-0.06537   Median :-0.03718
##                             Mean   : 0.00000   Mean   : 0.00000
##                             3rd Qu.: 0.52029   3rd Qu.: 0.24502
```

```
##                             Max.   : 1.03434   Max.   : 0.55056
##
##  DMSES.Physical     DMSES.Regimen       DMSES.Total          DK.10
##  Min.   :-1.28709   Min.   :-2.86027   Min.   :-5.6289   Min.   : 0.000
##  1st Qu.:-0.28906   1st Qu.: 0.07856   1st Qu.:-1.0268   1st Qu.: 4.224
##  Median : 0.01757   Median : 0.09450   Median :-0.0258   Median : 5.486
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 5.535
##  3rd Qu.: 0.29446   3rd Qu.: 0.10978   3rd Qu.: 1.0416   3rd Qu.: 6.919
##  Max.   : 0.79336   Max.   : 0.51344   Max.   : 2.5059   Max.   :10.000
##
##  alerta        bmi        sobrepeso obesidad  fecha.hba1c
##  No:333   Min.   :14.31   No:301    No:518   Min.   :2006-02-01
##  Sí:367   1st Qu.:22.91   Sí:399    Sí:182   1st Qu.:2016-03-01
##           Median :25.96                      Median :2016-05-04
##           Mean   :27.08                      Mean   :2016-03-07
##           3rd Qu.:30.14                      3rd Qu.:2016-06-07
##           Max.   :76.03                      Max.   :2016-07-04
##
##    fecha.ldl            fecha.hdl            fecha.trig
##  Min.   :2012-11-08   Min.   :2006-04-23   Min.   :2012-11-08
##  1st Qu.:2016-02-04   1st Qu.:2016-01-13   1st Qu.:2016-02-02
##  Median :2016-04-21   Median :2016-03-29   Median :2016-04-21
##  Mean   :2016-02-19   Mean   :2016-01-26   Mean   :2016-02-27
##  3rd Qu.:2016-05-31   3rd Qu.:2016-05-31   3rd Qu.:2016-05-31
##  Max.   :2016-09-03   Max.   :2016-12-09   Max.   :2016-12-21
##  NA's   :1
##     fecha.bp          age_out dmdura_out hba1c_out ldl_out hdl_out trig_out
##  Min.   :2006-04-27   0:692   0:692      0:666     0:674   0:683   0:665
##  1st Qu.:2016-04-22   1:  8   1:  8      1: 34     1: 26   1: 17   1: 35
##  Median :2016-05-19
##  Mean   :2016-05-14
##  3rd Qu.:2016-06-16
##  Max.   :2035-03-27
##  NA's   :1
##  sbp_out dbp_out DMSES.Diet_out DMSES.Monitor_out DMSES.Physical_out
##  0:684   0:687   0:700          0:700             0:696
##  1: 16   1: 13                                    1:  4
##
##
##
##
##
##  DMSES.Regimen_out DMSES.Total_out DK.10_out bmi_out     out
##  0:578             0:698           0:694     0:686   Length:700
##  1:122             1:  2           1:  6     1: 14   Class :character
##                                                      Mode  :character
##
##
##
##
```

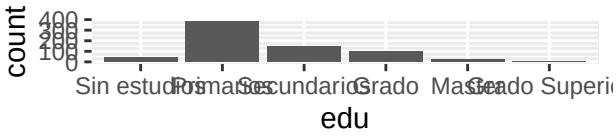
Los únicos valores faltantes se dan en las variables fecha.ldl y fecha.bp, un único caso para cada variable. De momento se decide no eliminar los registros.

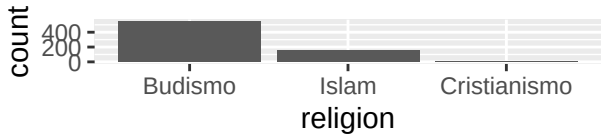
Analizamos con un poco más de detalle las variables categóricas

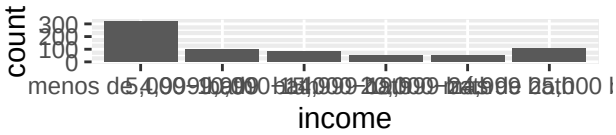

count
phupaman
srinakarin
wegaronrat
chula
hospid

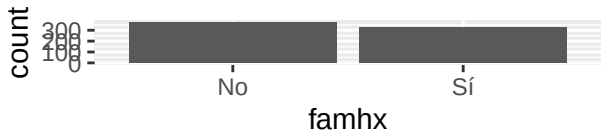

count
Hombre
Mujer
gender

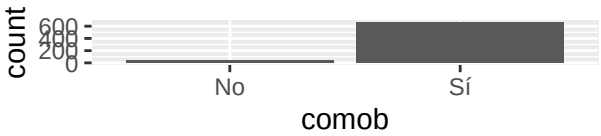

count
Soltero
´Casado
Viudo
Divorciado
Separado
mstatus

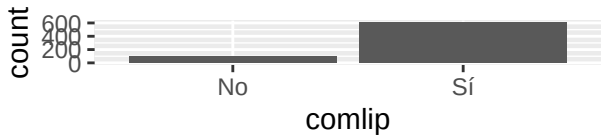

count
Sin estudios
Primarios
Secundarios
Grado
Master
edu

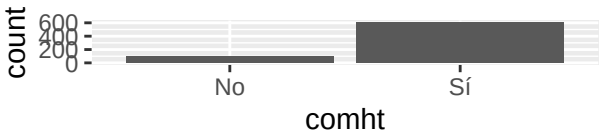

count
Budismo
Islam
Cristianismo
religion

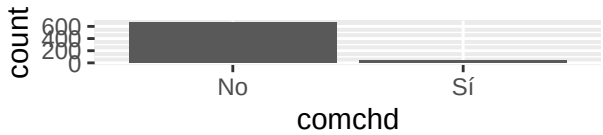

count
income

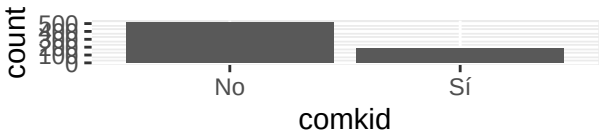

count
No
Sí
famhx

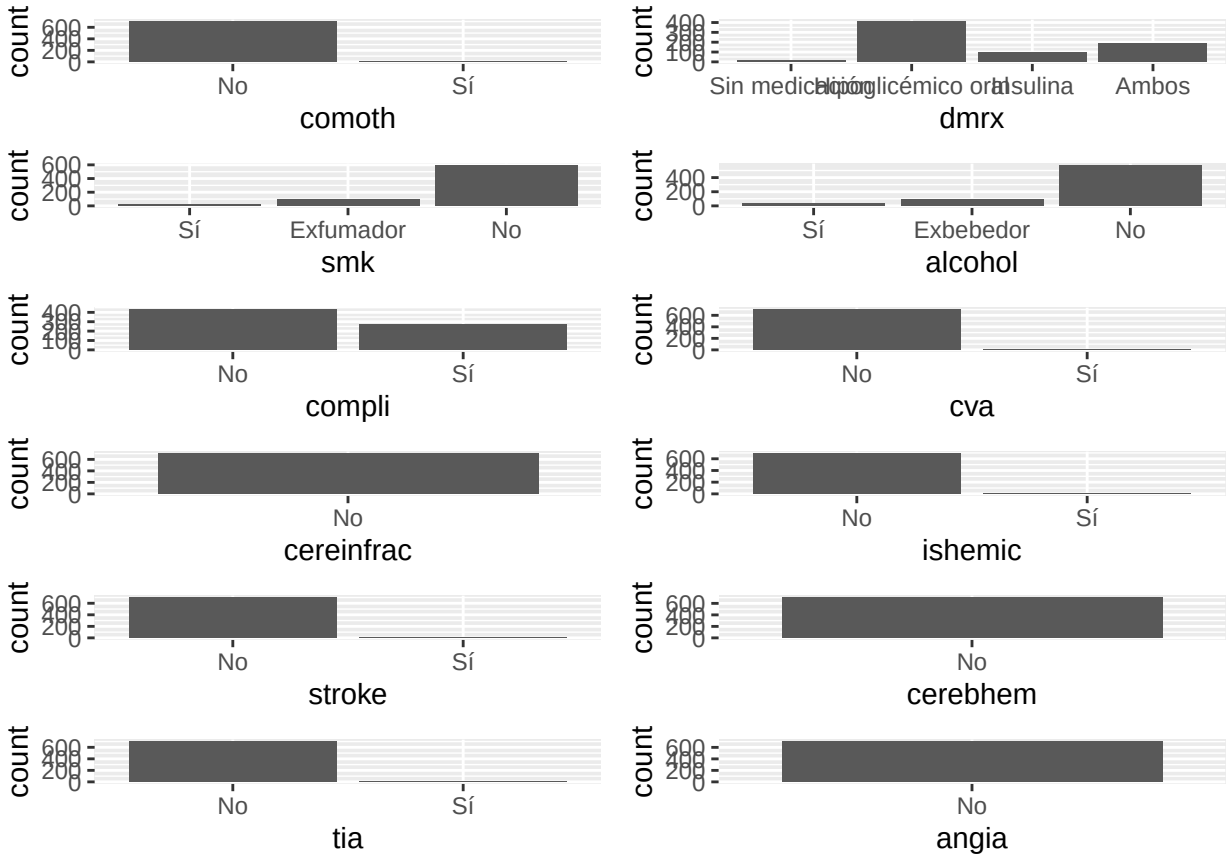

count
No
Sí
comob

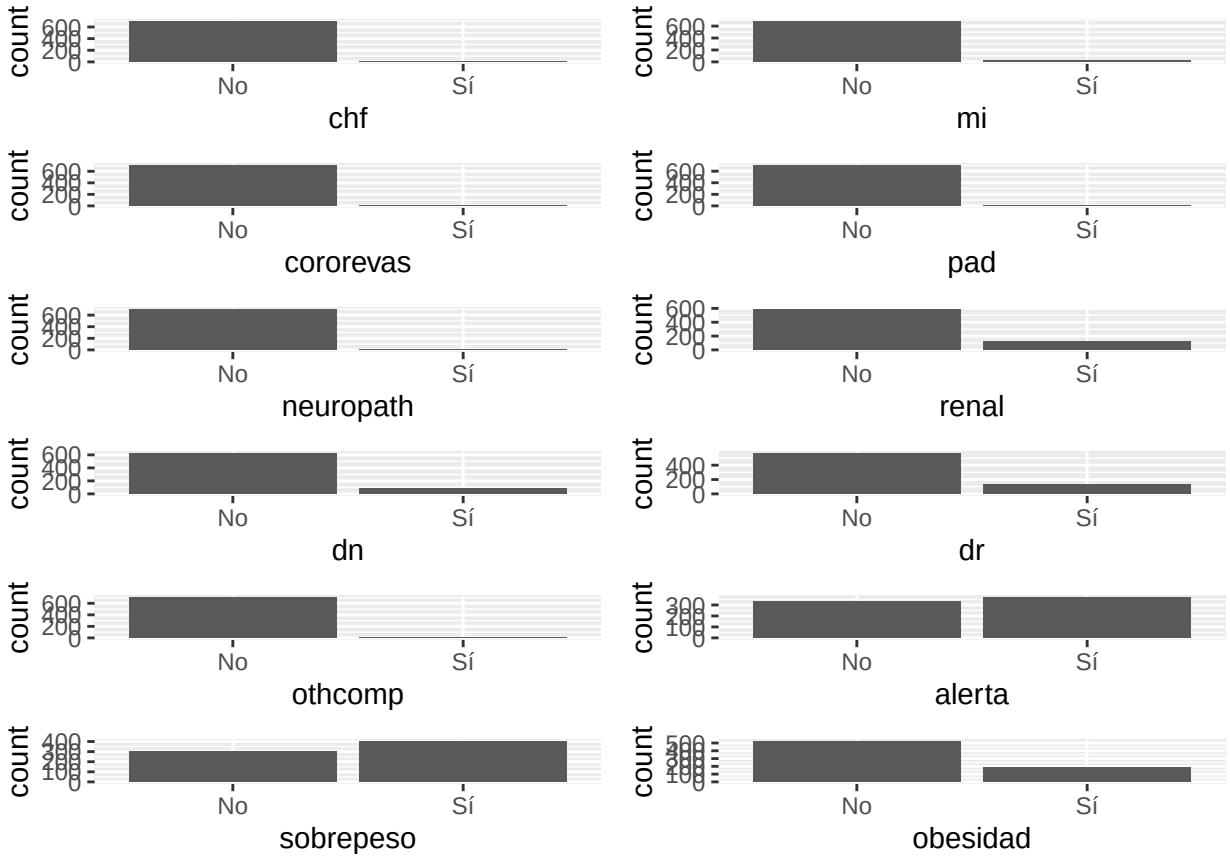

count
No
Sí
comlip

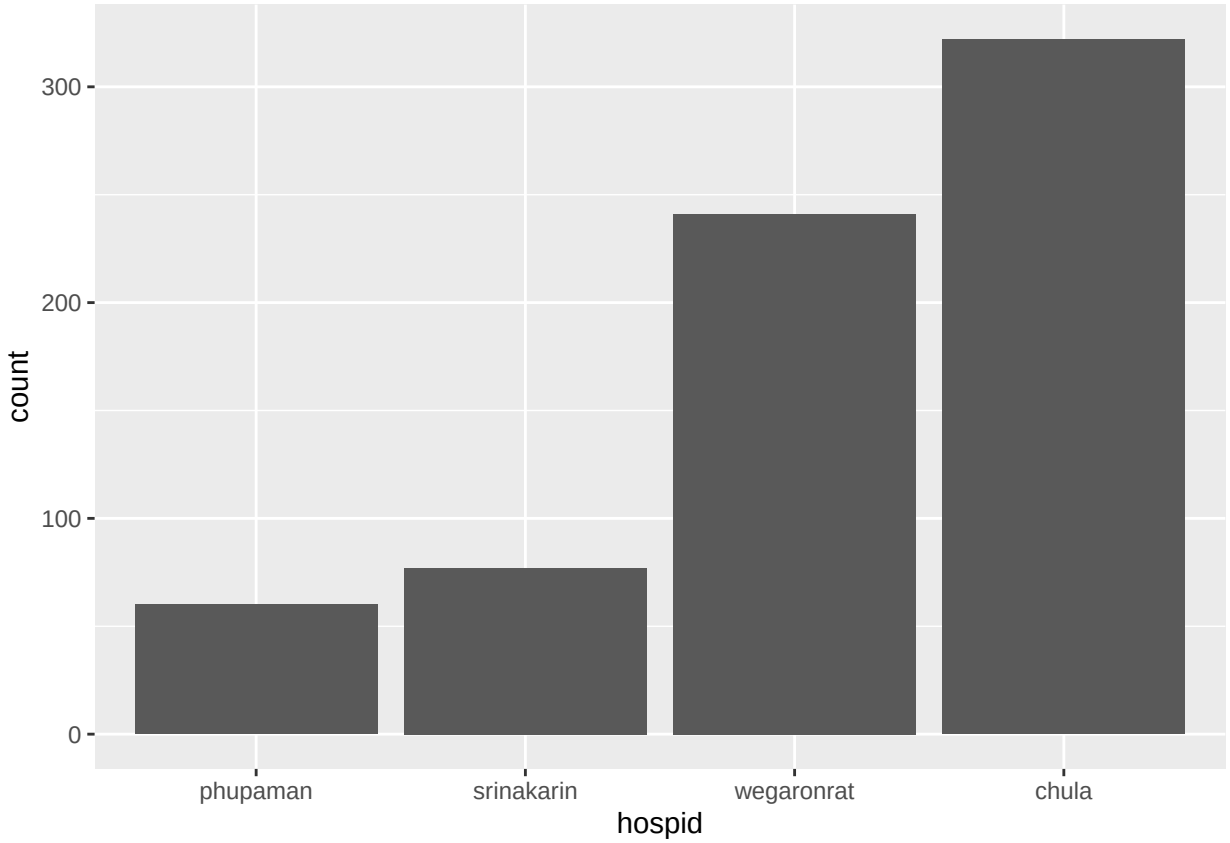

count
No
Sí
comht

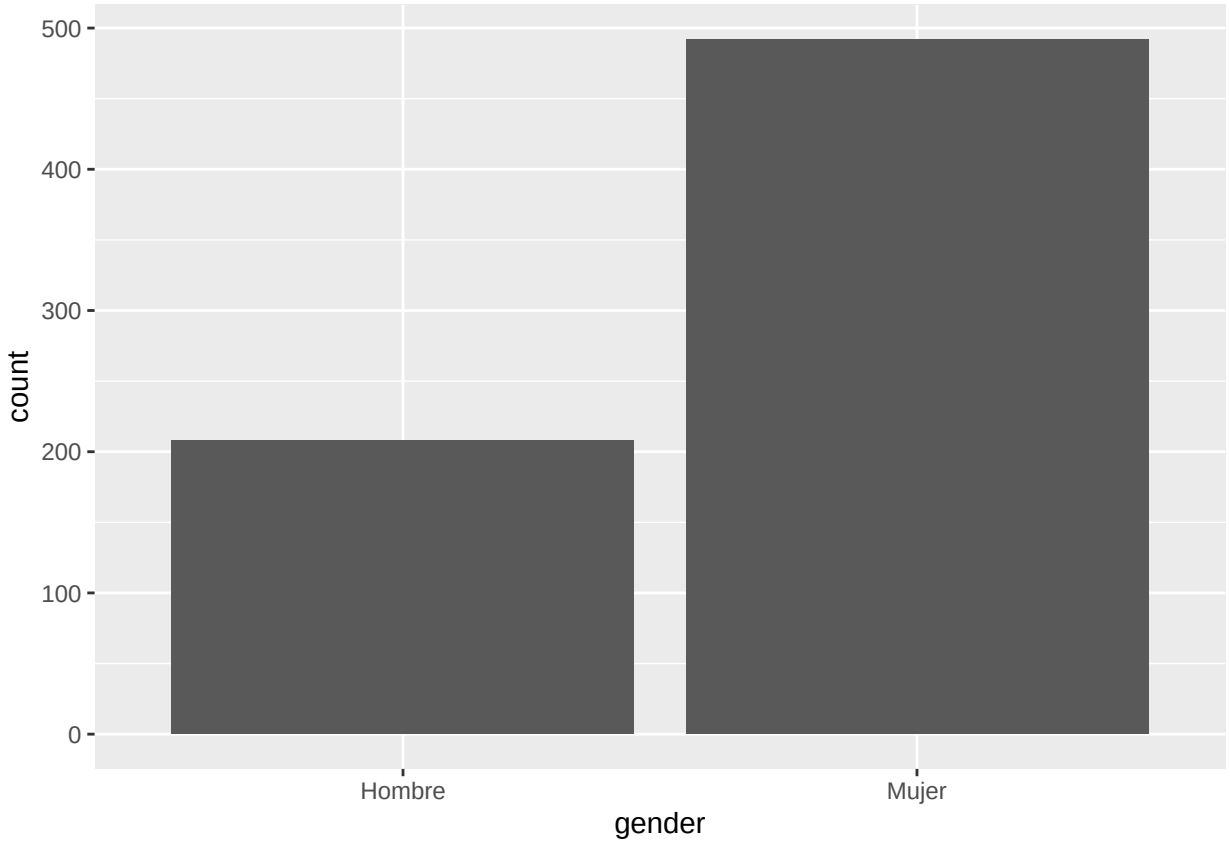

count
No
Sí
comchd

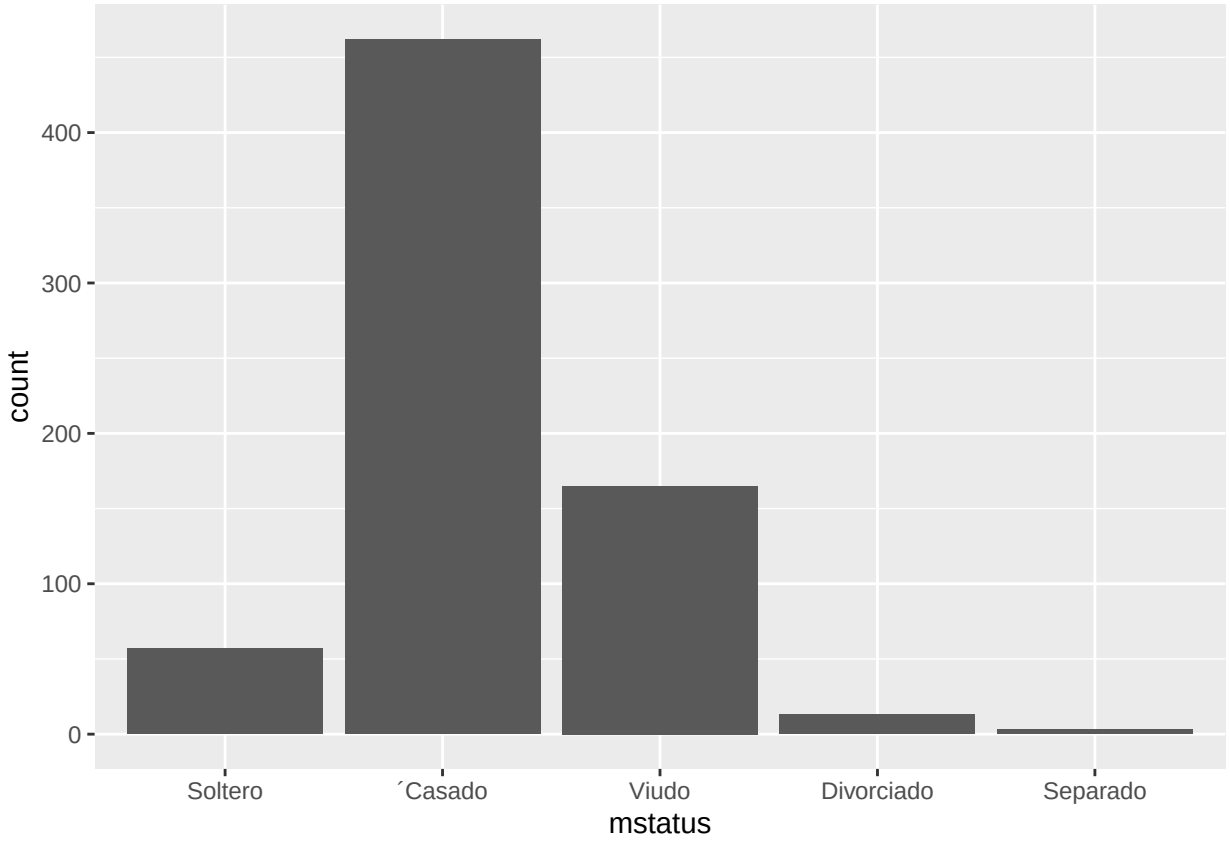

count
No
Sí
comkid

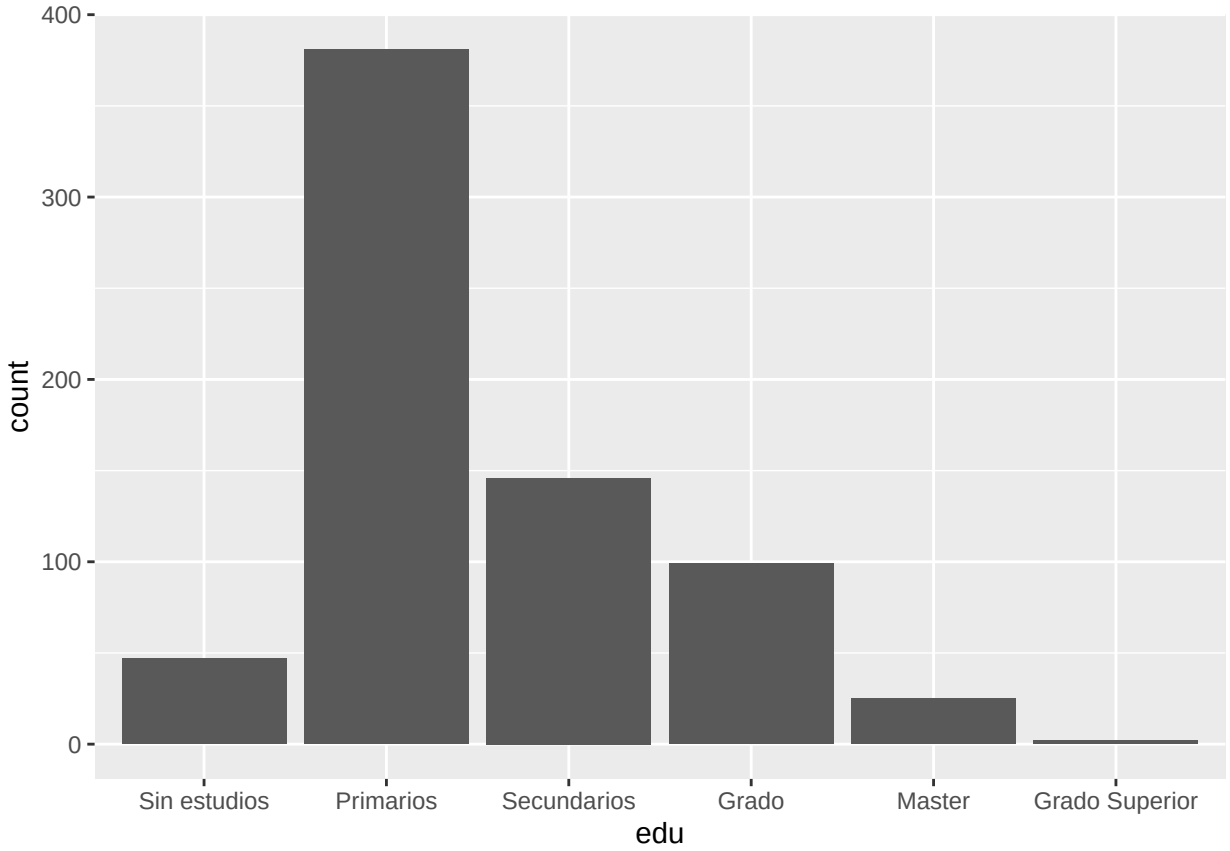

count
No
Sí
comoth
Sin medicación
Hipoglicémico oral
Insulina
Ambos
dmrx
Sí
Exfumador
No
smk
Sí
Exbebedor
No
alcohol
compli
cva
cereinfrac
ishemic
stroke
cerebhem
tia
angia

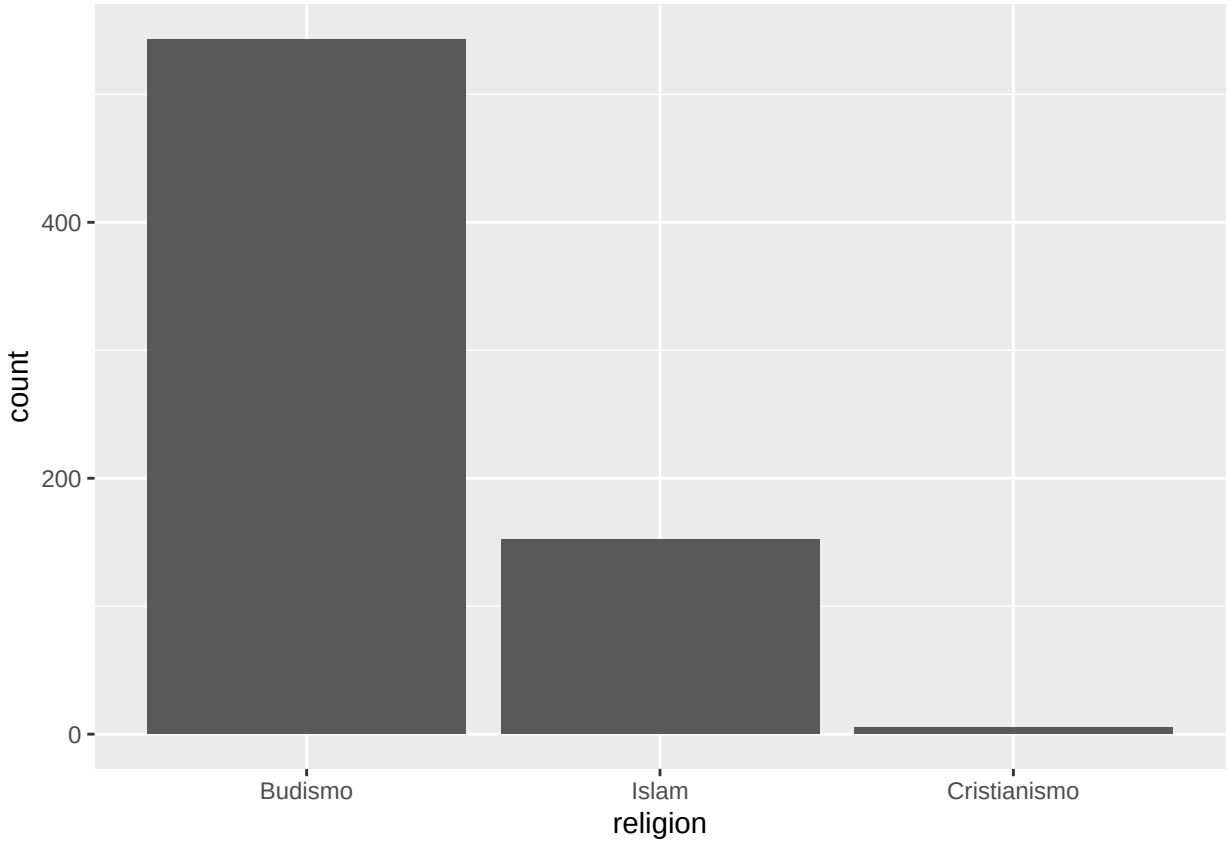

count
No
Sí
chf
mi
cororevas
pad
neuropath
renal
dn
dr
othcomp
alerta
sobrepeso
obesidad

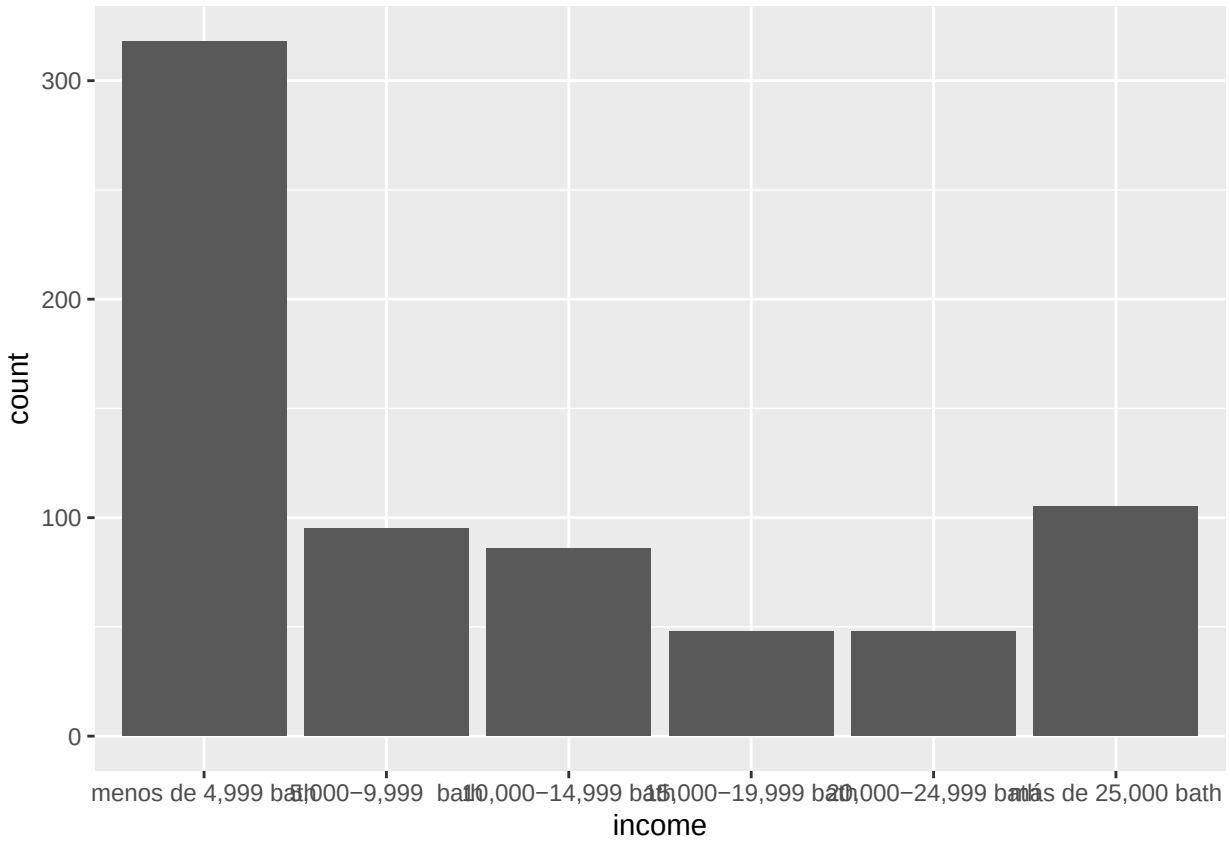

300
200
100
0
count
phupaman
srinakarin
wegaronrat
chula
hospid

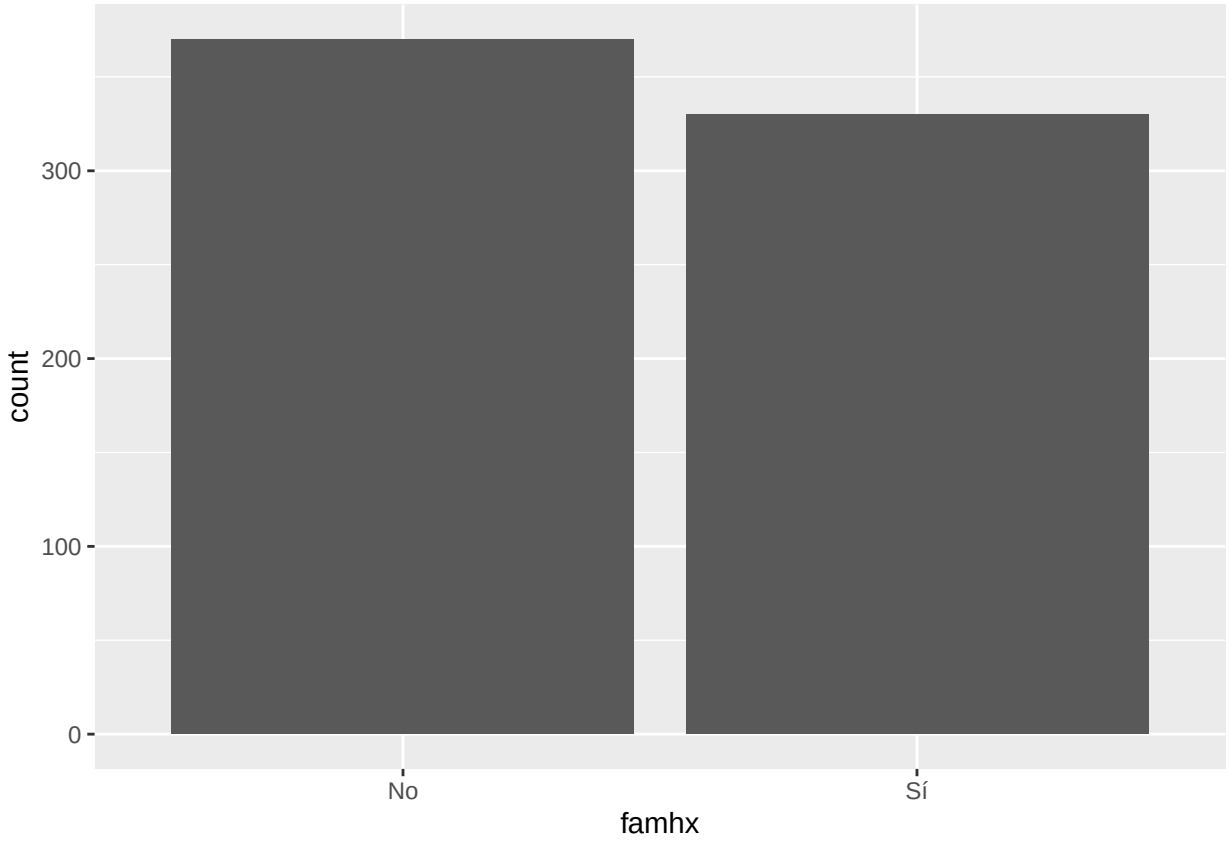

500
400
300
200
100
0
count
Hombre
Mujer
gender

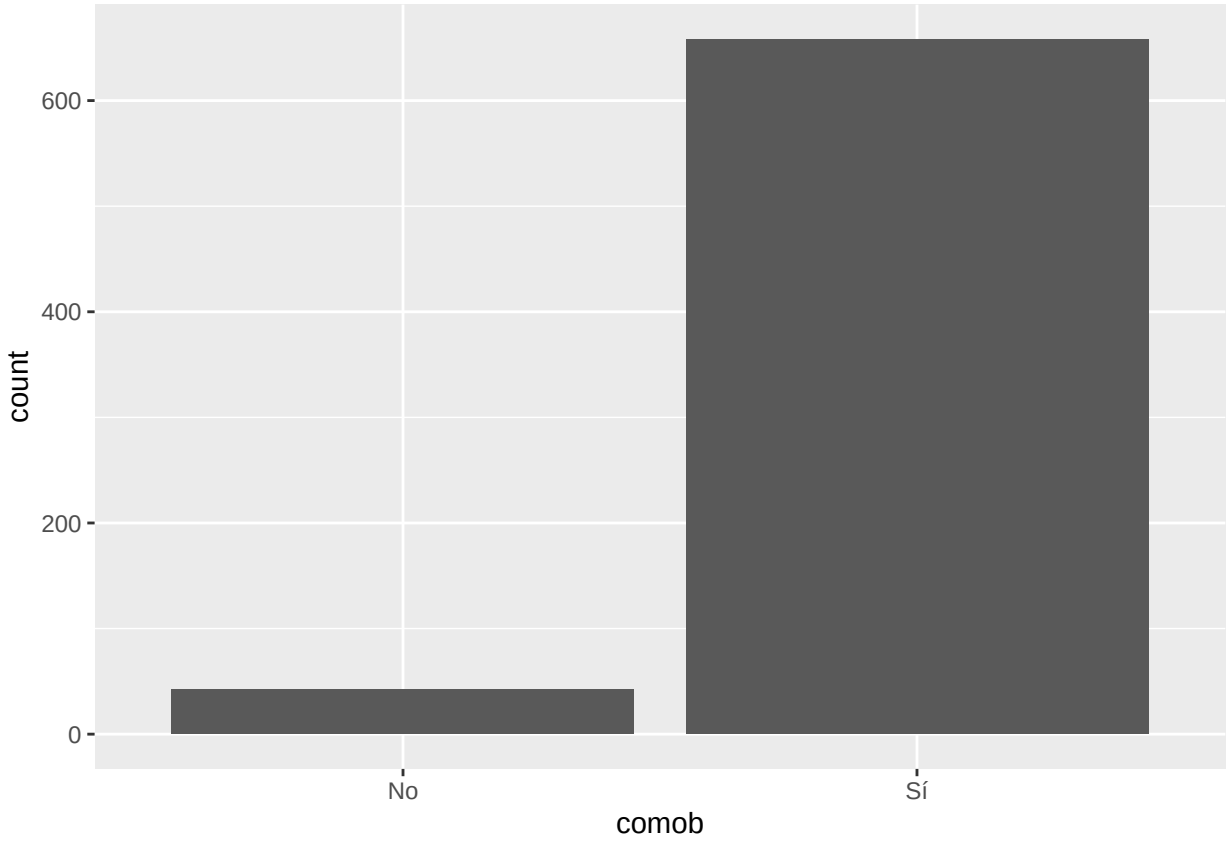

400
300
200
100
0
count
Soltero
´Casado
Viudo
Divorciado
Separado
mstatus

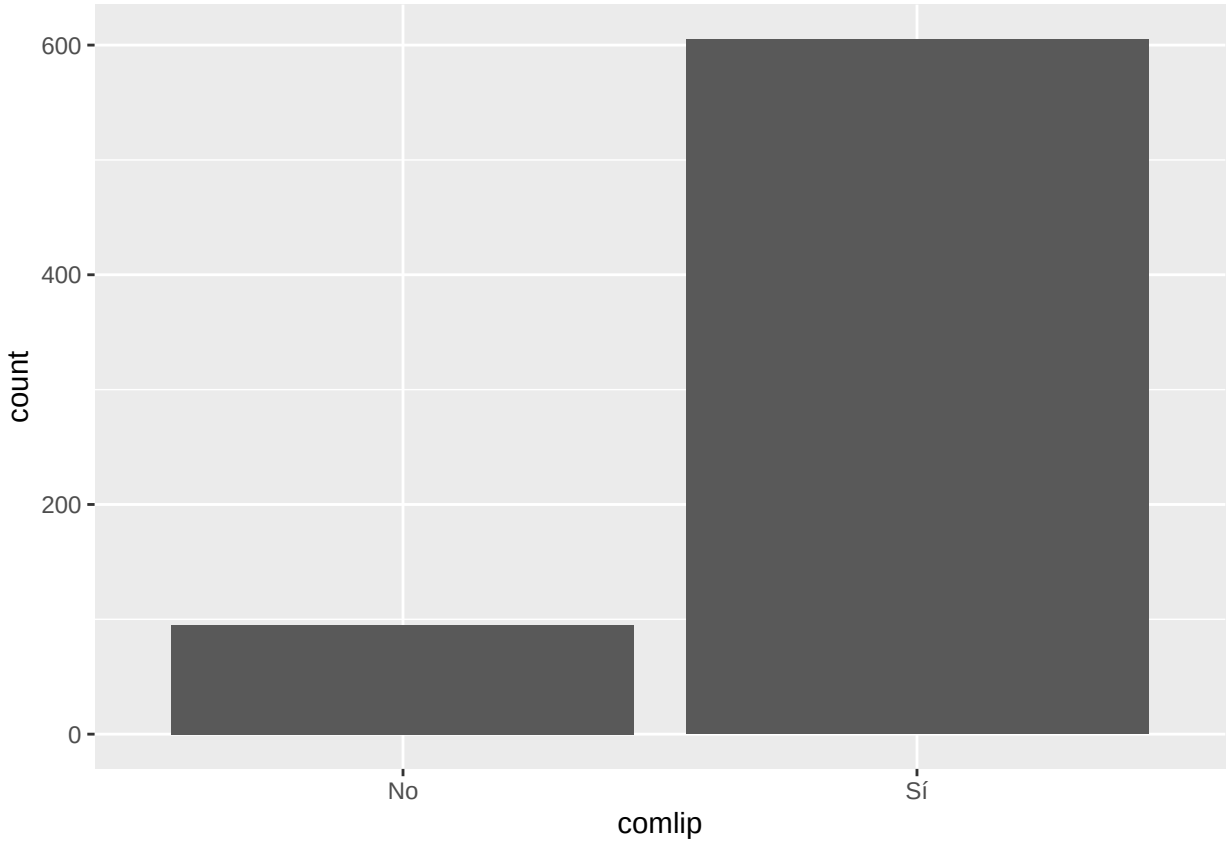

400
300
200
100
0
count
Sin estudios
Primarios
Secundarios
Grado
Master
Grado Superior
edu

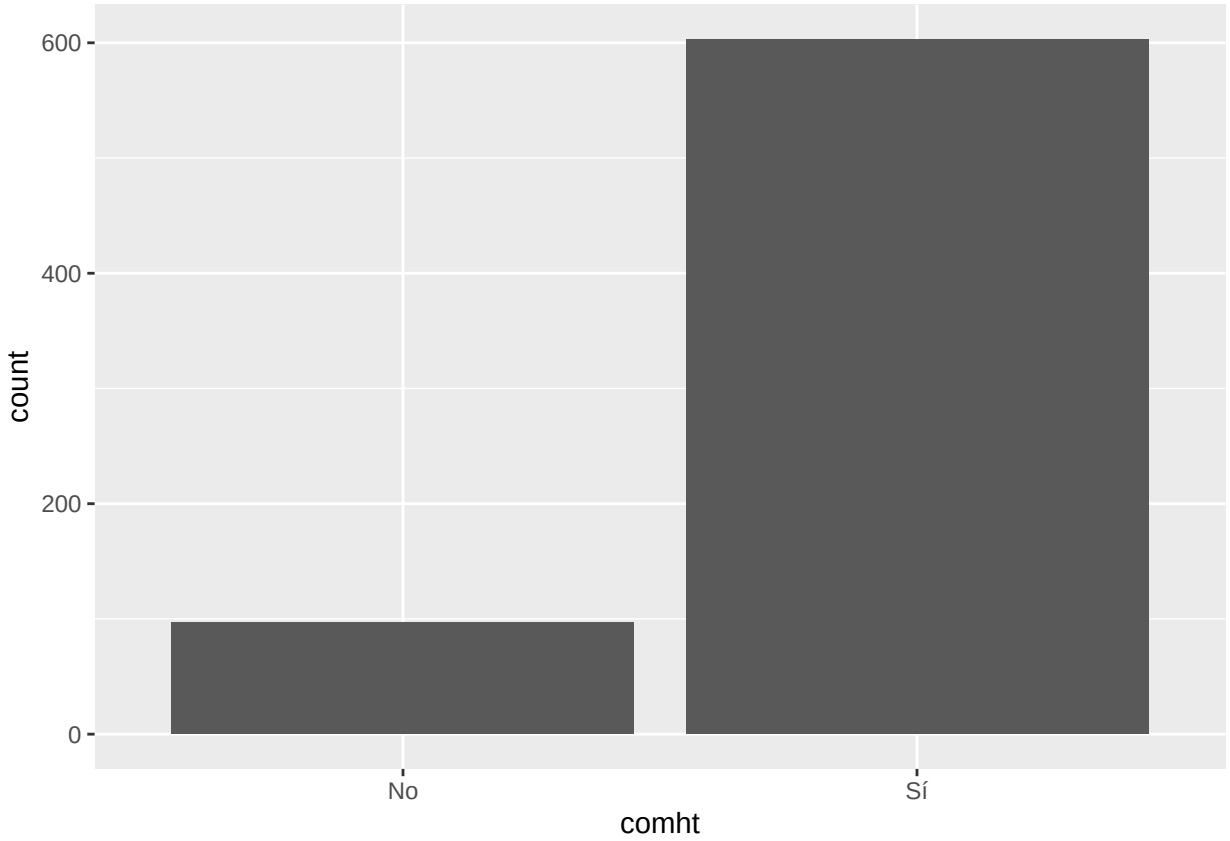

400
200
0
count
Budismo
Islam
Cristianismo
religion

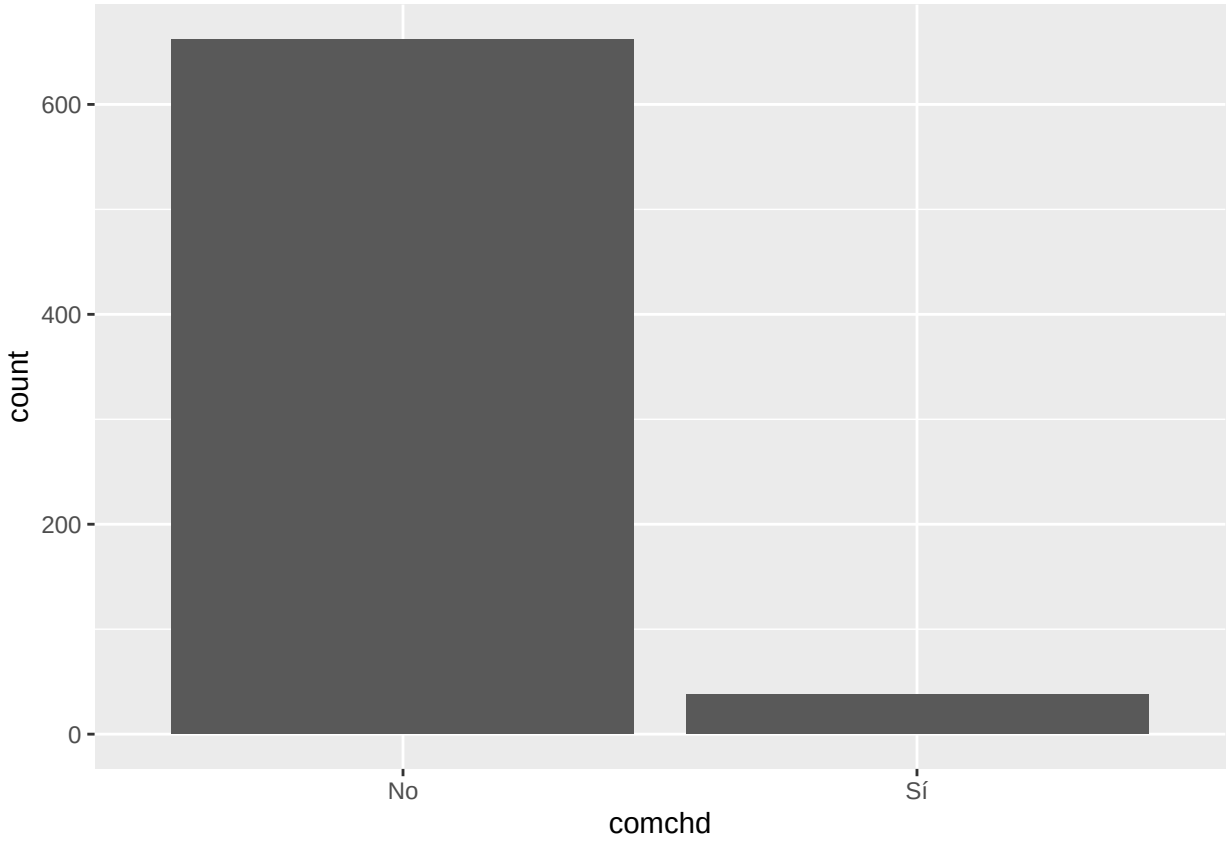

300
200
100
0
count
menos de 4,999 bath
5,000−9,999 bath
10,000−14,999 bath
15,000−19,999 bath
20,000−24,999 bath
más de 25,000 bath
income

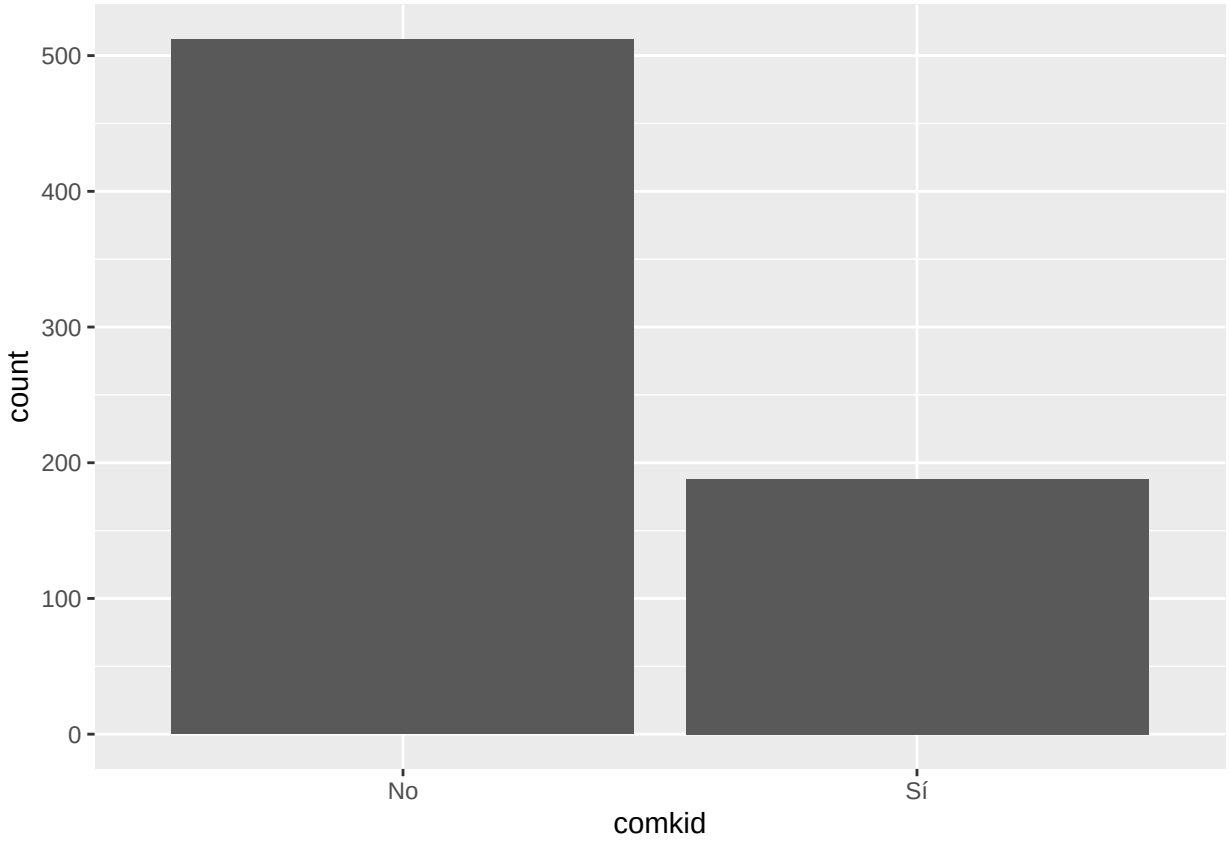

300
200
100
0
count
No
Sí
famhx

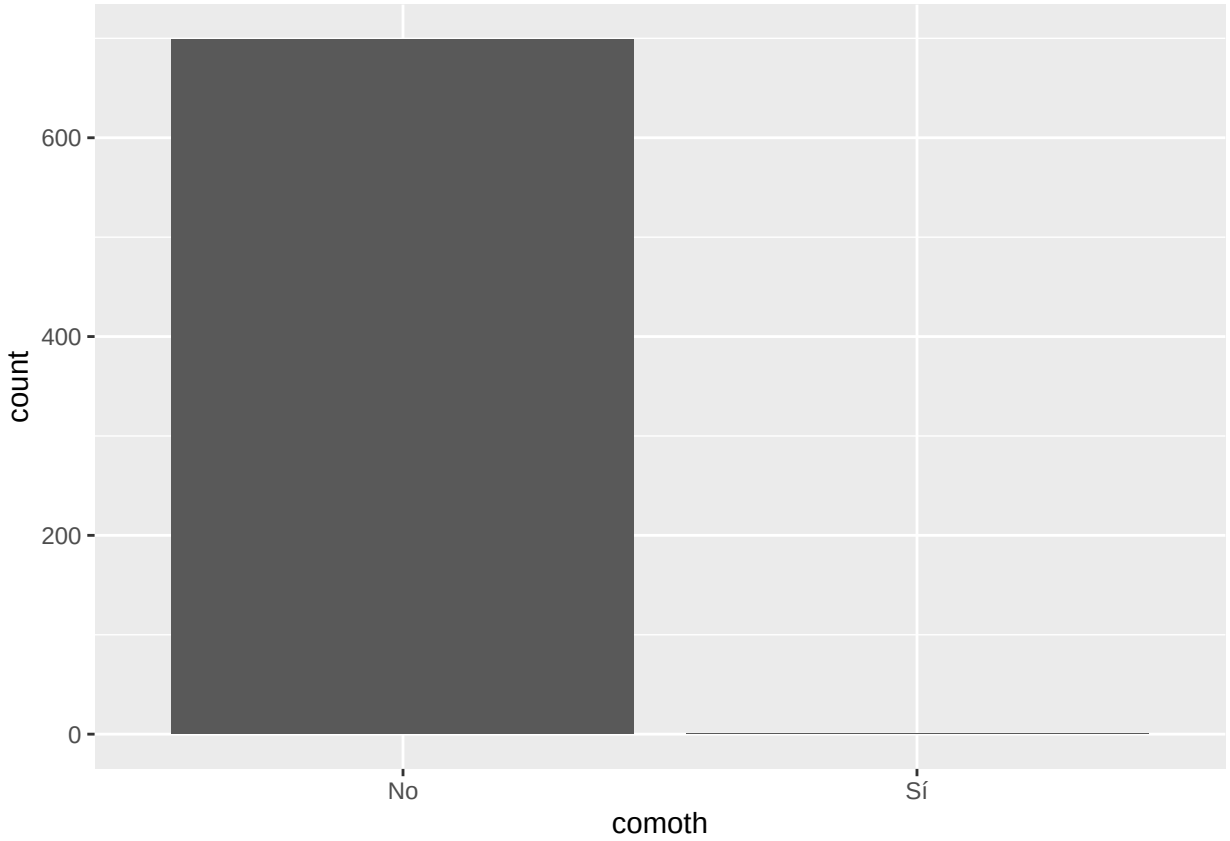

600
400
200
0
count
No
Sí
comob

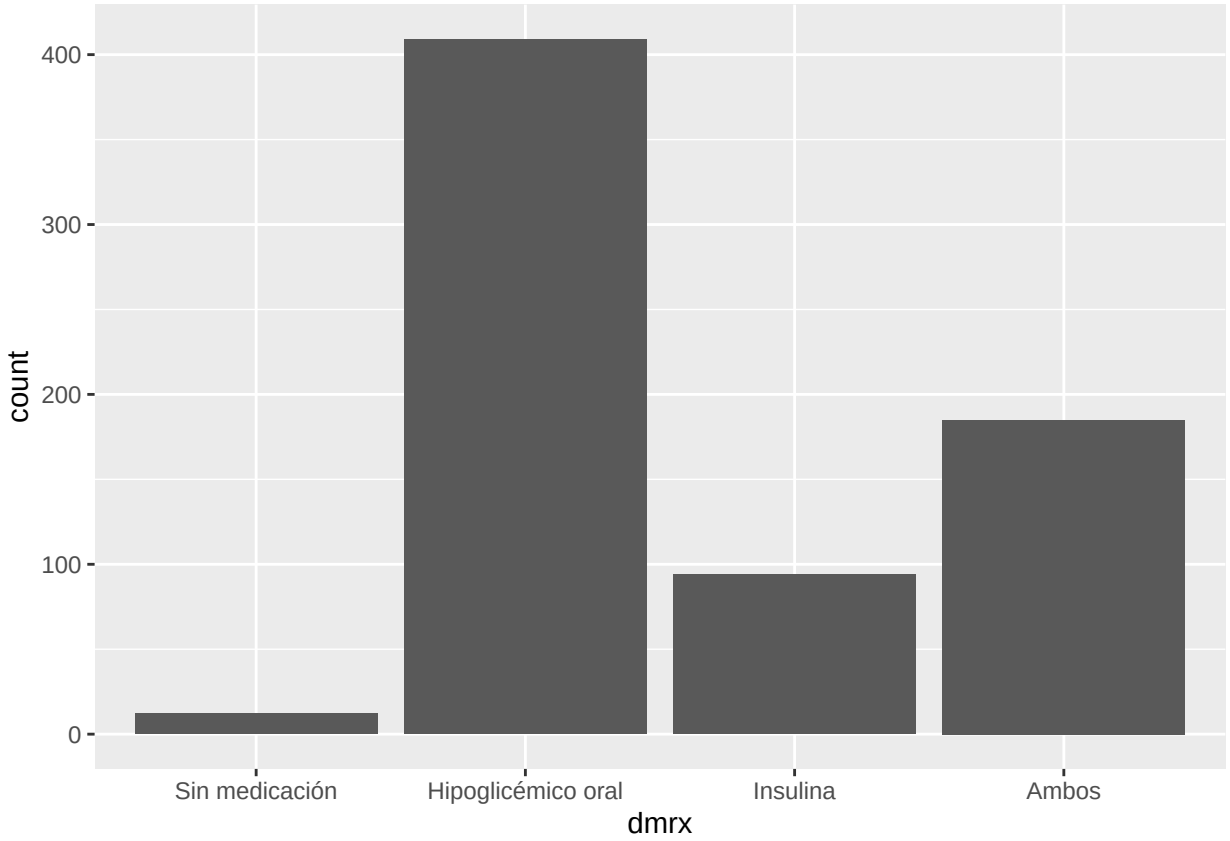

600
400
200
0
count
No
Sí
comlip

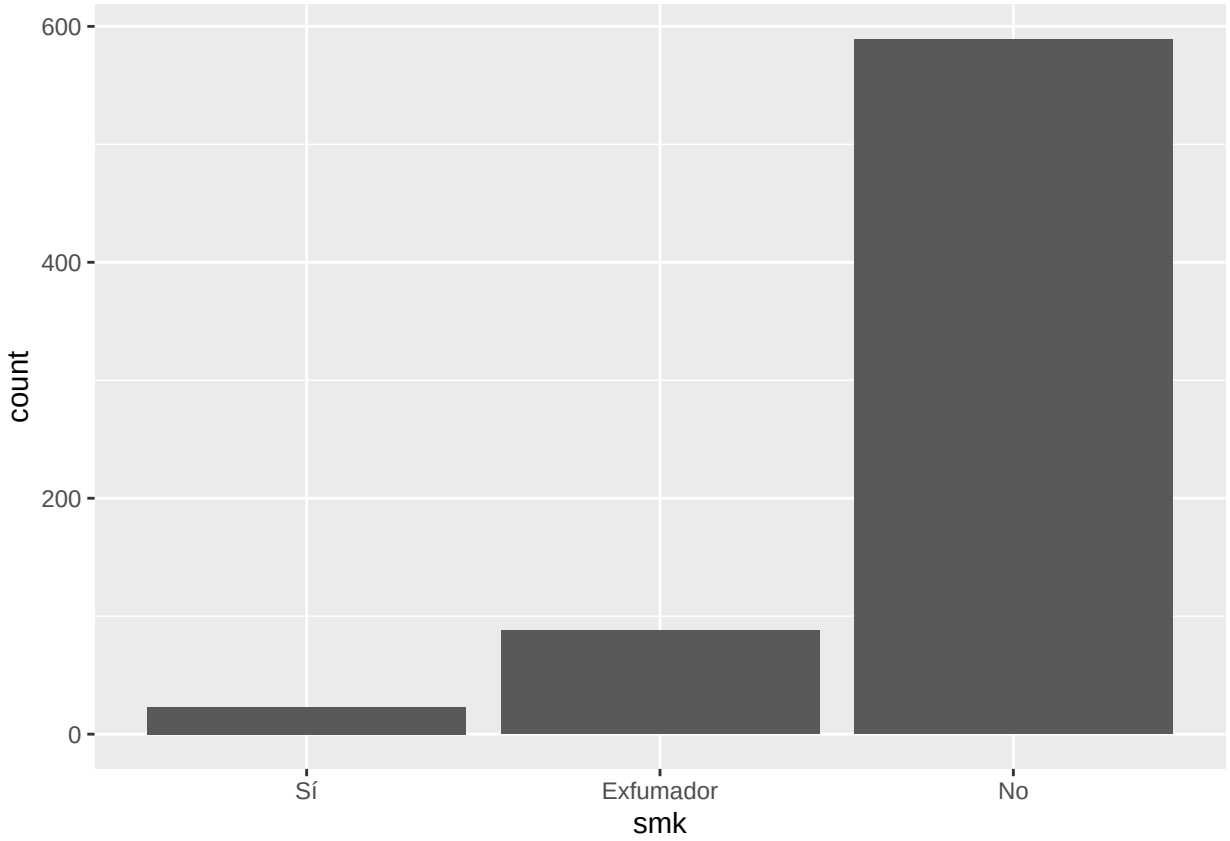

600
400
200
0
count
No
Sí
comht

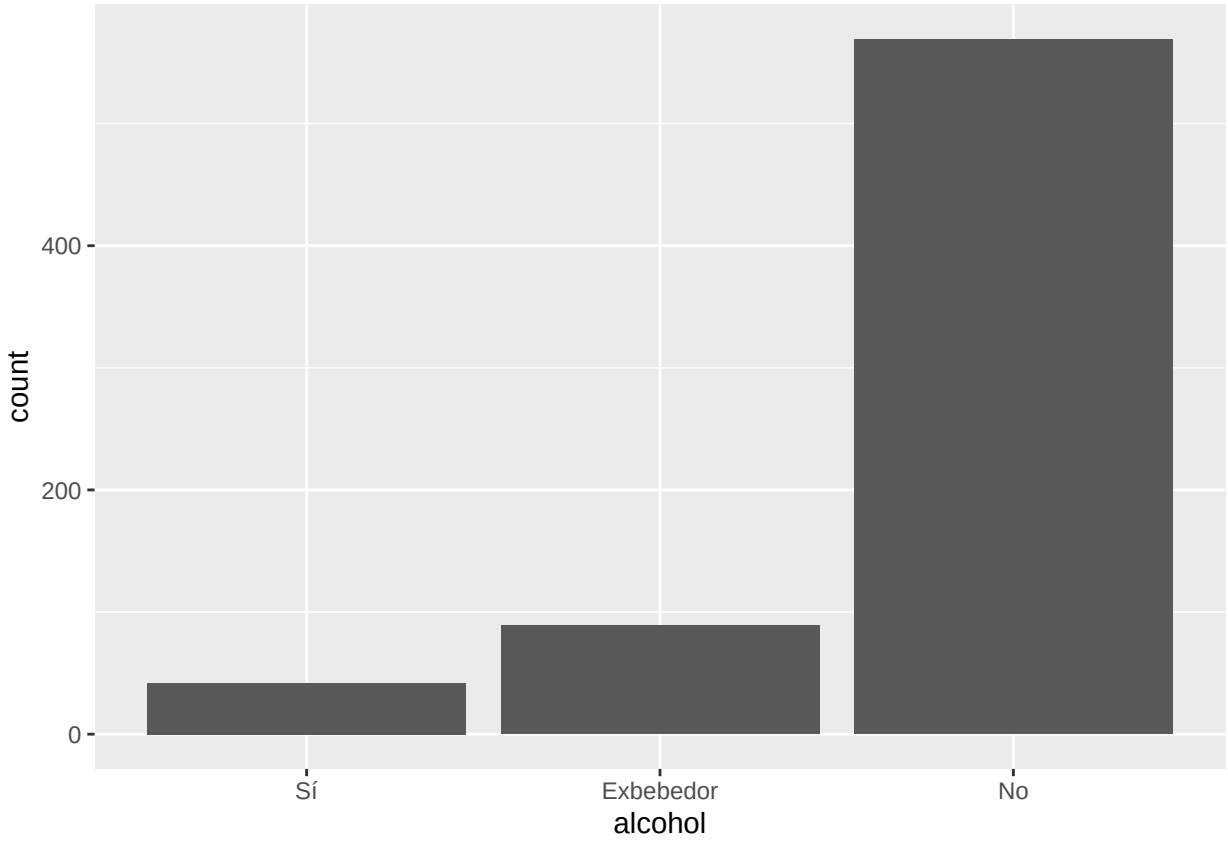

600
400
200
0
count
No
Sí
comchd

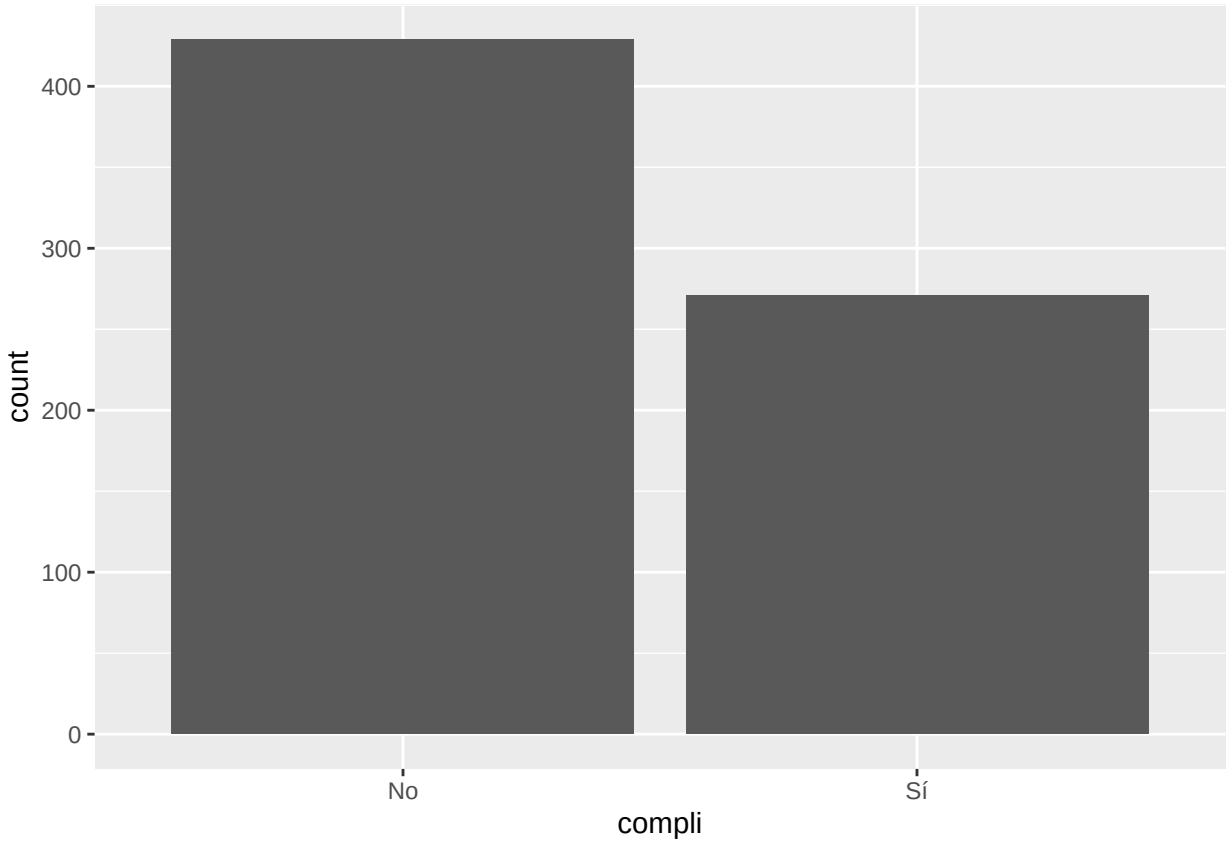

500
400
300
200
100
0
count
No
Sí
comkid

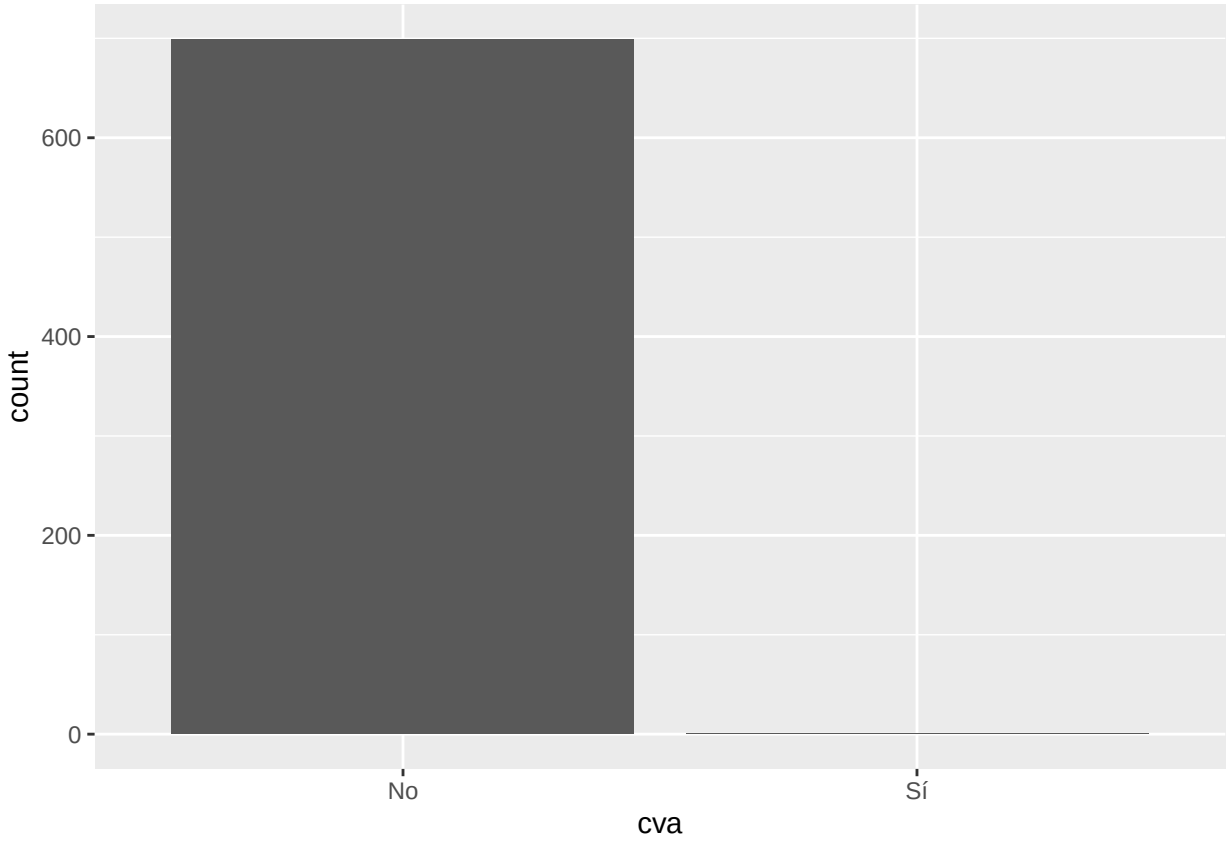

600
400
200
0
count
No
Sí
comoth

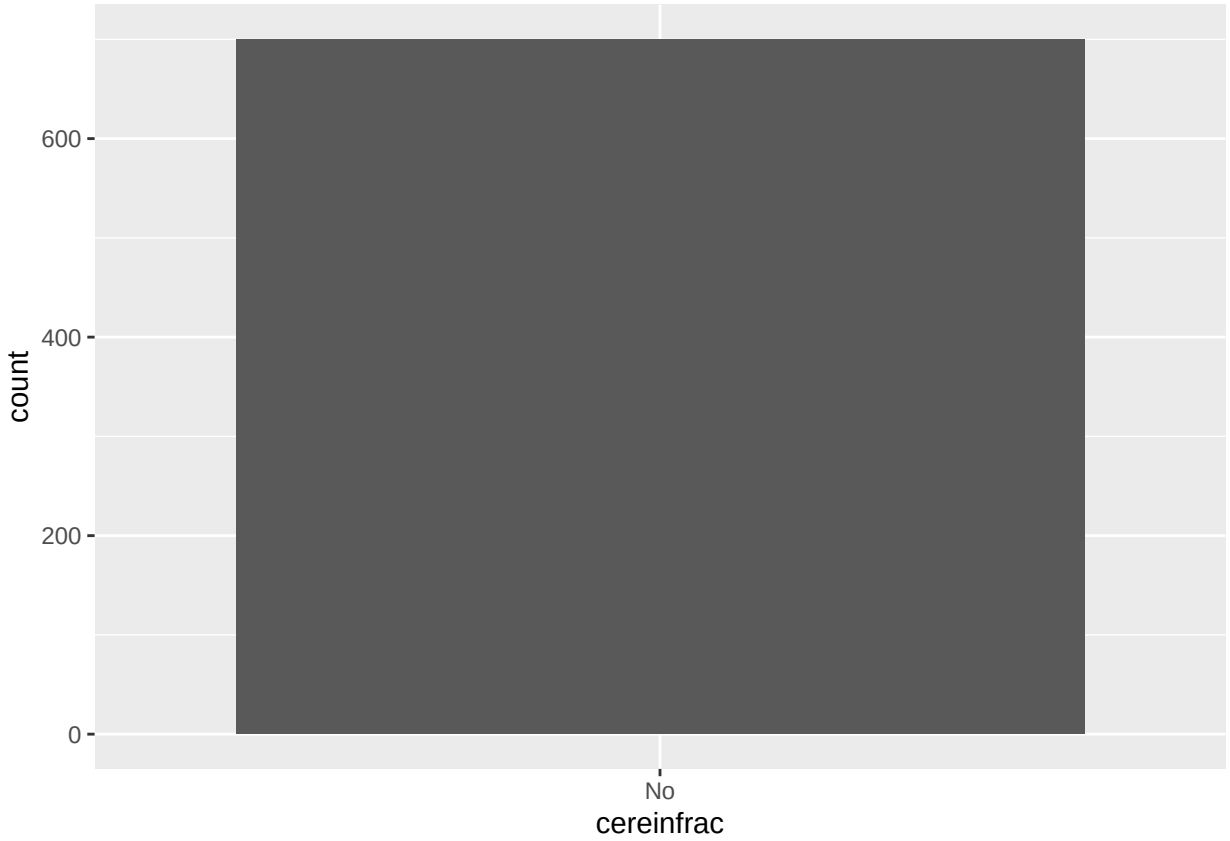

400
300
200
100
0
count
Sin medicación
Hipoglicémico oral
Insulina
Ambos
dmrx

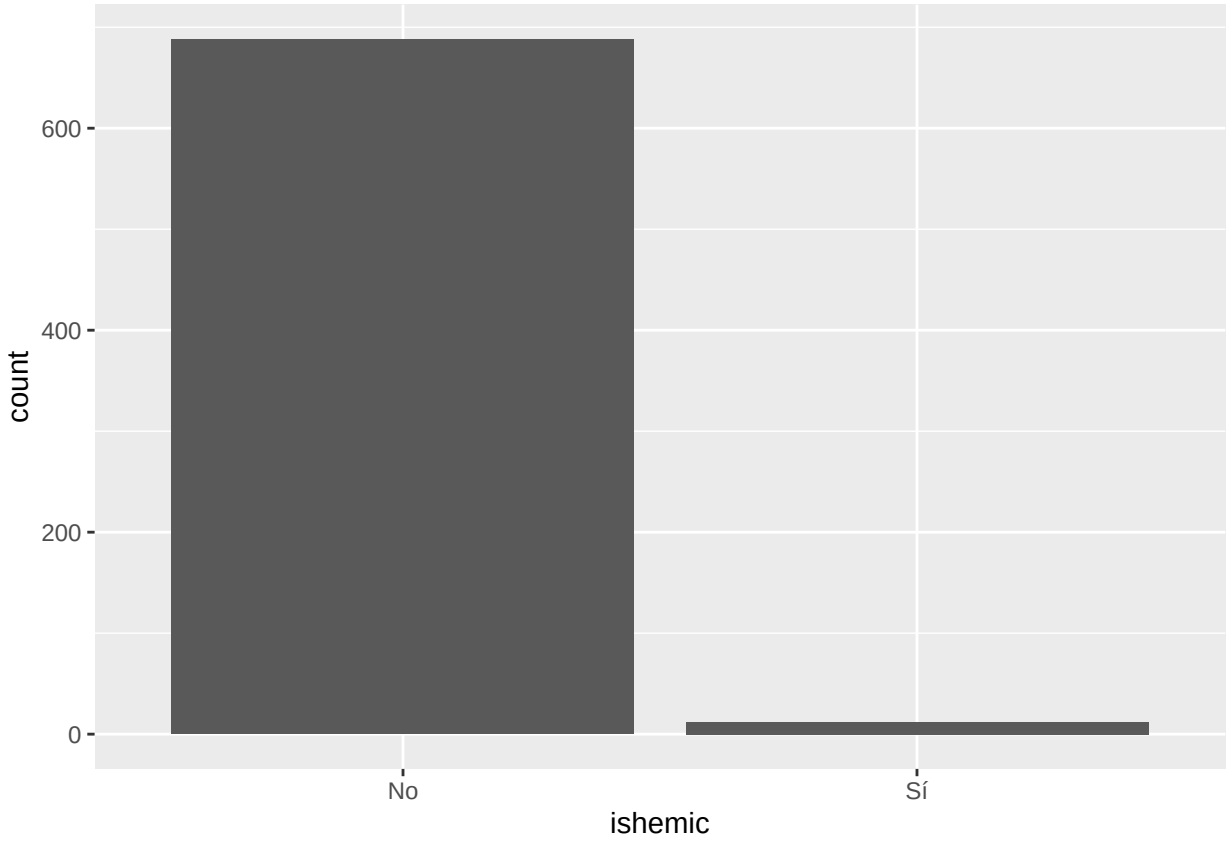

600
400
200
0
count
Sí
Exfumador
No
smk

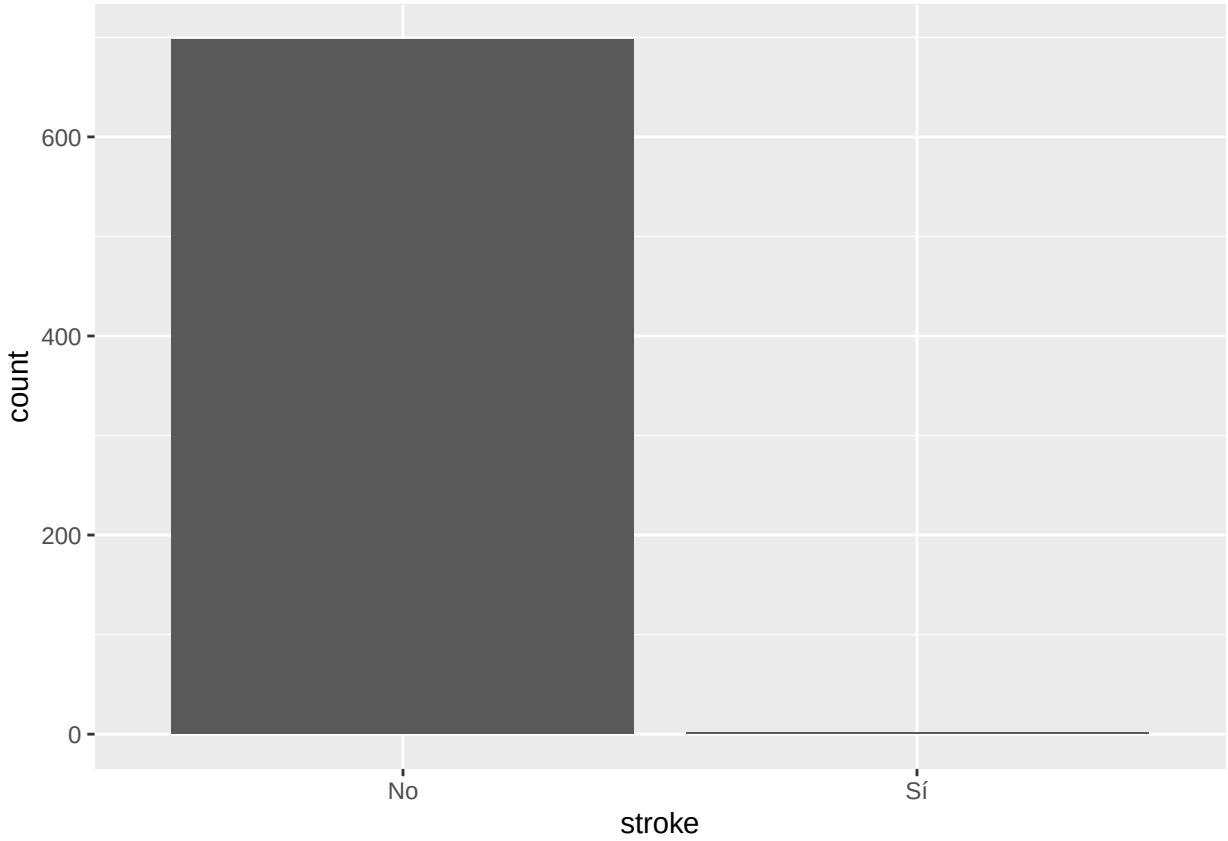

400
200
0
count
Sí
Exbebedor
No
alcohol

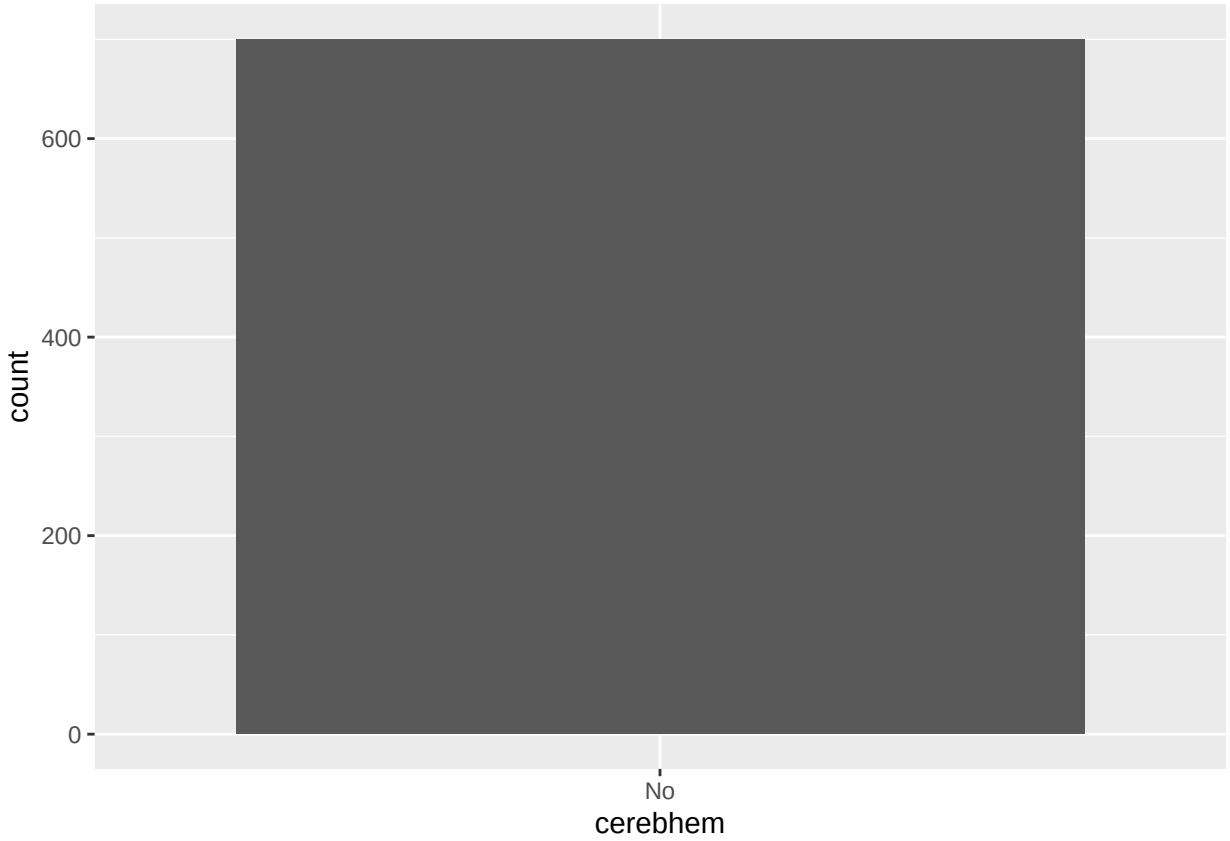

400
300
200
100
0
count
No
Sí
compli

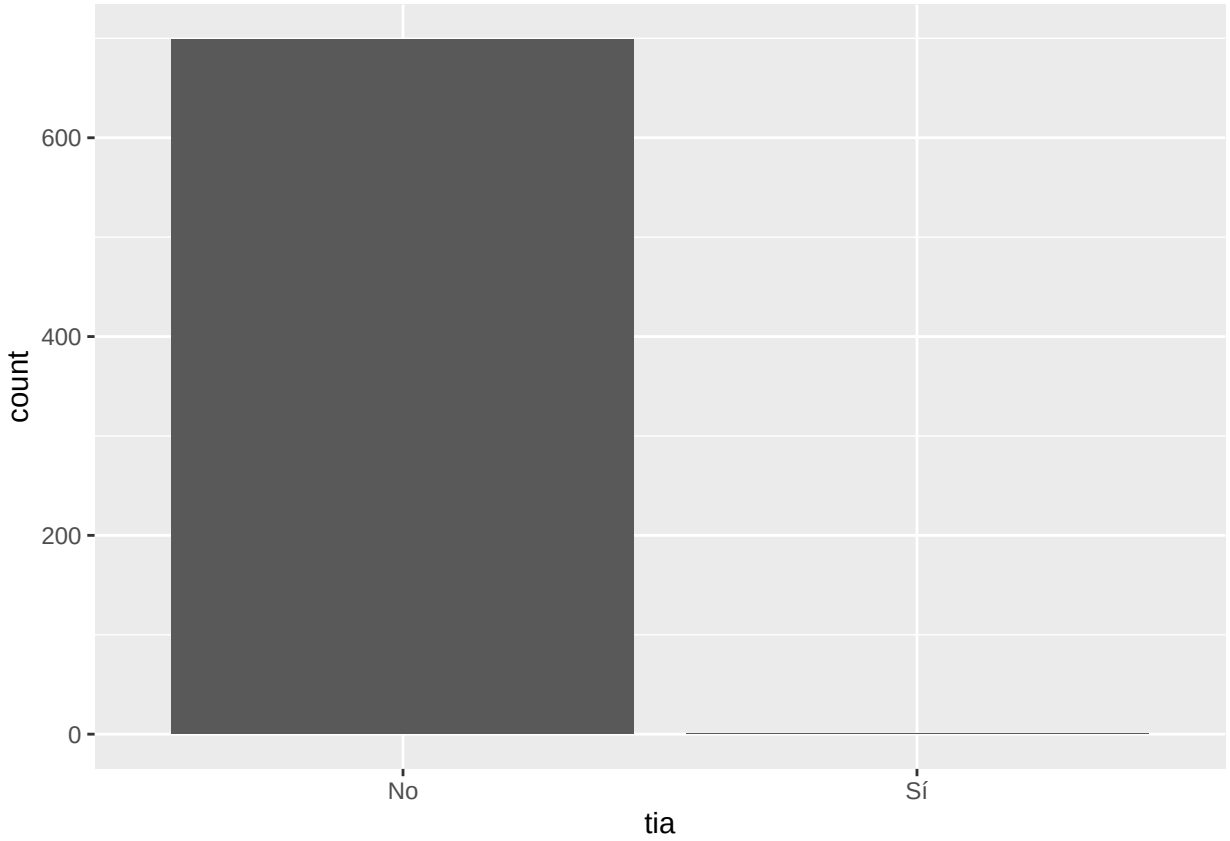

600
400
200
0
count
No
Sí
cva

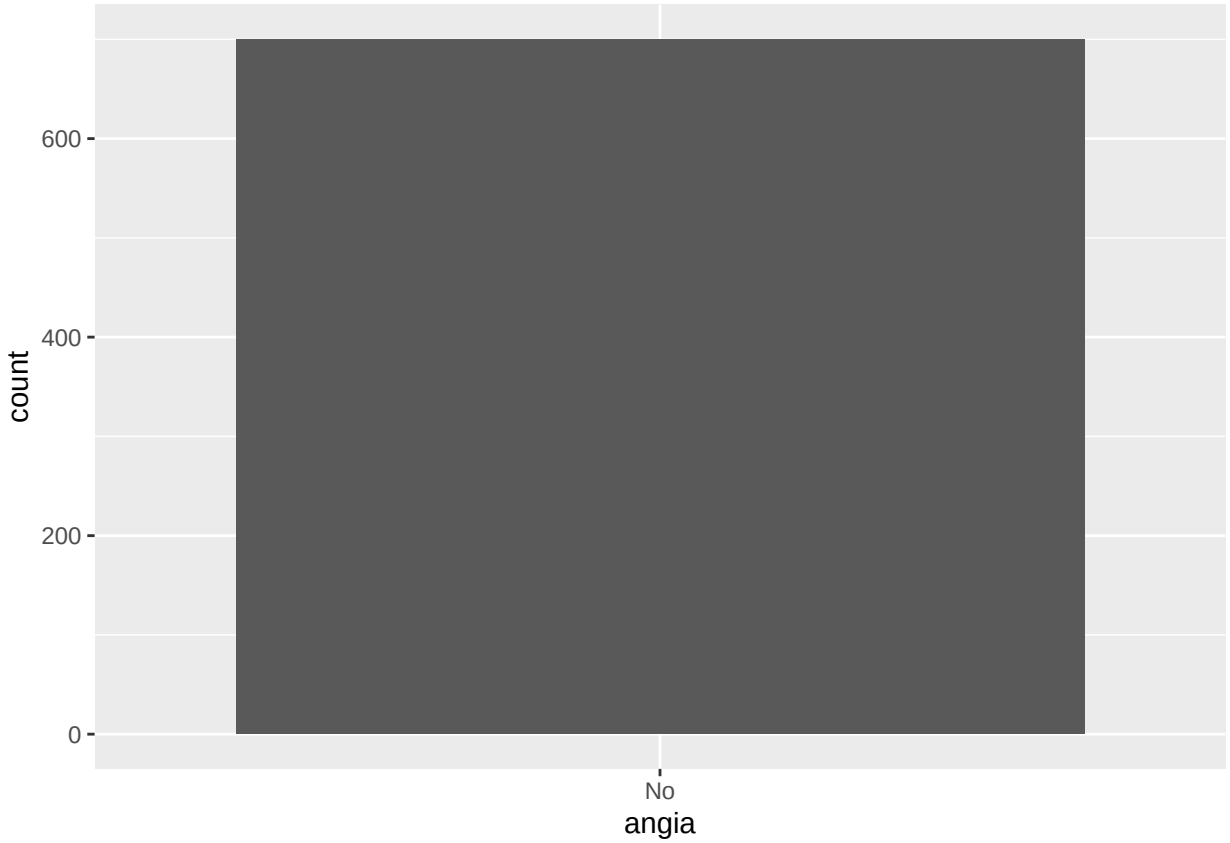

600
400
200
0
count
No
cereinfrac

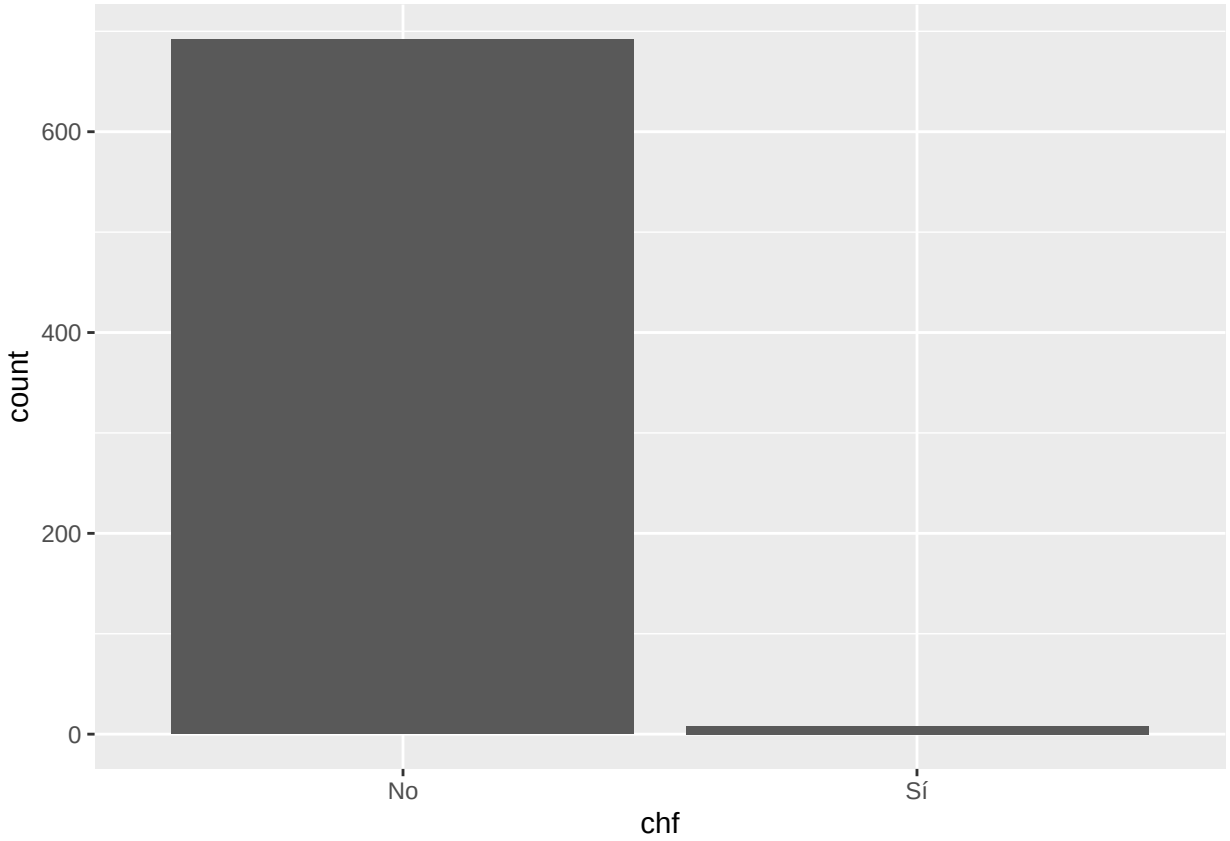

600
400
200
0
count
No
Sí
ishemic

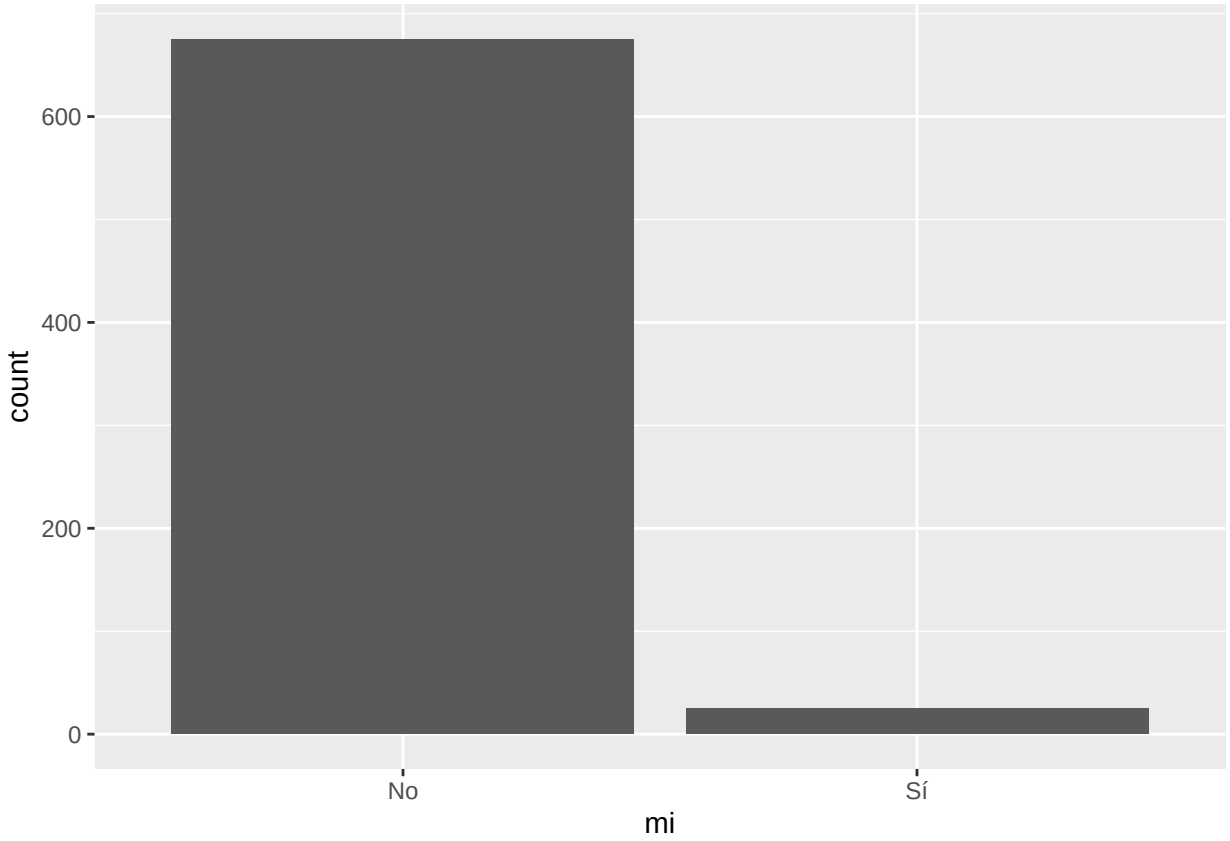

600
400
200
0
count
No
Sí
stroke

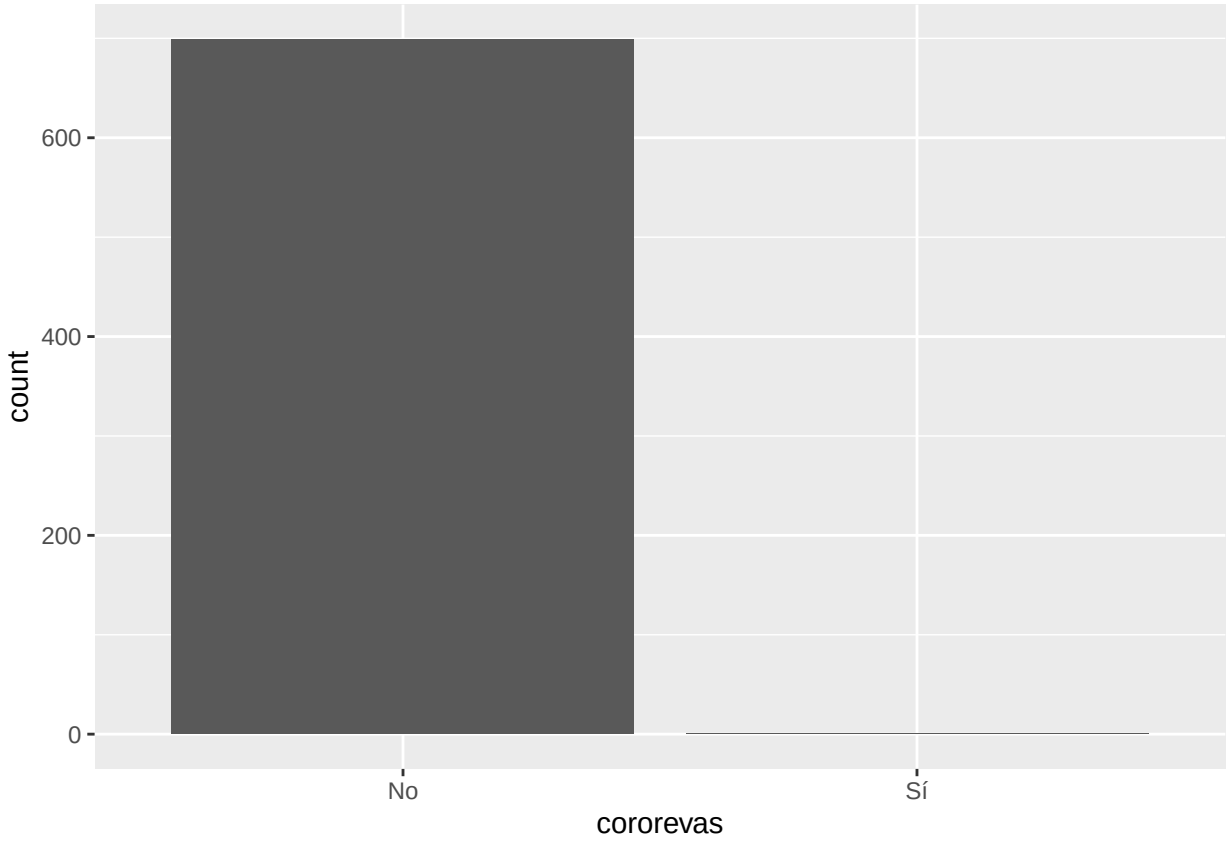

600
400
200
0
count
No
cerebhem

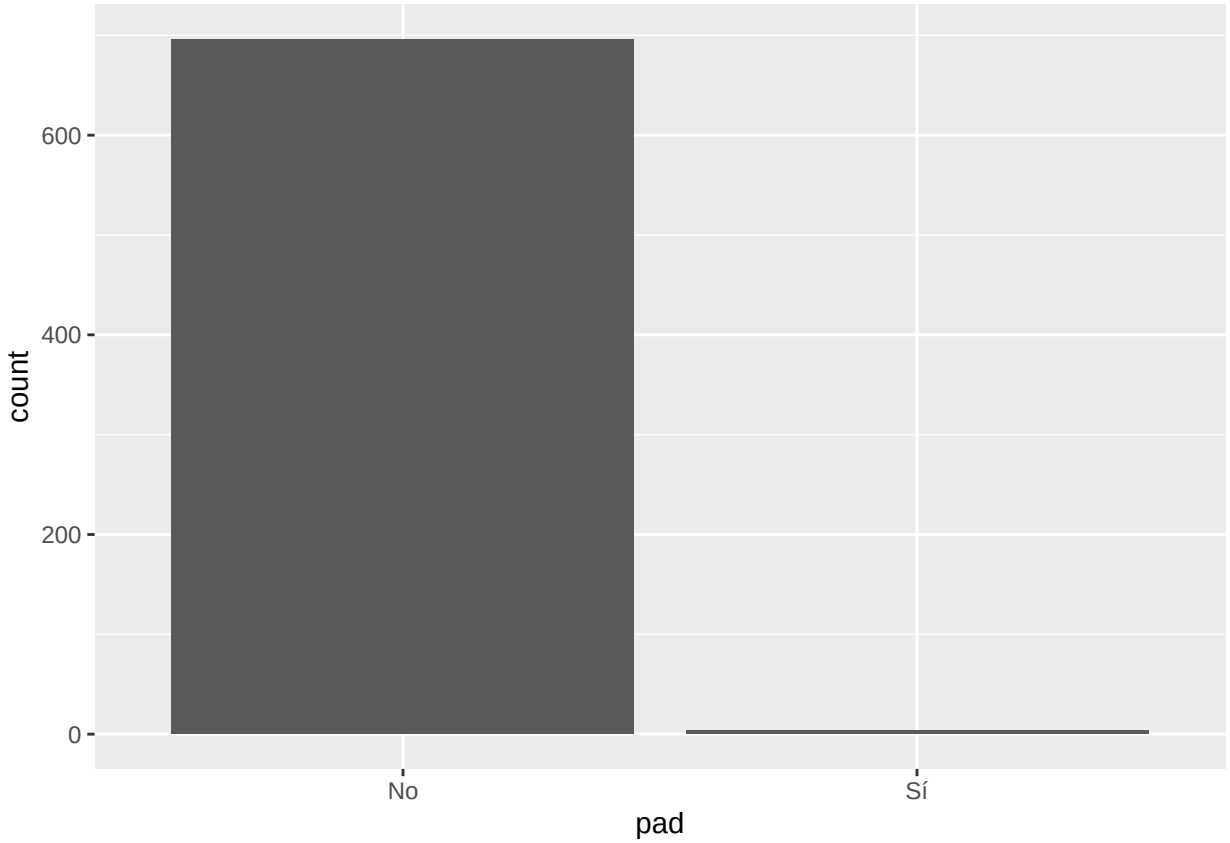

600
400
200
0
count
No
Sí
tia

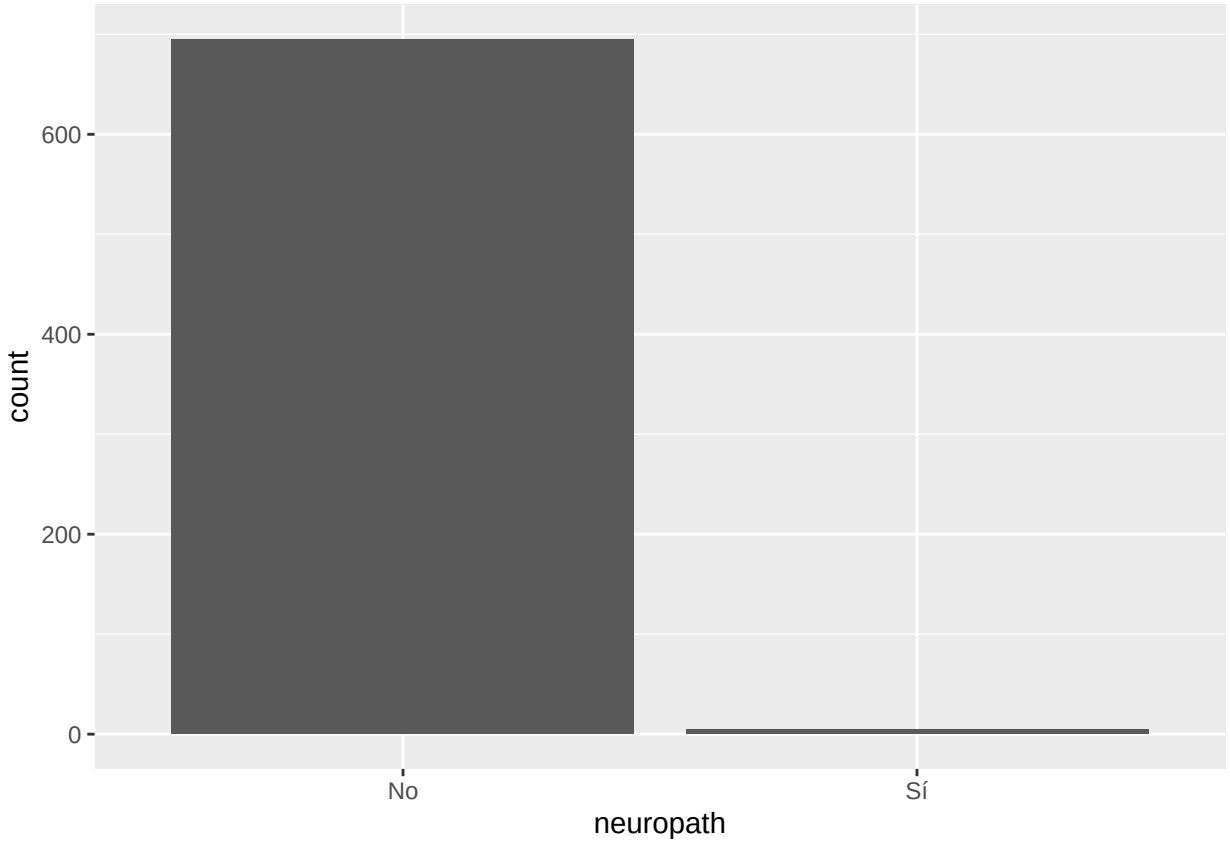

600
400
200
0
count
No
angia

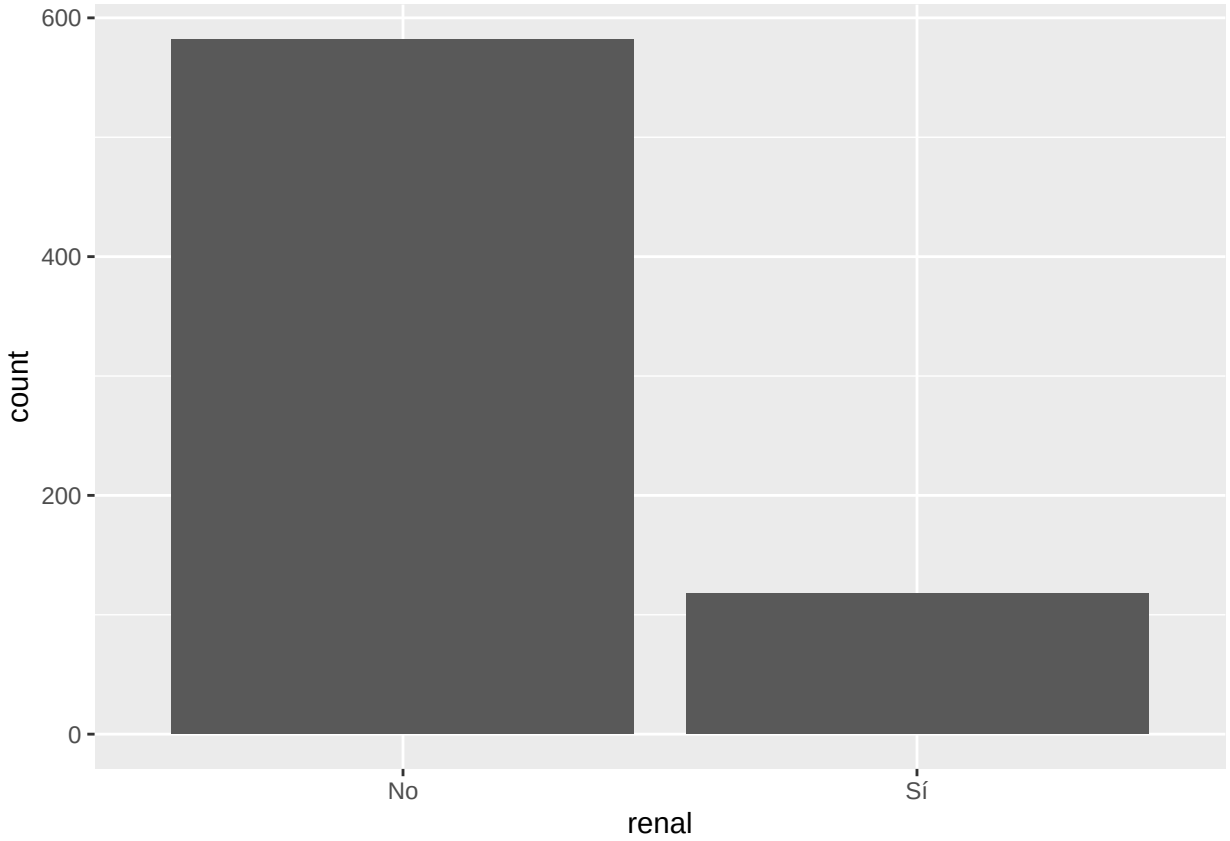

600
400
200
0
count
No
Sí
chf

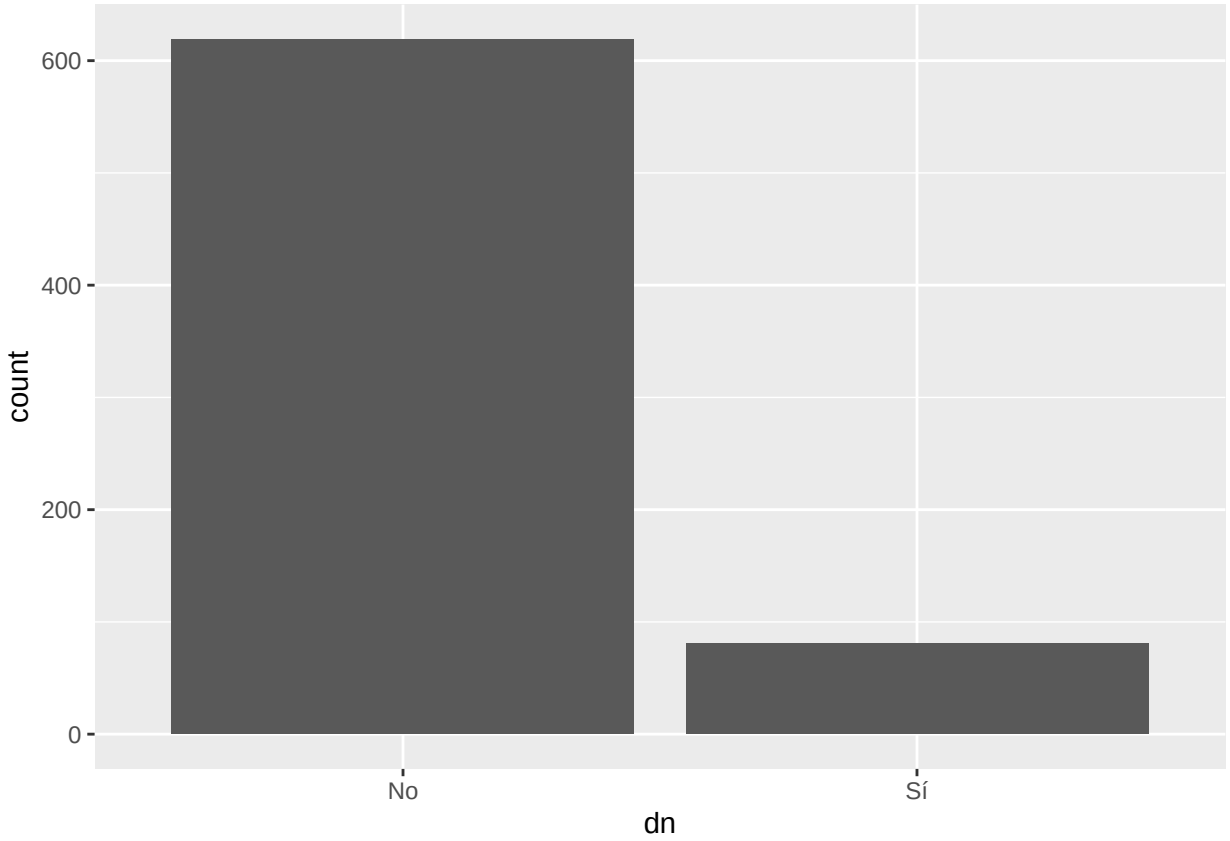

600
400
200
0
count
No
Sí
mi

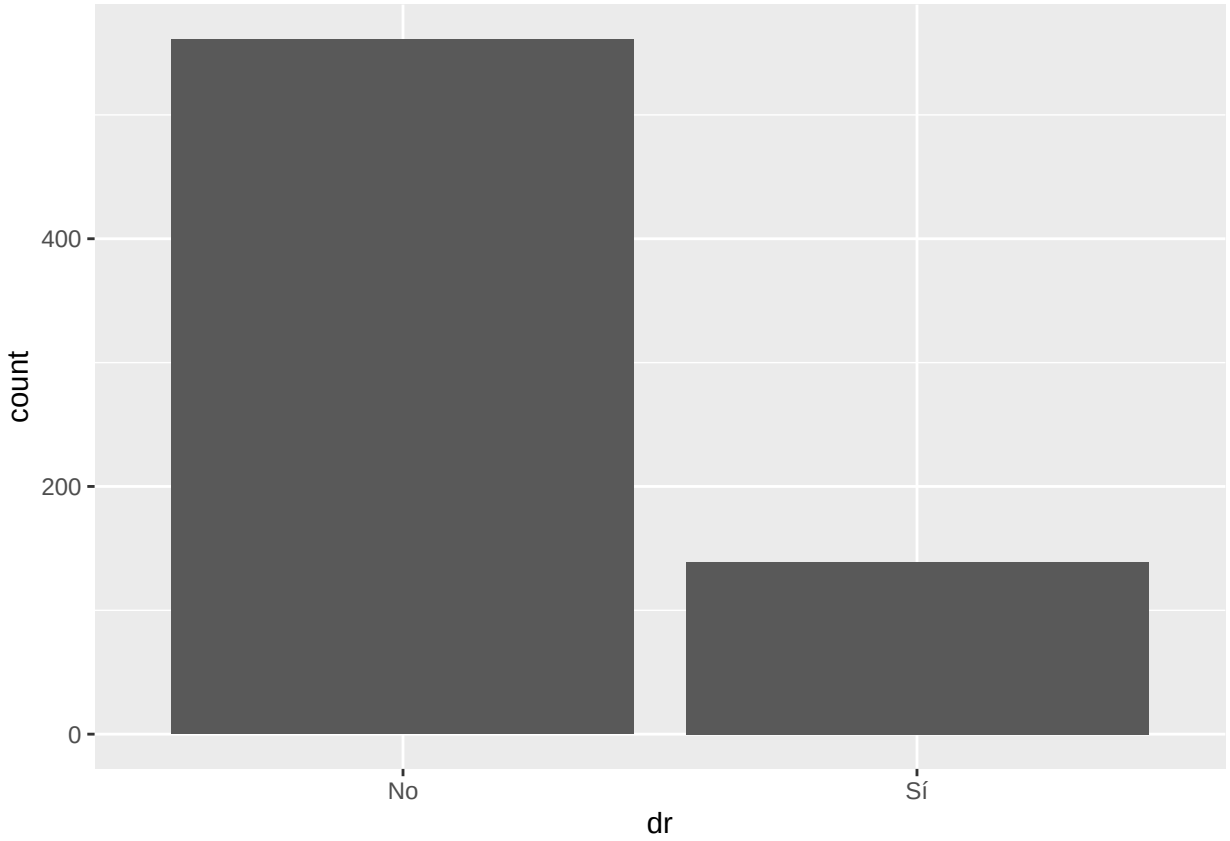

600
400
200
0
count
No
Sí
cororevas

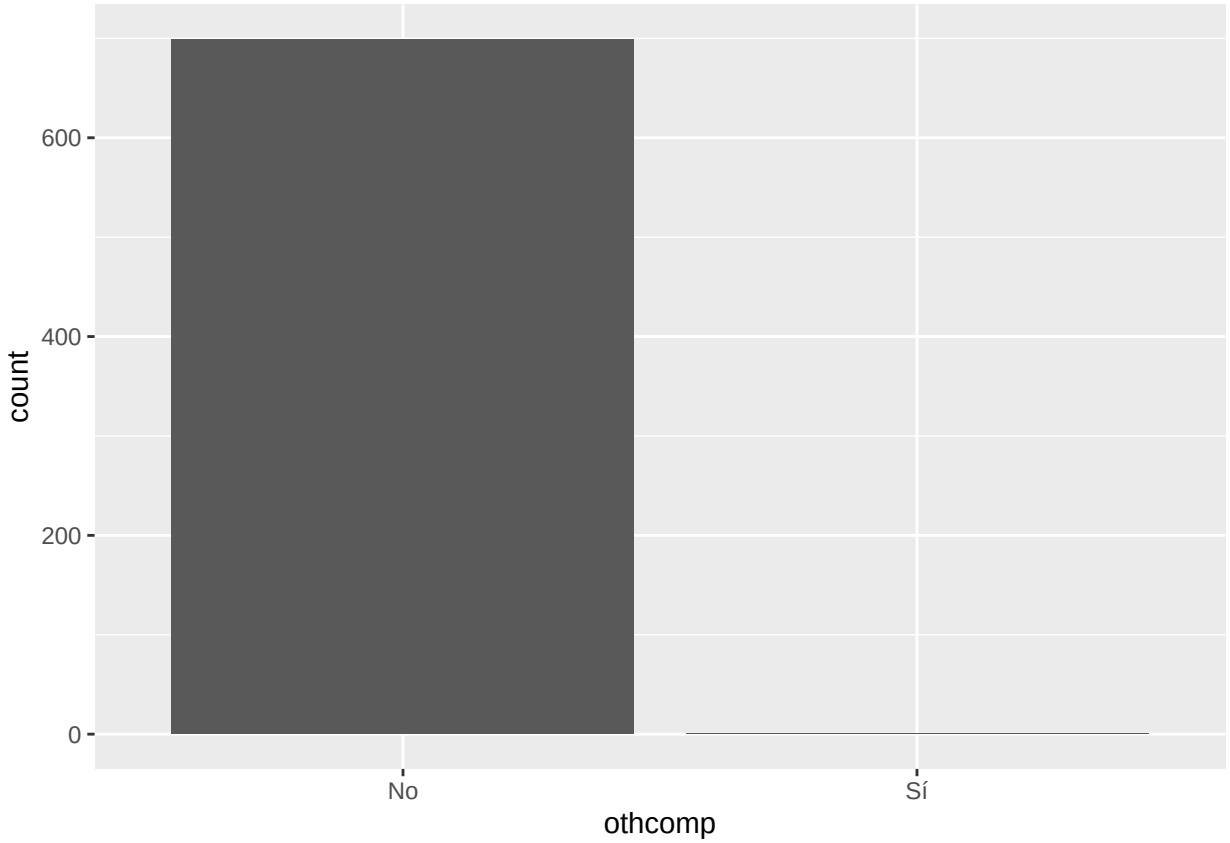

600
400
200
0
count
No
Sí
pad

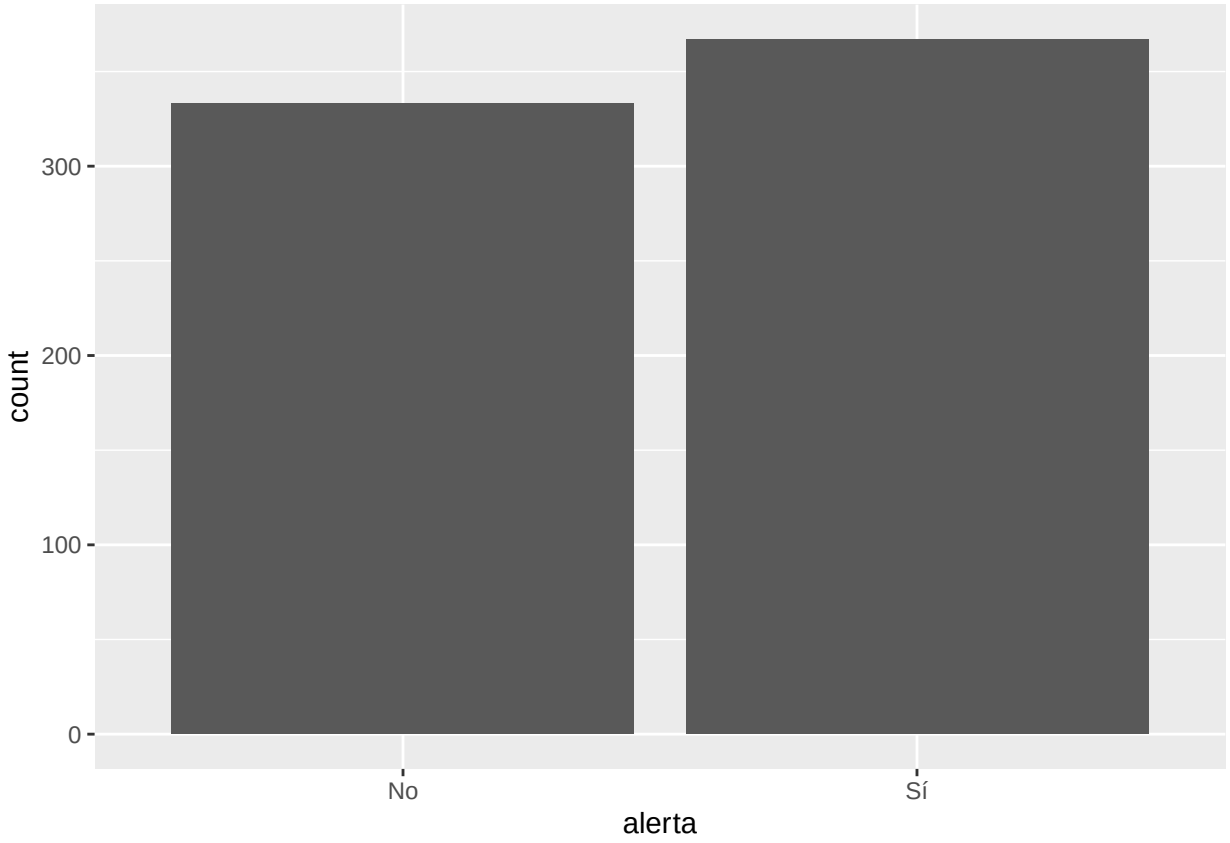

600
400
200
0
count
No
Sí
neuropath

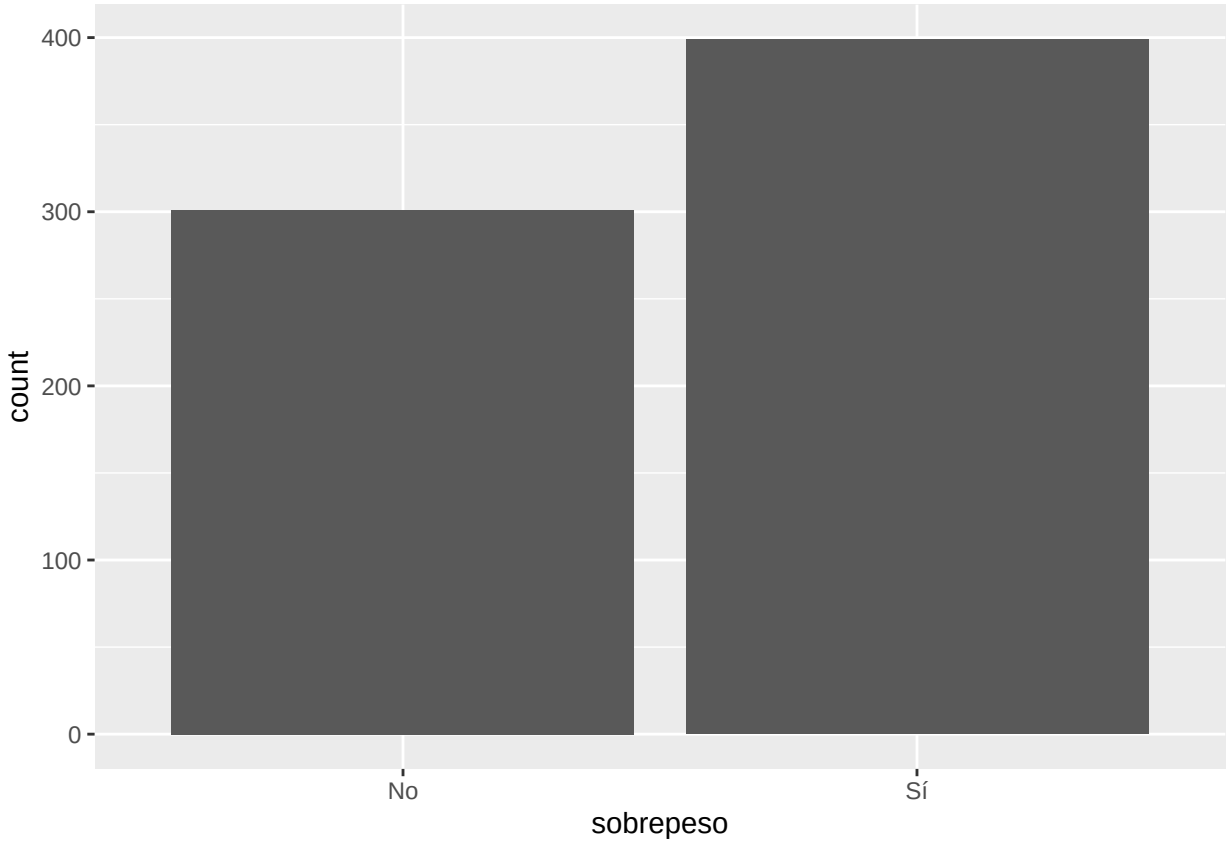

600
400
200
0
count
No
Sí
renal

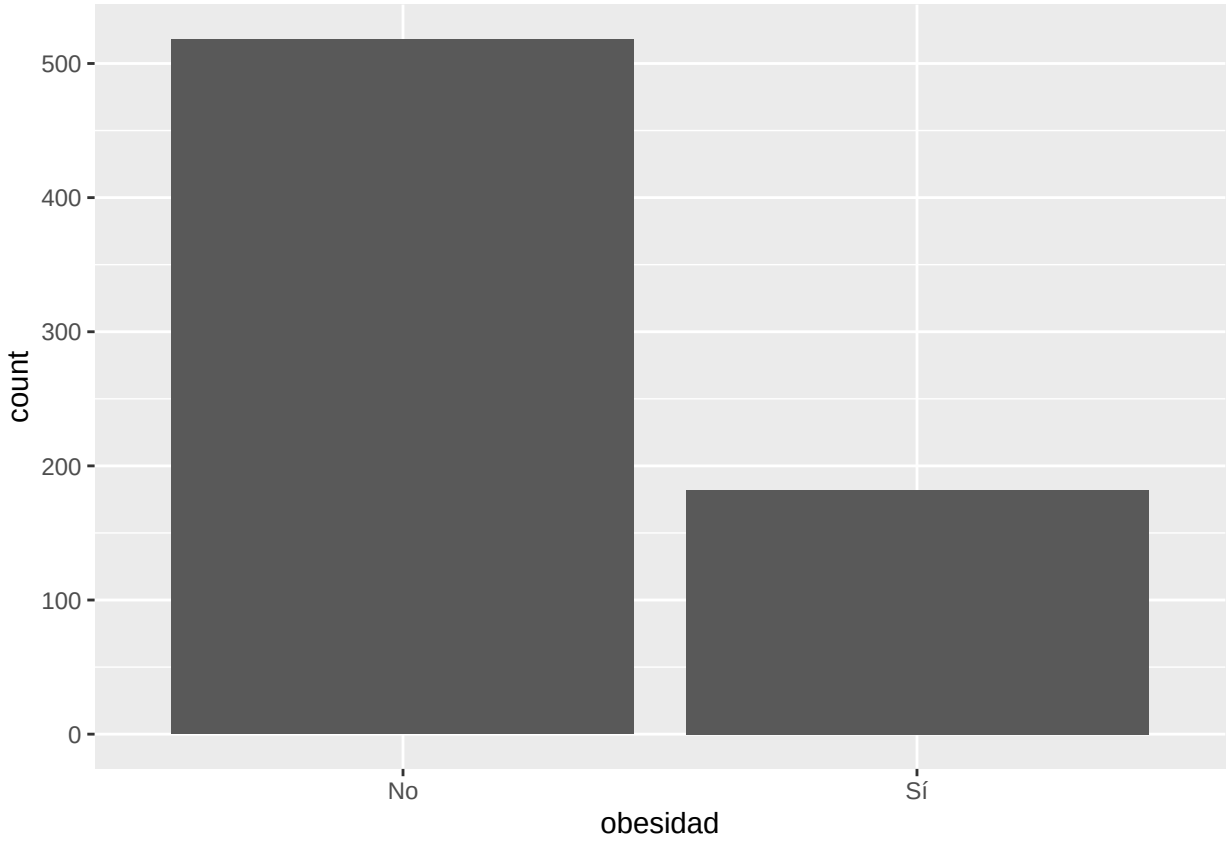

600
400
200
0
count
No
Sí
dn


400
200
0
count
No
Sí
dr


600
400
200
0
count
No
Sí
othcomp


300
200
100
0
count
No
Sí
alerta

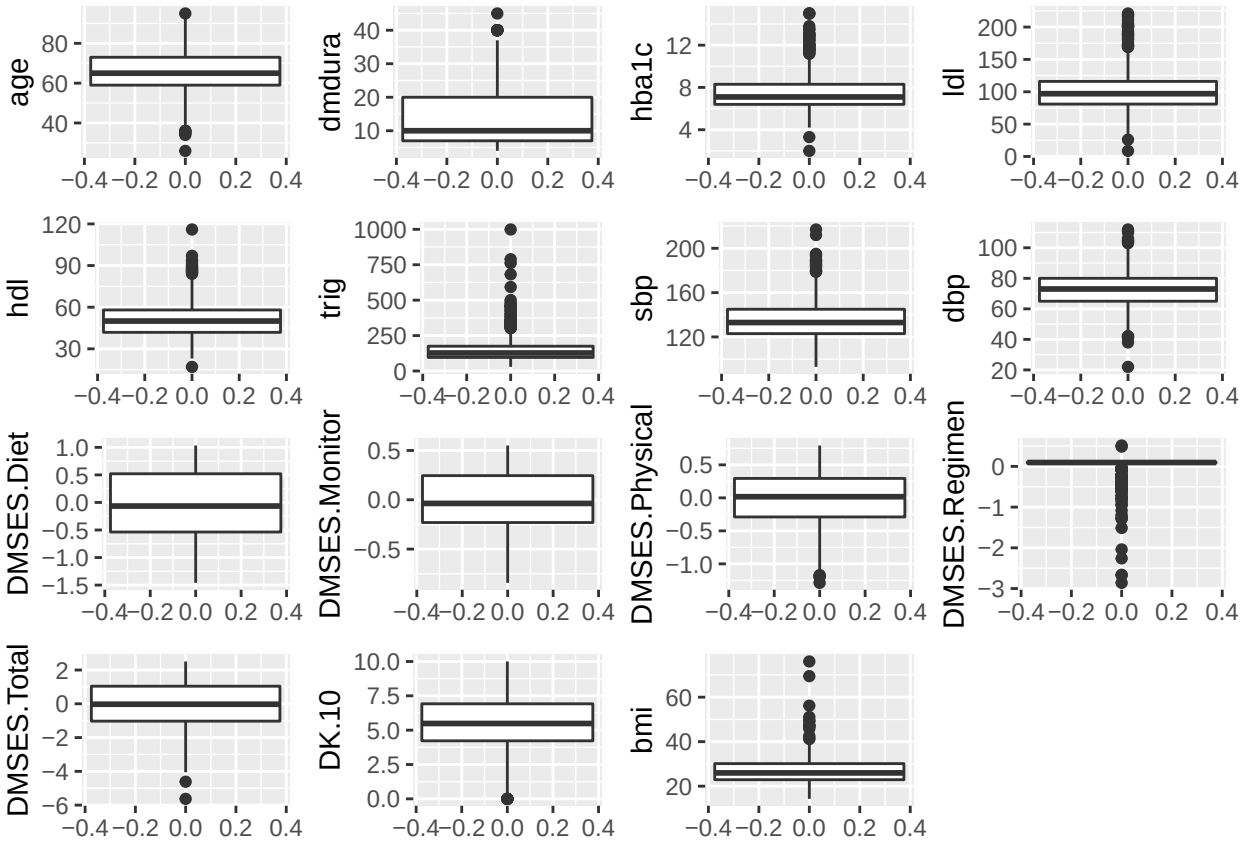

400
300
200
100
0
count
No
Sí
sobrepeso



Analizamos con un poco más de detalle las variables continuas.

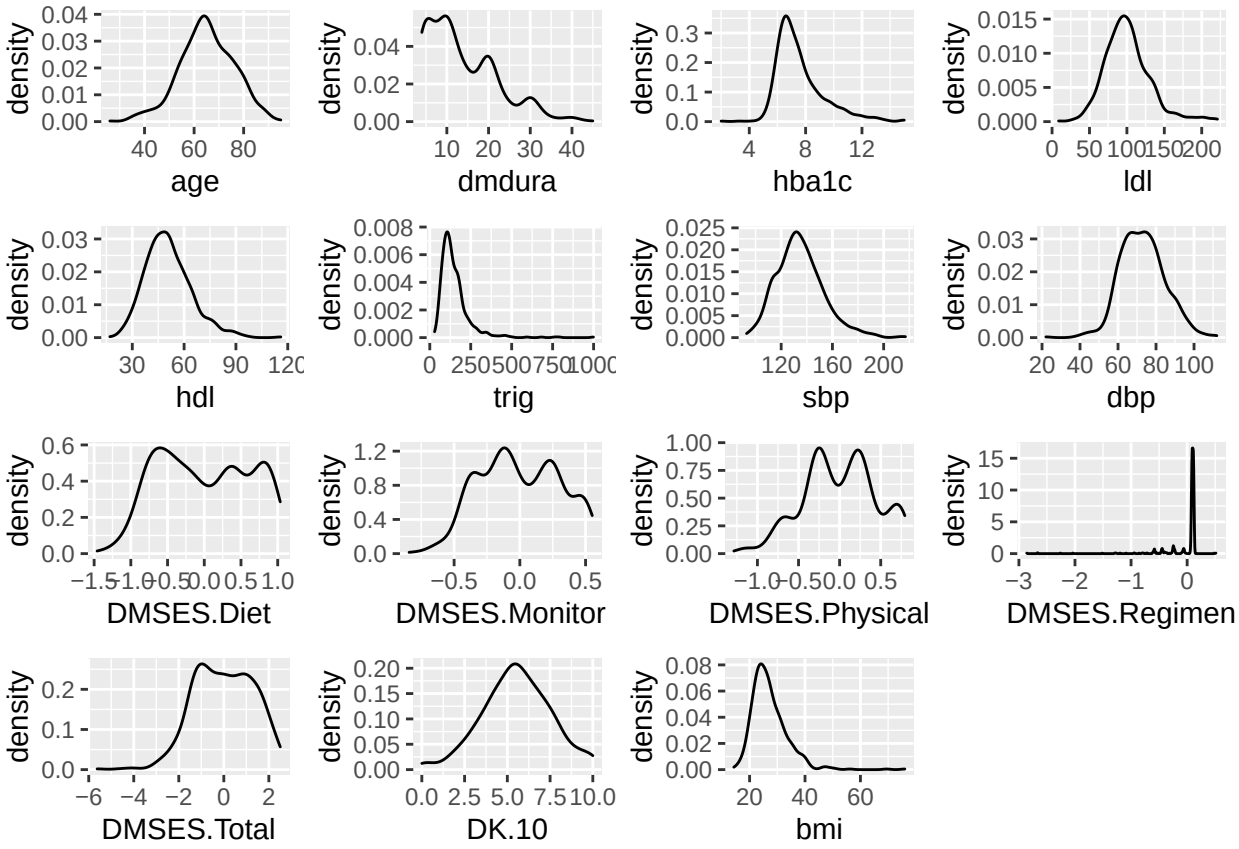
- Boxplot (con outliers)

```
bx_age            <- ggplot(datos,aes(y=age)) + geom_boxplot()
bx_dmdura         <- ggplot(datos,aes(y=dmdura)) + geom_boxplot()
bx_hba1c          <- ggplot(datos,aes(y=hba1c)) + geom_boxplot()
bx_ldl            <- ggplot(datos,aes(y=ldl)) + geom_boxplot()
bx_hdl            <- ggplot(datos,aes(y=hdl)) + geom_boxplot()
bx_trig           <- ggplot(datos,aes(y=trig)) + geom_boxplot()
bx_sbp            <- ggplot(datos,aes(y=sbp)) + geom_boxplot()
bx_dbp            <- ggplot(datos,aes(y=dbp)) + geom_boxplot()
bx_DMSES.Diet     <- ggplot(datos,aes(y=DMSES.Diet)) + geom_boxplot()
bx_DMSES.Monitor  <- ggplot(datos,aes(y=DMSES.Monitor)) + geom_boxplot()
bx_DMSES.Physical <- ggplot(datos,aes(y=DMSES.Physical)) + geom_boxplot()
bx_DMSES.Regimen  <- ggplot(datos,aes(y=DMSES.Regimen)) + geom_boxplot()
bx_DMSES.Total    <- ggplot(datos,aes(y=DMSES.Total)) + geom_boxplot()
bx_DK.10          <- ggplot(datos,aes(y=DK.10)) + geom_boxplot()
bx_bmi            <- ggplot(datos,aes(y=bmi)) + geom_boxplot()

grid.arrange(bx_age,bx_dmdura,bx_hba1c,bx_ldl,bx_hdl,bx_trig,bx_sbp,bx_dbp,
             bx_DMSES.Diet, bx_DMSES.Monitor, bx_DMSES.Physical, bx_DMSES.Regimen,
             bx_DMSES.Total, bx_DK.10, bx_bmi,ncol=4)
```



- Densidades

```
den_age             <- ggplot(datos,aes(x=age)) + geom_density()
den_dmdura          <- ggplot(datos,aes(x=dmdura)) + geom_density()
den_hba1c           <- ggplot(datos,aes(x=hba1c)) + geom_density()
den_ldl             <- ggplot(datos,aes(x=ldl)) + geom_density()
den_hdl             <- ggplot(datos,aes(x=hdl)) + geom_density()
den_trig            <- ggplot(datos,aes(x=trig)) + geom_density()
den_sbp             <- ggplot(datos,aes(x=sbp)) + geom_density()
den_dbp             <- ggplot(datos,aes(x=dbp)) + geom_density()
den_DMSES.Diet      <- ggplot(datos,aes(x=DMSES.Diet)) + geom_density()
den_DMSES.Monitor   <- ggplot(datos,aes(x=DMSES.Monitor)) + geom_density()
den_DMSES.Physical  <- ggplot(datos,aes(x=DMSES.Physical)) + geom_density()
den_DMSES.Regimen   <- ggplot(datos,aes(x=DMSES.Regimen)) + geom_density()
den_DMSES.Total     <- ggplot(datos,aes(x=DMSES.Total)) + geom_density()
den_DK.10           <- ggplot(datos,aes(x=DK.10)) + geom_density()
den_bmi             <- ggplot(datos,aes(x=bmi)) + geom_density()

grid.arrange(den_age,den_dmdura,den_hba1c,den_ldl,den_hdl,den_trig,den_sbp,den_dbp,
             den_DMSES.Diet, den_DMSES.Monitor, den_DMSES.Physical, den_DMSES.Regimen,
             den_DMSES.Total, den_DK.10, den_bmi,ncol=4)
```



- Normalidad

Ningun predictor sigue una distribución normal según el test de normalidad de Shapiro y según el QQ plot tenemos normalidad en edad, hdl, sbp, dbp, DMSES.Total y DK.10.

```
shapiro.test(datos$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$age
## W = 0.99459, p-value = 0.01385
```

```
shapiro.test(datos$dmdura)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$dmdura
## W = 0.90381, p-value < 2.2e-16
```

```
shapiro.test(datos$hba1c)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$hba1c
## W = 0.88872, p-value < 2.2e-16
```

```
shapiro.test(datos$ldl)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$ldl
## W = 0.95629, p-value = 1.435e-13
```

```
shapiro.test(datos$hdl)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$hdl
## W = 0.97306, p-value = 4.687e-10
```

```
shapiro.test(datos$trig)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$trig
## W = 0.73709, p-value < 2.2e-16
```

```
shapiro.test(datos$sbp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$sbp
## W = 0.97534, p-value = 1.774e-09
```

```
shapiro.test(datos$dbp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$dbp
## W = 0.99086, p-value = 0.0002488
```

```
shapiro.test(datos$DMSES.Diet)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$DMSES.Diet
## W = 0.94977, p-value = 1.089e-14
```

```
shapiro.test(datos$DMSES.Monitor)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$DMSES.Monitor
## W = 0.97283, p-value = 4.121e-10
```

```
shapiro.test(datos$DMSES.Physical)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$DMSES.Physical
## W = 0.9796, p-value = 2.638e-08
```

```
shapiro.test(datos$DMSES.Regimen)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$DMSES.Regimen
## W = 0.41931, p-value < 2.2e-16
```

```
shapiro.test(datos$DMSES.Total)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$DMSES.Total
## W = 0.98158, p-value = 1.038e-07
```

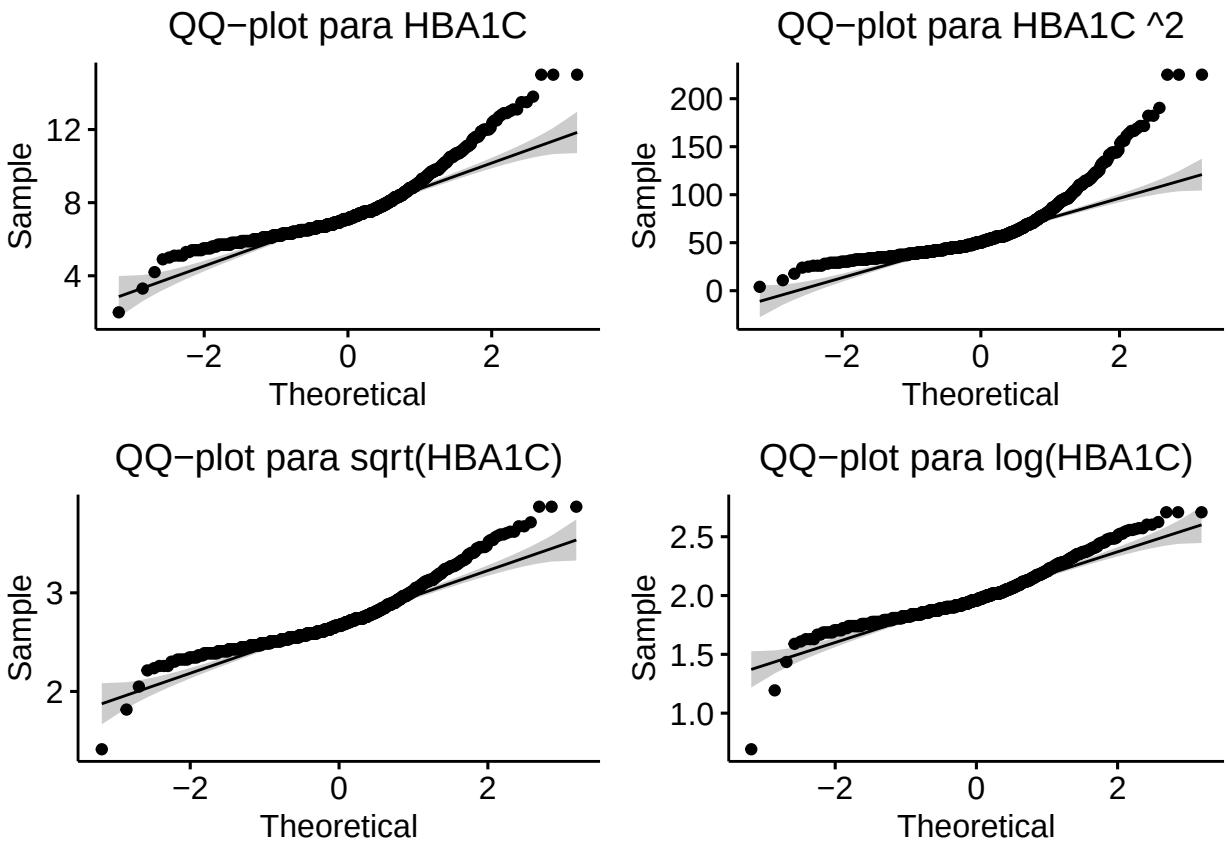
```
shapiro.test(datos$DK.10)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$DK.10
## W = 0.99402, p-value = 0.007196
```

```
shapiro.test(datos$bmi)
```

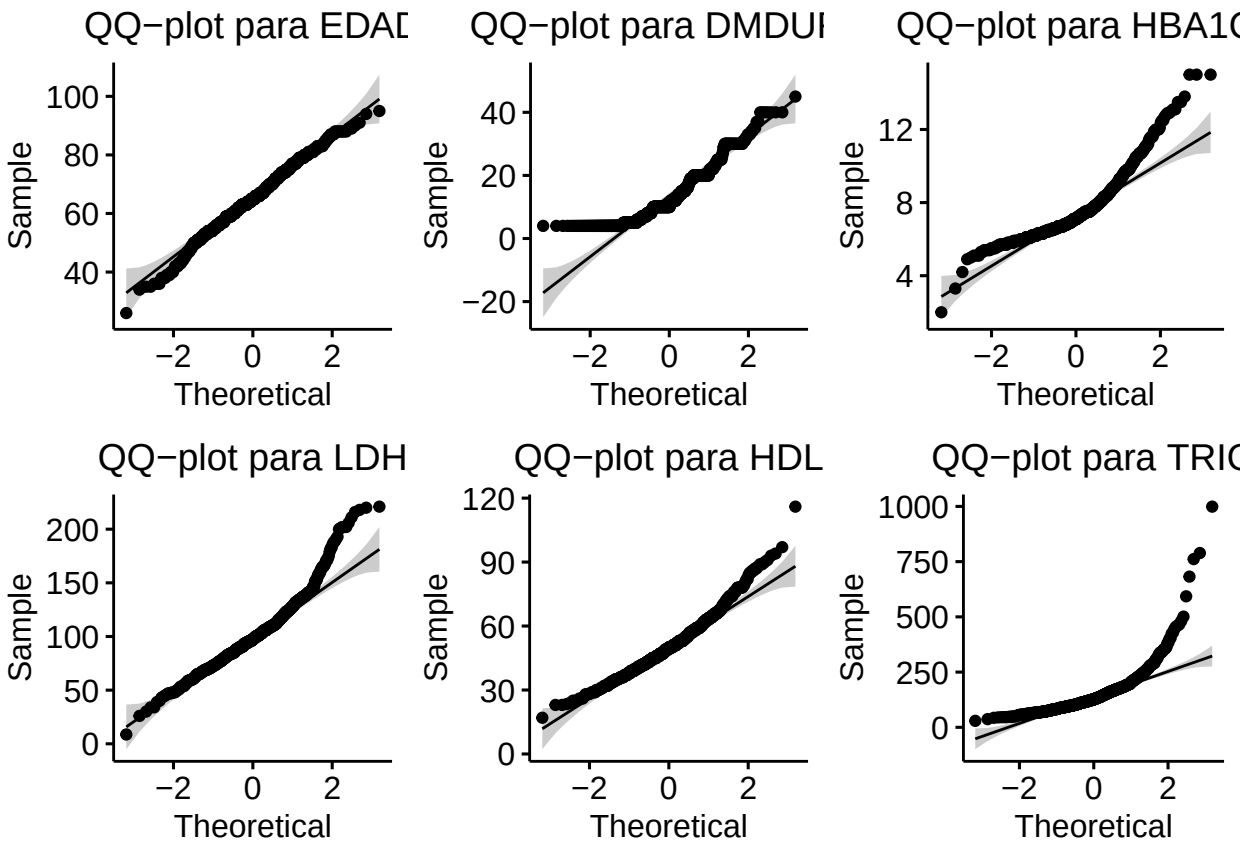
```
##
##  Shapiro-Wilk normality test
##
## data:  datos$bmi
## W = 0.88978, p-value < 2.2e-16
```

```
#ggqqplot(datos[,vars_cont],combine = TRUE,facet.by=c(5,3))

qq_hba1c          <- ggqqplot(datos$hba1c, main = "QQ-plot para HBA1C") + theme(plot.title = element_tex
qq_hba1c2          <- ggqqplot((datos$hba1c)**2, main = "QQ-plot para HBA1C ^2") + theme(plot.title = el
qq_hba1c3          <- ggqqplot(sqrt(datos$hba1c), main = "QQ-plot para sqrt(HBA1C)") + theme(plot.title
qq_hba1c4          <- ggqqplot(log(datos$hba1c), main = "QQ-plot para log(HBA1C)") + theme(plot.title =

grid.arrange(qq_hba1c,qq_hba1c2,qq_hba1c3,qq_hba1c4, ncol=2)
```

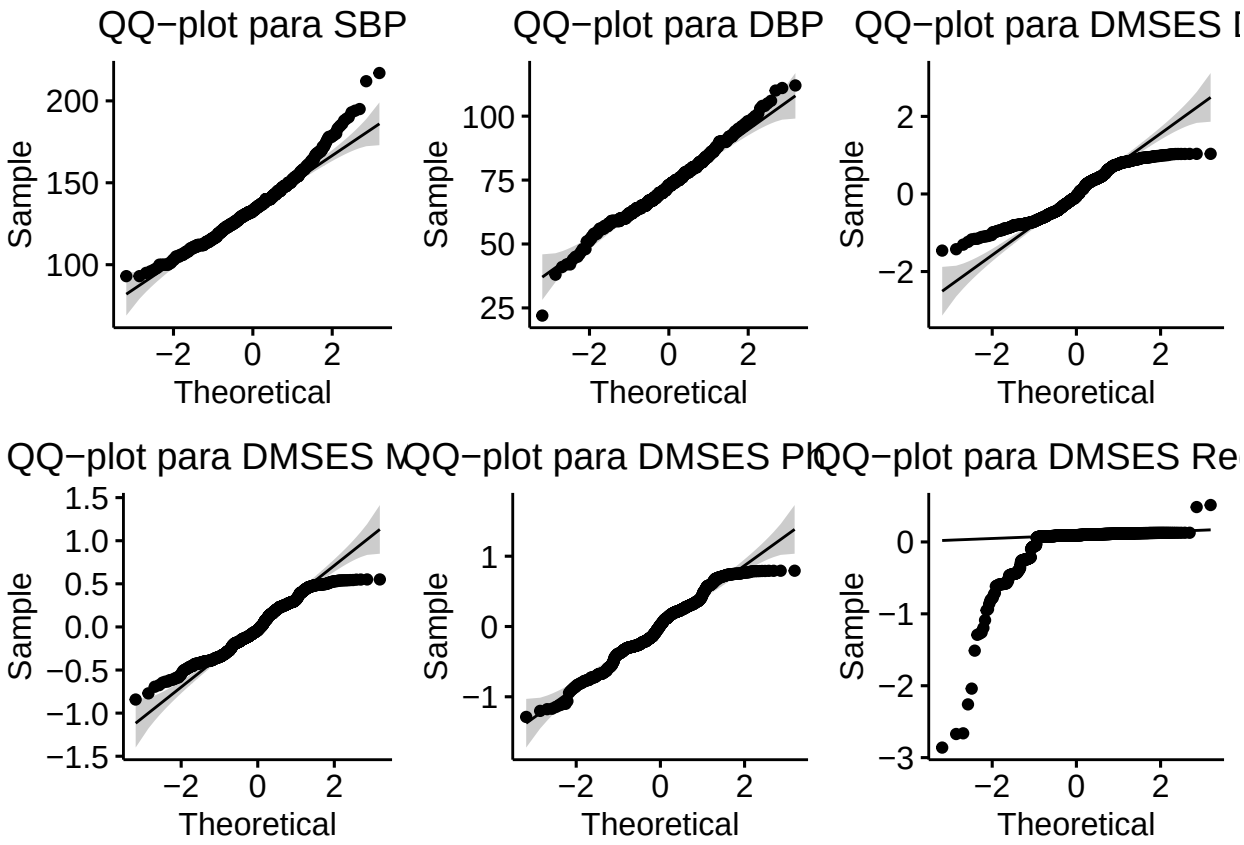



```
qq_age            <- ggqqplot(datos$age, main = "QQ-plot para EDAD") + theme(plot.title = element_text(
qq_dmdura         <- ggqqplot(datos$dmdura, main = "QQ-plot para DMDURA") + theme(plot.title = element_
qq_ldl            <- ggqqplot(datos$ldl, main = "QQ-plot para LDH") + theme(plot.title = element_text(h
qq_hdl            <- ggqqplot(datos$hdl, main = "QQ-plot para HDL") + theme(plot.title = element_text(h
qq_trig           <- ggqqplot(datos$trig, main = "QQ-plot para TRIG") + theme(plot.title = element_text
qq_sbp            <- ggqqplot(datos$sbp, main = "QQ-plot para SBP") + theme(plot.title = element_text(h
qq_dbp            <- ggqqplot(datos$dbp, main = "QQ-plot para DBP") + theme(plot.title = element_text(h
qq_DMSES.Diet     <- ggqqplot(datos$DMSES.Diet, main = "QQ-plot para DMSES Diet") + theme(plot.title =
qq_DMSES.Monitor  <- ggqqplot(datos$DMSES.Monitor, main = "QQ-plot para DMSES Monitor") + theme(plot.ti
qq_DMSES.Physical <- ggqqplot(datos$DMSES.Physical, main = "QQ-plot para DMSES Physical") + theme(plot.
```

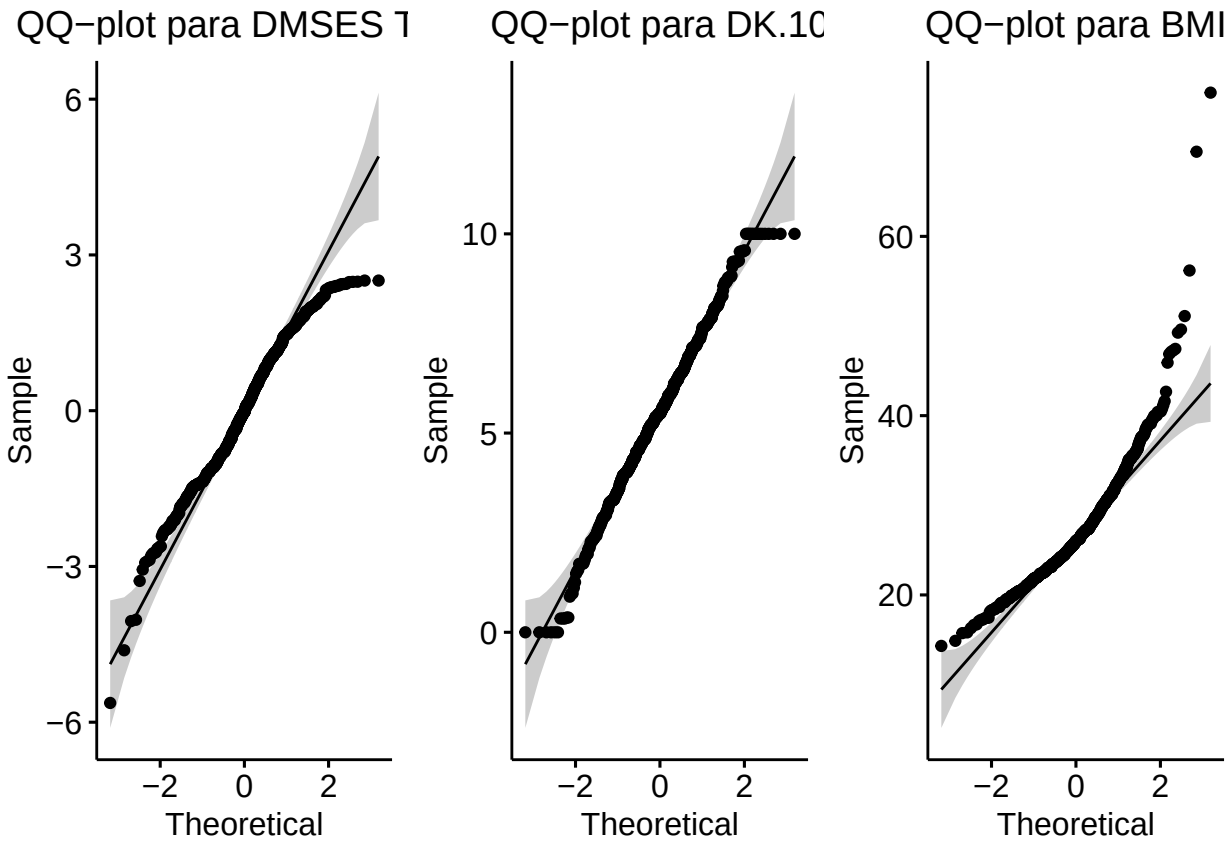
```
qq_DMSES.Regimen  <- ggqqplot(datos$DMSES.Regimen, main = "QQ-plot para DMSES Regimen") + theme(plot.ti
qq_DMSES.Total    <- ggqqplot(datos$DMSES.Total, main = "QQ-plot para DMSES Total") + theme(plot.title =
qq_DK.10          <- ggqqplot(datos$DK.10, main = "QQ-plot para DK.10") + theme(plot.title = element_tex
qq_bmi            <- ggqqplot(datos$bmi, main = "QQ-plot para BMI") + theme(plot.title = element_text(h

#Arrange all the plots onto one page

#plot_grid(qq_age, qq_dmdura, qq_hba1c, qq_ldl, qq_hdl, qq_trig,
#          qq_sbp, qq_dbp, qq_DMSES.Diet, qq_DMSES.Monitor,
#          qq_DMSES.Physical, qq_DMSES.Regimen, qq_DMSES.Total,
#          qq_DK.10, qq_bmi, nrow=5,ncol=3)

#grid.arrange(qq_age, qq_dmdura, qq_hba1c, qq_ldl, qq_hdl, qq_trig,
#             qq_sbp, qq_dbp, qq_DMSES.Diet, qq_DMSES.Monitor,
#             qq_DMSES.Physical, qq_DMSES.Regimen, qq_DMSES.Total,
#             qq_DK.10, qq_bmi, ncol=2)


grid.arrange(qq_age, qq_dmdura, qq_hba1c, qq_ldl, qq_hdl, qq_trig, ncol=3)
```




```
grid.arrange(qq_sbp, qq_dbp, qq_DMSES.Diet, qq_DMSES.Monitor,
             qq_DMSES.Physical, qq_DMSES.Regimen, ncol = 3)
```




```
grid.arrange(qq_DMSES.Total,qq_DK.10, qq_bmi, ncol=3)
```



1.3 Exploración multidimensional

Los niveles de correlación entre las puntuaciones DMSES y el resto de variables continuas son bajos a excepción de la variable hba1c. La mayor correlación entre hba1c y las puntuaciones DMSES se da con el factor dieta. Se detecta que las diferentes puntuaciones del cuestionario DMSES estan fuertemente relacionas, por lo que se puede considerar tomar sólo una para el análisis. La puntuación total y la de la parte de la dieta son las mejoras candidatas debido a que tienen una correlación más alta con hba1c.

```
dim(datos[,vars_cont])
```

```
## [1] 700  14
```

```
str(datos[,vars_cont])
```

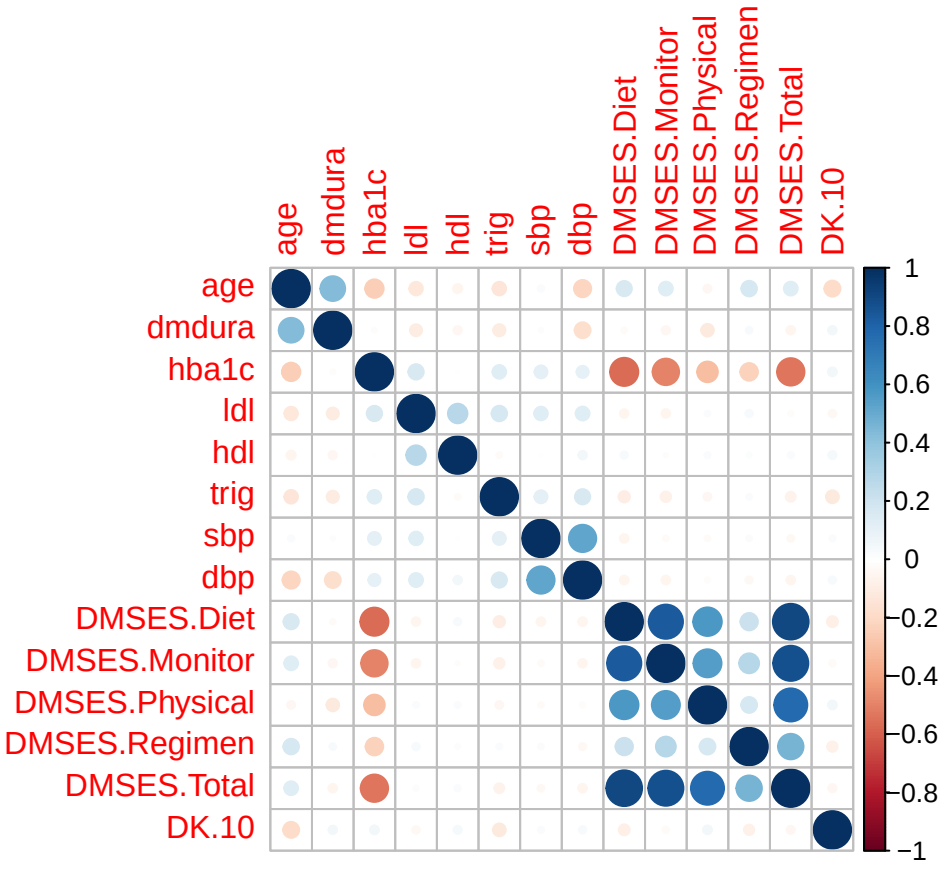
```
## 'data.frame':    700 obs. of  14 variables:
##  $ age           : int  51 55 47 54 71 55 59 64 58 69 ...
##  $ dmdura        : int  4 4 5 20 10 4 4 22 4 14 ...
##  $ hba1c         : num  9.7 7.1 8 7.2 8.8 8 9.8 6.3 6.7 7.7 ...
##  $ ldl           : num  123.7 89.8 119 202.1 58.6 ...
##  $ hdl           : num  45.2 49.5 50.7 47 35.7 35.3 42 44.9 46 35.6 ...
##  $ trig          : num  248.9 276.9 165.6 84.7 149.5 ...
##  $ sbp           : int  130 120 168 121 118 119 119 150 124 169 ...
##  $ dbp           : int  80 70 94 56 65 63 83 76 76 93 ...
##  $ DMSES.Diet    : num  -0.132233 -0.000414 1.031174 -1.102578 -1.064826 ...
```

```
##  $ DMSES.Monitor : num  -0.133 0.158 0.527 -0.168 -0.37 ...
##  $ DMSES.Physical: num  -0.1775 0.09218 0.79015 -0.00571 0.02158 ...
##  $ DMSES.Regimen : num  0.0909 0.106 -0.0232 0.0916 0.0824 ...
##  $ DMSES.Total   : num  -0.352 0.355 2.325 -1.184 -1.331 ...
##  $ DK.10         : num  5.77 6.66 4.27 5.19 5.65 ...
```

```
M <- cor(datos[,c(vars_cont)])
knitr::kable(round(M,2))
```

	age	dmdura	hba1c	ldl	hdl	trig	sbp	dbp	DMSES.Diet	DMSES.Monitor	DI
age	1.00	0.44	-0.24	-0.13	-0.06	-0.13	0.03	-0.21	0.16	0.14	
dmdura	0.44	1.00	-0.02	-0.10	-0.05	-0.11	-0.01	-0.18	-0.02	-0.05	
hba1c	-0.24	-0.02	1.00	0.17	0.00	0.13	0.12	0.11	-0.56	-0.50	
ldl	-0.13	-0.10	0.17	1.00	0.27	0.17	0.13	0.14	-0.05	-0.05	
hdl	-0.06	-0.05	0.00	0.27	1.00	-0.02	0.00	0.05	0.04	-0.01	
trig	-0.13	-0.11	0.13	0.17	-0.02	1.00	0.12	0.17	-0.09	-0.08	
sbp	0.03	-0.01	0.12	0.13	0.00	0.12	1.00	0.53	-0.05	-0.02	
dbp	-0.21	-0.18	0.11	0.14	0.05	0.17	0.53	1.00	-0.05	-0.05	
DMSES.Diet	0.16	-0.02	-0.56	-0.05	0.04	-0.09	-0.05	-0.05	1.00	0.83	
DMSES.Monitor	0.14	-0.05	-0.50	-0.05	-0.01	-0.08	-0.02	-0.05	0.83	1.00	
DMSES.Physical	-0.04	-0.11	-0.30	0.02	0.02	-0.04	-0.02	-0.02	0.58	0.56	
DMSES.Regimen	0.17	0.03	-0.22	0.04	0.02	0.02	0.02	-0.04	0.22	0.28	
DMSES.Total	0.13	-0.05	-0.53	-0.02	0.03	-0.07	-0.03	-0.05	0.91	0.88	
DK.10	-0.18	0.05	0.06	-0.04	0.05	-0.12	0.03	0.04	-0.08	-0.03	

```
corrplot(M,method="circle")
```



```r
# vamos a ordenar las correlaciones
flattenCorrMatrix <- function(cormat) {
  ut <- upper.tri(cormat)
  dd <- data.frame(
                   row = rownames(cormat)[row(cormat)[ut]],
                   column = rownames(cormat)[col(cormat)[ut]],
                   cor  =(cormat)[ut]
                   )
  dd <- dd[order(abs(dd$cor), decreasing = T),]
  dd$cor <- round(dd$cor, 3)
  dd
}

flattenCorrMatrix(M)
```

```
##                row          column    cor
## 75     DMSES.Diet    DMSES.Total  0.908
## 76   DMSES.Monitor    DMSES.Total  0.876
## 45     DMSES.Diet  DMSES.Monitor  0.832
## 77 DMSES.Physical    DMSES.Total  0.779
## 54     DMSES.Diet DMSES.Physical  0.578
## 31          hba1c     DMSES.Diet -0.563
## 55   DMSES.Monitor DMSES.Physical  0.558
## 69          hba1c    DMSES.Total -0.534
## 28            sbp            dbp  0.530
## 39          hba1c  DMSES.Monitor -0.499
## 78   DMSES.Regimen    DMSES.Total  0.461
## 1             age         dmdura  0.436
## 48          hba1c DMSES.Physical -0.303
## 65   DMSES.Monitor  DMSES.Regimen  0.281
## 10            ldl            hdl  0.271
## 2             age          hba1c -0.241
## 58          hba1c  DMSES.Regimen -0.221
## 64     DMSES.Diet  DMSES.Regimen  0.215
## 22            age            dbp -0.211
## 79            age          DK.10 -0.183
## 23         dmdura            dbp -0.179
## 66 DMSES.Physical  DMSES.Regimen  0.178
## 56            age  DMSES.Regimen  0.172
## 14            ldl           trig  0.170
## 6           hba1c            ldl  0.167
## 27           trig            dbp  0.165
## 29            age     DMSES.Diet  0.165
## 25            ldl            dbp  0.137
## 37            age  DMSES.Monitor  0.136
## 67            age    DMSES.Total  0.135
## 13          hba1c           trig  0.131
## 11            age           trig -0.131
## 19            ldl            sbp  0.131
## 4             age            ldl -0.129
## 21           trig            sbp  0.119
## 18          hba1c            sbp  0.118
## 84           trig          DK.10 -0.117
```

```
## 47         dmdura DMSES.Physical -0.113
## 24          hba1c            dbp  0.105
## 12         dmdura           trig -0.105
## 5          dmdura            ldl -0.102
## 34           trig     DMSES.Diet -0.090
## 87     DMSES.Diet          DK.10 -0.084
## 42           trig  DMSES.Monitor -0.078
## 90  DMSES.Regimen          DK.10 -0.076
## 72           trig    DMSES.Total -0.069
## 7             age            hdl -0.059
## 81          hba1c          DK.10  0.056
## 89 DMSES.Physical          DK.10  0.055
## 36            dbp     DMSES.Diet -0.054
## 40            ldl  DMSES.Monitor -0.053
## 74            dbp    DMSES.Total -0.053
## 35            sbp     DMSES.Diet -0.053
## 44            dbp  DMSES.Monitor -0.052
## 68         dmdura    DMSES.Total -0.052
## 26            hdl            dbp  0.051
## 32            ldl     DMSES.Diet -0.051
## 80         dmdura          DK.10  0.051
## 38         dmdura  DMSES.Monitor -0.049
## 8          dmdura            hdl -0.048
## 91    DMSES.Total          DK.10 -0.046
## 83            hdl          DK.10  0.046
## 46            age DMSES.Physical -0.043
## 51           trig DMSES.Physical -0.043
## 82            ldl          DK.10 -0.040
## 59            ldl  DMSES.Regimen  0.038
## 86            dbp          DK.10  0.038
## 33            hdl     DMSES.Diet  0.037
## 63            dbp  DMSES.Regimen -0.036
## 73            sbp    DMSES.Total -0.032
## 57         dmdura  DMSES.Regimen  0.031
## 88  DMSES.Monitor          DK.10 -0.030
## 16            age            sbp  0.028
## 71            hdl    DMSES.Total  0.027
## 85            sbp          DK.10  0.026
## 50            hdl DMSES.Physical  0.025
## 61           trig  DMSES.Regimen  0.024
## 15            hdl           trig -0.023
## 49            ldl DMSES.Physical  0.023
## 62            sbp  DMSES.Regimen  0.022
## 43            sbp  DMSES.Monitor -0.022
## 52            sbp DMSES.Physical -0.021
## 30         dmdura     DMSES.Diet -0.021
## 53            dbp DMSES.Physical -0.020
## 70            ldl    DMSES.Total -0.020
## 60            hdl  DMSES.Regimen  0.017
## 3          dmdura          hba1c -0.016
## 17         dmdura            sbp -0.012
## 41            hdl  DMSES.Monitor -0.012
## 20            hdl            sbp -0.005
## 9           hba1c            hdl  0.003
```

```r
cor(hba1c,DMSES.Diet)
```

```
## [1] -0.5630537
```

```r
cor(hba1c,DMSES.Physical)
```

```
## [1] -0.3027804
```

```r
cor(hba1c,DMSES.Monitor)
```

```
## [1] -0.4987917
```

```r
cor(hba1c,DMSES.Regimen)
```

```
## [1] -0.2206762
```

```r
cor(hba1c,DMSES.Total)
```

```
## [1] -0.5339001
```

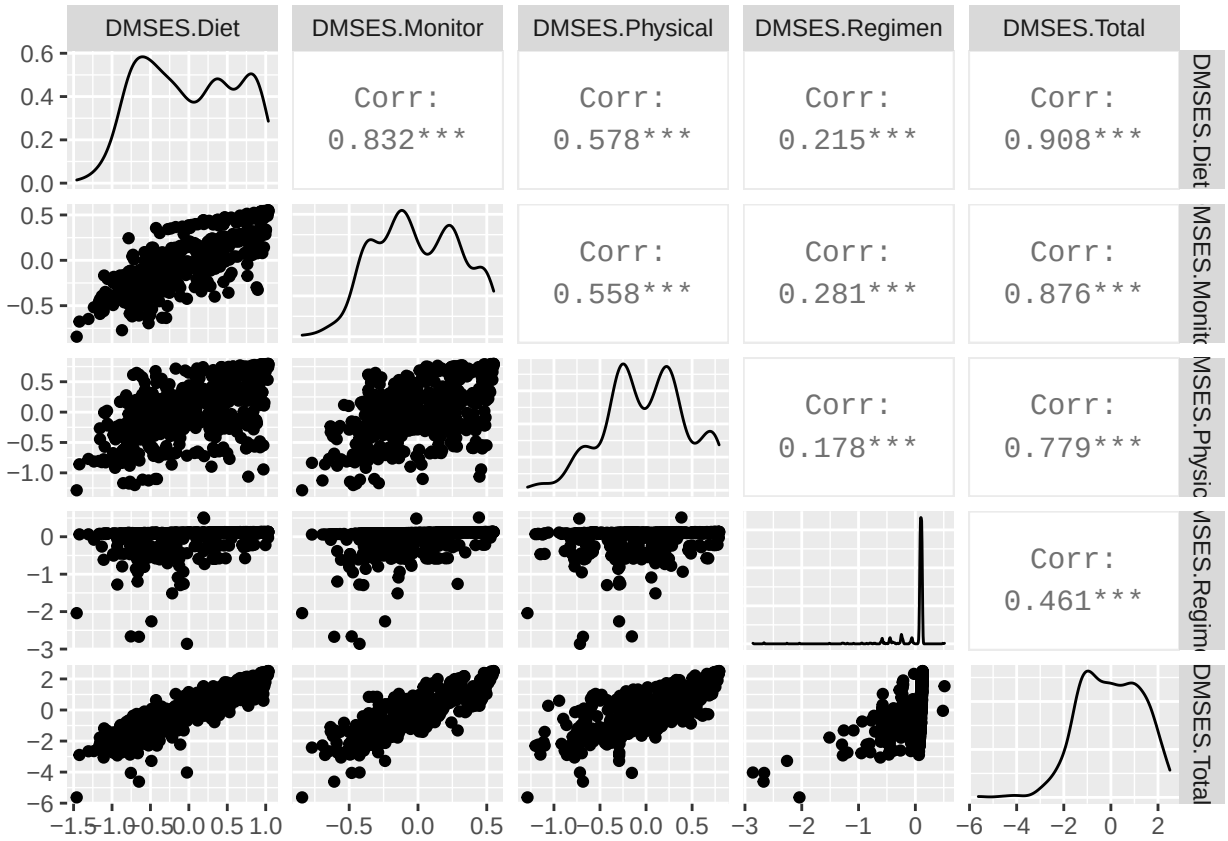
Podemos ver que las diferentes puntuaciones del cuestionario DMSES (total o de alguna de sus partes) están muy relacionadas. De entre todas las puntuaciones, la del factor dieta es la que mayor correlación tienen con la hba1c, por lo que parece razonable centrarse en el uso de sólo esa puntuación para buscar una relación con la hba1c y descartar el resto de puntuaciones del cuestionario DMSES del análisis.

Vemos en que medida están relacionadas las puntuaciones del cuestiorio DMSES entre sí.

```r
# demostracion DMSES total es la suma de los 4 subscores
max(DMSES.Total - (DMSES.Diet + DMSES.Monitor + DMSES.Physical + DMSES.Regimen))
```

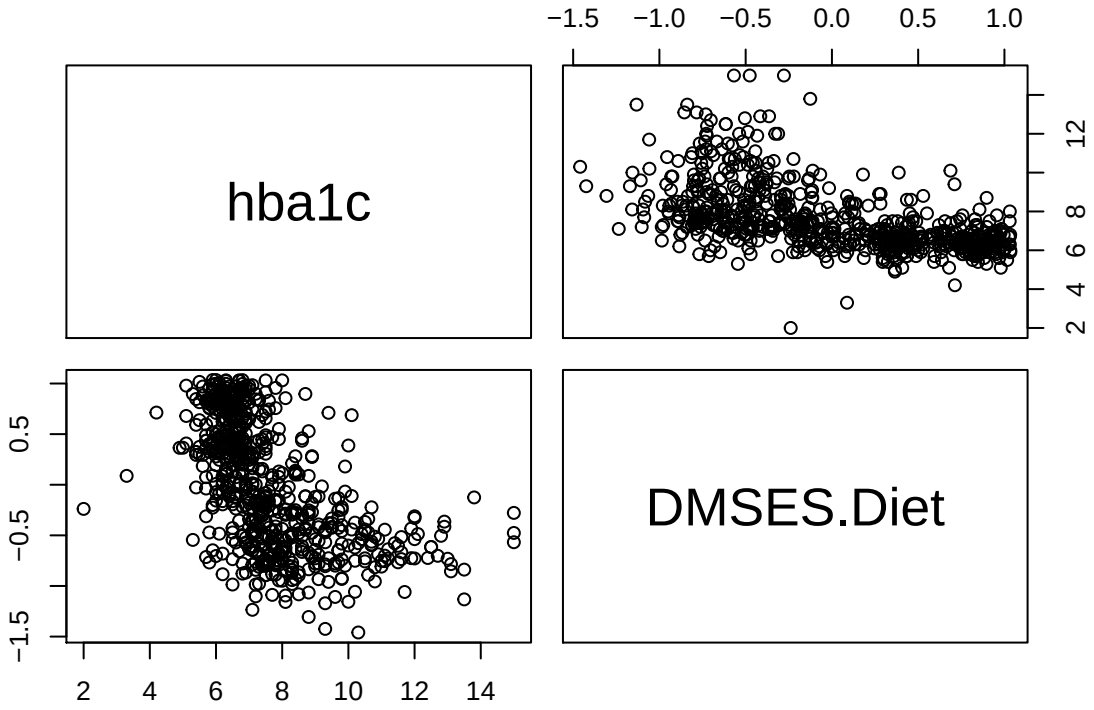
```
## [1] 7.105427e-15
```

```r
# relación entre los diferentes subscores
ggpairs(datos[,c("DMSES.Diet","DMSES.Monitor","DMSES.Physical","DMSES.Regimen","DMSES.Total")])
```

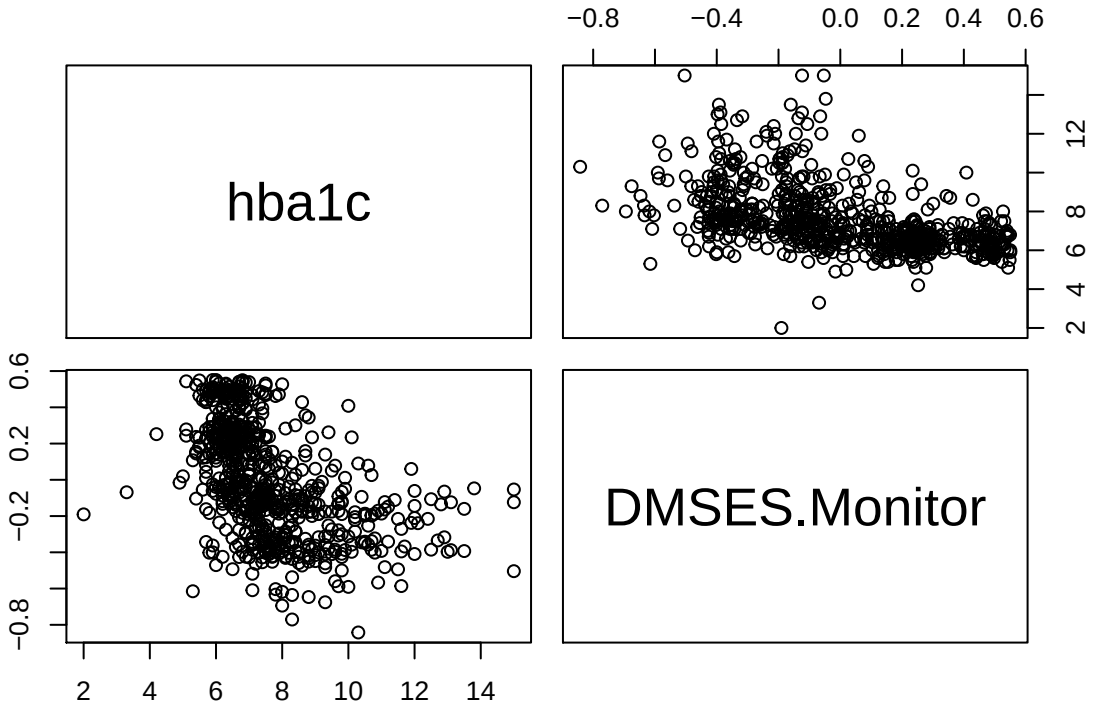


Al querer buscar las relaciones entre las puntuaciones del cuestionario DMSES (la puntuación total y la de las partes 'dieta', 'monitor', 'régimen' y 'estado físico') y el nivel de hba1c se observa que no hay una relación lineal a priori. Se valorará hacer transformaciones adecuadas para comprobar si se puede hallar una relación.

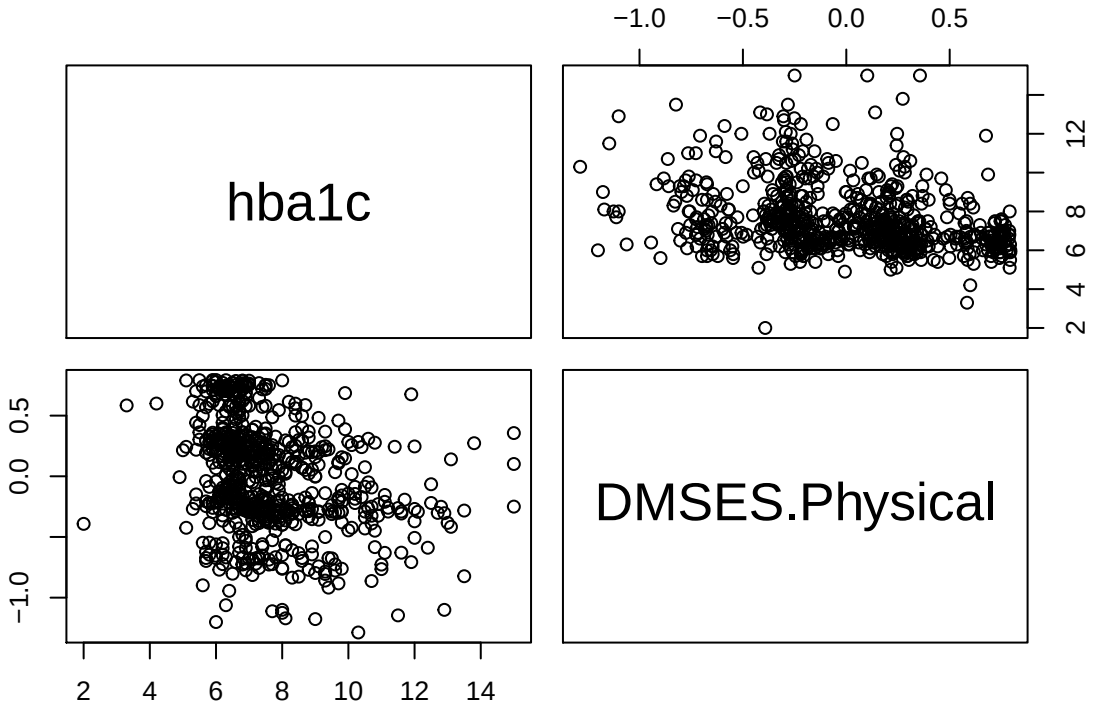
```
par(mfrow=c(2,3))
pairs(~ hba1c + DMSES.Diet ,data=datos)
```



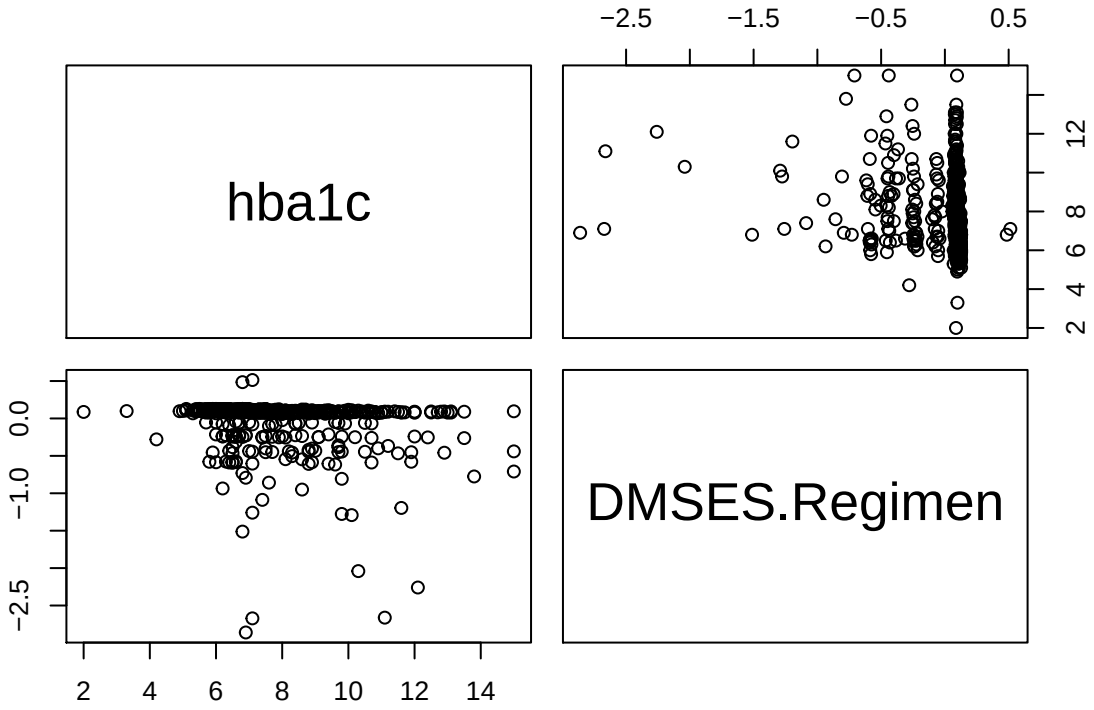

```
pairs(~ hba1c + DMSES.Monitor,data=datos)
```



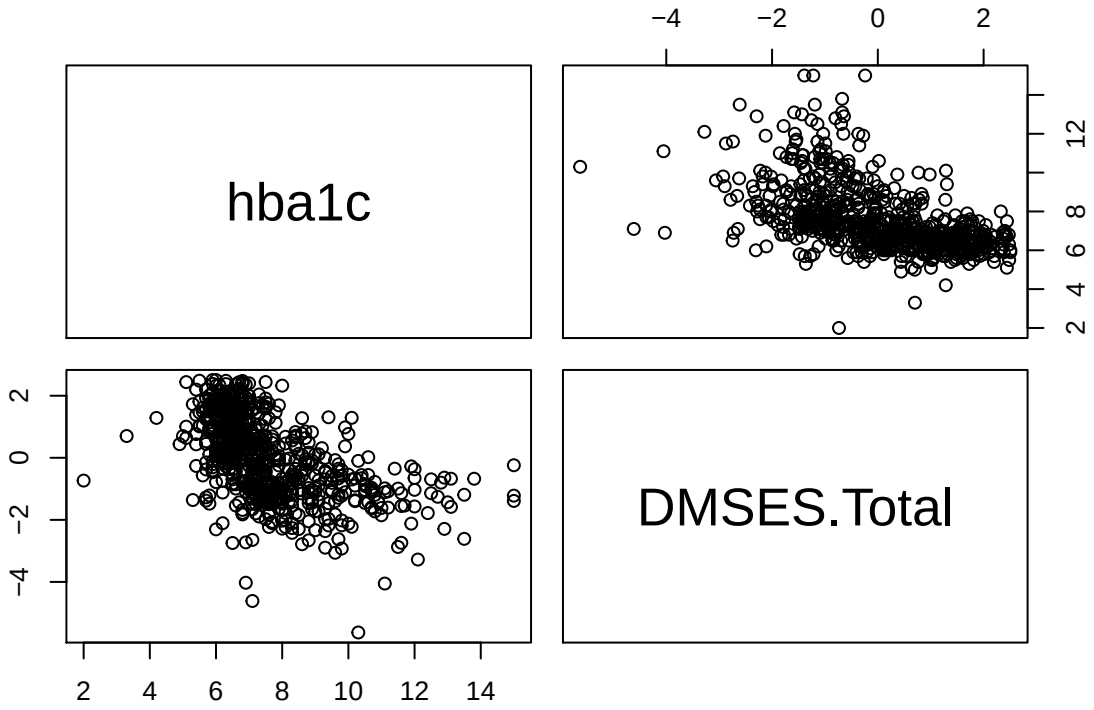

```
pairs(~ hba1c + DMSES.Physical,data=datos)
```




```
pairs(~ hba1c + DMSES.Regimen,data=datos)
```



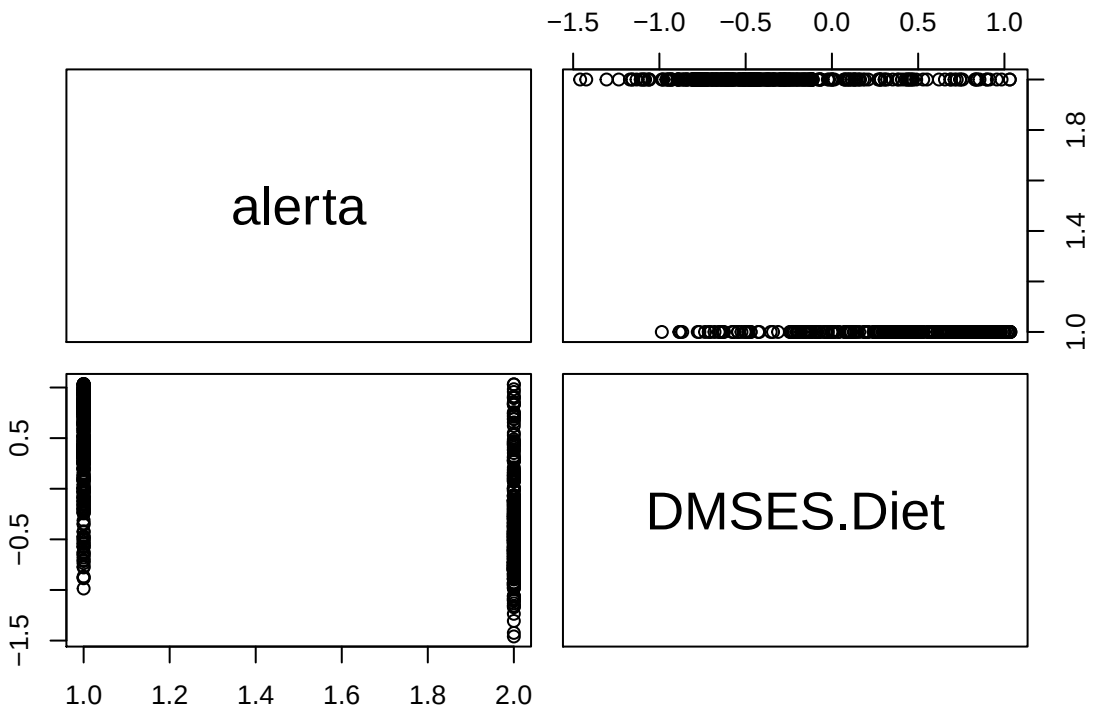

```
pairs(~ hba1c + DMSES.Total,data=datos)
```



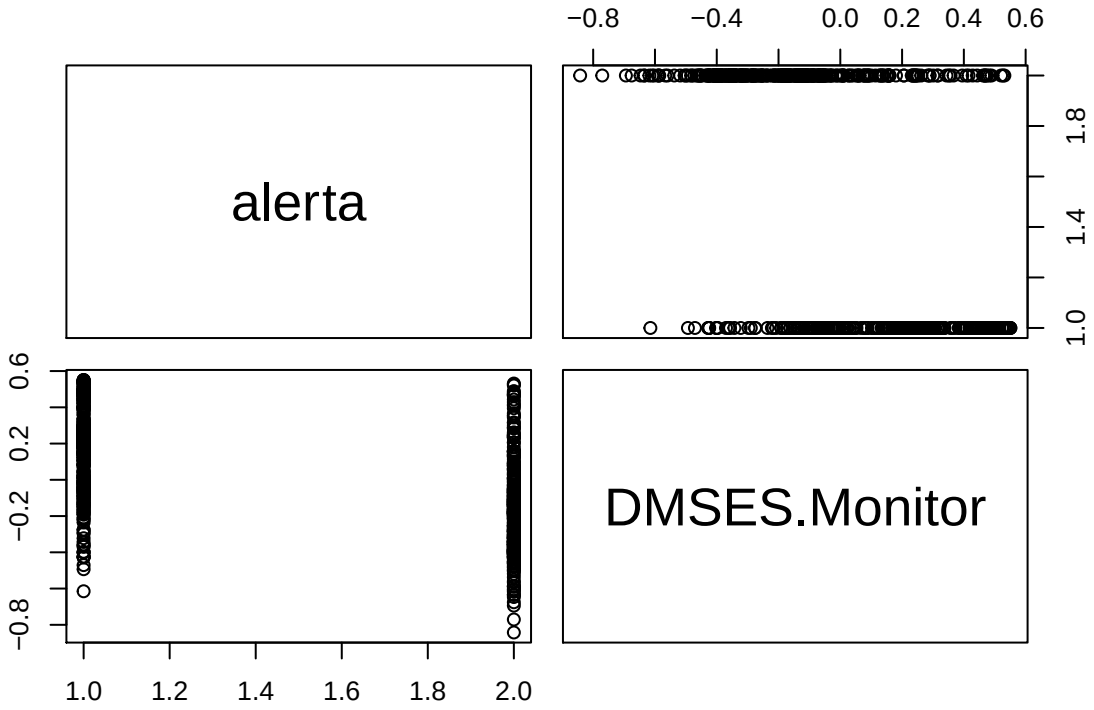

```
par(mfrow=c(1,1))
```

Tampoco se detecta una relación clara entre las puntuaciones del cuestionario DMSES y la variable categórica 'alerta'.

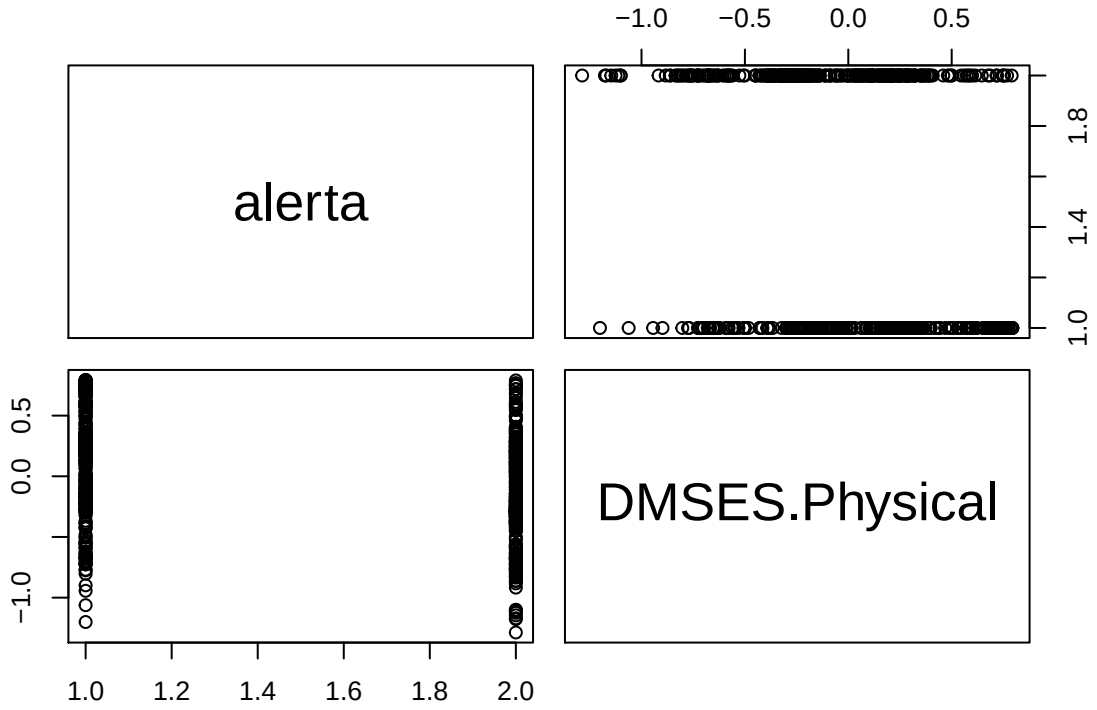
```
pairs(~ alerta + DMSES.Diet ,data=datos)
```



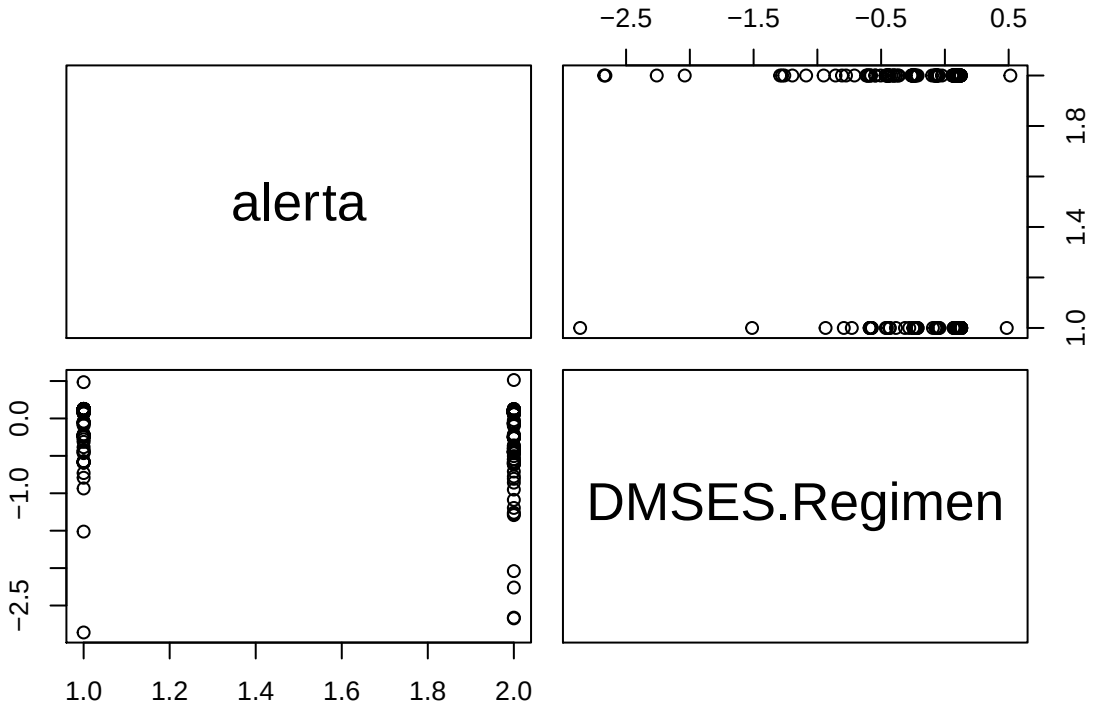

```
pairs(~ alerta + DMSES.Monitor,data=datos)
```



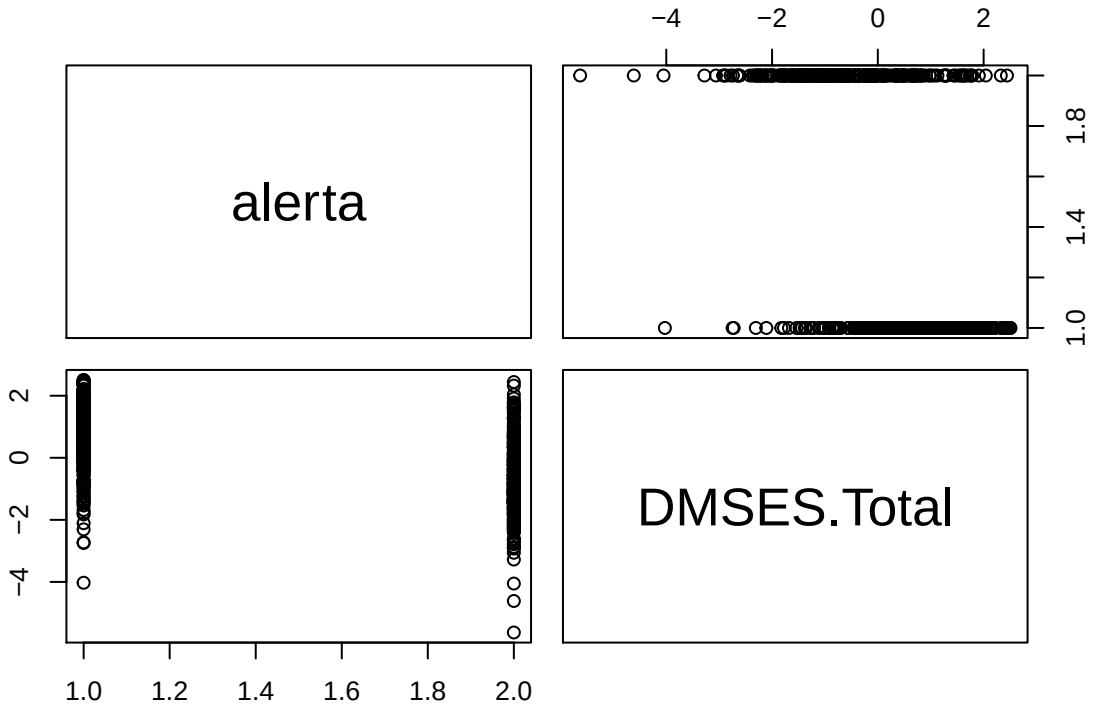

```
pairs(~ alerta + DMSES.Physical,data=datos)
```




```
pairs(~ alerta + DMSES.Regimen,data=datos)
```




```
pairs(~ alerta + DMSES.Total,data=datos)
```



Análisis multivariante de las variables discretas

Las variables discretas con poca variabilidad no se toman en cuenta (cva, cereinfrac, ishemic, stroke, cerebhem,tia, angia, chf, cororevas, pad, neuropath)

```
chisq.test(datos$alerta,datos$hospid)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$hospid
## X-squared = 21.203, df = 3, p-value = 9.551e-05
```

```
chisq.test(datos$alerta,datos$gender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$gender
## X-squared = 0.32616, df = 1, p-value = 0.5679
```

```
chisq.test(datos$alerta,datos$mstatus)
```

```
## Warning in chisq.test(datos$alerta, datos$mstatus): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$mstatus
## X-squared = 12.419, df = 4, p-value = 0.01449
```

```
chisq.test(datos$alerta,datos$edu)
```

```
## Warning in chisq.test(datos$alerta, datos$edu): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$edu
## X-squared = 8.9276, df = 5, p-value = 0.112
```

```
chisq.test(datos$alerta,datos$religion)
```

```
## Warning in chisq.test(datos$alerta, datos$religion): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$religion
## X-squared = 2.6512, df = 2, p-value = 0.2656
```

```
chisq.test(datos$alerta,datos$income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$income
## X-squared = 11.121, df = 5, p-value = 0.04904
```

```
chisq.test(datos$alerta,datos$famhx)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$famhx
## X-squared = 0.9669, df = 1, p-value = 0.3255
```

```
chisq.test(datos$alerta,datos$comob)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$comob
## X-squared = 1.2299, df = 1, p-value = 0.2674
```

```r
chisq.test(datos$alerta,datos$comlip)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$comlip
## X-squared = 1.0754, df = 1, p-value = 0.2997
```

```r
chisq.test(datos$alerta,datos$comht)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$comht
## X-squared = 0.63726, df = 1, p-value = 0.4247
```

```r
chisq.test(datos$alerta,datos$comchd)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$comchd
## X-squared = 4.8263, df = 1, p-value = 0.02803
```

```r
chisq.test(datos$alerta,datos$comkid)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$comkid
## X-squared = 39.775, df = 1, p-value = 2.849e-10
```

```r
chisq.test(datos$alerta,datos$comoth)
```

```
## Warning in chisq.test(datos$alerta, datos$comoth): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$comoth
## X-squared = 4.3425e-28, df = 1, p-value = 1
```

```r
chisq.test(datos$alerta,datos$dmrx)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$dmrx
## X-squared = 75.467, df = 3, p-value = 2.878e-16
```

```r
chisq.test(datos$alerta,datos$smk)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$smk
## X-squared = 0.77241, df = 2, p-value = 0.6796
```

```r
chisq.test(datos$alerta,datos$alcohol)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$alerta and datos$alcohol
## X-squared = 0.75205, df = 2, p-value = 0.6866
```

```r
chisq.test(datos$alerta,datos$compli)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$compli
## X-squared = 97.093, df = 1, p-value < 2.2e-16
```

```r
chisq.test(datos$alerta,datos$mi)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$mi
## X-squared = 6.7972, df = 1, p-value = 0.00913
```

```r
chisq.test(datos$alerta,datos$renal)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$renal
## X-squared = 22.828, df = 1, p-value = 1.772e-06
```

```r
chisq.test(datos$alerta,datos$dn)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$dn
## X-squared = 14.389, df = 1, p-value = 0.0001487
```

```r
chisq.test(datos$alerta,datos$dr)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$dr
## X-squared = 65.391, df = 1, p-value = 6.141e-16
```

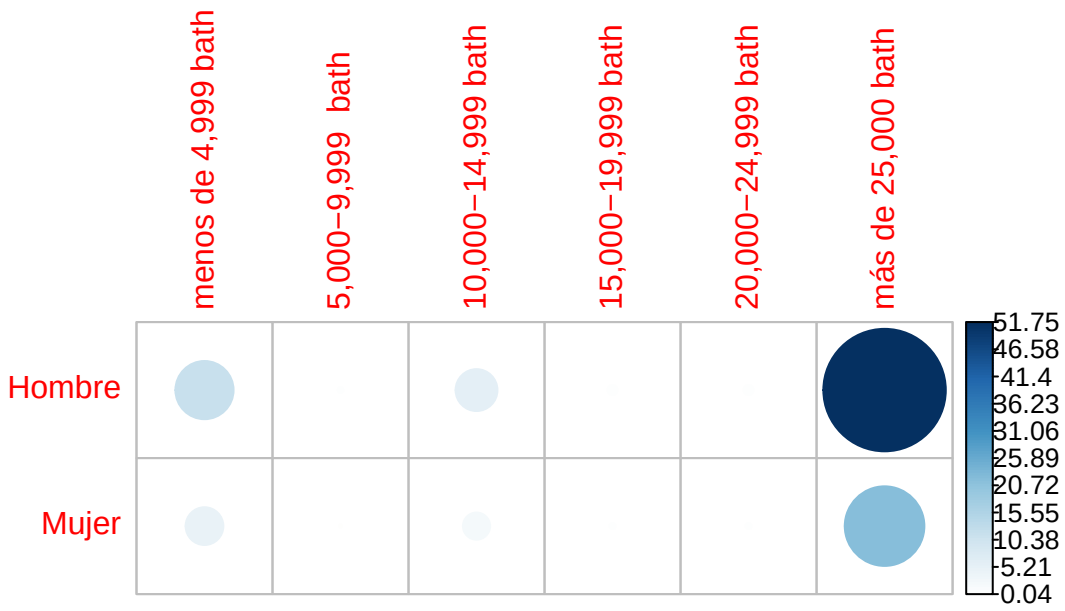
```r
chisq.test(datos$alerta,datos$sobrepeso)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$sobrepeso
## X-squared = 2.4841, df = 1, p-value = 0.115
```

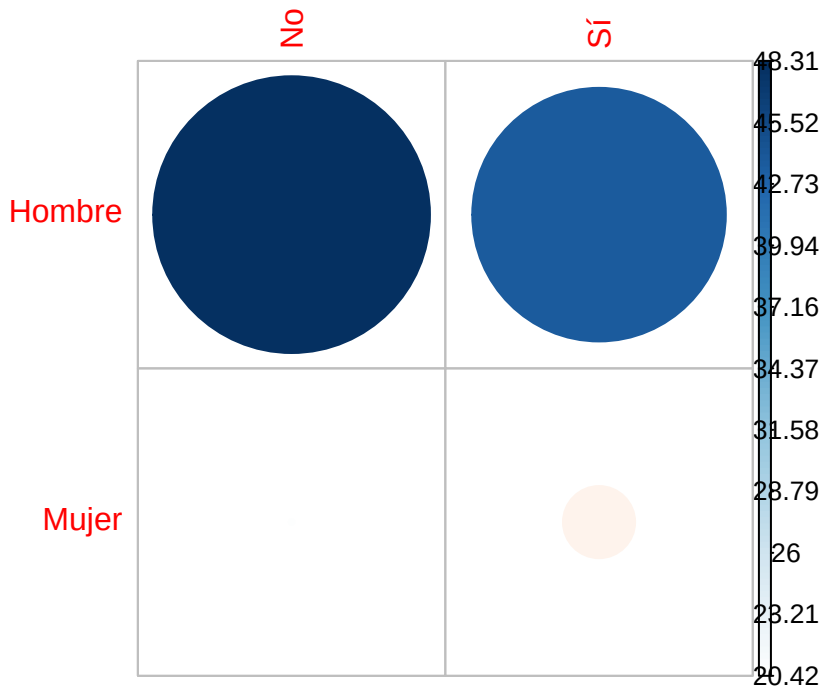
```r
chisq.test(datos$alerta,datos$obesidad)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$alerta and datos$obesidad
## X-squared = 0.1288, df = 1, p-value = 0.7197
```

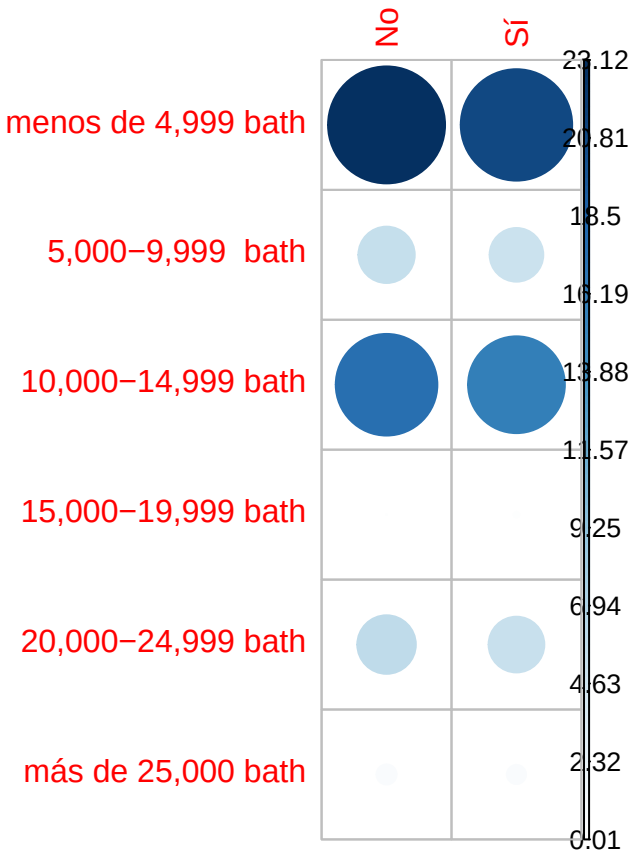
```r
# caso gráfico

chisq <- chisq.test(datos$gender,datos$income)
# Contibution in percentage (%)
contrib <- 100*chisq$residuals^2/chisq$statistic
# Visualize the contribution
corrplot(contrib, is.cor = FALSE)
```




```
chisq <- chisq.test(datos$gender,datos$alerta)
# Contibution in percentage (%)
contrib <- 100*chisq$residuals^2/chisq$statistic
# Visualize the contribution
corrplot(contrib, is.cor = FALSE)
```




```r
chisq <- chisq.test(datos$income,datos$alerta)
# Contibution in percentage (%)
contrib <- 100*chisq$residuals^2/chisq$statistic
# Visualize the contribution
corrplot(contrib, is.cor = FALSE)
```



Tenemos un dependencia entre las variables alerta y hospital, estado civil, ingresos, enfermedad coronaria, enfermedad renal, tipo de tratamiento médico d ela diabetes, complicación de la diabetes, problemas miocardio, insuficiancia renal, nefropatía y retinopatía.

2 Regresión lineal

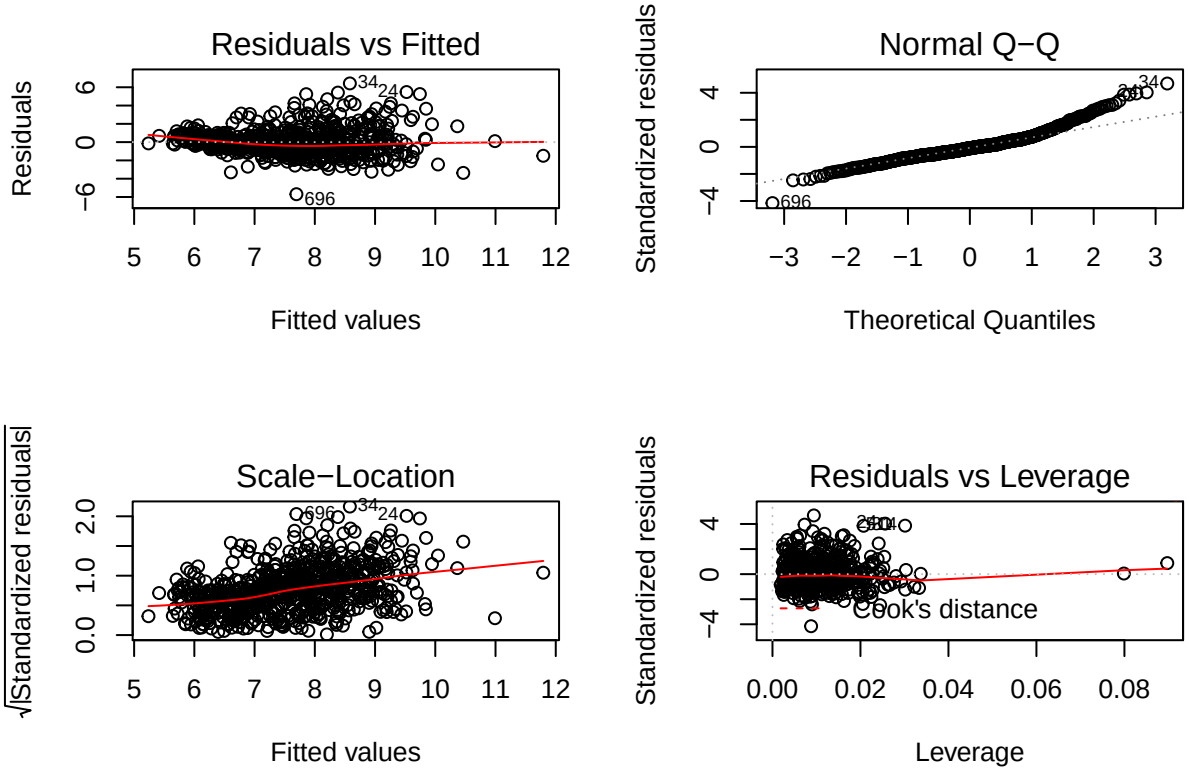
El objetivo de este apartado será comprobar si la puntuación del test DMSES (total o de alguno de sus apartados) puede predecir el valor de la variable 'hba1c'.

```
par(mfrow=c(2,2))
mod <- lm(hba1c ~ DMSES.Total + DK.10 + age + renal + ldl + bmi,data=datos)
(smod <- summary(mod))
```

```
##
## Call:
## lm(formula = hba1c ~ DMSES.Total + DK.10 + age + renal + ldl +
##     bmi, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6945 -0.8032 -0.1390  0.6161  6.4182
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.902466   0.543628  16.376  < 2e-16 ***
```

```
## DMSES.Total -0.631903   0.041203 -15.336  < 2e-16 ***
## DK.10        0.019167   0.026675   0.719   0.4727
## age         -0.028289   0.005113  -5.533 4.46e-08 ***
## renalSí      0.629150   0.143612   4.381 1.36e-05 ***
## ldl          0.007478   0.001714   4.363 1.48e-05 ***
## bmi         -0.016461   0.008569  -1.921   0.0551 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.377 on 693 degrees of freedom
## Multiple R-squared:  0.3538, Adjusted R-squared:  0.3482
## F-statistic: 63.24 on 6 and 693 DF,  p-value: < 2.2e-16
```

```
plot(mod)
```




```
vif(mod)
```

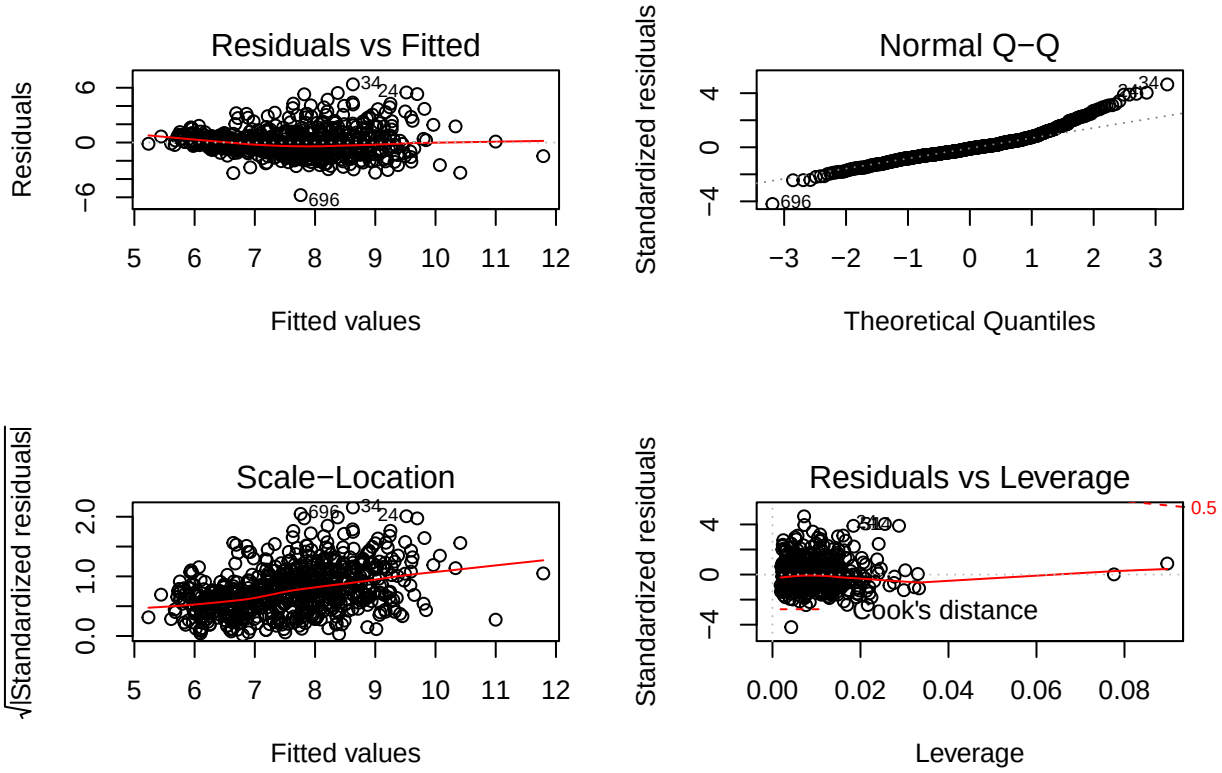
```
## DMSES.Total       DK.10         age     renalSí         ldl         bmi
##    1.076190    1.051559    1.153334    1.067002    1.021586    1.083572
```

```
# eliminamos DK.10
mod <- lm(hba1c ~ DMSES.Total +         age + renal + ldl + bmi,data=datos)
(smod <- summary(mod))
```

```
##
```

```
## Call:
## lm(formula = hba1c ~ DMSES.Total + age + renal + ldl + bmi, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7610 -0.7966 -0.1314  0.5947  6.3716
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.064169   0.494685  18.323  < 2e-16 ***
## DMSES.Total -0.633259   0.041145 -15.391  < 2e-16 ***
## age         -0.028930   0.005032  -5.749 1.35e-08 ***
## renalSí      0.618438   0.142786   4.331 1.70e-05 ***
## ldl          0.007401   0.001710   4.328 1.72e-05 ***
## bmi         -0.016621   0.008564  -1.941   0.0527 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.377 on 694 degrees of freedom
## Multiple R-squared:  0.3533, Adjusted R-squared:  0.3487
## F-statistic: 75.84 on 5 and 694 DF,  p-value: < 2.2e-16
```

```
plot(mod)
```



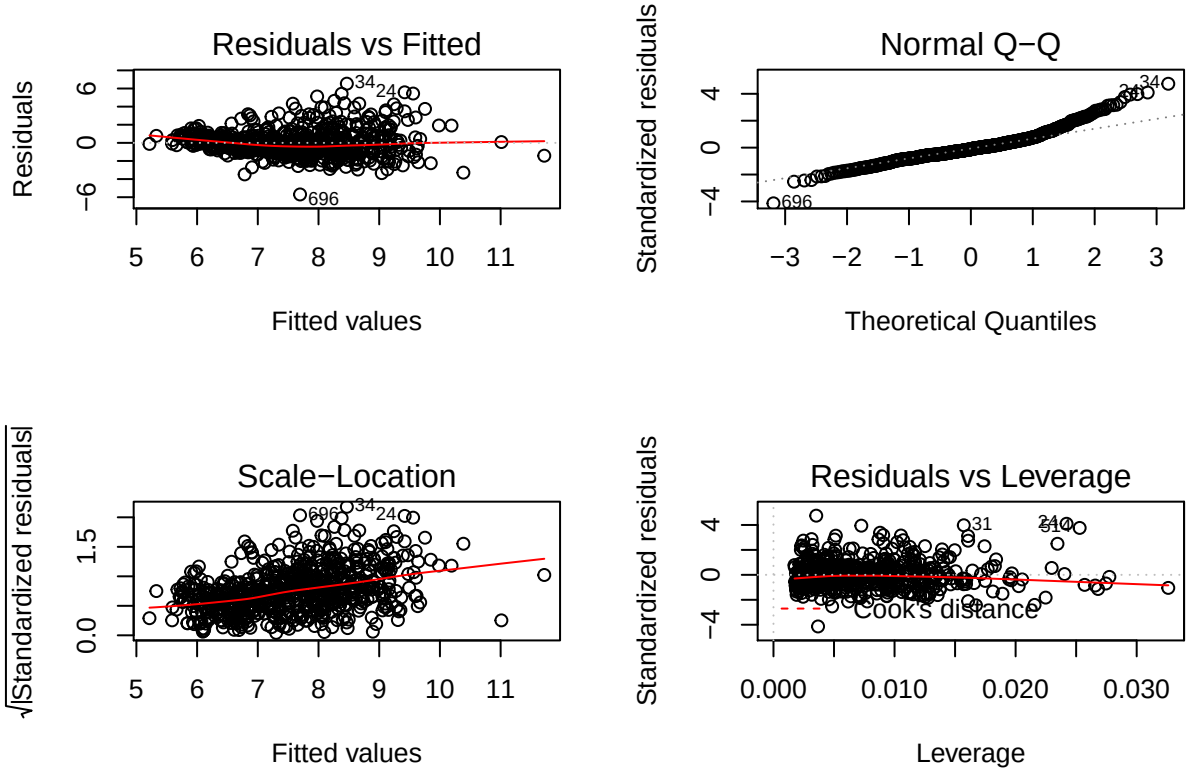
```r
vif(mod)
```

```
## DMSES.Total          age     renalSí          ldl          bmi
##    1.073931    1.118247    1.055502    1.017605    1.082847
```

```r
# eliminamos BMI
mod <- lm(hba1c ~ DMSES.Total +         age + renal + ldl      ,data=datos)
(smod <- summary(mod))
```

```
##
## Call:
## lm(formula = hba1c ~ DMSES.Total + age + renal + ldl, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6961 -0.7706 -0.1625  0.5901  6.5332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.454447   0.382888  22.081  < 2e-16 ***
## DMSES.Total -0.626578   0.041082 -15.252  < 2e-16 ***
## age         -0.026500   0.004884  -5.426 7.97e-08 ***
## renalSí      0.611629   0.143027   4.276 2.17e-05 ***
## ldl          0.007424   0.001713   4.334 1.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.379 on 695 degrees of freedom
## Multiple R-squared:  0.3498, Adjusted R-squared:  0.3461
## F-statistic: 93.48 on 4 and 695 DF,  p-value: < 2.2e-16
```

```r
plot(mod)
```




```
vif(mod)
```

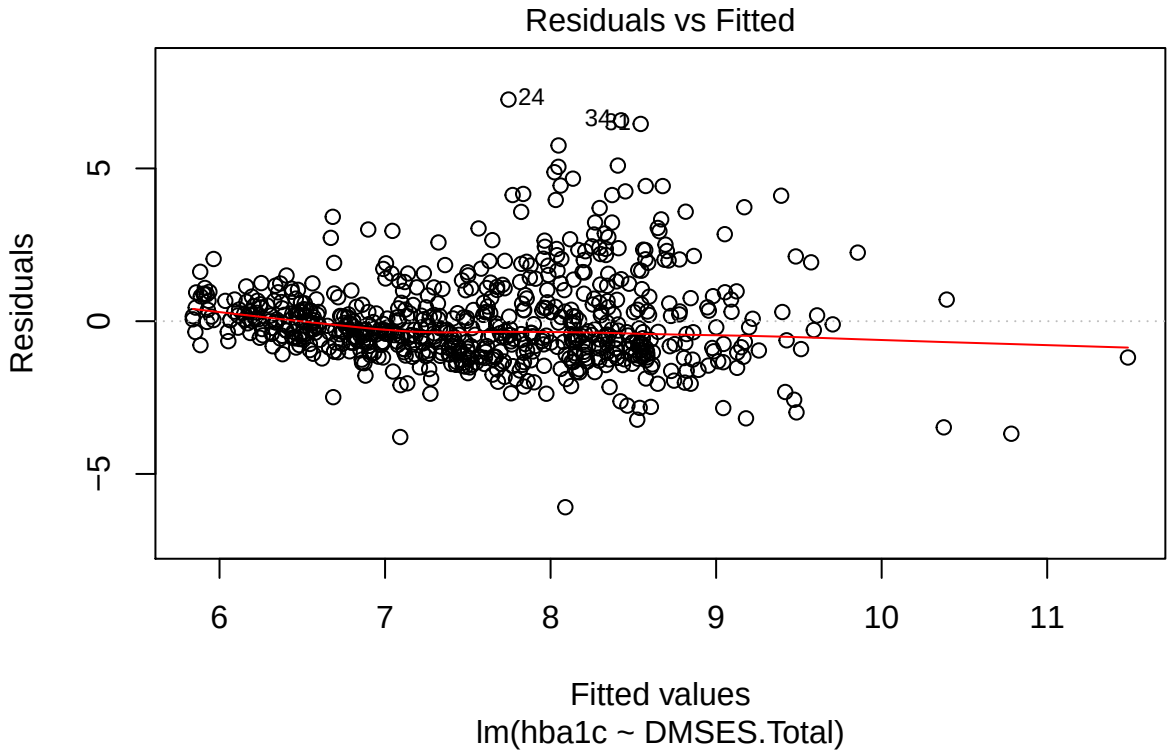
```
## DMSES.Total         age     renalSí         ldl
##    1.066413    1.049026    1.054865    1.017554
```

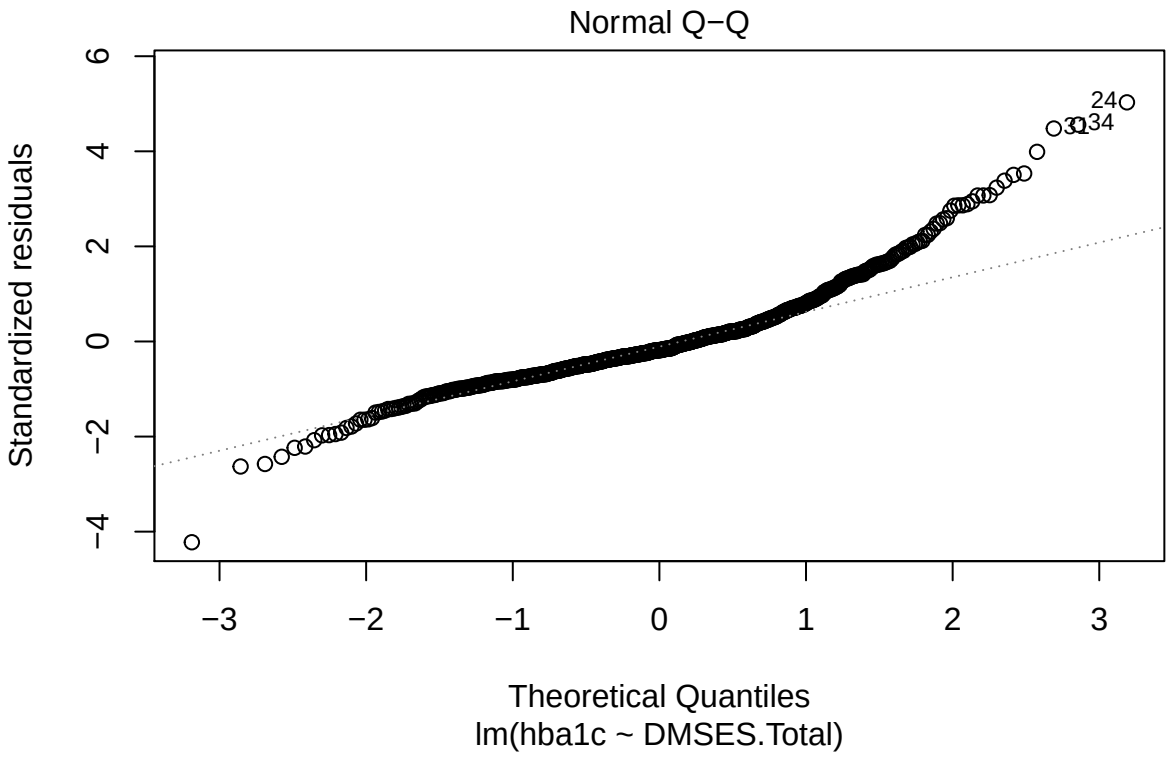
```
par(mfrow=c(1,1))


# sólo  DMSES.Total
mod <- lm(hba1c ~ DMSES.Total,data=datos)
(smod <- summary(mod))
```

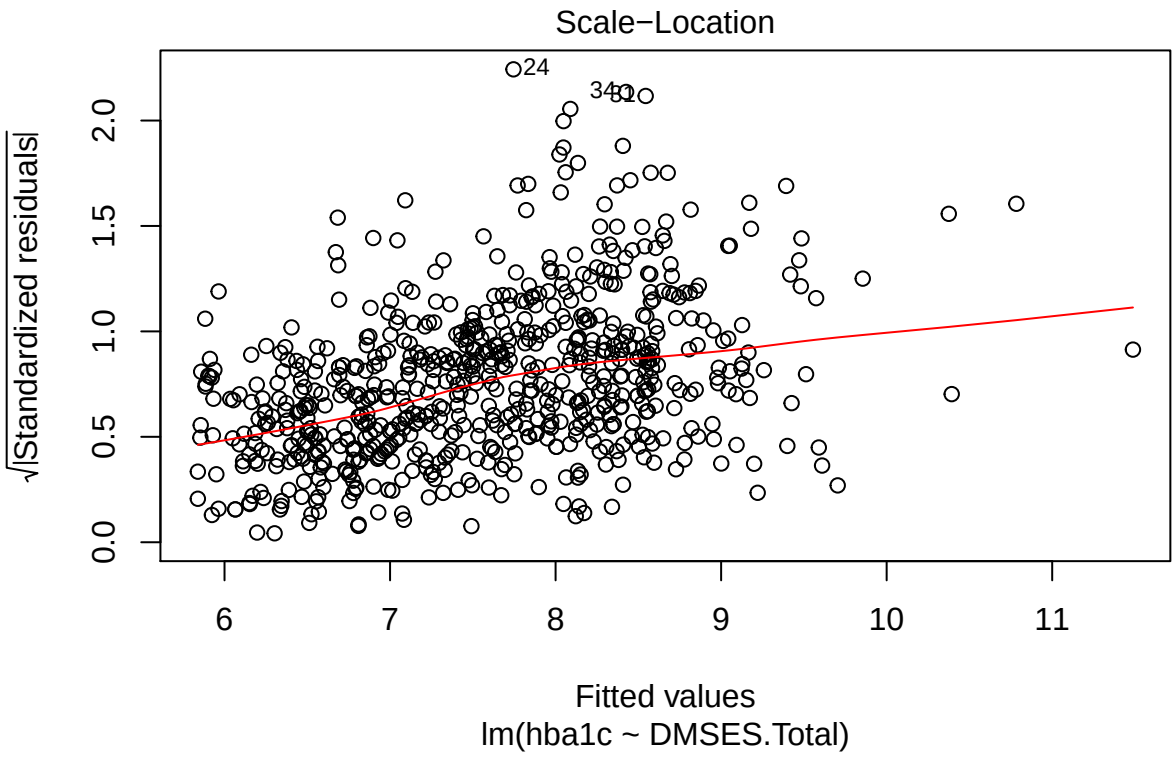
```
##
## Call:
## lm(formula = hba1c ~ DMSES.Total, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0890 -0.8658 -0.2634  0.5526  7.2544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.57900    0.05455  138.93   <2e-16 ***
## DMSES.Total -0.69443    0.04163  -16.68   <2e-16 ***
## ---
```

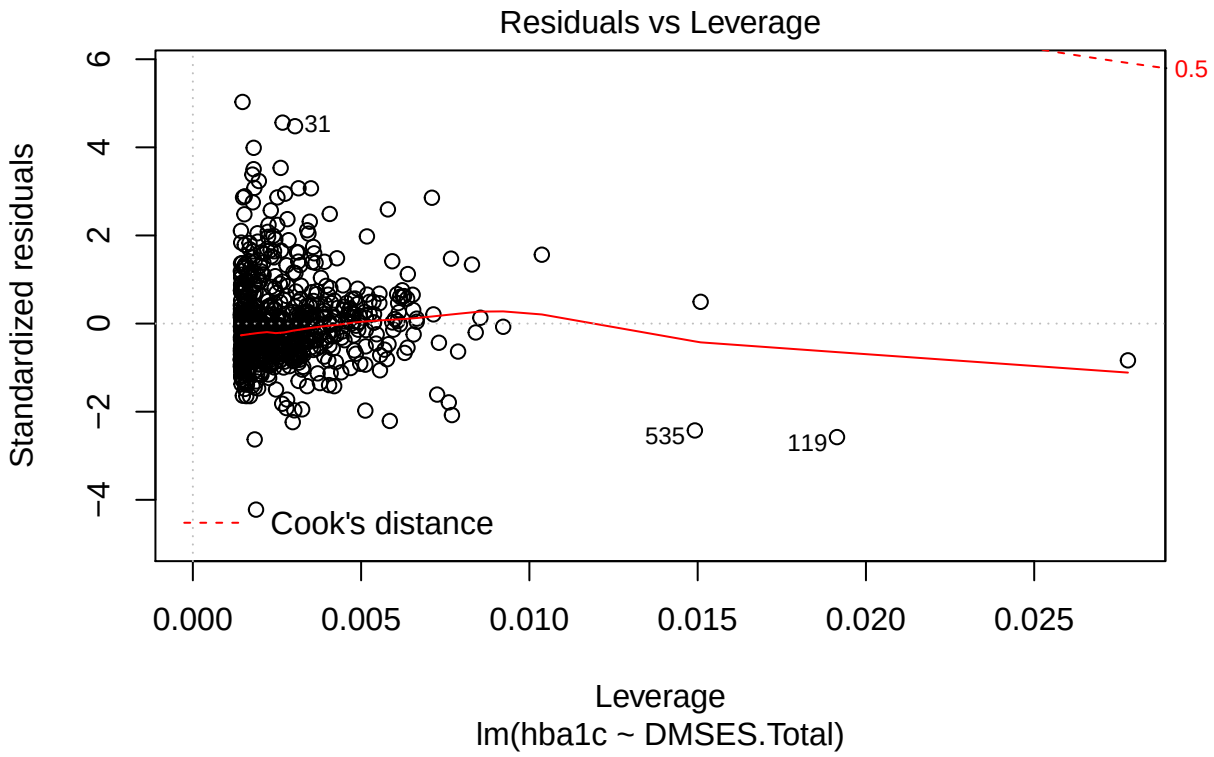
```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.443 on 698 degrees of freedom
## Multiple R-squared:  0.285,  Adjusted R-squared:  0.284
## F-statistic: 278.3 on 1 and 698 DF,  p-value: < 2.2e-16
```

```
plot(mod)
```




Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
24
34
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(hba1c ~ DMSES.Total)


Scale−Location
24
342
82
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
6
7
8
9
10
11
Fitted values
lm(hba1c ~ DMSES.Total)




```
vif(mod)
```

```
## DMSES.Total
##           1
```

```
par(mfrow=c(1,1))
```

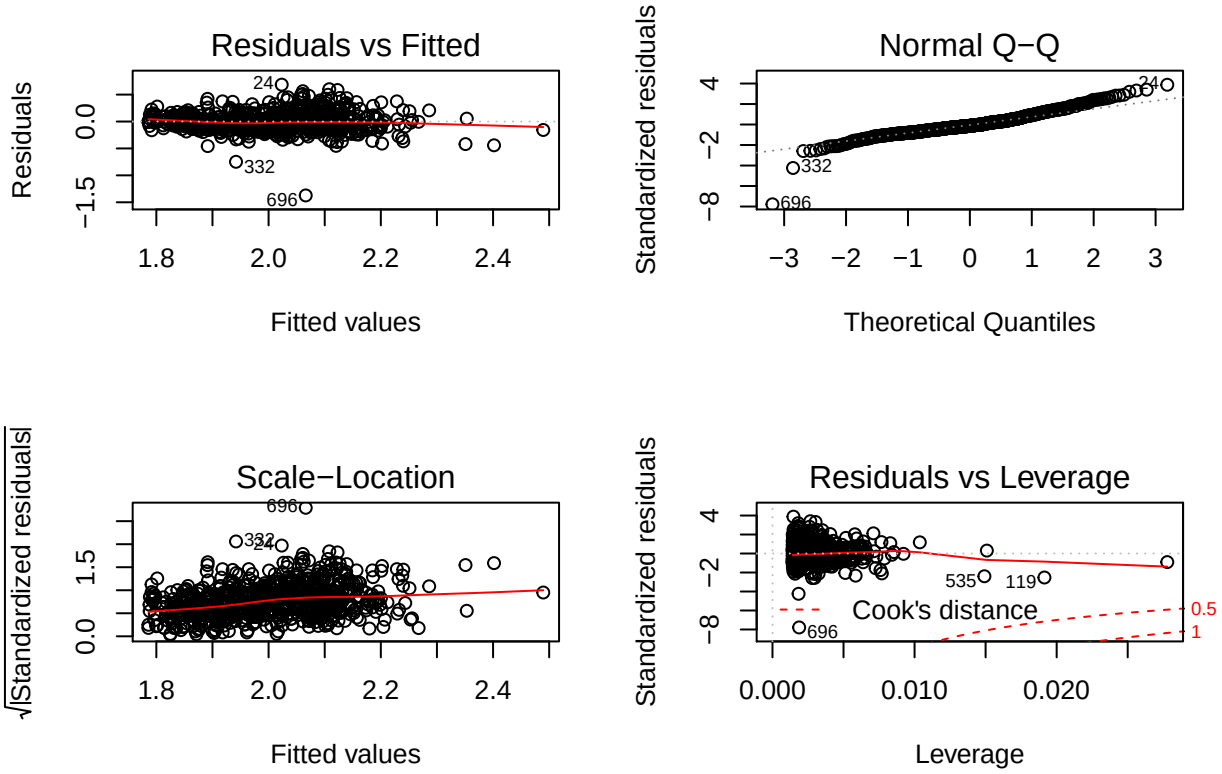
No se consigue tener un R2 elevado ni corregir la no normalidad de los residuos. Aplicamos logaritmos a 'hba1c' (variable estrictamente positiva) y se mejora un poco la situación en cuanto a normalidad pero el R2 sigue sin subir. Búsqueda de outliers pendiente.

```
# modelo solo con puntuación total
par(mfrow=c(2,2))
mod <- lm(log(hba1c) ~ DMSES.Total ,data=datos)
(smod <- summary(mod))
```

```
##
## Call:
## lm(formula = log(hba1c) ~ DMSES.Total, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37304 -0.10107 -0.02185  0.08619  0.68464
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.002665   0.006673  300.12   <2e-16 ***
## DMSES.Total -0.086487   0.005092  -16.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1765 on 698 degrees of freedom
## Multiple R-squared:  0.2924, Adjusted R-squared:  0.2914
## F-statistic: 288.5 on 1 and 698 DF,  p-value: < 2.2e-16
```

```
plot(mod)
```




```
vif(mod)
```

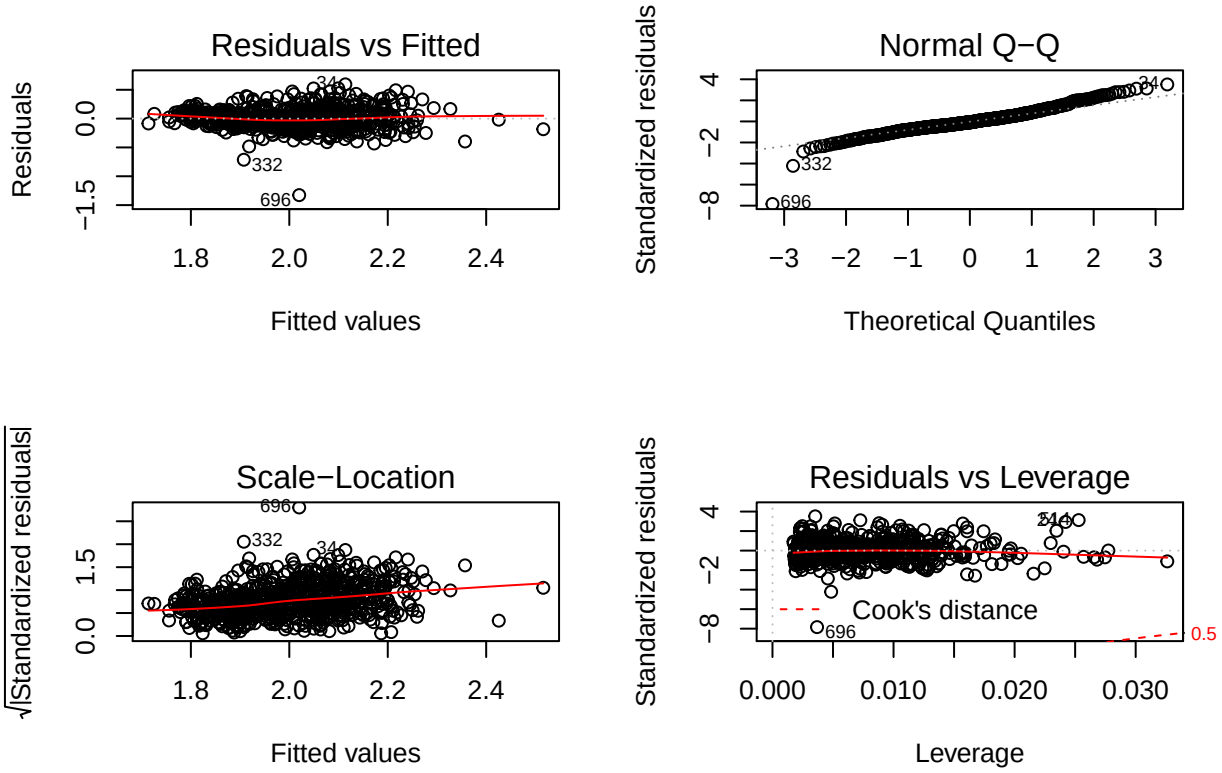
```
## DMSES.Total
##           1
```

```
par(mfrow=c(1,1))
```

```
par(mfrow=c(2,2))
mod <- lm(log(hba1c) ~ DMSES.Total +          age + renal + ldl       ,data=datos)
(smod <- summary(mod))
```

```
##
## Call:
## lm(formula = log(hba1c) ~ DMSES.Total + age + renal + ldl, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32692 -0.09223 -0.01355  0.08516  0.59417
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1135308  0.0470599  44.911  < 2e-16 ***
## DMSES.Total -0.0786385  0.0050494 -15.574  < 2e-16 ***
## age         -0.0031851  0.0006003  -5.306 1.51e-07 ***
## renalSí      0.0684713  0.0175791   3.895 0.000108 ***
## ldl          0.0008448  0.0002106   4.012 6.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1695 on 695 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3466
## F-statistic:  93.7 on 4 and 695 DF,  p-value: < 2.2e-16
```

```
plot(mod)
```



```r
vif(mod)
```

```
## DMSES.Total         age     renalSí         ldl
##    1.066413    1.049026    1.054865    1.017554
```

```r
par(mfrow=c(1,1))
```

```r
datos$dmrx1 <- as.factor(ifelse(datos$dmrx=="Sin medicación",1,0))
datos$dmrx2 <- as.factor(ifelse(datos$dmrx=="Hipoglicémico oral",1,0))
datos$dmrx3 <- as.factor(ifelse(datos$dmrx=="Insulina",1,0))

table(datos$obesidad)
```

```
##
##  No  Sí
## 518 182
```

```r
datos[1:10,c("dmrx","dmrx1","dmrx2","dmrx3")]
```

```
##                  dmrx dmrx1 dmrx2 dmrx3
## 1  Hipoglicémico oral     0     1     0
## 2  Hipoglicémico oral     0     1     0
## 3  Hipoglicémico oral     0     1     0
## 4               Ambos     0     0     0
## 5           Insulina      0     0     1
## 6  Hipoglicémico oral     0     1     0
## 7  Hipoglicémico oral     0     1     0
## 8  Hipoglicémico oral     0     1     0
## 9  Hipoglicémico oral     0     1     0
## 10              Ambos     0     0     0
```

```r
var.cat.inf   <- c("obesidad","sobrepeso","gender","renal","dmrx1","dmrx2","dmrx3","dn","dr")
var.cont.norm <- c("hba1c","age", "hdl", "sbp","dbp","DMSES.Total","DK.10")
var.interes   <- c(var.cat.inf,var.cont.norm)
var.interes
```

```
##  [1] "obesidad"    "sobrepeso"   "gender"      "renal"       "dmrx1"
##  [6] "dmrx2"       "dmrx3"       "dn"          "dr"          "hba1c"
## [11] "age"         "hdl"         "sbp"         "dbp"         "DMSES.Total"
## [16] "DK.10"
```

```r
datos1 <- datos[,var.interes]
dim(datos1)
```

```
## [1] 700  16
```

```r
# Fit the full model
#full.model <- lm(alerta ~., data = datos1)
#summary(full.model)
# Stepwise regression model
```

```r
#step.model <- stepAIC(full.model, direction = "both", trace = FALSE)
#summary(step.model)

require(leaps)
```

```
## Loading required package: leaps
```

```r
b <- regsubsets(hba1c~.,data=datos1,nvmax = 15)
rs <- summary(b)
rs$which
```

```
##    (Intercept) obesidadSí sobrepesoSí genderMujer renalSí dmrx11 dmrx21 dmrx31
## 1         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE
## 2         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE
## 3         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE
## 4         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE   TRUE  FALSE
## 5         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE   TRUE  FALSE
## 6         TRUE      FALSE        TRUE       FALSE   FALSE  FALSE   TRUE  FALSE
## 7         TRUE      FALSE        TRUE       FALSE   FALSE  FALSE   TRUE  FALSE
## 8         TRUE      FALSE        TRUE       FALSE   FALSE   TRUE   TRUE  FALSE
## 9         TRUE      FALSE        TRUE       FALSE    TRUE   TRUE   TRUE  FALSE
## 10        TRUE      FALSE        TRUE       FALSE    TRUE   TRUE   TRUE  FALSE
## 11        TRUE      FALSE        TRUE       FALSE    TRUE   TRUE   TRUE   TRUE
## 12        TRUE       TRUE        TRUE       FALSE    TRUE   TRUE   TRUE   TRUE
## 13        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE
## 14        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE
## 15        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE
##     dnSí  drSí   age   hdl   sbp   dbp DMSES.Total DK.10
## 1  FALSE FALSE FALSE FALSE FALSE FALSE        TRUE FALSE
## 2  FALSE  TRUE FALSE FALSE FALSE FALSE        TRUE FALSE
## 3  FALSE  TRUE  TRUE FALSE FALSE FALSE        TRUE FALSE
## 4  FALSE  TRUE  TRUE FALSE FALSE FALSE        TRUE FALSE
## 5  FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 6  FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 7   TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 8   TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 9   TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 10  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 11  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 12  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 13  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 14  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE  TRUE
## 15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE        TRUE  TRUE
```

```r
length(rs$rss)
```

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

```r
length(2:16)
```

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

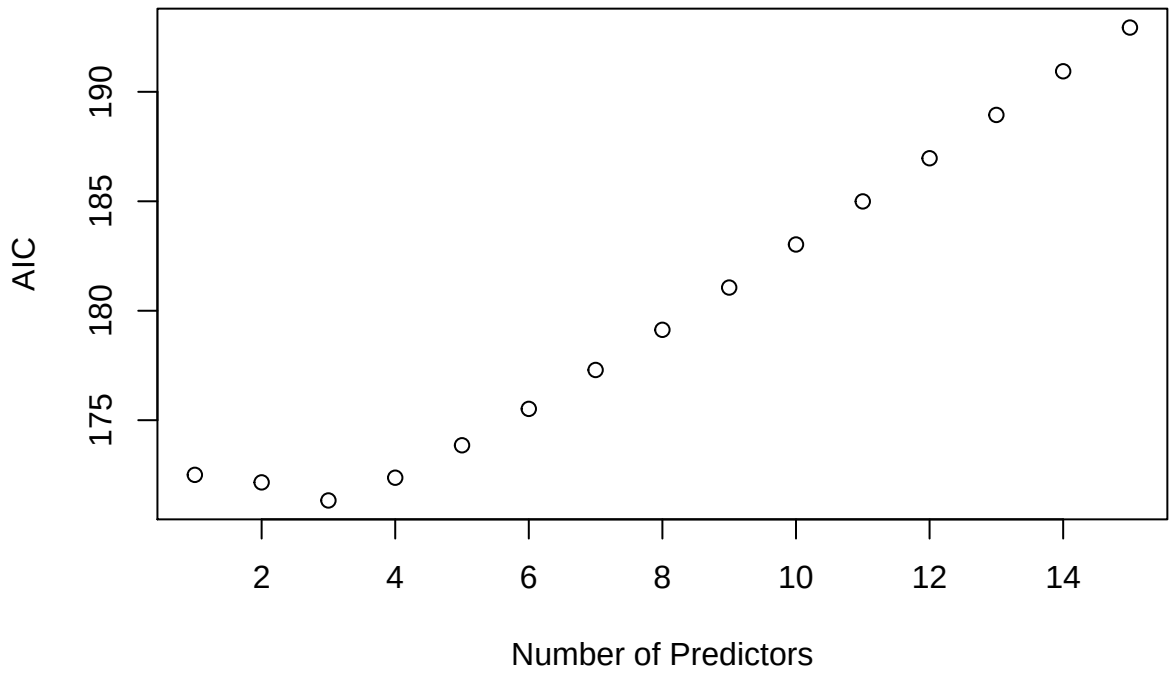
```
rs$rss
```

```
##  [1] 1454.025 1387.418 1311.201 1286.231 1272.945 1264.415 1258.775 1254.618
##  [9] 1252.876 1252.056 1251.262 1250.551 1250.030 1249.867 1249.804
```

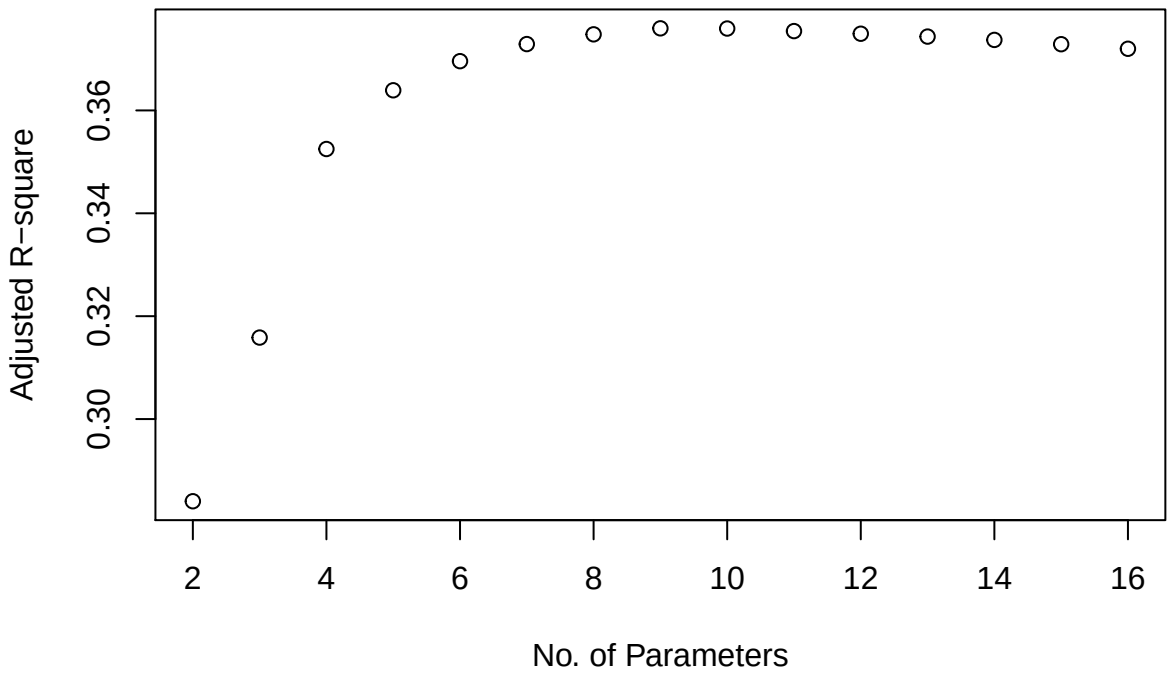
```
AIC <- 50*log(rs$rss/50) + (2:16)*2
dim(AIC)
```

```
## NULL
```

```
plot(AIC ~ I(1:15), ylab="AIC", xlab="Number of Predictors")
```




```
plot(2:16,rs$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")
```




```
which.max(rs$adjr2)
```

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

```
# datos recientes (suponiendo fecha frecuencia cardiaca es
# fecha de realizacion del cuestionario)
datos.recientes <- datos %>% filter(fecha.hba1c==fecha.bp)
datos.recientes <- datos.recientes[,var.interes]
dim(datos.recientes)
```

```
## [1] 309  16
```

```
b <- regsubsets(hba1c~.,data=datos.recientes,nvmax = 15)
rs <- summary(b)
rs$which
```

```
##    (Intercept) obesidadSí sobrepesoSí genderMujer renalSí dmrx11 dmrx21 dmrx31
## 1         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE
## 2         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE
## 3         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE
## 4         TRUE      FALSE       FALSE       FALSE    TRUE  FALSE  FALSE  FALSE
## 5         TRUE      FALSE       FALSE       FALSE    TRUE  FALSE  FALSE  FALSE
## 6         TRUE       TRUE       FALSE       FALSE    TRUE  FALSE  FALSE  FALSE
## 7         TRUE       TRUE       FALSE       FALSE    TRUE  FALSE  FALSE  FALSE
```

```
## 8         TRUE       TRUE       FALSE       FALSE    TRUE  FALSE   TRUE  FALSE
## 9         TRUE       TRUE        TRUE       FALSE    TRUE  FALSE   TRUE  FALSE
## 10        TRUE       TRUE        TRUE       FALSE    TRUE  FALSE   TRUE  FALSE
## 11        TRUE       TRUE        TRUE       FALSE    TRUE   TRUE   TRUE  FALSE
## 12        TRUE       TRUE        TRUE       FALSE    TRUE   TRUE   TRUE   TRUE
## 13        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE
## 14        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE
## 15        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE
##     dnSí  drSí   age   hdl   sbp   dbp DMSES.Total DK.10
## 1  FALSE FALSE FALSE FALSE FALSE FALSE        TRUE FALSE
## 2  FALSE  TRUE FALSE FALSE FALSE FALSE        TRUE FALSE
## 3  FALSE  TRUE  TRUE FALSE FALSE FALSE        TRUE FALSE
## 4  FALSE  TRUE  TRUE FALSE FALSE FALSE        TRUE FALSE
## 5  FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 6  FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 7   TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 8   TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 9   TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 10  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 11  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 12  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 13  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE        TRUE FALSE
## 15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE        TRUE  TRUE
```

```
length(rs$rss)
```

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

```
length(2:16)
```

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

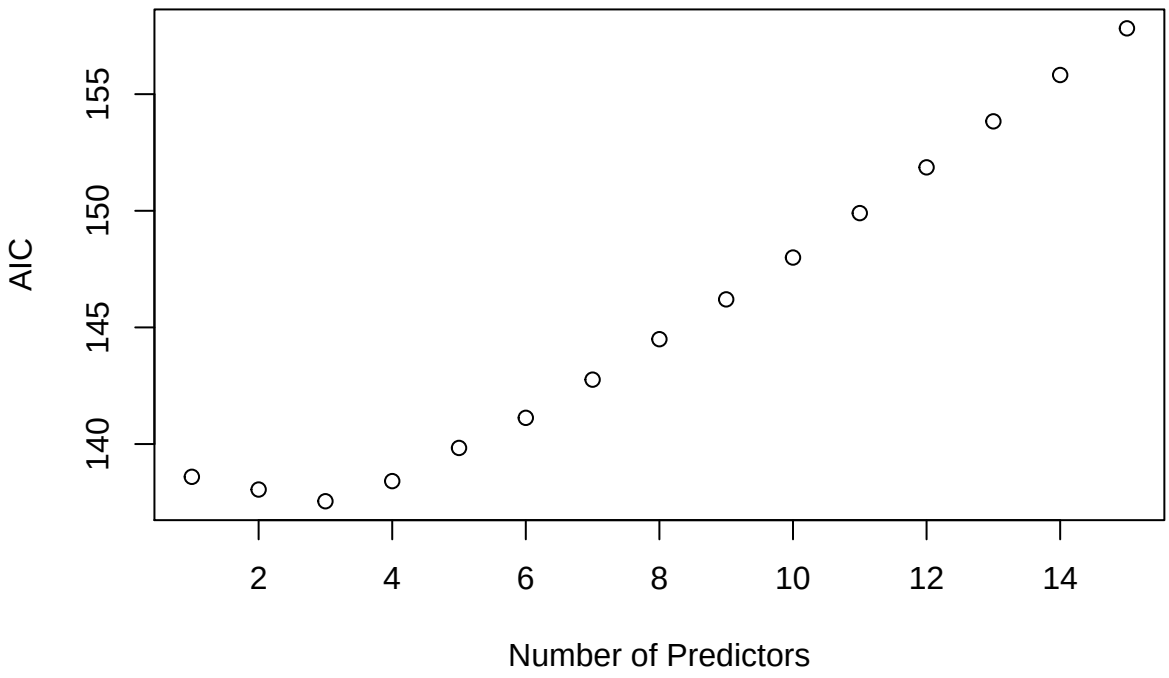
```
rs$rss
```

```
##  [1] 737.9831 701.3478 667.1215 652.1189 644.6347 635.5443 630.9557 627.6099
##  [9] 623.9445 621.3740 620.1989 619.7134 619.3812 619.2185 619.2185
```

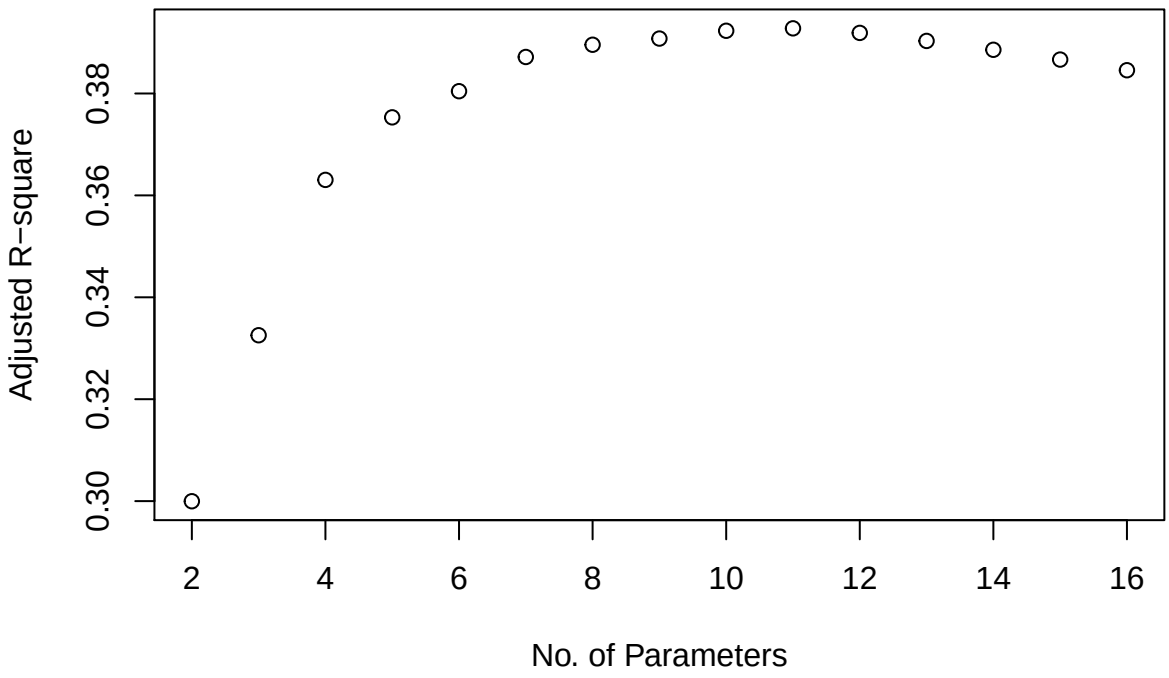
```
AIC <- 50*log(rs$rss/50) + (2:16)*2
dim(AIC)
```

```
## NULL
```

```
plot(AIC ~ I(1:15), ylab="AIC", xlab="Number of Predictors")
```




```
plot(2:16,rs$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")
```




```
which.max(rs$adjr2)
```

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

Utilizando datos recientes de la hemoglobina glicosilada no se mejora el R2 ajustado.

En resumen, no se ha encontrado un modelo lineal, o transformable en lineal con un R2 elevado (>70%). Utilizando únicamente la puntuación total del cuestionario DMSES tenemos un R2 del 28%. Añadiendo las variables edad, renal y LDL se alcanza el 35%.

Explorando los modelos con las 15 variables elegidas después de una criba de normalidad, para variables continuas, y de cierta variabilidad en las categóricas, tampoco se consigue un R2 elevado, por lo que por simplicidad los modelos con menos variables tendrían preferencia.

Los factores de inflación de la varianza se mantienten bajos en los modelos analizados.

3 Regresión logística

El objetivo de este apartado será comprobar si la puntuación del test DMSES (total o de alguno de sus apartados) puede ayudar a discriminar entre pacientes que controlan la diabetes y los que no.

3.1 Preparación de los ficheros para 'train' y 'test'

```r
# cálculo del numero de registros para el training
# según la proporción de registros dedicado al training
pct.train <- 0.80

n <- nrow(datos)
n.train <- floor(n*pct.train)

set.seed(123)
ind.training <- sample(1:n,n.train)
length(ind.training)
```

```
## [1] 560
```

```r
datos.train <- datos[ind.training,]
datos.test <- datos[-ind.training,]

dim(datos.train)
```

```
## [1] 560  76
```

```r
dim(datos.test)
```

```
## [1] 140  76
```

3.2 Modelo DMSES subscore Diet

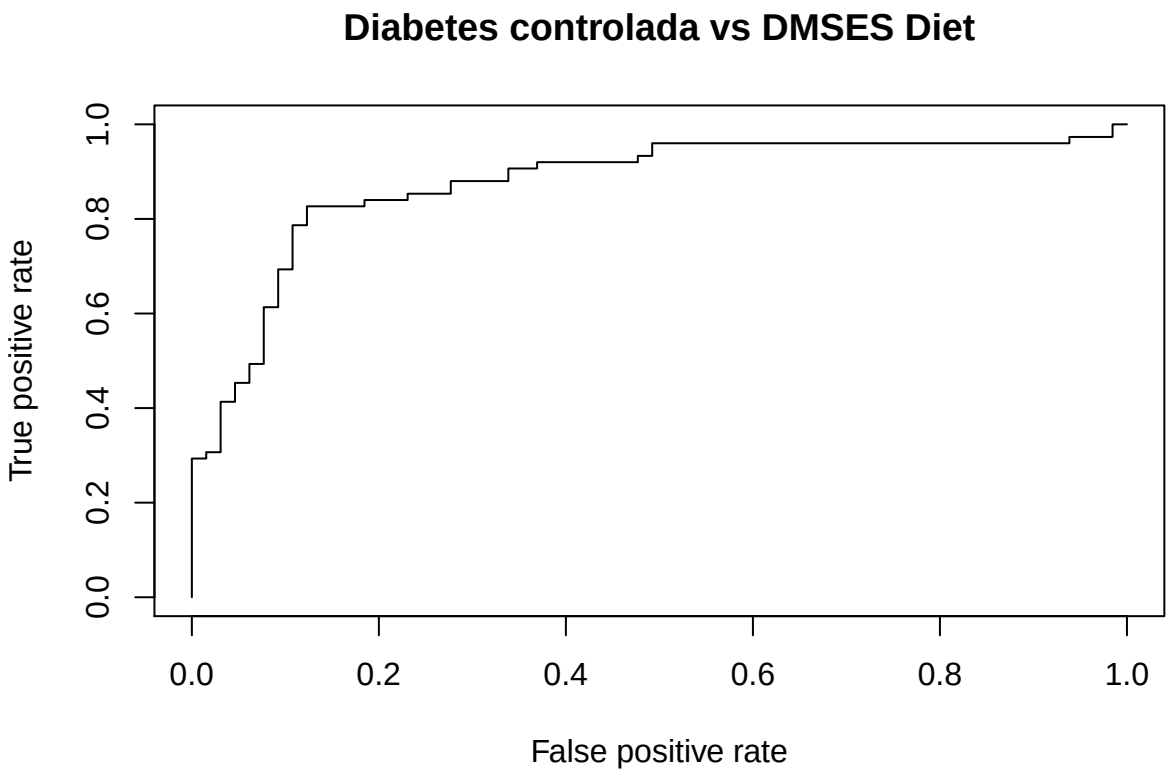
```r
logdiet <- glm(alerta ~ DMSES.Diet, data = datos.train, family = "binomial")
summary(logdiet)
```

```
##
## Call:
## glm(formula = alerta ~ DMSES.Diet, family = "binomial", data = datos.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4735  -0.7003   0.2690   0.6506   2.3641
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1358     0.1119   1.214    0.225
## DMSES.Diet   -2.9163     0.2274 -12.824   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 775.30  on 559  degrees of freedom
## Residual deviance: 501.94  on 558  degrees of freedom
## AIC: 505.94
##
## Number of Fisher Scoring iterations: 4
```

```r
predicciones <- predict(logdiet,newdata=datos.test[,-c(1)],type='response')
predicciones.categ <- ifelse(predicciones > 0.5,1,0)
predicciones.categ <- factor(predicciones.categ,levels=c(0,1),labels=c("No","Sí"))

# Confusion matrix
confusionMatrix(data=predicciones.categ, reference=datos.test$alerta)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Sí
##         No 50 12
##         Sí 15 63
##
##                Accuracy : 0.8071
##                  95% CI : (0.7319, 0.8689)
##     No Information Rate : 0.5357
##     P-Value [Acc > NIR] : 1.787e-11
##
##                   Kappa : 0.6111
##
##  Mcnemar's Test P-Value : 0.7003
##
##             Sensitivity : 0.7692
##             Specificity : 0.8400
##          Pos Pred Value : 0.8065
##          Neg Pred Value : 0.8077
##              Prevalence : 0.4643
##          Detection Rate : 0.3571
##    Detection Prevalence : 0.4429
##       Balanced Accuracy : 0.8046
##
##        'Positive' Class : No
##
```

```r
# ROC and AUC
pr <- prediction(predicciones, datos.test$alerta)
# TPR = sensitivity, FPR=specificity
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf,main = "Diabetes controlada vs DMSES Diet")
```




```
auc <- performance(pr, measure = "auc")
(auc <- auc@y.values[[1]])
```

```
## [1] 0.8754872
```

```
# significancia del modelo
anova(logdiet, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: alerta
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          559     775.30
## DMSES.Diet  1   273.36       558     501.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#
#
# hay mejora si añadimos algunos predictores?

logdiet <- glm(alerta ~ DMSES.Diet + bmi + dmdura + age, data = datos.train, family = "binomial")
summary(logdiet)
```

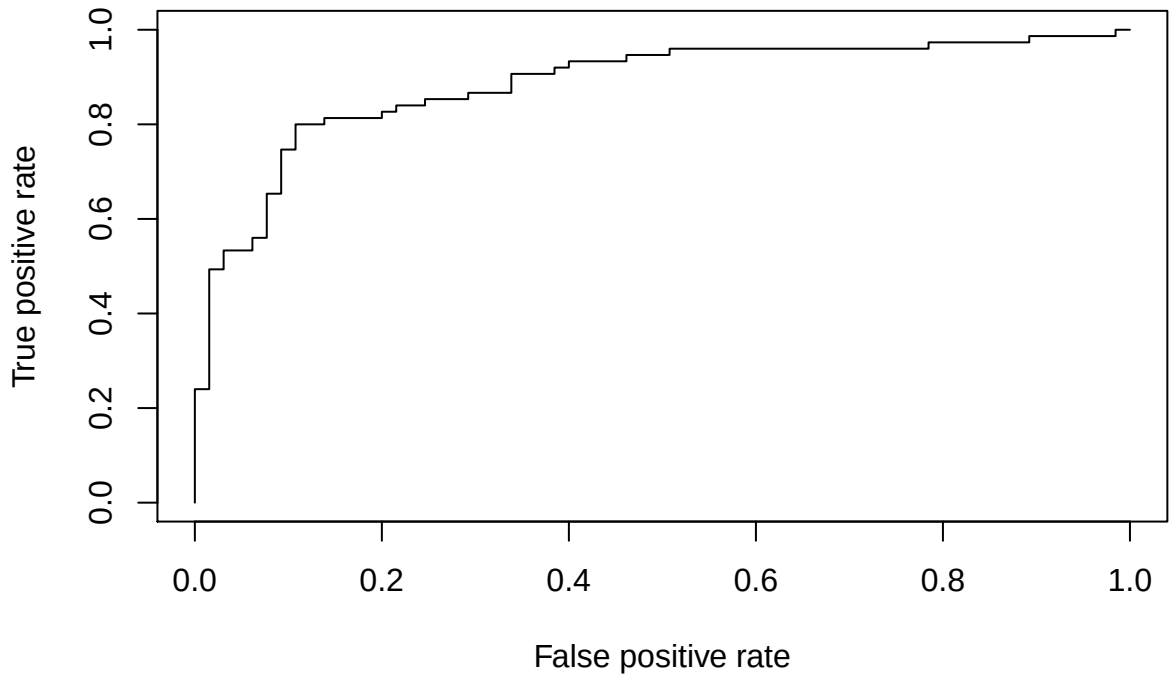
```
##
## Call:
## glm(formula = alerta ~ DMSES.Diet + bmi + dmdura + age, family = "binomial",
##     data = datos.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4451  -0.6475   0.2544   0.6422   2.3971
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.95170    0.97621   3.024   0.0025 **
## DMSES.Diet  -2.93378    0.23468 -12.501   <2e-16 ***
## bmi         -0.04063    0.01888  -2.153   0.0313 *
## dmdura       0.01003    0.01546   0.649   0.5166
## age         -0.02849    0.01212  -2.351   0.0187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 775.30  on 559  degrees of freedom
## Residual deviance: 493.11  on 555  degrees of freedom
## AIC: 503.11
##
## Number of Fisher Scoring iterations: 5
```

```r
predicciones <- predict(logdiet,newdata=datos.test[,-c(1)],type='response')
predicciones.categ <- ifelse(predicciones > 0.5,1,0)
predicciones.categ <- factor(predicciones.categ,levels=c(0,1),labels=c("No","Sí"))

# Confusion matrix
confusionMatrix(data=predicciones.categ, reference=datos.test$alerta)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Sí
##         No 51 13
##         Sí 14 62
##
##                Accuracy : 0.8071
##                  95% CI : (0.7319, 0.8689)
##     No Information Rate : 0.5357
##     P-Value [Acc > NIR] : 1.787e-11
```

```
##
##                  Kappa : 0.6119
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.7846
##             Specificity : 0.8267
##          Pos Pred Value : 0.7969
##          Neg Pred Value : 0.8158
##              Prevalence : 0.4643
##          Detection Rate : 0.3643
##    Detection Prevalence : 0.4571
##       Balanced Accuracy : 0.8056
##
##        'Positive' Class : No
##
```

```
# ROC and AUC
pr <- prediction(predicciones, datos.test$alerta)
# TPR = sensitivity, FPR=specificity
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
auc <- performance(pr, measure = "auc")
(auc <- auc@y.values[[1]])
```

```
## [1] 0.8847179
```

Añadir más variables no mejora sustancialmente el modelo logístico, y sólo con la puntuación DMSES.Diet ya se tenía un buen resultado.

3.3 Modelo DMSES subscore Monitor

```
logmonitor <- glm(alerta ~ DMSES.Monitor, data = datos.train, family = "binomial")
summary(logmonitor)
```

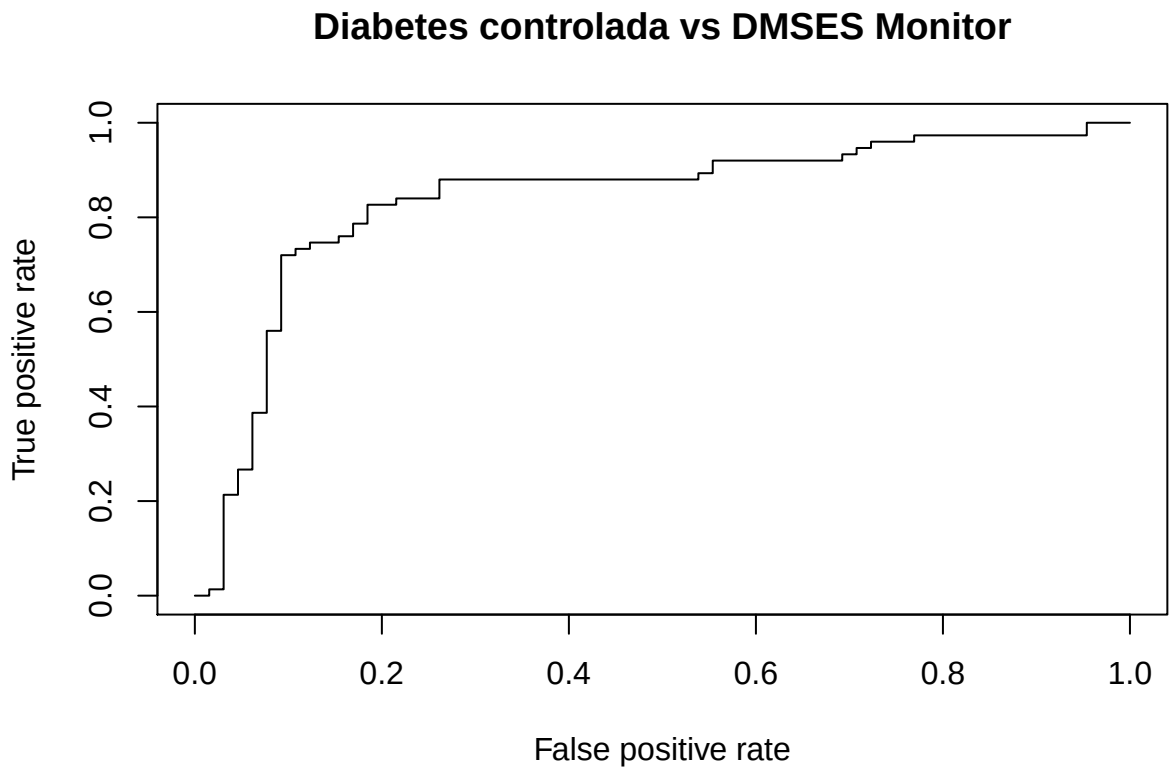
```
##
## Call:
## glm(formula = alerta ~ DMSES.Monitor, family = "binomial", data = datos.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3172  -0.7610   0.2982   0.8035   2.2939
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.09827    0.10398   0.945    0.345
## DMSES.Monitor -5.10044    0.43386 -11.756   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 775.30  on 559  degrees of freedom
## Residual deviance: 562.81  on 558  degrees of freedom
## AIC: 566.81
##
## Number of Fisher Scoring iterations: 4
```

```
predicciones <- predict(logmonitor,newdata=datos.test[,-c(1)],type='response')
predicciones.categ <- ifelse(predicciones > 0.5,1,0)
predicciones.categ <- factor(predicciones.categ,levels=c(0,1),labels=c("No","Sí"))

# Confusion matrix
confusionMatrix(data=predicciones.categ, reference=datos.test$alerta)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Sí
##         No 51 13
##         Sí 14 62
##
##                Accuracy : 0.8071
##                  95% CI : (0.7319, 0.8689)
##     No Information Rate : 0.5357
##     P-Value [Acc > NIR] : 1.787e-11
```

```
##
##                  Kappa : 0.6119
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.7846
##             Specificity : 0.8267
##          Pos Pred Value : 0.7969
##          Neg Pred Value : 0.8158
##              Prevalence : 0.4643
##          Detection Rate : 0.3643
##    Detection Prevalence : 0.4571
##       Balanced Accuracy : 0.8056
##
##        'Positive' Class : No
##
```

```
# ROC and AUC
pr <- prediction(predicciones, datos.test$alerta)
# TPR = sensitivity, FPR=specificity
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf,main = "Diabetes controlada vs DMSES Monitor")
```




```
auc <- performance(pr, measure = "auc")
(auc <- auc@y.values[[1]])
```

```
## [1] 0.8393846
```

```
# significancia del modelo
anova(logmonitor, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: alerta
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            559     775.30
## DMSES.Monitor  1   212.49       558     562.81 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Añadir más variables no mejora sustancialmente el modelo logístico, y sólo con la puntuación DMSES.Monitor ya se tenía un buen resultado.

3.4 Modelo DMSES subscore Regimen

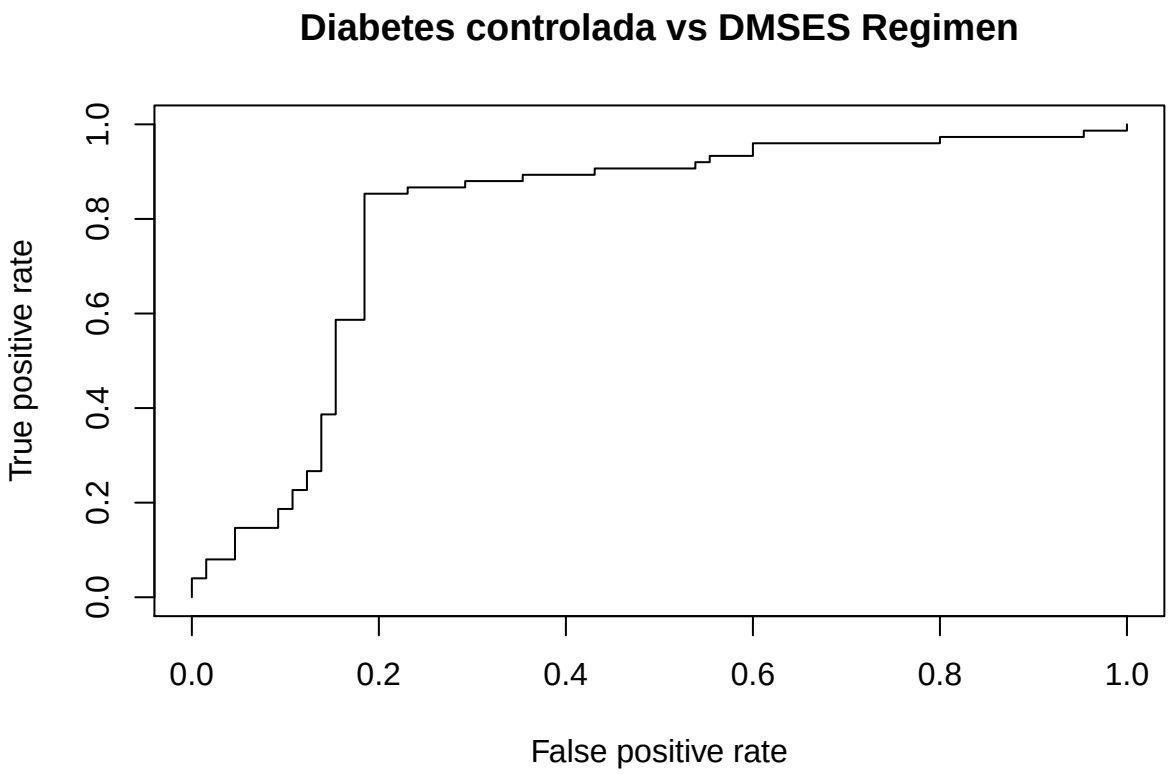
```
logregimen <- glm(alerta ~ DMSES.Regimen, data = datos.train, family = "binomial")
summary(logregimen)
```

```
##
## Call:
## glm(formula = alerta ~ DMSES.Regimen, family = "binomial", data = datos.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3280  -1.1738   0.9224   1.1724   1.1869
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.08850    0.08539   1.036  0.29998
## DMSES.Regimen -0.89237    0.34094  -2.617  0.00886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 775.30  on 559  degrees of freedom
## Residual deviance: 766.53  on 558  degrees of freedom
## AIC: 770.53
##
## Number of Fisher Scoring iterations: 3
```

```r
predicciones <- predict(logregimen,newdata=datos.test[,-c(1)],type='response')
predicciones.categ <- ifelse(predicciones > 0.5,1,0)
predicciones.categ <- factor(predicciones.categ,levels=c(0,1),labels=c("No","Sí"))

# Confusion matrix
confusionMatrix(data=predicciones.categ, reference=datos.test$alerta)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Sí
##         No 49 10
##         Sí 16 65
##
##                Accuracy : 0.8143
##                  95% CI : (0.7398, 0.875)
##     No Information Rate : 0.5357
##     P-Value [Acc > NIR] : 4.805e-12
##
##                   Kappa : 0.6244
##
##  Mcnemar's Test P-Value : 0.3268
##
##             Sensitivity : 0.7538
##             Specificity : 0.8667
##          Pos Pred Value : 0.8305
##          Neg Pred Value : 0.8025
##              Prevalence : 0.4643
##          Detection Rate : 0.3500
##    Detection Prevalence : 0.4214
##       Balanced Accuracy : 0.8103
##
##        'Positive' Class : No
##
```

```r
# ROC and AUC
pr <- prediction(predicciones, datos.test$alerta)
# TPR = sensitivity, FPR=specificity
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf,main = "Diabetes controlada vs DMSES Regimen")
```




```
auc <- performance(pr, measure = "auc")
(auc <- auc@y.values[[1]])
```

```
## [1] 0.8020513
```

```
# significancia del modelo
anova(logregimen, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: alerta
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                            559     775.30
## DMSES.Regimen  1   8.7661       558     766.53 0.003069 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Añadir más variables no mejora sustancialmente el modelo logístico, y sólo con la puntuación DMSES.Regimen ya se tenía un buen resultado.

3.5 Modelo DMSES subscore Physical

```
logphysical <- glm(alerta ~ DMSES.Physical, data = datos.train, family = "binomial")
summary(logphysical)
```

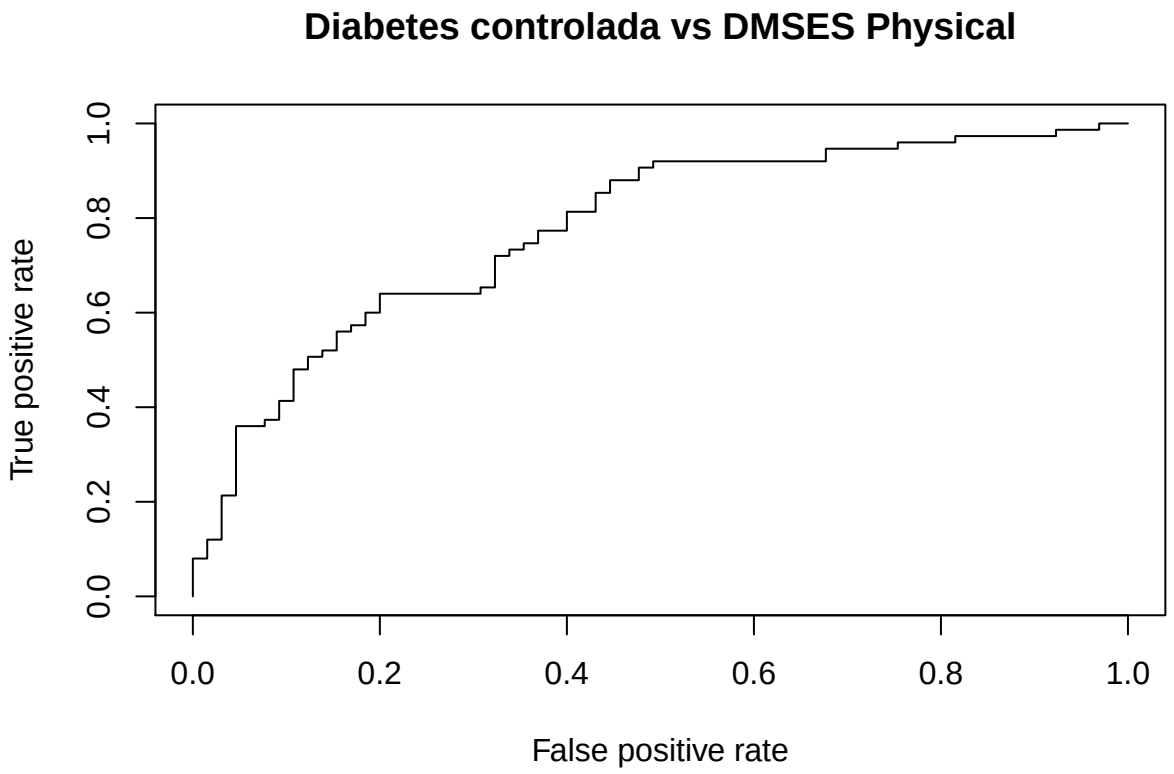
```
##
## Call:
## glm(formula = alerta ~ DMSES.Physical, family = "binomial", data = datos.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0110  -1.0739   0.6891   1.0350   1.6530
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.07947    0.08890   0.894    0.371
## DMSES.Physical -1.49921    0.21511  -6.970 3.18e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 775.3  on 559  degrees of freedom
## Residual deviance: 720.3  on 558  degrees of freedom
## AIC: 724.3
##
## Number of Fisher Scoring iterations: 4
```

```
predicciones <- predict(logphysical,newdata=datos.test[,-c(1)],type='response')
predicciones.categ <- ifelse(predicciones > 0.5,1,0)
predicciones.categ <- factor(predicciones.categ,levels=c(0,1),labels=c("No","Sí"))

# Confusion matrix
confusionMatrix(data=predicciones.categ, reference=datos.test$alerta)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Sí
##         No 45 26
##         Sí 20 49
##
##                Accuracy : 0.6714
##                  95% CI : (0.587, 0.7484)
##     No Information Rate : 0.5357
##     P-Value [Acc > NIR] : 0.00076
##
##                   Kappa : 0.3435
##
##  Mcnemar's Test P-Value : 0.46099
##
##             Sensitivity : 0.6923
```

```
##             Specificity : 0.6533
##          Pos Pred Value : 0.6338
##          Neg Pred Value : 0.7101
##              Prevalence : 0.4643
##          Detection Rate : 0.3214
##    Detection Prevalence : 0.5071
##       Balanced Accuracy : 0.6728
##
##        'Positive' Class : No
##
```

```
# ROC and AUC
pr <- prediction(predicciones, datos.test$alerta)
# TPR = sensitivity, FPR=specificity
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf,main = "Diabetes controlada vs DMSES Physical")
```




```
auc <- performance(pr, measure = "auc")
(auc <- auc@y.values[[1]])
```

```
## [1] 0.7780513
```

```
# significancia del modelo
anova(logphysical, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: alerta
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             559      775.3
## DMSES.Physical  1   54.997       558      720.3 1.207e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Añadir más variables no mejora sustancialmente el modelo logístico, y sólo con la puntuación DMSES.Physical ya se tenía un buen resultado.

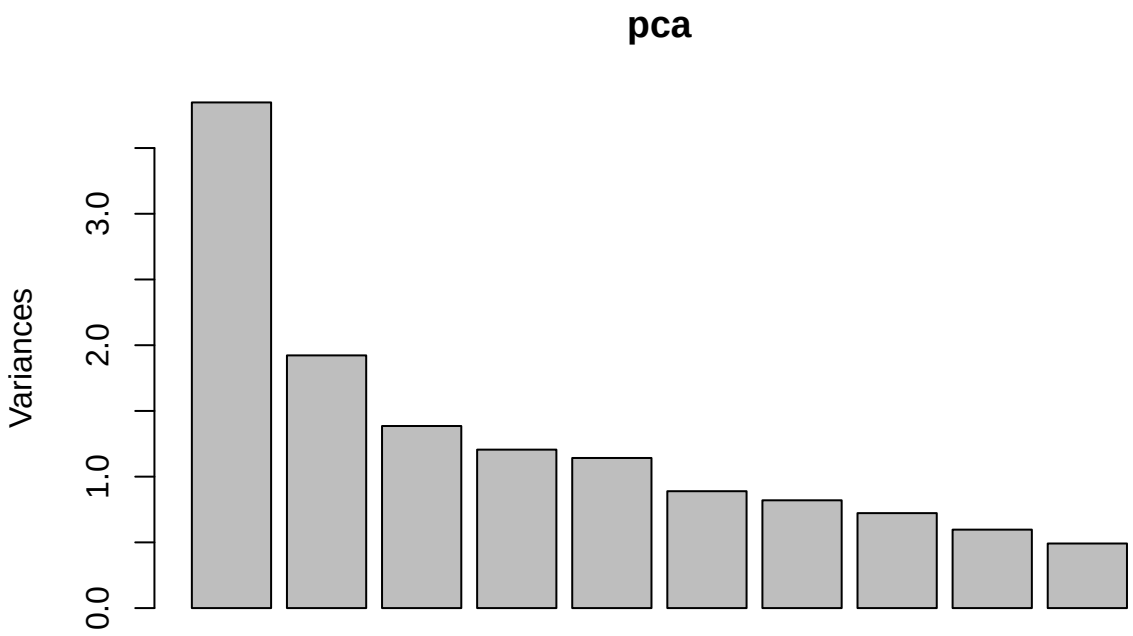
4. Análisis multivariante de datos

4.1 Análisis de componentes principales

```
pca <- prcomp(datos[,vars_cont],scale. = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.9614 1.3865 1.17696 1.0979 1.06861 0.94293 0.90553
## Proportion of Variance 0.2748 0.1373 0.09895 0.0861 0.08157 0.06351 0.05857
## Cumulative Proportion  0.2748 0.4121 0.51104 0.5971 0.67871 0.74221 0.80078
##                           PC8     PC9   PC10    PC11    PC12    PC13     PC14
## Standard deviation     0.8499 0.77263 0.7010 0.65595 0.62517 0.39650 1.46e-15
## Proportion of Variance 0.0516 0.04264 0.0351 0.03073 0.02792 0.01123 0.00e+00
## Cumulative Proportion  0.8524 0.89502 0.9301 0.96085 0.98877 1.00000 1.00e+00
```

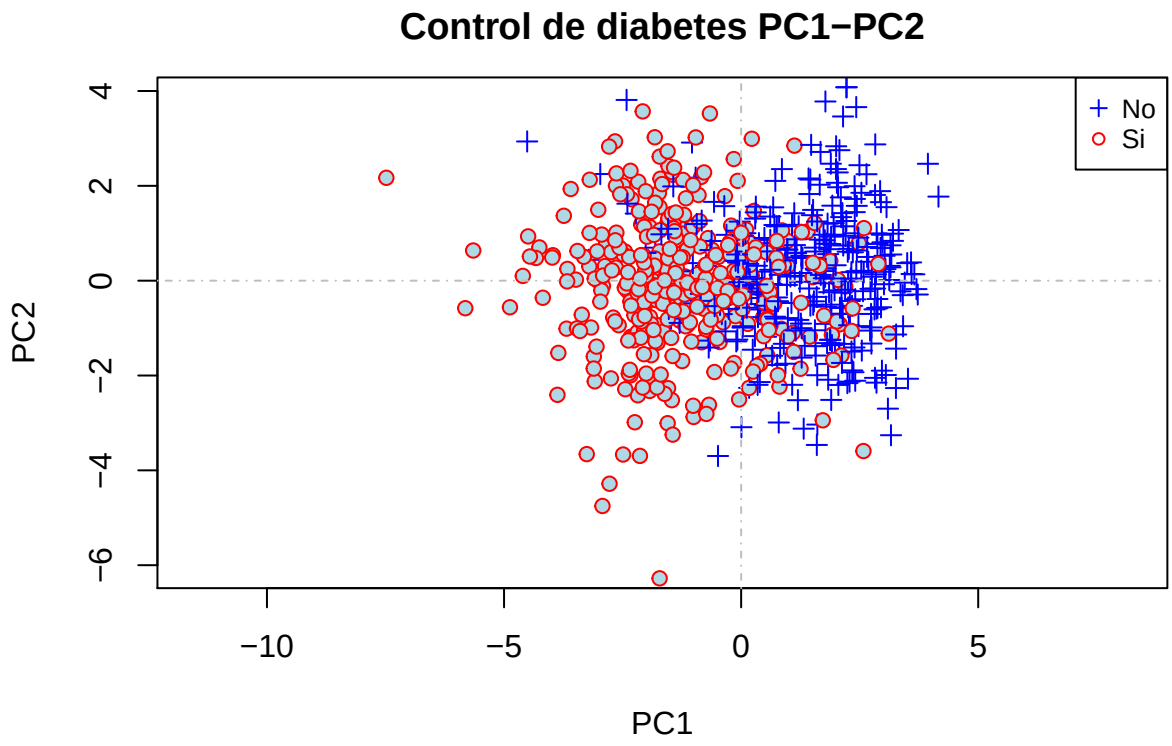
```
plot(pca)
```




```
round(pca$rotation[,1:10],2)
```

```
##                  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
## age             0.12  0.40 -0.51  0.11 -0.09  0.10 -0.08  0.01 -0.10  0.52
## dmdura          0.00  0.40 -0.39  0.02 -0.33 -0.25 -0.37 -0.02  0.18 -0.56
## hba1c          -0.34 -0.13  0.03  0.05 -0.02 -0.24 -0.11 -0.46  0.56  0.15
## ldl            -0.05 -0.35 -0.07  0.55 -0.19 -0.09 -0.18 -0.42 -0.56 -0.08
## hdl             0.00 -0.18  0.08  0.52 -0.51  0.30  0.07  0.43  0.38  0.08
## trig           -0.07 -0.28 -0.17  0.20  0.45 -0.43 -0.42  0.52  0.04  0.11
## sbp            -0.05 -0.36 -0.54 -0.30 -0.16  0.10  0.02 -0.10  0.04  0.34
## dbp            -0.07 -0.49 -0.32 -0.31 -0.07  0.19  0.09  0.12  0.01 -0.42
## DMSES.Diet      0.46 -0.05  0.01 -0.02  0.00  0.14 -0.19  0.02 -0.02 -0.07
## DMSES.Monitor   0.45 -0.07  0.01 -0.06  0.01  0.02 -0.11 -0.02 -0.02  0.00
## DMSES.Physical  0.37 -0.18  0.16 -0.06 -0.05 -0.14 -0.21 -0.25  0.33  0.13
## DMSES.Regimen   0.22 -0.01 -0.25  0.23  0.05 -0.49  0.73  0.00  0.08 -0.09
## DMSES.Total     0.50 -0.11  0.00  0.01  0.00 -0.09 -0.01 -0.08  0.11 -0.01
## DK.10          -0.04 -0.05  0.26 -0.36 -0.59 -0.50 -0.03  0.23 -0.26  0.21
```

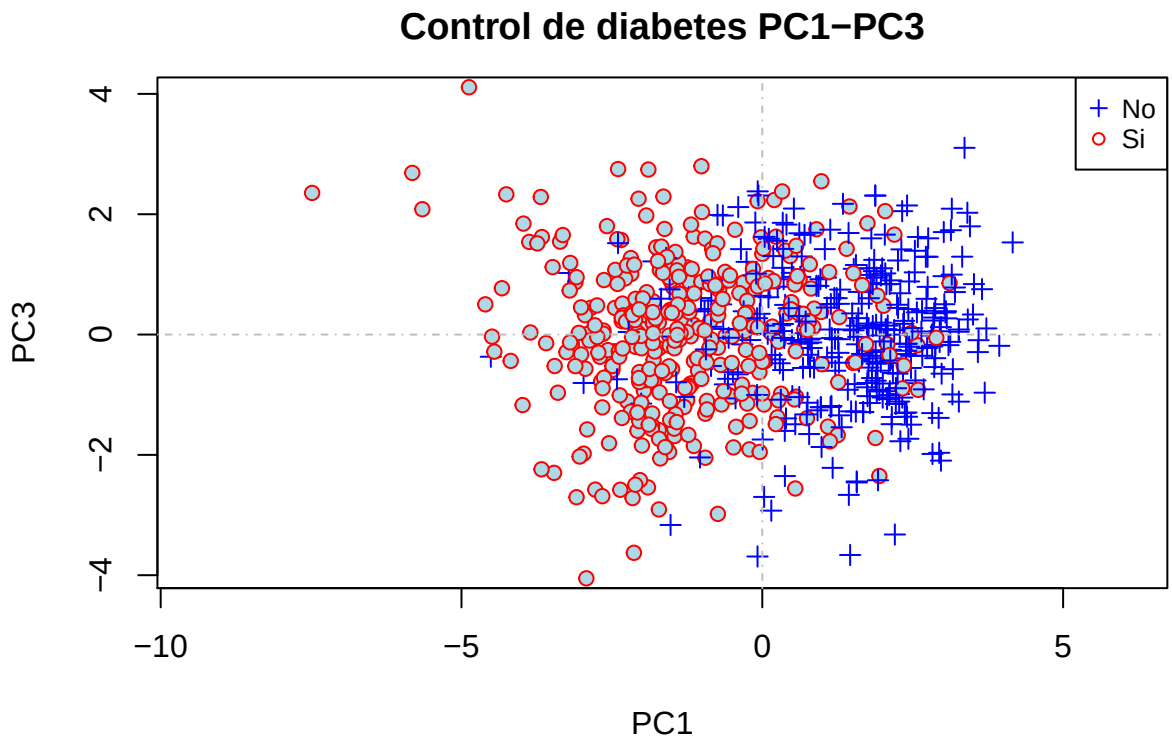
```
al <- as.numeric(datos$alerta)
#plot pc1-pc2
MASS::eqscplot(pca$x[,1], pca$x[,2], pch=c(3,21)[al],col=c("blue","red")[al],
         bg="lightblue",xlab="PC1",ylab="PC2")
abline(h=0,v=0,lty=4,col="gray")
legend("topright",pch=c(3,21),col=c("blue","red"), cex=0.8,legend=c("No","Si"))
title(main="Control de diabetes PC1-PC2",line=1)
```




```
round(summary(pca)$importance[ ,1:14],3)
```

```
##                            PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
## Standard deviation     1.961 1.386 1.177 1.098 1.069 0.943 0.906 0.850 0.773
## Proportion of Variance 0.275 0.137 0.099 0.086 0.082 0.064 0.059 0.052 0.043
## Cumulative Proportion  0.275 0.412 0.511 0.597 0.679 0.742 0.801 0.852 0.895
##                           PC10  PC11  PC12  PC13 PC14
## Standard deviation     0.701 0.656 0.625 0.396    0
## Proportion of Variance 0.035 0.031 0.028 0.011    0
## Cumulative Proportion  0.930 0.961 0.989 1.000    1
```

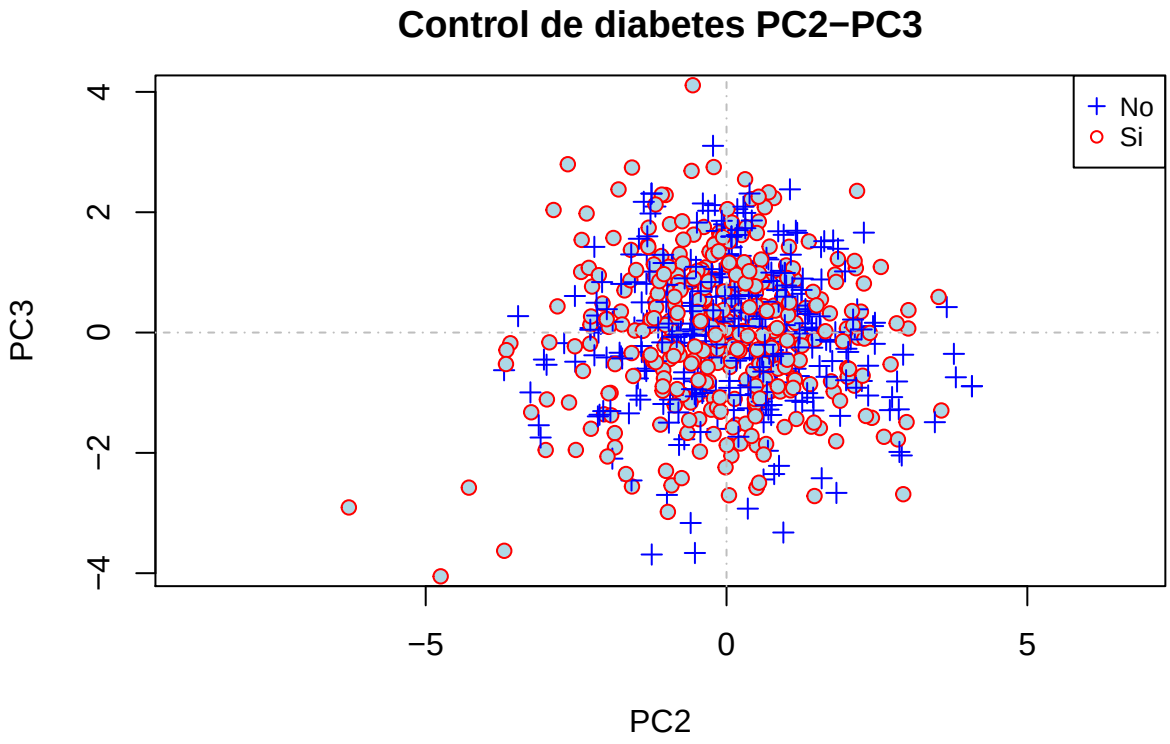
```
#plot pc1-pc3
MASS::eqscplot(pca$x[,1], pca$x[,3], pch=c(3,21)[al],col=c("blue","red")[al],
         bg="lightblue",xlab="PC1",ylab="PC3")
abline(h=0,v=0,lty=4,col="gray")
legend("topright",pch=c(3,21),col=c("blue","red"), cex=0.8,legend=c("No","Si"))
title(main="Control de diabetes PC1-PC3",line=1)
```




```
round(summary(pca)$importance[ ,1:14],3)
```

```
##                            PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
## Standard deviation     1.961 1.386 1.177 1.098 1.069 0.943 0.906 0.850 0.773
## Proportion of Variance 0.275 0.137 0.099 0.086 0.082 0.064 0.059 0.052 0.043
## Cumulative Proportion  0.275 0.412 0.511 0.597 0.679 0.742 0.801 0.852 0.895
##                           PC10  PC11  PC12  PC13 PC14
## Standard deviation     0.701 0.656 0.625 0.396    0
## Proportion of Variance 0.035 0.031 0.028 0.011    0
## Cumulative Proportion  0.930 0.961 0.989 1.000    1
```

```
#plot pc2-pc3
MASS::eqscplot(pca$x[,2], pca$x[,3], pch=c(3,21)[al],col=c("blue","red")[al],
         bg="lightblue",xlab="PC2",ylab="PC3")
abline(h=0,v=0,lty=4,col="gray")
legend("topright",pch=c(3,21),col=c("blue","red"), cex=0.8,legend=c("No","Si"))
title(main="Control de diabetes PC2-PC3",line=1)
```



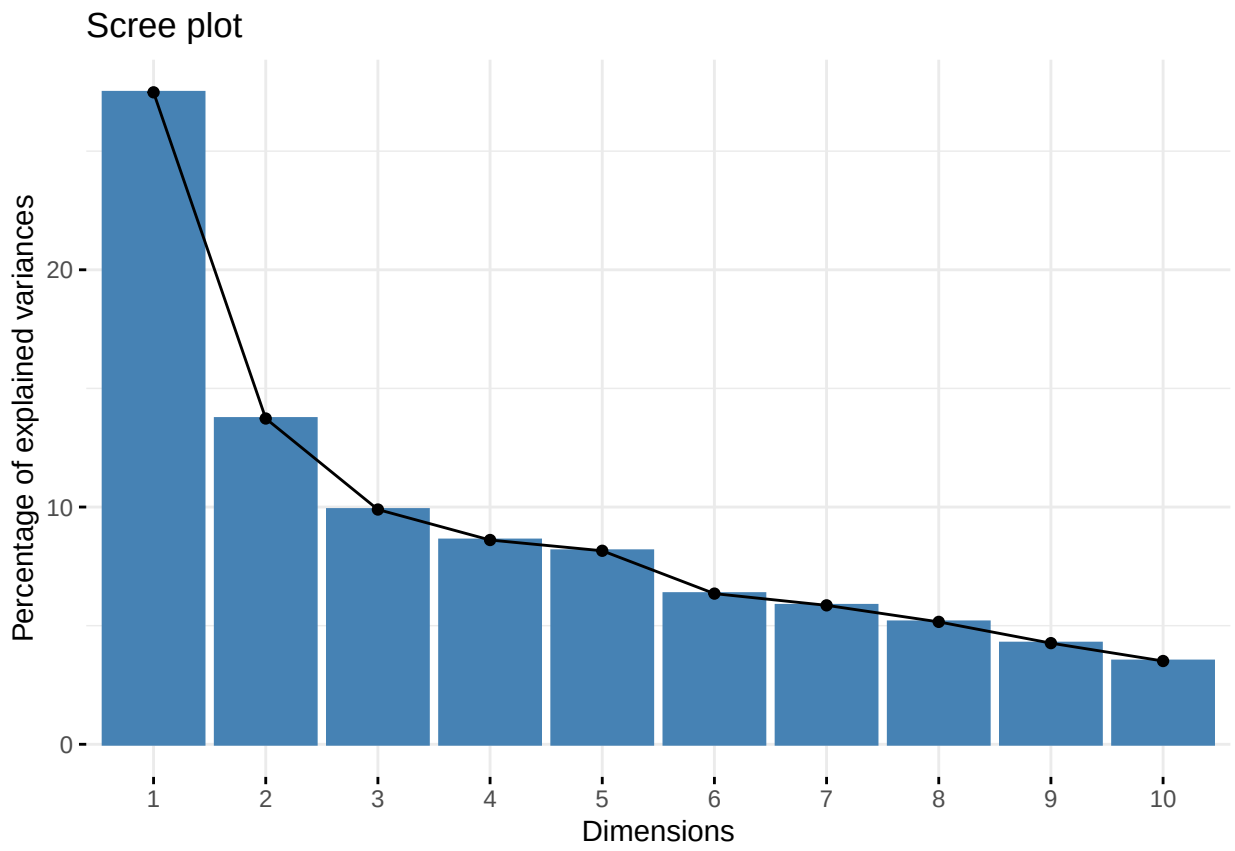

```
round(summary(pca)$importance[ ,1:14],3)
```

```
##                           PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
## Standard deviation     1.961 1.386 1.177 1.098 1.069 0.943 0.906 0.850 0.773
## Proportion of Variance 0.275 0.137 0.099 0.086 0.082 0.064 0.059 0.052 0.043
## Cumulative Proportion  0.275 0.412 0.511 0.597 0.679 0.742 0.801 0.852 0.895
##                          PC10  PC11  PC12  PC13 PC14
## Standard deviation     0.701 0.656 0.625 0.396    0
## Proportion of Variance 0.035 0.031 0.028 0.011    0
## Cumulative Proportion  0.930 0.961 0.989 1.000    1
```

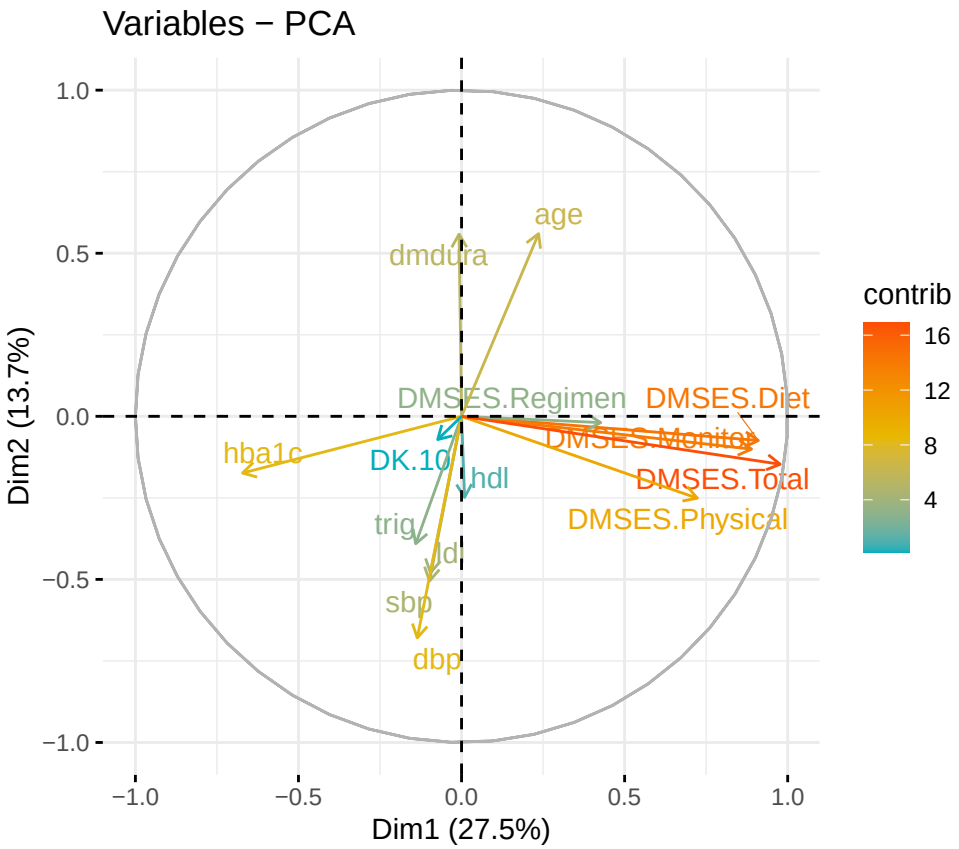
```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_eig(pca)
```




```
fviz_pca_var(pca,col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE )
```



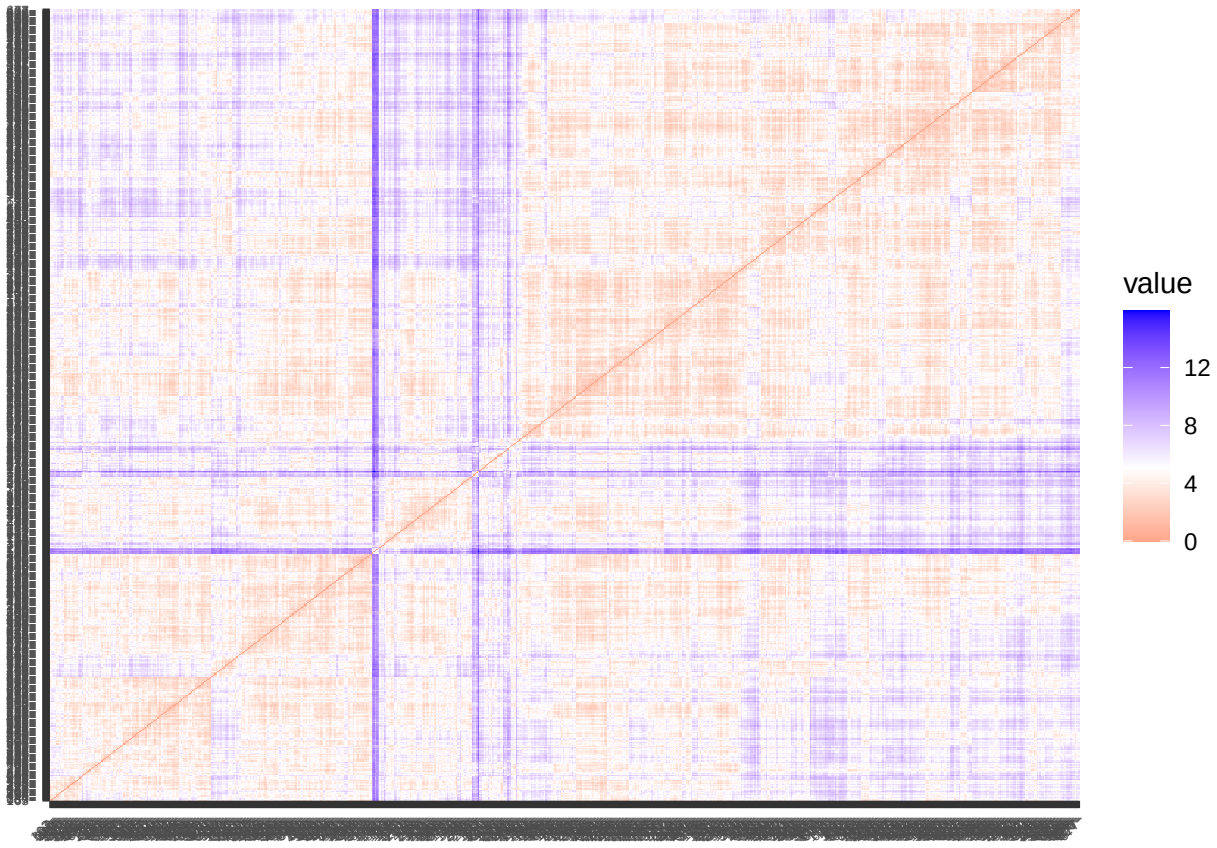
El analísis gráfico de componentes principales muestra la contribución de las variables en las dos primeras componentes principales. Se puede observar que en la segunda componente principal juegan un papel muy importante las puntuaciones del cuestionario DMSES por un lado y de manera opuesta la hemoglobina glicosilada (hb1ac). Este hecho podría indicar que mayor puntuación en el cuestionario DMSES implicaría una menor presencia de hemoglobina glicosilada.

4.2 Análisis de conglomerados

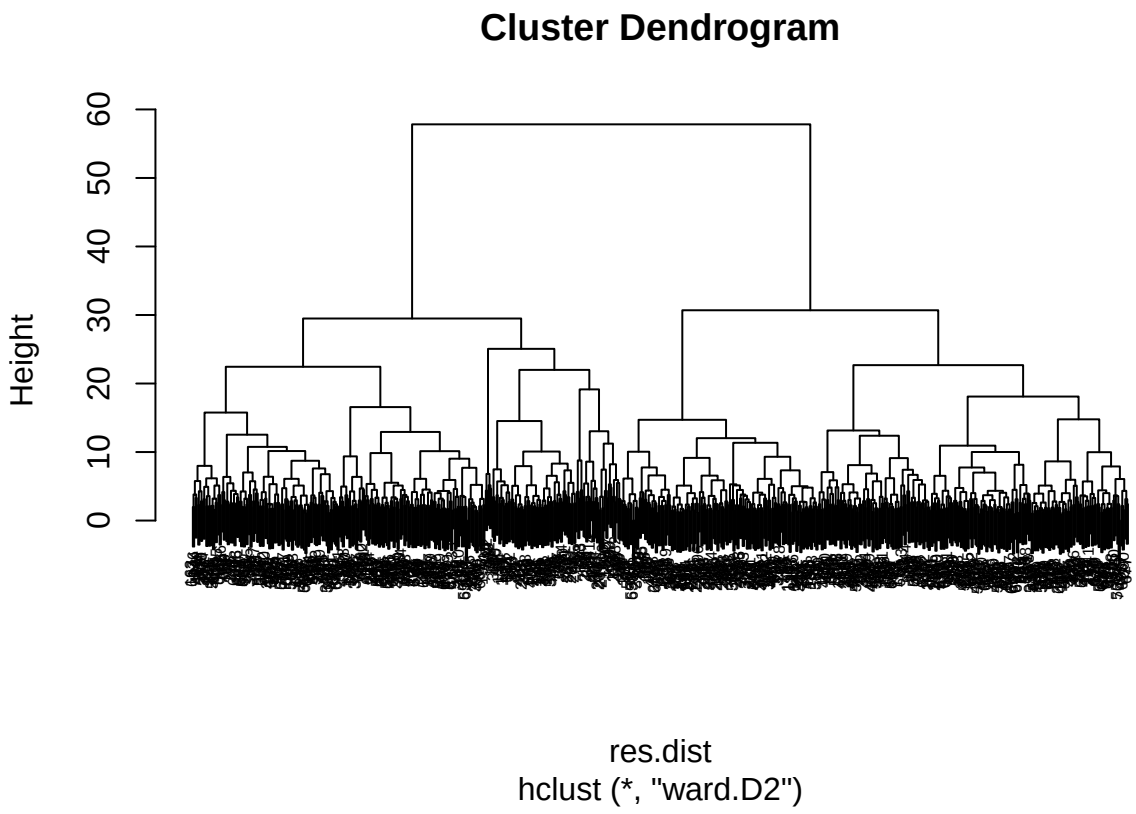
```
library(cluster)

# Fijamos la semilla aleatoria
set.seed(123)

df <- scale(datos[,vars_cont])

res.dist <- get_dist(df, method = "euclidean")
head(round(as.matrix(res.dist), 2))[, 1:5]
```

```
##      1    2    3    4    5
## 1 0.00 2.59 5.08 4.82 4.01
## 2 2.59 0.00 4.99 5.42 3.99
## 3 5.08 4.99 0.00 7.55 7.46
## 4 4.82 5.42 7.55 0.00 5.38
## 5 4.01 3.99 7.46 5.38 0.00
## 6 2.46 2.78 6.44 4.40 3.12
```

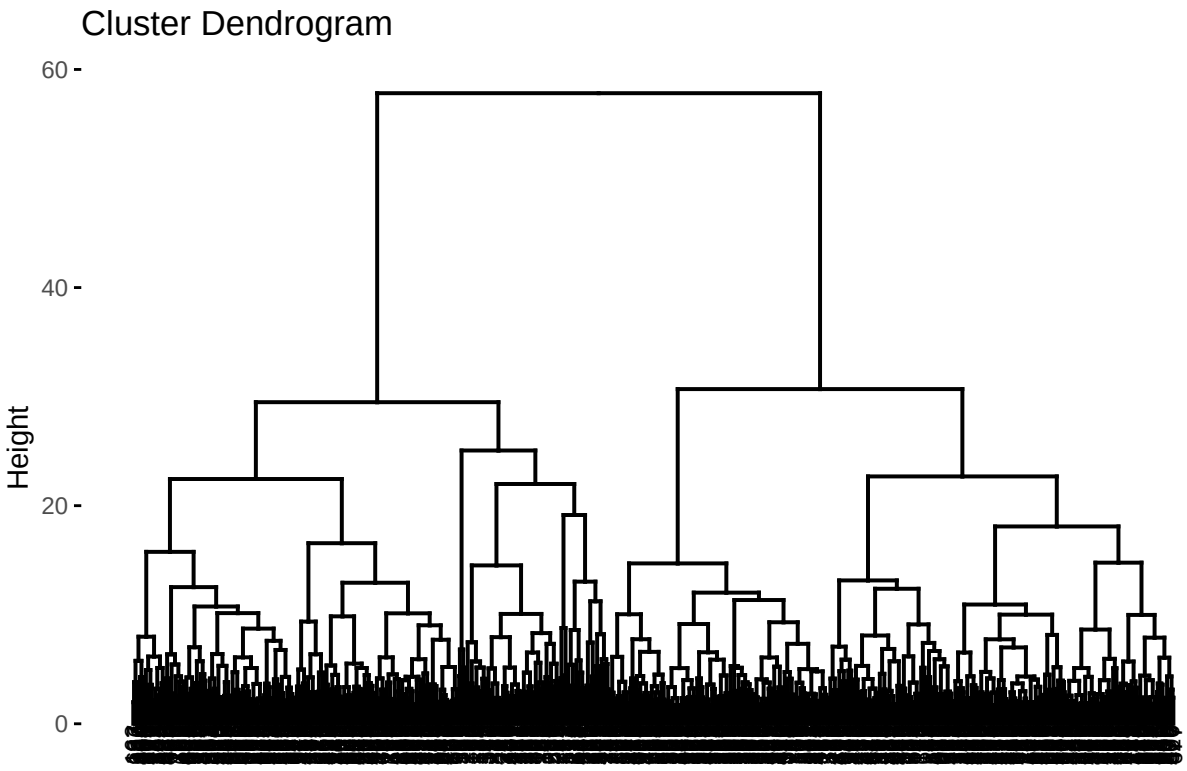
```r
# Visualize the dissimilarity matrix
fviz_dist(res.dist, lab_size = 5)
```



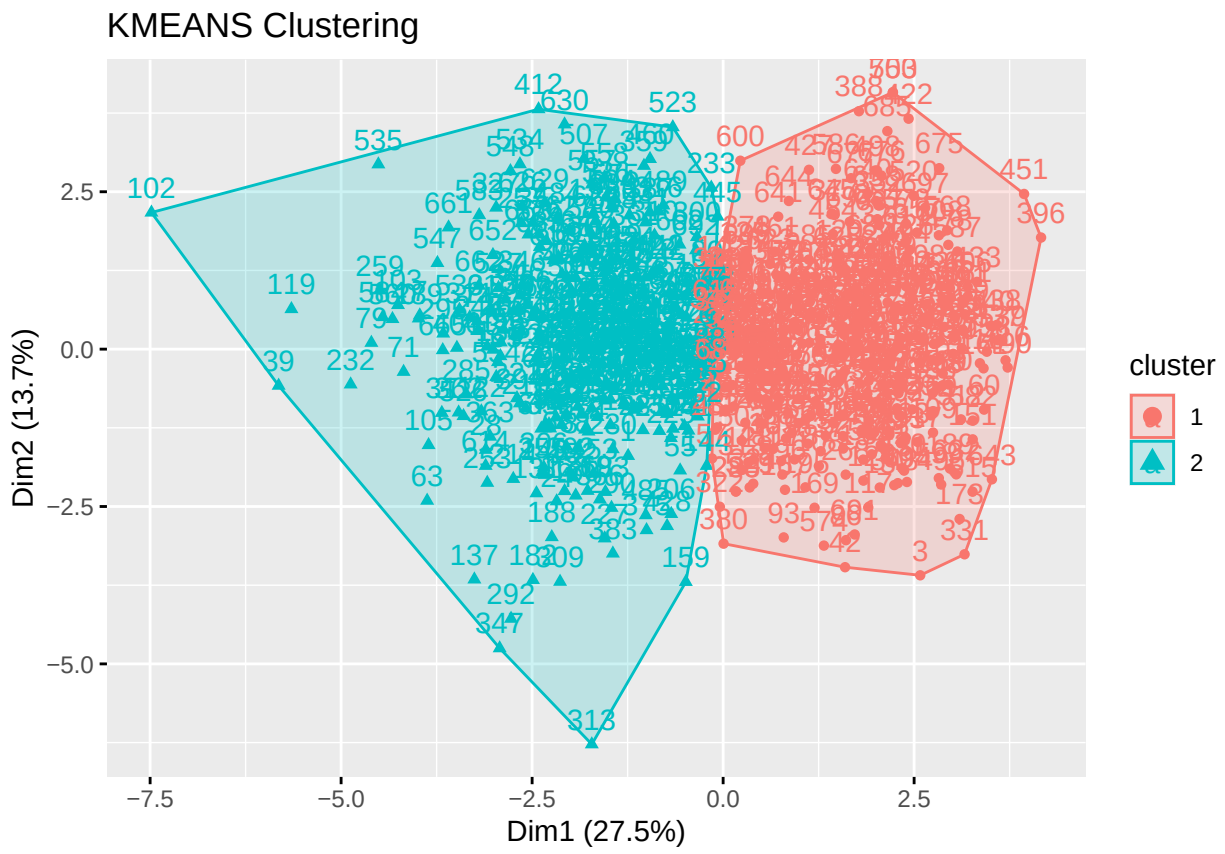

```r
# Compute hierarchical clustering
res.hc <- hclust(res.dist, method = "ward.D2")

# Visualize
plot(res.hc, cex = 0.5)
```



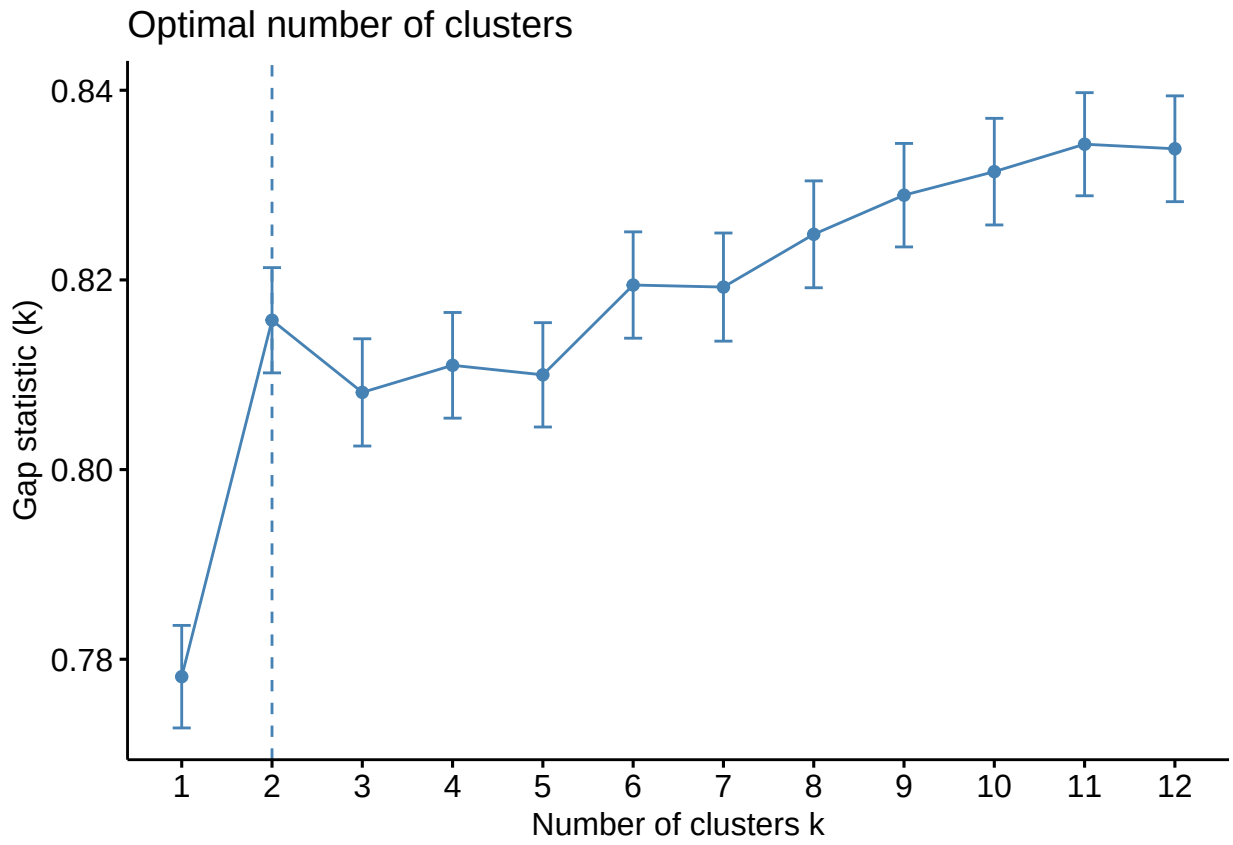

```
# dendograma
fviz_dend(res.hc, cex = 0.6, palette = "jco",
          rect = TRUE, rect_border = "jco", rect_fill = TRUE)
```




```
# Enhanced k-means clustering
res.km <- eclust(df, "kmeans", nstart = 25, k.max=12)
```

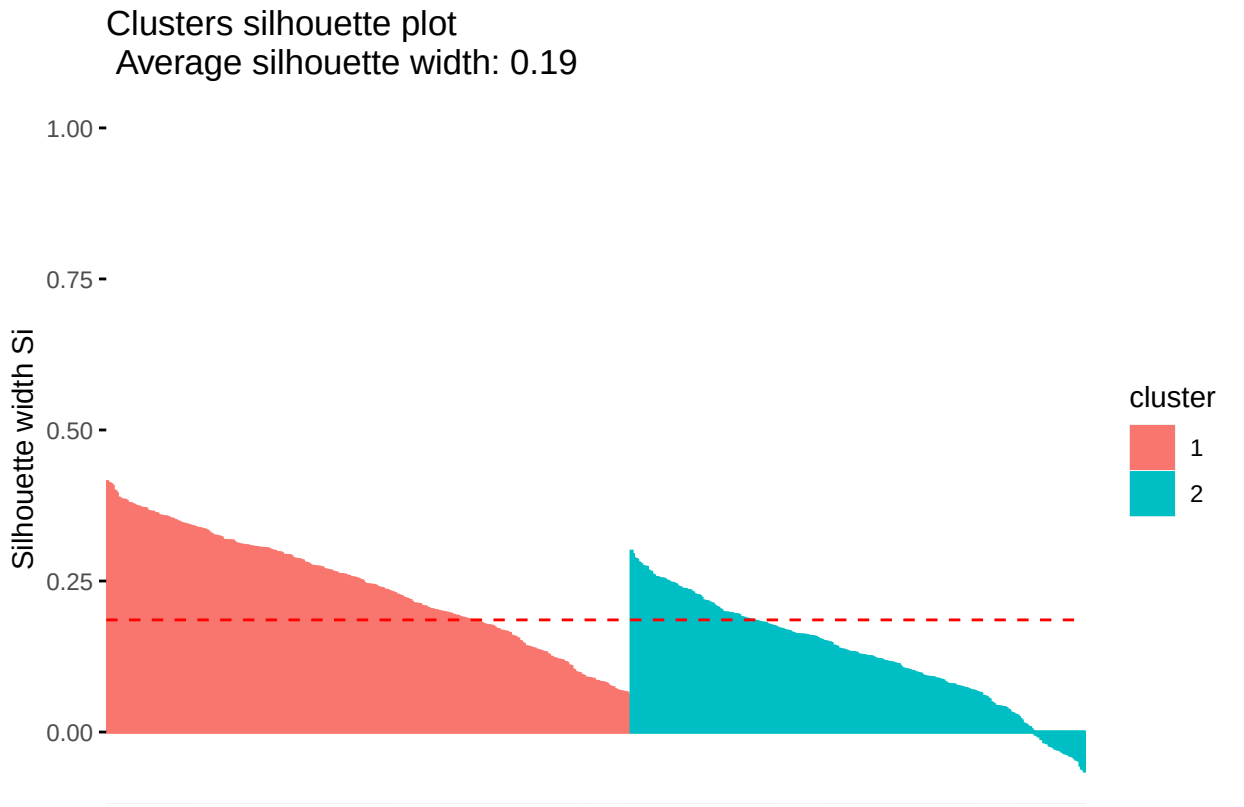



```
# Gap statistic plot
fviz_gap_stat(res.km$gap_stat)
```




```
# Silhouette plot
fviz_silhouette(res.km)
```

```
##   cluster size ave.sil.width
## 1       1  375          0.24
## 2       2  325          0.13
```




```
# Optimal number of clusters using gap statistics
res.km$nbclust
```

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

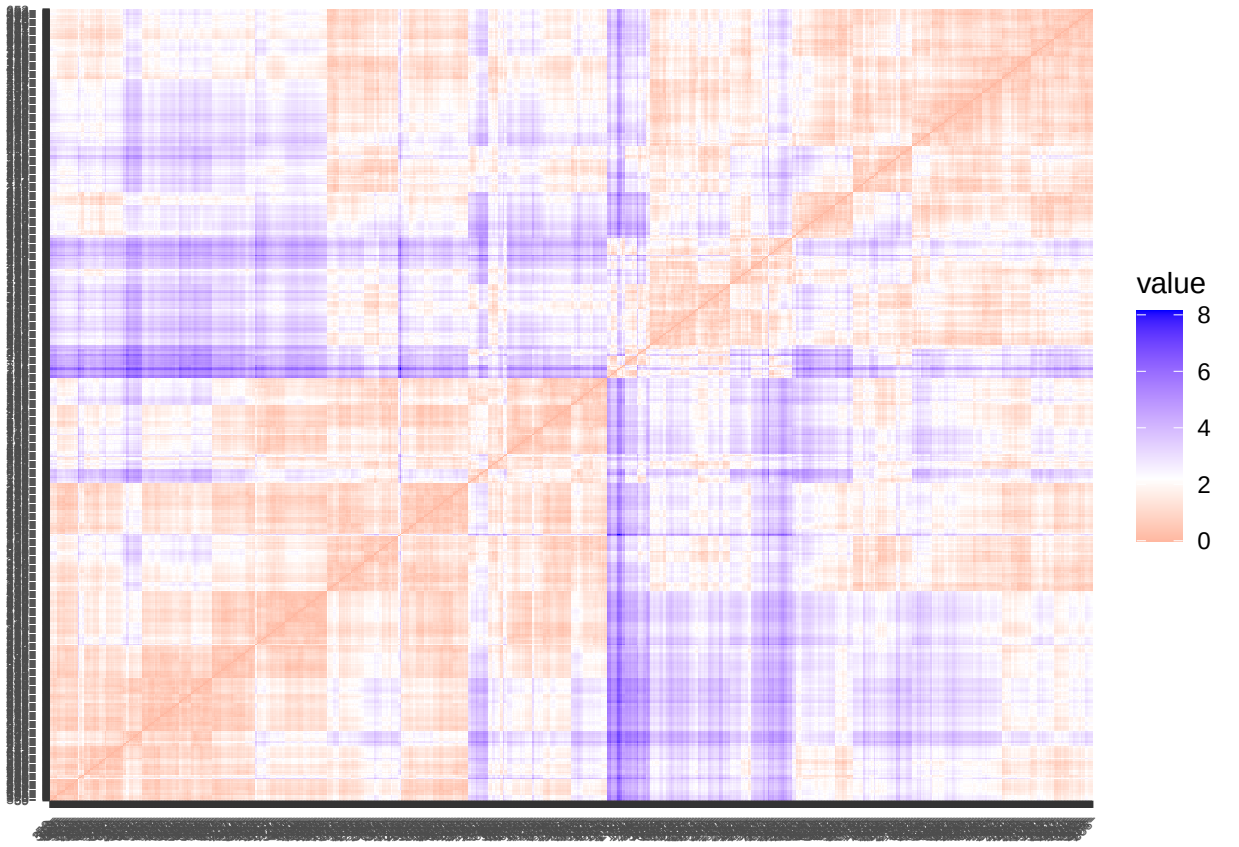
```
# Enhanced k-means clustering (forzando 11 conglomerados)
res.km11 <- eclust(df, FUNcluster = "kmeans",  nstart = 25, k = 11)

# Gap statistic plot
fviz_gap_stat(res.km11$gap_stat)

# Silhouette plot
fviz_silhouette(res.km11)
```

```
library(cluster)

# Fijamos la semilla aleatoria
set.seed(123)

df <- scale(datos[,c("hba1c","DMSES.Diet","age")])

res.dist <- get_dist(df, method = "euclidean")
head(round(as.matrix(res.dist), 2))[, 1:5]
```

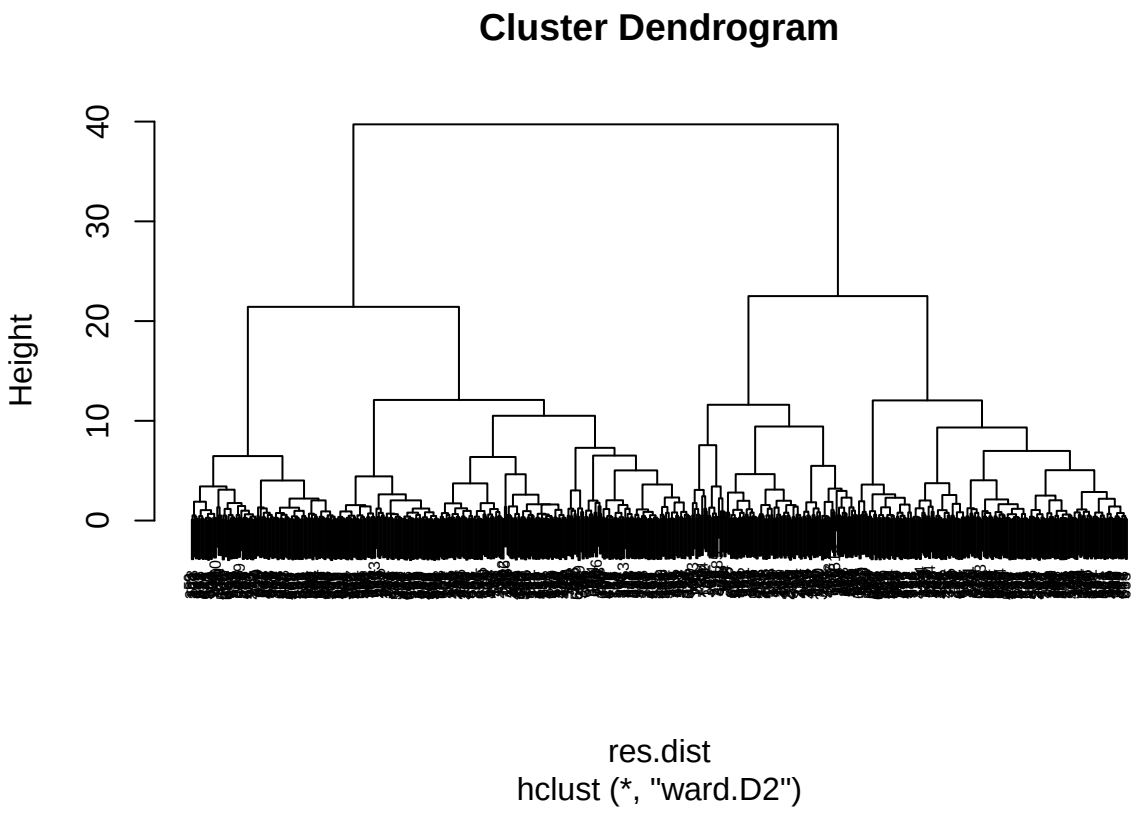
```
##      1    2    3    4    5
```

```
## 1 0.00 1.58 2.16 2.17 2.43
## 2 1.58 0.00 1.90 1.79 2.47
## 3 2.16 1.90 0.00 3.55 4.07
## 4 2.17 1.79 3.55 0.00 1.82
## 5 2.43 2.47 4.07 1.82 0.00
## 6 1.32 1.13 2.76 0.92 1.70
```

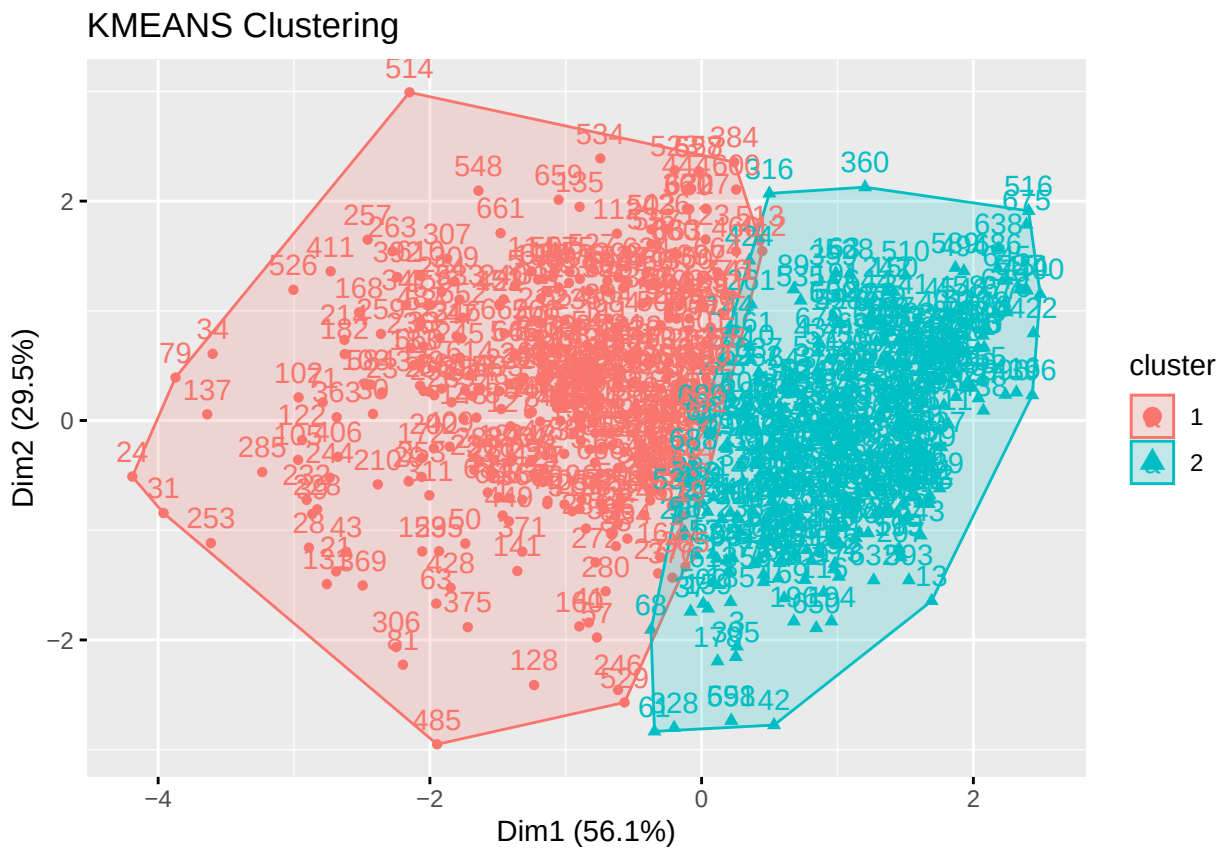
```r
# Visualize the dissimilarity matrix
fviz_dist(res.dist, lab_size = 5)
```



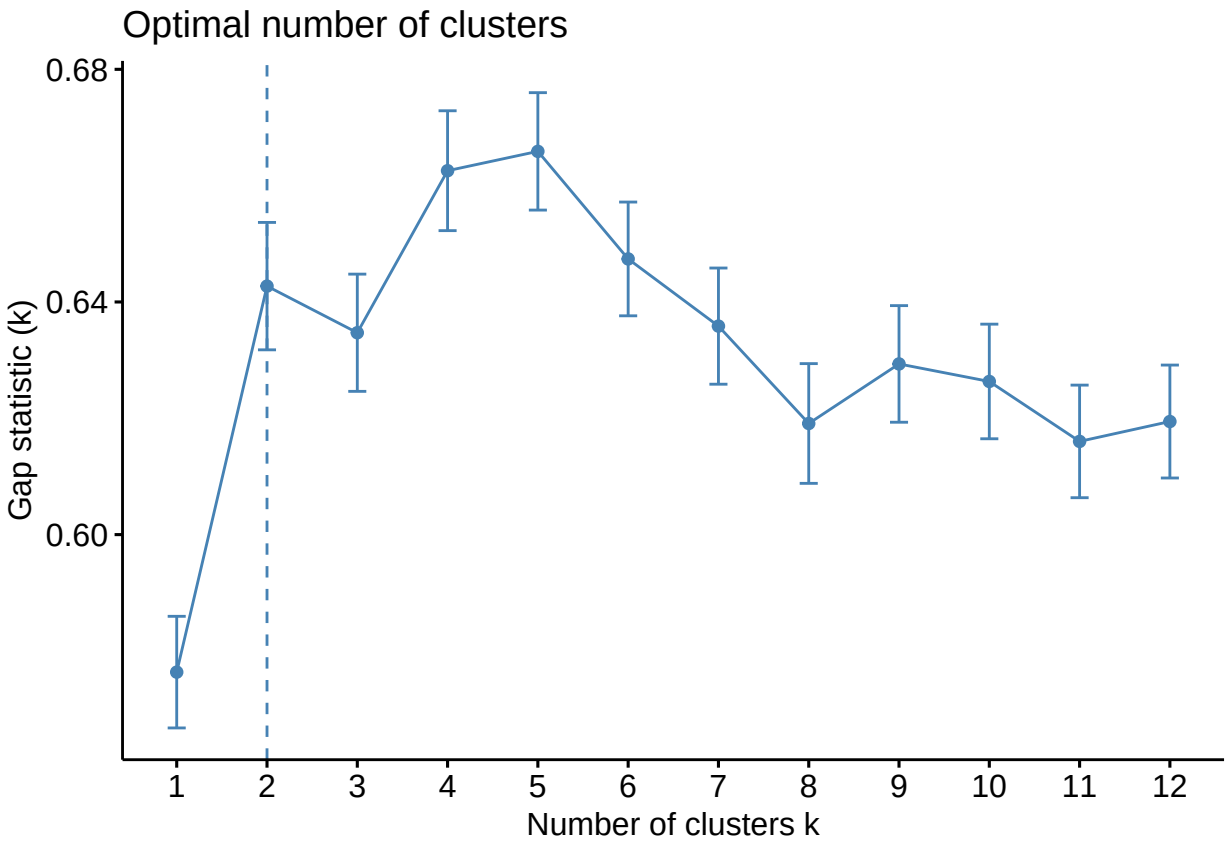

```r
# Compute hierarchical clustering
res.hc <- hclust(res.dist, method = "ward.D2")

# Visualize
plot(res.hc, cex = 0.5)
```




```
# Enhanced k-means clustering
res.km <- eclust(df, "kmeans", nstart = 25, k.max=12)
```

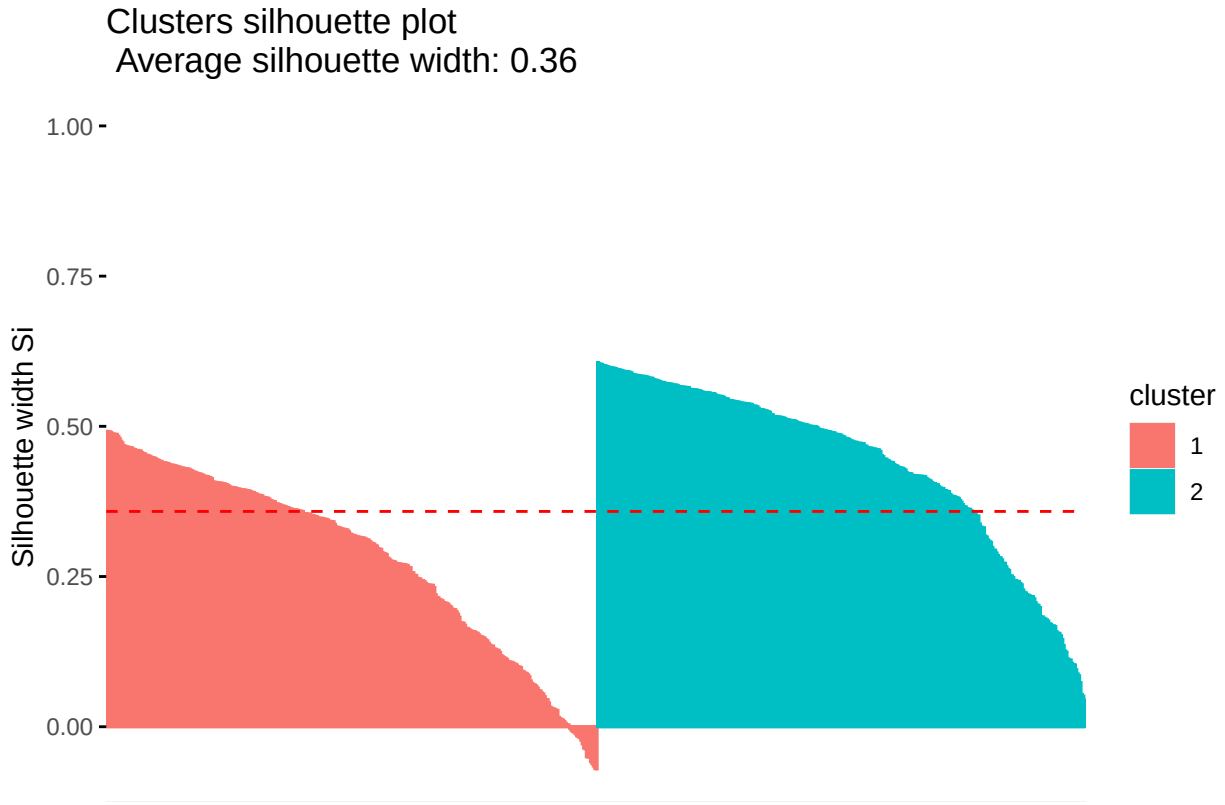



```
# Gap statistic plot
fviz_gap_stat(res.km$gap_stat)
```




```
# Silhouette plot
fviz_silhouette(res.km)
```

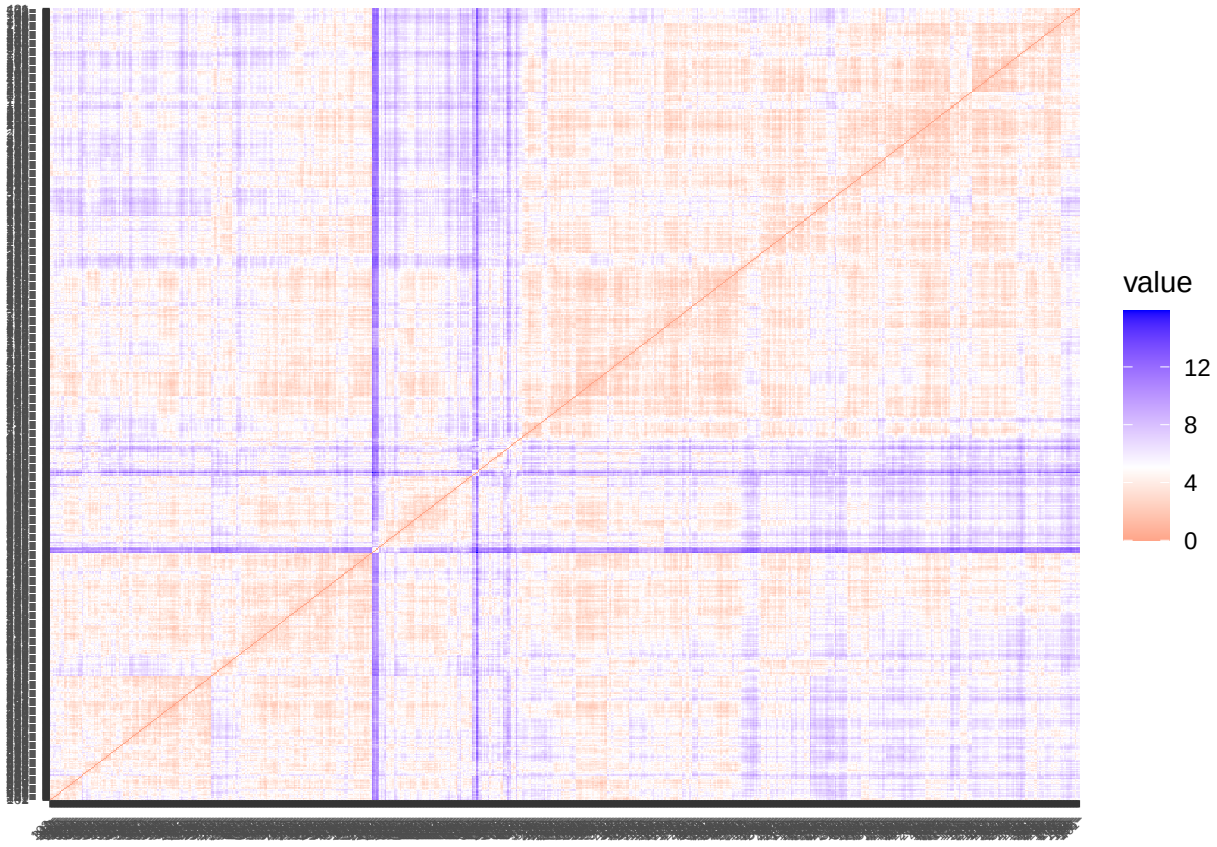
```
##   cluster size ave.sil.width
## 1       1  351          0.28
## 2       2  349          0.44
```



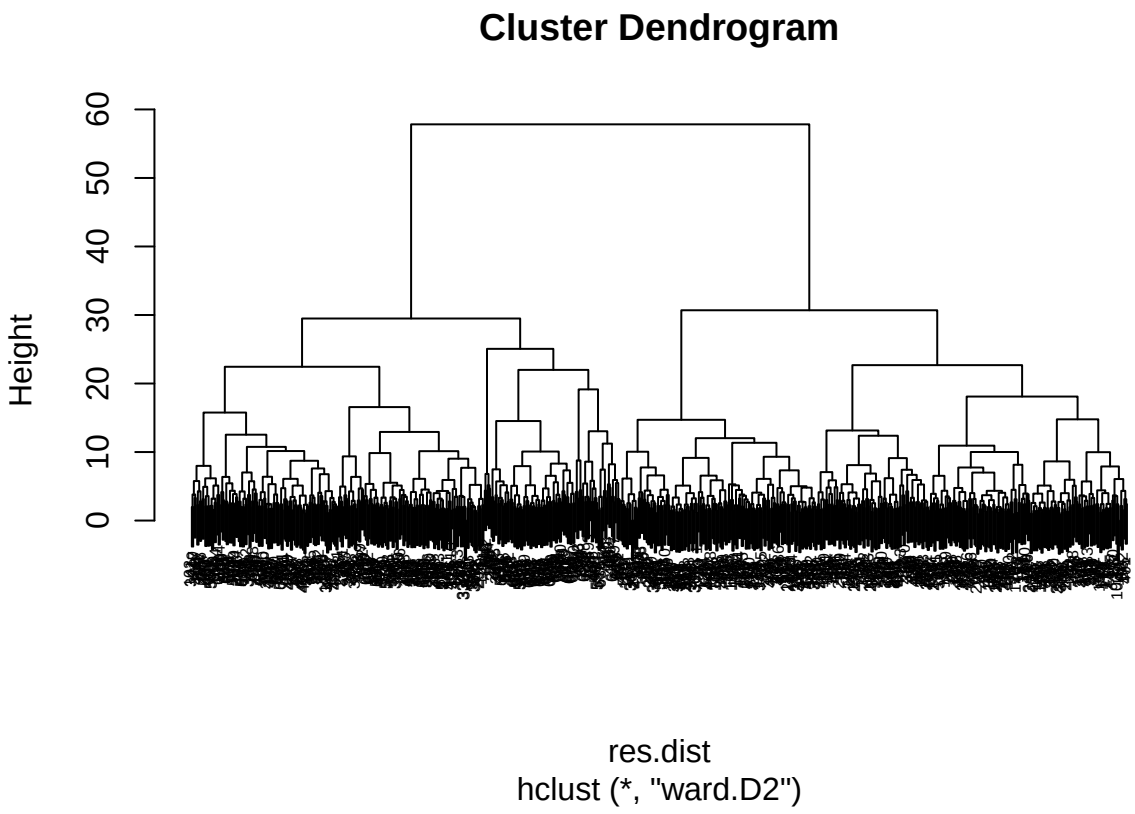

```
# Optimal number of clusters using gap statistics
res.km$nbclust
```

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

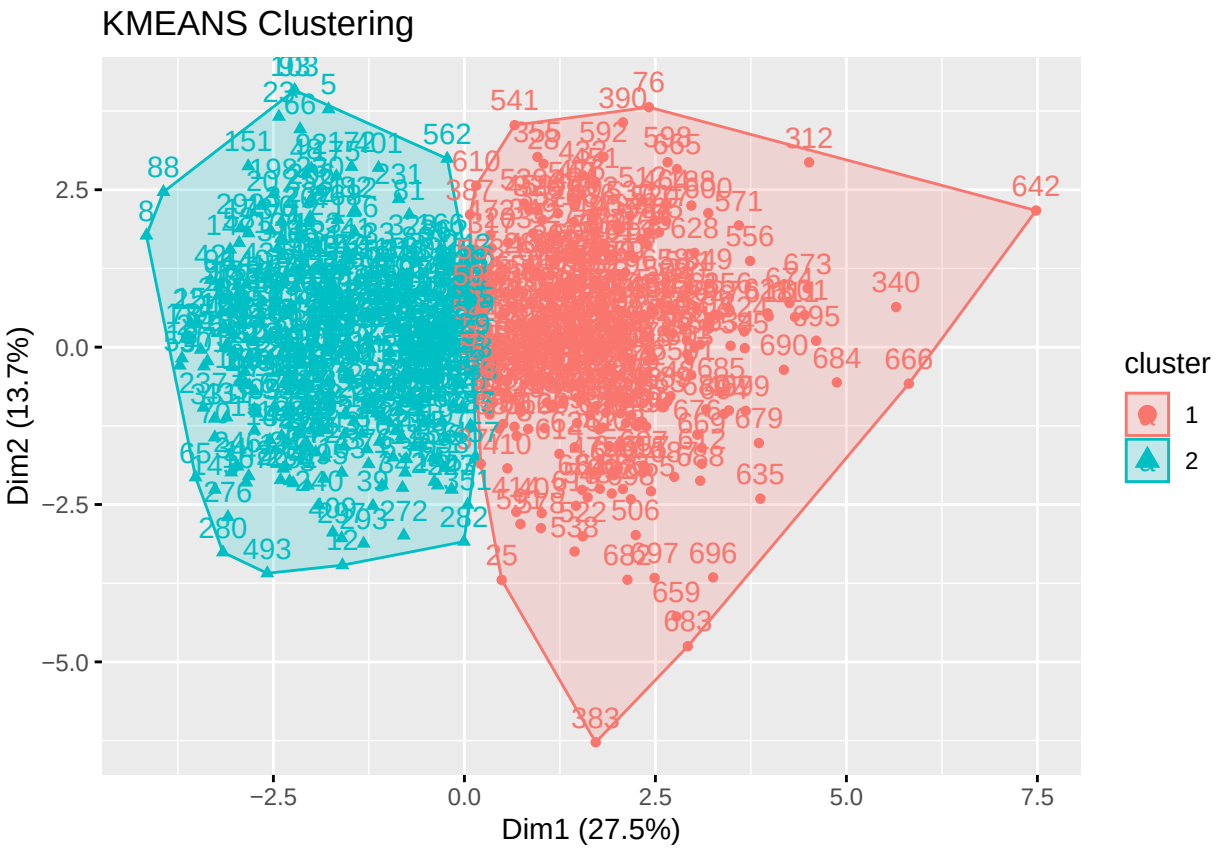
```
# ordenando por hba1c para tener los pacientes con diabetes controlada al principio y los otros al fina

datos.sort <- datos %>% arrange(hba1c)
df <- scale(datos.sort[,vars_cont])
res.dist <- get_dist(df, method = "euclidean")
fviz_dist(res.dist, lab_size = 5)
```




```
# Compute hierarchical clustering
res.hc <- hclust(res.dist, method = "ward.D2")
# Visualize
plot(res.hc, cex = 0.5)
```




```
# Enhanced k-means clustering
res.km <- eclust(df, "kmeans", nstart = 25, k.max=12)
```



4.3 Análisis discriminante

```
# Alerta vs Diet
fit.cv <- lda(alerta ~ DMSES.Diet, data=datos, CV=TRUE)
# Assess the accuracy of the prediction
# percent correct for each category of G
ct <- table(datos$alerta, fit.cv$class)
diag(prop.table(ct, 1))
```

```
##        No        Sí
## 0.7957958 0.8283379
```

```
prop.table(ct, 1)
```

```
##
##             No        Sí
##   No 0.7957958 0.2042042
##   Sí 0.1716621 0.8283379
```

```
# total percent correct
sum(diag(prop.table(ct)))
```

```
## [1] 0.8128571
```

```r
# Alerta vs Diet + age
fit.cv <- lda(alerta ~ DMSES.Diet + age, data=datos, CV=TRUE)
# Assess the accuracy of the prediction
# percent correct for each category of G
ct <- table(datos$alerta, fit.cv$class)
diag(prop.table(ct, 1))
```

```
##        No        Sí
## 0.8018018 0.8337875
```

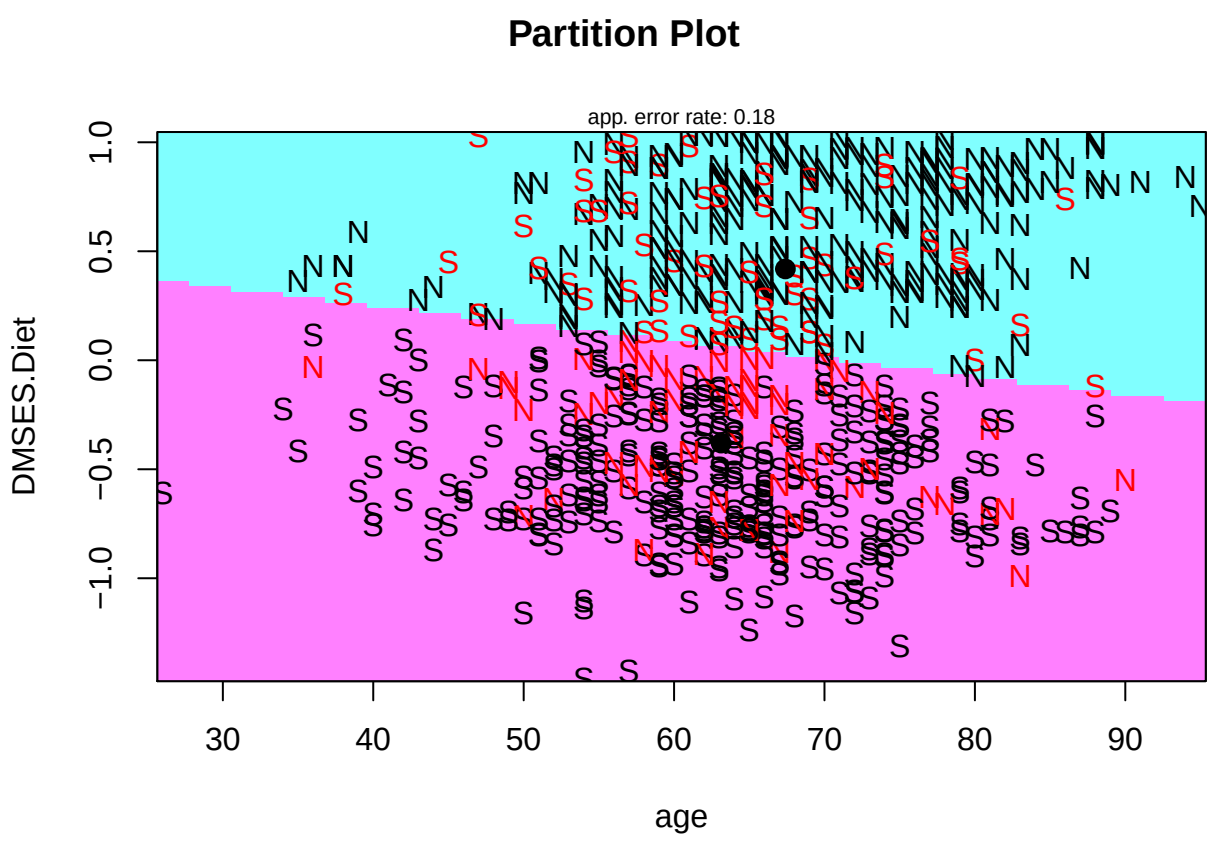
```r
prop.table(ct, 1)
```

```
##
##             No        Sí
##   No 0.8018018 0.1981982
##   Sí 0.1662125 0.8337875
```

```r
# total percent correct
sum(diag(prop.table(ct)))
```

```
## [1] 0.8185714
```

```r
partimat(alerta ~ DMSES.Diet + age, data=datos, method="lda")
```



```r
# Alerta vs otras variables
fit.cv <- lda(alerta ~ DMSES.Diet + age + ldl + hdl + trig + bmi, data=datos, CV=TRUE)
summary(fit.cv)
```

```
##           Length Class  Mode
## class      700   factor numeric
## posterior 1400   -none- numeric
## terms        3   terms  call
## call         4   -none- call
## xlevels      0   -none- list
```

```r
# Assess the accuracy of the prediction
# percent correct for each category of G
ct <- table(datos$alerta, fit.cv$class)
diag(prop.table(ct, 1))
```

```
##        No        Sí
## 0.8018018 0.8337875
```

```r
prop.table(ct, 1)
```

```
##
##             No        Sí
##   No 0.8018018 0.1981982
##   Sí 0.1662125 0.8337875
```

```r
# total percent correct
sum(diag(prop.table(ct)))
```

```
## [1] 0.8185714
```

```r
prop.table(ct)
```

```
##
##              No         Sí
##   No 0.38142857 0.09428571
##   Sí 0.08714286 0.43714286
```

```r
# Analisis discriminante cuadrático
datos.qda <- qda(alerta ~ DMSES.Diet + age, data=datos)
summary(datos.qda)
```

```
##         Length Class  Mode
## prior   2      -none- numeric
## counts  2      -none- numeric
## means   4      -none- numeric
## scaling 8      -none- numeric
## ldet    2      -none- numeric
## lev     2      -none- character
## N       1      -none- numeric
## call    3      -none- call
## terms   3      terms  call
## xlevels 0      -none- list
```

```r
clq <- predict(datos.qda, datos[,c("DMSES.Diet","age")])$class
ct <- table(datos$alerta, clq)
prop.table(ct)
```

```
##     clq
##              No         Sí
##   No 0.37285714 0.10285714
##   Sí 0.08285714 0.44142857
```