Examen año 2023

Code

Jose Valdes

2023-07-18

Hide

```
#limpio la memoria
rm( list= ls(all.names= TRUE) )  #remove all objects
gc( full= TRUE )                 #garbage collection
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 510321 27.3    1135937 60.7   644245 34.5
## Vcells 919144  7.1    8388608 64.0  1635140 12.5
```

Se realiza validación de la instalación de los paquetes necesarios para ejecutar el script

Hide

```
# Bibliotecas a cargar

check_packages <- function(packages) {
  if (all(packages %in% rownames(installed.packages()))) {
    TRUE
  } else{
    cat(
      "Instalar los siguientes packages antes de ejecutar el presente script\n",
      packages[!(packages %in% rownames(installed.packages()))],
      "\n"
    )
  }
}
packages_needed <- c("readxl","ggplot2","MVN","gridExtra","aod","MASS","carData","car","robustbase","l
eaps","olsrr","gamlss","lsr","ggpubr","lmtest","ResourceSelection","vcd","pROC","ROCR","randomFores
t","kableExtra","splitstackshape")

# Se llama a la funcion check_packages
check_packages(packages_needed)
```

```
## [1] TRUE
```

Hide

```
library(readxl)
library(ggplot2)
library(MVN)
library(gridExtra)
library(aod)
library(MASS)
library(carData)
library(car)
library(robustbase)
library(leaps)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

Hide

```
library(gamlss)
```

```
## Loading required package: splines
```

```
## Loading required package: gamlss.data
```

```
##
## Attaching package: 'gamlss.data'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
## Loading required package: gamlss.dist
```

```
## Loading required package: nlme
```

```
## Loading required package: parallel
```

```
##  **********   GAMLSS Version 5.4-12  **********
```

```
## For more on GAMLSS look at https://www.gamlss.com/
```

```
## Type gamlssNews() to see new features/changes/bug fixes.
```

Hide

```
library(lsr)
library(ggpubr)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Hide

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-6   2023-06-27
```

Hide

```
library(vcd)
```

```
## Loading required package: grid
```

Hide

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Hide

```
library(ROCR)
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

Hide

```
library(kableExtra)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 3 > 1' in coercion to 'logical(1)'
```

Hide

```
library(splitstackshape)
```

Funciones:

Función de cumplimientos de supuestos

Hide

```
#Funcion de cumplimientos de supuestos
Respuesta <- matrix(0, nrow = 8, ncol = 1)

cumplimientoSupuestos <- function(modeloLineal) {
  #Supuesto de normalidad
  Normalidad=shapiro.test(modeloLineal$residuals)
    if (Normalidad$p.value>0.01){
      Respuesta[1,1]="Los residuos del modelo son normales basado en el test de Shapiro"

      p_value <- Normalidad$p.value
      texto <- paste("En este caso, como el valor p (", p_value, ") es mayor que el nivel de significa
ncia (0.01), no se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo ta
nto, se puede considerar que los residuos del modelo siguen una distribución normal.")
      Respuesta[2,1] <- texto


    } else {

      Respuesta[1,1]="Los residuos del modelo no son normales basado en el test de Shapiro"

      p_value <- Normalidad$p.value
      texto <- paste("En este caso, como el valor p (", p_value, ") es menor que el nivel de significa
ncia (0.01), se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tant
o, se puede considerar que los residuos del modelo no siguen una distribución normal.")
      Respuesta[2,1] <- texto


    }

  Respuesta[3,1]="----------------------------------------------------------------------------------
----------------"

  #Supuesto de homocedastisidad
  homocedastisidad=bptest(modeloLineal)
    if (homocedastisidad$p.value>0.05){
      Respuesta[4,1]="Los errores del modelo son homocedastico basado en el test de Breusch-Pagan"
      p_value <- homocedastisidad$p.value
      texto <- paste("En este caso, como el valor p (", p_value, ") es mayor que el nivel de significa
ncia establecido (0.05), no se tiene suficiente evidencia para rechazar la hipótesis nula de homocedas
ticidad. Por lo tanto, se puede considerar que los errores del modelo tienen varianzas constantes (hom
ocedasticidad).")
      Respuesta[5,1] <- texto


    }else {
      Respuesta[4,1]="Los errores del modelo no son homocedastico basado en el test de Breusch-Pagan"
      p_value <- homocedastisidad$p.value
      texto <- paste("En este caso, como el valor p (", p_value, ") es menor que el nivel de significa
ncia establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de homocedastic
idad. Por lo tanto, se puede considerar que los errores del modelo no tienen varianzas constantes (het
erosedasticos).")
      Respuesta[5,1] <- texto
    }

   Respuesta[6,1]="----------------------------------------------------------------------------------
-----------------"
  #Supuesto de independencia
  independencia=dwtest(modeloLineal)
```

```
    if (independencia$p.value>0.05){
      Respuesta[7,1]="Los errores del modelo son independientes basado en el test de Durbin-Watson"
      p_value <- independencia$p.value
      texto <- paste("En este caso, como el valor p (", p_value, ") es mayor que el nivel de significa
ncia establecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independe
ncia de los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del
modelo.")
      Respuesta[8,1] <- texto
    }else {
      Respuesta[7,1]="Los errores del modelo no son independientes basado en el test de Durbin-Watson"
      p_value <- independencia$p.value
      texto <- paste("En este caso, como el valor p (", p_value, ") es menor que el nivel de significa
ncia establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de independenci
a de los errores. Por lo tanto, se puede concluir que existe autocorrelación en los errores del model
o.")
      Respuesta[8,1] <- texto
    }

  return(Respuesta)

}

#cumplimientoSupuestos(model5eje1)
```

función resumen de cumplimientos de los modelos

Hide

```
resumenCumplimiento <- function(cantidadModelos, ...) {
  modelos <- list(...)
  resultado <- matrix(0, nrow = cantidadModelos + 1, ncol = 5)

  # Nombre de la columna
  resultado[1, 1] <- "Modelos"

  # Obtener los nombres de los modelos como texto
  modelos_texto <- as.character(substitute(list(...)))[-1]

  for (i in 1:cantidadModelos) {
    # Asignar el nombre del modelo a la matriz resultado
    resultado[i + 1, 1] <- modelos_texto[i]

    # Nombre de la columna
    resultado[1, 2] <- "Normalidad"

    # Se obtiene el resultado de la función cumplimientoSupuestos
    resultado[i + 1, 2] <- cumplimientoSupuestos(modelos[[i]])[1, 1]

    #Reducción del texto
    if(resultado[i + 1, 2]=="Los residuos del modelo son normales basado en el test de Shapiro"){
      resultado[i + 1, 2] <- "Hay normalidad"
    }else{
      resultado[i + 1, 2] <- "No hay normalidad"
    }

    # Nombre de la columna 3
    resultado[1, 3] <- "Homocedasticidad"

    # Se obtiene el resultado de la función cumplimientoSupuestos
    resultado[i + 1, 3] <- cumplimientoSupuestos(modelos[[i]])[4, 1]
    resultado[i + 1, 3]
    #Reducción del texto
    if(resultado[i + 1, 3]=="Los errores del modelo son homocedastico basado en el test de Breusch-Pag
an"){
      resultado[i + 1, 3] <- "Hay homocedasticidad"
    }else{
      resultado[i + 1, 3] <- "No hay homocedasticidad"
    }


    # Nombre de la columna 4
    resultado[1, 4] <- "Independencia"

    # Se obtiene el resultado de la función cumplimientoSupuestos
    resultado[i + 1, 4] <- cumplimientoSupuestos(modelos[[i]])[7, 1]
    resultado[i + 1, 4]
    #Reducción del texto
    if(resultado[i + 1, 4]=="Los errores del modelo son independientes basado en el test de Durbin-Wat
son"){
      resultado[i + 1, 4] <- "Hay independencia"
    }else{
      resultado[i + 1, 4] <- "No hay independencia"
    }

    # Nombre de la columna
    resultado[1, 5] <- "Cumplimiento"
```

```
    if (resultado[i + 1, 2] == "Hay normalidad" & resultado[i + 1, 3]=="Hay homocedasticidad" & result
ado[i + 1, 4]=="Hay independencia") {
      resultado[i + 1, 5] <- "Si"
    } else {
      resultado[i + 1, 5] <- "No"
    }
  }

  return(resultado)

}
```

función para encontrar la mejor transformación box y cox (pendiente validar)

Hide

```
mejorBoxYCox<- function(independiente, dependiente,datos){
  box_cox_result <- boxcox(independiente ~ dependiente, lambda = -2:2, data = datos)

  # Se encuentra el valor óptimo de lambda que maximiza el logaritmo de verosimilitud
  best_box_cox <- box_cox_result$x[which.max(box_cox_result$y)]

  # Se ajusta un modelo de regresión lineal utilizando la variable dependiente elevada a la potencia ó
ptima de lambda (best_box_cox) como la variable de respuesta y la variable independiente si es distint
a de cero, si es cero se realiza el logaritmo en base de 10 de la variable independiente.
  if (best_box_cox==0){
    model <- lm(log10(independiente) ~ dependiente, data = datos)

  } else {
    model <- lm((independiente)^(best_box_cox) ~ dependiente, data = datos)
  }



  return(cumplimientoSupuestos(model))
}
```

función grafica con intervalos

Hide

```
graficaMLIntervalos <- function(modelolineal,independiente,dependiente,dataset){
  ICcompleto<-predict(modelolineal, interval="confidence",level=0.95)
  IPcompleto<-predict(modelolineal,newdata=data.frame(independiente), interval="prediction",level=0.9
5)
  datos<-data.frame(independiente,dependiente,IPcompleto)

 grafica <- ggplot(data = datos, mapping = aes(x = independiente, y = precio)) +
    geom_point(color = "firebrick", size = 2) +
    labs(title = "Diagrama de dispersión con bandas de confianza y predicción", x = "tamanio") +
    geom_line(aes(y=lwr), color="red" , linetype="dashed" ) +
    geom_line(aes(y=upr), color="red" , linetype="dashed" ) +
    geom_smooth(method = "lm", se = TRUE, color = "black") +
    theme_bw() + theme(plot.title = element_text(hjust = 0.5))

 return(grafica)
}
```

Función para obtener el error absoluto promedio de un modelo

Hide

```
# Función para obtener el error absoluto promedio de un modelo
get_mean_absolute_error <- function(model) {
  #return(mean(abs(model$residuals)))
  residuals <- as.numeric(model$residuals)
  return(mean(abs(residuals)))
}
```

Funcion obtener errorres de los modelos

Hide

```
evaluarErrores_modelos <- function(modelos) {
  # Crear un vector con los nombres de los modelos
  model_names <- names(modelos)

  # Calcular los valores de error absoluto promedio para cada modelo en la lista de modelos
  mean_absolute_error_values <- sapply(modelos, get_mean_absolute_error)

  # Obtener el índice del modelo con el menor valor de error absoluto promedio
  min_index <- which.min(mean_absolute_error_values)

  if (min_index > 0) {
    # Obtener el nombre del modelo con el menor valor de error absoluto promedio
    best_model_name <- model_names[min_index]

    # Obtener el valor del error absoluto promedio del mejor modelo
    best_error <- mean_absolute_error_values[min_index]

    # Imprimir el nombre del modelo y el valor del error absoluto promedio
    cat("El modelo con el menor error absoluto promedio es", best_model_name, "con un error de", best_
error, "\n")

    # Imprimir el modelo con el menor valor de error absoluto promedio
    cat("El modelo sería:\n")
    print(modelos[[min_index]])
  } else {
    cat("No se pudo determinar el mejor modelo debido a un error en los nombres de los modelos.\n")
  }
}
```

función curva ROC

Hide

```
curvaROC <- function(muestra,No.muestra,dataset,x,modelo){
 #muestra: numero de muestra de la población
 #No.muestra: No. de datos de la muestra
 #dataset: Conjunto de datos utilizado
 #x: Variable predictora
 #modelo: Modelo utilizado para hacer el ajuste.

  set.seed(20231)


  entrenamiento<-sample(1:muestra,No.muestra)

  validacion<-c(1:muestra)[-entrenamiento]


  # Crear una nueva variable binaria para representar 'x'
  dataset$bin_x <- ifelse(x == "Yes", 1, 0)

  dataset_train<-dataset[entrenamiento,]
  dataset_test<-dataset[validacion,]
  bin_x_train<-dataset$bin_x[entrenamiento]
  table(bin_x_train)

  bin_x_test<-dataset$bin_x[validacion]
  dataset_new<-data.frame(dataset_test)
  dataset_new$origen<-bin_x_test

  real <- dataset_new$origen


  dataset_new<-data.frame(dataset_test)
  dataset_new$origen<-bin_x_test
  # Se obtienen las probabilidades predichas para cada clase
  predicciones <- predict(object = modelo, newdata = dataset_new, type = "response")


  predic <-prediction(predicciones,real)
  perf <-  performance(predic, "tpr","fpr")

  g=plot(perf,
     main = "Curva ROC",
     xlab="Tasa de falsos positivos",
     ylab="Tasa de verdaderos positivos")
  abline(a=0,b=1,col="blue",lty=2)
  grid()
  auc <- as.numeric(performance(predic,"auc")@y.values)
  legend("bottomright",legend=paste(" AUC =",round(auc,4)))
  return(g)
}
```

función metricas

Hide

```r
metricas <- function(modelo, dataset, muestra, NoDeLaMuestra, bin_mora) {
  resultado <- matrix(0, nrow = 6, ncol = 2)

  set.seed(20231)

  entrenamiento <- sample(1:muestra, NoDeLaMuestra)
  validacion <- c(1:muestra)[-entrenamiento]

  dataset_train <- dataset[entrenamiento, ]
  dataset_test <- dataset[validacion, ]
  bin_x_train <- bin_mora[entrenamiento]
  bin_x_test <- bin_mora[validacion]

  dataset_new <- data.frame(dataset_test)
  dataset_new$origen <- bin_x_test

  predicciones <- predict(object = modelo, newdata = dataset_new, type = "response")
  predict_value <- predicciones

  pred <- predict_value > 0.5
  TP <- sum(bin_x_test[pred] == 1)
  TN <- sum(bin_x_test[!pred] == 0)
  FP <- sum(bin_x_test[pred] == 0)
  FN <- sum(bin_x_test[!pred] == 1)

  resultado[1, 1] <- "Metricas"
  resultado[1, 2] <- "Resultado"

  resultado[2, 1] <- "precision"
  resultado[2, 2] <- precision <- TP / (TP + FP)

  resultado[3, 1] <- "recalln"
  resultado[3, 2] <- recall <- TP / (TP + FN)

  resultado[4, 1] <- "f1_score"
  resultado[4, 2] <- f1_score <- (2 * precision * recall) / (precision + recall)

  pred_test_RegLog_0_1 <- ifelse(predict_value > 0.5, 1, 0)
  error_RegLog <- mean(bin_x_test != pred_test_RegLog_0_1) * 100

  resultado[5, 1] <- "error"
  resultado[5, 2] <- error_RegLog

  # Calcular matriz de confusión
  matriz_confusion <- table(predicciones, bin_x_test)

  # Calcular métricas
  #precision <- matriz_confusion[2, 2] / sum(matriz_confusion[, 2])
  #recall <- matriz_confusion[2, 2] / sum(matriz_confusion[2, ])
  #f1_score <- 2 * precision * recall / (precision + recall)
  exactitud <- sum(diag(matriz_confusion)) / sum(matriz_confusion)

  resultado[6, 1] <- "exactitud"
  resultado[6, 2] <- exactitud

  return(resultado)
}
```

función metricas para modelos random forest

Hide

```
calcularMetricasRandomForest <- function(modelo, datos, variable_objetivo) {
  # Realizar predicciones utilizando el modelo
  predicciones <- predict(modelo, datos)

  # Calcular matriz de confusión
  matriz_confusion <- table(predicciones, datos[[variable_objetivo]])

  # Calcular métricas
  precision <- matriz_confusion[2, 2] / sum(matriz_confusion[, 2])
  recall <- matriz_confusion[2, 2] / sum(matriz_confusion[2, ])
  f1_score <- 2 * precision * recall / (precision + recall)
  exactitud <- sum(diag(matriz_confusion)) / sum(matriz_confusion)


  # Crear matriz de resultados
  resultados <- matrix(0, nrow = 6, ncol = 2)
  resultados[1, ] <- c("Métrica", "Valor")
  resultados[2, ] <- c("Precision", precision)
  resultados[3, ] <- c("Recall", recall)
  resultados[4, ] <- c("F1-Score", f1_score)
  resultados[5, ] <- c("Exactitud", exactitud)

  pred_test_RegLog_0_1 <- ifelse(as.numeric(predicciones) > 0.5, 1, 0)
  error_RegLog <- mean(datos[[variable_objetivo]] != pred_test_RegLog_0_1) * 100

  resultados[6, 1] <- "error"
  resultados[6, 2] <- error_RegLog

  return(resultados)
}
```

función mostrar tablas visualmente agradables

Hide

```
packages <- c('mctest','olsrr','ggplot2','dplyr','readxl','MVN','stringr',
  'gridExtra','ragg','reshape2','aod','lmtest','kableExtra','nortest',
  'car','MASS','GGally','stats','ggrepel','broom', 'leaps', 'MPV', 'faraway','caret', 'lars','glmnet')

packLoad <- lapply(packages, require, character.only = TRUE)
```

```
## Loading required package: mctest
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'mctest'
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following object is masked from 'package:nlme':
##
##     collapse
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: stringr
```

```
## Loading required package: ragg
```

```
## Loading required package: reshape2
```

```
## Loading required package: nortest
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## Loading required package: ggrepel
```

```
## Loading required package: broom
```

```
## Loading required package: MPV
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'MPV'
```

```
## Loading required package: faraway
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'faraway'
```

```
## Loading required package: caret
```

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

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

```
## The following object is masked from 'package:gamlss':
##
##     calibration
```

```
## Loading required package: lars
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'lars'
```

```
## Loading required package: glmnet
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'glmnet'
```

Hide

```
library(DT,kableextra)


mostrarDF <- function(df, rows=5){
  datatable(df,
    options = list(
    pageLength = rows#,
    #dom = 'pt'
    )
  )
}
```

Ejercicio 1

Hide

```
data<-read.csv2("C:/Users/Josvaldes/Documents/Maestria/Austral/1ano/regresionAvanzada/TPRegresion/TPRegresion/examen/2023/data_pancreas_resumen.csv")
data$sexo=as.factor(data$sexo)
data$diagnosis=as.factor(data$diagnosi)
data$estadio=as.factor(data$estadio)
mostrarDF(data)
```

Show 5 entries Search:

	paciente	edad	sexo	diagnosis	estadio	creatinina	LYVE1	REG1B	TFF1
1	S1	33	F	normal	NO	1.83222	0.8932192	52.94884	654.282174
2	S10	81	F	normal	NO	0.97266	2.037585	94.46703	209.48825
3	S100	51	M	normal	NO	0.78039	0.1455889	102.366	461.141
4	S101	61	M	normal	NO	0.70122	0.00280488	60.579	142.95
5	S102	62	M	normal	NO	0.21489	0.00085956	65.54	41.088

Showing 1 to 5 of 382 entries Previous 1 2 3 4 5 ... 77 Next

Hide

```
library(splitstackshape)
set.seed(1082884412)
strat_data <- stratified(data, "diagnosis", 300/nrow(data))
```

Hide

```
dim(strat_data)
```

```
## [1] 300   9
```

1. Construya un modelo lineal simple para explicar el valor de la creatinina en función de alguna de las restantes variables numéricas y evalúe la bondad del ajuste.

Hide

```
modelEje1P1=lm(creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo + diagnosis + estadio,data=strat_data)
summary(modelEje1P1)
```

```
##
## Call:
## lm(formula = creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo +
##     diagnosis + estadio, data = strat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6307 -0.3915 -0.1269  0.3146  2.0990
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.041e+00  2.638e-01   3.945 0.000100 ***
## edad            -1.070e-02  3.062e-03  -3.496 0.000547 ***
## LYVE1            5.371e-02  1.268e-02   4.236 3.06e-05 ***
## REG1B           -1.121e-04  2.282e-04  -0.491 0.623561
## TFF1             1.863e-04  4.113e-05   4.529 8.65e-06 ***
## sexoM            1.393e-01  6.866e-02   2.029 0.043375 *
## diagnosisnormal  2.473e-01  1.720e-01   1.437 0.151693
## estadioII       -9.780e-02  1.751e-01  -0.558 0.576979
## estadioIII       1.305e-01  1.804e-01   0.724 0.469799
## estadioIV        1.475e-02  2.135e-01   0.069 0.944953
## estadioNO               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5759 on 290 degrees of freedom
## Multiple R-squared:  0.264,  Adjusted R-squared:  0.2412
## F-statistic: 11.56 on 9 and 290 DF,  p-value: 1.668e-15
```

Implementando un modelo lineal ajustado utilizando todas las variables numéricas y categóricas disponibles arroja los siguientes resultados:

La ecuación del modelo es: creatinina = 1.041 * (Intercept) - 0.0107 * edad + 0.05371 * LYVE1 - 0.0001121 * REG1B + 0.0001863 * TFF1 + 0.1393 * sexoM + 0.2473 * diagnosisnormal - 0.0978 * estadioII + 0.1305 * estadioIII + 0.01475 * estadioIV.

Los coeficientes estimados para las variables independientes indican el cambio esperado en el valor de la creatinina cuando cada variable aumenta en una unidad, manteniendo todas las demás variables constantes.

Los valores p asociados a los coeficientes proporcionan una medida de la significancia estadística de cada variable en el modelo. Un valor p menor que el nivel de significancia elegido (0.01) indica que la variable es estadísticamente significativa.

En este caso, las variables “edad”, “LYVE1”, “TFF1” y “sexoM” tienen valores p menores que 0.01, lo que indica que son variables significativas en relación con la creatinina.

La variable “edad” tiene un valor p de 0.000547, que es menor que el nivel de significancia del 1%. Esto indica que la edad tiene un efecto significativo en el valor de la creatinina. Por cada unidad adicional de edad, se espera una disminución promedio de 0.0107 en el valor de la creatinina, manteniendo todas las demás variables constantes.

La variable “LYVE1” tiene un valor p de 3.06e-05, que es menor que el nivel de significancia del 1%. Esto indica que la proteína LYVE1 tiene un efecto significativo en el valor de la creatinina. Por cada aumento de 1 unidad en la proteína LYVE1, se espera un aumento promedio de 0.05371 en el valor de la creatinina, manteniendo todas las demás variables constantes.

La variable “TFF1” tiene un valor p de 8.65e-06, que es menor que el nivel de significancia del 1%. Esto indica que la proteína TFF1 tiene un efecto significativo en el valor de la creatinina. Por cada aumento de 1 unidad en la proteína TFF1, se espera un aumento promedio de 0.0001863 en el valor de la creatinina, manteniendo todas las demás variables constantes.

La variable “sexoM” tiene un valor p de 0.043375, que es menor que el nivel de significancia del 1%. Esto indica que el sexo masculino tiene un efecto significativo en el valor de la creatinina. En promedio, se espera que los hombres tengan un aumento promedio de 0.1393 en el valor de la creatinina en comparación con las mujeres, manteniendo todas las demás variables constantes.

Las demás variables (REG1B, diagnosis y estadio) no son significativas a un nivel de significancia del 1%. Esto significa que no hay suficiente evidencia para indicar que estas variables tienen un efecto significativo en el valor de la creatinina en el modelo.

Se intenta un nuevo modelo escluyendo estas variables:

Hide

```
model2Eje1P1=lm(creatinina ~ edad + LYVE1 + TFF1 + sexo ,data=strat_data)
summary(model2Eje1P1)
```

```
##
## Call:
## lm(formula = creatinina ~ edad + LYVE1 + TFF1 + sexo, data = strat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8133 -0.4252 -0.1100  0.3159  2.2153
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.411e+00  1.820e-01   7.754 1.45e-13 ***
## edad        -1.386e-02  2.962e-03  -4.679 4.40e-06 ***
## LYVE1        4.286e-02  1.144e-02   3.745 0.000217 ***
## TFF1         1.571e-04  3.484e-05   4.510 9.38e-06 ***
## sexoM        1.197e-01  6.854e-02   1.746 0.081825 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5843 on 295 degrees of freedom
## Multiple R-squared:  0.2295, Adjusted R-squared:  0.2191
## F-statistic: 21.97 on 4 and 295 DF,  p-value: 6.937e-16
```

Se disminiye el r cuadro al excluir las variables no significartivas.

2. Realice un análisis diagnóstico y de puntos influyentes e indique si el modelo es adecuado.

Hide

```
# Se valida el cumplimiento de supuestos
cumplimientoSupuestos(modelEje1P1)
```

```
##      [,1]
## [1,] "Los residuos del modelo no son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 1.03208077592181e-08 ) es menor que el nivel de significancia
(0.01), se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto, se
puede considerar que los residuos del modelo no siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo no son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 9.26070649733302e-06 ) es menor que el nivel de significancia
establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.
Por lo tanto, se puede considerar que los errores del modelo no tienen varianzas constantes (heterosed
asticos)."
## [6,] "----------------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.761389739186759 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

Se observa que el primer modelo no cumple los supuestos.

modelo 2

Hide

```
cumplimientoSupuestos(model2Eje1P1)
```

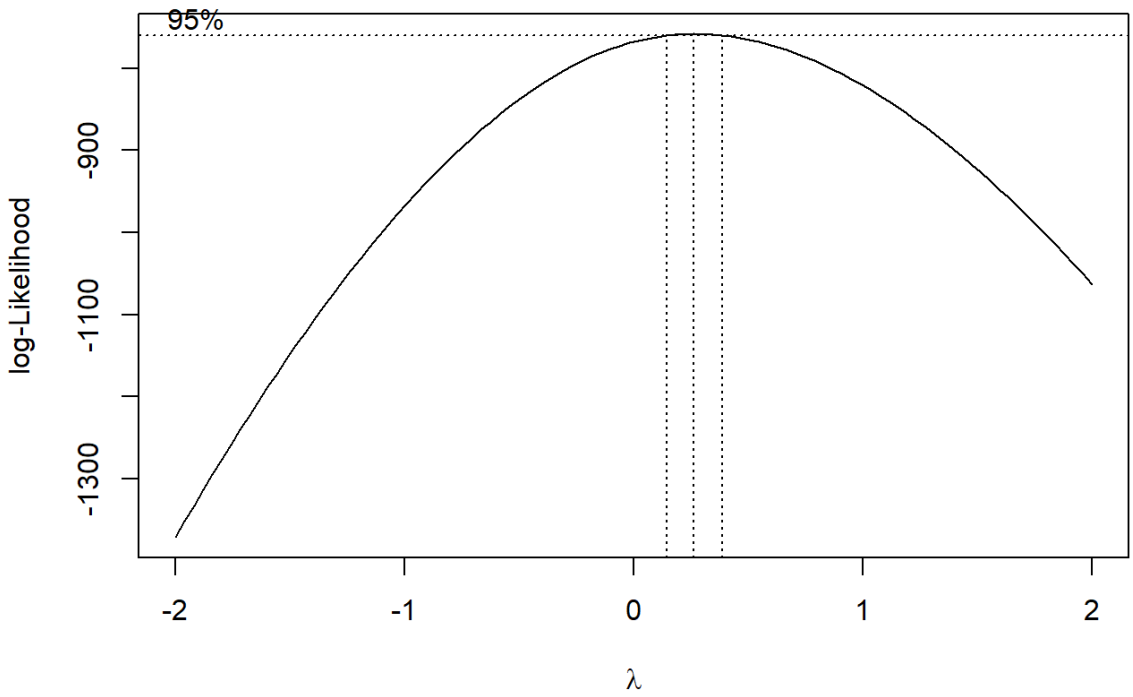
```
##      [,1]
## [1,] "Los residuos del modelo no son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 3.33646918049779e-09 ) es menor que el nivel de significancia
(0.01), se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto, se
puede considerar que los residuos del modelo no siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo no son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 1.19260458744812e-06 ) es menor que el nivel de significancia
establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.
Por lo tanto, se puede considerar que los errores del modelo no tienen varianzas constantes (heterosed
asticos)."
## [6,] "----------------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.447249432212957 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

El segundo modelo tampoco cumple los supuestos.

3. Realice una transformación de la variable respuesta para intentar lograr normalidad en la distribución de los residuos. Indique si el modelo con esta transformación resulta adecuado.

Hide

```
box_cox_result <- boxcox(creatinina ~ edad + LYVE1 + TFF1 + sexo , lambda = -2:2, data = strat_data)
```



Se valida el

lamda optimo

Hide

```
# Se encuentra el valor óptimo de lambda que maximiza el logaritmo de verosimilitud
best_box_cox <- box_cox_result$x[which.max(box_cox_result$y)]
best_box_cox
```

```
## [1] 0.2626263
```

Hide

```
# Se ajusta un modelo de regresión lineal utilizando la variable dependiente "precio" elevada a la pot
encia óptima de lambda (best_box_cox) como la variable de respuesta y la variable independiente "taman
io".
modeleje1P3 <- lm((creatinina)^(best_box_cox) ~ edad + LYVE1 + TFF1 + sexo, data = strat_data)

summary(modeleje1P3)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox) ~ edad + LYVE1 + TFF1 +
##     sexo, data = strat_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44386 -0.11493  0.00063  0.12662  0.39537
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0497016  0.0532712  19.705  < 2e-16 ***
## edad        -0.0036829  0.0008672  -4.247 2.91e-05 ***
## LYVE1        0.0142718  0.0033499   4.260 2.75e-05 ***
## TFF1         0.0000295  0.0000102   2.892  0.00411 **
## sexoM        0.0409534  0.0200628   2.041  0.04211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.171 on 295 degrees of freedom
## Multiple R-squared:  0.1922, Adjusted R-squared:  0.1813
## F-statistic: 17.55 on 4 and 295 DF,  p-value: 6.203e-13
```

Con la transformación siguen todas la variables signigicativas segun el test de wald y significativas segun el test F, el r cuadro disminuye.

Hide

```
cumplimientoSupuestos(modeleje1P3)
```

```
##      [,1]
## [1,] "Los residuos del modelo son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 0.161025671895042 ) es mayor que el nivel de significancia
(0.01), no se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto,
se puede considerar que los residuos del modelo siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 0.13525678913561 ) es mayor que el nivel de significancia est
ablecido (0.05), no se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.
Por lo tanto, se puede considerar que los errores del modelo tienen varianzas constantes (homocedastic
idad)."
## [6,] "----------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.582622472831813 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

Se cumplen los supuestos.

Se prueba la transformación con todas las variables por haber tenido un mejor r cuadro que dejando solo las variables significativas.

Hide

```
modele2je1P3 <- lm((creatinina)^(best_box_cox) ~ edad + LYVE1 + REG1B + TFF1 + sexo +
    diagnosis + estadio, data = strat_data)

summary(modele2je1P3)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox) ~ edad + LYVE1 + REG1B +
##     TFF1 + sexo + diagnosis + estadio, data = strat_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39782 -0.11758 -0.00054  0.11515  0.42403
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.476e-01  7.716e-02  12.281  < 2e-16 ***
## edad            -2.800e-03  8.958e-04  -3.126  0.00195 **
## LYVE1            1.749e-02  3.709e-03   4.716 3.75e-06 ***
## REG1B           -1.162e-05  6.676e-05  -0.174  0.86192
## TFF1             3.589e-05  1.203e-05   2.982  0.00310 **
## sexoM            4.628e-02  2.009e-02   2.304  0.02193 *
## diagnosisnormal  6.902e-02  5.032e-02   1.371  0.17129
## estadioII       -3.653e-02  5.123e-02  -0.713  0.47641
## estadioIII       2.563e-02  5.277e-02   0.486  0.62755
## estadioIV        1.526e-02  6.245e-02   0.244  0.80708
## estadioNO               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1685 on 290 degrees of freedom
## Multiple R-squared:  0.2294, Adjusted R-squared:  0.2054
## F-statistic:  9.59 on 9 and 290 DF,  p-value: 8.196e-13
```

Hide

```
cumplimientoSupuestos(modele2je1P3)
```

```
##      [,1]
## [1,] "Los residuos del modelo son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 0.421718332430314 ) es mayor que el nivel de significancia
(0.01), no se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto,
se puede considerar que los residuos del modelo siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 0.160896327029234 ) es mayor que el nivel de significancia es
tablecido (0.05), no se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticida
d. Por lo tanto, se puede considerar que los errores del modelo tienen varianzas constantes (homocedas
ticidad)."
## [6,] "----------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.761389739187196 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

Tambien cumple los supuestos con un mejor r cuadro

Resumen de los modelos utilizados

Hide

```
resumenCumplimiento(4, modelEje1P1,model2Eje1P1,modeleje1P3,modele2je1P3)
```

```
##      [,1]           [,2]                [,3]
## [1,] "Modelos"      "Normalidad"        "Homocedasticidad"
## [2,] "modelEje1P1"  "No hay normalidad" "No hay homocedasticidad"
## [3,] "model2Eje1P1" "No hay normalidad" "No hay homocedasticidad"
## [4,] "modeleje1P3"  "Hay normalidad"    "Hay homocedasticidad"
## [5,] "modele2je1P3" "Hay normalidad"    "Hay homocedasticidad"
##      [,4]                [,5]
## [1,] "Independencia"     "Cumplimiento"
## [2,] "Hay independencia" "No"
## [3,] "Hay independencia" "No"
## [4,] "Hay independencia" "Si"
## [5,] "Hay independencia" "Si"
```

Los dos modelos con transformación cumplen los supuestos, por tal razon son los que se siguen utilizando.

Puntos influyentes

Hide

```
summary(influence.measures(model = modele2je1P3))
```

```
## Potentially influential observations of
##   lm(formula = (creatinina)^(best_box_cox) ~ edad + LYVE1 + REG1B +      TFF1 + sexo + diagnosis +
estadio, data = strat_data) :
##
##     dfb.1_ dfb.edad dfb.LYVE dfb.REG1 dfb.TFF1 dfb.sexM dfb.dgns dfb.esII
## 52  -0.14   0.15     0.02    -0.01     0.01     0.13    -0.01     0.00
## 108 -0.17   0.20    -0.04     0.00     0.03     0.13    -0.01     0.01
## 148 -0.02   0.01     0.10    -0.07     0.03     0.01     0.01    -0.01
## 149 -0.14   0.24     0.58    -0.42    -0.66    -0.16     0.00    -0.09
## 150 -0.03   0.01    -0.01    -0.02     0.01     0.00     0.04     0.04
## 171  0.46  -0.15    -0.01    -0.06     0.03    -0.14    -0.53    -0.47
## 180  0.04  -0.02    -0.14     0.09    -0.03    -0.04    -0.02     0.01
## 185  0.00   0.00     0.00     0.00     0.01    -0.01     0.00     0.00
## 190  0.09   0.03    -0.01    -0.01     0.00    -0.04    -0.16    -0.16
## 194  0.04  -0.03     0.00    -0.09     0.02    -0.03    -0.02     0.00
## 197  0.00   0.01    -0.03     0.00     0.00    -0.02     0.00     0.00
## 199 -0.04   0.01     0.10     0.59    -0.41    -0.16     0.05     0.06
## 208  0.00   0.00     0.00     0.00     0.00     0.00     0.00     0.00
## 215  0.00   0.00    -0.01     0.01     0.01    -0.01     0.00     0.00
## 219  0.00   0.01    -0.02    -0.09     0.02     0.04    -0.01     0.01
## 223  0.00   0.00     0.00     0.03    -0.02     0.00     0.00     0.00
## 233  0.05  -0.06    -0.07    -0.09     0.07     0.11    -0.02     0.01
## 241 -0.08   0.16    -0.09    -0.05     0.01    -0.15    -0.01     0.00
## 245 -0.01   0.00    -0.16    -0.07     0.52    -0.02     0.02     0.00
## 246  0.07   0.03    -0.03    -0.03     0.03     0.03    -0.13    -0.13
## 252 -0.09   0.01     0.03     0.02    -0.03     0.03     0.13     0.12
## 256 -0.08  -0.02    -0.10    -0.05     0.06     0.06     0.15     0.17
## 259 -0.13  -0.01    -0.15     0.05    -0.01    -0.04     0.22     0.26
## 265  0.19  -0.13    -0.30    -0.07     0.08    -0.18    -0.10     0.02
## 268  0.01   0.00     0.02    -0.01    -0.02    -0.02     0.00     0.00
## 270  0.13  -0.16     0.05    -0.08     0.09    -0.08    -0.03    -0.01
## 271 -0.05   0.01     0.01    -0.01     0.01    -0.01     0.07     0.06
## 287 -0.10   0.03     0.00     0.02     0.00    -0.02     0.14     0.13
## 296  0.02  -0.02     0.02    -0.06     0.02    -0.01    -0.01     0.00
##     dfb.eIII dfb.esIV dffit   cov.r   cook.d hat
## 52   0.00    -0.01    -0.28    0.86_*  0.01   0.01
## 108  0.01    -0.01    -0.30    0.89_*  0.01   0.02
## 148  0.00    -0.02     0.13    1.27_*  0.00   0.19_*
## 149  0.00     0.01    -1.23_*  1.00    0.15   0.18_*
## 150  0.04     0.03    -0.05    1.15_*  0.00   0.10_*
## 171 -0.46    -0.38     0.56_*  1.02    0.03   0.09
## 180  0.02    -0.11    -0.27    1.11_*  0.01   0.09
## 185  0.00    -0.02    -0.03    1.11_*  0.00   0.06
## 190 -0.15    -0.13     0.18    1.12_*  0.00   0.09
## 194  0.00    -0.11    -0.22    1.10_*  0.00   0.08
## 197  0.01     0.05     0.08    1.12_*  0.00   0.08
## 199 -0.01     0.00     0.66_*  1.19_*  0.04   0.19_*
## 208  0.00     0.00     0.00    1.12_*  0.00   0.08
## 215  0.00    -0.03    -0.04    1.11_*  0.00   0.06
## 219 -0.02     0.01    -0.13    1.12_*  0.00   0.08
## 223  0.01     0.00     0.03    1.14_*  0.00   0.09
## 233  0.14     0.01     0.36    0.88_*  0.01   0.02
## 241  0.17     0.01     0.42    0.88_*  0.02   0.03
## 245 -0.02    -0.05     0.59_*  2.04_*  0.04   0.50_*
## 246 -0.12    -0.11     0.16    1.13_*  0.00   0.09
## 252  0.11     0.09    -0.14    1.13_*  0.00   0.09
## 256  0.17     0.15    -0.22    1.16_*  0.00   0.12_*
## 259  0.26     0.23    -0.32    1.13_*  0.01   0.11_*
```

```
## 265  0.04     0.47     0.79_*  0.89_*  0.06   0.08
## 268  0.00     0.04     0.07    1.11_*  0.00   0.07
## 270 -0.02     0.15     0.33    1.10_*  0.01   0.10
## 271  0.06     0.05    -0.07    1.13_*  0.00   0.09
## 287  0.13     0.10    -0.15    1.12_*  0.00   0.08
## 296  0.00    -0.04    -0.09    1.20_*  0.00   0.14_*
```

Existen varios puntos, se validan bajo ciertos criterios.

Hide

```
which(dfbetas(modele2je1P3)[,2]>1)
```

```
## named integer(0)
```

No existen puntos en este criterio.

Hide

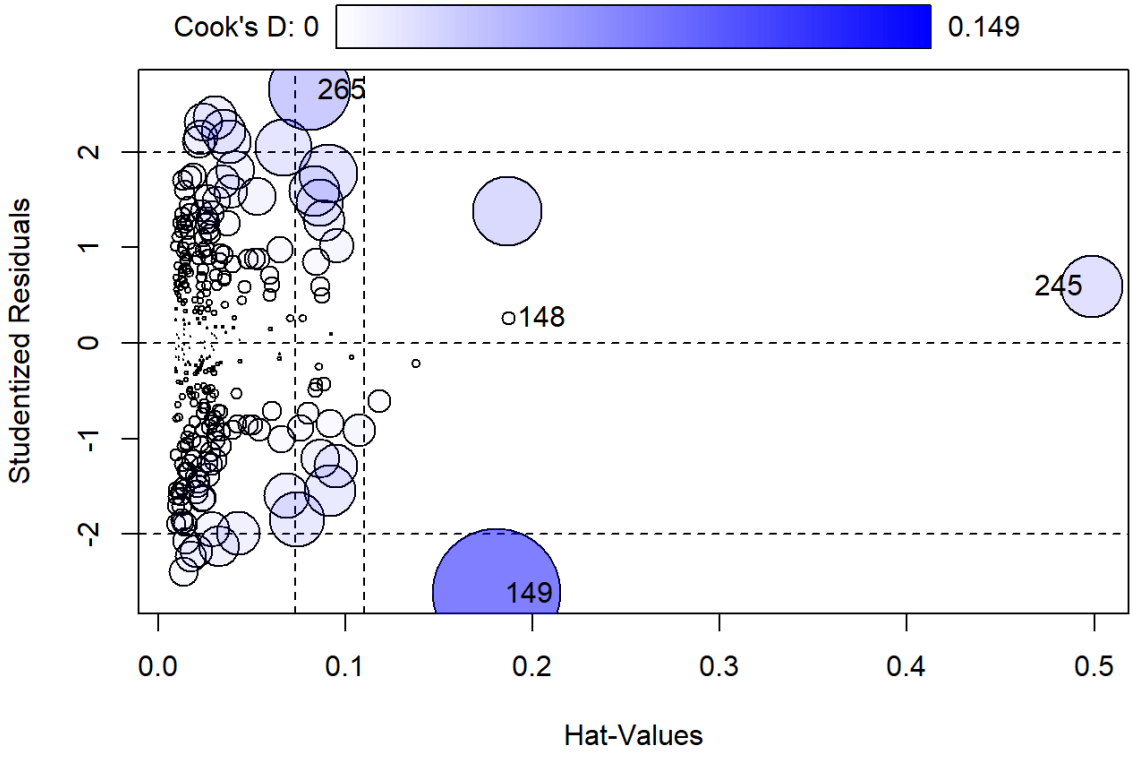
```
n<-length(strat_data$creatinina)
p<-length(modele2je1P3$coefficients)
which(dffits(modele2je1P3)>2 * sqrt(p / n))
```

```
## 171 199 214 225 234 241 245 247 264 265 280
## 171 199 214 225 234 241 245 247 264 265 280
```

Bajo este criterio existen 11 puntos incluyentes.

Hide

```
influencePlot(model = modele2je1P3)
```



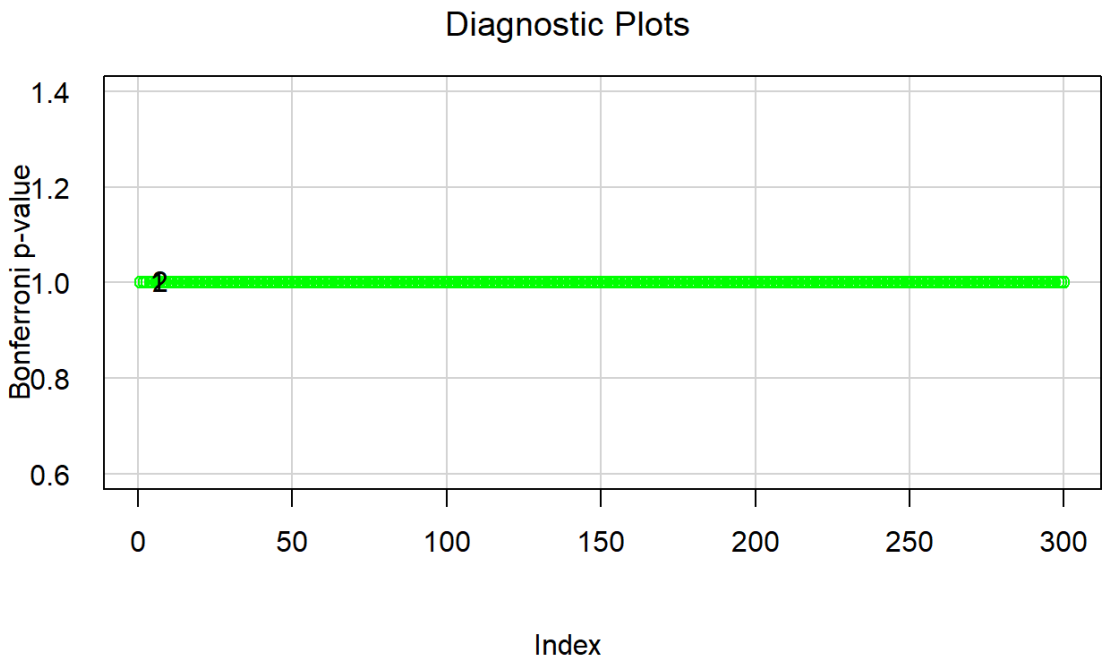
	StudRes <dbl>	Hat <dbl>	CookD <dbl>
148	0.2668966	0.18726260	0.001646569
149	-2.6202880	0.18077303	0.148501398
245	0.5951851	0.49867952	0.035316557
265	2.6521860	0.08078147	0.060556000

4 rows

Se observan algunos puntos del criterio anterior.

Hide

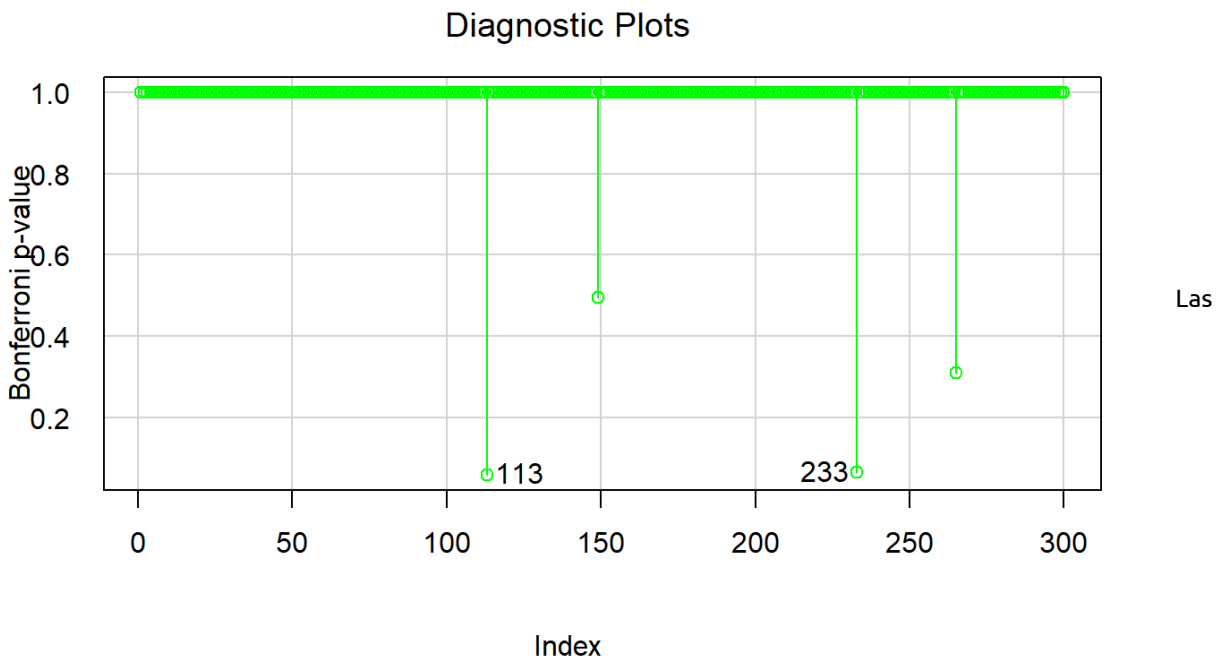
```
influenceIndexPlot(modele2je1P3, vars='Bonf', las=1,col='green')
```



Bajo el modelo transformado la prueba de bonferroni no detecta puntos influyentes, se valida sobre uno de los modelos lineales sin transformar.

Hide

```
influenceIndexPlot(modelEje1P1, vars='Bonf', las=1,col='green')
```



observaciones de uno de los modelos sin transformar son diferentes a las detectadas con los otros criterios, se entiende que la transformación disminuiria la influencia de esos puntos y por ello permite que se den los cumplimientos de los supuestos.

Hide

```
outlierTest(modele2je1P3)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 265 2.652186           0.008439           NA
```

Para el modelo transformado se observa que si se detecta un punto como outlier, este tambien es detectado en los anteriores criterior como influyente.

Hide

```
outlierTest(modelEje1P1)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 113 3.769781          0.0001981     0.059431
```

para el caso del modelo lineal sin tranformación se confirma que el punto 113 es inlfuyente y outlier.

4. Sin considerar la variable estadío, ajuste un modelo multivariado robusto para explicar el valor de la creatinina y estime el error absoluto medio cometido.

Hide

```
ajusterobEje1P4 <- lmrob(creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo + diagnosis, data = strat_dat
a)
summary(ajusterobEje1P4)
```

```
##
## Call:
## lmrob(formula = creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo + diagnosis,
##     data = strat_data)
##  \--> method = "MM"
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11718 -0.29008 -0.05214  0.32871  2.43182
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.950e-01  2.409e-01   2.470  0.01409 *
## edad            -5.864e-03  3.370e-03  -1.740  0.08290 .
## LYVE1            5.596e-02  1.981e-02   2.825  0.00505 **
## REG1B            2.345e-04  3.133e-04   0.749  0.45475
## TFF1             6.658e-05  9.216e-05   0.722  0.47064
## sexoM            1.746e-01  5.995e-02   2.912  0.00386 **
## diagnosisnormal  3.674e-01  8.195e-02   4.484 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 0.4505
## Multiple R-squared:  0.2126, Adjusted R-squared:  0.1964
## Convergence in 21 IRWLS iterations
##
## Robustness weights:
##  3 observations c(113,233,245) are outliers with |weight| = 0 ( < 0.00033);
##  16 weights are ~= 1. The remaining 281 ones are summarized as
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.001312 0.874800 0.952700 0.880400 0.984000 0.998800
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         3.333e-04
##             eps.x warn.limit.reject warn.limit.meanrw
##         2.427e-08         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

Hide

```
set.seed(1082884412) # fijamos una semilla para reproducibilidad
train <- strat_data %>% sample_frac(0.8)#separamos la base
test <- strat_data %>% setdiff(train)
ytest <- test$creatinina

modeloTRAINEje1P4 <- rlm(creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo + diagnosis, data=train, psi=
psi.huber)
predictTEST <- predict(modeloTRAINEje1P4, test)
RMSEEje1P4 <- caret::RMSE(predictTEST, ytest)
RMSEEje1P4
```

```
## [1] 0.676536
```

RMSE igual a 0.6765536.

Se calcula el RMSE para el mejor modelo lineal

Hide

```
modeloTRAIN2Eje1P4 <- lm((creatinina)^(best_box_cox) ~ edad + LYVE1 + REG1B + TFF1 + sexo +
    diagnosis, data=train)
predictTEST <- predict(modeloTRAIN2Eje1P4, test)
RMSE2Eje1P4 <- caret::RMSE(predictTEST, ytest)
RMSE2Eje1P4
```

```
## [1] 0.6741386
```

El RMSE del modelo lineal ajustado es levente menor que el del robusto.

5. Sin considerar la variable estadío, utilice un método de selección de variables para proponer un nuevo modelo multivariado que explique el valor de la creatinina. Estudie el cumplimiento de los supuestos y haga una transformación en caso de ser necesario. Analice los coeficientes del modelo final.

Hide

```
mejores_modelos <- regsubsets(creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo +
    diagnosis, data = strat_data, nvmax = 6)
summary(mejores_modelos)
```

```
## Subset selection object
## Call: regsubsets.formula(creatinina ~ edad + LYVE1 + REG1B + TFF1 +
##     sexo + diagnosis, data = strat_data, nvmax = 6)
## 6 Variables  (and intercept)
##                 Forced in Forced out
## edad                FALSE      FALSE
## LYVE1               FALSE      FALSE
## REG1B               FALSE      FALSE
## TFF1                FALSE      FALSE
## sexoM               FALSE      FALSE
## diagnosisnormal     FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          edad LYVE1 REG1B TFF1 sexoM diagnosisnormal
## 1  ( 1 ) " "  " "   " "   "*"  " "   " "
## 2  ( 1 ) "*"  " "   " "   "*"  " "   " "
## 3  ( 1 ) "*"  "*"   " "   "*"  " "   " "
## 4  ( 1 ) "*"  "*"   " "   "*"  " "   "*"
## 5  ( 1 ) "*"  "*"   " "   "*"  "*"   "*"
## 6  ( 1 ) "*"  "*"   "*"   "*"  "*"   "*"
```

El resultado muestra las variables que fueron incluidas en cada tamaño de subconjunto. Por ejemplo, el primer subconjunto incluye solo la variable “TFF1”, el segundo subconjunto incluye “edad” y “TFF1”, el tercer subconjunto incluye “edad”, “LYVE1” y “TFF1”, y así sucesivamente hasta el sexto subconjunto, que incluye todas las variables.

Hide

```
summary(mejores_modelos)$adjr2
```

```
## [1] 0.1377079 0.1723653 0.2136463 0.2296397 0.2387975 0.2369408
```

Hide

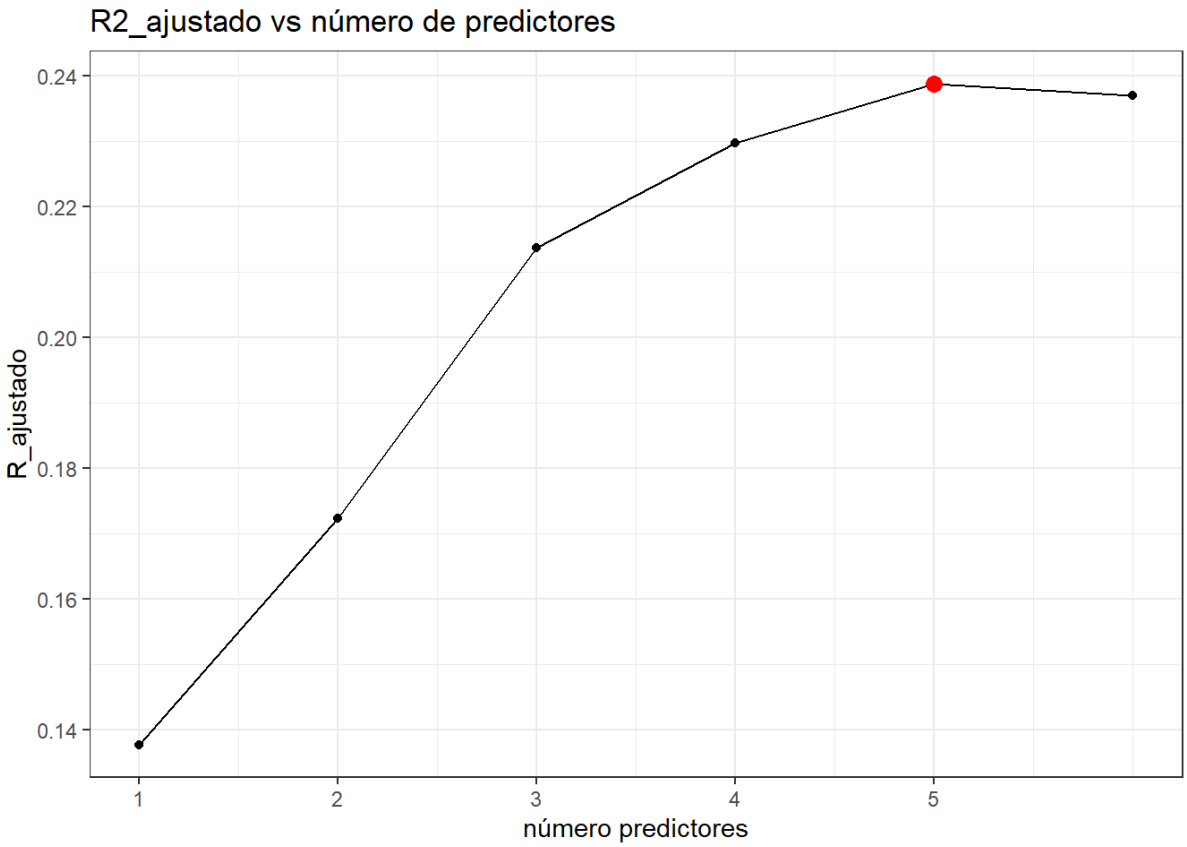
```
# se identifica qué modelo tiene el valor máximo de R ajustado
which.max(summary(mejores_modelos)$adjr2)
```

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

El mejor r cuadro es el modelo 5, r cuadro de 0.2387975.

Hide

```
library(ggplot2)
p <- ggplot(data = data.frame(n_predictores = 1:6, R_ajustado = summary(mejores_modelos)$adjr2), aes(x
= n_predictores, y = R_ajustado)) +
  geom_line() +
  geom_point()
# Se identifica en rojo el máximo
p <- p + geom_point(aes(x=n_predictores[which.max(summary(mejores_modelos)$adjr2)], y=R_ajustado[whic
h.max(summary(mejores_modelos)$adjr2)]), colour = "red", size = 3)
p <- p + scale_x_continuous(breaks = c(0:5)) + theme_bw() +
  labs(title = "R2_ajustado vs número de predictores", x = "número predictores")
p
```



La cantidad de predictores que maximiza el r cuadrado es con 5 predictores.

Hide

```
# coeficientes del modelo No. 5
coef(object = mejores_modelos, id = 5)
```

```
##     (Intercept)            edad           LYVE1            TFF1           sexoM
##    1.0773844348   -0.0115877675    0.0569890668    0.0001686631    0.1455385245
## diagnosisnormal
##    0.2523327868
```

Se confirmar con otros metodos de selección de variables

Hide

```
mejores_modelos_backward <- regsubsets(creatinina ~ edad + LYVE1 + REG1B + TFF1 +
    sexo + diagnosis, data = strat_data, nvmax = 6, method = "backward")
# se identifica el valor máximo de R ajustado
which.max(summary(mejores_modelos_backward)$adjr2)
```

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

Se sigue obteniendo el mismo modelo.

Hide

```
coef(object = mejores_modelos_backward, 5)
```

```
##     (Intercept)            edad           LYVE1            TFF1           sexoM
##    1.0773844348   -0.0115877675    0.0569890668    0.0001686631    0.1455385245
## diagnosisnormal
##    0.2523327868
```

igual.

Hide

```
lm.fit1 <- lm(creatinina ~ edad + LYVE1 + REG1B + TFF1 +
    sexo + diagnosis, data = strat_data)
k <- ols_step_all_possible(lm.fit1)

# AIC: Akaike Information Criteria
# SBIC: Sawa's Bayesian Information Criteria
# SBC: Schwarz Bayesian Criteria
# MSEP: Estimated error of prediction, assuming multivariate normality
# FPE: Final Prediction Error
# HSP: Hocking's Sp
# APC: Amemiya Prediction Criteria

k
```

	mindex	n	predictors	rsquare	adjr
	<int>	<int>	<chr>	<dbl>	<dbl>
4	1	1	TFF1	0.140591852	0.1377079324
2	2	1	LYVE1	0.108374444	0.1053824116
3	3	1	REG1B	0.061743567	0.0585950556
5	4	1	sexo	0.021998503	0.0187166189
1	5	1	edad	0.013014142	0.0097021092
6	6	1	diagnosis	0.003116231	-0.0002290162
9	7	2	edad TFF1	0.177901300	0.1723652822
7	8	2	edad LYVE1	0.169992804	0.1644035295
13	9	2	LYVE1 TFF1	0.160392139	0.1547382136
19	10	2	TFF1 sexo	0.157144529	0.1514687342

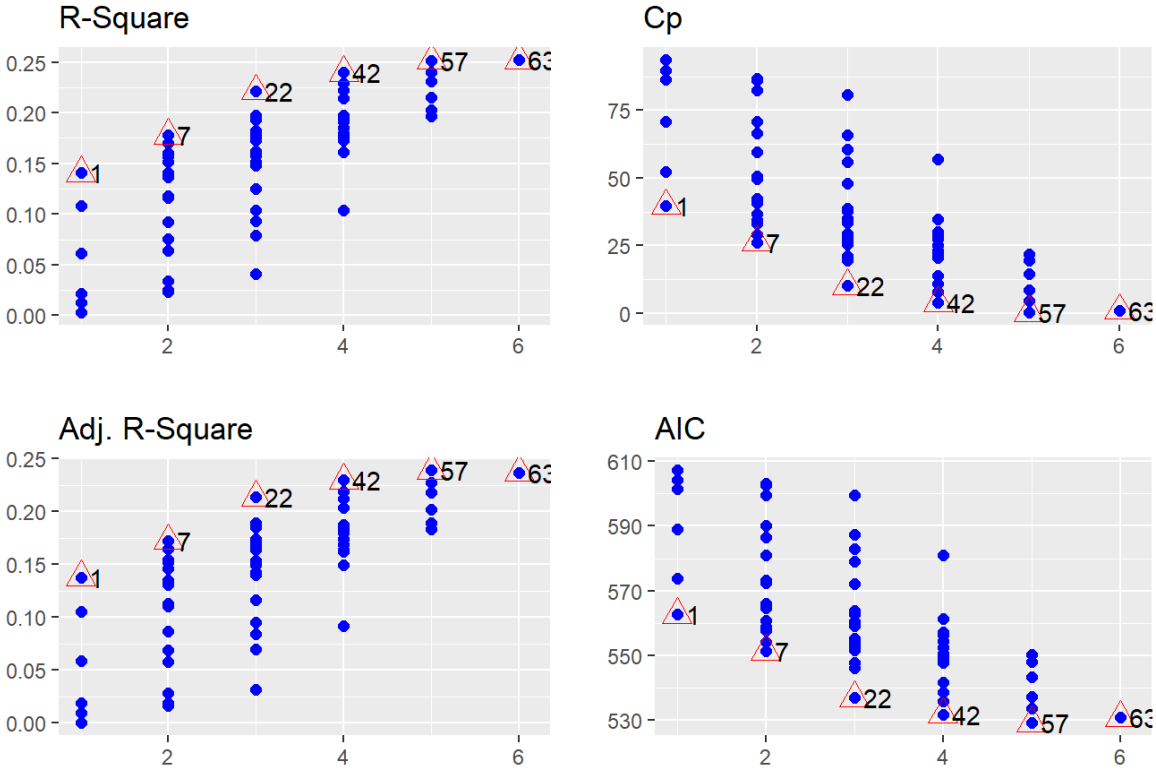
1-10 of 63 rows | 1-6 of 15 columns Previous 1 2 3 4 5 6 7 Next

Hide

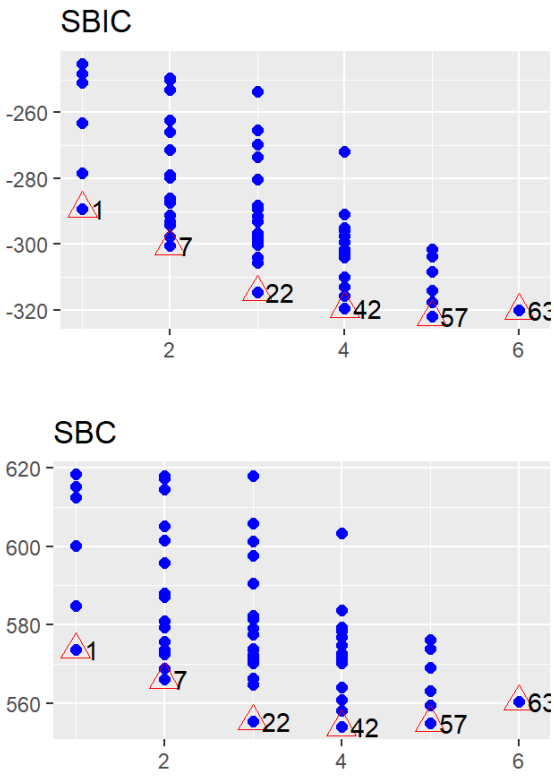
```
plot(k)# el eje horizontal representa la cantidad de variables utilizadas en cada modelo.
```

```
## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## ℹ Please use "none" instead.
## ℹ The deprecated feature was likely used in the olsrr package.
##   Please report the issue at <]8;;https://github.com/rsquaredacademy/olsrr/issueshttps://github.c
om/rsquaredacademy/olsrr/issues]8;;>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

page 1 of 2



page 2 of 2



Se confirma que las cinco variables utilizaas son las mejores para el modelo.

Hide

```
k_best <- ols_step_best_subset(lm.fit1)

# AIC: Akaike Information Criteria
# SBIC: Sawa's Bayesian Information Criteria
# SBC: Schwarz Bayesian Criteria
# MSEP: Estimated error of prediction, assuming multivariate normality
# FPE: Final Prediction Error
# HSP: Hocking's Sp
# APC: Amemiya Prediction Criteria

k_best
```

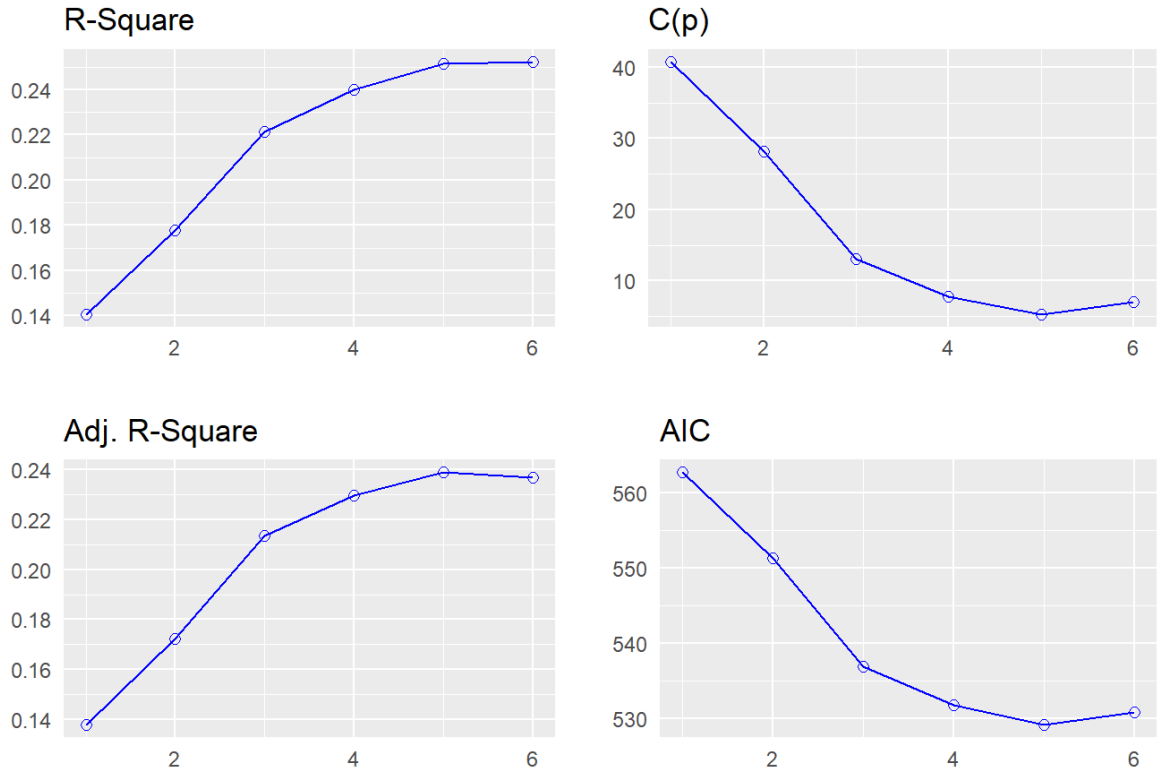
	mindex	n	predictors	rsquare	adjr	predrsq
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>
4	1	1	TFF1	0.1405919	0.1377079	0.1200567
9	2	2	edad TFF1	0.1779013	0.1723653	0.1515862
23	3	3	edad LYVE1 TFF1	0.2215361	0.2136463	0.1703968
46	4	4	edad LYVE1 TFF1 diagnosis	0.2399455	0.2296397	0.1880988
60	5	5	edad LYVE1 TFF1 sexo diagnosis	0.2515267	0.2387975	0.1920177
63	6	6	edad LYVE1 REG1B TFF1 sexo diagnosis	0.2522531	0.2369408	0.1824602

6 rows | 1-7 of 15 columns

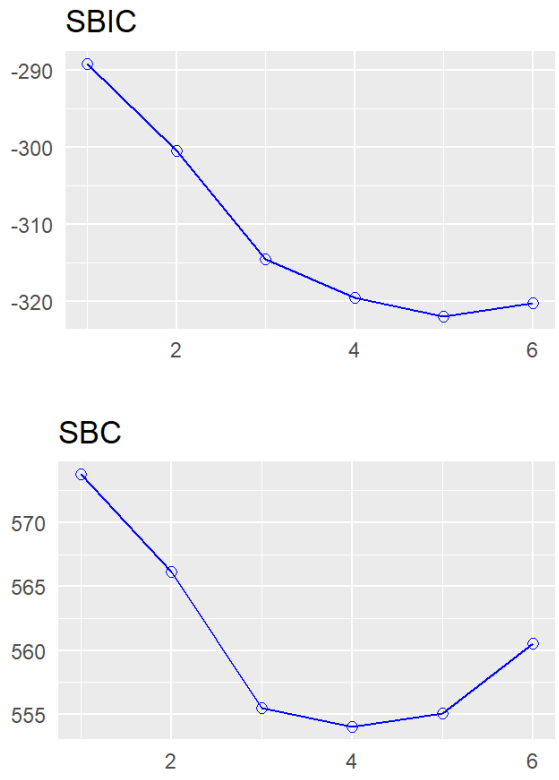
Hide

```
plot(k_best)# el eje horizontal representa la cantidad de variables utilizadas en cada modelo.
```

page 1 of 2



page 2 of 2



Se observa que por lo metodos indirectos se confirma que seleccionando 5 variables se maximiza el modelo, aunque en algunos muestra que puede ser algo parecido con 4 variables, habria que realizar mas pruebas, por lo pronto se mantienen las 5 variables.

En tal sentido, los modelos tranformados con 5 variables que cumplen los supuestos en los puntos anteriores estarian construidos con las mejores variables del conjunto de datos.

6. Estime los errores de predicción de los 4 modelos previos y compárelos. Cuál elegiría?

El RMSE se desarrollo en puntos anteriores entre el mejor modelo transformado (0.6741386) y el robusto (0.676536), se observo que son muy parecidos, por los resultados se puede concluir que se podria elegir cualquiera de los dos, la diferencia minima entre ambos es 0,0023974. Por ser menor se eligiria el modelo transformado con 5 variables.

7. Le parece adecuado un modelo GAMLSS en este caso? Justifique.

Hide

```
mod_OLS <- gamlss( formula = creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo +
    diagnosis, family = NO, data = strat_data, trace = FALSE)
summary(mod_OLS)
```

```
## ******************************************************************
## Family:  c("NO", "Normal")
##
## Call:  gamlss(formula = creatinina ~ edad + LYVE1 + REG1B +
##     TFF1 + sexo + diagnosis, family = NO, data = strat_data,
##     trace = FALSE)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.076e+00  2.103e-01   5.115 5.68e-07 ***
## edad            -1.150e-02  2.997e-03  -3.837 0.000153 ***
## LYVE1            5.804e-02  1.230e-02   4.718 3.70e-06 ***
## REG1B           -1.218e-04  2.257e-04  -0.540 0.589716
## TFF1             1.804e-04  4.055e-05   4.448 1.23e-05 ***
## sexoM            1.491e-01  6.784e-02   2.198 0.028751 *
## diagnosisnormal  2.495e-01  8.505e-02   2.933 0.003623 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.56079    0.04082  -13.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  300
## Degrees of Freedom for the fit:  8
##       Residual Deg. of Freedom:  292
##                       at cycle:  2
##
## Global Deviance:     514.887
##             AIC:     530.887
##             SBC:     560.5173
## ******************************************************************
```

Para el parámetro de ubicación (Mu):

La función de enlace utilizada es la identidad, lo que implica que los predictores se asocian linealmente con el valor esperado de la variable de respuesta “creatinina”. Los coeficientes estimados para las variables predictoras “edad”, “LYVE1”, “REG1B”, “TFF1”, “sexo” y “diagnosis” indican la dirección y la magnitud de la relación con el valor esperado de la creatinina. Los valores p asociados a cada coeficiente indican la significancia estadística de cada variable en relación con la creatinina. Un valor p menor que el nivel de significancia elegido (por ejemplo, 0.05) indica que la variable es estadísticamente significativa. Para el parámetro de escala (Sigma):

La función de enlace utilizada es log, lo que implica que el predictor correspondiente al parámetro de escala se asocia con el logaritmo de la desviación estándar de la variable de respuesta “creatinina”. El coeficiente estimado para el intercepto indica el logaritmo de la desviación estándar de la creatinina.

Hide

```
cumplimientoSupuestos(mod_OLS)
```

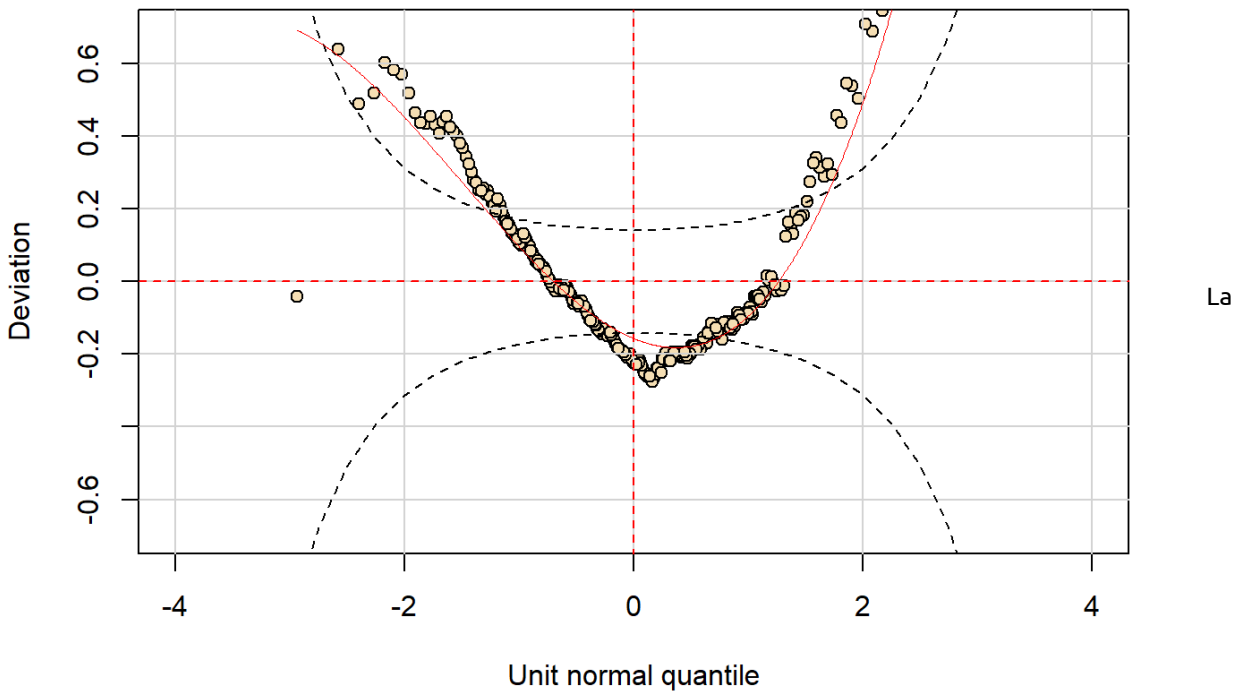
```
##      [,1]
## [1,] "Los residuos del modelo no son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 1.46512932034927e-09 ) es menor que el nivel de significancia
(0.01), se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto, se
puede considerar que los residuos del modelo no siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo no son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 3.82429094213541e-05 ) es menor que el nivel de significancia
establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.
Por lo tanto, se puede considerar que los errores del modelo no tienen varianzas constantes (heterosed
asticos)."
## [6,] "----------------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.483790431119234 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

No se cumplen los supuestos del modelo.

Hide

```
wp(mod_OLS)
```

```
## Warning in wp(mod_OLS): Some points are missed out
## increase the y limits using ylim.all
```



La

inspección visual del wormplot indica que este modelo no tiene los residuos dentro del rango de variación aceptable.

Hide

```
mod_GAMLSS <- gamlss(formula = creatinina ~ edad + LYVE1 + REG1B + TFF1 + sexo +
    diagnosis, sigma.formula = ~ edad + LYVE1 + REG1B + TFF1 + sexo +
    diagnosis, family = GA, data = strat_data, trace = FALSE )
summary(mod_GAMLSS)
```

```
## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:  gamlss(formula = creatinina ~ edad + LYVE1 + REG1B +
##     TFF1 + sexo + diagnosis, sigma.formula = ~edad +
##     LYVE1 + REG1B + TFF1 + sexo + diagnosis, family = GA,
##     data = strat_data, trace = FALSE)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.282e-02  2.391e-01   0.179  0.85801
## edad            -1.308e-02  3.314e-03  -3.945  0.00010 ***
## LYVE1            6.401e-02  1.037e-02   6.173 2.29e-09 ***
## REG1B            1.350e-04  2.072e-04   0.651  0.51530
## TFF1             1.039e-04  3.627e-05   2.865  0.00447 **
## sexoM            2.053e-01  7.501e-02   2.737  0.00659 **
## diagnosisnormal  3.154e-01  9.847e-02   3.203  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.190e-02  2.654e-01   0.158 0.874671
## edad            -1.597e-03  3.730e-03  -0.428 0.668864
## LYVE1           -5.886e-02  1.535e-02  -3.834 0.000155 ***
## REG1B           -5.427e-05  3.401e-04  -0.160 0.873328
## TFF1            -5.826e-06  6.918e-05  -0.084 0.932937
## sexoM            1.344e-02  8.142e-02   0.165 0.868966
## diagnosisnormal -3.302e-01  1.139e-01  -2.899 0.004036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  300
## Degrees of Freedom for the fit:  14
##       Residual Deg. of Freedom:  286
##                       at cycle:  6
##
## Global Deviance:     376.8024
##             AIC:     404.8024
##             SBC:     456.6554
## ******************************************************************
```

Se observan variables que no son significativas y se podrian retirar, antes se comprueba los cumplimientos del modelo.

Hide

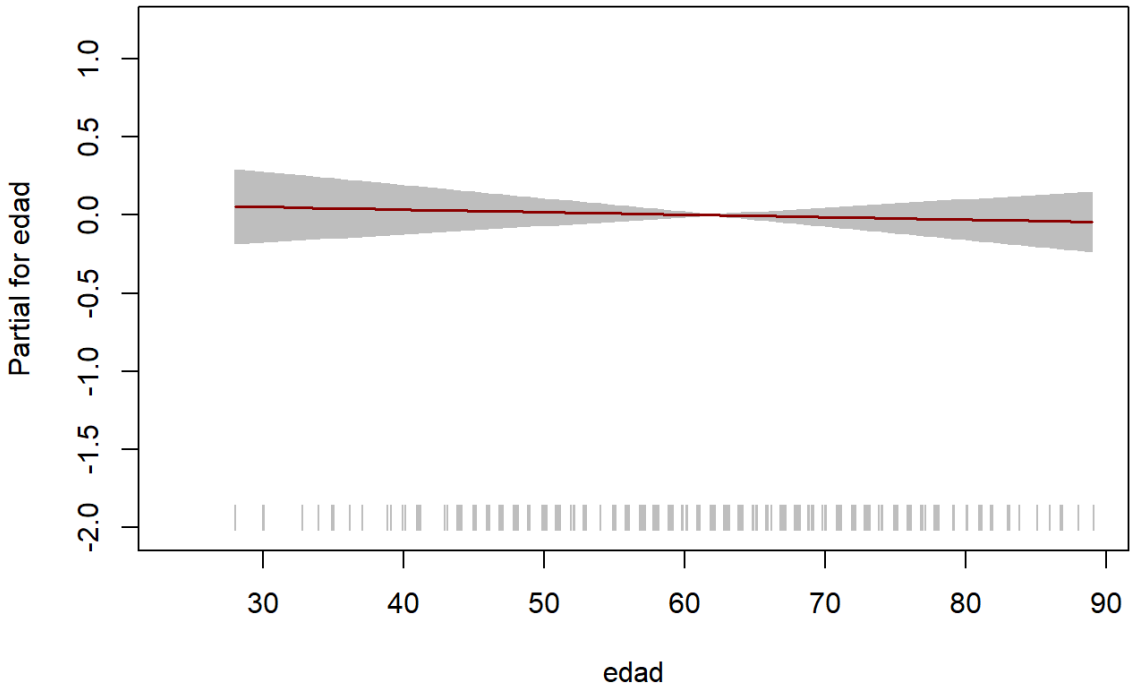
```
cumplimientoSupuestos(mod_GAMLSS)
```

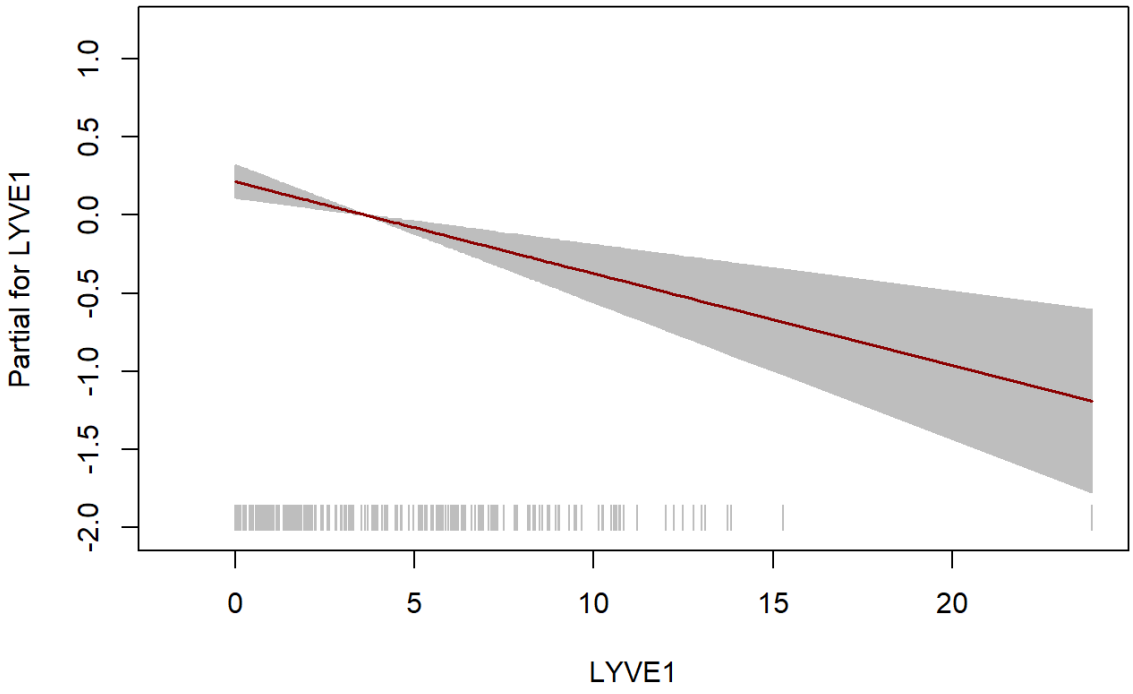
```
##      [,1]
## [1,] "Los residuos del modelo son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 0.502517872296299 ) es mayor que el nivel de significancia
(0.01), no se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto,
se puede considerar que los residuos del modelo siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo no son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 3.82429094213541e-05 ) es menor que el nivel de significancia
establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.
Por lo tanto, se puede considerar que los errores del modelo no tienen varianzas constantes (heterosed
asticos)."
## [6,] "----------------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.483790431119234 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

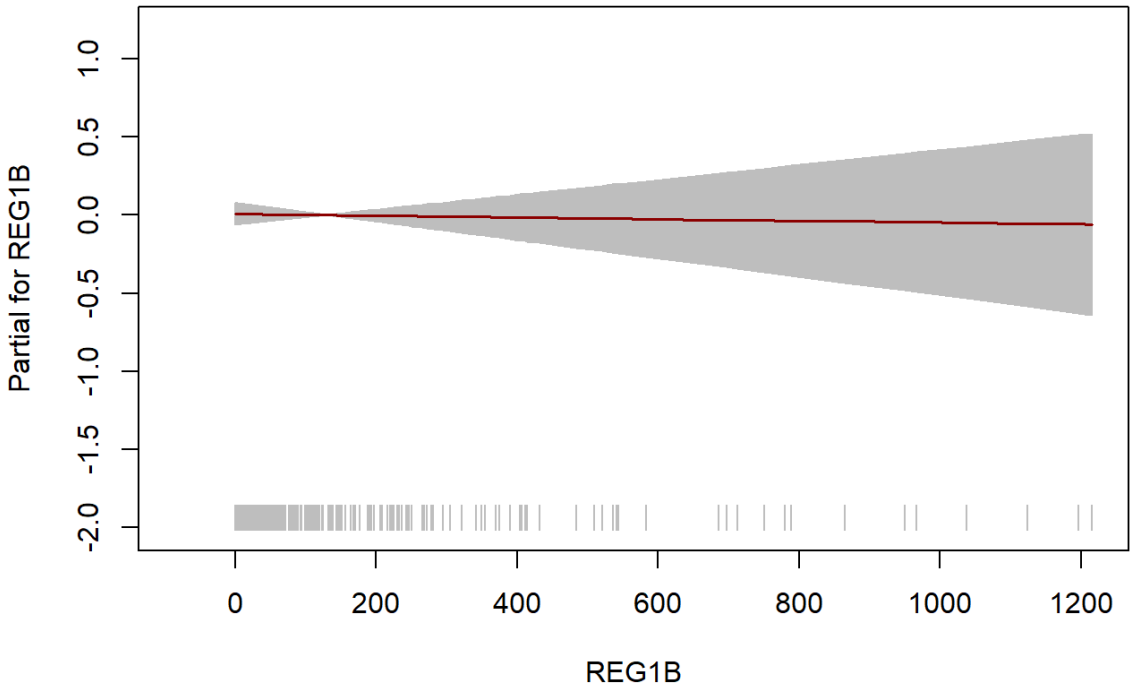
No se cumplen los supuestos.

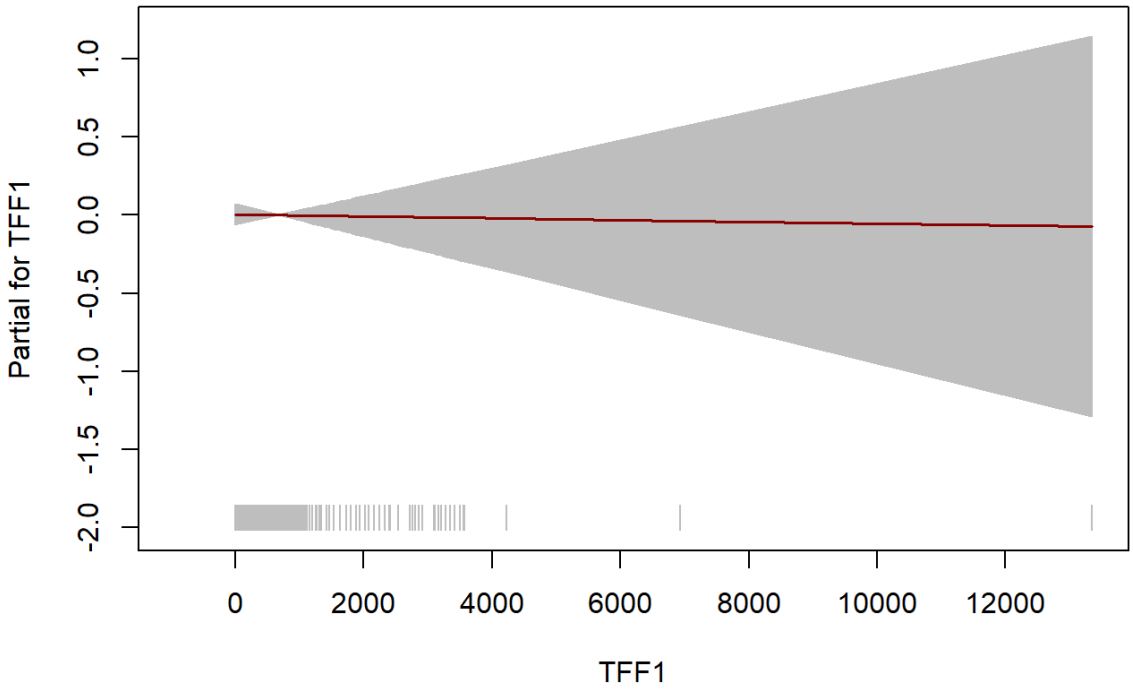
Hide

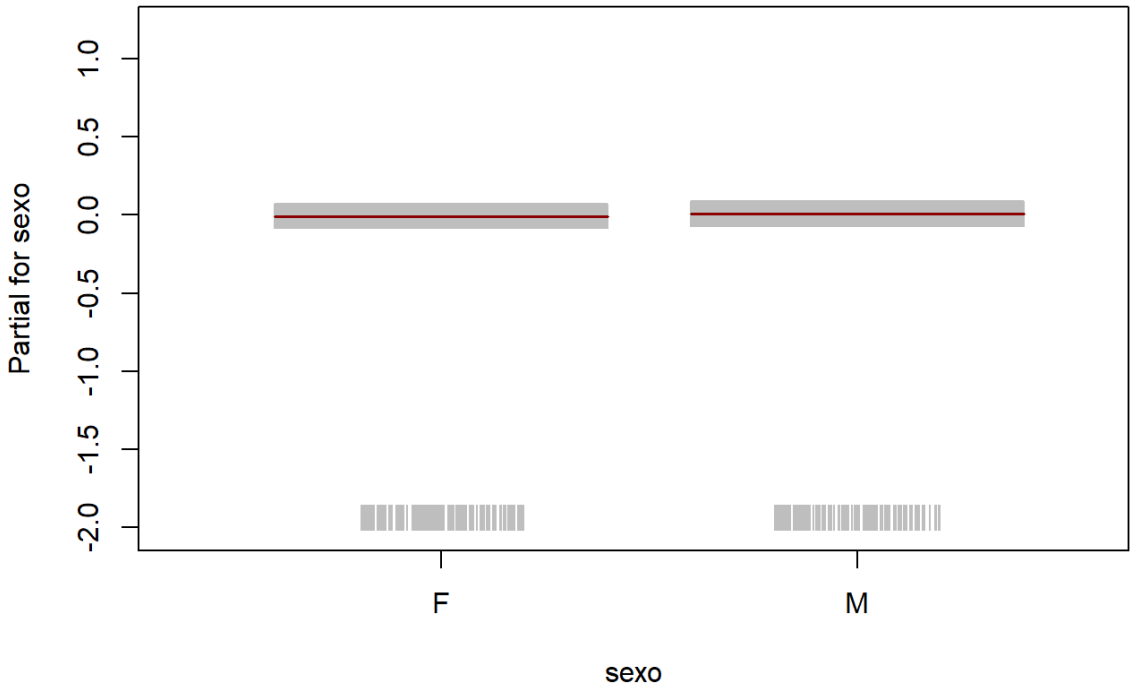
```
# Efecto individuales de los Predictores (GAMLSS)
term.plot(mod_GAMLSS, parameter = 'sigma',ask = FALSE, rug = TRUE)
```

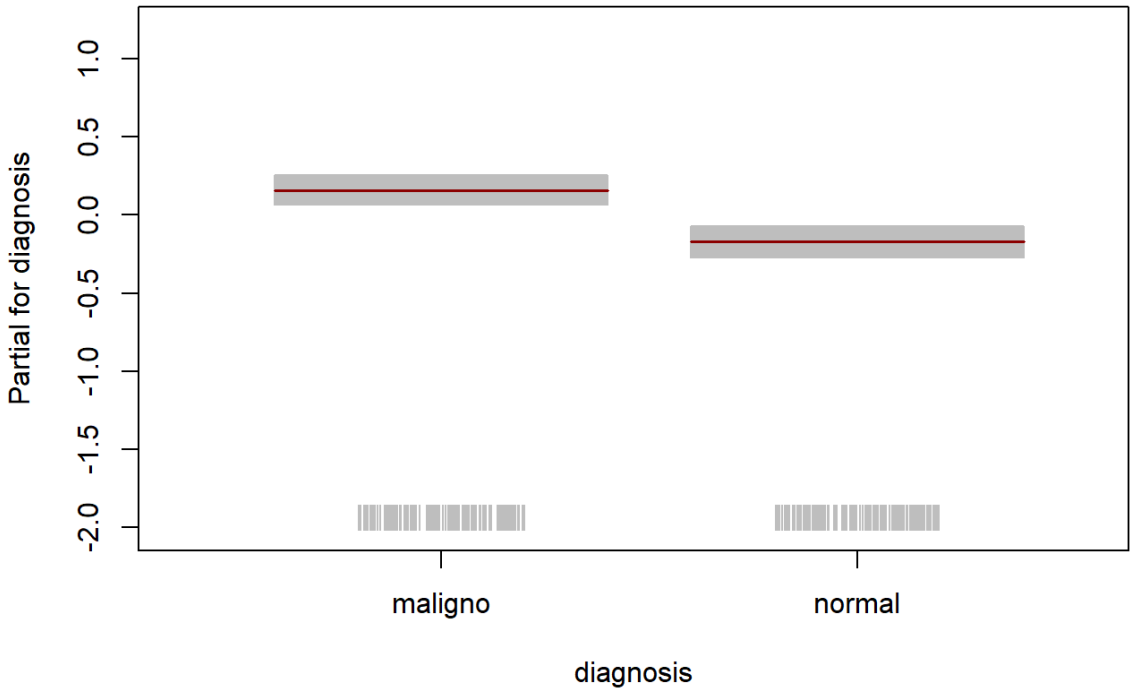

Partial for edad
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
30
40
50
60
70
80
90
edad


Partial for LYVE1
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
5
10
15
20
LYVE1


Partial for REG1B
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
200
400
600
800
1000
1200
REG1B


Partial for TFF1
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
2000
4000
6000
8000
10000
12000
TFF1





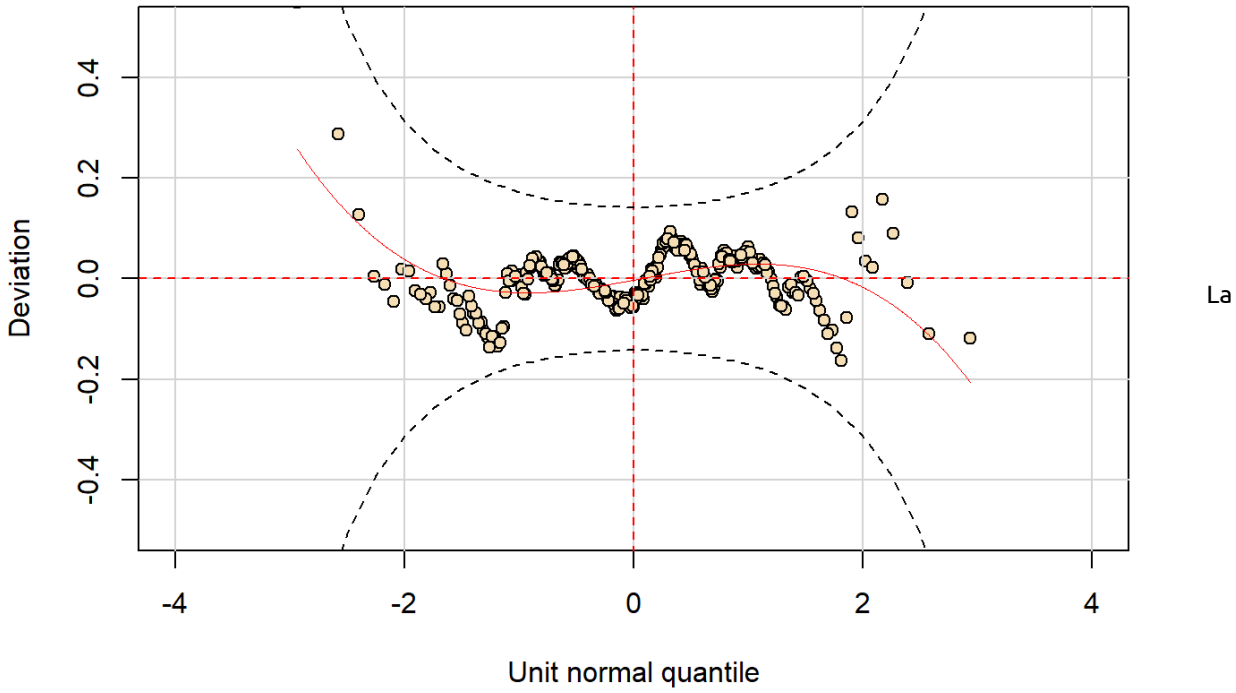
Se observa que la contribución de los predictores al modelo tiene un comportamiento lineal, esto nos dice que un modelo GAMLSS no es necesario.

Se intenta otro modelo retirando las variables que no son significativas.

Hide

```
# Worm plot de los residuos modelo 1
wp(mod_GAMLSS, ylim.all = 0.5)
```

```
## Warning in wp(mod_GAMLSS, ylim.all = 0.5): Some points are missed out
## increase the y limits using ylim.all
```



inspección visual del wormplot indica que este modelo tiene los residuos dentro del rango de variación aceptable.

Hide

```
mod_GAMLSS2 <- gamlss(formula = creatinina ~ edad + LYVE1 + TFF1 + sexo +
    diagnosis, sigma.formula = ~ LYVE1 +
    diagnosis, family = GA, data = strat_data, trace = FALSE )
summary(mod_GAMLSS2)
```

```
## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:  gamlss(formula = creatinina ~ edad + LYVE1 + TFF1 +
##     sexo + diagnosis, sigma.formula = ~LYVE1 + diagnosis,
##     family = GA, data = strat_data, trace = FALSE)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.208e-02  2.306e-01   0.269 0.787943
## edad            -1.309e-02  3.252e-03  -4.023 7.32e-05 ***
## LYVE1            6.369e-02  1.041e-02   6.120 3.01e-09 ***
## TFF1             1.168e-04  3.394e-05   3.442 0.000663 ***
## sexoM            2.085e-01  7.438e-02   2.804 0.005391 **
## diagnosisnormal  2.971e-01  9.336e-02   3.182 0.001620 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.07239    0.09757  -0.742  0.45871
## LYVE1           -0.05932    0.01365  -4.346 1.92e-05 ***
## diagnosisnormal -0.30062    0.10536  -2.853  0.00464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  300
## Degrees of Freedom for the fit:  9
##       Residual Deg. of Freedom:  291
##                       at cycle:  4
##
## Global Deviance:     377.6032
##             AIC:     395.6032
##             SBC:     428.9372
## ******************************************************************
```

Hide

```
cumplimientoSupuestos(mod_GAMLSS2 )
```

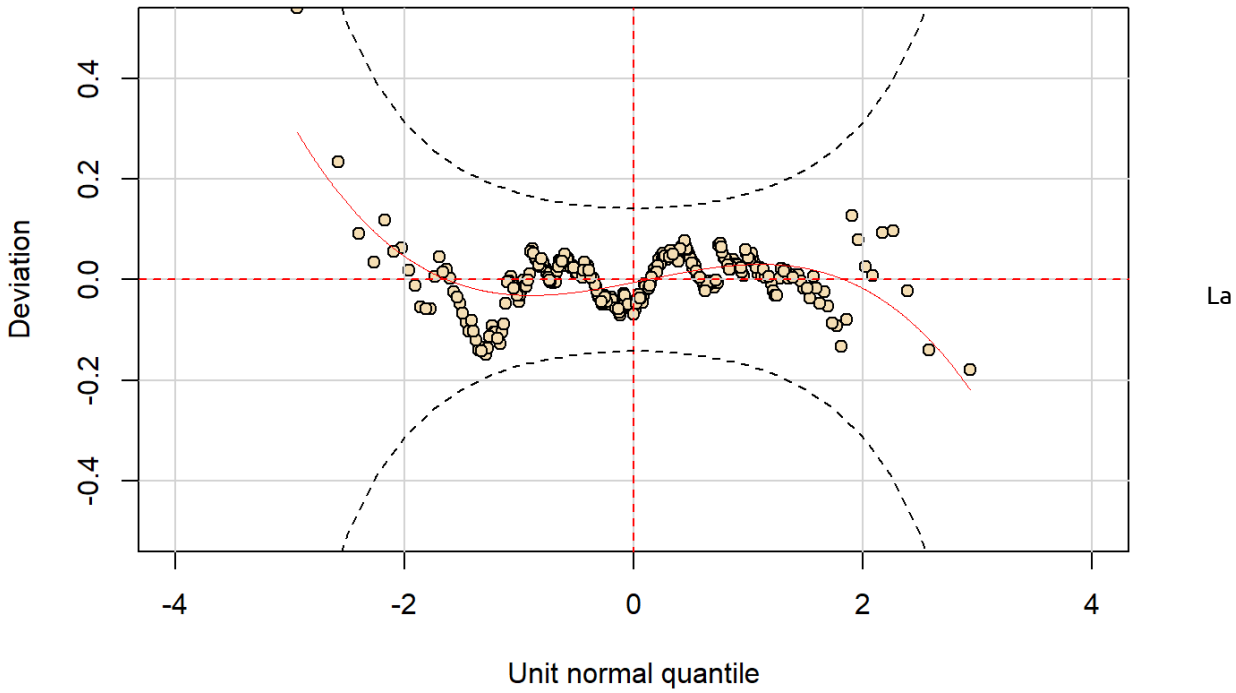
```
##      [,1]
## [1,] "Los residuos del modelo son normales basado en el test de Shapiro"
## [2,] "En este caso, como el valor p ( 0.533527029496373 ) es mayor que el nivel de significancia
(0.01), no se tiene suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto,
se puede considerar que los residuos del modelo siguen una distribución normal."
## [3,] "----------------------------------------------------------------------------------------
-------"
## [4,] "Los errores del modelo no son homocedastico basado en el test de Breusch-Pagan"
## [5,] "En este caso, como el valor p ( 1.36455061323056e-05 ) es menor que el nivel de significancia
establecido (0.05), se tiene suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.
Por lo tanto, se puede considerar que los errores del modelo no tienen varianzas constantes (heterosed
asticos)."
## [6,] "----------------------------------------------------------------------------------------
-------"
## [7,] "Los errores del modelo son independientes basado en el test de Durbin-Watson"
## [8,] "En este caso, como el valor p ( 0.519343605070592 ) es mayor que el nivel de significancia es
tablecido (0.05), se tiene suficiente evidencia para no rechazar la hipótesis nula de independencia de
los errores. Por lo tanto, se puede concluir que no existe autocorrelación en los errores del modelo."
```

No se cumple homocedasticidad. Para este tipo de modelos se flexibiliza el cumplimiento de todos los supuestos.

Hide

```
# Worm plot de los residuos modelo 1
wp(mod_GAMLSS2, ylim.all = 0.5)
```

```
## Warning in wp(mod_GAMLSS2, ylim.all = 0.5): Some points are missed out
## increase the y limits using ylim.all
```



inspección visual del wormplot indica que este modelo tiene los residuos dentro del rango de variación aceptable.

Comparación de modelos

Hide

```
#Comparamos los modelos ajustados:
GAIC(mod_OLS,mod_GAMLSS, mod_GAMLSS2)
```

	df <dbl>	AIC <dbl>
mod_GAMLSS2	9	395.6032
mod_GAMLSS	14	404.8024
mod_OLS	8	530.8870

3 rows

Se observa que de los modelos GAMLSS el mejor seria mod_GAMLSS2, esto por tener un menor AIC. El modelo GAMLSS2 es el que mejor explica la relación con el creatinina utilizando los mismos predictores.

En tal sentido, tambien se puede utilizar modelos GAMLSS debido a que con estos se puede analizar conjuntos de datos que tengan comportamientos no lineales (no es para el caso tratado) y exista presencia de heterocedasticidad (aspecto que se vio en el cumplimiento de supuestos).

Por el hecho de no presentar predictores no lineales se puede trabajar con el modelo lineal tranformado que cumple los supuestos.

Ejercicio 2

Estudie analítica y gráficamente si: ## 1. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto de la variable estadío.

Para evaluar si existen diferencias estadísticamente significativas en las medias de los valores de creatinina con respecto a la variable “estadío”, se puedes realizar un análisis de ANOVA (Análisis de Varianza) o una prueba no paramétrica, como la prueba de Kruskal-Wallis.

Hide

```
# Realizar análisis de ANOVA
modelo_anova <- aov(creatinina ~ estadio, data = strat_data)

# Obtener los resultados del análisis
resultados_anova <- summary(modelo_anova)

# Imprimir los resultados
print(resultados_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## estadio       4   4.41  1.1030   2.577 0.0377 *
## Residuals   295 126.29  0.4281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La significancia estadística se determina comparando el valor p con el nivel de significancia establecido (por ejemplo, 0.05). En este caso, como el valor p (0.0377) es menor que 0.05, se concluye que existe una diferencia estadísticamente significativa en las medias de los valores de creatinina entre los diferentes niveles de “estadio”.

Si no se cumple los supuestos del modelo anova se aplica la prueba de Kruskal-Wallis

Hide

```
# Realizar prueba de Kruskal-Wallis
resultado_kruskal <- kruskal.test(creatinina ~ estadio, data = strat_data)

# Imprimir el resultado
print(resultado_kruskal)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  creatinina by estadio
## Kruskal-Wallis chi-squared = 8.7815, df = 4, p-value = 0.0668
```

El valor p obtenido (0.0668) indica que no se alcanza un nivel de significancia estadística (por ejemplo, 0.05), lo cual significa que no hay suficiente evidencia para concluir que existen diferencias estadísticamente significativas en las medianas de los valores de creatinina entre los diferentes niveles de “estadio”.

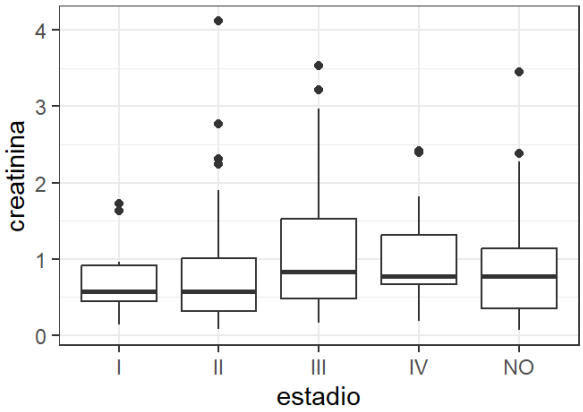
Sin embargo, es importante tener en cuenta que el valor p está cerca del umbral de significancia, por lo que es posible que haya cierta evidencia sugiriendo diferencias en las medianas. Dado que el valor p es mayor que 0.05, no se puede rechazar la hipótesis nula de que las medianas son iguales.

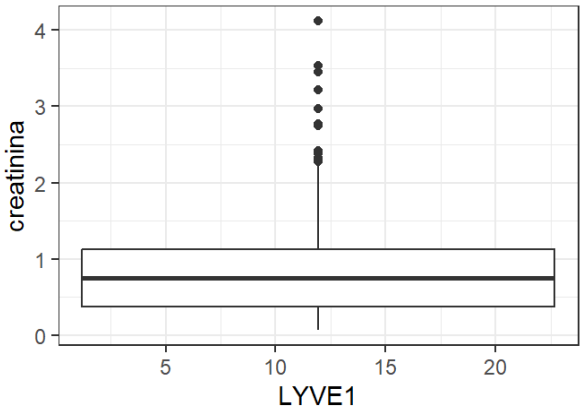
Hide

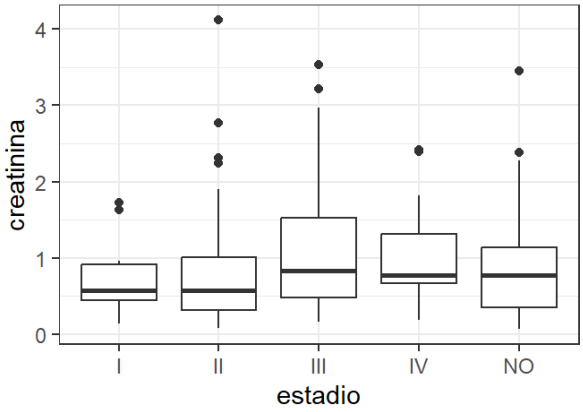
```
p1 <- ggplot(data = strat_data, mapping = aes(x = estadio, y = creatinina)) + geom_boxplot() +
  theme_bw()
p2 <- ggplot(data = strat_data, mapping = aes(x = LYVE1, y = creatinina)) + geom_boxplot() +
  theme_bw()
p3 <- ggplot(data = strat_data, mapping = aes(x = estadio, y = creatinina, colour = LYVE1)) +
  geom_boxplot() + theme_bw()
grid.arrange(p1, p2,p3, ncol = 2)
```

```
## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```







Graficamente se observa una leve diferencia entre las medianas de los boxplot, lo observado en los test.

2. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto de la variable estadío considerando sólo la base de pacientes enfermos.

Hide

```
# Filtrar la base de datos para obtener solo pacientes enfermos
enfermos <- subset(strat_data, diagnosis == "maligno")

# Realizar análisis de ANOVA en el subconjunto de pacientes enfermos
modelo_anova_enfermos <- aov(creatinina ~ estadio, data = enfermos)

# Obtener los resultados del análisis
resultados_anova_enfermos  <- summary(modelo_anova_enfermos )

# Imprimir los resultados
print(resultados_anova_enfermos )
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## estadio       3   4.00  1.3349   2.575 0.0561 .
## Residuals   152  78.81  0.5185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso, el valor p (0.0561) es ligeramente mayor que 0.05, pero se acerca a la significancia estadística. Esto sugiere que podría existir una diferencia en las medias de los valores de creatinina entre los diferentes niveles de “estadio” en el subconjunto de pacientes enfermos, aunque no es estadísticamente significativa de manera concluyente.

Hide

```
# Filtrar la base de datos para obtener solo pacientes enfermos
enfermos_kruskal <- subset(strat_data, diagnosis == "maligno")

# Realizar prueba de Kruskal-Wallis en el subconjunto de pacientes enfermos
resultado_kruskal_enfermos <- kruskal.test(creatinina ~ estadio, data = enfermos_kruskal )

# Imprimir el resultado
print(resultado_kruskal_enfermos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  creatinina by estadio
## Kruskal-Wallis chi-squared = 8.7051, df = 3, p-value = 0.03348
```

El valor p obtenido (0.03348) indica que se alcanza un nivel de significancia estadística (por ejemplo, 0.05), lo cual significa que existe evidencia suficiente para concluir que hay diferencias estadísticamente significativas en las medianas de los valores de creatinina entre los diferentes niveles de “estadio” en el subconjunto de pacientes enfermos.

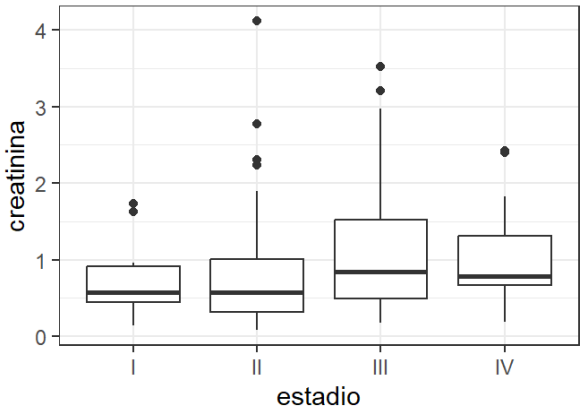
Por lo tanto, con base en la prueba de Kruskal-Wallis, se puede concluir que existe una diferencia estadísticamente significativa en las medianas de los valores de creatinina entre los diferentes niveles de “estadio” en el subconjunto de pacientes enfermos.

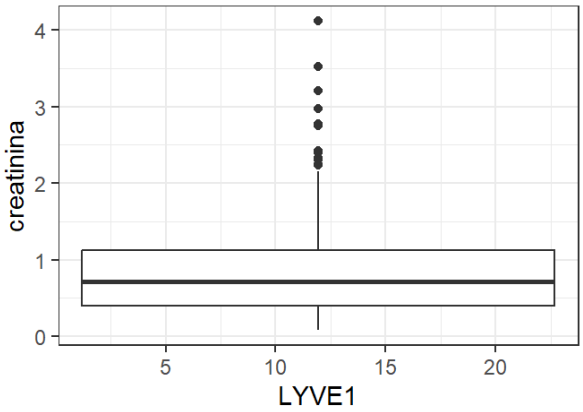
Hide

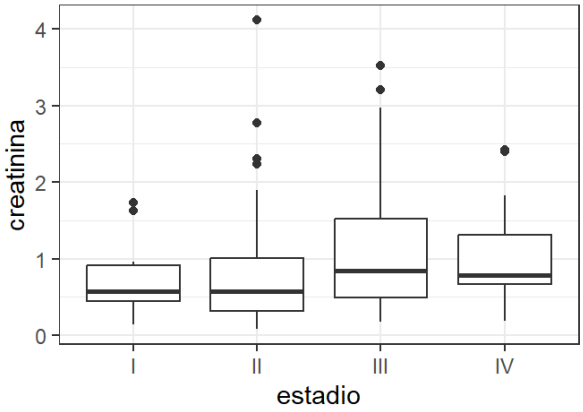
```
p11 <- ggplot(data = enfermos_kruskal, mapping = aes(x = estadio, y = creatinina)) + geom_boxplot() +
  theme_bw()
p22 <- ggplot(data = enfermos_kruskal, mapping = aes(x = LYVE1, y = creatinina)) + geom_boxplot() +
  theme_bw()
p33 <- ggplot(data = enfermos_kruskal, mapping = aes(x = estadio, y = creatinina, colour = LYVE1)) +
  geom_boxplot() + theme_bw()
grid.arrange(p11, p22, p33, ncol = 2)
```

```
## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```







Graficamente se observa que existe diferencia estadistica entre las medias de los niveles de los pacientes enfermos.

3. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto del sexo.

Hide

```
# Realizar análisis de ANOVA
modelo_anova_sexo <- aov(creatinina ~ sexo, data = strat_data)

# Obtener los resultados del análisis
resultados_anova_sexo <- summary(modelo_anova_sexo)

# Imprimir los resultados
print(resultados_anova_sexo)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sexo          1   2.88  2.8752   6.703 0.0101 *
## Residuals   298 127.82  0.4289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso, el valor p (0.0101) es menor que 0.05, por lo que se concluye que existe una diferencia estadísticamente significativa en las medias de los valores de creatinina entre hombres y mujeres.

4. la interacción entre estadío y sexo es significativa cuando se considera la base completa.

Hide

```
# Realizar análisis de ANOVA con interacción
modelo_anova_interaccion <- aov(creatinina ~ estadio * sexo, data = strat_data)

# Obtener los resultados del análisis
resultados_anova_interaccion <- summary(modelo_anova_interaccion)

# Imprimir los resultados
print(resultados_anova_interaccion)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## estadio        4   4.41  1.1030   2.651 0.0335 *
## sexo           1   2.13  2.1326   5.125 0.0243 *
## estadio:sexo   4   3.49  0.8716   2.095 0.0816 .
## Residuals    290 120.67  0.4161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La interacción entre las variables sexo y estadio son significativas teniendo un nivel de significancia de 0.1.Individualmente cumplen un nivel de significancia de 0.05.

5. se satisfacen los supuestos del modelo en 1, 2 y 3. En caso negativo intente una transformación adecuada sobre la variable respuesta en cada modelo y revise nuevamente los supuestos.

supuestos modelos anova

Hide

```
#modelo 1=modelo_anova
cumplimientoSupuestos(modelo_anova)[1,1] #La funcion es creada para modelos lineales, solo para compro bar la normalidad por shapiro
```

```
## [1] "Los residuos del modelo no son normales basado en el test de Shapiro"
```

Hide

```
leveneTest(creatinina~estadio,data=strat_data)
```

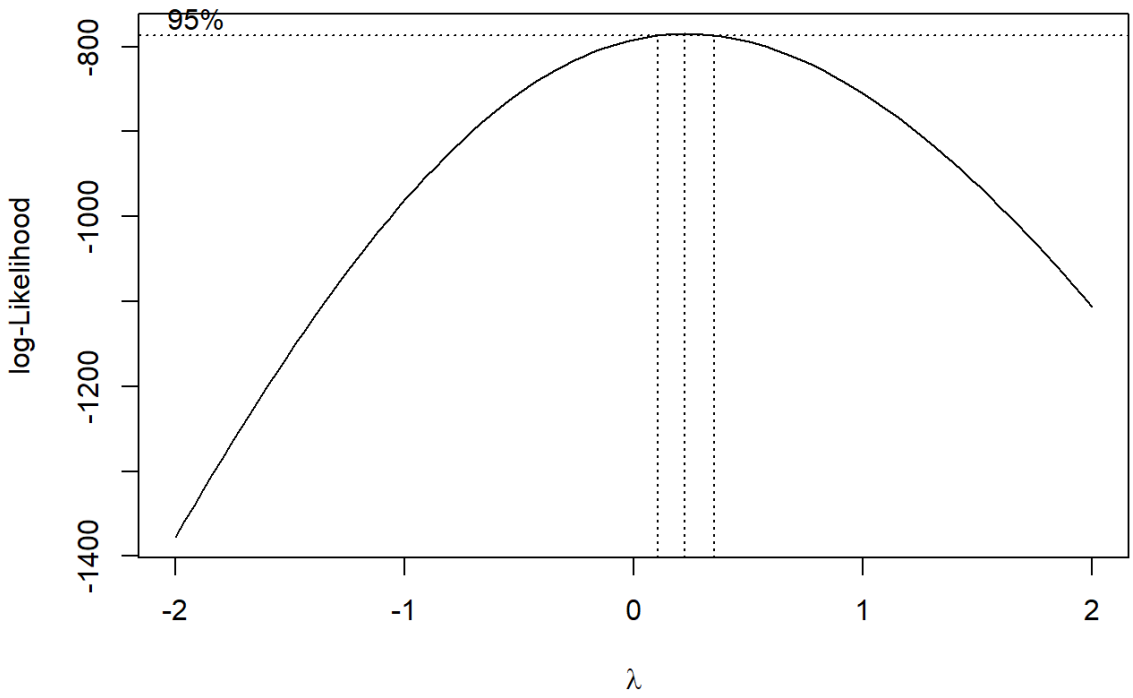
	Df <int>	F value <dbl>	Pr(>F) <dbl>
group	4	1.430301	0.2238923
	295	*NA*	*NA*

2 rows

Es homogenia la varianza

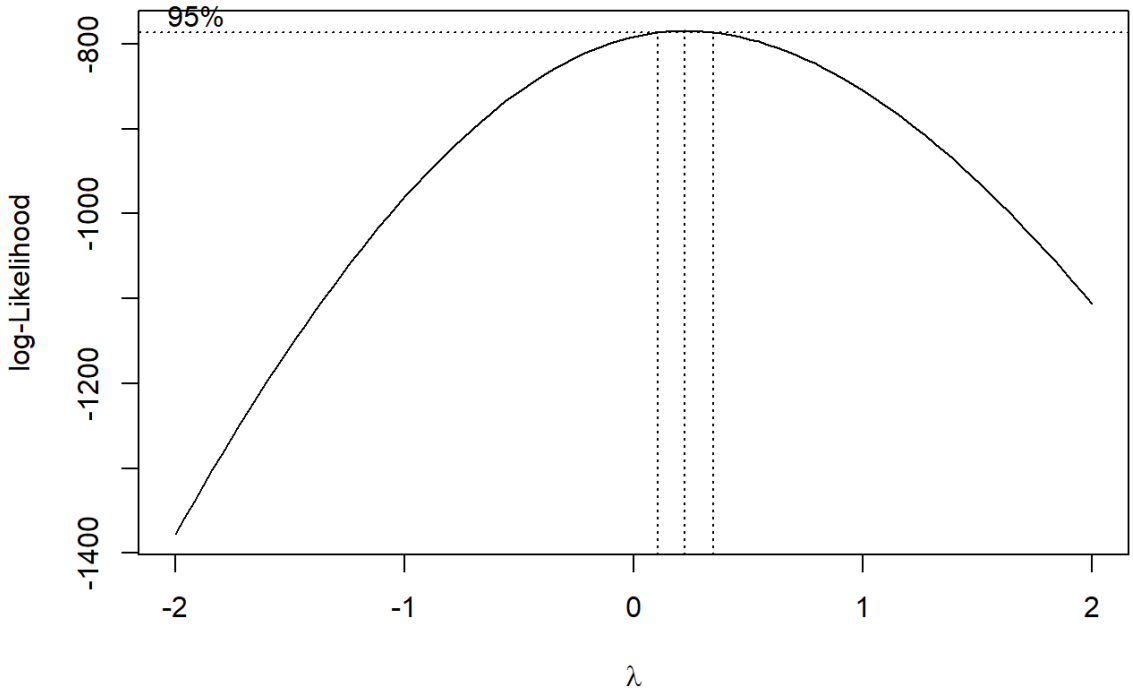
Hide

```
boxcox(creatinina~estadio,data=strat_data,plotit=TRUE)
```



Hide

```
box_cox_resultEje2P5 <- boxcox(creatinina ~ estadio , lambda = -2:2, data = strat_data)
```



Hide

```
best_box_coxEje2P5 <- box_cox_resultEje2P5$x[which.max(box_cox_result$y)]
best_box_coxEje2P5
```

```
## [1] 0.2626263
```

Hide

```
# Realizar análisis de ANOVA
modelo_anovaEje2P5 <- aov((creatinina)^(best_box_coxEje2P5) ~ estadio, data = strat_data)

# Obtener los resultados del análisis
resultados_anovaEje2P5 <- summary(modelo_anovaEje2P5)

# Imprimir los resultados
print(resultados_anovaEje2P5)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## estadio       4  0.363 0.09080   2.596 0.0366 *
## Residuals   295 10.319 0.03498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Singnificativa

Hide

```
cumplimientoSupuestos(modelo_anovaEje2P5)[1,1]
```

```
## [1] "Los residuos del modelo son normales basado en el test de Shapiro"
```

La transformación permite el cumplimiento de normalidad

Hide

```
#modelo 2=modelo_anova_enfermos
cumplimientoSupuestos(modelo_anova_enfermos)[1,1]
```

```
## [1] "Los residuos del modelo no son normales basado en el test de Shapiro"
```

Hide

```
leveneTest(creatinina~estadio,data=enfermos)
```

	Df <int>	F value <dbl>	Pr(>F) <dbl>
group	3	1.222374	0.3036152
	152	*NA*	*NA*

2 rows

Es homogenia la varianza

Hide

```
# Realizar análisis de ANOVA
modelo_anova2Eje2P5 <- aov((creatinina)^(best_box_coxEje2P5) ~ estadio, data = enfermos)

# Obtener los resultados del análisis
resultados_anova2Eje2P5 <- summary(modelo_anova2Eje2P5)

# Imprimir los resultados
print(resultados_anova2Eje2P5)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## estadio       3  0.344  0.1146   3.247 0.0236 *
## Residuals   152  5.366  0.0353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

significativo

Hide

```
cumplimientoSupuestos(modelo_anova2Eje2P5)[1,1]
```

```
## [1] "Los residuos del modelo son normales basado en el test de Shapiro"
```

Con la transformación se cumple normalidad.

Hide

```
#modelo 3=resultados_anova_sexo
cumplimientoSupuestos(modelo_anova_sexo)[1,1]
```

```
## [1] "Los residuos del modelo no son normales basado en el test de Shapiro"
```

Hide

```
leveneTest(creatinina~sexo,data=strat_data)
```

	Df <int>	F value <dbl>	Pr(>F) <dbl>
group	1	1.977132	0.1607347
	298	*NA*	*NA*

2 rows

Es homogenia la varianza

Hide

```
# Realizar análisis de ANOVA
modelo_anova3Eje2P5 <- aov((creatinina)^(best_box_coxEje2P5) ~ sexo, data = strat_data)

# Obtener los resultados del análisis
resultados_anova3Eje2P5 <- summary(modelo_anova3Eje2P5)

# Imprimir los resultados
print(resultados_anova3Eje2P5)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## sexo          1  0.297 0.29721   8.529 0.00376 **
## Residuals   298 10.385 0.03485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

significativo.

Hide

```
#modelo 3=resultados_anova_sexo
cumplimientoSupuestos(modelo_anova3Eje2P5)[1,1]
```

```
## [1] "Los residuos del modelo son normales basado en el test de Shapiro"
```

Con la transformación se cumple normalidad.

6. Obtenga conclusiones acerca de dónde se observan las diferencias si las hubiere.

Las diferencias se observan en el nivel III y IV de la variable estadio.

Ejercicio 3

1. Ajuste un modelo logístico para predecir el diagnóstico de cáncer de páncreas en función de las variables en la base que considere razonables.

Hide

```
# Crear una nueva variable binaria para representar 'diagnosis'
strat_data$bin_diagnosis <- ifelse(strat_data$diagnosis == "maligno", 1, 0)

# Ajustar el modelo logístico con la nueva variable binaria
modelo_logistico <- glm(bin_diagnosis ~ edad , data = strat_data, family = "binomial")

# Realizar el análisis del modelo
summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = bin_diagnosis ~ edad, family = "binomial", data = strat_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1050  -1.0593   0.6022   0.9890   1.9002
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.53167    0.73894  -6.133 8.64e-10 ***
## edad         0.07451    0.01176   6.334 2.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.41  on 299  degrees of freedom
## Residual deviance: 365.70  on 298  degrees of freedom
## AIC: 369.7
##
## Number of Fisher Scoring iterations: 4
```

El coeficiente estimado para “edad” es de 0.07451, lo que significa que por cada aumento de un año en la edad, la log-odds de tener un diagnóstico positivo de cáncer de páncreas aumenta en promedio 0.07451 unidades.

El coeficiente de intercepción es de -4.53167, lo que indica que cuando la edad es igual a cero (lo cual no tiene sentido práctico en este contexto), la log-odds de tener un diagnóstico positivo de cáncer de páncreas es de -4.53167.

Ambos coeficientes tienen errores estándar asociados y valores de p. El coeficiente para “edad” tiene un valor de p extremadamente bajo (2.38e-10), lo que indica que es altamente significativo y sugiere una relación estadísticamente significativa entre la edad y el diagnóstico de cáncer de páncreas.

La desviación residual es de 365.70 y el AIC (criterio de información de Akaike) es de 369.7. Estos valores se utilizan para comparar modelos alternativos, donde valores más bajos indican un mejor ajuste.

2. Evalúe la calidad de ajuste del modelo con al menos dos criterios distintos.

Hide

```
# Obtener la deviance nula
null_deviance <- as.numeric(anova(modelo_logistico, test = "Chisq")$Deviance[1])

# Obtener la deviance residual
deviance_residual <- deviance(modelo_logistico)

# Obtener los grados de libertad
null_df <- as.numeric(anova(modelo_logistico, test = "Chisq")$Df[1])
residual_df <- df.residual(modelo_logistico)

# Calcular el valor de p para el test de chi-cuadrado
p_value <- pchisq(null_deviance - deviance_residual, df = null_df - residual_df, lower.tail = FALSE)


# Calcular el AIC y BIC
AIC <- AIC(modelo_logistico)
BIC <- BIC(modelo_logistico)


print(paste("El el criterio de AIC del modelo logistico es : ",AIC))
```

```
## [1] "El el criterio de AIC del modelo logistico es :  369.699151699446"
```

Hide

```
print(paste("El el criterio de BIC del modelo logistico es : ",BIC))
```

```
## [1] "El el criterio de BIC del modelo logistico es :  377.106716648759"
```

Se calcula para el mejor modelo obtenido en los puntos anteriores

Hide

```
# Calcular el AIC y BIC
AIC2 <- AIC(modele2je1P3)
BIC2 <- BIC(modele2je1P3)


print(paste("El el criterio de AIC del modelo lineal transformado es : ",AIC2))
```

```
## [1] "El el criterio de AIC del modelo lineal transformado es :  -205.365444180481"
```

Hide

```
print(paste("El el criterio de BIC del modelo lineal transformado es : ",BIC2))
```

```
## [1] "El el criterio de BIC del modelo lineal transformado es :  -164.623836959263"
```

El resultado muestra que por los criterios de AIC y BIC es mejor el modelo lineal ajustado. Hay que tener en cuenta que el lineal usa 5 variables y el logistico solo 1.

3. Interprete los coeficientes del modelo elegido.

Hide

```
# Extraer los coeficientes del modelo logístico y asignarlos a variables individuales
intercepto <- modelo_logistico$coefficients["(Intercept)"]
coef_edad <- modelo_logistico$coefficients["edad"]


# Conversion razon odd (relación de probabilidades)
oddIntercepto=exp(intercepto)
oddEdad=exp(coef_edad)


# probabilidad

probabilidad_oddIntercepto=(oddIntercepto/(1+oddIntercepto))*100
probabilidad_oddEdad=(oddEdad/(1+oddEdad))*100


print(paste("El coefiente transformado del intercepto es: ",oddIntercepto, ", la probabilidad de éxito
del evento seria ", probabilidad_oddIntercepto, "%"))
```

```
## [1] "El coefiente transformado del intercepto es:  0.0107626805033775 , la probabilidad de éxito de
l evento seria  1.06480786350535 %"
```

Hide

```
print(paste("El coefiente transformado de la variable Edad es: ",oddEdad,", la probabilidad de éxito d
el evento seria ",probabilidad_oddEdad, "%"))
```

```
## [1] "El coefiente transformado de la variable Edad es:  1.0773590284445 , la probabilidad de éxito
del evento seria  51.861956151663 %"
```