# Preprocesado.R

user

2025-06-28

```
# Librerías ---------------------------------------------------------------
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'
```

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

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

```
library(readxl)
library(ggplot2)
library(GGally)
```

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

```
library(RcmdrMisc)
```

```
## Cargando paquete requerido: car
```

```
## Cargando paquete requerido: carData
```

```
##
## Adjuntando el paquete: 'car'
```

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

```
## Cargando paquete requerido: sandwich
```

```r
library(car)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.0.4     v tidyr     1.3.1
## v readr     2.1.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x car::recode()   masks dplyr::recode()
## x purrr::some()   masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(modelr)
library(devtools)
```

```
## Cargando paquete requerido: usethis
```

```r
library(conflicted)
library(MASS)
library(effects)
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
library(nortest)
library(DescTools)
library(lmtest)
```

```
## Cargando paquete requerido: zoo
##
## Adjuntando el paquete: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(MASS)
library(sjPlot)
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")`)
```

```r
library(olsrr)

Sys.setlocale("LC_ALL", "Spanish_Spain.utf8")
```

```
## [1] "LC_COLLATE=Spanish_Spain.utf8;LC_CTYPE=Spanish_Spain.utf8;LC_MONETARY=Spanish_Spain.utf8;LC_NUM
```

```r
# Importando datos ## no dejaba por medio del comando solo de read_exel##
setwd("C:/Users/user/3D Objects/Modelos/Proyecto_Icfes/Data")
datos <- readxl::read_excel("DATOS_conjunto.xlsx");datos
```

```
## # A tibble: 5,465 x 9
##    fami_tieneinternet punt_c_naturales punt_lectura_critica punt_matematicas
##    <chr>                        <dbl>                <dbl>            <dbl>
##  1 No                              44                   46               46
##  2 No                              51                   44               37
##  3 No                              39                   46               36
##  4 No                              30                   43               28
##  5 <NA>                            38                   50               48
##  6 <NA>                            47                   51               51
##  7 Si                              59                   47               66
##  8 No                              65                   71               64
##  9 No                              56                   67               58
## 10 No                              52                   47               43
## # i 5,455 more rows
## # i 5 more variables: estu_genero <chr>, estu_nse_individual <dbl>,
## #   cole_area_ubicacion <chr>, fami_estratovivienda <chr>,
## #   fami_educacionmadre <chr>
```

```r
### Modelo ANCOVA A DOS VIAS.

# Variables a analizar  ----------------------------------------------------

## Variable respuesta: Puntaje de matemáticas
## Variables explicativas (factores): nivel socio económico del
## Nivel del estudiante, nivel educativo de la madre
## Covariable: Puntaje de lectura critica

# Cambiando variables "estu_nse_individual" y "fami_educacionmadre"

datos$estu_nse_individual <- factor(datos$estu_nse_individual, ordered = TRUE)
datos$fami_educacionmadre <- factor(datos$fami_educacionmadre,
                                    levels = c("Ninguno", "Primaria incompleta", "Secundaria (Bachillera
                                               "Primaria completa", "Secundaria (Bachillerato) completa
                                               "Técnica o tecnológica incompleta", "Técnica o tecnológi
                                               "Educación profesional incompleta", "Educación profesiona
                                               "Postgrado", "No sabe"),
                                    labels = c("Educación básica (sin completar)",
                                               "Educación básica (sin completar)",
                                               "Educación básica (sin completar)",
                                               "Educación básica completa",
                                               "Educación básica completa",
                                               "Educación superior",
                                               "Educación superior",
                                               "Educación superior",
                                               "Educación superior",
                                               "Educación superior",
                                               "Educación básica (sin completar)"))  # "No sabe" a "Edu

# Convertir a factor ordinal
```

```r
datos$fami_educacionmadre <- ordered(datos$fami_educacionmadre,
                                     levels = c("Educación básica (sin completar)",
                                                "Educación básica completa",
                                                "Educación superior"))

## Procedemos a observar si se volvieron de carácter "factor" ordinal

str(datos)
```

```
## tibble [5,465 x 9] (S3: tbl_df/tbl/data.frame)
##  $ fami_tieneinternet  : chr [1:5465] "No" "No" "No" "No" ...
##  $ punt_c_naturales    : num [1:5465] 44 51 39 30 38 47 59 65 56 52 ...
##  $ punt_lectura_critica: num [1:5465] 46 44 46 43 50 51 47 71 67 47 ...
##  $ punt_matematicas    : num [1:5465] 46 37 36 28 48 51 66 64 58 43 ...
##  $ estu_genero         : chr [1:5465] "F" "M" "F" "F" ...
##  $ estu_nse_individual : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 1 3 1 3 NA NA 3 2 2 1 ...
##  $ cole_area_ubicacion : chr [1:5465] "URBANO" "RURAL" "RURAL" "URBANO" ...
##  $ fami_estratovivienda: chr [1:5465] "Estrato 1" "Estrato 2" "Estrato 2" "Estrato 4" ...
##  $ fami_educacionmadre : Ord.factor w/ 3 levels "Educación básica (sin completar)"<..: 1 2 1 2 NA NA
```

```r
# Escoger el mejo modelo ANCOVA ------------------------------------------

### Realizamos el modelo ANCOVA a dos vías y aplicamos la función "stepwise"

M1 <- lm(punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
                              punt_lectura_critica + punt_lectura_critica*fami_educacionmadre +
                              punt_lectura_critica*estu_nse_individual,
                            data = datos)
summary(M1)
```

```
##
## Call:
## lm(formula = punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
##     punt_lectura_critica + punt_lectura_critica * fami_educacionmadre +
##     punt_lectura_critica * estu_nse_individual, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.038  -5.468   0.002   5.656  39.452
##
## Coefficients:
##                                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                   6.87169    0.96965   7.087 1.57e-12
## fami_educacionmadre.L                         4.45684    1.43296   3.110  0.00188
## fami_educacionmadre.Q                         1.75839    1.18545   1.483  0.13806
## estu_nse_individual.L                        -2.79651    2.66780  -1.048  0.29458
## estu_nse_individual.Q                        -0.68338    1.95862  -0.349  0.72717
## estu_nse_individual.C                         1.20463    1.32254   0.911  0.36242
## punt_lectura_critica                          0.86986    0.01654  52.595  < 2e-16
## fami_educacionmadre.L:punt_lectura_critica -0.05959    0.02586  -2.304  0.02127
## fami_educacionmadre.Q:punt_lectura_critica -0.03339    0.02175  -1.535  0.12481
## estu_nse_individual.L:punt_lectura_critica  0.06972    0.04543   1.535  0.12493
## estu_nse_individual.Q:punt_lectura_critica  0.01281    0.03331   0.384  0.70064
```

```
## estu_nse_individual.C:punt_lectura_critica -0.02074    0.02334  -0.889  0.37431
##
## (Intercept)                                 ***
## fami_educacionmadre.L                       **
## fami_educacionmadre.Q
## estu_nse_individual.L
## estu_nse_individual.Q
## estu_nse_individual.C
## punt_lectura_critica                        ***
## fami_educacionmadre.L:punt_lectura_critica *
## fami_educacionmadre.Q:punt_lectura_critica
## estu_nse_individual.L:punt_lectura_critica
## estu_nse_individual.Q:punt_lectura_critica
## estu_nse_individual.C:punt_lectura_critica
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.54 on 4799 degrees of freedom
##   (654 observations deleted due to missingness)
## Multiple R-squared:  0.5327, Adjusted R-squared:  0.5316
## F-statistic: 497.3 on 11 and 4799 DF,  p-value: < 2.2e-16
```

```
### utilizamos "backward" <-> "pasos hacia atrás" (utilizar solo cuando sean datos
###                                                          Completos)
#library(RcmdrMisc)
#stepwise(M1,direction = "backward")


#### NOTA

# El error surge porque la base de datos contiene datos faltantes en las variables
# por ende para realizar este proceso "backward" lo ideal sería tener la base de datos,
# sin ningún dato faltante, recordar que la cantidad de datos faltantes, si supera el 5%
# de la muestra no es una buena opción imputar esos datos, así que, es mejor utilizar,
# otras técnicas. en este caso se trabajo con los datos faltantes, y se sabe que la variables
# explicativas son de carácter fijo.

M_final <- lm(punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
           punt_lectura_critica, data = datos)

summary(M_final)
```

```
##
## Call:
## lm(formula = punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
##     punt_lectura_critica, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.233  -5.452  -0.008   5.677  39.447
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            7.27098    0.70871  10.259  < 2e-16 ***
```

```
## fami_educacionmadre.L  1.16706    0.26226   4.450 8.78e-06 ***
## fami_educacionmadre.Q -0.09014    0.21610  -0.417   0.6766
## estu_nse_individual.L  1.05771    0.44654   2.369   0.0179 *
## estu_nse_individual.Q  0.05065    0.31738   0.160   0.8732
## estu_nse_individual.C  0.09418    0.22896   0.411   0.6809
## punt_lectura_critica   0.86258    0.01258  68.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.543 on 4804 degrees of freedom
##   (654 observations deleted due to missingness)
## Multiple R-squared:  0.5319, Adjusted R-squared:  0.5313
## F-statistic: 909.7 on 6 and 4804 DF,  p-value: < 2.2e-16
```

```
# Verificación de supuestos ------------------------------------------------

### Independencia de residuos (Errores).
## Prueba de durbin-watson
library(lmtest)
dwt(M_final) # o
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.005022934      2.009914    0.73
##  Alternative hypothesis: rho != 0
```

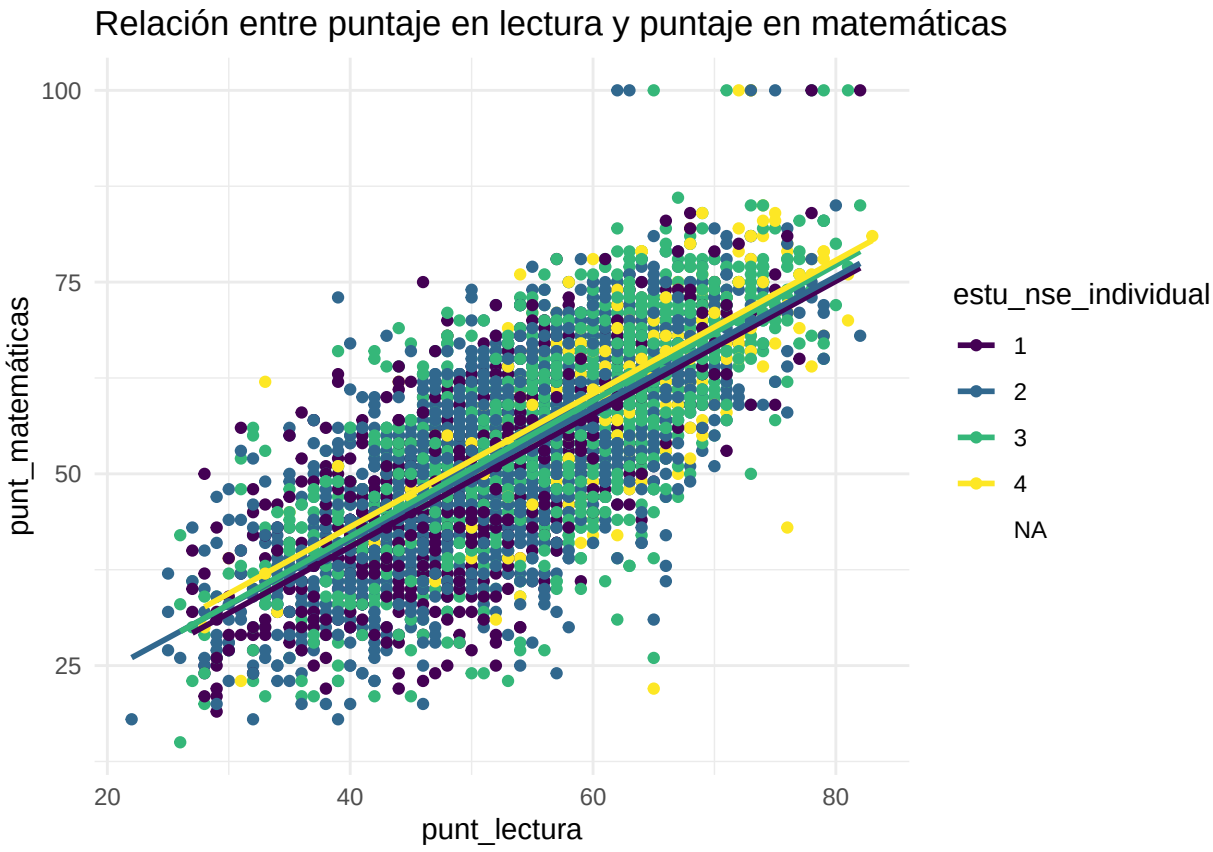
```
dwtest(M_final)
```

```
##
##  Durbin-Watson test
##
## data:  M_final
## DW = 2.0099, p-value = 0.6348
## alternative hypothesis: true autocorrelation is greater than 0
```

```
### Relación lineal (Variable respuestas vs Covariable) ajustada por un factor
# se verifica por medio de un gráfico.

ggplot(datos, aes(x = punt_lectura_critica, y = punt_matematicas,
                  color = estu_nse_individual)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(title = "Relación entre puntaje en lectura y puntaje en matemáticas",
       x = "punt_lectura", y = "punt_matemáticas")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 623 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



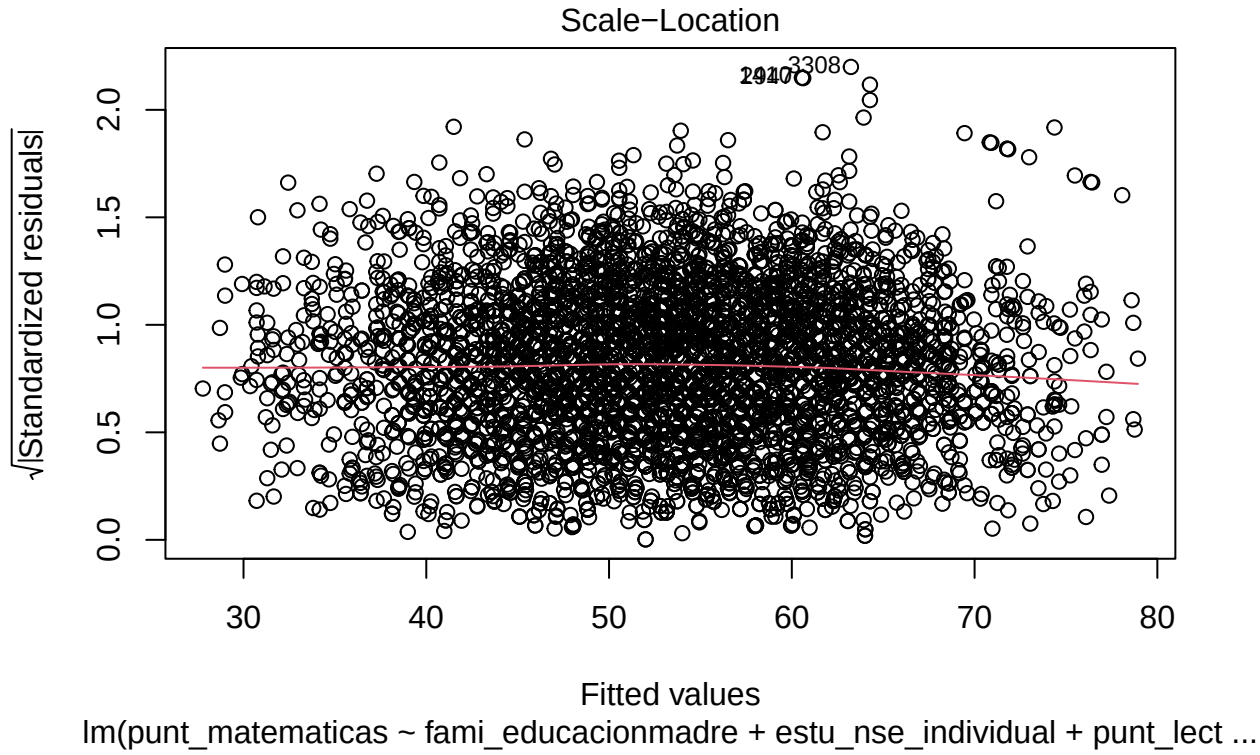

```
### Homogeneidad de varianzas (ANOVA) o pendientes.
## para comparar la homogeneidad de pendientes en este caso se realizo a partir
## de compara dos modelos, uno con y sin iteraciones, y comprobar el p-valor
## usando la función anova, este debe ser no significativo (P-valor > 0.05).

modelo_sin_interaccion <- lm(punt_matematicas ~ fami_educacionmadre +
        estu_nse_individual + punt_lectura_critica, data = datos)
modelo_con_interaccion <- lm(punt_matematicas ~ fami_educacionmadre +
        estu_nse_individual + punt_lectura_critica * fami_educacionmadre,
        data = datos)
anova(modelo_sin_interaccion,modelo_con_interaccion)
```

```
## Analysis of Variance Table
##
## Model 1: punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
##     punt_lectura_critica
## Model 2: punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
##     punt_lectura_critica * fami_educacionmadre
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4804 350629
## 2   4802 350368  2    261.29 1.7906  0.167
```

```
### Homocedasticidad y un gráfico.
bptest(M_final,studentize = TRUE)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  M_final
## BP = 3.1945, df = 6, p-value = 0.7841
```

```
plot(M_final, which = 3)
```



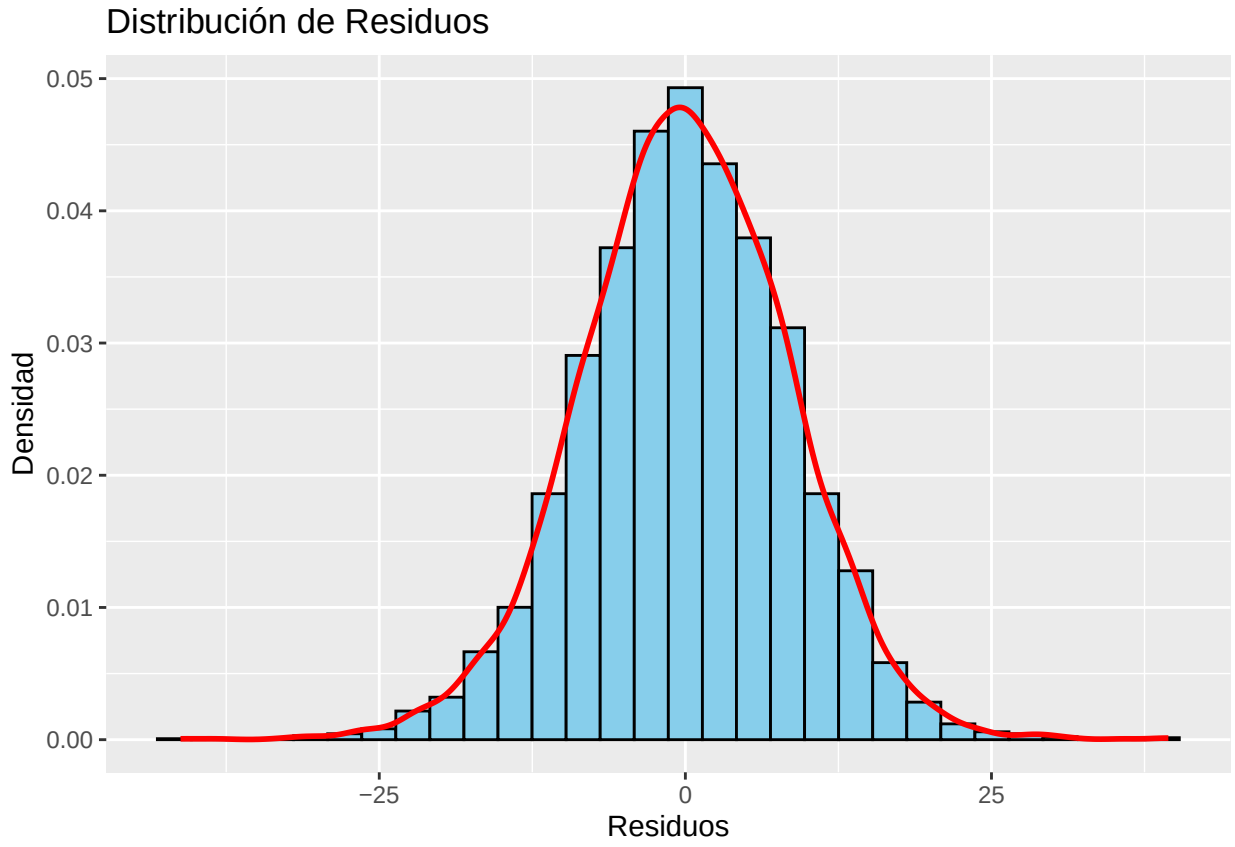

```
### Normalidad de residuos mediante gráficos y test de kolmo-gorov

ks.test(residuals(M_final), "pnorm", 0, sd(residuals(M_final)))
```

```
## Warning in ks.test.default(residuals(M_final), "pnorm", 0,
## sd(residuals(M_final))): ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

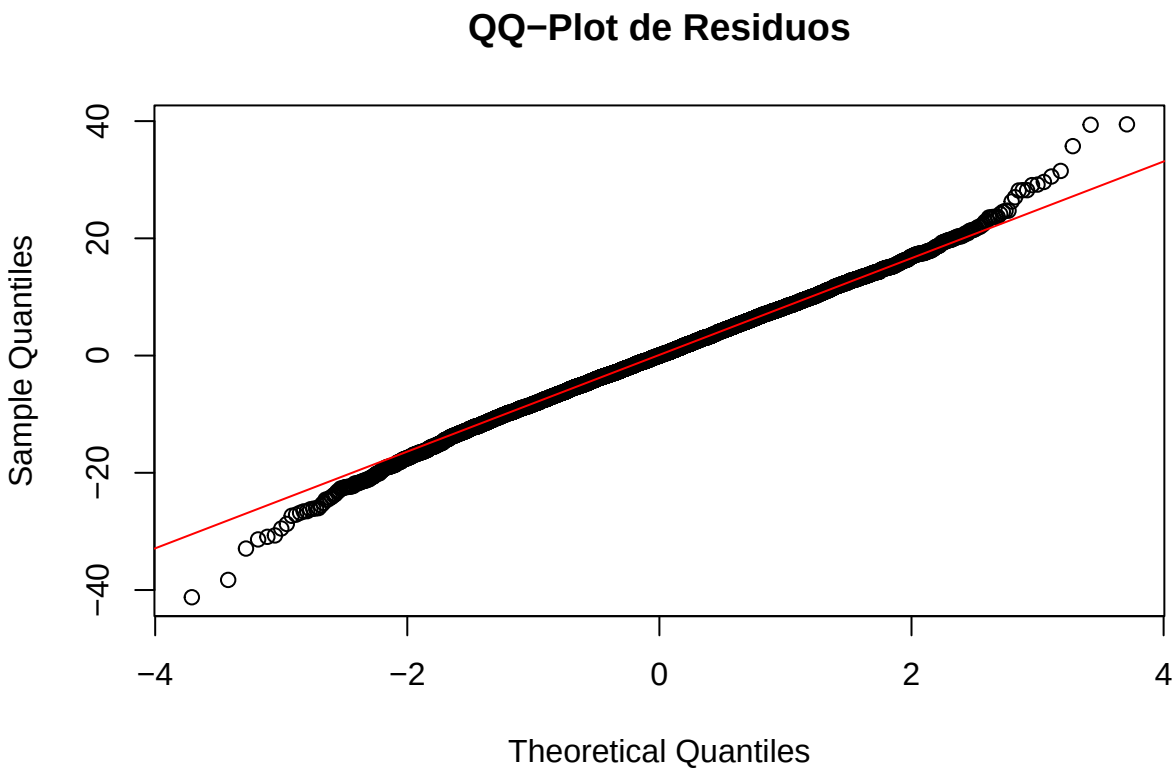
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  residuals(M_final)
## D = 0.012407, p-value = 0.4494
## alternative hypothesis: two-sided
```

```r
## Creación del objeto "residuos"
residuos <- residuals(M_final)

# Histograma
Hist_residuales <- ggplot(data.frame(Residuos = residuos), aes(x = Residuos)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
  geom_density(color = "red", linewidth = 1) +
  labs(title = "Distribución de Residuos", x = "Residuos", y = "Densidad");Hist_residuales
```

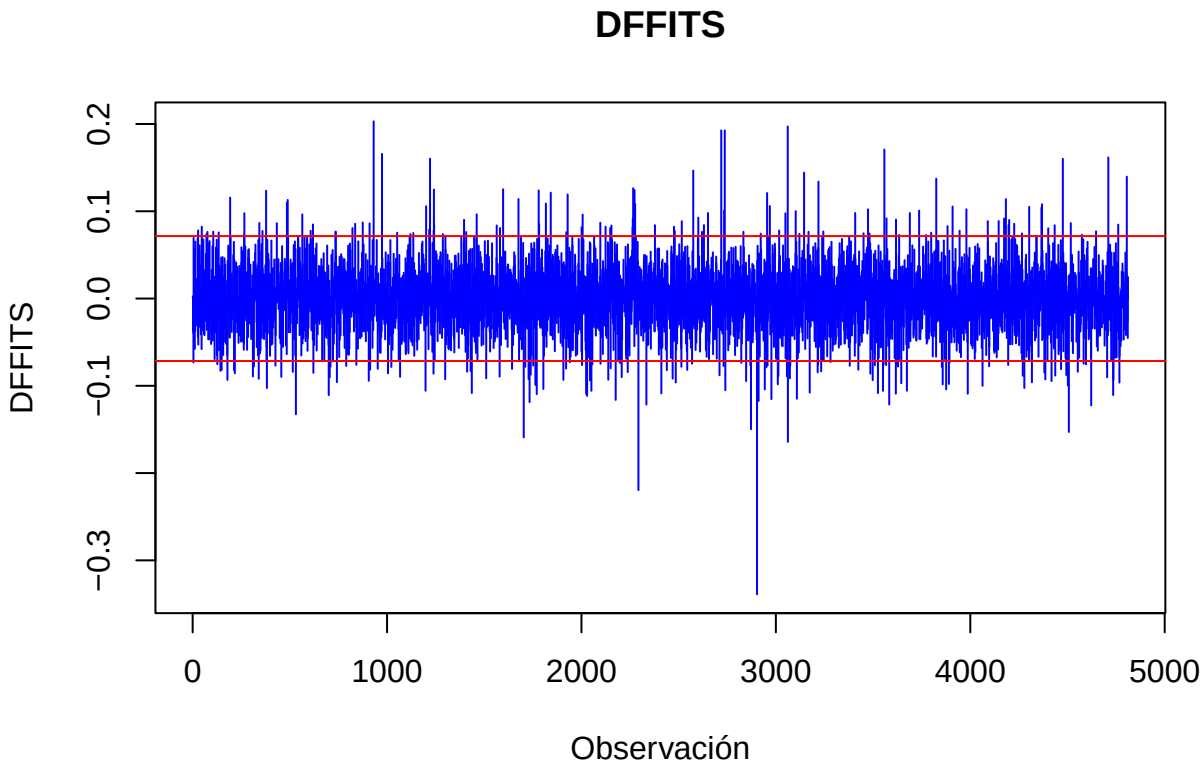
```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



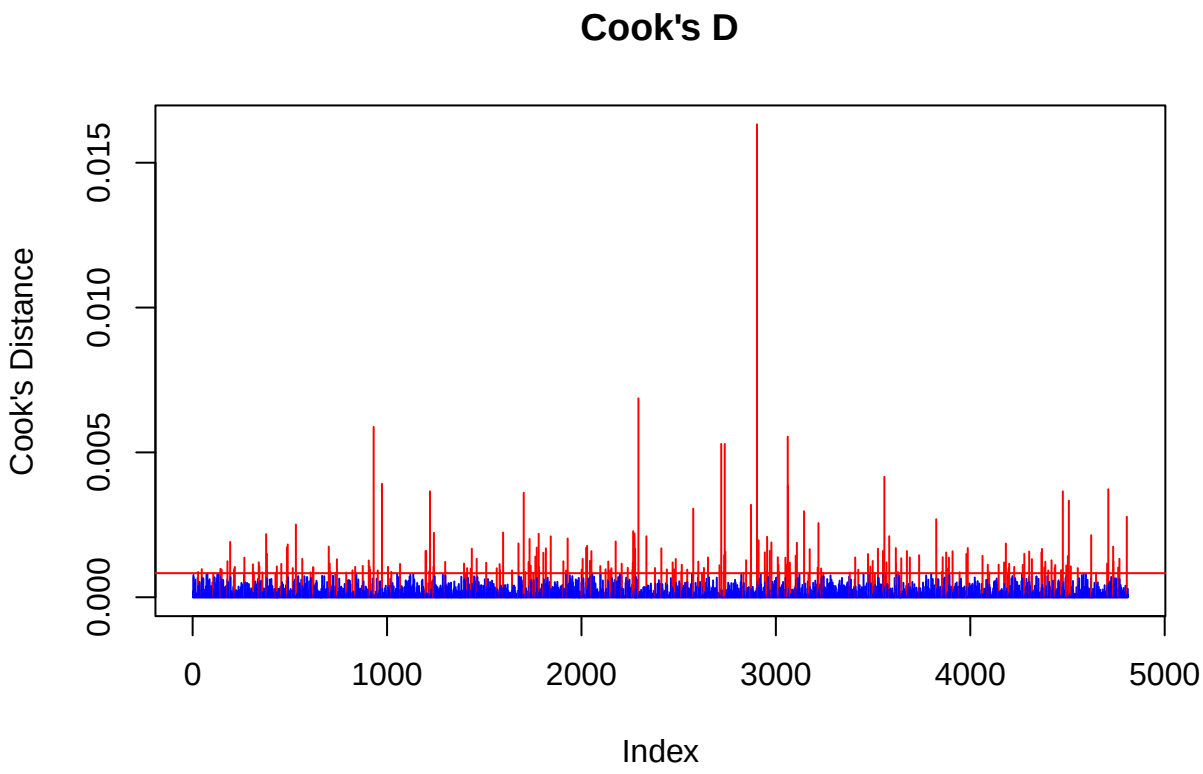

```r
# Q_Q
qqnorm(residuos, main = "QQ-Plot de Residuos")
qqline(residuos, col = "red")
```




```r
# Detección de datos influyentes ----------------------------------------
## Objetos para gráficas
influence_measures <- influence.measures(M_final)
dffits_values <- dffits(M_final)
cooks_d <- cooks.distance(M_final)
hat_values <- hatvalues(M_final)
stud_res <- rstudent(M_final)

## Gráfico DFFITS
plot(dffits_values, type = "h", col = "blue", main = "DFFITS", ylab = "DFFITS", xlab = "Observación")
abline(h = c(2*sqrt(length(coef(M_final))/nrow(datos)),
             -2*sqrt(length(coef(M_final))/nrow(datos))), col = "red")
```

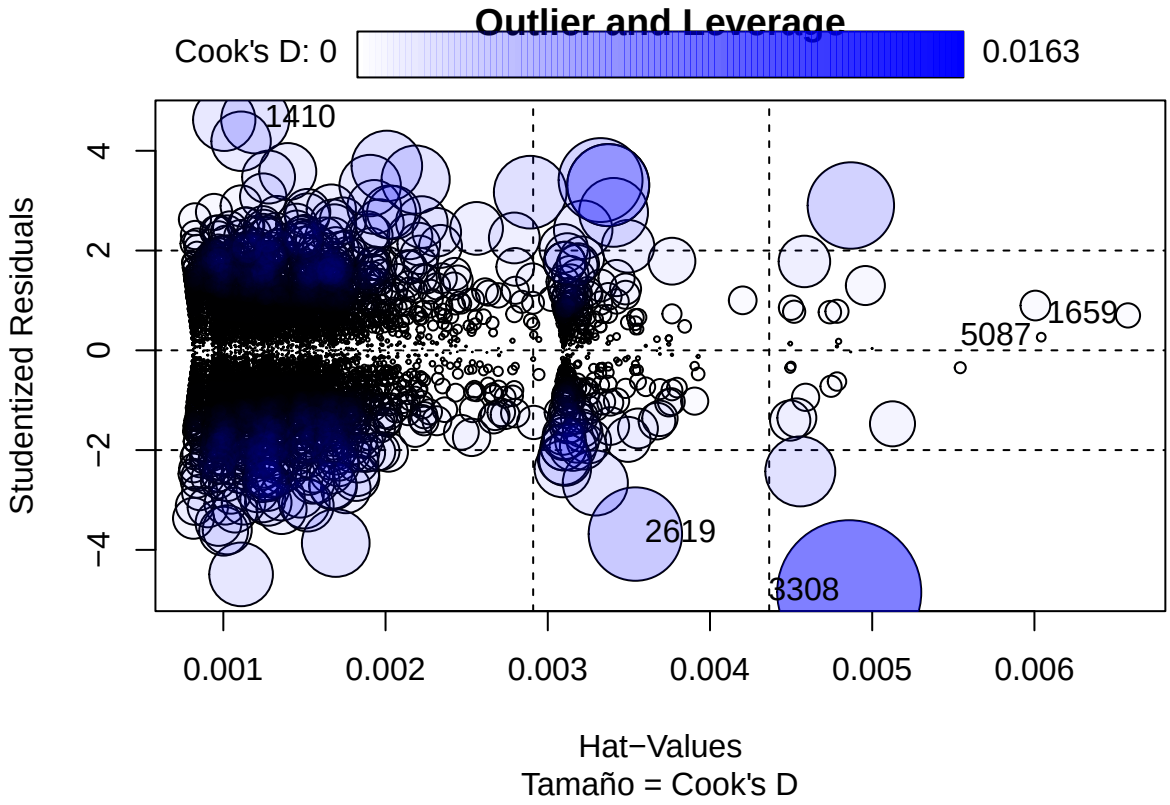



```
## Gráfico D-DCOOKS

plot(cooks_d, type = "h", col = ifelse(cooks_d > 4/length(cooks_d), "red", "blue"),
     main = "Cook's D", ylab = "Cook's Distance")
abline(h = 4/length(cooks_d), col = "red")
```



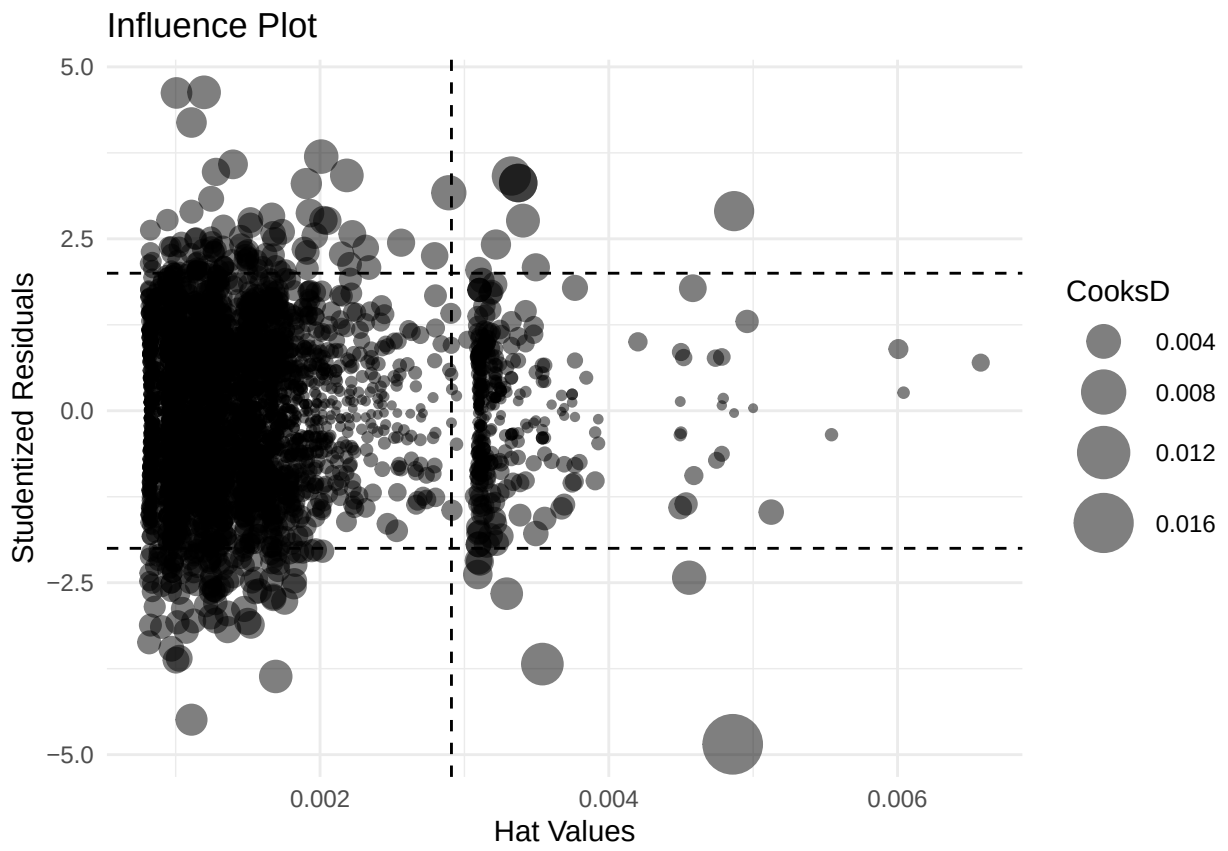

```
## Gráfico de ATÍPICOS Y APALANCAMIENTO
influencePlot(M_final, id.method = "identify", main = "Outlier and Leverage",
              sub = "Tamaño = Cook's D")
```

```
## Warning in plot.window(...): "id.method" es un parámetro gráfico inválido

## Warning in plot.xy(xy, type, ...): "id.method" es un parámetro gráfico inválido

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" es un
## parámetro gráfico inválido
## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" es un
## parámetro gráfico inválido

## Warning in box(...): "id.method" es un parámetro gráfico inválido

## Warning in title(...): "id.method" es un parámetro gráfico inválido

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" es un
## parámetro gráfico inválido
```




```
##          StudRes         Hat        CookD
## 1410  4.6299429 0.001196547 3.653081e-03
## 1659  0.6996334 0.006576510 4.629672e-04
## 2619 -3.6833921 0.003539911 6.867452e-03
## 3308 -4.8494227 0.004858268 1.632479e-02
## 5087  0.2630909 0.006041951 6.011835e-05
```

```r
## Gráfico de datos influyentes
df_influence <- data.frame(Observacion = 1:length(hat_values),
                           Hat = hat_values,
                           Residual = stud_res,
                           CooksD = cooks_d)

ggplot(df_influence, aes(x = Hat, y = Residual)) +
  geom_point(aes(size = CooksD), alpha = 0.5) +
  scale_size_continuous(range = c(1, 10)) +
  geom_hline(yintercept = c(-2, 2), linetype = "dashed") +
  geom_vline(xintercept = 2 * mean(hat_values), linetype = "dashed") +
  theme_minimal() +
  labs(title = "Influence Plot",
       x = "Hat Values",
       y = "Studentized Residuals")
```




```r
# Imputando los datos influyentes de la muestra ---------------------------

diagnosticos <- data.frame(
  observacion = as.numeric(rownames(M_final$model)),
  rstudent = rstudent(M_final),
  leverage = hatvalues(M_final),
  cooksD = cooks.distance(M_final),
  dffits = dffits(M_final)
)


n <- nrow(M_final$model)
p <- length(coef(M_final))

umbral_residuo <- 2
umbral_leverage <- 2 * p / n
umbral_cooks <- 4 / n
umbral_dffits <- 2 * sqrt(p / n)

outliers <- diagnosticos[
  abs(diagnosticos$rstudent) > umbral_residuo |
    diagnosticos$leverage > umbral_leverage |
    diagnosticos$cooksD > umbral_cooks |
    abs(diagnosticos$dffits) > umbral_dffits,
  "observacion"
]
```

```r
datos_limpios <- datos[-outliers, ]

summary(M_final)
```

```
##
## Call:
## lm(formula = punt_matematicas ~ fami_educacionmadre + estu_nse_individual +
##     punt_lectura_critica, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.233  -5.452  -0.008   5.677  39.447
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.27098    0.70871  10.259  < 2e-16 ***
## fami_educacionmadre.L    1.16706    0.26226   4.450 8.78e-06 ***
## fami_educacionmadre.Q   -0.09014    0.21610  -0.417   0.6766
## estu_nse_individual.L    1.05771    0.44654   2.369   0.0179 *
## estu_nse_individual.Q    0.05065    0.31738   0.160   0.8732
## estu_nse_individual.C    0.09418    0.22896   0.411   0.6809
## punt_lectura_critica     0.86258    0.01258  68.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.543 on 4804 degrees of freedom
##   (654 observations deleted due to missingness)
## Multiple R-squared:  0.5319, Adjusted R-squared:  0.5313
## F-statistic: 909.7 on 6 and 4804 DF,  p-value: < 2.2e-16
```

```r
## Observar la R^2 explicación de variabilidad de datos del modelo
```