Laboratorio 3: Analisis de Varianza (ANOVA) en R

Introduccion

En este análisis, exploraremos el rendimiento estudiantil en dos escuelas portuguesas utilizando un conjunto de datos que contiene una variedad de características demográficas, sociales y relacionadas con la escuela.

Nuestro objetivo principal es entender qué factores influyen en las calificaciones finales de los estudiantes.

Cargar los datos

```
datos <- read.csv("student-por.csv")
head(datos)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
## 6   mother          1         2        0        no    yes   no        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        4  0 11 11
## 2        2  9 11 11
## 3        6 12 13 12
## 4        0 14 14 14
## 5        0 11 13 13
## 6        6 12 12 13
```

Descrición de las variables

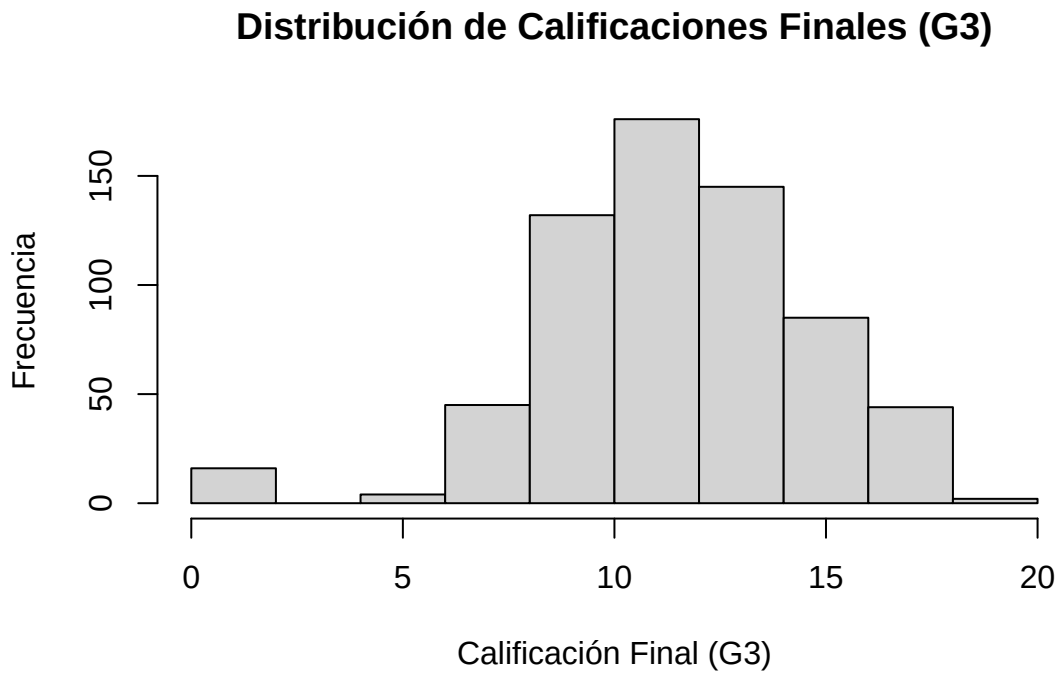
```
variables_categoricas <- c("school", "sex", "address")

variables_independientes <- c("G1", "G2", "absences", "failures", "studytime", "freetime", "goout", "Da

variable_dependiente <- "G3"
```

```r
residuos <- list()
```

Verificar normalidad

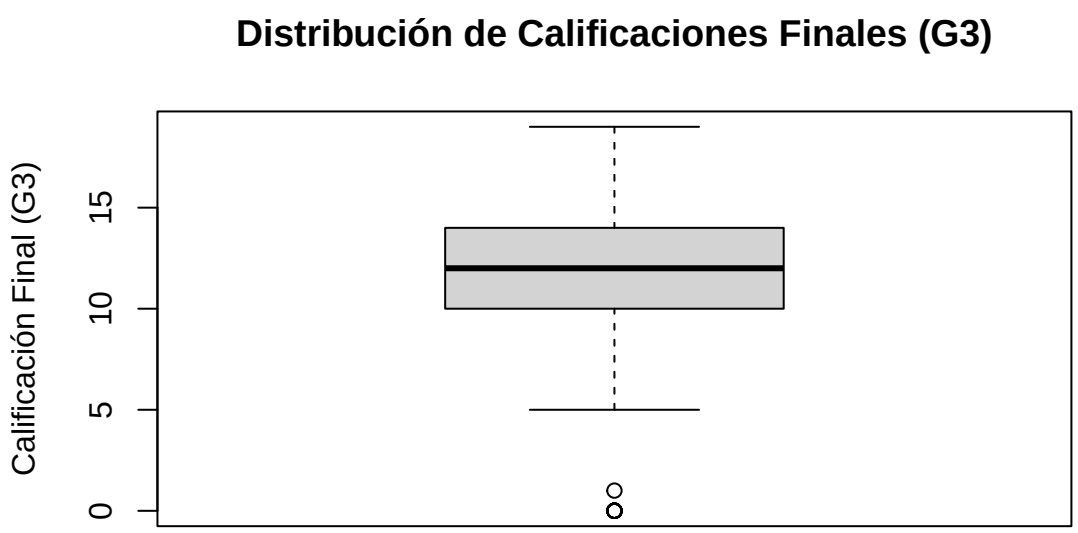
Histograma G3 sin transformar

```r
hist(datos$G3, main = "Distribución de Calificaciones Finales (G3)", xlab = "Calificación Final (G3)", y
```



Boxplot G3 sin transformar

```r
boxplot(datos$G3, main = "Distribución de Calificaciones Finales (G3)", ylab = "Calificación Final (G3)
```



Prueba Anderson-Darling para G3 sin transformar

```
if (!require(nortest)) { # Forma para descargar e instalar un paquete en caso de que no esté instalado
  install.packages("nortest")
  library(nortest)
}
```

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

```
# Realizar la prueba de Anderson Darling
ad_test_result <- ad.test(datos$G3) # nolint
print(ad_test_result)
```

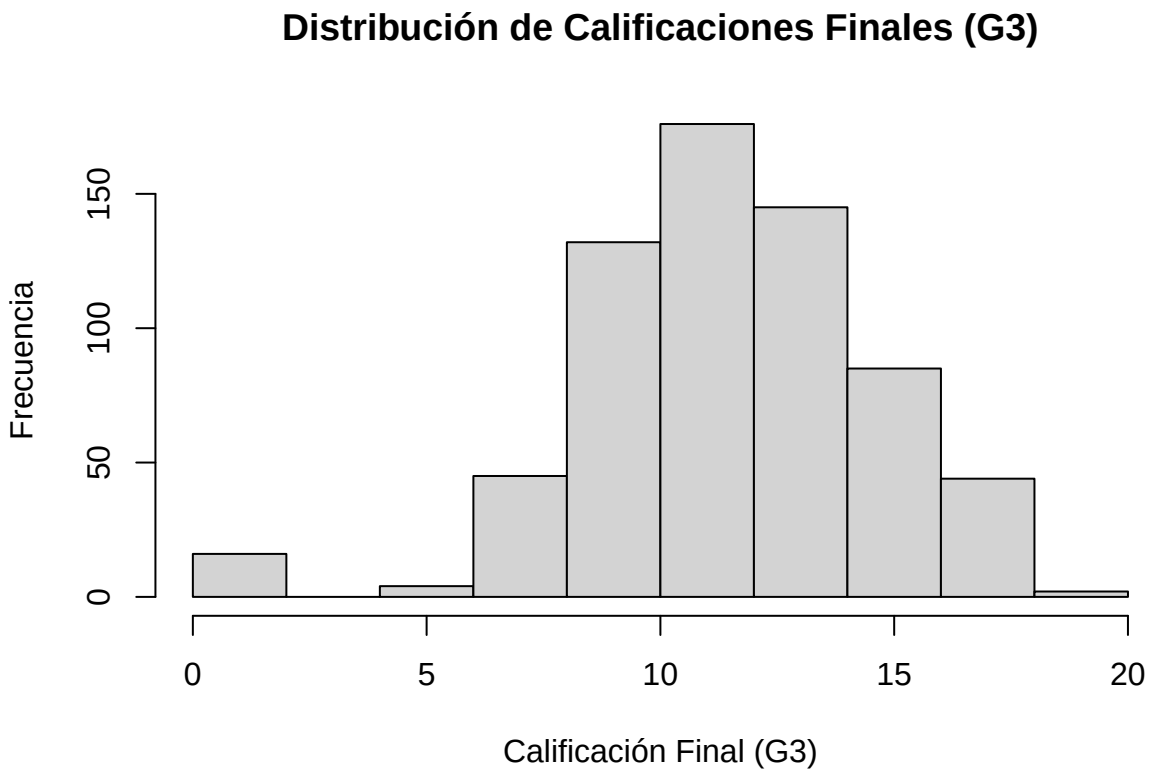
```
##
##  Anderson-Darling normality test
##
## data:  datos$G3
## A = 8.2336, p-value < 2.2e-16
```

A = 5.7993, p-value = 2.744e-14

Segun la prueba de Anderson-Darling podemos decir que la variable G3 no sigue una distribución normal dado que el p-value es menor a 0.05 y tiene una gran diferencia con el valor de prueba del estadistico.

Eliminar valores atípicos

```
#datos <- datos[datos$G3 != valor_atipico1]
#datos <- datos[datos$G3 != valor_atipico2]

hist(datos$G3, main = "Distribución de Calificaciones Finales (G3)", xlab = "Calificación Final (G3)", y
```



Homocedasticidad

Prueba de Levene para G3

```
if (!require(car)) {
  install.packages("car")
  library(car)
}
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```r
# Realizar la prueba de Levene para G3 vs. school
levene_test <- leveneTest(datos$G3 ~ datos$school) # Aqui podria ser cualquier variable categórica de i
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```r
cat("Resultados de la Prueba de Levene para G3 vs. school:\n")
```

```
## Resultados de la Prueba de Levene para G3 vs. school:
```

```r
levene_test  # No es necesario usar print aquí
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  12.706 0.0003913 ***
##       647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n")
```

```r
if (levene_test$`Pr(>F)`[1] < 0.05) {
  cat("La Prueba de Levene indica heterocedasticidad (p-value <", levene_test$`Pr(>F)`[1], ")\n") # nol
} else {
  cat("La Prueba de Levene no indica heterocedasticidad (p-value =", levene_test$`Pr(>F)`[1], ")\n") #
}
```

```
## La Prueba de Levene indica heterocedasticidad (p-value < 0.0003912727 )
```

```r
# Realizar ANOVA para G3 sin transformar
formula_anova_G3 <- as.formula("G3 ~ 1")  # nolint
resultado_anova_G3 <- aov(formula_anova_G3, data = datos) # nolint

cat("Resultados ANOVA para G3")
```

```
## Resultados ANOVA para G3
```

```r
summary(resultado_anova_G3)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Residuals   648   6763   10.44
```

```
        <!-- Df Sum Sq Mean Sq F value Pr(>F)
```

Residuals 615 3849 6.258 –>

Segun los resultados de la prueba de Levene, se observa que el valor p es mayor a 0.05, lo que sugiere que no hay evidencia estadística para rechazar la hipótesis nula de homogeneidad de varianzas. Por lo tanto, podemos concluir que no hay heterocedasticidad significativa en la variable G3 en función de la variable categórica school.

ANOVA

```r
resultados_anova <- list()

# Realizar un bucle para realizar ANOVAs para cada variable independiente
for (variable in variables_independientes) {
  formula_anova <- as.formula(paste("G3 ~", variable))
  resultado_anova <- aov(formula_anova, data = datos)
  resultados_anova[[variable]] <- summary(resultado_anova)
}

# Ver los resultados de los ANOVAs
for (variable in variables_independientes) {
  cat("Resultados ANOVA para", variable, ":\n")
  print(resultados_anova[[variable]])
  cat("\n")
}
```

```
## Resultados ANOVA para G1 :
##              Df Sum Sq Mean Sq F value Pr(>F)
## G1            1   4619    4619    1393 <2e-16 ***
## Residuals   647   2145       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para G2 :
##              Df Sum Sq Mean Sq F value Pr(>F)
## G2            1   5706    5706    3493 <2e-16 ***
## Residuals   647   1057       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para absences :
##              Df Sum Sq Mean Sq F value Pr(>F)
## absences      1     56   56.47   5.448 0.0199 *
## Residuals   647   6707   10.37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para failures :
##              Df Sum Sq Mean Sq F value Pr(>F)
## failures      1   1046  1046.3   118.4 <2e-16 ***
## Residuals   647   5717     8.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para studytime :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## studytime     1    422   422.0   43.06 1.09e-10 ***
## Residuals   647   6341     9.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Resultados ANOVA para freetime :
##              Df Sum Sq Mean Sq F value  Pr(>F)
## freetime      1    102   101.8    9.89 0.00174 **
## Residuals   647   6661    10.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para goout :
##              Df Sum Sq Mean Sq F value Pr(>F)
## goout         1     52   51.95   5.008 0.0256 *
## Residuals   647   6711   10.37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para Dalc :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Dalc          1    283  283.45    28.3 1.43e-07 ***
## Residuals   647   6480   10.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para Walc :
##              Df Sum Sq Mean Sq F value Pr(>F)
## Walc          1    211  210.97   20.83  6e-06 ***
## Residuals   647   6552   10.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para health :
##              Df Sum Sq Mean Sq F value Pr(>F)
## health        1     66   66.09   6.385 0.0117 *
## Residuals   647   6697   10.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Resultados ANOVA para age :
##              Df Sum Sq Mean Sq F value  Pr(>F)
## age           1     77   76.72   7.423 0.00661 **
## Residuals   647   6687   10.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se realizo un analisis de varianza ANOVA para determinar si existe una diferencia significativa de la variable dependiente G3 en funcion de cada una de las variables independientes. Donde se podra utilizar para comprender si las variables independientes son estadísticamente significativas para explicar las diferencias en las calificaciones finales (G3) en el modelo.

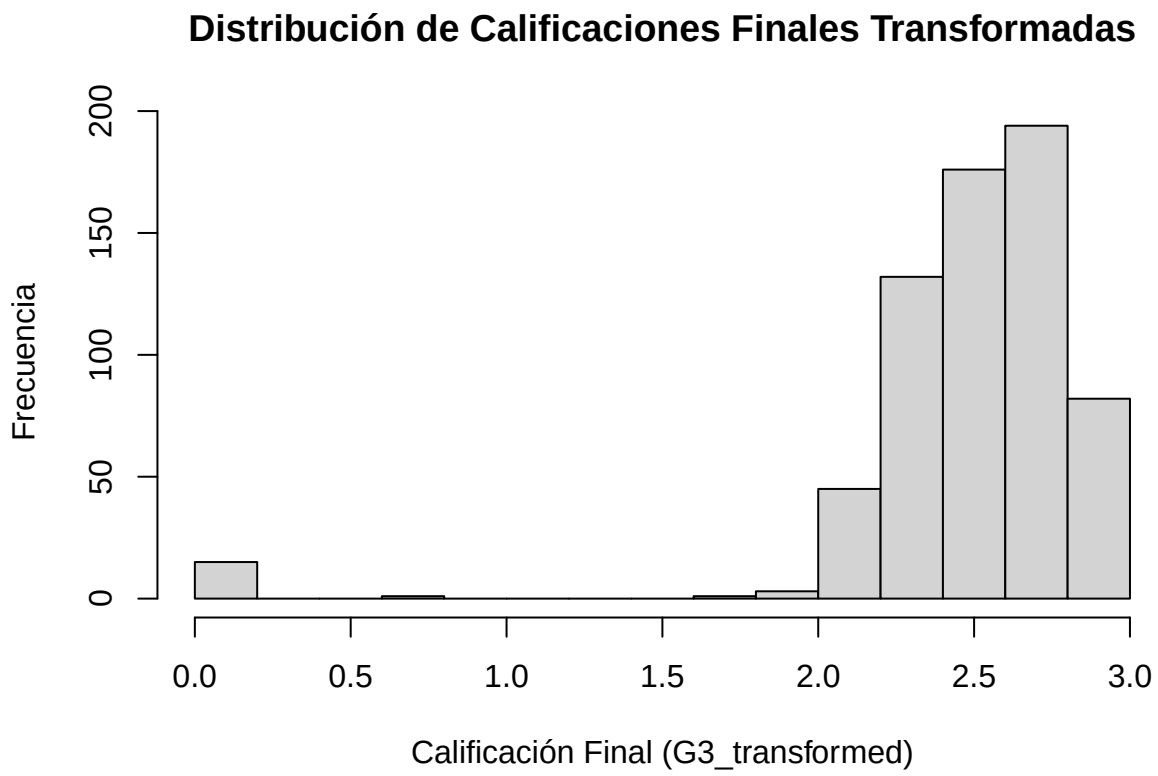
Según los resultados de los ANOVAs, se observa que las siguientes variables tienen un valor p menor a 0.05, lo que indica que son significativas para explicar las diferencias en las calificaciones finales (G3) en el modelo: G1, G2, absences, failures, studytime, goout, Dalc, Walc, health y age

Específicamente, las siguientes variables son las que tienen p-valores significativamente bajos y cercanos a 0.05
- 'absences' (p-valor = 0.0199) - 'freetime' (p-valor = 0.00174) - 'goout' (p-valor = 0.0256) - 'health' (p-valor = 0.0117) - 'age' (p-valor = 0.00661)

Por lo tanto, podemos concluir que estas variables son estadísticamente significativas para explicar las diferencias en las calificaciones finales (G3) en el modelo, mientras que las demás no aportan evidencia estadística significativa.

Transformación logaritmica de la variable G3 para normalizarla

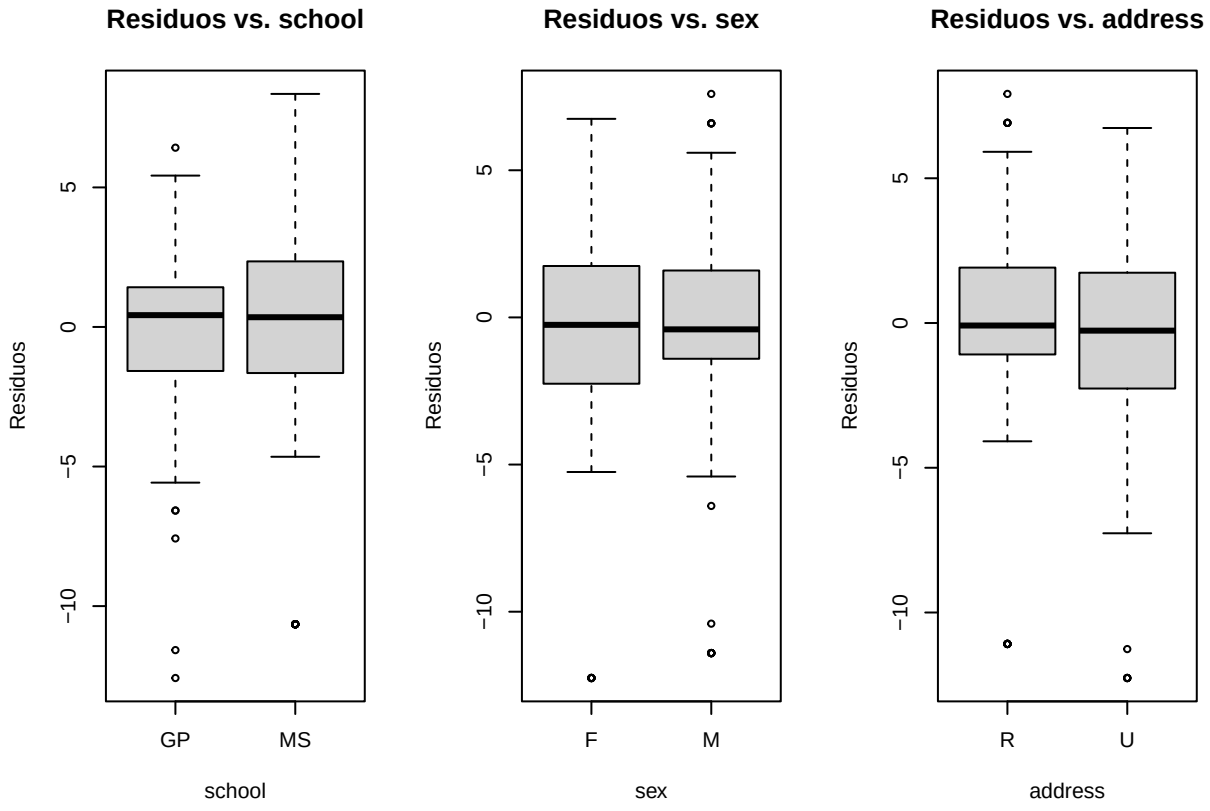
```r
constante <- 1  # Puedes ajustar esta constante según sea necesario
datos$G3_transformed <- log(datos$G3 + constante)

hist(datos$G3_transformed, main = "Distribución de Calificaciones Finales Transformadas", xlab = "Calif
```



Boxplot de residuos vs. variables categóricas

```r
# Ajustar el modelo lineal para G3 vs. school
modelo_school <- lm(G3 ~ school, data = datos)
residuos_school <- resid(modelo_school)

# Ajustar el modelo lineal para G3 vs. sex
modelo_sex <- lm(G3 ~ sex, data = datos)
residuos_sex <- resid(modelo_sex)

# Ajustar el modelo lineal para G3 vs. address
```

```r
modelo_address <- lm(G3 ~ address, data = datos)
residuos_address <- resid(modelo_address)

# Boxplot de residuos vs. variables categóricas
par(mfrow = c(1, 3))

boxplot(residuos_school ~ datos$school, main = "Residuos vs. school", xlab = "school", ylab = "Residuos
boxplot(residuos_sex ~ datos$sex, main = "Residuos vs. sex", xlab = "sex", ylab = "Residuos") # nolint
boxplot(residuos_address ~ datos$address, main = "Residuos vs. address", xlab = "address", ylab = "Resi
```



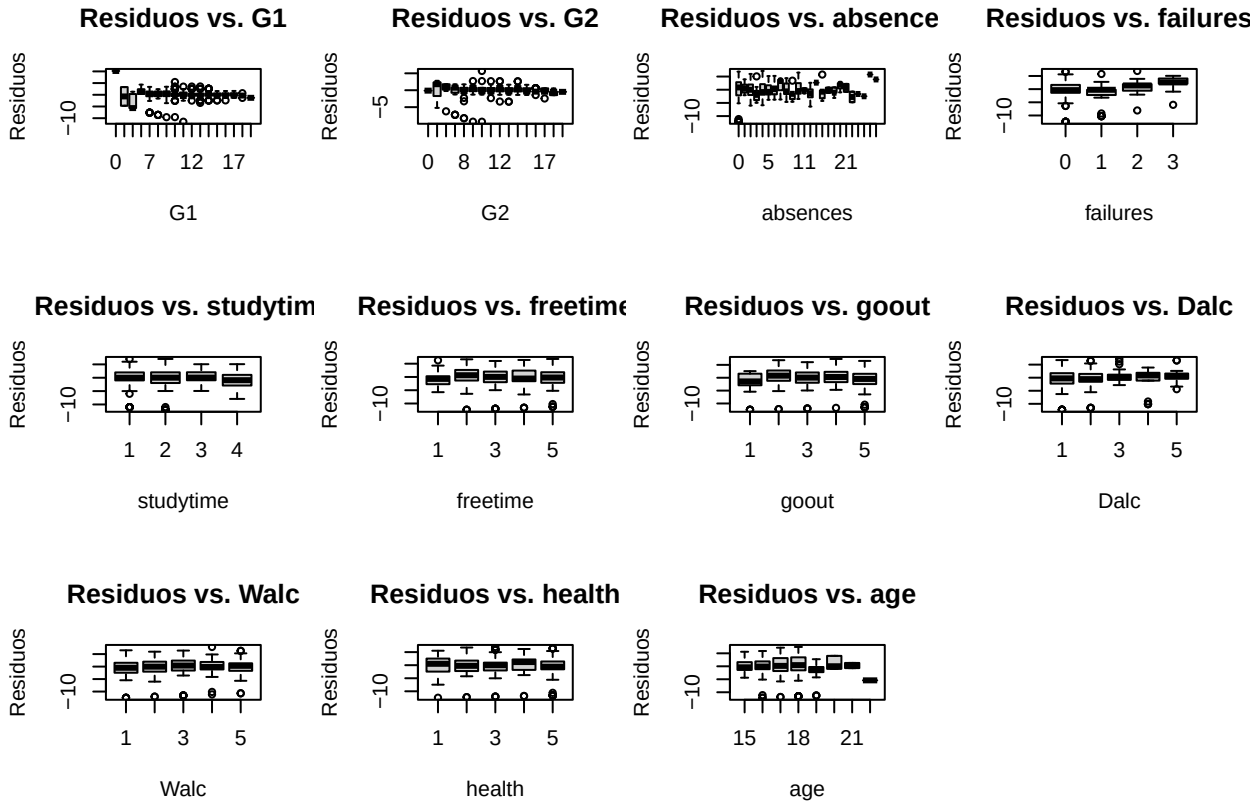

```r
par(mfrow = c(1, 1))  # Restaurar la disposición de gráficos
```

Al apreciar que los residuos tienen una dispersión similar se podria cumplir el supuesto de homocedasticidad. Pera para estas variables variables_categoricas

Boxplot de residuos vs. variables independientes

```r
residuos_independientes <- list()

# Ajustar modelos lineales para G3 vs. cada variable independiente y almacenar los residuos # nolint
for (variable in variables_independientes) {
  modelo <- lm(paste("G3 ~", variable), data = datos)
  residuos_independientes[[variable]] <- resid(modelo)
```

```r
}

# Crear boxplots de residuos vs. variables independientes
par(mfrow = c(3, 4)) # Esto crea una matriz de gráficos para todas las variables independientes # nolin

for (i in seq_along(variables_independientes)) {
  boxplot(residuos_independientes[[variables_independientes[i]]] ~ datos[[variables_independientes[i]]]
          main = paste("Residuos vs.", variables_independientes[i]),
          xlab = variables_independientes[i], ylab = "Residuos")
}

par(mfrow = c(1, 1)) # Restaurar la disposición de gráficos
```

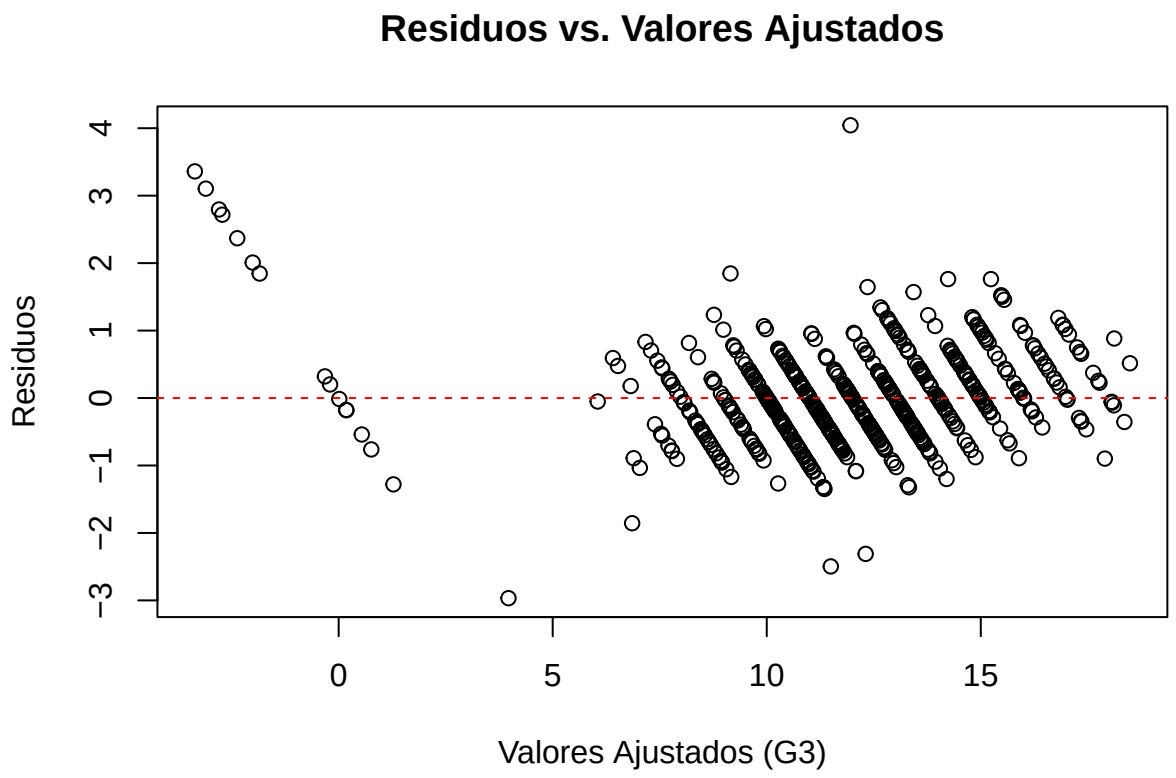


Tambien se puede apreciar que los residuos tienen una dispersión similar para las variables independientes, por lo que se podria cumplir el supuesto de homocedasticidad. pero en G1, G2 y absences se puede apreciar que los residuos tienen una dispersión diferente, por lo que no se cumple el supuesto de homocedasticidad.

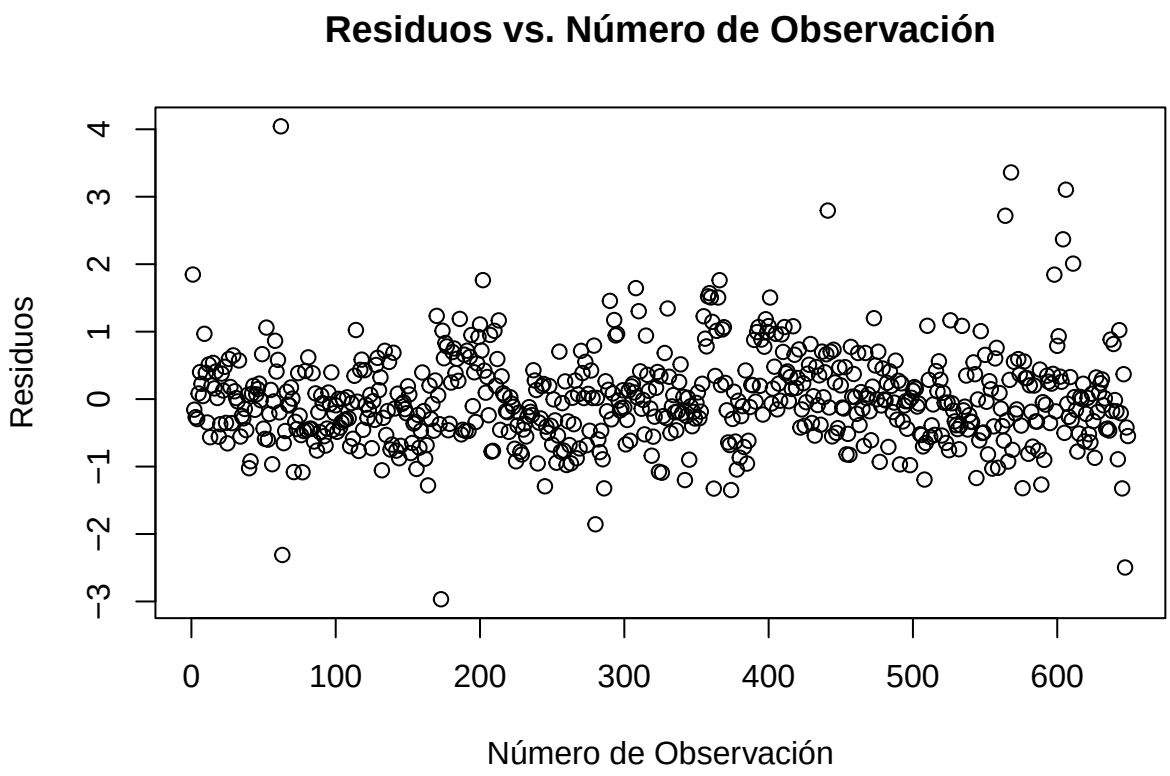
Independencia de observaciones

```r
# Ajustar el modelo lineal para G3
modelo_G3 <- lm(G3 ~ ., data = datos)

# Obtener los residuos y los valores ajustados
residuos_G3 <- resid(modelo_G3)
```

```r
valores_ajustados_G3 <- fitted(modelo_G3)

# Crear un gráfico de residuos vs. valores ajustados
plot(valores_ajustados_G3, residuos_G3, main = "Residuos vs. Valores Ajustados",
     xlab = "Valores Ajustados (G3)", ylab = "Residuos")

# Para agregar líneas horizontales
abline(h = 0, col = "red", lty = 2)
```




```r
plot(seq_along(datos$G3), residuos_G3, main = "Residuos vs. Número de Observación",
     xlab = "Número de Observación", ylab = "Residuos")
```



Con esto se puede apreciar que los residuos no tienen una distribución normal, por lo que no se cumple el supuesto de independencia de observaciones.

```
residuos <- residuals(modelo_G3)
valores_ajustados <- fitted(modelo_G3)

residuos_cuadrados <- residuos^2
modelo_breusch_pagan <- lm(residuos_cuadrados ~ valores_ajustados)

breusch_pagan_statistic <- summary(modelo_breusch_pagan)$fstatistic[1]
breusch_pagan_p_value <- 1 - pf(breusch_pagan_statistic, 1, length(residuos) - 2) # nolint

cat("Estadístico de prueba de Breusch-Pagan:", breusch_pagan_statistic, "\n")
```

```
## Estadístico de prueba de Breusch-Pagan: 63.13153
```

```
cat("Valor p del test de Breusch-Pagan:", breusch_pagan_p_value, "\n")
```

```
## Valor p del test de Breusch-Pagan: 8.65974e-15
```

Ivestigacion sobre ANOVA

Anova es un método estadístico que se utiliza para probar las diferencias entre dos o más medias. ANOVA se utiliza para probar la hipótesis nula de que las medias de dos o más grupos son iguales. ANOVA se utiliza en estadísticas, genética, ciencias de la conducta y otras áreas. Un ejemplo de ANOVA es el siguiente:

Para este caso de 3 calificaciones de 3 escuelas diferentes, se quiere saber si existe una diferencia significativa entre las calificaciones de las 3 escuelas. Para esto se realiza un ANOVA para determinar si existe una diferencia significativa de la variable dependiente G3 en funcion de cada una de las variables independientes. Donde se podra utilizar para comprender si las variables independientes son estadísticamente significativas para explicar las diferencias en las calificaciones finales (G3) en el modelo.

Conclusiones

- Se verifico si la variable G3 sigue una distribución normal, para esto se realizo un histograma y un boxplot, donde se observo que la variable G3 no sigue una distribución normal.
- Por otra parte al revisar los resultados del test de normalidad de Anderson se pudo comprobar que las variables de interes no distribuyen con normalidad pues, se tiene que todos los valores de p, para Age, Medu, G1,G2,G3 son muy pequeños en comparacion al valores del estadistico de prueba que son muy elevados en proporción con lo cual se puede concluir que los datos no siguen una distribución normal.
- Se realizo un analisis de varianza ANOVA para determinar si existe una diferencia significativa de la variable dependiente G3 en funcion de cada una de las variables independientes.
- Segun los resultados de los ANOVAs, se observa que las siguientes variables tienen un valor p menor a 0.05, lo que indica que son significativas para explicar las diferencias en las calificaciones finales (G3) en el modelo:
  - 'absences' (p-valor = 0.0199)
  - 'freetime' (p-valor = 0.00174)
  - 'goout' (p-valor = 0.0256)
  - 'health' (p-valor = 0.0117)
  - 'age' (p-valor = 0.00661)
- Con respecto de la homocedasticidad o heterocedasticidad, la varianza de los errores no es constante y varia de manera sistemática, pero al aplicar el test de normalidad de Breusch-Pagan se obtuvo que si se cumple el criterio de heterocedasticidad con un p-value = 0.002785 menor que 0.05 En otras palabras, la probabilidad de que la heterocedasticidad sea un hallazgo al azar es muy baja, lo que sugiere que es una característica real de los datos. Donde un valor alto del estadístico de prueba indica que es más probable que haya heterocedasticidad en los residuos.