Actividad 2.3 Regresión lineal Múltiple

Franco Mendoza Muraira A01383399

2023-11-07

```
library(plot3D)
```

```
## Warning: package 'plot3D' was built under R version 4.2.3
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

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

```
## Warning: package 'zoo' was built under R version 4.2.3
```

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

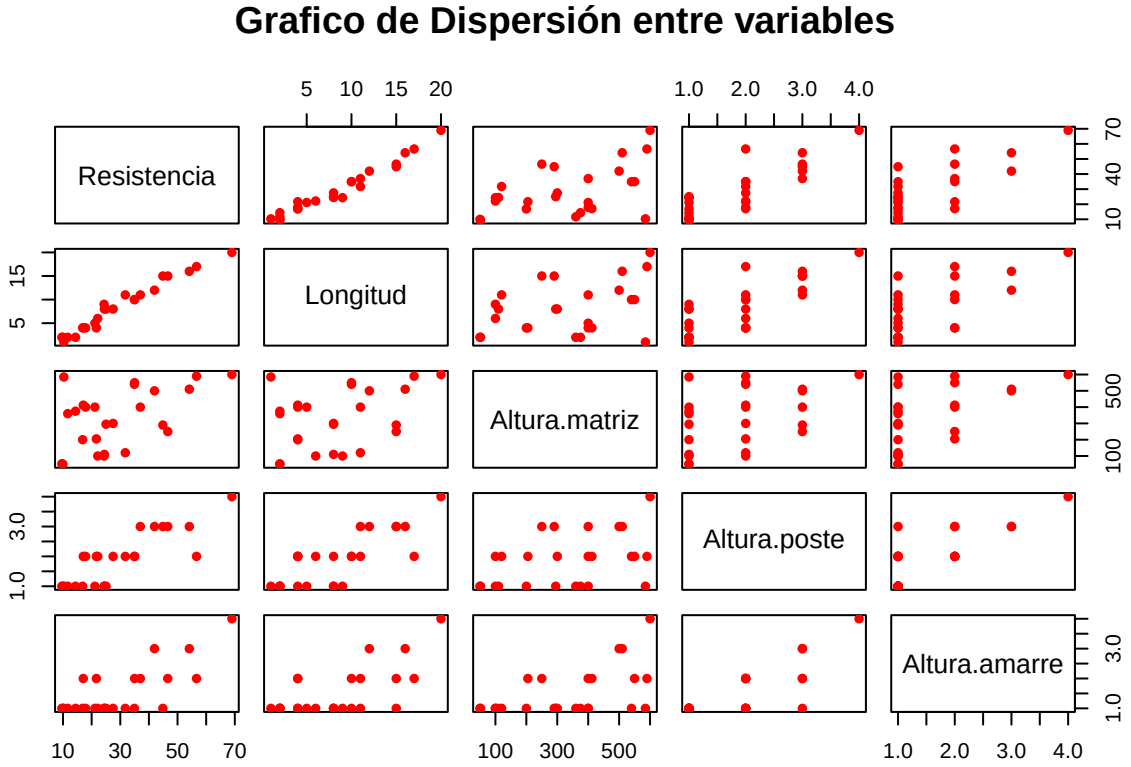
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

1. Análisis Exploratorio

```
##   Resistencia Longitud Altura.matriz Altura.poste Altura.amarre
## 1        9.95        2            50            1             1
## 2       24.45        8           110            1             1
## 3       31.75       11           120            2             1
```

Gráficos de dispersión entre variables

```
pairs(df, main = "Grafico de Dispersión entre variables",pch=16,col="red")
```



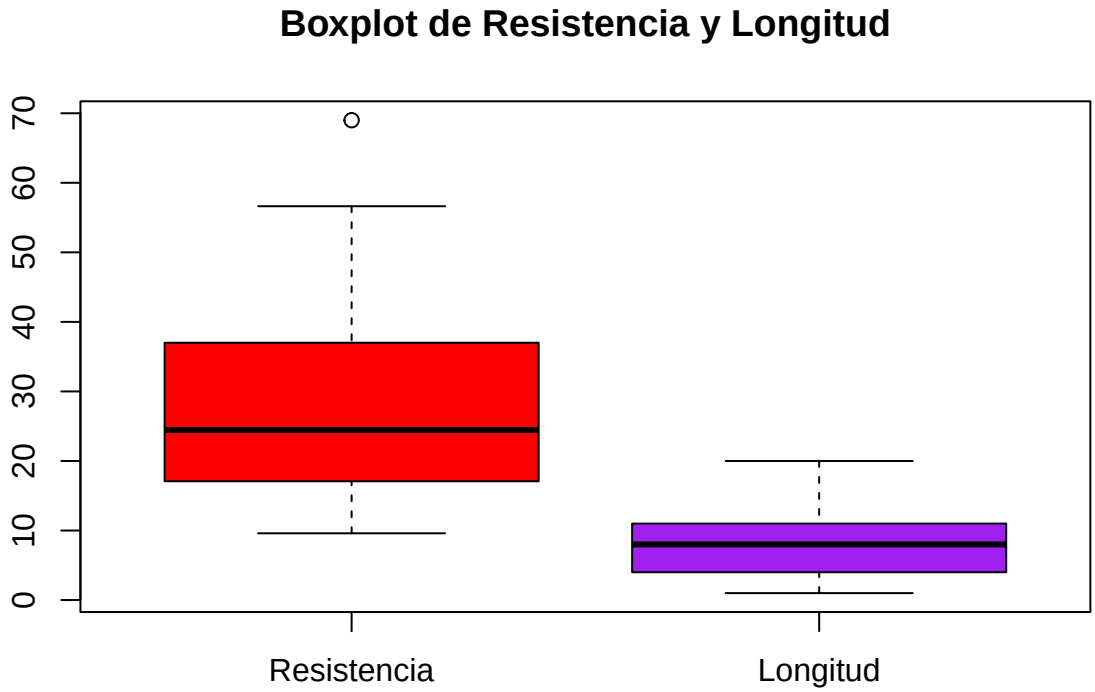
Matriz de correlación de los datos

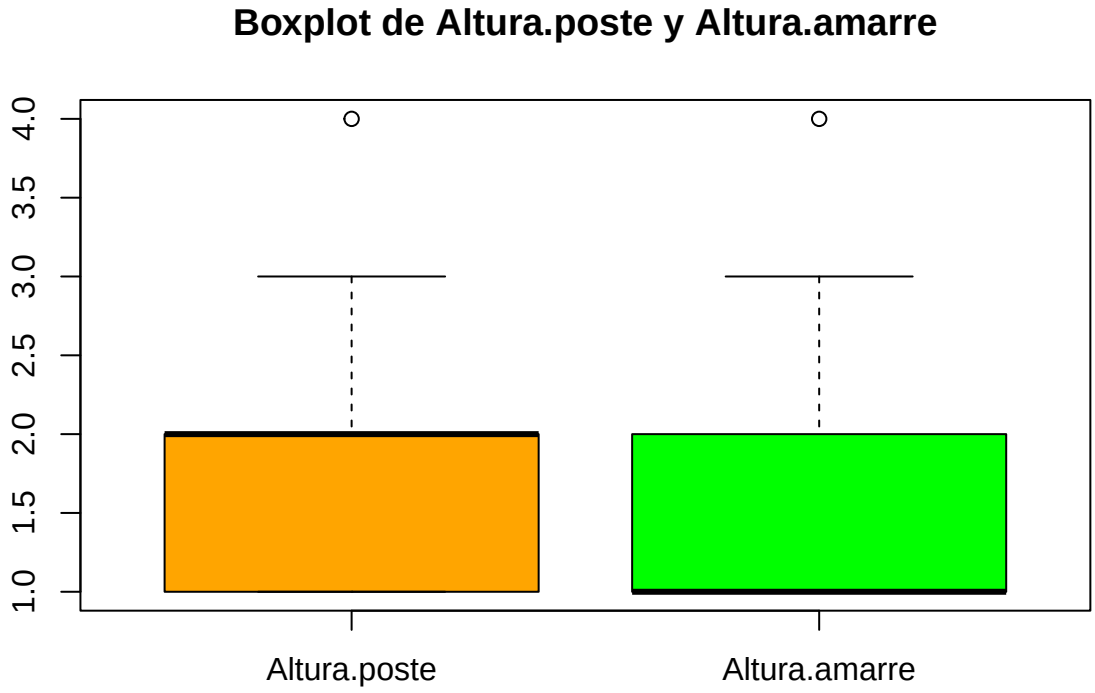
```
cor(df)
```

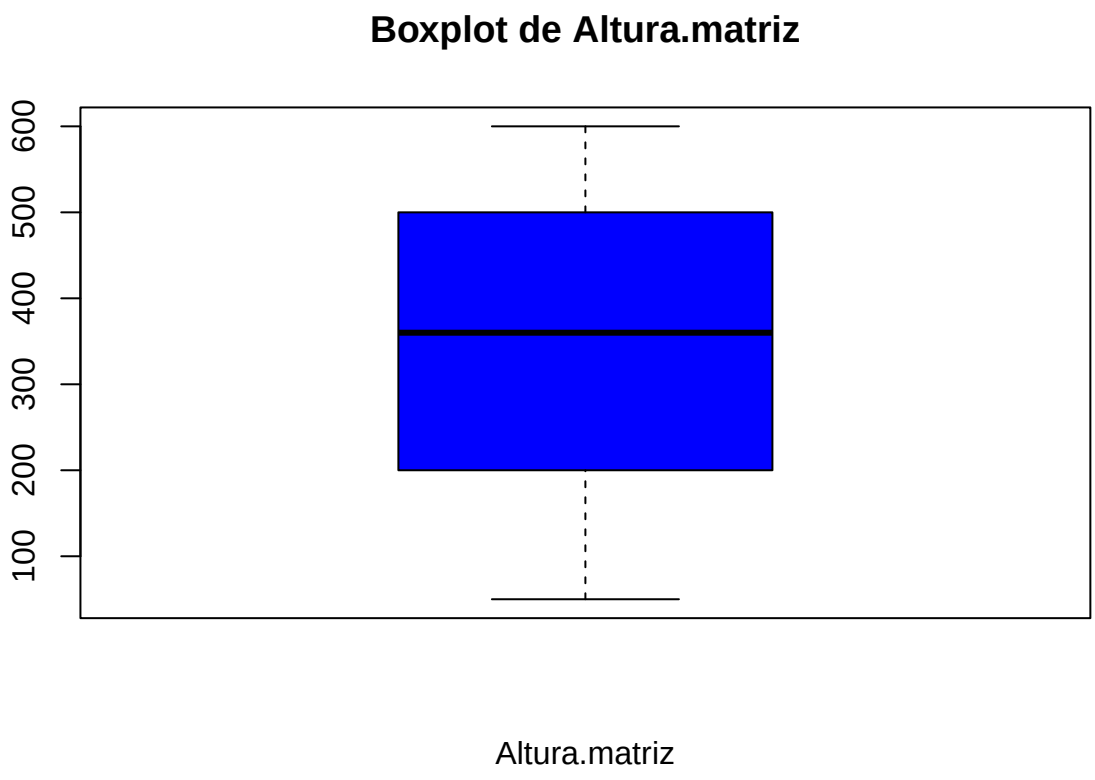
```
##                Resistencia  Longitud Altura.matriz Altura.poste Altura.amarre
## Resistencia      1.0000000 0.9818118     0.4928666    0.8356493     0.7483815
## Longitud         0.9818118 1.0000000     0.3784127    0.7950203     0.6560819
## Altura.matriz    0.4928666 0.3784127     1.0000000    0.4243451     0.5377305
## Altura.poste     0.8356493 0.7950203     0.4243451    1.0000000     0.7793701
## Altura.amarre    0.7483815 0.6560819     0.5377305    0.7793701     1.0000000
```

Podemos ver que la variable con mayor correlación a la variable a predecir (Resistencia) es la de Longitud, debido a esto la usaremos para nuestro modelo, y para no hacer uso de variables colineales, usaremos otra variable la cuál no tenfa una alta correlación a Longitud, escogiendo 2 variables diferentes para poder evaluar y crear un buen modelo, por esto escogeremos la variable Altura.matriz como nuestra segunda variable predictora.

Boxplots


Boxplot de Resistencia y Longitud
0
10
20
30
40
50
60
70
Resistencia
Longitud


Boxplot de Altura.poste y Altura.amarre
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Altura.poste
Altura.amarre



2. Método de mínimos cuadrados

Se va a predecir la variable dependiente de Resistencia, y para las variables independientes se usarán Longitud y Altura.matriz, esto debido a la alta correlación entre Resistencia y Longitud, y la falta de colinealidad entre Altura.matriz y Longitud, por lo que se podría hacer un buen modelo.

Betas

```
X = cbind(1,as.matrix(x))
Y = as.matrix(y)
betas = solve((t(X)%*%X))%*%t(X)%*%as.matrix(Y)
betas
```

```
##                Resistencia
##                 2.26379143
## Longitud        2.74426964
## Altura.matriz   0.01252781
```

Ecuación del modelo de regresión múltiple

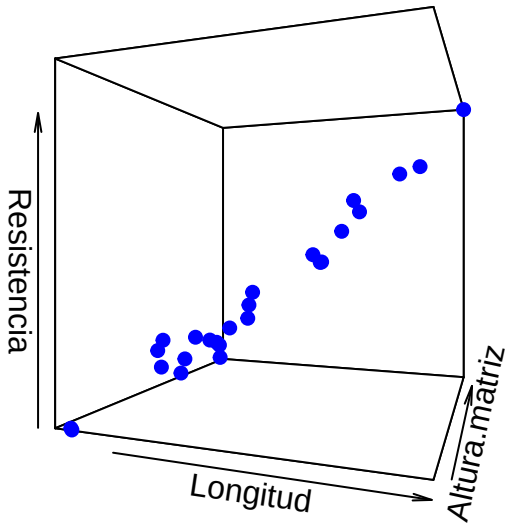
$Y = 2.2638 + 2.7443x_1 + 0.0125x_2$

3. Regresión lineal múltiple en R

```
model =lm(Y~x1+x2)
model
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##     2.26379      2.74427      0.01253
```

```
scatter3D(x1,x2,Y, col="blue", cex = 0.9 , pch=19, xlab = "Longitud" ,ylab= "Altura.matriz" ,zlab="Resis
```



Se puede ver que la fórmula del método de mínimos cuadrados nos da el mismo resultado que la función lm que nos proporciona R, y podemos ver como actúa el modelo y como se evalúa la resistencia a partir de las 2 variables predictoras en la gráfica anterior.

4. Evaluación del modelo

Nivel de Significancia: $\alpha = 0.05$

Colinealidad de las variables involucradas

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

```
cor.test(df$Longitud,df$Altura.matriz)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$Longitud and df$Altura.matriz
## t = 1.9606, df = 23, p-value = 0.06215
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01965744  0.67292598
## sample estimates:
##       cor
## 0.3784127
```

Debido a que el pvalue sobrepasa al nivel de significancia de 0.05 establecido anteriormente, no se rechaza la hipótesis nula de la colinealidad, y se concluye que no hay colinealidad entre las variables involucradas en el modelo.

Variabilidad explicada por el modelo

```
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.865 -1.542 -0.362  1.196  5.841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.263791   1.060066   2.136 0.044099 *
## x1          2.744270   0.093524  29.343  < 2e-16 ***
## x2          0.012528   0.002798   4.477 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.288 on 22 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9794
## F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

Esta función nos muestra un coeficiente de determinación alto de 0.9794, lo que nos dice que las variables usadas x_1 y x_2 explican al 97.94% la varaibilidad de nuestra variable dependiente y, la resistencia, lo cuál nos dice que el modelo realizado es bueno.

Significancia de beta_i

Todos los valores p de las β_i son menores al α de 0.05, por lo que se concluye que todos las betas son significantes, esto es bueno para nuestro modelo, ya que sabemos que las 2 variables son buenas para nuestro modelo, y se alejan lo suficiente del 0 para influir en él.

Economía del modelo y variabilidad explicada

Debido a que no se usan muchas variables, solo 2, podemos concluir que la economía del modelo es buena, la diferencia del r^2 con mas variables no supera por mucho la de este modelo por lo que no es necesario agregarlas.

5. Validación del modelo

Normalidad de los residuos

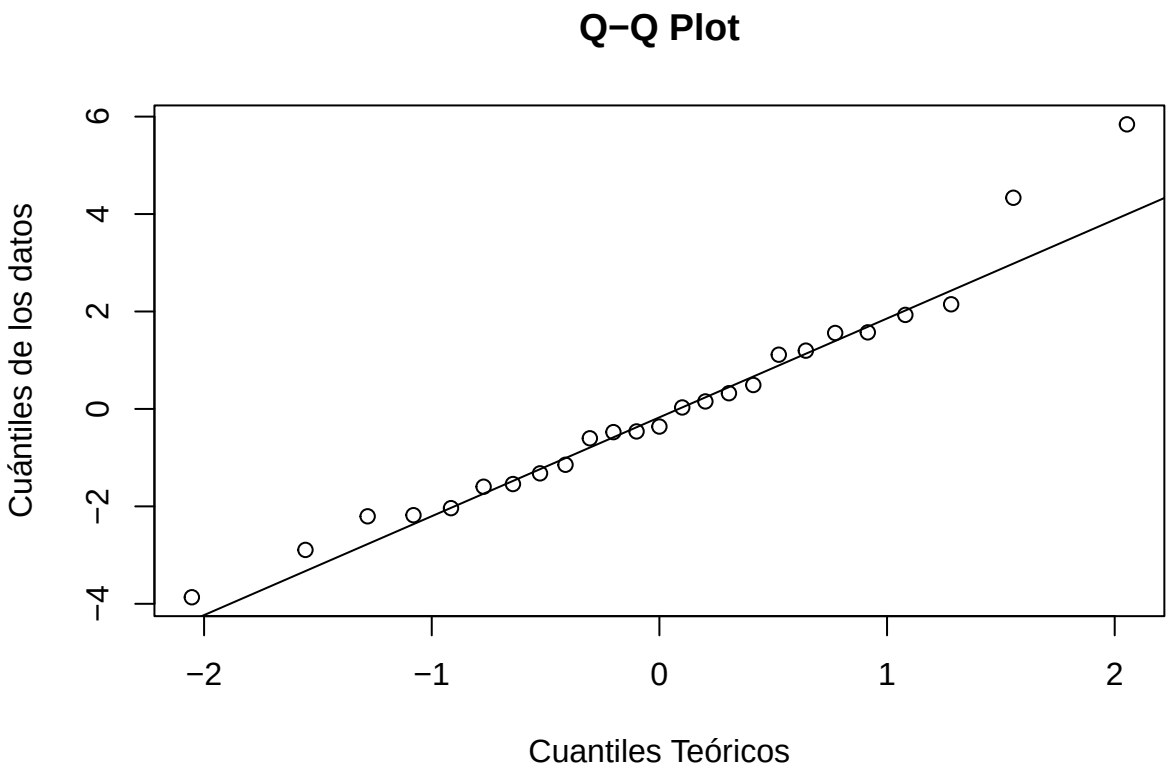
H_0 : Los residuos provienen de una distribución normal.

H_1 : Los residuos no provienen de una distribución normal.

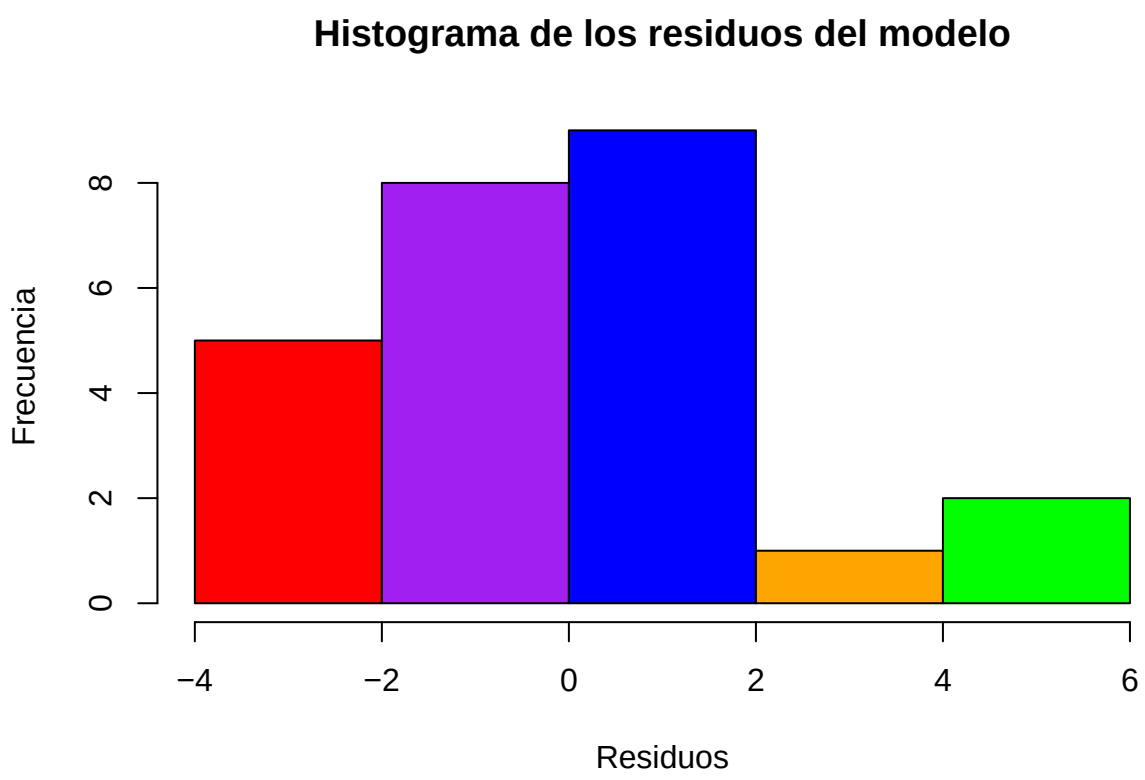
```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.95827, p-value = 0.381
```

```
qqnorm(model$residuals,main = "Q-Q Plot",xlab = "Cuantiles Teóricos",ylab= "Cuántiles de los datos")
qqline(model$residuals)
```




```
hist(model$residuals,main = "Histograma de los residuos del modelo",xlab = "Residuos",ylab = "Frecuencia
```



Viendo que el pvalue de la prueba de Shapiro es de 0.381, y siendo mayor que el α de 0.05, no podemos rechazar la hipótesis nula, y se concluye que se presenta una distribución normal en los resiudos del modelo.

Verificación de media cero

H_0 : Media $= 0$.

H_1 : Media $\neq 0$.

```
t.test(residuals(model))
```

```
##
##  One Sample t-test
##
## data:  residuals(model)
## t = 9.3671e-17, df = 24, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.9042509  0.9042509
## sample estimates:
##    mean of x
## 4.103976e-17
```

Debido a el pvalue de 1, no podemos rechazar la hipótesis nula, y concluimos que la media de los residuos si puede ser igual a 0.

Homocedasticidad

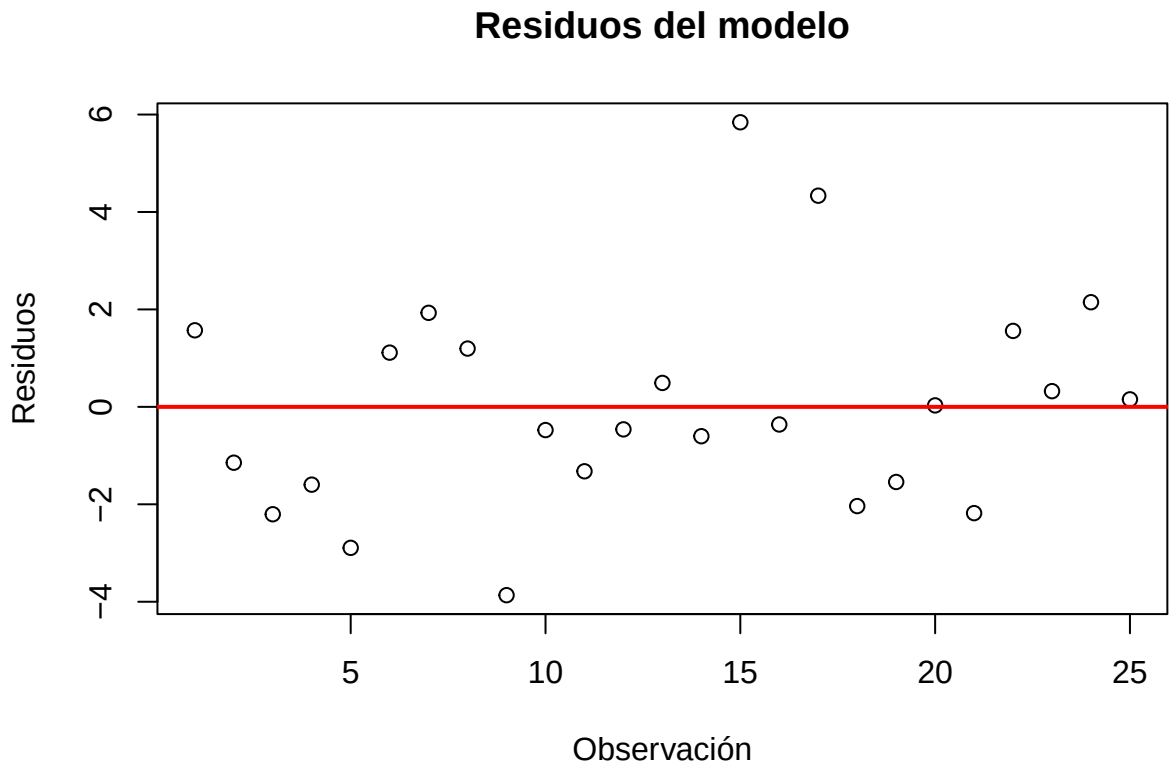
H_0 : La varianza de los errores es constante.

H_1 : La varianza de los errores no es constante.

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.66721, df = 2, p-value = 0.7163
```

```
plot(residuals(model),main = "Residuos del modelo", xlab = "Observación",ylab= "Residuos")
abline(h=0,col="red",lwd=2)
```



Debido a que el pvalue es muy alto, con un valor de 0.7163, no se puede rechazar la hipótesis nula, y se concluye que la varianza de los errores es constante, esto se puede ver en la gráfica superior.

Independencia

H_0 : Los residuos son independientes.

H_1 : Los residuos no son independientes.

```r
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.0972, p-value = 0.559
## alternative hypothesis: true autocorrelation is greater than 0
```

Debido a que el p value es alto, con un valor de 0.559, y mayor de 0.05, no hay suficiente evidencia para rechazar la hipótesis nula, por lo que se concluye que los residuos son independientes.

6. Conclusiones

En el análisis del modelo se pudieron ver resultados positivos en cuanto a sus supuestos, ya que quedaron satisfechos todos estos. En el trabajo se pudo hacer un análisis detallado de la regresión lineal múltiple usando las variables de Longitud y Altura.matriz.

El modelo de regresión lineal múltiple es sólido, ya que se ajusta bien a los datos y cumple con los supuestos fundamentales de la regresión lineal, lo que lo hace confiable para predecir la resistencia basada en las variables Longitud y Altura.matriz.