Actividad 2.3 Regresión lineal Múltiple

Raúl Correa Ocañas - A01722401

2024-08-16

Analiza la base de datosRes.csv Download datosRes.csv en donde se describen los datos recolectados en experimento realizado para estudiar la relación de la resistencia al desprendimiento de un alambre adherido (una medida de la cantidad de fuerza que se requiere para romper la unión) con algunas de las variables en un proceso de manufactura de semiconductores. Encuentra el mejor modelo de regresión múltiple que explique la variable dependiente.

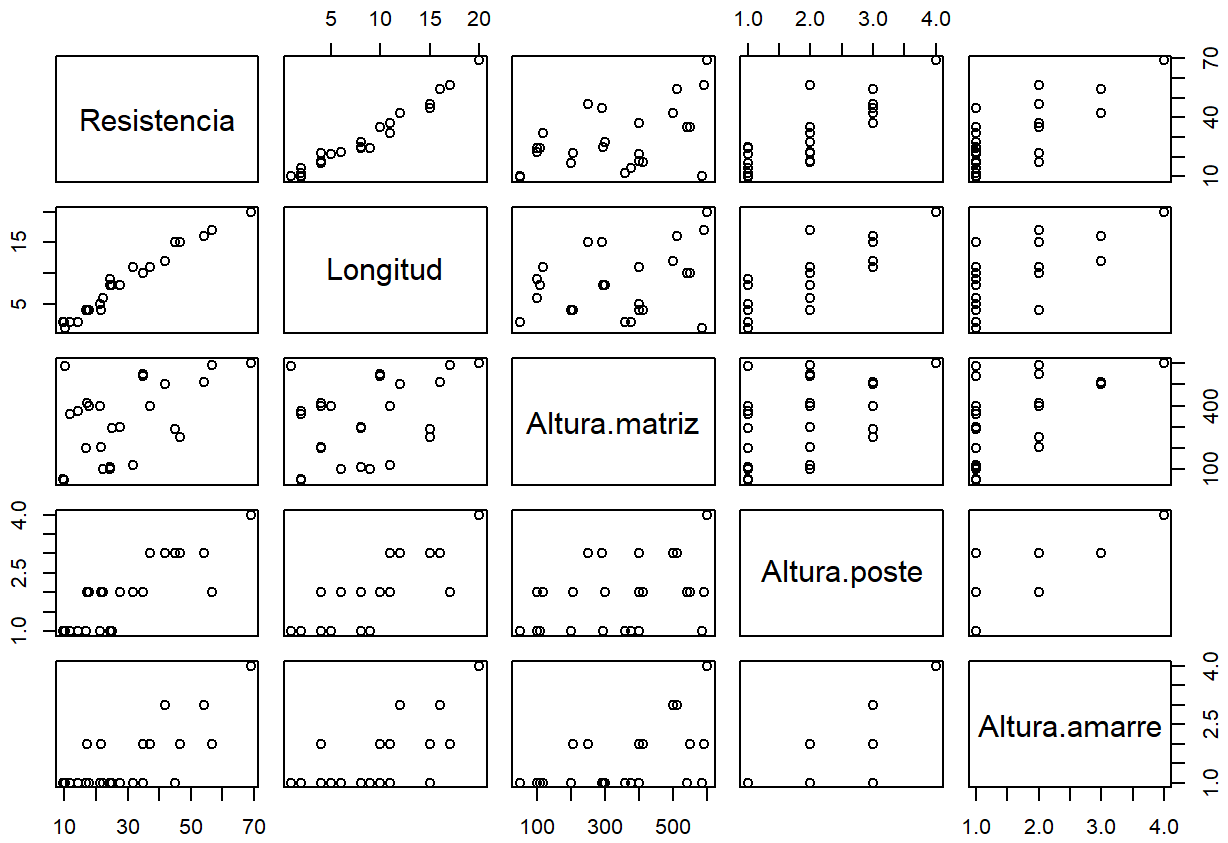
0. Importación de Datos

```
data_path = r"{../data/datosRes.csv}"
df = read.csv(data_path)
df
```

```
##    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
## 4        35.00       10           550            2             2
## 5        25.02        8           295            1             1
## 6        16.86        4           200            1             1
## 7        14.38        2           375            1             1
## 8         9.60        2            52            1             1
## 9        24.35        9           100            1             1
## 10       27.50        8           300            2             1
## 11       17.08        4           412            2             2
## 12       37.00       11           400            3             2
## 13       41.95       12           500            3             3
## 14       11.66        2           360            1             1
## 15       21.65        4           205            2             2
## 16       17.89        4           400            2             1
## 17       69.00       20           600            4             4
## 18       10.30        1           585            1             1
## 19       34.93       10           540            2             1
## 20       46.59       15           250            3             2
## 21       44.88       15           290            3             1
## 22       54.12       16           510            3             3
## 23       56.63       17           590            2             2
## 24       22.13        6           100            2             1
## 25       21.15        5           400            1             1
```

1. Análisis Exploratorio de Datos

```
pairs(df)
```

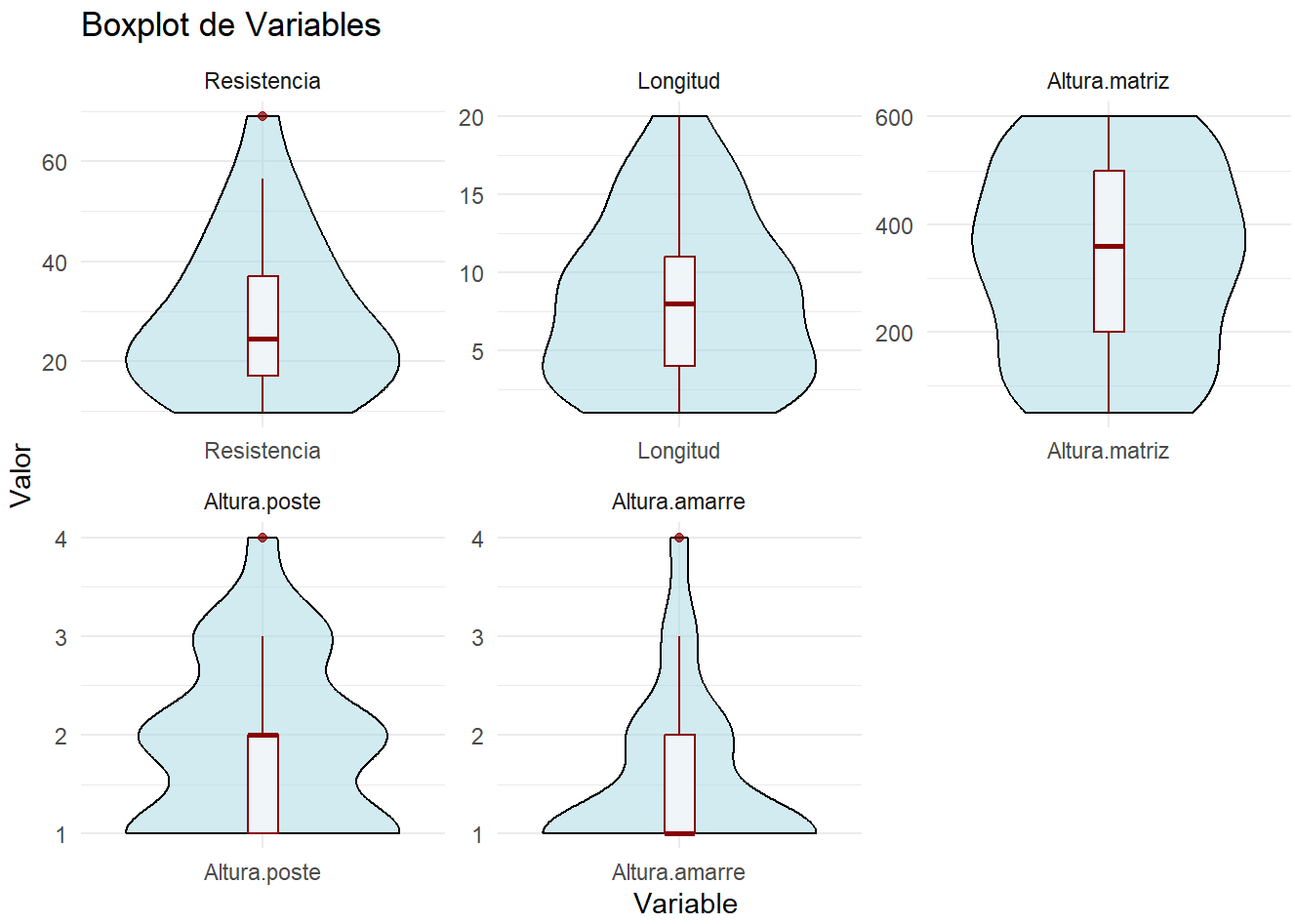



```
cor_matrix <- cor(df)
cor_melted <- melt(cor_matrix)
cor_melted_filtered <- cor_melted[upper.tri(cor_matrix, diag = FALSE), ]
cor_melted_filtered[order(-cor_melted_filtered$value), ]
```

```
##             Var1          Var2     value
## 6    Resistencia      Longitud 0.9818118
## 16   Resistencia  Altura.poste 0.8356493
## 17      Longitud  Altura.poste 0.7950203
## 24  Altura.poste Altura.amarre 0.7793701
## 21   Resistencia Altura.amarre 0.7483815
## 22      Longitud Altura.amarre 0.6560819
## 23 Altura.matriz Altura.amarre 0.5377305
## 11   Resistencia Altura.matriz 0.4928666
## 18 Altura.matriz  Altura.poste 0.4243451
## 12      Longitud Altura.matriz 0.3784127
```

```
ggplot(melt(df), aes(x=variable, y=value)) +
  geom_violin(fill = "lightblue", color = "black", alpha = 0.5) +
  geom_boxplot(width = 0.1, color = "darkred", alpha = 0.7) +
  theme_minimal() +
  labs(title="Boxplot de Variables", x="Variable", y="Valor") +
  facet_wrap(~ variable, scales = "free")
```

```
## No id variables; using all as measure variables
```




```
cor_matrix <- cor(df[, !(names(df) %in% "Resistencia")])
cor_melted <- melt(cor_matrix)
cor_melted_filtered <- cor_melted[upper.tri(cor_matrix, diag = FALSE), ]
cor_melted_filtered[order(cor_melted_filtered$value), ]
```

```
##             Var1          Var2     value
## 5       Longitud Altura.matriz 0.3784127
## 10 Altura.matriz  Altura.poste 0.4243451
## 14 Altura.matriz Altura.amarre 0.5377305
## 13      Longitud Altura.amarre 0.6560819
## 15  Altura.poste Altura.amarre 0.7793701
## 9       Longitud  Altura.poste 0.7950203
```

2. Método de Mínimos Cuadrados

```
# X <- cbind(1, df$Longitud, df$Altura.matriz, df$Altura.amarre)
X <- cbind(1, df$Longitud, df$Altura.amarre)

y <- df$Resistencia
beta <- solve(t(X) %*% X) %*% t(X) %*% y
beta
```

```
##          [,1]
## [1,] 2.645733
## [2,] 2.547728
## [3,] 3.548545
```

3. Regresión lineal múltiple en R

```
ols = lm(df$Resistencia ~ df$Longitud + df$Altura.amarre)
```

4. Evaluación del Modelo

```
vif(ols)
```

```
##      df$Longitud df$Altura.amarre
##         1.755752         1.755752
```

```
summary(ols)
```

```
##
## Call:
## lm(formula = df$Resistencia ~ df$Longitud + df$Altura.amarre)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7738 -1.5561  0.3703  1.5048  3.5758
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.6457     0.9448   2.800   0.0104 *
## df$Longitud         2.5477     0.1088  23.416  < 2e-16 ***
## df$Altura.amarre    3.5485     0.7136   4.973 5.63e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.17 on 22 degrees of freedom
## Multiple R-squared:  0.983,  Adjusted R-squared:  0.9815
## F-statistic: 637.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

4.1 Colinealidad de las variables involucradas

Las dos variables escogidas tienen un VIF de 1.76, por lo que se tiene baja certeza de colinealidad entre las variables. Probando modelos con todas las variables, se encontraron casos en donde el VIF para Altura Poste era casi 4, por lo que se buscó descartar con la sospecha de colinealidad. Al probar con las tres restantes, se tiene mayor certeza de ausencia de colinealidad.

4.2 Variabilidad explicada por el modelo (coeficiente de determinación)

Con las variables de Longitud y Altura Amarre, se explica el 98.15% de la variabilidad de los datos. Si se incluyen las variables Altura Poste y Altura Matriz, la variabilidad no podría ser tan distinta y la mejoría no es significativa, por lo que se evalúa que el modelo actual representa adecuadamente la variablidad de los datos.

4.3 Significancia del modelo: Valor p del modelo (F)

El modelo actual tiene un valor de F de 637.2, indicando que el modelo es significativo. Al incluir más variables se tendría una leve, pero mayor certeza de la significancia del modelo, mas no es necesario con el valor de p actual (menor a $2.2 * 10^{-16}$).

4.4 Significancia de Betas

Todas las betas son significantes, siendo especialmente significantes los coeficientes asociados especificamente a las variables predictoras. La beta de la intersección sigue siendo significativo, pero es menos significativo a comparación a los otros coeficientes.

4.5 Economía del modelo y variabilidad explicada (Coeficiente de determinación)

El modelo es viable y tiene un excelente porcentaje de variabilidad explicada. El tener ese nivel usando unicamente dos variables es un indicador de que el modelo es adecuado y sencillo.

5. Validación del Modelo

5.1 Normalidad de los residuos

Shapiro-Wilk normality test

- H_0 := La distribución de los errores es normal
- H_A := La distribución de los errores no es normal

```
shapiro.test(ols$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ols$residuals
## W = 0.97938, p-value = 0.8725
```

Con un valor de p de 0.8725 para la prueba de Shapiro-Wilk, no se tiene evidencia para rechazar la hipotesis nula, por lo que se puede inferir que la distribución de los errores es normal.

5.2 Verificación de media cero

T Test

- H_0 := La media de los errores es igual a 0.
- H_A := La media de los errores no es igual a 0.

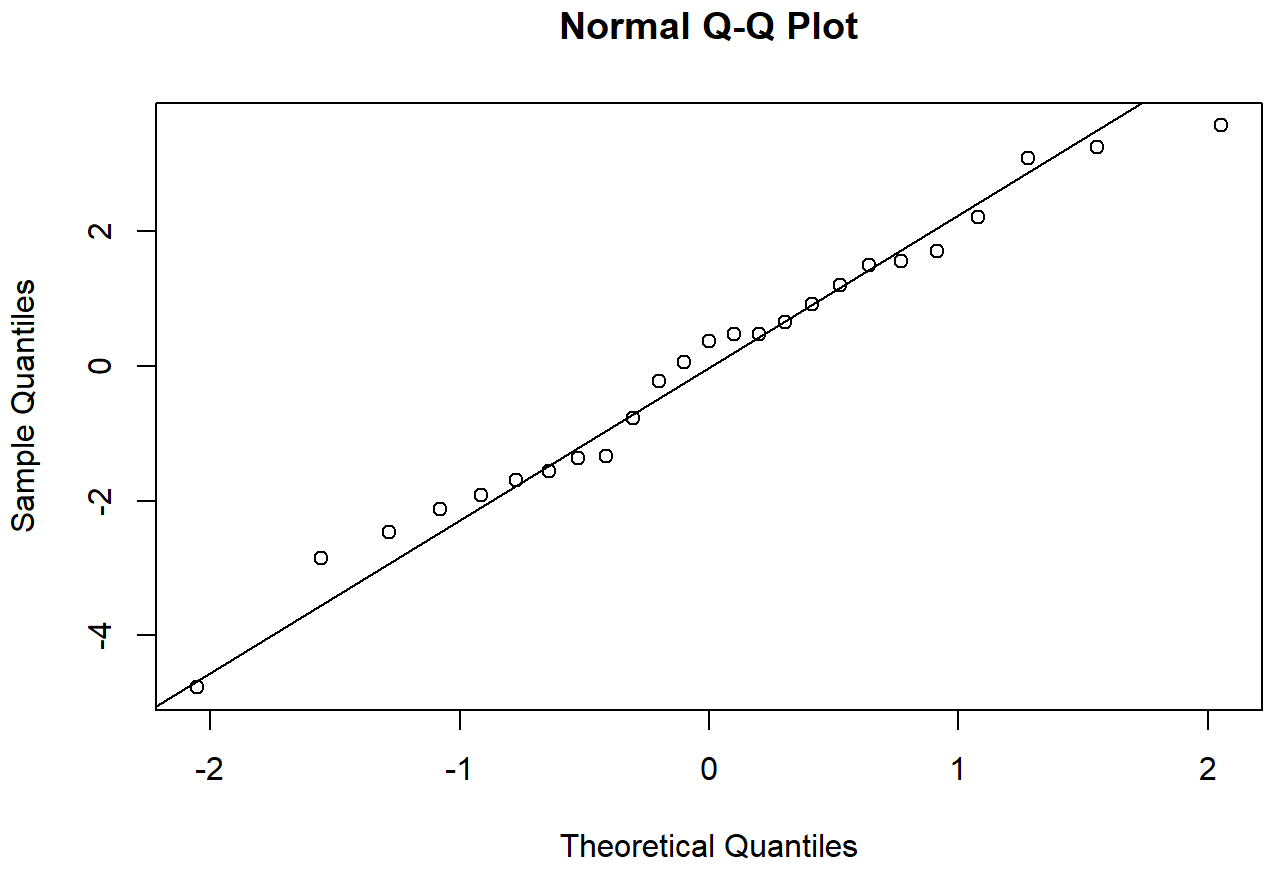
```
t.test(ols$residuals)
```

```
##
##  One Sample t-test
##
## data:  ols$residuals
## t = 9.3233e-17, df = 24, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.8576987  0.8576987
## sample estimates:
##    mean of x
## 3.874505e-17
```

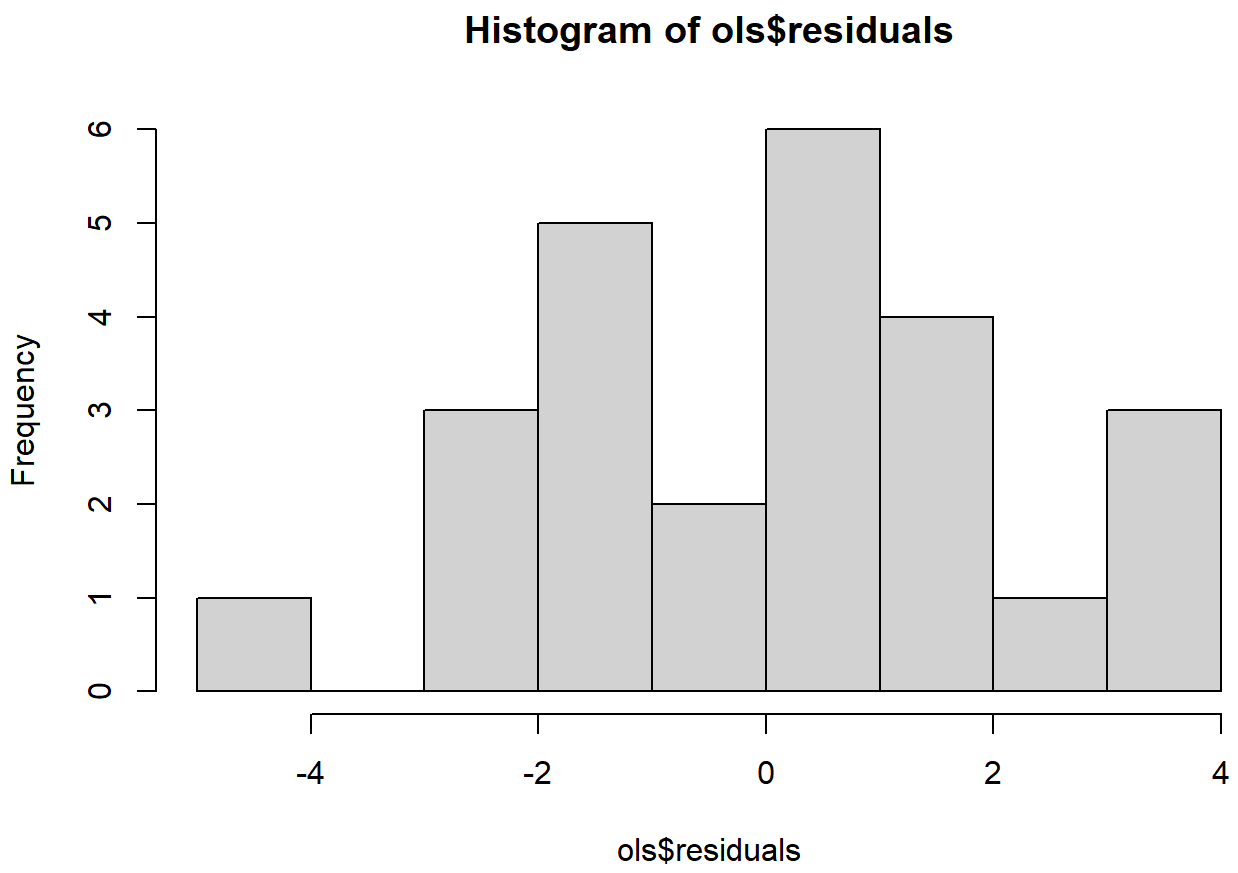
```
mean(ols$residuals)
```

```
## [1] 3.874505e-17
```

```
qqnorm(ols$residuals)
qqline(ols$residuals)
```




```
hist(ols$residuals)
```



Confirmando con gráficos y la prueba de T-student, se tiene un valor de p de “1”, indicando que no se tiene suficiente evidencia estadistica para rechazar la hipotesis nula. Por lo tanto, se infiere que la media de los residuos efectivamente es 0.

5.3 Homocedasticidad

Breusch-Pagan

- $H_0 :=$ Los datos tienen homocedasticidad.
- $H_A :=$ Los datos no tienen homocedasticidad.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.3.3
```

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

```
## Warning: package 'zoo' was built under R version 4.3.3
```

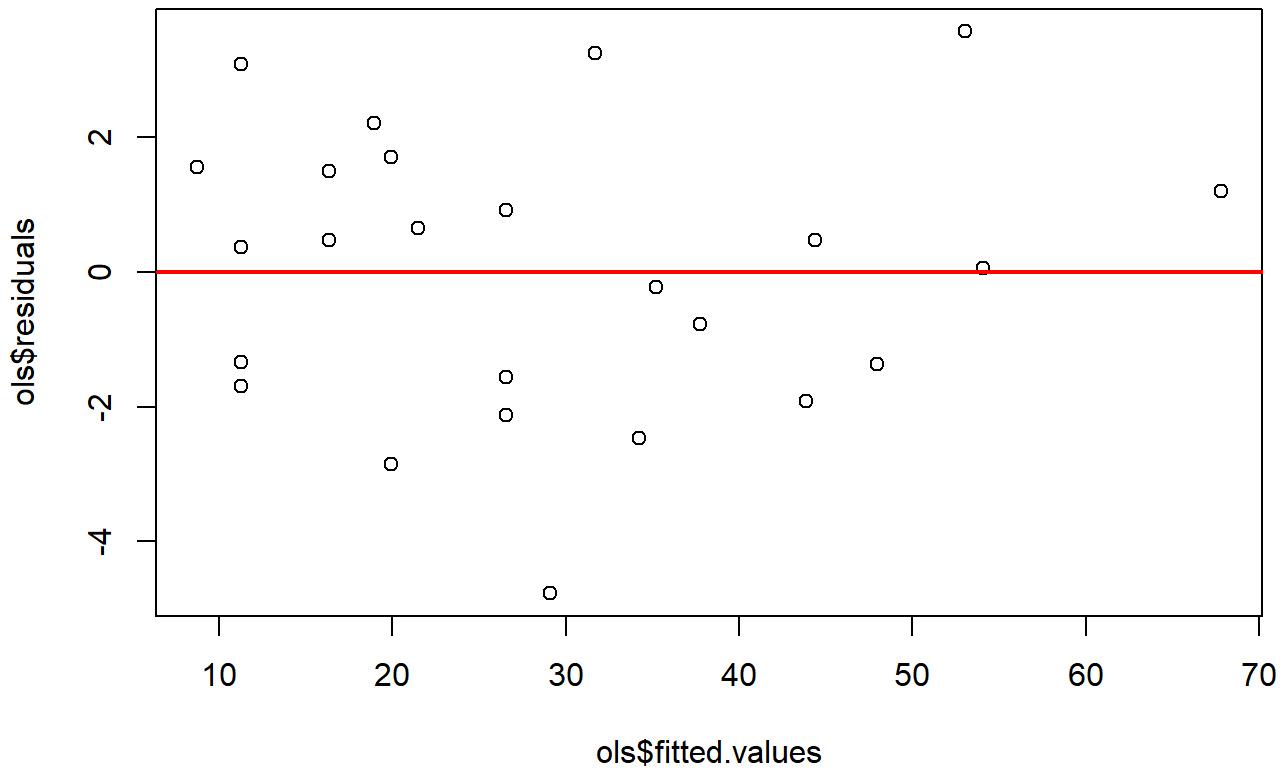
```
##
## Attaching package: 'zoo'
```

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

```
bptest(ols)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ols
## BP = 1.458, df = 2, p-value = 0.4824
```

```
plot(ols$fitted.values, ols$residuals)
abline(h=0, col = "red", lwd = 2)
```



Confirmando con gráficos y la prueba de Breusch-Pagan, el resultado es un valor de p de 0.4824, por lo que no se tiene suficiente evidencia estadistica para rechazar la hipotesis nula. Esto permite la inferencia de decir que los residuos tienen homocedasticidad.

5.4 Independencia

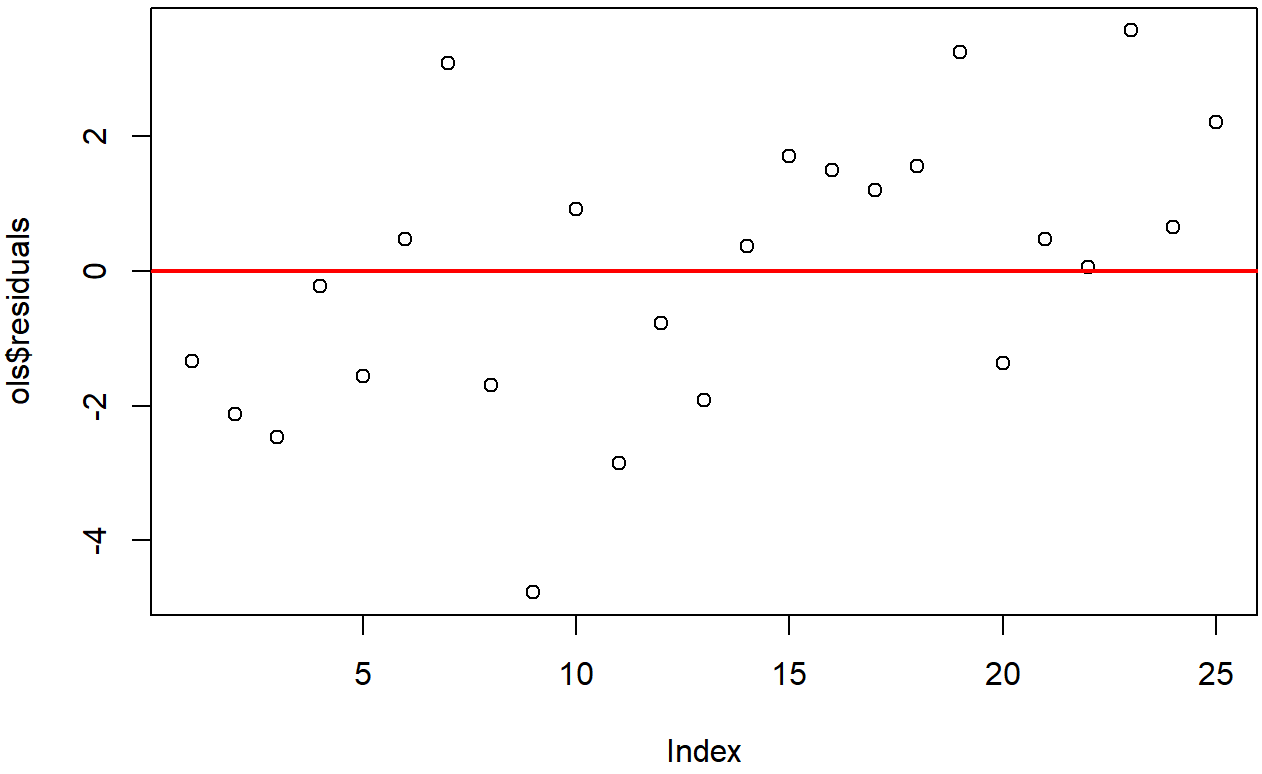
Durbin Watson

- $H_0 :=$ No existe autocorrelación en los datos
- $H_A :=$ Existe autocorrelacion en los datos.

```
dwtest(ols)
```

```
##
##  Durbin-Watson test
##
## data:  ols
## DW = 1.5614, p-value = 0.1171
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(ols$residuals)
abline(h=0, col = "red", lwd = 2)
```



Usando el gráfico se puede observar que podría haber alguna tendencia en la data, mas la prueba de Durbin-Watson indica que no se tiene suficiente evidencia estadística para rechazar la hipótesis nula. Esto indica que se puede inferir que no existe una autocorrelación en los residuos.

6. Conclusiones

El análisis de regresión lineal múltiple sobre el dataset determinó que las variables Longitud y Altura Amarre son las mejores predictoras de la resistencia al desprendimiento de un alambre adherido. El modelo final explica el 98.15% de la variabilidad en la resistencia, lo que indica un excelente ajuste. Todos los coeficientes son significativos, demostrando que ambas variables tienen un impacto relevante en la resistencia. No se reportan problemas de colinealidad, ya que los valores de VIF fueron bajos. Las pruebas de validación confirmaron la normalidad, homocedasticidad y la independencia de los residuos, lo que respalda la fiabilidad del modelo. El valor de F indica que el modelo es estadísticamente significativo. En resumen, el modelo es simple y eficaz para predecir la resistencia con un alto grado de precisión. Este resultado sugiere que no es necesario incluir variables adicionales para mejorar la calidad de las predicciones.