Actividad 2.2

Ricardo Kaleb Flores Alfonso

2024-09-24

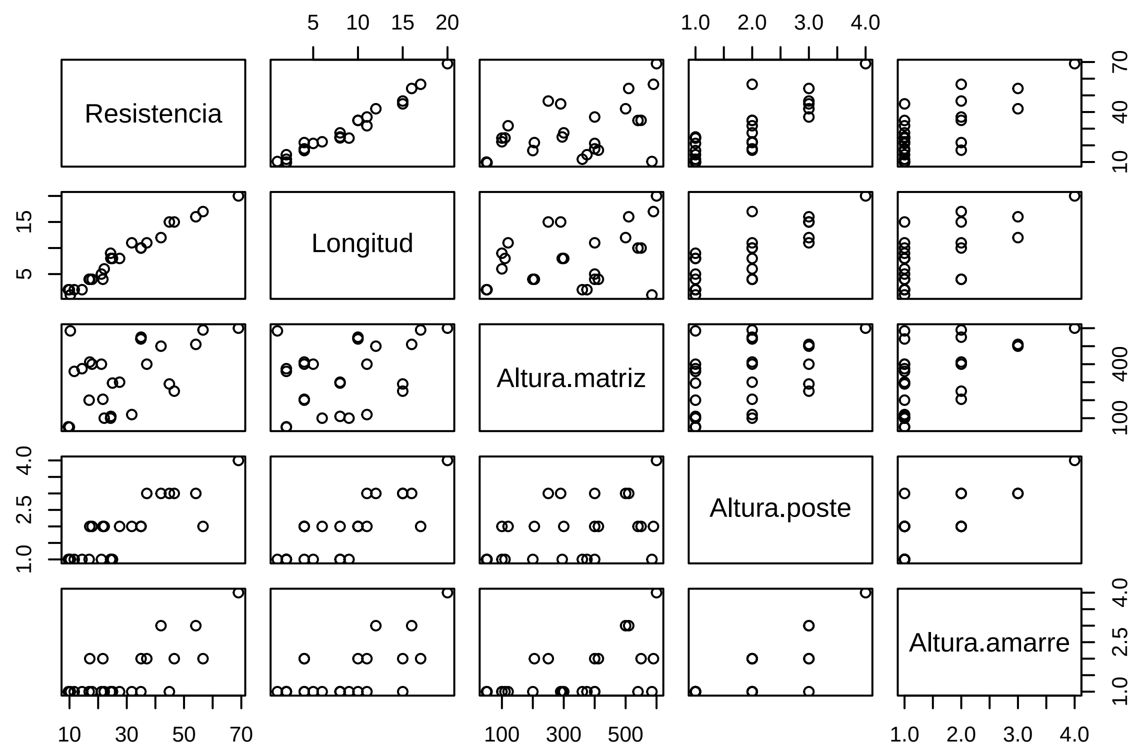
0) Importar librerias y base de datos

```
M <- read.csv("datosRes.csv")
```

1) Analisis exploratorio

1.1) Gráficos de dispersión

```
plot(M)
```



1.2) Obten la matriz de correlación

```
cor(M)
```

```
##                 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
```

4.1) Colinealidad de las variables

```
Y=M$Resistencia
x1=M$Longitud
x2=M$Altura.matriz
x3=M$Altura.poste
x4=M$Altura.amarre
R12=cor.test(x1,x2)
R13=cor.test(x1,x3)
R14=cor.test(x1,x4)
```

Prueba de hipótesis: Dada la hipotesis que:

- H_0 := No hay correlación
- H_A := Los datos tienen correlación

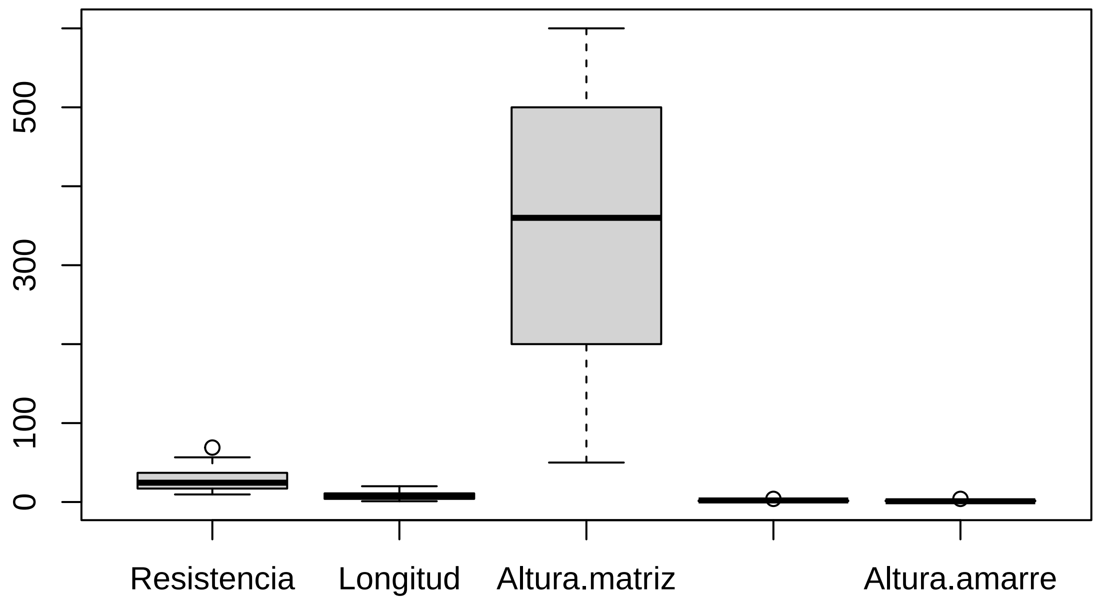
```
correlaciones=c(round(R12$estimate,3),round(R13$estimate,3),round(R14$estimate,3))
test=c(round(R12$statistic,3),round(R13$statistic,3),round(R14$statistic,3))
valorp=c(round(R12$p.value,9),round(R13$p.value,9),round(R14$p.value,9))
#b1=c(b1_1,b1_2,b1_3)

T=data.frame(correlaciones,test,valorp)
names(T)=c("r","t*","Valor p")
row.names(T)=c("r12","r13","r14")
T
```

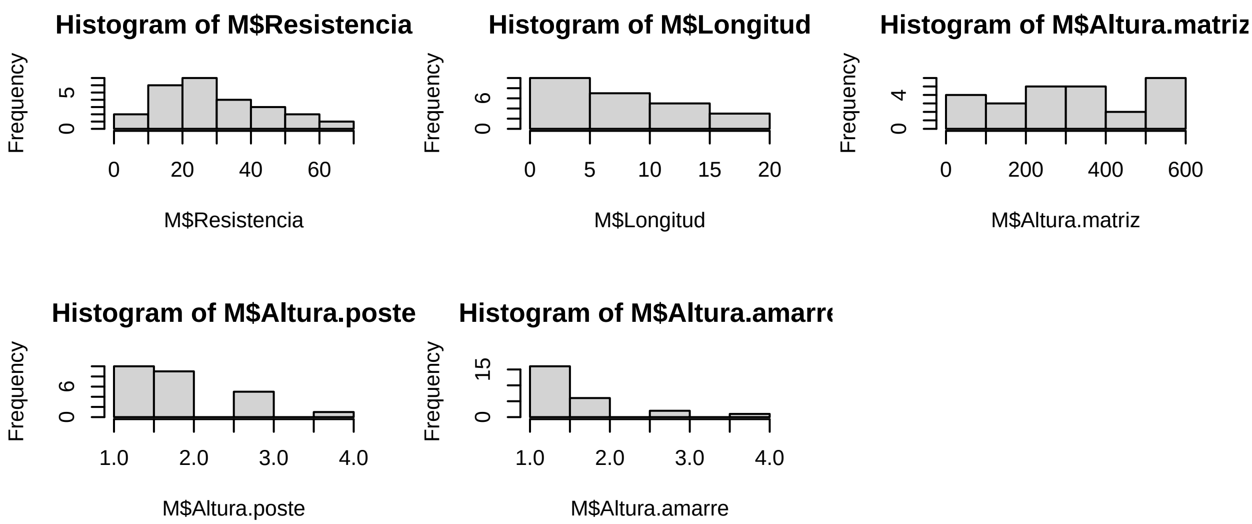
```
##         r    t*     Valor p
## r12 0.378 1.961 0.062146928
## r13 0.795 6.286 0.000002055
## r14 0.656 4.169 0.000369302
```

1.3) Ház graficos de Boxplot o histogramas para analizar las variables

```
boxplot(M)
```




```r
par(mfrow=c(3,3))
hist(M$Resistencia)
hist(M$Longitud)
hist(M$Altura.matriz)
hist(M$Altura.poste)
hist(M$Altura.amarre)
```



2) Método de mínimos cuadrados

```r
x <- cbind(1, M$Longitud, M$Altura.amarre)

y <- M$Resistencia

beta <- solve(t(x) %*% x) %*% t(x) %*% y
beta
```

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

3) Regresión Lineal múltiple

```r
#Modelo de regresión múltiple (reg)
reg=lm(Y~x1+x2)
reg$coefficients
```

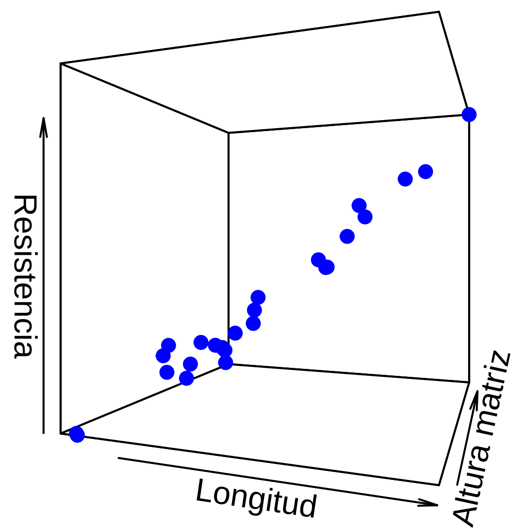
```
## (Intercept)          x1          x2
##  2.26379143  2.74426964  0.01252781
```

```r
cat("El modelo de regresión es Y=",reg$coefficients[1],"+",reg$coefficients[2],"X1+",reg$coefficients[3
```

```
## El modelo de regresión es Y= 2.263791 + 2.74427 X1+ 0.01252781 X2
```

4. Representación gráfica

```r
library("plot3D")
scatter3D(x1,x2,Y,col="blue",cex=0.9,pch=19,xlab="Longitud",ylab="Altura matriz",zlab="Resistencia",phi=
```



5. Coeficiente de determinación

```r
summary(reg)
```

```
##
## 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
```

El modelo explica el 97.94% de la variabilidad del modelo

6. Validación del modelo de regresión

Multicolinealidad

```r
#Se calcula el VIF
vif(reg)
```

```
##       x1       x2
## 1.167128 1.167128
```

Los valores de inflación de varianza en cada una de las variables predictoras son menores a 4, podemos suponer que el modelo no presenta evidencia de multicolinealidad

Significancia de los coeficiente de regresión

Analizaremos la significancia de los predictores X1 y X2

$H_0 : \beta_i = 0$ Y no depende de Xi $H_1 : \beta_i \neq 0$ Y depende de Xi

```r
summary(reg)
```

```
##
## 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
```

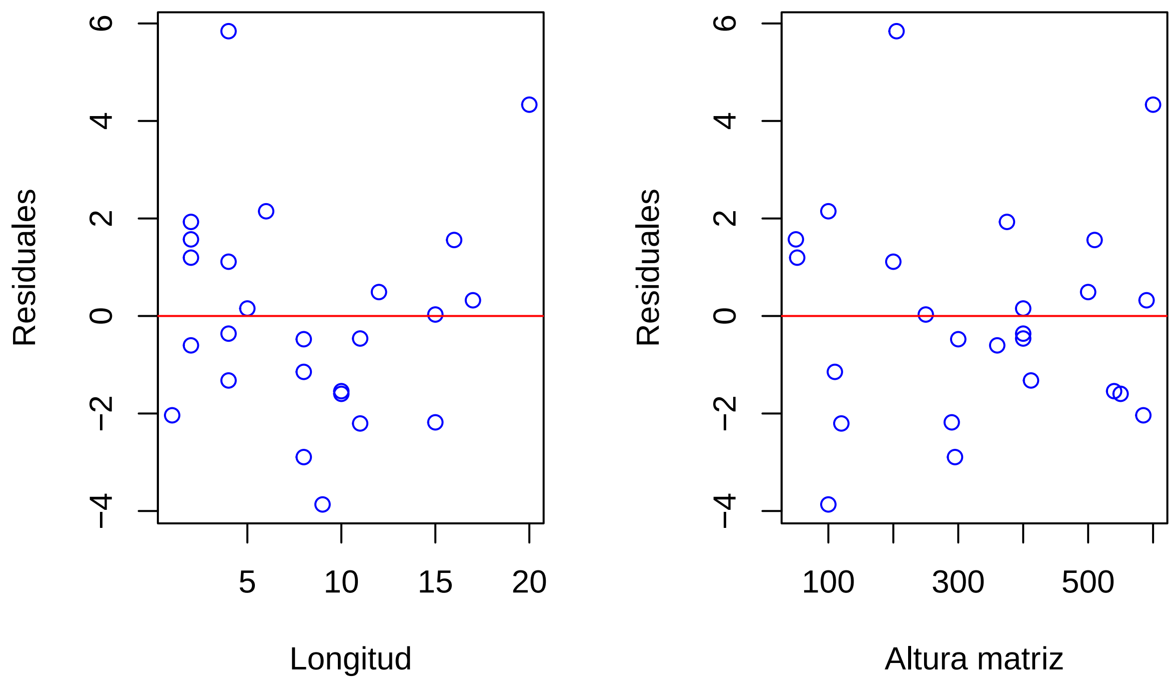
Relación lineal entre cada predictor y la variable de respuesta

Dado que los valores de la prueba de $p < 0.05$ Se rechaza la hipotesis nula, por lo tanto los coeficientes de x_1 y x_2 son mayores que 0

$H_0 : \beta_i = 0$ El modelo funcional es adecuado $H_1 : \beta_i \neq 0$ El modelo funcional no es adecuado

Graficaremos los residuos frente a cada variable independiente, esperando una distribución aleatoria alrededor de la media cero.

```r
par(mfrow=c(1,2))
plot(x1,reg$residuals, col="blue", ylab="Residuales",xlab="Longitud")
abline(h=0,col="red")
plot(x2,reg$residuals, col="blue", ylab="Residuales",xlab="Altura matriz")
abline(h=0,col="red")
```




```r
# Prueba de RESET de Ramsey
resettest(reg)
```

```
##
##  RESET test
##
## data:  reg
## RESET = 6.3418, df1 = 2, df2 = 20, p-value = 0.007362
```

A traves de la prueba de RESET de Ramsey, la cual nos da un valor de $p = 0.0073 < 0.05$ Con este resultado hay suficiente evidencia para rechazar la hipotesis nula, por lo que el modelo necesita de otras variables no lineales para explicar la variabilidad de la y. Sin embargo al graficar los puntos, estos se distribuyen de manera homogenea, por lo que es posible que los valores extremos esten afectando a la prueba.

Análisis de residuales

A continuación se validará si los residuales siguen una distribución Normal con media cero y varianza constante.

Media cero

Se realiza una prueba de hipótesis para media.

$H_0 : \mu_R = 0$

$H_1 : \mu_R \neq 0$

$\alpha = 0.05$

```
#prueba t.test con los residuales
t.test(reg$residuals,mu=0,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  reg$residuals
## t = 2.2037e-16, 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
## 9.655091e-17
```

Como el p valor es mayor que 0.05, por lo que no hay suficiente evidencia para demostrar que la media de los residuales es 0.

Normalidad de residuales

Utilizaremos un test de Normalidad

H_0: Los residuales (datos) tienen distribución Normal

H_1: No provienen de una distribuyen Normal

Puesto que se tienen <50 datos se aplicará la prueba Shapiro-Wilk

```
#Test de normalidad
shapiro.test(reg$residuals)
```

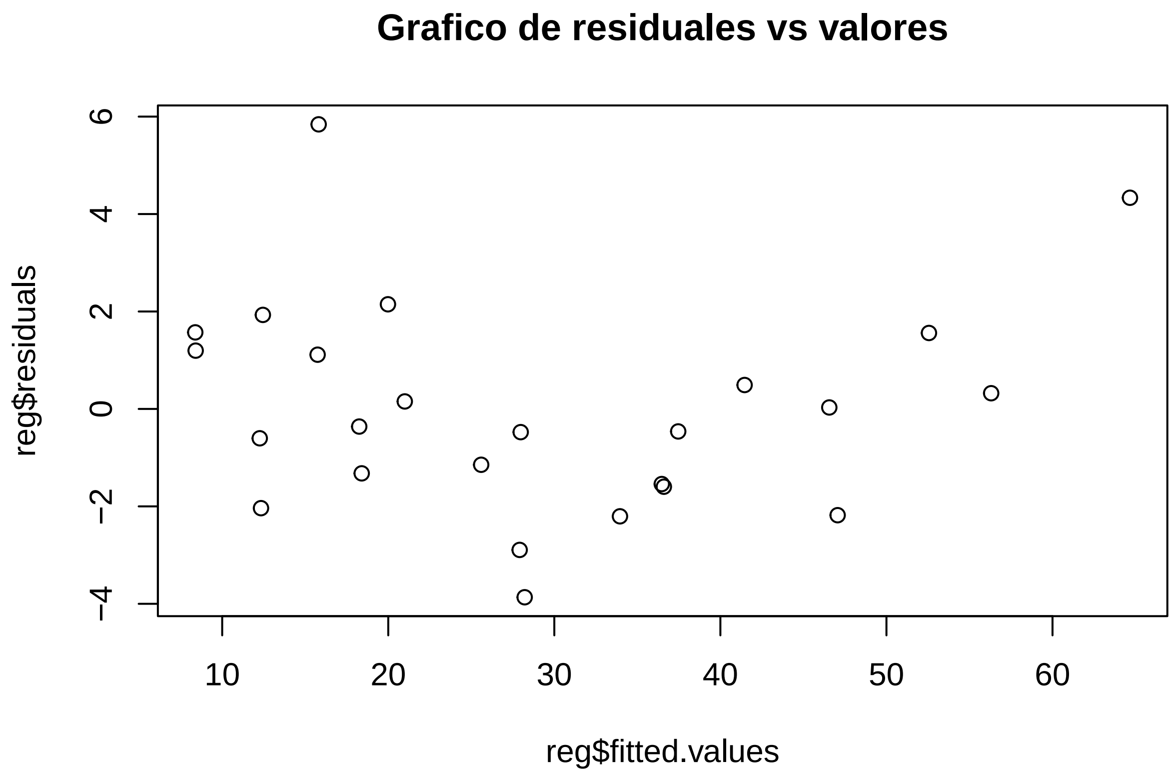
```
##
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.95827, p-value = 0.381
```

El p valor obtenido es mayor a 0.05, por lo que no hay suficiente evidencia para rechazar la hipotesis nula. Por lo que podemos suponer que los residuales tienen distribución normal

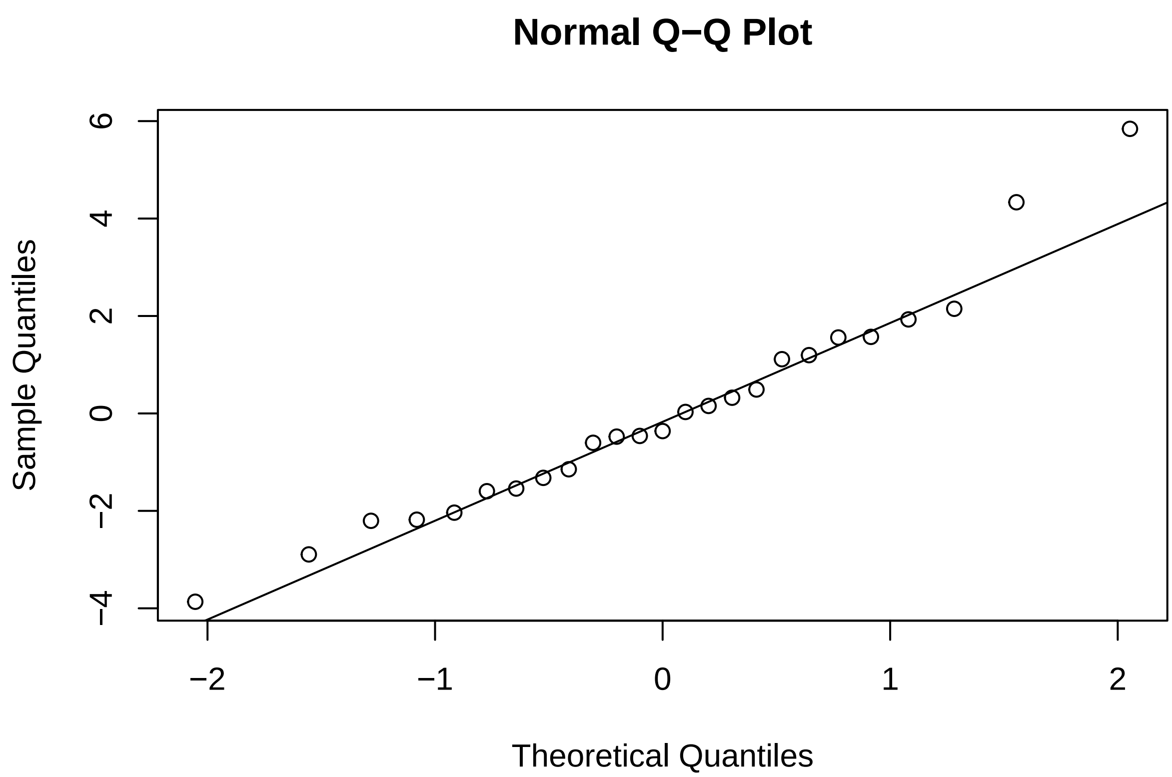
Homocedasticidad

La varianza de los residuos debe de ser constante en todo el rango de observaciones.Para comprobarlo graficaremos los residuales de las estimaciones de la regresión para observar si se distribuyen de forma aleatoria manteniendo una misma dispersión y sin ningún patrón específico a lo largo del eje horizontal (media cero).

```
#Gráfico de residuales frente a los valores predichos
plot(reg$fitted.values,reg$residuals,main="Grafico de residuales vs valores")
```




```r
qqnorm(reg$residuals)
qqline(reg$residuals)
```



Segun el gráfico de residuales vs valores, los datos estan distribuidos debajo del 0, es por esto que se puede concluir que la varianza no es constante para el modelo.

H_0: La varianza es constante homocedasticidad

H_1: La varianza no es constante hereocedasticidad

Se implementarán además la prueba de Breusch-Paga que evalúa si los residuales son función lineal de las covariables del modelo, además del test de White, la cual es una prueba más robusta que detecta formas no lineales de la heterocedasticidad.

```r
#Prueba Breusch-Pagan:
bptest(reg)
```

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

```r
#Prueba de White:
bptest(reg, varformula = ~ x1 * x2 + I(x1^2) + I(x2^2))
```

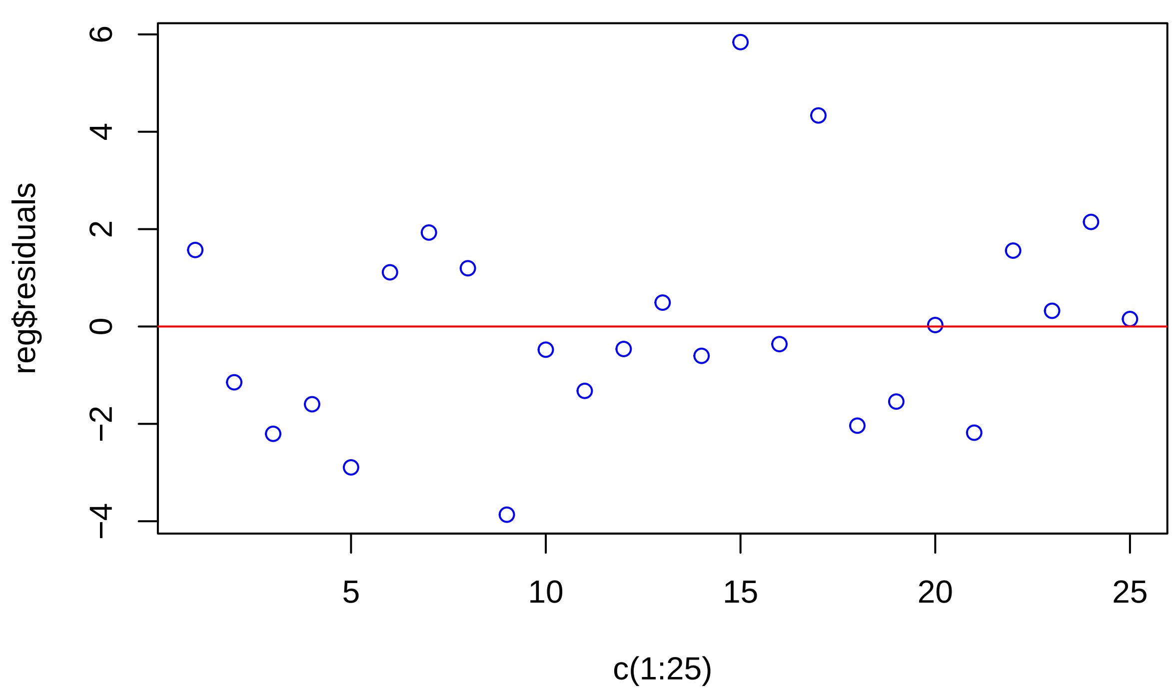
```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 2.0607, df = 5, p-value = 0.8407
```

El valor obtenido p , para la prueba de Breusch-Pagan y la prueba de White, es mayor que 0.05, por lo que no hay evidencia para rechazar la hipotesis nula, por lo que los datos tiene homocedasticidad, lo que significa que tienenn varianza constante.

Independencia

Se representarán los residuos ordenados acorde al tiempo de registro de las observaciones para observar si existe algún patrón.

```
n=length(M$Resistencia)
plot (c(1:25), reg$residuals, col="blue")
abline(h=0, col="red")
```



Adicionalmente, se realizará la prueba de Durbin-Watson para detectar presencia de autocorrelación de los residuos en un esquema autoregresivo de primer orden.

Por otra parte la prueba Breusch-Godfrey evalúa la autocorrelación de los residuos con un esquema autoregresivo con órdenes superiores.

H_0: Los residuales no estan correlacionados

H_1: Los residuales estan correlacionados

```
#Prueba Durbin-Watson
dwtest(reg)
```

```
##
##  Durbin-Watson test
##
## data:  reg
## DW = 2.0972, p-value = 0.559
```

```
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
#Prueba Breusch-Godfrey
bgtest(reg)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  reg
## LM test = 0.10479, df = 1, p-value = 0.7462
```

Dado que el valor p de ambas pruebas es mayor a 0.05, no hay evidencia suficiente para rechazar la hipotesis nula, por lo que los residuales no estan correlacionados

#Conclusiones

```r
b0=reg$coefficients[1]
b1=reg$coefficients[2]
b2=reg$coefficients[3]
summary(reg)
```

```
##
## 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
```

El modelo lineal múltiple: Y=2.2637914+2.7442696X1 + 0.0125278X2

Es capaz de explicar el 97.94% de la variabilidad observada en la resistencia. El test F muestra que el modelo es 572.2 para explicar la variablidad de la resistencia. Ademas los supuestos de multicolinealidad, normalidad de residuales, homocedasticidad e independencia, los cuales se cumplieron. De esta manera se puede concluir que el modelo es estadisticamente significativo, asi como puede ser usado para explicar los valores de Y.