Actividad 2.2 Análisis de varianza: Resistencia

Franco Mendoza Muraira A01383399

2023-11-06

```
df= read.csv(file="resistencia.csv")
df$Concentracion = as.factor(df$Concentracion)
```

```
c5 = subset(df$Resistencia,df$Concentracion==5)
c10 = subset(df$Resistencia,df$Concentracion==10)
c15 = subset(df$Resistencia,df$Concentracion==15)
c20 = subset(df$Resistencia,df$Concentracion==20)

data <- data.frame(Concentracion_5 = c5, Concentracion_10 = c10, Concentracion_15 = c15, Concentracion_

names(data) <- c("5", "10", "15", "20")

head(data)
```

```
##    5 10 15 20
## 1  7 12 14 19
## 2  8 17 18 25
## 3 15 13 19 22
## 4 11 18 17 23
## 5  9 19 16 18
## 6 10 15 18 20
```

1. Analisis Exploratorio

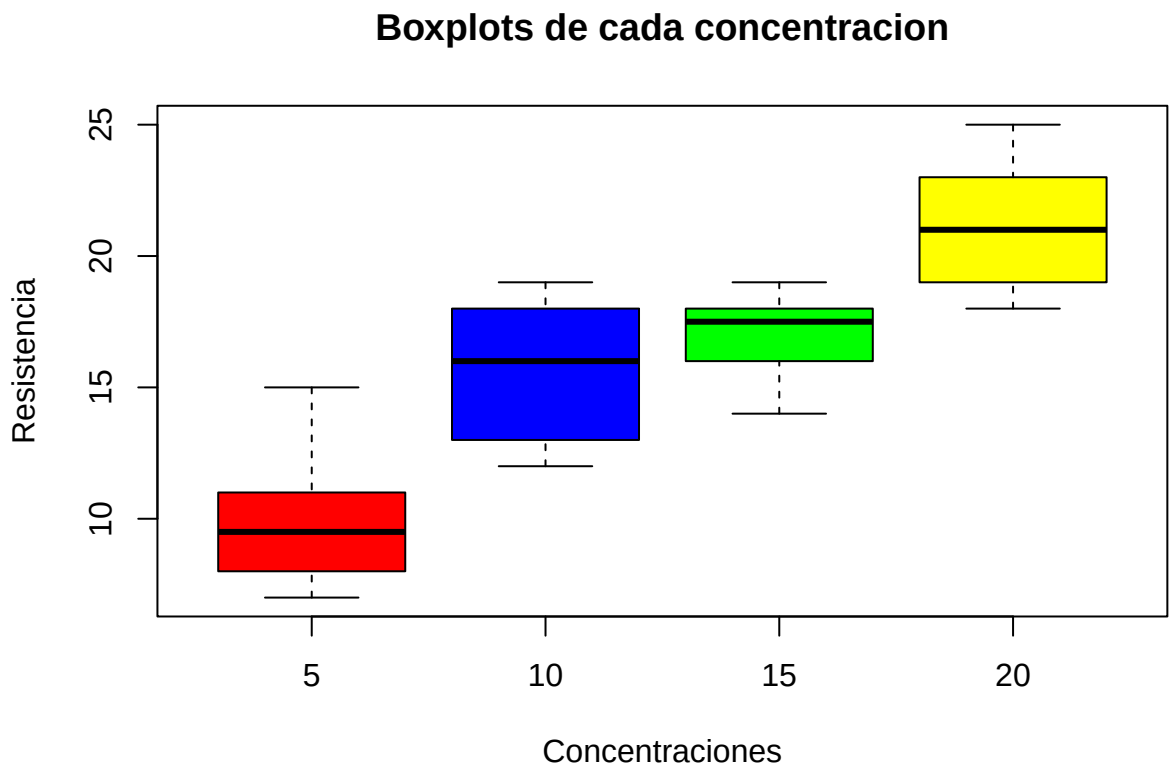
Medias

```
for (i in names(data)){
  cat("La media en la concentracion",i," es de ",mean(data[,i]),"\n")
}
```

```
## La media en la concentracion 5  es de  10
## La media en la concentracion 10  es de  15.66667
## La media en la concentracion 15  es de  17
## La media en la concentracion 20  es de  21.16667
```

Boxplots

```r
boxplot(data,col=c("red","blue","green","yellow"),main="Boxplots de cada concentracion",xlab="Concentra
```



En estos boxplots, los podemos comparar y observar que hay medias parecidas como las de 10 y 15, 15 teniendo poca variabilidad y las otras 3 teniendo datos mucho mas variables. En si a simple vista se ven diferentes las medias, pero se teiene que comprobar.

2. Hipótesis estadística

H_0: Todas las medias son iguales

H_1: Al menos hay una media diferente

Nivel de significancia (alfa) : 0.05

```r
alfa = 0.05
```

3. ANOVA: Suma de cuadrados medios

```r
SSTR = 0
SSE = 0
TSS = 0
for (i in names(data)){
```

```
  SSTR=SSTR+(mean(data[,i])-mean(colMeans(data)))^2
}
SSTR =SSTR*nrow(data)

for (i in names(data)) {
  for (j in data[[i]]) {
    SSE = SSE + (j - mean(data[[i]]))^2
  }
}

for (i in names(data)) {
  for (j in data[[i]]) {
    TSS = TSS + (j - mean(colMeans(data)))^2
  }
}
MSTR = SSTR/(ncol(data)-1)
MSE = SSE/(ncol(data)*(nrow(data)-1))
f = MSTR/MSE
crit=qf(1-alfa,ncol(data)-1,ncol(data)*(nrow(data)-1))
ANOVA = data.frame(
  Medida = c("SSTr (Inter grupos)","SSE (Intra grupos)","TSS (Total)","MSTR (Inter grupos)","MSE (Intra
  Valor = c(SSTR,SSE,TSS,MSTR,MSE,f,crit)
)
ANOVA
```

```
##                Medida      Valor
## 1 SSTr (Inter grupos) 382.791667
## 2  SSE (Intra grupos) 130.166667
## 3         TSS (Total) 512.958333
## 4 MSTR (Inter grupos) 127.597222
## 5  MSE (Intra grupos)   6.508333
## 6    f (Inter grupos)  19.605207
## 7   Regla de decisión   3.098391
```

Como se puede ver en la tabla, el valor f es mayor al de la regla de decisión, por lo que se tiene suficiente evidencia para rechazar H_0, lo que nos dice que al menos una de las medias es diferente, no todas son iguales.

4. ANOVA en R

```
mod_lm = lm(df$Resistencia~df$Concentracion)
anova_r = aov(df$Resistencia~df$Concentracion)
summary(anova_r)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## df$Concentracion  3  382.8  127.60   19.61 3.59e-06 ***
## Residuals        20  130.2    6.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

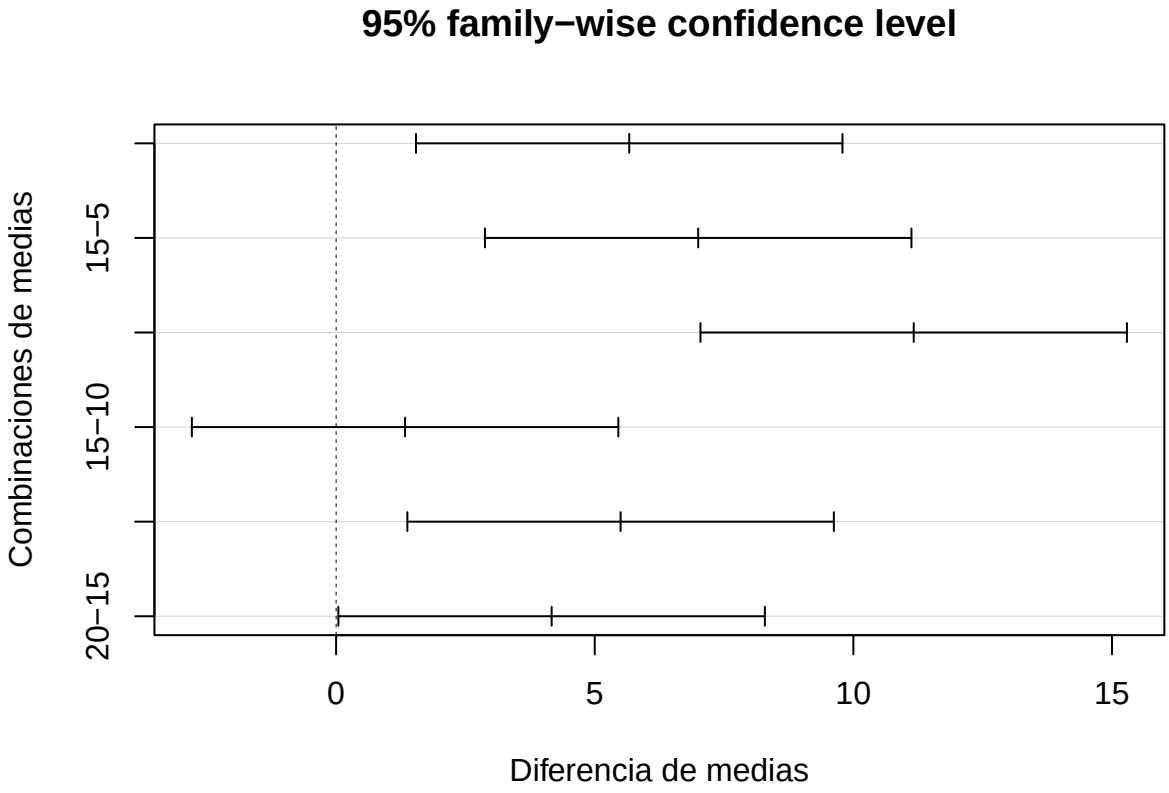
En el ANOVA con R se ve que se dieron los mismos resultados que calculados manualmente, lo que nos muestra que se hizo de manera correcta. En cuanto a las hipótesis, confirmamos lo que se mencionó previamente de rechazar H_0 debido al bajo valor de la p, siendo $3.59 * 10^{-6}$, lo cuál es menor al valor de α establecido como 0.05. Terminamos con la misma conclusión de que no todas las medias de las poblaciones son iguales.

5. Diferencias por pares

```
TukeyHSD(anova_r)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = df$Resistencia ~ df$Concentracion)
##
## $`df$Concentracion`
##            diff         lwr       upr     p adj
## 10-5   5.666667  1.54410408  9.789229 0.0051108
## 15-5   7.000000  2.87743741 11.122563 0.0006501
## 20-5  11.166667  7.04410408 15.289229 0.0000015
## 15-10  1.333333 -2.78922925  5.455896 0.8022275
## 20-10  5.500000  1.37743741  9.622563 0.0065966
## 20-15  4.166667  0.04410408  8.289229 0.0470251
```

```
tukey = TukeyHSD(anova_r,conf.level=1-alfa)
#plot(tukey,cex.axis=0.6,main="Prueba Tukey de diferencia de medias de las concentraciones",xlab="Difer
tuk_plot(tukey,"Diferencia de medias","Combinaciones de medias")
```



La prueba anterior hace la comparación de todas las combinaciones de diferencias de medias, y en la prueba con los valores de p ajustados pudimos ver que uno de los pares que podría considerarse que puede tener una media igual es la de 15 y 10 debido a su pvalue tan alto y en la gráfica sobrepasa la diferencia de 0 entre las 2 medias. Pero siendo que todas las demás son diferentes, nos quedamos con la conclusión de que se tiene que rechazar H_0.

6. Validación de supuestos

Normalidad

H_0 : Los residuos siguen una distribución normal.

H_1 : Los residuos no siguen una distribución normal.

```
library(nortest)
library(lmtest)
```

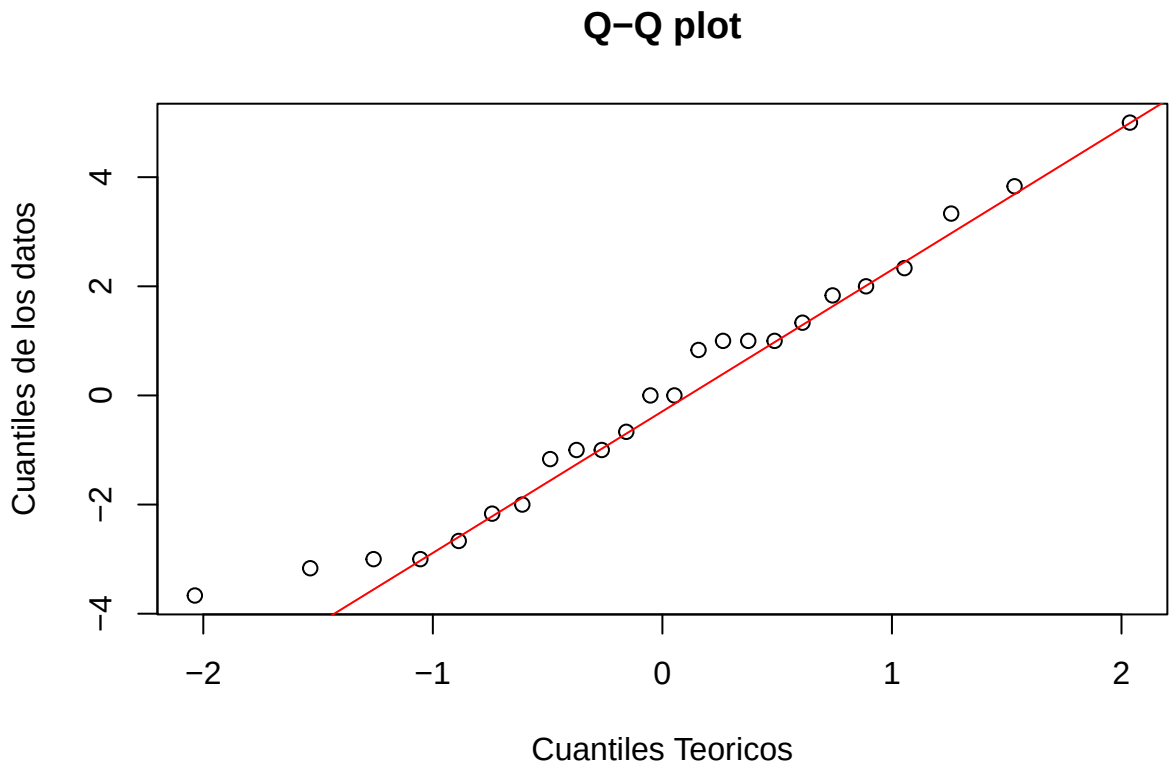
```
## Warning: package 'lmtest' was built under R version 4.2.3

## 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
```

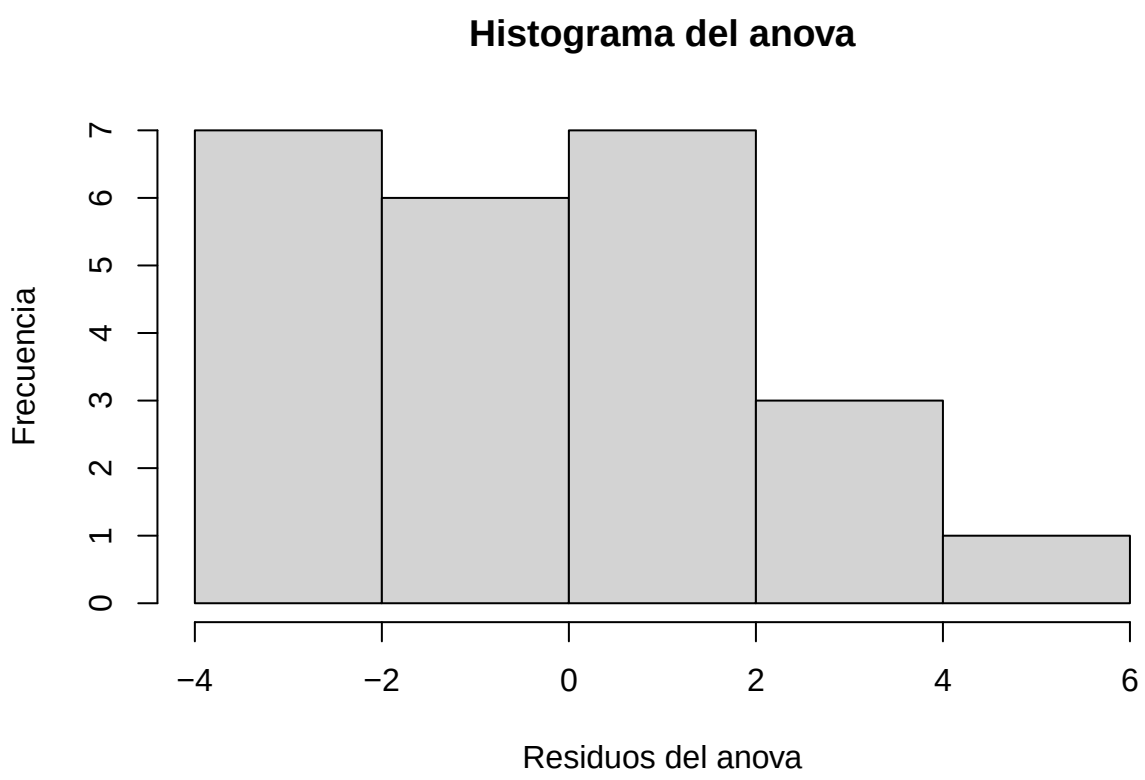
```
shapiro.test(residuals(anova_r))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_r)
## W = 0.96624, p-value = 0.5757
```

```
qqnorm(residuals(anova_r),main="Q-Q plot",ylab="Cuantiles de los datos",xlab="Cuantiles Teoricos")
qqline(residuals(anova_r),col="red")
```




```
hist(residuals(anova_r),main="Histograma del anova",xlab="Residuos del anova",ylab="Frecuencia")
```



Para revisar la normalidad de los residuos se hizo la prueba de Anderson-Darling, y se obtuvo un pvalue de 0.5757, el cual es mayor que el α por lo que no podemos rechazar H_0 y se concluye que los residuos sí cumplen una distribución normal.

Homocedasticidad

H_0 : La varianza de los errores es constante.

H_1 : La varianza de los errores no es constante.

```
bptest(anova_r)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  anova_r
## BP = 1.7746, df = 3, p-value = 0.6205
```

```
fisher.test(table(df))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(df)
## p-value = 0.9662
## alternative hypothesis: two.sided
```

Con la prueba de Breusch-Pagan podemos concluir que no hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad. Esto sugiere que los datos podrían tener una varianza constante en los diferentes niveles de las variables.

Y con la prueba de Fisher tampoco se puede rechazar la hipótesis nula, esto es debido a su alto pvalue de 0.9662, esto nos sirve para concluir lo mismo de que la varianza de los errores no es constante.

Independencia

H_0 : autocorrelacion en los residuos $= 0$

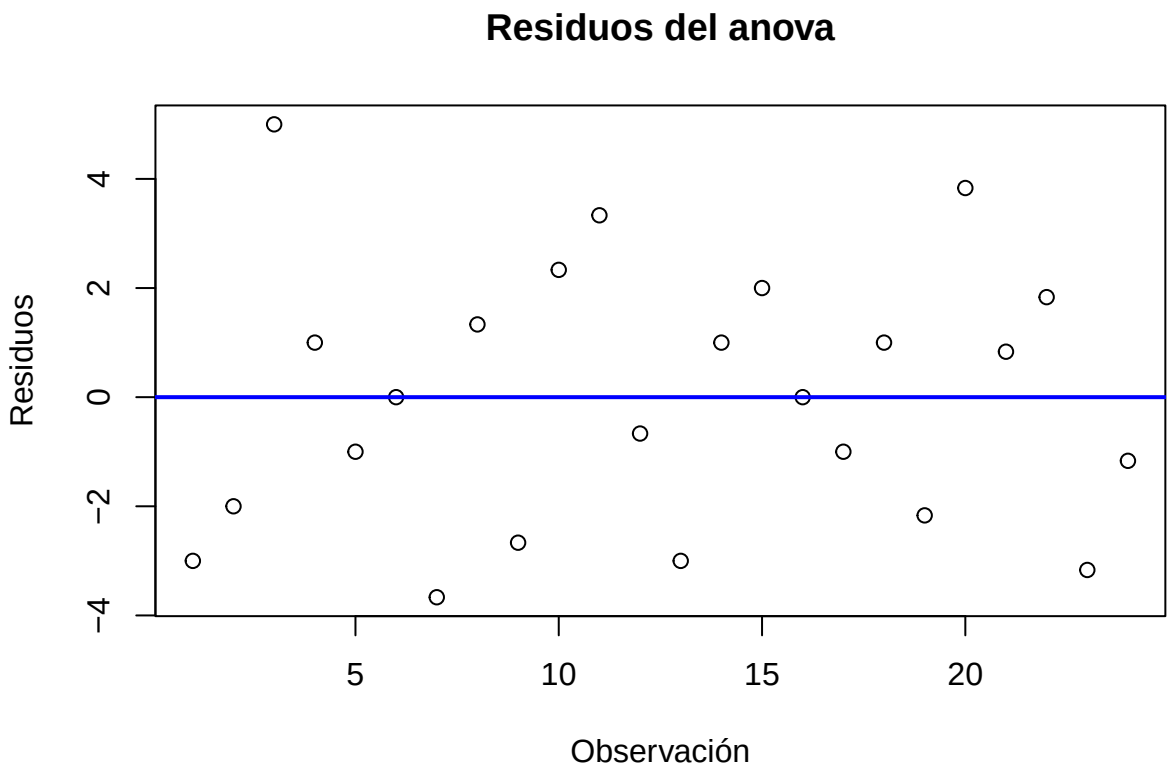
H_1 : autocorrelacion en los residuos $\neq 0$

```
dwtest(anova_r)
```

```
##
##  Durbin-Watson test
##
## data:  anova_r
## DW = 2.1812, p-value = 0.424
## alternative hypothesis: true autocorrelation is greater than 0
```

Debido a que el pvalue es de 0.424 y mayor que α, esto significa que no se rechaza H_0. Esto significa que no hay suficiente evidencia para decir que no hay autocorrelación en los residuos del modelo de regresión, lo que nos dice que hay independencia entre los residuos. En la siguiente gráfica se puede ver que los residuos están distribuidos de manera independiente del otro.

```
plot(residuals(anova_r),main="Residuos del anova",xlab="Observación",ylab="Residuos")
abline(h=0,col="blue",lwd=2)
```



7. Intervalos de confianza

```
for (col in names(data)) {
  result <- t.test(data[[col]])

  cat("Concentración:", col, "\n")
  cat("Intervalo de confianza del 95% para la media de resistencia:[",
      result$conf.int, "]\n\n")
}
```

```
## Concentración: 5
## Intervalo de confianza del 95% para la media de resistencia:[ 7.031748 12.96825 ]
##
## Concentración: 10
## Intervalo de confianza del 95% para la media de resistencia:[ 12.72325 18.61008 ]
##
## Concentración: 15
## Intervalo de confianza del 95% para la media de resistencia:[ 15.12271 18.87729 ]
##
## Concentración: 20
## Intervalo de confianza del 95% para la media de resistencia:[ 18.39674 23.93659 ]
```

Estos resultados podrían implicar que la concentración de madera dura puede influir en la resistencia del material, siendo la concentración de 15 la que muestra una resistencia más consistente, en comparación con

las otras concentraciones evaluadas en este estudio. Se puede ver que la concentracion de 5 muestra una alta variabilidad en las mediciones, al igual que en el de 20. Las concentraciones de 10 y 15 son más estrechos y tienen mayor precisión, la de 15 siendo la más estrecha de todas.

8. Conclusión

Con las pruebas estadísticas utilizadas pudimos llegar a diferentes conclusiones, la principal siendo comprobar si todas las medias de esta base de datos son iguales, para loq ue llegamos a concluir que no.

También pudimos observar como funciona el ANOVA, y la comparación inter grupos e intra grupos de ella. También pudimos ver las diferentes combinaciones de diferencias de medias en los diferentes pares de concentraciones con lo que pudimos ver que una de ellas si tenía una diferenencia de medias no significativa, la 10 y 15. Hicimos otras diferentes pruebas con los residuos del modelo, en donde pudimos observar normalidad en la distribución, homocedasticidad, y autocorrelación en ellos.

Podemos concluir que la concentración de madera dura si hace diferencia en la resistencia de el papel creado.