Actividad 2.2 Análisis de varianza: Resistencia

Raúl Correa Ocañas - A01722401

2024-08-12

Actividad 2.2 ANOVA

Un fabricante de papel para hacer bolsas comestibles se encuentra interesado en mejorar la resistencia a la tensión del producto. El departamento de ingeniería del producto piensa que la resistencia a la tensión es una función de la concentración de madera dura en la pulpa y que el rango de las concentraciones de madera dura de interés práctico está entre 5% y 20%.

El equipo de ingenieros responsables del estudio decide investigar cuatro niveles de concentración de madera dura 5%, 10%, 15% y 20%. Deciden hacer seis ejemplares de prueba con cada nivel de concentración utilizando una planta piloto. Las 24 muestras se prueban en orden aleatorio con una máquina de laboratorio para probar resistencia. A continuación, se muestran los datos de este experimento.

Concentración de madera dura (%)	Observaciones						Totales	Promedios
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	**15.96**

¿Hay alguna diferencia en la resistencia del papel causada por la concentración de madera dura?

0. Importación de los Datos

```
data_path = r"{..\data\resistencia.csv}"
df <- as.data.frame(read.csv(data_path))
df$Concentracion <- as.factor(df$Concentracion)
df
```

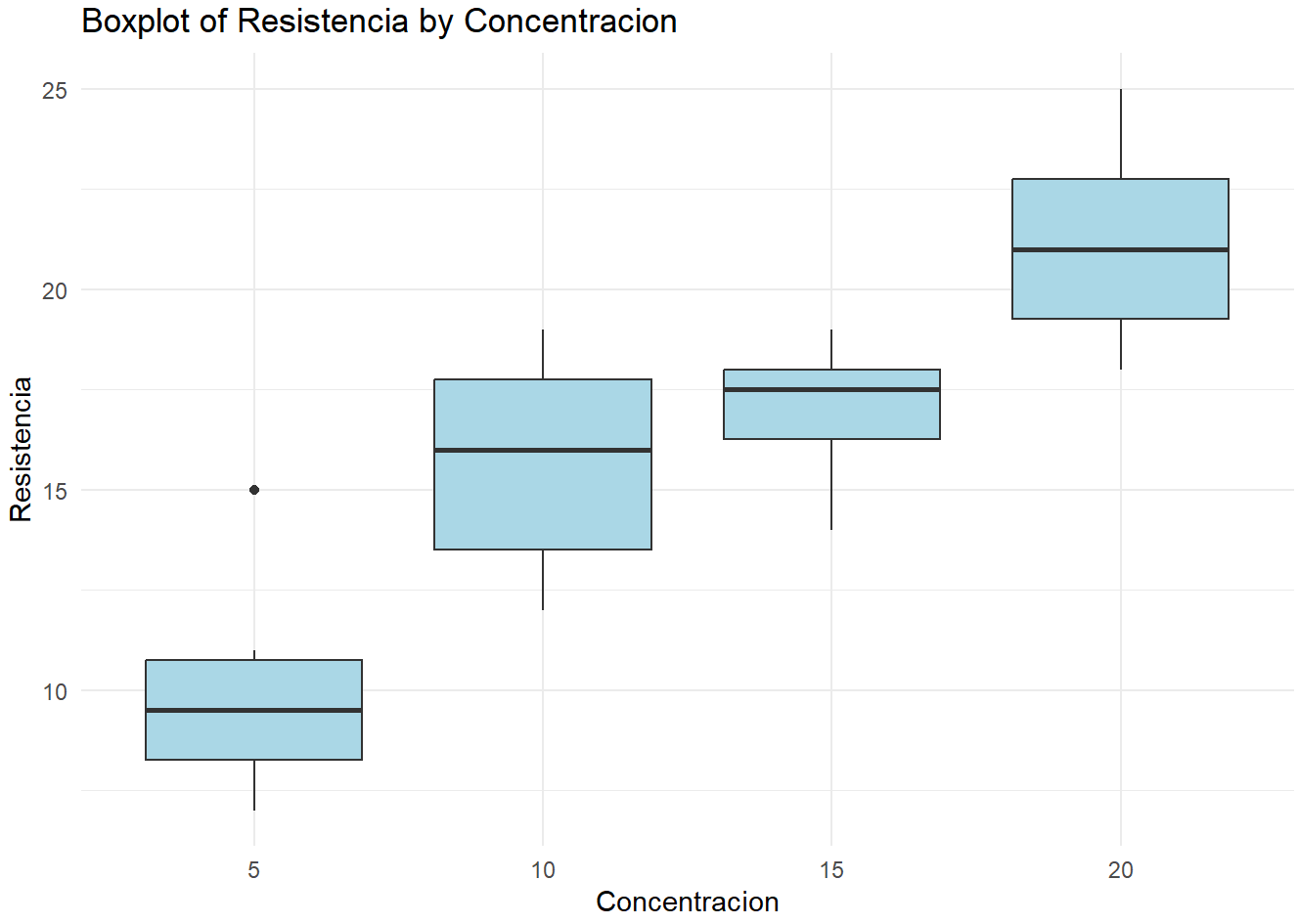
```
##    Resistencia Concentracion
## 1            7             5
## 2            8             5
## 3           15             5
## 4           11             5
## 5            9             5
## 6           10             5
## 7           12            10
## 8           17            10
## 9           13            10
## 10          18            10
## 11          19            10
## 12          15            10
## 13          14            15
## 14          18            15
## 15          19            15
## 16          17            15
## 17          16            15
## 18          18            15
## 19          19            20
## 20          25            20
## 21          22            20
## 22          23            20
## 23          18            20
## 24          20            20
```

1. Análisis Exploratorio de Datos

```
summary_df <- df %>%
  group_by(Concentracion) %>%
  summarize(
    Min = min(Resistencia, na.rm = TRUE),
    Q1 = quantile(Resistencia, 0.25),
    Q2 = quantile(Resistencia, 0.50),
    Mean = mean(Resistencia, na.rm = TRUE),
    Q3 = quantile(Resistencia, 0.75),
    Max = max(Resistencia, na.rm = TRUE),
    SD = sd(Resistencia, na.rm = TRUE),
    Skew = skewness(Resistencia, na.rm = TRUE),
    Kurtosis = kurtosis(Resistencia, na.rm = TRUE),
    count = n()
  )

summary_df
```

```
## # A tibble: 4 × 11
##   Concentracion   Min    Q1    Q2  Mean    Q3   Max    SD   Skew Kurtosis count
##   <fct>         <int> <dbl> <dbl> <dbl> <dbl> <int> <dbl>  <dbl>    <dbl> <int>
## 1 5                 7  8.25   9.5  10    10.8    15  2.83  0.663    -1.11     6
## 2 10               12 13.5   16    15.7  17.8    19  2.80 -0.124    -1.96     6
## 3 15               14 16.2   17.5  17    18      19  1.79 -0.524    -1.37     6
## 4 20               18 19.2   21    21.2  22.8    25  2.64  0.177    -1.79     6
```

```
ggplot(df, aes(x = as.factor(Concentracion), y = Resistencia)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Boxplot of Resistencia by Concentracion",
    x = "Concentracion",
    y = "Resistencia"
  ) +
  theme_minimal()
```



Al agrupar los datos según sus respectivas concentraciones, se puede observar una tendencia ascendiente en la media de las muestras a lo largo del aumento en la concentración. Si bien las medias fueran muy similares entre concentraciones, indicaría que la resistencia no varía según la concentración. Como se observa justamente lo contrario, incluso la posibilidad de una relacion lineal entre concentración y resistencia, es razonable argumentar que la resistencia a la tensión es una función de la concentración de madera dura en la pulpa.

2. Hipótesis Estadística

Siendo $i, j \in [5, 10, 15, 20]$ referencias a las concentraciones de cada grupo, las hipótesis planteadas serían las siguientes.

$$H_0 := \forall (i, j), \mu_i = \mu_j$$

$$H_A := \exists (i, j), \mu_i \neq \mu_j$$

En términos directos, la hipótesis nula implica que las medias entre grupos permanece constante y puede decirse que son las mismas. En el caso contrario, no se tiene suficiente evidencia estadística para afirmar que todas las medias son iguales. La hipótesis nula en este contexto negaría que la resistencia es una función de la concentración, mientras la hipótesis alterna indica que no se tiene suficiente evidencia estadística para afirmar la falta de una relación entre variables.

3. ANOVA: Suma de Cuadrados Medios

```
overall_mean = mean(df$Resistencia)
group_means = summarize(group_by(df, Concentracion), Mean = mean(Resistencia), Count = n())

# Errors
SSE = sum((df$Resistencia - group_means$Mean[match(df$Concentracion, group_means$Concentracio
n)])**2)

# Groups

SSR = sum((group_means$Mean - overall_mean)**2 * group_means$Count)

# Total
SST = SSR + SSE

k = length(group_means$Concentracion)
n = length(df$Resistencia)

df_groups = k - 1
df_error = n - k
df_total = n - 1

MSE = SSE / df_error
MSR = SSR / df_groups

f_stat = MSR / MSE

print(paste("SSE: ", SSE))
```

```
## [1] "SSE:  130.166666666667"
```

```
print(paste("SSR: ", SSR))
```

```
## [1] "SSR:  382.791666666667"
```

```
print(paste("SST: ", SST))
```

```
## [1] "SST:  512.958333333333"
```

```
print(paste("df Grupos: ", df_groups))
```

```
## [1] "df Grupos:  3"
```

```
print(paste("df Error: ", df_error))
```

```
## [1] "df Error:  20"
```

```
print(paste("df Total: ", df_total))
```

```
## [1] "df Total:  23"
```

```
# print(k)
# print(n)

print(paste("MSE: ", MSE))
```

```
## [1] "MSE:  6.50833333333333"
```

```
print(paste("MSR: ", MSR))
```

```
## [1] "MSR:  127.597222222222"
```

```
print(paste("F: ", f_stat))
```

```
## [1] "F:  19.6052069995732"
```

4. ANOVA en R

```
modelo = aov(Resistencia ~ Concentracion, df)
summary(modelo)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## 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
```

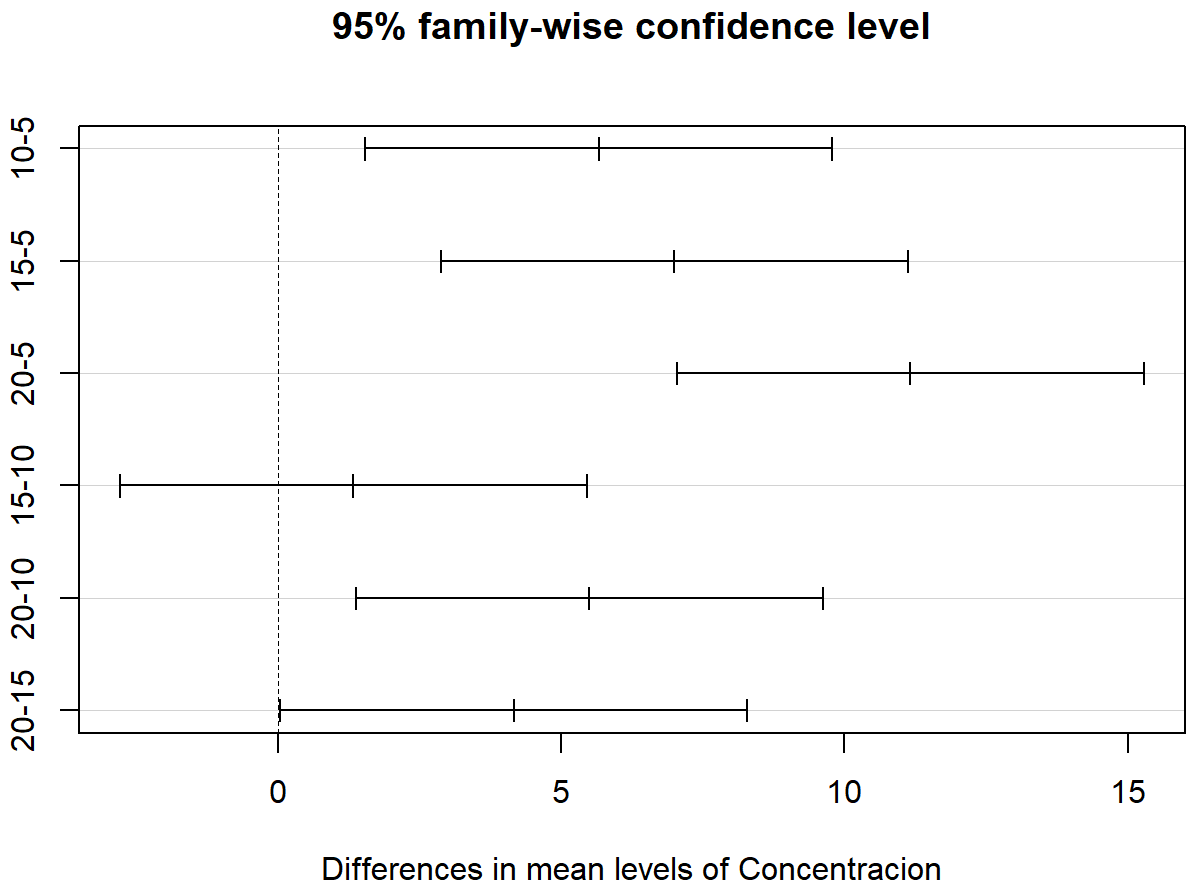
Tanto el cálculo manual como el resultado proporcionado por el método aov() de R, se tiene un estadístico de F de 19.61 que corresponde a un valor de p igual a 3.6×10^{-6} , indicando que se tiene evidencia estadística suficiente para afirmar que al menos un par de medias muestrales no son equivalentes.

5. Diferencias por Pares

```
TukeyHSD(modelo)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Resistencia ~ Concentracion, data = 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
```

```
plot(TukeyHSD(modelo, conf.level = 0.95))
```



6. Validación de Supuestos

6.1 Normalidad de los Errores

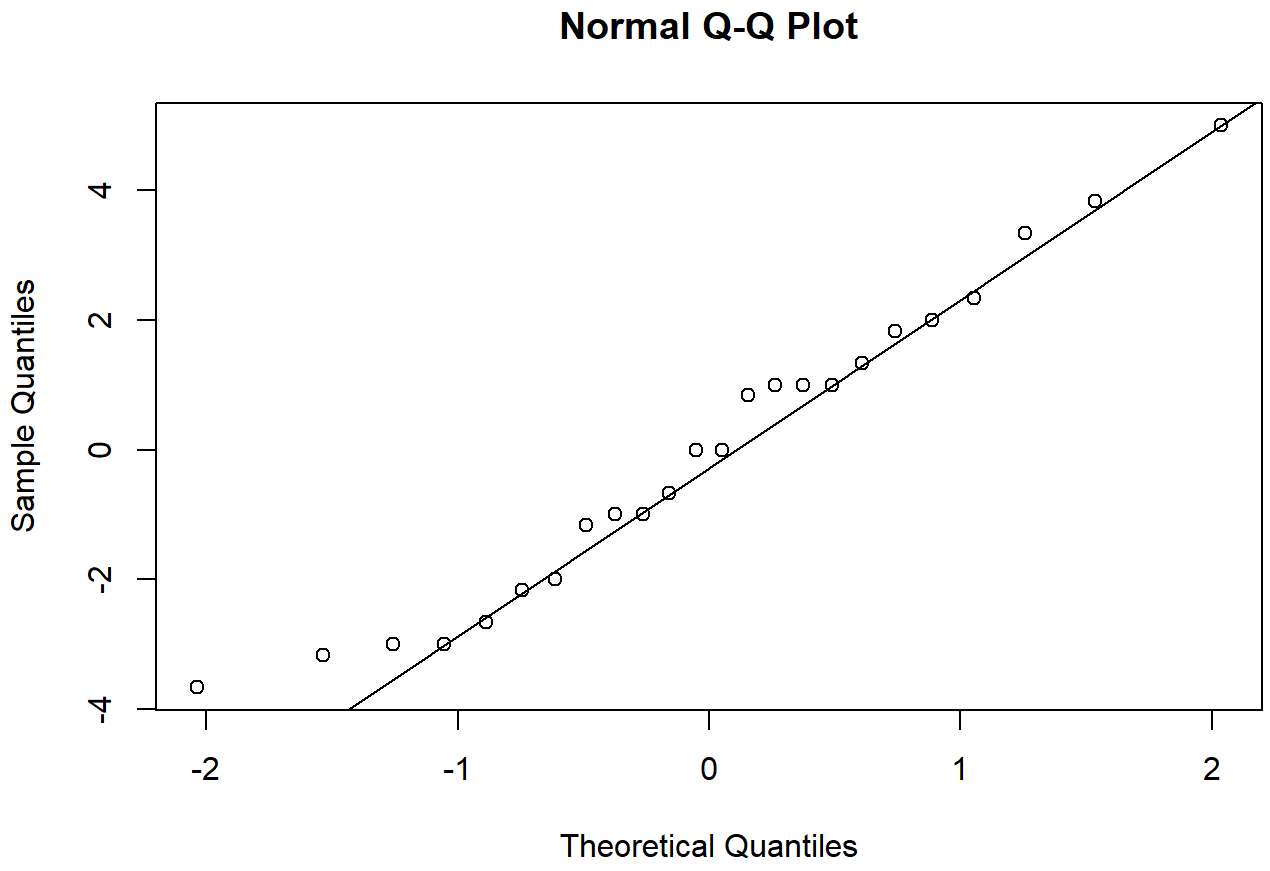
Shapiro-Wilk normality test

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

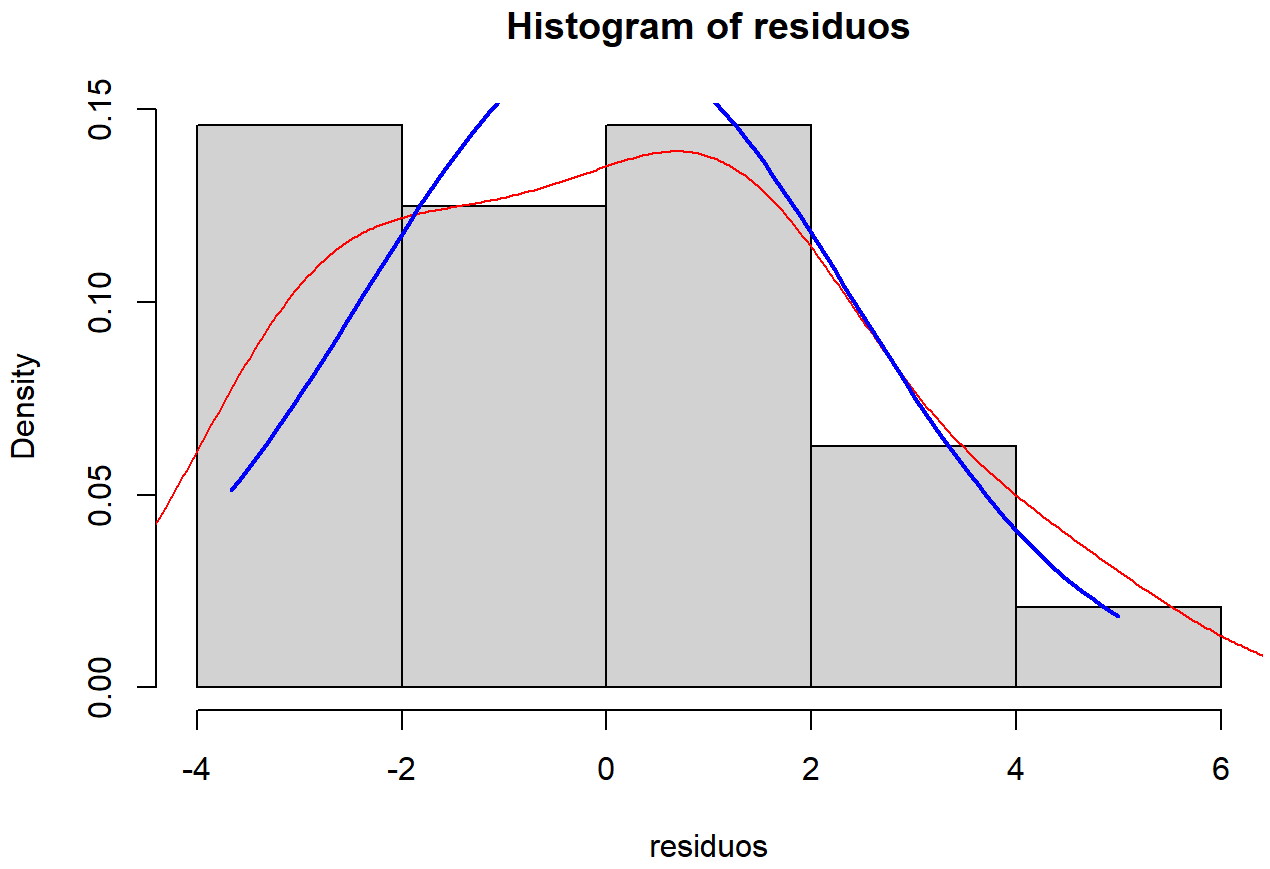
```
residuos = modelo$residuals

shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.96624, p-value = 0.5757
```

```
qqnorm(residuos)
qqline(residuos)
```




```
hist(residuos,freq=FALSE)
lines(density(residuos),col="red")
curve(dnorm(x,mean=mean(residuos), sd=sd(residuos)), from=min(residuos), to=max(residuos), add=T
RUE, col="blue",lwd=2)
```



6.2 Homocedasticidad

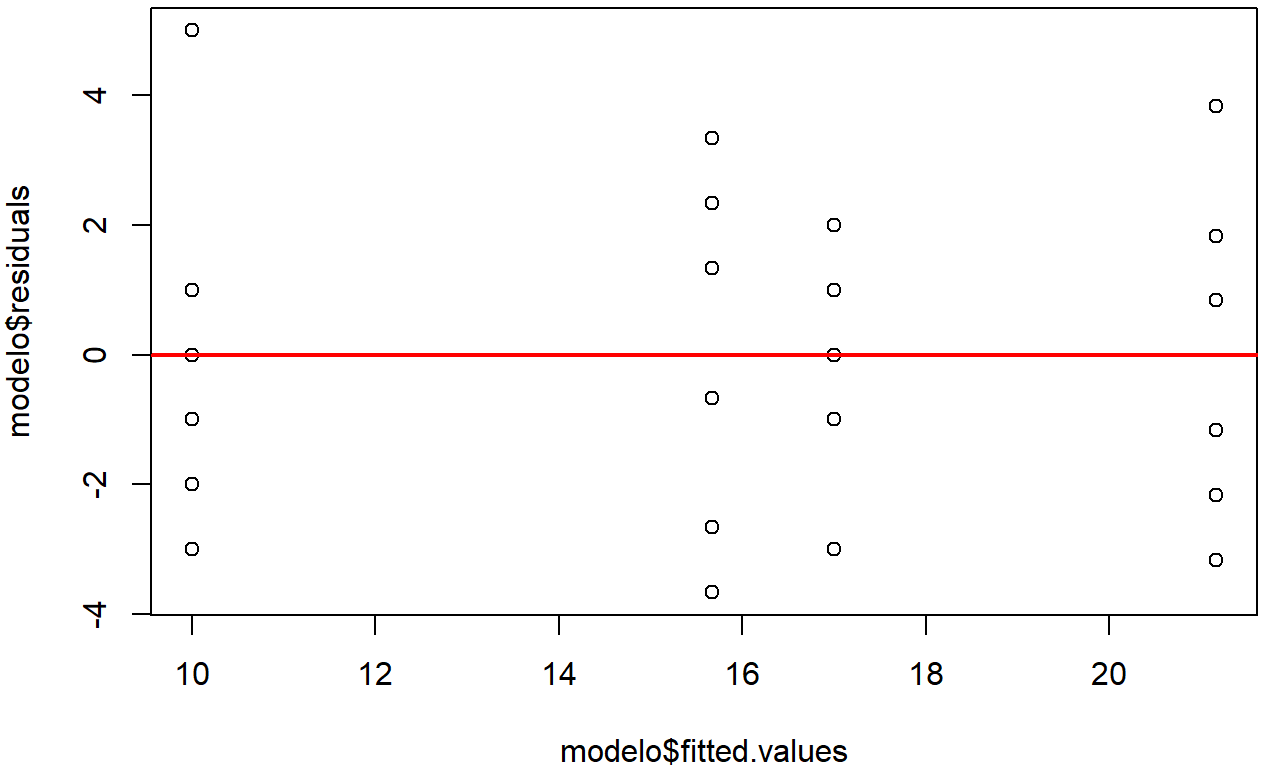
Bartlett Test of Homogeneity of Variances

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

```
bartlett.test(df$Resistencia, df$Concentracion)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df$Resistencia and df$Concentracion
## Bartlett's K-squared = 1.1352, df = 3, p-value = 0.7686
```

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



6.3 Independencia

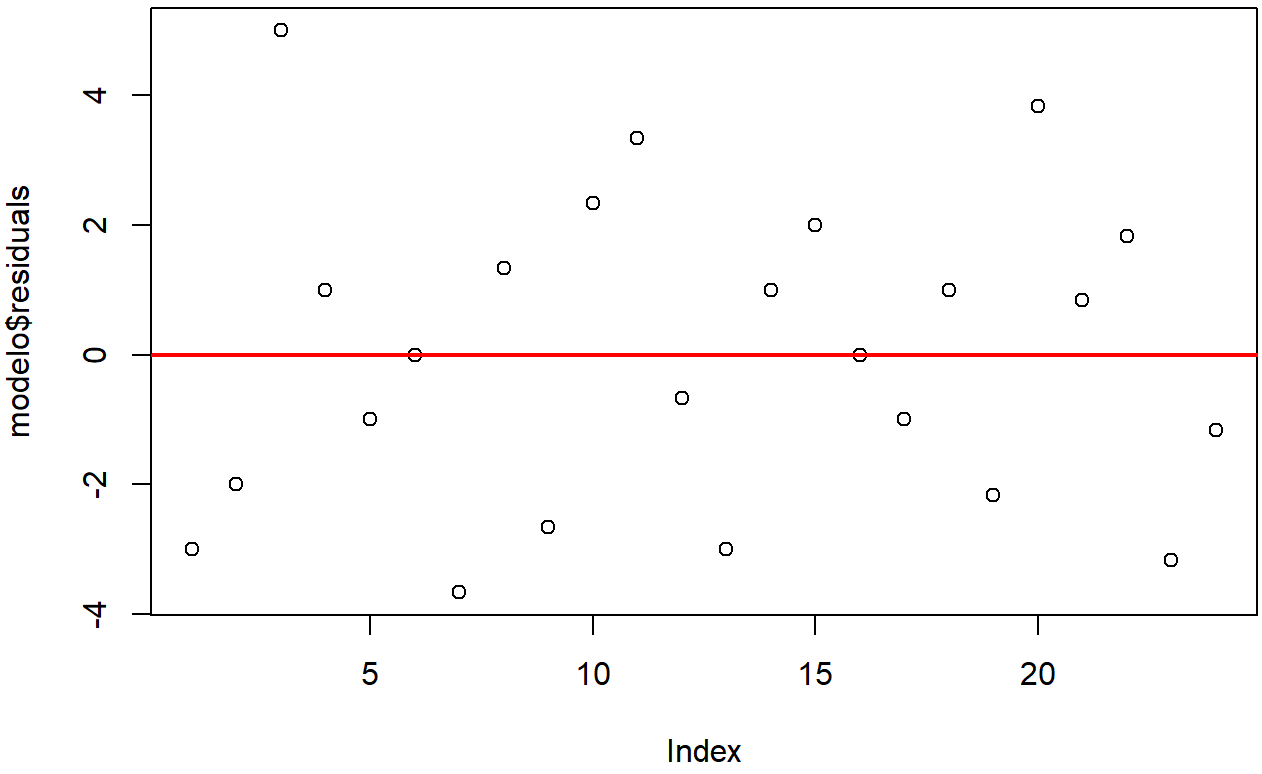
Durbin Watson

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

```
dwtest(modelo)
```

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

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



Ya que en ninguna de las pruebas se tiene un valor de p menor a 0.05, no se pueden rechazar las hipótesis nulas con las que verificamos que los supuestos se cumplen.

7. Intervalos de Confianza por Nivel

```
t_test_result <- df %>%
  group_by(Concentracion) %>%
  summarize(
    Conf_Lower = t.test(Resistencia)$conf.int[1],
    Mean = mean(Resistencia),
    Conf_Upper = t.test(Resistencia)$conf.int[2]
  )

t_test_result
```

```
## # A tibble: 4 × 4
##   Concentracion Conf_Lower  Mean Conf_Upper
##   <fct>              <dbl> <dbl>      <dbl>
## 1 5                   7.03  10         13.0
## 2 10                 12.7   15.7       18.6
## 3 15                 15.1   17         18.9
## 4 20                 18.4   21.2       23.9
```

8. Conclusión

El análisis de varianza ANOVA demostró que la concentración de madera dura en la pulpa afecta significativamente la resistencia a la tensión del papel, con un valor F de 19.61 y un valor p de 3.6×10^{-6}. El test post-hoc de Tukey reveló que las concentraciones más altas, especialmente el 20%, presentan una resistencia significativamente mayor comparado con las concentraciones más bajas. Las pruebas de validación indicaron que los supuestos del ANOVA se cumplieron adecuadamente. Estos resultados sugieren que aumentar la concentración de madera dura puede mejorar la resistencia del papel, lo que es relevante para optimizar la calidad del producto.