Anova

Arturo

2023-08-25

Autor: Arturo Garza Campuzano

Matrícula: A00828096

Anova

El rendimiento

En un instituto se han matriculado 36 estudiantes. Se desea explicar el rendimiento de ciencias naturales en función de dos variables: género y metodología de enseñanza. La metodología de enseñanza se analiza en tres niveles: explicación oral y realización del experimento (1er nivel) explicación oral e imágenes (2º nivel) y explicación oral (tercer nivel).

En los alumnos matriculados había el mismo número de chicos que de chicas, por lo que formamos dos grupos de 18 sujetos; en cada uno de ellos, el mismo profesor aplicará a grupos aleatorios de 6 estudiantes las 3 metodologías de estudio. A fin de curso los alumnos son sometidos a la misma prueba de rendimiento. Los resultados son los siguientes: https://drive.google.com/file/d/1W5u4rtusxCZ_v3z7PBwjS7zcAkcXrwYL/view?usp=sharing

¿Existe alguna influencia de la metodología de enseñanza y el género de los estudiantes en el rendimiento de los estudiantes?

Cargar datos

```r
# Introduciendo los datos
rendimiento = c(10,7,9,9,9,10,5,7,6,6,8,4,2,6,3,5,5,3,9,7,8,8,10,6,8,3,5,6,7,7,2,6,2,1,4,3)
metodo = c(rep("M1",6),rep("M2",6),rep("M3",6),rep("M1",6),rep("M2",6),rep("M3",6))
sexo = c(rep("h", 18), rep("m",18))
metodo = factor(metodo)
sexo = factor(sexo)
```

1. Establece las hipótesis estadísticas (tienen que ser 3).

Los factores que involucran al modelo son:

- F1: Método de enseñanza.
- F2: Sexo.

Bajo este modelo se tienen las siguientes hipótesis:

- Primera hipótesis:
  - $H_0 : \tau_i = 0$
  - $H_1 :$ algún $\tau_i \neq 0$
- Segunda hipótesis:
  - $H_0 : \alpha_j = 0$
  - $H_1 :$ algún $\alpha_j \neq 0$
- Tercera hipótesis:

- $H_0 : \tau_i \alpha_j = 0$
- H_1 : algún $\tau_i \alpha_j \neq 0$

2. Realiza el ANOVA para dos niveles con interacción.

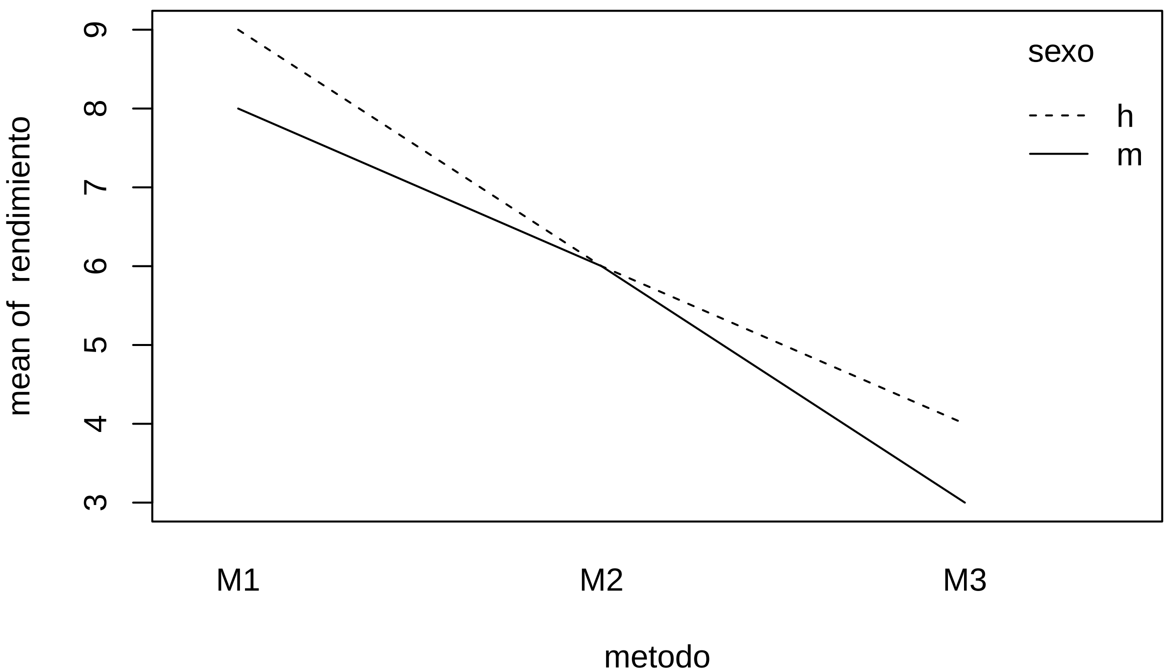
2.1 Haz la gráfica de interacción de dos factores.

Considerando a los dos factores más su interacción y algunos gráficos para observarla interacción entre los dos factores.

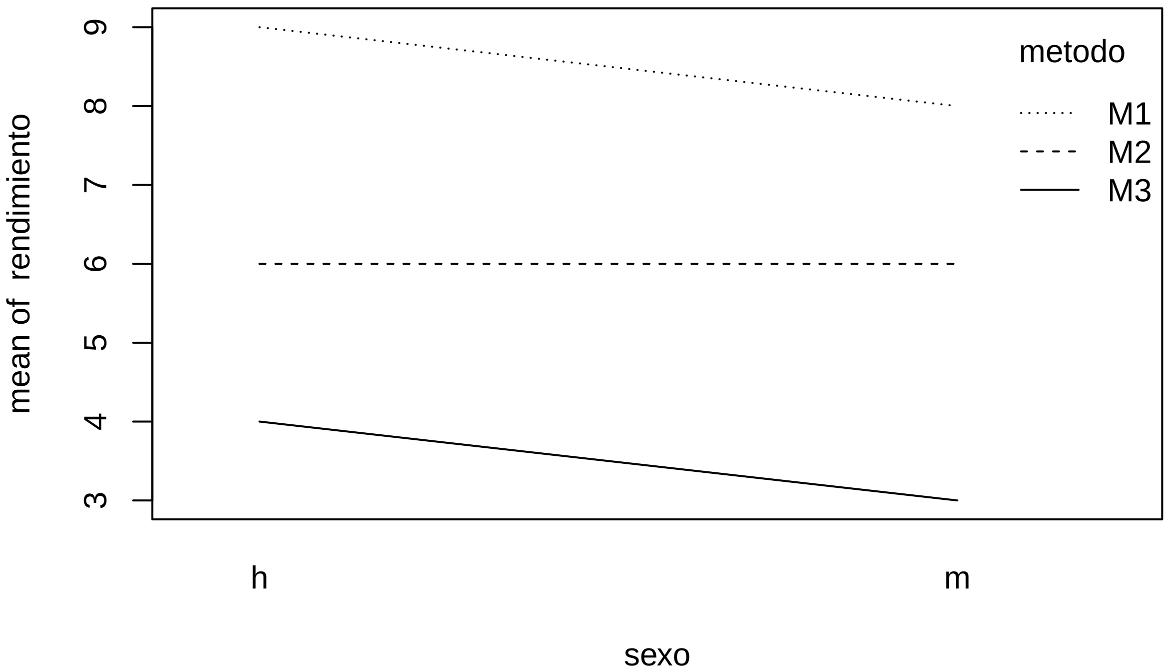
```r
A <- aov(rendimiento~metodo*sexo)
summary(A)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## metodo       2    150   75.00  32.143 3.47e-08 ***
## sexo         1      4    4.00   1.714    0.200    
## metodo:sexo  2      2    1.00   0.429    0.655    
## Residuals   30     70    2.33                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
interaction.plot(metodo,sexo,rendimiento)
```




```r
interaction.plot(sexo,metodo,rendimiento)
```



2.2 Interpreta el resultado desde la perspectiva estadística y en el contexto del problema.

Según el resumen del modelo, podemos realizar las siguientes observaciones:

- El estadístico F para el factor *metodo* es notablemente alto, y el valor p es extremadamente pequeño. Esto sugiere que el factor *metodo* tiene un efecto significativo en el rendimiento de los estudiantes.
- Por otro lado, en relación al factor *sexo*, el estadístico F es relativamente bajo, y el valor p es de 0.200. Esto indica que el factor *sexo* no tiene un efecto significativo en el rendimiento de los estudiantes.
- Además, la interacción entre los factores *metodo* y *sexo* no muestra significancia. El estadístico F es bajo, y el valor p es alto en este caso. Esto sugiere que no existen pruebas sólidas para concluir que la relación entre el método de enseñanza y el rendimiento varía significativamente según el sexo de los estudiantes.

En cuanto a las gráficas de interacción, podemos observar que tanto con el primer método de enseñanza como con el tercero, la diferencia en el rendimiento entre los sexos es la misma. Mientras que con el segundo método, se aprecia un rendimiento similar entre ambos sexos.

2.3 Escribe tus conclusiones parciales.

Después de realizar el ANOVA con el modelo completo, encontramos que $\tau_i \alpha_j$ fue no significativa (no hay efecto de interacción). No se rechaza la tercera hipótesis nula (H_0) y el modelo se reduce.

3. Realiza el ANOVA para dos niveles sin interacción.

En el modelo, se consideran sólo los efectos principales. Ya no se usa *, se usa +.

```
B<-aov(rendimiento~metodo+sexo)
summary(B)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## metodo       2    150   75.00  33.333 1.5e-08 ***
## sexo         1      4    4.00   1.778   0.192    
## Residuals   32     72    2.25                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.1 Haz el boxplot de rendimiento por sexo. Calcula la media para el rendimiento por sexo y método.

Para observar mejor los efectos de los factores principales, se calcula la media por nivel y se grafica por nivel. También se calcula la media general.

```
tapply(rendimiento,sexo,mean)
```

```
##        h        m
## 6.333333 5.666667
```

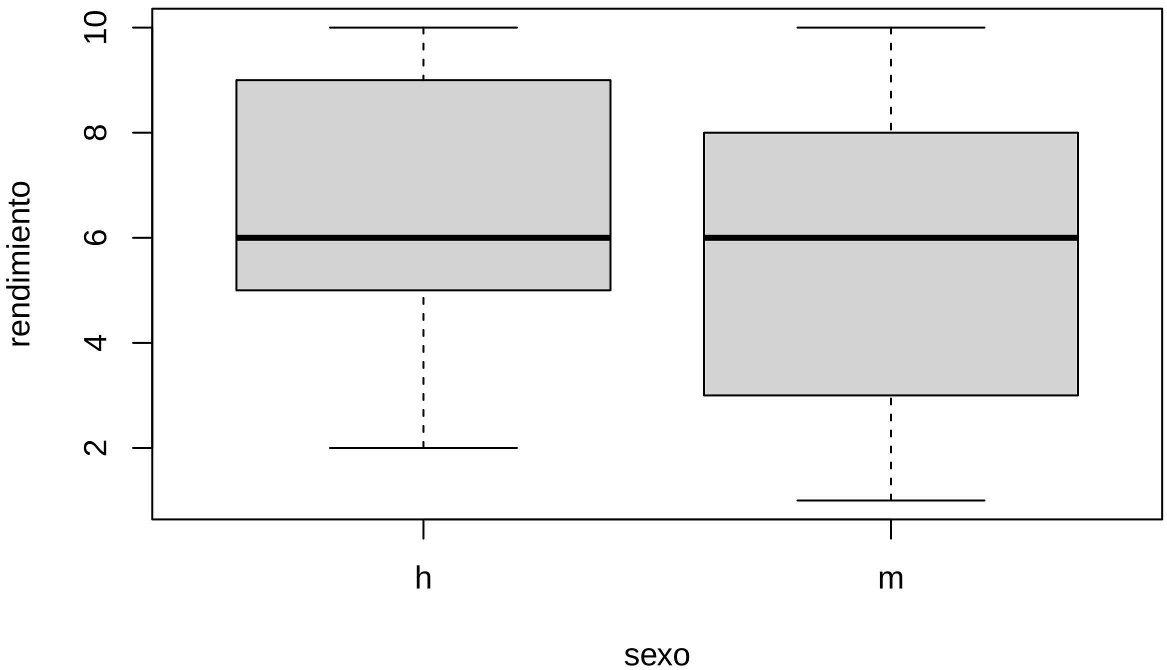
```
tapply(rendimiento,metodo,mean)
```

```
##  M1  M2  M3
## 8.5 6.0 3.5
```

```
M=mean(rendimiento)
M
```

```
## [1] 6
```

```
boxplot(rendimiento ~ sexo)
```



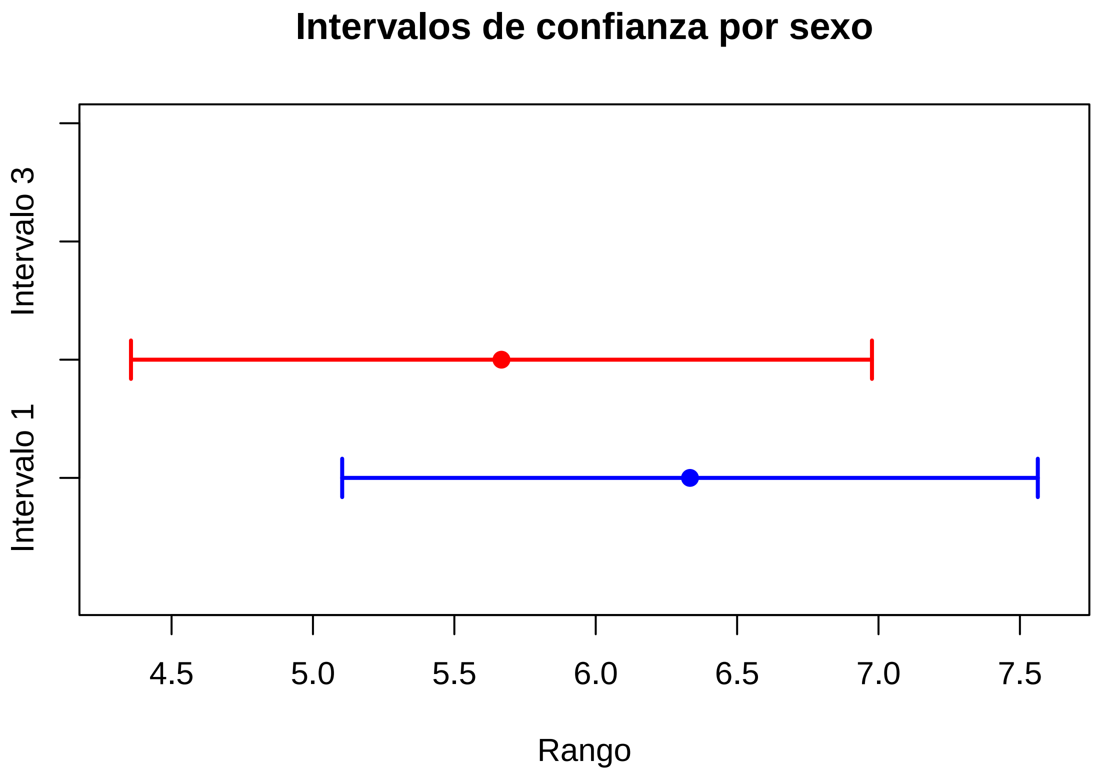
3.2 Haz los intervalos de confianza de rendimiento por sexo. Grafícalos.

```
# Intervalo de confianza para hombres
ci_hombres <- t.test(rendimiento[sexo == "h"])$conf.int
ci_mujeres <- t.test(rendimiento[sexo == "m"])$conf.int

# Obtener el valor máximo de los segmentos en y
max_y_ic <- max(ci_hombres[2], ci_mujeres[2])
min_y_ic <- min(ci_hombres[1], ci_mujeres[1])

plot(0, ylim=c(0, 4), xlim=c(min_y_ic -0.05, max_y_ic +0.05), yaxt="n", xlab="Rango", ylab="", main="In
axis(2, at=1:4, labels=paste("Intervalo", 1:4))
arrows(ci_hombres[1], 1, ci_hombres[2], 1, angle=90, code=3, length=0.1, lwd=2, col="blue")
arrows(ci_mujeres[1], 2, ci_mujeres[2], 2, angle=90, code=3, length=0.1, lwd=2, col="red")
```

```r
points(mean(ci_hombres), 1, pch=19, cex=1.1, col="blue")
points(mean(ci_mujeres), 2, pch=19, cex=1.1, col="red")
```



3.3 Interpreta el resultado desde la perspectiva estadística y en el contexto del problema.

De acuerdo con el resumen del modelo, se pueden realizar las siguientes observaciones:

- En relación al factor *metodo*, el estadístico F es considerablemente alto, y el valor p es extraordinariamente bajo. Esto sugiere que el factor *metodo* tiene un efecto significativo en el rendimiento de los estudiantes.
- Por otro lado, al considerar el factor *sexo*, el estadístico F es relativamente bajo, y el valor p es 0.192. Esto indica que el factor *sexo* no tiene un efecto significativo en el rendimiento de los estudiantes.

Al examinar tanto los diagramas de caja y bigotes como los intervalos de confianza, se aprecia que, en ambos casos, los niveles de rendimiento se superponen. Sin embargo, es notable que el rendimiento de los hombres es ligeramente superior al de las mujeres.

3.4 Escribe tus conclusiones parciales.

Después de realizar el ANOVA con el modelo con los efectos principales, encontramos que α_j fue no significativa (no hay efecto del sexo). No se rechaza la segunda hipótesis nula (H_0) y el modelo se reduce.

4. Realiza el ANOVA para un efecto principal.

```r
C<-aov(rendimiento~metodo)
summary(C)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## metodo       2    150    75.0   32.57 1.55e-08 ***
## Residuals   33     76     2.3                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.1 Haz el boxplot de rendimiento por método de enseñanza. Calcula la media.

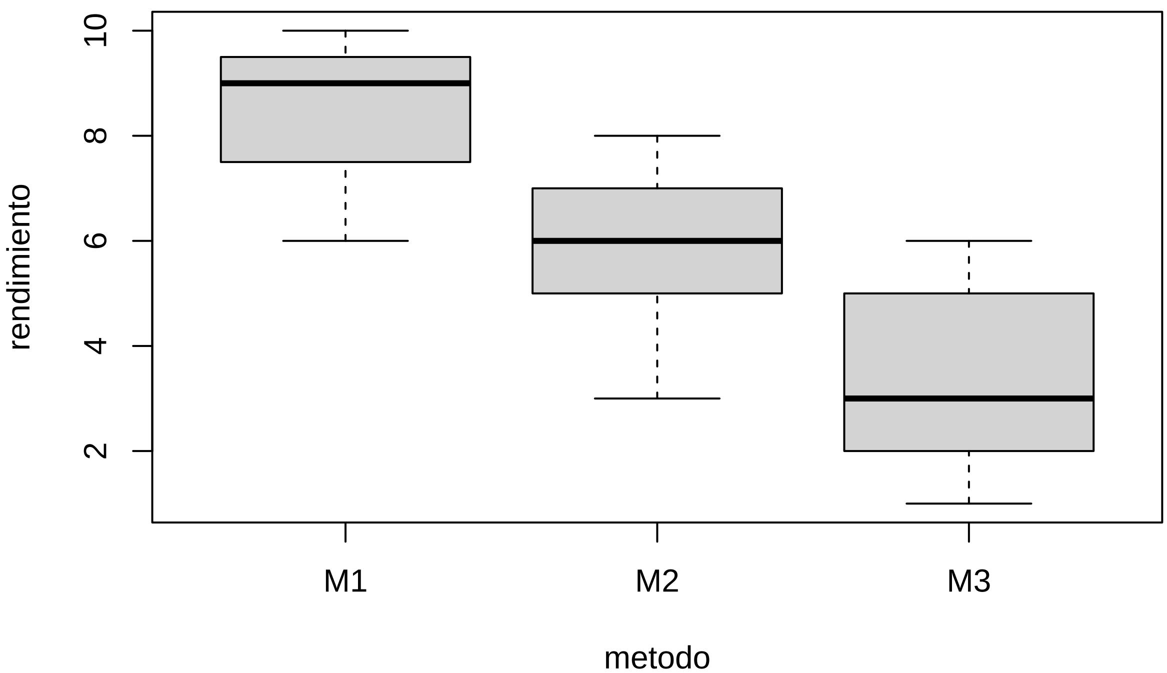
```r
tapply(rendimiento,metodo,mean)
```

```
##  M1  M2  M3
## 8.5 6.0 3.5
```

```r
mean(rendimiento)
```

```
## [1] 6
```

```r
boxplot(rendimiento ~ metodo)
```

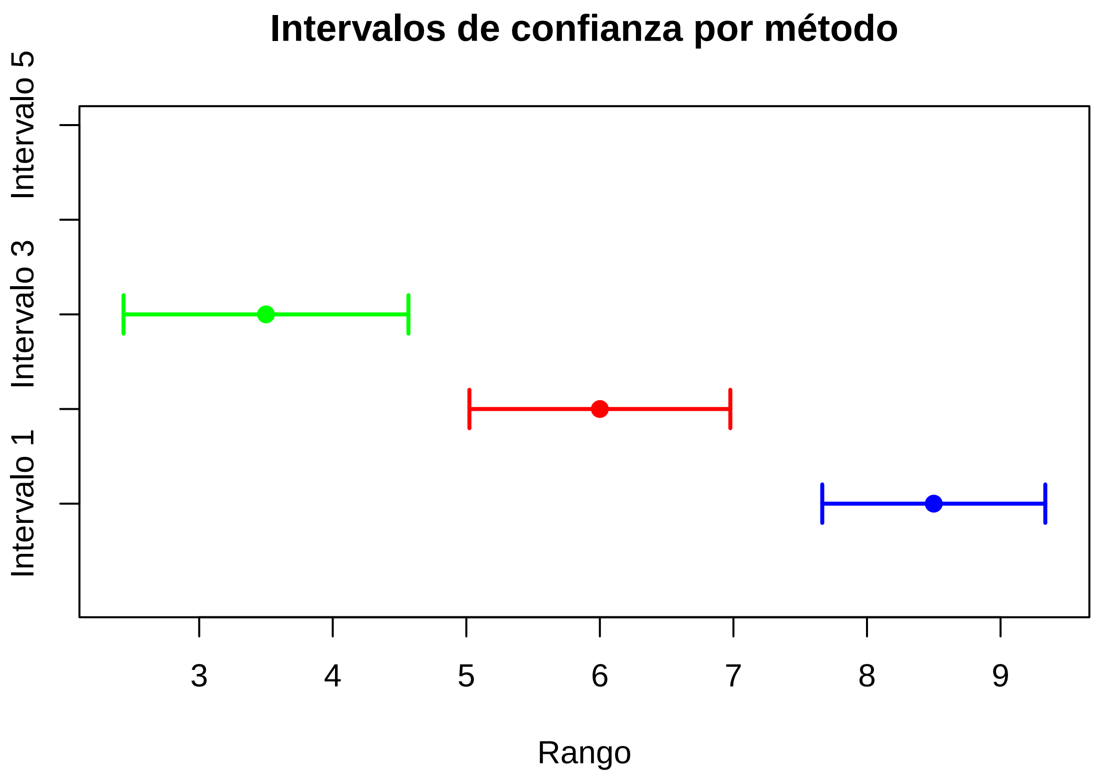


4.2 Haz los intervalos de confianza de rendimiento por método. Grafícalos.

```r
# Intervalo de confianza para hombres
ci_m1 <- t.test(rendimiento[metodo == "M1"])$conf.int
ci_m2 <- t.test(rendimiento[metodo == "M2"])$conf.int
ci_m3 <- t.test(rendimiento[metodo == "M3"])$conf.int

# Obtener el valor maximo de los segmentos en y
max_y_ic <- max(ci_m1[2], ci_m2[2], ci_m3[2])
min_y_ic <- min(ci_m1[1], ci_m2[1], ci_m3[1])

plot(0, ylim=c(0, 5), xlim=c(min_y_ic -0.05, max_y_ic +0.05), yaxt="n", xlab="Rango", ylab="", main="In
axis(2, at=1:5, labels=paste("Intervalo", 1:5))
arrows(ci_m1[1], 1, ci_m1[2], 1, angle=90, code=3, length=0.1, lwd=2, col="blue")
arrows(ci_m2[1], 2, ci_m2[2], 2, angle=90, code=3, length=0.1, lwd=2, col="red")
arrows(ci_m3[1], 3, ci_m3[2], 3, angle=90, code=3, length=0.1, lwd=2, col="green")
points(mean(ci_m1), 1, pch=19, cex=1.1, col="blue")
points(mean(ci_m2), 2, pch=19, cex=1.1, col="red")
points(mean(ci_m3), 3, pch=19, cex=1.1, col="green")
```

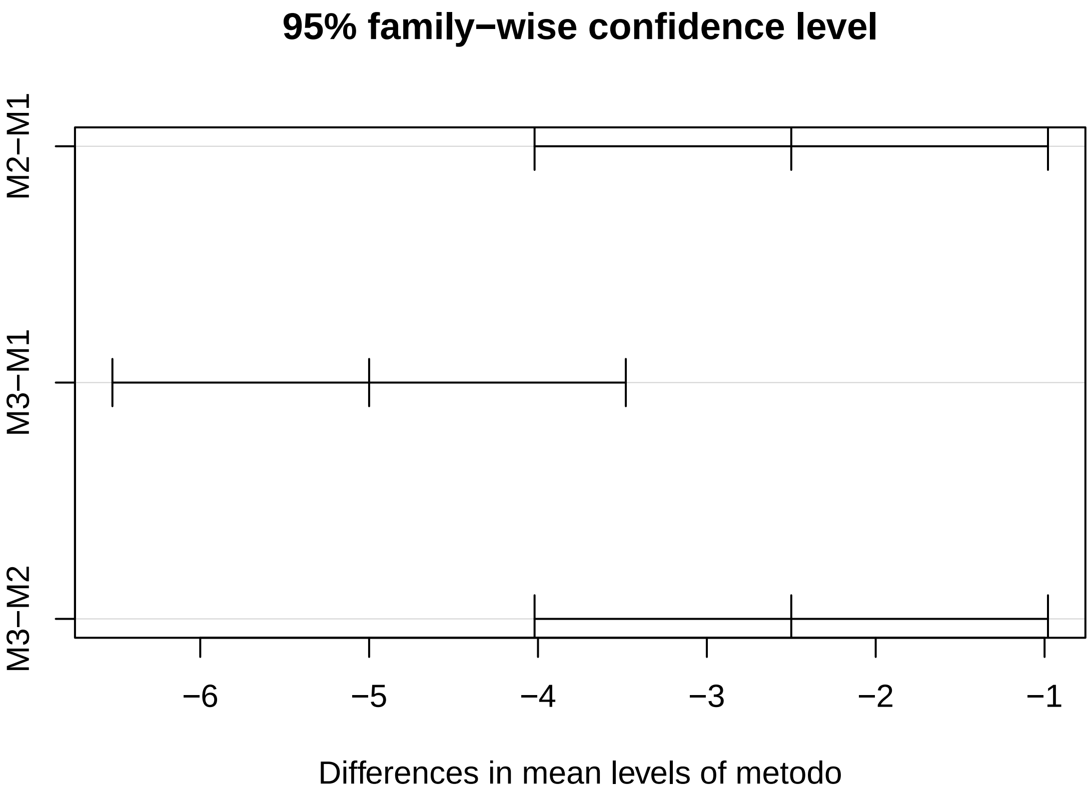


4.3 Realiza la prueba de comparaciones múltiples de Tukey. Grafica los intervalos de confianza de Tukey.

```
I = TukeyHSD(aov(rendimiento ~ metodo))
I
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rendimiento ~ metodo)
##
## $metodo
##       diff       lwr        upr     p adj
## M2-M1 -2.5 -4.020241 -0.9797592 0.0008674
## M3-M1 -5.0 -6.520241 -3.4797592 0.0000000
## M3-M2 -2.5 -4.020241 -0.9797592 0.0008674
```

```
plot(I) #Los intervalos de confianza se observan mejor si se grafican
```



4.4 Interpreta el resultado desde la perspectiva estadística y en el contexto del problema.

Se puede notar que el factor *metodo* exhibe un estadístico F bastante alto y un valor p extremadamente pequeño. Estos resultados sugieren firmemente que este factor tiene un efecto significativo en el rendimiento de los estudiantes.

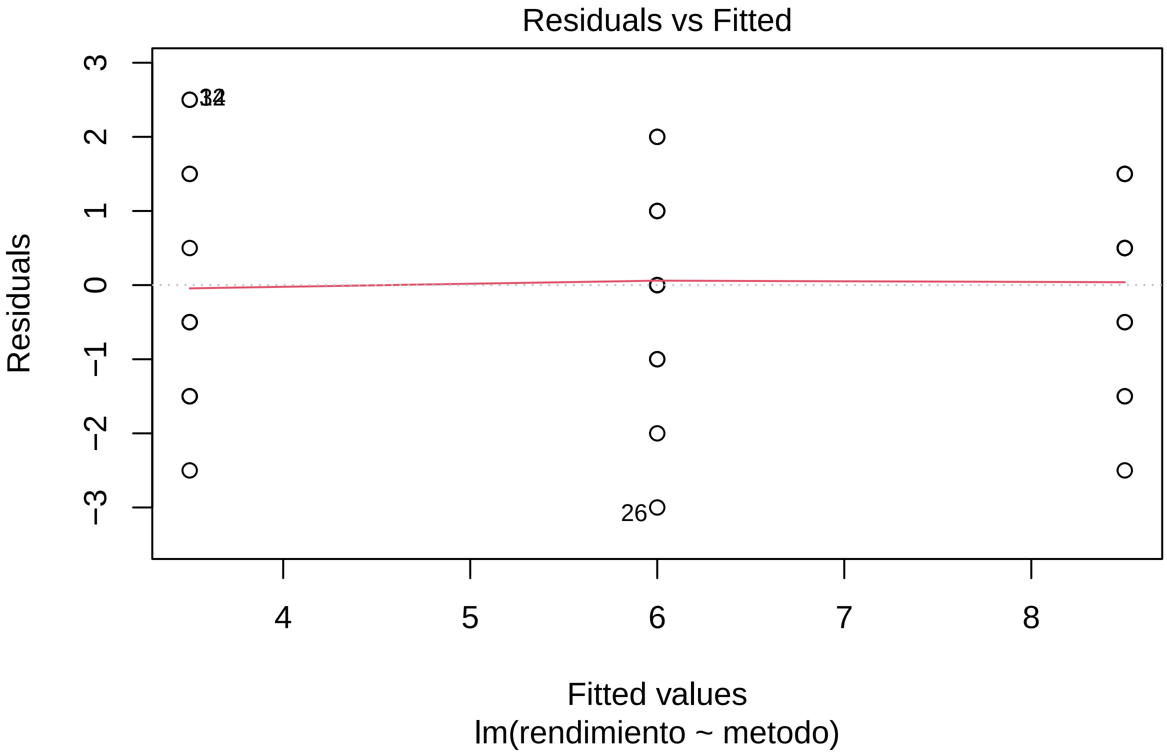
Al analizar tanto los diagramas de caja y bigotes como los intervalos de confianza, se evidencia que cada método presenta un rendimiento distinto.

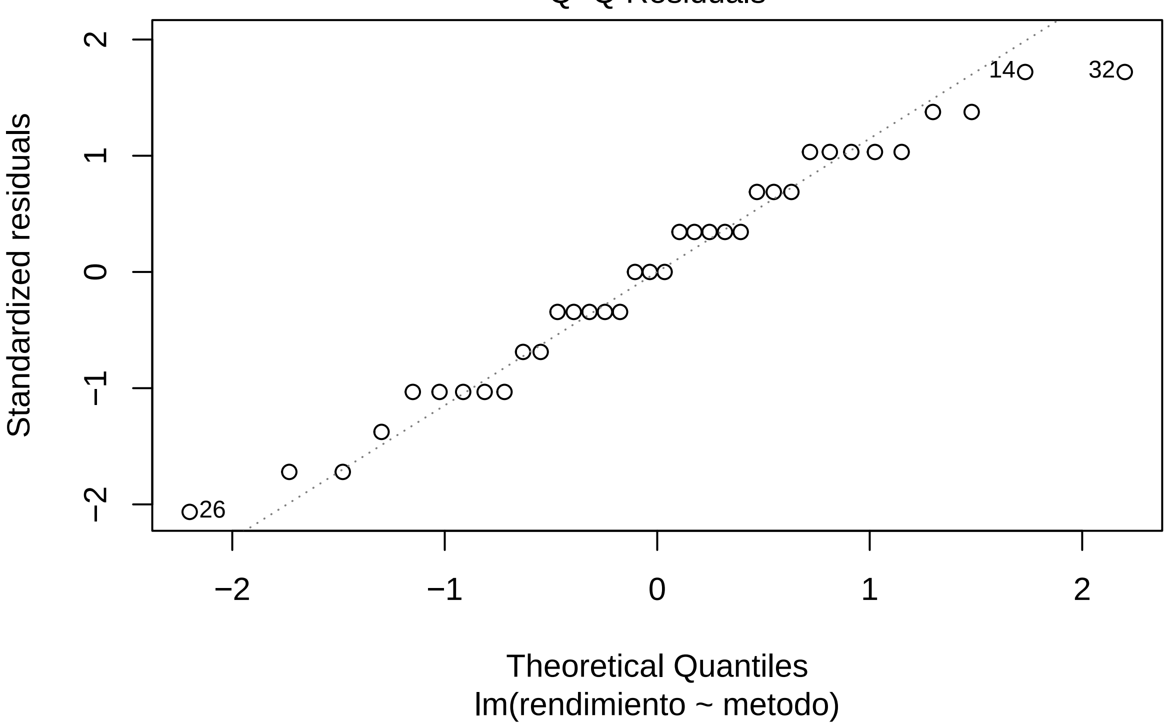
4.5 Escribe tus conclusiones parciales.

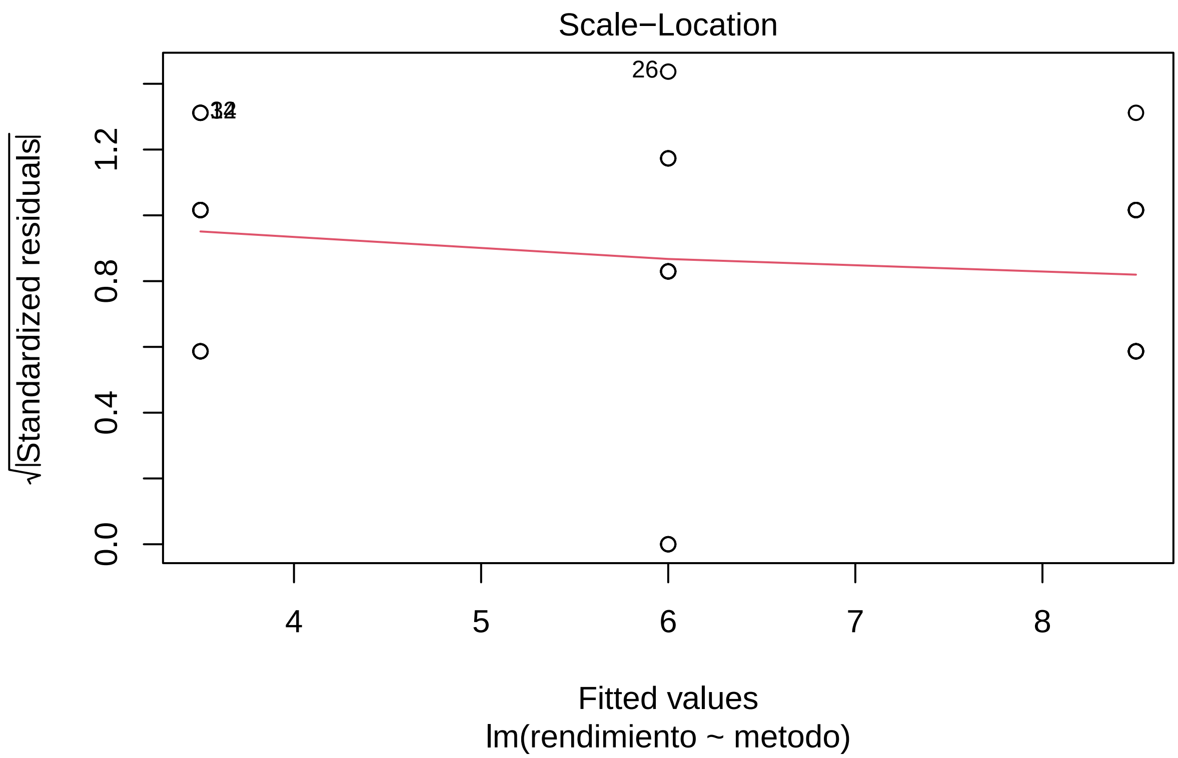
Se concluye que algún $\tau_i \neq 0$.

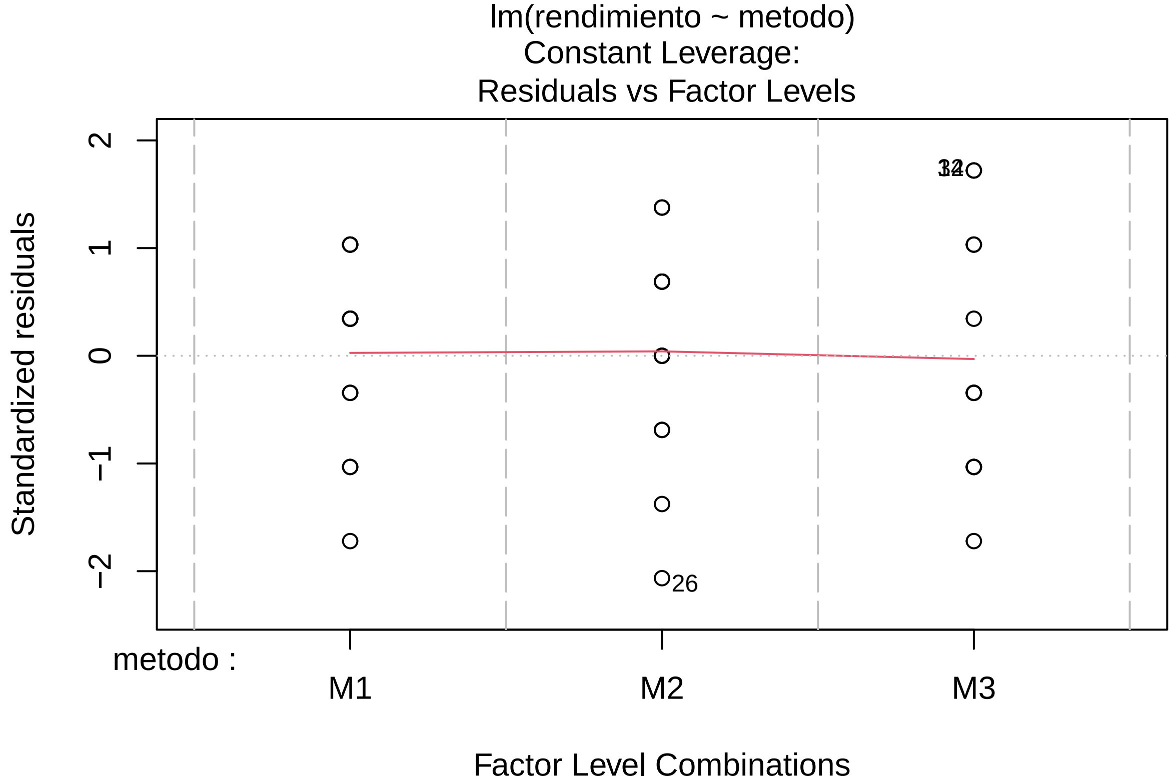
5. Comprueba la validez del modelo.

```
# Normailidad, homocedasticidad e independencia
plot(lm(rendimiento~metodo))
```


Residuals vs Fitted
Residuals
Fitted values
lm(rendimiento ~ metodo)
26


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(rendimiento ~ metodo)
14
32
26






```r
CD= 150/(150+76)
cat("El coeficiente de determinación es",CD)
```

```
## El coeficiente de determinación es 0.6637168
```

6. Concluye en el contexto del problema.

¿Se observaron efectos significativos? Si es así, ¿todos son diferentes? ¿Existen diferencias significativas entre los grupos? ¿Qué implicaciones tiene esto en el contexto del problema?

En nuestro análisis, encontramos que los tres métodos de enseñanza tienen un impacto significativamente diferente en el rendimiento de los estudiantes.

- El Método 3 muestra un efecto negativo, ya que disminuye el rendimiento en comparación con la media general.
- El Método 2, en contraste, no tiene un efecto significativo; es decir, no modifica el rendimiento de los estudiantes de manera apreciable.
- Finalmente, el Método 1 tiene un efecto positivo, ya que aumenta el rendimiento en comparación con la media general, lo que lo convierte en el método de enseñanza más efectivo.

El modelo utilizado logra explicar el 66.37% de la variación observada en el rendimiento. Esto subraya la importancia del método de enseñanza como un factor determinante en el rendimiento de los estudiantes, ya que fue el único factor significativo en el modelo. Sin embargo, el 32.73% restante de la variación podría deberse a otros factores no considerados en este estudio o simplemente a la aleatoriedad (error).

Es relevante destacar que el diseño del estudio fue equilibrado, con un número igual de datos en cada tratamiento, lo que lo hace robusto ante posibles problemas de heterocedasticidad.

Además, nuestros análisis de gráficos Q-Q y de residuos vs. valor esperado (ajustado) sugieren que los datos parecen cumplir con los supuestos de normalidad e independencia. Los errores presentan una media cercana a cero y una varianza constante, lo que respalda la validez de nuestro modelo.