6-anova

August 25, 2023

1 El rendimiento

Francisco Mestizo Hernández A01731549

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:

[53]:
```r
#Generamos un data frame con 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))
metodo = factor(metodo)
sexo = c(rep("H", 18), rep("M", 18))
sexo = factor(sexo)

M <- data.frame(Rendimiento = rendimiento, Metodo = metodo, Sexo = sexo)
head(M)
```

A data.frame: 6 × 3

	Rendimiento <dbl>	Metodo <fct>	Sexo <fct>
1	10	M1	H
2	7	M1	H
3	9	M1	H
4	9	M1	H
5	9	M1	H
6	10	M1	H

1.1 Planteamiento de las hipótesis estadísticas

Podemos ver que las tres variables que se están tomando en cuenta para el estudio son el rendimiento (la variable dependiente), el método de enseñanza y el sexo de los estudiantes.

Por lo tanto, tenemos dos factores en este problema. El método (τ) y el sexo (α). Además, para el modelo tenemos que considerar la interacción entre esas dos variables. Por lo tanto podemos plantear las siguientes hipótesis:

Primera hipótesis

$H_0 : \tau_i = 0$

H_1 : algún τ_i es distinto de cero

Segunda hipótesis

$H_0 : \alpha_i = 0$

H_1 : algún α_i es distinto de cero

Tercera hipótesis

$H_0 : \tau_i \alpha_j = 0$

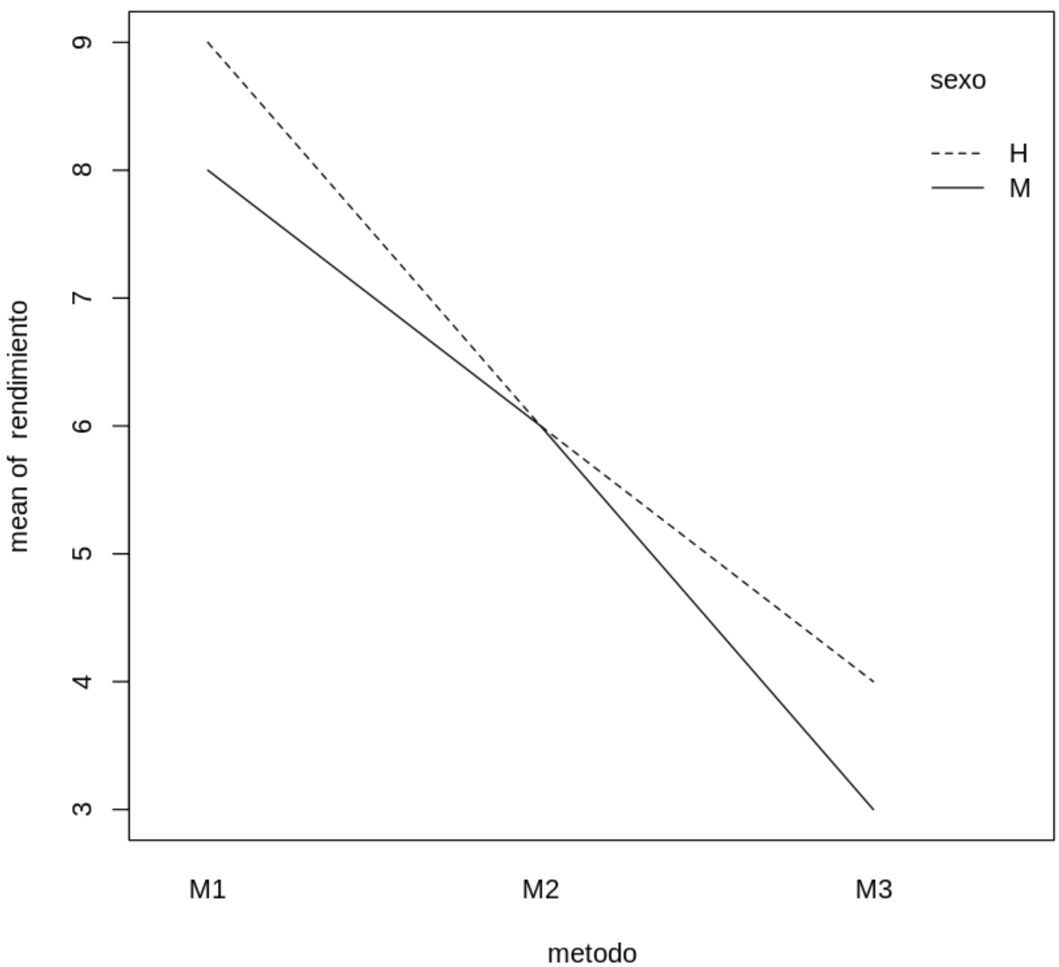
H_1 : algún $\tau_i \alpha_j$ es distinto de cero

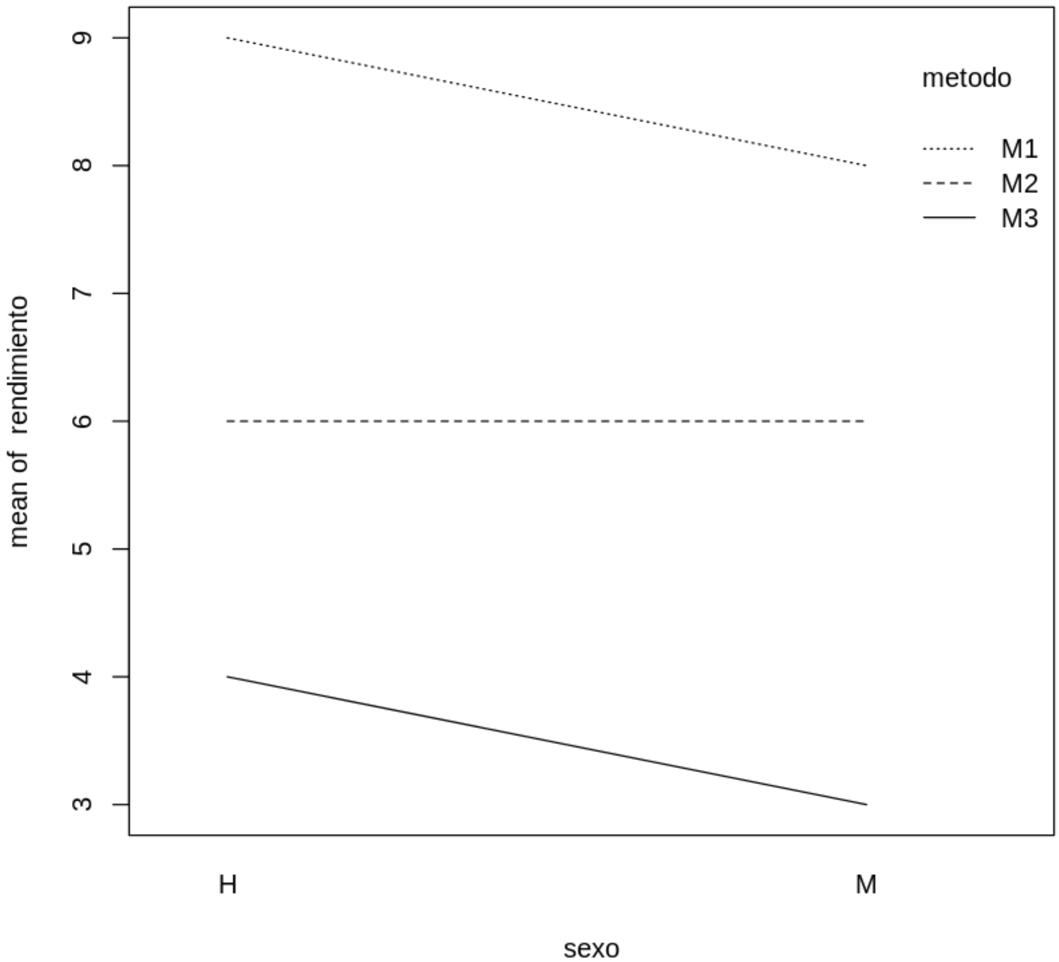
1.2 ANOVA con dos niveles y con interacción

La tercer hipótesis planteada, nos propone buscar si es significativa la interacción método-sexo para este modelo. Podemos realizar el ANOVA tomando en cuenta esta interacción:

[54]:
```
A<-aov(rendimiento~metodo*sexo)
summary(A)
interaction.plot(metodo, sexo, rendimiento)
interaction.plot(sexo, metodo, rendimiento)
```

```
            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
```


sexo
H
M
mean of rendimiento
9
8
7
6
5
4
3
M1
M2
M3
metodo



Podemos ver que en el summary, el valor p nos indica con los asteriscos cual es el mas significativo. Nos damos cuenta que la interacción metodo:sexo no es significativa ni importante, por eso tiene un valor p tan alto. Lo vamos a quitar, pero no quitaremos sexo para hacer el análisis sin la interacción. Entonces rechazamos h_0 para la tercer hipótesis planteada.

Además, esto se sustenta viendo las gráficas, ya que las lineas no tienen cruces y los metodos siguen el mismo orden para las dos sexos.

1.3 ANOVA para dos niveles sin interacción

Una vez que quitamos la interacción, podemos volver a realizar el ANOVA solamente con los dos factores independientes. Y al haber rechazado h0 en nuestra primer hipótesis, podemos ver que nos quedamos con:

Primera hipótesis

$H_0 : \tau_i = 0$

H_1 : algún τ_i es distinto de cero

Segunda hipótesis

$H_0 : \alpha_i = 0$

H_1 : algún α_i es distinto de cero

[55]:
```
B<-aov(rendimiento~metodo+sexo)
summary(B)

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

```
            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
```

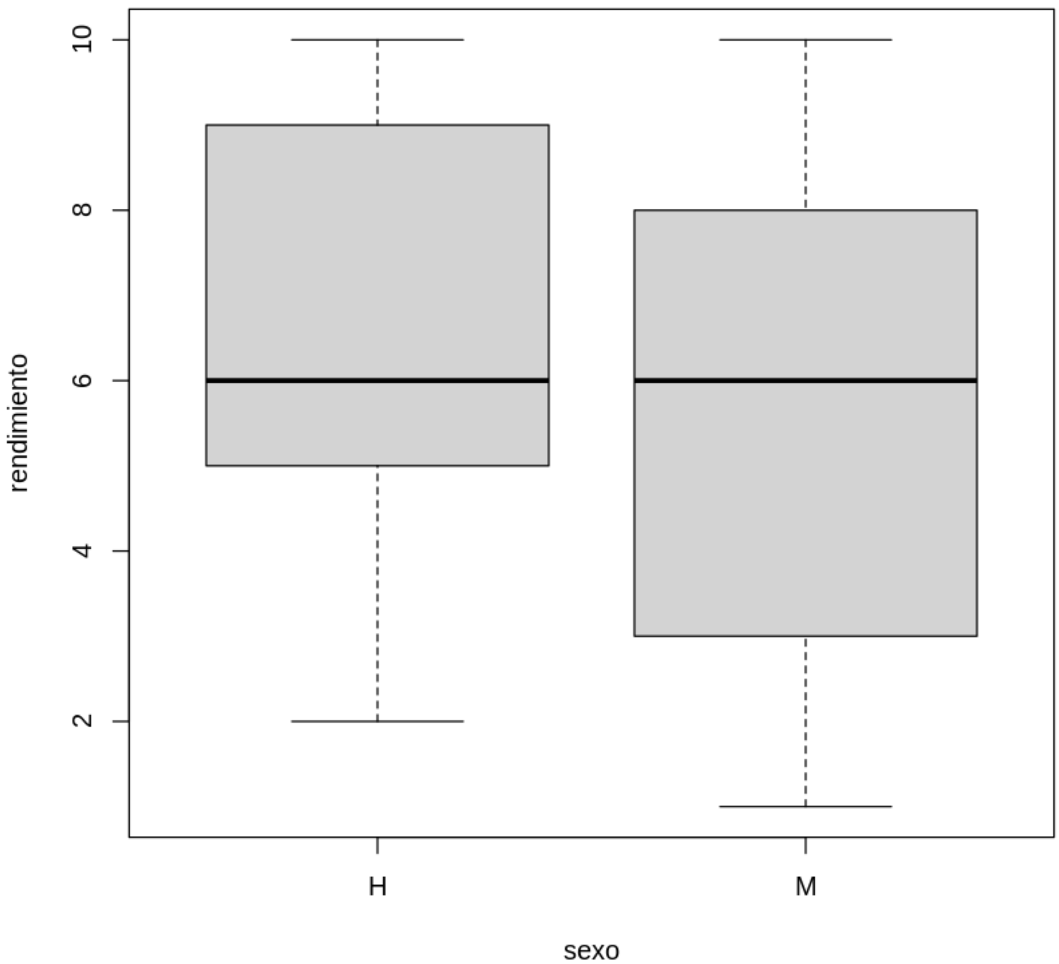
H	6.33333333333333	**M**	5.66666666666667

M1	8.5	**M2**	6	**M3**	3.5

Ahora, nos podemos dar cuenta que el valor p de sexo disminuyó al hacer el análisis sin la interacción. De todas formas el cambio no es tan importante. Seguimos viendo que la variable significativa es únicamente el método.

[56]:
```
m=mean(rendimiento)
cat("Media: ", m)
boxplot(rendimiento ~ sexo)
```

```
Media:  6
```



Además, si hacemos las gráficas de caja y bigote para hombres y mujeres, se sustenta la afirmación de que la variable de sexo no es significativa. Podemos ver que son prácticamente iguales.

[57]:
```
#Rendimiento para hombres

x = M$Rendimiento[M$Sexo == "H"]
media = mean(x)


alpha = 1 - 0.95 #Hacemos 1 menos el porcentaje de confianza
d = sd(x) #desviacion estandar
n = length(x) #numero de datos
```

```
z = abs(qt(alpha/2, n-1))
e = z*(d/n^0.5) #Este es el error

rango_inf = media - e
rango_sup = media + e

cat("El rango  para Hombres es: ", rango_inf, "   -   ", rango_sup, "\n")

#Rendimiento para mujeres

x = M$Rendimiento[M$Sexo == "M"]
media = mean(x)


alpha = 1 - 0.95 #Hacemos 1 menos el porcentaje de confianza
d = sd(x) #desviacion estandar
n = length(x) #numero de datos

z = abs(qt(alpha/2, n-1))
e = z*(d/n^0.5) #Este es el error

rango_inf = media - e
rango_sup = media + e

cat("El rango  para Mujeres es: ", rango_inf, "   -   ", rango_sup)
```

```
El rango  para Hombres es:  5.103347    -    7.56332
El rango  para Mujeres es:  4.356505    -    6.976828
```

Por ultimo, podemos ver los rangos de confianza para descartar la H_0 de la segunda hipótesis.

1.4 ANOVA para el efecto principal

Finalmente, podemos hacer el ANOVA con el efecto principal, el cuál es el método. Nuestras hipótesis quedan así:

Primera hipótesis

$H_0 : \tau_i = 0$

H_1 : algún τ_i es distinto de cero

[58]:
```
C<-aov(rendimiento~metodo)
summary(C)
tapply(rendimiento,metodo,mean)
cat("Media: ", mean(rendimiento), "\n")
```

```
            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
```

M1 8.5 **M2** 6 **M3** 3.5

```
Media:  6
```

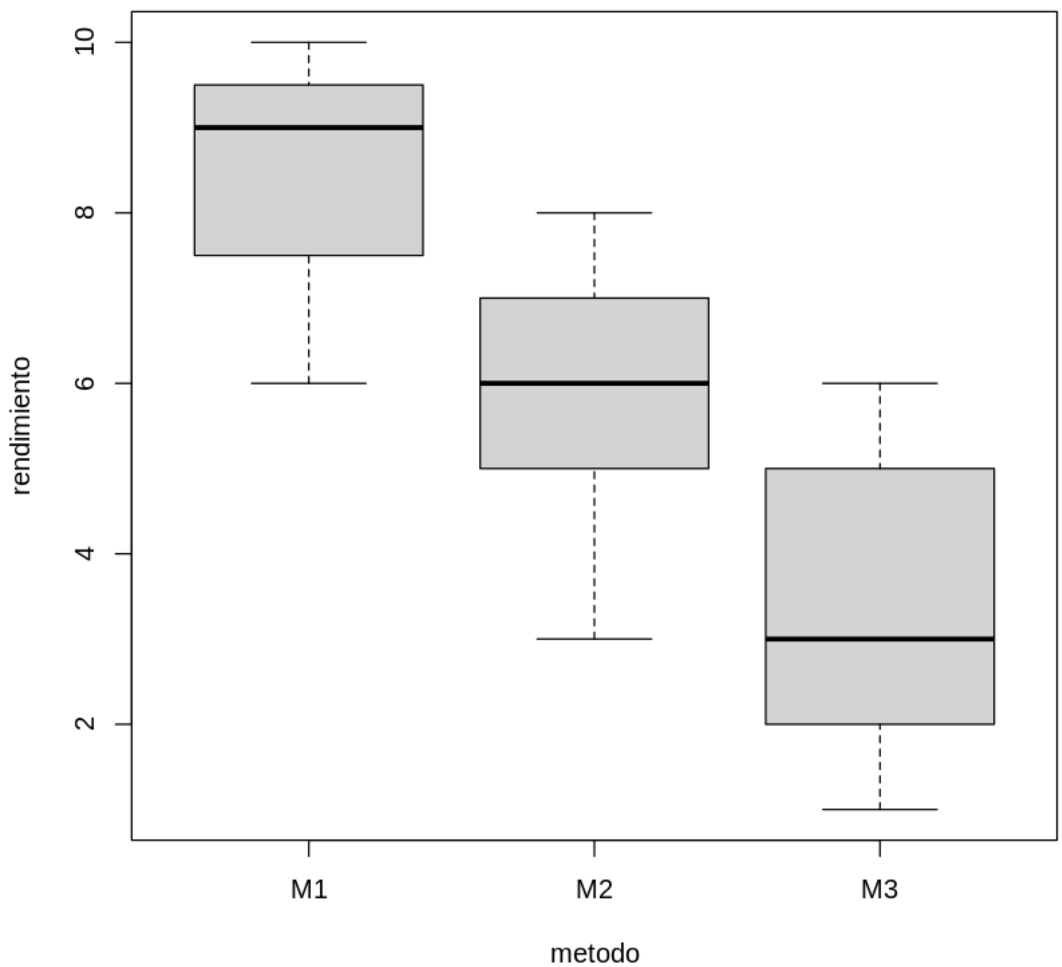
Podemos ver que al realizar el análisis únicamente con el método, sigue siento significativo. Esto lo podemos ver con el valor p resultante, que no rechaza la H_0.

[59]:
```
boxplot(rendimiento ~ metodo)
I = TukeyHSD(aov(rendimiento ~ metodo))
I
plot(I) #Los intervalos de confianza se observan
```

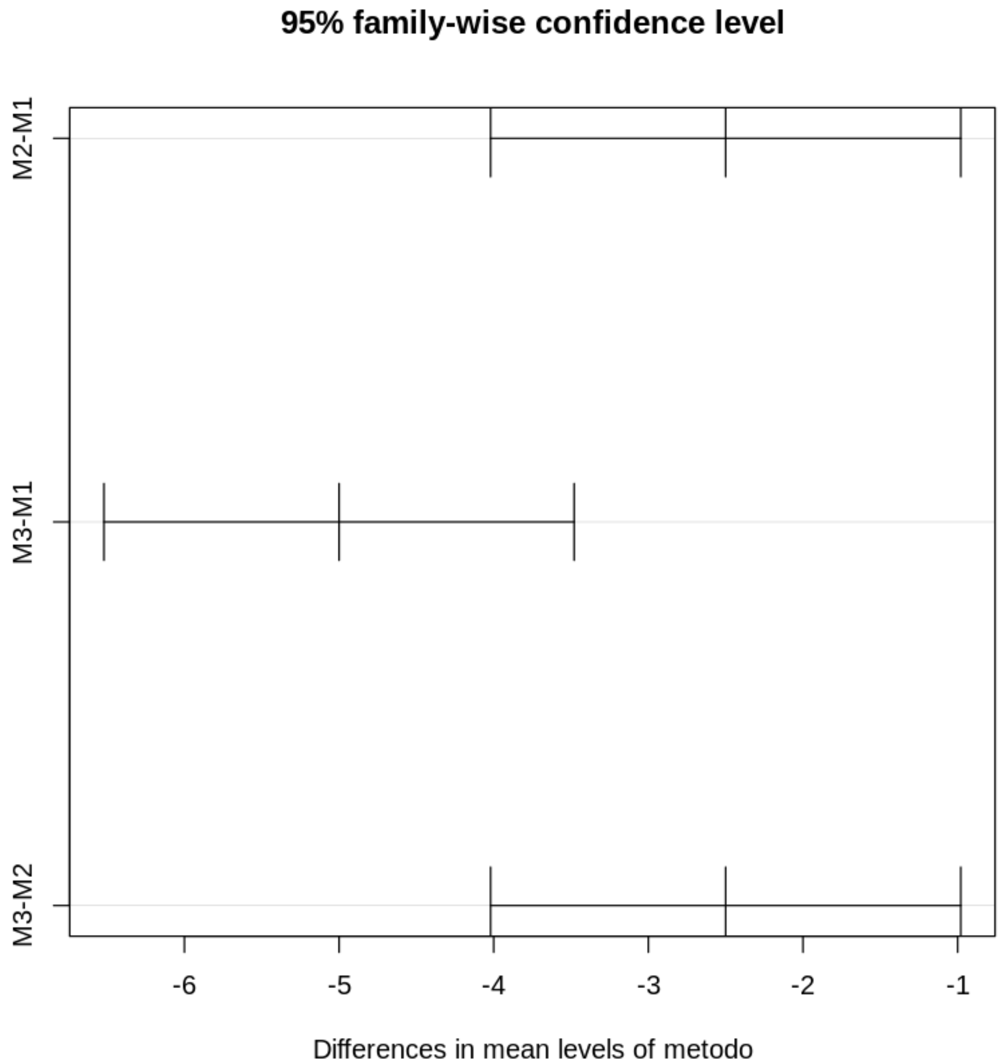
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

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
```



10
8
6
4
2
rendimiento
M1
M2
M3
metodo



Además, esto se sustenta viendo los dos gráficos superiores.

Con el primero podemos ver como los intervalos de confianza se encuentran bastante separados de modelo a modelo.

Y en el segundo gráfico podemos ver que las variables son diferentes entre ellas porque ninguno de los rangos pasa por el 0.

Por lo tanto, podemos decir que nuestra H_0 se mantiene:

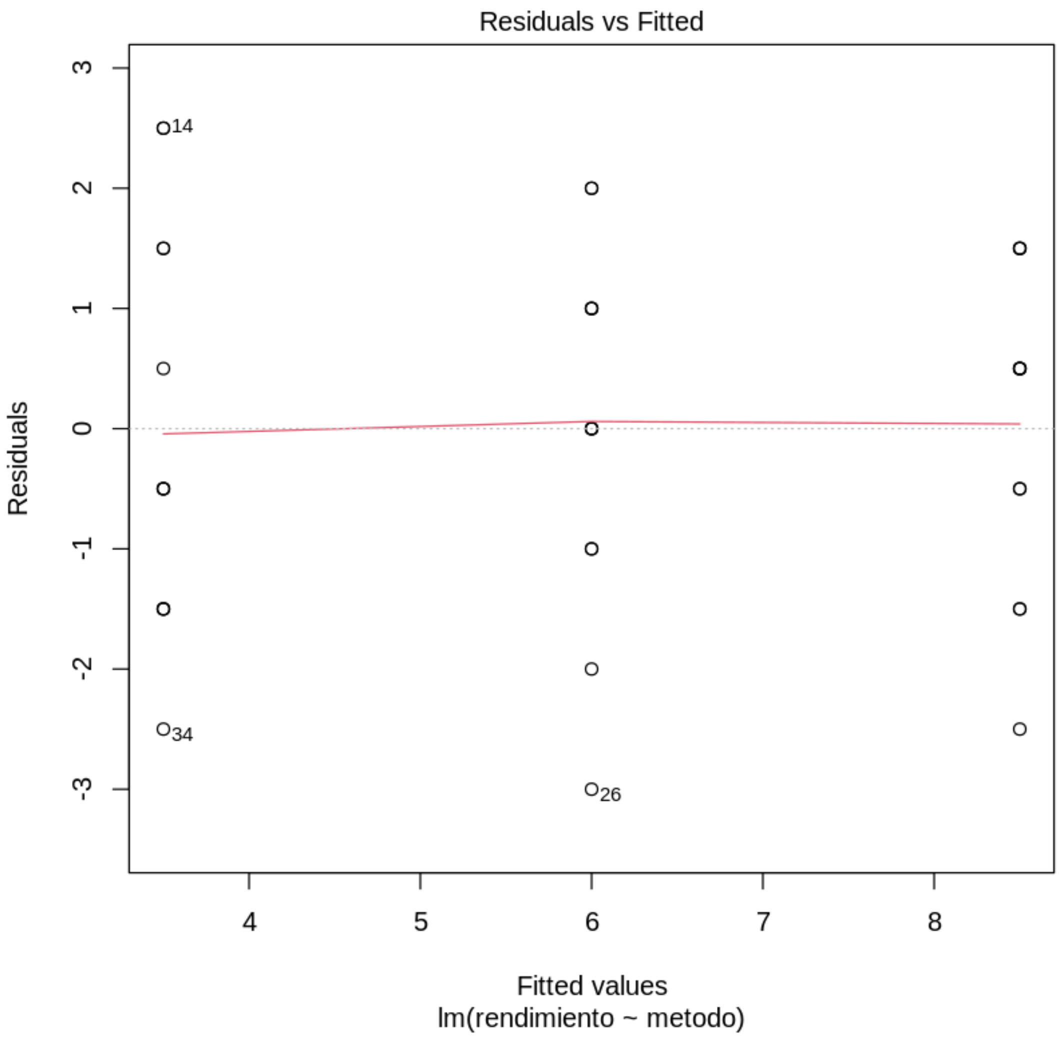
$H_0 : \tau_i = 0$

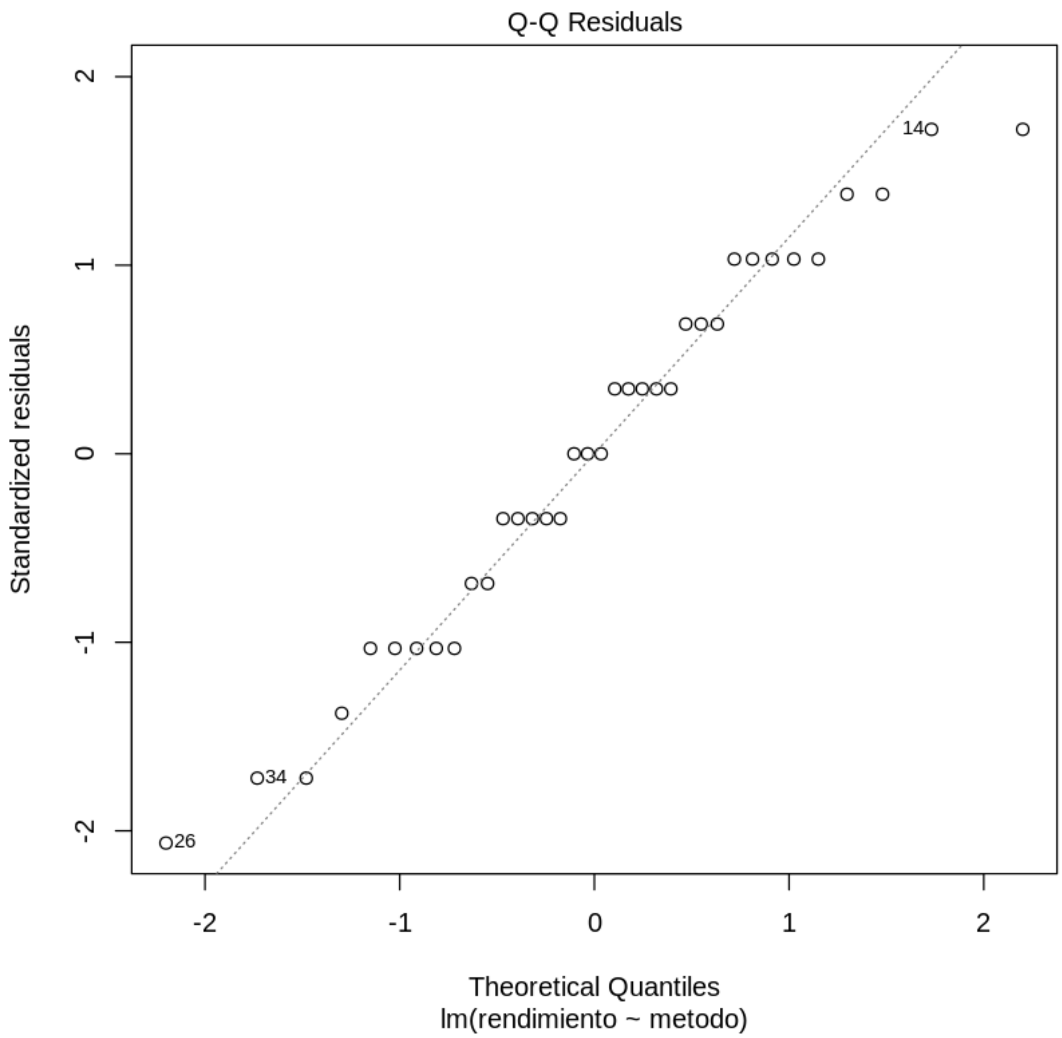
1.5 Validación del modelo

Podemos hacer al validación del modelo con la normalidad, homocedasticidad, independencia. Se muestran las gráficas para estos datos y posteriormente se realiza el análisis.

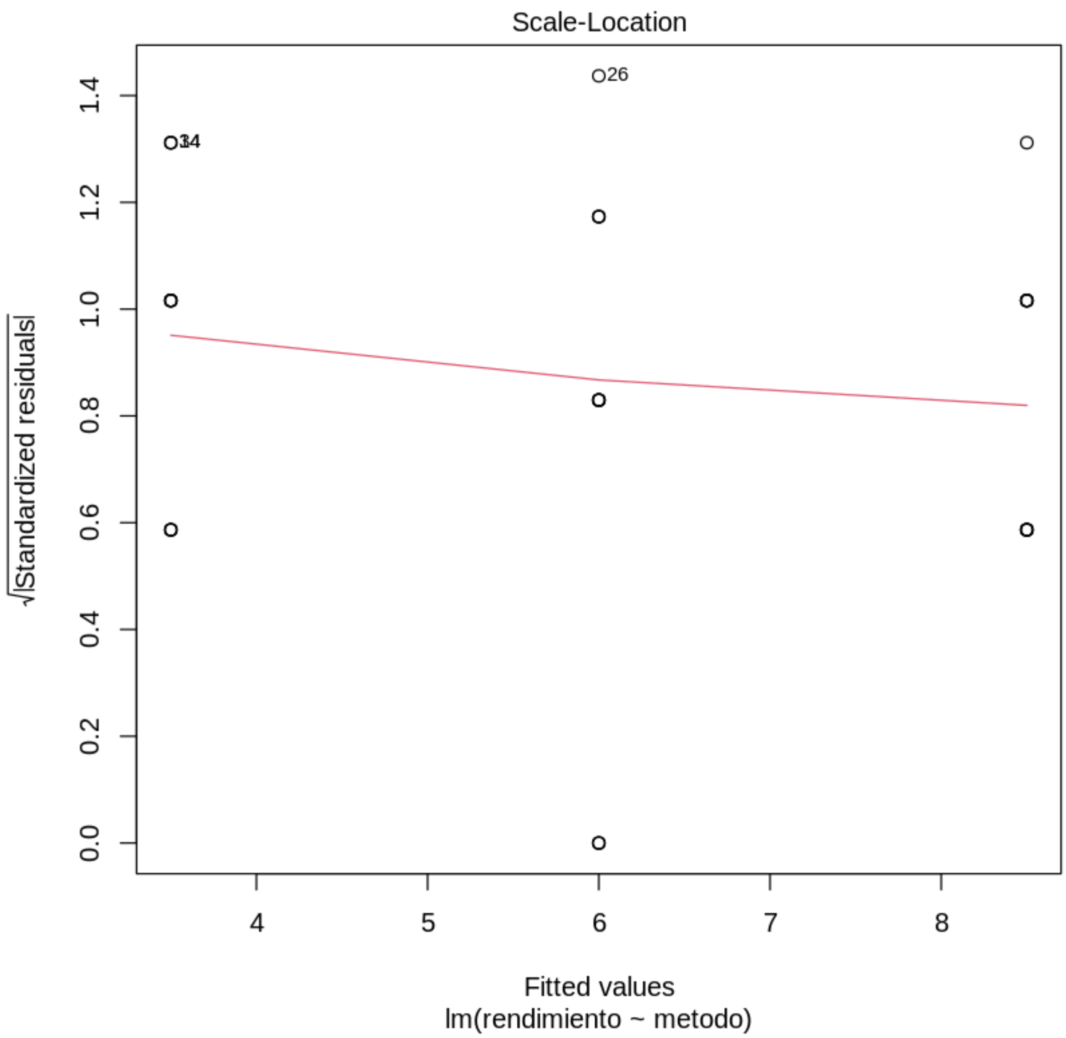
[60]:
```
plot(lm(rendimiento~metodo))
```

```
Coeficiente de determinación:  0.6637168
```

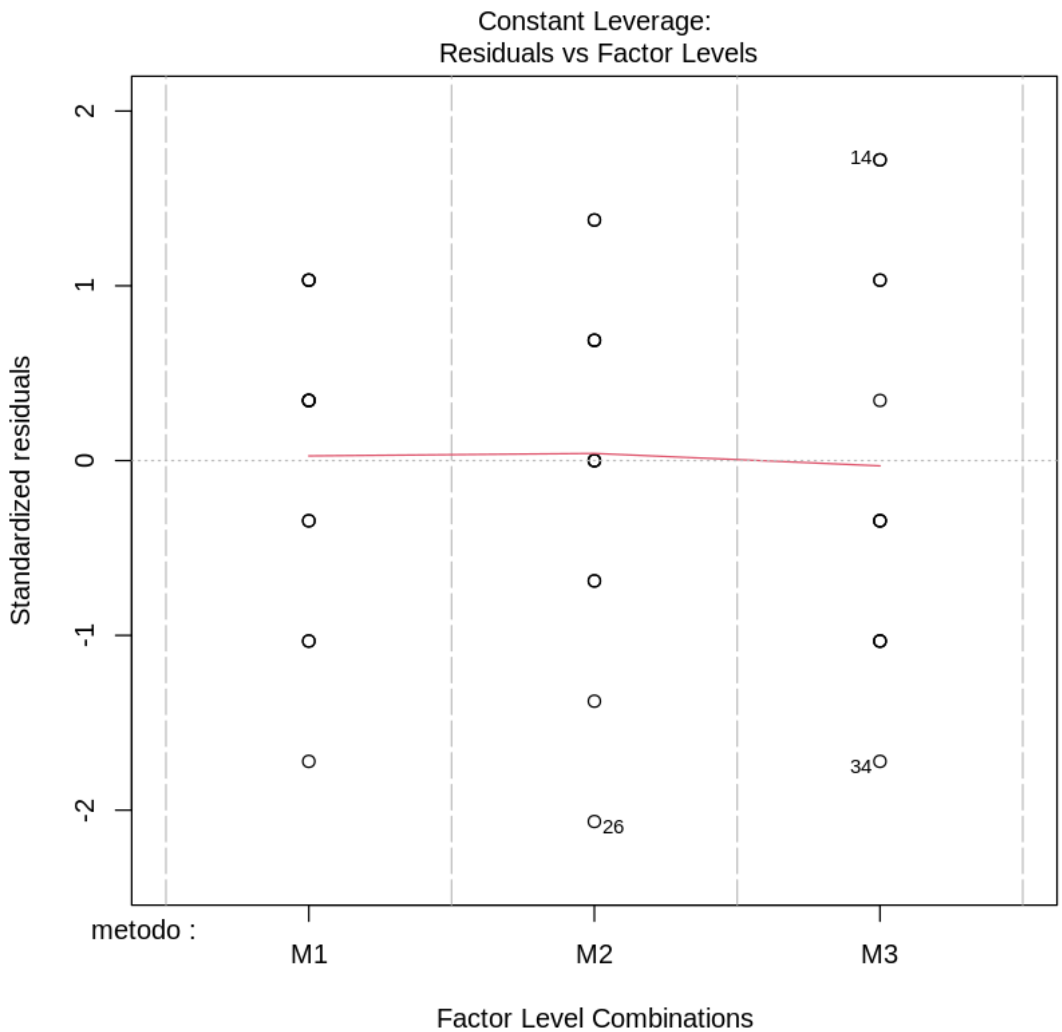




Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(rendimiento ~ metodo)
2
1
0
-1
-2
-2
-1
0
1
2
14
34
26



Scale-Location
26
34
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
√|Standardized residuals|
4
5
6
7
8
Fitted values
lm(rendimiento ~ metodo)



En la primer gráfica podemos ver que se precenta **homocedasticidad**. Todos los puntos tienen una varianza constante, ya que se encunetran distribuidos en todo el gráfico.

Al analizar el gráfico QQ podemos ver que los residuos se distribuyen como una normal (no hay colas muy pronunciadas). Por lo tanto, los datos también presentan **normalidad**.

En la ultima gráfica podemos ver que los residuos no tienen ningun comportamiento, por lo que se presenta **independencia**.

[62]:
```
CD= 150/(150+76) #coeficiente de determinación para el modelo.
cat("Coeficiente de determinación: ", CD)
```

0.663716814159292

Y para comprobar que los datos tienen un **comportamiento lineal** podemos ver el coefciente de determinación, que nos dice que el 66.37% de la varianza es explicada por el modelo.

1.6 Conclusiones

Podemos ver que para este modelo solamente fue necesario utilizar el método que usaron los niños, ya que el sexo no influyó en los cambios de calificaciones. Además, tampoco fue importante la relación entre el método usado y el sexo del estudiante.

Además, el modelo explica el 66.37% de la variación, mientras que el porcentaje se lo asignamos al error. De todas formas, podría haber otros factores que no estamos tomando en cuenta que afecten el porcentaje del error.