Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad M1. ANOVA

TC3006C.101 Inteligencia artificial avanzada para la ciencia de datos I

Profesores:

Ivan Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernandez

Antonio Carlos Bento

Frumencio Olivas Alvarez

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

25 de Agosto de 2023

Instrucciones

Resuelve las dos partes del problema "El rendimiento". Se encuentra en los apoyos de clase de "ANOVA". Para ello se te recomienda que sigas los siguientes pasos

Problema

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:

Chicos Chicas
Método 1 Método 2 Método 3 Método 1 Método 2 Método 3
10 5 2 9 8 2
7 7 6 7 3 6
9 6 3 8 5 2
9 6 5 8 6 1
9 8 5 10 7 4
10 4 3 6 7 3

In [ ]:
```
# Cargamos el lenguaje de R para utilizarlo en Google Colab
%load_ext rpy2.ipython
```

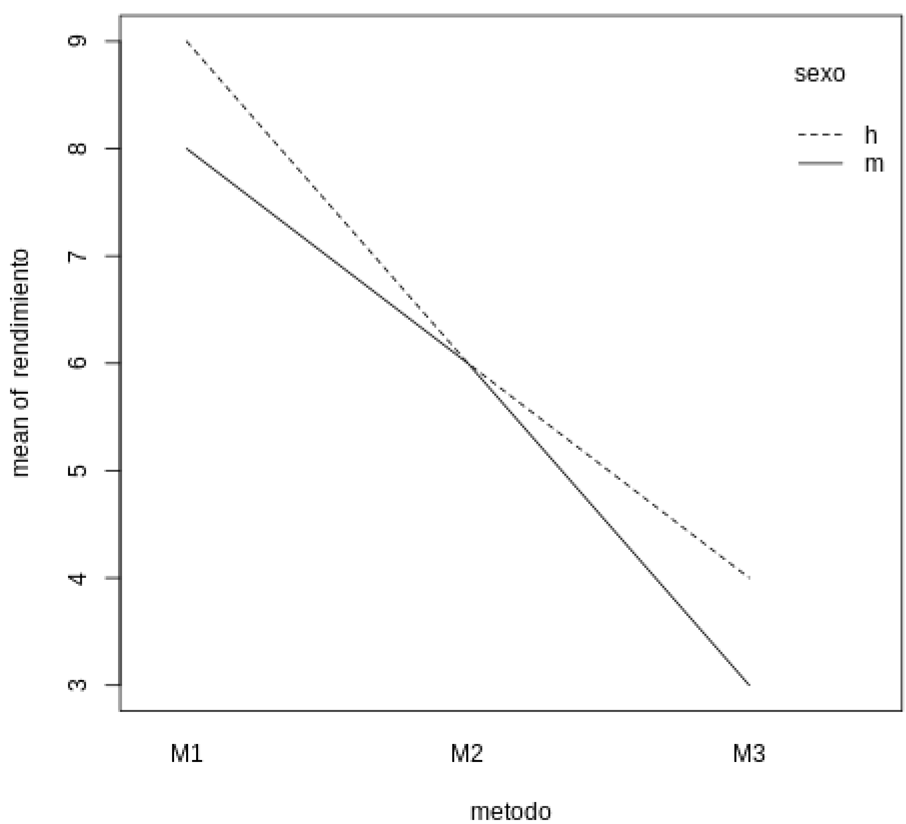
Establece las hipótesis estadísticas (tienen que ser 3).

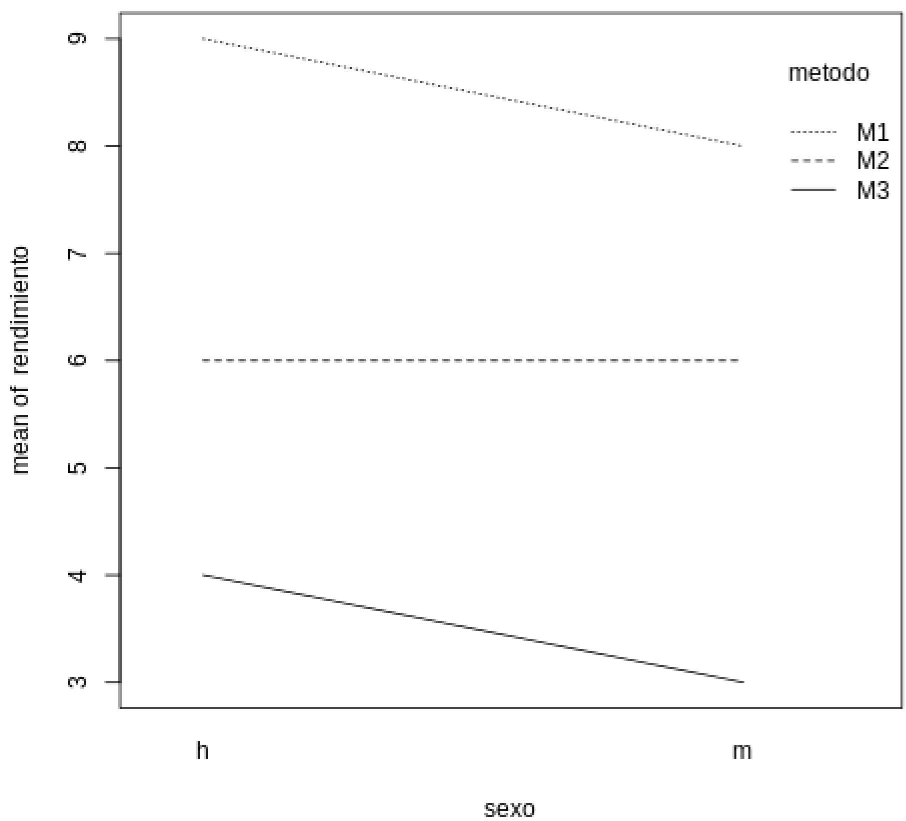
In [ ]:
```
%%R

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)
```

ANOVA (con interacción)

Considerando a los dos factores más su interacción y algunos gráficos para observar la interacción entre los dos factores

In [ ]:
```
%%R

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





F1: Método de enseñanza (tres niveles: M11 M2, M3)

F2: Sexo (dos niveles: hombre, mujer)

$Y_{ijk} = \mu + \tau_i + \alpha_j + \varepsilon_{ijk}$

Bajo este modelo:

Método

$\sum_{i=1}^{3} \tau_i = 0$

Sexo

$\sum_{j=1}^{2} \alpha_j = 0$

Después de realizar el ANOVA con el modelo completo, encontramos que $\tau\tau\ ii\ \alpha\alpha jj$ fue no significativa (no hay efecto de interacción). No se rechaza la tercera hipótesis nula (H 0) y el modelo se reduce.

Primera Hipótesis:

$H_0 : \tau_i = 0$ (no hay efecto del método de enseñanza)

H_1: algún τ_i es distinto de cero

Segunda Hipótesis:

$H_0 : \alpha_j = 0$ (no hay efecto del sexo)

H_1: algún α_j es distinto de cero

ANOVA (sin interacción)

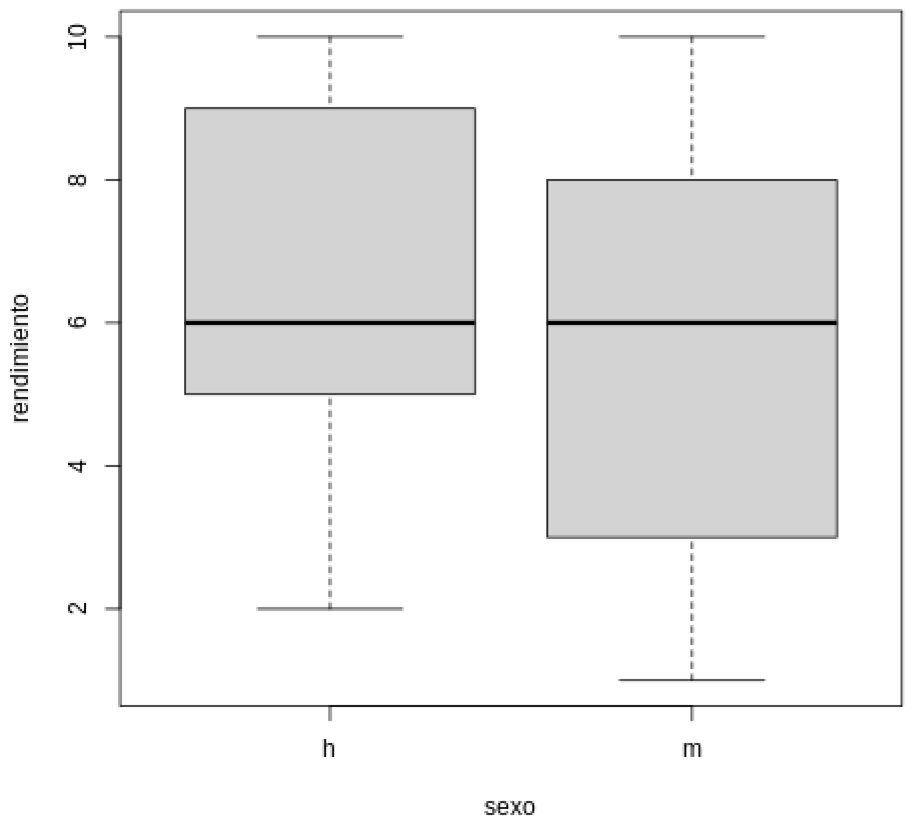
En el modelo, se consideran sólo los efectos principales. Ya no se usa *, se usa +.

In [ ]:
```
%%R

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
```

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.

In [ ]:
```
%%R

tapply(rendimiento,sexo,mean)
tapply(rendimiento,metodo,mean)
M=mean(rendimiento)
M
boxplot(rendimiento ~ sexo)
```



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

Bajo este modelo:

Método

$\sum_{i=1}^{3} \tau_i = 0$

Primera Hipótesis:

H_0: \tau_i = 0$ (no hay efecto del método de enseñanza)

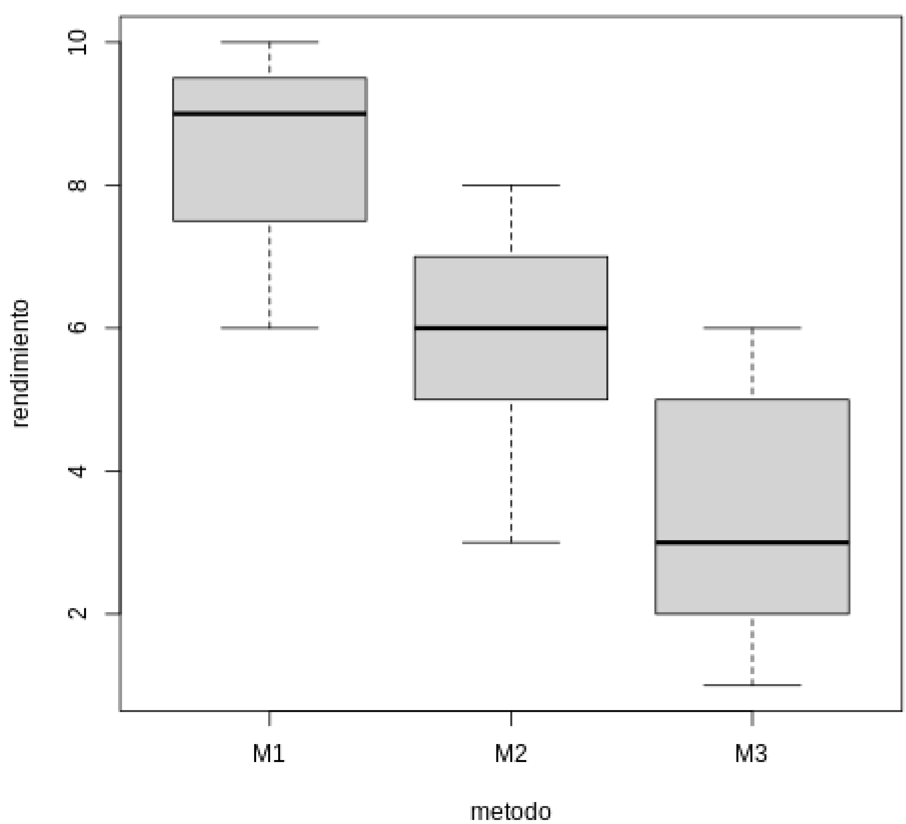
H_1: algún τ_i es distinto de cero

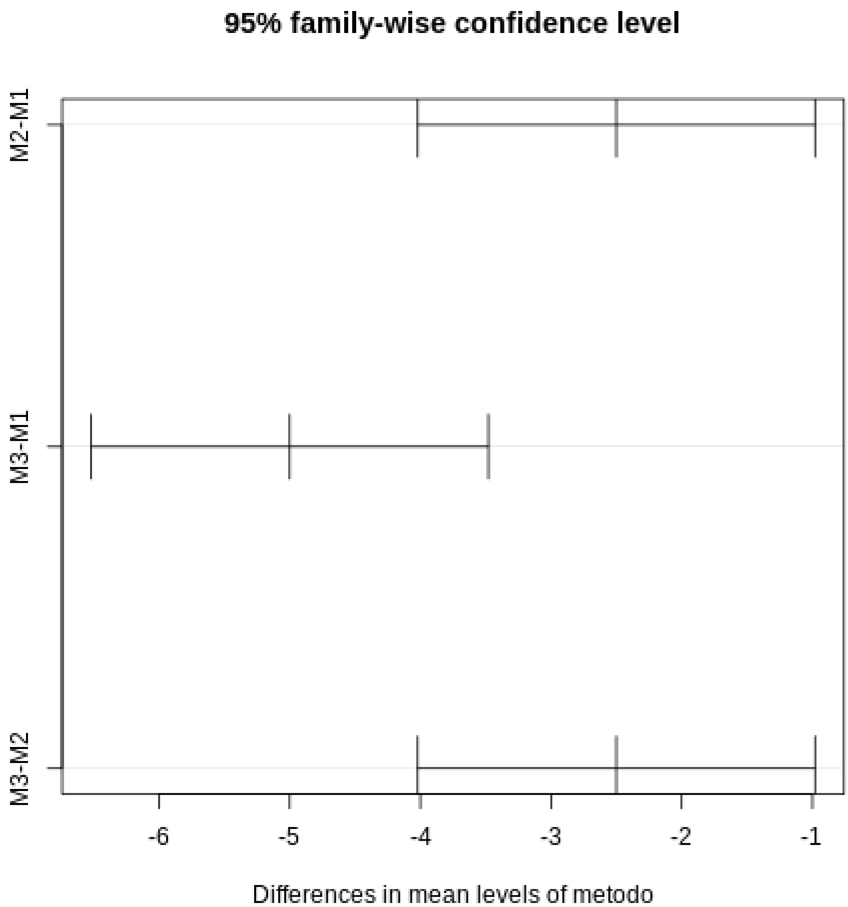
Una vez encontrado el modelo que solo contiene los efectos significativos, se analiza para detectar cuál es el efecto de los niveles del factor significativo y analizar la validez del modelo

ANOVA con un solo factor (el significativo)

En el modelo, se consideran sólo el efecto significativo.

In [ ]:
```
%%R

C<-aov(rendimiento~metodo)
summary(C)
tapply(rendimiento,metodo,mean)
mean(rendimiento)
boxplot(rendimiento ~ metodo)
I = TukeyHSD(aov(rendimiento ~ metodo))
I
plot(I) #Los intervalos de confianza se observan mejor si se grafican
```

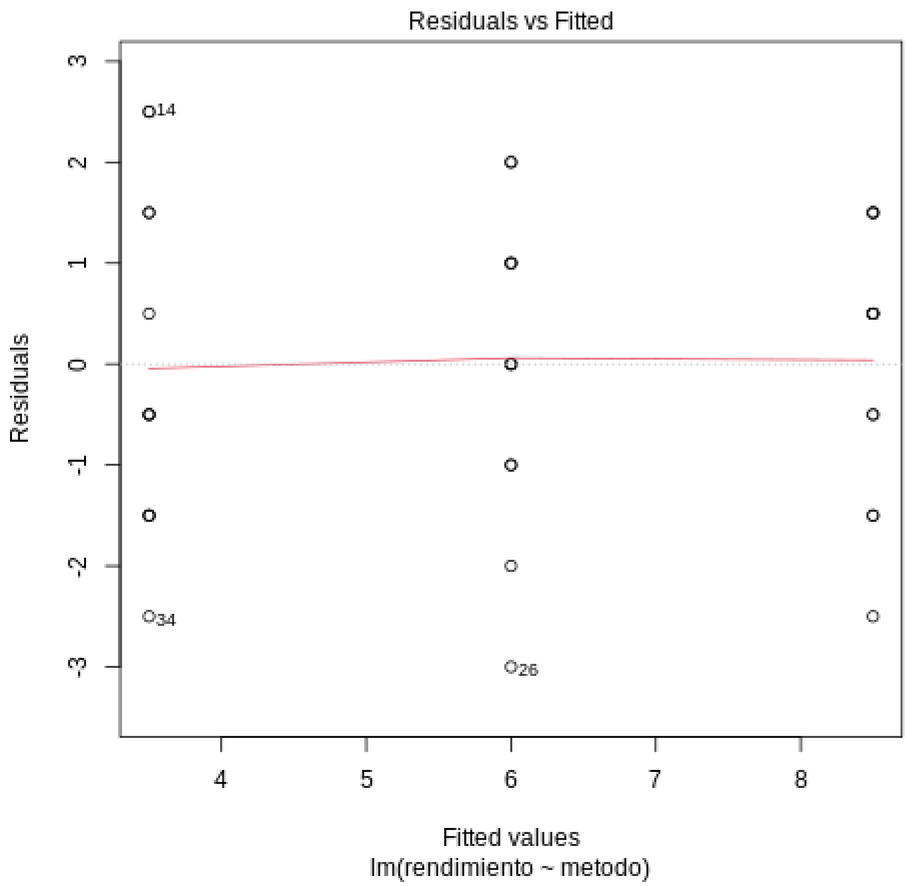


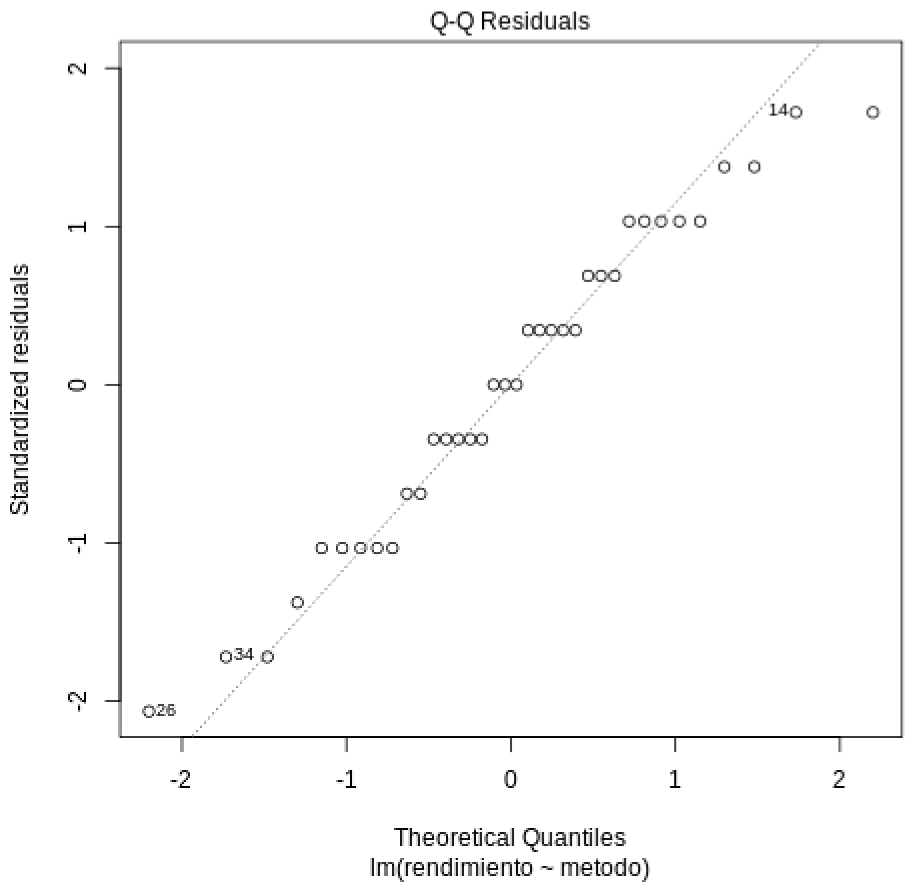


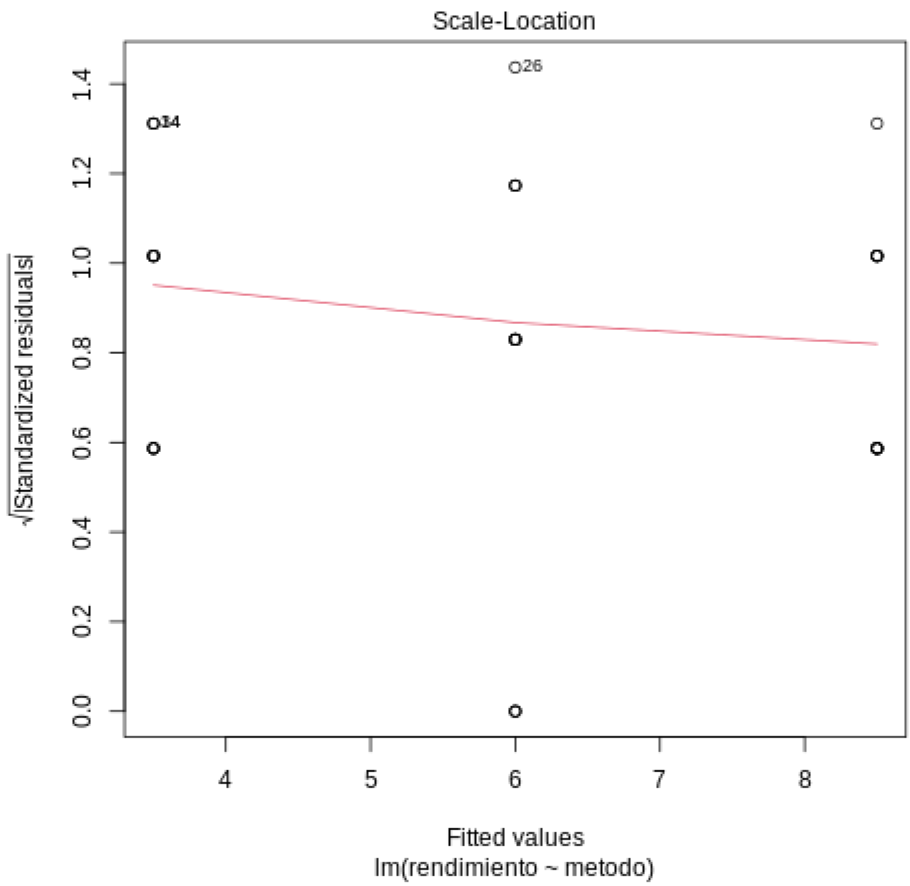
Análisis del Modelo

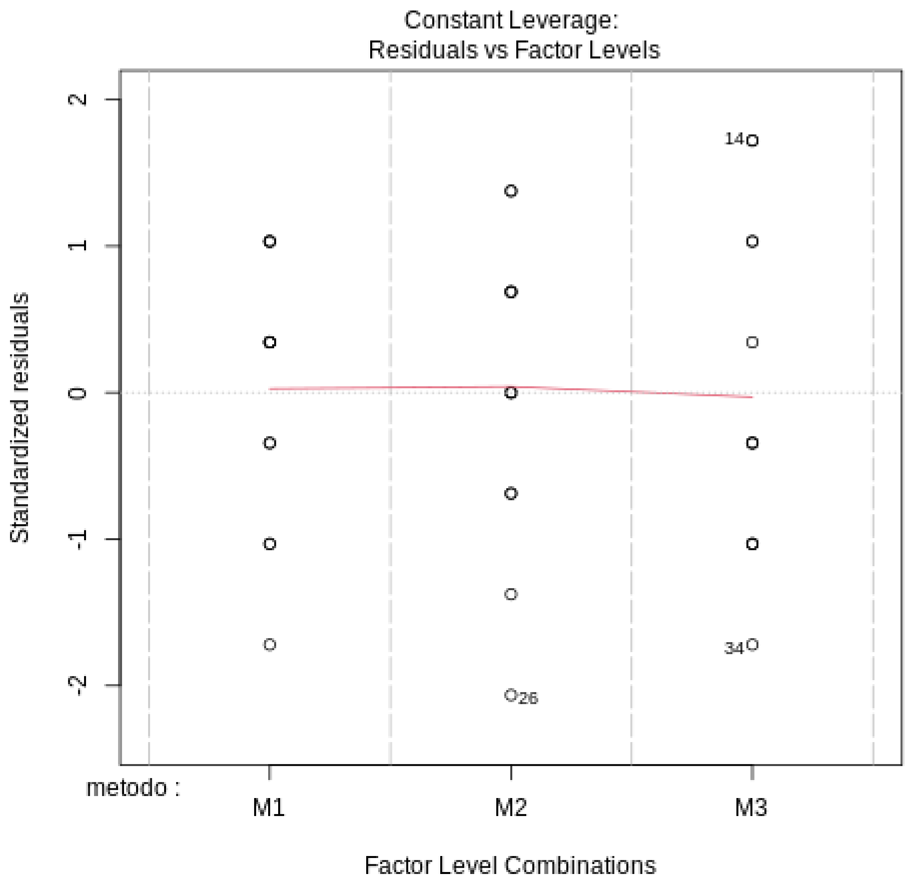
Se verifica la validez del modelo por medio de las gráficas de residuos y la gráfica de normalidad. También se pueden calcular los coeficientes de determinación del modelo para conocer la variación explicada por el modelo.

In [ ]:
```
%%R

plot(lm(rendimiento~metodo))
CD= 150/(150+76) #coeficiente de determinación para el modelo.
```


Residuals vs Fitted
Residuals
Fitted values
lm(rendimiento ~ metodo)
14
34
26
3
2
1
0
-1
-2
-3
4
5
6
7
8


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


Scale-Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
26
14
4
5
6
7
8
Fitted values
lm(rendimiento ~ metodo)



Se concluye: H_1: algún τ_i es distinto de cero.

Del cálculo de las medias se obtiene:

$\hat{\mu}_1 = 8.5$;

$\hat{\mu}_2 = 6.0$;

$\hat{\mu}_3 = 3.5$;

$\hat{\mu} = 6$

Entonces, los efectos para cada nivel son:

τ_1 = 2.5$;

τ_2 = 0.0$;

τ_3 = -2.5$;

Varianza Explicada por el modelo: 66.37% (SCTratamiento/SCtotales)

Solo el efecto del Método de enseñanza fue significativo y no hay diferencia en que los estudiantes sean niños o niñas. Se observó que los 3 métodos producen un efecto diferente en el rendimiento de los niños. El efecto del Método 3 es un método deficiente, puesto que se disminuye su rendimiento con respecto a la media general, el Método 2 no tiene efecto, es un método que no modifica el rendimiento de los estudiantes, y el Método 1 incrementa su rendimiento con respecto a la media general por lo que resulta ser el mejor método de enseñanza. El modelo explica el 66.37% de la variación. Por lo tanto, el Método de enseñanza es un factor determinante en el rendimiento de los estudiantes (puesto que es el único que fue significativo en el modelo), sin embargo, es posible que haya otros factores que expliquen el resto del porcentaje de variación (32.73%) y que en este modelo se le atribuye a la aleatoriedad (al error). El número de datos en cada tratamiento fue igual por lo que es un diseño equilibrado que es robusto a heterocedasticidad. De acuerdo al análisis de los gráficos Q-Q y de los residuos vs. el valor esperado (ajustado), los datos aparentemente cumplen con normalidad e independencia. También los errores tienen una media cero y variación constante.