ANOVA

AUTHOR
Alfredo García

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

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

Tabla de problema de rendimiento

```
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).

$H_0 : \tau_i = 0 \; H_1 : algún \; \tau_i \; es \; distinto \; de \; cero$

$H_0 : \alpha_j = 0 \; H_1 : algún \; \alpha_j \; es \; distinto \; de \; cero$

$H_0 : \tau_i \alpha_j = 0 \; H_1 : algún \; \tau_i \alpha_j \; es \; distinto \; de \; cero$

2. Realiza el ANOVA para dos niveles con interacción:

Haz la gráfica de interacción de dos factores.

Haz la gráfica de interacción de dos factores.

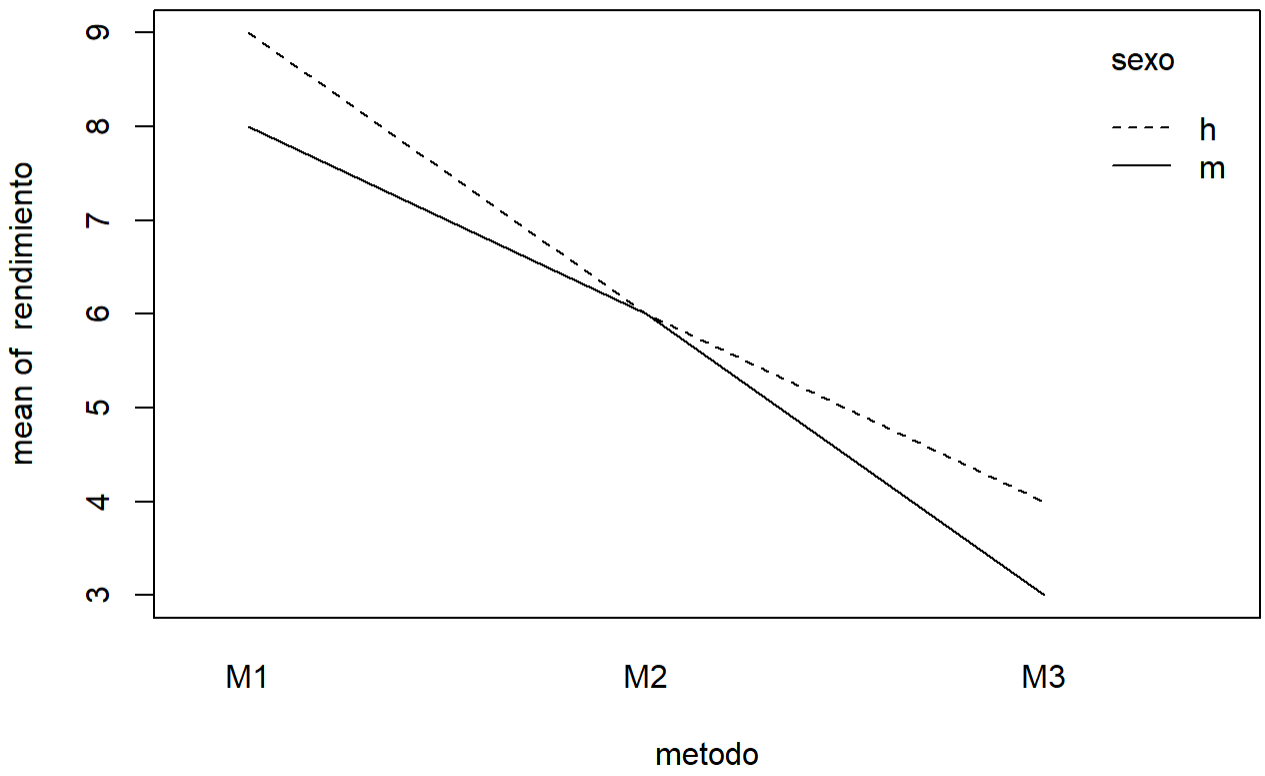
ANOVA (con interacción)

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

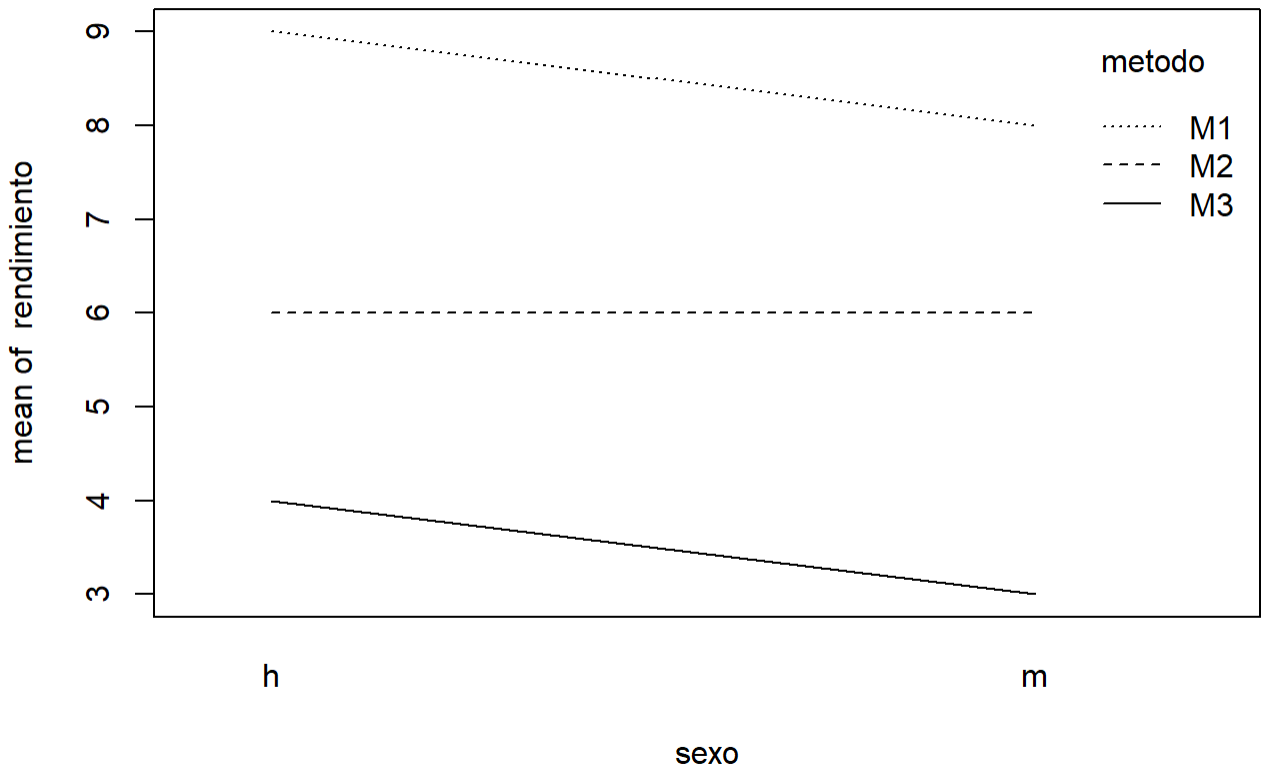
```
A = aov(rendimiento~metodo*sexo)
summary(A)
```

```
            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
```

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




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



3. Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. Escribe tus conclusiones parciales

De los resultados anteriores, observamos que la interacción entre "Método" y "Sexo" no es significativa, ya que el valor de p es 0.655, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula de que esta interacción no afecta significativamente el rendimiento. Por lo tanto, la tercera hipótesis del modelo se rechaza, ya que la suma de la interacción entre ambos factores no es igual a cero.

4. Realiza el ANOVA para dos niveles sin interacción.

ANOVA (sin interacción). Haz el boxplot de rendimiento por sexo. Calcula la media para el rendimiento por sexo y método.

En el modelo B, 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
```

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
```

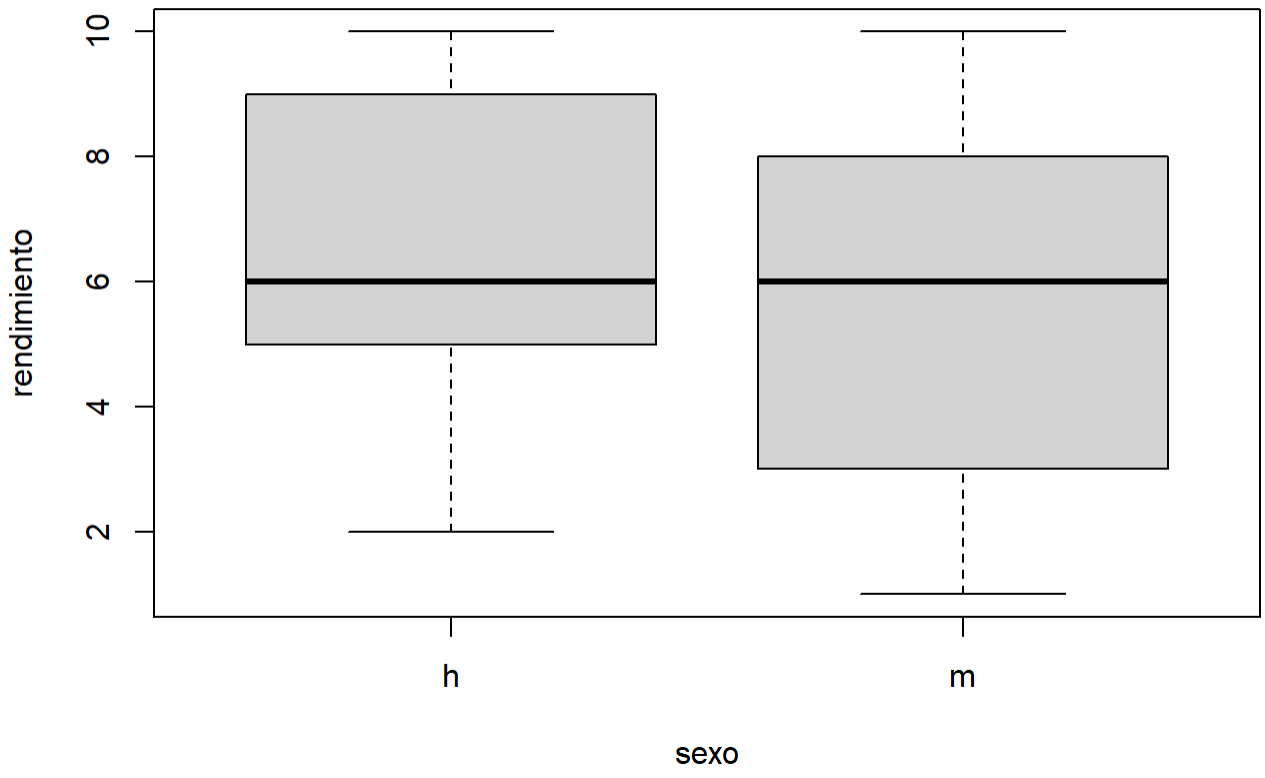
```
tapply(rendimiento,metodo,mean)
```

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

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

```
[1] 6
```

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

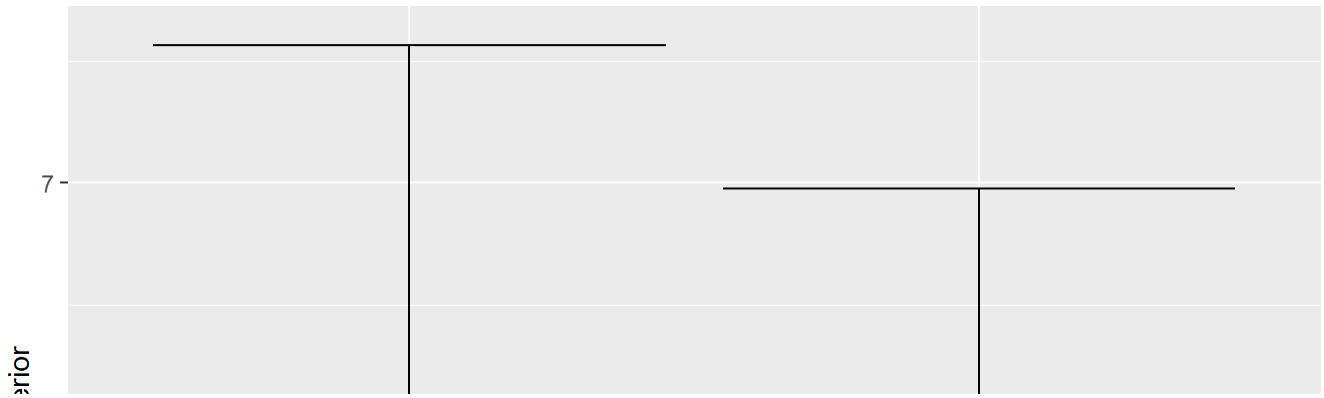


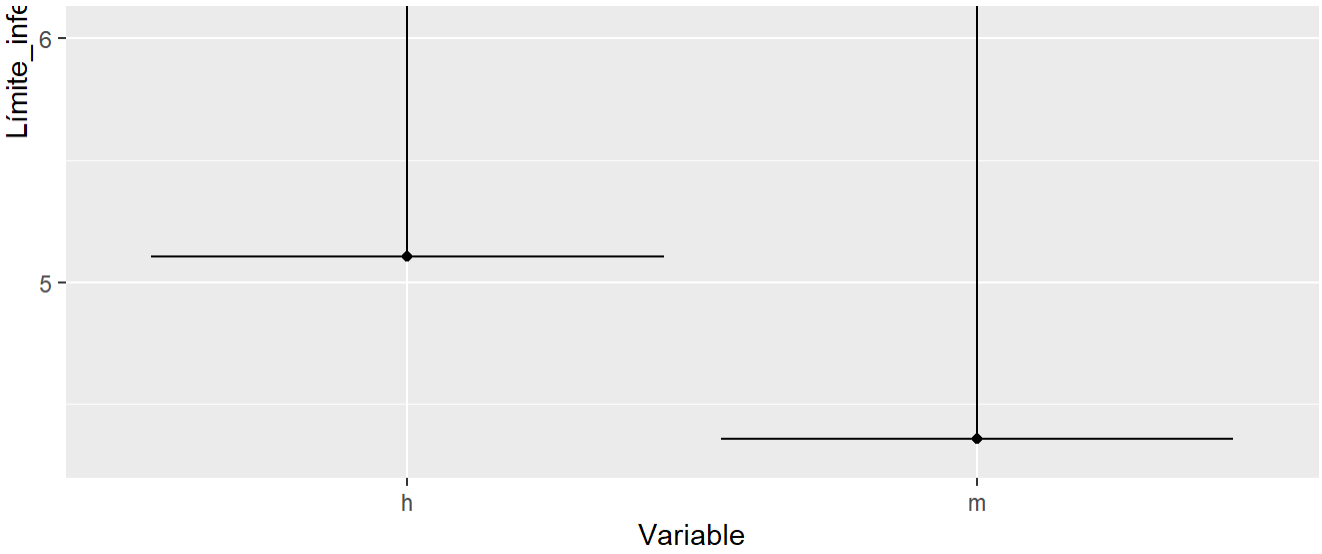
Haz los intervalos de confianza de rendimiento por sexo. Grafícalos

```
conf_int <- tapply(rendimiento, sexo, function(x) t.test(x)$conf.int)

# Crear un vector con los nombres de los grupos
grupos <- names(conf_int)

# Imprimir los intervalos de confianza
for (i in 1:length(grupos)) {
  cat("Intervalo de confianza para", grupos[i], ":\n")
  cat("  Límite inferior:", conf_int[[i]][1], "\n")
  cat("  Límite superior:", conf_int[[i]][2], "\n\n")
}
```

```
Intervalo de confianza para h :
  Límite inferior: 5.103347
  Límite superior: 7.56332

Intervalo de confianza para m :
  Límite inferior: 4.356505
  Límite superior: 6.976828
```

```
# Crea un dataframe con los valores
data <- data.frame(
  Variable = c("m", "h"),
  Límite_inferior = c(4.356505, 5.103347),
  Límite_superior = c(6.976828, 7.56332)
)

# Carga la librería ggplot2
library(ggplot2)

# Create the ggplot plot
ggplot(data, aes(x = Variable, y = Límite_inferior, ymin = Límite_inferior, ymax = Límite_super
  geom_point() +
  geom_errorbar()
```



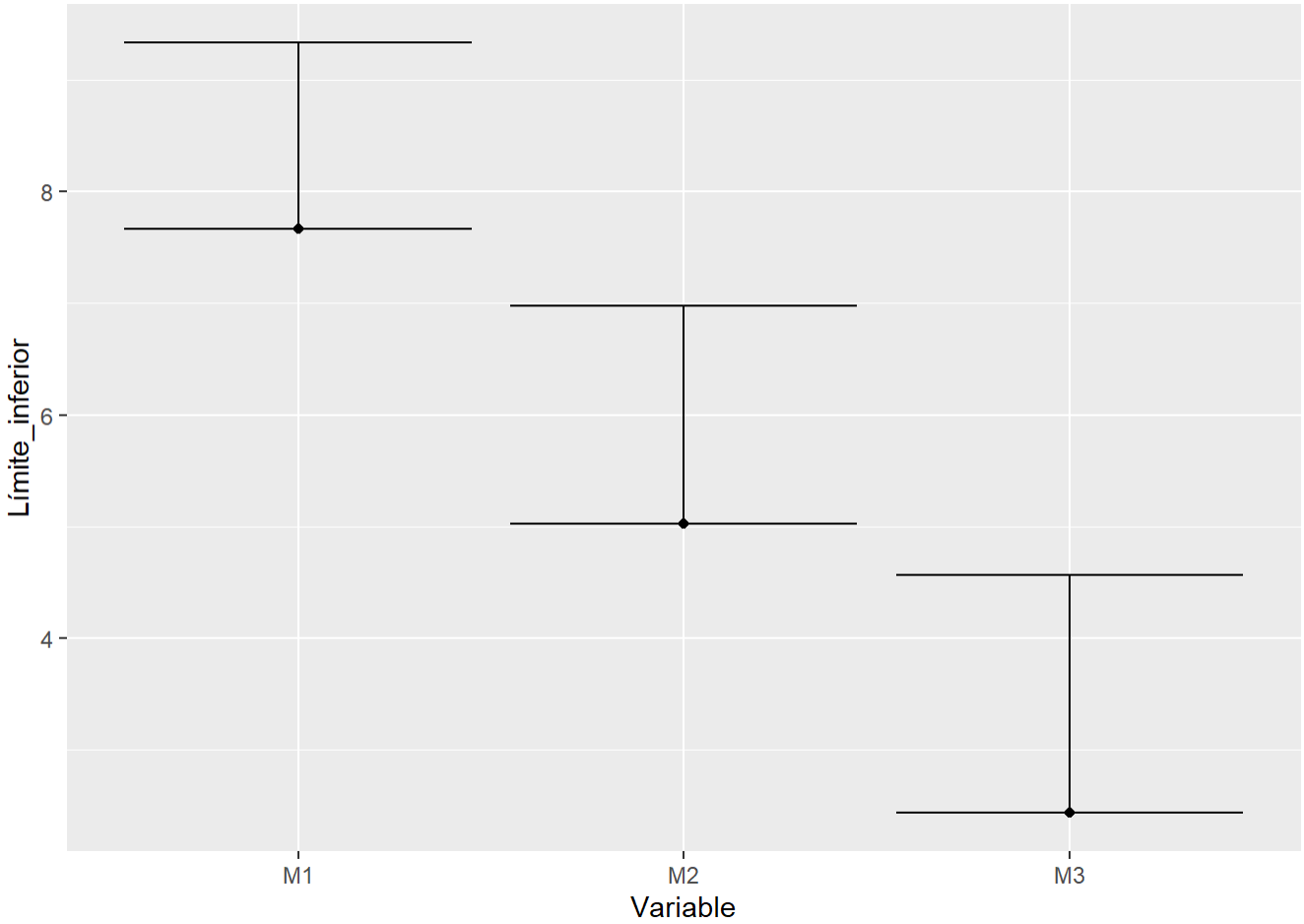



```
conf_int <- tapply(rendimiento, metodo, function(x) t.test(x)$conf.int)

# Crear un vector con los nombres de los grupos
grupos <- names(conf_int)

# Imprimir los intervalos de confianza
for (i in 1:length(grupos)) {
  cat("Intervalo de confianza para", grupos[i], ":\n")
  cat("  Límite inferior:", conf_int[[i]][1], "\n")
  cat("  Límite superior:", conf_int[[i]][2], "\n\n")
}
```

```
Intervalo de confianza para M1 :
  Límite inferior: 7.664961
  Límite superior: 9.335039

Intervalo de confianza para M2 :
  Límite inferior: 5.023175
  Límite superior: 6.976825

Intervalo de confianza para M3 :
  Límite inferior: 2.433377
  Límite superior: 4.566623
```

```
# Crea un dataframe con los valores
data <- data.frame(
  Variable = c("M1", "M2", "M3"),
  Límite_inferior = c(7.664961, 5.023175, 2.433377),
  Límite_superior = c(9.335039, 6.976825, 4.566623)
)


# Carga la librería ggplot2
```

```
library(ggplot2)

# Create the ggplot plot
ggplot(data, aes(x = Variable, y = Límite_inferior, ymin = Límite_inferior, ymax = Límite_super
  geom_point() +
  geom_errorbar()
```



Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. Escribe tus conclusiones parciales

El análisis muestra que el factor "método" tiene un efecto significativo en el rendimiento (p = 1.5e-08), mientras que el factor "sexo" no lo tiene (p = 0.192) lo que significa que no hay suficiente evidencia estadística para rechazar la hipótesis nula de que el género tiene un efecto significativo en el rendimiento. Las medias de rendimiento son aproximadamente 6.33 para hombres, 5.67 para mujeres, 8.5 para M1, 6.0 para M2 y 3.5 para M3, con una media global de 6.

Con lo cual pudimos ir dandonos una idea desde que calculamos los promedios y ver como cambiaban esto en funcion de si usabamos el sexo o el metodo, en donde vemos que las diferencias son bastante mas grandes cuando lo hacemos por metodo en comparacion a con sexo, además de que al sacar los intervalos de confianza podemos ver que cuando tomamos el sexo estos se traslapan en gran proporcion a comparacion de cuando usamos el metodo en donde ninguno se toca.

5. Realiza el ANOVA para un efecto principal

ANOVA con un solo factor (el significativo)

En el modelo, se consideran sólo el efecto significativo.

```
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
```

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

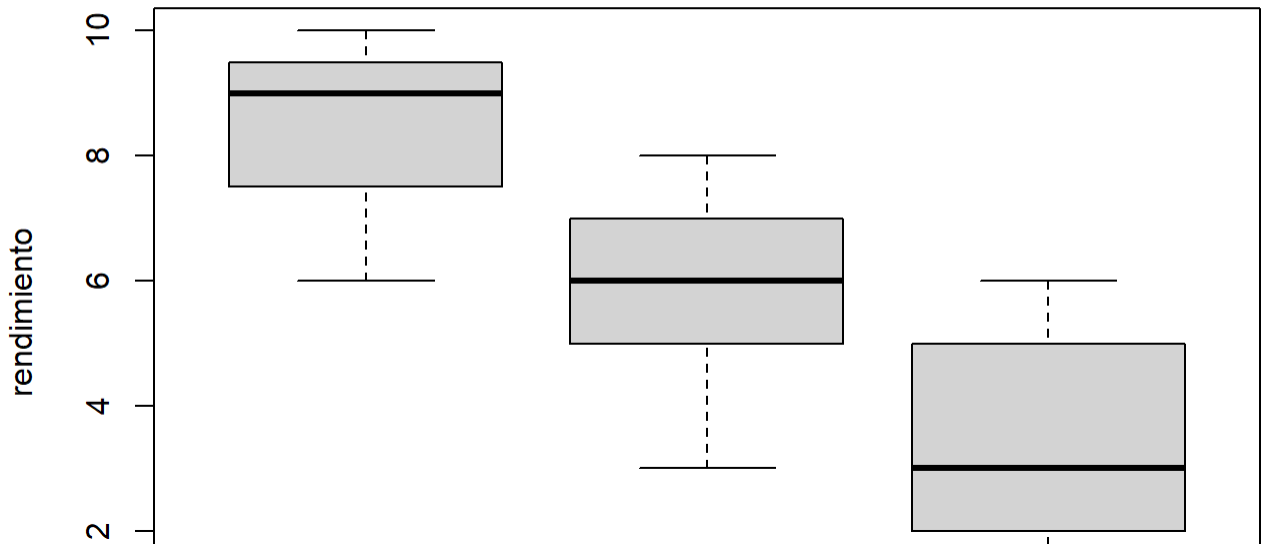
```
tapply(rendimiento,metodo,mean)
```

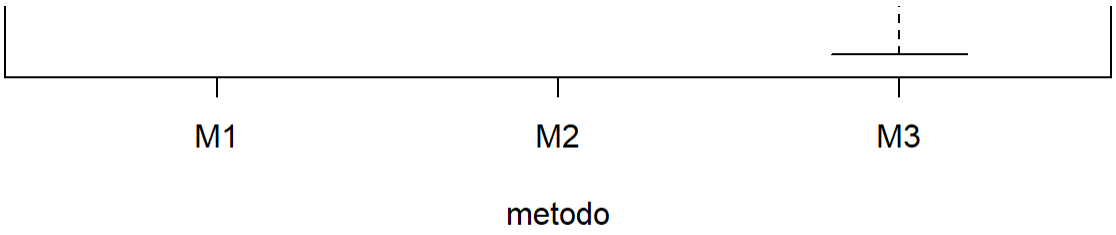
```
 M1  M2  M3
8.5 6.0 3.5
```

```
mean(rendimiento)
```

```
[1] 6
```

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





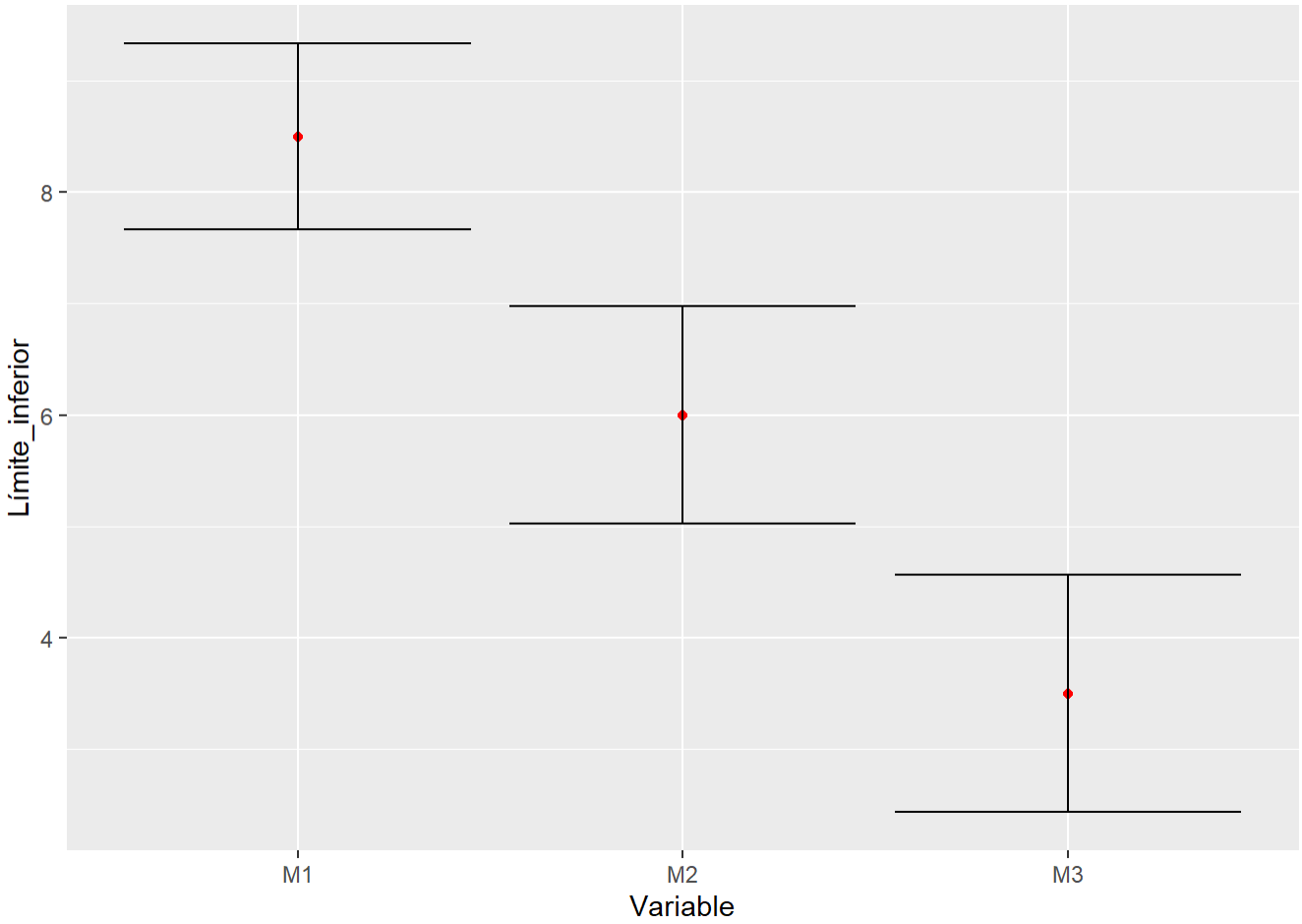
7. Haz los intervalos de confianza de rendimiento por método. Grafícalos

```
conf_int <- tapply(rendimiento, metodo, function(x) t.test(x)$conf.int)

# Crear un vector con los nombres de los grupos
grupos <- names(conf_int)

# Imprimir los intervalos de confianza
for (i in 1:length(grupos)) {
  cat("Intervalo de confianza para", grupos[i], ":\n")
  cat("  Límite inferior:", conf_int[[i]][1], "\n")
  cat("  Límite superior:", conf_int[[i]][2], "\n\n")
}
```

```
Intervalo de confianza para M1 :
  Límite inferior: 7.664961
  Límite superior: 9.335039

Intervalo de confianza para M2 :
  Límite inferior: 5.023175
  Límite superior: 6.976825

Intervalo de confianza para M3 :
  Límite inferior: 2.433377
  Límite superior: 4.566623
```

```
# Crea un dataframe con los valores
data <- data.frame(
  Variable = c("M1", "M2", "M3"),
  Límite_inferior = c(7.664961, 5.023175, 2.433377),
  Límite_superior = c(9.335039, 6.976825, 4.566623)
)


# Carga la librería ggplot2
library(ggplot2)

# Create the ggplot plot
ggplot(data, aes(x = Variable, y = Límite_inferior, ymin = Límite_inferior, ymax = Límite_super
```

```
  geom_point(aes(y = tapply(rendimiento,metodo,mean)), color = "red") +
  geom_errorbar()
```

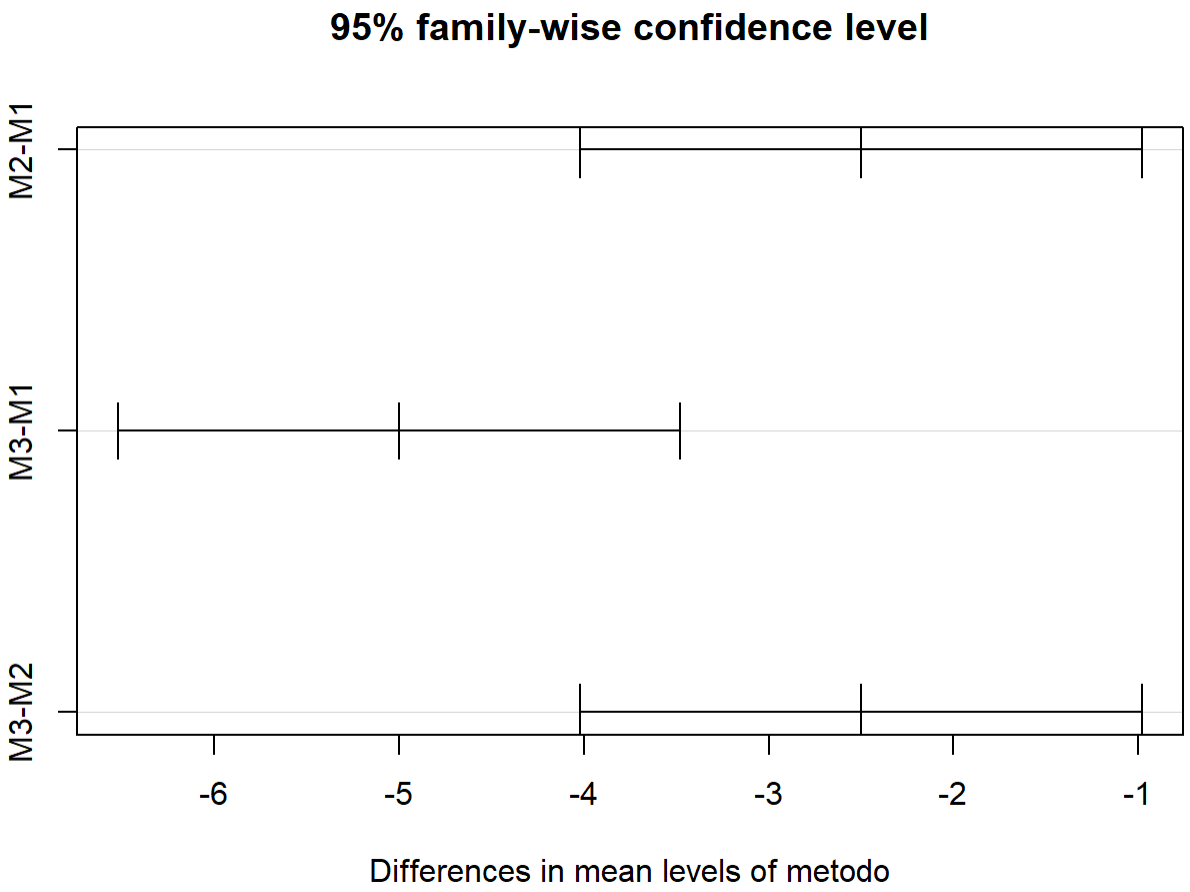


8. Realiza la prueba de comparaciones múltiples de Tukey. Grafica los intervalos de confianza de Tukey.

```
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
```



9. Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. Escribe tus conclusiones parciales

El análisis indica que los métodos de enseñanza afectan significativamente al rendimiento estudiantil. El Método 3 reduce el rendimiento, el Método 2 no tiene un impacto significativo y el Método 1 mejora el rendimiento en comparación con la media. El modelo explica el 66.37% de la variación en el rendimiento, siendo el método de enseñanza el único factor significativo de los 3 que comparamos esta vez. El 32.73% cae dentro de la variacion que no es explicada por nuestro modelo (metodo). El estudio tiene un diseño equilibrado y datos que siguen una distribución normal, respaldando las conclusiones del análisis.

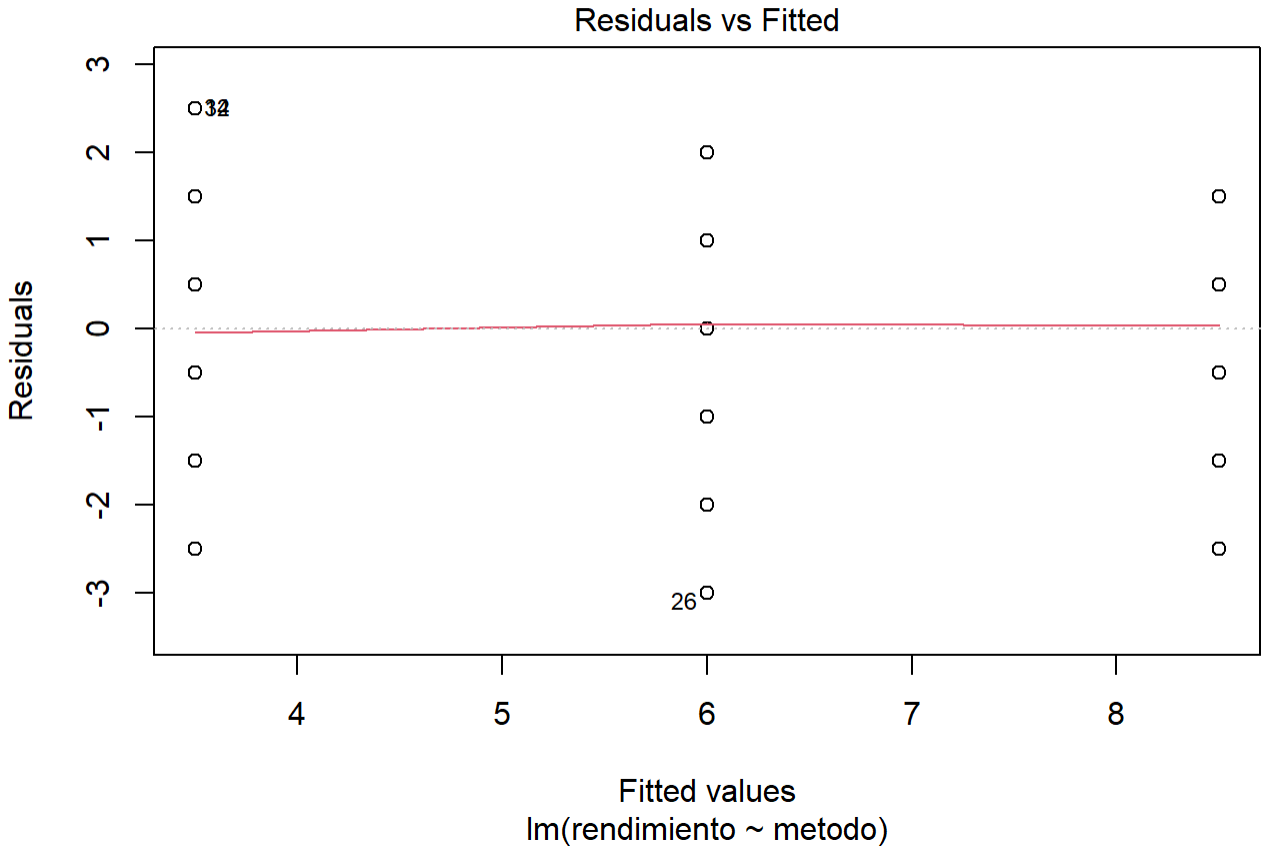
10. Comprueba la validez del modelo. Comprueba:

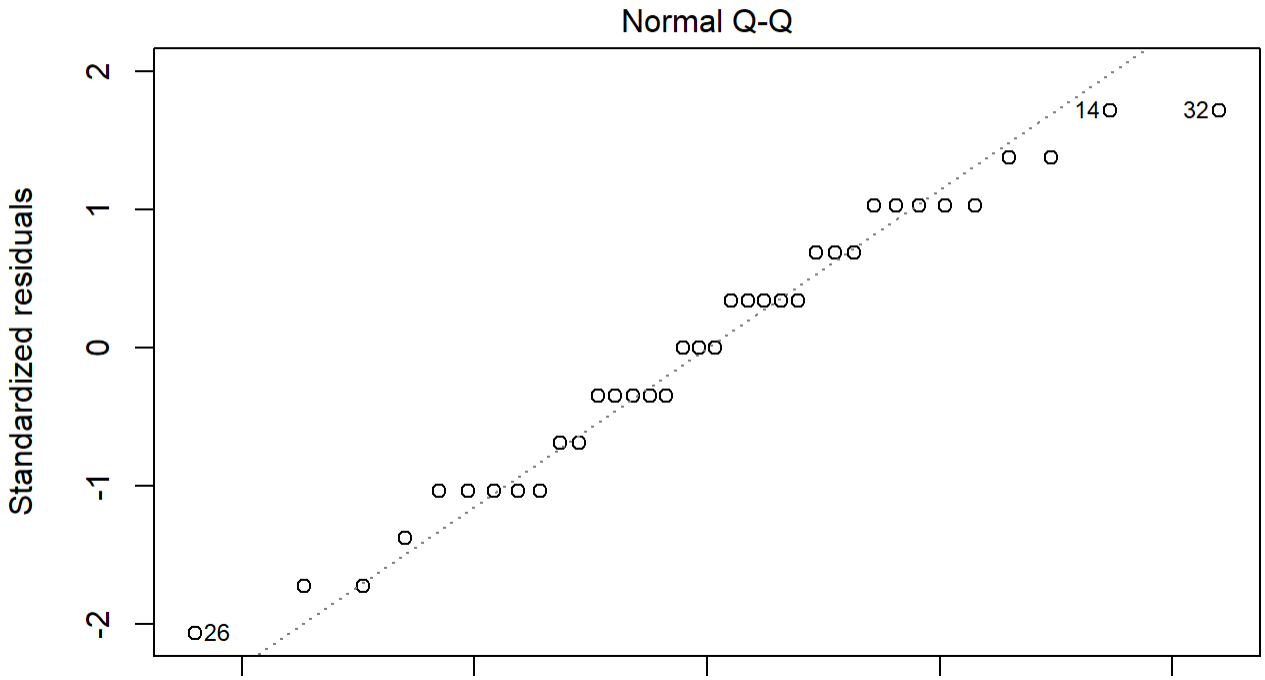
Normalidad Homocedasticidad Independencia Relación lineal entre las variables (coeficiente de determinación). Concluye en el contexto del problema.

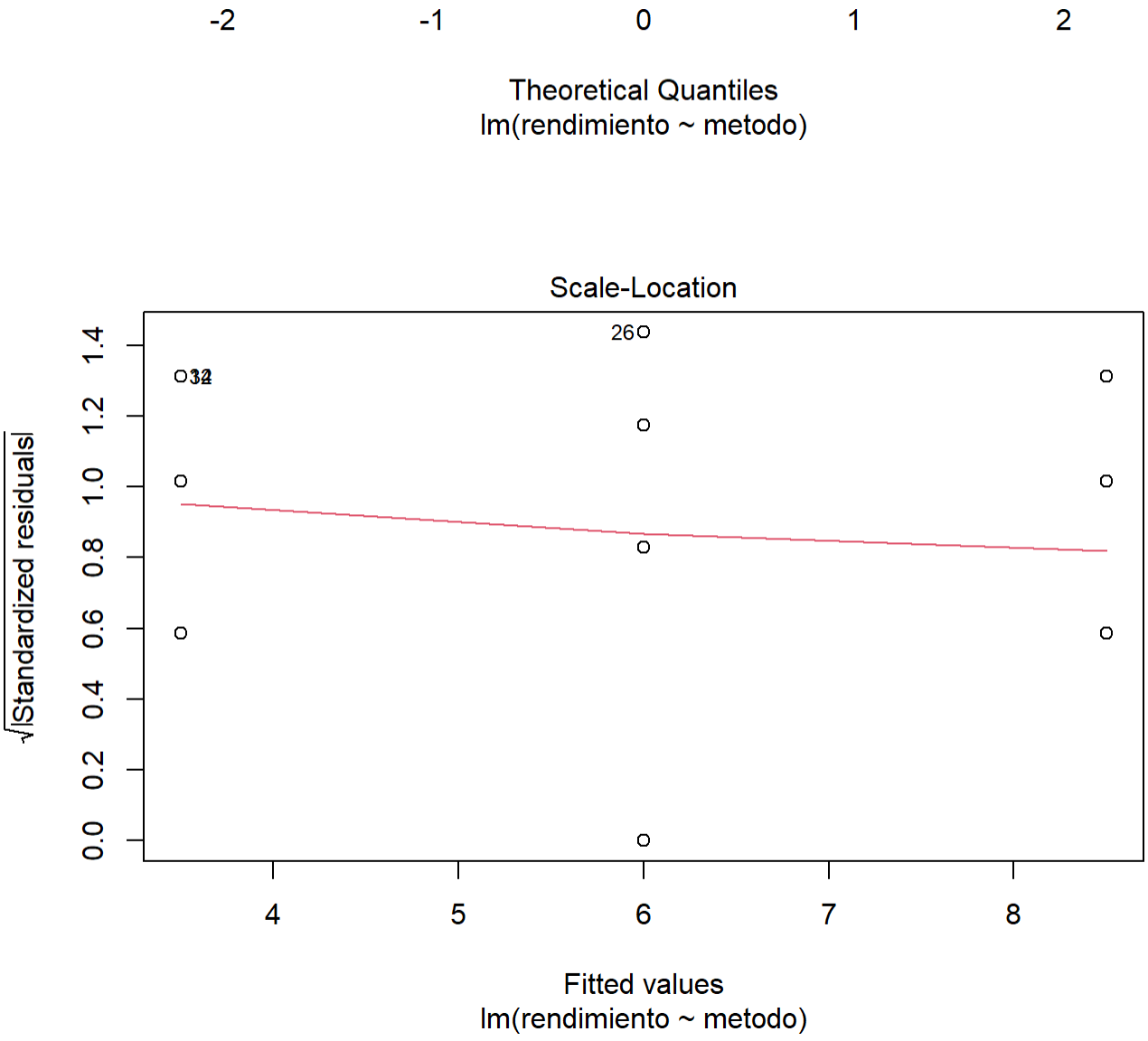
Análisis del modelo

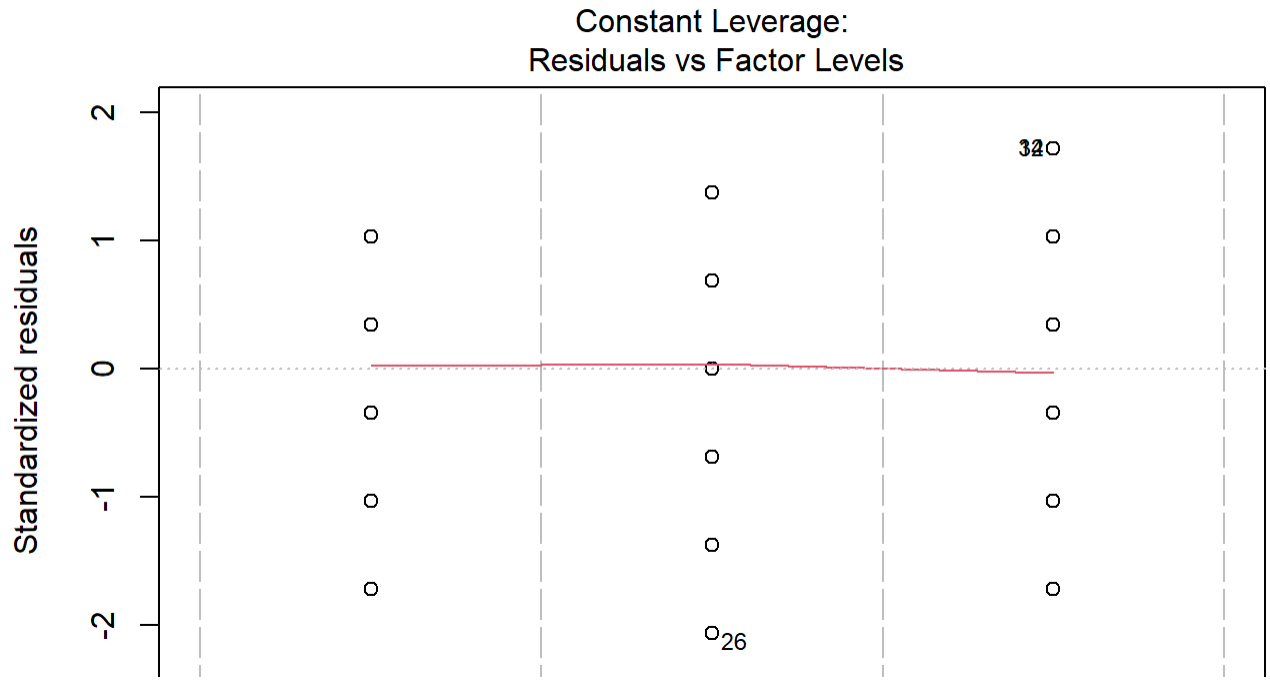
Se verifica la validez del modelo por medio de las gráficas de residuos y la gráfica de normalidad. También se pueden calcular los coeficientes de determinación del modelo para conocer la variación explicada por el modelo.

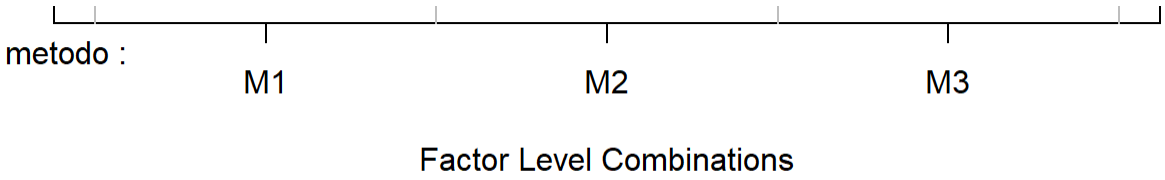
```
plot(lm(rendimiento~metodo))
```






-2
-1
0
1
2
Theoretical Quantiles
lm(rendimiento ~ metodo)
Scale-Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
26
4
5
6
7
8
Fitted values
lm(rendimiento ~ metodo)


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
2
1
0
-1
-2
26



Factor Level Combinations

```
CD= 150/(150+76) #coeficiente de
CD
```

```
[1] 0.6637168
```

Los 3 métodos de enseñanza tienen efectos distintos en el rendimiento de los niños. El Método 3 reduce el rendimiento, el Método 2 no tiene efecto y el Método 1 mejora el rendimiento en comparación con la media. El modelo explica el 66.37% de la variación, destacando el Método de enseñanza como un factor determinante (el único significativo). El 32.73% restante de variación se atribuye a otros factores o aleatoriedad. El estudio tiene un diseño equilibrado y los datos parecen seguir una distribución normal e independiente según los gráficos Q-Q y los residuos vs. el valor esperado. Los errores tienen media cero y variación constante, respaldando las conclusiones del análisis.