Trabajo Práctico P 6 - ANOVA

María Eugenia Fontecha

31 octubre, 2020

Ejercicio 1

La fasciolasis es una enfermedad parasitaria producida por el trematodo Fasciola hepática. Los trematodos adultos viven en el conducto biliar del huésped, donde segregan cantidades significativas de ciertos aminoácidos, en especial prolina. El huésped suele presentar anemia (reducción de los glóbulos rojos de la sangre), supuestamente por acción de los aminoácidos segregados. Para demostrarlo, se tomaron 40 ratas Wistar sanas de aproximadamente igual peso y edad, y se dividieron al azar en 4 grupos de 10 ratas cada uno. Se infundió material directamente en el conducto biliar de las ratas.

- **Las ratas de *grupo I* recibieron 20 milimoles de prolina disuelta en suero fisiológico;**
- **Las del *grupo II* recibieron un cóctel consistente en siete aminoácidos (excluyendo la prolina) segregados por el trematodo, también disuelto en suero fisiológico;**
- **Las del *grupo III* recibió lo mismo que el II más el agregado de 20 milimoles de prolina (simulando a lo segregado por el trematodo);**
- **El *grupo IV* sólo se trató con suero fisiológico.**

En todos los casos se tomó como variable el número de glóbulos rojos del huésped, expresados en millones/mm3 de sangre. Los datos resultantes se encuentran en el archivo *prolina.csv*.

1. **Describa la pregunta PICO. Escriba el modelo, en general y en términos del problema. ¿Cuál es el diseño utilizado en el experimento?**

- **P**: 40 ratas Wistar sanas aproximadamente de igual edad y peso.
- **I**: infusión de material en el conducto biliar de las ratas. Este material podía ser un suero de 7 aminoácidos sin prolina o cóctel de 7 aminoácidos más 20 milimoles de prolina, según el grupo.
- **C**: se compara con el grupo control, al que se le infundió suero fisiológico en el conducto biliar.
- **O**: número de glóbulos rojos, en $\frac{\text{millones}}{\text{mm}^3}$.

El modelo estadístico está dado por:

- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$,

donde:

- y_{ij} es la observación j del tratamiento i. En términos del problema, sería el número de glóbulos rojos de un ratón que se sometió a determinado tratamiento según su grupo.
- μ es la media de la población. En términos del problema, es la media del NGR de la población.
- α_i es el efecto del tratamiento. En este estudio, es el efecto que tiene la infusión de determinado material en el conducto biliar en el número de glóbulos rojos.
- i son los niveles de tratamiento. En términos del problema, son los cuatro tratamientos posibles: los tres de inyección de aminoácidos y el control.
- ϵ_{ij} es el error de la observación j y el tratamiento i. La variable aleatoria del error tiene una distribución aproximada a $N(0, \sigma^2)$.

El diseño utilizado en el experimiento es el de ensayo clínico, ya que se tiene una muestra de ratones y se la separa aleatoriamente en 4 grupos, de los cuales a 3 se les realiza una intervención y se compara con el cuarto grupo que es el de control.

2. **Describa gráfica y analíticamente los datos. ¿Identifica algún outlier?**

```
setwd(path)
```

```
prolina <- read.csv2('prolina.csv')
summary(prolina)
```

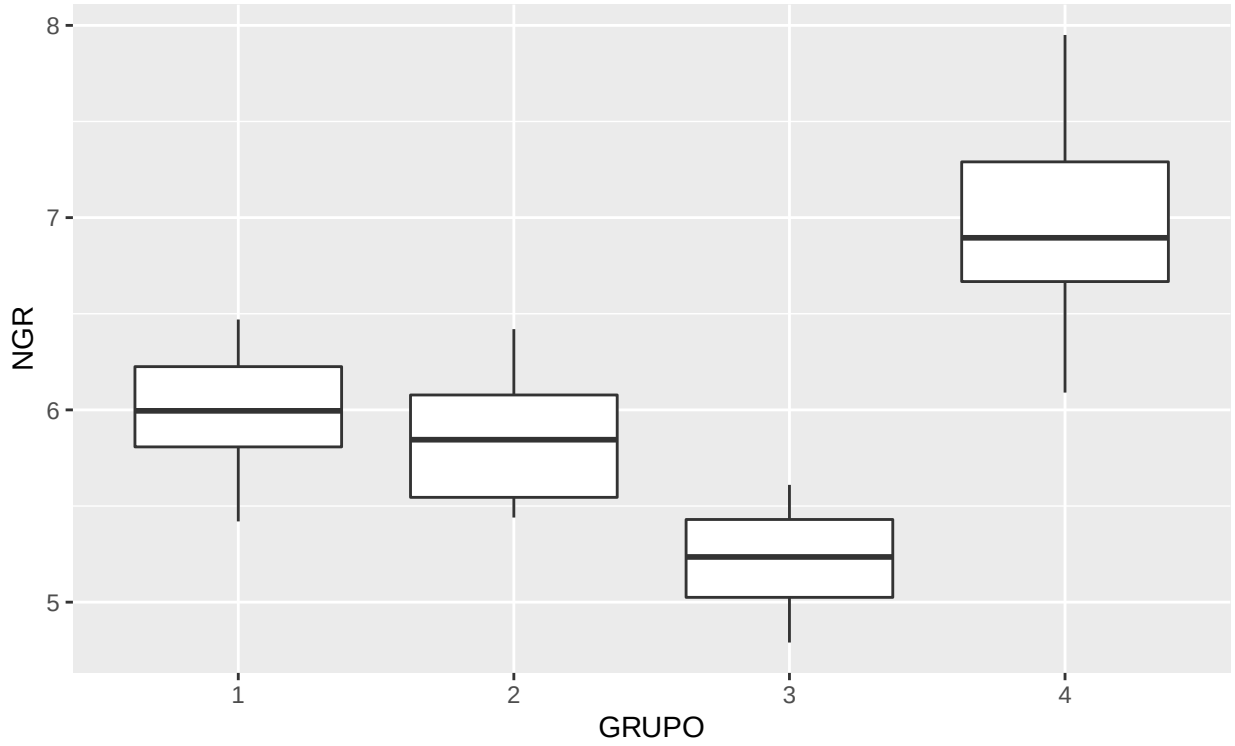
```
##      GRUPO          NGR
##  Min.   :1.00   Length:40
##  1st Qu.:1.75   Class :character
##  Median :2.50   Mode  :character
##  Mean   :2.50
##  3rd Qu.:3.25
##  Max.   :4.00
```

Convierto la variable GRUPO en factor y la variable NGR en numeric.

```
prolina$GRUPO <- as.factor(prolina$GRUPO)
prolina$NGR <- as.numeric(prolina$NGR)
summary(prolina)
```

```
##  GRUPO       NGR
##  1:10   Min.   :4.790
##  2:10   1st Qu.:5.492
##  3:10   Median :5.945
##  4:10   Mean   :6.014
##         3rd Qu.:6.433
##         Max.   :7.950
```

```
plt <- ggplot(prolina, aes(x=GRUPO ,y=NGR)) +
geom_boxplot()
plt
```



A simple vista se puede ver que el grupo 4 posee una mediana mayor a la de los demás grupos, que el grupo 3 posee una mediana menor a la de los demás y que el grupo 1 y 2 poseen una mediana bastante similar entre sí. Además, en todos los grupos la media y la mediana parecen ser similares, lo cual es un indicio de la normalidad de los datos por grupo. No se observan outliers.

```
variables = list( prolina$NGR[prolina$GRUPO == 1], prolina$NGR[prolina$GRUPO == 2],
                  prolina$NGR[prolina$GRUPO == 3],
                  prolina$NGR[prolina$GRUPO == 4] )

medidas = double(length = 12)

i = 1
for (var in variables){
  k = i+2
  medidas[i:k] = c(round(mean(var), digits=4), round(median(var),digits=4),
                   round(sd(var), digits=4))
  i = i +3
}
filas = c("Media", "Mediana", "Desvío Estándar")

tabla <- data.frame(GRUPO_1 = medidas[1:3],
                    GRUPO_2 = medidas[4:6],
                    GRUPO_3 = medidas[7:9],
                    GRUPO_4 = medidas[10:12], row.names = filas)
kable(tabla, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	GRUPO_1	GRUPO_2	GRUPO_3	GRUPO_4
Media	5.9820	5.8690	5.2270	6.9780
Mediana	5.9950	5.8450	5.2350	6.8950
Desvío Estándar	0.3318	0.3516	0.2708	0.5477

A partir de esta descripción analítica también podemos ver que los grupos 1 y 2 tienen valores similares, mientras que el grupo 3 tiene valores menores que los demás grupos y el grupo 4 tiene valores mayores.

3. **Plantee las hipótesis y efectúe el análisis de la varianza. Compruebe los supuestos de normalidad y homocedasticidad. Concluya, asumiendo un nivel de significación del 5%.**

Las hipotesis de ANOVA son:

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$
- H_1: algún $\mu_i \neq \mu$,

siendo μ_i la media de el grupo i y μ la media global.

Compruebo que se cumplan los supuestos de normalidad y homocedasticidad.

Normalidad de la variable respuesta

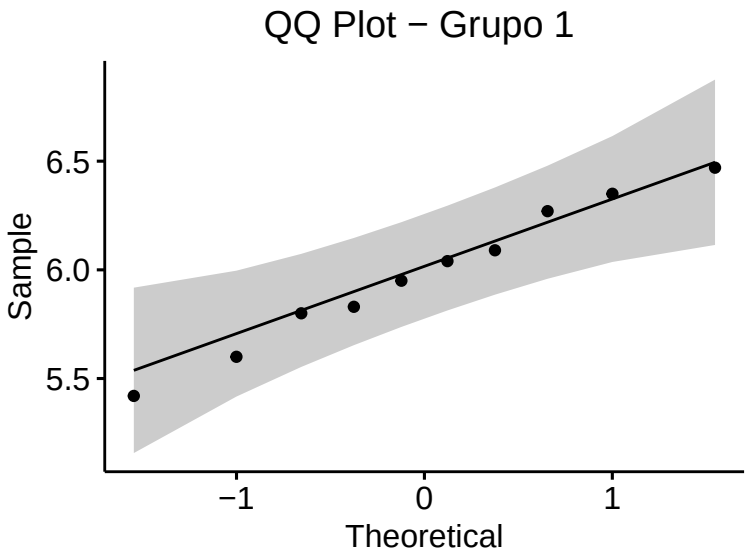
Grupo 1

```
grupo1 <- prolina$NGR[prolina$GRUPO == 1]

print(shapiro.test(grupo1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo1
## W = 0.98028, p-value = 0.9666
```

```
ggqqplot(grupo1) +
  ggtitle("QQ Plot - Grupo 1 ") +
  theme(plot.title = element_text(hjust = 0.5))
```



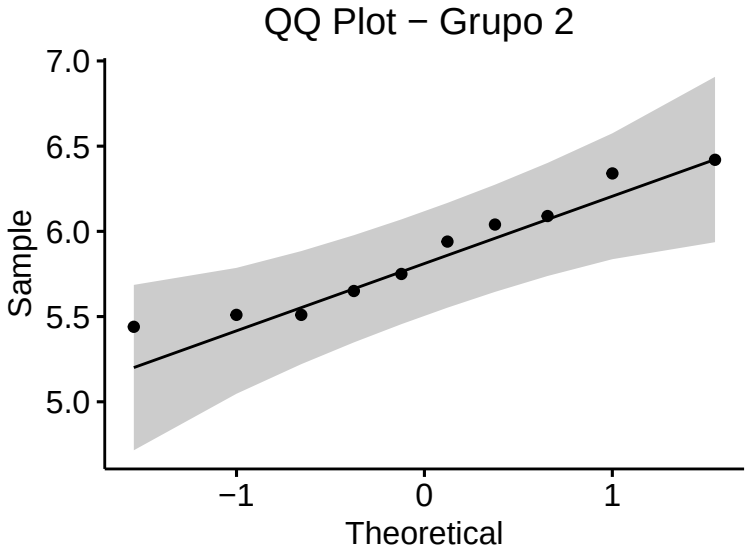
Grupo 2

```
grupo2 <- prolina$NGR[prolina$GRUPO == 2]

print(shapiro.test(grupo2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo2
## W = 0.92626, p-value = 0.4121
```

```
ggqqplot(grupo2) +
  ggtitle("QQ Plot - Grupo 2 ") +
  theme(plot.title = element_text(hjust = 0.5))
```



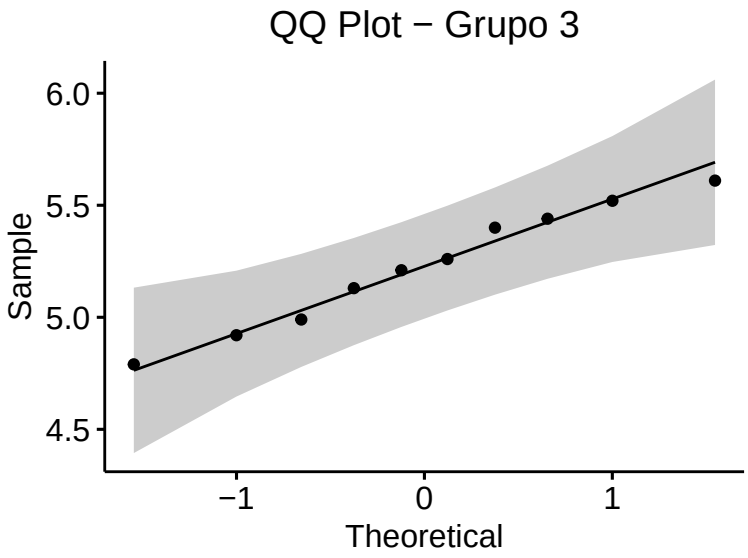
Grupo 3

```r
grupo3 <- prolina$NGR[prolina$GRUPO == 3]

print(shapiro.test(grupo3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo3
## W = 0.96833, p-value = 0.875
```

```r
ggqqplot(grupo3) +
  ggtitle("QQ Plot - Grupo 3") +
  theme(plot.title = element_text(hjust = 0.5))
```



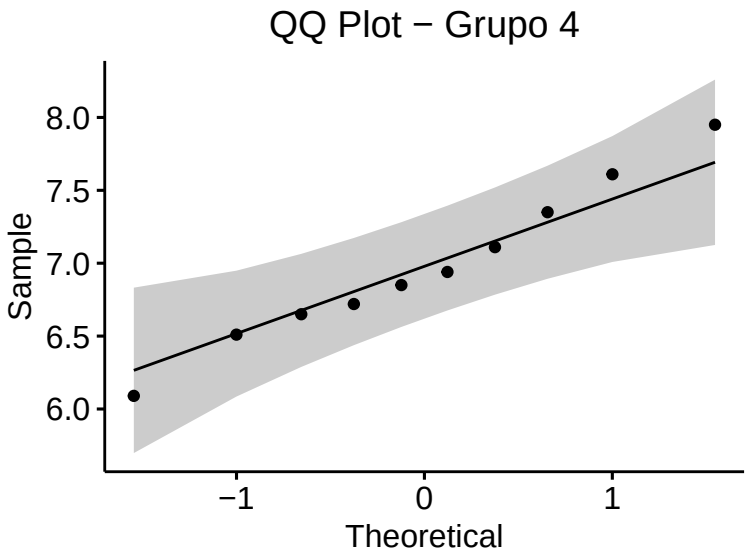
Grupo 4

```r
grupo4 <- prolina$NGR[prolina$GRUPO == 4]

print(shapiro.test(grupo4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo4
## W = 0.98482, p-value = 0.9857
```

```r
ggqqplot(grupo4) +
  ggtitle("QQ Plot - Grupo 4") +
  theme(plot.title = element_text(hjust = 0.5))
```



Se cumple el supuesto de normalidad para todo los puntos, como se puede ver con el método gráfico QQ Plot, donde todos los puntos siguen la recta teórica. Además, se refuerza con el método análitico, el test de Shapiro-Wilk, donde en todos los casos el valor-P fue mayor a 0.05, por lo que no se rechaza la H_0 que dice que la variable presenta distirbución normal.

Homocedasticidad

```
leveneTest(y = prolina$NGR, group = prolina$GRUPO)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.4173 0.2536
##       36
```

El test de Levene tiene como hipótesis nula que las varianzas de los grupos son homogéneas, y como se obtuvo un P-valor mayor a 0.05, no se rechaza esta hipótesis. Entonces, se cumple el supuesto de homocedasticidad.

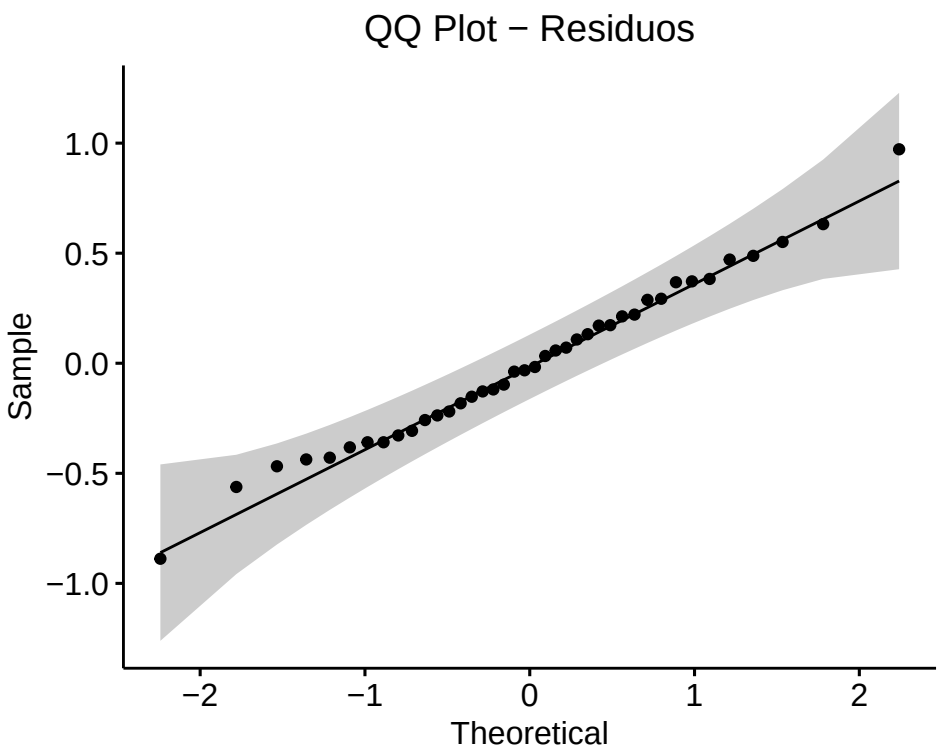
Normalidad de los errores aleatorios

Los errores se estiman con los residuos.

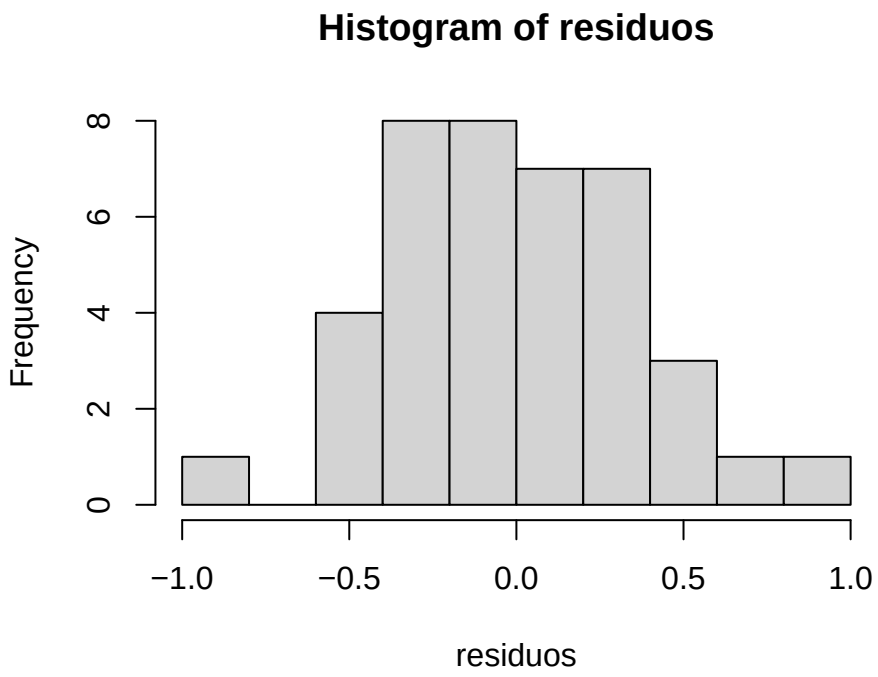
```
reg<-lm(prolina$NGR~prolina$GRUPO) # modelo lineal general
residuos<-residuals(reg) # Estima los residuos
print(shapiro.test(residuos))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.99124, p-value = 0.9872
```

```r
ggqqplot(residuos) +
  ggtitle("QQ Plot - Residuos") +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
hist(residuos)
```



Se puede ver tanto en el histograma como en el QQ Plot que los residuos presentan una distrubción normal. Además, con el test de normalidad también se puede concluir que los residuos tienen una distribución normal, porque el p-valor es mayor 0.05.

Test ANOVA

Una vez comprobados los supuestos, realizo el test ANOVA.

```
anova<-aov(prolina$NGR~prolina$GRUPO)
summary(anova)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## prolina$GRUPO  3 15.707   5.236   34.51 1.09e-10 ***
## Residuals     36  5.463   0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con un p-valor menor a 0.05 se rechaza la hipótesis nula, por lo que sabemos que alguna de las medias de los grupos es distinta de las demás, pero no sabemos cuál.

4. **Efectúe las comparaciones entre tratamientos que considere necesarias.**

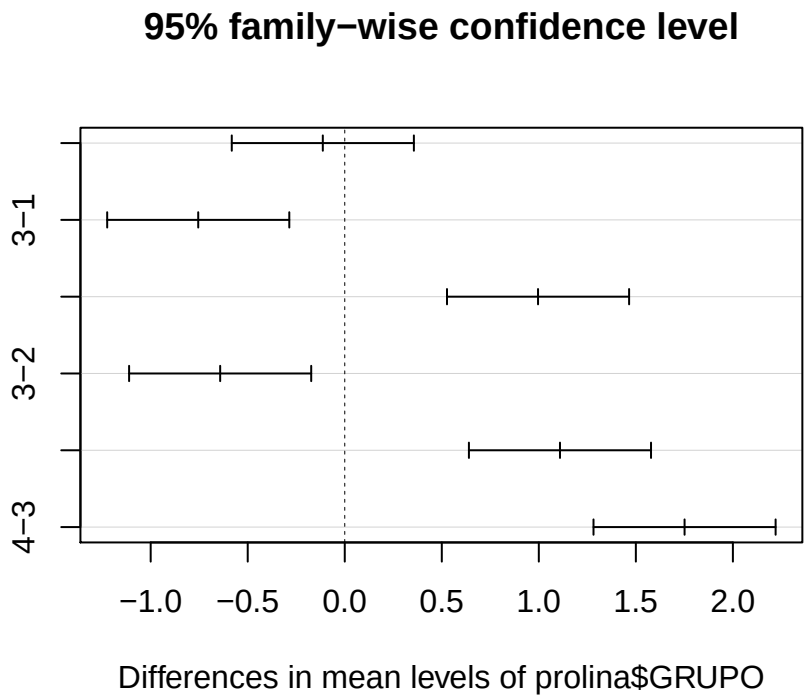
Como tengo 4 grupos, realizo una prueba de Tukey que me permite comparar todos los pares posibles de medias.

```r
Tukey <- TukeyHSD(anova, conf.level =  0.95)

print(Tukey)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = prolina$NGR ~ prolina$GRUPO)
##
## $`prolina$GRUPO`
##       diff        lwr        upr     p adj
## 2-1 -0.113 -0.5821776  0.3561776 0.9153301
## 3-1 -0.755 -1.2241776 -0.2858224 0.0006260
## 4-1  0.996  0.5268224  1.4651776 0.0000096
## 3-2 -0.642 -1.1111776 -0.1728224 0.0039845
## 4-2  1.109  0.6398224  1.5781776 0.0000013
## 4-3  1.751  1.2818224  2.2201776 0.0000000
```

```r
plot(Tukey)
```



El gráfico muestra de forma fácil de ver los intervalos de confianza de las diferencias entre las medias de cada uno de los pares posibles. Se puede ver que los grupos más similares entre sí son el 1 y el 2 ya que tienen el intervalo de confianza más cercano a 0, e incluso el 0 está dentro del intervalo de confianza. Además, vemos que todas las combinaciones del grupo 4 con los demás, son similares entre sí ya que los intervalos de confianza se superponen, y además son diferentes a los de las combinaciones entre los grupos 1, 2 y 3. También podemos ver que los intervalos de confianza de la diferencia de medias del grupo 3 con 1 y 2 se superponen entre sí, y se superponen un poco también con el del grupo 1 con 2.

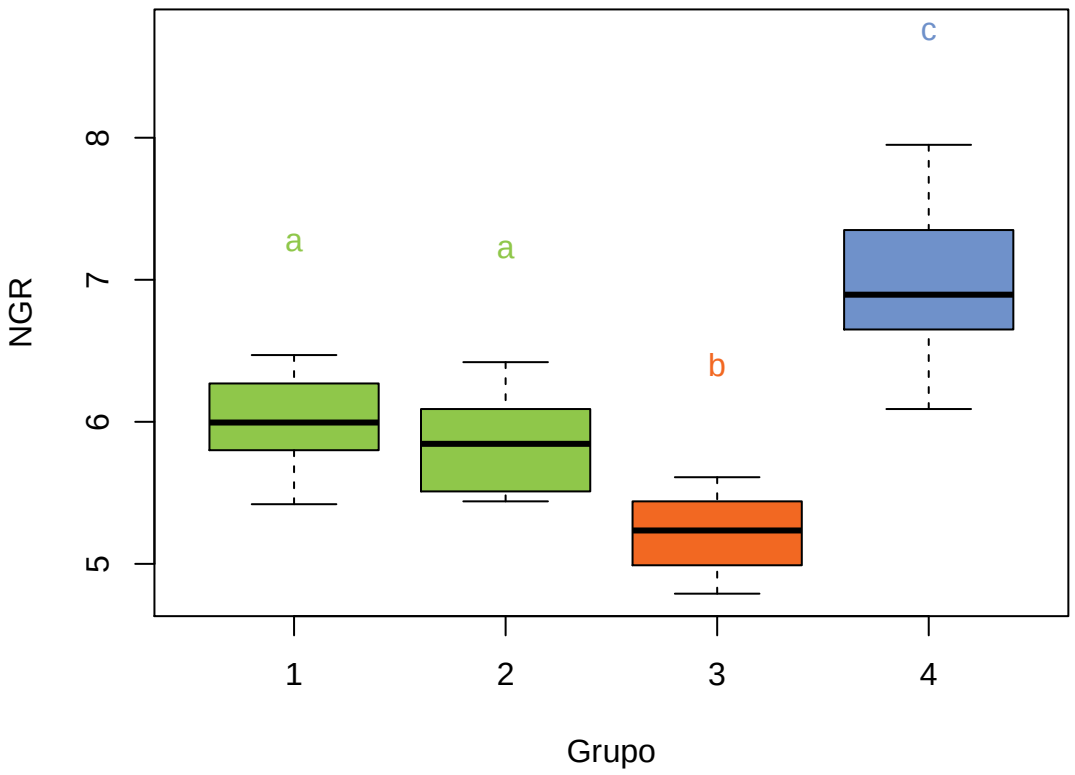
A partir de los resultados numéricos, podemos corrobar algunas de estas observaciones. Por ejemplo, vemos que para el grupo 4 todos las diferencias fueron estadísticamente significativas, con un p-valor menor a 0.05. En el caso del grupo 3, también arrojó diferencias significativas al compararse su media con las de los demás grupos, ya que los p-valores son menores a 0.05. Finalmente, las medias del grupo 1 y 2 no son diferentes ya que se obtuvo un p-valor mayor a 0.05. Entonces, se concluye que los grupos con media diferente son el 3 y el 4.

5. **¿Puede asegurarse que la prolina es la responsable de la anemia observada? Concluya con respecto al experimento. Represente gráficamente los resultados.**

Puede asegurarse que la prolina, en conjunto con los otros 7 aminoácidos, produce anemia, ya que disminuye en 1.751, con un intervalo de confianza del 95% entre 1.28 y 2.22, el recuento de glóbulos rojos. Además, la infusión de 20 milimoles de prolina disminuye el recuento de glóbulos rojos en 0.996, con un intervalo de confianza entre 0.53 y 1.46. Por otro lado, el cóctel de 7 aminoácidos sin prolina, también disminuye el recuento de glóbulos rojos, con una diferencia de 1.109 con un intervalo de confianza del 95% entre 0.64 y 1.58.

```
generate_label_df <- function(TUKEY, variable){
Tukey.levels <- TUKEY[[variable]][,4]
Tukey.labels <- data.frame(multcompLetters(Tukey.levels)['Letters'])
Tukey.labels$treatment=rownames(Tukey.labels)
Tukey.labels=Tukey.labels[order(Tukey.labels$treatment) , ]
return(Tukey.labels)
}
LABELS <- generate_label_df(Tukey , "prolina$GRUPO")
my_colors <- c(
rgb(143,199,74,maxColorValue = 255),
rgb(242,104,34,maxColorValue = 255),
rgb(111,145,202,maxColorValue = 255)
)

LABELS$Numbers = c(1,1,2,3)
a <- boxplot(prolina$NGR ~ prolina$GRUPO , ylim=c(min(prolina$NGR) , 1.1*max(prolina$NGR)),
col=my_colors[(LABELS[,3])] , ylab="NGR", xlab="Grupo")
over <- 0.1*max( a$stats[nrow(a$stats),] )
text( c(1:nlevels(prolina$GRUPO)) , a$stats[nrow(a$stats),]+over , LABELS[,1],
      col=my_colors[(LABELS[,3])] )
```



En este boxplot podemos ver una conclusión de los resultados del test de Tukey. El grupo 1 y 2 no tienen diferencia de s, por lo que se les asigna la misma etiqueta, mientras que al grupo 3 y 4 se les asignan otras dos etiquetas distintas, ya que sus medias son diferentes de las de los demás grupos.

Ejercicio 2

Se sabe que las radiaciones ionizantes utilizadas con fines terapéuticos tienen efectos secundarios a nivel vascular, como daño endotelial, fibrosis, etc. Se sospecha además que estos efectos aterogénicos de la radiación pueden ser más profundos si se combinan con hipercolesterolemia. A fin de estudiarlo, se utilizaron 24 ratones que se dividieron en 6 grupos de igual tamaño. Cada grupo fue sometido a una combinación de dosis de irradiación (4 Gy, 8 Gy o a una simulación de irradiación) y de dieta (estándar o rica en grasas). A las 18 semanas los ratones fueron sacrificados y se les extrajo la aorta, que fue analizada histológicamente. Se determinó el área aórtica lesionada (en µm2). Los datos se encuentran en el archivo *radiacion.csv*.

```
radiacion <- read.csv2('radiacion.csv')
summary(radiacion)
```

```
##    ï..RatÃ³n      Radiacion           Dieta              Ã.rea
##  Min.   : 1.00   Length:24          Length:24          Min.   :0.1400
##  1st Qu.: 6.75   Class :character   Class :character   1st Qu.:0.5775
##  Median :12.50   Mode  :character   Mode  :character   Median :0.9550
```

```
##  Mean   :12.50                                                  Mean   :2.0408
##  3rd Qu.:18.25                                                  3rd Qu.:3.1375
##  Max.   :24.00                                                  Max.   :8.1300
```

Primero voy a cambiar los nombres de las columnas que tienen caracteres especiales para facilitar el manejo de los datos.

```
names <- colnames(radiacion)
radiacion <- radiacion %>%
  rename(
    Raton = names[1],
    Area = names[4]
    )
summary(radiacion)
```

```
##      Raton        Radiacion            Dieta                Area
##  Min.   : 1.00   Length:24          Length:24          Min.   :0.1400
##  1st Qu.: 6.75   Class :character   Class :character   1st Qu.:0.5775
##  Median :12.50   Mode  :character   Mode  :character   Median :0.9550
##  Mean   :12.50                                         Mean   :2.0408
##  3rd Qu.:18.25                                         3rd Qu.:3.1375
##  Max.   :24.00                                         Max.   :8.1300
```

Ahora hago lo mismo con los valores de Dieta.

```
dietas <- unique(radiacion$Dieta)
```

Reemplazo de la siguiente forma:

- Dieta rica en grasas $\longrightarrow$ 1
- Dieta estándar $\longrightarrow$ 0

```
radiacion$Dieta[radiacion$Dieta == dietas[1]] <- 0
radiacion$Dieta[radiacion$Dieta == dietas[2]] <- 1
radiacion$Dieta <- as.factor(radiacion$Dieta)
str(radiacion$Dieta)
```

```
##  Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 2 1 1 ...
```

Cambio la variable Radiacion a tipo factor.

```
radiacion$Radiacion <- as.factor(radiacion$Radiacion)
str(radiacion$Radiacion)
```

```
##  Factor w/ 3 levels "0 Gy","4 Gy",..: 1 1 1 1 1 1 1 1 2 2 ...
```

1. **Describa la pregunta PICO. Escriba el modelo, en general y en términos del problema. ¿Cuál es el diseño utilizado en el experimento?**

- **P**: 24 ratones.

- **I**: irradiación de dosis 4 Gy o 8 Gy combinado con dieta rica en grasas o estándar.
- **C**: se compara con el grupo control, al que se alimentó con dieta estándar y que no recibió radiación.
- **O**: área aórtica lesionada, en μm^2.

El modelo estadístico está dado por:

- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$,

donde:

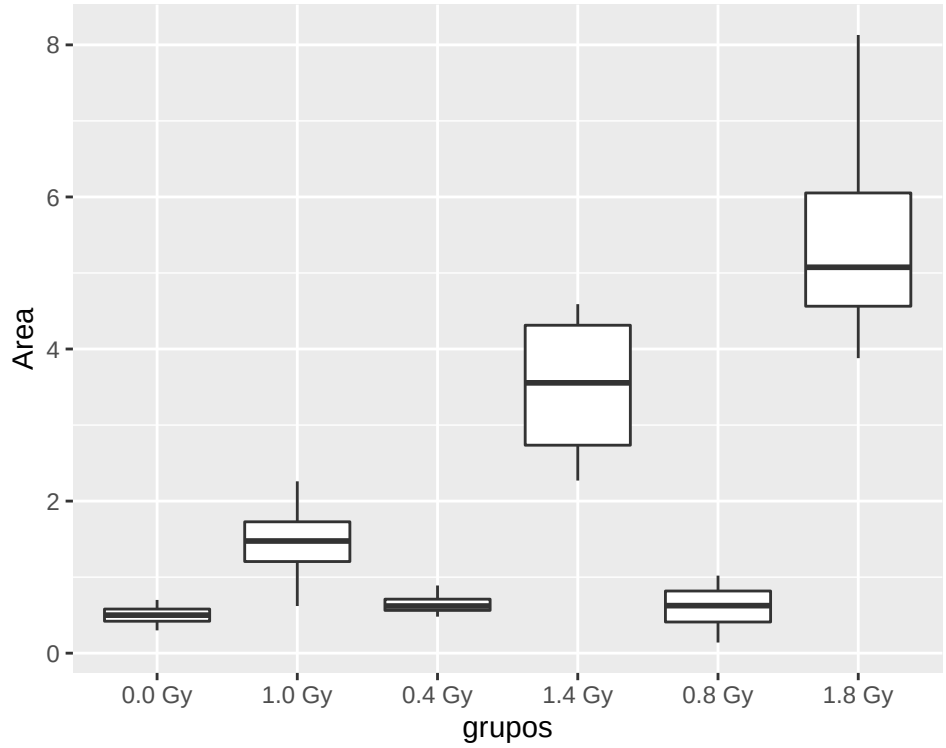
- y_{ij} es la observación j del tratamiento i. En términos del problema, sería el área aórtica lesionada de un ratón que se sometió a determinado tratamiento, es decir la combinación de la dosis de irradiación con la dieta recibida, según su grupo.
- μ es la media de la población. En términos del problema, es la media del área aórtica de la población.
- α_i es el efecto del tratamiento. En este estudio, es el efecto que tiene la dieta diaria y la radiación recibida en el tamaño del área aórtica lesionada.
- i son los niveles de tratamiento. En términos del problema, son los seis tratamientos posibles: dieta estándar + 0Gy, dieta estándar + 4Gy, dieta estándar + 8 Gy, dieta rica en grasas + 0Gy, dieta rica en grasas + 4Gy, dieta rica e grasas + 8Gy.
- ϵ_{ij} es el error de la observación j y el tratamiento i. La variable aleatoria del error tiene una distribución aproximada a $N(0, \sigma^2)$.

El diseño utilizado en este experimento es el de ensayo clínico, ya que se separa aleatoriamente la muestra de ratones y a cada grupo se lo interviene con un tratamiento diferente, excepto al grupo control, y luego se compara el outcome de los distintos grupos.

2. **Describa gráfica y analíticamente los datos.**

Boxplot de los datos

```
grupos = interaction(radiacion$Dieta, radiacion$Radiacion)
plt <- ggplot(radiacion, aes(x=grupos ,y=Area)) +
geom_boxplot()
plt
```



A simple vista en el boxplot podemos ver una similaridad en el área aórtica lesionada entre los grupos que fueron alimentados con dieta estándar, para todas las dosis de radiación. En cuanto a los grupos con dieta alta en grasas, se ve que hay un aumento en el área a medida que aumenta la dosis de radiación.

También se ve que en todos los casos la mediana es similar a la media, aunque es el grupo Dieta Estándar-0 Gy y en el grupo Dieta alta en grasas-8 Gy se ve una pequeña desviación.

```
df <- data.frame(Grupo = c("G1","G2","G3","G4","G5","G6"),
Media = aggregate(x =radiacion$Area, by =list(radiacion$Radiacion,radiacion$Dieta),
FUN=mean),
Desvio = aggregate(x= radiacion$Area, by =list(radiacion$Radiacion,radiacion$Dieta),
FUN=sd)[,3],
Mediana = aggregate(x = radiacion$Area, by =list(radiacion$Radiacion,radiacion$Dieta),
FUN=median)[,3],
row.names = NULL)
df <- df %>%
  rename(
    Dieta = Media.Group.2,
    Dosis = Media.Group.1,
    Media = Media.x
    )
kable(df, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

Grupo	Dosis	Dieta	Media	Desvio	Mediana
G1	0 Gy	0	0.5000	0.1665333	0.500
G2	4 Gy	0	0.6525	0.1732772	0.620
G3	8 Gy	0	0.6025	0.3743773	0.625
G4	0 Gy	1	1.4575	0.6726255	1.475
G5	4 Gy	1	3.4925	1.0941168	3.555
G6	8 Gy	1	5.5400	1.8310835	5.075

Comparando tanto las medias como las medianas podemos ver, igual que en el boxplot, que los grupos de Dieta Estándar (0) tienen valores similares, mientras que los grupos de Dieta alta en grasas (1) tienen valores que aumentan a medida que aumenta la dosis de radiación, e incluso en el caso de dosis 0, estos valores son mayores que los todos de los grupos de Dieta Estándar.

3. **Compruebe los supuestos de normalidad y homocedasticidad. De no comprobarse los supuestos realice la tranformación logarítmica de la variable área.**

Normalidad de los residuos

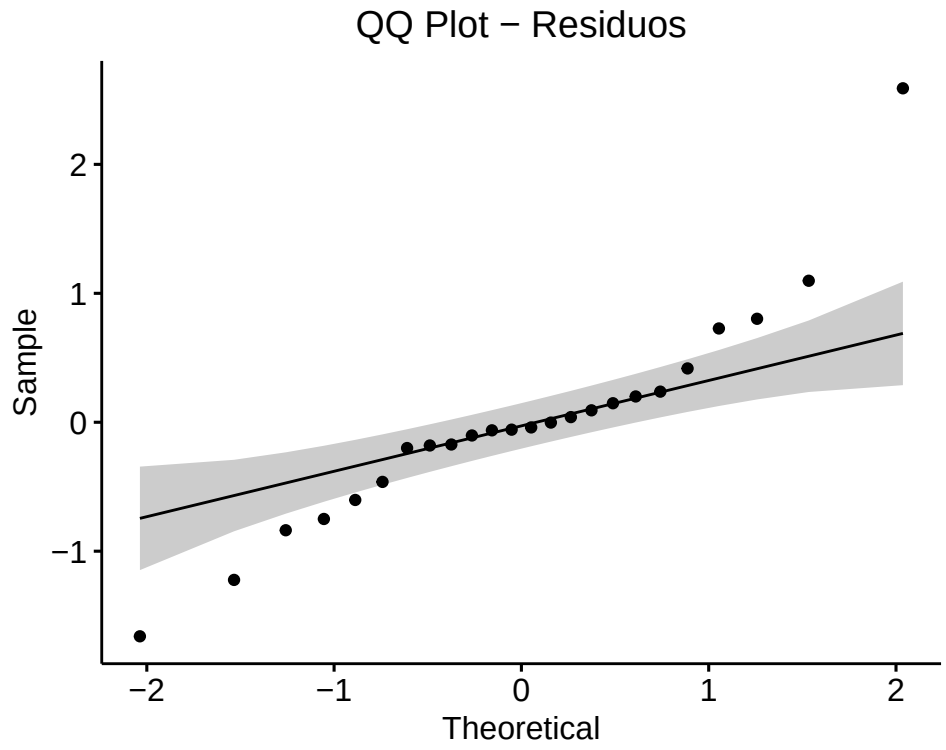
```
reg<-lm(radiacion$Area~radiacion$Radiacion*radiacion$Dieta) # modelo lineal general
residuos<-residuals(reg) # Estima los residuos
print(shapiro.test(residuos))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.90887, p-value = 0.03334
```

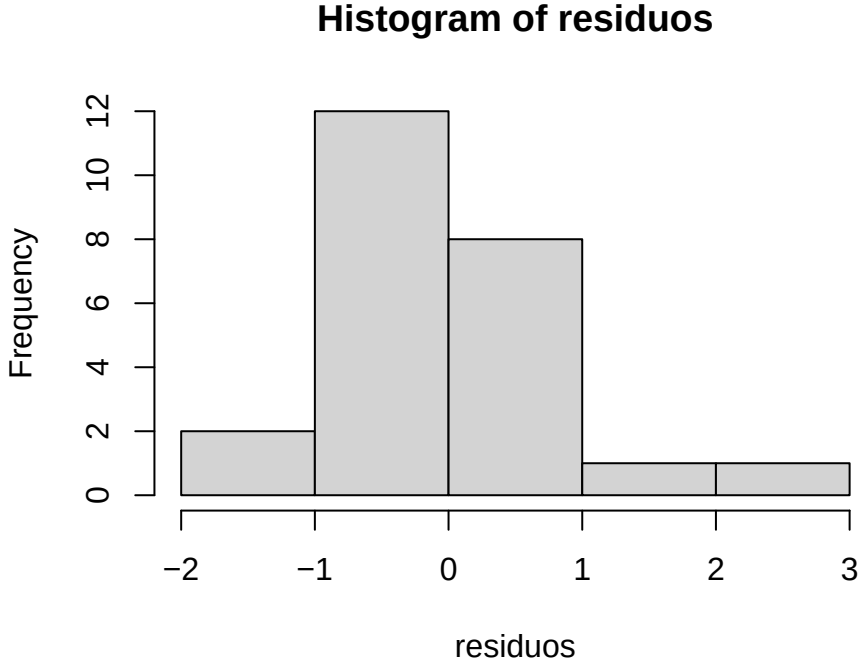
```
print(lillie.test(residuos))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuos
## D = 0.1782, p-value = 0.04717
```

```
ggqqplot(residuos) +
  ggtitle("QQ Plot - Residuos") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
hist(residuos)
```



Podemos ver que el supuesto de la distribución normal de los residuos no se cumple: el QQ-Plot muestra una clara desviación de la línea de tendencia esperada, especialmente en los extremos izquierdo y derecho.

Además, al realizar los tests estadísticos de Shapiro-Wilk y Lilliefors, se obtuvo en ambos casos un p-valor menor a 0.05, lo que implica que se rechace la H_0, entonces la distribución de los residuos no es normal.

Por no cumplirse el supuesto de normalidad de los residuos, realizo la trasformación logarítmica. Genero una nueva variable que es igual al logaritmo natural del área aórtica lesionado.

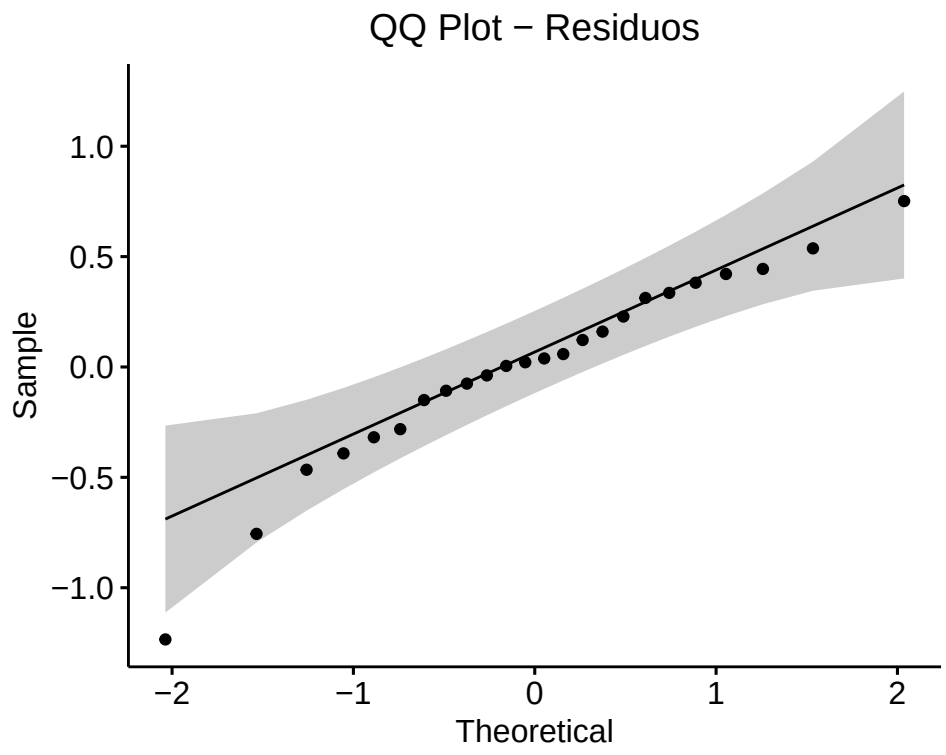
```
radiacion$Arealog <- log(radiacion$Area)
```

Compruebo nuevamente los supuestos.

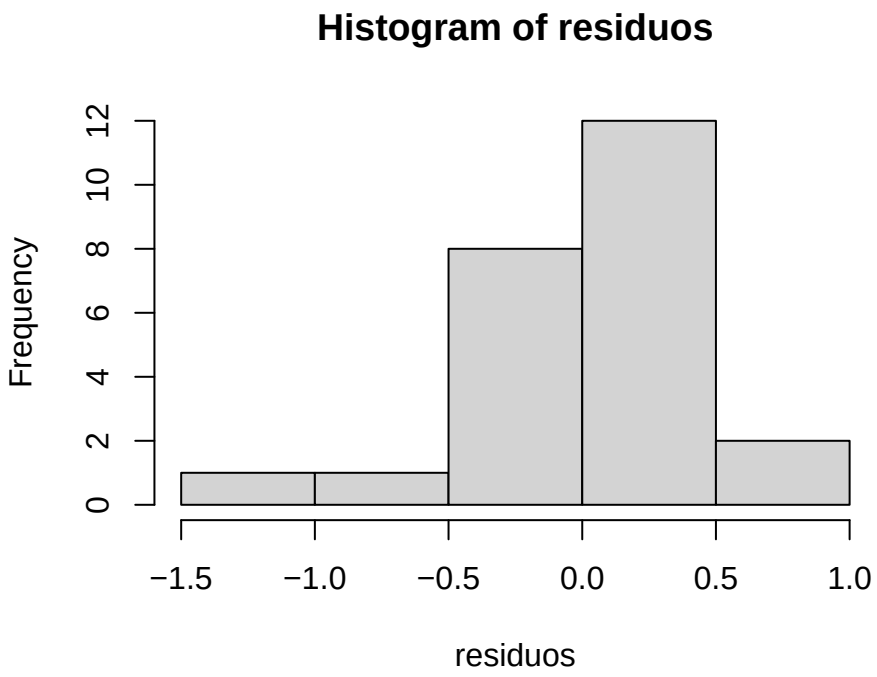
```
reg<-lm(radiacion$Arealog~radiacion$Radiacion*radiacion$Dieta) # modelo lineal general
residuos<-residuals(reg) # Estima los residuos
print(shapiro.test(residuos))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.95157, p-value = 0.2929
```

```
ggqqplot(residuos) +
  ggtitle("QQ Plot - Residuos") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
hist(residuos)
```



Con la tansformación logarítmica, ahora se puede ver que se cumple la normalidad de los residuos. En el QQ-Plot los puntos siguen a la línea de tendencia. Además, se rechaza la hipótesis alternativa del test de normalidad, ya que se obtuvo un p-valor mayor a 0.05. En cuanto a los histogramas, lo que puedo notar es que tienen una distribución similar, pero invertida.

Normalidad de la variable respuesta por grupos

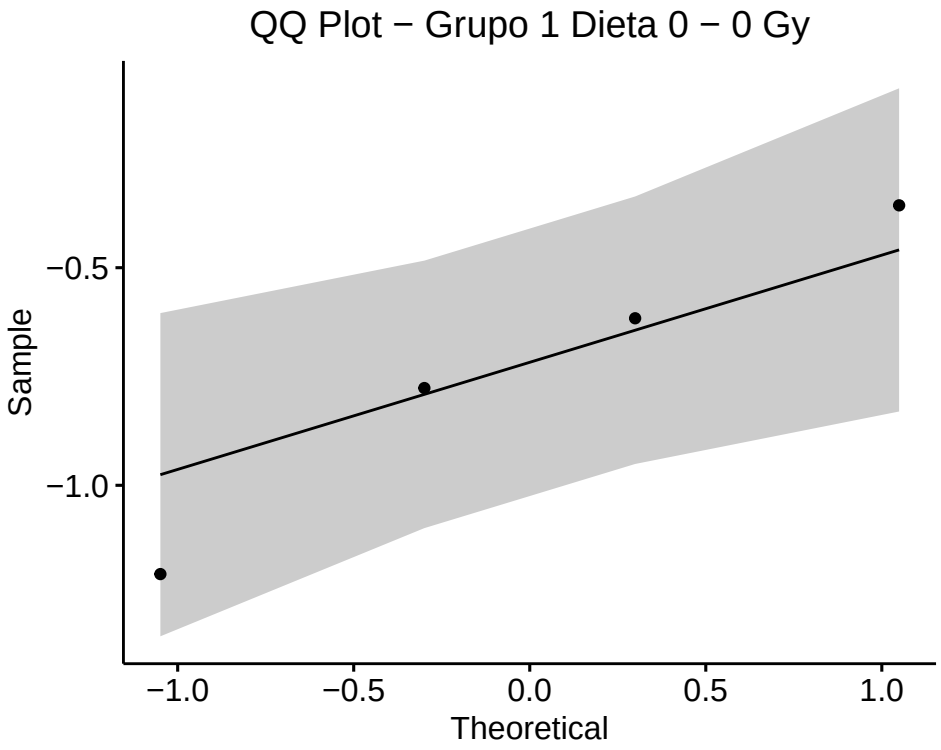
Grupo 1

```
grupo1 <- radiacion$Arealog[radiacion$Dieta == 0 & radiacion$Radiacion == '0 Gy' ]

print(shapiro.test(grupo1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo1
## W = 0.9789, p-value = 0.8955
```

```
ggqqplot(grupo1) +
  ggtitle("QQ Plot - Grupo 1 Dieta 0 - 0 Gy") +
  theme(plot.title = element_text(hjust = 0.5))
```



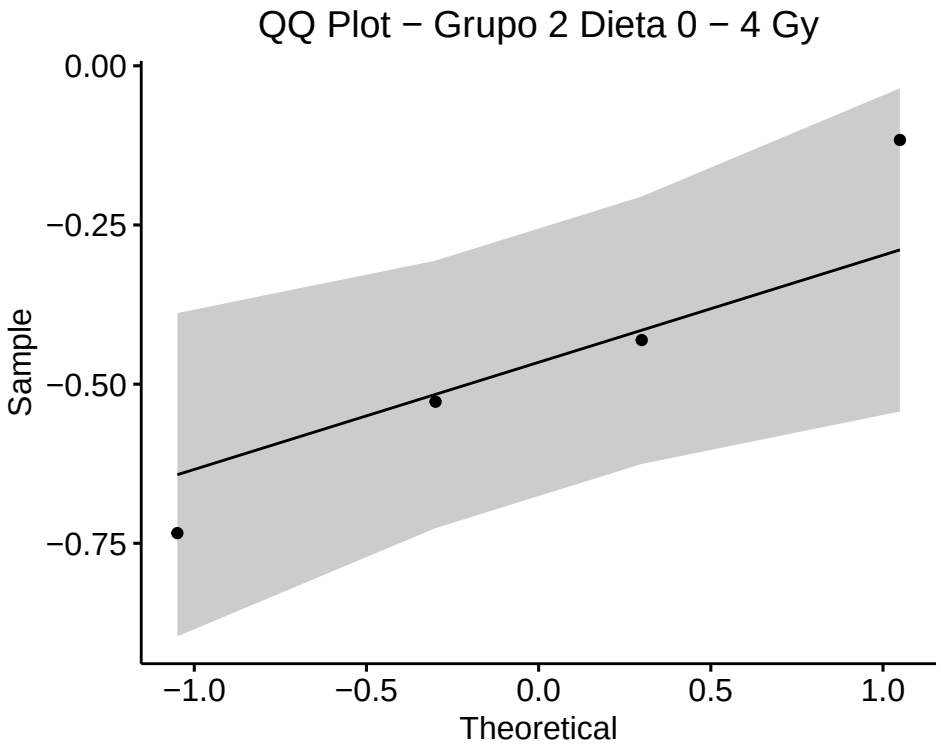
Grupo 2

```
grupo2 <- radiacion$Arealog[radiacion$Dieta == 0 & radiacion$Radiacion == '4 Gy' ]

print(shapiro.test(grupo2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo2
## W = 0.97872, p-value = 0.8945
```

```
ggqqplot(grupo2) +
  ggtitle("QQ Plot - Grupo 2 Dieta 0 - 4 Gy") +
  theme(plot.title = element_text(hjust = 0.5))
```



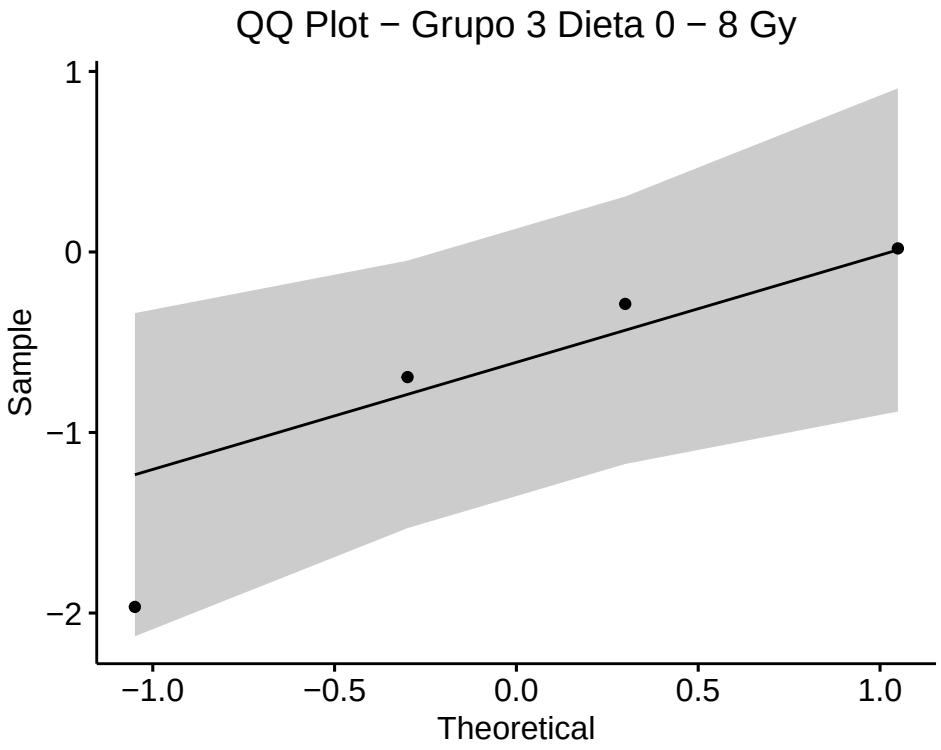
Grupo 3

```
grupo3 <- radiacion$Arealog[radiacion$Dieta == 0 & radiacion$Radiacion == '8 Gy' ]

print(shapiro.test(grupo3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo3
## W = 0.89694, p-value = 0.416
```

```
ggqqplot(grupo3) +
  ggtitle("QQ Plot - Grupo 3 Dieta 0 - 8 Gy") +
  theme(plot.title = element_text(hjust = 0.5))
```



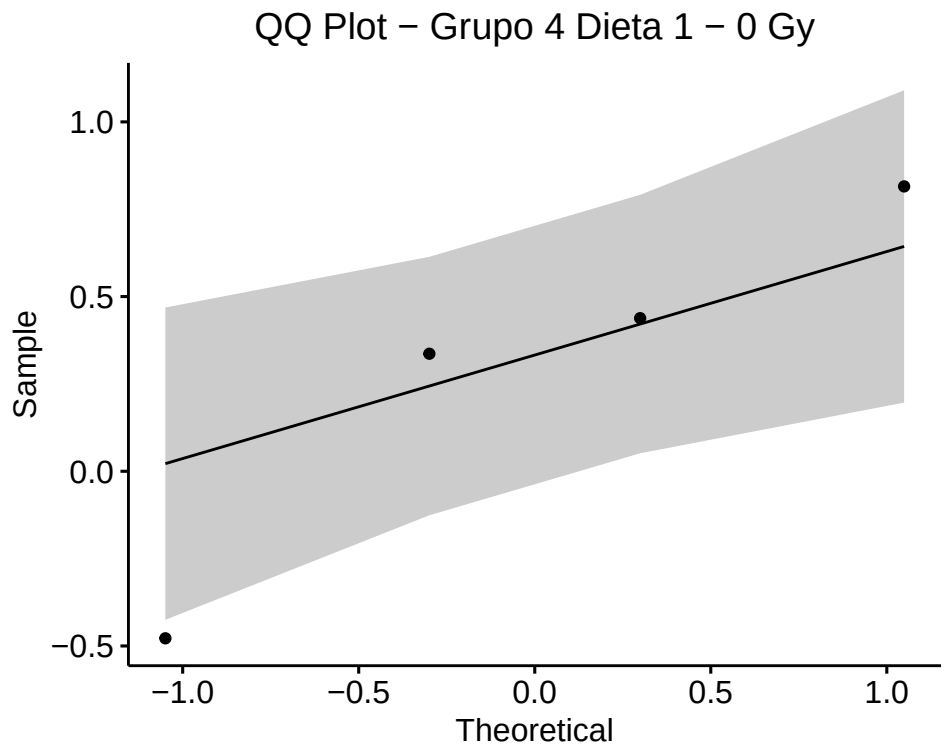
Grupo 4

```
grupo4 <- radiacion$Arealog[radiacion$Dieta == 1 & radiacion$Radiacion == '0 Gy' ]

print(shapiro.test(grupo4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo4
## W = 0.92253, p-value = 0.5512
```

```
ggqqplot(grupo4) +
  ggtitle("QQ Plot - Grupo 4 Dieta 1 - 0 Gy") +
  theme(plot.title = element_text(hjust = 0.5))
```



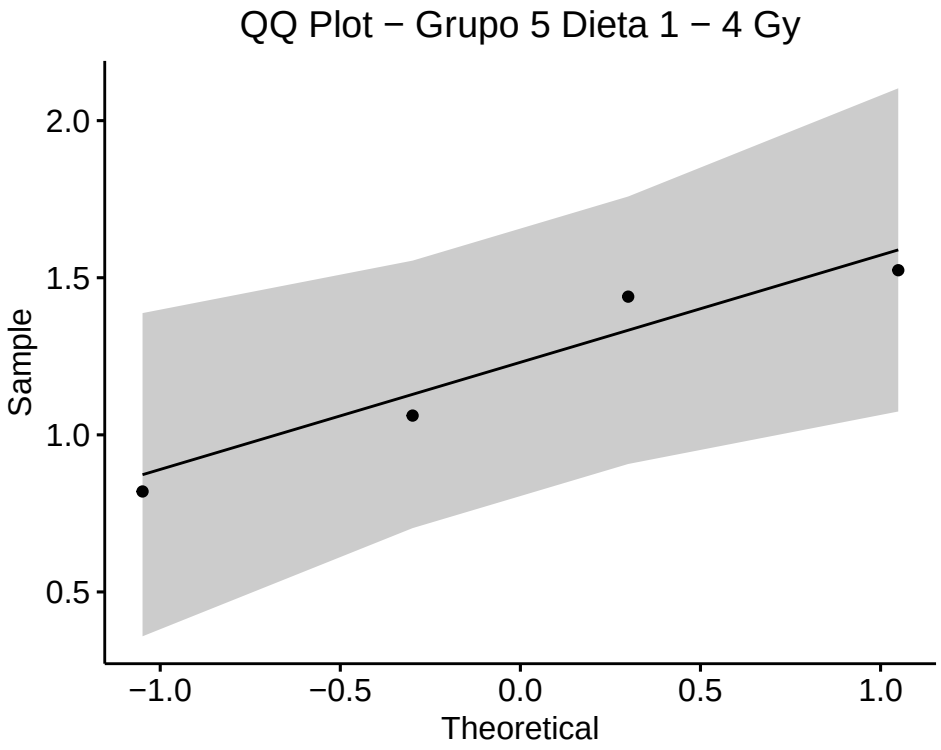
Grupo 5

```
grupo5 <- radiacion$Arealog[radiacion$Dieta == 1 & radiacion$Radiacion == '4 Gy' ]

print(shapiro.test(grupo5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo5
## W = 0.91814, p-value = 0.5266
```

```
ggqqplot(grupo5) +
  ggtitle("QQ Plot - Grupo 5 Dieta 1 - 4 Gy") +
  theme(plot.title = element_text(hjust = 0.5))
```



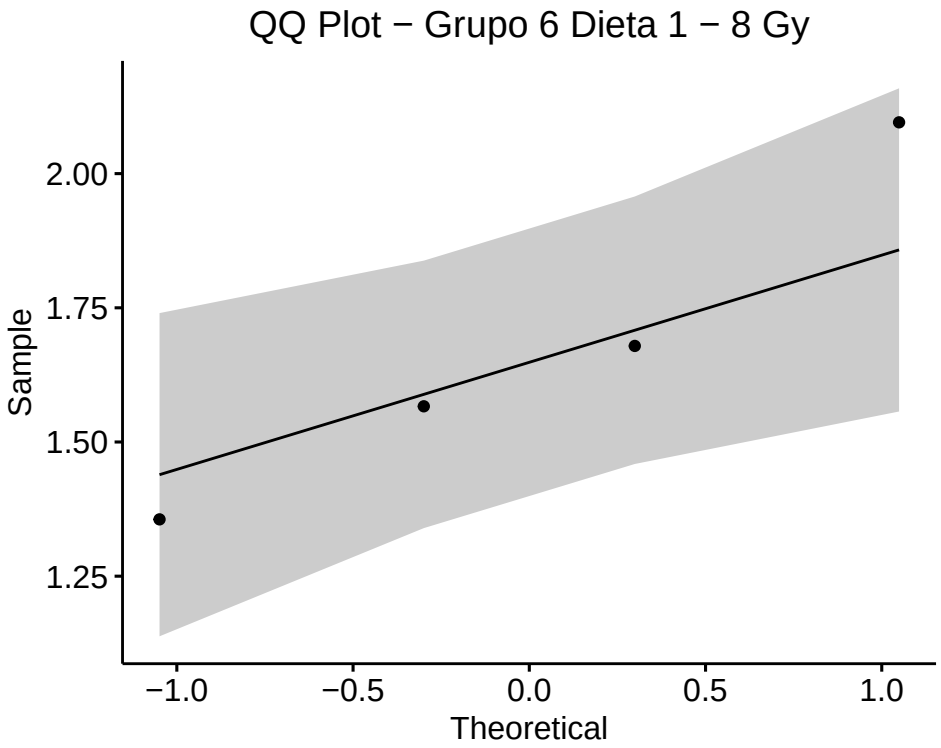
Grupo 6

```
grupo6 <- radiacion$Arealog[radiacion$Dieta == 1 & radiacion$Radiacion == '8 Gy' ]

print(shapiro.test(grupo6))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo6
## W = 0.95631, p-value = 0.7557
```

```
ggqqplot(grupo6) +
  ggtitle("QQ Plot - Grupo 6 Dieta 1 - 8 Gy") +
  theme(plot.title = element_text(hjust = 0.5))
```



El área aórtica lesionada tiene distrubución normal para todos los grupos de estudio ya que en todos los casos el p-valor del test de normalidad fue mayor a 0.05, aceptándose entonces la hipótesis nula de que la variable tiene distribución normal. También podemos ver en los QQ-Plots que los datos siguen la línea de normalidad. Por lo tanto, se cumple el supuesto de normalidad de la variable respuesta.

Homocedasticidad

Realizo el test de Levene para analizar si las varianzas de los grupos son homogéneas.

```
leveneTest(y = radiacion$Arealog~radiacion$Dieta*radiacion$Radiacion)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  0.8764 0.5164
##       18
```

Con un p-valor mayor a 0.05 se acepta la H_0 que dice que las varianzas de los grupos son homogéneas, por lo que también se cumple este supuesto.

4. **Plantee las hipótesis y efectúe el análisis de la varianza. Concluya, asumiendo un nivel de significación del 5%Plantee las hipótesis y efectúe el análisis de la varianza. Concluya, asumiendo un nivel de significación del 5%.**

Se comprobó que con la transformación logarítmica se cumplen todos los supuestos, por lo tanto puedo realizar el test ANOVA. Las hipótesis son:

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu$

- H_1: algún $\mu_i \neq \mu$,

siendo μ_i la media de el grupo i y μ la media global.

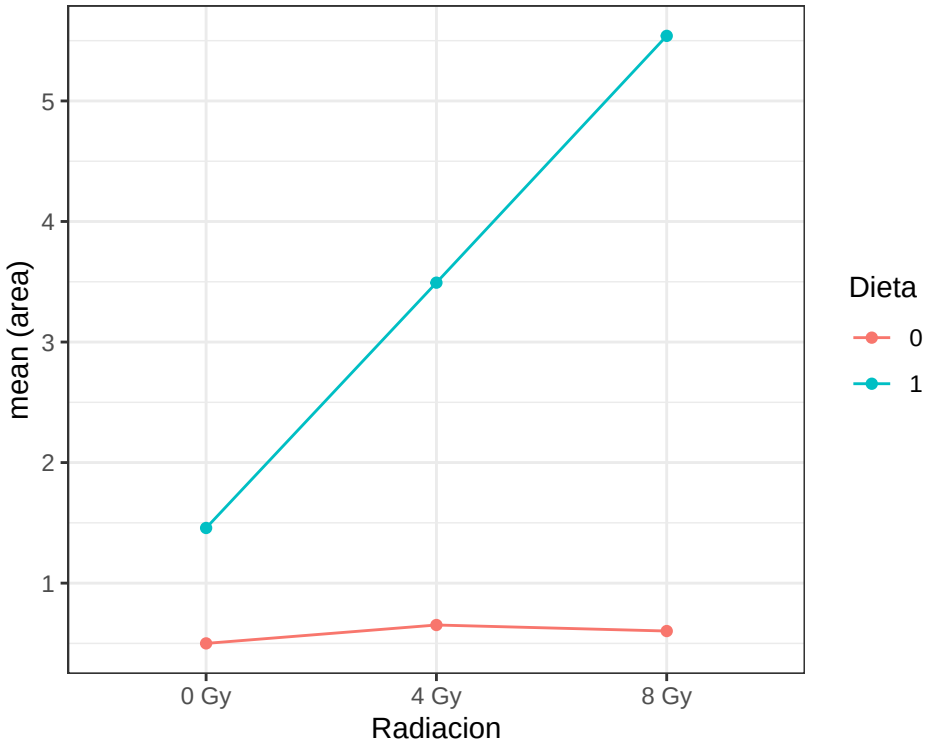
```
anova<-aov(radiacion$Area~radiacion$Radiacion*radiacion$Dieta)
summary(anova)
```

```
##                                      Df Sum Sq Mean Sq F value   Pr(>F)
## radiacion$Radiacion                   2  17.53    8.76  10.111  0.00114 **
## radiacion$Dieta                       1  50.87   50.87  58.689 4.52e-07 ***
## radiacion$Radiacion:radiacion$Dieta   2  15.86    7.93   9.147  0.00182 **
## Residuals                            18  15.60    0.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El test ANOVA dio un p-valor de 0.00182, menor a 0.05, por lo que se rechaza la hipótesis nula: al menos una de las medias es diferente al resto, aunque no sabemos cuál.

5. **Realice un grafico de medias y analice la posibilidad de interacción entre ambos factores.**

```
ggplot(data = radiacion, aes(x = Radiacion, y = Area, colour = Dieta,
group = Dieta)) +
stat_summary(fun = mean, geom = "point") +
stat_summary(fun = mean, geom = "line") +
labs(y = 'mean (area)') +
theme_bw()
```



El gráfico de medias muestra cómo varía la media del área aórtica lesionada de los grupos de ratones que se alimentaron con dieta estándar y de los grupos de ratones que se alimentaron con dieta rica en grasas, a medida que aumenta la dosis de radiación. Para el grupo de dieta estándar (línea rosa) se ve que hay un leve aumento de la media entre 0 Gy y 4 Gy, que luego decae apenas para la dosis de 8 Gy. En cambio, en el grupo de dieta rica en grasas (línea celeste) hay un aumento lineal importante de la media en función de la dosis irradiada. Se puede ver que incluso para 0 Gy de radiación, la media del área aórtica lesionada es mayor a las medias del grupo de dieta estándar, para todas las dosis de radiación.

A partir de este gráfico parece haber una relación lineal entre el área aórtica lesionada y la dosis de radiación cuando se las combina con una dieta rica en grasas, y que la radiación no tiene efectos significativos en el área aórtica lesionada cuando se somete a los ratones a una dieta estándar.

6. **Efectúe las comparaciones entre tratamientos que considere necesarias. ¿Qué conclusiones se pueden sacar al respecto?**

Realizo el test de Tukey para hacer la comparación de medias entre todas las combinaciones de grupos de tratamiento.

```
Tukey=TukeyHSD(x=anova, conf.level=0.95)
print(Tukey)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = radiacion$Area ~ radiacion$Radiacion * radiacion$Dieta)
##
## $`radiacion$Radiacion`
##              diff         lwr      upr     p adj
## 4 Gy-0 Gy 1.09375 -0.09425156 2.281752 0.0741775
## 8 Gy-0 Gy 2.09250  0.90449844 3.280502 0.0007789
## 8 Gy-4 Gy 0.99875 -0.18925156 2.186752 0.1086036
##
## $`radiacion$Dieta`
##         diff     lwr      upr p adj
## 1-0 2.911667 2.11317 3.710163 5e-07
##
## $`radiacion$Radiacion:radiacion$Dieta`
##                   diff         lwr      upr     p adj
## 4 Gy:0-0 Gy:0  0.1525 -1.93959904 2.244599 0.9998908
## 8 Gy:0-0 Gy:0  0.1025 -1.98959904 2.194599 0.9999847
## 0 Gy:1-0 Gy:0  0.9575 -1.13459904 3.049599 0.6952994
## 4 Gy:1-0 Gy:0  2.9925  0.90040096 5.084599 0.0029052
## 8 Gy:1-0 Gy:0  5.0400  2.94790096 7.132099 0.0000060
## 8 Gy:0-4 Gy:0 -0.0500 -2.14209904 2.042099 0.9999996
## 0 Gy:1-4 Gy:0  0.8050 -1.28709904 2.897099 0.8203025
## 4 Gy:1-4 Gy:0  2.8400  0.74790096 4.932099 0.0047478
## 8 Gy:1-4 Gy:0  4.8875  2.79540096 6.979599 0.0000091
## 0 Gy:1-8 Gy:0  0.8550 -1.23709904 2.947099 0.7819428
## 4 Gy:1-8 Gy:0  2.8900  0.79790096 4.982099 0.0040416
## 8 Gy:1-8 Gy:0  4.9375  2.84540096 7.029599 0.0000079
## 4 Gy:1-0 Gy:1  2.0350 -0.05709904 4.127099 0.0592850
## 8 Gy:1-0 Gy:1  4.0825  1.99040096 6.174599 0.0000947
## 8 Gy:1-4 Gy:1  2.0475 -0.04459904 4.139599 0.0571258
```

A partir del p-valor de cada uno de los pares de medias analizados, genero una tabla que resume los resultados del test de Tukey. Para cada combinación de tratamientos indico si la diferencia entre medias es estadísticamente significativa (p-valor menor a 0.05) con el texto 'SIG' o si no es estadísticamente significativa (p-valor mayor a 0.05) con 'no sig'.

```
filas = c('DietaEst_0Gy','DietaEst_4Gy','DietaEst_8Gy','DietaGra_0Gy','DietaGra_4Gy',
          'DietaGra_8Gy')
tabla <- data.frame(DietaEst_0Gy = c('-', 'no sig.','no sig','no sig','SIG', 'SIG'),
                    DietaEst_4Gy = c('no sig', '-','no sig','no sig','SIG', 'SIG'),
                    DietaEst_8Gy = c('no sig', 'no sig','-','no sig','SIG', 'SIG'),
                    DietaGra_0Gy = c('no sig', 'no sig','no sig','-','no sig', 'SIG'),
                    DietaGra_4Gy = c('SIG', 'SIG','SIG','no sig','-', 'no sig'),
                    DietaGra_8Gy = c('SIG', 'SIG','SIG','SIG','no sig', '-'),
                    row.names = filas)
kable(tabla, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	DietaEst_0Gy	DietaEst_4Gy	DietaEst_8Gy	DietaGra_0Gy	DietaGra_4Gy	DietaGra_8Gy
DietaEst_0Gy	-	no sig	no sig	no sig	SIG	SIG
DietaEst_4Gy	no sig.	-	no sig	no sig	SIG	SIG
DietaEst_8Gy	no sig	no sig	-	no sig	SIG	SIG
DietaGra_0Gy	no sig	no sig	no sig	-	no sig	SIG
DietaGra_4Gy	SIG	SIG	SIG	no sig	-	no sig
DietaGra_8Gy	SIG	SIG	SIG	SIG	no sig	-

Por lo tanto, se puede ver que entre todos los grupos con dieta estándar no hay diferencia significativa entre las medias y que, a su vez, tampoco existe una diferencia significativa entre estos grupos y el grupo de ratones que no recibió radiación y que mantuvo una dieta rica en grasas. Sin embargo, en este último caso la diferencia es más significativa que la de los grupos de dieta estándar entre sí.

Por otro lado, el grupo que tuvo mayores diferencias significativas entre las medias de los demás fue el de dieta rica en grasas y radiación de 8 Gy. Esto coincide lo que que se había observado anteriormente sobre el aumento lineal del área aórtica lesionada con la dosis de radiación en los grupos de dieta rica en grasas.

A continuación genero un boxplot con las etiquetas generadas según los resultados del test de Tukey.

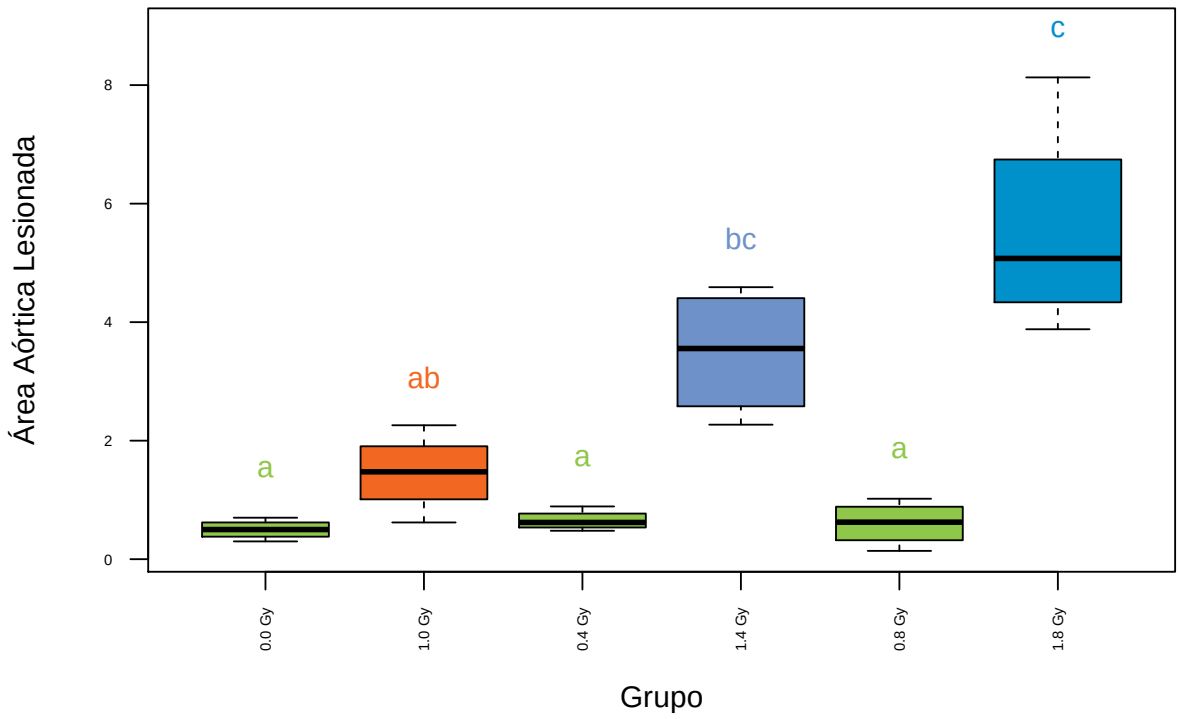
```
LABELS <- generate_label_df(Tukey , "radiacion$Radiacion:radiacion$Dieta")

my_colors <- c(
rgb(143,199,74,maxColorValue = 255),
rgb(242,104,34,maxColorValue = 255),
rgb(111,145,202,maxColorValue = 255),
rgb(0,145,202,maxColorValue = 255),
rgb(111,15,22,maxColorValue = 255),
rgb(11,165,202,maxColorValue = 255)
)

LABELS$Numbers = c(1,2,1,3,1,4)

a <- boxplot(radiacion$Area ~ radiacion$Dieta:radiacion$Radiacion ,
             ylim=c(min(radiacion$Area), 1.1*max(radiacion$Area)),
cex.axis=0.5, las=2,
```

```
col=my_colors[(LABELS[,3])]
, ylab="Área Aórtica Lesionada", xlab="Grupo")
over <- 0.1*max( a$stats[nrow(a$stats),] )
text( c(1:nlevels(radiacion$Radiacion:radiacion$Dieta)) , a$stats[nrow(a$stats),]+over ,
      LABELS[,1],
col=my_colors[(LABELS[,3])] )
```



Los tres grupos de dieta estándar no mostraron diferencias significativas entre sus medias, por lo que son etiquetados con la misma etiqueta. Esto indica que no hay una interacción entre la radiación y el área aórtica lesionada si se mantiene una dieta diaria estándar.

Por otro lado, los grupos con dieta rica en grasas poseen etiquetas únicas. El grupo de radiación 0 Gy no mostró una diferencia estadísticamente significativa con los grupos de dieta estándar y con el grupo de dieta rica en grasas y radiación de 4 Gy, sin embargo existe una diferencia suficientemente grande y por esta razón tiene la etiqueta 'ab'. Esto mismo sucede entre el grupo de dieta rica en grasas y radiación de 4 Gy y el de dieta rica en grasas y radiación de 8 Gy.

Finalmente, se puede concluir que existe una interacción entre la dosis de radiación y el área aórtica lesionada cuando se mantiene una dieta diaria rica en grasas, aumentado dicha área a medida que se aumenta la dosis.