Clase 12 Evaluación de supuestos pruebas paramétricas

Diplomado en Análisis de Datos y Modelamiento Predictivo con Aprendizaje Automático para la Acuicultura.

Dra. María Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

30 May 2023

PLAN DE LA CLASE

1.- Introducción

- Supuestos de los análisis paramétricos.
- Consecuencias de la violación de los supuestos.
- Métodos gráficos y análisis de residuos para evaluar supuestos.
- Pruebas de hipótesis para evaluar supuestos.

2.- Práctica con R y Rstudio cloud

- Evaluar supuestos de las pruebas paramétricas.
- Elaborar un reporte dinámico en formato html.

SUPUESTOS: INDEPENDENCIA

Independencia

Cada observación de la muestra no debe estar relacionada con otra observación de la muestra.

Si se viola este supuesto la prueba paramétrica ***NO*** *es válida.*

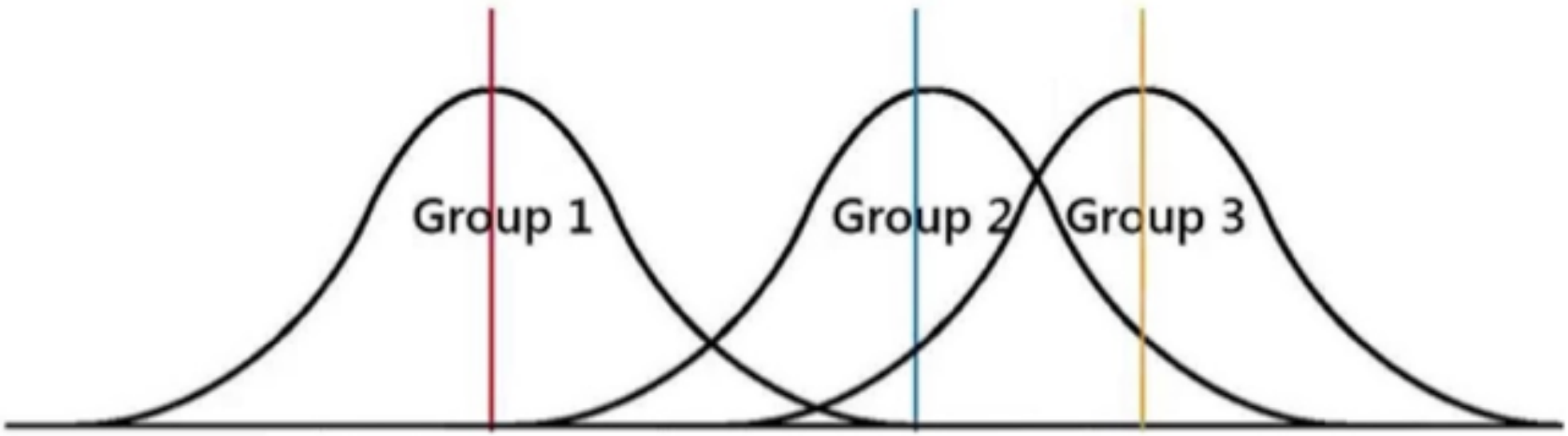
Ejemplo violación del supuesto

- Medidas repetidas en un mismo individuo (antes y después de un tratamiento).
- Observaciones están correlacionadas en el tiempo.
- Observaciones están correlacionadas en el espacio.

SUPUESTOS: HOMOGENEIDAD DE VARIANZAS

Homocedasticidad

En el caso de comparación de dos o más muestras éstas deben provenir de poblaciones con la misma varianza.

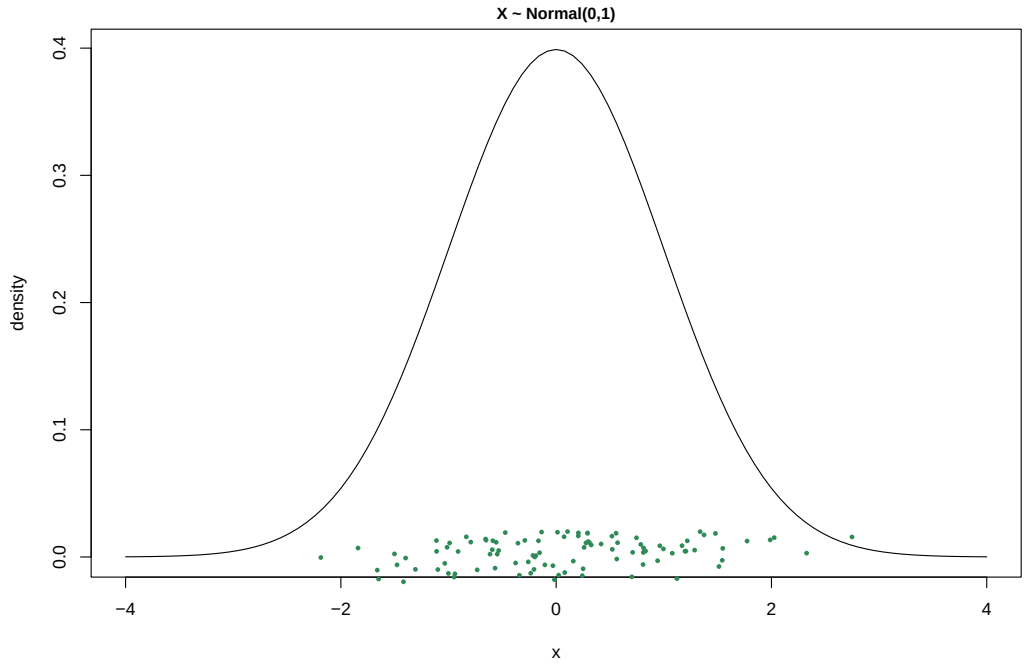


Alguna heterogeneidad es permitida, particularmente con $\boldsymbol{n > 30}$.

SUPUESTOS: NORMALIDAD

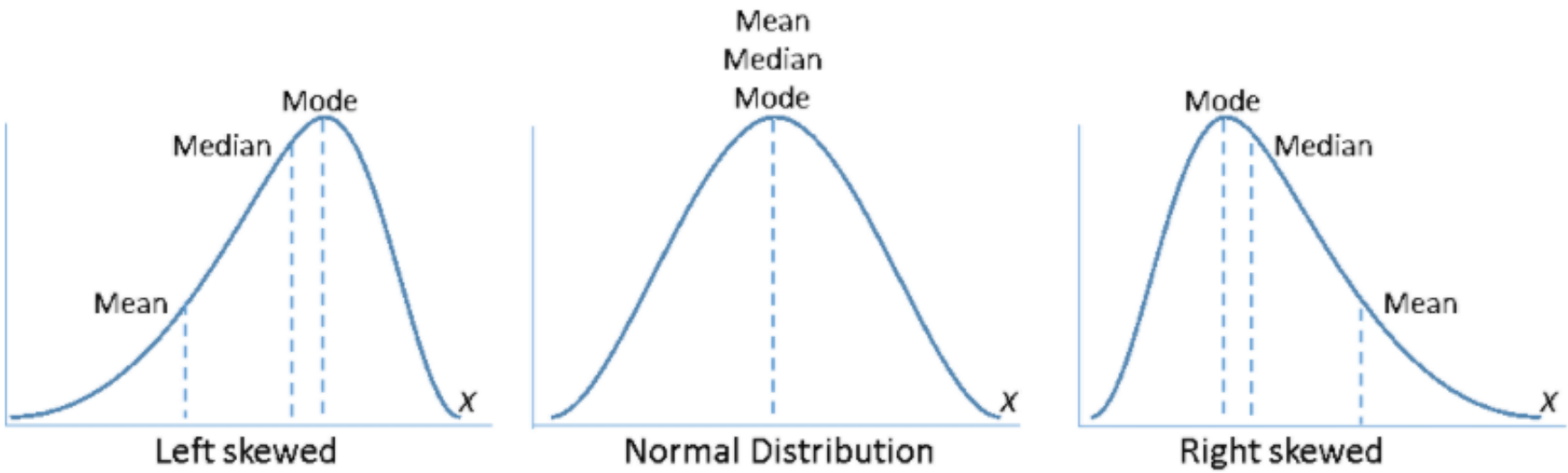
Normalidad

Los datos de muestreo se obtienen de una población que tiene distribución normal.



VIOLACIÓN DEL SUPUESTO DE NORMALIDAD

¿Cuál es el problema?
Cambia la probabilidad de rechazar la hipótesis nula.



En la práctica apróximadamente normal es suficiente, particularmente con $\boldsymbol{n > 30}$.

¿QUÉ SON LOS RESIDUALES?

Los residuales de un modelo se refieren a la diferencia entre los valores observados y los valores predichos por ese modelo. El valor predicho en un ANOVA se refiere a la media de cada nivel del efecto "Tratamiento".

	Peso	Tratamiento	Residuos	Predichos
1	4.2	Control	-12.76	16.96
2	11.5	Control	-5.46	16.96
3	7.3	Control	-9.66	16.96
4	5.8	Control	-11.16	16.96
5	6.4	Control	-10.56	16.96
31	15.2	Dieta 1	-5.46	20.66
32	21.5	Dieta 1	0.84	20.66
33	17.6	Dieta 1	-3.06	20.66
34	9.7	Dieta 1	-10.96	20.66
35	14.5	Dieta 1	-6.16	20.66

MÉTODO DE LOS RESIDUALES (GRÁFICOS)

Residuo = valor observado - valor predicho

$e = y$ - $\hat{y}$

Residuos en ANOVA

$$\sum_{i=1}^{n}(y - \hat{y})^2$$

Note que la suma de residuos representa la variabilidad no explicada por el modelo.

MÉTODOS PARA EVALUACIÓN DE SUPUESTOS 2

MÉTODO MEDIANTE PRUEBAS ESTADÍSTICAS

- INDEPENDENCIA: DURBIN-WATSON.
- HOMOGENEIDAD DE VARIANZAS: PRUEBA DE LEVENE.
- NORMALIDAD: PRUEBA DE SHAPIRO-WILKS.

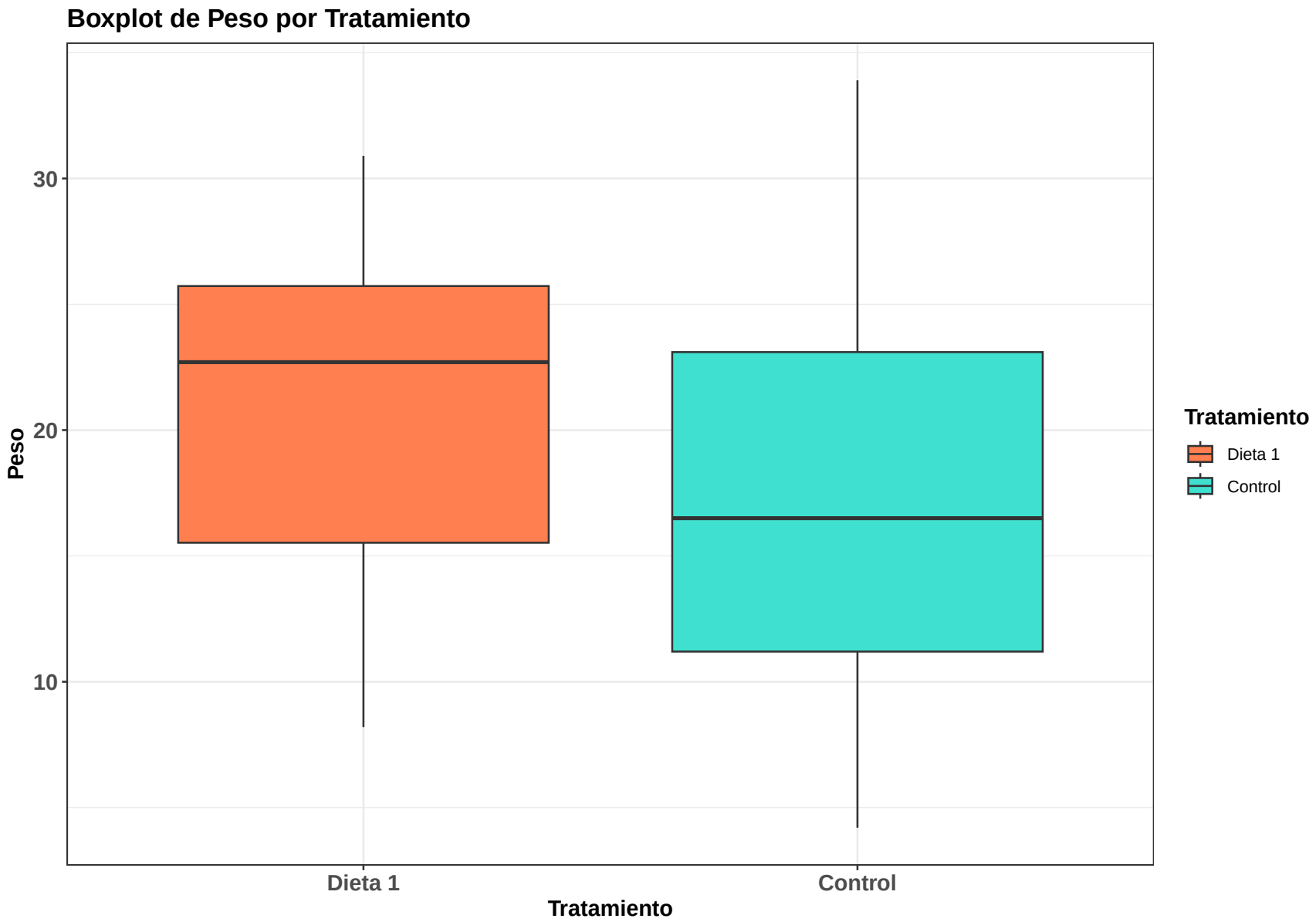
Regla de oro

1.- Primero evalúe independencia.

2.- Luego, homogeneidad de varianzas.

3.- Finalmente, normalidad.

ESTUDIO DE CASO



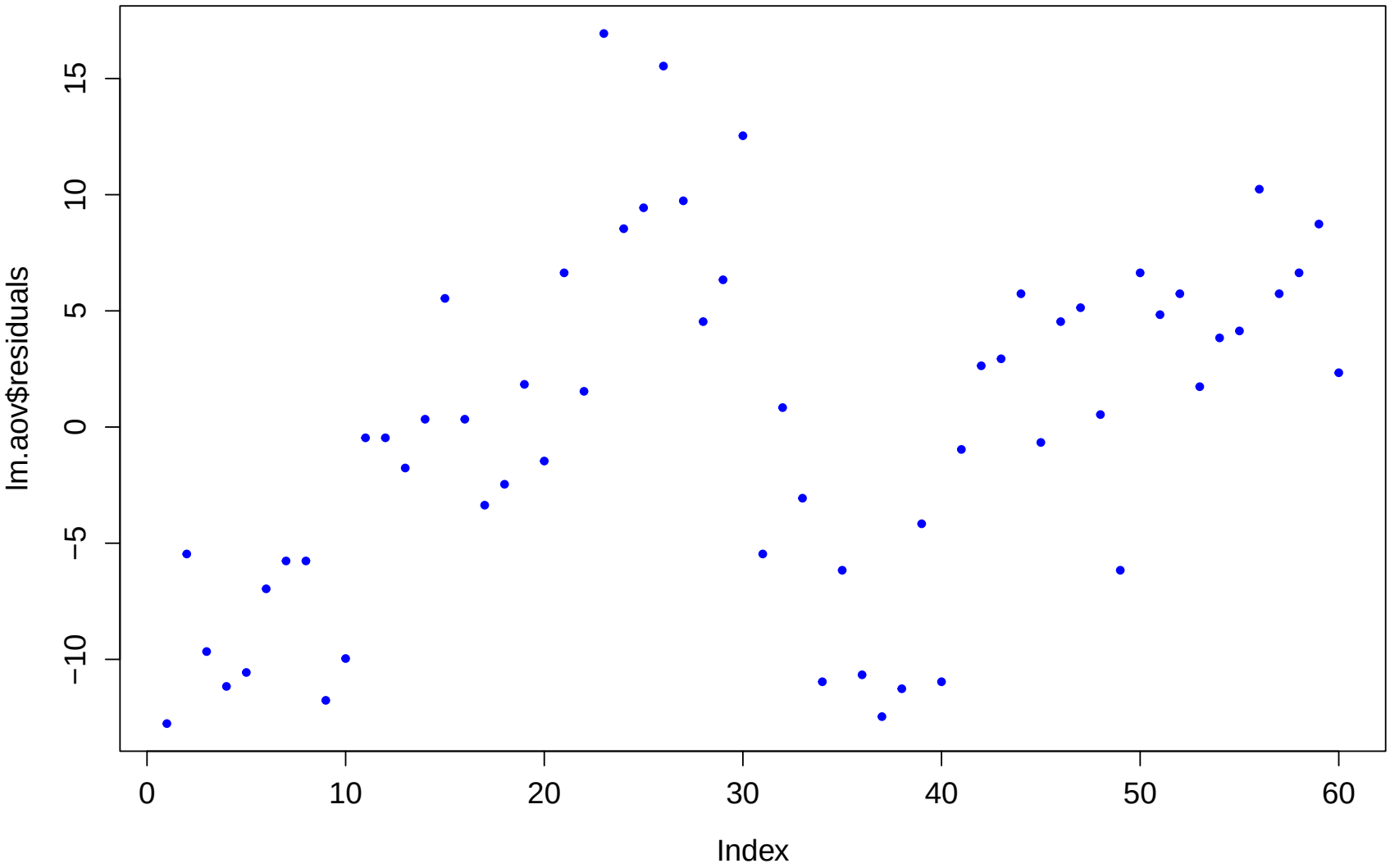
ANOVA

```
lm.aov <- lm(Peso ~ Tratamiento, data = my_data)
anova(lm.aov) %>% kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tratamiento	1	205.350	205.35	3.668	0.06
Residuals	58	3246.859	55.98	NA	NA

INDEPENDENCIA: ANÁLISIS DE RESIDUALES

```
plot(lm.aov$residuals, pch=20, col = "blue",
     cex.lab=1.25, cex.axis=1.25)
```



INDEPENDENCIA: PRUEBA DE DURBIN-WATSON

Hipótesis

$\mathbf{H_0}$: Son independientes o no existe autocorrelación.

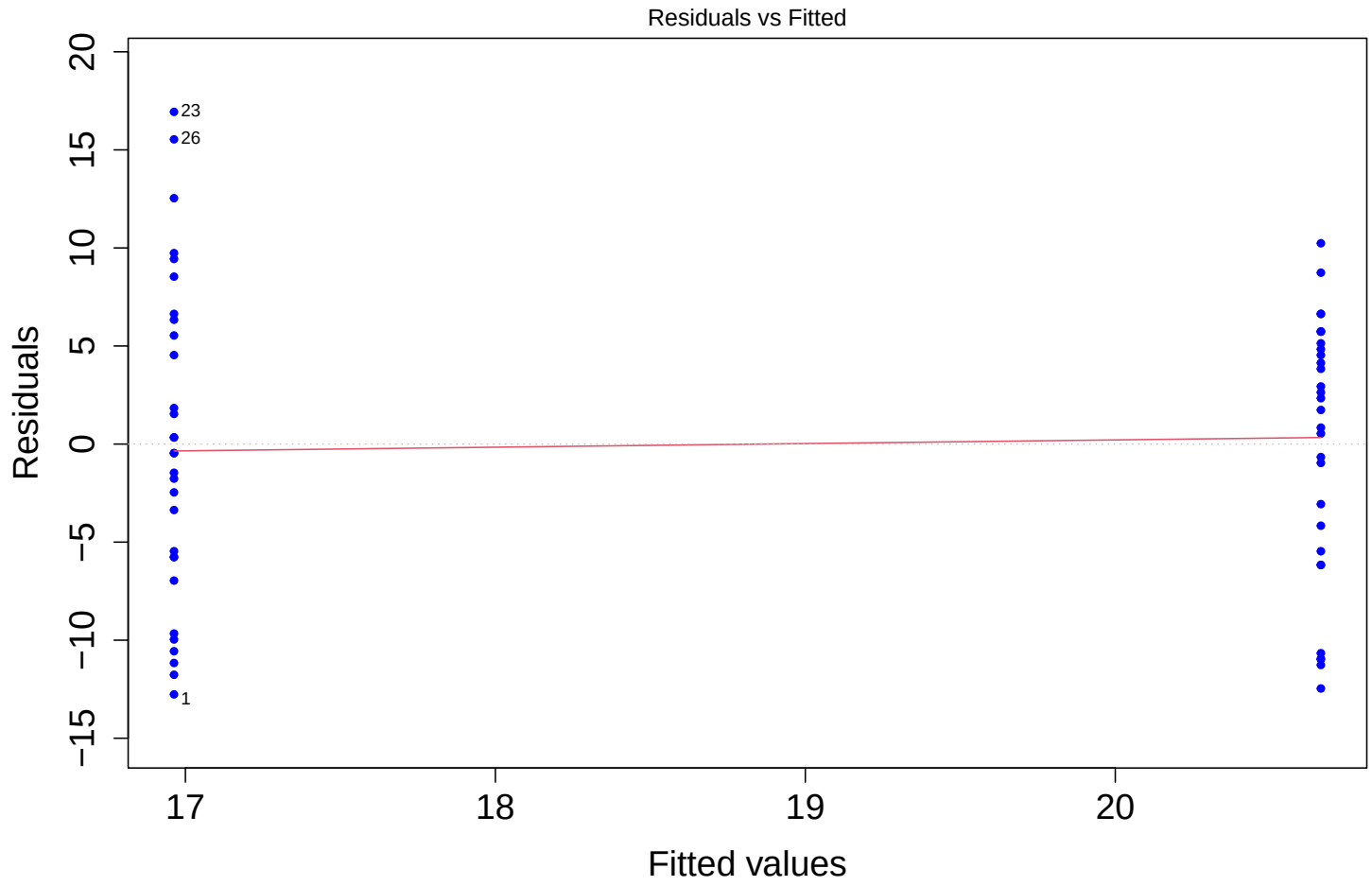
$\mathbf{H_A}$: No son independientes y existe autocorrelación.

```
dwtest(Peso ~ Tratamiento, data = my_data,
       alternative = c("two.sided"),
       iterations = 15) # library(lmtest)
```

```
##
##  Durbin-Watson test
##
## data:  Peso ~ Tratamiento
## DW = 0.61428, p-value = 1.166e-10
## alternative hypothesis: true autocorrelation is not 0
```

HOMOGENEIDAD DE VARIANZAS: ANÁLISIS DE RESIDUALES

```
plot(lm.aov, 1, pch=20, col = "blue",
     cex.lab=1.5, cex.axis=1.5, sub = "")
```



HOMOGENEIDAD DE VARIANZAS: PRUEBA DE LEVENE

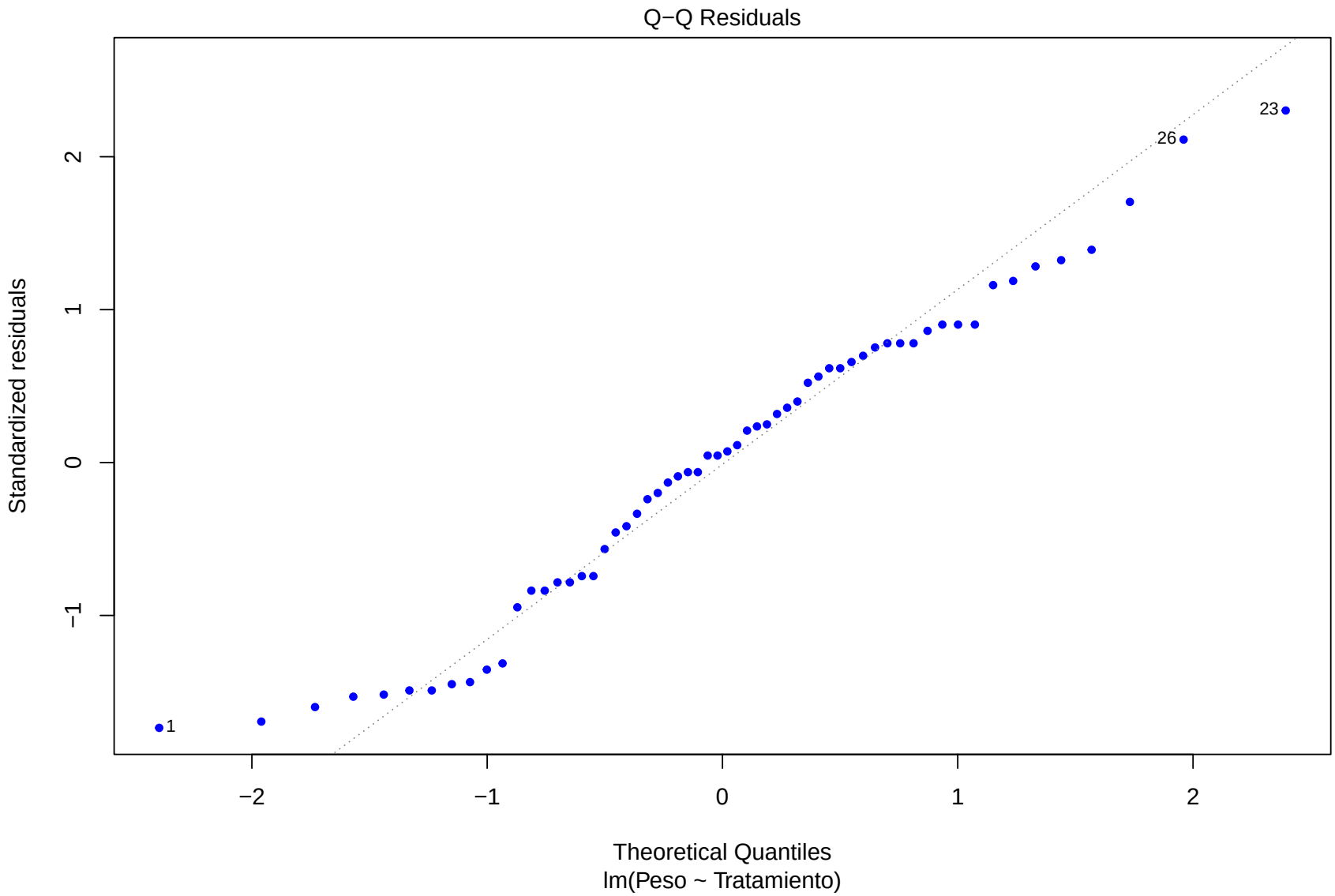
H_0: $\sigma_1^2 = \sigma_2^2$
H_A: $\sigma_1^2 \neq \sigma_2^2$

```
lv <- leveneTest(Peso ~ Tratamiento, data = my_data,
                 center = "median") # library(car)
lv %>% kable(digits = 3)
```

	Df	F value	Pr(>F)
group	1	1.214	0.275
	58	NA	NA

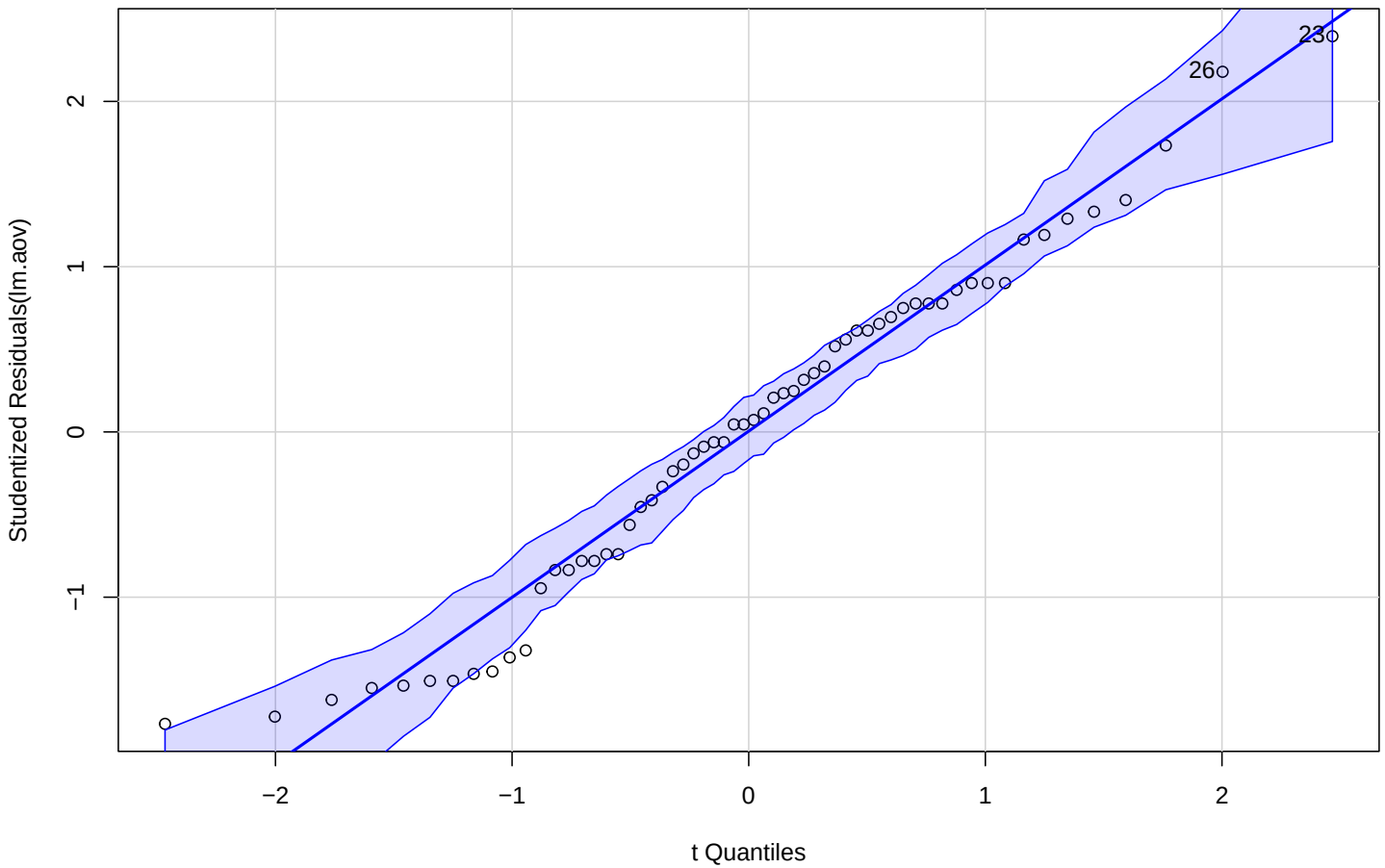
NORMALIDAD: ANÁLISIS DE RESIDUALES

```
plot(lm.aov, 2, pch=20, col = "blue")
```



NORMALIDAD: ANÁLISIS DE RESIDUALES 2

```
qqPlot(lm.aov) # library(car)
```




```
## [1] 23 26
```

NORMALIDAD: PRUEBA DE SHAPIRO-WILKS

$\mathbf{H_0}$: La distribución es normal.
$\mathbf{H_A}$: La distribución no es normal.

```
aov_residuals <- residuals(object = lm.aov)
shapiro.test(x= aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.96949, p-value = 0.1378
```

PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.
  Clase 12 - Evaluación de supuestos

RESUMEN DE LA CLASE

- **Teoría**
- Supuestos de los análisis paramétricos.
- Consecuencias de la violación de los supuestos.
- Interpretación de métodos gráficos, análisis de residuos y pruebas de hipótesis para evaluar supuestos.
- **Evaluación de supuestos**
  - Independencia.
  - Homocedasticidad.
  - Normalidad.