Instituto
de Ciencias
Antonio Brack

Programa de
Certificación Especializado

Data Science: Estadística y Análisis de Datos en R

Instructor: Blgo. Irwing S. Saldaña

Análisis Estadístico Básico con R

Semana 5

Pruebas estadísticas I

Instituto de Ciencias Antonio Brack

Aprenderás esta semana

- Test Paramétricos y No Paramétricos
- Test de normalidad
  - Simetría y Curtosis
  - KS Test
  - SW Test
  - AD Test
  - Q-Q Plot
- Comparación de grupos:
  - T-Student (una muestra, muestras independientes, varianzas iguales, varianzas desiguales)
  - Test de Wilcoxon de una muestra
  - Test de Rangos con Signo de Wilcoxon
  - Test U de MannWhitney
- Regresión lineal simple: Verificación de supuestos teóricos
- Outliers en regresiones lineales

Paramétrico vs No Paramétrico

Mis Datos: todo conjunto de datos tiene una distribución de probabilidades única, pero esta se asemeja a algúna distribución teórica. Nos preguntamos…

¿Siguen una distribución de probabilidades Normal?

sí

Métodos Paramétricos

no

¿Deseo o debo transformarlos (normalizarlos)?

sí

no

Métodos No Paramétricos

1er PC

(1) Test de normalidad
(2) Naturaleza de los datos

2do PC

(1) Transformación de datos
(2) Test de normalidad

Métodos estadísticos

Paramétricos	VS	No Paramétricos
• **Asumen normalidad** de los datos. • Trabajan sobre la comparación de **promedios** entre grupos. • Suelen tener **más poder estadístico**. Es más probable que se detecten diferencias significativas cuando realmente existen.		• **Es libre de asunciones** sobre la distribución de probabilidades de los datos. • Trabajan sobre la comparación de **medianas** entre grupos. • Suelen tener **poder estadístico limitado**, principalmente cuando las diferencias entre los grupos comparados son muy poco marcadas.

Pruebas de Normalidad

Abre el archivo **“R-Notebook-C2-S2.R”** y
trabajaremos en la sección **1. Pruebas de Normalidad**

Buscando la Normalidad

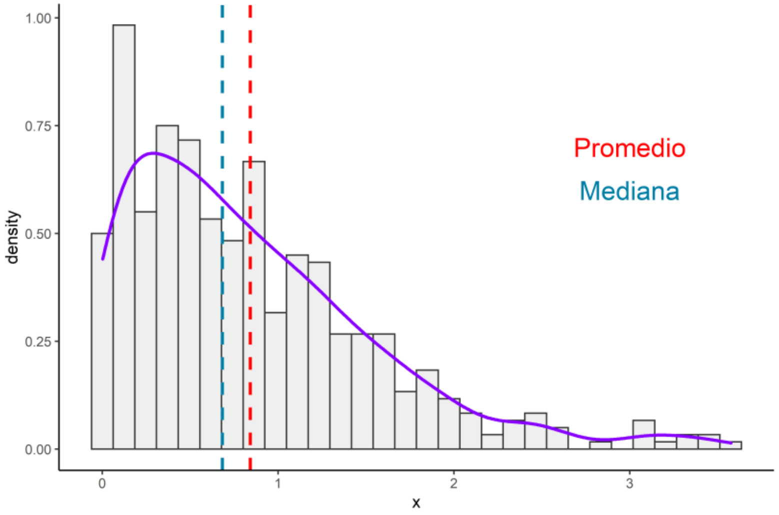
- 1. Simetría y Curtosis
- 2. Test de Normalidad
  - Kolmogorov-Smirnov
  - Shapiro Wilk
  - Anderson-Darling
- 3. Q-Q Plots (Gráficos Cuantil-Cuantil)

Alerta de aplicación

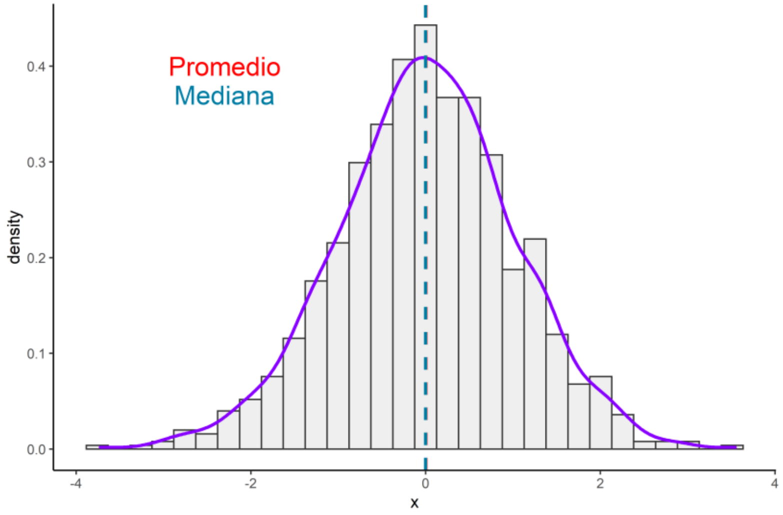
- Es poco útil comprobar la normalidad de un conjunto de datos (excepto test de comparación de uno o dos grupos).
- Por el contrario, la normalidad debe testearse de los residuales de los modelos.

1. Simetría y Curtosis

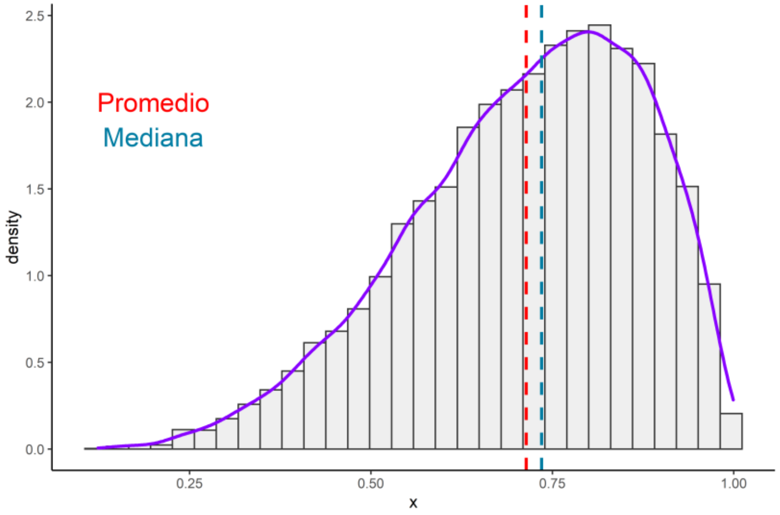
- **Simetría:** mide que tanto los datos se acumulan hacia uno u otro lado (s perfecto = 0).



Asimetría positiva:
S ≥ 0.5



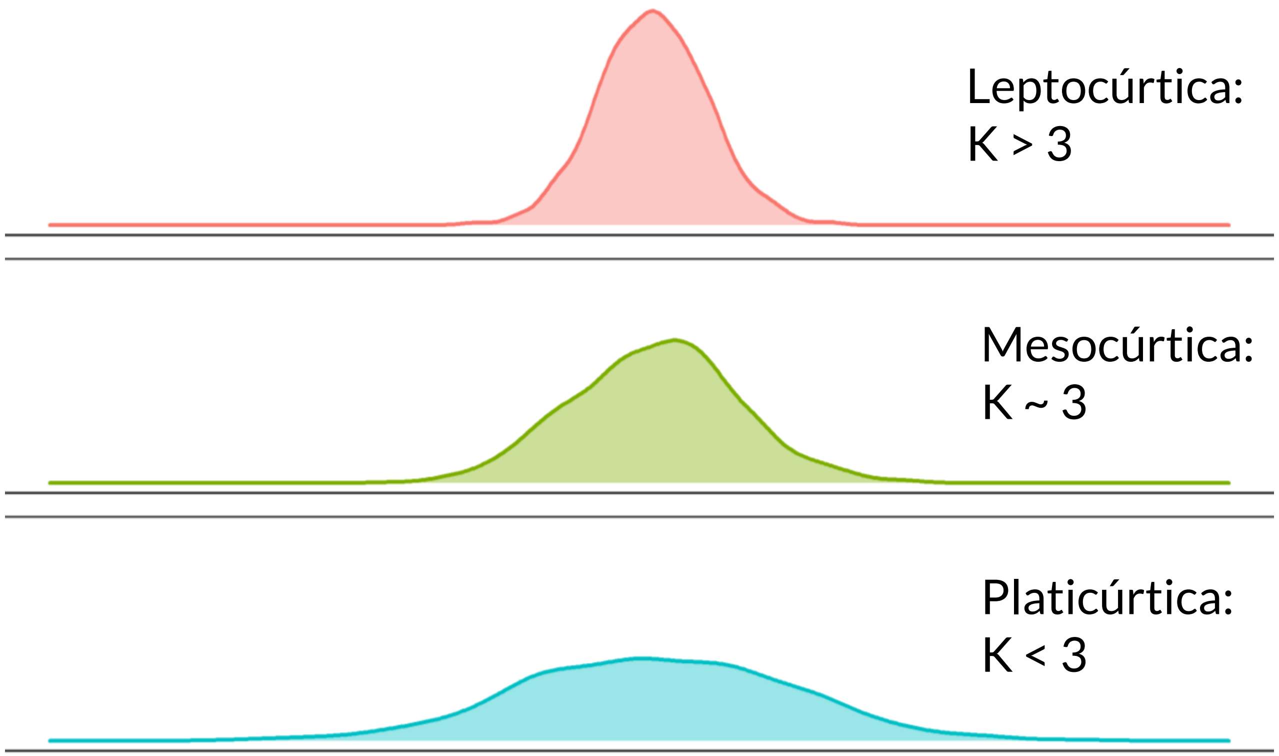
Simetría (normal) :
0.5 > S > -0.5



Asimetría negativa :
S ≤ -0.5

1. Simetría y Curtosis

- **Curtosis:** mide la frecuencia de los datos en el centro y colas de la curva.
    - Mientras concentrados en el centro estén, la curva se verá más elevada.
    - Mientras más uniformemente dispersos esté, más plana será la curva.



Funciones de Simetría y Curtosis

```
# Funciones del paquete e1071
library(moments)

# Cálculo de la simetría
skewness(vector)

# Cálculo de la curtosis
kurtosis(vector)
```

2. Test de Normalidad

Comprueban si la distribución empírica (eCDF) de un conjunto de datos encaja dentro de la distribución teórica Normal.

3. Normality Tests

Go to:

The normality tests are supplementary to the graphical assessment of normality (8). The main tests for the assessment of normality are Kolmogorov-Smirnov (K-S) test (7), Lilliefors corrected K-S test (7, 10), Shapiro-Wilk test (7, 10), Anderson-Darling test (7), Cramer-von Mises test (7), D'Agostino skewness test (7), Anscombe-Glynn kurtosis test (7), D'Agostino-Pearson omnibus test (7), and the Jarque-Bera test (7).

[Normality Tests for Statistical Analysis: A Guide for Non-Statisticians (nih.gov)]
[Full article: Comparisons of various types of normality tests (tandfonline.com)]

Kolmogorov-Smirnov
Shapiro Wilk
Anderson-Darling

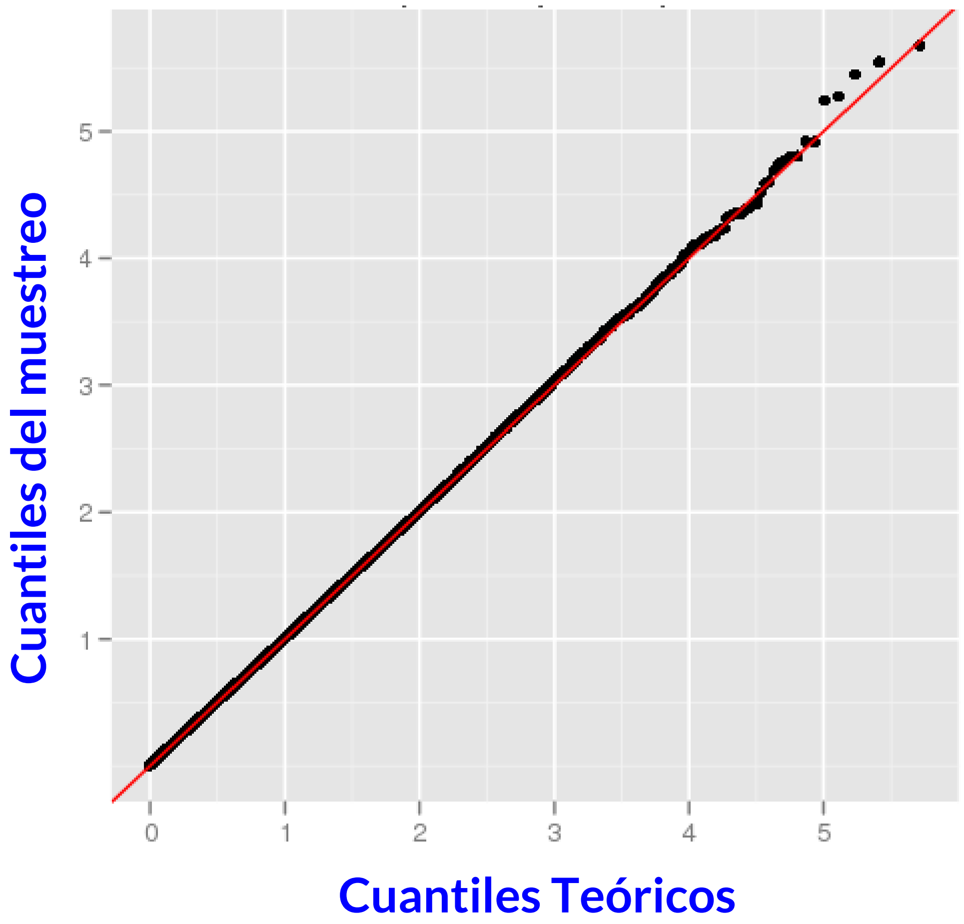
Funciones para el cálculo de S y K

```
# Test de Shapiro-Wilk
shapiro.test(vector)

# Test de Kolmogorov-Smirnov
ks.test(vector, "pnorm", mean(vector), sd(vector))

# Test de Anderson-Darling
nortest::ad.test(vector)
```

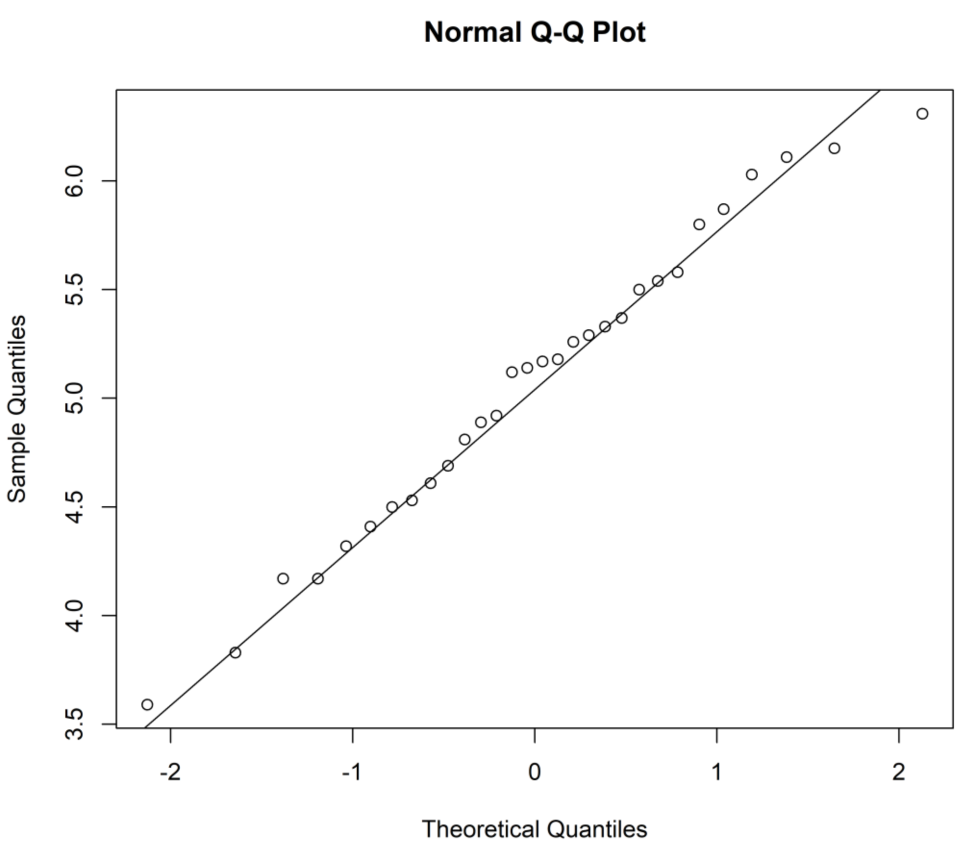
- Si tienes n muestral menor a 30 UM, utiliza SW.
- KS se ha dejado de usar por su poca confiabilidad en la mayoría de situaciones.
- Para n muestral mayor a 30 UM, es más confiable AD.

3. Q-Q plot

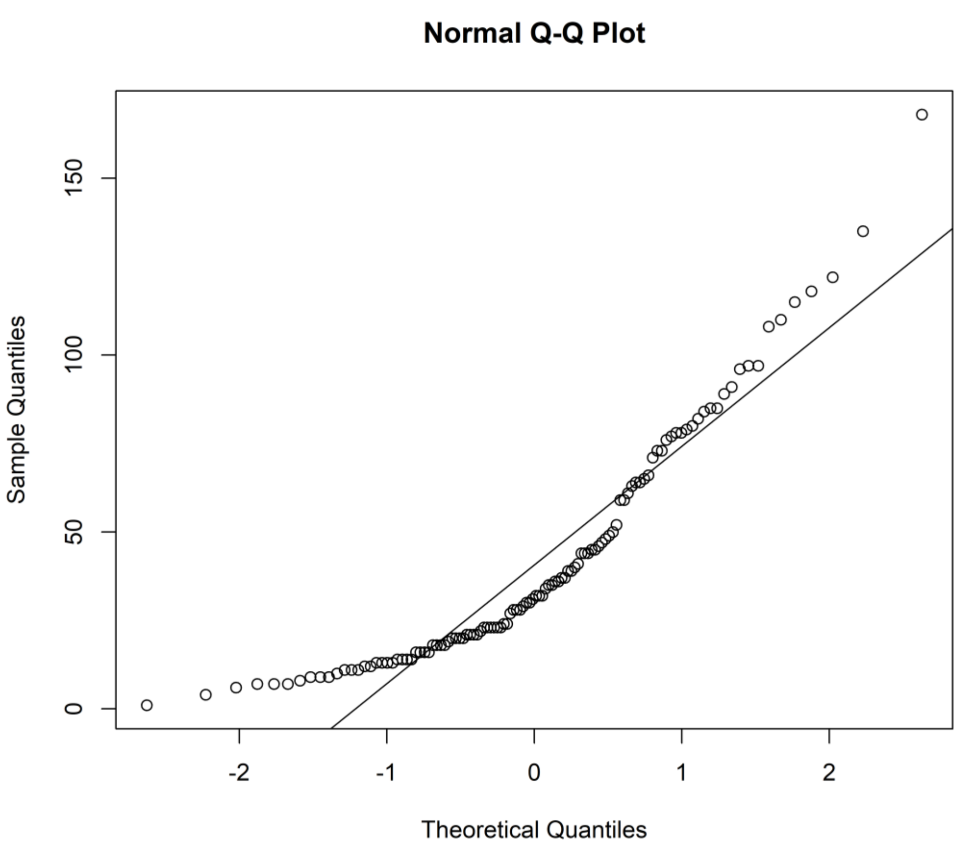
- Este gráfico enfrenta los cuantiles del muestreo vs los cuantiles teóricos de la distribución normal.
- Si ambos grupos de cuantiles son idénticos, se crea un **patrón de dispersión de puntos lineal**.
- Estos caen sobre una línea imaginaria perfecta con pendiente 1 y origen 0 (Línea Q-Q).
- Si los puntos no caen sobre la Línea Q-Q, entonces no hay normalidad.



Contraste de Normalidad y No Normalidad



Hay normalidad



NO hay normalidad

Funciones para calcular los cuantiles teóricos

```
# Calculan los cuantiles referentes a un conjunto de datos, usando como medidas sus estadísticos (promedio, desviación estándar por ejemplo para la distribución Normal

qnorm()

qpois()

qbinom()

qbeta()

qt()

qgamma()

qchisq()

qexp()
```

Construcción de un Q-Q Plot

```
# Calculo manual del qqplot a partir de "datos"
prob <- ppoints(datos)
q.teoricos <- qnorm(prob, mean(datos), sd(datos))

plot(q.teoricos, sort(datos))
abline(a=0, b=1)
```

```
# Usando las funciones básicas de R
qqnorm(datos)
qqline(datos)
```

Normalizar datasets

Abre el archivo **“R-Notebook-C2-S2.R”** y
trabajaremos en la sección **2. Normalización de datasets**

Normalizar conjuntos de datos

```r
# Utiliza la función bestNormalize, de la librería del mismo nombre
library(bestNormalize)
bestNormalize(BD)
```

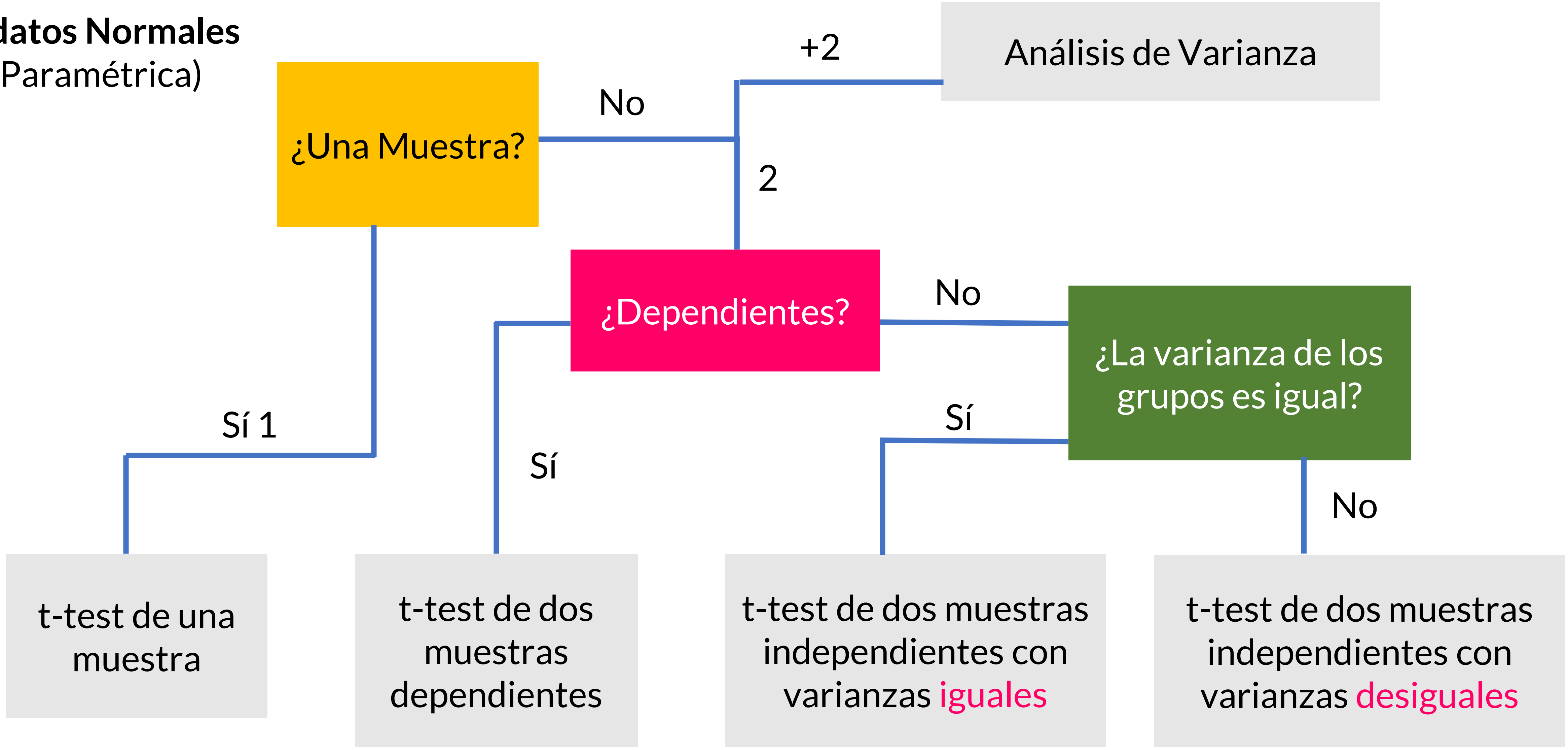
- Para normalizar los datos existen una serie de transformaciones que se les pueden aplicar.
- Desde un simple `sqrt()` hasta algoritmos complejos como Box-Cox o Yeo-Johnson.
- La función `bestNormalize()` seleccionará la mejor transformación para tus datos.

```
> bestNormalize(JN)
Best Normalizing transformation with 30 Observations
 Estimated Normality Statistics (Pearson P / df, lower => more normal):
 - arcsinh(x): 1.8
 - Box-Cox: 1.64
 - Center+scale: 1.6933
 - Exp(x): 2.8133
 - Log_b(x+a): 1.8
 - orderNorm (ORQ): 1.7467
 - sqrt(x + a): 1.6933
 - Yeo-Johnson: 1.64
Estimation method: Out-of-sample via CV with 10 folds and 5 repeats

Based off these, bestNormalize chose:
Standardized Box Cox Transformation with 30 nonmissing obs.:
 Estimated statistics:
 - lambda = 1.647643
 - mean (before standardization) = 93.38514
 - sd (before standardization) = 44.11489
```

Pruebas estadísticas para
Comparación entre grupos (Parte 1)

Abre el archivo **“R-Notebook-C2-S2.R”** y
trabajaremos en la sección **3. Comparación de grupos (parte 1)**

Para datos Normales
(Ruta Paramétrica)



Funciones en R: 1 y 2 muestras

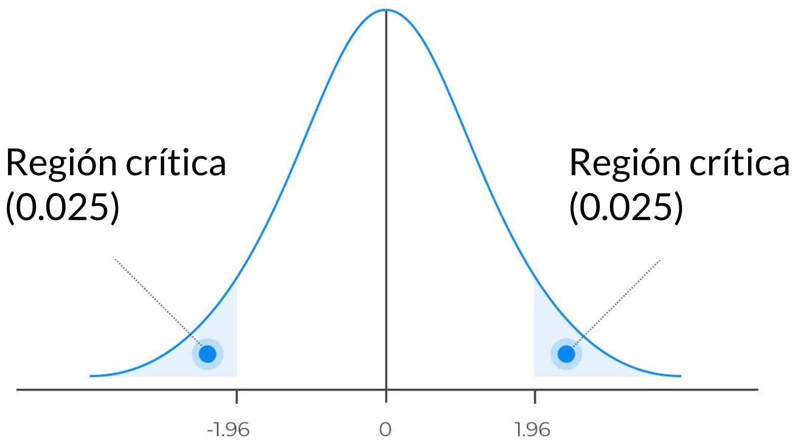
```r
# t-test de una muestra
t.test(x, mu = 0)

# t-test de dos muestras dependientes
t.test(x, y, paired = TRUE)

# t-test de dos muestras independientes con varianzas
iguales (pooled)
t.test(x, y, var.equal = TRUE)

# t-test de dos muestras independientes con varianzas
desiguales
t.test(x, y, var.equal = FALSE)
```

Argumento `alternative`

- Este argumento nos permite especificar si queremos hacer el test de T de dos colas ("two.sided") o de una cola ("less", "greater").
- Todas las formas de la función `t.test()` que revisamos en la lámina anterior pueden contener este argumento.
- Por defecto se encuentra activado "two.sided".

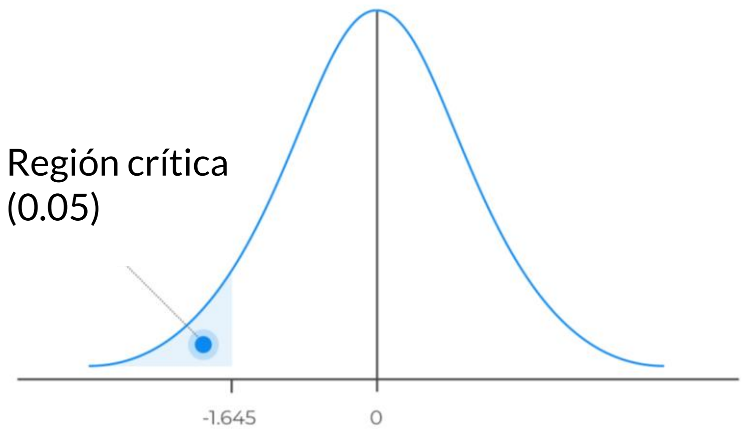
```r
# t-test de una muestra
t.test(x, mu = 0, alternative = "two.sided")

# t-test de dos muestras (todas las de la lámina anterior)
t.test(x, y, … , alternative = "two.sided")
```

Argumento `alternative` … Nos preguntamos



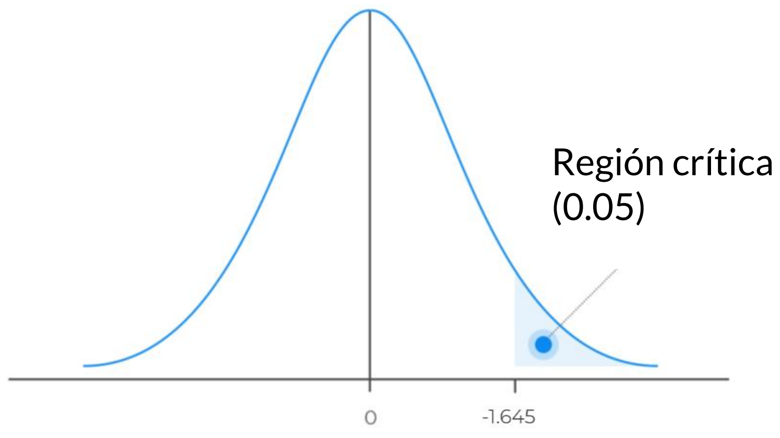
two.sided

¿Existen diferencias significativas entre el promedio del conjunto de dato x e y?



greater

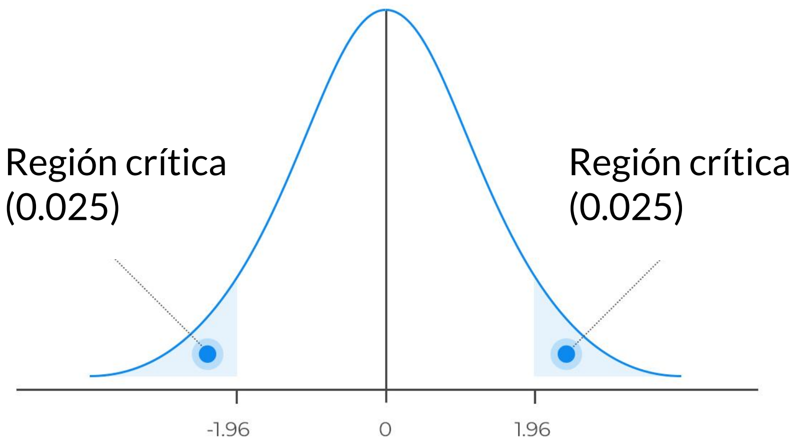
¿El promedio del conjunto de datos x es significativamente mayor que el de y?



less

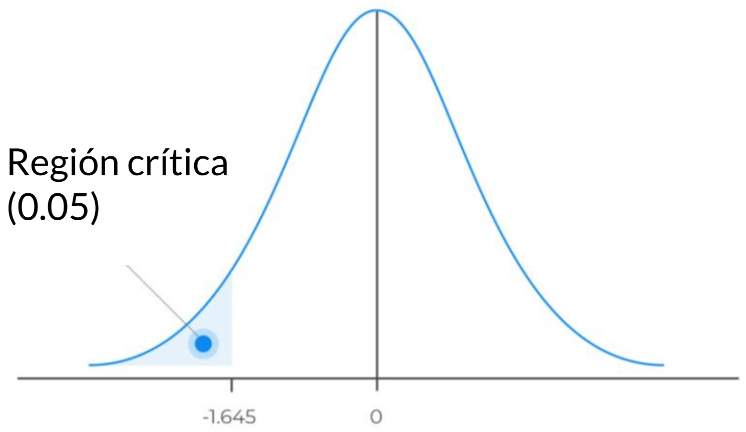
¿El promedio del conjunto de datos x es significativamente menor que el de y?

Argumento `alternative`



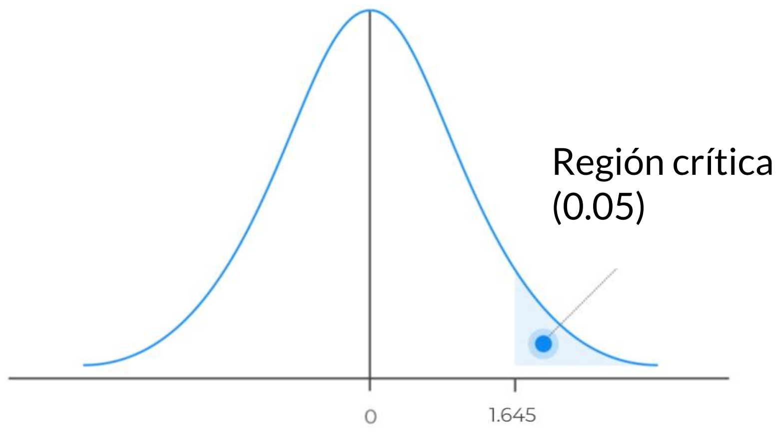
two.sided

Rechazamos la hipotesis nula si el estadístico es mayor que -1.96sd y menor que 1.96sd



greater

Rechazamos la hipotesis nula si el estadístico es mayor que -1.645sd.



less

Rechazamos la hipotesis nula si el estadístico es menor que 1.645sd.

Asunciones Teóricas de las Pruebas T

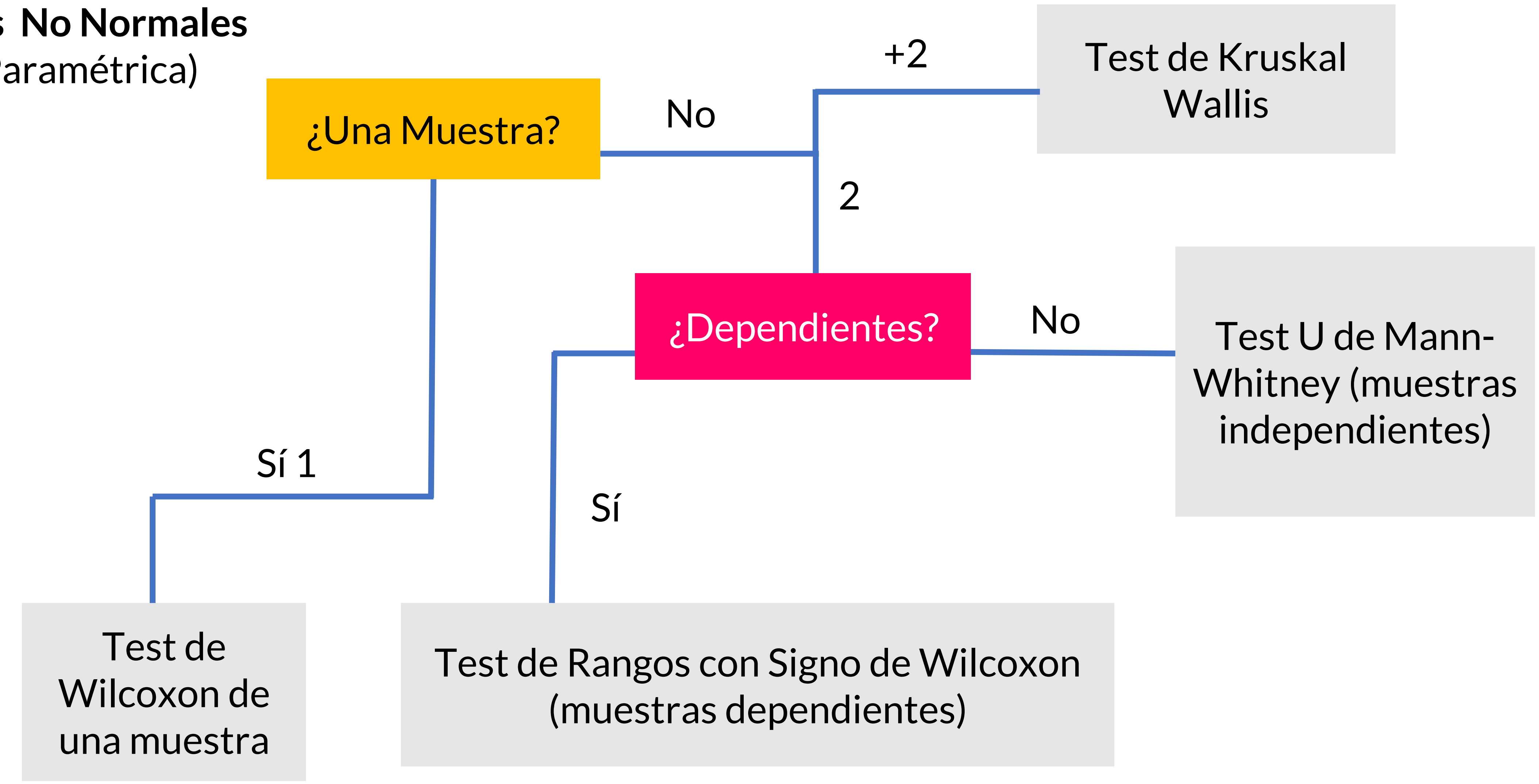
- **A1:** Los datos a analizar son resultados de mediciones (valores continuos).
- **A2:** Los datos a analiza fueron obtenidos por un muestreo aleatorio.
- **A3:** Debe haber homogeneidad de varianzas entre los grupos evaluados. No obstante, R lidia con esto aplicando la corrección de Welch cuando `var.equal = FALSE`.
- **A4:** Los datos de cada grupo tienen distribución normal (Test de Normalidad, Q-Q Plot). Para muestras >30 UM no hay necesidad de testear esta asunción, siempre y cuando se cumplan las otras.

Funciones en R: más de 2 muestras

```
# Anova
anova.res <- aov(y ~ x, data = BD)
anova.res <- anova(lm(y ~ x, data = BD))

summary(anova.res)

# Post Hoc
TukeyHSD(anova.res)
```

Lo veremos a detalle la siguiente semana cuando hablemos sobre ANOVAs.

Para datos No Normales
(Ruta No Paramétrica)



Funciones en R: 1 y 2 muestras

```r
# Test de Wilcoxon de una muestra
wilcox.test(x, mu = 0)

# Test de Rangos con Signo de Wilcoxon (muestras dependientes)
wilcox.test(x, y, paired = TRUE)

# Test U de Mann-Whitney (muestras independientes)
wilcox.test(x, y)
```

Aquí también podemos usar el **argumento `alternative`**, el cual tiene el mismo impacto en la "pregunta de estudio" que buscamos responder.

Funciones en R : más de 2 muestras

```
# Kruskall Wallis
Kruskal <- kruskal.test(y ~ x, data = datos)
Kruskal

# Post Hoc (metodo holm o Bonferroni)
pairwise.wilcox.test(x=respuesta, g=grupos, p.adjust.method="holm")
pairwise.wilcox.test(x=respuesta, g=grupos, p.adjust.method="bonferroni")
```

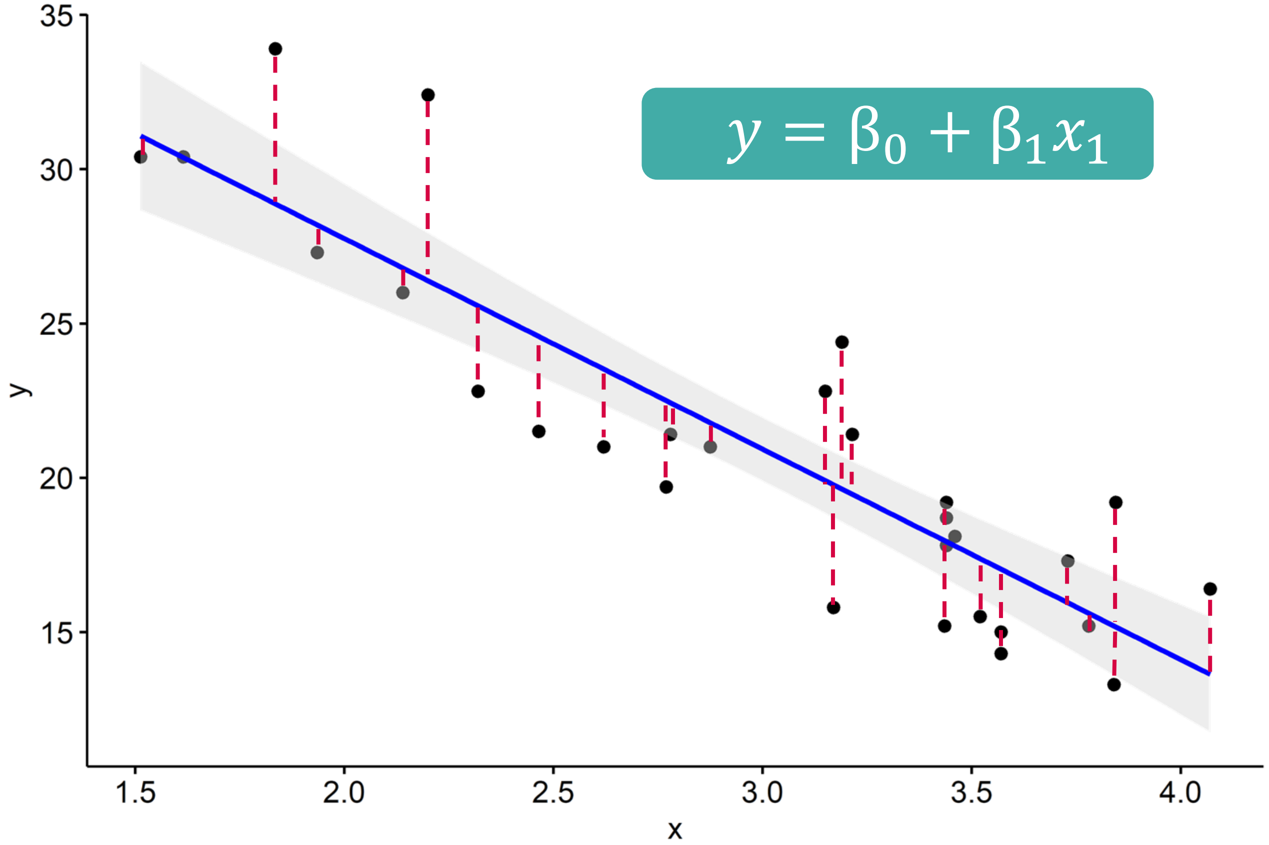
Lo veremos a detalle la siguiente semana cuando hablemos sobre ANOVAs.

Regresión lineal

Abre el archivo **“R-Notebook-C2-S2.R”** y
trabajaremos en la sección **4. Regresiones Lineales Simples**

Modelo Lineal

- Modelo = Abstracción de la realidad.
- Ecuación matemática que describe un fenómeno.



Estructura de la ecuación

y = variable respuesta (dependiente)

x = variable explicativa (independiente)

β_0 = intercepto de la curva

β_1 = pendiente de la curva

Modelo Lineal

- Los modelos lineales simples describen la relación entre dos variables como una línea.

Variable dependiente

Intercepto

Coeficiente de la Pendiente para x_1

Variable independiente

$$y = \beta_0 + \beta_1 x_1$$

Modelo Lineal

- Los modelos lineales simples describen la relación entre dos variables como una línea.

Variable **RESPUESTA**

Intercepto

Coeficiente de la Pendiente para x_1

Variable **EXPLICATIVA**

$$y = \beta_0 + \beta_1 x_1$$

Modelo Lineal

- Los modelos lineales simples describen la relación entre dos variables como una línea.

Debe ser CONTINUA

Intercepto

Coeficiente de la Pendiente para x_1

Puede ser CONTINUA/DISCRETA O CATEGÓRICA (ORDINAL O NOMINAL)

$$y = \beta_0 + \beta_1 x_1$$

Modelo Lineal

- Los modelos lineales simples describen la relación entre dos variables como una línea.

Debe ser **CONTINUA**

Intercepto

Coeficiente de la Pendiente para X

Puede ser **CONTINUA/DISCRETA O CATEGÓRICA**

$$y = \beta_0 + \beta_1 x_1$$

Continua/Discreta: el modelo linear prueba si hay una correlación significativa entre la variable explicativa y la respuesta

Categórica: el modelo linear prueba si hay diferencias significativas entre los nivel de la variable explicativa (x) respecto al valor promedio de la variable respuesta (y).

Modelo Lineal

- No toda la variación (Varianza) en un modelo lineal será explicada por la variable explicativa (X). El resto de la variación quedará como los residuales del modelo.

Variable **DEPENDIENTE, RESPUESTA**

Intercepto

Coeficiente de la Pendiente para x_1

Variable INDEPENDIENTE, EXPLICATIVA

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Término de Error (aproximado con el valor de los Residuales)

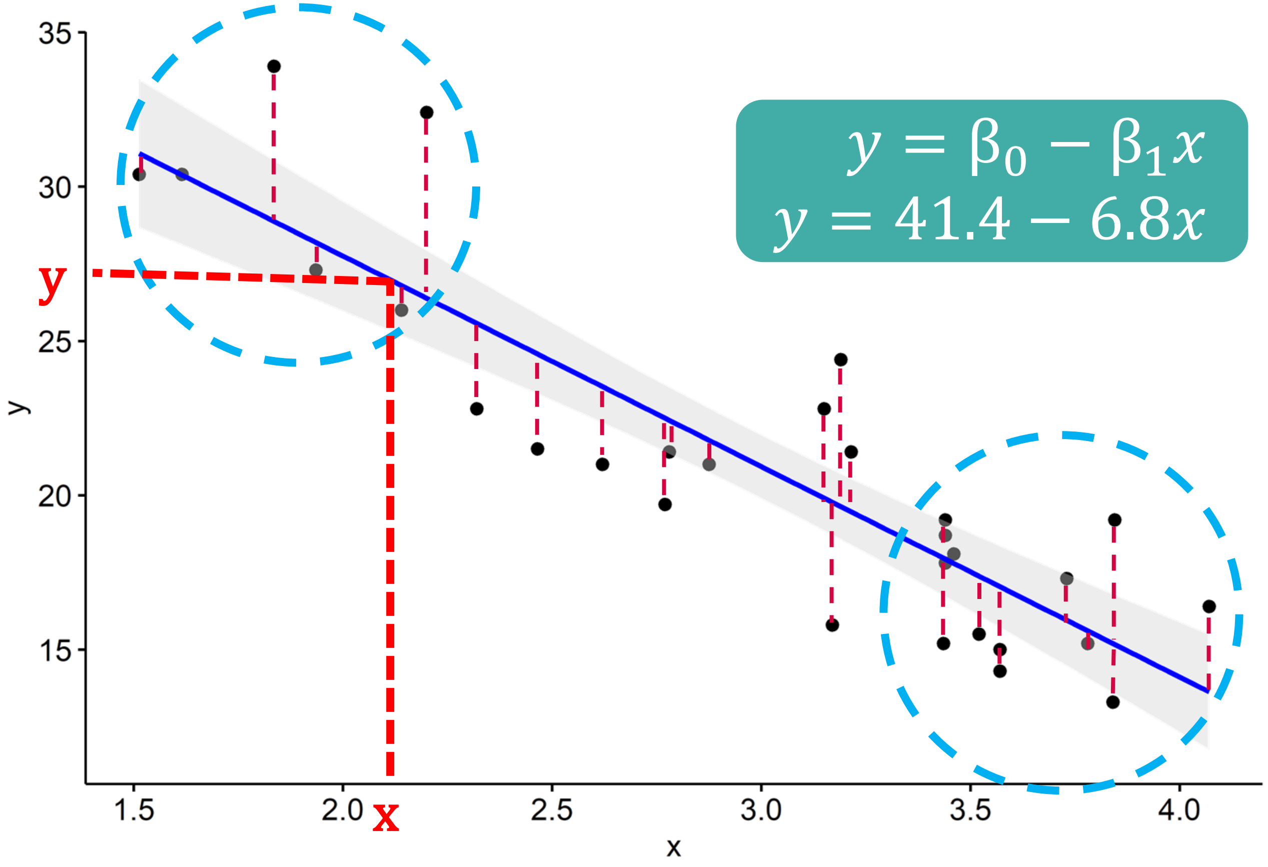
¿Qué nos dice un modelo lineal?

- Relación.
- ↑y ~ ↓x
- ↑x ~ ↓y
- Relación ≠ Causalidad.
- Gracias a la regresión, podemos predecir un **y desconocido** en base a un **x conocido**.

$$y = 41.4 - 6.8x$$

$$y = 41.4 - 6.8(2.25)$$

$$y = 26.1$$

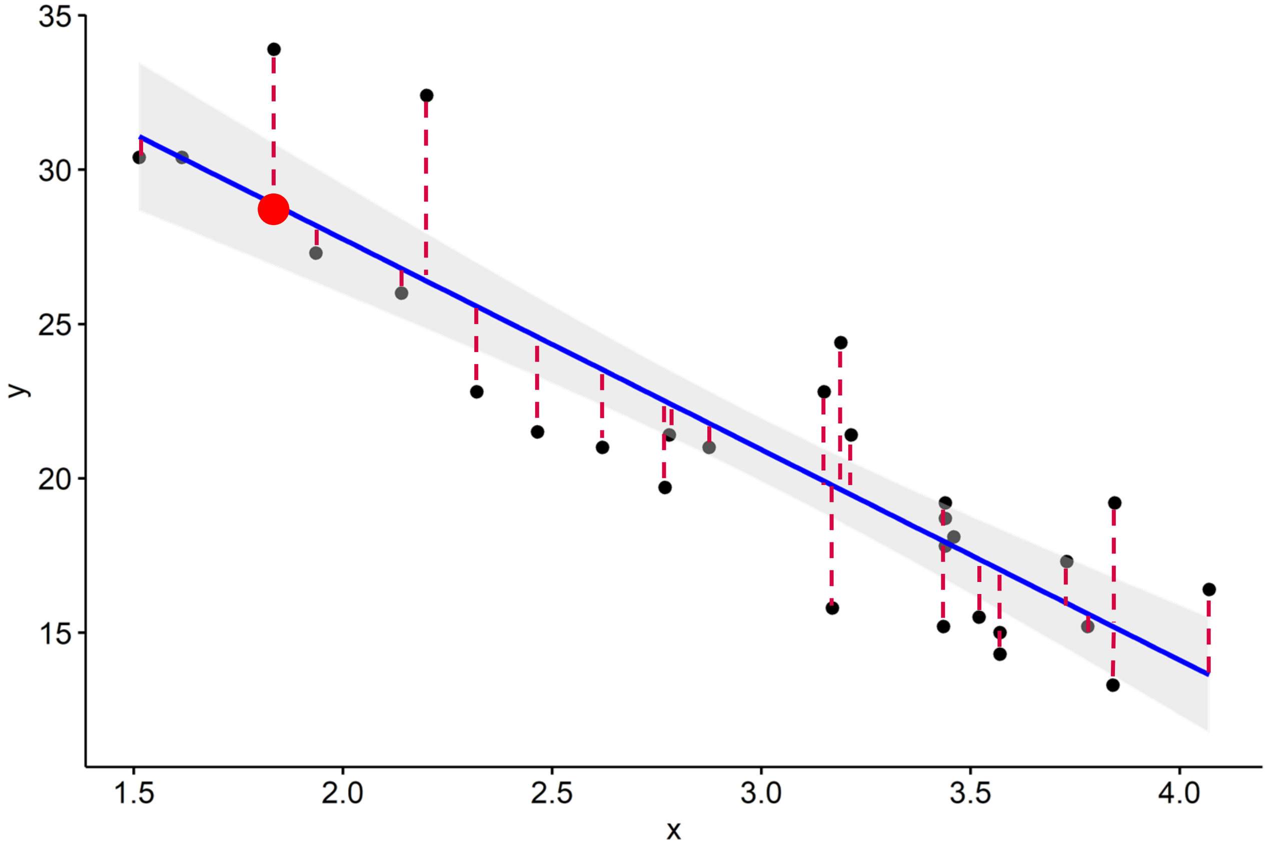


¿Qué busca el modelo lineal?

- Encontrar la mejor estimación de los parámetros (intercepto y coeficiente) y luego evaluar la bondad de ajuste del modelo (cuan bueno es).

Eso significa: reducir al mínimo los residuales en el modelo final.

```r
# Si deseamos obtener los
residuales de un modelo
residuals(modelo)
```



Residual = valor observado – valor predicho por la regresión

Tipos de modelos lineales

Método	Tipo de Var. Respuesta (Y)	Tipo de Var. Explicativa (X)	Número de Var. Explicativas	Número de Niveles
Regresión Lineal Simple	Continua	Continua	1	
t-Test	Continua	Categórica	1	2
ANOVA	Continua	Categórica	1 (ANOVA una vía), 2 (ANOVA de dos vías), o más	3 o más
ANCOVA	Continua	Continua y Categórica	2 o más	2 o más (si X es categórica)
Regresión Lineal Múltiple	Continua	Continua	2 o más	

Construcción de modelos lineales

- Modelo lineal simple

Coeficiente de la Pendiente para x_1

Intercepto

Variable RESPUESTA $\quad y = \beta_0 + \beta_1 x_1 \quad$ Variables EXPLICATIVAS

- Modelo lineal múltiple

Coeficientes de la Pendiente para cada variable explicativa

Intercepto

Variable RESPUESTA $\quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad$ Variables EXPLICATIVAS

Construcción de modelos lineales

```
# Modelo lineal simple
lm(y ~ x, data=BD)

# Modelo lineal múltiple (modelo aditivo, o de efectos principales)
lm(y ~ x1 + x2 + x3 + … + xn, data=BD)
```

Los efectos de cada x es independiente de las demás x.

```
# Modelo lineal múltiple (modelo de interacciones).
# Aquí lo ejemplificamos únicamente con dos variables explicativas.
lm(y ~ x1:x2, data=BD)

lm(y ~ x1 + x2 + x1:x2, data=BD)
lm(y ~ x1*x2, data=BD)
```

x1 y x2 tienen un efecto sinergético, los cambios de uno afecta a los cambios en el otro.

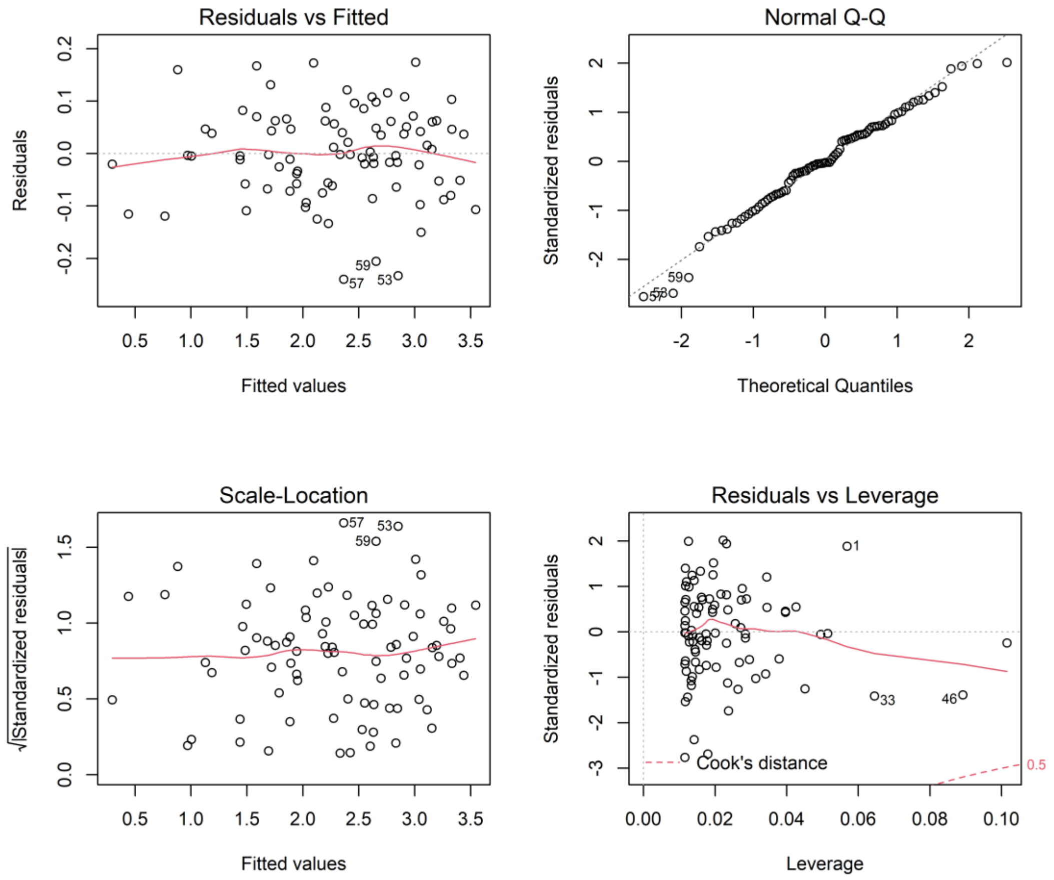
Asunciones Teóricas de la Regresión Lineal

- **A1:** la relación entre X e Y es lineal (ver gráfica de dispersión de puntos).
- **A2:** Los **residuales** son homocedásticos (el error no varía mucho cuando el valor del predictor varía, el error se mantiene constante a lo largo de los datos).
- **A3:** Los **residuales** son independientes (provienen de observaciones independientes)
- **A4:** Los **residuales** tienen distribución normal.

Validando las Asunciones Teóricas

- Vamos a ejecutar una regresión lineal simple en R y estudiaremos cómo es la validación de las asunciones teóricas.

```
# Gráficos de Diagnóstico
par(mfrow=c(2,2))
plot(modelo)
```



Validando las Asunciones Teóricas

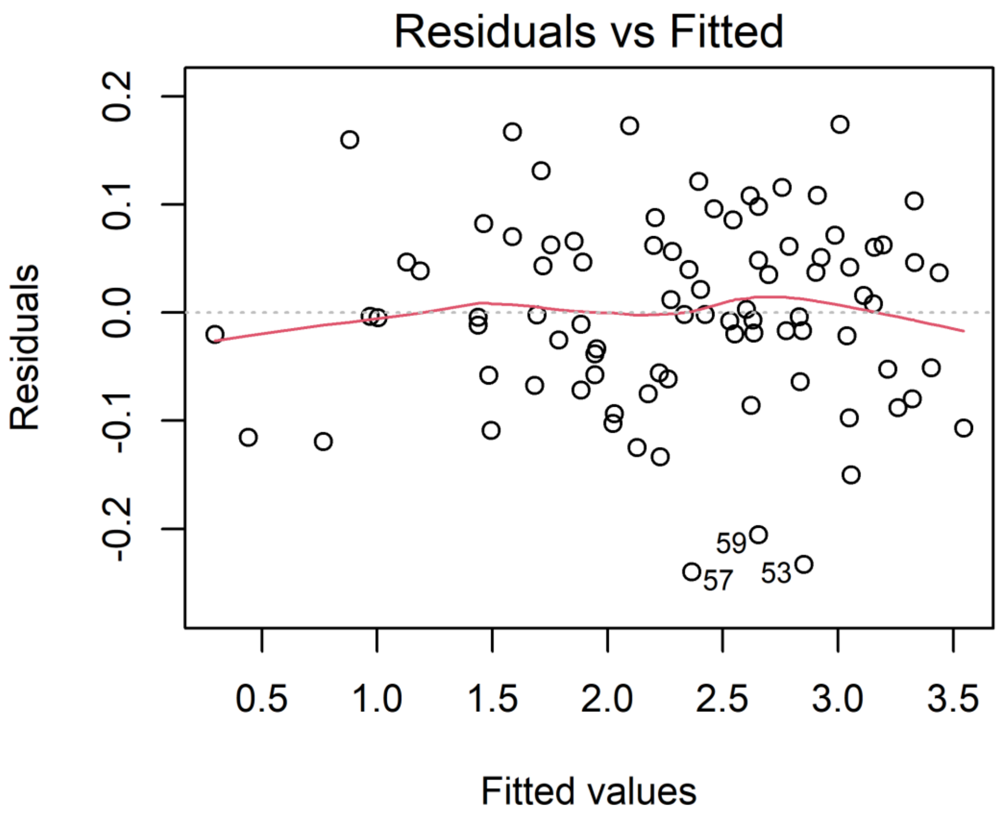


Gráfico 1 "Residuales vs Valores Estimados"

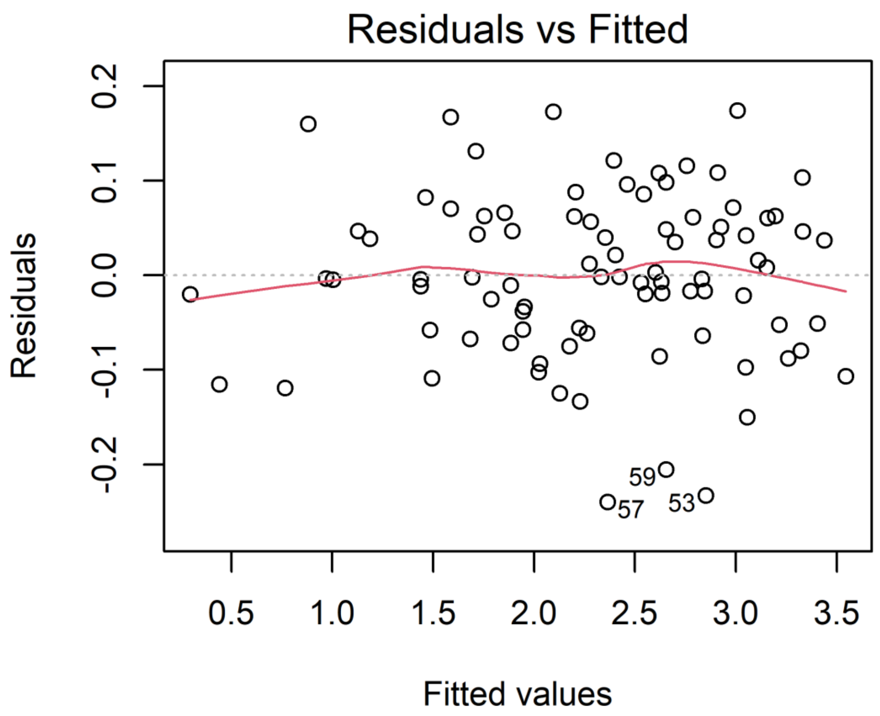
- Cada punto representa la distancia entre el valor de la variable respuesta y su valor estimado con el modelo.

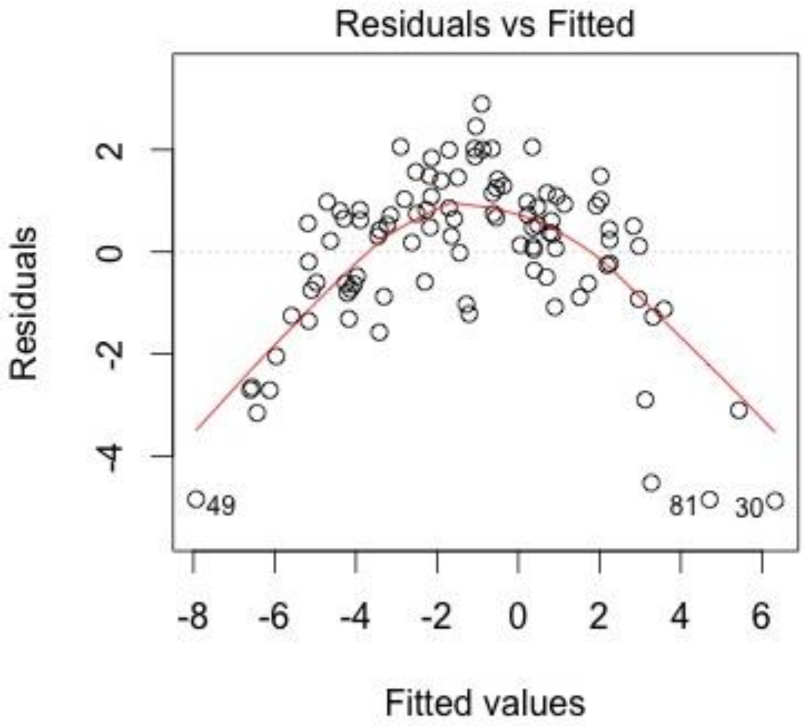
Interpretación

- Si los puntos se dispersan aleatoriamente alrededor de la línea horizontal 0, entonces indica linealidad de los residuales.

Validando las Asunciones Teóricas

Gráfico 1 "Residuales vs Valores Estimados": verificar linearidad





Validando las Asunciones Teóricas

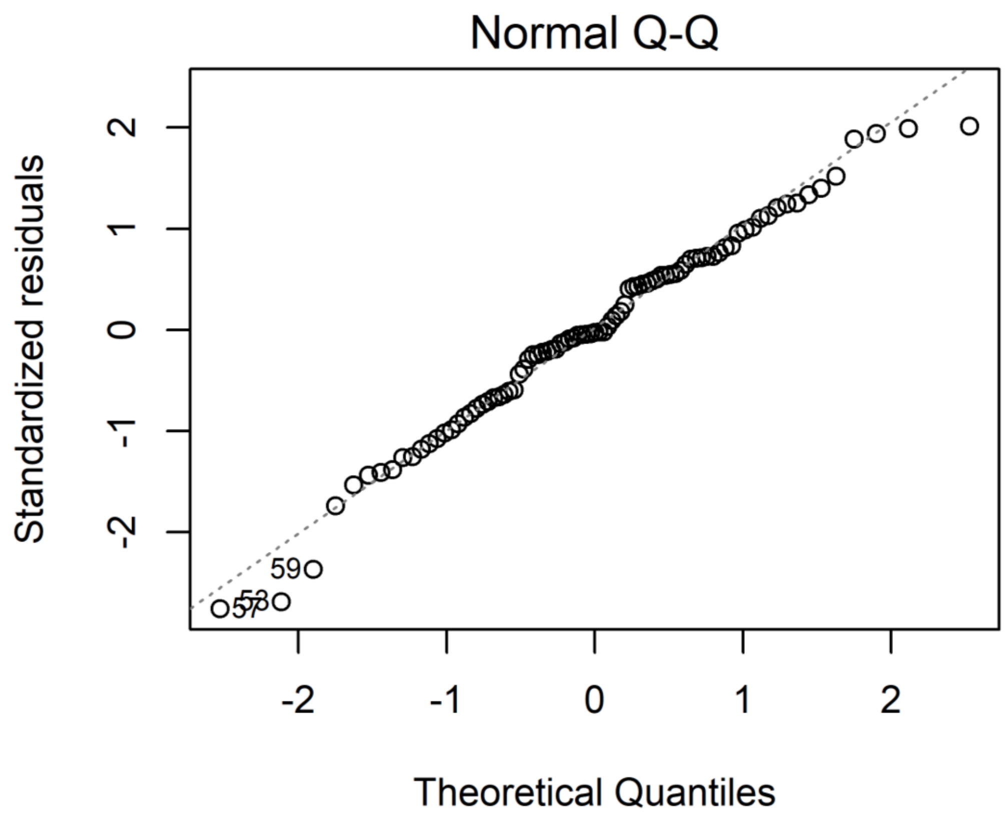


Gráfico 2 "Q-Q Plot"

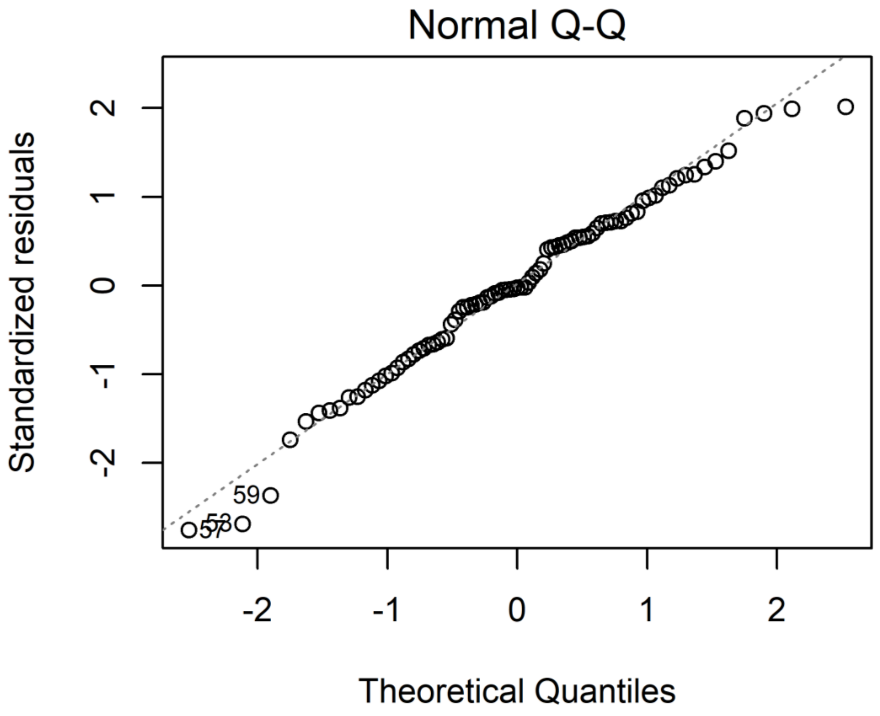
- Permite verificar la normalidad de los errores (residuales).

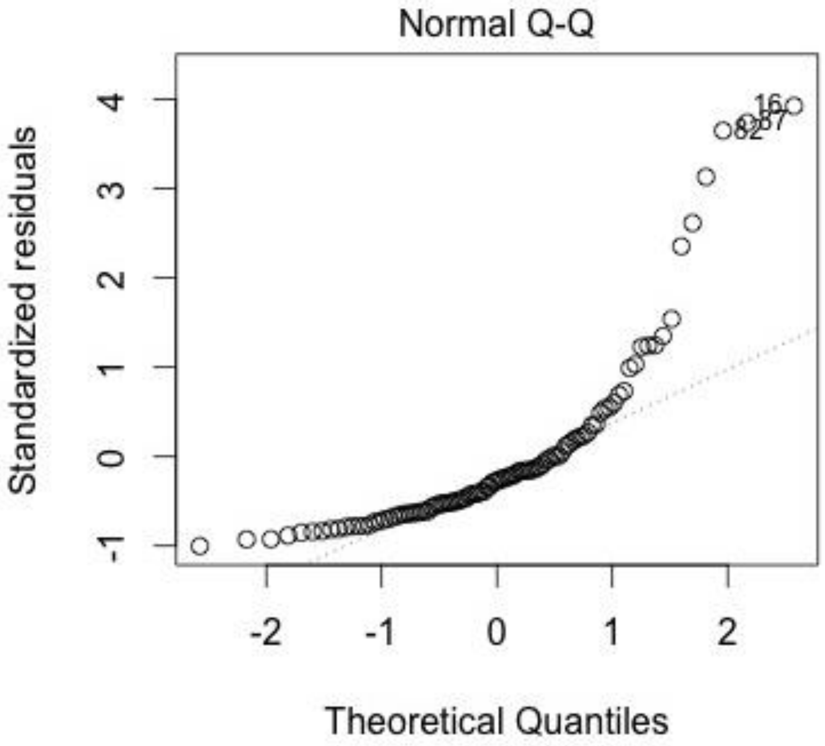
Interpretación

- Si los puntos se distribuyen sobre o muy cerca de la línea 1:1 de normalidad (la diagonal del Q-Q Plot), entonces hay normalidad.
- Puedo aceptar ciertas ligeras desviaciones, principalmente sobre los extremos de la distribución

Validando las Asunciones Teóricas

Gráfico 2 "Q-Q Plot": verificar normalidad





Validando las Asunciones Teóricas

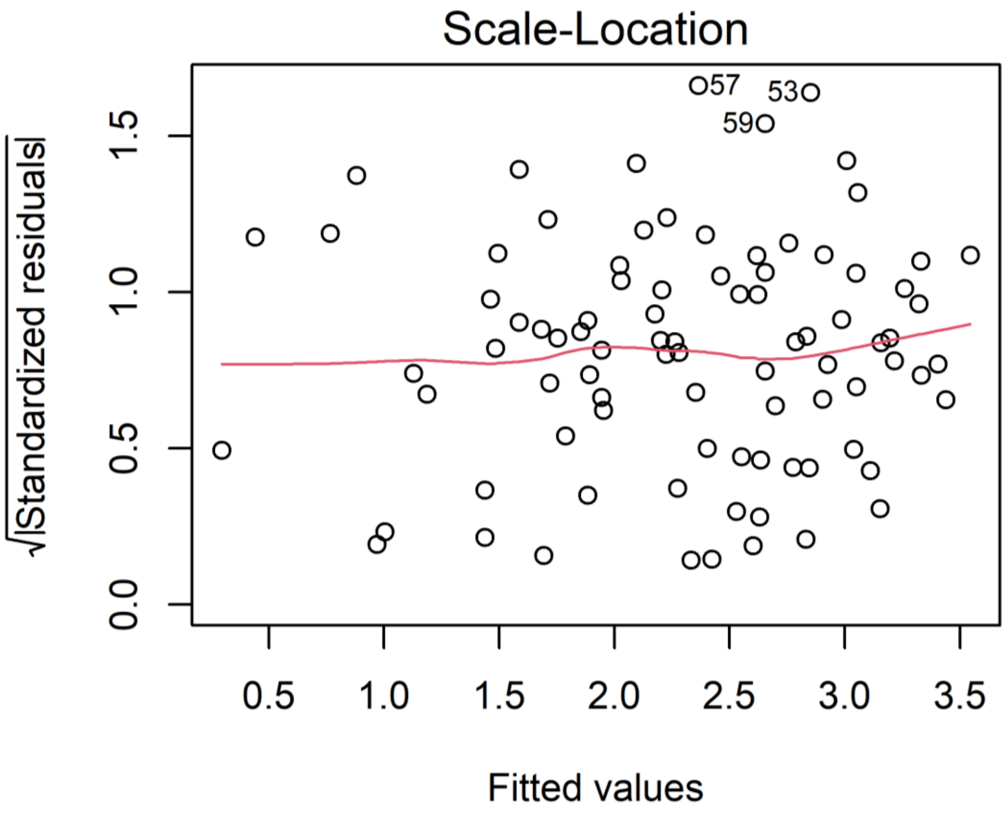


Gráfico 3 "de Escala-Localización"

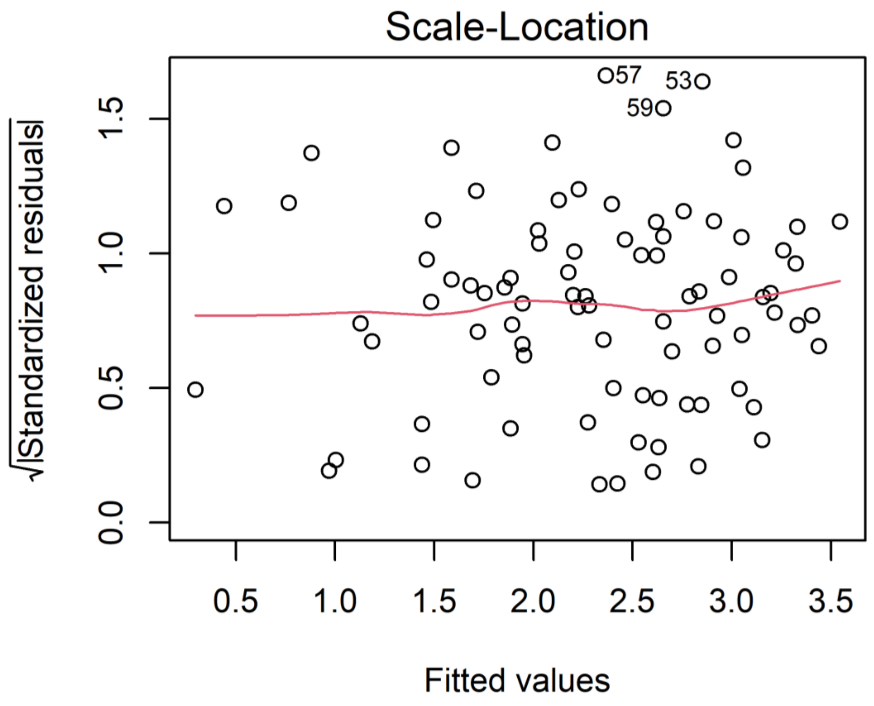
- Permite identifica si la dispersión de los errores (residuales) incrementa con los valores estimados con el modelo.

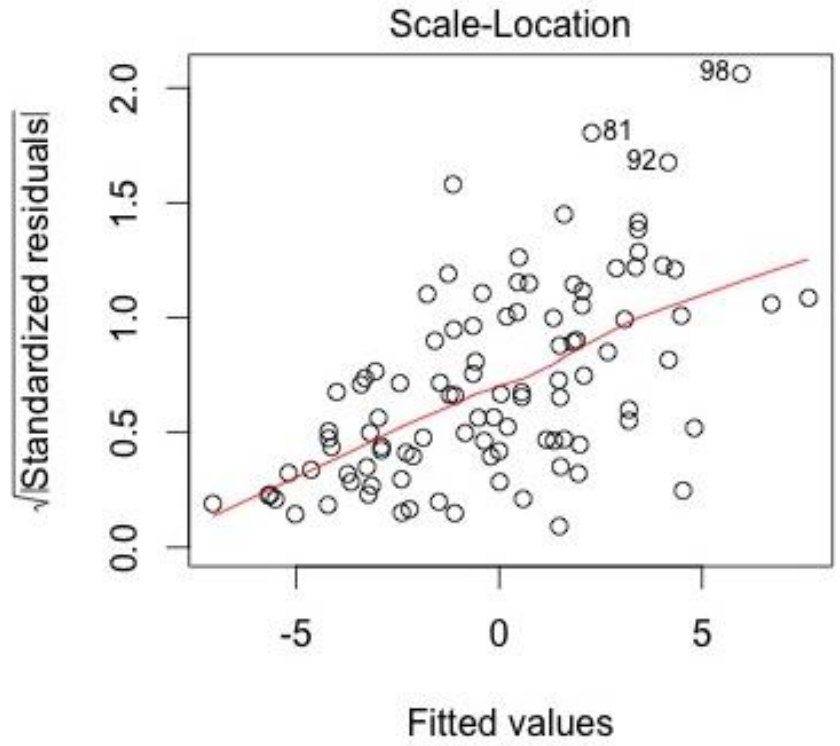
Interpretación

- Si los puntos no forman ningún patrón (se distribuyen aleatoriamente), y la línea roja que forman se mantiene relativamente horizontal, cumplo y confirmo la asunción de homogeneidad de varianza (Homocedasticidad).

Validando las Asunciones Teóricas

Gráfico 3 "de Escala-Localización": verificar homocedasticidad





Validando las Asunciones Teóricas

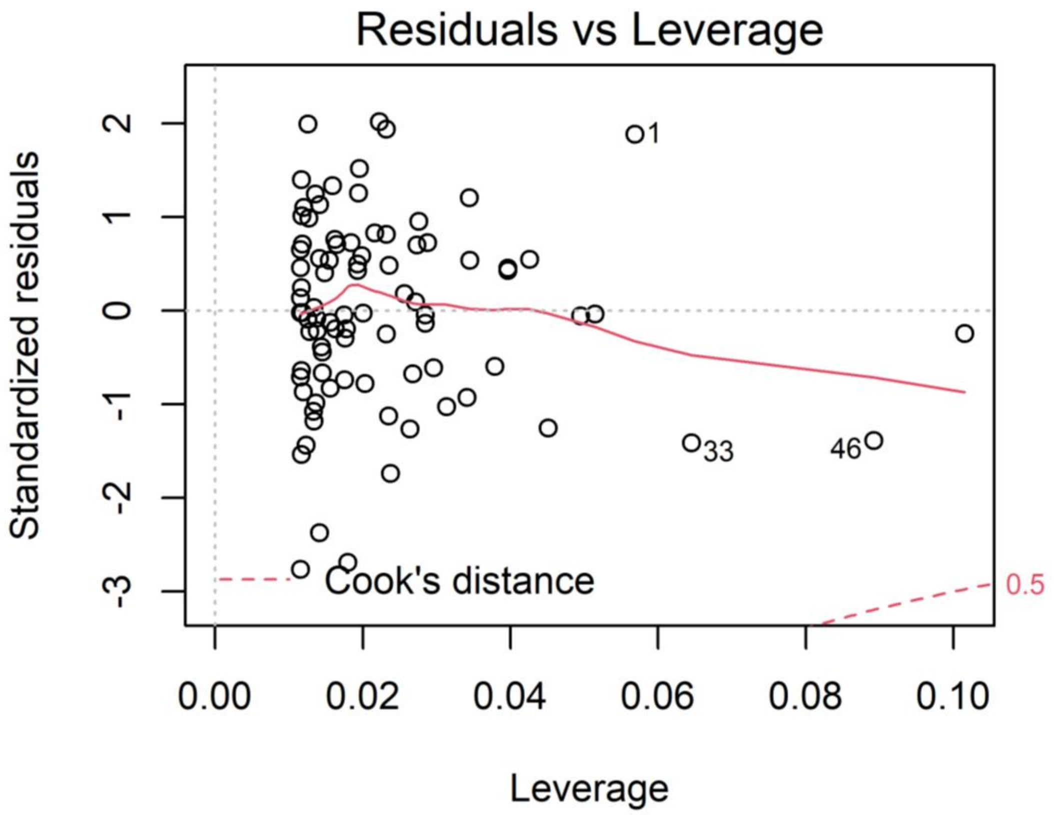


Gráfico 4 "Residuales vs Leverage"

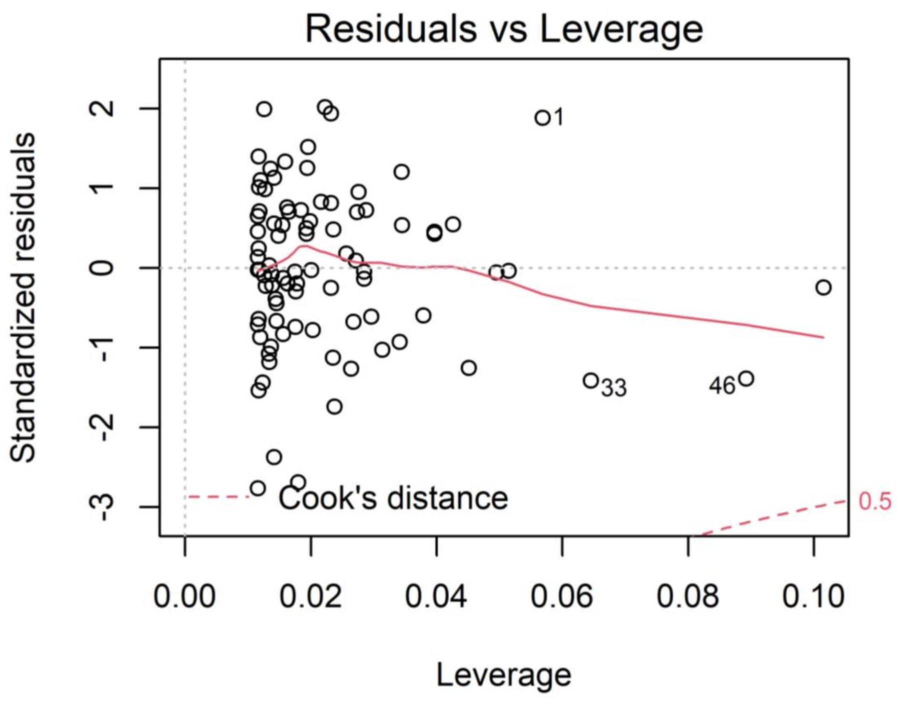
- No permite verificar ninguna asunción, pero identifica los **valores influyentes**, con diferencias muy marcadas respecto al resto de valores. Ojo, no todos los outliers terminan siendo valores influyentes en la regresión.

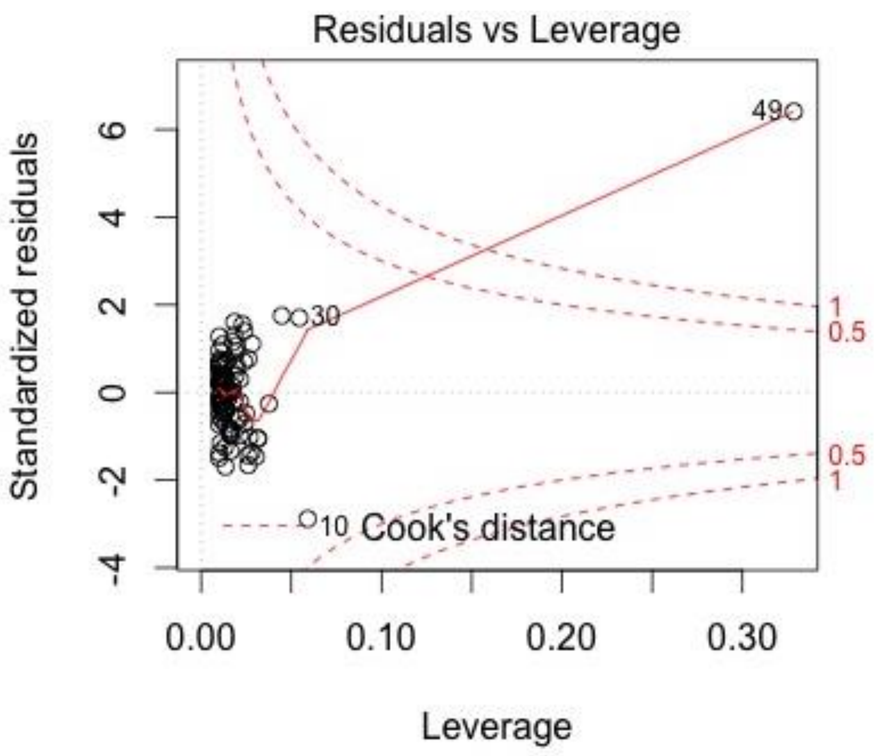
Interpretación

- Si un punto (dato o fila en la data frame) se acerca demasiado o sobrepasa la distancia de Cook (Cook's distance), definida por la **línea roja punteada**, debe ser removido.

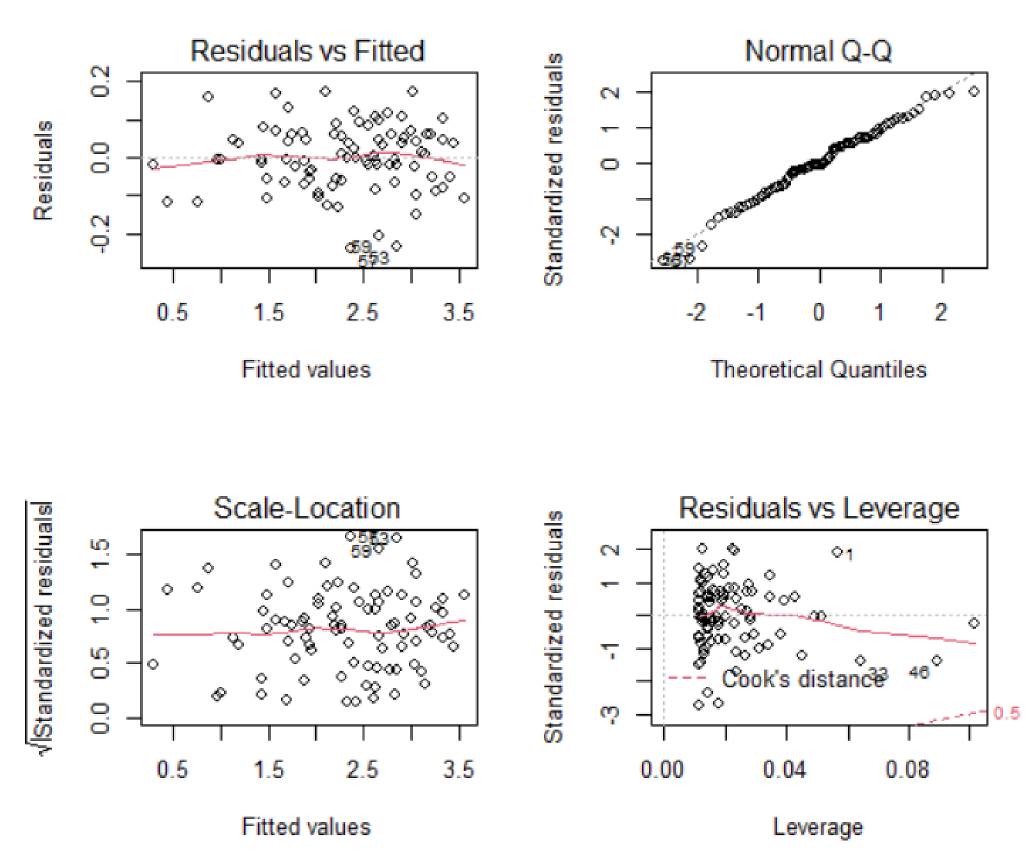
Validando las Asunciones Teóricas

Gráfico 3 "de Escala-Localización": identificar valores influyentes

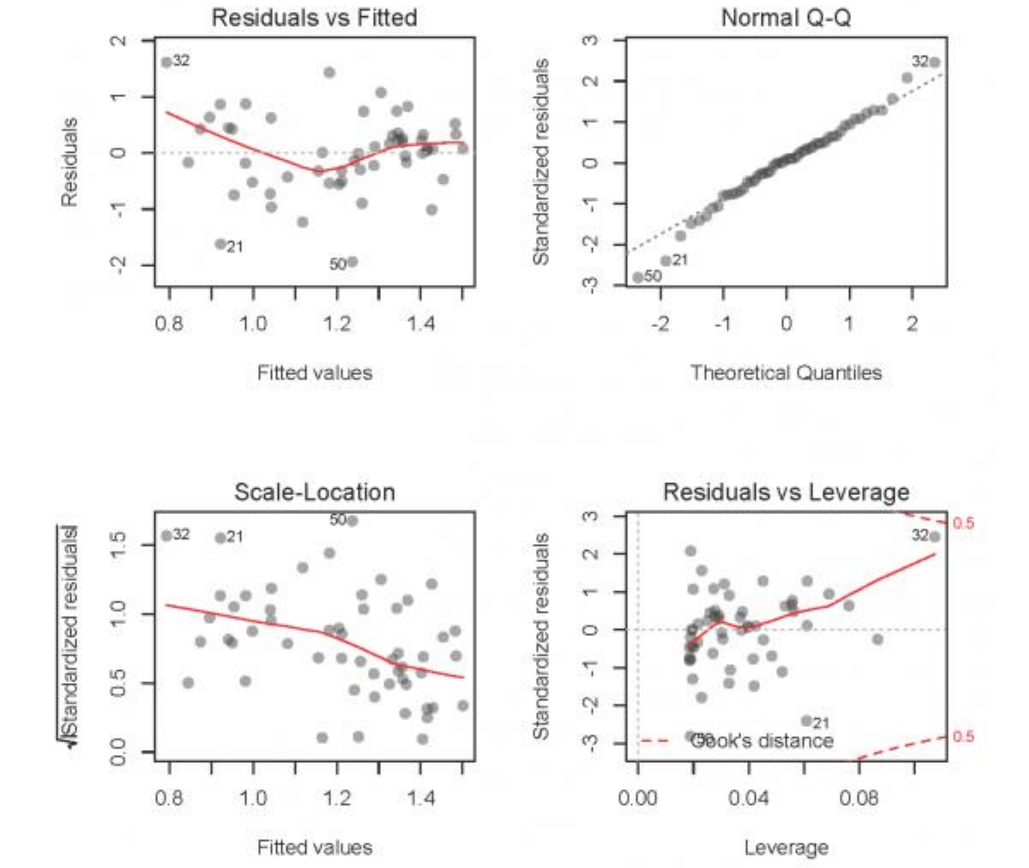




Validando las Asunciones Teóricas

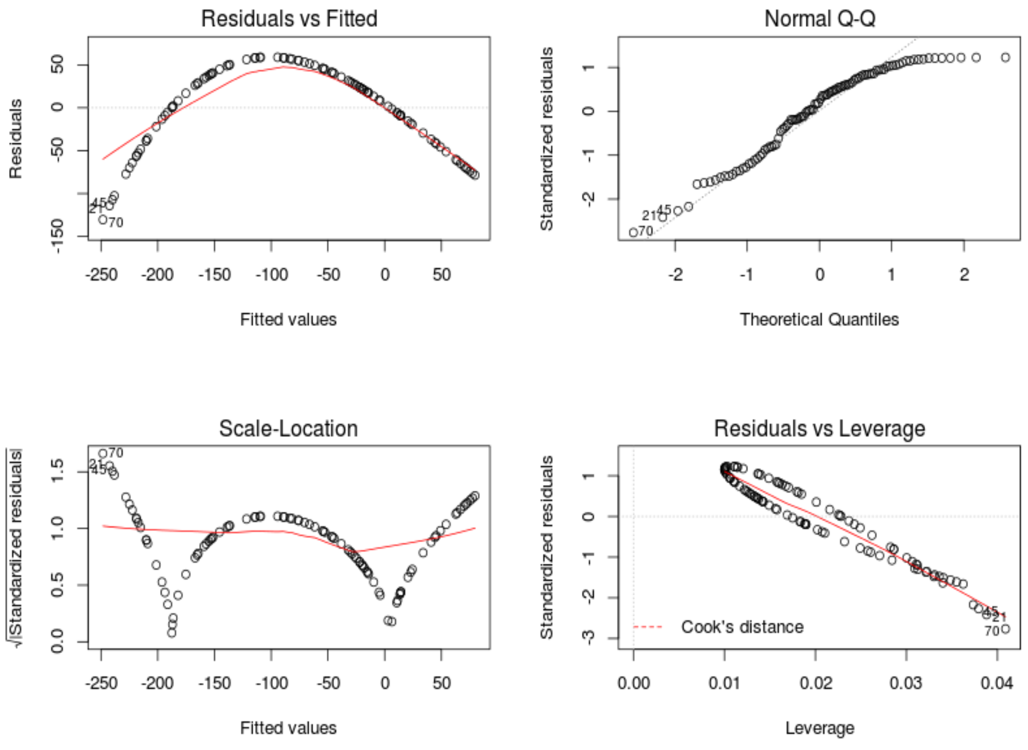


CASO IDEAL

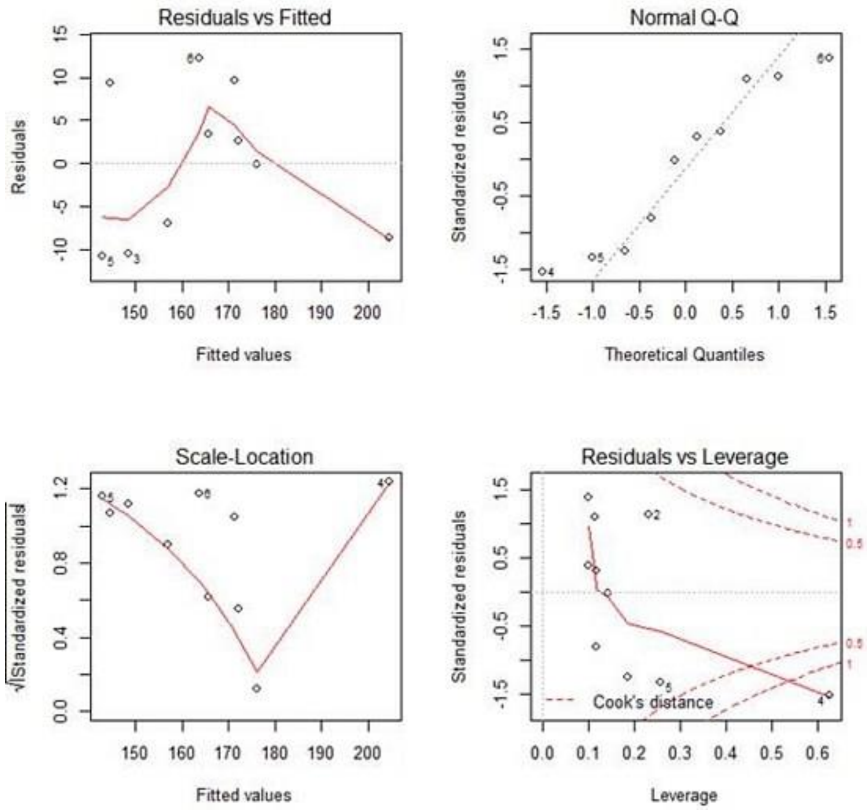


CASO ACEPTABLE

Validando las Asunciones Teóricas



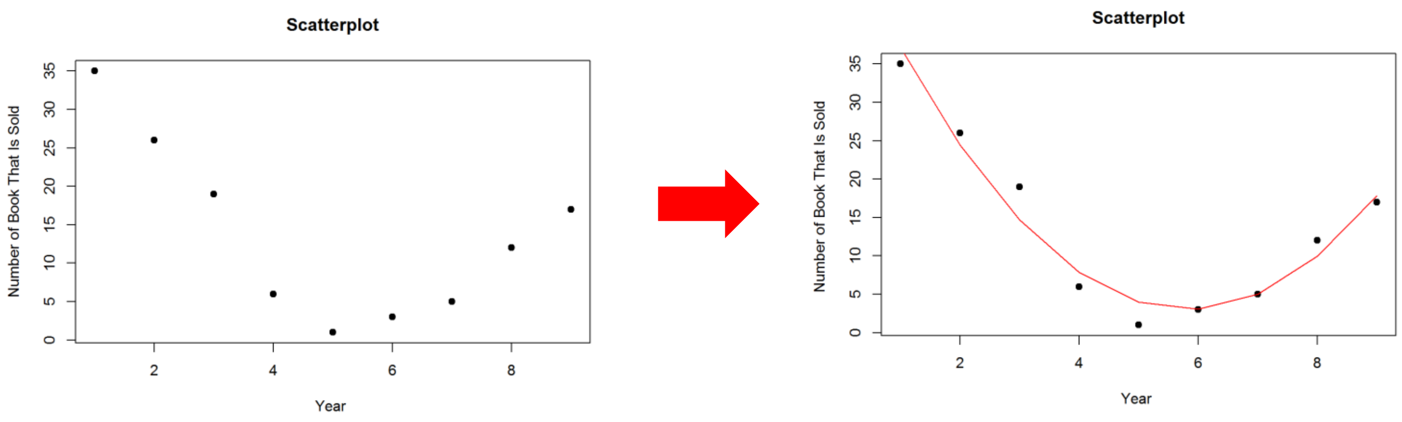
NO SE CUMPLEN LAS ASUNCIONES
(Usar otros métodos de modelamiento)



ASUNCIONES MUY POCO SUSTENTADAS

¿Qué sucede si no cumplo las asunciones?

- No significa que tu investigación está mal.
- No todo es un modelo lineal:
  - Puedes incorporar un término cuadrático en la ecuación.
  - O quizá una regresión lineal no era adecuada para tu variable explicativa (ver siguiente lámina).
  - O puedes realizar la transformación de la variable explicativa (normalizar) con `bestNormalize()`.
  - O quizá hay otras variables que no incluiste en el análisis que juegan un rol importante en el fenómeno estudiado.
- En última instancia, quizá sí hubo un sesgo de muestreo o diseño experimental más definido.



[How to Fit a Quadratic Curve to Data in R (rstudio-pubs-static.s3.amazonaws.com)]

Resultados del Modelo Lineal

```
> summary(modelo)

Call:
lm(formula = temp_anomaly ~ carbon_emissions, data = temp_carbon)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29704 -0.07938 -0.00903  0.09615  0.40084 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.591e-01  1.690e-02  -15.33   <2e-16 ***
carbon_emissions  9.994e-05  4.202e-06   23.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1338 on 133 degrees of freedom
  (133 observations deleted due to missingness)
Multiple R-squared:  0.8096,	Adjusted R-squared:  0.8082 
F-statistic: 565.7 on 1 and 133 DF,  p-value: < 2.2e-16
```

- **p-value:** significancia del modelo calculado con el valor F.
- **Adjusted R-squared:** varianza de y explicada por x en el modelo.
- **Coeficientes (Estimados, β):** intercepto, y pendiente de la variable explicativa x.
- **Error estándar (Std. Error):** mide la cantidad promedio en que los β varían del valor promedio real. Mientras más pequeño, mejor.
- **t value:** cantidad de sd del promedio se encuentra el coeficiente. Mientras más lejos de 0, mejor.
- **Pr(>|t|): p valor.** probabilidad de encontrar valores mayores igual al valor crítico t value. Es decir, la probabilidad de que la relación sea debida al azar y no a un proceso real.
- **Residual standard error:** es el término de error de la ecuación, mientras más pequeño, mejor el fit del modelo.

Resultados del Modelo Lineal: interpretación

```
> summary(modelo)

Call:
lm(formula = temp_anomaly ~ carbon_emissions, data = temp_carbon)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29704 -0.07938 -0.00903  0.09615  0.40084 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.591e-01  1.690e-02  -15.33   <2e-16 ***
carbon_emissions  9.994e-05  4.202e-06   23.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1338 on 133 degrees of freedom
  (133 observations deleted due to missingness)
Multiple R-squared:  0.8096,	Adjusted R-squared:  0.8082 
F-statistic: 565.7 on 1 and 133 DF,  p-value: < 2.2e-16
```

- Luego de verificar que el modelo es significativo, interpretamos,
  - El β del intercepto es el valor promedio esperado de y cuando tomamos en consideración todos los datos de x.
  - No es importante si es significativo o no.
  - El β de la variable explicativa x, debe ser significativo. Este β indica por cada unidad (1) que aumente x, y aumenta en β.

Más allá del modelo lineal

- La elección depende de la variable respuesta…
  - y = numérica (medición) …………………………. Modelos asociados a la regresión lineal LM
  - y = numérica (conteo) ……………………………. Regresión Logística de Poisson (y afines)
  - y = binaria (0,1) ………………………………………. Regresión Logística Binomial
  - y = proporciones (<0 & >1) ………………………. Regresión beta
  - y = factor …………………………………………………. Regresión Logística Multinomial o extension de R. Logística
- … algunas veces de la necesidad de extraer la variabilidad de variables de agrupamiento…
  - y = no gaussiana ……………………………………y = gaussiana ……………………………………………. LMM
  - .. GLMM
- … otras, requerimos lidiar con procesos de Poisson especiales.
  - y = con muchos ceros ………………………………. Regresión de Poisson con Ceros Inflados (o *Hurdle Models*)
  - y = con ausencia de ceros ………………………….. Regresión de Poisson con Ceros Truncados
  - y = con sobredispersión ……………………………. Regresión Quasipoisson o R. Binomial Negativa

Lidiando con outliers
en modelos lineales

Abre el archivo **“R-Notebook-C2-S2.R”** y
trabajaremos en la sección **5. Lidiando con Outliers**

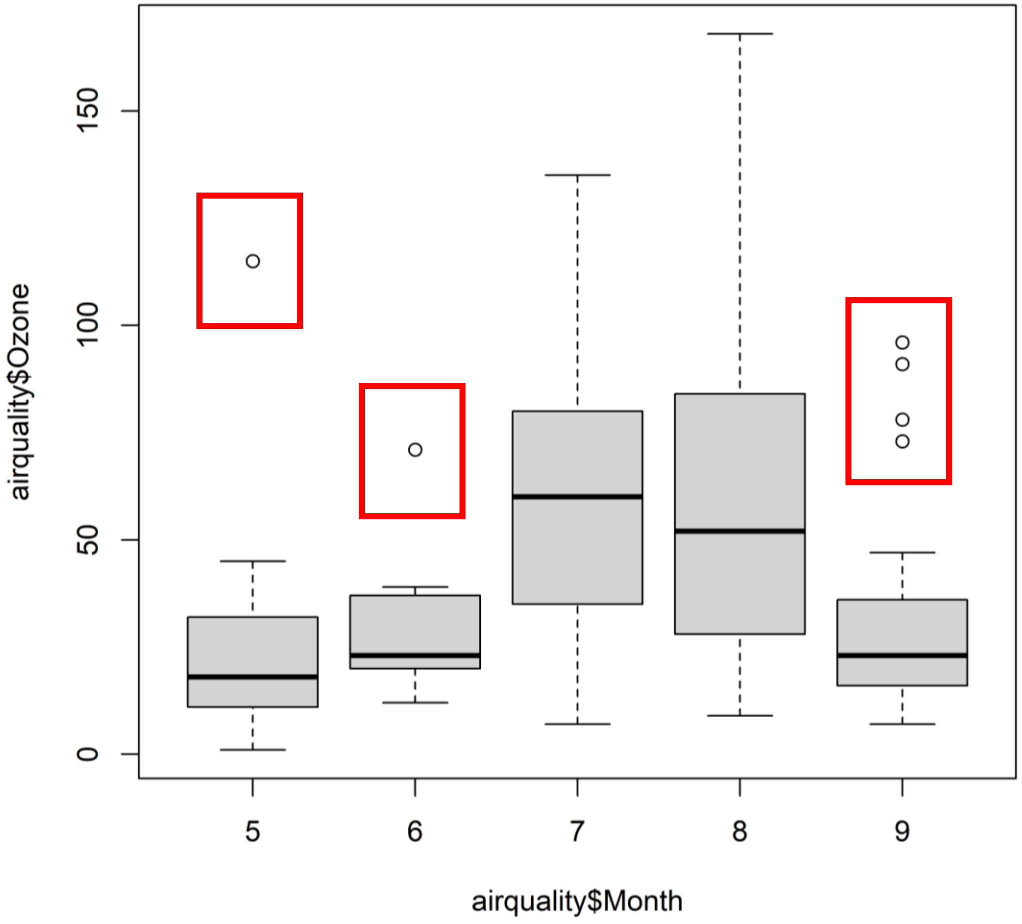
Outliers (valores atípicos, extremos)

- Es un valor que es muy diferente del resto de valores del conjunto de datos.
- Puede indicar tanto **error de muestreo**, **error de tipeo**, o, si no es un error, nos puede dar indicios de que algún **proceso** estaría generando valores extremos en mi población.
- Son muy problemáticos y modifican bastante los resultados y el tamaño del efecto (R^2) de las regresiones y deben ser removidos del conjunto de datos.

¿Cómo “se ven” los outliers?

- Outlier, valor que se encuentra a más de 1.5 veces la distancia intercuartil (distancia entre Q1 y Q3 ~ la caja).

```
# Identificar las filas de la Base de
datos BD que contienen outliers en una
columna definida
Library(rstatix)
identify_outliers(BD, Columna)
```



Instituto
de Ciencias
Antonio Brack