TÉCNICAS Y MÉTODOS DE MODELADO

17/11/2021

Técnicas de reducción de dimensiones:

- El análisis multivariante tiene sus orígenes en el siglo xx.
- Surge con la psicología que aplica técnicas que tratan de medir la inteligencia.
- Spearman (1904) y Pearson (1901) trataron de establecer una variable que midiera la inteligencia.
- **ANÁLISIS DE COMPONENTES.**

Análisis Multivariante:

- Se dedica al estudio de varias variables de modo simultáneo.
- Se consideran varios aspectos y trata de determinar la relación entre las medidas.

Variables:

- Nominales: Distinguen entre varias categorías, sin que exista ninguna jerarquía entre ellas.
- Ordinales: Distingue categorías para una variable y establece
- Intervalo: Combinación de las variables anteriores. Agrega sentido a la diferencia de valores.
- Razón: Son idénticas a las anteriores salvo que presentan un origen absoluto de medida

Técnicas multivariantes:

Método Dependiente:

La asociación entre las distintas variables, es decir, en las relaciones entre las mismas, donde parte de estas variables dependen o se miden en función de las otras -> **Interés predictivo**

Métodos Dependientes

- Regresión lineal: Estudia la dependencia de una variable en función de otras
- Análisis discriminante: Se busca una función lineal de varias variables que permita clasificar nuevas observaciones que se presenten
- Métodos log-lineales y logit: Se predicen números de apariciones en casillas en función de otras
- Análisis de correlación canónica: Se toma un grupo de variables y se trata de predecir sus valores en función de otro grupo de variables
- Análisis multivariante de la varianza: Se descompone la variabilidad en una medida de un conjunto de variables cuantitativas en función de otras variables

Método Independiente:

Se esta interesado en investigar las asociaciones que se presentan entre variables sin distinción de tipos entre ellas -> **Interés descriptivo**

Métodos Independientes

- Análisis de componentes principales (ACP): Se tienen n variables cuantitativas y se mezclan mediante combinaciones lineales reduciéndose a $p < n$ variables que resumen la información para facilitar la interpretación.
- Análisis factorial: Parecido al ACP aunque sólo se fija en explicar en términos de factores ocultos las variables originales, no tanto en reducir el número de variables.
- Multidimensional scaling: Busca mapas de los objetos, situándolos según una serie de métricas
- Análisis de correspondencias: Es parecido al análisis factorial, pero con variables categóricas exclusivamente.
- Análisis de cluster: Trata de identificar grupos naturales entre las observaciones según sus valores medidos por las variables.

CONCEPTOS BÁSICOS

Vector

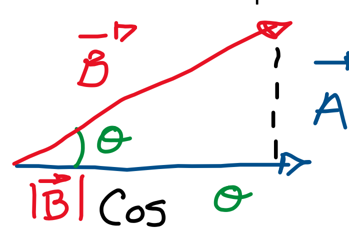
- Vector: Representación de los valores de una variable en una muestra de n elementos
- Modulo o noma: Longitud de un vector
- Suma o diferencia de vectores: Es un nuevo vector con componentes iguales a la suma (diferencia) de los componentes de los sumandos
  (La suma o diferencia es asociativa y conmutativa)

$$x+y=\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

- Covarianza: Mide la dependencia lineal entre dos variables
- Producto escalar: Herramienta para estudiar la relación entre 2 vectores (equivalente a la correlación). Es el escalar obtenido al sumar los productos de sus componentes

$$x'y = y'x = \sum_{i=1}^{n} x_i y_i$$

$$\cos\theta = \frac{x'y}{|x||y|}$$



Producto escalar de 2 vectores

Matriz

- Matriz: Conjunto de números acomodadas en filas y columnas
- Matriz transpuesta (A)’: Matriz que intercambia filas por las columnas

Operaciones de matrices

- Suma de matrices: Se realiza cuando dos matrices tienen la misma dimensión (Sumar los valores de las variables correspondientes)
- Producto de matrices: Solo es posible cuando el número de columnas de A es igual al numero de filas B
  a) A (B+C) = AB + AC
  b) (AB)’ = B’ A’
  c) AI = IA = A

  (A’ = Matriz transpuesta)

  (El producto de matrices no es conmutativo)

- Matriz simétrica: Son aquellas que tienen cada dila igual a la correspondiente columna A’=A

$$A = \begin{pmatrix} -2 & 1 & 3 \\ 1 & -3 & 2 \\ 3 & 2 & 0 \end{pmatrix} \qquad A^T = \begin{pmatrix} -2 & 1 & 3 \\ 1 & -3 & 2 \\ 3 & 2 & 0 \end{pmatrix}$$

- Matriz identidad (I): Tiene “ 1 “ en la diagonal y ceros fuera de ella

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Matriz cuadrada: Es una matriz cuadrada si el numero de filas es igual al numero de columnas, es decir, n = m

$$A_{2x2} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad A_{3x3} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}, \quad A_{nxn} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

- Determinante matriz (|A|): Es el resultado de restar la multiplicación de los elementos de la diagonal principal con la multiplicación de los elementos de la diagonal secundaria

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11} \cdot a_{22} - a_{21} \cdot a_{12}$$

$-$ $+$

Det

$$A = \begin{vmatrix} 4 & 1 \\ 6 & 2 \end{vmatrix}$$

$$det = |A| = (4)(2) - (1)(6)$$

$$|A| = 8 - 6 = \underline{2}$$

Determinante de una matriz

- El determinante de una matriz de varianzas y covarianzas es una medida global de la independencia entre las variables
- Cuanto mayor sea el determinante mayor es la independencia entre los vectores
- Traza de una matriz cuadrada: Es la suma de los elementos de la diagonal principal de la matriz

$$A = \begin{bmatrix} 5 & 7 & 0 \\ -1 & 4 & 3 \\ 0 & 2 & 5 \end{bmatrix} \qquad traz(A) = 5+4+5 = 14$$

- Matriz adjunta (Adj A): Es una matriz cuadrada de orden n. La adjunta de una matriz A es la transpuesta de la matriz cofactor de A

$$A = \begin{bmatrix} 2 & -1 \\ 7 & 3 \end{bmatrix} \qquad adj\ A = \begin{bmatrix} 3 & 1 \\ -7 & 2 \end{bmatrix}$$

- Matriz inversa (A-1): Está definida como aquella matriz que multiplicada por la original de por resultado la matriz identidad

$$A = \begin{bmatrix} 4 & -2 \\ 0 & 1 \end{bmatrix} \qquad A^{-1} = \frac{1}{det\ A}\ adj\ A$$

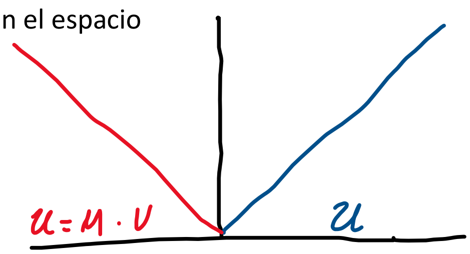
$$det\ A = 4$$

$$adj\ A = \begin{vmatrix} 1 & 2 \\ 0 & 4 \end{vmatrix} \qquad A^{-1} = \begin{bmatrix} 1/4 & 1/2 \\ 0 & 1 \end{bmatrix} \qquad A^{-1} = \frac{1}{4}\begin{bmatrix} 1 & 2 \\ 0 & 4 \end{bmatrix}$$

- Matriz ortogonal: Matriz cuadrada que representa un giro en el espacio

$$M = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \qquad M^T = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$M \cdot M^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad M_a = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 2 \end{pmatrix}$$



- Rango de una matriz: También llamado característica de una matriz es el orden máximo de sus valores no nulos

$$A_{2\times3} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \rightarrow submatriz \quad A_{2\times2} = \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix}$$

$$Det: |sub_{2\times2}| = 1\cdot5 - 2\cdot4 = 5 - 8 = -3 \neq 0 \rightarrow Rm\ A\ es\ 2$$

El rango de una matriz es buscar la mayor submatris y Calcular su determinante, det diferente de 0, entonces el rango será -#filas o #columnas

- Dada una matriz cuadrada se espera que cumpla ciertas propiedades
- Invariantes ante transformaciones lineales
- Que preserven la información existente
- Auto valores de una matriz: Conocidas como valores propios o raíces características
  Son medidas básicas de tamaño de una matriz, las cuales no se ven alteradas por transformaciones lineales de esta matriz

  $$A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

  (Tiene como autovalores 2,3 y 0. El 0 con multiplicidad de 2)

- Autovectores: Representan las direcciones características de la matriz y no son invariantes

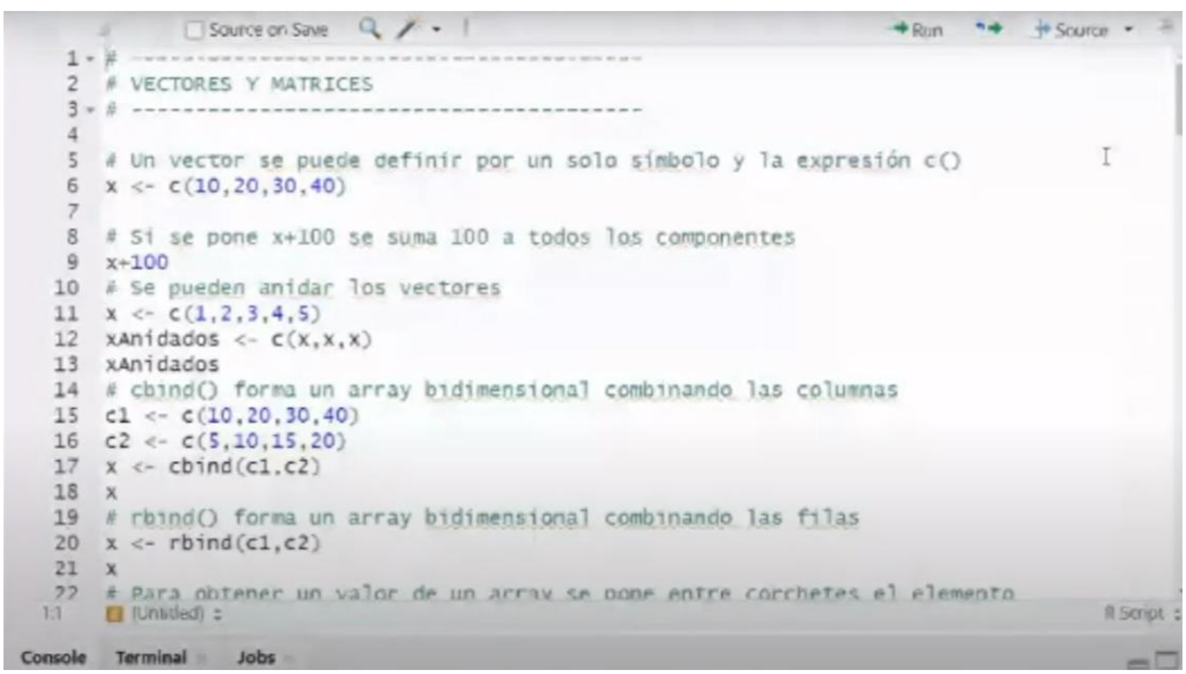
$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

(tiene como autovalores 1 y 2, 1 con multiplicidad de 2)

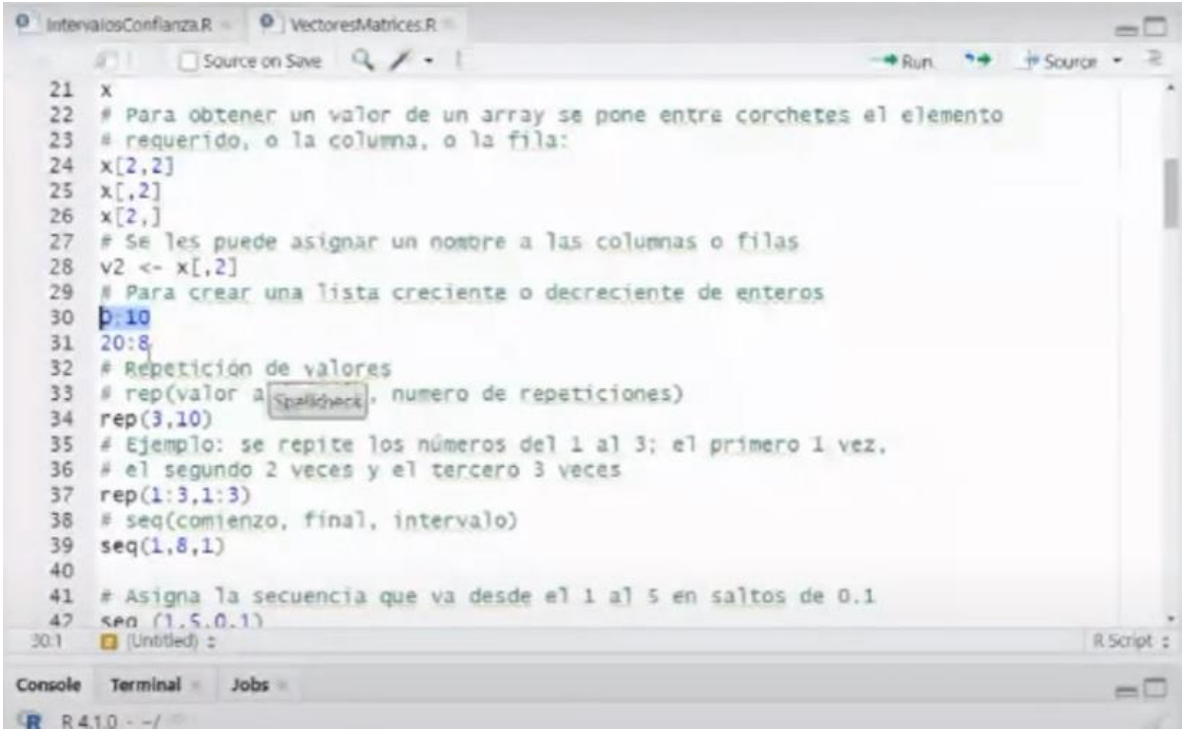
$$\Rightarrow \quad u1' = (1 \ \ 0 \ \ 0)' \qquad u2' = (0 \ \ 1 \ \ 0)' \qquad u3' = (0 \ \ 0 \ \ 2)'$$

vectores propios de la matriz A

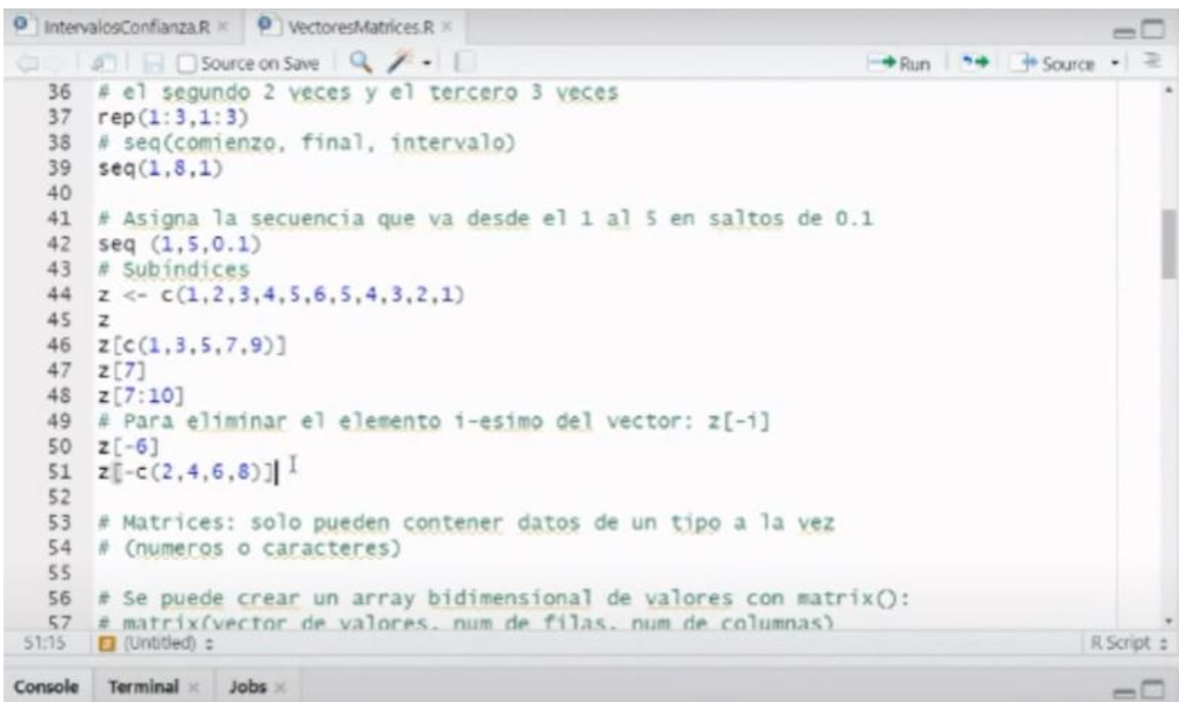
PROCEDIMIENTO EN R:




```r
# ----------------------------------------
# VECTORES Y MATRICES
# ----------------------------------------

# Un vector se puede definir por un solo símbolo y la expresión c()
x <- c(10,20,30,40)

# Si se pone x+100 se suma 100 a todos los componentes
x+100
# Se pueden anidar los vectores
x <- c(1,2,3,4,5)
xAnidados <- c(x,x,x)
xAnidados
# cbind() forma un array bidimensional combinando las columnas
c1 <- c(10,20,30,40)
c2 <- c(5,10,15,20)
x <- cbind(c1,c2)
x
# rbind() forma un array bidimensional combinando las filas
x <- rbind(c1,c2)
x
# Para obtener un valor de un array se pone entre corchetes el elemento
```



```r
x
# Para obtener un valor de un array se pone entre corchetes el elemento
# requerido, o la columna, o la fila:
x[2,2]
x[,2]
x[2,]
# Se les puede asignar un nombre a las columnas o filas
v2 <- x[,2]
# Para crear una lista creciente o decreciente de enteros
0:10
20:8
# Repeticion de valores
# rep(valor a Spellcheck, numero de repeticiones)
rep(3,10)
# Ejemplo: se repite los números del 1 al 3; el primero 1 vez,
# el segundo 2 veces y el tercero 3 veces
rep(1:3,1:3)
# seq(comienzo, final, intervalo)
seq(1,8,1)

# Asigna la secuencia que va desde el 1 al 5 en saltos de 0.1
seq (1,5,0.1)
```



```r
# el segundo 2 veces y el tercero 3 veces
rep(1:3,1:3)
# seq(comienzo, final, intervalo)
seq(1,8,1)

# Asigna la secuencia que va desde el 1 al 5 en saltos de 0.1
seq (1,5,0.1)
# Subíndices
z <- c(1,2,3,4,5,6,5,4,3,2,1)
z
z[c(1,3,5,7,9)]
z[7]
z[7:10]
# Para eliminar el elemento i-esimo del vector: z[-i]
z[-6]
z[-c(2,4,6,8)]

# Matrices: solo pueden contener datos de un tipo a la vez
# (numeros o caracteres)

# Se puede crear un array bidimensional de valores con matrix():
# matrix(vector de valores, num de filas, num de columnas)
```

```r
# Ejemplo: llenar la matriz por columnas
A <- matrix(1:12,3,4)
# Ejemplo: rellenar la matriz por filas
A <- matrix(1:12,3,4, byrow=T)
# Ejemplo: crea una matriz de 9's
matrix(9,3,4)
# Ejemplo se define la siguiente matriz por filas:
X <- matrix(c(1, -2, 3,
              4, -5, -6,
              7, 8, 9,
              0, 0, 10),
            4, 3, byrow=TRUE)
X
# Transpuesta de una matriz
t(X)

# Matriz diagonal
diag(B)

# Traza de una matriz
```

72:1 (Untitled) R Script

Console Terminal Jobs

```r
# Traza de una matriz

sum(diag(B))

# Comprobacion de que es simetrica una matriz
all(B == t(B))
C <- matrix(c(-5, 1, 3,
              1, 2, 6,
              3, 6, -4),
            3, 3, byrow=TRUE)
all(C == t(C))


# Definir una matriz diagonal
diag(c(6, -2, 0, 7))
```

87:1 (Untitled) R Script

Console Terminal Jobs

```r
# Definir una matriz identidad
diag(3)

# Definir una matriz de ceros
matrix(0, 4, 3)


# Operaciones entre matrices
A + B
A - B
-A
# Producto de un escalar por una matriz
3 * B
B * 3
# Producto escalar entre dos vectores
a <- c(2, 0, 1, 3)
b <- c(-1, 6, 0, 9)
a %*% b

# Producto de dos matrices
```

97:1 (Untitled) R Script

Console Terminal Jobs

```r
# Producto de dos matrices
A <- matrix(1:4, 2, 2, byrow=TRUE)
A
B <- matrix(c(0, 3, 2, 1), 2, 2, byrow=TRUE)
B
A %*% B
B %*% A
C <- matrix(1:6, 2, 3, byrow=TRUE)
C
I <- diag(3)
I
C %*% I
# Esto da error
```

```r
# Esto da error
I %*% C
# Inversa de una matriz
A <- matrix(c(2, 5, 1, 3), 2, 2, byrow=TRUE)
A
solve(A)
A %*% solve(A)
solve(A) %*% A

# Determinantes
A <- matrix(c(2, 5, 1, 3), 2, 2, byrow=TRUE)
det(A)

# Autovalores y Autovectores
```

```r
# Autovalores y Autovectores
A <- matrix(c(1, .5, .5, 1), 2, 2)
A
eigA <- eigen(A)
eigA
sum(eigA$values)
prod(eigA$values)
det(A)
# Rango de una matriz
# Calcular la descomposicion QR de la matriz
la.qr <- qr(A)
# Se listan los atributos del objeto anterior
names(la.qr)
# Se extrae el atributo rango
print(c("El rango de la matriz es",la.qr$rank),quote = F)
# El rango de una matriz cuadrada simetrica equivale al numero de
# autovalores distintos de 0:
autoval <- eigen(A, only.values = TRUE)
```

```r
prod(eigA$values)
det(A)
# Rango de una matriz
# Calcular la descomposicion QR de la matriz
la.qr <- qr(A)
# Se listan los atributos del objeto anterior
names(la.qr)
# Se extrae el atributo rango
print(c("El rango de la matriz es",la.qr$rank),quote = F)
# El rango de una matriz cuadrada simetrica equivale al numero de
# autovalores distintos de 0:
autoval <- eigen(A, only.values = TRUE)
rango <-length(autoval[[1]]>=1.e-10)
print(c("El rango de la matriz es",rango),quote = F)
```

Análisis de componentes principales

19/11/2021

- Es una técnica estadística multivariante de simplificación, que permite transformar un conjunto de variantes originales correlacionadas entre sí, en un conjunto sintético de variables no correlacionadas denominadas factores o componentes principales
- Es un método de reducción de dimensión
- Busca evitar información redundante
- Describe la relación entre las variables originales con respecto a las filas
- Facilitar la interpretación exploratoria de los datos y permite proponer un análisis estadística adecuada
- Técnica de ordenación que tiene por objetivo definir y resumir un conjunto de variables cuantitativas con una perdida mínima de información
- **Muchas variables:** Variables cuantitativas originales correlacionadas entre sí
- **Pocas variables:** Variables independientes ortogonales
- El primer componente principal, reúne el mayor porcentaje de la variación existente entre los datos existentes
- La segunda componente reúne la variación que no pudo ser explicada en el primer componente.
  (No. De variables = No. De componentes)

CONDICIONES PARA APLICAR EL ACP

- Los datos deben ser variables cuantitativas (cualitativos -> Escala Likert)
- El número de filas debe ser mayor al número de variables (columnas)

MATRICES UTILIZADAS ACP

- Matriz de correlación: Es utilizada cuando las variables de medida tienen unidades de medida iguales.
- Matriz de varianza: Es utilizada cuando las variables de medida tienen unidades de medida diferentes

PROCEDIMIENTO

- Obtención de los datos
- Obtención de los valores propios y la varianza
- Seleccionar los componentes principales
  a) Mayor varianza
  b) Matrices
- Analizarlos e interpretarlos

EJEMPLOS

Evaluación de 5 formulaciones de mermeladas de mora y sus efectos en las variables sensoriales: sabor, calor, aroma y textura.

¿Qué formulación de mermelada de mora presenta mejor perfil sensorial?

Mermelada	Sabor	Color	Aroma	Textura
M1	10	15	18	11
Don Serafin	16	17	24	16
M3	13	13	20	14
M4	12	20	10	25
M5	12	11	14	11

¿Cuántas componentes principales seleccionar?

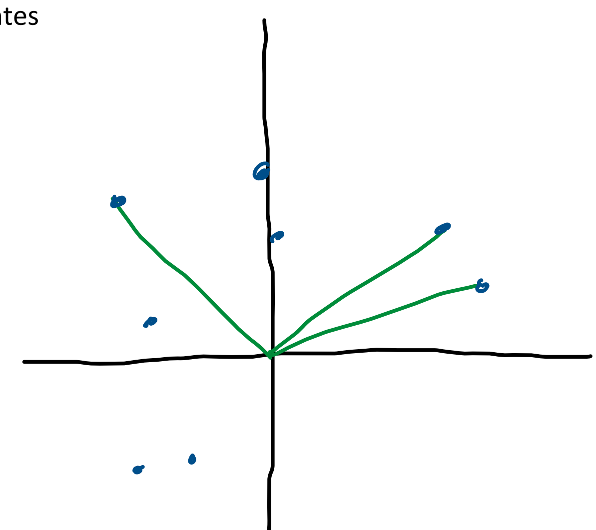
Varianza total explicada por las componentes

PC	Eligen value	% varianza
1	2.07538	51.884
2	1.56547	39.137
3	0.328326	8.2081
4	0.0308216	0.77054

} 91.02%

Matriz de correlación entre variables originales y componentes

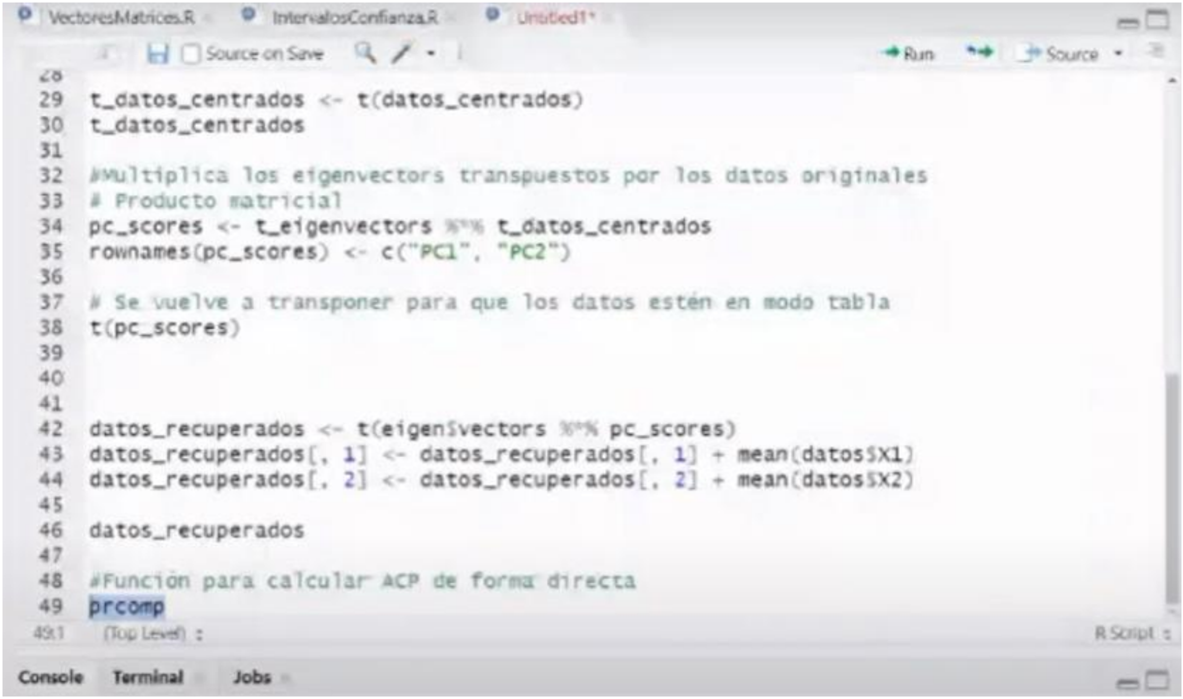
	PC1	PC2
Aroma	-0.0077965	0.7539
Color	0.61529	0.24673
Sabor	-0.4121	0.59128
Textura	0.67195	0.14545



PROCEDIMIENTO EN R:

```r
#Datos
datos <- data.frame(X1 = c(2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2, 1, 1.5, 1.1),
                    X2 = c(2.4, 0.7, 2.9, 2.2, 3.0, 2.7, 1.6, 1.1, 1.6, 0.9))

datos

#Normalización de datos
datos_centrados <- datos
datos_centrados$X1 <- datos$X1 - mean(datos$X1)
datos_centrados$X2 <- datos$X2 - mean(datos$X2)
datos_centrados

#Calculo de la matriz de correlacion
matriz_cov <- cov(datos_centrados)
matriz_cov

#Obtención de los valores propios de la matriz de covarianzas
eigen <- eigen(matriz_cov)
eigen$values

#Obtención de los vectores propios, Componentes principales ]
eigen$vectors
```

```r
eigen$values

#Obtención de los vectores propios, Componentes principales
eigen$vectors

#calcula el valor que toma cada componente para cada observación en función de las
#variables originales
t_eigenvectors <- t(eigen$vectors)
t_eigenvectors

t_datos_centrados <- t(datos_centrados)
t_datos_centrados

#Multiplica los eigenvectors transpuestos por los datos originales
# Producto matricial
pc_scores <- t_eigenvectors %*% t_datos_centrados
rownames(pc_scores) <- c("PC1", "PC2")

# Se vuelve a transponer para que los datos esten en modo tabla
t(pc_scores)
```


VectoresMatrices.R
IntervalosConfianza.R
Untitled1*
Source on Save
Run
Source
t_datos_centrados <- t(datos_centrados)
t_datos_centrados
#Multiplica los eigenvectors transpuestos por los datos originales
# Producto matricial
pc_scores <- t_eigenvectors %*% t_datos_centrados
rownames(pc_scores) <- c("PC1", "PC2")
# Se vuelve a transponer para que los datos estén en modo tabla
t(pc_scores)
datos_recuperados <- t(eigen$vectors %*% pc_scores)
datos_recuperados[, 1] <- datos_recuperados[, 1] + mean(datos$X1)
datos_recuperados[, 2] <- datos_recuperados[, 2] + mean(datos$X2)
datos_recuperados
#Función para calcular ACP de forma directa
prcomp
49:1
(Top Level)
R Script
Console
Terminal
Jobs

Análisis factorial

23/11/2021

Se leyó una lectura acerca de los papeles del Psicólogo:
https://www.redalyc.org/pdf/778/77812441003.pdf

Introducción

- El método de **Análisis factorial** fue desarrollado por **Harold Hotelling** en **1933.**
- Es el método más utilizado en investigaciones sociales y comerciales.
- Es conocido también como **factorización del eje principal.**
- El análisis factorial es un método de análisis multivariante de reducción de variables.
- Considera la varianza común de los factores.
- Genera combinaciones lineales llamada **factor.**

Análisis factorial

Método que intenta **identificar variables subyacentes** (factores), que expliquen la configuración de las **correlaciones** dentro de un conjunto de **variables observadas.**
También puede utilizarse para generar:

- Hipótesis relacionadas con los mecanismos causales
- Inspeccionar las variables para análisis subsiguientes

Consideraciones de aplicación del Análisis Factorial

- Las variables deben ser cuantitativos a nivel de *intervalo* o de *razón* - > **Datos.**
- El número de observaciones debe ser al menos 4 o 5 veces mayor que el número de variables.
- Los datos deben tener una distribución normal bivariada para cada pareja de variables y las observaciones deben de ser independientes
- Ningún factor único esta correlacionado con los demás, ni con los factores comunes

Conceptos importantes

- **Descriptivos univariados**: Para cada variable se muestra la media y desviación estándar
- **Comunalidad**: Porcentaje de varianzas de cada variable que explica el análisis factorial
- **Solución inicial**: Permite obtener las comunalidades iniciales, autovalores de la matriz analizada y los porcentajes de varianza asociados.
- **Matriz reproducida:** Es aquella que se obtiene a partir de la solución encontrada. En la diagonal de esta matriz se encuentran las comunalidades finales.
- **Media de Kaiser-Meyer-Olkin**: Permite comparar la magnitud de los coeficientes de correlación observados con la magnitud de los coeficientes de correlación Varía entre 0 y 1 Los valores KMO<0.5 indica que el análisis factorial no es adecuado realizarse
- **Prueba de esfericidad de Barlett**: Contrasta la hipótesis nula de que la matriz de correlaciones es una matriz identidad, en donde no existirán correlaciones significativas entre las variables y por tanto el método de análisis factorial no es adecuado.
- **Grafico de sedimentación**: Representación grafica de la magnitud de los valores. Es utilizado para determinar la cantidad optima de factores que deben presentarse en la solución

Procedimiento para el Análisis Factorial

1. Plantear el problema
2. Obtener matriz de correlación
3. Determinar el método de análisis factorial
4. Determinar el número de factores
5. Rotar los factores
6. Calcular las puntuaciones de los factores
7. Elegir variables sustitutas
8. Determinar el ajuste del modelo

Plantear el problema y obtener la matriz de correlación

Plantear el problema

- Establecer los objetivos del análisis factorial
- Especificar variables dependientes e independientes
- Obtener los datos

Matriz de correlación

- Para que el análisis sea adecuado, las variables tienen que estar correlacionadas
  Prueba de esfericidad de Bartlett

Determinar el método de análisis factorial

- El procedimiento de análisis ofrece un alto de grado de flexibilidad

Métodos de análisis factorial

- **Mínimos cuadrados no ponderados**: Genera una matriz de pesos que minimiza las sumas de los cuadrados de las diferencias entre las matrices de correlaciones, observada y producida, ignorando los elementos de la diagonal

- **Mínimos cuadrados generalizados:** Criterio igual al anterior, pondera los coeficientes de correlación inversa a la unicidad de las variables. Las variables con alta unicidad (baja comunalidad) tienen una influencia pequeña en el resultado final.

- **Máxima verosimilitud:** Calcula las estimaciones de los parámetros que con mayor probabilidad han producido la matriz de correlaciones observada. Genera un estadístico de bondad de ajuste que contrasta el grado de ajuste entre lo real y lo estimado

- **Análisis alfa:** Considera las variables incluidas en el análisis como una muestra al conjunto posible variable, intentado maximizar la fiabilidad de los factores respecto a la totalidad de las variables

- **Análisis de imágenes:** Mediante la correlación múltiple, estudia las partes comunes únicas de las variables observadas
  Imagen de variable: Parte común que tiene una variable con otras
  Anti imagen: Parte exclusiva de cada variable

Determinar el número de factores

- **A priori**: El investigador determina el número de componentes
- **Valor propio:** Cuando el valor propio es mayor a 1
- **Grafica de sedimentación:** Se visualiza las pendientes pronunciadas de factores de valores grandes
- **Otros:** Porcentaje de varianza, confiabilidad de división de mitades y pruebas de significancia

Rotar los factores

- La rotación transforma a la matriz factorial a matriz de patrones factoriales en una matriz más sencilla y fácil de interpretar
- La rotación redistribuye la varianza explicada por factores individuales
- Con la rotación de ejes solo cambia la varianza explicada. Las comunalidades y el porcentaje de varianza quedan inalterable (el mismo)

Métodos para realizar rotaciones

Métodos ortogonales

Gira los ejes ortogonalmente, en el mismo eje. Se utiliza cuando no existe relación entre los factores

- **Varimax:** Minimiza el numero de variables que tienen saturaciones altas en cada factor. Es el método de rotación más utilizado.
- **Quartimax**: Minimiza el numero de factores necesarios para explicar cada variable, concentrando la mayor parte de la varianza de cada factor y dejando próximas a 0 el resto de saturaciones.
- **Equamax**: Combinación del método Varimax que simplifica los factores y el método Quartimax que simplifica las variables

Métodos oblicuos

Considera que dos factores pueden explicar una misma realidad, que existe correlación entre factores y cada eje podría girar en un ángulo diferente.

- **Oblim directo**: Cuando delta es igual a 0. Cuando delta se va haciendo más negativos, los factores son menos oblicuos
- **Promax**: Se calcula más rápido que una rotación oblim directa. Es útil para un gran conjunto de datos.

Interpretar los resultados

- Consiste en identificar las variables que tienen cargas altas sobre el mismo factor, el cual puede interpretarse en términos de las variables que tienen mayor carga
- Se puede utilizar una gráfica de sedimentación en donde se empleen las cargas factoriales.

Calcular las puntuaciones

- El análisis factorial tiene su propio valor independiente.
- Si el objetivo es reducir las variables a un conjunto menos de variables compuestas (factores)
- Para utilizarlas en análisis posteriores.

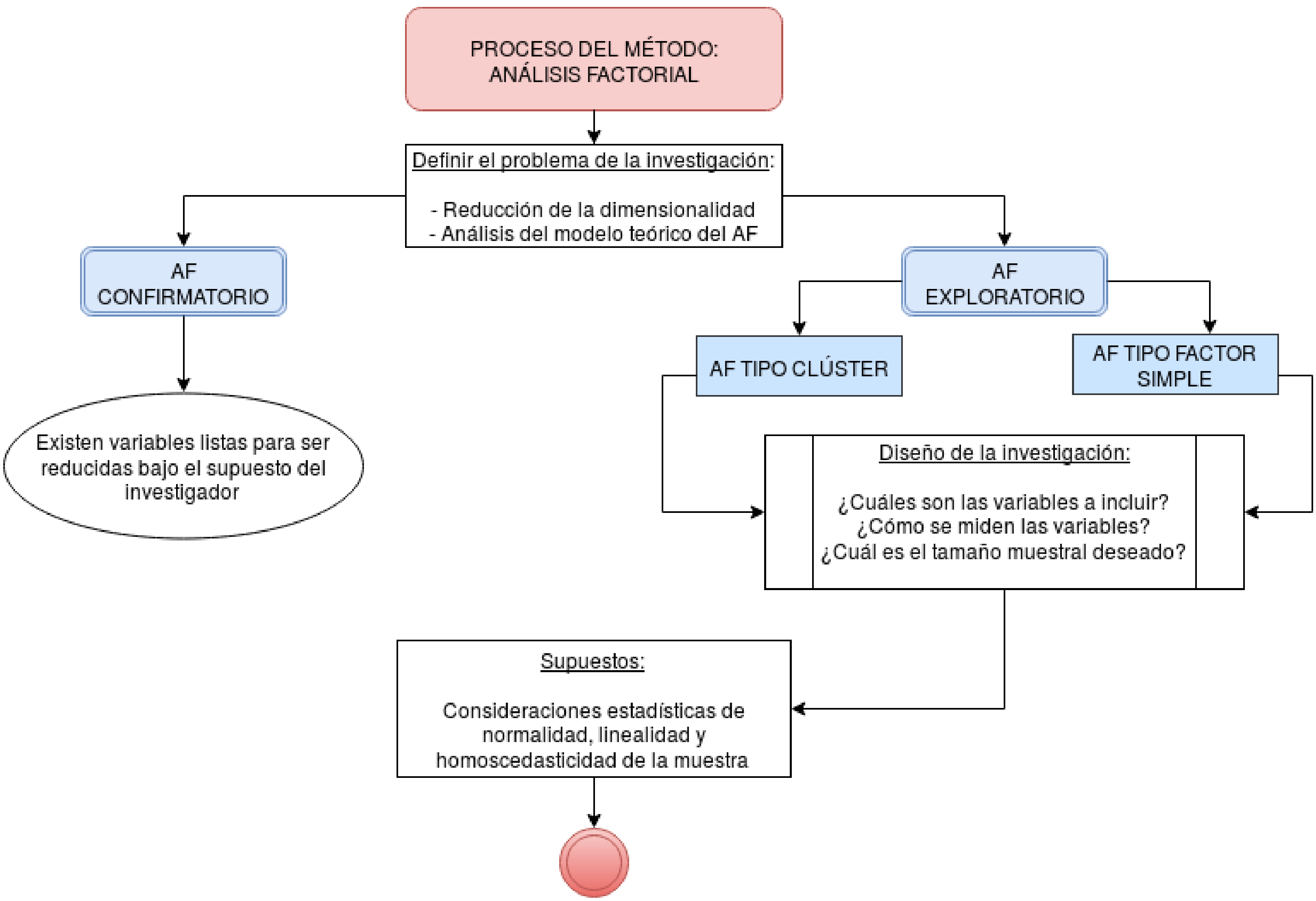
Elegir variables sustitutas

- Consiste en separar algunas variables originales para utilizarlas en análisis siguientes
- Para elegirse, se deben elegir para cada factor las variables con la carga mas alta para utilizarla como variable sustituta.

Determinar ajustes modelo

- Se examina las diferencias entre las correlaciones observadas y correlaciones reproducidas -> **Diferencia residual.**
- Si existen muchos residuos altos, el modelo no proporciona un buen ajuste para los datos

Diagrama de decisión – análisis factorial



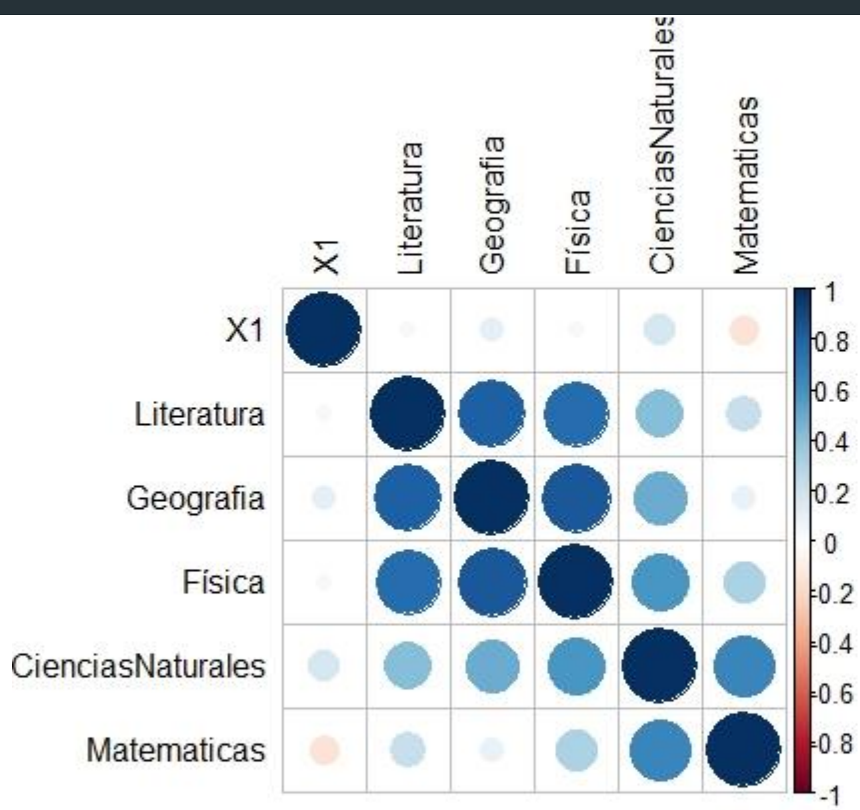
Se ejemplifico con el siguiente link en R
http://www.rubenjoserodriguez.com.ar/wp-content/uploads/2015/04/An%23U00e1lisis_Factorial_Test_-y_Escalas_Pedro_Morales_Vallejo.pdf
https://rpubs.com/marcelo-chavez/multivariado_1

PROCEDIMIENTO EN R

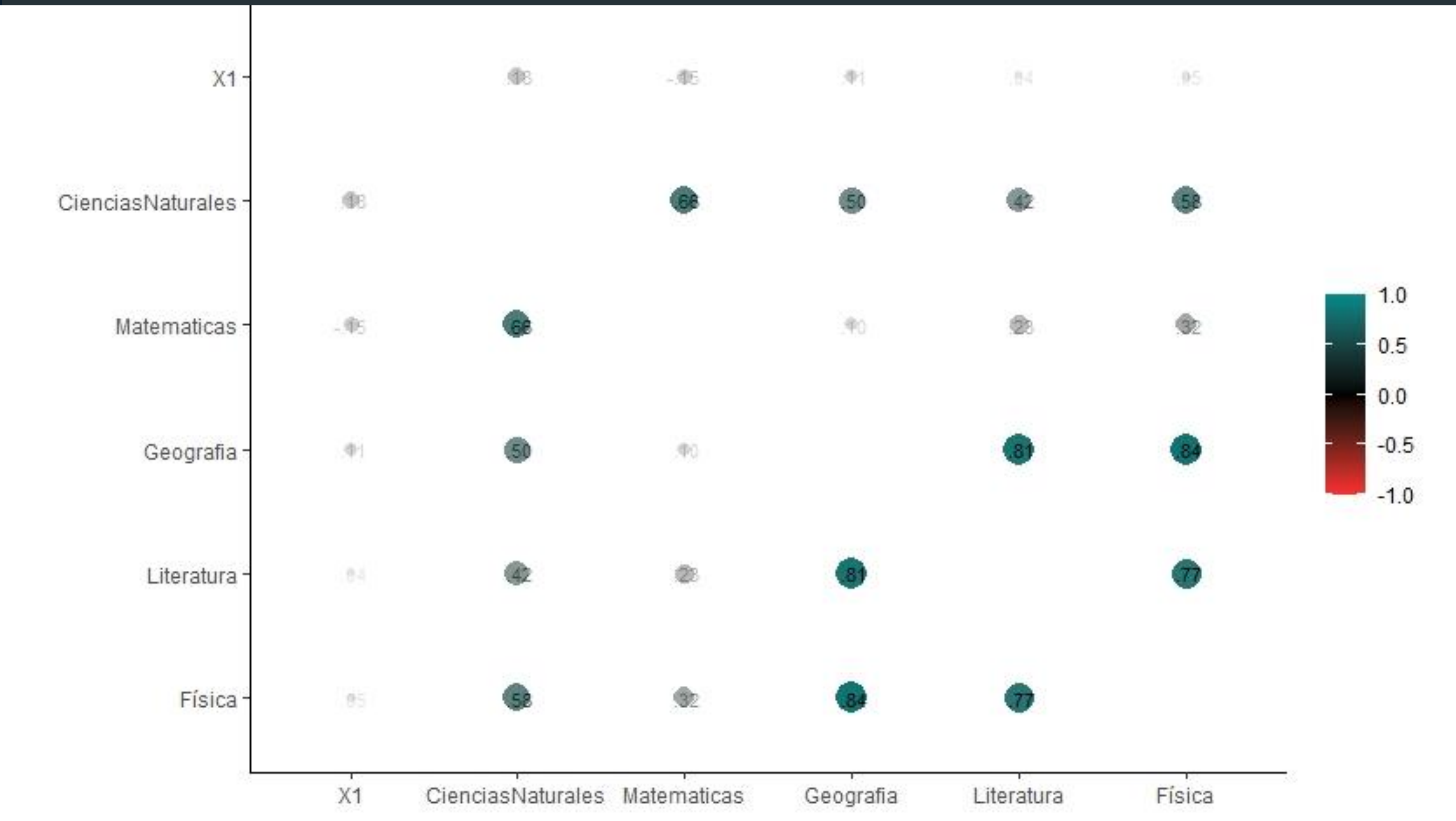
```r
# Análisis Factorial
# Ejemplo Clase
# 26/Noviembre/2021

# Definimos las librerias
library(openxlsx) #Librería que interactúa con MSExcel
library(corrplot) #Librería para el gráfico de correlaciones
library(corrr) #Otra opción de librería para el cálculo y gráfico de correlaciones
library(psych)
library(stats) #Librería del sistema base



estudiantes  <- read.xlsx(xlsxFile='C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/af.xlsx',
                          sheet = 'Hoja1') # Lectura de la BDD de acuerdo a su ubicación
estudiantes#Visualización de la tabla

matriz_correlaciones <- cor(estudiantes, use = "pairwise.complete.obs")
matriz_correlaciones

corrplot(cor(estudiantes), order = "hclust", tl.col='black', tl.cex=1) #Gráfico de las correlaciones

estudiantes_correlaciones <- correlate(estudiantes)  #Cálculo de un objeto de correlaciones
rplot(estudiantes_correlaciones, legend = TRUE, colours = c("firebrick1", "black",
                                                            "darkcyan"), print_cor = TRUE)  #Opción gráfica de las correlaciones

# Se puede conocere la presencia de multicolinealidad al evaluar la
# Determinante de la matriz de correlaciones de las variables ingresadas al
# estudio:
det(matriz_correlaciones)

#Por consiguiente vamos a ubicar el cálculo de los estimadores del Test de Bartlett y el MSA(KMO):
bartlett.test(estudiantes)

KMO(estudiantes)

factanal(estudiantes, factors = 2, rotation = "none")

factanal(estudiantes, factors = 2, rotation = "none", scores = "regression")$scores

puntuaciones <- factanal(estudiantes, factors = 2, rotation = "none", scores = "regression")$scores
estudiantes <- cbind(estudiantes, puntuaciones)
estudiantes$Factor1 <- round(((estudiantes$Factor1 - min(estudiantes$Factor1))/(max(estudiantes$Factor1) - min(estudiantes$Factor1))), 2)
estudiantes

hist(estudiantes$Factor1, freq = TRUE, main = "Gráfico de la Distribución del Factor 1",
     xlab = "Factor 1", ylab = "Frecuencia", col = "#009ACD")

```

RESULTADOS

```
> library(openxlsx) #Librería que interactúa con MSExcel
> library(corrplot) #Librería para el gráfico de correlaciones
> library(corrr) #Otra opción de librería para el cálculo y gráfico de correlaciones
> library(psych)
> library(stats) #Librería del sistema base
> estudiantes  <- read.xlsx(xlsxFile='C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/af.xlsx',
+                           sheet = 'Hoja1') # Lectura de la BDD de acuerdo a su ubicación
> estudiantes
   X1 CienciasNaturales Matematicas Geografia Literatura Física
1   1                 7           7         5          5      6
2   2                 5           5         6          6      5
3   3                 5           6         5          7      5
4   4                 6           8         5          6      6
5   5                 7           6         6          7      6
6   6                 4           4         6          7      6
7   7                 5           5         5          5      6
8   8                 5           6         5          5      5
9   9                 6           5         7          6      6
10 10                 6           5         6          6      6
11 11                 6           7         5          6      5
12 12                 5           5         4          5      4
13 13                 6           6         6          6      5
14 14                 8           7         8          8      8
15 15                 6           7         5          6      6
16 16                 4           3         4          4      4
17 17                 6           4         7          8      7
18 18                 6           6         7          7      7
19 19                 6           5         4          4      4
20 20                 7           7         6          7      6
> matriz_correlaciones <- cor(estudiantes, use = "pairwise.complete.obs")
> matriz_correlaciones
                           X1 CienciasNaturales Matematicas  Geografia Literatura     Física
X1                 1.00000000         0.1769981 -0.14818526 0.10557598 0.04269889 0.04704852
CienciasNaturales  0.17699808         1.0000000  0.65614980 0.49706742 0.42034118 0.58398469
Matematicas       -0.14818526         0.6561498  1.00000000 0.09908375 0.22951010 0.31711567
Geografia          0.10557598         0.4970674  0.09908375 1.00000000 0.81339038 0.84081035
Literatura         0.04269889         0.4203412  0.22951010 0.81339038 1.00000000 0.76622848
Física             0.04704852         0.5839847  0.31711567 0.84081035 0.76622848 1.00000000
> corrplot(cor(estudiantes), order = "hclust", tl.col='black', tl.cex=1)
> 
```



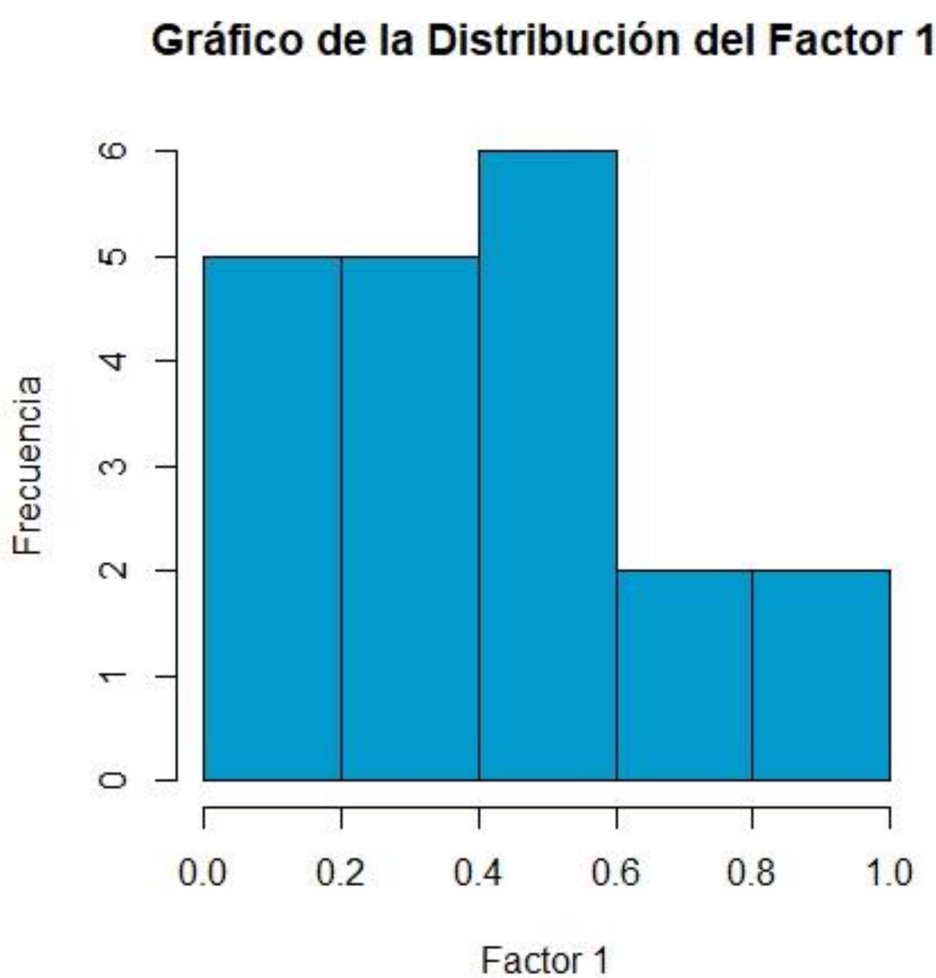
```
> estudiantes_correlaciones <- correlate(estudiantes)

Correlation method: 'pearson'
Missing treated using: 'pairwise.complete.obs'

> rplot(estudiantes_correlaciones, legend = TRUE, colours = c("firebrick1", "black",
+                                                             "darkcyan"), print_cor = TRUE)
Don't know how to automatically pick scale for object of type noquote. Defaulting to continuous.
Don't know how to automatically pick scale for object of type noquote. Defaulting to continuous.
>
```




```
> det(matriz_correlaciones)
[1] 0.02146713
> bartlett.test(estudiantes)

        Bartlett test of homogeneity of variances

data:  estudiantes
Bartlett's K-squared = 130.27, df = 5, p-value < 2.2e-16

> KMO(estudiantes)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = estudiantes)
Overall MSA =  0.63
MSA for each item =
               X1 CienciasNaturales       Matematicas         Geografia        Literatura            Física
             0.20              0.59              0.39              0.63              0.76              0.82
```

```
> factanal(estudiantes, factors = 2, rotation = "none")

Call:
factanal(x = estudiantes, factors = 2, rotation = "none")

Uniquenesses:
               X1 CienciasNaturales       Matematicas         Geografia        Literatura            Física
            0.962             0.377             0.005             0.041             0.291             0.204

Loadings:
                  Factor1 Factor2
X1                 0.133  -0.143
CienciasNaturales  0.417   0.670
Matematicas                0.997
Geografia          0.971   0.128
Literatura         0.803   0.253
Física             0.824   0.342

               Factor1 Factor2
SS loadings      2.459   1.661
Proportion Var   0.410   0.277
Cumulative Var   0.410   0.687

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 3.39 on 4 degrees of freedom.
The p-value is 0.494
> factanal(estudiantes, factors = 2, rotation = "none", scores = "regression")$scores
      Factor1    Factor2
1  -0.6314078  1.0189422
2   0.2692355 -0.5532315
3  -0.5176998  0.2186363
4  -0.7390925  1.7957397
5   0.4013216  0.2577276
6   0.5998192 -1.3328774
7  -0.3922924 -0.5635534
8  -0.6750642  0.2104977
9   1.1554370 -0.5234933
10  0.4418072 -0.5375273
11 -0.7093113  1.0047577
12 -1.3620321 -0.5940208
13  0.1605950  0.2362751
14  2.0598016  1.0949966
15 -0.5765776  1.0122282
16 -1.1683892 -2.1712948
17  1.6064904 -1.2911470
18  1.2187644  0.2692572
19 -1.4004864 -0.5899185
20  0.2590815  1.0380056
>
```

```
> puntuaciones <- factanal(estudiantes, factors = 2, rotation = "none", scores = "regression")$scores
> estudiantes <- cbind(estudiantes, puntuaciones)
> estudiantes$Factor1 <- round(((estudiantes$Factor1 - min(estudiantes$Factor1))/(max(estudiantes$Factor1) - min(estudiantes$Factor1))), 2)
> estudiantes$Factor1 <- round(((estudiantes$Factor1 - min(estudiantes$Factor1))/(max(estudiantes$Factor1) - min(estudiantes$Factor1))), 2)
> estudiantes
   X1 CienciasNaturales Matematicas Geografia Literatura Física Factor1    Factor2
1   1                 7           7         5          5      6    0.22  1.0189422
2   2                 5           5         6          6      5    0.48 -0.5532315
3   3                 5           6         5          7      5    0.26  0.2186363
4   4                 6           8         5          6      6    0.19  1.7957397
5   5                 7           6         6          7      6    0.52  0.2577276
6   6                 4           4         6          7      6    0.58 -1.3328774
7   7                 5           5         5          5      6    0.29 -0.5635534
8   8                 5           6         5          5      5    0.21  0.2104977
9   9                 6           5         7          6      6    0.74 -0.5234933
10 10                 6           5         6          6      6    0.53 -0.5375273
11 11                 6           7         5          6      5    0.20  1.0047577
12 12                 5           5         4          5      4    0.01 -0.5940208
13 13                 6           6         6          6      5    0.45  0.2362751
14 14                 8           7         8          8      8    1.00  1.0949966
15 15                 6           7         5          6      6    0.24  1.0122282
16 16                 4           3         4          4      4    0.07 -2.1712948
17 17                 6           4         7          8      7    0.87 -1.2911470
18 18                 6           6         7          7      7    0.76  0.2692572
19 19                 6           5         4          4      4    0.00 -0.5899185
20 20                 7           7         6          7      6    0.48  1.0380056
> hist(estudiantes$Factor1, freq = TRUE, main = "Gráfico de la Distribución del Factor 1",
+       xlab = "Factor 1", ylab = "Frecuencia", col = "#009ACD")
>
```



Análisis de conglomerados 1/12/2021

Análisis Multivariante

- Dependientes
- Independientes
  - ACP
  - Análisis factorial
  - **Análisis de Clúster**
  - Análisis de correspondencia
  - Multidimensional scaling

ANÁLISIS DE CLUSTER

- También llamado análisis de conglomerados
- Es una técnica de análisis multivariante que permite formar grupos (conglomerados o cluster)
- Herramienta de exploración de datos que se complementa con técnicas de visualización de los mismos (Jain & Dubes, 1988)
- Los grupos deben de ser lo más homogéneos al interior y heterogéneos fuera de sí
- El agrupamiento realizado se basa en métricas de distancia o medidas de similitud
- El análisis de cluster es la base para poder realizar otros estudios: Clasificación

DIFERENCIA ENTRE ARUPAMIENTO Y CLASIFICACIÓN

APLICACIONES

Área	*Aplicación*
Astronomía	Agrupamiento de galaxias
Mercadotecnia	Segmentación de mercado, investigación de mercado
Psicología	Tipos de personalidad
Biología	Taxonomía de seres vivos
Ciencias Ambientales	Agrupamiento de ríos
Sociología	Tipos de sociedades

PROCESO PARA EL ANÁLISIS DE CLÚSTER

1. Selección de variables
2. Detección de valores atípicos
3. Seleccionar la medida de distancia o similitud
4. Elegir el algoritmo de agrupamiento
5. Obtener los resultados
6. Valoración de resultados

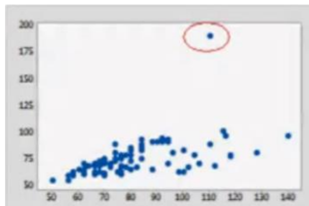
Seleccionar variables importantes

1.- SELECCIÓN DE VARIABLES

- Introducir variables irrelevantes aumenta la posibilidad de errores.
- Hay que utilizar algún criterio de selección:
  - Seleccionar variables que caracterizan a los objetos que se van a agrupar
  - Si el número de variables es muy grande aplicar un método de reducción de dimensiones

2.- DETECCIÓN DE VALORES ATÍPICOS

- El análisis de clúster es muy sensible a la presencia de objetos muy diferentes del resto (valores atípicos)



3.-SELECCIONAR LA MEDIDA DE SIMILITUD

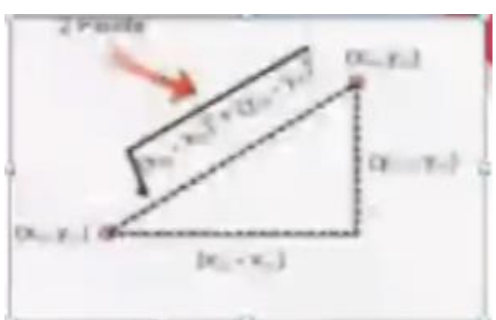
Distancia: Para datos cuantitativos

- Seleccionar la forma de medir la distancia/disimilitud entre objetos dependiendo de si los datos son cuantitativos o cualitativos

Medidas de similitud

- Métricas de distancia
  - ✓ Euclidiana: Es la distancia en línea recta o la trayectoria más corta posible entre dos puntos:
    a) Es necesario normalizar los datos antes de aplicarla
    b) Es útil cuando se tienen datos de baja dimensión y es importante medir la magnitud de los vectores
    c) También cuando se tienen columnas con ceros
    d) Los algoritmos KNN y HDBSCAN han mostrado buenos resultados

$$D_{Euc}(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}$$

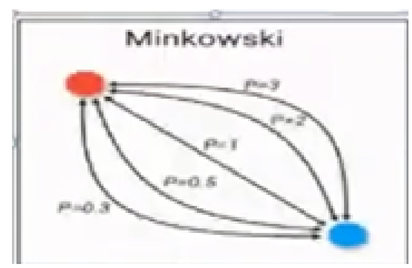


- ✓ Chebyshev: Es la mayor diferencia entre dos vectores a lo largo de cualquier dimensión de coordenadas
  a) Se utiliza para casos muy específicos
  b) Se utiliza para extraer el mínimo de movimientos que necesita un rey para ir de una casilla a otra
  c) Tiene su aplicación en la logística que se realiza en un almacén

$$D(x, y) = \max_i \left(|x_i - y_i|\right)$$

- ✓ Manhattan
- ✓ Minkowsky: Se puede utilizar en un espacio donde las distancias se pueden representar como un vector que tiene una longitud
  a) Es una medida más compleja que la Euclidiana, Chebyshev y Manhattan
  b) Recibe tres parámetros vector factor escalar y desigualdad triangular

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$



- ✓ Coseno: Es el coseno del ángulo entre dos vectores
  a) No considera la magnitud de los vectores, sino la dirección
  b) Se utiliza para datos que poseen una alta dimensionalidad
  c) Se utiliza para el análisis de texto (frecuencia de palabras)

$$D(x, y) = cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

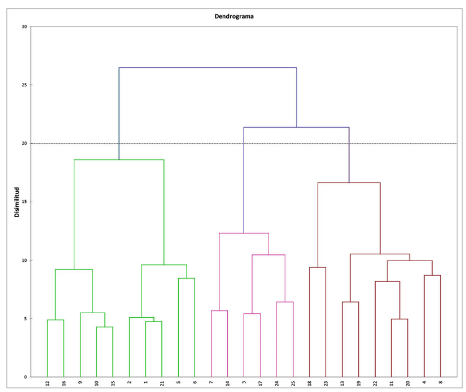
- Índices de similaridad

4.- ELEGIR EL ALGORITMO DE AGRUPAMIENTO

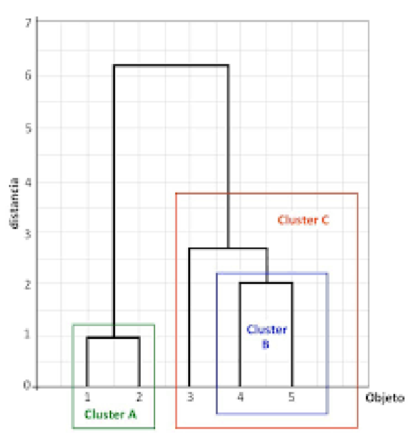
- Elegir el algoritmo para la formación de clúster (Procedimientos jerárquicos o procedimientos no jerárquicos)
  - Jerárquicos:
    a) Los grupos están anidados
    b) La agrupación final tiene un conjunto de grupos crecientes
  - No jerárquicos:
    a) Comienzan con una solución inicial, un numero de grupos **g** fijado y agrupa los objetos para obtener los **g** grupos

TIPOS DE ANÁLISIS DE CLÚSTER

- Jerárquicos:
  - Aglomerativos: Comienzan con tantos grupos como tantos objetos se tengan y en cada paso se recalculan las distancias entre los grupos existentes y se unen los dos grupos más similares. El algoritmo acaba con un clúster conteniendo todos los elementos



  - Divisivos: Comienzan con un grupo que engloba a todos los elementos y en cada paso se divide el grupo más heterogéneo. El algoritmo acaba con tantos grupos (de un elemento cada uno) como objetos se hayan agrupado



6.- VALORACIÓN DE RESULTADOS

- Comprobar que el modelo no ha definido clúster con un solo objeto o clúster con tamaños desiguales.
- Validar la calidad de los grupos obtenidos (Índices de Dunn y Davies-Boulding)

ALGORITMO JERÁRQUICO

- Algoritmo que permite trabajar con variables mixtas (cuantitativas, binarias o frecuencias)
- Se utiliza cuando no se conoce el número de grupos
- Cuando el número de objetos o individuos no es muy grande

MÉTODOS AGLOMERATIVOS

- **Enlace simple o vecino próximo:** Mide la proximidad entre dos grupos calculando la distancia entre sus objetos más próximos o la similitud entre sus objetos más semejantes.
- **Enlace completo o vecino más alejado:** Mide la proximidad entre dos grupos calculando la distancia entre sus objetos más lejanos o la similitud entre sus objetos menos semejantes.
- **Enlace medio inter grupo:** mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos.
- **Controlado y de la mediana:** Ambos métodos mide la proximidad entre dos grupos calculando la distancia entre sus centroides -> Método Ward

COMPARACIÓN DE LOS MÉTODOS AGLOMERATIVOS

Método Aglomerativo	**Resultados/Uso**
Enlace simple	Grupos encadenados
Enlace completo	Grupos compactos Menos sensible a outiliers que el enlace simple
Método Ward y enlace medio	Menos sensibles a outiliers
Método Ward	Grupos más compactos de igual tamaño

ALGORITMO

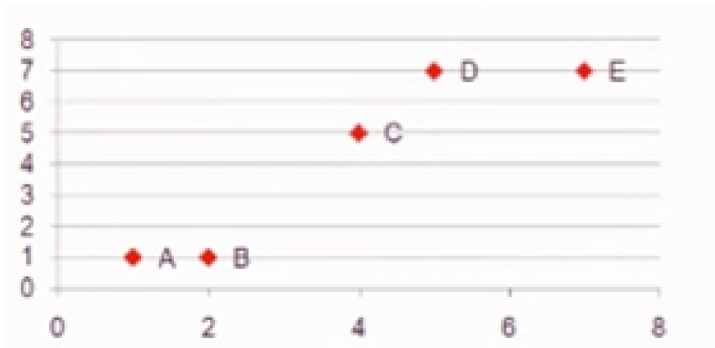
1. Inicio
2. Selección las variables
3. Detectar valores atípicos
4. Elegir una medida de similitud entre objetos -> Matriz de distancias
5. Buscar los grupos similares
6. Unir los grupos en un nuevo grupo
7. Calcular la distancia entre grupos
8. Repetir el paso 5
9. Fin

Ejemplo

Aplique el clustering jerárquico aglomerativo a los siguientes datos:

Individuo	X1	X2
A	**1**	**1**
B	**2**	**1**
C	**4**	**5**
D	**7**	**7**
e	**5**	**7**

Paso 2 y 3.- Seleccionar variables y detectar valores atípicos



Paso 4.- Aplicar la medida de distancia de similaridad

Paso 5.- Conformar el primer grupo con A y B

$$D_{Euc}(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}$$

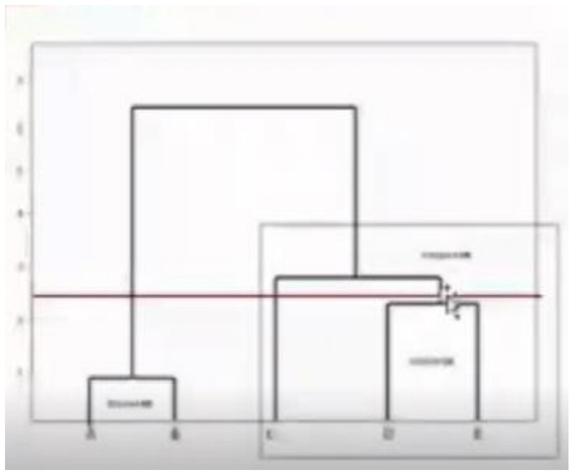
	A	B	C	D	E
A	**0**				
B	**1**	0			
C	5	4.5	0		
D	8.5	7.8	3.6	0	
E	7.2	6.7	2.2	2	0

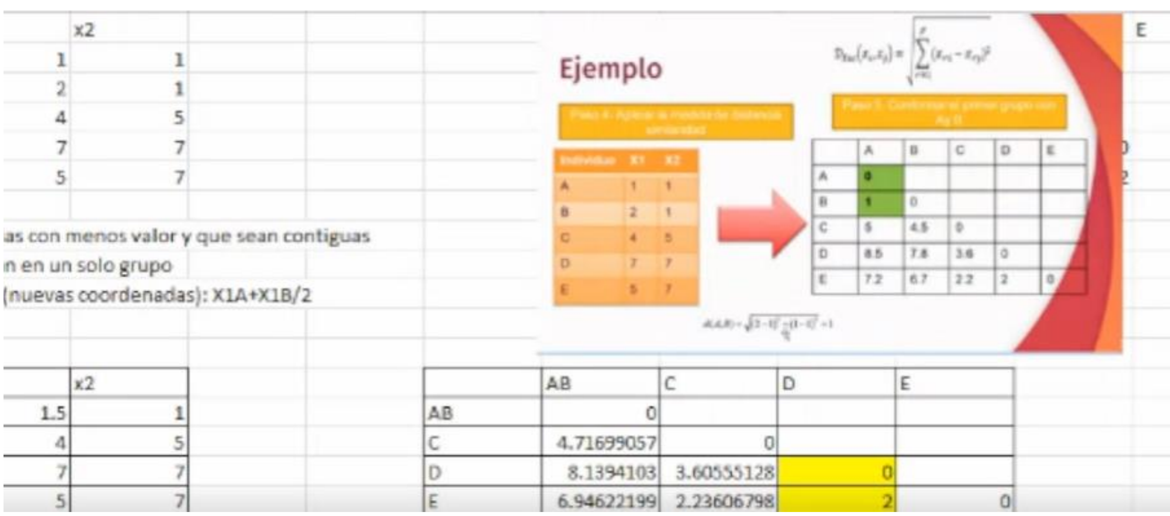
$$d(A,B) = \sqrt{(2-1)^2 + (1-1)^2} = 1$$

Paso 6.- Unir los individuos en un nuevo grupo

Individuo	X1	X2
AB	(1+2)/2 = 1.5	(1+1)/2= 1
C	4	5
D	7	7
E	5	7

Repetir desde el paso 4 hasta agrupar todos los individuos en un solo grupo





CONCLUSIÓN DEL ALGORITMO JERÁRQUICO

- El algoritmo jerárquico es conveniente cuando se tienen pocos datos.
- El número de grupos depende de la línea de corte del dendograma.
- Se recomienda confrontar los resultados aplicando otros algoritmos de agrupamiento.

CLUSTERING K-MEDIA

ALGORITMO K-MEDIAS

- Conocido como K-means
- Es un algoritmo de agrupamiento no jerárquico -> **aglomerativo**

CONSISTE EN:

- Se basa en medidas de distancias entre ellos en un conjunto de variables cuantitativas
- Asigna los individuos (casos) a un número fijo de grupos (clusters)
- Objeto -> maximizar la homogeneidad dentro de los grupos

CARACTERÍSTICAS:

- Permite agrupar un gran número de individuos
- Es sencillo (fácil de programar)
- Es uno de los algoritmos más utilizado
- Brinda resultados aceptables
- Se tiene que especificar el número de grupos (K)
- Se puede especificar el número de centroides de los grupos
- Es un algoritmo iterativo, busca:
  - Centroide de los K grupos
  - Asigna el individuo a un solo clúster

ALGORITMO

1.- Inicio

2.- Tomar al azar los k clúster iniciales y se calculan los centroides (medias) de los grupos

3.- Aplicar alguna medida de similaridad y calcular los centroides de los nuevos grupos

4.- Repetir los pasos 2 y 3 hasta que no se produzca reasignación

5.- Fin repetir

6.- FIN

EJEMPLO

Dado los siguientes datos aplique el algoritmo K-MEANS para agrupar los datos, considere K=2

IND	**X1**	**X2**
A	12	34
B	15	18
C	20	5
D	48	12

Paso 2.- Tomar al azar los K clúster iniciales y se calculan los centroides (medias) de los grupos

Grupo AB		Grupo CD	
X1	X2	X1	X2
(12+15)/2=13.5	(34+18)/2=26	(20+48)/2=34	(5+12)/2=8.5

Paso 3.- Aplicar alguna medida de similaridad y calcular los centroides de los nuevos grupos

Matriz de distancias

Distancia A-Grupo AB	Distancia A-Grupo CD
$d_{AAB} = \sqrt{(12-13.5)^2 + (34-26)^2} = 8.139$	$d_{ACD} = \sqrt{(12-34)^2 + (34-8.5)^2} = 33.678$
$d_{AAB} = \sqrt{(15-13.5)^2 + (18-26)^2} = 8.139$	$d_{AAB} = \sqrt{(15-34)^2 + (18-8.5)^2} = 21.24$

- El individuo A y B se mantienen en el grupo AB, pues tienen la menos distancia con los centroides de este grupo
- Si se reasignan individuos se calculan los nuevos centros

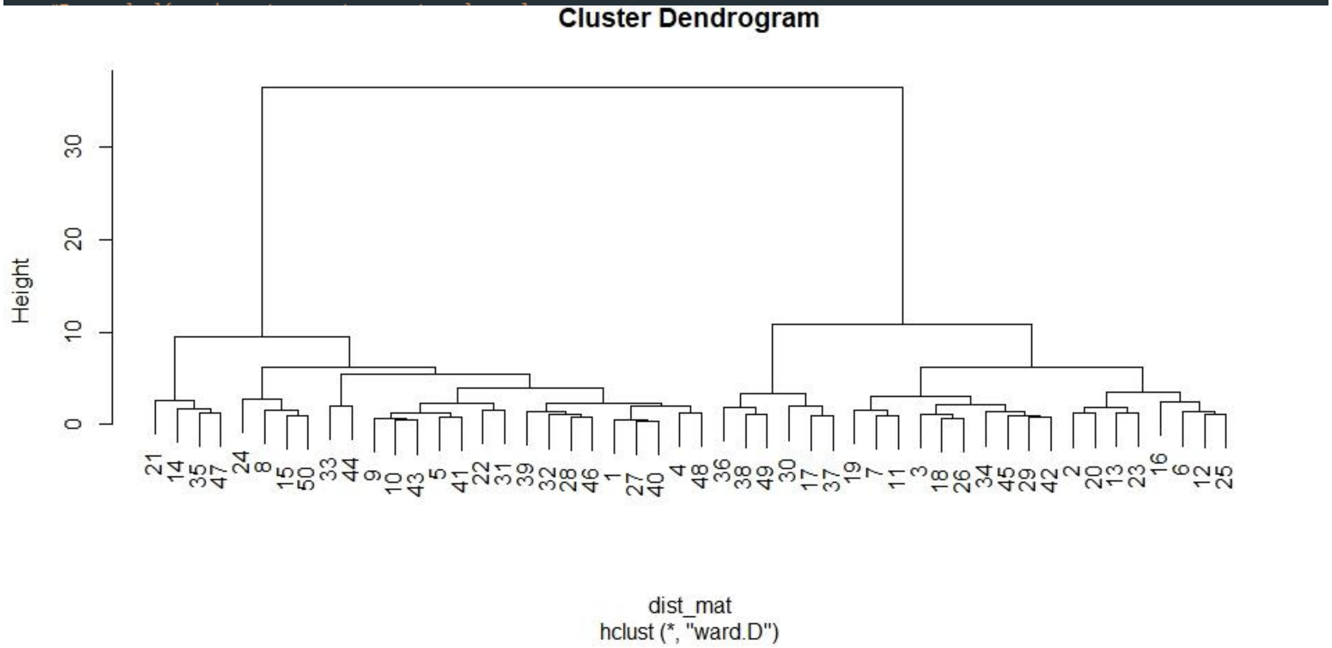
PROCEDIMIENTO EN R

```r
# Ejemplo de Clústering Jerárquico
# 03/12/2021

library(cluster)
principal <- function()
{

# Cargar los datos del archivo
textura_comidas <- read_csv("C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/food-texture.csv")
view(textura_comidas)
textura_comidas <- textura_comidas[, -1]

# Exlorar los datos
# Desplegar la estructura de los datos
str (textura_comidas)

# Obtención de medidas estadísticas
summary(textura_comidas)

# Valores NA
any(is.na(textura_comidas))

# Convertir los datos a un DataFrame
texComida <- as.data.frame(scale(textura_comidas))

# Obtención de medidas estadísticas del DF
summary(texComida)

# Se aplica la medida de distancia para obtener la matriz distancia
dist_mat <- dist(texComida, method = 'euclidean')

# Obtención de los grupos
grupos <- hclust(dist_mat, method = "ward.D")
plot(grupos)

#Trazar la línea de corte y mostrar rectangulo en los grupos
lineaCorte <- cutree(grupos, k=3)

 plot(grupos)
 rect.hclust(grupos, k = 3, border = 2:6)
 abline(h = 3, col = 'red')

}
```

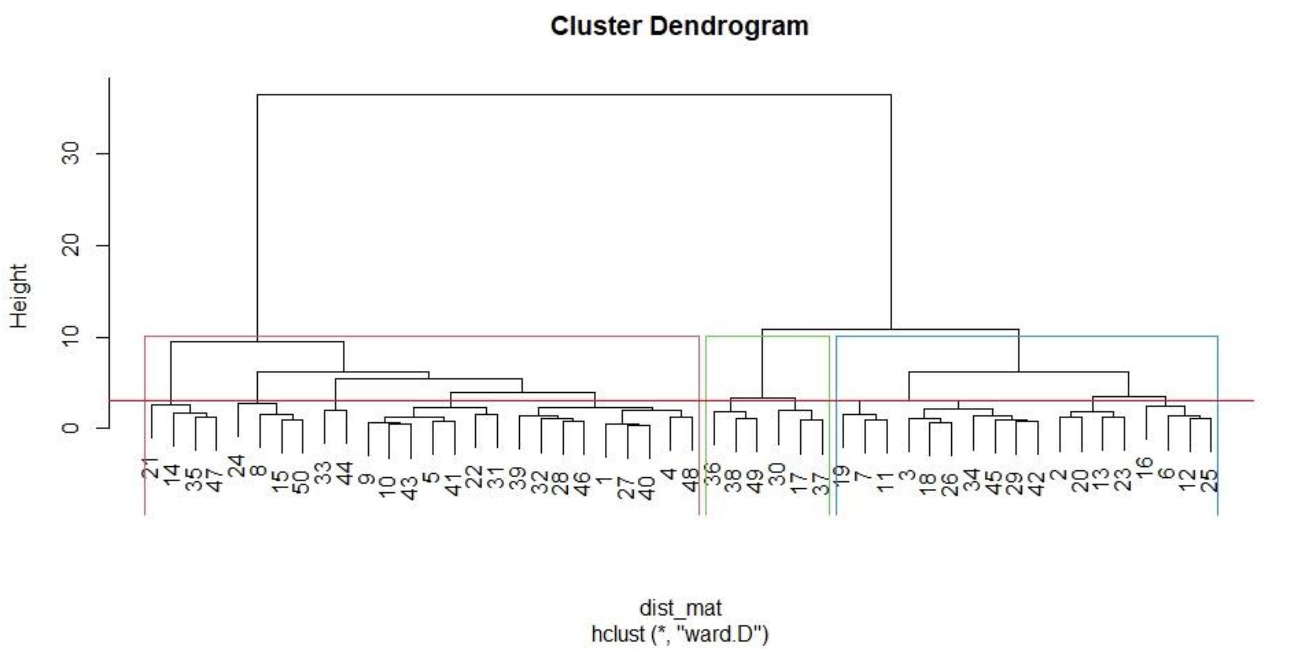
EJECUCIÓN

```
>  # Cargar los datos del archivo
>  textura_comidas <- read.csv("C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/food-texture.csv")
>  View(textura_comidas)
>  textura_comidas <- textura_comidas[, -1]
>  # Exlorar los datos
>  # Desplegar la estructura de los datos
>  str (textura_comidas)
'data.frame':	50 obs. of  5 variables:
 $ Oil     : num  16.5 17.7 16.2 16.7 16.3 19.1 18.4 17.5 15.7 16.4 ...
 $ Density : int  2955 2660 2870 2920 2975 2790 2750 2770 2955 2945 ...
 $ Crispy  : int  10 14 12 10 11 13 13 10 11 11 ...
 $ Fracture: int  23 9 17 31 26 16 17 26 23 24 ...
 $ Hardness: int  97 139 143 95 143 189 114 63 123 132 ...
>  # Obtención de medidas estadísticas
>  summary(textura_comidas)
      Oil           Density         Crispy         Fracture        Hardness
 Min.   :13.7   Min.   :2570   Min.   : 7.00   Min.   : 9.00   Min.   : 63.0
 1st Qu.:16.3   1st Qu.:2772   1st Qu.:10.00   1st Qu.:17.00   1st Qu.:107.2
 Median :16.9   Median :2868   Median :12.00   Median :21.00   Median :126.0
 Mean   :17.2   Mean   :2858   Mean   :11.52   Mean   :20.86   Mean   :128.2
 3rd Qu.:18.1   3rd Qu.:2945   3rd Qu.:13.00   3rd Qu.:25.00   3rd Qu.:143.8
 Max.   :21.2   Max.   :3125   Max.   :15.00   Max.   :33.00   Max.   :192.0
>  # Valores NA
>  any(is.na(textura_comidas))
[1] FALSE
>  # Convertir los datos a un DataFrame
>  texComida <- as.data.frame(scale(textura_comidas))
>  # Obtención de medidas estadísticas del DF
>  summary(texComida)
      Oil              Density             Crispy            Fracture            Hardness
 Min.   :-2.1997   Min.   :-2.31004   Min.   :-2.5457   Min.   :-2.16975   Min.   :-2.09396
 1st Qu.:-0.5666   1st Qu.:-0.68353   1st Qu.:-0.8561   1st Qu.:-0.70617   1st Qu.:-0.67239
 Median :-0.1897   Median : 0.07952   Median : 0.2703   Median : 0.02561   Median :-0.07003
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.5641   3rd Qu.: 0.70201   3rd Qu.: 0.8335   3rd Qu.: 0.75740   3rd Qu.: 0.50020
 Max.   : 2.5113   Max.   : 2.14779   Max.   : 1.9599   Max.   : 2.22097   Max.   : 2.05027
```

```
>   # Se aplica la medida de distancia para obtener la matriz distancia
>   dist_mat <- dist(texComida, method = 'euclidean')
>   # Obtención de los grupos
>   grupos <- hclust(dist_mat, method = "ward.D")
>   plot(grupos)
```




```
>   #Trazar la línea de corte y mostrar rectangulo en los grupos
>   lineaCorte <- cutree(grupos, k=3)
>   plot(grupos)
>   rect.hclust(grupos, k = 3, border = 2:6)
>   abline(h = 3, col = 'red')
>
```



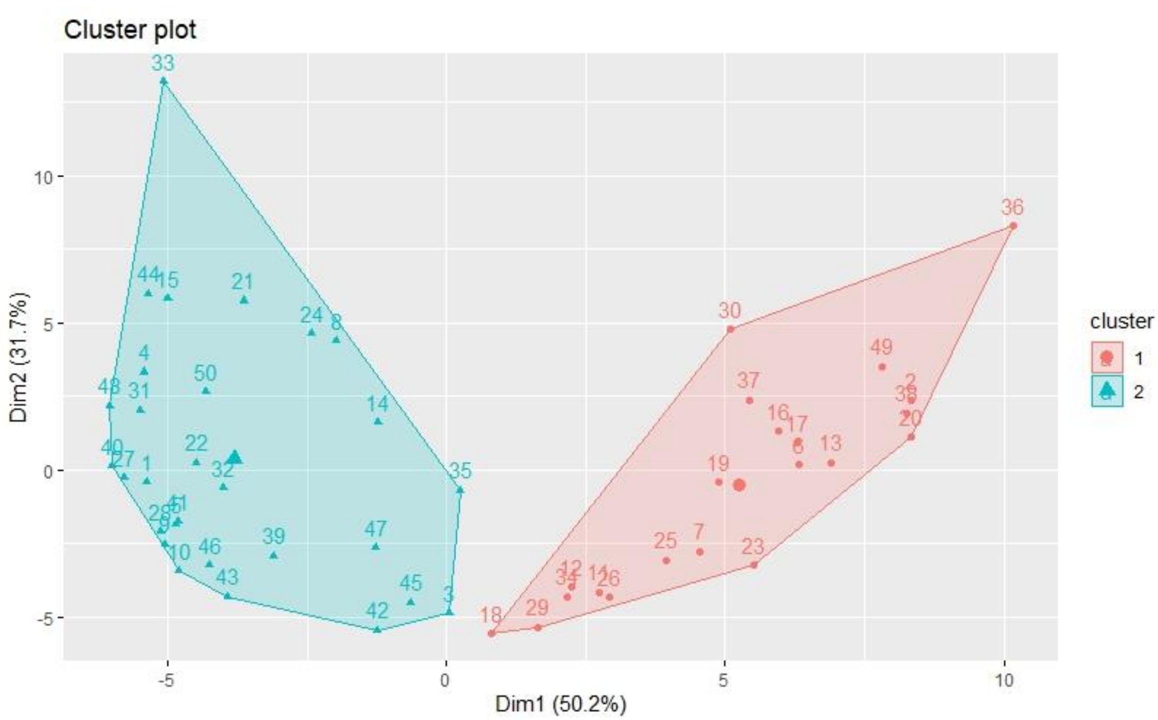
Algoritmo K-MEANS

```r
## Algoritmo K-MEANS##
library (tidyverse)  # Manipulación de datos
library (cluster)    # Algoritmos de clusterización
library (factoextra) # Algoritmos de clusterización y visualización
principale <- function()
{

# Cargar los datos del archivo
textura_comidas <- read.csv("C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/food-texture.csv")
View(textura_comidas)
textura_comidas <- textura_comidas[, -1]

# Explorar los datos
# Desplegar la estructura de los datos
str (textura_comidas)

# Obtención de medidas estadísticas
summary(textura_comidas)

# Valores NA
any(is.na(textura_comidas))

# Convertir los datos a un DataFrame
texComida <- as.data.frame(scale(textura_comidas))

# Obtención de medidas estadísticas del DF
summary(texComida)

# Se aplica la medida de distancia para obtener la matriz distancia
dist_mat <- dist(texComida, method = 'euclidean')


# Se aplica algoritmo K-Means
grupoK2 <- kmeans(dist_mat, centers = 2, nstart = 25)

# Obtención de la estructura de los datos K2
str(grupoK2)

# Impresión de los grupos
print(grupoK2)

# Obtención de gráfico de los grupos
fviz_cluster(grupoK2, data = d_mat)
}
```

EJECUCIÓN

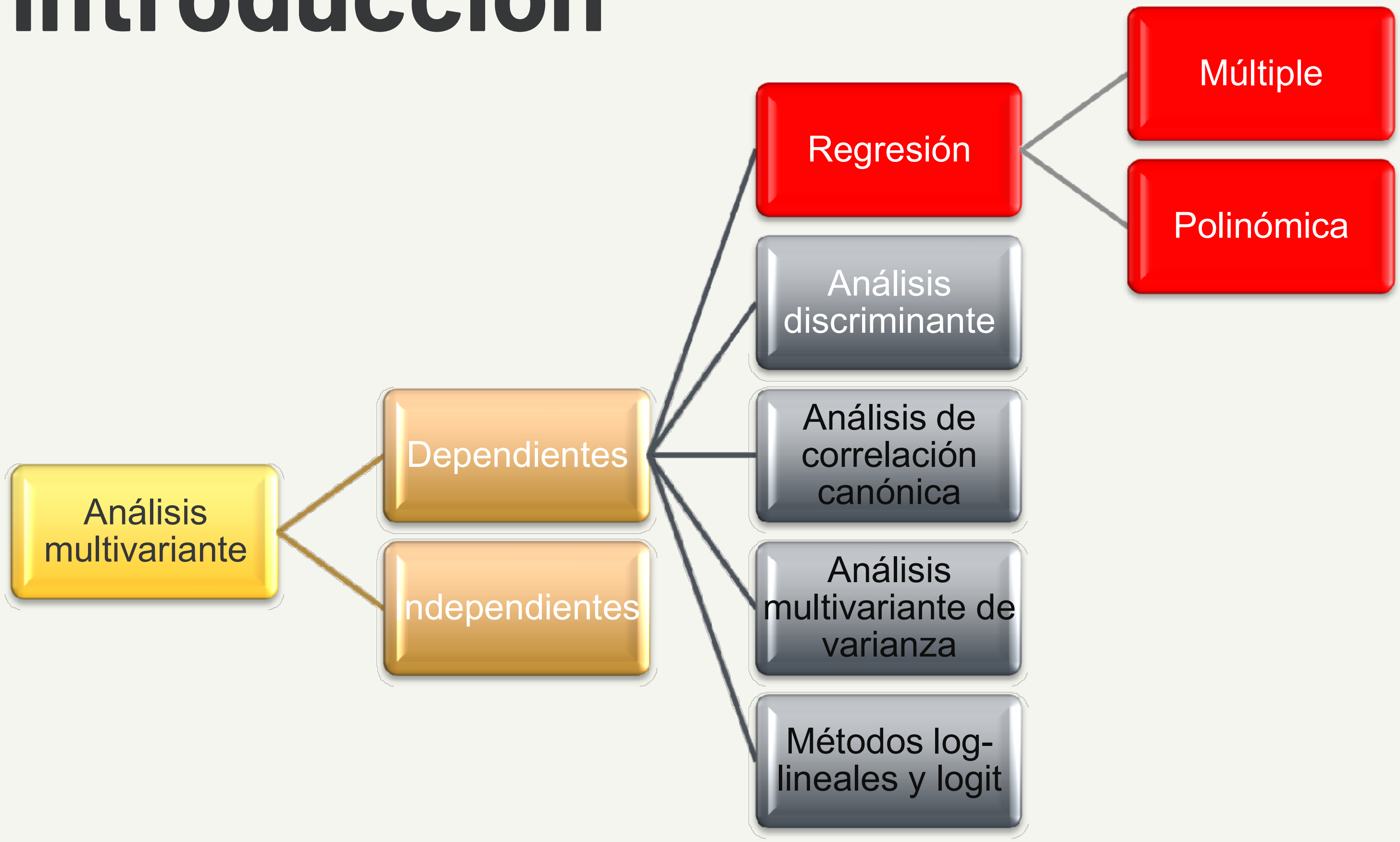
```
> # Cargar los datos del archivo
> textura_comidas <- read.csv("C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/food-texture.csv")
> View(textura_comidas)
> textura_comidas <- textura_comidas[, -1]
> # Exlorar los datos
> # Desplegar la estructura de los datos
> str (textura_comidas)
'data.frame':	50 obs. of  5 variables:
 $ Oil     : num  16.5 17.7 16.2 16.7 16.3 19.1 18.4 17.5 15.7 16.4 ...
 $ Density : int  2955 2660 2870 2920 2975 2790 2750 2770 2955 2945 ...
 $ Crispy  : int  10 14 12 10 11 13 13 10 11 11 ...
 $ Fracture: int  23 9 17 31 26 16 17 26 23 24 ...
 $ Hardness: int  97 139 143 95 143 189 114 63 123 132 ...
> # Obtención de medidas estadísticas
> summary(textura_comidas)
      Oil           Density         Crispy         Fracture        Hardness    
 Min.   :13.7   Min.   :2570   Min.   : 7.00   Min.   : 9.00   Min.   : 63.0  
 1st Qu.:16.3   1st Qu.:2772   1st Qu.:10.00   1st Qu.:17.00   1st Qu.:107.2  
 Median :16.9   Median :2868   Median :12.00   Median :21.00   Median :126.0  
 Mean   :17.2   Mean   :2858   Mean   :11.52   Mean   :20.86   Mean   :128.2  
 3rd Qu.:18.1   3rd Qu.:2945   3rd Qu.:13.00   3rd Qu.:25.00   3rd Qu.:143.8  
 Max.   :21.2   Max.   :3125   Max.   :15.00   Max.   :33.00   Max.   :192.0  
> # Valores NA
> any(is.na(textura_comidas))
[1] FALSE
> # Convertir los datos a un DataFrame
> texComida <- as.data.frame(scale(textura_comidas))
> # Obtención de medidas estadísticas del DF
> summary(texComida)
      Oil              Density             Crispy            Fracture            Hardness       
 Min.   :-2.1997   Min.   :-2.31004   Min.   :-2.5457   Min.   :-2.16975   Min.   :-2.09396  
 1st Qu.:-0.5666   1st Qu.:-0.68353   1st Qu.:-0.8561   1st Qu.:-0.70617   1st Qu.:-0.67239  
 Median :-0.1897   Median : 0.07952   Median : 0.2703   Median : 0.02561   Median :-0.07003  
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000  
 3rd Qu.: 0.5641   3rd Qu.: 0.70201   3rd Qu.: 0.8335   3rd Qu.: 0.75740   3rd Qu.: 0.50020  
 Max.   : 2.5113   Max.   : 2.14779   Max.   : 1.9599   Max.   : 2.22097   Max.   : 2.05027  
```

```
> # Se aplica la medida de distancia para obtener la matriz distancia
> dist_mat <- dist(texComida, method = 'euclidean')
> # Se aplica algoritmo K-Means
> grupoK2 <- kmeans(dist_mat, centers = 2, nstart = 25)
> # Obtención de la estructura de los datos K2
> str(grupoK2)
List of 9
 $ cluster     : Named int [1:50] 2 1 2 2 2 1 1 2 2 2 ...
  ..- attr(*, "names")= chr [1:50] "1" "2" "3" "4" ...
 $ centers     : num [1:2, 1:50] 3.55 1.7 2.22 4.46 2.29 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr [1:50] "1" "2" "3" "4" ...
 $ totss       : num 3468
 $ withinss    : num [1:2] 741 1116
 $ tot.withinss: num 1856
 $ betweenss   : num 1612
 $ size        : int [1:2] 21 29
 $ iter        : int 1
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> # Impresión de los grupos
> print(grupoK2)
K-means clustering with 2 clusters of sizes 21, 29

Cluster means:
         1        2        3        4        5        6        7        8        9       10       11       12       13       14       15
1 3.547150 2.219277 2.294133 4.187572 3.421034 2.215842 1.689055 3.807005 3.294872 3.091560 1.811418 2.100035 2.186721 3.833234 4.583806
2 1.700318 4.460662 2.276758 2.090544 1.720759 3.942471 3.164245 2.761688 1.590546 1.541174 2.764532 2.783534 4.057718 2.823897 2.433502
        16       17       18       19       20       21       22       23       24       25       26       27       28       29       30
1 2.616755 2.047307 1.990682 2.168857 1.954178 4.855855 3.808798 1.618287 4.009697 1.926085 1.865856 3.606805 3.249222 1.836029 2.841213
2 4.047126 3.849889 2.347395 3.518305 4.308582 2.860683 2.026553 3.362756 2.782286 3.144495 2.839834 1.634076 1.588081 2.477215 4.045295
        31       32       33       34       35       36       37       38       39       40       41       42       43       44       45
1 4.143842 3.340371 6.171338 1.899380 3.166682 2.868462 2.376594 2.033204 2.791262 3.730606 3.402125 2.255071 2.857528 4.909842 2.280572
2 2.007385 1.945795 3.368764 2.664192 2.821505 5.380226 3.836397 4.355407 1.827619 1.652520 1.733727 1.949338 1.613553 2.509730 2.140249
        46       47       48       49       50
1 2.944642 2.973519 4.190984 2.286701 3.861665
2 1.619222 2.325479 1.899220 4.402369 2.174200

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
 2  1  2  2  2  1  1  2  2  2  1  1  1  2  2  1  1  1  1  1  2  2  1  2  1  1  2  2  1  1  2  2  2  1  2  1  1  1  2  2  2  2  2  2  2  2  2  2
49 50
 1  2

Within cluster sum of squares by cluster:
[1]  740.662 1115.538
 (between_SS / total_SS =  46.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> # Obtención de gráfico de los grupos
> fviz_cluster(grupoK2, data = dist_mat)
>
```



Modelos predictivos

Introducción



Introducción

- En la regresión lineal simple se establece una relación entre una variable independiente (X) y dependiente(Y)

Ejemplos:

Afore-Años trabajados
Ventas-Cantidad de clientes
Altura –Masa
Gastos de publicidad-Ventas

$$\hat{y} = a + bx$$

Introducción

- Una variable depende no solo de otra variable, sino es multivariable, como:

- La inteligencia, depende de diversos factores como:
  - Genética
  - Contexto social
  - Contexto familiar
  - Ambiente

Regresión lineal múltiple

Regresión lineal múltiple

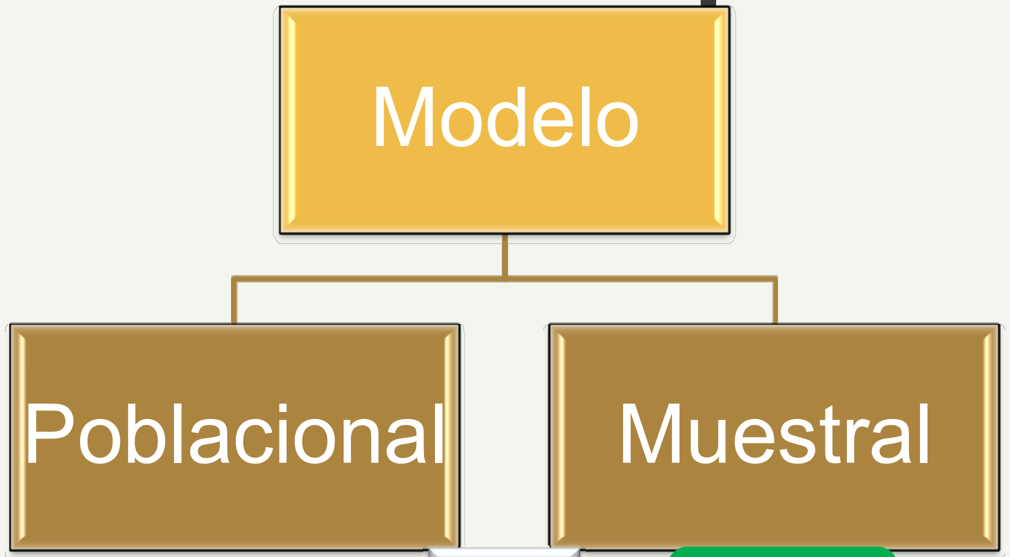
- También llamada modelo de regresión lineal múltiple
- Es una técnica estadística multivariante de tipo dependiente
- Varias **variables independientes (explicativas o regresoras)** Xi que influyen para explicar otra **variable dependiente Y**

Regresión lineal múltiple

Objetivo

- El análisis de regresión lineal múltiple permite modelar y predecir el comportamiento de la variable dependiente (Y) a través de la relación que hay con diversas variables explicativas (X)

Regresión lineal múltiple



Intercepto
Pendientes
Error

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k+\varepsilon$$

Intercepto
Pendientes
Error

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_kx_k+\varepsilon$$

Regresión lineal múltiple

Pendientes (β o b)

Estiman el cambio o valor promedio de y como b1 unidades por cada unidad de incremento de la variable explicativa (xi) manteniendo las otras variables constantes

Y intercepción (β_o ó b_o)

Estima el valor promedio de y cuando todas las variables xi son iguales a cero (suponiendo que el valor de cero está dentro de los rangos de valores que pueden tomar las xi)

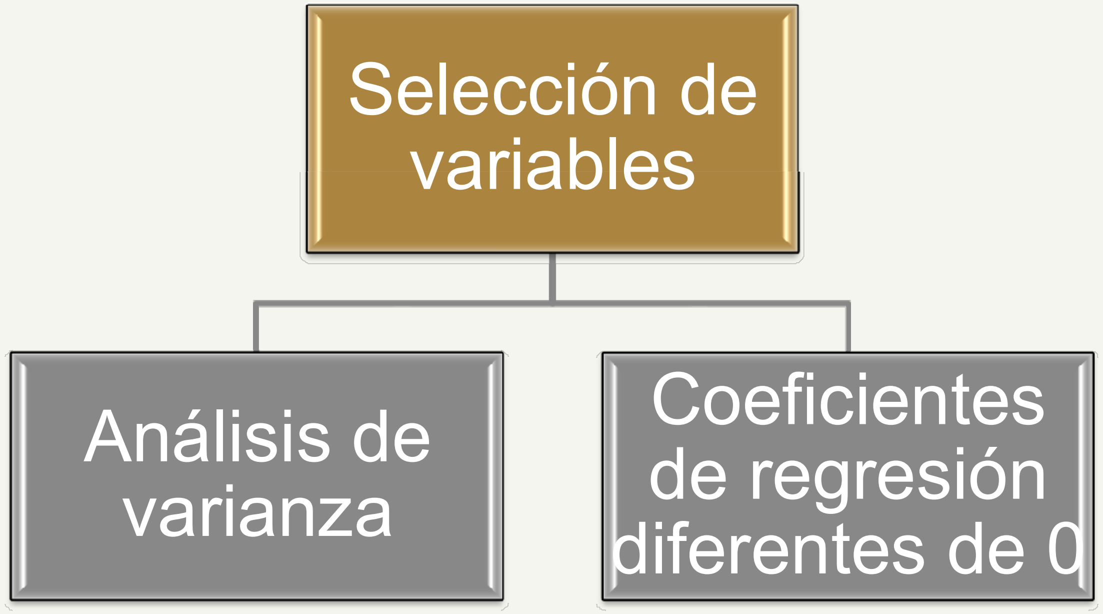
Cálculo de los coeficientes de regresión

Mínimos cuadrados

$$Min \sum (y_j - \hat{y}_j)^2$$

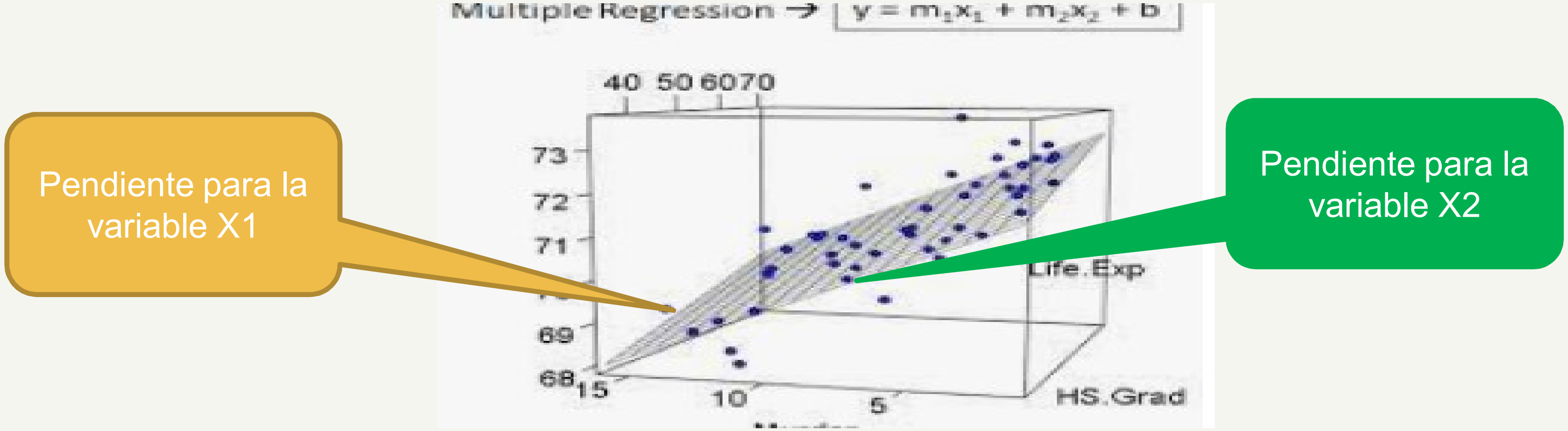
$$\boxed{B = (X' * X)^{-1} * X' * Y}$$

Selección de variables Xi en el modelo



Regresión lineal múltiple

- Cuando se tienen **dos variables** explicativas se llama **hiperplano de regresión**
- Los coeficientes del modelo deben elegirse de tal manera de se **minimice la varianza residual**



Regresión lineal múltiple

- Si las variables explicativas están muy relacionadas entre sí, tendrá un determinante con valor a cero
- Cuando se tiene una fuerte correlación entre las variables explicativas se tiene una **multicolinealidad**
- Cuando se presenta la multicolinealidad no se puede aplicar el método de mínimos cuadrados→ **Selección de variables explicativas**

Efectos de la multicolinealidad

- Varianzas y covarianzas grandes de los coeficientes (parámetros) estimados de regresión por mínimos cuadrados.

- Parámetros de regresión mal estimados

- Distintas muestras tomadas para los mismos valores de las variables explicativas

Regresión lineal múltiple

Para elegir las variables explicativas en el modelo hay que considerar lo siguiente:

- Ser variables de tipo cuantitativas
- No tener variables repetidas
- La relación de las variables explicativas con las variable dependiente debe ser lineal (proporcional)
- Relación del tamaño de la muestra $n = 10*k$

Regresión lineal múltiple

Linealidad

Los valores de la variable explicativa están generados por el modelo

$$\hat{y} = a + bx$$

Homocedasticidad

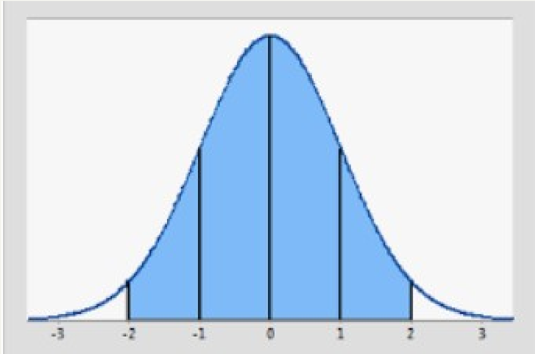
La varianza de los residuos deben ser iguales para todas las variables explicativas

Independencia de residuos

No debe haber una relación entre los valores predichos con los residuos

Coeficiente de Durbin-Watson
[1.25-2.5]

Regresión lineal múltiple



Normalidad

Los residuos se deben de ajustar a una curva normal

Pruebas de normalidad

No multicolinelidad

Fuerte correlación entre las variables explicativas

Máximo factor de inflación de varianza (VIF <10)
Media VIF aprox. 1

Tipo de variable

La variable explicativa (X) ordinal o cuantitativa, la variable Y siempre será cuantitativa

Recurso

- https://www.youtube.com/watch?v=wMg1HU6pfnk

Criterio para la evaluación del modelo

Valores bajos

Error cuadrático medio

$$S\varepsilon = \sqrt{\frac{\sum (y_i - yi)^2}{n}} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

- Entre menor sea el error es mejor el poder predictivo del modelo

Coeficiente de determinación Múltiple (R^2)

- Reporta la proporción de la variación total en y que es explicada por todas las variables

$$R^2 = \frac{SSR}{SST}$$

Suma de cuadrados de regresión

Suma total de cuadrados

Valores altos

- R^2 nunca decrece cuando una nueva variable x es agregada al modelo

R^2 ajustado

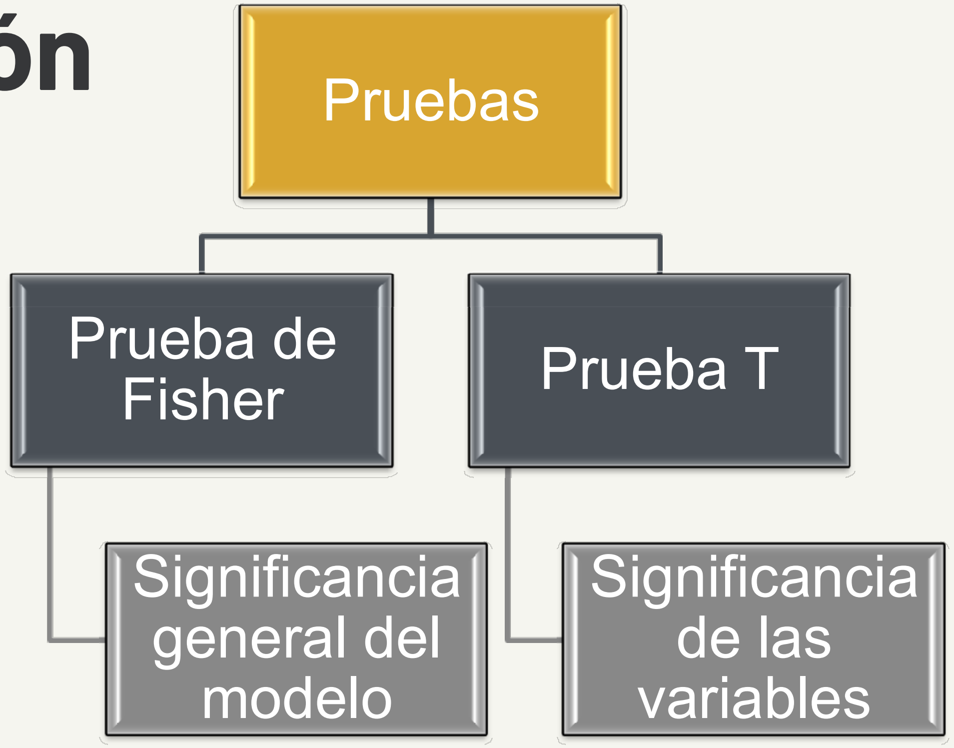
- Muestra la proporción explicada de la variación en y por las variables explicativas considerando la relación entre el tamaño de muestra y el número de variables independientes

$$R_A^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

Donde:
n: Tamaño muestral
k : número de variables explicativas

- Penaliza el uso excesivo de variables independientes no importantes
- Es más pequeña que R^2
- Es útil para la comparación de modelos

Pruebas aplicadas a la regresión



Prueba Fisher

- Se aplica la prueba para medir la significancia del modelo

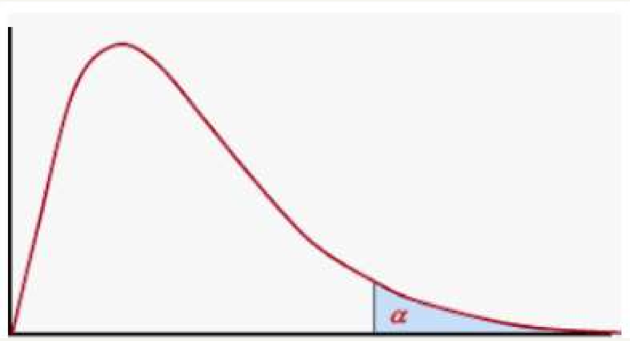
- Permite verificar si hay una relación lineal entre todas las variables x (consideradas en forma conjunta) ; así como, la variable y

- Se plantean dos hipótesis
  - $H_o = b1 = b2 = \ldots bk = 0$ (no hay relación lineal)
  - $H_A =$ al menos un $bi \neq 0$ (existe relación lineal entre y y la variable xi)

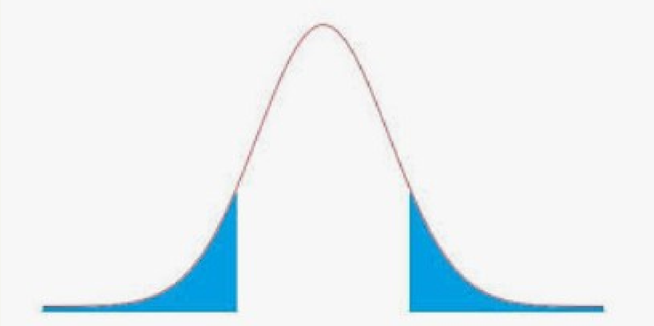
Prueba Fisher

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE}$$

Donde
k: grados de libertad
n: tamaño de la muestra



Significancia de cada variable



- Se utiliza la prueba t para evaluar la significancia de cada pendiente→ **relación lineal**

- **Hipótesis**
- H_o : βi = 0 → No hay relación lineal
- H_A: βi = 0 → Existe relación lineal entre las variables

Intervalo de confianza

$$bi \pm t_{\alpha/2,n-1} s_{bi}$$

Pruebas Multicolinealidad

- Se utiliza el Factor de Inflación de la Varianza (VIF) para medir la colinealidad

$$VIF = \frac{1}{1 - R_j^2}$$

Donde:
R^2: coeficiente de determinación de la regresión de la j variable independiente

VIF = 1→No hay multicolinealidad
VIF > 1 Multicolinealidad
VIF > 5 Multicolinealidad severa

Causas Multicolinealidad

Recolección de datos

- Datos pares de variables que tengan relación

Modelo sobredefinido

- Modelo con más variables predictoras que observaciones

Especificación del modelo

- Agregar términos polinomiales al modelo

Variables Dummies

- Son variables de tipo binarias (categóricas)
- Se les conoce como variables indicadoras
- Las intercepciones cambian si la variable es significativa
- Tienen pendiente al igual que una variable cuantitativa

Recurso Adicional

- https://yuasaavedraco.github.io/Docs/Regresi%C3%B3n lineal m%C3%BAltiple con R.html