Aplicación del Dendrograma

Sisi Guevara García

5/6/2022

Capítulo I

Introducción

El dendrograma es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud. El nivel de similitud se mide en el eje vertical (alternativamente se puede mostrar el nivel de distancia) y las diferentes observaciones se especifican en el eje horizontal.

En el presente trabajo se opto por utilizar este tipo de método de multivariado debido a su gran importancia, puesto que, permite apreciar claramente las relaciones de agrupación entre los datos e incluso entre grupos de ellos aunque no las relaciones de similitud o cercanía entre categorías.

Capítulo II

Descripción de la matriz

La matriz se obtuvo de la librería datasets que se encontraba precargada en R. Esta matriz es un conjunto de datos macroeconómicos que ofrece un ejemplo bien conocido de regresión altamente colineal.

Exploración de la matriz

```
library(cluster.datasets)
library(datasets)
library(knitr)

data("longley")
datos=longley[,-6]
kable(datos)
```

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Employed
1947	83.0	234.289	235.6	159.0	107.608	60.323
1948	88.5	259.426	232.5	145.6	108.632	61.122
1949	88.2	258.054	368.2	161.6	109.773	60.171
1950	89.5	284.599	335.1	165.0	110.929	61.187

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Employed
1951	96.2	328.975	209.9	309.9	112.075	63.221
1952	98.1	346.999	193.2	359.4	113.270	63.639
1953	99.0	365.385	187.0	354.7	115.094	64.989
1954	100.0	363.112	357.8	335.0	116.219	63.761
1955	101.2	397.469	290.4	304.8	117.388	66.019
1956	104.6	419.180	282.2	285.7	118.734	67.857
1957	108.4	442.769	293.6	279.8	120.445	68.169
1958	110.8	444.546	468.1	263.7	121.950	66.513
1959	112.6	482.704	381.3	255.2	123.366	68.655
1960	114.2	502.601	393.1	251.4	125.368	69.564
1961	115.7	518.173	480.6	257.2	127.852	69.331
1962	116.9	554.894	400.7	282.7	130.081	70.551

1.-Dimensión de la variable

La matriz cuenta con 16 observaciones y 6 variables.

```
dim(datos)
```

```
## [1] 16  6
```

2.- Nombre de las variables

```
names(datos)
```

```
## [1] "GNP.deflator" "GNP"          "Unemployed"   "Armed.Forces" "Population"
## [6] "Employed"
```

Las variables con las que cuenta la matriz "longley" son las siguientes:

- **GNP.deflator**: Deflactor de precios implícito del PNB (1954=100)
- **GNP**: Producto Nacional Bruto.
- **Unemployed**: Número de parados.
- **Armed Forces**: Número de personas en las fuerzas armadas.
- **Population**: Población 'no institucionalizada' mayor o igual a 14 años de edad.
- **Year**: El año (tiempo).
- **Employed**: El número de personas empleadas.

3.- Tipo de variables

El conjunto de sus variables son tipo numéricas.

```
str(datos)
```

```
## 'data.frame':    16 obs. of  6 variables:
##  $ GNP.deflator: num  83 88.5 88.2 89.5 96.2 ...
##  $ GNP         : num  234 259 258 285 329 ...
##  $ Unemployed  : num  236 232 368 335 210 ...
##  $ Armed.Forces: num  159 146 162 165 310 ...
##  $ Population  : num  108 109 110 111 112 ...
##  $ Employed    : num  60.3 61.1 60.2 61.2 63.2 ...
```

4.- Presencia de NA's

No se encontró ningún valor faltante.

```
anyNA(datos)
```

```
## [1] FALSE
```

5.- Tratamiento de la matriz

Debido a que la matriz se encuentra limpia y organizada se procederá a usarse tal cual esta para la realización del análisis.

Capútulo III

Metodología de análisis

A partir de la matriz de datos calcularemos la matriz de distancias para posteriormente generar el dendrograma. Se busca generar este debido al interés de agrupar los años con características similares y asimismo conocer la cantidad de grupos que se pueden formar a través de esta matriz según los resultados que muestre el dondrograma.

Capítulo IV

Resultados

Cálculo de la matriz de la distancia de Mahalonobis

Calculamos la distancia de Mahalanobis para las variables numéricas comprendidas.

Con la distancia de Mahalanobis podemos calcular la similitud que existe entre las variables teniendo en cuenta la correlación que hay entre ellas.

```
dist.datos<-dist(datos)
```

Proyección de la matriz de distancia redondeada

```
round(as.matrix(dist.datos),3)
```

```
##         1947    1948    1949    1950    1951    1952    1953    1954    1955
## 1947   0.000  29.206 134.856 111.899 180.553 234.377 241.206 250.757 226.600
## 1948  29.206   0.000 136.655 107.439 180.050 234.615 239.139 250.040 219.123
## 1949 134.856 136.655   0.000  42.613 228.386 278.897 286.023 203.486 215.074
## 1950 111.899 107.439  42.613   0.000 196.699 248.810 254.103 189.009 185.698
## 1951 180.553 180.050 228.386 196.699   0.000  55.310  62.267 153.953 106.107
## 1952 234.377 234.615 278.897 248.810  55.310   0.000  20.113 167.214 122.509
## 1953 241.206 239.139 286.023 254.103  62.267  20.113   0.000 171.958 119.257
## 1954 250.757 250.040 203.486 189.009 153.953 167.214 171.958   0.000  81.505
## 1955 226.600 219.123 215.074 185.698 106.107 122.509 119.257  81.505   0.000
## 1956 230.339 219.148 221.736 189.242 118.686 136.575 129.502 106.460  30.334
## 1957 249.584 236.522 232.899 201.036 145.273 160.519 151.951 116.712  52.470
## 1958 332.029 323.153 236.396 231.554 287.193 307.390 306.192 155.043 188.669
## 1959 305.593 291.310 245.463 224.188 237.547 254.936 248.361 146.480 134.761
## 1960 326.664 311.660 263.619 243.398 260.137 276.182 268.950 167.366 157.152
## 1961 389.639 377.008 301.002 292.005 335.437 350.569 345.705 213.517 230.959
## 1962 383.544 368.435 324.314 304.009 298.312 304.702 295.536 204.646 194.590
##         1956    1957    1958    1959    1960    1961    1962
## 1947 230.339 249.584 332.029 305.593 326.664 389.639 383.544
## 1948 219.148 236.522 323.153 291.310 311.660 377.008 368.435
## 1949 221.736 232.899 236.396 245.463 263.619 301.002 324.314
## 1950 189.242 201.036 231.554 224.188 243.398 292.005 304.009
## 1951 118.686 145.273 287.193 237.547 260.137 335.437 298.312
## 1952 136.575 160.519 307.390 254.936 276.182 350.569 304.702
## 1953 129.502 151.951 306.192 248.361 268.950 345.705 295.536
## 1954 106.460 116.712 155.043 146.480 167.366 213.517 204.646
## 1955  30.334  52.470 188.669 134.761 157.152 230.959 194.590
## 1956   0.000  27.179 189.042 121.953 143.434 224.016 180.989
## 1957  27.179   0.000 175.281  99.587 119.777 203.163 155.633
## 1958 189.042 175.281   0.000  95.249  95.808  75.407 131.149
## 1959 121.953  99.587  95.249   0.000  23.600 105.606  80.070
## 1960 143.434 119.777  95.808  23.600   0.000  89.111  61.664
## 1961 224.016 203.163  75.407 105.606  89.111   0.000  91.600
## 1962 180.989 155.633 131.149  80.070  61.664  91.600   0.000
```

Gráfico de la matriz de distancias

Los colores describen la distancia entre los sujetos, por lo que se muestra que mientras el gradiente sea mas rojo los sujetos estarán mas cerca. Al contrario de la tonalidad azul que muestra que las observaciones se encuentran mas alejadas.

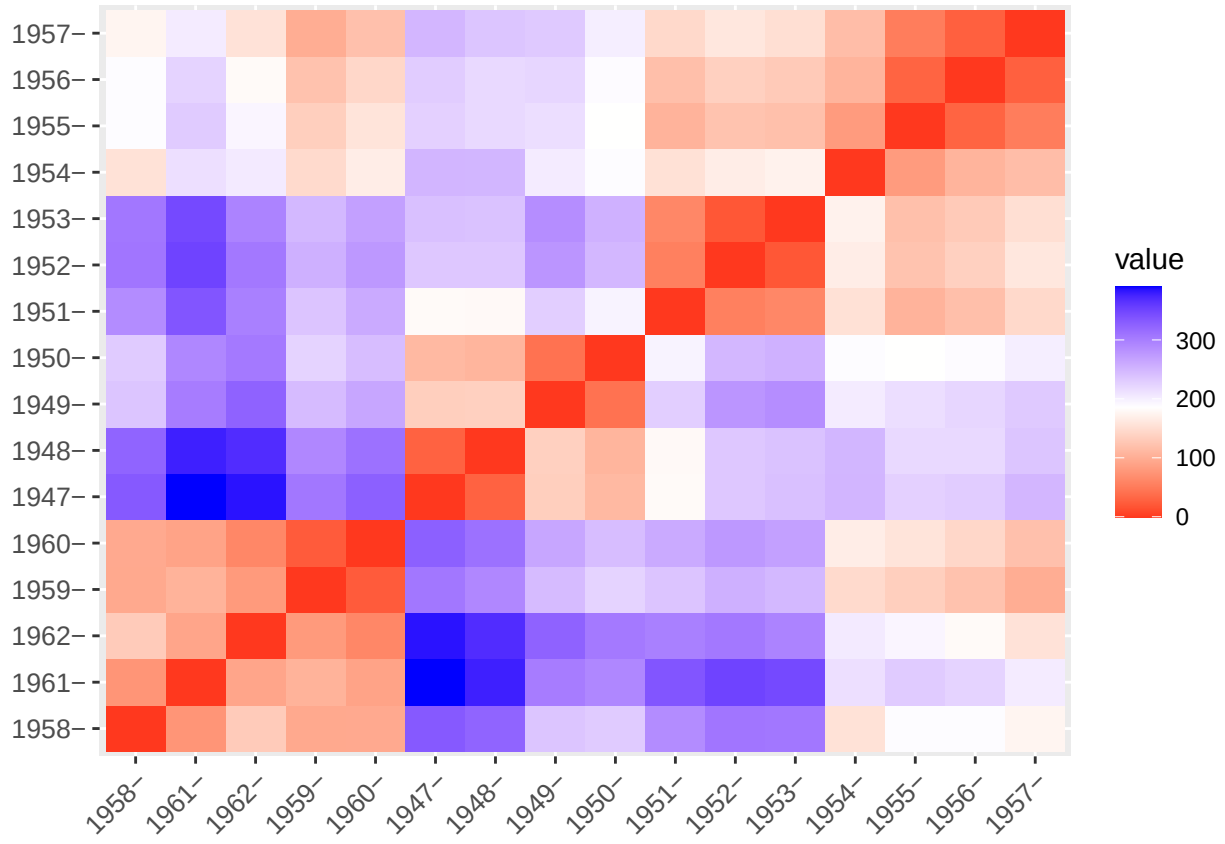
```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_dist(dist.obj = dist.datos, lab_size = 10)
```



Cálculo del dendrograma

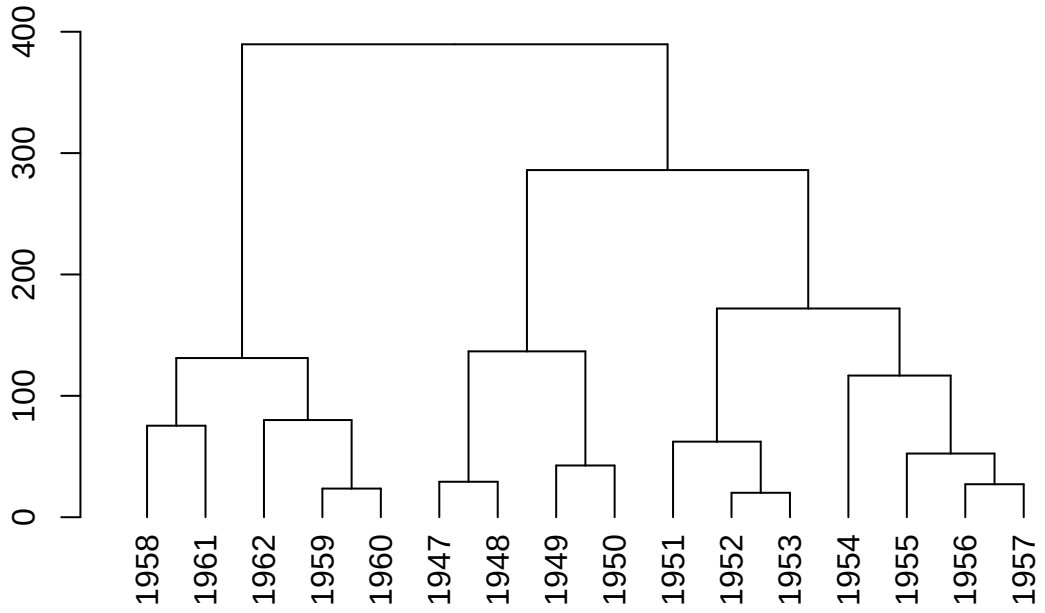
Se calcula el dendrograma para nuestras observaciones donde usaremos el método de agrupación por Clústers **"hclust"**, el cual nos ofrece una agrupación jerárquica.

```
dend.datos<-as.dendrogram(hclust(dist.datos))
```

Generación del dendrograma

En este dendrograma se puede visualizar que si se realiza un corte a la altura de 300 a 400 en la escala de la distancia podemos ver que se desprenden dos grupos y si se propusiera hacer un corte a la altura de 200 se hacen tres grupos. Posteriormente si se realiza el mismo proceso a la altura de 150 se puede observar que se realizan cuatro cortes.

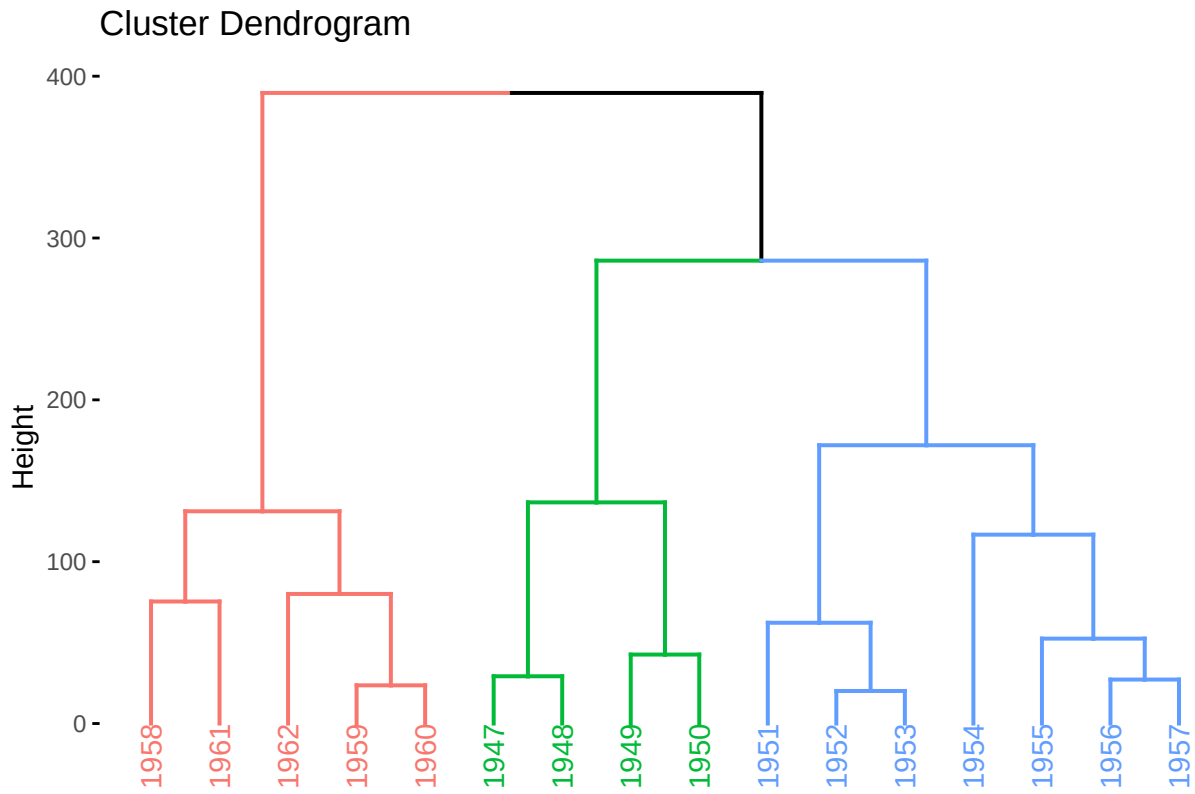
```
plot(dend.datos)
```



Dendrograma con tres particiones

En el siguente dendrograma se propuso realizar tres particiones debido al equilibrio entre la cantidad de grupos y sujetos que pertenecen a ellos.

```
library(factoextra)
library(ggplot2)
hc <- hclust(dist.datos)
fviz_dend(as.dendrogram(hc),
          k = 3)
```



Capítulo V

Conclusiones

Tras la realización del ejercicio del dendrograma, se realizaron tres particiones debido a con esa cantidad de cortes se encuentra una equilibrio entre la cantidad de grupos y la cantidad de sujetos de cada grupo. A la altura de 200 se generaron 3 grupos a los cuales el primer grupo pertenecen los sujetos 1958, 1961, 1962, 1959 y 1960. Por otro lado, en el segundo grupo se encuentran 1947, 1948, 1949 y 1950 y en el ultimo grupo se encuentra una mayor cantidad de sujetos, como lo son 1951, 1952, 1953, 1954, 1955, 1956 y 1957. Se concluye que son agrupados de esta manera debido a que tienen características muy similares en sus variables.

Referencias

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

J. W. Longley (1967) An appraisal of least-squares programs from the point of view of the user. Journal of the American Statistical Association 62, 819–841.