IDH

Ignacio Scuderi

28/12/2021

1.Introducción

El presente trabajo busca analizar los datos relativos al Índice de Desarrollo Humano (IDH), utilizando la metodología de componentes principales. El mismo se encuentra estructurado de la siguiente forma: luego de la presente introducción, continua la segunda parte correspondiente al importado y la limpieza de los datos. En la tercera parte, se indagan los datos mediante el uso de componentes principales.

El IDH es un índice compuesto calculado por la Organización de las Naciones Unidas (ONU), como medida multidimensional del nivel de desarrollo de los países. El mismo tiene en cuenta variables sanitarias, de expectativa de vida, distribución del ingreso, y producto bruto interno per cápita.

Se utilizará para el análisis el fichero *data_HDI_2019.csv*, que contiene los datos de IDH correspondientes a 189 países y 22 variables.

2.Importado y limpieza de los datos

En primer lugar se debe importar el fichero de trabajo:

```
IDH= read.csv("data_HDI_2019.csv", dec = ",", header = TRUE, sep = ";")
```

Identificación y eliminación de las filas con valores *NA*

```
IDH2=IDH %>% drop_na()
```

Reemplazo de la columna que indexa el listado de países por la columna *country*:

```
rownames(IDH2) = IDH2$COUNTRY
IDH2$COUNTRY= NULL
head(IDH2)
```

```
              HDI  LEB EYEDU MYEDU GNIpc  IHDI CHI IN_LE IN_EDU IN_INC
Norway      0.957 82.4  18.1  12.9 66494 0.899 6.0   3.0    2.3   12.6
Ireland     0.955 82.3  18.7  12.7 68371 0.885 7.2   3.4    3.3   15.0
Switzerland 0.955 83.8  16.3  13.4 69394 0.889 6.8   3.5    1.8   14.9
Iceland     0.949 83.0  19.1  12.8 54682 0.894 5.6   2.4    2.8   11.7
Germany     0.947 81.3  17.0  14.2 55314 0.869 7.9   3.8    2.3   17.7
Sweden      0.945 82.8  19.5  12.5 54508 0.882 6.5   2.9    3.7   13.0
            INC_40_POOR INC_10_RICH INC_1_RICH GINI   GII MMR ABR SSP_F P2EDU_F
Norway             23.2        21.6        9.4 27.0 0.045   2 5.1  40.8    95.4
Ireland            20.5        25.9       11.3 32.8 0.093   5 7.5  24.3    81.9
Switzerland        20.2        25.5       10.6 32.7 0.025   5 2.8  38.6    95.6
Iceland            23.7        22.5        7.6 26.8 0.058   4 6.3  38.1   100.0
Germany            20.4        24.6       12.5 31.9 0.084   7 8.1  31.6    95.9
Sweden             22.2        22.3        9.0 28.8 0.039   4 5.1  47.3    89.3
            P2EDU_M LFP_F LFP_M
```

```
Norway         94.9  60.4  67.2
Ireland        79.9  56.0  68.4
Switzerland    96.8  62.9  73.8
Iceland       100.0  70.8  79.2
Germany        96.3  55.3  66.6
Sweden         89.5  61.4  67.8
```

```
names(IDH2)
```

```
 [1] "HDI"          "LEB"          "EYEDU"        "MYEDU"        "GNIpc"
 [6] "IHDI"         "CHI"          "IN_LE"        "IN_EDU"       "IN_INC"
[11] "INC_40_POOR" "INC_10_RICH" "INC_1_RICH"  "GINI"         "GII"
[16] "MMR"          "ABR"          "SSP_F"        "P2EDU_F"      "P2EDU_M"
[21] "LFP_F"        "LFP_M"
```

Se observa en el paso anterior que la variable IDH esta incluida, por lo que antes de realizar el análisis de componentes principales, se debe excluir del dataframe. De igual modo, se procede a excluir IHDI ya que refleja el propio IDH ajustado por desigualdad.

```
var_excluidas= c("HDI", "IHDI")
IDH2 = select(IDH2, -var_excluidas)
names(IDH2)
```

```
 [1] "LEB"          "EYEDU"        "MYEDU"        "GNIpc"        "CHI"
 [6] "IN_LE"        "IN_EDU"       "IN_INC"       "INC_40_POOR" "INC_10_RICH"
[11] "INC_1_RICH"  "GINI"         "GII"          "MMR"          "ABR"
[16] "SSP_F"        "P2EDU_F"      "P2EDU_M"      "LFP_F"        "LFP_M"
```

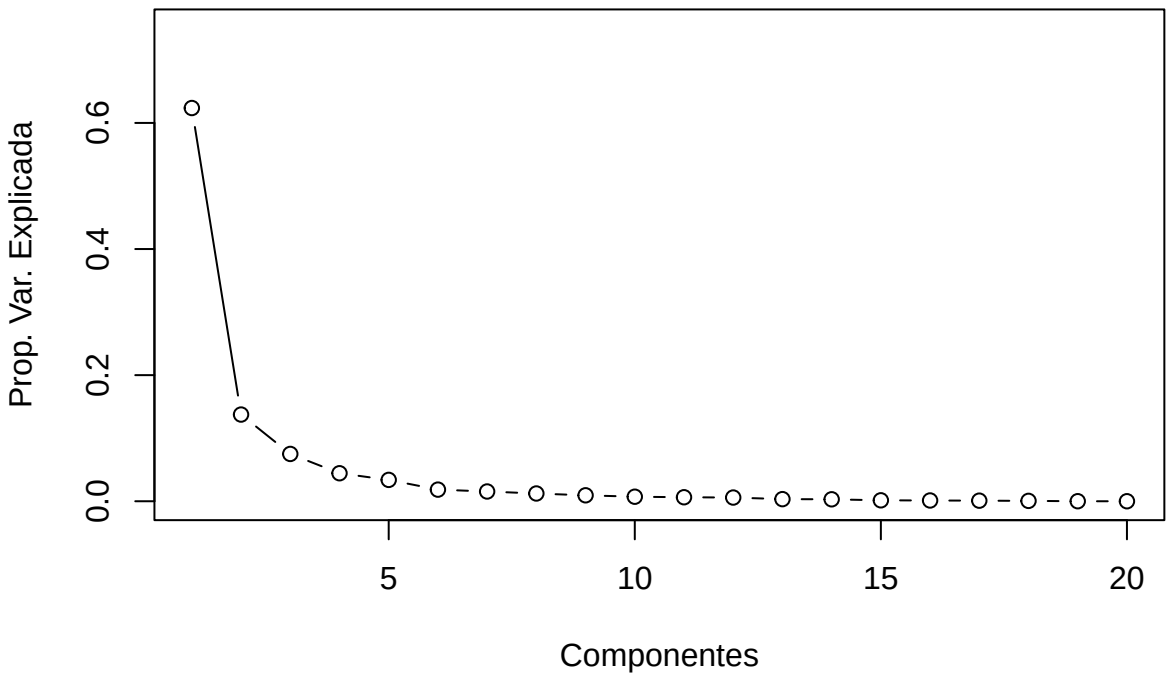
3. Componentes principales

3.1 Realiza una anàlisis de componenetes principales de los datos anteriores.

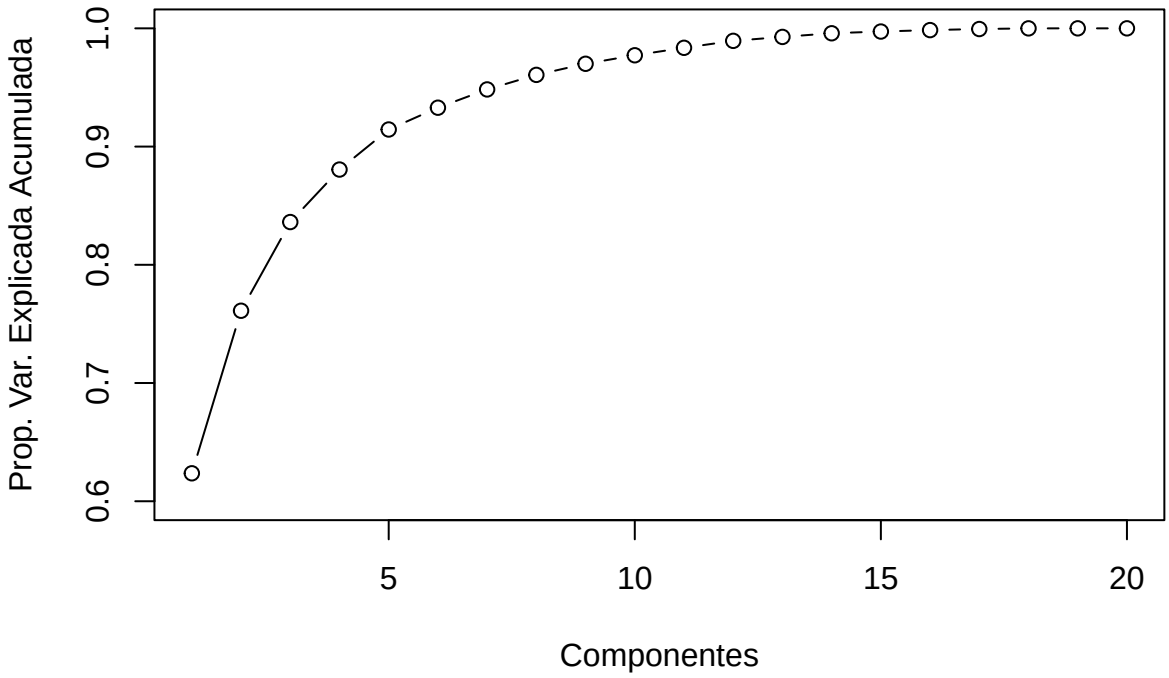
Se procede a realizar el análsis por componentes principales ajustando por centrado y escala.

```
set.seed(2021)
ACP= prcomp(IDH2, center=TRUE, scale=TRUE)
ACP.var=ACP$sdev^2
pve=ACP.var/sum(ACP.var)

plot(pve, xlab="Componentes", ylab="Prop. Var. Explicada", ylim=c(0,0.75),type='b')
```




```
plot(cumsum(pve), xlab="Componentes", ylab="Prop. Var. Explicada Acumulada", ylim=c(0.6,1),type='b')
```




```
summary(ACP)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     3.5317 1.6583 1.22500 0.94246 0.82294 0.60735 0.55599
Proportion of Variance 0.6236 0.1375 0.07503 0.04441 0.03386 0.01844 0.01546
Cumulative Proportion  0.6236 0.7611 0.83617 0.88058 0.91444 0.93288 0.94834
                          PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.4959 0.4336 0.37922 0.35531 0.34295 0.25804 0.24864
Proportion of Variance 0.0123 0.0094 0.00719 0.00631 0.00588 0.00333 0.00309
Cumulative Proportion  0.9606 0.9700 0.97723 0.98354 0.98942 0.99275 0.99584
                         PC15    PC16   PC17    PC18    PC19     PC20
Standard deviation     0.17119 0.15587 0.1342 0.10654 0.01325 0.002026
Proportion of Variance 0.00147 0.00121 0.0009 0.00057 0.00001 0.000000
Cumulative Proportion  0.99731 0.99852 0.9994 0.99999 1.00000 1.000000
```

Se pone de manifiesto que la primer componente explica el 62,3% de la varianza, mientras que la segunda el 13,7% (76,1% acumulada). A partir de allí cada componente adicional explica una proporción menor de la varianza. La tercera componente explica el 7,5% (83,6% acumulada), mientras que la cuarta un 4,4% (88% acumulada).

3.2 Interpreta la primera y la segunda componente principal a partir de los vectores de cargas.

Interpretación de la primera componente:

```
sort(ACP$rotation[,1], decreasing = TRUE)
```

```
        CHI       IN_LE         GII         ABR      IN_EDU         MMR
 0.27441506  0.26947843  0.26913413  0.25624316  0.24425779  0.23253952
INC_10_RICH        GINI      IN_INC  INC_1_RICH       LFP_M       LFP_F
 0.19939642  0.19769420  0.18334126  0.16632918  0.13916972  0.05937089
      SSP_F INC_40_POOR       EYEDU       GNIpc     P2EDU_M     P2EDU_F
-0.10905893 -0.19036125 -0.24481092 -0.24492060 -0.25264642 -0.25765378
      MYEDU         LEB
-0.26139287 -0.26578451
```

Se observa que las variables *CHI* (coeficiente de desigualdad humana), *IN_LE* (desigualdad en la esperanza de vida), y *GII* (índice de desigualdad de género) son los principales vectores de carga positivos de la componente 1, mientras que *LEB* (esperanza de vida al nacer), *MYEDU* (promedio de años de escolaridad), y *P2EDU_F* (Población femenina con al menos algunos estudios secundarios) son los principales vectores de carga negativos.

De este modo, cuanto *CHI*, *IN-LE*, O *GII* se incrementan, también lo hace la primera componente, mientras que cuando *LEB*, *MYEDU*, O *P2EDU_F* incrementan, la primera componente disminuye su valor.

Interpretación de la segunda componente:

```
sort(ACP$rotation[,2], decreasing = TRUE)
```

```
       GINI INC_10_RICH      IN_INC  INC_1_RICH       LFP_F       EYEDU
 0.41041871  0.39148314  0.37202211  0.31815437  0.19188451  0.17854473
      MYEDU     P2EDU_F     P2EDU_M       SSP_F       GNIpc       LFP_M
 0.16794024  0.14961493  0.13996807  0.12721126  0.08612089  0.06253594
        LEB         CHI         ABR       IN_LE         GII         MMR
 0.05839162 -0.03177961 -0.03241203 -0.10432570 -0.10515543 -0.16763374
     IN_EDU INC_40_POOR
-0.22105670 -0.40976617
```

Se observa que las variables *GINI* (coeficiente de Gini), *INC_10_RICH* (cuota de ingreso del 10 por ciento más rico), y *INC_1_RICH* (cuota de ingreso del 1 por ciento más rico) son los principales vectores de carga positivos de la componente 2, mientras que *INC_40_POOR* (cuota de ingreso del 40 por ciento más pobre), *IN_EDU* (desigualdad en la educación), y *MMR* (ratio de mortalidad materna) son los principales vectores de carga negativos.

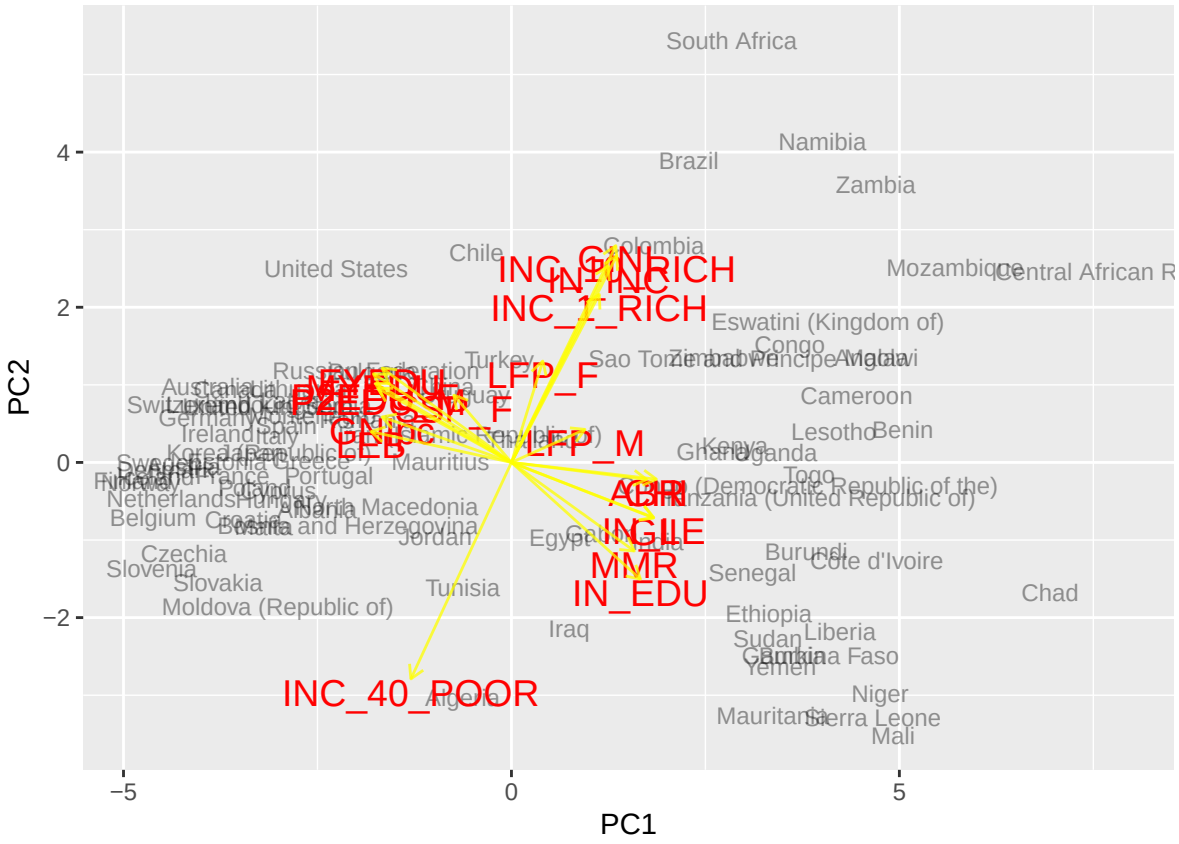
3.3 Interpreta el biplot de la primera y segunda componente principal. ¿Qué puedes decir de los scores? Recuerda que las primeras observaciones tiene el IDH elevado y las últimas lo tienen bajo.

Biplot de la primera y la segunda componente:

```
PCbiplot <- function(PC, x="PC1", y="PC2", colors=c('black', 'black', 'red', 'red')) {
  data <- data.frame(obsnames=row.names(PC$x), PC$x)
  plot <- ggplot(data, aes_string(x=x, y=y)) + geom_text(alpha=.4, size=3, aes(label=obsnames), color=c
  datapc <- data.frame(varnames=rownames(PC$rotation), PC$rotation)
  mult <- min(
    (max(data[,y]) - min(data[,y])/(max(datapc[,y])-min(datapc[,y]))),
    (max(data[,x]) - min(data[,x])/(max(datapc[,x])-min(datapc[,x])))
  )
  datapc <- transform(datapc,
                      v1 = .7 * mult * (get(x)),
                      v2 = .7 * mult * (get(y))
  )
```

```
  plot <- plot + coord_equal() + geom_text(data=datapc, aes(x=v1, y=v2, label=varnames), size = 5, vjust
  plot <- plot + geom_segment(data=datapc, aes(x=0, y=0, xend=v1, yend=v2), arrow=arrow(length=unit(0.2
  plot
}

PCbiplot(ACP, colors=c("black", "black", "red", "yellow"))
```



Los scores agrupan a los países de acuerdo al grado de similitud que poseen en los valores obtenidos. De este modo, se puede observar por ejemplo que los países con un IDH alto se encuentran agrupados en los valores negativos de la primera componente, mientras que los países con menor IDH se encuentran en el cuadrante de valores positivos de la primera componente principal.

3.4 ¿Con cuantas componentes te quedarías para representar de forma resumida toda la información contenida en los datos?

Si utilizamos la regla del codo y en base a los gráficos presentados en el inciso 3.1, deberíamos quedarnos con dos componentes, con lo cual explicamos el 76,1% de la varianza.

3.5 La ONU calcula el IDH dando el misma peso a los indicadores de salud (esperanza de vida al nacer), educación (años esperados de escolarización y la media de años de escolarización, para lo cual calcula la media aritmética de los índices simples calculados a partir de dichas dos variables) y nivel de vida (Renta Nacional Bruta per capita).3 La OCDE (2008), en su manual de elaboración de indicadores compuestos de sugiere, entre otras posibles metodologías de agregación de los índices simples, el uso de componentes principales. Calcula las componentes principales de las variables LEB, EYEDU, MYEDU y GNIpc.4 Estudia sus vectores de carga y comenta dichos resultados en función de cómo pondera el IDH los diferentes indicadores. ¿Con cuantas componentes principales te quedarías?

Calculo de los componentes principales de las variables seleccionadas:

```
variables_OCDE= select(IDH2, "LEB", "EYEDU", "MYEDU", "GNIpc")
ACP_OCDE= prcomp(variables_OCDE, center=TRUE, scale=TRUE)
summary(ACP_OCDE)
```

```
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.8793 0.42751 0.39013 0.36513
Proportion of Variance 0.8829 0.04569 0.03805 0.03333
Cumulative Proportion  0.8829 0.92862 0.96667 1.00000
```

Interpretación de la primera componente:

```
sort(ACP_OCDE$rotation[,1], decreasing = TRUE)
```

```
     GNIpc        LEB      EYEDU      MYEDU
-0.4953222 -0.5011125 -0.5015287 -0.5020071
```

Se ve que para la primera componente la variable *MYEDU* (promedio de años de escolaridad) y *EYEDU* (años previstos de escolaridad) son los vectores de carga negativos más importantes. No se registran vectores de carga positivos.

Interpretación de la segunda componente:

```
sort(ACP_OCDE$rotation[,2], decreasing = TRUE)
```

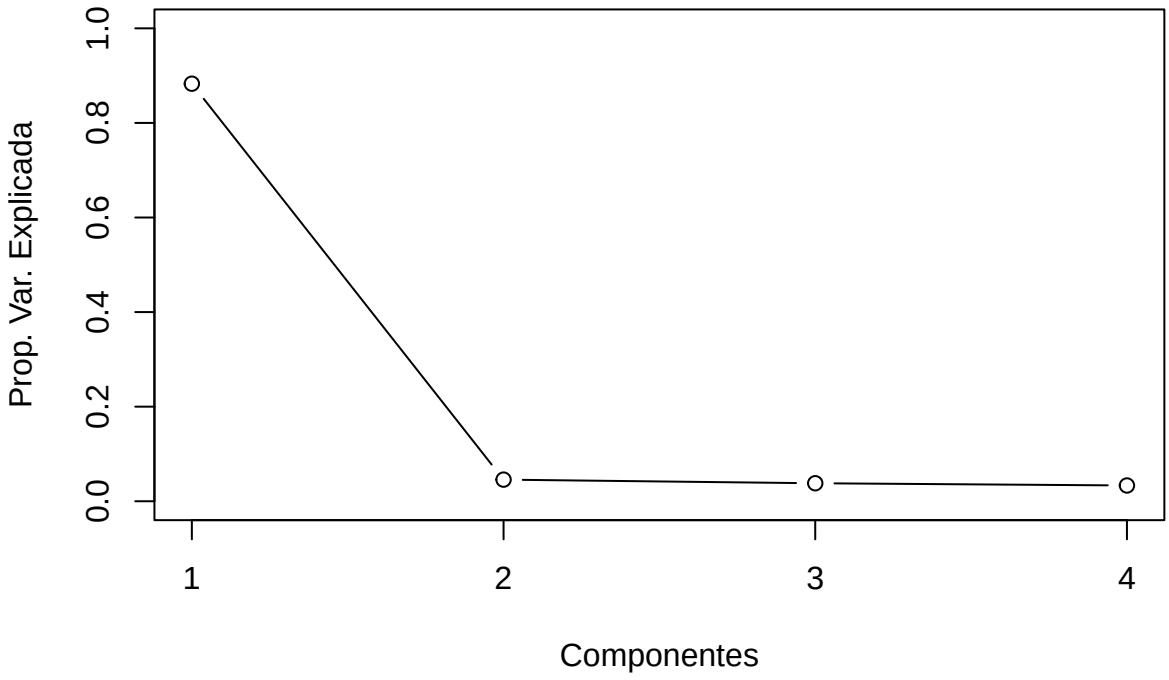
```
      EYEDU         LEB       MYEDU       GNIpc
 0.48465503  0.34429599 -0.03525034 -0.80332259
```

Se observa que *GNIpc* (renta nacional bruta (RNB) per cápita) es el vector de carga negativo más influyente, mientras que *EYEDU* (años previstos de escolaridad) es el vector de carga positivo más relevante.

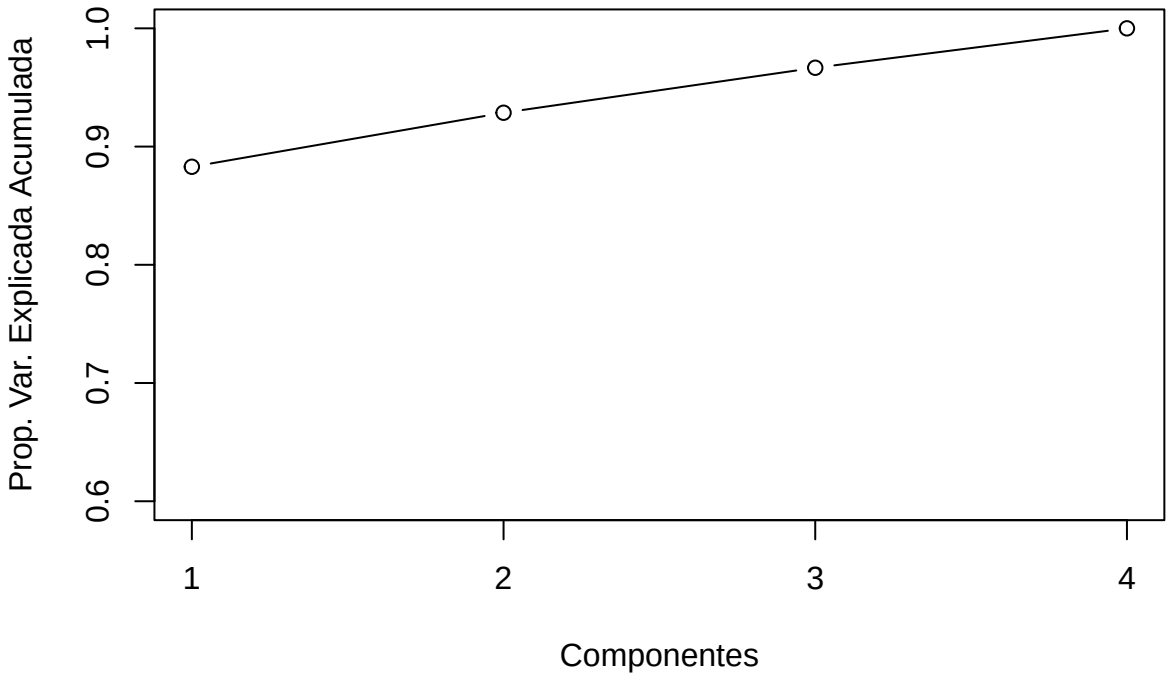
Análisis por método del codo para definir la cantidad de componentes principales a elegir:

```
ACP_OCDE.var=ACP_OCDE$sdev^2
pve_OCDE=ACP_OCDE.var/sum(ACP_OCDE.var)

y= plot(pve_OCDE, xlab="Componentes", ylab="Prop. Var. Explicada", ylim=c(0,1),xlim=c(1,4), xaxt='n',ty
  axis(side = 1, at=1:4)
```




```
z= plot(cumsum(pve_OCDE), xlab="Componentes", ylab="Prop. Var. Explicada Acumulada", ylim=c(0.6,1),xaxt=
  axis(side=1, at=1:4)
```



Se observa que la primera componente principal explica el 88% de la varianza, mientras que añadiendo la segunda el 4,5% (92,8% acumulada), y una tercera el 3,8% (96,6% acumulada). Siguiendo la regla del codo corresponde elegir dos componentes.