Universidad de los Andes

Taller métodos multivariantes

Análisis de componentes principales

Giovany Babativa-Márquez, PhD

Técnicas multivariantes: Métodos no supervisados

Considere el conjunto de datos “PaisesProteinas.sav”, el cual contiene información de 9 variables para 25 países europeos. Las variables representan el porcentaje de consumo de proteínas de Carnes rojas, carnes blancas, huevos, Leche, pescado, Cereales, féculas, frutos secos y, frutas y verduras.

1. Cargue el conjunto de datos a R. Recuerde que puede cargarlo directamente desde GitHub así:

```
options(scipen = 999)
library(pacman)

p_load(tidyverse, janitor, haven, GGally,
       FactoMineR, factoextra)

url <- "https://github.com/jgbabativam/AnaDatos/raw/main/datos/PaisesProteinas.sav"

datos <- read_sav(url)
```

2. Identifique las variables de análisis y prepare el conjunto de los datos.

3. ¿Es apropiado un análisis de componentes principales (PCA)?, explique y ejecute el método.

4. ¿Cuánta información del conjunto de los datos es explicado por las dos primeras dimensiones?. Presente un gráfico que lo ilustre.

5. Haga una gráfica del plano factorial generado por las dimensiones 1 y 2 para las variables.

6. Analice la correlación entre las variables dependiendo del ángulo generado por los vectores que las representan e indique qué podría decir de los países que se ubicarán en cada cuadrante.

7. ¿Considera que existe una asociación entre el consumo de pescado y de

8. Haga una gráfica del plano factorial generado por las dimensiones 1 y 2 para los países. Interprete.

9. ¿En cuáles países se consume menos carnes pero más frutos secos o frutos y cereales?.

10. Proyecte el biplot con el fin de realizar una representación simultánea de variables y países.

11. Si quisiera construir un perfil para Francia, ¿considera que las dimensiones más apropiadas para proyectarlo son la 1 y 2?. Ayuda: use el comando `res$ind$cos2` para identificar las dimensiones en donde se encuentra la mayor información de Francia.

12. De acuerdo con el plano de las variables, es posible construir un índice usando la primera dimensión, en donde un valor alto represente a los países con altos consumos de carne (bajos consumos de frutos secos), un valor bajo representará bajos consumos de carne (altos consumos de frutos secos), aunque el consumo de pescado no estará bien representado en ese índice. Explique la lógica del análisis anterior y construya el índice usando el siguiente código computacional:

```
index <- as.data.frame(res$ind$coord[,1]) |>
         rownames_to_column("Pais") |>
         rename(score = `res$ind$coord[, 1]`) |>
         mutate(Indice = round(GGally::rescale01(score)*100, 1)) |>
         select(-score)
```

13. Realice un gráfico de barras horizontal que presente en orden los países con el índice más alto hasta el más bajo. Concluya.