Análisis de Correspondencia en Datos Categóricos

Sebastián David Reyes Santamaria

17 de marzo de 2025

1. Análisis de Componentes Principales

Comenzamos con el análisis de componentes principales para cada conjunto de datos (hombres y mujeres). Para ello, realizamos la descomposición de la matriz de covarianzas S en sus eigenvectores y eigenvalores:

```
eigen() decomposition
$values
[1] 43.212400 14.958409 10.637096  2.629595

$vectors
           [,1]       [,2]       [,3]
[1,] -0.2245473 -0.06732947  0.2398902
[2,] -0.3155909  0.06695456  0.8980264
[3,] -0.7106910 -0.63913767 -0.2523383
[4,] -0.5872811  0.76320827 -0.2689361
           [,4]
[1,]  0.94207110
[2,] -0.29911201
[3,] -0.15081989
[4,] -0.01695292
```

Figura 1: Descomposición en hombres

```
eigen() decomposition
$values
[1] 48.955644 18.463514 13.537986  4.818058

$vectors
           [,1]       [,2]       [,3]
[1,] -0.2173746 -0.2730167  0.3733308
[2,] -0.3878472 -0.6205057  0.4658646
[3,] -0.6807847 -0.1707259 -0.7076914
[4,] -0.5821125  0.7150435  0.3778455
           [,4]
[1,]  0.85955471
[2,] -0.49751123
[3,]  0.08097994
[4,] -0.08420525
```

Figura 2: Descomposición en mujeres

Los valores propios representan la cantidad de varianza explicada por cada componente principal. En este caso, los valores son mayores para el conjunto de mujeres, lo que sugiere que las variables podrían estar más correlacionadas.

Visualizando el análisis de PCA junto a las observaciones de cada conjunto, obtenemos:

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

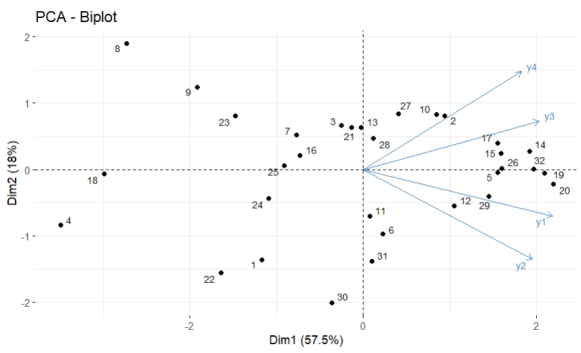

Figura 3: PCA en hombres

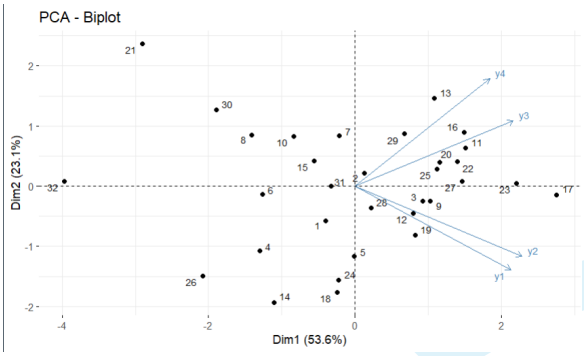

Figura 4: PCA en mujeres

Ambos análisis capturan aproximadamente el 75 % de la variabilidad total de los datos en sus dos primeras dimensiones. Según los resultados, las variables parecen formar dos grupos principales: (y_3, y_4) y (y_1, y_2). En el primer componente, la varianza explicada es mayor para los hombres (57.50 %) que para las mujeres (53.56 %), mientras que en el segundo componente, la varianza explicada es mayor en mujeres (23.10 %) que en hombres (17.97 %).

Selección del número de componentes principales

Para determinar la cantidad de componentes principales a retener, se emplearon diversos criterios, como la regla de los valores propios mayores a 1, el método del codo y el método paralelo. Como resultado, se decidió conservar una sola componente principal en ambos casos.

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

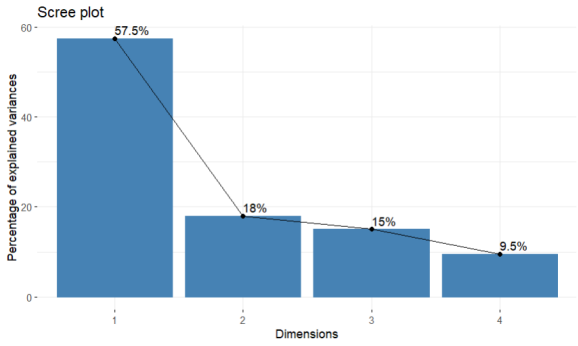

Figura 5: Método del codo en hombres

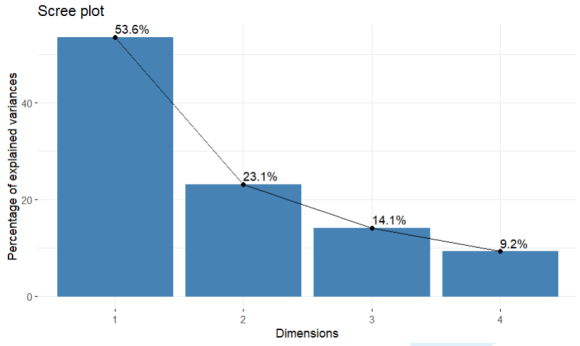

Figura 6: Método del codo en mujeres

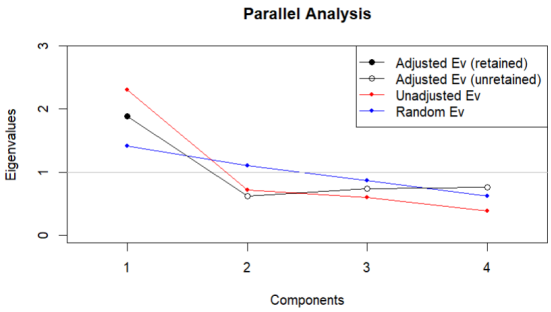

Figura 7: Método de paralelo en hombres

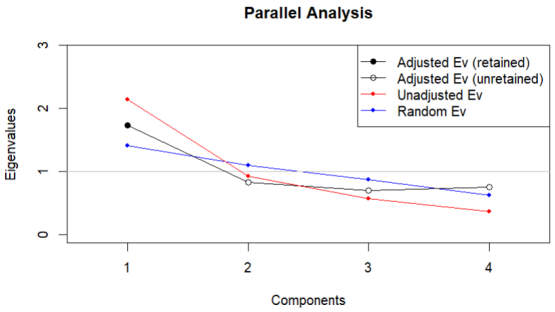

Figura 8: Método de paralelo en mujeres

Bajo la primera visión de considerar las componentes con valores propios mayores a 1, habría un poco de duda en el valor propio de la segunda componente en el PCA de mujeres que toma un valor de 0,924 (el más cercano a 1), pero tras analizar con los otros dos criterios vemos que se sigue sugiriendo seleccionar una única componente. Con el método del codo vemos una clara ruptura para hombres en la primera componente y más leve en el caso de las mujeres que mantienen más la forma deseada. También el análisis paralelo de Horn sugiere retener solo 1 componente en ambos casos, la línea de eigenvalores aleatorios cruza las otras líneas después del primer componente.

Interpretación de la Componente Principal

La primera componente se caracteriza por una oposición clara entre dos grupos de variables:

- y3 y y4 están positivamente correlacionadas entre sí y tienen una correlación positiva con Dim1.
- y1 y y2 están positivamente correlacionadas entre sí y tienen una correlación negativa con Dim1.

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

Los individuos que se encuentran en el lado positivo de Dim1 (derecha) tienden a tener valores altos en y3 y y4, y valores bajos en y1 y y2. Por otro lado, los individuos que se encuentran en el lado negativo de Dim1 (izquierda) tienden a tener valores altos en y1 y y2, y valores bajos en y3 y y4.

Contribución de las variables en la componente principal

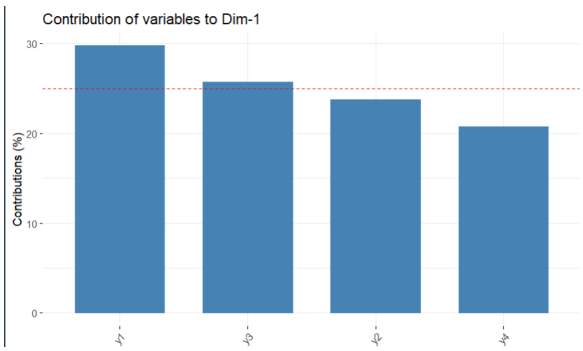

Figura 9: Contribución de cada variable en la primera componente principal para Hombres

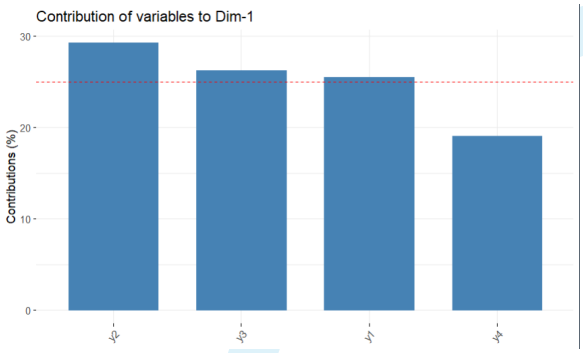

Figura 10: Contribución de cada variable en la primera componente principal para Mujeres

La variable dominante cambia: y1 en el primer conjunto vs y2 en el segundo, y4 consistentemente muestra la menor contribución, aunque sigue siendo relevante, sugiriendo que la primera componente principal captura un patrón robusto en los datos.

2. Análisis de Correspondencia

Para el análisis de correspondencia, primero se ingresan las tablas de contingencias en R:

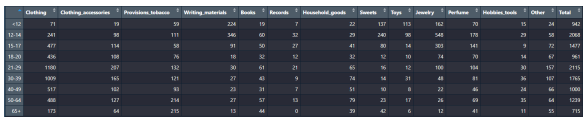

Figura 11: Tabla de contingencia para mujeres

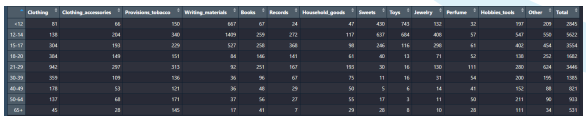

Figura 12: Tabla de contingencia para hombres

2.1. Análisis de correspondencia para mujeres

Cuantas dimensiones considerar:

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

Las dimensiones extraídas explican la varianza en los datos. La Dimensión 1 explica el 76.18 % de la varianza, seguida de la Dimensión 2 con el 15.07 %. En conjunto, las dos primeras dimensiones explican el 91.25 % de la varianza total de los datos, dado que es recomendable seleccionar un número suficiente de dimensiones que capturen una cantidad significativa de la varianza en los datos, sin incluir dimensiones que aporten muy poca varianza explicada. Seleccionaremos en el análisis estas 2 primeras dimensiones.

Principales hallazgos

Iniciando el análisis vemos que las categorías de edades (<12) y (12-14) tienen las contribuciones más altas a la Dimensión 1, lo que sugiere que estas categorías están fuertemente asociadas con las categorías de (Items) en esta dimensión. Las edades (+65) y (50-64) tienen contribuciones considerables a la Dimensión 2, lo que indica una asociación significativa con diferentes (Items) en esta dimensión.

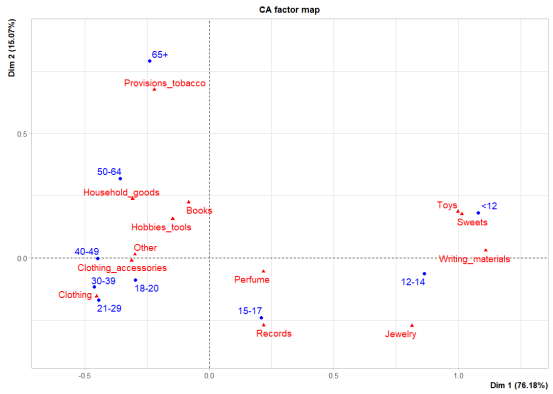

Figura 13: Análisis de correspondencia en mujeres

Al interpretar la dimensión 1, esta parece estar más relacionada con las diferencias por grupos de edad, especialmente con la edad más avanzada (65+), sugiriendo que las personas de 65 años o más tienen características o preferencias distintivas en comparación con otros grupos de edad. Las categorías más jóvenes, (15-17) y (18-20), tienen contribuciones más bajas en esta dimensión. La Dimensión 2 está más influenciada por el Ítem (Provisions tobacco), en edades (65+) y (50-64) mostrando una mayor asociación.

Podemos ver relaciones entre las categorías, por ejemplo, de que los robos registrados para edades (<12) y (12-14), tienden a ser de los Items (Toys, Sweets, Writing materials), para la edad (15-17) es más común el robo del item (Records), las edades (21-29) y (30-39) del Ítem (Clothing), la edad (50-64) con (Household goods) y finalmente la edad (65+) con (Provisions tobaco).

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

2.2. Análisis de correspondencia para hombres

Cuantas dimensiones considerar:

Similar al análisis anterior, la Dimensión 1 explica el 77.10 % de la varianza, seguida de la Dimensión 2 con el 11.45 %. En conjunto, las dos primeras dimensiones explican el 88.55 % de la varianza total de los datos. Seleccionaremos en el análisis estas 2 primeras dimensiones.

Principales hallazgos

Las categorías de edades (<12, 12-14, 20-29) y el Item (Toys) tienen las contribuciones más altas a la Dimensión 1. Por otro lado, la edad (15-17) y el Ítem (Records) tienen una contribución considerable a la Dimensión 2.

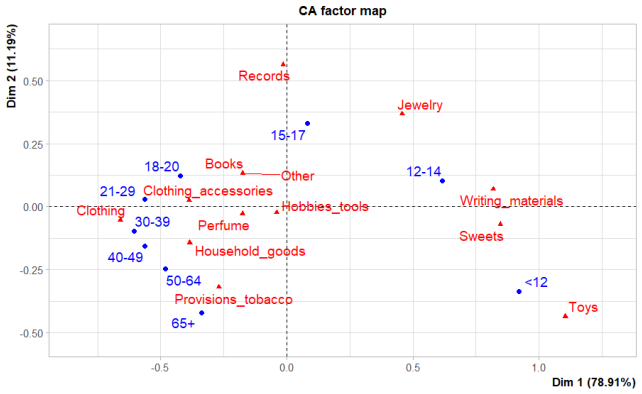

Figura 14: Enter Caption

Las edades más jóvenes (<12, 12-14) parecen estar más relacionadas con los Ítems (Toys, Sweets, Writing materials). Por otro lado, las edades (21-29, 30-39) parecen estar más asociadas con el Ítem (Clothing). Por otro lado, las edades más avanzadas, en particular (50-64, 65+) muestran una fuerte asociación con los Ítems (Household goods y Provisions tobacco), similar al caso de las mujeres.

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

2.3. Análisis de correspondencia conjunto

si bien las posiciones y las contribuciones de las categorías a cada dimensión varían, en general se mantiene una asociación similar entre las distintas categorías de forma general, como se ve en el gráfico. De igual forma, entre las dos primeras dimensiones se explica el 90.1 % de la varianza entre los datos.

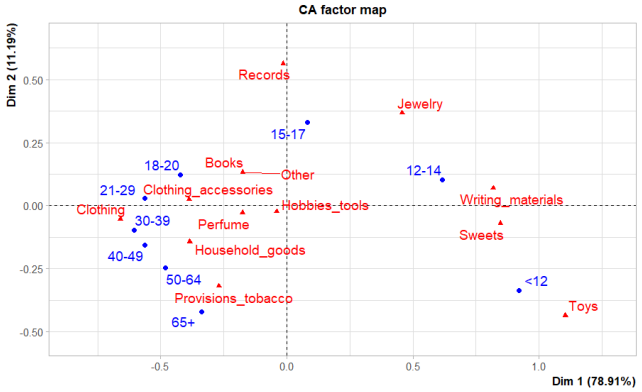

Figura 15: Análisis de correspondencia conjunto

las asociaciones más fuertes se mantienen, por ejemplo para la edad (<12) se encuentra más cercana al Ítem (Toys), la edad (12-14) a los Ítems (Sweets y Writing materials) o la edad (65+) a (Provisions tobacco), en general mantiene las relaciones dadas en los dos análisis de correspondencia por separado para hombres y mujeres

3. Código

Clik aqui para ir al repositorio en GitHub con el codigo

Nombre del archivo: CA_analysis_gender_age_groups.R

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec

Referencias

Heijden, P. G. M., Falguerolles, A., y De Leeuw, J. (2018). A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Loglinear Analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *38*(2), 249-273. https://doi.org/10.2307/2348058

Dirección: Avenida 12 de Octubre 1076 y Vicente Ramón Roca
Código postal: 170525 / **Teléfono:** (593-2) 299 1700
Quito - Ecuador / www.puce.edu.ec