Godinez Bravo Diego

Estadística Multivariada

Tarea 6

Ejercicio 1. Se clasificó de forma cruzada una muestra de 901 personas según tres categorías de ingresos y cuatro categorías de satisfacción laboral. Los resultados se dan en la siguiente tabla. Realice un análisis de correspondencia de estos datos. Interprete los resultados. ¿Pueden las asociaciones de los datos estar bien representadas en una dimensión? ¿La asociación de los datos en dos dimensiones es coherente con los resultados de la prueba χ^2?

	Job Satisfaction			
Income	Very dissatisfied	Somewhat dissatisfied	Moderately satisfied	Very satisfied
$< \$25,000$	42	62	184	207
$\$25,000 - \$50,000$	13	28	81	113
$> \$50,000$	7	18	54	92

Solución

Utilizando el lenguaje de programación R, se realizó un análisis de correspondencia para el conjunto de datos. Este análisis permitió visualizar la proyección conjunta de las filas y las columnas en un espacio bidimensional.

Al considerar una sola dimensión, las asociaciones entre los datos no pueden ser apreciadas en su totalidad. Para visualizar adecuadamente las relaciones entre los datos, se deben considerar dos dimensiones. Basándonos en las asociaciones de los datos en dos dimensiones (ver *gráfico derecho*), podemos afirmar que el primer componente separa las categorías de ingresos, donde los ingresos mayores a $\$25,000$ se desplazan hacia el lado izquierdo del gráfico (**Figura 1**).

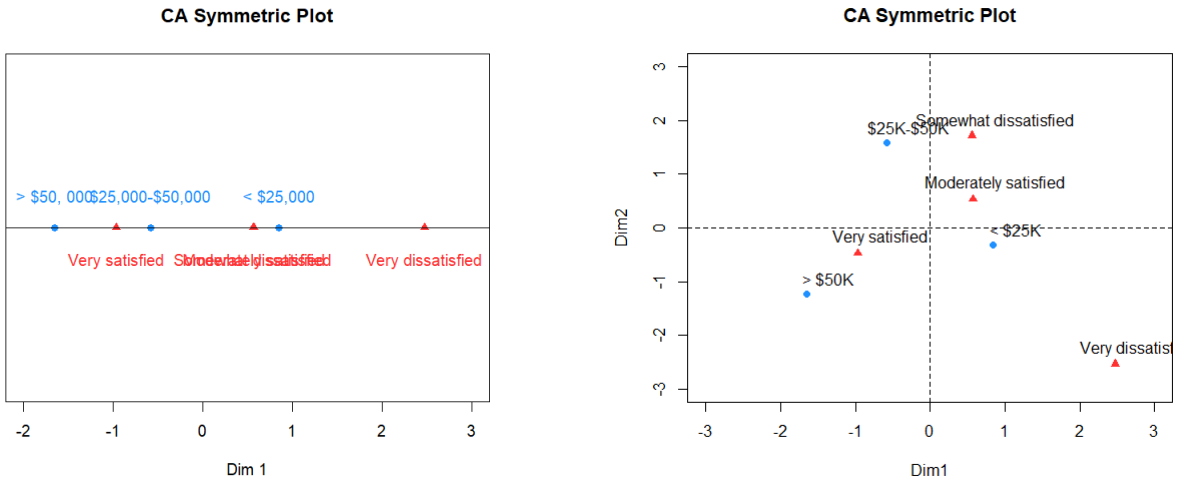

Figura 1. Representación gráfica de las proyecciones: en una dimensión (gráfico izquierdo), y en dos dimensiones (gráfico derecho).

Para evaluar la independencia entre las categorías de ingresos y las categorías de satisfacción laboral, se utilizó la prueba de χ^2. Los resultados obtenidos fueron los siguientes

data: satisfaction_data
X-squared = 10.407, df = 6, p-value = 0.1085

Considerando un nivel de significancia de $\alpha = 0{,}05$, concluimos que no hay suficiente evidencia para rechazar la hipótesis nula. Es decir, no podemos afirmar que existe una asociación significativa entre las categorías de ingresos y las categorías de satisfacción laboral, y cualquier patrón aparente en los datos podría ser atribuible al azar.

Ejercicio 2. El conjunto de datos **mundodes** representa 91 países en los que se han observado 6 variables: Razón de natalidad, Razón de mortalidad, mortalidad infantil, esperanza de vida en hombres, esperanza de vida en mujeres y PNB per cápita. Del conjunto de datos se ha tomado la esperanza de vida de hombres y de mujeres. Se han formado cuatro categorías tanto para la mujer como para el hombre. Se denotan por M1 y H1 a las esperanzas entre menos de 41 años a 50 años, M2 y H2, de 51 a 60 años, M3 y H3, de 61 a 70 años, y M4 y H4, para entre 71 a más de 80. La siguiente tabla de contingencia muestra las frecuencias de cada grupo.

Mujer/Hombre	H1	H2	H3	H4
M1	10	0	0	0
M2	7	12	0	0
M3	10	5	15	0
M4	10	0	23	19

Realice proyecciones por filas, por columnas y conjuntas de filas y columnas en un espacio de 2 dimensiones. Comprueba que en la proyección por filas las categorías están claramente separadas y que en el caso del hombre, las dos últimas categorías están muy cercanas. Comprueba en la proyección conjunta la cercanía de las categorías H3 con M3 y M4. ¿Qué se puede concluir de los resultados del análisis de correspondencia? ¿Son consistentes con los resultados de la prueba χ^2?

Solución

Haciendo uso del lenguaje de programación R, se realizó un análisis de correspondencia para el conjunto de datos, a partir del cual se visualizaron las proyecciones de las filas y columnas en un espacio de dos dimensiones.

El gráfico generado permite corroborar que, en la proyección por filas, las categorías están claramente separadas. Además, confirma que en la proyección por columnas, las categorías **H3** y **H4** se encuentran próximas entre sí (**Figura 2**).

Podemos interpretar que el componente 1 distingue entre la esperanza de vida de las personas menores de 60 años y la de las personas mayores de 60 años (**Figura 2**).

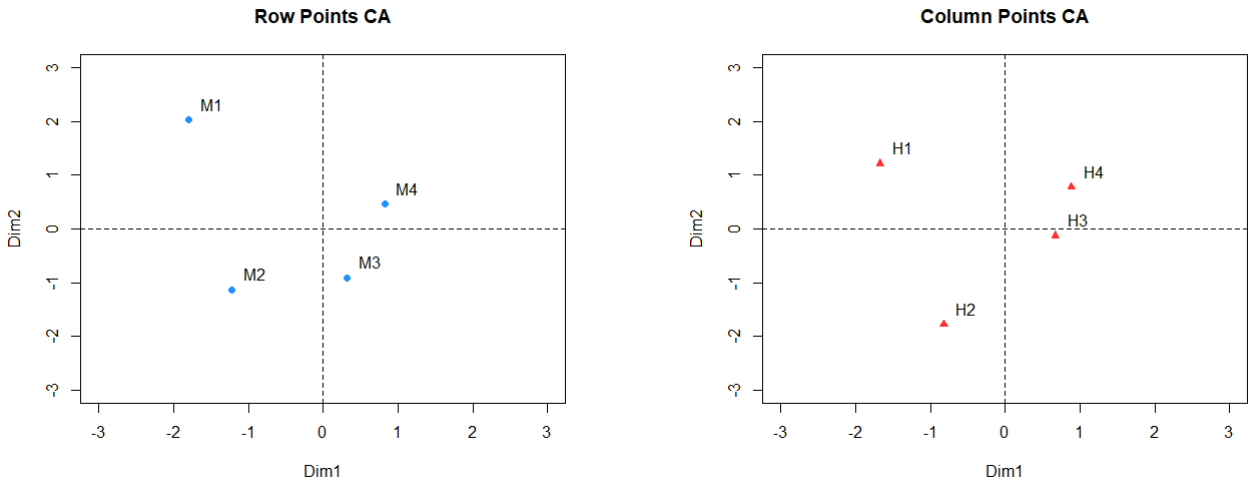

Figura 2. Representación gráfica de las proyecciones en dos dimensiones: por filas (gráfico izquierdo), y por columnas (gráfico derecho).

Al observar la proyecciones conjutas de filas y columnas, notamos la cercanía entre la categoría **H3** y las categorías **M3** y **M4** (**Figura 3**).

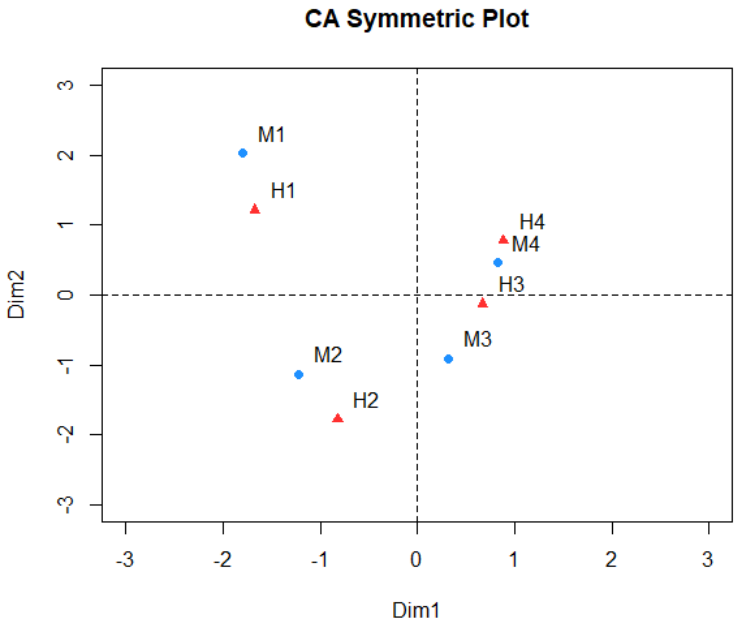

Figura 3. Representación gráfica del conjunto de datos en dos dimensiones.

Para evaluar la independencia entre los grupos de mujeres y hombres para los distintos rangos de edad se utilizó la prueba de χ^2. Se obtuvieron los siguientes resultados

$$\begin{gathered}\text{data: mundodes} \\ \text{X-squared} = 121.86, \quad \text{df} = 9, \quad \text{p-value} < 2.2\text{e-}16\end{gathered}$$

Considerando un nivel de significancia $\alpha = 0{,}05$, se concluye que estadísticamente las variables fila y columna se encuentran asociadas de manera significativa, es decir, se rechaza la hipótesis nula H_0.

Ejercicio 3. Realice un análisis de correspondencia de los datos de crímenes en Estados Unidos (archivo 'uscrime'). Este conjunto de datos consiste de 50 observaciones de 7 variables. Se reporta el número de crímenes en el año de 1985 para 50 estados de EUA, clasificados de acuerdo a 7 variables (X_3-X_9):

- X_1: land area
- X_2: population 1985
- X_3: murder
- X_4: rape
- X_5: robbery
- X_6: assault
- X_7: burglary
- X_8: larceny
- X_9: auto theft
- X_{10}: U.S. states region number
- X_{11}: U.S. states division number

a) Determine las contribuciones (inercias) para las 3 primeras dimensiones, ¿Cómo se puede interpretar la tercera dimensión? Intente identificar los estados con cada una de las 4 regiones a las que pertenecen, ¿Cree que las 4 regiones presentan diferente comportamiento respecto al tipo de crimen?

b) Agrupe los estados por región, de tal forma que cada región contenga la suma de las frecuencias de las 7 variables de los estados que la conforman. Realice de nuevo el análisis de correspondencia para las regiones y las variables, ¿Los resultados son congruentes con las conclusiones obtenidas en el inciso anterior?

Solución

Haciendo uso del lenguaje de programación R, se realizó un análisis de correspondencia para el conjunto de datos **uscrime**, a partir del cual se obtuvieron las contribuciones para las 3 primeras dimensiones. A continuación se resumen los valores de las inercias obtenidos

Inercias (valores propios)			
	1	2	3
Valor	0.0251	0.0126	0.0078
Porcentaje	49,18 %	24,71 %	15,45 %

Observamos que las dos primeras dimensiones representan aproximadamente $> 70\,\%$ de la inercia total. Por lo que una representación de las proyecciones en dos dimensiones debería de ser suficiente para mostrar las asociaciones existentes entre los datos.

Analizando los gráficos en tres dimensiones, observamos que la tercera componente divide algunas regiones de manera clare. Las regiones 1 y 3 se muestran en la parte inferior del eje z, mientras que la región 2 se encuentra en la parte superior del mismo. Sin embargo, no se observa un patrón aparente para el tipo de crimen, estos no muestran diferencias significativas a lo largo de la tercera dimensión. La excepción es el tipo de crimen 'robbery', el cual se aleja de las variables restantes (**Figura 4**, **Figura 5**).

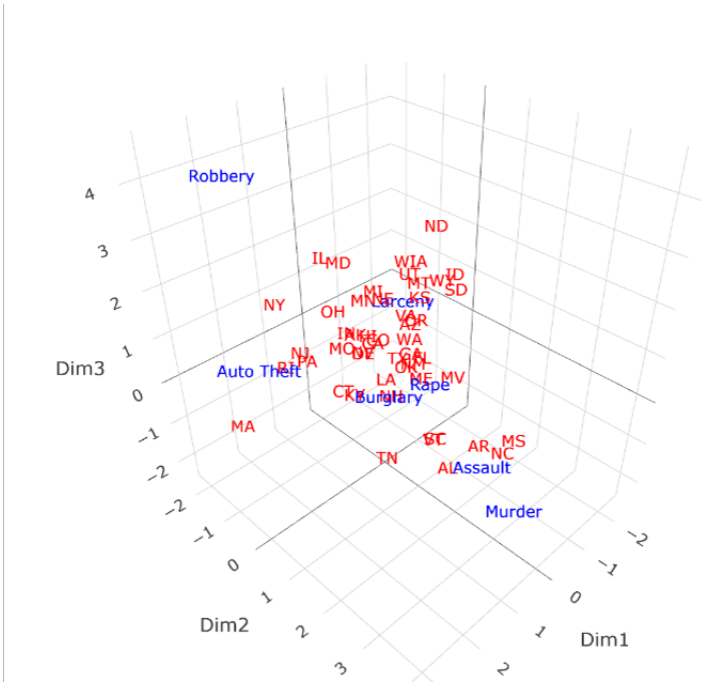

Figura 4. Representación gráfica del conjunto de datos **uscrimes** en tres dimensiones. Instancias etiquetadas utilizando el nombre del estado.

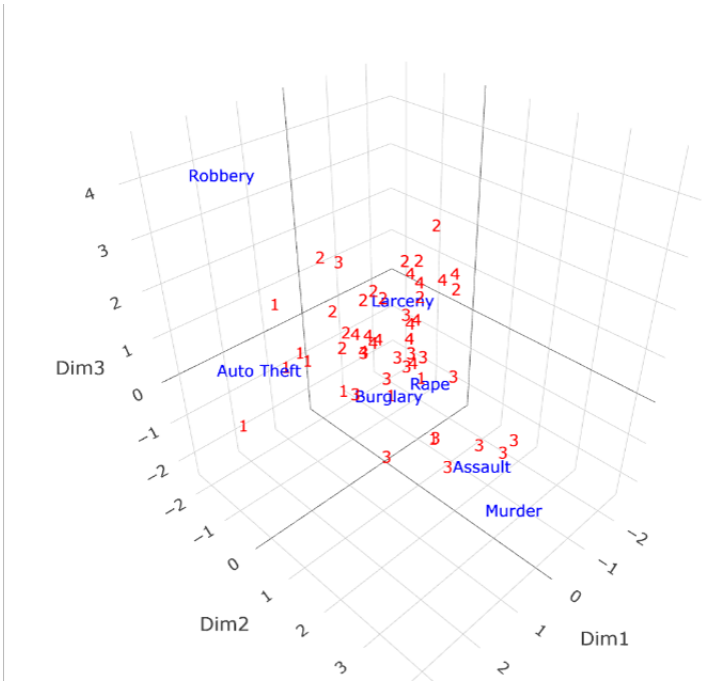

Figura 5. Representación gráfica del conjunto de datos **uscrimes** en tres dimensiones. Instancias etiquetadas utilizando el número de la región.

Solución

Al agrupar los datos por región, observamos que los resultados coinciden con lo estipulado previamente, las regiones se distribuyen a lo largo de la tercera dimensión. Sin embargo, en este caso, las regiones 1 y 4 se ubican en la parte inferior del eje, mientras que las regiones 2 y 3 se encuentran en la parte superior.

Tambíen se observa un patrón diferente para los tipos de crimen. Se observa una mayor separación de las variables a lo largo de la tercera dimensión (**Figura 6**).

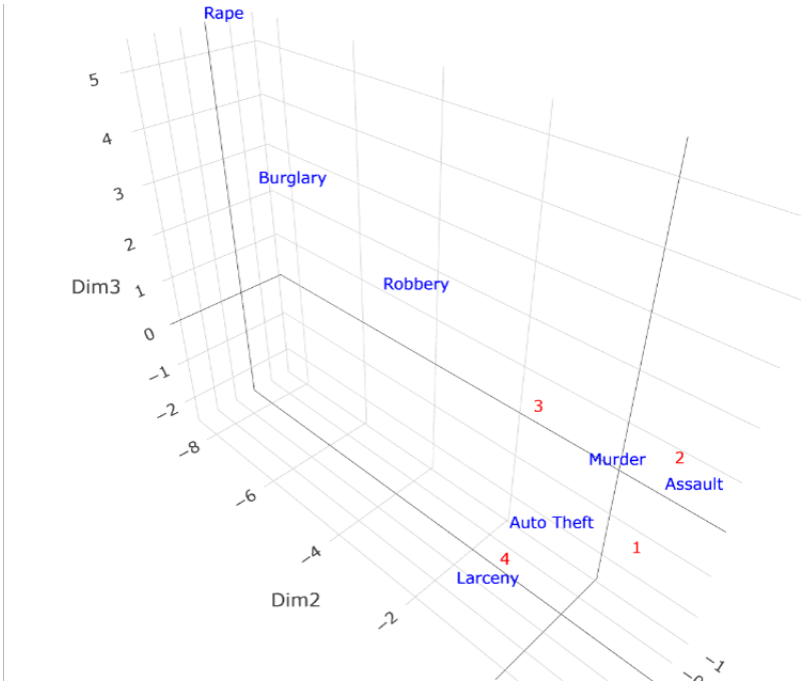

Figura 6. Representación gráfica del conjunto de datos **uscrimes** en tres dimensiones. Instancias agrupadas por región.

Ejercicio 4. Otra forma de derivar los resultados del análisis de correspondencia simple es encontrando una matriz $\hat{F}$ de dimensión $I \times J$ con rango reducido $s < \text{mín}(I, J)$ que aproxime a F minimizando el criterio de mínimos cuadrados ponderados:

$$tr\{D_r^{-\frac{1}{2}}(F - \hat{F})D_c^{-1}(F - \hat{F})' D_r^{-\frac{1}{2}}\}.$$

Usando el teorema de Eckart-Young, encuentre la matriz $\hat{F}$ que arroje la mejor aproximación de rango reducido a F en este sentido. Muestre que la mejor aproximación de 'rango 1' a F es la solución trivial $\hat{F} = rc'$.

Solución

Sea $A \in \mathbb{R}^{m \times n}$ una matriz real con $m \leq n$. Supongamos que

$$A = U\Sigma V^\top$$

es la descomposición en valores singulares de A. Recordemos que U y V son matrices ortogonales, y Σ es una matriz diagonal $m \times n$ con entradas $(\lambda_1, \lambda_2, \ldots, \lambda_m)$ tales que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$.

Afirmamos que la mejor aproximación de rango s a A en la norma espectral, denotada por $\|\cdot\|_2$, está dada por

$$B_s := \sum_{i=1}^{s} \lambda_i u_i v_i^\top$$

donde u_i y v_i denotan la i-ésima columna de U y V, respectivamente. Además minimiza

$$tr[(A - B)(A - B)']$$

de todas las matrices de tamaño $m \times k$ que tienen un rango no mayor que s.

Teorema de Eckart-Young.

Sea X una tabla con dimensión $I \times J$, cuyos elementos x_{ij} representan las frecuencias de aparición (o conteos) de la i-ésima categoría de X_1 y la j-ésima categoría de X_2. Se define n como el total de frecuencias en la tabla de datos X, y la matriz de frecuencias relativas $F = \{f_{ij}\}$ tal que

$$f_{ij} = \frac{x_{ij}}{n} \quad i = 1, 2, \ldots, I \quad j = 1, 2, \ldots, J$$

Se denota a la frecuencia relativa de la fila i de F como

$$r_i = \sum_{j=1}^{J} f_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n} \quad i = 1, 2, \ldots, I \quad r_{I\times 1} = F_{I\times J} 1_{J\times 1}$$

De manera similar se representa la frecuencia relativa de la columna j de F como

$$c_i = \sum_{i=1}^{I} f_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n} \quad j = 1, 2, \ldots, J \quad c_{J\times 1} = F_{J\times I} 1_{I\times 1}$$

Además, se consideran las matrices diagonales D_r y D_c definidas de la siguiente manera

$$D_r = diag(r_1, r_2, \ldots, r_I), \quad \text{y} \quad D_c = diag(c_1, c_2, \ldots, c_J)$$

Considerando lo anterior, definimos la matriz $B = D_r^{-1/2} F D_c^{-1/2}$. Sabemos que la mejor aproximación $\hat{B}$ está dada por los primeros s términos de la descomposición de valores singulares

$$D_r^{-1/2} \hat{F} D_c^{-1/2} = \sum_{k=1}^{s} \lambda_k u_k v_k'$$

Teorema de Eckart-Young,

donde

$$D_r^{-1/2} F D_c^{-1/2} v_k = \lambda_k u_k, \quad \text{y} \quad u_k' D_r^{-1/2} F D_c^{-1/2} = \lambda_k v_k'$$

y

$$|(D_r^{-1/2} F D_c^{-1/2})(D_r^{-1/2} F D_c^{-1/2})' - \lambda_k^2 I| = 0 \quad \text{para} \quad k = 1, \cdots, J$$

De manera que la mejor aproximación de B es

$$\hat{B} = D_r^{-1/2} \hat{F} D_c^{-1/2} = \sum_{k=1}^{s} \lambda_k u_k v_k'$$

Por lo tanto, podemos encontrar la aproximación para F de la siguiente manera

$$\hat{F} = D_r^{1/2} \hat{B} D_c^{-1/2} = \sum_{k=1}^{s} \lambda_k (D_r^{1/2} u_k)(D_c^{-1/2} v_k)'$$

Al considerar $u_k = D_r^{1/2} 1_{I\times 1}$ y $v_k = D_c^{1/2} 1_{J\times 1}$

$$D_r^{-1/2} F D_c^{-1/2} v_k = D_r^{-1/2} F D_c^{-1/2} (D_c^{1/2} 1_{J\times 1})$$

$$= D_r^{-1/2} F 1_{J\times 1}$$

$$= D_r^{-1/2} r$$

$$= D_r^{1/2} 1_{I\times 1}$$

$$= u_k$$

y

$$u_k' D_r^{-1/2} F D_c^{-1/2} = (D_r^{1/2} 1_{I\times 1})' D_r^{-1/2} F D_c^{-1/2}$$

$$= 1_{I\times 1}' F D_c^{-1/2}$$

$$= c' D_c^{-1/2}$$

$$= (D_c^{1/2} I_{J\times 1})'$$

$$= v_k'$$

Por lo que podemos afirmar que $(u_1, v_1) = (D_r^{1/2} 1_I, D_c^{1/2} 1_J)$ son vectores singulares asociados con los valores singulares $\lambda_k = 1$. Entonces

$$\hat{F} = (D_r^{1/2} \hat{B} D_c^{1/2}$$

$$= \sum_{k=1}^{s} \lambda_k (D_r^{1/2} u_k)(D_c^{1/2} v_k)'$$

$$= \lambda_1 (D_r^{1/2} u_1)(D_c^{1/2} v_1)'$$

$$= \lambda_1 (D_r^{1/2} D_r^{1/2} 1)(D_c^{1/2} D_c^{1/2} 1)'$$

$$= D_r 1_{I\times 1} 1_{J\times 1}^1 D_c$$

$$= rc'$$

De este modo podemos concluir que la mejor aproximación de rango 1 de F es la solución $\hat{F} = rc'$.