Análisis Multivariado para datos de la OPS

Luciano Garrido - Fiorella Lúngaro - Ana Vignolo

Curso Análisis Multivariado I 2022, Licenciatura en Estadística, Facultad de Ciencias Económicas y de Administración

Resumen

El presente trabajo fue elaborado en el marco de la unidad curricular Análisis Multivariado I con el objetivo de aplicar las distintas técnicas estudiadas a un conjunto de indicadores demográficos, socioeconómicos y del sistema sanitario publicados en 2019 para 45 países miembros de la Organización Panamericana de la Salud (OPS). En una primera parte, se realizó un Análisis Factorial, de forma de eliminar información redundante y reducir las dimensiones del problema. Posteriormente, se implementaron métodos de clusterización jerárquicos y se indagó sobre la posibilidad de obtener resultados similares a partir de una menor cantidad de variables. Para ello, se recurrió a la técnica de *Principal Feature Analysis*, basado en el algoritmo *K-means* de generación de *clusters*. Las variables surgidas a partir de este procedimiento fueron usadas, a su vez, para realizar un Análisis Discriminante logístico, con el objetivo de predecir si un país tendrá una esperanza de vida "alta" o "baja".

Presentación del problema

La base de datos original consta de 22 indicadores socioeconómicos, demográficos y del sistema sanitario para 48 países de América del Norte, América Latina y el Caribe. Para explorarlos, se dividió el análisis en dos partes:

- **Análisis Factorial:** Dado que solamente se contó con variables cuantitativas, se recurrió principalmente al Análisis de Componentes Principales (ACP). A partir del mismo, se buscó captar la mayor parte de la variabilidad de la nube de individuos en un primer eje principal.
- **Técnicas de clasificación:** Para agrupar los países, se utilizaron técnicas de clasificación supervisadas y no supervisadas. Así, se aplicaron métodos de clusterización jerárquicos. Además, se realizó un Análisis Discriminante logístico para predecir la probabilidad de que la esperanza de vida al nacer de los hombres en un cierto país se ubicara por encima de la mediana.

Datos faltantes

Para seis de las variables originales, el gran porcentaje de datos faltantes dificultó la posibilidad de aplicar cualquier procedimiento de imputación. Por otra parte, si se eliminaran todos los países con al menos un dato faltante, la base se reduciría a 17 observaciones. Esto equivale a algo más de un tercio de la matriz de datos original, por lo cual no parece una solución adecuada.
Como alternativa, se optó por eliminar todas las variables con más de un 20 % de datos faltantes, lo cual equivale a quitar todas las variables con 10 o más datos faltantes. De esta manera, se trabajó con una base de 16 variables.
Una vez hecho esto, se continuaron teniendo tres países con datos faltantes: Curazao, Islas Vírgenes de los Estados Unidos y Sain Maarten. Estos tres países del Caribe se caracterizan por su reducido tamaño y por no ser naciones independientes, con lo cual se consideró pertinente excluirlas del análisis. Así, se trabajó con 45 países.

Análisis Factorial

La técnica de ACP extrae la información esencial de los datos y la expresa a través de un nuevo conjunto de variables ortonormales denominadas componentes principales. Se busca maximizar la proporción de la variabilidad global captada por cada eje.
Por su parte, en la técnica de Análisis de Correspondencias Múltiples (ACM), se trabaja con variables categóricas. Así, se considera que los individuos serán similares en la medida que sus modalidades coincidan.

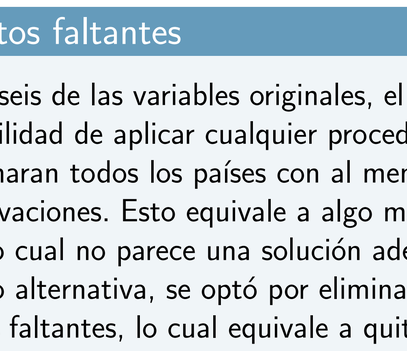

Figura 1: Análisis de Componentes Principales (individuo)

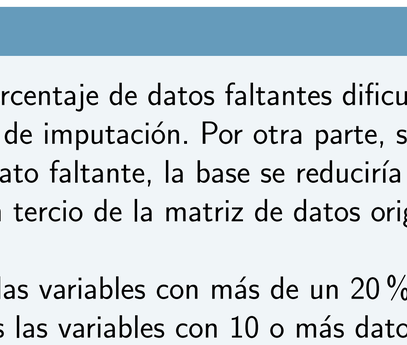

Figura 2: Análisis de Componentes Principales (variables)

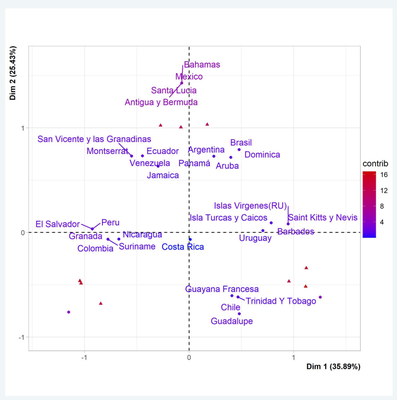

Figura 3: Análisis de Correspondencias Multiples (individuos)

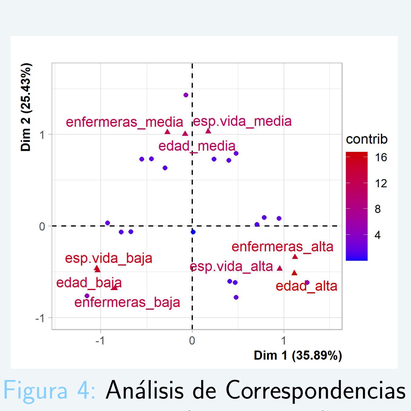

Figura 4: Análisis de Correspondencias Multiples (modalidades)

Conclusiones

- En el ACP, el primer eje principal logró explicar alrededor del 40 % de la variabilidad de la nube de puntos. Aun al considerarse una segunda dimensión, no se pudo llegar a captar el 60 % de la inercia global.
- Mientras que el método del vecino más lejano y de Ward sugieren trabajar con cinco clusters, el algoritmo del vecino más cercano se llega a tres grupos. En el método del vecino más cercano, Estados Unidos y Haití nunca llegan a juntarse con otras observaciones. Por su parte, los algoritmos del vecino más lejano y de Ward agrupan antes a las observaciones atípicas. Estados Unidos queda agrupado con otros países de gran tamaño como son Brasil y México.
- A través del método *Principal Feature Analysis*, se redujo la cantidad de variables consideradas de 16 a 4.
- Con las cuatro variables seleccionadas, se realizó un análisis discriminante para predecir si un país tendrá una esperanza de vida "alta" o "baja". Si bien las variables no resultaron ser individualmente significativas, el modelo presentó una capacidad predictiva aceptable.

Análisis de Clusters

El análisis de cluster es una técnica multivariada de clasificación no supervisada cuyo principal objetivo es agrupar individuos. Para ello, se busca que las observaciones dentro de un mismo cluster sean más parecidas entre sí que con respecto a las observaciones de los demás grupos. Para medir cuánto se asemejan dos individuos, es necesario definir una métrica o distancia. En este trabajo, se optó por utilizar la distancia euclídea y se aplicaron métodos de clusterización jerárquicos agregativos. Al inicio del proceso, cada observación corresponde a un cluster. En sucesivas etapas, se van agrupando las observaciones. Una vez que dos individuos se unen, no se separan hasta el final del proceso.

Reglas de detención para obtener la cantidad óptima de clusters

Cluster	Pseudot2	PseudoF	Silueta
2	0,57	5,04	0,45
3	−2,63	5,58	0,42
4	−1,85	4,52	0,25
5	−2,67	4,38	0,23
6	−3,02	4,13	0,24
7	−3,89	3,93	0,21
8	6,85	3,57	0,17
9	−1,33	4,47	0,21
10	−0,82	4,04	0,22
11	−1,98	4,21	0,25
12	3,99	4,15	0,22
13	−1,38	4,50	0,25
14	0,00	4,24	0,21
15	−0,53	3,91	0,24

Cuadro 1: Vecino más cercano

Cluster	Pseudot2	PseudoF	Silueta
2	−1,53	10,61	0,21
3	7,29	7,32	0,19
4	8,15	8,16	0,20
5	0,57	9,12	0,18
6	5,11	8,21	0,19
7	−0,59	8,38	0,20
8	−0,12	7,97	0,22
9	5,80	7,57	0,23
10	12,08	8,21	0,24
11	−1,51	9,92	0,27
12	3,05	9,86	0,29
13	−1,27	9,84	0,29
14	1,72	9,62	0,31
15	3,51	9,38	0,31

Cuadro 2: Vecino más lejano

Cluster	Pseudot2	PseudoF	Silueta
2	5,73	12,33	0,18
3	6,15	10,16	0,15
4	6,80	10,29	0,18
5	−1,88	9,57	0,17
6	4,53	10,02	0,20
7	4,89	10,22	0,22
8	0,18	9,75	0,18
9	−0,59	9,59	0,21
10	2,74	9,63	0,23
11	3,53	9,56	0,23
12	3,17	9,75	0,25
13	−0,63	9,62	0,25
14	2,91	9,60	0,27
15	−0,76	9,33	0,25

Cuadro 3: Ward

Dendrograma para cada algoritmo

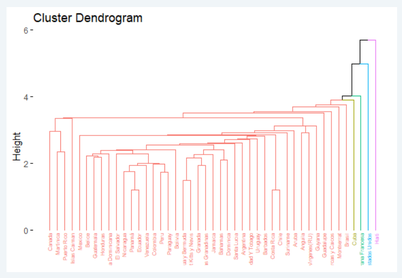

Figura 5: Vecino mas cercano

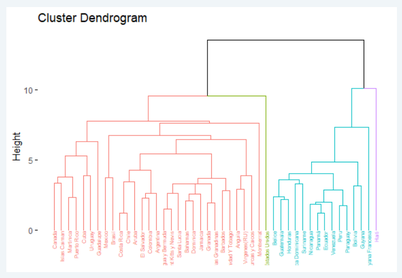

Figura 6: Vecino mas lejano

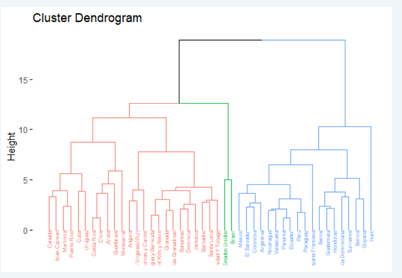

Figura 7: Ward

Análisis Discriminante

Mediante un Análisis Discriminante, se buscó predecir la probabilidad de que un país tuviera una esperanza de vida al nacer "alta" o "baja". Para ello, se categorizó la esperanza de vida de los hombres, *esp_vida_hombre*, en función de su mediana (73,7 años):

$$esp\_vida\_cat_i = \begin{cases} 1 & \text{si } esp\_vida\_hombres \geq 73,7 \\ 0 & \text{en otro caso} \end{cases} \quad (1)$$

Como variables explicativas, se consideraron cuatro indicadores obtenidos mediante el algoritmo de *Principal Feature Analysis*, a saber, la tasa de mortalidad neonatal (*mort_neonatal*), el porcentaje de la población menor de 15 años (*pob_menor_15*), la tasa de fecundidad adolescente (*tasa_fecund_adol*) y la cantidad anual de defunciones (*defunciones*).
Se rechazó la hipótesis nula en distintos tests de igualdad de medias de los dos grupos, por lo que tiene sentido realizar un Análisis Discriminante. Aunque no se rechaza la hipótesis nula de igualdad de varianzas, no se cumple el supuesto de multinormalidad. Por lo tanto, es preciso realizar un análisis logístico. En la figura 4, se presentan las estimaciones obtenidas.

Variable	Estimación	Error estándar	P-valor
Intercepto	6,98	2,42	0,00
mort_neonatal	−0,11	0,09	0,19
pob_menor_15	−0,21	0,11	0,06
tasa_fecund_adol	−0,01	0,02	0,48
defunciones	−0,00	0,00	0,59

Cuadro 4: Estimación del modelo logístico.

A pesar de que las variables no son individualmente significativas al 5 %, el p-valor correspondiente a la *deviance* del modelo estimado fue de 0,002. Por lo tanto, el modelo resulta ser globalmente significativo.

Figura 8: Curva ROC para el modelo estimado.

	Predicción 0	Predicción 1
Observado 0	0,76	0,24
Observado 1	0,29	0,71

Cuadro 5: Matriz de confusión en porcentajes por fila.

Dentro de la muestra, para un punto de corte óptimo de 0,55, se obtuvo una especificidad del 76 % una sensibilidad del 71 %. El error total fue del 27 %.

Referencias

[1] Blanco, J (2016). Introducción al Análisis Multivariado. Teoría y aplicaciones a la realidad latinoamericana. Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República.
[2] Greenacre, M (2017). Correspondence Analysis in Practice Chapman & Hall/CRC.
[3] Lu, Y and Cohen, I and Zhou, X and Tian, Q (2007). Feature Selection Using Principal Feature Analysis. ACM Multimedia, Augsburg, Germany.
[4] James, G and Witten, D and Hastie, D and Tibshirani, R (2013). An Introduction to Statistical Learning with Applications in R