Asignatura	Datos del alumno	Fecha
Estadística y R para Ciencias de la Salud	Adela Arias	19-06-2024
	Elena Díez García	
	Laura Montecino Fernández	
	Fernando Salgado Polo	

Análisis de un caso práctico en R

Conclusiones del análisis del caso práctico

Analizamos la base de datos de 131 alimentos y 19 nutrientes por PCA. A partir de este análisis, se pudo explicar el 11,2% de la varianza total con un primer componente principal (PC1) y el 3,5% con el segundo componente principal (PC2). El resto de los componentes explicaron menos del 3% de la varianza total (*Figura 1*). Para evaluar si el muestreo para realizar PCA fue correcto, realizamos el análisis de Kaiser-Meyer-Olkin (KMO), que mostró un MSA de 0.86. Dicho valor indica una adecuación correcta del muestreo para realizar PCA a partir de los alimentos y nutrientes.

A continuación, analizamos la contribución de los alimentos y nutrientes sobre PC1 y PC2. Para ello, evaluamos la correlación entre las variables y las dimensiones del PCA. Para PC1, los diez nutrientes que más contribuyeron fueron los siguientes: 4, 7, 2, 13, 10, 11, 12, 8, 14 y 19. En el caso de PC2, contribuyeron los nutrientes 6, 19, 16 y 15 y los alimentos 52, 27, 45, 20, 28 y 47 *(Figura 2 y 3).* Seguidamente, evaluamos la correlación de las variables originales con los componentes principales. Estas cargas no superaron en valor absoluto el 0,3 para PC1 o PC2, lo que sugiere una contribución moderada de los alimentos y nutrientes a los componentes principales.

En el análisis descriptivo (*Tabla 4*), analizamos la significación estadística de la prevalencia con algunas enfermedades y características fisiológicas de los grupos de individuos según PC1 y PC2. En el caso de las enfermedades, detectamos diferencias significativas entre los terciles de los grupos con distintos niveles de hipercolesterolemia en relación con PC1 ($p=0.026$) y PC2 ($p<0.001$). Asimismo, las diferencias en la prevalencia de hipertensión arterial fueron significativas para PC2 ($p<0.001$), pero no para PC1 ($p=0.2$). No obstante, estos dos componentes principales no proporcionaron diferencias significativas en la prevalencia de hipertrigliceridemia. Es reseñable que la agrupación por PC1 y PC2 no derivó en diferencias significativas con respecto al consumo de tabaco y los niveles de colesterol y HDL.

Por último, analizamos la prevalencia de la diabetes según los PC1 y PC2 junto con otras variables elegidas (sexo, edad, consumo de carne…) a través de un modelo de regresión logística *(Tabla 5).*

Prevalencia de la Diabetes según PC1 y PC2

- **PC1 Tercil 2:** Las personas que pertenecen a este componente tienen 1.88 veces más riesgo de desarrollar diabetes en comparación con los del tercil 1. Aunque este resultado no es estadísticamente significativo.
- **PC1 Tercil 3**: Las personas pertenecientes a este grupo tienen 1.16 veces más riesgo de padecer diabetes en comparación con el tercil 1. Este resultado no es estadísticamente significativo.
- **PC2 Tercil 2:** Las personas pertenecientes a este grupo tienen 1.88 veces más riesgo de tener diabetes de manera significativa en comparación con las del tercil 1.
- **PC2 Tercil 3**: Las personas pertenecientes a este grupo tienen 2.31 veces más riesgo de tener diabetes en comparación con el tercil 1. Este resultado es estadísticamente significativo.

Análisis de Otras Variables

- **Edad:**
  - Por cada año cumplido el riesgo de padecer diabetes aumenta de manera significativa.
- **Sexo**:
  - Las mujeres tienen menor riesgo de desarrollar diabetes en comparación con los hombres.
- **Hipertrigliceridemia**:
  - Menores de 25 años: Riesgo de diabetes es 2.83 veces mayor comparado con quienes nunca la desarrollan, aunque no es estadísticamente significativo.
  - Entre 25 y 64 años: Riesgo de diabetes es aproximadamente 3 veces mayor y este resultado es estadísticamente significativo.
  - Mayores de 65 años: Riesgo de diabetes es 2.57 veces mayor, pero no es estadísticamente significativo.
- **Hipercolesterolemia**:
  - Menores de 25 años: Riesgo de diabetes es 3.29 veces mayor comparado con quienes nunca la desarrollan.

Variables Sin Efecto Significativo

- El hábito tabáquico, el consumo de verduras y frutas no tienen un efecto significativo en la aparición de la diabetes.

Estos resultados muestran que ciertos factores como la edad al desarrollar hipertrigliceridemia e hipercolesterolemia, y la pertenencia al componente 2, están asociados con un mayor riesgo de desarrollar diabetes, mientras que otros factores como el consumo de verduras y frutas no tienen un efecto significativo.

Propuestas de mejora

Transformación de datos. Puesto que en este caso solo hemos empleado la estandarización de datos de la función *pca_result()*, habría que considerar otros métodos de transformación de datos que puedan aumentar la varianza explicada por los primeros componentes.

Inclusión de Más Componentes Principales. Al recoger las dos dimensiones que más contribuyen a la varianza total, no obtuvimos una representación elevada con dos dimensiones (14,7%), por lo que los datos no son óptimos para realizar un PCA. Convendría incluir una o dos dimensiones adicionales y determinar el efecto sobre el muestreo del PCA, así como los valores de significación para la prevalencia de enfermedades.

Análisis de Subgrupos. Se podría realizar PCA en subgrupos de datos específicos (por ejemplo, solo nutrientes o solo alimentos) para entender mejor las contribuciones individuales y su relación con las enfermedades.

Análisis Multivariante Adicional. Además de la regresión logística, se podría realizar un análisis discriminante para explorar relaciones causales y predictivas entre nutrientes, alimentos y enfermedades.