Tarea 2 - Análisis Multivariado

William Andrés Gómez Roa

2024-10-01

1. Datos de USairpollution

Los datos USairpollution del paquete HSAUR3 se refieren a la contaminación del aire en 41 ciudades de Estados Unidos. Se registraron las siguientes variables:

- **SO2**: Contenido de dióxido de azufre en el aire en microgramos por metro cúbico.
- **temp**: Temperatura media anual (en Fahrenheit).
- **manu**: Número de empresas manufactureras que emplean a 20 o más trabajadores.
- **popul**: Tamaño de la población (censo de 1970) en miles.
- **wind**: Velocidad media anual del viento en millas por hora.
- **precip**: Precipitación anual promedio en pulgadas.
- **predays**: Promedio de días con precipitación al año.

Los datos se recopilaron originalmente para investigar los determinantes de la contaminacíon, presumiblemente mediante una regresión del SO2 en las otras seis variables.

a. De acuerdo con los datos ¿usted trabajaría con la matriz de covarianzas o con la matriz de correlación? Justifique.

```
datos <- USairpollution[, -1]
```

```
S <- cov(datos)
R <- cor(datos)

S
```

```
##                temp        manu       popul        wind       precip    predays
## temp      52.239878   -773.9713   -262.3496  -3.6113537   32.8629884  -82.42616
## manu    -773.971341 317502.8902 311718.8140 191.5481098 -215.0199024 1968.95976
## popul   -262.349634 311718.8140 335371.8939 175.9300610 -178.0528902  645.98598
## wind      -3.611354    191.5481    175.9301   2.0410244   -0.2185311    6.21439
## precip    32.862988   -215.0199   -178.0529  -0.2185311  138.5693840  154.79290
## predays  -82.426159   1968.9598    645.9860   6.2143902  154.7929024  702.59024
```

```
R
```

```
##                temp        manu       popul        wind      precip     predays
## temp     1.00000000 -0.19004216 -0.06267813 -0.34973963  0.38625342 -0.43024212
```

```
## manu    -0.19004216  1.00000000  0.95526935  0.23794683 -0.03241688  0.13182930
## popul   -0.06267813  0.95526935  1.00000000  0.21264375 -0.02611873  0.04208319
## wind    -0.34973963  0.23794683  0.21264375  1.00000000 -0.01299438  0.16410559
## precip   0.38625342 -0.03241688 -0.02611873 -0.01299438  1.00000000  0.49609671
## predays -0.43024212  0.13182930  0.04208319  0.16410559  0.49609671  1.00000000
```

Ya que los datos tienen varianzas considerablemente diferentes y no están en la misma escala de comparación, es necesario hacer uso de la matriz de correlaciones (**R**) para eliminar el efecto causado por las diferentes escalas de medición entre las variables. Además, al utilizar la matriz de correlación, se facilita la comparación de la fuerza y dirección de la relación entre las variables, lo cual es especialmente importante en análisis como el Análisis de Componentes Principales (ACP), donde las variables con varianzas más grandes podrían influir desproporcionadamente en los resultados.

b. Verifique si el ACP para las últimas 6 variables es viable, en caso de serlo, ecriba las 6 componentes principales y sugiera un nombre para cada una.¿Con cuántas componentes trabajaría? Justifique.

Antes de realizar el ACP, es importante comprobar si los datos son adecuados para este análisis. Para esto podemos utilizar el test de esfericidad de Bartlett o la prueba de KMO (Kaiser-Meyer-Olkin) para evaluar la adecuación de la muestra.

```
library(psych)
```

Esfericidad de Bartlett

- **Hipótesis Nula (H0)**: La matriz de correlación es igual a la matriz de identidad (es decir, las variables son esféricamente distribuidas, lo que indica que no hay correlación entre ellas).
- **Hipótesis Alternativa (H1)**: La matriz de correlación no es igual a la matriz de identidad (es decir, hay correlación entre las variables).

```
cortest.bartlett(R, n= 41)
```

```
## $chisq
## [1] 159.2311
##
## $p.value
## [1] 3.501294e-26
##
## $df
## [1] 15
```

El resultado de la prueba de esfericidad de Bartlett es un p-valor < 0.05. Esto significa que podemos rechazar la hipótesis nula de que las variables son esféricas (es decir, que las correlaciones son iguales a cero). Esto sugiere que hay correlaciones significativas entre las variables y que es apropiado realizar un análisis factorial u otro análisis multivariado.

ACP : USairpollution

```
acp <- princomp(datos, cor = TRUE)

autovecotres <- acp$loadings
vectore_medias <- acp$center
autovecotres
```

```
##
## Loadings:
##         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## temp     0.330  0.128  0.672  0.306  0.558  0.136
## manu    -0.612  0.168  0.273 -0.137 -0.102  0.703
## popul   -0.578  0.222  0.350                -0.695
## wind    -0.354 -0.131 -0.297  0.869  0.113
## precip         -0.623  0.505  0.171 -0.568
## predays -0.238 -0.708        -0.311  0.580
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

Las cargas (loadings) indican cómo cada variable original contribuye a los componentes principales en el análisis:

- Componente 1: precip (-0.612) y popul (-0.578) tienen cargas negativas significativas, mientras que temp (0.330) y wind (-0.354) muestran contribuciones más bajas. Esto sugiere que las variables de precipitación y población están inversamente relacionadas con este componente.
- Componente 2: temp (0.672) y manu (0.306) tienen cargas positivas, lo que implica que un aumento en estas variables está asociado con un aumento en este componente, mientras que predays (-0.708) tiene una carga negativa, indicando una relación inversa.
- Componente 3: Las variables manu (0.558) y wind (0.505) contribuyen positivamente, mientras que temp (0.306) tiene una carga menor, lo que sugiere que este componente está relacionado con la manufactura y la velocidad del viento.
- Componente 4: La variable precip (0.869) tiene una carga muy alta, indicando que está fuertemente relacionada con este componente. Las otras variables tienen cargas cercanas a cero, lo que sugiere que este componente es principalmente influenciado por la precipitación.
- Componente 5: Aquí, la variable popul (0.350) tiene una carga positiva, pero es relativamente baja en comparación con las contribuciones de otros componentes. Las demás variables tienen cargas muy bajas.
- Componente 6: Este componente tiene cargas cercanas a cero para la mayoría de las variables, indicando que tiene poca relación con las variables originales, pero predays (-0.238) tiene una carga negativa. En términos de proporción de varianza, cada componente explica el 16.7% de la varianza total, sumando hasta el 100% acumulado con el último componente.

```r
summary(acp)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287
## Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511
## Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
##                            Comp.6
## Standard deviation     0.185599752
## Proportion of Variance 0.005741211
## Cumulative Proportion  1.000000000
```

Del anterior resumen podemos ver que los primeros tres componentes son los más informativos, mientras que los últimos componentes contribuyen poco a la explicación de la varianza en los datos.

- Desviación estándar: Cada componente tiene una desviación estándar que indica la cantidad de variabilidad que explica. Comp.1 tiene la mayor desviación estándar (1.4819), lo que sugiere que es el componente más importante. A medida que avanzamos a través de los componentes, la desviación estándar disminuye, con Comp.6 teniendo una desviación estándar muy baja (0.1856), indicando que aporta poco a la variabilidad total.
- Proporción de varianza: Comp.1 explica aproximadamente el 36.6% de la varianza total en los datos, seguido de Comp.2 (24.9%) y Comp.3 (23.2%). Estos tres primeros componentes en conjunto explican más del 84.7% de la varianza total, lo que sugiere que capturan la mayoría de la información presente en los datos. En contraste, Comp.6 solo explica el 0.57% de la varianza, lo que indica que su impacto es mínimo.
- Proporción acumulativa: La proporción acumulativa muestra cómo se acumula la varianza explicada a medida que se consideran más componentes. Después de Comp.1, el 61.6% de la varianza está explicada, y después de Comp.3, se alcanza el 84.8%. Con los primeros cuatro componentes, se explica el 97.5%, lo que resalta que se puede reducir significativamente la dimensionalidad sin perder mucha información.

2. Datos de Hijos

Los datos del archivo **Hijos.txt** corresponden a las medidas de la cabeza (en milímetros) de cada uno de los dos primeros hijos adultos en 25 familias. Estos datos fueron recopilados por **Frets (1921)**, y la pregunta de interés era si existe una relación entre las medidas de la cabeza de los pares de hijos.

a. Verifique si el ACP para la longitud de la cabeza (head1 y head2) es viable. En caso de ser viable, escriba las 2 componentes principales y sugiera un nombre para cada una. ¿Con cuántas componentes trabajaría? Justifique.

```
str(cabezas_hijos)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ head1: num  191 195 181 183 176 208 189 197 188 192 ...
##  $ head2: num  179 201 185 188 171 192 190 189 197 187 ...
```

Prueba de Esfericidad de Bartlett para la longitud las cabezas

```
R<-cor(cabezas_hijos)
cortest.bartlett(R, n= 25)
```

```
## $chisq
## [1] 15.82959
##
## $p.value
## [1] 6.931017e-05
##
## $df
## [1] 1
```

En este análisis, la prueba de esfericidad de Bartlett, con un valor p de $6.93 \times 10 - 5$ indica que los datos son adecuados para el análisis multivariado, lo que sugiere que las variables están correlacionadas de manera significativa.

ACP : Longitud de las cabezas

Se supone que las cabezas del hijo uno fueron medidas de la misma forma que las cabezas de los hijos 2, por lo que usaremos la matriz S para hacer el ACP.

```
acp <- princomp(cabezas_hijos, cor = FALSE)
summary(acp)
```

```
## Importance of components:
##                             Comp.1    Comp.2
## Standard deviation     12.6907660 5.2154059
## Proportion of Variance  0.8555135 0.1444865
## Cumulative Proportion   0.8555135 1.0000000
```

Dado que la primera componente (Comp.1) explica el 85.6% de la varianza, y la segunda (Comp.2) agrega un 14.4% adicional, se podría considerar trabajar únicamente con una componente (Comp.1) para simplificar el análisis, ya que captura la mayor parte de la información. Sin embargo, si se desea una comprensión más completa de los datos, podría considerarse mantener las dos componentes para investigar la variabilidad adicional que Comp.2 aporta.

b. Repita (a) para el ancho de la cabeza (breadth1 y breadth2)

```
str(ancho_cabezas)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ breadth1: num  155 149 148 153 144 157 150 159 152 150 ...
##  $ breadth2: num  145 152 149 149 142 152 149 152 159 151 ...
```

Prueba de Esfericidad de Bartlett para el ancho de las cabezas

```
R<-cor(ancho_cabezas)
cortest.bartlett(R, n= 25)
```

```
## $chisq
## [1] 15.68797
##
## $p.value
## [1] 7.469784e-05
##
## $df
## [1] 1
```

En este análisis, observando el valor p obtenido, la prueba de esfericidad de Bartlett indica que los datos son adecuados para el análisis multivariado, lo que sugiere que las variables están correlacionadas de manera significativa.

ACP : Ancho de las cabezas

Se supone que las cabezas del hijo uno fueron medidas de la misma forma que las cabezas de los hijos 2, por lo que usaremos la matriz S para hacer el ACP.

```r
acp <- princomp(ancho_cabezas, cor = FALSE)
summary(acp)
```

```
## Importance of components:
##                            Comp.1    Comp.2
## Standard deviation     9.0361670 3.7088659
## Proportion of Variance 0.8558225 0.1441775
## Cumulative Proportion  0.8558225 1.0000000
```

El Análisis de Componentes Principales (ACP) para el ancho de las cabezas se realiza utilizando la matriz de covarianza, asumiendo que las medidas han sido obtenidas de manera uniforme. Se identifican dos componentes principales: la Comp.1, que tiene una desviación estándar de 9.036 y explica el 85.6% de la varianza total, y se denomina "Variabilidad General del Ancho de la Cabeza", mientras que la Comp.2, con una desviación estándar de 3.709 y que explica el 14.4% de la varianza, se puede llamar "Diferencia Relativa entre Anchos de Hijos". Aunque es posible simplificar el análisis utilizando solo la Comp.1, considerar ambas componentes podría proporcionar una visión más completa de la variación en los datos, lo que dependerá del enfoque del análisis.

Los nombres asignados a las componentes principales reflejan su significado: "Variabilidad General del Ancho de la Cabeza" (Comp.1) se refiere a la primera componente, que captura el 85.6% de la varianza total, indicando que representa las características generales que afectan el ancho de las cabezas. En contraste, "Diferencia Relativa entre Anchos de Hijos" (Comp.2) destaca la variación adicional que explica el 14.4% restante de la varianza, señalando diferencias específicas entre los anchos de las cabezas de los hijos que no se reflejan en la variabilidad general.