Actividad 1.8 Análisis factorial II

Franco Mendoza Muraira A01383399

2023-11-17

1. Lea los datos y asegúrese que están límpios.

```
M = read.csv("cars93.csv")
M <- na.omit(M)
head(M)
```

```
##   V1 V2  V3  V4   V5   V6 V7 V8                          V9
## 1 18  8 307 130 3504 12.0 70  1 chevrolet chevelle malibu
## 2 15  8 350 165 3693 11.5 70  1          buick skylark 320
## 3 18  8 318 150 3436 11.0 70  1         plymouth satellite
## 4 16  8 304 150 3433 12.0 70  1              amc rebel sst
## 5 17  8 302 140 3449 10.5 70  1                ford torino
## 6 15  8 429 198 4341 10.0 70  1           ford galaxie 500
```

2. Reduzca la matriz de datos original a otra sólo de variables numéricas.

```
M1 = M[ , c(-8,-9)]
head(M1)
```

```
##   V1 V2  V3  V4   V5   V6 V7
## 1 18  8 307 130 3504 12.0 70
## 2 15  8 350 165 3693 11.5 70
## 3 18  8 318 150 3436 11.0 70
## 4 16  8 304 150 3433 12.0 70
## 5 17  8 302 140 3449 10.5 70
## 6 15  8 429 198 4341 10.0 70
```

3. Verifique si se cumple que los datos provienen de una población normal multivariada e interprete los resultados.

H_0 : Los datos multivariados siguen una distribución normal.

H_1 : Los datos multivariados no siguen una distribución normal.

```
result = mvn(M1, mvnTest = "mardia", alpha = 0.05)
result$multivariateNormality
```

```
##              Test        Statistic                p value Result
## 1 Mardia Skewness 980.306302410341 8.83518200960475e-153     NO
## 2 Mardia Kurtosis 18.0290946226721                      0     NO
## 3             MVN             <NA>                   <NA>     NO
```

```
result$univariateNormality
```

```
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    V1        3.5321  <0.001      NO
## 2 Anderson-Darling    V2       42.9803  <0.001      NO
## 3 Anderson-Darling    V3       17.4240  <0.001      NO
## 4 Anderson-Darling    V4       12.6748  <0.001      NO
## 5 Anderson-Darling    V5        7.2199  <0.001      NO
## 6 Anderson-Darling    V6        0.8379  0.0306      NO
## 7 Anderson-Darling    V7        5.1878  <0.001      NO
```

Viendo los resultados de la prueba de normalidad multivariada, concluimos que los datos no siguen una distribución normal multivariada, esto debido a la prueba de mardia en sesgo y curtosis saliendo con un pvalue muy bajo y por debajo del alpha de 0.05, por lo que no podemos rechazar la hipótesis nula.

4. Comprueben que hay suficiente correlación entre las variables dos a dos y en su conjunto:

a) Correlaciones por pares.

H_0 : No hay correlación significativa entre las variables 2 a 2.

H_1: Hay correlación significativa entre las variables 2 a 2.

```
alpha = 0.01
as.data.frame(cor(M1))
```

```
##            V1         V2         V3         V4         V5         V6         V7
## V1  1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  0.4233285  0.5805410
## V2 -0.7776175  1.0000000  0.9508233  0.8429834  0.8975273 -0.5046834 -0.3456474
## V3 -0.8051269  0.9508233  1.0000000  0.8972570  0.9329944 -0.5438005 -0.3698552
## V4 -0.7784268  0.8429834  0.8972570  1.0000000  0.8645377 -0.6891955 -0.4163615
## V5 -0.8322442  0.8975273  0.9329944  0.8645377  1.0000000 -0.4168392 -0.3091199
## V6  0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392  1.0000000  0.2903161
## V7  0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199  0.2903161  1.0000000
```

Podemos ver que hay altas correlaciones en las variables, muchas con correlaciones arriba de 80 y 90%, las únicas que no tienen valores tan altos como los mencionados son las V6 y V7.

```
corr = corr.test(M1,method = "pearson",alpha = alpha)
g = as.data.frame(corr$p)
g
```

```
##               V1            V2            V3            V4            V5
## V1  0.000000e+00  1.573660e-79  2.324899e-89  9.141586e-80 9.022944e-101
## V2  1.311384e-80  0.000000e+00 2.727135e-199 7.414214e-106 1.765217e-139
## V3  1.660642e-90 1.298636e-200  0.000000e+00 2.719600e-139 6.978168e-174
## V4  7.031989e-81 4.633884e-107 1.510889e-140  0.000000e+00 2.319390e-117
## V5 6.015296e-102 9.290616e-141 3.489084e-175 1.364347e-118  0.000000e+00
## V6  1.778576e-18  1.009001e-26  1.508540e-31  1.581886e-56  6.565616e-18
## V7  1.075794e-36  1.925683e-12  3.747957e-14  7.220175e-18  3.986518e-10
##              V6           V7
## V1 1.245003e-17 1.075794e-35
## V2 8.072006e-26 5.777048e-12
## V3 1.357686e-30 1.499183e-13
## V4 1.740075e-55 3.939369e-17
## V5 3.939369e-17 7.973035e-10
## V6 0.000000e+00 4.735475e-09
## V7 4.735475e-09 0.000000e+00
```

```
g<0.01
```

```
##      V1   V2   V3   V4   V5   V6   V7
## V1 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## V2 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## V3 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## V4 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## V5 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## V6 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## V7 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Podemos observar en esta tabla que todos los valores p son menores a nuestro alpha de 0.01, por lo que tenemos que rechazar la hipótesis nula, y por ende concluimos que si hay correlación significativa entre las variables 2 a 2.

b) Aplique la prueba de Kaiser-Meyer-Olkin (KMO) para correlaciones y compare el estadístico de prueba resultante con la escala siguiente y concluya.

0.00 a 0.49 inaceptable. 0.50 a 0.59 miserable. 0,60 a 0,69 mediocre. 0.70 a 0.79 medio. 0,80 a 0,89 meritorio. 0.90 a 1.00 maravilloso.

```
KMO(M1)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = M1)
## Overall MSA =  0.81
## MSA for each item =
##   V1   V2   V3   V4   V5   V6   V7
## 0.84 0.88 0.84 0.84 0.79 0.70 0.63
```

Podemos ver acorde a los resultados que nos dió la prueba KMO, que los primeras 4 variables tienen resultados meritorios acorde a la escala para sis correlaciones, las variables 5 y 6 nos dieron resultados medios, y por último la variable 7 nos dió nos dió un resultado mediocre. Esto nos dice que las primeras 4 tienen una alta relación entre ellas, mientras que las siguientes van bajando.

También nos dieron el resultado general MSA de 0.81 donde nos dice que el análisis factorial explica el 81% de la varianza de los datos.

6. Realicen un análisis de componentes principales y describan la proporción de varianza total explicada por cada componente.

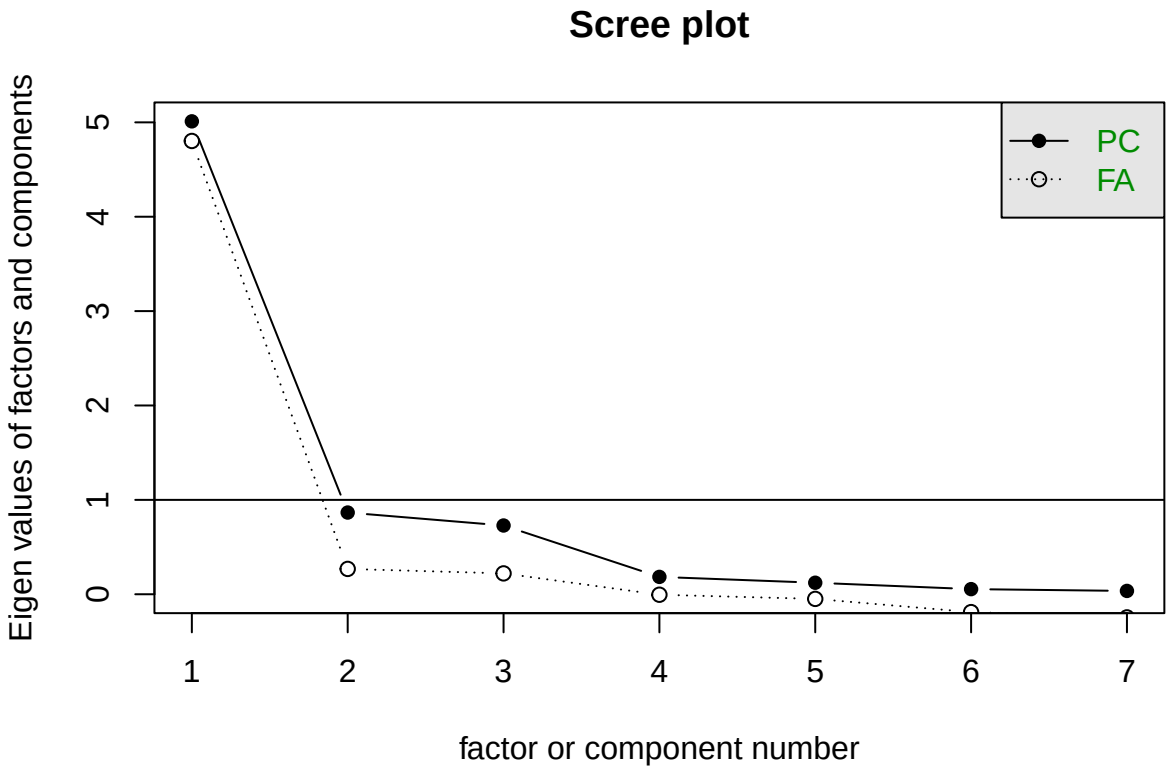
```
summary(prcomp(M1, scale = TRUE))
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.2384 0.9304 0.8535 0.42885 0.34917 0.23293 0.18786
## Proportion of Variance 0.7158 0.1237 0.1041 0.02627 0.01742 0.00775 0.00504
## Cumulative Proportion  0.7158 0.8395 0.9435 0.96979 0.98721 0.99496 1.00000
```

En este análisis de componentes principales podemos observar que el componente 1 ya explica el 71% de la varianza total, el componente 2 aumentando a la varianza total un 12%, la variable 3 un 10%, por lo que estas 3 componentes ya nos explican el 94.35%, las siguientes 4 nos explican menos de 2% cada una.

7. Con la ayuda del gráfico Scree y la tabla de distribución de la proporción acumulada de la varianza del punto anterior, decidan cuántos compontes son recomendables en este caso y que expliquen una mayoría de la varianza.

```
scree(cor(M1))
```



Viendo el gráfico scree, podemos ver que el codo de la gráfica se hace en el componente 3, por lo que eligiremos usar 3 componentes.

8. Realizar un análisis factorial según el método de máxima verosimilitud o componentes principales que convenga, así como dos modelos de rotación.

```
factanal=fa(r = cor(M1), nfactors = 3, fm = "ml")
```

```
## Loading required namespace: GPArotation
```

```
summary(factanal)
```

```
##
## Factor analysis with Call: fa(r = cor(M1), nfactors = 3, fm = "ml")
##
## Test of the hypothesis that 3 factors are sufficient.
## The degrees of freedom for the model is 3  and the objective function was  0.28
##
## The root mean square of the residuals (RMSA) is  0.02
## The df corrected root mean square of the residuals is  0.05
```

```
##
##  With factor correlations of
##       ML1   ML3   ML2
## ML1  1.00 -0.72 -0.48
## ML3 -0.72  1.00  0.37
## ML2 -0.48  0.37  1.00
```

```
quarti=fa(cor(M1), nfactors =3, rotate = "quartimax", fm ="ml")
summary(quarti)
```

```
##
## Factor analysis with Call: fa(r = cor(M1), nfactors = 3, rotate = "quartimax", fm = "ml")
##
## Test of the hypothesis that 3 factors are sufficient.
## The degrees of freedom for the model is 3  and the objective function was  0.28
##
## The root mean square of the residuals (RMSA) is  0.02
## The df corrected root mean square of the residuals is  0.05
```

```
quarti
```

```
## Factor Analysis using method =  ml
## Call: fa(r = cor(M1), nfactors = 3, rotate = "quartimax", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      ML1   ML2  ML3   h2    u2 com
## V1 -0.88 -0.02 0.47 1.00 0.005 1.5
## V2  0.95 -0.04 0.11 0.91 0.085 1.0
## V3  0.98 -0.07 0.12 0.98 0.018 1.0
## V4  0.90 -0.28 0.00 0.88 0.118 1.2
## V5  0.95  0.07 0.02 0.92 0.084 1.0
## V6 -0.50  0.87 0.01 1.00 0.005 1.6
## V7 -0.41  0.09 0.46 0.39 0.607 2.1
##
##                        ML1  ML2  ML3
## SS loadings           4.77 0.85 0.46
## Proportion Var        0.68 0.12 0.07
## Cumulative Var        0.68 0.80 0.87
## Proportion Explained  0.78 0.14 0.08
## Cumulative Proportion 0.78 0.92 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 3 factors are sufficient.
##
## df null model =  21  with the objective function =  8.91
## df of  the model are 3  and the objective function was  0.28
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.05
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 ML1  ML2  ML3
## Correlation of (regression) scores with factors   1.00 1.00 0.98
```

```
## Multiple R square of scores with factors          0.99 0.99 0.96
## Minimum correlation of possible factor scores     0.98 0.98 0.91
```

```
vari=fa(cor(M1), nfactors =3, rotate = "varimax", fm ="ml")
summary(vari)
```

```
##
## Factor analysis with Call: fa(r = cor(M1), nfactors = 3, rotate = "varimax", fm = "ml")
##
## Test of the hypothesis that 3 factors are sufficient.
## The degrees of freedom for the model is 3  and the objective function was  0.28
##
## The root mean square of the residuals (RMSA) is  0.02
## The df corrected root mean square of the residuals is  0.05
```

En todos los casos, el análisis sugiere que el modelo con 3 factores es adecuado. Esto se determina por la prueba de hipótesis de que tres factores son suficientes, junto con medidas de ajuste como la función objetivo, el RMSA (root mean square of the residuals) y el RMSA corregido. Los valores bajos en estos errores nos indican que hubo un buen ajuste.

La tabla de correlaciones entre factores muestra cómo están relacionados entre sí los factores extraídos. Los valores en esta tabla representan las correlaciones entre los factores. En el resultado, se observa que, por ejemplo, el factor ML1 tiene una correlación negativa moderada con ML3 y ML2, y así sucesivamente para los otros factores.

Ambos modelos de rotación (quartimax y varimax) muestran resultados similares en términos de la adecuación del modelo y la estructura de correlación entre factores. Esto sugiere que la estructura subyacente de los factores no cambia significativamente con la rotación. La diferencia principal entre estos modelos suele encontrarse en la interpretación de las cargas factoriales, pero en este caso, los resultados no parecen variar mucho entre las rotaciones.

9. Escriban las composiciónes lineales de las variables en función de los factores, según su análisis. Interprete factores e identifique variables que más influyen.

Usaremos el modelo varimax.

```
##
## Combinaciones lineales
```

```
##
## Factor 1 = -0.6016473 * V1 + 0.8850417 * V2 + 0.9100609 * V3 + 0.7420694 * V4 + 0.873624 * V5 + -0.25
##
## Factor 2 = 0.8850417 * V1 + 0.7420694 * V2 + -0.257181 * V3 + 0.7850875 * V4 + -0.2835003 * V5 + -0.3
##
## Factor 3 = 0.9100609 * V1 + -0.257181 * V2 + -0.2727054 * V3 + -0.3686343 * V4 + 0.1290894 * V5 + -0
```

Composiciones lineales de las variables en función de los factores:

Factor 1: - Variables que más influyen positivamente: V3, V5, V4, V2 - Variables que más influyen negativamente: V1, V6, V7

Factor 2: - Variables que más influyen positivamente: V1, V4, V7, V2 - Variables que más influyen negativamente: V3, V5, V6

Factor 3: - Variables que más influyen positivamente: V1, V6, V7, V3 - Variables que más influyen negativamente: V4, V2, V5

Interpretación de los factores:

- **Factor 1:** Está influenciado positivamente por variables como V3 (más fuerte), V5, V4 y V2, mientras que V1, V6 y V7 tienen una influencia negativa más fuerte en este factor.
- **Factor 2:** Se ve positivamente influenciado por variables como V1 (más fuerte), V4, V7 y V2, mientras que V3, V5 y V6 tienen una influencia negativa más fuerte en este factor.
- **Factor 3:** Muestra una fuerte influencia positiva de variables como V1 (más fuerte), V6, V7 y V3, mientras que V4, V2 y V5 tienen una influencia negativa más fuerte en este factor.

10. ¿Qué difierencias esenciales encuentran entre Componentes principales y Análisis factorial?

- **Componentes Principales:**
  - Método estadístico para reducir la dimensionalidad de los datos, busca maximizar la varianza de las variables originales.
  - No considera relaciones entre variables, busca nuevos ejes que representen la mayor variabilidad de los datos.
  - No asume estructura de factores latentes o subyacentes.
- **Análisis Factorial:**
  - Se enfoca en identificar factores latentes que expliquen las relaciones entre las variables observadas.
  - Busca comprender la estructura subyacente de los datos, identificando factores comunes que explican las correlaciones entre variables.
  - Ofrece cargas factoriales que indican cómo cada variable está relacionada con los factores identificados.