Actividad1_8

Ricardo Kaleb Flores Alfonso

2024-10-09

```
library(psych)
library(ggplot2)
```

```
##
## Adjuntando el paquete: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(polycor)
```

```
##
## Adjuntando el paquete: 'polycor'
```

```
## The following object is masked from 'package:psych':
##
##     polyserial
```

```
library(ggcorrplot)
```

1) Lea los datos y asegúrese que están límpios.

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

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

```
M2 <- M[ , purrr::map_lgl(M, is.numeric)]
```

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

```
library(MVN)
result = mvn(M2, mvnTest = "mardia", alpha = 0.05)
result$multivariateNormality
```

```
##              Test        Statistic               p value Result
## 1 Mardia Skewness 1168.24697920534 2.79443758861836e-171     NO
```

```
## 2 Mardia Kurtosis 13.7998161496409                    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
## 8 Anderson-Darling    V8       58.6897  <0.001      NO
```

Dados los valores obtenidos para las variables, se observa que el resultado de normalidad dado el valor de p obtenido, se rechaza la hipotesis nula, por lo que las variables no tienen una distribución normal. De igual manera el test de normalidad multivariada da un valor de p menor a 0.05, por lo que se rechaza la normalidad de la distribución multivariada.

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

a) Correlaciones por pares.

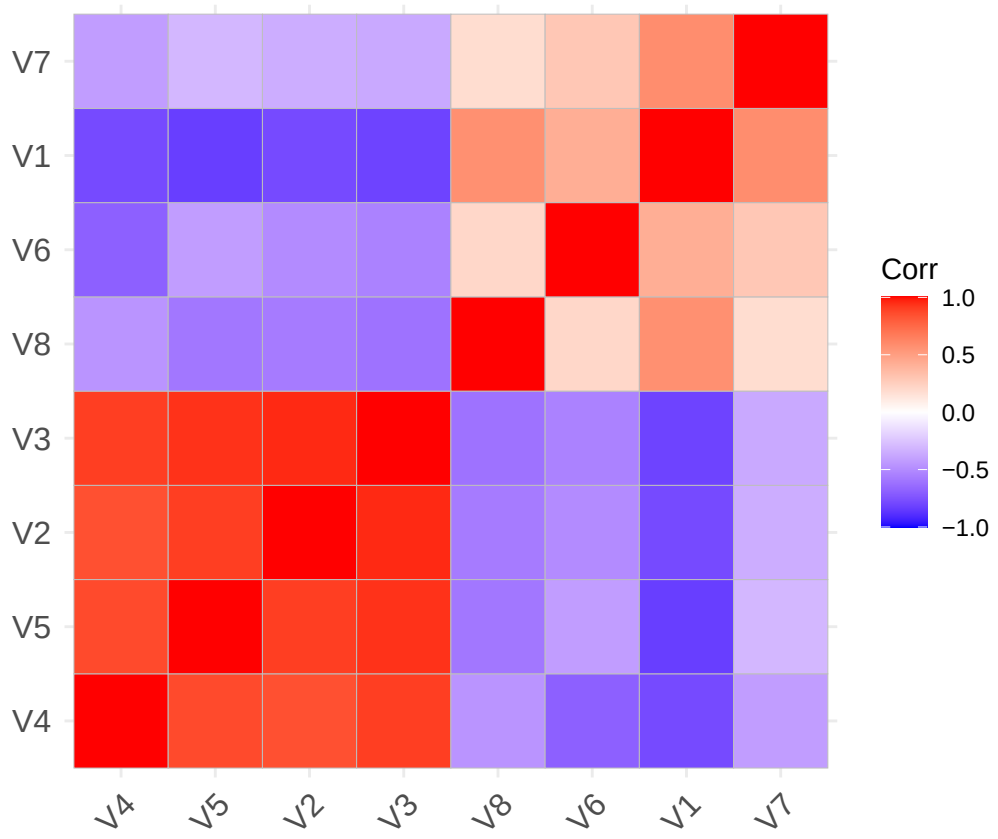
```
corr.test(M2,adjust="none")
```

```
## Call:corr.test(x = M2, adjust = "none")
## Correlation matrix
##       V1    V2    V3    V4    V5    V6    V7    V8
## V1  1.00 -0.78 -0.81 -0.78 -0.83  0.42  0.58  0.57
## V2 -0.78  1.00  0.95  0.84  0.90 -0.50 -0.35 -0.57
## V3 -0.81  0.95  1.00  0.90  0.93 -0.54 -0.37 -0.61
## V4 -0.78  0.84  0.90  1.00  0.86 -0.69 -0.42 -0.46
## V5 -0.83  0.90  0.93  0.86  1.00 -0.42 -0.31 -0.59
## V6  0.42 -0.50 -0.54 -0.69 -0.42  1.00  0.29  0.21
## V7  0.58 -0.35 -0.37 -0.42 -0.31  0.29  1.00  0.18
## V8  0.57 -0.57 -0.61 -0.46 -0.59  0.21  0.18  1.00
## Sample Size
## [1] 392
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##    V1 V2 V3 V4 V5 V6 V7 V8
## V1  0  0  0  0  0  0  0  0
## V2  0  0  0  0  0  0  0  0
## V3  0  0  0  0  0  0  0  0
## V4  0  0  0  0  0  0  0  0
## V5  0  0  0  0  0  0  0  0
## V6  0  0  0  0  0  0  0  0
## V7  0  0  0  0  0  0  0  0
## V8  0  0  0  0  0  0  0  0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Se observa que todos los valores cuentan con valor de p menor a 0.1, por lo que todas las variables muestran una correlación cierta entre ellas. De igual manera la correlación entre variables es fuerte, sin embargo esta

varia entre una correlación fuerte positiva o negativa.

```
mat_cor <- hetcor(M2)$correlations
ggcorrplot(mat_cor,hc.order = T)
```



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 1 aceptable para el análisis factorial

```
R = cor(M2)
K = KMO(R)
cat("El valor del estadístico es: ", K$MSA)
```

```
## El valor del estadístico es:  0.8188964
```

Es valor obtenido de K es 0.81, por lo que la correlación que existe es aceptable para un analísis factorial y será una buena opción realizarlo.

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

```
summary(prcomp(M2, scale = TRUE))
```

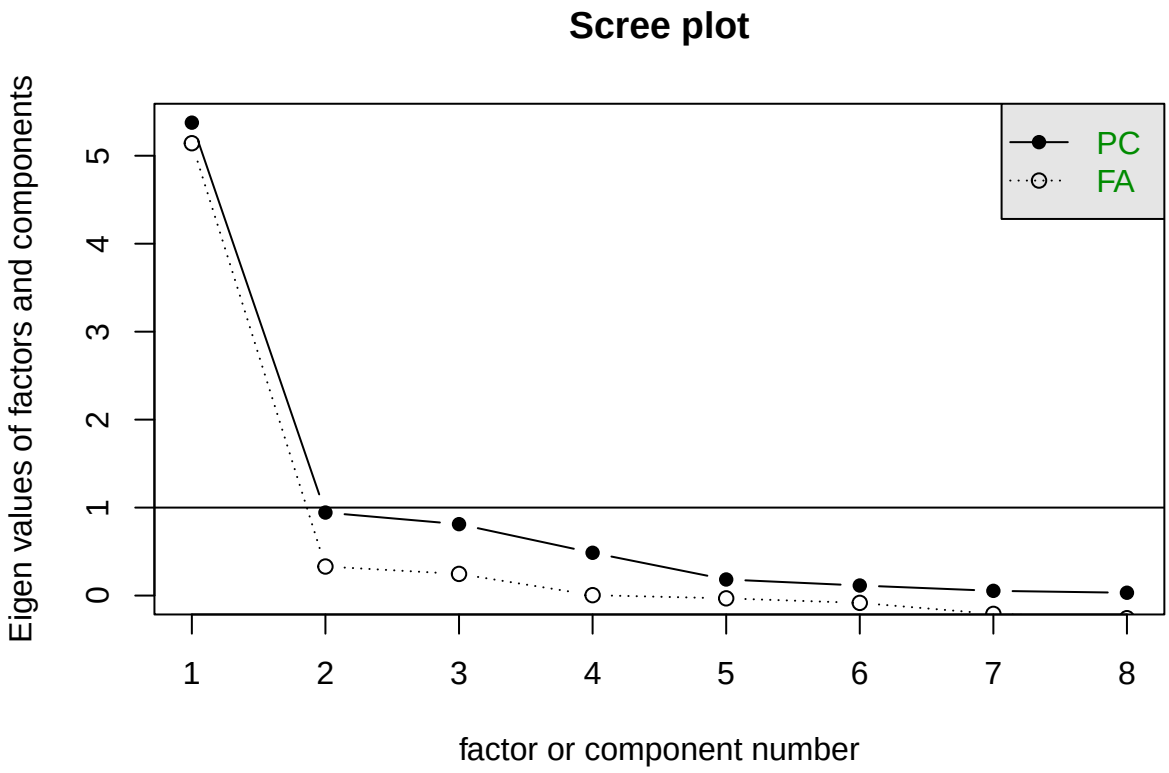
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
```

```
## Standard deviation     2.319 0.9714 0.9009 0.69725 0.42758 0.33812 0.23140
## Proportion of Variance 0.672 0.1180 0.1015 0.06077 0.02285 0.01429 0.00669
## Cumulative Proportion  0.672 0.7899 0.8914 0.95217 0.97502 0.98931 0.99600
##                           PC8
## Standard deviation     0.1788
## Proportion of Variance 0.0040
## Cumulative Proportion  1.0000
```

Se observa como los primeros 4 componentes explican el 95% de la varianza, esto nos permite reducir el analísis a estos cuatro componentes y explicar de una buena manera los datos.

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(R)
```



En este caso el gráfico de Scree, permite saber que el mejor número de componentes que se recomienda aceptar es hasta el segundo, sin embargo el número tres tambien podria ser tomado en cuenta.

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

```
quartimax = fa(R, nfactors =2, rotate = "quartimax", fm ="ml")
```

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

```
quartimax
```

```
## Factor Analysis using method =  ml
## Call: fa(r = R, nfactors = 2, rotate = "quartimax", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      ML1   ML2   h2    u2 com
## V1 -0.83  0.00 0.69 0.313 1.0
## V2  0.95  0.14 0.92 0.080 1.0
## V3  0.99  0.08 0.98 0.022 1.0
## V4  0.94 -0.34 1.00 0.005 1.3
## V5  0.94  0.06 0.90 0.105 1.0
## V6 -0.57  0.45 0.53 0.474 1.9
## V7 -0.39  0.15 0.17 0.825 1.3
## V8 -0.59 -0.29 0.44 0.560 1.5
##
##                        ML1  ML2
## SS loadings           5.16 0.45
## Proportion Var        0.65 0.06
## Cumulative Var        0.65 0.70
## Proportion Explained  0.92 0.08
## Cumulative Proportion 0.92 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  28  with the objective function =  9.54
## df of  the model are 13  and the objective function was  0.89
##
## The root mean square of the residuals (RMSR) is  0.06
## The df corrected root mean square of the residuals is  0.08
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                 ML1  ML2
## Correlation of (regression) scores with factors   1.00 0.96
## Multiple R square of scores with factors          0.99 0.91
## Minimum correlation of possible factor scores     0.98 0.83
```

En el caso de la rotación mediante máxima verosimilitud nos da los resutlados donde esta función nos permite obtener la cantidad de factores necesarios para explicar la varianza, este muestra que 2 es el número necesario de factores. De igual manera , de igual manera los scores obtenidos tienen valores altos de correlación.

```
varimax = fa(R, nfactors =2, rotate = "varimax", fm ="ml")
varimax
```

```
## Factor Analysis using method =  ml
## Call: fa(r = R, nfactors = 2, rotate = "varimax", fm = "ml")
```

```
## Standardized loadings (pattern matrix) based upon correlation matrix
##      ML2   ML1   h2    u2 com
## V1 -0.65 -0.51 0.69 0.313 1.9
## V2  0.83  0.48 0.92 0.080 1.6
## V3  0.83  0.55 0.98 0.022 1.7
## V4  0.53  0.85 1.00 0.005 1.7
## V5  0.78  0.53 0.90 0.105 1.8
## V6 -0.18 -0.70 0.53 0.474 1.1
## V7 -0.21 -0.36 0.17 0.825 1.6
## V8 -0.65 -0.13 0.44 0.560 1.1
##
##                        ML2  ML1
## SS loadings           3.19 2.43
## Proportion Var        0.40 0.30
## Cumulative Var        0.40 0.70
## Proportion Explained  0.57 0.43
## Cumulative Proportion 0.57 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  28  with the objective function =  9.54
## df of  the model are 13  and the objective function was  0.89
##
## The root mean square of the residuals (RMSR) is  0.06
## The df corrected root mean square of the residuals is  0.08
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                 ML2  ML1
## Correlation of (regression) scores with factors   0.97 0.98
## Multiple R square of scores with factors          0.94 0.96
## Minimum correlation of possible factor scores     0.88 0.93
```

En el caso de la rotación mediante componentes principales nos da los resutlados donde esta función nos permite obtener la cantidad de factores necesarios para explicar la varianza, este muestra que 2 es el número necesario de factores. De igual manera , de igual manera los scores obtenidos tienen valores altos de correlación.

Sin embargo la rotación obtenida por maxima verosimilitud, explica de mejor manera los datos, asi como tiene un mayor minimo de correlación entre los factores.

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

```
varimax$loadings
```

```
##
## Loadings:
##    ML2    ML1
## V1 -0.650 -0.514
## V2  0.832  0.477
## V3  0.825  0.545
```

```
## V4  0.528  0.846
## V5  0.783  0.532
## V6 -0.177 -0.703
## V7 -0.212 -0.360
## V8 -0.650 -0.133
##
##                  ML2   ML1
## SS loadings    3.186 2.430
## Proportion Var 0.398 0.304
## Cumulative Var 0.398 0.702
```

V1= -0.650*Factor2 - 0.514 Factor1 V2= 0.832*Factor2 + 0.477 Factor1 V3= 0.825*Factor2 + 0.545 Factor1 V4= 0.528*Factor2 + 0.846 Factor1 V5= 0.783*Factor2 + 0.532 Factor1 V6= -0.177*Factor2 - 0.703 Factor1 V7= -0.212*Factor2 - 0.360 Factor1 V8= -0.650*Factor2 - 0.133 Factor1

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

PCA se enfoca en maximizar la varianza total explicada por los componentes, esto trata de reducir la dimensionalidad del conjunto de datos sin preocuparse por las relaciones latentes entre las variables. Sin embargo análisis factorial busca descubrir factores latentes que expliquen las correlaciones entre las variables.

En PCA, los componentes principales son construidos para explicar la mayor cantidad posible de la varianza total de los datos. Esto hace que las combinaciones lineales obtenidas sean optimizadas para capturar la mayor cantidad de variación.

Los factores son elegidos para explicar las correlaciones entre las variables, no necesariamente la varianza total. esto nos permite agrupar las variables en menos factores.

Los componentes no siempre tienen una interpretación clara desde el punto de vista de las relaciones entre las variables originales.Por otro lado en el anlísis factorial el objetivo es encontrar factores que tengan una interpretación más directa.