Actividad 1.7 Análisis Factorial

Franco Mendoza Muraira A01383399

2023-11-15

```
##
## Attaching package: 'polycor'

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

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

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

##
## Attaching package: 'GPArotation'

## The following objects are masked from 'package:psych':
##
##     equamax, varimin
```

El problema del Lago

Realicen el análisis factorial, discutiendo y comentando los resultados obtendidos de:

```
##   Tamano Temperatura Proteinas Oxigeno LnProteinas LnTamano
## 1    7.5        13.8      65.9     2.8    4.188138 2.014903
## 2    5.8        13.8      20.3     2.8    3.010621 1.757858
## 3    8.0        13.8     136.6     2.8    4.917057 2.079442
## 4    8.0        13.8      70.5     2.8    4.255613 2.079442
## 5   10.0        13.8     117.8     2.8    4.768988 2.302585
```

1. Obtener la matriz de correlaciones y la matriz de valores p de significancia por pares. Interpreten los resultado en equipo.

```r
corr.test(df,adjust="none")
```

```
## Call:corr.test(x = df, adjust = "none")
## Correlation matrix
##             Tamano Temperatura Proteinas Oxigeno LnProteinas LnTamano
## Tamano        1.00        0.23      0.63   -0.34        0.84     0.97
## Temperatura   0.23        1.00     -0.22    0.34       -0.08     0.22
## Proteinas     0.63       -0.22      1.00   -0.37        0.78     0.57
## Oxigeno      -0.34        0.34     -0.37    1.00       -0.38    -0.31
## LnProteinas   0.84       -0.08      0.78   -0.38        1.00     0.86
## LnTamano      0.97        0.22      0.57   -0.31        0.86     1.00
## Sample Size
## [1] 90
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##             Tamano Temperatura Proteinas Oxigeno LnProteinas LnTamano
## Tamano        0.00        0.03      0.00       0        0.00     0.00
## Temperatura   0.03        0.00      0.04       0        0.46     0.04
## Proteinas     0.00        0.04      0.00       0        0.00     0.00
## Oxigeno       0.00        0.00      0.00       0        0.00     0.00
## LnProteinas   0.00        0.46      0.00       0        0.00     0.00
## LnTamano      0.00        0.04      0.00       0        0.00     0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Correlaciones:

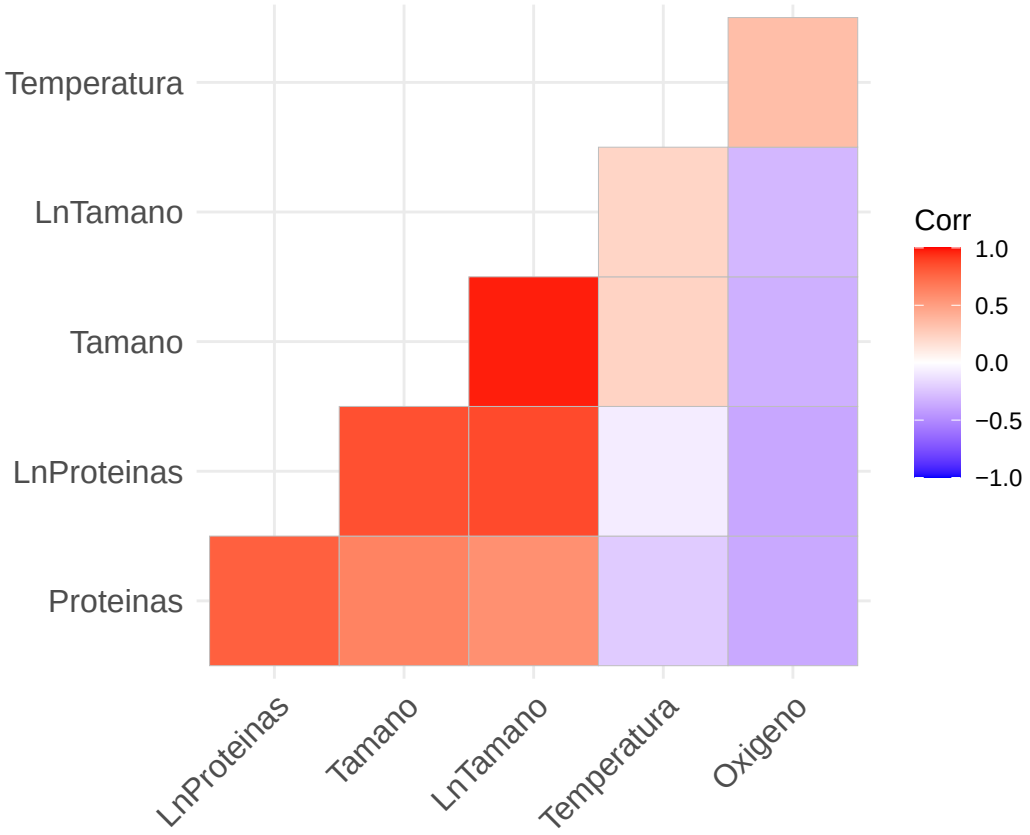
La variable "LnTamano" tiene una fuerte correlación positiva con "Tamano" (0.97), indicando una relación lineal positiva fuerte entre el logaritmo natural del tamaño y el tamaño. "Oxigeno" y "Proteinas" tienen una correlación negativa moderada (-0.37), sugiriendo una relación lineal negativa moderada entre el oxígeno y las proteínas.

Valores p:

La correlación entre "Tamano" y "Proteinas" es estadísticamente significativa (p = 0.00), indicando que la relación entre estas dos variables no es el resultado del azar. La correlación entre "Temperatura" y "LnProteinas" no es significativa (p = 0.46), lo que sugiere que la relación entre la temperatura y el logaritmo natural de las proteínas podría ser el resultado del azar. En resumen, estos resultados indican relaciones significativas y la fuerza de esas relaciones entre las variables en tu conjunto de datos.

2. Hacer una gráfica de la matriz de correlaciones. Hacer un comentario entre todos los del equipo de la gráfica.

```r
mat_cor <- hetcor(df)$correlations #matriz de correlación policórica
ggcorrplot(mat_cor,type="lower",hc.order = T)
```



Podemos ver en la gráfica que las variables con mas alta correlación son las de Tamano y Lntamano. Otras variables notorias con correlaciones altas son las de Lntamano y LnProteinas, Tamano y LnProteinas, y Proteinas y LnProteinas. Se puede ver debido a los nombres de las variables que las correlaciones tienen sentido, y que las variables con correlaciones solo son LnProteinas, Tamano, LnTamano y Proteinas todas entre sí.

3. Aplicar una prueba de correlación conjunta a los datos para verficar si es aplicable el Análisis Factorial y concluir.

```
check_sphericity_bartlett(df)
```

```
## # Test of Sphericity
##
## Bartlett's test of sphericity suggests that there is sufficient significant correlation in the data
```

```
# Para obtener valor del estadístico (chi-cuadrada), del parámetro (grados de libertad) y del valor p):
b = check_sphericity_bartlett(df)
b$chisq
```

```
## [1] 557.8585
```

```r
b$p
```

```
## [1] 3.135832e-109
```

```r
b$dof
```

```
## [1] 15
```

En la prueba de bartlett obtuvimos el estadistico de chi cuadrado de 557.68, los grados de libertad de 15, y el valor p de 3.1358e-19.

El resultado del valor p ($p < 0.001$) indica que hay suficiente evidencia para rechazar la hipótesis nula de que la matriz de correlación es una matriz de identidad. En otras palabras, hay correlaciones significativas en los datos. Por lo tanto, los resultados sugieren que es apropiado realizar un Análisis Factorial en los datos.

4. Otra prueba para, para comprobar si el análisis factorial es viable, y muy citada, es la prueba KMO. Aplíquela a estos datos, ¿contradice los resutados del inciso anterior?

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

```
## El valor del estadístico es:  0.6297281
```

La prueba KMO varía entre 0 y 1, y sabemos que los valores más cercanos a 1 indican una mejor adecuación para un análisis factorial. Con nuestro valor de KMO de 0.6297, esto sugiere una adecuación moderada para el análisis ya que con valores encima de 0.6 ya se considera adecuado, aunque valores más altos serían mejores.

5. Si los datos pasaron la prueba de los puntos anteriores 3 y 4, hacer un análisis factorial usando el criterio de máxima verosimilitud y el de mínimo residuo

```r
R = cor(df)
modelo1 = fa(R, nfactors =2, rotate = "none", fm = "mle")    # de máxima verosimilitud
modelo2 = fa(R, nfactors = 2, rotate = "none", fm = "minres") # Modelo de mínimo residuo
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```r
M1_commd = sort(modelo1$communality,decreasing = T)
M2_commd= sort(modelo2$communality,decreasing = T)
cbind(M1_commd,M2_commd)
```

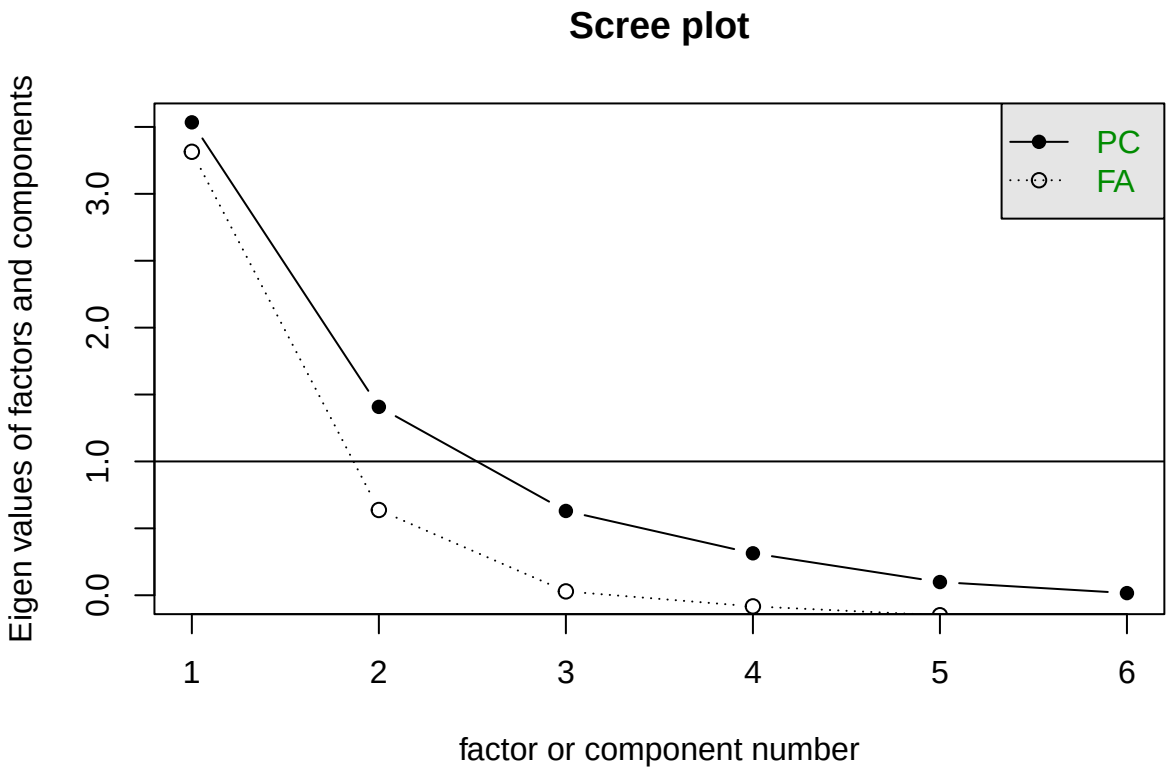
```
##              M1_commd  M2_commd
## LnTamano    0.9950455 0.9748176
## Tamano      0.9558283 0.9238582
## LnProteinas 0.9181230 0.9030553
## Proteinas   0.7708383 0.8453996
## Temperatura 0.3963254 0.6029645
## Oxigeno     0.2071029 0.2835476
```

Aquí se muestran las comunalidades, o la proporción de la varianza total de una variable observada que puede ser explicada por los factores del modelo, de las variables de los datos, podemos ver que las variables con las comunalidades más altas son las que tenían más relación entre sí en el análisis de correlación, mientras que las variables de Tempreatura y Oxigeno, las que tenían menos relación entre sí muestran comunalidades bajas.

6. Determine el número de factores adecuado según el criterio del gráfico de Cattell

```
R = cor(df)
scree(R) # se grafican los valores propios de R, y del análisis Factoria
```

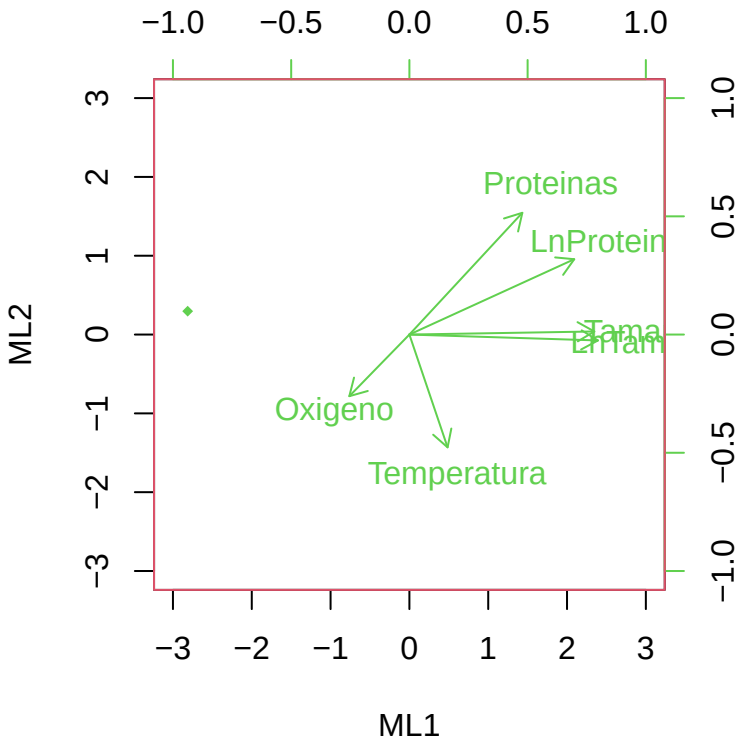
```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

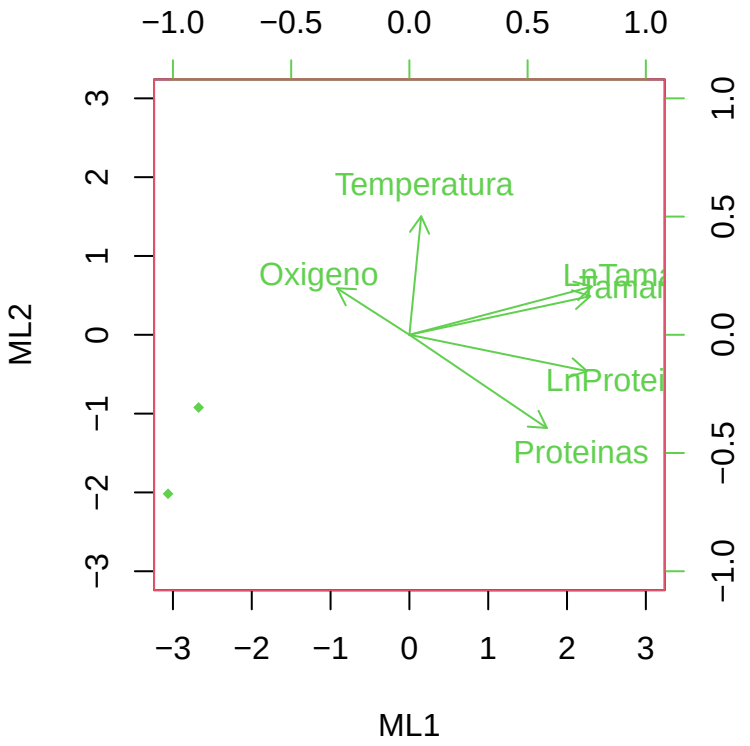


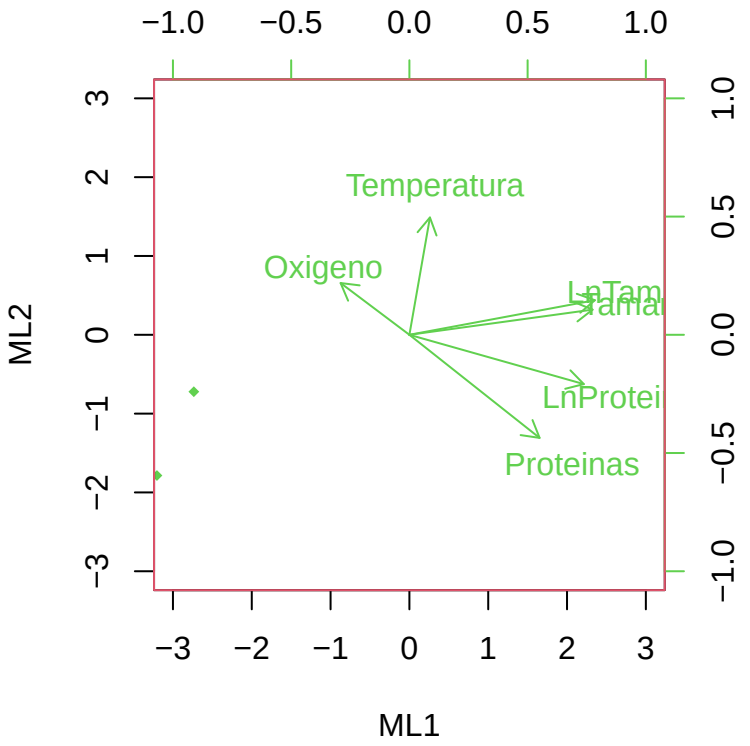
En la gráfica de scree, podemos ver que el numero de factores donde se hace el codo con los eigen valores de los factores es con 3 factores, por lo que se tomarían como los factores adecuados.

7. Realicen los gráficos correspondientes a la rotación Varimax y quartimax de los datos e interpreten en equipo los resultados.

```
R = cor(df)
rot = c("none", "varimax", "quartimax")
bi_mod = function(tipo){
biplot.psych(fa(df,nfactors = 2,fm="mle",rotate = tipo),main = "",col=c(2,3,4),pch = c(21,18),group = di
sapply(rot,bi_mod)
```




−1.0
−0.5
0.0
0.5
1.0
3
2
1
0
−1
−2
−3
ML2
Temperatura
Oxigeno
LnTama
LnProtei
Proteinas
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
ML1




```
## $none
## NULL
##
## $varimax
## NULL
##
## $quartimax
## NULL
```

En las gráficas podemos ver las variables respecto a los factores que se extrayeron del análisis factorial. También podemos llegar a la misma conclusión de las otras pruebas, en cuanto a la similitud en las variables de LnTamano y Tamano, y LnProteinas y Proteinas, mientras que las otras 2 variables de Oxigeno y Temperatura se encuentran alejadas de las demás, aunque relativamente cerca la una de la otra.

8. ¿Qué pueden concluir? ¿Resultó razonable para este caso el modelo de análisis factorial? Expliquen.

Considerando todos los análisis realizados hasta ahora, parece razonable aplicar el modelo de análisis factorial a este conjunto de datos. La prueba de Bartlett indicó que hay correlaciones significativas entre las variables, respaldando la idoneidad del análisis factorial. Aunque la prueba KMO mostró una adecuación moderada, lo cual sugiere que podría haber margen para mejorar la estructura del modelo.

El análisis factorial reveló relaciones consistentes entre ciertas variables, como la fuerte correlación entre "LnTamano" y "Tamano", así como entre "LnProteinas" y "Proteinas". Estas observaciones se reflejaron en las comunalidades más altas para estas variables.

Además, al utilizar el criterio de la gráfica de Cattell, se sugiere la inclusión de tres factores para modelar mejor los datos, lo cual se corroboró visualmente en las rotaciones Varimax y Quartimax. Estas rotaciones también mostraron la agrupación y separación de variables, destacando la similitud entre ciertas variables y la divergencia de otras.

En resumen, el análisis factorial parece un enfoque razonable para comprender la estructura de estos datos.