UNIVERSIDAD NACIONAL DE COLOMBIA

PREGRADO EN ESTADISTICA

Departamento de Estadística
Facultad de Ciencias

— Introducción al Análisis Multivariado —

Analisis de Factores Multivariados

Datos: Data calidad al dormir
tarea número 8

Integrantes:

Ricardo William Salazar Espinal C.C. 1017219472

Medellín, Colombia

Medellin, septiemnre 13 de 2024

Índice

Índice de Figuras **2**

Índice de Tablas **2**

1 Analisis Exploratorio de los datos **2**

2 Analisis Exploratorio AFE mediante ACP **5**

3 Prueba de Hipotesis datos multivariados **7**

4 Prueba de Hipotesis descomposicion matriz de correlaciones **8**

4.1 Descomposicion de matriz de correlacion mediante maxima verosimilitud . . 9

5 Analisis matrices residuales para ambos casos **10**

5.1 matriz residual con ACP 11

5.2 Metodo del MLE 11

6 Metodo Verimax para rotacion de datos **12**

6.1 Matriz residual metodo verimax 13

7 Graficos Factores-Scores Datos Estandarizados **13**

8 Conclusiones **14**

Referencias **15**

Índice de figuras

1 matriz de correlaciones 4
2 Grafico de codo ACP 6
3 Gráfico de Factores-Scores 1 y 2 (Transformación Varimax) 14

Índice de cuadros

1 Correlaciones Altamente Correlacionadas 5
2 Indices KMO 5
3 Variabilidad explicada por componentes principales 6
4 Valores de Componentes Principales por Variable 7
5 Resultados del Test Henze-Zirkler 8
6 Resultados del Estadístico de Razón de Verosimilitud Ajustado 9
7 Valores mediante MLE de los tres factores 9
8 Proporción de Varianza 10
9 Matriz residual ACP 11
10 Matriz residual MLE 11
11 Factores Varimax 12
12 Análisis de Factores - Estadísticas 12
13 Matriz de residuo metodo verimax 13

1 Analisis Exploratorio de los datos

lo primero es una descripcion de las variables que tiene nuestro dataset las cuales son:

ID de Persona: Un identificador para cada individuo. **Género**: El género de la persona (Masculino/Femenino). **Edad**: La edad de la persona en años. **Ocupación**: La ocupación o profesión de la persona. **Duración del Sueño** (horas): El número de horas que la persona duerme al día. **Calidad del Sueño** (escala: 1-10): Una calificación subjetiva de la calidad del sueño, que va de 1 a 10. **Nivel de Actividad Física** (minutos/día): El número de minutos que la persona realiza actividad física diariamente. **Nivel de Estrés** (escala: 1-10): Una calificación subjetiva del nivel de estrés experimentado por la persona, que va de 1 a 10. **Categoría de IMC**: La categoría de IMC (Índice de Masa Corporal) de la persona (por ejemplo, Bajo peso, Normal, Sobrepeso). **Presión Arterial** (sistólica/diastólica): La medición de la presión arterial de la persona, indicada como presión sistólica sobre presión diastólica. **Frecuencia Cardíaca** (ppm): La frecuencia cardíaca en reposo de la persona en

latidos por minuto. **Pasos Diarios**: El número de pasos que la persona da al día. **Trastorno del Sueño**: La presencia o ausencia de un trastorno del sueño en la persona (Ninguno, Insomnio, Apnea del Sueño).

Lo primero a notar es que variables como ID, Género, Ocupación, Categoría de IMC, Trastorno del Sueño y Presión Arterial son datos no categóricos o, en otras palabras, datos no numéricos. Por lo tanto, no tiene sentido realizar un análisis factorial para estas variables. En consecuencia, solo consideraremos un total de siete variables para este análisis, que son:

Edad , Duración del Sueño(Tsueño), Calidad del Sueño (Csueño), Frecuencia Cardíaca(Fcardiaca), Nivel de Estrés(estres), Nivel de Actividad Física(Afisica), Pasos Diarios(Npasos), las cuales estan denotadas mediante la clave entre los parentesis.

Como los datos presentan escalas diferentes para cada una de sus mediciones, lo primero que se realiza es una normalización de los datos.

Nuestro objetivo es reducir la dimensionalidad del conjunto de datos. Para ello, el análisis factorial de factores comunes nos permite expresar la matriz de varianzas y covarianzas de la siguiente manera:

$$S = \mathbf{L}\mathbf{L}^{\mathbf{T}} + \Psi$$

como nuestros datos tienen una escala diferente entre cada variable nos interesa primero normalizar los datos, por lo tanto el analisis no se basara en la matriz de varianzas y covarianzas si no en la matriz de correlaciones R, para ello proponemos la normalizacion de los datos como sigue:

$$\underline{\mathbf{z}} = \mathbf{V}^{-1/2}(\underline{\mathbf{x}} - \underline{\mu})$$

Luego la matriz de correlaciones R se puede representar por:

$$R = \mathbf{V}^{-1/2}\Sigma\mathbf{V}^{-1/2} = \mathbf{V}^{-1/2}\left[\mathbf{L}\mathbf{L}^{T} + \Psi\right]\mathbf{V}^{-1/2}$$

$$= \left(\mathbf{V}^{-1/2}\mathbf{L}\right)\left(\mathbf{V}^{-1/2}\mathbf{L}\right)^{T} + \mathbf{V}^{-1/2}\Psi\mathbf{V}^{-1/2}$$

Luego podemos escribir la matriz R como:

$$R = \mathbf{L_z}{\mathbf{L_z}}^{T} + \Psi_z$$

por lo tanto lo que se busca es poder describir la matriz de correlaciones mediantes esta descompocision.

como tenemos datos reales tenemos que la matriz de correlacion estimada se puede calcular como:

$$\hat{R} = \hat{\mathbf{L}_{\underline{\mathbf{z}}}}\mathbf{L}_{\underline{\mathbf{z}}}^{\mathbf{T}} \hat{+} \hat{\Psi_{\underline{z}}}$$

donde la matrices

$$\hat{\mathbf{L}} \; ; \; \hat{V^{\frac{-1}{2}}}$$

son obtenidos mediante los estimadores de maxima verosimilitud para las matrices $\mathbf{L}$ Y $V^{\frac{-1}{2}}$, respectivamente.

Dado que nuestro análisis busca utilizar estimadores basados en el análisis de factores comunes, lo primero que debemos hacer es determinar qué variables son adecuadas para este tipo de análisis. Para ello, es crucial que las variables estén altamente correlacionadas entre sí.

En primer lugar, realizaremos un gráfico de correlaciones para visualizar las relaciones entre las variables. Este análisis revela un aspecto importante: la variable "Edad" no muestra una alta correlación con ninguna de las otras seis variables. Por lo tanto, no es conveniente incluir la variable "Edad" en el análisis factorial, ya que su falta de relación con las demás variables podría dificultar su inclusión en alguno de los factores que se identifiquen.

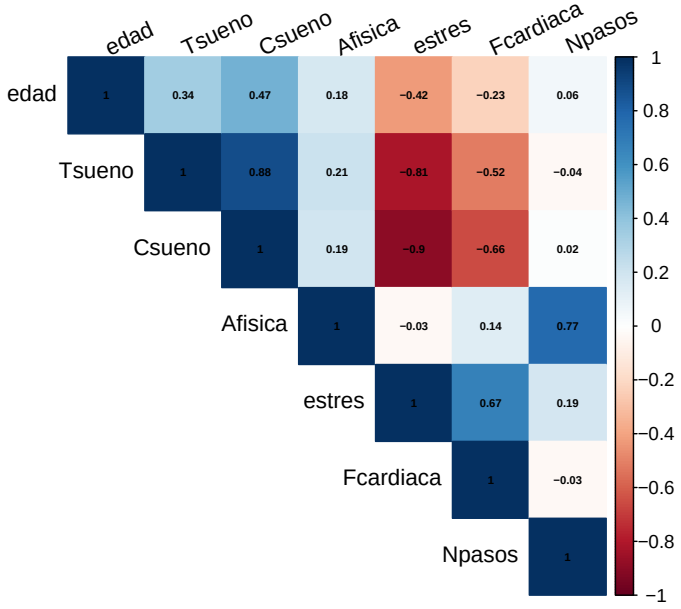

Figura 1: matriz de correlaciones

$|R_{i,j}| > 0.6$ como el valor que indica que las variables estan altamente correlacionadas, podemos observar estas correlaciones las cuales son altas.

Luego de este analisis inicial se decide no considerar a la variable edad como una variable que sea adecuada para realizar el analisis de factores ya que al no estar relacionada con ninguna otra variable no tendria sentido incluirla.

como podemos observar tenemos que los indice **KMO** son todos valores mayores a 0.5 por lo cual a un que no son el valor ideal, podemos realizar un analisis de factores de manera adecuada considerando las variables.

Ahora buscaremos realizar primero un analisis exploratorio de factores esto se realizara mediante la metodologia de **ACP**, o bien llamada metodo de las componentes principales.

Tabla 1: Correlaciones Altamente Correlacionadas

Variables	Correlación
Tsueno y Csueno	0.88
Tsueno y estres	-0.81
Csueno y estres	-0.90
Csueno y Fcardiaca	-0.66
estres y Fcardiaca	0.67
Afisica y Npasos	0.77

Tabla 2: Indices KMO

Variable	KMO
Tsueno	0.80
Csueno	0.73
Afisica	0.55
estres	0.70
Fcardiaca	0.55
Npasos	0.51

2 Analisis Exploratorio AFE mediante ACP

Habiendo seleccionado las variables para realizar el análisis factorial, el primer paso es utilizar la metodología de Análisis de Componentes Principales (ACP). Este método nos permite identificar el número de factores y determinar los posibles candidatos que conformarán cada uno de estos factores. La ACP sirve como una metodología inicial que nos permitirá llevar a cabo un análisis más profundo y obtener una mayor confirmación mediante el análisis factorial confirmatorio (AFC) posterior.

Es importante recordar que el método de ACP nos permite descomponer la matriz de varianza y covarianza en términos de una matriz P y una matriz Δ, donde $\Sigma = P\Delta P^{-1}$, donde la matriz P es la matriz de vectores propios (o autovectores) y la matriz Δ es la matriz diagonal de valores propios (o autovalores).

Del analisis de ACP sabemos que $\Sigma = [\sqrt{\lambda_1}\underline{e_1} \cdots \sqrt{\lambda_p}\underline{e_p}][\sqrt{\lambda_1}\underline{e_1} \cdots \sqrt{\lambda_p}\underline{e_p}]^T$, donde lo λ_i son los valores propios y los $\underline{e_i}$, son los vectores propios asociados al valor propio correspodiente.

el modelo consiste en seleccionar un numero $m < p$ tal que podemos aproximar a

$$\hat{R} \approx \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathbf{T}} + \hat{\Psi}$$

donde

$$\hat{\mathbf{L}} = [\sqrt{\lambda_1}\underline{e_1} \cdots \sqrt{\lambda_m}\underline{e_m}]$$

Es decir seleccionamos una cantidad de componenetes principales que expliquen cierta cantidad de variabilidad.

Tabla 3: Variabilidad explicada por componentes principales

Componentes principales	Variabilidad explicada
PC1	0.54202
PC2	0.30088
PC3	0.10443
PC4	0.02744
PC5	0.01557
PC6	0.00966

Al realizar un análisis de la varianza explicada por cada una de las componentes principales, observamos que las primeras tres componentes explican más del 90% de la varianza total. Esto indica que es adecuado utilizar tres componentes principales para llevar a cabo el análisis. En base a este resultado, propondremos un total de tres factores.

Este hallazgo se ilustra claramente en el gráfico de codo, donde se puede observar que la cantidad óptima de componentes es tres.

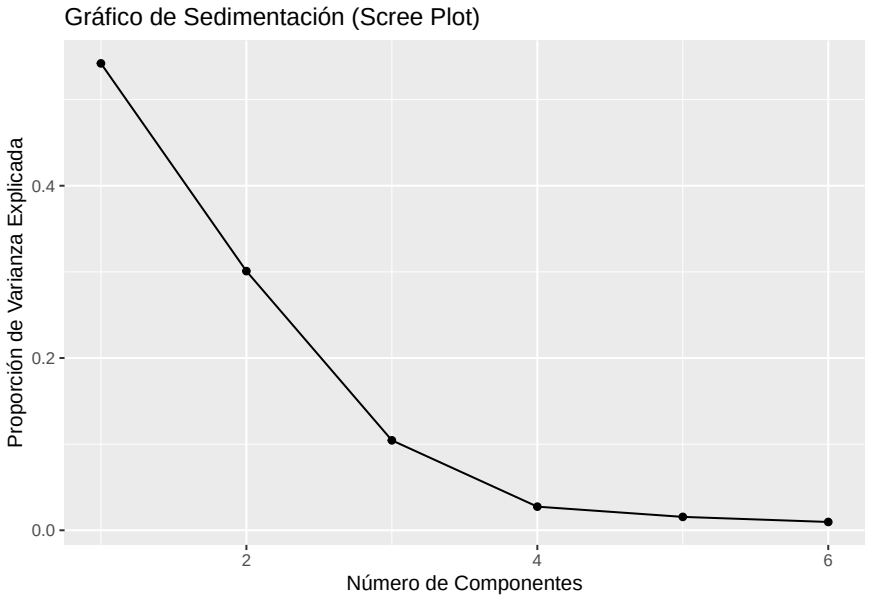

Figura 2: Grafico de codo ACP

Lo primero que se puede observar es que en la primera componente principal, las variables **Tsueno**, **Csueno** y **estres** son las tres variables más importantes. Por lo tanto, podemos considerar que estas tres variables conforman el primer factor. En la segunda componente principal, las variables **Afisica** y **Npasos** son las más relevantes, lo que sugiere que conforman el segundo factor. Finalmente, en la tercera componente principal, la variable **Fcardiaca** es

Tabla 4: Valores de Componentes Principales por Variable

	PC1	PC2	PC3	PC4	PC5	PC6
Tsueno	0.5030440	0.0526490	-0.3635828	-0.6959414	0.2416856	-0.2631274
Csueno	0.5363245	0.0544958	-0.1101193	0.0207875	-0.5601490	0.6189192
Afisica	0.0690446	0.7040041	-0.2694293	0.3146627	0.5026142	0.2745638
estres	-0.5239088	0.0966936	0.1191148	-0.6174825	0.0807695	0.5605117
Fcardiaca	-0.4239081	0.0956961	-0.7797581	0.0231069	-0.4203154	-0.1610045
Npasos	-0.0188983	0.6929135	0.4010960	-0.1854787	-0.4381911	-0.3636233

la única de peso significativo, por lo que podemos deducir que esta variable constituye el tercer factor.

Luego como podemos observar tenemos que las primeras tres componentes explican mas del 90% de la varianza por lo tanto proponemos en una primera instancia a la matrix $\hat{\mathbf{L}}$

$$\hat{L} = \begin{bmatrix} 0.50304400 & 0.05264898 & -0.3635828 \\ 0.5363245 & 0.05449582 & -0.1101193 \\ 0.06904462 & 0.70400410 & -0.2694293 \\ -0.52390876 & 0.09669356 & 0.1191148 \\ -0.42390814 & 0.09569614 & -0.7797581 \\ -0.01889834 & 0.69291350 & 0.4010960 \end{bmatrix}$$

De esta forma podriamos considerar a $\hat{R} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathbf{T}} + \hat{\Psi}$, donde podemos obtener a $\hat{\Psi} = R - \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathbf{T}}$.

Aunque este análisis exploratorio sugiere que es posible descomponer la matriz de correlaciones mediante esta técnica, el análisis por sí solo no es del todo confirmatorio. Por lo tanto, utilizaremos un método con mayor robustez, como el análisis de máxima verosimilitud. Este método nos permitirá confirmar los resultados de manera más precisa y confiable.

3 Prueba de Hipotesis datos multivariados

observamos anteriormente mediante el analisis de **ACP**, tenemos que el numero de factores adecuado es un total de 3 por lo tanto lo primero es proponer una prueba de **Hipotesis**, que nos permita ver si podemos describir a $R = \mathbf{L_z}\mathbf{L_z}^T + \Psi_z$. para ello debemos primero observar si los datos siguen una distribucion multivariada, para ello realizamos el test multivariado de Henze-Zirkler, el cual propone el siguient jeugo de hipotesis

H_0 : los datos son multivariados

H_a : los datos no siguen una distribucion multivariada

donde el umbral se porpone asi, si Valor p bajo $p - value < 0.05$ Sugiere que los datos no siguen una distribución normal multivariada. Valor p alto $p-value >= 0.05$ No hay evidencia suficiente para rechazar la hipótesis de normalidad multivariada.

Tabla 5: Resultados del Test Henze-Zirkler

Test	HZ	p_value	MVN
Henze-Zirkler	21.01629	0	NO

Al realizar esta prueba de hipótesis, observamos que los datos no cumplen con los requisitos para ser considerados como multivariados. Sin embargo, el análisis exploratorio había indicado que las tres primeras componentes principales segregaban adecuadamente las variables en tres factores distintos, explicando de manera significativa gran parte de la variabilidad en los datos. A pesar de que los datos no superan la prueba de hipótesis, se procede con la evaluación para determinar si un modelo con tres factores es el más adecuado.

4 Prueba de Hipotesis descomposicion matriz de correlaciones

A un que los datos no sigan una distribucion multivariada a un asi veremos si es posible descomponer la matriz, por lo tanto proponemos, proponemos la siguiente prueba de hipotesis.

$$H_0 : R_{p\times p} = \mathbf{L}_{\mathbf{z}p\times m}\mathbf{L}_{\mathbf{z}m\times p}^{T} + \Psi_{z_{p\times p}}$$

contra

$$H_1 : R_{p*p} : \text{es cualesquier otra mariz Def. +}$$

el estadistico de razon de la prueba es $\lambda = n\mathbf{Ln}(\frac{|\hat{\mathbf{R}}|}{|\mathbf{R}|})$

donde

$$\hat{R} = \hat{\mathbf{L}_{\underline{z}}}\hat{\mathbf{L}_{\underline{z}}^{\mathbf{T}}} + \hat{\Psi}_{\underline{z}}$$

es la matriz obtenida mediante los estimadores del MLE para $\mathbf{L}$ y Ψ, mientras que $|R|$ seria el determinante de la matriz de correlaciones obtenida.

es decir como tenemos 3 factores el determinante $\hat{R}$ es el producto de los determinantes de las matrices de correlaciones de cada factor, donde el primer factor tiene 3 varibles que la conforman el segundo factor tiene dos variables y el tercer factor solo tiene una.

Al calcular este estadistico de prueba y aproximar mediante un χ_r^2, donde $r = \frac{(p-m)^2-(p-m)}{2}$, en nuestro caso $p = 6$ y $m = 3$, tenemos un valor muy alto para el valor p luego no podemos rechazar la hipotesis nula y por consiguiente

$$R_{p\times p} = \mathbf{L}_{\mathbf{z}p\times m}\mathbf{L}_{\mathbf{z}m\times p}^{T} + \Psi_{z_{p\times p}}$$

Tabla 6: Resultados del Estadístico de Razón de Verosimilitud Ajustado

Estadístico	Valor	p.valor
Estadístico de Razón de Verosimilitud Ajustado	-613.7187	1

por lo tanto la matriz de correlaciones puede ser descrita mediante el modelo de 3 factores, que se habia indicado en el proceso exploratorio de **ACP**.

4.1 Descomposicion de matriz de correlacion mediante maxima verosimilitud

El método de máxima verosimilitud nos permite estimar tanto la matriz $\mathbf{L}$ como la matriz de errores Ψ, mediante $\hat{\mathbf{L}}$ y $\hat{\Psi}$ respectivamente. Como observamos en el apartado anterior, la matriz R se puede estimar mediante la relación

$$\hat{R_Z} = \hat{\mathbf{L_Z}}\hat{\mathbf{L_Z}}^T + \hat{\Psi_Z},$$

ya que las variables están estandarizadas.Además, para evitar múltiples elecciones de $\mathbf{L}$, se agrega la restricción

$$\mathbf{L}^T\Psi^{-1}\mathbf{L} = \Delta,$$

donde Δ es una matriz diagonal.

Mediante el metodo obtenemos a

$$\hat{\mathbf{L}}$$

y

$$\hat{\Psi}$$

como sigue:

Tabla 7: Valores mediante MLE de los tres factores

Variable	MR1	MR2	MR3
Tsueno	0.868	0.070	0.253
Csueno	0.968	0.080	0.129
Afisica	0.114	0.865	0.206
estres	-0.922	0.120	-0.109
Fcardiaca	-0.779	0.114	0.612
Npasos	-0.040	0.960	-0.273

Lo primero que podemos observar es que, para el primer factor, las variables **Tsueño**, **Csueño** y **estres** siguen siendo las más importantes. A pesar de esto, la variable **Fcardiaca** también

presenta un valor no despreciable. En la segunda componente, las variables **Afisica** y **Npasos** son nuevamente las más relevantes. Finalmente, en la tercera componente, **Fcardiaca** emerge como la variable más significativa.

El método de máxima verosimilitud confirma los resultados obtenidos con el método de Análisis de Componentes Principales (ACP), indicando que las variables con mayor peso en cada factor son las mismas en ambos métodos.

Tabla 8: Proporción de Varianza

Metric	MR1	MR2	MR3
SS loadings	3.161	1.709	0.584
Proportion Var	0.527	0.285	0.097
Cumulative Var	0.527	0.812	0.909

adicionalmente tenemos que la varianza explicada por estos 3 factores es de mas del 90%, donde la varianza explicada por la tercera componente es baja pero a un asi tenemos que la frecuencia cardiaca tiene una alta relacion con algunas variables.

En base a esto definimos la matriz de cargas y la matriz de factores unicos para el metodo de maxima verosimilitud como:

$$\hat{L} = \begin{bmatrix} 0.87 & 0.07 & 0.25 \\ 0.97 & 0.08 & 0.13 \\ 0.11 & 0.87 & 0.21 \\ -0.92 & 0.12 & -0.11 \\ -0.78 & 0.11 & 0.61 \\ -0.04 & 0.96 & -0.27 \end{bmatrix}$$

$$\hat{\Psi} = \begin{bmatrix} 0.178259097 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.040452019 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.196226609 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.124229723 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.004810059 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.001546367 \end{bmatrix}$$

5 Analisis matrices residuales para ambos casos

A continuación, procederemos a concluir cuál de los métodos es más adecuado: el Análisis de Componentes Principales (ACP) o el Método de Máxima Verosimilitud (MLE). Para ello, calcularemos las matrices residuales y examinaremos los valores fuera de la diagonal. La aproximación se considera buena siempre y cuando estos valores sean cercanos a cero.

Para ello calcularemos:

$$Residuo = R - \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathbf{T}}$$

5.1 matriz residual con ACP

Tenemos que para el metodo de la **ACP**, la matriz residual es de la forma

Tabla 9: Matriz residual ACP

	Tsueno	Csueno	Afisica	estres	Fcardiaca	Npasos
Tsueno	0.6119824	0.5705116	0.0426029	-0.5092566	-0.5917554	0.0793246
Csueno	0.5705116	0.6972600	0.0878315	-0.6099195	-0.5235939	0.0333346
Afisica	0.0426029	0.0878315	0.4270189	-0.0339410	-0.1112206	0.3942810
estres	-0.5092566	-0.6099195	-0.0339410	0.7019816	0.5315648	0.0621512
Fcardiaca	-0.5917554	-0.5235939	-0.1112206	0.5315648	0.2031214	0.2081290
Npasos	0.0793246	0.0333346	0.3942810	0.0621512	0.2081290	0.3586357

Como podemos observar los valores por fuera de la diagonal usando las primeras tres componentes principales no es nada bueno, por lo cual no se logra captar de manera adecuada las correlaciones entre las variables.

5.2 Metodo del MLE

Tenemos que para el metodo del **MLE**, la matriz residual es de la forma

Tabla 10: Matriz residual MLE

	Tsueno	Csueno	Afisica	estres	Fcardiaca	Npasos
Tsueno	0.1707000	-0.0021870	0.0070603	0.0177770	-0.0003549	-0.0013325
Csueno	-0.0021870	0.0335000	-0.0140035	0.0061480	0.0051353	0.0071914
Afisica	0.0070603	-0.0140035	0.1973000	-0.0109345	-0.0029290	-0.0000769
estres	0.0177770	0.0061480	-0.0109345	0.1138000	0.0004265	0.0106290
Fcardiaca	-0.0003549	0.0051353	-0.0029290	0.0063265	0.0074000	-0.0024086
Npasos	-0.0013325	0.0071914	-0.0000769	0.0051290	-0.0024086	0.0039000

En este caso, los valores fuera de la diagonal de las matrices residuales son cercanos a cero. Esto indica que el método de máxima verosimilitud ha logrado capturar adecuadamente la estructura de los datos, alcanzando una buena aproximación a lo que buscábamos.

Observando las matrices residuales en ambos casos, podemos notar que el método de máxima verosimilitud (MLE) proporciona la mejor aproximación, ya que captura adecuadamente la correlación entre las variables, con valores cercanos a cero fuera de la diagonal.

En contraste, los valores fuera de la diagonal del método de Análisis de Componentes Principales (ACP) no son cercanos a cero, lo que indica que la aproximación mediante este método es menos precisa.

6 Metodo Verimax para rotacion de datos

Teniendo en cuenta que el método de máxima verosimilitud (MLE) proporciona una buena captación de la correlación entre las variables, procederemos a aplicar el método de Rotación Varimax para este último.

En este método, se utiliza una matriz de transformación **T**, para maximizar la variabilidad de los cuadrados de las ponderaciones de cada factor. Esto se logra calculando $\mathbf{L}^* = \mathbf{LT}$, donde la matriz T es una matriz $m * m$ donde m es el numero de factores que se usaran.

Tabla 11: Factores Varimax

Variable	MR1	MR2	MR3
Tsueno	0.90	0.07	-0.12
Csueno	0.94	0.10	-0.27
Afisica	0.19	0.85	0.21
estres	-0.89	0.10	0.29
Fcardiaca	-0.46	0.05	0.88
Npasos	-0.14	0.98	-0.16

Al realizar la transformación mediante el método **Varimax**, obtenemos una nueva matriz $\mathbf{L}^*$, la cual presenta pesos diferentes. Observando esta matriz, podemos identificar qué variables son importantes para cada uno de los factores. En particular, notamos que la frecuencia cardíaca, que anteriormente tenía un peso significativo en el primer factor, ahora se vuelve mucho más relevante para el tercer factor, como se había comprendido inicialmente.

Tabla 12: Análisis de Factores - Estadísticas

Metric	MR1	MR2	MR3
SS loadings	2.73	1.70	1.02
Proportion Var	0.46	0.28	0.17
Cumulative Var	0.46	0.74	0.91
Proportion Explained	0.50	0.31	0.19
Cumulative Proportion	0.50	0.81	1.00

Ahora podemos observar que, mediante el método Varimax, obtenemos valores muy altos en términos de la varianza explicada, alcanzando un 91% de la varianza total. Adicionalmente, vemos que el primer factor explica el 50% de la varianza, lo que indica que es el factor principal en nuestro análisis.

6.1 Matriz residual metodo verimax

Nuevamente calculamos la matriz de residuos obtenida al aplicar el método Varimax. Observamos que la matriz resultante tiene valores muy próximos a cero. Por lo tanto, podemos concluir que esta transformación ha tenido un buen efecto en la captura de las correlaciones entre las variables.

Tabla 13: Matriz de residuo metodo verimax

Tsueno	Csueno	Afisica	estres	Fcardiaca	Npasos
0.1707000	-0.0021870	0.0070603	0.0177770	-0.0003549	-0.0013325
-0.0021870	0.0335000	-0.0140035	0.0061480	0.0051353	0.0071914
0.0070603	-0.0140035	0.1973000	-0.0109345	-0.0029290	-0.0000769
0.0177770	0.0061480	-0.0109345	0.1138000	0.0004265	0.0106290
-0.0003549	0.0051353	-0.0029290	0.0004265	0.0115000	-0.0029086
-0.0013325	0.0071914	-0.0000769	0.0106290	-0.0029086	-0.0056000

Adicionalmente podemos ver que algunos de los valores cambian donde la gran mayoria de valores fuera de la diagonal ahora llegan a ser incluso mas pequeños que cuando solo se realizo el metodo del **MLE**.

7 Graficos Factores-Scores Datos Estandarizados

Ahora observaremos un grafico para los Factores-Scores con los datos estandarizados, para ello graficaremos las dos columnas que contienen los factores 1 y 2.

Es decir estamos calculando los valores de:

$$\underline{\hat{f}^*} = (\mathbf{L}_{\underline{\mathbf{z}}}^{\mathbf{T}} \Psi^{-1} \mathbf{L}_{\underline{\mathbf{z}}})^{-1} \mathbf{L}_{\underline{\mathbf{z}}}^{\mathbf{T}} \Psi^{-1} \underline{z_j}$$

para posteriormente graficas los primeros dos factores.

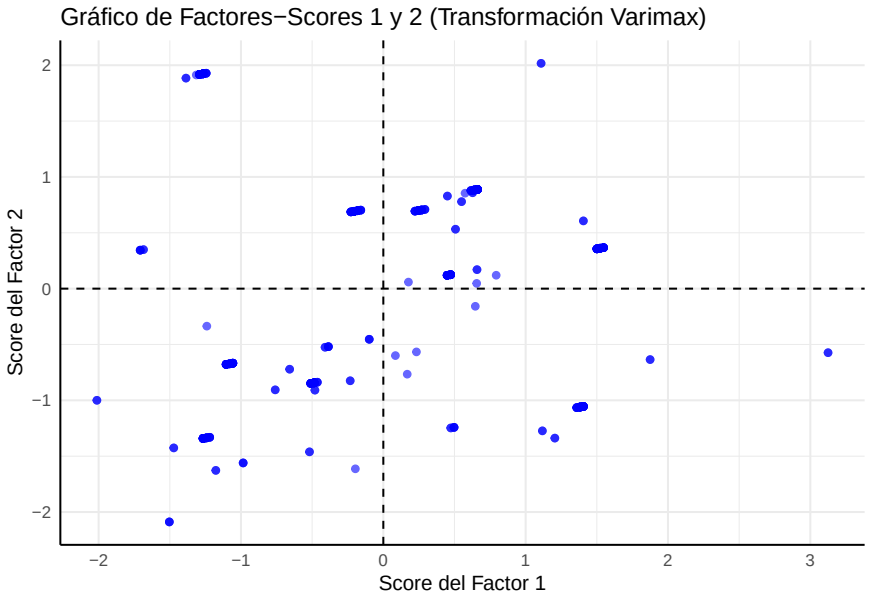

Figura 3: Gráfico de Factores-Scores 1 y 2 (Transformación Varimax)

Los datos parecen estar distribuidos en diferentes cuadrantes, lo que indica que las observaciones tienen diferentes combinaciones de los dos factores.

Como el propósito de la rotación Varimax es maximizar la varianza entre las variables en cada factor, buscando una estructura más interpretable. En este caso, la rotación parece haber logrado cierta diferenciación entre las observaciones a lo largo de ambos factores, con una clara separación entre los puntajes de factor 1 y factor 2.

8 Conclusiones

Se puede concluir que las seis variables estudiadas se pueden describir en términos de tres factores, que podemos definir de la siguiente manera: el primero como calidad del sueño, el segundo como cantidad de actividad física, y el tercero como frecuencia cardíaca. Por lo tanto, obtenemos una reducción de las seis variables originales a tres nuevas variables.

Durante el análisis, se observó que, mediante el método de Análisis de Componentes Principales (ACP), $\mathbf{LL^T} + \Psi$. Sin embargo, esta descomposición a través de las componentes principales generalmente no representa de manera óptima la matriz de varianzas y covarianzas, o en su defecto, la matriz de correlaciones.

El método de máxima verosimilitud requiere algunas condiciones, como la normalidad de los datos. Aunque en este trabajo no se cumplía esta condición, la descomposición se realizó de manera exitosa. Esto sugiere que, aunque en teoría se requiere dicho comportamiento para las variables, el método puede funcionar adecuadamente debido a la alta cantidad de datos o a la forma particular de la matriz de correlaciones.

El análisis factorial puede considerarse un paso posterior al análisis de componentes principales, ya que permite una interpretación más detallada de los datos. A través del análisis factorial, se puede observar de manera más precisa qué variables tienen mayor carga en cada factor, mientras que el método de ACP solo ofrece una combinación lineal de variables con sus respectivos pesos, sin proporcionar esta profundidad interpretativa.

Referencias

Jhonson, D., Richard And Wichern. (2007). *Applications of Multivariate Technique*. Pearson Education.

Luque-Calvo, P.L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. Disponible en http://destio.us.es/calvo.

Porras, J.C. (2016). Comparacion de pruebas de normalidad multivariada. *Anales Cientificos*, pp. 141-146. Universidad Nacional Agraria La Molina.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, **2**, 117-119.

UOM190346a. (2024). Sleep Health and Lifestyle Dataset.