Escuela Nacional de Estadística e Informática

SOFTWARE "R"

Lima – Perú

www.inei.gob.pe

Análisis Factorial

1. Introducción

El hecho de que algunas ideas requieran más de una pregunta para representarlas, en parte, genera una necesidad para combinar preguntas o variables. La clase social, por ejemplo, a menudo se representa mejor por un conjunto de preguntas que incluyen el ingreso, la educación y la ocupación. La necesidad de combinar preguntas también se debe, al hecho de que los conjuntos de preguntas que miden áreas, tan complejas, como el estilo de vida o la imagen contienen una redundancia considerable. Si las preguntas sobre el estilo de vida o, si los conjuntos de preguntas sobre imagen no se combinaran, el análisis sería en su mayor parte pesado y confuso. Una variable que se basa en una combinación de preguntas, puede ser tabulada del mismo modo que las preguntas o variables originales son tabuladas.

El análisis factorial es una técnica que sirve para combinar preguntas, creando de este modo nuevas variables. Esta técnica se encuentra dentro de lo que son las técnicas estructurales o interdependientes porque analizan la interdependencia entre las preguntas, las variables o los objetos. La meta es generar una comprensión de la estructura fundamental de las preguntas, variables u objetos y combinarlos en nuevas variables.

2. Ilustración de la técnica. Un ejemplo sencillo

Suponga que estamos interesados en determinar la forma en la que los estudiantes potenciales seleccionan las universidades. El primer paso podría ser determinar la forma en que las instituciones son percibidas y evaluadas por los estudiantes potenciales. Para generar preguntas relevantes, a los estudiantes se les podría pedir que hablaran informalmente de las escuelas. Más específicamente, a los estudiantes se les podría preguntar porqué una escuela es preferida o por qué dos son consideradas como similares. El resultado podría ser de 100 o más términos, tales como grande, buena facultad, costosa, buen clima, dormitorios, instalaciones, programas de deportes, aspectos sociales, impersonal, etc. Un segundo paso podría ser pedir a un grupo de estudiantes prospectivos que evaluaran la importancia que tiene para ellos cada uno de estos atributos. En este momento el análisis puede retroceder, sencillamente porque hay demasiados atributos o variables. Además, numerosos de estos atributos son redundantes, ya que miden solo la misma idea. Para determinar cuáles son redundantes y lo que están midiendo, el analista puede hacer uso del análisis factorial. Un resultado será un conjunto de variables (o factores) creados mediante la combinación de conjunto de atributos escolares.

En otra fase del estudio, los grupos de estudiantes podrían ser identificados de acuerdo con lo que están buscando en una universidad. Se podría hipotetizar que un grupo esta interesado en la atención individual; otro, en un costo bajo; otro en la proximidad a su casa; y aún otro, en la calidad de educación. Si tales grupos existen y pueden ser identificados, entonces sería posible aislar a varios de ellos, describirlos, y desarrollar un programa de comunicación (confeccionado a su interés) que pudiera ser dirigido hacia ellos.

3. Conceptos básicos del análisis factorial

La función del análisis factorial

El análisis factorial puede ser usado por los investigadores para realizar una variedad de tareas, incluyendo el análisis de agrupamiento y la escala multidimensional; sin embargo, tiene dos funciones principales en el análisis de información.

Una función consiste en identificar las ideas fundamentales de la información. Por consiguiente, las variables "impersonal" y "grande" de nuestro estudio escolar (punto 2) pueden ser, en realidad indicadores de la misma construcción teórica.

Un segundo papel del análisis factorial es, sencillamente, reducir el número de variables a un conjunto más manejable. Al reducir el número de variables, los procedimientos de análisis de factores tratan de retener tanto de la información como sea posible y, de hacer a las variables restantes tan significativas y tan fáciles de manipular como sea posible.

Los conceptos básicos del análisis factorial serán introducidos e ilustrados en el contexto de varios ejemplos. El primero es un estudio (hipotético) realizado por un banco para determinar si se deben desarrollar programas especiales de mercadotecnia para varios segmentos clave. Uno de los objetivos de estudio de investigación se refería a las actitudes hacia la banca. A los entrevistados se les preguntó en una escala de cero a nueve, acuerdo-desacuerdo, sobre las siguientes cuestiones:

a. Los bancos pequeños cobran menos que los bancos grandes. (variable 1)
b. Los bancos grandes tienen más probabilidades de cometer errores que los bancos pequeños. (variable 2)
c. Los cajeros no necesitan ser extremadamente corteses y amistosos, es suficiente con ser atentos. (variable 3)
d. Quiero ser personalmente conocido en mi banco y ser tratado con cortesía especial. (variable 4)
e. Después de ser tratado en una forma impersonal o descuidada por una institución financiera, nunca utilizaría esa organización otra vez. (variable 5)

Para propósitos ilustrativos, suponga que el estudio fue realizado a 15 personas. Un estudio real probablemente tendría un tamaño muestral de alrededor de 400 personas. Los datos del estudio se muestran en la figura 3.1. En esta figura también se muestran las correlaciones entre variables. Un programa de análisis factorial, generalmente, empezará calculando la matriz de correlaciones de variable por variable. De hecho, es del todo posible establecer directamente como insumo la matriz de correlaciones en lugar de los datos sin seleccionar. En cualquier caso, el programa de análisis factorial proporcionará como uno de sus productos la matriz de correlaciones. Es una buena idea examinar estas correlaciones para ver que información e hipótesis pueden obtenerse. ¿Qué correlaciones son las más grandes? ¿Qué implica esto?

DATOS DE INSUMO Y PUNTAJES DE LOS FACTORES

PERSONA	DATOS DE INSUMO					PUNTAJE DEL FACTOR	
	VARIABLE					FACTOR	
	1	2	3	4	5	1	2
1. Juan Echevarría	9	6	9	2	2	-.97	.79
2. María Sosa	4	6	2	6	7	1.02	-.18
3. Susana Gálvez	0	0	5	0	0	-.69	-2.20
4. José Andrade	2	2	0	9	9	1.52	-.22
5. Edgardo Barrantes	6	9	8	3	3	-.80	1.35
6. José Fernández	3	8	5	4	7	.35	.53
7. Toribio Moncada	4	5	6	3	6	.06	.06
8. Herman Palacios	8	6	8	2	2	-.78	.42
9. Miguel Torres	4	4	0	8	8	1.54	-.44
10. Branco Zuñiga	2	8	4	5	7	.48	.58
11. Gabriel López	1	2	6	0	0	-.88	-1.55
12. Alan Rivera	6	9	7	3	5	-.22	.95
13. Ricardo Yépez	6	7	1	7	8	1.48	-.10
14. Alicia Duarte	2	1	7	1	1	-.98	-.89
15. Susana Aliaga	9	7	9	2	1	-1.13	.92

VARIABLE	CORRELACIONES				
	VARIABLE				
	1	2	3	4	5
1	1.00	.61	.47	-.02	-.10
2		1.00	.73	.19	.32
3			1.00	-.83	-.77
4				1.00	.93
5					1.00

VARIABLE	CARGAS FACTORIALES		COMUNALIDAD
	FACTOR		
	1	2	
1	-.24	.72	.57
2	.06	.87	.75
3	-.94	.33	.99
4	.94	.21	.92
5	.93	.26	.93
Porcentaje de variación explicado	55%	36%	

FIGURA 3.1 Análisis Factorial de las actitudes hacia los bancos

¿Qué es un factor?

Para interpretar el resto de la figura 3.1, en primer lugar, es necesario entender el concepto de un factor. Las variables de insumo (entrada) con seguridad contendrán una redundancia. Varias pueden estar midiendo en parte la misma idea fundamental. Esta idea fundamental es lo que se denomina **factor**. Un factor es, por tanto, simplemente una variable o idea que no es directamente observable pero que necesita ser inferida de las variables de insumo (entrada). El factor también podría ser visualizado como un agrupamiento de aquellas variables de insumo que miden o que son indicadores del factor.

Puntajes del factor

Aunque el factor no es observable como las otras cinco variables, es aún una variable. Un producto de la mayoría de los programas de análisis factorial son la creación de valores para cada factor para todos los entrevistados. Estos valores de denominan **puntajes de factores** y se muestran en la figura 3.1 para los dos factores que se encontró que fundamentan a las cinco variables de insumo. De este modo, cada interrogado tiene un puntaje de factor sobre cada factor, en adición a la calificación del entrevistado sobre las cinco variables originales. El factor es una variable derivada, en el sentido de que el puntaje del factor se calcula partiendo de un conocimiento de las variables que están asociadas con él.

Cargas de los factores

¿Cómo se interpreta el factor si no es observable? La interpretación se basa en las **cargas de los factores**, las cuales son las correlaciones entre los factores y las variables originales. En la parte inferior de la figura 3.1 se muestran las cargas de los factores para nuestro estudio bancario. Por ejemplo, la correlación entre la variable 1 y el factor 1 es de -0.24. De este modo, las cargas de los factores proporcionan una indicación de qué variables originales están correlacionadas con cada factor y el grado de correlación. Esta información es usada entonces para identificar y para denominar a los factores no observables en forma subjetiva.

Claramente, las variables 3, 4 y 5 se combinan para definir al primer factor, el cual podría ser denominado como un factor “personal”. El segundo factor está correlacionado en una forma más alta con las variables 1 y 2. Podría denominarse como un factor de “banco pequeño”.

Comunalidades

Cada una e las cinco variables de insumo tiene asociada con ella una varianza que refleja la variación de los 15 entrevistados. El monto de la varianza de la variable 1 que se explica o que es explicado por los factores es la comunalidad de la variable 1 y se muestra en la figura 3.1 la cual es 57%. La comunalidad es el porcentaje de la varianza de una variable que contribuye a la correlación con otras variables o que es "común" a otras variables. En la figura 3.1 las variables 3, 4 y 5 tienen alta comunalidad; por consiguiente, su variación está representada en una forma bastante completa por los dos factores, mientras que la variable 1 tiene una comunalidad más baja. Exactamente más del 50% de la varianza de la variable 1 se debe a los dos factores

Explicación de la varianza

El porcentaje de la varianza explicada es una medida global que indica la cantidad de la varianza total original de las cinco variables que está representada por el factor. Así, el primer factor explica el 55% de la varianza total de las cinco variables, y el segundo factor el 36% más de la varianza. El estadístico de porcentaje de varianza explicada puede ser útil para evaluar y para interpretar un factor, tal como se ilustrará más adelante.

4. ¿Por qué se usa el análisis factorial sobre los datos?

Una razón consiste en obtener conocimientos de las variables a partir de los agrupamientos de variables que surjan. En particular, frecuentemente es posible identificar ideas fundamentales que pudieran tener un significado práctico y teórico. Otra razón consiste en reducir el número de preguntas o escalas a un número manejable.

5. Rotación de factores

El análisis factorial es un poco complicado (o se vuelve más interesante, dependiendo de su perspectiva) por el hecho de que es posible generar varias soluciones de análisis factorial (cargas y puntajes de factores) para cualquier conjunto de datos. Cada solución se denomina como una **rotación de factores** particular y es generada por un **esquema de rotación de factores**. Cada vez que los factores son rotados el patrón de cargas cambia, del mismo modo que lo hace la interpretación de los factores. Geométricamente, la rotación significa sencillamente que las dimensiones son rotadas. Aunque hay muchos métodos de rotación, la rotación varimax es la más común y será descrita posteriormente.

El análisis factorial básico "sin rotación" emplea el análisis de componentes principales (también denominado análisis de factor común) y será el primero en describirse. El objetivo de los componentes principales consiste en generar un primer actor que tendrá la máxima interpretación de la varianza. Posteriormente, con el primer factor y sus cargas fijas asociadas, los componentes principales localizarán un segundo factor que maximizará a la varianza explicada por este segundo factor. El procedimiento continúa hasta que hay tantos factores generados como variables o hasta que el análisis concluye que el número de factores útiles ha sido agotado. La determinación del número de factores a ser incluidos será considerada brevemente.

Cuando se usa el análisis de componentes principales, la interpretación de los factores puede ser difícil. El uso de la rotación varimax puede mejorar bastante la interpretación. Un estudio de las percepciones de 94 consumidores de una marca particular de café ilustrará este punto. Los consumidores, después de probar el café lo calificaron sobre 14 escalas de semántica diferencial.

Las calificaciones fueron analizadas factorialmente por el método de componentes principales, y los resultados se muestran del lado izquierdo de la figura 5.1. El primer factor explicó casi el 75% de la varianza y parece reflejar claramente una dimensión general de gusto-disgusto. Esta interpretación se ve apoyada por el hecho de que la escala 14 fue, de hecho, una calificación de preferencia general y tuvo una carga alta con el primer factor. Los tres factores restantes, en realidad, no contienen cargas de más de 0.40 y son difíciles de interpretar. Tal dificultad de interpretación no es poco común y motiva el uso de la rotación varimax.
La rotación varimax, probablemente sea el esquema de rotación más ampliamente usado, busca un conjunto de cargas de los factores de modo que cada factor tenga algunas cargas cercano a cero, y otras cargas cercanas a -1 ó +1. La lógica es que la interpretación es más sencilla cuando las correlaciones variable-factor están cercanas a -1 ó +1, indicando así una clara asociación entre la variable y el factor; o cercanas a cero, indicando una clara falta de asociación.

ATRIBUTOS DE CAFÉ	FACTORES NO ROTADOS				FACTORES ROTADOS (VARIMAX)				COMUNA-LIDADES
	1	2	3	4	1	2	3	4	
1. Sabor agradable	.86	-.01	-.20	.04	-.63	.38	.36	.34	.78
2. Sabor asombroso	.91	-.01	-.01	-.09	.48	.43	-.53	.38	.83
3. Sabor suave	.86	-.11	-.28	.00	-.70	.26	.38	.36	.83
4. Sabor caro	.91	.15	.00	-.10	.46	-.53	.54	.29	.87
5. Sabor de bienestar	.87	.00	-.31	.10	-.74	.38	.30	.32	.87
6. Sabor vivo	.93	-.03	-.02	-.16	.49	.43	-.59	.35	.90
7. Sabe a verdadero café	.90	-.02	.04	-.21	.42	.38	-.64	.37	.86
8. Un profundo sabor distinto	.77	.36	.11	.16	.31	-.74	.27	.22	.77
9. Sabe a recién hecho	.79	-.28	.24	-.09	.23	.24	.52	-.62	.76
10. Sabor dulce	.87	.25	.22	.17	.28	-.75	.33	.39	.89
11. Sabor puro y limpio	.89	.11	-.05	.10	.51	-.55	.36	.36	.82
12. Sabor tostado	.76	-.29	.04	.27	.43	.28	.16	-.67	.74
13. Sabor fresco	.84	-.27	.19	.12	.33	.32	.36	-.70	.83
14. Preferencia general	.90	.04	.08	-.23	.38	.43	-.65	.34	.86
Porcentaje varianza explicada	**74.6**	**3.4**	**2.7**	**2.6**	**22.9**	**21.3**	**20.4**	**18.7**	
Varianza explicada acumulada	**74.6**	**78.0**	**80.7**	**83.3**	**22.9**	**44.2**	**64.6**	**83.3**	

Figura 5.1 Cargas de los factores antes y después de la rotación

La porción derecha de la figura 5.1 muestra una solución rotada por varimax. Observe que, al igual que la solución de componentes principales sin rotar, un total del 83.3% de la varianza se explica por los cuatro factores. Además, las comunalidades son las mismas. Sin embargo, en la rotación varimax cada uno de los cuatro factores explica un monto sustancial de la varianza, mientras que en la solución sin rotar el primer factor explicaba casi la totalidad de la varianza. En este conjunto de datos la rotación varimax no fue exitosa para impulsar algunas cargas hacia cero.

Por consiguiente, la interpretación es aún un tanto difícil. Sin embargo, es posible proporcionar una interpretación de los primeros cuatro factores considerando las variables con cargas más grandes en los factores. Los factores podrían denominarse como:

Factor 1: Confort-suave

Factor 2: Cordialidad

Factor 3: Genuinidad

Factor 4: frescura

La rotación varimax conduce de este modo a una perspectiva muy diferente, la cual, incidentalmente, no es necesariamente la perspectiva más correcta. Un juicio subjetivo guiado por la teoría es necesario para determinar qué perspectiva, la de los factores sin rotar o la rotada, es más válida.

Algunas veces una solución rotada o no generará factores aparentemente interpretables aun cuando, en realidad, haya poca estructura presente. Para protegerse contra esta eventualidad es prudente dividir los datos a la mitad (aleatoriamente y cuando se pueda) y ejecutar un análisis factorial. Si los mismos factores surgen en cada uno, entonces alguna confianza de que los factores existen en realidad puede ser garantizada.

6. ¿Cuántos factores?

Puesto que el análisis de factores ha sido diseñado para reducir numerosas variables a un número más pequeño de factores o ideas, una pregunta central es: ¿Cuántos factores están involucrados en el modelo? Siempre es posible mantener la generación de factores hasta que haya tantos factores como variables originales, pero tal práctica frustraría, por lo menos, uno de los propósitos primarios de esta técnica.

Teóricamente, la respuesta a la pregunta es clara. Hay un cierto número de ideas que las variables de insumo están midiendo. Estas ideas son identificadas antes del análisis de nuestra teoría y conocimiento de la situación, y posteriormente los datos son analizados mediante factorización hasta que estas ideas surgen como factores. Desafortunadamente, nuestra teoría es rara vez bien definida y necesitamos explorar los datos. En consecuencia se añaden algunas reglas empíricas a la respuesta teórica.

La regla empírica que más se usa en los estudios de análisis factorial es que todos los factores incluidos (antes de la rotación) deben explicar por lo menos tanta varianza como una "variable promedio". En la figura 3.1 la variable promedio explicaría un quinto o el 20% de la varianza. En realidad, el segundo factor explicó el 36%, y el tercer factor, que no se mostró, explicó sólo el 7% de la varianza. En la figura 5.1 esta regla fue violada, puesto que con 14 variables la varianza explicada para una variable incluida debería ser igual a la de una variable promedio, la cual sería un catorceavo o 7%. La lógica es que si un factor es significativo y capaz de representar una ó más variables, entonces debería absorber por lo menos tanta varianza como una variable de insumo original promedio.

Solo el hecho de que exista una gran cantidad de varianza explicada, desde luego, no significa que un factor sea válido, significativo ó útil. Si una escala o pregunta irrelevante fuera repetida muchas veces, cada una con una pequeña modificación, un factor que fundamentara esas preguntas explicaría gran parte de la varianza, pero no sería una idea muy interesante porque las preguntas sobre las cuales se basó no fueron muy importantes.

Una regla empírica relacionada con este tópico consiste en buscar una fuerte caída en la varianza explicada entre dos factores (en la solución no rotada). Por ejemplo, si la varianza explicada por cinco factores (antes de la rotación) fuera del 40, del 30, del 20, del 6, y del 4%, hay una caída en la varianza explicada en el cuarto factor. Esta caída podría señalar la introducción de factores insignificantes y carentes de importancia relativa. Nuevamente, en la figura 5.1 la aplicación de esta regla empírica daría como resultado la consideración de un solo factor. Tal vez la regla más apropiada consiste en detener la factorización cuando los factores dejan de tener sentido. Eventualmente, los factores más pequeños representarán una variación aleatoria y debe esperarse que sean no interpretables. Contrariamente, si un factor que fuera excluido por una de las dos reglas empíricas fuera teóricamente interpretable y de interés práctico, o si diera como resultado factores rotados por varimax con estas cualidades, probablemente debería ser conservado. Resulta claro que la determinación del número de factores, al igual que la interpretación de los factores individuales, contiene un alto grado de subjetividad.

7. Algunos Ejemplos

7.1 Explorando el comportamiento de países en algunos de sus indicadores

En este ejemplo nos disponemos a explorar como los países, a ser analizados, se comportan en los valores de una serie de indicadores, con la finalidad de encontrar países con comportamientos semejantes y opuestos, y si se logra encontrar algún patrón que subyace del comportamiento natural.

Los indicadores que se utilizan en l análisis son los siguientes:

variable	descripción
pais esp_vida tasa_nata tasa_fecun pob_urb tasa_analf creci_veg tasa_mort_inf num_medicos pnb_pc pa_agric pa_ind pa_serv pob_15 pob_**65**	Nombre del país Esperanza de vida al nacer (años) Tasa de natalidad (% hab.) Tasa de fecundidad (hijos por m.) Población urbana (%) Tasa de analfabetismo (%) Crecimiento vegetativo (%) Tasa de mortalidad infantil (%) Número de médicos (% por hab.) PNB por habitante ($) Población activa: agricultura (%) Población activa: industria (%) Población activa: servicios (%) Población < de 15 años (%) Población > de 65 años (%)

Se trata de 170 países, cuyos valores de los indicadores analizados se encuentran en el archivo Matriz_Paises_Variables_Factorial.sav (formato spss) y, con el mismo nombre en formato excel.

Como bien se puede advertir, las variables son de tipo cuantitativas continuas, por lo que cumplen la condición de medida para ser involucradas en un análisis factorial.

Verificando si es adecuado realizar un Análisis Factorial

Como sabemos, un AF resultará adecuado cuando existan altas correlaciones entre las variables, que es cuando podemos suponer que se explican por factores comunes. El análisis de la matriz de correlaciones será pues el primer paso a dar.

Revisemos la matriz de correlaciones:

```
*/Sintaxis SPSS.
CORRELATIONS
  /VARIABLES=esp_vida tasa_nata tasa_fecun pob_urb tasa_analf creci_veg
  tasa_mort_inf num_medicos pnb_pc pa_agric pa_ind pa_serv pob_15 pob_65
  /PRINT=TWOTAIL SIG
  /MISSING=PAIRWISE .
```

	Esperanza de vida al nacer (años)	Tasa de natalidad (% hab.)	Tasa de fecundidad (hijos por m.)	Población urbana (%)	Tasa de analfabetismo (%)	Crecimiento vegetativo (%)	Tasa de mortalidad infantil (%)	Número de médicos (% por hab.)	PNB por habitante ($)	Población activa: agricultura (%)	Población activa: industria (%)	Población activa: servicios (%)	Población < de 15 años (%)	Población > de 65 años (%)
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	1.000	-0.849	-0.846	0.708	-0.833	-0.652	-0.932	0.646	0.562	-0.801	0.712	0.698	-0.803	0.653
2		1.000	0.966	-0.657	0.824	0.896	0.846	-0.736	-0.538	0.770	-0.780	-0.618	0.941	-0.819
3			1.000	-0.628	0.828	0.863	0.844	-0.694	-0.493	0.737	-0.728	-0.600	0.918	-0.774
4				1.000	-0.631	-0.500	-0.648	0.573	0.604	-0.862	0.616	0.815	-0.665	0.548
5					1.000	0.695	0.858	-0.668	-0.438	0.722	-0.661	-0.610	0.766	-0.639
6						1.000	0.675	-0.716	-0.474	0.633	-0.672	-0.507	0.893	-0.843
7							1.000	-0.655	-0.510	0.774	-0.706	-0.660	0.796	-0.649
8								1.000	0.401	-0.637	0.710	0.479	-0.772	0.743
9									1.000	-0.630	0.414	0.654	-0.630	0.629
10										1.000	-0.766	-0.929	0.788	-0.653
11											1.000	0.501	-0.761	0.655
12												1.000	-0.660	0.550
13													1.000	-0.887
14														1.000

Analizando la matriz de correlaciones podemos decir que existen variables con altas correlaciones entre si, como por ejemplo la variable 2 con la variable 3, que presentan una correlación de 0.966. Igualmente la variable 3 con la variable 13, que presentan una correlación de 0.918. Estas correlaciones observadas, nos hacen suponer la existencia de factores comunes y, por lo tanto, ser factible aplicar un análisis factorial.

Para estar seguros, estadísticamente, que la aplicación de un análisis factorial es adecuado, podemos basarnos en algunos test o pruebas, que el SPSS nos permite calcular, y que se encuentran dentro del procedimiento Factor.

- **Test de esfericidad de Bartlett**

Es necesario suponer la normalidad de las variables. Contrasta la H_0 de quela matriz de correlaciones es una matriz identidad (incorrelación lineal entre las variables). Si, como resultado del contraste, no pudiésemos rechazar esta H_0, y el tamaño de la muestra fuese razonablemente grande, deberíamos reconsiderar la realización de un AF, ya que las variables no están correlacionadas.

El estadístico de constraste del test de Bartlett es:

$B = -(n-1-(2p+5)/6)(\ln|R^*|)$

bajo la hipótesis nula resulta una X^2

donde:

p es el número de variables y $|R^*|$ es la determinante de la matriz de correlaciones muestrales.

- **Índice de KMO (Kaiser-Meyer-Olkin)**

KMO se calcula como:

donde:

r_{ji} son los coeficientes de correlación observada entre las variables j e i
a_{ji} son los coeficientes de correlación parcial entre las variables j e i

Estos coeficientes miden la correlación existente entre las variables j e i, una vez eliminada la influencia que las restantes variables ejercen sobre ellas. Estos pueden interpretarse como los efectos correspondientes a los factores comunes, y por tanto, al eliminarlos, a_{ji} representará la correlación entre los factores únicos de las dos variables, que teóricamente tendría que ser nula. Si hubiese correlación entre las variables (en cuyo caso resultaría apropiado un AF), estos coeficientes deberían estar próximos a 0, lo que arrojaría un KMO próximo a 1. Por el contrario, valores del KMO próximos a 0 desaconsejarían el AF.

Una regla comúnmente utilizada es:

- Si KMO < 0.5 no resultaría aceptable realizar un AF.
- Si 0.5 < KMO < 0.7 el grado de correlación es de nivel medio, y habría aceptación media para realizar un AF.
- Si KMO > 0.7 indicaría alta correlación y, por tanto, conveniencia de realizar un AF.

- **Medida de adecuación de la muestra para cada variable (MSA)**

Este índice es similar al KMO, pero para cada variable. La j-ésima variable de MSA viene dada por la siguiente expresión:

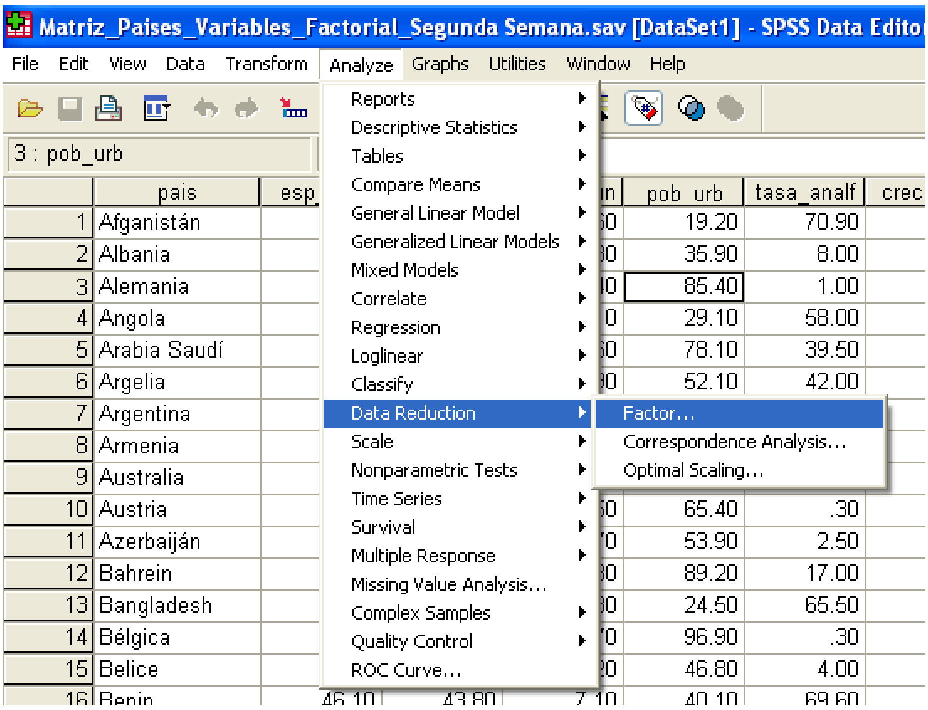
Si el valor del MSA fuera pequeño, no se aconsejaría el AF. Por el contrario, valores próximos a 1 indicarían que la variable X_j es adecuada para incluirla con el resto en un AF. En muchas ocasiones, se eliminan las variables con MSA muy bajo. Estos valores de MSA son los valores que aparecen en la diagonal principal de la matriz de correlación anti-imagen.

- **Correlación anti-imagen (AIC)**

El coeficiente de correlación anti-imagen es el negativo del coeficiente de correlación parcial entre dos variables. Si existiesen factores comunes, esperaríamos pequeños coeficientes de correlación parcial. Por ello, el AF es aplicable cuando en la matriz de correlaciones anti-imagen hay muchos coeficientes pequeños.

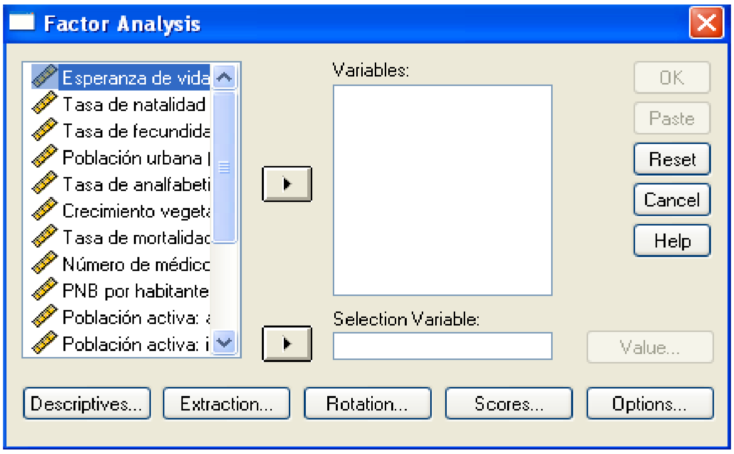
Hemos definido indicadores de los que nos podemos valer para decidir si es apropiado o no la realización de un Análisis Factorial. Procedemos ahora a calcular estos indicadores haciendo uso del SPSS.

Entramos al menú *Analyze*, seleccionamos el procedimiento *Data Reduction* y luego *Factor*, como se muestra en la imagen siguiente:

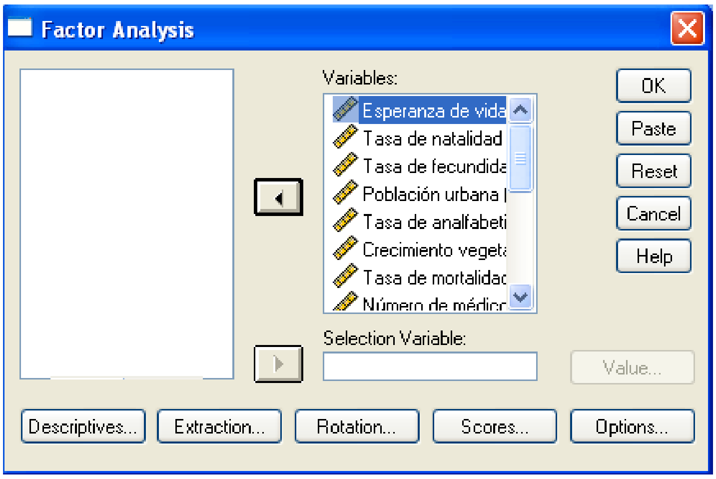


Matriz_Paises_Variables_Factorial_Segunda Semana.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
3 : pob_urb
pais
esp
pob_urb
tasa_analf
crec
1 Afganistán
2 Albania
3 Alemania
4 Angola
5 Arabia Saudí
6 Argelia
7 Argentina
8 Armenia
9 Australia
10 Austria
11 Azerbaiján
12 Bahrein
13 Bangladesh
14 Bélgica
15 Belice
16 Benin
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Quality Control
ROC Curve...
Factor...
Correspondence Analysis...
Optimal Scaling...
19.20 70.90
35.90 8.00
85.40 1.00
29.10 58.00
78.10 39.50
52.10 42.00
65.40 .30
53.90 2.50
89.20 17.00
24.50 65.50
96.90 .30
46.80 4.00

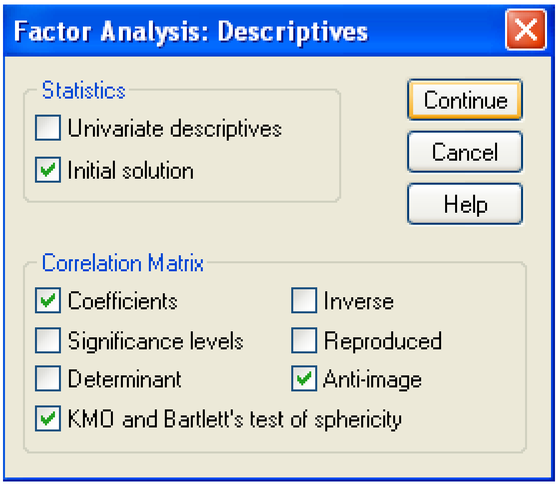
Luego aparecerá el siguiente cuadro de diálogo:



En el recuadro *Variables*, debe introducir todas las variables que ingresaran al análisis factorial, que para nuestro ejemplo son las 14 variables ya definidas anteriormente. Con lo cual el cuadro de diálogo anterior quedaría como sigue:



Luego le damos clik en el botón *Descriptives*, lo cual nos llevará al siguiente cuadro de diálogo, en donde debe activar las opciones que se indican en la imagen siguiente:



Le damos continuar, y en el cuadro de dialogo que aparece, le damos clik al botón *OK*, con lo que aceptamos la realización de una Análisis Factorial, con las demás opciones por defecto, que luego pasaremos a comentar y cambiar de ser necesario.

Entre los resultados que nos da el procedimiento de Análisis Factorial del SPSS está los indicadores de KMO y esfericidad de Bartlett, tal como se muestra en la siguiente imagen:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.902
Bartlett's Test of Sphericity	Approx. Chi-Square	3763.529
	df	91
	Sig.	.000

El índice de KMO es 0.902, muy cercano a 1, lo que nos permite decir que estamos frente a una conveniencia muy fuerte para la realización de un AF. Así mismo, la significancia asociada al valor de estadístico de contraste Bartlett es casi 0, lo que nos permite rechazar la hipótesis nula de que la matriz es una matriz identidad. Luego, ambos indicadores nos hacen decidir sobre que es adecuada la realización de un AF sobre los datos.
Igualmente, podemos analizar la matriz e correlación anti-imagen. Esta, también es proporcionada por el SPSS, y se muestra a continuación.

	Esperanza de vida al nacer (años)	Tasa de natalidad (% hab.)	Tasa de fecundidad (hijos por m.)	Población urbana (%)	Tasa de analfabetismo (%)	Crecimiento vegetativo (%)	Tasa de mortalidad infantil (%)	Número de médicos (% por hab.)	PNB por habitante ($)	Población activa: agricultura (%)	Población activa: industria (%)	Población activa: servicios (%)	Población < de 15 años (%)	Población > de 65 años (%)
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**

1	.922	.189	.106	-.114	.068	-.311	.612	-.046	-.110	-.148	-.153	-.171	.030	.040
2		.930	-.572	.094	-.040	-.421	-.038	-.075	.031	-.130	.006	-.162	-.130	.061
3			.934	.036	-.154	-.029	-.044	-.136	-.195	.032	-.071	.004	-.319	-.128
4				.952	.054	-.132	-.201	-.196	-.180	.298	.162	.101	-.077	.033
5					.973	-.066	-.306	.161	-.075	-.038	-.057	-.016	.128	.013
6						.931	.011	.099	-.107	.166	.115	.194	-.206	.198
7							.931	.050	.016	-.100	-.049	-.055	.013	-.036
8								.954	.224	.025	-.078	.071	.188	-.263
9									.917	-.132	-.064	-.181	.261	-.280
10										.796	.873	.949	-.115	-.131
11											.789	.854	-.020	-.088
12												.753	-.049	-.110
13													.948	.384
14														.942

Observamos de la anterior matriz de correlación anti-imagen, que los valores de la diagonal principal son muy cercanos a uno. Además sabemos, que estos valores corresponden a los indicadores MSA, que son las medidas de adecuación de la muestra para cada variable. Esto también nos dice que estamos frente a una excelente matriz de correlaciones para aplicar un AF.

Igualmente, y como ultima indicación, observamos que los coeficientes de correlación en la matriz anti-imagen son, en su gran mayoría, pequeños, lo que es otro indicativo de una adecuada realización de un AF.

Extracción de los Factores Comunes

Después de haber evaluado, analíticamente y por inspección de la matriz de correlaciones, sobre la adecuación de un AF, procedemos a la extracción de los factores comunes.

Existen distintos métodos de estimación de los coeficientes de la matriz factorial ***L***: los más comunes (para un AF exploratorio) son el método de las Componentes Principales y el método de los Ejes Factoriales.

Para proceder a la estimación de los coeficientes factoriales se parte de la identidad fundamental del AF proporcionada por Thurstone en 1947.

$$R = LL' + w$$

donde:

- R es la matriz de correlaciones entre variables
- W es la matriz de varianzas y covarianzas de los factores únicos

Consideramos una transformación de R: $R - w = LL' = R^*$ (matriz de correlación reducida) cuyos elementos diagonales son las comunalidades y el resto, los coeficientes de correlación lineal entre las variables originales. La idea consiste en determinar ***L*** partiendo de alguna estimación para ***R*,*** y a partir de ella calcular los coeficientes de la matriz ***L***. Podemos entonces optar por el método de los Componentes Principales.

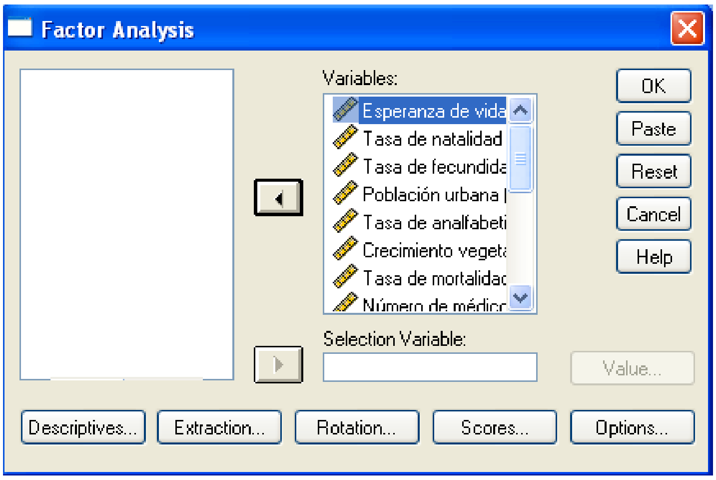
Método de los Componentes Principales (ACP)

El método de componentes principales se basa en suponer que los factores comunes explican el comportamiento de las variables originales en su totalidad de manera que el modelo es:

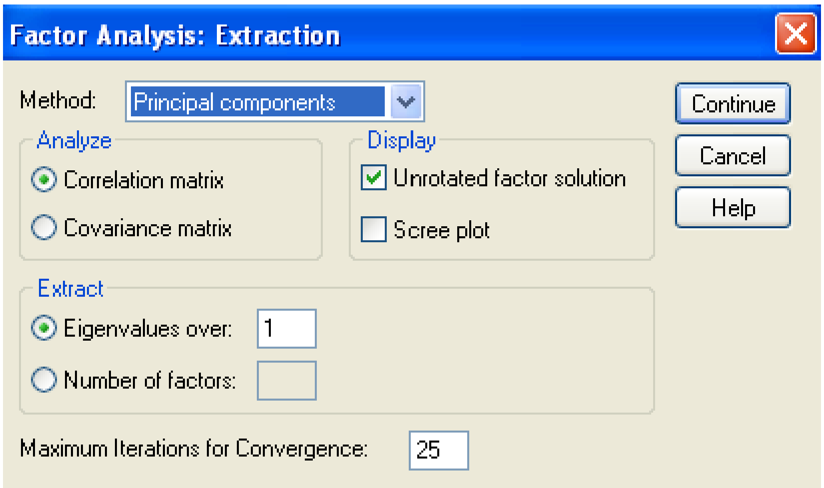
$$X = Lf$$

Las comunalidades iniciales de cada variable son, entonces, igual a 1, porque el 100% de la variabilidad de las **p** variables se explicará por los **p** factores. Evidentemente, carecería de interés sustituir las **p** variables originales por **p** factores que, en ocasiones, son de difícil interpretación. No obstante, si las correlaciones entre las **p** variables fuesen muy altas, sería de esperar que unos pocos factores explicasen gran parte de la variabilidad total. Supongamos que decidimos seleccionar **r** factores. La comunalidad final de cada variable indicará la proporción de variabilidad total que explican los **r** factores finalmente seleccionados.

Veamos entonces que factores nos extrae el procedimiento *Factor* del SPSS.



Recordemos que cuando definimos las variables que entrarían al AF, sólo hicimos cambios en la parte de *Descriptives* del cuadro de diálogo anterior. Por lo que, las opciones por defecto del método de extracción son las que el SPSS da por defecto. Veamos cuales son estas opciones por defecto. Le damos clik al botón *Extraction* y aparecerá el siguiente cuadro de diálogo:



Observamos en el cuadro de dialogo, que el método de extracción por defecto es el de Componentes Principales, que no lo cambiamos, por ser este el que usaremos como primer método. También vemos que, la matriz de análisis (por defecto) es el de los coeficientes de correlación. Esta opción es la requerida cuando las variables están medidas en diferentes unidades.

En lo que se refiere a los factores a ser extraídos, vemos que existen dos criterios, el que está como opción por defecto se refiere a que el procedimiento extraerá aquellos factores cuyos eigen values sean superiores a la unidad. Como bien se ve, también podemos especificar la cantidad de factores que queramos se extraigan, en cuyo caso deberemos seleccionar *Number of factors*. Las opciones de *Display* (mostrar), nos dice que mostrará la solución no rotada, la cual explicaremos más adelante. Además, nos permite especificar que se muestre el *Scree plot* el cual nos da la importancia de cada factor extraído en función de su eigen value. Seleccionemos esta opción.

Los resultados que nos da el SPSS, en lo referente a la extracción de los factores son:

Communalities

	Initial	Extraction
Esperanza de vida al nacer (años)	1.000	.819
Tasa de natalidad (% hab.)	1.000	.947
Tasa de fecundidad (hijos por m.)	1.000	.903
Población urbana (%)	1.000	.827
Tasa de analfabetismo (%)	1.000	.746
Crecimiento vegetativo (%)	1.000	.847
Tasa de mortalidad infantil (%)	1.000	.793
Número de médicos (% por hab.)	1.000	.695
PNB por habitante ($)	1.000	.595
Población activa: agricultura (%)	1.000	.930
Población activa: industria (%)	1.000	.680
Población activa: servicios (%)	1.000	.902
Población < de 15 años (%)	1.000	.928
Población > de 65 años (%)	1.000	.757

Extraction Method: Principal Component Analysis.

El cuadro anterior, muestra las comunalidades iniciales y después de la extracción. Como estamos haciendo uso del método de Componentes Principales, las comunalidades iniciales son 1 para todas las variables. Observamos también que, las comunalidades después de la extracción son lo suficientemente altas como para decir que los factores extraídos explican gran parte de la variabilidad de las variables.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.187	72.763	72.763	10.187	72.763	72.763
2	1.183	8.447	81.210	1.183	8.447	81.210
3	.752	5.375	86.585			
4	.522	3.727	90.311			
5	.345	2.467	92.778			
6	.314	2.243	95.021			
7	.195	1.390	96.411			
8	.165	1.180	97.590			
9	.125	.893	98.483			
10	.081	.581	99.064			
11	.057	.408	99.472			
12	.041	.292	99.764			
13	.025	.181	99.946			
14	.008	.054	100.000			

Extraction Method: Principal Component Analysis.

La imagen anterior, muestra el cuadro de los porcentajes de varianza explicado por cada factor. Como son 14 variables involucradas en el AF, entonces se consiguen extraer 14 factores. Sin embargo, sólo han sido extraídos para el modelo, dos factores, los cuales cumplen la condición de tener un eigenvalue superior a la unidad. Estos se muestran al lado derecho del cuadro. Además, del cuadro se desprende que estos dos factores extraídos para el modelo, explican, en conjunto, un 81,2% de la variabilidad total, siendo el primer factor el que explica un 72,8% y el segundo factor un 8,4%. Tal como se muestra en el cuadro anterior.

De acuerdo a la información dada en el cuadro de varianzas explicadas, un investigador podría decidir, quedarse con esos dos primeros factores, o añadir unos cuantos más, de manera que la explicación de la varianza aumente, hasta un porcentaje de, por ejemplo, 90%. Lo que significaría añadir los dos siguientes factores al modelo (factores 3 y 4).

Analicemos ahora la matriz de componentes. El cual se muestra a continuación.

Component Matrix[a]

	Component	
	1	2
Esperanza de vida al nacer (años)	-.903	.070
Tasa de natalidad (% hab.)	.950	.212
Tasa de fecundidad (hijos por m.)	.924	.221
Población urbana (%)	-.790	.451
Tasa de analfabetismo (%)	.860	.086
Crecimiento vegetativo (%)	.845	.364
Tasa de mortalidad infantil (%)	.891	.008
Número de médicos (% por hab.)	-.795	-.251
PNB por habitante ($)	-.664	.393
Población activa: agricultura (%)	.895	-.360
Población activa: industria (%)	-.815	-.129
Población activa: servicios (%)	-.772	.553
Población < de 15 años (%)	.951	.155
Población > de 65 años (%)	-.843	-.217

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

El cuadro anterior nos da lo que comúnmente se llama, cargas factoriales, que son las correlaciones que existen entre las variables y los factores extraídos para el modelo. Podemos observar, por ejemplo, que la variable *esperanza de vida al nacer* esta correlacionada negativamente con el factor 1 y casi nada con el factor 2. Igualmente, la variable *tasa de natalidad* esta correlacionada positivamente con el factor 1 y muy poco con el factor 2.

Casi todas las variables que entran al análisis tienen una correlación fuerte con el factor 1 y poca o muy poca con el factor 2, salvo las que observan una correlación de alrededor de 0.5.

Si utilizáramos, sólo el factor 1 para realizar un ordenamiento de los países analizados, quizás no lograríamos una adecuada ordenación, ya que el factor 1, sólo explica el 72,8% de la variabilidad. Por lo tanto, se hace necesario involucrar, para el ordenamiento, un segundo factor, lo que incrementará más información.

Sin embargo, si tratamos de interpretar el comportamiento de los países, en las variables de análisis, con estos dos factores extraídos, no se ve muy claro la influencia del factor 2. Esto por el hecho de que las cargas factoriales en el factor 2, son muy bajas. Quizás, para una mejor interpretación, necesitemos la rotación de los factores, de manera que las variables se carguen a uno y solo un factor, para la mejor interpretación.

Rotación de los factores con el objeto de facilitar su interpretación

La interpretación de los resultados del AF se basará en el análisis de las correlaciones entre las variables y los factores que como sabemos viene dado por las cargas factoriales.

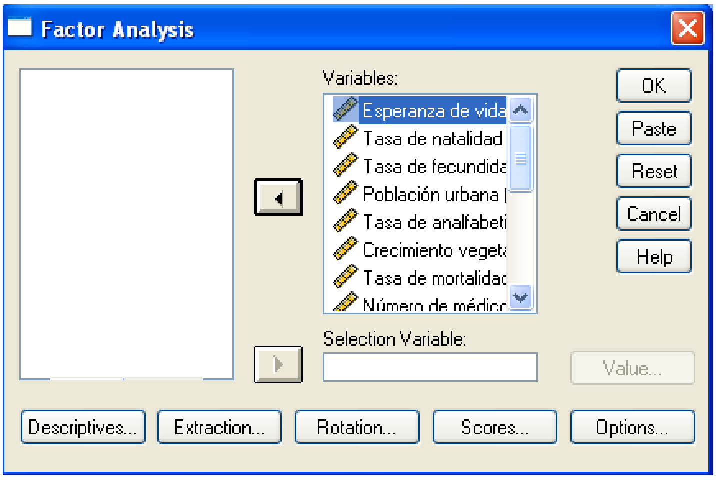
Para que dicha interpretación sea factible, es recomendable que:

- Las cargas factoriales de un factor con las variables estén cerca de 0 ó 1. Así, las variables con cargas próximas a 1 se explican en gran parte en el factor, mientras que las que tengan cargas próximas a 0 no se explican por el factor.
- Una variable debe tener cargas factoriales elevadas con un solo factor. Es deseable que la mayor parte de la variabilidad de una variable sea explicada por un solo factor.
- No debe haber factores con similares cargas factoriales para una misma variable.

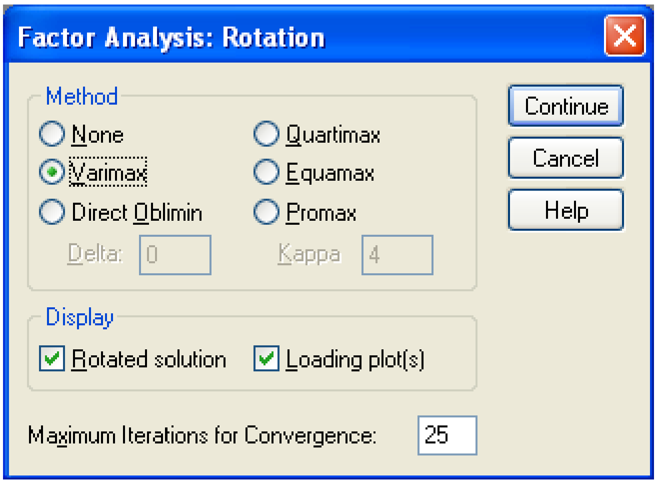
Así, si con la solución inicial no se consigue una fácil interpretación de los factores, éstos pueden ser rotados de manera que cada una de las variables tenga una correlación lo más próxima a 1 con un factor y a 0 con el resto de factores. Como hay menos factores que variables, conseguiremos que cada factor tenga altas correlaciones con un grupo de variables y baja con el resto. Si examinásemos las características de las variables de un grupo asociado a un factor, se podrían encontrar rasgos comunes que permitan identificar al factor y darle una denominación que responda a esos rasgos comunes. Así, conseguiremos desvelar la naturaleza de las interrelaciones existentes entre las variables originales. Los tipos de rotaciones más comunes son la ortogonal y oblicua.

La rotación ortogonal permite rotar los factores estimados inicialmente, de manera que se mantenga la incorrelación entre los mismos. El método más utilizado de rotación ortogonal es la varimax (varianza máxima), ideado por Kaiser. La rotación oblicua no mantiene la ortogonalidad de los factores, lo que nos lleva a aceptar que dos ó más factores expliquen a la vez una misma realidad. Las comunalidades finales de cada variable permanecen inalteradas con la rotación.

En el SPSS solicitamos un método de rotación, dándole clik al botón *Rotation* del siguiente cuadro de diálogo:



Al darle clik aparece el siguiente cuadro de diálogo:



En este cuadro, la opción por defecto es *None* (ninguna), pero para nuestro ejemplo seleccionamos *Varimax* y en *Display*, seleccionamos *Loading plot(s)*. Lo que nos dará la siguiente matriz de cargas factoriales rotadas.

Rotated Component Matrix[a]

	Component	
	1	2
Esperanza de vida al nacer (años)	-.676	.602
Tasa de natalidad (% hab.)	.884	-.407
Tasa de fecundidad (hijos por m.)	.869	-.384
Población urbana (%)	-.355	.838
Tasa de analfabetismo (%)	.736	-.453
Crecimiento vegetativo (%)	.893	-.223
Tasa de mortalidad infantil (%)	.713	-.533
Número de médicos (% por hab.)	-.785	.282
	-.290	.714

PNB por habitante ($)	.494	-.828
Población activa: agricultura (%)	-.726	.391
Población activa: industria (%)	-.279	.908
Población activa: servicios (%)	.850	-.453
Población < de 15 años (%)	-.802	.338

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

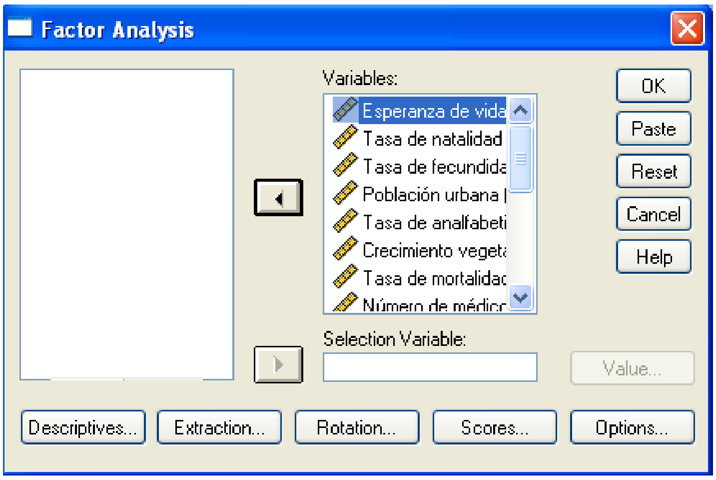
Observemos como después de rotar los factores, 4 variables ahora se cargan en el factor 2. Sin embargo, la variable *esperanza de vida al nacer*, que antes solo cargaba en el factor 1, ahora carga en igual intensidad en los dos factores. Pero aún así, la interpretación se hace mas clara.

Cálculo de las puntuaciones factoriales

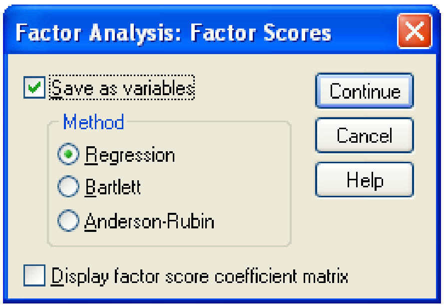
Una vez estimados los factores comunes, es importante calcular las puntuaciones de los sujetos (individuos u objetos) investigados para saber cuánto puntúan en cada factor. Estas puntuaciones son los valores que toman los factores extraídos para el modelo final. Con estas puntuaciones podremos:

- Sustituir los valores de las **p** variables originales para cada sujeto de la muestra por las puntuaciones factoriales obtenidas. En la medida en que el número de factores es menor que el número de variables iniciales, si el porcentaje de explicación de la varianza total fuese elevado, dichas puntuaciones factoriales podrían sustituir a las variables originales en muchos problemas de análisis o predicción. Además, muchas técnicas estadísticas se ven seriamente afectadas por la correlación entre las variables originales. En la medida que las puntuaciones factoriales estén incorrelacionadas podrán utilizarse en ulteriores análisis.
- Colocar a cada sujeto en una determinada posición en el espacio factorial y conocer qué sujetos son los más raros o extremos, dónde se ubican ciertos grupos de la muestra, los más jóvenes frente a los mayores; los de clase alta frente a los de clase media o baja; los creyentes frente a los no creyentes, etc. obteniendo en qué factores sobresalen unos y otros.

En el SPSS para pedir se generen y guarden los puntajes factoriales se entra a la opción de *Scores* en la siguiente ventana de diálogo:



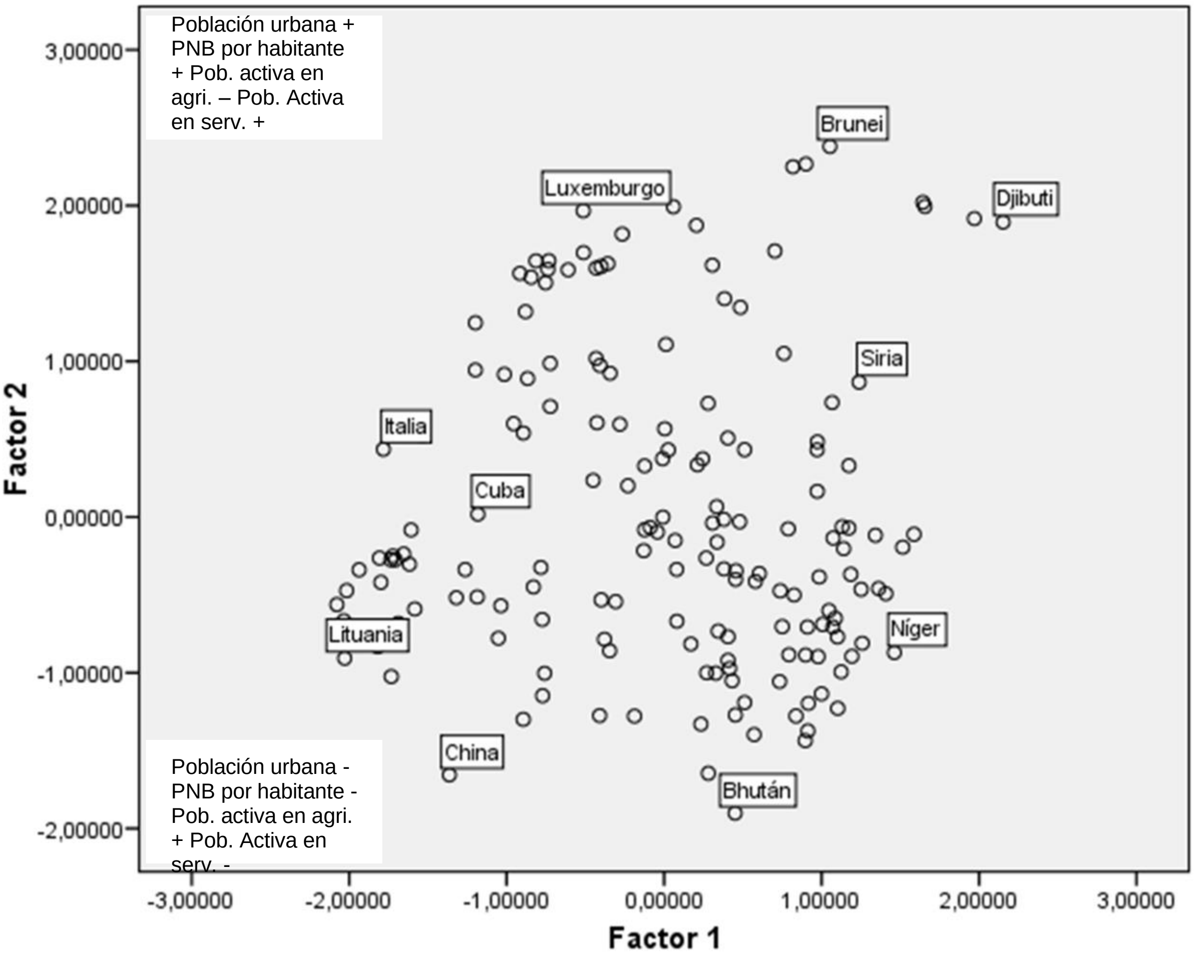
Al darle clik a *Scores*, aparecerá la siguiente ventana de diálogo:



Seleccionamos *Save as variables*, y dejamos la opción del método por defecto, que es el de regresión.

Al ejecutar el procedimiento, se generan dos variables nuevas en el archivo de datos del SPSS, las cuales son los factores con sus respectivas puntuaciones para casa país en el archivo. Los nombres de estas nuevas variables son: FAC1_1 y FAC2_1.

Con estas puntuaciones calculadas, pongamos a los países analizados en un plano factorial, lo que nos permitirá encontrar países con comportamientos similares o muy diferentes en base a las variables de estudio.



Tasa de natalidad –	Tasa de natalidad +
Tasa de fecundidad –	Tasa de fecundidad +
Tasa de analfabetismo –	Tasa de analfabetismo +
Crecimiento vegetativo -	Crecimiento vegetativo +
Tasa de mortalidad infantil -	Tasa de mortalidad infantil +
Número de médicos por hab.+	Número de médicos por hab. –
Población Activa industria +	Población Activa industria –
Población < 15 -	Población < 15 +
Población > 65 +	Población > 65 -