ÍNDICE

1. INTRODUCCIÓN AL ANÁLISIS FACTORIAL EXPLORATORIO (AFE)

1.1. Notas históricas

Aunque podemos considerar a Galton (1883) el pionero en estas técnicas, fueron dos discípulos suyos Spearman y Pearson los que pusieron las bases de las mismas cuando intentaron medir y definir objetivamente la inteligencia.

En 1904 Spearman publicó en el *American Journal of Psychology* un artículo titulado «General Intelligence Objetively Determined and Measured», en el que postulaba que bastaba una sola variable hipotética, un solo factor, para explicar las intercorrelaciones de un conjunto de tests cognoscitivos, a esta variable la denominó factor general de inteligencia, el factor «g». Realmente parece que esta hipótesis era plausible cuando se utilizaban las matrices de intercorrelaciones obtenidas a partir de los tests que había utilizado Spearman, pero cuando se consideraron otras matrices de correlaciones obtenidas a partir de otros tests se vio que la hipótesis de Spearman era bastante simplicista y fueron apareciendo otras hipótesis basadas en factores múltiples. Los precursores de estas nuevas hipótesis fueron L. L. Thurstone en América y C. Burt y G. H. Thomson en Gran Bretaña (Thurstone, 1931; 1935; 1947; Burt, 1949; Thomson, 1951).

1.2. Introducción al modelo de un solo Factor Común

No se trata de hacer una exposición exhaustiva del A.F. ya que, tal y como hemos planteado, nuestro objetivo es que el lector tenga una visión clara de la técnica, de los fundamentos del modelo, y de sus supuestos y limitaciones para que pueda utilizarla e interpretar los resultados obtenidos de una forma adecua-

da. No obstante, se incluye un ejemplo completo, realizado paso a paso, para aquellas personas que estén interesadas.

Cuando el investigador se enfrenta a la tarea de analizar un conjunto de datos observados, una de sus principales tareas es la formulación de un modelo estadístico teórico que permita hacer inferencias acerca de los mismos. Existen una gran variedad de modelos, pero debido a la complejidad de sus desarrollos matemáticos se suele elegir el modelo más simple, el modelo lineal.

El modelo del análisis factorial (A.F.) se desarrolla como una extensión de los modelos de regresión y de correlación múltiple y parcial, a su vez modelos derivados del modelo lineal general.

Vamos a tratar de explicarlo de forma muy sencilla partiendo del modelo de un factor común:

Supongamos que tenemos un conjunto de «n» variables observables correlacionadas entre sí. Supongamos también que las correlaciones entre cada dos de estas variables guardan una cierta proporcionalidad. Pues bien, el razonamiento que hizo Spearman (1904) fue pensar que la regularidad que había en esas correlaciones podía ser debida a la existencia de una variable independiente (V.I.) no observable (variable latente o factor) que fuera la causa de la variabilidad encontrada en las variables observadas, variables dependientes. Si suponemos, además, que la relación entre todas las variables es una relación lineal y que las puntuaciones de todas ellas se han transformado a puntuaciones típicas (se han estandarizado), se podrían derivar las siguientes ecuaciones para relacionar las variables observadas con el factor:

$$\begin{aligned} Z_1 &= a_1 f + e_1 \\ Z_2 &= a_2 f + e_2 \\ Z_n &= a_n f + e_n \end{aligned} \qquad [4.1]$$

donde:

Z_i = la puntuación típica de la variable observada "i"

f = la puntuación típica en el factor

a_i =el coeficiente de regresión en puntuaciones típicas de la variable sobre el factor [2]. Suelen recibir el nombre de «saturaciones o pesos factoriales».

[2] Recuérdese que el coeficiente de regresión en puntuaciones típicas es la correlación entre la variable independiente y la dependiente, en nuestro caso entre el factor (V.I.) y la variable observada (V.D.)

e_i = la parte de la variable dependiente (variable observada) que no viene explicada por el factor, se trata de una puntuación residual considerada como un término de error sobre la que se hacen los siguientes supuestos:

— Se distribuyen de forma aleatoria con (μ =0 y σ =1)

— No hay correlación entre la puntuación error y la puntuación debida al factor, ambas puntuaciones son independientes (ρ_{fe} =0)

— Los errores son independientes entre si ($\rho_{e_i e_j}$ =0)

Una representación del modelo de un factor común, suponiendo que hubiera 5 variables observables, podría ser la que se presenta en la figura 4 que aparece a continuación:

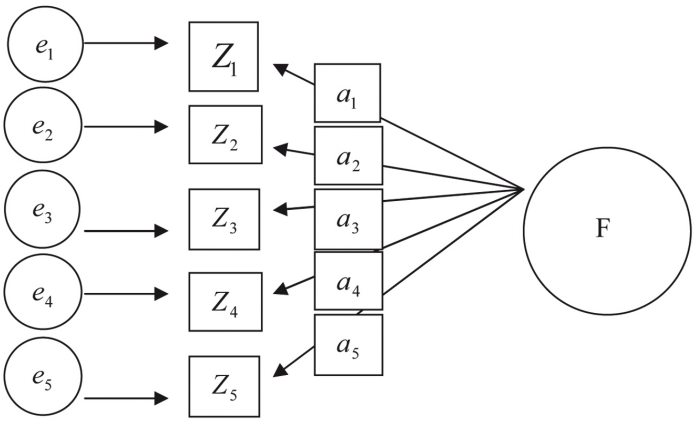

Figura 4: Modelo de un solo factor común

En la medida en que las saturaciones o pesos factoriales (a_i) sean mayores se puede decir que la variable observada (u observable) correspondiente tiene una mayor relación con la variable latente o factor.

Expuesto de esta manera se podría decir que:

— La puntuación en cada variable observada tiene dos componentes, una la determinada por la variable latente (factor) y otra un componente error que es independiente de la primera, puesto que uno de los supuestos de estas puntuaciones error es que no correlacionan con las del factor. En el término error se incluyen todos aquellos efectos que pueden estar incidiendo en los resultados pero que son ajenos a los efectos del factor; en general estos efectos pueden ser

debidos bien a errores de muestreo, de medida o bien a que el modelo especificado no es del todo correcto.

$$Z_i = a_i f + e_i$$
$$VarZ_i = Var(a_i f) + Var(e_i) = a_i^2 Var(f) + Var(e_i) \quad [4.2]$$

Como las dos componentes son independientes, la varianza de las puntuaciones en la variable observada (Z_i) será igual a la varianza de las puntuaciones error (varianza no explicada por el factor) más la explicada por el factor. Como la variable latente está multiplicada por a_i que es constante, su varianza queda multiplicada por esa constante al cuadrado.

Como tanto las puntuaciones *«f»* como las Z_i son puntuaciones típicas, su varianza es igual a la unidad, por lo tanto:

$$Var(Z_i) = 1 = a_i^2 + Var(e_i) \quad [4.3]$$

Dado que a_i es la correlación entre las puntuaciones de la variable observada «i» y el factor «f», que habíamos llamado *saturación o peso factorial,* elevada al cuadrado representa la proporción de la varianza de la variable observada «i» que puede ser explicada por el factor o variable latente y se denomina *comunalidad,* y $Var(e_i)$ es la proporción de la varianza de la variable observada que no depende del factor y se denomina *varianza error* y la designaremos por e_i^2. Por lo tanto la expresión anterior quedaría como sigue:

$$1 = a_i^2 + e_i^2.$$

Hemos visto que la correlación entre el factor y la variable observable venía dado por a_i (el peso factorial o saturación), vamos a ver ahora cómo se puede expresar la correlación entre dos variables observables en función de sus pesos factoriales.

La fórmula de la correlación en puntuaciones típicas es:[3]

$$r_{ij} = \frac{\sum Z_i Z_j}{N}$$

[3] Téngase en cuenta que la correlación entre dos variables es igual a la covarianza entre ellas dividida por el producto de sus desviaciones típicas. Pues bien, cuando se utilizan la escala de puntuaciones típicas, la correlación entre dos variables es igual que su covarianza puesto que las desviaciones típicas son igual a la unidad.

Si sustituimos Z_i y Z_j por sus ecuaciones correspondientes:

$$r_{ij} = \frac{\Sigma(a_i f + e_i)(a_j f + e_j)}{N} = \frac{a_i a_j \Sigma f.f + a_i \Sigma f.e_j + a_j \Sigma f.e_i + \Sigma e_i e_j}{N} = a_i a_j$$

teniendo en cuenta que el término $\frac{\Sigma f.f}{N} = 1$, puesto que es la varianza de la variable latente (factor) en puntuaciones típicas y que el resto de los términos son nulos porque representan las covarianzas entre los errores y la variable latente y la de los errores entre sí que hemos asumido que son independientes. Por lo tanto, la correlación entre dos variables observables se obtiene también a partir de las saturaciones (pesos factoriales) de cada variable en el factor común o variable latente.

De lo dicho anteriormente se deducen las similitudes entre el modelo de regresión lineal y el análisis factorial pero también algunas de sus diferencias, ambas son recogidas claramente en el trabajo de Ferrando y Anguiano-Carrasco (2010, pág. 19):

— En el modelo de regresión lineal la puntuación obtenida en una variable criterio viene explicada en parte por una combinación lineal ponderada de un conjunto de variables predictoras llamadas regresores, existiendo otra parte no explicada que sería el término error.

— En el Análisis factorial, cada una de las variables observadas puede considerarse como un criterio y los regresores o variables explicativas serían los factores que podrían ser comunes para todas las variables o para un subconjunto de las mismas.

— La diferencia más clara, por lo tanto, entre ambos modelos es que mientras que en el modelo de regresión las variables predictoras o regresores son variables observables, en el análisis factorial son variables inobservables o latentes, lo que implica que carezcan de una escala de medida determinada; de ahí la práctica común de utilizar una escala típica de media cero y varianza la unidad.

El modelo planteado por Spearman, el de un solo factor común, equivaldría al modelo de regresión lineal simple.

Ejemplo: Supongamos que tenemos las puntuaciones de un grupo de alumnos de la UNED en 5 variables observables que pueden ser los ítems de un test, cuyas intercorrelaciones son las que aparecen a continuación en la tabla 1:

Tabla 1. Matriz de correlaciones

	A(1)	B(2)	C(3)	D(4)	E(5)	
A(1)	1,00	0,78	0,80	0,75	0,70	Z_1
B(2)	0,78	1,00	0,73	0,61	0,58	Z_2
C(3)	0,80	0,73	1,00	0,48	0,60	Z_3
D(4)	0,75	0,62	0,48	1,00	0,65	Z_4
E(5)	0,70	0,58	0,60	0,65	1,00	Z_5
	Z_1	Z_2	Z_3	Z_4	Z_5	

En la diagonal principal aparecen las correlaciones de una variable consigo misma (la varianza cuando la variable está tipificada), en el resto de las casillas, la mayoría de las columnas son proporcionales entre sí.

Si aplicamos las ecuaciones del modelo de un factor común a la matriz de datos tendremos:

$$r_{12} = 0{,}78 = a_1 a_2$$
$$r_{13} = 0{,}80 = a_1 a_3$$
$$r_{14} = 0{,}75 = a_1 a_4$$
$$r_{15} = 0{,}70 = a_1 a_5$$
$$r_{23} = 0{,}73 = a_2 a_3$$
$$r_{24} = 0{,}61 = a_2 a_4$$
$$r_{25} = 058 = a_2 a_5$$
$$r_{34} = 0{,}48 = a_3 a_4$$
$$r_{35} = 0{,}60 = a_3 a_5$$
$$r_{45} = 0{,}65 = a_4 a_5$$

Podemos ir despejando las saturaciones en las distintas ecuaciones. Por ejemplo, en la primera y segunda ecuación despejamos a_1 y multiplicamos las dos expresiones:

$$a_1 = \frac{0{,}78}{a_2} \qquad a_1 = \frac{0{,}80}{a_3}$$

$$a_1^2 = \frac{0{,}78.0{,}80}{a_2 a_3} = \frac{0{,}78.0{,}80}{0{,}73} = 0{,}85$$

Así iríamos sacando todos los valores[4]:

$$a_1^2 = 0{,}85 \quad a_2^2 = 0{,}71 \quad a_3^2 = 0{,}75 \quad a_4^2 = 0{,}59 \quad a_5^2 = 0{,}81$$
$$a_1 = 0{,}92 \quad a_2 = 0{,}84 \quad a_3 = 0{,}87 \quad a_4 = 0{,}77 \quad a_5 = 0{,}90$$

y tendríamos lo siguiente:

$$e_2^2 = 1 - 0{,}71 = 0{,}29 \rightarrow e_2 = 0{,}54$$
$$e_3^2 = 1 - 0{,}75 = 0{,}25 \rightarrow e_3 = 0{,}50$$
$$e_4^2 = 1 - 0{,}59 = 0{,}41 \rightarrow e_4 = 0{,}64$$
$$e_5^2 = 1 - 0{,}81 = 0{,}19 \rightarrow e_5 = 0{,}44$$

Las ecuaciones del AF para las variables (ítems de un test) utilizadas serían:

$$Z_1 = 0{,}92\, f + 0{,}39$$
$$Z_2 = 0{,}84\, f + 0{,}54$$
$$Z_3 = 0{,}87\, f + 0{,}50$$
$$Z_4 = 0{,}77\, f + 0{,}64$$
$$Z_5 = 0{,}90\, f + 0{,}44$$

Ante estos resultados podremos decir que:

— Entre la variable observada 1 (A) y el factor o variable latente hay un 85% de varianza común o asociada (comunalidad), o lo que es lo mismo, el factor explica el 85% de la varianza de la variable A; queda un 15% de varianza que no es explicada por el factor, es independiente de la variable latente, y es lo que constituye el error.

— En la variable 2 (B) el porcentaje de varianza explicada por el factor es

[4] Dado que cada uno de los coeficientes se puede obtener a partir de varias ecuaciones y no siempre se obtienen los mismos resultados, la forma más adecuada de actuar para estabilizar los valores sería calcular la media de todos los valores posibles. Dejamos esta tarea para el lector puesto que a nosotros lo que nos interesa es que comprendan la lógica del método.

del 71%, la varianza común o asociada entre la variable observada y la latente, la comunalidad; el porcentaje de unicidad sería del 29%.

— En la variable 3 (C) la comunalidad es del 75% y la unicidad del 25%.

— En la variable 4 (D) la comunalidad representa el 59% y la unicidad el 41%.

Como puede observarse se trata de la variable que tiene una menor comunalidad con la variable latente (factor).

— En la variable 5 (E) el factor explica el 81% de su varianza quedando un 19% sin explicar.

Para ver la bondad de ajuste del modelo a los datos, se reproduce la matriz de correlaciones entre las variables a partir de las saturaciones encontradas[5]. Una vez hecho esto se obtiene la matriz de residuales restando de la matriz original los valores encontrados en la matriz que se ha reproducido. En la medida en que estos valores tiendan a cero, podremos decir que el ajuste del modelo es mejor.

La matriz reproducida se obtendría a partir de los siguientes valores (tabla 2):

Tabla 2. Matriz reproducida

$$\begin{bmatrix} a_1^2 & a_1a_2 & a_1a_3 & a_1a_4 & a_1a_5 \\ a_2a_1 & a_2^2 & a_2a_3 & a_2a_4 & a_2a_5 \\ a_3a_1 & a_3a_2 & a_3^2 & a_3a_4 & a_3a_5 \\ a_4a_1 & a_4a_2 & a_4a_3 & a_4^2 & a_4a_5 \\ a_5a_1 & a_5a_2 & a_5a_3 & a_5a_4 & a_5^2 \end{bmatrix} = \begin{bmatrix} 0{,}85 & 0{,}77 & 0{,}80 & 0{,}71 & 0{,}83 \\ 0{,}77 & 0{,}71 & 0{,}73 & 0{,}65 & 0{,}76 \\ 0{,}80 & 0{,}73 & 0{,}75 & 0{,}67 & 0{,}78 \\ 0{,}71 & 0{,}65 & 0{,}67 & 0{,}59 & 0{,}69 \\ 0{,}83 & 0{,}76 & 0{,}78 & 0{,}69 & 0{,}81 \end{bmatrix}$$

La matriz de residuales, prescindiendo de los valores de la diagonal se recoge en la tabla 3:

[5] Téngase en cuenta que se pueden encontrar bastantes discrepancias al no haber calculado la media de los valores posibles para las saturaciones. Se trata de un ejemplo y lo único que se pretende es mostrar el procedimiento.