Técnicas estadísticas multivariantes para la generación de variables latentes

*Carlos Poza Lara**

Resumen

Ante la necesidad de simplificar y medir adecuadamente determinados conceptos se hace necesario conocer el campo de las variables latentes y su explotación mediante el análisis multivariado. En estas líneas se hace referencia a la aplicación del análisis factorial como instrumento clave para generar variables latentes e indicadores.

Palabras clave

Variables latentes
Indicadores y análisis factorial.

1. El concepto de variable latente

En la vida real existen multitud de conceptos que son muy difíciles de definir y de medir *per se* y que, por tanto, necesitan de otros elementos para ser entendidos. Esto sucede todavía con más frecuencia en el campo de las ciencias sociales, debido al comportamiento complejo de las personas. Pensemos, por ejemplo, que tuviéramos que explicar qué es el talento o el bienestar. ¿No parece complicado encontrar una definición idónea?, ¿No dá la sensación de que podríamos construir varias definiciones de cada palabra?, vayamos más allá, ¿No tenemos la intuición de que si nos pudiéramos apoyar en otras ideas relacionadas seríamos capaces de afinar la definición? Pues sí, parece cierto. Cuando esto sucede podríamos decir que ese concepto está implícito en la suma de otras ideas, esto es, ese concepto es una variable latente.

* Doctor en Ciencias Económicas y Empresariales por la Universidad Complutense de Madrid, experto en análisis de datos en Investigación Social y de Mercados por la misma universidad. Actualmente es profesor de la Universidad Antonio de Nebrija.

Este artículo fué entregado el 4 de agosto de 2008 y su publicación aprobada por el Comité Editorial el 16 de agosto de 2008.

Continuemos con uno de los ejemplos: el bienestar es algo que depende de infinidad de elementos y es experimentado de una manera muy distinta según la persona de la que estemos hablando. Así, si tuviéramos que medir el bienestar de un individuo lo mejor sería, en vez de pedirle que valore de cero a diez cómo se siente (que también), preguntarle por su salud (si está enfermo), por su trabajo (si tiene empleo y está a gusto), por sus relaciones sociales (si mantiene un contacto fluido con su familia y amigos), por su nivel de ingresos (si su renta satisface sus necesidades), etc. Esto es, se trata de un concepto indirectamente observable mediante otros que sí son perceptibles o evidentes.

En definitiva, una variable latente es un tipo de variable que se caracteriza por mantener cierto grado de abstracción en su definición y que, por tanto, necesita de otros conceptos más concretos para precisarlo, de modo que se compone de numerosas variables que pretenden medir en detalle de qué se trata. También, se suele identificar como una variable directamente no observable medida o compuesta por variables directamente observables mucho más manejables. Además, es una forma de consolidar numerosa información en una sola variable.

Uno de los desarrollos analíticos más relevantes e innovadores de este tipo de variables fueron aplicados por Joreskog y Wold (1982) en el mundo de la economía hace ya unos años. El libro: "*Systems Under Indirect Observation: Causality, Structure, Prediction. Contribution to Economic Analysis*" es una referencia en este ámbito.

Este concepto también está íntimamente relacionado con la elaboración de indicadores

Abstract

Facing the real need for simplifying and measuring specific concepts, it is necessary to know the field of latent variables and its exploitation through the multivariable analysis. In this article, the application of factorial analysis is described as a key instrument to generate latent variables and indicators.

Key Words

Latent Variables
Indicators
Factorial Analysis.

puesto que, al fin y al cabo, un indicador trata de medir a través de una serie de elementos iniciales un concepto final. Utilizando el ejem-plo anterior es como si cuantificáramos la variable:

Bienestar = 0,25 ingresos + 0,25 salud + 0,25 trabajo + 0,25 relaciones sociales

2. La construcción de variables latentes mediante análisis multivariante

Este punto tiene la finalidad de describir una de las técnicas estadísticas multivariantes más utilizadas para la generación de variables latentes, es el análisis factorial. Veremos su utilidad, su metodología y algunos ejemplos.

Antes de adentrarnos en esta técnica es importante justificar el uso del análisis multivariante como método correcto para crear variables latentes o para producir indicadores. La idoneidad radica en la necesidad de apoyarnos en numerosas variables originales, combinarlas simultáneamente y definir las ponderaciones de forma no arbitraria.

Si bien existen diversas técnicas que podríamos utilizar para explotar las variables latentes, tales como el análisis de correspondencias múltiple, los modelos de clases latentes, la modelización de ecuaciones estructurales e incluso el análisis cluster, con el objetivo de ser precisos y concisos nos vamos a centrar única y exclusivamente en el análisis factorial (por su uso habitual).

A modo aclaratorio:

CUADRO 1
Generación de variables latentes

Técnica	Tipo de variable a utilizar	Paquete estadístico a utilizar
Análisis factorial	Cuantitativa	SPSS y Gandía BarbWin
Análisis de correspondencias múltiple	Cualitativa	SPSS
Análisis cluster	Ambas	SPS y Gandia BarbWin
Modelos de clases latentes	Cualitativa	Latent GOLD
Contrastación de interrelaciones entre variables latentes		
Regresión lineal múltiple	Cuantitativas	SPSS
Regresión logística	Ambas	SPSS
Modelado de ecuaciones estructurales	Cuantitativas	AMOS

Fuente. Elaboración propia

El Análisis Factorial (AF en adelante) es una técnica de reducción de datos. En ocasiones las bases de datos están integradas por variables en las que aparece una amplia redundancia en la información, técnicamente se dice que son variables con un elevado nivel de intercorrelación. Ello plantea el problema de la multicolinealidad que inutiliza la base para muchos modelos predictivos. Surge entonces la necesidad de eliminar la redundancia informativa o eliminar la multicolinealidad.

El AF va a permitirnos sustituir el conjunto original de variables por otro sensiblemente menor en número de variables no observables o hipotéticas, llamadas factores (o en nuestro caso variable latente). Son definidas como variables incorreladas (o con cierta correlación según el tipo de rotación aplicada) que explican los elevados niveles de intercorrelación presentes en la muestra. Estos factores, por tanto, amén de eliminar la multicolinealidad describen las relaciones entre las variables (Manuel, 2005).

A veces los factores son conocidos a priori y el diseño experimental se hace precisamente para obtener una puntuación para cada individuo en los diferentes factores. En este caso el análisis factorial recibe el nombre de "confirmatorio" y es el que habitualmente se utiliza para la generación de indicadores sintéticos, puesto que lo lógico es que sepamos de qué se trata el concepto. En otras situaciones, los factores no son conocidos y se trata de obtenerlos a partir del análisis. Diremos entonces que el análisis factorial es exploratorio.

El método del AF nos invita a seguir unos pasos para la correcta extracción de los resultados. Según Manuel (2005) y Visauta y Martori (2003) podrían ser los siguientes:

a. Evaluar si es apropiado con los datos disponibles ejecutar un análisis factorial

Tomando como primera condición que las variables sean numéricas, deberá haber una fuerte redundancia informativa en el conjunto de las seleccionadas. Dicho de otro modo deberá existir una fuerte correlación dentro de ciertos subconjuntos de variables pero muy pequeñas o nulas entre ellos.

Para desarrollar este apartado se deben obtener y evaluar la matriz de correlaciones de las variables (cuyos valores deberían ser mayores a 0,6 aproximadamente; y cuyos p-valores fueran inferiores a 0,05 con el objetivo de rechazar la hipótesis de correlación nula, lo cual no quiere decir que la correlación sea grande); la medida de adecuación muestral de Kaiser-Meyer-Olkin (KMO) (en este caso cuanto más se acerque a uno tanto más sentido tendrá aplicar el análisis factorial); y por último, aparece el Test de esfericidad de Bartlett (aquí se trata de contrastar la hipótesis de que la matriz de correlaciones es la identidad: si se rechaza, porque el p-valor es inferior a 0,05, estamos admitiendo que la correlación para cada pareja de variables no es nula y por lo tanto el análisis factorial es viable).

b. Obtención de los factores

En esta fase, dado el conjunto de variables intercorreladas el análisis factorial extrae un número de factores coincidente con el original de variables. Sin embargo, como éstas son internamente tipificadas por el método, la varianza global coincide con el número de variables. De esta varianza global cada factor recoge una cierta cantidad, es decir, explica una cierta proporción. Cuanto mayor sea la cantidad explicada más importante es el factor.

El método de Componentes Principales (de los más empleados) extrae secuencialmente los factores, de manera que cada uno de ellos está incorrelado (aunque depende del método de rotación) con todos los anteriores, de forma que la variabilidad recogida por los diferentes factores cada vez es menor. Así pues, se tenderá a despreciar los últimos factores dado que la variabilidad que recogen es pequeña y ahí es donde se consigue la reducción de la dimensionalidad del problema.

Ante esta secuencia, aparecen unos criterios para determinar el número de factores a conservar:

- Criterio de Kaiser: se conservarán aquellos factores con autovalor[1] mayor que uno.
- Gráfico de sedimentación: encontrar puntos de inflexión o saltos de importancia entre factores. Detectar un pico relevante da información sobre el rechazo de los factores siguientes.
- La lógica: basada en la posibilidad de describir el número de factores conservados.

Para identificar la lógica de los factores conservados utilizamos la matriz de componentes y la matriz de componentes rotados, donde se encuentran las variables directamente observables saturadas en los factores directamente no observables.

c. Rotación de los factores

La finalidad de la rotación no es otra sino la de ayudarnos a interpretar, en el supuesto de que no quede claro en la matriz de cargas factoriales no rotada el sentido y significado de los factores.

Existen distintos procedimientos de rotación, fundamentalmente se diferencian dos tipos: los ortogonales y los no ortogonales. Respecto al primer tipo se encuadra el método VARIMAX (trata de minimizar el número de variables que hay con pesos o saturaciones elevadas en cada factor, generando factores incorrelados entre sí), y respecto al segundo, destaca el PROMAX (mantiene cierto grado de correlación entre los factores conservados, muy útil cuando hablamos de sucesos en economía donde casi todo está interrelacionado).

Señalar que la rotación no afecta a la comunalidad[2] y al porcentaje de la varianza explicada por el modelo, aunque sí puede cambiar la de cada factor.

d. Obtención de las puntuaciones factoriales

Puesto que el objetivo fundamental es reducir un gran número de variables a un pequeño número de factores, es a veces aconsejable estimar las puntuaciones factoriales de cada individuo analizado[3], más aún cuando la finalidad es crear un indicador.

1 Entendido como el porcentaje que explica cada factor sobre el total de varianza explicada. Digamos, la importancia de cada factor en el total de la información que representan todas las variables.

2 Importancia de cada variable comparada con las demás utilizadas en el análisis. Así, estudiando las comunalidades de la extracción podemos valorar cuáles de las variables son peor explicadas por el modelo. En definitiva, es la proporción de la varianza de una variable que puede ser explicada por el modelo factorial obtenido.

3 Pensemos que nuestro objetivo de la investigación es obtener el bienestar de una población. El AF nos servirá para identificar a través de las variables originales y los factores (variables latentes) el bienestar de cada persona estudiada. Es decir, cada observación tendrá asignada una puntuación en cada factor, de forma que se podrán realizar comparaciones entre personas.

Como un factor no es otra cosa sino una combinación lineal de las variables originales, el sistema trata de obtener las puntuaciones factoriales de los individuos a través del valor estandarizado de las variables y el coeficiente de la puntuación factorial del factor j respecto de la variable i.

3. Ejemplos de aplicación del análisis factorial para la generación de variables latentes

A continuación exponemos dos ejemplos de generación de variables latentes mediante análisis factorial: uno exploratorio y otro confirmatorio.

EJEMPLO 1[4]: AF exploratorio

OBJETIVO: conocer la opinión de la población acerca de las causas más importantes que provocan una elevada tasa de paro. En nuestro caso, reducir todas las causas en pocos factores o variables latentes.

MÉTODO: cuestionario en el que se pregunta qué factores son los que consideran más importantes para explicar el paro. Se aplica un análisis factorial para reducir información en factores manejables.

VARIABLES: son numéricas de escala Likert. Van de 1 (poca importancia) a 5 (mucha importancia).

1. La crisis económica.
2. La política de empleo del gobierno.
3. La mala gestión de los empresarios.
4. La comodidad de la gente, que solo quiere buenos trabajos.
5. La falta de preparación del trabajador.
6. Las pocas ganas de trabajar de la gente.
7. El no saber buscar trabajo.
8. Que hay mucho pluriempleo.
9. Que el trabajo que hay no se reparte bien socialmente.

[4] Ejemplo proveniente de Pérez, (2004)

FASES:

a. Evaluar si es apropiado con los datos disponibles ejecutar un análisis factorial.
- Matriz de correlaciones: relativamente altas y la mayoría de ellas significativas.
- KMO: 0,712 (nivel aceptable).
- Test de Bartlett: significativo (se rechaza la hipótesis de que la matriz de correlaciones es una matriz de identidad).

b. Obtención de los factores
- Varianza total explicada: 58,32% (repartido entre F1 –27%–, F2 –18%– y F3 –12%–, aproximadamente).
- Criterio de Kaiser: autovalor mayor que uno.
- Gráfico de sedimentación: búsqueda de saltos o puntos de inflexión notables. Salto en el tercer factor.
- Utilización de la lógica en las interrelaciones: los factores creados así como su relación con las variables originales son lógicas.

c. Rotación de los factores
- Interpretabilidad: la rotación utilizada ha sido la varimax (ortogonal), por lo que los factores generados son incorrelados entre sí. Además, la composición de cada uno de ellos se interpreta con facilidad (esto es una máxima de esta técnica).

d. Obtención de las puntuaciones factoriales
- Indicador: cada factor se compone de una serie de variables originales, las cuales tienen un peso sobre la latente. De esta forma, cada factor se construye por combinación lineal a partir de la matriz de puntuaciones factoriales. Véase:

  F1 (Trabajador) = 0,364 * ganas + 0,354 * preparación + 0,343 * comodidad
  + 0,285 * búsqueda + resto variables menos relevante

  F2 (Gobierno y empresarios) = 0,532 * crisis + 0,518 * política de empleo
  + 0,313 * empresarios + resto variables menos relevante

  F3 (redistribución del trabajo) = 0,610 * reparto + 0,556 * pluriempleo
  + resto variables menos relevante

A tal efecto, el SPSS calcula la puntuación factorial de las tres dimensiones para cada individuo encuestado, en función de lo que hayan respondido, de forma que se puede conocer la percepción de las personas sobre los factores más importantes que explican el paro.

RESULTADO: después de aplicar el análisis, con sus diferentes etapas, las variables originales se han aglutinado en los siguientes factores (o variables latentes):

CUADRO 2
Identificación de factores

Factor 1 Trabajador	*Factor 2 Gobierno y empresarios*	*Factor 3 Redistribución del trabajo*
Ganas de trabajar	Crisis	Reparto
Comodidad	Política de empleo	Pluriempleo
Preparación	Empresarios	
Búsqueda		

Fuente: elaboración propia

EJEMPLO 2: AF confirmatorio

OBJETIVO: generar un indicador de bienestar.

MÉTODO: cuestionario realizado por Eurostat sobre aspectos relacionados con el bienestar y la calidad de vida. Se aplica un AF para construir la variable latente "bienestar", en forma de indicador.

VARIABLES: son numéricas contínuas y de escala Likert (desde 1 = muy insatisfecho hasta 6 = plenamente satisfecho).

1. Ingresos totales netos percibidos en el año anterior a la entrevista por el individuo (it_ind).

2. Ingresos mensuales netos actuales percibidos por el hogar (im_h).

3. Ingresos totales del hogar en el año anterior a la entrevista (it_h).

4. ¿Cuál es el grado de satisfacción en relación a su situación actual, respecto a su trabajo o actividad principal? (Sat_trab).

5. ¿Cuál es el grado de satisfacción en relación a su situación actual, respecto a las condiciones de la vivienda? (Sat_viv).

6. ¿Cuál es el grado de satisfacción en relación a su situación actual, respecto a su situación económica? (Sat_eco).

7. ¿Cuál es el grado de satisfacción en relación a su situación actual, respecto a la cantidad de tiempo que puede dedicar al ocio? (Sat_ocio).

FASES:

e. Evaluar si es apropiado con los datos disponibles ejecutar un análisis factorial.

- Matriz de correlaciones: las correlaciones presentadas son significativas.
- KMO: 0,706 (nivel aceptable).
- Test de Bartlett: significativo (se rechaza la hipótesis de que la matriz de correlaciones es una matriz de identidad).

f. Obtención de los factores

- Varianza total explicada: 63% (repartido entre F1 –39%– y F2 –24%– aproximadamente).
- Criterio de Kaiser: autovalor mayor que uno.
- Gráfico de sedimentación: búsqueda de saltos o puntos de inflexión notables. Salto en el segundo factor.
- Utilización de la lógica en las interrelaciones: los factores generados así como su relación con las variables originarias son lógicas.

g. Rotación de los factores

- Interpretabilidad: la rotación utilizada ha sido la promax (no ortogonal), por lo que los factores generados están parcialmente correlados entre sí. Nuevamente, la composición de cada uno de ellos se interpreta con facilidad.

h. Obtención de las puntuaciones factoriales

- Los indicadores podrían resumirse en:

F1 (Bienestar objetivo) = 0,393 * it_h + 0,374 * im_h + 0,333 * it_ind + resto no incluidos en el F1

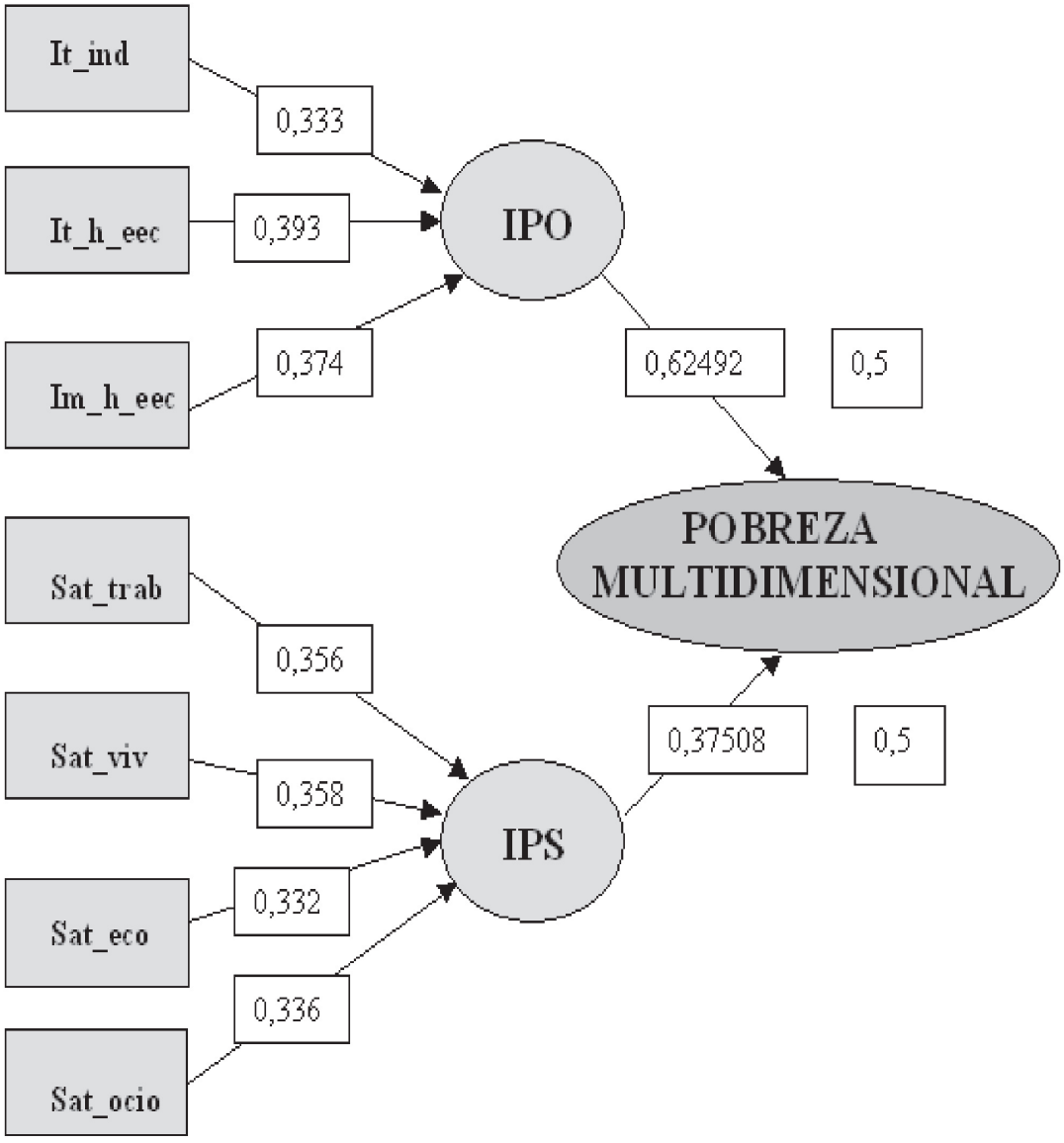
F2 (Bienestar subjetivo) = 0,358 * sat_viv + 0,356 * sat_trab + 0,336 * sat_ocio + 0,332 * sat_eco + resto

Estas variables latentes podrían ser consideradas como indicadores parciales, puesto que nuestro objetivo inicial era construir un indicador de "bienestar". De este modo, nos quedará por aglutinar ambos tipos de bienestar en un único concepto integral. Para ello, realizamos otro AF con la parte objetiva y subjetiva (se puede realizar porque todavía existe correlación entre los factores. Si hubiéramos aplicado al principio una rotación varimax serían incorrelados y, por tanto, no se podría ejecutar el segundo AF).

Para no repetir el proceso, el indicador queda definitivamente formado por:
Bienestar = 0,63 * F1 (bienestar objetivo) + 0,37 * F2 (bienestar subjetivo)

RESULTADO: la variable latente "bienestar" es medida por una parte objetiva y por otra subjetiva. Éstas a su vez cuantificadas por variables directamente observables.

DIAGRAMA 1
Composición de la variable bienestar



Fuente: elaboración propia

4. Conclusiones

- El concepto de variable latente puede ser definido como un tipo de variable indirectamente observable por otras originales. Puede ser utilizada prácticamente en cualquier campo de la investigación.

- Para obtener este tipo de variables se acude normalmente al análisis multivariante. El análisis factorial es una de las técnicas más utilizadas.

- Por último, el estudio de las variables latentes puede derivar en la elaboración de indicadores, muy interesantes en determinadas circunstancias.

5. Bibliografía

Busenitz, L.W.; Gómez, C. and J.W. Spencer, (2000) *"Country Institutional Profiles: Unlocking Entrepreneurial Phenomena"* in The Academy of Management Journal. Vol. 43, Nº 5, pp. 994-1003.

Hair, J.F.; Anderson, R.E.; Tatham, R.L. y W.C. Black, (1999) Análisis Multivariante. 5ª edición, Madrid, Prentice Hall Iberia.

Joreskog, K.G. and H. Wold, (1982) *Systems Under Indirect Observation: Causality, Structure, Prediction. Contributions to Economic Analysis*. Amsterdam: North-Holland, vol. 139, Part II.

Manuel, C.M., (2005) Análisis factorial. Madrid, Escuela de Estadística, Universidad Complutense de Madrid (mimeo).

Pérez, C., (2004) Técnicas de Análisis Multivariante de Datos. Aplicaciones con SPSS. Madrid, Pearson – Prentice Hall.

Poza Lara, C., (2007) Pobreza multidimensional: el caso específico español a través del Panel de Hogares de la Unión Europea. E-prints Complutense. Tesis doctoral, Madrid, Universidad Complutense de Madrid.

Visauta Vinacua, B. y J.C. Martori Cañas, (2003) Análisis estadístico con SPSS para Windows. Volumen II, Estadística multivariante, Madrid, McGraw-Hill.