# Modelos estructurales: aplicación en el análisis de datos categóricos

MSC Alexander Ibarra

## Resumen

Este artículo es un comentario breve sobre las especificidades del planteamiento de modelos estructurales en el caso de variables categóricas. En el se revisa el modo general de los modelos estructurales y luego se anotan algunos de los procedimientos específicos para la formulación de modelos causales con datos cuyo nivel de medida es ordinal o nominal.

**Palabras claves:** modelos estructurales, variables categóricas.

Es común el uso de variables categóricas en las investigaciones referentes a la salud y a la enfermedad, donde identificar de forma continua los procesos asociados resulta de especial dificultad (Davis y Oxford, 1997).

Las variables categóricas son aquellas que adquieren valores a partir de un grupo de etiquetas verbales, las cuales deben ser mutuamente excluyentes y exhaustivas, de forma tal que cada observación pertenezca solo a una de las categorías (Agresti, 1996). Pueden ser nominales u ordinales, y asignársele valores cuantitativos, aun cuando, del que se le asigne valores numéricos no se sigue que adquieran otras propiedades.

Para el análisis de variables categóricas se cuenta con técnicas inferenciales específicas, como el análisis discriminante, el análisis de cluster, la regresión logística y el análisis de tablas de contingencia mediante los *odds ratio*. Estas aproximaciones pueden ser entendidas como modelos causales, en la medida que se utilicen en un nivel más allá de lo descriptivo, hacia la puesta a prueba de modelos explicativos de las relaciones entre las variables.

Los modelos causales permiten especificar las relaciones entre las variables de forma *a priori* a la obtención de los datos, lo que permite disponer de un modelo de relaciones, con especificaciones cuantitativas respecto a la magnitud de las relaciones y sus direcciones. Este modelo es comparado con los datos observados según el grado de ajuste entre lo estimado desde el modelo y lo observado (Hall, 2007; Hair, Anderson, Tatham y Black, 2000).

Los modelos estructurales son uno de los tipos de modelos causales desarrollados para el estudio de variables latentes a partir de indicadores observados (Hair, et al., 2000), aun cuando se le han señalado problemas referentes a su carácter causal, refiriéndose a ellos solamente como útiles en la selección de aspectos según una estructura antecedente conocida (Hall, 2007).

Las diferentes técnicas perteneciente a la familia de modelos estructurales, como el análisis de estructura de covarianza, análisis de variable latente, y análisis factorial confirmatorio, comparten dos características: la estimación de dependencias múltiples y cruzadas y la capacidad de representar conceptos no observados o variables latentes, considerando el error de estimación (Hair, et al., 2000).

Los modelos estructurales tienen dos componentes: un modelo de medida y un modelo estructural (Skrondall y Rabe-Hesketh, 2005). El competente de medida especifica las relaciones entre los indicadores observados y las variables latentes, mientras que el componente estructural especifica las relaciones entre las variables latentes y las regresiones de éstas a los indicadores (Skrondall y Rabe-Hesketh, 2005).

El proceso de modelización de ecuaciones estructurales pasa por las siguientes fases (Hair, et al., 2000):

1. Desarrollo de un modelo estructural basado en la teoría. En esta fase se evalúan diferentes estrategias de modelización, como la confirmatoria, la de modelos rivales y la de desarrollo. La especificación adecuada de las variables es crucial en esta fase, puesto que resalta las variables relevantes según las relaciones que se quieran indagar con el modelo, para evitar así el error de especificación que sesga el papel de otras variables en el modelo.

2. A partir de la especificación de las variables y del establecimiento de relaciones causales se construye un diagrama de relaciones en el que se definen las variables endógenas y exógenas dentro del modelo. Este diagrama tiene como elementos fundamentales los constructos, los indicadores y las flechas que vinculan a los constructor entre sí, y entre éstos y los indicadores.

3. El diagrama de secuencias es traducido a un conjunto de ecuaciones estructurales que permiten especificar el modelo de medida. Esta fase permite formalizar las relaciones causales supuestas desde el modelo estructural. En la especificación del modelo de medida se utiliza el análisis factorial confirmatorio, determinando previamente qué variables definen a cada constructo o factor; donde las variables funcionan como indicadores observados de los constructos.

4. Elección del tipo de matriz de entrada, el cual puede ser de correlación o de varianza/covarianza, a partir de la cual se evalúa la adecuación de los diferentes elementos del modelo y se elige el método de estimación. Es aquí donde se requiere modificar el modelo de medida cuando se trata con indicadores categóricos, puesto que no es posible el cálculo de correlaciones. La normalidad en la distribución de los puntajes afecta la adecuación del chi cuadrado, utilizado en el cálculo del ajuste del modelo. Además, las variables categóricas no tienen una distribución normal, sino que pueden tener una distribución binomial, en el caso de las variables dicotómicas, o polinomial para variables categóricas con más de dos categorías.

5. Evaluación de la identificación del modelo.

6. Evaluar la estimación y la bondad de ajuste del modelo, para posteriormente pasar a la interpretación.

7. Se realizan las modificaciones del modelo, según el ajuste entre el modelo de medida y el estructural, evaluando el sustento teórico de cada modificación sugerida desde el modelo de medida, para finalmente disponer de un modelo final.

De todas estas fases, sólo el componente de medida requiere modificarse cuando los indicadores son categóricos, no así el estructural que puede permanecer esencialmente igual a cuando se dispone de indicadores cuantitativos (Skrondall y Rabe-Hesketh, 2005).

Con frecuencia, los indicadores de las variables latentes son obtenidos a partir de instrumentos con formato de respuestas tipo Likert, a partir de los cuales se extraen datos categóricos a nivel ordinal (DiStefano, 2002). Es un error común la utilización de los datos así obtenidos como variables continuas; aun cuando desde la teoría puede suponerse la existencia de un continuo real tras los indicadores ordinales, no siempre los instrumentos son acompañados de un desarrollo teórico que sustente tal suposición.

Los modelos estructurales suponen el uso de variables continuas con distribución normal; donde la estimación de los parámetros se realiza, generalmente, mediante la

estimación del valor paramétrico para el cual la probabilidad de los datos observados toman su mayor valor, o *maximum likelihood* (DiStefano, 2002).

La estimación de parámetros para las variables latentes, parámetros para los errores y los índices de ajuste *ad hoc* son las tres técnicas en las que se han centrado los investigadores para adaptar los modelos estructurales a las variables categóricas (DiStefano, 2002). A continuación se muestran problemas asociados al uso de estas tres técnicas:

1. La estimación de parámetros para las variables latentes mediante el método de máxima verosimilitud, utilizando variables categóricas con la técnica producto-momento de Pearson, tiende a generar desviaciones negativas en los parámetros conforme aumentan los niveles de asimetría y curtosis.

2. La estimación de parámetros para los errores utilizando la máxima verosimilitud y la técnica producto-momento de Pearson ha arrojado resultados contradictorios, con desviaciones negativas en unos casos y positivas en otros. Esta desviación tiende a decrecer conforme aumenta el tamaño muestral, y tiende a crecer mientras mayor es la anormalidad de la distribución de los ítems (DiStefano, 2002).

3. Con relación al índice de ajuste, DiStefano (2002) señala cómo el cálculo de la bondad de ajuste por método del chi cuadrado tiende a inflar su valor según aumenta la anormalidad de la distribución de los datos categóricos ordinales.

Formalmente, el modelo estructural es definido según la siguiente ecuación (Skrondall y Rabe-Hesketh, 2005):

$$\eta = \alpha + \beta\eta_j + \Gamma x_{1j} + \varepsilon_j$$

Cada variable latente ($\eta$) es definida por un vector que corresponde al intercepto ($\alpha$), una matriz de parámetros estructurales que corresponden a las relaciones entre las variables latentes ($\beta$), un parámetro de regresiones que corresponde a la matriz de regresiones desde las variables latentes a los indicadores ( $\Gamma x_{1j}$), y un vector de error ($\varepsilon_j$).

Disponer de indicadores categóricos requiere modificar el parámetro $\Gamma x_{1j}$, para incluir los valores referidos a cada una de las categorías. En este sentido, Flora y Curran (2004) proponen un método alternativo de estimación para el análisis factorial confirmatorio con datos ordinales, mientras que Muthén (1978) propone un método de estimación para variables dicotómicas.

El método propuesto por Muthén (1978) se basa en la noción de umbral, como límite entre las categorías, y permite la asignación de los casos a cada categoría según del lado del umbral en el que se observen los valores. A partir de esta noción de umbral se construye un modelo general de variable latente, el cual permite su aplicación tanto a variables continuas como a variables categóricas, sean estas ordenadas o no, y es lo que finalmente va a permitir la inclusión de variables categóricas dentro de los modelos estructurales.

Así, la matriz de correlación empleada en la estimación del parámetro es del tipo tetracórica, para variables dicotómicas, para las que se supone un continuo latente normalmente distribuido, o la policórica, para variables con más de dos categorías. Mediante estas técnicas es posible la estimación del parámetro $\Gamma\ x_{1j}$ dentro de la ecuación estructural.

## Referencias bibliográficas

Agresti, A. (1996) *An introduction to categorical data analysis.* Wiley & Sons: New York.

Davis, L. y Offord, K. (1997) "*Logistic regression*". *Journal of Personality Assessment. 68 (3)*, 497-507.

DiStefano, C. (2002) "*The impact of categorization with confirmatory factor analysis*". *Structural Equation Modeling, 9 (3)*, 327-346.

Forran, D. y Curran, P. (2004) "*An empirical evaluation of alternative methods of estimation for confirmatory factor análisis with ordinal data*". *Psychological Methods, 9 (4)*, 466-491.

Hair, J., Anderson, R., Tatham, R. y Black, W. (2000) *Análisis multivariante* (5ta edición). Prentice-Hall: Madrid.

Hall, N. (2007) "*Structural equations and causation*". *Philosophy Studies, 132,* 109-136.

Muthén, B. (1978) "*Contributions to factor analysis of dichotomous variables*". *Psicometrika, 43 (4)*, 551-560.

Skrondall, A. y Rabe-Hesketh, S. (2005) "*Structural equation modeling: categorical variables*". *Encyclopedia of Statistics in Behavioral Science.* Wiley & Sons: New York.