Modelo de Ecuación Estructural con Intervención de Variables Ordinales. Cálculo de sus Correlaciones

Aura López de Murillo*
María del Pilar Marín**
Alonso Arroyo***

* Ph.D. Profesora Pontificia Universidad Javeriana
Santiago de Cali - Colombia.
E-mail:alopez@puj.edu.co

** Estadística - Pontificia Universidad Javeriana -
Santiago de Cali - Colombia
E-mail:mmarin@puj.edu.co

*** Especialista en Sistemas de Información - Profesor Pontificia Universidad Javeriana - Santiago de Cali - Colombia
E-mail:arroyo60@terra.com.co

Grupo Métodos Cuantitativos para la Mejora de Procesos - MC - Facultad de Ingeniería - Pontificia Universidad Javeriana - Santiago de Cali - Colombia.

Fecha de recepción: Octubre 20 de 2002
Fecha de aprobación: Marzo 26 de 2003

1. RESUMEN

En este artículo se presenta una descripción de la técnica estadística multivariada denominada Modelo de Ecuación Estructural (SEM), así como una revisión de literatura para el cálculo de coeficientes de correlación cuando se utilizan variables ordinales.

La técnica SEM permite explicar las relaciones entre un conjunto de variables observadas en términos de un número más pequeño de variables no observadas llamadas factores o variables latentes, las cuales generan la estructura o patrón de relaciones entre las primeras.

Esta técnica ha sido usada en estudios sociales, econométricos y de mercados Price [1], Chau [2] pero sus aplicaciones a procesos industriales apenas comienza. López y Carrión [3] la utilizaron para la cuantificación de las relaciones entre las operaciones de un proceso de fabricación y las características de calidad generadas por dicho proceso. Las variables involucradas en esta aplicación fueron de naturaleza continua y con el supuesto de normalidad.

Los resultados de este estudio hacen parte de la fase inicial de un proyecto que busca explicar las relaciones existentes entre las distintas variables medidas en un proceso productivo, las cuales pueden ser de tipo ordinal, discreto o continuo.

Dado que el SEM requiere como datos de entrada la matriz de varianzas y covarianzas de las variables observadas, se ha realizado una revisión bibliográfica acerca del cálculo de las correlaciones entre variables ordinales, cuyos resultados son presentados en este artículo. Posteriormente serán usados en la etapa de aplicación a un caso real en un proceso de producción.

Palabras claves: Datos ordinales, variables policotómicas, dicotómicas, correlaciones tetracóricas, policóricas, biseriales y poliseriales, SEM, LISREL.

ABSTRACT

In this paper, we present a description of the multivariate technique called Structural Equation Modeling (SEM) and an overview of the calculation of the correlation coefficients using ordinal variables.

The Structural Equation Modeling technique allows us to explain the relations between a set of variables observed in terms of a smaller number of unobserved variables, called factors or latent variables, which generate the structure or pattern of the relations among the former.

This technique has been applied to social studies and market studies Price [1], Chau [2]; but the application on industrial process is just beginning. López and Carrión [3] used it and they quantified the relations between the operations of a manufacturing process, and the characteristics of quality generated by that process, using continuous variables with normal distribution.

The results of this study form part of an initial phase of a project which seeks to explain the relationship among the different variables in a production process, which can be ordinal, discrete of continuous.

Given that the SEM requires the variance and covariance matrix of the observed variables as access data, we have carried out a bibliographical check of the calculation of the correlations between the ordinal variables, and the results of this are presented in this article. They will later be used in the application stage in a real case in a production process.

Key Words: ordinal data, polichotomous variable, dichotomous variable, tetrachoric, polychoric, biserial an polyserial correlation, SEM, LISREL.

2. LA TÉCNICA SEM. ELEMENTOS QUE INTERVIENEN EN EL MODELO

El modelo de ecuación estructural (SEM) es una técnica multivariada que intenta explicar las relaciones entre un conjunto de variables observadas en términos de un número más pequeño de variables no observadas, llamadas factores o variables latentes. Hace parte de una familia de modelos conocidos como Análisis estructural de covarianza [4], Análisis de variables latentes [5], Análisis factorial confirmatorio, entre otros. El modelo asume que las variables no observadas generan la estructura o patrón entre las variables observadas.

Las variables observadas pueden ser endógenas

o exógenas. Las variables observadas exógenas son aquellas que no son "causadas" o predichas por otras variables del modelo mientras que las endógenas sí lo son. Las primeras se denominarán variables *x* y las segundas, variables *y*. Ambos tipos de variables se consideran medidas con error.

Los factores pueden de dos tipos: factores comunes y factores únicos. Los factores comunes pueden ser a su vez exógenos o endógenos. Los primeros son los factores x y los segundos los factores h. Se llaman factores comunes porque cada uno de ellos puede afectar a varias variables observadas. Ver figura 1.

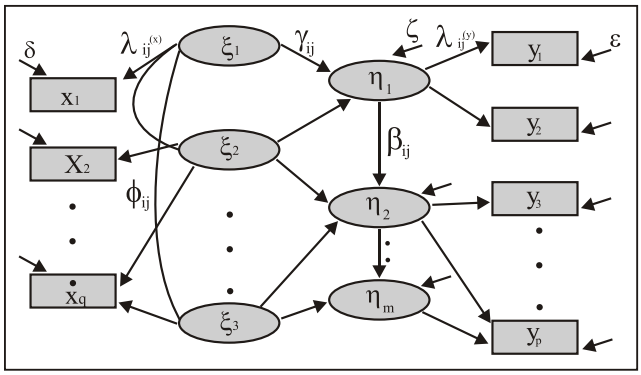


Figura 1. Elementos del Modelo de Ecuación Estructural

Los factores únicos como su nombre lo indica, afectan de manera única a las variables observadas y corresponden a sus errores de medida. Representan aquella parte de la variable observada que no es explicada por los factores comunes. Los factores únicos que afectan a las variables observadas exógenas se denominan factores δ, y los que afectan a las variables observadas endógenas, se denominan factores ε. También se consideran los factores únicos ζ los cuales representan los errores en las ecuaciones del modelo de la ecuación estructural, originados por la no inclusión en el modelo de variables relevantes.

El modelo matemático de estructura de la covarianza consta de tres submodelos, los dos primeros son modelos de mediciones, los cuales relacionan variables observadas con factores, y el tercero es un modelo estructural que relaciona factores endógenos entre sí y factores endógenos con factores exógenos:

- Submodelo de mediciones para las variables observadas exógenas, *x*:

$$X=\Lambda_x \xi + \delta \qquad (1)$$

- Submodelo de mediciones para las variables observadas endógenas, y:

$$y=\Lambda_y \eta + \varepsilon \qquad (2)$$

- Submodelo de ecuación estructural:

$$\eta= B\eta + \Gamma\xi + \zeta \qquad (3)$$

En el modelo intervienen 8 matrices: 4 de parámetros y 4 de varianzas y covarianzas. Las matrices de parámetros corresponden a las matrices de coeficientes de los modelos de medida o matrices de cargas factoriales Λ_x y Λ_y y las matrices de coeficientes estructurales B y Γ que intervienen en la tercera ecuación.

Las matrices de varianzas y covarianzas son: la matriz Φ de varianzas y covarianzas de los factores exógenos, la matriz de varianzas y covarianzas de los errores en las mediciones Θ_δ y Θ_ε respectivamente, y la matriz Ψ de varianzas y covarianzas de los errores en las ecuaciones.

Para la estimación de todos estos parámetros se requiere contar con una matriz de varianzas y covarianzas de las variables observadas, para lo cual se realizó la siguiente revisión sobre el cálculo de correlaciones, utilizando variables ordinales.

3. VARIABLES ORDINALES Y SUS CORRELACIONES

Las variables de escala ordinal, son variables de tipo cualitativo, con un nivel más estructurado que la escala nominal, pero mucho menor que el de las cuantitativas por el tipo de valores obtenidos. Por esta razón, se requiere aplicar un

tratamiento a estas variables para hacer posible la aplicación del coeficiente de correlación de Pearson (ρ) utilizado para la cuantificación de relaciones lineales entre variables cuantitativas.

Tal tratamiento transforma los valores de las variables categóricas en valores numéricos que representan la posición de cada dato después de haber sido ordenados en forma ascendente, constituyendo así una variable aleatoria. El empleo de rangos transforma entonces la escala categórica en numérica abriendo las puertas para el cálculo de sus coeficientes de correlación.

En la tabla 1 se presentan los diferentes coeficientes de correlación, según el tipo de variables relacionadas.

Variables	Correlación
Ambas variables son métricas	Correlación de Pearson
Ambas variables son binarias	Correlación tetracórica
Ambas variables son policotómicas*	Correlación policórica
Una variable métrica y una binaria	Correlación biserial
Una variable métrica y una policotómica	Correlación poliserial

*Ordinales con tres o más categorías

Tabla 1. Coeficientes de correlación según el tipo de variables Hair [6]

4. TRATAMIENTO DE LAS VARIABLES ORDINALES

Para cada variable ordinal x con s categorías se asume que hay una variable latente $\xi \sim N(0,1)$ cuya relación con la variable ordinal univariada es explicada de la siguiente manera, Moustaki [7]:

Si $x = i$ entonces $x \in$ categoría i. Valores bajos de x indican que x clasificó en categoría baja y que los valores asociados de ξ son también valores pequeños. Su conexión está dada por:

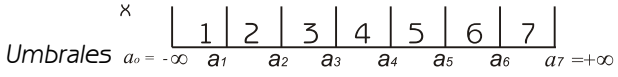
$$x = i \text{ si } a_{i-1} < \xi \leq a_i$$

en donde los parámetros a_i son llamados los umbrales para las variables ordinales, los cuales son estimados a partir de la expresión:

$$a_i = \Phi_1^{-1}\left(\sum_j^i (n_j / N)\right) = \Phi_1^{-1} P_i \qquad (4)$$

en la cual n_j es el número de observaciones en la categoría j, N el número total de observaciones de la variable x, P_j la proporción acumulada hasta la categoría i y Φ_1^{-1} es la inversa de la función de distribución normal univariada. El z correspondiente a $x=i$ es la media de ξ en el intérvalo $a_{i-1} < \xi \leq a_i$

Ejemplo: Sea x una variable ordinal con 7 categorías, es decir $s=7$



El número de observaciones n_j en cada categoría, se muestra en la tabla 2, siendo p_j la respectiva frecuencia relativa.

S	n_j	p_j	P_j
1	4	0.048	0.048
2	15	0.179	0.2270
3	6	0.071	0.2980
4	6	0.071	0.3690
5	11	0.131	0.5000
6	17	0.202	0.7020
7	25	0.298	1.0000

Tabla 2. Observaciones por categoría. Ejemplo

Si $x = 3 \Rightarrow a_2 < \xi \leq a_3$

$$\Rightarrow a_3 = \Phi_1^{-1}\left(\sum_{j=1}^{3}(n_j/N)\right) = \Phi_1^{-1}\left(\frac{4+15+6}{84}\right) = \Phi_1^{-1}(0.3) = -0.52 = z_3$$

De la misma forma se llega a calcular a_2 como $\Phi_1^{-1}(0.23) = -0.74$ y entonces $-0.74 < \xi < -0.52$

Y el z correspondiente a $x=3$ es la medida de ξ en este intérvalo.

4.1 Correlación policórica:

Sean x e y dos variables ordinales observadas, con s y r categorías respectivamente. Para tales variables ordinales se supone la existencia de sendas variables continuas subyacentes ξ y η

con distribución normal bivariada, Olsson [8]. Se define,

$$\begin{array}{ll} x = 1 \quad \text{si} \quad \xi < a_1 & y = 1 \quad \text{si} \quad \eta < b_1 \\ x = 2 \quad \text{si} \quad a_1 \le \xi < a_2 & y = 2 \quad \text{si} \quad b_1 \le \eta < b_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x = s \quad \text{si} \quad a_{s-1} \le \xi & y = r \quad \text{si} \quad b_{r-1} \le \eta \end{array}$$

En general, los umbrales de las variables *x* e *y* son llamados a_i y b_j respectivamente, con $a_o = b_o = -\infty$ y $a_s = b_r = +\infty$

4.2 El problema de estimación a resolver:
El problema consiste en llegar a estimar la correlación normal bivariada ρ entre ξ y η, cuando se dispone de datos categóricos, o una tabla de frecuencias. Este coeficiente de correlación fue denominado por Pearson, correlación policórica, el cual es una extensión de las correlaciones tetracóricas, Olsson [8]. Los parámetros a estimar son el coeficiente de correlación ρ y los umbrales *h* y *k* aplicando métodos numéricos a la siguiente función de densidad:

$$\Phi(h, k, \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \int_{-\infty}^{h}\int_{-\infty}^{k} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}} dxdy \quad (5)$$

Pearson expandió por series el lado derecho de esta ecuación, llegando a obtener una ecuación polinómica de ρ, cuyo grado depende del número de términos considerados en la expansión.

Utilizando el método de máxima verosimilitud, Olsson presenta dos formas para la estimación. Una primera forma consiste en estimar simultáneamente ρ y los umbrales. Por la segunda forma, llamada estimación en dos etapas, se estiman primero los umbrales a partir de la inversa de la función normal, evaluada en las proporciones marginales acumuladas de la tabla, y una vez obtenidos los umbrales, se procede a la estimación de ρ por máxima verosimilitud. Aunque la primera es más formal, la segunda tiene la ventaja de facilitar los cálculos numéricos.

4.3 Ecuaciones de máxima verosimilitud:
Los datos de partida son las frecuencias observadas n_{ij} de la tabla, con i= *1,2,...,s* y *j = 1,2,...,r*. Sea π_{ij} la probabilidad de que una observación caiga en la celda (*i, j*). La verosimilitud de la muestra está dada por:

$$L = C\prod_i^s \prod_j^r \pi_{ij}^{n_{ij}} \quad \text{siendo } C \text{ una constante} \qquad (6)$$

Tomando el logaritmo $l = \ln L = \ln C + \sum_{i=1}^{s} \sum_{j=1}^{r} n_{ij} \ln \pi_{ij}$

Y aplicando las definiciones dadas para los umbrales a_i y b_j, se tiene:

siendo Φ_2 la función de distribución normal bivariada con correlación ρ.

El procedimiento para estimar simultáneamente tanto el coeficiente de correlación ρ como los umbrales $a_1, a_2,...,a_{s-1}, b_1, b_2,..., b_{r-1}$, requiere hallar las derivadas parciales de *l* respecto de cada parámetro. Si se aplica la segunda forma, una vez estimados los umbrales, como se mostró en el ejemplo presentado, se procede a estimar ρ por máxima verosimilitud con base en estos umbrales estimados. Las ecuaciones aplicadas son:

$$a_i = \Phi \ (P_{i.})$$
$$b_j = \Phi \ (P_{.j})$$

Siendo p_{ij} la proporción observada en la celda (*i, j*), y $p_{i.}$ y $p_{.j}$ las proporciones marginales acumuladas de la tabla conjunta de distribución de frecuencias. Además Φ_1 es la función de distribución normal univariada.

$$P_{i.} = \sum_{k=1}^{i} \sum_{j=1}^{r} p_{kj}$$

Para facilitar los cálculos se ha desarrollado un software denominado LISREL (Linear Structural Relations) Jöreskog [9].

5. METODOLOGÍA PARA LA APLICACIÓN DEL SEM

La metodología para la aplicación del modelo propuesto por López y Carrión [3] tomada de la teoría del SEM, es bastante general y podrá ser aplicada también en este caso con variables ordinales. Sus pasos se enumeran enseguida:

1. Establecer con los expertos del proceso, el diagrama relacional que represente el sistema.
2. Identificar los elementos del proceso, tanto endógenos como exógenos.
3. Formular el modelo matemático.
4. Establecer las matrices del modelo y sus especificaciones.
5. Analizar la identificación del modelo.
5. Preparar la matriz de los datos de entrada y comprobar si es definida positiva. El paquete de computador hace esta verificación.
6. Estimar las matrices de parámetros y de varianzas y covarianzas que intervienen en el modelo. Interpretación.
7. Evaluar el modelo mediante el cálculo y análisis de estadísticas tales como: correlaciones de las estimaciones, residuales estandarizados, estadístico x^2, raíz del cuadrado medio residual, coeficientes de determinación de las variables observadas.
8. Analizar los resultados obtenidos.

6. IDENTIFICACIÓN DEL MODELO

Un modelo es identificado si tiene solución única; si los parámetros pueden ser determinados unívocamente; esto es, si y sólo si para dos vectores de parámetros θ_1 y θ_2 se cumple que sus matrices de varianzas y covarianzas son iguales, en el sentido de que su diferencia no resulta estadísticamente significativa. Everitt [10].

La identificación consiste en generar valores únicos para los parámetros B y Γ involucrados en el lado derecho de las ecuaciones estructurales (3). Si el modelo no es identificado habrá un número indeterminado de conjuntos de parámetros que podrían generar la matriz Σ de las variables observadas.

7. ESTIMACIONES DE LOS PARÁMETROS

Las estimaciones de los parámetros se obtienen minimizando una función de ajuste definida como siendo $S - \hat{\Sigma}$ siendo S la matriz de varianzas y covarianzas muestrales y $\hat{\Sigma}$ la estimación de la matriz de varianzas y covarianzas poblacionales Σ, la cual es expresada en función de los parámetros del modelo:

$$\hat{\Sigma}=\begin{bmatrix}\hat{\Sigma}_{yy} & \hat{\Sigma}_{yx}\\ \hat{\Sigma}_{xy} & \hat{\Sigma}_{xx}\end{bmatrix}=\left[\begin{array}{c|c}\hat{\Lambda}_y(\hat{\Pi}\hat{\Phi}\hat{\Pi}'+\hat{\Psi}^*)\hat{\Lambda}'_y+\hat{\Theta}_\varepsilon & \hat{\Lambda}_y\hat{\Pi}\hat{\Phi}\hat{\Lambda}'_x\\ \hline \hat{\Lambda}_x\hat{\Phi}\hat{\Pi}'\hat{\Lambda}'_y & \hat{\Lambda}_x\hat{\Phi}\hat{\Lambda}'_x+\hat{\Theta}_\delta\end{array}\right] \quad (7)$$

con $\Pi = A\Gamma$ y $A=(I-B)^{-1}$ siendo (I-B) una matriz no singular.

El lado izquierdo de esta ecuación es la matriz estimada de varianzas y covarianzas de las variables observadas **x** e **y**, las cuales están expresadas, respectivamente, en función de las variables no observadas o factores ξ y η, en las ecuaciones (1) y (2). De otro lado, mediante trabajo algebráico con esas ecuaciones y la ecuación (3) se llega a expresar la $COV[\eta]$ como $\Pi\Phi\Pi'+\Psi^*$, cuyas estimaciones aparecen en el lado derecho de la ecuación (7). Detalles de estos cálculos pueden ser consultados en [3].

La ecuación (7) expresa entonces la matriz estimada de varianzas y covarianzas de las variables observadas Σ, como una función de las estimaciones de las cuatro matrices de parámetros: Λ_x, Λ_y, B y Γ, y de las cuatro matrices de varianzas y covarianzas de los factores y los errores: Φ, Ψ, Θ_δ y Θ_ε.

En el análisis se utilizan las variables medidas como desviaciones respecto de su media, lo cual lleva a ecuaciones de regresión con intercepto

nulo, indicando un desplazamiento del origen pero sin afectar las covarianzas entre las variables, las cuales en consecuencia se calculan como el valor esperado del producto de tales variables.

Entre los métodos de estimación se pueden mencionar Jöreskog [11] : Mínimos cuadrados no ponderados (ULS), Mínimos cuadrados generalizados (GLS) y de Máxima verisimilitud (ML), los cuales minimizan la forma cuadrática F respecto de θ:

$$F(\theta) = (s - \hat{\sigma})' w^{-1} (s - \hat{\sigma})$$

Siendo s y σ vectores conformados por los elementos no redundantes de las matrices S y $\hat{\Sigma}$ respectivamente. La matriz W es una matriz definida positiva que tiene como dimensión el número de elementos no redundantes de S. Los diferentes métodos de estimación tienen en común que:

- $F(S,\hat{\Sigma}) \geq 0$
- $F(S,\hat{\Sigma}) = 0 \Leftrightarrow S = \hat{\Sigma}$
- $F(S,\hat{\Sigma})$ es continua en S y en $\hat{\Sigma}$
- Si W es la matriz idéntica, F será la $tr\,(S-\hat{\Sigma})^2$, y dará origen al método de estimación llamado de Mínimos cuadrados no ponderados ULS.

- Si W es una función de los elementos de S se genera el método de Mínimos cuadrados generalizados GLS.

- Si W es una función de los elementos de $\hat{\Sigma}$ el método originado será el de Máxima verosimilitud ML.

La diferencia entre los métodos GLS y ML es que el segundo va cambiando con cada actualización durante la pasada por computador, porque depende de las estimaciones que va realizando, mientras que el método GLS no lo hace porque utiliza en la matriz W los valores alimentados como datos que los toma como fijos.

8. CONCLUSIONES

Con esta revisión se ha logrado dar soporte teórico a la extensión de la técnica multivariada SEM (Structural Equation Modeling) para el rediseño de procesos industriales que generen datos tanto continuos como discretos y ordinales. Como ya se ha mencionado, en la primera aplicación se logró la cuantificación de las relaciones entre las operaciones de un proceso y las características de calidad generadas, utilizando variables continuas; con la extensión a variables ordinales se ha generalizado el modelo. El grupo de investigación se encuentra trabajando en su aplicación a datos reales, obtenidos de procesos industriales.

9. REFERENCIAS

[1] Price, Barbara, Barth Bruce. "A structural model relating process inputs and final product characteristics". Quality Engineering, 7(4), 693-704 (1995).

[2] Chau, Patrick. "Reexamining a model for evaluating information center success using a structural equation modelling approach". Decision Sciences Volume 28 Number 2 Spring 1997.

[3] López, Aura y Carrión Andrés. "Cuantificación de Relaciones entre Operaciones de un Proceso y Características de Calidad". Epiciclos, Publicación bianual de la Facultad de Ingeniería Pontificia Universidad JAVERIANA Cali. Volumen , número 1, Enero 2002.

[4] LONG, J. Scott. COVARIANCE STRUCTURE MODELS. An introduction to LISREL. Sage publications. Beverly Hills, California. 1983.

[5] EVERITT, B.S. An introduction to latent variable models. Chapman and Hall Ltd. Great Britain 1984.

6] Hair, Joseph. Multivariate data analysis. Fourth edition. Prentice Hall, New Jersey 1995.

[7] Moustaki, Irini. "A review of exploratory factor analysis for ordinal categorical data". In: Cudeck, Robert, Stephen Du, Sörbom Dag. Structural equation modeling: present and future. SSI Scientific Software International. USA 2001.

[8] Olsson, Ulf. "Maximum likelihood estimation of the polychoric correlation coefficient". Psychometrika-Vol. 44 No. 4, December 1979.

[9] Jöreskog, K./ Sörbom, D. "*LISREL 8: User´s reference guide*. SSI Scientific Software International USA 1996.

[10] EVERITT, B.S. An Introduction to Latent Variable Models. University Press Cambridge.Great Britain 1984.

[11] Jöreskog, Karl/ Sörbom, Dag. PRELIS 2: User´s reference guide. SSI Scientific Software International USA 1996