anales de psicología, 1992, *8* (1-2), 123-130

Aplicación del análisis exploratorio de datos en los sistemas de ecuaciones estructurales

Lluis Salafranca i Cosialls(*)
Montserrat Freixa i Blanxart
Joan Guàrdia i Olmos

Universidad de Barcelona

Resumen: Uno de los aspectos más controvertidos en la utilización de los modelos de ecuaciones estructurales, reside en la evaluación de las medidas de ajuste global como indicador de la validez del modelo propuesto. En este trabajo se propone estudiar el efecto de las anomalias en las distribuciones de las variables originales y, en consecuencia, la utilidad del Análisis Exploratório de Datos, como fase previa a la estimación de parámetros de máxima verosimilitud. Para ello, a partir de un modelo estructural simple, se simularon matrices de correlaciones para obtener un ajuste perfecto y un desajuste tambien perfecto, modificándose las distribuciones originales con la inclusión de valores extremos. Los datos finales muestran que la existencia de anomalias como las descritas afectan al grado de significación final, reduciendo su valor, mostrandose como modelos ajustados presentan datos de ajuste cercanos a su rechazo estadístico.
Palabras clave: Análisis exploratorio de datos, ecuaciones estructurales, metodología

Abstract: Uno de los aspectos más controvertidos en la utilización de los modelos de ecuaciones estructurales, reside en la evaluación de las medidas de ajuste global como indicador de la validez del modelo propuesto. En este trabajo se propone estudiar el efecto de las anomalias en las distribuciones de las variables originales y, en consecuencia, la utilidad del Análisis Exploratório de Datos, como fase previa a la estimación de parámetros de máxima verosimilitud. Para ello, a partir de un modelo estructural simple, se simularon matrices de correlaciones para obtener un ajuste perfecto y un desajuste tambien perfecto, modificándose las distribuciones originales con la inclusión de valores extremos. Los datos finales muestran que la existencia de anomalias como las descritas afectan al grado de significación final, reduciendo su valor, mostrandose como modelos ajustados presentan datos de ajuste cercanos a su rechazo estadístico.
Palabras clave: Análisis exploratorio de datos, ecuaciones estructurales, metodología

Introducción

Es suficientemente conocida la extrema sensibilidad de la fase de estimación de parámetros en la técnica de los sistemas de ecuaciones estructurales (Guardia, 1986). De hecho, se han efectuado algunos trabajos dirigidos a la determinación de los puntos críticos que pueden conducir a soluciones finales inadecuadas. Por ejemplo, se han realizado estudios en relación con la distribución normal de las variables (Bentler, 1983); respecto al tamaño de muestra (Boomsma, 1985); incluso en torno a la falta de convergencia (Gerbing y Anderson, 1985) entre otros. Todos ellos coinciden en establecer unas características generales (tamaño de muestra suficien-

(*)**Dirección**: Dept. de Metodología de les Ciències del Comportament. Facultat de Psicología. Divisió de Ciències de la Salut. Universitat de Barcelona. Zona Universitària. 08028 Barcelona (Spain).

te, distribuciones normales, etc...) que permiten la obtención de estimaciones de máxima verosimilitud que no se vean afectadas por elementos ajenos a la definición del modelo. Piénsese que se trata de evaluar el modelo que se proponga, no de obtener datos sesgados que puedan llevar a soluciones inadecuadas.

Sin embargo, a pesar de estas prevenciones parece plantearse una cuestión que no está aún claramente superada. Dadas las características de la estimaciones de parámetros completas ("full technique"), se han obtenido soluciones inadecuadas a partir de matrices de correlaciones simuladas para conseguir un ajuste adecuado. Fornell y Larcker (1983) simularon matrices de correlaciones para efectuar posteriores estimaciones de máxima verosimilitud, de forma que los coeficientes de correlación que las integraban ajustaran a un modelo muy simple. Los datos de ajuste finales mostraron que un porcentaje superior al 25% no ofrecieron datos de ajuste correctos. En una réplica parcial de dicho trabajo (Guàrdia y Salafranca, 1991), se han encontrado valores parecidos de desajuste en matrices simuladas para que se obtuvieran estimaciones correctas. Ello sugiere que las precauciones que se han mencionado anteriormente no son suficientes para asegurar una estimación exenta de problemas. Debe señalarse que en los sistemas de ecuaciones estructurales, los valores que reflejan el ajuste global del modelo propuesto no gozan de una contrastada fiabilidad (Satorra y Saris, 1985), lo cual nos lleva a determinar que la fase de ajuste no está suficientemente definida estadísticamente. Se han propuesto modificaciones o alternativas para completar los análisis de ajuste basados en el estadístico X^2 (Satorra y Saris, 1985) o en la significación de estadísticos complementarios como el coeficiente de determinación global (Fornell, 1983).

De todo ello se desprende una idea más a añadir a los intentos de evaluar sistemas de ecuaciones estructurales. Se basa en el hecho de que todas las consideraciones efectuadas se dirigen o bien a la propia técnica (modificaciones de los índices de ajuste por ejemplo) o a las características del dato en base a las necesidades de la técnica en cuestión (tamaño de muestra, distribución normal, ...). Pero en ningún caso se hace referencia a la "calidad de los datos" (en términos de su distribución) como una condición a considerar para que se obtengan estimaciones de máxima verosimilitud ajustadas.

No es difícil establecer una argumento secuencial que parte de esa "calidad del dato" para conectarse con el resultado final en la solución de un "path diagrama" cualquiera. En efecto, supongamos que partimos como dato inicial para la estimación de parámetros máximo verosímil de una matriz R de correlaciones. Toda alteración en ese dato será fatal en las posteriores fases del trabajo. La "calidad" de la correlación, pues, vendrá determinada en último extremo por la distribución original de las variables que integren ese coeficiente. Un valor de correlación puede estar afectado en cuanto al verdadero parámetro que representa como consecuencia de una distribución con presencia de valores extraños, alteraciones o anomalías que, por supuesto, no siempre se reflejarán en una distribución no normal y que, evidentemente, están al margen del tamaño de muestra empleado. Por tanto, además de las precauciones de todo tipo citadas anteriormente, será preciso acotar las posibles fuentes de distorsión mediante un exhaustivo estudio exploratorio de las variables antes de su tratamiento final mediante la técnica de los sistemas de ecuaciones estructurales. A este respecto, las técnicas propiciadas por el Análisis Exploratorio de Datos (E.D.A.) (Tukey, 1971; Freixa, et Als.., 1992) parecen adecuadas para efectuar ese estudio inicial, con lo cual su uso como fase previa a la técnica de los modelos estructurales pueda ser evaluada.

A partir de todo lo anterior, se desprende que en este trabajo se pretende evaluar empíricamente la posible conexión entre las anomalías en los datos originales y los resultados finales de ajuste global en los modelos estructurales, a la vez que realizar una valoración en el uso de las técnicas E.D.A. en este contexto.

Presentación de un modelo simulado

Para la obtención de datos relacionados con el objetivo anteriormente citado, se seleccionó un modelo estructural simple que contemplara los elementos básicos de un modelo estructural, es decir, un modelo de medida exógeno, un modelo de media endógeno y, por supuesto, una estructura de covarianza. En el modelo establecido se evitó plantear covarianzas en errores ni efectos no recursivos, puesto que se ha mostrado la repercusión que tales especificaciones tienen en la estimación de paráme-

tros (Cliff, 1983; Berry, 1984). El modelo así establecido, se representa en la Figura 1.

En una primera fase realizada dentro de un trabajo más amplio (Guàrdia y Salafranca, 1991), se efectuaron 120 simulaciones de matrices de correlaciones de orden 5x5 siguiendo para ello el formato empleado por Fornell y Larcker (1983). Estas matrices se simularon para que cumplieran las exigencias necesarias para una estimación de máxima verosimilitud (Van Driel, 1978) y se generaron de modo que algunas de ellas supusieran unos coeficientes de correlación que llevaran a un ajuste global perfecto en los tres submodelos (exógeno, endógeno y propiamente estructural) mientras que otras ajustaban o bien a alguno/s submodelos o, finalmente, a ninguno de los submodelos mencionados. Para este trabajo se han seleccionado solamente las matrices que aseguran una ajuste global perfecto o que ofrecen un desajuste global en todos los submodelos. En definitiva se dispuso de 30 matrices iniciales, la mitad de las cuales daban ajuste al modelo y la otra mitad desajuste.

Se recuperaron para esas treinta matrices las distribuciones originales de las variables para estudiarlas de forma exploratoria y establecer posibles conexiones con los resultados del estadístico X^2 de ajuste global. Somos conscientes de que ese estadístico no parece el instrumento ideal para la evaluación global del modelo, como ya se ha comentado en la introducción de este trabajo, pero hemos preferido su uso a la incorporación de estrategias alternativas, ya que este trabajo supone una primera fase y en consecuencia, la inclusión de nuevos aspectos merece un abordaje secuencial.

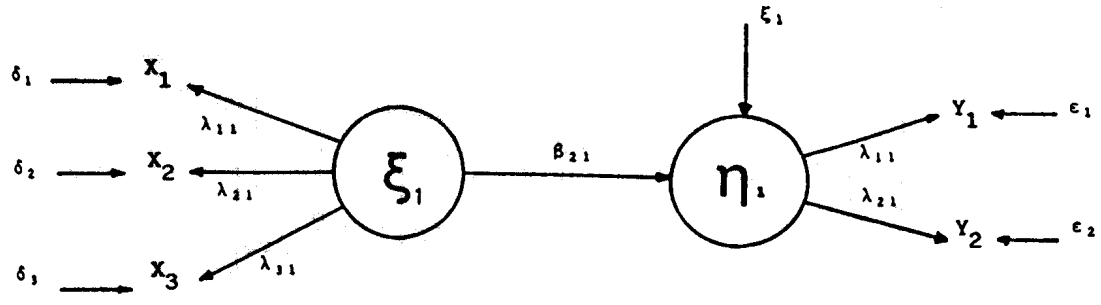

Figura 1.- "Path Diagrama" del Modelo simulado

Resultados

Todas las distribuciones simuladas presentaban, como se ha dicho, ajuste al modelo de la distribución normal, obteniéndose grados de significación entre 0.327 y 0.693 en la prueba de Kolmogorov-Smirnoff. Las 15 matrices de ajuste suponen 75 distribuciones a estudio (5 variables por matriz de correlaciones) y, evidentemente, se dispusieron de 75 distribuciones más las de las matrices de no ajuste. El promedio del grado de significación en la prueba de Kolmogorov para las primeras se situó en 0.394 (mediana de 0.412) y en las segundas de 0.413 (mediana de 0.396). Se puede asumir, en consecuencia, la no diferencia entre esos valores. Por otra parte, todas las distribuciones se efectuaron con un tamaño de muestra constante de 300 sujetos. Este valor asegura la no distorsión en la fase de estimación por muestra insuficiente. Por otra parte, como se ha mencionado, las treinta matrices eran susceptibles de estimación máximo verosímil (definidas positivamente e internamente consistentes).

Se manipularon las distribuciones originales para que contemplaran anomalías en su distribución. Tal manipulación consistió en provocar valores extremos de forma simétrica (tanto en la cola superior como en la inferior) de forma que un tercio de las distribuciones en cada condición (ajuste y no ajuste) presentaran un 5% de valores extremos (15 datos),

otro tercio un 10% (30 valores) y el tercio restante quedó sin modificar. Se seleccionó el 10% como valor superior ya que porcentajes superiores de manipulación se consideraron excesivamente perturbadores de la distribución original. Recuérdese que la importancia de la temática abordada reside en el hecho de evaluar la repercusión de la anomalía que por su aparente irrelevancia pasa inadvertida en la mayoría de casos y que la representación gráfica convencional de variables no permite recoge convenientemente. En la Figura 2 se muestran los diagramas de caja de una distribución de cada una de las condiciones definidas para establecer gráficamente el efecto de la manipulación realizada.

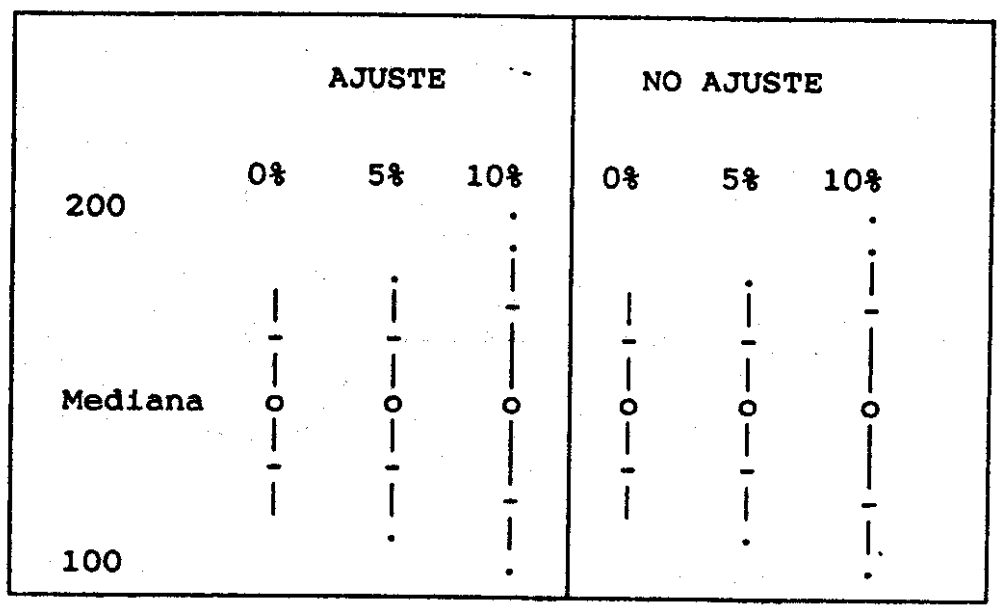

Figura 2.- Diagramas de Caja de algunas de las distribuciones.

Con las nuevas distribuciones (10 de ajuste y 10 de no ajuste) se repitieron los análisis de ajuste a la normalidad obteniéndose valores muy similares a los citados anteriormente, con lo que puede asegurarse que los porcentajes de valores extremos simulados no alteraron la distribución normal de las variables. Se consiguió un rango en los grados de significación entre 0.366 y 0.621, con una media de 0.466 y una mediana de 0.429. Por último se obtuvieron las estimaciones de parámetros y datos de ajuste global de X^2 para cada matriz, lo cual se efectuó mediante el programa informático Lisrel VII para microordenadores compatibles. La tabla número 1 muestra las medias de los grados de significación obtenidos para el estadístico X^2 para cada una de las condiciones.

Tabla 1: Medias de los grados de significación para cada una de las condiciones.

	SIN VALORES EXTREMOS	5% DE VALORES EXTREMOS	10% DE VALORES EXTREMOS	
AJUSTE	0.37	0.36	0.20	0.31
NO AJUSTE	0.03	0.03	0.02	0.02
	0.20	0.195	0.11	

El correspondiente Análisis de la Varianza (2x3) efectuado a partir de la tabla anterior indicó que tanto los efectos principales como la interacción de primer orden resultados altamente significativos.

Por lo que se refiere a las condiciones de ajuste mostró un efecto significativo (F=294.533; 1,24; p=0.000) lo que evidencia la adecuación en la simulación de las matrices de correlación según el procedimiento de Fornell y Larcker (1981). Como se desprende de los datos obtenidos el promedio en el grado de significación de X^2 se sitúa en 0.31 para las quince matrices en la condición de ajuste y en 0.03 en la condición de no ajuste.

El otro efecto principal, porcentaje de valores extremos, muestra también un efecto significativo (F=12.541; 2,24; p=0.000), situándose el promedio en el grado de significación en un valor de 0.20 en la distribución no manipulada, en un 0.19 para las distribuciones con un 5% de valores modificados y en 0.11 para un 10% de manipulación.

Por último, el efecto de interacción resultó igualmente significativo (F=9.484; 2,24; p=0.001), mostrándose su acción en la Figura 3.

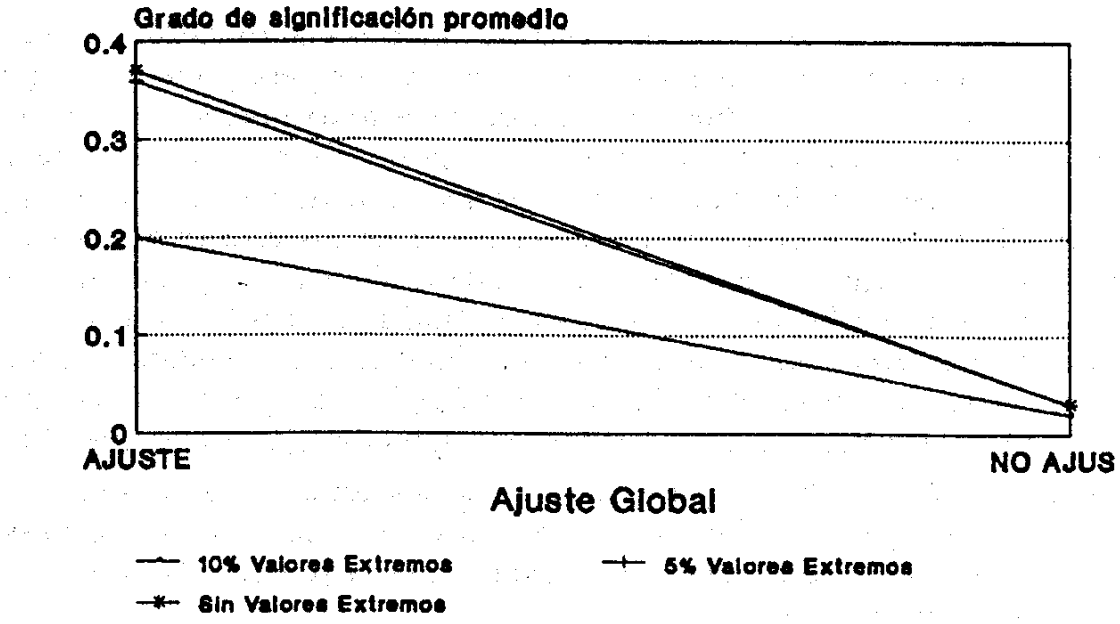

Figura 3: Efecto de interacción.

A partir de estos resultados se desprende la necesidad de un evaluación exhaustiva del efecto de las anomalías de las distribuciones originales en la estimación de parámetros para evaluar las consecuencias en el ajuste global en los sistemas de ecuaciones estructurales.

Discusión

Uno de los primeros aspectos a destacar, si bien se ha señalado la dificultad del ajuste global de los modelos estructurales, se centra en que en las matrices aquí empleadas no se ha constatado esa limitación. La significación encontrada y el sentido de la misma entre las condiciones de ajuste evidencia suficiente sensibilidad en el índice X^2 para la correcta evaluación global del modelo. Ello puede interpretarse, de acuerdo con algunos resultados presentados por Guàrdia y Salafranca (1991) como un efecto ligado a la simplicidad del modelo empleado. La limitación del valor X^2 puede verse aumentada por la complejidad del modelo. Es decir, las deficien-

cias en el uso de ese estadístico puede estar vinculado al número de parámetros a estimar en el modelo. De cualquier modo, en nuestro caso la simplicidad del modelo, el escaso número de variables y, consecuentemente, el poco número de parámetros ha favorecido una actuación correcta en el ajuste global, de acuerdo con la intención diseñada en la simulación de los datos. Por otro lado, no debe olvidarse que estamos utilizando modelos externos, es decir, de ajuste total o de desajuste total, lo cual favorece su reconocimiento en el ajuste global.

Por lo que se refiere a la influencia en el ajuste final por parte del porcentaje de valores modificados, parece claro pensar, a la vista de los datos, que el valor promedio en el grado de significación desciende ostensiblemente con el aumento del porcentaje de datos manipulados. En efecto, nótese que el grado de significación promedio en la condición del 10% se sitúa ligeramente por encima de la "mitad" del valor promedio conseguido en la condición de no modificación. En ambas condiciones extremas, los modelos serian, en general, aceptados como ajustados (no hacemos referencia aquí a su nivel de varianza explicada). Sin embargo, en la condición del 10% el valor promedio obtenido se sitúa claramente muy cerca del valor crítico de no ajuste (clásicamente situado en un valor de probabilidad de 0.10). Parece claro, en consecuencia que a pesar de simular matrices para conseguir un ajuste del modelo, la presencia de un 10% de valores extremos pone en serio peligro la posibilidad de detectar un modelo ajustado.

En la condición del 5% no se consigue una evidencia empírica de ese efecto, probablemente debido a que ese porcentaje de valores extremos no es suficiente para crear una auténtica anomalía. La prueba de Scheffé aplicada para comparar la primera y segunda condición no resultó significativa, mientras que la comparación de la segunda con la tercera si lo fue (p=0.000).

Por último, y de acuerdo con lo que se plantea en Satorra y Saris (1985) o en Cliff (1983), el problema del ajuste en los modelos estructurales se centra concretamente en el ajuste no en el desajuste. Ello se constata en nuestros datos si se analiza el efecto de interacción. En el se desprende que la condición de no ajuste no se ve afectada por la presencia de valores extremos. Ello sin embargo es precipitado, puesto que lo único evidente es que el modelo no es ajustado, ya sea por la simulación así efectuada como por la acción complementaria de los valores extremos. Pero parece más plausible pensar en el no ajuste como consecuencia de la simulación, puesto que se disponía de evidencia empírica anterior que así lo muestra (Guàrdia y Salafranca, 1991). Pero lo que es relevante, es el hecho de incidir en el decremento en el grado de significación a medida que aumentamos el porcentaje de valores extremos, en aquellas simulaciones dirigidas al ajuste. Así, la influencia de las anomalías en las distribuciones parecen estar ligadas al estudio del ajuste más que al del desajuste.

Esta última consideración hacer ver la importancia de los estudios exploratorios previos de la variables implicadas, puesto que la evidencia de anomalías y su subsanación (posibles transformaciones) permiten obtener grados de significación mayores con la repercusión que ello tiene con la posibilidad de cometer errores de tipo I.

Lógicamente, esto no es más que una primera evidencia, puesto que surgen cuestiones de debate más exhaustivo. Por ejemplo, el efecto concreto del porcentaje de valores extremos, repercusiones del tipo de transformación adoptada, importancia de la complejidad del modelo, asimetría de las alteraciones ...Como mínimo parece que la insistencia de las técnicas E.D.A. en conceptos como resistencia no parece gratuita en el estudio de sistemas de ecuaciones estructurales, lo cual, a nuestro entender, ya es suficiente argumento para recomendar que previamente a cualquier estimación de parámetros estructurales se someta a las variables a un estudio exploratorio. Asimismo, se percibe una excelente adecuación entre técnicas E.D.A. y estudio de los residuales en los sistemas de ecuaciones estructurales, que hasta el momento reciben un tratamiento un tanto convencional y no desligados del modelo lineal de la regresión.

Referencias

Bentler, P. (1983). Simultaneous equations systems as a moment structure models. *Journal of Econometrics*, 22, 13-42.

Berry, W.D. (1984). *Nonrecursive Causal Models*. Beverly Hills: Sage.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in Lisrel maximum likelihood estimation. *Psychometrika*, 50(2), 229-242.

Cliff, N.(1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18, 115-126.

Fornell, C. (1983). *Issues in the application of covariance structure analysis: A comment*. Manuscrito no publicado.

Fornell, C. & Larcker, D.F. (1981). Evaluating structural models with unobservables variables and measurement error. *Journal of Marketing Research*, 28, 39-50.

Freixa, M.; Salafranca, Ll.; Guàrdia, J.; Ferrer, R. y Turbany, J. (1992). *Análisis Exploratorio de Datos. Nuevas técnicas estadísticas*. Barcelona:PPU.

(Original recibido: 22-4-1991)
(Original aceptado: 18-11-1991)