

ESTIMACIÓN DEL TAMAÑO MUESTRAL REQUERIDO EN LA MODELIZACIÓN CON INDICADORES DE CALIDAD DE VIDA

(Estimated sample size required in the modeling with indices of quality of life)

Nahir Carballo¹, Edgar Jaimes², Neida Pineda², José G. Mendoza², Yolimar Garcés² e

Hilda

Rodríguez²

¹ Doctorante en Ambiente y Desarrollo. UNELLEZ San Carlos - Cojedes.

nahirdelc@gmail.com

² Grupo de Investigación de Suelo y Agua (GISA). ULA-NURR, Trujillo.

jaimes.5060@gmail.com;

pineida@ula.ve;

jgmendoz@ula.ve

Recibido: 11/01/16

Aceptado: 31/03/16

RESUMEN

El propósito de este trabajo es determinar el tamaño de muestra que sea suficiente para el ajuste de parámetros implícitos en la estructura de modelos multivariados basados en relaciones de causalidad. La investigación fue de tipo documental y experimental virtual. Para la simulación Monte Carlo, en el software estadística v.7, se utilizó una encuesta multipunto y discreta, con 80 ítems de medición de condiciones de vida urbana, con 408 casos. Los resultados mostraron, que el método no paramétrico de regresión, de distribución libre Asintótica, que no requieren el supuesto de multinormalidad, originaron una matriz de varianza-covarianza “no Gramiana”; igualmente el método robusto paramétrico de mínimos cuadrados generalizados-máxima verosimilitud, bajo el supuesto de data en escala categórica, originó una matriz de covarianza “no Gramiana”; por el contrario el simulado bajo el supuesto, de escala continua y multinormalidad, permitió realizar un excelente procedimiento de estimación muestral, indicando que con un mínimo de 150 casos, es un tamaño de muestra suficiente para el estudio de los indicadores de calidad en la comunidad del conjunto habitacional “Ezequiel Zamora” de San Carlos, estado Cojedes.

Palabras clave: Relaciones de causalidad, Tamaño de muestra, Simulación Monte Carlo, Calidad de Vida.

SUMMARY

The purpose of this study was to determine the sample size sufficient for setting parameters implicit in the structure of multivariate models based on causal relationships. The research was virtual documentary and experimental. A multipoint and discreet survey with 80 items measuring urban living conditions, with 408 cases was used for simulation Monte Carlo, in the software Statistica v. 7. The results showed that the nonparametric regression Asymptotic freely distributed, not requiring the assumption multinormality, they resulted in a variance-covariance matrix "not Grammian"; equally robust parametric method GLS-maximum likelihood under the assumption data in categorical scale, originated a covariance matrix "no Grammian"; by contrast the simulated under the assumption of continuous and multinormality scale allowed for an excellent procedure sample estimate, indicating that at least 150 cases, is a sufficient sample size for the study of the quality indicators in the community the housing complex "Ezequiel Zamora" in San Carlos, Cojedes state.

Keywords: Causal relationships, sample size, Monte Carlo Simulation, Quality of Life.

INTRODUCCIÓN

En el análisis de modelaje de relaciones de causalidad y en análisis multivariado en general, incluyendo el alfa de Cronbach (Gadermann, Guhn, & Zumbo, 2012), la data muestral requiere cumplir supuestos básicos, como son entre otros, normalidad multivariada, homocedasticidad, ausencia de valores extraños, de datos faltantes y de multicolinealidad entre ellos, así como de variables causales relevantes (Arias, 2008; Forero, Maydeu and Pujol, 2009; Hoyle, 2014); para ello, los datos requieren estar en escala numérica continua y de razón; sin embargo esto no siempre es posible, ya que en investigaciones en ciencias donde se miden datos de variables subyacentes, como la opinión, conducta o datos donde se categorizan medidas, se presentan escalas discretas, ya sean de conteo, nominales u ordinales. En esta situación, es necesario construir y utilizar matrices de correlación, que consideren la escala de medición de la variable involucrada, como es el caso de las matrices de Pearson, Poliserial, Policórica, de punto biseral, y la tetracórica; (Jöreskog, 2006, Choi, Peters & Mueller, 2010, Hoffmann, Stover, de la Iglesia y Liporace, 2013); que consideran las características de escala discreta. El propósito de este trabajo es determinar el tamaño de muestra que sea suficiente para el ajuste de parámetros implícitos en la estructura de modelos multivariados basados en relaciones de causalidad. Bajo estos supuestos está centrada la razón de esta investigación, en mostrar la implementación de una metodología de simulación Monte Carlo en la estimación del tamaño de muestra suficiente, con el software estadística v.7 (StatSoft, Inc., 2012), en el análisis de modelaje de relaciones de causalidad estructural, con datos discretizados de indicadores de calidad de vida tomados en la comunidad del conjunto habitacional “Ezequiel Zamora” de San Carlos.

REVISIÓN DE ANTECEDENTES

Los estudios que involucran modelación con relaciones de causalidad, constituyen métodos que tienen por objeto establecer evidencias de validez de la relación estructural entre indicadores con los factores (variables latentes o subyacentes) que explican un constructo de estudio. Para que estos procedimientos tengan lugar se debe partir de matrices de correlaciones que permitan estimar las covariaciones entre los elementos que componen la herramienta (Choi, Peters, & Mueller, 2010), siendo las más utilizadas las correlaciones

de Pearson, violándose en general sus supuestos paramétricos, cuando la escala es ordinal o dicotómica (Hoffmann, Stover, de la Iglesia y Liporace, 2013; Hoyle, 2014).

Para Lévy, Fuentes y González, (2006); Patiño y Muñoz, (2009); Ogasawara, (2011) y Hoffmann, Stover, de la Iglesia y Liporace, (2013), se debe evitar el uso generalizado de la matriz de correlación de Pearson, en estudios que involucran variables discretas. De esta manera, si todas ellas son continuas la matriz debe responder al modelo de Pearson. Si todas son dicotómicas se debe utilizar una matriz de correlación tetracórica (TCC). Si la totalidad es ordinal o algunas son ordinales y otras dicotómicas, corresponde el uso de una matriz de correlación policórica (PCC). Por otro lado, si algunas son continuas y otras son ordinales se emplea una matriz de correlación poliserial. De igual modo, se trabaja con una matriz de correlación punto biserial si se trata de variables continuas y dicotómicas.

Sin embargo, aun utilizando correlaciones policóricas y tetracóricas, se exige tener en cuenta una serie de supuestos propios de variables paramétricas, que de no cumplirse requieren de la adopción de ciertas precauciones metodológicas. Pueden mencionarse los de normalidad multivariada de los ítems, continuidad de la escala, presencia de ítems perturbadores estadísticos (varianzas nulas o no relevantes) y la relación entre amplitud de la muestra y número de variables observada; ya que estas condiciones pueden originar resultados extraños, que invalidan la investigación (Byrne, 2006).

Por otro lado, si no se adoptan precauciones metodológicas, la utilización de correlaciones policóricas y tetracóricas, suele producir con cierta frecuencia dos clases de problemas; a) la falta de convergencia, debido a que los métodos de ajuste no paramétricos generan una matriz varianza-covarianza no Gramiana y b) la ocurrencia de soluciones impropias (Cuttance, 2009). La no convergencia tiene que ver con la no estimación de valores positivos suficientemente pequeños obtenidos a partir de las interacciones entre los ítems y así obtener, con precisión, niveles de similitud entre el modelo propuesto y los datos provenientes de las variables observadas. Esto provoca que el algoritmo utilizado tenga dificultades para generar una matriz de covarianza que sea “definida positiva”, es decir que todos los valores sean superiores a cero, provocando un error de ejecución e interrumpiendo el análisis.

Los métodos de estimación más adecuados para el caso que se utilicen variables discretas serían los no paramétricos, como por ejemplo los de distribución libre asintótica (ADF), los robustos, tales como el de máxima verosimilitud robusto (ML Robusto) y mínimos cuadrados ponderados robustos (WLS robusto), (Bollen, 1989; Brown, 2006; Christoffersson, 1975; Flora & Curran, 2004; Holgado Tello et al., 2010; Jöreskog & Sörbom, 1999; Kline, 2005). Su uso se sustenta en que los resultados no sufren alteraciones significativas aun cuando se viole el supuesto de normalidad (Patiño Muñoz, 2009). No obstante, su aplicación requiere contar con muestras suficientemente amplias para obtener estimaciones consistentes ($N > 200$) (Bollen, 1989; Schermelleh-Engel, Moosbrugger & Müller, 2003; Hoyle, 2014).

Con base en las consideraciones anteriores, la estimación del tamaño de muestra suficiente (TMS), para la obtención de un modelo con excelente ajuste es complejo y tedioso, motivos por los cuales es recomendable utilizar procedimientos de simulación con software, como los que se utilizaron en este trabajo

MATERIALES Y MÉTODOS

La data utilizada para este estudio fueron colectados a través de un modelo de encuesta elaborada por Jaimes, (2015); misma que fue aplicada para evaluar la calidad de vida de 408 familias del área residencial “Los Iranés”, localizada en el municipio San Carlos, del estado Cojedes. En total se levantó información de ochenta (80) indicadores de calidad de vida, los cuales se estimaron con base en una escala discreta multipunto, con valores variables entre uno (01) y cinco (05).

Se aplicó un análisis estadístico descriptivo, basado en la estimación de promedios robustos, tales como el de Huber; el Bponderado de Tukey; el Redescendente de Hampel y el Onda de Andrews, con el objeto de poder determinar los valores atípicos e ítems perturbadores estadísticos

En lo referido a la estimación del tamaño de muestra suficiente, para el análisis de modelación de relaciones de causalidad estructural, se utilizó la simulación Monte Carlo, incluido en el paquete *statisticav.7*, mismo que se fundamenta en la construcción teórica del modelo operativo estructural, el cual se digitalizó con la sintaxis de lenguaje de rutas de causalidad PACH1, en el cuadro de dialogo del módulo de simulación Monte Carlo, antes indicado; y se sometió a las siguientes pruebas estadísticas:

- a) Carga de los factores: Estimado de correlación entre el factor y la variable, relativamente alto, iniciando con valores de arranque de 0,6;
- b) Valores iniciales de varianzas únicas que incluye valores para los errores experimentales o residuales, de 0,64;
- c) Correlaciones nulas entre factores y
- d) Un tamaño inicial de muestra relativamente pequeño.

Las simulaciones se realizaron con método GLS-ML, primeramente suponiendo escala continua y multivariada, y luego asumiendo una escala discreta y no multivariada. Así, cuando se ejecutó cada evento de simulación del algoritmo Monte Carlo, el procedimiento leyó las exigencias, tratando la data como una población de parámetros, generando muestras a partir de la matriz de correlación, correspondiente a los valores de arranque dados.

El procedimiento se ejecutó, en forma consecutiva, aumentando cada vez el tamaño de muestra; hasta que dejó de aparecer casos fuera del límite de restricción de desigualdad que usa el algoritmo. Cada resultado fue registrado (casos fuera de límite de restricción: casos boundary y tamaño de muestra ensayado, complementando el proceso con la estimación de la precisión de estimación de los parámetros del modelo, con graficas de parámetros (PAR_i) hasta que alcanzaron la normalidad de Gauss, con la consecuente disminución de la desviación estándar.

RESULTADOS Y DISCUSION

El análisis descriptivo con estimadores robustos, arrojó que, de los 80 ítems medidos, 19 son perturbadores estadísticos (PEs), ya que presentaron alta centralidad alrededor de su promedio robusto, y otros poseen varianza nula; estos PEs fueron sustraídos del análisis de confiabilidad de ítems alfa de Crombach y del modelamiento de relaciones de causalidad, y estudiados aparte. La confiabilidad interna de indicadores de calidad de vida (Tabla 1), para los 408 casos, muestra una relativa alta confiabilidad; es decir, de 71,72 %; con una inter correlación entre ítems de 25,68 %.

Tabla 1. Alfa de Crombach, calculado con matriz correlación policórica.

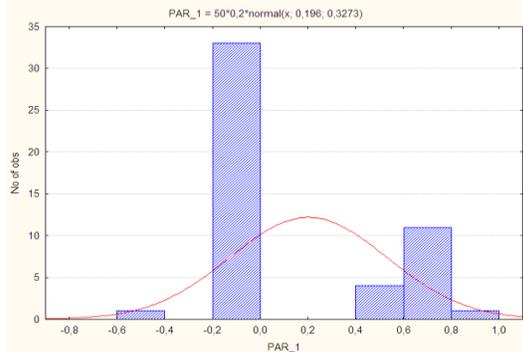
Variable Factor	Alfa de Crombach [%]		Correlación inter ítems [%]
	Regular	Estandarizado	
F1	57,26	59,00	15,44
F2	68,65	68,27	31,05
F3	62,40	63,27	22,76
F4	78,69	79,37	28,16
F5	72,28	72,24	19,88
F6	81,17	81,53	29,95
F7	78,18	78,57	32,55
Promedio	71,23	71,75	25,68
Total (F1-F7)	92,35	92,55	17,17

El modelo de relaciones de causalidad estructural se elaboró con base en el lenguaje PATH1, incluido en el aplicativo Monte Carlo (Referencia bibliográfica) utilizando los datos indicados en el cuadro 1. Teniendo como referencia las restricciones establecidas en la metodología; es decir, correlaciones nulas entre factores, correlación factor-ítems $> 0,6$ y varianza de error $< 0,64$; se muestran en la figura 1 el resumen de resultados obtenidos; a saber:

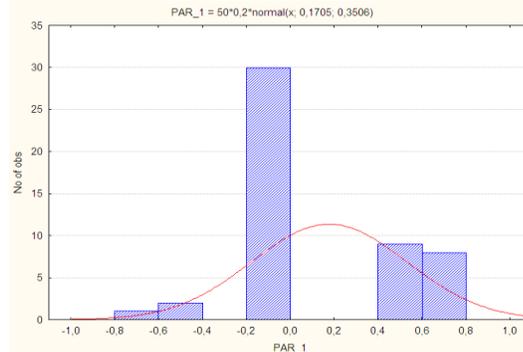
1.) cuando el Tamaño de Muestra Suficiente (TMS) alcanza las 80 familias encuestadas, se observa un amplio valor de Desviación Estándar (DE), el cual es un reflejo del grado de dispersión de los indicadores de calidad de vida que se evalúan en esas encuestas; 2.) Situación que se mantiene para un TMS = 85 y 90 casos y 3.) Cuestión que cambia, de manera significativa, para un TMS = 150 toda vez que se alcanza un valor adecuado de la DE = 0,073, valor de dispersión que refleja un buen ajuste de los indicadores de calidad de vida evaluados los cuales comienzan a mostrar una tendencia consistente hacia una distribución normal; misma que se mantiene, y se hace más evidente, cuando los TMS = 300 y 408 casos.

En síntesis, de acuerdo con estos resultados es obvio que la metodología utilizada en este estudio fue efectiva para determinar un tamaño de muestra suficiente y adecuado (TMS = 150) para validar los resultados que se puedan obtener a partir del modelo de encuesta propuesta por Jaimes (2015)

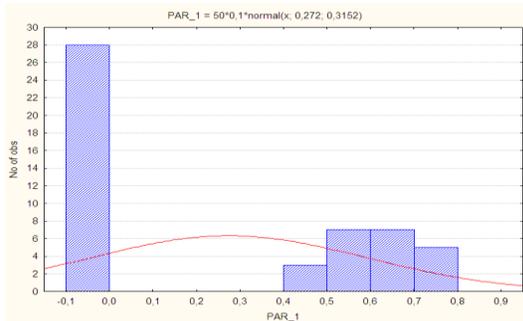
TMS= 80; Desviación estándar: 0,327
Parámetros fuera de restricción: 7



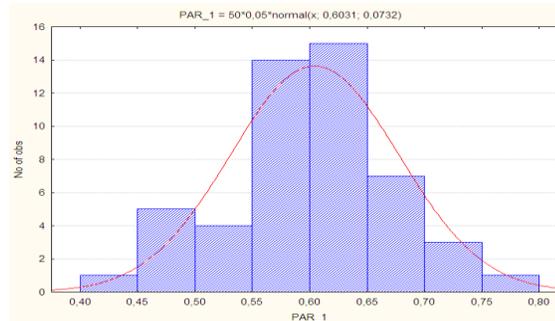
TMS= 85; Desviación estándar: 0,351
Parámetros fuera de restricción: 5



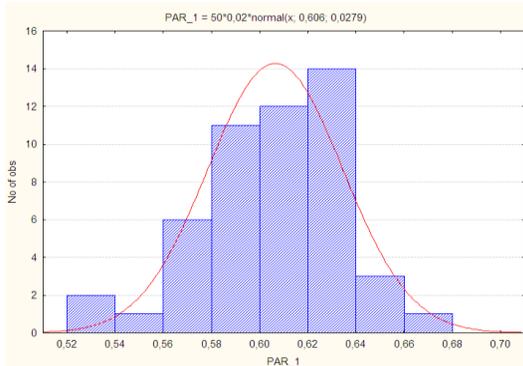
TMS = 90; Desviación estándar: 0,315
Parámetros fuera de restricción: 0



TMS = 150; Desviación estándar: 0,073
Parámetros fuera de restricción: 0



TMS = 300; Desviación estándar: 0,028
Parámetros fuera de restricción: 0



TMS = 408; Desviación estándar: 0,025
Parámetros fuera de restricción: 0

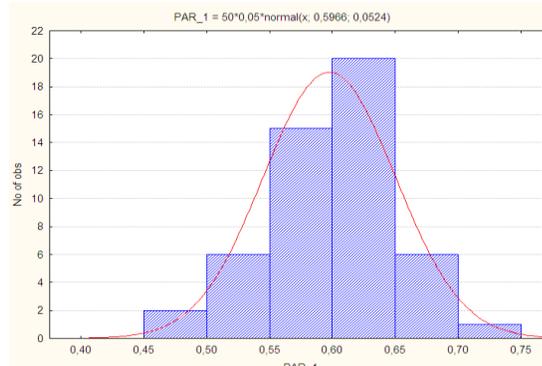


Figura 1. Resultados de la simulación Monte Carlo

CONCLUSIONES

La estimación de consistencia interna del instrumento alfa de Crombach, usando la matriz de correlación policóricas para datos discretizados, mostró excelentes resultados.

En relación al procedimiento de análisis descriptivo, usando estimadores robustos, estos permitieron caracterizar y descartar los ítems perturbadores estadísticos y determinar los ítems relevantes; que mejoró sustancialmente el tiempo computacional de simulación

Monte Carlo y la no aparición de resultados extraños, obteniéndose excelente convergencia del proceso de ajuste del modelo a los datos, con muy buena precisión; usando el método de regresión de mínimos cuadrados generales y el de máxima verosimilitud.

En cuanto a la metodología de simulación Monte Carlo, incluido en el paquete de *statistica 7*, utilizado en la estimación del tamaño de muestra suficiente (TMS) con datos en escala discreta para definir una modelación de relaciones de causalidad con base en indicadores de calidad de vida, mostró ser el adecuado en la obtención de los parámetros del modelo y con excelente precisión cuando se combina con el método robusto paramétrico GLS-ML (mínimos cuadrados generalizados-máxima verosimilitud), simulado bajo el supuesto, de escala continua y multinormalidad, ya que permitió realizar una excelente procedimiento de estimación muestral, obteniéndose un TMS = 150 casos.

Agradecimiento: Al CDCHTA-ULA por el apoyo financiero aprobado para el Proyecto NURR-C-588-15-01-B, con base en el cual se realizó el presente trabajo a través de la participación de los integrantes del GISA-NURR-ULA.

REFERENCIAS BIBLIOGRÁFICAS

- Arias, B. 2008. Desarrollo de un ejemplo de análisis factorial confirmatorio con LISREL, AMOS y SAS. <http://www.benitoarias.com/articulos/afc.pdf>.
- Bollen, K.A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.
- Brown, T. 2006. *Confirmatory Factor Analysis for Applied Research*. New York: Guildford Press.
- Byrne, B. 2006. *Structural Equation Modeling with EQS*. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.
- Choi, J, Peters, M., & Mueller, R. (2010). Correlational analysis of ordinal data: from Pearson's r to Bayesian polychoric correlation. *Asia Pacific Educ.Rev*, 11, 459-466. doi: 10.1007/s 12564-010-9096-y.
- Christoffersson, A.1975. Factor analysis of dichotomized variables. *Psychometrika*, 40 (1): 5-32.
- Cuttance, P. 2009. Issues and problems in the application of structural equation models. En P. Cuttance & R. Ecob (Eds.), *Structural Modeling* (pp. 241-280). New York: Cambridge University Press.
- Flora, D., & Curran, P. 2004. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4): 466-491.
- Forero, C. G., Maydeu-Olivares, A. and Gallardo-Pujol. D. 2009. Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. *Structural Equation Modeling*. 16:625-641. http://www.ub.edu/gdne/amaydeusp_archivos/uls%20vs%20dwls%20sem09.pdf.
- Gadermann, A. M., Guhn, M., & Zumbo, D. (2012). Estimating ordinal reliability for likert-tipe and ordinal item response data: a conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, 17(3), 1-13.

- Hoffmann, A. F., Stover, J. B., de la Iglesia, G. y Liporace, F. M. (2013). Correlaciones policóricas y tetracóricas en estudios factoriales exploratorias y confirmatorias. *Ciencias Psicológicas*, VII (2): 151–164.
- Holgado Tello, F., Chacón Moscoso, S., Barbero García, I., & Vila Abad, E. 2010. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, 44(1): 153-166.
- Hoyle, H. R. 2014. *Handbook of Structural Equation Modeling*. Guilford Publications P.740.
- Jaimes, E. 2015. Un modelo para evaluar la calidad de vida; en: *Análisis del Deterioro Agroecológico y Ambiental de la subcuenca del río Déleg, provincia del Cañar, República del Ecuador*. (Informe final, mimeografiado). Proyecto Prometeo- SENESCYT 28 p.
- Jöreskog, G. K. 2006. Factor analysis ordinal variables with full information maximum likelihood. <http://www.ssicentral.com/lisrel/techdocs/orfiml.pdf>
- Jöreskog, K.G., & Sörbom, D. 1999. *LISREL 8: user's reference guide*. Lincolnwood, IL: Scientific Software International, Inc.
- Lévy, M. J.P., Fuentes, M.T. y González, M.V. 2006. Optimización según estructuras de covarianzas. En J.P. Lévy Mangin y J. Varela (Eds.), *Modelización con estructuras de covarianzas en ciencias sociales* (pp. 11-30). Coruña: Netbiblo.
- Ogasawara, H. 2011. Asymptotic expansions of the distributions of the polyserial correlations coefficients. *Behaviormetrika*, 38(2):153-168.
- Patiño M. A. y Muñoz, Z. S. 2009. *Sistema de ecuaciones estructurales: una herramienta de investigación*. Centro Nacional de Evaluación para la Educación Superior, A.C. Ed. Ceneval. Mexico. D.C. 104. P.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. 2003. Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.