UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

ESCUELA DE ESTADÍSTICA

EFECTOS DE LA KURTOSIS EN LA ESTIMACIÓN DE MODELOS DE ECUACIONES ESTRUCTURALES BAJO DISTINTOS TAMAÑOS DE MUESTRA

CÉSAR ANDRÉS GAMBOA SANABRIA B12672

ANDRÉS ESTEBAN ARGUEDAS LEIVA B40535

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

RESUMEN

El Resumen

Palabras clave:SEM, simulación, kurtosis, lavaan, R

ABSTRACT

El resumen en inglés

Palabras clave:SEM, simulation, kurtosis, lavaan, R

Índice

1 INTRODUCCIÓN

1.1 Antecedentes

Los Modelos de Ecuaciones Estructurales (en adelante SEM, por sus siglas en inglés) representan un compendio de métodos estadísticos que buscan estimar y examinar las relaciones existentes entre varias mediciones fácilmente observables con conceptos más abstractos, denominados constructos, que no pueden ser medidos ni analizados de manera directa. Los SEM trabajan de una manera similar a los modelos de regresión más clásicos, pero representan una mejora pues analizan las relaciones causales lineales entre las variables involucradas al mismo tiempo que los errores de medición (Beran & Violato, 2010).

Los SEM están presentes en multitud de campos de investigación. Según Beran y Violato (2010), la cantidad de referencias a SEM en 1994 fueron de 164, aumentaron a 343 en el 2000 y llegaron a 742 en el 2009, lo cual es una señal de que muchos investigadores alrededor del mundo están mostrando cada vez más interés en este tipo de estudios, pues representan una potente herramienta para la investigación partiendo de la teoría sustantiva que poseen los diversos estudios.

Uno de los principales campos de aplicación de los SEM son las ciencias sociales, pues se busca explicar y/o predecir con un grado de validez el comportamiento específico de una o varias personas en grupo. Teniendo siempre en consideración (aunque de forma limitada) las condiciones que afectan a cada individuo involucrado en el estudio, así como las características propias de su entorno, los grupos de investigación pueden definir factores y relaciones latentes que se encuentran implícitas en el comportamiento humano. Este tipo de investigaciones permite entender los fenómenos no solo de forma descriptiva, sino que es posible también determinar relaciones de causalidad (Tarka, 2018).

Las variables indicadoras que sirven para construir las relaciones implícitas en cuestión, llamadas comúnmente constructos, pueden llegar a comportarse de manera muy diversa. Las ciencias sociales, al trabajar con seres humanos, es común trabajar con variables cuyo comportamiento es particularmente irregular, presentando valores muy distintos entre los sujetos de estudio, generando de esta manera que los indicadores de manera multivariada no sigan una distribución normal, lo cual representa un supuesto fundamental al trabajar con SEM (Sura-Fonseca, 2020). El no cumplimiento de este supuesto puede deberse, entre otras cosas, a medidas particularmente altas o bajas de una medida estadística en específico: La kurtosis.

1.2 El problema

Si al trabajar con un SEM no se cumple el supuesto de normalidad multivariada y además el modelo se estima vía máxima verosimilitud podría cometerse el error de sobreestimar el estadístico chi-cuadrado, el cual sirve de referencia para conocer la magnitud de la diferencia entre la matriz de covariancias estimadas por el modelo con la obtenida en la muestra. Lo anterior suele llevar al rechazar modelos que en realidad resumen bien la realidad y además a la subestimación de los errores asociados a los parámetros, lo cual genera interpretaciones inadecuadas en lo referente a la significancia de las relaciones planteadas por el modelo teórico. Considerar distintos niveles de kurtosis permite conocer el impacto que esta medida tiene sobre las estimaciones de un SEM dependiendo del tamaño de muestra utilizado (Muthen & Kaplan, 1992).

1.3 Objetivos del estudio

La presente investigación busca estudiar el efecto que tienen distintos niveles de kurtosis en varios tamaños de muestra sobre las estimaciones de un SEM. Para ello, tomando como base un estudio de la Universidad de California (Gao, Mokhtarian, & Johnston, 2008) se plantean los siguientes objetivos:

1.3.1 Objetivo general

Comparar mediante un estudio de simulación las estimaciones de modelos de ecuaciones estructurales en presencia de variables observadas con niveles de kurtosis de 0, 0.62, 6.65, 21.41 y 13.92 en tamaños de muestra de 50, 100, 200, 400 y 800.

1.3.2 Objetivos específicos

1) Definir como modelo poblacional el obtenido por Sura-Fonseca (2020) con dos variables exógenas y una endógena con tres variables indicadoras cada uno **(página 99 de la tesis)** como modelo de referencia teórico cuyas cargas factoriales se utilizarán para la generación de los datos simulados.

2) Medir el posible sesgo causado en la estimación de los modelos mediante el estadístico chi-cuadrado del modelo y la raíz del cuadrado medio de error de aproximación (RMSEA), la raíz de residuos de cuadrado medio estandarizado (SRMR) y el índice de bondad de ajuste (GFI).

3) Comparar los valores poblacionales de las cargas factoriales con los obtenidos en las simulaciones.

4) Publicar en una revista científica con revisión por pares el manuscrito final, en forma de un artículo científico.

1.4 Metodología de la investigación

AQUÍ VA UN RESUMEN DE LO QUE PONEMOS EN LA SECCIÓN DE METODOLOGÍA

1.5 Organización del estudio

AQUÍ VA UNA DESCRIPCIÓN BREVE DE CADA ETAPA DEL TRABAJO

2 METODOLOGÍA

2.1 Casos de simulación

2.2 Generación de datos con kurtosis

Los datos fueron simulados mediante la función `simulateData()` del paquete `lavaan` (Rosseel, 2012), el cual utiliza el método propuesto por Vale y Maurelli (Vale & Maurelli, 1983) para la simulación de datos no normales multivariados. Este método, comúnmente conocido como VM, se basa en el método propuesto por Fleishman (Fleishman, 1978), el cual, con base en una variable aleatoria distribuida como una normal estándar, permite simular una variable con un promedio, variancia, asimetría y kurtosis dada. El método VM permite especificar, adicionalmente, correlaciones entre las variables a estimar. Para utilizar el método de Fleishman, para generar una cierta variable aleatoria Y, se utiliza la siguiente ecuación:

$$Y = a + bX + cX^2 + dX^3 \tag{1}$$

donde $X \sim \mathcal{N}(0, 1)$. Es decir, se puede generar una variable no normal Y, con sus primeros cuatro momentos iguales a valores especificados, con base en los valores a, b, c y d de la ecuación 1, con base en una variable normal estándar X hasta su tercer potencia. Luego, para poder obtener los valores de a, b, c y d, se necesitan resolver las siguientes ecuaciones de forma simultánea:

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0 \tag{2}$$

$$2c(b^2 + 24bd + 105d^2 + 2) - \gamma_1 = 0 \tag{3}$$

$$24\left(bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)\right) - \gamma_2 = 0 \tag{4}$$

donde γ_1 es la asimetría deseada y γ_2 es la kurtosis deseada, además se define $a = -c$. Con base en las constantes calculadas a, b, c y d, además de una variable normal estándar, se puede simular variables no normales. Para poder generalizar el método de Fleishman a variables aleatorias multivariantes, Vale y Maurelli proponen una generalización. Esta se basa, para el caso bivariado, en la generación de dos variables aleatorias independientes, $X_1, X_2 \sim \mathcal{N}(0, 1)$, para la cuales se obtienen las constantes a, b, c y d, para cada una de dichas variables, como se describe en el método de Fleishman, obteniendo así el vector $w'_1 = (a_1, b_1, c_1, d_1)$, para el caso de X_1, y el vector $w'_2 = (a_2, b_2, c_2, d_2)$, para el caso de X_2. Además, se definen los vectores $x'_1 = (1, X_1, X_1^2, X_1^3)$ y $x'_2 = (1, X_2, X_2^2, X_2^3)$. Por lo tanto, se pueden crear variables no normales, Y_1 y Y_2, como:

$$Y_1 = w'_1 x_1$$

$$Y_2 = w'_2 x_2$$

donde se puede verificar que:

$$
\begin{aligned}
r_{Y_1,Y_2} =& \rho_{X_1,X_2}(b_1b_2 + 3b_1d_2 + 3d_1b_2 + 9d_1d_2) \\
&+ \rho^2_{X_1,X_2}(2c_1c_2) + \rho^3_{X_1,X_2}(6d_1d_2)
\end{aligned}
$$

Y resolviendo esta ecuación en términos de ρ_{X_1,X_2} se puede obtener una matriz de correlaciones para generar datos normales multivariados, que pueden ser transformados en variables no normales mediante el método de Fleishman.

2.3 Modelo teórico a estimar

2.4 Medidas de bondad de ajuste

2.5 Simulación y estimación

La simulación de los datos, junto con la estimación de los modelos, se realizó mediante el paquete `lavaan` (Rosseel, 2012) usando el software R (R Core Team, 2020) mediante la interfaz gráfica de RStudio (RStudio Team, 2015). Para e manejo de bases de datos y demás visualizaciones fueron utilizados los paquetes `ggplot2`(Wickham, 2016), `tidyr` (Wickham & Henry, 2020), `dplyr` (Wickham et al., 2020), `ggpubr` (Kassambara, 2020), `PerformanceAnalytics` (Peterson & Carl, 2020) y `kableExtra` (Zhu, 2019).

Para poder realizar la simulación deben seguirse varios pasos. Lo primero es definir el modelo teórico poblacional que van a seguir los datos simulados, como se describió en secciones anteriores este modelo cuenta con dos variables exógenas y una endógena, cada una con tres variables indicadoras. Los datos se generan mediante la función `simulateData()` la cuál requiere especificar varios argumentos, uno de ellos es el modelo poblacional, cuya sintaxis puede encontrarse en el Código 1. Los otros dos argumentos a especificar son el tamaño de muestra deseado y el nivel de kurtosis de interés, la definición de estos escenarios se obtiene mediante el Código 2 y se muestran en el cuadro 1:

Cuadro 1: Escenarios de simulación

kurtosis	n	kurtosis	n	kurtosis	n	kurtosis	n	kurtosis	n
0.00	50	0.00	100	0.00	200	0.00	400	0.00	800
0.62	50	0.62	100	0.62	200	0.62	400	0.62	800
6.65	50	6.65	100	6.65	200	6.65	400	6.65	800
21.41	50	21.41	100	21.41	200	21.41	400	21.41	800
13.92	50	13.92	100	13.92	200	13.92	400	13.92	800

Fuente: Elaboración propia a partir del estudio de la Universidad de California (2008)

Con estos escenarios definidos, se generaron entonces, para cada combinación de tamaño de muestra y kurtosis un total de 2000 conjuntos de datos para cada uno. La sintaxis necesaria para generar esto se muestra en el Código 3. Una vez que se obtuvieron estos conjuntos de datos, el siguiente paso es realizar la estimación de los SEM con cada uno de ellos; para ello es necesario definir un modelo de una forma similar a como se indicó en el Código 1, pero esta vez sin los valores de las cargas factoriales, pues se busca conocer las estimaciones a partir de los datos generados, este proceso se muestra en el Código 4.

3 RESULTADOS

3.1 Introducción

4 CONCLUSIONES Y RECOMENDACIONES

4.1 Introducción

4.2 Conclusiones

4.3 Recomendaciones

5 ANEXOS

5.1 Modelo poblacional para la función simulateData()

Código 1: Modelo poblacional para simular datos

```
modelo <- ' CA =~ 0.88*x1 + 0.77*x2 + 0.73*x3

            PHC =~ 0.85*x4 + 0.80*x5 + 0.82*x6

            VE =~ 0.62*x7 + 0.74*x8 + 0.72*x9


            VE ~ -0.01*CA + 0.41*PHC

          '
```

5.2 Escenarios de simulación para el tamaño de muestra y kurtosis

Código 2: Combinaciones de tamaño de muestra y kurtosis

```
casos <- expand.grid(kurtosis=c(0, 0.62, 6.65, 21.41, 13.92), n=c(50, 100, 200, 400, 800))
```

5.3 Simulación de datos

Código 3: Generación de datos para cada escenario

```
datos <- lapply(1:2000, function(x){

  data <- mapply(simulateData, sample.nobs=casos$n, kurtosis=casos$kurtosis,

                 MoreArgs = list(model=modelo),

                 SIMPLIFY = FALSE)


  names(data) <- paste("n", casos$n, "k", casos$kurtosis, sep="")

  data

})


casos_resultados <- expand.grid(x=1:length(datos), y=names(datos[[1]]))


nombres <- unique(casos_resultados$y) %>% paste

datos <- lapply(nombres, function(y){

  lapply(1:length(datos), function(x){

    datos[[x]][[y]]

  })

})
```

```
names(datos) <- nombres
```

5.4 Estimación de los SEM

Código 4: Definición del modelo y estimación para cada conjunto de datos

```
modelo <- ' CA =~ x1 + x2 + x3
            PHC =~ x4 + x5 + x6
            VE =~ x7 + x8 + x9
            VE ~ CA + PHC '

modelos <- lapply(datos, function(x){
  lapply(x, function(y){
    sem(modelo, data=y)
  })
})
```

6 REFERENCIAS

Beran, T. N., & Violato, C. (2010). Structural equation modeling in medical research: a primer. *BMC Research Notes*, *3*, 267-267. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2987867/#

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521-532. https://doi.org/10.1007/BF02293811

Gao, S., Mokhtarian, P. L., & Johnston, R. A. (2008). Nonnormality of Data in Structural Equation Models. *Transportation Research Record*, *2082*(1), 116-124. Recuperado de https://doi.org/10.3141/2082-14

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. Recuperado de https://CRAN.R-project.org/package=ggpubr

Muthen, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 19-30. Recuperado de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8317.1992.tb00975.x

Peterson, B. G., & Carl, P. (2020). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*. Recuperado de https://CRAN.R-project.org/package=PerformanceAnalytics

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Recuperado de https://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. Recuperado de http://www.jstatsoft.org/v48/i02/

RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Recuperado de http://www.rstudio.com/

Sura-Fonseca, R. (2020). *Modelos de ecuaciones estructurales: consecuencias de la asimetría positiva en los indicadores* *endógenos sobre las estimaciones puntuales de sus coeficientes y la bondad de ajuste*. Recuperado de http://www.kerwa.ucr.ac.cr/handle/10669/80716

Tarka, P. (2018). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity: International Journal of Methodology*, *52*(1), 313-354. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5794813/

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465-471. https://doi.org/10.1007/BF02293687

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Recuperado de https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. Recuperado de https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2020). *tidyr: Tidy Messy Data*. Recuperado de https://CRAN.R-project.org/package=tidyr

Zhu, H. (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. Recuperado de https://CRAN.R-project.org/package=kableExtra