Revista Colombiana de Estadística
Edición para autor(es)

The Effects of Kurtosis on the Estimation of Structural Equation Modelos Over Different Sample Sizes

Efectos de la Kurtosis en la estimación de Modelos de Ecuaciones Estructurales bajo distintos tamaños de muestra

César Gamboa-Sanabria[1,a], Andrés Arguedas-Leiva[1,b]

[1]School of Statistics, Faculty of Economical Sciences, University of Costa Rica, San José, Costa Rica

Abstract

Insert your abstract here.

***Key words*:** SEM, simulation, kurtosis, lavaan.

Resumen

Inserte su resumen aquí.

***Palabras clave*:** SEM, simulación, kurtosis, lavaan.

1. Introduction

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 causales 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). Para medir estas relaciones causales, los SEM cuentan con dos grandes componentes: 1) el modelo estructural, cuya función es cuantificar las relaciones causales presentes entre cada

[a]School of Statistics, University of Costa Rica. E-mail: info@cesargamboasanabria.com
[b]School of Statistics, University of Costa Rica. E-mail: andres.arguedasleiva@ucr.ac.cr

uno de los constructos planteados desde la teoría; y 2) un modelo de medición, cuyo objetivo último es brindar una descripción acerca de cuáles son los indicadores que sirven para medir los constructos en cuestión (Kaplan 2012).

Los SEM están presentes en multitud de campos de investigación como la psicología, la sociología, las políticas públicas y ciencias relacionadas a la familia (Tarka 2018), además, trabajos como el de Golob (2003) muestran la aplicación de los SEM en fenómenos económicos, o bien en investigación de mercados como sugieren los trabajos de Bagozzi (1980) y Chin, Peterson & Brown (2008). Según Beran & 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, además de las relaciones latentes y de causalidad entre ellos 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, las cuales se utilizan para construir los llamados 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), esta condición puede afectar negativamente la estimación del modelo y sus estadísticos de bondad de ajuste, llevando a pérdidas en la potencia (Foss, Jreskog & Olsson 2011) o al caso de descartar modelos que podrían ser adecuados solo por presentar un mal ajuste (Andreassen, Lorentzen & Olsson 2006). 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, que al día de hoy se mantiene como el método de estimación más extendido y popular, 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 a rechazar modelos que en realidad resumen bien la realidad para dar una mejor

explicación del por qué sucede un fenómeno, y además a la subestimación de los errores asociados a los parámetros, lo cual genera interpretaciones inadecuadas en lo referente a la significancia estadística de las relaciones planteadas por el modelo teórico.

Por otro lado, es posible toparse con conjuntos de datos que, en su conjunto, no presenten una distribución normal multivariada debido a la muy alta o muy baja concentración de datos alrededor de la zona central de su distribución. Este comportamiento se mide mediante un estadístico llamado kurtosis, que describe qué tan aplanada o empinada es la distribución, dependiendo de este estadístico, es posible saber si los datos atentan contra la presencia de una distribución normal. Trabajos como el de Sura-Fonseca (2020) o el de Andreassen et al. (2006) han abierto espacios de investigación para esta temática

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, se ha tomado tomado como base un estudio de la Universidad de California (Gao, Mokhtarian & Johnston 2008), por ser uno de los trabajos más recientes en cuanto a planteamiento de tamaños de muestra y kurtosis para la simulación de datos multivariados. Se plantean los siguientes objetivos:

1.3.1 Objetivo general

Comparar mediante un estudio de simulación los efectos en las estimaciones de cargas factoriales y medidas de ajuste de modelos de ecuaciones estructurales estimados mediante máxima verosimilitud 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, 120, 200 y 300.

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 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

De esta manera, el presente estudio consiste en en simular datos no normales multivariados con diferentes tamaños de muestra y kurtosis para la estimación de SEM tomando como punto modelo de referencia el obtenido por Sura-Fonseca (2020) para las habilidades cuantitativas, el cual consiste en dos variables exógenas y una endógena. Se realizaron 2000 conjuntos de datos para cada escenario de simulación y se comparan las estimaciones de tanto de las cargas factoriales como de varios estadísticos de bondad de ajuste que serán descritos más adelante.

2. Methodology

2.1 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 & Maurelli (1983) para la simulación de datos no normales multivariados. Este método, comúnmente conocido como VM, se basa en el método propuesto por 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:

$$\begin{aligned} b^2 + 6bd + 2c^2 + 15d^2 - 1 &= 0 \\ 2c(b^2 + 24bd + 105d^2 + 2) - \gamma_1 &= 0 \\ 24\left(bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)\right) - \gamma_2 &= 0 \end{aligned}$$

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:

$$\begin{aligned} r_{Y_1,Y_2} =& \rho_{X_1,X_2}(b_1 b_2 + 3b_1 d_2 + 3d_1 b_2 + 9d_1 d_2) \\ &+ \rho^2_{X_1,X_2}(2c_1 c_2) + \rho^3_{X_1,X_2}(6d_1 d_2) \end{aligned}$$

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.

2.2 Modelo a estimar

El modelo teórico utilizada para realizar las simulaciones es el presentado por Sura-Fonseca (2020), basado en datos de 155 estudiantes de la Universidad de Costa Rica, obtenidos de la Prueba de Habilidades Cuantitativas (PHC) del Instituto de Investigaciones Psicológicos (IIP) de dicha universidad y de un cuestionario autoadministrado aplicado a estos estudiantes. El modelo estimado está compuesto por dos variables exógenas (capital y habilidades cuantitativas) y una variable endógena (habilidades visoespaciales). Con respecto a estas variables: el capital se refiere al acceso y tenencia de ciertos bienes en los hogares de los estudiantes; las habilidades cuantitativas se refieren a la puntuación de los estudiantes en la prueba mencionada anteriormente; y las habilidades visoespaciales se refieren a la capacidad de los estudiantes para poder trabajar con objetos tridimensionales abstractos y poder manipularlos en la imaginación. Para cada una de estas variables latentes, se utilizó el método de parcelas para obtener tres variables indicadoras para cada uno de los constructos. Los resultados de la estimación de dicho modelo, presentados por Sura-Fonseca (2020), se presentan en la Figura 1.

2.3 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 el manejo de bases de datos y demás visualizaciones fueron utilizados los paquetes `ggplot2` (Wickham 2016), `tidyr` (Wickham & Henry 2020), `dplyr` (Wickham,

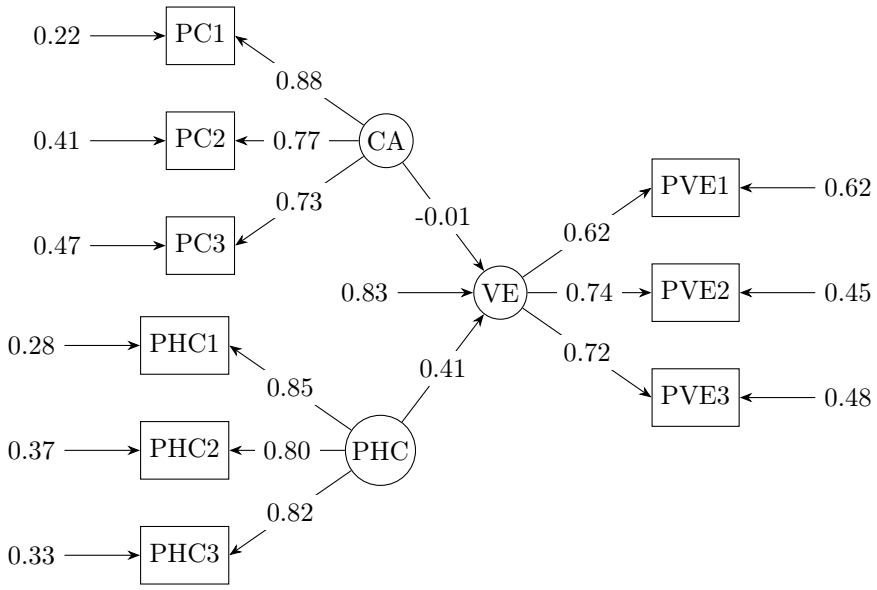

FIGURE 1: Modelo estimado sobre las habilidades cuantitativas

François, Henry & Müller 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 la sección anterior 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. 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 muestran en la Tabla 1:

TABLE 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 (Gao et al. 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. 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 sin los valores de las cargas factoriales, pues se busca conocer las estimaciones a partir de los datos generados.

2.4 Medidas de bondad de ajuste

Las medidas de bondad de ajuste utilizadas para comparar el ajuste de los modelos, para cada uno de los escenarios de simulación son: el estadístico chi-cuadrado, el RMSEA, el SRMR y el CFI.

2.4.1 Estadístico chi-cuadrado

El estadístico de chi-cuadrado busca cuantificar la diferencia que se presenta entre la matriz de covariancias de una muestra con la matriz de covariancias estimadas mediante un cierto modelo. Según Hu & Bentler (1999), su fórmula de cálculo viene dada por:

$$\chi^2 = (N-1)F_{min}$$

donde N es el tamaño de la muestra y F_{min} es el mínimo obtenido mediante la función de ajuste, la cual, normalmente, se asume que es la distribución normal multivariada, utilizando el método de máxima verosimilitud. Este estadístico tiene una distribución chi-cuadrado con grados de libertad igual a la cantidad de piezas de información única en la matriz de covariancias menos la cantidad de parámetros a estimar del modelo, bajo el supuesto de normalidad y, si este supuesto no se cumple, la distribución asintótica sigue siendo una chi-cuadrado con esos mismos grados de libertad. El estadístico chi-cuadrado es muy utilizado en los modelos de ecuaciones estructurales y da origen a la gran mayoría de las demás medidas de ajuste utilizadas en dichos modelos, aunque puede presentar algunos problemas ya que depende del tamaño de la muestra, por lo que, con muestras grandes tiende a ser significativo, mientras que con muestras pequeños tiende a no ser significativo (Kenny 2015).

2.4.2 RMSEA

El Error Cuadrático Medio de Aproximación (RMSEA por sus siglas en inglés) es una de las medidas de ajuste más conocidas y utilizadas en los modelos de ecuaciones estructurales. Su fórmula, según Hu & Bentler (1999), viene dada por:

$$RMSEA = \sqrt{\max\left\{\frac{\chi^2 - gl}{gl(N-1)}, 0\right\}}$$

donde χ^2 es el valor de la chi-cuadrado, gl son los grados de libertad, y N es el tamaño de la muestra. Por lo general, se considera un valor del RMSEA menor a 0.05 como un indicador de un buen ajuste, mientras que un valor mayor a 0.1 representa un mal ajuste del modelo (Kenny 2015).

2.4.3 SRMR

La Raíz Estandarizada del Error Cuadrático Medio (SRMR por sus siglas en inglés) es una medida de ajuste en la cual se comparan las diferencias entre las

covariancias estimadas y las de la muestra. La fórmula de cálculo, con base en Hu & Bentler (1999) es:

$$SRMR = \sqrt{\left(2\sum_{i=1}^{p}\sum_{j=1}^{i} \left((s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})\right)^2 \right)^2 /p(p+1)}$$

donde p es el número de variables observadas, s_{ij} son las covariancias observadas y $\hat{\sigma}_{ij}$ son las covariancias estimadas de las variables i y j. Dado que se están comparando las covariancias observadas y las estimadas, un valor de 0 indica un ajuste perfecto del modelo, pero, por lo general, se considera un valor menor a 0.08 como un indicador de un buen ajuste (Kenny 2015).

2.4.4 CFI

El Índice de Ajuste Comparativo (CFI por sus siglas en inglés) es una medida de ajuste que compara el valor de chi-cuadrado del modelo estimado con el valor de chi-cuadrado del modelo nulo, agregando una penalización por la cantidad de parámetros estimados. La fórmula de cálculo presentada por Hu & Bentler (1999) es la siguiente:

$$CFI = 1 - \left(\frac{\max\left\{(\chi^2_T - gl_T), 0\right\}}{\max\left\{(\chi^2_T - gl_T), (\chi^2_N - gl_N), 0\right\}} \right)$$

donde χ^2_T y χ^2_N son los valores del estadístico chi-cuadrado para el modelo estimado y el nulo, respectivamente, y gl_T y gl_N son los grados de libertad de los modelos estimado y nulo, respectivamente. Esta medida de ajuste puede tomar un valor entre 0 y 1 y se considera que el modelo tiene un buen ajuste cuando es mayor a 0.95, un buen ajuste cuando el valor está entre 0.9 y 0.95 y un mal ajuste cuando es menor que 0.9 (Kenny 2015).

3. Results

3.1 Introducción

4. Conclusions

4.1 Introducción

4.2 Conclusiones

4.3 Recomendaciones

5. Appendices

References

Andreassen, T. W., Lorentzen, B. G. & Olsson, U. H. (2006), 'The impact of non-normality and estimation methods in SEM on satisfaction research in marketing', *Quality and Quantity* **40**(1), 39–58.
*https://link.springer.com/article/10.1007/s11135-005-4510-y

Bagozzi, R. P. (1980), *Causal models in marketing*, Wiley, New York, NY.

Beran, T. N. & Violato, C. (2010), 'Structural equation modeling in medical research: A primer', *BMC Research Notes* **3**(1), 267.
*https://doi.org/10.1186/1756-0500-3-267

Chin, W. W., Peterson, R. A. & Brown, S. P. (2008), 'Structural equation modeling in marketing: Some practical reminders', *Journal of Marketing Theory and Practice* **16**(4), 287–298.

Fleishman, A. I. (1978), 'A method for simulating non-normal distributions', *Psychometrika* **43**(4), 521–532.
*https://link.springer.com/article/10.1007/BF02293811

Foss, T., Jreskog, K. G. & Olsson, U. H. (2011), 'Testing structural equation models: The effect of kurtosis', *Computational Statistics and Data Analysis* **55**(7), 2263–2275.

Gao, S., Mokhtarian, P. L. & Johnston, R. A. (2008), 'Nonnormality of data in structural equation models', *Transportation Research Record* (2082), 116–124.

Golob, T. F. (2003), 'Structural equation modeling for travel behavior research'.

Hu, L. T. & Bentler, P. M. (1999), 'Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives', *Structural Equation Modeling* **6**(1), 1–55.
*https://www.tandfonline.com/doi/abs/10.1080/10705519909540118

Kaplan, D. (2012), *Structural Equation Modeling (2nd ed.): Foundations and Extensions*, SAGE Publications, Inc.

Kassambara, A. (2020), *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.3.0.
*https://CRAN.R-project.org/package=ggpubr

Kenny, D. A. (2015), 'SEM: Fit (David A. Kenny)'.
*http://www.davidakenny.net/cm/fit.htm

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.
*https://onlinelibrary.wiley.com/doi/10.1111/j.2044-8317.1992.tb00975.x

Peterson, B. G. & Carl, P. (2020), *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*. R package version 2.0.4.
*https://CRAN.R-project.org/package=PerformanceAnalytics

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Rosseel, Y. (2012), 'lavaan: An R package for structural equation modeling', *Journal of Statistical Software* **48**(2), 1–36.
*http://www.jstatsoft.org/v48/i02/

RStudio Team (2015), *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.
*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, Master's thesis, University of Costa Rica.
*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 and Quantity* **52**(1), 313–354.
*https://doi.org/10.1007/s11135-017-0469-8

Vale, C. D. & Maurelli, V. A. (1983), 'Simulating multivariate nonnormal distributions', *Psychometrika* **48**(3), 465–471.
*https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/article/10.1007/BF02293687

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
*https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L. & Müller, K. (2020), *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5.
*https://CRAN.R-project.org/package=dplyr

Wickham, H. & Henry, L. (2020), *tidyr: Tidy Messy Data*. R package version 1.1.0.
*https://CRAN.R-project.org/package=tidyr

Zhu, H. (2019), *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0.
*https://CRAN.R-project.org/package=kableExtra