Grado en Estadística

TRABAJO FIN DE GRADO

Introducción a los Modelos de Ecuaciones Estructurales

Enrique Pérez Jiménez

Sevilla, Junio de 2023

Índice general

Prólogo

Cuando empecé a plantearme posibles temas sobre los que realizar mi TFG, tenía claro que debía ser alguno que permitiera una gran aplicación práctica con la que analizar y buscar solución a problemas reales. Por este motivo, cuando uno de mis tutores me sugirió como posible tema la modelización mediante ecuaciones estructurales y comencé a investigar un poco sobre ello, supe que podría ser un tema adecuado e interesante, además de tener una amplia utilización en diversos campos y, en particular, en ciencias sociales como la psicología, un área que siempre me ha resultado interesante.

Para finalizar este escrito me gustaría agradecer y dedicar este trabajo a mi familia por su paciencia y confianza, así como a todos los compañeros que de igual forma me han brindado su apoyo y ayuda en estos meses de trabajo. Por supuesto agradecer a mis tutores, D. José Luis Pino Mejías y D. Joaquín Antonio García de las Heras, por su predisposición, ayuda y orientación en la realización de este trabajo. Finalmente, agradecer también al profesor D. Pedro Luis Luque Calvo, por la realización de la plantilla basada en *R Markdown* con la que se ha construido este documento, y que ha sido de gran utilidad al permitir ahorrar tiempo y esfuerzo en infinidad de aspectos técnicos y estéticos.

Resumen

Los modelos de ecuaciones estructurales (en adelante SEM[1]) son una técnica estadística cuantitativa multivariante que tiene como objetivo principal describir la relación existente entre una serie de variables observables y otras no observables denominadas variables latentes. Esta técnica es muy utilizada en diversos campos, principalmente en ciencias sociales como la psicología.

En este trabajo, se tratarán de exponer los fundamentos que caracterizan a los SEM y se mostrarán algunas aplicaciones mediante el uso del software estadístico R. Para ello, el trabajo se ha dividido en cuatro bloques o capítulos. En el primer capítulo se describirá de manera breve en qué consisten este tipo de modelos, además de explicar el contexto en el que surgen e introducir algunos conceptos importantes para su formulación. También se describen de forma breve los tipos de SEM que existen.

En el segundo capítulo, se detallarán los pasos necesarios para la construcción de un modelo SEM, partiendo de la comprobación de las hipótesis previas que deben darse para ello, y finalizando con la evaluación del modelo y la obtención de conclusiones.

El tercer capítulo está dedicado a la aplicación práctica de los SEM mediante el uso del software R sobre dos conjuntos de datos. El primero de ellos trata sobre la expansión de la democracia y la industrialización en países en vías de desarrollo. En el segundo caso práctico se muestra la aplicación de los modelos SEM sobre un conjunto de datos cuyas observaciones están divididas en grupos o *clusters*.

Finalmente, el cuarto capítulo contiene las conclusiones, donde se hace balance de los principales resultados, tanto teóricos como prácticos, que se han obtenido a raíz de la realización de este trabajo.

[1]De sus siglas en inglés: *Structural Equation Models*

Abstract

Structural Equation Modeling (SEM) is a multivariate quantitative statistical technique whose main objective is to describe causal relationships between observable and unobservable variables, which are called latent variables. This technique is widely used in many areas, mainly in social sciences such as psychology.

In this work, we will try to explain the fundamentals that characterize the SEM and some applications will be shown by using the R statistical software. For this purpose, the work has been divided into four blocks or chapters. The first chapter will briefly describe what this type of modeling consists of, as well as explaining the context in which they arise and introducing some important concepts for their formulation. The types of SEM that we can found will also be briefly described.

In the second chapter, the necessary steps for the construction of a SEM model will be detailed, starting with the verification of the previous hypotheses that must be given for this purpose, and ending with the evaluation of the model and the drawing of conclusions.

The third chapter is focused on the practical application of SEM using R software on two datasets. The first one is about the expansion of democracy and industrialization in developing countries. The second practical case shows the application of SEM models on a dataset whose observations are divided into clusters.

Finally, the fourth chapter is devoted to conclusions, which take stock of the main results, both theoretical and practical, that have been obtained as a result of this work.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Motivación y origen de los SEM

Los modelos de ecuaciones estructurales (SEM) son una técnica estadística multivariante conocida también como análisis de estructura de covarianzas. Esta técnica busca estudiar la relación entre variables observadas y variables latentes (también llamadas factores o constructos). La cualidad que diferencia a los SEM del análisis factorial[1] clásico, es el hecho de que las variables latentes puedan a su vez depender de otras variables latentes. Por este motivo esta técnica puede considerarse como una extensión del mencionado análisis factorial.

La hipótesis principal sobre la que se basa la construcción de los SEM es que la matriz de varianzas y covarianzas poblacional es igual a la matriz de varianzas y covarianzas del modelo teórico. En otras palabras, si el modelo es correcto, se puede reproducir la matriz de varianzas y covarianzas poblacional a partir de la combinación de los parámetros del modelo:

$$H_0 : \Sigma = \Sigma(\boldsymbol{\theta}), \tag{1.1}$$

siendo Σ la matriz de varianzas y covarianzas poblacional, $\boldsymbol{\theta}$ un vector formado por los parámetros del modelo y $\Sigma(\boldsymbol{\theta})$ la matriz asociada al modelo teórico (obtenida como función de los parámetros del modelo).

Esta expresión general es muy usada en múltiples técnicas estadísticas. Los modelos de regresión, análisis factorial confirmatorio, análisis de correlación canónica, modelos ANOVA, etc., son casos especiales de (1.1). Para ilustrar esto podemos considerar la expresión de un modelo de regresión simple dadas dos variables aleatorias x e y:

$$y = \gamma x + \zeta, \tag{1.2}$$

donde γ es el coeficiente de regresión y ζ es una variable que mide posibles perturbaciones (error aleatorio) y de la que se supone $\mathbb{E}(\zeta) = 0$. Podemos expresar esta ecuación en términos de (1.1) de la siguiente forma:

[1] El análisis factorial tradicional tiene como objetivo esencial describir, si es posible, la estructura de covarianzas entre diversas variables en términos de un número pequeño de variables latentes o factores.

$$\begin{pmatrix} Var(y) & \\ Cov(x,y) & Var(x) \end{pmatrix} = \begin{pmatrix} \gamma^2\, Var(x) + Var(\zeta) & \\ \gamma\, Var(x) & Var(x) \end{pmatrix} \tag{1.3}$$

donde $Var()$ y $Cov()$, denotan la varianza y covarianza poblacional respectivamente de los elementos incluidos entre paréntesis. Así, expresando (1.3) en términos de (1.1), el primer miembro corresponde a Σ, mientras que el segundo miembro corresponde a $\Sigma(\boldsymbol{\theta})$, de modo que $\boldsymbol{\theta}$ es un vector que contiene a ζ, $Var(x)$ y $Var(\zeta)$ como parámetros.

Si en lugar de la expresión de un modelo de regresión simple, consideramos dos variables alteatorias x_1 y x_2, ambas indicadores de una variable latente o factor denotado por ξ, obtenemos la siguiente expresión de dependencia de las variables x_i con el factor ξ:

$$x_i = \xi + \delta_i \quad i = 1, 2 \tag{1.4}$$

con δ_i incorreladas $\forall\, i$ e incorreladas con ξ. Supondremos de nuevo $\mathbb{E}(\delta_1) = \mathbb{E}(\delta_2) = 0$. Así, podemos expresar (1.4) como caso particular de (1.1) de la siguiente forma:

$$\begin{pmatrix} Var(x_1) & \\ Cov(x_1, x_2) & Var(x_2) \end{pmatrix} = \begin{pmatrix} \phi + Var(\delta_1) & \\ \phi & \phi + Var(\delta_2) \end{pmatrix}$$

donde ϕ denota la varianza de la variable latente ξ. En este caso, el vector $\boldsymbol{\theta}$ está formado por tres elementos: ϕ, $Var(\delta_1)$ y $Var(\delta_2)$.

1.1.1. Origen de los SEM

No se puede dar una respuesta única a la pregunta de quién ideó o formalizó los modelos de ecuaciones estructurales, ya que son muchos los investigadores que contribuyeron a su desarrollo inicial, de igual forma que fueron muchos los que posteriormente siguieron desarrollando avances en el uso de este tipo de modelos, haciéndolos más generales y flexibles y aumentando por tanto su utilidad. Aún así, se tratará de dar una explicación del contexto en el que surgen los SEM, así como los principales pasos que se han dado hasta llegar a los modelos actualmente disponibles.

Para empezar a hablar de una primera aproximación a lo que hoy entendemos por modelos de ecuaciones estructurales, hay que remontarse a la década de 1920, cuando el genetista Sewall Wright (Merlose, EEUU, 1889 - Madison, EEUU, 1988) trató de emplear sistemas de ecuaciones simultáneas[2] para estudiar la influencia genética generacional entre compañeros de camada. Para ello, consideró una situación en la que se conocen los genes de los padres (causa) y sus efectos en los rasgos de los descendientes, de modo que la causalidad iba en una sola dirección. Esta situación se conoce como modelo de flujo causal no resursivo y entra dentro de la primera técnica que se desarrolló en el ámbito de los SEM, el conocido como *path analysis* (o análisis de ruta), que permitía representaciones gráficas en forma de grafo llamadas *path diagrams*. Esta técnica será explicada con mayor detalle más adelante. No obstante, en la Figura 1.1 se muestra un ejemplo de estos diagramas:

[2]Los sistemas de ecuaciones simultáneas están formados por dos o más ecuaciones que comparten las mismas variables y deben su nombre a que son resueltas al mismo tiempo.

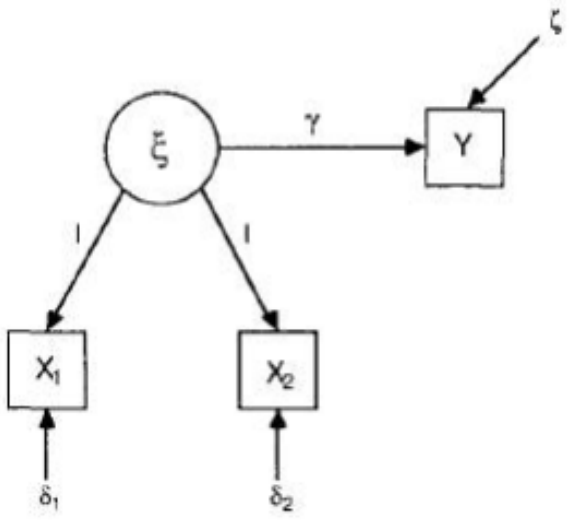

Figura 1.1: Ejemplo de *Path Diagram*.
Fuente: *Structural Equations with Latent Variables*. Bollen KA, 1989

No será hasta las década de los sesenta cuando investigadores como Blalock (1964) o Duncan (1966), trabajen sobre las ideas introducidas por Wright con el objetivo de adaptarlas al ámbito socioeconómico. No obstante, es cierto que anteriormente (1936), John Maynard Keynes, desarrolló un modelo en el ámbito de la economía que, si bien hacía uso de sistemas de ecuaciones lineales y podía considerarse como una extensión del análisis de regresión clásico, resultaba difícil encontrar técnicas útiles para estimar los parámetros del modelo. Por este motivo, algunos economistas y sociólogos consideran este modelo introducido por Keynes, como una extensión del *path analysis* formulado por Wright.

Posteriormente, otros autores siguen desarrollando modelos que, bien centrados en describir la correlación entre variables observables (Bock y Bargmann, 1966), o bien ampliando el análisis factorial clásico con el llamado análisis factorial confirmatorio general (Jöreskog, 1969), no eran capaces de relacionar variables observables y latentes para establecer un modelo más general. No fue hasta 1973 cuando el propio Karl Jöreskog, en colaboración con Sörbom desarrolla un modelo de ecuaciones lineales simultáneas con variables latentes. Este modelo además supuso un gran avance en la rama de la estadística computacional, ya que Jöreskog y Sörbom implementaron un programa basado en su modelo llamado LISREL[3], que permitió a futuros investigadores establecer hipótesis causales y comprobarlas mediante pruebas de bondad de ajuste chi-cuadrado.

Sin embargo, este programa presentaba algunos inconvenientes como la necesidad de separar las variables en función de si dependían de variables latendes exógenas o endógenas[4]. Esto motivó a Peter Bentler, en colaboración con Weeks, a introducir un nuevo modelo, para posteriormente implementar en 1985 su propio programa llamado EQS, que permitió llevar a cabo el estudio de los SEM sin la obligatoriedad de realizar la distinción anterior.

[3]De sus siglas en inglés: *Linear Structural Relations*

[4]Se definen las variables latentes endógenas como aquellas cuyo valor viene determinado por las relaciones entre variables exitentes dentro del propio modelo. Por el contrario, las variables exógenas son aquellas que vienen determinadas por agentes externos ajenos al modelo.

Para concluir este breve repaso histórico podemos decir que durante la década de los noventa hasta principios de siglo, son muchos los investigadores que han seguido estudiando el campo de los SEM, logrando avances significativos, tanto desde el punto de vista estadístico como computacional.

1.1.2. Causalidad y relación causal

Como se ha mencionado anteriormente, los SEM estudian relaciones causales entre variables observadas y variables latentes. Sin embargo, no resulta fácil dar una definición clara del concepto causalidad. Aún así, sí que podemos definir con cierta precisión el concepto de *relación causal* entre variables. Debemos tener en cuenta que este concepto y el de correlación no tienen una relación bicondicional, es decir, la correlación entre dos variables no implica necesariamente la existencia de una relación causal entre ambas. Partiendo de esta base, podemos dar la siguiente definición:

Definición 1.1: *Se dirá que existe una relación causal entre dos o más variables cuando además de la existencia de correlación, todo cambio en una variable suponga un efecto en la otra.*

Si tomamos como partida dos variables aleatorias x_1 e x_2, podemos representar el efecto causal de x_1 sobre x_2 mediante la siguiente expresión:

$$x_2 = \phi_{12} x_1 + \zeta, \tag{1.5}$$

siendo ϕ_{12}[5] la influencia esperada de x_1 sobre x_2 y ζ una perturbación de origen desconocido y de la que se supone $\mathbb{E}(\zeta) = 0$

1.1.3. Tipos de SEM

Llegados a este punto, y para cerrar este capítulo de introducción, vamos a resumir a continuación los tipos de SEM que podemos encontrar y que, fundamentalmente, se diferencian en la tipología de las variables que intervienen en ellos, las cuales se definirán en capítulos posteriores.

- ***Path analysis***, o modelos de variables observadas. Como su propio nombre indica, son aquellos formados únicamente por variables observadas, es decir, no cuentan con variables latentes. Estos modelos suponen una generalización de los modelos de regresión múltiple y por tanto consideran que existe un único flujo de causalidad unidireccional. Además, suponen máxima fiabilidad en la medición de las variables y por tanto asumen la no existencia de errores de medición. Cuentan con una representación gráfica llamada *path diagrams*, mencionada anteriormente y ejemplificada en la Figura 1.1

- **Análisis factorial**. Busca explicar la estructura de covarianzas entre distintas variables en términos del menor número de variables posibles. Podemos distinguir dos tipos de modelos de análisis factorial: análisis factorial *exploratorio*, cuyo objetivo

[5]Correspondiendo el primer subíndice a la variable causa y el segundo a la variable efecto.

es hallar precisamente esa estructura de covarianzas entre variables, y análisis factorial *confirmatorio*, que como su nombre indica, busca confirmar una estructura de covarianzas hallada previamente.

- **Modelo estructural general**. Constituye una generalización de los dos tipos anteriores, estudiando relaciones causales entre variables observadas y latentes, además de las relaciones entre las propias variables latentes.

1.2. Fundamentos de los SEM

Antes de detallar los procedimientos propios de la construcción y análisis de datos mediante SEM, es conveniente definir algunos conceptos clave a la hora de estudiar estos modelos, así como fijar y describir de antemano cuál será la notación utilizada para referirse a los distintos componentes del modelo, a fin de evitar confusiones posteriormente. En este caso, la notación que se empleará tanto en este capítulo como en los posteriores, será la introducida por Jöreskog, Wiley y Keesling en los años setenta, y que posteriormente se popularizó a raíz de su uso en el programa LISREL, mencionado en el capítulo anterior. Por este motivo, esta notación es conocida también como notación LISREL.

1.2.1. Tipos de variables en los SEM

Las variables presentes en los modelos de ecuaciones estructurales, pueden clasificarse en distintos tipos según su propia naturaleza, o según sus "interrelaciones" con otras variables, esto es, según si afectan y/o reciben efecto de otras variables. De este modo, podemos definir los siguientes tipos:

Definición 1.2 Variable observada: *Variables medidas en los individuos sobre un determinado fenómeno.*

Definición 1.3 Variable latente: *Características o conceptos unidimensionales que serían interesantes de medir, pero que no son observables, y que por tanto no están sujetas a error de medición. También son llamadas factores o constructos.*

Definición 1.4 Variable error: *Errores asociados a la medición de las variables, así como a la ausencia de otras en el modelo que pudieran suponer una perturbación en la medición de variables observadas. Al no ser medidas directamente, se consideran a su vez variables latentes.*

Por otro lado, si atendemos a las relaciones entre variables, podemos distinguir los siguientes tipos:

Definición 1.5 Variable endógena: *Variable que recibe efecto de otra variable del modelo. Un ejemplo sería la variable dependiente de un modelo de regresión. Van acompañadas del correspondiente error.*

Definición 1.6 Variable exógena: *Variable que afecta a otras pero que no recibe efecto de ninguna otra variable del modelo. Un ejemplo son las variables explicativas de un modelo de regresión*

Lógicamente estas dos clasificaciones no son excluyentes, sino que las variables se identifican dentro de ambas. Es decir, una variable latente, por ejemplo, será a su vez endógena

o exógena, del mismo modo que cualquier variable endógena o exógena, estará al mismo tiempo dentro de alguno de los tipos descritos en la primera distinción.

Estos distintos tipos de variables se denotarán de distinta forma, tanto al construir los modelos de forma analítica, como al representarlos gráficamente. Por este motivo, a continuación, vamos a hacer una síntesis de la notación[6] que se empleará en adelante en cada caso:

- Las variables observables endógenas se denotarán por y, mientras que las exógenas se denotarán por x.
- Las variables latentes endógenas se representarán como η, mientras que las exógenas se representarán como ξ.
- Los coeficientes γ representan el efecto directo de las variables latentes exógenas (ξ) sobre las variables latentes endógenas (η). Estos coeficientes están agrupados en una matriz denotada por Γ.
- Los coeficientes β representan el efecto de las variables latentes endógenas (η) sobre las propias variables latentes endógenas. Estos coeficientes se agrupan en una matriz denotada por B.
- Los coeficientes ϕ representan las correlaciones entre variables latentes exógenas (ξ). Estos coeficientes se agrupan en una matriz denotada por Φ.
- Los coeficientes λ representan la relación de cada variable latente con sus indicadores (observables). Estos coeficientes se agrupan en dos matrices, Λ_X o Λ_Y, según el tipo de variable observable que corresponda.
- En el caso de los errores, para las variables observables, se utilizarán los coeficientes ε y δ para referirse a los errores de medida asociados a variables endógenas y exógenas, respectivamente. Estos coeficientes se agrupan en las matrices Θ_ε y Θ_δ, respectivamente.
- Para las variables latentes, se emplearán los coeficientes ζ para denotar los errores asociados a variables endógenas. Estos coeficientes se agrupan en la matriz Ψ. Como se comentó anteriormente, las variables latentes exógenas no están sujetas a error de medición.

Observación 1.1: Los distintos coeficientes aparecerán generalmente acompañados de ciertos subíndices que determinarán las variables que relacionan. Por ejemplo, γ_{ij}, representará el efecto directo de la variable ξ_j sobre η_i, del mismo modo que ϕ_{ij} representará la correlación entre las variables ξ_i y ξ_j, etc.

1.2.2. Representación gráfica de los SEM: Diagramas de caminos

Los diagramas de caminos o *path diagrams*, son una forma de representar los SEM, tanto sus variables, como las relaciones existentes entre ellas. Estos diagramas, también

[6]Siguiendo la conocida como notación LISREL, como se mencionó anteriormente

llamados diagramas de flujo, diagramas causales, gráficos de rutas, etc., pueden ser de gran ayuda, ya que a partir de ellos se pueden inferir las ecuaciones que forman el modelo. Para ello, se deben seguir una serie de convenciones fijas a la hora de representarlos:

- Las variables observables se representan dentro de rectángulos.
- Las variables latentes se representan dentro de óvalos o círculos.
- Los errores, ya sean de medición o de estimación, no se representan dentro de rectángulos ni de círculos, si bien es cierto que según el programa que se utilice pueden aparecer como variables latentes.
- Las relaciones bidireccionales, tanto correlaciones como covarianzas, se representan como vectores curvos con una flecha en cada extremo.
- Los efectos causales se representan como una flecha recta, con origen en la variable predictora (causa) y final en la variable dependiente (efecto).
- Los distintos parámetros del modelo se ubicarán encima de la flecha correspondiente.

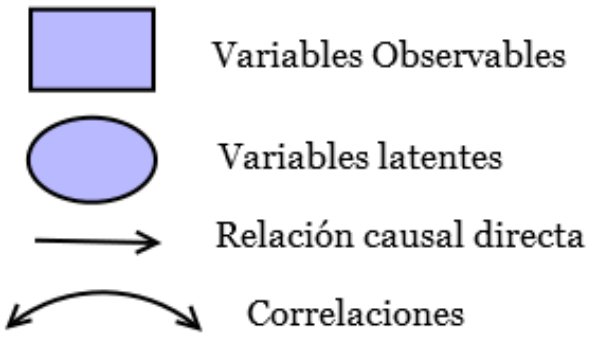

Figura 1.2: Notación básica: diagramas de caminos. Fuente: Elaboración propia

En la Figura 1.2, se puede ver un resumen de la nomenclatura empleada en los *path diagrams*

Una vez tratadas estas cuestiones de notación, es conveniente aclarar que, según las relaciones entre variables que estén presentes en el modelo estudiado, se pueden diferenciar dos subtipos: modelos *no recursivos*, que son aquellos en los que las relaciones causales son siempre directas, y los errores no están relacionados entre sí, y los *recursivos*, en los que pueden aparecer relaciones circulares y las perturbaciones pueden estar correlacionadas.

1.2.3. Otras relaciones entre variables

Hasta ahora, hemos definido las relaciones causales directas entre variables y las relaciones bidireccionales, pero existen otros tipos de relación que no se han mencionado y que son importantes a la hora de representar un SEM, y obtener posteriormente su desarrollo de forma analítica:

- **Relación espuria:** Involucra a tres variables. Se produce cuando, la covariación entre dos variables A y B, es debida en su totalidad o de forma parcial a la relación de ambas variables con una tercera C. Para ilustrar este concepto se propone el siguiente ejemplo: si medimos dos variables como inteligencia y estatura en preescolares, es probable que obtengamos una alta correlación. Sin embargo, hay una tercera variable (edad, desarrollo del niño) que es causa de las dos primeras y que produce esa relación. En este caso, la relación entre inteligencia y estatura sería una relación espuria.

- **Relación causal indirecta:** Se dice que existe una relación causal indirecta entre dos variables A y B cuando existe una tercera variable C que modula o mediatiza el efecto entre ambas, de forma que el efecto entre la primera y la segunda pasa a través de la tercera. Podemos representarlo de la forma: $A \rightarrow C \rightarrow B$. Un ejemplo podría ser la relación entre aptitud y rendimiento en el ámbito del deporte, con una tercera variable que fuera motivación. Se diría que existe una relación causal indirecta entre aptitud y rendimiento y se representaría:

$$\text{Aptitud} \rightarrow \text{Motivación} \rightarrow \text{Rendimiento}$$

- **Efectos totales o conjuntos:** Se produce con la combinación de las dos relaciones anteriores, siendo A y C variables exógenas. Al no especificar el tipo de relación existente entre estas variables, no podemos saber si C influye sobre la relación de A y B de forma espuria o de forma indirecta.

Capítulo 2

Construcción y evaluación de los SEM

2.1. Hipótesis previas

2.1.1. Normalidad

Como se verá en apartados posteriores, para poder llevar a cabo la estimación de los parámetros del modelo se emplean distintos métodos, siendo uno de los más frecuentes el método de la máxima verosimilitud. Por ello, para poder utilizar este método, es necesario que se satisfagan algunas condiciones o hipótesis que se deberán comprobar en los datos que se traten. Una de ellas es la distribución normal multivariante de las variables observadas del modelo:

Definición 2.1: *Se dice que una variable aleatoria k-dimensional* $\mathbf{X} = (X_1, \ldots, X_k)^t$, *sigue una distribución normal multivariante si su función de densidad conjunta viene dada por la expresión:*

$$f_X(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left\{-\frac{1}{2}\ (\mathbf{x} - \boldsymbol{\mu})^t\ \Sigma^{-1}\ (\mathbf{x} - \boldsymbol{\mu})\right\}, \text{con } x_i \in \mathbb{R}\ \forall i = 1, \ldots, k \tag{2.1}$$

y donde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^t$ con $\mu_i \in \mathbb{R}\ \forall i = 1, \ldots, k$ *y* Σ *es una matriz* $k \times k$ *simétrica y definida positiva.*

Observación 2.1: Se puede demostrar que si una variable aleatoria k-dimensional $\mathbf{X} = (X_1, \ldots, X_k)^t$, sigue una distribución normal multivariante, cada uno de los $X_i,\ i = 1, \ldots, k$ sigue una distribución normal univariante. El recíproco no es cierto.

En la práctica, para comprobar la normalidad de las variables observadas, disponemos de distintos tests como el de *Kolmogorov-Smirnov* o el de *Shapiro-Wilk*. Adicionalmente, resultará de utilidad realizar contrastes de asimetría y curtosis, como los propuestos por *Mardia* y que permitirán comprobar la normalidad multivariante de los datos.

2.1.2. Tamaño de la muestra

La cuestión de cuál es el tamaño que debe tener la muestra con la que se trabaje para que las estimaciones obtenidas por el modelo y por tanto las conclusiones a las que se llegue sean lo suficientemente fiables es a menudo fuente de conflicto, no sólo en el campo de los SEM, sino en toda la estadística aplicada en general. Con frecuencia no se dispone de un único criterio acerca del número de casos idóneo a considerar y, según la fuente consultada, se pueden ver distintas sugerencias al respecto. En el caso de los SEM ocurre lo mismo, aunque no hay un único criterio referente al tamaño de la muestra con la que se trabaje, sí vamos hacer un breve resumen de las propuestas que se pueden encontrar en la literatura, dividiéndolas en tres grupos o apartados:

- **Cantidad total de datos:** En la mayoría de publicaciones se sugiere un tamaño mínimo de $N = 200$ casos (Catena et al., 2003; Hair et al., 2014; Stevens, 2009). Sin embargo, otros autores consideran que incluso ese tamaño puede resultar insuficiente si el modelo es muy complejo, no existe normalidad multivariante, o se utilizan ciertos tipos de estimación (Kline, 2011). Por otra parte, en algunos casos se añade que considerar tamaños muestrales muy superiores a 200, podría tener el inconveniente de generar modelos excesivamente sensibles. (Hair et al., 2014).
- **Casos por parámetro:** En este caso, la recomendación más extendida es trabajar con muestras con al menos 100 casos para modelos con cinco constructos o menos, cada uno de ellos con al menos tres indicadores, y comunalidades superiores a 0.60 (Hair et al., 2014). Kline (2011) establece una forma de clasificar la "bondad" de la muestra en función del número de casos por parámetro (ver Tabla 2.1). Finalmente, otros autores aseguran que un tamaño de muestra de cinco casos por parámetro sería suficiente en un Análisis factorial confirmatorio (Worthington y Whittaker, 2006).

Tabla 2.1: Tamaño de la muestra necesario según el número de casos por parámetro propuesto por Kline (2011)

Casos por parámetro	Bondad de la muestra
20	Muestra ideal
10	Muestra menos ideal
<10	Muestra inapropiada

- **Casos por variable observada:** Catena et al. (2003), propusieron sumar el número de variables observadas y el número de variables latentes, de forma que un tamaño adecuado sería de ocho casos por el total de esta suma. Otra propuesta (Hair et al., 2014) es considerar un tamaño de 15 casos por indicador, lo que permitiría minimizar problemas relacionados con la ausencia de normalidad multivariante.

Para finalizar este apartado, es interesante destacar el estudio realizado por Marsh et al. (1988, 1996, 1998) mediante 35000 simulaciones realizadas mediante el método de Monte Carlo haciendo uso del programa LISREL, en el que pudieron establecer una regla para calcular el tamaño adecuado de una muestra, mediante la siguiente expresión:

$$N \geq 50r^2 - 450r + 1100$$

siendo r el número de indicadores por variable latente.

2.2. Construcción de los SEM

En este apartado vamos a llevar a cabo el desarrollo teórico de los distintos pasos que se han de seguir para trabajar con los SEM y que permitirán no sólo su construcción sino también realizar estimaciones de los distintos parámetros para posteriormente evaluar dichas estimaciones y obtener conclusiones en la práctica.

El número de pasos a seguir varía en función de la fuente consultada, pero en este caso se van a considerar los propuestos por *Kaplan* (2000) y *Kline* (2005), esto es:

1. Especificación del modelo
2. Identificación del modelo
3. Estimación de parámetros
4. Evaluación del ajuste
5. Reespecificación del modelo
6. Obtención de resultados e interpretación de los mismos.

Antes de comenzar el desarrollo teórico de cada uno de los apartados mencionados, es conveniente aclarar que éste se hará sobre el modelo más general de los descritos al final de primer capítulo. Este modelo general supone una síntesis de los otros dos tipos de modelos estructurales mencionados: el de variables observadas o *path analysis*, y el surgido a partir del análisis factorial.

2.2.1. Especificación del modelo

En primer lugar, vamos a partir de lo que se conoce como modelo de variables latentes o **modelo estructural**, y que servirá para identificar las relaciones entre variables latentes exógenas y endógenas. Se define de la siguiente manera:

Definición 2.2: *Un modelo estructual sigue la siguiente expresión en forma matricial:*

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2.2}$$

donde $\boldsymbol{\eta}$ es un vector $m \times 1$ de variables latentes endógenas, $\boldsymbol{\xi}$ es un vector $n \times 1$ de variables latentes exógenas, B es una matriz $m \times m$ que contiene los coeficientes que determinan la influencia de las variables latentes endógenas entre sí, y Γ es una matriz $m \times n$ que contiene los coeficientes correspondientes a los efectos de $\boldsymbol{\xi}$ sobre $\boldsymbol{\eta}$. $\boldsymbol{\zeta}$ es un vector $m \times 1$ que contiene los errores o perturbaciones.

Además, en los modelos estructurales dados por (2.2) se satisface que la matriz $(I - B)$ es no singular y los componentes del modelo cumplen las siguientes propiedades:

1. $\mathbb{E}(\boldsymbol{\eta}) = \mathbf{0}_{m\times 1}$
2. $\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}_{n\times 1}$
3. $\mathbb{E}(\boldsymbol{\zeta}) = \mathbf{0}_{m\times 1}$
4. $\boldsymbol{\xi}$ y $\boldsymbol{\zeta}$ están incorrelados.

Por este motivo para simplificar, a menudo los coeficientes η_i y ξ_j se escriben como desviaciones de su media. Esto supone que si consideramos η_i^* y ξ_j^* como las variables originales, se satisface que $\eta_i = \eta_i^* - \mathbb{E}(\eta_i^*)$ y $\xi_j = \xi_j^* - \mathbb{E}(\xi_j^*)$ $\forall\, i = 1, \ldots, m\,; \quad j = 1, \ldots, n.$

Este modelo se puede expresar también de la siguiente forma que recibe el nombre de **forma reducida**:

$$\boldsymbol{\eta} = (I - B)^{-1} + (\Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}), \tag{2.3}$$

siempre que se cumpla $|I - B| \neq 0$.

Finalmente, se asume que las perturbaciones ζ_i para una misma ecuación son homocedásticas ($Var(\zeta_i) = Var(\zeta_j)$) y no autocorrelacionadas ($Cov(\zeta_i, \zeta_j) = 0$) $\forall\, i, j = 1, \ldots, n$.

Una vez construido el modelo estructural, para continuar con la especificación de un modelo SEM generalizado, se debe definir lo que se conoce como *measurement model*, o modelo de medida, y que servirá para determinar las relaciones entre variables observadas y latentes. Estos modelos tendrán tantas ecuaciones como variables observadas haya:

Definición 2.3: *Se define el modelo de medida como aquel que viene dado por la expresión matricial:*

$$\begin{aligned} \boldsymbol{y} &= \Lambda_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ \boldsymbol{x} &= \Lambda_x\boldsymbol{\xi} + \boldsymbol{\delta} \end{aligned} \tag{2.4}$$

donde $\boldsymbol{x}$ e $\boldsymbol{y}$ son vectores de variables observadas $q \times 1$ y $p \times 1$ respectivamente, Λ_y y Λ_x son dos matrices $p \times m$ y $q \times n$, cuyos coeficientes determinan la relación de $\boldsymbol{y}$ con $\boldsymbol{\eta}$ y de $\boldsymbol{x}$ con $\boldsymbol{\xi}$, respectivamente, y $\boldsymbol{\varepsilon}$ y $\boldsymbol{\delta}$ son vectores $p \times 1$ y $q \times 1$ que representan los errores de medida asociados a $\boldsymbol{y}$ y a $\boldsymbol{x}$, respectivamente.

Se asume además que los errores de medida están incorrelados con $\boldsymbol{\eta}$, con $\boldsymbol{\xi}$, y entre ellos mismos. Además, se verifica:

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}) &= \mathbf{0}_{p\times 1} \\ \mathbb{E}(\boldsymbol{\delta}) &= \mathbf{0}_{q\times 1} \end{aligned}$$

Antes de finalizar con la especificación del modelo, es necesario definir algunos conceptos más que serán de gran importancia, ya que sobre ellos se construye la hipótesis principal sobre la que descansa toda la construcción de los SEM. En primer lugar, se definen las matrices de covarianzas entre variables observadas exógenas y endógenas:

Definición 2.4: *Se define la matriz de covarianzas entre variables observadas exógenas como:*

$$\Theta_{\boldsymbol{\delta}} = (\boldsymbol{\theta_{ij}})_{q\times q} = \mathbb{E}(\boldsymbol{\delta\delta^t}), \tag{2.5}$$

para todo modelo de medida (2.4).

Definición 2.5: *Se define la matriz de covarianzas entre variables observadas endógenas como:*

$$\Theta_{\boldsymbol{\varepsilon}} = (\boldsymbol{\theta_{ij}})_{p\times p} = \mathbb{E}(\boldsymbol{\varepsilon\varepsilon^t}), \tag{2.6}$$

para todo modelo de medida (2.4)

Estas definiciones suponen que si los errores de medida $\boldsymbol{\delta}$ y $\boldsymbol{\varepsilon}$ no están correlacionados, las matrices $\Theta_{\boldsymbol{\delta}}$ y $\Theta_{\boldsymbol{\varepsilon}}$ son matrices diagonales. Finalmente, y para concluir con la especificación del modelo SEM generalizado, se define la **matriz de covarianzas implicada**.

En el primer capítulo se vio que la hipótesis fundamental sobre la que se basa la construcción de los SEM era que la matriz de varianzas y covarianzas poblacional (Σ), era igual a la matriz de varianzas y covarianzas del modelo teórico ($\Sigma(\boldsymbol{\theta})$) (1.1). A partir de aquí definimos la matriz de covarianzas implicada de la siguiente forma:

Teorema 2.1: *La matriz de covarianzas implicada de un modelo SEM generalizado satisfaciendo:*

$$\Sigma = \Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \Sigma_{\boldsymbol{yy}}(\boldsymbol{\theta}) & \Sigma_{\boldsymbol{yx}}(\boldsymbol{\theta}) \\ \Sigma_{\boldsymbol{xy}}(\boldsymbol{\theta}) & \Sigma_{\boldsymbol{xx}}(\boldsymbol{\theta}) \end{pmatrix}, \tag{2.7}$$

donde $\Sigma_{\boldsymbol{yy}}(\boldsymbol{\theta})$, $\Sigma_{\boldsymbol{yx}}(\boldsymbol{\theta})$, $\Sigma_{\boldsymbol{xy}}(\boldsymbol{\theta})$ y $\Sigma_{\boldsymbol{xx}}(\boldsymbol{\theta})$, representan las matrices de covarianzas de las variables observadas determinadas por los respectivos subíndices, expresadas como función de parámetros desconocidos del modelo contenidos en el vector $\boldsymbol{\theta}$, viene dada por la expresión:

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \Lambda_{\boldsymbol{y}}(I-B)^{-1}(\Gamma\Phi\Gamma^t+\Psi)[(I-B)^{-1}]^t\Lambda_{\boldsymbol{y}}^t + \Theta_{\varepsilon} & \Lambda_{\boldsymbol{y}}(I-B)^{-}1\Gamma\Phi\Lambda_{\boldsymbol{x}}^t \\ \Lambda_{\boldsymbol{x}}\Phi\Gamma^t[(I-B)^{-1}]^t\Lambda_{\boldsymbol{y}}^t & \Lambda_{\boldsymbol{x}}\Phi\Lambda_{\boldsymbol{x}}^t + \Theta_{\delta} \end{pmatrix} \tag{2.8}$$

Demostración:

Se puede expresar $\Sigma_{\boldsymbol{yy}}(\boldsymbol{\theta})$ como:

$$\begin{aligned} \Sigma_{\boldsymbol{yy}}(\boldsymbol{\theta}) &= \mathbb{E}(\boldsymbol{yy^t}) \\ &= \mathbb{E}[(\Lambda_{\boldsymbol{y}}\boldsymbol{\eta}+\boldsymbol{\varepsilon})(\boldsymbol{\eta^t}\boldsymbol{\Lambda_y}^t+\boldsymbol{\varepsilon}^t)] \\ &= \Lambda_{\boldsymbol{y}}\mathbb{E}(\boldsymbol{\eta\eta^t})\Lambda_{\boldsymbol{y}}^t + \Theta_{\boldsymbol{\varepsilon}} \end{aligned}$$

Desarrollando ahora $\mathbb{E}(\boldsymbol{\eta\eta^t})$ haciendo uso de (2.3) y simplificando se obtiene:

$$\Sigma_{\boldsymbol{yy}}(\boldsymbol{\theta}) = \Lambda_{\boldsymbol{y}}(I-B)^{-1}(\Gamma\Phi\Gamma^t+\Psi)[(I-B)^{-1}]^t\Lambda_{\boldsymbol{y}}^t + \Theta_{\varepsilon}, \tag{2.9}$$

Haciendo ahora lo propio con $\Sigma_{\boldsymbol{yx}}(\boldsymbol{\theta})$:

$$\begin{aligned}\Sigma_{\boldsymbol{yx}}(\boldsymbol{\theta}) &= \mathbb{E}(\boldsymbol{yx^t}) \\ &= \mathbb{E}[(\Lambda_{\boldsymbol{y}}\boldsymbol{\eta}+\boldsymbol{\varepsilon})(\boldsymbol{\xi^t\Lambda_x}^t+\boldsymbol{\delta}^t)] \\ &= \Lambda_{\boldsymbol{y}}\mathbb{E}(\boldsymbol{\eta\xi^t})\Lambda_{\boldsymbol{x}}^t\end{aligned}$$

Empleando de nuevo (2.3) y simplificando se obtiene:

$$\Sigma_{\boldsymbol{yx}}(\boldsymbol{\theta}) = \Lambda_{\boldsymbol{y}}(I-B)^{-1}\Gamma\Phi\Lambda_{\boldsymbol{x}}^t \tag{2.10}$$

Haciendo uso de la propiedad $\Sigma_{\boldsymbol{xy}}(\boldsymbol{\theta}) = [\Sigma_{\boldsymbol{yx}}(\boldsymbol{\theta})]^t$ se tiene:

$$\Sigma_{\boldsymbol{xy}}(\boldsymbol{\theta}) = \Lambda_{\boldsymbol{x}}(I-B)^{-1}\Phi\Gamma^t[(I-B)^{-1}]^t\Lambda_{\boldsymbol{y}}^t \tag{2.11}$$

Finalmente, desarrollando $\Sigma_{\boldsymbol{xx}}(\boldsymbol{\theta})$:

$$\begin{aligned}\Sigma_{\boldsymbol{xx}}(\boldsymbol{\theta}) &= \mathbb{E}(\boldsymbol{xx^t}) \\ &= \mathbb{E}[(\Lambda_{\boldsymbol{x}}\boldsymbol{\xi}+\boldsymbol{\delta})(\boldsymbol{\xi^t\Lambda_x}^t+\boldsymbol{\delta}^t)] \\ &= \Lambda_{\boldsymbol{x}}\mathbb{E}(\boldsymbol{\xi\xi^t})\Lambda_{\boldsymbol{x}}^t+\Theta_{\boldsymbol{\delta}} \\ &= \Lambda_{\boldsymbol{x}}\Phi\Lambda_{\boldsymbol{x}}^t+\Theta_{\boldsymbol{\delta}}\end{aligned}$$

Luego hemos demostrado:

$$\Sigma_{\boldsymbol{xx}}(\boldsymbol{\theta}) = \Lambda_{\boldsymbol{x}}\Phi\Lambda_{\boldsymbol{x}}^t+\Theta_{\boldsymbol{\delta}} \tag{2.12}$$

Por tanto, haciendo uso de las expresiones (2.9), (2.10), (2.11), (2.12) se obtiene que que la matriz de covarianzas implicada viene dada por:

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix}\Lambda_{\boldsymbol{y}}(I-B)^{-1}(\Gamma\Phi\Gamma^t+\Psi)[(I-B)^{-1}]^t\Lambda_{\boldsymbol{y}}^t+\Theta_{\varepsilon} & \Lambda_{\boldsymbol{y}}(I-B)^{-}1\Gamma\Phi\Lambda_{\boldsymbol{x}}^t \\ \Lambda_{\boldsymbol{x}}\Phi\Gamma^t[(I-B)^{-1}]^t\Lambda_{\boldsymbol{y}}^t & \Lambda_{\boldsymbol{x}}\Phi\Lambda_{\boldsymbol{x}}^t+\Theta_{\delta}\end{pmatrix}$$

□

Corolario 2.1: *La matriz de covarianzas implicada de un modelo de variables observadas viene dada por la expresión:*

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix}(I-B)^{-1}(\Gamma\Phi\Gamma^t+\Psi)[(I-B)^{-1}]^t & (I-B)^{-1}\Gamma\Phi \\ \Phi\Gamma^t[(I-B)^{-1}]^t & \Phi\end{pmatrix} \tag{2.13}$$

Demostración:

Teniendo en cuenta que el modelo de variables observadas *(path analysis)*, es un caso particular del modelo general, tomando en (2.8) los valores $\Theta_{\boldsymbol{\varepsilon}} = 0$, $\Theta_{\boldsymbol{\delta}} = 0$,$\Lambda_{\boldsymbol{y}} = I_{p\times p}$ y $\Lambda_{\boldsymbol{x}} = I_{q\times q}$, se obtiene el resultado expuesto en (2.13). Esto es debido a que el modelo de variables observadas supone nulos los errores de medida, de donde se obtiene que $\boldsymbol{y} = \boldsymbol{\eta}$ y $\boldsymbol{x} = \boldsymbol{\xi}$.

□

Análogamente, se puede proceder para determinar la matriz de covarianzas implicada en un modelo de análisis factorial confirmatorio, que no es más que otro caso particular del modelo general:

Corolario 2.2: *La matriz de covarianzas implicada de un modelo de análisis factorial confirmatorio viene dada por la expresión:*

$$\Sigma(\boldsymbol{\theta}) = \Lambda_{\boldsymbol{x}} \Phi \Lambda_{\boldsymbol{x}}^t + \Theta_{\boldsymbol{\delta}} \tag{2.14}$$

Demostración:

Si se tiene en cuenta que el modelo de análisis factorial no estudia las relaciones entre variables latentes ($\boldsymbol{\eta}$ y $\boldsymbol{\xi}$), se puede asumir $\boldsymbol{B} = \boldsymbol{\Gamma} = 0$. En consecuencia, si no hay variables latentes endógenas, se puede tomar $\Lambda_{\boldsymbol{y}} = \Theta_{\boldsymbol{\varepsilon}} = \Psi = 0$, de modo que $\Sigma(\boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{t}})$, llegando a la expresión expuesta en (2.14) del mismo modo descrito en (2.12).

□

Fiabilidad y adecuación de los modelos de medida

Antes de continuar con la identificación del modelo, es conveniente (como en todo caso en que se realice una medición) comprobar la fiabilidad del modelo de medida que se ha construido. Para ello, se dispone de varios estadísticos que permiten llevar a cabo esta labor, no obstante se mostrará únicamente uno de ellos: el propuesto por Cronbach en (Cronbach, 1951), que recibe el nombre de Alpha de Cronbach. Se pueden consultar otros estadísticos que también son útiles para contrastar modelos de medida en (Bollen, 1989).

El estadístico Alpha de Cronbach es el método de estimación de fiabilidad más empleado en la actualidad, sobre todo en campos como la psicometría. Este método consigue solventar las dificultades que entrañaban otros métodos como el *split-halves* propuesto por Bollen. Antes de definir el estadístico, es necesario hacer lo propio con algunas propiedades de las medidas que se deben tener en cuenta:

Definición 2.6: *Dadas dos medidas $x_i = \alpha_i \tau + e_i$ y $x_j = \alpha_j \tau + e_j$, donde τ representa las puntuaciones reales que subyacen a las variables observadas x_i, x_j, y los errores e_i y e_j no están correlacionados se verifica:*

1. *Si $\alpha_i = \alpha_j = 1$ y $Var(e_i) = Var(e_j)$, entonces x_i y x_j son medidas equivalentes.*
2. *Si $\alpha_i = \alpha_j = 1$ y $Var(e_i) \neq Var(e_j)$, entonces x_i y x_j son medidas tau-equivalentes.*
3. *Si $\alpha_i \neq \alpha_j$ y $Var(e_i) \neq Var(e_j)$, entonces x_i y x_j son medidas congenéricas.*

Definición 2.7: *La fiabilidad o consistencia de una medida, $p_{x_i x_i}$, viene dada por la expresión:*

$$\rho_{x_i x_i} = \frac{\alpha_i^2 Var(\tau_i)}{Var(x_i)} \tag{2.15}$$

Proposición 2.1: *La fiabilidad de una medida, $\rho_{x_i x_i}$, verifica:*

$$\rho_{x_i x_i} = \rho^2_{x_i \tau_i} \tag{2.16}$$

Demostración:

Esta propiedad se basa en que, partiendo de la definición (2.15), la fiabilidad de una medida tau-equivalente (donde se verifica $\alpha_i = 1$) se puede expresar como el cociente $\frac{Var(\tau_i)}{Var(x_i)}$, de forma que se puede expresar la correlación entre las variables x_i y las puntuaciones reales τ_i, $\rho^2_{x_i \tau_i}$ de la siguiente manera:

$$\rho^2_{x_i \tau_i} = \frac{[Cov(x_i, \tau_i)]^2}{Var(x_i)Var(\tau_i)} = \frac{\alpha_i^2 [Var(\tau_i)]^2}{Var(x_i)Var(\tau_i)} = \frac{\alpha_i^2 Var(\tau_i)}{Var(x_i)} = \rho_{x_i x_i}$$

□

De este modo, $\rho_{x_i x_i}$ se puede interpretar como la varianza de x_i que es explicada por τ_i, con la varianza restante debida a los errores.

Ahora si se está en condiciones de definir el estadístico Alpha de Cronbach. Este coeficiente mide la fiabilidad de una suma de medidas tau-equivalentes o paralelas:

Definición 2.8 Alpha de Cronbach: *El coeficiente de fiabilidad de escala Alpha de Cronbach viene dado por la expresión:*

$$\alpha = \left(\frac{q}{q-1}\right)\left(1 - \frac{\sum_{i=1}^{q} Var(x_i)}{Var(H)}\right), \tag{2.17}$$

siendo q el número de ítems de la escala de medida, x_i las variables observadas del modelo y $H = \sum_{i=1}^{q} x_i$.

Haciendo uso de (2.16) y reformulando la ecuación resultante se demuestra que el coeficiente Alpha de Cronbach, verifica $\alpha = \rho_{HH}$, proporcionando así una fórmula general para la fiabilidad de la suma no ponderada de q medidas paralelas o tau-equivalentes. Dicha demostración puede consultarse en (Bollen, 1989).

El valor del estadístico Alpha de Cronbach estará entre 0 y 1, de modo que cuanto más cernano a 1 sea dicho valor, mejor será la fiabilidad o consistencia interna del modelo de medida que se esté considerando. El mínimo valor considerado "válido", como suele suceder en este tipo de ocasiones, varía según la fuente consultada, si bien es cierto que en general se suelen exigir valores superiores a 0.7, como puede verse en la tabla (2.2):

Tabla 2.2: Fiabilidad/Consistencia interna según el valor del estadístico α de Cronbach

Valor de α	Fiabilidad/Consistencia interna
$0.9 < \alpha < 1$	Muy buena
$0.8 < \alpha < 0.9$	Buena
$0.7 < \alpha < 0.8$	Aceptable
$\alpha < 0.7$	Inaceptable/Insuficiente

Análisis Factorial

Llegados a este punto, y dada su amplia presencia en la práctica, se ha creído conveniente hacer un inciso para tratar de forma algo más detallada uno de los casos particulares del modelo SEM generalizado: el modelo de análisis factorial. El objetivo de esta sección es principalmente mostrar algunos de los fundamentos teóricos en los que se sustenta el análisis factorial, y sobre todo presentar algunas de las técnicas que se emplearán posteriormente en la práctica para su aplicación.

Como se mencionó anteriormente, el objetivo fundamental del análisis factorial es describir, si es posible, la estructura de covarianzas entre variables en términos de un número menor[1] de variables no observables (variables latentes o factores). De este modo, existen dos técnicas diferentes. Por un lado, el análisis factorial *exploratorio*, trata de buscar o determinar precisamente esa estructura de covarianzas entre variables, de modo que no se parte de un modelo prefijado. En este caso, todas las variables latentes están influidas por las variables observadas, los errores de medida ($\boldsymbol{\delta}$) están incorrelados y a menudo se produce infraidentificación de parámetros. Por otro lado, el análisis factorial confirmatorio sí que parte de un modelo establecido (usualmente se suelen aplicar las dos técnicas de forma secuencial, de modo que el análisis factorial confirmatorio parte del modelo extraído de la aplicación del análisis factorial exploratorio para su "confirmación" o validación), y si una variable latente afecta a una variable observada, dicha influencia debe ser especificada.

En el caso de la aplicación práctica de este trabajo, se hará uso únicamente del análisis factorial confirmatorio, ya que se tratará de contrastar y evaluar una estructura de variables latentes establecida a priori de forma teórica. Para ello, se emplearán una serie de procedimientos y resultados que se detallan a continuación.

El primer paso, como se vio al principio de este capítulo, es comprobar que la muestra con la que se trabaja es adecuada para la aplicación de un análisis factorial. Para este fin, se debe evaluar en primer lugar el tamaño de la muestra. A continuación se deben calcular las correlaciones entre variables, para lo cual se suele construir la matriz de correlación. Se busca que las variables estén lo suficientemente correlacionadas como para poder aplicar un análisis factorial. El valor mínimo aceptado para poder continuar con la aplicación de la técnica varía según la fuente consultada, por ejemplo, Henson y Roberts propusieron que se debían buscar coeficientes de correlación superiores a 0.3 para que el análisis factorial tenga sentido, opinión que fue respaldada por Hair et al. en 1995. Estos últimos propusieron además una regla para evaluar los coeficientes de correlación entre variables del siguiente modo: ± 0.3 - bajo, ± 0.4 - importante, ± 0.5 - significativo.

Además, existen numerosas técnicas que permiten contrastar si una muestra es adecuada para realizar el análisis factorial. A continuación se describirán algunas de las más usadas y que están disponibles en la mayoría de paquetes estadísticos.

[1] Dado el hecho de que al aplicar un modelo de análisis factorial se reduce la dimensión del problema original a un número menor de factores, algunos autores engloban ésta técnica dentro de las conocidas como técnicas de reducción de la dimensión, si bien el objetivo del análisis factorial es más complejo, por lo que no debe confundirse con otros procedimientos totalmente diferentes, como el Análisis de Componentes Principales.

Prueba de esfericidad de Bartlett:

$$\begin{aligned} H_0 &: \Sigma_{q\times q} = \sigma^2 I_{q\times q} \\ H_1 &: \Sigma_{q\times q} \neq \sigma^2 I_{q\times q} \end{aligned} \tag{2.18}$$

El test de esfericidad de Bartlett contrasta la hipótesis nula de que la matriz de covarianzas es igual a la matriz identidad de dimensión $q \times q$ de modo que, de no haber evidencias en contra de dicha hipótesis, sería indicativo de que no existe relación significativa entre las variables objeto de estudio, por lo que no tendría sentido aplicar técnicas multivariantes y, en particular, un análisis factorial.

Sin embargo, este test es muy sensible a la falta de normalidad. Por este motivo surge una alternativa no paramétrica a esta prueba, el test de Fligner Killen que, al igual que el anterior, se basa en aproximaciones a una distribución χ^2_{q-1}, siendo q el número de ítems del vector aleatorio que se está analizando.

Estadístico de Keiser-Meyen Olkin: *Se define el estadístico KMO para la evaluación de la aplicabilidad de un análisis factorial, dado por la expresión:*

$$KMO = \frac{\sum_{i\neq j}\sum r_{ij}^2}{\sum_{i\neq j}\sum r_{ij}^2 + \sum_{i\neq j}\sum a_{ij}^2} \tag{2.19}$$

siendo r_{ij} los coeficientes de correlación y a_{ij} los coeficientes de correlación parcial.

Valores altos del estadístico KMO suponen una alta adecuación de la aplicación de un análisis factorial. En la siguiente tabla se muestran algunos valores propuestos por Kaiser que pueden ser tomados como referencia:

Tabla 2.3: Ajuste proporcionado por el estadístico KMO

Valor de KMO	Ajuste
$0.9 <$ KMO < 1	Muy bueno
$0.8 <$ KMO < 0.9	Meritorio
$0.7 <$ KMO < 0.8	Medio
$0.6 <$ KMO < 0.7	Mediocre
$0.5 <$ KMO < 0.6	Bajo
KMO ≤ 0.5	Insuficiente

El estadístico KMO proporciona una medida para el global de todas las variables del modelo. Si se desea contrastar variables individuales, se dispone del estadístico MSA:

Estadístico MSA: *El estadístico MSA_i mide la validez de la i-ésima variable para realizar un análisis factorial, y viene dado por la expresión:*

$$MSA_i = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} a_{ij}^2} \tag{2.20}$$

siendo r_{ij} los coeficientes de correlación y a_{ij} los coeficientes de correlación parcial.

La interpretación de los valores del estadístico MSA es análoga a la descrita en el caso del estadístico KMO.

Una vez que se ha comprobado que la muestra es adecuada para realizar un análisis factorial, se procede a extraer los factores, con el objetivo de hallar las cargas factoriales o coeficientes del modelo. Estas cargas factoriales se definen como los coeficientes que multiplican a los factores en la combinación lineal asociada a cada variable original. Sin embargo, ni esta matriz de cargas (denotada comúnmente por L) ni los factores son observables, lo que supone un problema de indeterminación desde dos puntos de vista. En primer lugar, los factores pueden explicarse con la misma precisión tanto en el caso de que estén correlacionados como en el caso de que no lo estén. Además, si el modelo factorial tiene solución, dichas soluciones serán infinitas, ya que para cualquier matriz ortogonal $T_{m\times m}$ (siendo m el número de factores) se puede determinar una solución. Este "problema" permitirá mejorar la interpretación de los factores comunes.

Existen numerosos métodos para la estimación de parámetros en un modelo factorial, siendo algunos de los más usados el de componentes principales (que será el que se utilice en el capítulo práctico de este trabajo), el de máxima verosimilitud, el de mínimos cuadrados ponderados, el de regresión, etc.

Otra cuestión a abordar es el número de factores a considerar. Para ello, se dispone de distintos procedimientos que se pueden aplicar en la práctica, como son la regla Kaiser (Guttman, 1954), según la cual se consideran tantos factores como autovalores de la matriz de varianzas y covarianzas sean mayores que uno. El razonamiento que reside detrás de esta regla es que un factor no debe explicar un porcentaje de varianza inferior al que hubiera explicado una variable de forma individual. Otra forma de determinar el número de factores, es el análisis del conocido como gráfico de sedimentación (Cattell, 1966), en el que se representa el tamaño de los autovalores, de forma que se intenta hallar el punto de inflexión en la gráfica donde se produce una reducción considerable de la cantidad de varianza explicada por dichos autovalores.

Ahora bien, se puede dar el caso de que algunas variables estén fuertemente relacionadas con más de un factor. En este caso suele ser útil la rotación de factores. Este procedimiento permite obtener un modelo de más fácil interpretación, para lo cual se maximizan las cargas factoriales de unos ítems y se minimizan las de otros. Se distinguen principalmente dos tipos de rotación: Rotación ortogonal, donde se realiza una rotación de los ejes coordenados produciéndose una transformación ortogonal de las cargas factoriales, y por tanto de los factores; y rotación oblicua, que se suele aplicar cuando no se obtienen buenos resultados con la rotación ortogonal. En la práctica, en este trabajo se solucionará este problema mediante el criterio *varimax*, que busca la representación con máxima variabilidad, maximizando la suma de las variabilidades para todos los factores.

2.2.2. Identificación del modelo

La fase de identificación del modelo es aquella que trata de comprobar, suponiendo correcto el modelo teórico, que todos los parámetros del modelo pueden ser estimados, lo cual ocurrirá si existe una solución única para cada uno de los parámetros. Por tanto, los parámetros en $\boldsymbol{\theta}$ están globalmente identificados si no existen dos vectores $\boldsymbol{\theta_1}$ y $\boldsymbol{\theta_2}$ tales que $\Sigma(\boldsymbol{\theta_1}) = \Sigma(\boldsymbol{\theta_2})$ a menos que $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$.

Se podrán distinguir hasta tres tipos de identificación posibles, en función del número de parámetros del modelo, y del número de elementos "no redundantes" en la matriz de varianzas y covarianzas:

Definición 2.9 modelo infraidentificado: *Se dice que un modelo está infraidentificado cuando la información contenida en Σ no es suficiente para estimar los parámetros* $\left(t > \frac{p+q(p+q+1)}{2}\right)$.

Definición 2.10 modelo saturado: *Se dice que un modelo está saturado si la información contenida en Σ es suficiente para estimar los parámetros, y la ecuación $\Sigma = \Sigma(\boldsymbol{\theta})$ tiene solución única* $\left(t = \frac{p+q(p+q+1)}{2}\right)$.

Definición 2.11 modelo sobreidentificado: *Se dice que un modelo está sobreidentificado si Σ contiene más información de la necesaria para estimar los parámetros del modelo, y la ecuación $\Sigma = \Sigma(\boldsymbol{\theta})$ tiene infinidas soluciones* $\left(t < \frac{p+q(p+q+1)}{2}\right)$. *El objetivo de todos los SEM sería trabajar con un modelo sobreidentificado.*

Siendo en los tres casos t el número de parámetros del modelo y $p+q$ el número total de variables observadas.

Ahora bien, ¿cuáles son los parámetros del modelo?. El vector $\boldsymbol{\theta}$ de parámetros del modelo está formado por:

- Varianzas y covarianzas de variables exógenas
- Coeficientes que relacionan las variables latentes con sus correspondientes indicadores
- Coeficientes de regresión (entre variables observadas o latentes).

Asimismo, estos parámetros pueden clasificarse en tres tipos:

- Parámetro libre: Son aquellos parámetros desconocidos que no tienen ninguna restricción para ser estimados.
- Parámetro fijo: Son aquellos a los que se les asigna previamente un cierto valor.
- Parámetro restringido: Son aquellos que al estimarse adquieren el valor de otro parámetro no fijo, o que pueden escribirse como función de parámetros libres.

Finalmente, para concluir este apartado de identificación de los SEM, se van a hacer algunas consideraciones referentes al primer caso de identificación del modelo: el de los modelos infraidentificados. Como se dijo anteriormente, esto se produce cuando la matriz Σ no tiene información suficiente para estimar los parámetros. Esto supondrá que dichas estimaciones, los correspondientes errores de estimación, o las conclusiones derivadas por ejemplo de realizar un test χ^2 no serán fiables. Con el objetivo de detectar en qué situación de identificación se encuentra un determinado modelo, en (Bollen, 1989) se proponen algunas técnicas que son útiles para saber si un modelo está infraidentificado, o si por el contrario su identificación es adecuada.

t-Rule

Teorema 2.2: *Una condición necesaria para la identificación de los parámetros del modelo viene dada por:*

$$t < \frac{1}{2}(p+q)(p+q+1), \tag{2.21}$$

siendo t el número de parámetros libres y no restringidos en $\boldsymbol{\theta}$, p el número de variables observadas en $\boldsymbol{y}$ y q el número de variables observadas en $\boldsymbol{x}$.

Los elementos "no redundantes" de $\Sigma = \Sigma(\boldsymbol{\theta})$ implican que habrá $\frac{1}{2}(p+q)(p+q+1)$ ecuaciones. Así, si el número de parámetros desconocidos en $\boldsymbol{\theta}$ excede el número de ecuaciones, la identificación del modelo no será posible. La demostración de este teorema puede consultarse en (Bollen, 1989).

Para conseguir la suficiencia de la que carece la *t-Rule*, se presenta el segundo de los procedimientos, la regla de los dos pasos:

Regla de los dos pasos

Como su propio nombre indica, esta técnica consta de dos pasos o fases para su aplicación. En el primer paso, se debe considerar el modelo como en el caso de un análisis factorial confirmatorio. Esto supone que las únicas relaciones entre las variables latentes que se considerarán son sus varianzas y covarianzas Φ. De esta forma, los términos B, Γ y Ψ son ignorados. Una vez reformulado el modelo como un análisis factorial confirmatorio, se determina si está identificado. Para ello, se dispone de varias técnicas que se detallan a continuación:

Regla de los tres indicadores: *Se dirá que un modelo de medida está correctamente identificado si, además de verificarse las condiciones del Teorema 2.2, se dan las siguientes condiciones:*

1. *Hay al menos tres indicadores por cada variable latente*
2. *Cada fila de $\Lambda_{\boldsymbol{x}}$ contiene un y sólo un elemento distinto de* 0
3. *$\Theta_{\boldsymbol{\delta}}$ es diagonal*

Regla de los dos indicadores: *Se dirá que un modelo de medida está correctamente identificado si, además de verificarse las condiciones del Teorema 2.2, se dan las siguientes condiciones:*

1. *Hay más de una variable latente, y a cada una de ellas le corresponden dos indicadores.*
2. *Cada indicador es causado por una única variable latente (cada fila de $\Lambda_{\boldsymbol{x}}$ tiene un único elemento $\neq 0$) y cada variable latente está correlada con al menos otra variable latente (ningún elemento de Φ es nulo).*
3. *$\Theta_{\boldsymbol{\delta}}$ es diagonal.*

El segundo paso consiste en examinar la ecuación de variables latentes del modelo estructural (dado por (2.2)), y considerarla como si se tratara de una ecuación estructural de variables observadas. Esto implica asumir que cada variable latente corresponde con una variable observada perfectamente medida. A continuación se determina si los términos B, Γ y Ψ están correctamente medidos. Para ello se proponen distintas técnicas que se explican a continuación:

Teorema 2.3 (Regla de la B nula): *En un modelo donde no hay relaciones entre variables endógenas (ninguna variable endógena afecta a otra del mismo tipo), la matriz B es nula. Esta condición es suficiente para identificar el modelo.*

Demostración:

Para establecer la identificación de un modelo donde $B = 0$, se va a tratar de demostrar que los parámetros en Γ, Φ y Ψ se pueden expresar como función de los parámetros identificados de Σ. Así, sustituyendo $B = 0$ en (2.8) se tiene:

$$\begin{pmatrix} \Sigma_{\boldsymbol{yy}} & \Sigma_{\boldsymbol{yx}} \\ \Sigma_{\boldsymbol{xy}} & \Sigma_{\boldsymbol{xx}} \end{pmatrix} = \begin{pmatrix} (\Gamma\Phi\Gamma^t + \Psi) & \Gamma\Phi \\ \Phi\Gamma^t & \Phi \end{pmatrix}, \tag{2.22}$$

de donde se deduce $\Sigma_{\boldsymbol{xx}} = \Phi$, lo que implica que Φ está bien identificado.

Centrándose ahora en el cuadrante inferior izquierdo se tiene:

$$\Phi\Gamma^t = \Sigma_{\boldsymbol{xy}} \Leftrightarrow \Sigma_{\boldsymbol{xx}}\Gamma^t = \Sigma_{\boldsymbol{xy}} \Leftrightarrow \Gamma^t = \Sigma_{\boldsymbol{xx}}^{-1}\Sigma_{\boldsymbol{xy}} \tag{2.23}$$

De esta forma, se ha podido expresar Γ como función de matrices de covarianzas identificadas, por lo que también está debidamente identificada. Finalmente, se realiza un proceso similar para Ψ:

$$\begin{aligned} \Psi &= \Sigma_{\boldsymbol{yy}} - \Gamma\Phi\Gamma^t \\ &= \Sigma_{\boldsymbol{yy}} - \Sigma_{\boldsymbol{yx}}\Sigma_{\boldsymbol{xx}}^{-1}\Sigma_{\boldsymbol{xx}}\Sigma_{\boldsymbol{xx}}^{-1}\Sigma_{\boldsymbol{xy}} \\ &= \Sigma_{\boldsymbol{yy}} - \Sigma_{\boldsymbol{yx}}\Sigma_{\boldsymbol{xx}}^{-1}\Sigma_{\boldsymbol{xy}} \end{aligned}$$

Luego hemos expresado Ψ como función de matrices de covarianzas identificadas y por tanto también está identificado. Se ha demostrado así que cuando $B = 0$, los parámetros desconocidos Φ, Γ y Ψ están identificados.

□

Para terminar este apartado, se va a exponer una última regla que, al igual que la anterior, supone una condición suficiente para la identificación de modelos.

Teorema 2.4 Regla recursiva: *Si B es una matriz triangular inferior y Ψ es una matriz diagonal, entonces el modelo está correctamente identificado.*

La demostración de este teorema se puede consultar en (Bollen, 1989).

Por tanto, aplicando la regla de los dos pasos se puede ver si el modelo de medida (1^{er}paso) y el modelo estructural (2^{o}paso) están correctamente identificados, comprobando así la identificación del modelo al completo. Existen otras reglas adicionales tales como la regla de condición de rango o la regla de condición de orden, que no se detallarán por exceder los objetivos de este trabajo. Aún así, si se deseara consultar esta información, se encuentra disponible en (Bollen, 1989).

2.2.3. Estimación del modelo

En esta etapa se estudirá la relación entre la matriz de covarianzas de las variables observadas y los parámetros del modelo. En la práctica, al ser desconocida la matriz de varianzas y covarianzas poblacional, ésta se estimará utilizando la correspondiente matriz de varianzas y covarianzas muestral, $\widehat{\Sigma}$, buscando que $\widehat{\Sigma}$ y Σ sean lo más similares posible. Para ello, se buscará minimizar una función de la matriz de residuos $(\widehat{\Sigma} - \Sigma)$, denotada por $F(\widehat{\Sigma}, \Sigma(\boldsymbol{\theta}))$ y llamada también **función de ajuste**.

Las funciones de ajuste que se presentan a continuación verifican las siguientes propiedades:

1. $F(\widehat{\Sigma}, \Sigma(\boldsymbol{\theta}))$ es un escalar.
2. $F(\widehat{\Sigma}, \Sigma(\boldsymbol{\theta})) \geq 0$
3. $F(\widehat{\Sigma}, \Sigma(\boldsymbol{\theta})) = 0 \Leftrightarrow \Sigma(\boldsymbol{\theta}) = \widehat{\Sigma}$
4. $F(\widehat{\Sigma}, \Sigma(\boldsymbol{\theta}))$ es continua en $\widehat{\Sigma}$ y en $\Sigma(\boldsymbol{\theta})$.

Según Brown (1984, 1966), los estimadores resultantes de minimizar funciones de ajuste que satisfacen las propiedades anteriores son estimadores consistentes de $\boldsymbol{\theta}$. A continuación, se detallarán los tres métodos que se suelen emplear con mayor frecuencia para realizar estimaciones en los SEM: máxima verosimilitud (ML), mínimos cuadrados generalizados (GLS) y mínimos cuadrados ponderados (WLS).

Estimación por Máxima Verosimilitud

En primer lugar, se debe tener en cuenta que para poder aplicar este método de estimación han de verificarse las siguientes restricciones:

1. $\Sigma(\boldsymbol{\theta})$ y $\widehat{\Sigma}$ son definidas positivas, y por tanto son no singulares.
2. $\boldsymbol{x}$ e $\boldsymbol{y}$ siguen una distribución normal multivariante, y $\widehat{\Sigma}$ sigue una distribución Wishart

Definición 2.17: *Se define la función de ajuste a minimizar para el método de máxima verosimilitud F_{ML} como aquella que viene dada por la expresión:*

$$F_{ML} = \log|\Sigma(\boldsymbol{\theta})| + \mathrm{tr}(\widehat{\Sigma}\Sigma^{-1}(\boldsymbol{\theta})) - \log|\widehat{\Sigma}| - (p+q) \tag{2.24}$$

La estimación por máxima verosimilitud parte considerando una muestra aleatoria de tamaño N de variables aleatorias independientes e idénticamente distribuidas de una variable Z. La función de densidad para cada $Z_i, \quad i = 1, \ldots, N$ es $f(Z_i; \boldsymbol{\theta})$, siendo $\boldsymbol{\theta}$ un parámetro fijo. Como las Z_i son independientes entre sí, la función de densidad conjunta viene dada por:

$$f(Z_1, \ldots, Z_n; \boldsymbol{\theta}) = f(Z_1; \boldsymbol{\theta}) f(Z_2; \boldsymbol{\theta}) \ldots f(Z_N; \boldsymbol{\theta}), \tag{2.25}$$

es decir, se puede expresar la función de densidad conjunta como producto de las correspondientes densidades marginales. Una vez que observamos un conjunto específico de valores para $Z_1, \ldots, Z_N$ en una muestra, podemos escribir la función:

$$L(\boldsymbol{\theta}; Z_1, \dots Z_N) = L(\boldsymbol{\theta}; Z_1)L(\boldsymbol{\theta}; Z_2) \dots L(\boldsymbol{\theta}; Z_N), \tag{2.26}$$

donde $L(\boldsymbol{\theta}; Z_i)$ es el valor de $f(Z_i; \boldsymbol{\theta})$ cuando Z_i está en su valor muestral. La ecuación (2.26) es la *función de verosimilitud*, abreviada comunmente por $L(\boldsymbol{\theta})$ o simplemente L. Aunque puedan parecer similares, las ecuaciones (2.25) y (2.26) tienen importantes diferencias. En (2.25), $\boldsymbol{\theta}$ es un parámetro fijo y las $Z_1, \dots, Z_N$ son variables aleatorias. Sin embargo en (2.26), las Z_i son valores fijos para una muestra determinada, y la magnitud de $L(\boldsymbol{\theta})$ es función de $\boldsymbol{\theta}$. Sea $\widehat{\boldsymbol{\theta}}$ un estimador de $\boldsymbol{\theta}$. En la estimación por máxima verosimilitud se pretende hallar el estimador $\widehat{\boldsymbol{\theta}}$ que maximice la probabilidad (o verosimilitud) de generar los valores muestrales dados de Z_i, esto es, se busca el $\widehat{\boldsymbol{\theta}}$ tal que maximice (2.26) para los valores muestrales $Z_1, \dots, Z_N$.

Para hallar el máximo de $L(\boldsymbol{\theta})$, a menudo se busca maximizar en su lugar $\log(L(\boldsymbol{\theta}))$, lo cual simplifica la búsqueda del máximo, y no afecta al valor resultante de $\boldsymbol{\theta}$ (ya que el logaritmo de un número es una función monótona de dicho número). A continuación se deriva $\log L(\boldsymbol{\theta})$ respecto a $\boldsymbol{\theta}$, igualando posteriormente el resultado a 0 y despejando el valor de $\boldsymbol{\theta}$. Basta calcular la segunda derivada respecto a $\boldsymbol{\theta}$ y comprobar que es estrictamente menor que 0 para asegurar que dicho valor de $\boldsymbol{\theta}$ es aquel que maximiza $\log L(\boldsymbol{\theta})$ y que por tanto se trata del estimador de máxima verosimilitud.

Como $\boldsymbol{x}$ e $\boldsymbol{y}$ son dos vectores que siguen una distribución normal multivariante, se pueden expresar como un único vector z $(p + q \times 1)$, siendo z su desviación. Su función de densidad viene dada por la siguiente expresión:

$$f(z; \Sigma) = (2\pi)^{-(p+q)/2}|\Sigma|^{-1/2}\exp\left[\left(-\frac{1}{2}\right) z^t \Sigma^{-1} z\right] \tag{2.27}$$

Para una muestra aleatoria de N observaciones independientes de z la función de densidad conjunta será:

$$f(z_1, z_2, \dots, z_N; \Sigma) = f(z_1; \Sigma) f(z_2; \Sigma) \dots f(z_N; \Sigma) \tag{2.28}$$

A partir de la observación de dicha muestra y haciendo uso de la hipótesis fundamental (1.1), la función de verosimilitud viene dada por:

$$L(\boldsymbol{\theta}) = (2\pi)^{-N(p+q)/2}|\Sigma(\boldsymbol{\theta})|^{-N/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{N} z_i^t \Sigma^{-1}(\boldsymbol{\theta}) z_i\right] \tag{2.29}$$

Tomando ahora logaritmos:

$$\log L(\boldsymbol{\theta}) = \frac{-N(p+q)}{2}\log(2\pi) - \left(\frac{N}{2}\right)\log|\Sigma(\boldsymbol{\theta})| - \left(\frac{1}{2}\right)\sum_{i=1}^{N} z_i^t \Sigma^{-1}(\boldsymbol{\theta}) z_i \tag{2.30}$$

Se puede reescribir el último término como sigue:

$$
\begin{aligned}
-\left(\frac{1}{2}\right)\sum_{i=1}^{N} z_i^t\Sigma^{-1}(\boldsymbol{\theta})z_i &= -\left(\frac{1}{2}\right)\sum_{i=1}^{N}\mathrm{tr}[z_i^t\Sigma^{-1}(\boldsymbol{\theta})z_i] \\
&= -\left(\frac{N}{2}\right)\sum_{i=1}^{N}\mathrm{tr}[N^{-1}z_i z_i^t\Sigma^{-1}(\boldsymbol{\theta})] \\
&= -\left(\frac{N}{2}\right)\mathrm{tr}[\widehat{\Sigma}^*\Sigma^{-1}(\boldsymbol{\theta})], \qquad (2.31)
\end{aligned}
$$

siendo $\widehat{\Sigma}^*$ el estimador de máxima verosimilitud muestral de la matriz de covarianzas. Asimismo, (2.30) se puede reescribir de la siguiente forma:

$$
\begin{aligned}
\log L(\boldsymbol{\theta}) &= C - \left(\frac{N}{2}\right)\log|\Sigma(\boldsymbol{\theta})| - \left(\frac{N}{2}\right)\mathrm{tr}[\widehat{\Sigma}^*\Sigma^{-1}(\boldsymbol{\theta})] \\
&= C - \frac{N}{2}\left\{\log|\Sigma(\boldsymbol{\theta})| + \mathrm{tr}[\widehat{\Sigma}^*\Sigma^{-1}(\boldsymbol{\theta})]\right\}, \qquad (2.32)
\end{aligned}
$$

siendo C una constante. Si comparamos las expresiones de (2.24) y (2.32) observamos que difieren en algunos aspectos, sin embargo, ninguno de estos tiene influencia en la elección del estimador $\widehat{\boldsymbol{\theta}}$. Así, ni la ausencia del término constante C en (2.24), ni la presencia de $(-\log|\widehat{\Sigma}| - (p+q))$ en (2.32) tiene impacto en la elección de $\widehat{\boldsymbol{\theta}}$ ya que, dada una muestra cualquiera, $\widehat{\Sigma}$ y $(p+q)$ son constantes. El único efecto que tendrá la presencia del término $(-N/2)$ en (2.32) pero no en (2.24) es que hará que se minimice en lugar de maximizar. La última diferencia entre las dos expresiones es que en (2.24) se trabaja con $\widehat{\Sigma}$, mientras que en (2.32) aparece la matriz $\widehat{\Sigma}^*$, si bien estas dos matrices serán iguales para tamaños muestrales suficientemente grandes, al verificarse $\widehat{\Sigma}^* = \frac{(N-1)}{N}\widehat{\Sigma}$.

Proposición 2.3: *Todo estimador de máxima verosimilitud satisfaciendo las correspondientes restricciones expuestas anteriormente, tiene las siguientes propiedades:*

1. *Es consistente, y por tanto, se cumple $\widehat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$ cuando $N \to \infty$ (el estimador converge al verdadero valor del parámetro al aumentar el tamaño muestral).*
2. *Es asintóticamente eficiente, por lo que su varianza es mínima.*
3. *Es asintóticamente insesgado. Por tanto se cumple $\mathbb{E}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.*

Para finalizar la estimación por máxima verosimilutud, es interesante mencionar que la función de ajuste F_{ML} permite realizar un contraste de hipótesis para comprobar el ajuste del modelo:

$$
\begin{cases} H_0: & \Sigma = \Sigma(\boldsymbol{\theta}) \\ H_1: & \Sigma \neq \Sigma(\boldsymbol{\theta}) \end{cases},
$$

de modo que, bajo H_0, el estadístico $(N-1)F_{ML}$ se distribuye asintóticamente según una distribución χ^2 con $\frac{1}{2}(p+q)(p+q+1) - t$ grados de libertad:

$$
(N-1)F_{ML} \overset{H_0}{\sim} \chi^2_{\frac{1}{2}(p+q)(p+q+1)-t} \qquad (2.33)
$$

Estimación por mínimos cuadrados ponderados

Este método de estimación, a diferencia del anterior permite trabajar con datos que no sigan una distribución normal multivariante, además de con variables ordinales o dicotómicas. Se partirá de nuevo definiendo la función de ajuste:

Definición 2.18: *Se define la función de ajuste para el método de mínimos cuadrados ponderados, F_{WLS} como aquella que viene dada por la expresión:*

$$F_{WLS} = [\boldsymbol{s} - \sigma(\boldsymbol{\theta})]^t W^{-1} [\boldsymbol{s} - \sigma(\boldsymbol{\theta})], \tag{2.34}$$

donde $\boldsymbol{s}$ es un vector $\frac{1}{2}(p+q)(p+q+1) \times 1$ formado por los elementos no redundantes de $\widehat{\Sigma}$, $\sigma(\boldsymbol{\theta})$ es un vector del mismo orden formado por los elementos de $\Sigma(\boldsymbol{\theta})$ y $\boldsymbol{\theta}$ es un vector $t \times 1$ de parámetros del modelo. W^{-1} es una matriz de pesos (o ponderaciones) $\frac{1}{2}(p+q)(p+q+1) \times \frac{1}{2}(p+q)(p+q+1)$ definida positiva.

La elección de $\boldsymbol{\theta}$ se hace buscando el objetivo de minimizar la suma ponderada del cuadrado de las desviaciones de $\boldsymbol{s}$ respecto a $\sigma(\boldsymbol{\theta})$. De este modo, el procedimiento es análogo al llevado a cabo al realizar la estimación por mínimos cuadrados ponderados en regresión, donde se busca minimizar el cuadrado de la diferencia entre la variable observada y la variable respuesta, mediante la selección de los coeficientes de regresión. La diferencia es que en este caso, los valores observados y "predichos" son covarianzas en lugar de valores individuales.

Al igual que ocurría con la estimación proporcionada por el método de máxima verosimilitud, el estimador $\widehat{\boldsymbol{\theta}}$ obtenido a través de F_{WLS} es un estimador consistente de $\boldsymbol{\theta}$. Además, un resultado relevante es el demostrado por *Brown (1982, 1984)*, según el cual, si W es un estimador consistente de la matriz de covarianzas asintótica de $\boldsymbol{s}$, entonces el estimador $\widehat{\boldsymbol{\theta}}$ proporcionado por F_{WLS} es asintóticamente eficiente.

Para evitar la sobreestimación del estadístico χ^2, es preciso trabajar con muestras de tamaño considerable, generalmente superiores a 5000 registros, lo cual supone una de las mayores desventajas de este método. Además, un número de variables observadas superior a 20, hace que la matriz de pesos W aumente de forma considerable, lo que hace más difícil su estimación. Para tratar de solucionar este problema, *Muthén* propuso en 1993 un método alternativo consistente en considerar únicamente la diagonal de la matriz W. Esta técnica se conoce como método de mínimos cuadrados ponderados diagonalizados, y su función de ajuste viene dada por la expresión:

$$F_{DWLS} = [\boldsymbol{s} - \sigma(\boldsymbol{\theta})]^t \mathrm{diag}(W^{-1}) [\boldsymbol{s} - \sigma(\boldsymbol{\theta})] \tag{2.35}$$

Estimación por mínimos cuadrados no ponderados

En primer lugar, como en los métodos anteriores se define la función de ajuste:

Definición 2.19: *Se define la función de ajuste para el método de mínimos cuadrados no ponderados como aquella dada por la expresión:*

$$F_{ULS} = \frac{1}{2} \mathrm{tr}[(S - \Sigma(\boldsymbol{\theta}))^2] \tag{2.36}$$

Esta función minimiza la suma de cruadrados de cada elemento de la matriz de residuos $(S - \Sigma(\boldsymbol{\theta}))$, al igual que ocurre en la regresión de mínimos cuadrados ordinarios (OLS). Sin

embargo, la diferencia entre estas dos técnicas es que con F_{ULS} se calculan las diferencias entre las varianzas y covarianzas muestrales "predichas" por el modelo, mientras que el método OLS calcula las diferencias entre los valores muestrales observados de la variable dependiente y los "predichos" por el modelo.

Entre las ventajas que presenta este método está el hecho de que, al igual que en los casos anteriores, el estimador de $\boldsymbol{\theta}$ que calcula es consistente, y esto sin necesidad de asumir que las variables observadas sigan ninguna distribución concreta (por lo que no se impone la normalidad multivariante como en el caso del método ML).

Respecto a las desventajas, el método ULS no proporciona el estimador asintóticamente más eficiente, ya que el calculado mediante máxima verosimilitud tiene una eficiencia mayor. Además, estos estimadores no son invariantes ante cambios de escala. Por ejemplo, los valores de F_{ULS} varían si se trabaja con la matriz de correlaciones en lugar de con la matriz de covarianzas. Finalmente, este método no posee a día de hoy ningún procedimiento que compruebe una posible sobreidentificación, mientras que el método ML dispone del propuesto por *Brown (1984, 1982)* para este fin.

Estimación por mínimos cuadrados generalizados

En este método se busca minimizar las diferencias cuadráticas entre los elementos observados de S y los correspondientes elementos "predichos" de $\Sigma(\boldsymbol{\theta})$, en un proceso análogo al realizado en el método OLS. La diferencia principal radica en que el método OLS considera los valores de la variable respuesta observados y "predichos" por el modelo para observaciones individuales, mientras que F_{GLS} pone el foco en las covarianzas observadas y "predichas."

Definición 2.20: *Se define la función de ajuste para el método de mínimos cuadrados generalizados como aquella que viene dada por la expresión:*

$$F_{GLS} = \frac{1}{2}\text{tr}[\{(S - \Sigma(\boldsymbol{\theta}))W^{-1}\}^2], \tag{2.37}$$

donde W^{-1} es una matriz de pesos o ponderaciones para la matriz de residuos. W^{-1} es una matriz aleatoria que converge en probabilidad a una matriz definida positiva cuando $N \to \infty$, o es una matriz de constantes definida positiva.

Además, F_{ULS}, vista en el apartado anterior y dada por (2.36) es un caso particular de F_{GLS} cuando $W^{-1} = I$. El estimador $\widehat{\boldsymbol{\theta}}$ obtenido a partir de F_{GLS}, con alguna matriz W^{-1} satisfaciendo las condiciones anteriores es un estimador consistente de $\boldsymbol{\theta}$. Sin embargo, no todas las elecciones posibles de la matriz W^{-1}, conducen a estimadores eficientes.

2.2.4. Evaluación del modelo

Esta es la etapa final en la construcción de modelos de ecuaciones estructurales, y posiblemente una de las más importantes ya que, en ella se determinará si el modelo que hemos construido en las etapas anteriores es realmente útil, y por tanto es capaz de identificar las relaciones entre todas las variables.

Para ello, se estudian las diferencias entre la matriz de covarianzas predicha por el modelo $\Sigma(\widehat{\boldsymbol{\theta}})$ y la matriz de covarianzas muestral S. No existe un consenso total sobre qué estadísticos son los más apropiados para llevar a cabo esta tarea, por lo que a continuación

se van a exponer algunos de los más utilizados, si bien se debe tener en cuenta que existen otros no mencionados en este trabajo y que de igual forma sirven para la evaluación de modelos.

Estadístico χ^2

Al igual que la mayoría de métodos de evaluación de modelos que se verán posteriormente, la prueba χ^2 de bondad de ajuste está dentro de las llamadas medidas de ajuste global. En (2.33) se propuso un estadístico que permitía evaluar la bondad del ajuste proporcionado por el modelo mediante un test χ^2. De este modo, valores significativamente elevados del estadístico suponen que las matrices S y $\Sigma(\hat{\boldsymbol{\theta}})$ son significativamente distintas, lo que conduciría a rechazar la hipótesis nula, y por tanto el modelo. Esto hace que se busquen valores "bajos" de χ^2, con el fin de no encontrar evidencias en contra de la hipótesis nula $H_0 : \Sigma = \Sigma(\boldsymbol{\theta})$. Generalmente, se suele fijar el umbral para validar o rechazar el modelo en un p-valor = 0.05, de modo que el ajuste ofrecido por el modelo sería adecuado si 0.05 < p-valor $\leq$ 1, si bien esto puede variar según el criterio del investigador.

Sin embargo, utilizar el p-valor como criterio conlleva algunas desventajas. La mayor radica en su alta dependencia del tamaño de la muestra ya que, debido a la definición del propio estadístico, valores muestrales elevados pueden hacer que el estadístico aumente su valor, lo cual llevará a rechazar H_0 y por ende el modelo. Esto hará que se incremente a su vez el error de tipo I. Para evitar este problema, se suele tomar como medida el cociente entre el valor del estadístico y sus grados de libertad, en lugar de únicamente los segundos. Se muestra una clasificación de valores comunes del estadístico χ^2/gl usados para evaluar el ajuste del modelo en la siguiente tabla:

Tabla 2.4: Bondad del ajuste del modelo según el valor del estadísitico χ^2

Valor de χ^2/gl	Bondad del ajuste del modelo
Menor que 1	Sobreidentificación
Entre 1 y 2	Ajuste muy bueno
Entre 2 y 3	Ajuste bueno/aceptable
Entre 3 y 5	Ajuste poco aceptable
Mayor que 5	Ajuste inaceptable

Índices de ajuste

Los índices de ajuste o *fit index*, forman parte de las llamadas medidas incrementales, y su objetivo es estudiar el ajuste logrado por un modelo, en comparación con su modelo base, esto es, el modelo más simple posible donde no existe asociación entre variables. En primer lugar, se muestra el índice de ajuste propuesto por *Bentler y Bonett (1980)*:

$$\begin{aligned} \triangle_1 &= \frac{F_b - F_m}{F_b} \\ &= \frac{\chi_b^2 - \chi_m^2}{\chi_b^2}, \end{aligned} \tag{2.38}$$

donde F_b es la función de ajuste del modelo base y F_m es la función de ajuste del modelo que se quiere evaluar. $\triangle_1$ toma valores entre 0 y 1, de modo que cuanto más cercano a

1 sea el valor, mejor será el ajuste, y cuanto más cercano a 0 el ajuste será peor. Como suele ocurrir en estos casos, no hay un valor límite para considerar un ajuste como bueno o aceptable, pero rara vez se aceptan modelos cuyo índice de ajuste sea inferior a 0.9. Una limitación de este índice es que no tiene en cuenta los grados de libertad de los estadísticos χ^2. Además, al no considerar tampoco el tamaño de la muestra con la que se trabaja, si se comparan los índices de ajuste de dos modelos con tamaños muestrales distintos, puede dar dar la impresión de que el mayor de ellos proporciona un mejor ajuste, incluso aunque ambos sean en realidad igual de "buenos" para reproducir el modelo teórico. Para tratar de solucionar esto, Bollen (1989) propone una segunda versión del índice de ajuste, dado por la siguiente expresión:

$$\begin{aligned} \triangle_2 &= \frac{F_b - F_m}{F_b - [\mathrm{gl}_m/(N-1)]} \\ &= \frac{\chi_b^2 - \chi_m^2}{\chi_b^2 - \mathrm{gl}_m}, \end{aligned} \tag{2.39}$$

Entre las propiedades de este segundo índice destacan las siguientes: 1) No se ve afectado por tamaños diferentes de muestra, 2) no está acotado al rango (0,1), 3) valores de $\triangle_2$ muy superiores a 1 podrían ser indicativos de un posible sobreajuste del modelo y 4) se verifica que $\triangle_1 < \triangle_2$, cuando el denominador de $\triangle_2$ es positivo.

Bollen propuso en 1986 un tercer índidce de ajuste, dado por la expresión:

$$\begin{aligned} \rho_1 &= \frac{(F_b/\mathrm{gl}_b) - (F_m/\mathrm{gl}_m)}{F_b/\mathrm{gl}_b} \\ &= \frac{(\chi_b^2/\mathrm{gl}_b) - (\chi_m^2/\mathrm{gl}_m)}{(\chi_b^2/\mathrm{gl}_b)}, \end{aligned} \tag{2.40}$$

El conjunto de valores que puede tomar ρ_1 es el mismo que en caso de $\triangle_1$, es decir, $0 \leq \rho_1 \leq 1$, de modo que cuanto más cercano a 1 sea su valor, mejor será el ajuste. Sin embargo, ρ_1 presenta el mismo problema que $\triangle_1$ respecto a su dependencia del tamaño de la muestra. Por este motivo, para finalizar este apartado sobre índices de ajuste, se presenta un último índice, anterior en fecha al último expuesto, que fue propuesto por Tucker y Lewis en 1973, y que disminuye la dependencia que sufría ρ_1 respecto al tamaño muestral.

$$\begin{aligned} \rho_2 &= \frac{(F_b/\mathrm{gl}_b) - (F_m/\mathrm{gl}_m)}{F_b/\mathrm{gl}_b - [1/(N-1)]} \\ &= \frac{(\chi_b^2/\mathrm{gl}_b) - (\chi_m^2/\mathrm{gl}_m)}{(\chi_b^2/\mathrm{gl}_b) - 1}, \end{aligned} \tag{2.41}$$

Al igual que en caso anterior, ρ_2 toma valores entre 0 y 1.

Root Mean Square Error of Aproximation

Este índice forma parte, al igual que el estadíscio χ^2 de las llamadas medidas de ajuste global. Se denotará como $RMSEA$, por sus siglas en inglés (raíz cuadrática media del error de aproximación en español), y estudia las diferencias entre la matriz de covarianzas dada por el modelo, y la matriz de covarianzas observadas. Se puede calcular mediante la siguiente expresión:

$$RMSEA = \sqrt{\frac{\chi_t^2 - \mathrm{gl}_t}{(N-1)\mathrm{gl}_t}} \tag{2.42}$$

Brown y Cudeck propusieron en 1993 un valor máximo de este índice de 0.05 para considerar el ajuste logrado por un modelo como aceptable. Valores superiores a 0.08 conducirían generalmente a rechazar el correspondiente modelo. Una desventaja es que el índice RMSEA tiende a rechazar modelo válidos cuando el tamaño muestral es bajo.

Root Mean Square Residual

Al igual que el estadístico χ^2 o el índice $RMSEA$, pertenece a las medidas de ajuste global, y estudia las diferencias entre la matriz de covarianzas observada y la misma predicha por el modelo. Su expresión viene dada por:

$$RMR = \left[2 \sum_{i=1}^{q} \sum_{j=1}^{i} \frac{(s_{ij} - \widehat{\sigma}_{ij})^2}{q(q+1)} \right]^{1/2}, \tag{2.43}$$

donde s_{ij} y $\widehat{\sigma}_{ij}$ son componentes pertenecientes a las matrices S y $\Sigma(\widehat{\boldsymbol{\theta}})$, respectivamente. Los valores que se suelen considerar para aceptar o rechazar un modelo son los mismos que en el caso del índice RMSEA, asumiendo como aceptable el ajuste de aquellos modelos con $RMR < 0.05$ y rechazando aquellos donde $RMR > 0.08$, si bien se insiste en que estos valores son orientativos.

2.2.5. Reespecificación del modelo y obtención de conclusiones

Es frecuente que el modelo propuesto no sea el que mejor ajusta. En este caso, se suelen buscar métodos para mejorar el ajuste conseguido, procediendo así con la reespecificación del modelo, añadiendo o eliminando parámetros a los estimados en el modelo original. No obstante, hay que tener en cuenta que estas modificaciones se deben realizar con precaución, y siempre primando las justificaciones teóricas por encima de las más deseables en cuanto a resultados. Por ello, es aconsejable, por ejemplo, realizar una validación cruzada estimando sobre un conjunto de datos distinto antes de aceptar el modelo modificado.

Para realizar la reespecificación, se deben analizar los ínidces de modificación, suyo valor representa la reducción que se produce en el estadístico χ^2 al estimar un coeficiente determinado. Hair et al. (2001) afirmaron que un valor superior a 3.84 supone una reducción significativa en el estadístico χ^2. Otro procedimiento consiste en obtener la matriz residual de la matriz de predicciones, en busca de residuos superiores a 2.58 (Hair et al., 2001), valor a partir del cual se pueden considerar estadísticamente significativos, lo que podría suponer un error de predicción sustancial.

Finalmente, se han de analizar los resultados obtenidos a partir del modelo seleccionado, validando o rechazando las hipótesis consideradas y finalizando así la investigación con la obtención de las conclusiones oportunas, tanto a nivel matemático, como en términos del problema real que se esté considerando.

Capítulo 3

Casos prácticos

En este capítulo se tratará de mostrar algunos ejemplos de aplicación de la modelización mediante ecuaciones estructurales a través de casos prácticos desarrollados en el software libre R. Para ello, se hará uso principalmente de la librería *lavaan (latent variables analysis)*, por ser la que cuenta con un uso más extendido en la actualidad. Además, el uso de esta librería se expone en (Rosseel, 2022), que es el manual que se ha seguido para la realización de los casos prácticos que se desarrollarán a continuación.

Antes de proceder con la exposición de los casos prácticos, se deben considerar dos aspectos en relación con la mencionada librería *lavaan*. En primer lugar, se ha de tener en cuenta que esta librería ofrece unas posibilidades de uso mucho más extensas de las mostradas en este trabajo, al permitir la realización de otras técnicas, tales como la construcción de modelos de análisis factorial, o modelos de *growth curves*, o curvas de crecimiento. En segundo lugar, para poder comprender cómo se ha llevado a cabo la construcción de los correspondientes modelos, se debe hacer una breve explicación de la sintaxis que emplea la propia librería para la especificación de los mismos.

Así, en función de si se desea especificar una regresión, la medición de una variable latente, o una correlación entre variables, se empleará una simbología u otra, como se describe en la Tabla (3.1)

Tabla 3.1: Nomenclatura librería *lavaan*

Tipo de fórmula	Operador
Medición de variable latente	=~
Regresión	~
(Co)Varianza / (Co)Varianza residual	~~
Intercept	~ 1

3.1. Conjunto de datos *PoliticalDemocracy*

En este primer caso práctico se va a analizar el conjunto de datos *PoliticalDemocracy*, perteneciente a la librería *lavaan*. Este conjunto de datos es conocido por ser el empleado por Kenneth A. Bollen para ilustrar la construcción de los SEM en (Bollen, 1989). El

conjunto de datos está formado por 75 observaciones de 11 variables, referentes a aspectos relacionados con la democracia y la industrialización en países en vías de desarrollo. Al tratarse de unos datos proporcionados por la propia librería, no han sido necesarias labores de limpieza o depuración de los datos, más allá de asignar nombre a las correspondientes variables, que se detallan a continuación:

Variable Latente	Variable Observada	Nombre
Industrialización en el año 1960 (*IND60*)	x1 - Producto Nacional Bruto per cápita en 1960	*GNP60*
	x2 - Consumo de energía per cápita en 1960	*energia60*
	x3 - Porcentaje de "fuerza laboral" de la industria en 1960	*porc-ind60*
Fuerza de la democracia en 1960 (*DEM60*)	y1 - Valoración de los expertos de la libertad de prensa en 1960	*libpren60*
	y2 - Libertad de la oposición política en 1960	*libopos60*
	y3 - "Limpieza" de las elecciones en 1960	*elecc60*
	y4 - "Efectividad del gobierno electo en 1960	*efect60*
Fuerza de la democracia en 1965 (*DEM65*)	y5 - Valoración de los expertos de la libertad de prensa en 1965	*libpren65*
	y6 - Libertad de la oposición política en 1965	*libopos65*
	y7 - "Limpieza" de las elecciones en 1965	*elecc65*
	y8 - "Efectividad del gobierno electo en 1965	*efect65*

Tabla 3.2: Variables latentes e indicadores bajo estudio. PoliticalDemocracy

3.1.1. Análisis Descriptivo

Si hacemos un primer resumen descriptivo de los datos, se observa que las variables observadas asociadas a los constructos *DEM60* y *DEM65* toman valores entre 0 y 10, salvo la variable *libpren60*, cuyo valor mínimo se sitúa en 1.25. Respecto a los indicadores asociados a la variable latente *IND60*, se observa que éstos toman valores más bajos que los anteriores, al no superarse en ningún caso el valor de 8. Estos valores vienen recogidos en la Tabla (3.3).

Tabla 3.3: Resumen descriptivo de las variables bajo estudio. PoliticalDemocracy

Variable	N	Media	Desviación Est.	Mín	Máx
libpren60	75	5.465	2.623	1.250	10.000
libopos60	75	4.256	3.947	0.000	10.000
elecc60	75	6.563	3.281	0.000	10.000
efect60	75	4.453	3.349	0.000	10.000
libpren65	75	5.136	2.613	0.000	10.000
libopos65	75	2.978	3.373	0.000	10.000
elecc65	75	6.196	3.286	0.000	10.000
efect65	75	4.043	3.246	0.000	10.000
GNP60	75	5.054	0.733	3.784	6.737
energia60	75	4.792	1.511	1.386	7.872
porc_ind60	75	3.558	1.406	1.002	6.425

No obstante, para obtener un mejor reflejo de la distribución media de los valores de cada una de las variables, se ha realizado la siguiente representación gráfica, en la que se muestran los valores medios de cada variable, agrupadas según el factor al que están asociados.

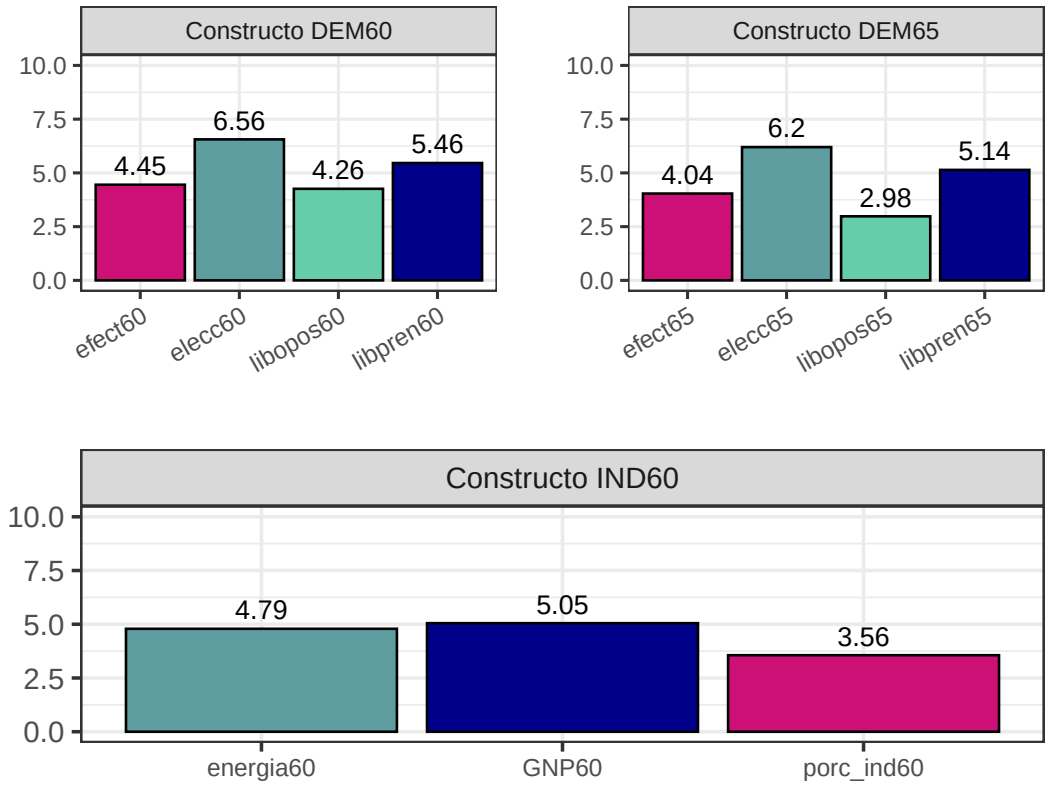

Figura 3.1: Distibución por medias muestrales de las variables bajo estudio. PoliticalDemocracy

Analizando la gráfica se observa que, en media, las variables que obtienen mejor y peor valoración en las mediciones realizadas en los años 1960 y 1965 son las que miden la "limpieza" de las elecciones y la libertad de la oposición, respectivamente. Si continuamos con la comparativa entre estos dos años, vemos que, en general, no se aprecian grandes cambios en las variables, salvo para la variable que mide la libertad de la oposición, en la que se observa una disminución de casi dos puntos desde la primera fecha a la segunda.

Finalmente, respecto a las variables asociadas al constructo *IND60*, se observa que el menor valor medio se obtiene para la variable *porc_ind60*. Sin embargo, esto no tiene por qué ser algo significativo en la comparación con el resto de variables, ya que tanto la variable *energia60* como la variable *GNP60*, son indicadores expresados en valor medio por habitante, mientras que la referida variable *porc_ind60* está expresada en forma de porcentaje.

3.1.2. Comprobación de hipótesis previas

En primer lugar, si atendemos al tamaño de la muestra con la que se está trabajando, vemos que este es de 75 registros. Teniendo en cuenta la regla propuesta por Catena et al. en 2003, un tamaño muestral adecuado sería de ocho registros por la suma del número de variables observadas (en este caso 11) y el número de variables latentes (en este caso 3), lo que equivaldría a un tamaño de 112 registros. Por otro lado, según la regla establecida por Marsh et al. a raíz del estudio publicado en 1998 basado en simulaciones realizadas a través del método de Montecarlo, resulta un tamaño muestral adecuado de unos 100 registros. Para nuestro conjunto de datos, en ninguno de los casos se alcanza el número sugerido, si bien se sabe que estos son valores orientativos, por lo que se continuará con el estudio de todos modos.

La segunda hipótesis a comprobar es la normalidad multivariante. Para ello, se realizan los tests de asimetría y curtosis propuestos por Mardia, arrojando los resultados disponibles en la tabla (3.4).

Tabla 3.4: Tests de Mardia sobre normalidad multivariante. PoliticalDemocracy

Prueba	Estadístico	p-valor	Normalidad
Mardia Skewness	344.4944	0.0101	NO
Mardia Kurtosis	-1.2217	0.2218	YES

Por tanto, de los dos tests propuestos por Mardia, únicamente acepta la normalidad multivariante el referente a los valores de curtosis. Además, si se realizan los test de Shapiro-Wilk (A.1) a cada una de las variables, se ve que únicamente se acepta la normalidad univariante para las variables *GNP60*, *energia60* y *porc_ind60*. Adicionalmente, se han representado los gráficos cuantil-cuantil (A.1) e histogramas (A.2) de todas las variables objeto de estudio, pudiéndose consultar estos resultados en el Apéndice A.

En consecuencia, es dudosa la presencia de normalidad multivariante en nuestros datos. La no satisfacción de esta hipótesis, condicionaría las estimaciones que se realicen posteriormente al no poder emplear el método de máxima verosimilitud, ya que como se vio en el capítulo 2, la normalidad multivariante es una condición necesaria para aplicar

esta técnica. No obstante, sería posible usar estimadores de máxima verosimilitud, si se hace uso de las cotas propuestas en (J. Curran, 1996), y cuyas condiciones se satisfacen si analizamos los datos descriptivos de las variables de la base de datos presentes en la tabla (3.5). Por tanto, consideramos satisfecha esta segunda hipótesis.

Tabla 3.5: Estadísticos descriptivos de las variables bajo estudio y niveles de asimetría y curtosis. PoliticalDemocracy

	n	Mean	Std.Dev	Median	Min	Max	25th	75th	Skew	Kurtosis
libpren60	75	5.4647	2.6227	5.4000	1.2500	10.0000	2.9000	7.5000	-0.0914	-1.1545
libopos60	75	4.2564	3.9471	3.3333	0.0000	10.0000	0.0000	8.2833	0.3187	-1.4673
elecc60	75	6.5631	3.2809	6.6667	0.0000	10.0000	3.7667	10.0000	-0.5940	-0.7191
efect60	75	4.4525	3.3495	3.3333	0.0000	10.0000	1.5815	6.6667	0.1177	-1.2127
libpren65	75	5.1363	2.6126	5.0000	0.0000	10.0000	3.6917	7.5000	-0.2280	-0.7782
libopos65	75	2.9781	3.3727	2.2333	0.0000	10.0000	0.0000	4.2069	0.8930	-0.4688
elecc65	75	6.1963	3.2862	6.6667	0.0000	10.0000	3.4777	10.0000	-0.5533	-0.7337
efect65	75	4.0434	3.2456	3.3333	0.0000	10.0000	1.3009	6.6667	0.4456	-0.9615
GNP60	75	5.0544	0.7329	5.0752	3.7842	6.7370	4.4773	5.5154	0.2539	-0.7539
energia60	75	4.7922	1.5107	4.9628	1.3863	7.8721	3.6632	5.8304	-0.3458	-0.5709
porc_ind60	75	3.5577	1.4057	3.5681	1.0017	6.4246	2.3002	4.5230	0.0838	-0.9361

Finalmente, para evaluar la consistencia interna de los datos, se ha calculado el índice Alpha de Cronbach, obteniendo un valor de 0.91, lo que supone una consistencia interna muy buena, por lo que se puede considerar que los datos son lo suficientemente fiables como para continuar con la construcción del modelo con las suficientes garantías.

3.1.3. Análisis de correlaciones

Como último paso previo a la construcción del modelo SEM, se van a estudiar las correlaciones entre las distintas variables observadas, para tratar de determinar si, a priori, tiene sentido la distribución de variables latentes que se ha expuesto anteriormente. Para ello, se ha construido la matriz de correlaciones muestral haciendo uso del coeficiente de correlación de Spearman. Al analizar dicha matriz (A.3), los resultados muestran que existe una correlación considerable entre los indicadores que definen las tres variables latentes presentes en el modelo (*DEM60*, *DEM65* e *IND60*). Dichas correlaciones son en todos los casos superiores a 0.45 para las variables observadas asociadas al primer constructo, 0.6 para las variables indicadoras del segundo constructo, y 0.78 para las variables que forman el tercer factor. Esto supone un buen punto de partida para construir el modelo. Sin embargo, correlaciones tan altas (sobre todo en el caso del último factor) podrían ser un indicio de multicolinealidad, si bien una de las ventajas de los SEM, es que generalmente son más "tolerantes" ante este problema que otros tipos de modelos como los de regresión.

A nivel general, se observa que no hay variables que no estén correlacionadas con el resto de variables, obteniendo para cada una de ellas coeficientes de correlación de al menos 0.2. Además, el determinante de la matriz de correlaciones muestral ha resultado ser de $6.42 \cdot 10^{-5}$, prácticamente nulo, lo cual significa que hay columnas y filas de dicha matriz entre las que existe una fuerte relación lineal. Esto indica que hay variables que están fuertemente relacionadas entre sí, lo que es otra buena señal para construir el modelo.

Finalmente, se ha realizado la prueba de esfericidad de Bartlett, obteniendo un p-valor $< 2.2 \cdot 10^{-16}$. Por tanto, se puede concluir que existe un subconjunto de variables interrelacionadas entre sí, por lo que tiene sentido aplicar técnicas multivariantes y, en particular, una modelización mediante ecuaciones estructurales.

3.1.4. Construcción y evaluación del modelo

Para llevar a cabo la construcción del modelo SEM, el primer paso es realizar su especificación. Para ello, se han especificado los correspondientes modelos de medida y estructural, describiendo las relaciones entre variables latentes y sus indicadores, y las relaciones entre las propias variables latentes, respectivamente. Asimismo, se han especificado las correspondientes covarianzas entre algunas de las variables, como por ejemplo la existente entre las variables que miden la libertad de prensa en los dos años estudiados, o entre las variables que miden la efectividad del gobierno elegido en dichos años. Cabe mencionar que un paso previo que podría realizarse, es validar la estructura de variables latentes que se ha construido mediante la aplicación de un análisis factorial confirmatorio. Esto se hace para garantizar que el modelo de medida que se está considerando es correcto. Sin embargo, dadas las limitaciones de este trabajo, se va a proceder a construir el modelo SEM directamente. A continuación se muestra el código correspondiente a la especificación y ajuste del modelo, el cual se ha llevado a cabo mediante la función *sem* del paquete *lavaan*.

```
modelo <- '
  # especificamos el modelo de medida
    IND60 =~ GNP60 + energia60 + porc_ind60
    DEM60 =~ libpren60 + libopos60 + elecc60 + efect60
    DEM65 =~ libpren65 + libopos65 + elecc65 + efect65
  # especificamos el modelo estructural
    DEM60 ~ IND60
    DEM65 ~ IND60 + DEM60
  # especificamos las correlaciones residuales
    libpren60 ~~ libpren65
    libopos60 ~~ efect60 + libopos65
    elecc60 ~~ elecc65
    efect60 ~~ efect65
    libopos65 ~~ efect65
'

ajuste <- sem(modelo, data = datos)
```

Una vez ajustado el modelo, se obtienen los siguientes resultados. En primer lugar, si evaluamos el ajuste proporcionado por el modelo, vemos que este es muy bueno para todos los índices de ajuste considerados. Estos índices has sido obtenidos mediante la función *fitMeasures* del paquete *lavaan* y pueden consultarse en la Tabla (3.6).

```
fitMeasures(ajuste, c("chisq", "df","cfi", "tli", "rmsea", "srmr"))
```

Tabla 3.6: Bondad de ajuste del modelo SEM. PoliticalDemocracy

chisq	df	CFI	TLI	RMSEA	RMSR
38.1252	35	0.9954	0.9927	0.0345	0.0444

Así, se observa que el ratio χ^2/df es de 1.0893, lo que supone un muy buen ajuste. De igual forma, tanto el índice de ajuste comparativo (CFI) como el índice de ajuste de Tucker y Lewis (TLI), toman valores muy cercanos a uno, lo que también es indicativo de la bondad de ajuste logrado por el modelo. Finalmente, los índices RMSEA y RMSR toman valores "bajos" y en ambos casos inferiores a 0.05, que es el valor máximo propuesto por Brown y Cudeck para considerar aceptable un modelo. Teniendo estos resultados en cuenta podemos concluir que el modelo ofrece un ajuste muy bueno.

Una vez ajustado el modelo, se puede obtener la matriz de covarianzas implicada asociada al mismo mediante la función *fitted*, como se mestra a continuación:

```
sigma_gorro <- fitted(ajuste)
```

Tabla 3.7: Matriz de covarianzas implicada del modelo. PoliticalDemocracy

	GNP60	energia60	porc_ind60	libpren60	libopos60	elecc60	efect60	libpren65	libopos65	elecc65	efect65
GNP60	0.5300	0.9778	0.8155	0.6650	0.8358	0.7034	0.8411	0.8135	0.9646	1.0409	1.0299
energia60	0.9778	2.2517	1.7781	1.4500	1.8223	1.5337	1.8340	1.7738	2.1032	2.2696	2.2455
porc_ind60	0.8155	1.7781	1.9497	1.2094	1.5199	1.2792	1.5296	1.4794	1.7541	1.8929	1.8728
libpren60	0.6650	1.4500	1.2094	6.8337	6.2112	5.2275	6.2509	5.1427	5.3582	5.7821	5.7208
libopos60	0.8358	1.8223	1.5199	6.2112	15.1788	6.5697	9.1689	5.6793	8.8867	7.2667	7.1896
elecc60	0.7034	1.5337	1.2792	5.2275	6.5697	10.5967	6.6117	4.7798	5.6674	6.9108	6.0510
efect60	0.8411	1.8340	1.5296	6.2509	9.1689	6.6117	11.0540	5.7156	6.7769	7.3132	7.5839
libpren65	0.8135	1.7738	1.4794	5.1427	5.6793	4.7798	5.7156	6.7730	5.2432	5.6581	5.5981
libopos65	0.9646	2.1032	1.7541	5.3582	8.8867	5.6674	6.7769	5.2432	11.1708	6.7088	7.9938
elecc65	1.0409	2.2696	1.8929	5.7821	7.2667	6.9108	7.3132	5.6581	6.7088	10.6710	7.1628
efect65	1.0299	2.2455	1.8728	5.7208	7.1896	6.0510	7.5839	5.5981	7.9938	7.1628	10.3410

Se dispone también de la función *lavInspect*, del paquete *lavaan*, que permite acceder, entre otra información, a las matrices que se han utilizado para la construcción del modelo, tales como las matrices Ψ, Θ, Λ, etc. Esta información se puede obtener con el siguiente código:

```
lavInspect(ajuste)
```

A continuación se va a proceder a representar gráficamente el *path diagram* del modelo, en el que se pueden observar las relaciones entre las distintas variables. Dicha representación se ha llevado a cabo mediante la función *semPaths* del paquete *semPlot*, en la que se han especificado además algunas cuestiones de estilo como se aprecia en el siguiente código:

```
library(semPlot)
semPaths(ajuste, whatLabels = "std",style="lisrel",
         layout="tree2",
         reorder=FALSE, optimizeLatRes=TRUE, edge.label.position=.5,
         edge.label.cex = 0.8, edge.color = "darkblue" )
```

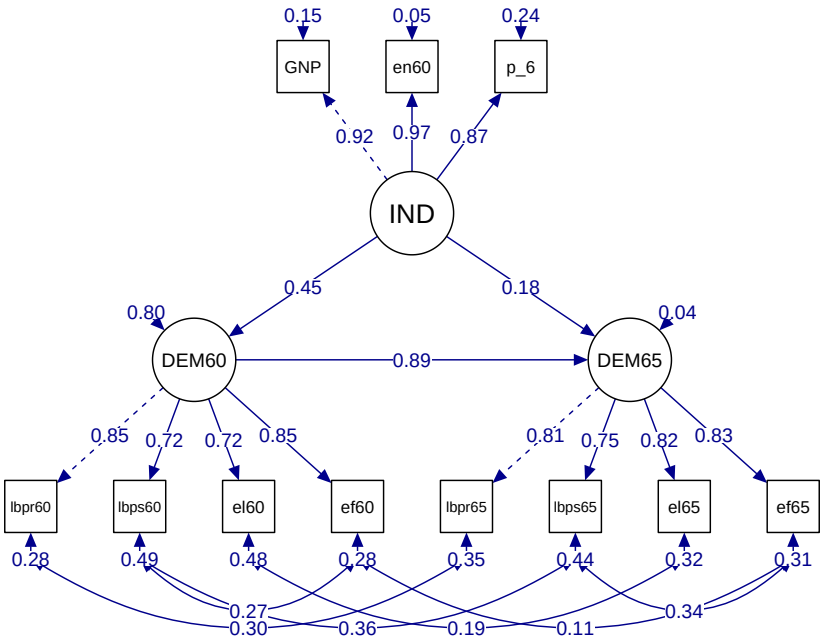

Figura 3.2: Path Diagram del modelo estudiado. PoliticalDemocracy

Así, analizando la Figura (3.2) a simple vista se aprecia que, a partir de las estimaciones estandarizadas presentes en el gráfico, la mayoría de las relaciones entre variables propuestas han resultado ser significativas, con la excepción de los parámetros correspondientes a la covarinanza entre las variables *elecc60* y *elecc65*, y la covarianza entre las variables *efect60* y *efect65*, cuyos coeficientes estimados estandarizados han resultado ser de 0.19 y 0.11, respectivamente.

Para confirmar esto, en la Tabla (3.8) se pueden consultar las estimaciones[1] y estimaciones estándarizadas para todas las relaciones entre variables propuestas, así como los errores estándar y los correspondientes p-valores. Esta tabla ha sido obtenida mediante la función *parameterEstimates* del paquete *lavaan*, como se muestra en el siguiente código:

```
parameterEstimates(ajuste, standardized = T)
```

[1]Todas las estimaciones se han obtenido empleando el Estimador de Máxima Verosimilitud

Tabla 3.8: Relaciones estimadas y estandarizadas. PoliticalDemocracy

Relación	Estimación	Estim. Estandarizada	Error Estándar	p-valor
IND60=~GNP60	1.0000	0.9199	0.0000	NA
IND60=~energia60	2.1804	0.9730	0.1385	0.0000
IND60=~porc_ind60	1.8185	0.8721	0.1520	0.0000
DEM60=~libpren60	1.0000	0.8504	0.0000	NA
DEM60=~libopos60	1.2567	0.7171	0.1824	0.0000
DEM60=~elecc60	1.0577	0.7223	0.1514	0.0000
DEM60=~efect60	1.2648	0.8457	0.1450	0.0000
DEM65=~libpren65	1.0000	0.8080	0.0000	NA
DEM65=~libopos65	1.1857	0.7460	0.1688	0.0000
DEM65=~elecc65	1.2795	0.8237	0.1599	0.0000
DEM65=~efect65	1.2659	0.8278	0.1581	0.0000
DEM60~IND60	1.4830	0.4467	0.3991	0.0002
DEM65~IND60	0.5723	0.1823	0.2213	0.0097
DEM65~DEM60	0.8373	0.8852	0.0984	0.0000
libpren60~~libpren65	0.6237	0.2958	0.3583	0.0818
libopos60~~efect60	1.3131	0.2726	0.7020	0.0614
libopos60~~libopos65	2.1529	0.3562	0.7338	0.0033
elecc60~~elecc65	0.7950	0.1906	0.6077	0.1908
efect60~~efect65	0.3482	0.1088	0.4422	0.4310
libopos65~~efect65	1.3562	0.3378	0.5683	0.0170
GNP60~~GNP60	0.0815	0.1539	0.0195	0.0000
energia60~~energia60	0.1198	0.0532	0.0697	0.0857
porc_ind60~~porc_ind60	0.4667	0.2394	0.0902	0.0000
libpren60~~libpren60	1.8914	0.2768	0.4444	0.0000
libopos60~~libopos60	7.3729	0.4857	1.3739	0.0000
elecc60~~elecc60	5.0675	0.4782	0.9517	0.0000
efect60~~efect60	3.1479	0.2848	0.7388	0.0000
libpren65~~libpren65	2.3510	0.3471	0.4802	0.0000
libopos65~~libopos65	4.9540	0.4435	0.9142	0.0000
elecc65~~elecc65	3.4314	0.3216	0.7128	0.0000
efect65~~efect65	3.2541	0.3147	0.6946	0.0000
IND60~~IND60	0.4484	1.0000	0.0867	0.0000
DEM60~~DEM60	3.9560	0.8004	0.9212	0.0000
DEM65~~DEM65	0.1725	0.0390	0.2148	0.4220

De esta forma, se confirma lo que se intuía a partir de las estimaciones estandarizadas observadas en la Figura (3.2), donde la mayoría de las relaciones propuestas aparentaban ser estadísticamente significativas. En la Tabla (3.8) aparecen asimismo resaltadas en rojo, aquellas relaciones en las que el correspondiente p-valor, indica que las mismas deben ser rechazadas como significativas, con un nivel de significación del 5 %. En todos los casos, estas relaciones corresponden a varianzas y covarianzas entre variables observadas.

Al analizar tabla anterior, se observa también que hay algunos valores de la columna correspondiente a los p-valores que aparecen ausentes, tomando el valor "NA" (valor establecido por R para referirse a valores perdidos o desconocidos). Esto es debido a las expresiones que utiliza la función *parameterEstimates* para calcular dichos p-valores. Así, estos son calculados mediante la expresión p-valor $= 2 \cdot (1 - \mathbb{P}(Z \leq |z|))$, siendo $Z \sim N(0, 1)$ y z el estadístico dado por la expresión $z = \text{est}/\text{se}$, donde est es la estimación del parámetro correspondiente y se el respectivo error estándar. Por tanto, en los casos en los que dicho error toma el valor 0, no es posible calcular el estadístico z, por lo que tampoco se puede calcular el correspondiente p-valor.

Si nos centramos en las relaciones entre variables latentes, vemos que todas ellas han resultado ser estadísticamente significativas. Esto permite sacar algunas conclusiones interesantes en términos del problema que se está tratando. En primer lugar, podemos decir que existe una relación causal significativa entre la "fuerza" de la democracia en el año 1965 en los países estudiados y la "fuerza" de la democracia que había en esos mismos países en el año 1960. De igual forma, existe una relación causal significativa entre la "fuerza" de democracia en el año 1965 y el nivel de industrialización que existía cinco años antes en los países bajo estudio. Estas relaciones aparecen detalladas en la Tabla (3.9):

Tabla 3.9: Relaciones estimadas y estandarizadas entre las variables latentes del modelo. PoliticalDemocracy

Relación	Estimación	Estim. Estandarizada	Error Estándar	p-valor
DEM60~IND60	1.4830	0.4467	0.3991	0.0002
DEM65~IND60	0.5723	0.1823	0.2213	0.0097
DEM65~DEM60	0.8373	0.8852	0.0984	0.0000

Por tanto, se puede afirmar que no existen evidencias en contra de la hipótesis de que el "nivel" de democracia que existe en un país, está fuertemente condicionado por el "nivel" de democracia que existía en dicho país en el pasado, así como por la industrialización que atravesaba el país en los años previos.

Finalmente, se puede concluir que existe una relación causal significativa entre la "fuerza" de la democracia en un país y el nivel de industrialización que atraviesa dicho país en ese mismo momento. Esta conclusión podría ser de utilidad para los dirigentes de un país en el que se considere que se debe reforzar la "fuerza" o el nivel de la democracia, de manera que invertir en industria podría ser una buena forma de conseguir dicho objetivo.

3.2. Conjunto de datos *Demo.twolevel*

Este segundo caso práctico nace con el objetivo de mostrar cómo se puede construir un modelo de ecuaciones estructurales cuando los datos con los que se está trabajando están agrupados en *clusters*, es decir, si tenemos un conjunto de datos con N observaciones, dichas observaciones estarán diferenciadas en k grupos, de modo que cada grupo contendrá un número n_i de observaciones. En estos casos, la librería *lavaan*, permite la construcción de modelos SEM multinivel, si bien la implementación de este tipo de modelos es

relativamente reciente y solo están recogidos los modelos con dos niveles. De esta forma, a la hora de especificar el modelo, se deben definir ambos niveles del siguiente modo: un primer nivel en el que se especificarán las relaciones entre variables *intraclase*, esto es, dentro del mismo *cluster*, y un segundo nivel con las relaciones *interclase*, es decir, entre los distintos grupos. Las especificaciones de cada uno de los niveles se llevan a cabo de forma totalmente análoga al caso de un solo nivel.

Para ilustrar el uso de este caso particular de SEM, se ha escogido el conjunto de datos *Demo.twolevel*, perteneciente a la librería *lavaan*. Este conjunto de datos está construido únicamente con el fin de mostrar el uso práctico de los modelos SEM multinivel por lo que, al contrario que en el primer caso práctico, los datos no corresponden a un problema real. Aún así, a continuación se describirán las variables que forman parte del conjunto de datos.

Los datos están formados por 2500 observaciones de 12 variables, una de las cuales corresponde al grupo o *cluster* al que pertenece cada observación. El resto de variables son 6 items (*y1 - y6*), tres covariables *x1 - x3 intraclase*, es decir, medidas dentro de un mismo grupo, y dos covariables *w1, w2 interclase*, es decir, medidas entre los distintos grupos. Estas dos últimas toman el mismo valor para todas las observaciones dentro de un mismo grupo.

3.2.1. Análisis descriptivo

En primer lugar, analizando los valores de la base de datos, vemos que todas las variables toman valores tanto positivos como negativos, a excepción lógicamente de la variable *cluster* que define los distintos grupos en los que están divididos los datos. Se ve cómo los ítems *y1 - y6*, toman valores aproximadamente entre -6 y 6, mientras que el resto de variables toman valores aproximados dentro del intervalo (-3, 3). La variable *cluster*, como se ha mencionado antes, define los 200 grupos en los que se dividen las observaciones, los cuales toman un tamaño de 5, 10, 15 o 20 observaciones. Podemos encontrar un primer resumen descriptivo de las variables bajo estudio en la Tabla (3.10)

Tabla 3.10: Resumen descriptivo de las variables del dataset *Demo.twolevel*

Variable	N	Media	Desviación Est.	Mín	Máx
y1	2,500	0.025	1.729	−7.950	5.991
y2	2,500	−0.024	1.505	−5.170	6.152
y3	2,500	−0.024	1.421	−6.002	5.915
y4	2,500	0.064	1.544	−5.426	5.385
y5	2,500	0.078	1.403	−5.499	5.137
y6	2,500	0.012	1.356	−6.613	5.482
x1	2,500	−0.007	0.991	−3.588	3.134
x2	2,500	−0.003	1.006	−3.959	3.676
x3	2,500	0.020	1.022	−3.586	3.861
w1	2,500	0.020	0.959	−2.410	2.474
w2	2,500	−0.091	0.967	−3.017	2.196
cluster	2,500	101.000	57.744	1	200

A continuación, para obtener una mejor visión de la distribución de los valores medios de las variables, se muestra una representación gráfica de los mismos, separados en distintos gráficos según la tipología de dichas variables.

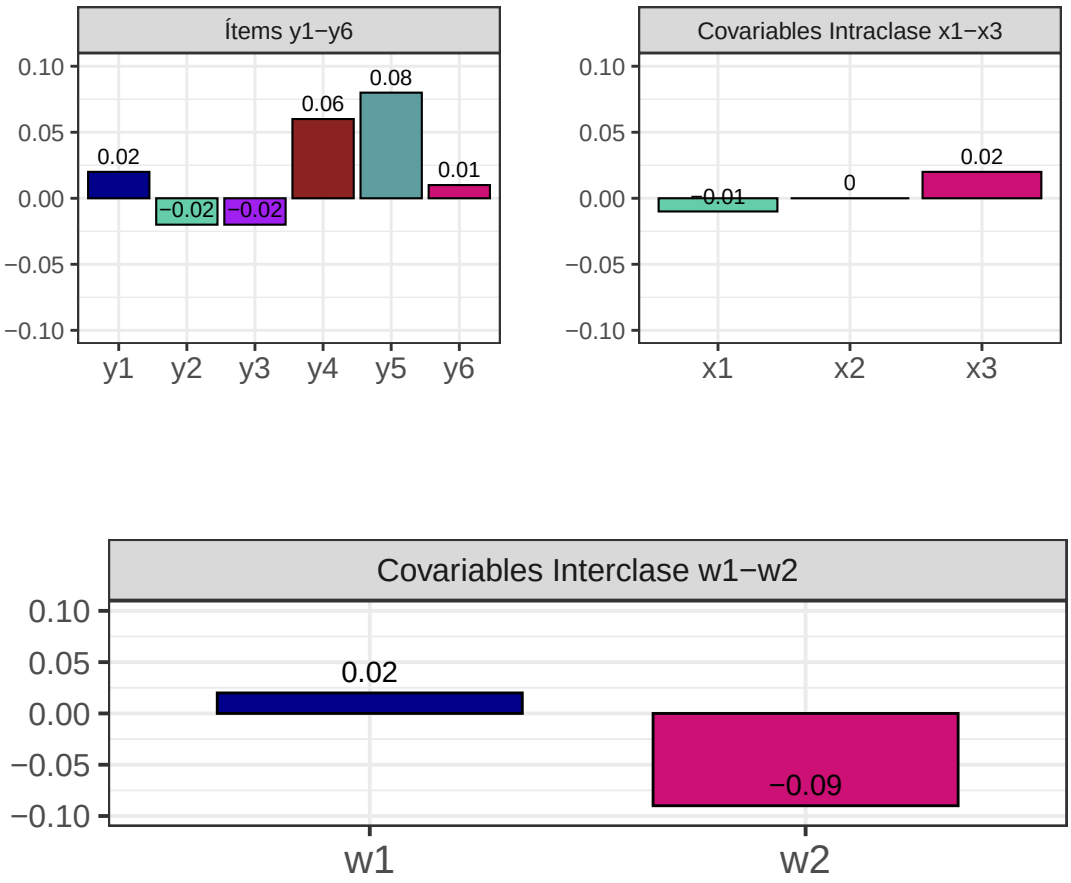

Figura 3.3: Distibución por medias muestrales de las variables bajo estudio. Demo.twolevel

Observando los gráficos anteriores se ve como las variables *y2, y3, x1* y *w1* toman valores medios negativos, mientras que el resto de variables tienen medias positivas. Llama también la atención que las tres variables "intraclase" (*x1, x2, x3*) tienen valores medios bastante cercanos a cero. Finalmente, comentar cómo las dos variables medidas entre los distintos grupos (*w1, w2*), toman valores medios de distinto signo, y se observa una cierta diferencia entre ambos valores medios. No obstante, al no saber con exactitud la naturaleza de las variables, es difícil determinar cuál puede ser el origen de esta diferencia.

3.2.2. Comprobación de hipótesis previas

En primer lugar, respecto al tamaño de la muestra, vemos que esta es de 2500 observaciones. Según la regla propuesta por Catena et al. en 2003, un tamaño muestral adecuado sería de ocho registros por la suma del número de variables observadas (en este caso 12) y el número de variables latentes (en este caso 2), lo que equivaldría a un tamaño de

112 registros. Por otro lado, si atendemos a la regla establecida por Marsh et al. a raíz del estudio publicado en 1998, basado en simulaciones realizadas a través del método de Montecarlo, resulta un tamaño muestral adecuado de unos 200 registros. En el caso de nuestro conjunto de datos, se ve que para ambos criterios se supera el tamaño muestral sugerido, por lo que podemos dar esta primera hipótesis por satisfecha.

La segunda hipótesis a comprobar es la normalidad multivariante. Para ello, se han realizado los tests de asimetría y curtosis propuestos por Mardia, cuyos resultados se pueden consultar en la Tabla (3.11)

Tabla 3.11: Tests de Mardia sobre normalidad multivariante. Demo.twolevel

Prueba	Estadístico	p-valor	Normalidad
Mardia Skewness	679.6170	0.0000	NO
Mardia Kurtosis	0.1603	0.8726	YES

Es decir, de los dos tests propuestos por Mardia, únicamente se acepta la normalidad multivariante en el correspondiente a los valores de curtosis. Se han realizado también los respectivos tests de Shapiro-Wilk (A.2) a cada variable por separado, aceptándose la normalidad univariante únicamente para los variables *y1, y2, y5, x1, x2* y *x3*. Finalmente, se han representado los histogramas (A.5) y gráficos cuantil-cuantil (A.4) para todas las variables de la base de datos. Todos estos resultados se pueden consultar en el apéndice A.

Por tanto, ante la dudosa presencia de normalidad multivariante en el conjunto de nuestros datos, y con el objetivo de poder emplear estimadores de máxima verosimilitud en pasos posteriores, se puede hacer uso de las cotas propuestas en (J. Curran, 1996), y cuya satisfacción se cumple según los resultados que se muestran en la Tabla (3.12). En consecuencia, damos por verificada esta segunda hipótesis.

Tabla 3.12: Estadísticos descriptivos de las variables bajo estudio y niveles de asimetría y curtosis Demo.twolevel

	n	Mean	Std.Dev	Median	Min	Max	25th	75th	Skew	Kurtosis
y1	2500	0.0247	1.7289	0.0269	-7.9504	5.9906	-1.1717	1.2090	-0.0046	-0.0031
y2	2500	-0.0242	1.5051	-0.0382	-5.1695	6.1518	-1.0275	0.9748	0.0275	-0.0156
y3	2500	-0.0236	1.4208	-0.0165	-6.0025	5.9146	-0.9464	0.8743	0.0302	0.4409
y4	2500	0.0641	1.5442	0.1088	-5.4263	5.3852	-0.9278	1.1214	-0.2000	0.1507
y5	2500	0.0779	1.4026	0.1004	-5.4988	5.1368	-0.8312	1.0038	-0.0492	0.0853
y6	2500	0.0123	1.3557	0.0177	-6.6129	5.4824	-0.8500	0.9161	-0.0764	0.3776
x1	2500	-0.0074	0.9913	0.0009	-3.5882	3.1337	-0.6513	0.6359	-0.0139	0.0892
x2	2500	-0.0029	1.0059	0.0264	-3.9594	3.6757	-0.6917	0.6479	0.0045	0.0881
x3	2500	0.0200	1.0225	0.0245	-3.5863	3.8609	-0.6883	0.6937	0.0176	-0.0740
w1	2500	0.0198	0.9588	0.0044	-2.4099	2.4741	-0.6142	0.7747	-0.1859	-0.4466
w2	2500	-0.0913	0.9665	-0.0817	-3.0175	2.1958	-0.8497	0.5624	0.0088	-0.2895
cluster	2500	101.0000	57.7437	100.5000	1.0000	200.0000	51.0000	151.0000	0.0000	-1.2017

3.2.3. Análisis de correlaciones

Como último paso previo a la construcción del modelo SEM multinivel, se han estudiado las correlaciones muestrales entre las variables observadas, con el fin de determinar, a priori, la posible estructura de variables latentes que se detallará posteriormente en la especificación del modelo. Para ello, se ha construido la matriz de correlaciones haciendo uso del coeficiente de correlación de Spearman (A.6). Analizando dicha matriz, vemos que existe una correlación significativa (al menos superior a 0.3) entre las variables *y1, y2 e y3*, del mismo modo que entre las variables *y4, y5 e y6*. Existe también cierta correlación, aunque con un menor grado de significancia, entre las variables *x1 y x2*, y las variables *y1, y2 e y3*. La existencia de correlación significativa entre algunas de las variables de la base de datos es un buen punto de partida para la construcción del modelo. Sin embargo, hay otras variables, como *x3, w1 y w2* que no parecen estar correlacionadas con ninguna otra variable de la base de datos. No obstante, como se ve en (Bollen, 1989), la no existencia de correlación no implica la ausencia de causalidad, por lo que mantendremos dichas variables en el estudio de todos modos.

Además, se ha calculado el determinante de la matriz de correlaciones, obteniendo un resultado de 0.2206. Lo ideal hubiera sido obtener un valor lo más cercano posible a 0, lo cual sería indicativo de que hay columnas y filas de la matriz de correlación entre las que existe una fuerte relación lineal, lo que sería otra buena señal de cara a la construcción del modelo. Aun así, se continuará con el estudio.

Finalmente, se ha realizado la prueba de esfericidad de Bartlett, obteniendo un p-valor $< 2.2 \cdot 10^{-16}$, por lo que podemos afirmar que existe un subconjunto de variables que están fuertemente interrelacionadas entre sí, por lo que tiene sentido aplicar técnicas multivariantes, y en particular, un modelo SEM multinivel.

3.2.4. Construcción y evaluación del modelo

Para la construcción de este modelo SEM multinivel, se comienza como es habitual por su especificación. Así, se han definido las respectivas relaciones entre variables para cada uno de los dos niveles comentados anteriormente. Tras el correspondiente análisis factorial exploratorio y su posterior validación mediante un análisis factorial confirmatorio se ha determinado una estructura de variables latentes formada por un factor *FW*, medido por los ítems *y1, y2 e y3* para el nivel 1 (intraclase), y un factor *FB* medido por los mismos ítems para el nivel 2. Asimismo, se han especificado las relaciones de regresión entre el factor *FW* y las covariables *x1, x2 y x3*, y entre el factor *FB* y las covariables *w1 y w2*. Como se mencionó anteriormente, la validación de los modelos de medida que se acaban de describir mediante los mencionados análisis factoriales no se muestra por escapar a los objetivos de este trabajo. Por este motivo, se ha procedido a ajustar el modelo SEM directamente. A continuación se muestra el código correspondiente al ajuste del modelo:

```
modelo2 <- '
    level: 1
        FW =~ y1 + y2 + y3
        FW ~ x1 + x2 + x3
    level: 2
```

```
        FB =~ y1 + y2 + y3
        FB ~ w1 + w2
'

ajuste2 <- sem(model = modelo2, data = Demo.twolevel, cluster = "cluster")
```

Como se ve en el código anterior, al tratarse de un modelo SEM multinivel, se debe incluir el argumento *cluster* en la función *sem* con el nombre de la variable que indica el grupo al que pertenece cada observación. Una vez realizado el ajuste, para evaluar la bondad del mimsmo, se han obtenido los principales índices de ajuste, esto es, estadístico χ^2, índice de ajuste comparativo (CFI), índice de Tucker-Lewis (TLI), así como los índices RMSEA y RMSR, obteniendo los resultados que se muestran en la Tabla (3.13). Al igual que la práctica anterior, para esta tarea se ha empleado la función *fitMeasures* del paquete *lavaan*.

Tabla 3.13: Bondad de ajuste del modelo SEM multinivel

chisq	df	CFI	TLI	RMSEA	RMSR
8.092	10	1	1	0	0.0375

Analizando los resultados se ve que el ajuste proporcionado por el modelo es muy bueno, obteniendo un valor de $0.80 < 1$ para el cociente χ^2/df, y unos valores de 1 para los índices CFI y TLI, y de 0 para el índice RMSEA, lo que supone un ajuste inmejorable. Además, el índice RSMR toma un valor de 0.0375, inferior a la cota establecida por Brown y Cudeck para considerar aceptable un modelo (0.05). Teniendo todo lo anterior en cuenta, podemos concluir que el modelo proporciona un muy buen ajuste.

Además, se dispone de las funciones descritas en la práctica anterior para la extracción de información sobre el modelo. Así, se pueden obtener las correlaciones y medias *inter-clase* e *instraclase* mediante la función *fitted*. Se muestra abajo el código necesario para su obtención, así como la salida proporcionada por el programa:

```
fitted(ajuste2)
```

```
## $within
## $within$cov
##        y1     y2     y3     x1     x2     x3
## y1  2.000
## y2  0.785  1.674
## y3  0.744  0.576  1.557
## x1  0.500  0.387  0.367  0.982
## x2  0.414  0.321  0.304  0.001  1.011
## x3  0.215  0.166  0.157 -0.006  0.008  1.045
##
## $within$mean
##     y1     y2     y3     x1     x2     x3
```

```
## -0.001 -0.001 -0.001 -0.007 -0.003  0.020
##
##
## $cluster
## $cluster$cov
##        y1     y2     y3     w1     w2
## y1  0.990
## y2  0.669  0.599
## y3  0.548  0.392  0.470
## w1  0.126  0.091  0.074  0.870
## w2  0.101  0.072  0.059 -0.128  0.931
##
## $cluster$mean
##     y1     y2     y3     w1     w2
##  0.021 -0.018 -0.044  0.052 -0.091
```

Esta información se puede obtener también mediante la orden `lavInspect(ajuste2, "h1")`, empleando el argumento "h1" para indicar que muestre los resultados del modelo sin restricciones (saturado). Asimismo, se pueden obtener las matrices empleadas para la construcción de modelo mediante la propia función *lavInspect*.

A continuación, con el fin de ver gráficamente las relaciones entre variables, así como para hacernos una primera idea de si dichas relaciones pueden ser significativas o no, la Figura 3.4 muestra una representación del *path diagram* del modelo anterior. En él aparecen también las estimaciones estandarizadas de cada parámetro. En esta ocasión, y dada las dificultades ofrecidas por la función *semPaths* para representar diagramas causales para modelos SEM multinivel debido a lo reciente de su implementación, se ha empleado en su lugar la función *lavaanPlot* del paquete del mismo nombre, como se muestra en el siguiente código:

```
lavaanPlot::lavaanPlot(model = ajuste2,
                       node_options = list(shape = "box",
                                           fontname = "Helvetica"),
           edge_options = list(color = "blue"),
           coefs = T, stand = T, covs = T)
```

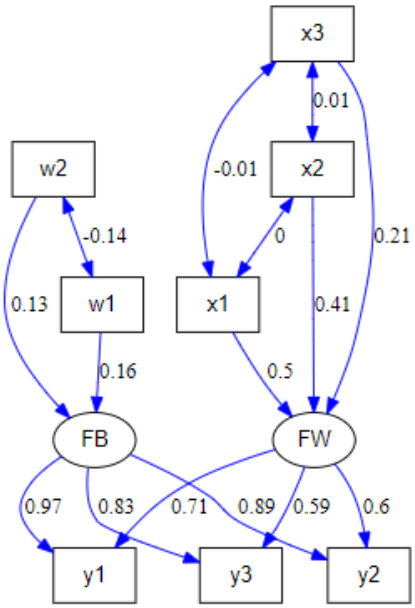

Figura 3.4: *Path Diagram* del modelo SEM multinivel

Observando el gráfico, parece que la mayor parte de las relaciones entre variables son significativas, a excepción de la formada por la regresión entre la variable latente *FB* y la covariable *w1*, así como las correspondientes a las covarianzas entre las variables *x1* y *x3*, entre las variables *x1* y *x2*, y entre las variables *x2* y *x3*.

Esto se puede confirmar si se observa la Tabla (3.14), obtenida de nuevo empleando la función *parameterEstimates* del paquete *lavaan*, y donde se recogen los parámetros estimados[2], así como las estimaciones estandarizadas, los errores estándar y los correspondientes p-valores.

Tabla 3.14: Relaciones estimadas y estandarizadas. Demo.twolevel

Relación	Estimación	Estim. Estandarizada	Error Estándar	p-valor
FW=~y1	1.0000	0.7121	0.0000	NA
FW=~y2	0.7742	0.6026	0.0342	0.0000
FW=~y3	0.7336	0.5921	0.0328	0.0000
FW~x1	0.5100	0.5019	0.0231	0.0000
FW~x2	0.4073	0.4068	0.0223	0.0000
FW~x3	0.2050	0.2081	0.0210	0.0000
y1~~y1	0.9859	0.4929	0.0457	0.0000
y2~~y2	1.0664	0.6369	0.0391	0.0000
y3~~y3	1.0110	0.6494	0.0366	0.0000
FW~~FW	0.5465	0.5389	0.0404	0.0000
x1~~x1	0.9823	1.0000	0.0000	NA
x1~~x2	0.0008	0.0008	0.0000	NA
x1~~x3	-0.0055	-0.0054	0.0000	NA

[2]Todas las estimaciones se han obtenido empleando el Estimador de Máxima Verosimilitud

Tabla 3.14: Relaciones estimadas y estandarizadas. Demo.twolevel (continuación)

Relación	Estimación	Estim. Estandarizada	Error Estándar	p-valor
x2~~x2	1.0115	1.0000	0.0000	NA
x2~~x3	0.0077	0.0075	0.0000	NA
x3~~x3	1.0450	1.0000	0.0000	NA
y1~1	0.0000	0.0000	0.0000	NA
y2~1	0.0000	0.0000	0.0000	NA
y3~1	0.0000	0.0000	0.0000	NA
x1~1	-0.0074	-0.0075	0.0000	NA
x2~1	-0.0029	-0.0029	0.0000	NA
x3~1	0.0200	0.0196	0.0000	NA
FW~1	0.0000	0.0000	0.0000	NA
FB=~y1	1.0000	0.9705	0.0000	NA
FB=~y2	0.7167	0.8942	0.0518	0.0000
FB=~y3	0.5869	0.8268	0.0476	0.0000
FB~w1	0.1647	0.1591	0.0787	0.0363
FB~w2	0.1310	0.1308	0.0764	0.0863
y1~~y1	0.0576	0.0581	0.0475	0.2251
y2~~y2	0.1201	0.2004	0.0314	0.0001
y3~~y3	0.1488	0.3165	0.0280	0.0000
FB~~FB	0.8989	0.9635	0.1184	0.0000
w1~~w1	0.8700	1.0000	0.0000	NA
w1~~w2	-0.1283	-0.1426	0.0000	NA
w2~~w2	0.9306	1.0000	0.0000	NA
y1~1	0.0245	0.0246	0.0748	0.7434
y2~1	-0.0161	-0.0209	0.0600	0.7879
y3~1	-0.0421	-0.0614	0.0542	0.4371
w1~1	0.0524	0.0562	0.0000	NA
w2~1	-0.0909	-0.0943	0.0000	NA
FB~1	0.0000	0.0000	0.0000	NA

Analizando la tabla se confirma lo que se intuía a partir del gráfico, es decir, la mayoría de las relaciones propuestas entre variables resultan ser estadísticamente significativas. Resaltadas en rojo aparecen aquellos parámetros que han resultado no significativos. Esto se ha determinado a partir del correspondiente p-valor, con un nivel de significación del 5 %. Así, se puede concluir, por ejemplo, que no existe una relación causal entre el factor *FB* y la variable *w2*. Tampoco han resultado significativos los parámetros correspondientes a la varianza de la variable *y1*, así como los interceptos asociados a las variables *y1*, *y2* e *y3*. Se observa también como, al igual que sucedía en el primer caso práctico, algunos de los valores de la columna que contiene los respectivos p-valores, aparecen sin ningún valor numérico, sino que aparece el valor "NA", empleado por R para referirse a datos perdidos o desconocidos. Esto sucede por haberse obtenido una estimación del error estándar (se) igual a 0, lo que impide el cálculo del estadístico z, y en consecuencia, de los p-valores asociados, como ya se detalló en el caso práctico anterior.

Antes de finalizar el estudio, se ha considerado interesante hacer un comentario sobre algunas diferencias que se han notado al ajustar ambos modelos (este último descrito y el del caso práctico anterior). En primer lugar, al calcular un resumen del modelo mediante la función *summary*, en ambos casos se obtiene información sobre el estimador empleado, así como el optimizador y el número de observaciones de la muestra. Sin embargo, en el caso del modelo SEM multinivel, se obtiene también información sobre el número de grupos en que están divididas las observaciones. Además, la información acerca de las estimaciones obtenidas está dividida en función de si se trata del nivel 1 (intra-grupos) o el nivel 2 (inter-grupos). Por último, tener en cuenta que, a diferencia del modelo construido para el primer caso práctico, en este caso las variables latentes, al pertenecer a niveles distintos, no interaccionan entre ellas, es decir, no existe relación causal alguna entre ambos factores.

Finalmente, hay que tener en cuenta que, al ser este un conjunto de datos ideado únicamente para ilustrar el funcionamiento en R de los SEM multinivel, el hecho de que no se trate de variables asociadas a un problema real, dificulta la extracción de conclusiones en términos de dicho problema, por lo que se concluye aquí el estudio.

Capítulo 4

Conclusiones

Este trabajo tenía como objetivo principal el estudio de los fundamentos de los SEM, así como mostrar ejemplos de aplicaciones prácticas de los mismos mediante librerías de R.

Para ello, se partió definiendo la hipótesis fundamental sobre la que se basan los SEM, esto es, minimizar la diferencia entre la matriz de varianzas y covarianzas poblacional, y la matriz de varianzas y covarianzas reproducida por los parámetros del modelo. También se ha presentado el marco histórico en el que surgen los SEM, así como el punto en el que se encuentra la investigación actual en torno a los mismos. Esto permite hacerse una idea de las líneas futuras de trabajo que se pueden tomar alrededor de este campo. Además, se ha hecho hincapié en el concepto de causalidad y sus diferencias con el de correlación, viéndose que el primero tiene una interpretación mucho más compleja que una "simple" correlación entre variables. Ha resultado importante también la presentación de los distintos tipo de SEM que existen, desde el más sencillo formado únicamente por variables observadas (*path analysis*), hasta el modelo general capaz de estudiar relaciones causales entre variables observadas y latentes, además de entre las propias variables latentes.

Ha quedado claro también que los SEM pueden ser una herramienta muy potente para validar teorías en campos como la sociología, si bien tienen algunas limitaciones, como su sensibilidad al tamaño de la muestra, o a la distribución seguida por los datos. En la construcción de los modelos propiamente dichos, ha resultado especialmente decisivo el realizar una buena especificación de los modelos estructural y de medida, pues de esta especificación depende en gran parte la validez de las estimaciones que se obtengan y, por tanto, de nuestro modelo.

En lo referente a las aplicaciones prácticas, en la primera de ellas se ha tratado un problema como la "fuerza" o expansión de la democracia y la industrialización en países en vías de desarrollo, llegándose a la conclusión de que el "nivel" de democracia en un país está fuertemente influenciado por el "nivel" de democracia que había en dicho país en el pasado, así como por el grado de industrialización que atraviesa dicho país. Como se comentó en el capítulo correspondiente a esta práctica, estas conclusiones pueden resultar de utilidad para dirigentes de países en los que se considere necesario reforzar la fuerza democrática, ya que invertir en industria podría ser un buen modo de alcanzar este objetivo.

Finalmente, en la segunda práctica se ha visto cómo el software R permite construir también modelos SEM en los casos en los que los datos tienen sus observaciones divididas

en grupos o *clusters*, empleando para este fin dos niveles: un nivel *intraclase*, es decir, dentro de cada grupo, y un nivel *interclase*, es decir, entre los distintos grupos.

Para finalizar, comentar que los modelos de ecuaciones estructurales, son mucho más extensos que lo abarcado en este trabajo, donde por limitación de extensión y objetivos, se han presentado únicamente los resultados más relevantes y, en el caso de las aplicaciones prácticas, los modelos que proporcionaban mejores estimaciones. Señalar por último que para la realización de este trabajo, se han empleado los conocimientos adquiridos a lo largo de todo el Grado en Estadística, de modo que su ejecución ha permitido ampliar los conocimientos alcanzados en algunas asignaturas como Análisis Multivariante, donde ya se introdujo un caso particular de los SEM, como es el análisis factorial.

Apéndice A

Apéndice: Hipótesis previas

A.1. Conjunto de datos *politicalDemocracy*

A.1.1. Normalidad univariante

Tabla A.1: Tests de Shapiro-Wilk a las variables del estudio. PoliticalDemocracy

Variable	Estadístico	p-valor	Normalidad
libpren60	0.9316	6e-04	No
libopos60	0.8311	<0.001	No
elecc60	0.8547	<0.001	No
efect60	0.8994	<0.001	No
libpren65	0.9598	0.0179	No
libopos65	0.8107	<0.001	No
elecc65	0.8715	<0.001	No
efect65	0.9006	<0.001	No
GNP60	0.9751	0.1454	Sí
energia60	0.9763	0.1715	Sí
porc_ind60	0.9734	0.1145	Sí

Gráficos Q-Q

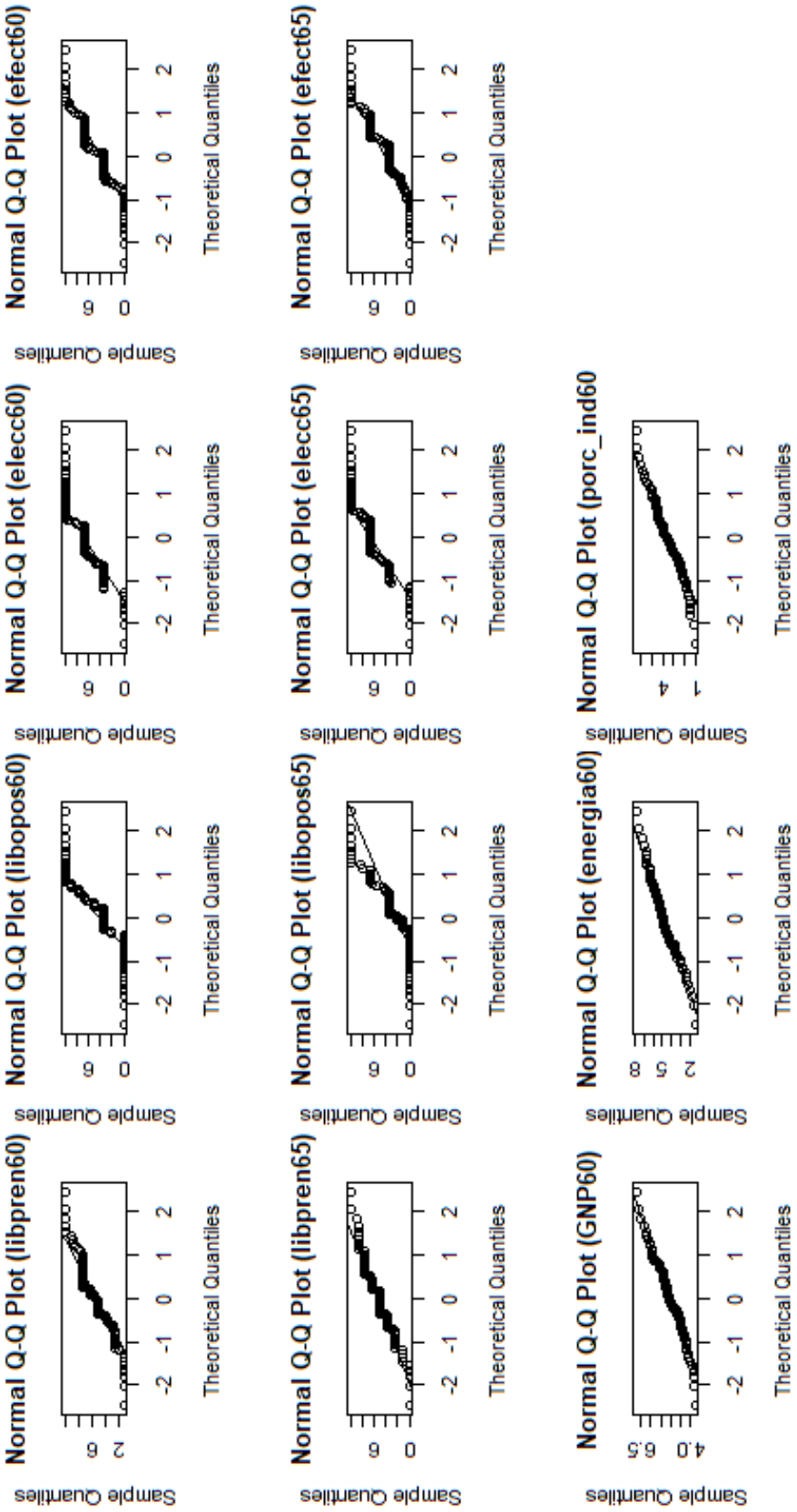

Figura A.1: Gráficos Cuantil-Cuantil de las variables bajo estudio. PoliticalDemocracy

Histogramas

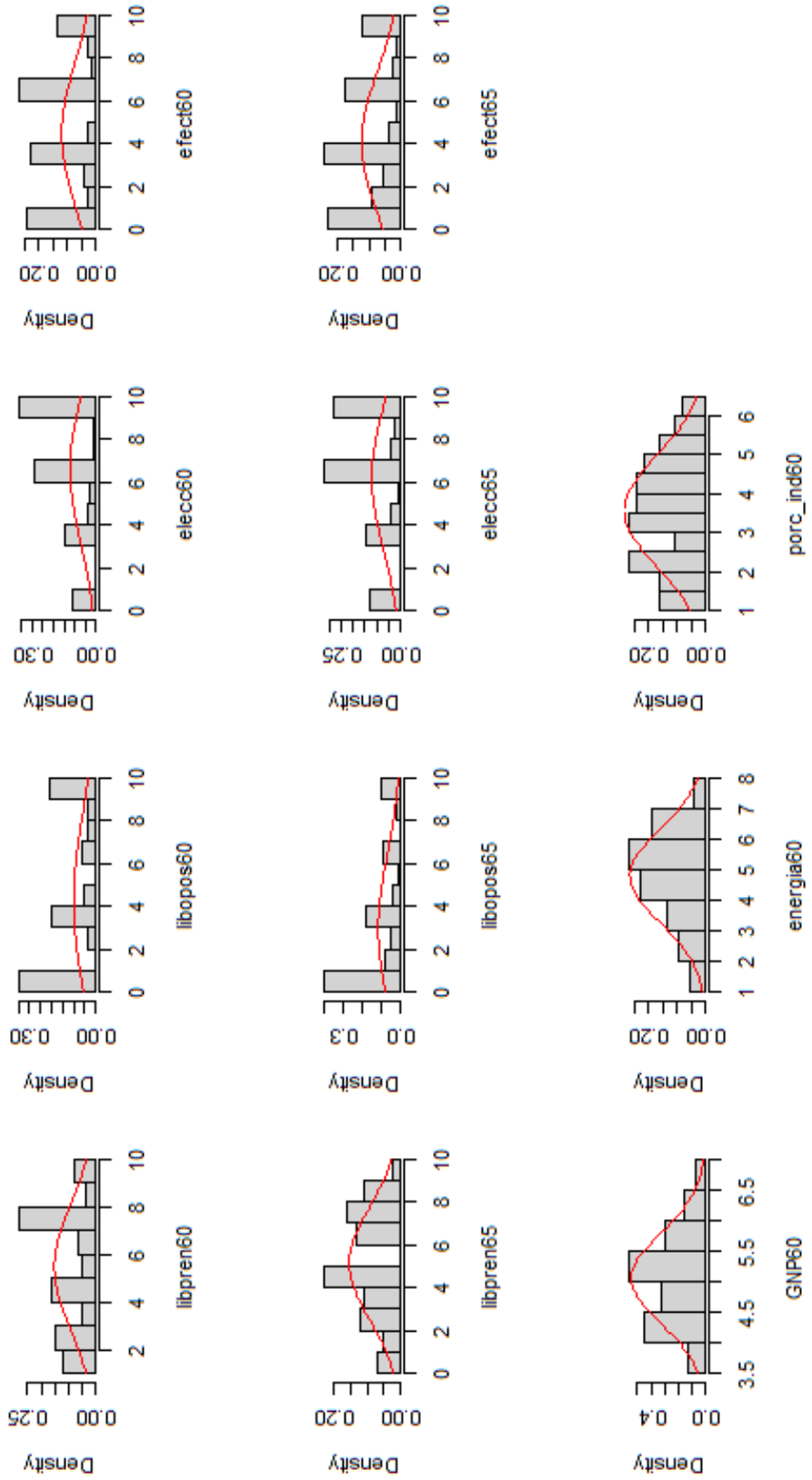

Figura A.2: Histogramas de las variables bajo estudio. PoliticalDemocracy

A.1.2. Matriz de correlaciones

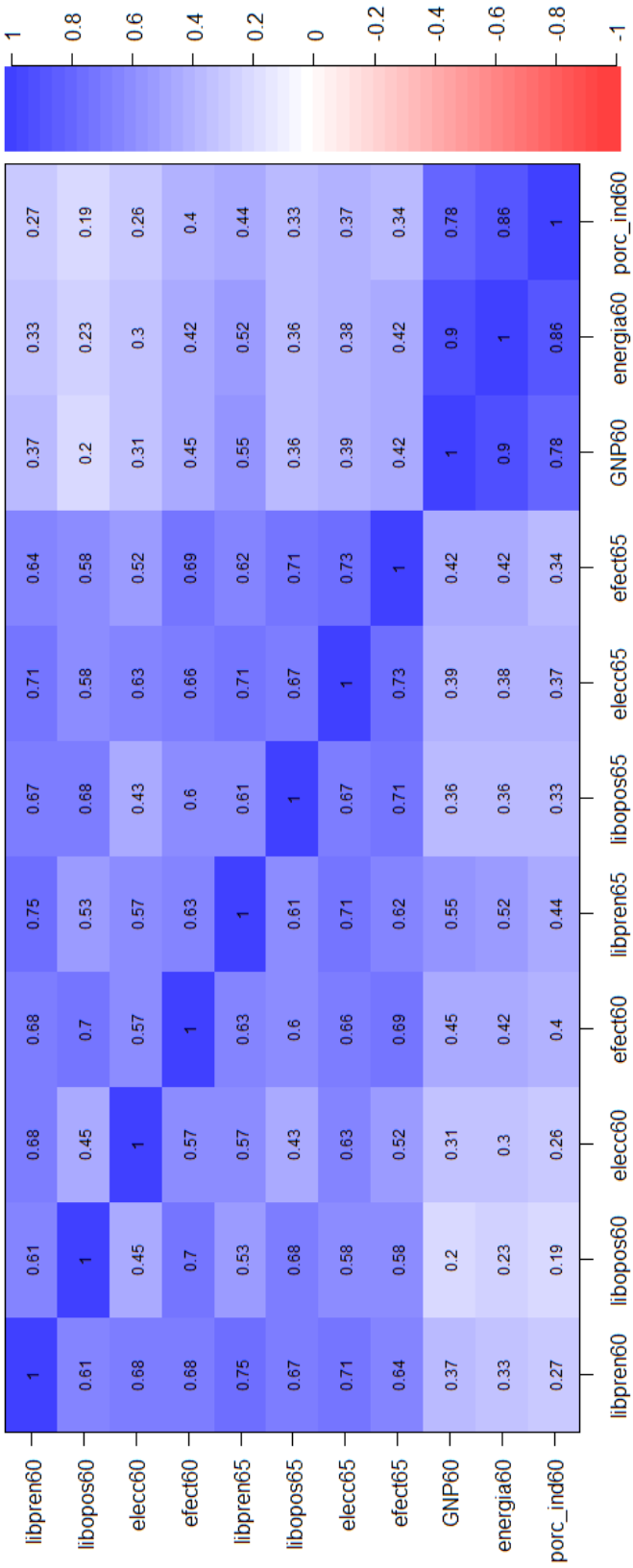

Figura A.3: Matriz de correlaciones. PoliticalDemocracy

A.2. Conjunto de datos *Demo.twolevel*

A.2.1. Normalidad univariante

Tabla A.2: Tests de Shapiro-Wilk a las variables del estudio. Demo.twolevel

Variable	Estadístico	p-valor	Normalidad
y1	0.9992	0.3667	Sí
y2	0.9997	0.9696	Sí
y3	0.9981	0.0051	No
y4	0.9972	2e-04	No
y5	0.9994	0.6226	Sí
y6	0.9984	0.0134	No
x1	0.9992	0.334	Sí
x2	0.9994	0.619	Sí
x3	0.9994	0.7138	Sí
w1	0.9892	<0.001	No
w2	0.9930	<0.001	No
cluster	0.9547	<0.001	No

Gráficos Q-Q

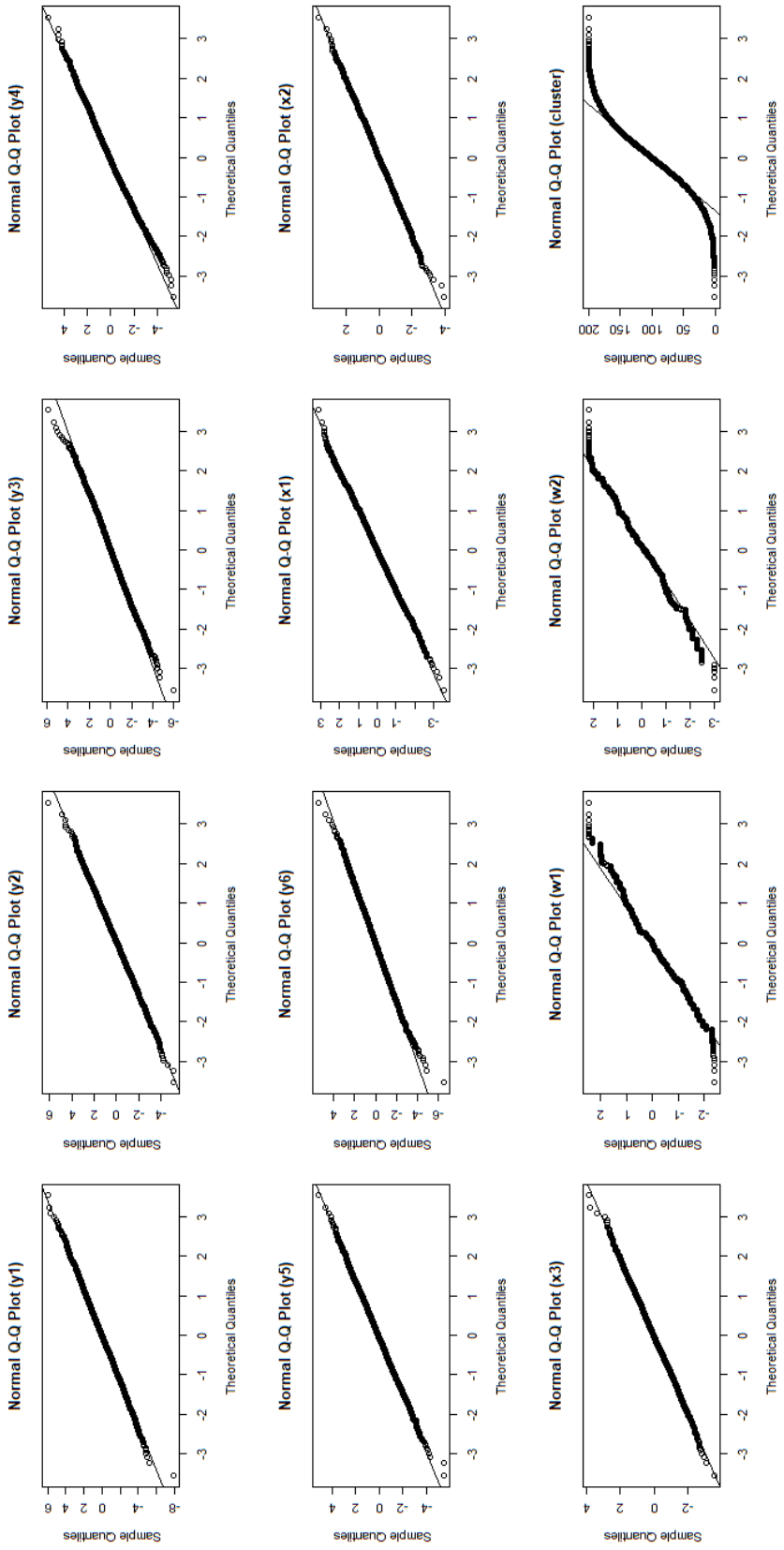

Figura A.4: Gráficos Cuantil-Cuantil de las variables bajo estudio. Demo.twolevel

Histogramas

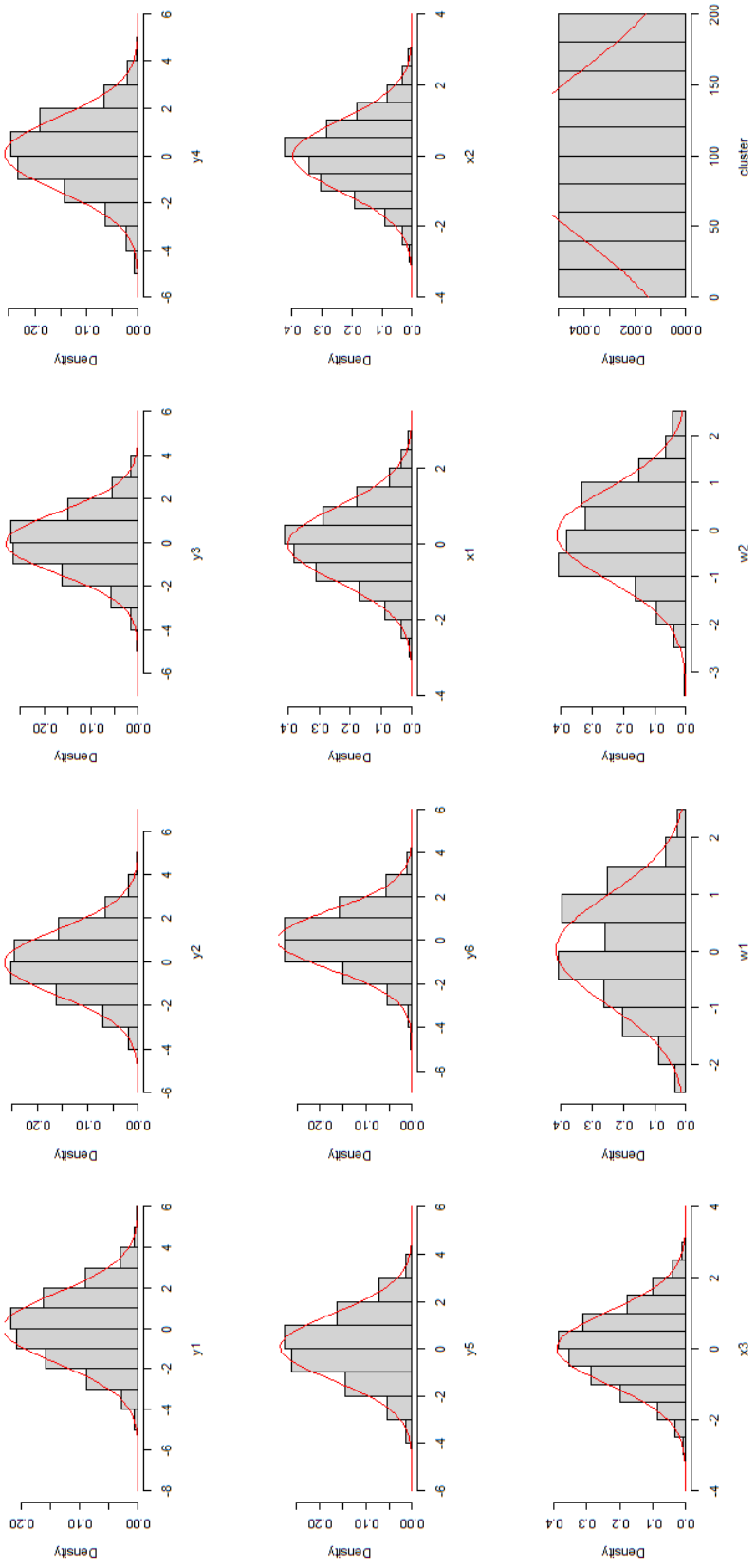

Figura A.5: Histogramas de las variables bajo estudio. Demo.twolevel

A.2.2. Matriz de correlaciones

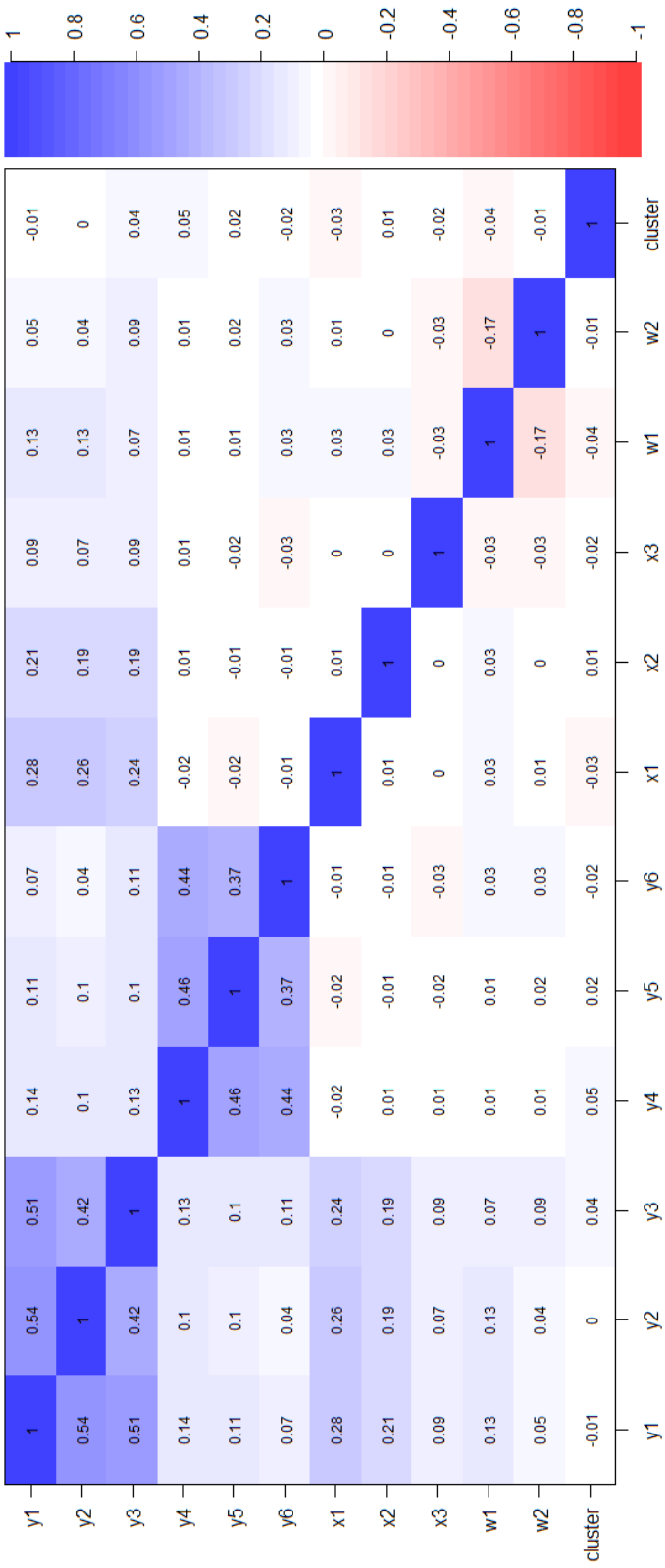

Figura A.6: Matriz de correlaciones. Demo.twolevel

Apéndice B

Apéndice: Código empleado

B.1. Conjunto de datos *PoliticalDemocracy*

```
####CARGA DE LIBRERÍAS UTILIZADAS####
library(lavaan)
library(ggplot2)
library(patchwork)
library(tidyverse)
library(MVN)
library(knitr)
library(kableExtra)
library(stargazer)
library(psych)
library(semPlot)
library(lavaanPlot)

########################################################
####CONJUNTO DE DATOS POLITICALDEMOCRACY###############
########################################################

####LECTURA DE LOS DATOS####

datos <- PoliticalDemocracy
names(datos) <- c("libpren60", "libopos60", "elecc60", "efect60",
                  "libpren65", "libopos65", "elecc65", "efect65",
                  "GNP60", "energia60", "porc_ind60")

####GRÁFICOS DE COLUMNAS####

medias1 <- apply(datos[,1:4], 2, mean)
medias2 <- apply(datos[,5:8], 2, mean)
medias3 <- apply(datos[,9:11], 2, mean)
```

```
grafico1 <- data.frame(variable = names(medias1),
                       medias = round(medias1, 2), row.names = NULL) %>%
  ggplot(aes(x = variable, y=medias,fill=variable)) +
  geom_bar(width = 0.9, stat="identity",
           position = position_dodge(),
           fill = c("blue4", "aquamarine3", "cadetblue", "deeppink3"),
           colour = "black"
           )+
  ylim(c(0,10)) +
  labs(x="", y= "") +
  labs(fill = "")+
  geom_text(aes(label=medias), vjust=-.5,
              position = position_dodge(0.9),  size=4.0
            ) +
  facet_wrap(~"Constructo DEM60")  +
  theme_bw(base_size = 13) +
  theme(axis.text.x = element_text(hjust = 1, vjust = 1,
                                   size = 10, angle = 30))

grafico2 <- data.frame(variable = names(medias2),
                       medias = round(medias2, 2), row.names = NULL) %>%
  ggplot(aes(x = variable, y=medias,fill=variable)) +
  geom_bar(width = 0.9, stat="identity",
           position = position_dodge(),
           fill = c("blue4", "aquamarine3", "cadetblue", "deeppink3"),
           colour = "black"
           )+
  ylim(c(0,10)) +
  labs(x="", y= "") +
  labs(fill = "")+
  geom_text(aes(label=medias), vjust=-.5,
              position = position_dodge(0.9),  size=4.0
            ) +
  facet_wrap(~"Constructo DEM65")  +
  theme_bw(base_size = 13) +
  theme(axis.text.x = element_text(hjust = 1, vjust = 1,
                                   size = 10, angle = 30))

grafico3 <- data.frame(variable = names(medias3),
                       medias = round(medias3, 2), row.names = NULL) %>%
  ggplot(aes(x = variable, y=medias,fill=variable)) +
  geom_bar(width = 0.9, stat="identity",
           position = position_dodge(),
           fill = c("blue4", "cadetblue", "deeppink3"), colour = "black"
           )+
  ylim(c(0,10)) +
  labs(x="", y= "") +
```

```
  labs(fill = "")+
  geom_text(aes(label=medias), vjust=-.5,
              position = position_dodge(0.9),  size=4.0
            ) +
  facet_wrap(~"Constructo IND60")  +
  theme_bw(base_size = 15) +
  theme(axis.text.x = element_text(vjust = 1, size = 10))

design <- "
  12
  33
"


grafico1 + grafico2 + grafico3 + plot_layout(design = design)

####ESTUDIO DESCRIPTIVO####
##Resumen descriptivo

stargazer(datos,header = FALSE,table.placement = "H")
#summary(datos)

####HIPÓTESIS PREVIAS####
##NORMALIDAD MULTIVARIANTE##
#Tests de Mardia

normalidad <- mvn(data = datos, mvnTest = "mardia")
norm_multi <- normalidad$multivariateNormality[-3,]

norm_multi[,2] <- as.numeric(as.vector(norm_multi[,2]))
norm_multi[,3] <- as.numeric(as.vector(norm_multi[,3]))

kable(norm_multi, caption = "\\label{tabla32}Tests de Mardia sobre
      normalidad multivariante", col.names = c("Prueba", "Estadístico",
                                               "p-valor", "Normalidad"),
      align = "c", booktabs = T, digits = 4) %>%
  kable_styling(latex_options = "H")

## Tabla de descriptivos acompañados de valores de asimetría y curtosis

normalidad$Descriptives

kable(normalidad$Descriptives, align = "c", format = "latex", booktabs=T,
    caption = "\\label{tabla33}Descriptivos de las variables bajo estudio
      y niveles de asimetría y curtosis", digits = 4) %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "condensed",
                                              "H"), position = "center")
```

```
##Normalidad Univariante
# Tests de Shapiro-Wilk

norm_univ <- mvn(data = datos, mvnTest = "mardia",
                 univariateTest = "SW")$univariateNormality[,-1]

norm_univ$Normality[norm_univ$Normality == "   YES   "] <- "Sí"
norm_univ$Normality[norm_univ$Normality == "   NO    "] <- "No"

kable(norm_univ, col.names = c("Variable", "Estadístico", "p-valor",
                               "Normalidad"), align = "c",
      caption = "\\label{tablashapiro}Tests de Shapiro-Wilk a las
      variables del estudio", booktabs = T) %>%
  kable_styling(latex_options = "H")

#Gráficos Q-Q

mvn(data = datos, mvnTest = "mardia", univariatePlot = "qqplot")

# Histogramas

mvn(data = datos, mvnTest = "mardia", univariatePlot = "histogram")

## Cálculo del Indicador Alfa de Cronbach
alpha(datos)

##ANÁLISIS DE CORRELACIONES##
#Matriz de correlaciones

mat_cor <- cor(datos, method = "spearman")
cor.plot(mat_cor, cex = 0.75)

#Determinante de la matriz de correlación

det(mat_cor)

#Prueba de esfericidad de Bartlett

bartlett.test(datos)

####CONSTRUCCIÓN Y EVALUACIÓN DEL MODELO

##Especificación del modelo

modelo <- '
  # especificamos el modelo de medida
    IND60 =~ GNP60 + energia60 + porc_ind60
```

```
    DEM60 =~ libpren60 + libopos60 + elecc60 + efect60
    DEM65 =~ libpren65 + libopos65 + elecc65 + efect65
  # especificamos el modelo estructural
    DEM60 ~ IND60
    DEM65 ~ IND60 + DEM60
  # especificamos las correlaciones residuales
    libpren60 ~~ libpren65
    libopos60 ~~ efect60 + libopos65
    elecc60 ~~ elecc65
    efect60 ~~ efect65
    libopos65 ~~ efect65
'

#Ajuste del modelo

ajuste <- sem(modelo, data = datos)

#Índices de ajuste

Indices_ajuste <- round(as.vector(fitMeasures(ajuste,
                                              c("chisq", "df", "cfi",
                                                "tli", "rmsea", "srmr")
                                              )), 4)


kable(cbind("chisq" = as.character(Indices_ajuste[1]),
            "df" = Indices_ajuste[2], "CFI" = Indices_ajuste[3],
            "TLI" = Indices_ajuste[4], "RMSEA" = Indices_ajuste[5],
            "RMSR" = Indices_ajuste[6]), align = "c", digits = 4,
      booktabs = T, caption = "\\label{tabla34}Bondad de ajuste
      del modelo SEM", format = "latex") %>%
  kable_styling(latex_options = "H") %>%
  row_spec(row = 1,
           background = "yellow")

## EXTRACCIÓN DE INFORMACIÓN DEL MODELO
# Matriz de covarianzas implicada

sigma_gorro <- fitted(ajuste)

kable(sigma_gorro, booktabs = T,
      caption = "\\label{covimp1}Matriz de covarianzas implicada
      del modelo. PoliticalDemocracy", digits = 4) %>%
  kable_styling(latex_options = c("H", "scale_down", "condensed"))

# Matrices empleadas en la construcción del modelo
```

```
lavInspect(ajuste)

#Representación gráfica del path diagram

semPaths(ajuste, whatLabels = "std",style="lisrel",
         layout="tree2",
         reorder=FALSE, optimizeLatRes=TRUE, edge.label.position=.5,
         edge.label.cex = 0.8, edge.color = "darkblue" )

## Estimación de los parámetros del modelo y presentación en tabla:

Estim <- parameterEstimates(ajuste, standardized=TRUE)
Relaciones <- paste0(Estim$lhs, Estim$op, Estim$rhs)
tabla35 <- data.frame(Relaciones, Estim$est, Estim$std.all,
                      Estim$se, Estim$pvalue)

kable(tabla35, col.names = c("Relación", "Estimación",
                             "Estim. Estandarizada", "Error Estándar",
                             "p-valor"),booktabs = T,
                             caption = "\\label{tabla35}Relaciones
      estimadas y estandarizadas", align = "c", digits = 4,
      longtable = T) %>%
  kable_styling(latex_options = c("scale_down", "condensed", "H",
                                  "repeat_header"),
                repeat_header_text = "(continuación)") %>%
  row_spec(row = c(15, 16, 18, 19, 22, 34),
           color = "red")

## Resumen del modelo

summary(ajuste2, standardized = T, fit.measures = T)
```

B.2. Conjunto de datos *Demo.twolevel*

```
##########################################################
####CONJUNTO DE DATOS DEMO.TWOLEVEL#####################
##########################################################


####LECTURA DE LOS DATOS####

Demo.twolevel

####GRÁFICOS DE COLUMNAS####
```

```
medias1_2 <- apply(Demo.twolevel[,1:6], 2, mean)
medias2_2 <- apply(Demo.twolevel[,7:9], 2, mean)

grafico1_2 <- data.frame(variable = names(medias1_2),
                         medias = round(medias1_2, 2),
                         row.names = NULL) %>%
  ggplot(aes(x = variable, y=medias,fill=variable)) +
  geom_bar(width = 0.9, stat="identity",
           position = position_dodge(),
           fill = c("blue4", "aquamarine3", "purple", "brown4",
                    "cadetblue", "deeppink3"), colour = "black"
           )+
  ylim(c(-.1,.1)) +
  labs(x="", y= "") +
  labs(fill = "")+

  geom_text(aes(label=medias), vjust=-.5,
              position = position_dodge(0.9),  size=4.0
            ) +
  facet_wrap(~"Ítems y1-y6")  +
  theme_bw(base_size = 15) +
  theme(axis.text.x = element_text(vjust = 1, size = 15))


grafico2_2 <- data.frame(variable = names(medias2_2),
                         medias = round(medias2_2, 2),
                         row.names = NULL) %>%
  ggplot(aes(x = variable, y=medias,fill=variable)) +
  geom_bar(width = 0.9, stat="identity",
           position = position_dodge(), fill = c("aquamarine3", "blue4",
                                                 "deeppink3"),
           colour = "black"
           )+
  ylim(c(-.1, .1)) +
  labs(x="", y= "") +
  labs(fill = "")+

  geom_text(aes(label=medias), vjust=-.5,
              position = position_dodge(0.9),  size=4.0
            ) +
  facet_wrap(~"Covariables Intraclase x1-x3")  +
  theme_bw(base_size = 15) +
  theme(axis.text.x = element_text(vjust = 1, size = 15))


grafico1_2 + grafico2_2
```

```
medias3_2 <- apply(Demo.twolevel[,10:11], 2, mean)

grafico3_2 <- data.frame(variable = names(medias3_2),
                         medias = round(medias3_2, 2),
                         row.names = NULL) %>%
  ggplot(aes(x = variable, y=medias,fill=variable)) +
  geom_bar(width = 0.7, stat="identity",
           position = position_dodge(),
           fill = c("blue4", "deeppink3"), colour = "black"
           )+
  ylim(c(-.1, .1)) +
  labs(x="", y= "") +
  labs(fill = "")+

  geom_text(aes(label=medias), vjust=-.5,
             position = position_dodge(0.9),  size=4.0
            ) +
  facet_wrap(~"Covariables Interclase w1-w2")  +
  theme_bw(base_size = 15) +
  theme(axis.text.x = element_text(vjust = 1, size = 15))

grafico3_2

####ESTUDIO DESCRIPTIVO####
##Resumen descriptivo

stargazer(Demo.twolevel,header = FALSE,table.placement = "H")
#summary(datos)

####HIPÓTESIS PREVIAS####
##NORMALIDAD MULTIVARIANTE##
#Tests de Mardia

normalidad2 <- mvn(data = Demo.twolevel, mvnTest = "mardia")
norm_multi2 <- normalidad2$multivariateNormality[-3,]

norm_multi2[,2] <- as.numeric(as.vector(norm_multi2[,2]))
norm_multi2[,3] <- as.numeric(as.vector(norm_multi2[,3]))

kable(norm_multi2, caption = "\\label{tabla38}Tests de Mardia sobre
      normalidad multivariante. Demo.twolevel",
      col.names = c("Prueba", "Estadístico", "p-valor", "Normalidad"),
      align = "c", booktabs = T,
      digits = 4) %>%
  kable_styling(latex_options = "H")

## Tabla de descriptivos acompañados de valores de asimetría y curtosis
```

```
normalidad2$Descriptives

kable(normalidad2$Descriptives, align = "c", format = "latex", booktabs=T,
      caption = "\\label{tabla39}Descriptivos de las variables bajo
      estudio y niveles de asimetría y curtosis Demo.twolevel",
      digits = 4) %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "condensed",
                                              "H"), position = "center")

##Normalidad Univariante
# Tests de Shapiro-Wilk

norm_univ2 <- mvn(data = Demo.twolevel, mvnTest = "mardia",
                  univariateTest = "SW")$univariateNormality[,-1]

norm_univ2$Normality[norm_univ2$Normality == "   YES    "] <- "Sí"
norm_univ2$Normality[norm_univ2$Normality == "   NO     "] <- "No"

kable(norm_univ2, col.names = c("Variable", "Estadístico", "p-valor",
                                "Normalidad"), align = "c",
      caption = "\\label{tablashapiro2}Tests de Shapiro-Wilk a las
      variables del estudio. Demo.twolevel", booktabs = T) %>%
  kable_styling(latex_options = "H")

#Gráficos Q-Q

mvn(data = Demo.twolevel, mvnTest = "mardia", univariatePlot = "qqplot")

# Histogramas

mvn(data = Demo.twolevel, mvnTest = "mardia",
    univariatePlot = "histogram")


##ANÁLISIS DE CORRELACIONES##
#Matriz de correlaciones

mat_cor2 <- cor(Demo.twolevel, method = "spearman")
cor.plot(mat_cor2, cex = 0.75)

#Determinante de la matriz de correlación

det(mat_cor2)

#Prueba de esfericidad de Bartlett

bartlett.test(Demo.twolevel)
```

```
####CONSTRUCCIÓN Y EVALUACIÓN DEL MODELO

##Especificación del modelo

modelo2 <- '
    level: 1
        FW =~ y1 + y2 + y3
        FW ~ x1 + x2 + x3
    level: 2
        FB =~ y1 + y2 + y3
        FB ~ w1 + w2
'

#Ajuste del modelo

ajuste2 <- sem(model = modelo2, data = Demo.twolevel, cluster = "cluster")

#Índices de ajuste

Indices_ajuste2 <- round(as.vector(fitMeasures(ajuste2,
                                               c("chisq", "df","cfi",
                                                 "tli", "rmsea", "srmr"))),
                         4)


kable(cbind("chisq" = as.character(Indices_ajuste2[1]),
            "df" = Indices_ajuste2[2], "CFI" = Indices_ajuste2[3],
            "TLI" = Indices_ajuste2[4], "RMSEA" = Indices_ajuste2[5],
            "RMSR" = Indices_ajuste2[6]), align = "c", digits = 4,
      booktabs = T, caption = "\\label{tabla310}Bondad de ajuste del
      modelo SEM multinivel", format = "latex") %>%
  kable_styling(latex_options = "H") %>%
  row_spec(row = 1,
           background = "yellow")

## EXTRACCIÓN DE INFORMACIÓN DEL MODELO
# Medias y covarianzas

fitted(ajuste2) #lavInspect(ajuste2, "h1")

# Matrices empleadas en la construcción del modelo

lavInspect(ajuste2)

#Representación gráfica del path diagram

lavaanPlot(model = ajuste2,
```

```
                        node_options = list(shape = "box",
                                            fontname = "Helvetica"),
           edge_options = list(color = "blue"),
           coefs = T, stand = T, covs = T)

## Estimación de los parámetros del modelo y presentación en tabla:

Estim2 <- parameterEstimates(ajuste2, standardized=TRUE)
Relaciones2 <- paste0(Estim2$lhs, Estim2$op, Estim2$rhs)
tabla311 <- data.frame(Relaciones2, Estim2$est, Estim2$std.all, Estim2$se,
                       Estim2$pvalue)

kable(tabla311, col.names = c("Relación", "Estimación",
                              "Estim. Estandarizada",
                              "Error Estándar", "p-valor"),
      booktabs = T, caption = "\\label{tabla311}Relaciones estimadas y
      estandarizadas. Demo.twolevel", align = "c", digits = 4,
      longtable = T) %>%
  kable_styling(latex_options = c("scale_down", "condensed", "H",
                                  "repeat_header"),
                repeat_header_text = "(continuación)") %>%
  row_spec(row = c(28, 29, 36, 37, 38),
           color = "red")

##Resumen del modelo

summary(ajuste2, standardized = T, fit.measures = T)
```

Bibliografía

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2023). *rmarkdown: Dynamic Documents for R*. R package version 2.22.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley and Sons, Incorporated, first edition.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Cupani, M. (2012). Análisis de ecuaciones estructurales: conceptos, etapas de desarrollo y un ejemplo de aplicación. *Revista Tesis*, 2(1):186–199.

Gutiérrez-Doña, B. (2008). *Modelos lineales estructurales: Conceptos básicos, aplicaciones y programación con LISREL*. Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica, first edition.

J. Curran, P. e. a. (1996). The robustness of tests statistics to nonormality and specification error in confirmatory factor analisys. *Psychological Methods*, 1(1):16–29.

Labraca, J. M. I. (2021). "modelos de ecuaciones estructurales". Disponible en http://repositorio.ual.es/handle/10835/13177.

Luque-Calvo, P. L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*.

Luque-Calvo, P. L. (2019). *Cómo crear Tablas de información en R Markdown*.

Manzano Patiño, A. (2018). Introducción a los modelos de ecuaciones estructurales. *Investigación en Educación Médica*, 7(25):67–72.

Maruyama, G. M. (1997). *Basics of Structural Equation Modeling*. SAGE Publications, Incorporated, first edition.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roseel, Y., Jorgensen, T. D., Roockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., and Du, H. (2023). *lavaan: Latent Variable Analysis*. R package version 0.6-15.

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rosseel, Y. (2022). *The lavaan tutorial*.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Ruiz, M. A. e. a. (2010). Modelos de ecuaciones estructurales. *Papeles del Psicólogo*, 31(1):34–45.

Techopedia (2017). "definition - what does business intelligence (bi) mean?". Disponible en https://www.techopedia.com/definition/345/business-intelligence-bi.

Tomás Vargas Halabí, R. M.-E. (2017). Tamaño de la muestra en modelos de ecuaciones estructurales con constructos latentes: Un método práctico. *Actualidades Investigativas en Educación*, 17(1):1–34.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., and Dunnington, D. (2023a). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.4.2.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023b). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.2.

Xie, Y. (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.43.