# Modelos de Ecuaciones Estructurales

Trabajo Fin de Grado

Autor:
Juan Miguel Iglesias Labraca

Tutor:
Fernando Reche Lorite

Grado en Matemáticas



Julio, 2021
Universidad de Almería

# *Índice general*

# *Abstract*

Structural equation modeling (SEM) is a multivariate statistical technique that, using multiple regression and factor analysis, aims to validate causal relationships between variables, whether observed or not. They stand out, mainly, for their applications in social sciences.

This work is presented as a synthesis of the main results used in SEM, as well as its subsequent application to a real practical case through the use of the statistical software R. To do this, we will begin by making a brief introduction where to establish the logic used and where to fix the historical context in which they arise. Later, the concept of causality and the types of structural equations will be presented.

The second chapter will establish the nomenclature used, as well as a series of basic results underlying path diagrams. In the third chapter, we will continue by studying the previous restrictions associated with the construction of the model. We must stand out the need for multivariate normality for the estimation of parameters by maximum likelihood, as well as a sample size appropriate to the number of observed variables in our model.

The fourth chapter stands as the mainstay of structural equation modeling. In it we will establish the theoretical foundation for its construction. We will differentiate four stages: specification, identification, estimation and evaluation.

Finally, we will develop a structural equation model with the intention of studying what influence the public perception of corruption has on institutional trust. At the same time, it will be studied if there is a possible causal relationship between socioeconomic status and the citizen's perceptions of corruption and performance of the Spanish democratic system, among others.

# *Resumen*

Los modelos de ecuaciones estructurales (SEM) son una técnica estadística multivariante que, haciendo uso de la regresión múltiple y el análisis factorial, tiene como objetivo validar relaciones causales entre variables, ya sean observadas o no. Destacan, principalmente, por sus aplicaciones en ciencias sociales.

Se presenta este trabajo como una síntesis de los principales resultados usados en SEM, así como su posterior aplicación a un caso práctico real mediante el uso del sofware estadístico R. Para ello, comenzaremos realizando una breve introducción donde establecer la lógica empleada y donde podamos fijar el contexto histórico en el que surgen. Posteriormente, se presentará el concepto de causalidad, así como los tipos de ecuaciones estructurales existentes.

En el segundo capítulo se establecerá la nomenclatura usada, así como una serie de resultados básicos subyacentes a los diagramas de caminos. Continuaremos ya, en el tercer capítulo, estudiando las restricciones previas asociadas a la construcción del modelo. De especial importancia será la necesidad de normalidad multivariante para la estimación de parámetros por máxima verosimilitud, así como un tamaño muestral adecuado al número de variables observadas de nuestro modelo.

El capítulo cuarto se erige como el pilar fundamental de los modelos de ecuaciones estructurales. En él estableceremos la fundamentación teórica para la construcción de los mismos diferenciando cuatro etapas: especificación, identificación, estimación y evaluación del modelo.

Finalmente, desarrollaremos un modelo de ecuaciones estructurales con la intención de estudiar qué influencia tiene la percepción ciudadana de la corrupción en la confianza institucional. A su vez, se estudiarán las posibles relaciones causales entre el estatus socioeconómico unipersonal y las percepciones ciudadanas de la corrupción y del funcionamiento del sistema democrático español, entre otros.

# 1 Introducción

## 1.1 Origen de los modelos de ecuaciones estructurales

Los modelos de ecuaciones estructurales son una herramienta estadística multivariante que tiene como objeto validar y cuantificar relaciones causales entre variables. Los SEM cobran importancia debido a que permiten especificar relaciones entre variables observadas y no observadas. Además, a diferencia de los modelos de regresión, tienen consideración explícita de los errores de medida, permitiendo modelar correlaciones entre ellos mediante ecuaciones lineales.

Veremos que la hipótesis fundamental sobre la que se construyen estos modelos consiste en que, si el modelo es correcto y conocemos los parámetros del mismo, la matriz de covarianzas poblacional podría ser reproducida exactamente a partir de la combinación de los parámetros del modelo:

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

donde $\sum$ es la matriz de covarianzas poblacionales entre las variables observadas, $\theta$ es un vector que contiene los parámetros del modelo y $\sum(\theta)$ es la matriz de covarianzas derivadas como una función de los parámetros contenidos en el vector $\theta$.

Nótese que el empleo de matrices de covarianzas en lugar de observaciones individuales, hace que modelos como regresión, ANOVA y modelos econométricos, entre otros, constituyan casos especiales de los modelos SEM (Bollen, 1989). Destacar además que los modelos SEM son usados, generalmente, como una técnica de análisis confirmatorio. Sin embargo, esto no implica que puedan ser usados con carácter exploratorio. Los principales ámbitos donde se emplean dichos modelos con la intención de validar resultados teóricos suelen ser la psicología, economía o biología.

### *Contexto histórico*

Los modelos de ecuaciones estructurales surgen como alternativa al análisis factorial, conocido hasta 1960 por ser el método predominante para establecer relaciones causales entre variables latentes y observadas. El análisis factorial, a diferencia de los SEM, es demasiado restrictivo. En concreto, si dos variables están correladas, este asume que es debido a la presencia de causas latentes comunes entre ellas. Consecuentemente, el análisis factorial exploratorio busca estas causas lo cual puede ser a veces engañoso.

En 1921 Sewell Wright desarrolló un nuevo enfoque basado en la correlación y la idea de establecer relaciones causales lineales. Hasta el momento, las correlaciones se usaban para, de forma exploratoria, descubrir qué relaciones existían entre variables. Wright fue aún más lejos, ya que con su modelo buscaba inferir dichas correlaciones para, posteriormente, compararlas con las correlaciones observadas. Dicha técnica, denominada *path analysis*, implementaba un sistema constituido por flechas que estable-

cía *caminos causales*[1] entre variables. Esta representación gráfica será conocida como *path diagrams*.

Por otro lado, en 1936 John Maynard Keynes elaboró un modelo económico haciendo uso de sistemas de ecuaciones lineales que relacionaban conjuntos distintos de variables. Muchos economistas vieron dicho modelo como una extensión del análisis de regresión y se valieron del álgebra matricial para representar dichas ecuaciones. Sin embargo, cuando buscaron métodos para estimar los parámetros de las ecuaciones estructurales, se dieron cuenta de que no podían identificar cuando estas tenían solución única. En base a esto, algunos sociólogos establecieron una relación entre las ecuaciones lineales simultáneas y el path analysis de Wright.

En 1966, Bock y Bargmann proponen un método denominado análisis de estructuras de covarianza. Dicho método se usará para validar modelos de ecuaciones estructurales lineales mediante la covariación entre variables. A su vez, Karl Jöreskog propone en 1969 un método denominado análisis factorial confirmatorio general, cuyo fin es determinar si el número de factores obtenidos y sus cargas se corresponden con los que cabría esperar a la luz de una teoría previa acerca de los datos. Además describiría la solución de máxima verosimilitud para la estimación de parámetros de dicho modelo. Sin embargo, no sería hasta 1970 cuando Jöreskog vio la relación entre ambos, proponiendo así un algoritmo de estimación de máxima veorisimilitud para obtener los parámetros en el modelo de Bock y Bargmann.

Finalmente en 1973, Jöreskog colaboró con Arthur S. Goldeberg y elaboró su propio modelo de ecuaciones lineales simultáneas con variables latentes. Además, lo dotó con un algoritmo computacional mediante el cual podrían realizar pruebas de bondad de ajuste chi-cuadrado. Este programa fue llamado LISREL (*linear structural relations*), combinando así el análisis factorial con las ecuaciones estructurales simultáneas.

LISREL proporcionó a los investigadores la oportunidad de pensar en términos de hipótesis causales, así como el hecho de poder probarlas. A raíz de la aparición de este programa, se desarrollaron numerosos artículos. Destacar, en particular, la creación de la revista *Structural Equation Modeling* en 1994.

A lo largo de los años, muchos han seguido la estela de Jöreskog con programas de su propia invención. El modelo de ecuaciones estructurales propuesto por Jöreskog fue formulado con álgebra matricial, lo cual se transmitió a LISREL, requiriendo así al usuario establecer de entrada su modelo en forma matricial. Además, surgiría otro problema, ya que se tenían que separar las variables según si dependian de variables latentes endógenas o exógenas[2]. Resultaba pues difícil formular un modelo en el que una variable dependiese tanto de unas como de otras.

[1] Los caminos causales o también conocidos como *causal paths*, son secuencias causales relativamente probables entre un conjunto complejo de causas y efectos potenciales.

[2] Una variable endógena es aquella cuyo valor está determinado por las relaciones establecidas dentro del modelo en el que está incluida. Por contra, el valor de una variable exógena está determinado por factores externos al modelo.

En este contexto, Bentler y Weeks proponen en 1982 un modelo en el que no se precisa de la separación entre variables de entrada. En él, cualquier variable observada podría depender tanto de variables endógenas como exógenas. Ahora bien, las distinguirán por ser funciones o no de otras variables, es decir, variables dependientes o independientes.

Una vez especificado el modelo elaborando una ecuación para cada variable dependiente (donde cada ecuación está compuesta por las causas directas que la afectan), podríamos distinguir fácilmente entre variables latentes y observadas. Bastará con ver si existe una ecuación para dicha variable o no. Si no existe dicha ecuación, podremos afirmar que nos encontramos ante una variable independiente.

De este modo, Bentler desarrolló y perfeccionó un programa llamado EQS entre los años 1989-1995, implementando así la idea de que dicho modelo podía ser descrito especificando ecuaciones lineales para cada variable dependiente del sistema. Además, con la ayuda de sus programadores, consiguió elaborar un método mediante el cual podía especificar el modelo computacionalmente a través de un path diagram.

A día de hoy, LISREL y EQS se consideran programas obsoletos para el estudio de SEM. En contraposición, podemos encontrar otros programas vigentes tales como: *Amos*, *Steiger's EZPath* y *SEPath*, *McDonald's COSAN*, *Neale's Mx* y *Fox's SEM* para R.

Las últimas contribuciones al campo de los modelos de ecuaciones estructurales serían realizadas por Pearl, Spirtes, Glymour y Scheines entre los años 1993-2000 construyendo la teoría causal en el contexto de la teoría de grafos, introduciendo así conceptos como los grafos acíclicos dirigidos, el uso de la condición de Markov, condición de minimalidad o fidelidad y d-separación, entre otros. Una revisión metodológica de estos conceptos puede encontrarse en [12].

Concluimos esta breve introducción destacando el continuo crecimiento y profundización conceptual, estadística y computacional a la que se han encontrado sometidos los SEM a lo largo de todo el siglo XX.

## *Naturaleza de la causalidad*

A día de hoy, no existe una definición clara de causalidad. De hecho, Nagel estableció en 1965 una hipótesis que sigue siendo aceptada a día de hoy: «No hay una única explicación correcta para el término causa». Organizaremos pues este capítulo entorno al significado de causalidad en términos de modelos de ecuaciones estructurales.

Por norma, no podremos inferir una conexión causal a partir de la correlación entre dos variables. A lo sumo, podríamos afirmar que la covariación entre dos variables implica que ciertos valores de una, están asociados a ciertos valores de otra. Llegamos así a la siguiente definición:

**Definición 1.1.** ***(Relación causal)*** *Diremos que existe una relación causal entre dos o más variables cuando además de la existencia de covariación, todo cambio en una variable suponga un efecto en la otra.*

Obsérvese que la covariación define una relación simétrica entre variables mientras que, por lo general, toda relación causal será asimétrica.

Consideraremos pues como punto de partida dos variables $x_1$ e $y_1$. Para representar el efecto causal de $x_1$ sobre $y_1$, suponiendo que este sea lineal, utilizaremos la siguiente expresión:

$$y_1 = \gamma_{11} x_1 + \zeta_1 \ , \tag{1.2}$$

donde $\gamma_{11}$[3] consiste en la influencia esperada de $x_1$ en $y_1$, y donde $\zeta_1$ es una perturbación desconocida con $E(\zeta_1) = 0$. Obsérvese que, en este modelo, ninguna otra causa de $y_1$ está correlacionada con $x_1$. Optaremos así por una definición de causa que implica la correlación entre variables, el aislamiento de causas externas y el establecimiento de la dirección del efecto.

En el caso que nos compete, los conceptos de aislamiento y correlación son los más difíciles de evaluar. Para establecer $x_1$ como una causa de $y_1$, $x_1$ tiene que estar aislada de $\zeta_1$. Como $\zeta_1$ es una perturbación no observada, no tenemos control directo sobre ella. Sin embargo, haremos suposiciones acerca de su comportamiento creando así una condición de pseudo-aislamiento.

**Definición 1.2.** ***(Condición de pseudo-aislamiento)*** *Diremos que una variable está pseudo-aislada si no existe una covariación entre el término de perturbación y todas las variables explicativas*[4] *incluidas en el modelo.*

Por otro lado, la creedibilidad de una relación causal depende finalmente de la dirección del efecto. Esto es, si se afirma que $x_1$ es la causa de $y_1$, entonces la variación de $x_1$ debe variar $y_1$, pero la variación de $y_1$ no debe influir en $x_1$ necesariamente.

En 1977, Hume estableció la necesidad de prioridad temporal como una cuarta condición para establecer una relación causal. Esto es, la presunta causa debe preceder al efecto. Sin embargo, esta y aquellas definiciones de causalidad que dependen de que el efecto sea perfectamente predecible, carecen de sentido. Esto se debe a que, de ser así, no podrían establecerse relaciones simultáneas. Además, en caso de aceptar dicho intervalo temporal, ¿qué amplitud ha de tener? Optaremos pues, por una visión probabilística del concepto de causalidad (Suppes, 1970).

Si bien es cierto, en algunos casos aceptaremos un desfase temporal entre la causa y el efecto. Si ese desfase es menor que el intervalo anteriormente mencionado, la causa estará establecida en el mismo periodo temporal que el efecto, generándose así una relación de retroalimentación entre ambas. Esto es, un efecto futuro puede causar un evento pasado.

Lo razonado hasta ahora en la ecuación (1.2) es extrapolable al caso multivariante ya que, generalmente, un experimento aleatorio necesitará de variables múltiples que

[3] En primer lugar, escribiremos el subíndice de la variable causa. Seguidamente, haremos lo propio con el subíndice de la variable efecto.

[4] También conocidas en regresión como variables independientes. Son aquellas que usamos para explicar, describir o predecir las variables dependientes.

lo expliquen:

$$\begin{aligned} y_1 &= \gamma_{11}x_1 + \gamma_{12}x_2 + \cdots + \gamma_{1q}x_q + \zeta_1 \ , \\ y_1 &= \mathbf{\Gamma_1}\boldsymbol{x} + \zeta_1 \ , \end{aligned} \tag{1.3}$$

donde $\mathbf{\Gamma_1}$ es un vector $1 \times q$ de $[\gamma_{11} \dots \Gamma_{1q}]$ y $E(\zeta_1) = 0$. Asumimos además que la condición de pseudo-aislamiento es $COV(\mathbf{x}, \zeta_1) = 0$ y que la correlación viene dada por $\gamma_{1j}$, que es el cambio esperado en $y_1$ para un cambio en $x_j$ cuando las demás $x_i$ $(\forall i \neq j)$ se mantienen constantes. Como no tenemos control sobre $\zeta_1$, este cambio nunca será exacto.

Hasta el momento, hemos estado tratando con modelos compuestos por variables observadas[5]. Sin embargo, esto es extrapolable a todos los modelos de ecuaciones estructurales. Obsérvese, por ejemplo, un modelo de medida para una variable observada $x_1$ y las variables latentes[6] $\xi_i$ para todo i=1,2,...,n que la afectan:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots + \lambda_{1n}\xi_n + \delta_1 \ , \\ x_1 &= \mathbf{\Lambda_1}\boldsymbol{\xi} + \delta_1 \ , \end{aligned} \tag{1.4}$$

donde $\mathbf{\Lambda_1}$ es un vector $1 \times p$ de $[\lambda_{11} \dots \lambda_{1p}]$ y $E(\delta_1) = 0$. Razonando de igual modo, la condición de pseudo-aislamiento viene dada por $COV(\boldsymbol{\xi}, \delta_1) = 0$ y la condición de correlación por $\lambda_{1j}$. Se entiende además que se verifica la condición de la dirección causal.

## 1.2 Tipos de modelos de ecuaciones estructurales

Para hacer una correcta síntesis de los modelos de ecuaciones estructurales debemos diferenciar entre tres tipos:

1. Los ***modelos de variables observadas***, también conocidos como *path analysis*, no incluyen variables latentes y asumen la no existencia de errores de medición. Consisten en una generalización del modelo de regresión múltiple, ya que tienen en cuenta múltiples ecuaciones de regresión que pueden ser estimadas simultáneamente.

2. El ***análisis factorial*** es una técnica estadística de reducción de datos que se usa para explicar las correlaciones entre variables observadas haciendo uso del menor número de variables latentes posible. Diferenciaremos entre *análisis factorial confirmatorio* y *análisis factorial exploratorio* dependiendo de si nuestro objetivo es confirmar una estructura factorial previamente definida, o en el caso de que no se considere ninguna, poder hallarla.

3. El ***modelo de ecuaciones estructurales general*** consiste en una generalización de los dos modelos previamente expuestos. Permite evaluar la relación entre variables observadas y latentes, así como las relaciones entre constructos.

En el caso que nos atañe, dedicaremos este trabajo a la exposición de los principales fundamentos y técnicas usadas para la construcción del modelo más generalista.

[5] Indicador de un fenómeno dado en el estudio previo a la definición del modelo. Ver definición 2.1.
[6] Variable no observada en el estudio previo a la definición del modelo. Ver definición 2.2.

# 2

# Principios de los modelos de ecuaciones estructurales

## 2.1 Tipos de variables en los modelos de ecuaciones estructurales

Para poder formular correctamente el modelo necesitaremos de una notación estándar a partir de la cual diferenciar los objetos del estudio. Esta sección estará dedicada a dar nombre a las principales variables que nos encontraremos en la elaboración del modelo causal. En primer lugar, y de acuerdo con lo anteriormente expuesto, el principal objetivo de los SEM consiste en establecer relaciones causales entre indicadores medibles y aquellas variables que no se pueden observar directamente.

**Definición 2.1.** ***(Variable observada o indicador)*** *Variable obtenida mediante la medición de un fenómeno dado.*

**Definición 2.2.** ***(Variable latente o no observada)*** *Variable del modelo no medida y, por ende, libre de errores de medición. También conocida como factor.*

Por otro lado, indiferentemente de si una variable es observada o no, tendremos que hacer una apreciación dependiendo de si esta es causa o efecto de otra variable. El hecho de que una variable sea efecto de otra, no excluye que pueda ser causa de una tercera variable. Esto lo veremos más adelante. Distringuiremos así entre variables endógenas y exógenas.

**Definición 2.3.** ***(Variable exógena)*** *Variable no causada por ninguna otra variable del modelo. Generalmente, pueden ser causas de otras variables. Conocidas, en regresión, como variables independientes.*

**Definición 2.4.** ***(Variable endógena)*** *Variable causada por otra variable del modelo. Una variable endógena puede causar a su vez otra variable endógena. Por analogía, diremos que son las variables dependientes en los modelos de regresión. Esta debe ir acompañada siempre de un error.*

Por último, destacar la existencia de una variable que refleje el error cometido a la hora de elaborar nuestro modelo. Distinguiremos entre errores de medida (a la hora de estimar las variables observadas de nuestro modelo) y errores de predicción (cuando el error se fundamente en los cálculos matemáticos derivados de las suposiciones hechas a la hora de hipotetizar el modelo).

**Definición 2.5.** ***(Variable error)*** *Variable latente que representa tanto a los errores de medida como a las variables no contempladas en el modelo y que pueden afectar a la medición de una variable observada. El error asociado a una variable dependiente es conocido como error de predicción.*

La notación que emplearemos a lo largo de este trabajo para hacer referencia a estas variables consistirá en aquella desarrollada por Jöreskog (1973, 1977), Wiley (1973) y Keesling (1972), posteriormente popularizada en LISREL:

- Las variables exógenas observables se escribirán como $x$ y las variables exógenas latentes como $\xi$.
- Las variables endógenas observables se representarán como $y$ mientras que las endógenas latentes como $\eta$.
- Los errores de medida de las variables observables exógenas se representan como $\delta$, los de las variables endógenas como $\varepsilon$ y los errores que afectan a las variables latentes endógenas como $\zeta$. Las variables latentes exógenas no estarán afectadas por error.
- Los coeficientes que representan la relación de una variable latente (endógena o exógena) con sus indicadores se denotan mediante $\lambda$, acompañada de un subíndice que indique el tipo de variable ($\mathbf{x}$ o $\mathbf{y}$) si fuera necesario distinguirlos.
- Los efectos de una variable endógena sobre otra endógena se representan por un coeficiente $\boldsymbol{B}$, los efectos de una variable exógena sobre otra endógena se representan por un coeficiente $\gamma$.
- Las covarianzas entre los errores de medida se representan como $\theta$, acompañado del subíndice $\delta$ o $\varepsilon$, según al error que se refieran. Las covarianzas entre variables latentes exógenas se representan como $\boldsymbol{\Phi}$.
- Las covarianzas entre los errores de variables endógenas latentes se representan como $\boldsymbol{\Psi}$.

Ahora bien, establecer las hipótesis de partida para la elaboración de nuestro modelo puede resultar algo complejo si no lo vemos gráficamente. Es por ello que surgen los digramas de caminos.

## 2.2 Diagramas de caminos

La complejidad de los modelos de ecuaciones estructurales precisa de una representación simplificadora. En 1921, Sewell Wright mostró que, mediante los diagramas de caminos, o también conocidos como grafos dirigidos, era posible ajustar el modelo. Los diagramas de caminos se rigen por las siguientes normas:

- Las variables observadas serán representadas mediante cuadrados y las variables latentes mediante círculos.
- Los errores, ya sean perturbaciones o variables residuales, carecerán de estructuras visuales que los contengan (aunque algunos programas los dibujan como variables latentes).
- Los caminos causales directos serán representados mediante flechas apuntando desde la variable causa hasta la variable efecto. Si no existe un camino entre dos variables, entonces no existe una relación causal entre ambas.
- Las relaciones entre variables de tipo covariante (correlaciones y covarianzas) se representarán como vectores curvos con una flecha en cada extremo.

En la siguiente imagen podemos observar un resumen de los principales símbolos empleados para denotar a las principales variables y relaciones de un diagrama de caminos:

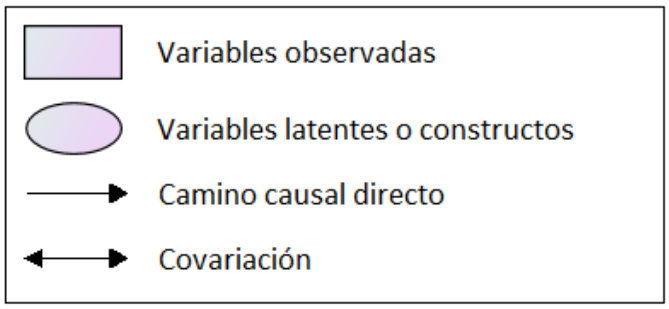


Figura 2.1: Nomenclatura diagramas de caminos

En base a lo anteriormente expuesto y antes de continuar, debemos diferenciar entre los conceptos de perturbación y variable residual. Una perturbación representa influencias en el modelo tales como errores de medida, y son estableciendo una analogía, parecidas a los factores únicos del análisis factorial. Sin embargo, las perturbaciones pueden contener tanto errores sistemáticos como no sistemáticos. Además, normalmente se suponen incorreladas entre sí y con otras variables exógenas. En cuanto a los residuos, destacar que son el resultado de las operaciones matemáticas realizadas en el modelo.

Por último, y en base a nuestro diagrama de caminos, podremos diferenciar entre dos tipos de modelos: recursivos o no recursivos.

**Definición 2.6.** ***(Modelo recursivo)*** *Modelo en el que las perturbaciones no están relacionadas y donde los efectos causales son unidireccionales.*

**Definición 2.7.** ***(Modelo no recursivo)*** *Modelo en el que puede existir referencias circulares o cuyos errores pueden estar correlacionados.*

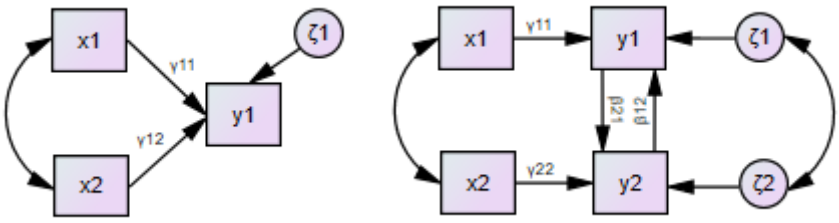


Figura 2.2: A la izquierda un modelo recursivo. A la derecha un modelo no recursivo.

Identificar y desarrollar un modelo de ecuaciones estructurales no resulta tarea sencilla. Es por ello, que muchos investigadores comienzan dibujando el diagrama de caminos, que posteriormente convierten a un sistema de ecuaciones lineales válido para usarlo computacionalmente. Sin embargo, antes de que veamos cómo establecer

dichas ecuaciones, necesitaremos establecer el tipo de relaciones causales patentes entre las variables y otros parámetros del modelo.

### *Tipos de relaciones entre variables*

Para identificar los tipos de relaciones causales inherentes en los modelos de ecuaciones estructurales supondremos dos variables A y B cuya covarianza sea no nula. Las diferentes formas de covariar entre dos variables son:

- A causa B: relación directa en la que asumimos un modelo de regresión donde A es la variable causa y B la variable efecto.
- B causa A: relación análoga a la anterior en la que podemos establecer B como la variable causa y A como la variable efecto.
- A causa B y B causa A: ocurre cuando existe reciprocidad entre las relaciones directas anteriormente mencionadas.
- Relación espuria entre A y B: diremos que A y B están espuriamente relacionadas si ambas poseen una causa común. Es decir, existe una variable C que causa A y B.
- A causa C y C causa B: relación indirecta entre A y B debido a la participación de una tercera variable C entre ambas.
- Relación espuria + relación indirecta: en este caso A y C serán variables exógenas. No se especificará el tipo de relación existente entre ambas y, por ende, no podremos distinguir si C influye indirectamente o espuriamente en la covariación de A y B. Denominaremos a esta relación como efecto conjunto.

En la siguiente figura podemos ver una representación gráfica de los tipos de relaciones que nos podemos encontrar entre las variables de nuestro modelo:

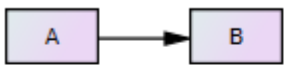


Figura 2.1: A causa B


Figura 2.2: B causa A

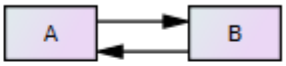


Figura 2.3:
A causa B y B causa A

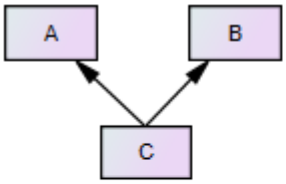


Figura 2.4: Relación
espuria entre A y B

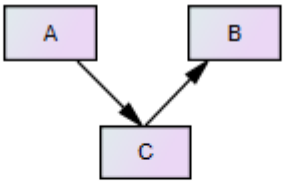


Figura 2.5:
A causa C y C causa B

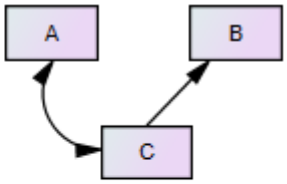


Figura 2.6:
Efecto conjunto

Figura 2.3: Tipos de relaciones entre variables

### *Reglas de descomposición de los diagramas de caminos*

Una vez determinadas las fuentes de covariación de nuestro modelo, debemos establecer la relación entre los parámetros y las covariaciones entre variables. Surgen así las llamadas reglas de descomposición que nos permitirán, en última instancia, estimar el valor de los parámetros a partir de las relaciones entre variables.

Consideraremos a las varianzas y covarianzas de las variables exógenas como parámetros del modelo. Las reglas de descomposición para derivar el resto de varianzas y covarianzas son:

1. La covarianza ente dos variables es igual a la suma de los efectos directos, indirectos, espurios y conjuntos. Estos efectos se representarán mediante flechas. El origen puede ser tanto una de las dos variables (efectos directos e indirectos), una tercera variable (efectos espúreos), o una covarianza entre variables exógenas (efectos conjuntos). El efecto es el resultado de multiplicar la varianza de la variable de partida (o covarianza en su caso) por todos los parámetros asociados a las flechas recorridas hasta unir las variables de interés. No podremos pasar más de una vez por la misma variable.

2. La varianza de una variable dependiente es igual a la varianza del término de perturbación más la varianza explicada por las otras variables del modelo. Esta varianza explicada puede expresarse a su vez en función de todas las variables explicativas con efecto directo sobre la dependiente, como suma de todos los productos entre estos efectos directos y las covarianzas entre la variable dependiente y la explicativa relacionadas por dichos efectos.

# 3

# Hipótesis previas

## Multinormalidad de las variables observadas

En la sección 4.3 veremos que la etapa de estimación precisa de ciertos métodos capaces de aproximar los valores de los parámetros del modelo. Entre ellos destacaremos la estimación por máxima verosimilitud. Una de las hipótesis previas para poder aplicar este método reside en la necesidad de que las variables observadas de nuestro modelo sigan de forma conjunta una distribución normal multivariante:

**Definición 3.1.** *Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ una variable aleatoria k-dimensional. Diremos que $\boldsymbol{X}$ sigue una distribución* ***normal multivariante*** *si su densidad de probabilidad conjunta viene dada por:*

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\sum|}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}) \cdot \sum^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^T\right\}, \quad x_i \in \mathbb{R} \ \ \forall i\,, \tag{3.1}$$

*donde $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ con $\mu_i \in \mathbb{R}$ para cada $i$, y $\sum$ es una matriz cuadrada de orden k simétrica y definida positiva*[1].

**Proposición 3.1.** *Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ una variable aleatoria k-dimensional. Si $\boldsymbol{X}$ sigue una distribución normal multivariante, entonces $X_i \ \forall i = 1, 2, \ldots, k$ sigue una distribución normal univariante. El recíproco no es cierto.*

La normalidad univariante es una condición necesaria pero no suficiente para que se satisfaga la multinormalidad. Debido a esto, necesitaremos comprobar, en primer lugar, que todas las variables observadas se distribuyen normalmente. Para ello podemos hacer uso de los tests de *Kolmogorov-Smirnov-Lilliefors* o *Shapiro-Wilk* y los *contrastes de asimetría y curtosis*. En referencia a estos últimos, Mardia (1970, 1974, 1985) propone tests multivariantes de curtosis y asimetría con la intención de certificar si un conjunto de variables observadas sigue una distribución normal multivariante:

**Definición 3.2.** ***(Test de Mardia)*** *Sea N es el número total de observaciones; $\boldsymbol{z_i}$ y $\boldsymbol{z_j}$ los vectores columna con valores de todas las variables para las observaciones i-ésima y j-ésima, repectivamente; $\overline{\boldsymbol{z}}$ el vector columna de las medias muestrales y $\boldsymbol{S^{-1}}$ la matriz inversa de la matriz de varianzas-covarianzas muestral. Entonces, los tests multivariantes de curtosis y asimetría son:*

1. <u>*Test de asimetría*</u>

$$b_{1,p} = \left(\frac{1}{N^2}\right) \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{(\boldsymbol{z_i} - \overline{\boldsymbol{z}})^t \cdot \boldsymbol{S^{-1}} \cdot (\boldsymbol{z_j} - \overline{\boldsymbol{z}})\right\}^3. \tag{3.2}$$

2. <u>*Test de curtosis*</u>

$$b_{1,p} = \left(\frac{1}{N}\right) \cdot \sum_{i=1}^{N} \left\{(\boldsymbol{z_i} - \overline{\boldsymbol{z}})^t \cdot \boldsymbol{S^{-1}} \cdot (\boldsymbol{z_i} - \overline{\boldsymbol{z}})\right\}^2. \tag{3.3}$$

[1] Por ser definida positiva, la forma cuadrática $(\boldsymbol{x} - \boldsymbol{\mu}) \cdot \sum^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^T > 0$ y dado que es simétrica y definida positiva, $|\sum| > 0$

Los estadísticos de contraste $W(b_{1,p})$ y $W(b_{2,p})$ [2] obtenidos a partir de (3.2) y (3.3) se distribuyen asintóticamente según una distribución normal estándar. Además, podemos realizar un contraste conjunto de simetría y mesocurtosis multivariante mediante el siguiente estadístico:

**Definición 3.3.** ***(Estadístico de Mardia y Foster)*** *El test general verificando las hipótesis de no asímetría o exceso de curtosis*

$$K^2 = W(b_{1,p})^2 + W(b_{2,p})^2 \tag{3.4}$$

*se aproxima a una distribución $\mathcal{X}^2$ con 2 grados de libertad. Donde $W(b_{1,p})$ y $W(b_{2,p})$ son las transformaciones de Wilson-Hilferty de $b_{1,p}$ y $b_{2,p}$.*

Ante la dificultad de obtener datos distribuidos según una normalidad multivariante, la práctica más común reside en comprobar que los valores de asimetría y curtosis de cada variable no superan los valores 3 y 10 respectivamente (Curran, West y Finch, 1996).

## *Tamaño de la muestra*

A día de hoy siguen existiendo discrepancias debido al tamaño mínimo que debe tener una muestra. Desde 1990, gran cantidad de investigadores han propuesto un tamaño mínimo de 10 registros por cada variable observada (Jayaram, Kannan & Tan, 2004), mientras que otros sugieren un ratio de 5 registros por parámetro libre[3] (Bentler, 1989). Goodhue, et al. (2006, 2007) estudiaron la regla de los 10 mediante simulaciones de Monte Carlo para comparar muestras de tamaño 40, 90, 150 y 200. Sin embargo, los resultados fueron poco concluyentes:

> «*In fact, for simple [SEM] models with normally distributed data and relatively reliable measures, none of the three techniques have adequate power to detect small or medium effects at small sample sizes ... These findings run counter to extant suggestions in MIS literature*»

Se deduce así que no existe una regla general para la elección del tamaño de la muestra. Si bien es cierto, rara vez se considerarán muestras con $N < 200$. Para finalizar, destacar que Marsh et al (1988, 1996, 1998) llevaron a cabo 35 000 simulaciones de Monte Carlo haciendo uso de LISREL y obtuvieron que el tamaño adecuado para una muestra debe ser:

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

donde $r$ es el número de indicadores por variable latente.

[2] Estos estadísticos, así como su obtención, pueden ser consultados en la página 424 de [7].
[3] Consultar definición 4.13.

# 4

# Construcción del modelo SEM

## 4.1 Especificación del modelo

Comenzaremos la construcción del modelo mediante la obtención de una estructura que identifique las relaciones entre variables latentes exógenas y endógenas. Dicha estructura, llamada modelo estructural, consistirá en un conjunto de hipótesis acerca de las relaciones direccionales entre constructos. En segundo lugar, construiremos los modelos de medida que indicarán cómo medir las variables latentes.

### *Modelo estructural*

**Definición 4.1.** ***(Modelo estructural)*** *La formulación matricial del modelo estructural viene dada por:*

$$\eta = \boldsymbol{B}\eta + \boldsymbol{\Gamma}\xi + \zeta\,, \tag{4.1}$$

*donde $\eta$ es un vector (p x 1) de variables endógenas latentes; $\xi$ es un vector (q x 1) de variables exógenas latentes; $\boldsymbol{B}$ es una matriz (q x q) de coeficientes que relacionan variables latentes endógenas; $\boldsymbol{\Gamma}$ es una matriz (p x q) de coeficientes $\Gamma_{ij}$ de saturación de las variables latentes endógenas con exógenas; $\zeta$ es un vector (q x 1) de errores o términos de perturbación.*

**Definición 4.2.** ***(Condiciones modelo estructural)*** *Todo modelo satisfaciendo (4.1) cumple las siguientes condiciones:*

$$E(\eta) = 0, \quad E(\xi) = 0, \quad E(\zeta) = 0, \quad COV(\zeta, \xi) = 0, \tag{4.2}$$

*donde $E(\eta) = 0$ y $E(\xi) = 0$ indican que $\eta_i$ en $\eta$ y $\xi_i$ en $\xi$ entran desviadas de su media. Esto es, si $\eta_i^*$ y $\xi_i^*$ son las variables originales, entonces $\eta_i = \eta_i^* - E(\eta_i^*)$ para todo $i = 1,\ldots,m$, y $\xi_i = \xi_i^* - E(\xi_i^*)$ para todo $i = 1,\ldots,n$.*

**Proposición 4.1.** ***(Forma reducida)*** *Todo modelo estructural satisfaciendo $|\boldsymbol{I}-\boldsymbol{B}| \neq 0$ puede expresarse como:*

$$\eta = (\boldsymbol{I}-\boldsymbol{B})^{-1} \cdot (\boldsymbol{\Gamma}\xi + \zeta), \tag{4.3}$$

*donde los elementos de las matrices $\boldsymbol{B}$ y $\boldsymbol{\Gamma}$ podrán ser fijados con la intención de identificar el modelo.*

Asumiremos además que los $\zeta_i$ con $i = 1,\ldots,m$ son homocedásticos y no autocorrelacionados. Es decir, para una misma ecuación las perturbaciones asociadas a dos observaciones no están correlacionadas y tienen la misma varianza. Sin embargo, hemos de tener en cuenta que los SEM están descritos por múltiples ecuaciones en las que las perturbaciones de dos ecuaciones distintas pueden estar correlacionadas y tener varianzas distintas.

**Definición 4.3.** ***(Matriz de covarianzas entre variables latentes exógenas)*** *Para el modelo estructural (4.1) se define*

$$\boldsymbol{\Phi} = (\phi_{ij})_{nxn} = E(\xi\xi^{\boldsymbol{t}}) \tag{4.4}$$

*como una matriz n x n de covarianzas entre variables latentes exógenas.*

**Definición 4.4.** *(Matriz de covarianzas entre perturbaciones) Para el modelo estructural* (4.1) *se define*

$$\boldsymbol{\Psi} = (\psi_{ij})_{mxm} = E(\zeta\zeta^t) \tag{4.5}$$

*como una matriz $m \times m$ de covarianzas entre los errores de las variables endógenas latentes. Si dichos errores no están correlacionados, entonces $\boldsymbol{\Psi}$ es diagonal.*

## *Modelo de medida*

El modelo de medida consta de al menos una variable latente definida mediante indicadores (o variables observadas) susceptibles a errores de medida. Esto es, los modelos de medida son aquellos que relacionan variables latentes con variables observadas. De hecho, todo modelo de medida consta de tantas ecuaciones como variables observadas haya. Destacar además que diferenciaremos entre un sistema de ecuaciones que definirá las variables exógenas y otro que hará lo propio con las variables endógenas.

**Definición 4.5.** ***(Modelo de medida)*** *La formulación matricial del modelo de medida viene dada por:*

$$\left.\begin{array}{ll} \boldsymbol{x} & = \boldsymbol{\Lambda}_x \xi + \delta \\ \boldsymbol{y} & = \boldsymbol{\Lambda}_y \eta + \varepsilon \end{array}\right\}, \tag{4.6}$$

*donde $\boldsymbol{x}$ e $\boldsymbol{y}$ son dos vectores (q x 1) y (p x 1) de variables observables; $\boldsymbol{\Lambda}_x$ y $\boldsymbol{\Lambda}_y$ son dos matrices de orden (q x n) y (p x m) que contienen las saturaciones de las variables observables en las variables exógenas y endógenas respectivamente; $\xi$ es un vector (n x 1) de variables exógenas; $\eta$ es un vector (m x 1) de variables endógenas; $\delta$ y $\varepsilon$ son dos vectores (q x 1) y (p x 1) de errores de medida.*

**Definición 4.6.** ***(Condiciones modelo de medida)*** *Todo modelo satisfaciendo* (4.5) *cumple las siguientes condiciones:*

$$E(\eta) = 0, \quad E(\xi) = 0, \quad E(\varepsilon) = 0, \quad E(\delta) = 0. \tag{4.7}$$

*Además, $\varepsilon$ y $\delta$ no están correlacionados entre ellos mismos ni con $\xi$ y $\eta$.*

**Definición 4.7.** ***(Matriz de covarianzas entre variables exógenas observables)*** *Para el modelo de medida* (4.5) *se define*

$$\boldsymbol{\Theta}_\delta = (\theta_{ij})_{qxq} = E(\delta\delta^t) \tag{4.8}$$

*como la matriz $(q \times q)$ de covarianzas entre variables exógenas observables.*

**Definición 4.8.** ***(Matriz de covarianzas entre variables endógenas observables)*** *Para el modelo de medida* (4.5) *se define*

$$\boldsymbol{\Theta}_\varepsilon = (\theta_{ij})_{pxp} = E(\varepsilon\varepsilon^t) \tag{4.9}$$

*como una matriz (p x p) de covarianzas entre las variables endógenas observables. Si los errores no están correlacionados, entonces $\boldsymbol{\Theta}_\delta$ y $\boldsymbol{\Theta}_\varepsilon$ son matrices diagonales.*

La definción de un buen modelo de medida previo es imprescindible para la construcción de un modelo estructural. Para llevar a cabo dicha labor haremos uso del análisis factorial.

## *Matriz de covarianzas implicada*

**Teorema 4.1.** *La matriz de covarianzas implicada para el caso general satisfaciendo*

$$\sum = \sum(\theta) = \begin{pmatrix} \sum_{yy}(\theta) & 0 \\ \sum_{xy}(\theta) & \sum_{xx}(\theta) \end{pmatrix}, \tag{4.10}$$

*donde* $\sum_{yy}(\theta)$, $\sum_{xy}(\theta)$ *y* $\sum_{xx}(\theta)$ *corresponden a las matrices de covarianza de los respectivos subíndices, viene dada por:*

$$\sum(\theta) = \begin{pmatrix} \Delta_y(I-B)^{-1}(\Gamma\Phi\Gamma^t + \Psi)[(I-B)^{-1}]^t\Delta_y^t + \Theta_\varepsilon & \Delta_y(I-B)^{-1}\Gamma\Phi\Delta_x^t \\ \Delta_x\Phi\Gamma^t[(I-B)^{-1}]^t\Delta_y^t & \Delta_x\Phi\Delta_x^t + \Theta_\delta \end{pmatrix}. \tag{4.11}$$

Demostración:
Si consideramos en primer lugar la matriz de covarianzas implicada $\sum_{yy}(\theta)$ tenemos que:

$$\begin{aligned} \sum_{yy}(\theta) &= E(yy^t) \\ &= E[(\Delta_y\eta + \varepsilon)(\Delta_y\eta + \varepsilon)^t] \\ &= E[(\Delta_y\eta + \varepsilon)(\eta^t\Delta_y^t + \varepsilon^t)] \\ &= \Delta_y E(\eta\eta^t)\Delta_y^t + \Theta_\varepsilon. \end{aligned} \tag{4.12}$$

Obsérvese que haciendo uso de (4.3) tenemos que:

$$\sum_{yy}(\theta) = \Delta_y(I-B)^{-1}(\Gamma\Phi\Gamma^t + \Psi)[(I-B)^{-1}]^t\Delta_y^t + \Theta_\varepsilon. \tag{4.13}$$

En cuanto a la matriz de covarianza implicada de y y x, $\sum_{yx}(\theta)$:

$$\begin{aligned} \sum_{yx}(\theta) &= E(yx^t) \\ &= E[(\Delta_y\eta + \varepsilon)(\xi^t\Delta_x^t + \delta^t)] \\ &= \Delta_y E(\eta\xi^t)\Delta_x^t. \end{aligned} \tag{4.14}$$

De nuevo, haciendo uso de (4.3) se prueba que:

$$\sum_{yx}(\theta) = \Delta_y(I-B)^{-1}\Gamma\Phi\Delta_x^t \tag{4.15}$$

y

$$\sum_{xx}(\theta) = \Delta_x\Phi\Delta_x^t + \Theta_\delta. \tag{4.16}$$

Ahora bien, haciendo uso de las ecuaciones (4.13), (4.15) y (4.16), la matriz de covarianza implicada de x e y es:

$$\sum(\theta) = \begin{pmatrix} \Delta_y(I-B)^{-1}(\Gamma\Phi\Gamma^t + \Psi)[(I-B)^{-1}]^t\Delta_y^t + \Theta_\varepsilon & \Delta_y(I-B)^{-1}\Gamma\Phi\Delta_x^t \\ \Delta_x\Phi\Gamma^t[(I-B)^{-1}]^t\Delta_y^t & \Delta_x\Phi\Delta_x^t + \Theta_\delta \end{pmatrix}. \tag{4.17}$$

Llegando así a la matriz implicada por el modelo estructural general. Hemos hecho uso de que $\sum_{xy}(\theta) = [\sum_{yx}(\theta)]^t$.

■

**Corolario 4.1.1.** *La matriz de covarianzas implicada para el modelo de variables observadas satisfaciendo* (4.10) *viene dada por:*

$$\sum(\theta) = \begin{pmatrix} (\boldsymbol{I}-\boldsymbol{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^t + \boldsymbol{\Psi})[(\boldsymbol{I}-\boldsymbol{B})^{-1}]^t & (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}^t[(\boldsymbol{I}-\boldsymbol{B})^{-1}]^t & \boldsymbol{\Phi} \end{pmatrix}. \tag{4.18}$$

Demostración:
El modelo de ecuaciones estructurales con variables observadas no es más que una simplificación del caso general. Esto se debe a que el modelo de variables observadas asume que todos los errores son nulos. Llegamos así a que $\mathbf{y} = \eta$ y que $\mathbf{x} = \xi$. Ahora bien, sustituyendo $\boldsymbol{\Theta}_\varepsilon = 0$, $\boldsymbol{\Theta}_\delta = 0$, $\boldsymbol{\Lambda}_y = \boldsymbol{I}_p$ y $\boldsymbol{\Lambda}_x = \boldsymbol{I}_q$ en (4.17) se concluye el resultado. ■

**Corolario 4.1.2.** *La matriz de covarianzas implicada para el modelo de análisis factorial confirmatorio satisfaciendo* (4.10) *viene dada por:*

$$\sum(\theta) = \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x^t + \boldsymbol{\Theta}_\delta. \tag{4.19}$$

Demostración:
Como el modelo factorial confirmatorio no plantea hipótesis acerca de las relaciones direccionales entre $\eta$ y $\xi$, entonces $\mathbf{B}$ y $\boldsymbol{\Gamma}$ son nulas. De igual modo, al no haber variables latentes endógenas, los términos $\boldsymbol{\Lambda}_y$, $\boldsymbol{\Theta}_\varepsilon$ y $\boldsymbol{\Psi}$ también lo son. Por tanto:

$$\begin{aligned} \sum(\theta) &= E(\boldsymbol{x}\boldsymbol{x}^t) \\ &= E[(\boldsymbol{\Lambda}_x\xi + \delta)(\xi^t\boldsymbol{\Lambda}_x^t + \delta^t)] \\ &= \boldsymbol{\Lambda}_x E(\xi\xi^t)\boldsymbol{\Lambda}_x^t + \boldsymbol{\Theta}_\delta \\ &= \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x^t + \boldsymbol{\Theta}_\delta. \end{aligned} \tag{4.20}$$

■

## *Análisis factorial*

El análisis factorial es una técnica estadística multivariante de reducción de dimensionalidad que se usa para explicar las covarianzas y correlaciones entre variables observadas a partir de una número menor de variables latentes. Distinguiremos entre *análisis factorial exploratorio* (EFA) y *análisis factorial confirmatorio* (CFA).

El análisis factorial exploratorio no parte de un modelo prefijado, por lo que el número de variables latentes no está predefinido. Generalmente, todas las variables latentes afectan a todas las variables observadas, los errores de medida no están correlacionados y la infraespecificación de parámetros es algo común. Por contra, el análisis factorial confirmatorio si parte de un modelo dado, por lo que el número de variables latentes será fijado con anterioridad por el analista. Además, toda influencia de una variable latente en una variable observada ha debido ser también previamente especificada. Destacar que algunos efectos directos de variables latentes son fijados o simplemente tienen valor nulo. Los errores de medida pueden estar correlacionados y las covarianzas de las variables latentes pueden ser estimadas, siendo la identificación

de parámetros necesaria.

Para el desarrollo de este trabajo se hará uso del análisis factorial con la intención de proponer y validar los constructos teóricamente establecidos. En esta sección se tratará de realizar una breve descripción de los principales resultados usados, aunque debido a la brevedad de este documento, se propone al lector consultar [15] o [4]. Siguiendo lo dicho por Williams, Onsman y Brown en 2010, todo análisis factorial debe comprender las siguientes fases:

**¿Es la muestra apropiada para realizar un análisis factorial?**

En primer lugar, siguiendo lo establecido en el capítulo anterior, comprobaremos que el tamaño de la muestra satisface el mínimo requerido. Por otro lado, Henson y Roberts establecieron que debemos inspeccionar la matriz de correlaciones en busca de coeficientes mayores a 0.3, ya que, de lo contrario, el análisis factorial carece de sentido. Esta opinión fue respaldada por Hair et al. (1995) mediante la elaboración de una regla por la cual se establecen dichos coeficientes como ±0.3 = mínimo, ±0.4 = importante, ±0.5 = significativo. En tercer lugar, previa extracción de los factores, debemos confirmar que la muestra es válida para realizar el análisis factorial. Existen gran cantidad de test para llevar a cabo este proceso. Sin embargo, en este documento se mencionarán solo cuatro de ellos: test de adecuación muestral de Kaiser-Meyer-Olkin (KMO), el test MSA, la prueba de Esfericidad de Bartlett y el test de Fligner-Killeen .

El test KMO nos proporciona un indicador de la aplicabilidad de una análisis factorial. Valores altos indican alta adecuación de un AF.

**Definición 4.9.** ***(Test de Keiser-Meyer Olkin)***

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2}, \tag{4.21}$$

*donde $r_{ij}$ son los coeficientes de correlación y $a_{ij}$ los coeficientes de correlación parcial.*

La interpretación del test KMO puede verse reflejada en la siguiente tabla:

| Valor de $\mathcal{X}^2/df$ | Ajuste |
|---|---|
| $0{,}9 < KMO \leq 1$ | Muy bueno |
| $0{,}8 < KMO \leq 0{,}9$ | Meritorio |
| $0{,}7 < KMO \leq 0{,}8$ | Mediano |
| $0{,}6 < KMO \leq 0{,}7$ | Mediocre |
| $0{,}5 < KMO \leq 0{,}6$ | Bajo |
| $KMO \leq 0{,}5$ | Inaceptable |

Cuadro 4.1: Ajuste del estadístico KMO

El test MSA es un coeficiente muy parecido al KMO pero proporciona una medida individual de cada variable. Si en alguna variable el valor es bajo, podemos considerar eliminarla del análisis:

**Definición 4.10.** ***(Test MSA)***

$$MSA_i = \frac{\sum_{j\neq i} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} a_{ij}^2}, \tag{4.22}$$

*donde $r_{ij}$ son los coeficientes de correlación y $a_{ij}$ los coeficientes de corrlación parcial.*

La prueba de esfericidad de Bartlett contrasta la hipótesis nula de que la matriz de correlaciones es una matriz identidad, en cuyo caso no existirían correlaciones significativas entre las variables y el modelo factorial no tendría cabida.

**Definición 4.11.** ***(Prueba de esfericidad de Bartlett)*** *Dado el contraste de hipótesis:*

$$\begin{aligned} &H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2, \\ &H_1 : \sigma_i^2 \neq \sigma_j^2, \end{aligned} \tag{4.23}$$

*se define el estadístico:*

$$T = \frac{(N-k)\cdot ln(s_p^2) - \sum_{i=1} k(N_i - 1)\cdot ln(s_i^2)}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i-1)) - 1/(N-k))}, \tag{4.24}$$

*donde $s_i^2$ es la varianza del término i-ésimo, N es tamaño de la muestra, $N_i$ es el tamaño de la muestra del grupo i-ésimo, k es el número de grupos, y $s_p^2$ es la varianza agrupada* [1].

Este estadístico sigue una distribución $\mathcal{X}^2$ con k-1 grados de libertad, siendo muy sensible a la falta de normalidad. Cuando trabajemos con muestras no normales, haremos uso del Test de Fligner-Killen, prueba no paramétrica que presume de robustez frente a dichas desviaciones (Mandasky, 1988). Mencionar que este estadístico sigue, bajo $H_0$, una distribución $\mathcal{X}_{k-1}^2$.

**Definición 4.12.** ***(Test de Fligner-Killen)*** *Dado el contraste de hipótesis (4.23), se define el estadístico:*

$$FK = \frac{\sum_{i=1}^{k} n_i(\overline{\alpha_i} - \overline{\alpha})^2}{\sum_{j=1}^{N}(\alpha_{N,j} - \overline{\alpha})^2/(n-1)}, \tag{4.25}$$

*siendo $n_i$ el tamaño de la i-ésima muestra; $\alpha_{N,j} = \Phi^{-1}\left(\frac{1}{2} + \frac{j}{2(N+1)}\right)$ para todo $j = 1,\ldots,N$ donde $\Phi^{-1}(p)$ es el percentil 100p de la distribución N(0,1); $\alpha_i = \sum_{j\in G_i} \frac{\alpha_{N,j}}{n_i}$ donde $G_i$ denota la muestra de la población i y $\overline{\alpha} = \sum_{j=1}^{N} \frac{\alpha_{N,j}}{N}$.*

**Extracción de factores**

El objetivo de la extracción consiste en encontrar los coeficientes del modelo o cargas factoriales. Partimos de que ni la matriz de cargas ni los factores son magnitudes observables lo que plantea un problema de indeterminación:

- Un conjunto de factores puede explicarse con la misma precisión tanto con factores correlacionados como incorrelacionados.

[1] Se define la varianza agrupada como la media de las varianzas del grupo: $s_p^2 = \sum_{i=1}^{k}(N_i-1)\cdot s_i^2/(N-k)$.

- Los factores no pueden determinarse de forma única: el modelo factorial es invariante frente a rotaciones ortogonales.

Existen numerosos métodos para realizar dicha extracción: análisis de componentes principales, máxima verosimilitud, mínimos cuadrados generalizados, mínimos cuadrados ponderados, etc... En el caso práctico a desarrollar en este trabajo, se hará uso de la factorización de ejes principales.

Ahora bien, para establecer el número de factores a extraer tendremos en cuenta el *criterio de Kaiser para autolavores mayores que 1* y el *gráfico de sedimentación*. El criterio establecido por Kaiser, como su propio nombre indica, consiste en establecer tantos factores como autovalores de la matriz de varianzas-covarianzas existen y son mayores que 1. Por otro lado, el gráfico de sedimentación (Cattell, 1966) consiste en una representación gráfica del tamaño de los autovalores. Buscaremos el punto de inflexión donde los autovalores dejan de explicar una cantidad significativa de varianza.

**Rotación de factores**

Otra consideración que debemos hacer a la hora de decidir cuántos factores vamos a analizar es si las variables pueden estar relacionadas con más de un factor. La rotación maximiza las cargas factoriales de unos items y minimiza las de otros, obteniendo así una solución más simple de interpretar. Diferenciaremos dos tipos principales de rotación: *rotación ortogonal* o *rotación oblicua*.

- Rotación ortogonal: las rotaciones respetan los ángulos y los factores permanecen incorrelados. En este caso, las cargas factoriales se interpretan como correlaciones entre factores y las variables.
- Rotación oblícua: los factores están correlados y no pueden interpretarse independientemente. Solo se utilizarán estas rotaciones si no obtenemos buenos resultados con una rotación ortogonal.

En el caso que nos compete, haremos uso, o bien de ninguna rotación, o de la rotación ortogonal *varimax* (Kaiser, 1958).

## *Fiabilidad de los modelos de medida*

Todo estudio en el que exista una componente de medida debe preocuparse por la fiabilidad de la medición realizada. Para comprobar dicha precisión o fiabilidad, haremos uso del estadístico Alpha de Cronbach propuesto en [9]. Otros estadísticos más restrictivos como el *Método test-reset*, *alternative form* o *split-halves* pueden ser consultados en [7].

**Definición 4.13.** *Supuestas dos medidas* $x_i = \alpha_i + \tau_i + e_i$ *y* $x_j = \alpha_j \tau_j + e_j$, *donde* $e_i$ *y* $e_j$ *no están correlacionados y* $\tau_i = \tau_j$.

1. *Si* $\alpha_i = \alpha_j = 1$ *y* $VAR(e_i) = VAR(e_j)$, *entonces* $x_i$ *y* $x_j$ *son* ***medidas paralelas***.
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 4.14.** ***(Fiabilidad)*** *La fiabilidad o consistencia de una medida $p_{x_i,x_j}$ equivale a*

$$p_{x_i,x_j} = \frac{\alpha_i^2 VAR(\tau_i)}{VAR(x_i)}. \tag{4.26}$$

**Proposición 4.2.** *Dada la fiabilidad de una medida $p_{x_i,x_j}$, se cumple que:*

$$p^2_{x_i,\tau_i} = p_{x_i,x_j}. \tag{4.27}$$

Demostración:

$$p^2_{x_i,\tau_i} = \frac{[COV(x_i,\tau_i)]^2}{VAR(x_i)VAR(\tau_i)} = \frac{\alpha_i^2[VAR(\tau_i)]^2}{VAR(x_i)VAR(\tau_i)} = \frac{\alpha_i^2 VAR(\tau_i)}{VAR(x_i)} = p_{x_i x_j}. \tag{4.28}$$

■

El alpha de Cronbach surge ante la necesidad de paliar las críticas del método *Split-halves*, siendo actualmente el método de estimación de fiabilidad más empleado en psicometría. Este, medirá la fiabilidad de una suma de medidas paralelas o tau-equivalentes.

**Lema 4.1.** ***(Alfa de Cronbach)*** *El coeficiente de fiabilidad de escala Alfa de Cronbach viene dado por:*

$$\alpha = \left(\frac{q}{q-1}\right)\left(1 - \frac{\sum_{i=1}^{q} VAR(x_i)}{VAR(H)}\right), \tag{4.29}$$

*donde q es el número de items de la escala.*

Demostración:
Sea $H = \sum_{i=1}^{q} x_i = H$, donde $x_i$ son las variables observadas de un modelo. Partiendo de (4.27):

$$p^2_{\tau_1,H} = \frac{[COV(\tau_1,H)]^2}{VAR(\tau_1)VAR(H)} = \frac{[COV(\tau_1, x_1 + x_2 + \cdots + x_q)]^2}{VAR(\tau_1)VAR(H)} = \frac{[COV(\tau_1, q\tau_1 + \sum_{i=1}^{q} e_i)]^2}{VAR(\tau_1)VAR(H)}$$
$$= \frac{[qVAR(\tau_1]^2}{VAR(\tau_1)VAR(H)} = \frac{q^2 VAR(\tau_1)}{VAR(H)} = p_{HH}.$$

Reformulando lo obtenido en la ecuación anterior, llegamos a que el alfa de Cronbach vendrá dada por:

$$p^2_{HH} = \frac{q^2 VAR(\tau_1)}{VAR(H)} = \frac{q(q-1)qVAR(\tau_1)}{(q-1)VAR(H)} = \left(\frac{q}{q-1}\right)\left(\frac{q^2 VAR(\tau_1) - qVAR(\tau_1)}{VAR(H)}\right)$$
$$= \left(\frac{q}{q-1}\right)\left(\frac{q^2 VAR(\tau_1) + \sum_{i=1}^{q} VAR(e_i) - qVAR(\tau_1) + \sum_{i=1}^{q} VAR(e_i)}{VAR(H)}\right)$$
$$= \left(\frac{q}{q-1}\right)\left(\frac{VAR(H) - [qVAR(\tau_i) + \sum_{i=1}^{q} VAR(e_i)]}{VAR(H)}\right)$$
$$= \left(\frac{q}{q-1}\right)\left(1 - \frac{\sum_{i=1}^{q} VAR(x_i)}{VAR(H)}\right).$$

■

Se podría comprobar además que el alfa de Cronbach proporciona un límite inferior para la fiabilidad en el caso de mediciones congenéricas. Las reglas mediante las que se regirá este coeficiente son:

| Valor de $\alpha$ | Ajuste |
|---|---|
| $0{,}9 < \alpha < 1$ | Muy buen ajuste |
| $0{,}8 < \alpha < 0{,}9$ | Buen ajuste |
| $0{,}7 < \alpha < 0{,}8$ | Ajuste aceptable |
| $0 < \alpha < 0{,}7$ | Ajuste insuficiente |

Cuadro 4.2: Reglas del Alfa de Cronbach.

## 4.2 Identificación del modelo

La etapa de identificación de parámetros consiste en ver si la matriz de covarianzas de las variables observadas contiene suficiente información como para que, a la hora de estimar los parámetros, no haya inconsistencias.

**Definición 4.15.** *Un parámetro de $\theta$ está identificado si puede escribirse como función de uno o varios elementos no redundantes de $\sum$ y esta función conduce a una solución única. Si todos los parámetros de $\theta$ están identificados, entonces el modelo está identificado.*

A la hora de identificar un modelo nos encontraremos principalmente con dos problemas. En primer lugar, en los modelos que implican una mayor complejidad a la hora de relacionar sus variables latentes y observadas, podemos caer en la infraespecificación de variables. En segundo lugar, debemos centrarnos en la consistencia del modelo ante grandes volúmenes de datos y distribuciones normales multivariantes.

**Definición 4.16.** ***(Tipos de identificación)*** *Distinguiremos tres tipos de identificación dependiendo de si la cantidad de elementos no reduntantes en la matriz de covarianzas de las variables observadas es menor, igual o mayor que el número de parámetros desconocidos en $\theta$:*

1. *Modelo infraidentificado ($t > n(n+1)/2$): la información contenida en $\sum$ no es suficiente para estimar los parámetros.*
2. *Modelo saturado ($t = n(n+1)/2$): la información contenida en $\sum$ es suficiente para la estimación de los parámetros. La solución del sistema $\sum(\theta) = \sum$ es única.*
3. *Modelo sobreidentificado ($t < n(n+1)/2$): tenemos más información en $\sum$ de la necesitada. El sistema determinado por $\sum(\theta) = \sum$ tiene infinitas soluciones. Es la solución óptima.*

*Donde t es el número de parámetros y n el número de variables observadas.*

Ahora bien, antes de poder continuar con la identificación del modelo, necesitamos establecer los tipos de parámetros presentes en el mismo:

**Definición 4.17.** ***(Tipos de parámetros)***

1. *Parámetro libre: Parámetro desconocido no restringido para ser estimado.*

2. *Parámetro fijo: Parámetro conocido al que se le agina previamente un valor dado.*

3. *Parámetro restringido: Parámetro que al estimarse, adquiere el mismo valor que otro parámetro no fijo, o que puede escribirse en función de otros parámetros libres.*

**Definición 4.18.** ***(Parámetros del modelo)***

1. *Varianzas y covarianzas de variables exógenas.*

2. *Coeficientes que conectan las variables latentes con sus indicadores.*

3. *Coeficientes de regresión entre variables observadas o latentes.*

**Nota:** Las varianzas y covarianzas entre variables endógenas y las covarianzas entre variables endógenas y exógenas no son parámetros del modelo.

## *Infraidentificación en modelos de ecuaciones estructurales*

Si un modelo está infraespecificado no podemos confiar en la estimación de sus parámetros, errores o en su test chi-cuadrado. Encontraremos principalmente dos tipos de infraespecificación: estructural y empírica. Diremos que un modelo está infraespecificado estructuralmente cuando el problema resida en su propia construcción. Por contra, diremos que un modelo está infraespecificado empíricamente cuando el problema se encuentre en los valores particulares encontrados en los datos.

En 1989, Bollen confeccionó varias reglas mediante las cuales poder juzgar si un modelo está estructuralmente definido. Estas reglas suponen algunas restricciones previas para garantizar que el modelo posee una escala de medida. Generalmente, asumiremos que, o bien algunos coeficientes estructurales $\lambda_{ij}$ son fijos con valor constante 1, o bien que la varianza de algunas variables latentes vale 1.

### *T-Rule*

**Teorema 4.2.** ***(T-Rule)*** *Sea $t$ el número de parámetros en $\boldsymbol{\theta}$, $p$ el número de variables de $y$ y $q$ el número de variables de $\boldsymbol{x}$. Una condición necesaria para la identificación de los parámetros consiste en:*

$$t \leq \frac{(p+q)(p+q+1)}{2}. \tag{4.30}$$

Esta condición es necesaria pero no suficiente, ya que puede dar lugar a modelos infraidentificados. Para aportar la suficiencia de la que precisa esta regla, haremos uso de la comúnmente conocida como regla de los dos pasos:

*Regla de los dos pasos*

En primer lugar, obviamos las relaciones entre variables latentes para estudiar la correcta identificación del modelo de medida. El modelo queda reformulado como un análisis factorial confirmatorio y son aplicables las siguientes reglas:

**Proposición 4.3.** ***(Regla de los tres indicadores)*** *Si además del teorema (4.2) se cumple que:*

1. *Hay al menos tres indicadores por cada variable latente y cada indicador es causado únicamente por una única variable latente. Es decir, cada fila de $\mathbf{\Lambda_x}$ tiene un único elemento distinto de cero.*

2. *No hay correlación entre errores. Equivalentemente, $\mathbf{\Theta_\delta}$ es diagonal.*

*Entonces el modelo de medida está correctamente identificado.*

**Proposición 4.4.** ***(Regla de los dos indicadores)*** *Si además del teorema (4.2) se cumple que:*

1. *Hay más de una variable latente y a cada una le corresponden dos indicadores.*

2. *Cada indicador es causado únicamente por una variable latente y cada variable latente está correlada con al menos otra variable latente. Es decir, cada fila de $\mathbf{\Lambda_x}$ tiene un único elemento distinto de cero y ningún elemento en $\mathbf{\Phi}$ es nulo.*

3. *No hay correlación entre errores. Equivalentemente, $\mathbf{\Theta_\delta}$ es diagonal.*

*Entonces el modelo de medida está correctamente identificado.*

En segundo lugar, una vez hemos verificado la identificación del modelo de medida, pasamos a comprobar lo propio con el modelo estructural. Consideremos, por ejemplo, un modelo de múltiples ecuaciones en el que ninguna variable endógena afecte a otra variable endógena. Entonces, la matriz B será nula.

**Teorema 4.3.** ***(Regla de B nula)*** *Si el modelo de variables observadas no formula relaciones direccionales entre variables endógenas, entonces la matriz $\boldsymbol{B}$ es nula, siendo esta una condición suficiente para la identificación del modelo.*

Demostración:

Para establecer la identificación de un modelo donde B es cero, veremos que los parámetros desconocidos $\mathbf{\Gamma}$, $\mathbf{\Phi}$ y $\mathbf{\Psi}$ se pueden expresar en función de los parámetros de $\sum$. Si sustituimos B=0 en la ecuación (4.17) llegamos a que:

$$\begin{pmatrix} \sum_{yy} & \sum_{yx} \\ \sum_{xy} & \sum_{xx} \end{pmatrix} = \begin{pmatrix} (\mathbf{\Gamma\Phi\Gamma}^t + \mathbf{\Psi}) & \mathbf{\Gamma\Phi} \\ \mathbf{\Phi\Gamma}^t & \mathbf{\Phi} \end{pmatrix}, \tag{4.31}$$

de donde se obtiene que $\sum_{xx} = \mathbf{\Phi}$, por lo que $\mathbf{\Phi}$ está definido. Ahora bien,

$$\mathbf{\Phi\Gamma}^t = \sum_{xy} \Leftrightarrow \sum_{xx} \mathbf{\Gamma}^t = \sum_{xy} \Leftrightarrow \mathbf{\Gamma}^t = (\sum_{xx})^{-1} \cdot \sum_{xy}. \tag{4.32}$$

Lo que confirma que $\mathbf{\Gamma}$ se puede expresar en función de matrices de covarianzas, y por ende, que está identificada. Por último, realizando un proceso análogo para $\mathbf{\Psi}$:

$$\begin{aligned}\mathbf{\Psi} &= \sum_{yy} - \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^t \\ &= \sum_{yy} - \sum_{yx} \cdot (\sum_{xx})^{-1} \cdot \sum_{xx} \cdot (\sum_{xx})^{-1} \cdot \sum_{xy} \\ &= \sum_{yy} - \sum_{yx} \cdot (\sum_{xx})^{-1} \cdot \sum_{xy}.\end{aligned} \tag{4.33}$$

Llegamos así a que, en efecto, cuando $\mathbf{B} = 0$, los elementos $\mathbf{\Phi}$, $\mathbf{\Psi}$ y $\mathbf{\Gamma}$ pueden ser escritos en función de matrices de covarianza identificadas.

■

Por último destacaremos la existencia de una última condición suficiente, aunque no necesaria, para identifcar un modelo estructural:

**Teorema 4.4.** ***(Regla recursiva)*** *Si $\boldsymbol{B}$ es una matriz triangular inferior y $\mathbf{\Psi}$ una matriz diagonal entonces el modelo está correctamente identificado.*

Demostración:
Veamos que en efecto $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$ y $\mathbf{\Psi}$ están identificadas en modelos recursivos. Para ello usaremos una notación ligeramente distinta. La i-ésima ecuación de un modelo recursivo es:

$$y_i = [\boldsymbol{B}_i^t \mid \boldsymbol{\gamma}_i^t]\boldsymbol{z_i} + \boldsymbol{\zeta_i}, \tag{4.34}$$

donde $y_i$ es la variable dependiente y $\boldsymbol{\zeta_i}$ es la perturbación de la i-ésima ecuación. El vector $\boldsymbol{B}_i^t$ es la i-ésima fila de $\mathbf{B}$ quitando todos los valores nulos y dejando solo los parámetros libres; $\boldsymbol{\gamma}_i^t$ hace lo propio con $\mathbf{\Gamma}$; y $\boldsymbol{z_i}$ es el conjunto de variables en $\mathbf{x}$ e $\mathbf{y}$ que tienen efectos directos sobre $y_i$. Multiplicando a ambos lados de (4.34) por $z_i^t$ llegamos a que:

$$\boldsymbol{\sigma}_{y_i z_i}^{\boldsymbol{t}} = [\boldsymbol{B}_i^t \mid \boldsymbol{\gamma}_i^t]\sum_{z_i z_i} + \boldsymbol{\sigma}_{\zeta_i z_i}^{\boldsymbol{t}}, \tag{4.35}$$

donde $\boldsymbol{\sigma}_{y_i z_i}^{\boldsymbol{t}}$ es el vector de covarianzas de $y_i$ con las variables explicativas, $\sum_{z_i z_i}$ es la matriz no singular de covarianzas de $\boldsymbol{z_i}$, y $\boldsymbol{\sigma}_{\zeta_i z_i}^{\boldsymbol{t}}$ es el vector de covarianzas de $\boldsymbol{\zeta_i}$ con las variables explicativas de la i-ésima ecuación. Bajo el supuesto de que $\boldsymbol{\zeta_i}$ está incorrelado con las variables explicativas de la i-ésima ecuación, llegamos a que $\boldsymbol{\sigma}_{\zeta_i z_i}^{\boldsymbol{t}} = 0$. Luego, podemos expresar (4.35) como sigue:

$$[\boldsymbol{B}_i^t \mid \boldsymbol{\gamma}_i^t] = \boldsymbol{\sigma}_{y_i z_i}^{\boldsymbol{t}} \cdot \left(\sum_{z_i z_i}\right)^{-1}. \tag{4.36}$$

Como las covarianzas de las variables observadas están identificadas, entonces $[\boldsymbol{B}_i^t \mid \boldsymbol{\gamma}_i^t]$ también lo está. Ahora bien, para la identificación de $\mathbf{\Psi}_{ii}$, multiplicaremos a ambos lados de (4.34) por $y_i^t$ y tomaremos esperanzas. Llegamos así a que:

$$VAR(y_i) = [\boldsymbol{B}_i^t \mid \mathbf{\Gamma}_i^t]\sum_{z_i z_i}\left[\frac{\boldsymbol{B}_i}{\mathbf{\Gamma}_i}\right] + \mathbf{\Psi}_{ii}. \tag{4.37}$$

Despejando $\mathbf{\Psi}_{ii}$ y sustituyendo $[\boldsymbol{B}_i^t \mid \boldsymbol{\gamma}_i^t]$ por (4.36), se sigue que:

$$\mathbf{\Psi}_{ii} = VAR(y_i) - \sigma_{y_i z_i}^t \cdot \left( \sum_{z_i z_i} \right)^{-1} \sigma_{z_i y_i}. \tag{4.38}$$

Como $\mathbf{\Psi}_{ii}$ se puede expresar en función de variables identificadas, ella también lo estará. Llegamos así a que $\mathbf{B}$, $\mathbf{\Gamma}$ y $\mathbf{\Psi}$ están identificadas. De (4.31) se desprende además que $\mathbf{\Phi} = \sum_{xx}$. En definitiva, todos los parámetros de un modelo recursivo estarán correctamente identificados.

■

En definitiva, si la primera etapa muestra que los parámetros del modelo de medida están indentificados y la segunda etapa muestra lo propio en el modelo estructural, podemos afirmar que el modelo está identificado.

Existen además otras reglas similares a estas, como la condición de orden y la condición de rango. Sin embargo, exceden los propósitos de este sucinto documento. Si el lector quiere saber más acerca de ellas puede hacer uso de [7].

## 4.3 Estimación del modelo

La etapa de estimación deriva de la relación entre la matriz de covarianzas de las variables observadas y los parámetros del modelo. En la práctica, al no conocer ni disponer de los valores poblacionales de las varianzas y covarianzas, debemos estimar dichos valores mediante una muestra. Esto es, construiremos la matriz de covarianzas muestral de las variables observadas, $\mathbf{S}$, mediante una estimación de $\sum$. De este modo $\sum(\hat{\theta})$ y $\mathbf{S}$ serán lo más similares posible. Para que dicha relación sea certera, minimizaremos una función de la matriz de residuos $(S - \sum)$. Representaremos dicha función como $F(S, \sum(\theta))$, ya que compara los valores de la matriz de varianza-covarianza muestral con la matriz implicada desarrollada a partir de los parámetros del modelo. De este modo, diremos que toda función de ajuste satisface que:

$$F\left(S, \sum(\theta)\right) = \left(S - \sum(\theta)\right)^t \cdot W \cdot \left(S - \sum(\theta)\right). \tag{4.39}$$

Donde $W$ es la matriz de pesos para la matriz de residuos. Ahora bien, las principales propiedades que cumple toda función de ajuste son:

**Propiedades 4.1.** ***(Función de ajuste)***

1. $F(S, \sum(\theta))$ *es un escalar.*
2. $F(S, \sum(\theta)) \geq 0$.
3. $F(S, \sum(\theta)) = 0 \Leftrightarrow \sum(\theta) = S$.
4. $F(S, \sum(\theta))$ *es continua en* $S$ *y* $\sum(\theta)$.

En 1984, Browne demostraría que al minimizar una función que cumpla dichas propiedades se obtienen estimadores consistentes. Ahora bien, los métodos de estimación más usados en SEM son máxima verosimilitud (ML), mínimos cuadrados ordinarios (ULS) y minimos cuadrados generalizados (GLS) o mínimos cuadrados ponderados (WLS).

## *Estimación por máxima verosimilitud*

**Propiedades 4.2.** ***(Restricciones previas)*** *Para poder aplicar la estimación por máxima verosimilitud ha de verificarse que:*

1. $\sum(\boldsymbol{\theta})$ *y* $\boldsymbol{S}$ *son no singulares.*
2. $\boldsymbol{x}$ *e* $\boldsymbol{y}$ *siguen una distribución normal multivariante.*

**Definición 4.19.** ***(Función de ajuste*** $F_{ML}$***)*** *La función de ajuste, satisfaciendo las propiedades (4.2), para la estimación por máxima verosimilitud es:*

$$F_{ML} = log\left|\sum(\boldsymbol{\theta})\right| + tr\left(\boldsymbol{S}\cdot\left(\sum(\boldsymbol{\theta})\right)\right)^{-1} - log|\boldsymbol{S}| - (p+q). \tag{4.40}$$

Sea $N$ el tamaño de una muestra independientemente distribuida para una variable aleatoria $\mathbf{Z}$. La función de densidad para cada $Z_i$ $(i = 1, 2 \ldots N)$ vendrá dada por $f(Z_i;\boldsymbol{\theta})$ donde $\theta$ es fijo. Como $Z_i$ son independientes, entonces la densidad conjunta es el producto de las densidades marginales:

$$f(Z_1, Z_2, \ldots, Z_n;\boldsymbol{\theta}) = f(Z_1;\boldsymbol{\theta})\cdot f(Z_2;\boldsymbol{\theta})\cdot \ldots \cdot f(Z_n;\boldsymbol{\theta}). \tag{4.41}$$

Ahora bien,

$$L(\boldsymbol{\theta}; Z_1, Z_2, \ldots, Z_n) = L(\boldsymbol{\theta}; Z_1)\cdot L(\boldsymbol{\theta}; Z_2)\cdot \ldots \cdot L(\boldsymbol{\theta}; Z_n), \tag{4.42}$$

donde $L(\boldsymbol{\theta}; Z_i)$ es el valor de $f(Z_i;\boldsymbol{\theta})$ cuando $Z_i$ es un valor muestral.
Sea $\hat{\boldsymbol{\theta}}$ un estimador de $\boldsymbol{\theta}$. Como $\mathbf{x}$ e $\mathbf{y}$ son dos variables aleatorias distribuidas según una distribución normal multivariante, podemos expresarlas mediante un único vector $(p+q)\times 1$ de $\mathbf{z}$, donde $\mathbf{z}$ es su desviación. La función de densidad vendrá dada por:

$$f\left(\boldsymbol{z};\sum\right) = (2\pi)^{\frac{-(p+q)}{2}}\cdot\left|\sum\right|^{-1/2}\cdot exp\left[\left(\frac{-1}{2}\right)\cdot\boldsymbol{z}^t\cdot\left(\sum\right)^{-1}\cdot\boldsymbol{z}\right]. \tag{4.43}$$

Para una muestra aleatoria de tamaño N de observaciones independientes de $\mathbf{z}$, la densidad conjunta vendrá dada por:

$$f\left(\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_N};\sum\right) = f\left(\boldsymbol{z_1};\sum\right)\cdot f\left(\boldsymbol{z_2};\sum\right)\cdot\cdots\cdot f\left(\boldsymbol{z_n};\sum\right). \tag{4.44}$$

La función de verosimilitud quedará por tanto como:

$$L(\boldsymbol{\theta}) = (2\pi)^{\frac{-N(p+q)}{2}}\cdot\left|\sum(\boldsymbol{\theta})\right|^{-N/2}\cdot exp\left[\left(\frac{-1}{2}\right)\cdot\sum_{i=1}^{N}\boldsymbol{z_i^t}\cdot\left(\sum(\boldsymbol{\theta})\right)^{-1}\cdot\boldsymbol{z_i}\right], \tag{4.45}$$

donde hemos hecho uso de la hipótesis fundamental $\sum = \sum(\theta)$. Aplicando logaritmos llegamos a que:

$$logL(\boldsymbol{\theta}) = \frac{-N(p+q)}{2} \cdot log(2\pi) - \left(\frac{N}{2}\right) \cdot log\left|\sum(\boldsymbol{\theta})\right| - \frac{1}{2} \cdot \sum_{i=1}^{N} z_i^t \cdot \left(\sum(\boldsymbol{\theta})\right)^{-1} \cdot z_i. \tag{4.46}$$

Además:

$$\begin{aligned} -\frac{1}{2} \cdot \sum_{i=1}^{N} z_i^t \cdot \left(\sum(\boldsymbol{\theta})\right)^{-1} \cdot z_i &= -\frac{1}{2} \cdot \sum_{i=1}^{N} tr\left[z_i^t \cdot \left(\sum(\boldsymbol{\theta})\right)^{-1} \cdot z_i\right] \\ &= -\frac{N}{2} \cdot \sum_{i=1}^{N} tr\left[N^{-1} \cdot z_i^t \cdot \left(\sum(\boldsymbol{\theta})\right)^{-1} \cdot z_i\right] \\ &= -\frac{N}{2} \cdot tr\left[S^* \left(\sum(\boldsymbol{\theta})\right)^{-1}\right], \end{aligned} \tag{4.47}$$

con $S^*$ siendo el estimador máximo verosimil de la matriz de covarianzas.
Obsérve que si K es una constante, entonces podemos reescribir $logL(\theta)$ como sigue:

$$\begin{aligned} logL(\boldsymbol{\theta}) &= K - (\frac{N}{2}) \cdot log|\sum(\boldsymbol{\theta})| - (\frac{N}{2}) \cdot tr[S^*(\sum(\boldsymbol{\theta}))^{-1}] \\ &= K - (\frac{N}{2})\{log|\sum(\boldsymbol{\theta})| - tr[S^*(\sum(\boldsymbol{\theta}))^{-1}]\}. \end{aligned} \tag{4.48}$$

Si comparamos (4.48) con (4.40) podemos encontrar algunas diferencias. Sin embargo, ambas expresiones llevan al mismo estimador $\hat{\boldsymbol{\theta}}$, ya que en el caso que nos compete, los términos constantes $K = -log|\boldsymbol{S}| - (p+q)$ no influyen en su obtención. La presencia de $(-N/2)$ en (4.48) hará que minimicemos en lugar de maximizar dicha función. Por último, debemos tener en cuenta que $\boldsymbol{S}^* = [(N-1)/N] \cdot \mathbf{S}$, por lo que estas matrices serán idénticas en muestras grandes.

**Proposición 4.5.** *Todo estimador máximo verosimil, satisfaciendo las propiedades* (4.2)*, es consistente, eficiente e insesgado.*

La consistencia nos asegura que al aumentar el tamaño de la muestra, el estimador converge al verdadero valor del parámetro. Por otro lado, la eficiencia indica que el estimador $\hat{\boldsymbol{\theta}}$ tiene varianza mínima. Por último, al ser insesgado se tiene que $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.

Generalmente, obtener estos estimadores analíticamente puede resultar muy costoso. Es por ello que recurriremos a métodos numéricos para estimarlos computacionalmente. Algunos de estos métodos pueden ser Newton-Raphson o Gauss-Newton.

Para finalizar, destacar que $F_{ML}$ permite, mediante un contraste de hipótesis, comprobar el ajuste global del modelo:

**Proposición 4.6.** *Sea N el tamaño de la muestra y $F_{ML}$ el valor de la función minimizada por máxima verosimilitud. Entonces, dado el contraste de hipótesis:*

$$\begin{aligned} H_0 &: \sum = \sum(\boldsymbol{\theta}), \\ H_1 &: \sum \neq \sum(\boldsymbol{\theta}). \end{aligned} \tag{4.49}$$

*Se cumple que, bajo la hipótesis nula $H_0$, la distribución asintótica del estadístico $(N-1)F_{ML}$ sigue una distribución $\chi^2$ con $\frac{1}{2}(p+q)(p+q+1)-t$ grados de libertad.*

## *Estimación por mínimos cuadrados ponderados (diagonalizados)*

**Definición 4.20.** ***(Función de ajuste** $F_{WLS}$**)** La función de ajuste para la estimación por mínimos cuadrados ponderados es:*

$$F_{WLS} = [\boldsymbol{s} - \sigma(\boldsymbol{\theta})]^t \boldsymbol{W}^{-1} [\boldsymbol{s} - \sigma(\boldsymbol{\theta})], \tag{4.50}$$

*donde s es un vector de $\frac{1}{2}(p+q)(p+q-1)$ elementos no redundantes de $\boldsymbol{S}$, $\sigma(\theta)$ es el vector de elementos de $\sum(\theta)$ y $\boldsymbol{\theta}$ es el vector de parámetros libres. $\boldsymbol{W^{-1}}$ es un a matriz de $\frac{1}{2}(p+q)(p+q-1) x \frac{1}{2}(p+q)(p+q-1)$ de ponderaciones definida positiva.*

La estimación por mínimos cuadrados ponderados, también conocida como distribución asintóticamente libre, ofrece la posibilidad de trabajar con variables ordinales, dicotómicas o datos que, simplemente, no satisfacen la condición de multinormalidad.

La elección de $\boldsymbol{\theta}$ se hará con la intención de minimizar la suma ponderada del cuadrado de los residuos. Este procedimiento es análogo al de la estimación por mínimos cuadrados ponderados en regresión, con la salvedad de que aquí, los residuos provienen de las diferencias entre las varianzas y covarianzas predichas y observadas, mientras que en regresión, este proceso se lleva a cabo mediante la elección de los coeficientes. Destacar además que, al igual que ocurría en $F_{ML}$, la función $F_{WLS}$ da lugar a estimadores consistentes. De hecho, Browne (1982,1984) probaría que si $\boldsymbol{W}$ es la matriz de covarianzas asintótica de $\boldsymbol{S}$, entonces dicho estimador es asintóticamente eficiente. Mencionar que según Bollen (1989), la opción óptima de $\boldsymbol{W}$ es la matriz de covarianzas de las covarianzas muestrales.

Este método tiene como desventaja que precisa de muestras grandes[2], ya que de otro modo, el estadístico $\mathcal{X}^2$ podría ser sobreestimado. Además, cuando el número de variables observadas es mayor o igual a 20, la matriz de pesos aumenta considerablemente, dificultando su estimación. Este error sería tenido en cuenta y corregido por Muthén (1993) al considerar únicamente la diagonal de la matriz de pesos.

**Definición 4.21.** ***(Función de ajuste** $F_{DWLS}$**)** La función de ajuste para la estimación por mínimos cuadrados ponderados diagonalizados es:*

$$F_{DWLS} = [\boldsymbol{s} - \sigma(\boldsymbol{\theta})]^t diag(\boldsymbol{W}^{-1}) [\boldsymbol{s} - \sigma(\boldsymbol{\theta})]. \tag{4.51}$$

## *Estimación por mínimos cuadrados no ponderados*

**Definición 4.22.** *(Función de ajuste $F_{ULS}$) La función de ajuste para la estimación de mínimos cuadrados no ponderados es:*

$$F_{ULS} = \frac{1}{2} tr\left[\boldsymbol{S} - \left(\sum(\boldsymbol{\theta})\right)^2\right]. \tag{4.52}$$

[2]Muestras del orden de 5000 registros.

La función $F_{ULS}$ minimiza la suma de cuadrados de los elementos de la matriz de residuos $(\boldsymbol{S} - \sum(\boldsymbol{\theta}))$ análogamente a la regresión de mínimos cuadrados ordinarios (OLS). La principal diferencia reside en que $F_{ULS}$ calcula las diferencias entre varianzas y covarianzas muestrales predichas por el modelo, en lugar de las diferencias entre los valores muestrales de las variables observadas y los predichos por el modelo.

El hecho de minimizar la función $F_{ULS}$ conlleva la estimación consistente de $\boldsymbol{\theta}$ sin que esta precise de una distribución multinormal para las variables observadas. Sin embargo, el principal handicap de este método de estimación reside en que los estimadores obtenidos no son eficientes.

### *Estimación por mínimos cuadrados generalizados*

**Definición 4.23.** ***(Función de ajuste*** $F_{GLS}$***)*** *La función de ajuste para la estimación por mínimos cuadrados generalizados es:*

$$F_{GLS} = \frac{1}{2} tr\left(\left\{\left[\boldsymbol{S} - \sum(\boldsymbol{\theta})\right]\boldsymbol{W^{-1}}\right\}^2\right), \tag{4.53}$$

donde $\boldsymbol{W^{-1}} = (\boldsymbol{S} \otimes \boldsymbol{S})^{-1}$ es la matriz de pesos para la matriz de residuos.

**Propiedades 4.3**

1. $\boldsymbol{W^{-1}}$ converge en probabilidad a una matriz definida positiva cuando $N \longrightarrow \infty$, o es una matriz definida positiva de constantes.
2. Si $\boldsymbol{W^{-1}} = \boldsymbol{I}$, entonces $F_{ULS} = F_{GLS}$.

La función $F_{GLS}$ minimiza las desviaciones cuadráticas entre los elementos observados de $\boldsymbol{S}$ y los correspondientes elementos predictores de $\sum(\boldsymbol{\theta})$. Obsérvese además que los estimadores $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ son estimadores consistentes bajo la condición de multinormalidad. Por contra, no todas las elecciones de $\boldsymbol{W^{-1}}$ guiarán a estimadores eficientes.

## 4.4 Evaluación e interpretación del modelo

Una vez hemos conseguido identificar y estimar los parámetros del modelo, debemos efectuar un diagnóstico acerca de la bondad de ajuste del mismo. La etapa de evaluación e interpretación es quizás la más importante de todas. El objetivo de esta etapa es determinar si el modelo es correcto o no. Entendemos por modelo correcto aquel que identifica correctamente las relaciones entre variables sin omitir parámetro alguno. Un modelo correcto debe ser capaz de reproducir fidedignamente la realidad.

Evaluar el ajuste de nuestro modelo consiste, al fin y al cabo, en comprobar las diferencias existentes entre la matriz de covarianzas $\sum(\hat{\boldsymbol{\theta}})$ predicha y la matriz de covarianzas muestral **S**. Para llevar a cabo este proceso utilizaremos estadísticos globales. En concreto, debido a las desavenencias entre autores para indicar cuáles son los más efectivos, nos vemos obligados a exponer solo algunos[3]: $\boldsymbol{\chi^2}$, **CFI**, **TLI**, **RMSEA** y **SRMR**.

[3]Para más información consultar Bollen (1989) o Bentler (2004), entre otros.

## *Bondad de ajuste del modelo*

### *Estadístico $\chi^2$*

Este índice pertenece a las medidas de ajuste global. Afirmaremos, bajo las hipótesis propuestas en la proposición (4.6), que valores significativos (generalmente asociados a errores de significación bajos) del estadístico $\boldsymbol{\mathcal{X}^2}$ implican que las matrices $\mathbf{S}$ y $\sum(\hat{\boldsymbol{\theta}})$ difieren. Rechazaremos pues, bajo estas condiciones, la hipótesis nula y, por ende, el modelo. Buscaremos por tanto, valores de $\boldsymbol{\mathcal{X}^2}$ bajos (asociados a altos niveles de significación) que nos impidan rechazar la hipótesis nula $H_0 : \sum = \sum(\boldsymbol{\theta})$. En particular, aceptaremos $H_0$ cuando $0.05 < p-valor \leq 1$ (muy buen ajuste) o, en su defecto, cuando $0.01 \leq p-valor \leq 0.05$ (buen ajuste).

Una de las principales ventajas del índice de bondad de ajuste $\mathcal{X}^2$ es que permite que las conclusiones obtenidas respecto de la muestra, sean generalizadas a la población. Sin embargo, su principal inconveniente es que depende en gran medida del tamaño de la muestra. Es decir, a partir de muestras de más de 200 sujetos, muchos modelos presentarán valores de p $<0.001$ y tendrán que ser rechazados. Esto hace que el uso del $p-valor$ sea desaconsejable. Un mejor indicador es el ratio $\mathcal{X}^2/df$, donde $df$ hace referencia a los grados de libertad del modelo.

Destacar que el estadístico $\boldsymbol{\mathcal{X}^2}$ también se ve afectado cuando se viola la condición de normalidad multivariante. Además, a mayor complejidad del modelo, mayor probabilidad de que el test lo acabe aceptando. En el siguiente cuadro podemos observar las reglas que seguiremos para comprobar la bondad de ajuste del estadístico $\mathcal{X}^2/df$:

| Valor de $\mathcal{X}^2/df$ | Ajuste |
|---|---|
| $\mathcal{X}^2/df < 1$ | Modelo sobreidentificado |
| $1 < \mathcal{X}^2/df < 2$ | Muy buen ajuste |
| $2 < \mathcal{X}^2/df < 3$ | Buen ajuste |
| $3 < \mathcal{X}^2/df < 5$ | Ajuste raramente aceptable |
| $5 < \mathcal{X}^2/df$ | Ajuste insuficiente |

Cuadro 4.3: Ajuste del ratio $\mathcal{X}^2/df$ .

### *Comparative Fit Index (CFI)*

El **CFI** (índice de ajuste comparativo) es un índice perteneciente a las medidas incrementales de ajuste basadas en la comparación de modelos. Estos índices buscan comparar la mejoría de bondad de ajuste entre un modelo y su modelo base [4]. Definiremos el CFI como:

$$\boldsymbol{CFI} = \frac{(\mathcal{X}_b^2 - df_b) - (\mathcal{X}_t^2 - df_t)}{\mathcal{X}_b^2 - df_b}. \tag{4.54}$$

[4] Un modelo base es aquel en el que no existe asociación entre variables.

$df$ hace referencia a los grados de libertad y los subíndices $b$ y $t$ se refieren al modelo base y al modelo teórico respectivamente. El CFI, que toma valores entre 0 y 1, se regirá por las siguientes reglas:

| Valor de $\mathcal{X}^2$ | Ajuste |
|---|---|
| $0.95 < CFI < 1$ | Muy buen ajuste |
| $0.9 < CFI < 0.95$ | Buen ajuste |
| $0 < CFI < 0.9$ | Ajuste insuficiente |

Cuadro 4.4: Ajuste del Comparative Fit Index (CFI).

### *Root Mean Square Error Of Aproximation (RMSEA)*

El índice **RMSEA** (raíz cuadrática media del error de aproximación), al igual que el estadístico $\boldsymbol{\mathcal{X}^2}$, es un índice de ajuste global. Este índice estudia la discrepancia entre la matriz de covarianzas reproducida por el modelo y la matriz de covarianzas observadas en términos poblacionales. Definiremos el RMSEA como:

$$\boldsymbol{RMSEA} = \sqrt{\frac{\mathcal{X}_t^2 - df_t}{(N-1)\cdot df_t}}. \tag{4.55}$$

Podemos interpretrar este indice como el error de aproximación medio por grado de libertad. Valores entorno al 0.05 se consideran aceptables (Browne y Cudeck, 1993). Destacar además que Hu y Bentler (1999) probarían que, entre otros, el índice RMSEA es propenso a rechazar modelos correctos cuando el tamaño de la muestra es pequeño. En la siguiente tabla podemos observar el tipo de ajuste del modelo dependiendo del valor de este índice:

| Valor de $\mathcal{X}^2$ | Ajuste |
|---|---|
| $0 < RMSEA < 0.05$ | Muy buen ajuste |
| $0.05 < RMSEA < 0.08$ | Buen ajuste |
| $0.08 < RMSEA$ | Ajuste insuficiente |

Cuadro 4.5: Ajuste del Root Mean Square Error Of Aproximation (RMSEA).

### *Tucker-Lewis Index (TLI)*

El índice **TLI**, también conocido como Non-Normed Fit Index (NNFI), consiste en una medida de ajuste incremental usada para comparar un modelo teórico con un modelo base. Definiremos el TLI como:

$$\boldsymbol{TLI} = \frac{\mathcal{X}_b^2/df_b - \mathcal{X}_t^2/df_t}{\mathcal{X}_b^2/df_b - 1}. \tag{4.56}$$

La bondad de ajuste del modelo en base a los distintos valores que pueda tomar este índice se deduce de la siguiente tabla:

| Valor de $\mathcal{X}^2$ | Ajuste |
|---|---|
| $0.95 < TLI < 1$ | Muy buen ajuste |
| $0.9 < TLI < 0.95$ | Buen ajuste |
| $0 < TLI < 0.9$ | Ajuste insuficiente |

Cuadro 4.6: Ajuste del Tucker-Lewis index (TLI).

### *Standarized Root Mean Square Residual Index (SRMR)*

El índice **SRMR**, al igual que el RMSEA o el estadístico $\mathcal{X}^2$, es una medida de ajuste global. Esta, se puede interpretar como la diferencia estandarizada entre la matriz de varianzas-covarianzas observada y la matriz varianzas-covarianzas predicha. El SRMR puede calcularse como:

$$\boldsymbol{SRMR} = \sqrt{2 \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{n(n+1)}}, \tag{4.57}$$

donde $s_{ij}$ y $\hat{\sigma}_{ij}$ son componentes de $\mathbf{S}$ y $\sum(\hat{\theta})$ respectivamente. Nótese además que $n$ es el número de variables observadas de nuestro modelo.

En la siguiente tabla podemos observar el ajuste de nuestro modelo en base a los valores del SRMR:

| Valor de $\mathcal{X}^2$ | Ajuste |
|---|---|
| $0 < SRMR < 0.05$ | Muy buen ajuste |
| $0.05 < SRMR < 0.08$ | Buen ajuste |
| $0.08 < SRMR$ | Ajuste insuficiente |

Cuadro 4.7: Ajuste del Standarized Root Mean Square Residual Index (SRMR).

# 5 Corrupción y confianza en las instituciones

## 5.1 Modelos de ecuaciones estructurales con R

Antes de comenzar con el caso práctico, haremos una breve descripción del software empleado. **R** es un entorno y lenguaje de programación distribuido según licencia GNU; es decir, es un software libre y gratuito. Fue diseñado por Robert Gentleman y Ross Ihaka en 1993 a modo de reimplementación del lenguaje **S**, aunque actualmente es el resultado de la colaboración de toda una comunidad de usuarios[1]. Se caracteriza por su uso en estadística computacional, siendo sus principales aplicaciones el aprendizaje automático o la minería de datos, entre otras.

Destacar que **R** es un lenguaje de programación multiplataforma orientado a objetos. Esto es, está disponible en múltiples sistemas operativos y las variables, datos, funciones, etc... que lo componen, se guardan en memoria en forma de objetos para su posterior manipulación. Otra de sus ventajas reside en su flexibilidad para integrarse con distintas bases de datos. Para finalizar, mencionar que posee una gran capacidad para la documentación y la graficación debido a la posibilidad de poder cargar diferentes bibliotecas o paquetes con dichas funcionalidades.

Para este caso práctico haremos uso del entorno de desarrollo integrado *RStudio*. Los paquetes que vamos a mencionar a lo largo de este proyecto pueden encontrarse en el repositorio *Comprehensive R Archive Network (CRAN)* con dirección `http://CRAN.R-project.org/`.

### *Paquetes empleados*

- ***lavaan***: el paquete lavaan, acrónimo de *la*tent *va*riable *an*alysis, proporciona herramientas que pueden ser usadas para explorar, estimar y entender una amplia familia de modelos de variables latentes, incluyendo el análisis factorial, modelos de ecuaciones estructurales, modelos longitudinales o multinivel, entre otros. La especificación de los modelos se realizará de acuerdo a una sencilla sintaxis común. Dicha sintaxis propuesta por Rossel en 2012, puede verse reflejada en la siguiente tabla:

| Tipo de fórmula | Operador |
|---|---|
| Variable latente | =∼ |
| Regresión | ∼ |
| (Co)varianza (Residual) | ∼∼ |
| Intercepto | ∼ 1 |

Cuadro 5.1: Nomenclatura para la especificación de un modelo en lavaan

Diferenciaremos entre las funciones *cfa* y *sem* dependiendo de si queremos realizar un análisis factorial confirmatorio o un modelo de ecuaciones estructurales.

[1] Dichos usuarios pueden publicar paquetes con la intención de extender su configuración básica.

Destacar que estas funciones, a diferencia de la función *lavaan*, fijan automáticamente el valor de algunos parámetros del modelo[2] con la intención de facilitar su correcta identificación. En el caso de emplear la función *lavaan*, hay que ser especialmente cuidadosos a la hora de especificar el modelo. Por último, mencionar que las funciones *summary()* y *parametersEstimates()* aportan toda la información referente a la bondad de ajuste de nuestro modelo, así como la estimación de sus parámetros.

- ***semPlot***: paquete usado para la representación gráfica de los diagramas de caminos. Usaremos la función *semPaths* cuyo principal argumento consiste en el objeto asociado al modelo previamente calculado con el paquete *lavaan*.
- ***psych***: paquete centrado en la investigación psicológica mediante el análisis de datos multivariantes. Resultará de gran practicidad al permitirnos realizar la estimación del estadístico Alfa de Cronbach (*alpha*), obtener los principales estadísticos descriptivos (*describe*), representar la matriz de correlaciones (*cor.plot*), realizar análisis factoriales (*fa*), estimar el gráfico de sedimentación (*scree*) o calcular la prueba de adecuación muestral KMO (*KMO*).
- ***MVN***: paquete empleado para la estimación de los test de kurtosis y asimetría de Mardía. Utilizaremos la función *mvn*.
- ***Otras librerías***: *ggplot2*, *car*, *scales*, *gridExtra* o *parameters*.

## 5.2 Planteamiento del problema

Las instituciones son las reglas de juego que rigen una sociedad; imponen las limitaciones ideadas por las personas que dan forma a la interacción humana, garantizando el orden y contribuyendo a la construcción de conocimiento (North, 1993). Además, reducen la incertidumbre y proporcionan una estructura a la vida diaria. Por lo tanto, la calidad de las mismas es clave para el desarrollo de los países y su bienestar.

La corrupción por su parte, vista como un abuso de poder para el beneficio propio, ha sido y es uno de los mayores problemas a los que se ha enfrentado la sociedad española en las últimas décadas. En concreto, el *Transparency International* publicaría en 2009 el *Índice de percepción de la corrupción* (IPC), mediante el cual se estudiaría la transparencia del sector público en un total de 180 paises. España aparece en este ranking en el puesto 34 con una puntuación de 6.1, bajando por quinto año consecutivo (en 2006 figuraba en el puesto 23 con una puntuación de 6.8), evidenciándose así un claro auge en la percepción ciudadana de la corrupción.

Si los ciudadanos dudan de las instituciones, o consideran que están plagadas de corrupción, es poco probable que confíen en éstas o en quienes las encabezan, así como en sus representantes y sus políticos (Morris y Klesner, 2010). En nuestro caso, el deterioro institucional español, claramente afectado por tramas de corruptela, no se produce durante la crisis económica de 2009, sino en los años del boom económico e

[2]Generalmente, se fijan los coeficientes de regresión asociados a la primera variable observada especifiifacada en la construcción de las variables latentes.

inmobiliario. Esto deja un marco teórico propicio para estudiar con perspectiva cómo la confianza en las instituciones de la sociedad de 2009 se ve claramente afectada por los numerosos escándalos protagonizados por políticos y empresarios de los años precedentes.

Por otro lado, la democracia se define como un sistema político que defiende la soberanía del pueblo y el derecho del mismo a elegir y controlar a sus gobernantes. Se puede intuir, por tanto, que el correcto funcionamiento de esta puede tener efectos directos sobre nuestra percepción de la corrupción. De hecho, algunos investigadores que han estudiado esta relación (Warren, 2005; Bohara, Mitchel y Mittendorff, 2004) consideran que la relación entre la corrupción y la democracia es de causalidad: a mayor democracia se reducen las percepciones de corrupción.

A su vez, se puede pensar que la estratificación social puede influir en la percepción que, como ciudadanos, tenemos acerca de la corrupción o del funcionamiento de la democracia. De hecho, Huntigton (1972) y Wolfe (1980), establecen que la corrupción tiende a manifestarse especialmente cuando una clase o grupo ha obtenido el poder económico sin su correspondiente poder político. Esto es, generalmente, debería apreciarse una mayor percepción de la corrupción en aquellas clases sociales más altas. De igual modo, directa o indirectamente, deberíamos apreciar una clara relación entre la clase social y la confianza que depositamos en las instituciones.

Se presenta por tanto este caso práctico con la intención de validar las relaciones expuestas con anterioridad mediante un modelo de ecuaciones estructurales. Las hipótesis que planteamos por tanto en este proyecto son:

| | Hipótesis |
|---|---|
| $H_1$ | La percepción que tenemos acerca del funcionamiento de la democracia influye en nuestra percepción de la corrupción. |
| $H_2$ | La percepción que tenemos acerca del funcionamiento de la democracia influye en nuestra confianza en las instituciones. |
| $H_3$ | La percepción que tenemos de la corrupción influye en nuestra confianza en las instituciones. |
| $H_4$ | Nuestro estatus socioeconómico influye en la percepción que tenemos acerca de la corrupción. |
| $H_5$ | Nuestro estatus socioeconómico influye en nuestra confianza en las instituciones. |
| $H_6$ | Nuestro estatus socioeconómico influye en la percepción que tenemos acerca del funcionamiento de la democracia. |

Cuadro 5.2: Hipótesis a contrastar en el caso práctico

Para llevar a cabo este estudio se ha tomado una muestra de 335 registros obtenida del Centro de Investigaciones Sociológicas (CIS). Dicha encuesta acerca de *Ética Pública y Corrupción*, con número de estudio 2826, fue encargada y realizada por la Fundación e Instituto Universitario Ortega y Gasset el 16 de diciembre de 2009. Para agilizar el proceso, ha sido previamente tratada con Excel eliminando los datos no concluyentes

(No sabe/No contesta). Las variable objeto de nuestro estudio así como la forma en las que han sido medidas, pueden verse reflejadas en la siguiente tabla:

| Variable latente (1) | Variable latente (2) | Variable observada | Nombre |
|---|---|---|---|
| | | P43aa - Estatus socioeconómico (1-5) de la persona entrevistada. | *socioeconom* |
| | | P5 - Escala de satisfacción (0-10) con el funcionamiento de la democracia en España. | *democracia* |
| | P12 - Escala de confianza (0-10) en las instituciones (*confinstitu*). | P1201 - Los partidos políticos.<br>P1202 - El Gobierno central.<br>P1203 - El Poder judicial.<br>P1204 - Los gobiernos autonómicos.<br>P1205 - Los medios de comunicación.<br>P1206 - Los ayuntamientos y gobiernos locales. | *ppoliticos*<br>*gobcentral*<br>*poderjud*<br>*gobautonom*<br>*medcomun*<br>*goblocal* |
| Corrupción (*corrupcion*) | P22 - Grado de extensión de la corrupción (1-6) en distintos ámbitos de la Administración pública (*corrupcionpub*). | P2201 - Las fuerzas de seguridad.<br>P2202 - La administración de justicia.<br>P2204 - Las autoridades que otorgan contrarios públicos/subvenciones.<br>P2205 - Los inspectores/as de sanidad, urbanismo, etc.<br>P2206 - Las autoridades que conceden permisos y licencias de obras.<br>P2207 - Los/as trabajadores/as de las administraciones públicas. | *fuerzasseg*<br>*adminjust*<br>*subvenciones*<br>*inspectores*<br>*licencias*<br>*adminpub* |
| | P23 - Grado de extensión de la corrupción (1-6) en distintos sectores (*corrupcionos*). | P2301 - Empresarios/as.<br>P2302 - Sindicatos.<br>P2303 - La banca.<br>P2304 - ONGs.<br>P2305 - Medios de comunicación. | *empresarios*<br>*sindicatos*<br>*banca*<br>*ongs*<br>*medios* |
| | P24 - Grado de extensión de la corrupción (1-6) en distintos niveles políticos (*corrupcionpo*). | P2401 - La política local.<br>P2402 - La política autonómica.<br>P2403 - La política nacional.<br>P2404 - La política europea. | *corlocal*<br>*corautonom*<br>*cornacional*<br>*coreuropea* |

Cuadro 5.3: Variables e identificadores para el caso práctico

## 5.3 Resultados

### *Comprobación de las hipótesis previas*

De acuerdo con lo establecido en la sección 3, debemos comprobar que el tamaño de la muestra es lo suficientemente grande como para poder efectuar un modelo de ecuaciones estructurales. En nuestro caso, haciendo uso de la propuesta de Jayaram et al (2004), al tener 23 variables observadas deberíamos tener una muestra de al menos

230 registros. Al haber seleccionado 325, podemos concluir que se satisface dicho tamaño.

Veamos en segundo lugar si se satisface la condición de normalidad multivariante en nuestros datos. Nótese que las variables de partida son de tipo discreto, por lo que se presupone que una distribución de tipo continuo, como es la normal, no se satisfacerá. Los test de asimetría y curtosis de Mardia (Cuadro 5.4) obtenidos mediante la función mvn del paquete MVN muestran que, como era de esperar, no se satisface dicha condición. Es más, los tests de Shapiro-Wilk presentes en el anexo A.1 indican que además, tampoco existe normalidad univariada.

| Prueba | Estadístico | P-valor | Resultado |
|---|---|---|---|
| Test de Asimetría | 3983.22 | 9.55e-94 | No |
| Test de Curtosis | 22.88 | 0 | No |

Cuadro 5.4: Tests de Mardia para estudiar la normalidad multivariante

Para este caso práctico y ante la dificultad de obtener una muestra normal multivariada, se ha optado por emplear una estimación por mínimos cuadrados ponderados diagonalizados (DWLS). Sin embargo, si el lector quisiera realizar dicha estimación mediante máxima verosimilitud, podría hacer uso de las cotas establecidas por Curran, West y Finch en 1996. Obsérvese en el cuadro 5.5 que dichas condiciones se satisfacen en los datos objeto de nuestro estudio. Además, este fenómeno puede verse reflejado en los gráficos Q-Q e histogramas presentes en el anexo A.1.

| | n | Mean | Std.Dev | Median | Min | Max | 25th | 75th | Skew | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|
| democracia | 325 | 5.184615 | 2.547647 | 5 | 0 | 10 | 3 | 7 | -0.272930742 | -0.5944535 |
| ppoliticos | 325 | 3.707692 | 2.232774 | 4 | 0 | 9 | 2 | 5 | -0.128386978 | -0.6802062 |
| gobcentral | 325 | 3.735385 | 2.475337 | 4 | 0 | 10 | 2 | 5 | -0.105491496 | -0.9525863 |
| poderjud | 325 | 4.095385 | 2.659738 | 5 | 0 | 10 | 2 | 6 | -0.116950637 | -1.0694860 |
| gobautonom | 325 | 4.132308 | 2.405825 | 5 | 0 | 10 | 3 | 6 | -0.210824606 | -0.6497646 |
| medcomun | 325 | 4.313846 | 2.324047 | 5 | 0 | 9 | 3 | 6 | -0.316630409 | -0.6190541 |
| goblocal | 325 | 4.086154 | 2.471809 | 5 | 0 | 10 | 2 | 6 | -0.146254364 | -0.7553123 |
| fuerzasseg | 325 | 2.852308 | 1.267971 | 3 | 1 | 6 | 2 | 4 | -0.012300326 | -1.2872379 |
| adminjust | 325 | 2.729231 | 1.199501 | 2 | 1 | 6 | 2 | 4 | 0.123243251 | -1.1836282 |
| subvenciones | 325 | 1.966154 | 1.025339 | 2 | 1 | 5 | 1 | 2 | 0.957933626 | 0.0736367 |
| inspectores | 325 | 2.501538 | 1.218744 | 2 | 1 | 6 | 2 | 4 | 0.434966589 | -0.9432603 |
| licencias | 325 | 1.950769 | 1.050003 | 2 | 1 | 5 | 1 | 2 | 1.070606791 | 0.3113306 |
| adminpub | 325 | 2.956923 | 1.311625 | 3 | 1 | 6 | 2 | 4 | 0.062805841 | -1.0956849 |
| empresarios | 325 | 2.440000 | 1.083091 | 2 | 1 | 5 | 2 | 3 | 0.357467227 | -1.0021482 |
| sindicatos | 325 | 2.689231 | 1.221733 | 2 | 1 | 6 | 2 | 4 | 0.191725727 | -1.1425005 |
| banca | 325 | 2.372308 | 1.194058 | 2 | 1 | 6 | 1 | 3 | 0.578827573 | -0.7119996 |
| ongs | 325 | 3.378462 | 1.235345 | 4 | 1 | 6 | 2 | 4 | -0.156695952 | -0.9194734 |
| medios | 325 | 2.766154 | 1.199573 | 2 | 1 | 6 | 2 | 4 | 0.273141916 | -0.9235205 |
| corlocal | 325 | 2.372308 | 1.141191 | 2 | 1 | 5 | 2 | 3 | 0.494689456 | -0.8556445 |
| corautonom | 325 | 2.390769 | 1.134983 | 2 | 1 | 6 | 2 | 3 | 0.522585507 | -0.6175539 |
| cornacional | 325 | 2.230769 | 1.050709 | 2 | 1 | 5 | 1 | 3 | 0.581905543 | -0.6135537 |
| coreuropea | 325 | 2.723077 | 1.208352 | 3 | 1 | 5 | 2 | 4 | 0.101009335 | -1.1893290 |
| socioeconom | 325 | 2.846154 | 1.381433 | 3 | 1 | 5 | 2 | 4 | -0.004019421 | -1.3711541 |

Cuadro 5.5: Tests de Mardia para estudiar la normalidad multivariante

Analicemos ahora si los datos empleados son lo suficientemente consistentes. Para ello, usaremos el estadístico Alfa de Cronbach propuesto en la sección 3. En el caso que nos compete, el Alfa de Cronbach global de nuestra muestra posee un valor de 0.89, resultado considerable para afrontar con garantías la construcción del modelo.

## *Análisis exploratorio*

En cuanto a la distribución de nuestros datos, remarcar que la encuesta ha sido estratificada por cuotas de sexo y edad, habiendo participado un total de 195 hombres (60 %) y 130 mujeres (40 %). Obsérvese que nos encontramos ante una muestra donde las personas de mediana edad predominan, oscilando la media de edad entre los 43 y 44 años. Destacar además que ningún menor de edad ha participado en el estudio y que la persona más longeva tiene 89 años.

Otro factor a tener en cuenta es que el nivel educativo asociado a nuestra muestra tiende a estabilizarse entorno a los niveles medios, siendo el valor predominante el de personas con certificación primaria (34.5 %). Aún así, puede observarse una clara presencia de personas con títulos universitarios y/o superiores (10.2 % + 17.5 %). Por otro lado, nótese la clara predominancia de personas con trabajo activo (56.6 %) frente a aquellas en situación de desempleo (16.6 %). Estos datos reproducen fielmente la realidad, ya que, en 2009 la tasa de desempleo alcanzaba un 18.83 % con un total de 4.326.500 personas desempleadas.

Se presenta la siguiente gráfica como un resumen de las medias de las principales variables observadas escogidas para la realización de un posterior análisis factorial. Han sido divididas según la variable latente que miden.

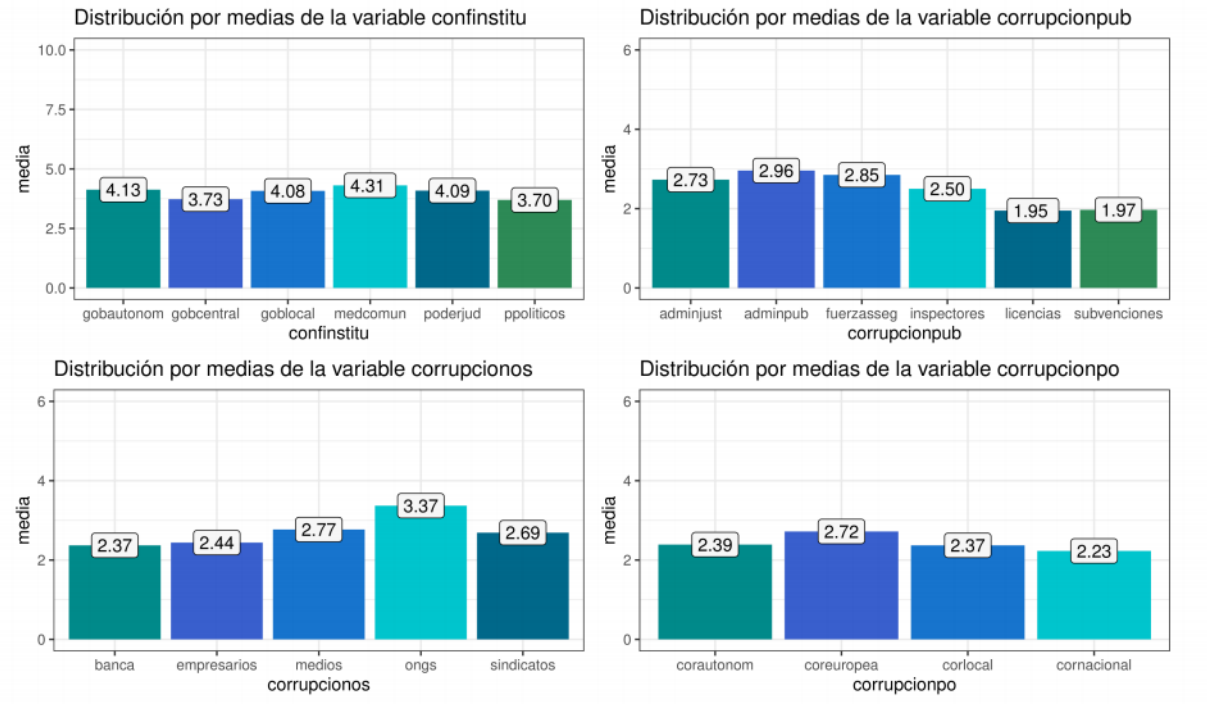


Figura 5.1: Distribución por medias de las variables de nuestro estudio

Si nos centramos en las variables observadas asociadas a la variable latente *confinstitu*, podemos observar un claro suspenso de las instituciones (4/11). Nótese que ninguna variable supera un 4.31, siendo el 5.5 el valor mínimo para considerar que existe una confianza real en las mismas. Dentro de ellas, distinguiremos una menor confianza en los partidos políticos (*ppoliticos*) y el gobierno central (*gobcentral*).

En lineas generales, las variables observadas asociadas a la variable latente *corrupcion*, muestran claros indicios (2.55/6)[3] de una situación preocupante. Diferenciaremos tres tipos dependiendo del sector el que nos situemos. Comenzaremos viendo que la percepción de la corrupción pública aflora sobre todo en las autoridades que otorgan contratos públicos/subvenciones (1.97/6) y en las autoridades que conceden permisos y licencias de obras (1.95/6). No obstante ninguna de las otras instituciones se salva. En cuanto a la percepción de la corrupción en otros sectores, destacar que salvo en el caso de las ONGs (3.37/6), el resto de variables observadas muestran indicios de corrupción. Por último mencionar que, aunque la presencia de corrupción a nivel político sea la mayor de las tres, resulta curioso que la percepción de la misma (2.43/6) difiera en tan poca medida del resto, siendo posiblemente la más cuestionada en 2009.

Se muestra ahora la distribución de la muestra en función de la percepción del funcionamiento de la democracia y el estatus socioeonómico unipersonal:

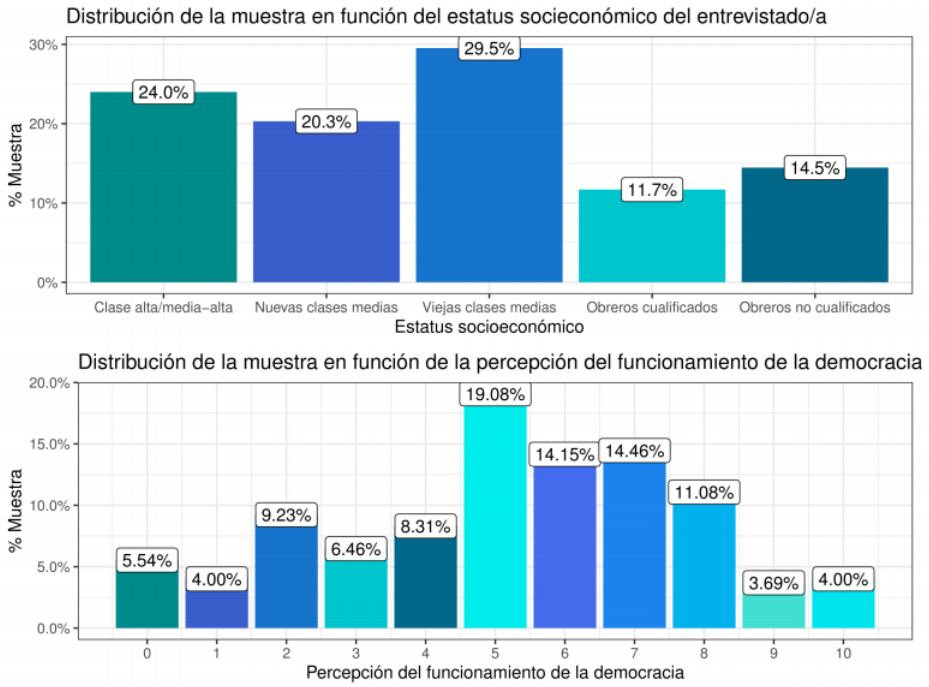


Figura 5.2: Distribución muestral en función de la percepción del funcionamiento de la democracia y el estatus socioecnomómico

[3]En este caso, el valor 1 hace referencia a una gran extensión de la corrupción y el valor 6 a la no existencia de la misma.

Nótese que la muestra sugiere una población perteneciente a clases sociales medias-altas, siendo las viejas clases medias (29,5 %) la más frecuente. En cuanto a la percepción del funcionamiento de la democracia, destacar que existen opiniones muy dispares. Sin embargo, puede apreciarse una ligera estabilización de los datos entorno al aprobado (5.18/11).

## *Análisis de correlaciones*

Realizaremos ahora un análisis de correlaciones que nos ayude a identificar si la distribución por constructos expuesta con anterioridad tiene fundamento matemático o no. Para llevarlo a cabo usaremos el coeficiente de correlación de Spearman, elegido por encontrarnos ante variables de tipo discreto. La matriz que recoge dicha información (A.3) muestra una presencia considerable[4] de correlación entre las variables observadas que definen los constructos *confinstitu* y *corrupcion*, siendo esto un buen indicio para llevar a cabo el análisis factorial. De igual modo, nótese que la variable *socioeconom* muestra una baja relación con el resto de variables. Aún así, la ausencia de correlación no prueba la ausencia de causalidad (Bollen, 1989). Por otro lado, si estudiamos el determinante de la matriz de correlaciones ($3{,}77 \cdot 10^{-5}$), podemos ver que existe una fuerte dependencia lineal entre las variables observadas. Descartándose la posibilidad de que estas sean combinación lineal.

Ahora bien, diremos que existe multicolinealidad cuando las variables observadas estén altamente correlacionadas, lo cual conlleva que las estimaciones de los parámetros en los modelos lineales sean muy inestables. Obsérvese que no existen indicios de multicolinealidad en nuestra muestra, ya que a simple vista no se observan coeficientes de correlación superiores a 0.7. Para cerciorarnos de que dicha hipótesis es cierta, haremos uso del Factor de Inflación de la Varianza (VIF). Este estadístico mostraría presencia de multicolinealidad si alguna de las variables presentara un valor superior a 10. Por lo tanto, a la vista de que todas las variables de nuestro estudio oscilan entre 1.102 y 3.151, podemos concluir que no existe multicolinealidad entre las mismas.

```
##   democracia   ppoliticos   gobcentral     poderjud   gobautonom     medcomun
##     1.435068     2.223504     2.108872     2.040416     2.197113     1.584642
##     goblocal   fuerzasseg    adminjust subvenciones  inspectores    licencias
##     2.186818     2.235818     2.421627     2.005613     2.151673     1.919842
##     adminpub  empresarios   sindicatos        banca         ongs       medios
##     1.744466     1.541339     1.974659     1.774146     1.780670     1.951868
##     corlocal   corautonom  cornacional   coreuropea  socioeconom
##     2.338454     3.151028     2.520453     2.288313     1.102560
```

## *Análisis factorial confirmatorio*

Para poder llevar a cabo un modelo de ecuaciones estructurales debemos asegurarnos que los modelos de medida que lo conforman están bien construidos. Según lo expuesto en la sección 4, especificaremos dichos constructos mediante el uso del análisis factorial. Se presentará por tanto en esta sección un análisis factorial confirmatorio

[4] Hablaremos de una presencia considerable cuando el índice de correlación sea superior a 0.3

para las escalas *corrupcion* y *confinstitu*, validando así las estructuras propuestas. Destacar además que estas estructuras fueron diseñadas de acuerdo al criterio del autor y no existe una teoría subyacente tras ellas, solo meras suposiciones basadas en la disposición de las preguntas de la encuesta.

### *Análisis factorial confirmatorio de la variable corrupcion*

La variable *corrupcion* ha sido medida mediante el uso de 16 items que dependen de nuestra percepción de la corrupción política, administrativa y en otros sectores. Destacar que, en primera instancia, nos enfrentamos ante un análisis factorial de segundo orden, compuesto por un factor principal (*corrupcion*) que depende a su vez de otros tres factores (*corrupcionpub*, *corrupcionpo* y *corrupcionos*).

Para probar la estructura factorial asignada a los items relacionados con la corrupción, debemos garantizar que nos encontramos en disposición de aplicar un análisis factorial confirmatorio. Obsérvese que tanto el resultado de la prueba de adecuación muestral KMO (0.88), como los del MSA (0.84-0.92) y la prueba de Fligner-Killen (*p-valor* = $8.9 \cdot 10^{-4}$ <0.05) arrojan resultados significativos como para llevar a cabo dicho análisis.

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datos[, c(8:23)])
## Overall MSA =  0.88
## MSA for each item =
##   fuerzasseg    adminjust subvenciones  inspectores    licencias     adminpub
##         0.87         0.88         0.87         0.89         0.85         0.89
##  empresarios   sindicatos        banca         ongs       medios     corlocal
##         0.85         0.89         0.89         0.87         0.90         0.85
##   corautonom  cornacional   coreuropea  socioeconom
##         0.84         0.89         0.92         0.49
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  datos[, c(8:23)]
## Fligner-Killeen:med chi-squared = 141.52, df = 15, p-value < 2.2e-16
```

La obtención del número de factores en los que dividiremos la variable *corrupcion* se realizará de acuerdo al gráfico de sedimentación. Este gráfico nos hará plantearnos la estructura factorial a elegir, ya que, aunque el criterio de Kaiser afirme que el número óptimo de factores es 4, no se aprecia una diferencia considerable de varianza total explicada cuando *nfactor* = 2 o 3. Es por ello, y ante la imposibilidad de formar cuatro factores consistentes, que se opta finalmente por *nfactor*=3.

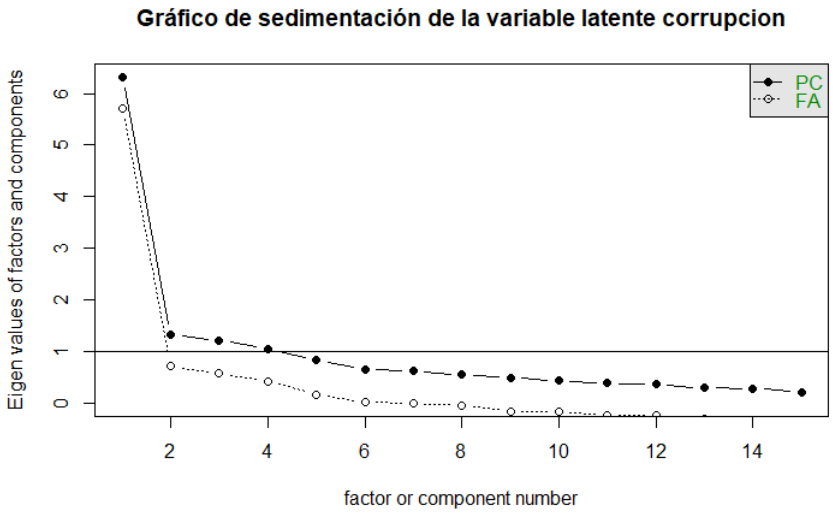


Figura 5.3: Gráfico de sedimentación de la variable *corrupcion*

Para llevar a cabo la extracción factorial se hará uso del método del Factor Principal, buscando extraer así los factores que expliquen la mayor cantidad de varianza posible (49,58 %[5]). Además, se aplicará una rotación ortogonal varimax con la intención de maximizar los pesos de cada factor (Cuadro 5.6). Obsérvese que los tres factores obtenidos reafirman las suposiciones teóricas realizadas.

| Variable | *corrupcionpub* | *corrupcionpo* | *corrupcionos* |
|---|---|---|---|
| inspectores | 0.74 | | |
| adminjust | 0.60 | | |
| adminpub | 0.59 | | |
| subvenciones | 0.55 | | |
| fuerzasseg | 0.55 | | |
| licencias | 0.46 | | |
| corautonom | | 0.86 | |
| corlocal | | 0.64 | |
| cornacional | | 0.64 | |
| coreuropea | | 0.61 | |
| medios | | | 0.73 |
| banca | | | 0.57 |
| sindicatos | | | 0.54 |
| ongs | | | 0.54 |
| empresarios | | | 0.50 |

Cuadro 5.6: Pesos factoriales de la variable *corrupcion*

Garantizadas las primeras suposiciones, haremos uso de la función *cfa* del paquete *lavaan* para llevar a cabo el análisis factorial confirmatorio. Recordemos que la estima-

[5]En ciencias sociales es común que la cantidad de varianza total explicada oscile entorno a un 40 %-60 %. De hecho, según Henson y Roberts (2006) la proporción media de varianza explicada por los factores suele ser de un 52,03 %

ción se realizará mediante mínimos cuadrados ponderados diagonalizados. El constucto final puede verse reflejado en la siguiente imagen:

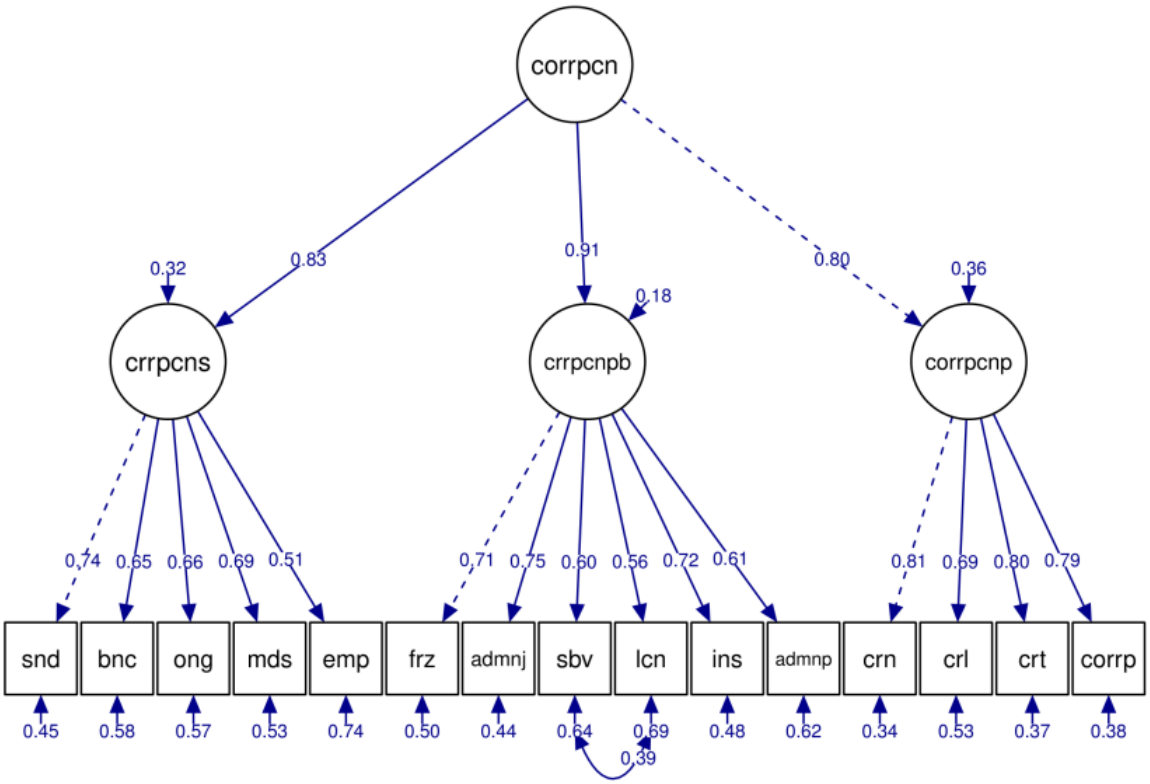


Figura 5.4: Diagrama del constructo *corrupcion*

Ahora bien, para ajustar correctamente el modelo de medida se han realizado sendos análisis factoriales confirmatorios a las variables latentes de primer orden. Destacar que los tests KMO, MSA y Fligner-Killen (A.3) aportan resultados significativos para llevarlos a cabo. Además, nótese que los estadísticos de bondad de ajuste del modelo (Cuadro 5.7) indican un muy buen ajuste del mismo. Mencionar que para que el ajuste de la variable *corrupcionpub* fuese óptimo, ha sido necesaria la inclusión de una covarianza entre los errores de medición de las variables observadas *subvenciones* y *licencias*. Dicha covarianza ha sido estimada mediante la función *modificationIndices()* del paquete *lavaan*.

| Variable | $X^2$ | df | $X^2/df$ | CFI | TLI | RMSEA | SRMR | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| *corrupcionpo* | 5.52 | 2 | 2.76 | 0.994 | 0.983 | 0.074 | 0.044 | 0.85 |
| *corrupcionpub* | 19.43 | 8 | 2,4287 | 0.988 | 0.977 | 0.066 | 0.054 | 0.83 |
| *corrupcionos* | 13.07 | 5 | 2.614 | 0.986 | 0.972 | 0.071 | 0.053 | 0.79 |
| *corrupcion* | 105.47 | 87 | 1.212 | 0.996 | 0.995 | 0.025 | 0.054 | 0.9 |

Cuadro 5.7: Bondad de ajuste de las variables asociadas al constructo *corrupcion*

Destacar, por último, que los estadísticos Alfa de Cronbach asociados a la variable *corrupcion* y a los factores de primer orden asociados a ella, presentan valores superiores a 0.79, lo que denota una gran consistencia de los datos empleados en la medición.

### *Análisis factorial confirmatorio de la variable confinstitu*

El constructo *confinstitu* ha sido medido gracias a 6 variables observadas que miden nuestra confianza en partidos políticos (*ppoliticos*), gobierno central (*gobcentral*), poder judicial (*poderjud*), gobiernos autonómicos (*gobautonom*), gobiernos locales (*goblocal*) y medios de comunicación (*medcomun*). Realizaremos un proceso análogo al anterior verificando que se puede aplicar un análisis factorial confirmatorio a este constructo. Las pruebas de adecuación muestral KMO (0.84), MSA (0.83-0.89) y la prueba de Fligner-Killen ($p - valor = 2.2 \cdot 10^{-16} < 0.05$), ofrecen valores significativos.

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datos[, c(2:7)])
## Overall MSA =  0.86
## MSA for each item =
## ppoliticos gobcentral   poderjud gobautonom   medcomun   goblocal
##       0.84       0.83       0.89       0.87       0.89       0.84
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  datos[, c(2:7)]
## Fligner-Killeen:med chi-squared = 20.759, df = 5, p-value = 0.0008996
```

Haciendo uso de la propuesta de Kaiser, el gráfico de sedimentación (Figura 5.5) confirma la presencia de un solo factor.

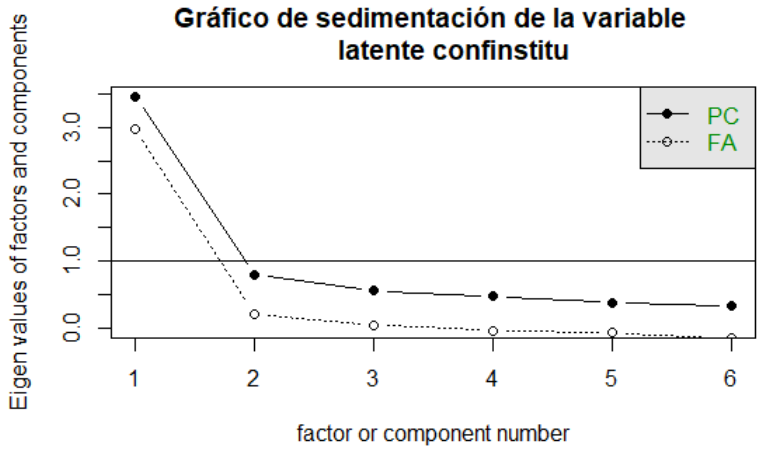


Figura 5.5: Gráfico de sedimentación del constructo *confinstitu*

Ahora bien, los pesos factoriales de las variables observadas que conforman la variable *confinstitu* pueden verse reflejados en la siguiente tabla:

| Variable | *confinstitu* |
|---|---|
| gobautonom | 0.78 |
| ppoliticos | 0.77 |
| poderjud | 0.73 |
| goblocal | 0.71 |
| gobcentral | 0.68 |
| medcomun | 0.52 |

Cuadro 5.8: Pesos factoriales de la variable *confinstitu*

Definimos de este modo el constructo que medirá nuestra confianza en las instituciones. Haciendo uso de la función *semPaths* del paquete *semPlot*, podemos visualizar el constructo *confinstitu* final:

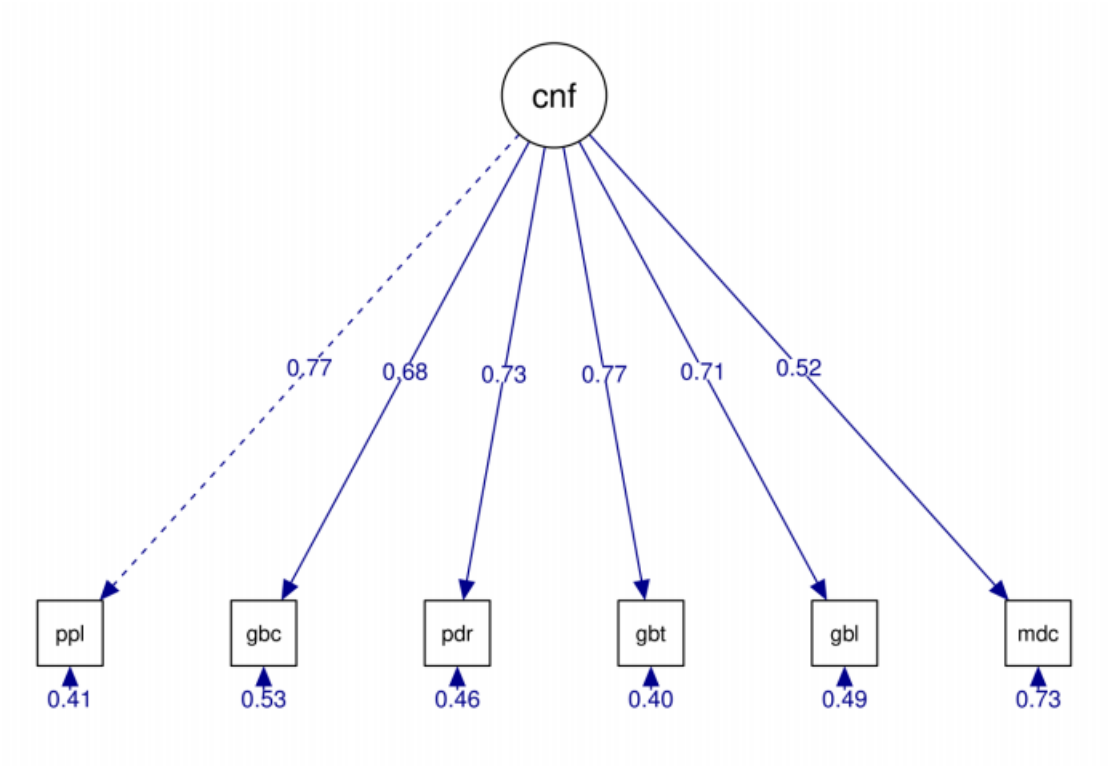


Figura 5.6: Diagrama del constructo *confinstitu*

Obsérvese que al igual que ocurriera en el constructo corrupcion, todos los coeficientes de regresión presentan un valor superior a 0.5. Veamos ahora que el modelo ajusta correctamente:

| Variable | $X^2$ | df | $X^2/df$ | CFI | TLI | RMSEA | SRMR | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| *confinstitu* | 10.74 | 9 | 1.193 | 0.998 | 0.997 | 0.024 | 0.043 | 0.84 |

Cuadro 5.9: Bondad de ajuste de las variables asociadas al constructo *confinstitu*

Finalmente, estudiaremos el valor del alfa de Cronbach de la variable *confinstitu* para comprobar la fiabilidad de la escala considerada. Este, tiene un valor de 0.84, por lo que podemos afirmar que los datos son fiables.

## *Modelo de ecuaciones estructurales general*

Garantizada la construcción de los modelos de medida, se construye un modelo de ecuaciones estructurales general capaz de representar las relaciones descritas al comienzo de este capítulo. Haciendo uso nuevamente de las funciones *sem* y *semPaths* de los paquetes *lavaan* y *semPlot* respectivamente, llegamos al diagrama de caminos presente en la figura 5.7. Nótese una óptima bondad de ajuste de nuestro modelo, donde todos los coeficientes presentan valores incluidos en las cotas preestablecidas:

| | $X^2$ | df | $X^2/df$ | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|---|---|
| SEM General | 233.809 | 222 | 1.053 | 0.999 | 0.998 | 0.013 | 0.053 |

Cuadro 5.10: Bondad de ajuste del modelo de ecuaciones estructurales general

Obsérvese gráficamente que todas las relaciones propuestas, salvo aquellas derivadas de la variable observada *socioeconom*, aparentan ser estadísticamente significativas. Es más, fijándonos en los coeficientes estandarizados de la figura 5.7, se puede observar la inexistencia de una relación causal entre las variables *corrupcion* y *socioeconom*. Se presenta de este modo una tabla con los principales resultados obtenidos al llevar a cabo la estimación de parámetros.

| Relación | Estandar | Estimación | Error | P-valor |
|---|---|---|---|---|
| corrupcion ~ socioeconom | 0.005 | 0.002 | 0.016 | 0.882 |
| democracia ~ socioeconom | -0.067 | -0.124 | 0.095 | 0.193 |
| confinstitu ~ socioeconom | -0.113 | -0.139 | 0.047 | 0.003 |
| corrupcion ~ democracia | 0.372 | 0.104 | 0.010 | 0.000 |
| confinstitu ~ democracia | 0.324 | 0.216 | 0.034 | 0.000 |
| confinstitu ~ corrupcion | 0.481 | 1.150 | 0.088 | 0.000 |

Cuadro 5.11: Relaciones estimadas y estandarizadas

Como se puede observar, además de la anteriormente descrita, tampoco existe una relación causal entre la variable *democracia* y *socioeconom*. Resulta curioso que, a la vista de los resultados obtenidos, exista una relación causal entre la variable *socioeconom* y *confinstitu*. De todos modos, se entiende que puede ser debido a la muestra seleccionada y, por tanto, nos limitaremos a decir que no existe suficiente evidencia estadística como para poder rechazar dicha hipótesis.

Concluimos así este caso práctico remarcando que no existe una relación causal entre la clase social a la que pertenecemos y la precepción de la corrupción o del sistema democrático, rechazando, entre otras, la hipótesis propuesta por Huntigton en 1972.

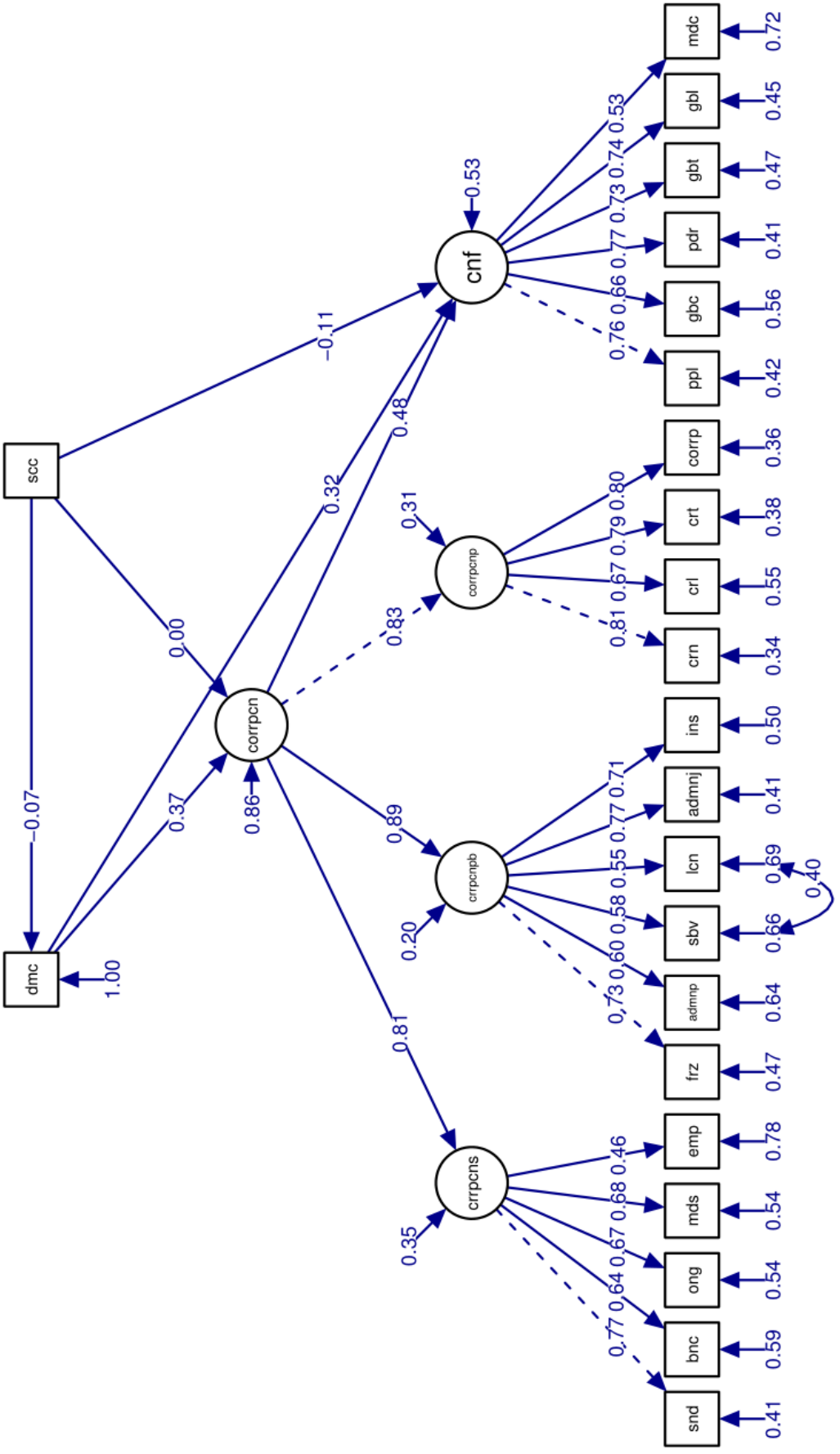


Figura 5.7: Modelo de ecuaciones estructurales general

# 6

# *Conclusiones*

El presente documento tenía como objetivo presentar los principales resultados usados en el modelado de ecuaciones estructurales, así como investigar con datos reales una aplicación práctica de los mismos.

En primer lugar, se ha establecido la hipótesis fundamental sobre la que se construyen los modelos de ecuaciones estructurales. Esto es, minimizar la diferencia entre la matriz de covarianzas poblacional y la matriz de covarianzas reproducida. Además, se presenta el marco histórico en el que se engloban los SEM con la intención de conocer cómo surgen y hacia dónde van. Diferenciaremos también los términos correlación y causalidad, pues la definición de esta, como hemos podido observar, es mucho más compleja que una simple correlación entre variables. En segundo lugar, se realizará una breve introducción a los distintos tipos de modelos de ecuaciones estructurales, decantándonos finalmente por realizar una revisión teórica del modelo más generalista. Se presenta de igual modo, la principal nomenclatura usada y su representación mediante diagramas de caminos, haciendo especial hincapié en los tipos de relaciones existentes entre variables.

Por otro lado, destacamos que los modelos de ecuaciones estructurales son una herramienta estadística muy potente para la validación de teorías en ciencias sociales. Sin embargo, sufren muchas limitaciones. Obsérvese por ejemplo que son muy sensibles al tamaño muestral, a la linealidad o a la distribución que siguen los datos. En cuanto a la construcción de los mismos, detacar la correcta validación de los componentes estructural y de medida, puesto que si no son correctamente especificados e identificados, la estimación de sus parámetros carece de sentido y, por ende, nuestro modelo. Remarcar la importancia del teorema 4.1, pues se erige como el principal sosten de este, nuestro proyecto.

Por último, mencionar que el caso práctico nace con la intención de conocer y medir las consecuencias de un tema tan candente como es la corrupción. Quedarán validadas estadísticamente las hipótesis que afirman que nuestra percepción de la corrupción y nuestra confianza en las instituciones se ven claramente afectadas por nuestra percepción del funcionamiento de la democracia. A su vez, diremos que la percepción que tenemos de la corrupción afecta directamente e indirectamente a nuestra confianza en las instituciones. Sin embargo, según lo que hemos podido observar, no existe suficiente evidencia estadística como para afirmar que el factor socioeconómico tenga influencia en nuestras percepciones de la corrupción y del funcionamiento democrático.

En lo referente a la elaboración de este proyecto, destacar que se han aplicado los conocimientos adiquiridos a lo largo del Grado en Matemáticas, siendo especialmente decisivo el papel de la asignatura *Análisis de datos*. Mencionar además, que los modelos de ecuaciones estructurales abarcan mucho más de lo expuesto en este proyecto. Sin embargo, debido a la limitación de su extensión, solo se han presentado los resultados más relevantes y el modelo con mejores estimaciones. Para finalizar, agradecer a mi tutor, Fernando Reche Lorite, su colaboración a lo largo de este proyecto.

# *Bibliografía*


[1] B. Shipley, *Cause and Correlation in Biology. A User's Guide to Path Analysis, Structural Equations and Causal Inference*, Cambridge University Press, 2004.

[2] B. Williams, A. Onsman, T. Brown, *Exploratoy factor analysis: A five-step guide for novices*, Journal of Emergency Primary Health Care (JEPHC), 8 (2010), 1-14.

[3] C. D. Rivera Ramírez, *Un modelo de ecuaciones estructurales para el escalamiento multidimensional de datos asimétricos*. Granada: Universidad de Granada, 2018.

[4] D. Peña, *Análisis de datos multivariantes*, Editorial McGraw Hill, 2002.

[5] J. C. Westland, *Lower bounds on sample size in structural equation modeling*. Electronic Commerce Research and Applications, 9 (2010), 476–487.

[6] J. M. Batista Foguet, G.C. Gallart, *Modelos de ecuaciones estructurales*, Cuadernos de Estadística, Editorial La Muralla, S.A., 2012.

[7] K. A. Bollen, *Structural equations with latent variables*, John Wiley  Sons, 1989.

[8] K. G. Jöreskog, U. H. Olsson, F. Y. Wallentin, *Multivariate Analysis with LISREL*, Springer Series in Statistics, 2016.

[9] L. J. Cronbach, *Coefficient alpha and the internal structure of the tests*. Psychometrika, 16 (1951), 297-334.

[10] Ledyard R Tucker and Charles Lewis, *A reliability coeficient for maximum likelihood factor analysis**. Psychometrika, 38 (1973), 1-10.

[11] M. A. Verdugo, M. Crespo, M. Badía, B. Arias, *Metodología en la investigación sobre discapacidad. Introducción al uso de las ecuaciones estructurales*, Publicaciones del INICO, 2008.

[12] S. A. Mulaik, *Linear Causal Modeling with Structural Equations*, Chapman & Hall/CRC, 2009.

[13] S. Epskamp, *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output*. R package version 1.1.2 (2019). https://CRAN.R-project.org/package=semPlot.

[14] S. Yáñez Canal, M.C. Jaramillo Elorza, J.C Correa Morales, *Una revisión de medidas multivariadas de asimetría y Kurtosis para pruebas de multinormalidad*, Revista Colombiana de Estadística, 22 (1999), 5-16.

[15] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley series in probability and mathematical statistics, 2003.

[16] Y. Rosseel, *lavaan: An R Package for Structural Equation Modeling*. Journal of Statistical Software, 48 (2012), 1-36. https://www.jstatsoft.org/v48/i02/.

[17] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.6 (2021), https://CRAN.R-project.org/package=psych.

# Anexos

## A.1 Código empleado

```
#Fijamos el directorio activo
setwd("C:/Users/6001638/Desktop")

#Cargamos los paquetes que vamos a necesitar
library(psych)
library(ggplot2)
library(car)
library(lavaan)
library(semPlot)
library(MVN)
library(scales)
library(dplyr)
library(gridExtra)
library(parameters)

#Leemos el archivo CSV
datos <- read.table("C:/Users/6001638/Desktop/muestraTFG.csv",sep=';',
                    header = TRUE, dec='.')

#Convertimos a numéricos los valores character
datos <- setNames(data.frame(lapply(datos, as.numeric)),
                  colnames(datos))

#datos <- datos[,-c(10)]
#datos <- datos[,-c(24:28)]

#*************************************************************************
#COMPROBACIÓN DE HIPÓTESIS PREVIAS
#*************************************************************************

#Aplicamos la función mvn para estudiar la normalidad de los datos y
#sus valores de kurtosis y asimetría
norm <- mvn(datos, mvnTest = "mardia")
norma.multivariada <- as.data.frame(norma$multivariateNormality)
norma.multivariada <- norma.multivariada[-c(3),]
norma <- as.data.frame(norma$Descriptives)
norma

#Obtenemos el qqplot y el histograma de cada variable
qqplot <- mvn(data = datos, mvnTest = "mardia",
univariatePlot = "qqplot")
```

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

#Estudio de correlaciones entre variables observadas
matriz.cor <- cor(datos, method = "spearman")
cor.plot(matriz.cor,cex=0.75)

#Consistencia de la muestra
psych::alpha(datos)

#*************************************************************************
#********************** CONSTRUCTO CONFINSTITU ***************************
#*************************************************************************

#Prueba KMO y de Fligner-Killeen para la variable confinstitu
KMO(datos[,c(2:7)])
fligner.test(datos[,c(2:7)])

#Gráfico de sedimentación de la variable latente confinstitu
scree(datos[,c(2:5,7)], main="Gráfico de sedimentación de la variable
      latente confinstitu")

#Cargas factoriales de la variable confinstitu
fac.confinstitu <- psych::fa(datos[,c(2:7)], nfactor = 1,
rotate = "none", fm="pa") %>%
  model_parameters(sort = TRUE, threshold = "max")
fac.confinstitu

#Realizamos el análisis factorial confirmatorio de la variable
#confinstitu
confinstitu <- 'confinstitu =~ ppoliticos + gobcentral + poderjud +
           gobautonom + goblocal + medcomun'
ajuste.confinstitu <- cfa(confinstitu, data = datos, estimator = "dwls",
                          meanstructure = FALSE)
summary(ajuste.confinstitu, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste.confinstitu, c("chisq", "df", "cfi", "tli",
           "rmsea", "srmr"))

#Hacemos uso de la función semPaths para representar el constructo
semPaths(ajuste.confinstitu, whatLabels = "std",style="lisrel",
        layout="tree2",
         reorder=FALSE, optimizeLatRes=TRUE, edge.label.position=.5,
         edge.label.cex = 0.8, edge.color = "darkblue" )

#Alpha de Cronbach de la variable confinstitu
psych::alpha(datos[,c(2:5,7)])
```

```r
#*************************************************************************
#************************ CONSTRUCTO CORRUPCION ************************
#*************************************************************************


#************************ VARIABLE CORRUPCIONPO ************************

#Prueba KMO y de Fligner-Killeen para la variable corrupcionpo
KMO(datos[,c(19:22)])
fligner.test(datos[,c(19:22)])

#Gráfico de sedimentación de la variable corrupcionpo
scree(datos[,c(19:22)], main="Gráfico de sedimentación de la variable
      latente corrupcion")

#Especificación del constructo corrupcionpo
corrupcionpo <- 'corrupcionpo=~cornacional+ corlocal +
                 corautonom + coreuropea'

#Análisis factorial confirmatorio de la variable latente corrupcionpo
ajuste.corrupcionpo <- cfa(corrupcionpo, data = datos,
                       estimator = "dwls", meanstructure = FALSE)

#Bondad de ajuste de la variable latente corrupcionpo
summary(ajuste.corrupcionpo, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste.corrupcionpo, c("chisq", "df", "cfi", "tli",
            "rmsea", "srmr"))

#Representamos el constructo corrupcionpo
semPaths(ajuste.corrupcionpo, whatLabels = "std")

#Parámetros estimados del modelo corrupcionpo
parameterestimates(ajuste.corrupcionpo)

#Alpha de Cronbach de la variable corrupcionpo
psych::alpha(datos[,c(19:22)])



#************************ VARIABLE CORRUPCIONPUB ************************

#Prueba KMO y de Fligner-Killeen para la variable corrupcionpub
KMO(datos[,c(8:13)])
fligner.test(datos[,c(8:13)])

#Gráfico de sedimentación de la variable corrupcionpub
```

```
scree(datos[,c(8:13)], main="Gráfico de sedimentación de la variable
      latente corrupcion")

#Especificación de la variable latente corrupcionpub
corrupcionpub <- 'corrupcionpub =~ fuerzasseg + adminpub +
                  subvenciones + licencias+ adminjust +
                  inspectores'

#Análisis factorial confirmatorio de la variable corrupcionpub
ajuste.corrupcionpub <- cfa(corrupcionpub, data = datos,
                            estimator = "dwls", meanstructure = FALSE,
                            std.lv = TRUE)

#Bondand de ajuste de la variable latente corrupcionpub
summary(ajuste.corrupcionpub, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste.corrupcionpub, c("chisq", "df","cfi", "tli",
                                    "rmsea", "srmr"))

#Se presentan los índices de modificación de la variable coorrupcionpub
modificationIndices(ajuste.corrupcionpub)

#Reformulación del constructo
corrupcionpub <- 'corrupcionpub =~ fuerzasseg + adminpub +
                  subvenciones + licencias+ adminjust +
                  inspectores
                  subvenciones ~~ licencias'
ajuste.corrupcionpub <- cfa(corrupcionpub, data = datos,
                            estimator = "dwls", meanstructure = FALSE,
                            std.lv = TRUE)
summary(ajuste.corrupcionpub, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste.corrupcionpub, c("chisq", "df","cfi", "tli",
                                    "rmsea", "srmr"))

#Representamos el constructo corrupcionpub
semPaths(ajuste.corrupcionpub, whatLabels = "std")

#Estimación de parámetros de la variable corrupcionpub
parameterestimates(ajuste.corrupcionpub)

#Alfa de cronbach de la variable corrupcionpub
psych::alpha(datos[,c(8:13)])



#************************ VARIABLE CORRUPCIONOS ************************

#Prueba KMO y de Fligner-Killeen para la variable corrupcionpub
```

```
KMO(datos[,c(14:18)])
fligner.test(datos[,c(14:18)])

#Gráfico de sedimentación de la variable corrupcionpub
scree(datos[,c(8:13)], main="Gráfico de sedimentación de la variable
      latente corrupcion")

#Especificación de la variable latente corrupcionos
corrupcionos <- 'corrupcionos =~ sindicatos + banca + ongs + medios +
                 empresarios'
ajuste.corrupcionos <- cfa(corrupcionos, data = datos,
                           estimator = "dwls", meanstructure = FALSE,
                           std.lv = TRUE)

#Bondad de ajuste de la variable latente corrupcionos
summary(ajuste.corrupcionos, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste.corrupcionos, c("chisq", "df","cfi", "tli",
            "rmsea", "srmr"))

#Estimación de parametros de la variable corrupcionos
parameterestimates(ajuste.corrupcionos)

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

#Alfa de cronbach de la variable corrupcionos
psych::alpha(datos[,c(14:18)])



#**ANÁLISIS FACTORIAL CONFIRMATORIO GENERAL DE LA VARIABLE CORRUPCIÓN**

#Prueba KMO y de Fligner-Killeen de la variable corrupcion
KMO(datos[,c(8:22)])
fligner.test(datos[,c(8:22)])

#Gráfico de sedimentación de la variable corrupcion
scree(datos[,c(8:22)], main="Gráfico de sedimentación de la variable
      latente corrupcion")

#Cargas factoriales de la variable corrupcion
corrupcion <- psych::fa(datos[,c(8:22)], nfactor = 3,
                        rotate = "varimax", fm="pa") %>%
                        model_parameters(sort = TRUE, threshold = "max")
```

```
corrupcion

#Especificamos la variable latente corrupcion
modelocor <- ' corrupcion =~ corrupcionpo + corrupcionpub + corrupcionos

               corrupcionos=~ sindicatos + banca + ongs + medios +
                              empresarios

               corrupcionpub =~ fuerzasseg + adminjust + subvenciones +
               licencias + inspectores + adminpub
               subvenciones ~~ licencias

               corrupcionpo =~  cornacional +corlocal + corautonom +
                                coreuropea'

#Análisis factorial confirmatorio de la variable latente corrupcion
ajuste.cor <- cfa(modelocor, orthogonal = FALSE, data = datos,
                  estimator = "dwls", meanstructure = FALSE)

#Bondad de ajuste del constructo corrupcion
summary(ajuste.cor, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste.cor, c("chisq", "df","cfi", "tli", "rmsea", "srmr"))

#Estimación de parámetros del constructo corrupcion
parameterestimates(ajuste.cor, standardized = TRUE)

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

#Alfa de cronbach de la variable corrupcion
psych::alpha(datos[,c(8:22)])



#*************************************************************************
#******************** MODELO DE EC. ESTRUCTURALES ************************
#*************************************************************************

modelo <- 'corrupcion =~ corrupcionpo + corrupcionpub + corrupcionos

           corrupcionos=~ sindicatos + banca + ongs + medios +
                          empresarios

           corrupcionpub =~ fuerzasseg + adminpub + subvenciones +
           licencias+  adminjust + inspectores
```

```
            subvenciones ~~ licencias

            corrupcionpo =~ cornacional + corlocal + corautonom +
                            coreuropea

            confinstitu =~ ppoliticos + gobcentral + poderjud +
                           gobautonom + goblocal + medcomun

            corrupcion ~ socioeconom
            confinstitu ~ socioeconom
            democracia ~ socioeconom

            corrupcion ~ democracia
            confinstitu ~ democracia

            confinstitu ~ corrupcion'

#Modelo de ecuaciones estructurales general
ajuste <- sem(modelo, orthogonal = FALSE, data = datos,
              estimator = "dwls", meanstructure = FALSE)

#Bondad de ajuste del modelo de ecuaciones estrcuturales general
summary(ajuste, standardized = TRUE, fit.measures = TRUE)
fitMeasures(ajuste, c("chisq", "df","cfi", "tli", "rmsea", "srmr"))

#Representamos el modelo de ecuaciones estructurales general
semPaths(ajuste, whatLabels = "std",style="lisrel",layout="tree3",
         sizeMan = 6, node.width = 0.5,sizeLat =8, reorder=FALSE,
         optimizeLatRes=TRUE, edge.label.position=.7,
         edge.label.cex = 0.5, edge.color = "darkblue",rotation=1 )

#Estimación de parámetros del modelo de ecuaciones estructurales general
parameterEstimates(ajuste, standardized=TRUE)
```

## A.2 Comprobación de las hipótesis previas

### *Tests de Shapiro-Wilk*

| Variable | Estadístico | p-valor | Normalidad |
|---|---|---|---|
| democracia | 0.9622 | 0.001 | No |
| ppoliticos | 0.9468 | 0.001 | No |
| gobcentral | 0.9338 | 0.001 | No |
| poderjud | 0.9379 | 0.001 | No |
| gobautonom | 0.9458 | 0.001 | No |
| medcomun | 0.9453 | 0.001 | No |
| goblocal | 0.9478 | 0.001 | No |
| fuerzasseg | 0.8704 | 0.001 | No |
| adminjust | 0.8729 | 0.001 | No |
| subvenciones | 0.8704 | 0.001 | No |
| inspectores | 0.8702 | 0.001 | No |
| licencias | 0.7907 | 0.001 | No |
| adminpub | 0.8917 | 0.001 | No |
| empresarios | 0.8627 | 0.001 | No |
| sindicatos | 0.8802 | 0.001 | No |
| banca | 0.8631 | 0.001 | No |
| ongs | 0.9111 | 0.001 | No |
| medios | 0.8942 | 0.001 | No |
| corlocal | 0.8638 | 0.001 | No |
| corautonom | 0.8783 | 0.001 | No |
| cornacional | 0.8574 | 0.001 | No |
| coreuropea | 0.8872 | 0.001 | No |
| socioeconom | 0.8764 | 0.001 | No |

Cuadro A.1: Tests de Shapiro-Wilk

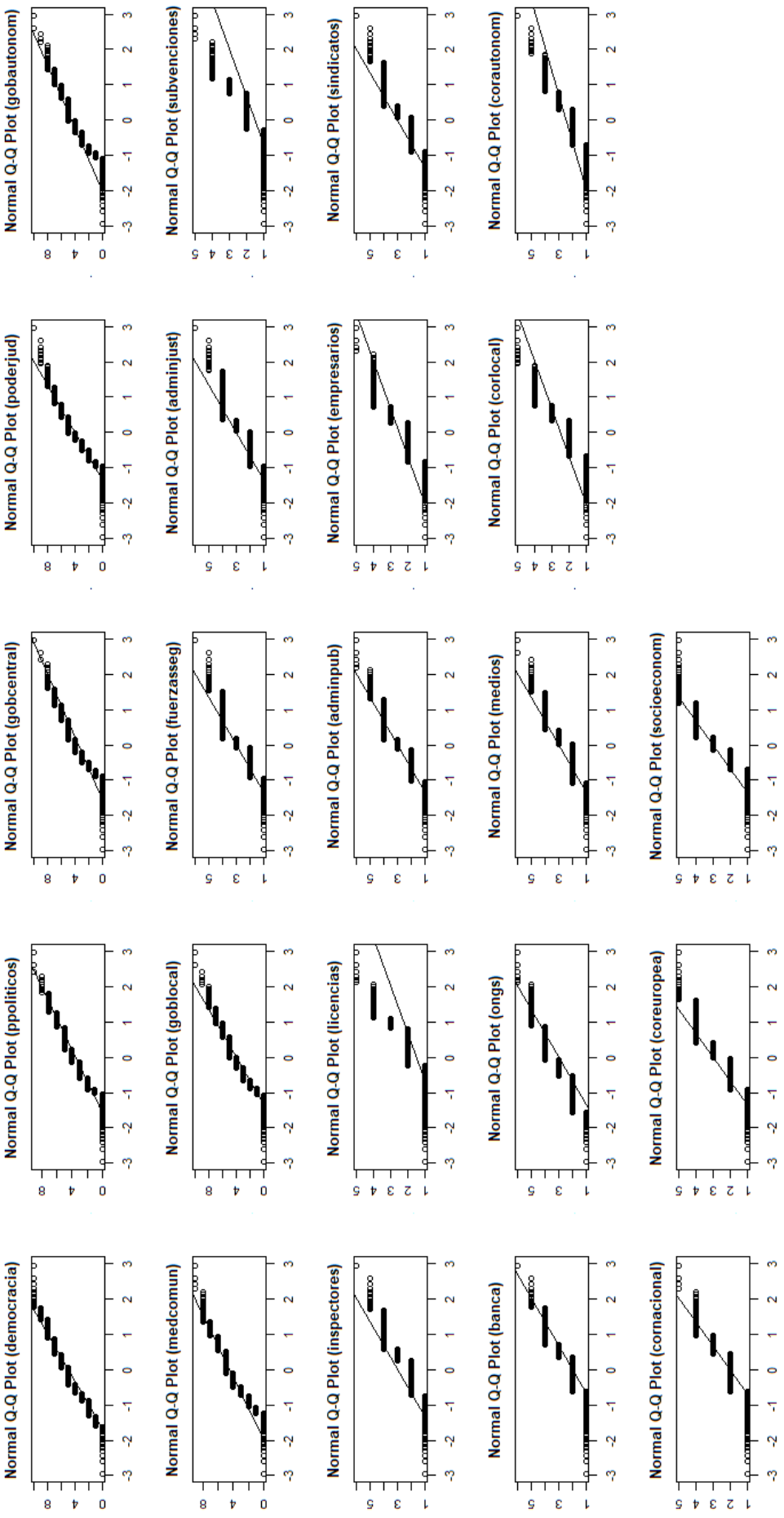


Figura A.1: QQplot de las variables

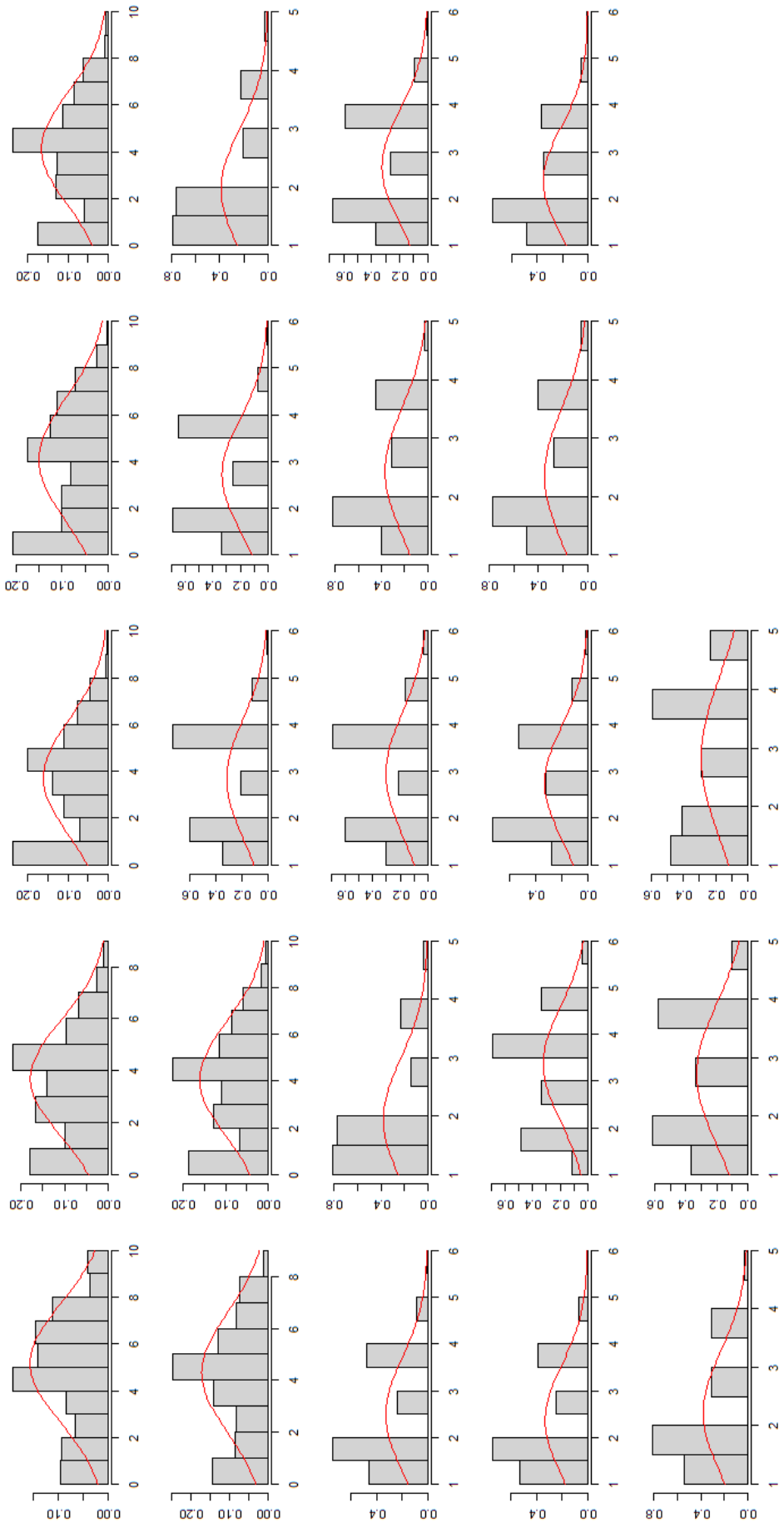

Figura A.2: Histrogramas de las variables

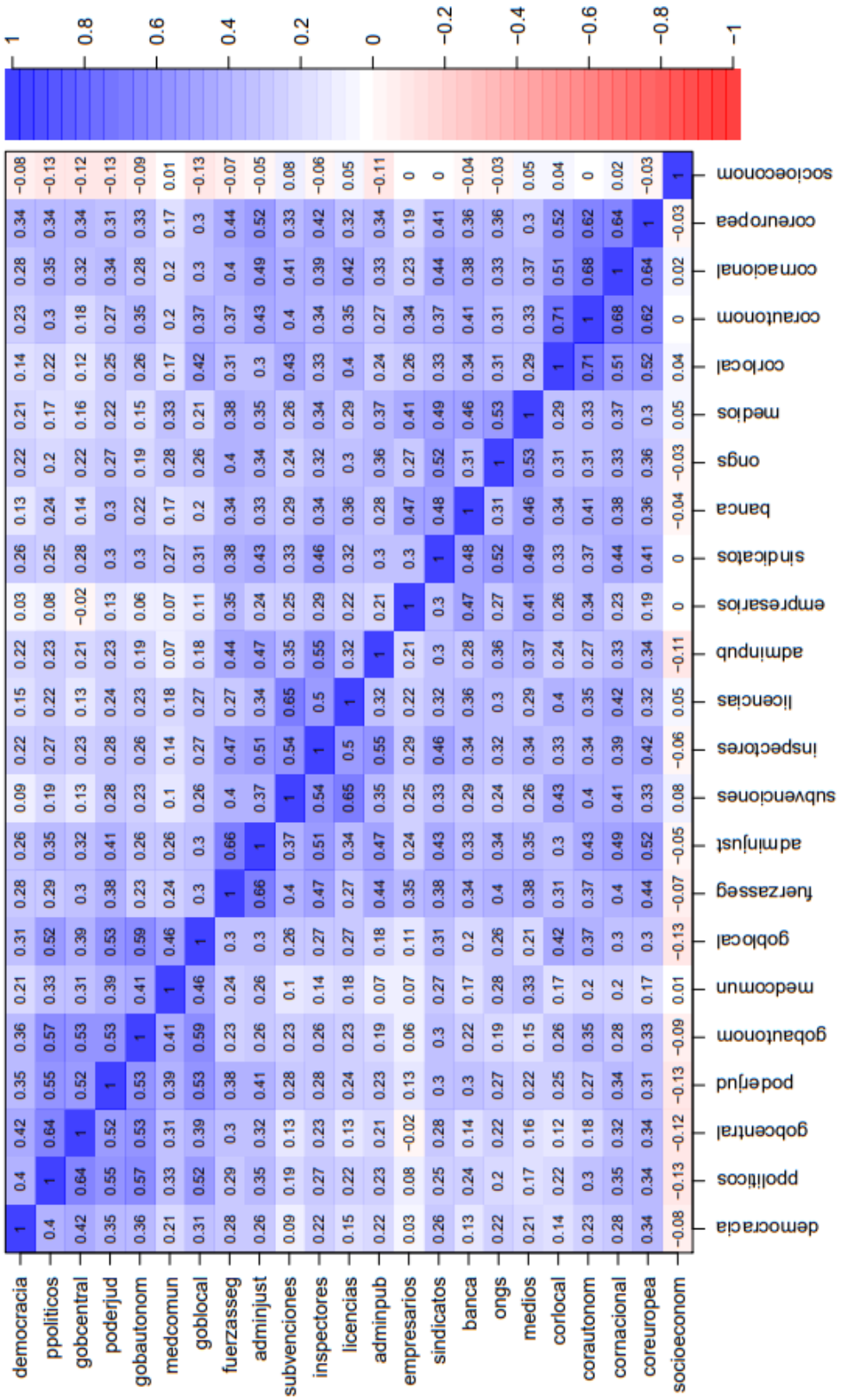


Figura A.3: Matriz de correlaciones

## A.3 Pruebas para la aplicación del análisis factorial

```
#Prueba KMO, MSA y Fligner-Killen de la variable corrupcionpub
KMO(datos[,c(8:13)])

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datos[, c(8:13)])
## Overall MSA =  0.8
## MSA for each item =
##   fuerzasseg    adminjust subvenciones  inspectores    licencias     adminpub
##         0.77         0.78         0.78         0.84         0.76         0.85

fligner.test(datos[,c(8:13)])

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  datos[, c(8:13)]
## Fligner-Killeen:med chi-squared = 76.684, df = 5, p-value = 4.141e-15
```

```
#Prueba KMO, MSA y Fligner-Killen de la variable corrupcionpo
KMO(datos[,c(14:18)])

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datos[, c(14:18)])
## Overall MSA =  0.78
## MSA for each item =
## empresarios  sindicatos       banca        ongs      medios
##        0.79        0.78        0.78        0.75        0.80

fligner.test(datos[,c(14:18)])

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  datos[, c(14:18)]
## Fligner-Killeen:med chi-squared = 17.255, df = 4, p-value = 0.001725
```

```
#Prueba KMO Y Fligner-Killen de la variable corrupcionos
KMO(datos[,c(19:22)])

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datos[, c(19:22)])
## Overall MSA =  0.76
## MSA for each item =
##    corlocal  corautonom cornacional  coreuropea
##        0.76        0.73        0.77        0.82

fligner.test(datos[,c(19:22)])

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  datos[, c(19:22)]
## Fligner-Killeen:med chi-squared = 17.342, df = 3, p-value = 0.0006011
```