



UNIVERSIDAD NACIONAL DE COLOMBIA

Modelos de Ecuaciones Estructurales: Conceptos y Aplicaciones

Trabajo presentado como requisito para optar el título de:
Magíster en Matemática Aplicada

Presentado por:
Yehidy Brand Uribe

Directora:
PhD. Nubia Esteban Duarte
Profesora Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

Universidad Nacional de Colombia
Facultad de Ciencias Exactas y Naturales, Departamento de Matemáticas
Manizales, Colombia
2021

Agradecimientos

En este largo camino de aprendizaje se han involucrado muchas personas las cuales me han brindado su gran amistad, acompañamiento y apoyo incondicional para cumplir esta etapa de mi vida.

En primera instancia quiero agradecer a Dios por darme esos factores importantes para luchar, los cuales son vida y salud.

Agradezco a mi Familia, Padre, Madre y Hermanas por siempre estar brindado su valioso apoyo, dándome ánimo en los momentos que tuve errores y mostrando que todo lo que deseé se puede lograr con mucha dedicación esfuerzo y sobre todo disciplina.

Agradezco a mis amigos incondicionales los cuales siempre estuvieron apoyándome en cada paso a dar y de parte de su gran corazón siempre tuve un sabio consejo.

El término gratitud es la palabra más adecuada para expresar lo que siento en el momento de pensar en todo el apoyo por parte de mi directora de Tesis, profesora Nubia Esteban Duarte, gracias por todo el tiempo dedicado y acompañamiento.

Agradezco a las Profesoras Támara Pessanha Taporoski y Sabine Pompeia por colaborarme con la base de datos y las discusiones para entender sobre estudios de cognición.

Gracias a mi pareja por la motivación expresada en los momentos de perder el rumbo, es el ingrediente perfecto para poder lograr alcanzar este dichoso y muy merecido logro.

Resumen

Modelos de Ecuaciones Estructurales: Conceptos y Aplicaciones

La metodología de ecuaciones estructurales (SEM, por sus siglas en Inglés, *Structural Equation Modeling*) es una técnica estadística multivariada, relativamente nueva, que combina tanto la regresión lineal múltiple como el análisis factorial. Permite al investigador no solo evaluar las muy complejas interrelaciones de dependencia entre las variables, sino también incorporar los efectos del error de medida sobre los coeficientes estructurales al mismo tiempo.

A través de la revisión de artículos científicos, se destaca la importancia de los modelos de ecuaciones estructurales en las más diversas áreas del conocimiento, particularmente sus aplicaciones se han extendido al área de Neuropsicología, siendo propuestos varios modelos para explicar los dominios cognitivos, pero a pesar de este avance, todavía hay varias líneas de investigación donde se hace relevante el estudio y la aplicación de la teoría asociada a ecuaciones estructurales.

En este trabajo se realiza una formalización teórica asociada a la metodología de ecuaciones estructurales partiendo desde la fundamentación matemática del análisis factorial. Estas metodologías son aplicadas a un conjunto de datos reales del Proyecto “Corazones de Baependi” (Processo Fapesp 2007/58150-7) del Laboratorio de Genética y Cardiología Molecular (Incor/USP). Entre los objetivos de este Proyecto están involucradas investigaciones que hacen referencia a las funciones cognitivas que son las capacidades mentales de una persona relacionadas a la memoria, la atención, el lenguaje y un aspecto muy importante que son las **funciones ejecutivas** (Friedman y Miyake, 2017) que será el foco de la aplicación del presente trabajo.

Palabras Clave: Análisis factorial, ecuaciones estructurales, funciones ejecutivas, secuencia de números.

Abstract

Structural Equation Models: Concepts and Applications

The Structural Equation Modeling methodology is a relatively new multivariate statistical technique that combines both Multiple Line Regression and Factor Analysis. It allows the researcher not only to evaluate the very complex internal relationship of the dependency between the variables, but also to incorporate the effects of the measurement of the error on the structural coefficients at the same time. The importance of structural equation models in the most diverse areas of knowledge is highlighted, through the review of scientific articles. In particular, its applications have been extended to the area of Neuropsychology, with various models being proposed to explain cognitive domains, but despite these advances, there are several lines of research where the application of the theory of structural equations becomes relevant.

In this thesis a theoretical formalization associated with the Structural Equations methodology is carried out based on the mathematical foundation of Factorial Analysis. These methodologies are applied to a set of real data from the “Corazones de Baependi” / “Baependi Hearts” Project (Processo Fapesp 2007 / 58150-7) of the Molecular Genetics and Cardiology Laboratory (Incor / USP). Among the objectives of this Project there are involved investigations that reference the cognitive functions that are of the mental capacities of a person which refers to memory, attention, language and some very important aspects as executive functions (Friedman and Miyake, 2017) that are the ones that will be the focus of the application of the current thesis.

Keywords: Factorial analysis, structural equations, executive functions, number sequence.

Índice general

Agradecimientos	2
Resumen	3
Abstract	4
1. Introducción	11
1.1. Antecedentes	12
1.1.1. Justificación	16
2. Algunos conceptos del área de Psicología	18
2.1. Definición e ilustración de las funciones ejecutivas	18
2.2. Ilustración de la obtención de las variables	20
3. Análisis Factorial	24
3.1. Análisis Factorial Ortogonal	25
3.1.1. Supuestos del modelo	26
3.1.2. Diagrama de caminos de un modelo factorial ortogonal . . .	27
3.2. Algunos resultados sobre el modelo de análisis factorial	27
3.3. Métodos de estimación	30
3.3.1. El método de Componentes Principales	30
3.3.2. El método de Máxima Verosimilitud	32
3.4. Rotación de factores	36

3.5. Análisis de la matriz de correlación	39
3.5.1. Test de Esfericidad de Bartlett	39
3.5.2. Matriz Anti-imagen	40
3.5.3. Coeficiente de correlación parcial	40
3.5.4. Medidas de adecuación de la muestra	41
4. Modelo de Ecuaciones Estructurales	43
4.1. Tipos de variable	44
4.2. Modelo Matemático	45
4.2.1. Modelo Estructural	45
4.2.2. Modelo de Medida.	45
4.3. Supuestos del modelo	46
4.4. Diagrama de caminos	47
4.5. Efecto Directo, Indirecto y Total	53
4.5.1. Indicadores reflectivos y formativos	55
4.6. Matriz de covarianza implícita	56
4.7. Identificación del Modelo	60
4.8. Métodos de Estimación	62
4.8.1. Método de Máxima Verosimilitud	63
4.8.2. Mínimos Cuadrados Generalizados	64
4.8.3. Mínimos Cuadrados no Ponderados	66
4.9. Diagnóstico de ajuste del modelo	67
4.9.1. Ajuste General del modelo	67
4.9.2. Medidas de ajuste alternativas	69
5. Aplicación	74
5.1. Descripción del proyecto	74
5.2. Análisis Factorial - Aplicación	75
5.3. Modelo de ecuaciones estructurales que incluye correlaciones entre las funciones ejecutivas	83

ÍNDICE GENERAL	7
5.4. Modelo alternativo de ecuaciones estructurales: Relaciones de causalidad entre las funciones ejecutivas	88
5.4.1. Identificación del modelo	90
6. Conclusiones	93
A. Álgebra de matrices	96
A.1. Matriz traspuesta y propiedades	96
A.2. Derivación matricial	96
B. Algoritmos de optimización	98
B.1. Los métodos de la segunda derivada	99
B.1.1. Método de Newton-Raphson	99
B.1.2. Método de Quadratic Hill-Climbing (<i>Goldfeld-Quandt</i>)	99
C. Programas y Paquetes Estadísticos	101
C.1. LISREL:	101
C.2. AMOS	102
C.3. Programa R	108
D. Descripción de las variables	111
E. Modelo incluyendo edad y escolaridad	113
F. Modelos SEM: Primeros ensayos	115

Índice de tablas

3.1. Matriz Σ	34
3.2. Cargas factoriales	34
3.3. Resumen de los cálculos presentados en el ejemplo asociado a análisis factorial	35
3.4. Medida de KMO	42
4.1. Fórmulas de efectos (Bollen, 1989)	55
5.1. Estadísticas descriptivas de las variables incluidas en el estudio (N=717)	76
5.2. KMO y prueba de Bartlett	76
5.3. MSA para cada variable	77
5.4. Varianza total explicada: Método de extracción por máxima verosimilitud	77
5.5. Tabla de cargas factoriales sin rotación	78
5.6. Matriz de cargas factoriales utilizando el método de rotación varimax	78
5.7. Estimación de parámetros	85
5.8. Estimación de parámetros- covarianza	85
5.9. Índices de bondad de ajuste	86
5.10. Estimación de parámetros SEM	91
5.11. Estimación de parámetros- covarianza-SEM	91
5.12. Índices de bondad de ajuste	92

Índice de figuras

3.1. Diagrama de camino de un modelo de análisis factorial ortogonal	27
4.1. Representación gráfica de las variables del modelo	48
4.2. Diagrama de caminos para un modelo de ecuaciones estructurales completo	49
4.3. Diagrama de caminos para el modelo estructural	50
4.4. Diagrama de caminos para el modelo de medida	51
4.5. Relación simple	53
4.6. Relación recíproca	53
4.7. Relación espúrea	54
4.8. Relación indirecta	54
4.9. Diagrama de indicador reflectivo para el modelo de medida	56
4.10. Diagrama de indicador formativo para el modelo de medida	56
5.1. Variables asociadas a cada uno de los factores	79
5.2. Histograma y Boxplot de las variables asociadas con Inhibición	81
5.3. Histograma y Boxplot de las variables asociadas con Actualización	82
5.4. Histograma y Boxplot de las variables asociadas con Alternancia	83
5.5. Diagrama de caminos obtenido para los datos del proyecto de Bae- pendi	84
5.6. Diagrama de caminos modelo alternativo	88

5.7. Diagrama de caminos modelo alternativo: Estimación de los parámetros	89
C.1. Panel principal AMOS	102
C.2. Modelo de Ecuaciones Estructurales	103
C.3. Iconos clave en AMOS	104
C.4. Menú opciones AMOS	104
C.5. Menú de propiedades de las variables utilizadas en el modelo	105
C.6. Menú Análisis en AMOS	106
C.7. Modelo de ecuaciones estructurales estimado por máxima verosimilitud	107
C.8. Bondad de ajuste	107
E.1. Diagrama de caminos con las variables : edad y escolaridad	114
F.1. Diagrama de caminos con la variable combinación	116
F.2. Diagrama de caminos modelo fallido con la variable combinación	117
F.3. Diagrama de caminos con aciertos	118
F.4. Diagrama de caminos con aciertos y efectos	119

Capítulo 1

Introducción

Los modelos de ecuaciones estructurales conforman una técnica estadística multivariada actual, con una serie de ventajas sobre las técnicas más comúnmente usadas (como regresión múltiple y análisis factorial). Permite la prueba de teorías sobre diferentes fenómenos, incorpora el uso no sólo de variables observables sino de variables latentes. Permite que se evalúen diversas relaciones posibles entre las variables observables y latentes, sean dependientes o independientes, (Kline, 2011).

La regresión múltiple, el análisis factorial, el análisis multivariante de la varianza, el análisis discriminante y otras técnicas comunes proporcionan al investigador un conjunto de poderosas herramientas a la hora de tratar con un amplio abanico de cuestiones prácticas y teóricas, pero cada técnica puede examinar sólo una relación al mismo tiempo. Incluso las técnicas que tienen en cuenta varias variables dependientes, como el análisis multivariante de la varianza y el análisis canónico, siguen representando sólo una única relación entre las variables dependientes e independientes. Sin embargo, el investigador se enfrenta a un conjunto de cuestiones interrelacionadas, por ejemplo, ¿cuántas variables determinan la imagen de una tienda? ¿cómo se combina esa imagen con otras variables para afectar las decisiones de compra? y finalmente, ¿cómo se convierte la satisfacción con la tienda en la lealtad a largo plazo? Esta serie de asuntos tienen importancia tanto teórica como práctica y ninguna de estas metodologías permite tratar con todas estas cuestiones a la vez mediante una técnica exhaustiva. Por esta razón se estudia la técnica de modelos de ecuaciones estructurales como una extensión de varias técnicas multivariadas, entre ellas la regresión múltiple y el análisis factorial, (Hair y cols., 1999).

1.1. Antecedentes

Los modelos de ecuaciones estructurales surgen en el año 1921 por el genetista Sewall Wright, quien los implementó en la genética, con el fin de analizar un conjunto de relaciones entre una o más variables independientes, desarrollando una forma de romper las altas correlaciones observadas en un sistema de ecuaciones matemáticas que describían sus hipótesis respecto a unas relaciones causales, estas relaciones entre las variables fueron representadas en un “path diagram” (diagrama de caminos), también conocido como “Análisis *Path*”, un método que permite evaluar el ajuste de modelos teóricos en los que se propone un conjunto de relaciones de dependencia entre variables.

En los años 70 otros investigadores reconocieron la importancia del diagrama de caminos y comenzaron a trabajar en esta dirección, como Lawley (1940). Boudon (1965) y Duncan (1969) propusieron otra variante de análisis multivariante vinculada a los diagramas de caminos para ponderar los efectos causales. Blalock (1969) planteó que todo modelo causal se basa en una serie de supuestos y se refiere a un número limitado de variables. Se postula en ellos una relación de causa-efecto (o de causación recíproca) entre variables (independientes, dependientes), buscando mantener aislados los posibles efectos de otras variables causales.

Jöreskog (1978), desarrolló modelos de ecuaciones estructurales e incorporó las características del análisis *Path* (diagrama de caminos), sus logros se deben a la combinación de elementos del análisis de caminos y el análisis factorial estableciendo las relaciones de las variables. Esta conjunción de planteamientos dio lugar a lo que ahora se llama modelo de medida, que relaciona variables latentes con variables observadas en las de ecuaciones estructurales, y el análisis *Path*, con el que se establecen las relaciones entre variables latentes, se conoce como modelo de estructura, también dedujo la base de los modelos de ecuaciones estructurales, en los que se incorpora el error de medida y las relaciones entre variables latentes.

Matemáticamente, estos modelos son más complejos de estimar que otros modelos multivariados como los de regresión o análisis factorial exploratorio y por ello su uso no se extendió hasta 1973, momento en el que apareció el programa de análisis LISREL (por sus siglas en Inglés, *Linear Structural Relations*) (Jöreskog, 1973). El programa LISREL fue perfeccionado, dando lugar al LISREL VI (Jöreskog y Sörbom, 1986), que ofrecía una mayor variedad de métodos de estimación. El EQS (Abreviatura de *Structural Equations*) (Bentler, 1985) es el otro paquete utilizado tradicionalmente para este tipo de análisis. En la actualidad, existen otros programas de estimación en entorno gráfico, como el AMOS (por

sus siglas en Inglés, *Analysis of Moment Structures*) (Arbuckle, 1997). En la actualidad, es posible utilizar también el programa STATA y el programa R para estimar estos modelos (StataCorp., 2013) y (R Core Team, 2017).

Desde hace algunos años, también mencionado en Hair y cols. (1999), los modelos de ecuaciones estructurales se han utilizado en casi todos los campos de estudio concebibles, incluyendo la educación, el marketing, la psicología, la sociología, la salud, la demografía, el comportamiento organizacional, la biología e incluso la genética. Las razones de su atractivo para áreas tan diversas son dobles:

1. Proporciona un método directo de tratar con múltiples relaciones simultáneamente a la vez que se da eficiencia estadística,
2. Es notable la capacidad para evaluar las relaciones exhaustivamente y proporcionar una transición desde el análisis exploratorio al confirmatorio.

Algunos estudios mas recientes se describen a continuación. En Zamora (2008) se presenta una secuencia de modelos matemáticos y estadísticos útiles para describir las relaciones entre varias variables que permiten analizar un fenómeno de interés, teniendo en cuenta la estructura de covarianzas existente entre ellas. Plantea ocho aspectos del estado de salud reagrupables en dos dimensiones de segundo orden: salud física y salud mental tanto en hombres como en mujeres, a pesar de que existen diferencias en algunas variables con respecto al género. También en Mateos y Morales (2011), los trabajos son dedicados a la teoría y estructura de los modelos de ecuaciones estructurales. Casas (2016) utiliza estos modelos para determinar el índice de satisfacción al cliente en toda Europa, estimando los parámetros del modelo con ecuaciones lineales simultáneas. En el área de calidad, Cuautle (2010) utiliza estos modelos para proponer un modelo de gestión integral para las empresas mexicanas de manufactura relacionado a la transferencia del conocimiento, Maynez (2011) los implementó en transferencia de conocimiento organizacional como fuente de ventaja competitiva sostenible: modelo integrador de factores y estrategias.

En la educación, se están implementando distintas investigaciones como la de Cervelló y cols. (2004), donde analizan la motivación de los alumnos en las clases de educación física y su percepción del clima motivacional establecido por el profesor en el aula. Lozano (2007) valida un modelo para determinar las dificultades a las que se enfrentan los estudiantes en el momento de decidir qué carrera estudiar. Moneta y cols. (2018) implementaron estos modelos en una propuesta para explicar el desempeño académico en *e-learning*, se centraron en el proceso de aprendizaje bajo entornos virtuales (*e-learning*) como principal

variable latente endógena del modelo, postulando la importancia de los conocimientos previos, la motivación, las competencias digitales, la autorregulación y los procesos de interacción. Madrigal y Ávila (2019) utilizaron investigaciones sociales enfocadas al comportamiento de compra de los *Millennials*, se realizó una selección y análisis de artículos que abordan investigaciones aplicadas a los jóvenes y que además implican la incidencia tecnológica en ellos.

Por otro lado, en el área de psicología, se han publicado varios artículos desde diferentes campos de investigación, por ejemplo investigaciones que hacen referencia a las funciones cognitivas que son las capacidades mentales de una persona relacionadas a la memoria, la atención, el lenguaje, y un aspecto muy importante que son las **funciones ejecutivas** que será el foco de la aplicación del presente trabajo.

Para entender un poco sobre el campo de las **funciones ejecutivas**, en primer lugar se define **la memoria de trabajo** que es uno de los constructos cognitivos más influyentes y que más atención ha recibido en las últimas décadas. Puede definirse como la capacidad para mantener temporalmente activa la información para su utilización en diferentes actividades cognitivas como comprender o pensar. Esta capacidad es importante en tanto que se relaciona con el éxito de la persona en muchas actividades y tareas cognitivas, (Pelegriña y cols., 2016). Ahora, **funciones ejecutivas** son habilidades cognitivas involucradas en la formación, planificación, selección, mantenimiento, seguimiento e implementación de metas que se mantienen temporalmente en la memoria de trabajo. Por lo tanto, son esenciales para la abstracción, control del pensamiento, emociones, el comportamiento y la acción (Miyake y Friedman, 2017), lo que explica por qué están asociadas o predicen muchos resultados de la vida, como el éxito en el aprendizaje y rendimiento académico así como también profesional, el comportamiento interpersonal y social, la toma de riesgos, consumo de drogas, así como salud física y mental (Moffitt y cols., 2011; Baggeta y Alexander, 2016).

Existen muchos modelos de **funciones ejecutivas**, el más citado de los cuales (Baggeta y Alexander, 2016) es el modelo de unidad y diversidad de funciones ejecutivas (Miyake y cols., 2012; Friedman y Miyake, 2017). Este modelo propone que existen tres dominios ejecutivos correlacionados, a nivel de variables latentes, es decir, variables que se obtienen matemáticamente de los puntajes en más de una tarea que involucra procesos cognitivos similares. Los dominios investigados en el modelo antes mencionado son:

- Inhibición de respuestas automáticas, es decir, la capacidad de inhibir el comportamiento automático, en otras palabras poder concentrarse en una tarea ignorando fuentes de distracción;
- Alternancia, la capacidad de cambiar entre dos o más tareas;
- Actualización, la capacidad de actualizar los contenidos almacenados en la memoria de trabajo para mantener temporalmente sólo la información que es relevante para alcanzar un objetivo determinado.

Aunque se ha encontrado que las variables latentes de estos dominios son distinguibles (diversidad), comparten una habilidad común que se denomina unidad de funciones ejecutivas. Este fraccionamiento de dominios ejecutivos se ha replicado en muchos tipos de muestras, refleja fundamentos fisiológicos distinguibles que parecen ser en gran parte heredados genéticamente (ver Miyake y Friedman, 2012; Friedman y Miyake, 2017).

En relación a funciones ejecutivas, hay publicaciones en algunos ámbitos, por ejemplo Lawson y cols. (2014) proporcionan pruebas consistentes de que el nivel socioeconómico se relaciona sistemáticamente con la función ejecutiva, y los niños con un nivel socioeconómico bajo muestran un peor desempeño en muchas tareas y enfoques metodológicos.

Por otro lado, Friedman y Miyake (2017) definen las funciones ejecutivas y revisan la evidencia de que en múltiples edades y poblaciones, las funciones ejecutivas comúnmente estudiadas (a) están fuertemente correlacionadas cuando se miden con variables latentes; (b) no son lo mismo que inteligencia general; (c) son altamente heredables a nivel latente y aparentemente también altamente poligénicas o heredables.

Sin embargo, en la literatura hay varias discusiones al respecto de las definiciones y conceptos asociados con funciones ejecutivas que es un área de estudio que está en rápida expansión de desarrollo, los conceptos y las etiquetas verbales a menudo se usan indistintamente y las tareas a veces se proponen como medidas de una u otra construcción psicológica sin suficiente evidencia de apoyo. A este respecto, en el artículo de Morra y cols. (2018) se presenta un estudio que, se espera, pueda contribuir a corregir la claridad terminológica en un campo de investigación floreciente actualmente, como ellos lo manifiestan.

Un estudio que ha sido ampliamente difundido en la literatura en Miyake y cols. (2000) que envuelve las funciones ejecutivas o dominios cognitivos (inhibición, alternancia y actualización) fue realizado en la Universidad de Colorado, en que 137 estudiantes universitarios recibieron crédito parcial del curso por participar en el estudio. Los participantes completaron 9 tareas, tres asociadas

a cada dominio y mediante un estudio de ecuaciones estructurales encontraron asociaciones entre las tres tareas para cada dominio y aún, más importante dentro de los resultados del modelo de ecuaciones estructurales se destaca la correlación significativa desde el punto de vista estadístico de estos tres factores o funciones ejecutivas.

En el contexto de este problema, con el objetivo de replicar este estudio a habitantes de una población brasilera, varios investigadores del área de Neuropsicología tomaron medidas de varias tareas encaminadas a explicar estos factores y al mismo tiempo estudiar la correlación de las tres funciones ejecutivas en la población blanco de estudio. Así, dentro del contexto de este problema se utilizará la teoría asociada al análisis factorial así como la teoría de ecuaciones estructurales para dar respuesta al problema planteado y así contribuir más desde el punto de vista teórico en el área de Neuropsicología.

Los participantes del estudio pertenecen al proyecto “Corazones de Baependi” (Processo Fapesp 2007/58150-7) conducido por el Laboratorio de Genética y Cardiología Molecular (Incor/USP) (de Oliveira y cols., 2008). El proyecto tiene como objetivo general encontrar determinantes genéticos que modulan o regulan enfermedades cardiovasculares evaluando la influencia de factores genéticos y ambientales. Sin embargo, un estudio que está contribuyendo al proyecto es en el área de Neuropsicología que tiene varios frentes de estudio como lo son estudios de depresión y ansiedad y últimamente estudios asociados al área de cognición, específicamente lo relacionado con funciones ejecutivas, que es el foco de este estudio. Al igual que en el estudio presentado en Miyake y cols. (2000), a los 717 participantes, de la población brasilera, se les realizaron 8 pruebas o tareas asociadas con las funciones ejecutivas y sobre estas 8 tareas es que se centra la aplicación de este estudio. La descripción de las pruebas o tareas y variables analizadas se encuentra en el Capítulo 2 y el Capítulo 5 de Aplicaciones.

1.1.1. Justificación

En la literatura existen muchas investigaciones en las más diversas áreas en las que en sus análisis utilizan modelos de ecuaciones estructurales, pero en la gran mayoría se deja totalmente implícita la fundamentación teórica. Siendo así, este trabajo surge de la necesidad de adquirir una fundamentación matemática sólida de la metodología de ecuaciones estructurales partiendo desde el análisis factorial.

Por otro lado, en el área de Psicología Cognitiva existe el objetivo de investigar la utilidad de las tareas o *test* con el objetivo de estudiar el modelo utilizando

datos reales de la población en estudio, se pueden replicar al estudio consagrado en la literatura donde ya existen asociaciones o correlaciones (Miyake y cols., 2000), lo cual es de gran utilidad para continuar con investigaciones como lo es un estudio de heredabilidad y asociaciones con otras variables como lo son la edad y el grado de escolaridad de los individuos participantes. Se pretende comprobar o refutar esta teoría, que como se acaba de anotar, es relevante para que el grupo de investigación pueda dar continuidad a otros estudios de gran interés para establecer asociaciones con otras áreas.

Como se mencionó anteriormente, para la aplicación del presente trabajo, existen datos de investigaciones que dan la oportunidad de fundamentar la teoría y realizar las respectivas aplicaciones, específicamente, se utilizarán datos del Proyecto “Corazones de Baependi” (Processo Fapesp 2007/58150-7) del laboratorio de Genética y Cardiología Molecular (Incor/USP), (Oliveira y cols., 2008, Giolo y cols., 2009).

El objetivo general y los objetivos específicos son presentados a continuación:

Objetivo General

El presente trabajo tiene como objetivo general la formalización matemática de la teoría asociada a modelos de ecuaciones estructurales estableciendo ejemplos ilustrativos de la teoría y aplicaciones a datos reales que envuelven variables del campo de la Psicología.

Objetivos Específicos

1. Fundamentar matemáticamente la teoría asociada a modelos de ecuaciones estructurales, teniendo en cuenta el método multivariado de análisis factorial.
2. Establecer patrones que permitan el estudio de las funciones de cognición asociadas con las funciones ejecutivas.
3. Investigar sobre los programas que se utilizan para llevar a cabo las aplicaciones de los modelos de ecuaciones estructurales, entre ellos el Programa **SPSS** y los paquetes *lavaan* y *SEM* del Programa **R**.
4. Plantear el modelo de ecuaciones estructurales al objeto de estudio como ilustración de la teoría formalizada, desde el modelo propuesto hasta la validación y conclusiones.

Capítulo 2

Algunos conceptos del área de Psicología

En este corto capítulo serán introducidos algunos conceptos importantes del área de Psicología, los cuales ayudarán en la comprensión de la aplicación de la teoría abordada en este trabajo, a saber aplicaciones del análisis factorial y ecuaciones estructurales a un problema muy específico del área de Neuropsicología. También se definirá la forma de obtención de las 8 variables que se utilizarán en este estudio. Los psicólogos aportan su conocimiento teórico en la materia y los datos experimentales de sus estudios, mientras que los matemáticos aportan su experiencia en estrategias de modelización y técnicas estadísticas, con la esperanza de que de esta simbiosis surjan modelos útiles y robustos.

2.1. Definición e ilustración de las funciones ejecutivas

Memoria de trabajo: La memoria de trabajo es uno de los componentes cognitivos que más está involucrado en el aprendizaje humano. Se puede describir como el conjunto de procesos mentales encargados del almacenamiento y la manipulación de la información de manera temporal (Baddeley y Hitch, 1974). Igualmente, está íntimamente relacionada con la memoria de largo plazo, ya que ambas poseen múltiples vínculos para recuperar y procesar información ya almacenada. Gracias a estas características, la memoria de trabajo representa un elemento crucial para diferentes tipos de investigación, (Baddeley, 2012).

Miyake y cols. (2000) descubrieron que la capacidad cognitiva de una persona podría ser explicada dentro de tres variables latentes dependientes entre sí, llamadas **funciones ejecutivas**. Como anotado anteriormente, las funciones ejecutivas son habilidades cognitivas involucradas en la formación, planificación, selección, mantenimiento, seguimiento e implementación de metas que se mantienen temporalmente en la memoria de trabajo (Miyake y Friedman, 2017). Específicamente las funciones ejecutivas son: *inhibición, alternancia y actualización*. Estas funciones ejecutivas al ser funciones de nivel inferior en comparación con otras funciones ejecutivas (funciones de planificación) pueden definirse operativamente de una manera bastante precisa ya que para su descripción normalmente se dispone de una serie de tareas cognitivas relativamente simples y bien estudiadas. Sin embargo, es posible que estas funciones estén implicadas en la ejecución y rendimiento de pruebas ejecutivas más complejas.

A continuación se definen las funciones ejecutivas de forma específica y se ilustran con algunos ejemplos prácticos.

- **Inhibición** : La inhibición o control inhibitorio, es definida como la capacidad del ser humano para controlar las respuestas impulsivas (o automáticas) y generar respuestas mediadas por la atención y el razonamiento. Más específicamente es conseguir mantener latente alguna cosa que es periférica y no es el foco de lo que se está realizando. En términos de medición de la inhibición, una prueba o tarea consiste en medir la capacidad que un individuo tiene para aleatorizar. Por ejemplo, se le presenta una secuencia numérica 1, 2, 3, 4; es de esperarse que al indicarle a esa persona que continúe con la secuencia de números escoja aquel que está definido por la sucesión (5) a pesar de que se le brinda total libertad a la hora de escoger el siguiente número. La tendencia es a continuar la secuencia con el número 5, pero el individuo tiene que inhibir el 5 (que es lo que se tiene en la cabeza) para poder decir otro número lo más alejado posible dentro del rango de valores establecido. De manera que si la respuesta es un número distinto al esperado la persona tendrá una nota de inhibición alta. (La tarea se repite un número determinado de veces).
- **Actualización**: Consiste en la sustitución de los contenidos mantenidos en memoria. Un ejemplo de actualización se da en la lectura. Cuando una persona va leyendo, va actualizando la información por otra, específicamente actualizando las letras y las palabras a lo largo del tiempo. En un momento está leyendo una información, cambia esa proyección por otra información y

no ha borrado de su memoria la anterior, sino que actualiza con la nueva información.

- **Alternancia:** Capacidad para cambiar de manera flexible las operaciones o *test* mentales, es decir, alternar entre tareas.

Por ejemplo, una persona necesita ir al mercado y requiere ir con su hijo, entonces debe estar pendiente al mismo tiempo de la lista del mercado y del bienestar del niño, por lo tanto, está alternando constantemente entre estas dos tareas. Así, la alternancia es la capacidad de alternar actividades en el menor tiempo posible.

Otro ejemplo muy práctico es cuando una persona está escribiendo y alguien le interrumpe durante la ejecución de su tarea, la persona deja de escribir y le presta atención a su interlocutor hasta el momento en el que regresa a su actividad. Si en el momento en que recupera su actividad está perdida, entonces se dice que su capacidad de alternancia es baja. De manera que cuanto mayor sea la capacidad de alternancia, menos será la distracción.

Según las definiciones y ejemplos anteriores, se puede inferir que para ejecutar por ejemplo una tarea 1, el individuo debe focalizar la atención ignorando otros estímulos ajenos (se debe inhibir), pero la Alternancia permite cambiar de tarea exitosamente y para lograrlo se debe cambiar el foco entre una tarea y otra. Por su parte la actualización permite saber la fase de ejecución de la tarea 1, por lo que terminar esta tarea satisfactoriamente requiere de actualización e inhibición y la alternancia exige tener un grado de inhibición y actualización para poder cambiar de tareas de forma exitosa. Lo anterior, que expresa las relaciones entre las tres funciones ejecutivas, en términos estadísticos se puede escribir que se esperarían correlaciones significativas entre las funciones ejecutivas.

2.2. Ilustración de la obtención de las variables

Los números aleatorios son la base esencial de una simulación. Usualmente, toda la aleatoriedad involucrada en el modelo se obtiene a partir de un generador de números aleatorios que produce una sucesión de valores que son realizaciones de una secuencia de variables aleatorias independientes e idénticamente distribuidas.

En el caso del presente estudio, área de Neuropsicología, los investigadores utilizaron el Programa *RgCalc* como generador de números aleatorios. Específicamente, *RgCalc* (para Windows95) es un programa informático para ayudar a la

investigación psicológica referente al desempeño de la generación aleatoria en los humanos. Está diseñado para evaluar el orden en una secuencia de elementos (es decir, medir la falta de aleatoriedad) y fue escrito por Towse y Neil (1998).

Como se describió anteriormente, el programa *RgCalc* está diseñado para el análisis de varios tipos de situaciones de generación aleatoria empleadas en la literatura psicológica. Siendo así, con la ayuda de este programa se obtuvieron las medidas de 8 pruebas o tareas aplicadas a la población en estudio. Esas pruebas o tareas son específicamente las 8 variables que se espera que estén asociadas con las tres funciones ejecutivas (inhibición, alternancia y actualización). A continuación se presenta una breve descripción de las variables utilizadas para esta aplicación y en el Apéndice D se presentan las fórmulas que el Programa usa en la generación aleatoria. Las secuencias van de 0 a 100 y una condición fundamental es que están diseñados para que el individuo en sus respuestas no cambie de decena

- **TPI:** *Turning Point Index (TPI)* o índice de punto de cambio, calcula el número de respuestas que, como valores numéricos, marcan un cambio entre secuencias ascendentes y descendentes (es decir, puntos que representan picos). Por ejemplo se tiene los siguientes casos donde puede ocurrir que haya un TPI:
 - En la secuencia de números "1,2,3,4,5,6,3", hay un único punto de cambio en la respuesta "6", ya que la secuencia comienza a descender en este punto.
 - Si se tiene la secuencia de números "9,8,7,6,5,8,9", se puede apreciar entonces dos TPI, en las respuestas "5", ya que en este punto la secuencia empieza a ascender.
 - Un último caso en donde puede ocurrir un TPI, es en la secuencia "3,4,5,5,5,4", contiene un único punto de cambio en la respuesta repetida "5".
- **Adyacencia:** Lo que busca esta medida de frecuencia es que no hayan pares consecutivos, es decir, en la secuencia "1,2,4,6,7" lo ideal es no responder los pares consecutivos de la forma "1,2" y "6,7", ya que la idea principal es que la persona responda datos lo más aleatoriamente posible.
- **Runs:** Es la varianza de una secuencia cuando se presente un TPI, y si esta varianza es alta, significa que la persona está logrando ser lo más aleatoria

posible. Por ejemplo, tomando los 3 casos donde puede ocurrir un TPI, se tiene que:

- En la secuencia "1,2,3,4,5,6,3" se halla la varianza de la secuencia antes de que se generó el TPI, es decir, en la secuencia "1,2,3,4,5,6", la varianza es igual a 3,5.
 - Tomando la secuencia "9,8,7,6,5,8,7,9", por tanto se calcula la varianza de "9,8,7,6,5,8,7", así la varianza es 1,80.
 - Por último, en la secuencia "3,4,5,5,5,4", se calcula la varianza de "3,4,5,5,5", la cual es 0,80.
- **Coupon:** Un coupon es la reorganización de una secuencia, por ejemplo el objetivo es tomar números aleatorios y si la respuesta es la secuencia "1,2,3,4,5", en esta secuencia no se está logrando ser aleatorio, ya que se está haciendo una secuencia predefinida de términos, ahí la medida de Coupon es baja.
 - **Redundancia:** La redundancia consiste en los números que más se repiten en una secuencia, por ejemplo si una persona responde "1,2,2,2,3,4,5,5,5,5", está siendo redundante en los números "2" y "5", por ello si la persona es menos redundante eso quiere decir que está logrando ser aleatorio en sus respuestas.
 - **Repgap (Repetición Gap):** La repetición Gap consiste en tomar el promedio de las distancias de repetición de los números, por ejemplo si se tiene la secuencia "1,2,3,2,1", la persona en esta secuencia repitió el número "2", entonces la distancia entre esta repetición es sólo un número, correspondiente al número "3". Tomando ahora la secuencia "2,4,6,8,10,6", en este caso la distancia es 2 ya que hay 2 números en medio para volver a repetir el número 6, por tanto la Repetición Gap es el promedio entre estas 2 distancias.
 - **ef_pmerrorsalt:** Consiste en una medida de cambio ejecutivo que consiste en 3 subtareas en las que se presentan listas de números de dos dígitos. La primera (tarea más) y la segunda (tarea menos), incluye 15 números de dos dígitos cada una. Los individuos debían sumar tres (tarea más) y restar tres (tarea menos), respectivamente, a cada número de dos dígitos lo más rápido posible, anotando las respuestas junto a cada número de dos dígitos a medida que avanzaba la tarea. La tercera (tarea alterna), se trata de incluir 30 números de dos dígitos y los participantes debían alternar

entre sumar y restar tres a cada número de dos dígitos manteniendo el orden en mente sin pistas. Los números de dos dígitos se eligieron al azar entre el 10 y el 99, con la salvedad de que se excluyeron los números que terminaban en 8 y 9 cuando había que sumar tres (por ejemplo, $19+3 = 22$, en este caso el participante cambiaría de decena, lo cual no es permitido) y los números que terminaban en uno y dos cuando había que restar tres (por ejemplo, $32 - 3 = 29$, cambiaría de decena y tampoco es permitido).

- **ef_pmtimecost:** Tiempo empleado en las tres tareas (más, menos y alterna). A partir de la batería de cuentas se suma el tiempo para hacer las pruebas de la tarea más y tarea menos y se resta el tiempo de la tarea alterna.

Estas 8 variables que representan las pruebas o *test*, son las variables que se pretenden relacionar con las tres funciones ejecutivas a saber inhibición, alternancia y actualización. El desarrollo de la aplicación se presentará en el Capítulo 5.

Capítulo 3

Análisis Factorial

El análisis factorial es una técnica estadística multivariada en que a partir de la estructura de dependencia entre las variables de interés, en general representada por las correlaciones o covarianzas entre las variables, permite la creación de un conjunto menor de variables (variables latentes o factores) obtenidas a partir de las variables originales. El objetivo principal de estos modelos es estudiar la estructura de correlación entre un grupo de variables, asumiendo que la asociación entre las variables puede ser explicada por una o más variables latentes, que en el caso del análisis factorial se les conoce como factores. Dicho de otra manera, la correlación entre el grupo de variables se explica por la presencia de los factores subyacentes a ellas (Díaz y Morales, 2012).

Mediante esta técnica es posible saber cuánto cada factor está asociado a cada variable y cuánto el conjunto general de factores explica la variabilidad total del conjunto de datos originales.

Hay dos tipos de análisis factorial los cuales son:

1. **Análisis factorial exploratorio:** Este análisis permite simplificar la información a partir de la matriz de correlación, explorar con mayor precisión los constructos o variables latentes y variables observadas. Se aplica con el objetivo de comprobar hasta qué punto estos instrumentos o los ítems que los conforman representan adecuadamente los constructos latentes de interés. En otras palabras la variabilidad de un conjunto de variables es explicada por un número más reducido de factores y supone que cualquier indicador o variable puede estar asociado a cualquier factor. Este tipo de análisis es el más común ya que el investigador lo emplea como técnica exploratoria o descriptiva para determinar el número adecuado

de factores comunes y descubrir cuáles variables de medición son indicadores razonables de los factores.

2. **Análisis factorial confirmatorio:** Permite contrastar un modelo construido con antelación, es decir, el investigador especifica el número de factores y el patrón de relación entre el indicador y las cargas factoriales de antemano.

En el análisis factorial para interpretar bien los factores se utiliza una rotación de ejes, la rotación se hace necesaria cuando en ocasiones es difícil interpretar el significado de los factores a partir de la matriz de cargas factoriales, sobre todo si aparecen varios factores compartiendo variables, es decir, puede haber distintas variables que muestran correlaciones altas con varios factores. Los métodos de rotación ortogonales más utilizados son: varimax, cuartimax y equimax que serán discutidos en la Sección 3.4.

3.1. Análisis Factorial Ortogonal

Sea $\mathbf{X} = (X_1, X_2, \dots, X_p)$ el conjunto de las p variables observadas, se asumen relacionadas con un número dado de variables latentes $\mathbf{F} = F_1, F_2, \dots, F_m$, llamadas factores comunes y además las fuentes de variación $\boldsymbol{\varepsilon} = \varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ llamadas errores, algunas veces factores específicos.

El modelo factorial ortogonal es dado por:

$$\begin{aligned} \mathbf{X}_1 - \boldsymbol{\mu}_1 &= \lambda_{11}\mathbf{F}_1 + \lambda_{12}\mathbf{F}_2 + \dots + \lambda_{1m}\mathbf{F}_m + \boldsymbol{\varepsilon}_1 \\ \mathbf{X}_2 - \boldsymbol{\mu}_2 &= \lambda_{21}\mathbf{F}_1 + \lambda_{22}\mathbf{F}_2 + \dots + \lambda_{2m}\mathbf{F}_m + \boldsymbol{\varepsilon}_2 \\ &\vdots \\ \mathbf{X}_p - \boldsymbol{\mu}_p &= \lambda_{p1}\mathbf{F}_1 + \lambda_{p2}\mathbf{F}_2 + \dots + \lambda_{pm}\mathbf{F}_m + \boldsymbol{\varepsilon}_p, \end{aligned}$$

donde:

- λ_{ij} cargas factoriales.
- $(\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_m)$, con $m < p$ son los factores comunes.
- $(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \dots, \boldsymbol{\varepsilon}_p)$ los factores únicos o específicos.

En forma matricial

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_p \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \vdots \\ \mathbf{F}_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_p \end{bmatrix}$$

$$\mathbf{X}_{(p \times 1)} - \boldsymbol{\mu} = \boldsymbol{\Lambda}_{(p \times m)} \mathbf{F}_{(m \times 1)} + \boldsymbol{\varepsilon}_{(p \times 1)}.$$

3.1.1. Supuestos del modelo

Suponga que:

1. $E[\mathbf{X}] = \boldsymbol{\mu}$ vector de la media poblacional ($p \times 1$).
2. $Cov[\mathbf{X}] = \boldsymbol{\Sigma}$ matriz de dimensión ($p \times p$), definida positiva de covarianzas de la población.
3. $E[\mathbf{F}] = 0$, y $E[\boldsymbol{\varepsilon}] = 0$.
4. $Cov[\mathbf{F}] = E[\mathbf{F}\mathbf{F}'] = \mathbf{I}_{m \times m}$.
5. $E[\boldsymbol{\varepsilon}] = 0$.
6. $Cov[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Psi}_{p \times p}$ con $\boldsymbol{\Psi} = diag(\psi_1, \dots, \psi_p)$, para $i = 1, \dots, p$. Corresponde a las variables específicas o unicidades.
7. $Cov[\boldsymbol{\varepsilon}, \mathbf{F}] = E[\boldsymbol{\varepsilon}\mathbf{F}'] = 0$.

Con ello los vectores \mathbf{F}_j con $j = 1, \dots, m$ y $\boldsymbol{\varepsilon}_i$ con $i = 1, \dots, p$ son no correlacionados (variables aleatorias independientes cuya covarianza es igual a cero), la matriz $\boldsymbol{\Psi}$ implica que los errores $\boldsymbol{\varepsilon}_i$ son heterocedásticos y no correlacionados.

Finalmente los elementos $\mathbf{X}, \mathbf{F}, \boldsymbol{\varepsilon}$ y $\boldsymbol{\Lambda}$ representan el modelo lineal ortogonal:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\varepsilon},$$

donde $\boldsymbol{\mu}$ media de la variable i , \mathbf{F} factores, $\boldsymbol{\Lambda}$ cargas factoriales y $\boldsymbol{\varepsilon}$ especificidad.

Las cargas factoriales indican la importancia que cada constructo tiene en la determinación del valor de cada variable.

Los factores específicos dan cuenta de la parte de cada variable que no es explicada por los factores comunes.

3.1.2. Diagrama de caminos de un modelo factorial ortogonal

La Figura 3.1 presenta un ejemplo de diagrama de caminos donde las variables observadas están representadas por cuadrados, los factores o variables latentes por círculos, sus respectivos errores sin cuadrados ni círculos y las flechas van de una variable independiente a una variable dependiente.

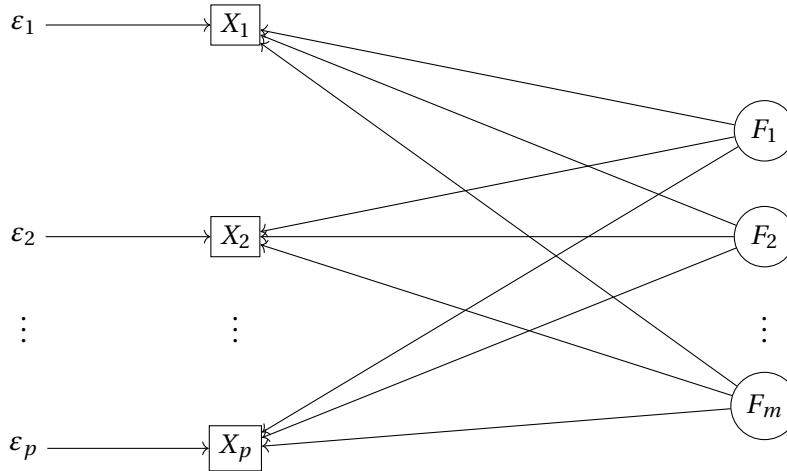


Figura 3.1: Diagrama de camino de un modelo de análisis factorial ortogonal

3.2. Algunos resultados sobre el modelo de análisis factorial

1. Si se dan todos los supuestos del modelo, la matriz de covarianzas de \mathbf{X} , Σ en el modelo de análisis factorial se puede expresar de la forma:

$$\begin{aligned}
 \Sigma &= Cov(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] \\
 &= E[(\Lambda\mathbf{F} + \boldsymbol{\varepsilon})(\Lambda\mathbf{F} + \boldsymbol{\varepsilon})'] \\
 &= E[(\Lambda\mathbf{F} + \boldsymbol{\varepsilon})(\Lambda\mathbf{F}')' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\
 &= E[(\Lambda\mathbf{F} + \boldsymbol{\varepsilon})(\mathbf{F}'\Lambda') + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\
 &= E[(\Lambda\mathbf{F})(\mathbf{F}'\Lambda') + \boldsymbol{\varepsilon}(\mathbf{F}'\Lambda') + \Lambda\mathbf{F}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\
 &= \Lambda E[\mathbf{F}\mathbf{F}']\Lambda' + \Lambda E[\mathbf{F}\boldsymbol{\varepsilon}'] + E[\boldsymbol{\varepsilon}\mathbf{F}']\Lambda' + E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\
 &= \Lambda\Lambda' + \Psi
 \end{aligned}$$

donde Σ es la matriz de correlación poblacional de las variables $(\mathbf{X}_1, \dots, \mathbf{X}_p)$, variables observadas y $\Psi = diag(\psi_i)$ es la diagonal de las especificidades.

Esto implica que la covarianza entre \mathbf{X} y \mathbf{F} tiene la forma:

$$\begin{aligned} Cov[\mathbf{X}, \mathbf{F}] &= E[(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}'] \\ &= E[(\boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\varepsilon})\mathbf{F}'] \\ &= E[\boldsymbol{\Lambda}\mathbf{F}\mathbf{F}' + \boldsymbol{\varepsilon}\mathbf{F}'] \\ &= \boldsymbol{\Lambda}E[\mathbf{F}\mathbf{F}'] + E[\boldsymbol{\varepsilon}\mathbf{F}'] \\ &= \boldsymbol{\Lambda}E[\mathbf{F}\mathbf{F}'] \\ &= \boldsymbol{\Lambda}\mathbf{I} \\ &= \boldsymbol{\Lambda}, \end{aligned}$$

lo que permite hacer una interpretación clara de la matriz de las cargas factoriales: Sus elementos son las covarianzas de variables originales de la población con los factores \mathbf{F} .

2. En un análisis factorial ortogonal, se verifica:

- La varianza de \mathbf{X}_i , y la covarianza de \mathbf{X}_i con \mathbf{X}_k se expresa como:

$$\begin{aligned} Var(\mathbf{X}_i) &= \lambda_{i1}^2 + \dots + \lambda_{im}^2 + \boldsymbol{\psi}_i. \\ Cov(\mathbf{X}_i, \mathbf{X}_k) &= \lambda_{i1}\lambda_{k1} + \dots + \lambda_{im}\lambda_{km}. \end{aligned}$$

- La interpretación de las cargas factoriales viene dada por:

$$Cov(\mathbf{X}_i, \mathbf{F}_j) = \lambda_{ij},$$

donde

$$\begin{aligned} Cov(\mathbf{X}_i, \mathbf{F}_j) &= Cov(\lambda_{i1}\mathbf{F}_1 + \dots + \lambda_{ij}\mathbf{F}_j + \dots + \lambda_{im}\mathbf{F}_m + \boldsymbol{\varepsilon}_i, \mathbf{F}_j) \\ &= Cov(\lambda_{i1}\mathbf{F}_1, \mathbf{F}_j) + \dots + Cov(\lambda_{ij}\mathbf{F}_j, \mathbf{F}_j) + \dots + Cov(\lambda_{im}\mathbf{F}_m, \mathbf{F}_j) \\ &\quad + Cov(\boldsymbol{\varepsilon}_i, \mathbf{F}_j) \\ &= 0 + Cov(\lambda_{ij}\mathbf{F}_j, \mathbf{F}_j) \\ &= \lambda_{ij}Cov(\mathbf{F}_j, \mathbf{F}_j) \\ &= \lambda_{ij}, \end{aligned}$$

por tanto las cargas factoriales son las covarianzas medidas entre las variables observadas y factores comunes.

- Para facilitar la interpretación de un factor común, se calcula la correlación entre el factor y cada variable observada

$$Corr(\mathbf{X}_i, \mathbf{F}_j) = \rho(\mathbf{X}_i, \mathbf{F}_j) = \frac{Cov(\mathbf{X}_i, \mathbf{F}_j)}{\sqrt{Var(\mathbf{X}_i)}\sqrt{Var(\mathbf{F}_j)}} = \frac{\lambda_{ij}}{\sigma_i}.$$

3. En análisis factorial ortogonal, se debe tener en cuenta que la varianza puede ser descompuesta en 2 términos, a saber:

$$\begin{aligned}
 \sigma_i^2 &= \text{Var}(\mathbf{X}_i) \\
 &= \text{Var}(\lambda_{i1}\mathbf{F}_1 + \lambda_{i2}\mathbf{F}_2 + \cdots + \lambda_{im}\mathbf{F}_m + \boldsymbol{\varepsilon}_i) \\
 &= \text{Var}(\lambda_{i1}\mathbf{F}_1) + \text{Var}(\lambda_{i2}\mathbf{F}_2) + \cdots + \text{Var}(\lambda_{im}\mathbf{F}_m) + \text{Var}(\boldsymbol{\varepsilon}_i) \\
 &= \lambda_{i1}^2 \text{Var}(\mathbf{F}_1) + \lambda_{i2}^2 \text{Var}(\mathbf{F}_2) + \cdots + \lambda_{im}^2 \text{Var}(\mathbf{F}_m) + \text{Var}(\boldsymbol{\varepsilon}_i) \\
 &= \underbrace{\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2}_{\text{Comunalidades}} + \underbrace{\boldsymbol{\psi}_i}_{\text{Especificidad}} .
 \end{aligned}$$

También se puede ver las comunalidades como $c_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2$, es la parte de la varianza de \mathbf{X}_i explicada por los factores y la especificidad es la parte de la varianza de \mathbf{X}_i no explicada por los factores.

En resumen, las varianzas se expresan como:

$$\begin{aligned}
 \text{Var}(\mathbf{X}_1) &= \lambda_{11}^2 + \lambda_{12}^2 + \cdots + \lambda_{1m}^2 + \boldsymbol{\psi}_1 \\
 \text{Var}(\mathbf{X}_2) &= \lambda_{21}^2 + \lambda_{22}^2 + \cdots + \lambda_{2m}^2 + \boldsymbol{\psi}_2 \\
 &\vdots \\
 \text{Var}(\mathbf{X}_i) &= \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \boldsymbol{\psi}_i \\
 &\vdots \\
 \text{Var}(\mathbf{X}_p) &= \lambda_{p1}^2 + \lambda_{p2}^2 + \cdots + \lambda_{pm}^2 + \boldsymbol{\psi}_p .
 \end{aligned}$$

También se puede escribir como:

$$\sigma_i^2 = c_i^2 + \boldsymbol{\psi}_i \quad \text{con } i = 1, 2, \dots, p.$$

Para facilitar la interpretación de las comunalidades, se sugiere la utilización del término entre $[0, 1]$, así:

$$\bar{c}_i^2 = \frac{c_i^2}{\sigma_i^2},$$

que indica la proporción de la variabilidad de \mathbf{X}_i que es explicada por los factores comunes. Cuanto más próxima de 1 (100%) mejor es el ajuste del modelo.

Uno de los principales objetivos del análisis factorial es la representación de la estructura de dependencia de los datos, una vez estén estandarizados, es decir :

- Al realizar el análisis sobre un conjunto de datos con varianzas de magnitudes diferentes se pueden tener dificultades en la explicación de esa dependencia.
- En casos en que exista una diferencia grande entre las varianzas de las variables originales, se sugiere que el análisis sea realizado sobre las variables estandarizadas.
- En este caso, las covarianzas corresponden a las correlaciones entre las variables originales.

De este modo, el análisis se puede desarrollar a partir de la matriz de correlación \mathbf{R} definida por

$$\mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}.$$

3.3. Métodos de estimación

Una vez que se especifican las variables y se prepara la matriz de correlación, ya se puede aplicar el análisis factorial. Para ello es necesario definir el método de estimación de los factores, existen varios métodos para estimar las ponderaciones factoriales λ_{ij} y la matriz $\mathbf{\Psi}$ de especificidades (Díaz y Morales, 2012).

En esta sección se explican los métodos más comunes, entre ellos está la estimación por "Componentes Principales", no necesita suponer normalidad de los datos originales representados por la matriz \mathbf{X} , por otra parte el método de estimación por "Máxima Verosimilitud" necesita suponer normalidad multivariada (modelo factorial de Lawley-Maxwell).

También existe el método del "Factor Principal" que es conocido en el análisis factorial, (Johnson y Wichern, 1998).

3.3.1. El método de Componentes Principales

Dada la matriz $\mathbf{\Sigma}$ la matriz de covarianza de $(\mathbf{X}_1, \dots, \mathbf{X}_p)$, sean $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ los valores propios y los vectores propios e_1, e_2, \dots, e_p obtenidos por la descomposición espectral de la matriz de covarianzas $\mathbf{\Sigma}$ (Johnson y Wichern, 1998).

Nota: En este trabajo se denota el valor propio por λ_p , con un sólo subíndice, para no confundir con λ_{ij} , cargas factoriales del modelo definido en la Sección 3.1 .

Tomando la matriz de covarianza Σ , se puede descomponer de la siguiente forma:

$$\Sigma_{p \times p} = \lambda_1 e_1 e_1' + \dots + \lambda_p e_p e_p' = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \dots & \sqrt{\lambda_p} e_p \end{bmatrix}_{p \times p} \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix}_{p \times p} = \Lambda \Lambda'.$$

Por tal motivo tomando tantos factores como variables se obtiene que la matriz de especificidades es cero, a saber

$$\Sigma = \Lambda \Lambda' + 0 = \Lambda \Lambda'.$$

Del análisis de componentes principales, por el Teorema de Descomposición Espectral dice que existe una matriz ortogonal P y otra matriz diagonal Λ ver en Díaz y Morales (2012), tales que

$$\begin{aligned} \Sigma &= P \Lambda P' \\ &= P \Lambda^{1/2} (P \Lambda^{1/2})' \\ &= P \Lambda^{1/2} \Lambda^{1/2} P' \\ &= \Lambda \Lambda' + 0, \end{aligned}$$

con $\Lambda = P \Lambda^{1/2}$, $P P' = I$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ y $P = [e_1 \dots e_p]$.

Con ello se busca factorizar la matriz Σ de tal manera que existan tantos factores como variables ($m = p$), y se obtiene $\psi_i = 0$, para todo $i = 1, \dots, p$ y las cargas factoriales valen $\sqrt{\lambda_k} e_k$, con λ_k valores propios, situados en la columna k -ésima de la matriz de cargas Λ .

En resumen, salvo el factor $\sqrt{\lambda_k}$, las cargas factoriales del k -ésimo factor son los coeficientes de la k -ésima componente principal, dado que, en efecto, los coeficientes de la k -ésima componente principal es el vector e_k . Cuando $m = p$, se obtiene una representación exacta $\Sigma = \Lambda \Lambda'$ de tipo factorial, con tantos factores como variables. Pero esto no es objeto del análisis factorial, se pretende fundamentalmente explicar la correlación entre variables con un número significativamente menor de factores que de variables.

Para ello se desprecian las $p - m$ contribuciones debidas a los $p - m$ "últimos" valores propios $\lambda_{m+1} \geq \lambda_{m+2} \geq \dots \geq \lambda_p$, es decir :

$$\Sigma_{p \times p} = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \dots & \sqrt{\lambda_m} e_m \end{bmatrix}_{p \times m} \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \vdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix}_{m \times p} = \Lambda \Lambda' + \Psi.$$

De forma más específica,

$$\Sigma_{p \times p} = \Lambda \Lambda' + \Psi = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \cdots & \sqrt{\lambda_m} e_m \end{bmatrix}_{p \times m} \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \vdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix}_{m \times p} + \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \Psi_p \end{bmatrix}_{p \times p}$$

donde $\psi_i = \sigma_{ii} - \sum_{j=1}^m \lambda_{ij}^2$, $i = 1, \dots, p$, teniendo en cuenta que las comunalidades son los elementos de la diagonal principal de $\Lambda \Lambda'$, por tanto están dadas por $(\sqrt{\lambda_1} e_1, \dots, \sqrt{\lambda_m} e_m)$.

3.3.2. El método de Máxima Verosimilitud

Del modelo de factores ortogonales se tiene que

$$\mathbf{X} = \boldsymbol{\mu} + \Lambda + \boldsymbol{\varepsilon}, \quad \Sigma = \Lambda \Lambda' + \Psi.$$

Sea $\mathbf{X}_1, \dots, \mathbf{X}_p$ una muestra aleatoria tal que $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, es decir, \mathbf{X} (variables originales) se supone distribuida Normal Multivariada, donde Σ es una matriz simétrica definida positiva. Entonces, los estimadores de máxima verosimilitud están dados por (Johnson y Wichern 1998):

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i,$$

$$S_n = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{X})(\mathbf{X}_i - \bar{X})'.$$

Además por los supuestos del modelo se tiene que $E[\boldsymbol{\varepsilon}] = 0$, $E[\mathbf{F}] = 0$, $Cov[\boldsymbol{\varepsilon}] = \Psi$ y $Cov[\mathbf{F}] = \mathbf{I}$, entonces se puede suponer que \mathbf{F} y $\boldsymbol{\varepsilon}$ tienen distribuciones Normales multivariadas independientes, tal que

$$\mathbf{F} \sim N(0, \mathbf{I}),$$

$$\boldsymbol{\varepsilon} \sim N(0, \Psi),$$

y se pueden obtener estimaciones de máxima verosimilitud de las varianzas específicas. Cuando \mathbf{F}_j y $\boldsymbol{\varepsilon}_j$ son conjuntamente normales, las observaciones

$\mathbf{X}_j - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{F}_j + \boldsymbol{\varepsilon}_j$ son normales, entonces la función de densidad de probabilidad está dada por:

$$\begin{aligned} L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\mu}) &= \prod_{j=1}^n \left\{ (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{X}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu}) \right] \right\} \\ &= \left\{ (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}|^{-\frac{n}{2}} \exp \left[-\frac{1}{2} \sum_{j=1}^n (\mathbf{X}_j - \boldsymbol{\mu})' (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} (\mathbf{X}_j - \boldsymbol{\mu}) \right] \right\} \end{aligned}$$

donde al maximizar $L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\mu})$, y volverse cero esta función, es claro que permite concluir que $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$. Por tanto el problema queda reducido a

$$(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Psi}}) = \underset{\boldsymbol{\Lambda}, \boldsymbol{\Psi}}{\text{máx}} L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \hat{\boldsymbol{\mu}}) = \underset{\boldsymbol{\Lambda}, \boldsymbol{\Psi}}{\text{máx}} L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \bar{\mathbf{X}}).$$

Sea $\mathcal{L} = \ln L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \bar{\mathbf{X}})$ el logaritmo de la función de verosimilitud. Entonces

$$\begin{aligned} \mathcal{L} &= -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln |\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}| - \frac{1}{2} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})' (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \\ &= -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln |\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}| - \frac{1}{2} \text{tr}(n(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} S_n) \\ &= -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \{ \ln |\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}| - \text{tr}[(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} S_n] \}, \end{aligned}$$

la anterior ecuación depende de $\boldsymbol{\Lambda}$ y $\boldsymbol{\Psi}$ a través de $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. Este modelo aún no está bien definido, debido a la multiplicidad de opciones para $\boldsymbol{\Lambda}$ que posibilitan las transformaciones ortogonales. Es deseable hacer que $L(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\mu})$ esté bien definido imponiendo la condición de unicidad computacionalmente conveniente:

$$\boldsymbol{\Lambda}'\boldsymbol{\Psi}\boldsymbol{\Lambda} = \boldsymbol{\Delta}.$$

Las estimaciones de máxima verosimilitud $\boldsymbol{\Lambda}$ y $\boldsymbol{\Psi}$ y denotados $\hat{\boldsymbol{\Lambda}}$ y $\hat{\boldsymbol{\Psi}}$, en el modelo de análisis factorial ortogonal, con la condición de unicidad, deben obtenerse mediante la maximización numérica. Afortunadamente, ahora existen programas informáticos eficientes que permiten obtener estas estimaciones con bastante facilidad.

Nota: Una demostración completa de este método se puede ver en Hortal (2012).

Ejemplo: En este ejemplo se ilustra la teoría realacionada con el análisis factorial, el ejemplo es tomado de Barroso y Artes (2003).

Suponga que el conjunto de datos conformado por 4 variables X_1, X_2, X_3, X_4 presenta la siguiente matriz de covarianza Σ :

Tabla 3.1: Matriz Σ

	X_1	X_2	X_3	X_4
X_1	4.00	0.80	1.68	2.16
X_2	0.80	2.00	0.44	0.96
X_3	1.68	0.44	2.00	1.32
X_4	2.16	0.96	1.32	2.00

La matriz de las cargas factoriales Λ es dada por:

Tabla 3.2: Cargas factoriales

	F_1	F_2
	1.60	0.40
	0.20	1.20
	1.00	0.20
	1.20	0.60

La matriz de especificidades Ψ es:

$$\Psi = \begin{bmatrix} 1,28 & 0,00 & 0,00 & 0,00 \\ 0,00 & 0,52 & 0,00 & 0,00 \\ 0,00 & 0,00 & 0,96 & 0,00 \\ 0,00 & 0,00 & 0,00 & 0,20 \end{bmatrix}.$$

El modelo factorial ortogonal con las cargas factoriales y en función de dos Factores es dado por.:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} = \begin{bmatrix} 1,60 & 0,40 \\ 0,20 & 1,20 \\ 1,00 & 0,20 \\ 1,20 & 0,60 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}.$$

Más específicamente el modelo puede ser escrito de la siguiente manera:

$$\begin{aligned} \mathbf{X}_1 - \boldsymbol{\mu}_1 &= 1,60\mathbf{F}_1 + 0,40\mathbf{F}_2 + \boldsymbol{\varepsilon}_1 \\ \mathbf{X}_2 - \boldsymbol{\mu}_2 &= 0,20\mathbf{F}_1 + 1,20\mathbf{F}_2 + \boldsymbol{\varepsilon}_2 \\ \mathbf{X}_3 - \boldsymbol{\mu}_3 &= 1,00\mathbf{F}_1 + 0,20\mathbf{F}_2 + \boldsymbol{\varepsilon}_3 \\ \mathbf{X}_4 - \boldsymbol{\mu}_4 &= 1,20\mathbf{F}_1 + 0,60\mathbf{F}_2 + \boldsymbol{\varepsilon}_4. \end{aligned}$$

En la Tabla 3.3 se presentan las variables, las cargas factoriales al cuadrado, la communalidad, el porcentaje de explicación de los dos factores y la especificidad para cada una de las 4 variables.

Tabla 3.3: Resumen de los cálculos presentados en el ejemplo asociado a análisis factorial

\mathbf{X}_i	λ_{ij}^2	c_i^2	Porcentaje de explicación	ψ_i
\mathbf{X}_1	$\lambda_{11}^2 + \lambda_{12}^2 = 1,60^2 + 0,40^2$	2.72	$\bar{c}_1^2 = \frac{c_1^2}{\sigma_1^2} = \frac{2,72}{4} = 0,68, 68\%$	1.28
\mathbf{X}_2	$\lambda_{21}^2 + \lambda_{22}^2 = 0,20^2 + 1,20^2$	1.48	$\bar{c}_2^2 = \frac{c_2^2}{\sigma_2^2} = \frac{1,48}{2} = 0,74, 74\%$	0.52
\mathbf{X}_3	$\lambda_{31}^2 + \lambda_{32}^2 = 1,00^2 + 0,20^2$	1.04	$\bar{c}_3^2 = \frac{c_3^2}{\sigma_3^2} = \frac{1,04}{2} = 0,52, 52\%$	0.96
\mathbf{X}_4	$\lambda_{41}^2 + \lambda_{42}^2 = 1,20^2 + 0,60^2$	1.80	$\bar{c}_4^2 = \frac{c_4^2}{\sigma_4^2} = \frac{1,80}{2} = 0,90, 90\%$	0.20

Se puede observar que la variable mejor explicada por los factores es \mathbf{X}_4 ya que explica el 90% de variabilidad y la variable que menos explica los factores es \mathbf{X}_3 ya que explica el 52% de la variabilidad, ya que cuanto más próxima a 1 (100%) mejor es el ajuste del modelo.

Definiendo la varianza total como:

$$\sigma_T^2 = \sum_{i=1}^p \sigma_i^2 = 4 + 2 + 2 + 2 = 10$$

Ahora la variabilidad por parte de cada factor está dada por:

$$\sigma_{F_i}^2 = \sum_{i=1}^p \lambda_{ip}^2 = \lambda_{11}^2 + \lambda_{21}^2 + \dots + \lambda_{1p}^2, \quad (3.1)$$

y

$$\frac{\sum_{i=1}^p \lambda_{ip}^2}{\sigma_T^2}, \quad (3.2)$$

es la proporción de la variabilidad total de datos que es explicada por el conjunto de los factores comunes.

Aplicando la ecuación 3.1 y 3.2, para hallar la explicación de la variabilidad total por parte de cada factor, se tiene que:

- Para el factor 1

$$\sigma_{F_1}^2 = \lambda_{11}^2 + \lambda_{21}^2 + \lambda_{31}^2 + \lambda_{41}^2 = 1,60^2 + 0,20^2 + 1,00^2 + 1,20^2,$$

entonces

$$\frac{\sum_{i=1}^p \lambda_{i1}^2}{\sigma_T^2} = \frac{\sigma_{F_1}^2}{\sigma_T^2} = \frac{5,04}{10} = 0,504.$$

Luego el factor 1 explica el 50% de la variabilidad total de las variables.

- Para el factor 2

$$\sigma_{F_2}^2 = \lambda_{12}^2 + \lambda_{22}^2 + \lambda_{32}^2 + \lambda_{42}^2 = 0,40^2 + 1,20^2 + 0,20^2 + 0,60^2 = 2,$$

entonces

$$\frac{\sum_{i=1}^p \lambda_{i2}^2}{\sigma_T^2} = \frac{\sigma_{F_2}^2}{\sigma_T^2} = \frac{2}{10} = 0,2.$$

Por lo tanto el factor 2 explica el 20% de la variabilidad total de las variables.

3.4. Rotación de factores

En esta sección se describe principalmente el método de rotación más usado el cual es varimax, en relación a las rotaciones cuartimax, equimax y rotaciones oblicuas se puede encontrar una formalización matemática en Díaz y Morales (2012).

1. **Varimax:** Rotación ortogonal, también llamada rotación *Kaiser- Varimax*. La rotación varimax es una técnica estadística que se utiliza en análisis factorial para intentar aclarar la relación entre los factores. Generalmente, el proceso consiste en ajustar las coordenadas de los datos que resultan de un análisis de componentes principales. El ajuste, o la rotación, tiene por objeto maximizar la varianza compartida entre los elementos. Al maximizar la varianza compartida, los resultados representan de forma simplificada de como se correlacionan los datos con cada componente principal. Maximizar la varianza generalmente significa aumentar la correlación al cuadrado de los ítems relacionados con un factor, mientras se disminuye la correlación en cualquier otro factor. En otras palabras, la rotación varimax simplifica las cargas de los ítems eliminando el término medio e identificando más específicamente el factor sobre el que se cargan los datos.

Por otro lado, el objetivo de la rotación varimax es determinar una matriz **T**, tal que cualquier factor tenga algunas variables con ponderación alta y otras con ponderación baja. Esto se logra mediante la maximización de la varianza asociada con los cuadrados de las ponderaciones ($\hat{\lambda}_{ij}$) sobre todas las variables, con la restricción de que la comunalidad para cada variable no se altere (Díaz y Morales, 2012).

Por tanto, para un factor dado j

$$V_j = \frac{\sum_{i=1}^p (\hat{\lambda}_{ij}^2 - \overline{\hat{\lambda}_j^2})^2}{p} = \frac{\sum_{i=1}^p \hat{\lambda}_{ij}^4 - \left(\sum_{i=1}^p \hat{\lambda}_{ij}^2 \right)^2}{p^2},$$

donde V_j es la varianza de las comunalidades de las variables dentro del

factor j y $\overline{\hat{\lambda}_j^2} = \frac{\sum_{i=1}^p \hat{\lambda}_{ij}^2}{p}$, es el promedio de los cuadrados para las ponderaciones del factor j . La varianza para todos los factores está dada por

$$\begin{aligned}
 V &= \sum_{j=1}^k V_j = \sum_{j=1}^k \left(\frac{\sum_{i=1}^p \hat{\lambda}_{ij}^4 - \left(\sum_{i=1}^p \hat{\lambda}_{ij}^2 \right)^2}{p^2} \right) \\
 &= \frac{\sum_{j=1}^k \sum_{i=1}^p \hat{\lambda}_{ij}^4}{p} - \frac{\sum_{j=1}^k \left(\sum_{i=1}^p \hat{\lambda}_{ij}^2 \right)^2}{p^2}.
 \end{aligned}$$

Como el número de variables permanece constante, la maximización se hace sobre la siguiente ecuación:

$$V = \frac{1}{p} \left(\sum_{j=1}^k \sum_{i=1}^p \hat{\lambda}_{ij}^4 - \frac{\sum_{j=1}^k \left(\sum_{i=1}^p \hat{\lambda}_{ij}^2 \right)^2}{p} \right).$$

La matriz ortogonal \mathbf{T} , se obtiene de tal forma que la ecuación anterior sea máxima, con la restricción de que la comunalidad para cada variable permanezca constante.

2. **Cuartimax:** La rotación cuartimax maximiza la varianza de las cargas factoriales al cuadrado en cada variable, es decir, simplifica las filas de la matriz de carga. En otras palabras, la rotación cuartimax minimiza el número de factores necesarios para explicar cada variable. Mediante la rotación cuartimax, a menudo se incluye un gran número de variables en el primer factor rotado, por lo que este método no suele ser útil en muchas investigaciones.
3. **Equimax:** Método de rotación que es combinación del método varimax, que simplifica los factores, y el método cuartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.

Una forma más formal de definir las rotaciones cuartimax, equimax y rotaciones oblicuas se pueden encontrar en Johnson y Wichern (1998).

3.5. Análisis de la matriz de correlación

La finalidad de analizar la matriz de correlación $\mathbf{R} = (r_{ij})$ donde r_{ij} es la correlación muestral observada entre las variables \mathbf{X}_i y \mathbf{X}_j , es comprobar si sus características son las adecuadas para realizar un análisis factorial.

Uno de los requisitos que deben cumplirse es que las variables se encuentren altamente intercorrelacionadas. También se espera que las variables que tengan correlación muy alta entre sí la tengan con el mismo factor o factores.

En consecuencia, si las correlaciones entre todas las variables son bajas, tal vez no sea apropiado el análisis factorial.

3.5.1. Test de Esfericidad de Bartlett

Se utiliza para probar la hipótesis nula que afirma que las variables no están correlacionadas en la población. Comprueba si la matriz de correlaciones es una matriz de identidad \mathbf{I} , es decir ausencia de correlación significativa entre las variables. Si se confirma la hipótesis nula $H_0 : \mathbf{R} = \mathbf{I}$, es decir, que el determinante de la matriz de correlaciones es 1, $H_0 : |\mathbf{R}| = 1$, se comprueba que las variables no están correlacionadas.

El *test* de esfericidad de Barlett se obtiene mediante una transformación del determinante de la matriz de correlación. El estadístico del *test* viene dado por:

$$\chi^2 = - \left(n - 1 - \frac{1}{6}(2p + 5) \right) \ln |\mathbf{R}|,$$

donde n es el número de individuos de la muestra, p número de variables y \mathbf{R} matriz de correlaciones.

Bajo la hipótesis nula, el estadístico se distribuye asintóticamente según una $\chi^2_{p(p-1)/2}$. Si se acepta la hipótesis nula ($p > 0,05$) significa que las variables no están intercorrelacionadas y por tanto no tiene mucho sentido llevar a cabo un análisis factorial.

Por el contrario, si con el *test* de Barlett se obtienen valores altos de χ^2 , o un determinante bajo, hay variables con correlaciones altas (un determinante próximo a cero indica que una o más variables podrían ser expresadas como combinación lineal de otras variables).

En definitiva, si el estadístico del *test* toma valores grandes (o un determinante próximo a cero) se rechaza la hipótesis nula con cierto grado de significancia.

En caso de aceptarse la hipótesis nula, las variables no están correlacionadas y debería reconsiderarse la aplicación de un análisis factorial.

3.5.2. Matriz Anti-imagen

Una de las premisas de un análisis factorial es que exista una estructura de dependencia clara entre las variables en estudio. En el modelo propuesto, esa estructura se expresa a través de una matriz de covarianza o de correlación. La existencia de tal estructura implica que una variable puede, dentro de ciertos límites, ser prevista por las demás. Para verificar ese hecho se pueden calcular los coeficientes de la correlación parcial entre los pares de variables, eliminando el efecto de las demás variables, se espera que los valores obtenidos sean bajos.

La matriz de covarianzas anti-imagen contiene los negativos de las covarianzas parciales y la matriz de correlaciones anti-imagen contiene los coeficientes de correlación parcial cambiados de signo (la correlación entre dos variables se parcializa teniendo en cuenta el resto de variables incluidas en el análisis).

3.5.3. Coeficiente de correlación parcial

Son correlaciones simples y de primer orden, esto es, relacionando una variable con otra y eliminando el efecto de una tercera. Si tenemos \mathbf{X}_1 , \mathbf{X}_2 y \mathbf{X}_3 la expresión que relaciona \mathbf{X}_1 con \mathbf{X}_2 eliminando \mathbf{X}_3 de la manera siguiente:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}.$$

La correlación parcial mide la variación conjunta que se da entre una variable independiente y una variable dependiente, controlando los efectos que sobre esa variación pudiera ejercer otra independiente. La correlación parcial se expresa en términos de coeficiente de correlación de Pearson, y en la fórmula se separa con un punto la variable controlada de las variables dependiente e independiente (Johnson y Wichern, 1998):

$$r_{ij,k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}.$$

3.5.4. Medidas de adecuación de la muestra

- El coeficiente de correlación parcial es un indicador del grado de relación entre dos variables, eliminando la influencia del resto de variables.
- Si las variables comparten factores comunes, el coeficiente de correlación parcial entre pares de variables es bajo, puesto que se eliminan los efectos lineales de las otras variables.
- Las correlaciones parciales son estimaciones de las correlaciones entre los factores únicos, debiendo ser próxima a cero cuando el Análisis Factorial es adecuado, dado que se supone que los factores únicos no son correlacionados entre sí.
- En definitiva, si existe un número elevado de coeficientes de correlación parcial distintos de cero, se interpreta que las hipótesis del modelo no son compatibles con los datos.
- Una manera de cuantificar este hecho, consiste en examinar la medida de adecuación de la muestra propuesta por *Kaiser-Meyer-Olkin* (KMO). Este indicador mide la adecuación de un conjunto de datos, se obtiene por

$$KMO = \frac{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2}{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2 + \sum_{i=1}^p \sum_{j=1}^p a_{ij}^2},$$

donde r_{ij}^2 es el coeficiente de correlación simple entre las variables \mathbf{X}_i y \mathbf{X}_j , a_{ij} es el coeficiente de correlación parcial entre las mismas variables \mathbf{X}_i y \mathbf{X}_j .

El índice KMO se utiliza para comparar las magnitudes de los coeficientes de correlación parcial, de forma que cuánto más pequeño sea su valor, mayor será el valor de los coeficientes de correlación parciales a_{ij} y, en consecuencia, menos apropiado es realizar un análisis factorial (Díaz y Morales, 2012).

Kaiser-Meyer-Olkin para realizar un análisis factorial propone las siguientes medidas

Tabla 3.4: Medida de KMO

$KMO \geq 0,75$	Bien
$KMO \geq 0,5$	Aceptable
$KMO < 0,5$	Inaceptable

Otra medida de adecuación es MSA, (*Measure of Sampling adequacy*) esta medida es bastante similar al KMO. La medida MSA debe ser calculada separadamente para cada variable. El objetivo es verificar si una variable dada puede ser explicada por las demás (lo que es esperado en un análisis factorial). Valores bajos de MSA_i son indicios de que la respectiva variable puede ser retirada del modelo sin mayores prejuicios.

La medida es similar al KMO, presentando una lógica semejante. Este índice puede ser interpretado utilizando los mismos límites que la medida KMO, y se denota por:

$$MSA_i = \frac{\sum_{j=1}^p r_{ij}^2}{\sum_{j=1}^p r_{ij}^2 + \sum_{j=1}^p a_{ij}^2}.$$

Capítulo 4

Modelo de Ecuaciones Estructurales

Según Hair y cols. (2009) el modelado de ecuaciones estructurales es una agrupación de modelos estadísticos que buscan explicar las relaciones entre múltiples variables. Para ello, se utiliza una serie de ecuaciones que explican las interrelaciones entre las variables, como una serie de ecuaciones de regresión múltiple. Estas ecuaciones describirán los constructos, que a su vez son factores latentes compuestos por múltiples variables.

Para el autor, todos los modelos de ecuaciones estructurales son distintos debido a tres factores:

- (i) La estimación de relaciones de dependencia múltiples;
- (ii) La capacidad de representar conceptos no observados en estas relaciones y la corrección del error de medición en el proceso;
- (iii) La definición de un modelo estructural para la representación de un conjunto de relaciones.

Para representar lo descrito de forma teórica, un modelo general en la metodología de ecuaciones estructurales consta de dos modelos: El modelo de medida y el modelo estructural. El modelo de medida representa la relación entre las variables observadas y las variables latentes, mientras que el modelo estructural representa solamente la relación entre las variables latentes.

Es importante señalar que la definición de modelos de medición y estructurales está completamente elaborada y controlada por investigadores. Sin embar-

go, con un soporte teórico adecuado, SEM se convierte en una poderosa herramienta analítica para muchos profesionales que estudian relaciones complejas en muchas áreas (Hair y cols., 2009).

En este capítulo se presentan los principales fundamentos teóricos de los modelos de ecuaciones estructurales. En primer lugar se describen los tipos de variables para seguir con la formulación matemática del modelo, describiendo específicamente el modelo estructural y el modelo de medición, estableciendo los respectivos supuestos. También se presenta la notación general, representada a través del diagrama de caminos. Finalmente se presentan las ideas de efectos del modelo, la identificabilidad, los métodos de estimación y ajuste del modelo.

4.1. Tipos de variable

El modelo de ecuaciones estructurales tiene los siguientes tipos de variables (Hair y cols., 2009; Bollen, 1989):

- **Variables observadas:** Son las variables que pueden ser medidas de manera directa, también llamadas variables indicadoras.
- **Variables latentes:** También llamadas factores o variables no observadas; son aquellas variables que se infieren a partir de las variables observadas y no se pueden medir de manera directa. Las variables latentes o factores se clasifican en:
 1. **Variables latentes endógenas:** Son las variables latentes dependientes ya que reciben el efecto de otras variables. La variable dependiente de un modelo de regresión es endógena. Toda variable endógena debe ir acompañada de un error.
 2. **Variables latentes exógenas:** Son las variables latentes independientes ya que tienen efecto sobre otras variables pero no reciben efecto de ninguna de ellas. Las variables independientes de un modelo de regresión son exógenas.
 3. **Variable error:** Representa tanto los errores asociados a la medición de una variable como el conjunto de variables que no han sido contempladas en el modelo y que pueden afectar la medición de una variable observada. Se considera que son variables de tipo latente por no ser observables directamente. El error asociado a la variable dependiente representa el error de predicción.

4.2. Modelo Matemático

Un modelo de ecuaciones estructurales completo consta de dos partes fundamentales: el modelo de medida y el modelo de relaciones estructurales, recordando que el modelo de medida representa la relación entre las variables observadas y las variables latentes, mientras que el modelo estructural representa solamente la relación entre las variables latentes. A continuación se escriben estas definiciones en términos matemáticos.

4.2.1. Modelo Estructural

Representa las relaciones causales entre las variables latentes, se define por:

$$\boldsymbol{\eta}_{(m \times 1)} = \mathbf{B}_{(m \times m)} \boldsymbol{\eta}_{(m \times 1)} + \boldsymbol{\Gamma}_{(m \times n)} \boldsymbol{\xi}_{(n \times 1)} + \boldsymbol{\zeta}_{(m \times 1)}, \quad (4.1)$$

donde

1. $\boldsymbol{\eta}$ es el vector de las variables latentes endógenas.
2. $\boldsymbol{\xi}$ es el vector de las variables latentes exógenas.
3. $\mathbf{B}_{(m \times m)}$ es la matriz de coeficientes que representa los efectos directos de las variables $\boldsymbol{\eta}$ con otras variables $\boldsymbol{\eta}$.
4. $\boldsymbol{\Gamma}_{(m \times n)}$ es la matriz de coeficientes que representan los efectos directos de la variables $\boldsymbol{\xi}$ con las variables $\boldsymbol{\eta}$.
5. $\boldsymbol{\zeta}$ es el vector aleatorio de error de las ecuaciones estructurales.

4.2.2. Modelo de Medida.

El modelo de medida representa las relaciones de las variables latentes (factores) con las variables observadas. Específicamente, el modelo de medida contiene la manera en que cada constructo latente está medido mediante sus indicadores observables, los errores que afectan a las mediciones y las relaciones que se espera encontrar entre los constructos cuando éstos están relacionados entre sí.

En un modelo completo hay dos modelos de medida, uno para las variables predictoras y otro para las variables dependientes, a saber:

$$\mathbf{y}_{(p \times 1)} = \boldsymbol{\Lambda}_{\mathbf{y}} \boldsymbol{\eta}_{(m \times 1)} + \boldsymbol{\epsilon}_{(p \times 1)} \quad (4.2)$$

$$\mathbf{x}_{(q \times 1)} = \mathbf{\Lambda}_x(q \times n)\boldsymbol{\xi}_{(n \times 1)} + \boldsymbol{\delta}_{(q \times 1)} \quad (4.3)$$

donde

1. $\boldsymbol{\eta}$ es el vector de las variables latentes endógenas.
2. \mathbf{y} e \mathbf{x} son los vectores de las variables observadas.
3. $\mathbf{\Lambda}_y$ es la matriz de coeficientes que representa el impacto de las variables $\boldsymbol{\eta}$ con las variables \mathbf{y} .
4. $\mathbf{\Lambda}_x$ es la matriz de coeficientes que representa el impacto de las variables $\boldsymbol{\xi}$ con las variables \mathbf{x} .
5. $\boldsymbol{\xi}$ es el vector de las variables latentes exógenas.
6. $\boldsymbol{\epsilon}$ y $\boldsymbol{\delta}$ son los vectores de error de medición.

4.3. Supuestos del modelo

1. $\boldsymbol{\zeta}$ y $\boldsymbol{\xi}$ son no-correlacionados.
2. $(\mathbf{I} - \mathbf{B})$ matriz no singular.
3. $E[\boldsymbol{\zeta}] = 0$.
4. $E[\boldsymbol{\eta}] = 0$.
5. $E[\boldsymbol{\xi}] = 0$.
6. $\boldsymbol{\epsilon}$ es no-correlacionados con $\boldsymbol{\eta}$.
7. $\boldsymbol{\delta}$ es no-correlacionado con $\boldsymbol{\xi}$.
8. $\boldsymbol{\epsilon}$ y $\boldsymbol{\delta}$ no correlacionados, pero pueden tener sus respectivas variables correlacionadas entre sí.
9. $\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}$ y $\boldsymbol{\delta}$ mutuamente no-correlacionados.

Las matrices de covarianzas de las variables latentes exógenas son dadas por:

1. $\boldsymbol{\Phi}_{(n \times n)} = E[\boldsymbol{\xi}\boldsymbol{\xi}']$ - Matriz de covarianzas de $\boldsymbol{\xi}$.
2. $\boldsymbol{\Psi}_{(m \times m)} = E[\boldsymbol{\zeta}\boldsymbol{\zeta}']$ - Matriz de covarianzas de $\boldsymbol{\zeta}$.

Las matrices de covarianzas de los errores de medición son:

1. $\Theta_{\epsilon(p \times p)} = E[\epsilon\epsilon']$ - Matriz de covarianzas de ϵ .
2. $\Theta_{\delta(q \times q)} = E[\delta\delta']$ - Matriz de covarianzas de δ .

4.4. Diagrama de caminos

Para representar un modelo causal y las relaciones que se desea incluir, se acostumbra utilizar diagramas similares a los diagramas de flujo. Estos diagramas se denominan diagramas causales, gráfico de rutas o diagrama de caminos. En el decorrer de este trabajo se le llamará diagrama de caminos.

El diagrama de caminos de un modelo es su representación gráfica y es de gran ayuda a la hora de especificar el modelo y los parámetros contenidos en él. De hecho, los programas actuales permiten realizar la definición del modelo en su totalidad al representarlo en el interfaz gráfico. A partir del diagrama estructural, algunas veces, el propio programa deriva las ecuaciones del modelo e informa de las restricciones necesarias para que esté completamente identificado.

El diagrama de caminos sigue unas convenciones particulares que es necesario conocer para poder derivar las ecuaciones correspondientes:

- Las variables observadas son representadas por cuadrados o rectángulos.
- Las variables latentes o factores son representadas por círculos o elipses.
- Los errores (sean de medición o de predicción) se representan sin rectángulos ni círculos (aunque algunos programas las dibujan como variables latentes).
- Las relaciones bidireccionales (correlaciones y covarianzas) se representan como vectores curvos con una flecha en cada extremo.
- Una flecha en un sólo sentido entre dos variables indica el efecto directo de una variable sobre otra.
- Una flecha en doble sentido indica una influencia recíproca entre las variables.

- Cualquier efecto estructural se representa como una flecha recta, cuyo origen es la variable predictora y cuyo final, donde se encuentra la punta de la flecha, es la variable dependiente.
- Los parámetros del modelo se representan sobre la flecha correspondiente.
- Cualquier variable que reciba efecto de otras variables del modelo deberá incluir también un término de error.
- Aunque no es necesario que el usuario lo especifique, los programas suelen incluir, junto a cada variable, su varianza y, si se trata de una variable dependiente, su correspondiente proporción de varianza explicada.

La representación gráfica y las convenciones a ser utilizadas son descritas en la Figura 4.1. Cabe recordar que esas convenciones se utilizan desde hace mucho tiempo, (Bollen, 1989 y Hair y cols., 2009) y se continúan utilizando en muchos artículos publicados en las diferentes áreas del conocimiento.

La representación gráfica es

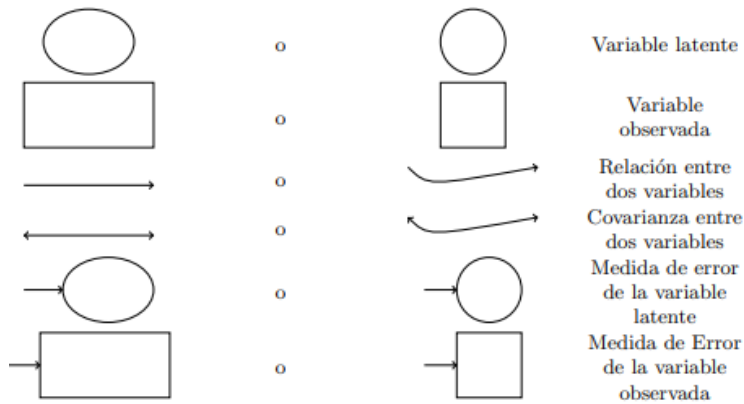


Figura 4.1: Representación gráfica de las variables del modelo

Así mismo, el diagrama de caminos completo es presentado en la Figura 4.2 donde se condensan las ideas que fueron presentadas en los *ítems* anteriores.

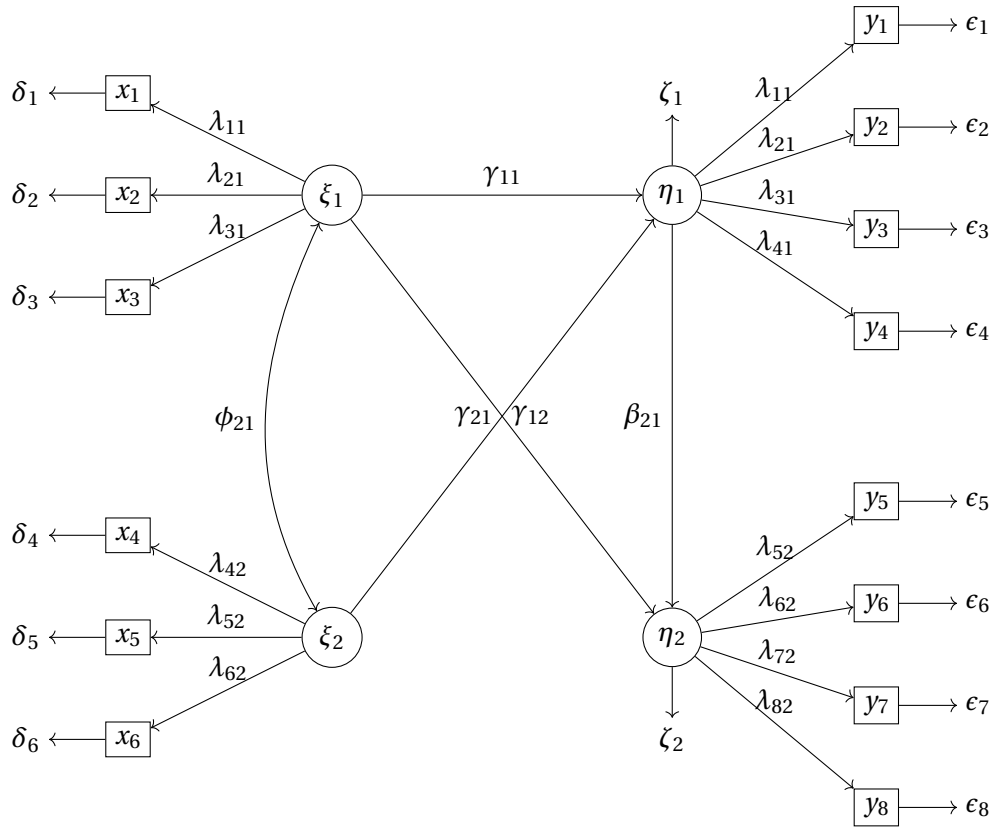


Figura 4.2: Diagrama de caminos para un modelo de ecuaciones estructurales completo

Partiendo del diagrama de caminos que representa el modelo de ecuaciones estructurales completo (Figura 4.2), se puede desprender el diagrama de caminos para el modelo estructural, representado específicamente en la Figura 4.3.

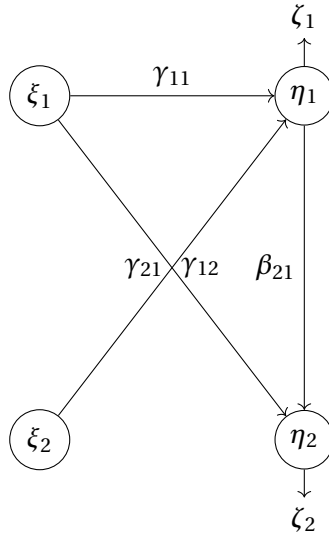


Figura 4.3: Diagrama de caminos para el modelo estructural

Por tanto el modelo estructural puede ser escrito de forma matricial de la siguiente forma:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

que es equivalente a las siguientes ecuaciones:

$$\begin{aligned} \eta_1 &= \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1 \\ \eta_2 &= \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2. \end{aligned}$$

Así mismo, partiendo del diagrama de caminos que representa el modelo de ecuaciones estructurales completo (Figura 4.2), se representa el diagrama de caminos para el modelo de medida, en la Figura 4.4.

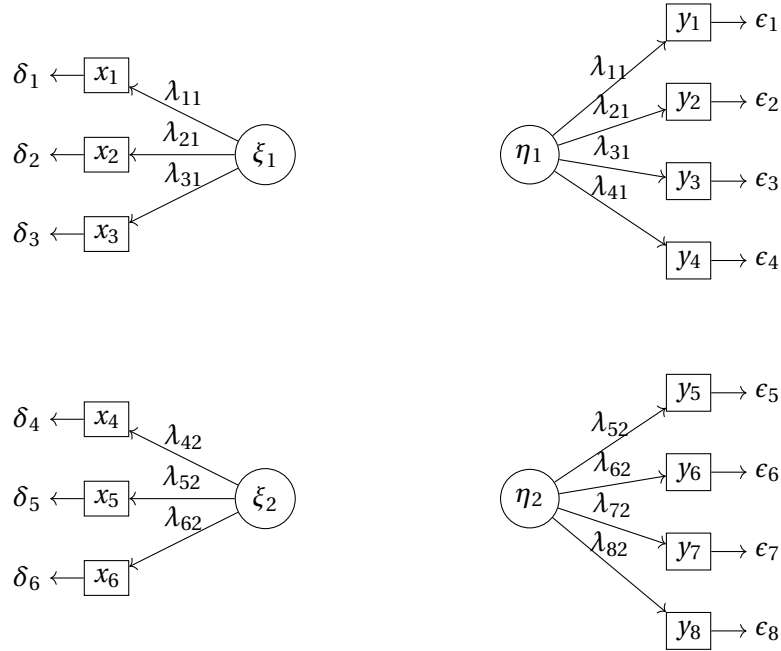


Figura 4.4: Diagrama de caminos para el modelo de medida

Así, se tienen las siguientes ecuaciones tanto para las variables \mathbf{x} y \mathbf{y}

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix},$$

donde los \mathbf{x}_i se escriben de forma específica:

$$\begin{aligned} \mathbf{x}_1 &= \lambda_{11}\xi_1 + \delta_1 \\ \mathbf{x}_2 &= \lambda_{21}\xi_1 + \delta_2 \\ \mathbf{x}_3 &= \lambda_{31}\xi_1 + \delta_3 \\ \mathbf{x}_4 &= \lambda_{42}\xi_2 + \delta_4 \\ \mathbf{x}_5 &= \lambda_{52}\xi_2 + \delta_5 \\ \mathbf{x}_6 &= \lambda_{62}\xi_2 + \delta_6, \end{aligned}$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \\ \mathbf{y}_5 \\ \mathbf{y}_6 \\ \mathbf{y}_7 \\ \mathbf{y}_8 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{41} & 0 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \\ 0 & \lambda_{72} \\ 0 & \lambda_{82} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \boldsymbol{\epsilon}_3 \\ \boldsymbol{\epsilon}_4 \\ \boldsymbol{\epsilon}_5 \\ \boldsymbol{\epsilon}_6 \\ \boldsymbol{\epsilon}_7 \\ \boldsymbol{\epsilon}_8 \end{bmatrix},$$

las variables \mathbf{y}_i como:

$$\begin{aligned} \mathbf{y}_1 &= \lambda_{11}\boldsymbol{\eta}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{y}_2 &= \lambda_{21}\boldsymbol{\eta}_1 + \boldsymbol{\epsilon}_2 \\ \mathbf{y}_3 &= \lambda_{31}\boldsymbol{\eta}_1 + \boldsymbol{\epsilon}_3 \\ \mathbf{y}_4 &= \lambda_{41}\boldsymbol{\eta}_1 + \boldsymbol{\epsilon}_4 \\ \mathbf{y}_5 &= \lambda_{52}\boldsymbol{\eta}_2 + \boldsymbol{\epsilon}_5 \\ \mathbf{y}_6 &= \lambda_{62}\boldsymbol{\eta}_2 + \boldsymbol{\epsilon}_6 \\ \mathbf{y}_7 &= \lambda_{72}\boldsymbol{\eta}_2 + \boldsymbol{\epsilon}_7 \\ \mathbf{y}_8 &= \lambda_{82}\boldsymbol{\eta}_2 + \boldsymbol{\epsilon}_8, \end{aligned}$$

y las matrices de covarianza están dadas por:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{21} \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}$$

$$\boldsymbol{\Theta}_\delta = \begin{bmatrix} \theta_{\delta_{11}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{\delta_{22}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{\delta_{33}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{\delta_{44}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{\delta_{55}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{\delta_{66}} \end{bmatrix} \quad \boldsymbol{\Theta}_\epsilon = \begin{bmatrix} \theta_{\epsilon_{11}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{\epsilon_{22}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{\epsilon_{33}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{\epsilon_{44}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{\epsilon_{55}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{\epsilon_{66}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{\epsilon_{77}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{\epsilon_{88}} \end{bmatrix}.$$

4.5. Efecto Directo, Indirecto y Total

La covarianza entre las variables X e Y es negativa cuando X crece e Y decrece. Hay una relación positiva cuando X crece e Y también y no existe relación entre las variables X e Y cuando la correlación entre ellas es cero. Sin embargo, la covariación y causalidad no son la misma cosa. Cuando se observa una alta relación (covariación) entre dos variables, no interpretarla como una relación causal entre ambas.

En la Sección 4.4 se determinó el principal objetivo del diagrama de caminos y para explicar mejor este concepto, a continuación se habla de los tipos de relación que puede llevar a que dos variables X e Y covaríen:

1. **Relación directa:** En la Figura 4.5, X e Y pueden covariar si X causa Y , esto quiere decir, asume un modelo de regresión en el que Y fuese la variable respuesta y X la variable predictora, de igual forma si Y causa X pueden covariar, lo cual sería una regresión de X sobre Y .



Figura 4.5: Relación simple

2. **Relación recíproca:** En la Figura 4.6, la relación causal entre dos variables puede ser recíproca o unidireccional. Cuando la relación es recíproca (bidireccional) la variable causa es a su vez efecto de la otra. Este tipo de relaciones se representa como dos flechas separadas orientadas en sentidos contrarios.

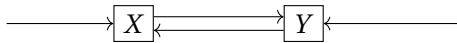


Figura 4.6: Relación recíproca

3. **Relación espúrea:** En una relación espúrea la relación comprende al menos tres variables, en la Figura 4.7, se observa que X e Y covarían si tienen una causa común Z . Una relación espúrea se refiere a la existencia de covariación entre dos variables que es debida, total o parcialmente, a la relación común de ambas variables con una tercera. Esta es la razón por la cual la covariación entre dos variables puede ser muy elevada y, sin embargo, ser nula su relación causal.

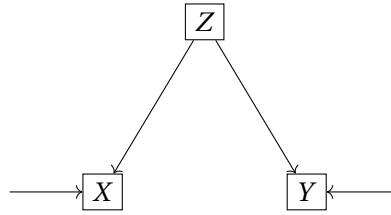


Figura 4.7: Relación espúrea

4. **Relación indirecta:** Una relación causal indirecta implica tres variables como se observa en la Figura 4.8 ya que X e Y también covarían si están relacionadas a través de una tercera variable Z . Existe una relación indirecta entre dos variables cuando una tercera variable modula o mediatiza el efecto entre ambas. Es decir, cuando el efecto entre la primera y la segunda pasa a través de la tercera. A las variables que median en una relación indirecta se les denomina también variables moduladoras.

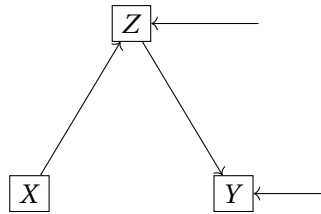


Figura 4.8: Relación indirecta

Una vez establecidos los posibles tipos de covariación o covarianza entre las variables del modelo; se define entonces una serie de reglas de descomposición que permiten establecer las relaciones entre las covariaciones y los parámetros de éste, (Batista y Coenders, 2000).

Por tanto en un diagrama de caminos se tienen los siguientes efectos:

1. **Efecto directo:** Influencia de una variable en otra sin intermedio de ninguna variable.
2. **Efecto indirecto:** Influencia de una variable con otra a través de al menos otra variable.
3. **Efecto total:** Suma del efecto directo e indirecto.

La Tabla 4.1 resume la descomposición de los efectos para el modelo general de ecuaciones estructurales con variables latentes. Estas descomposiciones pueden especializarse para tratar cualquiera de los modelos. Por ejemplo para ilustrar estos efectos, se considera la Figura 4.2 presentada en la Sección 4.4. Un efecto directo de ξ_1 sobre η_1 es γ_{11} , un efecto indirecto de ξ_1 sobre η_1 es $\gamma_{11} * \beta_{21}$ y el efecto total es $\gamma_{11} + \gamma_{11} * \beta_{21}$.

Tabla 4.1: Fórmulas de efectos (Bollen, 1989)

Efecto de ξ	Efectos en		
	η	y	x
Directo	Γ	0	Λ_x
Indirecto	$(I - B)^{-1}\Gamma - \Gamma$	$\Lambda_y(I - B)^{-1}\Gamma$	0
Total	$(I - B)^{-1}\Gamma$	$\Lambda_y(I - B)^{-1}\Gamma$	Λ_x
Efecto de η	η	y	x
Directo	B	Λ_y	0
Indirecto	$(I - B)^{-1}I - B$	$\Lambda_y(I - B)^{-1} - \Lambda_y$	0
Total	$(I - B)^{-1} - I$	$\Lambda_y(I - B)^{-1}$	0

4.5.1. Indicadores reflectivos y formativos

Un indicador formativo se diferencia de uno reflectivo en que el indicador formativo afecta a la variable latente, mientras que en el indicador reflectivo es la variable latente la que produce un efecto en el indicador. Por tanto se relacionados causalmente con la variable latente de manera opuesta.

En la Figura 4.9 se puede ver un modelo reflectivo ya que la variable latente es la causa de las variables o medidas observadas, en cambio en la Figura 4.10 se aprecia el indicar formativo ya que causa una variable latente, así que un cambio en ella no es acompañado por un cambio en todos sus indicadores. (Bollen, 1989).

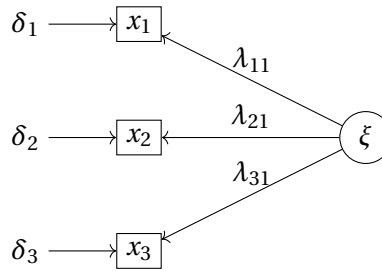


Figura 4.9: Diagrama de indicador reflectivo para el modelo de medida

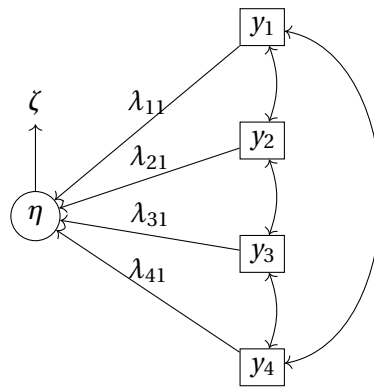


Figura 4.10: Diagrama de indicador formativo para el modelo de medida

4.6. Matriz de covarianza implícita

Los parámetros del modelo están presentes en 8 matrices: \mathbf{B} , $\mathbf{\Gamma}$, $\mathbf{\Lambda}_y$, $\mathbf{\Lambda}_x$, $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Theta}_e$, $\mathbf{\Theta}_\delta$. El vector que contiene todos esos parámetros se denota por θ . Específicamente los parámetros son los coeficientes que acompañan las variables latentes en el modelo estructural y de medida, varianzas y covarianzas de las variables latentes exógenas, errores de medida y errores del modelo de estructura.

La premisa para modelar las ecuaciones estructurales se puede enunciar de la siguiente forma:

Si el modelo estuviera correcto y si los parámetros fuesen conocidos, la matriz poblacional de las covarianzas podría ser reproducida exactamente a partir de la combinación de los parámetros del modelo, (Bollen, 1989).

Si se denota la matriz Σ como la matriz de covarianza poblacional de las variables observadas \mathbf{x} y \mathbf{y} , θ el vector que contiene los parámetros del modelo y $\Sigma(\theta)$ la matriz de covarianzas implícitas en función de los parámetros del modelo, la premisa en forma estadística puede ser enunciada mediante la siguiente hipótesis:

$$H_0 : \Sigma = \Sigma(\theta). \quad (4.4)$$

Como Σ es la matriz de covarianzas de las variables observadas, y por álgebra de matrices se puede escribir Σ como función de los parámetros del modelo:

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}. \quad (4.5)$$

Después de haber enunciado la matriz $\Sigma(\theta)$ en la ecuación 4.5 se procede a describir cada uno de sus componentes en función de los parámetros del modelo, (Matsueda, 2012).

En primer lugar se describe la matriz implícita de \mathbf{y} :

$$\begin{aligned} \Sigma_{yy}(\theta) &= E[\mathbf{y}\mathbf{y}'] \\ &= E[(\Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon})'] \quad \text{Por la ecuación 4.2} \\ &= E[(\Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\eta}'\Lambda_y' + \boldsymbol{\epsilon}')] \\ &= E[\Lambda_y\boldsymbol{\eta}\boldsymbol{\eta}'\Lambda_y' + \Lambda_y\boldsymbol{\eta}\boldsymbol{\epsilon}' + \boldsymbol{\epsilon}\boldsymbol{\eta}'\Lambda_y' + \boldsymbol{\epsilon}\boldsymbol{\epsilon}'] \quad (4.6) \\ &= \Lambda_y E[\boldsymbol{\eta}\boldsymbol{\eta}']\Lambda_y' + \Lambda_y E[\boldsymbol{\eta}\boldsymbol{\epsilon}'] + E[\boldsymbol{\epsilon}\boldsymbol{\eta}']\Lambda_y' + E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] \\ &= \Lambda_y E[\boldsymbol{\eta}\boldsymbol{\eta}']\Lambda_y' + 0 + 0 + \Theta_\epsilon \quad \text{Supuestos del modelo} \\ &= \Lambda_y E[\boldsymbol{\eta}\boldsymbol{\eta}']\Lambda_y' + \Theta_\epsilon. \end{aligned}$$

Por lo tanto por la ecuación 4.1 donde se presenta el modelo de forma matricial:

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{B}\boldsymbol{\eta} + \Gamma\xi + \zeta \\ \boldsymbol{\eta} - \mathbf{B}\boldsymbol{\eta} &= \Gamma\xi + \zeta \\ \boldsymbol{\eta}(\mathbf{I} - \mathbf{B}) &= \Gamma\xi + \zeta \\ \boldsymbol{\eta} &= (\mathbf{I} - \mathbf{B})^{-1}(\Gamma\xi + \zeta). \end{aligned} \quad (4.7)$$

Tomando la ecuación anterior 4.6 y reemplazando por 4.7 se tiene que:

$$\begin{aligned}
\Sigma_{yy}(\theta) &= \Lambda_y E[\eta\eta'] \Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y E[(I - B)^{-1}(\Gamma\xi + \zeta)((I - B)^{-1}(\Gamma\xi + \zeta))'] \Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y (I - B)^{-1} E[(\Gamma\xi + \zeta)(\Gamma\xi + \zeta)' (I - B)^{-1}] \Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y (I - B)^{-1} E[(\Gamma\xi + \zeta)(\xi'\Gamma' + \zeta')] (I - B)^{-1} \Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y (I - B)^{-1} \{E[\xi\xi']\Gamma' + E[\xi\xi'] + E[\zeta\xi']\Gamma' + E[\zeta\xi']\} (I - B)^{-1} \Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y (I - B)^{-1} \{\Gamma\Phi\Gamma' + \Psi\} (I - B)^{-1} \Lambda_y' + \Theta_\epsilon \quad \text{Supuestos del modelo.}
\end{aligned} \tag{4.8}$$

En segundo lugar se describe la matriz implícita de \mathbf{x} , la cual está dada por:

$$\begin{aligned}
\Sigma_{xx}(\theta) &= E[\mathbf{x}\mathbf{x}'] \\
&= E[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)'] \quad \text{Por la ecuación 4.3} \\
&= E[(\Lambda_x\xi + \delta)(\xi'\Lambda_x' + \delta')] \\
&= E[\Lambda_x\xi\xi'\Lambda_x' + \Lambda_x\xi\delta' + \delta\xi'\Lambda_x' + \delta\delta'] \\
&= \Lambda_x E[\xi\xi']\Lambda_x' + \Lambda_x E[\xi\delta'] + E[\delta\xi']\Lambda_x' + E[\delta\delta'] \\
&= \Lambda_x\Phi\Lambda_x' + 0 + 0 + \Theta_\delta \quad \text{Supuestos del modelo} \\
&= \Lambda_x\Phi\Lambda_x' + \Theta_\delta.
\end{aligned} \tag{4.9}$$

Ahora la matriz de covarianza implícita de \mathbf{y} con \mathbf{x} está dada por:

$$\begin{aligned}
\Sigma_{yx}(\theta) &= E[\mathbf{y}\mathbf{x}'] \\
&= E[(\Lambda_y\eta + \epsilon)(\Lambda_x\xi + \delta)'] \quad \text{Por ecuación 4.2, 4.3} \\
&= E[(\Lambda_y\eta + \epsilon)(\xi'\Lambda_x' + \delta')] \quad \text{Por el apéndice A.2, A.4} \\
&= E[\Lambda_y\eta\xi'\Lambda_x' + \Lambda_y\eta\delta' + \epsilon\xi'\Lambda_x' + \epsilon\delta'] \\
&= \Lambda_y E[\eta\xi']\Lambda_x' + \Lambda_y E[\eta\delta'] + E[\epsilon\xi']\Lambda_x' + E[\epsilon\delta'] \\
&= \Lambda_y E[\eta\xi']\Lambda_x' + 0 + 0 + 0 \quad \text{Supuestos del modelo} \\
&= \Lambda_y E[((I - B)^{-1}(\Gamma\xi + \zeta))\xi']\Lambda_x' \quad \text{Por ecuación 4.7} \\
&= \Lambda_y (I - B)^{-1} E[\Gamma\xi\xi' + \zeta\xi']\Lambda_x' \\
&= \Lambda_y (I - B)^{-1} \{\Gamma E[\xi\xi'] + E[\zeta\xi']\}\Lambda_x' \\
&= \Lambda_y (I - B)^{-1} \{\Gamma E[\xi\xi'] + E[\zeta\xi']\}\Lambda_x' \\
&= \Lambda_y (I - B)^{-1} \{\Gamma\Phi + 0\}\Lambda_x' \quad \text{Supuestos del modelo} \\
&= \Lambda_y (I - B)^{-1} \Gamma\Phi\Lambda_x'.
\end{aligned} \tag{4.10}$$

Por último la covarianza implícita de \mathbf{x} con \mathbf{y} :

$$\begin{aligned}
\Sigma_{xy}(\theta) &= E[\mathbf{x}\mathbf{y}'] \\
&= E[(\Lambda_x\boldsymbol{\xi} + \boldsymbol{\delta})(\Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon})'] \quad \text{Por ecuación 4.2, 4.3} \\
&= E[(\Lambda_x\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\eta}'\Lambda_y' + \boldsymbol{\epsilon}')] \\
&= E[\Lambda_x\boldsymbol{\xi}\boldsymbol{\eta}'\Lambda_y' + \Lambda_x\boldsymbol{\xi}\boldsymbol{\epsilon}' + \boldsymbol{\delta}\boldsymbol{\eta}'\Lambda_y' + \boldsymbol{\delta}\boldsymbol{\epsilon}'] \\
&= \Lambda_x E[\boldsymbol{\xi}\boldsymbol{\eta}']\Lambda_y' + \Lambda_x E[\boldsymbol{\xi}\boldsymbol{\epsilon}'] + E[\boldsymbol{\delta}\boldsymbol{\eta}']\Lambda_y' + E[\boldsymbol{\delta}'\boldsymbol{\epsilon}] \\
&= \Lambda_x E[\boldsymbol{\xi}\boldsymbol{\eta}']\Lambda_y' + 0 + 0 + 0 \quad \text{Supuestos del modelo} \\
&= \Lambda_x E[\boldsymbol{\xi}((\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}))']\Lambda_y' \quad \text{Por ecuación 4.7} \\
&= \Lambda_x E[\boldsymbol{\xi}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})'(\mathbf{I} - \mathbf{B})^{-1}]\Lambda_y' \\
&= \Lambda_x E[\boldsymbol{\xi}(\boldsymbol{\xi}'\boldsymbol{\Gamma}' + \boldsymbol{\zeta}')](\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' \\
&= \Lambda_x E[\boldsymbol{\xi}(\boldsymbol{\xi}'\boldsymbol{\Gamma}' + \boldsymbol{\zeta}')](\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' \\
&= \Lambda_x \{E[\boldsymbol{\xi}\boldsymbol{\xi}']\boldsymbol{\Gamma}' + E[\boldsymbol{\xi}\boldsymbol{\zeta}']\}(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' \\
&= \Lambda_x \{\boldsymbol{\Phi}\boldsymbol{\Gamma}' + 0\}(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' \quad \text{Supuestos del modelo} \\
&= \Lambda_x \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y'.
\end{aligned} \tag{4.11}$$

En el Apéndice A se presentan algunas fórmulas para poder entender mejor estas demostraciones, específicamente las ecuaciones A.2 y A.4.

Ahora, con los resultados (4.8), (4.9), (4.10) y (4.11) se llega a la matriz que fue descrita en la ecuación 4.5 pero ahora escrita en términos de los parámetros del modelo, específicamente se tiene:

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\{\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\}(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' + \boldsymbol{\Theta}_\epsilon & \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\Lambda_x' \\ \Lambda_x\boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' & \Lambda_x\boldsymbol{\Phi}\Lambda_x' + \boldsymbol{\Theta}_\delta \end{bmatrix},$$

Por tanto, la hipótesis definida en la ecuación 4.4, puede ser formulada como:

$$\mathbf{H}_0 : \boldsymbol{\Sigma} = \Sigma(\theta) = \begin{bmatrix} \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\{\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\}(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' + \boldsymbol{\Theta}_\epsilon & \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\Lambda_x' \\ \Lambda_x\boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' & \Lambda_x\boldsymbol{\Phi}\Lambda_x' + \boldsymbol{\Theta}_\delta \end{bmatrix}. \tag{4.12}$$

Se debe tener en cuenta que no se conoce la covarianza real de la población, por tanto se toma $\mathbf{S} = \boldsymbol{\Sigma}(\theta)$ para encontrar las estimaciones de los parámetros del modelo, donde \mathbf{S} es la matriz de covarianza muestral.

La estimación se realiza buscando valores para los parámetros desconocidos θ para que $\boldsymbol{\Sigma}(\theta)$ coincida con la matriz de covarianza de las variables observadas.

Tomando la hipótesis formulada en 4.12, para comprobar la igualdad, se deben estimar los respectivos parámetros. El objetivo de la estimación es obtener los valores de los parámetros que permitan mantener esta igualdad con los datos muestrales como será visto en la Sección 4.8. En la próxima sección se abordará el problema de identificabilidad del modelo.

4.7. Identificación del Modelo

Dado un modelo de ecuaciones estructurales, no siempre se puede encontrar un único vector de parámetros θ que optimice la solución del problema. Partiendo de la notación definida en la sección anterior, el modelo de ecuaciones estructurales se dice identificable cuando se cumple la siguiente condición:

$$\Sigma(\theta_1) = \Sigma(\theta_2) \Rightarrow \theta_1 = \theta_2.$$

La identificación tiene como objetivo demostrar que los parámetros desconocidos del modelo son funciones de uno o varios elementos no redundantes de Σ y esta función conduce a una solución única. Si todos los elementos del vector θ están identificados el modelo está identificado. Para determinar la identificación del modelo a partir de esta definición, cada elemento de θ debe ser resuelto en términos de los elementos de la matriz Σ a partir de la ecuación $\Sigma = \Sigma(\theta)$ (dada en la hipótesis presentada en la ecuación 4.4). Si se tienen k variables observadas el número de ecuaciones que se derivan de la igualdad anterior es $\frac{1}{2}k(k+1)$, con lo cual si t es el número de parámetros tres situaciones son posibles:

1. $t = \frac{1}{2}k(k+1)$. En este caso, la información contenida en Σ es suficiente para estimar los parámetros y existe una solución única para los elementos en θ . Si un modelo satisface la condición se le conoce como **saturado**.
2. $t < \frac{1}{2}k(k+1)$. Aquí hay más información en Σ de la requerida para estimar los parámetros y el modelo se denomina **sobreidentificado**. En este caso, el sistema determinado $\Sigma = \Sigma(\theta)$ (4.4) posee infinitas soluciones. En el proceso de estimación se escoge la solución que minimice una función de discrepancia (ver Sección 4.8 $F(\mathbf{S}, \Sigma)$). En este sentido la escogencia es óptima.
3. $t > \frac{1}{2}k(k+1)$. En este caso la información contenida en Σ es insuficiente para estimar los parámetros y se dice que el modelo es **no identificado** o **sub identificado**.

Para establecer la identificación de parámetros en modelos sencillos, en ocasiones el álgebra matricial puede ser de gran ayuda. En el caso de ecuaciones estructurales, donde se tienen muchos parámetros para estimar, se ha establecido un conjunto de reglas para ayudar en la identificación. Los parámetros a estimar son los parámetros libres¹ ² ³ de las 8 matrices paramétricas que fueron escritas al inicio de la Sección 4.6.

Antes de describir las reglas de identificación, se presenta un resumen de la propuesta de Bentler (1995) que es útil para identificar los parámetros del modelo:

1. Todas las varianzas y covarianzas de las variables exógenas son parámetros del modelo.
2. Todas las cargas factoriales que conectan a las variables latentes con sus indicadores son parámetros del modelo.
3. Todos los coeficientes de regresión entre variables observadas o entre variables latentes son parámetros del modelo.
4. Las varianzas y covarianzas entre variables endógenas y las covarianzas entre variables endógenas y exógenas, nunca son parámetros del modelo.

Algunas reglas que son de uso común para establecer la identificabilidad de un modelo de ecuaciones estructurales se presentan a continuación.

Regla de los grados de Libertad, también conocida como regla t

La información disponible para la identificación corresponde a los elementos no redundantes de Σ . Debido a que hay p variables en \mathbf{y} y q variables en \mathbf{x} la matriz Σ es de orden $(p + q) \times (p + q)$ y como la matriz es simétrica hay

$$\frac{(p + q)(p + q + 1)}{2},$$

en consecuencia para escribir los parámetros en θ en función de los elementos de Σ se dispone de $\frac{(p + q)(p + q + 1)}{2}$ ecuaciones. Por lo tanto, una condición necesaria aunque no suficiente para la identificación de los parámetros es que

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

donde t denota el número de elementos de θ .

Esta regla es bastante general en el sentido que puede aplicarse a cualquier modelo SEM. Su mayor utilidad es que permite identificar un modelo no identificado. Aún cuando su cumplimiento no garantiza que el modelo pueda identificarse, por cuanto la

¹Un *parámetro libre* es aquel que no es fijo ni restringido.

²Un *parámetro fijo* es aquel al que se le asigna el valor de una constante conocida.

³Un *parámetro restringido* es aquel que se escribe en función de otros parámetros libres.

condición no es suficiente, su incumplimiento indicaría que el modelo no es identificable por cuanto la condición es necesaria.

Otras alternativas al problema de identificabilidad del modelo pueden ser encontradas en Jöreskog y Sörbom (1986). Una discusión al respecto de la identificación de modelos se puede encontrar en (Kline, 2011). Para una revisión más extensa puede consultar, por ejemplo, Bollen (1989); Bekke y cols. (1994); Rigdon (1995); Brito y Pearl (2002); Bollen y Davis (2009) entre otros.

4.8. Métodos de Estimación

En las secciones anteriores se ha mencionado el problema de estimación. La estimación del modelo de ecuaciones estructurales completo se basa en las relaciones entre la matriz de covarianza y los parámetros del modelo contenidos en el vector θ . Si el modelo en estudio es correcto y se conocen los parámetros de la población, se tiene que $\Sigma = \Sigma(\theta)$, sin embargo se usa la matriz de covarianza muestral S como una estimación de Σ . Por tanto los parámetros del vector θ son estimados de tal manera que la matriz de covarianza estimada $\Sigma(\theta)$ esté tan próxima de S como sea posible. La evaluación de tal proximidad sigue lo que siempre se realiza en estadística, a saber, minimizar una función de la matriz de residuos ($S - \Sigma$). Dado lo anterior, los parámetros en θ se estiman de tal forma que minimicen una función del tipo $F(S, \Sigma)$ (Raykov y Marcoulides, 2006) que se denominan funciones de discrepancia y cumplen las siguientes propiedades:

1. $F(S, \Sigma)$ es un escalar.
2. $F(S, \Sigma) \geq 0$.
3. $F(S, \Sigma) = 0$ si y sólo si $\Sigma = S$.
4. $F(S, \Sigma)$ es continua en Σ y S .

Minimizar una función de ajuste que satisfaga tales condiciones conduce a estimadores consistentes de θ , (Browne, 1984).

Asumiendo que el modelo está identificado, se deben estimar los parámetros libres. Esta estimación está basada en procedimientos iterativos. Los métodos de estimación más empleados, a través de la función $F(S, \Sigma)$, son Máxima Verosimilitud (*Maximum Likelihood* o ML), Mínimos Cuadrados no Ponderados (*Unweighted Least Squares* o ULS), Mínimos Cuadrados Ponderados (*Weighted Least Squares* o WLS), Mínimos Cuadrados Generalizados (*Generalised Least Squares* o GLS) o Mínimos Cuadrados Parciales (*Partial Least Squares* o PLS).

4.8.1. Método de Máxima Verosimilitud

El método de Máxima Verosimilitud (ML), construido para la prueba de hipótesis $H_0 : \Sigma = \Sigma(\theta)$ frente $H_0 : \Sigma \neq \Sigma(\theta)$, bajo el supuesto de que los datos se distribuyen normalmente⁴, es uno de los métodos más utilizados en ecuaciones estructurales. La función $F(\mathbf{S}, \Sigma)$ adopta la expresión:

$$F_{ML}(\mathbf{S}, \Sigma(\theta)) = \ln \det \Sigma(\theta) - \ln \det \mathbf{S} + \text{tr}(\mathbf{S}\Sigma^{-1}(\theta)) - (p + q), \quad (4.13)$$

donde $\Sigma(\theta)$ y \mathbf{S} se asumen definidas positivas, con lo cual sus determinantes son mayores que cero y así, todos los términos del lado derecho de la ecuación 4.13 están bien definidos.

Para maximizar la función de Verosimilitud descrita en la ecuación 4.13 se deriva parcialmente con respecto a θ y usando A.11 y A.12, definidas en el Apéndice A de expresiones matriciales, se tiene:

$$\begin{aligned} \frac{\partial}{\partial \theta} F_{ML}(\mathbf{S}, \Sigma(\theta)) &= \text{tr} \left[\Sigma^{-1}(\theta) \frac{\partial}{\partial \theta} \Sigma(\theta) \right] - \text{tr} \left[\Sigma^{-1}(\theta) \mathbf{S} \Sigma^{-1}(\theta) \frac{\partial}{\partial \theta} \Sigma(\theta) \right] \\ &= \text{tr} \left\{ \Sigma^{-1}(\theta) - \Sigma^{-1}(\theta) \mathbf{S} \Sigma^{-1}(\theta) \right\} \frac{\partial}{\partial \theta} \Sigma(\theta) \\ &= \text{tr} \left\{ [\mathbf{I} - \Sigma^{-1}(\theta) \mathbf{S}] \Sigma^{-1}(\theta) \frac{\partial}{\partial \theta} \Sigma(\theta) \right\} \\ &= \text{tr} \left\{ \Sigma^{-1}(\theta) [\Sigma^{-1}(\theta) - \mathbf{S}] \Sigma^{-1}(\theta) \frac{\partial}{\partial \theta} \Sigma(\theta) \right\}. \end{aligned}$$

La derivada parcial de la ecuación 4.13 respecto θ_u y θ_v , dos componentes cualquiera de θ , es:

$$\frac{\partial}{\partial \theta_u \partial \theta_v} F_{ML}(\mathbf{S}, \Sigma(\theta)) = \text{tr} \left\{ \left[\Sigma^{-1}(\theta) \frac{\partial}{\partial \theta_u} \Sigma(\theta) \right] \left[\Sigma^{-1}(\theta) \frac{\partial}{\partial \theta_v} \Sigma(\theta) \right] \right\}.$$

De esta forma, se aplica un proceso de optimización para estimar los parámetros. Así, una estimación óptima $\hat{\theta}$ maximiza la verosimilitud del modelo, que está asociado a minimizar la distancia $d(\Sigma(\hat{\theta}), \mathbf{S})$ entre la matriz de covarianzas muestral \mathbf{S} y las covarianzas poblacionales inferidas $\Sigma(\hat{\theta})$, aunque no debe confundirse con un ajuste de datos (Kline, 2011).

La maximización de $F_{ML}(\mathbf{S}, \Sigma(\theta))$ lógicamente cumple con las propiedades de los estimadores de Máxima Verosimilitud, es decir, conduce a estimadores asintóticamente normales, invariantes, consistentes, eficientes e insesgados siempre que \mathbf{y} y \mathbf{x} sigan una distribución normal multivariada y el modelo esté sobre identificado.

⁴Los detalles se pueden encontrar en Bollen (1989), con la única necesidad de manipular las expresiones para que la prueba sea cero cuando $S = \Sigma(\theta)$

- La consistencia asegura que al aumentar el tamaño de la muestra, el estimador converge al verdadero valor del parámetro.
- Por otra parte, la eficiencia indica que la distribución del parámetro tiene varianza mínima.
- Al ser insesgado, se tiene que $E[\theta] = \theta$, lo que asegura que el estimador se distribuye alrededor del parámetro.
- Como los estimadores son asintóticamente normales, cuando aumenta el tamaño de la muestra, es posible realizar pruebas de significancia estadística para cada uno de los parámetros estimados. Además de las pruebas individuales, es posible realizar una prueba para el ajuste general del modelo.

El supuesto de normalidad multivariada de las variables observadas, por lo general, no se satisface completamente con datos reales. Las pruebas estadísticas de adecuación del modelo se ven afectadas por la no normalidad (Kaplan, 2000). Para abordar este problema, Satorra y Bentler (1988, 1994, 2001, 2010) desarrollaron pruebas estadísticas para corregir la prueba de bondad de ajuste por el sesgo producido por las desviaciones de la normal multivariada. No obstante, algunas investigaciones (por ejemplo, Raykov y Widaman, 1995) han mostrado que $F_{ML}(\mathbf{S}, \boldsymbol{\Sigma}(\theta))$ puede ser empleada aún con leves desviaciones de la normalidad, en particular, si el interés primario es la estimación de parámetros.

Debido al carácter asintótico, se tiene que para valores grandes de n

$$(n-1)F_{ML} \sim \chi^2 \left(\frac{1}{2}(p+q)(p+q+1) - t \right),$$

donde p es el número de variables observadas de $\boldsymbol{\eta}$, q es el número de variables observadas de $\boldsymbol{\xi}$ y t es el número de parámetros libres del modelo.

Esto provee una prueba de ajuste global del modelo, conocida como prueba de bondad de ajuste χ^2 , para contrastar la hipótesis establecida en 4.4, a saber, $\mathbf{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\theta)$.

En este caso, si no se rechaza la hipótesis nula, se tiene evidencia de que el modelo tuvo un buen ajuste. Es importante señalar que esta prueba debe usarse para un tamaño de muestra suficientemente grande y que las variables tengan la normalidad multivariada.

4.8.2. Mínimos Cuadrados Generalizados

La estimación de mínimos cuadrados generalizados tiene como objetivo minimizar la diferencia entre los valores observados en \mathbf{S} y los elementos correspondientes predichos en $\boldsymbol{\Sigma}(\theta)$. La función a minimizar es:

$$F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) = \frac{1}{2} \text{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\mathbf{W}^{-1}\}^2, \quad (4.14)$$

donde \mathbf{W}^{-1} es una matriz de peso para la matriz residual. Minimizando la función en 4.14 se obtienen los estimadores de mínimos cuadrados generalizados para los parámetros del modelo. Se pueden utilizar varias matrices de pesos, sin embargo, la más habitual es $\mathbf{W}^{-1} = \mathbf{S}^{-1}$ lo que lleva a la obtención de estimadores consistentes, eficientes y con distribución asintóticamente normal multivariada. También con esta selección de la matriz de pesos se obtiene una distribución para el estimador $(n-1)F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta))$ idéntica a la obtenida en el caso de la función asociada a la verosimilitud $F_{ML}(\mathbf{S}, \boldsymbol{\Sigma}(\theta))$, bajo la suposición de que el modelo es correcto, a saber:

$$(n-1)F_{GLS} \sim \chi^2 \left(\frac{1}{2}(p+q)(p+q+1) - t \right).$$

Bajo la suposición de normal multivariada tanto los estimadores de máxima verosimilitud como los estimadores de mínimos cuadrados generalizados son óptimos en el sentido de ser más precisos en grandes muestras, (Bistaffa 2010).

Así, la ecuación 4.14 puede ser escrita como:

$$F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) = \frac{1}{2} \text{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\mathbf{S}^{-1}\}^2. \quad (4.15)$$

La derivada parcial de la ecuación 4.15 con respecto a θ es dada por:

$$\begin{aligned} \frac{\partial}{\partial \theta} F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) &= \frac{1}{2} \text{tr} \left\{ 2[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\mathbf{S}^{-1} \frac{\partial}{\partial \theta} [\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\mathbf{S}^{-1} \right\} \\ &= \text{tr} \left\{ [\mathbf{S}^{-1}[\boldsymbol{\Sigma}(\theta) - \mathbf{S}]\mathbf{S}^{-1} \frac{\partial}{\partial \theta} \boldsymbol{\Sigma}(\theta)] \right\}, \end{aligned} \quad (4.16)$$

ya que la derivada parcial de \mathbf{S} con respecto a θ es igual a la matriz cuadrada nula de orden $p+q$ variables observadas.

De la derivada parcial de la ecuación 4.16, se procede a hacer la derivada parcial de la ecuación 4.15 respecto θ_u y θ_v , para dos componentes cualquiera de θ se obtiene:

$$\frac{\partial}{\partial \theta_u \partial \theta_v} F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) = \text{tr} \left\{ \left[\mathbf{S}^{-1} \frac{\partial}{\partial \theta_u} \boldsymbol{\Sigma}(\theta) \right] \left[\mathbf{S}^{-1} \frac{\partial}{\partial \theta_v} \boldsymbol{\Sigma}(\theta) \right] \right\}.$$

Las estimaciones obtenidas por el método GLS tienen las mismas propiedades que las estimaciones obtenidas por el método ML. Bajo la suposición de normalidad multivariada, tanto los estimadores ML como los estimadores GLS son excelentes cuando el tamaño de la muestra es grande (Bollen, 1989).

4.8.3. Mínimos Cuadrados no Ponderados

La función de discrepancia de mínimos cuadrados no ponderados es un caso particular de la función de mínimos cuadrados generalizados para $\mathbf{W} = \mathbf{I}$, luego puede ser escrita a partir de la siguiente expresión:

$$F_{ULS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) = \frac{1}{2} \text{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]^2\}. \quad (4.17)$$

Minimizando esta expresión se obtienen los estimadores de mínimos cuadrados no ponderados para los parámetros del modelo. Esta función minimiza la suma de cuadrados de cada elemento de la matriz residual $[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]$. Aunque es la más simple de todas las funciones de discrepancia, determina estimadores consistentes y no requiere el supuesto de que las variables observadas siguen alguna distribución particular. Por otro lado, no cuenta con los estimadores asintóticamente más eficientes, no tiene una escala invariante y no se puede evaluar la prueba χ^2 , (Bollen, 1989). La ponderación impuesta sobre la matriz de residuos es el único factor que diferencia $F_{ULS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta))$ de $F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta))$.

$$\begin{aligned} F_{ULS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) &= \frac{1}{2} \text{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]'[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\} \\ &= \frac{1}{2} \text{tr}\{[\mathbf{S}' - \boldsymbol{\Sigma}'(\theta)][\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\} \\ &= \frac{1}{2} \text{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]^2\}. \end{aligned} \quad (4.18)$$

Por tanto, usando las ecuaciones del Apéndice A, A.8 y A.9 la derivada parcial de la función anterior está dada por

$$\frac{\partial}{\partial \theta} F_{ULS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) = \frac{1}{2} \text{tr} \left\{ 2[\mathbf{S} - \boldsymbol{\Sigma}(\theta)] \frac{\partial}{\partial \theta} [\mathbf{S} - \boldsymbol{\Sigma}(\theta)] \right\} - \text{tr} \left\{ [\boldsymbol{\Sigma}(\theta) - \mathbf{S}] \frac{\partial}{\partial \theta} \boldsymbol{\Sigma}(\theta) \right\},$$

ya que la derivada parcial de \mathbf{S} con respecto a θ es igual a la matriz cuadrada nula de orden $p + q$ variables observadas.

De la derivada parcial de ecuación 4.8.3, se procede a obtener la derivada parcial de la ecuación 4.18 respecto θ_u y θ_v , de nuevo, para dos componentes cualquiera de θ es:

$$\frac{\partial}{\partial \theta_u \partial \theta_v} F_{ULS}(\mathbf{S}, \boldsymbol{\Sigma}(\theta)) = \text{tr} \left\{ \left[\frac{\partial}{\partial \theta_u} \boldsymbol{\Sigma}(\theta) \right] \left[\frac{\partial}{\partial \theta_v} \boldsymbol{\Sigma}(\theta) \right] \right\}$$

donde "tr" es la traza o la suma de los elementos de la diagonal de una matriz.

4.9. Diagnóstico de ajuste del modelo

Tras la etapa de estimación se dispone del vector $\hat{\theta}$ que optimiza el modelo de ecuaciones estructurales. Entonces, surge la pregunta relevante: ¿se puede aceptar el modelo teórico como representativo de los datos? No existe un consenso sobre qué índice de bondad de ajuste es mejor emplear, por lo que lo preferible es observar varios de los índices disponibles.

Se recomienda el Artículo de *Some Clarifications and Recommendations on Fit Indices*, *Psy 523/623* (Newson, 2012) donde se proporciona una descripción general simplificada de algunos de los índices de ajuste SEM más conocidos para ayudar a dar sentido a la vertiginosa variedad de medidas de ajuste del modelo. Allí se incluye un número considerable de opiniones profesionales y son una representación razonable de los estándares de práctica actuales de los usuarios de modelos de ecuaciones estructurales.

En esta sección se presentan algunos índices que ayudan a evaluar el ajuste de un modelo propuesto. Es importante que antes de realizar cualquier interpretación del modelo estimado, se evalúe si los resultados obtenidos son válidos.

4.9.1. Ajuste General del modelo

El ajuste general tiene como objetivo verificar que las medidas de ajuste evalúan si la matriz de covarianza ajustada por el modelo, $\Sigma(\hat{\theta})$, es consistente con la matriz de covarianza muestral, \mathbf{S} .

Tiene la ventaja de ser medidas que evalúan el modelo en su conjunto, permitiendo detectar fallas del modelo que no se pueden detectar en las pruebas de parámetros estimados, sin embargo tienen las desventajas de no poder ser aplicadas en modelos exactamente identificados y puede diferir de los ajustes de los componentes del modelo, (Bistaffa, 2010 y Bollen, 1989).

Raíz cuadrada media del error

Un residuo individual es dado por $(s_{ij} - \hat{\sigma}_{ij})$ siendo s_{ij} un elemento de \mathbf{S} y $\hat{\sigma}_{ij}$ un elemento de $\Sigma(\hat{\theta})$. Si el residuo es positivo es porque el modelo subestima la covarianza entre dos variables, si es negativo es porque la covarianza estimada es demasiado elevada. Idealmente todos los residuos deberían estar próximos de cero. A partir de este concepto se define la Raíz del Cuadrado Residual Medio (RMR, en inglés, *Root Mean-square Residual*) que en términos generales es una medida que se utiliza para cuantificar las diferencias entre $\Sigma(\hat{\theta})$ y \mathbf{S} . La expresión matemática es dada por:

$$RMR = \left[2 \sum_{i=1}^k \sum_{j=1}^k \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{(p+q)(p+q+1)} \right]^{\frac{1}{2}}.$$

Si es usada una matriz de correlaciones, el RMR varía entre 0 y 1, idealmente debería asumir valores inferiores a 0.05. Si se usa una matriz de covarianzas, el RMR es siempre positivo y no tiene límite superior, sólomente pudiendo ser interpretado en relación a la dimensión de las varianzas y covarianzas observadas en la matriz \mathbf{S} (se vuelve más fácil si todas las variables observadas están estandarizadas).

Residuos estandarizados

Jöreskog y Sorbom (1986) desarrollaron residuos estandarizados, que toman en cuenta el tamaño de la muestra y corrigen las variables para que estén todas en la misma escala. La fórmula para calcular los residuos estandarizados es

$$e_{i,j} = \frac{(s_{ij} - \hat{\sigma}_{ij})}{\left[\frac{1}{n}(\hat{\sigma}_{ii}\hat{\sigma}_{jj} + \hat{\sigma}_{ij})\right]^{\frac{1}{2}}}$$

Los autores indican que si existen muchos residuos estandarizados con módulo mayor que 2 o 3 hay indicios de mala especificación del modelo. Así, al contrario de lo que ocurre cuando son utilizados los residuos calculados directamente, los residuos estandarizados no solamente ponderan su medida por el tamaño muestral sino que también estandarizan la influencia de variables con diferentes escalas.

Prueba χ^2

En la Sección 4.8 se definió la función de discrepancia $F(\mathbf{S}, \mathbf{\Sigma})$, la cual varía de acuerdo al método de estimación utilizado. Se supone que si el modelo está correctamente especificado, si la muestra es suficientemente grande, si se analiza una matriz de covarianzas y si las variables observadas siguen una distribución normal multivariada $(n-1)F(\mathbf{S}, \mathbf{\Sigma})$ tiene una distribución asintótica χ^2 .

Si se verifican los supuestos mencionados, la estadística χ^2 puede ser utilizada para probar la hipótesis $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}(\theta)$ definida en la ecuación 4.4 (probar la hipótesis de que el modelo especificado sí reproduce la estructura de varianzas y covarianzas en la población). Los grados de libertad de la distribución χ^2 son dados por $\frac{1}{2}(p+q)(p+q+1) - t$ siendo $p+q$ el número total de variables observadas y t el número total de parámetros a estimar.

Vale recordar que para no rechazar la hipótesis nula, de que el modelo especificado reproduce la estructura de asociaciones en la población, el p-valor tiene que ser mayor o igual al nivel de significancia α deseado.

Se resalta que en la práctica la prueba chi-cuadrado χ^2 no es considerada, por muchos investigadores, como un buen índice de ajuste, por los siguientes factores:

- Tamaño muestral: Modelos ajustados con pequeñas muestras pueden ser aceptados, es difícil obtener un valor de la prueba χ^2 que no se acepte.

- El estadístico es sensible a tamaños muestrales superiores a 200 observaciones (Long, 1983). Por tal motivo al aumentar el N, es probable que el valor p de la prueba estadística sea significativo lo que conlleva a rechazar un modelo que sí presenta un buen ajuste de los datos (Hu y Bentler, 1999).
- Modelos con más variables, tienden a tener valores altos de la prueba χ^2 .
- La prueba χ^2 es afectada por la distribución de las variables. Variables con alta asimetría y curtosis tienden a incrementar el valor de la prueba.

4.9.2. Medidas de ajuste alternativas

Raíz cuadrada media estándar

La raíz cuadrada media estándar (SRMR, en inglés, *Standardized Root Mean Square Residual*) está dado por:

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

El índice presentado asume que el modelo estimado es correcto y señala que el modelo es correcto a medida que se aproxima a cero, por lo que el incremento de los residuales indica empeoramiento en el ajuste del modelo. Los valores de *SRMR* menores a 0,10 son generalmente considerados favorables.

Raíz del error cuadrático medio de aproximación

La raíz del error cuadrático medio de aproximación (RMSEA, del *Root Mean Square Error of Approximation*) se denota por:

$$RMSEA = \sqrt{\frac{\hat{F}_0}{g.l.p}}$$

donde $\hat{F}_0 = \max\left(\frac{\hat{C} - g.l.p}{n}, 0\right)$ y $\hat{C} = n \times F_p$ donde F_p y $g.l.p$ son respectivamente la función de discrepancia y el número de grados de libertad, ambos referidos al modelo del investigador.

Indica el grado en el que el modelo se ajusta a la matriz de covarianza de la muestra. Melhado (2004) señala que los valores por debajo de 0,05 es un buen ajuste, los valores entre 0,05 y 0,10 indican un ajuste razonable y los valores por encima de 0,10 indican un ajuste deficiente.

Índice de ajuste normalizado

El Índice de Ajuste Normalizado (o NFI, del *Normalized fit index*) mide la reducción de la función de discrepancia al pasar del modelo saturado al modelo del investigador.

$$NFI = \frac{F_S - F_P}{F_P},$$

donde F_S y F_P son los valores de las funciones de discrepancia de los modelos saturados y del investigador, respectivamente. Este indicador asume valores en $[0; 1]$ y sugiere que el modelo está bien a medida que se acerca a la unidad, con 0,9 como punto de corte (Bentler y Bonett, 1980). Su uso requiere cuidado, porque el tamaño de la muestra influye en las esferas matemáticas de NFI y también tiende a aumentar con la adición de parámetros, como ocurre con el coeficiente de determinación del análisis de regresión (Melhado, 2004).

Índice de ajuste corregido

Para mejorar el NFI, el Índice de Ajuste Corregido (o IFI, del *Incremental Fit Index*) es una propuesta que busca reducir la influencia del tamaño de la muestra y la adición de parámetros. Está dado por:

$$IFI = \frac{F_S - F_P}{F_S - [g.l.p / (n - 1)]},$$

donde $g.l.p$ es el número de grados de libertad del modelo del investigador, mientras que F_S y F_P son los valores respectivos de las funciones de discrepancia de los modelos saturados y del investigador.

El indicador sugiere que los valores cercanos a 1 representan modelos válidos, pero no necesariamente varía en $[0; 1]$, lo que sugiere problemas de sobreajuste cuando se encuentran valores muy grandes (Melhado, 2004). Además, las diferencias entre NFI e IFI disminuyen a medida que aumenta el tamaño de la muestra (De Souza, 2013).

Índice de Tucker-Lewis o índice de ajuste no normalizado

El índice de Tucker-Lewis (o TLI, del *Tucker-Lewis Index*), también se conoce como el índice de ajuste no normalizado (o NNFI, del índice de ajuste no normalizado), tiene como objetivo medir la reducción de la función de discrepancia cuando se pasa del modelo saturado al modelo del investigador, considerando los grados de libertad. Se denota como:

$$TLI = NNFI = \frac{(F_S / g.l.s) - (F_P / g.l.p)}{(F_S / g.l.s) - [1 / (n - 1)]},$$

donde F_S y F_P son las funciones de discrepancia respectivas de los modelos saturados y del investigador, mientras que $g.l.s$ y $g.l.p$ son los grados de libertad correspondientes que varían para los modelos mencionados. Según Melhado (2004), valores en $[0,9;1]$ indican el ajuste sin depender del tamaño de la muestra, pero el índice no está restringido a $[0;1]$. El mismo autor también está convencido de que los valores negativos sugieren problemas de mala especificación y que los valores superiores a la unidad despiertan sospechas de justificación.

Índice de reajuste comparativo

También asumiendo valores en $[0;1]$, Widaman & Thompson (2003) presentan la siguiente expresión para el Índice de ajuste comparativo (o CFI, de *Comparative Fit Index*):

$$CFI = 1 - \frac{\max(\{F_P - [g.l.p/(n-1)], 0\}}{\max(\{F_P - [g.l.p/(n-1)], \{F_S - [g.l.s/(n-1)]\}, 0)}$$

donde F_S y F_P son las funciones de discrepancia respectivas de los modelos saturados y del investigador, mientras que $g.l.s$ y $g.l.p$ son los grados de libertad correspondientes que varían para los modelos mencionados. Sin estar influenciado por el tamaño de la muestra, pero sí afectado por la adición de variables, el indicador CFI califica el modelo como: malo, si es $[0; 0,90]$; bueno, cuando está entre $(0,90; 0,95]$; o muy bueno, si pertenece a $(0,95; 1]$.

Índice de no centralidad relativa

Widaman & Thompson (2003) describen el índice de no centralidad relativa (o RNI, de *Relative Noncentrality Index*) a partir de la siguiente expresión:

$$RNI = 1 - \frac{F_P - [g.l.p/(n-1)]}{F_S - [g.l.s/(n-1)]}$$

donde F_S y F_P son las funciones de discrepancia respectivas de los modelos saturados y del investigador, mientras que $g.l.s$ y $g.l.p$ son los grados de libertad correspondientes que varían para los modelos mencionados. Los mismos autores justifican que valores superiores a 0,95 son aceptables.

Índices de calidad de ajuste

El índice de calidad de ajuste (o GFI, *Goodness-of-Fit Index*) tiene diferentes formas, que varían según el método de estimación. Las siguientes son las versiones de GFI para ML, ULS y GLS:

$$GFI_{ML} = 1 - \frac{tr[(\Sigma^{-1}(\hat{\theta})S - I)^2]}{tr[(\Sigma^{-1}(\hat{\theta})S)^2]},$$

$$GFI_{ULS} = 1 - \frac{tr[(S - \Sigma(\hat{\theta}))^2]}{tr[S^2]},$$

$$GFI_{GLS} = 1 - \frac{tr[(I - \Sigma(\hat{\theta})S^{-1})^2]}{p + q},$$

donde $\hat{\theta}$ es un vector de estimaciones del vector de parámetros θ , $\Sigma(\hat{\theta})$ es la matriz de covarianza $\Sigma(\theta)$ evaluada en el vector $\hat{\theta}$ que minimiza la función de discrepancia F considerada, S es la matriz de covarianza de $z = [y' \ x']'$, $tr(\cdot)$ designa la función de seguimiento, I es la matriz de identidad, p es el número de componentes de y y q es el número de componentes de x .

Independientemente de la función de discrepancia, los indicadores GFI buscan medir el número relativo de covarianza S que predice $\Sigma(\hat{\theta})$. Eventualmente pueden asumir valores negativos y tener la siguiente característica: $GFI = 1$ si, y sólo si, $S = \Sigma(\hat{\theta})$. En palabras, el ajuste de SEM es perfecto solo cuando se obtiene un GFI unitario (Tanaka y Huba, 1985).

Con base en revisiones de literatura y experiencias prácticas, Marôco (2010) sugiere a los valores de GFI la siguiente interpretación: si $GFI < 0,80$, es mala; si $0,80 \leq GFI < 0,90$, es pobre; si $0,90 \leq GFI < 0,95$, es bueno; y si $GFI \geq 0,95$, es muy bueno.

Índices de calidad de ajuste corregidos

Con el fin de resolver el GFI a los grados de libertad y al número de variables observadas en el SEM, se ha propuesto el Índice de Bondad de Ajuste Ajustado (AGFI) y también según el método de estimación elegido, con las siguientes versiones para GFI para ML, ULS y GLS:

$$AGFI_{ML} = 1 - \left[\frac{(p + q)(p + q + 1)}{2g.l.p} \right] (1 - GFI_{ML}),$$

$$AGFI_{ULS} = 1 - \left[\frac{(p + q)(p + q + 1)}{2g.l.p} \right] (1 - GFI_{ULS}),$$

$$AGFI_{GLS} = 1 - \left[\frac{(p+q)(p+q+1)}{2g.l.p} \right] (1 - GFI_{GLS}),$$

donde que $g.l.p$ es el número de grados de libertad del modelo del investigador, p es el número de variables endógenas, q es el número de variables exógenas y, GFI_{ML} , GFI_{ULS} y GFI_{GLS} secuencialmente son los Indicadores GFI justificados por ML, ULS y GLS. El indicador AGFI también puede asumir valores negativos, así como el GFI. Además, se ve fácilmente que $GFI = 1$ si, y sólo si, $AGFI = 1$.

Capítulo 5

Aplicación

5.1. Descripción del proyecto

Con el objetivo de aplicar la teoría descrita en los Capítulos 3 y 4 se han utilizado variables asociadas a un estudio de Neuropsicología, un área de bastante interés para diferentes grupos de investigación. Los datos utilizados son reales y fueron obtenidos a través del proyecto “Corazones de Baependi” (Processo Fapesp 2007/58150-7) conducido por el Laboratorio de Genética y Cardiología Molecular (Incor/USP) (de Oliveira y cols., 2008). El proyecto tiene como objetivo general encontrar determinantes genéticos que modulan o regulan enfermedades cardiovasculares evaluando la influencia de factores genéticos y ambientales.

El proyecto ha tenido varias fases para la recolección de datos. Una primera fase fue llevada a cabo entre diciembre de 2005 y enero de 2006 en que un total de 81 familias fueron muestreadas en el municipio de Baependi, Minas Gerais, Brasil. En esa primera fase del estudio, varios fenótipos (o variables) de los individuos fueron evaluados, por ejemplo, glicemia, colesterol total, presión arterial, fracción de lipoproteínas, triglicéridos, entre otras. También, el ADN fue extraído de las muestras de sangre de cada individuo y almacenado para genotipar posteriormente.

Para los estudios en el área de Psicología, los datos fueron recolectados entre abril de 2013 y marzo de 2016 en la estación de investigación permanente del estudio, con sede en el centro de Baependi. Se recopilaban datos de 1.735 individuos pertenecientes a 134 familias extendidas. El protocolo de estudio fue aprobado por el comité de Ética del Hospital de las Clínicas, Universidad de São Paulo, Brasil, y cumplió con los estándares éticos internacionales sobre experimentación humana. A continuación, de forma resumida se describe la metodología utilizada en la toma de datos. Se invitó a participar a todos los miembros de la familia extendida mayores de 18 años que fueron seleccionados al azar de 11 de los 12 distritos censales en Baependi. Todos los participantes proporcionaron su consentimiento informado por escrito. Para las personas analfabetas, el investigador leyó el contenido en voz alta y recogió la autorización de un repre-

sentante legal de acuerdo con la decisión del individuo de participar en el estudio. Los dos examinadores son psicólogos capacitados para supervisión durante la recopilación de datos. Una descripción detallada del estudio de Baependi, puede ser encontrada en Egan y cols. (2016).

Un frente de investigación, del proyecto en mención, que se ha venido desarrollando es en el área de Neuropsicología en que varios investigadores expertos en el área han estado recolectando datos y analizando, por ejemplo estudios de depresión y ansiedad para saber cuánto de estas enfermedades se le atribuyen a factores genéticos (Taporoski y cols., 2015); por otro lado estudios de influencia genética en factores cognitivos (Taporoski y cols., 2015); influencia de factores genéticos y ambientales sobre cronotipo o preferencia diurna (von Schantz y cols., 2015) y estudios acerca del insomnio (Ahmed y cols., 2019) entre otros.

En la actualidad, los investigadores de esta área están interesados en dar continuidad al estudio en el área de Cognición, específicamente sobre las funciones ejecutivas, que son las habilidades cognitivas involucradas en la formación, planificación, selección, mantenimiento, seguimiento e implementación de metas que se mantienen temporalmente en la memoria de trabajo, (Friedman y Miyake, 2017). Cabe anotar que las definiciones pertinentes a esta aplicación fueron presentadas en el Capítulo 2.

En las siguientes secciones se describen las variables utilizadas en el estudio, se realizará la aplicación del análisis factorial y luego se plantea el modelo de ecuaciones estructurales que junto con los investigadores del área se propone. Se especificarán los resultados, validación del modelo y conclusiones.

5.2. Análisis Factorial - Aplicación

En primer lugar, como fue especificado en el Capítulo 2, las variables que se analizarán son: **TPI, adyacencia, runs, redundancia, ep_pmerrorsalt, ep_pmtimecost, coupon y repgap**. El número de individuos en la muestra fue de 717, no hay valores faltantes en ninguno de los individuos para las 8 variables descritas anteriormente.

En la Tabla 5.1 de forma muy resumida se pueden ver las mediciones (Media, Desviación Estándar, Rango, Asimetría y Curtosis) para las 8 tareas utilizadas que representan las variables de análisis; se pueden obtener las siguientes descripciones:

- Casi todas las medidas tienen una asimetría positiva.
- El coeficiente de curtosis de las variables: runs, coupon, redundancia, repgap y ef_pmerrorsalt presentan un valor mayor a 3, por tanto la distribución para estas variables es Leptocúrtica, por otro lado las variables TPI, adyacencia y ef_pmtimecost tiene un valor menor a 3, su distribución es Platicúrtica.

- Sólo la variable runs tiene una desviación estándar baja, por ello, se tiene que la mayoría de los datos están próximos a la media y las variables redundancia y repgap también poseen una desviación estándar relativamente baja con respecto al comportamiento de las desviaciones estándar de las demás variables.
- También se observa que la variable TPI presenta un coeficiente de variabilidad bajo, pero también se puede observar que las variables runs, redundancia, repgap, ep_pmtimecost y ef_pmerrorsalt tiene un coeficiente de variación muy alto.

Tabla 5.1: Estadísticas descriptivas de las variables incluidas en el estudio (N=717)

Variable (tarea)	Media	Des. Estándar	CV	Rango	Asimetría	Curtosis
TPI	82,8740	14,3012	17,2565	32,250 a 135,9375	0,018	0,431
adyacencia	28,4239	11,8833	41,8078	2,04 a 70,410	0,382	0,058
runs	1,2448	0,6928	55,6555	0,080 a 6,4667	2,958	16,139
coupon	17,3773	6,6271	38,1365	0 a 79	3,829	27,512
redundancia	1,4808	1,3709	92,5783	0,0617 a 11,298	2,768	10,752
repgap	2,47	1,588	64,2914	0 a 13	1,122	4,156
ef_pmtimecost	31,44	22,959	73,0248	-63 a 151	0,904	2,474
ef_pmerrorsalt	2,56	3,041	118,7890	0 a 29	2,272	10,080

Conociendo estas medidas descriptivas de las variables, se procede a realizar el análisis factorial para descubrir cómo es la asociación con posibles factores latentes. Se escogieron 3 factores recomendados por los investigadores. Cabe resaltar que este análisis factorial se toma como una base para el posterior planteamiento de un modelo de ecuaciones estructurales. Se realizó la prueba de normalidad multivariada para el conjunto de datos, se utilizó la prueba de Shapiro- Wilk para datos multivariados y esta no dio una significancia estadística, sin embargo se procede con el análisis, dado que se tiene un tamaño muestral grande.

Resultados del análisis factorial

En primer lugar en la Tabla 5.2 se presentan algunas medidas de adecuación del modelo de análisis factorial. Se observa que el índice *Kaiser-Meyer-Olkin* (KMO) es de 0,637 lo cual indica que es aceptable proceder con el análisis factorial de estos datos. La prueba de esfericidad de Bartlett que presentó un valor p de la prueba estadística < 0,05, indica que la correlación entre las variables es válida.

Tabla 5.2: KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-MeyerOlkin(KMO)		0,637
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	1108,888
	gl	28
	sig	$p < 0,01$

Por otro lado, el MSA del modelo fue de 0,64. En la Tabla 5.3 se presenta el MSA por cada una de las variables donde se observa que la medida es mayor que 0,5 lo cual indica que es procedente incluir las 8 variables en el ajuste del modelo.

Tabla 5.3: MSA para cada variable

TPI	adyacencia	runs	coupon	redundancia	repgap	ef_pmerrorsalt	ef_pmtimecost
0,60	0,66	0,71	0,61	0,63	0,64	0,56	0,62

En la Tabla 5.4, se muestra un listado de los autovalores de la matriz de varianzas-covarianzas y del porcentaje de varianza que representa cada uno de los factores. Los autovalores expresan la cantidad de la varianza total que es explicada por cada factor y los porcentajes de varianza explicada asociados a cada factor, se obtienen dividiendo su correspondiente autovalor por la suma de los autovalores. Por defecto, se extraen tantos factores como autovalores mayores que 1. En este estudio hay tres factores que consiguen explicar un 42,644% de la varianza total.

Tabla 5.4: Varianza total explicada: Método de extracción por máxima verosimilitud

Factor	Varianza total explicada					
	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de Var	% acumulado	Total	% de Var	% acumulado
1	2,231	27,883	27,883	1,798	22,478	22,478
2	1,760	21,999	49,882	1,243	15,535	38,013
3	1,113	13,910	63,792	0,371	4,631	42,644

Cargas factoriales para cada factor sin aplicar ninguna rotación

En la Tabla 5.5 se presentan las variables asociadas con los tres factores. Se observa que hay confusión en las variables que están asociadas a cada factor, por lo tanto se procede a realizar una rotación varimax, apoyados en lo que indica la literatura (Díaz y Morales 2012).

Tabla 5.5: Tabla de cargas factoriales sin rotación

Variables	Factores		
	Factor 1	Factor 2	Factor 3
TPI	-0,938	-0,126	0,931
adyacencia	0,081	-0,135	0,890
runs	0,053	-0,053	0,118
coupon	0,001	0,706	0,364
redundancia	-0,043	0,483	-0,056
repgap	0,004	0,277	-0,204
ef_pmerrorsalt	0,009	0,011	0,624
ef_pmtimecost	0,010	0,005	1,154

Cargas factoriales para cada factor aplicando la rotación varimax

Después de realizar la rotación, se puede observar en la Tabla 5.6 una clara agrupación de patrones donde prevalecen variables que definen los factores. Específicamente:

- **Factor 1:** TPI, adyacencia, runs.
- **Factor 2:** coupon, redundancia, repgap.
- **Factor 3:** ef_pmerrorsalt, ef_pmtimecost.

Tabla 5.6: Matriz de cargas factoriales utilizando el método de rotación varimax

Variables	Factores		
	Factor 1	Factor 2	Factor 3
TPI	-0,970	-0,017	0,029
adyacencia	0,677	-0,149	0,221
runs	0,598	-0,079	0,040
coupon	0,005	0,740	0,086
redundancia	-0,042	0,652	-0,017
repgap	0,055	0,490	-0,081
ef_pmerrorsalt	0,140	0,024	0,295
ef_pmtimecost	0,129	0,009	0,467

En la Figura 5.1 se puede observar de forma gráfica las asociaciones presentadas en la Tabla 5.6, donde se visualiza mejor la asociación de las variables con los respectivos factores, las cargas factoriales se aproximaron a un sólo dígito decimal.

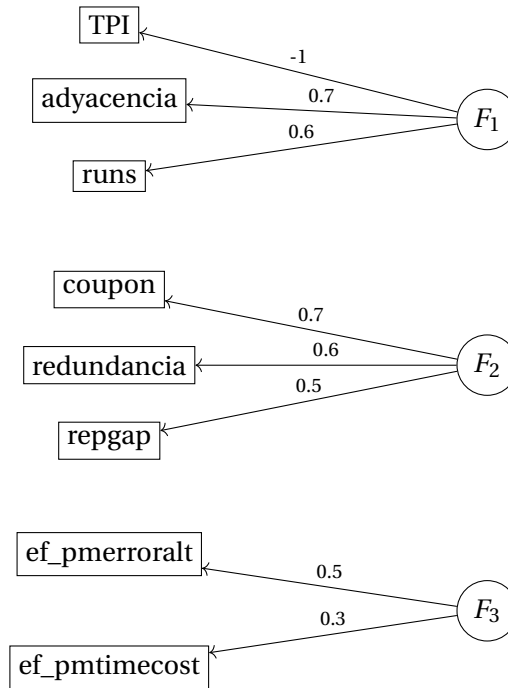


Figura 5.1: Variables asociadas a cada uno de los factores

Las cargas factoriales escritas en forma matricial son dadas en la matriz 5.1, de acuerdo con la teoría descrita en la Sección 3.1 de análisis factorial. Así, de la Tabla 5.6 y de la Figura 5.1 se tiene la matriz de las cargas factoriales:

$$\begin{bmatrix}
 -1 & -0,017 & 0,029 \\
 \mathbf{0,677} & -0,149 & 0,221 \\
 \mathbf{0,598} & -0,079 & 0,040 \\
 0,005 & \mathbf{0,740} & 0,086 \\
 -0,042 & \mathbf{0,652} & -0,017 \\
 0,055 & \mathbf{0,490} & -0,081 \\
 0,0140 & 0,024 & \mathbf{0,295} \\
 0,129 & 0,009 & \mathbf{0,467}
 \end{bmatrix}
 \begin{bmatrix}
 F_1 \\
 F_2 \\
 F_3
 \end{bmatrix}. \tag{5.1}$$

Cabe recordar que al tener las variables asociadas con los factores, se procede a dar un significado apropiado de acuerdo con las variables que los conforman. Afortunadamente, la literatura ya ha descrito estos 3 factores, que en este caso se tiene apoyo del estudio presentado en Miyake y cols. (2000). Específicamente el factor 1 corresponde a

Inhibición, el factor 2 corresponde a Actualización y finalmente el factor 3 a Alternancia. El significado de cada uno de estos factores fue presentado en el Capítulo 2, donde se les conoce con el nombre de **funciones ejecutivas**. Una vez definidos los factores, a continuación se presentan las descriptivas del conjunto de variables asociadas con cada factor.

En primer lugar se describirán las variables asociadas con inhibición. En la Figura 5.2 se presentan el histograma y el boxplot de estas variables.

De acuerdo con la variable conocida como **TPI** (puntos de inflexión entre secuencias ascendentes o descendentes) se observa que la mayoría de personas obtuvieron puntos de inflexión inferiores al 100% en las secuencias de números, con una media de 82,87%, además se concluye que pocos participantes poseen un número de **TPI** superior al 100%, es decir pocos individuos lograron hacer demasiados puntos de inflexión. Esta población fue muy heterogénea en esta tarea, observando que hubo dificultad para realizar bastantes cambios en la secuencia de números. En el Apéndice D se presenta la fórmula del TPI, es el valor observado entre el valor esperado y este resultado se multiplica por 100. Se obtienen resultados mayores que 100 cuando el valor observado es mayor que el valor esperado.

Adyacencia, esta medida busca que los participantes no respondan pares de números consecutivos. En el Apéndice D se presenta la fórmula, esta medida se calcula como el número de pares adyacentes entre el número de pares de respuesta y este resultado se multiplica por 100, donde este resultado oscila entre 0% (no hay pares vecinos) y 100% (compuesto en totalidad por pares vecinos). Se puede observar en el histograma que las respuestas de los individuos oscilan entre 0% y 70%. También solo unos pocos individuos alcanzaron valores mayores a 60, es decir, estas personas lograron no decir muchos pares de números consecutivos.

Runs, hace referencia a la varianza de la secuencia cuando se presenta un **TPI**, en el boxplot se presentan varios puntos atípicos, es decir, algunos participantes lograron hacer muchos cambios en sus respuestas respecto a las secuencias. Se puede apreciar una desviación estándar baja en esta variable esto se tiene ya que las respuestas de estos individuos no se alejan mucho de su valor promedio cuando se consideran valores inferiores a 3.

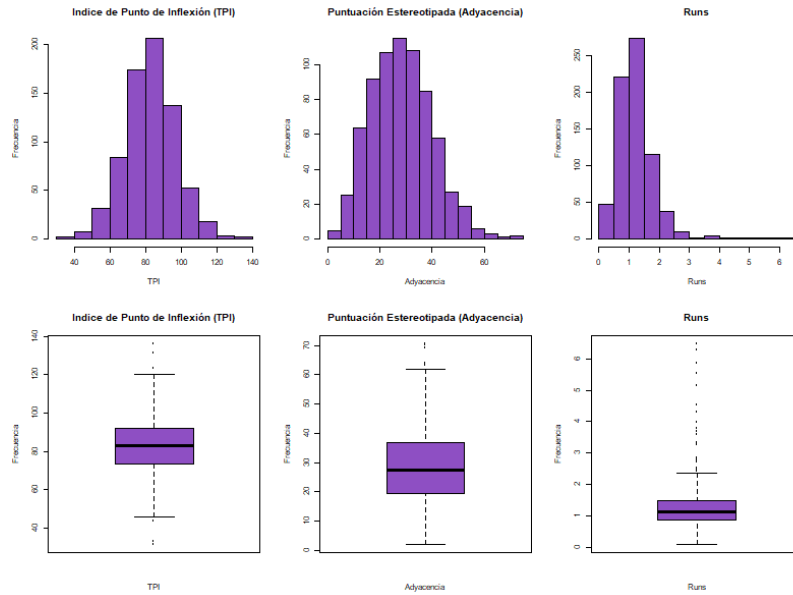


Figura 5.2: Histograma y Boxplot de las variables asociadas con Inhibición

En segundo lugar se describirán las variables asociadas con actualización. En la Figura 5.3 se puede apreciar el comportamiento de las variables asociadas con este factor.

Coupon, esta medida utiliza todas las alternativas de respuesta disponibles, el objetivo de esta medida es tomar números aleatorios, en el histograma se observa que solo pocos participantes obtuvieron una medida de coupon alta mayor a 60, lo que indica que fueron capaces de de aleatorizar secuencias.

Redundancia, en el Apéndice D se presenta la fórmula de esta medida y cómo se hallan los valores de redundancia, donde 0% indica que no se utiliza la misma opción de respuesta en la secuencia, no se repite un mismo número y el 100% se utiliza la misma respuesta en la secuencia, se repite un mismo número. Se observa en el histograma que la media de Redundancia está próxima a 2, lo que indica que gran parte de los individuos al dar sus respuestas no repitieron un mismo número varias veces, pero también se puede apreciar puntos atípicos, es decir, algunos de ellos fueron redundantes al nombrar algún número varias veces.

Repgap, esta variable es la distancia de repetición de números en la secuencia. Se puede observar que la mayoría de individuos se concentran entre 0 y 2, en el boxplot se nota que pocos individuos obtuvieron un valor alto de distancia de repetición.

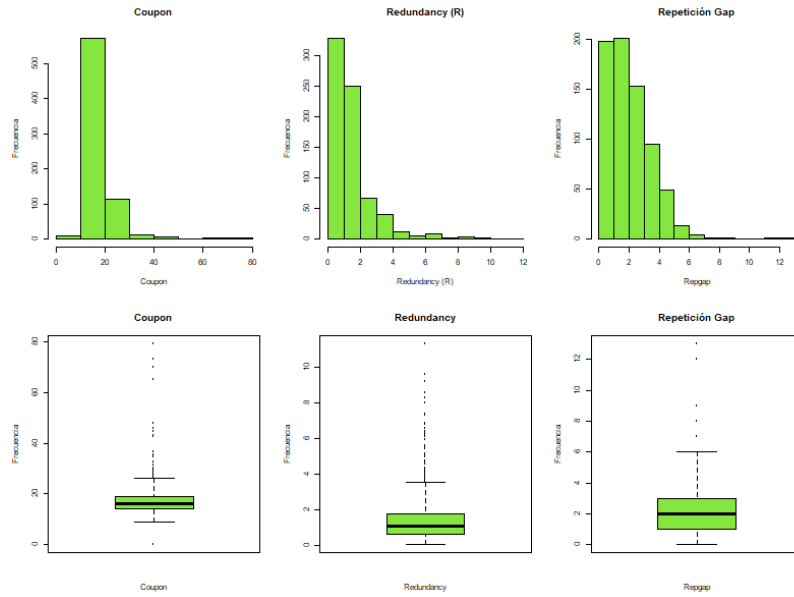


Figura 5.3: Histograma y Boxplot de las variables asociadas con Actualización

Por último, en la Figura 5.4 se puede apreciar el comportamiento de las variables asociadas con Alternancia, que en este caso son apenas dos variables.

Se puede observar en el histograma de la variable **ef_pmerrorsalt** que muchos de los individuos obtuvieron pocos errores de alternancia se puede observar que los valores de la mayoría de individuos están entre 0 y 10 errores, también se puede apreciar del boxplot varios puntos atípicos donde se ve que un participante tuvo 29 errores de 29, es decir, esta persona no cumplió con el *test* de alternar entre más y menos. También de la variable **ef_pmtimecost** mide el tiempo que demora en decir la secuencia, se puede ver muchos puntos atípicos, por tanto el tiempo de respuesta de muchos participantes se salió del promedio.

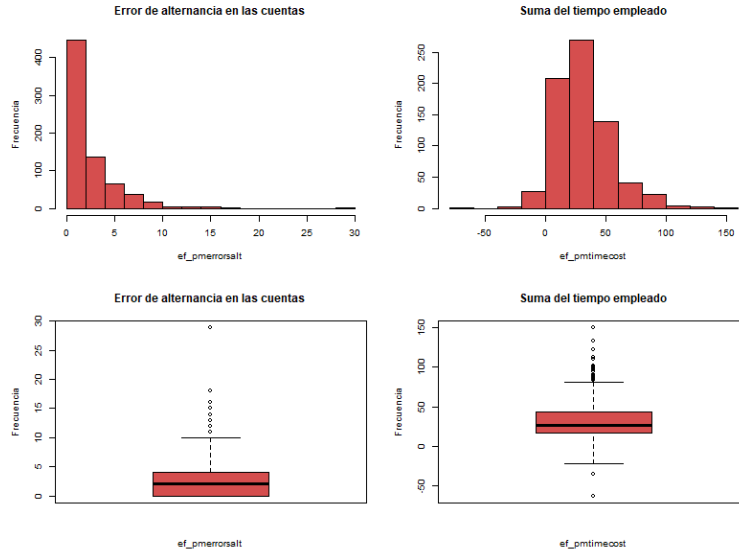


Figura 5.4: Histograma y Boxplot de las variables asociadas con Alternancia

5.3. Modelo de ecuaciones estructurales que incluye correlaciones entre las funciones ejecutivas

En esta sección se plantea el modelo sugerido en la literatura al conjunto de datos del proyecto Baependi. Cabe recordar que lo que se busca es emplear el modelo de Miyake y cols. (2000) que fue realizado a 137 estudiantes en la Universidad de Colorado (Boulder), adaptándolo a la población de Baependi para ver la correlación entre las funciones ejecutivas.

El diagrama de caminos de la Figura 5.5 representa el modelo planteado a partir de las variables en estudio. Como fue ilustrado en el análisis factorial se tienen los tres factores que corresponden a las funciones ejecutivas representadas mediante los círculos (variables latentes) y los rectángulos representan las tareas individuales (variables observadas). Las flechas curvas de doble sentido representan las correlaciones entre las variables latentes. Ambos modelos representan las tres variables latentes, a saber, Alternancia, Actualización e Inhibición. Se observa que todos estos tres factores están correlacionados.

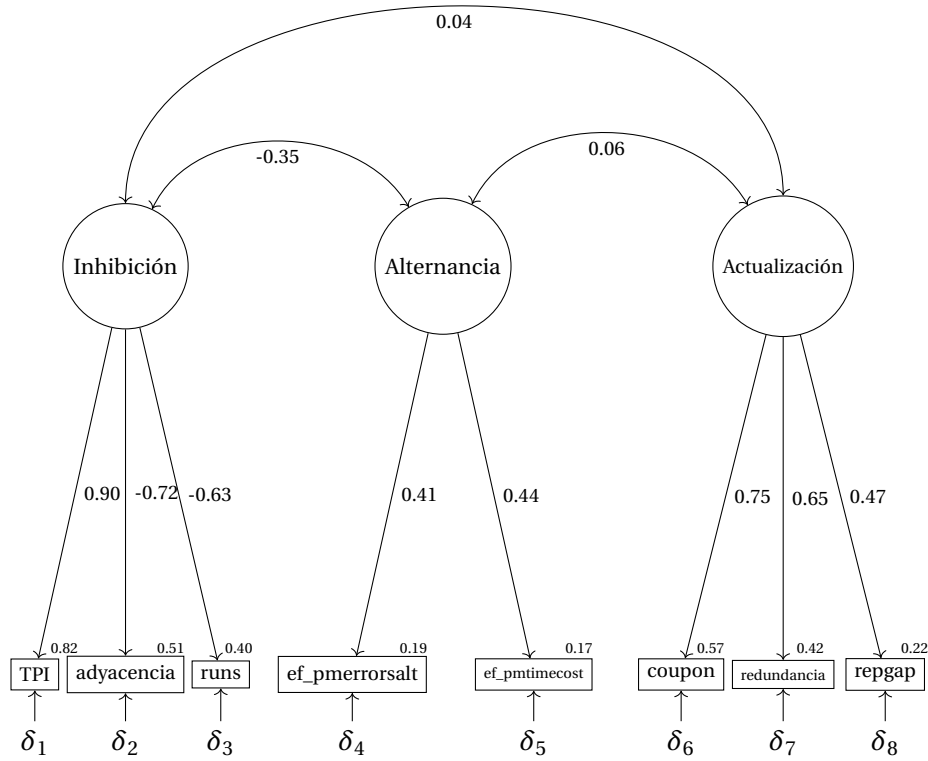


Figura 5.5: Diagrama de caminos obtenido para los datos del proyecto de Baependi

Específicamente se puede ver en la Figura 5.5 la correlación negativa de -0,35 y significativa (valor p de la prueba estadística < 0,01), de las variables latentes Inhibición y Alternancia. Esto puede ocurrir ya que si falla la Alternancia falla la Inhibición y viceversa; la Alternancia necesita cierto grado de Inhibición, pero si un individuo se inhibe por completo no podría alternar entre diferentes tareas. La correlación entre Alternancia y Actualización es baja de 0,06 y no es significativa con un valor p de la prueba estadística es igual a 0,464, esto pudo ocurrir debido a la baja escolaridad de los individuos que realizaron estas pruebas y que tal vez no están acostumbrados a realizar operaciones con secuencias numéricas. También la correlación entre Inhibición y Actualización es baja con 0,04 no significativa su valor p de la prueba estadística es igual a 0,458.

Para obtener las estimaciones del modelo, representado mediante el diagrama de caminos, en la Figura 5.5, se dejan aquellas variables cuya carga factorial en valores estimados sea mayor a 0,4. Tomando los resultados obtenidos por en el programa AMOS, se obtienen las estimaciones de los parámetros en las Tablas 5.7 y 5.8.

En la la Tabla 5.7 se muestran las flechas con una sola dirección, son las cargas factoriales estandarizadas. Las cargas factoriales son todas significativas con un valor p de la prueba estadística $< 0,05$, son equivalentes a los coeficientes de regresión estandarizados (pesos beta) estimados por el método de máxima verosimilitud.

Tabla 5.7: Estimación de parámetros

	Estándar	Estimado	Error	valor p
ef_pmttimecost ← Alternancia	0,439	1,000		
ef_pmerrorsalt ← Alternancia	0,409	0,123	0,043	0,004
TPI ← Inhibición	0,903	1,000		
runs ← Inhibición	-0,633	-0,033	0,002	* * *
adyacencia ← Inhibición	-0,717	-0,660	0,41	* * *
coupon ← Actualización	0,752	1,000		
redundancia ← Actualización	0,650	0,179	0,021	* * *
repgap ← Actualización	0,468	0,149	0,018	* * *

En la Tabla 5.8 se pueden observar las estimaciones de las correlaciones entre las tres funciones ejecutivas. Las flechas bidireccionales tienen coeficientes de correlación junto a ellas que indican correlaciones de las variables latentes, para este modelo sólo se encontró significativa la correlación de las variables latentes Inhibición y Alternancia como fue dicho anteriormente y mostrado en la Figura 5.5.

Tabla 5.8: Estimación de parámetros- covarianza

	Estándar	Estimado	Error	valor p
Inhibición ↔ Alternancia	-0,347	-45,196	12,003	* * *
Actualización ↔ Alternancia	0,060	2,995	4,095	0,464
Actualización ↔ Inhibición	0,036	2,305	3,104	0,458

En la Tabla 5.9 se muestran los índices de bondad de ajuste presentados en la Sección 4.10.1 como la Prueba χ^2 , CFI, IFI, NFI, I y RMSA. También se puede observar: **NPAR** es el número de parámetros distintos a estimar, **CMIN** es el estadístico χ^2 , **DF** son los grados de libertad y **CMIN/DF** es la razón de χ^2 sobre los grados de libertad, se emplea para juzgar si el valor del estadístico χ^2 es grande o pequeño.

Una vez obtenidas las estimaciones de los parámetros del modelo presentado en el diagrama de caminos de la Figura 5.5, es necesario evaluar la bondad ajuste del modelo. Las

medidas de bondad ajuste analizadas son el estadístico χ^2 y el Error de Aproximación Cuadrático Medio (RMSEA). El estadístico χ^2 refleja si el modelo se ajusta a los datos. En la Tabla 5.9 se observa un valor p de la prueba estadística $< 0,05$ en el estadístico χ^2 , en la Sección 4.10.1 se presentaron las características de la Prueba χ^2 , donde la muestra es sensible a tamaños mayores de 200 observaciones, se está rechazando la hipótesis nula. En relación con lo anterior se deben considerar otros índices de bondad de ajuste para determinar la adecuación del modelo.

Tabla 5.9: Índices de bondad de ajuste

Modelo	NPAR	CMIN	DF	p-valor	CMIN/DF
Base	27	69,993	17	0,000	4,117
Saturado	44	0,000	0		
Independiente	16	11114,336	28	0,000	39,798
	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Base	0,937	0,897	0,952	0,920	0,951
Saturado	1,000		1,000		1,000
Independiente	0,000	0,000	0,000	0,000	0,000
	RMSEA	LO 90	HI 90	PCLOSE	
Base	0,066	0,050	0,082	0,046	
Independiente	0,233	0,221	0,245	0,000	

También se emplea el indicador **RMSEA**, el cual no es sensible al tamaño muestral e indica si el ajuste del modelo es bueno. Valores inferiores a 0,08 indican buen ajuste, se obtiene un RMSEA=0,066, lo cual es aceptable. Por otra parte, las medidas de ajuste incremental son: NFI, IFI y CFI. El índice NFI mide la reducción proporcional del estadístico χ^2 . En el modelo planteado en este estudio el NFI=0,937, lo cual es aceptable ya que alcanza valores cercanos a 1. El CFI=0,951 lo que indica un ajuste razonable debido a que su valor es cercano a 1. Así mismo, el Índice de Ajuste Incremental (IFI) compara el ajuste de dos o más modelos, lo cual en este caso el IFI=0,952, indicando un ajuste aceptable del mismo.

Como conclusión de este resultado, se puede afirmar que no fue del todo satisfactorio para los investigadores del área. Se esperaban correlaciones significativas entre todas las funciones ejecutivas (Inhibición, Alternancia y Actualización). En el estudio presentado en Miyake y cols. (2000) esas correlaciones fueron significativas, posiblemente debido a que los individuos a quienes se les aplicaron los instrumentos de medida eran estudiantes universitarios, por tanto una población homogénea pero con un tamaño muestral bajo (137). En contrapartida, en la población de Baependi se tomó una muestra grande (717) pero es una población con un nivel bajo de escolaridad y en general muy heterogénea, lo cual pudo haber influenciado de forma contundente los resultados del estudio, a pesar de un primer esfuerzo de la adaptación de los instrumentos de medida por parte de los expertos que llevan a cabo el estudio.

Se resalta el hecho de que los resultados presentados en esta sección acompañan la fase inicial del estudio de Cognición y han permitido descubrir fallas en los instrumentos de medida (aplicación de pruebas o tareas), sin embargo pueden ayudar a direccionar el rediseño de las pruebas o tareas de tal modo que midan, de manera más eficiente, las funciones ejecutivas en poblaciones con las características de Baependi. Se espera que con la nueva batería de pruebas se incluyan las variables de edad y escolaridad en el modelo de ecuaciones estructurales. Un acercamiento a ese modelo, con las pruebas o tareas actuales, se presenta en el Apéndice E.

El resultado anterior fue replicado a partir de lo propuesto en la literatura, sin embargo se están estudiando otras propuestas de relaciones entre las variables latentes o funciones ejecutivas. Una de estas propuestas es mostrada en la siguiente sección, la cual ilustra un modelo de ecuaciones estructurales completo y que ayuda a aplicar la teoría del Capítulo 4.

5.4. Modelo alternativo de ecuaciones estructurales: Relaciones de causalidad entre las funciones ejecutivas

En esta sección se presentará el modelo de ecuaciones estructurales aplicando la teoría vista en el Capítulo 4, tomando el conjunto de datos del proyecto de Baependi.

En primer lugar, se representa el diagrama de caminos en la Figura 5.6 donde se establecieron las relaciones causales entre las variables observables y variables latentes. De donde se tienen dos variables latentes exógenas que son Alternancia e Inhibición y una variable endógena Actualización, se tienen 8 variables observadas las cuales son: **TPI**, **adyacencia**, **runs**, **coupon**, **redundancia**, **repgap**, **ep_pmerrorsalt** y **ep_pmtimecost**. El diagrama de caminos, con todos los indicadores y errores se presenta en la Figura 5.7.

En el modelo propuesto, que a la vez se presenta en el diagrama de caminos de la Figura 5.6 se definen las siguientes hipótesis estructurales:

- H₁**: La Alternancia e Inhibición están significativamente correlacionadas.
- H₂**: La Alternancia tiene un efecto directo y positivo sobre la Actualización.
- H₃**: La Inhibición tiene un efecto directo y positivo sobre la Actualización.

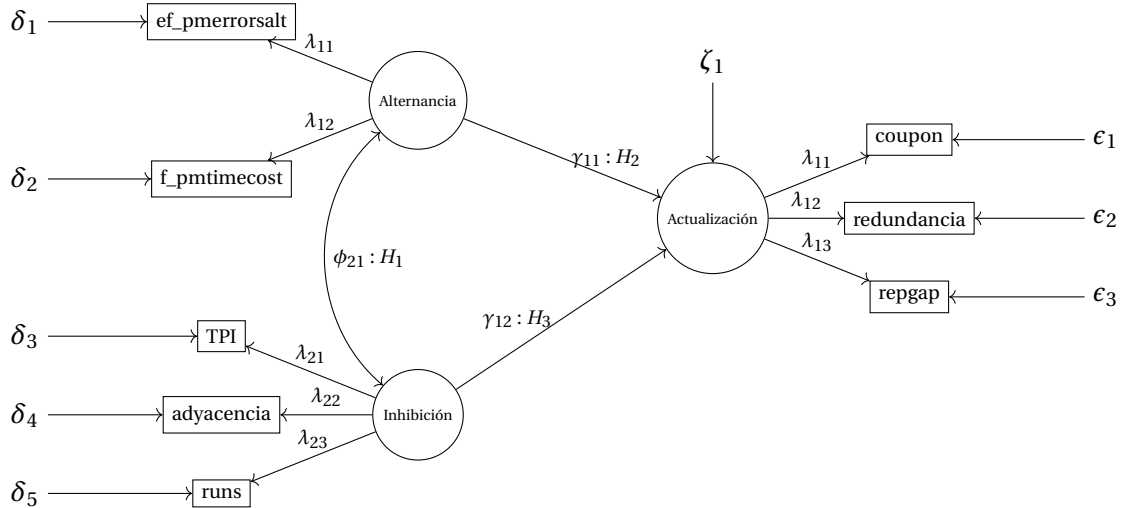


Figura 5.6: Diagrama de caminos modelo alternativo

El modelo presentado en el diagrama de caminos (Figura 5.7), se observan las estimaciones de los parámetros planteados en el modelo representado en el diagrama de caminos de la Figura 5.6, donde se empleó el método de Máxima Verosimilitud utilizando el programa AMOS versión 23 para estas estimaciones.

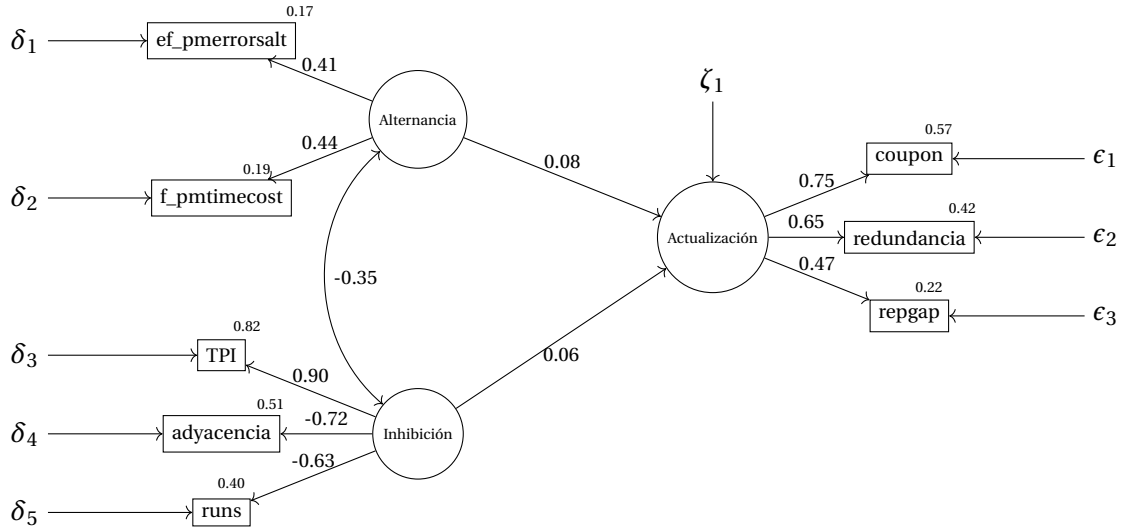


Figura 5.7: Diagrama de caminos modelo alternativo: Estimación de los parámetros

Del anterior diagrama de caminos de la Figura 5.7 se tiene el modelo estructural dado por las ecuaciones:

$$[Actualización] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} [Actualización] + \begin{bmatrix} 0,08 & 0,06 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Alternancia \\ Inhibición \end{bmatrix} + \begin{bmatrix} \zeta \\ 0 \end{bmatrix}. \quad (5.2)$$

Su respectivo modelo de medida de la siguiente forma:

$$\begin{bmatrix} ef_pmerrorsalt \\ ef_pmtimecost \\ TPI \\ adyacencia \\ runs \end{bmatrix} = \begin{bmatrix} 0,41 & 0 \\ 0,44 & 0 \\ 0 & 0,090 \\ 0 & -0,72 \\ 0 & -0,63 \end{bmatrix} \begin{bmatrix} Alternancia \\ Inhibición \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix} \quad (5.3)$$

y

$$\begin{bmatrix} coupon \\ redundancia \\ repgap \end{bmatrix} = \begin{bmatrix} 0,75 \\ 0,65 \\ 0,47 \end{bmatrix} [Actualización] + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}. \quad (5.4)$$

Las matrices de covarianza de los errores son:

$$\Theta_{\delta} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \Theta_{\epsilon} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Por lo tanto se tiene el respectivo modelo estructural, de medida y las matrices de covarianza de los errores del modelo alternativo de ecuaciones estructurales

5.4.1. Identificación del modelo

Para la identificación se utilizará la regla de los grados de libertad, la teoría se puede ver en la Sección 4.7. Consiste en calcular el número de elementos no redundantes en la matriz de covarianzas de las variables observadas el cual debe ser mayor que el número de parámetros desconocidos en θ .

Esta regla permite conocer cuándo un modelo no es identificado, aunque no permite discernir si es sobre identificado o no, así que adicionalmente al cálculo de la regla t se debe tener en cuenta que el valor de los grados de libertad (gl) pueden ayudar a discernir el tipo de modelo de la siguiente manera (Romero, 2015):

1. $gl < 0$: Modelos nunca identificados, los parámetros podrían tomar infinitos valores, razón por la cual son indeterminados.
2. $gl = 0$: Posiblemente identificados, modelos en los que puede existir una única solución para los parámetros, que ajuste la matriz Σ .
3. $gl > 0$: Posiblemente sobre identificados, modelos que incluyen menos parámetros que varianzas y covarianzas tiene Σ .

El modelo ilustrado en el diagrama de caminos de la Figura 5.7, es un modelo sobreidentificado, ya que se tiene $p = 3$, variables observadas de la variables endógena Actualización y $q = 5$ variables observadas de las variable exógenas Inhibición y Alternancia por lo tanto analizando los grados de libertad se tiene:

$$\frac{(p+q)(p+q+1)}{2} = \frac{(3+5)(3+5+1)}{2} = \frac{(3+5)(3+5+1)}{2} = 36.$$

A primera vista y asumiendo que todas las variables observadas no son correlacionadas, el modelo se encontraría sobreidentificado, ya que analizando los grados de libertad se tienen $gl = 36$. Sin embargo, no existe una prueba estadística para realizar la identificación del modelo, así como tampoco existe una condición necesaria, suficiente y general para hacerlo, aunque existen ciertas normas según el tipo de modelo estructural que se esté tratando.

Dado que este modelo contiene variables latentes, el cumplimiento de las siguientes condiciones simultáneas, son suficientes para aceptar el modelo como identificado (Bollen y Davis, 2009):

- Cada variable latente tiene al menos dos variables observadas que se relacionan con una sola variable latente.
- Cada variable latente tiene al menos una variable observada con un efecto directo distinto de cero que se usa para fijar la escala de la variable latente.

La Tabla 5.10, muestra las estimaciones de los parámetros, tanto estándar, como no estándar, para todas las relaciones causales del modelo. Además, siempre aparece el valor p de la prueba estadística asociado a cada parámetro. En la notación empleada, cuando aparecen (* * *) significa el valor p de la prueba estadística $< 0,01$ que es lo deseado.

Tabla 5.10: Estimación de parámetros SEM

	Estándar	Estimado	Error	valor p
Actualización \leftarrow Inhibición	0,064	0,025	0,024	0,294
Actualización \leftarrow Alternancia	0,082	0,040	0,047	0,387
ef_pmtimecost \leftarrow Alternancia	0,439	1,000		
ef_pmerrorsalt \leftarrow Alternancia	0,409	0,123	0,043	0,004
TPI \leftarrow Inhibición	0,903	1,000		* * *
adyacencia \leftarrow Inhibición	-0,717	-0,660	0,042	* * *
runs \leftarrow Inhibición	-0,633	-0,033	0,002	* * *
coupon \leftarrow Actualización	0,752	1,000		
redundancia \leftarrow Actualización	0,650	0,179	0,021	* * *
repgap \leftarrow Actualización	0,468	0,149	0,018	* * *

En la Tabla 5.11 muestra la correlación de las variables Alternancia e Inhibición, tanto estándar, como no estándar, el error y el valor p de la prueba estadística.

Tabla 5.11: Estimación de parámetros- covarianza-SEM

	Estándar	Estimado	Error	valor p
Alternancia \leftrightarrow Inhibición	-0,347	-45,196	12,104	* * *

Teniendo en cuenta los resultados obtenidos y presentados en el diagrama de caminos de la Figura 5.7 y con los resultados de la Tabla 5.11 se puede ver la correlación negativa de $\phi_{21} = -0,35$ y significativa (valor p de la prueba estadística $< 0,01$), por tanto no se rechaza la hipótesis H_1 , es decir que las variables latentes exógenas Inhibición y Alternancia están significativamente correlacionadas.

Es posible identificar que no hay un efecto directo del factor latente de Inhibición en el factor Actualización ya que $\gamma_{11} = 0,06$; valor p de la prueba estadística = 0,294, se rechaza la hipótesis H_2 , también no hay un efecto directo del factor latente de Inhibición en el factor Actualización ya que $\gamma_{12} = 0,08$; el valor p de la prueba estadística es igual a 0,387, se rechaza la hipótesis H_3 .

En la Tabla 5.12 se muestran los índices de bondad de ajuste del modelo de ecuaciones estructurales de la Figura 5.7. El estadístico $\chi^2(17, N = 717) = 69,993$, el valor p de la prueba estadística $< 0,001$, por tanto se rechaza la hipótesis $H_0 = \Sigma(\theta)$ que el modelo especificado sí reproduce la estructura de varianzas y covarianzas en la población; en la Sección 4.10.1 se comenta que este estadístico es sensible a muestras grandes, casi siempre rechazando la hipótesis nula, por lo cual se hace necesario tener en cuenta los índices de bondad $CFI = 0,951 > 0,9$, $RMSEA = 0,066 < 0,08$, $NFI = 0,937 > 0,9$ y $IFI = 0,920 > 0,9$, donde se puede ver que todos son aceptables, lo cual se concluye que el modelo tiene un ajuste satisfactorio.

Tabla 5.12: Índices de bondad de ajuste

Modelo	NPAR	CMIN	DF	p-valor	CMIN/DF
Base	27	69,993	17	0,000	4,117
Saturado	44	0,000	0		
Independiente	16	11114,336	28	0,000	39,798
	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Base	0,937	0,897	0,952	0,920	0,951
Saturado	1,000		1,000		1,000
Independiente	0,000	0,000	0,000	0,000	0,000
	RMSEA	LO 90	HI 90	PCLOSE	
Base	0,066	0,050	0,082	0,046	
Independiente	0,233	0,221	0,245	0,000	

En conclusión, el modelo planteado tuvo un ajuste satisfactorio donde se observó que las variables Inhibición y Alternancia tienen una correlación negativa estadísticamente significativa, también se pudo constatar que las relaciones causales entre las variables Inhibición y Actualización así como Alternancia y Actualización son muy bajas. Por ahora el planteamiento de un modelo con relaciones causales es apenas una sugerencia para abordar el problema desde otro punto de vista de causalidad, no apenas considerando la correlación entre las funciones ejecutivas, que es el foco principal de la investigación.

Capítulo 6

Conclusiones

En este trabajo de tesis de maestría se logró la formalización matemática de la teoría estadística asociada con análisis factorial y ecuaciones estructurales. Esta teoría fue aplicada a los datos sobre funciones ejecutivas del Proyecto “Corazones de Baependi”. Con los hallazgos de este trabajo los investigadores de dicho proyecto tuvieron la formulación de una teoría estadística completa que les permitirá desarrollar nuevas formas de medida apropiadas para la población de Baependi, en relación a detectar asociaciones con las funciones ejecutivas. Además les permitirá continuar con varios frentes de este estudio que se desprenden de los análisis de modelos de ecuaciones estructurales.

En el primer Capítulo se presentó el siguiente objetivo general:

“Formalizar matemáticamente la teoría asociada a modelos de ecuaciones estructurales estableciendo ejemplos ilustrativos de la teoría y aplicaciones a datos reales que involucren variables del campo de la Psicología”.

Para lograrlo se plantearon 4 objetivos específicos que a continuación se describe cómo se alcanzaron:

1. La teoría de análisis factorial y de ecuaciones estructurales se fundamentó matemáticamente y se presenta de forma ordenada y sistemática en los Capítulos 3 y 4 de este documento. Particularmente en el Capítulo 3 se compiló y se presentó de forma sistemática, desde las diversas fuentes, una sustentación matemática del método multivariado de análisis factorial. En el Capítulo 4 se recopiló y ordenó la teoría asociada a ecuaciones estructurales. Además, se completaron algunas demostraciones presentadas en la literatura. Lo anterior permite tener una amplia visión de la robustez de esta herramienta estadística para ser aplicada teniendo en cuenta la formalización matemática que, como se anotó, falta en muchas aplicaciones de diferentes áreas. Desde la matemática se tiene un fundamento teórico necesario, pero al ser metodologías estadísticas asociadas a diversas aplicaciones, se requiere profundizar más en el área para lograr casos particulares que se puedan extender y asociar a otras teorías.

2. La aplicación de la teoría estadística fundamentada matemáticamente es la medición de las tareas o pruebas para probar asociación y correlación de estas tareas con las funciones ejecutivas. A este respecto, en el Capítulo 2, se presentaron los conceptos fundamentales de las funciones ejecutivas desde la Psicología, de manera tal que se logre entender cómo se diseñan estas pruebas con el objetivo de encontrar relaciones con áreas de Cognición, específicamente con las funciones ejecutivas. Además, se hace referencia al comportamiento esperado de acuerdo a estudios previos (vale destacar que existen varios tipos de pruebas para medir el desempeño de las funciones ejecutivas en los individuos).
3. Para la aplicación se usaron los paquetes *lavaan* del Programa R y paralelamente el Programa SPSS, específicamente el Paquete AMOS. Los códigos usados y el paso a paso de cómo se desarrolla, se muestra en el Apéndice C. Si bien es una teoría que ya se ha trabajado en la estadística y se han utilizado estas herramientas, la extensión a cualquier tipo de aplicación no es muy clara y por ello se presentó el Apéndice mencionado.
4. Como se trató en el Capítulo 1 y en la primera parte del Capítulo 5, el Proyecto Baependi recopila información genética, datos de familia, datos de enfermedades cardiovasculares y datos psicológicos de población heterogénea en todos los ámbitos (raza, edad, escolaridad, etc). La aplicación que se hizo en este trabajo de investigación corresponde a una primera aproximación de datos heterogéneos para funciones ejecutivas. Al aplicar la teoría se encontraron los siguientes hallazgos:
 - Existe una correlación significativa entre Inhibición y Alternancia que corresponde a $-0,35$.
 - No existen correlaciones significativas entre las otras funciones ejecutivas, a saber, entre Inhibición y Actualización así como entre Alternancia y Actualización.
 - Al comparar los estudios se encuentran diferencias marcadas entre la homogeneidad de la población y el tamaño muestral.
5. Como se mencionó, esta primera aproximación buscaba mejorar las pruebas para medir las funciones ejecutivas en poblaciones heterogéneas, pues las pruebas existentes en la literatura son de poblaciones homogéneas en todo sentido a diferencia de la población del Proyecto Baependi. Desde la teoría cognitiva se sabe que las funciones ejecutivas se correlacionan significativamente. Ante los hallazgos encontrados, los investigadores del Proyecto Baependi, decidieron modificar las pruebas y volverlas a aplicar antes de formular las ecuaciones estructurales nuevamente incluyendo muchas más variables. De cualquier modo en el Capítulo 5, Sección 5.4, se presenta una aproximación que envuelve relaciones causales entre las funciones ejecutivas lo cual ayudó a visualizar otra forma de análisis, además de ilustrar un modelo de ecuaciones estructurales completo. También en el Apéndice E se presentó un diagrama de caminos con las variables de edad y escolaridad como una primera visualización de lo que se pretende realizar con las nuevas tareas que el grupo de investigación formulará.

6. En el Apéndice F se presenta el diagrama de caminos de algunos modelos propuestos, donde de forma breve se analizan los resultados explicando el por qué no contribuyen en el desarrollo de la investigación, pero que ayudaron a entender mejor el problema y fueron fundamentales para la toma de decisiones y realizar las conclusiones de este trabajo.

De acuerdo al trabajo elaborado se realizan las siguientes recomendaciones para dar continuidad:

- Implementar nuevas tareas donde se incluyan las variables de edad y escolaridad en la población de Baependi.
- Como un trabajo complementario se trata de incorporar la teoría vista con los modelos mixtos para encontrar el coeficiente de correlación intraclase, que el área de genética se conoce como heredabilidad, la cual explica a variación fenotípica observable debida a efectos genéticos y tal vez encontrar genes asociados con cognición.
- Tener en cuenta las variables latentes como depresión y ansiedad, entre otras variables asociadas con cognición para establecer relaciones de causalidad con las funciones ejecutivas.

Apéndice A

Álgebra de matrices

A.1. Matriz traspuesta y propiedades

Sea A una matriz de dimensión $m \times n$, se denota al elemento de la fila i y columna j como $A_{i,j}$, siendo $i < m$ y $j < n$, entonces, se define la matriz traspuesta de A como la matriz A' de dimensión $n \times m$ tal que $A'_{j,i} = A_{i,j}$

1. Traspuesta de la traspuesta

$$(A')' = A \tag{A.1}$$

2. Traspuesta de la suma

$$(A + B)' = A' + B' \tag{A.2}$$

3. Traspuesta del producto de un escalar por una matriz

$$(\alpha A)' = \alpha A' \tag{A.3}$$

4. Traspuesta del producto

$$(AB)' = B' A' \tag{A.4}$$

5. Una matriz es igual que su traspuesta, si y solo si, es simétrica

$$A \text{ simétrica si y sólo si } A = A'. \tag{A.5}$$

A.2. Derivación matricial

El cálculo matricial permite una notación más compacta a la hora de realizar cálculo multivariado, en el cual las funciones pueden ser de un tipo más amplio, ya que la variable puede ser vectorial o matricial.

Sea x un escalar, $A_{n \times r}$ y $B_{p \times q}$ matrices, $C_{n \times n}$ y $D_{n \times n}$ matrices cuadradas no singulares, con C constantes. Tenemos las siguientes propiedades (Mulaik, 2009a):

1. **Propiedad 1:** La deriva de A con respecto a x es una matriz cuyos elementos son

$$\forall (i, j) \in \{1, \dots, n\} \times \{1, \dots, r\}. \quad \frac{\partial}{\partial x} A = \left[\frac{\partial}{\partial x} a_{i,j} \right]_{n \times r} \quad (\text{A.6})$$

2. **Propiedad 2:** Derivada del producto de dos matrices $A_{n \times r}$ y $B_{p \times q}$ en relación a x

$$\frac{\partial}{\partial x} (A + B) = \frac{\partial}{\partial x} A + \frac{\partial}{\partial x} B \quad (\text{A.7})$$

3. **Propiedad 3:** Derivada de la suma de dos matrices $A_{n \times r}$ y $B_{p \times q}$ con respecto a x . Si $r = p$ entonces

$$\frac{\partial}{\partial x} (AB) = \frac{\partial A}{\partial x} B + \frac{\partial B}{\partial x} A \quad (\text{A.8})$$

4. **Propiedad 4:** La derivada de la traza de producto de una matriz rectangular por una matriz cuadrada, constante y no singular con respecto a x :

$$\frac{\partial}{\partial x} \text{tr}(AC) = \left[\text{tr} \left(\frac{\partial}{\partial x} A \right) \right] C \quad (\text{A.9})$$

5. **Propiedad 5:** La derivada de matrices cuadradas no singulares con respecto a x :

$$\frac{\partial}{\partial x} D^{-1} = -D^{-1} \left(\frac{\partial}{\partial x} D \right) D^{-1} \quad (\text{A.10})$$

6. **Propiedad 6:** La derivada de la traza del producto de la inversa de una matriz cuadrada por una matriz constante también cuadrada con respecto a x :

$$\frac{\partial}{\partial x} (D^{-1}C) = -\text{tr} \left(D^{-1} \left(\frac{\partial}{\partial x} D \right) D^{-1}C \right) \quad (\text{A.11})$$

7. **Propiedad 7:** La derivada del logaritmo natural del determinante de una matriz cuadrada no singular en relación con x :

$$\frac{\partial}{\partial x} \ln \det D = -[\text{tr}(D^{-1})] \left(\frac{\partial}{\partial x} D \right). \quad (\text{A.12})$$

Apéndice B

Algoritmos de optimización

Puede resultar útil revisar brevemente varios algoritmos básicos de optimización utilizados en la estimación no lineal. Se recuerda que el problema al que se enfrenta la estimación no lineal es encontrar los valores de los parámetros que optimizan (maximizan o minimizan) una función objetivo, como por ejemplo la función $F(\mathbf{S}, \boldsymbol{\Sigma}(\theta))$ dada en la Sección 4.8.

Los algoritmos de optimización iterativa funcionan tomando un conjunto inicial de valores para los parámetros, $\theta^{(0)}$, y luego realizando cálculos basados en estos valores para obtener un mejor conjunto de valores de los parámetros, $\theta^{(1)}$. Este proceso se repite para $\theta^{(2)}$, $\theta^{(3)}$ y así sucesivamente hasta que la función objetivo deja de mejorar entre iteraciones.

El proceso de optimización consta de tres partes principales:

1. La obtención de los valores iniciales de los parámetros,
2. La actualización del vector de parámetros candidatos en cada iteración,
3. La determinación de cuándo se alcanza el óptimo.

Si la función objetivo es globalmente cóncava, de modo que hay un único máximo, cualquier algoritmo que mejore el vector de parámetros en cada iteración acabará encontrando este máximo (suponiendo que el tamaño de los pasos dados no sea despreciable). Si la función objetivo no es globalmente cóncava, diferentes algoritmos pueden encontrar diferentes máximos locales, pero todos los algoritmos iterativos sufrirán el mismo problema de no poder distinguir un máximo local de uno global.

Lo principal que distingue a los distintos algoritmos es la rapidez con la que encuentran el máximo. Desgraciadamente, no hay reglas fijas y rápidas. Para algunos problemas, un método puede ser más rápido, para otros problemas puede no serlo.

Los algoritmos pueden clasificarse a grandes rasgos en tres tipos: métodos de segunda derivada, métodos de primera derivada y métodos sin derivada. Los métodos de segunda derivada evalúan los valores actuales de los parámetros y las primeras y segundas

derivadas de la función objetivo para cada observación. Los métodos de primera derivada sólo utilizan las primeras derivadas de la función objetivo durante el proceso de iteración. Como su nombre indica, los métodos sin derivadas no calculan las derivadas. A continuación se presentan los métodos de la segunda derivada que son utilizados para estimadores de Máxima Verosimilitud (Everitt, 1984).

B.1. Los métodos de la segunda derivada

Los métodos de segunda derivada para los modelos binarios ordenados, pueden estimar el modelo de ecuaciones estructurales utilizando Newton-Raphson o la *quadratic hill-climbing*.

El problema consiste en minimizar una de las funciones de discrepancia vistas en la Sección 4.8 de acuerdo a los métodos de estimación con respecto a los parámetros libres en el modelo de la investigación (Everitt, 1984).

B.1.1. Método de Newton-Raphson

Los valores candidatos para los parámetros $\theta^{(1)}$ pueden obtenerse utilizando el método de Newton-Raphson mediante la linealización de las condiciones de primer orden $\frac{\partial F}{\partial \theta}$ en los valores actuales de los parámetros, $\theta^{(k)}$:

$$\begin{aligned} g^{(k)} + H^{(k)}(\theta^{(k+1)} - \theta^{(k)}) &= 0 \\ \theta^{(k+1)} &= \theta^{(k)} + (H^{(k)})^{-1} g^{(k)} \end{aligned} \quad (\text{B.1})$$

donde g es el vector gradiente $\frac{\partial F}{\partial \theta}$ y H es la matriz hessiana $\frac{\partial^2 F}{\partial \theta^2}$.

Si la función es cuadrática, Newton-Raphson encontrará el máximo en una sola iteración. Si la función no es cuadrática, el éxito del algoritmo dependerá de lo bien que una aproximación cuadrática local capte la forma de la función.

B.1.2. Método de Quadratic Hill-Climbing (*Goldfeld-Quandt*)

Este método, que es una variación directa de Newton-Raphson, se atribuye a veces a Goldfeld y Quandt. La *Quadratic Hill-Climbing* modifica el algoritmo de Newton-Raphson añadiendo una matriz de corrección al hessiano. El algoritmo de actualización de Quadratic Hill-Climbing viene dado por:

$$\theta^{(k+1)} = \theta^{(k)} + (H^{(k)})^{-1} g^{(k)} \quad \text{donde } -H^{(k)} = -H^{(k)} \alpha I, \quad (\text{B.2})$$

donde I es la matriz de identidad y α es un número positivo que es elegido por el algoritmo.

El efecto de esta modificación es empujar las estimaciones de los parámetros en la dirección del vector gradiente. La idea es que cuando esté lejos del máximo, la aproximación cuadrática local a la función puede ser una mala guía para su forma general, por lo que puede ser mejor seguir simplemente el gradiente. La corrección puede proporcionar un mejor rendimiento en lugares alejados del óptimo y permite calcular el vector de dirección en los casos en que el hessiano es casi singular.

Se debe tener en cuenta que los errores estándar asintóticos siempre se calculan a partir del hessiano no modificado una vez que se alcanza la convergencia.

Apéndice C

Programas y Paquetes Estadísticos

Existen varios paquetes estadísticos para ecuaciones estructurales, con los cuales buscan calcular y analizar las relaciones entre las variables, algunos programas con sus paquetes estadísticos son:

C.1. LISREL:

LISREL (del inglés *Linear Structural Relationships*) es considerado sinónimo de SEM ya que se desarrolló específicamente para el empleo de esta técnica y por ser pionero en su aplicación (Jöreskog y Sörborn, 1986; SPSS, 1990, 1993).

El modelo LISREL, sus métodos y software se han convertido en sinónimo de modelos de ecuaciones estructurales. SEM permite a los investigadores en ciencias sociales, ciencias de gestión, ciencias del comportamiento, ciencias biológicas, ciencias de la educación y otros campos de investigación, evaluar empíricamente sus teorías. Estas teorías suelen formularse como modelos teóricos para variables observadas y latentes. Si se recopilan datos para las variables observadas del modelo teórico, el programa LISREL puede utilizarse para adaptar el modelo a los datos.

Hoy LISREL ya no se limita a SEM, ya que en su última versión incluye las siguientes aplicaciones estadísticas:

1. LISREL para el modelado de ecuaciones estructurales,
2. PRELIS para la manipulación de datos y análisis estadístico básico,
3. MULTILEV para el modelado lineal y no lineal jerárquico,

4. SURVEYGLIM para el modelado lineal generalizado,
5. MAPGLIM para el modelado lineal generalizado para datos multinivel.

LISREL es una aplicación para el modelado de ecuaciones estructurales estándar y multinivel. Estos métodos están disponibles para los datos de encuestas complejas completas e incompletas sobre variables categóricas y continuas, así como datos de muestras aleatorias simples completas e incompletas sobre variables categóricas y continuas.

C.2. AMOS

AMOS: (del inglés *Analysis of Moment Structures*), este programa trabaja bajo la plataforma del SPSS (*Statistical Package for the Social Science*), es un sistema global para el análisis estadístico de datos. AMOS utiliza el modelado de ecuaciones estructurales para confirmar y explicar los modelos conceptuales que tratan las actitudes, percepciones y otros factores que determinan el comportamiento de las variables, es decir, se puede evaluar el ajuste del modelo y realizar cualquier modificación, (Arbuckle, 1997).

A continuación se hace una breve introducción del Programa AMOS.

En la Figura C.1 se muestra la pantalla principal del Programa.

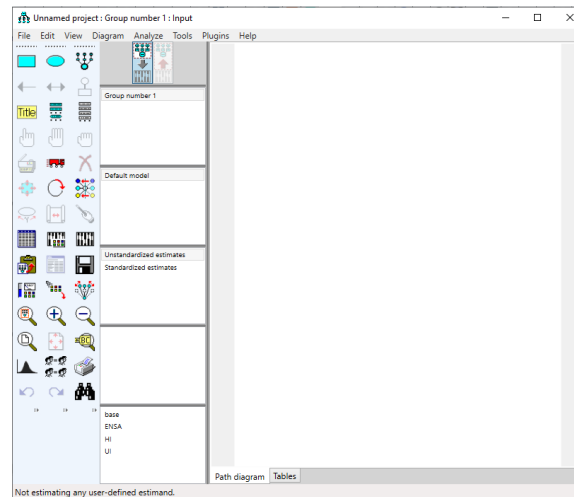


Figura C.1: Panel principal AMOS

El uso de SEM en 5 pasos utilizando la interfaz gráfica de AMOS, se da a continuación:

- Paso 1:** Como muestra la Figura C.2 los rectángulos representan las variables observables, y los círculos las variables latentes. Además, las variables correspondientes a los errores no se representaban por círculos y cuadrados pero al no ser observables, en el programa AMOS estas variables se representan también como un círculo, (estos círculos se representan con un tamaño bastante inferior al de las variables latentes).

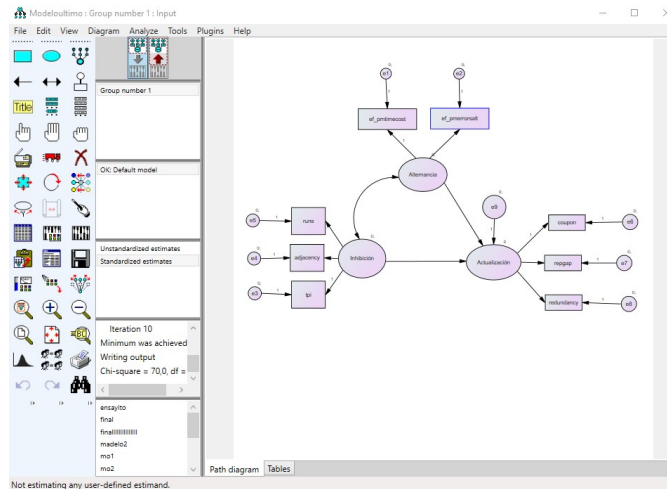


Figura C.2: Modelo de Ecuaciones Estructurales

- Paso2:** Como se ve en la Figura C.3, se presenta el menú en el cual se muestran las figuras de un rectángulo y un círculo que representan las variables, así como las flechas, tanto unidireccionales como bidireccionales, que representan las covarianzas y las regresiones entre las variables, además de otras opciones como puede ser la opción de seleccionar, eliminar, entre otras.

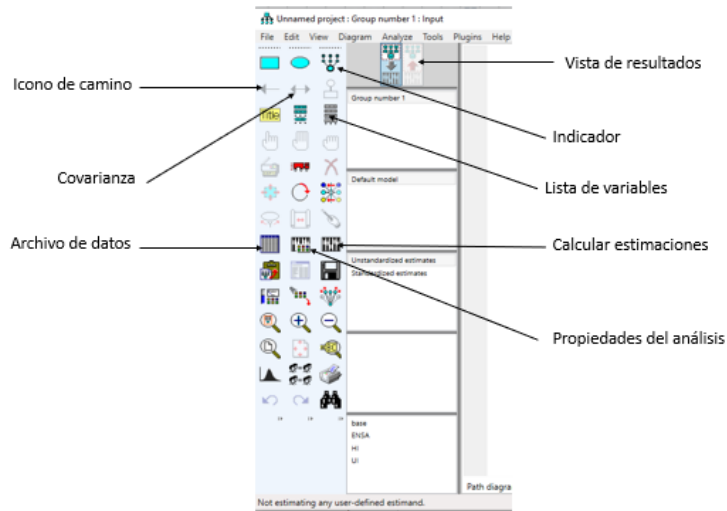


Figura C.3: Iconos clave en AMOS

- **Paso 3:** En la Figura C.4 se observan la variables consideradas en el modelo y en la Figura C.5 se puede cambiar las opciones de cada una de las variables observadas y también latentes. Para ello se hace doble *clik* sobre la variable (ya sea una variable observable o no observable) se muestra una ventana en la cual cambia el texto que muestra la variable, tamaño, el estilo de la fuente, entre otras. Además da la opción de fijar una varianza para la variable (en la pestaña de Parameters).

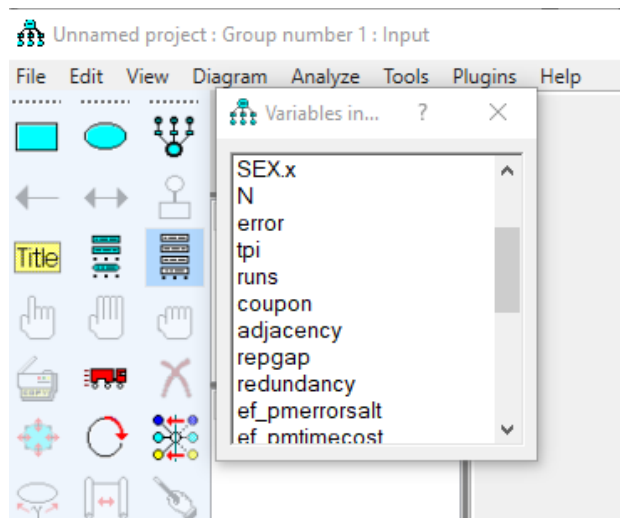


Figura C.4: Menú opciones AMOS

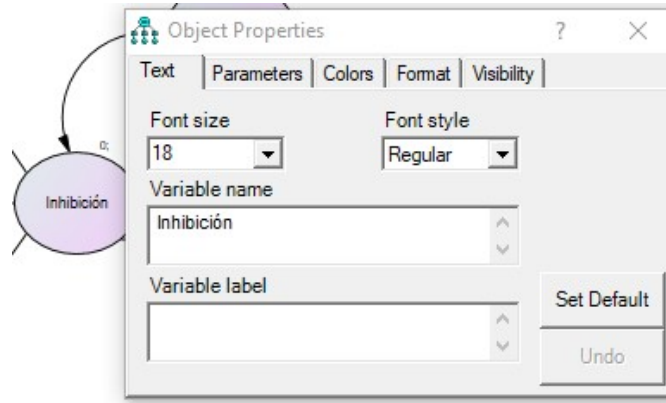


Figura C.5: Menú de propiedades de las variables utilizadas en el modelo

- **Paso 4:** En la Figura C.6 se observa la ventana del menú para analizar el modelo. En la ventana de estimación se puede colocar el método de estimación del modelo, ya sea el de máxima verosimilitud, mínimos cuadrados ponderados o mínimos cuadrados generalizados, también se pueden ver las ventanas como *Output*, donde están todas las opciones de salida que ofrece el programa. Ver las correlaciones, estimaciones de los parámetros así como diferentes opciones como *test* de normalidad o la matriz observada.

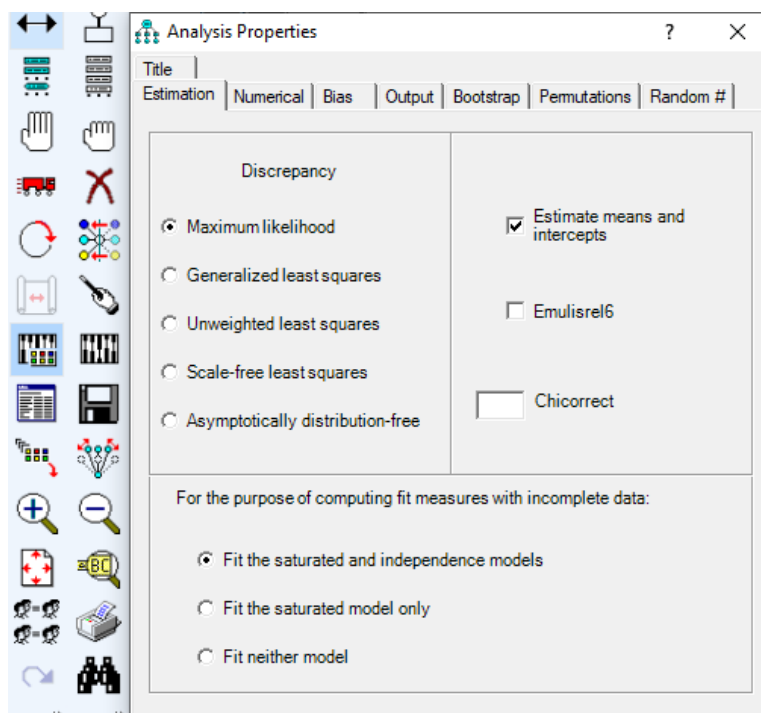


Figura C.6: Menú Análisis en AMOS

- **Paso 5:** En la Figura C.7 se observa el modelo, donde se pueden ver las estimaciones de los parámetros, todas ellas están estandarizadas para que la interpretación del modelo sea más sencilla.

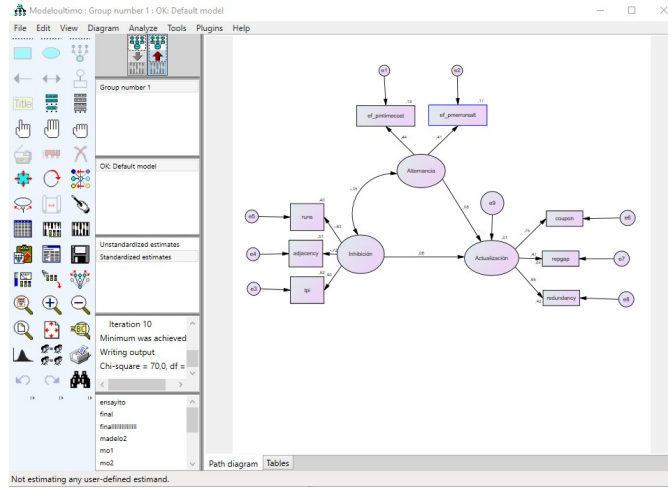


Figura C.7: Modelo de ecuaciones estructurales estimado por máxima verosimilitud

- **Paso 6:** Finalmente en la figura C.8 se tiene la hoja de resultado de las estimaciones.

Model Fit Summary					
CMIN					
Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	27	69,993	17	,000	4,117
Saturated model	44	,000	0		
Independence model	16	1114,336	28	,000	39,798
Baseline Comparisons					
Model	NFI	RFI	IFI	TLI	CFI
Default model	,937	,897	,952	,920	,951
Saturated model	1,000	1,000	1,000	1,000	1,000
Independence model	,000	,000	,000	,000	,000
Parsimony-Adjusted Measures					
Model	PRATIO	PNFI	PCFI		
Default model	,607	,569	,578		
Saturated model	,000	,000	,000		
Independence model	1,000	,000	,000		
NCP					
Model	NCP	LO 90	HI 90		
Default model	52,993	30,878	82,666		
Saturated model	,000	,000	,000		
Independence model	1086,336	980,867	1199,194		
FMN					

Figura C.8: Bondad de ajuste

C.3. Programa R

Programa R: Es un software para análisis estadístico, este es libre, puede extenderse fácilmente mediante paquetes que han creado muchos investigadores. Este hecho confiere a **R** un gran potencial para su uso en análisis de datos, este programa tiene una interfaz de usuario poco amigable y obliga a sus usuarios a escoger el paquete más adecuado entre los diferentes desarrollos para una misma técnica, (R Core Team, 2017).

Algunos paquetes para el análisis de modelos de ecuaciones estructurales son:

- **Paquete lavaan:** Paquete para el modelado de ecuaciones estructurales, el objetivo de éste es proporcionar una colección de herramientas que se pueden utilizar para explorar, estimar y entender una amplia familia de modelos variables latentes, incluyendo análisis de factores, ecuación estructural, longitudinal, entre otros, (Rossee, 2012). Sin embargo, el desarrollo del lavaan no ha hecho más que hacer mucho para alcanzar este ambicioso objetivo. El desarrollo de lavaan se ha centrado en el modelado de ecuaciones estructurales con variables observadas continuas.

Lavaan tiene 3 funciones para estimar modelos de ecuaciones estructurales, la primera de ellas es la función `cfa()`, la cual es enfocada en modelos de análisis factorial confirmatorio, la segunda función `sem()` se emplea para modelos estructurales y por último la función `lavaan()` permite estimar modelos de Ecuaciones Estructurales en los que se especifica el modelo completo.

Tomando La Aplicación del Capítulo 5, se hará la estimación del modelo de ecuaciones estructurales aplicando este paquete.

```
library(lavaan) #Se utiliza para estimar una gran variedad
#de modelos estadísticos multivariados, como
#el análisis de caminos, el Análisis Factorial confirmatorio,
#el modelado de Ecuaciones Estructurales y
#los modelos de curva de crecimiento.

library(semPlot) #Paquete que se utiliza para la generación de gráficos
#a partir del modelo evaluado.

library(semTools) #Paquete que se utiliza para el cálculo de fiabilidad
#a partir del modelo evaluado

indices<- read.table("./modelo_final.csv", sep=";",dec=",", head=T)
#Base de datos

model <- '
Alternancia =~ ef_pmtimecost + ef_pmerrorsalt
Inhibición =~ runs + adjacency + tpi
Actualización =~ coupon + redundancy + repgap
```

```

###regresiones

Actualización ~ Inhibición
Actualización ~ Alternancia
,

# Ajuste del modelo y salidas del modelo
fit <- sem(model, data=indices) #obtiene el estadístico básico
#de la prueba de chi-cuadrado.
summary(fit, standardized=TRUE) #El argumento estandarizado = VERDADERO
# aumenta la salida con valores de parámetros estandarizados.
coef(fit) #Extrae los valores estimados de los parámetros
#libres en el modelo, junto con sus nombres.

parameterEstimates(fit) #Obtiene intervalos de confianza para los
# parámetros estimados.
Est <- parameterEstimates(fit, ci=FALSE, standardized=TRUE)
subset(Est, op == "~")

MI <- modificationIndices(fit)
subset(MI, mi > 10)

#####
model <- '
Alternancia =~ ef_pmtimecost + ef_pmerrorsalt
Inhibición =~ runs + adjacency + tpi
Actualización =~ coupon + redundancy + repgap
###regresiones

Actualización ~ Inhibición
Actualización ~ Alternancia
,

fit.equal <- sem(model.equal, data=indices)
summary(fit.equal)

anova(fit, fit.equal)

fitMeasures(fit, c("cfi", "rmsea")) #Obtener indicadores de
#bondad de ajuste del modelo
inspect(fit)9
inspect(fit, what="start")

#Diagrama de Caminos del modelo
semPaths(fit)

```

```
semPaths(fit,"mod","est", edge.color = "black", style = "lisrel",  
residScale = 8)
```

- **Paquete SEM:** Los modelos de ecuaciones estructurales combinan modelos de medición (por ejemplo, de fiabilidad) con modelos estructurales (por ejemplo, de regresión). El paquete SEM, desarrollado por Fox J (2002), permite realizar algunos modelos de ecuaciones estructurales básicos. Para utilizarlo, añade el paquete SEM mediante el gestor de paquetes. El modelo de ecuaciones estructurales puede considerarse como una regresión corregida por la atenuación. El paquete SEM proporciona facilidades básicas de modelado de ecuaciones estructurales en el programa **R**, incluyendo la capacidad de ajustar ecuaciones estructurales en modelos de variables observadas mediante mínimos cuadrados de dos etapas y de ajustar ecuaciones latentes por Máxima Verosimilitud con información completa, asumiendo la multinormalidad.

Apéndice D

Descripción de las variables

Se consideran los diferentes métodos que se han desarrollado para cuantificar la generación aleatoria e incorporar estas escalas de medición en un programa informático (Windows95) llamado *RgCalc*. El programa está diseñado para ser apropiado para el análisis de varios tipos de situaciones de generación aleatoria empleadas en la literatura, principalmente en el análisis de datos psicológicos (Towse y Neil, 1998).

- **TPI:**

El *Turning Point Index* (TPI) o índice de punto de inflexión trata de calcular el número de respuestas que, como valores numéricos, marcan un cambio entre secuencias ascendentes y descendentes (es decir, puntos que representan picos). El número de puntos de giro se compara con un valor teórico dado por:

$$TP_{expected} = \frac{2}{3}(n - 2),$$

donde n es el número de respuestas aleatorias en el conjunto.

El valor de TPI se da como una puntuación porcentual, lo que indica la correspondencia entre los valores observados y esperados:

$$TPI = 100 \times \frac{TP_{observed}}{TP_{expected}}.$$

Así, los valores superiores al 100% indican que se produjeron demasiados puntos de giro (en relación con una distribución teórica de respuestas aleatorias), mientras que los valores inferiores al 100% indican menos puntos de giro de los esperados.

- **Adyacencia:** La medida A , o de adjacency, (a veces denominada puntuación este-reotipada) es una medida más específica o centrada en la frecuencia y se calcula como:

$$A = 100 \times \frac{\text{números de pares adyacentes}}{\text{número de pares de respuesta}}.$$

Las puntuaciones oscilaron entre el 0%, en el caso de que no hay pares vecinos, y el 100%, si el conjunto está compuesto en su totalidad por dichos pares.

- **Redundancia:** La redundancia R en la secuencia se encuentra determinando el grado de desviación de la generación de información ideal, expresando este valor como una puntuación porcentual,

$$R = 100 \times \left(1 - \frac{H_{\text{single}}}{H_{\text{max}}} \right),$$

donde la cantidad de información de primer orden que se proporciona por una secuencia se calcula como $H_{\text{single}} = \log_2 n - \frac{1}{n} (\sum n_i \log_2 n_i)$ donde n es el número de respuestas aleatorias en el conjunto. Para un conjunto de respuestas con unas alternativas, la máxima cantidad de información que es posible generar es n_i el número de ocurrencias de la i -ésima alternativa de respuesta (los cálculos son omitidos donde $n_i = 0$) y para un conjunto de respuestas con a diferentes alternativas, la máxima cantidad de información que es posible generar es $H_{\text{max}} = \log_2 a$.

Por lo tanto, una puntuación R de 0% indica que no hay redundancia (no se utiliza la misma opción de respuesta), y una puntuación R de 100% indica redundancia completa (se utiliza la misma opción de respuesta en todas partes).

Apéndice E

Modelo incluyendo edad y escolaridad

En este corto Apéndice simplemente se presenta el diagrama de caminos de un modelo de ecuaciones estructurales incluyendo 7 de las variables que se han considerado en el estudio (no se incluyó la variable *ef_pmtimecost* porque en el análisis factorial no se asocia con ningún factor) y dos variables que son muy importantes en el análisis, a saber: edad y escolaridad. Se recuerda que no se ha hecho un análisis riguroso de un modelo de ecuaciones estructurales con estas variables ya que los expertos del área están revisando las tareas asociadas a las funciones ejecutivas con el fin de, tal vez, obtener correlaciones significativas en el estudio de Baependi. Una vez logren tener estas nuevas variables se procederá al planteamiento del modelo de ecuaciones estructurales.

En la Figura E.1 se muestra el diagrama de caminos teniendo en cuenta las variables de edad y escolaridad. Se puede observar que Inhibición y Alternancia continúan significativamente correlacionadas (aunque el valor fue más bajo que el valor en el modelo inicial, Sección 5.3) y las otras correlaciones siguen bajas sin ser significativas estadísticamente. También se puede resaltar que edad y escolaridad están relacionadas con Alternancia.

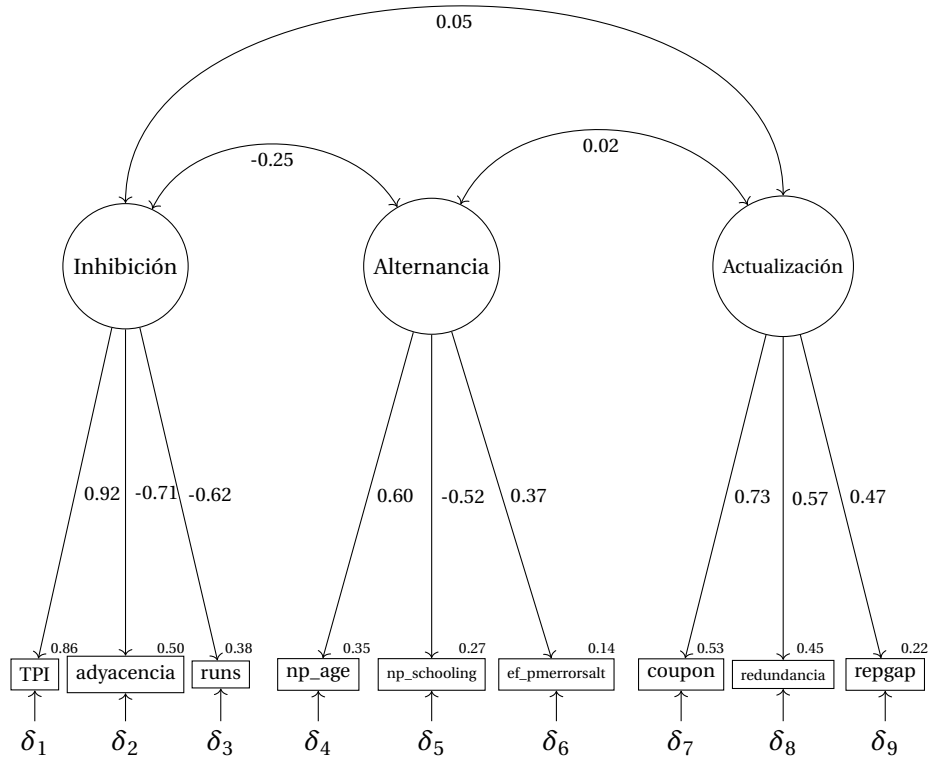


Figura E.1: Diagrama de caminos con las variables : edad y escolaridad

En este primer planteamiento incluyendo las variables edad y escolaridad las pruebas de bondad de ajuste indican que el modelo es adecuado (resultados no presentados), pero como se ha anotado anteriormente, estas conclusiones definitivas se tendrán una vez se incluyan las tareas revisadas. Se espera tener un modelo de ecuaciones estructurales donde, además de las variables edad y escolaridad, se incluyan en el estudio otras variables como depresión y ansiedad, entre otras variables asociadas con cognición.

Apéndice F

Modelos SEM: Primeros ensayos

En este Apéndice se presentan algunos de los primeros ensayos de los modelos planteados como una primera visualización al análisis factorial y ecuaciones estructurales, vale resaltar como en toda aplicación se realizaron muchos ensayos incluyendo variables que no se han mencionado en este trabajo, las mencionaremos en forma general para tener algunas ideas de algunos planteamientos en el área de Psicología y donde se explica de manera breve el porque se descartaron.

En el diagrama de caminos que se presenta en la Figura F.1, se tomaron las siguientes variables para hacer una combinación lineal de ellas con fin de definir un resumen al cual se le asigna la variable tiempo de costo. Las variables involucradas en la combinación lineal fueron:

1. **ef_pmerrorsar**: Alternar errores durante la batería de cuentas alternativas.
2. **ef_pluserrors**: Aciertos aritméticos durante la batería de cuentas de menos.
3. **ef_minuserrors**: Errores aritméticos durante la batería de cuentas de menos.

La variable ef_pmcombinación es el resultado de hacer la siguiente operación: $ef_pmcombinacion = ef_pmerrorsar - (ef_pluserrors + ef_minuserrors)$, es decir, se toman los errores durante la batería de cuentas alternativas, se le resta la suma de los aciertos aritméticos durante la batería de cuentas de menos y los errores aritméticos durante la batería de cuentas menos.

Los expertos en el área sugirieron tomar esa combinación para descubrir si conseguía explicar el factor de Alternancia, pero lastimosamente como veremos a continuación esa hipótesis se rechazó.

El modelo representado en el diagrama de caminos ilustrado en la Figura F.1, se obtuvo buen ajuste del modelo (resultados no presentados), pero la variable ef_pmcombinacion presentó una carga factorial relacionada con Alternancia muy baja (como observado en la Figura la carga factorial es de 0,20) además el porcentaje de varianza explicada de la

variable ef_pmcombinacion es muy baja (0,04). Por este motivo este modelo fue descartado ya que se esperaba mejores resultados.

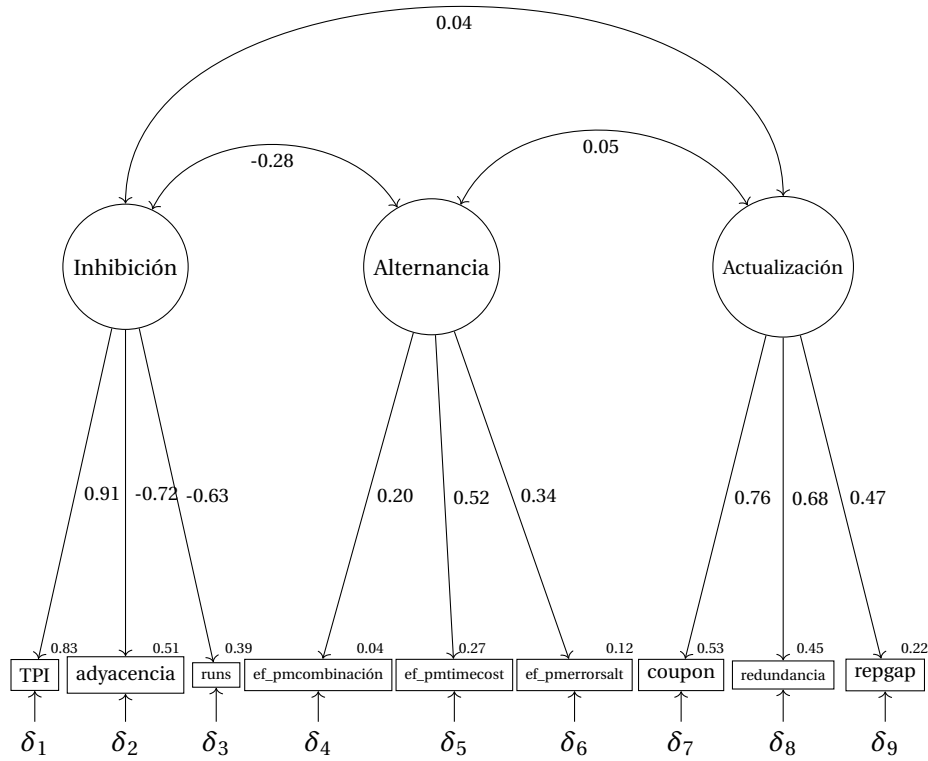


Figura F1: Diagrama de caminos con la variable combinación

Continuando con los modelos ensayados, se optó por incluir algunas relaciones causales entre los tres factores junto con algunos errores, esto con el fin de proponer una nueva forma de análisis para que los expertos determinaran si tenía sentido en relación a la investigación. Bajo estas consideraciones, en la Figura F2 se puede observar el diagrama de caminos del modelo de ecuaciones estructurales con relaciones causales tomando las variables anteriormente mencionadas.

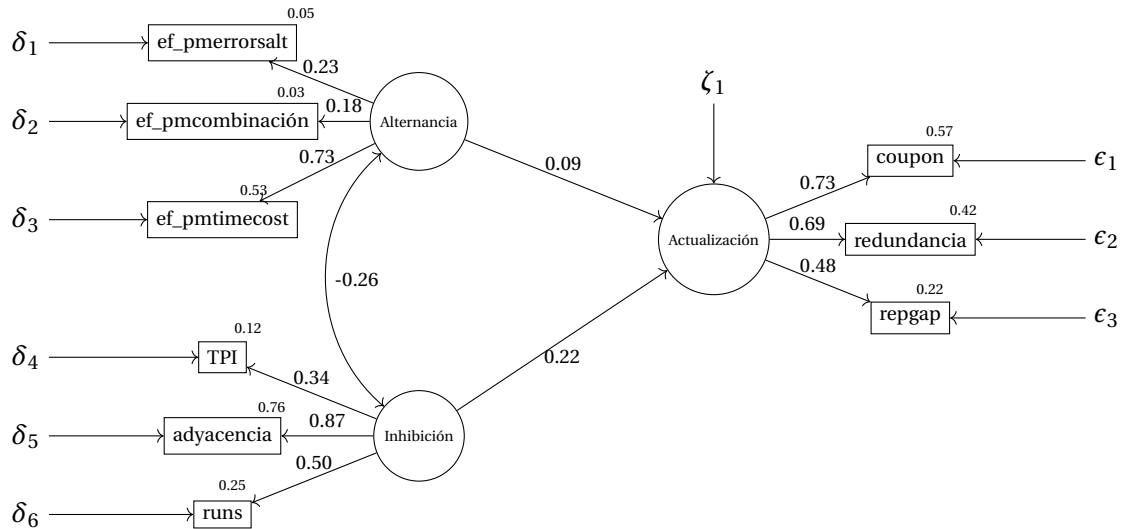


Figura F2: Diagrama de caminos modelo fallido con la variable combinación

Observando la Figura F2 se puede observar la carga factorial de la variable ef_pmcombinación de 0,18, además el porcentaje de varianza explicada por la variable es muy bajo (0,03) por tanto los investigadores del área descartaron esta variable y este modelo.

También se propuso el modelo con las variables de los aciertos aritméticos que obtuvieron los individuos en las cuentas más, menos y alternadas. Este modelo se propuso para mirar correlaciones positivas y tal vez significativas. Siendo así, las variables que se incluyeron con la esperanza que estuvieran asociadas con el factor Alternancia fueron:

1. **ef_plusright:** Aciertos atiméticos durante la batería de cuentas de más.
2. **ef_minusright:** Aciertos atiméticos durante la bateria de cuentas de menos.
3. **ef_pmrightari:** Aciertos aritméticos durante la batería de cuentas alternadas.
4. **ef_pmtimeall:** Suma del tiempo de costo en las tres baterías (más, menos e alternadas)

En la Figura F3 se presenta el diagrama de caminos con las variables de los aciertos aritméticos, se puede observar la correlación positiva de la variables latentes Inhibición y Alternancia de 0,21 y la correlación positiva pero baja entre Inhibición y Alternancia de 0,05. Este modelo tuvo buen ajuste (los resultados los cuales no se presentan).

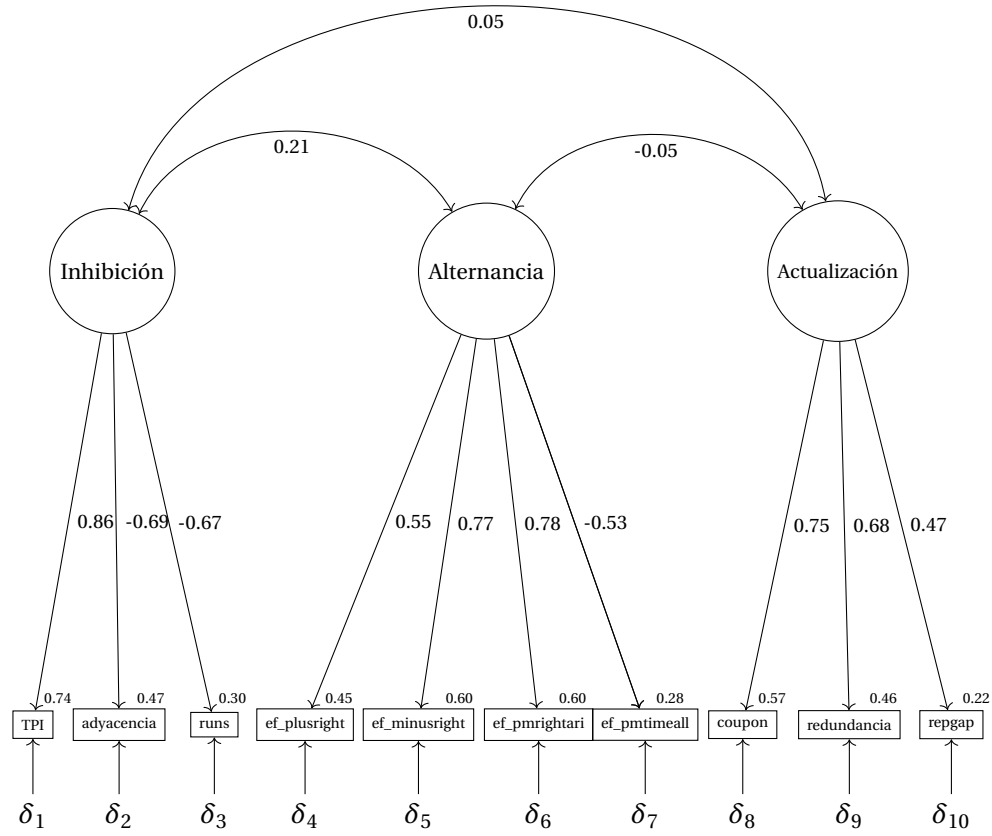


Figura F3: Diagrama de caminos con aciertos

Por último, teniendo en cuenta el modelo considerado anteriormente, se sugirió observar las relaciones causales entre estos factores, pero definitivamente a nivel de la aplicación no aportan conclusiones relevantes para la investigación

Los expertos del área descartaron este modelo ya que ellos recomendaron que para la variable latente de Alternancia trabajar con las variable de error ef_pmerrorsalt y el ef_pmtimecost. También se propuso el modelo de ecuaciones estructurales con relaciones causales con las variables de los aciertos que obtuvieron los individuos presentado en la Figura F4.

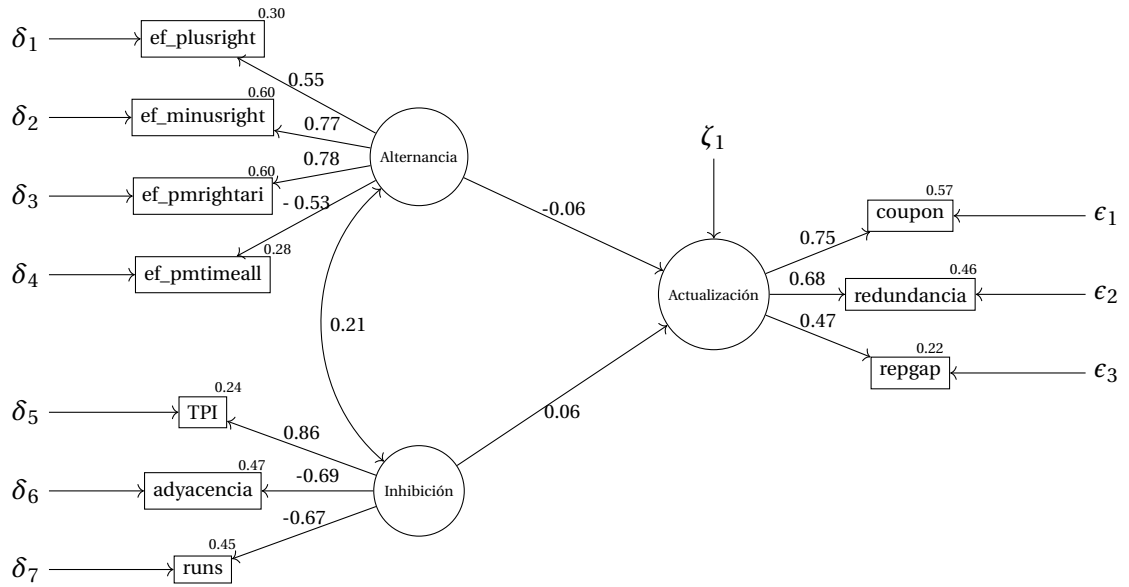


Figura F4: Diagrama de caminos con aciertos y efectos

A través de los planteamientos anteriores, se percibe la necesidad de trabajar conjuntamente con expertos en el área, ya que mismo que los modelos tenían un buen ajuste estadístico, los resultados no cumplieron con las conclusiones a las que los expertos querían llegar, a saber que los 3 factores tuvieran correlaciones altas y significativas. De cualquier forma estos análisis planteados fueron de gran ayuda para tener un modelo final y poder contribuir con sugerencias importantes para dar continuidad a la investigación.

Bibliografía

- [1] **Ahmed S., Taporoski T., Gómez L., Ruiz E., Bejamini E., Horimoto A y cols., (2019).** Data from the brazilian baependi heart study cohort yield new insights into the genetic epidemiology of insomnia. *BMJ Open Respiratory Research*. Vol. 6.
- [2] **Arbuckle, J. L. (1997).** Amos Users' Guide. Version 3.6. Chicago: SmallWaters Corporation.
- [3] **Baddeley A. D. y Hitch G. (1974).** Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.
- [4] **Baddeley, A. D. (2012).** Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, 63(1), 1–29. <http://doi.org/10.1146/annurev-psych-120710-100422>.
- [5] **Baggetta P. y Alexander P. A. (2016).** Conceptualization and operationalization of executive function. *Mind, Brain, and Education*, Vol. 10, no. 1, pp. 10-33.
- [6] **Barroso L. P. y Artes R. (2003).** Análise Multivariada. Minicurso do 10 Simposio de Estatística Aplicada a Experimentação Agronômica. RBRAS - SEAGRO. Lavras, UFLA.
- [7] **Batista F., Joan M. y Coenders G. (2000).** Modelos de Ecuaciones Estructurales. Editorial La Muralla. Madrid (Spain).
- [8] **Bekker, P. A. (1994).** Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, Vol. 62, no. 3, pp. 657–681. JSTOR. www.jstor.org/stable/2951662.
- [9] **Bentler P. M. y Bonett D. G. (1980).** Significance tests and Goodness-of-Fit in the analysis of covariance structures. *Psychological Bulletin*. Vol. 88, no. 3. University of California, Los Angeles, CA.
- [10] **Bentler, P. M. (1995).** Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Sage Publications, Inc.
- [11] **Bistaffa, B. C. (2010).** Incorporacao de indicadores categóricos ordinais em modelos de equações estruturais, São Paulo.

- [12] **Blalock, H Jr.(1969)**. Recasting verbal theories as causal models. En: Theory Construction, Englewood Cliffs, N.J.: Prentice-Hall. Chicago / Turabian.
- [13] **Bollen, K. A. (1989)**. Structural Equations with Latent Variable. Wiley, New York, USA.
- [14] **Bollen K. A. y Davis W. R. (2009)**. Two Rules of Identification for Structural Equation Models, Structural Equation Modeling: A Multidisciplinary Journal, 16:3, 523-536, DOI: 10.1080/10705510903008261.
- [15] **Boudon, R. (1965)**. A method of linear causal analysis: Dependence analysis. American Sociological Review, Vol. 30, no. 3, 1965, pp. 365–374. JSTOR, www.jstor.org/stable/2090717.
- [16] **Briseño C., Herrera R., Enders J. y Fernández A. (2005)**. Estudio de riesgos ergonómicos y satisfacción laboral en el personal de enfermería. Rev Salud Pública (Córdoba); 9(1):53-59.
- [17] **Brito C. y Pearl J. (2002)**. A New Identification Condition for Recursive Models With Correlated Errors, Structural Equation Modeling: A Multidisciplinary Journal, 9:4, 459-474, https://doi.org/10.1207/S15328007SEM0904_1. University of California, Los Angeles, CA.
- [18] **Browne, M. W. (1984)**. Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37(1), 62–83. <https://doi.org/10.1111/j.2044-8317.1984.tb00789.x>.
- [19] **Casas, M. (2016)**. Los modelos de ecuaciones estructurales y su aplicación en el índice europeo de satisfacción al cliente. Madrid: Universidad San Pablo-CEU.
- [20] **Cervelló E., Iglesias D. y cols., (2004)**. Aplicación de modelos de ecuaciones estructurales al estudio de la motivación de los alumnos en las clases de educación física. Rev Educ (Madrid); (335):371-383.
- [21] **Cuautle, L. (2010)**. Creación de una propuesta de modelo de sistema de gestión integral para las empresas mexicanas de manufactura [Tesis doctoral]. Puebla: Universidad Popular Autónoma del Estado de Puebla.
- [22] **Cupani, M. (2012)**. Análisis de Ecuaciones Estructurales: conceptos, etapas de desarrollo y un ejemplo de aplicación. Revista Tesis, 2, 186-199. Extraído de <https://revistas.psi.unc.edu.ar/index.php/tesis/article/view/2884>.
- [23] **de Oliveira C. M., Pereira A. C., De Andrade M., Soler J. M., y Krieger J. E. (2008)**. Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. BMC medical genetics, 9(1), 1-8.
- [24] **De Souza, R. (2013)**. Contribuições à análise de outliers em modelos de equações estruturais. <http://www.teses.usp.br/teses/disponiveis/45/45133/tde-19062013-135858>. São Paulo.
- [25] **Díaz L. G. y Morales M. A. (2012)**. Análisis estadístico de datos Multivariados. Universidad Nacional de Colombia.

- [26] **Duncan, O. D. (1969)**. Introduction to Structural Equation Models. Nueva York: Academic Press.
- [27] **Egan K. y cols., (2016)**. Timing and quality of sleep in a rural Brazilian family-based cohort, the Baependi Heart Study. *Sci Rep* 6, 39283. <https://doi.org/10.1038/srep39283>.
- [28] **Escobedo M. T., Hernández J. A. y cols., (2016)**. Modelos de ecuaciones estructurales: Características, fases, construcción, aplicación y resultados. Universidad Autónoma de Ciudad Juárez, Chihuahua, México.
- [29] **Everitt, (1984)**. Everitt. An Introduction to Latent Variable Models. Chapman and Hall, London.
- [30] **Fernandez V., (2004)**. Tesis doctoral Relaciones encontradas entre las dimensiones de las estructuras organizacionales y los componentes del constructo capacidad de absorción: El caso de empresas ubicadas en el territorio español. Barcelona.
- [31] **Fox J., (2002)**. TEACHER'S CORNER: Structural Equation Modeling With the sem Package in R, *Structural Equation Modeling: A Multidisciplinary Journal*, 13:3, 465-486, DOI: 10.1207/s15328007sem1303_7.
- [32] **Friedman N. P. y Miyake A. (2017)**. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, 86, 186-204.
- [33] **Giolo S., Pereira A., De Andrade M., Oliveira C., Krieger J. y Soler J. (2009)**. Genetic analysis of age-at-onset for cardiovascular risk factors in a Brazilian family study. *Human Heredity*. 68(2), 131-138.
- [34] **Hair J. E., Anderson R. E., Tatham R. L. y Black W. C. (1999)**. Análisis multivariante. 5. edición. PRENTICE HALL, Madrid.
- [35] **Hatcher, L. (2003)**. A Step-by-Step Approach to using SAS for Factor Analysis and Structural Equation Modeling. Cary: SAS Institute Inc.
- [36] **Hortal, G. (2012)**. Trabajo de Investigación. Master Universitario en Estadística Aplicada. El modelo de Análisis Factorial: teoría y aplicaciones. Universidad de Granada. Departamento de Estadística e I.O.
- [37] **Hu L. y Bentler P. (1999)**. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal*, 6:1, 1-55, DOI: 10.1080/10705519909540118.
- [38] **Johnson R. A. y Wichern D. W. (1998)**. Applied Multivariate Statistical Analysis. University of Wisconsin. Madison.
- [39] **Jöreskog, K. G. (1973)**. Linear structural relationships (LISREL) in family research. *Journal of Marriage and Family*, Vol. 50, no. 4, 1988, pp. 937-948. JSTOR, www.jstor.org/stable/352105.
- [40] **Jöreskog, K. G. (1978)**. Structural analysis of covariance and correlation matrices. *Psychometrika* 43, 443-477 (1978). <https://doi.org/10.1007/BF02293808>.

- [41] **Jöreskog K. G. y Sörbom D. (1986)**. LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods. Mooresville, IN: Scientific Software, Inc.
- [42] **Kaplan, D. (2000)**. Structural Equation Modeling: Foundations and Extensions, Vol. 10 de Advanced Quantitative Techniques in the Social Sciences, Sage Publications, California.
- [43] **Kline, R. B. (2011)**. Principles and practice of structural equation modeling. 2nd ed. New York: The Guilford Press.
- [44] **Keith, T. Z. (2015)**. Multiple regression and beyond: An introduction to multiple regression and structural equation modeling. 2nd ed. New York: Taylor & Francis.
- [45] **Lawson G. M., Hook C. J., Hackman D. A., Farah M. J., Griffin J. A., Freund L. S. y McCardle P. (2014)**. Socioeconomic status and neurocognitive development: Executive function. Executive Function in Preschool Children: Integrating Measurement, Neurodevelopment, and Translational Research. American Psychological Association, 1-28.
- [46] **Latif, S. A. (2000)**. Modelagem de equações estruturais, Dissertação de Mestrado. Instituto de Matemática e Estatística - USP, São Paulo.
- [47] **Lawley, D. N. (1940)**. The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society Edinburgh; 60:64-82.
- [48] **Long, J. (1983)**. Confirmatory factor analysis a preface to LISREL. Newbury Park: Sage Publications.
- [49] **Lozano, S. (2007)**. Validación de un modelo de medida de las dificultades en los procesos de toma de decisiones sobre la carrera profesional. Rev Educ (Madrid); (343):325-351.
- [50] **Madrigal M. S. y Ávila F. C. (2019)**. Modelos de Ecuaciones Estructurales en Investigaciones Sociales enfocadas al comportamiento de compra de los Millennials. Revista Espacios. Vol. 40, no. 11, pág. 13
- [51] **Manzano, A. P. (2017)**. Introducción a los modelos de ecuaciones estructurales. Investigación en Educación médica. Vol. 7, no. 25. México. <https://doi.org/10.1016/j.riem.2017.11.002>.
- [52] **Marôco, J. (2010)**. Análise de Equações Estruturais: Fundamentos Teóricos, Software e Aplicações. Rep ortNumber, Pêro Pinheiro.
- [53] **Mateos G. y Morales A. (2011)**. Los modelos de ecuaciones estructurales: una revisión histórica sobre sus orígenes y desarrollo. Ciencia & trabajo. 18. 16-22. 10.4067/S0718-24492016000100004.
- [54] **Matsueda, R. L. (2012)**. The general LISREL model. Lecture notes of SEM class at the University of Washington.

- [55] **Maynez, A. (2011)**. La transferencia de conocimiento organizacional como fuente de ventaja competitiva sostenible: modelo integrador de factores y estrategias. [Tesis doctoral]. Puebla: Universidad Popular Autónoma del Estado de Puebla.
- [56] **Miyake A. y Friedman N. P. (2012)**. The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current directions in psychological science*. 21. 8-14. 10.1177/0963721411429458.
- [57] **Miyake A., Friedman N. P. y cols., (2000)**. The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cogn Psychol*. 2000 Aug;41(1):49-100. doi: 10.1006/cogp.1999.0734. PMID: 10945922.
- [58] **Moffitt T. E., Arseneault L., Belsky D., Dickson N., Hancox R. J., Harrington H. y cols., (2011)**. A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7), 2693-2698.
- [59] **Moneta A. M., Gonzalez M., Arrieta M. y Britos V. (2018)**. Propuesta de un modelo de ecuaciones estructurales para explicar el desempeño académico en e-learning. CRUC-IUA UNDEF. <https://rdu.iaa.edu.ar/handle/123456789/1768>.
- [60] **Morra S., Panesi S., Traverso L., y Usai M. C. (2018)**. Which tasks measure what? *Journal of Experimental Child Psychology*, Vol.167, pp. 246-258, ISSN 0022-0965. <https://doi.org/10.1016/j.jecp.2017.11.004>.
- [61] **Mulaik, (2009a)**. Foundations of Factor Analysis. Chapman y Hall/CRC.
- [62] **Muthén, B. (1983)**. Latent variable structural equation modeling with categorical data. *Journal of Econometrics*; 22:43-65.
- [63] **Nachtigall C., Kroehne U., Funke F. y Steyer R. (2003)**. (Why) should we use SEM? Pros and cons of structural equation modeling. *Methods of Psychological Research Online*; 8:1-22.
- [64] **Newson, J. (2012)**. Some Clarifications and Recommendations on Fit Indices. Course notes of SEM class at Portland State University.
- [65] **Oliveira M., Pereira A., De Andrade M., Soler J. y Krieger J. (2008)**. Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. *BMC Medical Genetics*. 9:32, 881-888.
- [66] **Pelegrina S., Lechuga M. T., Castellanos M. C. y Elosúa M. R. (2016)**. Memoria de trabajo. Madrid: Alianza Editorial S.A.
- [67] **R Core Team (2017)**. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- [68] **Raykov T. y Marcoulides G. A. (2006)**. A first course in structural equation modeling (2nd ed.). Lawrence Erlbaum Associates Publishers.

- [69] **Raykov T. y Widaman K. F. (1995)**. Issues in applied structural equation modeling research, *Structural Equation Modeling: A Multidisciplinary Journal*, 2:4, 289-318, DOI: 10.1080/10705519509540017.
- [70] **Rigdon, E. E. (1995)**. A Necessary and Sufficient Identification Rule for Structural Models Estimated in Practice, *Multivariate Behavioral Research*, 30:3, 359-383, DOI: 10.1207/s15327906mbr3003_4.
- [71] **Romero, R. (2015)**. Modelo de lealtad a partir de un análisis de ecuaciones estructurales. *Comunicaciones en Estadística*, Vol. 9, no. 2, pp. 165–197. Universidad Santo Tomás.
- [72] **Rosseel, Y. (2012)**. Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. <http://www.jstatsoft.org/v48/i02/>.
- [73] **Salgueiro, M. F. (2012)**. Modelos com Equações Estruturais. Sociedade Portuguesa de Estatística. ISBN: 978-972-8890-26-1.
- [74] **SPSS, (1993)**. SPSS® LISREL® 7 and PRELIS®: User's Guide and Reference. Chicago, IL: SPSS.
- [75] **StataCorp (2013)**. Stata longitudinal-data/panel-data reference manual release 13. Stata Press Publication, StataCorp LP, College Station, Texas.
- [76] **Satorra A. y Bentler P. (1988)**. Scaling corrections for statistics in covariance structure analysis (UCLA Statistics Series #2). Los Angeles, CA: University of California.
- [77] **Satorra A. y Bentler P. (1988)**. Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye y C.C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- [78] **Satorra A. y Bentler P. (2001)**. A scaled difference chisquare test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.
- [79] **Satorra A. y Bentler P. (2010)**. Ensuring Positiveness of the Scaled Difference Chisquare Test Statistic. *Psychometrika*, 75, 243-248.
- [80] **StataCorp., (2013)**. Stata 13 Base Reference Manual. College Station, TX: Stata Press
- [81] **Tanaka, J. S. y Huba, G. J. (1985)**. A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, 38, 197–201.
- [82] **Taporoski T. P., Negrao A. B., Horimoto A. R., Duarte N. E., Alvim R. O., De Oliveira C. M. y cols., (2015)**. Shared Genetic Factors of Anxiety and Depression Symptoms in a Brazilian Family-Based Cohort, the Baependi Heart Study. *PloS one*. 10. e0144255. 10.1371/journal.pone.0144255.
- [83] **von Schantz M. y cols., (2015)**. Distribution and heritability of diurnal preference (chronotype) in a rural Brazilian family-based cohort, the Baependi study. *Sci. Rep.* 5, 9214; DOI:10.1038/srep09214.

- [84] **Westland, J. C. (2015).** Structural equation models: From paths to networks. Springer: 175.
- [85] **Widaman K. F y Thompson J. S. (2003).** On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8(1), 16–37.
- [86] **Zamora S. y Schiattino I.(2008).** Modelos de ecuaciones estructurales ¿Qué es eso? *Ciencia y Trabajo*, ISSN-e 0718-2449, no. 29, pp. 106-110.
- [87] **Zamora S., Cazorla. L. y cols., (2010).** Análisis factorial: una técnica para evaluar la dimensionalidad de las pruebas. Cuaderno técnico 6. México DE.