



UNIVERSIDAD NACIONAL AUTÓNOMA  
DE MÉXICO

---

---

FACULTAD DE CIENCIAS

Modelos de Ecuaciones Estructurales:  
una aplicación en *Business Analytics*

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuario

P R E S E N T A:

REYNALDO DAVID RUIZ JIMÉNEZ

DIRECTORA DE TESIS:

DRA. LIZBETH NARANJO ALBARRÁN

Ciudad Universitaria, 2017.

**CDMX**





Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



# Índice general

<b>Introducción</b>	<b>vii</b>
<b>1. Antecedentes</b>	<b>1</b>
1.1. Correlación y causalidad . . . . .	1
1.2. Variables latentes . . . . .	3
1.3. Diagrama de trayectorias ( <i>Path Diagram</i> ) . . . . .	4
1.4. Tipos de relaciones . . . . .	5
<b>2. Análisis de factores confirmatorio</b>	<b>9</b>
2.1. Introducción . . . . .	9
2.2. Planteamiento del modelo . . . . .	11
2.3. Identificación del modelo AFC . . . . .	14
2.4. Estimación de parámetros . . . . .	16
2.4.1. Estimación por mínimos cuadrados no ponderados . . . . .	17
2.4.2. Estimación por mínimos cuadrados generalizados . . . . .	18
2.4.3. Estimación por máxima verosimilitud . . . . .	18
2.4.4. Estimación por la teoría de la distribución elíptica . . . . .	18
2.5. Pruebas de bondad de ajuste . . . . .	18
2.5.1. Matriz residual de covarianzas . . . . .	19
2.5.2. Estadística Ji-cuadrada para el contraste global del modelo . . . . .	19
2.5.3. Estadísticas <i>ad hoc</i> . . . . .	19
2.6. Reespecificación del modelo . . . . .	22
2.6.1. Contraste del multiplicador de Lagrange . . . . .	23
2.6.2. Contraste de Wald . . . . .	24
2.7. Paquete <code>sem</code> (software R) . . . . .	25
<b>3. Modelos de ecuaciones estructurales</b>	<b>31</b>

3.1. Introducción . . . . .	31
3.1.1. Aplicaciones . . . . .	31
3.1.2. ¿Qué son los MEE? . . . . .	31
3.1.3. Componentes de los MEE . . . . .	32
3.2. Representación matemática del modelo . . . . .	34
3.3. Identificación del modelo . . . . .	37
3.3.1. Tipos de parámetros . . . . .	37
3.3.2. Restricciones del modelo . . . . .	38
3.4. Estimación . . . . .	39
3.4.1. Mínimos cuadrados generalizados . . . . .	42
3.4.2. Máxima verosimilitud . . . . .	45
3.5. Bondad de ajuste . . . . .	47
3.5.1. Ajuste global del modelo . . . . .	47
3.5.2. Análisis de validez . . . . .	50
3.5.3. Ajuste del modelo estructural . . . . .	53
3.5.4. Reespecificación del modelo . . . . .	54
3.6. Descomposición del modelo . . . . .	54
3.7. Ejemplo . . . . .	56
<b>4. Aplicación</b>	<b>63</b>
4.1. Antecedentes . . . . .	63
4.2. Objetivo de la aplicación del Análisis de Trayectoria . . . . .	64
4.3. Desarrollo del modelo . . . . .	64
4.3.1. Conocimiento de la Marca . . . . .	65
4.3.2. Recopilación y estructuración de datos . . . . .	65
4.3.3. Análisis descriptivo de las variables . . . . .	67
4.3.4. Planteamiento del modelo . . . . .	73
4.3.5. Resultados y bondad de ajuste del modelo . . . . .	74
4.3.6. Discusión y comentarios finales . . . . .	78
<b>5. Conclusiones</b>	<b>81</b>
<b>A. El producto Kronecker</b>	<b>83</b>
A.1. Definiciones y notaciones . . . . .	83
<b>B. Transformaciones en medios</b>	<b>85</b>

B.1. Efecto Lag . . . . .	85
B.2. Efecto Adstock . . . . .	86
B.2.1. Modelos Adstock . . . . .	86
<b>Bibliografía</b>	<b>89</b>



# Introducción

El Análisis de Regresión Lineal ha sido una herramienta relevante en casi cualquier análisis de relación lineal, ya que es posible encontrar un modelo matemático que asocie al menos una variable independiente a una dependiente de manera lineal. Por otro lado, existen otras técnicas cuyo objetivo es reducir la dimensión de la información, tal como el Análisis de Componentes Principales (CP), el Análisis de Factores Exploratorio (AFE) o el Análisis de Factores Confirmatorio (AFC).

Estas técnicas estadísticas han tenido una participación importante para cualquier área, principalmente en esta época que se genera una gran cantidad de datos porque permiten analizar y explicarlos, estableciendo modelos matemáticos que dejan entender la asociación entre ellos.

En particular, el modelo de regresión muestra la relación lineal entre una variable dependiente y al menos una independiente, sin embargo, como lo hemos mencionado, estamos en una era en que hay una gran generación de información y tener relaciones simples pueden no explicar de manera adecuada su comportamiento dinámico ya que podrían presentarse múltiples relaciones, es decir, puede haber más de una variable dependiente y que a su vez ellas se vuelvan variables explicativas de otras variables.

Si bien el analista puede postular varios modelos de regresión para la explicación de más de una variable dependiente, sin embargo, no se puede tener un ajuste global de más de dos modelos de manera simultánea, sino de cada uno de ellos de manera separada, lo que haría que se esté visualizando la realidad de forma parcial, y no de forma global, y que podría terminar en resultados sesgados.

En particular, para la investigación de mercados o de medios de comunicación, la limitación se presenta al hacer preguntas como: ¿cuántas variables determinan la imagen de una marca o una tienda?, ¿cómo se combina esa imagen con otras variables para afectar las decisiones de compra y la satisfacción con la tienda?, ¿cómo se convierte la satisfacción con la tienda en la lealtad a largo plazo?, etc.

La técnica estadística que resuelve este tipo de conflictos son los Modelos de Ecuaciones Estructurales (MEE) y en esta tesis se revisa bajo qué condiciones podrá ser aplicada, cuáles son los resultados obtenidos y cuál sería su interpretación.

Este trabajo presenta el desarrollo teórico de esta técnica y algunos ejemplos aplicados.

El objetivo de este trabajo es entender lo que podrían responder estas técnicas estadísticas, generar un interés para aplicar este tipo de análisis, así como mostrar una de las áreas de aplicación.

El contenido del trabajo queda de la siguiente manera:

- En el primer capítulo se mencionan conceptos básicos necesarios para plantear conceptos más complejos estudiados en los siguientes capítulos.
- En el segundo capítulo se describe el AFC que es la base teórica para el desarrollo de los MEE.
- Para el tercer capítulo se revisan los MEE. En los capítulos dos y tres se revisan identificación del modelo, métodos de estimación, estadísticos de bondad de ajuste, estadísticos de modificación del modelo, etc.
- En el cuarto capítulo se desarrolla una aplicación basada en el Análisis de Trayectoria o *Path Analysis*, que es un caso particular de MEE en que sólo se trabaja con variables observadas y se puede obtener un modelo global en que interactivamente las variables independientes se vuelven dependientes. Esta aplicación fue un estudio real con el que se tomaron decisiones de inversión en publicidad.
- Finalmente, en el capítulo 5 se concluye con los resultados más relevantes de este trabajo, mencionando las ventajas y desventajas de esta técnica.
- Se anexan dos apéndices. En el apéndice A se mencionan las definiciones, teoremas y lemas que se utilizan para la estimación de los parámetros de MEE. En el apéndice B se da una explicación acerca del término *Adstock* y el tiempo de retardo entre el mensaje y el acto de compra.

# Capítulo 1

## Antecedentes

### 1.1. Correlación y causalidad

Dentro del modelo de Análisis de Factores Confirmatorio (AFC)<sup>1</sup> y del Modelo de Ecuaciones Estructurales (MEE) se trabajará con dos conceptos que son fundamentales: la *correlación* y la *causalidad*, es importante dejar claro que estos conceptos son diferentes. A continuación se darán las definiciones de ellos.

Antes de plantear la definición de correlación lo haremos primero con la covarianza, la cual no tiene una interpretación estandarizada, es decir, la covarianza estará en la escala en que están las variables. Por lo tanto, en la mayoría de las ocasiones se usa la correlación la cual se define a partir de la covarianza.

**Definición 1.1.1.** Sean  $X$  y  $Y$  dos variables aleatorias, la covarianza de  $X$  y  $Y$ , denotada por  $cov(X, Y)$ , se define como:

$$cov(X, Y) = E[(X - E(X))(Y - E(Y))],$$

donde  $E(X)$  y  $E(Y)$  son las esperanzas de las respectivas variables aleatorias, las cuales deben ser finitas.<sup>2</sup>

Ya expuesta la definición de covarianza, es posible definir el coeficiente de correlación. Se dará la definición de coeficiente de correlación ya que es la unidad de medida de la correlación entre dos variables.

**Definición 1.1.2.** El coeficiente de correlación de las variables aleatorias  $X$  y  $Y$ , denotado por  $\rho(X, Y)$ , está definido como

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}.$$

Naturalmente, para esta definición se necesita suponer que las varianzas son finitas.<sup>3</sup>

---

<sup>1</sup>Tambi n conocido como An lisis Factorial Confirmatorio

<sup>2</sup>Luis Rinc n (2007), *Curso Intermedio de Probabilidad*, p gina 157.

<sup>3</sup>Luis Rinc n (2007), *Curso Intermedio de Probabilidad*, p gina 159.

Como ya habíamos mencionado, el coeficiente de correlación es una unidad de medición de la correlación, así de este modo, mencionaremos tres resultados que nos ayudarán a asociar el coeficiente de correlación con la correlación entre dos variables:

1. Si  $X$  y  $Y$  son independientes, entonces  $\rho(X, Y) = 0$ ;
2.  $-1 \leq \rho(X, Y) \leq 1$ ;
3.  $|\rho(X, Y)| = 1$  si y sólo si, existen constantes  $a$  y  $b$  tales que, con probabilidad uno,  $Y = aX + b$ . Además,  $a > 0$  si  $\rho(X, Y) = 1$ , y  $a < 0$  si  $\rho(X, Y) = -1$ .

Dados los resultados anteriores, la interpretación de la correlación queda expresada de la siguiente manera: *La correlación es el grado de asociación lineal que hay entre dos variables.* Cuando el coeficiente de correlación está cerca de  $-1$  implica que las variables tienen una relación inversamente proporcional, y cuando está cerca de  $1$  es directamente proporcional. También podemos decir que cuanto  $\rho$  esté más cerca de  $1$  o  $-1$  la probabilidad de que haya una relación lineal es más grande, mientras que cuando esté cerca de cero nada es posible concluir. Por último, existe la seguridad que cuando las variables son independientes, siempre se cumple que  $\rho = 0$ , sin embargo, si  $\rho = 0$  no es posible concluir que las variables son independientes.

Dado que la mayoría de las investigaciones se hacen con más de una variable es necesario tener una expresión que muestre el grado de correlación que hay entre todas las variables por parejas, para esto a continuación damos la definición de matriz de correlaciones.

**Definición 1.1.3.** Sean  $X_1, X_2, X_3, \dots, X_n$  variables aleatorias, tal que la matriz de correlación se define como el arreglo de correlaciones por filas y columnas, donde en la intersección de la  $i$ -ésima fila y la  $j$ -ésima columna se encuentra la correlación de las variables  $X_i$  y  $X_j$ , es decir,  $\rho(X_i, X_j)$ . En términos matemáticos, la matriz de correlaciones queda definida por:

$$\begin{pmatrix} 1 & \rho(X_1, X_2) & \rho(X_1, X_3) & \cdots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & 1 & \rho(X_2, X_3) & \cdots & \rho(X_2, X_n) \\ \rho(X_3, X_1) & \rho(X_3, X_2) & 1 & \cdots & \rho(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(X_n, X_1) & \rho(X_n, X_2) & \rho(X_n, X_3) & \cdots & 1 \end{pmatrix}.$$

La matriz de correlaciones es simétrica y positiva definida, además de que en la diagonal se encuentran 1's, esto se debe a que la correlación de cualquier variable aleatoria consigo misma es 1.

La matriz de varianzas-covarianzas de la cual haremos uso en el desarrollo de los siguientes capítulos.

**Definición 1.1.4.** Sean  $X_1, X_2, X_3, \dots, X_n$  variables aleatorias, la matriz de varianzas y covarianzas queda definida por:

$$\Sigma = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \text{cov}(X_1, X_3) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \text{cov}(X_2, X_3) & \cdots & \text{cov}(X_2, X_n) \\ \text{cov}(X_3, X_1) & \text{cov}(X_3, X_2) & \text{var}(X_3) & \cdots & \text{cov}(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \text{cov}(X_n, X_3) & \cdots & \text{var}(X_n) \end{pmatrix}.$$

Es importante mencionar que la matriz de varianzas-covarianzas, aunque no está normalizada, conserva las mismas propiedades que la matriz de correlaciones, por ejemplo que también es simetría y positiva definida.

Antes de mencionar cualquier otro punto, daremos un estimador puntual del coeficiente de correlación<sup>4</sup>:

$$\widehat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2 \sum_{k=1}^n (y_k - \bar{y})^2}}. \quad (1.1)$$

Por tanto, la estimación de la matriz de correlación muestral está formada por los estimadores del coeficiente de correlación entre cada variable.

Una vez que se ha definido la correlación es importante mencionar que en la mayoría de las ocasiones el investigador erróneamente concluye que existe una conexión causal a partir de la correlación, esta conclusión no es completamente cierta ya que la correlación se refiere a cómo están asociados, además de que, como ya se había mencionado anteriormente, la correlación es simétrica mientras que una relación causal no lo es. En otras palabras, la idea que se debe tomar es que la correlación no implica causalidad, ayuda a evaluarla, pero no la determina. Los requisitos adicionales para evaluar causalidad son: *dirección* de la asociación y el *aislamiento* de otras posibles causas; además la causalidad se determina a partir del conocimiento de área.

Con estas condiciones es posible conceptualizar a la causalidad como: “Una relación causal indica que todo cambio en una variable llamada *causa* refleja un cambio forzoso en la otra variable llamada *efecto*”.

## 1.2. Variables latentes

Uno de los elementos más importantes para los MEE es el planteamiento de conceptos que no pueden ser medidos directamente, pero que sabemos que tienen presencia en la investigación, a estos conceptos se les conoce como *variables latentes*. Estas variables desempeñan un papel importante en el desarrollo de los (MEE). En esta sección daremos una idea intuitiva sobre este tipo de variables y una definición de ellas. Más adelante con el desarrollo de este trabajo se darán elementos para que quede claro el concepto.

Un concepto que se puede dar sobre las *variables latentes*, de manera informal, es que son indicadores que no pueden ser medidos u observados directamente, pero influyen en los modelos estadísticos y son construidos a partir de variables medidas u observadas partiendo del Análisis de Factores Conformatorio (AFC). Éstas pueden definirse formalmente como:

**Definición 1.2.1.** *Las variables latentes son la herramienta a la que se acude cuando se tiene la hipótesis sobre la existencia de conceptos que no se pueden medir directamente sino a través de indicadores de ellos, es decir, de variables observadas*<sup>5</sup>.

Algunos de los ejemplos que se pueden dar para entender a las variables latentes son: nivel de inteligencia, calidad de vida, nivel de violencia y rendimiento escolar, entre otras.

<sup>4</sup>Alfonso De Lara (2005), *Medición y Control de Riesgos Financieros*, página 35.

<sup>5</sup>Ignacio Méndez Ramírez, *SEM 1 Introducción y Panorama*.

Hay que tener en cuenta que las variables latentes siempre serán causa de las variables observadas, sin embargo, las variables latentes siempre son estimadas a través de las variables observadas. Esta idea quedará más clara en el siguiente capítulo con el AFC.

Un ejemplo es: Supongamos que tenemos las evaluaciones de materias tomadas por un grupo de estudiantes de cierto nivel educativo, ¿qué factor o factores definen las calificaciones? Claro que hay distintos, están aquellos que se pueden medir y otros que no; hagamos énfasis en uno de ellos: la inteligencia. Este es un factor que no es posible medirlo de manera directa, es decir, no se tiene una unidad de medida de inteligencia. Sin embargo es posible cuantificarla ya que sabemos que es un factor que determina el aprendizaje de un alumno, es decir, las calificaciones se vuelven una consecuencia de la inteligencia y por tanto al medirlas también se está cuantificando la inteligencia.

Cabe señalar que a través del AFC es posible que las variables latentes puedan descomponerse en dos factores: los llamados *factores comunes*, son los que comparten efectos con más de una variable, y los llamados *factores específicos o errores*, son los que afectan sólo a una variable observable. Estos factores pueden producirse debido a un error en la medida de la variable observada o en el muestreo.

### 1.3. Diagrama de trayectorias (*Path Diagram*)

El *Diagrama de Trayectorias* o *Path Diagram* juega un papel elemental en el planteamiento de los MEE y AFC, ya que estas técnicas suelen convertirse en un conflicto considerable cuando el número de variables observadas y latentes son muchas y por tanto es necesario tener alguna representación gráfica del modelo. Como en la mayoría de las investigaciones se analiza un número considerable de variables, es común recurrir al planteamiento de manera visual, ya que éste es más sencillo y manipulable. Los MEE y AFC no son la excepción, en esta sección se darán la definición y las reglas de planteamiento de los Diagramas de Trayectorias, los cuales serán ejemplificados. Pero antes de esto, se darán algunas definiciones de los componentes de estos diagramas.

**Definición 1.3.1.** *El Diagrama de Trayectorias se define como la representación gráfica de las relaciones causales y de las correlaciones entre variables latentes y variables medidas.*<sup>6</sup>

Otra definición para los Diagramas de Trayectorias es:

**Definición 1.3.2.** *Los Diagramas de Trayectorias son como diagramas de flujo. En ellos se muestran las variables medidas y latentes interconectadas con líneas que son usadas para indicar el flujo causal.*

En el Diagrama de Trayectorias es posible clasificar las variables en dos tipos, *endógenas* y *exógenas*, las cuales definimos a continuación:

**Definición 1.3.3.** *Las variables endógenas son aquellas a las que le llega una flecha en un Diagrama de Trayectorias, es decir, son variables dependientes. Estas variables son afectadas por un término de perturbación aleatorio y que simplemente se incluye en el diagrama como una flecha adicional apuntando a la variable endógena.*<sup>7</sup>

<sup>6</sup>John Loehlin (1992), *Latent Variable Models. An Introduction to Factor, Path, and Structural Analysis*, página 2.

<sup>7</sup>Joan Batista Foguet y Germà Coenders Gallart (2000), *Modelos de ecuaciones estructurales. Modelos para el análisis de relaciones causales*, página 3.

**Definición 1.3.4.** *Las variables exógenas son aquellas a donde ninguna flecha llega en un Diagrama de Trayectorias, es decir, son variables independientes.*

Otra clasificación que se toma en cuenta en el Diagrama de Trayectorias es entre *variables observadas o medibles* y las *variables latentes*. Las variables medibles se representan por medio de un cuadro, mientras que las variables latentes se representan por medio de un círculo.

A continuación mencionaremos un conjunto de reglas sencillas para poder construir un Diagrama de Trayectoria. Estas reglas fueron obtenidas del texto de Batista et al. (2000).

1. La relación causal entre las variables se indica por una flecha cuyo origen es la variable causa y finaliza en la variable efecto. Cada flecha está afectada por un parámetro o coeficiente llamado “*path*”, que indica la magnitud del efecto entre ambas variables. Cuando entre dos variables no existe flecha alguna, el efecto entre éstas es cero.
2. La covarianza entre dos variables exógenas o dos términos de perturbación sin interpretación causal alguna se representa por una flecha bidireccional que las une en ambos sentidos. El parámetro asociado es el valor de la covarianza.
3. Cuando una teoría en estudio incluye tanto variables observadas como latentes, las primeras se representan en cuadros, y las segundas en círculos. Las relaciones causales y covarianzas mencionadas en los puntos anteriores se aplican también a las relaciones entre variables latentes. Las llamadas *relaciones de mediciones* conectan a las variables observadas con las latentes, con flechas unidireccionales con origen en las latentes. Las observadas están afectadas por un término de error aleatorio de medición que simplemente se incluye en el diagrama como una flecha adicional apuntando a la variable observable. La covarianza entre dos errores de medición se representa por una flecha bidireccional que los une en ambos sentidos.

Estas reglas permiten representar teorías causales y de medición de forma equivalente como hacen los sistemas de ecuaciones, siempre y cuando todas las relaciones causales propuestas sean lineales y se cumpla que:

1. Todas las relaciones causales estén representadas en el diagrama.
2. Todas las variables que sean causa de endógenas estén incluidas en el diagrama.
3. El diagrama sea lo más parco o sencillo posible y contenga solamente aquellas relaciones que puedan justificarse en bases teóricas o técnicas, depende del conocimiento del área por parte del investigador o su intuición.

## 1.4. Tipos de relaciones

Además de las reglas mencionadas anteriormente es importante especificar los tipos de relaciones causales que pueden existir entre dos o tres variables según sea el caso.

La representación de una relación entre dos variables se hará a través de una flecha bidireccional. El origen de la flecha parte de la variable causa para terminar en la variable dependiente. Las relaciones en que la variable causa no se muestra, es porque se refiere al término aleatorio que afecta a la variable dependiente.

Sean  $x$ ,  $y$  y  $z$ , tres variables medidas que definen las siguientes relaciones causales:

1. Las **relaciones directas** las podemos describir como aquellas relaciones en las que sólo participan dos variables. Donde una causa a la otra sin que haya algún otro factor. Suponiendo que  $x$  es causa de  $y$  podemos visualizar el ejemplo de esta relación en la figura 1.1. La flecha adicional apuntando a  $y$  indica que la variable  $y$  está afectada por un término de error aleatorio.

Este tipo de relaciones particularmente pueden ser *recíprocas*, es decir, que  $x$  cause  $y$  y  $y$  cause  $x$ , dicho de otra manera,  $x$  y  $y$  están correlacionadas. Este tipo de relaciones es posible visualizarlo en la figura 1.2. Además  $y$  está afectada por un error aleatorio (denotado por flecha adicional).

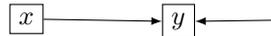


Figura 1.1: Relación directa,  $x$  es causa de  $y$ .

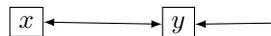


Figura 1.2: Relación recíproca.

2. Hay otro tipo de relación llamada **relación espúrea**, la cual se caracteriza por implicar a una tercera variable  $z$  tal que es una causa común de ambas. Un ejemplo de esta relación se puede visualizar en la figura 1.3, en donde  $x$  y  $y$  adicionalmente están afectadas por términos aleatorios (denotado por flechas adicionales).

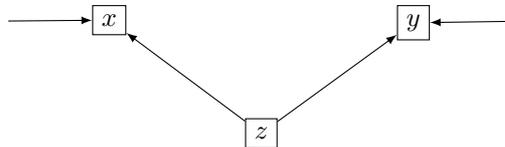


Figura 1.3: Relación espúrea.

3. Además de la relación espúrea, que involucra tres variables, también existe la **relación indirecta**, esta relación se caracteriza ya que  $x$  es causa de  $y$ , pero a través de una tercera variable  $z$ , la relación se ejemplifica en la figura 1.4. Además,  $y$  y  $z$  dependen de efectos aleatorios (flechas adicionales).
4. Además de estas tres relaciones existe una cuarta que se puede tomar como la combinación de las relaciones *indirecta* y *espúrea* ya que en esta relación  $x$  y  $z$  son ambas exógenas y carecen de mecanismo causal explícito que las relacione entre sí. No se especifica de qué tipo es la relación entre  $x$  y  $z$  y se deja la covariación entre ambas como no explicada. Además, su principal característica es la indeterminación que se refleja en la imposibilidad de discernir si  $z$  contribuye a la causalidad entre  $x$  y  $y$  por vía indirecta o espúrea, esta relación es posible visualizarla en la figura 1.5.

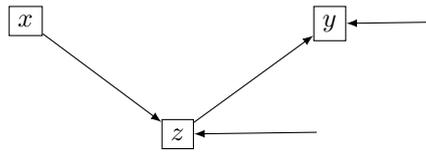


Figura 1.4: Relación indirecta.

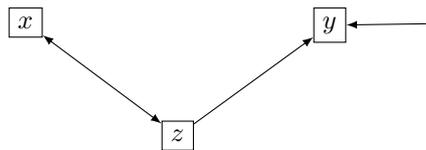


Figura 1.5: Combinación entre la relación indirecta y espúrea.



## Capítulo 2

# Análisis de factores confirmatorio

### 2.1. Introducción

En general un Análisis de Factores consiste en estudiar las variables latentes (conocidas también como factores) hipotéticas que pueden ser generadas por un conjunto de variables medidas.

Un *Análisis de factores Exploratorio* (AFE) consiste en determinar el número de factores que son necesarios para explicar de forma correcta las relaciones existentes, dado un conjunto de variables medidas.

El *Análisis de factores Confirmatorio* (AFC) es una técnica estadística cuyo objetivo es el de confirmar una hipótesis sobre los Factores o Variables Latentes a partir de un conjunto de variables medibles.

El AFC nos ayudará a entender los *Modelos de Ecuaciones Estructurales* (MEE), ya que la herramienta estadística que los resuelve es esencialmente la misma. Además, el AFC es un caso particular de los MEE. Esta afirmación podremos explicarla en el capítulo posterior donde estudiaremos los MEE.

Para entender de una manera más sencilla el AFC se expondrá un problema y a partir de él se planteará la parte teórica de esta técnica estadística.

Supongamos que se desea hacer un estudio con el objetivo de responder si la inteligencia determina el aprovechamiento académico de los estudiantes, es decir, a partir de las calificaciones obtenidas nos interesa determinar si la inteligencia se acumula en un único factor o en más factores, es decir, si hay uno o más tipos de inteligencia. Para comenzar con el desarrollo del problema vamos a suponer que tenemos un conjunto de datos con las calificaciones de 275 alumnos de las materias: *Lenguaje (L)*, *Filosofía (FSF)*, *Historia (H)*, *Matemáticas (M)*, *Física (FSC)* y *Química (Q)*. La tabla 2.1 muestra la matriz de correlaciones de las calificaciones antes mencionadas.

En caso de que no se cuente con información sobre las relaciones que hay entre las variables medibles (las calificaciones) y las latentes (el tipo de inteligencia), se tendrá que recurrir al Análisis de Factores Exploratorio (AFE) para determinar los factores que se podrían obtener partiendo de las variables medidas; ya que el objetivo del AFE es encontrar la agrupación que podrían formar las variables medidas. Ya que en esta tesis no se hablará del AFE, debemos tener un conocimiento previo sobre el tema que se trabajará.

	L	FSF	H	M	FSC	Q
$X_1 = L$	1					
$X_2 = FSF$	0.493	1				
$X_3 = H$	0.401	0.314	1			
$X_4 = M$	0.278	0.347	0.147	1		
$X_5 = FSC$	0.317	0.318	0.183	0.587	1	
$X_6 = Q$	0.284	0.327	0.179	0.463	0.453	1

Tabla 2.1: Matriz de correlaciones entre las calificaciones de los 275 alumnos.

Partiendo de literatura psicopedagógica y en algunos estudios realizados, sabemos que hay dos tipos de inteligencia denominadas como: *inteligencia cualitativa* ( $\xi_1$ ), que está enfocada a estudios como lenguaje, historia y filosofía, en otras palabras esta inteligencia se orienta a las humanidades; el otro tipo de inteligencia se denomina como *inteligencia cuantitativa* ( $\xi_2$ ), la cual se dirige a materias como matemáticas, física y química, es decir, a las ciencias exactas.

Entonces se deben plantear dos variables latentes:  $\xi_1$  y  $\xi_2$ . En la figura 2.1 se plantean las relaciones que hay entre las calificaciones obtenidas y los tipos de inteligencia.

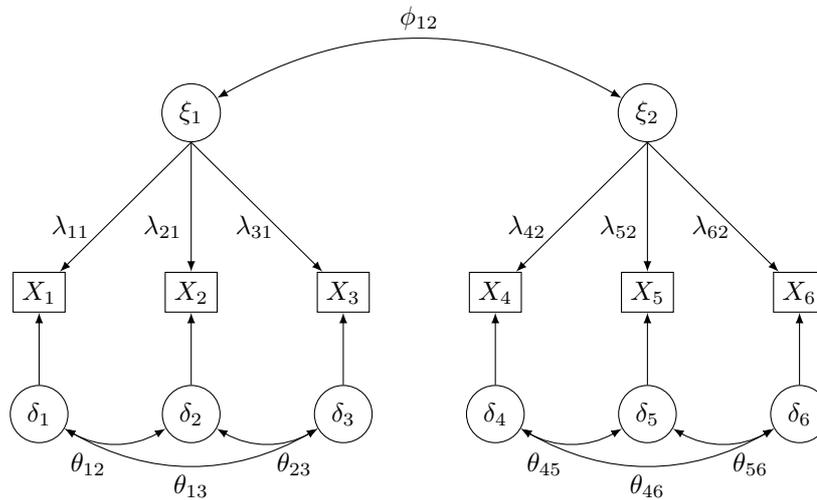


Figura 2.1: Modelo de análisis de factores confirmatorio (AFC).

Por otro lado, el investigador podría plantearse dos hipótesis: hay involucramiento entre los tipos de inteligencia o bien existe una inteligencia global que es causa de las dos mencionadas anteriormente. La primera hipótesis se infiere asociando las variables latentes con una covarianza ( $\phi_{12}$ ). En la segunda se debe plantear una tercera variable latente ( $\xi_3$ ), esto se puede visualizar en la figura 2.2.

Finalmente, para ambos modelos se cree que los errores ( $\delta_1, \dots, \delta_6$ ) de las variables medibles asociadas a un factor común están correlacionados. Para validar esto se estimarán las correlaciones entre los errores. En los diagramas de trayectorias se visualizan con  $\theta$ 's. En específico, como las variables medidas  $X_1$ ,  $X_2$  y  $X_3$  están asociadas al factor  $\xi_1$ , entonces sus correspondientes errores  $\delta_1$ ,  $\delta_2$  y  $\delta_3$  están correlacionados, y estas correlaciones se denotan como  $\theta_{12}$ ,  $\theta_{13}$  y  $\theta_{23}$ ,

respectivamente. Lo mismo sucede con las variables medidas  $X_4$ ,  $X_5$  y  $X_6$ , las cuales están asociadas al factor  $\xi_2$ , entonces sus correspondientes errores  $\delta_4$ ,  $\delta_5$  y  $\delta_6$  están correlacionados, y estas correlaciones se denotan como  $\theta_{45}$ ,  $\theta_{46}$  y  $\theta_{56}$ , respectivamente.

Además, en el modelo definido en la figura 2.2, se tienen las cargas  $\beta_{13}$  y  $\beta_{23}$  que relacionan los factores  $\xi_1$  y  $\xi_3$  y los factores  $\xi_2$  y  $\xi_3$ , respectivamente. Adicionalmente, los factores  $\xi_1$  y  $\xi_2$  dependen de los errores aleatorios  $\varsigma_1$  y  $\varsigma_2$ , respectivamente.

Ahora, el trabajo consiste en estimar las  $\lambda_{ij}$ , las cuales definimos como cargas de los factores. En el segundo modelo se puede notar que en lugar de establecer una relación de correlación se establece una relación de causalidad entre los tipos de inteligencia. En este cambio, el primer modelo deja de ser un modelo de AFC para pasar al segundo modelo que se convierte en un MEE por lo que estimarlo no es parte de este capítulo.

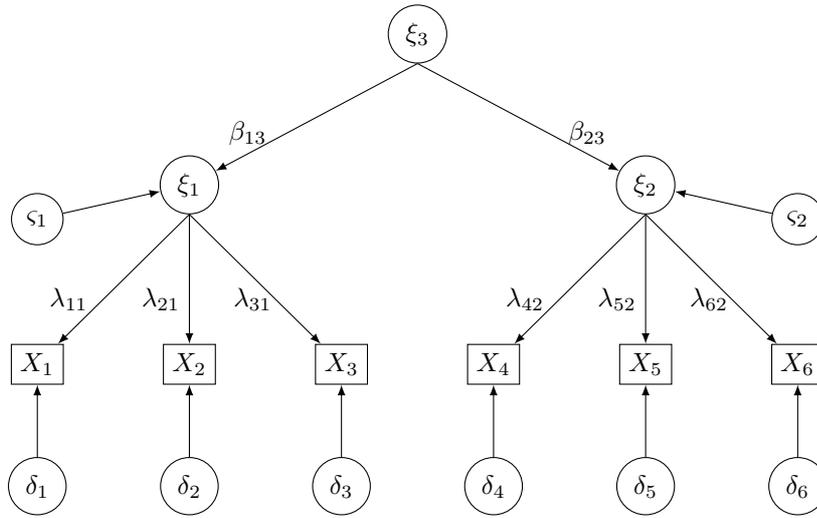


Figura 2.2: Modelo de estructural de covarianzas.

## 2.2. Planteamiento del modelo

El modelo matemático del AFC está dado por:

$$\mathbf{X} = \mathbf{\Lambda}\xi + \delta. \quad (2.1)$$

Entonces partiendo del modelo de la figura 2.1 plantearemos las ecuaciones que lo definen para que a partir de aquí se lleve a la forma matricial. La expresión quedará determinada por:

$$\begin{aligned} X_1 &= \lambda_{11}\xi_1 + \delta_1, & X_4 &= \lambda_{42}\xi_2 + \delta_4, \\ X_2 &= \lambda_{21}\xi_1 + \delta_2, & X_5 &= \lambda_{52}\xi_2 + \delta_5, \\ X_3 &= \lambda_{31}\xi_1 + \delta_3, & X_6 &= \lambda_{62}\xi_2 + \delta_6. \end{aligned} \quad (2.2)$$

La idea de plantear estas ecuaciones parte de la misma idea que en un modelo de regresión, es decir, la variable causa se iguala a la suma de carga del factor por el factor común (variable latente) más un error en la medición. Si recurrimos a la notación matricial se tiene la expresión:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \quad (2.3)$$

Ya que partimos de (2.3) para dar la expresión matricial en (2.1) pasaremos a definir los tamaños de cada matriz para tener la idea general del número de variables, cargas de los factores y errores que habrá en cualquier modelo.

1.  $\mathbf{X}$  es un vector  $q \times 1$ , donde  $q$  es el número de variables observadas.
2.  $\xi$  es un vector  $s \times 1$ , donde el modelo contiene  $s$  factores comunes o variables latentes.
3.  $\Lambda$  es una matriz  $q \times s$  que contiene las cargas de los factores de las variables latentes.
4.  $\delta$  es el vector  $q \times 1$  que contiene los errores o factores específicos de las variables medidas.

Una observación importante es que el número de variables medidas es mayor que el número de variables latentes ( $q > s$ ). Este principio parte de la idea que se desea reducir el número de variables o encontrar una dimensión menor.

Con la finalidad de tener expresiones más sencillas supondremos que las variables latentes, variables observadas y errores están expresadas como desviaciones sobre la media, es decir, la esperanza es cero, lo cual no afecta a las varianzas y covarianzas entre las variables. Matemáticamente lo anterior se expresa como:

$$E[\mathbf{X}] = 0, \quad E[\xi] = 0, \quad E[\delta] = 0. \quad (2.4)$$

Ahora, procederemos a calcular la varianza de  $\mathbf{X}$ , la cual quedará expresada por  $\Sigma$ , el cálculo estará dado por:

$$\Sigma = E[\mathbf{X}\mathbf{X}'] = E[(\Lambda\xi + \delta)(\Lambda\xi + \delta)']. \quad (2.5)$$

Teniendo en cuenta que la transpuesta de una suma de matrices es la suma de las matrices transpuestas y que la transpuesta de un producto es el producto de las matrices transpuestas en orden inverso, por tanto, de la expresión (2.5) tenemos que:

$$\Sigma = E[\mathbf{X}\mathbf{X}'] = E[(\Lambda\xi + \delta)(\xi'\Lambda' + \delta')]. \quad (2.6)$$

Ahora, vamos a tomar el argumento de las esperanzas, desarrollarlo y sintetizarlo, por tanto tenemos:

$$(\Lambda\xi + \delta)(\xi'\Lambda' + \delta') = \Lambda\xi\xi'\Lambda' + \Lambda\xi\delta' + \delta\xi'\Lambda' + \delta\delta'. \quad (2.7)$$

De nuevo volviendo a aplicar la esperanza a la expresión (2.7) tenemos:

$$E[\Lambda\xi\xi'\Lambda' + \Lambda\xi\delta' + \delta\xi'\Lambda' + \delta\delta'].$$

Ahora, dado que la esperanza es un operador lineal, es decir, abre sumas y saca escalares, tenemos que:

$$\begin{aligned} E[\mathbf{\Lambda}\xi\xi'\mathbf{\Lambda}'] + E[\mathbf{\Lambda}\xi\delta'] + E[\delta\xi'\mathbf{\Lambda}'] + E[\delta\delta'] \\ = \mathbf{\Lambda}E[\xi\xi']\mathbf{\Lambda}' + \mathbf{\Lambda}E[\xi\delta'] + E[\delta\xi']\mathbf{\Lambda}' + E[\delta\delta']. \end{aligned} \quad (2.8)$$

Es importante señalar que  $\mathbf{\Lambda}$  se toma como una matriz de escalares ya que son los parámetros del modelo AFC, en otras palabras, es el impacto que tienen las variables latentes sobre las variables medidas y se debe tener en cuenta que los parámetros estimados son variables aleatorias. Por otro lado, se tiene que la esperanza de un vector aleatorio multiplicado por su transpuesto es la varianza del vector (cuando el vector aleatorio tiene media cero), y la esperanza de la multiplicación de dos vectores aleatorios, uno de ellos transpuesto, es la covarianza de los vectores (cuando los vectores aleatorios tienen media cero). Por tanto tenemos de la expresión que:

$$E[\xi\xi'] = \Phi, \quad E[\delta\delta'] = \Theta. \quad (2.9)$$

Por otro lado, se asume que los factores comunes y específicos no están correlacionados, es decir:

$$E[\xi\delta'] = E[\xi'\delta] = \mathbf{0}.$$

Por lo tanto, por lo último mencionado y las expresiones (2.6) y (2.9) tenemos que la matriz de varianzas-covarianzas de las variables medidas es igual:

$$\Sigma = \mathbf{\Lambda}\Phi\mathbf{\Lambda}' + \Theta. \quad (2.10)$$

La expresión (2.10) es importante ya que a partir de ella se calcularán los estimadores del modelo y se sabrá si el modelo es identificable o no. Antes de pasar este tema mencionaremos algunos aspectos que deben ser considerados de la expresión (2.10):

1. La matriz  $\Sigma$  que es de tamaño  $q$  tiene  $\frac{q(q+1)}{2}$  elementos no redundantes, ya que es una matriz de varianzas-covarianzas, lo que implica que es una matriz simétrica con  $\frac{q(q+1)}{2}$  elementos posiblemente distintos.
2. La matriz  $\mathbf{\Lambda}$ , es de tamaño  $q \times s$ , la cual contiene las cargas de los factores del modelo AFC.
3. La matriz de varianzas-covarianzas de los factores comunes  $\Phi$ , tiene  $\frac{s(s+1)}{2}$  elementos no redundantes.
4. Para la matriz de varianzas-covarianzas  $\Theta$  de los factores específicos se tienen  $\frac{q(q+1)}{2}$  elementos no redundantes. Pero considerando los supuestos del modelo de factores, en donde los errores son independientes, esta matriz se reduce a una matriz diagonal.

Por lo tanto, se puede notar que los  $(\frac{q(q+1)}{2})$  elementos de la matriz  $\Sigma$  se expresan en función de  $[qs + \frac{s(s+1)}{2} + \frac{q(q+1)}{2}]$  parámetros de las matrices  $\Sigma$ ,  $\Phi$  y  $\Theta$ . De esta manera los parámetros que se deben estimar estarán vinculados mediante la expresión (2.10).

Por otro lado, recordemos que el modelo de la figura 2.2 tiene ciertos tipos de restricciones, tales como que algunas cargas de los factores son cero y la correlación entre algunos factores específicos

es cero, estas restricciones matemáticamente se expresan de la siguiente manera:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \underline{0} \\ \lambda_{21} & \underline{0} \\ \lambda_{31} & \underline{0} \\ \underline{0} & \lambda_{42} \\ \underline{0} & \lambda_{52} \\ \underline{0} & \lambda_{62} \end{bmatrix}, \mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \underline{0} & \underline{0} & \underline{0} \\ \theta_{12} & \theta_{22} & \theta_{23} & \underline{0} & \underline{0} & \underline{0} \\ \theta_{13} & \theta_{23} & \theta_{33} & \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} & \theta_{44} & \theta_{45} & \theta_{46} \\ \underline{0} & \underline{0} & \underline{0} & \theta_{45} & \theta_{55} & \theta_{56} \\ \underline{0} & \underline{0} & \underline{0} & \theta_{46} & \theta_{56} & \theta_{66} \end{bmatrix},$$

donde los ceros subrayados refieren a la especificación particular del modelo, por lo que si el planteamiento del modelo cambia, la configuración de las hipótesis cambiará.

### 2.3. Identificación del modelo AFC

Antes de estimar los parámetros del modelo es necesario detenerse un poco para observar lo que se denomina *identificación del modelo*. Note que la matriz de varianzas-covarianzas  $\mathbf{\Sigma}$  se obtiene a partir de las matrices  $\mathbf{\Lambda}$ ,  $\mathbf{\Phi}$  y  $\mathbf{\Theta}$ , la cual debe ser lo más aproximada a la matriz de varianzas-covarianzas muestral obtenida de las variables medidas, sin embargo, antes de pasar a la parte de *estimación* es necesario presentar la *identificación* del modelo.

De la misma manera que en un sistema de ecuaciones lineales, la existencia y número de soluciones está en función del número de incógnitas y ecuaciones que tiene el sistema. La *identificación* consiste en determinar el número de parámetros a estimar. La identificación implica determinar la existencia y unicidad de los parámetros planteados.

El problema de que un modelo no esté identificado impacta en los resultados de las estimaciones, los cuales impactan en interpretaciones carentes de sentido.

Las restricciones se establecen desde el planteamiento del modelo, cuando se decide cuáles cargas de factores son distintas de cero (qué relaciones existen) y qué factores comunes y específicos están correlacionados y cuáles no.

Existen versiones distintas sobre las restricciones que tienen que darse para que un modelo sea identificado, sin embargo, sólo se mencionarán algunas, las cuales deben ser validadas antes de estimar los parámetros:

1. Comparar el número de datos disponibles en la matriz de varianzas y covarianzas con el número de parámetros que han de estimarse. Los datos son siempre las varianzas-covarianzas muestrales, y hemos visto que existen  $\left(\frac{q(q+1)}{2}\right)$ . El número de parámetros a estimar es  $\left[qs + \frac{s(s+1)}{2} + \frac{q(q+1)}{2}\right]$ , donde  $s$  es el número de factores comunes. Por tanto se debe plantear, *al menos*  $qs + \left[\frac{s(s+1)}{2}\right]$  restricciones para que el sistema de ecuaciones formado para la estimación de los datos tenga solución. La diferencia entre el número de elementos distintos en la matriz de varianzas covarianzas y el número de parámetros que se estimarán son los grados de libertad.
2. Establecer una escala para los factores comunes. Esto se consigue poniendo restricciones acerca de las estimaciones de algunas cargas de los factores. En específico, se fija la varianza de cada factor común a un valor fijo o con el coeficiente de regresión (cargas de factores) de una de las variables observadas que cargan sobre cada factor a un valor fijo, por ejemplo 1. Si esto no se hace se produce el denominado problema de indeterminación entre la varianza

y las cargas de factores, es decir, no sería posible distinguir los valores entre las varianzas de cada factor y las cargas de los factores.

3. Asegurar la identificabilidad de la parte del modelo que contiene la relación entre las variables observadas y los factores. Para ello debe analizarse el número de factores y el número de variables observadas que cargan sobre cada factor. Si sólo hay un factor, el modelo puede estar identificado si el factor tiene al menos tres variables con cargas no nulas sobre él. Si hay dos o más factores, se examina el número de variables observadas de cada factor, si tiene tres o más variables que cargan sobre el factor, el modelo puede estar identificado si los errores asociados con los indicadores no están correlacionados entre sí, cada variable carga sólo sobre un factor y los factores pueden covariar entre ellos. Si sólo hay dos indicadores por factor, el modelo puede estar identificado si los errores asociados con cada indicador no están correlacionados, cada indicador carga sólo sobre un factor y ninguna de las covarianzas entre los factores es igual a cero.
4. Fijar arbitrariamente el coeficiente de regresión del término de error a un valor predeterminado, por ejemplo 1<sup>1</sup>.

Retomando el modelo de la figura 2.1 tenemos  $6(6 + 1)/2 = 21$  varianzas y covarianzas de las variables observadas y por tanto tenemos que estimar  $6 \times 2 + (\frac{2 \times 3}{2}) + (\frac{6 \times 7}{2}) = 36$  parámetros, los cuales se descomponen en 12 coeficientes de regresión (cargas de los factores), las varianzas de los 2 factores comunes, la covarianza entre ellos, 6 coeficientes de regresión entre las variables observadas y los factores específicos, las 6 varianzas de los factores específicos y las 15 covarianzas entre esos factores específicos.

Las cifras anteriores son brutas así que a medida que se establecen las restricciones antes mencionadas tendremos lo siguiente. El primer establecimiento será la escala de los factores comunes, para este modelo partiremos de las variables  $X_1$  y  $X_4$  donde fijaremos las cargas de factores a 1, es decir,  $\lambda_{11} = 1$  y  $\lambda_{42} = 1$ , y el resto de los valores los dejamos libres para ser estimados. Acto seguido es asegurar la identificabilidad del modelo en que participan las variables observadas y latentes. En este modelo existen dos factores comunes y por cada uno de ellos tres variables observadas. Por tanto, como se mencionó anteriormente, los errores de las variables  $\delta$  no están correlacionados, es decir  $\theta_{ij} = 0$ , y cada uno de ellos tienen un coeficiente entre ellos igual a 1. Por otra parte, la correlación entre los factores comunes sí está permitida y ya que la correlación es simétrica, i.e,  $\phi_{12}$ .

Una vez expuestas las restricciones estipuladas las preguntas son: ¿el modelo está identificado o sobre identificado? ¿puede ser sometido a contraste? ¿con cuántos grados de libertad se disponen? Respondiendo a cada uno de estos puntos, tenemos 21 datos provenientes de las variables observadas, mientras que los parámetros a estimar son:

- Un estimador de la covarianza de los factores comunes.
- 2 varianzas, una por cada factor.
- 4 coeficientes correspondientes a las cargas de los factores.
- 6 varianzas, una por cada error de las variables medibles.

---

<sup>1</sup>En el modelo de la figura 2.3, como puede notarse, los coeficientes correspondientes al término de error ( $\delta$ ) son 1. Sin embargo, en algunos *softwares* para el ajuste del AFC se permite fijar los coeficientes  $\delta$  a valores distintos de 1.

Por tanto, tenemos que el número de parámetros a estimar son 13, mientras que los datos son 21, por lo tanto tenemos 8 grados de libertad y por lo tanto el modelo está identificado y puede ser sometido a un contraste.

Las restricciones se hacen en medida de la importancia y coherencia que el modelo debe presentar, es decir, se restringen los parámetros que pueden ser sustituidos a través de otra explicación. Por ejemplo, se sustituyen las correlaciones entre los factores específicos por las correlaciones entre los factores comunes, ya que se puede notar que no tiene sentido volver a poner una explicación estadística dos veces. Por otro lado, los coeficientes del impacto entre los factores específicos y las variables observadas tienen que ser uno ya que es una hipótesis muy común que se trae desde el modelo de regresión lineal. Por último, fijamos al menos uno de los parámetros entre las variables latentes y las observadas bajo la misma idea de los factores comunes ya que son un caso particular de las variables latentes.

El diagrama presentado en la figura 2.3 contiene las restricciones que se mencionan anteriormente. En la figura, los parámetros que se igualan a un asterisco \* significa que son parámetros a estimar.

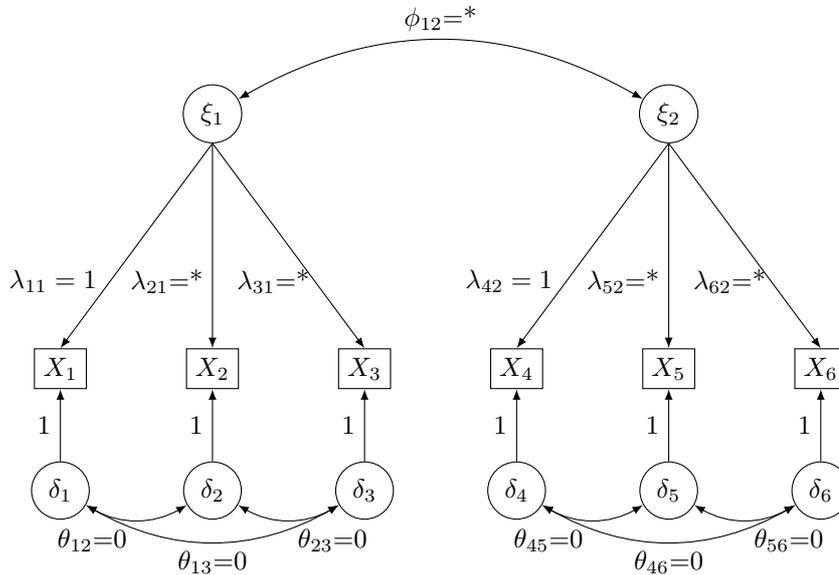


Figura 2.3: Modelo de AFC identificado.

Una vez que hemos mencionado que un modelo AFC debe tener restricciones para que éste pueda ser identificado pasaremos a la estimación de los parámetros que se deben estimar.

## 2.4. Estimación de parámetros

La idea general para la estimación de los parámetros del modelo AFC consiste en que la matriz de varianzas-covarianzas  $\Sigma$  que se define en (2.10) debería ser próxima a la matriz  $S$  formada por los elementos de (1.1). Recordando, la matriz  $\Sigma$  está expresada en términos de los parámetros que hay que estimar, por lo tanto, lo que buscamos es una expresión en la cual los argumentos sean  $\Sigma$  y  $S$  y podamos minimizar esta diferencia a través de esa expresión. De esta función se

obtendrá un conjunto de parámetros estimados que serán los elementos de la matriz  $\hat{\Sigma}$ .

En este trabajo presentaremos los distintos métodos de estimación, expresión matemática, cómo funcionan, sus ventajas y desventajas.

Retomando, la ecuación (2.10) la cual expresa el modelo que se desea estimar, tenemos:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta.$$

Se debe considerar que esta expresión es bruta, es decir, no está restringida y por lo tanto podría tener problemas de identificación del modelo. A continuación se presenta la expresión que denota a las matrices que contienen los parámetros que se desean estimar y además cumplen con las restricciones del modelo para que esté identificado:

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Phi}\hat{\Lambda}' + \hat{\Theta}. \quad (2.11)$$

Antes de estimarlos es necesario que la expresión (2.11) considere sólo aquellas matrices que cumplan con las restricciones impuestas por el investigador, estas matrices están denotadas por:  $\Lambda^*$ ,  $\Phi^*$  y  $\Theta^*$ , y permiten obtener una estimación de la matriz de varianzas-covarianzas poblacional  $\Sigma^*$  en la ecuación (2.11). Si esta matriz se aproxima de una manera adecuada a la matriz  $S$  diremos que las matrices  $\hat{\Lambda}^*$ ,  $\hat{\Phi}^*$  y  $\hat{\Theta}^*$  son consistentes con los datos de  $S$  y que los elementos de estas matrices pueden ser buenos estimadores de los parámetros del modelo.

La expresión mencionada se define como la *función de ajuste*, la cual será un indicador para saber qué tan próxima está  $\Sigma$  de  $S$ . Se denota a esta función como  $F(S; \Sigma)$  para todas las matrices  $\Sigma$  que cumplan con las restricciones para la identificabilidad del modelo.

En otras palabras, supongamos que las matrices  $\Sigma_1^*$  y  $\Sigma_2^*$  cumplen con las restricciones de identificabilidad del modelo, tales que  $F(S; \Sigma_1^*) < F(S; \Sigma_2^*)$ , entonces concluimos que  $\Sigma_1^*$  está más próxima a  $S$  que  $\Sigma_2^*$ , entonces las matrices  $\hat{\Lambda}^*$ ,  $\hat{\Phi}^*$  y  $\hat{\Theta}^*$  ayudan a minimizar la función de ajuste y por lo tanto son los estimadores de  $\Lambda^*$ ,  $\Phi^*$  y  $\Theta^*$ .

A continuación se describirán los siguientes métodos de estimación: *mínimos cuadrados no ponderados*, *mínimos cuadrados generalizados*, *máxima verosimilitud* y *estimación por la teoría de la distribución elíptica*.

### 2.4.1. Estimación por mínimos cuadrados no ponderados

La estimación por mínimos cuadrados no ponderados, o ULS por sus siglas en inglés (*Unweighted Least Squares*), se da cuando se minimiza la función:

$$F_{ULS}(S; \Sigma^*) = \frac{1}{2}tr[(S - \Sigma^*)^2], \quad (2.12)$$

donde  $tr$  representa la traza de una matriz, es decir, la suma de los elementos sobre la diagonal.

Las ventajas que se presentan en este método es que no es necesario suponer algún tipo de distribución probabilística por lo que cuando la hipótesis de normalidad es violada se puede recurrir a este método, pero se debe tener en consideración el problema de la dependencia de las unidades de medida, por lo tanto se deben tomar como datos de partida la matriz de varianzas-covarianzas estandarizada, es decir, la matriz de correlación.

Por otro lado, las desventajas consisten en que no es posible hacer contrastes estadísticos basados en este tipo de estimación y que los estimadores dependen de la escala de medida de las variables observadas, es decir, no alcanzaríamos el mínimo si tuvieramos diferentes unidades de medida, en caso de tener el mismo experimento y mismo modelo.

### 2.4.2. Estimación por mínimos cuadrados generalizados

La estimación por mínimos cuadrados generalizados, o GLS por su nombre en inglés (*Generalized Least Squares*), consiste en ponderar la diferencia de las matrices  $\Sigma^*$  y  $S$  respecto a la inversa de la matriz  $S$ , en otras palabras esto es:

$$F_{GLS}(S; \Sigma^*) = \frac{1}{2} tr[(S - \Sigma^*)S^{-1}]^2. \quad (2.13)$$

### 2.4.3. Estimación por máxima verosimilitud

Para usar este método de estimación las variables medidas deben seguir una función de densidad Normal. La función que debe ser minimizada para estimar los parámetros por el método de máxima verosimilitud (ML por sus siglas en inglés, *Maximum Likelihood*) es la siguiente:

$$F_{ML}(S, \Sigma^*) = tr(S\Sigma^{*-1}) + [\log |\Sigma^*| - \log |S|] - q, \quad (2.14)$$

donde el primer sumando de la expresión ya se ha explicado anteriormente, el segundo término consiste en restar los logaritmos naturales de los determinantes de las matrices de varianzas-covarianzas poblacional y muestral, y por último tenemos a  $q$  que es el número de variables medidas en el modelo. Cuando las matrices  $\Sigma^*$  y  $S$  sean muy parecidas entonces el producto de una de ellas y la inversa de la otra se aproximará a la matriz identidad; siendo su traza próxima a  $q$ . Por otra parte, los logaritmos de los determinantes se aproximarán a cero.

A continuación se expondrá una función más, que es poco conocida a nivel de aplicación ya que los softwares estadísticos no la utilizan.

### 2.4.4. Estimación por la teoría de la distribución elíptica

La estimación EDT, por su siglas en inglés (*Elliptical Distribution Theory*), se basa en la distribución de probabilidad de este nombre. La distribución normal multivariada es un caso particular de esta familia con parámetros de curtosis igual a cero. En este caso, la función a minimizar tiene la forma:

$$F_{EDT}(S; \Sigma^*) = \frac{1}{2}(k+1)^{-1} tr[(S - \Sigma^*)W^{-1}]^2 - \delta tr[(S - \Sigma^*)W^{-1}], \quad (2.15)$$

siendo  $k$  y  $\delta$  funciones de curtosis y de ponderación, y  $W$  es algún estimador consistente de  $\Sigma$ .

## 2.5. Pruebas de bondad de ajuste

En el siguiente paso en la investigación estadística del modelo AFC nos gustaría saber qué tanto se ajusta un modelo a los datos muestrales. En estadística, entre los modelos encontrados y los datos muestrales, siempre existen diferencias, las cuales son inevitables y esto es una situación que no podemos controlar. Sin embargo, lo que sí podemos controlar es qué grado de tolerancia tendrán estas diferencias.

A continuación se expondrán algunos indicadores del grado de ajuste del modelo propuesto y los contrastes que se deben realizar.

### 2.5.1. Matriz residual de covarianzas

Como se mencionó anteriormente la idea en general es encontrar  $\Sigma$  tal que sea lo más próxima a  $S$ , esto se refleja en que la diferencia de las matrices anteriores debería ser lo más cercana a cero, es decir, tener una matriz nula. La matriz obtenida de la diferencia de  $S$  y  $\Sigma$  se denomina *matriz residual de covarianzas*. Los valores de esta matriz deben ser pequeños ante un buen modelo propuesto, sin embargo, los residuos grandes asociados a algunos parámetros pueden indicar que han sido mal especificados, lo cual afectaría desfavorablemente al ajuste global del modelo.

La manera de interpretar la matriz residual consiste en observar en qué entrada se encuentran los mayores errores para analizar la relación causal que origina este residuo. Otra manera de validar esta matriz es obtener el promedio de la diagonal y evaluar si es alto o bajo. Esta decisión se deja al investigador, pues como ya se había mencionado, el margen de error depende del criterio del ajuste de la investigación que se está llevando a cabo.

### 2.5.2. Estadística Ji-cuadrada para el contraste global del modelo

Esta parte consiste en presentar las hipótesis a contrastar, los estadísticos de prueba y las regiones de rechazo. Como ya se ha descrito anteriormente,  $\Sigma$  es la matriz de varianzas-covarianzas del vector  $X$ , el cual está condicionado al modelo visto en (2.1), la estimación de esta matriz se ha denotado por  $\hat{\Sigma}$ , y el modelo está sujeto a un conjunto de restricciones denominando a  $\hat{\Sigma}^*$  la matriz que cumple estas condiciones. Ahora definiremos a  $\Sigma_{nc}$  la matriz no condicionada al modelo. En el caso de que el modelo sea adecuado para explicar el comportamiento de las variables medidas (u observadas) y las variables latentes, ambas matrices serán iguales, esto se traduce en la siguiente expresión.

$$H_0 : \Sigma_{nc} = \Sigma \quad \text{vs.} \quad \Sigma_{nc} \neq \Sigma. \quad (2.16)$$

La estadística de prueba está dada por:

$$y = N \times F_{ML}^0, \quad (2.17)$$

donde  $N$  es el número de datos y  $F_{ML}^0$  es el valor que toma la función de ajuste ML de la expresión (2.14), esta expresión se distribuye, bajo la hipótesis nula, como una Ji-cuadrada con  $\frac{1}{2}q(q+1) - k$  grados de libertad, tal que  $q$  representa el número de variables medidas y  $k$  el número de parámetros a estimar.

Se rechaza  $H_0$  si el  $p\text{-value} = p[Y \geq y]$  es menor o igual al nivel de significancia  $\alpha$ , o dicho de otra forma no se rechaza  $H_0$  si el  $p\text{-value}$  es mayor a  $\alpha$ , donde  $\alpha$  es la probabilidad de rechazar  $H_0$  cuando  $H_0$  es cierta, o de rechazar el modelo cuando el modelo es adecuado. Normalmente los softwares estadísticos presentan este valor denotado con una  $p$ .

### 2.5.3. Estadísticas *ad hoc*

Además del estadístico mencionado anteriormente, hay dos conjuntos de estadísticos conocidos como: los *índices comparativos de ajuste o de incremento* y los *índices de ajuste absoluto*.

El primer conjunto de índices comparan el modelo estimado contra el modelo independiente, es decir, aquel que no tiene relaciones entre las variables. Algunos de estos índices son: NFI, NNFI, CFI, IFI, MFI, AIC, entre otros.

La evaluación de los índices NFI, NNFI, CFI, IFI y MFI consiste en que deben estar lo más cercano a 1, o dicho de otra forma, si el analista está evaluando distintos modelos estimados, el mejor de ellos es aquel que su índice sea mayor al otro. Estos índices se encuentran en el intervalo (0,1) (Manzano, 2013). A continuación se da la expresión de ellos.

### Índice NFI

El índice NFI (*Normed Fit Index*) compara el valor del estadístico  $\chi^2$  del modelo teórico con el modelo independiente:

$$NFI = \frac{\chi_{indep}^2 - \chi_{nulo}^2}{\chi_{indep}^2}.$$

Este índice entre más cercano a 1 esté, el modelo estimado es más satisfactorio (Loehlin, 2004).

Algunos autores señalan que este índice tiende a subestimar el ajuste del modelo si las muestras son pequeñas, de tal manera que se presentan algunos índices modificados que solucionan estas situaciones. Estos índices se presentan a continuación.

### Índice NNFI

El *Nonnormed Fit Index* (NNFI) incorpora los grados de libertad de los modelos teórico e independiente. De esta manera se evita la subestimación del ajuste. El índice está dado por:

$$NNFI = \frac{\chi_{indep}^2 - \frac{gl_{indep}}{gl_{nulo}} \chi_{nulo}^2}{\chi_{indep}^2 - gl_{indep}}.$$

### Índice CFI

El índice CFI (*Comparative Fit Index*) corrige por el número de grados de libertad del siguiente modo:

$$CFI = \left[ \frac{(\chi_{indep}^2 - gl_{indep}) - (\chi_{nulo}^2 - gl_{nulo})}{(\chi_{indep}^2 - gl_{indep})} \right].$$

### Índice IFI

El índice IFI (*Incremental Fit Index*) pretende corregir la posibilidad de que el NNFI tome valores por encima del intervalo (0,1). Para ello se formula como:

$$IFI = \frac{\chi_{indep}^2 - \chi_{nulo}^2}{\chi_{indep}^2 - gl_{nulo}}.$$

### Índice MFI

El índice MFI (*McDonald Fit Index*) entraría en los denominados índices de ajuste absoluto; fue propuesto por McDonald y Marsh (1990). En contraposición a los anteriores que hemos denominado comparativos, por basarse en poner en relación el modelo teórico con el independiente. El

MFI sólo toma en consideración a la  $\chi^2$  del modelo teórico y responde a la siguiente expresión:

$$MFI = \exp \left\{ -\frac{1}{2} \frac{\chi_{nulo}^2 - gl_{nulo}}{N} \right\},$$

donde  $N$  es el tamaño de la muestra; el resto de los elementos ya definidos con anterioridad.

### Índice AIC

Este índice pertenece a un nuevo grupo llamado *índices de grado de parsimonia* ya que toman en cuenta no sólo la bondad de ajuste del estadístico sino también el número de parámetros a estimar. El índice *AIC* por sus siglas en inglés, *Akaike Information Criterion*, se define por:

$$AIC = \chi_{nulo}^2 - 2gl_{nulo} \quad (2.18)$$

Cuestión ahora ¿Cuál es el valor que debe tomar este índice? Debe ser lo *suficientemente pequeño*, sin embargo, como el índice no está ajustado a un intervalo (0,1), entendemos por suficientemente pequeño cuando se compara con otros modelos teóricos, es decir, este número nos ayuda a comparar modelos entre sí. Por tanto, se recomienda que el AIC se acompañe del modelo independiente, el cual es la base de cualquier modelo teórico. Entonces, cuanto mayor sea la diferencia entre el modelo teórico y el independiente el modelo teórico es mejor.

Los índices de ajuste absoluto miden directamente el ajuste del modelo, es decir, miden en qué grado el modelo estimado predice la matriz de varianzas covarianza observada. Estos índices normalmente se encuentran entre 0 y 1.

### Índice GFI

El índice GFI, por sus siglas en inglés *Goodness of Fit Index*, es una razón entre los elementos ponderados de la matriz de covarianzas poblacional estimada y los elementos ponderados de la matriz de covarianzas muestral estimada directamente. Concretamente, su expresión es la siguiente:

$$GFI = \frac{tr(\hat{\sigma}'W\hat{\sigma})}{tr(\hat{s}'W\hat{s})},$$

donde el vector  $\hat{\sigma}$  contiene las varianzas de la matriz de varianzas-covarianzas estimada y el vector  $s$  las de la matriz muestral. La matriz  $W$  es una matriz de ponderación que varía en función del método de estimación elegido: la matriz identidad en el *ULS*, la matriz de covarianzas estimada en el *ML*, etc.

### Índice AGFI

El índice AGFI por sus siglas en inglés *Adjusted Goodness of Fit Index* es una corrección del anterior que se hace en función del número de parámetros que se han de estimar ( $k$ ) y el número de datos disponibles ( $d$ ). El índice se define de la siguiente manera:

$$AGFI = 1 - \frac{1 - GFI}{1 - k/d}.$$

### Índice CAIC

El *Consistent AIC* (CAIC) es la corrección al AIC. Su expresión es la siguiente:

$$CAIC = \chi_{nulo}^2 - (\log N + 1)gl_{nulo}.$$

### Índice RMR

Este índice está basado en los residuos, es decir, los promedios en las diferencias de las varianzas y covarianzas muestrales y las estimadas que se derivan del modelo. El índice está definido por:

$$RMR = \sqrt{\frac{\sum_{i=1}^q \sum_{j=1}^q (s_{ij} - \sigma_{ij})^2}{q(q+1)/2}}.$$

Dado que los residuos están afectados por la escala en que se miden las variables se recomienda usar el SRMR, el cual es la estandarización del RMR y toma valores en el intervalo (0,1) y lo recomendable es que tome valores inferiores a 0.05.

### Indicador RMSEA

RMSEA (*Root Mean Square Error Approximation*) por sus siglas en inglés, es un indicador que estima la escasez en el modelo ajustado comparado a una perfecta saturación del modelo (Ullman, 2006); éste se puede interpretar como el error de aproximación medio por grado de libertad.

El RMSEA es una medida no central que depende del tamaño de la muestra y los grados de libertad del modelo estimado (Browne & Cudeck, 1993; Steiger, 1990) y se calcula como la diferencia entre estos. Sí la diferencia es menor a cero, el RMSEA se toma como cero. La expresión del RMSEA está dado por:

$$RMSEA = \sqrt{\frac{\chi_{nulo}^2 - gl_{nulo}}{(N-1)gl_{nulo}}}. \quad (2.19)$$

Si el indicador toma valores menores a 0.05 indica que el modelo tiene un buen ajuste, lo deseable es que el indicador se acerque a 0.

El investigador debe hacer una revisión de todos los indicadores ya que no hay una regla en que se mencione aquellos estadísticos en los que el investigador debe enfocarse para determinar si el modelo tiene un buen ajuste. Algunos autores mencionan (Mejía y Cornejo, 2010) que los indicadores que más se emplean son: GFI, SRMR, AGFI, CFI y RMSEA.

## 2.6. Reespecificación del modelo

Antes de proceder de aplicar alguna reespecificación del modelo se debe tener presente que existen dos causas principales para hacerlas: (i) mejorar el ajuste y (ii) contrastar alguna hipótesis teórica. Esto se menciona ya que hacer reespecificaciones puede generar resultados que reflejen un panorama y no una situación general. Estadísticamente para sustentar estas reespecificaciones

existe el *multiplicador de Lagrange* y el *contraste Wald* que nos indican qué relaciones causales pueden añadirse o eliminarse respectivamente<sup>2</sup>.

Es científicamente incorrecto modificar un modelo simplemente por mejorar su ajuste, ya que el cambio debe ser sustentado a través de un marco teórico en el cual sea posible justificar la modificación y que ésta sea interpretable. Finalmente, antes de pasar a describir los contrastes, planteamos las siguientes recomendaciones extraídas del trabajo de McCallum, Roznowski y Necowitz (1992):

1. Utilizar muestras grandes. Los modelos basados en pocas observaciones (digamos, algunos autores considera que en menos de 150 casos) llevan a modelos finales poco estables si las modificaciones se basan en los datos y no en la teoría.
2. Hacer pocas modificaciones. Es posible que las primeras modificaciones pueden estar derivadas de un modelo que refleje las relaciones poblacionales; mientras que las siguientes, una vez aplicadas las anteriores podrían reflejar relaciones específicas de la muestra.
3. Realizar sólo aquellos cambios que puedan ser interpretados desde una perspectiva teórica o tengan soporte en trabajos precedentes. En todo caso, se deben detallar todos los cambios realizados sobre el modelo inicial en el informe del trabajo final.
4. Seguir un procedimiento paralelo de especificación. Siempre que sea posible, el investigador debería trabajar con dos muestras independientes. Si las dos muestras desembocan en las mismas modificaciones del modelo, se podrá tener una mayor confianza en la estabilidad del mismo.
5. Comparar modelos alternativos desde el principio. Más que proponer un modelo e ir modificándolo, puede ser conveniente en algunas ocasiones plantear modelos alternativos y determinar con cuál se obtiene un mejor ajuste.
6. Finalmente, describir detalladamente las limitaciones del modelo. La mayoría de los trabajos que se publican están basados en una única muestra sobre los que se efectúan sucesivas modificaciones basadas en los datos hasta lograr un ajuste razonable. Si se sigue este enfoque, sería recomendable que en el trabajo se mencionaran todas estas circunstancias.

### 2.6.1. Contraste del multiplicador de Lagrange

El contraste del multiplicador de Lagrange permite evaluar la mejora que se obtiene al añadir una relación causal o una nueva covarianza al modelo teórico, es decir, toma aquellas relaciones que no definió el investigador y estima el valor que tomaría la carga de los factores en caso de incluir la relación.

Adicional a esto el *Multiplicador de Lagrange* da el valor *p-value* de la relación propuesta. Este *p-value* está asociado a una prueba de hipótesis nula que consiste en que la carga de los factores del factor común  $j$  no es relevante sobre la variable medible  $i$ , contra la hipótesis alternativa que la carga de los factores del factor común  $j$  es relevante sobre la variable medible  $i$ , es decir:

$$H_0 : \lambda_{ij} = 0 \quad vs. \quad H_a : \lambda_{ij} \neq 0$$

---

<sup>2</sup>Algunos autores señalan que el AFC no tiene un carácter tan estrictamente *confirmatorio* (Arbuckle, 2000; Cribbie, 2007; Hancock, 1999; Loehlin, 2004) y que podrían tomarse como exploratorios en aquellos casos en que la teoría no es suficientemente sólida.

El estadístico sigue una distribución Ji-cuadrada con un grado de libertad,  $\chi_{(1)}^2$ , por lo que se rechaza  $H_0$  si  $T \geq w_{1-\alpha}$ , donde  $T$  es la estadística de prueba,  $w_{1-\alpha}$  es el cuantil  $1 - \alpha$  de una distribución  $\chi_{(1)}^2$ , y  $\alpha$  es el nivel de significancia de la prueba.

Una vez que se han definido qué parámetros son candidatos a formar parte del modelo inicial, es preciso tener una aproximación del valor de la estimación que tomará el coeficiente de la nueva relación. Ya que se tiene una propuesta de la relación que se debe incluir y la aproximación que tomará la estimación, ahora es trabajo del investigador analizar si existe una justificación teórica o empírica para añadir la relación propuesta. Para apoyar su criterio el investigador puede ayudarse del posible valor que puede tomar la estimación.

Existen dos versiones del contraste del multiplicador de Lagrange, la versión univariada y multivariada. Las diferencias radican en que en la univariada se muestra el incremento que se produce en la estadística de prueba si se introdujera, por separado, cada una de las relaciones consideradas. Esta aproximación tiene un problema, como no considera las covarianzas entre los distintos parámetros estimados, harán que aparezcan como candidatos a añadirse al modelo y mejorar su ajuste. Por otra parte, el enfoque multivariado, identifica el parámetro que conduciría a un mayor incremento en la estadística de prueba del modelo. Después de tener en consideración su introducción, se muestra el parámetro cuya inclusión causaría el siguiente mayor incremento.

### 2.6.2. Contraste de Wald

Mientras que el contraste de Lagrange plantea si se deberían añadir nuevas relaciones, el contraste Wald se orienta a resolver el problema si se deberían eliminar algunas relaciones existentes, en otras palabras, confirmar la validez del modelo. La eliminación o integración de parámetros podría afectar la significancia que hay en un principio, por lo que se recomienda primero añadir los parámetros que indiquen los multiplicadores de Lagrange y después, sobre este modelo modificado, hacer el contraste de Wald para eliminar los innecesarios.

El estadístico que se usa para este contraste queda determinado de la siguiente manera:

$$cw = \frac{(\theta - \hat{\theta})^2}{SE^2}, \quad (2.20)$$

donde  $\theta$  denota el parámetro de interés,  $\hat{\theta}$  el estimador máximo verosímil de  $\theta$  y  $SE$  es el error estándar de este estimador máximo verosímil.

Además este estadístico sigue la distribución  $\chi_{(1)}^2$  y la hipótesis que se desea verificar es:

$$H_0 : \theta = 0 \quad vs. \quad H_a : \theta \neq 0 .$$

Por lo tanto, la regla de decisión queda determinada por:

$$\text{Se rechaza } H_0 \text{ si: } cw > w_{1-\alpha}, \quad (2.21)$$

donde  $w_{1-\alpha}$  es el cuantil  $1 - \alpha$  de una distribución Ji-cuadrada con un grado de libertad, tal que,  $P(X \geq w_{1-\alpha}) = 1 - \alpha$ , donde  $X \sim \chi_{(1)}^2$ .

Por último, tanto el contraste de Lagrange como el de Wald son procedimientos en los cuales el error tipo I suele sobreestimarse. Por esta razón se recomienda ser conservadores con el nivel de significancia considerado. Por ejemplo, para Lagrange sería  $\alpha = 0.01$  y para Wald  $\alpha = 0.05$ .

## 2.7. Paquete sem (software R)

A continuación se presenta el uso del paquete `sem` del software R para el ajuste de los modelos AFC. Las funciones que se usan se mencionarán a medida en que se vayan utilizando para el ajuste del ejemplo acerca de la inteligencia y calificaciones de seis materias, presentado en las secciones anteriores, cuyo modelo AFC identificado se presenta en la figura 2.3.

El primer paso es llamar el paquete `sem`:

```
library(sem)
```

Definimos los nombres de las variables observadas:

```
name.ej = c("L", "FSF", "H", "M", "FSC", "Q")
```

Definir las desviaciones estándar de los datos:

```
sd.ej <- c(1.090, 0.590, 0.980, 1.100, 0.410, 1.110)
```

Definir la matriz de correlaciones, la función `readMoments` permite introducir matrices de forma triangular inferior, como se muestra a continuación:

```
cor.ej <- readMoments(diag=TRUE, names=name.ej)
1
0.493 1
0.401 0.314 1
0.278 0.347 0.147 1
0.317 0.318 0.183 0.587 1
0.284 0.327 0.179 0.463 0.453 1
```

Ahora podemos calcular la matriz de varianzas-covarianzas,  $cov(X_i, X_j) = cor(X_i, X_j) \times sd(X_i) \times sd(X_j)$ :

```
cov.ej = matrix(0,6,6, dimnames=list(name.ej,name.ej))
for(i1 in 1:6){
  for(i2 in 1:i1){
    cov.ej[i1,i2] = cor.ej[i1,i2]*sd.ej[i1]*sd.ej[i2]
    cov.ej[i2,i1] = cov.ej[i1,i2]
  }
}
```

Definir el modelo AFC identificado, para esto existen dos funciones que nos permiten definir el modelo: `specifyModel` y `specifyEquations`.

La forma en que trabaja la función `specifyModel` es a través de tres entradas por cada renglón. La primera consiste en definir la relación entre las variables que se especifican en esa línea (sólo se puede trabajar con dos variables por línea) y estas relaciones son de causalidad o correlación varianza o covarianza. La forma de declarar las relaciones es similar a como se hace en el *diagrama de trayectorias*, es decir, de la variable causa sale una flecha `->` que llega a la variable efecto, mientras que para determinar las varianzas o covarianzas la flecha debe ser bidireccional `<->`. La

segunda entrada especifica si se va a calcular el parámetro o no, en caso de que se estime deberá ir el nombre que se le asignará al parámetro, y en caso de que no sea calculado se debe poner un NA. La tercera y última entrada consiste en que, si se va a estimar el parámetro de la relación, entonces se debe poner un NA, pero en caso de no estimarlo se debe poner el valor de la constante al que se va a fijar (generalmente se iguala a 1). Cabe mencionar que la determinación de estimar o dejar fijo el valor de los parámetros se establece cuando se hace la *Identificación del Modelo*.

```

modelo.ej <- specifyModel()
XI1 -> L, NA, 1, ### Relaciones Variables-Factores
XI1 -> FSF, lam21, NA,
XI1 -> H, lam31, NA,
XI2 -> M, NA, 1,
XI2 -> FSC, lam52, NA,
XI2 -> Q, lam62, NA,
L <-> L, the11, NA, ### Varianzas Observaciones
FSF <-> FSF, the22, NA,
H <-> H, the33, NA,
M <-> M, the44, NA,
FSC <-> FSC, the55, NA,
Q <-> Q, the66, NA,
XI1 <-> XI1, phi11, NA, ### Varianzas Factores
XI2 <-> XI2, phi22, NA,
XI1 <-> XI2, phi12, NA, ### Covarianzas Factores

```

La función `specifyEquations` permite definir de otra manera el modelo AFC identificado, en la cual se definen las ecuaciones que relacionan las variables observadas y los factores (como en la expresión (2.3)), las varianzas (a través de  $V$ ) y covarianzas (a través de  $C$ ) de las variables observadas y factores:

```

modelo.ej <- specifyEquations()
L = 1*XI1 ### Relaciones Variables-Factores
FSF = lam21*XI1
H = lam31*XI1
M = 1*XI2
FSC = lam52*XI2
Q = lam62*XI2
V(L) = the11 ### Varianzas Observaciones
V(FSF) = the22
V(H) = the33
V(M) = the44
V(FSC) = the55
V(Q) = the66
V(XI1) = phi11 ### Varianzas Factores
V(XI2) = phi22
C(XI1,XI2) = phi12 ### Covarianzas Factores

```

Estimar el modelo AFC de los 275 alumnos, la función que usa el paquete `sem` es la misma para ajustar los modelos AFC y MEE (porque los modelos AFC son un caso particular de los MEE):

```
afc.ej <- sem(modelo.ej, cov.ej, 275)
```

Los resultados del ajuste se pueden obtener con `summary(afc.ej)`:

```

Model Chisquare = 8.842   Df = 8   Pr(>Chisq) = 0.3558
Chisquare (null model) = 392.82   Df = 15
Goodness-of-fit index = 0.98898
Adjusted goodness-of-fit index = 0.97108
RMSEA index = 0.019599   90% CI: (NA, 0.075256)
Bentler-Bonnett NFI = 0.97749
Tucker-Lewis NNFI = 0.99582
Bentler CFI = 0.99777
SRMR = 0.030809
AIC = 34.842
AICc = 10.237
BIC = 81.86
CAIC = -44.092

```

#### Normalized Residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.14000	-0.21300	0.00000	0.00154	0.20500	1.29000

#### R-square for Endogenous Variables

L	FSF	H	M	FSC	Q
0.5347	0.4744	0.2413	0.5766	0.5777	0.3778

#### Parameter Estimates

	Estimate	Std Error	z value	Pr(> z )	
lam21	0.509834	0.068272	7.4676	8.1644e-14	FSF <--- XI1
lam31	0.604047	0.095627	6.3167	2.6723e-10	H <--- XI1
lam52	0.373072	0.039403	9.4681	2.8497e-21	FSC <--- XI2
lam62	0.816837	0.095506	8.5528	1.2017e-17	Q <--- XI2
the11	0.552848	0.088252	6.2644	3.7420e-10	L <--> L
the22	0.182979	0.025109	7.2872	3.1640e-13	FSF <--> FSF
the33	0.728614	0.070855	10.2832	8.3895e-25	H <--> H
the44	0.512313	0.075028	6.8283	8.5954e-12	M <--> M
the55	0.070994	0.010426	6.8096	9.7863e-12	FSC <--> FSC
the66	0.766588	0.079401	9.6547	4.6958e-22	Q <--> Q
phi11	0.635252	0.116748	5.4412	5.2911e-08	XI1 <--> XI1
phi22	0.697687	0.111734	6.2442	4.2607e-10	XI2 <--> XI2
phi12	0.387961	0.067905	5.7133	1.1082e-08	XI2 <--> XI1

Iterations = 71

Es habitual obtener la solución estandarizada del AFC, esto es, aquella en que se recalculan los estimadores para asegurar que las varianzas de los factores comunes y de las variables observadas son igual a la unidad. A continuación se muestra el código en R para obtener los parámetros estimados<sup>3</sup>:

```
stdCoef(afc.ej)
```

<sup>3</sup>VF hace referencia al valor fijo para la identificación del modelo.

		Std. Estimate			
1	VF	0.732	L	<---	XI1
2	lam21	0.688	FSF	<---	XI1
3	lam31	0.492	H	<---	XI1
4	VF	0.759	M	<---	XI2
5	lam52	0.760	FSC	<---	XI2
6	lam62	0.615	Q	<---	XI2
7	the11	0.465	L	<-->	L
8	the22	0.525	FSF	<-->	FSF
9	the33	0.758	H	<-->	H
10	the44	0.423	M	<-->	M
11	the55	0.422	FSC	<-->	FSC
12	the66	0.622	Q	<-->	Q
13	phi11	1.000	XI1	<-->	XI1
14	phi22	1.000	XI2	<-->	XI2
15	phi12	0.582	XI2	<-->	XI1

Como se puede observar, los parámetros que en un principio se habían fijado a 1 ahora están estandarizados. Los resultados se muestran en la figura 2.4.

Revisando el ajuste del modelo, se tiene que: El *p-value* asociado a la prueba Ji-Cuadrada es mayor a 0.05, lo que significa que no se rechaza  $H_0$ . Los índices GFI, AGFI, NFI, NNFI y CFI son mayores a 0.9 y además se encuentran muy próximos a la unidad, tal como deben ser ante un buen modelo. Finalmente, el índice RMSEA es menor a 0.05. Por lo tanto, podemos concluir qué modelo tiene un buen ajuste global.

En general, para todos los parámetros estimados el *p-value* muestra ser mucho menor a 0.05 por lo tanto, estadísticamente son distintos de cero.

Revisando el diagrama podemos darnos cuenta que la inteligencia cuantitativa tiene mayor impacto en las habilidades Literarias mientras que la inteligencia cualitativa impacta casi en misma proporción en las habilidades de Matemáticas y Física

Como puntos complementarios a los resultados anteriores en R es posible obtener un conjunto de atributos del modelo, el software enlista estas propiedades con `attributes(afc.ej)` y lo que se obtendrá es información adicional del modelo: **S** matriz de varianzas-covarianzas observada, **C** matriz de varianzas-covarianzas obtenidas por el modelo, **A** matriz  $\Lambda$ , **P** matriz  $\Theta$ , entre otros.

Otra función importante dentro del paquete `sem` es `modIndices`, ésta permite obtener una estadística de prueba usando los multiplicadores de Lagrange para fijar y restringir parámetros en el modelo.

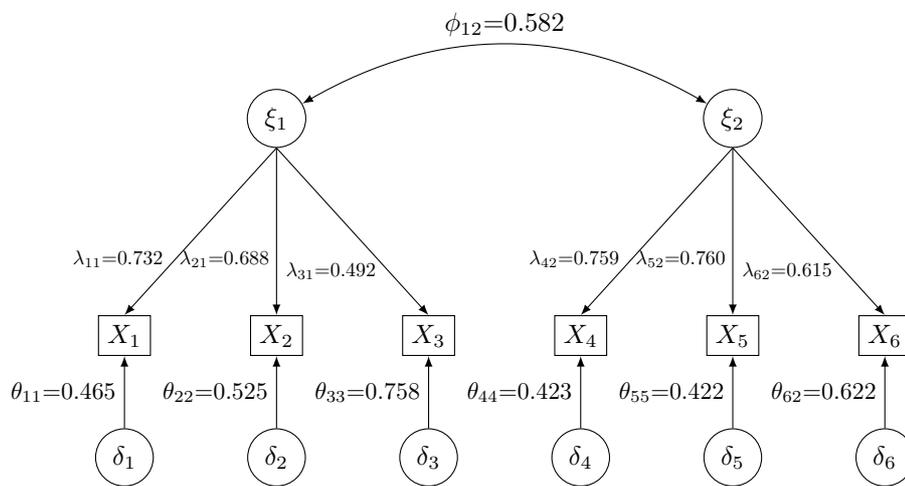


Figura 2.4: Modelo de AFC estimado.



## Capítulo 3

# Modelos de ecuaciones estructurales

### 3.1. Introducción

#### 3.1.1. Aplicaciones

Los modelos de ecuaciones estructurales (MEE) tienen un gran número de aplicaciones en distintas disciplinas como psicología, medicina, econometría, sociología, finanzas, investigación de mercados, entre otras, debido a la posibilidad de plantear más de una relación de dependencia y la complejidad del modelo ya que las variables dependientes también se convierten en independientes. Los MEE cuentan con la habilidad de incorporar variables latentes y variables observadas y como consecuencia tener los errores de medida y de predicción.

Como se mencionó en el primer capítulo se debe tener claro que la existencia de correlación entre dos variables no implica necesariamente la existencia de una relación causal. Sin embargo, la existencia de una relación causal sí implica la presencia de correlación. Por lo tanto, la correlación juega un papel muy importante en el desarrollo de los MEE ya que ayuda a la evaluación de las relaciones que se requieren validar.

Ahora bien, técnicamente los MEE son una extensión del AFC y por tanto la mayoría de los conceptos teóricos ya han sido definidos, entonces en este capítulo habrá elementos que sólo se mencionarán como referencia, y en otros sí habrá un desarrollo más profundo.

#### 3.1.2. ¿Qué son los MEE?

*“Los Modelos de Ecuaciones Estructurales (MEE) tienen como objetivo explicar las relaciones entre un conjunto de variables observadas basándose en las relaciones de un número más pequeño de variables latentes. Estas relaciones se caracterizan mediante la covarianza de las variables observadas. Por tanto, se tiene que estos modelos son una herramienta poderosa para el estudio de las relaciones causales en algún área de interés”.*

Las estimaciones de los parámetros de las relaciones, varianzas y covarianzas, índices de bondad de ajuste y estadísticos de reespecificación son heredados del AFC, por lo que el desarrollo técnico

para los MEE conserva la misma idea.

De manera formal es posible definir los MEE de la siguiente manera:

**Definición 3.1.1.** *Los MEE son un sistema de variables cuya organización y relación se establece previamente a partir de una o más hipótesis. En estos modelos puede manejarse dos o más variables dependientes, cuya combinación según un esquema teórico contribuyen a confirmar las relaciones de dependencia causal entre estas variables; además, la técnica es capaz de sugerir relaciones que los datos respaldan como significativas con el objetivo de obtener un mejor modelo.*

Es importante aclarar que, por lo general, estos modelos tienen un carácter confirmatorio, así que, aunque se puede hacer un análisis para mejorar el modelo, esto corresponde a la aplicación del modelo, no a los objetivos de MEE.

La diferencia entre un AFC y un MEE es: el AFC sólo confirma la veracidad de variables latentes que son efecto directo de las variables observadas, mientras que un MEE extiende su capacidad a validar las relaciones planteadas entre las variables latentes.

Para entender el concepto y el planteamiento matemático de esta técnica se expone un ejemplo a continuación.

### 3.1.3. Componentes de los MEE

A continuación se enuncia un ejemplo (Uriel y Manzano, 2005) basado en la investigación de Bagozzi (1980) en su trabajo con título *Performance and Satisfaction in an Industrial Sales Force: An Examination of Their Antecedents and Simultaneity*. Tomamos el ejemplo para mostrar cómo desarrollar un MEE, mostrar el código de desarrollo en R, interpretar los indicadores de bondad de ajuste y para interpretar el modelo. Los resultados mostrados no deberían ser tomados para hacer conclusiones verídicas.

**Ejemplo 3.1.1.** *Entre los directores de una fuerza de venta existe el convencimiento de que hay una relación entre el desempeño, o eficacia de los vendedores en su trabajo (variable latente  $\eta_1$ ) y la satisfacción que éstos declaran tener con su empleo (variable latente  $\eta_2$ ). Pero no hay una clara idea sobre si el desempeño es causa de la satisfacción o la satisfacción es causa del desempeño. Asimismo, la influencia de una variable latente sobre la otra puede no estar causada por la relación causal directa entre ambas, sino deberse a la existencia de unos antecedentes comunes que afectan a los individuos y que se sintetiza en: motivación para el logro ( $\xi_1$ ), es decir, el valor que cada individuo da a las recompensas que puede recibir por hacer un buen trabajo, la autoestima ( $\xi_2$ ), es decir, el buen concepto que uno tiene de sí mismo y que si considera que puede influir sobre el buen desempeño de un trabajo, y la inteligencia verbal ( $\xi_3$ ), esto es, la habilidad cognitiva para percibir fácilmente los contenidos de las conversaciones, las instrucciones escritas y otras formas de comunicación, lo que hará que se pueda reaccionar adecuadamente ante las objeciones de los clientes y organizar correctamente las tareas de ventas. El diagrama de trayectoria que representa las relaciones mencionadas está dado por la figura 3.1.*

El ejemplo que se plantea, a diferencia del análisis de regresión (incluso del múltiple), es que en esta técnica sí se permite la multicolinealidad<sup>1</sup>; esta característica se puede notar al relacionar las variables latentes  $\xi_1$ ,  $\xi_2$  y  $\xi_3$ , y al tomar en cuenta que sí se pueden correlacionar ya sea de manera directa o indirecta a través de las variables latentes dependientes  $\eta_1$  y  $\eta_2$ .

<sup>1</sup>Se define como la dependencia lineal entre las variables regresoras, Montgomery et al. (2007).

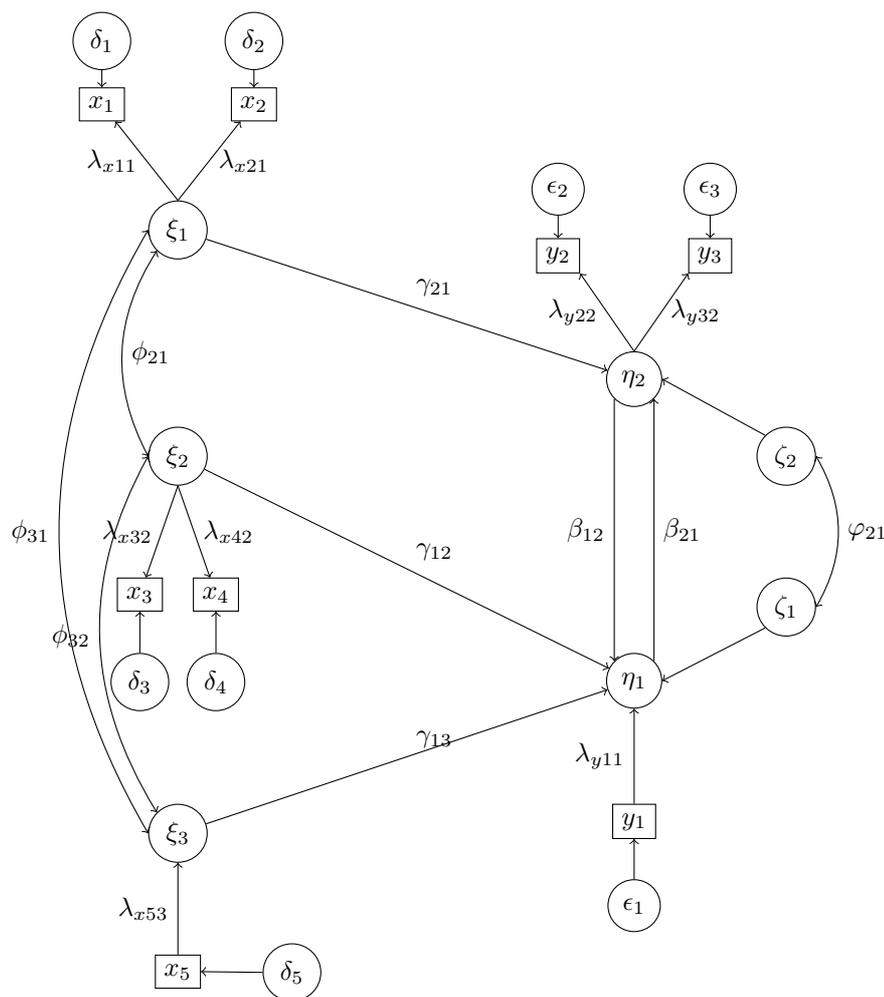


Figura 3.1: Modelo de desempeño y satisfacción de la fuerza de ventas.

Otra parte sobresaliente de los MEE es que las relaciones van sobre las variables medidas y variables latentes, y no sólo sobre las medidas, esto es con el objetivo de disminuir la dimensionalidad de los datos y/o enfocarse sobre variables que no se pueden medir directamente pero que se sabe de su existencia. Sin embargo, no sólo es posible estudiar las relaciones entre variables latentes, puede ser que un conjunto de variables latentes dependan de variables medidas, que un conjunto de variables medidas dependa de una o más variables latentes, o en un caso extremo, que sólo se desee inferir relaciones entre variables medidas, en este caso se estaría aplicando un *Análisis de Trayectoria*.

También es posible identificar la parte del AFC, en la cual se infieren las variables latentes a partir de las variables medidas, esto puede verse en la figura 3.2, lo que vendría siendo el *modelo de medida*.

Por otro lado, se presentan las relaciones que se forman con las variables latentes, esto es posible visualizarlo en la figura 3.3, el cual corresponde al *componente estructural* del modelo.

De manera general se define el *Componente Estructural* como aquel que incluye las múltiples relaciones que existen sólo entre las variables latentes y, el *Modelo de Medida* es aquel en que las variables observadas actúan como medida de las latentes para su construcción.

En los MEE hay dos tipos de errores para cada uno de los componentes del modelo, para el componente estructural se asocia el *error de predicción* que como su nombre lo dice: el error contiene aquellos efectos en las variables dependientes que no son predichos por las variables independientes. Por otra parte están los *errores de medición* que son reflejo de una deficiente recolección de datos debida al muestreo, sesgos en las respuestas o cualquier otro motivo que hagan que las medidas sean imperfectas, es por eso que actúan en las variables medidas, para dar solución a estos casos.

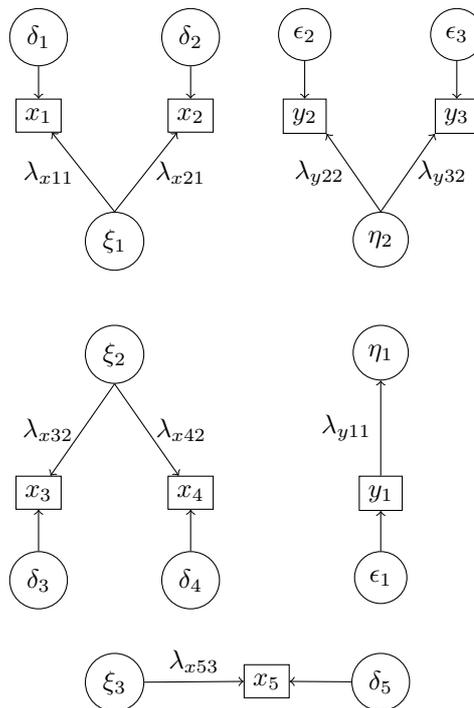


Figura 3.2: Componente medible del modelo de desempeño y satisfacción de la fuerza de ventas.

### 3.2. Representación matemática del modelo

Esta sección comienza planteando las ecuaciones que resuelven el modelo del ejemplo mencionado. Lo primero que se plantea es el modelo de medida. Es posible notar que en el *diagrama de trayectoria* las variables observadas se denotan por  $x$ 's y  $y$ 's. Las  $x$ 's definen a las variables latentes que han de ser independientes (las  $\xi$ 's), y las  $y$ 's definen aquellas variables latentes que

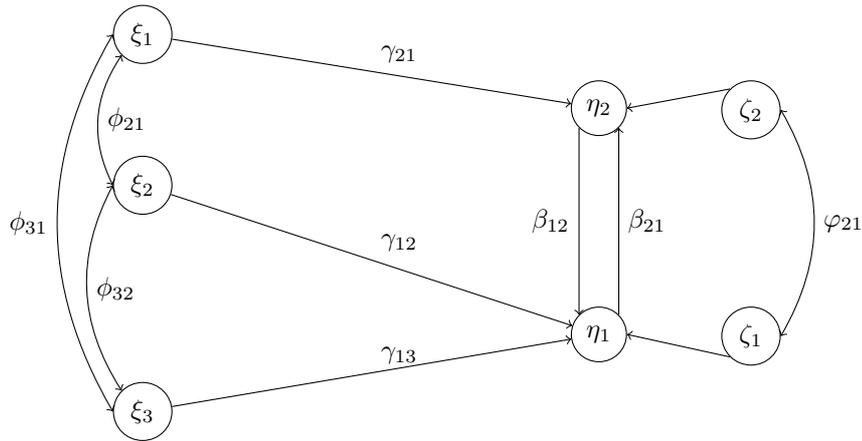


Figura 3.3: Componente estructural del modelo de desempeño y satisfacción de la fuerza de ventas.

han de ser dependientes (las  $\eta$ 's). Las ecuaciones que representan el modelo son:

$$\begin{aligned} x_1 &= \lambda_{x11}\xi_1 + \delta_1, & y_1 &= \lambda_{y11}\eta_1 + \epsilon_1, \\ x_2 &= \lambda_{x21}\xi_1 + \delta_2, & y_2 &= \lambda_{y22}\eta_2 + \epsilon_2, \\ x_3 &= \lambda_{x32}\xi_2 + \delta_3, & y_3 &= \lambda_{y32}\eta_2 + \epsilon_3. \\ x_4 &= \lambda_{x42}\xi_2 + \delta_4, \\ x_5 &= \lambda_{x53}\xi_3 + \delta_5, \end{aligned}$$

Para el componente estructural las ecuaciones del ejemplo están dadas por:

$$\begin{aligned} \eta_1 &= \gamma_{12}\xi_2 + \gamma_{13}\xi_3 + \beta_{12}\eta_2 + \varsigma_1, \\ \eta_2 &= \gamma_{21}\xi_1 + \beta_{21}\eta_1 + \varsigma_2. \end{aligned}$$

Para el componente de medida es aceptable dividirlo en dos: una parte para las variables que van a los factores latentes independientes (serían las  $x$ 's) y otra hacia los dependientes (serían las  $y$ 's), entonces a partir de esto la notación matricial quedará de la siguiente manera:

El componente de medida que compone los factores independientes es:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} \lambda_{x11} & 0 & 0 \\ \lambda_{x21} & 0 & 0 \\ 0 & \lambda_{x32} & 0 \\ 0 & \lambda_{x42} & 0 \\ 0 & 0 & \lambda_{x53} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix}.$$

Para los factores dependientes, el componente de medida queda determinado por:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{y11} & 0 \\ 0 & \lambda_{y22} \\ 0 & \lambda_{y32} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}.$$

Y finalmente el componente estructural se reescribe como:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \varsigma_1 \\ \varsigma_2 \end{bmatrix}.$$

Una vez que se ha presentado la forma matricial de manera particular, podemos hacer la formalización matemática en notación matricial, la cual quedará de la siguiente manera:

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \quad (3.1)$$

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad (3.2)$$

$$\boldsymbol{\eta} = \mathbf{\Gamma} \boldsymbol{\xi} + \mathbf{B} \boldsymbol{\eta} + \boldsymbol{\varsigma}. \quad (3.3)$$

Por tanto, tenemos que las ecuaciones (3.1), (3.2) y (3.3) son las que representan en forma genérica un MEE. En particular, para el ejemplo, los tamaños de las matrices quedan de la siguiente manera:  $X_{5 \times 1}$ ,  $\Lambda_{x5 \times 3}$ ,  $\xi_{3 \times 1}$ ,  $\delta_{5 \times 1}$ ,  $Y_{3 \times 1}$ ,  $\Lambda_{y3 \times 2}$ ,  $\eta_{2 \times 1}$ ,  $\epsilon_{3 \times 1}$ ,  $\Gamma_{2 \times 3}$ ,  $B_{2 \times 2}$  y  $\varsigma_{2 \times 1}$ .

La ecuación (3.1) plantea los factores latentes que se convierten en independientes en el componente estructural, mientras que (3.2) plantea a los factores latentes que han de ser los dependientes, y por último (3.3) plantea el componente estructural, i.e., las relaciones causales entre variables latentes. Otra forma de ver esta última ecuación es:

$$\boldsymbol{\eta} - \mathbf{B} \boldsymbol{\eta} = \mathbf{\Gamma} \boldsymbol{\xi} + \boldsymbol{\varsigma}.$$

Factorizando  $\boldsymbol{\eta}$  se tiene que:

$$(\mathbf{I} - \mathbf{B}) \boldsymbol{\eta} = \mathbf{\Gamma} \boldsymbol{\xi} + \boldsymbol{\varsigma}.$$

Sea  $\ddot{\mathbf{B}} = (\mathbf{I} - \mathbf{B})$  tal que la expresión anterior se puede escribir como:

$$\ddot{\mathbf{B}} \boldsymbol{\eta} = \mathbf{\Gamma} \boldsymbol{\xi} + \boldsymbol{\varsigma},$$

que también se puede expresar como:

$$\boldsymbol{\eta} = \ddot{\mathbf{B}}^{-1} \mathbf{\Gamma} \boldsymbol{\xi} + \ddot{\mathbf{B}}^{-1} \boldsymbol{\varsigma}. \quad (3.4)$$

La ecuación (3.4) es adecuada para deducir resultados e interpretar, dado que un valor negativo en  $\ddot{\mathbf{B}}$  indica relaciones positivas entre los factores dependientes, mientras que un valor positivo en  $\ddot{\mathbf{B}}$  indicaría relaciones negativas.

Se debe tener claro acerca de las variables que participan en un MEE, el número de elementos que tiene cada matriz y el vector que representa a cada variable. La tabla 3.1 tiene la información necesaria y de manera resumida lo que se debe tener presente.

Una vez presentado el modelo matemático que representa a los MEE, es momento de dar paso a la estadística y establecer las hipótesis del modelo, las cuales se mencionan a continuación:

1. Los factores comunes y específicos están medidos en desviaciones respecto a su media, por lo que:

$$\mathbf{E}[\boldsymbol{\eta}] = \mathbf{E}[\boldsymbol{\xi}] = \mathbf{E}[\boldsymbol{\varsigma}] = \mathbf{0}.$$

Esto no afecta la generabilidad del modelo puesto que los parámetros estructurales contenidos en  $\mathbf{B}$  y  $\mathbf{\Gamma}$  no están afectadas por esta hipótesis.

Matriz	Dim.	Varianzas	Descripción
$\eta$	$r \times 1$	$\Phi_\eta = \mathbf{E}[\eta\eta']$	Matriz de factores dependientes
$\xi$	$s \times 1$	$\Phi_\xi = \mathbf{E}[\xi\xi']$	Matriz de factores independientes
$\varsigma$	$r \times 1$	$\Psi = \mathbf{E}[\varsigma\varsigma']$	Errores de predicción
$\mathbf{B}$	$r \times r$		Efectos directos de $\eta$ sobre $\eta$
$\ddot{\mathbf{B}}$	$r \times r$		$(\mathbf{I} - \mathbf{B})$
$\Gamma$	$r \times s$		Efectos directos de $\xi$ sobre $\eta$
$\mathbf{x}$	$q \times 1$	$\Sigma_{\mathbf{xx}} = \mathbf{E}[\mathbf{xx}']$	Variables observadas independientes
$\mathbf{y}$	$p \times 1$	$\Sigma_{\mathbf{yy}} = \mathbf{E}[\mathbf{yy}']$	Variables observadas dependientes
$\Lambda_{\mathbf{x}}$	$q \times s$		Cargas de los factores de $\mathbf{x}$ sobre $\xi$
$\Lambda_{\mathbf{y}}$	$p \times r$		Cargas de los factores de $\mathbf{y}$ sobre $\eta$
$\delta$	$q \times 1$	$\Theta_\delta = \mathbf{E}[\delta\delta']$	Errores de medición para las variables $\mathbf{x}$
$\epsilon$	$p \times 1$	$\Theta_\epsilon = \mathbf{E}[\epsilon\epsilon']$	Errores de medición para las variables $\mathbf{y}$

Tabla 3.1: Definición por variable.

2. Los términos de error de los factores dependientes y los factores independientes no están correlacionados entre sí, es decir, del modelo estructural, los factores específicos y comunes no están correlacionados. Matemáticamente se expresa como:

$$\mathbf{E}[\xi'\varsigma] = \mathbf{E}[\xi\varsigma'] = \mathbf{0}.$$

3. Suponemos que  $\ddot{\mathbf{B}}$  es no singular<sup>2</sup>, lo cual es una condición para que (3.4) tenga solución.  
4. Las variables están medidas en desviaciones sobre la media, por lo que:

$$E[\mathbf{x}] = E[\delta] = E[\mathbf{y}] = E[\epsilon] = \mathbf{0}.$$

5. Los factores y los términos de error que se presentan en una relación no están correlacionados, es decir:

$$E[\xi\delta'] = E[\delta\xi'] = E[\eta\epsilon'] = E[\epsilon\eta'] = \mathbf{0}.$$

6. Los términos de error de las variables observadas están no correlacionados entre sí, en otras palabras:

$$E[\delta\epsilon'] = E[\epsilon\delta'] = \mathbf{0}.$$

### 3.3. Identificación del modelo

#### 3.3.1. Tipos de parámetros

Para entender la identificación del modelo se mencionarán primero los tipos de parámetros que hay en un MEE; los hay de tres tipos: *libre*, *fijos* y *de restricción*. Los *libres* son aquellos que su valor depende del proceso de estimación y hacen referencia a las varianzas y covarianzas de las variables independientes, los coeficientes de las relaciones causales entre las variables latentes y

<sup>2</sup>Una matriz singular es aquella que no tiene matriz inversa.

las cargas de los factores. Los *fixos* son aquellos que no participan en el proceso de estimación ya que sus valores se fijan a una constante. Finalmente tenemos los de *restricción* que son aquellos sobre los que se expresa una conjetura acerca de los valores; esencialmente estas presunciones se establecen en términos de una hipótesis, por lo que se igualan a un valor particular (cero por ejemplo), o se asume que son iguales a otro u otros parámetros del modelo.

La idea de la identificación de un MEE se hereda de la técnica del AFC, el cual en esencia es la misma, sólo que en el MEE se complica debido a que el número de relaciones es mucho mayor y por lo tanto los parámetros a estimar son mucho más.

### 3.3.2. Restricciones del modelo

En el MEE se tiene un sistema de ecuaciones como en álgebra lineal y de la misma manera debe haber restricciones para que la estimación del modelo sea posible.

En los modelos estudiados existen dos tipos de restricciones que son: las *necesarias* y las que son *necesarias y suficientes*. Las necesarias consisten en hacer las siguientes validaciones:

1. La matriz de información, que se determina a través de la segunda derivada con respecto a los parámetros que se han de estimar de la función de ajuste, debe ser definida positiva. Si el modelo está identificado, el rango<sup>3</sup> de la matriz mencionada debe ser igual al número de parámetros libres del modelo. Esto es equivalente a la comprobación del modelo de regresión lineal, que el rango de la matriz de varianzas-covarianzas de los regresores debe ser igual al número de regresores. Algunos autores ponen en tela de juicio que esta condición se clasifique como necesaria, sin embargo, los paquetes estadísticos hacen el cálculo para esta restricción.
2. Otro enfoque supone evaluar el Jacobiano (derivadas de primer orden de la función de ajuste con respecto a los parámetros libres), demostrando que en caso de ser definida positiva el modelo estaría identificado. Este enfoque aún no ha sido implementado en los programas estadísticos y es por ello no aplicable.

La primera condición se evalúa con una regla conocida como la *Regla t*. Para aplicar esta regla se necesita saber el número de parámetros libres, es decir, aquellos que han de estimarse y se debe tener claro el número de variables observadas con que se cuenta para poder determinar el número de elementos distintos en la matriz de varianzas-covarianzas. Sea  $t$  el número de parámetros a estimar,  $q$  el número de variables observadas ( $x$ 's) que determinan las variables latentes independientes y  $p$  el número de variables observadas ( $y$ 's) que determinan las variables latentes dependientes; entonces el modelo es estimable sí se cumple lo siguiente:

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

Algunos autores (Manzano y Zamora 2013) consideran esta condición simplemente como necesaria más no suficiente mientras que Uriel y Joaquín (2005) la consideran como necesaria y suficiente.

Por otro lado se tienen las condiciones/restricciones *necesarias y suficientes*, a las cuales se recurre más, ya que el modo de empleo es práctico y fácil. Algunos autores las clasifican por tipo y algunos simplemente los mencionan sin clasificar.

<sup>3</sup>El rango de una matriz es el número de columnas linealmente independientes.

- Se debe establecer la escala de los factores dependientes e independientes, esto se consigue fijando a un valor arbitrario la carga de los factores (usualmente se fija  $\lambda_{ij} = 1$ ) o fijando la varianza factorial a un valor fijo (usualmente igual a uno). Este último caso sólo aplica para los factores independientes, ya que en los factores dependientes la varianza no es un parámetro a estimar directamente. NO ENTIENDO COMO CAMBIARLA
- Para un modelo que tiene un solo factor común, la condición necesaria para la identificabilidad del modelo consiste en que haya al menos tres variables observadas que carguen sobre él, estas cargas deben ser distintas a cero y que los errores asociados no estén correlacionados.
- Si existe más de un factor común entonces se debe identificar cuántas variables observadas cargan por factor para establecer las restricciones. En caso de que haya dos variables por cada factor entonces se debe aplicar la *Regla de dos variables*, la cual indica que para que un modelo esté identificado debe suceder: (i) los errores asociados con cada indicador no deben estar correlacionados ( $cov(\delta_i, \delta_j) = cov(\epsilon_i, \epsilon_j) = 0$ ), (ii) cada variable sólo carga sobre un factor (esta restricción en la presentación matricial se visualiza en la matriz de cargas en la que por cada renglón sólo debe haber un elemento distinto de cero), y (iii) no se permite el supuesto de que haya factores comunes que no se correlacionen, es decir, no se permite fijar a cero las covarianzas entre factores. En caso de que haya tres o más variables por factor las restricciones quedan determinadas por la regla conocida como: *Regla de tres variables observadas*, la cual es muy similar a la regla anterior ya que las condiciones (i) y (ii) son iguales. La diferencia estriba en que los factores comunes tienen libertad de correlacionar o no, lo cual depende de los criterios que establezca el investigador. En ambas reglas las cargas de los factores deben ser distintas de cero. Por otro lado, la siguiente regla comienza con la cuestión ¿Qué sucede cuando sólo se tiene una sola variable observable que carga sobre un factor? ¿Cómo se interpreta? No hay un nombre conocido para la regla, pero ésta consiste en que la variable observable sólo debe cargar sobre un factor y las varianzas de los términos de error se deben fijar a cero.

De esto se deriva que los grados de libertad se obtienen de la diferencia del número de elementos diferentes en la matriz de varianzas-covarianzas de las variables observadas y de los parámetros que se pretenden estimar, esto se expresa en la ecuación:

$$gl = \frac{(p+q)(p+q+1)}{2} - t.$$

### 3.4. Estimación

De la misma manera que se estimaron los parámetros en el AFC, a través de la matriz de varianzas-covarianzas teórica y de los datos con la minimización de la función objetivo, se hará en los MEE. Sin embargo, en MEE no es tan evidente el planteamiento de la matriz de varianzas-covarianzas, ya que en el AFC sólo se tenía una ecuación medible donde se encuentran todos los parámetros que se han de estimar, y en el MEE hay tres ecuaciones, es decir, está formado por dos ecuaciones medidas y una ecuación estructural. Por tal motivo ahora los parámetros no están contenidos en una sola ecuación, sino en tres. Por tanto, se debe partir de una expresión que contenga todos los parámetros que se necesiten estimar. Si sustituimos (3.4) en (3.2) se tiene:

$$\mathbf{y} = \Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\zeta) + \epsilon. \quad (3.5)$$

Ahora la información de los parámetros del componente estructural se contempla en la ecuación medible referente a las variables latentes endógenas. El siguiente paso es capturar los elementos de las dos componentes medidas y esto se obtiene con la matriz de varianzas-covarianzas de  $\mathbf{x}$  y  $\mathbf{y}$ , es decir, debemos calcular:

$$\Sigma_{\mathbf{y}\mathbf{x}} = \Sigma_{\mathbf{x}\mathbf{y}}^t = \begin{bmatrix} var(\mathbf{y}) & cov(\mathbf{y}, \mathbf{x}) \\ cov(\mathbf{x}, \mathbf{y}) & var(\mathbf{x}) \end{bmatrix} \quad (3.6)$$

La tarea ahora es calcular cada uno de los elementos de la matriz antes mencionada. Partiendo de (3.5) el cálculo de  $var(\mathbf{y})$  está dado por:

$$\begin{aligned} var(\mathbf{y}) &= E[\mathbf{y}\mathbf{y}^t] - E[\mathbf{y}]E[\mathbf{y}^t] = E[\mathbf{y}\mathbf{y}^t] - 0 \\ &= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon]^t \right\} \\ &= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma)^t\Lambda_y^t + \epsilon^t] \right\} \\ &= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][((\ddot{B}^{-1}\Gamma\xi)^t + \varsigma^t(\ddot{B}^{-1})^t)\Lambda_y^t + \epsilon^t] \right\} \\ &= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][(\xi^t\Gamma^t(\ddot{B}^{-1})^t + \varsigma^t(\ddot{B}^{-1})^t)\Lambda_y^t + \epsilon^t] \right\} \\ &= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t + \epsilon^t] \right\} \\ &= E \left\{ [\Lambda_y\ddot{B}^{-1}\Gamma\xi][\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t + \epsilon^t] \right\} \\ &\quad + E \left\{ [\Lambda_y\ddot{B}^{-1}\varsigma][\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t + \epsilon^t] \right\} \\ &\quad + E \left\{ [\epsilon][\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t + \epsilon^t] \right\} \\ &= E[\Lambda_y\ddot{B}^{-1}\Gamma\xi\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t] + E[\Lambda_y\ddot{B}^{-1}\Gamma\xi\varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t] + E[\Lambda_y\ddot{B}^{-1}\Gamma\xi\epsilon^t] \\ &\quad + E[\Lambda_y\ddot{B}^{-1}\varsigma\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t] + E[\Lambda_y\ddot{B}^{-1}\varsigma\varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t] + E[\Lambda_y\ddot{B}^{-1}\varsigma\epsilon^t] \\ &\quad + E[\epsilon\xi^t\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t] + E[\epsilon\varsigma^t(\ddot{B}^{-1})^t\Lambda_y^t] + E[\epsilon\epsilon^t] \end{aligned} \quad (3.7)$$

usando las propiedades de la esperanza,

$$\begin{aligned} var(\mathbf{y}) &= \Lambda_y\ddot{B}^{-1}\Gamma E[\xi\xi^t]\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \Lambda_y\ddot{B}^{-1}\Gamma E[\xi\varsigma^t](\ddot{B}^{-1})^t\Lambda_y^t + \Lambda_y\ddot{B}^{-1}\Gamma E[\xi\epsilon^t] \\ &\quad + \Lambda_y\ddot{B}^{-1} E[\varsigma\xi^t]\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \Lambda_y\ddot{B}^{-1} E[\varsigma\varsigma^t](\ddot{B}^{-1})^t\Lambda_y^t + \Lambda_y\ddot{B}^{-1} E[\varsigma\epsilon^t] \\ &\quad + E[\epsilon\xi^t]\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + E[\epsilon\varsigma^t](\ddot{B}^{-1})^t\Lambda_y^t + E[\epsilon\epsilon^t] \end{aligned}$$

De acuerdo a las hipótesis planteadas anteriormente sobre la correlación entre los errores de medición y las variables latentes, los únicos elementos que sobreviven de (3.7) son:

$$var(\mathbf{y}) = \Lambda_y\ddot{B}^{-1}\Gamma\Phi_\xi\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t + \Lambda_y\ddot{B}^{-1}\Psi_\varsigma(\ddot{B}^{-1})^t\Lambda_y^t + \Theta_\epsilon \quad (3.8)$$

Por lo tanto, factorizando algunos términos de la expresión (3.8) se tiene que:

$$var(\mathbf{y}) = \Lambda_y\ddot{B}^{-1}(\Gamma\Phi_\xi\Gamma^t + \Psi_\varsigma)(\ddot{B}^{-1})^t\Lambda_y^t + \Theta_\epsilon \quad (3.9)$$

El cálculo de  $cov(\mathbf{y}, \mathbf{x})$  está dado por:

$$\begin{aligned}
cov(\mathbf{y}, \mathbf{x}) &= E[\mathbf{y}\mathbf{x}^t] - E[\mathbf{y}]E[\mathbf{x}] = E[\mathbf{y}\mathbf{x}^t] - 0 \\
&= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][\Lambda_x\xi + \delta]^t \right\} \\
&= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon][\xi^t\Lambda_x^t + \delta^t] \right\} \\
&= E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon](\xi^t\Lambda_x^t) \right\} \\
&\quad + E \left\{ [\Lambda_y(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\varsigma) + \epsilon](\delta^t) \right\} \\
&= E \left\{ [\Lambda_y\ddot{B}^{-1}\Gamma\xi + \Lambda_y\ddot{B}^{-1}\varsigma + \epsilon](\xi^t\Lambda_x^t) \right\} \\
&\quad + E \left\{ [\Lambda_y\ddot{B}^{-1}\Gamma\xi + \Lambda_y\ddot{B}^{-1}\varsigma + \epsilon](\delta^t) \right\} \\
&= E[\Lambda_y\ddot{B}^{-1}\Gamma\xi\xi^t\Lambda_x^t + \Lambda_y\ddot{B}^{-1}\varsigma\xi^t\Lambda_x^t + \epsilon\xi^t\Lambda_x^t] \\
&\quad + E[\Lambda_y\ddot{B}^{-1}\Gamma\xi\delta^t + \Lambda_y\ddot{B}^{-1}\varsigma\delta^t + \epsilon\delta^t] \\
&= \Lambda_y\ddot{B}^{-1}\Gamma E[\xi\xi^t]\Lambda_x^t + \Lambda_y\ddot{B}^{-1}E[\varsigma\xi^t]\Lambda_x^t + E[\epsilon\xi^t]\Lambda_x^t \\
&\quad + \Lambda_y\ddot{B}^{-1}\Gamma E[\xi\delta^t] + \Lambda_y\ddot{B}^{-1}E[\varsigma\delta^t] + E[\epsilon\delta^t]
\end{aligned} \tag{3.10}$$

Nuevamente, aplicando las hipótesis sobre la correlación entre los factores independientes y dependientes a la expresión (3.10) se tiene que:

$$cov(\mathbf{y}, \mathbf{x}) = \Lambda_y\ddot{B}^{-1}\Gamma\Phi_\xi\Lambda_x^t \tag{3.11}$$

Por definición, sabemos que la covarianza es simétrica, es decir  $cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{y}, \mathbf{x})$ , por lo anterior y por la expresión (3.11) la  $cov(\mathbf{x}, \mathbf{y})$  está dada por:

$$cov(\mathbf{x}, \mathbf{y}) = \Lambda_x\Phi_\xi\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t \tag{3.12}$$

Hasta este momento se han calculado tres de los cuatro elementos de la matriz de varianzas-covarianzas, hace falta calcular  $var(\mathbf{x})$ , la cual está dada por:

$$\begin{aligned}
var(\mathbf{x}) &= E[\mathbf{x}\mathbf{x}^t] - E[\mathbf{x}]E[\mathbf{x}^t] = E[\mathbf{x}\mathbf{x}^t] - 0 \\
&= E[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)^t] \\
&= E[(\Lambda_x\xi + \delta)(\xi^t\Lambda_x^t + \delta^t)] \\
&= E[(\Lambda_x\xi + \delta)(\xi^t\Lambda_x^t) + (\Lambda_x\xi + \delta)\delta^t] \\
&= E[(\Lambda_x\xi\xi^t\Lambda_x^t + \delta\xi^t\Lambda_x^t)] + E[(\Lambda_x\xi\delta^t + \delta\delta^t)] \\
&= \Lambda_x E[\xi\xi^t]\Lambda_x^t + E[\delta\xi^t]\Lambda_x^t + \Lambda_x E[\xi\delta^t] + E[\delta\delta^t]
\end{aligned} \tag{3.13}$$

Nuevamente, aplicando las hipótesis de correlación entre factores latentes y comunes a (3.13) se tiene:

$$var(\mathbf{x}) = \Lambda_x E[\xi\xi^t]\Lambda_x^t + E[\delta\delta^t] = \Lambda_x\Phi_\xi\Lambda_x^t + \Theta_\delta \tag{3.14}$$

Por lo tanto, de las expresiones (3.9), (3.11), (3.12) y (3.14) se tiene que la matriz de varianzas-covarianzas de (3.6) es equivalente a:

$$\Sigma_{\mathbf{y}\mathbf{x}} = \begin{bmatrix} \Lambda_y\ddot{B}^{-1}(\Gamma\Phi_\xi\Gamma^t + \Psi_\varsigma)(\ddot{B}^{-1})^t\Lambda_y^t + \Theta_\epsilon & \Lambda_y\ddot{B}^{-1}\Gamma\Phi_\xi\Lambda_x^t \\ \Lambda_x\Phi_\xi\Gamma^t(\ddot{B}^{-1})^t\Lambda_y^t & \Lambda_x\Phi_\xi\Lambda_x^t + \Theta_\delta \end{bmatrix} \tag{3.15}$$

La matriz de varianzas-covarianzas incorpora la información que se desea estimar de las dos ecuaciones medidas y el componente estructural. La siguiente tarea consiste en definir la función objetivo necesaria para la estimación de los parámetros la cual se obtendrá a través del método de mínimos cuadrados generalizados (GLS) o máxima verosimilitud (LM).

En el capítulo anterior se mencionaron algunas de estas funciones, en este capítulo se presenta su desarrollo. Para esto se necesita un producto matricial especial llamado *Producto Kronecker*, el cual junto con algunas de sus propiedades se definen de manera formal en el apéndice.

### 3.4.1. Mínimos cuadrados generalizados

El análisis se hará sobre las dos ecuaciones medidas que se definieron en (3.1) y (3.2) por lo que supondremos al vector  $(x^t, y^t)$ , tal que tiene esperanza cero y una matriz de varianzas-covarianzas determinada por (3.15) y que denotaremos por  $\Sigma$ . Por construcción esta matriz es función de los parámetros que conforman el modelo. Por ejemplo: cargas de los factores, varianzas, covarianzas entre factores comunes y específicos.

Supongamos que esta matriz la podemos ver como una función de  $\theta_0$  vector que contiene los parámetros que se deben estimar, es decir, este vector hace que el modelo esté identificado ya que contiene las restricciones del modelo.

Por simplicidad en notación diremos que la matriz  $\Sigma$  es función de  $\theta$ , tal que  $\Sigma_0 = \Sigma(\theta_0)$  cumple las siguientes propiedades:

1.  $\Sigma_0$  es definida positiva<sup>4</sup>.
2. El vector  $\theta_0$  es un punto interior en el espacio paramétrico, i.e., existe una vecindad de  $\theta$  que se encuentra completamente contenida en el espacio paramétrico.
3. El modelo es identificado, i.e., si  $\Sigma(\theta_0) = \Sigma(\theta^*) \Rightarrow \theta_0 = \theta^*$ .
4. Todas las derivadas de los tres primeros ordenes de  $\Sigma$  con respecto a los elementos de  $\theta$  son continuas y acotadas en una vecindad de  $\theta_0$ .
5. La matriz  $q \times p^2$  de las derivadas parciales  $\frac{\partial \Sigma}{\partial \theta}$  es de rango completo en una vecindad de  $\theta_0$ .

La condición 1 implica que la distribución de las variables observadas no es degenerada, mientras que la condición 2 implica que los elementos de  $\Sigma$  son una función continua de los parámetros  $\theta$ . Esta condición y la 4 son importantes para la aplicación del teorema del valor medio al vector de parámetros y de esta manera demostrar ciertas propiedades asintóticas. La condición 3 es planteada para la unicidad del estimador de  $\theta$ , es necesaria para los procesos de interacción, los cuales serán divergentes sin esta condición. La condición 5 evita la colinealidad entre los elementos de  $\theta$ ; la identificación del modelo juega un papel importante, lo cual está asociado a la condición 3. Estos supuestos deben ser satisfechos por los modelos que se analizarán.

Por simplicidad en la notación, cuando hablemos de la variable medible  $x$  nos referiremos a las variables medidas conjuntas  $(x, y)^t$ . Por lo tanto, sea:  $\{x_1, x_2, \dots, x_p\}$  una muestra de v.a.i.i.d.  $\forall i \in \{1, \dots, p\}$  de acuerdo a una muestra aleatoria  $N(0, \Sigma_0)$ , el tamaño de cada muestra es  $n$ .

Se debe considerar, tanto para la teoría como para la aplicación de los MEE, que el tamaño de la muestra debe ser significativamente más grande que  $p^* = \frac{p(p+1)}{2}$ , i.e., el número de observaciones

<sup>4</sup>Una matriz es definida positiva si todos sus eigenvalores son positivos, o si todos los determinantes de sus menores principales son positivos.

debe ser significativamente mayor al número de elementos distintos en la matriz  $\Sigma_0$  el cual es una función del número de variables a analizar. Por lo tanto, el tamaño de muestra debe ser considerablemente mucho mayor que el número de variables a analizar.

Se debe tomar en cuenta que un modelo que tenga muchas variables implica un modelo muy parametrizado lo que podría ocasionar dificultades en la distribución Ji-cuadrada y la demanda de un número mayor de observaciones.

Un estimador de la matriz  $\Sigma_0$  es  $S$  y queda determinada por:

$$S = \begin{bmatrix} S_{11} & \dots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \dots & S_{pp} \end{bmatrix}$$

tal que cada elemento de la matriz se define como:

$$S_{kh} = \frac{\sum_{i=1}^n (x_{ki} - \bar{x}_k)(x_{hi} - \bar{x}_h)}{n-1}$$

$\forall k, h \in \{1, \dots, p\}$ . Se puede notar que cuando  $k \neq h$  estamos hablando de la  $cov(x_k, x_h)$  y cuando  $k = h$  se habla de la  $var(x_k)$ .

La matriz  $S$  antes definida es un estimador insesgado de  $\Sigma_0$ , entonces por esto y el teorema central del límite multivariado se tiene que:

$$\sqrt{n}\{vecs(S - \Sigma_0)\} \sim N(0, 2K_p^T(\Sigma_0 \otimes \Sigma_0)K_p) \quad (3.16)$$

donde  $vecs$  de una matriz  $B_{p \times p}$  es<sup>5</sup>:

$$vecs(B) = (b_{11}, b_{21}, b_{22}, b_{31}, b_{32}, b_{33}, \dots, b_{p1}, \dots, b_{pp})^T \quad (3.17)$$

De la ecuación (3.16) se tiene que los elementos de  $S$  son funciones cuadráticas de las  $x'_i s$  y pueden ser expresados por  $\Sigma_0$  (el segundo momento de las  $x'_i s$ ). Esta suposición es fundamental para el desarrollo del análisis de estructura de covarianza bajo una hipótesis normal.

La ecuación (3.16), da lugar a plantear la siguiente ecuación de regresión no lineal:

$$vecs(S) = vecs(\Sigma_0) + \epsilon'$$

donde  $\epsilon'$  es el vector residual, el cual tiene una distribución asintótica normal  $N(0, 2n^{-1}K_p^t(\Sigma_0 \otimes \Sigma_0)K_p)$ . La definición de la función *mínimos cuadrados generalizados* (GLS) es motivada de la función cuadrática de los residuos:

$$[vecs(S - \Sigma)]^t [2n^{-1}K_p^t(\Sigma_0 \otimes \Sigma_0)K_p]^{-1} [vecs(S - \Sigma)] \quad (3.18)$$

Se puede demostrar que la inversa de  $K_p^t(\Sigma_0 \otimes \Sigma_0)K_p$  es  $K_p^{-1}(\Sigma_0 \otimes \Sigma_0)(K_p^{-1})^t$ . Por tanto, sustituyendo este resultado en (3.18), con la ecuación (A.2) y del lema (A.1.1) del apéndice se tiene:

$$\begin{aligned} n2^{-1}[vecs(S - \Sigma)]^t K_p^{-1}(\Sigma_0 \otimes \Sigma_0)(K_p^{-1})^t [vecs(S - \Sigma)] \\ = n2^{-1}[vec(S - \Sigma)]^t (\Sigma_0 \otimes \Sigma_0) [vec(S - \Sigma)] \end{aligned}$$

<sup>5</sup>En el lema A.1.1 del apéndice se muestra la definición de  $vecs$  y  $vec$ .

Esta forma cuadrática residual es una función de  $\Sigma_0^{-1}$ . Como  $\Sigma_0$  es desconocida se debe reemplazar por una matriz positiva definida  $V_{p \times p}$  y considerar la siguiente función GLS:

$$\begin{aligned} G(\theta) &= 2^{-1}[\text{vec}(S - \Sigma)]^t (V \otimes V) [\text{vec}(S - \Sigma)] \\ &= 2n^{-1} \text{tr}(S - \Sigma)V^2 \end{aligned}$$

La matriz  $V$  puede ser una matriz constante positiva definida o una matriz estocástica que converge a una matriz positiva definida  $V^*$ . Cuando  $V$  es igual a la matriz identidad, la función GLS se reduce a una función de mínimos cuadrados (nótese que se hace el mismo supuesto que en análisis de regresión lineal al solicitar que la varianza de los errores sea constante).

**Definición 3.4.1.** *El estimador GLS  $\hat{\theta}$  de  $\theta$ , es el vector que minimiza  $G(\theta)$ .*

Basado en las técnicas del cálculo de matrices, el vector gradiente  $\dot{G}(\theta)$  y la matriz Hessiana  $\ddot{G}(\theta)$  de  $G(\theta)$  pueden ser calculados. Los resultados de las derivadas están dados por el siguiente lema.

**Lema 3.4.1.** 1.  $\dot{G}(\theta) = \frac{\partial G(\theta)}{\partial \theta} = -\dot{\Sigma}_0 (V \otimes V) \text{vec}(S - \Sigma)$

2.  $\ddot{G}(\theta) = \frac{\partial^2 G(\theta)}{\partial \theta \partial \theta} = \dot{\Sigma}(\theta) (V \otimes V) \dot{\Sigma}(\theta)^t - \ddot{\Sigma}(\theta) \{I_q \otimes ((V \otimes V) \text{vec}(S - \Sigma))\}$  donde  $I_q$  es una matriz identidad cuadrada de tamaño  $q$ .

Entonces  $G(\hat{\theta})$  es el mínimo de  $G(\theta)$  pues cumple con  $\dot{G}(\hat{\theta}) = 0$  y  $\ddot{G}(\hat{\theta})$  es positiva definida dentro de una vecindad lo suficientemente pequeña de  $\hat{\theta}$ . Por otra parte, se tiene que  $S$  converge en probabilidad a  $\Sigma_0$ . Con esto y del segundo postulado del lema 3.4.1 tenemos:

$$\ddot{G}(\theta_0) \xrightarrow{p} \Omega(V^*) = \dot{\Sigma}(\theta_0) (V^* \otimes V^*) \dot{\Sigma}(\theta_0)^t$$

Como  $\dot{\Sigma}(\theta_0)$  es de rango completo y  $(V^* \otimes V^*)$  es positiva definida; entonces  $\Omega(V^*)$  y la matriz Hessiana  $\ddot{G}(\theta_0)$  son positivas definidas. Estas matrices son importantes para derivar la distribución asintótica de  $\hat{\theta}_0$ .

### Propiedades asintóticas de los estimadores GLS

Las propiedades asintóticas presentadas en esta parte son de ayuda para la evaluación de la estimación del modelo.

**Teorema 3.4.1.** *El estimador GLS es consistente, es decir,  $\hat{\theta} \xrightarrow{p} \theta_0$ .*

Lo que el teorema 3.4.1 nos dice es que si  $n$  es lo suficientemente grande, el estimador GLS  $\hat{\theta}$  debería ser cercano al vector de parámetros  $\theta_0$ .

Esta propiedad da seguridad en el estimador GLS. La distribución de  $\hat{\theta}$  está dada por:

**Teorema 3.4.2.** *Sea  $C(\theta_0) = 2\Omega(V^*)^{-1}\Omega(V^*\Sigma_0V^*)\Omega(V^*)^{-1}$ , donde  $\Omega(V^*) = \dot{\Sigma}(\theta_0)(V^* \otimes V^*)\dot{\Sigma}(\theta_0)^t$  entonces:*

$$\sqrt{n}(\hat{\theta} - \theta_0) \sim N[0, C(\theta_0)].$$

La matriz de varianzas-covarianzas de  $\hat{\theta}$  es bastante complicada. Sin embargo, si  $V$  se selecciona de tal manera que  $V^* = \Sigma_0^{-1}$ , la matriz de varianzas-covarianzas asintótica de  $\hat{\theta}$  se simplifica significativamente. Por otra parte, el estimador resultante de GLS tendrá varianza mínima, de acuerdo al siguiente corolario.

**Corolario 3.4.1.** Si  $V \xrightarrow{p} V^* = \Sigma_0^{-1}$ , entonces

$$\sqrt{n}(\hat{\theta} - \theta_0) \sim N[0, 2\Omega(\Sigma_0^{-1})^{-1}]$$

Por otra parte,  $C(\theta_0) - 2\Omega(\Sigma_0^{-1})^{-1}$  es una matriz positiva definida y es el error estándar de  $\hat{\theta}$ .

Para obtener un buen estimador de la matriz de varianzas-covarianzas se debe a la elección de la matriz  $V^*$  y se le llamó el estimador GLS con  $V^* = \Sigma_0^{-1}$  el mejor estimador GLS. Claramente, si  $V = S^{-1}$ , entonces  $V^* = \Sigma_0^{-1}$ . Por lo tanto,  $S^{-1}$  es una selección lógica para la matriz ponderada  $V$  en  $G(\theta)$ .

Sea  $\Sigma = \Sigma(\hat{\theta})$  y  $\hat{\Omega}(\hat{\Sigma}^{-1}) = \dot{\Sigma}(\hat{\theta})(\hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1})\dot{\Sigma}(\hat{\theta})^t$ , donde  $\hat{\Omega}(\hat{\Sigma}^{-1})$  converge en probabilidad a  $\hat{\Omega}(\Sigma_0^{-1})$ , el error estándar de los elementos en  $\hat{\theta}$  puede obtenerse vía la correspondiente diagonal de los elementos de  $2n^{-1}\hat{\Omega}(\hat{\Sigma}^{-1})$ .

Otra inferencia estadística de  $\theta$  sería la prueba  $t$  o intervalo de confianza de un parámetro individual y puede obtenerse de métodos multivariados estandarizados.

Para la evaluación de bondad de ajuste, el siguiente teorema ayuda a obtener una evaluación estadística asintótica para el modelo hipotético.

**Teorema 3.4.3.** Si  $V^* = \Sigma_0^{-1}$  entonces  $nG(\hat{\theta}) \sim \chi_{(p^*-q)}^2$  donde  $p^* = \frac{p(p+1)}{2}$  tal que  $p$  es el número de variables observadas que forman parte del modelo.

Basado en el teorema 3.4.3, el modelo  $\Sigma(\theta)$  se rechaza si  $nG(\hat{\theta})$  es más grande que  $\chi_{(p^*-q)}^2(\alpha)$ , que es el valor crítico de una  $\chi^2$  con  $(p^* - q)$  grados de libertad y un nivel de error tipo I de  $\alpha$ .

El modelo propuesto  $\Sigma(\theta)$  no se rechaza si ocurre que  $nG(\hat{\theta})$  es menor que  $\chi_{(p^*-q)}^2(\alpha)$  sin embargo, esto no significa que  $\Sigma(\theta)$  pueda ser aceptado como un único modelo correcto.

En la práctica común de MEE, el modelo propuesto es considerado viable si  $\Sigma(\theta)$  se rechaza, sin embargo, no se debe tomar como el mejor o único modelo.

### 3.4.2. Máxima verosimilitud

Dado un vector de muestras  $\{x_1, x_2, \dots, x_p\}$ , cada muestra es de tamaño  $n$  con distribución normal multivariada  $N[\mu_0, \Sigma_0]$ . Una distribución exacta de la matriz de varianzas-covarianzas  $S$  sería una distribución *Wishart*<sup>6</sup> con una función de densidad:

$$f(S|\Sigma_0) = C \frac{\exp\left\{-\frac{(n-1)}{2}tr(S\Sigma_0^{-1})\right\}}{|\Sigma_0|^{(n-1)/2}}, \quad (3.19)$$

donde  $C$  es una constante de integración. Por lo tanto, el negativo de la función de máxima verosimilitud es:

$$-\ln C + \frac{(n-1)}{2}[tr(S\Sigma_0^{-1}) + \ln |\Sigma_0^{-1}|],$$

donde  $\Sigma_0^{-1} = \Sigma(\theta)^{-1}$  y  $\theta$  es el vector de parámetros tal como se trata en la estimación GLS.

Definiendo a  $C$  en función de  $S$  y  $p$  y haciendo el cálculo de las derivadas parciales correspondientes se llega a la función de discrepancia comúnmente usada en MEE:

$$F(\theta) = \ln |\Sigma(\theta)| + tr[S\Sigma(\theta)^{-1}] - \ln |S| - p$$

<sup>6</sup>Consultar Anderson, 2003.

**Definición 3.4.2.** El vector  $\hat{\theta}_M$  que minimiza  $F(\theta)$  se define como el estimador de  $\theta_0$ .

El vector estimado  $\hat{\theta}_M$  es un estimador consistente de  $\theta_0$  ya que  $\hat{\theta}_M \xrightarrow{P} \theta_0$ , el cual cumple las condiciones del siguiente lema:

**Lema 3.4.2.** 1.  $\dot{F}(\theta) = -\dot{\Sigma}(\theta)(\Sigma^{-1} \otimes \Sigma^{-1})\text{vec}(S - \Sigma)$

$$2. \ddot{F}(\theta) = \frac{\partial^2 F(\theta)}{\partial \theta \partial \theta} = -\dot{\Sigma}(\theta)(\Sigma^{-1} \otimes \Sigma^{-1})\dot{\Sigma}(\theta)^t - 2\dot{\Sigma}(\theta)\{(\Sigma^{-1} \otimes \Sigma^{-1})(S - \Sigma)\Sigma^{-1}\}\dot{\Sigma}(\theta)^t - \ddot{\Sigma}(\theta)\{I_q \otimes ((\Sigma^{-1} \otimes \Sigma^{-1})\text{vec}(S - \Sigma))\}$$

Similar a  $\ddot{G}(\theta)$  en la estimación GLS, para  $\ddot{F}(\theta)$  es también posible definir una vecindad lo suficientemente pequeña de  $\hat{\theta}_M$ . Por otra parte, se sigue de la segunda condición del lema 3.4.2, que:

$$\ddot{F}(\hat{\theta}_M) \xrightarrow{P} \dot{\Sigma}(\theta)(\Sigma_0^{-1} \otimes \Sigma_0^{-1})\dot{\Sigma}(\theta_0)^t = \Omega(\Sigma_0^{-1}) \quad (3.20)$$

y

$$E[\ddot{F}(\theta)] = \Omega(\Sigma_0^{-1})$$

Por lo tanto,  $\Omega(\Sigma_0^{-1})$  es la matriz de información de  $F(\theta)$  evaluado en  $\Omega(\Sigma_0^{-1})$ . Entonces, de la primer condición del lema 3.4.2 y de la ecuación (3.20) se tiene que:

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \xrightarrow{P} \Omega(\Sigma_0^{-1})\dot{\Sigma}(\theta_0)(\Sigma_0^{-1} \otimes \Sigma_0^{-1})(\sqrt{n}\text{vec}(S - \Sigma_0))$$

Se puede demostrar que  $\sqrt{n}(\hat{\theta} - \hat{\theta}_M)$  converge en probabilidad a:

$$\{\Omega(V^*)^{-1}\dot{\Sigma}(\theta_0)(V^* \otimes V^*) - \Omega(\Sigma_0^{-1})^{-1}\dot{\Sigma}(\theta_0)(\Sigma_0^{-1} \otimes \Sigma_0^{-1})\}n^{1/2}\text{vec}(S - \Sigma_0)$$

Como  $\{\Omega(V^*)^{-1}\dot{\Sigma}(\theta_0)(V^* \otimes V^*) - \Omega(\Sigma_0^{-1})^{-1}\dot{\Sigma}(\theta_0)(\Sigma_0^{-1} \otimes \Sigma_0^{-1})\}$  converge a cero en probabilidad y  $n^{1/2}\text{vec}(S - \Sigma_0)$  converge en distribución y  $n^{1/2}(\hat{\Omega} - \hat{\Omega}_M)$  converge a cero en probabilidad. Por lo tanto para el mejor estimador  $\hat{\Omega}$  con  $V^* = \Sigma_0^{-1}$ .

Máxima verosimilitud también ofrece un camino para la evaluación de bondad de ajuste del modelo, la cual puede ser planteada a través de la distribución asintótica con el criterio de razón de verosimilitudes, entonces se tiene que:

La hipótesis nula: la matriz de varianzas-covarianzas de  $X$  se ajusta a la matriz del modelo hipotético  $\Sigma(\theta)$ , contra la hipótesis alternativa: la matriz del modelo hipotético es cualquier otra matriz positiva definida.

Para contrastar las hipótesis se recurre al criterio de razón de verosimilitudes, tal que la expresión queda establecida como:

$$LR = \frac{f(S|\Sigma(\hat{\theta}_M))}{f(S|(n-1)n^{-1}S)}, \quad (3.21)$$

donde  $(n-1)n^{-1}S$  es el estimador de máxima verosimilitud de la matriz de varianzas-covarianzas bajo el espacio parametral general. Ahora bien por (3.19) se tiene que (3.21) se puede expresar como:

$$-2 \ln LR = -2 \ln \left[ \frac{|\Sigma(\hat{\theta}_M)|^{-(n-1)/2} \exp \frac{-(n-1)}{2} \text{tr} S \Sigma(\hat{\theta}_M)^{-1}/2}{|(n-1)n^{-1}S|^{-(n-1)/2} \exp \frac{-(n-1)}{2} \text{tr} S(n-1)n^{-1}S)^{-1}/2} \right]$$

¿Cómo se distribuye este cociente? Para dar respuesta nos ayudamos de dos resultados: el primero es el siguiente teorema:

**Teorema 3.4.4.** *La distribución asintótica de  $nF(\hat{\theta}_M)$  es  $\chi^2$  con  $p^* - q$  grados de libertad, es decir,*

$$nF(\hat{\theta}_M) \sim \chi^2_{(p^*-q)}$$

El otro resultado es que (3.21) tiende en probabilidad a  $nF(\hat{\theta}_M)$ , entonces, por transición, se tiene que la razón de verosimilitudes se distribuye como una  $\chi^2$  con  $(p^* - q)$  grados de libertad.

Para terminar con esta sección es posible demostrar que la estimación de los parámetros a través de los métodos de estimación revisados son estadísticamente equivalentes. Este resultado se sustenta en el siguiente teorema.

**Teorema 3.4.5.** *El mejor estimador GLS y el estimador ML son asintóticamente equivalentes tal que:*

$$n^{1/2}(\hat{\Omega} - \hat{\Omega}_M) \xrightarrow{p} 0$$

De este teorema se tiene que:

**Corolario 3.4.2.** 1.  $n^{1/2}(\hat{\Omega} - \hat{\Omega}_M) \sim N[0, \Omega(\Sigma_0^{-1})^{-1}]$

2.  $nG(\hat{\Omega}) \sim \chi^2_{(p^*-q)}$

3.  $nF(\hat{\Omega}_M) \sim \chi^2_{(p^*-q)}$

Las propiedades asintóticas de las aproximaciones GLS y ML dependen fundamentalmente de la validación del resultado sobre la distribución asintótica de la matriz de varianzas-covarianzas muestral, por lo que es necesario revisar que las siguientes condiciones se satisfagan:

1. Las variables son independientes e idénticamente distribuidas.
2. Las variables provienen de una distribución normal.

## 3.5. Bondad de ajuste

Como en toda técnica estadística existe la necesidad de evaluar el ajuste del modelo obtenido; por ejemplo, en el Análisis de Regresión la bondad de ajuste consiste en el análisis de la  $R^2$ , análisis de la tabla ANOVA, análisis de residuos, multicolinealidad, entre otras; para el caso de MEE también hay técnicas para la evaluación del ajuste y son similares al AFC. En esta sección mostraremos la explicación y aplicación de las técnicas para la evaluación del modelo MEE.

### 3.5.1. Ajuste global del modelo

Como se mencionó anteriormente, la estimación de MEE es a través de la adaptación de la matriz de varianzas y covarianzas  $\Sigma(\theta)$ , la cual está restringida al modelo propuesto y a  $S$  (que es la estimación de la matriz de varianzas-covarianzas). Hay tres tipos de medidas para el ajuste global: *métricas de ajuste absoluto*, *ajuste relativo* y *ajuste de parsimonia*, las cuales se explican a continuación.

### Medidas de ajuste absoluto

Nuestra medida de discrepancia está dada por la función de ajuste (la cual depende de la selección del método de estimación) y que tiene una distribución  $\chi^2_{(p^*-q)}$  de acuerdo al corolario 3.4.2, esto quiere decir que cuanto menos sea la diferencia entre ambas matrices se podrá entender que el modelo reproduce el comportamiento de los datos observados de manera aceptable.

Dentro de las métricas de ajuste absoluto se encuentra la siguiente prueba de hipótesis conocida como la *Ji-Cuadrada* la cual establece la hipótesis nula:

$$H_0 : \Sigma(\theta) = S$$

y la hipótesis alternativa:

$$H_0 : \Sigma(\theta) = A$$

donde la matriz  $A$  es cualquier matriz positiva definida distinta a  $S$ . En MEE no se debe rechazar la hipótesis nula (éste es un objetivo usual en pruebas de bondad de ajuste), lo cual implica que el modelo propuesto ajusta adecuadamente a los datos, por lo tanto, el *p-value* asociado al estadístico debe ser mayor al nivel de significancia ( $\alpha$ ) propuesto por el investigador. En caso contrario, se concluye que el modelo propuesto no es consistente con los datos observados.

Se deben tener en cuenta tres factores que pueden influir de manera negativa en esta prueba y por lo que esta prueba ha caído en desuso:

1. Violación de la condición de normalidad en las variables medidas.
2. Complejidad del modelo. El estadístico decrece cuando incrementan los parámetros que se tienen que estimar, i.e., para modelos muy parametrizados y de gran complejidad tiende a producir valores más pequeños de este estadístico en comparación con modelos más sencillos dada la reducción de los grados de libertad empleados. Pero si los modelos muy parametrizados tienden a producir valores pequeños de la estadística de prueba, entonces crece la probabilidad de aceptar el modelo. En otras palabras, cuanto mayor sea la complejidad del modelo, mayor es la probabilidad de que la prueba no rechace el modelo.
3. Ante tamaños de muestra pequeños, la prueba no es capaz de detectar diferencias significativas, aceptando con alta probabilidad modelos que no se ajustan bien a los datos. Por otra parte, ante tamaños de muestra grandes, la estadística de prueba tiende a ser estadísticamente significativa, rechazando el modelo que en realidad se aleja muy poco de los datos observados.

Debido a estos tres factores cuando se evalúa la bondad de ajuste del modelo sería un error tomar una decisión basándose únicamente en esta prueba de hipótesis, así que se debe tomar en cuenta la evaluación de métricas distintas a ésta para la evaluación del modelo.

Dentro de las métricas de ajuste absoluto también se encuentra la evaluación de la estadística Ji-cuadrada central, la raíz del error cuadrático medio *RMSEA* (*Root Mean Square Error of Approximation*), la cual mide la discrepancia entre la matriz reproducida por la población y no en términos de la muestra. Para el RMSEA, valores por debajo de 0.05 indican un buen ajuste del modelo aceptable, aunque si el índice está entre 0.05 y 0.08 (Hair et al., 2000) el modelo podría ser aceptado.

Otra medida de evaluación absoluta es *GFI* (*Goodness of Fit Index*) que reporta la variabilidad explicada por el modelo (como la  $R^2$  del análisis de regresión lineal<sup>7</sup>), entonces  $GFI \in [0, 1]$ , donde la cota inferior del intervalo indica que no hay buen ajuste y la cota superior indica un ajuste perfecto, es decir, valores más cercanos a 1 son preferibles.

El índice *AGFI* (*Adjusted Goodness of Fit Index*), que es el GFI ajustado. Es ajustado ya que toma en cuenta el número de parámetros a estimar y el número de datos disponibles. Tanto el GFI como el AGFI deberían estar por arriba de 0.9 para concluir en un buen modelo (Raykov y Marcoulides, 2006).

### Índices de ajuste incrementales

En este grupo se incluyen una serie de índices que comparan la mejoría en la bondad de ajuste de un modelo con la bondad de ajuste de un modelo base. El modelo base, comúnmente usado para realizar las comparaciones, es el modelo nulo o independiente, en el que se supone que las variables no están relacionadas.

En este sentido, aunque el ajuste del modelo diseñado no sea perfecto, será una mejor aproximación a la realidad en tanto cuanto mejore el ajuste del modelo nulo.

Dentro de este conjunto están los siguientes índices: el índice de ajuste normado *NFI* (*Normed Fit Index*), que mide la reducción proporcional en la función de ajuste cuando se pasa del modelo nulo al modelo propuesto; el índice de ajuste no normado *NNFI* (*Nonormed Fit Index*), el cual compara el ajuste por grado de libertad entre el modelo propuesto y el nulo; el índice de ajuste comparativo *CFI* (*Comparative Fit Index*), el cual indica un buen ajuste próximo a la unidad.

Los índices antes mencionados también deberían estar por arriba de 0.9 para concluir en un buen modelo (Raykov y Marcoulides, 2006).

### Índices de ajuste de parsimonia

Hay otro grupo de indicadores de ajuste llamado de *Parsimonia*, estos indicadores ayudan a determinar dentro de un conjunto de modelos estimados cuál es el mejor modelo considerando su ajuste relativo a su complejidad. A continuación se definen los índices: *PGFI* y *PNFI*.

*PGFI* por sus siglas en inglés (*Parsimony Goodness of Fit Index*), consiste en el ajuste del GFI basado en la parsimonia del modelo estimado, el cual se expresa como sigue:

$$PGFI = \frac{gl_T}{p^*} \times GFI \quad (3.22)$$

donde  $p^* = \frac{p(p+1)}{2}$ , tal que  $p$  es el número de variables medidas con que se está modelando.

*PNFI* (*Parsimony Normed Fit Index*) es similar al NFI, pero considerando los grados de libertad usados para alcanzar el nivel de ajuste. La expresión de este índice es:

$$PNFI = \frac{gl_T}{gl_I} \times NFI \quad (3.23)$$

Valores elevados (cercaos a 1) de PNFI son mejores y su uso principal consiste en la comparación de modelos con diferentes grados de libertad. Se utiliza para comparar modelos alternativos y

---

<sup>7</sup>Manzano y Zamora, 2013.

no existen niveles recomendados en que se deba encontrar este índice. Sin embargo, cuando se compara entre modelos, se recomienda escoger aquel modelo con menor PNFI tal que la diferencia sea menor a 0.09 al del modelo independiente.

El criterio de información de *Akaike* o *AIC* (*Akaike Information Criterion*), cuya utilidad reside en comparar modelos que poseen diferente número de variables. La expresión para calcular este índice se dió en el capítulo anterior.

Los valores de AIC cercanos a cero indican un mejor ajuste y una mayor parsimonia; ya que un valor reducido de AIC se produce cuando los valores de la estadística Ji-cuadrada se consiguen con muy pocos coeficientes estimados. Con ellos obtenemos, no sólo un buen ajuste de las covarianzas y correlaciones previstas y observadas, sino también un modelo no propenso al *sobreaajuste*.

El índice de validación cruzada esperada (*ECVI*) es una aproximación a la bondad de ajuste que conseguiría el modelo estimado en otra muestra del mismo tamaño. Basándose en la matriz de covarianzas de la muestra, tiene en cuenta el tamaño de muestra efectiva y la diferencia que podría esperarse en otra muestra. El *ECVI* también toma en cuenta el número de parámetros estimados tanto para los modelos de medida como para los estructurales.

El *ECVI* no tiene un rango especificado aceptable de valores, pero se utiliza en la comparación entre modelos alternativos, prefiriendo a aquel cuyos índices en conjunto sean los de menor valor.

En forma de resumen se presenta la tabla 3.2 con los índices de bondad de ajuste, su expresión y su criterio para un buen ajuste (niveles de ajuste aceptables).

### 3.5.2. Análisis de validez

Comencemos entendiendo la confiabilidad de una variable latente. Ésta indica que los distintos ítems que construyen a la latente al estar muy correlacionados entre sí están midiendo la misma variable latente. Sin embargo, esto no implica que la variable latente es la que se debe estar midiendo, es decir, no sabemos si es válida.

Que una variable latente sea confiable no implica que sea válida, pero que una variable no sea confiable implica que no es válida, por lo tanto, la confiabilidad se convierte en una condición necesaria pero no suficiente para medir la validez de la escala.

La confiabilidad de una variable latente se medirá a través de un estadístico conocido como *Coefficiente Alpha de Cronbach* el cual se definirá a continuación.

#### Coefficiente Alpha de Cronbach

Supongamos que tenemos una variable latente  $Y$  construida por  $k$  variables medidas  $(x_1, x_2, \dots, x_k)$  con cierta escala  $v$  de la cual se desea calcular su fiabilidad. Las variables medidas que la construyen tienen asociada una matriz de varianzas-covarianzas<sup>8</sup>  $\Sigma$ :

$$\Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \dots & \sigma_{kk} \end{bmatrix}$$

Los elementos de la diagonal de la matriz anterior expresan la varianza de cada variable medida,

<sup>8</sup>Los desarrollos que se realicen para la matriz de varianzas-covarianzas también se aplican para matriz de correlación, ya que la primera es una generalización de esta última.

Índice	Expresión	Niveles de ajuste aceptable
$\chi^2$	$\omega = N \times F(\theta_0)$	$p[\omega \leq \omega_\alpha] \geq \alpha$
RMSEA	$\sqrt{\frac{NCP}{N \times gl_T}}$ donde $NCP = \begin{cases} \chi^2 - gl_T \\ 0 \end{cases}$	$RMSEA \leq 0.08$ ( $RMSEA \leq 0.05$ preferible)
RMR	$\sqrt{\frac{\sum_{i=1}^q \sum_{j=1}^i (s_{ij} - \hat{\sigma}_{ij})^2}{q(q+1)/2}}$	Valores cercanos a cero
SRMR	$\sqrt{\sum_{i=1}^q \sum_{j=1}^q \rho_{ij}^2}$	Valores cercanos a cero
GFI	$\frac{tr(\hat{\sigma}'W\hat{\sigma})}{tr(\hat{s}'W\hat{s})}$	Similar a $R^2$ de regresión lineal
AGFI	$1 - \frac{1-GFI}{1-k/p^*}$	Valores entre 0.9 y 1 son deseables
CFI	$\left[ \frac{(\chi_I^2 - gl_I) - (\chi_T^2 - gl_T)}{(\chi_T^2 - gl_T)} \right]$	Valores entre 0.9 y 1 son deseables
NFI	$\frac{\chi_I^2 - \chi_T^2}{\chi_I^2}$	Valores entre 0.9 y 1 son deseables
NNFI(TLI)	$\frac{\chi_I^2 - \frac{gl_I}{gl_T} \chi_T^2}{\chi_I^2 - gl_I}$	Valores entre 0.9 y 1 son deseables
PNFI	$\frac{gl_T}{gl_I} \times NFI$	Diferencia sea menor a 0.09
PGFI	$\frac{gl_T}{p^*} \times GFI$	Valores entre 0.9 y 1 son deseables
AIC	$\chi_T^2 - 2gl_T$	El de menor valor entre modelos
ECVI	$\frac{\chi^2}{N-1} \frac{2k}{N-1}$	El de menor valor entre modelos

Tabla 3.2: Resumen de métricas para la bondad de ajuste de los MEE.

mientras que los que están fuera de la diagonal la covarianza o correlación (según sea el caso) entre cada par de variables medidas. Una propiedad que también es útil es que la suma de los elementos de la matriz da como resultado la varianza total de la escala  $v$ .

Pues bien, el *Coefficiente Alpha de Cronbach* se define como la porción de la varianza total de la escala  $v$  atribuible a la variable latente  $y$ . Cuanto mayor sea este coeficiente se entiende que la escala representa mejor a la variable latente en su escala, esta porción de varianza se le identifica como *varianza común*.

La parte de la varianza total que no explica la variable latente es la varianza de cada ítem y es a razón de los errores de medición que hay sobre cada variable medible, a esta porción de la varianza se denomina *varianza específica* y por tanto la siguiente igualdad:

$$varianza_{total} = varianza_{común} + varianza_{específica}$$

Por lo que se sigue que la varianza común se puede calcular como:

$$\alpha_c = \frac{\sum_i \sum_j \sigma_{ij} - \sum \sigma_{gh}}{\sum_i \sum_j \sigma_{ij}} = 1 - \frac{\sum \sigma_{gh}}{\sum_i \sum_j \sigma_{ij}}$$

$\forall g = h$ , es decir, sólo se suman los elementos de la diagonal. Se puede notar que cada término de error provoca varianza en cada ítem por separado y esos errores no están correlacionados con ningún otro error, por lo tanto determinamos que cada variable medida sólo varía en función de la varianza común y la varianza específica.

En la expresión anterior hace falta introducir una última corrección para tener la expresión del Alpha de Cronbach. El número total de elementos de la matriz de varianzas-covarianzas que es  $k^2$ ,  $k$  el número de elementos que hay en la diagonal, entonces el número de elementos que están fuera de la diagonal es:  $k^2 - k = k(k - 1)$ . Entonces, se tiene que la expresión anterior es una fracción con un numerador basado en  $k$  variables y un denominador basado en  $k^2$  valores.

Para ajustar los cálculos de tal forma que el cociente exprese las magnitudes relativas, más que un número absoluto que hay en el numerador y denominador, se corrige la expresión anterior para contrastar el efecto de la diferencia por  $k^2/(k^2 - k)$  y factorizando  $k$  se tiene que es equivalente a  $k/(k - 1)$ , entonces:

$$\alpha_c = \frac{k}{(k - 1)} \left( 1 - \frac{\sum_i \sum_j \sigma_{ij}}{\sum_i \sigma_{gh}} \right) \quad (3.24)$$

La anterior expresión indica qué porción de la varianza total está provocada por la variable latente corrigiendo este valor por el número de casos que intervienen en los cálculos.

En caso de que se esté analizando la matriz de correlaciones y no la de varianzas-covarianzas, denominaremos a la media de las varianzas como  $v$  y la media de las covarianzas como  $c$ , entonces como hay  $k$  varianzas y  $(k^2 - k)$  covarianzas se puede expresar (3.24) como:

$$\alpha_{cc} = \left( \frac{k}{k - 1} \right) \left( 1 - \frac{kv}{kv + (k^2 - k)c} \right)$$

la cual se puede resumir como:

$$\alpha_{cc} = \frac{kv}{v + (k - 1)c}$$

Ahora bien, como lo que se busca es la expresión que utilice correlaciones, entonces, si se estandarizan las varianzas y covarianzas, estas últimas pasan a expresarse como  $\rho$  y las varianzas vendrían siendo 1, por tanto:

$$\alpha_{cc} = \frac{k}{1 + (k - 1)\rho}$$

donde  $\rho$  es el coeficiente de correlación entre todos los ítems que conforman la escala. La expresión es mejor conocida como *Spearman-Brown* (Croker y Algina, 1986).

La interpretación de este coeficiente es:

- Si la escala se encuentra en desarrollos preliminares, el valor a un nivel de 0.7 indica un buen ajuste.
- Si la escala ha tenido ya depuraciones, el valor no debe ser menor a 0.8.
- Si se va a tomar algún tipo de decisión, el valor no debería bajar de 0.9.

### Fiabilidad compuesta

Hasta el momento se ha hablado de evaluar la fiabilidad de una variable latente por separado, sin embargo, esto no atrapa la influencia sobre la fiabilidad del resto de los constructos. Por esta razón, Fornell y Larcker (1981) proponen el cálculo de la fiabilidad compuesta (IFC) para cada factor, el cual se interpreta de la misma manera que el Alpha de Cronbach, la expresión es la siguiente:

$$IFC_i = \frac{(\sum L_{ij})^2}{(\sum L_{ij})^2 + \sum var(E_{ij})}$$

donde  $L_{ij}$  es la carga de los factores estandarizada de cada uno de los  $j$  indicadores que cargan sobre el factor  $i$  y  $var(E_{ij})$  es la varianza del término de error asociado a cada uno de los  $j$  indicadores del factor  $i$ . Se puede calcular, también como:

$$var(E_{ij}) = 1 - L_{ij}^2$$

Para obtener esta información es necesario realizar el análisis de factores confirmatorio entre los instrumentos de medida.

### Índice de varianza extraída

Fornell y Lacker (1981) presentan este índice IVE como la relación entre la varianza que es capturada por un factor  $i$  en relación a la varianza total debida al error de medida de ese factor, esto es:

$$IVE_i = \frac{\sum L_{ij}^2}{\sum L_{ij}^2 + \sum var(E_{ij})}$$

Nótese que la única diferencia con el IFC es que cada carga de los factores estandarizada es primero elevada al cuadrado antes de ser sumadas. Fornell y Larcker (1981) sugieren que es deseable que el constructo tenga valores de IVE iguales o superiores a 0.5, es decir, que sean superior la varianza capturada por el factor que la debida al error de medida.

### 3.5.3. Ajuste del modelo estructural

El análisis de trayectoria *Path Analysis* es el que modela las relaciones entre las variables latentes, o lo que es lo mismo, el modelo de estructura, en donde se incorporan el error de medida y la relación entre variables latentes.

Para evaluar el ajuste del modelo estructural, corresponde, en primer lugar, analizar la significancia alcanzada por los coeficientes estimados. De este modo, cualquier parámetro estimado debe ser estadísticamente diferente de cero.

Consecuentemente, un parámetro no significativo indicaría que la relación propuesta no tiene efecto sustancial, por lo que debería ser eliminado y, a continuación, reformular el modelo. Cabe aclarar que los MEE tienen carácter confirmatorio, y no exploratorio, por lo que esto último, de eliminar parámetros y reformular el modelo, no siempre puede hacerse.

Hay que tener en cuenta que la magnitud de los coeficientes no está únicamente determinada por la significancia de los parámetros, ya que hay que considerar otros factores como el tamaño muestral y la varianza de las variables dependientes e independientes. Cuanto mayor es la magnitud

de la relación y el tamaño muestral y cuanto menor es la varianza de las variables dependientes e independientes, mayor es la probabilidad de obtener una relación significativa.

### 3.5.4. Reespecificación del modelo

En ocasiones el investigador se ve en la necesidad de modificar el modelo estimado principalmente por:

1. Mejorar el ajuste del modelo obtenido.
2. Contrastar alguna hipótesis teórica.

Y tal como se mencionó anteriormente, se deben tener ciertas precauciones al hacer algún tipo de reespecificación, ya que los resultados obtenidos pueden ser sólo un efecto muestral y no poblacional.

Para lograr la reespecificación se tienen tres herramientas:

- Significatividad de los parámetros. Tal como en regresión ayuda a evaluar si una relación es significativa.
- Multiplicadores de Lagrange. Propone nuevas relaciones que son sustentadas por los datos.
- Contraste Wald. Tiene el mismo objetivo del primer punto, ayuda a evaluar que las relaciones sean soportadas por los datos, si no hay evidencia muestral que los datos soporten relaciones que deben ser retiradas.

## 3.6. Descomposición del modelo

Otra propiedad relevante de los modelos de ecuaciones estructurales es que permite calcular el *efecto indirecto y total* que puede tener una variable sobre otra, y no sólo el *efecto directo* como en el análisis de regresión. A continuación definimos los tipos de efectos:

- *Directo*: es la influencia que tiene una variable sobre otra, que se da de manera directa en el *diagrama de trayectoria* por medio de la flecha que une dos variables.
- *Indirecto*: es la influencia que tiene una variable sobre otra, pero en cuya trayectoria hay al menos una variable intermedia que las une.
- *Total*: es la suma del efecto directo e indirecto y permite cuantificar el cambio que se observa en la variable en que se produjo el efecto (la que recibe la flecha), inducido por un cambio en la variable que lo causa (variable que sale la flecha), independiente de los mecanismos por los cuales se haya producido dicho cambio.

Para calcular el efecto total de  $\eta$  sobre  $\eta$  se da la siguiente definición:

$$T_{\eta\eta} = \sum_{k=1}^{\infty} B^k$$

Note que  $T_{\eta\eta}$  está definida sí y sólo sí la ecuación anterior converge a una matriz con elementos finitos.

En general para los sistemas recursivos,  $B$  es una matriz triangular inferior. Si  $B$  es triangular inferior su  $k$ -ésima potencia es cero para  $k \geq m$ , donde  $m$  es el número de relaciones directas; por tanto, los efectos totales estarán definidos.

Los efectos directos de  $\eta$  sobre  $\eta$  están dados por los coeficientes entre las variables endógenas, es decir:

$$D_{\eta\eta} = B$$

Por lo tanto, los efectos indirectos de  $\eta$  sobre  $\eta$  quedan definidos por:

$$I_{\eta\eta} = \sum_{k=1}^{\infty} B^k - B = \sum_{k=2}^{\infty} B^k$$

Para los modelos no recursivos se requiere que  $B^{k+1} \xrightarrow{k \rightarrow \infty} 0$ , de esta manera queda definido  $T_{\eta\eta}$  y puede ser expresada como:

$$T_{\eta\eta} = (I - B)^{-1} - I$$

Entonces los efectos indirectos están dados por:

$$I_{\eta\eta} = (I - B)^{-1} - I - B$$

Lo anterior fue para calcular los efectos de  $\eta$  sobre  $\eta$ , sin embargo, queda pendiente el efecto del resto de las relaciones.

Los efectos totales de  $\xi$  sobre  $\eta$ , los cuales están dados por la siguiente expresión:

$$T_{\eta\xi} = (I - B)^{-1}\Gamma$$

El efecto directo está dado por los coeficientes de la matriz  $\Gamma$ , esto implica que los efectos indirectos quedan definidos por:

$$I_{\eta\xi} = (I - B)^{-1}\Gamma - \Gamma = ((I - B)^{-1} - I)\Gamma$$

Se puede notar que los efectos indirectos de  $\xi$  sobre  $\eta$  es igual al producto de los efectos totales de  $\eta$  sobre  $\eta$  para modelos no recursivos multiplicado por los efectos directos de  $\xi$  sobre  $\eta$ .

Por otra parte, para el modelo de medida se muestra que el efecto total que tiene  $\xi$  sobre  $y$  es:

$$T_{\xi y} = \Lambda_y(I - B)^{-1}\Gamma$$

El efecto total de  $\eta$  sobre  $y$  es:

$$T_{y\eta} = \Lambda_y(I - B)^{-1}$$

y el efecto indirecto está dado por:

$$I_{y\eta} = \Lambda_y(I - B)^{-1} - \Lambda_y = \Lambda_y((I - B)^{-1} - I)$$

Como ya hemos mencionado los efectos totales e indirectos están definidos bajo ciertas condiciones, es decir, el eigenvalor de  $B$  deber ser menor a uno, siendo ésta una condición suficiente pero no necesaria; por último estas condiciones no son necesarias para la existencia de  $T_{\xi y}$ ,  $T_{\eta y}$  y  $T_{\xi \eta}$ .

Se ha mencionado una técnica para la estimación de los efectos con la finalidad de que haya claridad de lo que significa, cómo se interpreta y qué condiciones se necesitan. Sin embargo, cuando se tenga un modelo de MEE no hay la necesidad de calcular los efectos de forma manual ya que algunos paquetes estadísticos los proporcionan; entre ellos están MPLUS, SAS, LISREL y AMOS.

Actualmente en marketing estos efectos son utilizados para obtener el ROI (*Return of Investment*) a través de las inversiones que se hacen en los distintos medios de comunicación *vs.* las ventas obtenidas.

Para tener una idea clara y pueda aterrizar ideas de lo que se ha hablado en este capítulo, en la próxima sección se dará un ejemplo de un MEE.

### 3.7. Ejemplo

Para clarificar conceptos e ideas se desarrolla y analiza un Modelo de Ecuaciones Estructurales, el cual fue enunciado en el ejemplo 3.1.1.

La *Especificación* del modelo fue hecha al plantear el *diagrama de trayectoria* 3.1. Por otro lado, en la *Identificación* del modelo sabemos que hay 8 variables medidas, por lo tanto el número de datos distintos en la matriz de correlación es  $8(8 + 1)/2 = 36$ . Ahora bien, antes de aplicar restricciones se tendrían que estimar los siguientes parámetros:

- 8 varianzas de los términos de error de medida, **vare**.
- 2 varianzas de los errores de predicción, **vard**.
- 3 varianzas de los factores independientes, **varf**.
- 8 coeficientes de regresión entre las variables observadas y sus términos de error.
- 2 coeficientes de regresión entre los factores dependientes y sus términos de error.
- 4 covarianzas, 3 entre los factores independientes y 1 entre los factores dependientes, **cff** y **cdd**.
- 8 coeficientes de regresión entre los factores y las variables observadas, **lvf**.
- 5 coeficientes entre factores independientes y dependientes, **pff**.

En total se tendrían que estimar 40 parámetros, i.e., no se tendrían grados de libertad. Esto sería un problema ya que se tendría un modelo subidentificado y no se podrían hacer las estimaciones. Sin embargo, este resultado no es definitivo ya que faltan las restricciones del modelo, las cuales quedarían de la siguiente manera:

- Se fijan a 1 todos los coeficientes de regresión entre variables medidas o latentes y los términos de error.

- Se establecen las escalas de las variables latentes, i.e., los coeficientes de:  $v_1$  a  $f_1$ ,  $v_2$  a  $f_2$ ,  $v_4$  a  $f_3$ ,  $v_6$  a  $f_4$  y  $v_8$  a  $f_5$  se fijan a 1.
- Se fijan a cero las varianzas de los términos de error asociadas a las variables observadas cuando un factor está explicado por una de ellas, sería el caso de  $v_1$  y  $v_8$ .

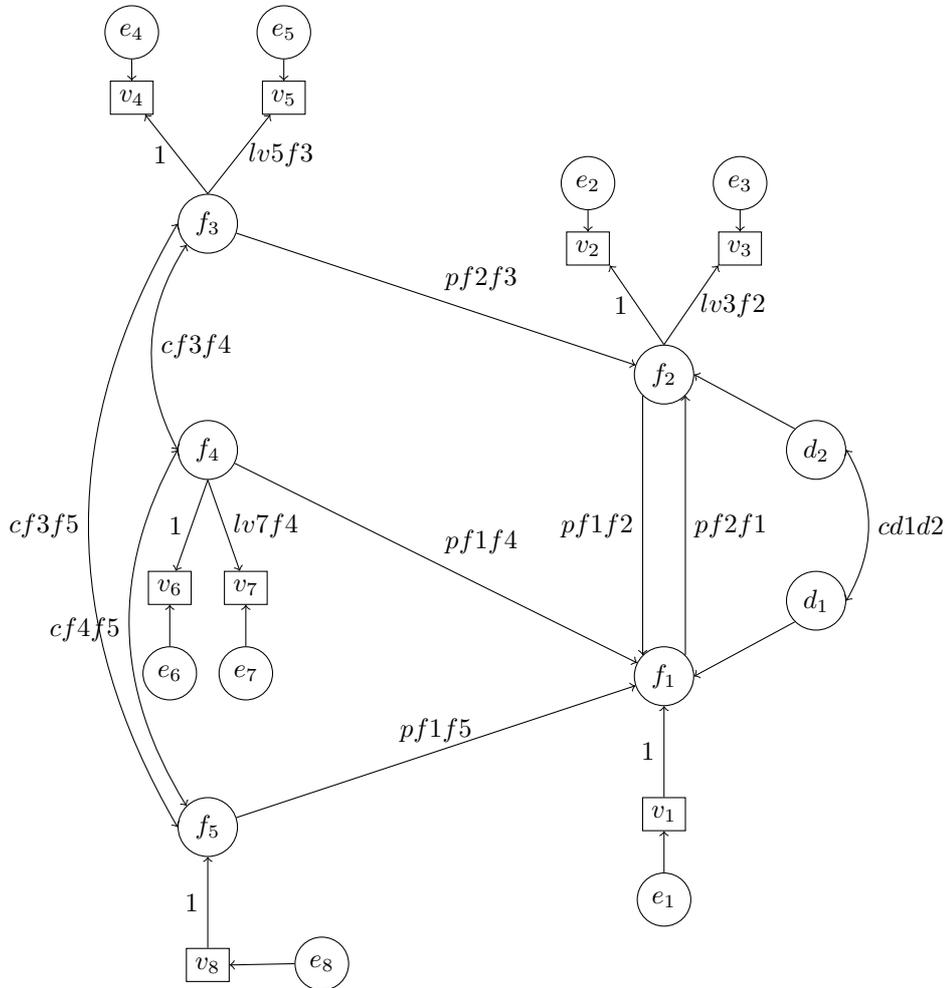


Figura 3.4: Modelo de desempeño y satisfacción de la fuerza de ventas con restricciones.

Con estas restricciones el número de parámetros a estimar ha pasado de 40 a 23 y por tanto, se tiene que el modelo está identificado con 13 grados de libertad ( $36 - 23$ ). Las restricciones mencionadas se ven reflejadas en el *diagrama de trayectoria* 3.4.

Para la *Estimación* se recurre a la librería `sem` de R. A continuación se muestra los datos de entrada necesarios para la estimación de los parámetros.

Comenzamos llamando a la librería y cargando la matriz de correlación:

```
library(sem)
```

```

corr.ejem <- readMoments(diag=TRUE,
names=c('v1','v2','v3','v4','v5','v6','v7','v8'))
1,
0.418, 1,
0.394, 0.627, 1,
0.129, 0.202, 0.266, 1,
0.189, 0.284, 0.208, 0.365, 1,
0.544, 0.281, 0.324, 0.201, 0.161, 1,
0.507, 0.225, 0.314, 0.172, 0.174, 0.546, 1,
-0.357, -0.156, -0.038, -0.199, -0.277, -0.294, -0.174, 1

```

Para la definición del modelo se usa la función *specifyModel*. Para el ejemplo, el modelo se especifica de la siguiente manera:

```

mod.cap3 <- specifyModel()
f1 -> v1, NA, 1   ### Relaciones Factores-VARIABLES Observadas
f2 -> v2, NA, 1
f2 -> v3, lv3f2, NA
f3 -> v4, NA, 1
f3 -> v5, lv5f3, NA
f4 -> v6, NA, 1
f4 -> v7, lv7f4, NA
f5 -> v8, NA, 1
f1 -> f2, pf2f1, NA   ### Relaciones Factores
f2 -> f1, pf1f2, NA
f3 -> f2, pf2f3, NA
f4 -> f1, pf1f4, NA
f5 -> f1, pf1f5, NA
f1 <-> f1, vard1, NA   ### Varianzas Errores de Prediccion
f2 <-> f2, vard2, NA
f3 <-> f3, varf3, NA   ### Varianzas Factores Independientes
f4 <-> f4, varf4, NA
f5 <-> f5, varf5, NA
v1 <-> v1, NA, 0   ### Varianzas Errores de Medida
v2 <-> v2, vare2, NA
v3 <-> v3, vare3, NA
v4 <-> v4, vare4, NA
v5 <-> v5, vare5, NA
v6 <-> v6, vare6, NA
v7 <-> v7, vare7, NA
v8 <-> v8, NA, 0
f1 <-> f2, cd1d2, NA   ### Covarianzas Factores Dependientes
f3 <-> f4, cf3f4, NA   ### Covarianzas Factores Independientes
f3 <-> f5, cf3f5, NA
f4 <-> f5, cf4f5, NA

```

Para estimar los parámetros necesarios se usa la función *sem*, esta función recibe tres argumentos: modelo, matriz de correlaciones o covarianzas y el tamaño de la muestra aleatoria.

```

ejem.cap3 <- sem(mod.cap3, corr.ejem, N=122)

```

Para visualizar los resultados obtenidos se recurre a la función *summary*, la cual muestra los indicadores de Bondad de Ajuste y la estimación de los parámetros de las relaciones, las varianzas de los errores de medida y las covarianzas o correlaciones de los errores de predicción.

```
summary(ejem.cap3)
```

Y por último, para estimar los estadísticos que indican las modificaciones sugeridas al modelo, recurrimos a:

```
modIndices(ejem.cap3)
```

Para la Bondad de Ajuste es necesario revisar la siguiente información:

```
Model Chisquare = 15.059   Df = 13   Pr(>Chisq) = 0.30369
Chisquare (null model) = 256.5   Df = 28
Goodness-of-fit index = 0.96984
Adjusted goodness-of-fit index = 0.91649
RMSEA index = 0.036179   90% CI: (NA, 0.1007)
Bentler-Bonnett NFI = 0.94129
Tucker-Lewis NNFI = 0.98059
Bentler CFI = 0.99099
SRMR = 0.041668
AIC = 61.059
AICc = 26.324
BIC = 125.55
CAIC = -60.393
```

#### Normalized Residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.567	-0.117	0.000	0.126	0.208	1.780

#### R-square for Endogenous Variables

f1	v1	f2	v2	v3	v4	v5
0.3999	1.0000	0.3810	0.6719	0.5851	0.3428	0.4133
	v6	v7	v8			
0.5912	0.5043	1.0000				

Revisando los índices de ajuste incremental, se tiene que para la prueba Ji-cuadrada no se rechaza la hipótesis nula ya que el *p-value* es mayor a un nivel de significancia del 5%, i.e., no hay evidencia de que la matriz  $\Sigma$  estimada por el modelo y la matriz estimada  $S$  sean distintas. Por otra parte, se tiene que el GFI y AGFI son cercanos a 1 por lo que indican un buen ajuste, y finalmente el RMSEA es menor a 0.05 por lo que también indica un buen ajuste.

Ahora revisando los índices de ajuste incremental se tiene que NFI, NNFI y CFI son muy cercanos a uno. Por lo tanto hay evidencia que el ajuste del modelo es bueno y el modelo es confiable.

A pesar de que no hay evidencia para rechazar a este modelo en lo general, también es necesario revisar los parámetros de las relaciones y su significancia, para esto es necesario revisar los siguientes resultados:

```

Parameter Estimates
      Estimate Std Error z value Pr(>|z|)
lv3f2  0.93320 0.162166   5.75461 8.6844e-09 v3 <--- f2
lv5f3  1.09801 0.346912   3.16509 1.5504e-03 v5 <--- f3
lv7f4  0.92355 0.149236   6.18851 6.0737e-10 v7 <--- f4
pf1f2 -0.25828 0.487238  -0.53009 5.9605e-01 f1 <--- f2
pf1f4  0.95906 0.276981   3.46256 5.3507e-04 f1 <--- f4
pf1f5 -0.17870 0.102187  -1.74872 8.0339e-02 f1 <--- f5
pf2f1  0.30815 0.146193   2.10785 3.5044e-02 f2 <--- f1
pf2f3  0.51886 0.228624   2.26950 2.3238e-02 f2 <--- f3
varf3  0.34279 0.143252   2.39294 1.6714e-02 f3 <--> f3
varf4  0.59120 0.141490   4.17838 2.9360e-05 f4 <--> f4
varf5  1.00000 0.128565   7.77817 7.3578e-15 f5 <--> f5
vard1  0.60008 0.299007   2.00691 4.4759e-02 f1 <--> f1
vard2  0.41588 0.117329   3.54456 3.9326e-04 f2 <--> f2
vare2  0.32812 0.110616   2.96627 3.0144e-03 v2 <--> v2
vare3  0.41488 0.103804   3.99681 6.4201e-05 v3 <--> v3
vare4  0.65721 0.135169   4.86210 1.1615e-06 v4 <--> v4
vare5  0.58672 0.147631   3.97423 7.0608e-05 v5 <--> v5
vare6  0.40880 0.094948   4.30554 1.6658e-05 v6 <--> v6
vare7  0.49574 0.092796   5.34228 9.1783e-08 v7 <--> v7
cf3f4  0.18233 0.076023   2.39837 1.6468e-02 f4 <--> f3
cf4f5 -0.25040 0.087560  -2.85973 4.2400e-03 f5 <--> f4
cf3f5 -0.20418 0.081727  -2.49827 1.2480e-02 f5 <--> f3
cd1d2  0.14203 0.242876   0.58477 5.5870e-01 f2 <--> f1

```

Iterations = 68

Es posible notar que hay tres parámetros menores a 2: pf1f2, pf1f5 y cd1d2. Revisemos qué relación representan en el *diagrama de trayectoria* y determinemos si existe la posibilidad de eliminarlas del modelo:

- pf1f2, indica la causalidad que tiene f2 sobre f1, i.e., no hay evidencia estadística para determinar que la *Satisfacción* cause el *Desempeño*. No existe impedimento para eliminar esta relación ya que al final del día la idea de estimar una relación recíproca fue con la idea de inferir qué causa qué. Por lo tanto, el parámetro se considera significativa.
- pf1f5, indica la causalidad que tiene f5 sobre f1, la significancia de este parámetro está en el umbral para descartarlo del modelo, sin embargo, si fuésemos un poco flexibles, podríamos no eliminar esta relación que representa el impacto que tienen la *Inteligencia Verbal* sobre el *Desempeño*. Ahora bien, en caso de dejarla, lo que este parámetro nos indica es que a mayor inteligencia el desempeño disminuye; esta interpretación es justificable ya que el individuo puede llegar a encontrar aburrido y poco desafiante su trabajo, llevándole a poner un menor esfuerzo en él.
- cd1d2, este parámetro representa la covarianza entre los términos de error de los factores dependientes, en caso de que fuese significativa nos estaría diciendo que existen factores asociados compartiendo una variación común que no ha sido explicada en la relaciones expresadas en el modelo.

Con este análisis es posible dejar fuera el *Constraste Wald* ya que R no lo proporciona. En algunos otros software como SAS o AMOS este estadístico si está disponible, sin embargo, no es indispensable tenerlo.

Se debe tomar en cuenta que los motivos para modificar un modelo son para mejorar el ajuste o contrastar alguna hipótesis en particular. Hay que tener precaución ya que las relaciones que se aumenten o quiten deben estar sustentadas por un marco teórico.

Para la inclusión de relaciones se revisan los Multiplicadores de Lagrange. En R se extraen con la sentencia *modIndicates*. Para nuestro ejemplo queda de la siguiente manera:

```
modIndices(ejem.cap3)

5 largest modification indices, A matrix:
f5<-v3  f4<-v3  v8<-v3  v3<-f5  v3<-v8
7.190381 6.747341 6.722907 4.941608 4.941608

5 largest modification indices, P matrix:
f5<->v3  v8<->v3  f4<->v3  v5<->v4  f3<->f2
5.948519 5.191360 4.675407 4.012981 4.012967
```

Las relaciones que se podrían incluir al modelo para obtener una mejora significativa, eso implica que la probabilidad acumulada del estadístico dado por  $R$  sea menor a sea menor a 0.01 (o 0.05 en caso permitir un nivel de significancia mayor) de acuerdo a una distribución Ji-cuadrada con 1 grado de libertad, por lo tanto las relaciones que mejoran el modelo son:

```
5 largest modification indices, A matrix:
f5<-v3  f4<-v3  v8<-v3
7.190381 6.747341 6.722907
```

La primer propuesta es una relación de v3 a f5, sin embargo, esta relación ya está contemplada dentro del modelo a través de f5 pasando por f1 y terminando en f2. La segunda propuesta es similar a la primera, pero en vez de llegar a f5 llega a f4 por lo que no tiene sentido incluir esta relación. Finalmente, la tercer relación es una copia de la primer relación propuesta por lo que tampoco se agregará al modelo.

Para el caso de los parámetros de correlación ninguno de ellos es significativo. Por lo tanto no hay evidencia suficiente para incluir nuevas relaciones.

Corriendo nuevamente el programa ajustado con las modificaciones mencionadas. Los indices de Bondad de Ajuste quedan de la siguiente manera:

```
Model Chisquare = 15.4  Df = 15 Pr(>Chisq) = 0.42302
Chisquare (null model) = 256.5  Df = 28
Goodness-of-fit index = 0.96913
Adjusted goodness-of-fit index = 0.92591
RMSEA index = 0.014842  90% CI: (NA, 0.087811)
Bentler-Bonnett NFI = 0.93996
Tucker-Lewis NNFI = 0.99673
Bentler CFI = 0.99825
SRMR = 0.041933
```

AIC = 57.4  
 AICc = 24.64  
 BIC = 116.28  
 CAIC = -71.66

Comparando los resultados tenemos que las diferencias más significativas están en el  $p$ -value de la prueba Ji-cuadrada al pasar de 0.304 a 0.423 y el índice RMSEA al pasar de 0.0361 a 0.0148. El modelo quedaría expresado como en la figura 3.5.

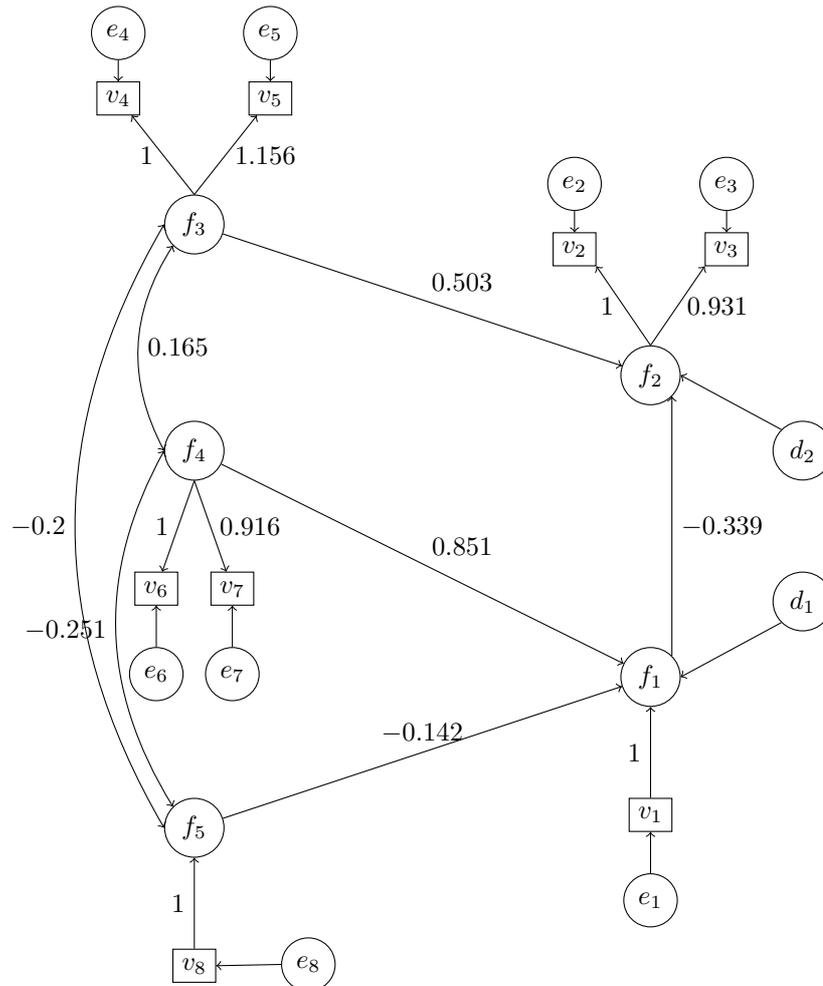


Figura 3.5: Modelo final de desempeño y satisfacción de la fuerza de ventas.

# Capítulo 4

## Aplicación

### 4.1. Antecedentes

Anteriormente los medios tradicionales, tales como Televisión, Radio, Prensa, Revista y Espectaculares, eran los principales canales de comunicación de una marca de consumo o servicio al público, esto hacía que la forma de medir los beneficios de la publicidad en los resultados de mercado (ventas, visitas a un sitio Web, recordación de marca, etc.) fueran a través de una relación completamente directa y se podía utilizar un modelo de regresión lineal. En las áreas de mercadotecnia y negocios, al procedimiento de aplicar el Análisis de Regresión para obtener los resultados mencionados se le conoce como *Marketing Mix Modeling* (MMM) (Tellis, 2006).

El objetivo del MMM es cuantificar el impacto de los esfuerzos realizados (industriales, publicitarios, producto, etc.) a los resultados de negocio. Por lo tanto, debido a que estos esfuerzos tienen una inversión y un beneficio económico es posible cuantificar el retorno de inversión (RoI) y finalmente obtener aprendizajes sobre los medios más eficientes y poder construir estrategias de negocio, en particular, planes de inversión publicitaria tal que, genere un mayor volumen de ventas a un menor costo.

Sin embargo, hoy en día el uso de las redes sociales y la publicidad en internet han revolucionado la forma en que las marcas se comunican con el consumidor, ya que hay diversos puntos de contacto que interactúan entre sí (véase la figura 4.1).

Con este nuevo comportamiento de la publicidad en marketing, el MMM que se mencionó anteriormente, es obsoleto ya que sólo permite cuantificar las relaciones directas dejando las relaciones indirectas (relaciones que interactúan entre sí) fuera del análisis de los datos.

Una técnica estadística que permite analizar las interrelaciones en el mismo modelo como en la figura 4.1, es el *Análisis de Trayectoria* (AT). Sin embargo, como se mencionó en capítulos anteriores, el AT es una de las técnicas que no se había explotado debido a la limitación tecnológica de la época en que fue desarrollada. Los Modelos de Ecuaciones Estructurales (MEE) colaboraron en este sentido ya que los softwares estadísticos que calculan los MEE también calculan el AT.

El objetivo de marcas de consumo son las ventas, y en algunas ocasiones la recordación de marca es el objetivo de todos los esfuerzos publicitarios. En este ejercicio solo analizaremos los efectos que tiene la publicidad a las ventas.

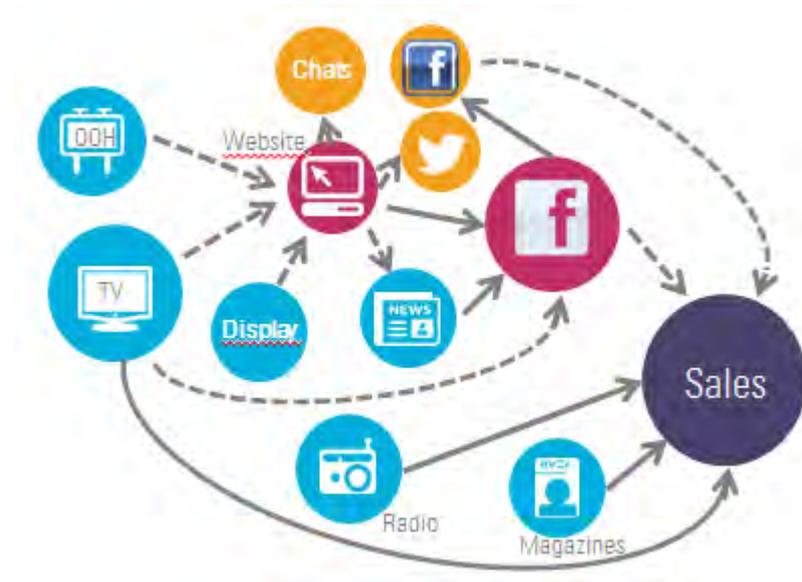


Figura 4.1: Nuevos puntos de contacto publicitarios.

## 4.2. Objetivo de la aplicación del Análisis de Trayectoria

Adicionalmente a confirmar la hipótesis sobre la trayectoria que tienen los efectos publicitarios hacia las ventas, hay otros beneficios relevantes que se obtienen de este análisis versus un análisis de regresión, estos son:

1. Estimar los efectos directos e indirectos de las variables incluidas en el modelo hacia las ventas, a través de la descomposición del MEE.
2. Encontrar los puntos de contacto más relevantes entre el consumidor y la marca.
3. Encontrar la ruta más efectiva entre la inversión realizada en medios publicitarios y las ventas.

Los resultados obtenidos del análisis deben tener una visión de negocios en aras de tomar decisiones en colaboración con otras áreas para una aplicación efectiva y que ayuden a los objetivos de negocio.

## 4.3. Desarrollo del modelo

Para explicar el desarrollo del modelo se dividirá en las siguientes fases:

1. Conocimiento de la Marca.
2. Recopilación y estructuración de datos.
3. Análisis descriptivo de las variables.

4. Planteamiento del modelo (aplicación del Análisis de Trayectoria).
5. Resultados y bondad de ajuste del modelo.
6. Discusión y comentarios finales.

#### 4.3.1. Conocimiento de la Marca

Esta etapa es para conocer las costumbres de la marca, quién es su consumidor, hábitos de ellos, periodos estacionales, problemáticas del producto, tipo de comunicación que ha realizado, puntos de venta relevantes, principales competidores, etc.

A partir de esta información es relevante entender la tendencia y el comportamiento de las ventas, así como sus puntos atípicos y los esfuerzos que ha realizado la marca. Entiéndase por esfuerzos: sitios web, publicaciones en redes sociales, promociones, eventos, asistencias del uso de un producto, etc.

Una vez que se tiene conocimiento del comportamiento de la marca se comienza con la recopilación de la información con que se cuenta, definir su temporalidad y las unidades y escalas en que se mide cada una de ellas.

El ejercicio que se presenta es para una marca de Queso-Crema, la cual presenta más de tres cuartas partes de sus ventas en los canales de autoservicio. Cuenta con redes sociales, tales como Facebook y Twitter donde realiza publicaciones de promociones o diversos usos del producto, un sitio web en el que su actividad se centra en comunicar recetas de cocina sobre el uso del producto, concursos o promociones, canal de YouTube en el que sube videos explicando el proceso de elaboración de los platillos que se pueden hacer con su producto y, por último, se han abierto distintos canales de comunicación entre las usuarios de sus productos tales como Whatsapp o Call Center.

Por tanto, la principal comunicación que tiene la marca es en medios digitales y es enfocada al uso de las diversas aplicaciones que tiene el producto.

#### 4.3.2. Recopilación y estructuración de datos

Esta es una de las etapas pilares para la construcción del modelo, pues datos erróneos dan resultados incorrectos. La recopilación de la información puede ser un proceso largo, ya que podría encontrarse en diversas fuentes, como Excel, Power Point o incluso hasta en PDF, pues si bien es cierto que el fenómeno *Big Data* está de moda, no se cuenta con una cultura de mantener la información estructurada.

Un problema que podría presentarse al realizar el análisis es la poca disponibilidad de información, pues a pesar de toda la información que genera el mundo digital, cuando la información se estructura en las condiciones para la alimentación del modelo, los datos se reducen de manera considerable. Otra razón puede ser porque los esfuerzos realizados son recientes (menores de 2 años) y los datos son de forma mensual lo que se traduce en 24 observaciones.

Por tanto, para generar el número adecuado de observaciones en la estimación del modelo se tomarán observaciones semanales desde enero de 2011 a junio de 2013. Cada observación refiere a la actividad acumulada en ese periodo. Por ejemplo, las ventas de la semana del 9 de enero de 2011 corresponden a las ventas totales. En el caso del precio y la distribución se interpreta como un promedio.

En la tabla 4.1 se presentan las variables que estaremos trabajando y el re-etiquetado que se utilizará en cada una de las variables en R. A continuación, se da una breve explicación de cada una de ellas:

1. Ventas. Muestra los kilogramos del queso que se vendió al menudeo en las cadenas de autoservicio.
2. Precio. Precio promedio por cada kg que se registró en las cadenas de autoservicio.
3. Distribución. Porcentaje del total de tiendas de autoservicio en que el producto está disponible.
4. Inversión Digital.
5. Inversión Exteriores/Revistas.
6. Inversión Tv Paga.
7. Inversión Tv Abierta.
8. Fans en Facebook. Se registra el número de personas que siguen a la marca en dicha red social.
9. Asistencia Online. El número de personas que recibieron ayuda o ideas para hacer un platillo con el producto, a través de videos en YouTube, o que visitaron la página electrónica de la marca.
10. Asistencia en Tiempo Real. Número de personas que realizaron alguna llamada al *Call Center*, o que enviaron un mensaje de *Whatsapp*, o que chatearon en línea.
11. Ruido en Internet. Número de comentarios que realiza cualquier usuario de internet en cualquier plataforma.

<b>Tipo de Variables</b>	<b>Nombre de Variables</b>	$v_i$
Beneficios de mercado	Ventas	$v_1$
Información del producto	Precio	$v_2$
	Distribución	$v_3$
Información de medios	Inversión Digital	$v_4$
	Inversión Exteriores/Revistas	$v_5$
	Inversión Tv Paga	$v_6$
	Inversión Tv Abierta	$v_7$
	Facebook Fans	$v_8$
	A. Online	$v_9$
	A. Tiempo Real	$v_{10}$
	Ruido en internet	$v_{11}$

Tabla 4.1: Variables para el Análisis de Trayectorias.

Las variables que son de inversión, hacen referencia a la cantidad monetaria en pesos mexicanos que se realizó para tener publicidad en el medio que lleva el nombre de la variable. Esta inversión,

es el resultado de un análisis estratégico considerando diversos objetivos como: recordación de marca, ventas, consideración de compra, incrementar algún atributo de marca, etc.

Cabe mencionar que no necesariamente existe una metodología para la asignación de inversión y en algunas ocasiones se decide a través de la experiencia de los directivos de la empresa.

Por otro lado, el impacto de la inversión publicitaria en las ventas, se asume que tiene un efecto adicional una vez que la publicidad impacta en el consumidor; este efecto es conocido como *Adstock*, el cual se interpreta como un efecto de recordación que mantiene el consumidor de la publicidad vista. La forma de medirlo es suponer que hay un porcentaje de la inversión que sigue impactando en el consumidor.

También se asume que la publicidad no tiene un efecto inmediato, es decir, si hoy el consumidor está expuesto a la publicidad, no necesariamente acude al autoservicio a adquirir el producto, este efecto se conoce como *Lag*. En este caso un lag 1 se interpreta como: la publicidad se convierte en acto de compra una semana después de haber realizado la inversión; la forma de aplicarlo en la información es desfazar los datos una semana.

En el apéndice B, se describen con mayor detalle sobre estas transformaciones, las cuales se aplican a las inversiones que se incluyen en el modelo. En ambos casos, el porcentaje de recordación y retardo de consumo son dados por los analistas que planifican la inversión.

Se crean dos tablas de datos, una de ellas tendrá la información sin realizar alguna transformación en las variables de inversión, esto para realizar un análisis exploratorio y entender los niveles de inversión que se han realizado. La otra tabla, contiene las variables de inversión con las transformaciones; esta tabla es la que se usa para la estimación del modelo. Ambas tablas de datos tienen las variables que se mencionan en la tabla 4.1.

### 4.3.3. Análisis descriptivo de las variables

En general, los datos que el analista debe tener claro son: el valor promedio que toma cada una de las variables, la desviación estándar, valor mínimo y máximo; en este caso algunas veces es subjetivo y se tendría que revisar si alguno de estos valores podría ser un punto atípico.

En el caso de los puntos atípicos, se dan por dos principales motivos: (1) error en la captura de la información, o (2) un comportamiento distinto al de mercado. Para cualquiera de los dos puntos, lo primero es consultar al proveedor de la información con la finalidad de entender qué sucedió.

En caso de que el punto atípico se deba a (1), hay dos soluciones: una es sustituirlos por los datos correctos que será dado por el proveedor de los datos, o hacer un promedio de los datos de las mismas semanas en los diferentes años y revisar si los datos mantienen la tendencia. También existen procesos especiales para abordar el problema de datos faltantes, a este proceso se le conoce como imputación de la información, ver por ejemplo Allison (2003).

Para el caso (2), la atipicidad puede suceder debido a épocas estacionales, campañas especiales. Un ejemplo son las ventas de vino en el periodo de "*Guadalupe-Reyes*", el hecho de que un envase de refresco lleve tu nombre, etc.

A través de la experiencia en otros mercados de la aplicación de estos modelos, la agencia ha sugerido un cambio en los valores atípicos que se puedan encontrar debido a que no es un comportamiento natural del mercado, es decir, en el caso puntual del pico más alto en ventas (incremento de hasta el 52% de una semana a otra) se debió a una promoción realizada por una cadena de supermercados y no por una oferta general hecha por la marca o que se deba a un

impacto de la comunicación. En el caso de las inversiones en medios de comunicación, aquellas considerablemente altas se deben a temas de administración de presupuestos o negociación de costos por lo que no necesariamente, ese gasto se relaciona con la calidad del impacto que pueda tener el mensaje publicado.

La agencia cuenta con un criterio para la detección de puntos atípicos. Este es establecido a través de la experiencia al realizar diversos modelos estadísticos similares a éste. Se considera punto atípico si el máximo de la serie es mayor que el promedio (sin considerar ceros) más tres veces la desviación estándar.

En caso de encontrar puntos atípicos, lo recomendable es disminuirlos al máximo nivel aceptado, es decir, al promedio (sin considerar ceros) más tres veces la desviación estándar calculado con los puntos atípicos detectados.

Se ajusta a este nuevo máximo nivel, ya que sí es necesario tener un alto valor, porque éste se debió a una acción extraordinaria y que podría ayudar a cuantificar un resultado en caso de que se realice alguna otra estrategia similar.

En la tabla 4.2 se muestran los estadísticos básicos de las variables y el límite para determinar si hay puntos atípicos. Para las atipicidades encontradas se mostrará una gráfica de líneas con el objetivo de visualizar cuáles son los puntos que sobrepasan los límites permitidos. También se muestra una gráfica de dispersión de la variable que presenta atipicidades versus las ventas para poder visualizar cómo es afectada la pendiente de la recta en caso de dejar los outliers.

Las variables que presentan puntos atípicos son: Ventas, Ext/Rev, Tv Paga y R. Internet, las cuales se grafican en las figuras 4.2, 4.3, 4.4, y 4.5, y que analizaremos a continuación.

En la figura 4.2 se puede ver los puntos atípicos en ventas que se presentan en las últimas semanas del año. Esto se asocia a las fiestas decembrinas; hay dos casos: el primero es en 2012 donde, no hay evidencia de que las ventas hayan sido ajenas a un comportamiento no natural del mercado, por lo que la observación se queda sin cambios. En el segundo caso sí se encontró un pico de ventas y tal como se mencionó anteriormente, este incremental en ventas se debió a una acción promocional de una cadena de autoservicios en particular y no necesariamente a un acto publicitario.

En la figura 4.3 se puede observar la inversión de Exteriores y Revistas; se presentan dos sema-

Variable	Prom S/C	Desv. Std	Máximo	Límite de Outlier
Ventas ( $v_1$ )	132,792.71	19,911.97	235,965.13	192,528.63
Precio ( $v_2$ )	17.24	0.85	19.23	19.78
Dist ( $v_3$ )	94.83	1.18	98.00	98.38
Digital ( $v_4$ )	47,961.17	31,197.87	138,375.62	141,554.79
Ext/Rev ( $v_5$ )	376,649.99	208,106.16	1,274,210.00	1,000,968.48
Tv Paga ( $v_6$ )	350,334.25	241,146.67	1,114,899.60	1073,774.25
Tv Abierta ( $v_7$ )	1,127,652.94	688,429.23	3,013,006.03	3,192,940.64
F. Fans ( $v_8$ )	36,276.20	30,129.53	124,008.00	126,664.78
A. Online ( $v_9$ )	35,956.42	12,905.57	70,396.00	74,673.11
A.T.Real ( $v_{10}$ )	341.38	246.68	878.00	1,081.41
R. Internet ( $v_{11}$ )	2,764.57	5,093.82	34,509.00	18,046.03

Tabla 4.2: Estadísticas descriptivas.

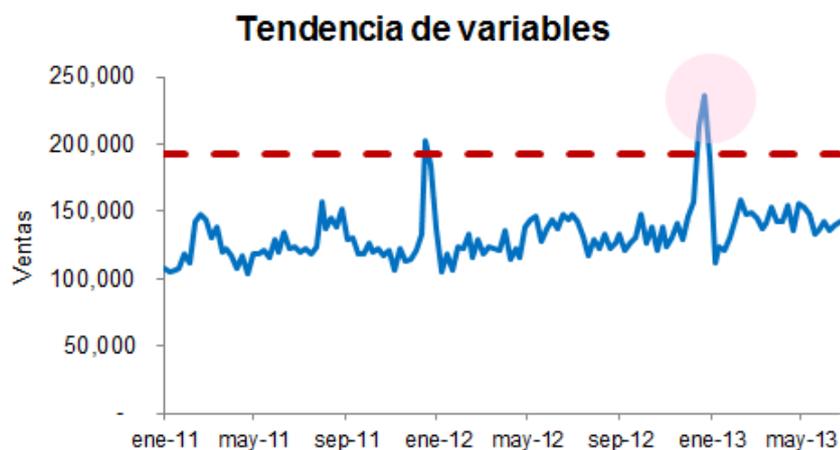


Figura 4.2: Tendencia de ventas.

nas atípicas, la mayor de ellas se presenta en 27 de enero de 2013 y la otra el 31 de marzo del mismo año. Estas inversiones están asociadas a una negociación inadecuada con un *Network* en los costos de los espacios ofrecidos, por lo tanto no hay una justificación orientada a un resultado de negocio por lo que se deberían ajustar los picos al nivel permitido.

En la figura 4.4 se visualiza la inversión de Tv de Paga, hay un punto atípico en que se encuentra el 6 de mayo de 2012, esta inversión estuvo asociada a un mensaje en especial por el día de las madres, por lo que es un movimiento natural del mercado. En este caso, la inversión sobrepasa un 4%, en caso de hacer la reducción al límite permitido no habría una diferencia relevante por lo que se quedará con el dato original.

La figura 4.5 presenta el caso de Ruido en Internet. Las semanas atípicas son las últimas dos del periodo de análisis. Se revisó la información con el proveedor y confirmó un error en ella, sin embargo, no se cuenta con los datos correctos por lo que las observaciones serán ajustadas al máximo permitido.

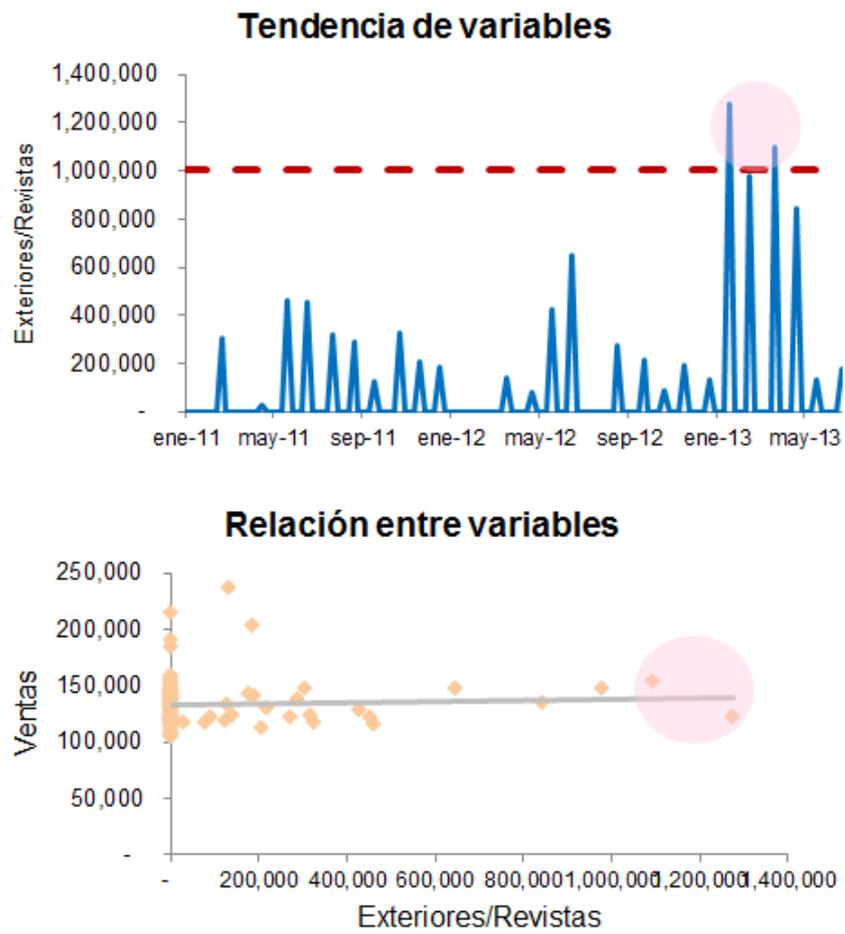


Figura 4.3: Tendencia de inversión en Exteriores y Revistas.

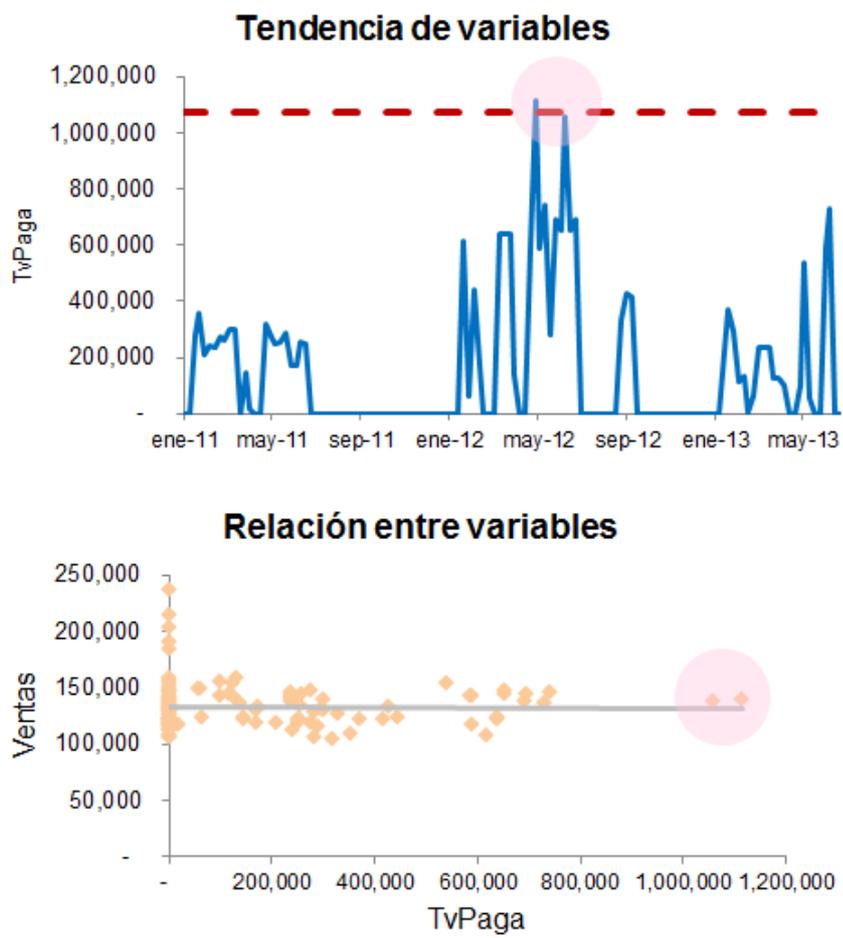


Figura 4.4: Tendencia de inversión en Televisión de Paga.

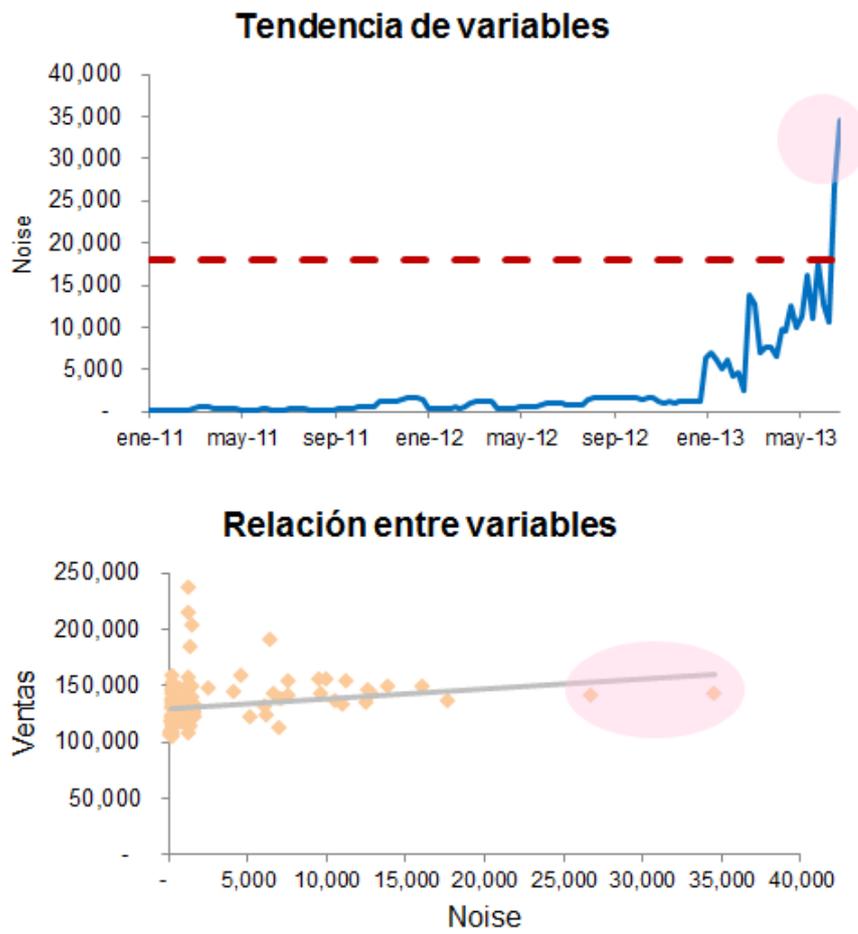


Figura 4.5: Tendencia de Ruido en Internet.

#### 4.3.4. Planteamiento del modelo

El modelo que se desea estimar se muestra en la figura 4.6, el cual es construido a partir del conocimiento estratégico que tienen los especialistas en marketing, el cliente y el analista estadístico.

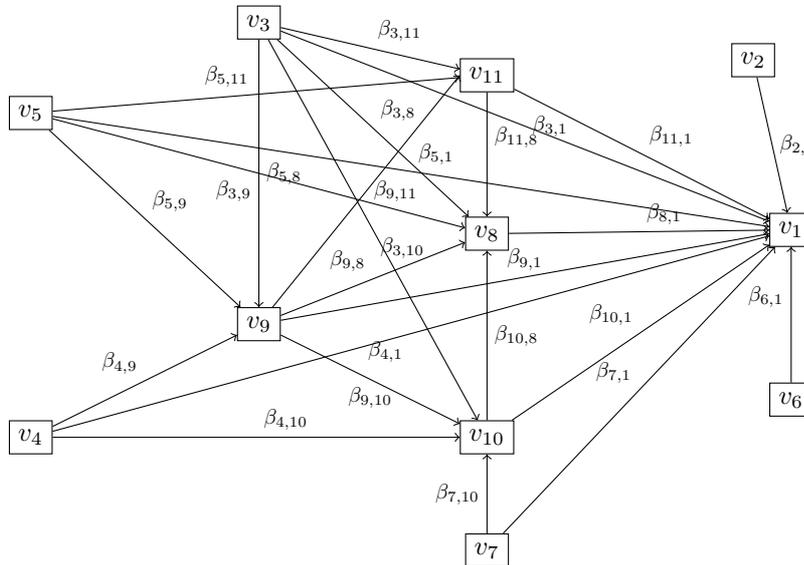


Figura 4.6: Modelo teórico publicitario para Queso-Crema.

En el diagrama de la figura 4.6, las inversiones en medios ( $v_4, \dots, v_7$ ) deben ser puramente exógenas ya que se supone que un especialista es quien las determina y no dependen de ninguna otra variable. Por otro lado, no todas las variables tienen un efecto directo hacia las ventas debido a que su objetivo es otro, por lo que tendrán un efecto indirecto.

A continuación se da un resumen conceptual del modelo:

- La distribución de la marca ( $v_3$ ) ayuda a incrementar los niveles de Fans ( $v_8$ ), asistencia Online ( $v_9$ ), asistencia en tiempo real ( $v_{10}$ ) y el ruido en internet ( $v_{11}$ ), así como las ventas ( $v_1$ ). Esto tiene sentido, ya que si el consumidor encuentra el producto en el punto de venta entonces ayuda a tener interacción con el resto de puntos de contacto.
- La inversión en TV de Paga ( $v_6$ ) sólo mantiene una estrategia para generar ventas, por lo tanto, sólo se confirma la relación a venta ( $v_1$ ).
- Como se podrá apreciar, el precio ( $v_2$ ) sólo tiene relación con las ventas ya que no hay asociación a ninguna otra variable publicitaria. La hipótesis es que a medida que el precio aumenta, las ventas decrecen.
- En general, todas las inversiones con excepción de TV de Paga (es decir,  $v_4, v_5$  y  $v_7$ ) tienen una relación directa e indirecta. Esto a través del resto de puntos de contactos.

Para la estimación del modelo se utilizará la matriz de correlación. Al utilizar las correlaciones, los resultados del modelo serán obtenidos con coeficientes estandarizados y esto facilita la comparación entre ellos en el diagrama de trayectorias. La matriz de correlaciones se muestra en la

tabla 4.3. Estas correlaciones están calculadas sin los datos atípicos encontrados en la sección de las estadísticas descriptivas.

	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$	$v_6$	$v_7$	$v_8$	$v_9$	$v_{10}$	$v_{11}$
$v_1$	1	-0.14	0.4	0.31	0.28	-0.02	0.21	0.39	0.44	0.36	0.22
$v_2$	-0.14	1	-0.01	0.13	-0.29	-0.03	-0.01	-0.02	-0.02	0.06	-0.09
$v_3$	0.4	-0.01	1	0.09	0.29	0.07	0.29	0.69	0.44	0.48	0.43
$v_4$	0.31	0.13	0.09	1	-0.14	0.28	0.22	0.06	0.32	0.38	-0.13
$v_5$	0.28	-0.29	0.29	-0.14	1	-0.04	0.17	0.54	0.29	-0.06	0.5
$v_6$	-0.02	-0.03	0.07	0.28	-0.04	1	0.02	0.05	0.12	0.18	-0.04
$v_7$	0.21	-0.01	0.29	0.22	0.17	0.02	1	0.25	0.21	0.22	0.13
$v_8$	0.39	-0.02	0.69	0.06	0.54	0.05	0.25	1	0.62	0.39	0.8
$v_9$	0.44	-0.02	0.44	0.32	0.29	0.12	0.21	0.62	1	0.39	0.34
$v_{10}$	0.36	0.06	0.48	0.38	-0.06	0.18	0.22	0.39	0.39	1	-0.01
$v_{11}$	0.22	-0.09	0.43	-0.13	0.5	-0.04	0.13	0.8	0.34	-0.01	1

Tabla 4.3: Matriz de correlaciones del modelo de Queso-Crema.

#### 4.3.5. Resultados y bondad de ajuste del modelo

Utilizando el software R con el paquete `sem`, el modelo planteado en el diagrama 4.6 se obtuvo una estimación satisfactoria ya que con 31 interacciones se consigue la convergencia en la solución del sistema de ecuaciones.

Para llegar a un modelo que pueda ser aceptado como explicación del comportamiento de los medios publicitarios a las ventas es necesario evaluar distintas aristas de él, es decir, tenemos que evaluar:

1. El signo y la significancia de los coeficientes de cada relación.
2. Los principales indicadores de bondad de ajuste.
3. Las contribuciones de cada una de las variables.

Los resultados se muestran en la tabla 4.4, éstos se muestran organizados por *Secciones*, las cuales están asociadas a las variables endógenas.

La última sección nos dice que el *Ruido en internet* ( $v_{11}$ ) está explicada por la Distribución del producto ( $v_3$ ), la inversión en Exterior&Revistas ( $v_5$ ) y la Asistencia Online ( $v_9$ ). La variable que más aporta en este caso es la inversión en Exteriores&Revista, siendo ésta una conexión interesante porque comunicar la marca en medios fuera de casa motiva a generar conversación en la red. La inversión Online en su relación con Ruido en Internet no es significativa. En este caso, se debe verificar con el equipo de estrategia ya que, estadísticamente no hay evidencia que la actividad esté generando un incremento.

En la penúltima sección, la *A. Tiempo Real* ( $v_{10}$ ) se construye debido a la Distribución del producto ( $v_3$ ), la inversión en digital ( $v_4$ ), la inversión en Tv Abierta ( $v_7$ ) y la Asistencia Online ( $v_9$ ). La distribución del producto es la que más beneficia a la Asistencia en Tiempo Real. Esto tiene sentido ya que para tener interacción con la marca, una primer interacción en el punto de venta se muestra al ser eficiente. En el caso de la relación de TV Abierta a Asistencia en Tiempo Real, es una relación no significativa. La inversión Online en su relación con Asistencia

Endógenas	Exógenas	Coef.	Std.Err	Z-value	p-value
Ventas ( $v_1$ )	Precio ( $v_2$ )	-0.109	0.074	-1.483	0.138
	Dist ( $v_3$ )	0.235	0.102	2.303	0.021
	Digital ( $v_4$ )	0.254	0.086	2.971	0.003
	Ext/Rev ( $v_5$ )	0.211	0.094	2.248	0.025
	Tv Paga ( $v_6$ )	-0.145	0.073	-1.975	0.048
	Tv Abierta ( $v_7$ )	0.005	0.076	0.062	0.951
	F. Fans ( $v_8$ )	-0.264	0.222	-1.189	0.234
	A. Online ( $v_9$ )	0.235	0.097	2.414	0.016
	A.T.Real ( $v_{10}$ )	0.203	0.101	2.004	0.045
R. Internet ( $v_{11}$ )	0.166	0.154	1.079	0.281	
F. Fans ( $v_8$ )	Dist ( $v_3$ )	0.200	0.035	5.657	0
	Ext/Rev ( $v_5$ )	0.152	0.032	4.681	0
	A. Online ( $v_9$ )	0.202	0.033	6.195	0
	A.T.Real ( $v_{10}$ )	0.234	0.033	7.182	0
	R. Internet ( $v_{11}$ )	0.575	0.034	16.771	0
A. Online ( $v_9$ )	Dist ( $v_3$ )	0.345	0.076	4.559	0
	Digital ( $v_4$ )	0.326	0.073	4.452	0
	Ext/Rev ( $v_5$ )	0.239	0.076	3.144	0.002
A.T.Real ( $v_{10}$ )	Dist ( $v_3$ )	0.404	0.081	4.989	0
	Digital ( $v_4$ )	0.302	0.076	3.987	0
	Tv Abierta ( $v_7$ )	0.008	0.075	0.110	0.912
	A. Online ( $v_9$ )	0.112	0.083	1.346	0.178
R. Internet ( $v_{11}$ )	Dist ( $v_3$ )	0.274	0.080	3.414	0.001
	Ext/Rev ( $v_5$ )	0.394	0.075	5.242	0
	A. Online ( $v_9$ )	0.099	0.080	1.225	0.221

Tabla 4.4: Estimación del Análisis de Trayectoria.

en Tiempo Real no es significativo. En este caso, se deben verificar estas relaciones con el equipo de estrategia ya que, estadísticamente, no hay evidencia que estas actividades estén generando incremento significativo.

La *Asistencia Online* ( $v_9$ ) es consecuencia de la Distribución de producto ( $v_3$ ), Inversión en Digital ( $v_4$ ) e inversión Exterior&Revistas ( $v_5$ ). En general, el impacto de las tres es similar, estando ligeramente por arriba la distribución del producto.

Los Fans ( $v_8$ ) se construyen a través de la Distribución del producto ( $v_3$ ), la inversión en Exteriores&Revistas ( $v_5$ ), la Asistencia Online ( $v_9$ ), la Asistencia en Tiempo Real ( $v_{10}$ ) y el Ruido en internet ( $v_{11}$ ). El Ruido en internet muestra tener el mayor impacto para que el consumidor se haga Fan en la cuenta de Facebook de la marca, lo cual tiene sentido, ya que las diferentes conversaciones en Internet motivan a buscar el Fanpage y a ser Fan.

Con respecto a la sección de ventas ( $v_1$ ), si los coeficientes mantienen un signo positivo indican que toda acción ayuda a la construcción de ventas y de los puntos de contacto intermedios, como Fans o Ruido en internet. Con respecto a la relación entre Precio y Ventas, tiene sentido que tenga signo negativo, su interpretación es que a mayor aumento de precio hay una disminución de ventas.

En la tabla 4.4 podemos ver que adicional a Precio, Tv de Paga y Fans mantienen un signo

Endógenas	Exógenas	Efecto Total	Efecto Directo	Efecto Indirecto
Ventas( $v_1$ )	Precio( $v_2$ )	-0.109	-0.109	0
	Dist( $v_3$ )	0.313	0.235	0.077
	Digital( $v_4$ )	0.362	0.254	0.108
	Ext/Rev( $v_5$ )	0.224	0.211	0.013
	Tv Paga( $v_6$ )	-0.145	-0.145	0
	Tv Abierta( $v_7$ )	0.006	0.005	0.001
	F. Fans( $v_8$ )	-0.264	-0.264	0
	A. Online( $v_9$ )	0.199	0.235	-0.036
	A.T.Real( $v_{10}$ )	0.142	0.203	-0.062
	R. Internet( $v_{11}$ )	0.015	0.166	-0.152
	F. Fans( $v_8$ )	Dist( $v_3$ )	0.55	0.2
Digital( $v_4$ )		0.164	0	0.164
Ext/Rev( $v_5$ )		0.447	0.152	0.295
Tv Paga( $v_6$ )		0.002	0	0.002
Tv Abierta( $v_7$ )		0.285	0.202	0.083
F. Fans( $v_8$ )		0.234	0.234	0
A. Online( $v_9$ )		0.575	0.575	0
A. Online( $v_9$ )	Dist( $v_3$ )	0.345	0.345	0
	Digital( $v_4$ )	0.326	0.326	0
	Ext/Rev( $v_5$ )	0.239	0.239	0
A.T.Real( $v_{10}$ )	Dist( $v_3$ )	0.443	0.404	0.039
	Digital( $v_4$ )	0.339	0.302	0.036
	Ext/Rev( $v_5$ )	0.027	0	0.027
	Tv Abierta( $v_7$ )	0.008	0.008	0
	A. Online( $v_9$ )	0.112	0.112	0
R. Internet( $v_{11}$ )	Dist( $v_3$ )	0.308	0.274	0.034
	Digital( $v_4$ )	0.032	0	0.032
	Ext/Rev( $v_5$ )	0.418	0.394	0.024
	A. Online( $v_9$ )	0.099	0.099	0

Tabla 4.5: Efectos Totales, Directos e Indirectos del Análisis de Trayectoria para las ventas de Queso-Crema.

negativo en relación con ventas. Para el caso de Tv de Paga, el parámetro está cerca de no ser significativo a un nivel del 5%; mientras que los Fans no lo son. La relación negativa de los Fans a las ventas se puede deber a que las estrategias usadas hasta el momento podrían estar motivando un indicador de salud de marca, como lo es Recordación de marca. Por otro lado, Tv Abierta es una variable no significativa. En este caso, el equipo que planea la comunicación en este medio confirmó que TV, al ser un medio con mayor penetración <sup>1</sup>, es usado para que el consumidor conozca la marca.

En la tabla 4.5 está el cálculo de los efectos Totales, Directos e Indirectos. En ella se observa que hay efectos adicionales a los Directos en los Fans ( $v_8$ ) solo con efectos Indirectos de la inversión en Digital ( $v_4$ ) y de Tv Abierta ( $v_7$ ) que son variables que impactan a través de la Asistencia en Tiempo Real ( $v_{10}$ ). Mismo caso para la Asistencia en Tiempo Real ( $v_{10}$ ) con efectos Indirectos de

<sup>1</sup>Porcentaje de personas u hogares que físicamente están en posibilidad de ser expuestos a ese medio (Domínguez y Hermo, 2007).

la inversión en Exteriores y Revistas ( $v_5$ ) por medio de la Asistencia Online ( $v_9$ ). Finalmente, el Ruido en Internet ( $v_{11}$ ) es impactado por la Inversión en Digital ( $v_4$ ) a través de la asistencia Online ( $v_9$ ).

Por otra parte, se observa que hay tres efectos indirectos negativos sobre las ventas, lo que ocasiona que el efecto Total sea menor que el efecto Directo, esto sucede para las variables Asistencia Online ( $v_9$ ), Asistencia en Tiempo Real ( $v_{10}$ ) y el Ruido en Internet ( $v_{11}$ ). Estos efectos son negativos debido a que ellas tienen un efecto Directo en los Fans ( $v_8$ ) y estos impactan de manera negativa en las ventas.

En el caso de la evaluación estadística de bondad de ajuste, en la tabla 4.6 podemos observar que el  $p$ -value asociado a la prueba Ji-cuadrada es 0.046, lo cual no está por arriba de 0.05, sin embargo, está muy cerca, por lo tanto con un  $p$ -value igual a 0.046 la matriz de correlación muestral es igual a la matriz de correlación conformada con los parámetros del modelo. En el caso del indicador CFI, tienen un valor de 0.98; lo que es muy cercano a 1. SRMR, que debería ser menor o igual que 5%, está en el límite.

Indicador	Valor
Number of observations	130
Estimator	ML
Minimum Function Test Statistic	25.3
Degrees of freedom	15
p-value (Chi-square)	0.046
Model test baseline model:	0
Minimum Function Test Statistic	548.828
Degrees of freedom	40
P-value	0
User model versus baseline model:	0
Comparative Fit Index (CFI)	0.98
Tucker-Lewis Index (TLI)	0.946
Loglikelihood and Information Criteria:	0
Loglikelihood user model (H0)	-1731.662
Loglikelihood unrestricted model (H1)	-1719.012
Number of free parameters	30
Akaike (AIC)	3523.324
Bayesian (BIC)	3609.35
Sample-size adjusted Bayesian (BIC)	3514.467
Root Mean Square Error of Approximation:	0
RMSEA	0.073
90 Percent Confidence Interval	[0.010-0.120]
P-value RMSEA $\leq$ 0.05	0.205
Standardized Root Mean Square Residual:	0
SRMR	0.05

Tabla 4.6: Indicadores de bondad de ajuste.

El RMSEA es mayor a 0.073, que es mayor a 0.05 pero no mayor a 0.10. Esto representa 0.023 mayor a lo recomendado por muchos autores. Sin embargo, algunos autores como Browne y Cudeck (1993) recomendaron que “un valor del RMSEA menor o igual que 0.05 indicaría un buen ajuste del modelo en relación a sus grados de libertad”, y que “un valor del RMSEA menor

o igual que 0.08 indicaría un error razonable de aproximación”, además de que “no es deseable emplear un modelo con un RMSEA mayor que 0.1”. Otro trabajo interesante sobre el punto de corte óptimo del RMSEA es el de Chen et al. (2008). Ellos hicieron un amplio estudio de simulación y concluyen que “un punto de corte del RMSEA de 0.05 rechaza demasiados modelos válidos con tamaños de muestra pequeños ( $n \leq 100$ )”, además dicen que “el punto de corte 0.05 funciona mejor en tamaños de muestra grandes (aunque tiene a sobre-aceptar cuando  $n \geq 800$ )”, también mencionan que “cuando el intervalo de confianza del RMSEA se usa conjuntamente al estimador puntual del RMSEA, no hay una justificación para usar de manera estricta el punto de corte 0.05”. En conclusión, considero que un RMSEA de 0.073 es aceptable, siendo un error moderado para el modelo propuesto, y dado que el número de observaciones en la muestra es de  $n = 130$ .

En resumen, el modelo presenta algunos índices que no lo están favoreciendo, sin embargo, la diferencia entre lo aceptable y los resultados es poca. Este modelo podría presentarse mencionando que hay un leve grado de error mayor al recomendado en la teoría estadística.

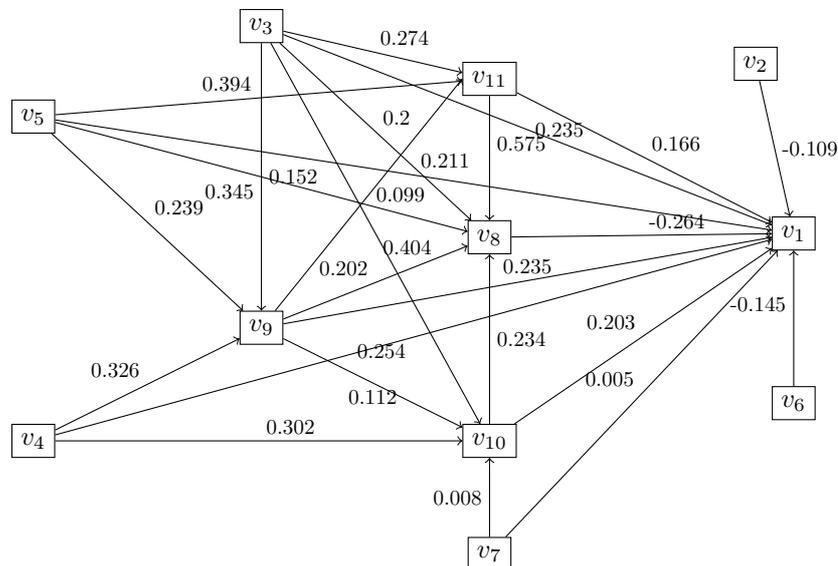


Figura 4.7: Modelo teórico publicitario para Queso-Crema.

#### 4.3.6. Discusión y comentarios finales

Estadísticamente el modelo muestra un ajuste marginal, es decir, algunos indicadores difieren un poco más de lo permitido por la literatura estadística. Sin embargo, sí nos quedaremos con el modelo estimado teniendo en cuenta que hay una ligera diferencia con la recomendación teórica y los resultados obtenidos.

Por lo tanto, podemos tomar el diagrama de la figura 4.7 como una representación del comportamiento de los esfuerzos publicitarios. Con este modelo los expertos en el área de planificación de medios podrían generar estrategias con un menor costo y con mayor impacto, o en otro caso, se podrían realizar cambios en los tipos de mensajes emitidos ya que los actuales no cumplen con los objetivos establecidos.

Dentro del modelo estimado, hay relaciones que no se esperaba tuvieran un efecto negativo a las ventas, como es el caso de la inversión de TV de Paga o los Fans. En este caso, las primeras acciones es generar un cambio de mensaje en estos medios. Este cambio de mensaje debe ser paulatino, y no al 100 %, ya que primero se debe explorar qué estrategia sería la adecuada para el medio y, segundo, dejar de hacer lo que se hace en la actualidad podría ocasionar una disminución considerable en aquellos indicadores que sí estén afectados por este medio.

En la figura 4.7 podemos darnos cuenta que la ruta con mayor impacto es a través de la distribución del producto, lo cual tiene sentido, ya que si el producto es encontrado en distintos puntos de venta generará que el consumidor interactúe con la marca en distintos puntos de contacto, en particular, en redes sociales.

Se debe tener en cuenta que, incrementar 1 PP el porcentaje de cobertura del producto tiene un costo mayor que tener una serie de spots, es decir, con la inversión para incrementar 1 PP la cobertura se podrían tener distintos spots en cada medio. Este monto de inversión debería ser evaluado en cada medio y en la cobertura, y determinar dónde se generaría mayor volumen de ventas.

En el caso de redes sociales, nos damos cuenta que los Fans tienen un impacto negativo. Como ya se mencionó, esto se debe a que una de las estrategias que se hacen en Facebook es enviar mensajes orientados a la construcción de imagen de marca.

Inversión Digital es el medio pagado, no tendrá el mayor número de relaciones a ventas, ya sea directas o indirectas, pero todas sus relaciones impactan de manera positiva. Mientras que, la inversión de Exteriores&Revistas tiene mayor número de relaciones, sin embargo, la mayoría de ellas van a Fb fans; esto implica que no todo el impacto de la inversión termina en una venta. Por lo tanto, se recomienda generar una estrategia de comunicación enfocada en medios digitales ya que tiene un impacto mayor a ventas.

Además de enfocarse en inversión Digital por ser el medio con mayor reacción y que hay un crecimiento en la penetración del medio, se debe tomar en cuenta que el contenido de la comunicación debe ser relevante debido a que es un medio en que el usuario decide si puede ver o no la publicidad y esto podría mermar el impacto de la comunicación.

En el caso de los efectos Indirectos y Totales, es importante tener la medición de estos efectos ya que, como se mencionó al inicio del capítulo, el impacto de la publicidad es ya sobre distintos puntos de contacto y la interacción que tiene el consumidor en ellos provoca que los anuncios tengan un impacto mayor. Teniendo claridad de estos impactos, la actividad en medios de comunicación puede ser optimizada invirtiendo en aquellos puntos más rentables.



# Capítulo 5

## Conclusiones

En esta tesis se mostraron los resultados y conceptos necesarios para entender el Análisis de Factores Confirmatorio (AFC) y los Modelos de Ecuaciones Estructurales (MME), tratando de mantener formalidad estadística y a la vez de ser simples para que no sólo profesionales con formación matemática puedan consultarla.

En los capítulos 2 y 3 se mostró un ejemplo extraído de la literatura estadística para asociar y reforzar los conceptos mencionados. Estos ejemplos están asociados a la investigación del comportamiento humano y no necesariamente a datos cuyos resultados del modelo pudieran ser aplicados a los negocios, de aquí la motivación de generar una aplicación así.

### Los resultados

Finalmente, en el capítulo 4 se realizó una aplicación del Análisis de Trayectoria, que es un caso especial del Modelo de Ecuaciones Estructurales, enfocado a datos de inversión en medios, interacciones que el consumidor muestra en las redes sociales de la marca y estrategias de cobertura - precio para explicar las ventas del producto. El modelo pudo estimarse con múltiples relaciones entre variables, es decir, tal que hubo más de una variable dependiente y que a su vez pasó a ser una variable exógena.

Adicionalmente, el modelo muestra un buen ajuste, es decir, en general está dentro de lo indicado en la teoría estadística. Algunos parámetros no fueron significativos, sin embargo, se decide reportar el modelo obtenido ya que es una evidencia que la relación teórica es no significativa y no se debería plantear una estrategia que mantenga esa relación.

### La aportación

El modelo encontrado, se puede describir como vanguardista ya que en México es de los primeros en el área de investigación de mercados o medios que da solución a las múltiples relaciones con que la información interactúa.

El modelo colaboró relevantemente para realizar planes de inversión publicitaria y determinar la mejor opción de inversión, es decir, seleccionar aquellos puntos de contacto que tengan mejores relaciones tanto directa como indirectamente en las ventas y en otros puntos de contacto en que no se puede invertir de manera directa. Por ejemplo: “Me gusta”, “Compartir” o “Comentar”

las publicaciones de la marca en redes sociales.

### **Las áreas de oportunidad**

Este modelo carece de información como de CRM, actividad en YouTube, Email a usuarios registrados, indicadores de salud de marca. Por lo que podría tener mejoras y poder extraer más aprendizajes del mismo.

Los datos presentaron algunas problemáticas tales como: datos reportados como ceros en las inversiones; estos pudieron ser transformados bajo la hipótesis de que existe un nivel de recordación.

Por otro lado, algunas hipótesis no pudieron ser verificadas debido a que los Modelos de Ecuaciones Estructurales pueden presentar dificultades al usar variables categóricas, pues lo mejor es usar variables continuas para echar mano de los estadísticos de varianza-covarianza; esto como objetivo de precisión en la estimación de los parámetros a través de la matriz de varianza-covarianza muestral. En el caso de variables categóricas se pueden obtener correlaciones policóricas, ver, por ejemplo, Yuan et al. (2011).

Probablemente el tamaño muestral usado (130 observaciones) se podría considerar pequeño, sin embargo, lo recomendable es que haya entre 100 y 200 observaciones para evitar un deterioro en los valores estadísticos de bondad de ajuste debido al tamaño grande de la muestra, ver, por ejemplo, Iacobucci (2010).

En caso de tomar el tamaño muestral sugerido, se recomienda trabajar con modelos que incluyan entre 10-20 variables. Podría establecerse una tasa superior a las 10 observaciones por cada variable medible.

Finalmente, los datos utilizados presentan una asociación temporal debido a la naturaleza con que son medidos y recopilados. Esto implicaría que el modelo obtenido podría ser cuestionable. Sin embargo, existen Modelos de Ecuaciones Estructurales que consideran la estructura temporal de los datos o datos longitudinales, ver, por ejemplo, Little (2013).

### **Recursos adicionales**

Se recomienda revisar otros temas que podrían ser el siguiente paso a esta tesis o ser una opción alternativa de análisis con múltiples relaciones, por ejemplo: Análisis de Redes Neuronales o Análisis de Regresión Canónica.

# Apéndice A

## El producto Kronecker

### A.1. Definiciones y notaciones

**Definición A.1.1.** Sean  $A_{p \times q}$  y  $B_{r \times s}$  cualesquiera matrices, tal que definimos el producto de Kronecker de  $A$  y  $B$  como:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{pmatrix}.$$

De tal forma que el producto  $A \otimes B$  es una matriz bloque de  $pr$  por  $qs$ , por ejemplo en la posición  $(k, h)$  tiene asociado el bloque  $a_{kh}B$ , es decir, un real por la matriz  $B$ . Debemos tener en cuenta que, si  $A$  es un escalar, el producto se reduce a un producto ordinario de un escalar por una matriz.

Una vez definido el producto de Kronecker, podemos empezar a definir una serie de propiedades para este producto.

**Lema A.1.1.** Sean  $A_1, A_2, A_3$  y  $A_4$  matrices tales que  $A_1A_3$  y  $A_2A_4$  son productos (producto ordinario de matrices) bien definidos entonces:

1.  $(A_1 \otimes A_2)(A_3 \otimes A_4) = (A_1A_3) \otimes (A_2A_4)$ .
2.  $(A_1 \otimes A_2)^t = A_1^t \otimes A_2^t$ .
3.  $(A_1 \otimes A_2)^{-1} = A_1^{-1} \otimes A_2^{-1}$  donde  $A_1$  y  $A_2$  son matrices no singulares.
4. Si  $A$  es una matriz positiva definida, entonces tenemos que  $A \otimes A$  también es positiva definida.

Por otro lado para cualquier matriz  $A_{p \times q}$  y cualquier matriz simétrica  $B_{p \times p}$ , se pueden definir las transformaciones vec y vecs como:

1.  $\text{vec}A = (a_{11}, a_{12}, \dots, a_{1q}, a_{21}, a_{22}, \dots, a_{2q}, \dots, a_{p1}, a_{p2}, \dots, a_{pq})^T$ .

$$2. \text{vecs}B = (b_{11}, b_{21}, b_{22}, b_{31}, b_{32}, b_{33}, \dots, b_{p1}, \dots, b_{pp})^T.$$

En otras palabras,  $\text{vec}A$  es un vector columna de dimensión  $(pq \times 1)$  formado por los renglones secuenciales de  $A$ , y  $\text{vecs}B$  es un vector columna de dimensión  $(p^* \times 1)$ , donde  $p^* = \frac{p(p+1)}{2}$ , formado por los elementos de la matriz triangular inferior  $B$ . A partir de este operador se tienen las siguientes propiedades del producto de Kronecker:

**Lema A.1.2.** 1.  $(A_1 \otimes A_2)\text{vec}C = \text{vec}A_2CA_1^t$  donde  $A_1$ ,  $A_2$  y  $C$  son matrices tales que  $A_1CA_2^t$  está bien definido.

2.  $(\text{vec}C_1)^t(A_1 \otimes A_2)(\text{vec}C_2) = \text{tr} C_1^tA_1C_2A_2^t$ , donde  $A_1$ ,  $A_2$ ,  $C_1$  y  $C_2$  son matrices tales que el producto  $C_1^tA_1C_2A_2^t$  está bien definido.

Con estas definiciones podemos mostrar que  $\text{vecs}B$  y  $\text{vec}B$  se relacionan a través de la siguiente expresión (Browne 1974):

$$\text{vecs}B = K_p^t \text{vec}B, \quad (\text{A.1})$$

donde  $K_p$  es una matriz de tamaño  $(p^2 \times p^*)$  y sus elementos están definidos por:

$$K_p(ij, kh) = \frac{\delta_{ik}\delta_{jh} + \delta_{ih}\delta_{jk}}{2}, \quad i \leq p, \quad j \leq p, \quad k \leq h \leq p,$$

de tal manera que  $\delta_{ij}$  representa la delta de Kronecker, la cual se define como:

$$\delta_{ik} = \begin{cases} 1 & \text{si } i = k \\ 0 & \text{si } i \neq k \end{cases}.$$

Por lo tanto  $K_p^t$  expresa a  $\text{vec}B$  en términos de  $\text{vecs}B$  la cual sólo contiene elementos no redundantes en la parte triangular inferior partiendo de una matriz simétrica.

El rango de  $K_p$  es  $p^*$  y la “inversa izquierda” es:

$$K_p^- = (K_p^t K_p)^{-1} K_p^t$$

Esta inversa izquierda puede usarse para expresar  $\text{vec}B$  en términos de  $\text{vecs}B$  como sigue:

$$\text{vec}B = (K_p^-)^t \text{vecs}B. \quad (\text{A.2})$$

Otra matriz importante para desarrollar la estimación es  $M_p$  de tamaño  $p^2 \times p^2$  y se define como:

$$M_p = K_p K_p^- = K_p (K_p^t K_p)^{-1} K_p^t.$$

**Lema A.1.3.** Las propiedades de la matriz  $M_p$  son:

1.  $M_p$  es una matriz simétrica e idempotente de rango  $p^*$ , i.e.  $M_p = M_p M_p$ .
2.  $M_p \mathbf{K}_p = \mathbf{K}_p$ .
3.  $M_p \text{vec}B = \text{vec}B$ , para cualquier matriz simétrica  $B$ .
4.  $M_p(A \otimes A) = (A \otimes A) \mathbf{M}_q$ , para cualquier matriz  $A_{p \times q}$ <sup>1</sup>.

<sup>1</sup>Para consultar más propiedades de las matrices  $M_p$  y  $\mathbf{K}_p$  consultar Browne, 1974 y Bentler y Lee, 1975.

## Apéndice B

# Transformaciones en medios

### B.1. Efecto Lag

**Definición B.1.1.** *Es el tiempo de desfase que pasa entre el momento de la publicidad y el acto de compra.*

Es frecuente que la publicidad no afecte de manera inmediata al consumidor para que éste compre de inmediato el producto anunciado. Mencionado esto se realiza un desfase de la información. En la figura B.1.

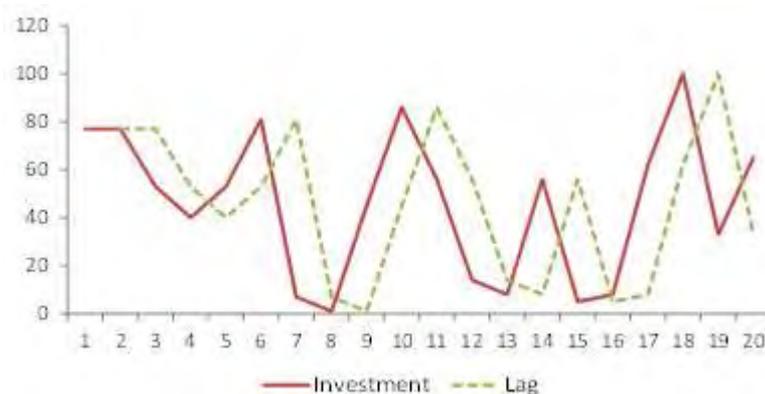


Figura B.1: Visualización del efecto *Lag*.

Este tipo de transformaciones dependen en gran medida de la categoría que se este modelando. Por ejemplo, una bebida es probable que la adquiera dentro de los siguientes días, mientras que para un auto la decisión de compra toma más tiempo.

## B.2. Efecto Adstock

**Definición B.2.1.** *Aquella transformación que habilita la inclusión de efectos no lineales en la actividad publicitaria de cualquier medio a los beneficios de mercado se le conoce como Adstock<sup>1</sup>.*

La transformación se aplica porque la actividad publicitaria construye recordación en el consumidor y por lo tanto cada nueva exposición a los mensajes incrementa la recordación o motivación en  $\frac{1}{2}$ . En caso de que no haya otra exposición de la actividad anterior disminuye la recordación o motivación de compra, en la figura B.2 se muestra el comportamiento de este efecto.

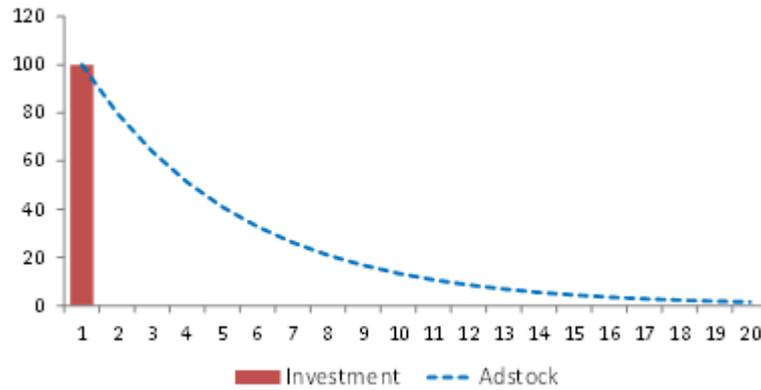


Figura B.2: Visualización del efecto *Adstock*.

Existen diversas formas de calcular el efecto *decay* del Adstock, a continuación se mencionan<sup>2</sup>.

### B.2.1. Modelos Adstock

#### Efecto Decay Simple

$$A_t = I_t + \lambda A_{t-1} \quad (\text{B.1})$$

donde

- $t$  es la semana o mes de referencia tal que  $t \in \mathbb{N}$ .
- $A_t$  es la inversión del tiempo  $t$  con el efecto Adstock.
- $I_t$  es la inversión en el tiempo  $t$ .
- $\lambda$  es el porcentaje de recordación o efecto *decay*.

La expresión recursiva del modelo queda como:

$$A_t = I_t + \lambda I_{t-1} + \lambda^2 I_{t-2} + \dots + \lambda^{(t-1)} I_1 \quad (\text{B.2})$$

<sup>1</sup>Broadbent, S. (1979).

<sup>2</sup>Fry et al. (2000).

o

$$A_t = \sum_{i=0}^t \lambda^{t-i} I_{t-i} \quad (\text{B.3})$$

### Efecto Decay Logarítmico

Este efecto es aplicando logaritmo a la actividad de ese periodo, es decir:

$$A_t = \log I_t + \lambda A_{t-1} \quad (\text{B.4})$$

Este es un modelo no lineal del Adstock además de que no permite medir la saturación de la actividad del medio.

### Efecto Decay Exponencial Negativo

El siguiente modelo aplica una distribución exponencial negativa usando dos parámetros, el efecto decay Adstock  $\lambda$  y una tasa de aprendizaje o saturación  $\beta$

$$A_t = (1 - \exp -\beta I_t) + \lambda A_{t-1} \quad (\text{B.5})$$

Este modelo es más flexible con diferentes valores de  $\beta$  puede ser empíricamente probado en una respuesta modelada correctamente medido el nivel de saturación.

### Efecto Decay Logística

Usando una distribución logística en lugar de la exponencial negativa el comportamiento del efecto Adstock será en forma de  $S$

$$A_t = \frac{1}{(1 + \exp -\beta I_t)} + \lambda A_{t-1} \quad (\text{B.6})$$



# Bibliografía

- [1] Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, **112**(4): 545-557.
- [2] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 3rd ed..
- [3] Arbuckle, J. L. (2000). *Exploratory structural equation modeling*. Fordham University. Department of Psychology colloquium series.
- [4] Bagozzi, R. P. (1980). Performance and satisfaction in an industrial sales force: an examination of their antecedents and simultaneity, *Journal of Marketing*, **44**, 65-77.
- [5] Batista Forguet, J. M., Coenders Gallart, G. (2000). *Modelos de Ecuaciones Estructurales. Modelos para el análisis de relaciones causales*. España: La Muralla.
- [6] Bentler, P. M., Lee, S. Y. (1975). Some extensions of matrix calculus. *General Systems*, **20**, 145-150.
- [7] Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- [8] Broadbent, S. (1979). One Way TV Advertisements Work. *Journal of the Market Research Society*, **23**(3).
- [9] Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structure. *South African Statistical Journal*, **8**, 1-24.
- [10] Browne, M. W., Cudeck, R. (1993). Alternative ways of assessing model fit. En: Bollen, K. A., Long, J. S. (eds.), *Testing Structural Equation Models*. Newbury Park: Sage Publications, pp 136-162.
- [11] Chen, F., Curran, P. J., Bollen, K. A., Kirby, J. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Social Methods & Research*, **36**(4): 462-494.
- [12] Cribbie, R. A. (2007). *Multiplicity Control in Structural Equation Modeling*. Structural Equation Modeling: A Multidisciplinary Journal, **14**, 98-112.
- [13] Croker, L., Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- [14] De Lara Jaro, A. (2008). *Medición y control de riesgos financieros*. 3ra ed., Limusa.

- [15] Domínguez Doncel, A., Hermo Gutiérrez, S. (2007). *Métricas del marketing*. ESIC Editorial.
- [16] Fornell, C., Larcker, D. F. (1981). Evaluating structural equations models with unobservable variables and measurement error. *Journal of Marketing Research*, **18**, 39-50.
- [17] Fox, J. (2006). Structural Equation Modeling With the sem Package in R. *Structural Equation Modeling*, **13**(3), 465-486.
- [18] Fry, T.R.L., Broadbent, S., Dixon, J.M. (2000). Estimating Advertising Half-life and the Data Interval Bias. *Journal of Targeting, Measurement & Analysis in Marketing*, **8**, 314-334.
- [19] García Veiga, M. A. (2011). *Análisis Causal con Ecuaciones Estructurales de la Satisfacción Ciudadana con los Servicios Municipales*. Proyecto Fin de Máster, Universidad de Santiago de Compostela.
- [20] Hair, J. F., Anderson, R. W., Tatham, R. L., Black, W. C. (2000). *Multivariate Data Analysis*. Prentice Hall, 5th ed..
- [21] Hancock, G. R. (1999). A sequential Scheffé-type respecification procedure for controlling Type I error in exploratory structural equation model modification. *Structural Equation Modeling: A Multidisciplinary Journal*, **6**, 158-168.
- [22] Hu, L.T., Bentler, P. M., Kano, Y. (1992). Can test statistics in covariance structure analyses be trusted? *Psychological Bulletin*, **112**, 351-362.
- [23] Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology* **20**, 90-98.
- [24] Lee, Sik-Yum (2007). *Structural Equation Modeling. A Bayesian approach*. England: John Wiley & Sons.
- [25] Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: The Guilford Press.
- [26] Loehlin, J. C. (2004). *Latent Variable Models. An Introduction to factor, path, and structural equation analysis*. New Jersey: Lawrence Erlbaum Associates Publishers, 2nd ed..
- [27] Manzano Patiño, A., Zamora Muñoz, S. (2013). *Sistema de ecuaciones estructurales: una herramienta de investigación*. Cuaderno técnico 4. Centro Nacional de Evaluación para la Educación Superior, CENEVAL, México.
- [28] McDonald, R. P., Marsh, H.W. (1990). Choosing a multivariate model: noncentrality and goodness of fit. *Psychological Bulletin*, **107**, 247-255.
- [29] Mejía Puente, M., Cornejo Sánchez, C. (2010). Aplicación del modelo de ecuaciones estructurales a la gestión del conocimiento. *Latin American and Caribbean Journal of Engineering Education*, **4**(1).
- [30] Méndez Ramírez, I. Modelos de Ecuaciones Estructurales. Notas de clase.  
<http://www.dppe.iimas.unam.mx/nacho/chooseFiles.php?elec=1>
- [31] Montgomery, D. C., Peck, E. A., Vining, G. G. (2007). *Introducción al Análisis de Regresión Lineal*. 3ra ed., Patria.

- [32] Raykov, T., Marcoulides, G. A. (2006) *A First Course in Structural Equation Modeling*. Lawrence Erlbaum Associates, 2nd ed..
- [33] Rincón, L. (2007). *Curso Intermedio de Probabilidad*. México: Las Prensas de Ciencias.
- [34] Rosseel, Y. (2011). *lavaan: an R package for structural equation modeling and more*. Department of Data Analysis, Ghent University, Psychoco 2011-Tubingen.
- [35] Schumacker, R. E., Lomax, R. G. (2010). *A beginner's guide to structural equation modeling*. New York: Routledge.
- [36] Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, **25**, 173-180.
- [37] Ullman, J. B. (2006). Structural equation modeling: reviewing the basics and moving forward. *Journal of Personality Assessment*, **87**(1), 35-50.
- [38] Uriel Jiménez, E., Aldás Manzano, J. (2005). *Análisis Multivariante Aplicado. Aplicaciones al marketing, investigación de mercados, economía, dirección de empresas y turismo*. Thomson.
- [39] Yuan, K.-H., Wu, R., Bentler, P. M. (2011). Ridge structural equation modeling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology* **64**(0 1).
- [40] Tellis, Gerard J. (2006). *Handbook of Marketing Research*. Rajiv Grover, Thousand Oaks: Sage Publication, 506-522.