ÍNDICE

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

CAPÍTULO 1
ANÁLISIS DE SENDEROS

1.1. INTRODUCCIÓN

El fundador del análisis de senderos fue Sewell Wright, biométrico, cuyos trabajos se empezaron a publicar en el año 1921.

El análisis de senderos es una técnica matemática que proviene del campo de las ciencias naturales, y su utilización se extendió al campo de las ciencias sociales y de la Sociología alrededor del año 1960, principalmente en los Estados Unidos, actualmente se emplea con mayor frecuencia y en muchos lugares alrededor del mundo.

El análisis de senderos es una técnica similar a la regresión pero con poder explicativo, que estudia los efectos directos e indirectos en un conjunto de variables observables.

El estudio del análisis de senderos involucra la estimación de relaciones causales[1] entre variables observadas. Sin embargo, el dato básico del análisis de senderos es la covarianza, que incluye correlación. Generalmente se conoce que "correlación no implica causalidad". Este principio es adecuado porque aunque una correlación sustancial indicaría una relación causal, las variables pueden también asociarse de manera que no tengan causalidad[2].

Los modelos de senderos intentan explicar por qué las variables observadas están correlacionadas. Parte de esta explicación puede suponer efectos causales. Además, podría reflejar una presunta relación de no causalidad, como una falsa

---

[1] Una unión causal es una relación de causa-efecto de una variable sobre otra, en el diagrama de senderos se la representa mediante flechas. Dos variables están unidas por una relación de causalidad cuando una variable influye en otra, de tal manera que una modificación en la primera produce o da lugar a una modificación en la segunda.

[2] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 93-94.

asociación entre variables observadas debido a causas comunes. La finalidad principal del análisis de senderos es estimar aspectos causales versus no causales de correlaciones observadas.

1.2. CONCEPTOS BÁSICOS

1.2.1. ELEMENTOS DEL ANÁLISIS DE SENDEROS

Los elementos que componen este análisis son:

- El diagrama de senderos
- El modelo de senderos
- Las ecuaciones estructurales
- Los coeficientes de Wright

El Diagrama de Senderos. Es un gráfico en donde se encuentran representadas las relaciones de causalidad que se supone existen en un conjunto de variables.

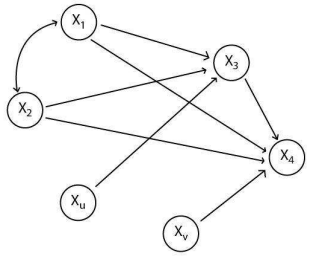

GRAFICO 1.1 Ejemplo del diagrama de senderos[3].

Como se observa en el gráfico 1.1, las variables se encuentran representadas por círculos (o en algunos casos por cuadrados) que encierran el nombre de la variable, y las flechas rectas que tienen una sola dirección, relacionan las

[3] (13) SIERRA, "Ciencias Sociales: Análisis Estadístico y Modelos Matemáticos", pág. 287.

variables, la flecha comienza en la variable independiente que influye y su punta termina en la variable dependiente o influida.

Las flechas con línea curva y de doble punta, representa la posible correlación entre las variables independientes (ó variables exógenas).

En el gráfico 1.1 se pueden distinguir tres tipos de variables: endógenas (dependientes), exógenas (independientes), y residuales. Las variables endógenas son x_3 y x_4, que figuran en el modelo influidas por otras variables. Exógenas son x_1 y x_2, variables que no dependen de ninguna otra, es decir, son en cierto modo variables externas al modelo. Residuales son las variables x_u y x_v, que influyen en ciertas variables del modelo (en este caso influyen en x_3 y x_4) y representan a los factores no observados.

El Modelo de Senderos y las Ecuaciones Estructurales. El conjunto de ecuaciones estructurales, que procede de un diagrama de senderos, se denomina modelo de senderos.

Cualquier variable que sea dependiente de otra o de otras variables, en el diagrama de senderos, se la puede expresar como una función lineal de las variables independientes.

De esta manera, considerando el diagrama de senderos del gráfico 1.1, las variables x_3 y x_4 se las puede expresar como función de x_1, x_2, x_u y x_1, x_2, x_3, x_v, respectivamente, originándose así las siguientes ecuaciones estructurales:

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, x_u$$

$$x_4 = p_{41}\, x_1 + p_{42}\, x_2 + p_{43}\, x_3 + p_{4v}\, x_v$$

De manera general, las x_i en las ecuaciones estructurales dentro del análisis de senderos, representan funciones de variables y los parámetros se denominan coeficientes de Wright.

Las condiciones que debe reunir un modelo en el análisis de senderos, son las siguientes:

- **El modelo debe ser completo.** Cada variable dependiente se la debe considerar explícitamente como completamente determinada por alguna combinación de variables en el modelo, caso contrario, cuando las variables no mantengan la determinación completa por las variables medidas, se debe introducir una variable residual no correlacionada con otras variables determinantes del modelo.

- **El modelo debe ser recursivo[4].** Esto se cumple cuando las relaciones entre las variables que forman el modelo son asimétricas. En este tipo de modelo, dos variables no pueden ser a la vez causa y efecto una de la otra, es decir, el efecto de una causa no puede ser al mismo tiempo causa de su causa.

- **El modelo debe ser lineal.** Esta condición se cumple cuando las relaciones que unen las variables que forman el modelo pueden ser representadas por ecuaciones lineales. Cuando estas ecuaciones son de otro tipo, se las debe transformar en lineales.

- **Relaciones de causa-efecto entre las variables del modelo.** Todas las variables del modelo deben estar unidas por relaciones de causalidad, esto se debe al hecho de que el análisis de senderos es una técnica que analiza las estructuras causales.

- **Las variables del modelo deben ser de tipo cuantitativo y continuas**. En el caso de que se empleen variables cualitativas se las debe poder cuantificar.

- **Variables residuales o errores.** Representan, los errores de medición o las variables que podrían influir en el modelo pero que no se encuentran incluidas en él. Estas variables o errores no se encuentran correlacionadas[5] entre sí y ejercen influencia sobre una sola variable del modelo.

---

[4] Una explicación más detallada de los modelos recursivos se encuentra en la siguiente sección (1.3).

[5] Se adopta la hipótesis de que los errores no tienen correlación entre sí.

Coeficientes de Wright. En el análisis de senderos, los parámetros de las ecuaciones estructurales reciben el nombre de coeficientes de Wright. De igual manera que en el análisis de regresión, estos coeficientes de Wright son los que constituyen las incógnitas, y su valor se determina resolviendo el sistema de ecuaciones estructurales del modelo.

Por tanto, el análisis de senderos consiste en la determinación de estos coeficientes.

Una vez que se identifican a los coeficientes de Wright, los valores obtenidos figuran en el diagrama de senderos. En el caso de las flechas orientadas hacia las variables endógenas, se coloca los valores de los coeficientes de Wright y, en las variables exógenas, se colocan los valores de los coeficientes de correlación simple de orden cero.

Estos coeficientes se denotan por p_{ij}. El primer subíndice i, representa la variable dependiente, y el segundo j, la variable independiente[6].

1.3. MODELOS RECURSIVOS Y NO RECURSIVOS

Se consideran las siguientes notaciones: X para variables exógenas observadas, Y para variables endógenas observadas, D para las variables residuales (por ejemplo, variables exógenas no observadas),  y  para, respectivamente, las varianzas y covarianzas de variables exógenas, $\rightarrow$ para los presuntos efectos causales directos, y $\leftrightarrows$ para los efectos causales recíprocos.

Existen básicamente dos tipos de modelos de senderos: recursivos y no recursivos.

Los modelos recursivos tienen dos características básicas: sus residuos (o perturbaciones) están no correlacionados, y todos los efectos causales son

---

[6] (13) SIERRA, "Ciencias Sociales: Análisis Estadístico y Modelos Matemáticos", pág. 286-288.

unidireccionales. Los modelos no recursivos tienen lazos de retroalimentación o pueden tener perturbaciones correlacionadas. El modelo de la figura (a) del gráfico 1.2 es recursivo porque sus perturbaciones son independientes y ninguna variables es causa y efecto de otra variable, directa o indirectamente. Por ejemplo, X_1, X_2 y Y_1 están especificadas como causas directas o indirectas de Y_2, pero Y_2 no tiene efecto de retorno sobre una de sus presuntas causas. Por el contrario, la figura (b) del gráfico 1.2 es no recursivo, porque tiene lazos de retroalimentación directos en el cual Y_1 y Y_2 están especificadas como causa y efecto de cada uno ($Y_1 \leftrightarrows Y_2$). Se puede apreciar también en el mismo modelo (figura (b)) que hay una correlación de perturbación. Los modelos con lazos de retroalimentación, tales como $Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_1$, son también no recursivos.

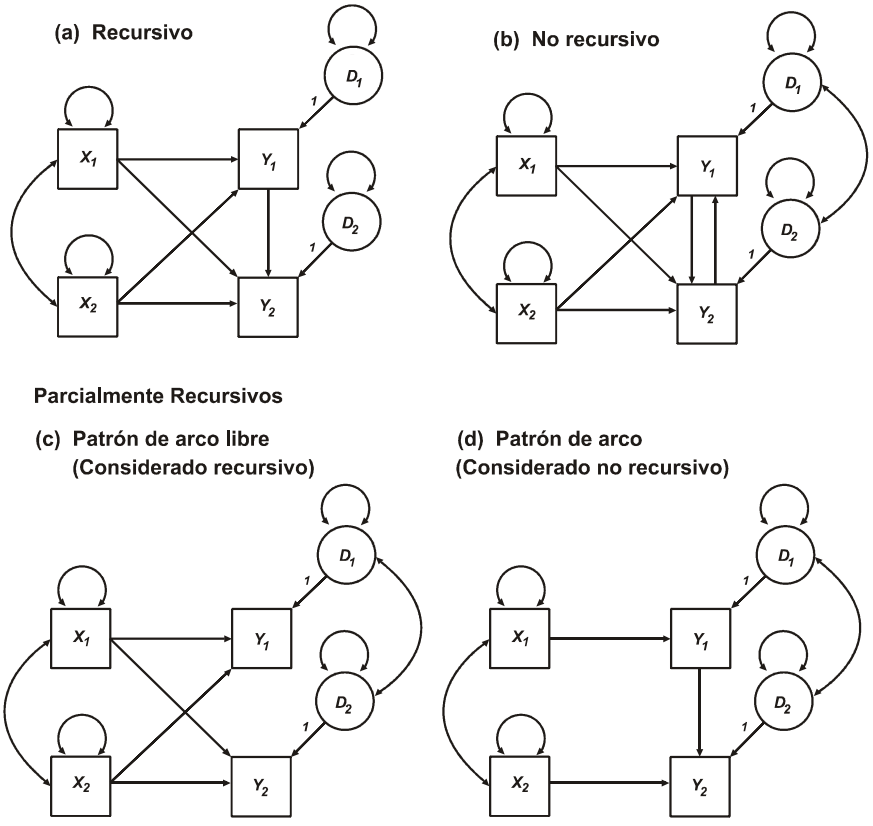

GRÁFICO 1.2 Ejemplos de modelos de senderos recursivos, no recursivos y parcialmente recursivos[7].

[7] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 103.

Otro tipo de modelo de senderos, es el que tiene efectos direccionales y perturbaciones correlacionadas, las figuras (c) y (d) del gráfico 1.2 son ejemplos de éstos.

A continuación se describe el número y tipos de parámetros para cada uno de los modelos de senderos del gráfico 1.2.

Modelo	Varianzas	Covarianzas	Efectos directos sobre variables endógenas			Total
(a) Recursivo	X_1, X_2 D_1, D_2	$X_1 \cup X_2$	$X_1 \rightarrow Y_1$ $X_1 \rightarrow Y_2$	$Y_1 \rightarrow Y_2$	$X_2 \rightarrow Y_1$ $X_2 \rightarrow Y_2$	10
(b) No recursivo	X_1, X_2 D_1, D_2	$X_1 \cup X_2$ $D_1 \cup D_2$	$X_1 \rightarrow Y_1$ $X_1 \rightarrow Y_2$	$Y_1 \rightarrow Y_2$ $Y_2 \rightarrow Y_1$	$X_2 \rightarrow Y_1$ $X_2 \rightarrow Y_2$	12
(c) Parcialmente recursivo (considerado recursivo)	X_1, X_2 D_1, D_2	$X_1 \cup X_2$ $D_1 \cup D_2$	$X_1 \rightarrow Y_1$ $X_1 \rightarrow Y_2$		$X_2 \rightarrow Y_1$ $X_2 \rightarrow Y_2$	10
(d) Parcialmente recursivo (considerado no recursivo)	X_1, X_2 D_1, D_2	$X_1 \cup X_2$ $D_1 \cup D_2$	$X_1 \rightarrow Y_1$	$Y_1 \rightarrow Y_2$	$X_2 \rightarrow Y_2$	9

TABLA 1.1 Descripción de los modelos de senderos del gráfico 1.2[8].

A los modelos no recursivos se los suele también llamar parcialmente recursivos. Los modelos parcialmente recursivos con patrón de arco libre de correlaciones de perturbación pueden ser tratados en el análisis como modelos recursivos. Un patrón de arco libre significa que las perturbaciones correlacionadas están restringidas a pares de variables endógenas sin efectos directos entre ellas, por ejemplo, el modelo de la figura (c). Por el contrario, los modelos parcialmente recursivos con patrón de arco de correlaciones de perturbación podrían tratarse en el análisis como modelos no recursivos. Un patrón de arco significa que la

[8] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 106.

perturbación correlacionada ocurre con efectos directos entre las variables endógenas.

Las suposiciones en los modelos recursivos, de que todos los efectos causales son unidireccionales y que las perturbaciones son independientes cuando hay efectos directos entre las variables endógenas, simplifica su análisis[9].

1.4. CONSTRUCCIÓN DEL DIAGRAMA DE SENDEROS

En la mayoría de ocasiones, la decisión sobre qué incluir en el modelo de senderos se toma más por la experiencia del investigador que por publicaciones (conceptos teóricos). Algunas veces hay mucha información por analizar, es decir, hay quizá muchas variables causales potenciales en el estudio, que se vuelve imposible incluir todas. Para tratar este problema, el investigador debe confiar en su juicio acerca de qué variables son las más cruciales para el estudio.

El error de especificación al omitir variables causales en un modelo de senderos, tiene la misma consecuencia que omitir predictores en una ecuación de regresión.

El primer paso en la construcción de un diagrama de senderos es diferenciar las variables que no están influenciadas por otras variables en el modelo (variables exógenas) y las variables que están afectadas por otras (variables endógenas). A cada una de las variables endógenas se le asocia una variable residual.

Para construir un diagrama de senderos, se siguen los siguientes pasos (un sendero se representa con una flecha dirigida):

- Para cada variable dependiente (endógena), se dibuja una flecha recta desde cada una de sus fuentes.
- De igual manera, para cada variable dependiente, se dibuja una flecha recta desde su residuo.

---

[9] (10) KLINE, “Principles and Practice of Structural Equation Modeling”, pág. 102-104.

- Entre cada par de variables independientes (exógenas) que tengan correlación distinta de cero, se dibuja una flecha curva con doble punta.

La flecha curva para la correlación indica la simetría de un coeficiente de correlación. Las otras conexiones o uniones que se muestran con la flecha de una punta son direccionales.

Para una mejor visualización, se considera el gráfico 1.3 como ejemplo de un diagrama de senderos para variables de causa-efecto.

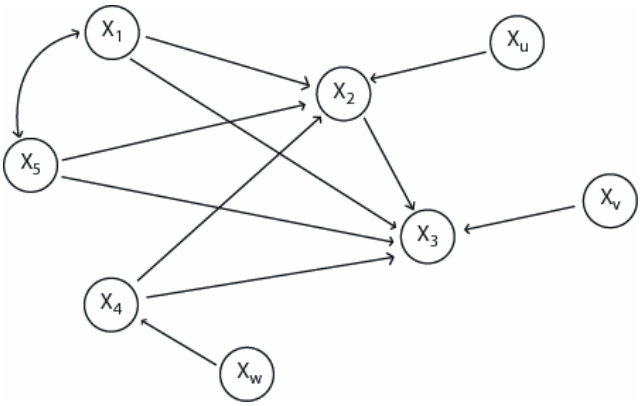

GRÁFICO 1.3 Diagrama de senderos.

Como se puede ver, en el diagrama de senderos del gráfico 1.3, las variables endógenas son x_2, x_3 y x_4, las variables exógenas son x_1 y x_5 y las variables residuales son x_u, x_v y x_w.

1.5. IDENTIFICACIÓN DE LOS COEFICIENTES DE WRIGHT

Se realiza la determinación numérica o identificación de los parámetros o coeficientes de las ecuaciones una vez establecido el sistema de ecuaciones del modelo recursivo, esta determinación se realiza en función de los coeficientes de correlación, denotados por r_{ij}.

Como primer paso, se debe realizar una transformación matemática de las ecuaciones originales, de tal manera que las variables de dichas ecuaciones estén expresadas en términos de los coeficientes de correlación correspondientes. Esta transformación, cuando se trata de variables estandarizadas, se basa en la igualdad entre la covarianza y su coeficiente de correlación.

Se multiplican todas las ecuaciones del sistema por x_g, para todos los g<i, siendo i el subíndice de la variable dependiente de la ecuación.

Para explicar de mejor manera, se considera la siguiente ecuación, que proviene del diagrama de senderos del gráfico 1.1:

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, x_u$$

Se la multiplica por x_1 y se tiene:

$$x_3\, x_1 = p_{31}\, x_1^2 + p_{32}\, x_2\, x_1 + p_{3u}\, x_u\, x_1$$

A continuación, se toman las esperanzas matemáticas de las variables consideradas en la ecuación anterior y se obtiene la siguiente ecuación:

$$E(x_3\, x_1) = p_{31}\, E(x_1^2) + p_{32}\, E(x_2\, x_1) + p_{3u}\, E(x_u\, x_1),$$

en donde $p_{32}\, E(x_2\, x_1)$ es igual a $p_{32}\, Cov(x_2\, x_1)$ y $E(x_1^2) = 1$.

Así mismo, la $E(x_u\, x_1)$ es cero, debido a que la correlación de los errores con las variables es nula.

De la ecuación de esperanzas matemáticas que se obtuvo, se puede pasar a esta otra ecuación:

$$Cov(x_3\, x_1) = p_{31}\, Cov(x_1\, x_1) + p_{32}\, Cov(x_2\, x_1) + 0$$

Si se considera que:

$$Cov\,(x_2\, x_1) = r_{21}; \qquad Cov(x_3\, x_1) = r_{31}; \qquad Cov(x_1\, x_1) = 1$$

Entonces, se obtiene la ecuación siguiente:

$$r_{31} = p_{31} + p_{32}\, r_{21}$$

Los coeficientes de Wright estandarizados son los p_{ij} y los coeficientes de correlación son los términos restantes.

Continuando el proceso, ahora se multiplica por x_2. a la misma ecuación que se trabajó anteriormente:

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, x_u$$

$$x_3\, x_2 = p_{31}\, x_1\, x_2 + p_{32}\, x_2^2 + p_{3u}\, x_u\, x_2$$

Y se realiza el mismo proceso anterior:

$$E(x_3\, x_2) = p_{31}\, E(x_1\, x_2) + p_{32}\, E(x_2^2) + p_{3u}\, E(x_u\, x_2),$$

en donde $p_{31}\, E(x_1\, x_2)$ es igual a $p_{31}\, Cov(x_1\, x_2)$, $E(x_2^2) = 1$ y $E(x_u\, x_2)$ es cero.

De este modo:

$$Cov(x_3\, x_2) = p_{31}\, Cov(x_1\, x_2) + p_{32}\, Cov(x_2\, x_2) + 0$$

Considerando que:

$$Cov\,(x_1\, x_2) = r_{12}; \qquad Cov(x_3\, x_2) = r_{32}; \qquad Cov(x_2\, x_2) = 1$$

Ahora, el resultado es:

$$r_{32} = p_{31}\, r_{12} + p_{32}$$

Así, se obtiene un conjunto de dos ecuaciones con dos incógnitas:

$$r_{31} = p_{31} + p_{32}\, r_{21}$$

$$r_{32} = p_{31}\, r_{12} + p_{32}$$

en donde p_{31} y p_{32} son los coeficientes de Wright.

La ecuación utilizada en este proceso ($x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, x_u$) es identificable[10], pero el coeficiente residual p_{3u} no.

Para determinar si p_{3u} es identificable, se multiplica por x_u a la ecuación ya conocida:

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, x_u$$

Realizando el mismo proceso anterior, se obtiene que:

$$r_{3u} = p_{3u}.$$

Multiplicando por x_3 a la misma ecuación, se obtiene:

$$1 = p_{31}\, r_{13} + p_{32}\, r_{32} + p_{3u}\, r_{3u}$$

Como $r_{3u} = p_{3u}$, la ecuación anterior es igual a:

$$1 = p_{31}\, r_{13} + p_{32}\, r_{23} + p_{3u}^2$$

Ahora, se despeja el valor de p_{3u}^2:

$$p_{3u}^2 = 1 - p_{31}\, r_{31} - p_{32}\, r_{32},$$

$$p_{3u} = \sqrt{1 - p_{31} r_{31} - p_{32} r_{32}}$$

De esta manera queda identificada la ecuación.

Continuando con la segunda ecuación, se procederá como se ha hecho con la primera ecuación, con la diferencia de que en esta ocasión se debe multiplicar a la ecuación tres veces, por x_1, x_2 y x_3.

$$x_4 = p_{41}\, x_1 + p_{42}\, x_2 + p_{43}\, x_3 + p_{4v}\, x_v$$

---

[10] Una ecuación es identificable si existe información suficiente para ser resuelta y determinar sus parámetros.

Realizando dichas multiplicaciones, se obtendrán tres ecuaciones con tres incógnitas, así, los tres coeficientes p_{ij} se los puede identificar y resolver.

Con la resolución e identificación del coeficiente p_{4v}, se procede de igual manera que con p_{3u} en el caso anterior.

De manera general, se debe multiplicar los dos lados de la ecuación a ser identificada y resuelta por todos los x_i inferiores al x_i dado. Por lo tanto habrá $i-1$ multiplicaciones, con lo cual se obtendrá un sistema de $i-1$ ecuaciones con $i-1$ incógnitas.

La fórmula general para los coeficientes residuales es:

$$p_{ie} = \sqrt{1 - \Sigma p_{iq} r_{iq}} \ .$$

A continuación se describen las fórmulas que dan directamente los valores de p, para la segunda y tercera ecuación:

Para la segunda ecuación:
$$p_{31} = \frac{\left(r_{13} - r_{12} r_{23}\right)}{\left(1 - r_{12}^2\right)}$$

$$p_{32} = \frac{\left(r_{23} - r_{12} r_{13}\right)}{\left(1 - r_{12}^2\right)}$$

Para la tercera ecuación:

Siendo:
$$p = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}$$

$$p_{41} = \frac{\begin{vmatrix} r_{14} & r_{12} & r_{13} \\ r_{24} & 1 & r_{23} \\ r_{34} & r_{23} & 1 \end{vmatrix}}{p}$$

$$p_{42} = \frac{\begin{vmatrix} 1 & r_{14} & r_{13} \\ r_{12} & r_{24} & r_{23} \\ r_{13} & r_{34} & 1 \end{vmatrix}}{p}$$

$$p_{43} = \frac{\begin{vmatrix} 1 & r_{12} & r_{14} \\ r_{12} & 1 & r_{24} \\ r_{13} & r_{23} & r_{34} \end{vmatrix}}{p}$$ [11]

El número de casos no tiene gran importancia en la identificación de los modelos de senderos. El rol que cumple el tamaño muestral aquí, es básicamente el mismo que para otros tipos de métodos estadísticos: los resultados derivados de muestras más grandes tienen menor error que de muestras más pequeñas.

Por lo tanto, algunas pautas sobre la determinación del tamaño muestral en los métodos de estimación, para establecer qué tan grande debe ser la muestra en orden de los resultados son: pequeña, N<100; mediana, N entre 100 y 200; grande, N>200. De este modo, un tamaño muestral de 200 o más, podría ser necesario para modelos de senderos muy complicados o complejos.

No existen estándares escritos sobre la relación entre el tamaño muestral y la complejidad de un modelo de senderos. Un modelo de senderos con 20 parámetros debería tener mínimo un tamaño muestral de 200 casos.

1.6. EFECTO TOTAL: DIRECTO E INDIRECTO

Se puede destacar, en el análisis de senderos, el efecto total directo y el total indirecto.

---

[11] (13) SIERRA, "Ciencias Sociales: Análisis Estadístico y Modelos Matemáticos", pág. 288-293.

La correlación de orden cero entre dos variables, se llama efecto total, e indica el efecto conjunto de una variable sobre otra variable dependiente de ella a través de todos los posibles caminos, directos o indirectos.

El coeficiente de Wright p_{ij} es el efecto total directo, y la diferencia entre el coeficiente de correlación cero r_{ij} y el coeficiente de Wright p_{ij} es el efecto total indirecto.

Por ejemplo, el efecto total indirecto ETI, en el caso de las variables x_1 y x_2 viene dado por:

$$ETI = r_{12} - p_{12}$$

A continuación se explica el procedimiento mediante el cual se puede expresar el efecto total indirecto en término de los coeficientes de Wright, y su determinación.

a) Se inicia el estudio con la matriz de correlaciones del conjunto de variables a ser tratadas.
b) Según el conocimiento que se tenga de las relaciones de influencia entre estas variables, se formulan las hipótesis.

c) Se grafica el correspondiente diagrama de senderos.

d) Se aplica la fórmula básica del análisis de senderos para formar las ecuaciones de los coeficientes de correlación en función de éstos y de los coeficientes p_{ij}.

e) Se obtienen los valores p_{ij} al resolver estas ecuaciones.

f) Se encuentran los coeficientes de los errores aplicando la correspondiente fórmula.

g) Se calculan los efectos indirectos de cada variable.

h) Si se desea conocer todas las influencias o efectos directos e indirectos de una variable sobre otra, se desarrollará la ecuación del análisis de senderos que una ambas variables en función solamente de los

coeficientes de Wright, es decir, se sustituyen los términos r_{ij} que entren en la ecuación por los p_{ij} correspondientes.

1.7. INTERPRETACIÓN

La obtención de los coeficientes de Wright es quizá el objetivo primordial del análisis de senderos, en donde se involucra tanto a las variables explícitas[12] del modelo como a las implícitas[13] (errores), y se refiere también a la determinación de los efectos directos e indirectos de cada variable independiente de una ecuación sobre las variables dependientes de ella.

Además, los resultados del análisis de senderos se basan en los coeficientes estandarizados que se obtienen mediante regresión múltiple, y que muestran, en las variables explícitas, la influencia de la variable a la que se refieren en la variación de la variable dependiente, cuando se mantienen constantes las variables restantes del modelo.

El valor de los coeficientes de Wright de las variables residuales, al cuadrado, significa la variación en la variable dependiente no explicada por las restantes variables de la ecuación, variación que generalmente se atribuye a las variables implícitas o errores. Entonces, la diferencia entre 1 y el cuadrado de este coeficiente de Wright, indica la varianza en la variable dependiente de la ecuación, expresada en su conjunto por las variables explícitas de ésta ecuación.

Para que un modelo de senderos sea un buen modelo, se debe saber si este modelo tiene o no solución.

El grado en que el modelo se ajusta a los datos usados en el modelo, tiene relación con la varianza total del modelo.

---

[12] Son las variables endógenas y exógenas del modelo.

[13] Variables residuales.

En ciertas ocasiones se puede obtener un sistema de ecuaciones sobre-identificado[14] debido a que existen muchas variables en el modelo. Por lo tanto, si con las ecuaciones redundantes se obtienen resultados muy divergentes, se tiene entonces que el modelo no se ajusta bien a los datos.

Considerado este significado de los coeficientes de Wright, su interpretación se puede fundamentar en:

a) el resultado de los efectos directos e indirectos de cada variable sobre las variables dependientes;

b) el análisis de la influencia comparada de las distintas variables del modelo;

c) la comparación de estos coeficientes, con los coeficientes de correlación total o de orden cero; y,

d) la consideración, en cada caso, de las varianzas explicadas.

[14] Existe un exceso de información, por tanto, no es posible hallar una solución que satisfaga a la vez a todas las ecuaciones del sistema. En este caso, pueden elegirse subsistemas que den soluciones al sistema, pero éstas serán diversas y tantas como subsistemas se contruyan.

CAPÍTULO 2
ANÁLISIS FACTORIAL

2.1. INTRODUCCIÓN

El análisis factorial es una técnica utilizada para explicar un conjunto de variables observables mediante un número reducido de variables no observables llamadas factores.

Básicamente, el modelo factorial sigue el siguiente fundamento: "se consideran variables que puedan agruparse por sus correlaciones; es decir, se supone que las variables están dentro de un grupo particular, correlacionadas entre sí, pero tienen pequeñas correlaciones con variables en un grupo diferente".

Entonces es concebible que cada grupo de variables represente una estructura simple, o factor, que es responsable por las correlaciones observadas. Por ejemplo, las correlaciones de un grupo de resultados en pruebas de francés, inglés, matemáticas y música, sugiere un factor "inteligencia". Un segundo grupo de variables, representado por resultados de pruebas físicas, podría corresponder a otro factor. Este es el tipo de estructura que el análisis factorial busca comprobar[15].

2.2. MODELO FACTORIAL ORTOGONAL

Se considera inicialmente un vector aleatorio observable X, con *p* componentes, media μ y matriz de covarianza $\sum$. En el modelo del análisis factorial, X es linealmente dependiente de las variables aleatorias no observadas F_1, F_2, ..., F_m, llamadas *factores comunes*, y de los p valores de variación ε_1, ε_2, ..., ε_p, llamados *errores* o *factores específicos*. Entonces, el modelo del análisis factorial es:

---

[15] (7) JOHNSON, "Applied Multivariate Statistical Analysis", pág. 514.

$$X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1$$

$$X_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2$$

$$\vdots$$

$$X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p \qquad (2.1)$$

en notación matricial:

$$\underset{(px1)}{X - \mu} = \underset{(pxm)}{L}\ \underset{(mx1)}{F} + \underset{(px1)}{\varepsilon} \qquad (2.2)$$

Se llama *carga* de la i-ésima variable del factor j-ésimo al coeficiente ℓ_{ij}, por lo tanto, la *matriz de cargas factoriales* es la matriz L. Es importante mencionar que el i-ésimo factor específico ε_i está asociado únicamente a la i-ésima respuesta Xi. Las p desviaciones $X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p$ se encuentran expresadas en términos de $p + m$ variables aleatorias no observables $F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$.

En la práctica, el vector X tiene varias interpretaciones dependiendo del objetivo a estudiarse, por ejemplo, en Psicología, X puede expresar los p resultados de una prueba para medir el grado de inteligencia de una persona, así mismo, en Marketing, X puede consistir de p respuestas de una encuesta para medir los niveles de satisfacción de clientes. En estos casos, estas p medidas pueden explicarse con factores comunes que representarían el nivel de atracción de un producto o la imagen de una marca.

Verificar directamente el modelo factorial es incierto, cuando se consideran variables no observables, y se parte de las observaciones $X_1, X_2, \ldots, X_p$, Pero, cuando se trabaja con algunas suposiciones adicionales a los vectores aleatorios F y ε, se puede obtener una relación de covarianza en el modelo (2.2), la misma que puede ser comprobada.

Suponiendo que:

$$\underset{(mx1)}{E(F)} = 0 \qquad Cov(F) = E[FF'] = \underset{(mxm)}{I},$$

$$E(\varepsilon) = \underset{(px1)}{0}, \qquad Cov(\varepsilon) = E[\varepsilon\varepsilon'] = \underset{(pxp)}{\psi} = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & \psi_p \end{bmatrix} \qquad (2.3)$$

y que F y ε son independientes, se tiene que $Cov(\varepsilon, F) = E(\varepsilon F') = 0$.

Con estas suposiciones y la ecuación descrita en (2.2) se obtiene el *modelo factorial ortogonal*.

El modelo factorial ortogonal tiene una estructura de covarianza, como se ve a continuación:

$$(X - \mu)(X - \mu)' = (LF + \varepsilon)(LF + \varepsilon)'$$

$$= (LF + \varepsilon)((LF)' + \varepsilon')$$

$$= LF(LF)' + \varepsilon(LF)' + LF\varepsilon' + \varepsilon\varepsilon'$$

Por tanto, debido a la suposición señalada en (2.3) se tiene:

$$\Sigma = Cov(X) = E(X - \mu)(X - \mu)'$$

$$= L\,E(FF')L' + E(\varepsilon F')L' + L\,E(F\varepsilon') + E(\varepsilon\varepsilon')$$

$$= LL' + \psi \qquad (2.4)$$

ó

$$Var(X_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i$$

$$Cov(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}$$

Además, considerando (2.2) y (2.3), se tiene que: $(X - \mu)F' = (LF + \varepsilon)F' = LFF' + \varepsilon F'$, por lo tanto,

$$Cov(X,F) = E(X - \mu)F' = L\,E(FF') + E(\varepsilon F') = L, \qquad (2.5)$$

ó $Cov(X_i, F_j) = \ell_{ij}$

De esta manera se tiene que el modelo $X - \mu = LF + \varepsilon$ es *lineal* en los factores comunes.

Si las p "respuestas" de X son relativas a los factores, pero las relaciones no son lineales como en $X_1 - \mu_1 = \ell_{11}F_1F_3 + \varepsilon_1$, $X_2 - \mu_2 = \ell_{21}F_2F_3 + \varepsilon_2$, …, entonces, la estructura de covarianza $LL' + \psi$ descrita en (2.4) no se cumple.

Además, de la i-ésima variable dada por los m factores comunes, se obtiene una parte de la varianza que se llama la i-ésima *comunalidad*, que está dada por la suma del cuadrado de las cargas de la i-ésima variable de los *m* factores comunes y se denota como h_i^2. Esta parte de la $Var\ (X_i) = \sigma_{ii}$, dada por los factores específicos, se la denomina más frecuentemente como *unicidad*, o *varianza específica*. Así, de la ecuación (2.4) se tiene:

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + ... + \ell_{im}^2 \qquad (2.6)$$

y

$$\sigma_{ii} = h_i^2 + \psi_i, \qquad i = 1,2,..., p$$ [16]

2.3. MÉTODOS DE ESTIMACIÓN

Con el análisis factorial, dadas las observaciones $x_1, x_2, ..., x_n$ de p variables correlacionadas, se intenta conocer si el modelo factorial descrito en la ecuación 2.2, con un pequeño número de factores, representa adecuadamente a los datos.

La matriz de covarianza muestral S es un estimador de la matriz de covarianza de la población Σ que es desconocida, pero, si sus elementos de fuera de la diagonal son pequeños (de la matriz S) o si los elementos de la matriz de correlación

[16] (7) JOHNSON, "Applied Multivariate Statistical Analysis", pág. 515-518.

muestral R son cero, entonces las variables no están relacionadas y el análisis factorial no funcionaría. Cuando esto sucede, los factores específicos son primordiales en el análisis, puesto que determinar la importancia de los factores comunes es un objetivo muy importante en el análisis factorial.

A continuación se explica el método de componentes principales y el método de máxima verosimilitud que son dos de los métodos más usados de estimación de parámetros.

2.3.1. MÉTODO DE COMPONENTES PRINCIPALES

La idea principal del método de componentes principales es encontrar una aproximación $\hat{\Psi}$, de la matriz de varianzas específicas Ψ.

Inicialmente se tiene que $\sum$ la matriz de covarianza, que tiene pares de valores propios-vectores propios (λ_i, e_i) con $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, se descompone:

$$\sum = \lambda_1 e_1 e'_1 + \lambda_1 e_1 e'_1 + \ldots + \lambda_p e_p e'_p$$

$$= \left[\sqrt{\lambda_1}e_1 \vdots \sqrt{\lambda_2}e_2 \vdots \cdots \vdots \sqrt{\lambda_p}e_p\right] \begin{bmatrix} \sqrt{\lambda_1}e'_1 \\ \cdots \\ \sqrt{\lambda_2}e'_2 \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_p}e'_p \end{bmatrix} \qquad (2.7)$$

Al tener tantos factores como variables ($m = p$) y varianzas específicas $\psi_i = 0$ para todo i, se logra mantener la estructura de covarianza descrita para el modelo de análisis factorial. La j-ésima columna de la matriz de carga, está dada por $\sqrt{\lambda_j}e_j$, de este modo, se puede escribir:

$$\underset{(pxp)}{\sum} = \underset{(pxp)}{L}\ \underset{(pxp)}{L'} + 0 = \underset{(pxp)}{LL'} \qquad (2.8)$$

Cabe resaltar que además del factor escalar $\sqrt{\lambda_j}$, las cargas factoriales en el j-ésimo factor, son los coeficientes para el j-ésimo componente principal de la población.

El modelo que se expresa en la ecuación (2.8) representa el análisis factorial de Σ, que es exacto, pero no es muy útil en la práctica debido a que emplea tantas variables como factores comunes y además no permite ninguna variación en los factores específicos ε, de (2.2). Son preferibles los modelos que expliquen la estructura de covarianza expresada con pocos factores comunes. Es decir, cuando los últimos valores propios $(p - m)$, son pequeños, se omite la contribución de $\lambda_{m+1}e_{m+1}e'_{m+1}+\cdots+\lambda_p e_p e'_p$ para Σ en (2.7). Prescindiendo de esta contribución, se tiene:

$$\Sigma = \left[\sqrt{\lambda_1}e_1 \vdots \sqrt{\lambda_2}e_2 \vdots \cdots \vdots \sqrt{\lambda_m}e_m\right] \begin{bmatrix} \sqrt{\lambda_1}e'_1 \\ \cdots \\ \sqrt{\lambda_2}e'_2 \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_m}e'_m \end{bmatrix} = \underset{(pxm)}{L} \quad \underset{(mxp)}{L'} \qquad (2.9)$$

La representación de (2.9), asume que los factores específicos ε en (2.2) son de menor importancia y pueden ser ignorados al factorizar Σ. Si no se ignoran estos factores específicos, sus varianzas pueden ser los elementos de la diagonal de Σ - LL'.

En el caso en el cual no se ignoran los factores específicos en el modelo, se tiene entonces que la aproximación es:

$$\Sigma = LL'+\Psi = \left[\sqrt{\lambda_1}e_1 \vdots \sqrt{\lambda_2}e_2 \vdots \cdots \vdots \sqrt{\lambda_m}e_m\right] \begin{bmatrix} \sqrt{\lambda_1}e'_1 \\ \cdots \\ \sqrt{\lambda_2}e'_2 \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_m}e'_m \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \qquad (2.10)$$

donde $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$, para *i = 1, 2, ..., p*.

Para aplicar al conjunto de datos $x_1, x_2, ..., x_n$, primero se centran las observaciones sustrayéndolas la media muestral $\bar{x}$.

$$x_j - \bar{x} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} x_{j1} - \bar{x}_1 \\ x_{j2} - \bar{x}_2 \\ \vdots \\ x_{jp} - \bar{x}_p \end{bmatrix} \qquad j = 1, 2, ..., n \qquad (2.11)$$

Las observaciones centradas de (2.11) tienen la misma matriz de covarianza muestral S que las observaciones originales.

Cuando las unidades de medida de las variables son diferentes, es preferible trabajar con variables estandarizadas:

$$z_j = \begin{bmatrix} \dfrac{(x_{j1} - \bar{x}_1)}{\sqrt{s_{11}}} \\ \dfrac{(x_{j2} - \bar{x}_2)}{\sqrt{s_{22}}} \\ \vdots \\ \dfrac{(x_{jp} - \bar{x}_p)}{\sqrt{s_{pp}}} \end{bmatrix}, \qquad j = 1, 2, ..., n$$

cuya matriz de covarianza muestral será la matriz de correlación muestral R de las observaciones $x_1, x_2, ..., x_n$. Esta estandarización evita que si hay una variable con varianza grande, ésta influya en la obtención de las cargas factoriales.

La estimación de parámetros mediante *el método de componentes principales* es la descrita en la ecuación (2.10), cuando la aplicamos a la matriz de covarianza muestral S o a la matriz de correlación muestral R.

Si se define la matriz de covarianza muestral S en términos de sus pares de valores propios-vectores propios $\left(\hat{\lambda}_1, \hat{e}_1\right), \left(\hat{\lambda}_2, \hat{e}_2\right), \cdots, \left(\hat{\lambda}_p, \hat{e}_p\right)$, en donde

$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Siendo m el número de factores comunes, $m < p$, entonces la matriz cargas factoriales estimados $\left\{\hat{\ell}_{ij}\right\}$ es:

$$\tilde{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{e}_2 \vdots \cdots \vdots \sqrt{\hat{\lambda}_m}\hat{e}_m\right] \qquad (2.12)$$

Las varianzas específicas estimadas resultan de los elementos de la diagonal de la matriz $S - \tilde{L}\tilde{L}'$:

$$\tilde{\Psi} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix} \text{ con } \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2 \qquad (2.13)$$

Las comunalidades estimadas se expresan como:

$$\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \tilde{\ell}_{i2}^2 + \cdots + \tilde{\ell}_{im}^2 \qquad (2.14)$$

En la solución que se obtiene mediante el método de componentes principales, las cargas estimadas para un factor dado no cambian mientras el número de factores se incrementan. Por ejemplo, si m = 1, $\tilde{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1\right]$ y si m = 2, $\tilde{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{e}_2\right]$, en donde los pares de valores propios-vectores propios para S (o R) son $\left(\hat{\lambda}_1, \hat{e}_1\right)$ y $\left(\hat{\lambda}_2, \hat{e}_2\right)$.

Como se indica anteriormente en la definición de $\tilde{\psi}_i$, los elementos de la diagonal de S son iguales a los elementos de la diagonal de $\tilde{L}\tilde{L}'+\tilde{\Psi}$. No obstante, los elementos de fuera de la diagonal de S no son los mismos de $\tilde{L}\tilde{L}'+\tilde{\Psi}$. Entonces, si el número m de factores comunes no se determina a priori, la elección de m puede basarse sobre los valores propios estimados, de igual manera que en análisis de componentes principales, como se explica a continuación.

La *matriz residual:*

$$S - \left(\tilde{L}\tilde{L}'+\tilde{\Psi}\right) \qquad (2.15)$$

resultante de la aproximación de S, en este caso, mediante el método de componentes principales. Todos sus elementos de la diagonal son cero, y si los otros elementos son también pequeños, sería posible tomar m factores del modelo como fuera más conveniente. De esta manera, se tendría que:

$$\text{La suma del cuadrado de } \left(S-\left(\tilde{L}\tilde{L}'+\Psi\right)\right) \leq \lambda_{m-1}^2+\cdots\lambda_p^2 \quad (2.16)$$

Por lo tanto, un valor pequeño para la suma de los cuadrados de los valores propios omitidos, implica un valor pequeño para la suma de los cuadrados de los errores de aproximación.

Es preferible que las contribuciones de los primeros factores a las varianzas muestrales de las variables sean grandes. La contribución a la varianza muestral s_{ii} del primer factor común es $\tilde{\ell}_{i1}^2$. La contribución a la varianza muestral total s_{11} + s_{22} +...+ s_{pp} = $tr\,(S)$, del primer factor común es por lo tanto:

$$\tilde{\ell}_{11}^2+\tilde{\ell}_{21}^2+\cdots+\tilde{\ell}_{p1}^2=\left(\sqrt{\hat{\lambda}_1}\hat{e}_1\right)'\left(\sqrt{\hat{\lambda}_1}\hat{e}_1\right)=\hat{\lambda}_1$$

en donde el vector propio $\hat{e}_1$ tiene longitud uno.

De manera general, un método heurístico para establecer el número apropiado de factores comunes es considerar la proporción de la varianza muestral total debido al j-ésimo factor, que es:

i) $\hat{\lambda}_j \Big/ \sum_{j=1}^{p} s_{jj}$, para un análisis factorial de S. (2.17)

ii) $\hat{\lambda}_j \Big/ p$, para un análisis factorial de R.

2.3.2. MÉTODO DE MÁXIMA VEROSIMILITUD

Se pueden obtener las estimaciones de máxima verosimilitud de las cargas factoriales y las varianzas específicas, si se asume que los factores comunes F y los factores específicos ε, tienen una distribución normal. Si F_j y ε_j tienen

distribución normal conjunta, entonces las observaciones $Xj - \mu = LF_j + \varepsilon_j$ tienen también una distribución normal, y la función de verosimilitud en este caso es:

$$L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left\{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]\right\} \quad (2.18)$$

$$=(2\pi)^{-\frac{(n-1)p}{2}}|\Sigma|^{-\frac{(n-1)}{2}}\exp\left\{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'\right)\right]\right\}(2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\left(\frac{n}{2}\right)(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)\right]$$

la cual depende de L y Ψ mediante $\Sigma = LL'+\Psi$. Como se puede ver, este modelo no está todavía bien definido, debido a que L, mediante transformaciones ortogonales, tiene diversidad de opciones. Por este motivo, es preferible definir L, con una *condición de unicidad:*

$$L' \Psi^{-1} L = \Delta \quad \text{una matriz diagonal} \qquad (2.19)$$

La maximización numérica de (2.18) proporciona las estimaciones de máxima verosimilitud de $\hat{L}$ y $\hat{\Psi}$. Actualmente, se puede contar con varios programas de computación que fácilmente realizan esta maximización y la consecuente estimación de parámetros.

A continuación se describe un hecho importante, sobre los estimadores de máxima verosimilitud.

Sea $X_1, X_2,\ldots,X_n$ una muestra aleatoria con una distribución normal $N_p(\mu, \Sigma)$, en donde la matriz de covarianza para los m factores comunes del modelo descrito en (2.2) es $\Sigma = LL' + \Psi$. Los estimadores de máxima verosimilitud $\hat{L}$, $\hat{\Psi}$ y $\hat{\mu} = \bar{x}$ maximizan (2.18) condicionados a la matriz diagonal $\hat{L}'\hat{\Psi}^{-1}\hat{L}$.

Las comunalidades estimadas, mediante el método de máxima verosimilitud son:

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \cdots + \hat{\ell}_{im}^2 \qquad \text{para } i = 1,2,\ldots,p \qquad (2.20)$$

Por lo tanto, la proporción de la varianza muestral total debido al j-ésimo factor es:

$$\frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}}$$ [17]

2.4. ROTACIÓN FACTORIAL

Cuando las cargas iniciales, no pueden ser interpretadas, es preferible rotarlas hasta conseguir una estructura más simple.

Para lograr esta simplificación, se realiza una transformación ortogonal[18] a las cargas iniciales, obteniéndose las cargas factoriales, las cuales tienen la misma habilidad para reproducir la matriz de covarianza (o correlación). Esta transformación ortogonal de las cargas factoriales, así como también la transformación ortogonal implícita de los factores, se llama *rotación factorial*.

Se considera inicialmente a $\hat{L}$, la matriz $p \; x \; m$ de cargas factoriales estimadas, obtenida mediante cualquier método de estimación, entonces:

$$\hat{L}^* = \hat{L}T, \qquad \text{donde } TT' = T'T = I \tag{2.21}$$

será la matriz $p \; x \; m$ de cargas rotadas. Como se puede ver a continuación, tanto la matriz de covarianza estimada:

$$\hat{L}\hat{L}' + \hat{\Psi} = \hat{L}TT'\hat{L} + \hat{\Psi} = \hat{L}^*\hat{L}^{*\prime} + \hat{\Psi} \tag{2.22}$$

---

[17] (7) JOHNSON, "Applied Multivariate Statistical Analysis", pág. 521-531.

[18] Una transformación ortogonal, según el álgebra de matrices, es una rotación rígida del eje de coordenadas.

así como también, la matriz residual, $S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L}^*\hat{L}^{*\prime} - \hat{\Psi}$ no cambian. De igual manera, las varianzas específicas $\hat{\psi}_i$, y por lo tanto, las comunalidades $\hat{h}_i^2$, se mantienen inalterables. Por estas razones, no importaría si se obtiene $\hat{L}$ o $\hat{L}^*$.

Se usan métodos gráficos y analíticos para realizar una rotación ortogonal y de esta manera obtener una estructura simple. Una transformación gráfica se realiza cuando $m = 2$, o hay dos factores comunes a la vez. En este caso, los factores comunes no correlacionados se consideran como vectores unitarios a lo largo de los ejes de coordenadas perpendiculares. Gráficamente, el par de cargas factoriales $(\hat{\ell}_{i1}, \hat{\ell}_{i2})$ da *p* puntos, en donde, a cada variable le corresponde un punto. Se puede entonces visualizar el eje de coordenadas rotado, a través del ángulo denotado ϕ. Luego, las nuevas cargas rotadas $\hat{\ell}^*_{ij}$ están determinadas por la siguiente ecuación:

$$\underset{(px2)}{\hat{L}^*} = \underset{(px2)}{\hat{L}} \; \underset{(2x2)}{T} \qquad (2.23)$$

En donde:

$$\begin{cases} T = \begin{bmatrix} \cos\phi & sen\phi \\ -sen\phi & \cos\phi \end{bmatrix} \begin{matrix} rotación & en \\ sentido & horario \end{matrix} \\ T = \begin{bmatrix} \cos\phi & -sen\phi \\ sen\phi & \cos\phi \end{bmatrix} \begin{matrix} rotación & en \\ sentido & antihorario \end{matrix} \end{cases}$$

Cuando se realiza un análisis gráfico en dos dimensiones, por lo general no se usa la expresión (2.23), ya que en este caso, es fácil encontrar el grupo de variables que permite identificar los factores comunes, por lo tanto no es necesario estudiar las magnitudes de las cargas factoriales.

Cuando $m > 2$, es preferible estudiar las magnitudes de las cargas *rotadas* para obtener una interpretación significativa de los datos originales, debido a que las orientaciones no son fácilmente visibles.

En ciertos casos, se utiliza una medida analítica de estructura simple conocida como el criterio varimax.

Sean $\tilde{\ell}_{ij}^{*} = \hat{\ell}_{ij}^{*} / \hat{h}_i$, los coeficientes escalares rotados de la raíz cuadrada de las comunalidades. Entonces, el criterio varimax selecciona la transformación ortogonal T que haga que $V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*4} - \left(\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*2}\right)^2 \Big/ p\right]$ sea tan grande como sea posible.

Luego de realizar esta transformación, se determina T; para mantener las comunalidades originales se multiplican las cargas $\tilde{\ell}_{ij}^{*}$ por $\hat{h}_i$.

Al maximizar V, se "distribuyen" los cuadrados de las cargas en cada factor tanto como se pueda. De esta manera, se espera encontrar grupos de coeficientes grandes en cualquier columna de la matriz de cargas rotadas $\hat{L}^{*}$.

Las rotaciones varimax de las cargas factoriales, que se obtienen mediante distintos métodos (componentes principales, máxima verosimilitud, etc.), generalmente, no coincidirán. Además, la muestra de las cargas rotadas puede cambiar considerablemente si los factores comunes adicionales se incluyen en la rotación. Si existe un solo factor dominante, éste no será relevante al realizar cualquier rotación ortogonal.[19]

2.5. PUNTUACIONES FACTORIALES

Las puntuaciones factoriales son los valores estimados de los factores comunes, aunque generalmente en el análisis factorial interesa conocer solamente los parámetros del modelo factorial, en algunas ocasiones es necesario conocer las puntuaciones factoriales, para usarlos como datos iniciales para un subsiguiente análisis.

---

[19] (7) JOHNSON, "Applied Multivariate Statistical Analysis", pág. 540-544.

Se estiman las puntuaciones factoriales a partir de valores para vectores factoriales aleatorios no observados F_j, $j = 1,2,\ldots,n$. Es decir, las puntuaciones factoriales se las denota como $\hat{f}_j$, que es la estimación de los valores f_j obtenidos de F_j (para el j-ésimo caso).

Ya que las cantidades no observadas f_j y ε_j son mayores al valor observado x_j, el proceso de estimación para obtener las puntuaciones factoriales se vuelve complejo. Para simplificar este problema, se considerará los siguientes puntos:

1. Tratar a las cargas factoriales estimadas $\hat{\ell}_{ij}$ y a las varianzas específicas estimadas $\hat{\psi}_i$ como si fueran valores verdaderos.

2. Es preferible utilizar transformaciones lineales de los datos originales, centrados o estandarizados. Para calcular las puntuaciones factoriales se debe usar las cargas estimadas rotadas, en lugar de las cargas estimadas originales.

A continuación se explican dos métodos de estimación usados para obtener las puntuaciones factoriales.

2.5.1. MÉTODO DE LOS MÍNIMOS CUADRADOS PONDERADOS

Se parte del ya conocido modelo factorial:

$$\underset{(px1)}{X} - \underset{(px1)}{\mu} = \underset{(pxm)}{L}\ \underset{(mx1)}{F} + \underset{(px1)}{\varepsilon}$$

Se supone inicialmente que el vector μ, la matriz cargas factoriales L, y la varianza específica Ψ, son conocidas.

Además, se tiene que los errores son los factores específicos $\varepsilon' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p]$, cuya $Var(\varepsilon_i) = \psi_i$, $i=1,2,\ldots,p$. Luego:

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \varepsilon \Psi^{-1} \varepsilon = (x - \mu - Lf)\Psi^{-1}(x - \mu - Lf) \qquad (2.24)$$

Para minimizar (2.24), se usan las estimaciones $\hat{f}$ de f:

$$\hat{f} = \left(L'\Psi^{-1}L\right)^{-1}L'\Psi^{-1}(x-\mu) \qquad (2.25)$$

Mediante (2.25), se toman las estimaciones $\hat{L}$, $\hat{\Psi}$, y $\hat{\mu} = \bar{x}$, como valores verdaderos, luego, las puntuaciones factoriales para el j-ésimo caso es:

$$\hat{f}_j = \left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\hat{L}'\hat{\Psi}^{-1}\left(x_j - \bar{x}\right) \qquad (2.26)$$

Cuando se usa el método de máxima verosimilitud para obtener $\hat{L}$ y $\hat{\Psi}$, estas estimaciones satisfacen la condición de unicidad: $\hat{L}'\hat{\Psi}^{-1}\hat{L} = \hat{\Delta}$, una matriz diagonal. Se tiene entonces que las puntuaciones factoriales obtenidas mediante el método de los mínimos cuadrados ponderados a partir de estimaciones de máxima verosimilitud son:

$$\hat{f}_j = \left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\hat{L}'\hat{\Psi}^{-1}\left(x_j - \hat{\mu}\right)$$

$$= \hat{\Delta}^{-1}\hat{L}'\hat{\Psi}^{-1}\left(x_j - \bar{x}\right), \qquad j = 1,2,...,n \qquad (2.27)$$

Las puntuaciones factoriales obtenidas mediante (2.27) tienen media muestral 0 y covarianza muestral nula.

Si hay rotación factorial, entonces para obtener las puntuaciones factoriales mediante (2.27), denotadas ahora por $\hat{f}_j^*$, se usan las cargas rotadas $\hat{L}^* = \hat{L}T$ en lugar de las cargas originales. $\hat{f}_j^*$, está relacionada a $\hat{f}_j$ mediante $\hat{f}_j^* = T'f_j$, $j = 1,2,...,n$.

Cuando se estiman las cargas factoriales mediante el método de componentes principales, las puntuaciones factoriales se generan usando el método de mínimos cuadrados no ponderados. Entonces, las puntuaciones factoriales son:

$$\hat{f}_j = \left(\hat{L}'\hat{L}\right)^{-1}\hat{L}'\left(x_j - \bar{x}\right)$$

Ahora, el valor de L es $L = \left[\sqrt{\hat{\lambda}_1} e_1 \vdots \sqrt{\hat{\lambda}_2} e_2 \vdots \cdots \vdots \sqrt{\hat{\lambda}_m} e_m \right]$, entonces:

$$\hat{f}_j = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}} \hat{e}_1' (x_j - \bar{x}) \\ \frac{1}{\sqrt{\hat{\lambda}_2}} \hat{e}_2' (x_j - \bar{x}) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}} \hat{e}_m' (x_j - \bar{x}) \end{bmatrix} \qquad (2.28)$$

Para estas puntuaciones factoriales, se tiene que media muestral $\frac{1}{n}\sum_{j=1}^{n} \hat{f}_j = 0$ y la covarianza muestral $\frac{1}{n-1}\sum_{j=1}^{n} \hat{f}_j \hat{f}_j' = I$.

2.5.2. MÉTODO DE REGRESIÓN

Se considera el modelo factorial original $X - \mu = LF + \varepsilon$. Cuando los factores comunes F y los factores específicos ε (o errores), tienen distribución normal conjunta, el modelo factorial $X - \mu = LF + \varepsilon$ también tiene distribución normal $N_p(O, LL' + \Psi)$. De igual manera, la distribución conjunta de $(X - \mu)$ y F es una normal $N_{m+p}(O, \Sigma^*)$, en donde

$$\underset{(m+p)x(m+p)}{\Sigma^*} = \begin{bmatrix} \Sigma = LL' + \Psi & \vdots & L \\ (pxp) & \vdots & (pxm) \\ \cdots & \cdots & \cdots \\ L' & \vdots & I \\ (mxp) & \vdots & (mxm) \end{bmatrix} \qquad (2.29)$$

Y O es un vector $(m+p)x1$ de ceros.

La distribución condicional de $F|x$ es una normal con

media = $E(F|x) = L'\Sigma^{-1}(x - \mu) = L'(LL' + \Psi)^{-1}(x - \mu)$ (2.30)

y

covarianza = $Cov\,(F|x) = I - L'\Sigma^{-1}L = I - L'(LL' + \Psi)^{-1}L$ (2.31)

$L'(LL' + \Psi)^{-1}$ en (2.30) son los coeficientes de regresión de los factores en las variables, cuyas estimaciones generan las puntuaciones factoriales. Por lo tanto, dado cualquier vector de observaciones x_j, y considerando las estimaciones de máxima verosimilitud $\hat{L}$ y $\hat{\Psi}$ como valores verdaderos, se tiene:

$$\hat{f}_j = \hat{L}'\hat{\Sigma}^{-1}(x_j - \bar{x}) = \hat{L}'(\hat{L}\hat{L}' + \hat{\Psi})^{-1}(x_j - \bar{x}), \quad j = 1,2,\ldots,n \qquad (2.32)$$

Se puede simplificar este cálculo del j-ésimo vector de puntuación factorial, $\hat{f}_j$, utilizando la matriz de identidad I, así:

$$\underset{(mxp)}{\hat{L}'}\ \underset{(pxp)}{(\hat{L}\hat{L}' + \hat{\Psi})^{-1}} = \underset{(mxm)}{(I + \hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}}\ \underset{(mxp)}{\hat{L}'}\ \underset{(mxp)(pxp)}{\hat{\Psi}^{-1}} \qquad (2.33)$$

Mediante esta identidad es posible comparar las puntuaciones factoriales en (2.32), generadas mediante regresión (denotadas como $\hat{f}_j^R$), con aquellas generadas mediante el método de los mínimos cuadrados ponderados (denotadas como $\hat{f}_j^{MC}$), de esta manera:

$$\hat{f}_j^{MC} = (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}(I + \hat{L}'\hat{\Psi}^{-1}\hat{L})f_j^R = \left(I + (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}\right)f_j^R \qquad (2.34)$$

Si las estimaciones se realizan mediante el método de máxima verosimilitud, teniendo $(\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1} = \hat{\Delta}^{-1}$, y si los elementos de esta matriz diagonal son cercanos a cero, los métodos de regresión y de mínimos cuadrados generan puntuaciones factoriales muy similares.

Si se desea determinar el número de factores con mayor precisión, se debe calcular las puntuaciones factoriales de (2.32) usando S (la matriz de covarianza muestral original) en lugar de $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi}$, de esta manera:

$$\hat{f}_j = \hat{L}'S^{-1}(x_j - \bar{x}), \quad j = 1,2,\ldots,n \qquad (2.35)$$

En el caso en que se usen las cargas rotadas $\hat{L}^* = \hat{L}T$ en lugar de las cargas originales en (2.35), la puntuación factorial $\hat{f}_j^*$ estará relacionada a $\hat{f}_j$ mediante

$$\hat{f}_j^* = T'\hat{f}_j, \qquad j = 1,2,\ldots,n.$$[20]

2.6. PERSPECTIVAS Y ESTRATEGIAS DEL ANÁLISIS FACTORIAL

Uno de los principales problemas que se presenta en el desarrollo del análisis factorial es la elección de m, el número de factores comunes. El tener una muestra grande en un modelo proporciona un m válido, siempre y cuando los datos tengan una distribución normal. Generalmente en el análisis factorial la elección correcta de m, se basa en algunas combinaciones de (1) la proporción de la varianza muestral, (2) el conocimiento que se tenga de la materia, y (3) la coherencia en los resultados.

Para que los resultados de un análisis factorial sean más satisfactorios, es recomendable realizar las rotaciones con más de un método y verificar que los resultados posean la misma estructura factorial.

Para obtener una solución confiable del análisis factorial se sugiere:

a) Utilizar el método de componentes principales en el análisis factorial como primera opción para analizar los datos, ya que no se requiere que R o S sean no singulares.

   Indagar observaciones sospechosas dibujando las puntuaciones factoriales. Calcular las puntuaciones estandarizadas para cada observación.

---

[20] (7) JOHNSON, "Applied Multivariate Statistical Analysis", pág. 550-554.

Intentar una rotación varimax.

b) Utilizar el método de máxima verosimilitud en el análisis factorial, con una rotación varimax.

c) De los dos análisis factoriales antes escritos, comparar sus soluciones.

   Verificar si las cargas se agrupan de la misma manera.

   Realizar un gráfico en donde se contrasten las puntuaciones factoriales obtenidas mediante componentes principales con las puntuaciones obtenidas mediante máxima verosimilitud.

d) Repetir los tres pasos anteriores para otro número de factores comunes m y verificar qué tanto influyen los factores extra en la interpretación de los datos.

e) Cuando se tiene un conjunto grande de datos, se los puede dividir en la mitad, y de esta manera, se realiza en cada parte un análisis factorial. Luego se compara los resultados obtenidos con cada parte y con los obtenidos de todo el conjunto de datos, para de este modo, verificar si la solución varía. (Los datos pueden dividirse aleatoriamente o se puede poner la primera mitad en un grupo y la otra mitad en otro grupo.)

2.6.1. APLICACIÓN DEL ANÁLISIS FACTORIAL

Para la realización de esta aplicación del análisis factorial, se utilizan seis variables que provienen de una medición realizada a 1907 vehículos del parque automotriz de Quito sobre los niveles de contaminación que emanan de monóxido de carbono (CO), hidrocarburos (HC) y óxido de nitrógeno (NOx). Para obtener los resultados de este análisis de utiliza como ayuda al programa SPSS 12.

Las variables a estudiarse se encuentran denotadas de la siguiente manera:

1. CO_alta: Cantidad emanada de monóxido de carbono, cuando el automóvil está encendido y con una mínima aceleración.

2. CO_baja: Cantidad emanada de monóxido de carbono, cuando el automóvil está encendido.

3. HC_alta: Cantidad emanada de hidrocarburos, cuando el automóvil está encendido y con una mínima aceleración.

4. HC_baja: Cantidad emanada de hidrocarburos, cuando el automóvil está encendido.

5. NOx_alta: Cantidad emanada de óxido de nitrógeno, cuando el automóvil está encendido y con una mínima aceleración.

6. NOx_baja: Cantidad emanada de óxido de nitrógeno, cuando el automóvil está encendido.

El objetivo primordial de este análisis es obtener factores que puedan explicar la relación entre las variables anteriormente descritas.

En la siguiente tabla 2.1, se encuentran los estadísticos, las medias, las desviaciones estándar y los tamaños muestrales para las variables incluidas en este análisis.

Estadísticos descriptivos

	Media	Deviación estándar	N del análisis
CO_baja	1,3650	1,6801	1907
HC_baja	297,0991	282,1861	1907
CO_alta	1,4799	1,7657	1907
HC_alta	297,4305	300,9216	1907
NOx_alta	1,0097	,1266	1907
NOx_baja	1,0293	,1367	1907

TABLA 2.1 Estadísticos descriptivos.

El siguiente paso es obtener la matriz de correlaciones, que describe las relaciones bivariadas en las que intervienen todas las variables. La matriz de correlaciones es la matriz en la que se basa el análisis factorial, y ayuda para apreciar la naturaleza de las relaciones entre las variables, tomadas de dos en dos.

Matriz de correlaciones

		CO_baja	HC_baja	CO_alta	HC_alta	NOx_alta	NOx_baja
Correlación	CO_baja	1,000	,576	,899	,539	-,408	-,393
	HC_baja	,576	1,000	,585	,880	-,104	-,049
	CO_alta	,899	,585	1,000	,560	-,456	-,303
	HC_alta	,539	,880	,560	1,000	-,071	-,048
	NOx_alta	-,408	-,104	-,456	-,071	1,000	,877
	NOx_baja	-,393	-,049	-,303	-,048	,877	1,000

TABLA 2.2 Matriz de correlaciones.

En la tabla 2.2, se puede ver que las variables HC_alta y HC_baja tienen una correlación muy alta entre sí, pero éstas variables tienen correlaciones bajas con las variables NOx_alta y NOx_baja. Esta matriz de correlaciones contiene 15 correlaciones, pero no por ello es fácil ver todos los posibles patrones que podrían presentarse.

La siguiente tabla 2.3, presenta las comunalidades. Inicialmente, las comunalidades son 1.00, la varianza para cada variable en forma de puntuación estándar. Sin embargo, una vez que se ha tomado en cuenta el número de factores, las comunalidades bajan y reflejan, para cada variable, la proporción de su varianza que se explica con los factores.

Comunalidades

	Inicial	Extracción
CO_baja	1,000	,798
HC_baja	1,000	,850
CO_alta	1,000	,803
HC_alta	1,000	,831
NOx_alta	1,000	,913
NOx_baja	1,000	,887

Método de extraction: Análisis de componentes principales.

TABLA 2.3 Comunalidades.

Se puede observar en la tabla anterior, que la comunalidad para la primera variable CO_baja es 0.798, esto quiere decir que el análisis factorial ha explicado 79.8% de su varianza. Si la comunidad para una variable es baja (<0.5), implicaría que el análisis factorial no explica gran parte de la varianza asociada a esa

variable. Es decir, la variable no tiene mucho en común con las restantes variables del análisis. Esto podría deberse a que posiblemente la variable es muy distinta de las demás variables, en cuanto a que en realidad está midiendo algo diferente, o que, se extrajo un número insuficiente de factores. Por estas razones, es importante considerar las estimaciones de comunalidad al interpretar un análisis factorial.

La tabla 2.4, presenta los valores propios para el análisis y las estimaciones de la varianza explicada por la solución final, utilizando el método de análisis de componentes principales.

Varianza total explicada

Componente	Valores propios iniciales			Sumas de las saturaciones al cuadrado de la extracción			Sumas de las saturaciones al cuadrado de la rotación		
	Total	% de Varianza	% acumulado	Total	% de Varianza	% acumulado	Total	% de Varianza	% acumulado
1	3,331	55,509	55,509	3,331	55,509	55,509	2,913	48,546	48,546
2	1,752	29,198	84,707	1,752	29,198	84,707	2,170	36,161	84,707
3	,581	9,687	94,393						
4	,179	2,978	97,372						
5	,120	2,003	99,374						
6	3,753E-02	,626	100,000						

Método de extracción: Análisis de componentes principales

TABLA 2.4 Varianza total explicada.

En la primera parte de esta tabla llamada “Valores propios iniciales” se pueden ver los valores propios, el porcentaje de la varianza y el porcentaje acumulado de la varianza para cada factor, en orden según la magnitud de los valores propios. El primer valor propio es 3.331, y explica 55.509% de la varianza. Se puede ver también que todos los valores propios son mayores que cero y que su suma es 6.

La sección “Suma de las saturaciones al cuadrado de la extracción” reproduce esta información para el número de factores extraídos en el análisis (dos, en este caso). Además, se puede ver que las sumas de las saturaciones al cuadrado de la extracción son iguales a los valores propios.

La sección “Suma de las saturaciones al cuadrado de la rotación” presenta la misma información para los factores rotados. Se puede observar que a pesar de que estas sumas son diferentes a las reportadas en “suma de las saturaciones al

cuadrado de la extracción", su suma (2.913 + 2.170) es igual a la suma de los valores propios (3.331 + 1.752).

En este ejemplo se extrajeron dos factores, ya que se extraen todos los factores con valores propios mayores a uno, y en este caso se tiene que son dos.

Para sustentar el hecho de que se extraen dos factores, se presenta a continuación el gráfico de sedimentación, que es un gráfico de los valores propios contra los factores. Este gráfico permite determinar el número de factores que mejor representan toda la varianza significativa descrita por la matriz de correlaciones.

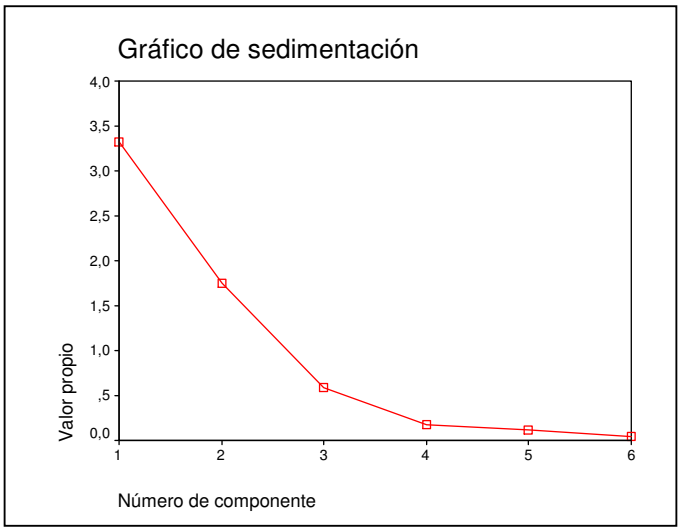

GRAFICO 2.1 Gráfico de sedimentación.

Como se puede observar, el gráfico 2.1 tiene una flexión en el factor 3, lo que significa que dos factores explican la principal varianza significativa en la matriz de correlaciones, debido a que, para el factor 3 y subsiguientes, la cantidad de varianza explicada es baja y prácticamente equivalente.

El siguiente paso es determinar la matriz de factores iniciales de componentes principales, que contiene las correlaciones de cada variable con cada componente principal.

Se puede apreciar en la tabla 2.5 que la matriz de factores iniciales (llamada Matriz de componentes en el SPSS) para este caso consiste de dos factores.

Matriz de componentes[a]

	Componente	
	1	2
CO_baja	,893	1,767E-02
HC_baja	,761	,520
CO_alta	,895	4,737E-02
HC_alta	,738	,535
NOx_alta	-,583	,757
NOx_baja	-,518	,787

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos.

TABLA 2.5 Matriz de componentes.

Como se puede ver en la tabla 2.5, el primer factor tiene saturaciones positivas y altas por parte de las cuatro primeras variables, y el segundo factor contiene saturaciones positivas altas únicamente por parte de las dos últimas variables.

El valor resultante de la sumatoria del cuadrado de las saturaciones factoriales de cada factor es igual al valor propio para ese factor. Es decir, $0.893^2 + 0.761^2 + \ldots + (-0.518)^2 = 3.331$. Por lo tanto, el valor propio es la suma de los cuadrados de las saturaciones factoriales para cada factor de componentes principales.

También, la sumatoria del cuadrado de cada saturación factorial para los factores de cada variable, es la comunalidad para esa variable. Es decir, $0.893^2 + (1.767E-02)^2 = 0.798$. Por lo tanto, la comunalidad de una variable es la suma de los cuadrados de las saturaciones factoriales para esa variable.

Luego de obtener la matriz de factores iniciales, es posible rotarla para producir una solución más fácil de interpretar.

Para este ejemplo se utilizó el método de rotación ortogonal Varimax y los resultados de la matriz de componentes rotados, figuran el la tabla 2.6.

Esta matriz de componentes rotados contiene saturaciones positivas altas por parte de las cuatro primeras variables, para el primer factor y el segundo factor contiene saturaciones positivas altas por parte de las dos últimas variables.

Matriz de componentes rotados[a]

	Componente	
	1	2
CO_baja	,775	-,444
HC_baja	,921	5,448E-02
CO_alta	,792	-,420
HC_alta	,908	7,904E-02
NOx_alta	-,111	,949
NOx_baja	-3,92E-02	,941

Método de extracción: Análisis de componentes principales.
Método de Rotación: Varimax con Normalización Kaiser.

a. La rotación ha convergido en 3 iteraciones

TABLA 2.6 Matriz de componentes rotados.

La matriz de componentes rotados se la interpreta considerando cada factor y determinando qué tienen en común todas las variables que presentan saturaciones altas en ese factor y qué no tienen en común con las variables que presentan saturaciones bajas.

Es importante primero determinar qué constituye una carga alta, en algunos casos se utiliza 0.30 como punto de corte y en otros casos 0.50 dependiendo del tamaño de la muestra (100 o más). Para este caso se utilizará 0.50.

El factor 1 contiene saturaciones altas (mayores que ±0.50) de las cuatro primeras variables, esto se debería posiblemente a que las cuatro primeras variables correspondientes a las emisiones de CO y HC contienen niveles de contaminación mayores que de NOx.

El factor 2 obtiene saturaciones altas de las variables correspondientes a las emisiones de NOx, que son niveles de contaminación “menores o más bajos” comparándolas con la contaminación que produce la emanación de CO y HC.

El siguiente gráfico 2.2, es el gráfico de las componentes en el espacio rotado, donde se aprecia las posiciones de las variables respecto a los ejes factoriales rotados. Se puede distinguir que todas las variables están bien representadas

sobre el plano, debido a que se encuentran próximas al borde del círculo de radio unidad, y no están cerca del origen.

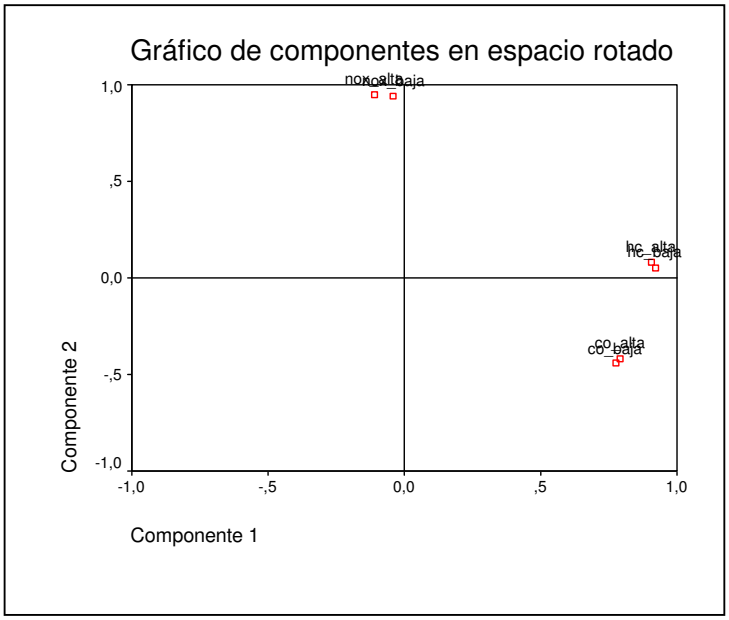

GRAFICO 2.2 Gráfico de componentes en espacio rotado.

Adicionalmente, en el análisis factorial, es posible calcular las puntuaciones factoriales para representar los factores en cuestión. También se las podrá usar para representar los factores en análisis posteriores, como por ejemplo en un análisis de regresión múltiple.

A continuación, en la tabla 2.7 se encuentra la matriz de coeficientes de las puntuaciones factoriales obtenidas para cada variable, a partir de la matriz de factores rotada. Es decir, los coeficientes que permiten expresar cada factor como combinación lineal de todas las variables.

Matriz de coeficientes de las puntuaciones en las componentes

	Componente	
	1	2
CO_baja	,235	-,129
HC_baja	,349	,137
CO_alta	,244	-,115
HC_alta	,347	,148
NOx_alta	,072	,461
NOx_baja	,098	,465

Método de extracción: Análisis de componentes principales.
Método de Rotación: Varimax con Normalización Kaiser.

TABLA 2.7 Matriz de coeficientes de las puntuaciones factoriales.

Para concluir este ejemplo de aplicación, se puede decir que el primer factor extraído está formado por variables que emiten un mayor nivel de contaminación al ambiente, mientras que las dos últimas variables no tienen mayor repercusión en la contaminación ambiental y son las variables que componen el segundo factor.

CAPÍTULO 3
MODELOS DE ECUACIONES ESTRUCTURALES

3.1. IINTRODUCCIÓN

Los modelos de ecuaciones estructurales son conjuntos de ecuaciones lineales, utilizados para especificar fenómenos en términos de sus variables de causa-efecto.

Estos modelos son, el resultado de la evolución y unión de la metodología desarrollada en el análisis de senderos y en el análisis factorial.

La modelización de ecuaciones estructurales incorpora variables no observables directamente, llamadas variables latentes o constructos, que sólo pueden ser medidas a través de otras variables directamente observables.

Los modelos de ecuaciones estructurales están formados por: los modelos estructurales, compuestos por el análisis de senderos, y los modelos de medida, que son el análisis de variables latentes o no observables.

Los modelos de ecuaciones estructurales constituyen una poderosa herramienta de análisis, cuyo verdadero valor está en usar simultáneamente variables observadas y variables latentes. En el desarrollo de un modelo de ecuaciones estructurales es necesario que se lleve a cabo cuatro fases: la especificación, la identificación, la estimación y la evaluación e interpretación de dicho modelo.[21]

Los modelos de ecuaciones estructurales son una eficaz técnica de análisis multivariante y son particularmente de gran ayuda en las Ciencias Sociales y del comportamiento, y suelen usarse en el estudio de las relaciones entre las áreas sociales y los logros obtenidos; por ejemplo, la discriminación en los empleos, la eficacia de programas de acción social, etc.

---

[21] (12) LUQUE, “Técnicas de Análisis de Datos en Investigación de Mercados”, pág. 489-492.

3.2. ESPECIFICACIÓN DEL MODELO GENERAL DE ECUACIONES ESTRUCTURALES

Un modelo de ecuaciones estructurales es un modelo de senderos completo, el cual se lo puede describir mediante un diagrama de senderos. El modelo de ecuaciones estructurales difiere de un modelo de senderos simple en que se pueden utilizar variables latentes en su análisis.

Los modelos de ecuaciones estructurales incorporan errores de medida, como consecuencia de no medir perfectamente las variables latentes mediante las variables observadas, ya sean debido a los encuestados, o debido al investigador. Las personas expuestas a un cuestionario pueden dar respuestas inexactas a las cuestiones planteadas, bien por no querer decir la verdad, bien por desconocimiento u otros motivos. También el investigador contribuye al error de medida al intentar medir conceptos teóricos, tales cono actitudes, comportamientos, opiniones, etc., mediante una serie de ítems en un cuestionario.

En los modelos de ecuaciones estructurales, las variables del modelo que son variables latentes, son medidas por un conjunto de indicadores.

Las variables latentes en los modelos de ecuaciones estructurales son análogas a los factores en el análisis factorial. Tanto las variables latentes como los factores son funciones estadísticas de un conjunto de variables medibles.

Una ventaja de la modelización de ecuaciones estructurales es que se pueden analizar datos experimentales, no experimentales o ambos a la vez.

De todos los tipos de modelos de ecuaciones estructurales, los modelos más generales son los modelos de regresión estructural, que pueden verse como una síntesis de los modelos de senderos y los de medida.

La especificación de un modelo de regresión estructural permite probar hipótesis sobre patrones de efectos causales, como en el análisis de senderos. A diferencia de los modelos de senderos, estos efectos pueden involucrar variables latentes ya que un modelo de regresión estructural también incluye un modelo de medida que

contiene variables observadas como indicadores de factores subyacentes, del mismo modo que en el análisis factorial.

La capacidad para probar hipótesis sobre relaciones de medida y estructurales, con un modelo simple ofrece mucha flexibilidad.

Para detallar ciertas características de los modelos de regresión estructural, se presentan a continuación, algunos tipos de estos modelos.

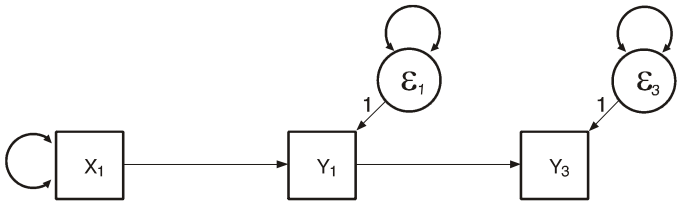

GRÁFICO 3.1 Ejemplo de un modelo de senderos[22].

El gráfico (3.1) representa un modelo estructural con variables observadas, es decir, un modelo de senderos, que caracteriza el uso de una medida simple de cada constructo. La variable exógena de este modelo, X_1, se asume que está medida sin error, suposición que en la práctica es violada; no se necesita esta suposición para las variables endógenas en el modelo, el error de medición en Y_1 o en Y_3 se presenta en las variables residuales ε_1, ε_3.

[22] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 210.

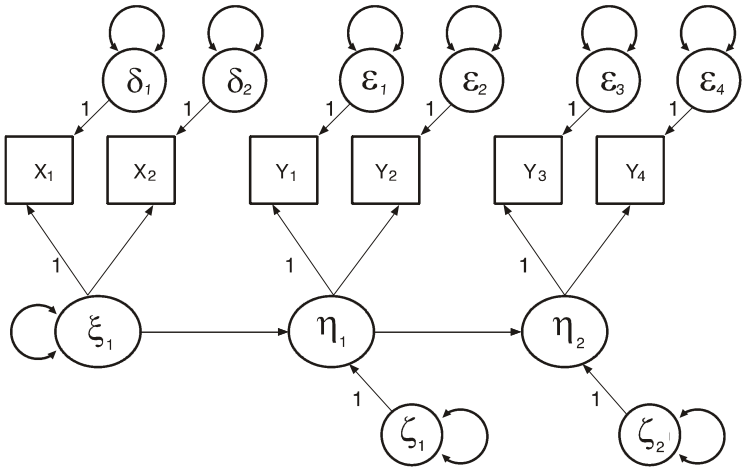

GRÁFICO 3.2 Ejemplo de un modelo de regresión estructural[23].

El modelo del gráfico (3.2) es un modelo de regresión estructural con componentes de medida y estructural. Su modelo de medida tiene las mismas tres variables observadas representadas en el modelo de senderos, X_1, Y_1, y Y_3. A diferencia del modelo de senderos, cada una de estas tres variables está especificada como un par de indicadores de un factor subyacente. Consecuentemente, todas las variables observadas del gráfico (3.2) tienen error de medición. Este modelo de regresión estructural tiene también componente estructural que describe el mismo patrón básico de efectos causales directos e indirectos como en el modelo de senderos pero entre variables latentes en lugar de variables observadas. Cada variable endógena latente de este modelo de regresión estructural tiene una variable residual (ε_1, ε_2, ε_3, ε_4). A diferencia del modelo de senderos, las variables residuales del modelo de regresión estructural reflejan solamente causas omitidas y no errores de medición. Por la misma razón, estimaciones de efectos directos (es decir, los coeficientes de senderos) para el modelo de regresión estructural ($\xi_1 \rightarrow \eta_1$, $\eta_1 \rightarrow \eta_2$) están corregidos para errores de medición, pero estimaciones de efectos directos para el modelo de senderos ($X_1 \rightarrow Y_1$, $Y_1 \rightarrow Y_3$) no.

---

[23] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 210.

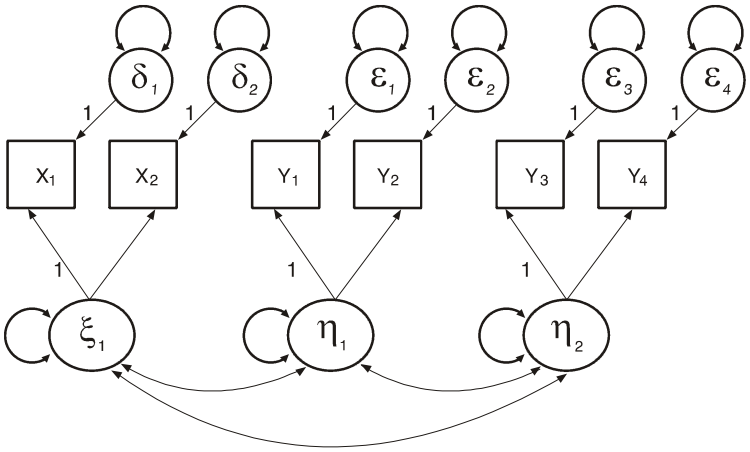

GRÁFICO 3.3 Ejemplo de un modelo de análisis factorial confirmatorio[24].

En el gráfico (3.3) se encuentra representado un modelo de análisis factorial confirmatorio. Este modelo presenta el enfoque del indicador-múltiple[25] de medición, donde todas las asociaciones entre los factores están especificadas como no analizadas.

El modelo de regresión estructural del gráfico (3.2) podría describirse como "completamente latente" porque todas las variables en este modelo estructural son latentes. Sin embargo, esta característica es deseable porque implica una medida de indicador-múltiple; también es posible representar en los modelos de regresión estructural una variable observada que es un indicador simple de un constructo. Tales modelos podrían llamarse "parcialmente latentes" porque por lo menos una variable del modelo estructural en un indicador simple.[26]

---

[24] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 210.

[25] Cuando se tiene más de una variable observada en un modelo de regresión estructural, estas variables son indicadores múltiples de los constructos.

[26] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 209-211.

De manera general, un modelo de ecuaciones estructurales se lo puede plantear de diferentes maneras: mediante un diagrama, matricialmente o escribiendo un sistema de ecuaciones simultáneas.

A continuación se encuentra el gráfico 3.4 en el que se describe un modelo causal hipotético de ecuaciones estructurales y sus componentes:

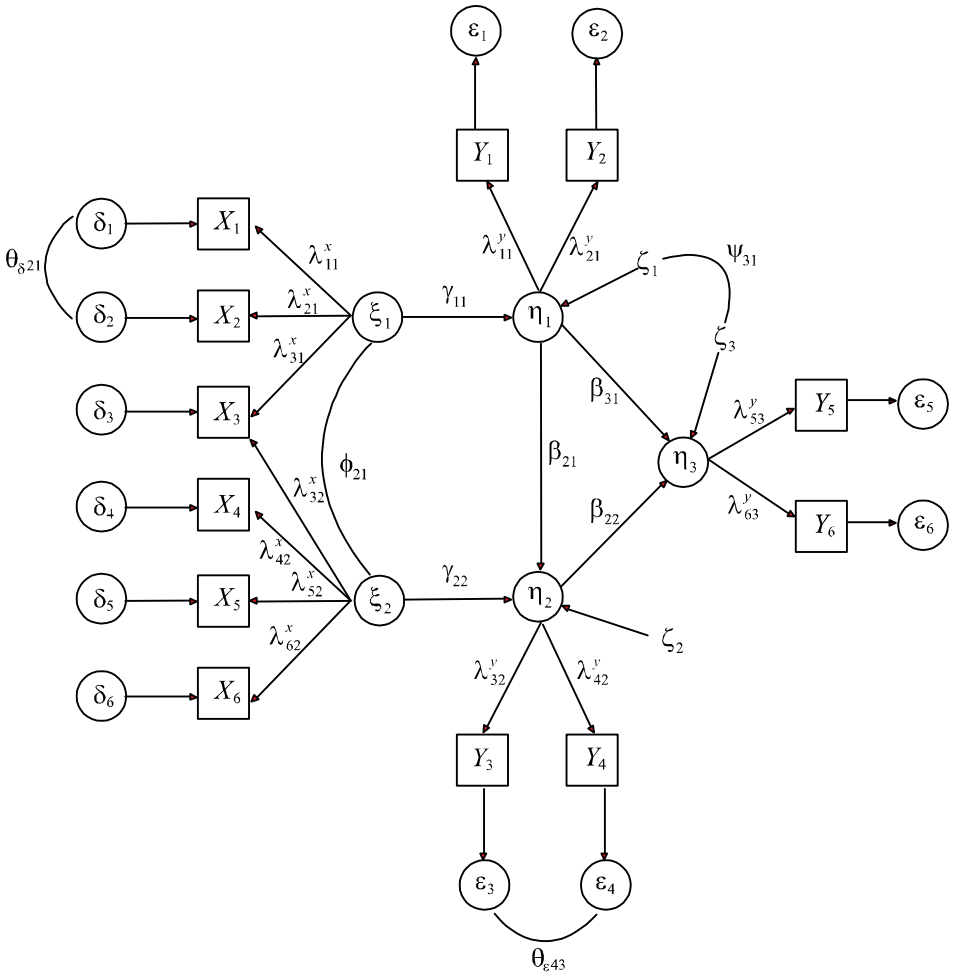

GRÁFICO 3.4 Modelo causal hipotético[27].

Cuyos respectivos componentes son:

- Variables latentes: endógenas η_1, η_2, η_3, exógenas ξ_1, ξ_2,

[27] (12) LUQUE, "Técnicas de Análisis de Datos en Investigación de Mercados", pág. 495.

- Variables observadas: endógenas Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, exógenas X_1, X_2, X_3, X_4, X_5, X_6,

- Errores de medida: de variables observadas endógenas ε_1, ε_2, ε_3, ε_4, ε_5, ε_6, de variables observadas exógenas δ_1, δ_2, δ_3, δ_4, δ_5, δ_6,

- Coeficientes de correlación: θ_ε, θ_δ, correlacionan a los errores de medida.

- Términos de perturbación: ζ_1, ζ_2, ζ_3, los cuales incluyen los efectos de las variables omitidas, los errores de medida y la aleatoriedad del proceso especificado. La variación en el término de perturbación se denota por ψ y la covariación entre los términos de perturbación i-ésimo y j-ésimo se denota por ψ_{ij}.

- Coeficientes de regresión: λ_x, λ_y, que relacionan las variables latentes con las observadas.

- Coeficientes de regresión: γ_{11}, γ_{22}, β_{21}, β_{22}, β_{31}, ϕ_{21} que relacionan las variables latentes entre sí, y las variables observadas entre sí.

El modelo causal hipotético del gráfico 3.4, está compuesto por 3 variables latentes endógenas (η_1, η_2, η_3), 2 variables latentes exógenas (ξ_1, ξ_2) y por 12 variables observadas o indicadores (X_1, X_2, X_3, X_4, X_5, X_6, Y_1, Y_2, Y_3, Y_4, Y_5, Y_6).

Las variables latentes exógenas son medidas a través de las variables observadas X, mientras que las variables latentes endógenas son medidas mediante las variables observadas Y. Como se aprecia en el gráfico 3.4, lo normal es que las relaciones causales partan de las variables latentes hacia las observadas; a éstas se les denomina indicadores reflexivos. Aunque es muy poco usual cabe la posibilidad del caso inverso, es decir, que las variables observadas influyan sobre las latentes, considerando a éstas como indicadores agregados establecidos a partir de indicadores parciales.

Por otro lado, al no medirse perfectamente los conceptos teóricos del modelo a través de las variables observadas se producen errores de medida, representados mediante δ para las variables X y ε para las variables Y. De igual manera, cuando

se trata de explicar una variable latente a través de otras, se produce un término de perturbación o error estructural ζ que incluye los efectos de las variables desconocidas, las variables omitidas en el modelo, los errores de medida y la aleatoriedad del proceso especificado. Como muestra el gráfico 3.4, las variables exógenas (ξ) no presentan término de perturbación al considerarse variables independientes.

Se considera también en este modelo hipotético que el error de medida δ_1 está correlacionado con el error de medida δ_2, lo que sucede en ocasiones. Esto es frecuente en estudios de carácter longitudinal en los que se aplica una misma medida en varios intervalos de tiempo diferentes. Esta correlación se representa en el diagrama de senderos con la letra griega θ y a través de una curva entre los dos errores. Así mismo, se considera que los términos de perturbación de las variables η_1 y η_3 presentan una covariación diferente a 0, siendo representada por Ψ_{31}.

Las flechas unidireccionales entre dos variables indican una influencia directa de una variable sobre la otra, siendo los parámetros asociados a cada flecha los coeficientes que representan la relación entre las variables. Cada parámetro lleva dos subíndices, el primero corresponde a la variable de llegada de la flecha (efecto) y el segundo a la variable de salida (causa). Los parámetros que expresan la relación entre las variables latentes exógenas y su medida se representan mediante la letra lamdba con un superíndice X (λ^X), mientras que los parámetros entre las variables latentes endógenas y su medida se representa de la misma forma pero con el superíndice Y (λ^Y). Paralelamente, el parámetro que representa la relación entre una variable latente exógena y una endógena se indica a través de la letra gamma (γ), y la relación entre dos variables latentes endógenas se representa mediante la letra beta (β). Por último, la covariación entre las variables exógenas se representa mediante una línea curva y la letra phi (ϕ).

El modelo de ecuaciones estructurales, como ya se indicó, está compuesto por dos sub-modelos que pueden expresarse de forma matricial:

a) Modelo estructural: representa una red de relaciones en forma de un conjunto de ecuaciones lineales que enlazan variables latentes endógenas con variables latentes exógenas. El modelo estructural se considera como una extensión de la regresión, estableciendo tantas ecuaciones como variables latentes endógenas haya. Así, cada constructo endógeno es la variable dependiente de la ecuación y el resto de constructos endógenos y exógenos relacionados son las variables independientes.

La representación del modelo estructural en forma de ecuación es:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \qquad (3.1)$$

Donde: η: vector *m x 1* de variables latentes endógenas,

B: matriz *m x m* de coeficientes de las variables endógenas,

Γ: matriz *m x k* de coeficientes de las variables exógenas,

ξ: vector *k x 1* de variables latentes exógenas,

ζ: vector *m x 1* de términos de perturbación aleatoria.

La representación en forma de ecuación para el modelo del gráfico 3.4 es:

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \gamma_{22}\xi_2 + \beta_{21}\eta_1 + \zeta_2$$
$$\eta_3 = \beta_{31}\eta_1 + \beta_{32}\eta_2 + \zeta_3$$

Aquí, existen tres únicas variables latentes endógenas representadas en tres ecuaciones que incluyen los términos de perturbación. La variable η_1, que en la primera ecuación es dependiente, en la segunda y tercera se considera independiente.

En forma matricial, el modelo estructural del gráfico 3.4 es:

$$\begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix} = \begin{pmatrix}0 & 0 & 0\\ \beta_{21} & 0 & 0\\ \beta_{31} & \beta_{32} & 0\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix} + \begin{pmatrix}\gamma_{11} & 0\\ 0 & \gamma_{22}\\ 0 & 0\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + \begin{pmatrix}\zeta_1\\ \zeta_2\\ \zeta_3\end{pmatrix}$$

b) Modelo de medida: especifica las variables observadas o indicadores, que van a servir para medir los conceptos teóricos propuestos en el modelo estructural. Para poner en forma de ecuaciones el modelo de medida, se establecerán tantas como indicadores posea el modelo representado en el diagrama causal, con independencia de que sean exógenos (X) o endógenos (Y). Estas ecuaciones son:

$$y = \Lambda_y \eta + \varepsilon \qquad (3.2)$$

$$x = \Lambda_x \xi + \delta \qquad (3.3)$$

Donde: η: vector *m x 1* de variables latentes endógenas,

ξ: vector *k x 1* de variables latentes exógenas,

Λ_x: matriz *q x k* de coeficientes de variables exógenas,

Λ_y: matriz *p x m* de coeficientes de variables endógenas,

δ: vector *q x 1* de errores de medición para los indicadores exógenos,

ε: vector *p x 1* de errores de medición para los indicadores endógenos.

Las ecuaciones del modelo de medida del gráfico 3.4 son:

$$\begin{aligned} Y_1 &= \lambda_{11}^{y}\eta_1 + \varepsilon_1 \\ Y_2 &= \lambda_{21}^{y}\eta_1 + \varepsilon_2 \\ Y_3 &= \lambda_{32}^{y}\eta_2 + \varepsilon_3 \\ Y_4 &= \lambda_{42}^{y}\eta_2 + \varepsilon_4 \\ Y_5 &= \lambda_{53}^{y}\eta_3 + \varepsilon_5 \\ Y_6 &= \lambda_{63}^{y}\eta_3 + \varepsilon_6 \end{aligned}$$

Modelo de medida de indicadores endógenos

$$\begin{aligned} X_1 &= \lambda_{11}^{x}\xi_1 + \delta_1 \\ X_2 &= \lambda_{21}^{x}\xi_1 + \delta_2 \\ X_3 &= \lambda_{31}^{x}\xi_1 + \lambda_{32}^{x}\xi_2 + \delta_3 \\ X_4 &= \lambda_{42}^{x}\xi_2 + \delta_4 \\ X_5 &= \lambda_{52}^{x}\xi_2 + \delta_5 \\ X_6 &= \lambda_{62}^{x}\xi_2 + \delta_6 \end{aligned}$$

Modelo de medida de indicadores exógenos

En este caso hay 12 ecuaciones, de las cuales 6 establecen la medida de las variables latentes exógenas y las otras 6 la medida de las 3 variables latentes endógenas. También se puede apreciar cómo la variable X_3 sirve para medir los dos constructos exógenos, lo que no es recomendable ni usual, pero puede darse siempre bajo un razonamiento teórico.

La forma matricial para el modelo de medida del gráfico 3.4 es:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} \lambda_{11}^{y} & 0 & 0 \\ \lambda_{21}^{y} & 0 & 0 \\ 0 & \lambda_{32}^{y} & 0 \\ 0 & \lambda_{42}^{y} & 0 \\ 0 & 0 & \lambda_{53}^{y} \\ 0 & 0 & \lambda_{63}^{y} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$ Modelo de medida de indicadores endógenos

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix} = \begin{pmatrix} \lambda_{11}^{x} & 0 \\ \lambda_{21}^{x} & 0 \\ \lambda_{31}^{x} & \lambda_{32}^{x} \\ 0 & \lambda_{42}^{x} \\ 0 & \lambda_{52}^{x} \\ 0 & \lambda_{62}^{x} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{pmatrix}$$ Modelo de medida de indicadores exógenos[28]

Es importante considerar también que los modelos de ecuaciones estructurales son un caso especial del modelo de estructura de covarianza general.

La matriz de covarianza para y y x , que se obtiene sustituyendo la ecuación (3.1) en la (3.2):

$$\mathrm{y} = \Lambda_{\mathrm{y}}\, \mathrm{B} + \Lambda_{\mathrm{y}}\, \Gamma\xi + \Lambda_{\mathrm{y}}\, \zeta + \varepsilon$$

Reescribiendo la ecuación anterior de tal manera que las variables endógenas estén en un lado de la ecuación y las variables exógenas en el otro lado, se tiene:

$$(\mathrm{I-B})\mathrm{y} = \Lambda_{\mathrm{y}}\, \Gamma\xi + \Lambda_{\mathrm{y}}\, \zeta + \varepsilon$$

---

[28] (12) LUQUE, "Técnicas de Análisis de Datos en Investigación de Mercados", pág. 495-499.

Asumiendo que (I-B) es no singular y por tanto su inversa existe, la ecuación anterior puede reescribirse como:

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

Esta última ecuación puede representarse en términos de medias, varianzas y covarianzas. Para ello, sea Ω un vector que contiene los parámetros estructurales del modelo, en este caso $\Omega = (B,\Gamma,\Psi,\Phi)$.

Además, cov(x) = E(x´x) = Φ, y E(ζ) = 0.

$$\Sigma_{xx} = \text{Cov}\,(xx') = \Lambda_x E(\xi\xi')\Lambda'_x + E(\delta\delta') = \Lambda_x\Phi\Lambda'_x + \Theta_\delta$$

De este modo,

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \qquad (3.4)$$

$$= \begin{bmatrix} \Lambda_y (I-B)^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-B)'^{-1}\Lambda_y + \Theta_\varepsilon & \Lambda_y (I-B)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi(I-B)'^{-1}\Lambda_y' & \Lambda_x\Phi\Lambda_x'+\Theta_\delta \end{bmatrix},$$

donde: Φ : matriz de covarianzas $k\ x\ k$ de variables latentes exógenas,

Ψ : matriz de covarianzas $m\ x\ m$ de los términos de perturbación,

Θ_ε : matriz de covarianzas de los errores de medición ε, y,

Θ_δ : matriz de covarianzas de los errores de medición δ.

Para el modelo del gráfico 3.4, estas matrices son:

$$\Phi = \begin{pmatrix} & \xi_1 & \xi_2 \\ \xi_1 & - & \\ \xi_2 & \phi_{21} & - \end{pmatrix}$$ Matriz de covarianzas 2 x 2 de variables latentes exógenas.

$$\Psi = \begin{pmatrix} & \zeta_1 & \zeta_2 & \zeta_3 \\ \zeta_1 & - & & \\ \zeta_2 & & - & \\ \zeta_3 & \psi_{31} & & - \end{pmatrix}$$

Matriz de covarianzas 3 x 3 de los términos de perturbación.

$$\Theta_\varepsilon = \begin{pmatrix} & \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & \varepsilon_5 & \varepsilon_6 \\ \varepsilon_1 & - & & & & & \\ \varepsilon_2 & & - & & & & \\ \varepsilon_3 & & & - & & & \\ \varepsilon_4 & & & \theta_{\varepsilon 43} & - & & \\ \varepsilon_5 & & & & & - & \\ \varepsilon_6 & & & & & & - \end{pmatrix}$$

Matriz de covarianzas 6 x 6 de los errores de medición ε.

$$\Theta_\delta = \begin{pmatrix} & \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 \\ \delta_1 & - & & & & & \\ \delta_2 & \theta_{\delta 21} & - & & & & \\ \delta_3 & & & - & & & \\ \delta_4 & & & & - & & \\ \delta_5 & & & & & - & \\ \delta_6 & & & & & & - \end{pmatrix}$$

Matriz de covarianzas 6 x 6 de los errores de medición δ.

En términos del vector de parámetros Ω, se tiene que $\Omega = (\Lambda_y, \Lambda_x, \Theta_\varepsilon, \Theta_\delta, \Phi, B, \Gamma, \Psi)$.

3.3. IDENTIFICACIÓN DE LOS MODELOS DE ECUACIONES ESTRUCTURALES

Un modelo se dice que está identificado, si los parámetros del modelo completo (modelo estructural y modelo de medida, juntos) pueden estimarse a partir de los elementos de la matriz de covarianza de las variables observadas. Cuando se combinan los modelos: estructural y el de medida, en uno solo, se puede agregar el nuevo conjunto de condiciones de identificación al modelo de ecuaciones estructurales.

Como primer paso, se fija la métrica de las variables latentes exógenas poniendo una carga de 1.0 en cada columna de la matriz Λ_x, o poniendo 1.0 en los elementos de la diagonal de la matriz Φ. De igual manera se fija la métrica de las variables latentes endógenas poniendo una carga de 1.0 en cada columna de la matriz Λ_y. Una vez que se ha determinado la métrica de las variables latentes, se puede considerar un conjunto de reglas que sirven para identificar los modelos de ecuaciones estructurales.

La primera es la *regla del conteo*. Se denotará al número total de variables con $s = p + q$, siendo p las variables endógenas y q las exógenas. Luego, $\frac{1}{2} s(s+1)$ es el número de elementos no redundantes en Σ. Además, se denota al número total de parámetros a ser estimados en el modelo como t, entonces, para realizar la identificación del modelo se debe tener la siguiente condición necesaria $t \leq \frac{1}{2} s(s+1)$.

- Si se tiene la igualdad, se dice que el modelo está *identificado*.
- Si t es estrictamente menor que $\frac{1}{2} s(s+1)$, se dice que el modelo está *sobre identificado*.
- Si t es mayor que $\frac{1}{2} s(s+1)$, entonces el modelo *no está identificado*.

Otro método de identificación es la *regla de los dos pasos*, esta regla trata los modelos de ecuaciones estructurales como modelos de análisis factorial restringidos.

Lo primero que se debe hacer es reparametrizar al modelo de ecuaciones estructurales como un modelo de análisis factorial confirmatorio, de este modo, los elementos en Φ pueden trasladarse a elementos en Γ y B. Luego, evaluar este modelo de análisis factorial confirmatorio con los requerimientos necesarios para saber si el modelo está identificado; para lo cual, el modelo de análisis factorial confirmatorio debe cumplir con dos condiciones necesarias: (1) que el número de parámetros libres sea menor o igual al número de elementos de la matriz de varianzas-covarianzas (es decir, que los grados de libertad sea mayor o igual a 0),

y (2) cada variable latente, que incluya errores de medida y factores, debe tener escala.

El segundo paso consiste en tratar al modelo estructural como si éste fuera entre variables observadas. Si el modelo de variables observadas es recursivo, entonces el modelo estructural está identificado; caso contrario, si el modelo de variables observadas no es recursivo, entonces se debe evaluar nuevamente el modelo estructural con los requerimientos necesarios para su identificación.

A continuación, se presenta un ejemplo ilustrativo de la regla de los dos pasos para la identificación de un modelo de ecuaciones estructurales.

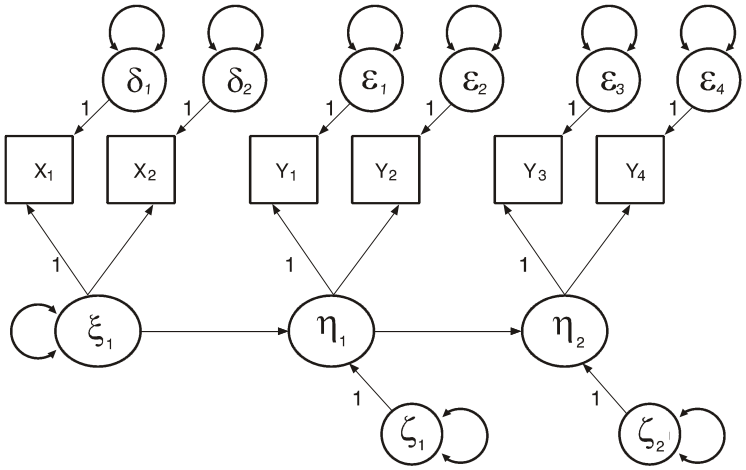

GRÁFICO 3.5 Modelo estructural original[29].

El modelo del gráfico 3.5 reúne los requerimientos necesarios porque cada variable latente está escalada y hay más covarianzas que parámetros libres. Específicamente, con 6 variables observadas, hay 6(7)/2 = 21 covarianzas disponibles para estimar estos 14 parámetros del modelo, que incluyen 9 varianzas de variables exógenas (el factor ξ_1, 6 errores de medida, y 2 perturbaciones) y 5 efectos directos en las variables endógenas (3 factores de carga –uno por factor-, y 2 senderos: $\xi_1 \rightarrow \eta_1$, $\eta_1 \rightarrow \eta_2$); por lo tanto, los grados de libertad del modelo es igual a 7.

---

[29] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 214.

Sin embargo, todavía no se conoce si el modelo del gráfico 3.5 está identificado, para averiguarlo, se aplicará la regla de los dos pasos. La reespecificación de este modelo de ecuaciones estructurales a un modelo de análisis factorial confirmatorio se muestra en el gráfico 3.6.

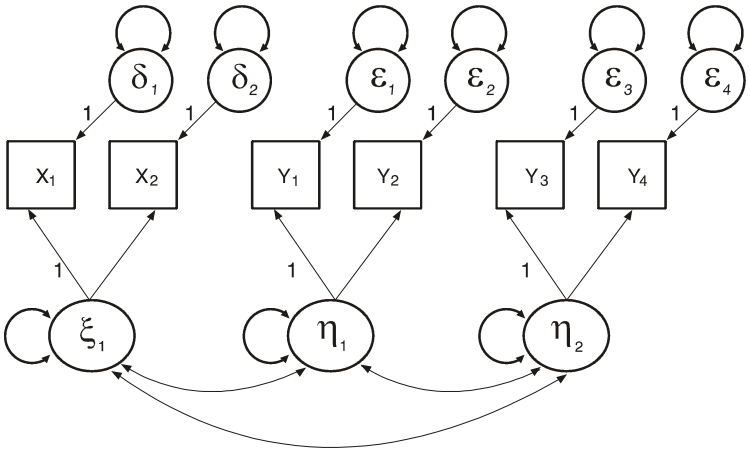

GRÁFICO 3.6 Modelo reespecificado como un modelo de análisis factorial confirmatorio[30].

Debido a que este modelo de análisis factorial confirmatorio tiene al menos dos indicadores por factor, está identificado. Por esta razón, la primera parte de la regla de los dos pasos está efectuada. La parte estructural del modelo de ecuaciones estructurales se indica en el gráfico 3.7.

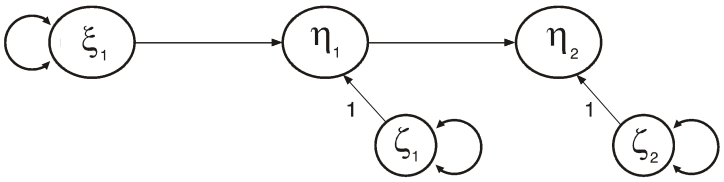

GRÁFICO 3.7 Modelo estructural[31].

[30] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 214.

[31] (10) KLINE, "Principles and Practice of Structural Equation Modeling", pág. 214.

Debido a que este modelo es recursivo, está también identificado. Ya que el modelo estructural original del gráfico 3.5 reúne la condición suficiente de la regla de los dos pasos, está por lo tanto, identificado.

Las dos reglas, anteriores son condiciones de identificación necesarias o suficientes, pero no ambas. Si no se tiene condiciones suficientes para la identificación, se podría intentar resolver el modelo con los parámetros de la forma estructural en términos de los parámetros de la forma reducida.[32]

3.4. PRUEBAS E INTERPRETACIÓN EN LOS MODELOS DE ECUACIONES ESTRUCTURALES

Partiendo de que el modelo está identificado, el siguiente paso es realizar la estimación del modelo de ecuaciones estructurales.

Para realizar la estimación de los parámetros del modelo, hay que decidir la forma que tendrán los datos de entrada. Los modelos de ecuaciones estructurales pueden usar como datos iniciales la matriz de varianzas-covarianzas o la matriz de correlaciones de las variables observadas. Tradicionalmente los modelos de ecuaciones estructurales han sido formulados para usar la matriz de varianzas-covarianzas; no obstante, es posible emplear también la matriz de correlaciones.

El uso de la matriz de varianzas-covarianzas presenta la ventaja de proporcionar comparaciones válidas entre diferentes poblaciones y muestras, debido a que los coeficientes estimados conservan la unidad de medida de los indicadores. Toda la teoría estadística clásica de los modelos de ecuaciones estructurales está basada en las propiedades distributivas de los elementos de la matriz de covarianzas, de manera que si se utiliza la matriz de correlaciones, que presenta una distribución diferente, podrían producirse problemas como la obtención de errores estándar equívocos. Sin embargo, existen ciertos paquetes estadísticos que están

---

[32] (10) KLINE, “Principles and Practice of Structural Equation Modeling”, pág. 212-215.

programados para proporcionar errores estándar correctos, independientemente de que se emplee una u otra matriz, lo que elimina este problema.

Por otra parte, la utilización de la matriz de varianzas-covarianzas presenta un pequeño inconveniente en cuanto a la interpretación de los resultados: los coeficientes deben ser interpretados en términos de las unidades de medida para los constructos. No obstante, esta dificultad se corrige obteniendo a posteriori los coeficientes estandarizados.

El uso de la matriz de correlaciones, radica en el hecho de que tiene un rango común de variación entre -1 y +1, lo que hace posible la comparación directa de los coeficientes dentro de un modelo. El uso de esta matriz es adecuado sólo cuando se pretende comprender el patrón de las relaciones entre los conceptos teóricos establecidos, pero no como prueba rigurosa de la teoría. En estos casos, se debe tener cuidado al momento de generalizar los resultados obtenidos a otras situaciones, puesto que los parámetros obtenidos mediante la matriz de correlaciones, si bien son correctos, no presentan unas adecuadas pruebas de significación al producir errores estándar inadecuados, a no ser que se emplee algún paquete informático que elimine ese inconveniente.

Por lo tanto, es recomendable partir de la matriz de varianzas-covarianzas y obtener posteriormente las soluciones estandarizadas para facilitar la interpretación de los resultados.

Se consideran, las ecuaciones (3.1), (3.2) y (3.3) que son los submodelos estructural y de medida que conforman un modelo de ecuaciones estructurales:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta$$

$$y = \Lambda_y \eta + \varepsilon$$

$$x = \Lambda_x \xi + \delta$$

Los parámetros del modelo son:

- Las varianzas y covarianzas de las variables exógenas de Φ (matriz de covarianza $k\ x\ k$ de variables latentes exógenas).

  Para el modelo del gráfico 3.4, esta matriz es:

$$\Phi = \begin{pmatrix} & \xi_1 & \xi_2 \\ \xi_1 & - & \\ \xi_2 & \phi_{21} & - \end{pmatrix}$$ Matriz de covarianzas 2 x 2 de variables latentes exógenas.

- Las varianzas y covarianzas de los términos de perturbación de Ψ (matriz de covarianzas $m\ x\ m$ de los términos de perturbación).

  Para el modelo del gráfico 3.4, esta matriz es:

$$\Psi = \begin{pmatrix} & \zeta_1 & \zeta_2 & \zeta_3 \\ \zeta_1 & - & & \\ \zeta_2 & & - & \\ \zeta_3 & \psi_{31} & & - \end{pmatrix}$$ Matriz de covarianzas 3 x 3 de los términos de perturbación.

- Los coeficientes de regresión de B (matriz $m\ x\ m$ de coeficientes de las variables endógenas) y Γ (matriz $m\ x\ k$ de coeficientes de las variables exógenas).

  Para el modelo del gráfico 3.4, estas matrices son:

$$\mathrm{B} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix}$$ Matriz 3 x 3 de coeficientes de las variables endógenas.

$$\Gamma = \begin{pmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \\ 0 & 0 \end{pmatrix}$$ Matriz 3 x 2 de coeficientes de las variables exógenas.

Además se debe considerar al vector de parámetros, denotado como Ω, que contiene los parámetros antes descritos juntos: $\Omega = (\Lambda_y, \Lambda_x, \Theta_\varepsilon, \Theta_\delta, \Phi, \mathrm{B}, \Gamma, \Psi)$.

El objetivo es obtener las estimaciones del vector de parámetros Ω, denotado como $\hat{\Omega}$, que minimice la función de discrepancia $F(S,\hat{\Sigma})$, donde $\hat{\Sigma}=\Sigma(\hat{\Omega})$ es la matriz de varianza-covarianza de las estimaciones del modelo llamada matriz de varianza-covarianza ajustada. La función $F(S,\hat{\Sigma})$ es un escalar que mide la discrepancia (distancia) entre la matriz de varianza-covarianza muestral S y la matriz de varianza-covarianza ajustada $\hat{\Sigma}$ y puede caracterizarse por las siguientes propiedades:

1. $F(S,\hat{\Sigma})\geq 0$,
2. $F(S,\hat{\Sigma})=0$, sí y solo sí $\hat{\Sigma}=S$,
3. $F(S,\hat{\Sigma})$ es una función contínua en S y $\hat{\Sigma}$.

Esta función de ajuste vendrá dada por la siguiente expresión:

$$F=[S-\Sigma(\Omega)]W^{-1}[S-\Sigma(\Omega)]$$

donde:

S: es la matriz de varianzas-covarianzas de la muestra,

$\Sigma(\Omega)$: es la matriz de varianzas-covarianzas predichas por el modelo,

W: es una matriz de ponderaciones que puede tomar diversas formas dependiendo de la distribución que tengan las variables observadas.

Si se asume que la distribución muestral de dichas variables es normal multivariante, entonces la función de ajuste tomará la siguiente forma:

$$F_{NORMAL}=2^{-1}Traza[(S-\Sigma(\Omega))W_2]^2$$

donde:

S: es la matriz de varianzas-covarianzas de la muestra,

$\Sigma(\Omega)$: es la matriz de varianzas-covarianzas predicha por el modelo,

W_2: es una matriz que puede tomar diversas formas en función del tipo de método de estimación que se escoja:

$W_2 = \Sigma^{-1}$: Máxima verosimilitud,

$W_2 = S^{-1}$: Mínimos cuadrados generalizados,

$W_2 = I$: Mínimos cuadrados no ponderados.[33]

3.4.1. MÉTODO DE MÁXIMA VEROSIMILITUD

El método de máxima verosimilitud fue propuesto originalmente por Koopmans, Rubin y Leipnik (1950) como un método de estimación para modelos econométricos de ecuaciones simultáneas. Luego, Jöreskog (1973) desarrolló la estimación de máxima verosimilitud para modelos de ecuaciones estructurales.

Sea m el número de variables endógenas y k el número de variables exógenas de un modelo de ecuaciones estructurales, entonces se puede representar un modelo de ecuaciones estructurales completo como:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta$$

$$y = \Lambda_y \eta + \varepsilon$$

$$x = \Lambda_x \xi + \delta$$

Se considera, inicialmente, al vector z de respuestas observadas (que contiene a x e y) basado en una muestra de $n=N-1$ observaciones con la correspondiente matriz de varianzas-covarianzas de la muestra S, que estima la matriz de varianzas-covarianzas de la población Σ. La estimación con el método de máxima verosimilitud supone que las observaciones se derivan de una población que sigue una distribución normal multivariante. La función de densidad normal de z puede denotarse:

---

[33] (12) LUQUE, "Técnicas de Análisis de Datos en Investigación de Mercados", pág. 504-506.

$$f(z) = (2\pi)^{-(m+k)/2} |\Sigma|^{1/2} \exp\left[-\frac{1}{2} z \Sigma^{-1} z \right]. \qquad (3.6)$$

Esta función de densidad normal, describe la distribución para cada observación en la muestra. Asumiendo que las N observaciones son independientes entre sí, la función de densidad conjunta puede escribirse como el producto de las densidades individuales:

$$f(z_1, z_2, \cdots, z_N) = f(z_1) f(z_2) \cdots f(z_N). \qquad (3.7)$$

Si la ecuación (3.6) representa la función de densidad normal para una muestra simple, entonces el producto dado en la ecuación (3.7) puede escribirse como:

$$L(\Omega) = (2\pi)^{-N(m+k)/2} |\Sigma(\Omega)|^{-N/2} \exp\left[-\frac{1}{2} \sum_{i=1}^{N} z_i \Sigma^{-1}(\Omega) z_i \right], \qquad (3.8)$$

donde $L(\Omega)$ se define como la *verosimilitud* de la muestra.

Con el fin de simplificar y sin perder generalidad, es preferible tomar el logaritmo de la ecuación (3.8), resultando el log de verosimilitud:

$$\log L(\Omega) = \frac{-N(m+k)}{2} \log(2\pi) - \left(\frac{N}{2} \right) \log|\Sigma(\Omega)| - \left(\frac{N}{2} \right) tr\left[\mathrm{T} \Sigma^{-1}(\Omega) \right]. \qquad (3.9)$$

El último término de la ecuación (3.9) se debe a que el término en el corchete de la ecuación (3.8) es un escalar, y la traza de un escalar es otro escalar. Por lo tanto, haciendo referencia al último término de la ecuación (3.8) se tiene

$$-\left(\frac{1}{2} \right) \sum_{i=1}^{N} z_i \Sigma^{-1}(\Omega) z_i = -\left(\frac{1}{2} \right) \sum_{i=1}^{N} tr\left[z_i \Sigma(\Omega) z_i \right].$$

Multiplicando y dividiendo por N, y sabiendo que $tr(ABC) = tr(CAB)$:

$$-\left(\frac{1}{2} \right) \sum_{i=1}^{N} tr\left[z_i \Sigma(\Omega) z_i \right] = -\left(\frac{N}{2} \right) \sum_{i=1}^{N} tr\left[N^{-1} z_i z_i \Sigma^{-1}(\Omega) \right] = -\left(\frac{N}{2} \right) tr\left[\mathrm{T} \Sigma^{-1}(\Omega) \right],$$

donde T es la matriz de covarianzas muestrales basada en N más que en $n=N-1$.

El siguiente paso es maximizar la ecuación (3.9) con respecto a los parámetros del modelo. Para maximizar el log de verosimilitud en la ecuación (3.9) se obtiene primeramente las derivadas con respecto a los parámetros del modelo, se iguala las derivadas a cero y se resuelve.

La ecuación (3.9) contiene el término constante: $((-N(m+k))/2)\log(2\pi)$, el cual no tiene parámetros del modelo, por lo tanto, se lo puede ignorar ya que no tendrá consecuencias cuando se obtengan las derivadas. Nótese además, que la diferencia entre T (basada en N) y la estimación insesgada usual S (basada en $n=N-1$) es insignificante en muestras grandes. Entonces se puede rescribir la ecuación (3.9) así:

$$\log L(\Omega) = -\frac{N}{2}\left\{\log|\Sigma(\Omega)| + tr\left[S\Sigma^{-1}(\Omega)\right]\right\} \qquad (3.10)$$

El inconveniente que se tiene con la ecuación (3.10) es que no tiene la propiedad de la función de discrepancia como se describió anteriormente. Por ejemplo, si $S=\Sigma$ entonces el segundo término de la derecha de la ecuación (3.10) será una matriz de identidad de orden $m+k$ y la traza será igual a $m+k$. Sin embargo, si la ecuación (3.10) tiene la propiedad de la función de discrepancia, la diferencia entre el primer término y el segundo término no será igual a cero como se requeriría en este caso. Para dar a la ecuación (3.10) la propiedad de la función de discrepancia, se requiere aumentar términos que no dependan en el modelo de parámetros y que por lo tanto no se involucren en la diferenciación. Inicialmente, se puede quitar el término $-(N/2)$, de este modo se estaría minimizando la función antes que maximizándola. Luego, se aumentan términos que no dependan del modelo de parámetros. Así:

$$F_{MV} = \log|\Sigma(\Omega)| + tr\left[S\Sigma^{-1}(\Omega)\right] - \log|S| - t \qquad (3.11)$$

donde t es el número total de variables en x e y, es decir, $t=m+k$. Puede verse que si el modelo se ajusta perfectamente, el primer y tercer términos suman cero, y el segundo y cuarto términos suman cero, consecuentemente la ecuación (3.11) es una función ajustada.

Se puede también obtener la matriz de covarianza de las estimaciones, con el fin de obtener las estimaciones del modelo de parámetros. Sea Ω el vector *rx1* del modelo de parámetros estimado, entonces la matriz de covarianzas asintótica de Ω puede escribirse como

$$a\operatorname{cov}(\Omega) = \left\{ - E\left[\frac{\partial^2 \log L(\Omega)}{\partial \Omega \partial \Omega'} \right] \right\}^{-1}$$

donde, la expresión dentro de los corchetes se llama matriz de información de Fisher, denotada como $\mathrm{I}(\Omega)$. A partir de aquí, se pueden obtener los errores estándar, sacando la raíz cuadrada a los elementos de la diagonal de la matriz de covarianzas asintótica de las estimaciones.[34]

Los estimadores de máxima verosimilitud son no sesgados, es decir, si se extrajera un número infinito de muestras de cien o más casos y se calculara cada vez el valor de estos estimadores, el valor medio de los mismos sería el correspondiente a la población total.

Además, para muestras lo suficientemente grandes, el método de estimación por máxima verosimilitud proporciona estimadores eficientes; es decir, que si una vez obtenidas todas esas muestras se calcula la desviación típica de esos valores, se obtiene un valor mínimo comparado con el que se obtendría con otros métodos.

El inconveniente que tiene la función de máxima verosimilitud es su sensibilidad al tamaño muestral, de forma que a medida que la muestra va aumentando, dicha función se va haciendo más sensible para detectar diferencias entre los datos.

El tamaño de la muestra ideal para aplicar esta técnica está entre 100 y 200. Cuando se tiene una muestra demasiado amplia (más de 400), incluso modelos que se ajustan bien a los datos van a presentar diferencias significativas entre la matriz de datos originales y la matriz estimada.

---

[34] (8) KAPLAN, “Structural Equation Modeling: Fundations and Extensions”, pág. 25-27.

3.4.2. MÉTODO DE MÍNIMOS CUADRADOS GENERALIZADOS

De igual manera que el método de estimación de máxima verosimilitud, este método proporciona estimadores no sesgados y eficientes de los parámetros del modelo, aunque con muestras pequeñas dichos estimadores tiene un sesgo próximo a cero.

Nuevamente, se asume que los datos tienen una distribución normal multivariante. La forma general para la función de ajuste de los mínimos cuadrados generalizados (MCG) se escribe así

$$F_{MCG} = [S - \Sigma(\Omega)]W^{-1}[S - \Sigma(\Omega)],$$

donde W^{-1} es una matriz de ponderaciones que pondera las desviaciones $S - \Sigma(\Omega)$ en términos de sus varianzas y covarianzas con otros elementos.

El método de estimación de mínimos cuadrados ponderados (MCP) se lo considera dentro del estimador de MCG. En la estimación de los MCP es importante escoger la matriz de ponderaciones W^{-1}. Aquí, se considerará dos opciones para W. La primera opción es $W^{-1} = \mathrm{I}$, la matriz identidad. Utilizando esta opción, los MCP se reducen a mínimos cuadrados no ponderados (MCNP). Los MCNP operan de manera similar a los mínimos cuadrados ordinarios, en cuanto se ignora el potencial para los errores de heteroelasticidad. Además, aunque los MCNP se los conoce para producir las estimaciones insesgadas del modelo de parámetros, los MCNP no es la opción más eficiente de estimación.

Para corregir el problema que se tiene con los MCNP, se puede elegir $W^{-1} = S^{-1}$. Esta elección es la más común para W^{-1}. Con $W^{-1} = S^{-1}$, la función de ajuste de las MCG se puede escribir:

$$F_{MCG} = \frac{1}{2} tr\left[S^{-1}(S - \Sigma)\right]^2 = \frac{1}{2} tr\left(\mathrm{I} - S^{-1}\Sigma\right)^2.$$

Bajo la suposición de normalidad multivariante, la función de ajuste de los MCG tiene las mismas propiedades asintóticas de MV, es decir, el estimador de MCG es asintóticamente normal y asintóticamente eficiente.

3.4.3. MÉTODO DE MÍNIMOS CUADRADOS NO PONDERADOS

Ésta es otra alternativa que existe para la función de discrepancia, a pesar de ser muy poco común su uso. El procedimiento de estimación por mínimos cuadrados no ponderados (MCNP) presenta la siguiente función de ajuste:

$$F_{MCNP} = \frac{1}{2} tr\left[(S - \Sigma)^2\right]$$

En este caso se minimiza la suma de cuadrados de cada elemento en la matriz de residuos (S - Σ), ponderando implícitamente todos los elementos de dicha matriz como si tuvieran las mismas varianzas y covarianzas con otros elementos, debido a que $W_2 = I$. Esto difiere con las F_{MV} y F_{MCG} en que éstas ponderan los elementos de la matriz residual de acuerdo a sus varianzas y covarianzas con otros elementos, ya sea utilizando Σ^{-1} o su estimador consistente S^{-1}. Por lo tanto, el nivel de exigencia en cuando a la multinormalidad de la distribución de la muestra es bastante menor que para las F_{MV} y F_{MCG}, lo que implica que las estimaciones obtenidas de esta forma sean poco eficientes.

Por otro lado, cabe resaltar que este procedimiento de estimación es dependiente de la escala de medida, difiriendo sus valores de función del tipo de matriz de entrada (correlaciones o varianzas-covarianzas). No obstante, esta función de ajuste tiene la ventaja de su facilidad de cálculo y comprensión intuitiva.

Cuando los datos de la muestra no sigan una distribución normal multivariante, la teoría y los estudios de simulación llevados a cabo han mostrado que se produce un sesgo estadístico χ^2 (MV o MCG) y en los errores estándar de las estimaciones de los parámetros, aunque no parece afectar a las propias estimaciones de los parámetros que son altamente consistentes.

Por lo tanto, antes de escoger el método de estimación se debe analizar la distribución muestral de las variables observadas, estudiando las características de la multinormalidad, asimetría y curtosis a través de una serie de tests y de coeficientes disponibles para tal efecto, como el coeficiente de normalidad de Mardia (1970) o los tests de asimetría y curtosis multivariante que proporcionan algunos programas estadísticos.

3.4.4. MÉTODO DE MÍNIMOS CUADRADOS PONDERADOS

La función de ajuste según el método de estimación por mínimos cuadrados ponderados (MCP) será:

$$F_{MCP} = [S - \Sigma(\Omega)]W^{-1}[S - \Sigma(\Omega)]$$

De esta manera, se van a obtener las estimaciones de los parámetros minimizando la suma ponderada de las diferencias entre las varianzas-covarianzas de las variables observadas y las varianzas-covarianzas predichas por el modelo, o lo que es igual, la función de MCP se expresa como la suma ponderada de los residuos al cuadrado, siendo la matriz de ponderación W la matriz de covarianzas de los residuos.

Esta expresión es igual a la función de discrepancia general que se describe al inicio de este capítulo, que bajo los supuestos de multinormalidad se derivaba en las funciones antes descritas F_{MV}, F_{MCG} y F_{MCNP}, y que no son más que casos específicos de la propia F_{MCP}.

Dicha función supone una serie de ventajas e inconvenientes. Entre las ventajas, se puede destacar que son mínimas las suposiciones sobre la distribución muestral de las variables observadas, y de ahí que también se denominen como funciones de discrepancia asintóticamente libres de distribución, y que proporciona estimaciones eficientes de los parámetros.

El principal inconveniente de la F_{MCP} es que al tener una forma de distribución asintóticamente libre requiere que se invierta la matriz W, lo que se va complicando a medida que el número de variables observadas incrementa (por ejemplo, en el caso de que haya doce variables observadas, la matriz a invertir sería de orden 78 x 78). Además, otro importante problema es que esta aproximación requiere que el tamaño muestral sea lo suficientemente amplio como para que la función de ajuste pueda converger y dar una solución óptima, de manera que si la muestra es demasiado pequeña no se podrá llevar a cabo dicho método de estimación.

Por estas razones, es importante considerar lo siguiente: si las desviaciones que se producen de la normalidad no son muy importantes, es preferible emplear procedimientos de estimación más simples tales como MV, MCG, o incluso MCNP. La estrategia más prudente será comparar los resultados de estos métodos con los obtenidos para MCP, de manera que se extraigan conclusiones acordes con los planteamientos teóricos previamente establecidos.[35]

3.4.5. COMPROBACIÓN DE LOS PARÁMETROS

Una característica de la estimación de MV y MCG es que se puede probar la hipótesis de que el modelo se ajusta a los datos. Considerando nuevamente la ecuación (3.10):

$$\log L(\Omega) = -\frac{N}{2}\left\{\log|\Sigma(\Omega)| + tr\left[S\Sigma^{-1}(\Omega)\right]\right\}$$

que es el log de verosimilitud bajo la hipótesis nula de que el modelo especificado se ajusta a la población. La correspondiente hipótesis alternativa es que Σ es cualquier matriz simétrica definida positiva. Bajo esta hipótesis alternativa, el log de verosimilitud alcanza su máximo con S como estimador de Σ. Por lo tanto, el log de verosimilitud bajo la hipótesis alternativa, denotado como L_a, se denota:

$$\log L_a = -\frac{n}{2}\log|S| + tr\left(SS^{-1}\right)$$

$$= -\frac{n}{2}\log|S| + tr(\mathrm{I})$$

$$= -\frac{n}{2}\log|S| + t$$

El estadístico para probar la hipótesis nula que el modelo se ajusta a la población se llama *prueba de la razón de verosimilitud* (RV) y se la puede expresar como:

---

[35] (12) LUQUE, "Técnicas de Análisis de Datos en Investigación de Mercados", pág. 507-510.

$$-2\log\frac{L_0}{L_a} = -2\log L_0 + 2\log L_a$$

$$= n\left[\log|\Sigma| + tr\left(\Sigma^{-1}S\right)\right] - n\left(\log|S| + t\right)$$

$$= n\left[\log|\Sigma| + tr\left(\Sigma^{-1}S\right) - \log|S| - t\right].$$

Como se puede ver, el último término de la igualdad de la ecuación anterior es igual a $n*F_{MV}$.

La distribución para una muestra grande de la prueba de la razón de verosimilitud es una ji-cuadrado con grados de libertad dados por la diferencia entre el número de elementos no redundantes en Σ y el número de parámetros libres en el modelo. La prueba de la razón de verosimilitud ji-cuadrado se usa para probar la hipótesis nula de que la matriz de covarianza de la población posee la estructura implicada por el modelo, en contraste con la hipótesis alternativa de que Σ es una matriz simétrica definida positiva, arbitraria.

En los modelos de ecuaciones estructurales, la prueba de ajuste del modelo, basada en las estadísticas de la razón de verosimilitud ji-cuadrado, tiene ahora muchos más grados de libertad que lo usual. De manera general, es posible particionar el número total de grados de libertad en aquellos basados en restricciones de la parte medible del modelo y en aquellos basados en restricciones de la parte estructural del modelo. Usualmente, los grados de libertad de la parte medible del modelo son mayores que aquellos de la parte estructural del modelo. Luego, se compara la parte estructural del modelo, que usualmente es el enfoque principal, con la parte medible del modelo, que por lo general sirve para proporcionar estimados insesgados de los parámetros del modelo estructural; de este modo, es posible desechar un modelo estructural relativamente bien ajustado, debido a un modelo de medición poco desarrollado.

3.4.6. EVALUACIÓN E INTERPRETACIÓN DEL MODELO

Una vez que el modelo ha sido identificado y estimado, el siguiente paso consistirá en evaluar lo bien que nuestros datos se han ajustado al modelo

propuesto. Esta evaluación debe realizarse a tres niveles: evaluación del ajuste del modelo global, evaluación del ajuste el modelo de medida y evaluación del ajuste del modelo estructural.

- **Ajuste global del modelo**

Existirá un ajuste perfecto cuando haya una correspondencia perfecta entre la matriz reproducida por el modelo y la matriz de observaciones. La evaluación del ajuste de un modelo de ecuaciones estructurales no es algo sencillo y único, habiéndose desarrollado multitud de medidas que en conjunto permiten analizar su bondad y adecuación. Existen tres tipos de medidas de ajuste global: medidas absolutas de ajuste, medidas incrementales de ajuste y medidas de ajuste de parsimonia.

a. Las medidas absolutas de ajuste determinan el grado en que el modelo globalmente (modelo de medida y modelo estructural) predice la matriz de datos inicial. Las principales medidas absolutas de ajuste empleadas son las siguientes:

Estadístico ji-cuadrado, es una de las medidas de bondad de ajuste más comunes y utilizadas. Se trata de una prueba estadística (basado en la distribución χ^2) que mide la distancia existente entre la matriz de datos inicial y la matriz estimada por el modelo.

Si se cumplen todas las suposiciones necesarias para poder aplicar la prueba (distribución muestral multinormal) y el tamaño muestral es lo suficientemente amplio, la prueba funciona de la siguiente manera:

$$H_0: S = \Sigma$$

$$H_1: S \neq \Sigma$$

La hipótesis nula es que la matriz de observaciones (S) y la matriz estimada (Σ) son iguales, mientras que la hipótesis alternativa es que son diferentes.

Así, valores altos del estadístico χ^2 llevarán a rechazar la hipótesis nula y valores bajos a no rechazarla. Con esta prueba se trata que no existan

diferencias significativas entre ambas matrices, por lo que para no rechazar la hipótesis nula el nivel de significación debe ser superior a 0.05 o 0.01 dependiendo de la exigencia que se haya propuesto a la prueba.

Un importante inconveniente, es que para muestras suficientemente amplias (más de 400) se incrementa la probabilidad de rechazar el modelo aunque las diferencias entre las matrices sea mínima y, por otro lado, cuando el tamaño muestral es pequeño (menos de 100) la prueba mostrará un ajuste aceptable, aunque en realidad existan importantes diferencias entre dichas matrices.

Estadístico ji-cuadrado no centrado (Noncentrality Parameter, NCP), es igual al estadístico χ^2 corregido por los grados de libertad; estando de esta forma menos afectado por el tamaño muestral:

$$\text{NCP} = \chi^2 - \text{grados de libertad}$$

El número de grados de libertad (gl) para un modelo propuesto se calcula de la siguiente forma:

$$gl = \frac{1}{2}[(k+t)(k+t+1)] - p$$

Donde: k: número de indicadores exógenos

t: número de indicadores endógenos

p: número de parámetros estimados en el modelo

Al igual que para la ji-cuadrado, se consideran aceptables los valores que son lo más próximo posible a 0.

Raíz Cuadrada del Error Medio Cuadrático* (*Root Mean Square Error of Approximation, RMSEA), es una medida de ajuste introducida por Steiger (1990) para intentar eliminar el inconveniente que presentaba la χ^2 cuando la muestra era lo suficientemente grande. Aquí, la discrepancia entre la matriz de observaciones inicial y la matriz estimada por el modelo, está

medida en términos de la población y no en términos de la muestra. Así, describe la diferencia de las matrices por grado de libertad, es decir, la bondad de ajuste que debería ser esperada si el modelo fuera estimado en la población y no en la muestra:

$$RMSEA = \left[\frac{Máx\left\{\chi^2 - (gl/N-1), 0\right\}}{gl} \right]^{-1/2}$$

Valores del RMSEA inferiores a 0.05, e incluso a 0.08, son indicativos de un buen ajuste del modelo en la población.

Además de la estimación puntual que ciertos programas como el AMOS o LISREL realizan del RMSEA, también proporcionan una estimación por intervalo y una prueba estadística para RMSEA≤0.05 (denominado **PCLOSE** en AMOS, que proporciona una prueba de ajuste exacto), que resultan muy útiles para el proceso de evaluación.

Índice de Bondad de Ajuste (Goodness of Fit Index, GFI), es un índice de la variabilidad explicada por el modelo, oscilando sus valores entre 0 (pobre ajuste) y 1 (perfecto ajuste). Es más independiente del tamaño de la muestra y menos sensible que la χ^2 a las desviaciones de la normalidad. Es análogo en interpretación al R^2 en la regresión múltiple y no existe ningún límite a partir del cual se pueda afirmar que el ajuste es bueno, si bien valores superiores a 0.90 y a 0.95 indicarían un ajuste aceptable. En el procedimiento de ajuste MV, el GFI viene definido por la siguiente expresión:

$$GFI = 1 - \frac{tr\left(\Sigma^{-1}S - I\right)^2}{tr\left(\Sigma^{-1}S\right)^2}$$

donde I es la matriz identidad.

El GFI está afectado por el tamaño de la muestra y por el número de indicadores, por lo que a veces puede resultar equívoco un determinado valor de este indicador. Para solventar este inconveniente, puede usarse el

Índice de Bondad de Ajuste Relativo (Relative Goodness of Fit Index, RGFI) que viene dado por el cociente entre el GFI estimado por el modelo y el GFI esperado en función del tamaño muestral y del número de indicadores que posea el modelo (EGFI):

$$EGFI = \frac{1}{1 + [2gl/(k+t)n]}$$

El EGFI desciende a medida que incrementa el número de indicadores y aumenta con el tamaño muestral:

$$RGFI = \frac{GFI}{EGFI}$$

De este modo, se puede evaluar la bondad de ajuste del modelo a través de RGFI, obteniendo así, una medida relativa de bondad, que tiene en cuenta el tamaño muestral y el número de indicadores, considerándose en la práctica adecuados aquellos modelos que tengan un RGFI superior a 0.90.

b. Las medidas incrementales de ajuste comparan el modelo propuesto con un modelo nulo o básico que se toma de referencia y que, tradicionalmente, suele ser aquel que estipula una falta absoluta de asociación entre las variables del modelo; se trata, por lo tanto, de comparar nuestro modelo con el peor modelo posible. Dentro de estos índices incrementales se puede destacar los siguientes:

Índice de Bondad de Ajuste Ajustado (Adjusted Goodness of Fit Index AGFI), es otra de las medidas tradicionales que junto con la χ^2, el GFI, y otras medidas absolutas de ajuste, se han utilizado para evaluar la bondad de ajuste en los modelos de ecuaciones estructural. No existen límites exactos a partir de los cuales poder afirmar la idoneidad de un modelo, en la experiencia práctica se considera que valores superiores a 0.90 son indicativos de un buen ajuste del modelo a los datos:

$$AGFI = 1 - \left[\frac{(k+t)(k+t+1)}{2gl} \right] (1 - GFI)$$

Al igual que el GFI, este índice está afectado por el tamaño muestral y por el número de indicadores, por lo que será más adecuado obtener el valor relativo del AGFI (RAGFI). De igual forma que para el caso del GFI, habrá que calcular el valor esperado del AGFI (EAGFI), que se obtendrá sustituyendo en la anterior fórmula el GFI por el EGFI. Seguidamente, el RAGFI se calculará dividiendo el AGFI entre el EAGFI. Este valor nos proporcionará una base más adecuada para valorar la bondad de ajuste del modelo, eliminando de esta forma el efecto del tamaño muestral y el número de indicadores. El valor límite para la aceptación del modelo suele establecerse en 0.80, si bien ha de tenerse en cuenta que estos límites son arbitrarios y que deben servir sólo como orientación, ya que lo adecuado es que sean utilizados comparando determinados modelos alternativos ajustados a un mismo conjunto de datos.

Índice de Ajuste Normado (Normed Fit Index, NFI), es otra medida incremental de ajuste que compara el modelo propuesto y el nulo. En realidad mide la reducción proporcional en función de ajuste cuando se pasa del modelo nulo al propuesto. El rango de variación de este índice también está entre 0 y 1, siendo recomendable valores superiores a 0.90:

$$NFI = \frac{\left(\chi^2_{\text{Modelo Nulo}} - \chi^2_{\text{Modelo Propuesto}}\right)}{\chi^2_{\text{Modelo Nulo}}}$$

Presenta algunos problemas: primero, al no tener en cuenta los grados de libertad, el valor de la ji-cuadrado del modelo propuesto se reduce añadiendo más parámetros, con lo que el índice aumenta no por el hecho de un buen ajuste a los datos, sino porque disminuye el número de grados de libertad; segundo, el valor del NFI varía en función del tamaño muestral, es decir, es más grande a medida que la muestra aumenta. Estos inconvenientes hace que sea inapropiado para comparar modelos alternativos con diferente número de parámetros y de tamaño muestral.

Índice de Ajuste No Normado (Nonnormed Fit Index, NNFI), este índice, también denominado Índice de Tucker y Lewis (*Tucker-Lewis Index,* ***TLI***), compara el ajuste por grado de libertad del modelo propuesto y nulo. Viene a resolver los inconvenientes que presentaba el NFI, ya que por una parte, al considerar los grados de libertad de los modelos, se elimina el problema del sobreajuste como consecuencia del número de parámetros y, por otra, estudios de simulación han hallado que este índice está muy débilmente relacionado con el tamaño muestral. El NNFI tiende a 1 para modelos con muy buen ajuste, considerándose aceptables valores superiores a 0.90:

$$NNFI = \frac{\left(\chi^2_{\text{Modelo Nulo}} / gl_{\text{Modelo Nulo}}\right) - \left(\chi^2_{\text{Modelo Propuesto}} / gl_{\text{Modelo Propuesto}}\right)}{\left(\chi^2_{\text{Modelo Nulo}} / gl_{\text{Modelo Nulo}}\right) - 1}$$

Índice de Ajuste Incremental (Incremental Fit Index, IFI)*,* es otro índice que elimina en parte los inconvenientes del NFI, propuesto por Bollen (1988):

$$IFI = \frac{\chi^2_{\text{Modelo Nulo}} - \chi^2_{\text{Modelo Propuesto}}}{\chi^2_{\text{Modelo Nulo}} - gl_{\text{Modelo Propuesto}}}$$

En igualdad de condiciones, el IFI es mayor para muestras pequeñas que para muestras grandes, lo que contrarresta la tendencia al azar del NFI para muestras grandes. La introducción en el denominador de los grados de libertad del modelo propuesto implica que si se tiene dos modelos con los mismos valores para la χ^2 del modelo nulo y propuesto, el que tenga menos parámetros presentará un valor más alto para el IFI, siendo, por tanto, el más adecuado. Se consideran aceptables valores próximos a la unidad, si bien su valor puede ser mayor que 1 en determinadas ocasiones.

Índice de Ajuste Relativo (Relative Fit Index, RFI), también fue introducido por Bollen (1986), siendo muy similar al NNFI con la única diferencia de que en el denominador no se le resta la unidad al cociente entre la χ^2 y los grados de libertad para el modelo nulo:

$$RFI = \frac{\left(\chi^2_{\text{Modelo Nulo}} / gl_{\text{Modelo Nulo}}\right) - \left(\chi^2_{\text{Modelo Propuesto}} / gl_{\text{Modelo Propuesto}}\right)}{\left(\chi^2_{\text{Modelo Nulo}} / gl_{\text{Modelo Nulo}}\right)}$$

De esta manera, se premia a los modelos con más parsimonia o más simples, si bien también depende del tamaño muestral.

Este índice proporciona valores próximos a la unidad a medida que el modelo va alcanzando un buen ajuste.

Índice de Ajuste Comparativo (Comparative Fit Index, CFI), introducido por Bentler (1990), indica un buen ajuste del modelo para valores próximos a 1.

$$CFI = 1 - \frac{Máx\left[\left(\chi^2_{\text{Modelo Propuesto}} - gl_{\text{ModeloPropuesto}}\right), 0\right]}{Máx\left[\left(\chi^2_{\text{Modelo Nulo}} - gl_{\text{Modelo Nulo}}\right), \left(\chi^2_{\text{Modelo Propuesto}} - gl_{\text{Modelo Propuesto}}\right), 0\right]}$$

c. Las medidas de ajuste de parsimonia. La parsimonia de un modelo es el grado en que alcanza ajuste para cada coeficiente o parámetros estimado, de tal manera que estas medidas relacionan la bondad del modelo con el número de coeficientes estimados requeridos para alcanzar ese nivel de ajuste. Lo que se pretende es obtener una medida del nivel de ajuste por coeficiente estimado, evitando el sobreajuste del modelo con coeficientes innecesarios.

   Al igual que la mayoría de índices, no se dispone de ningún test estadístico asociado a ellos, por que su uso es más adecuado comparando modelos alternativos. Dentro de estas medidas se pueden destacar las siguientes:

   Índice de Ajuste Normado Parsimonioso (Parsimonious Normed Fit Index, PNFI), es similar al NFI, pero teniendo en cuenta el número de grados de libertad usados para alcanzar el nivel de ajuste. Como el nivel de parsimonia ideal sería 1 grado de libertad por coeficiente estimado, lo que interesa es conseguir altos valores de parsimonia, es decir, altos valores de este índice. Cuando se comparan modelos alternativos, diferencias en sus valores del PNFI entre 0.06 y 0.09 resultan importantes:

$$PNFI = \left(\frac{gl_{\text{Modelo Propuesto}}}{gl_{\text{Modelo Nulo}}} \right) * NFI$$

Índice de Bondad de Ajuste Parsimonioso (Parsimonious Goodness of Fit Index, PGFI), consiste en el ajuste del GFI de manera similar al AGFI, pero basado en la parsimonia del modelo estimado. Nuevamente, son preferibles valores altos de este índice:

$$PGFI = \frac{gl_{\text{Modelo Propuesto}}}{(k+t)(k+t+1)/2}$$

Ji-cuadrado Normado, esta medida propuesta en 1969 por Jöreskog, consiste en el valor del estadístico ji-cuadrado dividido por los grados de libertad. Presenta el mismo inconveniente que se comenta anteriormente para ji-cuadrado, el ser muy sensible al tamaño muestral, sin embargo, al considerar los grados de libertad permite evaluar aquellos modelos *sobreajustados* (valores inferiores a la unidad) y aquellos que no presentan un ajuste suficiente a los datos (valores superiores a 2.3 o incluso 5).

Criterio de Información de Akaike (Akaike Information Criterion, AIC), sirve para comparar modelos que poseen diferente número de variables latentes. Cuando se obtienen valores pequeños de la χ^2 con pocos parámetros, esta medida será muy pequeña (alrededor de 0) indicando una alta parsimonia:

$$AIC = \chi^2 + 2p$$

Una transformación del AIC fue propuesta por Bozdogan (1987), teniendo prácticamente las mismas implicaciones:

$$CAIC = \chi^2 + [1 + \ln(N)]p$$

siendo N el tamaño muestral.

N Crítico (Critical N, CN), el CN (Hoetler, 1983) sugiere el tamaño que una muestra debe alcanzar en orden a aceptar el ajuste de un modelo dado

sobre una base estadística. Se recomienda valores de al menos 200 para este índice, ya que el valor de 200 es un razonable punto de inicio sugiriendo que las diferencias entre la matriz de covarianzas de la población y la matriz de covarianzas observada son triviales:

$$CN = \frac{\chi^2_{\text{Percentil}(1-\alpha)}}{\chi^2_{\text{Modelo Propuesto}}} + 1$$

- **Ajuste del modelo de medida**

Los índices descritos anteriormente, sirven para evaluar el ajuste global de un modelo de ecuaciones estructurales al considerar conjuntamente el modelo de medida y el modelo estructural. Si sólo se analizan dichos índices, puede ocurrir que se obtenga una medida de ajuste global con unos límites aceptables pero con algunos de los parámetros estimados no significativos. Por tal motivo, se debe revisar por separado tanto el ajuste del modelo de medida como el del modelo estructural.

Para realizar el ajuste del modelo de medida, el paso inicial consiste en examinar la significación estadística de cada carga obtenida entre el indicador y la variable latente. Una carga no significativa (valor t inferior a 1.96 para α=0.05 si el investigador no ha especificado el signo de la relación, prueba de 2 colas, o valor t inferior a 1.645 si el investigador ha impuesto un signo concreto al parámetro a estimar, prueba de 1 cola), indica que ese valor es estadísticamente igual a 0, lo que supone que el indicador no explica nada de la variable latente. Ante esto, se debe eliminar o transformar dicho indicador.

Una vez comprobada la significación de las cargas, el siguiente paso es comprobar la fiabilidad de cada uno de los indicadores así como la fiabilidad compuesta del constructo. La varianza total de un indicador puede ser descompuesta en dos partes: la que tiene en común con la variable latente a la que mide y la que se debe al error. Por lo tanto, dicha fiabilidad, proporcionada para cada indicador por el programa computacional que se escoja (LISREL, AMOS,…), será la proporción de varianza que tiene en común con el constructo y equivalente a la comunalidad en el análisis factorial

exploratorio. Se considera que un indicador debería tener al menos un 50% de su varianza en común con la variable latente, estableciendo así como límite de aceptación para la fiabilidad el valor del 0.50.

Todos los indicadores deberán tener una alta consistencia interna, es decir, han de ser una medida válida del concepto a estudiar. Dicha consistencia interna va a ser medida a través de la fiabilidad compuesta del constructo, dada por la siguiente expresión:

$$\text{Fiabilidad} = \frac{(\Sigma\text{Cargas estandarizadas})^2}{(\Sigma\text{Cargas estandarizadas})^2 + (\Sigma\text{Errores de medida})}$$

El límite que se considera aceptable para esta medida de la fiabilidad compuesta es de 0.70, si bien no es un límite rígido, ya que depende del tipo de estudio que se lleve a cabo. Así, en estudios de carácter exploratorio incluso valores por debajo de dicho límite se consideran adecuados.

Otra medida que normalmente se utiliza para evaluar el ajuste del modelo de medida es la *varianza extraída*. Indica la cantidad global de varianza en los indicadores explicada por la variable latente. Si esta valor es alto (superior a 0.50), se considera que los indicadores miden adecuadamente dicha variable latente. Se trata de una medida complementaria a la de la fiabilidad compuesta y su cálculo viene dado por la expresión siguiente:

$$\text{Varianza extraída} = \frac{\Sigma\text{Cargas estandarizadas}^2}{\Sigma\text{Cargas estandarizadas}^2 + \Sigma\text{Errores de medida}}$$

- Ajuste del modelo estructural

Así como se ha examinado detenidamente el modelo de medida, se debe hacer los mismo con el modelo estructural estimado, independientemente de que las medidas de ajuste global indiquen unos valores aceptables. Lo primero a analizar en un modelo estructural es la significación alcanzada por los coeficientes estimados. Así, cualquier parámetro estimado debe ser estadísticamente diferente de cero, o lo que es igual, si consideramos un nivel de significación de 0.05, el valor t ha de alcanzar 1.96. Un parámetro no

significativo indicaría que la relación propuesta no tiene ningún efecto sustancial, por lo que debería ser eliminada y el modelo reformulado. Para eliminar los parámetros no significativos ha de sugerirse un proceso paso a paso en el que no se deben eliminar de una vez todos esos parámetros, ya que cada vez que se elimina uno de ellos cambia la estructura del modelo y un parámetro no significativo en un paso puede serlo en otro paso posterior. El nivel de exigencia más elevado consistirá en no aceptar el modelo estructural como válido salvo que todos los parámetros sean significativos y en el sentido esperado.

Otra alternativa adicional para evaluar el modelo estructural es revisar los coeficientes de fiabilidad de las ecuaciones estructurales (R^2) y la matriz de correlaciones estandarizadas entre las variables latentes (para el caso en el que se haya utilizado como matriz inicial la matriz de varianzas-covarianzas). Es también de gran utilidad revisar las correlaciones existentes entre las variables latentes, puesto que si son muy elevadas (más de 0.90 o incluso 0.80) significa que tales variables están explicando información redundante y que no representan constructos diferentes. En suma, habrá que volver a especificar el modelo, eliminando alguna de esas variables redundantes.

La interpretación del modelo se hará con arreglo a la estructura teórica en que se ha basado su especificación y a los diversos coeficientes o parámetros estimados, analizando si se corresponden en magnitud y en sentido (positivo o negativo) con las propuestas planteadas por la teoría. La magnitud de los coeficientes no está únicamente determinada por su significación estadística puesto que depende además de 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 estadísticamente significativa).

De igual manera, es necesario decidir si se usan los coeficientes estandarizados o sin estandarizar al proceder a la interpretación del modelo. Cuando se utiliza la

matriz de varianzas-covarianzas los coeficientes que se obtienen dependen de la escala de medida que tengan los indicadores, lo cual complica bastante el proceso de comparación así como la interpretación del modelo. Este problema ha llevado a buscar algún procedimiento para transformar dichos coeficientes. El mas adecuado es estandarizar los coeficientes obtenidos para ponerlos en una escala 1- +1, multiplicando cada uno de ellos por la razón entre la desviación típica de la variable independiente y la desviación típica de la variable dependiente. Otro procedimiento cada vez más extendido, pero no por ello del todo correcto, es emplear la matriz de correlaciones para estimar los coeficientes del modelo de forma que así se elimina el problema de escala y se facilita la comparación y la interpretación del modelo. No obstante, el uso de la matriz de correlaciones presenta una serie de inconvenientes que se describen anteriormente. La mejor opción será emplear la matriz de varianzas-covarianzas para la estimación del modelo y, posteriormente calcular los coeficientes estandarizados.

Frecuentemente, el proceso de evaluación del modelo desemboca en la modificación del mismo, para lo cual el programa computacional que se utilice proporciona ayuda a través de una serie de indicadores. Es importante señalar que nunca se deben hacer modificaciones de un modelo sin que se tenga una explicación suficientemente basada en la teoría.[36]

3.5. MODELIZACIÓN DE GRUPOS MÚLTIPLES: SIN ESTRUCTURA DE MEDIAS

Se extiende el modelo, de un grupo simple a situaciones de grupo múltiple.

El caso del grupo simple considera la unión del análisis de senderos y el análisis factorial e introduce el problema de la identificación para el modelo de ecuaciones estructurales completo (modelo estructural y modelo de medida).

---

[36] (12) LUQUE, “Técnicas de Análisis de Datos en Investigación de Mercados”, pág. 513-525.

El caso del grupo múltiple, es un enfoque tradicional para probar restricciones a través de grupos.

3.5.1. ESPECIFICACIÓN Y PRUEBAS EN GRUPOS MÚLTIPLES

Se considera inicialmente el problema de evaluar la comparabilidad de estructuras factoriales entre grupos. Jöreskog (1971) sugiere una estrategia para evaluar la comparabilidad de estructuras factoriales entre grupos basada en pruebas de hipótesis restringidas en incremento.

Se considera, el modelo de medida de variables exógenas, para analizar estructuras factoriales con la finalidad de compararlas, por lo tanto, el modelo usado para relacionar medidas observadas a factores es el *modelo de análisis factorial lineal* que se denota:

$$x = \Lambda_x \xi + \delta \qquad (3.9)$$

donde, x: es un vector q x 1 de respuestas observadas (podría tratarse de un cuestionario que contiene q preguntas hechas a un número N de individuos).

Λ_x: es una matriz q x k de factores de regresión ponderados (cargas),

ξ: es un vector k x 1 de k factores comunes, y

δ: es un vector q x 1 de variables únicas que contiene en conjunto al error de medida y al error específico.

La ecuación (3.9) expresa las variables observadas en términos de un conjunto ponderado de factores comunes y de un vector de variables únicas.

El modelo factorial descrito en la ecuación (3.9) se lo describe ahora, con un índice de grupo $g = 1,2,...,G$:

$$x_g = \Lambda_{xg} \xi_g + \delta_g, \qquad (3.10)$$

donde: x_g: es un vector de medidas observadas;

Λ_{xg}: es una matriz de factores de cargas;

ξ_g: es un vector de factores comunes; y,

δ_g : vector de variables únicas.

Suponiendo independencia entre las muestras y asumiendo que los valores de las variables están relacionados a una población normal multivariante, la función de log-verosimilitud, para el grupo g, es:

$$\log L_0(\Omega)_g = -\frac{n_g}{2}\log|\Sigma_g| + tr\left(S_g\Sigma_g^{-1}\right), \tag{3.11}$$

al realizar una sumatoria se tiene:

$$\log L_0(\Omega) = \sum_{g=1}^{G}\log L_0(\Omega). \tag{3.12}$$

Para obtener la función de máxima verosimilitud, F_{MV}, se minimiza la función de la ecuación (3.11), cuyo resultado es:

$$F_{MV} = \log|\Sigma| + tr\left(S\Sigma^{-1}\right) - \log|S| - q. \tag{3.13}$$

Resultando la especificación del modelo y las supocisiones requeridas, la primera prueba de hipótesis que se sugiere realizar es, la de igualdad de las matrices de covarianza a través de grupos. Para este primer paso, no se impone ninguna estructura especial al modelo. La idea es más bien, determinar si las matrices de covarianza difieren. La hipótesis nula para este primer paso es:

$$H_\Sigma : \Sigma_1 = \Sigma_2 = ... = \Sigma_G. \tag{3.14}$$

Para comprobar esta hipótesis, se utiliza la prueba M de Box, la cual está asintóticamente distribuida como una ji-cuadrado, de este modo:

$$M = n\log|S| - \sum_{g-1}^{G} n_g \log|S_g|, \tag{3.15}$$

y con grados de libertad iguales a:

$$gl_{\Sigma} = 1/2(g-1)q(q+1). \qquad (3.16)$$

En el procedimiento multivariante estándar, como el análisis multivariante de varianza, lo usual es retener la hipótesis nula de igualdad de las matrices de covarianza. En las pruebas propuestas por Jöreskog (1971), reteniendo $H_{0\Sigma}$ sugeriría un procedimiento con un análisis usando la matriz de covarianza agrupada y un examen discontinuo de diferencias de grupo.

Si la hipótesis de igualdad de covarianzas se rechaza, el siguiente paso es realizar otra prueba de hipótesis, pero ahora para probar la igualdad del número de factores, sin considerar el patrón específico de cargas fijas y libres. La hipótesis nula, siendo k un número específico de factores, es:

$$H_k: k_1 = k_2 = ... = k_G, \qquad (3.17)$$

Esta prueba suele realizarse como una separación no restringida de modelos de análisis factorial, en donde cada modelo se prueba utilizando una ji-cuadrado con $gl_k = ½ [(q - k)^2 - (q + k)]$ grados de libertad. Pero como la prueba estadística ji-cuadrado es independiente, éstas pueden sumarse, obteniendo así una prueba global ji-cuadrado para probar la igualdad del número de factores, con grados de libertad:

$$gl_k = ½ [(q - k)^2 - (q + k)]. \qquad (3.18)$$

Si esta hipótesis se rechaza, entonces se detiene la prueba y los análisis pueden tener lugar dentro de grupos.

En cambio, si la hipótesis de igualdad del número de factores no se rechaza, el siguiente paso es realizar una prueba de igualdad de factores de carga, llamada comúnmente, prueba de *invarianza factorial* y su hipótesis nula es:

$$H_{\Lambda}: \Lambda_1 = \Lambda_2 = ... = \Lambda_G. \qquad (3.19)$$

La prueba de invarianza factorial, se obtiene poniendo contenidos iguales a través de grupos para los elementos comunes de Λ y permitiendo que los parámetros

permanezcan libres a través de los grupos. De esta manera, el resultado es un estadístico ji-cuadrado que puede evaluarse con grados de libertad:

$$gl_{\Lambda} = ½\ gq\ (q+1) - qk + q - ½\ qk\ (k+1) - gq \qquad (3.20)$$

Si la invarianza factorial se rechaza, se detiene la prueba en este punto.

Por el contrario, si la invarianza factorial no se rechaza, lo siguiente, es evaluar la igualdad de los factores de carga y las variables únicas. Según la notación de Jöreskog (1971) es:

$$H_{\Lambda\Theta}: \Lambda_1 = \Lambda_2 = ... = \Lambda_G, \qquad (3.21)$$

$$\Theta_1 = \Theta_2 = ... = \Theta_G.$$

La hipótesis nula de la ecuación (3.21) se obtiene poniendo iguales contenidos en los grupos, para elementos comunes de Λ y Θ.

Nuevamente se obtiene una ji-cuadrado, que puede evaluarse con grados de libertad iguales a:

$$gl_{\Lambda\Theta} = ½\ gq\ (q+1) - qk + q - ½\ gk\ (k+1) - q \qquad (3.22)$$

Si la prueba se rechaza, entonces se detiene el procedimiento en ese momento.

Por último, se puede probar la invarianza de todos los parámetros a través de los grupos. Esta hipótesis nula es:

$$H_{\Lambda\Theta\Phi}: \Lambda_1 = \Lambda_2 = ... = \Lambda_G, \qquad (3.23)$$

$$\Theta_1 = \Theta_2 = ... = \Theta_G,$$

$$\Phi_1 = \Phi_2 = \ldots = \Phi_G.$$

Esta prueba de hipótesis, utiliza la matriz de covarianza de muestras agrupadas.

En este caso también se evalúa un estadístico ji-cuadrado, con grados de libertad:

$$gl_{\Lambda\Theta\Phi} = ½\ q\ (q+1) - qk + q - ½\ gk\ (k+1) - q. \qquad (3.24)$$

De esta manera, esta estrategia modelada se puede extender al modelo general de ecuaciones estructurales, aumentando constantes que pertenezcan a la igualdad de los coeficientes estructurales B, Γ y Ψ.

Según Jöreskog (1971), la hipótesis descrita en la ecuación (3.23) es más fuerte que la hipótesis en (3.14) de igualdad de matrices de covarianza, ya que ésta última incluye casos donde Σ no necesariamente indica un modelo factorial común.

3.6. MODELIZACIÓN DE GRUPOS MÚLTIPLES: USANDO MEDIAS

La covarianza es el dato básico de los modelos de ecuaciones estructurales, pero ésta no contiene información sobre medias. Si únicamente las covarianzas son analizadas, entonces, todas las variables observadas tienen media centrada para que las variables latentes deban tener media cero. Algunas veces esta pérdida de información es demasiado restrictiva (por ejemplo, las medias de variables medibles puede esperarse que difieran).

Las medias son estimadas en los modelos de ecuaciones estructurales agregando una estructura de medias a la estructura básica de covarianza del modelo. Los datos de entrada para el análisis de un modelo con una estructura de medias son las covarianzas y las medias.

Un enfoque de la modelización de ecuaciones estructurales al análisis de medias se distingue por la capacidad de probar hipótesis sobre las medias de las variables latentes y la estructura de covarianza de los términos de error.

3.6.1. ESPECIFICACIÓN Y PRUEBAS EN LA ESTRUCTURA DE MEDIAS

Para realizar una estimación de las diferencias entre grupos de las medias de variables latentes, se debe expandir el modelo factorial, incorporando coeficientes de intersección. Expandiendo el modelo factorial descrito en la ecuación (3.10): $x_g = \Lambda_{xg}\xi_g + \delta_g$, se tiene:

$$x_g = \tau_x + \Lambda_{xg}\xi_g + \delta_g, \qquad (3.25)$$

donde τ_x es un vector de coeficientes de intersección q-dimensional. La definición de los términos restantes se mantiene igual. Para este caso, donde se tiene estructura de medias, se aumenta la suposición de que:

$$E(x_g) = \tau_x + \Lambda_{xg}E(\xi_g) \qquad (3.26)$$

$$= \tau_x + \Lambda_{xg}\kappa_g,$$

donde κ_g es un vector k-dimensional de factores de medias para el grupo g.

3.6.2. IDENTIFICACIÓN DEL MODELO CON ESTRUCTURA DE MEDIAS

Los parámetros de un modelo con una estructura de medias incluyen las medias de las variables exógenas, los coeficientes de intersección de las variables endógenas, y el número de parámetros de la parte de covarianza del modelo contado de manera usual para aquel tipo de modelo.

Una regla simple para el conteo del número total de varianzas-covarianzas disponibles para estimar los parámetros de un modelo con estructura de medias es $v(v+3)/2$, donde v es el número de variables observadas. El valor de esta expresión da el número total de varianzas, de covarianzas no redundantes, y de medias de variables observadas. Por ejemplo, si se tiene 5 variables observadas, entonces hay 5(8)/2=20 varianzas-covarianzas, lo cual incluye 5 medias y 15 varianzas y covarianzas (5(6)/2).

Para que una estructura de medias esté identificada, el número de sus parámetros (las medias de variables exógenas y los coeficientes de intersección de las variables endógenas) no puede exceder el número total de medias de las variables observadas. Además, la condición de identificación de una estructura de medias debe considerarse separadamente de la estructura de covarianza; esto es, una estructura de covarianza sobre identificada no identificará una estructura de medias subidentificada. De igual manera, una estructura de medias sobre identificada no puede remediar una estructura de covarianza subidentificada. Si la

estructura de medias está identificada, tendrá tantos parámetros libres como medias observadas.

Por lo tanto,

- El modelo de medias (es decir, los efectos totales de la constante), será exactamente igual a las correspondientes medias observadas, y
- El ajuste del modelo con estructura de covarianza únicamente, será idéntico a aquel modelo con ambas estructuras: de covarianza y de medias.

Es importante señalar también, que el objetivo general del análisis con estructura de medias es evaluar la diferencia entre grupos de factores de medias.

Cuando el modelo de medida es igual en los grupos, generalmente se asume que la invarianza factorial se mantiene; sin embargo, esto sucede únicamente si hay una selección aleatoria de observaciones y una asignación aleatoria de grupos. Bajo esta suposición de invarianza factorial, el modelo en (3.26) no está identificado.

Entonces, se puede aumentar un vector k-dimensional, denotado como d, para κ_g y de τ_x se sustrae Λd, de esta manera se tiene:

$$E(x_g) = \tau_x - \Lambda d + \Lambda_x(\kappa_g + d) \qquad (3.27)$$

$$= \tau_x + \Lambda\kappa_g .$$

Puesto que d es un vector k-dimensional, solamente si se aumentan k restricciones, el modelo en (3.26) está identificado. Una manera de lograr esto es poniendo κ_g = 0, lo cual fija k restricciones. De aquí, el factor de medias estimado restante es interpretado como diferencias del grupo g.

3.6.3. ESTIMACIÓN DEL MODELO CON ESTRUCTURA DE MEDIAS

La mayoría de métodos de estimación descritos anteriormente para analizar modelos con estructura de covarianza únicamente, pueden aplicarse a modelos

que tienen ambas estructuras: de medias y de covarianza. Esto incluye la estimación de máxima verosimilitud, el método más general en la modelización de ecuaciones estructurales.

Sin embargo, no todos los índices de ajuste estandarizados con estructura de covarianza solamente, puede utilizarse para modelos con ambas estructuras: de medias y de covarianza o talvez calculados solo para la parte de covarianza del modelo. Este se cumple especialmente para los índices de ajuste incrementales, tal como el índice de ajuste comparativo (comparative fit index, CFI), midiendo la mejora relativa en el ajuste del modelo estudiado sobre el modelo nulo. Cuando son analizadas únicamente las covarianzas, el modelo nulo es típicamente el modelo de independencia, el cual asume cero a las covarianzas de la población.

No obstante, el modelo de independencia es más difícil definir cuando se analizan las medias y las covarianzas juntas. Por ejemplo, un modelo de independencia donde todas las medias y covarianzas son ajustadas e igualadas a cero puede ser muy irreal. Un modelo de independencia alternativo permite que las medias de las variables observadas sean estimadas libremente (es decir, las medias no se asume que son cero).

3.7. UN MODELO ALTERNATIVO PARA ESTIMAR LAS DIFERENCIAS DE GRUPOS

En esta parte se considerará un caso especial del modelo de ecuaciones estructurales, llamado modelo MIMIC (multiple indicators and multiple causes), que fue propuesto por Jöreskog y Goldberger (1975) y sirve para los indicadores múltiples y el modelo de causas múltiples.

Siendo x, un vector que contiene códigos ficticios, que representan el número de miembros en los grupos, al modelo MIMIC se lo denota como:

$$y = \Lambda_y \eta + \varepsilon,$$

$$\eta = \Gamma x + \zeta, \qquad (3.28)$$

$$x \equiv \xi.$$

Modificando a la matriz $\Lambda_x = I$, matriz de identidad $q\ x\ q$, y a $\Theta_\delta = 0$, matriz nula, se obtiene la identidad entre x y ξ.

No hay reglas especiales de identificación que estén asociadas con el modelo MIMIC, se procede de la misma manera que con el modelo general de ecuaciones estructurales. De igual manera, para estimar los parámetros del modelo MIMIC se realiza el mismo procedimiento que para estimar los parámetros del modelo general de ecuaciones estructurales.

3.7.1. EXTENSIONES DEL MODELO MIMIC

El modelo MIMIC es uno de los casos especiales más flexibles del modelo general de ecuaciones estructurales, utilizado regularmente para solucionar problemas en las ciencias sociales y del comportamiento.

Al modelo MIMIC se lo puede incorporar cualquier tipo de variables exógenas (continuas o categóricas). Inicialmente se considera el vector de variables exógenas x, el cual puede codificarse para representar el análisis ortogonal de varianza diseñando vectores, es decir, integrando diseños experimentales a manera de variable latente.

Una especificación interesante del modelo MIMIC, se deriva del trabajo realizado por Muthén (1989) sobre estimación de parámetros en poblaciones heterogéneas. Entre otras cosas, Muthén extendió la especificación del modelo MIMIC para permitir la regresión de indicadores, así como de factores de variables exógenas. De este modo, se puede conocer si hay diferencias de grupo en términos específicos, por sobre el factor.

En esta especificación extendida, primero se considera el modelo completo de ecuaciones estructurales:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta,$$

$$y = \Lambda_y \eta + \varepsilon, \qquad (3.29)$$

$$x = \Lambda_x \xi + \delta.$$

Siendo Λ_x = I, y Θ_δ = 0, de igual manera que en el modelo regular MIMIC, además, se tiene que para este modelo extendido, Λ_y = I, y Θ_ε = 0, esta especificación mantiene a las cargas como elementos en B. La métrica de las variables latentes puede determinarse, poniendo a una carga 1.0, del mismo modo que en el modelo básico MIMIC.

En esta parametrización extendida, las cargas se mantienen en la matriz B, y uno de los elementos de B puede fijarse (o ponerse) en 1.0. En esta parametrización, la matriz Γ contiene las regresiones del factor, así como sus indicadores en las variables exógenas. Por ejemplo, en el caso de un factor simple, el vector ζ contiene $p+1$ elementos, en donde los p primeros elementos están asociados de manera única con los elementos de ε y el último elemento es el término de perturbación ζ.

3.8. PROBLEMAS DE INFERENCIA CAUSAL EN MODELOS DE GRUPOS MÚLTIPLES

En esta sección se considerará el problema de selección no aleatoria, para tratar problemas de invarianza factorial. También se consideran métodos que contienen mecanismos de selección no aleatoria en modelos de variables latentes para diferencias de grupos.

3.8.1. EL PROBLEMA DE LA INVARIANZA FACTORIAL

El problema de la invarianza factorial hace referencia de hasta qué punto se asume que un modelo factorial contiene una población padre que a su vez contiene una subpoblación formada por medias, con algún criterio de selección.

Se considera el modelo factorial analítico para explicar el problema de la invarianza factorial. Inicialmente se asume que $E(\xi\delta')=0$ y que $E(\delta\delta') = \Theta$. La matriz de covarianza de las q variables, se denota:

$$\Sigma = \Lambda\Phi\Lambda'+\Theta\,, \qquad (3.30)$$

donde $\sum$: matriz $q\ x\ q$ de covarianza de la población,

$\Phi = E(\xi\xi')$: matriz $k\ x\ k$ de covarianza factorial, y

Θ: matriz diagonal $q\ x\ q$ de varianzas únicas.

También se asume que $E(\xi) = \kappa$; es el vector $q\ x\ q$ de medias observadas. Por lo tanto, μ puede modelarse como:

$$\mu = \tau_x + \Lambda_\kappa, \qquad (3.31)$$

donde τ_x: vector $q\ x\ 1$ de interceptos medibles, y

κ: vector $k\ x\ 1$ de medias factoriales.

Sea z una selección de variables y $f(z)$ la función de selección, que determina la selección de una subpoblación de la población padre.

Meredith (1993) hace distinción entre dos tipos de invarianza factorial, la *invarianza factorial fuerte* y la *invarianza factorial estricta*. Cualquiera que fuera el caso, ciertas suposiciones deben mantenerse. Estas suposiciones son: que un modelo factorial se mantiene en la población padre, que la covarianza condicional de los factores y la unicidad dan la función $f(z)$ como un vector cero. Bajo estas dos suposiciones, la invarianza factorial fuerte implica que para cada subpoblación, denotada como s, se tiene que:

$$\mu_s = \tau_x + \Lambda\kappa_s\,, \qquad (3.32)$$

y

$$\sum_s = \Lambda\Phi_s\Lambda'+\Theta_s\,. \qquad (3.33)$$

Las ecuaciones (3.32) y (3.33) significan que bajo la invarianza factorial fuerte, tanto los interceptos estructurales como los factores de carga, se mantienen sin variación alguna en los grupos; pero en sentido factorial, tanto la matriz de covarianza factorial como la matriz de covarianza de unicidad, pueden diferir.

La invarianza factorial estricta contiene la ecuación (3.33), pero en este caso la matriz de varianzas únicas permanece constante en las subpoblaciones, es decir:

$$\sum_s = \Lambda\Phi_s\Lambda' + \Theta. \qquad (3.34)$$

En conclusión, la aplicación práctica de la invarianza factorial al modelamiento de múltiples grupos, se refleja en pruebas explícitas donde existen mecanismos de selección.[37]

[37] (8) KAPLAN, "Structural Equation Modeling: Fundations and Extensions", pág. 63-75.

CAPÍTULO 4
APLICACIÓN

4.1. INTRODUCCIÓN

Se aplica el análisis de senderos, el análisis factorial y la modelización de ecuaciones estructurales, a un conjunto de variables, de este modo se podrá comprender de mejor manera la teoría expuesta en los tres capítulos anteriores.

Para la aplicación, se eligieron ocho variables: Nuevos Conocimientos, Carga Lectiva, Volumen Global de Trabajo, Interés, Infraestructura, Importancia, Matriculados, e Índice de Participación.

Para llevar a cabo la recolección de datos, se realizó una encuesta durante los años lectivos 2002/2003 y 2003/2004, a todos los alumnos (N=242) del ciclo diversificado (cuarto, quinto y sexto cursos) de la Unidad Educativa "Santa Juana de Chantal" de la ciudad de Otavalo.

Con la información obtenida se aplica el análisis de senderos y se determinan las relaciones de causalidad entre las ocho variables, para obtener el diagrama de senderos respectivo y establecer el sistema de ecuaciones estructurales correspondiente a las variables estudiadas.

Se resuelve el sistema y se obtiene los coeficientes de Wright de las ecuaciones, se realiza la determinación numérica de las ecuaciones en función de los coeficientes de la matriz de correlación de las ocho variables observadas. Una vez que se identifican los coeficientes de Wright, los valores figuran el diagrama de senderos, y por último se obtienen los efectos indirectos de las variables del sistema, para su posterior interpretación.

A través de la aplicación del análisis factorial se logra explicar las ocho variables observadas mediante dos factores. Para determinar estos factores se realizó un estudio con ayuda del programa SPSS 12.

Se utiliza el método de componentes principales y se desarrollan las siguientes etapas: se calcula las relaciones entre las ocho variables por medio de la matriz

de correlaciones; se realiza la extracción de los factores que describen los componentes principales de la varianza en la matriz de correlación y constituyen la matriz de factores iniciales, en este caso se obtienen dos factores; se rota los dos factores descritos en la matriz de factores iniciales para producir una estructura más susceptible de interpretación, mediante el procedimiento de rotación ortogonal para soluciones ortogonales Varimax; y, se calculan las puntuaciones factoriales para expresar cada factor como combinación lineal de las ocho variables estudiadas.

Para la aplicación de los modelos de ecuaciones estructurales, se une la metodología desarrollada en el análisis de senderos y en el análisis factorial.

Para el desarrollo del modelo de ecuaciones estructurales se lleva a cabo cuatro fases: la especificación, la identificación, la estimación y la evaluación e interpretación de dicho modelo.

En la etapa de especificación del modelo, se aplican los conocimientos teóricos del fenómeno estudiado al planteamiento de las ecuaciones matemáticas relativas a los efectos causales de las variables latentes y a las expresiones que las relacionan con los indicadores o variables observables; se plantea el modelo de ecuaciones estructurales mediante el diagrama de senderos utilizando el programa AMOS 4.0 (versión estudiantil) y con este programa se realiza también el desarrollo de las siguientes tres etapas de la modelización de ecuaciones estructurales.

El segundo paso es realizar la identificación del modelo, para asegurar que pueden ser estimados los parámetros del modelo; el modelo está identificado si todos los parámetros lo están, es decir, si existe una solución única para cada uno de los parámetros estimados; es importante señalar que el modelo de ecuaciones estructurales planteado es recursivo, por lo tanto, está identificado, siendo este tipo de modelo el que no contienen efectos circulares o recíprocos entre sus variables.

El siguiente paso es realizar la estimación del modelo, con el programa AMOS, mediante el método de máxima verosimilitud. Partiendo de que el modelo está

identificado, cada uno de los parámetros tendrá un valor único. Este proceso de estimación consiste en la obtención de aquellos valores de los parámetros que ajusten lo mejor posible a la matriz observada, por aquellos que la reproducen.

Una vez que se ha identificado y estimado el modelo, se procede a evaluar qué tanto se han ajustado los datos de esta aplicación al modelo propuesto, para lo cual se realiza una evaluación del ajuste del modelo global con varias pruebas de ajuste. Debido a que los índices de bondad de ajuste evaluados presentaron valores buenos (óptimos), las posibilidades de modificar el modelo son nulas, ya que una modificación no presentaría una mejora sustancial; finalmente se realiza una interpretación de los resultados obtenidos con la finalidad de sustentar la idea principal de esta aplicación.

Como ya se mencionó, se utiliza el programa AMOS 4.0 (Versión estudiantil) para la parte de las Ecuaciones Estructurales y los programas SPSS 12, NCSS 6.0 y Microsoft Excel, para realizar los Análisis de Senderos y Factorial.

4.2. APLICACIÓN DEL ANÁLISIS DE SENDEROS

Para la aplicación de la teoría expuesta, se determinan inicialmente ocho variables de la siguiente manera:

- Variables

  Para obtener los datos de las seis primeras variables se utiliza la encuesta realizada a los alumnos, cuantificando sus respuestas con cifras entre 1 y 5.

  Las dos últimas variables no hacen referencia a ninguna pregunta de la encuesta, y son: variable Matriculados, que contiene el número de estudiantes

matriculados en cada curso, y variable Índice de Participación, que es el porcentaje de los estudiantes que colaboraron en la encuesta.[38]

- Distribución de frecuencias

Se muestra en el gráfico 4.1, la distribución de frecuencias de los datos de las seis primeras variables que corresponden a las respuestas que se obtuvieron en la encuesta.

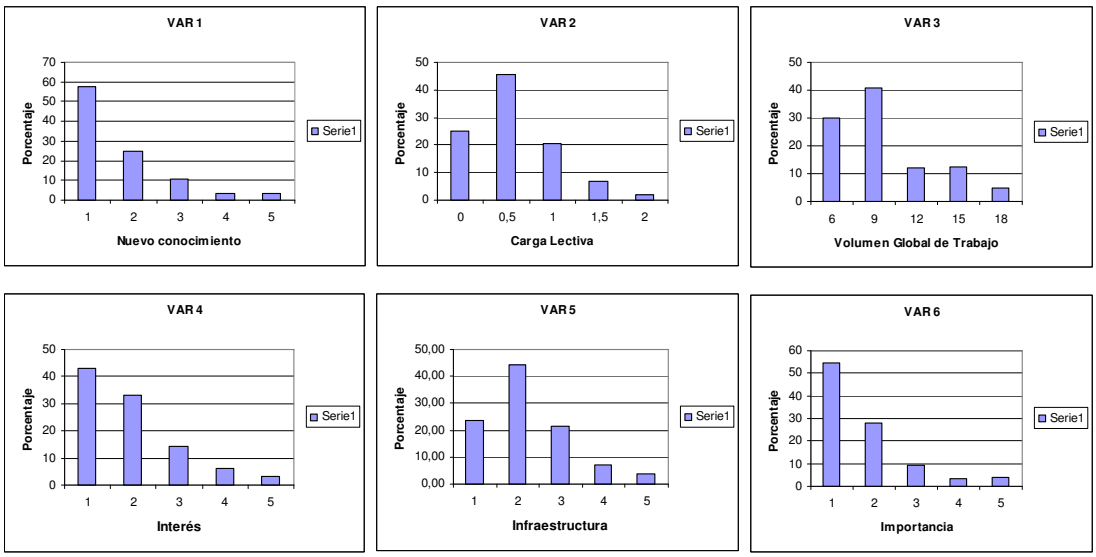

GRAFICO 4.1 Histograma de variables analizadas en la encuesta.

Las variables 1, 4 y 6 presentan una asimetría positiva mayor a uno, es decir, una mayor dispersión hacia la derecha y mayor concentración a la izquierda; lo cual indica que hay más estudiantes que han contestado muy positivamente antes que negativamente. El resto de variables presenta una asimetría igualmente positiva pero menor a uno, e indica de igual manera que los estudiantes han contestado en su mayoría de manera positiva a las preguntas realizadas en la encuesta.

---

[38] BALBUENA y CASAS, "Aplicación del análisis factorial a la valoración por parte de los estudiantes de las asignaturas de la ETSICCO de Barcelona en sus distintas titulaciones".

- Estadísticos

La tabla 4.1, muestra los estadísticos numéricos correspondientes a las ocho variables observadas. Como se aprecia, el coeficiente de variación de todas las variables es inferior a uno, lo cual indica que los datos son bastante homogéneos.

	Nuevos Conocimientos	Carga Lectiva	Volumen Global de Tabajo	Interés	Infraestructura	Importancia	Matriculados	Índice de Participación
Media	1.6846	0.5747	9.6198	1.9336	2.2314	1.7479	84.000	0.3775
Mediana	1.0000	0.5000	9.0000	2.0000	2.0000	1.0000	82.000	0.3056
Error Típico	0.0639	0.0303	0.2186	0.0682	0.0648	0.0675	4.9461	0.0141
Desv. Estándar	0.9917	0.4706	3.4006	1.0586	1.0083	1.0498	24.2308	0.2027
Varianza	0.9835	0.2215	11.5644	1.1206	1.0168	1.1022	587.130	0.0411
Asimetría	1.5976	0.7856	0.8615	1.1321	0.8198	1.5828	0.1344	1.9771
Coef. Variación	0,5887	0,8189	0.3535	0,5475	0,4519	0.6006	0.5370	0.5435
Mín.	1.00	0.00	6.00	1.00	1.00	1.00	56.00	0.21
Máx.	5.00	2.00	18.00	5.00	5.00	5.00	114.00	1.00

TABLA 4.1 Estadísticos numéricos.

A continuación se encuentra la matriz de correlaciones que procede de las variables antes descritas, en la tabla 4.2.

Matriz de correlación[a]

		Nuevos Conocimientos	Carga Lectiva	Volumen Global de Trabajo	Interés	Infraestructura	Importancia	Matriculados	Indice de Participación
Correlación	Nuevos Conocimientos	1,000	,681	,756	,773	,704	,946	,094	,094
	Carga Lectiva	,681	1,000	,920	,781	,951	,682	,149	,149
	Volumen Globa de Trabajo	,756	,920	1,000	,830	,878	,758	,074	,074
	Interés	,773	,781	,830	1,000	,767	,798	,121	,122
	Infraestructura	,704	,951	,878	,767	1,000	,700	,097	,097
	Importancia	,946	,682	,758	,798	,700	1,000	,167	,167
	Matriculados	,094	,149	,074	,121	,097	,167	1,000	1,000
	Indice de Participación	,094	,149	,074	,122	,097	,167	1,000	1,000

a. Determinante = 3,676E-10

TABLA 4.2 Matriz de correlación.

Para facilitar los cálculos en la realización de la aplicación del análisis de senderos, se denota a las variables:

x_1, a la variable Nuevos Conocimientos;

x_2, a la variable Carga Lectiva;

x_3, a la variable Volumen Global de Trabajo;

x_4, a la variable Interés;

x_5, a la variable Infraestructura;

x_6, a la variable Importancia;

x_7, a la variable Matriculados;

x_8, a la variable Índice de Participación.

- Se establece las relaciones de influencia que se puedan considerar probables entre las variables antes indicadas.

Mediante la matriz de correlaciones, se puede observar, que la variable Nuevos Conocimientos, x_1, influye sobre las variables x_2, x_3, x_4, x_5 y x_6.

También, la variable Carga Lectiva, x_2, influye en la variable Volumen Global de Trabajo x_3, en la Interés x_4 y en la variable Infraestructura x_5.

La variable x_3, Volumen Global de Trabajo, influye sobre las variables x_4, x_5 y x_6.

La variable Interés, x_4, influye sobre las variables x_5 y x_6.

Y, la variable Infraestructura x_5, influye sobre la variable Importancia, x_6.

La variable Importancia, x_6, influye sobre las variables x_7 y x_8.

La variable Matriculados, x_7, influye sobre la variable Índice de Participación x_8.

Por lo tanto, el diagrama de senderos es el siguiente:

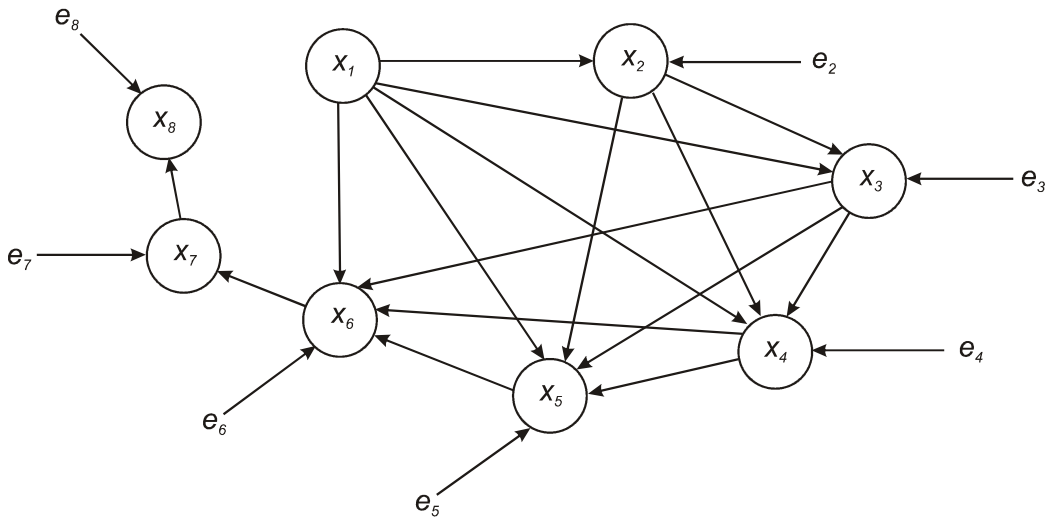

GRAFICO 4.2 Diagrama de senderos.

En el diagrama de senderos del gráfico 4.1 se distinguen las siguientes variables: endógenas que son x_2, x_3, x_4, x_5, x_6, x_7, x_8; exógena que es x_1, y residuales que son e_2, e_3, e_4, e_5, e_6.

El gráfico 4.1, indica la influencia en las variables endógenas del modelo, por parte de otras variables no conocidas y de los errores de medida, generalmente llamadas variables residuales, mediante flechas que unen cada variable e_i con la correspondiente variable endógena.

- Partiendo del diagrama de senderos anterior, se desarrollará el correspondiente conjunto de ecuaciones estructurales.

Para obtener el conjunto de ecuaciones estructurales, se debe tener en consideración que cualquier variable que según el diagrama sea dependiente de otra u otras, por acabar en ella una o más puntas de flecha, se puede expresar como una función de las variables de las cuales parten las flechas, en este caso, se puede ver claramente que en la variable x_2, acaban dos

puntas de flecha, de x_1 y de e_2, por lo tanto, x_2 se puede expresar como una función lineal de x_1 y e_2.

Así mismo, x_3 puede ser expresada como función lineal de las variables x_1, x_2 y e_3; x_4, de las variables x_1, x_2, x_3 y e_4; x_5 se la puede expresar en función de x_1, x_2, x_3, x_4 y e_5; x_6 se puede expresar en función de las variables x_1, x_3, x_4, x_5 y e_6; x_7, de las variables x_6 y e_7; y por último x_8 se puede expresar en función de x_7 y e_8.

Por tanto, las correspondientes ecuaciones estructurales son:

$x_2 = p_{21}\, x_1 + p_{2e}\, e_2$

$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3e}\, e_3$

$x_4 = p_{41}\, x_1 + p_{42}\, x_2 + p_{43}\, x_3 + p_{4e}\, e_4$

$x_5 = p_{51}\, x_1 + p_{52}\, x_2 + p_{53}\, x_3 + p_{54}\, x_4 + p_{5e}\, e_5$

$x_6 = p_{61}\, x_1 + p_{63}\, x_3 + p_{64}\, x_4 + p_{65}\, x_5 + p_{6e}\, e_6$

$x_7 = p_{76}\, x_6 + p_{7e}\, e_7$

$x_8 = p_{87}\, x_7 + p_{8e}\, e_8$

- Considerando las ecuaciones estructurales anteriores, se formulará en términos de los coeficientes de Wright p_{ij} y de los coeficientes de correlación r_{ij}, las ecuaciones del análisis de senderos que se puedan derivar de las mismas ecuaciones estructurales.

Se partirá de la ecuación del teorema básico del análisis de senderos:

$$r_{ij} = \sum_{q} p_{iq} r_{jq}$$

en donde, i y j son dos variables del sistema y q indica que se han de sumar todas las variables q, cuyos caminos conduzcan directamente a x_i, siendo $q<i$.

Considerando esta fórmula, se tiene que los p_{iq} son los de la ecuación estructural correspondiente y los r_{iq} están formados por el primer subíndice de r_{ij} en la ecuación que se trate en cada caso y el q, está formado por toda la serie de números q, siendo q = i-1. Durante este desarrollo, los p_{ie} se hacen cero todos.

Por tanto, los sistemas de ecuaciones correspondientes para cada caso son:

Primer sistema: $r_{21} = p_{21}$

Segundo sistema: $r_{31} = p_{31} + p_{32}\, r_{12}$

$r_{32} = p_{31}\, r_{12} + p_{32}$

Tercer sistema: $r_{41} = p_{41} + p_{42}\, r_{12} + p_{43}\, r_{13}$

$r_{42} = p_{41}\, r_{12} + p_{42} + p_{43}\, r_{23}$

$r_{43} = p_{41}\, r_{13} + p_{42}\, r_{23} + p_{43}$

Cuarto sistema: $r_{51} = p_{51} + p_{52}\, r_{12} + p_{53}\, r_{13} + p_{54}\, r_{14}$

$r_{52} = p_{51}\, r_{12} + p_{52} + p_{53}\, r_{23} + p_{54}\, r_{24}$

$r_{53} = p_{51}\, r_{13} + p_{52}\, r_{23} + p_{53} + p_{54}\, r_{34}$

$r_{54} = p_{51}\, r_{14} + p_{52}\, r_{24} + p_{53}\, r_{34} + p_{54}$

Quinto sistema: $r_{61} = p_{61} + p_{63}\, r_{13} + p_{64}\, r_{14} + p_{65}\, r_{15}$

$r_{63} = p_{61}\, r_{13} + p_{63} + p_{64}\, r_{34} + p_{65}\, r_{35}$

$r_{64} = p_{61}\, r_{14} + p_{63}\, r_{34} + p_{64} + p_{65}\, r_{45}$

$r_{65} = p_{61}\, r_{15} + p_{63}\, r_{35} + p_{64}\, r_{45} + p_{65}$

Sexto sistema: $r_{76} = p_{76}$

Séptimo sistema: $r_{87} = p_{87}$

Ahora, se sustituye los r_{ij} que aparecen en los sistemas de ecuaciones estructurales anteriores, por los valores de éstos según la matriz de correlaciones, para de esta manera obtener los coeficientes p_{ij} que se pretende identificar.

Primer sistema: $0.681 = p_{21}$

Segundo sistema: $0.756 = p_{31} + 0.681\ p_{32}$

$0.920 = 0.681\ p_{31} + p_{32}$

Tercer sistema: $0.773 = p_{41} + 0.681\ p_{42} + 0.756\ p_{43}$

$0.781 = 0.681\ p_{41} + p_{42} + 0.920\ p_{43}$

$0.830 = 0.756\ p_{41} + 0.920\ p_{42} + p_{43}$

Cuarto sistema: $0.704 = p_{51} + 0.681\ p_{52} + 0.756\ p_{53} + 0.773\ p_{54}$

$0.951 = 0.681\ p_{51} + p_{52} + 0.920\ p_{53} + 0.781\ p_{54}$

$0.878 = 0.756\ p_{51} + 0.920\ p_{52} + p_{53} + 0.830\ p_{54}$

$0.767 = 0.773\ p_{51} + 0.781\ p_{52} + 0.830\ p_{53} + p_{54}$

Quinto sistema: $0.946 = p_{61} + 0.756\ p_{63} + 0.773\ p_{64} + 0.704\ p_{65}$

$0.758 = 0.756\ p_{61} + p_{63} + 0.830\ p_{64} + 0.878\ p_{65}$

$0.798 = 0.773\ p_{61} + 0.830\ p_{63} + p_{64} + 0.767\ p_{65}$

$0.700 = 0.704\ p_{61} + 0.878\ p_{63} + 0.767\ p_{64} + p_{65}$

Sexto sistema: $0.167 = p_{76}$

Séptimo sistema: $1.000 = p_{87}$

- Se resolverá los sistemas de ecuaciones anteriores, para hallar los valores de los coeficientes p_{ij}.

El primer sistema de ecuaciones está resuelto:

$$p_{21} = 0.681$$

Se resolverán los sistemas restantes, utilizando las fórmulas que dan directamente los valores de p:

Segundo Sistema: $0.756 = p_{31} + 0.681\ p_{32}$

$$0.920 = 0.681\ p_{31} + p_{32}$$

$$p_{31} = \frac{(r_{13} - r_{12}r_{23})}{(1 - r_{12}^2)} = \frac{0.756 - (0.681)(0.920)}{(1 - 0.681^2)} = \frac{0.756 - 0.6265}{0.5362} = 0.2414$$

$$p_{32} = \frac{(r_{23} - r_{12}r_{13})}{(1 - r_{12}^2)} = \frac{0.920 - (0.681)(0.756)}{(1 - 0.681^2)} = \frac{0.756 - 0.3148}{0.5362} = 0.7556$$

Por lo tanto, se tiene: $p_{31} = 0.2414$, y $p_{32} = 0.7556$.

Tercer sistema: $0.773 = p_{41} + 0.681\ p_{42} + 0.756\ p_{43}$

$$0.781 = 0.681\ p_{41} + p_{42} + 0.920\ p_{43}$$

$$0.830 = 0.756\ p_{41} + 0.920\ p_{42} + p_{43}$$

Se calcula primero el valor de p,

$$p = \begin{vmatrix} 1 & r_{12} & r_{14} \\ r_{12} & 1 & r_{24} \\ r_{13} & r_{23} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0.681 & 0.756 \\ 0.681 & 1 & 0.920 \\ 0.756 & 0.920 & 1 \end{vmatrix} = 0.0656$$

Ahora se usa las fórmulas para obtener directamente los valores de p que corresponden a este tercer sistema:

$$p_{41} = \frac{\begin{vmatrix} r_{14} & r_{12} & r_{13} \\ r_{24} & 1 & r_{23} \\ r_{34} & r_{23} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 0.773 & 0.681 & 0.756 \\ 0.781 & 1 & 0.920 \\ 0.830 & 0.920 & 1 \end{vmatrix}}{0.0656} = \frac{0.0226}{0.0656} = 0.3446$$

$$p_{42} = \frac{\begin{vmatrix} 1 & r_{14} & r_{13} \\ r_{12} & r_{24} & r_{23} \\ r_{13} & r_{34} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.773 & 0.756 \\ 0.681 & 0.781 & 0.920 \\ 0.756 & 0.830 & 1 \end{vmatrix}}{0.0656} = \frac{0.0096}{0.0656} = 0.1459$$

$$p_{43} = \frac{\begin{vmatrix} 1 & r_{12} & r_{14} \\ r_{12} & 1 & r_{24} \\ r_{13} & r_{23} & r_{34} \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.681 & 0.773 \\ 0.681 & 1 & 0.781 \\ 0.756 & 0.920 & 0.830 \end{vmatrix}}{0.0656} = \frac{0.0286}{0.0656} = 0.4353$$

Por tanto, se tiene: $p_{41} = 0.3446$, $p_{42} = 0.1459$ y $p_{43} = 0.4353$.

Cuarto sistema:

$$0.704 = p_{51} + 0.681\ p_{52} + 0.756\ p_{53} + 0.773\ p_{54}$$

$$0.951 = 0.681\ p_{51} + p_{52} + 0.920\ p_{53} + 0.781\ p_{54}$$

$$0.878 = 0.756\ p_{51} + 0.920\ p_{52} + p_{53} + 0.830\ p_{54}$$

$$0.767 = 0.773\ p_{51} + 0.781\ p_{52} + 0.830\ p_{53} + p_{54}$$

Se realizan los cálculos como se hizo anteriormente con las correspondientes fórmulas para este caso.

$$p = \begin{vmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{12} & 1 & r_{23} & r_{24} \\ r_{13} & r_{23} & 1 & r_{34} \\ r_{14} & r_{24} & r_{43} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0.681 & 0.756 & 0.773 \\ 0.681 & 1 & 0.920 & 0.781 \\ 0.756 & 0.920 & 1 & 0.830 \\ 0.773 & 0.781 & 0.830 & 1 \end{vmatrix} = 0.0170$$

$$p_{51} = \frac{\begin{vmatrix} r_{15} & r_{12} & r_{13} & r_{14} \\ r_{25} & 1 & r_{23} & r_{24} \\ r_{35} & r_{32} & 1 & r_{34} \\ r_{45} & r_{42} & r_{34} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 0.704 & 0.681 & 0.756 & 0.773 \\ 0.951 & 1 & 0.920 & 0.781 \\ 0.878 & 0.920 & 1 & 0.830 \\ 0.767 & 0.781 & 0.830 & 1 \end{vmatrix}}{0.0170} = \frac{0.0021}{0.0170} = 0.1212$$

$$p_{52} = \frac{\begin{vmatrix} 1 & r_{15} & r_{13} & r_{14} \\ r_{12} & r_{25} & r_{23} & r_{24} \\ r_{13} & r_{35} & 1 & r_{34} \\ r_{14} & r_{45} & r_{34} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.704 & 0.756 & 0.773 \\ 0.681 & 0.951 & 0.920 & 0.781 \\ 0.756 & 0.878 & 1 & 0.830 \\ 0.773 & 0.767 & 0.830 & 1 \end{vmatrix}}{0.0170} = \frac{0.0160}{0.0170} = 0.9425$$

$$p_{53} = \frac{\begin{vmatrix} 1 & r_{12} & r_{15} & r_{14} \\ r_{12} & 1 & r_{25} & r_{24} \\ r_{13} & r_{32} & r_{35} & r_{34} \\ r_{14} & r_{42} & r_{45} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.681 & 0.704 & 0.773 \\ 0.681 & 1 & 0.951 & 0.781 \\ 0.756 & 0.920 & 0.878 & 0.830 \\ 0.773 & 0.781 & 0.767 & 1 \end{vmatrix}}{0.0170} = \frac{-0.0016}{0.0170} = -0.0920$$

$$p_{54} = \frac{\begin{vmatrix} 1 & r_{12} & r_{13} & r_{15} \\ r_{12} & 1 & r_{23} & r_{25} \\ r_{13} & r_{32} & 1 & r_{35} \\ r_{14} & r_{42} & r_{34} & r_{45} \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.681 & 0.756 & 0.704 \\ 0.681 & 1 & 0.920 & 0.951 \\ 0.756 & 0.920 & 1 & 0.878 \\ 0.773 & 0.781 & 0.830 & 0.767 \end{vmatrix}}{0.0170} = \frac{0.0002}{0.0170} = 0.0136$$

Por tanto, se tiene: $p_{51} = 0.1212$, $p_{52} = 0.9425$, $p_{53} = -0.0920$ y $p_{54} = 0.0136$.

Quinto sistema:

$$0.946 = p_{61} + 0.756\, p_{63} + 0.773\, p_{64} + 0.704\, p_{65}$$

$$0.758 = 0.756\, p_{61} + p_{63} + 0.830\, p_{64} + 0.878\, p_{65}$$

$$0.798 = 0.773\, p_{61} + 0.830\, p_{63} + p_{64} + 0.767\, p_{65}$$

$$0.700 = 0.704\, p_{61} + 0.878\, p_{63} + 0.767\, p_{64} + p_{65}$$

Se procede igual que en los casos anteriores:

$$p = \begin{vmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & 1 & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & 1 & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{43} & 1 & r_{45} \\ r_{15} & r_{25} & r_{53} & r_{54} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0.681 & 0.756 & 0.773 & 0.704 \\ 0.681 & 1 & 0.920 & 0.781 & 0.951 \\ 0.756 & 0.920 & 1 & 0.830 & 0.878 \\ 0.773 & 0.781 & 0.830 & 1 & 0.767 \\ 0.704 & 0.951 & 0.878 & 0.767 & 1 \end{vmatrix} = 0.0015$$

$$p_{61} = \frac{\begin{vmatrix} r_{16} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{26} & 1 & r_{23} & r_{24} & r_{25} \\ r_{36} & r_{32} & 1 & r_{34} & r_{35} \\ r_{46} & r_{42} & r_{43} & 1 & r_{45} \\ r_{56} & r_{52} & r_{53} & r_{54} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 0.946 & 0.681 & 0.756 & 0.776 & 0.704 \\ 0.682 & 1 & 0.920 & 0.781 & 0.951 \\ 0.758 & 0.920 & 1 & 0.830 & 0.878 \\ 0.798 & 0.781 & 0.830 & 1 & 0.767 \\ 0.700 & 0.951 & 0.878 & 0.767 & 1 \end{vmatrix}}{0.0015} = \frac{0.0012}{0.0015} = 0.8111$$

$$p_{63} = \frac{\begin{vmatrix} 1 & r_{12} & r_{16} & r_{14} & r_{15} \\ r_{12} & 1 & r_{26} & r_{24} & r_{25} \\ r_{13} & r_{32} & r_{36} & r_{34} & r_{35} \\ r_{14} & r_{42} & r_{46} & 1 & r_{45} \\ r_{15} & r_{52} & r_{56} & r_{54} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.681 & 0.946 & 0.776 & 0.704 \\ 0.681 & 1 & 0.682 & 0.781 & 0.951 \\ 0.756 & 0.920 & 0.758 & 0.830 & 0.878 \\ 0.773 & 0.781 & 0.798 & 1 & 0.767 \\ 0.704 & 0.951 & 0.700 & 0.767 & 1 \end{vmatrix}}{0.0015} = \frac{6.996E-05}{0.0015} = 0.0465$$

$$p_{64} = \frac{\begin{vmatrix} 1 & r_{12} & r_{13} & r_{16} & r_{15} \\ r_{12} & 1 & r_{23} & r_{26} & r_{25} \\ r_{13} & r_{32} & 1 & r_{36} & r_{35} \\ r_{14} & r_{42} & r_{43} & r_{46} & r_{45} \\ r_{15} & r_{52} & r_{53} & r_{56} & 1 \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.681 & 0.756 & 0.946 & 0.704 \\ 0.681 & 1 & 0.920 & 0.682 & 0.951 \\ 0.756 & 0.920 & 1 & 0.758 & 0.878 \\ 0.773 & 0.781 & 0.830 & 0.798 & 0.767 \\ 0.704 & 0.951 & 0.878 & 0.700 & 1 \end{vmatrix}}{0.0015} = \frac{0.0002}{0.0015} = 0.1646$$

$$p_{65} = \frac{\begin{vmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{16} \\ r_{12} & 1 & r_{23} & r_{24} & r_{26} \\ r_{13} & r_{32} & 1 & r_{34} & r_{36} \\ r_{14} & r_{42} & r_{43} & 1 & r_{46} \\ r_{15} & r_{52} & r_{53} & r_{54} & r_{56} \end{vmatrix}}{p} = \frac{\begin{vmatrix} 1 & 0.681 & 0.756 & 0.776 & 0.946 \\ 0.681 & 1 & 0.920 & 0.781 & 0.682 \\ 0.756 & 0.920 & 1 & 0.830 & 0.758 \\ 0.773 & 0.781 & 0.830 & 1 & 0.798 \\ 0.704 & 0.951 & 0.878 & 0.767 & 0.700 \end{vmatrix}}{0.0015} = \frac{2.456E-05}{0.0015} = 0.0163$$

Por tanto, se tiene: $p_{61} = 0.8111$, $p_{63} = 0.0465$, $p_{64} = 0.1646$ y $p_{65} = 0.0163$.

De igual manera que el primer sistema, el sexto y séptimo sistemas también están resueltos:

$p_{76} = 0.167$

$p_{87} = 1.000$

- Continuando con este desarrollo, se hallará los coeficientes de Wright de las variables residuales o errores, y luego se formará el sistema de ecuaciones para dibujar el diagrama de senderos completo.

Primera ecuación:

$$p_{2e} = \sqrt{1 - p_{21}r_{21}} = \sqrt{1 - 0.681(0.681)} = 0.732$$

Segunda ecuación:

$$p_{3e} = \sqrt{1 - p_{31}r_{31} - p_{32}r_{32}} = \sqrt{1 - (0.2414)(0.756) - (0.7556)(0.920)} = 0.350$$

Tercera ecuación:

$$p_{4e} = \sqrt{1 - p_{41}r_{41} - p_{42}r_{42} - p_{43}r_{43}}$$

$$= \sqrt{1 - (0.3446)(0.773) - (0.1459)(0.781) - (0.4353)(0.830)} = 0.508$$

Cuarta ecuación:

$$p_{5e} = \sqrt{1 - p_{51}r_{51} - p_{52}r_{52} - p_{53}r_{53} - p_{54}r_{54}}$$

$$= \sqrt{1 - (0.1212)(0.704) - (0.9425)(0.951) - (-0.0920)(0.878) - (0.0136)(0.767)} = 0.298$$

Quinta ecuación:

$$p_{6e} = \sqrt{1 - p_{61}r_{61} - p_{63}r_{63} - p_{64}r_{64} - p_{65}r_{65}}$$

$$= \sqrt{1 - (0.8111)(0.946) - (0.0465)(0.758) - (0.1646)(0.798) - (0.0163)(0.700)} = 0.234$$

Sexta ecuación:

$$p_{7e} = \sqrt{1 - p_{76} r_{76}} = \sqrt{1 - (0.167)(0.167)} = 0.986$$

Séptima ecuación:

$$p_{8e} = \sqrt{1 - p_{87} r_{87}} = \sqrt{1 - (1.000)(1.000)} = 0$$

Por lo tanto, el sistema de ecuaciones del modelo es:

$x_2 = 0.681\ x_1 + 0.732\ e_2$

$x_3 = 0.241\ x_1 + 0.756\ x_2 + 0.350\ e_3$

$x_4 = 0.345\ x_1 + 0.146\ x_2 + 0.435\ x_3 + 0.508\ e_4$

$x_5 = 0.121\ x_1 + 0.943\ x_2 - 0.092\ x_3 + 0.014\ x_4 + 0.298\ e_5$

$x_6 = 0.811\ x_1 + 0.047\ x_3 + 0.165\ x_4 + 0.016\ x_5 + 0.234\ e_6$

$x_7 = 0.167\ x_6 + 0.986\ e_7$

$x_8 = 1.000\ x_7$

En consecuencia, el diagrama de senderos completo es el siguiente:

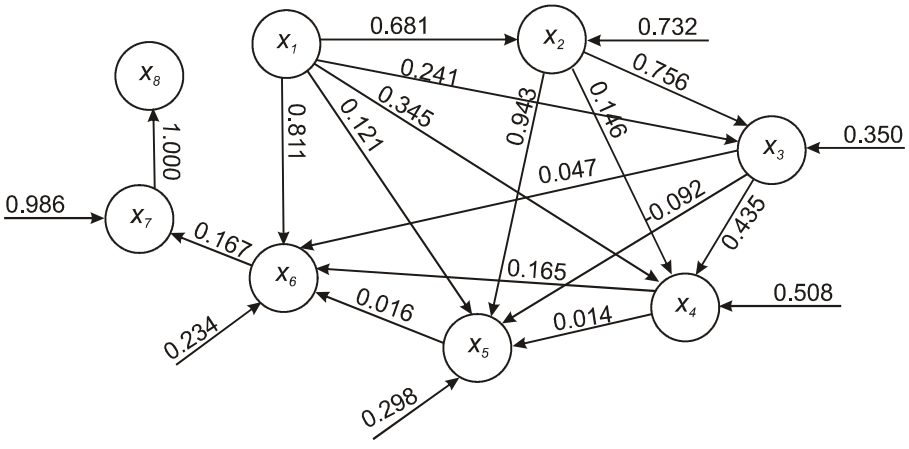

GRAFICO 4.3 Diagrama de senderos completo.

- Se hallará los efectos indirectos de las variables del sistema, para lo cual se usará la fórmula general: Efecto indirecto de x_i sobre $x_j = r_{ij} – p_{ij}$.

Efecto indirecto de x_1 sobre $x_2 = 0.681 – 0.681 = 0$

Efecto indirecto de x_1 sobre $x_3 = 0.756 – 0.241 = 0.515$

Efecto indirecto de x_1 sobre $x_4 = 0.773 – 0.345 = 0.428$

Efecto indirecto de x_1 sobre $x_5 = 0.704 – 0.121 = 0.583$

Efecto indirecto de x_1 sobre $x_6 = 0.946 – 0.811 = 0.135$

Efecto indirecto de x_2 sobre $x_3 = 0.920 – 0.756 = 0.164$

Efecto indirecto de x_2 sobre $x_4 = 0.781 – 0.146 = 0.635$

Efecto indirecto de x_2 sobre $x_5 = 0.951 – 0.943 = 0.008$

Efecto indirecto de x_3 sobre $x_4 = 0.830 – 0.435 = 0.395$

Efecto indirecto de x_3 sobre $x_5 = 0.878 – (-0.092) = 0.970$

Efecto indirecto de x_3 sobre $x_6 = 0.758 – 0.047 = 0.711$

Efecto indirecto de x_4 sobre $x_5 = 0.767 – 0.014 = 0.753$

Efecto indirecto de x_4 sobre $x_6 = 0.798 – 0.165 = 0.633$

Efecto indirecto de x_5 sobre $x_6 = 0.700 – 0.016 = 0.684$

Efecto indirecto de x_6 sobre $x_7 = 0.097 – 0.167 = -0.070$

Efecto indirecto de x_7 sobre $x_8 = 1.000 – 1.000 = 0.000$

- Interpretación de los resultados obtenidos:

Es importante mencionar que en el caso de las variables x_4, Interés, y la variable x_5, Infraestructura existe una mayor dimensión de los coeficientes de correlación que la de los coeficientes de Wright ($r_{54} = 0.767$, $p_{54} = 0.014$); esto

significa que mientras que la asociación total entre dichas variables es importante, 0.767, cuando se controlan las demás variables dicha asociación se vuelve insignificante y disminuye a 0.014.

Sucede lo mismo, entre las parejas de variables x_3 y x_6, Volumen Global de Trabajo e Importancia, en donde se tiene $r_{63} = 0.758$ y $p_{63} = 0.047$; y en x_5 y x_6, Infraestructura e Importancia, donde $r_{65} = 0.700$ y $p_{65} = 0.016$.

Se tiene también un caso particular entre las variables x_3 y x_5, Volumen Global de Trabajo e Infraestructura, en donde a un r_{53} positivo de 0.878 le corresponde un p_{53} negativo de –0.092. Esto quiere decir que la asociación positiva entre las variables es falsa y debido a la fuerte influencia por parte de las otras variables del modelo, se obtiene realmente una asociación negativa cuando se controlan estas variables.

4.3. APLICACIÓN DEL ANÁLISIS FACTORIAL

Se realiza un análisis factorial para determinar qué variables están altamente relacionadas con un factor.

De entre los resultados que se obtuvieron, como se podrá observar, se destacan dos factores claramente diferenciados, al primer factor se lo llamará RENDIMIENTO y al segundo factor, AFLUENCIA.

Una de las razones por lo cual se utiliza el análisis factorial, es para descubrir agrupaciones de variables de tal manera que las variables de cada grupo estén altamente correlacionadas, y los grupos están relativamente incorrelacionados. De esta manera se consigue reducir el número de variables incorrelacionadas a un número menor de factores no correlacionados, que permiten explicar la mayor parte de variabilidad de cada una de las variables. El análisis que se presenta a continuación, se basa en el estudio de la matriz de correlaciones de las variables observadas y su propósito es interpretar esta matriz a partir del menor número de factores posibles.

- Variables

Se usarán las ocho variables que se encuentran explicadas de manera detallada al inicio de este capítulo: Nuevos Conocimientos, Carga Lectiva, Volumen Global de Trabajo, Interés, Infraestructura, Importancia, Matriculados, e Índice de Participación.

- Matriz de correlaciones

Como se sabe, el análisis factorial se basa en la interpretación de la matriz de correlaciones. En la tabla 4.3 aparece la matriz de correlación de todas las variables analizadas, que indica las correlaciones lineales de cada pareja de variable.

Se observa que los coeficientes de correlación son altos y el determinante de esta matriz es 3,676E-10, muy cercano a cero, lo cual es un indicio de que estas variables están altamente correlacionadas entre ellas, y que el análisis factorial, en principio es adecuado.

Matriz de correlación[a]

		Nuevos Conocimientos	Carga Lectiva	Volumen Global de Trabajo	Interés	Infraestructura	Importancia	Matriculados	Indice de Participación
Correlación	Nuevos Conocimientos	1,000	,681	,756	,773	,704	,946	,094	,094
	Carga Lectiva	,681	1,000	,920	,781	,951	,682	,149	,149
	Volumen Global de Trabajo	,756	,920	1,000	,830	,878	,758	,074	,074
	Interés	,773	,781	,830	1,000	,767	,798	,121	,122
	Infraestructura	,704	,951	,878	,767	1,000	,700	,097	,097
	Importancia	,946	,682	,758	,798	,700	1,000	,167	,167
	Matriculados	,094	,149	,074	,121	,097	,167	1,000	1,000
	Indice de Participación	,094	,149	,074	,122	,097	,167	1,000	1,000

a. Determinante = 3,676E-10

TABLA 4.3 Matriz de correlación.

Otro indicador de la magnitud de la relación lineal entre las variables es el índice KMO (Kaiser-Meyer-Olkin) y el contraste de esfericidad de Bartlett, que es muy usado para probar si la matriz de correlaciones es la identidad. El índice KMO se considera aceptable si se encuentra dentro del intervalo 0.7 a

0.8, y cuando la prueba de esfericidad da valores ji-cuadrado altos y significaciones casi nulas, se puede rechazar la hipótesis de que la matriz de correlaciones sea la identidad. En la tabla 4.4 se aprecia que el índice KMO es igual a 0.771, el valor ji-cuadrado es 4399.121 y la significación es de 0.000.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,771
Prueba de esfericidad de Bartlett.	Chi-cuadrado aproximado	4399,121
	gl	28
	Sig.	,000

TABLA 4.4 KMO y prueba de Bartlett.

Continuando con este análisis, se tiene en la tabla 4.5, las matrices anti-imagen[39] de covarianzas y correlaciones entre todas las variables analizadas.

En la diagonal de la matriz de correlaciones se tiene los coeficientes MSA (medida de adecuación muestral) que vienen a ser los KMO, pero en este caso para cada variable por separado. Aquí interesan los valores diagonales cercanos a uno, mientras que el resto de coeficientes es mejor cuanto más pequeños sean.

[39] Matrices Anti-Imagen (en Análisis Factorial): La matriz de correlación anti-imagen contiene los coeficientes de correlación parciales negativos, y la matriz de covarianza anti-imagen contiene las covarianzas parciales negativas. En un buen modelo factorial, la mayoría de los elementos de la diagonal serán pequeños. La medida de muestro adecuada para una variable se despliega en la diagonal de la matriz de correlación anti-imagen.

Matrices Anti-imagen

		Nuevos Conocimientos	Carga Lectiva	Volumen Global de Trabajo	Interés	Infraestructura	Importancia	Matriculados	Indice de Participación
Covarianza Anti-imagen	Nuevos Conocimientos	9,583E-02	2,625E-03	-6,76E-03	6,613E-03	-7,476E-03	-7,448E-02	5,743E-05	-5,726E-05
	Carga Lectiva	2,625E-03	5,597E-02	-4,51E-02	-8,16E-03	-5,399E-02	8,083E-03	-2,249E-05	2,228E-05
	Volumen Global de Trabajo	-6,764E-03	-4,51E-02	,105	-3,37E-02	1,217E-02	-1,090E-02	4,569E-05	-4,546E-05
	Interés	6,613E-03	-8,16E-03	-3,37E-02	,219	-3,821E-03	-4,125E-02	2,417E-04	-2,417E-04
	Infraestructura	-7,476E-03	-5,40E-02	1,217E-02	-3,82E-03	8,607E-02	-4,137E-03	-1,901E-05	1,918E-05
	Importancia	-7,448E-02	8,083E-03	-1,09E-02	-4,13E-02	-4,137E-03	8,229E-02	-1,231E-04	1,229E-04
	Matriculados	5,743E-05	-2,25E-05	4,569E-05	2,417E-04	-1,901E-05	-1,231E-04	3,019E-06	-3,019E-06
	Indice de Participación	-5,726E-05	2,228E-05	-4,55E-05	-2,42E-04	1,918E-05	1,229E-04	-3,019E-06	3,020E-06
Correlación Anti-imagen	Nuevos Conocimientos	,804[a]	3,584E-02	-6,73E-02	4,562E-02	-8,232E-02	-,839	,107	-,106
	Carga Lectiva	3,584E-02	,773[a]	-,588	-7,37E-02	-,778	,119	-5,470E-02	5,419E-02
	Volumen Global de Trabajo	-6,732E-02	-,588	,887[a]	-,222	,128	-,117	8,102E-02	-8,061E-02
	Interés	4,562E-02	-7,37E-02	-,222	,906[a]	-2,781E-02	-,307	,297	-,297
	Infraestructura	-8,232E-02	-,778	,128	-2,78E-02	,837[a]	-4,916E-02	-3,730E-02	3,763E-02
	Importancia	-,839	,119	-,117	-,307	-4,916E-02	,767[a]	-,247	,246
	Matriculados	,107	-5,47E-02	8,102E-02	,297	-3,730E-02	-,247	,482[a]	-1,000
	Indice de Participación	-,106	5,419E-02	-8,06E-02	-,297	3,763E-02	,246	-1,000	,482[a]

a. Medida de adecuación muestral (MSA)

TABLA 4.5 Matrices anti-imagen.

Del análisis realizado hasta este momento, se puede resumir lo siguiente:

Coeficientes de correlación altamente significativos en su mayoría.

El determinante de la matriz de correlaciones muy bajo: 3,676E-10.

El índice KMO 0.771, aceptable.

El resultado de la prueba de Bartlett con un ji-cuadrado = 4399.121 y una significación p = 0.000.

Las MSA bastante altas en la diagonal de la matriz de correlaciones anti-imagen.

Por lo tanto, se podría concluir que el análisis factorial que se describe a continuación resulta adecuado y puede proporcionar buenas conclusiones aceptables.

- Extracción de factores

Todos los factores que se obtienen con el método de componentes principales se indican en la tabla 4.6.

En la sección “Valores propios iniciales” se presenta los valores propios, el porcentaje de la varianza y el porcentaje acumulado de varianza para cada factor, en orden según la magnitud de los valores propios. Se puede ver en los resultados que hay únicamente dos componentes o factores con un valor propio inicial superior a 1. El primer factor explica un 62.89% de la varianza de los datos, mientras que el segundo factor un 24.41% de la varianza, es decir que de manera conjunta pueden explicar un 87.3%, lo cual puede interpretarse como un porcentaje aceptable. Esto implica que, con menos de un 15% de pérdida de información, se puede expresar cada una de las variables como combinación lineal de estos dos factores o variables latentes.

La sección “Sumas de las saturaciones al cuadrado de la extracción” reproduce esta información para el número de factores extraídos en el análisis (dos, en este caso). Se puede apreciar en este caso, que las sumas de las saturaciones al cuadrado son idénticas a los valores propios.

La sección “Sumas de las saturaciones al cuadrado de la rotación” presenta la misma información para los factores rotados. Se puede observar que las sumas de las saturaciones al cuadrado son diferentes de las reportadas en la sección “sumas de las saturaciones al cuadrado de la extracción”, pero que su suma (4.965 + 2.020) es igual a la suma de los valores propios (5.031 + 1.953).

Se puede observar también, que antes y después de la rotación, los dos factores explican 87.304% de la varianza, aunque individualmente el porcentaje de la varianza explicada por cada factor difiere antes y después de la rotación.

Varianza total explicada

Componente	Valores propios iniciales			Sumas de las saturaciones al cuadrado de la extracción			Sumas de las saturaciones al cuadrado de la rotación		
	Total	% de la Varianza	% acumulado	Total	% de la Varianza	% acumulado	Total	% de la Varianza	% acumulado
1	5,031	62,889	62,889	5,031	62,889	62,889	4,965	62,059	62,059
2	1,953	24,414	87,304	1,953	24,414	87,304	2,020	25,245	87,304
3	,601	7,511	94,814						
4	,222	2,772	97,586						
5	,109	1,364	98,950						
6	4,965E-02	,621	99,571						
7	3,433E-02	,429	100,000						
8	1,510E-06	1,887E-05	100,000						

Método de Extracción: Análisis de Componentes Principales.

TABLA 4.6 Varianza total explicada.

En el gráfico 4.3 (Gráfico de sedimentación) se encuentra una representación gráfica de estos resultados, en la abscisa figuran el número total de factores y en la ordenada el valor propio de cada uno de ellos.

Mediante este gráfico es posible determinar el número de factores que mejor representan toda la varianza significativa descrita por la matriz de correlaciones.

Se puede apreciar en el gráfico 4.3 que hay dos factores que explican la principal varianza significativa en la matriz de correlaciones. Ya que, a partir del factor 3 y los subsiguientes factores, la cantidad de la varianza explicada es baja y prácticamente equivalente.

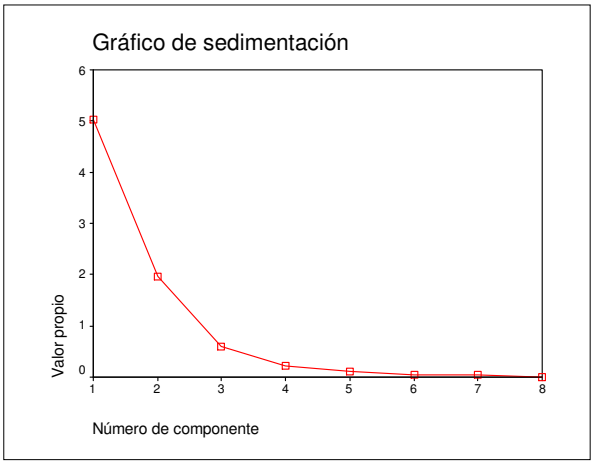

GRÁFICO 4.4 Representación gráfica de los dos primeros factores.

En el análisis factorial lo óptimo es encontrar un modelo en el cual todas las variables saturen en algún factor, es decir, pesos factoriales altos en uno y bajos en el resto. La tabla 4.7 indica los coeficientes usados para expresar cada variable en términos de los dos factores del modelo.

La matriz de componentes de la tabla 4.7 tiene en el primer factor (componente 1) pesos factoriales positivos por parte de todas las variables, mientras que en el segundo factor (componente 2) tiene una combinaciones de pesos factoriales positivos y negativos. Este patrón es usual de los análisis de componentes principales, a menos que haya un número sustancial de correlaciones negativas en la matriz de correlación.

Matriz de componentes[a]

	Componente	
	1	2
Nuevos Conocimientos	,883	-8,88E-02
Carga Lectiva	,919	-4,95E-02
Volumen Global de Trabajo	,935	-,129
Interés	,903	-6,85E-02
Infraestructura	,912	-,100
Importancia	,893	-1,45E-02
Matriculados	,210	,978
Indice de Participación	,210	,978

Método de extracción: Análisis de Componentes Principales.

a. 2 componentes extraídos.

TABLA 4.7 Matriz de componentes.

- Rotación de factores

En el análisis de componentes principales, los factores se extraen en orden según la cantidad de la varianza que explican (tabla 4.7), no en términos de qué tan bien describen realmente las relaciones entre las variables. Por tal motivo, se recomienda rotar los ejes de modo que se acerquen más a los puntos que representan las variables.

El propósito de la rotación es poder interpretar el sentido y significado de los factores extraídos y así, obtener una descripción más simple de las relaciones entre las variables.

Empleando el método de componentes principales, la matriz de pesos factoriales tanto la no rotada (tabla 4.7) como la rotada, con procedimientos ortogonales (Varimax, Quartimax y Equamax), y con procedimientos no ortogonales (Promax y Oblim) son prácticamente las mismas. En la tabla 4.8 se encuentra la matriz rotada con el procedimiento Varimax.

Matriz de componentes rotados[a]

	Componente	
	1	2
Nuevos Conocimientos	,886	4,179E-02
Carga Lectiva	,916	8,608E-02
Volumen Global de Trabajo	,944	9,765E-03
Interés	,903	6,486E-02
Infraestructura	,917	3,490E-02
Importancia	,886	,117
Matriculados	6,416E-02	,998
Indice de Participación	6,421E-02	,998

Método de extracción: Análisis de Componentes Principales.
Método de Rotación: Varimax con Normalización Kaiser.

a. La rotación ha convergido en 3

TABLA 4.8 Matriz de componentes rotados.

A partir de la información que se encuentra en la tabla 4.8, se puede realizar la siguiente asignación:

Rendimiento (Factor 1)		Afluencia (Factor 2)	
Variables	Volumen Global de Trabajo	Variables	Matriculados
	Infraestructura		Índice de Participación
	Carga Lectiva		
	Interés		
	Nuevos Conocimientos		
	Importancia		

TABLA 4.9 Asignación de variables en el factor correspondiente.

Así, se tiene que el primer factor está representado por las seis primeras variables y el segundo factor por las variables Matriculados e Índice de Participación. Los dos factores claramente diferenciados son llamados RENDIMIENTO y AFLUENCIA.

A continuación, en el gráfico 4.4 se observa las posiciones de las variables respecto a los ejes factoriales rotados. Se puede apreciar que todas ellas están bien representadas sobre el plano, ya que están próximas al borde del círculo de radio unidad, y ninguna está cerca del origen.

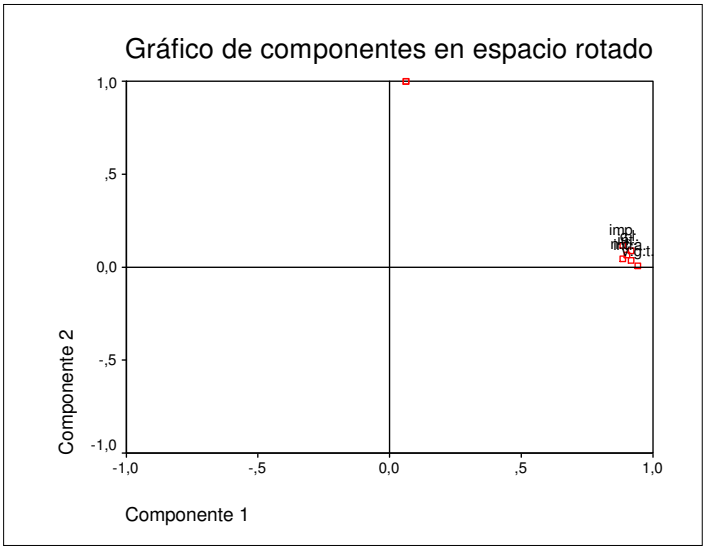

GRÁFICO 4.5 Gráfico de componentes en espacio rotado.

- Puntuaciones factoriales

Luego de realizar un análisis de componentes principales, hay ocasiones en que conviene calcular puntuaciones para representar los factores en cuestión. Dichas puntuaciones se pueden formar a partir de la matriz original de componentes principales o partir de la matriz rotada.

La característica importante de las puntuaciones es que se correlacionan entre sí de la misma forma en que lo hacen los factores correspondientes.

En la tabla 4.10 se encuentra la matriz de coeficientes para obtener las puntuaciones factoriales obtenidas por cada variable. Es decir, los coeficientes que permiten expresar cada factor como combinación lineal de todas las variables.

Como se puede ver en la tabla, los pesos de las seis primeras variables contribuyen en la formación del primer factor (Rendimiento) y las variables Matriculados e Índice de Participación no contribuyen prácticamente en nada (-0.32 de coeficiente en ambos casos). Para la formación del segundo factor (Afluencia) contribuyen los pesos de las dos últimas variables: Matriculados e Índice de Participación (0.501 de coeficiente en ambos casos).

Matriz de coeficientes de las puntuaciones en las componentes

	Componente	
	1	2
Nuevos Conocimientos	,180	-,019
Carga Lectiva	,184	,002
Volumen Global de Trabajo	,194	-,038
Interés	,183	-,008
Infraestructura	,187	-,024
Importancia	,177	,019
Matriculados	-,032	,501
Indice de Participación	-,032	,501

Método de extracción: Análisis de Componentes Principales.
Método de Rotación: Varimax con Normalización Kaiser.

TABLA 4.10 Matriz de coeficientes para el cálculo de las puntuaciones en las componentes.

Se obtuvieron dos factores claramente diferenciados: Rendimiento (Factor 1) que contiene pesos altos de las seis primeras variables.

Y, Afluencia (Factor 2) que contiene pesos altos de las variables Matriculados e Índice de Participación, es decir, estas dos variables tienen una correlación elevada con el Factor 2.

Puesto que se trata de una solución Varimax, se concluye también que los dos factores, Rendimiento y Afluencia, son mutuamente independientes.

4.4. APLICACIÓN DE LOS MODELOS DE ECUACIONES ESTRUCTURALES

El hallazgo de relaciones causales entre las variables objeto de estudio, es uno de los propósitos de las investigaciones empíricas, esto es posible cuando se trabaja con conceptos experimentalmente controlables como los fenómenos físicos, sin embargo sobre las variables analizadas en las ciencias sociales y del comportamiento, no es posible ejercer un control, por lo que es necesario realizar otro tipo de análisis metodológico, llamado en este caso, modelización de ecuaciones estructurales.

Como se expone en la teoría, la modelización según ecuaciones estructurales sigue una metodología que pasa por las siguientes etapas: especificación, identificación, estimación de parámetros, evaluación del ajuste, reespecificacón del modelo (en el caso de ser necesario) e interpretación de los resultados.

Para llevar a cabo la aplicación se realiza una modelización con las ocho variables que se ha venido usando en los anteriores análisis: Nuevos Conocimientos, Carga Lectiva, Volumen Global de Trabajo, Interés, Infraestructura, Importancia, Matriculados, e Índice de Participación.

- **Especificación del modelo**

  El diagrama de senderos del modelo de ecuaciones estructurales, en donde se encuentran las variables latentes incorporadas se muestra en el gráfico 4.5.

  En el diagrama de senderos, el modelo planteado presenta tres variables latentes y ocho indicadores. NC es variable latente exógena y, RENDIMIENTO y AFLUENCIA son variables latentes endógenas. Los ocho indicadores corresponden a las ocho variables que sirven para medir las variables latentes consideradas.

Por tanto se tiene, 7 variables observadas endógenas y 1 variable observada exógena.

Se considera también los errores de medida: 7 de las variables observadas endógenas y 1 de la variable observada exógena.

Por último, se considera también 2 términos de perturbación de las variables latentes endógenas.

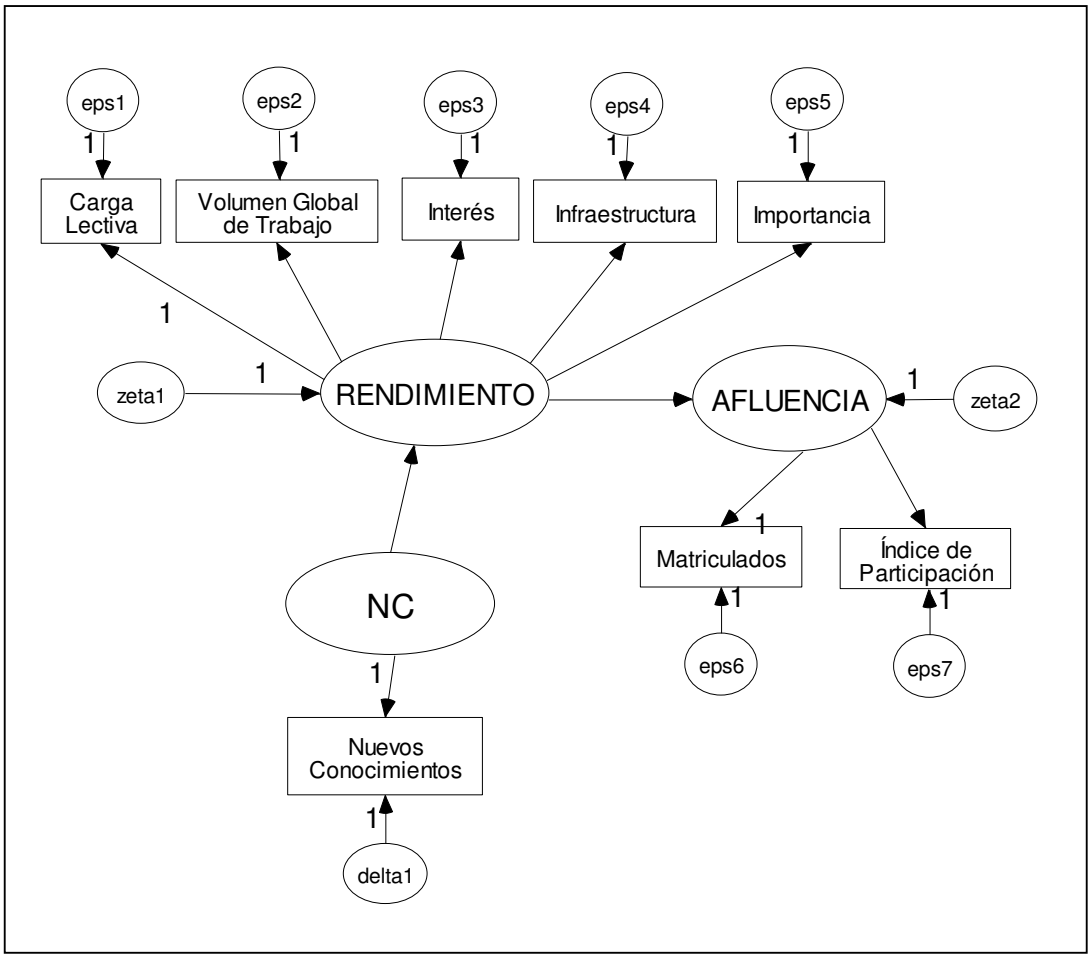

GRÁFICO 4.6 Diagrama de senderos.

El siguiente paso es traducir el diagrama de senderos a ecuaciones estructurales, distinguiendo tanto el modelo de medida como el modelo estructural.

Las ecuaciones estructurales para el modelo de medida del diagrama de senderos que se presenta en el gráfico 4.5 son:

$$\text{Nuevos Conocimientos} = \lambda_1^{x}\ \text{NC} + \delta_1$$

$$\text{Carga Lectiva} = \lambda_1^{y}\ \text{RENDIMIENTO} + \varepsilon_1$$

$$\text{Volumen Global de Trabajo} = \lambda_2^{y}\ \text{RENDIMIENTO} + \varepsilon_2$$

$$\text{Interés} = \lambda_3^{y}\ \text{RENDIMIENTO} + \varepsilon_3$$

$$\text{Infraestructura} = \lambda_4^{y}\ \text{RENDIMIENTO} + \varepsilon_4$$

$$\text{Importancia} = \lambda_5^{y}\ \text{RENDIMIENTO} + \varepsilon_5$$

$$\text{Matriculados} = \lambda_6^{y}\ \text{AFLUENCIA} + \varepsilon_6$$

$$\text{Índice de Participación} = \lambda_7^{y}\ \text{AFLUENCIA} + \varepsilon_7$$

Las ecuaciones estructurales para el modelo estructural del diagrama de senderos que se presenta en el gráfico 4.5 son:

$$\text{RENDIMIENTO} = \gamma_1\ \text{NC} + \zeta_1$$

$$\text{AFLUENCIA} = \beta_1\ \text{RENDIMIENTO} + \zeta_2$$

- **Identificación del modelo**

La siguiente fase es comprobar si la matriz de entrada permite obtener estimaciones únicas de los parámetros no conocidos. Utilizando la regla del conteo, se comprobará la condición necesaria que viene dada por la siguiente expresión:

$$t \leq \tfrac{1}{2}\,(p+q)(p+q+1)$$

$$t = 1(\lambda_x) + 7(\lambda_y) + 1(\gamma) + 1(\beta) + 1(\delta) + 7(\varepsilon) + 2(\zeta) = 20$$

$$\tfrac{1}{2}\,(7+1)(7+1+1) = 72$$

Como, $20 \leq 72$, el modelo cumple la condición necesaria para estar identificado.

Se tiene que el modelo a estudiarse es recursivo, esto es, cuando ninguna variable en el modelo tiene efecto sobre sí mismo. Es decir, en el diagrama de senderos del modelo, no es posible empezar en cualquier variable y, siguiendo el camino de las flechas de una cabeza, se regresa a la misma variable. Debido a que el modelo es recursivo, se cumple la condición suficiente de identificación del modelo.

- **Estimación del modelo**

El método de estimación a utilizar para estimar el modelo es el de máxima verosimilitud.

La siguiente tabla 4.11 contiene los pesos de la regresión: se estima los pesos de la regresión, el error estándar aproximado (S.E.) y la proporción crítica (C.R.), que es la estimación del parámetro dividida por una estimación de su error estándar.

Si las suposiciones apropiadas de la distribución se reúnen, la proporción crítica tiene una distribución normal bajo la hipótesis nula de que el parámetro tiene un valor de la población de cero. Por ejemplo, si una estimación tiene una proporción crítica mayor que dos (en valor absoluto), la estimación está significativamente diferente del cero en el nivel 0.05. Incluso sin las suposiciones para la distribución, las proporciones críticas tienen la interpretación siguiente: para cualquier parámetro no forzado, el cuadrado de su proporción crítica es, aproximadamente, la cantidad por la que la estadística del ji-cuadrado aumentaría si el análisis se repitiera con ese parámetro cambiado a cero.

Estimaciones de Máxima Verosimilitud

Pesos de la Regresión:

			Estimación	S.E.	C.R.
RENDIMIENTO	<--	NC	0,614	0,056	10,876
AFLUENCIA	<--	RENDIMIENTO	0,705	0,054	13,163
Índice de Participación	<--	AFLUENCIA	1,000		
Matriculados	<--	AFLUENCIA	25,218	0,505	21,243
Importancia	<--	RENDIMIENTO	0,888	0,041	21,413
Infraestructura	<--	RENDIMIENTO	1,000		
Interés	<--	RENDIMIENTO	1,054	0,029	36,588
Volumen Global de Trabajo	<--	RENDIMIENTO	3,390	0,084	40,269
Carga Lectiva	<--	RENDIMIENTO	0,472	0,012	38,262
Nuevos Conocimientos	<--	NC	5,331	0,430	12,403

TABLA 4.11 Pesos de la regresión.

La estandarización de los pesos de la regresión, que es la estandarización de los pesos estimados de la regresión, figura en la tabla 4.12.

Estandarización de los Pesos de la Regresión:

			Estimación
RENDIMIENTO	<--	NC	0,567
AFLUENCIA	<--	RENDIMIENTO	0,663
Índice de Participación	<--	AFLUENCIA	0,840
Matriculados	<--	AFLUENCIA	0,996
Importancia	<--	RENDIMIENTO	0,812
Infraestructura	<--	RENDIMIENTO	0,963
Interés	<--	RENDIMIENTO	0,955
Volumen Global de Trabajo	<--	RENDIMIENTO	0,968
Carga Lectiva	<--	RENDIMIENTO	0,961
Nuevos Conocimientos	<--	NC	0,649

TABLA 4.12 Estandarización de los pesos de la regresión.

La tabla 4.13 despliega las varianzas: la estimación de las varianzas entre las variables exógenas, el error estándar aproximado (S.E.) y la proporción crítica (C.R.).

Varianzas:

	Estimación	S.E.	C.R.
NC	6,656	0,640	10,398
zeta1	5,301	0,472	11,230
zeta2	3,737	0,387	9,653
eps1	4,010	0,343	11,700
eps2	3,187	0,271	11,757
eps3	3,696	0,373	9,908
eps4	3,622	0,292	12,414
eps5	4,073	0,308	11,761
eps6	19,047	52,022	8,861
eps7	0,939	0,092	10,194
delta1	2.944	0.499	5.900

TABLA 4.13 Varianzas.

El siguiente paso es determinar los Efectos Totales, es decir, el efecto total de cada columna de variables en cada fila de variables. Posteriormente se obtienen los Efectos Totales Estandarizados, es decir, el efecto total de cada columna de variables en cada fila de variables después de estandarizar todas las variables. Los resultados se presentan en la tabla 4.14.

Estimaciones – Efectos Totales

	NC	RENDIMIENTO	AFLUENCIA
RENDIMIENTO	0,614	0,000	0,000
AFLUENCIA	0,000	0,705	0,000
Nuevos Conocimientos	0,331	0,000	0,000
Índice de Participación	0,000	0,598	0,849
Matriculados	0,000	0,705	1,000
Importancia	0,615	1,016	0,000
Infraestructura	0,507	1,000	0,000
Interés	0,546	1,054	0,000
Volumen Global de Trabajo	0,614	3,390	0,000
Carga Lectiva	0,586	0,472	0,000

Estimaciones – Efectos Totales Estandarizados

	NC	RENDIMIENTO	AFLUENCIA
RENDIMIENTO	0,567	0,000	0,000
AFLUENCIA	0,000	0,663	0,000
Nuevos Conocimientos	0,649	0,000	0,000
Índice de Participación	0,000	0,529	0,798
Matriculados	0,000	0,557	0,840
Importancia	0,451	0,789	0,000
Infraestructura	0,408	0,757	0,000
Interés	0,460	0,812	0,000
Volumen Global de Trabajo	0,461	0,813	0,000
Carga Lectiva	0,515	0,873	0,000

TABLA 4.14 Efectos totales, efectos totales estandarizados.

Los Efectos Directos, es decir, el efecto directo de cada columna de variables en cada fila de variables, y los Efectos Directos Estandarizados, que es, el efecto directo de cada columna de variables en cada fila de variables después de estandarizar todas las variables, se muestran en la siguiente tabla 4.15.

Estimación – Efectos Directos

	NC	RENDIMIENTO	AFLUENCIA
RENDIMIENTO	0,614	0,000	0,000
AFLUENCIA	0,000	0,705	0,000
Nuevos Conocimientos	0,331	0,000	0,000
Índice de Participación	0,000	0,000	0,849
Matriculados	0,000	0,000	1,000
Importancia	0,000	1,016	0,000
Infraestructura	0,000	1,000	0,000
Interés	0,000	1,054	0,000
Volumen Global de Trabajo	0,000	3,390	0,000
Carga Lectiva	0,000	0,472	0,000

Estimación – Efectos Directos Estandarizados

	NC	RENDIMIENTO	AFLUENCIA
RENDIMIENTO	0,567	0,000	0,000
AFLUENCIA	0,000	0,663	0,000
Nuevos Conocimientos	0,649	0,000	0,000
Índice de Participación	0,000	0,000	0,798
Matriculados	0,000	0,000	0,840
Importancia	0,000	0,789	0,000
Infraestructura	0,000	0,757	0,000
Interés	0,000	0,812	0,000
Volumen Global de Trabajo	0,000	0,813	0,000

TABLA 4.15 Efectos directos, efectos directos estandarizados.

La siguiente tabla 4.16, indica los Efectos Indirectos, es decir, el efecto indirecto de cada columna de variables en cada fila de variables. A continuación se encuentran los Efectos Indirectos estandarizados, que es, el efecto indirecto de cada columna de variables en cada fila de variables después de estandarizar todas las variables. Como se puede ver claramente, todos los resultados son nulos, para este caso.

Estimación – Efectos Indirectos

	NC	RENDIMIENTO	AFLUENCIA
RENDIMIENTO	0,000	0,000	0,000
AFLUENCIA	0,000	0,000	0,000
Nuevos Conocimientos	0,000	0,000	0,000
Índice de Participación	0,000	0,000	0,000
Matriculados	0,000	0,000	0,000
Importancia	0,615	1,016	0,000
Infraestructura	0,507	1,000	0,000
Interés	0,546	1,054	0,000
Volumen Global de Trabajo	0,614	3,390	0,000
Carga Lectiva	0,586	0,472	0,000

Estimación – Efectos Indirectos Estandarizados

	NC	RENDIMIENTO	AFLUENCIA
RENDIMIENTO	0,000	0,000	0,000
AFLUENCIA	0,000	0,000	0,000
Nuevos Conocimientos	0,000	0,000	0,000
Índice de Participación	0,000	0,000	0,000
Matriculados	0,000	0,000	0,000
Importancia	0,451	0,789	0,000
Infraestructura	0,408	0,757	0,000
Interés	0,460	0,812	0,000
Volumen Global de Trabajo	0,461	0,813	0,000
Carga Lectiva	0,515	0,873	0,000

TABLA 4.16 Efectos indirectos, efectos indirectos estandarizados.

Por último, se realiza el cálculo del mínimo valor, $\hat{C}$, de la función de discrepancia (CMIN). Se minimiza la función de discrepancia (Browne, 1982, 1984) de la forma:

$$C(\alpha, a) = [N - r]\left(\frac{\sum_{g=1}^{G} N^{(g)} f\left(\mu^{(g)}, \Sigma^{(g)}; \bar{x}^{(g)}, S^{(g)}\right)}{N}\right) = [N - r]F(\alpha, a).$$

Además, se presenta el número de grados de libertad (DF) para probar el modelo, df = d = p-q, donde p es el número de momentos muestrales y q es el

número de parámetros distintos, y el valor de P que es la probabilidad de conseguir una discrepancia tan grande como ocurra con la muestra presente (bajo una apropiada suposición y suponiendo un modelo correctamente especificado). Es decir, P es un "valor p" por probar la hipótesis que el modelo se ajuste perfectamente en la población.

La tabla 4.17 presenta el resumen de los modelos con los datos antes descritos:

Resumen de los modelos

	CMIN	DF	P	CMIN/DF
El modelo	71,544	18	0,000	3,975
Modelo Saturado	0,000	0		
Modelo de Independencia	2131,790	18	0,000	101,514

TABLA 4.17 Resumen de los modelos.

- **Evaluación del modelo**

Una vez identificado y estimado el modelo, se procede a evaluarlo para determinar si los datos se ajustan al modelo propuesto, mediante un ajuste global del modelo.

En la tabla 4.18 se despliega los valores resultantes de los índices absolutos de ajuste: GFI, NCP, RMSEA y PCLOSE.

En la tabla 4.19 se encuentran los valores resultantes de los índices incrementales de ajuste: AGFI, NFI, RFI, IFI, TLI (NNFI) y CFI.

Y, en la tabla 4.20 se muestra los valores resultantes de los índices de ajuste de parsimonia: PNFI, PGFI, AIC, CAIC y CN.

Medida de Ajuste	El modelo	Saturado	Independencia	Macro
GFI	0,975	1,000	0,494	GFI
Noncentrality parameter estimate	35,544	0,000	1116,790	NCP
NCP límite inferior	21,936	0,000	868,786	NCPLO
NCP límite superior	56,603	0,000	1272,133	NCPHI
RMSEA	0,108		0,389	RMSEA
RMSEA límite inferior	0,087		0,375	RMSEALO
RMSEA límite superior	0,132		0,403	RMSEAHI
Valor P	0,000		0,000	PCLOSE

TABLA 4.18 Índices absolutos de ajuste.

En cuanto a los índices absolutos de ajuste, el modelo propuesto, presenta de manera general un buen ajuste, ya que, se tiene valores de GFI, RMSEA y PCLOSE dentro de los límites de aceptación establecidos, que son: para GFI valores superiores a 0.95 (resultado 0.975), RMSEA inferiores a 0.08 (el resultado de RMSEA, evaluado al límite superior del intervalo de confianza al 90%, 0.087, es el mejor valor obtenido en este índice, pero los otros dos valores de 0.108 y 0.132 de RMSEA y RMSEA evaluado al límite inferior del intervalo de confianza al 90%, respectivamente, no presentan un desvío muy significativo del límite de aceptación y PCLOSE que prueba que el RMSEA sea menor o igual a 0.05 tiene un valor de 0.000.

Los valores resultantes del NCP no son cercanos a cero, lo que implicaría que existen diferencias entre la matriz de observaciones y la matriz estimada, pero se debe considerar que este índice considera el tamaño muestral y los grados de libertad y éstos valores son quizá la respuesta a que se obtengan estos resultados.

Medida de Ajuste	El modelo	Saturado	Independencia	Macro
Adjusted GFI	0,913		0,292	AGFI
Normed fit index	0,966	1,000	0,000	NFI
Relative fit index	0,916		0,000	RFI
Incremental fit index	0,969	1,000	0,000	IFI
Tucker-Lewis index	0,923		0,000	TLI
Comparative fit index	0,969	1,000	0,000	CFI

TABLA 4.19 Índices incrementales de ajuste.

En cuanto a los índices incrementales de ajuste obtenidos, se puede ver que todos los valores resultantes superan los límites de aceptación, el 0.90 para el caso del AGFI, el NFI y el TLI, y valores muy próximos a la unidad para el caso del RFI, el IFI y el CFI. Lo que de manera general indica un muy buen ajuste del modelo.

Medida de Ajuste	El modelo	Saturado	Independencia	Macro
Parsimony-adjusted NFI	0,387	0,000	0,000	PNFI
Parsimony-adjusted GFI	0,279		0,353	PGFI
Akaike information criterion (AIC)	101,544	42,000	2143,790	AIC
Consistent AIC	189,104	164,584	2178,814	CAIC
Hoelter .05 index	164,000		11,000	HFIVE
Hoelter .01 index	219,000		14,000	HONE

TABLA 4.20 Índices de ajuste de parsimonia.

Los índices de ajuste de parsimonia son los que tiene más sentido dentro de una estrategia de modelización competitiva por ofrecer medidas del ajuste del modelo por coeficiente estimado. De manera general todos los valores resultantes obtenidos están dentro de los límites esperados.

- **Interpretación y modificación del modelo**

El último paso, una vez demostrada la adecuación del modelo a los datos, consistirá en interpretar dicho modelo de acuerdo con la literatura al respecto en que se ha basado su especificación. Antes de realizar la interpretación se ha de comprobar que el modelo no tiene capacidad de mejora, pues en caso contrario habrá que plantear las modificaciones oportunas.

Debido a que todos los índices de bondad de ajuste anteriormente evaluados presentaban valores muy buenos, las posibilidades de modificación del modelo son escasas. Esto se comprueba con el hecho de que el programa AMOS no presenta ningún índice de modificación para el modelo.

CAPÍTULO 5
CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

- El análisis de senderos es una herramienta muy importante y poderosa para identificar variables de efecto - causa dentro de un modelo general propuesto; además, permite la interpretación de las relaciones entre variables.

- La idea principal del análisis factorial es que la medida de cualquier variable se compone potencialmente de una serie de variables latentes. Por lo tanto, el objetivo del análisis factorial es revelar esas variables latentes subyacentes, también llamadas factores.

  El análisis factorial se diferencia de otros análisis estadísticos en que no se ocupa de la manera en que alguna variable de tratamiento influya en cierta conducta; el análisis factorial se ocupa de las relaciones entre variables.

- Los modelos de ecuaciones estructurales son la unión del análisis de senderos y el análisis factorial. Estos modelos constan de dos partes: el modelo de medida y el modelo estructural. El modelo de medida se lo estima primero, y la matriz de covarianza resultante entre los factores sirve para luego estimar los coeficientes estructurales entre las variables latentes.

- Mediante una aplicación de toda la teoría expuesta en el presente proyecto de titulación, se vuelve más comprensible la metodología de estudio ya que se puede observar el desarrollo de cada proceso para obtener un modelo que se ajuste a los datos estudiados.

- En la aplicación del análisis de senderos se obtienen las relaciones de causalidad entre las ocho variables a estudiarse, se establece el diagrama de senderos respectivo y se resuelve el sistema de ecuaciones estructurales con el fin de determinar los coeficientes de Wright.

- A través de la aplicación del análisis factorial, se obtuvieron dos factores: Motivación y Presión, que explican cómo se correlacionan las variables objeto de estudio.

- Mediante la aplicación de los modelos de ecuaciones estructurales, se confirma que el modelo propuesto se ajusta bien a los datos, debido a que los índices de ajuste presentan buenos resultados.

5.2. RECOMENDACIONES

- Debido al poder del análisis para identificar variables de efecto - causa dentro de un modelo, se recomienda el uso de esta herramienta para estudiarr problemas de tipo social, psicológicos, ambientales, etc. Actualmente este método es muy útil, debido a la creciente demanda que tiene la sociedad por resolver problemas cotidianos importantes, por lo tanto no se debe olvidar la utilidad que puede proporcionarnos el análisis de senderos para resolver dichos problemas.

- Cuando se desee investigar las relaciones entre un conjunto de variables, es ideal utilizar el análisis factorial, ya que de esta manera se obtienen todas las relaciones posibles existentes en un grupo de variables. Otra ventaja del análisis factorial, razón por la cual se recomienda su uso, radica en el hecho de que, cuando no se conocen los factores "a priori" de un modelo, es factible realizar un análisis factorial, del mismo modo que cuando se propone "a priori" un modelo, según el cual hay unos factores que representan a las variables originales.

- El poder que tiene la modelización de ecuaciones estructurales se basa en el hecho de que se utiliza la teoría del análisis de senderos y del análisis factorial, por tal motivo, los modelos de ecuaciones estructurales es la herramienta más adecuada para resolver problemas de tipo social, científico, entre otras, se recomienda utilizarlo con la finalidad de describir las relaciones

entre variables, comprendiendo el rol que tienen las relaciones causales dentro de una análisis estadístico.

- Cuando se realiza una modelización de ecuaciones estructurales, se puede plantear más de un modelo inicial (dos o tres), y en el momento de realizar la evaluación de los modelos ver cuál se ajusta mejor a los datos.

BIBLIOGRAFÍA

1. BALBUENA Camino y CASAS Joan, “Aplicación del análisis factorial a la valoración por parte de los estudiantes de las asignaturas de la ETSICCP de Barcelona en sus distintas titulaciones”, http://www-ma3.upc.es/users/balbuena/CiudadReal.pdf

2. CASAS Mercedes, “Los modelos de ecuaciones estructurales y su aplicación en el índice europeo de satisfacción al cliente”, http://www.uned.es/asepuma2002/c29r.htm

3. DUNCAN Otis, “Introduction to Structural Equation Models”, Academic Press Inc., New York, EE.UU., 1975.

4. FOX John, “Structural Equation Models”, http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-sems.pdf

5. GARDNER Robert, “Estadística para Psicología Usando SPSS para Windows”, Prentice Hall, México, 2003.

6. HÄRDLE Wolfgang y SIMAR Leopold, “Applied Multivariate Statistical Analysis”, http://www.xplore-stat.de/ebooks/ebooks.htm

7. JOHNSON Richard, “Applied Multivariate Statistical Analysis”, Prentice Hall, EE.UU., 2001.

8. KAPLAN David, “Structural Equation Modeling: Fundations and Extensions”, Sage Publications, California, EE.UU., 2000.

9. KENNY David, “Structural Equation Modeling”, http://users.rcn.com/dakenny/causalm.htm

10. KLINE Rex, “Principles and Practice of Structural Equation Modeling”, The Guilford Press, New York, 2005.

11. LOEHLIN John, “Latent Variable Models”, Lawrence Erlbaum Associates, New Jersey, 2004.

12. LUQUE Teodoro, “Técnicas de Análisis de Datos en Investigación de Mercados”, Ediciones Pirámide, Madrid, 2000.

13. SIERRA Restituto, “Ciencias Sociales: Análisis Estadístico y Modelos Matemáticos”, Paraninfo S.A., Madrid, España, 1981.