



UNIVERSIDAD NACIONAL DE INGENIERÍA  
Facultad de Ingeniería Económica, Estadística y CCSS

**Ecuaciones Estructurales con variables dicotómicas: Modelo aplicado a los factores que determinan la comprensión lectora de los alumnos de 6to grado de Lima Metropolitana**

Informe de Suficiencia

Para obtener el grado de Ingeniero Estadístico en la Escuela Profesional de Ingeniería Estadística-UNI, bajo la dirección del profesor Ms.ME

por

Bach. Pilar Denisse Villena Guzmán

- 2014 -

## RESUMEN

En la actualidad, los modelos de ecuaciones estructurales aplicados a estudiar los factores que influyen en la comprensión lectora, usan solo variables continuas debido a los supuestos que se deben cumplir para aplicar esta técnica. Así, si en su información cuentan con variables dicotómicas, éstas son transformadas a variables continuas mediante diversos métodos. Por esta razón el objetivo de este estudio es explorar y desarrollar un modelo de ecuaciones estructurales con variables dicotómicas, que permita explicar los factores que influyen en la comprensión lectora de los alumnos de Lima Metropolitana. Para este fin, el análisis hace uso de la estadística Bayesiana. Se utiliza información de las pruebas Crecer dirigidas a los niños de 6to de primaria, y además encuestas tanto a los alumnos como padres y directores; todas estas pruebas fueron respaldadas y administradas por el Ministerio de Educación. Usando sólo variables dicotómicas se construyó un modelo de ecuaciones estructurales que nos permite obtener los siguientes resultados: i) el factor socioeconómico es el que presenta mayor relación con la comprensión lectora, confirmando así estudios realizados con anterioridad; ii) el factor entorno escolar presenta una relación opuesta a la comprensión lectora iv) y por último se presenta y propone un modelo que relaciona la comprensión lectora con el factor socioeconómico, entorno familiar y programas sociales.

Palabras clave: Ecuaciones Estructurales, Estadística Bayesiana, comprensión lectora.

## INDICE

I)	INTRODUCCION.....	4
II)	EL PROBLEMA DE INVESTIGACIÓN.....	5
III)	PREGUNTA Y OBJETIVO.....	6
IV)	IMPORTANCIA/JUSTIFICACIÓN .....	7
V)	ANTECEDENTES.....	7
VI)	MARCO TEÓRICO	
	6.1) Con respecto al tema de investigación.....	10
	6.1.1 Definiciones Básicas.....	10
	6.1.2 Variables que influyen en la comprensión lectora.....	13
	6.2) Con respecto al modelo.....	15
	6.2.1 Modelo de Ecuaciones Estructurales.....	15
	6.2.2 Estimación Bayesiana de los Modelos de Ecuaciones Estructurales.....	20
	6.2.3 Modelo de Ecuaciones Estructurales con Variables Dicotómicas.....	24
VII)	METODOLOGÍA .....	27
	7.1) Metodología del MINEDU.....	27
	7.2) Instrumentos .....	28
	7.3) Procedimientos.....	29
VIII)	RESULTADOS.....	31
IX)	CONCLUSIONES.....	34
X)	REFERENCIAS BIBLIOGRÁFICAS.....	37
	ANEXO A: Cuadros convergencia.....	40
	ANEXO B: Código OpenBugs.....	41
	ANEXO C: Datos(muestra).....	42
	ANEXO D:Encuesta MINEDU.....	43

## I) INTRODUCCION

La importancia de la lectura en la vida de cualquier persona es innegable, forma parte del proceso de aprendizaje, brinda acceso a la información ampliando el bagaje cultural, da facilidad para exponer el propio pensamiento, etc. (Junior Cabral, La importancia de la lectura 2007<sup>1</sup>)

Dado su nivel de importancia, desde hace muchos años se ha estudiado en diferentes países, cuáles son los factores que influyen en la comprensión lectora.

En el Perú, a raíz de las cifras anunciadas por el Ministerio de Educación (MINEDU), donde la UNESCO y el Programa para la Evaluación Internacional de alumnos (PISA) indican que nuestro país presenta la mayor proporción de estudiantes (54.1%) situados en el nivel más bajo de comprensión lectora, también se ha desarrollado a fondo este tema.

Todos estos estudios están enfocados en estudiar la relación existente entre la comprensión lectora y factores como conocimiento del vocabulario, situación socioeconómica, entorno familiar, motivación personal, etc.

Para obtener esta información se hizo uso de encuestas, exámenes reconocidos y validados internacionalmente, etc. Algunas de las variables obtenidas de estos instrumentos, son continuas, sin embargo, la mayoría de éstas son binarias, ordinales, y son tratadas como si fueran continuas ó transformadas mediante diferentes métodos al tipo continua, para así cumplir con los supuestos de las diversas técnicas estadísticas existentes. Cualquiera de estos procedimientos no debería llevar a problemas muy serios, siempre y cuando se cumplan otros requisitos con estas variables, sin embargo no se estaría considerando información importante.

Por todo esto, el presente documento (usando información procedente del MINEDU), busca desarrollar una solución alternativa, usando un modelo de Ecuaciones Estructurales, con enfoque bayesiano para variables dicotómicas para estudiar la relación entre la comprensión lectora y los factores que podrían influir en su desarrollo.

## II) El problema de investigación

El presente trabajo enfoca una doble problemática. La primera es del tipo social, donde los resultados obtenidos en las pruebas realizadas por organismos internacionales como UNESCO en el Perú, nos permiten tener una idea de lo grave de la situación (Resultados provenientes Ministerio del Perú, Lineamientos de política 2004 – 2006):

- Más del 10% de la población puede considerarse como analfabeta
- El rendimiento de los estudiantes peruanos está por debajo del promedio del grupo de países latinoamericanos.

Resultados más actuales (Evaluación Censal de estudiantes de 2do Grado 2012, resultados presentados por la ministra de educación de ese entonces Patricia Salas), nos indican que solo un 31% de estudiantes alcanza un nivel satisfactorio en comprensión de lectora.



Ministerio de Educación 2012

Además de estas cifras, existe una realidad grave en nuestro país, y es que mucho se confunde el hecho de saber leer con el hecho de comprender lo que se lee. En las aulas escolares del Perú, se da por entendido que todos los estudiantes saben leer, sin embargo, solo un pequeño grupo llega a comprender lo que lee. Todas estas razones han hecho que la comprensión lectora sea considerada como asunto de gran importancia, no solo por el gobierno del Perú, sino también por la empresa privada, como el Banco Continental quien lanzó su programa “Leer es estar adelante” (Fundación BBVA Continental)

Además se hicieron diversos estudios para revertir esta situación, explorando todas las variables que podrían influir en la comprensión lectora.

Y es ahí donde empieza la segunda problemática en este tema.

Todos los estudios realizados sobre este tema usan información proveniente de encuestas y exámenes validados internacionalmente, donde las variables obtenidas son de diversos tipos: continuos, categóricas, binarias, etc.

Sin embargo, dado los supuestos de las técnicas estadísticas como las ecuaciones estructurales donde se exige la normalidad de las variables, muchas de éstas son tratadas como tales, cuando no lo son, o muchas veces, transformadas a continuas asignándoles determinados puntajes según la respuesta.

Si bien este tratamiento de la información no lleva a problemas serios siempre y cuando se realice de manera correcta, el presente documento busca explorar y proponer una solución alternativa haciendo uso de la estadística bayesiana, que nos permita usar la información tal y como son obtenidas.

En resumen, nuestra problemática de estudio es el bajo nivel en comprensión lectora que presentan los alumnos de lima metropolitana.

Y nuestro problema de investigación, es la dificultad de usar variables binarias en un modelo de ecuaciones estructurales.

Viendo la realidad sobre estas dos problemáticas, formulamos a continuación nuestra pregunta de investigación

### **III) Pregunta y Objetivo de la Investigación**

#### *Pregunta principal*

¿Es posible desarrollar un modelo de Ecuaciones Estructurales para estudiar los posibles factores que influyan en la comprensión lectora, usando solo variables dicotómicas?

#### *Objetivo principal*

Proponer un modelo de ecuaciones estructurales con variables dicotómicas, que permita explicar los factores que influyen en la comprensión lectora de los alumnos de Lima Metropolitana.

#### **IV) Importancia y/o justificación**

Son muchos los estudios que se han efectuado para explicar los factores que influyen en la comprensión lectora, todos ellos adecuando la información obtenida a los requisitos de las diversas técnicas estadísticas existentes.

Con el presente trabajo se busca proponer el uso de ecuaciones estructurales con enfoque bayesiano- muy desarrollado en los últimos años- en el campo de la educación, de tal manera que se puedan superar algunas limitaciones existentes, como que solo se pueda usar variables continuas, cuando la mayoría de la información que se recoge no es de este tipo.

Se busca además estudiar un poco más los posibles factores que afectan la comprensión lectora de los alumnos de Lima, Perú.

#### **V) Antecedentes**

El modelo de ecuaciones estructurales es una técnica estadística que ha aumentado su popularidad desde que fue concebido por primera vez por Wright (1918), quien desarrolló el Análisis Path para analizar la teoría genética en el campo de la biología. En los años 70, tuvo un resurgimiento, particularmente en la sociología y la econometría y luego se diversificó a otras disciplinas como la psicología, ciencias políticas y en la educación (Timothy Theo 2009).

En los últimos años, el uso de las Ecuaciones Estructurales se ha incrementado entre los investigadores de la educación. Esto se debe en parte al fácil acceso a la literatura y fuentes de este tema, que pueden ser fácilmente encontrados en internet.

En los años 90, a nivel mundial, se hicieron diversos estudios para explicar la relación entre la comprensión lectora y diversos factores como conocimiento de vocabulario, situación socioeconómica, etc, en alumnos de diversos grados de primaria; todas estas investigaciones fueron publicadas en la revista Scientific Studies of Reading. Algunos de ellos son:

- 1) Título: Structural Equation Model for Reading Comprehension Based on Background, Phonemic, and Strategy Knowledge.

Autor: Victor L. Willson & William H. Rupley 1997

Año: 1997.

Se buscó desarrollar un modelo de ecuaciones estructurales para comprensión lectora basado en el conocimiento fonético, conocimiento previo del texto y estrategia usada.

Conclusiones: Los resultados revelan que para los grados de 2do y 3ro primaria, la comprensión lectora está básicamente guiada por el conocimiento fonético y en segundo lugar por el conocimiento previo del texto. Esta misma variable, es de mayor relevancia para el 4to grado de primaria, y la estrategia para leer, empieza a tener relevancia para la comprensión lectora. En los grados superiores, la estrategia sobre cómo leer y qué leer, empieza a dominar la predicción para la comprensión lectora. Esta relación aparentemente es estable, y el conocimiento previo se vuelve irrelevante en el 6to grado.

- 2) Título: Relationships between Word Knowledge and Reading Comprehension in Third-Grade Children.

Autor: Kendra R. Tannenbaum, Joseph K. Torgesen & Richard K. Wagner,

Año: 2006.

Lugar: Estados Unidos.

Se buscó desarrollar un modelo de ecuaciones estructurales que estudiara la relación entre el conocimiento de la palabra y comprensión lectora en niños de 3er grado.

Conclusiones: Un análisis factorial confirmatorio, un modelo de ecuaciones estructurales y un análisis de regresión jerárquico demuestran que un modelo de dos factores: amplitud de vocabulario y profundidad/fluidez de vocabulario, es el que mejor se ajusta a los datos. La amplitud tiene una relación más fuerte con la comprensión lectora que el otro factor. Sin embargo ambas variables tienen una influencia significativa en la comprensión lectora



En el Perú, también se han hecho este tipo de investigaciones, pero la mayoría de ellas del tipo descriptivo como “Actitudes hacia la lectura y niveles de comprensión lectora en estudiantes de sexto grado de Primaria” (Ana Cubas Barrueto, Lima-Perú 2007)

Sin embargo, sí existe en el Perú, un estudio de la comprensión lectora y los factores que influyen, con el uso de ecuaciones estructurales

3) Título : Socio-cultural variables and Reading literacy acquisition in Lima, Perú

Autor: Silvia Morales

Año: 2009

Lugar: Lima, Perú.

Este estudio hace una investigación detallada acerca de la situación de la comprensión lectora en el Perú, y analiza diversos factores socioculturales, socioeconómicos, etc, en alumnos de 4to y 5to grado de primaria de Lima. Tiene en cuenta además, la situación de los alumnos al comienzo y final del año, y escoge niños de escuelas que pertenezcan a dos zonas de diferente nivel social.

Para hacer este estudio se usaron adaptaciones de test reconocidos internacionalmente y diversas encuestas.

Estas últimas presentaban variables de todo tipo, siendo predominantes las del tipo categórica y en muchos casos binarias. Para poder usar las ecuaciones estructurales se le asignó distintos puntajes a cada pregunta y así se pudo formar una variable continua.

Conclusiones: El modelo de ecuación estructural mostró que el desarrollo de la comprensión lectora fue influenciado por la habilidad de los niños por decodificar palabras, su vocabulario y la motivación para leer. Además, el género y la madurez intelectual, así como el clima de alfabetización del hogar del niño y la situación son aparentemente muy importantes para predecir el desarrollo de la comprensión lectora, directa o indirectamente.

Éste último estudio, constituye el antecedente más directo al presente documento, donde se buscará poder hacer uso de las ecuaciones estructurales con las variables en su forma original.

## **VI) Marco Teórico**

### **6.1) Con respecto al tema de investigación**

#### **6.1.1) Definiciones Básicas:**

A fin de comprender mejor el problema identificado y la solución a presentar, se deben tener en cuenta los siguientes conceptos:

**Lectura.-** La lectura es un proceso complejo mediante el cual un lector construye el significado de un texto. La lectura es un proceso sociocultural, porque es aprendido en el marco de una comunidad que la considera relevante, y su apropiación se inicia aún antes de la educación formal, desde los primeros contactos con el mundo escrito y con las conductas de lectoescritura de los adultos (Anderson y Teale, 2007).

#### **Comprensión lectora**

La comprensión de un texto supone la construcción por parte del lector de una representación mental del texto, dentro de todas las posibles representaciones de este. En este proceso constructivo juegan un rol principal los conocimientos previos que tiene el lector. Estos le permitirán aproximarse con mayor o menor éxito al sentido de lo que el autor quiso transmitir (Gómez-Palacio, 1993). En ese sentido, la comprensión del texto también implica una actividad estratégica por parte del lector. Es decir, pone en marcha una serie de procesos que implican la elaboración de inferencias, conjeturas, interpretaciones, etc. desde el mismo inicio de la lectura, o incluso antes (Van Dijk, 1998). Estos procesos son automáticos y casi inconscientes en lectores expertos, sin embargo, son susceptibles de ser enseñados y aprendidos.

**CRECER.-** es el proyecto de evaluación del rendimiento escolar a cargo de la UMC (Unidad de Medición de Calidad Educativa) que aplica periódicamente pruebas en algunas áreas de desarrollo o cursos (i.e., Comunicación Integral, Lógico-Matemática) y encuestas (para directores, profesores, padres de familia y alumnos) para retroalimentar los programas y proyectos desarrollados por el MINEDU (Ministerio de Educación), y contribuir con la formulación de las

políticas educativas. En el área de comunicación integral, evalúa la comprensión lectora, que busca examinar la capacidad del alumno de lograr:

- a.- Identificar los hechos o ideas principales de textos leídos (4 ítems)
- b.- Extraer información específica de un texto (4 ítems)
- c.- Descubrir la intencionalidad del autor (2 ítems)
- d.- Reconocer las marcas de cohesión textual (1 ítem)

Al término del sexto grado el niño debe haber desarrollado dos grandes competencias en comprensión de lectura de textos funcionales e informativos y otra a la lectura de textos literarios. En la prueba cuyos resultados estamos evaluando, se utilizó un texto informativo sobre las riquezas marinas, y lo que se buscó saber fue cuán capaces eran los estudiantes de este año “para leer con sentido crítico textos informativos y si podían identificar ideas, imágenes y datos importantes. La información presentada en el texto era muy probablemente nueva para el alumno ya que se trata de información algo difícil de manejar con nombres complejos como plancton, necton, etc. Con todo esto se buscaba conocer la capacidad de los niños que terminan su primaria, para procesar información que encontrarán en la secundaria. Se usó el siguiente texto:

*Lee atentamente el siguiente texto y luego responde las preguntas 1,2,3,4,5,6,7 y 8:*

### LAS RIQUEZAS MARINAS

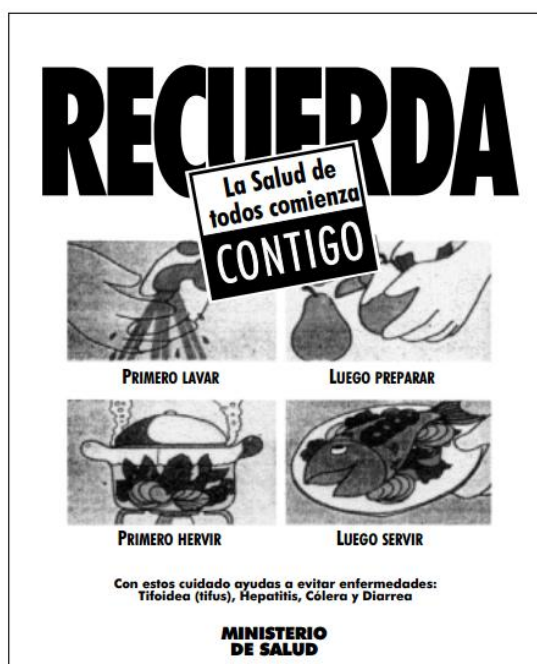
Las aguas del mar contienen muchas variedades de organismos vivos. Éstos suelen dividirse en tres grandes clases, conocidas con los nombres de plancton, necton y bentos.

El **plancton** se compone de distintos tipos de plantas y animales, generalmente diminutos, que flotan cerca de la superficie del mar y se dejan arrastrar por los vientos y las corrientes de agua. El **necton** es el conjunto de seres marinos que se mueven por sus propios medios, independientemente de los vientos y las corrientes marinas. Entre los más conocidos se puede mencionar a los peces, los calamares, las tortugas marinas, las focas y las ballenas. Por último, el **bentos** está constituido por organismos que habitan en el suelo marino ya sea cerca de la costa o en el fondo de los océanos. Algunos se adhieren a las rocas o a la arcilla del fondo del océano, mientras que otros se arrastran sobre ellas o cavan túneles en sus arenas. Entre sus miles de variedades conocidas están las esponjas, erizos, ostras y estrellas de mar.

Los organismos que conforman el plancton abundan de un modo prodigioso en las aguas marinas. En su mayoría son traslúcidos y por eso, se confunden con el color de las aguas. Algunos de ellos son algas y bacterias, mientras que otros pueden ser pequeñísimos crustáceos y moluscos. Los peces, así como muchos otros animales marinos, se alimentan de plancton. Sin éste morirían, y los seres humanos se verían privados de una de sus más importantes fuentes de alimentos...

(Adaptación )

Además de esto, también se busca evaluar la comprensión de lectura en textos ícono-verbales, que son aquellos cuyos mensajes se expresan complementariamente en imágenes y textos y también surge la necesidad de aprender a leerlos para acceder a la información que contienen. Se busca que identifiquen el tipo de texto por sus rasgos distintivos, que reconozcan su intencionalidad y el mensaje en sí.



#### 6.1.2) Variables que influyen en la comprensión lectora

Son muchos los investigadores que han estudiados los diversos factores que pueden influir en la comprensión lectora (Defior, 1996, Kendra Tannembaum 2009, Thorne y Pinzás 1988, etc)

Algunos de ellos, como Mayor, Suengas y Gonzáles Márquez, en un estudio hecho en el 2004, la agrupan de la siguiente manera

Variables contextuales.- donde se encuentran los libros que se emplean, el contexto escolar, en entorno familiar, y el ambiente sociocultural

Variables subjetivas.- donde se encuentran el conocimiento previo del lector, las estrategias de aprendizaje, motivación y la memoria del trabajo.

Variables de actividad.- son aquellas que se refieren al tipo de texto, a las diferentes metas que se traza el lector y la adecuación entre metas y recursos.

El presente trabajo se enfocará en el primer grupo donde se evaluarán los siguientes factores:

##### - **Factor socioeconómico.-**

Se sabe que la diversidad sociocultural tiene gran impacto en el desarrollo de las habilidades en la lectura de los estudiantes. Asimismo se puede afirmar que

mientras la familia del niño cuente con suficientes recursos económicos, éste podrá acceder a más fuentes de aprendizaje, etc.

Siendo este factor considerado por muchos estudios como el más importante, ha sido incluido en el presente estudio.

#### **- Factor Entorno Familiar :**

Muchos investigadores definen el ambiente de alfabetización en el hogar, como, entre otras cosas: la disponibilidad de materiales para la alfabetización en el hogar, el número de lecturas hecha con los niños, leerles en voz alta, la guía de lectura y el estímulo suministrado por los padres, el comportamiento de los padres como modelo de lectura y su interés expresado por la lectura (Baker, Scher) Además se ha encontrado que en hogares donde los padres leen hay mayor porcentaje de niños que leen de manera independiente (fuera del colegio). Sabemos que si los padres leen, y les leen a sus hijos y además los incentivan a leer están teniendo buena influencia en sus hijos lo que probablemente conducirá a que en el futuro sean buenos lectores. (Baker 1999 Beginning Readers' motivation for Reading in relation to parental beliefs, and home Reading experiences; Kirsten D. M. Kowalewski: Reading Engagement: What Influences The Choice to Read?)

Si bien pudiera parecer que está muy relacionado al factor socioeconómico, muchos estudios afirman que sí deben ser evaluados por separados. Aunque las familias de clases altas tienen mayores probabilidades de obtener mejores resultados en el campo de la educación, se ha llegado a demostrar cómo el número de libros con los que cuenta el hogar ejerce un efecto independiente al socioeconómico (Eduardo Velez). Una de las posibles razones que explique esto es, según el sociólogo peruano Martín Benavides, es que el gusto por la cultura esté distribuido homogéneamente entre diversos grupos sociales.

Por estas razones, éste factor ha sido incluido en el presente análisis

#### **- Factor Programas Sociales Escolares.-**

En diversos estudios se ha tenido en cuenta el contexto escolar en los resultados académicos de los alumnos. Para este trabajo, y dada la

disponibilidad de información, se evalúa específicamente las variables donde el Gobierno tenga influencia directa en el colegio mediante programas sociales como el vaso de leche, asistencia médica y mobiliario escolar. Sobre este punto en particular no se ha encontrado mayores investigaciones hechas a nivel nacional, es por eso que ha sido incluido en el presente estudio.

## **6.2) Con respecto al modelo**

### **6.2.1) Modelo de Ecuaciones Estructurales**

En los campos del comportamiento, educación, medicina y las ciencias sociales usualmente son mencionadas dos tipos de variables: las observadas (manifiestas) y las latentes.

Las variables observadas son aquellas que pueden ser medidas directamente, tales como ingresos, puntajes de exámenes, presión sanguínea, peso, etc. Todos estos datos son registros de mediciones a diferentes individuos. A menudo, también es necesario tratar con las variables del segundo tipo, las latentes, que son aquellas que no pueden ser medidas directamente por una variable observada. Ejemplos: la inteligencia, personalidad, habilidad matemática, ansiedad, comportamiento de compra, etc. En la práctica, las características de una variable latente pueden ser medidas parcialmente mediante una combinación lineal de algunas variables observadas. Por ejemplo, la habilidad cuantitativa de los estudiantes de secundaria, puede reflejarse por sus puntajes en los exámenes de matemáticas, física y química, etc.

En las investigaciones es importante establecer un modelo apropiado para evaluar una serie de hipótesis acerca del impacto de variables latentes y observadas en otras variables, y tomar en cuenta los errores de medición. Las ecuaciones estructurales (Structural Equation Models, SEMs, por sus siglas en inglés) son ampliamente reconocidas como la técnica estadística más importante para cumplir este propósito.

Ejemplo de Aplicaciones: pueden ser aplicados para investigación de mercados para establecer interrelaciones entre la demanda y la oferta y la actitud y

comportamiento de los clientes; en las ciencias del ambiente para investigar como la salud es afectada por el aire y la contaminación del agua; en la educación para medir el desarrollo de la inteligencia y su relación con la personalidad y el ambiente escolar; en la medicina para analizar la calidad de la vida ó el impacto del comportamiento del médico en el uso de medicinas.

El Modelo de ecuaciones estructurales tiene dos componentes. El primero, es un análisis factorial confirmatorio el cual relaciona las variables latentes con todas sus variables observadas correspondientes y toma en cuenta el error en la medición. Este componente puede ser considerado como un modelo de regresión que relaciona las variables observadas con un pequeño número de variables latentes. El segundo componente es una regresión tipo ecuación estructural, el cual relaciona las variables latentes endógenas (dependientes) con los términos lineales de algunas variables latentes endógenas y exógenas (independientes). Como las variables latentes son aleatorias, no pueden ser analizadas directamente por técnicas de una regresión común que están basadas en las observaciones. Sin embargo, conceptualmente, los SEMs son formulados por, un modelo muy conocido por nosotros, tipo regresión, por lo que son relativamente fáciles de aplicar.

#### Análisis de la estructura de la covarianza

Para los SEMs estándar, la matriz de la covarianza del vector aleatorio observado “y” contiene todos los parámetros desconocidos en el modelo, por lo tanto los métodos clásicos para analizar SEM se enfocan en la matriz de covarianza muestral S y no en los vectores aleatorios individuales  $y_i$ . Esto involucra la formulación de la estructura de la covarianza  $\Sigma(\theta)$ , que es una matriz función del vector de parámetros desconocidos  $\theta$ ; la estimación de  $\theta$  minimizando (ó maximizando) alguna función objetivo que mida la discrepancia entre S y  $\Sigma(\theta)$ , tales como la función de máxima verosimilitud (ML) o la función generalizada de mínimos cuadrados (GLS); y la obtención de los estadísticos de bondad de ajuste para asegurar si  $\Sigma(\theta)$  se ajusta a S. Como este tipo de análisis enfatiza la matriz de covarianza poblacional y la muestral, es llamado comúnmente Análisis de Estructura de la Covarianza. Este análisis depende en gran medida de la normalidad asintótica de S. Cuando la distribución del vector



de variables aleatorias observadas es normal multivariada y el tamaño de la muestra es razonablemente grande, podemos decir que la distribución asintótica de  $S$  se aproxima con precisión a esta distribución, y por lo tanto esta aproximación da buenos resultados. Sin embargo, bajo condiciones ligeramente más complejas, que son comunes en las investigaciones, la aproximación del análisis de estructura de la covarianza de  $S$  no es efectiva y se puede encontrar problemas teóricos y computacionales (Sik Yum Lee).

Debido a la presencia de los términos no lineales de las variables latentes, las variables latentes endógenas y las variables observadas relacionadas en  $y_i$  no son distribuidas normalmente, y por lo tanto la matriz de covarianza muestral no es adecuada para modelar este tipo de relaciones.

En resumen, para datos dicotómicos o categóricos, la matriz de covarianza muestral no puede ser usada.

La siguiente simbología, usada en el libro Structural Equation Modeling de Sik Yum Lee, será usada a partir de ahora

<b>Símbolo</b>	<b>Significado</b>
$u, v, x, y$	Vectores aleatorios observados
$\omega$	Vector aleatorio latente en la ecuación de medida
$\eta$	Vector latente endógeno en la ecuación estructural
$\xi$	Vector latente exógeno en la ecuación estructural
$\epsilon, \delta$	Vector aleatorio de errores
$\Lambda$	Matriz carga factores en la ecuación de medida
$\Pi, \Gamma, B$	Matrices de coeficientes de regresión en la ecuación estructural
$\Phi$	Matriz de covarianza de variables latentes
$\Psi_\epsilon, \Psi_\delta$	Matriz de covarianza diagonal de la medida del error con valores $\psi_{\epsilon k}, \psi_{\delta k}$
$\alpha_{0\epsilon k}, \beta_{0\epsilon k}, \alpha_{0\delta k}, \beta_{0\delta k}$	Hiperparámetros en la distribución de gamma de $\psi_{\epsilon k}, \psi_{\delta k}$ respectivamente
$R_0, \rho_0$	Hiperparámetros en la distribución Wishart relacionadas a la distribución a priori de $\Phi$

$\Lambda_{0k}, H_{0yk}$	Hiperparámetros en la distribución normal multivariada relacionada a la distribución a priori de la k-ésima fila de $\Lambda$ en la ecuación de medida con y
$\Lambda_{0\omega k}, H_{0\omega k}$	Hiperparámetros en la distribución normal multivariada relacionada a la distribución a priori de la k-ésima fila de $\Lambda_k$ en la ecuación estructural
$I_q$	Matriz identidad q x q

### Modelos básicos de ecuaciones estructurales

Históricamente, el análisis factorial es el SEM básico que fue desarrollado por psicómetras, para estudiar las relaciones internas de un conjunto de variables. El principal objetivo de este modelo es que para un conjunto de variables, queremos encontrar un número más pequeño de factores latentes no relacionados entre sí, que se tendrán en cuenta para la intercorrelaciones de las variables respuesta, de tal manera que cuando los factores latentes son removidos parcialmente de las variables respuestas, no permanece ninguna correlación entre ellas.

- **El modelo Análisis Factorial Exploratorio**, está definido por un vector aleatorio  $X$  que satisface la siguiente ecuación ineal

$$x = \Lambda \xi + \epsilon$$

donde  $\Lambda(p \times q)$  es una matriz de coeficientes factoriales,  $\xi(q \times 1)$  es un vector aleatorio de factores comunes latentes y  $\epsilon(p \times 1)$  es un vector aleatorio de errores. Se asume que  $\xi$  se distribuye como  $N(0, I)$  y  $\epsilon$  se distribuye como  $N(0, \Psi_\epsilon)$  donde  $\Psi$  es una matriz diagonal y  $\xi$  no está correlacionado con  $\epsilon$ . Generalmente  $q$  es mucho más pequeño que  $p$ . El vector aleatorio de variables observadas  $x$  se distribuye como  $N(0, \Sigma)$ , donde

$$\Sigma = \Lambda \Lambda' + \Psi_\epsilon$$

Claramente, la  $\text{Cov}(x, \xi) = \Lambda$ . De aquí que las correlaciones entre los factores latentes y las variables observadas son dadas por los elementos en la matriz de cargas factoriales  $\Lambda$ . La varianza de la k-ésima variable observada es igual a

$$\sigma_{kk} = \lambda_{k1}^2 + \dots + \lambda_{kq}^2 + \psi_{\epsilon k}$$

donde  $\lambda_{kh}$  y  $\psi_{ek}$  son: el (k,h)-ésimo elemento de  $\Lambda$  y el k-elemento de  $\Psi_e$  respectivamente.

El valor  $\lambda_{k1}^2 + \dots + \lambda_{kq}^2$  es llamado comunalidad, y representa el porcentaje de la varianza explicada por los factores latentes

## - Modelo Análisis Factorial Confirmatorio

### Modelo Lisrel

En el análisis Factorial Confirmatorio (CFA) las correlaciones entre las variables latentes pueden ser evaluadas por su matriz de covarianza; sin embargo, las variables latentes nunca pueden formar parte de una regresión para estimar otra variable. El objetivo de desarrollar los SEMs es generalizar el modelo CFA, para evaluar como las variables latentes influyen una en otra en diversas maneras. Una típica representación de los SEMs es el modelo LISREL, el cual consiste de 2 componentes: modelo de medición y modelo estructural.

El modelo de medición, está definido por los siguientes modelos CFA

$$x_1 = \Lambda_1 \eta + \epsilon_1$$

$$x_2 = \Lambda_2 \xi + \epsilon_2$$

Donde  $x_1(r \times 1)$  y  $x_2(s \times 1)$  son vectores aleatorios de variables observadas, los cuáles son los indicadores respectivos para  $\eta$  y  $\xi$ ,  $\Lambda_1(r \times q_1)$  y  $\Lambda_2(s \times q_2)$  son las matrices de coeficientes,  $\epsilon_1(r \times 1)$  y  $\epsilon_2(s \times 1)$  son los errores de medida. Se asume que  $\epsilon_1$  y  $\epsilon_2$  no están correlacionados con  $\eta$  y  $\xi$ ,  $\delta$  y las distribuciones de estos vectores aleatorios son normales con media cero.

Modelo estructural.- Especifica las ecuaciones causales lineales entre las variables latentes del modelo

$$\eta = \Pi \eta + \Gamma \xi + \delta$$

donde  $\eta(q_1 \times 1)$  es un vector de variables latentes endógenas y  $\xi(q_2 \times 1)$  es un vector de variables latentes exógenas,  $\Pi(q_1 \times q_1)$  y  $\Gamma(q_2 \times q_1)$  son las matrices de coeficientes de regresión que representan los efectos entre  $\eta$  y  $\xi$ , y  $\delta$  es el término de perturbación aleatoria.

### 6.2.2) Estimación Bayesiana de los Modelos de Ecuaciones Estructurales

La teoría estadística que está asociada con los métodos de estimación de Mínimos Cuadrados y máxima Verosimilitud es desarrollada en la base de la matriz de covarianza muestral  $S$ . Por ejemplo la forma de la función de mínimos cuadrados y la derivación de las propiedades asintóticas de su estimador, depende principalmente de la distribución de  $S$ . Por esta razón estas aproximaciones serán correctas solo bajo ciertas suposiciones

Para que estas aproximaciones sean correctas se asumen que las observaciones aleatorias están distribuidas idénticamente e independientemente de acuerdo a una distribución normal. Sin embargo, sabemos que en mundo real es complicado, y que los supuestos requeridos no pueden ser cumplidos en muchos problemas de investigación. Por lo tanto, existe una fuerte necesidad de desarrollar nuevos métodos estadísticos para el manejo de los modelos más generales y estructuras de datos complejas.

El objetivo básico de este documento es presentar un enfoque bayesiano para analizar no sólo los SEM estándar, sino también sus generalizaciones útiles que se han desarrollado en los últimos años

En contraste con el enfoque de análisis de covarianza estructura existente, nos centramos en el uso de las observaciones en lugar de la matriz de covarianza de la muestra.

Para resolver las dificultades que son inducidos por la complejidad del modelo y de los datos, se sigue la siguiente estrategia: ***Primero se trata las variables latentes en el modelo, y las mediciones latentes (tales como los datos que faltan, o las mediciones continuas que están asociados con los datos discretos) como datos perdidos.*** A continuación, se analiza el modelo sobre la base del conjunto de datos completos, que contiene los datos observados y los datos perdidos, mediante la aplicación de algunas herramientas de gran alcance en el cálculo estadístico.

Como el conjunto de datos completo es mucho más fácil de manejar, las dificultades que son inducidos por las complejidades del modelo y los datos se hacen más ligeras.

El atractivo básico de la aproximación bayesiana es su flexibilidad para usar información a priori muy útil para conseguir mejores resultados. En muchas investigaciones, se puede contar con muy buena información a priori de diferentes fuentes, por ejemplo el conocimiento de expertos y análisis pasados. Para situaciones sin información previa precisa, se puede usar algún tipo de distribución a prioris no informativas. En estos casos la precisión con el enfoque bayesiano es muy cerca al del estimado mediante máxima verosimilitud

Además, se sabe que para las aproximaciones mediante máxima verosimilitud, es necesario contar con un número de datos bastante grande. En el campo de los SEM's se han hecho numerosos estudios para verificar esto y se concluye que las propiedades de los estadísticos no son robustos para muestras pequeñas (Boomsma, 1982; Chou, Bentler and Satorra, 1991, e.g.). Por el contrario, como ha sido remarcado por muchos artículos importantes en análisis Bayesiana de SEM, los métodos bayesianos dependen menos de la teoría asintótica y por lo tanto tienen el potencial para dar resultados confiables aún con muestras pequeñas (Ansari and Jedidi, 2000, e.g.)

La distribución a posteriori de parámetros y variables latentes pueden ser estimados usando un número suficientemente largo de observaciones que son simulados de la distribución a posteriori de los parámetros desconocidos a través de técnicas eficientes en cómputo estadístico como las Cadenas de Markov Monte Carlo (MCMC).

Las medias, quantiles, de esta distribución a posteriori pueden ser estimados de las observaciones simuladas. Estas cantidades son útiles para hacer inferencias estadísticas. Por ejemplo, los estimados bayesianos de los parámetros desconocidos y las variables latentes pueden ser obtenidas de la correspondiente media muestral de la distribución a posteriori.

### **Estimación Bayesiana**

Sea  $M$  un SEM arbitrario, con un vector de parámetros desconocidos  $\theta$ , y sea  $Y$  el conjunto de datos observados de tamaño " $n$ ". En una aproximación no bayesiana, por ejemplo Máxima Verosimilitud,  $\theta$  no es considerado como aleatorio, En la aproximación bayesiana  $\theta$  es considerado aleatorio con una

distribución (llamada distribución a priori ) y una función de densidad asociada (a priori), sea  $p(\theta, M)$  (Berger, 1985). Para simplificar las nomenclaturas, usaremos  $p(\theta)$  para denotar  $p(\theta, M)$ . La inferencia Bayesiana está basada en los datos observados  $Y$  y la distribución a priori de  $\theta$ .

Sea  $p(Y, \theta | M)$  la función de densidad de probabilidad de la distribución conjunta  $Y$  y  $\theta$  dado  $M$ . el comportamiento de  $\theta$ , dado  $Y$ , es totalmente descrito por la probabilidad condicional de  $\theta$  dado  $Y$ . Esta distribución condicional es llamada distribución a posteriori de  $\theta$ .

Sea  $p(\theta | Y, M)$  la función de densidad de la distribución a posteriori , que es llamada función de densidad a posteriori,

La función a posteriori de  $\theta$  ó su densidad, cumple el rol más importante en el modelo de análisis Bayesiano. Basado en una muy conocida identidad en las probabilidades, tenemos que

$$p(Y, \theta | M) = p(Y | \theta, M) p(\theta) = p(\theta | Y, M) p(Y | M)$$

Como  $p(Y | M)$  no depende de  $\theta$ , y puede ser considerado como una constant con  $Y$  fijo, tenemos que:

$$\text{Log } p(\theta | Y, M) \propto \log p(Y | \theta, M) + \log p(\theta)$$

Notemos que  $p(\theta | Y, M)$  puede ser considerado como la función de verosimilitud, porque es la probabilidad de densidad de  $y_1, \dots, y_n$  condicionado al vector de parámetros  $\theta$ .

De la última ecuación observamos que la función de densidad posterior incorpora la información muestral y la a priori a través de la función de verosimilitud  $p(Y | \theta, M)$  y la función de densidad a priori  $p(\theta)$ . Notamos también que  $p(Y | \theta, M)$  depende del tamaño de la muestra, mientras que  $p(\theta)$  no. Cuando el tamaño de la muestra es muy grande,  $\log(Y | \theta, M)$ , podría ser muy largo y por lo tanto  $\log p(Y | \theta, M)$  domina  $\log p(\theta)$ . En esta situación, la distribución a priori de  $\theta$  juega un papel menos importante y la función de densidad a posteriori  $\log p(\theta | Y, M)$  está cerca al logaritmo de la función de verosimilitud. Por lo tanto los métodos bayesianos y los de MV son asintóticamente equivalentes, y también

las estimaciones bayesianas tienen las mismas propiedades óptimas que las estimaciones de MV.

Cuando el tamaño de la muestra es pequeño ó moderado, la distribución a priori de  $\theta$  tiene un papel importante en la aprox. Bayesiana. Por lo tanto, en los problemas de investigación donde los tamaños de las muestras son pequeños, la distribución a priori de  $\theta$  puede ser incorporada al análisis bayesiano a través de la distribución a priori de  $\theta$  con el fin de conseguir mejores resultados. En muchos casos, los investigadores pueden tener muy buena información a priori de expertos de análisis similares del pasado ó de otras fuentes. Además se pueden conseguir resultados más exactos mediante la incorporación de información a priori a través de la distribución a priori de  $\theta$ .

De esta manera, la selección de la densidad a priori es un tema importante en el análisis bayesiano.

### **Distribución a priori**

La distribución a priori de  $\theta$  representa la distribución de los posibles valores de los parámetros, de la que  $\theta$  se ha elaborado. Básicamente hay 2 clases: distribuciones a prioris no informativas y las informativas.

Las no informativas se usan en situaciones donde la distribución a priori no tiene población, cuando se tiene muy poca información a priori y de aquí que la distribución a priori tenga una importancia mínima en la distribución a posteriori. La densidad a priori asociada es considerada vaga, difusa, ó no informativa, por ejemplo densidad que es proporcional a una constante o tiene una varianza extremadamente grande. En este caso la estimación bayesiana no es afectada por información externa a los datos observados.

Para la distribución a priori informativa, podemos tener conocimiento previo acerca de su distribución, ya sea de datos relacionados o del conocimiento de expertos. Usualmente una distribución a priori informativa tiene sus propios parámetros que son llamados hiperparámetros.

### 6.2.3) Modelo de Ecuaciones Estructurales con variables dicotómicas.

Un tipo de variable discreta que se encuentra con frecuencia en la investigación del comportamiento, educativa, médica y social es la variable dicotómica (sólo consta de dos categorías). Este tipo de variables surgen cuando se hacen preguntas para seleccionar Si/No acerca de la presencia de síntomas, “Sentirse mejor” “Sentirse peor” acerca de los efectos de un medicamento, De acuerdo o desacuerdo con el comportamiento de la policía, Satisfactorio ó Insatisfactorio acerca de un servicio público. Los valores numéricos usados a menudo son “0” y “1” ó “1” y “2”. Cuando analizamos datos dicotómicos, el requisito básico en SEM claramente no se cumple, y un análisis riguroso que tenga en cuenta esto es necesario.

En muchas investigaciones, particularmente en educación, es importante explorar y determinar un pequeño número de factores latentes intrínsecos bajo un número de test ítems. El análisis de ítems es un modelo importante que ha sido propuesto para explicar la estructura de los factores.

En 1981, Bock and Aitkin hicieron un importante avance con el MV Full Information Item Factor (FIIF). Su estrategia para resolver las dificultades de cálculo debido a las variables dicotómicas fue tratar los factores latentes como datos perdidos y aplicar el algoritmo EM(maximum expectation). En 1996 Meng and Schilling mejoraron el algoritmo EM usando el método de Montecarlo (MCEM). Recientemente Lee y Song (2003) desarrollaron una enfoque bayesiano para un SEM general con variables dicotómicas compuesto de una ecuación de medida y una estructura, basados en el muestreo de Gibbs. En este tipo de análisis se hace hincapié en la relación entre las variables observadas y las latentes, ó de la estructura de la covarianza en lugar del vector de medias.

#### Análisis Bayesiano.

Consideremos otra vez, un SEM que está compuesto de una ecuación de medida y una estructural. La primera está definida por el siguiente modelo de Análisis Factorial Confirmatorio

$$y_i = \mu + \Lambda\omega_i + \epsilon_i, \quad i=1, \dots, n. \quad (1)$$



donde  $y_i$  es un vector de variables observadas ( $p \times 1$ ),  $\mu$  es un vector de medias ( $p \times 1$ ),  $\omega_i$  es el vector de residuales, y  $\Lambda$  es la matriz de coeficientes factoriales desconocidos ( $p \times q$ ). Se asume que para  $i=1 \dots n$ ,  $\omega_i$  se distribuye  $N(0, \Phi)$ ,  $\epsilon_i$  está independiente distribuido como  $N(0, \Psi_E)$  donde  $\Psi_E$  es una matriz diagonal, y  $\omega_i$  y  $\epsilon_i$  no están correlacionados. Sea el vector  $\omega_i$  particionado en  $(\eta_i', \xi_i')'$ , donde  $\eta_i$  y  $\xi_i$  son vectores de variables latentes de  $q_1 \times 1$  y  $q_2 \times 1$ . Supongamos además que estos vectores latentes satisfacen la siguiente ecuación estructural

$$\eta_i = \Pi \eta_i + \Gamma \xi_i + \delta_i \quad (2)$$

Donde  $\Pi(q_1 \times q_1)$  y  $\Gamma(q_2 \times q_2)$  son matrices de coeficientes de parámetros desconocidos y  $\delta_i$  está distribuido  $N(0, \Psi_\delta)$  donde  $\Psi_\delta$  es una matriz diagonal y  $\xi_i$  y  $\delta_i$  son independientes. Se asume que

$\Pi_0 = I - \Pi$  es no singular, y su determinante una constante independiente, diferente de cero y con los elementos de  $\Pi$ . Por conveniencia sea  $\Lambda_\omega = (\Pi, \Gamma)$  entonces la ecuación estructural puede ser:

$$\eta_i = \Lambda_\omega \omega_i + \delta_i \quad (3)$$

Ahora supongamos que el valor exacto de  $y_i$  no está disponible y que su información está dada por un vector de variables observadas dicotómicas  $z_i = (z_{i1}, \dots, z_{ip})'$  de tal manera que para  $k=1 \dots p$ ,

$$z_{ik} = \begin{cases} 1 & \text{si } y_{ik} > 0, \\ 0 & \text{otro caso} \end{cases} \quad (4)$$

El conjunto de datos observados es  $\{z_i; i = 1 \dots n\}$

Como la función de densidad de  $z_i$  involucra una integral multidimensional, el análisis de SEM con variables dicotómicas no es sencillo.

Consideremos la relación entre el modelo de análisis factorial definido por la ecuación 1 con las variables dicotómicas de  $z$ . Sea  $\Lambda_k'$ ,  $\mu_k$  y  $\psi_{ek}$  la  $k$ -ésima fila de  $\Lambda$ , el  $k$ -ésimo elemento de  $\mu$ , y el  $k$ -ésimo elemento diagonal de  $\Psi_E$ , respectivamente. Entonces de la ecuación 4 decimos que

$$\begin{aligned} \Pr(Z_{ik}=1 | \omega_i, \mu_k, \Lambda_k, \psi_{ek}) &= \Pr(y_{ik} > 0 | \omega_i, \mu_k, \Lambda_k, \psi_{ek}) \\ &= \Phi^*[(\Lambda_k / \psi_{ek}^{1/2}) \omega_i + \mu_k / \psi_{ek}^{1/2}] \end{aligned} \quad (5)$$

Donde  $\Phi^*$  es la función de distribución de  $N(0,1)$ .). Notemos que  $\mu_k, \Lambda_k, \psi_{\epsilon k}$  no son estimables porque  $C\Lambda_k/(C\psi_{\epsilon k}^{1/2}) = \Lambda_k/\psi_{\epsilon k}^{1/2}$  y  $C\mu_k/(C\psi_{\epsilon k}^{1/2}) = \mu_k/\psi_{\epsilon k}^{1/2}$  para cualquier constante positiva  $C$ . Hay muchas maneras de resolver este problema. Algunos investigadores sugieren fijar  $\psi_{\epsilon k} = 1$ . Escogemos 1 por conveniencia, ya que escoger otro valor sería igual que una solución que involucre un cambio de escala. Por lo tanto nuestras ecuaciones de medidas y estructuras no son identificadas.

Sea,  $Z = (z_1, \dots, z_n)$  los datos observados de variables dicotómicas, y nuestro objetivo desarrollar un procedimiento bayesiano para estimar  $\theta$ , que contenga  $a = \mu, \Lambda, \Lambda_\omega, \Phi, \Psi_\delta$  basados en  $Z$ . Analizamos la función de densidad posterior de  $\theta$  dado  $Z$

$$\text{Log } p(\theta | Z) \propto \log p(Z|\theta) + \log p(\theta)$$

Donde  $p(Z|\theta)$  es la función de verosimilitud, y  $p(\theta)$  es la densidad a priori de  $\theta$ . Debido a la complejidad del modelo y la naturaleza de los datos,  $p(Z|\theta)$  involucra cálculos integrales muy grandes y la distribución a posteriori es muy complicada.

Por esta razón la técnica de aumento (Tanner y Wong, 1987) es otra vez empleada en el análisis a posteriori. Sea  $\Omega = (\omega_1, \dots, \omega_n)$  la matriz de variables latentes del modelo y  $Y = (y_1, \dots, y_n)$  la matriz de mediciones latentes continuas, que subyacen a la matriz de datos observados dicotómicos  $Z$ . En el análisis los datos observados, son aumentados con  $\Omega$  y  $Y$  los cuales pueden ser considerados como datos perdidos dicotómicos para formar un conjunto de datos completo  $(Z, \Omega, Y)$ .

Una vez más, se muestra un gran número de observaciones de  $p(\theta, \Omega, Y|Z)$  mediante el muestreo de Gibbs.

Para implementar el muestreo de Gibbs, empezamos con valores iniciales de  $(\theta^0, \Omega^0, Y^0)$ , simulamos  $(\theta^{(1)}, \Omega^{(1)}, Y^{(1)})$  y continuamos. A la  $m$ -ésima iteración con valores  $\theta^{(m)}, \Omega^{(m)}, Y^{(m)}$ :

Paso (a): Generar  $\Omega^{(m+1)}$  de  $p(\Omega | \theta^m, Y^m, Z)$ ;

Paso (b): Generar  $\theta^{(m+1)}$  de  $p(\theta | \Omega^{(m+1)}, Y^m, Z)$ ;

Paso (c): Generar  $Y^{(m+1)}$  de  $p(Y | \theta^{(m+1)}, \Omega^{(m+1)}, Z)$ .

Bajo condiciones leves y después de un número suficientemente largo de iteraciones la distribución conjunta de  $(\theta^{(m)}, \Omega^{(m)}, Y^{(m)})$  converge en una tasa exponencial a la distribución a posteriori deseada  $[\theta, \Omega, Y | Z]$ . La convergencia del muestreo de Gibbs puede ser revelada mediante gráficos de la secuencia simulada de los parámetros individuales y/o los valores EPSR (estimated potential scale reduction) correspondientes a los parámetros que se calculan secuencialmente mientras la ejecución procede.

## VII) Metodología

Para realizar este estudio se usó **datos reales extraídos de la página web del Ministerio de Educación del Perú (MINEDU), de la evaluación Nacional de Rendimiento Estudiantil Crecer** (Crecer con calidad y equidad en el rendimiento).

Este estudio es de tipo exploratorio, ya que se está explorando una técnica poco aplicada en el Perú, en el campo de la educación. Además, también es un estudio correlacional, puesto que se quiere demostrar que existe una relación entre ciertos factores (entorno escolar, socioeconómico, entorno familiar) y los resultados de comprensión lectora en los niños de 6to grado de primaria de lima metropolitana.

Se está usando fuentes de información secundaria, ya que los datos que se usan los que fueron recogidos en estudios previamente realizados, en este caso, del Ministerio de Educación del Perú.

### 7.1-) Metodología del MINEDU

Para el estudio original se evaluó a una muestra representativa de centros educativos polidocentes urbanos en el ámbito nacional de 17000 alumnos (de cada grado observado) de todo el Perú (estudiantes de centros educativos polidocentes completos de áreas urbanas tanto estatales como no estatales).

Los objetivos de este estudio original fueron:

- a) Evaluar a los estudiantes de primaria y secundaria.
- b). Recoger información sobre los factores asociados al rendimiento, con el fin de identificar a aquellos que muestran una mayor asociación con los resultados de aprendizaje de los estudiantes evaluados.

En el área de primaria se evaluaron a los estudiantes de 4to y 6to grado, mientras que en secundaria fueron los alumnos de 4to y 5to año. En cada grado fueron evaluados alrededor de 17000 estudiantes distribuidos en aproximadamente 578 centros educativos. El diseño muestral del proyecto, permitió tener una representatividad a escala nacional; se utilizó un sistema de muestreo estratificado bietápico y de conglomerados, siendo las escuelas las unidades seleccionadas en la 2da etapa. Sin embargo, para el presente estudio sólo se consideran a los **alumnos de 6to grado de primaria de Lima Metropolitana**, siendo un total de 3762 alumnos (Hombres: 50.2%; Mujeres: 49.8%)

Como además el estudio buscaba encontrar los factores asociados al rendimiento, se definieron tres poblaciones, la de padres de familia de los alumnos, profesores y directores. Estas poblaciones no fueron objeto especial de diseño, sino que estuvieron subordinadas a la población principal (la de los estudiantes que dieron las pruebas)

La información sobre los estudiantes fue recogida usando una encuesta sociodemográfica para los mismos alumnos, para los padres de familia ó apoderados, y para los directores. Se muestra un ejemplo en el ANEXO C.

## **7.2) Instrumentos**

A continuación, se detallarán los instrumentos que fueron usados por el MINEDU a fin de conseguir la información que se usó para el presente estudio:

*Encuestas.* - para poder recoger información sobre los factores asociados a la comprensión lectora, se aplicaron cuestionarios dirigidos al estudiante,

director, profesor y padres de familia. Pueden observarse estas encuestas en los links del MINEDU<sup>24</sup>

*Evaluación Comprensión Lectora.*- Para poder evaluar la comprensión lectora, se usaron las pruebas nacionales Crecer, que son pruebas de selección múltiples basadas en el modelo de normas. En este estudio se evaluaron a estudiantes de cuarto y sexto grados en las áreas curriculares de matemática, ciencias sociales, ciencias naturales y comunicación integral. Dentro de esta última categoría se encuentra un conjunto de preguntas que miden la comprensión lectora en todos sus aspectos (ver marco teórico)

Para verificar la validez de la prueba- que en forma sencilla se podría decir que es la medida en que ésta mide el constructo que se pretende medir (ó el grado en que la evidencia acumulada soporta las interpretaciones derivadas de los puntajes obtenidos en las pruebas)- se siguieron los siguientes criterios: el juicio de expertos, análisis de unidimensionalidad de las preguntas que componen las pruebas (la propiedad de una prueba de medir únicamente un constructo (unicidad de la prueba medible).), confiabilidad de las pruebas, y otras características como el nivel de dificultad, el grado de discriminación y los índices de no respuesta.

En cuanto a la confiabilidad de una prueba, que mide el grado en que es consistente con los puntajes que de ella se obtienen, se usa generalmente el coeficiente “alfa” de Cronbach. Todas las pruebas arrojaron altos índices de consistencia interna, estando el rango en cada una de las categorías observadas entre 0.73 y 0.85.

En cuanto a la unidimensionalidad, las pruebas arrojaron un grado de 70% (porcentaje de la varianza explicada de la primera dimensión), por lo que se concluye que el conjunto de preguntas mide un solo aspecto

### **7.3) Procedimiento**

En el estudio original, del cual, se aplicó a los estudiantes las pruebas Crecer, y además se les proporcionó la encuesta destinada a recoger la

información sobre posibles factores asociados. De igual manera, otra encuesta fue proporcionada a los profesores, directores y padres de familia.

La base de datos completa obtenida, cuenta con 16925 registros. Sin embargo, al estar este estudio centrado en los estudiantes de Lima, esta total disminuye a 3762.

Se realizó además un filtro por la edad, de tal manera que solo hubiera estudiantes entre 10 y 12 años (se excluye 674 casos por tener datos perdidos).

	Frequency	Percent	Valid Percent
Valid			
10	200	5.3	5.4
11	1765	46.9	47.6
12	1066	28.3	28.8
13	288	7.7	7.8
14	109	2.9	2.9
15	38	1.0	1.0
16	18	.5	.5
17	6	.2	.2
18	7	.2	.2
97	120	3.2	3.2
99	88	2.3	2.4
Total	3705	98.5	100.0
Missing			
System	57	1.5	
Total	3762	100.0	

Otras exclusiones se hicieron por datos perdidos, quedándonos con un total de 1862 registros. De ellos el porcentaje de hombres es de 47% y el de mujeres es de 53%.

La base de datos final, cuenta con 10 variables dicotómicas, que son mostradas en el siguiente cuadro de operacionalización de las variables. La variable dependiente es un indicador de si comprende o no. Las restantes 9 son las variables independientes, siendo las 3 siguientes variables manifiestas usadas como indicadores para la primera variable latente que llamaremos:

“Entorno Familiar”, las siguientes 3, explican la variable latente: “Factor Socioeconómico” y las últimas 3 variables explican el “Entorno Escolar”

Variable Dependientes	Descripción de la variable	Indicador	Unidades	Escala
Comprende	Indicador de si el alumno tiene un nivel satisfactorio en comprensión lectora	Pruebas "Crecer"	1:Si ,0:No	Binaria
Variables Independientes	Descripción de la variable	Indicador	Unidades	Escala
AyudaPadre	Indicador de si el alumno recibe ayuda del padre	Encuesta para alumnos	1:Si ,0:No	Binaria
Con libros	Indicador de si el alumno cuenta con la suficiente cantidad de libros en casa(fuera de textos escolares, periódicos, revistas)	Encuesta para alumnos	1:Si ,0:No	Binaria
No molestado	Indicador de si el alumno cuenta con un lugar tranquilo donde pueda leer	Encuesta para alumnos	1:Si ,0:No	Binaria
Padresconeduc	Indicador de si los padres del alumno cuentan con educación secundaria completa	Encuesta para padres	1:Si ,0:No	Binaria
ConElectro	Indicador de si el alumnos posee los electrodomésticos básicos en el hogar	Encuesta para padres	1:Si ,0:No	Binaria
Serv_basicos	Indicador de si el alumno cuenta con agua/luz/desagüe en el hogar	Encuesta para padres	1:Si ,0:No	Binaria
Desayuno	Indicador de si el colegio donde estudia el alumno cuenta con el programa desayuno escolar	Encuesta para directores	1:Si ,0:No	Binaria
Asist_medica	Indicador de si el colegio donde estudia el alumno cuenta con Asistencia Médica	Encuesta para directores	1:Si ,0:No	Binaria
Mobiliarioescolar	Indicador de si el colegio donde estudia el alumno cuenta con Mobiliario Escolar	Encuesta para directores	1:Si ,0:No	Binaria

Operacionalización de las variables

Puede verse la estructura de los datos en el Anexo C, donde se presenta una muestra.

El análisis fue desarrollado en el programa de uso libre OpenBUGS, un software diseñado para el análisis Bayesiano de modelos estadísticos complejos usando el método de Cadenas de Markov Monte Carlo.

## VIII) Resultados

El conjunto final de datos dicotómicos es analizado usando un modelo como el definido en las ecuaciones (1) y (2) del marco teórico. Pueden ser consideradas otras estructuras para la matriz de cargas, sin embargo usamos ésta que nos da factores latentes no superpuestos para una clara interpretación.

Por lo tanto se usará la siguiente matriz de cargas  $\Lambda$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \lambda_{32} & \lambda_{42} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{63} & \lambda_{73} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{94} & \lambda_{104} \end{pmatrix}$$

Donde  $\lambda_{ij}$  son los parámetros de factor de carga desconocidos, y los unos y ceros son fijados en la estimación para conseguir un modelo.

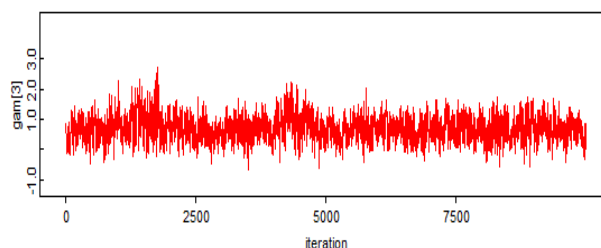
Resumiendo, se tiene la variable “Comprensión Lectora” ( $\eta$ ) factores no superpuestos (variables latentes), entorno familiar ( $\xi_1$ ), factor socioeconómico ( $\xi_2$ ) y Programas Escolares ( $\xi_3$ ).

Cómo nuestro principal interés es estudiar los efectos lineales de los 3 factores latentes en “comprensión lectora”, la ecuación estructural es escogida para ser un modelo de regresión lineal que relaciona  $\eta$  en  $\xi_1, \xi_2, \xi_3$  es

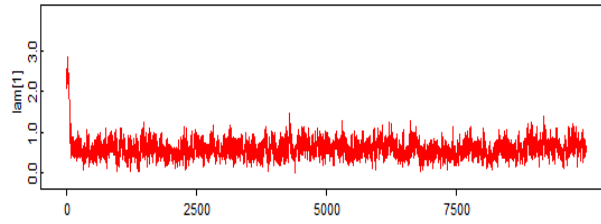
$$\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_3 + \delta$$

donde  $\gamma_1, \gamma_2$  y  $\gamma_3$  son parámetros desconocidos. Otros parámetros desconocidos: la varianza y covarianza de  $\xi_1, \xi_2$ , y  $\xi_3$ :  $\phi_{11}, \phi_{22}, \phi_{33}, \phi_{12}, \phi_{13}, \phi_{23}$  y la varianza de  $\delta, \psi_\delta$ . En total hay 16 parámetros desconocidos

Con la ayuda del programa Winbugs se usará el algoritmo de Gibbs, para obtener los estimados bayesianos de los parámetros. Observamos que el estimador de Gibbs en la estimación real converge en aproximadamente 5000 iteraciones. Para asegurar la convergencia, usamos los gráficos de los valores de algunos parámetros y el número de iteraciones (A continuación se presentan algunos gráficos, la totalidad de ellos pueden ser vistos en el anexo A)







Para obtener los estimadores bayesianos, se recogen 5000 observaciones después de descartar las primeras 5000 iteraciones. Los resultados son presentados en la siguiente tabla.

Parámetro		mean	sd	MC_error
$\gamma_1$	gam[1]	-0.1426	0.4105	0.02818
$\gamma_2$	gam[2]	2.142	0.4851	0.03977
$\gamma_3$	gam[3]	-0.6777	0.3596	0.01842
$\lambda_{32}$	lam[1]	0.5857	0.1943	0.01211
$\lambda_{42}$	lam[2]	1.928	0.3171	0.02546
$\lambda_{63}$	lam[3]	2.311	0.4238	0.03691
$\lambda_{73}$	lam[4]	3.122	0.3969	0.03147
$\lambda_{94}$	lam[5]	3.55	0.5471	0.0406
$\lambda_{10\ 4}$	lam[6]	2.952	0.4466	0.03117
$\phi_{11}$	phx[1,1]	0.1671	0.03685	0.003313
$\phi_{12}$	phx[1,2]	0.1154	0.01995	0.001491
$\phi_{13}$	phx[1,3]	-0.0176	0.01052	0.0006521
$\phi_{22}$	phx[2,2]	0.1422	0.03004	0.002775
$\phi_{23}$	phx[2,3]	-0.02345	0.009185	0.0006255
$\phi_{33}$	phx[3,3]	0.06967	0.01574	0.001311
$\psi_\delta$	sgd	1.468	0.5643	0.05993

Basados en estos estimados,  $\hat{\Sigma}$  y  $\Phi$  la matriz de covarianzas estimadas del vector aleatorio observado  $y$ , y el vector aleatorio latente  $\omega$ , pueden ser calculados.

Los resultados más importantes:

De los estimados de  $\gamma_1$ ,  $\gamma_2$  y  $\gamma_3$  en la ecuación estructural, vemos que  $\xi_1$ ,  $\xi_2$ ,  $\xi_3$  tiene efectos significativos en  $\eta$ . Por otra parte, si observamos la ecuación estructural:

$$\eta = -0.14 \xi_1 + 2.1 \xi_2 - 0.67 \xi_3$$

Notamos que el factor que más influye en la comprensión lectora es el socioeconómico, hecho que viene a confirmar todos los estudios hechos en el Perú y a nivel mundial. Además vemos que en este caso, el factor “entorno familiar” no presenta mayor relevancia. Mientras que el factor “Programas sociales escolares”, presenta una relación negativa con nuestra variable objetivo. Lo cual a simple vista puede llamar la atención, sin embargo esto se debe a que estos programas sociales están asociados a los niveles económicos bajos, lo que nos lleva a observar que estos programas, por sí solos, no contribuyen a una mejora en la comprensión lectora.

Analizando la tabla de estimados, vemos que la correlación entre  $\xi_2$   $\xi_3$  es  $-0.0235$ . Como no se cuenta con la información de estos estimados completamente standarizados, solo podemos decir que existe una relación inversa entre ambos factores, es decir que el factor socioeconómico esta correlacionado negativamente con los programas sociales. De igual manera vemos que la correlación entre  $\xi_1$  y  $\xi_2$  es  $0.11$  lo cual nos dice que mientras mayor sea el nivel socioeconómico, mejor será el entorno familiar.

## **IX) Conclusiones**

El objetivo principal del presente trabajo ha sido explorar y proponer un modelo de ecuaciones estructurales para explicar los factores que influyen en la comprensión lectora, usando sólo variables dicotómicas.

Para tal fin se hizo uso de la estadística bayesiana, usando además información recogida por el Ministerio de educación. El presente documento propone un modelo que relaciona la comprensión lectora con los siguientes 3 factores: Factor socioeconómico, factor entorno familiar, y factor Programas Escolares, de donde se puede deducir diversas conclusiones.

En primer lugar, se cumplió con el objetivo principal, creando un modelo de ecuaciones estructurales para explicar la comprensión lectora, solo con variables dicotómicas.

El factor socioeconómico ha demostrado ser el que presenta mayor influencia en los resultados de comprensión lectora en los alumnos de 6to grado de Lima Metropolitana, hecho que viene a concordar con todos los estudios hechos hasta el momento sobre este tema. Una de las razones del papel importantísimo de este factor puede deberse a que los niños de nivel económico más alto tienen acceso a una mejor calidad de educación.

Lo cual nos lleva a concluir que mientras no se haga nada por mejorar el nivel de vida de los niños de los niveles socioeconómicos más bajos, es muy difícil que los resultados en la comprensión lectora mejore.

El factor Programas escolares, tiene un efecto pequeño (negativo) en la comprensión lectora. Una de las razones por la que podría pasar esto es que estos programas impartidos por el gobierno se realizan en colegios de las zonas más pobres de Lima. Esto podemos corroborarlo analizando la correlación entre este factor y el socioeconómico, el cual es negativo.

Este resultado en particular, tiene mucha importancia, ya que nos dice que estos programas por sí solos, no ayudan a mejorar la comprensión lectora de los alumnos, sino que deberían ir acompañados además por una estrategia educacional. Lo cual no quiere decir que estos programas no sean útiles, ni dejen de impartirse, ya que ayudan en otros aspectos como nutrición, etc. pero debe dárseles un mejor enfoque.

El tercer factor: “Entorno familiar” no presenta influencia relevante en la comprensión lectora en este estudio. Presenta sin embargo una correlación positiva con el factor socioeconómico, lo cual confirma otros estudios donde se afirma que en los niveles socioeconómicos más altos los familiares del niño se preocupan más por su educación que en los niveles más bajos.

Es importante recalcar que este estudio tiene algunas limitaciones ya que no se han tenido en cuenta otros factores que también pudieron resultar importantes como el tipo de colegio (estatal, particular), etc.

Además, queda como tarea pendiente para futuros estudios, realizar un modelo de ecuaciones estructurales para explicar la comprensión lectora, usando no solo un tipo de variable, sino todos los diversos existentes (continua, ordinal, binaria).

## **X) Referencias Bibliográficas**

- 1) Junior Cabral, La importancia de la lectura.
- 2) Resultados de la Evaluación Censal de Estudiantes de Segundo Grado 2012. Dirección URL:  
[http://www.minedu.gob.pe/DelInteres/xtras/conferencia\\_de\\_prensa\\_ece\\_ministra\\_-\\_version\\_final\\_02.04.13.pdf](http://www.minedu.gob.pe/DelInteres/xtras/conferencia_de_prensa_ece_ministra_-_version_final_02.04.13.pdf).
- 3) Fundación BBVA Continental, Leer es estar adelante. Dirección URL:  
<http://www.leer.pe/>
- 4) Structural Equation Modeling in Educational Research, Concepts and Applications, Timothy Theo.
- 5) Sik Yum Lee: Structural Equation Modeling, A Bayesian Approach, (2007)
- 6) Silvia Morales: Socio-Cultural variables and Reading literacy acquisition in Lima, Peru (2009)
- 7) Anderson, A. y W. Teale (2007) La lectoescritura como práctica cultural.
- 8) Ferreiro y M. Gómez-Palacio (Eds.), Nuevas perspectivas sobre los procesos de lectura y escritura. Madrid: Siglo XXI
- 9) Gómez-Palacio, M. (1993) Indicadores de la comprensión lectora. Washington: OEA.
- 10) Van Dijk, T. (1998). Estructura y funciones del discurso. Madrid: Siglo XXI. Dirección URL:  
<http://www.discursos.org/oldbooks/Teun%20A%20van%20Dijk%20-%20Las%20Estructuras%20y%20Funciones%20del%20Discurso.pdf>
- 11) Evaluación Censal de Estudiantes (ECE), Segundo grado de primaria, marco de trabajo. Dirección URL:  
[http://www2.minedu.gob.pe/umc/ece/Marco\\_de\\_Trabajo\\_ECE.pdf](http://www2.minedu.gob.pe/umc/ece/Marco_de_Trabajo_ECE.pdf)

- 12) Arturo Calderón, Cholly Farro, Jorge Bazán: Diseño Muestral en la aplicación nacional CRECER 1998.
- 13) Martín Benavides: Para explicar las diferencias en el rendimiento en matemática de cuarto grado en el Perú Urbano: Análisis de Resultados a partir de un modelo básico.
- 14) Sylvia Defior Citoler: Las dificultades del aprendizaje: un enfoque cognitivo: Lectura, escritura, matemáticas, 1996.
- 15) Thorne, C & Pinzás: Factors affecting reading achievement in Perú.
- 16) González, A. (2004). Estrategias de comprensión lectora.
- 17) Linda Baker, Deborah Scher & Kirsten Mackler: Home and family influences on motivations for reading.
- 18) Ann Kennedy, Katleen Trong: Influence of the Home Literacy environment on reading motivation and reading comprehension.
- Dirección URL:
- [http://www.iea.nl/fileadmin/user\\_upload/IRC/IRC\\_2010/Papers/IRC2010\\_Kennedy\\_Trong.pdf](http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2010/Papers/IRC2010_Kennedy_Trong.pdf)
- 19) Eduardo Velez: Factores que afectan el rendimiento académico en la educación primaria; Martín Benavides: Para explicar las diferencias en el rendimiento en matemática de cuarto grado en el Perú Urbano
- 20) Anders Skrondal, Sophia Rabe-Hesketh: Structural Equation Modeling: Categorical Variables (2005).
- 21) Scheines, R. Hoijtink, and Boomsma (1999) Bayesian estimation and testing of structural equation models. Dirección URL:
- <http://www.hss.cmu.edu/philosophy/scheines/psychometrika.pdf>

22) Kendra R. Tannembaum, Joseph, K. Torgesen, Richard K. Wagner (2006): Relationships between Word Knowledge and Reading Comprehension in Third Grade Children, Scientific Studies of Reading, 10:4, 381-398.

23) Ana Cecilia Cubas Barrueto: Actitudes hacia la lectura y niveles de comprensión lectora en estudiantes de sexto grado de primaria.

24) Encuestas Ministerio de Educacion. Link:

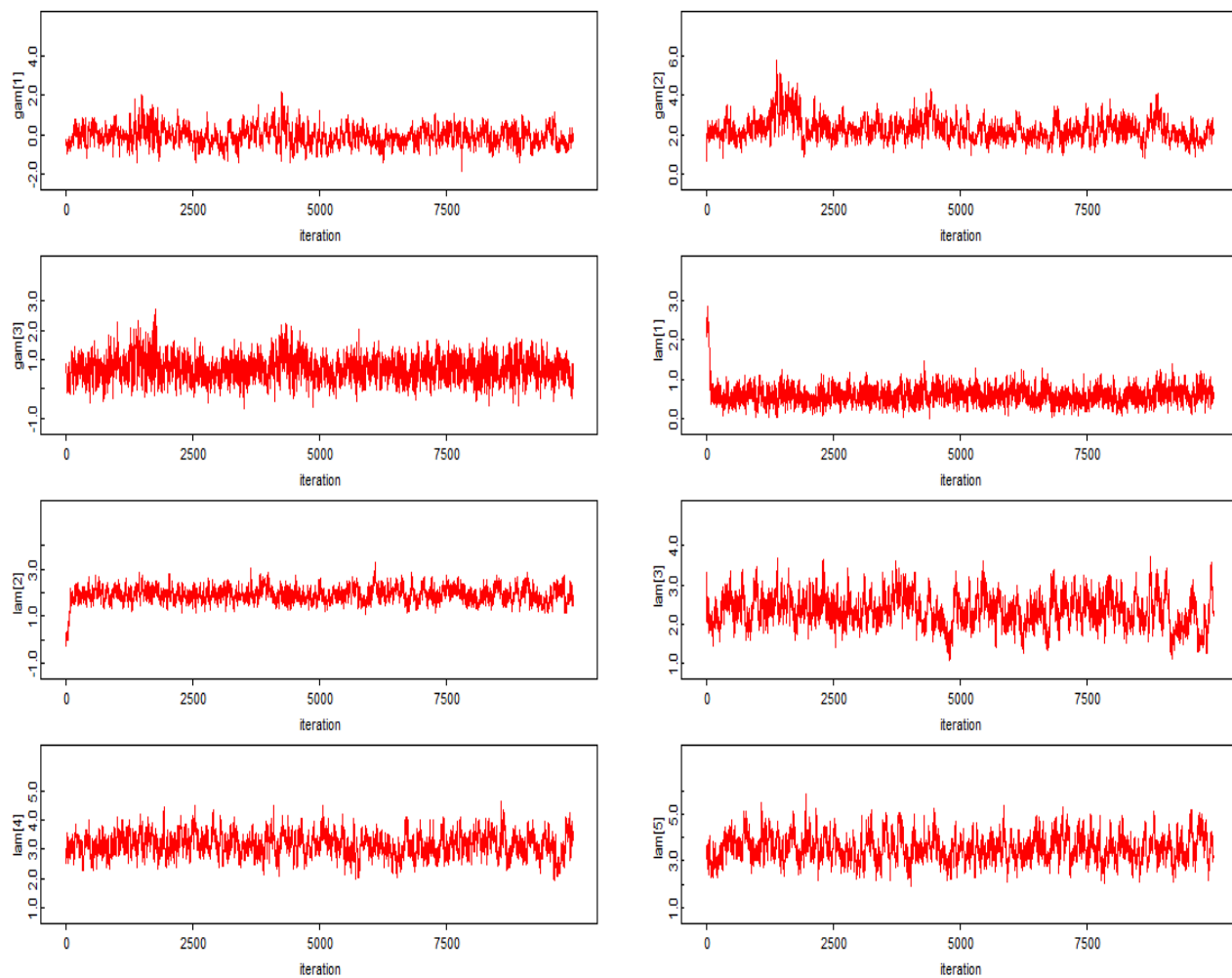
[http://www2.minedu.gob.pe/umc/admin/images/archivosprincipal/archivo\\_pri\\_32.pdf](http://www2.minedu.gob.pe/umc/admin/images/archivosprincipal/archivo_pri_32.pdf)

[http://www2.minedu.gob.pe/umc/admin/images/archivosprincipal/archivo\\_pri\\_35.pdf](http://www2.minedu.gob.pe/umc/admin/images/archivosprincipal/archivo_pri_35.pdf)

[http://www2.minedu.gob.pe/umc/admin/images/archivosprincipal/archivo\\_pri\\_34.pdf](http://www2.minedu.gob.pe/umc/admin/images/archivosprincipal/archivo_pri_34.pdf)

# ANEXO A

## CUADROS DE CONVERGENCIA





## ANEXO B

### CÓDIGO OPENBUGS

```

model{
for(i in 1:N){

for(j in 1:P){y[i,j]~dnorm(mu[i,j],psi[j])l(low[z[i,j]+1],high[z[i,j]+1])}
mu[i,1]<-eta[i]
mu[i,2]<-xi[i,1]
mu[i,3]<-lam[1]*xi[i,1]
mu[i,4]<-lam[2]*xi[i,1]
mu[i,5]<-xi[i,2]
mu[i,6]<-lam[3]*xi[i,2]
mu[i,7]<-lam[4]*xi[i,2]
mu[i,8]<-xi[i,3]
mu[i,9]<-lam[5]*xi[i,3]
mu[i,10]<-lam[6]*xi[i,3]

xi[i,1:3]~dmnorm(u[1:3],phi[1:3,1:3])
eta[i]~dnorm(nu[i],psd)
nu[i]<-gam[1]*xi[i,1]+gam[2]*xi[i,2]+gam[3]*xi[i,3]
} #end of i
for(j in 1:P){psi[j]<-1.0}
for(j in 1:3){u[j]<-0.0}
#priors on loadings and coefficients
lam[1]~dnorm(3.12,4.0) lam[2]~dnorm(3.12,4.0) lam[3]~dnorm(3.32,4.0)
lam[4]~dnorm(3.10,4.0) lam[5]~dnorm(4.30,4.0) lam[6]~dnorm(3.14,4.0)

var.gam<-4.0*psd
gam[1]~dnorm(-1.0,var.gam) gam[2]~dnorm(0.86,var.gam) gam[3]~dnorm(0.9,var.gam)
#priors on precisions
psd~dgamma(8.0, 10.0)
sgd<-1/psd
phi[1:3,1:3]~dwish(R[1:3,1:3], 9)
phx[1:3,1:3]<-inverse(phi[1:3,1:3])
}

```

ANEXO C  
(Muestra 40 primeros casos)

Comprende, ayudatareas, Nomolestado, Conlibros, Serv\_basicos, ConElectro, Padresconeduc, Asist\_medic, Desayuno, Mobiliarioescolar

0,0,1,0,0,0,0,1,1,0,  
0,0,1,0,0,0,1,1,1,0,  
0,0,1,0,0,0,0,1,1,0,  
0,1,1,0,0,0,0,1,0,0,  
0,0,1,0,1,0,0,1,0,0,  
0,1,0,1,1,0,0,1,0,0,  
0,1,1,0,1,0,0,1,1,1,  
0,0,1,1,0,0,1,1,0,0,  
0,0,1,1,1,0,0,1,0,0,  
0,0,1,0,0,0,0,1,0,1,  
0,0,1,1,1,0,0,1,0,1,  
0,0,0,0,1,0,1,1,0,1,  
0,1,1,1,1,0,1,1,1,1,  
0,0,1,0,0,0,1,1,1,1,  
0,1,0,1,1,1,1,1,1,0,  
0,1,0,0,1,0,1,1,1,0,  
0,1,1,0,1,0,0,1,1,0,  
0,0,0,0,0,0,0,1,1,1,  
0,1,0,0,0,0,1,1,1,1,  
0,0,1,0,1,0,1,1,1,1,  
0,1,1,1,1,0,0,1,1,1,  
0,0,1,1,1,0,0,1,1,1,  
0,1,0,1,1,0,0,1,1,1,  
0,0,1,0,1,1,0,1,1,1,  
0,0,0,0,0,0,0,1,0,1,  
0,0,0,1,1,0,1,1,1,1,  
0,0,1,1,1,1,0,1,1,1,  
1,1,1,1,0,0,0,1,0,1,  
0,1,1,0,0,1,0,1,0,1,  
0,1,1,1,1,0,0,1,0,1,  
0,0,1,0,0,0,0,1,0,1,  
0,0,1,0,1,0,0,1,1,1,  
1,0,1,1,1,1,1,1,1,1,  
1,0,0,0,1,0,1,1,1,1,  
1,0,1,0,1,0,0,1,1,1,  
0,1,0,0,1,1,1,1,0,1,  
1,0,0,1,0,0,1,1,1,1,