Las **finanzas públicas modernas** se constituyen en un punto de encuentro al que concurren varias disciplinas, que van desde la filosofía y sociología del estado -considerando los aspectos técnicos y sociológicos de las finanzas públicas- pasando por la metodología de la investigación, y así hasta llegar a las tecnologías de la informática y las comunicaciones. No se concibe que este encuentro sea otro que el de la **multi e inter disciplina**; es decir, que no se concibe que las disciplinas se presenten de una manera aislada, sino que se integran generando ensambles asociados a problemáticas generales –las que son el objeto de las finanzas públicas–. El perfil de un **investigador en finanzas públicas modernas** requiere una serie de competencias que para su aprendizaje necesitan de procesos formativos, que se dan a partir no sólo del estudio, sino que indudablemente requieren del ejercicio mismo de la investigación; esto es muy claro, sobre todo, en las competencias para la investigación (se asume que no se pude aprender a hacer investigación, sin hacer investigación). Así entonces, el escenario de formación de un investigador en finanzas públicas sólo se concibe en un encuentro armonioso de vertientes disciplinarias asociadas al **proceso de investigación**. Entre estas vertientes disciplinarias ensambladas destacan las que atienden el instrumental metodológico, donde se ubica a la **metodología estadística**. Es decir, que la estadística se presenta, en este contexto, aunada a los procesos de diseño y desarrollo de la investigación en finanzas públicas –que necesariamente deben culminar con la elaboración de productos, como reportes para presentaciones en congresos o artículos científicos–.

En este libro desarrollamos una serie de temáticas de metodología estadística ilustradas con problemas de finanzas públicas y presentamos los resultados –en la forma de **artículos científicos**- de varios ejercicios de aplicación en problemas relevantes de esta multi e inter disciplina.

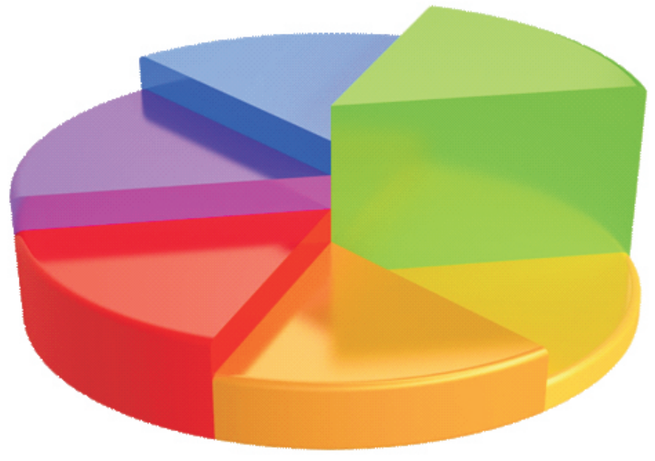
ISBN: 978-607-00-5322-1
Xalapa, Veracruz, México

Ojeda
Velasco
Cruz
Tapia

Metodología Estadística Aplicada
a las Finanzas Públicas

METODOLOGÍA ESTADÍSTICA APLICADA A LAS FINANZAS PÚBLICAS

Mario Miguel Ojeda Ramírez
Fernando Velasco Luna
Cecilia Cruz López
Patricia Tapia Blázquez



2011

Xalapa, Ver., México
Diciembre 2011

ISBN: 978-607-00-5322-1

Metodología Estadística Aplicada a las Finanzas Públicas

Mario Miguel Ojeda Ramírez
Fernando Velasco Luna
Cecilia Cruz López
Patricia Tapia Blásquez

Edición y Formación: M.C. Cecilia Cruz López

Título:	Metodología Estadística Aplicada a las Finanzas Públicas / Mario Miguel Ojeda Ramírez, Fernando Velasco Luna, Cecilia Cruz López, Patricia Tapia Blásquez.
Edición:	Primera edición.
Pie de imprenta:	Xalapa, Veracruz, México, 2011.
Descripción física:	285 p.
Serie:	(Libros universitarios)
Nota:	Incluye bibliografías.
ISBN:	978-607-00-5322-1
Materias:	Metodología estadística Estadística multivariada Finanzas públicas
Autores:	Ojeda Ramírez, Mario Miguel. Velasco Luna, Fernando. Cruz López, Cecilia. Tapia Blásquez, Patricia.

Primera edición, diciembre 2011
ISBN: 978-607-00-5322-1

Impreso en México
Printed in Mexico

Contenido

Lista de Figuras

Introducción

Las finanzas públicas modernas se constituyen en un punto de encuentro al que concurren varias disciplinas, que van desde la filosofía y sociología del estado -considerando los aspectos técnicos y sociológicos de las finanzas públicas- pasando por la metodología de la investigación, y así hasta llegar a las tecnologías de la informática y las comunicaciones. No se concibe que este encuentro sea otro que el de la multidisciplina; es decir, que no se concibe que las disciplinas se presenten de una manera aislada, sino que se integran generando ensambles asociados a problemáticas generales –las que son el objeto de las finanzas públicas–. De esta forma en el perfil de un investigador en finanzas públicas modernas podemos identificar una serie de competencias que requieren para su aprendizaje de procesos formativos, que se dan a partir no sólo del estudio, sino que indudablemente requieren del ejercicio mismo de la investigación; esto es muy claro, sobre todo, en las competencias para la investigación (se asume que no se puede aprender a hacer investigación, sin hacer investigación). Así entonces, el escenario de formación de un investigador en finanzas públicas sólo se concibe en un encuentro armonioso de vertientes disciplinarias asociadas al proceso de investigación. Entre estas vertientes disciplinarias ensambladas con el proceso de investigación destacan las que atienden el instrumental metodológico, donde se ubica a la metodología estadística. Es decir, que la estadística se presenta aunada a los procesos de diseño y desarrollo de la investigación en finanzas públicas –que necesariamente deben culminar con la elaboración de productos, como reportes para presentaciones en congresos o artículos científicos–.

Se sabe y se reconoce ampliamente que la estadística es una herramienta fundamental para la realización de procesos de investigación en ciencias fácticas que utilizan la investigación cuantitativa. Los diseños estadísticos son los principios y procedimientos que permiten obtener los datos pertinentes, acorde a las restricciones –de tiempo y recursos- y para suplir las necesidades de información –que se hacen explicitas a través de las preguntas de investigación–. Los estudios observacionales y los de muestreo son generalmente los tipos generales de diseños estadísticos a los que se hace referencia cuando se protocoliza una investigación en el área de las finanzas públicas. En cada caso hay que especificar algunos elementos clave como la fuente de los datos, la población

objetivo, las unidades de estudio, las variables a medir, las escalas y los métodos de medición, el tamaño de la muestra, etc. La caracterización adecuada de estos elementos define el diseño particular de la investigación y establece la estructura de la base de datos con la que se van a realizar los análisis; a partir de una clara definición de estos elementos se puede bosquejar la metodología de análisis estadístico; es decir, los pasos a seguir para realizar el procesamiento de los datos, dónde otra vez las preguntas de investigación son la guía fundamental. Todo esto se establece en el protocolo de investigación, que debe incluir de un marco conceptual, un marco teórico, una revisión de antecedentes –lo que se llama el estudio del estado de la cuestión- y una clara definición de objetivos, seguida de una precisa delimitación del problema en estudio. Si vemos así diseccionada esta fase de la investigación, ya considerando los principios y las técnicas de la metodología estadística, podremos entender cómo se ensambla la metodología estadística al proceso de investigación en finanzas públicas.

Ahora bien, debemos considerar que la metodología estadística comprende tres grandes pasos en el desarrollo de una investigación: 1) el diseño adecuado para la obtención de datos; 2) el análisis de éstos; y 3) la interpretación y presentación de los resultados en forma apropiada. Todo esto se deberá definir y protocolizar en un documento, que es precisamente el protocolo de la investigación. En este sentido el diseño de la investigación es la guía que conduce todo el proceso; desempeña el mismo papel que el itinerario en un viaje, es el que lleva al investigador de un punto inicial u origen, al sitio final o resultados. Asimismo, conduce a la formulación de la metodología que se utilizará para obtener los datos de acuerdo con las necesidades de información. Entre los criterios que se emplean para formular la metodología de trabajo está que los datos se colecten de la manera más rápida, económica y sencilla; es necesario también conducir un procedimiento para garantizar la calidad de los datos. El análisis de los datos procede a partir de una serie de métodos y procedimientos para explotar los datos de manera tal que sea posible extraer de ellos la información relevante, tal que resuelva las preguntas que dieron origen al estudio –las llamadas preguntas de investigación–. Finalmente, en la interpretación y presentación de los resultados, una serie de principios y procedimientos de la estadística proporcionan los lineamientos generales para elaborar los formatos de presentación y

elaboración de tablas y figuras –incluyendo bajo este rubro a lo que también se llama cuadros y gráficas–, además de proporcionar los elementos para construir los juicios de valor a partir de los resultados de los análisis estadísticos.

En este sentido el investigador de las finanzas públicas en su enfoque moderno requiere de una formación sólida en estadística –pero claramente en este enfoque integral– que implica contar con un marco conceptual y una serie de motivaciones que propicien una reflexión y una evaluación positiva hacia esta metodología, buscando con todo esto un cambio de actitud hacia el uso de las técnicas y métodos estadísticos en el proceso de investigación en ciencias sociales –en particular en la inter y trans disciplina que implica la investigación en finanzas públicas–. Así, la adquisición de las competencias de un usuario de la metodología estadística en este contexto, requiere la observancia de una serie de habilidades para identificar en este marco problemas de finanzas públicas de índole estadística –hay que decir que muchos lo son–, y proponer estrategias generales de solución. En suma, en la formación del investigador en finanzas públicas modernas, se busca dotarlo de las competencias para diseñar y desarrollar estudios estadísticos, con énfasis en estudios observacionales y de muestreo, que consideren el uso de técnicas exploratorias univariadas y multivariadas, pero también el uso de la modelación estadística, implementadas todas estas técnicas a partir de software estadístico. Concretamente los objetivos de formación para un investigador en finanzas públicas modernas serían:

- Analizar el proceso de aplicación de la estadística en el contexto de investigaciones sociales, y particularmente en el contexto de los problemas de las finanzas públicas modernas.
- Identificar las fases del proceso del diseño estadístico, clasificando y caracterizando los diferentes tipos de estudios estadísticos.
- Caracterizar particularmente el proceso de diseño y análisis de un estudio observacional y de un estudio de muestreo en el marco de las finanzas públicas.
- Diseñar e implementar el proceso de obtención de datos y verificación de la calidad de los mismos.

- Caracterizar los elementos de una estrategia para el análisis estadístico de los datos en una investigación particular, considerando como referencia el análisis inicial y el análisis definitivo –el que se puede incluir el proceso de modelación estadística–, en presencia de facilidades computacionales.
- Identificar y caracterizar los elementos distintivos del análisis multivariado de naturaleza exploratoria y descriptiva.
- Diseñar e implementar procesos de aplicación que impliquen las técnicas estadísticas multivariantes de naturaleza exploratoria.
- Adquirir las habilidades para plantear, ajustar e interpretar modelos estocásticos, particularmente modelos estadísticos lineales.
- Diseñar y desarrollar la presentación de resultados de la investigación en formato de presentación en congresos y escritura de artículos científicos.

Es precisamente con este enfoque que se ha venido desarrollando el curso taller de estadística aplicada a las finanzas públicas, el cual se ha impartido en las cinco últimas ediciones del doctorado. Los resultados han ido mejorándose, al igual que el diseño y desarrollo de esta experiencia educativa; hemos llegado hasta el nivel de tener una integración de materiales de estudio, prácticas de investigación y de uso de software estadístico, lo que ha hecho que los estudiantes tengan una guía cada vez más precisa –de lo que deben hacer y cómo lo deben hacer– en su proceso formativo.

Presentamos aquí la primera edición de una suerte de memoria de esta experiencia, la cual hemos titulado Metodología Estadística Aplicada a las Finanzas Públicas. Está integrada por cuatro partes: en la primera presentamos aspectos generales del diseño estadístico y del que llamamos análisis estadístico básico; en la segunda parte se hace una presentación de las técnicas multivariantes de naturaleza exploratoria; la tercera la dedicamos a la modelación estadística con énfasis a la modelación lineal multinivel, que encuentra una gran veta de aplicación en las problemáticas de las finanzas públicas; la cuarta parte es una selección de artículos que presentan aplicaciones concretas.

El diseño y conducción de este proyecto ha corrido bajo mi responsabilidad, pero debo reconocer que no podría haberlo hecho si no hubiese contado con el apoyo de Fernando Velasco, Patricia Tapia y Cecilia Cruz –en reconocimiento a su trabajo aparecen como coautores–; ellos han compilado y escrito versiones anteriores, y han corregido y mejorado, los materiales que aquí presentamos. Varios de los artículos que incluimos en la última parte fueron realizados en versiones preliminares por algunos de los estudiantes del Doctorado en Finanzas Públicas en su última generación que cursó esta experiencia educativa, pero fueron complementados y mejorados por algunos de nosotros, incluida la participación de Yesenia Zavaleta, que colaboró en el equipo; es por tal motivo que los artículos mencionados aparecen firmados por más de un autor. En esta colección de trabajos hay dos que se desarrollaron fuera del contexto del curso, que son el que coautora Roberto Gallardo y en el que aparecemos Patricia Tapia y yo. Los criterios de revisión e inclusión me los reservé, y por tanto soy enteramente responsable de los que aparecen, y por supuesto de la exclusión de algunos otros que inicialmente consideramos para la memoria, pero que no se llegaron a incluir.

Finalmente debo agradecer la colaboración de la doctora Minerva Montero, de la Academia de Ciencias de Cuba –que realizó una estancia de investigación en la Universidad Veracruzana mientras trabajábamos en el proyecto–; ella nos ayudó mucho, diseñando estrategias de análisis de datos, particularmente de modelación multinivel, leyendo los materiales y los artículos, dando sugerencias de mejora, y también revisando las mejoras; en fin, que debo reconocer que no aparece como coautora solamente porque expresamente así lo decidió. Un agradecimiento especial a Roberto Gallardo quien leyó la versión final de los materiales y dio algunas sugerencias que atendimos de última hora. También reconozco al coordinador del doctorado, Julio Cesar Sosa, quien siempre nos animó a concluir este libro. Agradeceremos de antemano las observaciones y sugerencias para mejorar este material en ediciones futuras.

Xalapa, Veracruz, México, octubre de 2011.

Mario Miguel Ojeda Ramírez.

I. Metodología Estadística

1.1 La buena cultura estadística

El conocimiento se obtiene mediante un proceso de estudio al que llamaremos genéricamente investigación. El conocimiento derivado de investigaciones fácticas o factuales implica que se defina un problema, que se establezcan preguntas de investigación, que se definan necesidades específicas de información, lo que lleva a requerir datos, y es en este contexto donde aparece también la necesidad de utilizar la metodología estadística. A menudo se piensa que usar la metodología estadística para el desarrollo de una investigación es una tarea compleja que solamente se puede llevar a cabo con el apoyo de un asesor estadístico. Sin embargo, el avance tecnológico de las últimas décadas ha traído un desarrollo rápido de software estadístico, de uso fácil, que ha hecho que los investigadores, a veces en equipo con consultores estadísticos, logren usar adecuadamente la estadística en sus investigaciones; de esta manera se ha ido derribando el mito de que la estadística es una disciplina compleja que solamente puede ser aplicada correctamente por especialistas estadísticos.

La metodología estadística, siguiendo la concepción de que esta es empleada en el desarrollo de investigaciones fácticas, según Ojeda y Velasco (2010, p.1), "…es la disciplina que se encarga de la captación, manejo y presentación de información numérica, que de acuerdo a algún objetivo definido en el contexto de una investigación o estudio se requiere." La mayoría de las definiciones coinciden en que a través de la estadística se obtienen los datos, se procesan y finalmente se presenta la información relevante para la investigación; si bien es cierto, podemos decir entonces que el procedimiento de la buena cultura estadística es precisamente lo que la mayoría de las definiciones de estadística afirman que es: una ciencia que recolecta, procesa, analiza y presenta conclusiones emanadas de los resultados de diferentes procesamientos de un conjunto de datos.

La aplicación de la estadística está compuesta por cuatro fases fundamentales que se interrelacionan entre sí, y que juntas definen a la estadística como tal. La primera es la que se conoce como *Recopilación de los datos*, la segunda *Procesamiento de la información*, la tercera *Análisis de los datos* y la cuarta *Presentación e interpretación de*

los resultados. Esta serie de fases es la guía a seguir cuando se inicia una investigación; el proceso de cada fase se describe enseguida.

1.1.1 Recopilación de los datos

Este es el punto inicial de la metodología estadística en la investigación y en esta fase se define el diseño de la recolección de los datos, en la cual se toman en cuenta las necesidades de información planteadas en el estudio. El diseño debe proporcionar al investigador una forma rápida y eficaz de obtener los datos al menor costo posible y con la garantía de que la información arrojada en el proceso sea válida para el estudio.

Antes de comenzar con la recolección de los datos se deben identificar claramente a las unidades de estudio; es decir, se deberá definir la población objetivo y los casos o la muestra con la que se trabajará durante la investigación. Se recomienda definir cuál es la unidad de estudio y cuál es el colectivo; este primer paso es fundamental ya que permite entender bien qué es lo que se va a medir y sobre quién se va a medir, que es una unidad o entidad concreta y debidamente delimitada.

Una vez definida la población objetivo se determinan las características a medir en cada unidad; esto es lo que se conoce como medición; por ejemplo, si suponemos que se desea hacer un estudio sobre la microempresa en el país, la población objetivo sería todas las microempresas registradas en el país y la unidad de estudio sería una microempresa en particular; en esta unidad se pueden medir varias cosas, el número de empleados, las ventas diarias, los gastos mensuales, etc. Al concepto que se mide se le llama variable y al resultado de la medición para una unidad particular se le llama dato. El dato se llama univariado si es de una sola variable y multivariado si es de más de una variable.

Los datos son clasificados en cualitativos y cuantitativos, esta clasificación del dato es muy importante ya que el análisis estadístico más adecuado para alguna variable depende de esta clasificación.

Los datos cualitativos son etiquetas o nombres asignados a un atributo de cada unidad de estudio. Por ejemplo, en una microempresa el nombre, el giro, el RFC, etc, son datos cualitativos. Como los datos son cualitativos a la variable “nombre de la empresa” se le llama variable cualitativa.

Los datos cuantitativos indican cuánto o cuántos. Por ejemplo, nuevamente en la microempresa, el número de trabajadores, ventas mensuales, gastos de administración, etc., son cuantitativos. Por lo anterior, a la variable "número de trabajadores" se le llama variable cuantitativa. Los datos están asociados a una escala de medición, usualmente se utilizan cuatro: la nominal, la ordinal, la de intervalo y la de razón. Los datos que se generan con escala nominal únicamente permiten contar cuántos individuos hay en cada categoría y se pueden hacer representaciones comparando las frecuencias relativas o absolutas de las categorías. La escala ordinal tienen un elemento adicional de importancia en muchas investigaciones: el orden. Datos que se generan con características como la opinión respecto a algún asunto, se pueden registrar en una escala ordinal. Para este caso se podrían definir las categorías como "favorable", "neutra" y "desfavorable". Estas categorías podrían codificarse con números como 1, 2 y 3. Es claro que aquí entre los números 1 y 2, hay un significado de orden, pero no se sabe qué tanto menos es "neutra" que "favorable".

Las escalas de intervalo y de razón, sirven para registrar datos cuantitativos; la primera tiene una característica importante: el cero no significa ausencia de la característica de interés, sino más bien es un valor que tiene un significado específico. La escala de grados Fahrenheit es un ejemplo de este tipo. La última es la de razón, y en ella la ausencia de la característica de interés se registra con el cero; aquí tienen sentido las proporciones o razones. Con esta escala se registran variables como longitudes, cantidades, pesos, volúmenes, etcétera.

Los datos pueden ser recolectados de diversas formas, ya sea por un proceso de medición directa, a través de una encuesta, con el diseño de un experimento, o bien a través de sistemas de captación de información gubernamental. En los primeros casos hablamos de información captada en fuentes directas y en el segundo de fuentes secundarias. La forma en que se recolectan los datos se definirá en base a la naturaleza, objetivos y restricciones de la investigación.

1.1.2 Procesamiento de los datos

Esta fase se refiere a la organización de los datos, de tal forma que puedan ser analizados y procesados eficientemente; es decir, esta es la fase cuando se crea una base de datos. En el

proceso de la metodología estadística los datos representan la materia prima con la que se trabaja, es por eso que el primer paso en una investigación es recolectarlos; una vez recolectados se procede a elaborar las bases de datos con las que se trabajará durante toda la investigación.

La elaboración de una base de datos no es tarea sencilla, porque ésta debe ser estructurada dependiendo del tipo de análisis que se realizará y en muchas ocasiones en función del paquete estadístico que se utilizará. La forma general en que se organizan es a través de una matriz de doble entrada en la que las unidades de estudio son las filas de la matriz y las variables medidas en ellos son las columnas. Esquemáticamente la estructura de la matriz de datos se muestra en la Figura 1.1.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Figura 1.1. Estructura general de la matriz de datos.

Ojeda y Velasco (2010) recomiendan diseñar la base de datos de acuerdo a las necesidades de la investigación, verificando la calidad de los datos y procurando minimizar los errores de captura, así como seleccionar el paquete estadístico apropiado para el análisis.

Las bases de datos usadas para los artículos presentados en esta memoria fueron obtenidas a través de medios de captación de información gubernamental, el más consultado fue el del INEGI, que cuenta con información financiera, precios, trabajo, etc. Generalmente la captura se hace en una hoja de cálculo para la rápida manipulación de los datos y posteriormente la base se exporta al paquete estadístico que se vaya a usar en la investigación. Esto ayuda porque hoy en día casi cualquier egresado de la educación universitaria maneja estas herramientas. Cada disciplina tiene un paquete estadístico de preferencia; por ejemplo, el SPSS es usado en las investigaciones de tipo social, el Epi-Info es usado en la epidemiología, el Minitab es preferido por los especialistas en control de la calidad, etc., pero se puede hacer uso de varios paquetes a la vez, siempre y cuando se tenga acceso a ellos y los conocimientos necesarios para usarlos. Algunos investigadores

son buenos manipulando paquetes estadísticos, porque su operación no es muy difícil, pero en muchas ocasiones no saben cómo interpretar las salidas, por lo que se ven en la necesidad de acudir con un asesor estadístico.

1.1.3 Análisis de datos

En esta fase se da una serie de procedimientos para manipular los datos a fin de transformarlos en información relevante para la investigación. El análisis debe estar en función de los objetivos de la investigación, debido a que, de un conjunto de datos se puede obtener infinidad de información, pero sólo si ésta se asocia con los objetivos, entonces se le da racionalidad.

El análisis de los datos se divide en dos partes: Análisis inicial de los datos y análisis estadístico formal, que también es llamado análisis definitivo; en el primero se hace una descripción de los datos, es también llamado análisis descriptivo o exploratorio, sirve para observar el comportamiento general de los datos respecto a patrones de tendencia y variabilidad; implica una serie de procedimientos gráficos y numéricos, de conteo y la obtención de tablas de frecuencia y porcentajes para tener la primera información sobre el tema que se está estudiando. También se puede realizar un análisis para cada columna de la matriz de datos, llamado análisis marginal, y algunos estudios de asociación, con lo cual se tiene generalmente la base de las primeras conclusiones sobre el estudio. También hay análisis bivariados o cruzados, esto implica la selección de una serie de preguntas de interés, las cuales posibilitan identificar las variables a cruzar.

Ojeda y Velasco, (2010) recomiendan que para hacer un análisis estadístico hay que entender con claridad la estructura de la matriz de datos y la naturaleza de la información de los datos mismos; además de que se deben llevar a cabo varios análisis marginales e ir construyendo poco a poco juicios sobre la población de referencia u objetivo.

El análisis formal se basa en el planteamiento de modelos y técnicas estadísticas que se determinan con el análisis inicial y con las que se pueda llevar a cabo la inferencia estadística. Algunos de estos análisis son: los estudios multivariados, la modelación estadística, análisis de regresión simple o múltiple, técnicas de modelación estadística, de estadística no paramétrica, entre otros. Chatfield (1995) plantea una serie de reglas para

analizar datos en el contexto de un estudio o investigación en general, las cuales se presentan a continuación.

Seis reglas básicas para analizar datos:

1. No intentar analizar los datos antes de tener un entendimiento claro de qué es lo que se está midiendo y por qué; tratando, además de encontrar si existe información anterior o primaria acerca de los posibles efectos que pueda introducir cada variable en el comportamiento general del problema o fenómeno. En este orden de ideas, el analista de los datos deberá hacerse muchas preguntas con la finalidad de: clarificar los objetivos del estudio o análisis del problema; conocer el significado de cada variable y las unidades en que se están midiendo; conocer el significado de los símbolos especiales que se estén utilizando (si los hay); y si existen experiencias similares que aporten información complementaria sobre el problema o fenómeno en cuestión, que apoyen los análisis, se deberá acceder a la revisión de antecedentes.
2. Conocer cómo fueron recolectados los datos. Aquí se destaca básicamente la importancia de conocer, la forma de obtención de los datos; si hubo un proceso de aleatorización que garantice la confiabilidad de las mediciones. Si los datos provienen de un proceso no aleatorizado propiamente, posiblemente sólo sea justificado realizar un análisis descriptivo simple, lo cual tendrá que ser explícitamente indicado. Hay muchas técnicas estadísticas que se soportan sobre supuestos restrictivos, que de no cumplirse le restan validez a los resultados.
3. Especificar la estructura de los datos, siendo importante aquí contestar las siguientes preguntas: ¿Son suficientes las observaciones para explicar el problema o fenómeno? ¿Son muchas o pocas las variables explicativas? En esta parte es necesario distinguir los diferentes tipos de variables que se vayan a estudiar, definiendo si son variables controlables o variables respuesta, etc. Además debe hacerse una clasificación de variables por tipo de medida o escala, y por la naturaleza: continuas o discretas, cualitativas o binarias. Todo ello porque los análisis resultantes dependen críticamente de la estructura que guarden los datos.

4. Examinar los datos en una forma exploratoria antes de tratar de intentar un análisis más complejo. Para llevar a efecto este análisis es necesario el cálculo de estadísticas básicas y el ajustar gráficas de funciones a los datos en cualquier forma que aparezca apropiada, haciendo esto para cada variable separadamente (y en algunos casos para pares de ellas). Se recomienda el uso de histogramas, diagramas de cajas y alambres, así como diagramas de dispersión, de tallos y hojas, para hacerse una idea de la distribución que pueda suponerse para los datos, así como también para tratar de observar los efectos de los valores faltantes o valores extremos, ya que pueden afectar los posibles análisis.
5. Ser coherente al tener siempre presente la procedencia de los datos y contar con una teoría que sustente la definición de la relación entre las variables implicadas en el fenómeno de estudio, con la finalidad de obtener resultados coherentes que brinden información de acuerdo al contexto del problema.
6. Reportar los resultados de tal forma que éstos reflejen claramente el proceso llevado a cabo con el análisis de los datos, además de sustentarlos con el marco teórico que defina la relación entre las variables analizadas y que conduzca a una correcta interpretación de los mismos.

1.1.4. Presentación e interpretación de los resultados

En esta última fase se proporcionan una serie de indicaciones para la presentación de los resultados de la investigación a través de un informe final mediante tablas y gráficas; es así como se dan los elementos necesarios para construir aseveraciones válidas y confiables en base a los resultados arrojados en el análisis. Las tablas y figuras deben ser etiquetadas, las primeras deben llevar el título a la cabeza y las segundas lo deben tener al pie de la gráfica, y ambas deben ir numeradas consecutivamente por separado; es decir, las tablas su numeración consecutiva y las figuras la suya. En las figuras se incluyen los diagramas, las gráficas y los esquemas. El documento final debe contener una estructura general con al menos los siguientes apartados: Introducción, Metodología, Resultados y Conclusiones.

Cuando se elabora el informe final no hace falta presentar en él todo lo que se hizo en el análisis, más bien se debe seleccionar lo relevante; es decir, aquello que conteste a las preguntas de investigación. Finalmente se interpretan los resultados y se ubican en el

contexto del fenómeno en cuestión mostrando todo lo que se obtuvo con la investigación, así como el nuevo conocimiento que se adquirió a través de ella. En caso de presentarlo como un artículo científico se deberán tomar en cuenta las indicaciones específicas de la revista en la que se piensa publicar y seguir el formato que para tal fin se indica.

1.2 Tipos de estudio

Dentro de la metodología estadística existen tres tipos de estudios: los observacionales, los experimentales y los de muestreo. En este sentido Ojeda *et al.*, (2004, p.50) explican "En los primeros, las unidades de estudio están dadas en la investigación, de tal forma que el investigador sólo las observa en las características de interés. En estos estudios se recurre a expedientes, a fuentes secundarias o a veces también se hace toma directa de datos. En los estudios experimentales, el investigador agrupa las unidades de estudio mediante un mecanismo aleatorio y asigna un tratamiento para cada grupo. Por otro lado, en los estudios de muestreo las unidades de estudio son una muestra (aleatoria o no aleatoria) de un colectivo mayor llamado población de muestreo". Para fines de esta memoria solamente se describen los estudios observacionales y los de muestreo, ya que son los requeridos regularmente en las investigaciones en finanzas públicas.

1.2.1 Estudios observacionales

Los estudios observacionales son aquellos en los que sólo se observan los sujetos; es decir, no existe ninguna manipulación de ellos en el estudio sino que sólo se miden los efectos de las variables de estudio y se analizan. Éstos, a su vez, se dividen en transversales y longitudinales, que se describen a continuación:

Transversales. Son aquellos en los que no existe continuidad en el tiempo; es decir, los datos se colectan en un único momento dado. El objetivo de este tipo de estudio es describir los sujetos bajo estudio en una o más variables observadas; son muy usados en el área de ciencias de la salud, ya que a través de ellos se analiza la incidencia de una enfermedad. Algunos ejemplos de estudios transversales son: el nivel de satisfacción de un cliente en una inversión bursátil, la prevalencia de un fondo de acciones en altos índices de la bolsa de valores, el nivel de marginación de los municipios de un estado, etc. Los transversales, a su vez se clasifican en tres tipos: descriptivos, exploratorios y

correlacionales. Una descripción más detallada de estos estudios se presenta en Ojeda, *et al.*, (2004).

Longitudinales. Son aquellos en los que se analizan los sujetos bajo estudio a través de diferentes periodos de tiempo, con la finalidad de observar si existen cambios en ellos; es decir, se hacen las mediciones de los mismos sujetos en distintos tiempos, estudiándose la evolución que presentan respecto a la variable medida. Como ejemplo de estudios longitudinales podemos mencionar la evolución de indicadores del sistema financiero mexicano en el siglo XX. Al igual que los transversales los estudios longitudinales se dividen en: longitudinales de tendencia y de evolución de grupo.

De tendencia. Este tipo de estudios analizan cambios de una variable a través del tiempo en una población en general; por ejemplo: Se desea medir la percepción de satisfacción de usuarios a los que se les presta un servicio de transporte público en una ciudad, se mide la variable al inicio del estudio tomando una muestra de usuarios del transporte, un mes después se vuelve a hacer el estudio tomando otra muestra de usuarios; dos meses después se vuelve a hacer el estudio con otra muestra de usuarios y así sucesivamente se va midiendo mes con mes hasta tener una evolución de la variable percepción de satisfacción durante un año.

De evolución de grupo. En estos estudios la evolución en el tiempo se mide a un grupo de sujetos y no a un grupo variable como en el caso anterior. Por ejemplo, se desea medir el nivel de evolución en el indicador de desarrollo humano de un grupo de municipios, –por ejemplo, municipios gobernados por cierto partido durante varios periodos de gobierno–.

1.2.2 Estudios de muestreo

El muestreo en las investigaciones se usa cuando se requiere un estudio rápido y económico o quizá no se tienen los recursos necesarios para estudiar a toda la población objetivo, por lo que a través de ciertos procedimientos bien diseñados se selecciona solamente una parte de los sujetos o unidades de la población –que en este caso se llama población de muestreo– y se miden las variables de interés en ellos. Estos métodos si se realizan correctamente garantizan validez en los resultados y es posible hacer inferencias hacia la población.

Los pasos recomendados para llevar a cabo un estudio de muestreo son: definir la población objetivo, identificar un marco de muestreo (que consiste en listados de la población), seleccionar un diseño muestral acorde a las características del estudio, determinar un tamaño de muestra que garantice la validez externa, selección de la muestra siguiendo las medidas de aleatoriedad, levantamiento de los datos (previa capacitación de los encuestadores –si esto procede– para así garantizar validez de resultados), análisis de la información recabada –que implica el procesamiento y análisis estadístico de los datos– para finalmente tomar decisiones sustentados en los resultados.

En la metodología estadística existen dos tipos de estudios de muestreo: el probabilístico en el que cada elemento de la población tiene una probabilidad conocida de ser seleccionado en la muestra, y el no probabilístico, donde se usa cualquier proceso de selección de una muestra (que obviamente no satisface la característica de un muestreo probabilístico). Aquí solamente se describen los esquemas de tipo probabilístico.

Muestreo aleatorio simple. Este tipo de muestreo es el más común entre todos los métodos de muestreo y el más utilizado, pero solamente es recomendable cuando la población es homogénea con respecto a ciertas variables definidas en el estudio. Una muestra aleatoria simple de tamaño n, es una muestra seleccionada de tal manera que cada muestra posible de tamaño n tenga la misma probabilidad de ser seleccionada bajo un método aleatorio. Esto es como si fuera una rifa donde cada elemento tiene un sólo boleto. El procedimiento para seleccionarla se describe mediante el siguiente ejemplo: Supongamos que se enumeran a 2000 usuarios de una institución bancaria en una ciudad, asignándoles un número progresivo, es decir, 1, 2, 3, … , 2000, en el orden en que aparecen en el archivo de usuarios del banco. Se seleccionan números aleatorios mediante un generador o una tabla para números aleatorios; y los resultados de los números aleatorios fueron 8, 20, 789, 12, 1213, etc. Entonces el primer seleccionado en la muestra es el usuario con el número 8 de la base de datos del banco, el segundo es el usuario con el número 20 y así sucesivamente hasta completar digamos 55 usuarios, que es el tamaño de muestra calculado. Esquemáticamente se puede representar como se observa en la Figura 1.2.

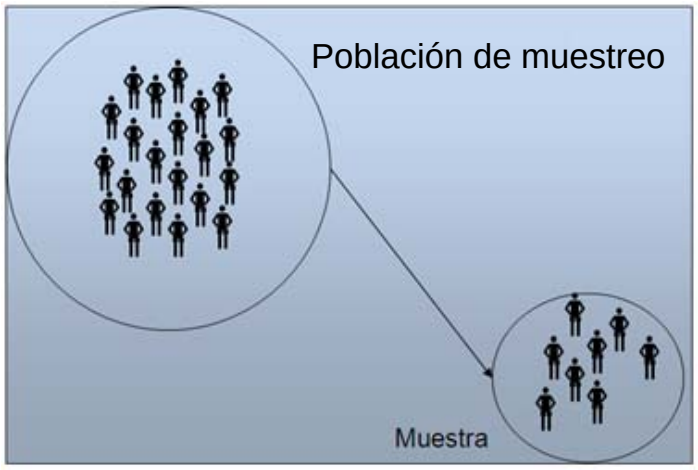


Figura 1.2. Esquema de un muestreo aleatorio simple.

Muestreo sistemático. Una muestra sistemática es obtenida cuando los elementos son seleccionados en una manera sistemática (el mismo número de orden) en grupos que aparecen en una secuencia. La forma de la selección depende del número de elementos incluidos en la población y del tamaño de la muestra. El número de elementos en la población es, primero, dividido por el número deseado en la muestra. El cociente indicará si cada décimo, cada onceavo, o cada centésimo elemento en la población es seleccionado en la muestra. Los N elementos de la población están numerados del 1 al N en cierto orden; y únicamente el primer elemento de la muestra (que es el primer elemento del grupo) es seleccionado al azar; por lo tanto, una muestra sistemática puede dar la misma precisión de estimación que una muestra aleatoria simple cuando los elementos en la población están ordenados al azar.

Para extraer una muestra de tamaño n dividimos a la población en n grupos de tamaño k, donde k = N/n, elegimos aleatoriamente un número entre 1 y k, digamos j y de esta manera la muestra sistemática queda conformada por el elemento j, j+k, j+2k,…, j+(n-1)k. Si durante el muestreo un sujeto seleccionado no quiere participar en el estudio se pierde la aleatoriedad, por lo que hay que volver a elegir aleatoriamente un número entre 1 y k y seguir con el procedimiento ya mencionado. El esquema de este tipo de muestreo se ve ilustrado en la Figura 1.3 para grupos de tamaño 4.

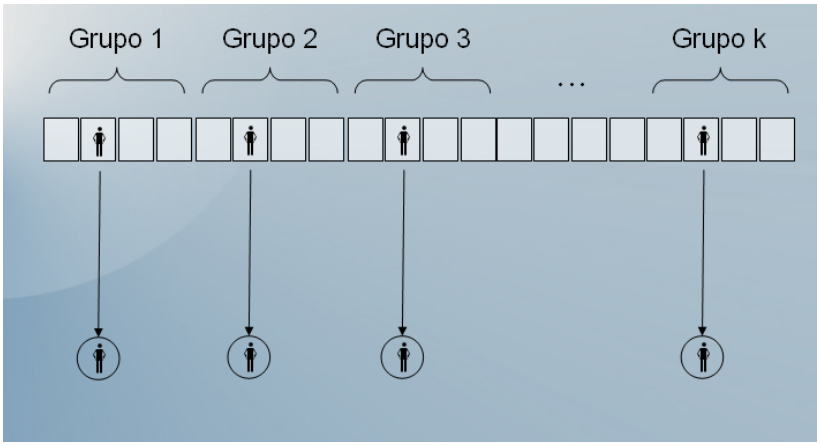


Figura 1.3. Esquema de un muestro sistemático para grupos de tamaño 4.

Muestreo estratificado. Cuando se tiene una población no homogénea es conveniente usar un muestreo de tipo estratificado; éste consiste en dividir a la población en varios grupos, llamados estratos, que garantizan una población dividida en grupos homogéneos respecto a ciertas características. El procedimiento consiste en seleccionar aleatoriamente en cada estrato una muestra que puede ser proporcional al tamaño del estrato en relación con la población.

Los casos en los que conviene usar el muestreo estratificado son: para protegerse de obtener una muestra no representativa; cuando para el estudio es conveniente estudiar sub-poblaciones con precisión; si una muestra estratificada puede ser menos costosa que una muestra aleatoria simple; cuando una muestra estratificada da estimaciones más precisas que una muestra aleatoria simple. En la Figura 1.4 se representa es esquema de un muestreo estratificado.

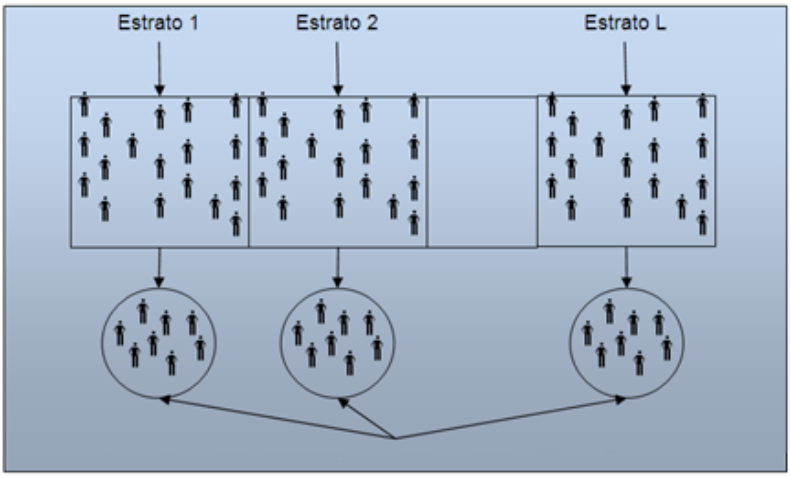


Figura 1.4. Esquema de un muestro estratificado.

Muestreo por conglomerados. En este tipo de muestreo se divide a la población en grupos que se encuentran agrupados naturalmente y que son llamados conglomerados. Una vez identificados se selecciona una porción de los grupos de manera aleatoria. Finalmente, se censan los grupos seleccionados; es decir, se toman todos los elementos. Bajo este método, aunque no todos los grupos son muestreados, cada grupo tiene una igual probabilidad de ser seleccionado, por lo tanto, la muestra es aleatoria. En la Figura 1.5 se presenta el esquema de un muestreo por conglomerados; cabe hacer notar que de cada conglomerado seleccionado se puede realizar una muestra aleatoria simple, con lo que se tendría un muestreo bietápico por conglomerados.

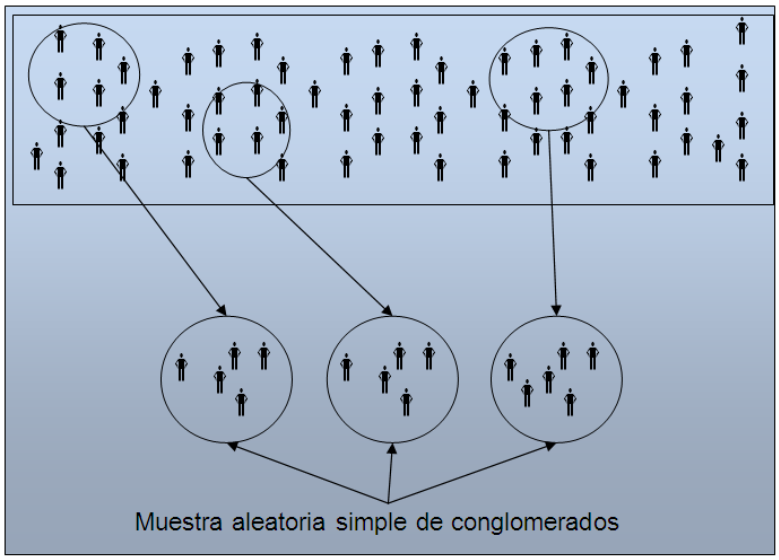


Figura 1.5. Esquema de un muestro por conglomerados.

Es menester señalar que los tipos de muestreo se pueden combinar; por ejemplo, haciendo un esquema de selección sistemática de viviendas en manzanas seleccionadas en un muestreo estratificado por conglomerados.

Determinación del tamaño de muestra. En la metodología del muestreo hay varios aspectos a considerar para el cálculo del tamaño de una muestra; a continuación se presentan algunos factores generales que se deben considerar en cualquier tipo de muestreo: Identificar la(s) variable(s) a medir; la variabilidad en la población considerando las principales variables de interés; los objetivos inferenciales del estudio; y los recursos disponibles para realizar el muestreo.

La siguiente fórmula se usa para calcular una muestra aleatoria simple cuando se va a estimar una media o promedio con un nivel de significancia α; este valor usualmente se fija en α = (0.1, 0.05, 0.01), que implica significancia baja, media o alta.

$$n = \frac{Z_\alpha^2 \sigma^2}{\varepsilon^2},$$

donde Z_α es un valor de tablas y dá el nivel de significancia α establecido para la inferencia; los valores para α = (0.1, 0.05, 0.01), son Z_α= (1.645, 1.96, 2.576); σ^2 es el valor de la varianza de la variable de interés (que se estima en estudios previos o con la muestra piloto); y ε es la precisión que se desea para la estimación (en función de la escala y los valores de la variable de interés).

La siguiente fórmula se usa para calcular una muestra aleatoria simple cuando se va a estimar una proporción:

$$n = \frac{Z_\alpha^2 pq}{\varepsilon^2},$$

donde: Z_α es un valor de tablas y da el nivel de significancia para la inferencia; p es la proporción que se desea estimar (este valor se asume p = 0.5 si se ignora totalmente; se puede usar un valor obtenido en estudios previos); q = 1 – p; y ε es la precisión que se desea para la estimación (se asume como un valor de (0.1, 0.05, 0.01), baja, media y alta precisión).

1.3 Análisis exploratorio en SPSS

SPSS (Statistical Package for the Social Sciencies, 2010) es un paquete estadístico que funciona en ambiente Windows XP y Vista; además, a partir de la versión 16 en 2007, se desarrolló SPSS para Macintosh y una versión para Linux. SPSS combina facilidades de manejo de base de datos, elaboración de análisis estadístico y realización de gráficos de alta resolución. En esta sección se presenta una breve introducción al paquete y una serie de instrucciones para la creación y el desarrollo de gráficos y la utilización de algunas herramientas estadísticas con las que se puede realizar un análisis exploratorio de datos.

Para iniciar la ejecución del programa, se elige primero SPSS desde Inicio/Programas/SPSS para Windows/SPSS 15.0 para Windows y nos aparece la siguiente ventana que nos muestra el panel del menú principal, como se muestra en la Figura 1.6.

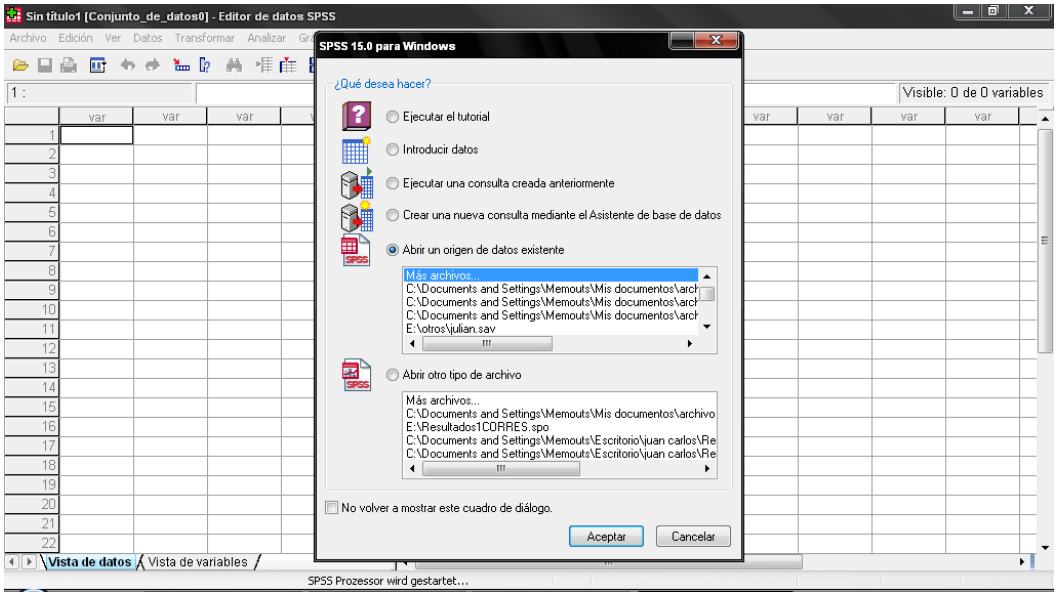


Figura 1.6. Panel inicial del paquete SPSS.

Anteriormente se mencionó que las bases de datos usadas en las investigaciones presentadas en esta memoria fueron obtenidas de fuentes de captación gubernamental, por lo que todas las bases de datos están capturadas en Excel. Para importar un archivo de Excel, se le da click en Archivo/abrir/datos y en la pestaña que dice tipo de archivo se selecciona la opción Excel(*.xls); ver Figura 1.7.

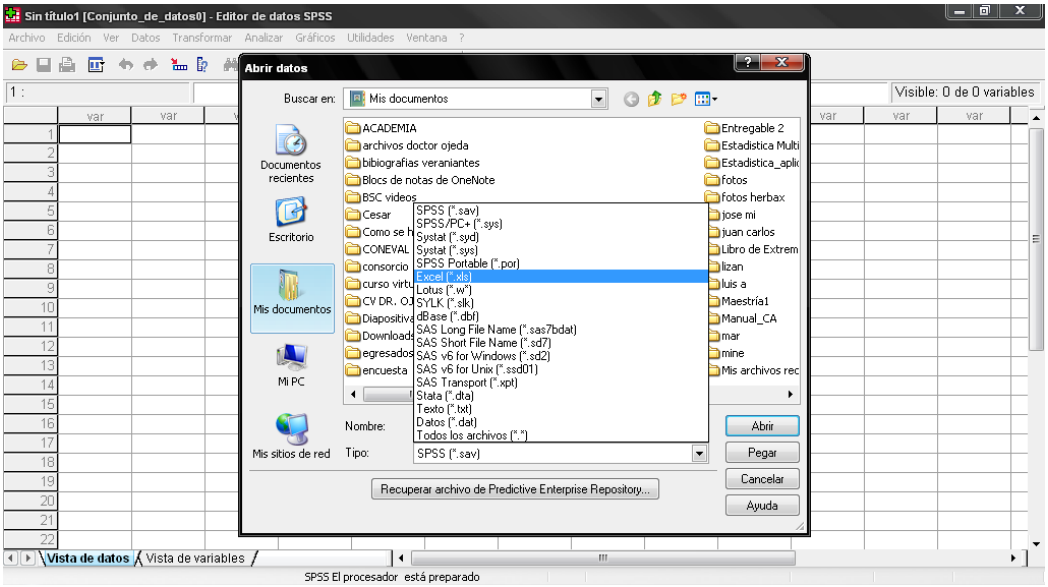


Figura 1.7. Exportar archivos con extensión *.xls de Excel.

Se elige el archivo de Excel en donde se tienen los datos y aparece la pantalla que se muestra en la Figura 1.8.

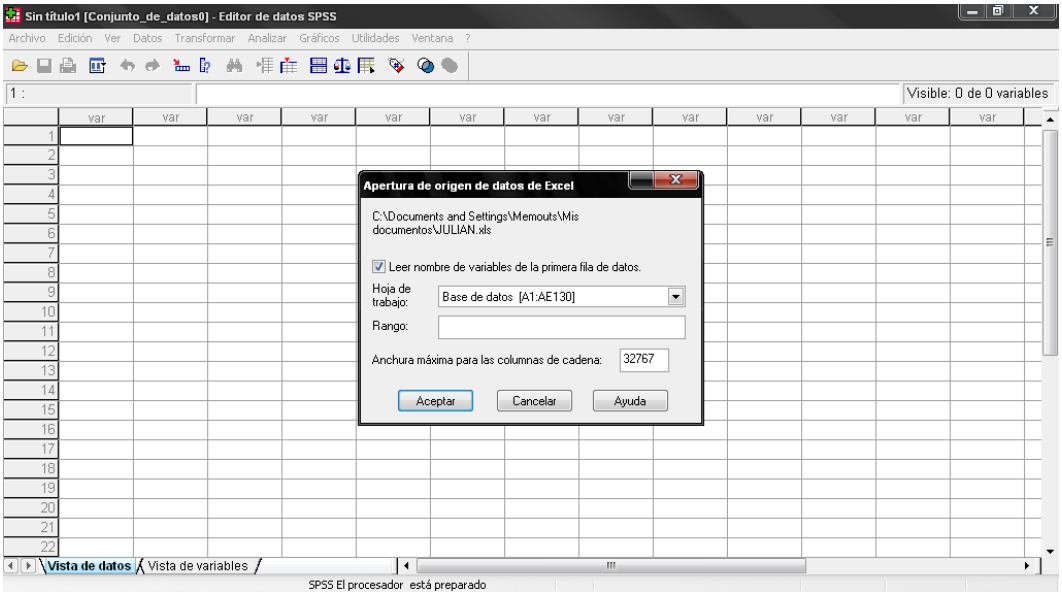
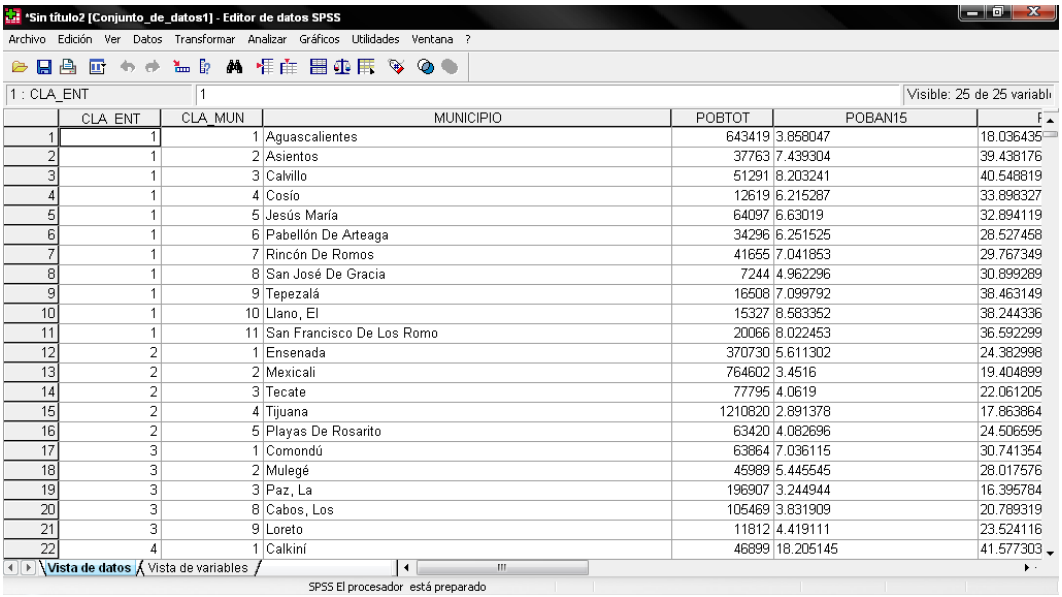


Figura 1.8. Esquema que muestra la apertura de datos de archivos Excel.

Se le da aceptar y aparece la base de datos con la que se trabajará (Ver Figura 1.9).



	CLA_ENT	CLA_MUN	MUNICIPIO	POBTOT	POBAN15	
1	1	1	Aguascalientes	643419	3.858047	18.036435
2	1	2	Asientos	37763	7.439304	39.438176
3	1	3	Calvillo	51291	8.203241	40.548819
4	1	4	Cosío	12619	6.215287	33.898327
5	1	5	Jesús María	64097	6.63019	32.894119
6	1	6	Pabellón De Arteaga	34296	6.251525	28.527458
7	1	7	Rincón De Romos	41655	7.041853	29.767349
8	1	8	San José De Gracia	7244	4.962296	30.899289
9	1	9	Tepezalá	16508	7.099792	38.463149
10	1	10	Llano, El	15327	8.583352	38.244336
11	1	11	San Francisco De Los Romo	20066	8.022453	36.592299
12	2	1	Ensenada	370730	5.611302	24.382998
13	2	2	Mexicali	764602	3.4516	19.404899
14	2	3	Tecate	77795	4.0619	22.061205
15	2	4	Tijuana	1210820	2.891378	17.863864
16	2	5	Playas De Rosarito	63420	4.082696	24.506595
17	3	1	Comondú	63864	7.036115	30.741354
18	3	2	Mulegé	45989	5.445545	28.017576
19	3	3	Paz, La	196907	3.244944	16.395784
20	3	8	Cabos, Los	105469	3.831909	20.789319
21	3	9	Loreto	11812	4.419111	23.524116
22	4	1	Calkiní	46899	18.205145	41.577303

Figura 1.9. Base de datos en SPSS importada de un archivo Excel.

1.3.1 Técnicas para explorar datos

Para ejemplificar cada una de las herramientas usadas en el análisis exploratorio se usará una base de datos que muestra el índice de marginación obtenido en 2010 y que considera los 2,443 municipios de toda la república mexicana.

Diagrama de barras. La gráfica de barras es la representación más útil para datos nominales u ordinales. Ésta consiste en barras verticales u horizontales que representan la frecuencia de las observaciones en categorías específicas. Para obtener la gráfica damos un click en *Graficos/interactivos*, seleccionamos *barras* y arrastramos la variable que deseamos graficar al segundo campo como se ve en la Figura 1.10; la variable que se arrastró fue GRADO que representa el grado de marginación en todos los municipios de la República Mexicana en 2010, y se le da click en aceptar.

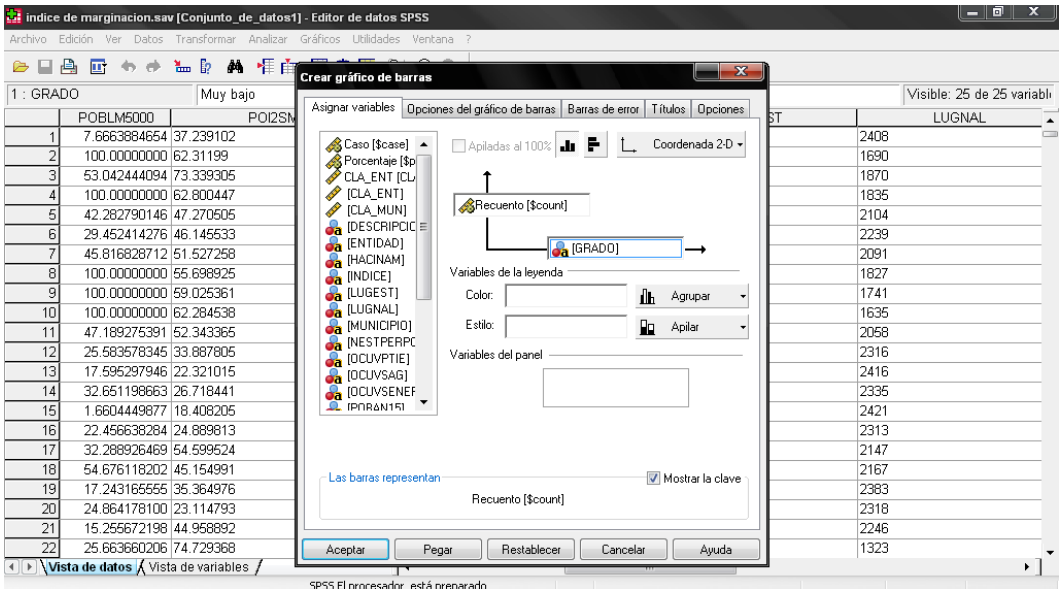


Figura 1.10. Creación de un gráfico de barras en SPSS.

El resultado de la gráfica es el que se muestra en la Figura 1.11.

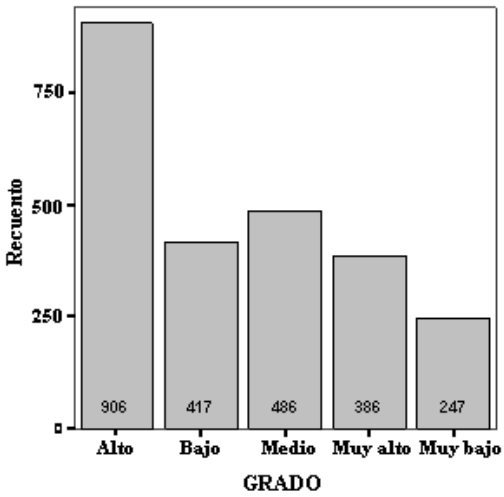


Figura 1.11. Gráfico de barras que muestra el grado de marginación por municipio en la República Mexicana en 2010.

Diagrama circular o de sectores. El gráfico circular consiste en representar proporcionalmente en un círculo la frecuencia o porcentaje de cada una de las categorías; se recomienda para variables con no más de 5 categorías. Para obtener la gráfica damos un click en *Grafico*, seleccionamos *Interactivo*/*sectores*/*simple* y posteriormente arrastramos la variable a graficar tal como se muestra en la Figura 1.12. Nuevamente tomamos la variable GRADO que representa el grado de marginación en los municipios de la República Mexicana.

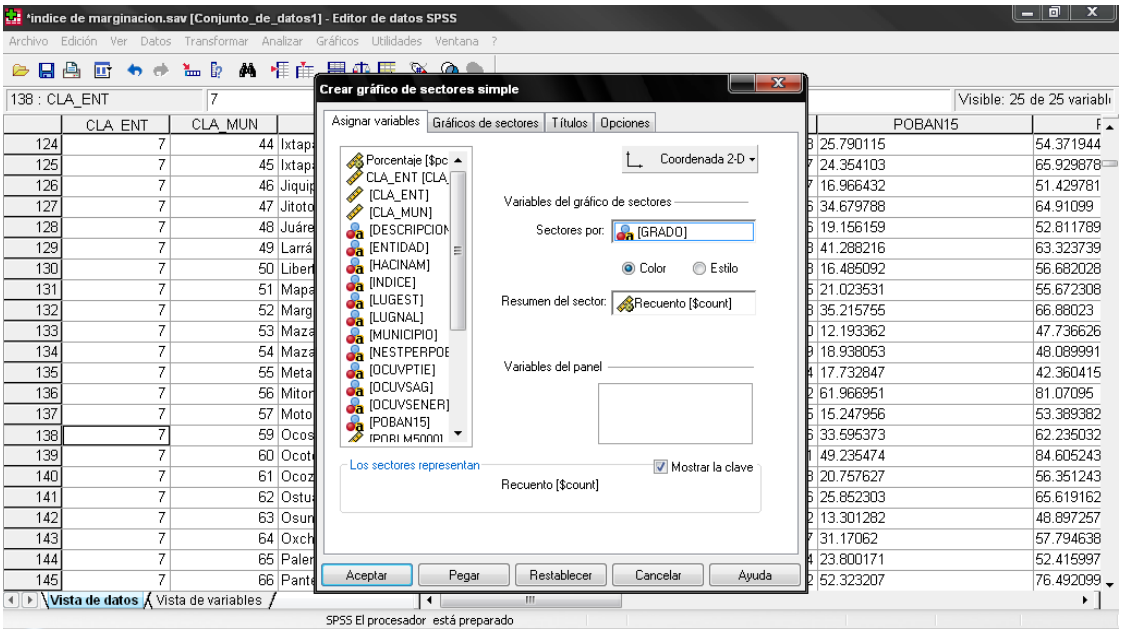


Figura 1.12. Creación de un gráfico de sectores.

La grafica resultante es la que se muestra en la Figura 1.13.

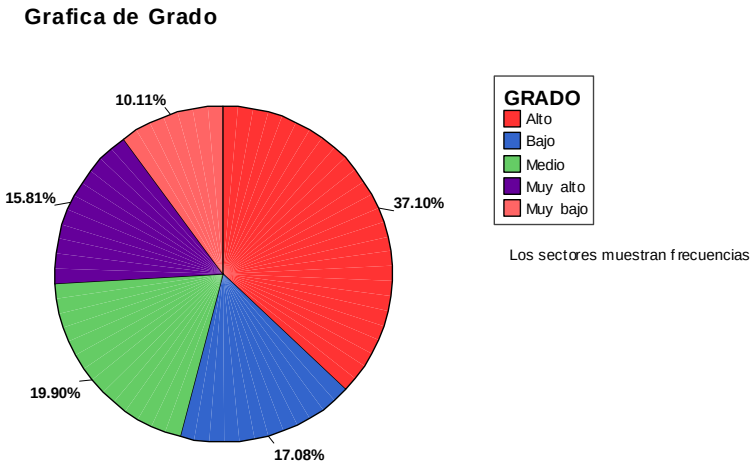


Figura 1.13. Gráfico de sectores que muestra el grado de marginación de los 2,443 municipios de la República Mexicana en 2010.

Diagrama de cajas y alambres. Este gráfico es un ingenioso despliegue de los estadísticos de orden más importantes en un grupo de datos en una escala de intervalo o de razón. Se grafican además de los cuartiles primero y tercero, la mediana y el valor mínimo así como también el máximo. Para realizar este diagrama seleccionamos la opción *Gráficos* desde el menú principal, *cuadros de dialogo antiguos*/*diagramas de caja* y aparece la pantalla que se muestra en la Figura 1.14.

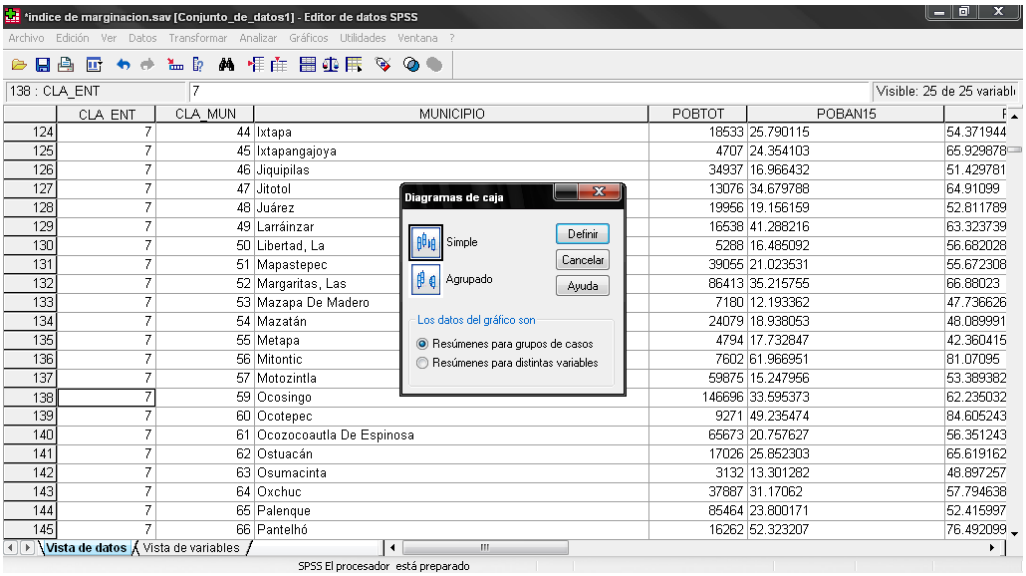


Figura 1.14. Selección del tipo de diagrama de caja.

Se elige la opción de *Resúmenes para distintas variables* y se le da click en *Definir*, en la Figura 1.15 se muestra la ventana que aparece, se selecciona la variable a graficar que en este caso es POBLACIÓN TOTAL y se da aceptar.

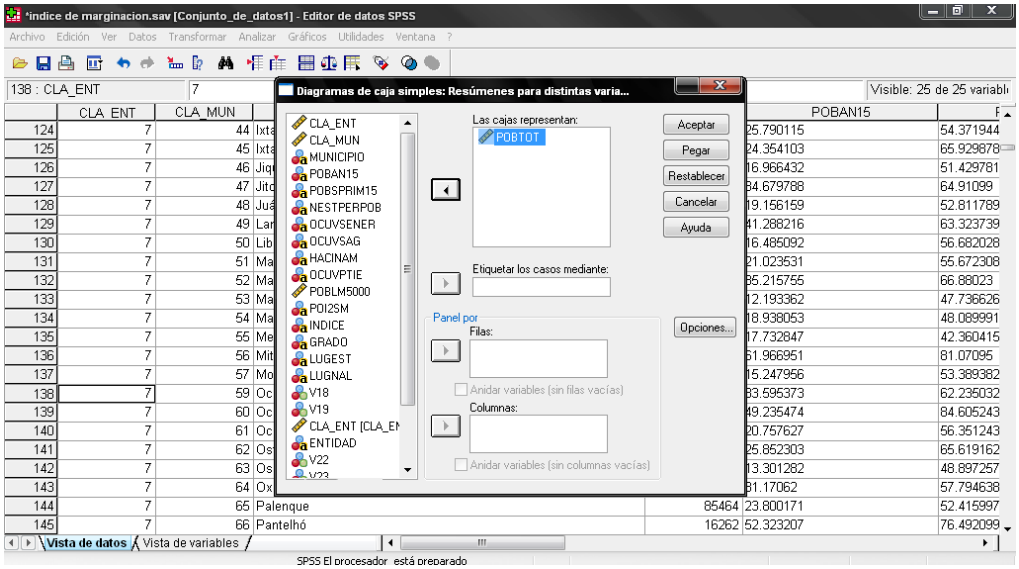


Figura 1.15. Ventajas que muestran el procedimiento de creación de un diagrama de caja.

La grafica de cajas se puede observar en la Figura 1.16a donde pueden verse los casos de municipios considerados como atípicos; asimismo se observa que en este tipo de gráfico no se puede ver la forma de la distribución, pero si su tendencia y la dispersión, por lo que se convierte en una magnifica opción para realizar análisis comparativos. En la Figura 1.16b se muestra el análisis comparativo del PIB por estado.

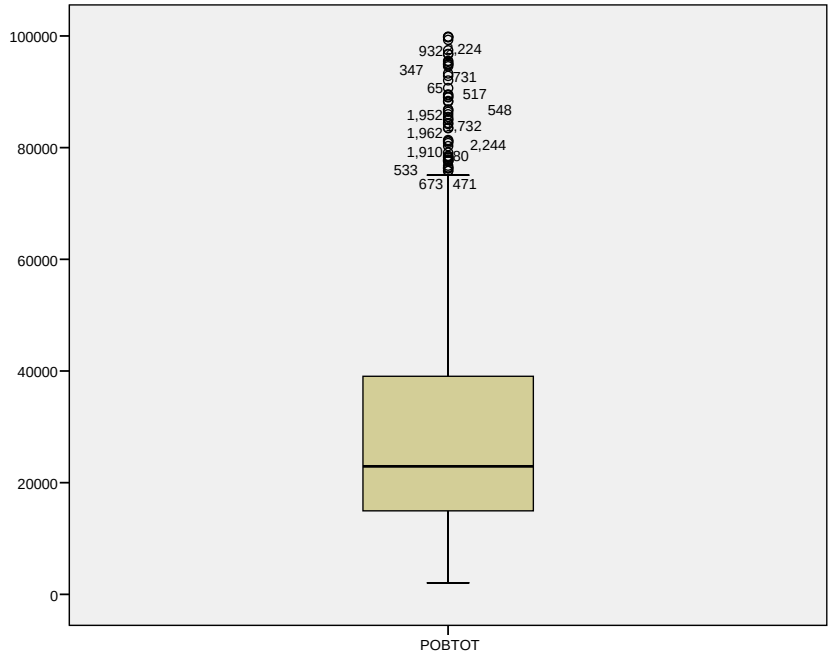


Figura 1.16a. Diagrama de caja de la población total de los 2443 municipios de la República Mexicana en 2010.

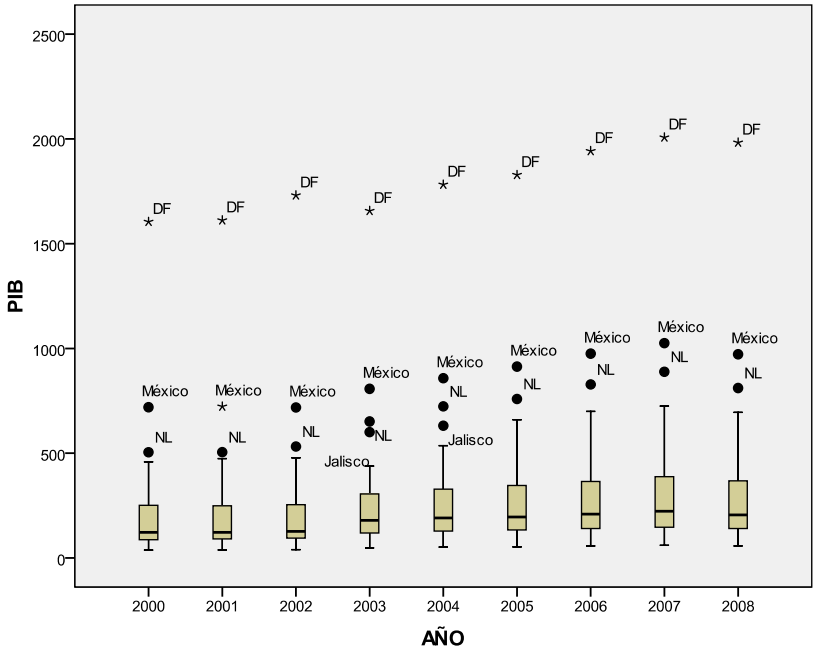


Figura 1.16b. Diagrama de cajas comparativo del PIB en los 32 estados de la República Mexicana del año 2000 al 2008.

Histograma. El histograma es una gráfica de barras sin espaciamiento entre ellas; esto se debe a que los datos deben pertenecer a variables continuas. Se recomienda su empleo para problemas con grandes cantidades de datos ($n > 50$) y que presenten una variación que permita realizar la agrupación de los datos. Por intervalos de clase (esto lo hace automáticamente el paquete). Para realizar este gráfico seleccionamos la opción *Gráficos* desde el menú principal, *Interactivos*/*histograma* y aparece la siguiente pantalla (ver Figura 1.17) en la cual arrastramos la variable a graficar y se le da aceptar.

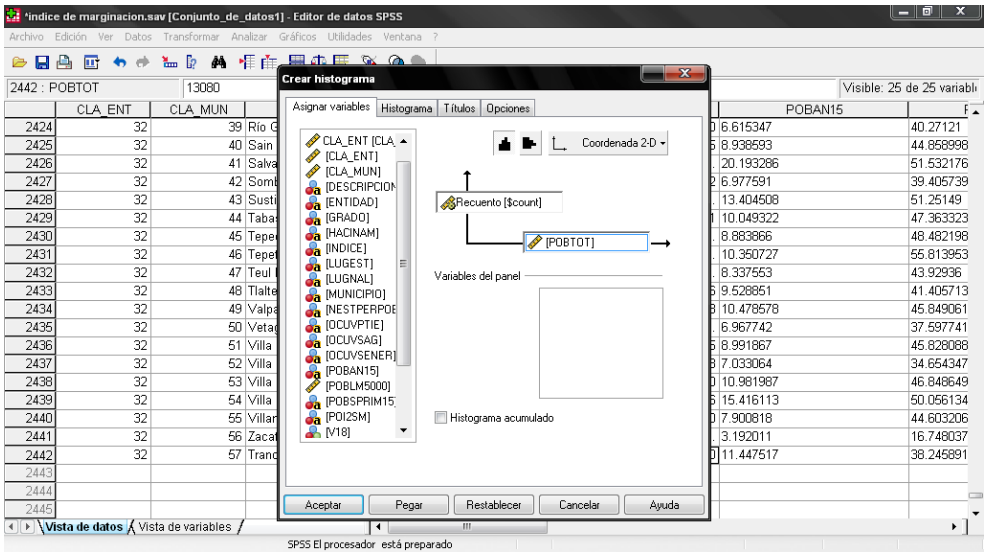


Figura 1.17. Creación de un histograma.

El histograma queda representado como se muestra en la Figura 1.18.

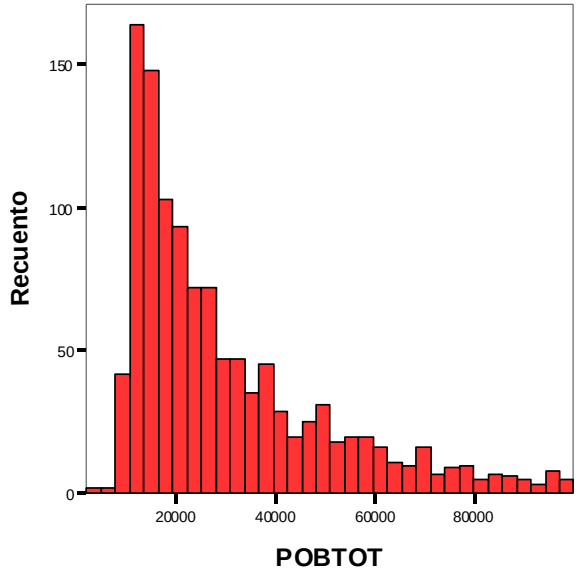


Figura 1.18. Histograma de la población total de los 2443 municipios de la República Mexicana en 2010.

Diagrama de dispersión. El Diagrama de dispersión es un gráfico que permite una visión rápida de la forma e intensidad de la asociación entre un par de variables X y Y; se pide que ambas variables sean continuas, aunque X puede ser discreta o alguna que indique la pertenencia a un grupo. Para realizar este tipo de gráfico seleccionamos la opción *Gráficos* desde el menú principal, *Cuadros de dialogo antiguos/dispersión/puntos* y aparece la pantalla que se muestra en la Figura 1.1.9; se da click en *Dispersión simple.*

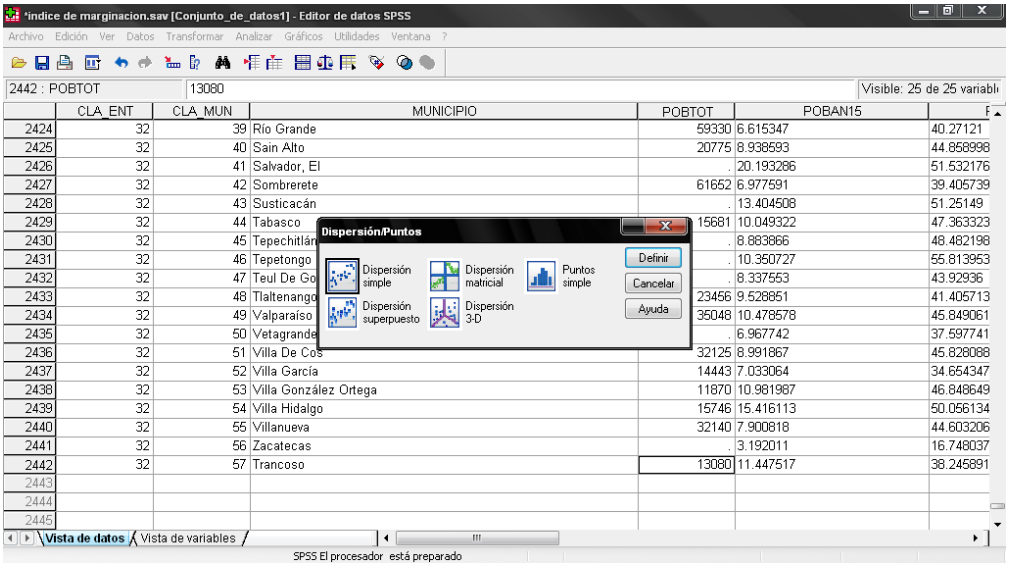


Figura 1.19. Selección del tipo de diagrama de dispersión.

Se selecciona la variable X y Y tal como se muestra en la pantalla que se muestra en la Figura 1.20.

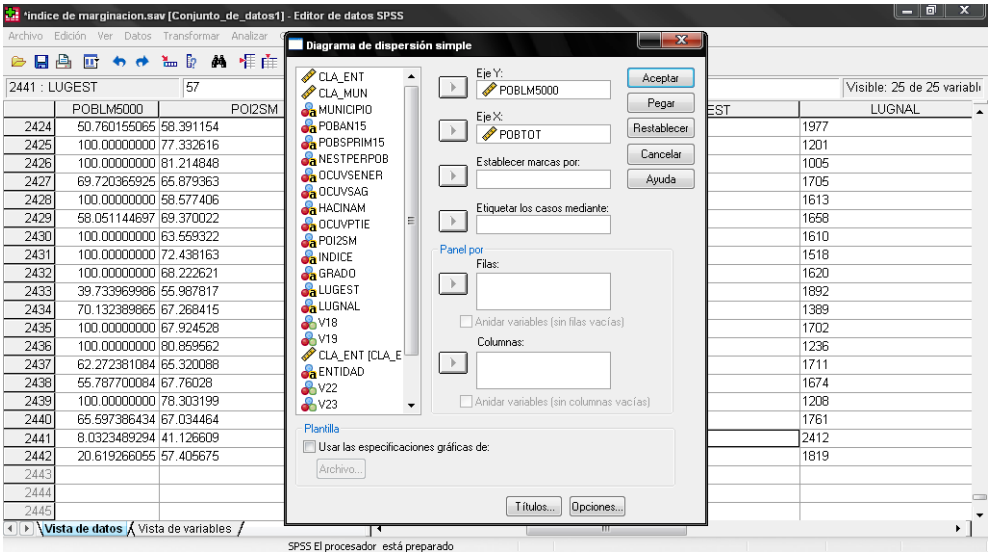


Figura 1.20. Creación de un diagrama de dispersión.

La gráfica que se obtiene se muestra en la Figura 1.21.

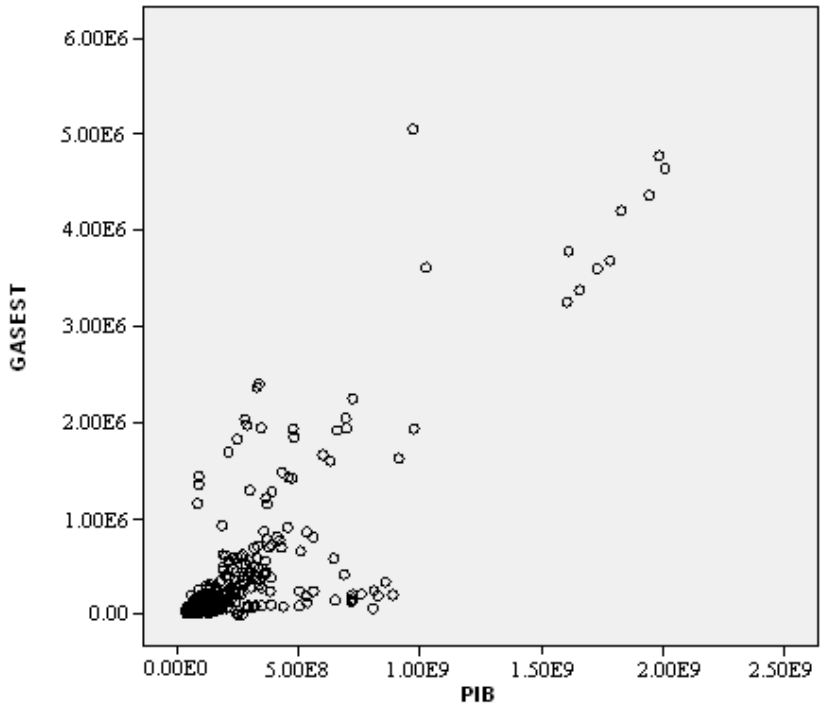


Figura 1.21. Diagrama de dispersión del PIB contra gasto total por estado en 2010.

Existen más gráficas que se pueden obtener para el análisis exploratorio, pero las gráficas presentadas aquí son las más comunes. Para mayores detalles y otros despliegues gráficos y análisis numéricos sencillos ver Ojeda y Behar (2006).

1.3.2 Ventajas del paquete SPSS

El SPSS es el software más usado en las ciencias sociales, es muy popular ya que tiene una capacidad muy buena para trabajar con bases de datos de gran tamaño. También permite la recodificación de las variables y registros según las necesidades del usuario. Por sus capacidades puede competir con paquetes licenciados como el SAS, Statistica, Stata, así como también con software libre como el R. El sistema de módulos de SPSS provee toda una serie de capacidades adicionales a las existentes en el sistema base. También cuenta con un sistema de archivos, cuyo principal propósito es que el manejo los archivos sea amigable y permita múltiples operaciones acorde a las necesidades del usuario. Para más información se recomienda (Lara, 2011).

Los archivos de datos se guardan con extensión .SAV y tiene un sistema de archivos de salida con extensión .SPO; las salidas pueden ser exportadas en formato

HTML, RTF o TXT; las nuevas versiones incorporan exportación a PDF, XLS y DOC. Cuenta con ficheros scripts, que son usados por usuarios avanzados para generar rutinas que permiten automatizar procesos muy largos y complejos. Estos procesos suelen ser parte de las salidas estándar de los comandos del SPSS, aunque parten de estas salidas. La funcionalidad de los scripts ha sido ahora asumida por la inserción del lenguaje de programación Python en las rutinas de sintaxis del SPSS. Cuando se instala SPSS trae un determinado número de ejemplos y bases de datos que son usados para ilustrar algunos de los ejemplos de uso del programa.

Su uso es sencillo y además las salidas son muy claras y regularmente tienen alguna explicación que permite entender mejor el resultado, cosa que otros paquetes no tienen. Y el plus de este programa, además de las ventajas ya mencionadas, es que tiene un modulo para obtener muestras complejas donde se calculan tamaños de muestra, y además si existe la base de datos o el marco de muestreo en un archivo del sistema la muestra es seleccionada aleatoriamente.

II. Análisis Multivariado

2.1 Aspectos generales

El análisis multivariado o multivariante es la rama de la estadística que permite analizar simultáneamente conjuntos amplios de variables medidas sobre cada unidad de estudio. El investigador en la mayoría de las ocasiones tiene la necesidad de estudiar medidas múltiples para poder dar solución a su investigación. Por ejemplo, para poder describir el comportamiento de la situación en salud de un país, se tienen que estudiar variables como los tipos de enfermedades, el grado de conocimiento en educación sexual de los ciudadanos, la esperanza de vida, el índice de mortalidad, entre otros. En finanzas públicas, el gobierno podría estar interesado en aplicar tres diferentes programas de desarrollo estatal a las entidades federativas para lo cual desea conocer la manera en que éstas podrían formar tres grupos de acuerdo a un conjunto de variables de interés. En el caso de la economía, la identificación de las dimensiones que intervienen en el desarrollo económico, construcción de índices, entre otras. Además de estas áreas de aplicación, el análisis multivariado tiene cabida en diversas disciplinas, como biología, ciencias sociales, ingeniería, agricultura, economía, medicina, entre otras.

Algunas de las razones por las cuales hay que aplicar técnicas multivariadas son: (1) Los fenómenos bajo estudio son de naturaleza multivariada. En la mayoría de las investigaciones, existe la necesidad del análisis no solo de una variable de estudio, sino de múltiples variables; (2) Existe correlación entre las variables. Cuando se tienen varias variables en la mayoría de las ocasiones existe correlación entre éstas, lo cual conlleva a que la información del fenómeno de estudio dada por las variables sea más difícil de obtener. Existen técnicas multivariadas las cuales ayudan a solucionar este aspecto; (3) Conclusiones más adecuadas. Las técnicas multivariadas ayudan a dar conclusiones del grupo de variables al analizarlas como un conjunto de variables, y no a dar conclusiones erróneas al tratar cada variable individualmente; y (4) Computadoras y disponibilidad de las técnicas en paquetes estadísticos. El desarrollo de los ordenadores con capacidad de almacenamiento y potencia de procesamiento suficiente, acompañados de mayor cantidad de software estadístico cada vez más fácil de usar.

Entre los objetivos del análisis multivariado se encuentran: (1) la reducción de la dimensionalidad del problema; es decir, resumir los datos mediante un pequeño conjunto de nuevas variables construidas como combinaciones de las variables originales, tratando de perder la mínima información sobre el fenómeno de estudio, presente en las variables originales. Así, al tener un número menor de variables, sin perder información, se realiza una mejor descripción del fenómeno; (2) La identificación de conglomerados. Encontrar grupos existentes en los datos, tales grupos serán formados por unidades que sean semejantes; por ejemplo, países con variables de salud semejantes, alumnos con aprovechamiento escolar semejante, etc; (3) La clasificación de unidades de estudio. Si ya se cuenta con grupos de unidades semejantes, se tiene una o más unidades, en ocasiones existe la necesidad de ubicar estas unidades en los grupos ya definidos con anterioridad, es decir es necesaria la clasificación de nuevas variables en grupos definidos con anterioridad; y (4) La relación entre conjuntos de variables. En ocasiones el investigador necesita conocer si existe relación entre dos o más conjuntos de variables y, de existir relación, se desea cuantificar tal relación, caracterizarla y por supuesto, interpretarla.

2.1.1. Matriz de datos

Supóngase que se han observado 5 mediciones de variables relacionadas con las finanzas públicas en las 32 entidades federativas del país durante el año 2011; digamos el PIB, el gasto en salud, el índice de desarrollo humano, los ingresos de la Comisión Federal de Electricidad, y el gasto en seguridad. El conjunto de las 5 variables forman una matriz multivariada (en el caso de 2 se denomina bivariada). En general se tienen p mediciones (variables) medidas en n sujetos (personas, entidades federativas, países, empresas), a partir de estas p variables medidas en cada una de las n unidades de estudio se tiene una matriz de observaciones, denominada matriz de datos. Cada entrada de la matriz denota la medición de una de las p variables a una de las n unidades de estudio; esta matriz es de orden $n \times p$, se denota por $\mathbf{X}$ y está dada por el arreglo genérico que se muestra en la Figura 2.1.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots \\ x_{(n\text{-}1)1} & x_{(n\text{-}1)2} & \cdots & x_{(n\text{-}1)p} \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Figura 2.1 Matriz de datos.

En este caso se tienen *n* filas y *p* columnas, en donde las *n* filas denotan a las *n* unidades de estudio, y las *p* columnas denotan las *p* variables que se midieron a cada una de las unidades de estudio. La i-ésima fila $\mathrm{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}$, está formada por las mediciones de las *p* variables, para la i-ésima unidad de estudio. Mientras que si se toma la j-ésima columna $\mathrm{x}_j^t = \begin{bmatrix} x_{1j} & x_{2j} & \cdots & x_{nj} \end{bmatrix}$, ésta está formada por las *n* mediciones correspondientes a las *n* unidades de estudio para la *j*-ésima variable.

Ejemplo (Gasto en salud). Se tiene información sobre el gasto en salud que se destinó en el 2008 en cada una de las 32 entidades federativas por el Ramo33; el gasto en salud de cada estado, las personas aseguradas y las personas no aseguradas; estas cantidades están expresadas en miles de millones de pesos. La matriz de datos se presenta continuación:

	Entidad	Ramo33	GasEst	GasAse	GasNoAse
1	Aguascalientes	695.53	243.18	1438.55	1497.57
2	Baja California	735.35	420.15	1495.35	1582.52
3	Baja California Sur	773.92	585.87	1819.46	1674.31
4	Campeche	753.63	807.20	2027.86	1735.88
5	Chiapas	1626.63	66.97	2479.97	1668.64
6	Chihuahua	1615.30	33.99	2839.40	1789.96
7	Coahuila	1683.73	270.31	3238.06	1860.60
8	Colima	1884.14	350.81	3708.12	1970.90

Esta matriz está formada por 32 renglones, uno por cada entidad federativa, y por 5 columnas, la primera corresponde al nombre de la entidad federativa, la segunda al gasto en salud por el Ramo33, la tercera al gasto en salud estatal, la cuarta al gasto en salud correspondiente a las personas aseguradas y la quinta columna al gasto en salud correspondiente a las personas no aseguradas.

El primer renglón corresponde a la entidad federativa de Aguascalientes y así sucesivamente con el resto de las entidades.

	Entidad	Ramo33	GasEst	GasAse	GasNoAse
1	Aguascalientes	695.53	243.18	1438.55	1497.57
2	Baja California	735.35	420.15	1495.35	1582.52

Como se aprecia en la tabla anterior, en este estado el gasto en salud en el año 2008 correspondiente al Ramo33 fue de 695.53 miles de millones de pesos, mientras que el gasto estatal en salud fue de 243.18 miles de millones de pesos. El segundo renglón corresponde a la entidad federativa de Baja California, en donde en el 2008 el gasto en salud de las personas no aseguradas fue de 1582.82 miles de millones de pesos.

Ejemplo (Municipios indígenas). Se recabó información de 50 municipios indígenas de la entidad federativa Veracruz, relacionada al Índice de Desarrollo Humano (IDH), al Fondo de Apoyo a la Infraestructura Social Municipal (FAISM) y al Fondo de Fortalecimiento Municipal (FORTAMUN).

	Municipio	IDH	FAISM	FORTAMUN
1	Astacinga	106.5	20.06	6.98
2	Atlahuilco	120.3	28.77	10.19
3	Ixhuatlancillo	70.8	29.65	14.49
4	Magdalena	97.8	5.45	3.72
5	Zongolica	125.5	50.61	10.42
6	Tequila	29.7	31.33	18.34
7	Texhuacán	114.6	17.65	5.22

Esta matriz está formada por 50 renglones, uno por cada municipio indígena, y por 4 columnas, la primera corresponde al municipio, la segunda al IDH municipal, la tercera al monto de FAISM destinado a un municipio, y la cuarta al monto de FORTAMUN. El renglón 1 corresponde al municipio indígena de Astacinga, como puede verse el IDH en este municipio es de 106.5, se tiene un monto otorgado por el FAISM de 20.06 millones de pesos; y un monto otorgado por el FORTAMUN de 6.98 millones de pesos.

2.1.2. Estadísticas descriptivas

Cuando se tienen datos multivariados se debe de estudiar el comportamiento del conjunto de individuos y variables en forma de un todo. Al igual que en el caso univariado, este comportamiento se estudia por medio de las estadísticas descriptivas. Algunas de las estadísticas descriptivas en el caso multivariado son las siguientes:

Vector de medias. Cuando se trabaja con una variable de estudio lo más representativo del comportamiento respecto a la tendencia central de la variable es su media, la cual está dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

En el caso multivariado ocurre lo mismo, pero al existir p variables de estudio se tienen p medias, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$, una por cada variable, que está dada por:

$$\bar{x}_p = \frac{\sum_{i=1}^{n} x_{ip}}{n}.$$

Con estas p medias se construye el vector de medias:

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Ejemplo (Gasto en salud). Para los datos sobre el gasto en salud que se destina por el Ramo33, el gasto estatal, las personas aseguradas y las personas no aseguradas, se obtiene el vector de medias que se muestra a continuación:

$$\bar{\mathbf{X}} = \begin{bmatrix} 1175.65 \\ 379.75 \\ 2396.15 \\ 4066.88 \end{bmatrix}$$

Del vector de medias se tiene que en promedio el gasto en salud de las personas no aseguradas es el mayor; que en promedio el menor gasto en salud es el proporcionado por el estado. Mientras que el gasto en salud promedio estatal por entidad es de 2396 millones de pesos.

Ejemplo (Municipios indígenas). Respecto a la información de 50 municipios indígenas de la entidad federativa Veracruz donde se registró el IDH, la aportación del FAISM y del FORTAMUN se obtiene el vector de medias de estas variables, el cual se muestra a continuación:

$$\bar{\mathbf{x}} = \begin{bmatrix} 0.08 \\ 72.50 \\ 32.73 \end{bmatrix}$$

En este caso, la media del IDH en los 50 municipios es de 0.08, mientras que 72.5 millones de pesos es el valor de la media del FAISM por municipio y la media del FORTAMUM tiene un valor de 32.73 millones de pesos.

Matriz de varianzas y covarianzas. En el caso univariado la segunda medida que nos da información sobre el comportamiento de la variable de estudio es la varianza, la cual para las variables está dada por:

$$s_{ii} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_i \right)^2}{\left(n - 1 \right)}$$

Y para el caso de covarianza sería para p variables:

$$s_{jk} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)\left(x_{ik} - \bar{x}_k \right)}{n}$$

En el caso bivariado; es decir, de tener 2 variables de estudio X_1 y X_2, se obtienen las varianzas de cada variable s_{11} y s_{22}, y además la covarianza entre éstas s_{12}. En el caso de tener p variables de estudio se tienen p varianzas, una por cada variable, y $p(p-1)/2$ covarianzas, las cuales se ponen en un matriz de orden p x p, a la cual se le denomina

matriz de varianzas y covarianzas y es denotada por $\mathbf{S}$. En la Figura 2.2 se presenta la matriz de varianzas y covarianzas.

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix}$$

Figura 2.2 Matriz de varianzas y covarianzas.

donde $s_{ii} = s_i^2$. Puede verse que la matriz de varianzas y covarianzas es una matriz simétrica, en su diagonal principal tiene las varianzas s_i^2 de cada una de las p variables y fuera de ésta las covarianzas s_{jk} entre cada par de variables bajo estudio.

Ejemplo (Gasto en salud). Tenemos la información sobre el gasto en salud que se destina por el Ramo33, el gasto estatal, las personas aseguradas y las personas no aseguradas, se obtiene la matriz de varianzas y covarianzas de estas variables, la cual se muestra a continuación:

$$\mathbf{S} = \begin{bmatrix} 281628 & 46513 & 739946 & -398765 \\ 46513 & 108734 & 201342 & 110120 \\ 739946 & 201342 & 2125110 & -1042276 \\ -398765 & 110120 & -1042276 & 4241334 \end{bmatrix}$$

De la matriz de varianzas y covarianzas se interpreta que la variable gasto en salud de las personas no aseguradas es el que presenta mayor variabilidad y gasto en salud del estado la que presenta una menor variabilidad. En cuanto a la variable gasto en salud del Ramo33, se aprecia se encuentra relacionado en forma negativa con el gasto de las personas no aseguradas, mientras que el gasto en salud de las personas aseguradas se encuentra relacionado en forma negativa con el gasto de las personas no aseguradas. Cabe hacer notar que tanto las varianzas como las covarianzas son cantidades difíciles de interpretar ya que su valor depende de la escala. Ante esta necesidad surge la Matriz de correlaciones, que se presenta a continuación.

Matriz de correlación. Otra medida descriptiva de las variables es su correlación lineal, entre cada par de variables, la cual para las variables X_j y X_k está dada por:

$$r_{jk} = \frac{s_{jk}}{s_j s_k}.$$

El valor de este coeficiente está comprendido entre -1 y 1. Cuando $r = 1$, se dice que la correlación lineal es perfecta directa o positiva. Si $r = 0$, no existe correlación lineal y cuando $r = -1$, la correlación lineal es perfecta inversa o negativa. En el caso de tener p variables de estudio se tienen p correlaciones de valor 1 (que es la correlación de la variable consigo misma), una por cada variable, y $p(p-1)/2$ correlaciones r_{ij} entre cada par de variables, las cuales se ponen en una matriz de orden $p \times p$, a la cual se le denomina matriz de correlaciones y es denotada por **R**; es decir, al igual que la matriz de varianzas y covarianzas, la matriz de correlaciones es una matriz simétrica, en su diagonal principal tiene unos y fuera de ésta los coeficientes de correlación r_{jk} entre cada par de variables bajo estudio.

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Figura 2.3. Matriz de correlaciones.

La relación entre la matriz de correlaciones y la matriz de varianzas y covarianzas está dada por:

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} \qquad \mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2},$$

donde $\mathbf{D} = \text{diagonal}(s_{11}, s_{22}, ..., s_{pp})$, es decir **D** es una matriz diagonal cuyas entradas son las varianzas de las p variables de estudio. Es inmediato que $\boldsymbol{D}^{-1/2} = \left(s_{11}^{-1/2}, s_{22}^{-1/2}, \cdots, s_{pp}^{-1/2}\right)$.

Ejemplo (Gasto en salud). Se obtiene la matriz de correlaciones de las variables del ejemplo de gasto en salud que venimos trabajando, la cual se muestra a continuación:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.266 & 0.956 & -0.365 \\ 0.266 & 1 & 0.419 & 0.162 \\ 0.956 & 0.419 & 1 & -0.347 \\ -0.365 & 0.162 & -0.347 & 1 \end{bmatrix}$$

De la matriz de correlaciones se aprecia que la variable gasto en salud del Ramo33 está altamente relacionada (0.956) con el gasto de las personas aseguradas. Asimismo, se sabe que esta variable se encuentra muy poco relacionada en forma lineal con el gasto estatal, así como con el gasto en salud de las personas no aseguradas, y que no está relacionada en forma lineal con ninguno de los otros tipos de gasto en salud. Es importante tener presente que la correlación lineal es una medida sobre el grado de asociación lineal entre dos variables, sin importar cuál es la causa y cuál es el efecto, se trata de la dependencia entre la variación de las variables.

Medidas de variabilidad global. Es conocido que en la matriz de varianzas y covarianzas **S** está la información relacionada con la dispersión de los dato; en ocasiones resulta de utilidad concentrar dicha información en una sola cantidad, por ejemplo cuando el objetivo es comparar distintos conjuntos de variables una etapa es obtener medidas de la variabilidad promedio, algunas de las medidas promedio de la variabilidad son: la Varianza total, la Varianza media, la Varianza generalizada y la Desviación típica generalizada, las cuales se describen a continuación:

- *Varianza total* (VT). Se define la varianza total como la traza de la matriz de varianzas y covarianzas. Es decir, $\mathrm{VT} = \sum_{j=1}^{p} s_j^2$.
- *Varianza media*. Se define como la varianza total entre el número de variables; es decir, $\frac{1}{p}\sum_{j=1}^{p} s_j^2$.
- *Varianza generalizada* (VG). Se define como el determinante de la matriz de varianzas y covarianzas. Es decir, $\mathrm{VG} = |\mathrm{S}|$.
- *Desviación típica generalizada*. Se define como la raíz cuadrada de la varianza generalizada.

Estas medidas son de gran utilidad en la comparación de grupos y sobre todo, para sustentar procesos de inferencia estadística multivariada.

2.1.3. Análisis multivariado gráfico

Al igual que en el análisis univariado, para estudiar el comportamiento de las variables la primera etapa es realizar un análisis exploratorio de los datos –a través de la obtención de una distribución de frecuencias de la elaboración de histogramas, graficas de cajas y alambres– para cada una de las variables. Como segunda etapa de un análisis gráfico se busca estudiar el comportamiento en forma bivariada para lo cual se realiza, para el caso de dos variables cuantitativas, un gráfico de dispersión o correlograma. Cuando se tienen más de dos variables en vez de realizar un gráfico de dispersión para cada par de variables, se realiza el llamado gráfico de matriz (matrix plot) o grafico de escalera, que es un grafico en el cual se presentan los diagramas de dispersión para cada par de variables que intervienen en el estudio. Siendo éste un grafico simétrico. Así, si se tienen 3 variables se tendrán 3 gráficos de dispersión que formarán el correlograma, si se tienen 4 variables habrá 6 gráficos de dispersión en el correlograma; en general si se tiene p variables habrá $p(p-1)/2$ gráficos de dispersión distintos.

Ejemplo (Gasto en salud). Se tiene información sobre el gasto en salud en el 2008 que se destinó en cada una de las 32 entidades federativas por el Ramo33, el gasto estatal (GasEst), las personas aseguradas (GasAse) y las personas no aseguradas (GasNoAse); estas cantidades están expresadas en miles de millones de pesos. El despliegue gráfico correspondiente se presenta a continuación en la Figura 2.4.

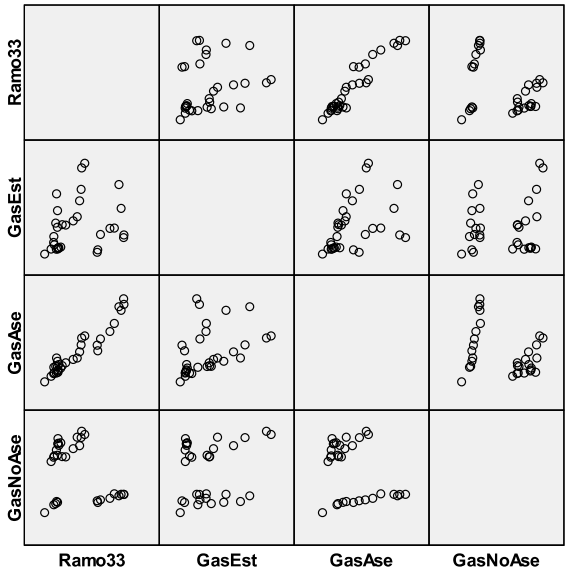


Figura 2.4. Gráfico de matriz para las variables de tipo de gasto en salud 2002.

Se observa que la variable Ramo33 se encuentra altamente relacionada con el gasto en salud de las personas aseguradas (GasAse), mientras que existe dos posibles grupos del gasto en salud por parte del Ramo33 respecto al gasto en salud de las personas no aseguradas (GasNoAse). Así también, el gasto en salud del Estado (GasEst) con el gasto en salud de las personas no aseguradas presenta dos posibles grupos, en ambos grupos se presenta una relación lineal, por lo cual esta relación no se puede detectar por medio del coeficiente de correlación. Para este estudio si observamos todos los diagramas de dispersión se puede ver la formación de dos grupos de entidades federativas.

Ejemplo (Ingresos paraestatales). Se tiene información sobre los ingresos durante el periodo 2003-2008 para el sector Primario (SERCPRIM), la Comisión Federal de Electricidad (CFE) y el Sector Petrolero (PEMEX); estas cantidades están expresadas en miles de millones de pesos. El matrix plot se presenta a continuación en la Figura 2.5.

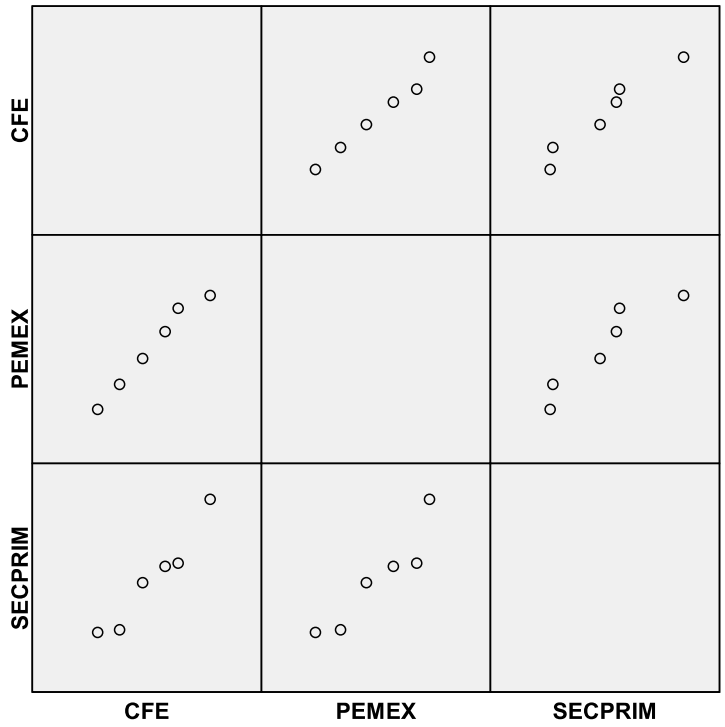


Figura 2.5. Gráfico de matriz de los ingresos del sector primario, PEMEX y CFE. Periodo 2003-2008.

Se observa que los ingresos de la CFE están altamente correlacionados con los ingresos tanto del sector petrolero (PEMEX), así como del sector primario (SECPRIM). El mismo comportamiento se observa de los ingresos del sector petrolero con los ingresos del sector primario.

2.1.4. Descripción de técnicas multivariadas

Las técnicas multivariadas que se utilizan con mayor frecuencia, se describen a continuación:

- *Análisis de conglomerados*. Si el interés es la agrupación de las unidades de estudio en grupos homogéneos de acuerdo a las variables de estudio, la técnica de conglomerados es la adecuada. Esta técnica forma grupos tales que las unidades dentro de cada grupo sean semejantes y aquéllas en grupos distintos no lo sean.
- *Análisis de correspondencia*. Esta técnica multivariada permite la visualización gráfica de tablas de contingencia, con el objetivo de poder identificar relaciones entre las categorías (niveles) de dos o más variables.
- *Análisis de componentes principales*. Si el objetivo es la reducción de dimensionalidad de un problema la técnica de componentes principales es la adecuada. Esta técnica se basa en la construcción de nuevas variables las cuales son combinaciones lineales de las variables originales, pero estas nuevas variables mantienen la información que sobre el fenómeno de estudio tienen las variables originales.
- *Análisis de correlación canónica.* El objetivo del análisis de correlación canónica es determinar la existencia de asociación entre dos conjuntos de variables, usando combinaciones lineales de las variables de cada conjunto haciendo máximo el coeficiente de correlación.

A continuación se presentan en forma breve estas técnicas, motivando al lector a que para un estudio más profundo de cada técnica revise en forma detallada las referencias. Además se presenta la manera en que se ejecuta la técnica en el software estadístico SPSS, haciendo hincapié en que hay disponibilidad de muchos otros paquetes estadísticos, en los cuales se pueden realizar aplicaciones de las técnicas mencionadas.

2.2 Análisis de conglomerados

Supóngase que a un número de empresas se les han medido un conjunto de variables y el interés está en poder agrupar a las empresas en clases, de modo tal que las empresas en

cada una de las clases sean más similares, de acuerdo a las variables estudiadas, a aquéllas que están en otra clase. También puede ser de interés para el gobierno que a partir de un conjunto de variables se puedan formar clases de entidades federativas que sean más similares entre sí dentro de cada grupo, que aquéllos que se encuentran en otro grupo respecto a un conjunto de variables económicas. En general, supóngase que se tienen *n* unidades de estudio a las cuales se les ha medido un conjunto de *p* variables, y el interés está en formar, a partir de estas *n* unidades, grupos tal que aquellas unidades dentro de cada grupo sean más similares entre sí, de acuerdo a las p variables bajo estudio, que aquellas unidades en otro grupo.

Este tipo de problemas pueden ser resueltos a partir del análisis de conglomerados o también conocido como "cluster analysis" o "análisis de cúmulos". El objetivo del análisis de cúmulos es la formación de grupos homogéneos de unidades de estudio en función de las similitudes entre estas unidades de acuerdo a un conjunto de variables medidas en cada una de las unidades de estudio. La pregunta es, si es posible tener una clasificación o agrupación que permita dividir las unidades de estudio, tal que las unidades que se encuentren dentro o formen un grupo sean semejantes entre sí, y que los grupos formados sean tan diferentes como sea posible.

El punto de partida es tener *n* unidades de estudio, tal que a cada una de las unidades se le ha medido *p* variables, con estas observaciones se forma la matriz de datos, la cual está dada por:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots \\ x_{(n\text{-}1)1} & x_{(n\text{-}1)2} & \cdots & x_{(n\text{-}1)p} \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

donde $\mathrm{x}_i^t = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}$ denota las observaciones de la *i*-ésima unidad de estudio. El análisis de conglomerados va a tomar en cuenta las mediciones de las *p* variables de cada una de las unidades de estudio para formar los grupos. Se dispone de datos y el objetivo es agruparlos en cúmulos o conglomerados, de tal manera que cada una de las unidades

pertenezca a uno, y sólo uno de los conglomerados; y que toda unidad de estudio se encuentre en un conglomerado; es decir, cada unidad quede clasificada.

2.2.1. Distancias

Para realizar un análisis de conglomerados, primero se debe determinar la distancia entre dos unidades para posteriormente la distancia entre dos grupos. Así se van a comparar las unidades i y j, las cuales tienen como vectores de datos a $\mathbf{x}_i = \lfloor x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip} \rfloor$ y $\mathbf{x}_j = \lfloor x_{j1} \quad x_{j2} \quad \cdots \quad x_{jp} \rfloor$, respectivamente. De las distancias disponibles se pueden mencionar las siguientes.

La distancia euclidiana. Se define como la raíz cuadrada de la suma de las diferencias al cuadrado de las p variables de las unidades i y j, es conocida también como distancia métrica.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}$$

La distancia de Minkowski. La distancia euclidiana es un caso particular de esta distancia cuando $\alpha = 2$

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{1/\alpha}}$$

La distancia de Mahalanobis. Esta distancia requiere la estimación de la matriz de varianzas y covarianzas

$$d_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

La distancia de Pearson. Como su nombre lo indica se basa en el coeficiente de correlación de Pearson

$$d_{ij} = r_{ij}$$

Ejemplo (Ingresos paraestatales). Se tiene información sobre los ingresos durante el periodo 2002-2007 para el sector Primario (SERCPRIM), la Comisión Federal de Electricidad (CFE) y el Sector Petrolero (PEMEX), estas cantidades están expresadas en

miles de millones de pesos. Se presenta la matriz de distancias para los años del periodo en estudio:

Caso	Distancia Euclidiana al cuadrado					
	2002	2003	2004	2005	2006	2007
2002	0.000	2.524	10.532	24.567	41.142	53.561
2003	2.524	0.000	2.799	11.403	23.337	33.017
2004	10.532	2.799	0.000	2.932	10.072	16.611
2005	24.567	11.403	2.932	0.000	2.142	5.648
2006	41.142	23.337	10.072	2.142	0.000	1.028
2007	53.561	33.017	16.611	5.648	1.028	0.000

Figura 2.6. Matriz de distancias.

De la matriz de distancias se tiene que los años 2006 y 2007 son los más similares respecto a los ingresos de los tres sectores, ya que la distancia entre estos dos años es la menor de todas; asimismo, la distancia entre los años 2002 y 2003 es pequeña, pudiendo decir que el comportamiento en estos dos años es muy similar respecto a los ingresos de las tres paraestatales; lo mismo ocurre entre los años 2003 y 2004. Mientras que los años que son más distintos de acuerdo a los ingresos de los tres sectores son el 2002 y 2007.

2.2.2. Métodos de agrupación

Los métodos de formación de grupos son de dos tipos:

- *Métodos jerárquicos*. Cada agrupación obtenida en cada paso es el resultado de agrupar varios grupos obtenidos en pasos anteriores; en esta situación es posible visualizar las agrupaciones intermedias cuando se pasa de un nivel a otro. Inicialmente cada unidad de estudio es un grupo en sí mismo y finalmente todas las unidades forman un sólo grupo.
- *Métodos no jerárquicos*. En este método los grupos no se forman a partir de grupos más pequeños. El número de grupos se establece de antemano y las unidades se clasifican en uno de estos grupos, de tal forma que las unidades en un grupo sean más homogéneas entre sí que aquéllas en otro grupo.

2.2.3. Algoritmos de agrupamiento

Métodos jerárquicos:

Vecino más próximo o más cercano. Este método se inicia con n agrupamientos, un agrupamiento por cada unidad de estudio; a partir de una distancia se agrupan las dos unidades más cercanas. A continuación se define la distancia entre este grupo y cualquier unidad como, la distancia mínima entre las unidades del grupo y la unidad. Se continúan formando grupos hasta que todas las unidades se encuentren en un solo grupo.

Vecino más lejano. La distancia entre los grupos se define como aquélla entre las unidades más alejadas.

Método del centroide. La distancia entre los agrupamientos se define como la distancia entre las medias de los grupos.

Método del promedio. La distancia entre los grupos se define como el promedio de todas las distancias entre todas las parejas posibles de unidades en cada grupo.

Métodos no jerárquicos:

K-medias. En este algoritmo se da el número *k*, que es el número de grupos a formar. Cada uno de los *k* grupos está caracterizado por su media, así a partir de las medias de los k grupos, cada unidad se asigna a aquel grupo cuya media este más cercana a esta unidad. Después de que se asignaron las unidades a los k grupos se vuelven a calcular las medias de cada grupo, y así sucesivamente hasta que se cumple un criterio de paro, que está determinado por un proceso de optimización del cociente entre las varianzas dentro y entre los grupos. Así se van a formar grupos tal que la varianza entre grupos sea máxima y dentro de los grupos la varianza sea mínima; es decir, la varianza de las unidades dentro de cada uno de los k grupos se minimiza.

2.2.4. Dendrograma

El dendrograma es una representación grafica del resultado del proceso de agrupamiento en forma de árbol. Es decir, se utiliza el dendrograma para representar la estructura jerárquica de la formación de los grupos.

Ejemplo (*Gasto en salud*). Se tiene información sobre el gasto en salud que se destinó en el 2008 para cada una de las 32 entidades federativas por el Ramo33; aparte del gasto estatal, se tiene el número de las personas aseguradas y las personas no aseguradas. El dendrograma de las 32 entidades federativas respecto al gasto en salud en el 2008 se presenta a continuación:

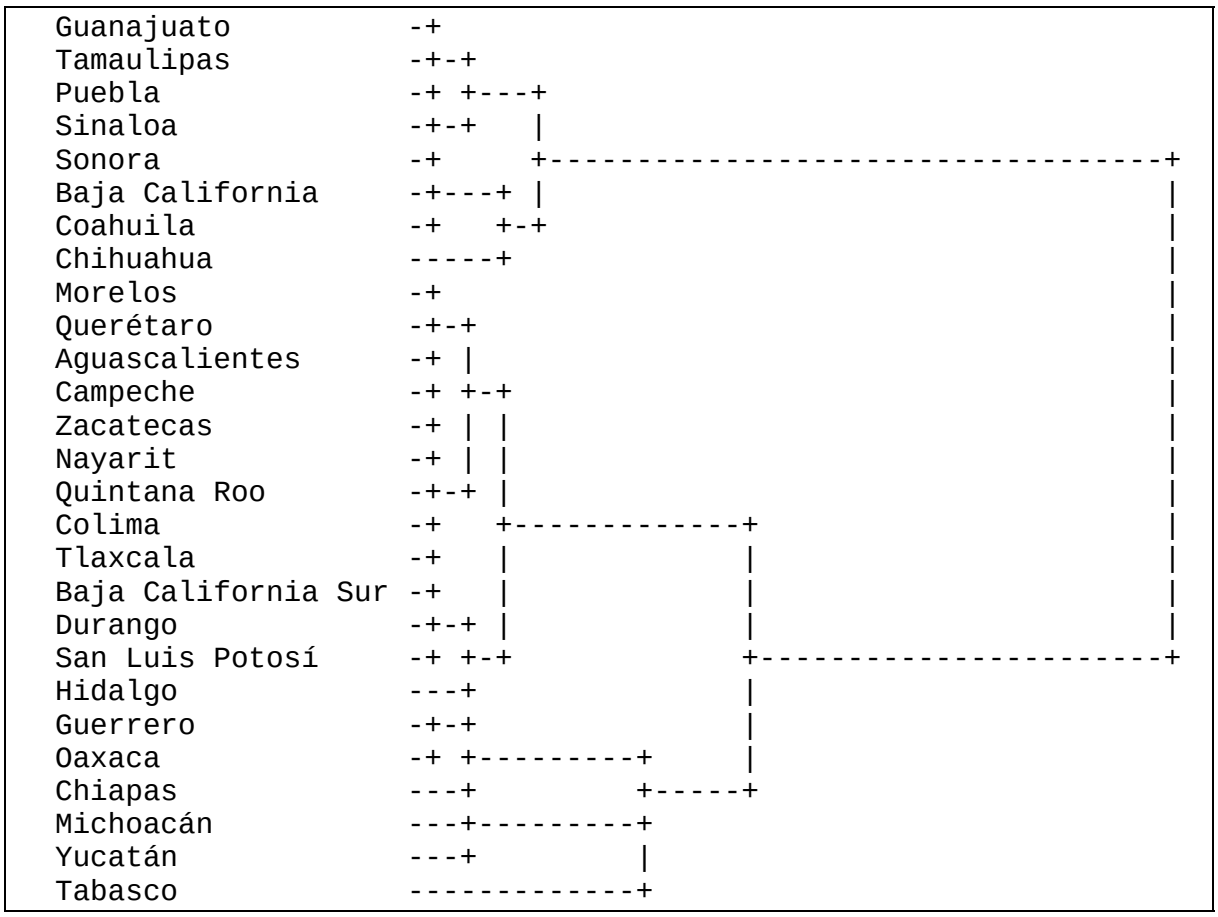


Figura 2.7. Dendrograma de Gasto en Salud 2008.

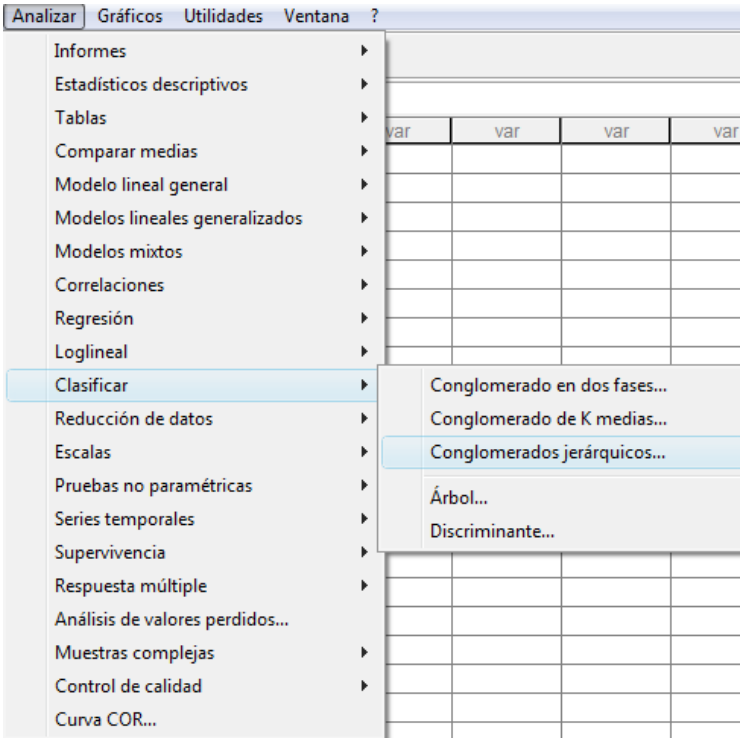
Se observa en la Figura 2.7 que las entidades en las que el valor de las variables en estudio para el gasto en salud es muy distinto, son: Michoacán, Yucatán y Tabasco, que junto con Chiapas y Oaxaca forman un grupo. Los otros dos grupos de entidades más compactos son bien identificados, al hacer un corte en el dendograma.

Caso práctico. En la entidad federativa de Veracruz, se tiene información sobre 19 municipios de acuerdo a algunas variables; el gobierno del estado de Veracruz desea agrupar a estos municipios de acuerdo a información obtenida, con el propósito de diseñar e implantar estrategias especiales de acuerdo a las agrupaciones. Para tal efecto ha

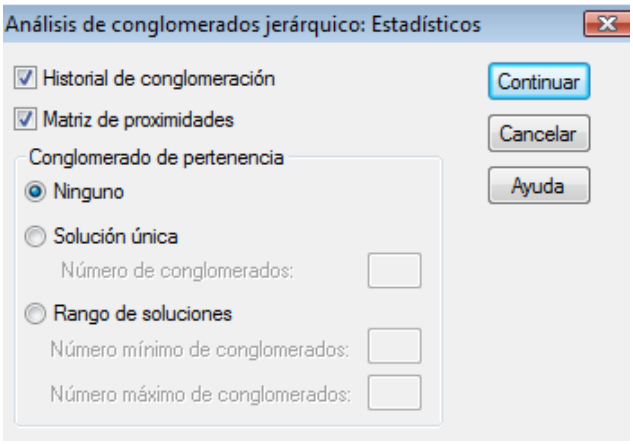
registrado las siguientes variables obtenidas en cada municipio: el Índice de Desarrollo Humano (IDH), el Fondo de Apoyo a la Infraestructura Social Municipal (FAISM) y el Fondo de Fortalecimiento Municipal (FORTAMUN).

Para realizar este ejercicio se ejemplificará su aplicación con el Software Estadístico SPSS:

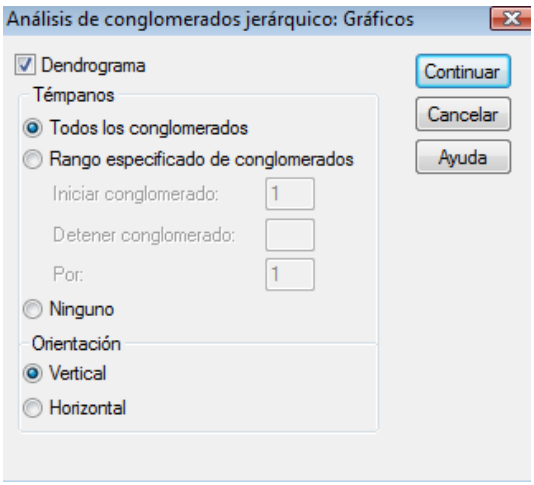
Dendograma. En la ventana de Analizar, se elige la opción *Clasificar*, y se elige la opción *Conglomerados jerárquicos*



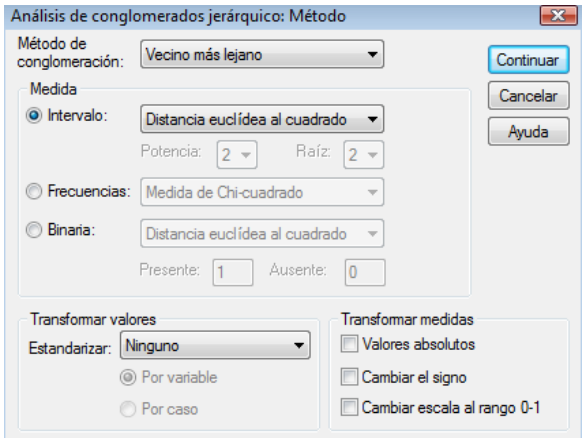
En la ventana de *Análisis de conglomerados jerárquicos* se debe elegir la ventana *Estadísticos*; al abrirse esta ventana elegir *Historial de conglomeración* y *Matriz de proximidades*.



En la ventana de *Análisis de conglomerados jerárquicos* elegir la ventana *Gráficos*, al abrirse esta ventana elegir *Dendrograma.*



En la ventana de *Análisis de conglomerados jerárquicos* elegir la ventana *Método*, al abrirse esta ventana elegir el *Método de conglomeración*, y la distancia a utilizar.



Salida obtenida: En primer lugar se muestra la matriz de distancias de las *n* unidades de estudio, en este caso los municipios.

Municipio	1	2	3	4	5	6	7
1	0.00	276.70	1422.96	299.65	1306.05	6154.42	74.51
2	276.70	0.00	2469.53	1091.96	503.87	8281.30	181.00
3	1422.96	2469.53	0.00	1430.52	3448.03	1706.83	2148.49
4	299.65	1091.96	1430.51	0.00	2851.15	5521.00	433.17
5	1306.04	503.86	3448.02	2851.15	0.00	9612.06	1232.31
6	6154.41	8281.29	1706.82	5521.00	9612.06	0.00	7567.36
7	74.51	181.00	2148.48	433.17	1232.31	7567.36	0.00
8	2315.92	4136.14	830.69	1362.68	6286.83	1896.41	3042.82
9	3232.35	4606.12	614.04	3886.30	4998.28	1016.82	4606.85
10	5669.82	5019.86	4373.87	7905.13	3055.64	7471.20	6397.05
11	993.23	1019.67	944.28	1989.68	913.88	4662.98	1408.57
12	1010.32	2204.70	226.11	691.65	3652.83	2367.50	1584.55
13	405.26	1218.50	366.31	382.10	2424.50	3436.57	811.90
14	4722.66	7084.03	1566.88	3491.21	9325.31	789.03	5830.32
15	26239.66	24125.20	22539.92	30972.48	18599.30	24684.38	27554.32
16	2725.97	3548.48	439.68	3237.31	3751.53	1575.67	3657.08
17	2497.66	3851.29	154.38	2342.09	4889.11	838.69	3432.81
18	1499.58	2743.31	63.06	1278.32	3970.09	1607.49	2225.27
19	1750.70	2772.32	31.88	1856.38	3594.84	1504.25	2545.60

Figura 2.8. Matriz de distancias de 19 municipios veracruzanos.

Se presenta sólo una parte de la matriz de distancias (ver Figura 2.8). Se observa que el municipio etiquetado como “1” es semejante al municipio etiquetado como “7”, también que estos dos municipios son semejantes al municipio 2; se espera que estos tres municipios pertenezcan a un grupo en una etapa temprana de la formación de los grupos. Los municipios “2” y “14” son muy distintos. Se observa que el municipio “15” es muy distinto a los demás municipios respecto las variables medidas, por lo que se puede pensar que este municipio solo formará un grupo, o bien se unirá a un grupo en una de las etapas finales. También a continuación se obtiene el historial de formación de los grupos en cada una de las etapas.

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado			Conglomerado		
	1	2		1	2	
1	3	19	31.879	0	0	5
2	1	7	41.513	0	0	6
3	9	16	81.907	0	0	11
4	12	18	98.539	0	0	7
5	3	17	131.928	1	0	8
6	1	2	228.852	2	0	10
7	12	13	248.64	4	0	8
8	3	12	385.736	5	7	11
9	8	14	502.937	0	0	13
10	1	4	608.261	6	0	14
11	3	9	837.958	8	3	15
12	5	11	913.883	0	0	14
13	6	8	1342.72	0	9	15
14	1	5	1413.068	10	12	16
15	3	6	1543.271	11	13	16
16	1	3	3219.773	14	15	17
17	1	10	5396.562	16	0	18
18	1	15	23235.123	17	0	0

Figura 2.9. Historial de conglomeración.

En la etapa 1 los municipios “3” y “19” formaron el primer grupo; en la etapa 2 se formó el grupo con los municipios “1” y “7”; en la etapa 3 los municipios “9” y “16” formaron un grupo; en la etapa 4 los municipios “12” y “18” se unieron, y en la etapa 5 al grupo formado en la etapa 1 se le unió el municipio “17”, tal como se indica en el primer reglón de la tabla “Historial de conglomeración”. En la etapa 6, al grupo formado en la etapa 2 se le unió el municipio 2, tal como se indica en el renglón 2. Por último se unen los municipios 10 y 15 (Figura 2.9). A continuación se obtendrá el dendrograma, el cual puede ser horizontal o vertical.

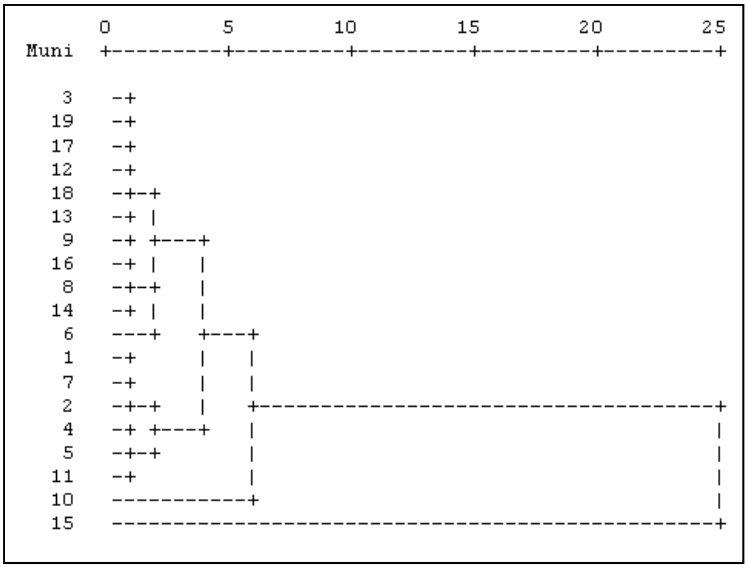


Figura 2.10. Dendrograma por municipio.

Municipios semejantes de acuerdo a las variables medidas son "3" y "19", "1" y "7", el municipio "15" es muy distinto a los demás municipios respecto a las variables bajo estudio, así como el municipio "10". Si el interés es formar 3 grupos, estos serían los grupos formados por los municipios $A = \{10\}$, $B = \{15\}$ y el grupo C estaría formado por todos los demás municipios (ver Figura 2.10). Resulta, entonces, más conveniente formar 4 grupo: dos, que sería el A y el B, y el C se partiría en 2, que serían C_1={1,7,2,4,5,11} y C_2={3,19,17,12,18,13,9,16,8,14,6}.

2.3 Análisis de correspondencia simple

El objetivo del análisis de correspondencias es la reducción de la dimensionalidad y la representación grafica de la relación "correspondencia" existente entre dos o más variables categóricas. Es decir, es una técnica que permite analizar la asociación desplegando un mapa de correspondencia entre las categorías de las variables. En el grafico se muestra una distancia entre las categorías de las variables. Esto permite identificar patrones de asociación entre las categorías de las variables y, con esto, identificar patrones de asociación en las unidades de estudio.

2.3.1. Tablas de contingencia

El punto de partida de un análisis de correspondencia es una tabla de contingencia. Considérese dos variables A y B con tres y cuatro categorías respectivamente; denótese por

n_{ij} el número de unidades que corresponden al mismo tiempo a la categoría i de la variable A y a la categoría j de la variable B. Una tabla de contingencia es una tabla de conteos (frecuencias absolutas) de dos entradas, donde el conteo n_{ij} es reportado. En el caso de las dos variables A y B mencionadas se tiene una tabla de contingencia 4×3.

Tabla 2.1. Tabla de contingencia.

$A \mid B$	A_1	A_2	A_3	Suma renglón
B_1	n_{11}	n_{12}	n_{13}	n_{1+}
B_2	n_{21}	n_{22}	n_{23}	n_{2+}
B_3	n_{31}	n_{32}	n_{33}	n_{3+}
B_4	n_{41}	n_{42}	n_{43}	n_{4+}
Suma columna	n_{+1}	n_{+2}	n_{+3}	n_{++}

En forma general una tabla de contingencia ($I \times J$) es un arreglo de I renglones y J columnas, en donde las entradas son las frecuencias absolutas de dos variables cualitativas de n elementos. La primera de las variables cuenta con J categorías o niveles de la variable, mientras que la segunda variable cuenta con I categorías. Cada una de las n unidades de estudio se puede clasificar en una y sólo una de las J categorías de la primera variable y una y sólo una de las I categorías de la segunda variable. Esta tabla de contingencia tiene información de las variables y una técnica para estudiar la asociación entre las categorías es el análisis de correspondencia.

Ejemplo (Partido y marginación). Se tiene una información sobre el partido gobernante y sobre el grado de marginación en 110 municipios del país. Estos datos se pueden presentar en la siguiente tabla de contingencia. El interés en este caso es conocer si existe relación entre el grado de marginación del municipio y el partido gobernante.

Tabla 2.2. Tabla de contingencia de partido político contra marginación.

	BAJO	*MEDIO*	*ALTO*	Suma renglón
PRI	13	10	8	31
PAN	10	8	12	30
PRD	12	9	8	29
OTRO	7	8	5	20
Suma columna	42	35	33	110

De la Tabla 2.2, se tiene que en 12 municipios el partido gobernante es el *PAN* y se tiene un grado de marginación *ALTO*, mientras que en 29 municipios el partido gobernante es el *PRD*.

La manera usual de llevar a cabo un análisis estadístico de una tabla de contingencia es por medio de pruebas estadísticas de asociación que están basadas en la estadística Chi-cuadrada χ^2, con los grados de libertad correspondientes a los renglones y columnas de la tabla de contingencia. El análisis de correspondencia es una técnica que se usa para analizar los renglones y columnas de una tabla de contingencia, y muestra simultáneamente las relaciones existentes entre los renglones y entre las columnas; estas relaciones se presentan en una gráfica que muestra a los renglones y columnas de la tabla como puntos en el plano cartesiano.

Para realizar el análisis de correspondencias a partir de la tabla de contingencia se obtiene la tabla de frecuencias relativas, la cual está formada por los cocientes n_{ij}/n_{++}; es decir, las entradas de la tabla de frecuencias relativas son cada una de las entradas de la tabla de contingencia entre el toral *n*. Esta matriz de frecuencias relativas se denota por *F*. En el caso de una tabla de contingencia 4×3 está dada por medio de:

	A_1	A_2	A_3	Suma renglón
B_1	$\frac{n_{11}}{n_{++}}$	$\frac{n_{12}}{n_{++}}$	$\frac{n_{13}}{n_{++}}$	$\frac{n_{1+}}{n_{++}}$
B_2	$\frac{n_{21}}{n_{++}}$	$\frac{n_{22}}{n_{++}}$	$\frac{n_{23}}{n_{++}}$	$\frac{n_{2+}}{n_{++}}$
B_3	$\frac{n_{31}}{n_{++}}$	$\frac{n_{32}}{n_{++}}$	$\frac{n_{33}}{n_{++}}$	$\frac{n_{3+}}{n_{++}}$
B_4	$\frac{n_{41}}{n_{++}}$	$\frac{n_{42}}{n_{++}}$	$\frac{n_{43}}{n_{++}}$	$\frac{n_{4+}}{n_{++}}$
Suma columna	$\frac{n_{+1}}{n_{++}}$	$\frac{n_{+2}}{n_{++}}$	$\frac{n_{+3}}{n_{++}}$	$\frac{n_{++}}{n_{++}}$

Ejemplo (Partido y marginación). Se tiene una información sobre el partido gobernante y sobre el grado de marginación en 110 municipios del país. Estos datos se pueden presentar en la siguiente tabla de contingencia de frecuencias relativas.

	BAJO	*MEDIO*	*ALTO*	Suma renglón
PRI	0.1182	0.0909	0.0727	0.2818
PAN	0.0909	0.0727	0.1091	0.2727
PRD	0.1091	0.0818	0.0727	0.2636
OTRO	0.0636	0.0727	0.0456	0.1819
Suma columna	0.3818	0.3181	0.3001	1

En general en una tabla de frecuencias relativas, que proviene de una tabla de contingencia (*I* X *J*), se tienen *I* renglones y *J* columnas. Las *I* filas se pueden tomar como *I* puntos en R^J y el objetivo del análisis de correspondencia es obtener una representación de estos *I* puntos en R^J en un espacio de dimensión menor y así poder observar las distancias entre éstas. Por ello es que el análisis de correspondencia es en este sentido similar al análisis de componentes principales, técnica que revisaremos más adelante.

Cabe destacar, que no todos los renglones en una tabla de contingencia tienen el mismo peso; se debe de tomar en cuenta el número de casos que contiene cada renglón. Y al estudiar la asociación se debe dar mayor peso a los renglones que contienen más casos. Para comparar ya sea dos renglones o dos columnas en una tabla de contingencia se deben de comparar los porcentajes y no los valores originales n_{ij}; estos porcentajes se denominan perfil renglón y perfil columna.

2.3.2. Perfil renglón (columna)

El perfil del *i*-ésimo renglón, está definido como n_{ij}/n_{i+}; esto es, cada una de las entradas que conforman el renglón se dividen entre el total que corresponde a ese renglón. El perfil para la *j*-ésima columna se define por n_{ij}/n_{j+}; esto es, cada una de las entradas que conforman la columna se dividen entre el total que corresponde a esa columna. Así se tienen *J* perfiles renglón e *I* perfiles columna.

Para los datos que aparecen en la tabla 2.2 se tienen 4 perfiles renglón y 3 perfiles columna; en este caso, para construir el primer perfil renglón se debe de tener n_{1+}, que tiene el valor de 31; así el primer perfil renglón está dado por:

$\frac{13}{31}$	$\frac{10}{31}$	$\frac{8}{31}$

Mientras que el tercer perfil columna está dado por:

8/33
12/33
8/33
5/33

El objetivo del análisis de correspondencia es la representación de estos perfiles usando un número de dimensiones (ejes principales) que sea lo menor posible (generalmente 2 ó 3) y a la vez se busca conservar la mayor información presente en todas las dimensiones. Así, el papel que juegan los ejes principales o dimensiones usadas, es condensar la mayor cantidad posible de información que sobre la variabilidad entre perfiles renglón y perfiles columna tiene la tabla de contingencia. Por lo general se busca que sean dos ejes principales para representar gráficamente los perfiles y así poder resaltar las relaciones entre ellos. Recuérdese que en vez de estudiar la relación entre los valores originales n_{ij} en el análisis de correspondencia se estudia la relación entre los perfiles, y los ejes principales tienen la propiedad de permitir estudiar las relaciones entre los perfiles renglón y los perfiles columna de manera simultánea.

Inercia. Un concepto de suma importancia del análisis de correspondencia es la inercia, la cual es una medida de la dispersión o variabilidad de los perfiles. La inercia es una medida de la variación explicada y es el cuadrado del valor propio, que indican la contribución relativa de cada dimensión en la explicación de la variación de las categorías; esto es, que a mayor inercia mayor es la distancia entre los perfiles. El número de dimensiones se puede elegir de acuerdo al porcentaje acumulado de inercia asociada a las dimensiones. La bondad de la representación de los perfiles será mayor cuanto más sea la inercia explicada por los ejes principales; es decir, que la representación gráfica es de mayor calidad en la medida que la inercia es mayor.

En el análisis de correspondencia el interés está en:

- Comparar los perfiles renglón; es decir, representar la variabilidad entre los renglones.

- Comparar los perfiles columna; es decir, de manera equivalente al estudio de los renglones.
- Investigar las asociaciones de los perfiles renglón y perfiles columna; el interés está en la representación de la correspondencia entre las categorías de los renglones y columnas. Esto sirve para identificar patrones de asociación.

Para llevar a cabo la comparación de perfiles es necesario tener una distancia; en este caso la distancia que se usa para la comparación de perfiles es la distancia Chi-cuadrada; dos perfiles renglón (columna) son parecidos si producen distancias pequeñas y si dos perfiles renglón (columna) son diferentes producen distancias grandes. Al formarse un grafico de los perfiles renglón y los perfiles columna se puede observar la dispersión entre los perfiles, tanto renglón como columna.

2.3.3. Reglas de interpretación

En el análisis de correspondencia se grafican los perfiles renglón y los perfiles columna en el plano generado por los primeros ejes principales (comúnmente dos ejes principales). La interpretación de los resultados del análisis de correspondencia simple se basa en tres elementos:

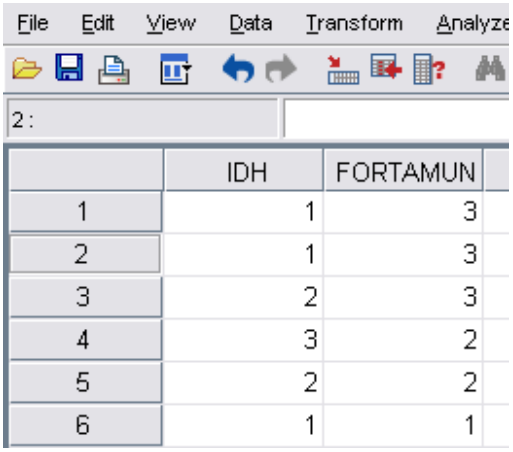
1. Si dos perfiles de renglón (columna) tienen una estructura semejante, su ubicación geométrica en el plano generado por los dos primeros ejes principales, será próxima. Lo inverso no siempre es cierto, a menos que la calidad de la representación de los perfiles sea muy buena (la inercia sea cercana a 100%).
2. La cercanía de un punto fila con un punto columna sólo se puede interpretar si están alejados del origen. Geométricamente, un perfil de renglón tenderá a estar en una posición geométrica la cual corresponde a las categorías de la variable en las columnas que son prominentes en dicho perfil de renglones. Sin embargo, en general no se debe interpretar las distancias entre los punto fila con los puntos columna; a no ser que están más allá del cuadrante (-1, 1) en ambos ejes.
3. Cuando un perfil renglón (columna) es próximo (parecido) al perfil renglón (columna) medio, es decir, tiene un comportamiento medio, se ubicará próximo al origen.

Caso práctico: Se recabó información de 862 municipios relacionada al Índice de Desarrollo Humano (IDH) y al Fondo de Fortalecimiento Municipal (FORTAMUN). El objetivo es conocer si las categorías del IDH están asociadas a algunas de las categorías del FORTAMUN.

IDH 1. Menor de 0.3 2. De 0.3 a 0.7 3. Mayor a 0.7

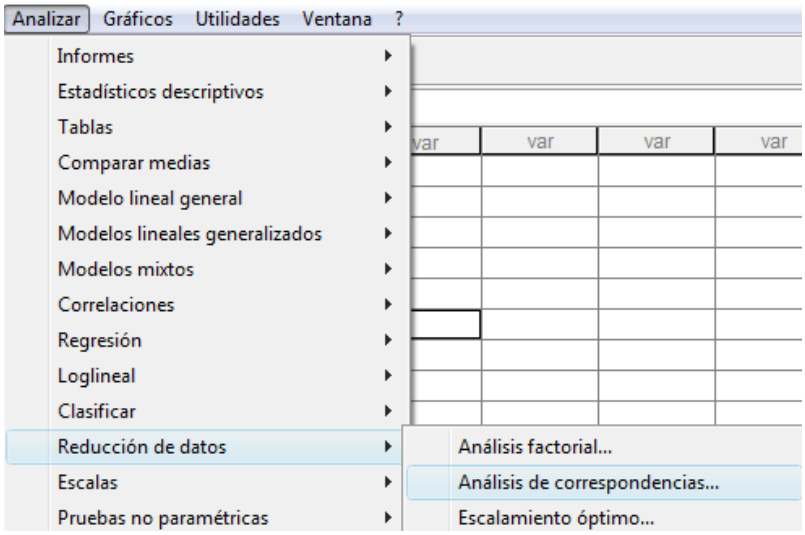
FORTAMUN 1. Menor de 5,500 2. De 5,500 a 12,000 3. Mayor a 12,000

Paso 1. Se forma la base de datos en SPSS de la siguiente manera

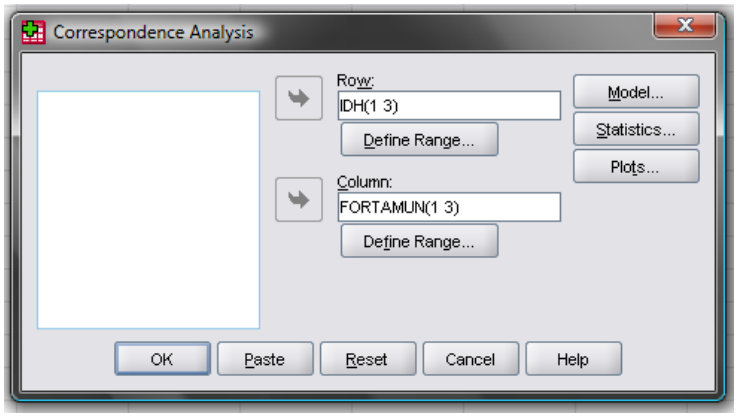


	IDH	FORTAMUN
1	1	3
2	1	3
3	2	3
4	3	2
5	2	2
6	1	1

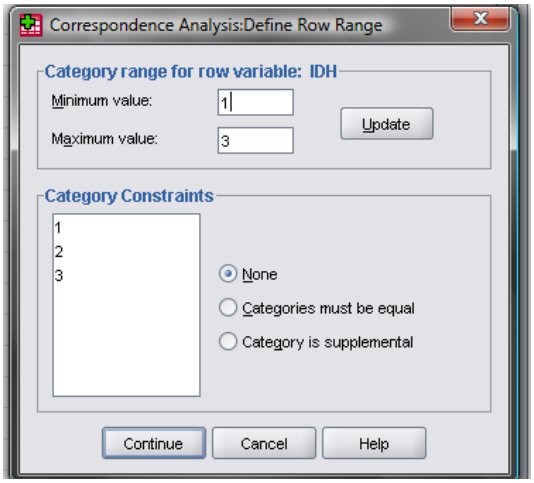
Paso 2. En la ventana de Analizar, se elige la opción Reducción de datos, y se elige la opción Análisis de correspondencias



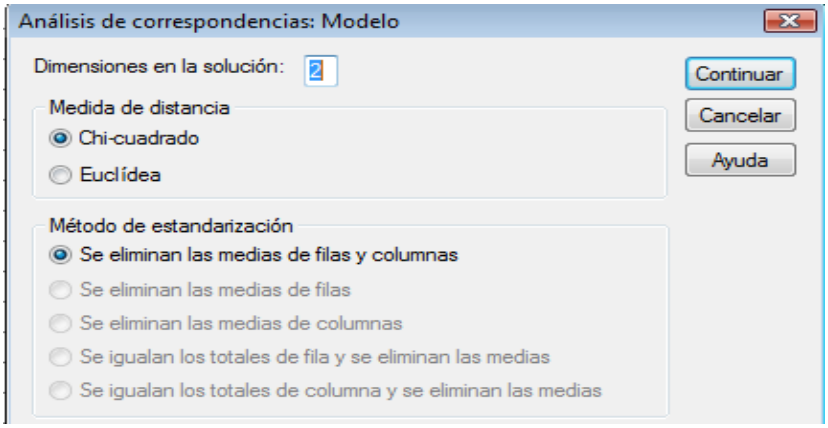
Paso 3. Se teclean las variables



Paso 4. Se teclear el número de categorías de cada una de las variables

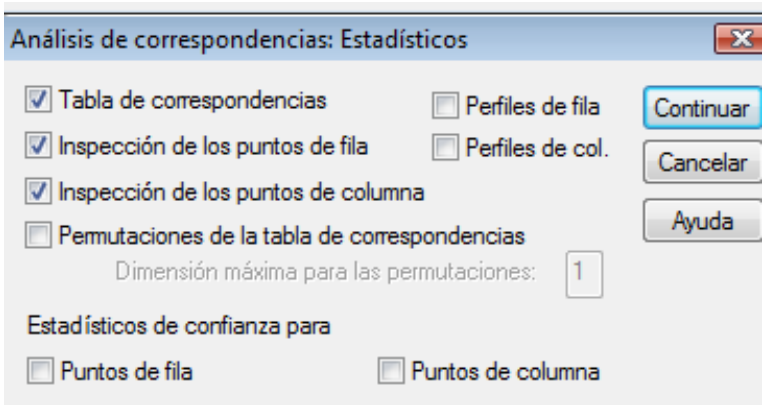


Paso 5. Se teclean el número de dimensiones y se elige la distancia Chi-cuadrada

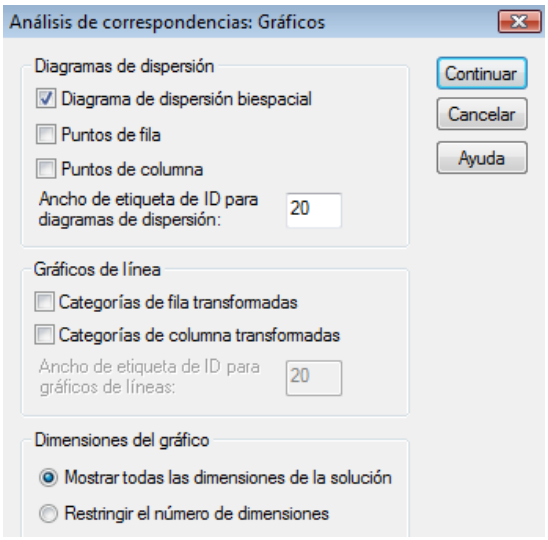


Dimensión Menor número de renglones (columnas) - 1

Paso 6. Se teclean los estadísticos a obtener



Paso 7. Se elige el grafico de dispersión



La salida que arrojo el paquete se muestra a continuación:

IDH	FORTAMUM			
	1	2	3	Total
1	79	118	59	256
2	238	138	28	404
3	18	46	138	202
Total	335	302	225	862

Figura 2.11. Tabla de correspondencia del IDH contra FORTAMUN.

De la Figura 2.11, se puede saber que hay en el estudio 79 municipios con un IDH menor a 5 y cuyo nivel de FORTAMUN es menor a 5,500. Asimismo, hay 302 municipios con FORTAMUN de entre 5,500 y 12,000. A continuación se presenta la tabla resumen, en la cual se muestra las dimensiones, la inercia asociada a cada dimensión y la prueba de significancia.

Dimensión	Valor singular	Inercia	Chi cuadrada	Sig.	Proporción de inercia	
					Explicada	Acumulada
1	0.576	0.332			0.935	0.935
2	0.152	0.023			0.065	1.000
Total		0.355	306.041	0.000	1.000	1.000

El valor del estadístico de prueba es de 306.041 con un p-value de 0.000, que a un nivel de significancia de 0.05 indica dependencia entre el IDH y el FORTAMUN. La proporción de la inercia indica que con la primera dimensión se tiene aproximadamente un 94% de la variación explicada de los perfiles y con la segunda dimensión se tiene el 100%.

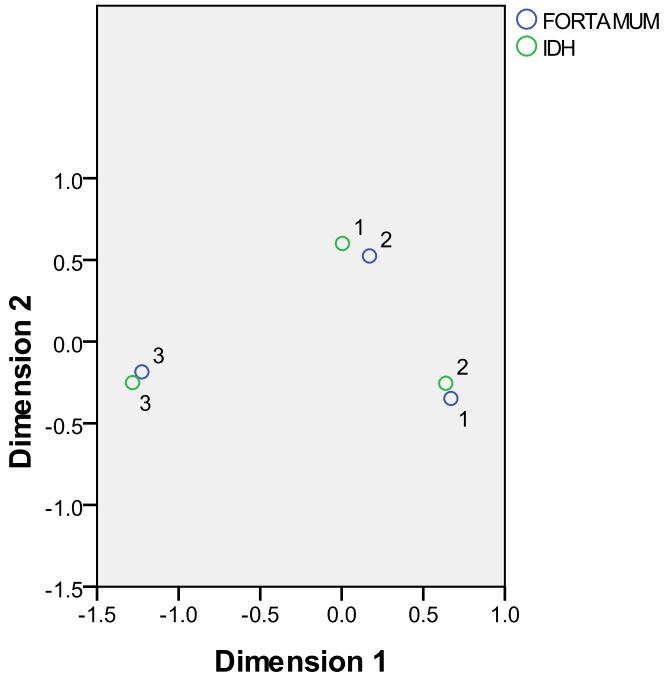


Figura 2.12. Gráfico de correspondencias entre el IDH con el FORTAMUN.

Se observa en la Figura 2.12, que la categoría del IDH "Mayor de 7", y la categoría de FORTAMUN "Mayor a 12,000" están juntas y alejadas del origen, es decir fuera del rectángulo (1,1), (1,-1), (-1,1) y (-1,-1). Por lo que existe asociación entre tales categorías. Aunque las categorías "1" de IDH y "2" de FORTAMUN se encuentran próximas se debe de destacar que no es válido dar una interpretación de esta cercanía ya que los puntos se encuentran dentro del rectángulo mencionado. Similarmente ocurre con las categorías "2" de IDH y "1" de FORTAMUN. Así, los municipios cuyo FORTAMUN es mayor a 12,000 tienen un IDH mayor de 7. Por lo que podría pensarse que un municipio con FORTAMUN mayor a 12,000 se encuentra asociado a un valor de IDH mayor a 7.

2.4 Análisis de componentes principales

El análisis de componentes principales (ACP) es uno de los métodos multivariados más simples, y más usados, y por ello de los más importantes. El objetivo en este análisis es a partir de un conjunto de variables cuantitativas obtener otro conjunto de nuevas variables, denominadas "los componentes principales", tales que estas nuevas variables nos faciliten el análisis de la variabilidad de los elementos del colectivo y también la correlación lineal entre las variables originales; todo sin perder la información relevante en los datos originales. En algunos estudios la información que se tiene depende de muchas variables, que además están correlacionadas; en este marco se desea trabajar con un número menor de variables; en esta parte –que se denomina reducción de la dimensionalidad del problema– es donde el ACP resulta de gran utilidad. Esta técnica es usada en Economía para definir índices de desarrollo social, económico, urbanístico, de marginación, de desarrollo humano, etc. Así también es una técnica de gran uso en otras áreas como Biología, Agronomía, Educación, entre otras.

2.4.1. Estrategias de uso del análisis de componentes principales

El ACP transforma un conjunto de variables correlacionadas en un conjunto menor de variables no correlacionadas que se denominan "los componentes principales". También es útil cuando se desea que las unidades de estudio se organicen en subgrupos. También puede ser de utilidad para resolver el problema de multicolinealidad, que es un problema que se presenta en la regresión múltiple.

Los principales objetivos del ACP son: Estudiar la estructura de asociación entre variables, reducir la dimensionalidad del problema, explorar agrupación y discriminación en un espacio reducido, construir índices o nuevas variables para futuros análisis.

2.4.2. Procedimiento

La construcción de los componentes principales que son combinaciones lineales de las variables originales:

$$\begin{aligned} C_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ C_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ C_p &= a_{k1}X_1 + a_{k2}X_2 + \cdots + a_{kp}X_p, \end{aligned}$$

sigue un procedimiento de optimización. Se tendrán tantos componentes principales como variables originales; es decir, al principio se tiene las variables originales $X_1, X_2, \cdots, X_P$, las cuales se transforman en $C_1, C_2, \cdots, C_P$, que tienen la misma información sobre la variabilidad del estudio que las variables originales, pero con la ventaja de que no están correlacionadas. Cada uno de los componentes principales $C_1, C_2, \cdots, C_P$, tiene un porcentaje de la varianza, pero estas varianzas están ordenadas de manera decreciente –la primera es la más grande, y así sucesivamente– garantizando que la suma de la varianza de los componentes sea igual a la suma de las varianzas de las variables originales. Los paquetes estadísticos presentan en su salida cada componente y la varianza asociada, además del porcentaje de la varianza total. Por lo que, como la estrategia de reducción de dimensionalidad es trabajar con un número menor de variables que el número de variables originales, el investigador fija un porcentaje alto –digamos 80 ó 90 por ciento– y entonces seleccionamos el número de componentes que de manera acumulada cubran este porcentaje.

Para aplicar correctamente el ACP se tienen que considerar tres observaciones:

1. Cuando todas las variables están en la misma escala el ACP se lleva a cabo a partir de la matriz de varianzas y covarianzas **S**, mientras que cuando las variables están en diferente escala se debe de usar la matriz de correlaciones **R**.

2. Para que tenga caso llevar a cabo un ACP las variables originales deben de presentar correlación. De no ser así el número de componentes principales será casi similar al número de variables originales.
3. Como medida de la cantidad de información incorporada en un componente se utiliza su varianza. Así a mayor varianza del componente principal implica que tiene mayor información de las variables originales.

Como ya se mencionó, los componentes principales se construyen de manera que el primer componente principal tiene la máxima varianza posible, el segundo componente principal la segunda mayor varianza posible que no fue explicada por el primer componente principal, y así hasta el último. Otro atributo importante es que los componentes principales son variables no correlacionadas.

Ya se especificó que la varianza de los componentes principales indica la importancia de cada uno de estos, lo que se denota por:

$$Var(C_1) \geq Var(C_2) \geq \cdots \geq Var(C_p),$$

o definiendo $\lambda_i = Var(C_i)$, se tiene que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. En el paquete estadístico se presenta la siguiente tabla:

j	λ_j	% varianza parcial	% varianza acumulada
1	λ_1	$\frac{\lambda_1}{\sum \lambda_i} \times 100$	$\frac{\lambda_1}{\sum \lambda_i} \times 100$
2	λ_2	$\frac{\lambda_2}{\sum \lambda_i} \times 100$	$\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} \times 100$
⋮	⋮	⋮	⋮
p	λ_p	$\frac{\lambda_p}{\sum \lambda_i} \times 100$	$\frac{\sum \lambda_i}{\sum \lambda_i} \times 100$

En esta tabla se puede ver que se indica el valor del valor propio, el porcentaje de la varianza total atribuido a cada componente principal, así como el porcentaje acumulado.

Los pesos de las variables originales en cada uno de los componentes principales son los coeficientes a_{ij} utilizados en la construcción de los componentes principales como

combinaciones lineales; es decir, el peso que tiene la j-ésima variable X_j en el i-ésimo componente Y_i está dado por a_{ij}. Así la matriz de pesos está dada por:

CP	X_1	X_2	···	X_p
C_1	a_{11}	a_{12}	···	a_{1p}
C_2	a_{21}	a_{22}	···	a_{2p}
...	...	...	...	...
C_p	a_{p1}	a_{p2}	···	a_{pp}

Para interpretar cada CP se observan los pesos; una variable "pesa" en el CP si el peso correspondiente es mayor a la mitad del valor absoluto del peso mayor. Los signos de los pesos permiten interpretar las correlaciones. Para saber cuál es el número correcto de los CP se utilizan tres criterios:

1. Los que acumulan un porcentaje de varianza especificado. Los datos de la varianza explicada son muy importantes para saber cuántos componentes principales se van a utilizar, con lo que se debe de decidir en función de la proporción de la varianza acumulada. Un porcentaje de 80% se considera bueno, así que si con los dos primeros componentes principales se explica un 79% y con los tres primeros un 84% es preferible quedarse con dos componentes, aunque esto depende del tipo de aplicación, y por tanto de la variabilidad en los datos.
2. Los que tengan un valor característico mayor que 1 (si se usa **R**). Cuando se utiliza la matriz de varianzas y covarianzas elegir aquellos valores mayores a la varianza media $\sum \lambda_i/p$.
3. Graficar λ_i contra j. Seleccionar los componentes hasta que los restantes tengan aproximadamente el mismo valor de λ_i. La idea es buscar un codo en el grafico; es decir, un punto a partir del cual los valores propios son aproximadamente iguales. Existe un grafico, llamado gráfico de sedimentación, en el cual se presentan los valores propios y se observa a partir de cuál se tiene el codo (ver Figura 2.13).

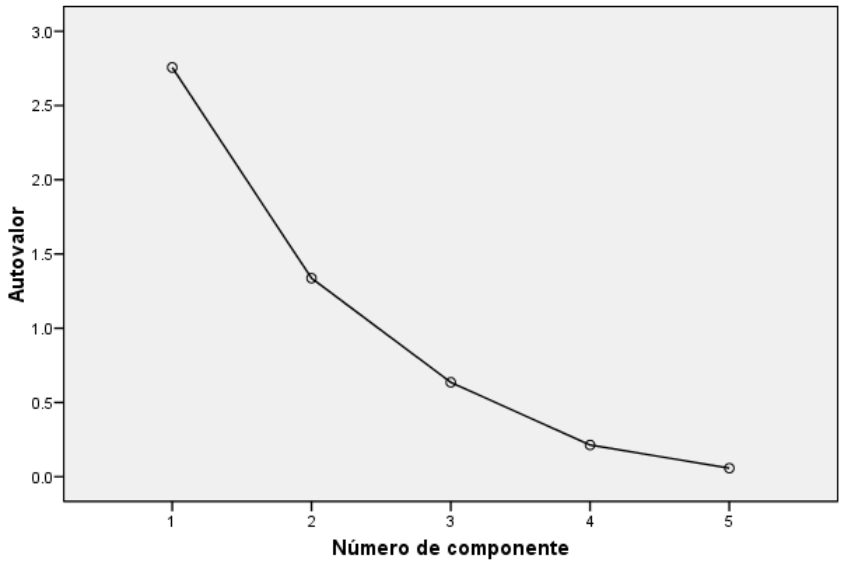
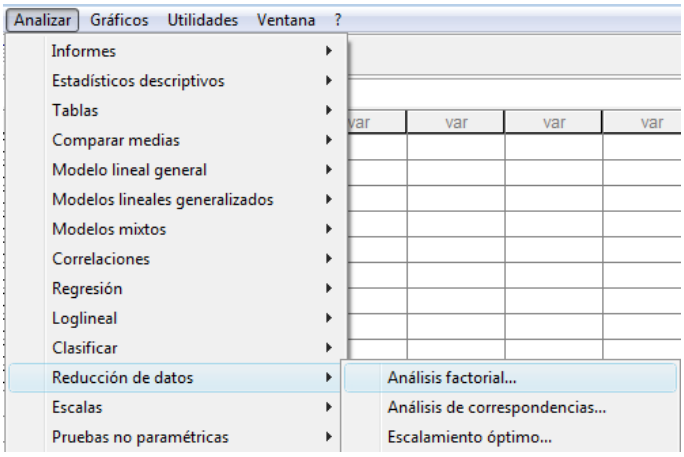


Figura 2.13. Gráfico de sedimentación.

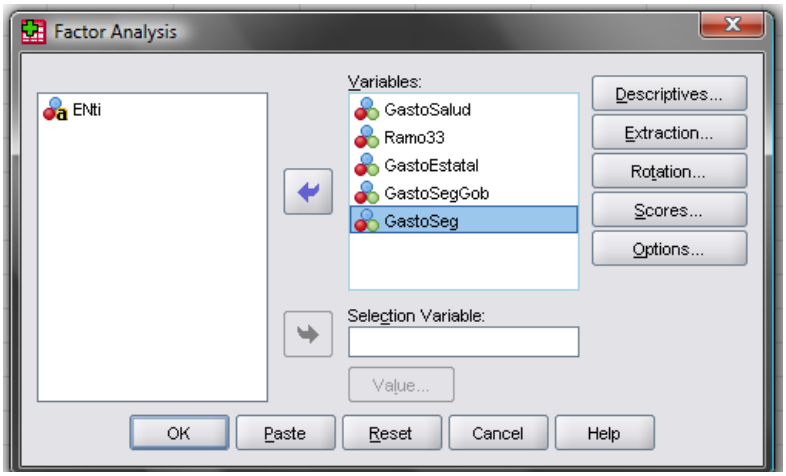
Los componentes principales son "nuevas variables" a las que hay que "dar nombre" y saber interpretar. Un aspecto clave en ACP es esta interpretación, ya que no viene dada a priori, sino que será deducida tras observar la relación de los componentes con las variables iniciales. Habrá entonces, que estudiar tanto el signo como la magnitud de las correlaciones. Esto no siempre es fácil, y depende en gran medida del conocimiento que el investigador tenga sobre las correlaciones entre las variables originales que describen el fenómeno de estudio. La interpretación se da a partir de los pesos y en ella se tiene que describir la naturaleza de cada componente, lo que se hace mediante la identificación de las variables originales que están asociadas con el componente; es decir, las variables que tienen coeficientes altos (se dice "los que pesan") en el componente.

Caso práctico: Se tiene información sobre 5 tipos de gastos que realizan las 32 entidades federativas del país; el objetivo es reducir el número de variables para poder realizar estudios posteriores y a la vez ver la posible formación de grupos de entidades para implementar programas federales referentes a la gasto. Los gastos de los que se tiene información son: Gasto en Seguridad, Gasto en la Vivienda, Gasto del Gobierno, Gasto en Salud, y Gasto en Obras. Los pasos en el paquete estadístico SPSS son:

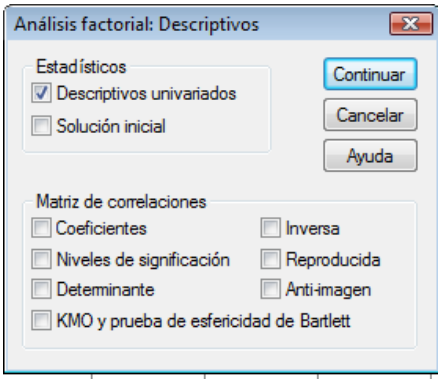
Paso 1. En la ventana de *Analizar*, se elige la opción *Reducción de datos*, y se elige la opción *Análisis factorial.*



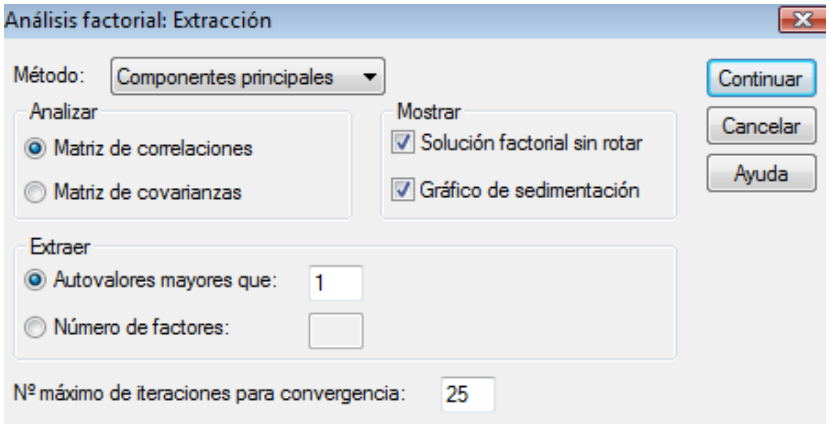
Paso 2. En la ventana de *Análisis factorial*, se eligen las variables originales para el análisis de componentes principales.



Paso 3. En la ventana de *Análisis factorial*, se da click en la opción *Descriptivo*, y se elige la opción *Descriptivos univariados.*



Paso 4. En la ventana de *Análisis factorial*, se da click en la opción *Extracción*, y se elige la opción *Matriz de correlaciones*; al igual se hace en *Gráfico de sedimentación*, y se teclea el número de componentes deseados, o en su defecto aquellos mayores a cierto valor.



Salida: En la siguiente tabla se presenta la salida del paquete, que muestra los valores propios, el porcentaje de varianza de cada componente y el porcentaje de varianza acumulada.

j	λ_j	% varianza parcial	% varianza acumulada
1	3.073	61.451	61.451
2	1.052	21.048	82.499
3	0.818	16.362	98.862
4	0.570	1.138	100
5	0.000	0.000	100

Tomando el criterio del valor característico mayor que 1, el número de componentes principales es de 2, pero el valor del tercer componente principal es de 0.818. Se observa que con los dos primeros componentes principales se tiene aproximadamente el 82% de la información del comportamiento de los gastos de la entidad federativa, mientras que si se toman los tres primeros componentes se tiene aproximadamente un 99%. Así que considerando los dos criterios, en este caso nos quedaremos con los primeros componentes principales.

En la siguiente tabla se presentan los pesos, dado que se eligieron los dos primeros componentes:

Variable	CP1	CP2
Seguridad	0.413	0.886
Vivienda	0.962	0.214
Gobierno	0.108	0.185
Salud	0.916	0.222
Obras	0.119	0.984

Para el primer componente principal tomando el valor más alto de 0.962 se tiene que su mitad es de 0.481, así que los valores que son mayores de éste en valor absoluto son los coeficientes de las variables Gasto en Vivienda y Gasto en Salud. Mientras que para el segundo componente principal el valor referencia es de 0.984 y su mitad es de un valor de 0.492, así que las variables con un valor mayor a 0.492 son gasto en Seguridad y Gasto en Obras. Como se observa que la variable Gasto en Gobierno no pesa significativamente, podemos concluir que este gasto es bastante homogéneo, y no contribuye a la distinción entre entidades.

El primer componente principal podría denominarse “Componente ciudadano”, ya que están involucradas las variables del Gasto en la Vivienda y el Gasto en Salud, gastos que se dirigen a la ciudadanía. El segundo componente principal podría denominarse “Componente público”, ya que intervienen los gastos en servicios fundamentalmente orientados a un servicio general. Mientras que en el tercer componente principal intervienen los Gastos del Gobierno, así que este componente principal podría llamarse Componente del Gobierno.

De esta manera, a partir del ACP se han podido reducir el número de variables para este estudio relacionado a los gastos de las entidades federativas, de 5 a 2. Los dos componentes principales que se eligieron son: 1. Componente del Gasto en el Ciudadano; y 2. Componente del Gasto Público.

Además, es de interés la formación de agrupaciones de las 32 entidades federativas para la implementación de programas relacionados al gasto. El gráfico de dispersión de los dos primeros componentes principales permite la formación de grupos; aquí podemos identificar dos grupos que se distinguen del resto de entidades.

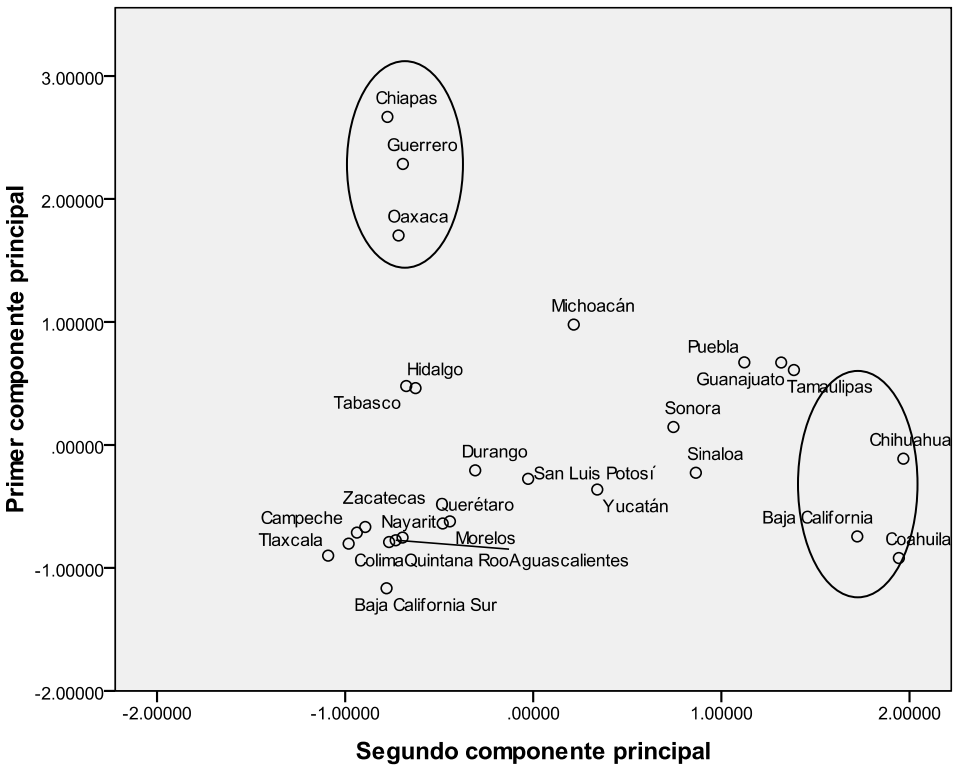


Figura 2.14. Gráfico de dispersión para los componentes principales obtenidos en el análisis.

2.5 Análisis de Correlación Canónica

En ocasiones el investigador necesita conocer si existe relación entre dos conjuntos de variables y de existir poder cuantificar tal relación. El análisis de correlación canónica permite identificar y cuantificar la asociación de tipo lineal entre dos conjuntos de variables con información multidimensional. Es necesario que las variables dentro de cada uno de los grupos sean homogéneas entre sí. El análisis de correlación canónica encuentra subgrupos de variables de un conjunto que están asociadas con subgrupos de variables del otro conjunto. Esta asociación no está dada entre las variables sino a través de combinaciones lineales de las variables de cada uno de los conjuntos; es a lo que llamaríamos una red de relaciones.

El análisis de correlación canónica responde a las preguntas

1. ¿Existe relación entre los conjuntos de las variables? ¿Cuál es la red de asociaciones?
2. ¿Cuántas parejas de variables canónicas significativas existen?

2.5.1. Procedimiento

El punto de partida es dos conjuntos de variables, que se pide que sean al menos en escala ordinal. Se tiene un conjunto de p variables $X_1, X_2, \cdots, X_p$ y el otro con q variables $Y_1, Y_2, \cdots, Y_q$.

A partir de estos conjuntos de variables se forman *k* combinaciones lineales de las *p* variables *X*;

$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ U_2 &= a_{21}X_1 + b_{22}X_2 + \cdots + b_{2p}X_p \\ &\vdots \\ U_k &= a_{k1}Y_1 + a_{k2}X_2 + \cdots + a_{kp}X_p, \end{aligned}$$

así como *k* combinaciones lineales de las *q* variables *Y*;

$$\begin{aligned} V_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\ V_2 &= b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\ &\vdots \\ V_k &= b_{k1}Y_1 + b_{k2}Y_2 + \cdots + b_{kq}Y_q. \end{aligned}$$

Así se forman *k* parejas de combinaciones lineales $(U_1, V_1), (U_2, V_2), \cdots, (U_k, V_k)$. A estas *k* parejas de combinaciones lineales se les denomina variables canónicas; es decir, las variables canónicas son nuevas variables las cuales son combinaciones lineales de las variables originales. El número *k* de parejas de variables canónicas es igual al menor número de variables en cada uno de los grupos, es decir, *k* es el menor entre *p* y *q*.

El criterio que se usa para medir la relación existente entre estas parejas de combinaciones es la correlación de Pearson ρ, que se denomina coeficiente de correlación canónica, aunque es un valor que siempre es positivo. Las correlaciones canónicas al cuadrado se denominan raíces canónicas o autovalores.

La manera en que se forman las combinaciones lineales es buscando que la correlación entre la primera pareja de variables canónicas (U_1, V_1) sea la máxima; es decir, entre todas las combinaciones lineales posibles de las variables *X* y todas las posibles combinaciones lineales de las variables originales *Y*, se eligen como primeras variables canónicas al par de combinaciones lineales las cuales presenten la mayor correlación. Esto se realiza mediante un proceso de optimización numérica. La manera en que se elige la

segunda pareja de variables canónicas (U_2, V_2) es eligiendo aquellas combinaciones lineales restantes que presenten la mayor correlación, sujetas a que están incorrelacionadas con U_1 y V_1; la tercera pareja se forma tal que entre las combinaciones lineales restantes sean las que presentan la mayor correlación, sujetas a que están incorrelacionadas con U_1, V_1, U_2 y V_2; y así sucesivamente.

Así:

$$\begin{aligned} \rho_1 &= \text{corr}(U_1, V_1) \\ \rho_2 &= \text{corr}(U_2, V_2) \\ &\vdots \\ \rho_k &= \text{corr}(U_k, V_k), \end{aligned}$$

teniéndose que $\rho_1 > \rho_2 > \cdots > \rho_k$.

Cabe hacer hincapié que la correlación no es entre las variables originales $X_1, X_2, \cdots, X_p$ y $Y_1, Y_2, \cdots, Y_q$, sino que es la correlación entre combinaciones lineales de estas variables. Las correlaciones canónicas son coeficientes de correlación de Pearson, aunque, como ya se dijo, sólo toman valores positivos. Una manera de interpretar los valores de las correlaciones canónicas es:

$$\begin{aligned} 0.0 \leq p < 0.3 &\text{ baja} \\ 0.3 \leq p < 0.5 &\text{ leve} \\ 0.5 \leq p < 0.7 &\text{ moderada} \\ 0.7 \leq p < 0.9 &\text{ alta} \\ p \geq 0.9 &\text{ muy alta} \end{aligned}$$

Al igual que cualquier investigación que utiliza otras técnicas estadísticas, la práctica más común es analizar las funciones cuyos coeficientes de correlación canónica son estadísticamente significativos para un nivel, normalmente se toma un nivel de significancia de 0.05 o menor, aunque a veces se puede tomar incluso un valor menor que 0.1.

2.5.2. Interpretación de las variables canónicas

La realización de las interpretaciones comprende el examen de las variables canónicas para determinar la importancia relativa de cada una de las variables originales en las relaciones

canónicas. Los coeficientes canónicos a_{ij} y b_{k1} para cada variable indican la importancia de cada variable en la combinación lineal, de manera análoga a los pesos en la técnica de componentes principales.

Las variables con ponderaciones relativamente mayores contribuyen más al valor de la variable canónica y viceversa. Igualmente, las variables cuyas ponderaciones tienen signos contrarios presentan una relación inversa unas de otras y las variables con ponderaciones del mismo signo presentan una relación directa. Una interpretación de esta información permite obtener conclusiones respecto a la intensidad y sentido de las redes de correlación.

2.5.3. Coeficiente de redundancia

El coeficiente de redundancia es la proporción de la varianza promedio de un conjunto de variables que es explicada por la variable canónica del otro conjunto. Así, un coeficiente es la proporción de la varianza promedio del primer conjunto de variables que es explicada por la variable canónica del segundo conjunto, y otro la proporción de la varianza promedio del segundo conjunto de variables que es explicada por la variable canónica del primer conjunto.

Las raíces canónicas, es decir, las correlaciones canónicas al cuadrado, representan la cantidad de la varianza de una variable canónica explicada por la otra variable canónica, lo cual se puede pensar como la cantidad de varianza compartida entre las dos variables canónicas. El índice de redundancia es el equivalente de calcular el coeficiente de correlación múltiple al cuadrado entre un conjunto de variables y cada una de las variables en el otro conjunto, y después promediar estos coeficientes al cuadrado para obtener un R^2 promedio. Por lo que el índice de redundancia es análogo al estadístico R^2 en el caso de la regresión lineal.

Caso práctico: Se tiene información sobre variables relacionadas al Desarrollo Humano (Y_1 = Esperanza de vida; Y_2 = Tasa de alfabetización; y Y_3 = Tasa de matriculación) y las variables del Financiamiento (X_1 = Producto Interno Bruto y X_2 = Ingreso per cápita). El objetivo del estudio es conocer si existe relación entre componentes básicos del desarrollo humano y variables monetarias relacionadas con el financiamiento en salud, en las entidades federativas del país.

Se hace uso del paquete SAS para llevar a cabo el análisis de correlación canónica. La forma en que se meten las observaciones para formar una base de datos en SAS, es:

```
data canoli;
input   Unidad Y1 Y2 Y3 X1 X2;
cards;
77.20   96.97   75.30   22816  0.8913
76.80   96.58   62.31   15837  0.8534
76.30   96.28   65.54   12434  0.8401
75.80   95.12   66.56   12965  0.8355
……
……
76.20   96.03   63.29   11262  0.8329
76.30   95.69   65.62   11040  0.8323
76.40   95.11   65.66   10933  0.8310
76.10   95.50   66.91   10377  0.8285
```

El procedimiento usado en SAS para llevar a cabo un análisis de Correlación Canónica es el CANCORR

```
proc cancorr redundancy;
  var X1 X2
  with Y1 Y2 Y3;
run;
```

Salida

1. Las correlaciones canónicas, sus raíces cuadradas, así como su valor *p*

Correlación canónica	Correlación canónica al cuadrado	Valor *p*
0.850430	0.723232	<0.001
0.348646	0.121554	.097038

La primera correlación canónica es de 0.85, resultando significativa, se tiene que es una correlación alta. La segunda correlación canónica es de 0.35, la cual es baja, pero resultó no significativa. Así que se tiene sólo una pareja de variables canónicas.

2. Las correlaciones entre las variables X y las variables canónicas

	U_1	U_2
X_1	0.537496	0.843266
X_2	-0.833505	0.552512

De lo cual se tiene que en la primera variable canónica U_1, el ingreso per cápita es la variable de mayor peso. Además de que el Producto interno bruto y el Ingreso per cápita presentan una relación inversa.

3. Las correlaciones entre las variables *Y* y sus variables canónicas es

	V_1	V_2
Y_1	0.997820	0.001815
Y_2	0.930213	0.339041
Y_3	0.218030	-0.014672

De lo cual se tiene que en la primera variable canónica V_1 la Esperanza de vida y la Tasa de alfabetización son las variables de mayor peso. Además se tiene que las tres variables relacionadas al desarrollo humano presentan una relación directa. Los índices de redundancia son: *X* con las *V*, 0.3557; y *Y* con las *U*, 0.4601. Los dos índices de redundancia son bajos, implicando que la capacidad de explicar la variación de las variables del grupo del desarrollo humano que tiene la variable canónica de las variables de financiamiento en salud es baja, e igualmente la capacidad de predicción que tiene la variable canónica del grupo de financiamiento en salud para explicar la variación de las variables del grupo de la desarrollo humano.

III. Modelación Estadística

Dentro del campo de las Finanzas públicas, como en muchas otras áreas del conocimiento, se suscitan diversas interrogantes de investigación que pueden ser abordadas a partir de estrategias de modelación estadística. La modelación estadística se puede considerar como un área de estudio y especialización, en la que convergen los aspectos teóricos, metodológicos y computacionales de los modelos estadísticos, considerando éstos en el marco de un proceso que pretende postular, ajustar y evaluar la capacidad y sensibilidad del modelo para describir una relación causa efecto, sobre un conjunto de unidades de estudio (Ojeda, 1993). El objetivo que se pretende alcanzar, a través de la aplicación de los modelos estadísticos, es modelar la realidad; es decir, tener la posibilidad de sustentar las posibles relaciones que existen en el marco del fenómeno que se desea estudiar, obteniendo una herramienta para describir, analizar, predecir y hasta pronosticar tendencias o valores particulares.

Aunque hay diferentes definiciones de modelo, la modelación estadística parte del concepto de modelo matemático, que de manera muy general, se puede concebir como una abstracción en la que se aspira a estudiar y entender desde otra perspectiva un fenómeno en el que subyace una relación causa-efecto entre dos o más variables**:** una dependiente, a la que se denominará *Y,* y otra independiente, que se denotará como *X*; la relación que se estudia es del tipo *X* implica *Y*. En este sentido, el modelo expresa un postulado acerca de esta relación que se expresa a través de una formulación matemática. En un modelo matemático se deben identificar una o varias variables dependientes y una o varias variables independientes. Ahora bien, es posible utilizar diferentes tipos de funciones matemáticas para proponer un modelo que representa a una variable respuesta como función de una o más variables independientes; estos tipos de funciones matemáticas las podemos identificar como modelos determinísticos. En general, se denota por un modelo determinístico a la relación funcional

$$y = f(x),$$

donde y es un escalar o vector y x es también un escalar o vector. En el contexto de la estadística se les llama modelos determinísticos pues establecen una relación funcional exacta; es decir, dados los valores de x es posible obtener exactamente los valores de y.

Estos modelos puedes ser lineales o no lineales. Sin embargo, como es bien sabido, la estadística estudia fenómenos aleatorios a partir de la teoría de la probabilidad, la cual permite manejar la incertidumbre, un elemento presente en todas las esferas de la vida. Así, en estadística se trabaja con lo que se llama variables aleatorias, las cuales son cantidades que se describen por una distribución de probabilidades. El concepto clave para definir y entender muchos procesos de inferencia estadística es el de modelo estadístico, el cual para una observación *i*-ésima (*i*=1,2,3.,…,*n*), se denota su valor en Y, agregándole el subíndice *i*, (y_i), y su valor en x de igual forma (x_i). Esta observación se concibe en dos componentes, que se denominan genéricamente la parte sistemática ($f(x)$) y la parte aleatoria (ε); por lo tanto, el modelo estadístico se formaliza con una ecuación del tipo:

$$y_i = f(x_i) + \varepsilon_i \; ; \qquad i = 1,2, \dots, n$$

que representa las observaciones sobre las *n* unidades de estudio, donde la parte sistemática explica la respuesta a partir de las condiciones x_i. Como ε_i es no observable, esta cantidad se asume como desconocida. El problema es, en consecuencia, doble: El investigador no sabe exactamente cómo una variable o un conjunto de p variables dependientes $(x_1, x_2, \cdots, x_p)$ afectan a la variable independiente *Y*, por lo tanto no conoce $f(x_1, x_2, \cdots, x_p)$; bueno, no exactamente, porque su conocimiento teórico de los procesos biológicos, físicos, y en general de la realidad, le permiten plantear modelos determinísticos como candidatos a modelar la relación bajo estudio. Suponiendo que el investigador tuviese un modelo propuesto $y = f(x_1, x_2, \cdots, x_p) + \varepsilon$, no sabe cuáles son los parámetros que se corresponden con las mediciones específicas que tiene de cada unidad; así el problema de modelación se plantea a partir de una matriz de datos como se aprecia en la Tabla 3.1:

Tabla 3.1. Datos para modelar.

Unidad	$X_1, X_2, \cdots, X_p$	Y
1	$x_{11}\ x_{12} \dots x_{1p}$	y_1
2	$x_{21}\ x_{22} \dots x_{2p}$	y_2
⋮	⋮	⋮
n	$x_{n1}\ x_{n2} \dots x_{np}$	y_n

Por lo tanto, se busca que el modelo permita conocer una estimación del valor de los parámetros que mejor describa los datos, así como tener un referente de la variabilidad que existe en los mismos, y así mismo lograr explicar la mayor variabilidad que los datos expresan.

3.1 ¿Qué es modelar estadísticamente?

Aunque la respuesta más lógica a la pregunta sobre qué es modelar estadísticamente, pareciera que es definir un modelo para explicar la relación que existe entre las variables, está respuesta sería incorrecta, porque modelar estadísticamente es llevar a cabo un proceso que no sólo está relacionado con la especificación del modelo matemático y la interpretación de los resultados una vez que este modelo ha sido ajustado. Así, entonces modelar estadísticamente es, primeramente construir un marco teórico que sustente la definición de la relación entre las variables implicadas en el fenómeno de estudio; así como sustentar la relación causal que existe entre ellas; posteriormente se selecciona el conjunto de unidades de estudio sobre las cuales se realizarán los respectivos análisis. Se realiza la obtención de los datos, para posteriormente hacer un análisis exploratorio de los mismos y poder postular un modelo razonable en función de los objetivos y la relación entre las variables; en consecuencia con los procedimientos de estimación del modelo, se procede a ajustar ese modelo a los datos, para hacer un diagnóstico y evaluar si los resultados se corresponden con lo esperado y finalmente realizar la interpretación del modelo. En resumen la modelación estadística aplicada a un proceso de investigación implica los pasos descritos anteriormente. Una vez que se ha planteado la pregunta de investigación, se revisan los objetivos del proyecto y se traducen a objetivos estadísticos. A partir de este momento se inicia con la recolección de los datos, se debe tener en cuenta si son suficientes para el análisis, cuál es su estructura, qué tipos de variables se midieron a las unidades de estudio, entre otras. Posteriormente, antes de especificar la relación causal entre las variables, se recomienda hacer un análisis exploratorio[1] para observar el comportamiento de los datos e identificar el modelo que sea el más adecuado; este proceso se debe de realizar de una forma iterativa e interactiva.

[1] El análisis exploratorio implica la elaboración de gráficos, estadísticas descriptivas, análisis comparativos, análisis bivariados, multivariados e inferencia informal, esto último puede implicar a su vez, el ajuste de modelos para identificar tendencias o asociaciones.

Una vez que se ha postulado el modelo, siempre ante la evidencia mostrada por los datos en el análisis exploratorio, se realiza el ajuste. El proceso de ajuste del modelo puede llevar a la identificación de datos atípicos, los cuales podrían estar asociados a puntos influyentes en el ajuste del modelo; es decir, pueden ser observaciones que de eliminarse cambiarían significativamente el modelo ajustado.

Habiendo realizado el diagnóstico del modelo ajustado y tomada la decisión de que es efectivamente un modelo adecuado, viene la etapa de interpretación y uso del modelo. Muchas veces el modelo se usa simplemente para describir el fenómeno bajo estudio de manera sintetizada. Como en realidad el modelo ajustado es el patrón de comportamiento de la relación entre dos o más variables, descrita a través de los datos, éste reemplaza a los datos y puede ser más fácil de interpretar. Sin embargo, a veces el modelo ajustado se usará para hacer predicciones de valores esperados de Y para algún valor específico x de X; en este caso se recomiendan precauciones si el valor x está fuera del rango de valores en el modelo ajustado; es decir, hay que tener cuidado especial si se va a realizar una extrapolación. Finalmente, la verificación de la adecuación del modelo implica la revisión de algunos indicadores involucrados en el proceso de estimación y prueba de hipótesis. En resumen, la modelación estadística es conducir un proceso de obtención de conocimiento.

3.1.1 Retos del modelador

Actualmente, la modelación estadística cuenta con los respaldos tecnológicos y metodológicos que le dan una gran viabilidad como un área de desarrollo de la estadística aplicada y ha permitido que sea utilizada en distintas áreas del conocimiento. Sin embargo, no basta con hacer uso de la tecnología y aplicar la teoría. Los principales retos del modelador se centran en que se debe tener la capacidad para postular un modelo razonable; es decir, un modelo adecuado a los objetivos que se plantean y a la naturaleza de los datos recabados; verificar el cumplimiento de los supuestos con suficiente sustento, para tener la certeza de que las estimaciones del modelo no son insesgadas; y finalmente, interpretar y usar adecuadamente el modelo ajustado, lo que significa que se deben obtener conclusiones útiles que aporten un conocimiento relevante al fenómeno bajo estudio. De lo contrario, como señalan Skrondall y Rabe-Hesketh (2004), se puede abusar de cualquier método estadístico, si se hacen especificaciones o interpretaciones inadecuadas del modelo.

3.1.2 ¿Para qué sirve un modelo?

Lo más importante al construir o desarrollar un modelo estadístico es el uso que se le dará. Comúnmente, una inquietud, un problema de investigación o una necesidad específica de mejoramiento dan origen a la realización de un estudio, en cuyo contexto se debe ajustar y evaluar un modelo. El investigador debe tener claro cuáles son los principios y limitaciones bajo las que podría utilizar el modelo ajustado. Estas limitaciones frecuentemente están relacionadas con el uso final que se le dará al modelo. Los modelos estadísticos se usan para varios propósitos, incluyendo los siguientes: la descripción de datos, la estimación de parámetros, predicción y calibración.

Muchos investigadores y técnicos frecuentemente usan las ecuaciones para resumir o describir un conjunto de datos. A veces, problemas de determinación de los niveles óptimos de un proceso pueden ser resueltos por modelos estadísticos. Muchas aplicaciones de modelación involucran la estimación de la variable respuesta para valores no observados en las variables explicativas, lo que puede usarse también con propósitos de control de procesos, así como para probar hipótesis, comparar grupos y realizar calibraciones. Sin embargo, si ya se ha especificado un modelo y lo que se busca es la estimación de parámetros, las preguntas entonces son ¿cuáles son los valores de los parámetros que mejor describen a los datos?, y ¿qué porcentaje de la explicación de la variabilidad se logra con el modelo? y ¿cuál es el porcentaje de la variabilidad no explicada? Para dar respuestas a estas interrogantes, se pueden utilizar diversos modelos estadísticos. En las siguientes secciones se introducirá al lector en los modelos de regresión lineal simple y múltiple y en los modelos multinivel, todos estos ejemplos de modelos estadísticos que gozan de una amplia popularidad.

3.2 Modelos de Regresión

El objetivo de los modelos estadísticos es encontrar qué relación existe entre una o un conjunto de variables independientes ($X_1, X_2, X_3 \ldots X_p$) y una o un conjunto de variables dependientes ($Y_1, Y_2, Y_3 \ldots Y_k$) . Si hablamos del caso más simple: una X y una Y, primero se analizan las dos características de estudio para las unidades de la muestra, de tal manera que se tienen un par de observaciones para cada unidad (x_i, y_i) (i=1,2,...,n).

Posteriormente, se representan dichos valores en un plano cartesiano, formándose un diagrama de dispersión o nube de puntos. Así, cada unidad está representada por un punto en el gráfico de coordenadas (x_i, y_i). De esta manera, se obtiene una primera aproximación de la relación entre las variables. El tipo de relación que se puede observar se presenta en la Figura 3.1.

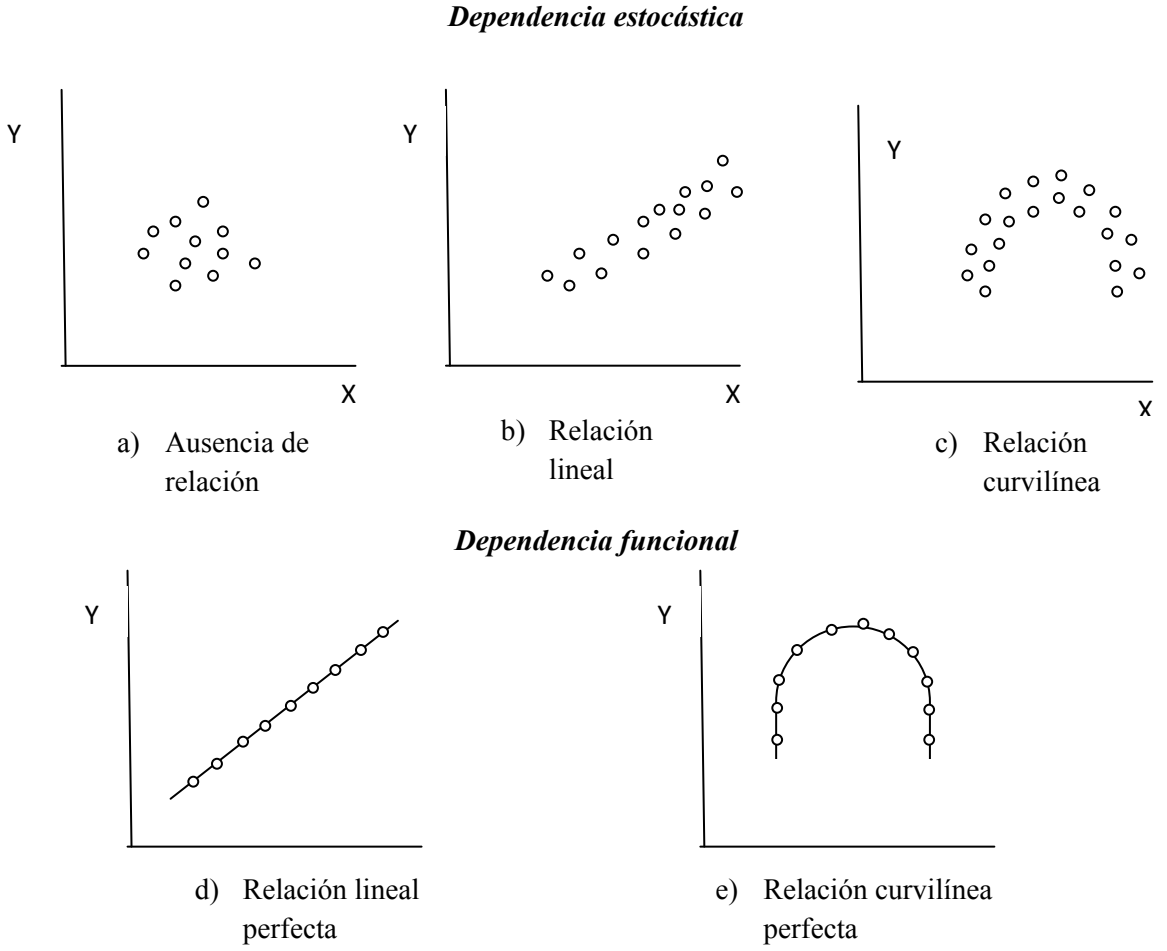


Figura 3.1. Tipos de relación entre dos variables *X* y *Y*.

No siempre se obtienen relaciones perfectas, como en el caso de los ejemplos d) y e), en el que los puntos del diagrama de dispersión correspondiente aparecen sobre la función $y = f(x)$, sino otro tipo de dependencia o relación menos rigurosa conocida como dependencia estocástica, como se aprecia en los casos b) y c), por lo que, la relación entre X y Y, se puede expresar para el ejemplo b), de la forma $y = a + bx + \varepsilon$ donde ε es un error que refleja precisamente la diferencia entre los puntos con respecto al valor sobre la recta de regresión verdadera. Por esta razón, es un valor no observable, porque la recta de

regresión verdadera es precisamente desconocida. En el caso del ejemplo a) se muestra la ausencia de relación.

Cuando se presenta una relación de dependencia estocástica, se distinguen dos tipos de técnicas: El Análisis de correlación y el Análisis de regresión. En este capítulo, se estudiarán únicamente los modelos de regresión, pero de manera general diremos que el Análisis de correlación tiene como finalidad estudiar la existencia de dependencia estocástica entre las variables, y cuál es el grado de tal dependencia. Por su parte, el Análisis de regresión responde a interrogantes como cuál es el tipo de dependencia entre dos variables y si pueden estimarse los valores de Y, a partir de los de X y con qué tanta precisión. A continuación se presentan las características, fundamentos, propiedades y tipos de los modelos de regresión.

3.2.1 Modelos de regresión lineal

El Análisis de regresión es una de las técnicas de uso más frecuente para el tratamiento de datos multifactoriales, que ha tenido diversas aplicaciones y que es utilizada en distintos campos del conocimiento, incluyendo la Ingeniería, las Ciencias físicas, las Ciencias biológicas, las Ciencias sociales y la Economía, entre otras. Esta expansión de las aplicaciones se ha debido principalmente al respaldo teórico, metodológico y computacional con el que se cuenta para su uso y, también, porque los modelos de regresión son una buena aproximación a relaciones funcionales más complejas.

De manera general, se dice que existe regresión de los valores de una variable con respecto a otra, cuando hay alguna línea, llamada línea de regresión, que se ajusta en cierto grado a la nube de puntos. En una definición moderna del Análisis de regresión, en términos generales se puede decir que se trata de la relación de dependencia de una variable (variable dependiente o respuesta, Y), con respecto a una o más variables (variables independientes o explicativas, X), con el objetivo de estimar o predecir la media o valor promedio de la variable dependiente con base en los valores conocidos o fijados (en muestras repetidas) de las variables independientes. En un Análisis de regresión primordialmente se busca estimar los parámetros desconocidos, pero también comprobar la adecuación del modelo y la calidad del ajuste logrado. Es decir, se trata de un proceso

iterativo, en el que los datos conducen a un modelo y se produce un ajuste del modelo a los datos, por la vía de la estimación de los parámetros del modelo.

Estructura del modelo: La ecuación que representa la línea recta es: $y = \beta_0 + \beta_1 x$ en un análisis de regresión, se busca ajustar una línea recta a la relación que existe entre dos variables, donde β_1 representa la pendiente de la línea, y β_0 es el intercepto, o el punto en el cual la recta corta el eje de las Y; en otras palabras, el valor de y cuando $x = 0$. La pendiente se interpreta como el cambio en la media de y por un cambio unitario en x, y el intercepto como la altura de la línea de regresión para un valor dado de x. En estadística β_0 y β_1, son llamados coeficientes. El coeficiente de una variable es una cantidad que lo multiplica, por lo que en este caso, β_1 es el coeficiente de la variable explicativa x, y β_0 es el coeficiente de una variable que equivale a 1 para cada observación (usualmente denominada la "constante").

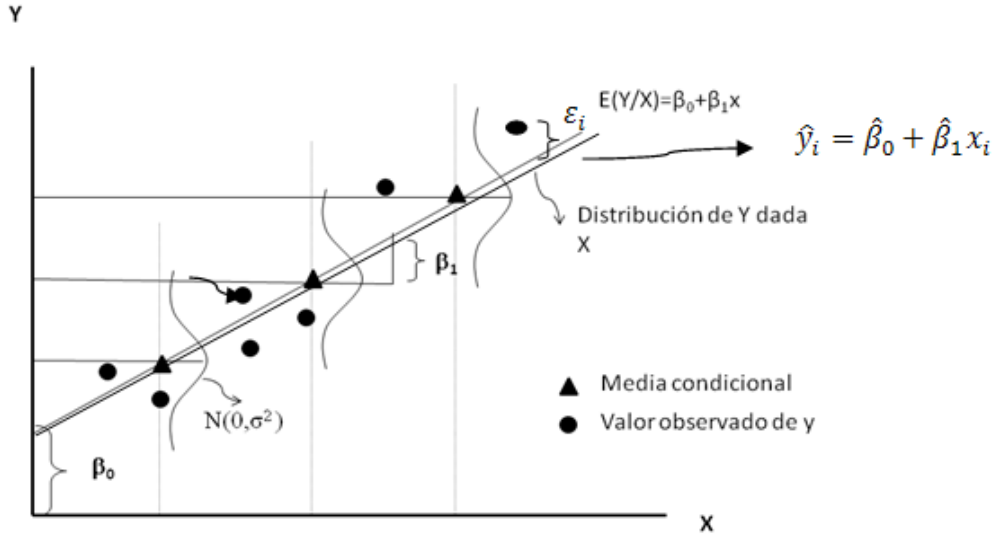


Figura 3.2. El modelo de regresión lineal simple y los datos observados con la recta ajustada.

No obstante, como se está trabajando con datos reales, y como se aprecia en la Figura 3.2, éstos valores observados no caen exactamente sobre una recta, por lo que se debe modificar la ecuación y tomar en cuenta los residuos, diferencia entre el valor observado de y, con respecto a la línea recta $(\beta_0 + \beta_1 x)$, lo que se conoce como "error", y se representa con la letra ε. El error se puede definir como la contribución a Y de todos aquellos factores que no se pueden observar durante la obtención de la información o recolección de los datos. Estos errores se originan por el uso de inadecuados instrumentos de medición en la obtención de los datos, que también pueden ocurrir en el momento del

registro de datos, o bien, representar el efecto de otras variables que no están incluidas en el modelo. Por lo tanto, para explicar esas fluctuaciones es considerado el término ε en el modelo. Más adelante se mostrarán los supuestos del modelo que se corroboran a través del análisis de los residuos, los cuales son denotados por $e = y_i - \widehat{y_i}$, donde y_i es el valor estimado una vez que se ajusta el modelo; es decir, una vez obtenidos los estimados β_0 y β_1 , y sustituidos para el valor de x_i .

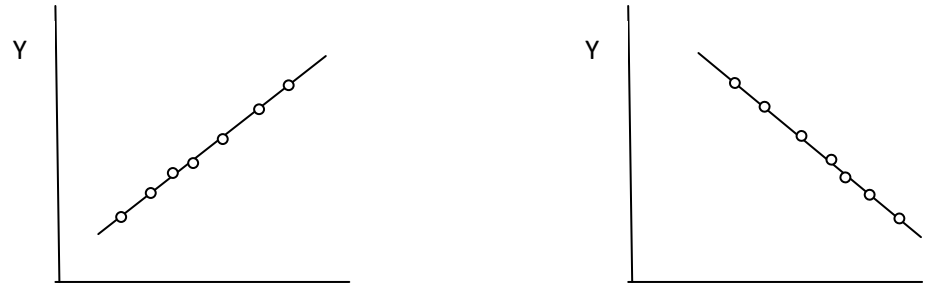
$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.0}$$

Para una observación *i*-ésima (*i*=1,2,3,…,*n*), se denota su valor en *Y*, agregándole el subíndice *i*, (y_i), y su valor en x de igual forma (x_i) Entonces para una observación i, la relación lineal entre Y y X, se expresa como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \cdots, n \tag{3.1}$$

3.2.2 Modelo de regresión lineal simple

La ecuación 3.1 representa un modelo de regresión lineal; en este caso que sólo se tiene una variable explicativa, se conoce como modelo de regresión simple, el cual se establece para un conjunto de *n* observaciones en (*X*,*Y*); es decir, para los datos $(x_1, y_1), (x_2 y_2), \ldots, (x_n, y_n)$. En la recta de regresión, la relación puede ser directa o inversa como se aprecia en la Figura 3.3, si el coeficiente $\beta > 0$, entonces cuando X aumenta, Y también lo hace (relación directa); y si $\beta < 0$, entonces cuando X aumenta, Y disminuye (relación inversa).



a) $\beta > 0$ b) $\beta < 0$

Figura 3.3. Signo de la pendiente en una recta de regresión.

Ahora bien, la estimación de los parámetros de regresión (β_0 y β_1), puede realizarse a través de diferentes métodos. Sin embargo, el más utilizado es el método de Mínimos cuadrados ordinarios (MCO), en el cual la suma de los residuos al cuadrado es minimizada. Es decir, se estima β_0 y β_1 , tales que la suma de los cuadrados de las diferencias entre los valores de y_i y la línea recta sea mínima. Entonces el criterio puede expresarse como la minimización de:

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

donde *D* equivale a la suma de los residuos al cuadrado, que en el método de MCO, busca que sea la mínima, y y_i son los valores estimados según el modelo $y = \beta_0 + \beta_1 x$; es decir, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, donde $\hat{\beta}_0$ y $\hat{\beta}_1$ serían los estimados. En la actualidad la mayoría de los paquetes de programas estadísticos ofrecen opciones para realizar estimaciones y evaluaciones de los modelos de regresión lineal, por lo que no se ahondará en la notación matemática (Véase Montgomery, *et al*. (2004)).

Precisando, entonces, al aplicar el método de mínimos cuadrados a los datos recabados, se obtiene una estimación de los valores de los parámetros de la población. Estas estimaciones se denotan por $\hat{\beta}_0$ y $\hat{\beta}_1$; con lo que

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (3.2)$$
$$\text{y } e_i = (y_i - \hat{y}_i).$$

La ecuación 3.2, representa la línea de regresión ajustada. El valor esperado de $\hat{y}_i$, es el punto de la línea que corresponde a x_i. Esto se puede ver en la Figura 3.2.

Prueba de hipótesis: Hasta ahora se ha visto que uno de los objetivos del análisis de regresión es conocer si existe una relación lineal entre las variables bajo estudio a través del análisis de los datos recabados; esto implica estimar el valor de los parámetros que se desconocen de la población, que se representa por la recta de regresión verdadera. En otras palabras, se busca verificar si la relación entre las variables es estadísticamente significativa. Para ello, se recurre a realizar una prueba de hipótesis.

La prueba de hipótesis comienza con una suposición, denominada hipótesis, que se hace en torno a un (o varios) parámetro del modelo estadístico que se asume describe a la

población. Posteriormente se reúnen los datos, se calculan las estadísticas y en base a estos valores, con cierto grado de probabilidad, se dice si la estimación del parámetro supuesto de la población es razonablemente aproximada a su verdadero valor. La diferencia entre el parámetro supuesto de la población y el estadístico calculado a partir de los datos no suele ser ni tan grande que automáticamente se rechace la hipótesis, ni tan pequeña que de inmediato se acepte. Por esta razón se requiere realizar el procedimiento de prueba de hipótesis, lo que permite tomar decisiones sustentadas con los datos. Formalmente en una prueba de hipótesis se tienen los siguientes cuatro elementos:

- *Hipótesis nula* (H_0): Esta es la hipótesis a probar, generalmente es una aseveración en el sentido de que un parámetro tiene un valor específico.
- *Hipótesis alterna*tiva (H_1): Esta hipótesis, sobre la cual se enfoca la atención, es una aseveración sobre el mismo parámetro poblacional que se utiliza en la hipótesis nula. Generalmente se especifica que el parámetro poblacional tiene un valor diferente al establecido en la hipótesis nula.
- *Estadístico de prueba*: Es una función de las mediciones o datos sobre la cual se fundamenta la decisión "no se rechaza H_0", o bien "se rechaza H_0".
- *Región de rechazo*: especifica los valores del estadístico de prueba para los cuales se rechaza la hipótesis nula.

Además de los elementos anteriores, para llevar a cabo una prueba de hipótesis se deben definir el error estándar (ES) y los intervalos de confianza. El primero se refiere a la desviación estándar de la distribución del estadístico y constituye una medida de variación, para la que si se obtienen grandes valores indica una mayor incertidumbre sobre el verdadero valor poblacional. Los intervalos de confianza son un rango de valores que contendrán el valor verdadero del parámetro poblacional, con una probabilidad asignada. Así, se tiene que el intervalo será el parámetro estimado $\widehat{\beta_1} \pm$ una medida de variación.

Para aplicar los conceptos anteriores supóngase lo siguiente, se desea probar que no existe relación entre las variables bajo estudio. La hipótesis nula para esta prueba está dada por:

$$H_0: \beta_1 = 0$$

y la hipótesis alterna, será entonces que sí existe relación:

$$H_1: \beta_1 \neq 0$$

La prueba que se planteó es de tipo bilateral, pues no interesa determinar si el valor del parámetro es mayor o menor que un valor determinado; simplemente se plantea que sea diferente de cero. Es decir, que exista relación. El estadístico de prueba se calcula de la siguiente manera, se calcula de la siguiente manera:

$$\frac{\hat{\beta}_1}{ES(\hat{\beta}_1)}$$

donde $ES(\hat{\beta}_1)$ es el error estándar de $\hat{\beta}_1$; que tiene la siguiente expresión:

$$ES(\beta_1) = \hat{\sigma}/\sqrt{s_{XX}},$$

con

$$\hat{\sigma}^2 = \frac{SSE}{(n-2)} = D/(n-2).$$

Es decir, se calcula con la varianza estimada y la suma corregida de cuadrados.

$$S_{XX} = (X_i - \bar{X})^2.$$

El estadístico de prueba se compara con un valor de tabas de la distribución Normal Estándar o de una distribución *t* de Student (si el tamaño de muestra es pequeño). Para rechazar la hipótesis nula de que no hay relación entre las variables, se sigue el criterio que se muestra en la Figura 3.4; como se trata de una prueba bilateral, el valor del estadístico de prueba debe ser mayor que un valor crítico $c_{\alpha/2}$ o menor que - $c_{\alpha/2}$. En la Figura 3.4, se muestran los distintos tipos de prueba y su zona de rechazo y no rechazo.

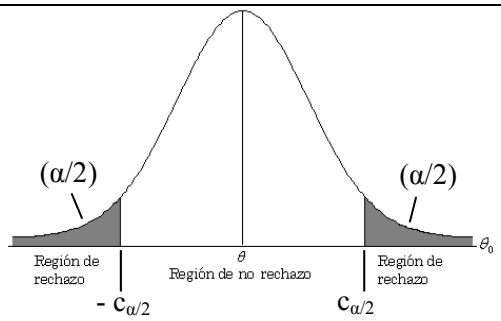
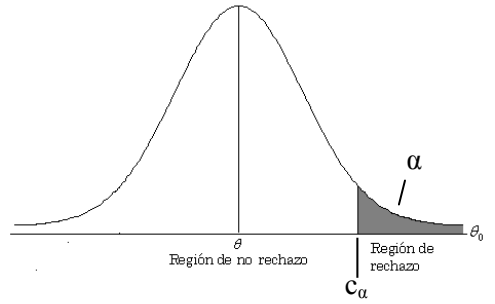
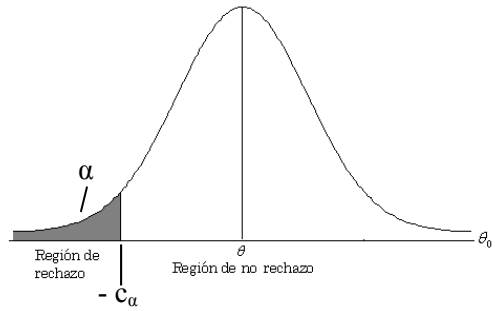
TIPO DE PRUEBA	REPRESENTACIÓN GRÁFICA	CRITERIO DE RECHAZO O NO RECHAZO
a) Prueba bilateral $H_1 : \beta_1 \neq 0$		Se rechaza H_0, si $\left\|\frac{\hat{\beta}_1}{ES(\hat{\beta}_1)}\right\| > c_{\alpha/2}$ o $\left\|\frac{\hat{\beta}_1}{ES(\hat{\beta}_1)}\right\| < -c_{\alpha/2}$
b) Prueba de cola derecha $H_1 : \beta_1 > 0$		Se rechaza H_0, si $\left\|\frac{\hat{\beta}_1}{ES(\hat{\beta}_1)}\right\| > c_\alpha$ El área de rechazo es (c_α,∞), El área de no rechazo $(-\infty , c_\alpha)$
c) Prueba de cola izquierda $H_1 : \beta_1 < 0$		Se rechaza H_0, si $\left\|\frac{\hat{\beta}_1}{ES(\hat{\beta}_1)}\right\| < -c_\alpha$

Figura 3.4. Pruebas bilaterales y unilaterales para el coeficiente de regresión.

No obstante, el criterio de rechazo no es perfecto y el estadístico de prueba que se utiliza es asumido como una variable aleatoria, en este caso con distribución normal. El mismo procedimiento realizado puede conducir a diferentes conclusiones para distintas colecciones de datos (decimos para diferentes muestras). En general no se tiene la certeza de que la decisión tomada sea la correcta, de tal manera que es posible que se esté

cometiendo algún tipo de error. Los errores que pueden presentarse, se conocen como Error tipo I y Error tipo II.

- *Error tipo I*: se comete cuando se rechaza H_0 siendo verdadera. La probabilidad de un error tipo I se denota por α, conocido también como nivel de significancia.

$$\alpha = P(\text{Error tipo } I) = P(\text{rechazar } H_0 / H_0 \text{ verdadera})$$

- *Error tipo II*: Se comete si no se rechaza H_0 cuando es verdadera H_1. La probabilidad de cometer un error tipo II se denota por β.

$$\beta = P(\text{Error tipo } II) = P(\text{no rechazar } H_0 / H_0 \text{ falsa})$$

La probabilidad de un Error tipo I, α, suele denominarse nivel de significancia asociado con una prueba; este término se originó de la manera siguiente: la probabilidad del valor observado del estadístico de prueba, o de algún valor que se contraponga aún más a la hipótesis nula, mide, en cierta manera, el peso de la evidencia a favor del rechazo de la hipótesis nula. A pesar de que se recomiendan valores pequeños para α, la selección de α para aplicarlo en un análisis es arbitraria. Un investigador podría escoger hacer una prueba con α = 0.05, mientras que otro podría preferir α = 0.01. Por lo tanto es posible que al analizar los mismos datos, una persona concluiría que se tendría que rechazar la hipótesis nula a un nivel de significancia α = 0.05, mientras que otra persona decide que no se puede rechazar la hipótesis nula con α = 0.01. Además, se utilizan muchas veces valores de α de 0.05 o bien 0.01 por costumbre, y no por considerar de manera cuidadosa las consecuencias de cometer un Error tipo I. Hoy en día, con la disponibilidad de los paquetes computacionales se busca que p (el valor estimado del nivel de significancia regularmente llamado *p-value*) resulte un valor más pequeño que 0.1. Un criterio heurístico es el siguiente:

1) Rechazar H_0: con alta significancia si $p \leq 0.01$;
2) Rechazar H_0: con significancia moderada si $0.01 < p \leq 0.005$;
3) Rechazar H_0: con baja significancia si $0.05 < p \leq 0.1$; y
4) No rechazar H_0: si $p > 0.1$

La potencia de la prueba, se refiere a la probabilidad de rechazar correctamente la hipótesis nula, cuando ésta es falsa.

$$potencia = 1 - \beta = 1 - \Pr\,(\text{Error tipo } II)$$

Finalmente el valor p (valor de probabilidad o p-value) de una prueba, es el nivel de significancia $\hat{x}$ de la prueba. Esta cantidad es un estadístico que representa el mínimo valor de α para el cual los datos observados indican que se tendría que rechazar la hipótesis nula, dados los datos para un nivel de significancia dado, se busca que el valor de β, sea tan pequeño como sea posible. Este valor, permitirá decidir si se rechaza o no se rechaza la hipótesis nula. Si la selección de α en un experimento es mayor o igual al valor p, se rechaza H_0. De otra manera, si α es menor que el valor p, no se puede rechazar H_0.

La aproximación del valor p como ayuda en la toma de decisiones es bastante usual debido a que casi en todos los paquetes computacionales que proporcionan el cálculo de la prueba de hipótesis imprimen el valor p. Así entonces, se puede decir que la decisión de rechazar o no una hipótesis se puede basar en la observación de p, utilizando la regla heurística que se presentó anteriormente.

Supuestos del modelo: Para que el método de estimación MCO, garantice que $\hat{\beta}_0$ y $\hat{\beta}_1$ sean estimadores insesgados de los parámetros β_0 y β_1, es necesario que se cumplan lo siguientes supuestos sobre los residuos e_i, que se conocen también como las condiciones de Gauss-Markov.

1. Los errores o residuos siguen de manera independiente unos de otros, una distribución de probabilidad normal con media cero y varianza constante; esto es que. $\varepsilon_i \sim N(0,\sigma^2)$ para cada *i*.
2. La varianza de los errores es constante, cualquiera que sea el valor de *X*. Esto significa que si tomamos un segmento de un diagrama de dispersión de *Y* con respecto a *X*, en cualquier valor de *X*, los valores de *Y* deberán tener la misma variación con respecto a cualquier otro valor de *X*. Si la varianza es constante, se dice que los errores son homocedásticos, el caso contrario se conoce como heterocedasticidad; $Var(\varepsilon_i) = \sigma^2, \ \forall i$.
3. Los errores no están autocorrelacionados; es decir, son independientes. La correlación entre los errores puede darse cuando las observaciones se repiten o si

las unidades están de alguna manera agrupadas (por ejemplo, estudiantes dentro de escuelas). Si se detecta que los errores están correlacionados, el modelo de regresión necesita ser modificado y en caso de anidamiento en los datos, se puede recurrir al modelo multinivel, que se describe en el capítulo posterior. En resumen se escribe que $\varepsilon_i \sim NI\ (0, \sigma^2)\ \forall i$, con $\mathrm{E}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$.

El análisis e inspección de estos supuestos se realiza a través de distintos procedimientos, principalmente a través del uso de gráficos. Pero también existen pruebas formales que se aplican sobre los residuos e_i. Para analizar gráficamente los residuos se han propuesto diagramas de dispersión contra los valores predichos o contra los valores de alguna variable *X*. Dado que entre estos y tales variables no debería existir asociación alguna, es decir; los residuos, deben distribuirse homogéneamente alrededor del hiperplano de regresión, no deben variar de forma sistemática y la varianza ha de ser constante, cualquier patrón diferente de uno aleatorio sería indicativo del incumplimiento de los supuestos, y por tanto, se corre el riesgo de obtener estimaciones que no sean insesgadas. Adelante se explicará un poco más sobre este tema.

El coeficiente de determinación y la Varianza explicada y no explicada. Los modelos estadísticos presentan una estructura común y están formados por una parte fija y una parte aleatoria. En el caso de la regresión simple:

$$y = \underbrace{\beta_0 + \beta_1 x_i +}_{\text{Parte fija}} \underbrace{\ \varepsilon}_{\text{Parte aleatoria}}$$

La parte fija representa la pendiente y el intercepto de la línea recta que define la relación, mientras que la parte aleatoria engloba aquéllos factores que no son controlables por el modelador; es decir, la parte de la variabilidad no explicada (varianza no explicada). Por otro lado, la variabilidad en *Y* que es explicada por *X* se denomina varianza explicada.
Un estadístico que se utiliza para evaluar la adecuación del modelo y que se relaciona con la variabilidad, es el Coeficiente de Determinación, que se denota por R2. Esta medida se interpreta como la proporción de la varianza total en *Y* que puede ser explicada por la variabilidad en *X*. También se interpreta como el valor que indica qué tanto se corresponden los datos ajustados con los datos reales; es decir, es una medida de la capacidad de bondad de ajuste del modelo. Este coeficiente R2 toma valores entre 0 y 1,

cuando es 0, quiere decir que no hay relación, contrario al valor 1, que indica una relación perfecta. Un modelo de regresión estimado cuyo coeficiente de determinación se aproxima a 1, significa que está bien especificado, pues la mayor parte de la variabilidad de Y se explica por ese modelo. En el caso del modelo de regresión simple, R2 equivale al cuadrado del Coeficiente de correlación de Pearson.

3.2.3 Modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple es una extensión natural de la regresión lineal simple, al caso en el que se tiene más de una variable explicatoria; es decir, el modelo es postulado considerando que a la respuesta Y contribuyen p variables explicatorias $X_1, X_2, \cdots, X_p$. Así si se tienen datos:

$$\begin{gathered}(y_1, x_{11}, x_{12}, \cdots, x_{1p}) \\ (y_2, x_{21}, x_{22}, \cdots, x_{2p}) \\ \vdots \\ (y_n, x_{n1}, x_{n2}, \cdots, x_{np})\end{gathered}$$

Entonces el modelo propuesto es:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i; \;\; i = 1,2, \cdots, n, \tag{3.3}$$

donde β_0, es el valor esperado de Y, cuando $x_1 = 0$, $x_2 = 0$, … , $x_p = 0$. La interpretación de los coeficientes β_1, β_2, …, β_p, varía respecto al modelo de regresión lineal simple y se realiza de la siguiente manera: β_1 es el coeficiente de X_1, el cual se interpreta como el cambio en y por un cambio unitario en x, manteniendo las demás variables constantes. Igualmente, β_2 es el coeficiente de X_2 y se interpreta como el cambio en Y por un cambio unitario en X_2, manteniendo el resto constante, y así de la misma manera para β_p.

El modelo de regresión lineal múltiple estipulado en la ecuación (3.3), es llamado lineal por la linealidad sobre los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ (exponente 1 en todos ellos). La expresión en notación matricial del modelo, queda de la siguiente forma:

$$\boldsymbol{Y = X\beta + \varepsilon.}$$

En donde:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \cdots & x_{1p} \\ 1 & x_{21} & x_{22} \cdots & x_{2p} \\ \vdots & \vdots & \vdots \ddots & \vdots \\ 1 & x_{n1} & x_{n2} \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\boldsymbol{X}$ es una matriz de $n \times (p+1)$ con $(p+1) \leq n$.

En el modelo anterior el componente aleatorio ε se asume una variable aleatoria n-variada distribuida normalmente con media cero y varianza σ^2; es decir, $\varepsilon \sim N_n(0, \sigma^2 I_n)$, donde I_n es la matriz identidad de orden *n.* Para la postulación del modelo, se supone que los ensayos o casos son independientes, con la misma distribución; es decir, se supone que $\varepsilon_i \sim NI(0, \sigma^2),\ i = 1,2, \ldots, n$, que tiene la explicación equivalente a la que se dio para la regresión lineal simple.

La estimación de los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, al igual que en la regresión lineal simple se realiza mediante el método mínimos cuadrados ordinarios; los detalles teóricos y deducciones matemáticas pueden verse en Montgomery *et al*. (2004).

Prueba de hipótesis: El planteamiento de la prueba de hipótesis es similar al descrito en el modelo de regresión simple, sólo que la hipótesis general bajo la que se construye el modelo de regresión lineal múltiple, es que las variables $\boldsymbol{X_1, X_2, \cdots, X_p}$, contribuyen significativamente de manera conjunta para explicar *Y*. A ésta se le llama la hipótesis de la regresión, que implica el rechazo de la hipótesis nula (H_0), a favor de la alternativa (H_1):

$$H_0 : \beta_j = 0 \quad \text{para} \quad j = 1,2,..., p$$

$$\text{v s}$$

$$H_1 : \beta_j \neq 0 \quad \text{para al menos una } j.$$

Esto quiere decir que la regresión se declarará significativa si al menos una de las variables $X_1, X_2, \ldots X_p$, está contribuyendo a la explicación de la variable *Y*; para probar esta hipótesis se construye la tabla de análisis de la varianza, que toma la forma general mostrada en la Tabla 3.2, en la que se calcula la suma de cuadrados de los residuos (SSE) y la suma de cuadrados explicada por el modelo (SSR). En este caso se utiliza el estadístico F_c, que sirve para constatar la hipótesis de adecuación del modelo. Si el valor de F_c resulta

ser grande (con un valor de probabilidad pequeño), declaramos que existe suficiente evidencia para concluir que el modelo es, en principio, adecuado.

Tabla 3.2. Tabla general de análisis de la varianza para el caso de la regresión lineal múltiple**.**

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrado Medio	Estadístico F_c
Regresión	$(p-1)$	SSR	$CMR = \frac{SSR}{p-1}$	$F_c = \frac{CMR}{CME}$
Error	$(n-p-1)$	SSE	$CME = \frac{SSE}{n-p-1} = \hat{\sigma}^2$	
Total	$(n-1)$			

Supuestos del modelo: En el modelo de regresión múltiple, los residuos ahora se calculan en presencia de los factores $X_1, X_2,\ldots, X_p$ que predicen a la variable Y; no obstante, los errores deben cumplir los mismos supuestos que en la regresión simple para garantizar la correcta estimación de los parámetros. Además de los supuestos mencionados en el caso de la regresión lineal simple, en el modelo de regresión múltiple, al estar incluida en el modelo más de una variable explicatoria, se debe cumplir que no exista relación entre estas variables; o se generaría un problema que es denominado de multicolinealidad, el cual afecta la precisión con la que se estiman los parámetros, y puede ser tan grave que genere patologías importantes, lo cual obliga a su adecuado diagnóstico y correspondiente tratamiento.

Una exploración inicial del supuesto de multicolinealidad, consiste en observar la matriz de correlación entre las variables $X_1, X_2, \ldots, X_p$. Si alguna o algunas de las correlaciones resultan ser mayores que 0.7 puede haber un problema de multicolinealidad. Usualmente los paquetes estadísticos proporcionan un diagnostico de multicolinealidad, que incluye algunas medidas del efecto que ocasiona sobre la precisión de las estimaciones, como se mencionó.

Cuando se tienen problemas de multicolinealidad se puede generar una sobrestimación de las varianzas y los errores estándar; las magnitudes de los coeficientes pueden ser diferentes a lo esperado; los signos podrían resultar contrarios a lo que se esperaría a partir de la teoría que explica el fenómeno y las pruebas estadísticas pueden

arrojar resultados contradictorios; así entonces, el problema de multicolinealidad es un problema que puede afectar seriamente a la selección del mejor modelo, por lo que se debe diagnosticar previamente a cualquier proceso de este tipo. La solución para un problema con multicolinealidad puede ser muy simple o puede ser muy compleja, en dependencia del futuro uso del modelo o de la posibilidad de obtener nuevos datos adicionales. Las soluciones señaladas para este problema son: la eliminación de algunas variables explicativas para romper la estructura de asociación, o bien, incluir más valores de cada una de ellas; transformarlas en otras variables no multicolineales o usando otros métodos de estimación como regresión Ridge o en componentes principales (Ver Gunst y Mason (1980)).

Coeficiente de determinación R^2*:* En el modelo de regresión simple, el Coeficiente de determinación R^2, como ya se explicó, es la proporción de la varianza en Y que es explicada por la variable X. En el caso de que existan más de una variable explicativa como sucede en el modelo de regresión múltiple, el Coeficiente de determinación será ahora, la proporción de la variabilidad de Y que es explicada por todas las variables del modelo. Un problema que surge con este coeficiente es que su valor se incrementa a medida que se aumenta el número de variables del modelo. Por lo tanto, en el modelo de regresión múltiple se recomienda utilizar una medida que se denomina R^2 ajustada. Este indicador considera el número de variables que tiene el modelo, así entonces, se trata de una medida de ajuste del modelo, que es penalizada por la complejidad del mismo, ponderando la cantidad de variables en función de la cantidad de datos. En este sentido si R^2 y R^2 ajustada son muy parecidas quiere decir que el tamaño de muestra es suficiente para el tamaño de la muestra de estudio.

Verificación de los supuestos: La verificación del cumplimiento de los supuestos que garantice la adecuada estimación del modelo, se realiza a través del análisis de los residuos. Para llevar a cabo este análisis, se elaboran varios tipos de gráficos como los que se presentan en la Figura 3.5, los cuales estarían reflejando heterocedasticidad en los residuos (incisos b, c, d). En a) se presenta el patrón adecuado.

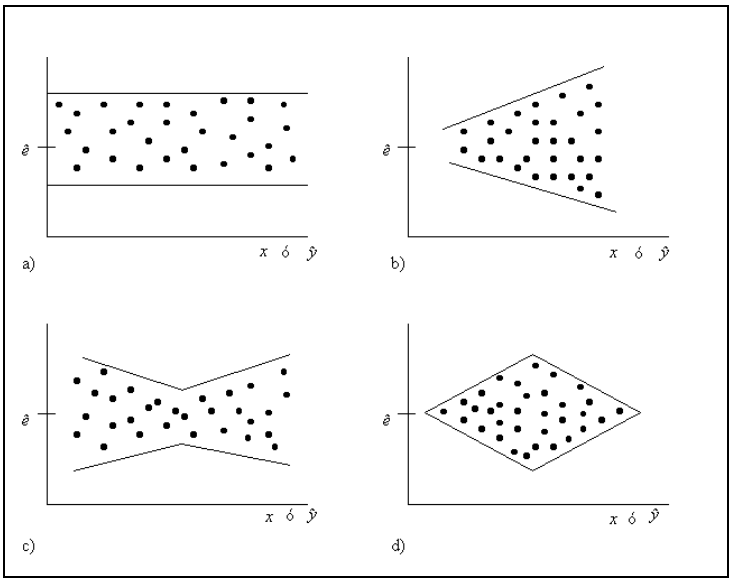


Figura 3.5. Gráficos con indicativos de problemas en el supuesto de homogeneidad de varianzas, excepto el que se presenta en el inciso a).

Otro aspecto importante relativo a los supuestos es el de la normalidad, que se requiere para garantizar la eficiencia de las pruebas de hipótesis y aunque no es un supuesto muy importante, ya que un tamaño de muestra grande puede atenuar los problemas que desviaciones de la normalidad ocasionen, sí debemos hacer una verificación de esta suposición. Para tener una idea de la razonabilidad de este supuesto podemos explorar los residuos a través de gráficos como histograma con curva ajustada, P-Plot y Q-Plot, o usando diagramas sencillos como los de tallos y hojas, los de dispersión o los de caja. En la Figura 3.6 se presentan cada uno de estos gráficos y diagramas, cuando los datos tienen una apariencia de normalidad razonable.

La presencia de observaciones atípicas también puede afectar a la bondad del ajuste del modelo. A veces, la atipicidad de un dato se observa en un gráfico de dispersión, otras veces es necesario ajustar el modelo y observar los residuos para identificarlo. La idea de elevar al cuadrado los residuos ayuda mucho, puesto que permite que en el gráfico se acentúen los valores de residuos grandes.

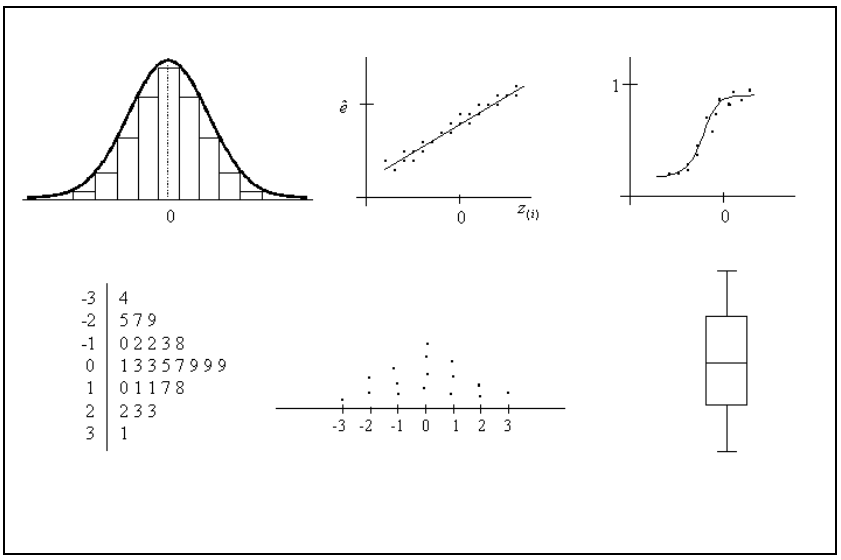


Figura 3.6. Diferentes despliegues gráficos que muestran razonabilidad en el supuesto de normalidad para un conjunto de datos.

Un punto atípico puede ser de diferente naturaleza, pero en general debe ser evaluado respecto al patrón determinado por el modelo. Por tal motivo una forma de identificar puntos atípicos para el modelo es construir una banda de predicción o confianza para los datos, como se muestra en la Figura 3.7.

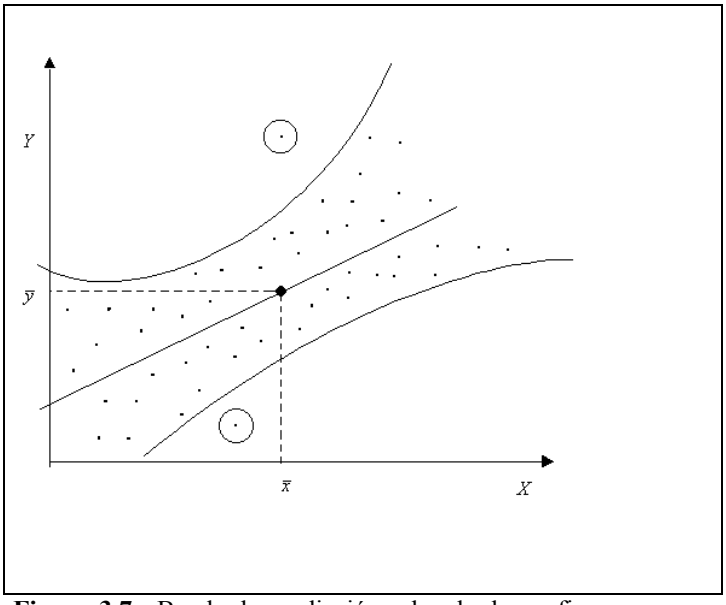


Figura 3.7. Banda de predicción o banda de confianza para un modelo ajustado mostrando dos observaciones claramente atípicas.

El problema de identificación de casos atípicos en un ajuste es de suma importancia en la regresión. En la Figura 3.7 podemos ver que hay un par de puntos que caen fuera del patrón esperado, definido por la banda de predicción. Esa sería una primera señal de que posiblemente esos casos son atípicos (outliers). Hay varios criterios, basados en varios tipos de residuos, que nos pueden guiar en la identificación concreta de puntos atípicos. Véase Barnet y Lewis (1994).

3.2.4 Análisis de regresión múltiple en SPSS

Para realizar un ejercicio de aplicación de los modelos de regresión utilizando el software estadístico SPSS, se utilizará la base de datos del artículo publicado en esta memoria, sobre el efecto del Fondo de Aportaciones para la Infraestructura Social de los Municipios en la variable respuesta, el Índice de Rezago Social de los municipios indígenas del estado de Veracruz (Véase sección 4.6). La base de datos contiene las siguientes variables con 50 observaciones:

Nombre de la variable	Descripción y categorización
IRIS	Índice de Rezago en Infraestructura Social. Diferencia entre 2005 y 2000 en cada municipio.
FAISM	Fondo de aportaciones para la Infraestructura de los Municipios medido en millones de pesos. (2000-2005).
REGIÓN	Región en la que se encuentra ubicado el municipio. 1=Zongolica 2=Huasteca 3=Popoluca 4=Totonaca
MUNICIPIO	Identificación del municipio indígena.

El objetivo es determinar si el FAISM que se destina a cada municipio indígena de Veracruz ha contribuido a explicar la diferencia entre el IRIS que presentaban los municipios en el año 2000 respecto a la cifra alcanzada en el 2005, así como analizar si la región a la que pertenece cada municipio influye en esta diferencia.

De esta manera, se asume una relación lineal entre el IRIS, como variable respuesta y el FAISM como variable explicatoria, que permite estimar el valor de los parámetros y ver como se afecta el IRIS ante un cambio en el FAISM. También se incluye la variable categórica de región, con la finalidad de conocer si hay diferencia en la relación entre el

IRIS y el FASIM dependiendo de la región a la que pertenezca el municipio. El modelo queda especificado de la siguiente manera:

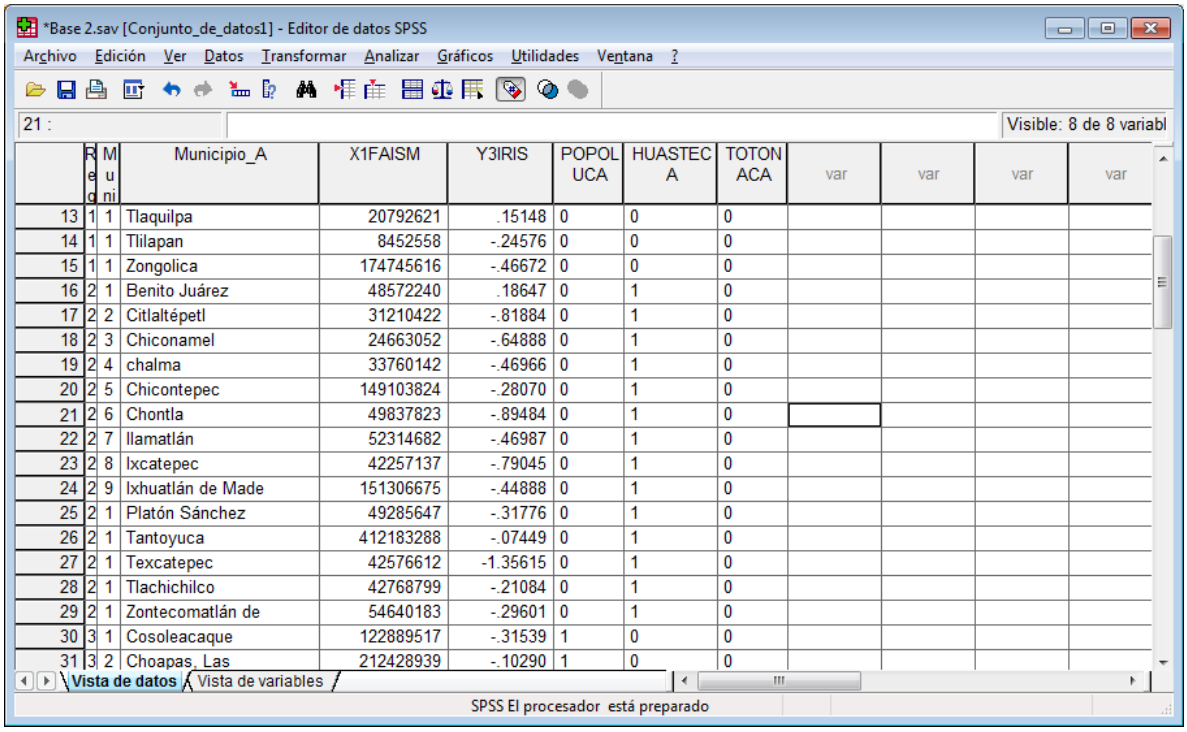
$$y_i = \beta_0 + \beta_1 \text{FAISM}_1 + \beta_2 \text{POPOLUCA}_2 + \beta_3 \text{HUASTECA}_3 + \beta_4 \text{TOTONACA}_4 + \varepsilon_i$$

$i = 1,2,..50$

$$\varepsilon_i \sim (\text{N}, \sigma^2)$$

Donde *yi* representa el IRIS para cada municipio *i*, β_0 representa el intercepto o el valor del IRIS cuando el FAISM es 0, β_1 es la pendiente y mide el cambio en el promedio del IRIS, cuando hay un cambio unitario en el FAISM. En este caso, como se está incluyendo la variable región, una variable cualitativa con 4 categorías: Zongolica, Huasteca, Popoluca y Totonaca, se deben crear 3 variables dummy en la base de datos, que sirven para indicar si el municipio pertenece a determinada región, utilizando una región como la categoría de referencia.

Para proceder a ejecutar el modelo, en el programa SPSS, primero se debe abrir la base de datos en el software SPSS, siguiendo las indicaciones presentadas en la sección 1.3. Ahora bien, se deben crear 3 variables dummy, como aparece en la ventana:

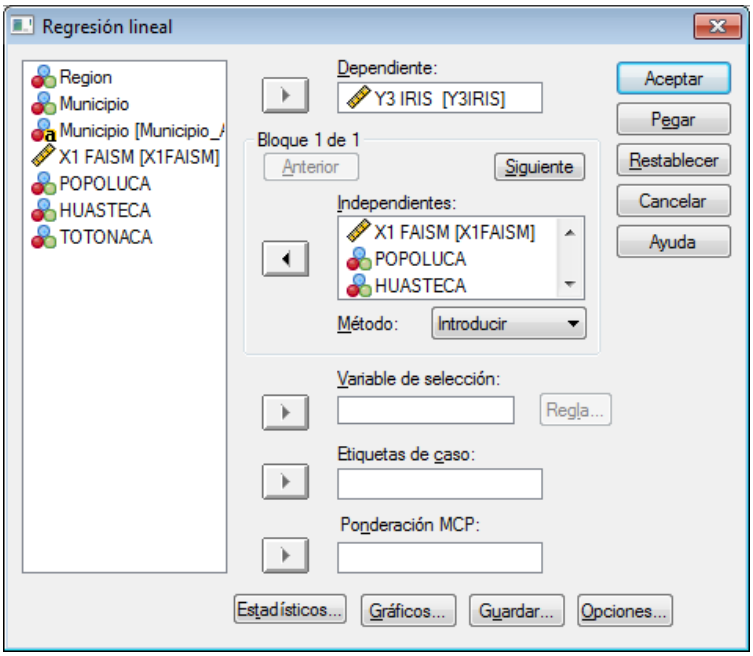


Así se tiene que en el modelo, se hizo la diferenciación a través de la siguiente categorización y se tomó la región Zongolica como categoría de referencia. Estas tres variables dummy fueron codificadas de la siguiente manera:

ZONGOLICA categoría de referencia.

- Si el municipio pertenece a la región HUASTECA, se codifica con 1, 0 si pertenece a la Zongolica, Popoluca o Totonaca.
- Si el municipio pertenece a la región POPOLUCA, se codifica con 1, 0 si pertenece a la Zongolica, Popoluca o Totonaca.
- Si el municipio pertenece a la región TOTONACA, se codifica con 1, 0 si pertenece a la Zongolica, Popoluca o Totonaca.

Para ejecutar el modelo de Regresión linean en SPSS, se debe ir al menú de *Analizar*, seleccionar la opción de *Regresión Lineal*, en la que aparecerá la siguiente ventana:



Una vez desplegadas las opciones, se selecciona la variable dependiente, que en este caso se trata del IRIS, así como las variables explicatorias: el FAISM y cada una de las variables dummy que se crearon para indicar la región a la que pertenece el municipio. El

método que se elegirá para este ejercicio es el que maneja por default el programa. Al darle click *Aceptar*, aparecerá la salida mostrando los resultados del modelo:

En el cuadro *Resumen del modelo*, muestra el coeficiente de determinación analizado en la sección 3.2.2, que indica la variabilidad explicada por el modelo. Para este ejercicio, se tiene que se trata de un 60%.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.597[a]	.356	.299	.34789215

a. Variables predictoras: (Constante), TOTONACA, X1 FAISM, POPOLUCA, HUASTECA

Asimismo, aparece el valor de los coeficientes del modelo que se ejecutó:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-.290	.093		-3.116	.003
	X1 FAISM	1.60E-009	.000	.323	2.607	.012
	POPOLUCA	-.383	.144	-.373	-2.658	.011
	HUASTECA	-.338	.132	-.369	-2.555	.014
	TOTONACA	-.668	.142	-.673	-4.706	.000

a. Variable dependiente: Y3 IRIS

Para concluir que variables resultan significativas y contribuyen a explicar a la variable dependiente, se recurre al valor Sig, que equivale al valor de probabilidad explicado en la sección 3.2.2. En este ejemplo, todas las variables resultan significativas a un nivel de confianza α=0.05. Esto quiere decir que el FAISM, sí influye en el IRIS, es decir, por cada millón de pesos que se aumente el FAISM, la diferencia entre el IRIS 2005 y 2000 de los municipios aumentará en .0000000016 unidades. Esta conclusión se aplica para todas las regiones.

Para el caso particular de cada región, que se crearon 3 variables dummy, se observa que las tres resultaron significativas a un nivel del 5%, por lo que se interpreta que el promedio del IRIS para Zongolica (la categoría de referencia) es -.290, mientras que para la región Huasteca es de -.628 (-.338+(-.290); para la Popoluca -0.628 (-.383+(-.290) y finalmente para la Totonaca -.958 (-.668 + (-.290). En este caso se obtiene un intercepto negativo porque indica que, en promedio en cada región, la diferencia entre el IRIS del 2005 y 2000 ha sido negativa, es decir, el valor del IRIS ha disminuido, respecto al que se tenía en el año 2000.

También se muestra, con el valor del estadístico *F* con 4 grados de libertad, que se rechaza la hipótesis de que alguno de los parámetros sea igual a 0, por lo que se concluye que las variables incluidas en el modelo contribuyen a explicar el comportamiento del Índice de rezago en Infraestructura Social de los municipios indígenas del Estado de Veracruz.

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3.009	4	.752	6.215	.000[a]
	Residual	5.446	45	.121		
	Total	8.455	49			

a. Variables predictoras: (Constante), TOTONACA, X1 FAISM, POPOLUCA, HUASTECA

b. Variable dependiente: Y3 IRIS

1.3. Modelos multinivel

Las muestras o poblaciones que tienen estructuras complejas en sus clasificaciones y anidamientos son bastante comunes en diferentes áreas de las Ciencias sociales, como en educación (se estudian estudiantes agrupados en escuelas, escuelas en zonas, etc.), en salud (pacientes, hospitales, regiones, etc.) y en Economía (estudios longitudinales, de grupos anidados de empresas, economía comparada de países, etc.). Esta situación se presenta particularmente en los estudios que abordan las finanzas públicas, donde se analizan comúnmente variables que se miden sobre las entidades federativas, las cuales a su vez están formadas (y los datos se desagregan) por los municipios, y a veces es necesario llegar hasta el nivel de áreas geoestadísticas básicas (AGEB´s). Cuando el caso es el de las entidades federativas que se estudian en un periodo de varios años, se tiene un conjunto de

series de tiempo (una para cada entidad), lo cual constituye una muestra anidada (años en entidades). En fin, que las estructuras de datos y poblaciones de referencia ordenadas jerárquicamente es muy frecuente, con lo que los problemas –llamados multinivel- plantean la necesidad del uso de metodologías de modelación estadística adecuadas. Para tratar este tipo de problemas la metodología estadística cuenta con una serie de técnicas, métodos y modelos que en la actualidad están bien definidos y se encuentran disponibles junto con el software que permite su adecuada aplicación para plantear y resolver problemas de este tipo, a través de ajuste de modelos, estimación de parámetros y de prueba de hipótesis, amén de la aplicación de técnicas exploratorias en los análisis preliminares.

La modelación multinivel ha adquirido especial atención desde finales de la década de los ochenta, aunque sus orígenes se remontan varios años atrás. Estos modelos fueron diseñados para analizar un fenómeno a partir de una o varias variables respuesta, considerando variables explicativas de diferentes niveles simultáneamente, para lo que se plantea y ajusta un modelo estadístico que apropiadamente incluye las diversas dependencias en los diferentes niveles. Los modelos multinivel incluyen una amplia gama de generalizaciones, pero son más conocidos y están bien estudiados los Modelos lineales multinivel, también llamados en la literatura científica como: Modelos de componentes de la varianza (Dempster, Rubin y Tsutakawa, 1981; Longford 1987), Modelos de coeficientes aleatorios (Rosenberg, 1973; de Leeuw y Kreft, 1986; Longford 1995), Modelos lineales jerárquicos (Raudenbush y Bryk, 1982, 1986), Modelos multinivel (Goldstein,1987; Mason *et al*., 1983) y Modelos de efectos mixtos (Laird y Ware, 1982; Milliken, Stroup y Wolfinger, 1996).

El interés que provocó el desarrollo de la modelación multinivel en la comunidad científica, ha acelerado su aplicación en diferentes disciplinas, tales como la Sociología, donde se introdujo el concepto de efecto contextual en este campo (Hox, 2002), en la Medicina con el Meta-análisis (Glass, 1976), los estudios de medias repetidas y curvas de crecimiento en las Ciencias del comportamiento (Laird y Ware, 1983), entre otras. Actualmente, su aplicación se ha extendido a diversas áreas del conocimiento. Como señala Bryk y Raudenbaush (1992), cuando se combinan con la gran cantidad de software disponible, esta expansión en la modelación ha inspirado toda una serie de nuevas aplicaciones.

El objetivo de este capítulo es introducir al lector en los conceptos básicos de la modelación multinivel. En una primera parte se sientan las bases de la modelación multinivel y se tratan las consecuencias de ignorar el anidamiento de los datos. Seguido se muestran los tipos de estructura multinivel y clasificaciones que pueden ser analizadas aplicando esta metodología. Posteriormente, se presenta la especificación del modelo, desde su versión más simple, hasta llegar a los modelos más generales pasando por el de de intercepto y pendientes aleatorias. Se explica en qué consisten los efectos contextuales, que constituye una de las aportaciones de este tipo de metodología comparada con otras técnicas de modelación estadística más básicas, así como la obtención del Coeficiente de correlación intraclase, indicador que brinda información sobre la proporción de la variabilidad que es explicada por el modelo en su versión multinivel. Asimismo, se trabaja con el análisis de los residuos asociados al ajuste del modelo. Finalmente, se presenta un ejemplo de aplicación de este tipo de metodología utilizando el software estadístico diseñado especialmente para este tipo de modelación, el paquete MLWin (Rasbash *et al*., 2009).

3.3.1 Introducción a los modelos lineales multinivel

Como se mencionó en capítulos previos, la parte fundamental de un análisis de datos son las unidades de estudio. Éstas se definen como el conjunto de observaciones de las cuales obtenemos información y a través de las cuales los valores medidos variarán. Las unidades pueden ser de varios tipos de acuerdo al contexto del problema. Sin embargo, en el caso de la modelación multinivel tienen una característica fundamental y ésta es que se encuentran anidadas, estructuradas o agrupadas en un cierto número de niveles o clasificaciones. Por ejemplo, estudiantes que se encuentran agrupados en clases, escuelas, vecindarios; entidades o provincias que pertenecen a países; trabajadores en empresas; árboles en bosques; pacientes en hospitales, etc. Frecuentemente, se estudian estas unidades no considerando su estructura de anidamiento, pero al omitir que pueden estar organizadas en un sistema jerárquico y pertenecer a diferentes niveles de clasificación, se puede llegar a incurrir en un problema conceptual y metodológico.

Para ilustrar lo anterior, supóngase que se desea estudiar qué factores influyen en el tiempo (medido en meses) que les toma a los estudiantes de doctorado en el país obtener su

grado académico. La forma más usual de abordar el problema, sería seleccionando una muestra de individuos quienes estuvieran realizando su doctorado en algún programa universitario. Las variables seleccionadas como explicativas podrían ser: el género de los individuos, es decir, interesaría saber si el ser hombre o mujer influye en que un estudiante termine sus estudios más pronto que otros; la edad; el tiempo dedicado a esta actividad, si es tiempo completo o trabajan paralelamente, y su rendimiento durante el programa. También se incluiría como variable explicativa la universidad a la que asisten. En este ejemplo, las unidades de estudio son los individuos; si al ajustar el modelo de dos niveles, resultan significativas las variables género, edad y universidad y se concluye que las universidades tienen una baja eficiencia terminal en sus programas de doctorado por la edad y el género de los estudiantes que reciben, se incurre en una "falacia atomística" (Alker, 1969), pues se están infiriendo relaciones a nivel grupal de relaciones a nivel individual.

Si por el contrario, las unidades de estudio fueran las universidades y se contemplaran otras variables explicativas a este nivel grupal, como el tipo de universidad (pública o privada), la duración del programa académico, si está registrado en algún padrón de calidad y el área de conocimiento, y se elaboran conclusiones a nivel individual, por ejemplo, exponer que los factores que influyen en que los estudiantes terminen sus estudios de doctorado en un menor tiempo, se deben al tipo de universidad a la que asisten y al registro del programa en un padrón de excelencia[2], se incurre en una "falacia ecológica" (Robinson, 1950*),* lo que se traduce en el error de interpretar los resultados de grupo (universidades) como si se aplicaran a los individuos (estudiantes). En otras palabras, se comete esta falacia al elaborar conclusiones a nivel individual considerando información agregada.

Hox (2002) ha identificado particularmente dos problemas por ignorar la estructura jerárquica de los datos: se pierde información y el análisis es menos robusto. Las pruebas estadísticas ordinarias, tratan los valores de los datos desagregados como valores independientes de la muestra, lo que origina que los errores estándar sean pequeños y esto a su vez conduce a pruebas de hipótesis significativas, cuando realmente no lo son.

[2] Esta conclusión puede ser válida si el modelo especificado fuera el adecuado. Por ello, resulta sumamente importante la correcta definición de las unidades de estudio y sus clasificaciones, así como la apropiada selección e inclusión de variables.

Los límites grupales en la realidad frecuentemente son confusos y arbitrarios y la asignación de variables no siempre es obvia y simple. Los modelos multinivel tienen el propósito de subsanar esta problemática y analizar los datos considerando la estructura jerárquica de los mismos, al modelar la realidad con la existencia de diferentes niveles de variación (Rasbash *et al*., 2009). Por dicha razón, los modelos multinivel se aplican principalmente a datos que presentan una estructura jerárquica, es decir que se encuentran estructurados en un cierto número de niveles o clasificaciones. Estas condiciones permiten tener una mejor comprensión de la variabilidad de los datos, pues se logra conocer la varianza entre las unidades de un mismo grupo y la varianza entre los grupos, condición limitada en un análisis de regresión no multinivel, donde sólo hay un tipo de error ε_i. Esta forma de modelación de la varianza, en varios niveles, proporciona un marco más sólido que permite generar un amplio espectro de preguntas sobre el problema en cuestión, tales como los efectos contextuales, que pueden ser sumamente importantes en el problema de investigación.

3.3.2 Estructuras jerárquicas y clasificaciones

En las estructuras jerárquicas, el tipo de organización de datos se origina cuando las unidades de nivel más bajo se anidan o agrupan en unidades de nivel más alto. Retomando el ejemplo anterior planteado en la sección previa, en el que el objetivo es especificar un modelo que permita explicar los factores que influyen en el tiempo que requieren los estudiantes de doctorado para titularse, un adecuado diseño definirá el modelo considerando como unidades de estudio en un primer nivel a los estudiantes, agrupados en distintas universidades, también unidades de estudio pero en un segundo nivel. Esto representa una estructura jerárquica de dos niveles, lo que significa que las unidades pertenecen a un grupo de anidamiento, ya sea en un primer nivel, segundo nivel o más. Es decir, los estudiantes sólo realizan un programa de doctorado en una universidad. De acuerdo con Rasbash (2008), las estructuras jerárquicas pueden ser representadas por diagramas de unidad o diagramas de clasificación para tener una mejor comprensión del problema.

Los diagramas de unidad, tienen el objetivo de mostrar la estructura subyacente del problema de investigación, en términos de las unidades primarias. Los puntos en el diagrama son las unidades de la población específica, como se observa en la Figura 3.8.

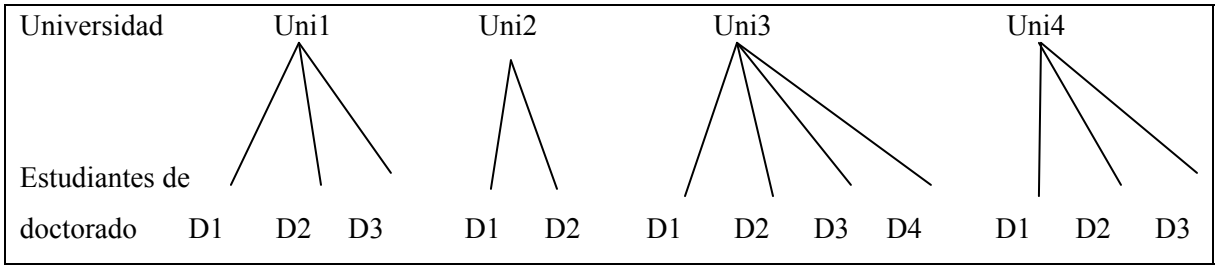


Figura 3.8. Diagramas de unidad para una estructura jerárquica de dos niveles; estudiantes de doctorado en 4 universidades**.**

Por su parte, los diagramas de clasificación son más utilizados cuando la población objetivo tiene una estructura compleja, pues son más abstractos y tienen un nodo por cada nivel que se une a través de una flecha, como se muestra en la Figura 3.9.

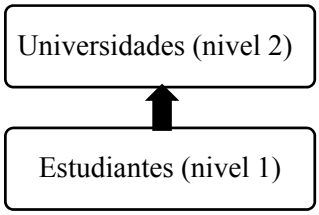


Figura 3.9. Diagrama de clasificación para una estructura jerárquica de dos niveles; estudiantes en universidades.

En el diagrama de unidad de la Figura 3.8, se puede apreciar que las universidades no tienen el mismo número de estudiantes. La universidad 1 tiene 3 y la número 2 registra 4 estudiantes. Esto significa que los datos no están balanceados. Una de las características de los modelos multinivel es que no requieren que los grupos sean del mismo tamaño, lo cual es muy frecuente en los problemas multinivel reales.

Modelar datos con una estructura jerárquica de dos niveles como la que se visualiza en los diagramas, permite responder a un amplio espectro de preguntas que enriquecen la labor de investigación y que resultaría erróneo resolver utilizando otras técnicas más básicas. Al aplicar una modelación de dos niveles para estudiar los factores que influyen en el tiempo de titulación de los estudiantes de doctorado en el país, con información sobre los

estudiantes en un primer nivel y teniendo variables explicativas sobre las universidades en las que realizan sus estudios (nivel 2), permite tener una comprensión más holística del problema que si se utilizara una regresión ordinaria.

Evidentemente, las características individuales de los estudiantes influyen en el tiempo que les toma para la obtención de su grado, pero también es importante considerar el contexto; es decir, contemplar la existencia de otros factores, como las características de las universidades en las que realizaron su programa, que indudablemente tiene un efecto en la variable respuesta. En este sentido, el modelo permite conocer la variabilidad del tiempo de obtención del grado entre las universidades y dentro de una misma universidad; si la cantidad de tiempo varía entre las universidades públicas o privadas; o saber si el género es un factor determinante en la obtención del grado de los estudiantes y varias interrogantes más que pueden determinarse a través de un análisis multinivel.

Las estructuras jerárquicas de dos o más niveles se pueden analizar con modelación multinivel; pero además, se puede modelar la realidad social no sólo como la interrelación de las unidades dentro de una misma clasificación, sino también de diferentes clasificaciones. Es decir, los datos siguen una estructura jerárquica particular en la que un mismo individuo pertenece a diferentes niveles de clasificación. Así se tiene que la modelación multinivel puede ser aplicada tanto a estructuras jerárquicas o de anidamiento como las descritas anteriormente, en las que las unidades pertenecen a un sólo sistema de clasificación, como a estructuras de clasificación cruzada o estructuras múltiples, donde los individuos están incluidos en más de un nivel de clasificación simultáneamente. En este apartado, sólo se abordarán las estructuras jerárquicas o de anidamiento; para ahondar en el tema sobre la aplicación de modelación multinivel a estructuras no jerárquicas véase Rasbash (2008).

Hasta ahora, se han ejemplificado los anidamientos de datos de dos niveles cuya estructura obedece a la naturaleza propia de los datos. Por ejemplo, los estudiantes agrupados en escuelas, las ciudades en países, los pacientes en hospitales, etc. Sin embargo, la definición de una estructura jerárquica de datos para ser analizada utilizando modelación multinivel, no necesariamente debe obedecer a una cuestión natural de los datos.

También es factible que esta estructura pueda ser impuesta a través de un diseño estadístico y de recolección de datos; tal es el caso de los datos de panel o medidas

repetidas. Un conjunto de datos de panel contiene información de múltiples unidades individuales a lo largo de un periodo de tiempo. De esta manera, se tiene una medida (nivel 1) que varía en el tiempo en un número de individuos o del fenómeno en cuestión (nivel 2), lo que significa que las unidades medidas están anidadas dentro de los individuos. Este diseño se utiliza cuando se desea analizar la variación entre los individuos y sus patrones de crecimiento.

En la Tabla 3.3 se observa un conjunto de datos de panel para un estudio de dos niveles, en el que, continuando con el tema de Educación Superior y el estudio de los estudiantes de doctorado y la obtención del grado, se presentan datos ficticios[3] del número de titulados que se han registrado en algunas universidades del país en los últimos cuatro años.

Tabla 3.3. Ejemplo de datos ficticios de panel para un estudio de dos niveles.

Niveles		**Variable respuesta**	**Variables explicatorias**	
Año (1) ***i***	**Universidad (2)** ***j***	**Titulados de doctorado$_{ij}$**	**Alumnos inscritos$_{ij}$**	**Tipo de universidad$_{j}$**
2004	U. Veracruzana	18	20	Pública
2005	U. Veracruzana	31	27	Pública
2006	U. Veracruzana	48	47	Pública
2007	U. Veracruzana	52	52	Pública
2004	U. de las Americas	40	39	Particular
2005	U. de las Americas	59	57	Particular
2006	U. de las Americas	72	72	Particular
2007	U. de las Americas	81	78	Particular

Fuente: Elaboración propia con datos hipotéticos

3.3.3. Relevancia de los modelos multinivel

La importancia de estos modelos radica en que se puede tener una mejor comprensión de la variabilidad de los datos, pues permite conocer la varianza entre las unidades de un mismo grupo y entre grupos. Esta línea de investigación es muy potente, pues otras técnicas de análisis estadístico no permiten obtener esta información. Retomando el ejemplo que se ha presentado, si se utiliza un modelo de dos niveles, es posible llegar a conocer la variación

[3] Los datos son ficticios porque no se obtuvieron de una fuente real. Su objetivo es sólo ejemplificar la estructura de la base de datos multinivel.

que existe entre las universidades y cuánta de esta variabilidad es explicada por las variables seleccionadas a este nivel.

Por otro lado, se tiene que al ajustarse un modelo de un solo nivel (Regresión ordinaria), se ignorarían los efectos de agrupamiento y por lo tanto, se obtendrían estimadores sesgados que conducirían a inferencias erróneas. En los casos en que se opta por introducir variables indicadoras para considerar el efecto del grupo, se restringe el análisis al número de grupos de la muestra y el número de parámetros adicionales a estimar también aumentará. Los efectos de las variables explicativas a nivel de grupo no pueden ser estimados simultáneamente utilizando los residuos del agrupamiento (Steele, 2008), ni es posible calcular un solo parámetro que refleje esta información. Las técnicas usuales no están diseñadas para dividir la variación de esta manera y sólo estiman un término para explicar esta diferencia, al que se le denomina error. En la modelación multinivel esta variación presenta una estructura relevante susceptible de ser analizada y que aporta mucha información al problema.

3.3.4. Variables y niveles

Definición y clasificación: Una de las principales cuestiones que surgen cuando se diseña el estudio estadístico para un modelo multinivel, es definir cuando una variable debe ser tratada como nivel o como variable explicativa. Un nivel es una clasificación aleatoria de unidades que puede ser considerada como una muestra aleatoria de una población (Goldstein, 1991). Por ejemplo, los estudiantes y las universidades del ejemplo, constituyen una muestra aleatoria de todos los estudiantes de doctorado que estudian en el país y de las universidades que ofrecen este tipo de estudios de posgrado (población). Por ello, estudiantes y universidades son considerados niveles y no variables explicativas. Por su parte, las variables explicativas que no son continuas, tienen un número de categoría fijas y no hay una población de la que hayan sido muestreadas. Así tenemos que hay dos tipos de clasificaciones para los efectos: Fijos y Aleatorios. La distinción entre este tipo de clasificaciones tiene importantes alcances sobre cómo incluir las variables en el diseño estadístico. Rasbash (2008) señala que un nivel en un modelo jerárquico debe necesariamente corresponder a una clasificación aleatoria. De obedecer a una clasificación fija, será tratada como variable explicativa.

3.3.5 Tamaño de muestra en los modelos multinivel

El número de unidades que deben ser incluidas en cada nivel del modelo, es una de las preguntas más frecuentes cuando se utiliza este tipo de metodología. La respuesta a esta interrogante estará en función principalmente de los intereses del investigador y de las unidades de estudio. Si el objetivo es estudiar la variación entre las universidades del país respecto al tiempo que tardan sus estudiantes de doctorado en obtener el grado, se necesitará información de varias universidades con el objetivo de obtener estimadores confiables. Esto significa que no se podría utilizar información sólo de dos universidades aunque se tuvieran datos de 500 estudiantes titulados en esa universidad. Goldstein (1999) recomienda que dada la magnitud de los efectos que es común encontrar entre las diferencias de las escuelas, se requiere información de al menos 25 centros escolares para proporcionar un estimador preciso de la varianza entre las escuelas. Por su parte, Snijders y Boslíe (1993) señala que la robusticidad de las pruebas estadísticas usualmente depende del tamaño de la muestra y ha diseñado un software especializado, llamado PinT, de las siglas de Power Analysis in Two Level Designs para la determinación del tamaño de muestra óptimo en diseños multinivel (Véase Snijders, 2005).

3.3.6. Estructura del modelo multinivel

El modelo multinivel busca estimar los parámetros desconocidos (intercepto y pendiente), pero además la varianza dentro de un grupo σ^2 y la varianza entre los grupos σ^2_{u0}. La estimación de los coeficientes puede realizarse a través de diferentes enfoques como el de Máxima verosimilitud o Estimación bayesiana, y utilizando diversos algoritmos como el de Mínimos cuadrados generalizados iterativos (MCGI) (Goldstein, 1999), el de Fisher-Scoring (Longford,1987) y el algoritmo EM (Lindley y Smith, 1972). Actualmente, existen diversos paquetes estadísticos para el cálculo de los coeficientes.

Antes de presentar el modelo multinivel, partamos del modelo de regresión ordinaria más simple. En un modelo de regresión ordinaria para un solo nivel, sin considerar variables explicativas, la ecuación es:

$$y_i = \beta_0 + \varepsilon_i. \qquad (3.4)$$

Donde y_i es el valor que toma la variable respuesta para la *i*-ésima observación (i = 1, 2,…, n), el intercepto ó β_0 representa el promedio de y en la población, y ε_i es el "error" para la *i*-ésima observación; esto es la diferencia entre el valor observado de y con respecto a la media poblacional (Véase Figura 3.10), siendo uno de los supuestos básicos de este modelo, que los residuos se distribuyen como una normal de media cero y varianza constante $\varepsilon_i \sim N(0, \sigma^2)$. La varianza resume la variabilidad alrededor de la media. Entre más grande sea este valor, la diferencia con respecto a la media se incrementa.

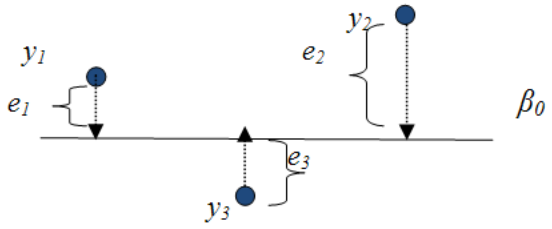


Figura 3.10. Residuos para tres puntos de un modelo de un solo nivel respecto a la media.

Ahora se introducirá el segundo nivel al modelo anterior. Supóngase que se tiene un conjunto de individuos en el nivel 1, anidados en grupos en el nivel 2. Para expresar algebraicamente esta relación, se añade el subíndice j a la respuesta de y_i, de esta manera y_{ij} representa el valor de y para el *i*-ésimo individuo en el *j*-ésimo grupo. Como se mencionó anteriormente, el modelo multinivel permite estimar la variabilidad entre los individuos de un mismo grupo y la variabilidad entre los grupos. Por lo tanto, el error se dividirá en dos componentes[4], correspondiente a estas dos variaciones. Los errores entre los grupos se denotan como u_j y entre los individuos como ε_{ij}. Integrando estos elementos al modelo (3.4), se origina la siguiente expresión:

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij} \tag{3.5}$$

donde β_0 ahora representa la media general de y para todos los grupos, u_j es la diferencia entre la media del grupo j y la media global (Véase Figura 3.11). En este caso, la media del grupo j es $\beta_0 + u_j$. Para los errores en el nivel 1, ε_{ij} representa la diferencia entre los valores de y para el *i*-ésimo individuo con respecto a la media de su grupo, $\varepsilon_{ij} = y_{ij} - (\beta_0 + u_j)$.

[4] Por esta razón, los modelos multinivel también son conocidos ampliamente como Modelos de componentes de la varianza.

Tal como en los modelos de regresión, se asume que ambos errores se distribuyen como una normal con media cero y varianza constante; es decir, $u_j \sim N(0, \sigma_u^2)$ y $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

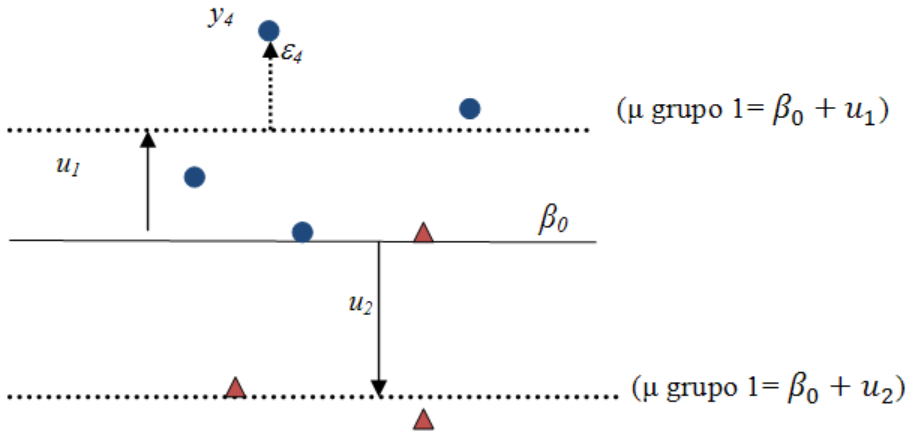


Figura 3.11. Errores a nivel individual y grupal en un modelo de dos niveles.

El modelo también puede ser expresado de la siguiente forma:

Nivel 1 $\quad y_{ij} = \beta_{0j} + \varepsilon_{ij}$
Nivel 2 $\quad \beta_{0j} = \beta_0 + u_j$
Combinado $\quad y_{ij} = \beta_0 + u_j + \varepsilon_{ij}$

$u_j \sim N(0, \sigma_u^2)$
$\varepsilon_{ij} \sim N(0, \sigma_e^2)$

Los parámetros a estimar son $\beta_0, \sigma_u^2 \ \sigma_e^2$.

Modelo multinivel de intercepto aleatorio: Hasta el momento, se ha estructurado el modelo sólo con el intercepto para ver el comportamiento de la variable respuesta debido sólo a la variabilidad entre los grupos o niveles y dentro de cada grupo. Ahora se añadirá una variable explicativa en el nivel 1. Supóngase que se tiene una variable continua explicativa en el nivel 1 denotada por $x_{ij.}$ El subíndice *ij* en *x*, indica que los valores de *x* cambian de observación a observación dentro de un grupo. El modelo queda especificado de la siguiente forma:

$$\underbrace{y_{ij}} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{parte fija}} + \underbrace{u_j + \varepsilon_{ij}}_{\substack{\text{parte} \\ \text{aleatoria}}} \qquad (3.6)$$

En la expresión 3.6, la relación global entre x y y está representada por una línea recta en la que β_0 muestra el intercepto o la altura de esta línea para el valor esperado de la variable respuesta dada una variable explicativa, y la pendiente o β_1 constituye el cambio de la media de la variable respuesta para un cambio unitario de la variable explicativa. Se debe tener presente lo que se observa en la Figura 3.11, el intercepto para un grupo dado j está definido por la relación $\beta_0 + u_j$.

De esta forma, se tiene que como un tipo de modelo estadístico, el modelo multinivel está compuesto por dos partes: una fija y otra aleatoria como se observa en 3.6 La parte fija muestra la relación entre la media de y y la variable explicativa y el componente aleatorio contiene los residuos del nivel 1 y del nivel 2. Usualmente, este modelo se conoce como modelo de intercepto aleatorio, porque el intercepto de la línea de regresión puede variar entre los grupos, pero la pendiente se asume fija para cada grupo. Gráficamente esto significa que se tendrán líneas de regresión para cada grupo paralelas entre sí, tal como se observa en la Figura 3.12. Por esta razón, también se puede especificar de la siguiente manera:

$$\begin{aligned} &\text{Nivel 1} && y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \\ &\text{Nivel 2} && \beta_{0j} = \beta_0 + u_j \\ &\text{Combinado} && y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij} \end{aligned} \qquad (3.7)$$

$u_j \sim N(0, \sigma_u^2)$
$\varepsilon_{ij} \sim N(0, \sigma_e^2)$

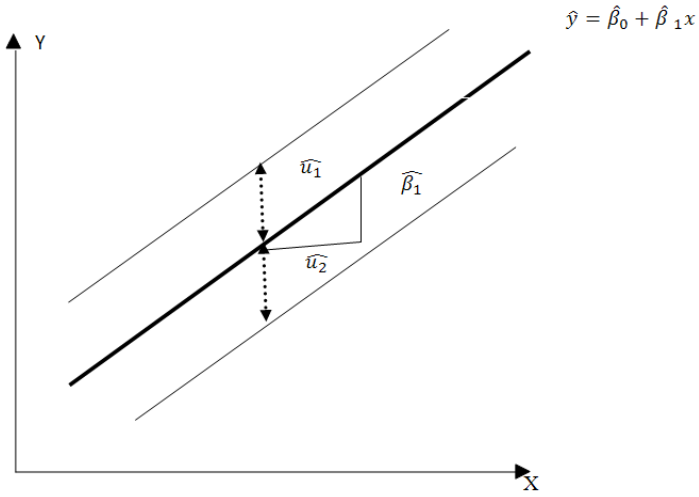


Figura 3.12. Representación gráfica de un modelo de intercepto aleatorio.

Modelo multinivel de coeficientes aleatorios: En el modelo anterior (3.7), la pendiente β_1 se mantenía fija para todos los grupos, pero supóngase que ésta varía aleatoriamente entre los grupos, lo que nos conduce a un modelo de pendiente aleatoria:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij} \qquad (3.8)$$

que también puede ser escrito como:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u): \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

Como se aprecia en la expresión 3.8, se han agregado nuevos términos a la ecuación, dentro del componente aleatorio. Ahora se tiene $u_{1j}x_{ij}$, y se añadió el subíndice 0 al término $u_{j.}$. Asimismo, los supuestos se han modificado, pues ahora se asume que los errores u_{0j} y u_{1j}, se distribuyen como una normal bivariada con media cero y varianza σ^2_{u0}, σ^2_{u1} , y covarianza σ_{u01}, que es la covarianza entre los interceptos de grupo y las pendientes. Ahora la pendiente de la línea de regresión global es β_1 y la pendiente para cada grupo j es $\beta_1 + u_{1j}$, por lo que la interpretación de los coeficientes cambia, como se aprecia en la Figura 3.13.

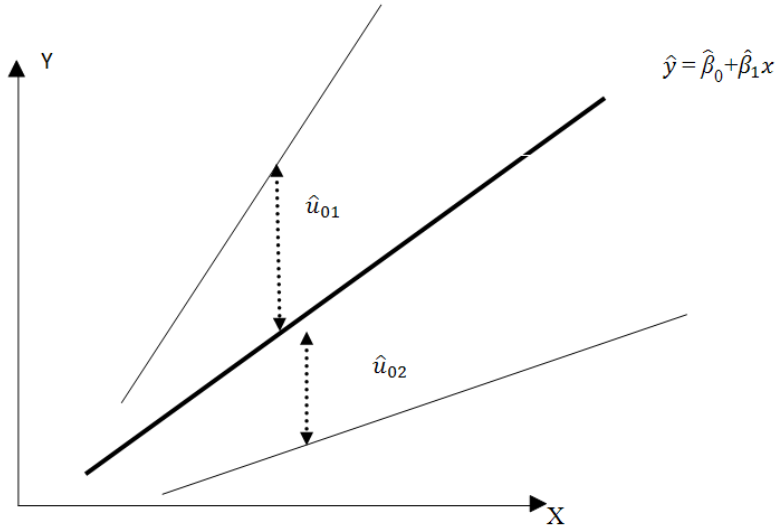


Figura 3.13. Representación gráfica de un modelo con pendiente aleatoria de dos niveles.

Efectos contextuales. Variables explicativas al segundo nivel: Una de las ventajas que ofrecen los modelos multinivel es, como se ha descrito anteriormente, la posibilidad de conocer los efectos que tienen las variables explicativas de grupo a nivel 2 en la variable respuesta. Las variables a este nivel se definen como variables contextuales y por tanto, sus efectos en Y se conocen como efectos contextuales (Steele, 2008). Retomando el ejemplo inicial de los estudiantes de doctorado, al plantearse un modelo multinivel y definir variables explicativas a nivel universidad, se puede conocer el efecto del contexto en el problema. Esto significa que el fenómeno del tiempo que les toma a los estudiantes obtener su título no sólo depende de factores individuales, sino que también tienen un efecto importante las características de las universidades donde realizaron su programa.

Introducir las variables contextuales en el modelo, es muy similar al procedimiento realizado anteriormente, donde se incluyó una variable explicativa al nivel 1. Sin embargo, en la estructura de los datos, es importante tener presente que las variables explicativa a nivel 2 tienen un valor constante dentro de cada grupo. El modelo toma la siguiente forma:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{2j} + u_j + \varepsilon_{ij} \tag{3.9}$$

Como se observa en la expresión 3.9, x_2 representa la variable explicatoria a nivel 2 y sólo tiene el subíndice j, pues como se mencionó, sus valores no varían de observación en observación dentro de las unidades de nivel 2.

3.3.7. Modelo de regresión para datos con dos niveles en notación matricial

Definiendo

$$\mathbf{Y}_j = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{n_j j} \end{bmatrix}; \mathbf{X}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{n_j j} \end{bmatrix} = \begin{bmatrix} 1 & x_{11j} & x_{21j} & \cdots & x_{m1j} \\ 1 & x_{12j} & x_{22j} & \cdots & x_{m2j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n_j j} & x_{2n_j j} & \cdots & x_{mn_j j} \end{bmatrix}; \boldsymbol{\varepsilon}_j = \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{n_j j} \end{bmatrix}; \quad \boldsymbol{\beta}_j = \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \vdots \\ \beta_{mj} \end{bmatrix};$$

En forma matricial el modelo nivel 1 toma la forma:

$$\mathrm{Y}_j = \mathrm{X}_j \beta_j + \varepsilon_j ; \; j = 1,...,J , \tag{3.10}$$

donde $\mathbf{Y}_j$ es el vector respuesta, n_j x 1, $\mathbf{X}_j$ es la matriz de variables explicativas a nivel 1 de orden $n_j \text{ x} (m+1)$, $\boldsymbol{\beta}_j$ es el vector de parámetros de orden $(m+1)$ x 1 y $\mathbf{e}_j$ es un vector de errores aleatorios, n_j x 1. Se supone $\text{E}\left(\mathbf{e}_j\right)=\mathbf{0}$, $\text{Var}\left(\mathbf{e}_j\right)=\sigma^2\,\mathbf{I}_{n_j}$ y el supuesto de normalidad. Definiendo

$$\mathbf{W}_j=\begin{bmatrix} 1 & w_{1j} & w_{2j} & \cdots & w_{qj} & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & w_{1j} & w_{2j} & \cdots & w_{qj} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ & & \vdots & & & & \vdots & & & & \ddots & & & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & w_{1j} & w_{2j} & \cdots & w_{qj} \end{bmatrix},\text{y}$$

$$\beta = \left[\beta_{00},\beta_{01},...,\beta_{0q},\beta_{10},\beta_{11},...,\beta_{1q},...,\ \beta_{m0},\beta_{m1},...,\beta_{mq}\right]^T;\quad \mathbf{u}_j=\begin{bmatrix} \text{u}_{0j} \\ \text{u}_{1j} \\ \vdots \\ \text{u}_{mj} \end{bmatrix}.$$

En forma matricial el modelo nivel 2 tiene la forma:

$$\boldsymbol{\beta}_j=\mathbf{W}_j\boldsymbol{\beta}+\mathbf{u}_j;\quad j=1,...,J\,, \tag{3.11}$$

donde $\mathbf{W}_j$ es la matriz de variables explicativas a nivel 2, de orden $(m+1)$ x $(q+1)(m+1)$, $\boldsymbol{\beta}$ es el vector $(m+1)(q+1)$ x 1 de coeficientes fijos, y $\mathbf{u}_j$ es el vector de errores aleatorios del nivel 2 de orden $(m+1)$ x 1. Supóngase $\text{E}\left(\mathbf{u}_j\right)=0,$ y

$$\text{Var}\left(\mathbf{u}_{\text{j}}\right)=\boldsymbol{\Omega}=\begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} & \cdots & \sigma_{u0m} \\ \sigma_{u10} & \sigma_{u1}^2 & \cdots & \sigma_{u1m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{um0} & \sigma_{um1} & \cdots & \sigma_{um}^2 \end{bmatrix}, \tag{3.12}$$

además del supuesto de normalidad.

En forma matricial el modelo combinado para la j-ésima unidad de nivel 2 toma la forma:

$$\mathbf{Y}_j = \mathbf{X}_j \mathbf{W}_j \boldsymbol{\beta} + \mathbf{X}_j \mathbf{u}_j + \boldsymbol{\varepsilon}_j ; \ \ j = 1,...,J ,$$

$$\mathrm{E}\left(\mathbf{Y}_j\right) = \mathbf{X}_j \mathbf{W}_j \boldsymbol{\beta} ,$$

$$\mathbf{V}_j = \mathrm{Var}\left(\mathbf{Y}_j\right) = \mathbf{X}_j \boldsymbol{\Omega} \mathbf{X}_j^T + \sigma_e^2 \mathbf{I}_{n_j} . \tag{3.13}$$

El modelo de interés es el modelo combinado del modelo nivel 1 con una variable explicatoria a nivel 1, x_{1ij}

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \varepsilon_{ij} ,$$

$$\mathrm{E}\left(\varepsilon_{ij}\right) = 0 , \quad \mathrm{Var}\left(\varepsilon_{ij}\right) = \sigma_e^2 , \tag{3.14}$$

y del modelo nivel 2 con una variable explicatoria a nivel 2 , w_{1j}

$$\beta_{0j} = \beta_{00} + \beta_{01} w_{1j} + u_{0j}; \qquad \beta_{1j} = \beta_{10} + \beta_{11} w_{1j} + u_{1j} ,$$

$$\mathrm{E}\left(u_{0j}\right) = 0 , \ \ \mathrm{Var}\left(u_{0j}\right) = \sigma_{u0}^2 , \ \ \mathrm{Cov}\left(u_{0j}, u_{1j}\right) = \sigma_{u01} ,$$

$$\mathrm{E}\left(u_{1j}\right) = 0 , \ \ \mathrm{Var}\left(u_{1j}\right) = \sigma_{u1}^2 , \qquad . \tag{3.15}$$

el cual tiene la forma:

$$y_{ij} = \left(\beta_{00} + \beta_{01} w_{1j} + u_{0j}\right) + \left(\beta_{10} + \beta_{11} w_{1j} + u_{1j}\right) x_{1ij} + \varepsilon_{ij} ,$$

$$y_{ij} = \left(\beta_{00} + \beta_{01} w_{1j} + \beta_{10} x_{1ij} + \beta_{11} w_{1j} x_{1ij}\right) + \left(u_{1j} x_{1ij} + u_{0j} + \varepsilon_{ij}\right),$$

$$\mathrm{Var}\left(\varepsilon_{ij}\right) = \sigma_e^2 , \quad \mathrm{Var}\left(u_{0j}\right) = \sigma_{u0}^2 , \ \ \mathrm{Var}\left(u_{1j}\right) = \sigma_{u1}^2 , \tag{3.16}$$

$$\mathrm{Cov}\left(u_{0j}, u_{1j}\right) = \sigma_{u01} , \quad \mathrm{Cov}\left(u_{kj}, \varepsilon_{ij}\right) = 0 .$$

del modelo (3.16) se tiene

$$\begin{aligned}\mathrm{Var}(\mathrm{y}_{ij}) &= \mathrm{Var}(u_{1j}\mathrm{x}_{1ij} + u_{0j} + \varepsilon_{ij}),\\ &= \sigma^2_{u0} + \sigma^2_{u1}x^2_{1ij} + 2\sigma_{u01}\mathrm{x}_{1ij} + \sigma^2_e.\end{aligned} \tag{3.17}$$

De (3.17) se tiene que en forma matricial el modelo combinado para la j-ésima unidad de nivel 2 toma la forma:

$$\begin{aligned}\mathbf{Y}_j &= \mathbf{X}_j\mathbf{W}_j\boldsymbol{\beta} + \mathbf{X}_j\mathbf{u}_j + \mathbf{e}_j;\ j = 1,...,J,\\ \mathrm{E}(\mathbf{Y}_j) &= \mathbf{X}_j\mathbf{W}_j\boldsymbol{\beta},\\ \mathbf{V}_j &= \mathrm{Var}(\mathbf{Y}_j) = \mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j^T + \sigma^2_e\mathbf{I}_{n_j},\end{aligned} \tag{3.18}$$

donde

$$\mathrm{Var}(\mathbf{u}_j) = \boldsymbol{\Omega} = \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} & \cdots & \sigma_{u0m} \\ \sigma_{u10} & \sigma^2_{u1} & \cdots & \sigma_{u1m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{um0} & \sigma_{um1} & \cdots & \sigma^2_{um} \end{bmatrix}. \tag{3.19}$$

Definiendo

$$\mathbf{Y} = \begin{bmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \\ \vdots \\ \boldsymbol{Y_J} \end{bmatrix};\ u = \begin{bmatrix} \boldsymbol{u_1} \\ \boldsymbol{u_2} \\ \vdots \\ \boldsymbol{u_J} \end{bmatrix};\ \varepsilon = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_J} \end{bmatrix};\ \mathbf{X} = \mathrm{diag}(\mathbf{X}_j);\ \text{y}\ \mathbf{W} = \begin{bmatrix} \boldsymbol{W_1} \\ \boldsymbol{W_2} \\ \vdots \\ \boldsymbol{W_J} \end{bmatrix} \tag{3.20}$$

donde $\mathrm{diag}(\mathbf{A}_j)$ representa los términos diagonales por matriz bloque, con $\mathbf{A}_j$ ($j = 1,...,J$) en el $j-$ésimo bloque de la diagonal. El modelo lineal jerárquico toma la forma:

$$\mathbf{Y} = \mathbf{XW}\boldsymbol{\beta} + \mathbf{Xu} + \boldsymbol{\varepsilon}, \tag{3.21}$$

el cual se denomina modelo lineal general jerárquico.

La matriz de varianzas y covarianzas tiene la forma

$$\mathbf{V} = \text{Var}(\mathbf{Y}) = \mathbf{X}[\text{diag}(\boldsymbol{\Omega})]\mathbf{X}^T + \text{diag}(\sigma_e^2 \mathbf{I}_{n_j}). \quad (3.22)$$

Definiendo

$$\text{Var}(\boldsymbol{\varepsilon}) = \mathbf{R} = \text{diag}(\sigma_e^2 \mathbf{I}_{n_j}) \quad \text{y} \quad \text{Var}(\mathbf{u}) = \text{G} = \text{diag}(\boldsymbol{\Omega}), \quad (3.23)$$

la matriz de varianzas y covarianzas tiene la forma:

$$\mathbf{V} = \text{Var}(\mathbf{Y}) = \mathbf{XGX}^T + \mathbf{R}. \quad (3.24)$$

3.3.8. El coeficiente de correlación intraclase

Uno de los indicadores que se calculan a través de los componentes de la varianza en un modelo multinivel de intercepto aleatorio, es el coeficiente de correlación intraclase. Este coeficiente mide el punto en el cual los valores de Y en las observaciones de un mismo nivel, se asemejan entre sí, comparada con aquéllas observaciones de diferentes grupos. Se obtiene al dividir la variabilidad entre los niveles o grupos y la variabilidad total; es decir:

$$CCI = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_e^2}$$

De esta manera, el coeficiente de correlación intraclase representa la proporción de la variación total de los residuos que es explicada debido a las diferencias entre los grupos. También se conoce como el Coeficiente de partición de la varianza, VCP, por sus siglas en inglés (Variance Partition Coefficient). El coeficiente puede tomar valores entre 0 y 1; si es igual a 0 significa que no hay diferencias entre los grupos, y si es igual a 1, no hay diferencias dentro del grupo. Supóngase que se obtiene un coeficiente de correlación intraclase de 0.3, esto quiere decir que el 30% de la variación de los datos se da entre los grupos y el 70% entre las unidades de nivel 1.

3.3.9. Análisis de residuos

Como se ha mencionado en los modelos especificados anteriormente, se estableció que los componentes aleatorios deben cumplir ciertos supuestos para validar el modelo. Esto es, se

asume que u_j y ε_{ij}, se distribuyen normal con media cero y varianza constante, $u_j \sim N(0, \sigma_{u_0}^2)$ y $\varepsilon_{ij} \sim N(0, \sigma_e^2)$. El cumplimiento de este supuesto se realiza a través del análisis gráfico de los residuos.

El residuo es la diferencia entre el valor observado de Y y el valor esperado $\hat{y}$. En un modelo de regresión ordinaria, se estiman los residuos simplemente obteniendo la diferencia $y - \hat{y}$. En el caso del modelo multinivel de coeficientes aleatorios, como se tienen residuos en cada nivel, se necesita de un procedimiento un poco más complejo. Supóngase que y_{ij} es el valor observado del *i*-ésimo individuo para el *j*-ésimo grupo, mientras que $\hat{y}_{ij}$ representa los valores esperados de la regresión. Los residuos primarios son $r_{ij} = y_{ij} - \hat{y}_{ij}$. El residuo primario para el j-ésimo nivel es el promedio de r_{ij} para los individuos de cada nivel (r_{+j}). Por lo tanto, los residuos en el nivel 2 se obtienen multiplicando r_{+j}, por el siguiente factor:

$$\hat{u}_{0j} = \frac{\hat{\sigma}_{u0}^2}{\hat{\sigma}_{u0}^2 + \hat{\sigma}_e^2 / n_j} \, r_{+j},$$

donde n_j es el número de unidades dentro de cada nivel.

Este multiplicador r_{+j}, se conoce como el residuo reducido y siempre será menor o igual que 1. Una vez estimados los residuos a nivel 2, se pueden estimar los residuos a nivel 1 por la siguiente fórmula:

$$\hat{\varepsilon}_{ij} = r_{ij} - \hat{u}_{0j}$$

Los paquetes estadísticos especializados en modelación multinivel, calculan los residuos, tanto los crudos como los estandarizados, para todos los niveles. También se obtiene gráficos como el que se muestra en la Figura 3.14, en el que se corroboran si se distribuyen como una normal.

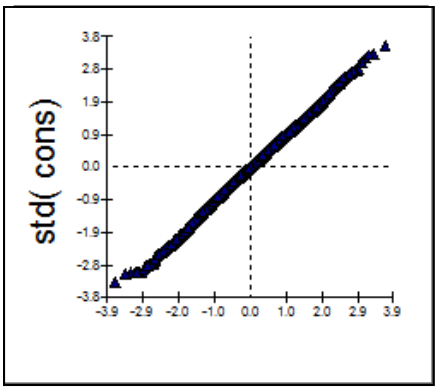


Figura 3.14. Gráfico de los residuos estandarizados.

3.3.10. Software para modelación multinivel

En los apartados anteriores se presentaron de manera sintetizada los fundamentos de la modelación multinivel y se mencionaron las ventajas que representa este tipo de metodología, así como el por qué estos modelos han adquirido especial relevancia en los últimos años. Sin embargo, como señalan Bryk y Raudenbaush (1992), el auge de la modelación multinivel se debe a las nuevas aplicaciones que se han desarrollado y a la gran cantidad de software disponible para su ejecución. Dentro de los primeros paquetes estadísticos que aparecieron para modelación multinivel, se encuentran: HLM de SSI (Hierarchical Linear and Nonlinear Modelling de Scientific Software International por sus siglas en Inglés), Proc Mixed de SAS (Statistical Analysis System) y Mlwin, (Multilevel Modelling for Windows), desarrollado en el Centro de Modelación Multinivel en la Universidad de Bristol (Rasbash *et al*., 1989)[5], por mencionar los más relevantes. Todos estos programas fueron incorporando nuevos algoritmos para el desarrollo de los modelos y actualmente se tiene versiones más actualizadas de ellos, HLM 7.0 y Mlwin 2.17.

En esta sección, se utilizará el software Mlwin, por su accesibilidad en el manejo de comandos, para exponer cómo se maneja una base de datos jerárquicos y cómo se procesan los datos al ajustar el modelo, así como la interpretación de la salida y las herramientas disponibles para su análisis. Los datos que se utilizan en este ejemplo fueron tomados del Informe del Programa de las Naciones para el Desarrollo (PNUD, 2005) en México. La

[5] http://www.bristol.ac.uk/cmm/

base contiene información sobre el Índice de Desarrollo Humano (IDH)[6] de los 2 443 municipios con los que cuenta el país, pertenecientes a las 32 entidades federativas, por lo que se tiene una estructura de anidamiento de los datos, al estar los municipios agrupados en estados (véase Figura 3.15). De acuerdo a la Organización de las Naciones Unidas, el IDH es un indicador que mide la calidad de vida del ser humano en el medio en el que se desenvuelve. El objetivo de este ejercicio será analizar si factores como el Índice de Potenciación de Género (IPG), que mide el grado de participación activa (económica y política) de hombres y mujeres en el país; el Índice de Empleo, la población (POB) y el Ingreso per cápita (INGPC) de cada municipio contribuyen a explicar las diferencias en el Índice de Desarrollo Humano alcanzado por cada uno. Asimismo, se estudiará si el contexto en el que se encuentra el municipio también tiene una incidencia en este índice, es decir, ¿Los municipios registran un mayor o menor índice de IDH de acuerdo al estado al que pertenecen?, ¿Las entidades federativas que tiene mayor riqueza del país, por el Producto Interno Bruto (PIB) que generan, cuentan con municipios con IDH más altos? Las respuestas a estas preguntas, se obtendrán al aplicar un modelo de dos niveles utilizando como herramienta el software Mlwin.

El primer paso es familiarizarse con el programa para abrir o capturar la base de datos que se utilizará para este ejercicio. Debido a que el software MlwiN está diseñado para trabajar en un sistema operativo de Windows, su funcionamiento también es a través del uso de ventanas; al iniciar el programa aparece una interface que muestra el menú y la barra de tareas, como en cualquier otra paquetería (Véase Figura 3.15).

[6] El objetivo de este ejercicio es presentar un ejemplo de aplicación de los modelos multinivel utilizando un software estadístico especializado. Para mayor información sobre cómo se mide el Índice de Desarrollo Humano, véase www.undp.org.mx.

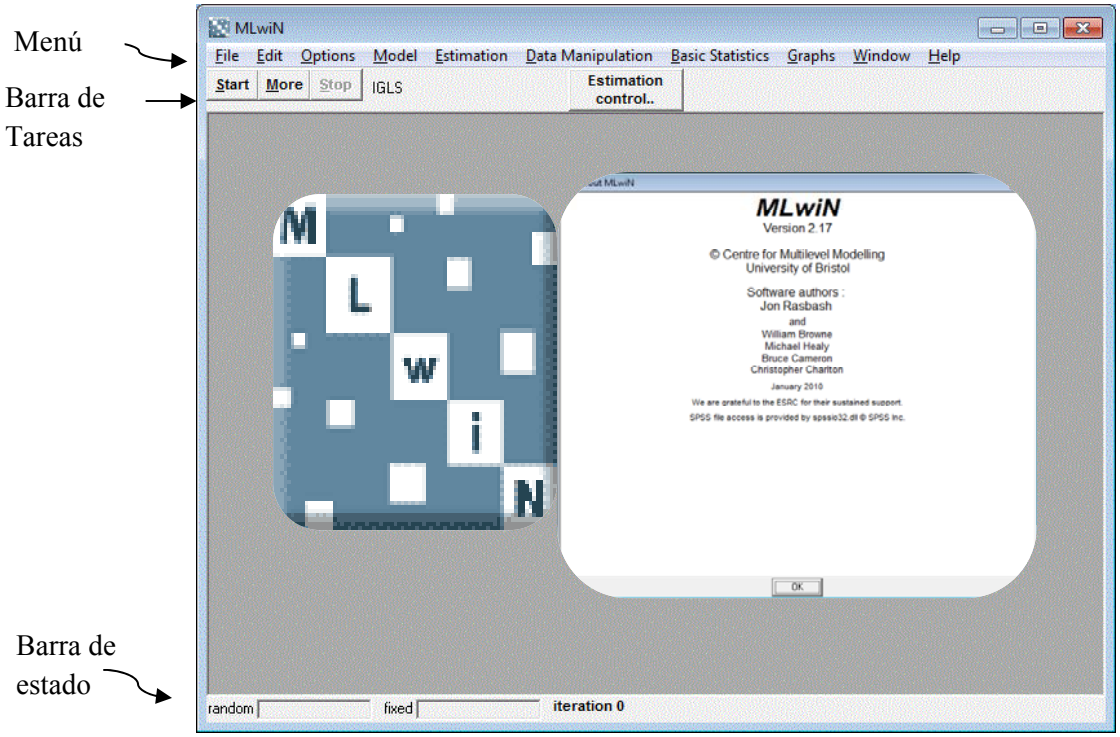


Figura 3.15. Ventana principal del software Mlwin.

En la barra de tareas, aparecen los botones relacionados con la estimación y control del modelo, que se utilizarán más adelante, debajo de ésta aparece la superficie de trabajo donde se irán abriendo las ventanas con la base de datos, la estimación del modelo, gráficos, etc., de acuerdo a la tarea que se vaya a realizar. Finalmente, aparece la barra de estado, que indica el progreso del procedimiento de estimación iterativo del modelo.

Ahora bien, hay tres formas principales de disponer de los datos para trabajarlos en el programa. Una de ellas es a través de la importación del archivo, es importante mencionar que Mlwin sólo puede importar o exportar datos numéricos. La versión 2.17 del programa, admite la importación de archivos de Stata (*.dta), SPSS (*.sav), y Minitab (*.mtw), permitiendo también guardar las hojas de trabajo en estos formatos[7].

La segunda opción es capturar los datos directamente en la hoja de trabajo, a través del menú de Manejo de Datos *(Data Manipulation*) o, finalmente, copiando los datos de otra paquetería como EXCEL o SPSS, utilizando el menú de Edición (*Edit*). Si se opta por esta vía, lo recomendable primero es configurar en la hoja de trabajo (*worksheet*), el

[7] También puede leer archivos de texto en formato ASCII, pero requiere un procedimiento especial. Véase Rasbash (2010), A user´s guide to Mlwin.

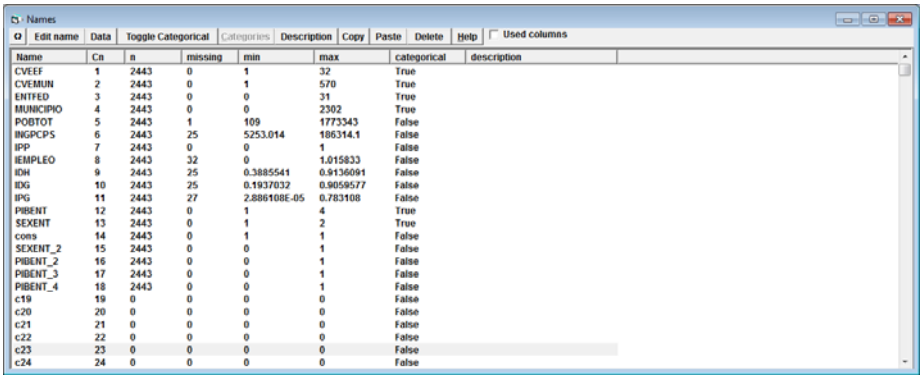
número de filas y columnas que se pegarán cuando se trata de bases de datos muy grandes. El procedimiento que se sigue es:

1. Ir al menú de Edición (*Edit*)
2. Seleccionar la opción Pegar (*Paste*)
3. Aparecerá la ventana *Paste View Window.*
4. Marcar la opción Usar la primera fila como nombre de variable (*Use first row as names*), en caso de que así se haya capturado en la base original.
5. Seleccionar Pegar (*Paste*).

Para este ejercicio, se cuenta con la base de datos en SPSS, por lo que se optó por importar el archivo:

1. Ir al menú de Archivo (*File*).
2. Seleccionar abrir hoja de trabajo (*Open worksheet*).
3. Seleccionar la ubicación del archivo que se desea importar.
4. Abrir (*Open*).

Al realizar los pasos anteriores, aparecerá la ventana de trabajo siguiente, que es la misma que aparece en cualquier forma que se haya utilizado para introducir los datos.



Name	Cn	n	missing	min	max	categorical	description
CVEEF	1	2443	0	1	32	True	
CVEMUN	2	2443	0	1	570	True	
ENTFED	3	2443	0	0	31	True	
MUNICIPIO	4	2443	0	0	2302	True	
POBTOT	5	2443	1	109	1773343	False	
INGPCPS	6	2443	25	5253.014	186314.1	False	
IPP	7	2443	0	0	1	False	
IEMPLEO	8	2443	32	0	1.015833	False	
IDH	9	2443	25	0.3885541	0.9136091	False	
IDG	10	2443	25	0.1937032	0.9059577	False	
IPG	11	2443	27	2.886108E-05	0.783108	False	
PIBENT	12	2443	0	1	4	True	
SEXENT	13	2443	0	1	2	True	
cons	14	2443	0	1	1	False	
SEXENT_2	15	2443	0	0	1	False	
PIBENT_2	16	2443	0	0	1	False	
PIBENT_3	17	2443	0	0	1	False	
PIBENT_4	18	2443	0	0	1	False	
c19	19	0	0	0	0	False	
c20	20	0	0	0	0	False	
c21	21	0	0	0	0	False	
c22	22	0	0	0	0	False	
c23	23	0	0	0	0	False	
c24	24	0	0	0	0	False	

La hoja de trabajo de Mlwin presenta la información de las variables en columnas, como se aprecia en la imagen anterior. Debajo de nombre (*Name*), aparecen las variables que contiene la base, el número de elementos de cada columna, que en este caso reporta

2,443 registros correspondientes a la información de cada uno de los municipios representados en el conjunto de datos. También, proporciona información sobre los datos faltantes (*missing values*), el valor mínimo y máximo registrado en cada variable, así como la indicación de si es categórica o no. En la base de datos, hay varios municipios de los cuales no se cuenta información sobre su IDH, por lo que aparecen datos faltantes.

El número de variables que se tienen para este análisis son 7, que se describen en la tabla 3.4, más una variable adicional que debe crearse, para poder ejecutar el modelo (Véase sección 3.3.7 de este capítulo). La variable contiene sólo una columna de 1 y se le llama cons. Uno de los comandos para crearla se presenta a continuación[8]:

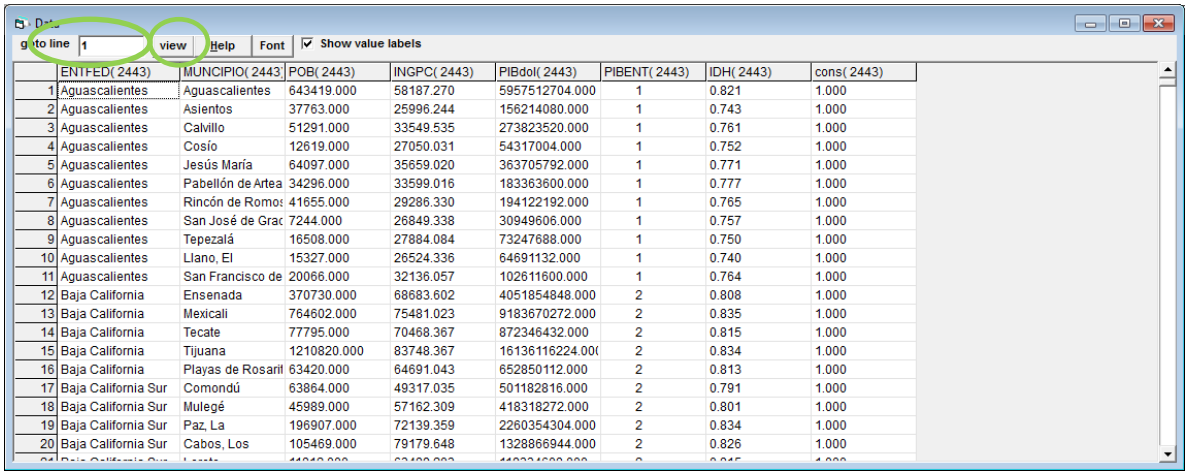
1. Ir al menú Archivo (*File*)
2. Seleccionar *New macro*.
3. Introducir la siguiente instrucción en la ventana:
   Code 1 1 2443 c8
   Name c8 ”cons”
4. Ejecutar (*Execute*)

Tabla 3.4. Descripción de las Variables incluidas en la base de datos idh.sav.

Variable	Descripción	Tipo
ENTFED	Identifica la entidad federativa	Categórica
MUNICIPIO	Identifica el municipio de cada entidad	Categórica
IDH	El Índice de Desarrollo Humano calculado para cada municipio en el año 2005.	De razón
POB	Población total de cada municipio	De razón
IPG	Índice de Potenciación de Género	De razón
INGPC	Ingreso per cápita promedio calculado para cada municipio en el año 2005.	De razón
PIBdol	Producto Interno Bruto de cada municipio en el año 2005 medido en dólares.	De razón
PIBENT	Clasificación de la entidad federativa de acuerdo al PIB que genera: 1PIB per cápita muy bajo (<-2σ) 2 PIB per cápita (-2σ <-1σ) 3 PIB per cápita . (-1 σ <1σ) 4 PIB per cápita (>2σ)	Categórica
Cons	Columna de unos.	

[8] Para conocer otros comandos, Véase (Snijders, 2003). Example session Mlwin.

Para poder ver los valores contenidos en cada variable de la base de datos, se debe ir al menú de Manejo de datos (*Data Manipulation*) y seleccionar la opción de vista o edición (*View or edit data*), y aparecerá la siguiente ventana. Utilizando la opción ver (*View*), se pueden seleccionar sólo las variables que se deseen consultar y con la opción ir a fila (*Go to line*), examinar algún caso en particular



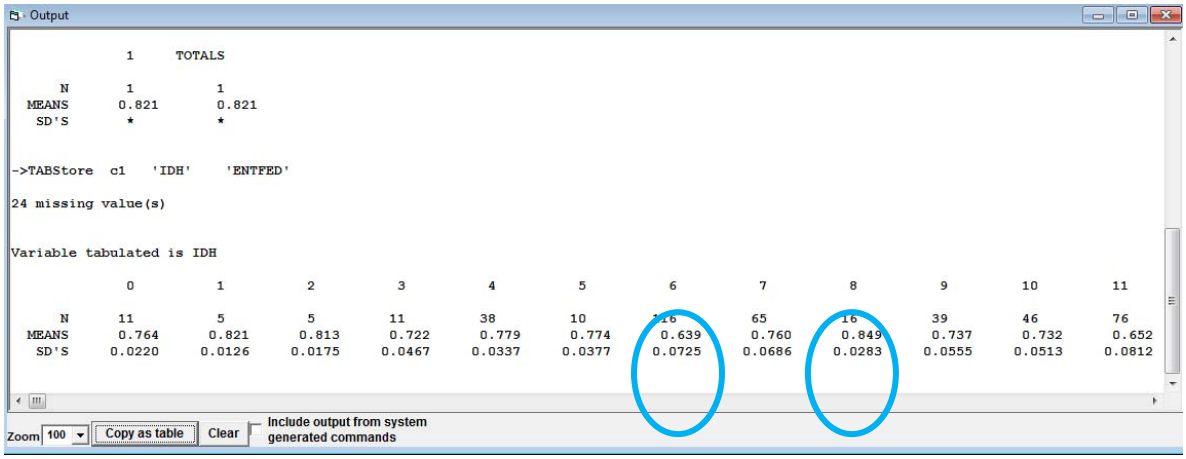
	ENTFED(2443)	MUNCIPIO(2443)	POB(2443)	INGPC(2443)	PIBdol(2443)	PIBENT(2443)	IDH(2443)	cons(2443)
1	Aguascalientes	Aguascalientes	643419.000	58187.270	5957512704.000	1	0.821	1.000
2	Aguascalientes	Asientos	37763.000	25996.244	156214080.000	1	0.743	1.000
3	Aguascalientes	Calvillo	51291.000	33549.535	273823520.000	1	0.761	1.000
4	Aguascalientes	Cosío	12619.000	27050.031	54317004.000	1	0.752	1.000
5	Aguascalientes	Jesús María	64097.000	35659.020	363705792.000	1	0.771	1.000
6	Aguascalientes	Pabellón de Artea	34296.000	33599.016	183363600.000	1	0.777	1.000
7	Aguascalientes	Rincón de Romos	41655.000	29286.330	194122192.000	1	0.765	1.000
8	Aguascalientes	San José de Grac	7244.000	26849.338	30949606.000	1	0.757	1.000
9	Aguascalientes	Tepezalá	16508.000	27884.084	73247688.000	1	0.750	1.000
10	Aguascalientes	Llano, El	15327.000	26524.336	64691132.000	1	0.740	1.000
11	Aguascalientes	San Francisco de	20066.000	32136.057	102611600.000	1	0.764	1.000
12	Baja California	Ensenada	370730.000	68683.602	4051854848.000	2	0.808	1.000
13	Baja California	Mexicali	764602.000	75481.023	9183670272.000	2	0.835	1.000
14	Baja California	Tecate	77795.000	70468.367	872346432.000	2	0.815	1.000
15	Baja California	Tijuana	1210820.000	83748.367	16136116224.00(	2	0.834	1.000
16	Baja California	Playas de Rosarit	63420.000	64691.043	652850112.000	2	0.813	1.000
17	Baja California Sur	Comondú	63864.000	49317.035	501182816.000	2	0.791	1.000
18	Baja California Sur	Mulegé	45989.000	57162.309	418318272.000	2	0.801	1.000
19	Baja California Sur	Paz, La	196907.000	72139.359	2260354304.000	2	0.834	1.000
20	Baja California Sur	Cabos, Los	105469.000	79179.648	1328866944.000	2	0.826	1.000

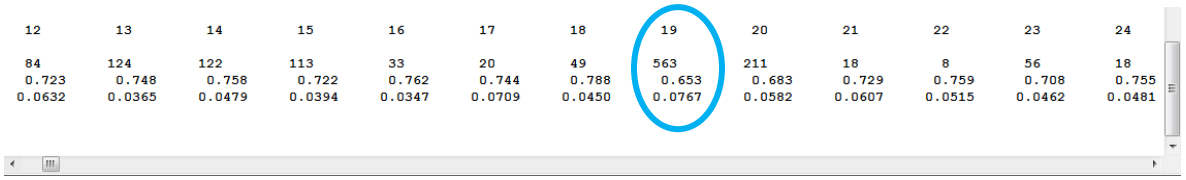
Habiendo completado la información de la base de datos, para explorarlos antes de proceder a la especificación del modelo, se recurre al menú de Estadísticas Básicas (*Basic Statistics*), donde se pueden tabular las frecuencias, promedios, obtener estadísticas descriptivas, correlaciones y comparación de medias entre grupos, a través del análisis de varianza (ANOVA). Por ejemplo, con la base de datos de IDH, se obtendrán los promedios del Índice de Desarrollo Humano de los municipios por entidad federativa. Para ello, se ejecutan los siguientes pasos:

1. Ir al menú de Estadísticas Básicas (*Basic Statistics*), seleccionar la opción Tabular (Tabulate).
2. En la opción de salida (*Output mode*), seleccionar la media (*Means*).
3. En la lista desplegable junto a la columna variable (*Variate Column*), seleccionar la variable de interés (*IDH*).
4. Dar click en Tabular (Tabulate).

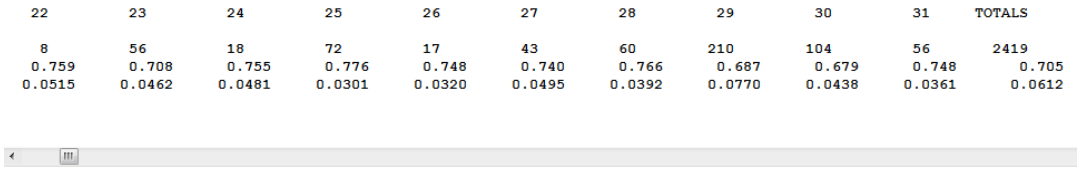
Aparecerá la ventana que se muestra a continuación, en la que se indica para cada grupo o entidad, el número total de unidades que agrupa (N), el promedio de la variable en

ese grupo (MEANS) y la desviación estándar (SD's). Los resultados se muestran de manera horizontal, por lo que para consultar el resto de las entidades se debe ocupar la barra de desplazamiento. En este caso, se observa que el grupo 0, la entidad Aguascalientes cuenta con un total de 11 municipios que presentan en promedio un Índice de Desarrollo Humano de 0.764, con una desviación estándar de 0.0220.





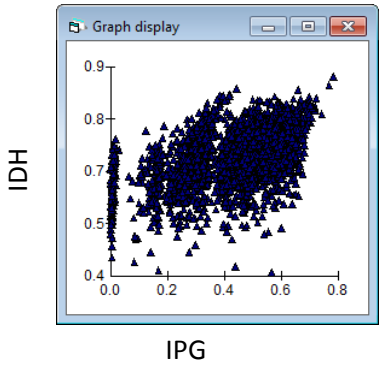
Visualizando toda la información, se observa que la entidad federativa que tiene el mayor número de municipios en el país es Oaxaca (grupo 19), con 563 más 7 municipios de los cuales no se tiene información de IDH, un total de 570 municipios. La entidad con el IDH más alto es el Distrito Federal (grupo 8), con un promedio de 0.849 y la entidad con el IDH más bajo, es Chiapas con 0.63 (grupo 6). Igualmente, se aprecia que las entidades con mayor dispersión en sus valores de IDH, son Chiapas, Veracruz y Oaxaca. Es decir, estas entidades cuentan con municipios tanto con alto desarrollo como con muy bajo desarrollo. Este análisis permite ir explorando el comportamiento de la variable de interés IDH a nivel entidad federativa.



Por otro lado, continuando con la exploración, Mlwin ofrece el menú de Gráficos (*Graphs*), para realizar gráficos que permitan observar la relación entre la variable respuesta, en este caso el IDH y las respectivas variables explicativas (se iniciará con el IPG[9]),El software no cuenta con muchos tipos de gráfico, como otros paquetes estadísticos (SPSS,Statistica,Stata), los gráficos que se elaboran en Mlwin, están primordialmente diseñados para observar relaciones entre variables, analizar el comportamiento de los residuos y la detección de datos atípicos. Si se desea elaborar gráficos como el diagrama de cajas y alambres, se debe recurrir a alguno de los paquetes mencionados anteriormente. El procedimiento a seguir para la elaboración de gráfico de dispersión entre el IDH y el IPG es:

1. Ir al menú Gráficos (*Graphs*).
2. Seleccionar gráfico personalizado (*Customised graph*).
3. En la ventana que aparecerá (*plot what?*), indicar la variable *Y,* y la variable *X* que se desea graficar.
4. Si se quiere agrupar, seleccionar el grupo en el menú desplegable (*Group*).
5. Las pestañas siguientes son para cambiar el formato del gráfico.
6. Dar click en Aplicar (*Apply*)

Una vez ejecutados los pasos anteriores, seleccionando el IDH (Variable *Y*) y el IPG (Variable *X*), el gráfico que aparece se muestra a continuación:

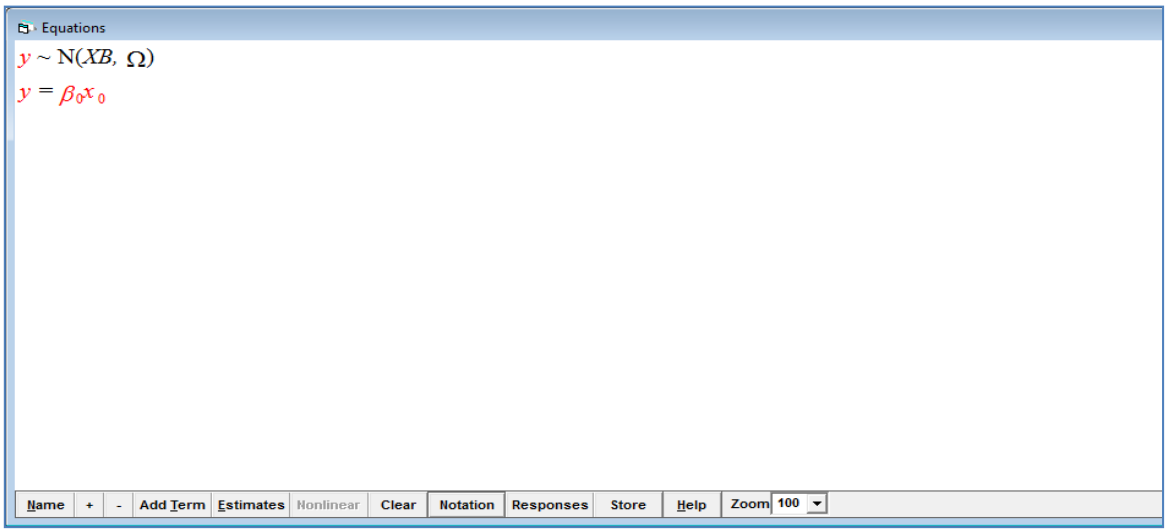


En el gráfico anterior, se observa que puede existir una relación directa entre el IDH y el IPG, lo que indica que las poblaciones donde la participación económica y política de hombres y mujeres es más igualitaria, presentan un Índice de Desarrollo Humano más alto.

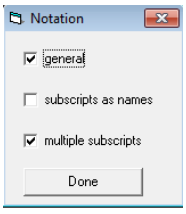
[9] El Índice de Potenciación de Género es un indicador social que mide el nivel de oportunidades de las mujeres. Por tanto, mide también las desigualdades en tres dimensiones de participación de las mujeres. Es elaborado por el Programa de las Naciones Unidas para el Desarrollo (PNUD

Sin embargo, para ver la magnitud del efecto y analizar si este hecho se presenta para todas las entidades federativas del país, a pesar de su diversidad y desigualdad económica, política y social, se especificará el modelo de dos niveles. En la hoja de trabajo, Mlwin presenta el menú Model (Modelos), que despliega todas las opciones relativas con la especificación del modelo: definición de la ecuación, comparación de modelos, predicciones, análisis de residuos, definición de la estructura jerárquica, etc.

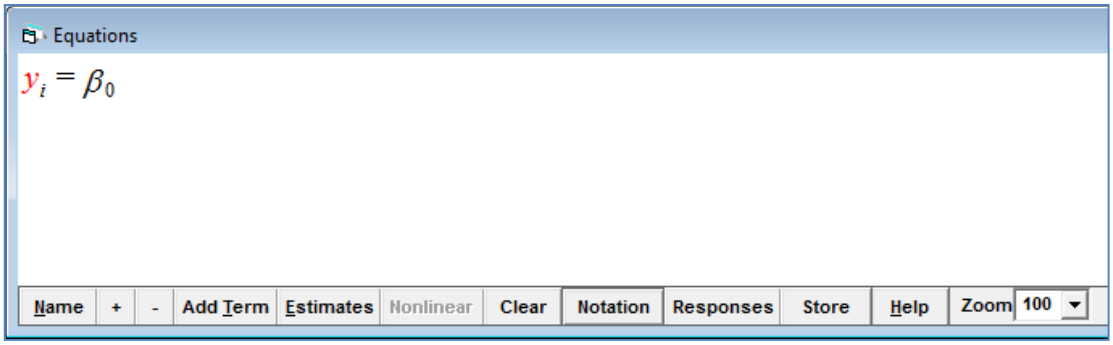
Se comenzará por plantear el modelo más sencillo, pero antes, se muestra la ventana de trabajo de la ecuación y las herramientas que se utilizan, donde se ajustan los modelos y aparecen sus resultados. Para ello, se debe ir al menú Modelo (*Model)* y seleccionar de la lista, la primera opción de Ecuaciones (*Equations*). Aparecerá la ventana siguiente:



Siempre que se abre la ventana, automáticamente aparecen los supuestos del modelo multinivel y las variables y parámetros en rojo, indicando que no han sido especificados, pero como se iniciará del modelo más sencillo, se cambiará este modo. Para ello, en la pestaña de Notación (*Notation*), se desmarcan las opciones que aparecen:



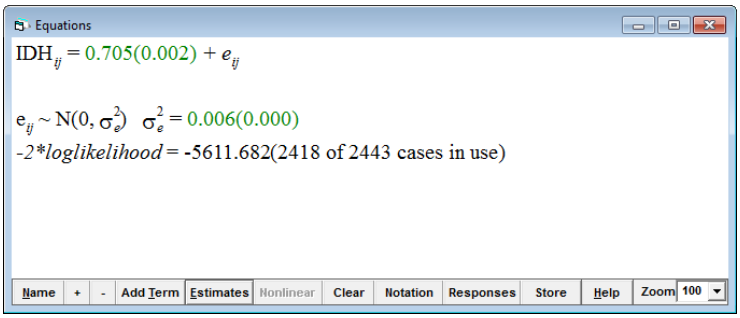
Y la ventana ahora se muestra de esta manera:



Las pestañas que aparecen en la parte inferior, se utilizan para lo siguiente:

- *Name*: Al hacer click, en la ventana aparecerán los nombres de las variables en lugar de la notación algebraica.
- *Add term*: Se utiliza para agregar las variables explicatorias.
- *Estimates*: Presenta los valores estimados de los parámetros.
- *Clear*: Limpia la pantalla.
- *Notation*: Para mostrar la notación más simple y para seleccionar que se reemplacen los subíndices *ij* por el nombre de los niveles.
- *Responses*: Con esta opción se trabajan modelos multinivel multivariados, cuando se tienen más de dos variables respuestas (Y).
- *Store*: Para guardar los resultados del modelo.
- *Help*: Muestra el menú de ayuda.
- *Zoom*: Para ampliar o disminuir el tamaño de la ventana.

Para introducir la variable respuesta, se da click directamente en Y, apareciendo una ventanita con un menú desplegable, en el que se selecciona la variable de interés (IDH), así como los niveles de clasificación según corresponda. En este ejercicio, sólo se utilizan dos: el municipio en el primer nivel y las entidades federativas como segundo nivel. Una vez realizada esta operación, se da click en iniciar la estimación (*Start*), ubicado en la barra de tareas y posteriormente en valores estimados (*Estimate*s), en la parte inferior de la ventana de trabajo. Se mostrarán los siguientes resultados:

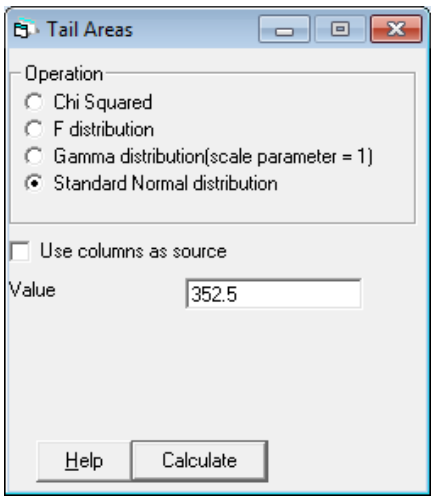


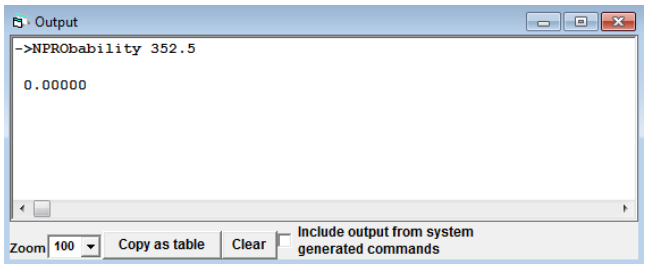
El modelo especificado no contiene ninguna variable explicatoria, sólo reporta el valor del intercepto. La cifra que aparece entre paréntesis es el error estándar del parámetro, que se utiliza para realizar la prueba de hipótesis que determina si el parámetro es significativamente diferente de cero. Es decir, si tiene o no efecto en el modelo. Para este ejercicio, el planteamiento es el siguiente:

$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0$$

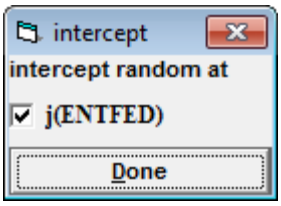
El test estadístico se calcula como $\hat{\beta}_0 / SE(\hat{\beta}_0)$. Utilizando los datos del ejercicio, esto equivale a 0.705/0.002= 352.2, valor que se compara con el valor de tablas de la Normal. Para ello, se recurre el menú de Estadísticas básicas (*Basic Statistics*) y se selecciona prueba de colas (*Tail Areas*). Se indica la distribución que interesa, en este caso la Normal stándar y se captura el valor que obtuvimos de la división. Al dar click en calcular (*Calculate*) aparecerá el resultado:



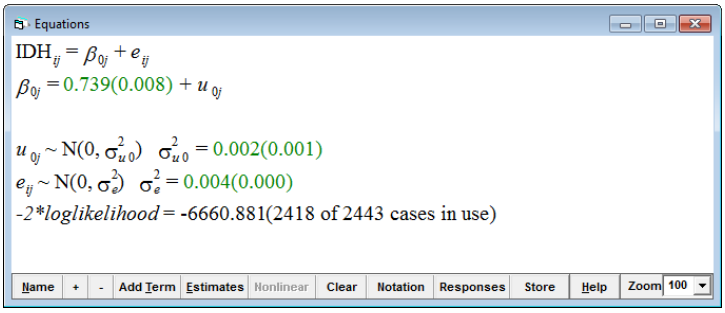


En este ejercicio, el valor que se obtuvo es 0.0000, por lo que se rechaza la hipótesis nula y se concluye que el intercepto es significativo. Esto quiere decir, que el promedio del IDH de los municipios mexicanos es de 0.750, con una varianza estimada de 0.006.

Ahora se introducirá el segundo nivel en el modelo, para ver la variabilidad que existe entre las entidades federativas. Para ello, se da click directamente en el parámetro β, y se indica marcando el recuadro, que el intercepto varíe entre las unidades del segundo nivel (entidades federativas).



El modelo cambiará, al introducirse automáticamente la variabilidad en el segundo nivel (σ^2_{u0}.), como se muestra a continuación:

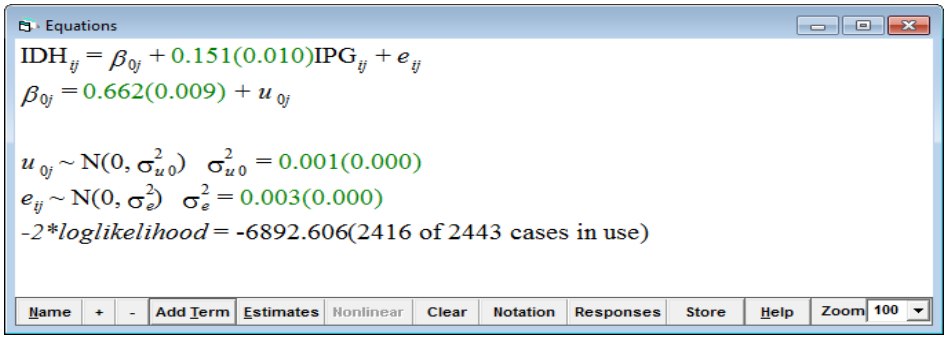


Se aprecia que el valor del intercepto también cambió, pero sigue siendo significativo, lo que no sucede con las varianzas en ambos niveles. Aplicando el procedimiento anterior, se obtiene que la variabilidad entre los municipios no resulta significativa y la variabilidad entre las entidades sí a un 2%[10]. Es decir, no hay diferencias significativas entre los valores del Índice de Desarrollo Humano de los municipios en las entidades federativas del país.

El siguiente paso será añadir la variable explicativa, Índice de Potenciación de Género (IPG), para determinar si influye en el IDH. El modelo que se especificará es un modelo de dos niveles con el intercepto aleatorio, lo que significa que el promedio del IDH

[10] Obteniendo el valor del estadístico de prueba y comparando con el valor de tablas, no se rechaza a un 0.0227.

variará en cada entidad federativa, por la cantidad de σ^2_{u0}, la cual se asume se distribuye como una normal (Véase sección 3.3.6).



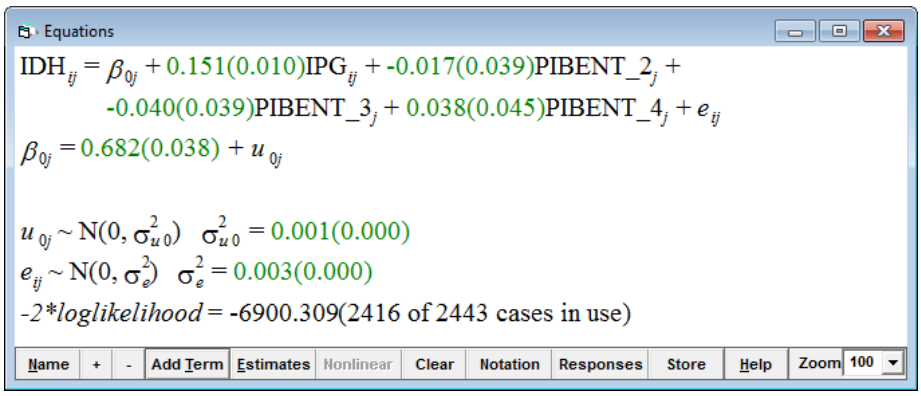
Como se aprecia en el recuadro, la variable Índice de Potenciación de Género (IPG), resulta significativa e influye positivamente en el IDH de los municipios mexicanos. Es decir, al aumentar en una unidad la participación igualitaria en hombres y mujeres, a través del IPG, se esperaría mejorar el IDH de las poblaciones en 0.151unidades y por tanto se eleva la calidad de vida. Esto significa que hay que reducir la brecha de desigualdad entre hombres y mujeres, en los distintos ámbitos de participación ciudadana de los individuos para alcanzar un mejor desarrollo humano.

Se calcula el Coeficiente de correlación intraclase como se describió en el aparatado 3.3.8, para conocer la variabilidad que es explicada por el modelo a nivel de las entidades. En este caso, el coeficiente es igual a $\widehat{CCI} = \frac{\hat{\sigma}^2_{u0}}{\hat{\sigma}^2_{u0}+\hat{\sigma}_e} = \frac{0.001}{0.001+0.003} = 0.25$. Por lo que se tiene, que alrededor de un 25% de la variabilidad en el IDH de los municipios mexicanos puede ser atribuida a las diferencias entre las entidades federativas.

El valor de la prueba de verosimilitudes (loglikelihood ratio test), permite comparar los modelos. Si se obtiene la diferencia entre los valores obtenidos, -6660.88-(-6892.606) = 232, el cual es comparada con una distribución χ^2 con un grado de libertad, se concluye que existe una variación entre las entidades federativas.

Una de las aportaciones de la modelación multinivel frente a otras técnicas de modelación, es que permite conocer los efectos contextuales. Es decir, si variables a nivel de grupo tienen un efecto en los resultados del nivel 1. En este ejercicio, permitirá analizar si variables a nivel entidad federativa, contribuyen a explicar las diferencias entre el IDH de los municipios del país. Se utiliza la variable PIB, clasificado en 4 categorías, que indican

la riqueza del estado: Entidades federativas con clasificación 4, reportan un valor de PIB muy alto (2σ por encima de la media), contrario a las entidades con valor de 1 (véase Tabla 3.4).



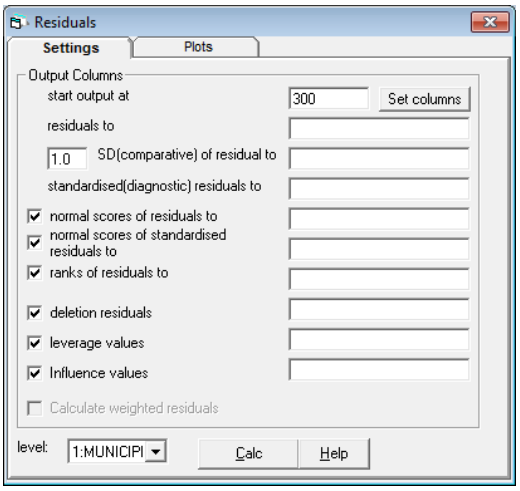
En la salida del modelo anterior, se observa[11] que ninguna de las categorías (variables dummy) relativas al PIB de la entidad, resultaron significativas, lo que se traduce en que hay municipios con un alto IDH y viceversa, tanto en entidades con PIB altos, como en entidades con PIB bajos.

Para validar los resultados obtenidos en el modelo y saber si se están cumpliendo las premisas de linealidad en las relaciones, se deben cumplir los supuestos de normalidad en los residuos. De acuerdo con McCafe y Moore (2000), se deben siempre examinar los residuos como una ayuda para determinar si el modelo de regresión es apropiado a los datos. Para verificar ese supuesto de normalidad en el modelo multinivel que se planteó, en Mlwin primero se deben calcular los residuos a través del siguiente procedimiento:

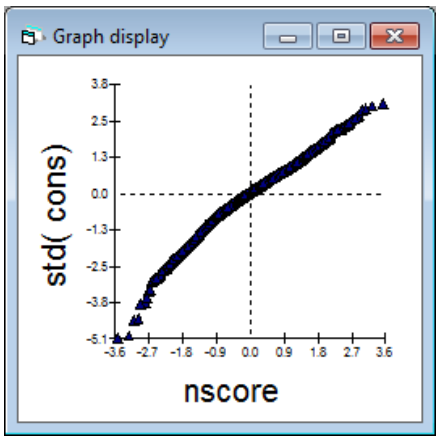
1. Ir al Menú del Modelo (*Model*)
2. Seleccionar la opción de Residuos (*Residuals*).
3. Indicar las columnas en las que se presentaran el resultado de los cálculos del residuos.
4. Seleccionar el nivel de análisis en la parte inferior de la ventana.
5. En la pestaña *Plots*, seleccionar el gráfico que se desea inspeccionar.
6. Dar click en Calcular (*Calc*)

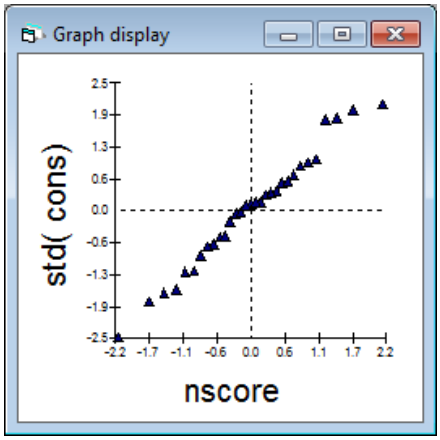
Aparecerá la siguiente ventana:

[11] Se realiza el procedimiento de contrastar el valor del estadístico de prueba con el valor de tablas

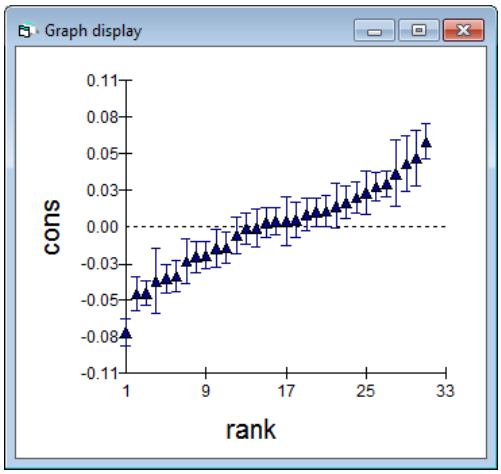


Si se seleccionó graficar los residuos estandarizados contra los valores normales, se producirán los gráficos de estilo Q-Q plot, para verificar que siguen una distribución normal, esto se realiza a través de una inspección visual para observar que los datos se ajustan a una línea recta. Los gráficos que se calcularon para este ejercicio, se presentan a continuación. El primero muestra los residuos a nivel de los municipios y el segundo a nivel de las entidades federativas. Se puede apreciar que los datos se ajustan a una línea recta, por lo que corroboran el supuesto de normalidad.





También se puede generar el llamado gráfico de oruga (*Caterpillar plot)*, que permite graficar los residuos en orden ascendiente, con un intervalo de confianza del 95%.



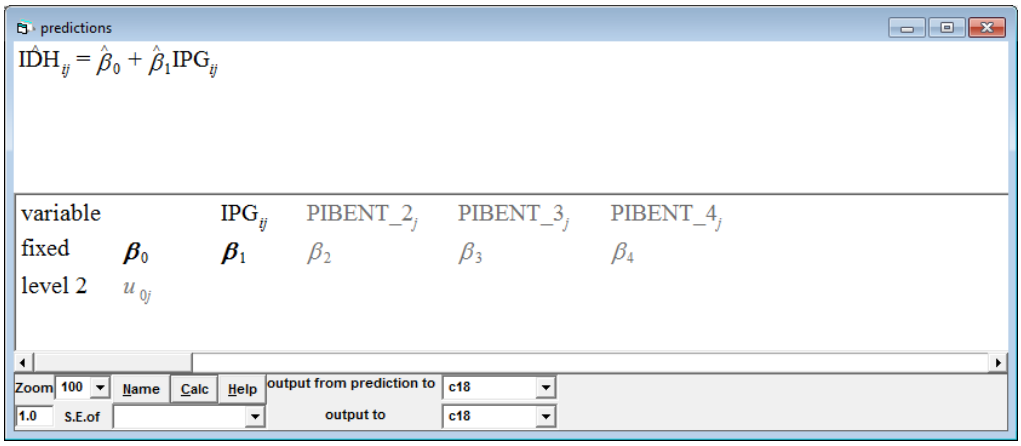
En el gráfico anterior, se observan los residuos correspondientes a las 32 entidades federativas del nivel 2. Mirando los intervalos de confianza alrededor de estas, se puede observar un grupo de 15 entidades federativas, al inicio y al final del gráfico, que no se traslapan con el 0, lo que significa que estas entidades difieren significativamente del promedio a un 5%. Es importante tener en cuenta que los residuos representan la diferencia entre la media de cada entidad y el promedio general pronosticado por el parámetro fijo β_0.

El modelo que se especificó para llevar a cabo este ejercicio, fue un modelo de dos niveles con intercepto aleatorio (Véase 3.3.6, sección b). Una vez que se calcularon los residuos, se pueden elaborar las predicciones para cada entidad federativa. El primer paso es calcular la línea de predicción promedio que se genera de la parte fija del modelo: el intercepto y la pendiente, β_0 y β_1, respectivamente. La ecuación de predicción es:

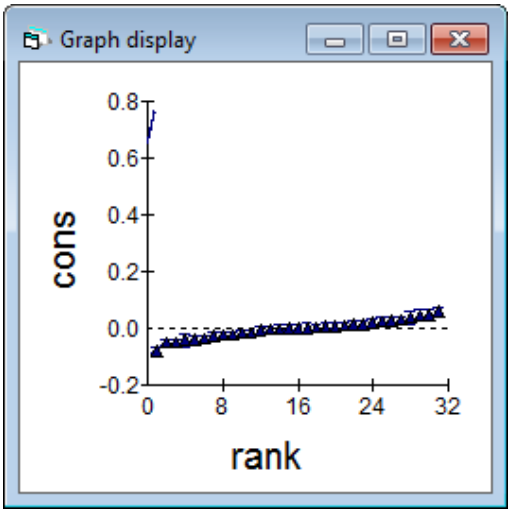
$$\widehat{IDH}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 IPG_{ij}$$

Para calcularla en Mlwin, se siguen estos pasos:

1. Ir al menú Modelo (*Model*) y seleccionar Predicciones (*Predictions*).
2. En la ventana que aparecerá, dar click en los parámetros de interés β_0 y β_1. No se incluye la parte aleatoria.
3. Especificar el número de columna en que se generarán los valores pronosticados en *Output to.*
4. Dar click en Calcular



El siguiente paso será graficar las predicciones que se obtuvieron con los residuos calculados anteriormente. Para ello, se debe ir a Gráficos personalizados (*Customised Graphs*) y seleccionar las columnas donde tenemos los respectivos valores.



Los valores de la línea recta graficada son:

$$\hat{IDH}_{ij} = 0.662 + 1.51 IPG_{ij}$$

La línea de predicción para la entidad federativa *j*-ésima parte de la recta anterior + una cantidad $u_{0j.}$ Este residuo del segundo nivel modifica el valor del intercepto, más no el de la pendiente porque se determinó como fijo. Por lo tanto, las líneas de predicción para las 32 entidades federativas deben ser paralelas. La ecuación de predicción para la *j*-ésima entidad federativa es:

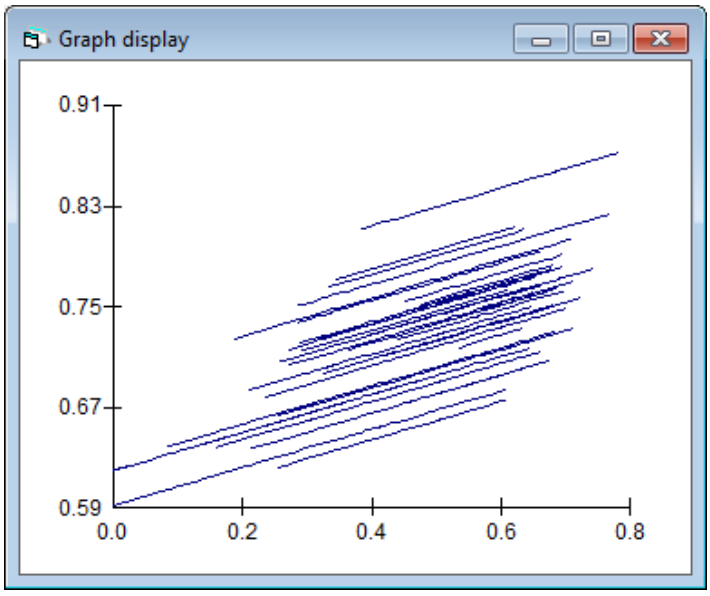
$$\widehat{IDH}_{ij} = (0.662 + \hat{u}_{0j}) + 1.51IPG_{ij}$$

Para conocer la línea de predicción de las entidades de Aguascalientes y Baja California por ejemplo, sólo se deben sustituir los valores en la ecuación anterior. El valor u_{oj} corresponde a los residuos de nivel 2. En la hoja de trabajo, en la sección de Ver datos (*View or edit data*) se pueden obtener estos valores de los residuos, dependiendo la columna en la que se hayan calculado. Para este ejercicio los valores y las ecuaciones generadas para ambas entidades son:

$$\widehat{IDH}_{ij} = (0.662 + 0.016) + 1.51IPG_{ij} \text{ Aguascalientes}$$

$$\widehat{IDH}_{ij} = (0.662 + 0.058) + 1.51IPG_{ij} \text{ Caja California}$$

Con esta información, se pueden calcular las líneas de predicción para las 32 entidades federativas y tener un panorama de la variabilidad que existe entre las entidades federativas:



Referencias

Barnet, V. y Lewis, T. (1994). *Outliers in Statistical Data*. 3rd. Edition. Wiley, NewYork.

Bryk, A.S. y Raudenbush, S.W. (1992). *Hierarchichal Linear Models*. Newbury Park, California: Sage.

Chatfield, C. (1995). *Problem Solving: a Statistician's Guide*. Second edition, Champman and Hall, London, UK.

De Leeuw, J. y Meijer, E. (2008). *Handbook of Multilevel Analysis*. Springer, New York.

Dempster, A.P.; Laird, N.M y Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, B, 39(1): 1–38.

Fox, J. (1997). *Análisis Factoriales Simples y Múltiples*. Universidad del País Vasco, España.

Goldstein, H. (1999). *Multilevel Statistical Models*. London. First Internet Edition.

Hair, F. J., Anderson, R. E., Tatham, R. L. and Black, W. C. (1999). *Analysis Multivariante*. Prentice-Hall, Inc., New York.

Hox, J.J. (2002). *Multilevel Analysis: Techniques and Applications*. Rotuledge. Taylor and Francis Group.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. Academic Press, Inc., New York.

Johnson, D. E. (1999). *Métodos Multivariados Aplicados al Análisis de Datos*. International Thompson Editores, México.

Johnson, D. E. (2000). *Métodos Multivariados Aplicados al Análisis de Datos*. Internacional Thomson Editores, S.A. de C. V., México.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariante Statistical Analysis*. Prentice Hall, Upper Saddle River, N. J., USA.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford. Clarendon Press.

McCabe, G y Moore, D. (2000) *Introduction to the Practice of Statistics*, 3ra. Edición, New York, W. H. Freeman and Company.

Montgomery, D., Peck, E. y Geoffrey Vining. (2004). *Introducción al Análisis de Regresión Lineal*. CECSA. Primera reimpresión. México.

Morrison, D. F. (1990). *Multivariate Statistical Methods*. Third Edition. McGraw Hill, New York.

Ojeda, M. M. (1988). Análisis de datos. *La Ciencia y el Hombre* (1), 121-133.

Ojeda, M. M. (1994). La importancia de una buena cultura estadística en la investigación. *La Ciencia y el Hombre* (17), 143-152.

Ojeda, M.M. y Behar, R. (2006). Estadística, productividad y calidad. SEV, Gobierno del Estado de Veracruz.

Ojeda, M. M., Díaz, C. J. E., Apodaca, C., Trujillo, I. (2004). *Metodología de Diseño Estadístico*. Xalapa, México: Universidad Veracruzana.

Ojeda, M. M., y Velasco, F. (2010). Aplicaciones de la Estadística en el Área Biomédica. *Investigación aplicada a la salud. Una mirada desde la investigación de operaciones*. P. 1-23 México: Editorial Ultradigital Press, S. A.

Ojeda, M.M. (1993). *La modelación estadística*. Foro de Matemáticas del Sureste. Universidad Juárez Autónoma de Tabasco.

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw Hill, Madrid.

Rasbash, J. (2008). *Multilevel Structures and Classifications*. Centre for Multilevel Modelling. University of Bristol. New York, W. H. Freeman and Company: 724.

Rasbash, J., Steele, F., Browne, W.J. y Goldstein, H. (2009). *A User's Guide to Mlwin*. Version 2.10. Centre for Multilevel Modelling. University of Bristol.

Rencher, A. C. (1995). *Methods of Multivariante Analysis*. John Wiley, New York.

Robinson, W.S. (2009). Ecological correlations and the behaviour of individuals. *International Journal on Epidemiology*. 38 (2): 337-341.

Seber, G. A. F. (1990). *Multivariate Observations*. John Wiley, New York.

Sharma, S. (1996). *Applied Multivariante Techniques*. John Wiley, New York.

Skrondall, A. y Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* Champman &Hall /CRC. First Edition.

Snijders, T.A.B. (2005). Power and Sample Size in Multilevel Linear Models; in B.S. Everitt and D.C. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science*. Chicester (etc.): Wiley (3): 1570-1573.

Snijders, T.A.B. and Bosker, R.J., (1993). Standard errors and sample sizes for two-level research, *Journal of Educational Statistics* (18), 237-259.

Steele, F. (2008). *Introduction to Multilevel Modelling Concepts*. Centre for Multilevel Modelling. University of Bristol.

IV. Artículos

Olivera- Gómez, D. A. y Cruz-López, C. (2011). Construcción de un índice de competencias para el desarrollo de un modelo de atención empresarial. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Cruz-López, C. y Herrera-Jiménez, J. (2011). Análisis de la industria del calzado en el periodo 1999-2009. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Eguía-Casis, A. y Zavaleta-Sánchez, Y. (2011). Análisis del mercado ocupacional en México durante el periodo 2005-2009. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Mayo-Lara, D. y Velasco-Luna, F. (2011). Análisis del gasto en salud y su relación con el crecimiento económico de México en el periodo 2000-2008. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Nuñez-Herrera, O. O. y Velasco-Luna, F. (2011). Influencia del sector eléctrico y petrolero en la producción primaria 2003-2008. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Abad-Espíndola, A. y Tapia-Blasquez, P. (2011). Evaluación del Fondo de Aportaciones para la Infraestructura Social Municipal (FAISM) en el combate al rezago en infraestructura social de los municipios indígenas de Veracruz en el periodo 2000-2005. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Ojeda-Ramírez, M. M. y González-Hernández, A. (2011). Un análisis del impacto del Programa de Apoyos Directos al Campo (PROCAMPO) en la productividad del campo veracruzano, periodo 2002-2008. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Galván-Herrera, A. A. y Tapia-Blasquez, P. (2011). Becas PRONABES: Una mirada a su evolución e impacto en el fortalecimiento del Desarrollo Humano 2002-2007. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Gallardo-del Angel, R. y Aguilar-López, A. (2011). Causalidad entre los ingresos y egresos de los gobiernos locales de México. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Tapia-Blasquez, P., Ojeda-Ramírez, M. M. y Tapia-Blasquez, E. (2011). Efecto de los contextos escolares en los resultados de la prueba ENLACE 2009: Un Análisis multinivel. *Metodología Estadística Aplicada a la Finanzas Públicas*.

Construcción de un índice de competencias para el desarrollo de un modelo de atención empresarial

Daniel Armando Olivera Gómez
Cecilia Cruz López

Resumen

En esta investigación se realiza un análisis de componentes principales para crear un modelo de atención, dirigido a la mejora de las competencias en la gestión empresarial. Este enfoque toma en cuenta que el entorno empresarial es multivariable y complejo, y priorizará la incidencia de las variables estudiadas en la supervivencia y desarrollo de las empresas en el corto, mediano y largo plazo. El tratamiento estadístico permitió reducir la dimensión de los datos obteniéndose una tipología empírica del perfil del empresario MIPYME en Veracruz. Los datos empleados corresponden a una base de datos de un muestreo probabilístico estratificado. Lo anterior permitió construir una escala para evaluar y/o comparar las necesidades de formación en las competencias a partir de rasgos cuantitativos referentes a las variables de estudio y las prioridades de formación que mejor incidan el éxito de la empresa. Se obtuvo un solo índice final que permitió la clasificación, el ordenamiento y la estimación de las diferencias entre las competencias de las MIPYME veracruzanas a través de un único valor representativo para la delineación de un modelo de atención basado en la formación de competencias para los empresarios.

Palabras Clave*: Análisis de Componentes principales, Gestión empresarial, Formación empresarial.*

Abstract

In this investigation an analysis of main components (AMP) is realised to create a model of attention directed to the improvement of the competitions in enterprise management (CEM) in this sector, that takes into account that the enterprise surroundings are multi-variate and complex, and therefore it would prioritize the incidence of those variables in the survival and development of the companies in the short, medium and long term. The statistical treatment allowed to synthesize and to reduce the data to do them comprehensible and to be used to obtain an empirical typology of the profile of the MIPyME enterprises in Veracruz; for it the one data base of a stratified probabilistic sampling of the investigation of Olivera *et al* in 2009 was used. The previous thing allowed to construct a scale to evaluate and/or to compare the needs of formation in the Cge from referring quantitative characteristics to the study variables and the priorities of formation that affect the success of the company more. A single final index was obtained that allowed the classification, the ordering and the estimation of the differences between the CEM of Veracruz´s MIPyME through a unique representative value for the delineation of a model of attention based on the formation of competitions for the enterprises of the sector in the management.

Keywords*: Training, Competitions, Principal Component Analisys.*

Introducción

Por la participación del sector de la Micro Pequeña y Mediana Empresa (MIPyME) en la economía estatal, su desarrollo se vincula a los múltiples beneficios que tiene en lo social y económico, principalmente en aspectos como la disminución del desempleo, la reducción de problemas sociales, el consumo y la reactivación económica. A pesar de ello, en la investigación de Olivera *et al* (2009) se detectó en este sector vulnerable, el poco desarrollo de las Competencias en Gestión Empresarial (CGE) -*Dimensión estratégica; Dimensión Técnica; Dimensión Mercadológica; Dimensión Capital Humano; Dimensión fiscal; Dimensión Responsabilidad Social Empresarial (RSE) y Dimensión Información*- convirtiéndose en un factor determinante para su supervivencia y desarrollo.

Ante esta carencia, se decidió realizar una investigación retomando la base de datos usada en Olivera *et al* (2009), que determinaba la factibilidad de la implementación del Departamento de Investigación Aplicada a la Gestión Empresarial (DIAGEM), del Instituto de Investigaciones y Estudios Superiores de las Ciencias Administrativas (IIESCA) de la Universidad Veracruzana (UV) para medir el grado de competencias en los empresarios MIPyME en el estado de Veracruz, como diagnóstico para establecer un modelo de atención dirigido a la mejora de las CGE en este sector, que tomara en cuenta que el entorno empresarial es multivariable y complejo, y por lo tanto priorizara la incidencia de esas variables en la supervivencia y desarrollo de las empresas en el corto, mediano y largo plazo.

El Análisis de Componentes Principales (ACP) como técnica exploratoria, permitirá hacer un estudio multivariado de las competencias en gestión empresarial, ofreciendo una visión de las posiblesrelaciones entre las variables y explicando éstas en términos de sus dimensiones subyacentes comunes (componentes); esto supone un procedimiento inductivo, que en ocasiones permite establecer hipótesis probabilísticas acerca de las variables o de los datos (Ojeda, 1990).

En particular, el objetivo del ACP es condensarla información contenida en las variables originalesen un pequeño número de índices que son combinaciones lineales de las variables originales. Si se logra encontrar una tipología empírica de la estructura de las variables consideradas,entonces los componentes pueden interpretarsecomo categorías

descriptivas que permitirán clasificar los empresarios por sus perfiles similaresen cuanto al grado de necesidad de CGE, permitiendo hacer comparaciones dentro del quehacer empresarial.

Lo anterior permitió evaluar las necesidades de formación en competencias de gestión empresarial según las prioridades que más inciden en el éxito de la empresa.

Así, en el presente trabajo, el ACP se utiliza para explorar posibles relaciones entre diez indicadores sobre CGE, identificar estructuras latentes en los mismos y construir escalas aditivas que resuman parcialmente la información contenida en todos ellos. La elección del ACP se debe a las ventajas que ofrece esta técnica para reducir un amplio número de variables en un conjunto menor de factores o componentes no correlacionados entre sí que den cuenta, de manera óptima, de los diferentes porcentajes de varianza común existente entre las variables originales, logrando un índice final que permita la clasificación, el ordenamiento y la estimación de las diferencias entre las CGE de las MIPyME veracruzana a través de un único valor representativo.

El objetivo general de este trabajo se resume entonces en la construcción de un índice que caracteriza las necesidades de CGE de la los empresarios de las MIPyME del estado de Veracruz. Con ello se sintetiza y prioriza las variables de formación cn lo que se construye un modelo de atención en el DIAGEM-IIESCA-UV para coadyuvar a la supervivencia y permanencia de este sector vulnerable.

Mora (1999) define la gestión como el conjunto de actividades necesarias para desarrollar un proceso o para lograr un producto determinado. Desde un enfoque gerencial, la gestión se plantea como una función institucional global e integradora de todas las fuerzas que conforman una organización (Valencia, 2002). El concepto de gestión agrupa cuatro funciones fundamentales para el desempeño de la empresa: a) planificación; b) organización; c) dirección; y d) control (Rubio, 2006). Teniendo en cuenta lo anterior, es evidente el beneficio de llevar a cabo la administración de empresas en base al concepto de gestión. El enfoque de gestión por competencias viene a resolver la problemática del desarrollo del empresario al considerarse compleja la interrelación con su papel clave en la empresa, con mayores responsabilidades y con multitud de funciones.

Llorens et al (2005) argumentan que las competencias pueden ser vistas como las características subyacentes en una persona que están causalmente relacionadas con los comportamientos y la acción exitosa en su actividad profesional; se refiere a las capacidades, conocimientos, destrezas, habilidades y actitudes necesarias para el pleno desempeño de una ocupación o cargo para la adecuada resolución de las funciones requeridas que se adquieren mediante una formación.

La formación continua para responder a los problemas de las MIPyME es una de las más importantes estrategias de desarrollo de recursos humanos que las organizaciones empresariales tienen en sus manos, hasta el punto que, en un entorno cambiante y competitivo como el que nos rodea, la formación se convierte en factor clave del éxito empresarial.

Se debe transmitir una cultura de formación que lleve a las MIPyME a reexaminar con frecuencia las competencias profesionales necesarias, la situación de sus trabajadores frente a esas competencias y las actuaciones en materia de formación para cubrir los posibles desajustes. Desde esta perspectiva, la acción de las instituciones educativas puede contribuir a mejorar el entorno empresarial a fin de generar beneficios para el conjunto de la sociedad (Banco Mundial, 2004; Brunetti, Kisunko y Weder, 1998; Schiffer y Weder, 2001).

Varela y Bedoya (2009), resaltan que el desarrollo de empresarios cimentado en competencias empresariales asegura el éxito y permanencia de la empresa joven; elaborando un modelo conceptual de desarrollo empresarial basado en competencias. Este autor menciona que la formación de empresarios es un proceso en el cual intervienen un sin número de variables sociales, culturales, psicológicas y económicas que contribuyen, con un conjunto de conocimientos específicos, a desarrollar una serie de competencias que buscan lograr que este empresario en formación tenga altas probabilidades de convertirse en un empresario exitoso, capaz de generar riqueza y desarrollo social a lo largo de su vida. Lo fundamental de sus aportaciones es que consideran inadecuado que la educación empresarial sea un bien franquiciable y que el modelo de una institución en un entorno especifico, sea necesariamente válido para otras instituciones en otros entornos diferentes; pues las características educacionales de cada institución: sus estudiantes, su entorno socio-

económico, su cultura, sus percepciones y todas las variables de desarrollo humano pueden ser muy diferentes.

Llorens *et al* (2005), publicaron un estudio del análisis de las competencias profesionales y su relación con las estructuras organizativas de la empresa. Mencionan que las principales formas de conseguir una ventaja competitiva sostenible, se asocian a las competencias de los trabajadores. Su trabajo plantea, la evaluación de las competencias claves que aseguren el cumplimiento de sus objetivos, atendiendo al enfoque de las capacidades dinámicas como método de establecimiento estratégico. En consecuencia, las estructuras organizativas también deberán sufrir variaciones y redefiniciones para adaptarse a los nuevos requerimientos de los sectores.

Gutiérrez y De Pablos (2008) publicaron una metodología de análisis y evaluación de la gestión por competencias en el ámbito empresarial y su aplicación a la universidad; Delinearon fases importantes como la evaluación y el desarrollo de las competencias y abordaron retos de las IES ante cambios en la docencia, investigación y transferencia de resultados hacia la sociedad, y cómo las universidades han iniciado un proceso de cambio para adaptarse a este nuevo entorno, sobre todo en el aspecto metodológico. Esta adaptación se centra en el aprendizaje y en las competencias. Además debatieron la viabilidad de extrapolar los modelos de gestión por competencias propios del mundo empresarial a la enseñanza superior universitaria.

En el estudio de Huerta et al (2009), se discute un análisis de las competencias gerenciales en el sector público de salud. El estudio se centró en el análisis sobre la efectividad de la metodología para transformar en oportunidades de formación determinadas situaciones de resolución de problemas, ya que el desarrollo y fortalecimiento de las competencias es un aspecto clave para el éxito organizacional. En este trabajo se aplicó una autoevaluación por competencias, concluyendo que existe una carencia y necesidad de formación en las CGE y que es cada vez más imperante la identificación y evaluación de tales habilidades.

Como competencias específicas, las CGE, reflejan conocimientos teóricos y procedimientos propios y concretos de gestión empresarial. Para el presente estudio se definen las CGE involucradas en Olivera *et al*, (2009).

1. *Dimensión estratégica.* Esta competencia ayuda a determinar si en el empresario hay un sentido de actuar que lleve a una *Planeación Estratégica*, entendida como el lineamiento general de acción que se elige para llegar al objetivo planteado ligado a la Misión y Visión; y una *Planeación Táctica*, o conjunto de acciones y métodos que se requieren para alcanzar los objetivos planteados, como pueden ser los planes de acción con metas establecidas; y por último, una *Planeación Operativa*, es decir, medios específicos que deben ser utilizados para llevar a cabo los planes de acción y así, alcanzar las metas inmediatas o resultados específicos.
2. *Dimensión Técnica.* Se refiere a los aspectos que intervienen en la toma de decisiones sobre la capacidad de producción a corto, medio y largo plazo. A corto plazo deben tomarse en cuenta factores que permitan realizar la programación de la producción. A medio plazo, más de un año, los factores que inciden en la toma de decisión son todas las líneas de productos de la empresa, condicionadas por los planes a largo plazo. A largo plazo las decisiones son relativas a la capacidad de las instalaciones y a su localización.
3. *Dimensión Mercadológica.* Aborda los objetivos de las políticas de producto, precio, plaza y promoción que se pretenden lograr con la planeación, ejecución, investigación y control de la mercadotecnia; cómo se van a alcanzar las metas y con qué recursos; la planeación de las actividades de mercadotecnia que se van a implementar y los métodos de control y monitoreo que se van a utilizar para realizar los ajustes que sean necesarios.
4. *Dimensión Capital Humano*. Resalta la capacidad de administrar y dirigir el recurso humano como factor crítico de éxito o de riesgo en la empresa, abarcando el proceso de selección, incorporación, inducción, retribución, organización, motivación, evaluación, preservación, capacitación, desarrollo y control.
5. *Dimensión Jurídica.* Se refiere a la aplicación y atención del conjunto de normas relativas a los comerciantes en el ejercicio de su profesión, a los actos de comercio legalmente calificados como tales y a las relaciones jurídicas derivadas de la

realización de estos; en base al derecho mercantil, fiscal, corporativo, administrativo y de seguridad social.

6. *Dimensión Financiera.* Describe el análisis de la estructura financiera y consecuentemente de la demanda de medios financieros en la empresa, así como del aprovisionamiento de medios o fuentes de financiación, su racionamiento o planificación y presupuesto financiero, como también la política financiera de la empresa condicionada por sus objetivos.
7. *Dimensión tecnológica.* Aborda la gestión del uso de nuevas tecnologías de información y comunicación para la toma de decisiones; La tecnología ofrece la manipulación de datos para proveer información clara, precisa y confiable que sea utilizada para la toma de decisiones oportuna y acertada.
8. *Dimensión fiscal.* Está encaminada a que el empresario pueda aprovechar las diversas opciones que en materia fiscal existen para contribuir de manera proporcional y equitativa dentro de un marco de auténtica legalidad.
9. *Dimensión Responsabilidad Social Empresarial (RSE).* Refleja la contribución activa y voluntaria al mejoramiento social, económico y ambiental por parte de la empresa, englobando un conjunto de prácticas y estrategias que implementa el empresario.
10. *Dimensión Información*. Manejo como herramienta fundamental los sistemas de información de diferente ámbito, para atender de una manera eficiente, ágil y oportuna las necesidades de los clientes, hacer un uso óptimo de los recursos y cuidar el entorno ambiental; además de visualizar las oportunidades y amenazas del entorno externo de la empresa.

En la siguiente sección se presenta la metodología usada en la investigación, donde se describe la base de datos y la estrategia de análisis estadístico, posteriormente se presentan los resultados con la técnica de componentes principales y en la última sección se detallan las conclusiones a las que se llegó.

Metodología

En esta investigación se utilizó la base de datos presentada en Olivera *et al* (2009), que contiene información de 105 empresas MIPYME veracruzanas censadas hasta el 2005, de las cuales 101 pertenecían a la Micro; 3 a la Pequeña y 1 a la Mediana empresa.

La base de datos se construyó con un instrumento autoadministrado dirigido al empresario con características MIPYME para detectar necesidades de CGE. De un total de 100 categorías de estudio, medidas en una escala de Lickert del 1 al 5, donde la menor puntuación reflejaba mayor carencia de CGE se construyeron 10 indicadores que se muestran en la siguiente tabla:

Tabla 1. Indicadores

Dimensión
Estratégicas
Técnicas
Mercadológicas
Recursos humanos
Jurídicas
Financieras
Administrativas
Fiscal
Sistemas
Responsabilidad empresarial

El tratamiento estadístico aplicado fue inicialmente un análisis exploratorio a las diez variables indicadoras, posteriormente se realizó en el Análisis de Componentes Principales, para reducir la dimensionalidad de las diez variables y construir una escala que resuma parcialmente la información contenida en todos ellas, para ello se usó el paquete estadístico SPSS versión 15.0 para Windows.

Resultados y discusión

Se realizó inicialmente un análisis exploratorio en el que se obtuvieron las medias de los 10 indicadores, recordemos que a menor puntuación se carece o es menor de la CGE, por lo que (ver Tabla 2) la CGE que presenta una puntuación más alta es la dimensión de ESTRATEGIA y la que tiene menor puntuación es la de RESPONSABILIDAD.

Tabla 2. Estadísticas descriptivas

Dimensión	Media
ESTRATEGICA	3.833
TECNICA	3.034
MERCADO	3.173
HUMANO	3.032
JURÍDICA	2.794
FINANCIERA	3.273
TECNOLOGICA	3.036
FISCAL	2.851
RESPONSABILIDAD	2.72
Num. de casos válidos	105

Después de aplicar la técnica de ACP se despreciaron las componentes con una pequeña proporción de varianza. De las 10 componentes se decidió retener tres, los cuales explican el 58.47 % de la variación total. En la Tabla 3 se explica la composición de estos tres componentes.

Tabla 3. Descripción de los factores artificiales.

INDICADOR	DESCRIPCIÓN
Componente de Factor Interno* (*CFI)	Factor que agrupa las variables relacionadas con aspectos de procesos internos operativos y estratégicos, agrupando las competencias en la dimensión estratégica, humano, jurídica y financiera.
Componente de Factor Externo* (*CFE)	Factor que agrupa las variables relacionadas con aspectos de clientes y proveedores agrupadas en las dimensiones técnica, de mercado y responsabilidad social.
Componente de Factor Innovación y Desarrollo (CFID)	Factor que agrupa las variables relacionadas con aspectos de tecnología y de sistemas de información comerciales, fiscales, contables, sociodemográficos y económicos que agrupa las dimensiones tecnológica, fiscal e información.

Las figuras siguientes presentan el diagrama de empresarios en los planos factoriales.

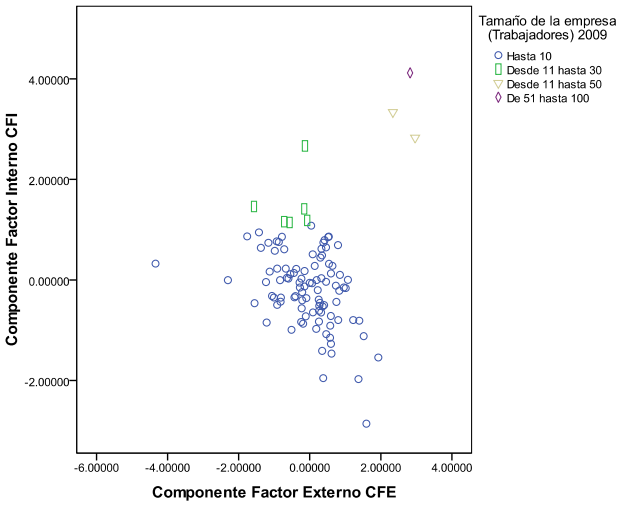


Figura 1. Componente Factor Interno CFI y Componente Factor Externo CFE por tamaño de empresa

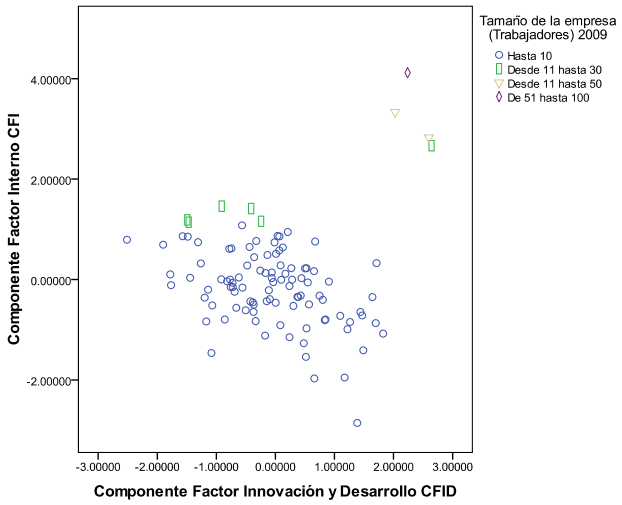


Figura 2. Componente Factor Interno CFI y Componente Factor Innovación y Desarrollo CFID por tamaño de empresa

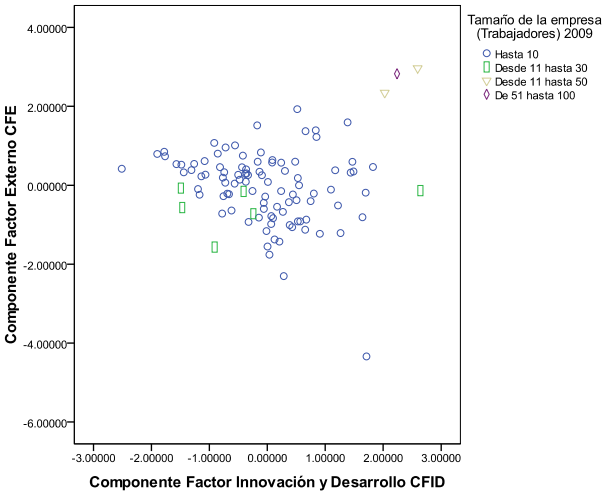


Figura 3. Componente Factor Externo CFE y Componente Factor Innovación y Desarrollo CFID por tamaño de empresa

Para la primera componente (Figura 1) que agrupa las variables relacionadas con aspectos de procesos internos operativos y estratégicos agrupando las competencias en la dimensión estratégica, humano, jurídica y financiera, se observa que la empresa mediana se separa del resto (esquina superior derecha) por tener un dominio en estos factores que el resto de las demás, al igual que sucede con las dos empresas pequeñas que la acompañan. Para la segunda (Figura 2) componente que agrupa las variables relacionadas con aspectos de clientes y proveedores agrupadas en las dimensiones técnica, de mercado y responsabilidad social, vemos el mismo comportamiento que en la Figura 1, para la empresa mediana y las dos pequeñas. Finalmente para la tercera componente (Figura 3) que agrupa las variables relacionadas con aspectos de tecnología y de sistemas de información comerciales, fiscales, contables, sociodemográficos y económicos que agrupa las dimensiones tecnológica, fiscal e información se observa lo mismo el destacado alejamiento de la empresa mediana.

En base a los resultados de la caracterización de los componentes realizado en el gráfico de empresarios según el tamaño de empresa procediéndose a la interpretación de los perfiles de cada de ellos. Por la posición de los empresarios en empresas de 51 a 100 trabajadores, y de 11 a 51 trabajadores en el plano factorial, este perfil de empresario se caracteriza por muy poca necesidad de Competencias de Gestión Empresarial en el componente *CFI,* componente *CFE* y componente *CFID*. El perfil de los empresarios de empresas de un tamaño de 11 hasta 30 trabajadores requieren mayor competencias en el componente *CFE* pero requieren de más competencias en el componente *CFID* que en el componente *CFI*. El perfil de los empresarios de empresas con hasta 10 trabajadores requieren competencias tanto en competencias *CFE* como en competencias *CFI* pero caracterizadas la mayoría de ellas en una gran necesidad de competencias *CFI* y *CFID*. (Ver Tabla 4).

Tabla 4. Necesidades de competencias en empresarios por empresas según tamaño.

Empresas de:	NECESIDAD DE COMPETENCIAS		
	CFI	CFE	CFID
Hasta 10 trabajadores	SI	SI	SI
Desde 11 hasta 30 trabajadores	NO	SI	SI
Desde 11 hasta 50 trabajadores	NO	NO	NO
De 51 hasta 100 trabajadores	NO	NO	NO

Con estos coeficientes se obtiene el factor o índice de las necesidades de CGE como una combinación lineal de los indicadores estandarizados. Este factor conlleva una ordenación de los empresarios MIPyME sobre estas necesidades ya que está constituido en una escala de intervalo. Esta cualidad del factor permite agrupar a los empresarios en nueve conjuntos claramente definidos, de acuerdo a la Técnica de Estratificación Óptima desarrollada por Dalenius y Hodges (2000).

La aplicación de este método estadístico lleva a dividir el recorrido del factor en subintervalos en puntos de corte (incompetente, muy poco competente, medio competente, competente, muy competente, alto competente, muy alto competente y sobresaliente). De esta manera, cada programa se ubicará según se ubique el valor de su factor

Para obtener una apreciación más clara se construyó un Índice de Competencias en Gestión Empresarial (ICGE) observe la Tabla 5.

Tabla 5. Índice de competencias

Grado	Casos	Rango	Número de trabajadores	Tamaño de empresa
Incompetente	27	(-0.776197, -0.42210)	Hasta 10	Microempresa
Muy poco competente	35	(-0.422110, -0.06800)	Hasta 10	Microempresa
Poco competente	20	(-0.068004, 0.28609)	Hasta 10	Microempresa
Medio competente	12	(0.286093, 0.64019)	Hasta 10	Microempresa
Competente	5	(0.640180, 0.99429)	Hasta 10	Microempresa
Muy competente	2	(0.994288, 1.13005)	Hasta 10	Microempresa
Alto competente	1	(1.130049, 6.57824)	Desde 11 hasta 30	Pequeña empresa
Muy alto competente	2	(6.578251, 14.0205)	Desde 11 hasta 50	Pequeña empresa
Sobresaliente	1	(14.020528, 16.1671)	De 51 hasta 100	Mediana empresa
	105			

Los resultados destacan que los programas de formación en gestión empresarial dirigidos a la MIPyME deben tomar en cuenta que este sector agrupa empresas con diferencias marcadas en sus características y en sus necesidades de competencias. Como marcan algunos autores en sus estudios, el desarrollo de competencias en CGE debe partir de la identificación de sus necesidades específicas y esto es un elemento fundamental para asegurar el éxito y permanencia de la empresa en el mercado, como lo menciona Varela y Bedoya (2009). La microempresa veracruzana tiene carencias en sus competencias, lo que significa una desigualdad de la competitividad muy marcada con respecto a la pequeña y mediana empresa, por lo cual debe ser intervenida con programas específicos basados en sus necesidades. Llorens *et al* (2005) le atribuyen al cambio en las estructuras de la empresa lograda por la formación del empresario un peso importante para elevar la competitividad de la empresa. Por ello la importancia de la evaluación y desarrollo de la formación del empresario en CGE.

Conclusiones

Los resultados alcanzados han permitido ilustrar, con una base de datos relativamente reducida, la aplicación de una técnica multivariable, el ACP, en el tratamiento de datos procedentes de las necesidades en Competencias de Gestión Empresarial (CGE) de MIPyME del estado de Veracruz. Por tanto, el ACP es una herramienta óptima para la identificación y caracterización del perfil del empresario en cuanto a las necesidades de CGE. El tratamiento estadístico permitió delinear una estrategia de atención basado en la formación de competencias. (Figura 4).

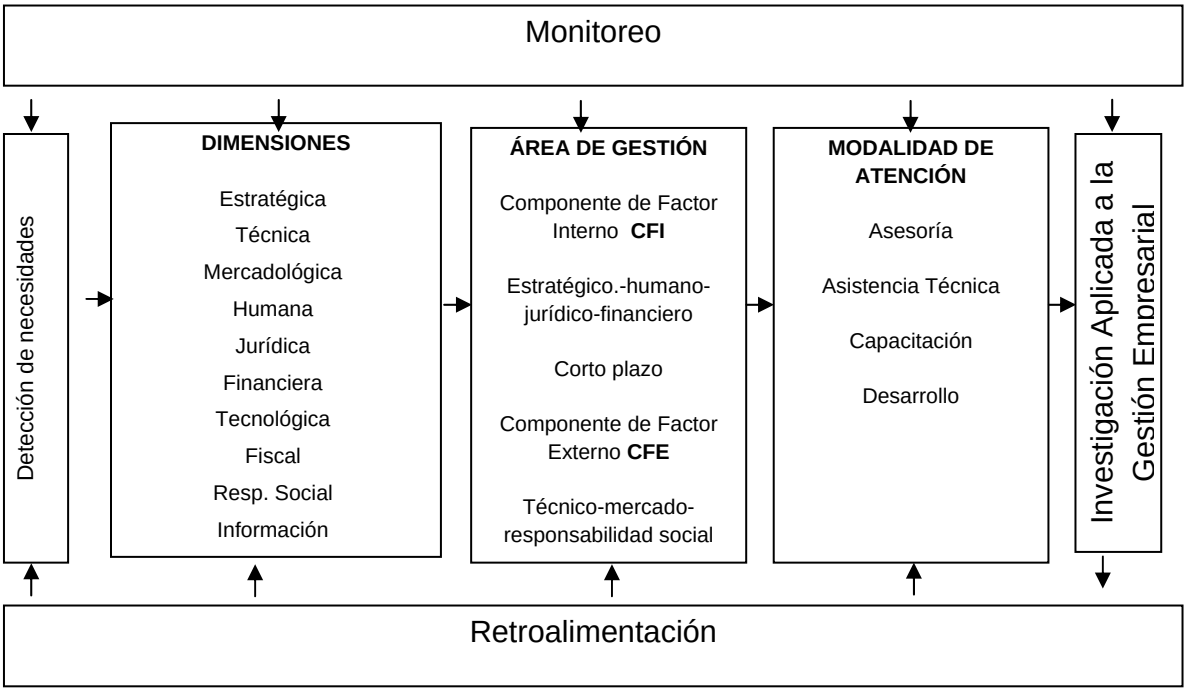


Figura. 4. Modelo de Atención Empresarial del DIAGEM

Basado en los resultados el DIAGEM-IIESCA propone un diseño de modelo específico para la atención basada en la formación de competencias de gestión empresarial de empresarios MIPyME en el estado de Veracruz, teniendo en cuenta el perfil de necesidades.

En el componente CFI que se refiere a las competencias en factores internos como son el estratégico, el financiero, el de capital humano y el jurídico, y que inciden más en la supervivencia y permanencia de la microempresa se propone la formación de este rublo acorto plazo, antes que hacer una intervención dirigida a elementos externos en el

componente CFE que agrupa a los factores técnicos, de mercado y de responsabilidad social y que tienen que ver con la relación directa del cliente y proveedores y que pueden desarrollarse a mediano plazo. Definitivamente el componente CFID referente a la innovación y desarrollo que engloba las dimensiones de tecnología, información y fiscal, su incidencia no amerita una atención prioritaria por lo que se considera a largo plazo.

El desarrollo de competencias de factores internos (CFI), competencias de factores externos (CFE) y competencias de factores de innovación y desarrollo (CFID) y busca generar un empresario integrado al contexto empresarial. El tipo y el nivel de desarrollo de cada competencia debe estar asociada a la modalidad de atención y ello definirá la extensión de las fases y los recursos requeridos.

El modelo se visualiza con la posibilidad de tener etapas de retroalimentación y no se basa en conceptos curriculares, sino en términos de desarrollo de competencias y por lo tanto no es un proceso con duración definida.

Referencias

Banco Mundial, (2004). Small and Medium Enterprise (SME*). World Bank Group review of small business*

Brunetti, A., Kisunko G. y Weder B. (1998). A Note on the Institutional Bias Against Small, Local Firms in Less Development Countries, *World Bank Policy Research Working Paper*, Washington, D.C., World Bank.

Gutiérrez S. y De Pablos C. (2009). *Analysis and evaluation of competency Management in the business world and its application to university*. Facultad de Ciencias Jurídicas y Sociales, Departamento de Economía de la Empresa. Universidad Rey Juan Carlos

Huerta P., Leyton C. y Saldia H. (2009). *Analysing a public health service network's managerial competente*. Facultad de Ciencias Empresariales, Universidad del Bío-Bío Concepción- Chile.

Kish, L. (1995) *Diseño estadístico para la investigación*. Centro de Investigaciones Sociológicas.

Llorens G., Olivella J., y Llinas J. (2005). Análisis de las competencias profesionales y de las estructuras organizativas en el entorno de las tecnologías de la información y la comunicación. *IX Congreso de Ingeniería de Organización*. Gijón, España

Mora, J. (1999). Transformación y gestión curricular. *Memorias Seminario Taller Evaluación y Gestión Curricular*, Univ. De Antioquia.

Ojeda, M. M. (1990). Componentes principales: ideas para su uso adecuado como una técnica exploratoria. *Revista Investigación Operacional*. Universidad de la Habana. Cuba. 11 (1), 37-44.

Olivera D. A., Cano M., y Domínguez B. (2009). Estudio de factibilidad para la implementación de un Departamento de Investigación Aplicada a la Gestión empresarial. *IIESCA-Universidad Veracruzana*. México

Rubio, D. P. (2006). *Introducción a la Gestión Empresarial*. Instituto Europeo de Gestión Empresarial.

Schiffer, M. y Weder B. (2001). Firm Size and the Business Environment: Worldwide Survey Results*, Discussion Paper*, Núm. 43. Washington, D.C., International Finance Corporation.

Valencia, C. (2002). Gerencia de Proyectos. *Seminario para profesores U. de A. México*.

Varela R. y Bedoya O., (2009). Modelo conceptual de desarrollo empresarial basado en Competencias. *Centro de Desarrollo del Espíritu Empresarial*. Universidad ICESI

Análisis de la industria del calzado en el periodo 1999-2009

Cecilia Cruz López
Julián Herrera Jiménez

Resumen

Se realizó un análisis de la industria nacional del calzado y de las diferentes regiones en las que se encuentra distribuida dentro de la república mexicana. Para tal efecto, se utilizaron datos agregados a nivel entidad federativa, con ellos, se analizó el comportamiento de la industria a nivel nacional y se realizaron comparaciones entre las regiones. Se aplicó la prueba de Kruskal Wallis con el propósito de evaluar el desarrollo económico de la industria, encontrándose que ésta, atraviesa por un grave estancamiento económico. Se demostró, a través de la aplicación de la prueba de la Mediana, que en el periodo de estudio se presentaron cambios importantes en las regiones que conforman la industria del calzado. Por último, se utilizó la prueba de Dunn que permitió conocer que las regiones que resultaron con diferencias significativas en el número de unidades económicas, de población ocupada y de producción bruta, fueron el Sur-Sureste, Centro y Noroeste.

Palabras clave: Prueba de Kruskal Wallis, Prueba de la Mediana, Prueba de Dunn.

Abstract

An analysis of the Mexican footwear industry and the different regions where it is distributed. To this end, we used aggregate data at state, with them, we examined the behavior of the industry nationwide and made comparisons between regions. We applied the Kruskal Wallis test with the purpose of evaluating the economic development of the industry, finding it, going through a severe economic stagnation. It was demonstrated through the application of the Median test, which in the study period there were significant changes in the regions that make up the footwear industry. Finally, we used the Dunn test which indicated that the regions were significant differences in the number of economic units of production employed population and gross, were the South-Southeast, Central and Northwest.

Keywords: Kruskal Wallis test, median test, Dunn test.

Introducción

A través de la historia, el desarrollo económico se ha estudiado ampliamente (Kelsen, 1949; Porrúa, 1954; Serra, 1964). En México, estudios recientes han profundizado en el desarrollo de las industrias (Cordero *and* Domínguez, 2007; Dussel, 2007; Macías, 2007). Se ha demostrado que el país atraviesa por un acentuado estancamiento económico que repercute en la disparidad de la distribución del ingreso entre los mexicanos. Se ha demostrado también que gran parte del problema radica en el inadecuado financiamiento dispuesto para las micro, pequeñas y medianas empresas (Garrido *and* Prior, 2007; Cuamatzin, 2007; Orlik, 2007; Mántey, 2007), lo cual, se encuentra directamente relacionado con los bajos niveles de productividad que ha presentado la industria del calzado en los años recientes (Rendón *and* Morales, 2001; Ojeda, 2007). Asimismo, Hernández, (2007) hace un diagnóstico del desempeño de la industria mexicana del calzado, en el cual presenta la problemática que enfrenta este sector y las pocas posibilidades que tiene de competir en el mercado globalizado, ya que se está viendo amenazada por países como Brasil, Argentina, Taiwan y Corea, por esa razón los empresarios mexicanos del calzado muestran pesimismo ante la competencia y tienden a desaparecer no sólo del mercado internacional sino del mercado interno.

El trabajo tiene como objetivo conocer la evolución que tuvo la industria del calzado en el periodo 1999-2009 realizando un análisis de sus unidades económicas, su población ocupada y su producción para cada una de las regiones del país e identificando aquellas cuyas diferencias resulten más significativas. El artículo inicia con una descripción general del planteamiento del problema, los objetivos, algunos resultados y las conclusiones más relevantes. Se continúa con una sección dedicada a la metodología utilizada. Se explica la naturaleza de la base de datos y la estrategia estadística que se llevó a cabo. Se presenta además, un apartado que contiene los resultados obtenidos, los cuales son comparados con las afirmaciones extraídas de los autores mencionados en la parte introductoria. Se culmina con una serie de conclusiones que describen el conocimiento generado y precisa si el objetivo del artículo fue alcanzado.

Metodología

Se realizó un estudio de la industria del calzado, la información se obtuvo de los censos económicos del sistema automatizado de información censal del INEGI. Los variables utilizadas en la base de datos son: número de unidades económicas, población ocupada y producción bruta de los 32 estados de la República Mexicana en 3 periodos de tiempo, 1999, 2004 a 2009. Para llevar a cabo el análisis de la base de datos se recurrió a la distribución territorial del país de los establecimientos del ramo de la industria del calzado, según las cinco regiones definidas por el Ejecutivo Federal en el Plan Nacional de Desarrollo 2001-2006 que se presentan en la Tabla 1.

Tabla 1. Distribución territorial de la Industria Mexicana del Calzado

Región	Estados
1.- Sur-Sureste	Campeche, Chiapas, Guerrero, Oaxaca, Quintana Roo, Tabasco, Veracruz y Yucatán.
2.- Centro-Occidente	Aguascalientes, Colima, Guanajuato, Jalisco, Michoacán, Nayarit, Querétaro, San Luis Potosí y Zacatecas.
3.- Centro	Distrito Federal, Hidalgo, México, Morelos, Puebla y Tlaxcala.
4.- Noreste	Coahuila, Durango, Nuevo León y Tamaulipas.
5.- Noroeste	Baja California, Baja California Sur, Chihuahua, Sinaloa y Sonora.

La estrategia de análisis estadístico dio inicio con un análisis exploratorio de los datos de las industrias que consistió en la elaboración de histogramas de las variables de estudio. Posteriormente, se aplicó la prueba de Kruskal Wallis a los datos de las 5 regiones, tomando como K medianas para comparar los años referentes a los censos económicos

efectuados en 1999, 2004 y 2009. Asimismo, se implementó la prueba de la Mediana en las cinco regiones industriales para probar diferencias entre ellas. Por último, se efectuó la prueba de Dunn como herramienta de comparación múltiple para identificar cuáles fueron las regiones divergentes.

Resultados y Discusión

Los histogramas mostraron que las tres variables de análisis presentaban distribuciones sesgadas, por lo que se aplicó la prueba Kolmogorov-Smirnov y se comprobó que ninguna de las variables de estudio presentaba normalidad ($P < 0.01$). Por esta razón, el tratamiento de las mismas quedó reservado para las pruebas no paramétricas

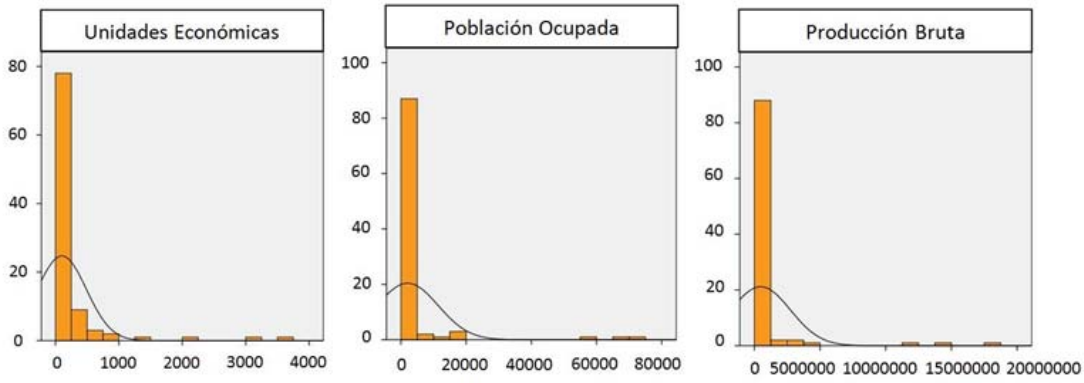


Figura 1. Histogramas para las variables de estudio.

Se aplicó la prueba de Kruskal Wallis para comprobar que no existieron diferencias significativas en el número de unidades económicas ($p = 0.686$), ni en la población ocupada ($p = 0.612$), ni en la producción bruta ($p = 0.932$), para la industria en los años 1999, 2004 y 2009. Estos resultados demuestran que la rama del calzado sufrió recesión económica en el periodo de estudio. Asimismo, se confirma que esta industria comparte el estancamiento económico por el que atraviesa el país, según lo mencionado por Cordero *y* Domínguez, (2007); Dussel, (2007); y Macías (2007).

Con respecto al análisis por regiones que conforman la industria del calzado, si se encontraron diferencias significativas con la prueba de la mediana entre el número de

unidades económicas ($p < 0.01$), la población ocupada ($p < 0.01$), y la producción bruta ($p = 0.01$). Estas diferencias pueden observarse gráficamente en la Figura 2.

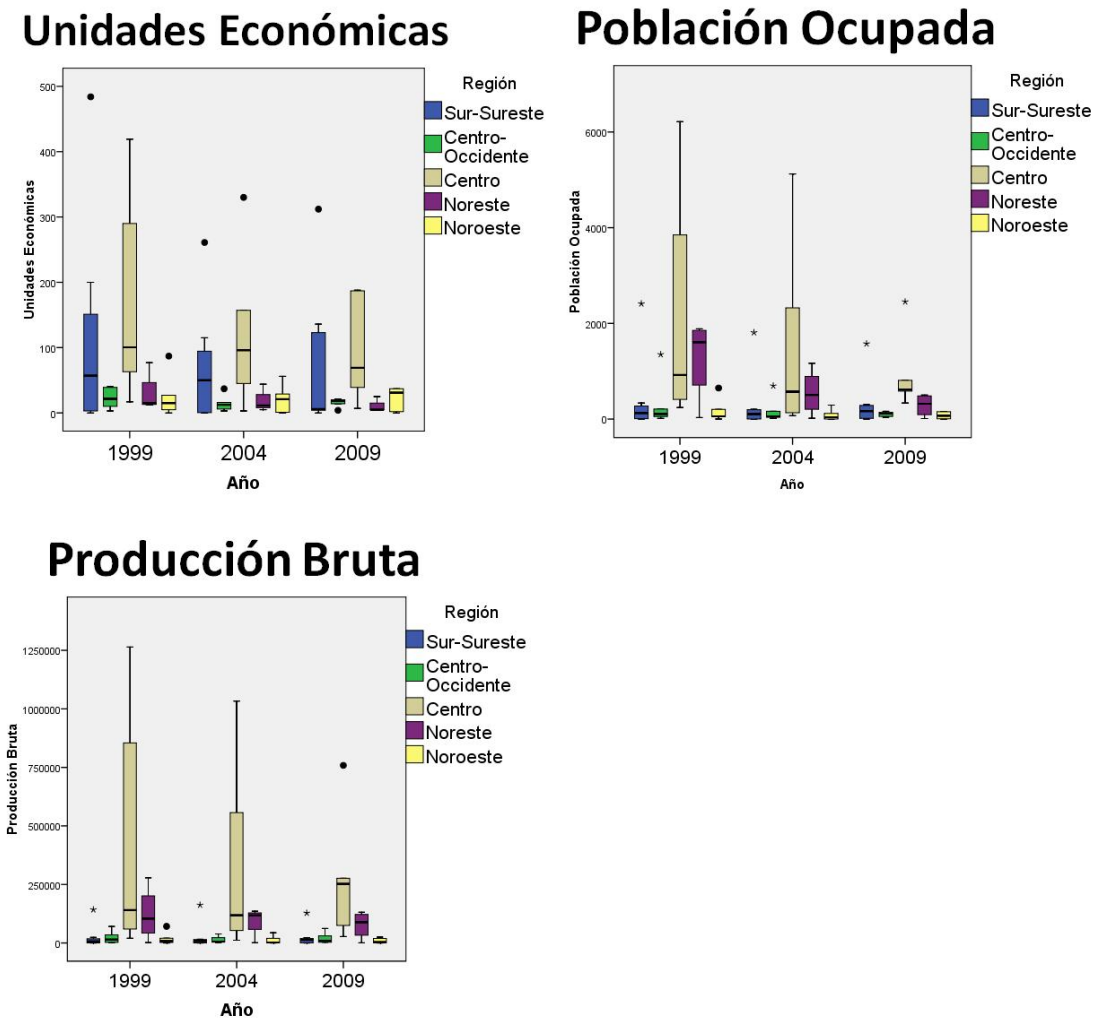


Figura 2. Gráfica de cajas comparativa por año y por región.

Como la prueba de la mediana resulto significativa se realizó la prueba de Dunn para encontrar entre qué parejas de regiones se encontraron las diferencias y se hizo evidente lo siguiente: En cuanto a unidades económicas una de las diferencias resultantes fue entre la región Centro con la región Noreste ($p = 0.04$) y también la región Centro con la Noroeste ($p=0.02$), ya que el número de empresas del centro es mayor un 97.37%, mientras que las industrias del Noreste y Noroeste tienen un 66.38% y un 20.14% respectivamente.

Con respecto del personal ocupado, la región Centro, presentó diferencias significativas con las regiones Sur-Sureste ($p = 0.01$) y Noroeste ($p < 0.01$). El centro incremento su población ocupada un 18.90%, mientras que el Sur-Sureste redujo sus empleos un 24.79%, de la misma manera, el Noroeste escindió del 59.62% de su personal.

Asimismo, el Centro difirió del Sur-Sureste ($p < 0.01$) y del Noroeste ($p < 0.01$) en cuanto a la producción bruta. En este sentido, la región Centro-Occidente también presentó diferencias significativas con respecto del Noroeste ($P = 0.04$). Las explicaciones son las siguientes: el Centro-Occidente y el Centro incrementaron su nivel de producción en un 29.40% y un 11.33%, mientras que el Sur-Sureste y el Noroeste decrecieron el 5.47% y el 52.33% de su producción bruta.

Los resultados permiten observar que las unidades económicas ubicadas en el Sur-Sureste y Centro-Occidente han permanecido constantes, mientras que las empresas pertenecientes al Noreste y Noroeste han disminuido drásticamente. La población ocupada en las regiones Centro-Occidente y Noroeste también ha permanecido constante. Asimismo, el personal del Sur-Sureste y Noroeste disminuyó de una manera muy pronunciada. Con respecto de la producción bruta, el Sur-Sureste y del Noroeste decreció significativamente, mientras que la producción del Noroeste no registró cambios importantes. Por último, la región centro, aunque incrementó de manera notable el número de sus empresas y de su personal ocupado, tal incremento no se vio reflejado en la producción bruta, situación que supone que el centro mantiene una baja productividad puesto que aunque incrementó su número de unidades económicos y de personal ocupado genera niveles muy bajos de producción bruta. Todos estos datos hacen evidente la baja productividad de la industria en todas las regiones, tal como lo muestran Rendón *and* Morales, (2001); Hernández, (2007) y Ojeda, (2007).

Conclusiones

Se concluye que la industria nacional del calzado no ha evolucionado positivamente a través de los años contenidos en el periodo de estudio, a partir de lo cual se demuestra que los esfuerzos estatales en materia de desarrollo económico no resultaron suficientes para evitar la recesión de la rama productiva. Asimismo, se descubrió que los problemas de

estancamiento económico de la industria son compartidos por todas las regiones que la integran, siendo las regiones Sur-Sureste y Noroeste, aquellas que se vieron más perjudicadas en la pérdida de empresas, de personal ocupado y en la reducción de su producción bruta. Cabe mencionar que las regiones Centro-Occidente y Noreste aunque no presentaron pérdidas sustanciales en las variables de estudio, tampoco presentaron muestras de desarrollo económico. Con respecto de la región Centro, se encontró que si bien incrementó en gran medida sus unidades económicas y personal ocupado, no registró incrementos significativos en su producción, lo que significa que el crecimiento de esta región no obedeció al desarrollo de su industria.

Referencias

Cuamátzin, F. (2007). Los retos de la inversión pública en México. *Finanzas Públicas para el desarrollo*. P. 54-76. UNAM. México.

Cordero, M. E. y Domínguez, L. (2007). ¿Puede México aplicar una política industrial?: Márgenes en el TLC y la OCDE. *Política industrial manufacturera*. P. 23-36. UNAM. México.

Dussel, E. (2007). Política Industrial y Microempresa: lineamientos generales. *Política industrial manufacturera*. P. 54-64. UNAM. México.

Garrido, C. y Prior, F. (2007). Bancarización y Microfinanzas: sistemas financieros para las mipymes como un dilema central para el desarrollo económico en México. *Financiamiento del crecimiento económico.* P. 57-77. UNAM. México.

Hernández, E. P., (2007). Retos y Perspectivas de la Industria Mexicana del Calzado ante la apertura comercial: el impacto de la competencia con china. *Estudios sobre estado y sociedad*. P. 95-121. Vol. 13. No. 40.

Mántey de Anguiano, G. (2007). Política bancaria para el crecimiento con estabilidad. *Financiamiento para el desarrollo*. P. 152-155. UNAM. México.

Tello, C. (2007). Política Económica: Finanzas Públicas. *Finanzas públicas para el desarrollo*. P. 17-36. UNAM. México.

Kelsen, H. (1949). *Teoría General del Derecho y el Estado*. 2a. Ed. México. Universidad Nacional Autónoma de México.

Kruskal, H., and Wallis, A. Use of ranks in one criterion variance analysis. *Journal of American Statistics Asociation*. P. 584-618. Vol. 47. No. 260.

Ojeda, J. (2007). Ventaja competitiva: Retos de las Pymes en la industria del Calzado. Revista *Venezolana de Gerencia*. P. 513-533. Vol. 12. No. 40.

Levi, N. (2007). Financiamiento del crecimiento y disponibilidad de créditos bancarios. *Financiamiento del desarrollo económico.* UNAM. México.

Porrúa, F. (1978). *Teoría del Estado*. 11 Ed. México: Editorial Porrúa, S. A.

Rendón, A. y Morales, A. (2001). Modelos econométricos para analizar el impacto de variables económicas en la competitividad de la industria del calzado. *Política y Cultura*. Universidad Autónoma Metropolitana. México. No. 15.

Serra, A. (1964). *Ciencia Política: la Proyección Actual de la Teoría del Estado*. 4ta. Ed. México: Editorial Porrúa, S. A.

Análisis del mercado ocupacional en México durante el periodo 2005-2009

Alicia Eguía Casis
Yesenia Zavaleta Sánchez

Resumen

Este trabajo estudia el comportamiento de las variables del mercado ocupacional: nivel de ingresos, posición ocupacional y sector económico, durante los años 2005 al 2009, en México. Para la investigación se aplicó la técnica estadística de Análisis de Correspondencia Simple a una base de datos construida a partir de la Encuesta Nacional de Ocupación y Empleo (ENOE). El nivel de ingreso de la población ocupada no ha sufrido cambios significativos en su integración. La posición ocupacional ha permanecido estable. Además, el sector económico al igual que las otras variables, no presenta cambios significativos en su composición. Los resultados muestran que no existe una asociación significativa entre las variables que conforman el mercado ocupacional en México y los trimestres del periodo sujeto de evaluación.

***Palabras clave*:** Población Económicamente Activa, Población No Económicamente Activa, Población Ocupada, Análisis de correspondencia.

Abstract

This paper studies the behavior of the labor market variables: income, occupational status and economic sector in Mexico, during the years 2005 to 2009. For the research was applied the statistical technique of simple correspondence analysis to a database built from the National Survey of Occupation and Employment (ENOE). The income level of the employed population has undergone significant changes in their integration. Occupational status has remained stable. In addition, the economic sector, as well as other variables, does not significant changes in its composition. The results show no significant association between the variables that make the job market in Mexico and the quarters of the period under evaluation.

Keywords: Economically active population, Population not economically active, Employed population, Correspondence analysis.

Introducción

En México, como en otros países, las políticas del mercado laboral se han convertido en un importante instrumento de política económica y social. A estas políticas se ha canalizado un volumen considerable de recursos públicos, destinados a aliviar la pérdida de ingresos por desempleo, mejorar las habilidades de la fuerza de trabajo desocupada y facilitar el encuentro cualitativo y cuantitativo entre oferta y demanda de trabajo. La mezcla de programas ha ido variando, en función de la coyuntura económica y de la propia experiencia del país en programas de esta naturaleza.

El empleo es fundamental para la generación de bienes y servicios de las naciones, mediante el mismo las economías crecen y se logra el desarrollo de las naciones, por lo que el estudio de las variables que integran el mercado laboral de las naciones toma relevancia. El empleo no sólo mejora el nivel de ingresos y por tanto el nivel de desarrollo humano, también contribuye a la salud mental de las personas y a su favorable integración en la sociedad. El conocer la situación ocupacional de la población permite adecuar la legislación y diseñar políticas económicas, fiscales, laborales y demográficas, que atiendan al contexto y circunstancias particulares de una nación o región.

Para efectos de las encuestas sobre el mercado de trabajo, a la población ocupada se le clasifica en: 1) población en edad de trabajar: personas de 14 años y más; y 2) personas que aún no tienen edad de trabajar: menores de 14 años. De acuerdo al marco normativo de los sistemas de contabilidad nacional, la población en edad de trabajar a su vez está conformada por dos grupos: la Población Económicamente Activa (PEA) que incluye a las personas que participan en la generación de los bienes y servicios que produce el país; y por la Población No Económicamente Activa (PNEA), que se compone por quienes no realizan una actividad que contribuya a incrementar la producción del país. Las amas de casa son un ejemplo de personas que integran la PNEA, sin que ello implique que sus actividades sean intrascendentes, pues contribuyen al bienestar social; también en el segmento de la PNEA se encuentran los estudiantes, quienes se preparan para incorporarse al mercado ocupacional, las personas que prestan servicios gratuitos a la comunidad y los incapacitados.

La PEA desde la perspectiva ocupacional está constituida por los oferentes de servicios laborales, los cuales pueden estar: ocupados (aquellas personas que se encuentran produciendo bienes y servicios, durante un periodo de referencia especificado) y desocupados (los individuos que se encuentran sin desempeñar un trabajo pero en búsqueda del mismo, en un lapso determinado). Éstos son oferentes de mano de obra, por lo que ejercen presión sobre el mercado laboral. El no trabajar no implica ser desocupado, pues para formar parte del grupo de los desocupados se debe buscar activamente un trabajo. También es posible que personas que se encuentren ocupadas busquen un empleo, en tal caso no forman parte de los desocupados. De ahí que, la desocupación abierta no sea, ni pretenda ser, la magnitud que exprese cuánta gente necesita trabajar en un lugar y momento determinados o la medida de cuán grande es el déficit de oportunidades laborales: en realidad lo que la desocupación abierta indica es la magnitud de la población que se comporta como buscadora de trabajo.

La ENOE revela que la población de 14 años o más en México con relación a la población total pasó del 70.9% al 73.5%, del año 2005 al año 2009. La población de 14 años o más era en el primer trimestre de 2005 de 73,448,358 personas y en el último trimestre de 2009 de 79,312,278, lo que equivale a un incremento del 7.98%, mientras que el aumento de la población total fue del 4.2%. El crecimiento de la población de 14 o más años se debe a factores como la reducción de morbilidad y las tasas de natalidad registradas en la década de los noventas del siglo pasado. Si bien es cierto que no todos las personas que tienen 14 años o más se incorporan a la Población Económicamente Activa (PEA), dicho incremento representa un reto para la creación de suficientes fuentes de empleo en el presente y en el futuro cercano.

La PEA sufrió un incremento del primer trimestre del año 2005 al último trimestre de 2009, pasando de 42,215,661 personas a 47,041,909, lo que representa un aumento del 11.43% lo que contrasta con la PNEA, la cual pasó de 31,232,697 personas a 31,270,849, representando un incremento de sólo el 3.32%. Lo anterior revela un crecimiento mucho más acelerado de personas que se incorporan como demandantes de un trabajo que el de las personas que no participan en los procesos productivos. La composición de la PEA y

PNEA ha variado a lo largo de los años desde 2009 a 2005, representando en el último trimestre del año el 59.31% la PEA y el 40.69% la PNEA.

El incremento de la PEA evidencia la necesidad de brindar educación de calidad a los jóvenes y niños que les permita su inserción en un mercado laboral. La creación de nuevos bienes y servicios, y la mejora de los procesos productivos son acciones que harán competitivas a las generaciones que se incorporan a la PEA en un mundo globalizado, en el que sólo aquellos que ofrezcan calidad y costos adecuados podrán crecer y garantizar su continuidad en el mercado.

Aparicio (2007) realizó un estudio en el que se aborda la importancia de las políticas activas del mercado laborar (ALMP, Active Labour Market Policies por sus siglas en inglés) para reducir los niveles de desempleo originados por una recesión o estancamiento de la economía. En particular, se revisa la experiencia mexicana entre 1995 y 2005. El artículo presenta los aspectos teóricos del mercado de trabajo, del desempleo y de las ALMP para luego explicar la forma en que estas políticas son aplicadas en México. Para este fin se ajustó un modelo de regresión lineal múltiple entre la tasa de desempleo (variable dependiente) y dos de los programas más importantes en México de reinserción de desempleados al mercado de trabajo de las ALMP (Programas: Bécate y Bolsa de Trabajo, variables independientes) el cual reveló que existe una relación inversa, estadísticamente significativa, entre el desempleo y las ALMP en México. En otras palabras, que la tasa de desempleo disminuye a medida que se incrementa el número de becas de capacitación para desempleados del programa Bécate, y a medida que aumenta la cantidad de desempleados que son colocados en un empleo a través del servicio de Bolsa de Trabajo.

El objetivo de este estudio es analizar si existe un comportamiento que determine alguna relación en tres de las variables que integran el mercado ocupacional, en México, en los periodos trimestrales de los años 2005 a 2009, como son: 1) posición en la ocupación; 2) sector de actividad económica; 3) nivel de ingreso.

Metodología

Para llevar a cabo el propósito de esta investigación se analizaron los datos de la ENOE, que es una encuesta aplicada a los hogares mexicanos por el Instituto Nacional de Estadística y Geografía (INEGI). Los datos de la ENOE muestrean a la población ocupada en todo el territorio nacional. Se cuenta con la información trimestral en el periodo que comprende los años 2005 a 2009 de cada una de las variables sujetas a análisis y que conforman el mercado laborar en México; posición ocupacional, sector económico y nivel de ingreso. Es decir, para cada una de las variables se cuenta con 20 registros en términos de porcentajes. No se utilizaron bases de datos por estado de la república ni se hace diferenciación entre hombres y mujeres. Con fundamento en los datos proporcionados por la ENOE se elaboraron tablas de contingencia que agrupaban a la población ocupada. Las variables sujetas al análisis de correspondencia son: 1) posición en la ocupación; 2) nivel de ingreso y 3) sector de actividad económica, durante los trimestres del periodo 2005 a 2009 (Ver Anexo Tabla 1). Cabe mencionar que, para llevar a cabo el análisis se utilizó como medio de apoyo el software estadístico SPSS.

Con el propósito de hacer comparable la información sobre ocupación, desocupación y empleo producida por diferentes naciones, la Organización para la Cooperación y el Desarrollo Económico (OCDE) emitió lineamientos para la formulación de estadísticas laborales, los cuales fueron tomados en consideración para el diseño y establecimiento de la Encuesta Nacional de Ocupación y Empleo (ENOE). La aplicación de la encuesta en México se inició a partir de 2005 y dejó sin efecto a las encuestas que en materia del trabajo se llevaban a cabo hasta el momento: la Encuesta Nacional de Empleo Urbano (ENEU) y la Encuesta Nacional de Empleo (ENE). Actualmente se está realizando la conversión de la información obtenida de las encuestas vigentes antes del 2005 a los criterios de la ENOE para permitir la comparación de indicadores en series de tiempo.

Para la realización de este artículo también se recurrió a los informes elaborados por la Conferencia Internacional de Estadísticos del Trabajo, dependiente de la Organización Internacional del Trabajo, la cual emite resoluciones sobre estadísticas en

materia laboral, con el propósito de homogenizar la metodología y contribuir a la calidad de este tipo de información. También se consultaron publicaciones emitidas por el INEGI, relacionadas con la formulación y evolución de estadísticas sobre el mercado ocupacional. Se realizó un análisis exploratorio de las variables que conforman el mercado laboral en el periodo del 2005 a 2009, el cual consistió en calcular estadísticas descriptivas y elaborar algunos gráficos de líneas para analizar la tendencia de cada una de las variables a lo largo del tiempo. Posteriormente, se utilizó la técnica de análisis de Correspondencia Simple para identificar la existencia o no de alguna asociación entre cada una de las variables del mercado ocupacional en el periodo analizado.

Resultados y Discusión

Del análisis preliminar de los datos es posible observar el comportamiento de cada una de las variables en el periodo analizado que incluye los trimestres del 2005 al 2009. Con respecto a la posición ocupacional se pudo determinar que el porcentaje de los trabajadores subordinados del total de población ocupada oscila entre el 63% y el 67%, en cuanto a los trabajadores por cuenta propia representan una parte de la población ocupada entre el 21% y el 24%. Los trabajadores no remunerados se ubican en un porcentaje alrededor del 7% y los empleadores entre el 4 y 5% (Ver figura 1). De manera general se aprecia que continúan siendo los trabajadores subordinados los que representan la mayoría de la forma de ocupación en México, representando en el último trimestre de 2009 el 65.3% de la población ocupada, seguido por un 23.1% de personas que prestan servicios en forma independiente. Sólo el 4.5% de la población ocupada es empleadora.

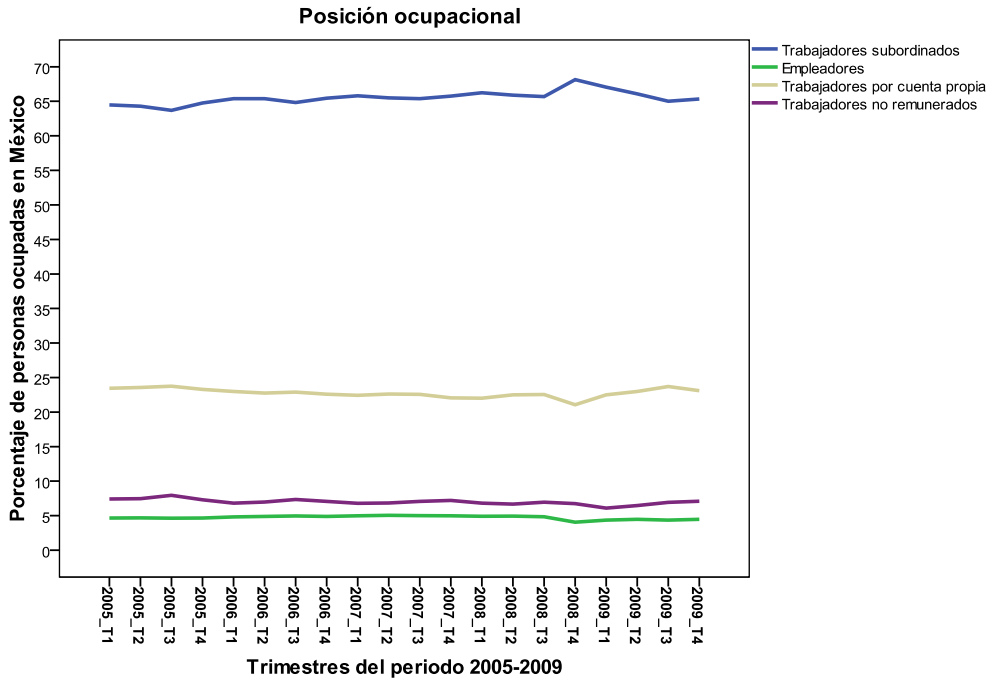


Figura 1: Gráfico de líneas de la posición ocupacional en periodo de tiempo trimestral (2005-2009) por trimestre de la población en México.

Con respecto al nivel de ingreso de la población ocupada en México, se observa que la población que recibió más de un salario mínimo hasta dos, mostró el cambio más significativo en los trimestres I de 2005 y III de 2008 pasando del 23.9% de la población ocupada al 19.9%. La población ocupada que obtuvo más de dos salarios mínimos hasta tres, registró la mayor variabilidad en los mismos trimestres y años, siendo de 19.1% a de 23.5% de la población ocupada (Ver figura 2). En el último trimestre de 2009, un escaso 9% de la población ocupada es la que percibe ingresos de más de cinco salarios mínimos, concentrándose la mayor parte de la población en un nivel de ingresos entre un salario mínimo hasta tres salarios. El patrón de comportamiento del nivel de ingreso entre las personas que perciben más de uno hasta dos salarios mínimos es opuesto al de las personas que perciben más de dos hasta tres salarios mínimos.

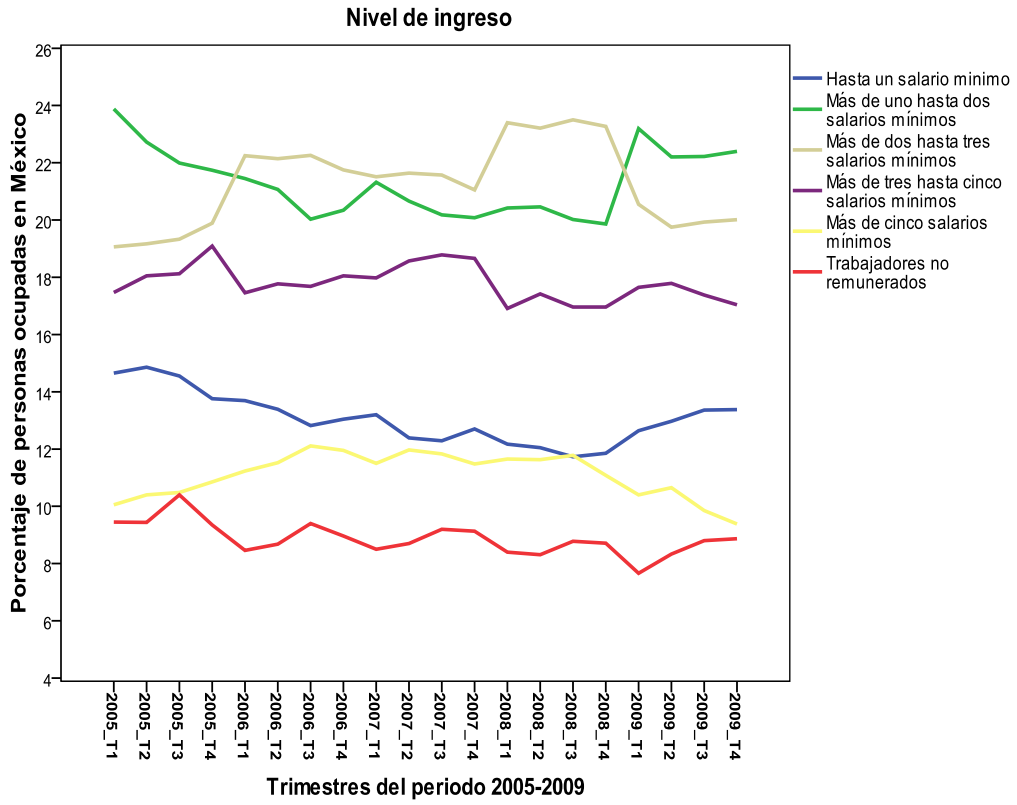


Figura 2: Gráfico de líneas del nivel de ingreso en periodo de tiempo trimestral (2005-2009) de la población en México.

Posteriormente, al analizar a la población ocupada en México por sector económico se observó que es en el sector primario donde se concentra el menor porcentaje de la población empleada, la cual oscila entre un 12.7% y de 15.42% de la población total. Seguido de este sector, encontramos que el sector industrial o secundario es el que absorbe entre un 23.4% y 26.1%, de la población ocupada. Y, como era de esperarse, es en el sector terciario o de servicios en donde se concentra la mayor parte de la población ocupada en México, la cual oscila un 59.2% y de 62.2% del total de la población (Ver figura 3). El sector terciario es el que emplea al mayor número de la población ocupada en México, seguido por el sector secundario y el primario, respectivamente.

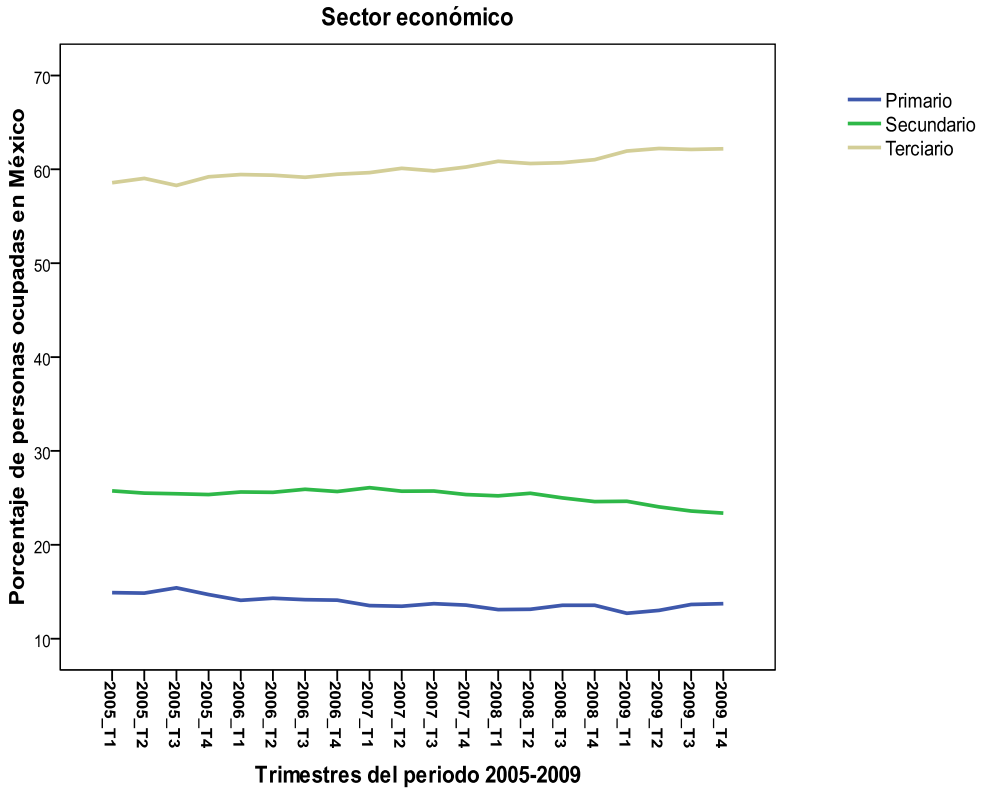


Figura 3: Gráfico de líneas del sector económico en el periodo de tiempo trimestral (2005-2009) de la población en México.

Por otra parte, seguido del análisis preliminar, se procedió a efectuar el Análisis de Correspondencia Simple entre cada una de las variables anteriormente analizadas en el periodo trimestral del 2005-2009. Primeramente, como resultado del análisis de correspondencia entre la posición ocupacional y los trimestres del periodo analizado, se aprecia que la mayoría de los puntos se encuentran muy cercanos al origen, de ahí que la variabilidad en las diferentes categorías que conforman la posición ocupacional ha sido muy reducida. No se observó ningún patrón de comportamiento en las diferentes categorías de la población ocupada en los trimestres (Ver figura 4).

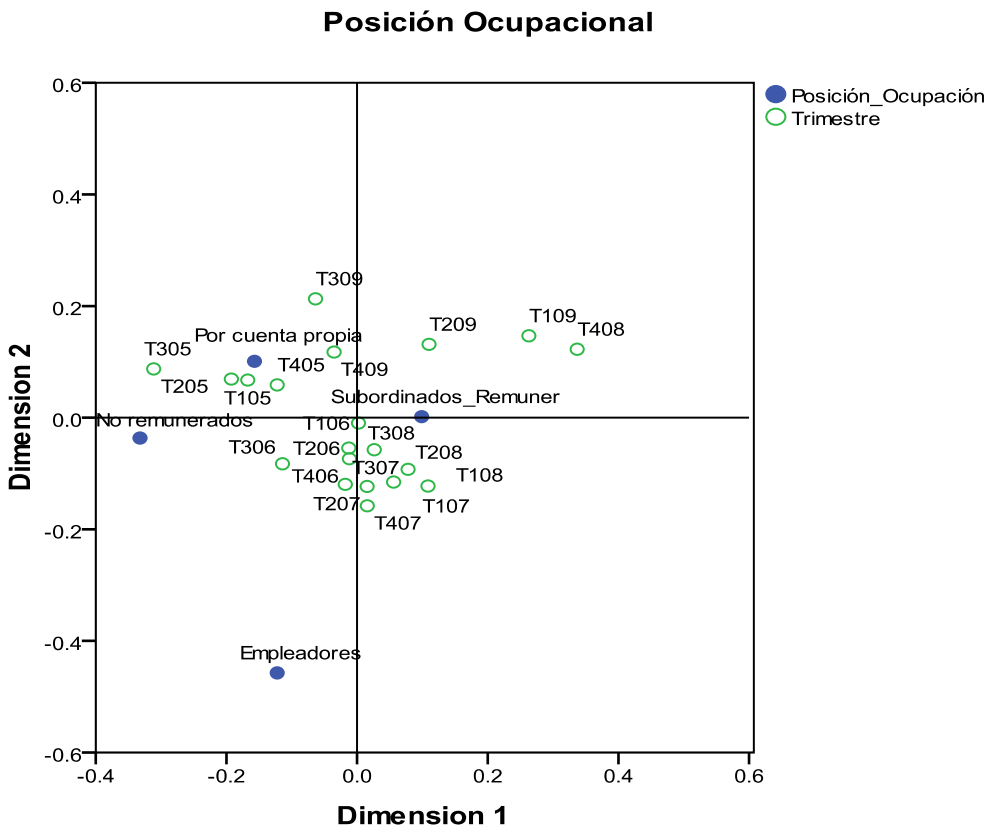


Figura 4: Gráfico de correspondencia simple en dos dimensiones de la posición ocupacional en el periodo de tiempo trimestral (2005-2009) de la población en México.

Conclusiones

Al cabo de la realización de este trabajo, una vez efectuados los análisis correspondientes para cumplir con el objetivo establecido, se obtuvo la evidencia necesaria para poder concluir que no existe una asociación o relación entre cada una de las variables analizadas que conforman el mercado ocupacional en México y los trimestres del periodo sujeto de evaluación. Las tres variables analizadas no reflejaron cambios o variaciones significativas en el lapso de tiempo estudiado. No se detectó alguna tendencia de crecimiento o decrecimiento, lo que muestra un mercado ocupacional que no ha cambiado en su integración; haciendo necesario la implementación de políticas que favorezcan el incremento de los puestos laborales al mismo tiempo que mejoren el nivel de percepciones.

La integración de la posición ocupacional ha permanecido estable, evidenciándose la necesidad de fomentar el espíritu emprendedor de las generaciones que se han incorporado recientemente a la población económicamente activa (PEA) y de establecer políticas de fomento y apoyo para la creación de pequeñas y medianas empresas. El nivel de ingreso de la población ocupada no ha sufrido cambios significativos en su integración, por lo que se hace necesario implementar políticas que promuevan no sólo la generación de nuevos puestos de trabajo, sino también el incremento de los niveles salariales. El incremento de puestos laborales debe ser acompañado con políticas de promoción de la productividad que permitan mejorar el nivel de ingresos de la población ocupada en México. Por último, el sector económico, al igual que las otras variables, no presentó cambios significativos en su composición siendo estable en los trimestres del periodo analizado y revelando la vocación de nuestro país al desarrollo del sector terciario.

En base a lo anterior, es importante mencionar que las estadísticas laborales son un instrumento que permite diseñar políticas públicas que contribuyen a la generación de nuevos empleos y al incremento de los ingresos de la población que se encuentra ocupada, coadyuvando al desarrollo económico de las naciones. Por lo que, el fomento del enfoque emprendedor debe ser una prioridad nacional, tanto de dependencias gubernamentales, como de instituciones educativas. Los Fondos de Apoyo para la Micro, Pequeña y Mediana Empresa (FONDO PYME) deben identificar proyectos con potencial de desarrollo y brindarles el apoyo financiero, en tecnologías de la información y comunicación (TIC´S) y administrativo necesario para que inicien sus operaciones, o mejoren su planta o procesos a fin de hacerlos más eficientes.

Referencias

Aparicio, A. (2007) *Reducción del desempleo a través de ALMP: el caso de México*. Análisis Económico, num. 47, vol. XXI, UAM, México.

Consejo Nacional de Población. (2001) *La Población de México en el Nuevo Siglo*. México, D.F.

Deville, J. C. y Malinvaud, E. (1983) *Data Analysis in Official Socio-economic Statistics*. Institut National de la Statistique et des Etudes Economiques, Paris. Journal of the Royal Statistical Society. Series A (General), Vol. 146, No. 4, pp.335-361.

Encuesta Nacional de Ocupación y Empleo (ENOE) *Indicadores Trimestrales de 2005 a 2009* en www.stps.gob.mx .

Fernández, F. J. (2002) *El uso del análisis de correspondencia simple como ayuda en la interpretación del dato en arqueología. Un caso de estudio*. Boletín Antropológico, Universidad de los Andes. Mérida, Venezuela.

Instituto Nacional de Estadística, Geografía e Informática. (2005) *Encuesta Nacional de Ocupación: ENOE.* Aguascalientes, Ags. México.

Instituto Nacional de Estadística, Geografía e Informática. Cómo *se hace la ENOE. Métodos y procedimientos*. (2007) Aguascalientes, Ags. México.

Ojeda, M. M., Díaz, J., Apodaca, C. y Trujillo, I. (2004). *Metodología de Diseño Estadístico.* Universidad Veracruzana. Xalapa, Ver. México.

Partida, V. (2008) *Proyecciones de la Población Económicamente Activa de México y de las Entidades Federativas, 2005-2050*. (eds.) Consejo Nacional de Población. México, D.F.

Subsecretaría de Empleo y Política Laboral, Secretaría del Trabajo y Previsión Social. Estadísticas Laborales en México, en www.observatorio**laboral**.gob.mx/pdf/**Estadisticas_Laborales**.pdf

ANEXO

Tabla 1. Descripción de las variables bajo estudio y sus categorías.

Variable	Categoría	Código	Descripción
Posición en el mercado ocupacional	Trabajadores		Se refiere a quienes desempeñan un trabajo subordinado a cambio de un salario
	Trabajadores por cuenta propia		Son aquellos que prestan servicios en forma independiente
	Empleadores		Incluye a quienes realizan un trabajo y tienen personal bajo su dirección
	Trabajadores no remunerados		Son los que laboran en una parcela o negocio familiar sin percibir una remuneración.
Nivel de ingreso	Hasta un salario mínimo	Salario<=1	
	Más de uno hasta dos salarios mínimos	1<Salario<=2	
	Más de dos hasta tres salarios mínimos	2<Salario<=3	
	Más de tres hasta cinco salarios mínimos	3<Salario<=5	
	Más de cinco salarios mínimos	Salario>5	
	Trabajadores no remunerados	No recibe	Este rubro incluye a trabajadores dependientes no remunerados y a trabajadores dedicados a actividades de subsistencia.
Sector de actividad económica	Primario		Este sector comprende agricultura, ganadería, silvicultura, caza y pesca
	Secundario		Este sector abarca la industria extractiva y de la electricidad y manufacturera
	Terciario		El sector incluye comercio, restaurantes y alojamiento, transportes, comunicaciones, servicios profesionales, financieros, sociales, gobierno y organismos internacionales.
Periodos Trimestrales	2005	T1	Primer trimestre
	2006	T2	Segundo trimestre
	2007	T3	Tercer trimestre
	2008	T4	Cuarto trimestre
	2009		

Análisis del gasto en salud y su relación con el crecimiento económico de México en el periodo 2000-2008

Dolores Mayo Lara
Fernando Velasco Luna

Resumen

En la presente investigación se analiza la evolución del gasto público en salud (*GPS*) y del Producto Interno Bruto (*PIB*), por entidad federativa. Paralelamente, a través de una modelación jerárquica se determinó la relación entre *GPS* y el *PIB*, para analizar si hay variabilidad entre esta relación durante el periodo 2000-2008 por entidad federativa, teniendo como resultado, que sí existe una relación directa entre el gasto en salud y el crecimiento económico, y que sí hay variabilidad significativa entre los años bajo estudio y entre las entidades federativas. Por lo anterior, se concluye que los factores que explican esta variabilidad se encuentran a nivel de cada entidad federativa y en cada año, lo que implica la necesidad de realizar estudios con otras variables explicatorias a estos niveles.

Palabras clave: Producto Interno Bruto, Gasto en Salud, modelo lineal jerárquico, entidad federativa.

Abstract

In this research, we analyze the evolution of public expenditure on health (*GPS*) and the Gross Domestic Product (*PIB*), both nationally and by state. In parallel, we apply a multilevel linear model to determine the relationship between GPS and *PIB*. The results indicate that there is a direct relationship between health spending and economic growth, and that there is significant variability between the years under study and among the states. Therefore, we conclude that the factors that explain this variability was found at the level of each state and each year, which implies the need for studies with other explanatory variables at these levels.

Keywords: Gross Domestic Product, health spending, hierarchical linear model, federal entity.

Introducción

La salud es un estado de bienestar muy importante para toda la población, en la medida en que refleja las condiciones físicas de potencialidad que permiten a los individuos desempeñar sus actividades sin que factores de enfermedad se interpongan y disminuyan su rendimiento, por lo tanto, la salud tiene un papel determinante y decisivo para el buen desempeño de la economía. Así pues, mientras al concepto de salud le corresponde directamente la conservación de la vida humana y el desarrollo de capacidades y potencialidades de la propia colectividad, su concepción teórico-práctico asume la identificación de necesidades, intervenciones y propuestas para su mantenimiento o en su caso para su cambio.

Desde la perspectiva de la doctrina francesa de las finanzas públicas, la función principal del Estado debe estar orientada hacia el ámbito social traducido en la satisfacción de necesidades colectivas. Es decir, el Estado debe realizar acciones encaminadas a proteger y salvaguardar las necesidades de la población (Gaudemet, 1996). Por tanto, es necesario que el Estado organice sistemas y recursos con el fin de detectar, ordenar, controlar y vigilar situaciones y hechos que puedan tener un efecto directo sobre las condiciones de salud de la población. En México, de acuerdo con lo establecido en la Constitución Política de los Estados Unidos Mexicanos, el Estado tiene la responsabilidad de emitir leyes conducentes para regular las acciones de salud en general. Desde tiempos remotos, la economía mexicana ha estado y está en la búsqueda constante para establecer soluciones a los problemas que imperan en el país, problemas que han evolucionado y que cada vez van tornándose más complicados debido a diversos factores.

El tema de la salud y su relación con la economía es en sí basto, por lo que ha sido estudiado por diversos autores desde diversos enfoques. Por su parte, la Fundación Mexicana para la Salud (Funsalud, 2006), realizó un estudio de la situación de salud en México, en el cual considera de suma importancia avanzar hacia el objetivo de diseñar e implantar una política pública de Estado en materia de salud. En dicha reforma de Estado, Funsalud propone una reforma financiera para alcanzar el acceso universal a la protección financiera y ordenar el financiamiento del sistema, que tenga como meta movilizar recursos

adicionales para la salud y garantizar un financiamiento justo de la atención evitando gastos de los hogares en esta materia.

Molina y Carvajal (2003) realizaron un estudio a través de un modelo econométrico, en el cual, analizaron el financiamiento público en los gastos de salud y su nivel de eficiencia medida a través de los principales indicadores macroeconómicos de salud, de ingreso y variabilidad. Estos autores concluyeron que el objetivo del mejor funcionamiento del sistema de salud no se logrará con base a la disminución del gasto público en salud. La disminución de la participación pública ha provocado un incremento en la desigualdad en el ingreso y el deterioro de los niveles de salud. Gálvez (2008), por su parte, realizó una investigación en la cual concluyó que la salud y la economía constituyen un binomio que se relaciona de forma activa. Posiblemente no exista una decisión en salud que no tenga una implicación económica. La interacción entre la economía y la salud se puede apreciar desde dos perspectivas diferentes. La primera se evidencia a través del impacto que tiene el sistema de salud como condicionante del bienestar de la población, como determinante de la productividad del trabajo y en la formación de capital humano; la segunda, a través de la influencia del sistema de salud de manera cuantitativa y cualitativa en el crecimiento de la economía nacional, lo que refuerza su importancia como sector económico.

El objetivo general de este estudio es explicar cuál ha sido el impacto del gasto en salud en la economía nacional, así como analizar la evolución del *PIB* y del *GPS* en México durante el periodo 2000-2008; observar el comportamiento del *PIB* por entidad durante este periodo, así como analizar el comportamiento del *GPS* para las 32 entidades de la República Mexicana, e identificar qué tipo de relación guarda el *GPS* y el *PIB*.

El presente trabajo esta divido en tres secciones; en la primera se encuentra la introducción al tema y al problema abordado en la investigación; en la sección segunda se describe la metodología implementada para la elaboración de esta investigación. En la sección tercera se presentan los resultados obtenidos con el estudio realizado y la discusión sobre investigaciones previas en esta materia. Por último se presentan las conclusiones a las que se llegaron tras realizar los análisis correspondientes.

Metodología

Este es un estudio observacional, en el cual se estudia la relación que existe entre el *GPS* y el *PIB*. Los datos utilizados para la realización del análisis fueron tomados de la página electrónica del Sistema Nacional de Información en Salud (SINAIS), entidad dependiente de la Secretaría de Salud (SSA), así como del Instituto Nacional de Estadística y Geografía (INEGI), contando con una muestra de 288 observaciones, correspondientes a la información del *PIB* y del *GPS*, en el periodo 2000-2008, de cada entidad federativa.

Se realizó un análisis exploratorio sobre el comportamiento del *GPS* y el *PIB* desde 2000 hasta 2008, en cada una de las 32 entidades federativas. Se utilizó un gráfico de perfil en el tiempo para observar el comportamiento del *PIB* a lo largo del periodo de estudio entre las 32 entidades; así también se emplearon gráficas de cajas y alambres del *PIB* y del *GPS* considerando las 32 entidades federativas. Todo esto permitió hacer una descripción en el tiempo y en todo el país.

Dado que la información que se obtuvo presenta una estructura de anidamiento, y se desea modelar la relación existente entre el *PIB* de cada entidad por año con el *GPS* de cada entidad por año, se hizo uso de la modelación jerárquica, haciendo uso de un modelo de dos niveles (Goldstein, 1999; Raudenbush y Bryk, 2002), como unidades de nivel 1 se tomaron los 9 años que comprende este estudio y como unidades de nivel 2 las 32 entidades federativas (Figura 1).

Figura 1. Diagramas de unidad para la estructura jerárquica de los datos bajo estudio.

A través de la modelación jerárquica, se pretende tener una mejor comprensión de la variabilidad del *PIB*, pues permite conocer la varianza entre los años y las entidades federativas respecto al *PIB*, tomando en consideración la posible relación con el tiempo y el *GPS*. El modelo para ajustar la relación está dado por medio de la ecuación:

$$
\begin{aligned}
&PIB_{ij} = \beta_{0j} + \beta_1 TIEMPO + \beta_2 GPS_{ij} + e_{ij} \\
&\beta_{0j} = \beta_0 + u_{0j} && i = 1,...,9 \\
&e_{ij} \sim N\left(0, \sigma_e^2\right) && j = 1,...,32 \\
&u_{0j} \sim N\left(0, \sigma_{u0}^2\right)
\end{aligned}
$$

donde β_0 denota el intercepto o la media global del *PIB* para todas las entidades federativas en todos los años; β_1 y β_2 constituyen la pendiente o el cambio en la media del *PIB*, cuando hay un cambio unitario en cada variable explicatoria *TIEMPO*, y *GPS*, respectivamente, manteniendo las otras variables constantes, e_{ij} denota el error aleatorio correspondiente a la i-ésima unidad de nivel 1 en la j-ésima unidad de nivel 2 y u_{0j} denota el j-ésimo error aleatorio a nivel 2. Con este modelo, lo que interesa es conocer si alguna variable como el *TIEMPO* o el *GPS* influyen en el comportamiento del *PIB*. Para validar los resultados del modelo, se comprobó el cumplimiento de los supuestos de normalidad de los errores en los dos niveles,

Resultados y discusión

Del análisis preliminar de los datos obtenidos es posible observar el comportamiento del *PIB* para las 32 entidades de la República Mexicana, durante el periodo 2000-2008. Se tiene que de las 32 entidades federativas las que registraron mayores niveles de crecimiento económico fueron el Distrito Federal, Estado de México y Nuevo León, ocupando el primer, segundo y tercer lugar respectivamente, durante el periodo de estudio 2000-2008, lo anterior se puede apreciar en la Figura 2. Además se observa que el *PIB* presenta un ligero incremento a través de los años, este incremento se da en la media aunque no en la varianza como se puede apreciar en la Figura 2.

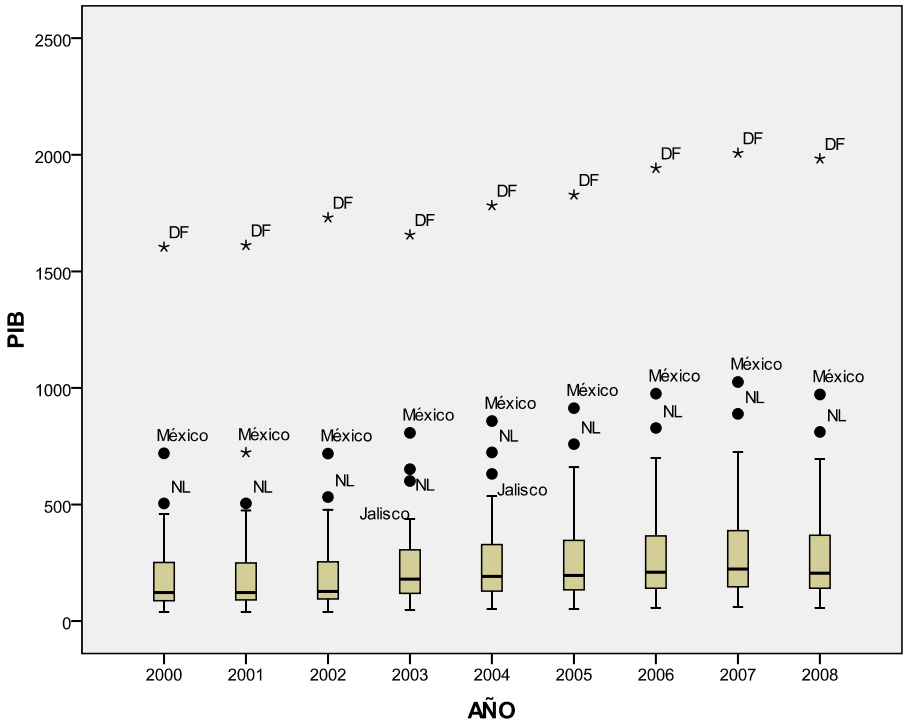


Figura 2. Comportamiento del *PIB* durante el periodo 2000-2008.

El *GPS* ejercido en el interior de la república mexicana tuvo un comportamiento similar al *PIB*, aquí las entidades que destinaron mayores monto de recursos a la salud fueron el Distrito Federal, Estado de México y Jalisco, durante el periodo 2000-2005, ocupando el primer, segundo y tercer lugar respectivamente, mientras que para el periodo 2006 2008 fueron las entidades federativas del Distrito Federal y México, lo anterior se aprecia en la Figura 3. Además se observa que el *GPS* presenta un ligero incremento a través de los años, este incremento se da tanto en la media como en la varianza como se puede apreciar en la Figura 3.

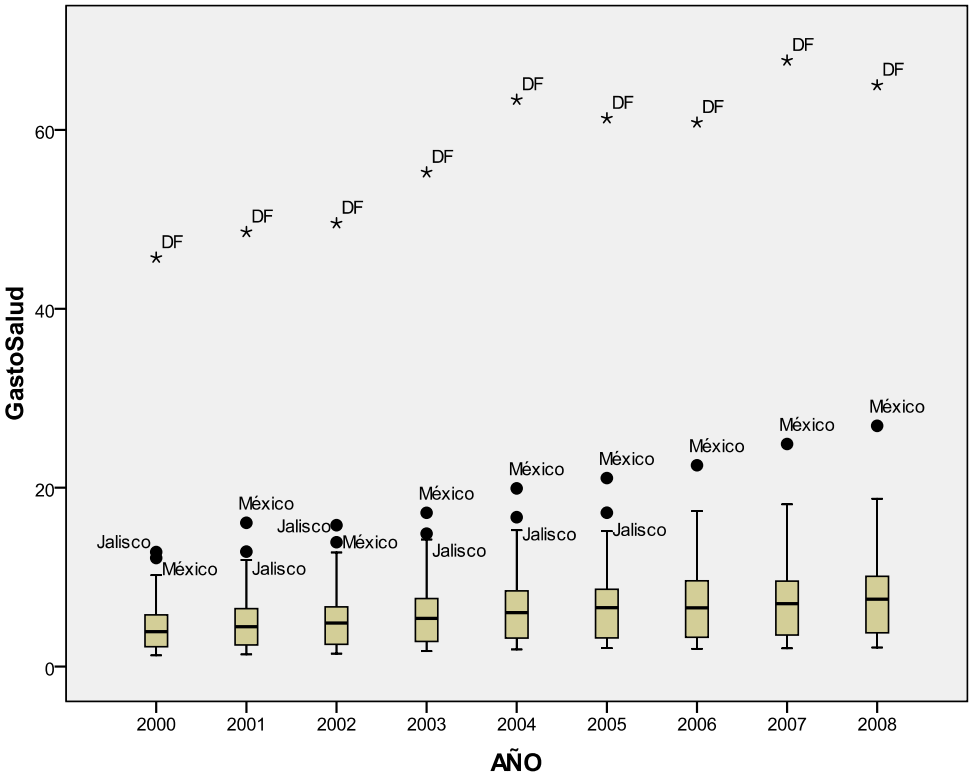


Figura 3. Comportamiento del *GPS* durante el periodo 2000-2008.

De las gráficas de dispersión realizadas a las 32 entidades fue posible determinar que existe relación directa entre el *GPS* y el *PIB*, tal como lo muestra la Figura 4, la cual muestra la relación existente en la entidad federativa del Estado de México. Cabe destacar que el comportamiento fue similar en las 32 entidades federativas, por lo cual únicamente se presenta la grafica referente a esta entidad.

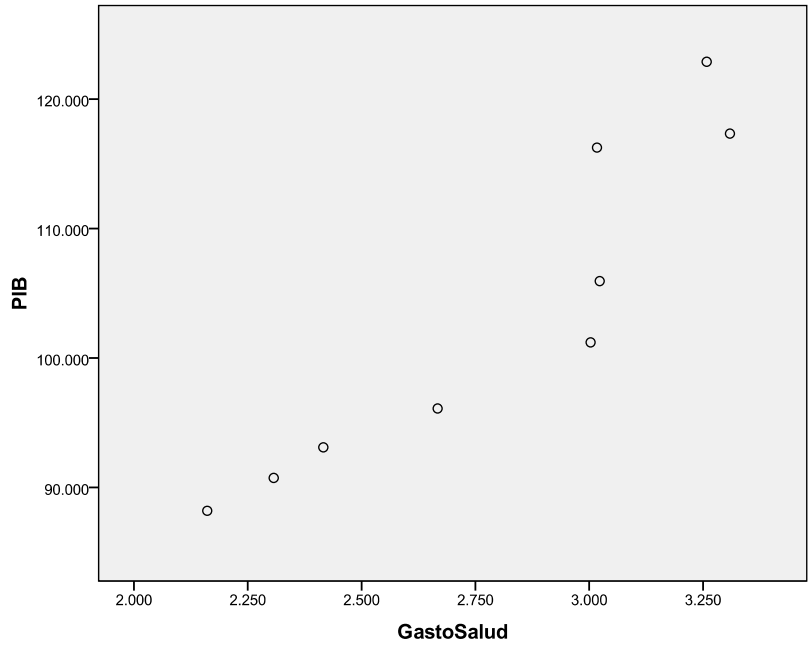


Figura 4. Estado de México: Relación entre el *GPS* y el *PIB* 2000-2008.

Al relacionar el *PIB* con los años del periodo de estudio para cada entidad federativa, se aprecia en la Figura 5 una primera aproximación de la relación lineal existente entre ambas variables. De la Figura 5, se observa una tendencia de crecimiento a través de los años del *PIB*, además se observa que hay una variabilidad entre las entidades respecto al *PIB* la cual se mantiene durante el periodo de estudio 2000-2008.

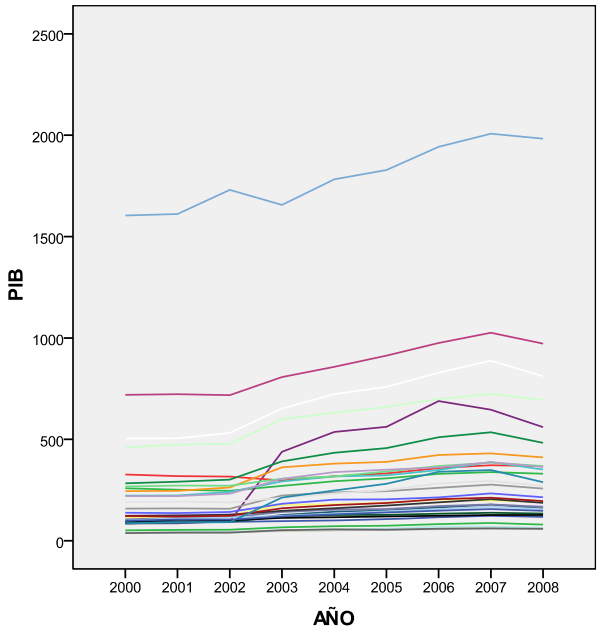


Figura 5. Relación entre el *PIB* y el tiempo (2002-2008) para cada entidad federativa.

Para corroborar los factores que contribuyen a explicar por qué hay variación entre los años y entre las entidades federativas respecto al *PIB*, se ajustaron 4 modelos multinivel, utilizando el método de Mínimos Cuadrados Generalizados Iterativos. Los resultados de las estimaciones se muestran en la tabla 1. En el modelo (1), modelo intercepto aleatorio, los resultados del ajuste muestran que se tiene un *PIB* en promedio de 292.745 millones de pesos en cada entidad federativa por año, además de que existe variación tanto entre los años como entre las entidades federativas, El porcentaje de la variabilidad del *PIB* atribuida a las entidades federativas es de aproximadamente el 95% y solo un 5% a los años. En el modelo (2) se introdujo la variable *AÑOS* como variable explicatoria, se mantuvo fija la pendiente y el intercepto aleatorio, los resultados del ajuste muestran que la variable *TIEMPO* sí resulta significativa, es decir, que cada año el *PIB* de las entidades federativas se incrementa en promedio en 18.69 millones de pesos. También se observa que la variación entre los años y entre las entidades es significativa. Sin embargo, la varianza entre las entidades se mantiene alta (57.495), mientras la varianza del *PIB* a nivel de los años disminuyo de 5461.39 a 3257.7.

Tabla 1. Resultados de las estimaciones.

	Modelo intercepto aleatorio (1)	Modelo intercepto aleatorio con el *TIEMPO* (2)	Modelo intercepto aleatorio con el *TIEMPO* y *GPS* (3)	Modelo intercepto aleatorio con el *TIEMPO* y *GPS,* sin datos atípicos influyentes (4)
Parámetros fijos				
β_0 (Intercepto)	292.745 (58.00)	198.05 (58.2)	259.9 (22.3)	276.65
β_1 (*TIEMPO*)		18.69 (1.303)	6.32 (1.401)	1.639
β_2 (*GPS*)			29.97 (1.49)	32.04
Componente de la varianza				
Nivel 2 σ^2_{u0}	107054.023	106952.922	14933.551	5008
Nivel 1 σ^2_e	5461.385	3257.703	1136.818	894.96
Deviance				
-2*loglikelihood	4167.062	3329.02	2903.018	2723.89

En el modelo (3), se introdujo adicionalmente el *GPS* y se modela como fija. En los resultados mostrados en la Tabla 1, se observa que el *GPS* es significativo, esto quiere decir, que ante un cambio unitario en el *GPS* de cada entidad, el *PIB* se incrementan en 29.97 millones de pesos, manteniendo la variable *TIEMPO* fija, cabe destacar que al introducir la variable *GPS* al modelo, el *TIEMPO* influye de manera distinta, ahora cada año el *PIB* de cada entidad federativa se incrementa en promedio en 6.32 millones de pesos. Al comparar los modelos (2) y (3), se aprecia que la varianza a nivel entidad disminuyó de 106952.9 a 14933.6, y la varianza a nivel año también presenta una reducción de 3257.7 a 1136.8. También hay una disminución en el valor de la deviance de 3329 a 2903 es decir una reducción de 426, que al compararlo con una distribución χ^2 con 1 grado de libertad, resulta significativa. Lo que indica que el modelo (3) es más adecuado para el ajuste de los datos.

Sin embargo, en un análisis de los residuos, se observó la presencia de datos atípicos, de una entidad: Campeche. En la Figura 6, se observa el comportamiento de los residuos y la presencia del dato atípico al extremo inferior del gráfico.

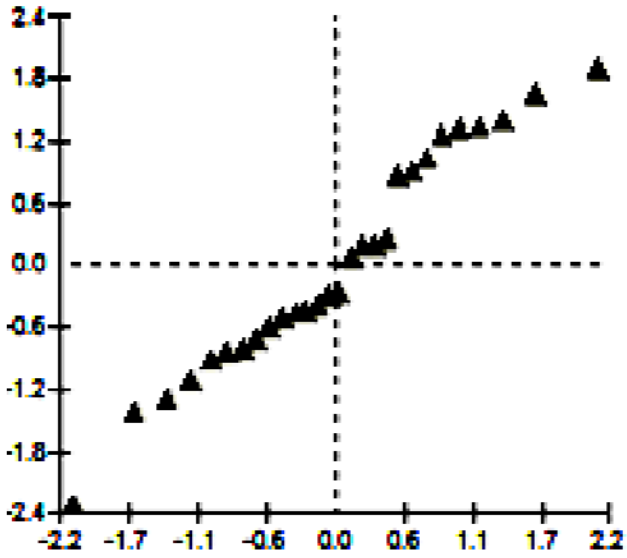


Figura 6. Gráfico de los residuos del modelo.

Para analizar si éste dato altera las estimaciones del modelo planteado, se incluyó esta entidad como variable indicadora, obteniéndose como resultado que en el nuevo

modelo, el valor de la deviance disminuyó de 2903 a 2723.89, 179 puntos con 1 grado de libertad por el parámetro extra incluido, lo que resulta significativo.

Comparando este modelo (4), con los modelos anteriores, se observa que el modelo (4) es el que mejor se ajusta a los datos, pues se reduce la varianza entre las entidades considerablemente de 14933.6 a 5008, y la varianza entre los años disminuye de 1137 a 895. Así, el *TIEMPO* y el *GPS* presentan influencia en el comportamiento del *PIB*, siendo que por cada año transcurrido del periodo 2000 a 2008 el *PIB* se incrementó en aproximadamente 1.64 millones de pesos, y que por cada millón de pesos que se incrementó el *GPS* el *PIB* se incrementó en aproximadamente 32 millones de pesos.

Con los resultados obtenido en el presente estudio, se deja claro que la salud no es sólo una necesidad más de la población, sino que además constituye una necesidad de la propia economía, por lo tanto, si lo que se busca es elevar los niveles de productividad y con ello mejorar la actividad económica del país es primordial poner atención en la salud de la población. Por lo que, el Estado debe postular y poner en su agenda las discusiones y reformas necesarias a fin de garantizar a la población el respaldo de un servicio de salud adecuado a sus necesidades. Esto, debido a que mejores resultados en materia de salud, beneficia el progreso social de la población en general.

Cabe mencionar que la situación de salud de una población entra directamente en las medidas sobre el nivel de vida en términos económicos al ponerse de manifiesto los términos de prestación de servicios, cobertura, accesibilidad, calidad y monto; en donde este último, representa el factor clave en la dotación de servicios de salud, puesto que si los recursos son insuficientes, los responsables de la salud pasarán serias dificultades a la hora de cumplir su misión, por lo que es recomendable tener un adecuado nivel socioeconómico en términos generales que permitan solventar la actividad de cualquier sector salud.

Además, se deben de conocer todos y cada uno de los elementos de la población como su composición, estructura, densidad, etc., esto con la finalidad de que las medidas que se tomen sean las correctas para su funcionamiento social y adecuado económicamente hablando, es preciso conocer el monto de recursos necesarios para efectuar dichas acciones.

Conclusiones

El presente estudio se basó en el enfoque de las finanzas públicas de la doctrina francesa, reforzando así la participación del Estado como proveedor del servicio de salud para la población. Las conclusiones presentadas en el presente apartado, son resultado del estudio realizado al *GPS* y su relación con el crecimiento económico, concluyendo que la salud y la economía están unidas y se complementan entre sí, ya que un nivel adecuado de salud permite a otros sectores de la economía un buen funcionamiento y a su vez, una economía que garantice el sostén de la estructura económica de la población hace que se mantengan funcionando las instituciones prestadoras de los servicios de salud, lo que corrobora la hipótesis de Gálvez (2008).

Por otro lado, el *GPS* es considerado una inversión en capital humano, mismo que trae múltiples beneficios como son: mejora en la productividad, reducción de incapacidades y mejora en la calidad de vida de la población, lo que posibilitaría un mayor crecimiento económico y mayor bienestar social. Con base en lo anterior, es posible hacer las siguientes recomendaciones: Es preciso contar con el financiamiento adecuado para la impartición de servicios de salud, es decir, mantener y mejorar los niveles de gasto público y de inversión en salud, puesto que aún siguen teniendo el carácter de bien público.

Los ejes de las acciones de salud siempre han sido incrementar los recursos materiales del sector salud, sin embargo, es necesario tener en cuenta los cambios registrados en la población, cambios que van desde su tamaño hasta su estructura y ubicación geográfica, por lo que es primordial que el futuro en materia de salud, esté fundado en la realidad económica, ya que eso permitirá conocer el verdadero impacto sobre la población en términos reales.

Resulta necesario seguir impulsando políticas de racionalización del gasto, ya que la dinámica económica continua siendo precaria, por lo que el gasto en salud tiene que ser destinado para lo que se necesita realmente y que puede catalogarse en un primer momento como ampliación de la cobertura y después en mantenimiento del propio sector. Así mismo, es necesario buscar nuevas alternativas que tengan que ver con la inversión en medicina preventiva, ya que la mayor parte del *GPS* es destinado a curación.

Referencias

Fundación Mexicana de la Salud. (2006). Salud en México 2006-2012

Gálvez, G.A. (2003). Economía de la salud en el contexto de la salud pública cubana. *Revista Cubana Salud Pública*. 29(4).

Gaudemet, P-M. et Moliner, J. (1997). *Finances Publiques, tome II emprunt/fiscalité*, Montchestien, France.

Goldstein, H. (1999). *Multilevel Statistical Models*. London. First Internet Edition.

Instituto Nacional de Estadística y Geografía (2010). Consultado el día 24 de enero de 2011, desde http://www.inegi.gob.mx

López, L.M.C. (1993). *Salud Pública*. México: Interamericana, McGraw-Hill.

Molina, S. R. y Carbajal de Nova, C. (2003). Salud y Desigualdad, el caso de México. *UAM-Unidad Iztapalapa.*

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Second edition, Newbury Park, CA: Sage.

Sistema Nacional de Información en Salud (2010). Consultado el día 8 de noviembre de 2010, desde http://sinais.salud.gob.mx/indicadores/index.html.

Influencia del sector eléctrico y petrolero en la producción primaria 2003-2008

Oscar Omar Núñez Herrera
Fernando Velasco Luna

Resumen

En este trabajo se analizó la relación que existe entre los ingresos del sector eléctrico y del sector petrolero en relación con el sector primario; para la obtención de la información se recurre a una base de datos del INEGI, respecto a los ingresos de petróleo y de energía eléctrica desde 2003 hasta 2008, en las 32 entidades federativas. Se aplicaron una serie de modelos multinivel para analizar la influencia del tiempo, de los ingresos del sector eléctrico y de los ingresos del sector petrolero en el sector primario y determinar si existe variabilidad entre las entidades federativas y los 6 años del periodo de estudio. Se encontró que sólo el tiempo y los ingresos del sector eléctrico influyen en el comportamiento de los ingresos del sector primario, y que existe variabilidad tanto a nivel entidad federativa como entre los años del periodo 2003 2008. Además que la entidad federativa Jalisco presenta un comportamiento distinto a las demás entidades federativas en los ingresos del sector primario.

Palabras clave: Sector primario, electricidad, petróleo, ingresos, entidad federativa.

Abstract

This paper analyzes the relationship between electric sector income and sector petroleum income in relation to the primary sector, to obtain the information we use a database INEGI, on income of oil and electricity from 2003 to 2008 in the 32 states. We applied a series of multilevel models to analyze the influence of time, the electricity sector income and income from the oil sector in the primary sector and whether there is variability among states and the 6 year study period. It was found that only time and power sector income influence the behavior of the primary sector income, and that variability exists both as a federal entity between the period 2003 2008. In addition to the federal entity Jalisco has a different response to other federal entity in the primary sector income.

***Keywords*:** Primary sector, electric sector, oil, income, federal entity.

Introducción

Dentro de la planeación económica y política de los ingresos públicos se puede hablar de recursos provenientes de la tributación que aporta la población con el fin de cumplir sus funciones públicas, o bien, por el producto de los ingresos que le otorgan los entes estatales que aprovechan recursos de la nación, como es el caso en México de Petróleos Mexicanos (*PEMEX*) y Comisión Federal de Electricidad (*CFE*).

La teoría que justifica este tipo de intervención del Estado a través de una empresa pública es la del Estado de Bienestar de Keynes (1981), para poder estabilizar los impactos negativos que ha traído consigo el capitalismo mismo, no tanto de una manera "prudente" sino necesaria a las necesidades de cada país, haciendo alusión a un análisis multi e interdisciplinario de todos los factores que inciden en las finanzas públicas de dicho país, con lo que se alinea a la visión de las Finanzas Públicas Modernas.

Hoy en día, uno de los sectores que más atención requieren por parte del Estado Mexicano es el sector primario (*SECPRIM*), dado que la desigualdad regional, la rápida liberalización comercial y la creciente generación e incorporación de innovaciones tecnológicas, han arrasado con la producción agropecuaria nacional, ocasionándose un abultado déficit agropecuario externo de acuerdo a la FAO (2011) de aproximadamente 3.5 millones de dólares anuales durante 2000 a 2003, lo cual refleja el estado de dependencia alimentaria en que se ha desenvuelto la economía mexicana por más de dos décadas.

Ello ha dado como resultado la agudización de los niveles de pobreza rural y un crecimiento sin precedente en los flujos migratorios del campo mexicano hacia los Estados Unidos, cuya magnitud se estimó por el Consejo Nacional de Población (2004) en 28 mil mexicanos por año hacia los Estados Unidos en la década de los setenta, en 138 mil en la de los setenta, 235 mil anuales durante las dos décadas siguientes y en 390 mil mexicanos por año para el periodo 2000-2002. Ante esta situación, la intervención del Estado Mexicano a partir del año 2002 ha sido a través de programas de estímulos económicos para los agricultores, como es el caso del Programa de Apoyos Directos al Campo (*PROCAMPO*), los cuales consisten en la aportación de una cantidad no muy decorosa para los agricultores siempre y cuando estos hayan destinado sus parcelas a la producción de algún grano básico de los establecidos en los ordenamientos o reglamentos respectivos. En este contexto, y ante la existencia de entes paraestatales que podrían intervenir y estimular la producción del

SECPRIM, y al mismo tiempo mejorar las condiciones de vida de la población rural del país, se cuestiona primeramente si la evolución de los resultados económicos de la *CFE* y *PEMEX* ha sido mejor que la del *SECPRIM* en México, si existe variabilidad entre estos resultados y si la producción del campo pudiera ser afectada por estos sectores energéticos.

Dentro de los antecedentes que tiene el presente estudio se puede enunciar las aportaciones de Willars (1989), quien realiza un estudio sobre la interrelación del sector petrolero y la economía mexicana en el periodo comprendido entre 1980 y 1987, encontrando que la economía nacional dependía considerablemente (90 %) de la producción petrolera, dado que no sólo se reflejaba en las contribuciones extraídas de la producción petrolera, sino también en la dependencia creciente de productos petroleros para producir bienes y servicios. Ante esta situación la deficiente canalización de recursos para la inversión en la actividad petrolera, reducían la capacidad de Petróleos Mexicanos para mantener sus entonces niveles de producción.

Aburto y Hudlet (1989) realizan un estudio sobre la estructura del balance energético, las participaciones del sector energético en la economía y la política nacional de energéticos, obteniendo que la participación del sector energético en el Producto Interno Bruto asciende de 2.8% en 1960 a 5.9% en 1985. En particular determina que el sector eléctrico manifiesta durante ese periodo una actividad más activa con una tasa de crecimiento de 1.9% para 1985, mientras que la de las ramas de petroquímica básica y refinación alcanzaron el 1.5% de crecimiento en ese mismo año.

Respecto a la producción del *SECPRIM* en México, los autores Del Valle y Lina (1996) realizan un estudio sobre la participación de este sector en el Producto Interno Bruto, del que obtiene una pérdida de importancia cuantitativa al pasar de un 8.2% al inicio de la década de los ochentas, a un 7.3% a finales de la misma década, fenómeno que se justifica con factores como la reducción del apoyo por parte del gobierno en inversión tecnológica agrícola, el desánimo para sembrar por parte de los campesinos, y el desplazamiento poblacional hacia las urbes.

López (2004) señala a la adopción de las tecnologías agrícolas contemporáneas, la expansión del riego, la apertura de nuevas tierras al cultivo y el desarrollo de sistemas de agricultura empresarial, como los principales detonantes de la desigualdad que presenta el campo mexicano; en el caso de Veracruz, la contribución del campo al PIB agropecuario

nacional ha perdido importancia, dado que en 1970 era de 10.3% , decreció en la década de los ochentas a 8.5%, y en la década de los noventas esa participación fue de tan solo el 7.5% anual.

En la segunda sección de este artículo se desarrolla un estudio de modelación multinivel para analizar la posible influencia que tienen los resultados de las paraestatales en el SECPRIM, se aplicaron 5 modelos multinivel para ver la influencia de las variables, tiempo, ingresos del sector energético e ingresos del sector eléctrico sobre los ingresos del *SECPRIM*, obteniéndose que el tiempo y los ingresos del sector eléctrico influyen sobre el comportamiento de los ingresos del *SECPRIM* durante el periodo 2003-2008, y que además existe variabilidad en las entidades federativas así como entre los años. Destacándose que la entidad federativa de Jalisco presenta un comportamiento distinto a las otras entidades federativas respecto a los ingresos del *SECPRIM* durante el periodo de estudio. En la tercera sección se discute sobre el irregular comportamiento del *SECPRIM* por los programas de estimulo económico que obtuvo del gobierno, comparado con un crecimiento positivo del sector eléctrico, mismo que se desarrolla por un ente paraestatal, y se destaca la superioridad de los productos petroleros en sus resultados y a su vez su autonomía respecto a las demás variables de estudio, concluyéndose un crecimiento sostenido del campo mexicano por parte del Estado, sin resultados suficientes para un desarrollo económico y una afinidad entre el comportamiento del *SECPRIM* con el eléctrico.

Metodología

Este es un estudio observacional, en el cual se estudia la relación que existe entre los ingresos del sector petrolero y de la energía eléctrica, con los ingresos del *SECPRIM*. La información se obtiene de la base de datos del INEGI, respecto a los ingresos de ventas totales de petróleo y de energía eléctrica de 2003 a 2008, así como los ingresos del *SECPRIM* en el mismo periodo.

Se realizó un análisis exploratorio sobre el comportamiento de los ingresos de *SECPRIM*, de los ingresos del sector petrolero y de los ingresos del sector energético desde 2003 hasta 2008, de cada una de las 32 entidades federativas. Se utilizó un gráfico de perfil en el tiempo para observar el comportamiento de los ingresos del *SECPRIM* a lo largo del periodo de estudio entre las 32 entidades; así también se emplearon gráficas de cajas y

alambres de los ingresos del *SECPRIM* y los ingresos del sector eléctrico. Todo esto permitió hacer una descripción en el tiempo y en todo el país.

Dado que la información que se obtuvo presenta una estructura de anidamiento, y se desea modelar la relación existente entre los ingresos del *SECPRIM* con los ingresos del sector eléctrico y del sector energético, se hizo uso de la modelación jerárquica, haciendo uso de un modelo de dos niveles (Goldstein, 1999; Raudenbush y Bryk, 2002). Como unidades de nivel 1 se tomaron los 6 años que comprende este estudio y como unidades de nivel 2 se tomaron a las 32 entidades federativas (Figura 1).

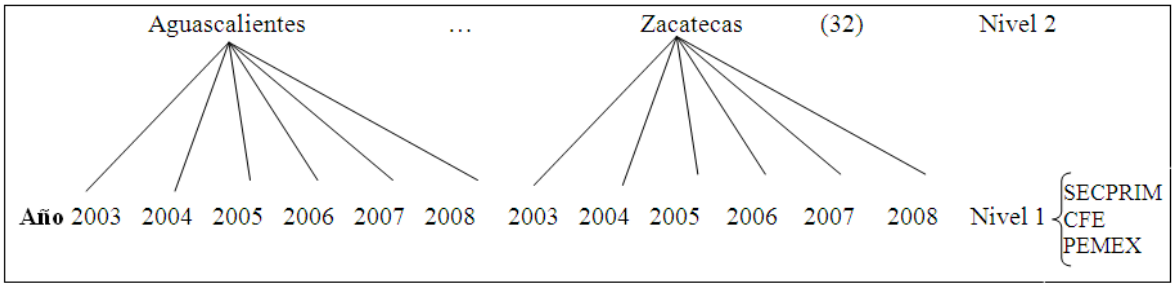


Figura 1. Diagramas de unidad para la estructura jerárquica de los datos bajo estudio.

A través de la modelación jerárquica, se pretende tener una mejor comprensión de la variabilidad de los ingresos del *SECPRIM*, pues permite conocer la varianza entre los años y las entidades federativas respecto a los ingresos del *SECPRIM*, tomando en consideración la posible relación con los ingresos del sector eléctrico y los ingresos del sector petrolero. El modelo propuesto está dado por la siguiente ecuación:

$$SECPRIM_{ij} = \beta_{0j} + \beta_1 TIEMPO_{ij} + \beta_2 CFE_{ij} + \beta_3 PEMEX_{ij} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j} \qquad i = 1,...,6$$
$$e_{ij} \sim N(0, \sigma_e^2) \qquad j = 1,...,32$$
$$u_{0j} \sim N(0, \sigma_u^2)$$

donde β_0 denota el intercepto o la media global de los ingresos del sector primario para todas las entidades federativas en todos los años; β_1, β_2 y β_3 constituyen la pendiente o el cambio en la media de los ingresos del *SECPRIM*, cuando hay un cambio unitario en cada variable explicatoria tiempo (*TIEMPO*), ingresos de la *CFE* e ingresos de *PEMEX*,

respectivamente, manteniendo las otras variables constantes, e_{ij} denota el error aleatorio correspondiente a la i-ésima unidad de nivel 1 en la j-ésima unidad de nivel 2 y u_{0j} denota el j-ésimo error aleatorio a nivel 2. Con este modelo, lo que interesa es conocer si alguna variable como *TIEMPO*, los ingresos de la *CFE* de la entidad o los ingresos de *PEMEX*, influyen en el comportamiento de los ingresos de *SECPRIM*. Para validar los resultados del modelo, se comprobó el cumplimiento de los supuestos de normalidad de los errores en los dos niveles,

Resultados y discusión

En la Figura 2 se aprecia que los ingresos del *SECPRIM* se mantuvieron constantes en el periodo de 2003 a 2008, tanto en mediana como en varianza. Siendo el valor de la mediana de aproximadamente de 8 miles de millones de pesos. Además, se observa que la entidad federativa de Jalisco durante el periodo de estudio siempre tuvo ingresos del *SECPRIM* por arriba de las demás entidades federativas y que además se presentó un incremento en sus ingresos a lo largo del periodo de estudio.

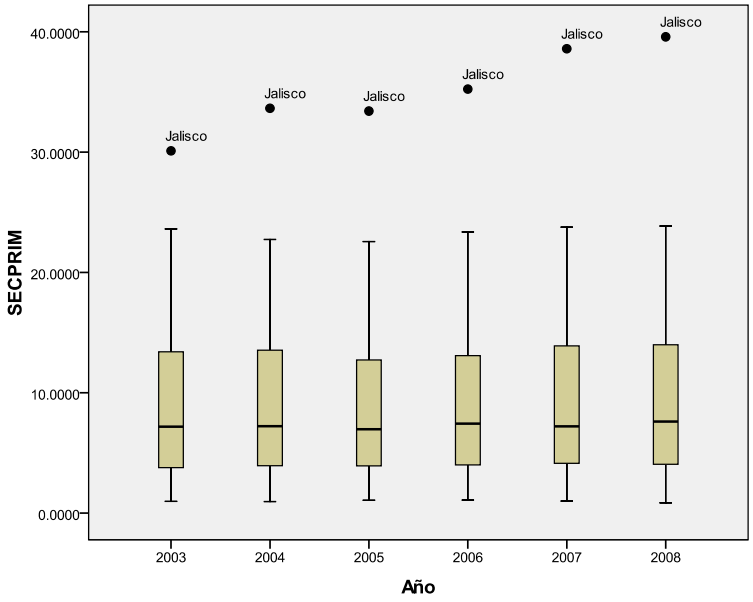


Figura 2. Tendencia y variabilidad de los ingresos del *SECPRIM* (2003-2008).

En la Figura 3 se aprecia que los ingresos de la *CFE* presentan una tendencia creciente durante el periodo de 2003 a 2008, tanto en mediana como en varianza. Además,

se observa que las entidades federativas de México y el Distrito Federal presentan mayores ingresos de la *CFE* durante los años 2003, 2004 y 2005.

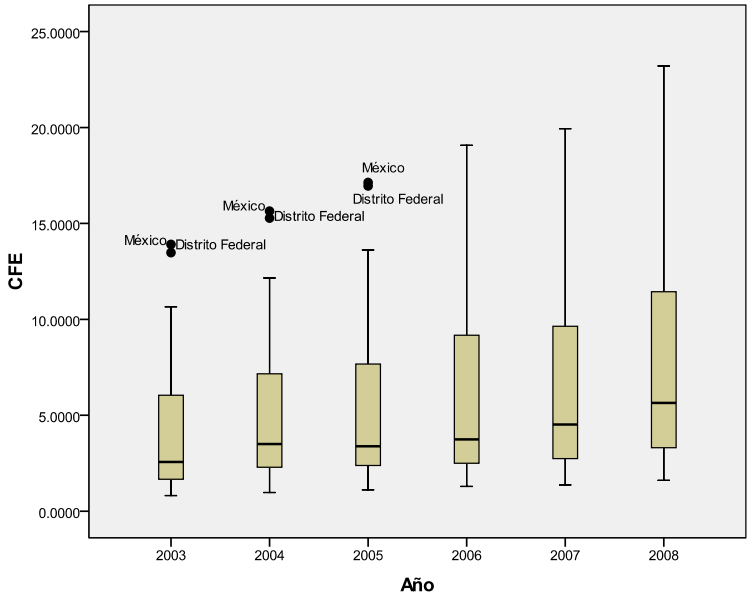


Figura 3. Tendencia y variabilidad de los ingresos de la *CFE* (2003-2008).

Al relacionar los ingresos del *SECPRIM* , con los años del periodo de estudio para cada entidad federativa, se aprecia una primera aproximación de la relación lineal existente entre *SECPRIM* y el tiempo. De la Figura 4, se tiene que no se observa una tendencia a través de los años, pero se observa que hay una variabilidad entre las entidades respecto a los ingresos del *SECPRIM* la cual se mantiene durante el periodo de estudio.

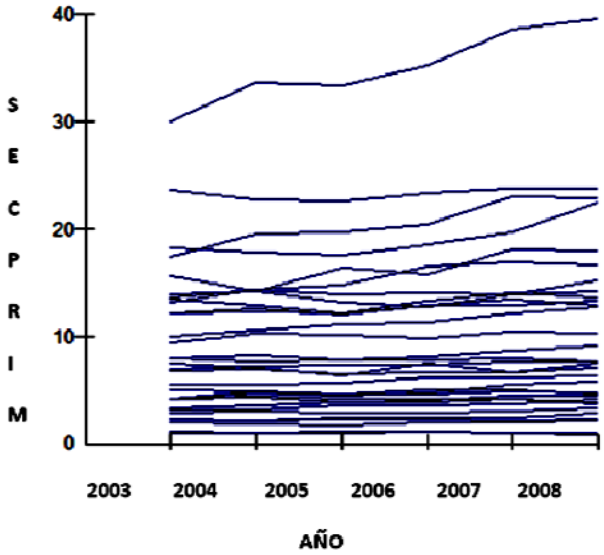


Figura 4. Relación entre los ingresos del *SECPRIM* y los años del periodo de estudio (2003-2008).

Para corroborar los factores que contribuyen a explicar la posible variación entre los años y entre las entidades federativas respecto a los ingresos del *SECPRIM*, se ajustaron 5 modelos multinivel, utilizando el método de Mínimos Cuadrados Generalizados Iterativos. Los resultados de las estimaciones se muestran en la Tabla 1. En modelo (1), modelo intercepto aleatorio, los resultados del ajuste muestran que se tiene un ingreso promedio de 9,457 millones de pesos en cada entidad federativa por año, además de que existe variación tanto entre los años como entre las entidades federativas, siendo aproximadamente el 98% de la variación de los ingresos atribuida a las entidades federativas. En el modelo (2) se introdujo la variable años como variable explicatoria, se mantuvo fija la pendiente y el intercepto aleatorio, los resultados del ajuste muestran que la variable *TIEMPO* sí resulta significativa, es decir, que cada año los ingresos del *SECPRIM* se incrementa en promedio en 0.254 miles de millones de pesos. También se observa que la variación entre los años y entre las entidades es significativa. Sin embargo, la varianza entre las entidades se mantiene alta (57.495), mientras la varianza de los ingresos del *SECPRIM* a nivel de los años disminuyó de 0.93 a 0.71.

Tabla 1. Resultados de las estimaciones.

	Modelo intercepto aleatorio (1)	Modelo intercepto aleatorio con el *TIEMPO* (2)	Modelo intercepto aleatorio con el *TIEMPO* y la *CFE* (3)	Modelo intercepto aleatorio con el *TIEMPO*, la *CFE* y *PEMEX* (4)	Modelo intercepto con el *TIEMPO*, la *CFE* y Jalisco (5)
Parámetros fijos					
β_0 (Intercepto)	9.475 (1.342)	6.811 (1.393)	8.882 (1.304)	7.503 (1.821)	8.872 (1.037)
β_1 ($TIEMPO$)		0.254 (0.036)	0.169 (0.053)	0.364 (0.377)	0.172 (0.052)
β_2 (CFE)			0.121 (0.056)	0.117 (0.056)	0.117 (0.055)
β_4 ($PEMEX$)				**-0.130 (0.250)**	

$\beta_{0\ Jalisco}$ (*Jalisco*)					25.733 (5.877)
Comp. de la varianza					
Nivel 2 σ^2_{u0}	57.458 (14.634)	57.495 (14.633)	53.203 (13.297)	53.721 (13.313)	33.229 (8.333)
Nivel 1 σ^2_e	0.933 (0.104)	0.712 (0.0799)	0.694 (0.078)	0.693 (0.077)	0.649 (0.078)
-2*logVerosimilitud	717.4	678.0	670.994	670.723	665.936

En el modelo (3), se introdujo adicionalmente los ingresos de la *CFE* y se modeló como fija. En los resultados mostrados en la tabla 1, se observa que la variable *CFE* es significativa, esto quiere decir, que ante un cambio unitario en los ingresos de la *CFE* de cada entidad, los ingresos del *SECPRIM* se incrementan en 0.121 miles de millones de pesos, manteniendo la variable *TIEMPO* constante, cabe destacar que al introducir la variable *CFE* al modelo, el *TIEMPO* influye de manera distinta, ahora cada año los ingresos del *SECPRIM* se incrementa en 0.169 miles de millones de pesos en promedio. Al comparar los modelos (2) y (3), se aprecia que la varianza a nivel entidad disminuyó de 57.495 a 53.203, y la varianza a nivel año también presenta una reducción de 0.712 a 0.694. También hay una disminución en el valor de la deviance de 678 a 670.994, es decir una reducción de 7.006, que al compararlo con una distribución χ^2 con 1 grado de libertad, resulta significativa. Lo que indica que el modelo (3) está mejor ajustado a los datos.

En el modelo (4) se introdujo adicionalmente los ingresos de *PEMEX* y se modela como fija. En los resultados mostrados en la tabla 1, se observa que la variable *PEMEX* no es significativa. Al comparar los modelos (3) y (4) hay una disminución en el valor de la deviance de 670.994 a 670.723 es decir una reducción de 0.271, que al compararlo con una distribución χ^2 con 1 grado de libertad, resulta no significativa. Lo que indica que los ingresos del sector petrolero no ayudan a explicar el comportamiento de los ingresos del *SECPRIM*, en el periodo bajo estudio.

Una vez que se ha realizado el ajuste del modelo, es importante corroborar el cumplimiento de los supuestos y realizar un diagnóstico de los datos atípicos. Como se había observado en la figura 4, la entidad federativa de Jalisco tiene un ingreso en el sector primario distinto, lo que pudiera tener una influencia en el modelo, lo que podría alterar el

valor de las estimaciones. Con tal finalidad, se obtuvo el gráfico de los residuos a nivel entidad, ver Figura 5.

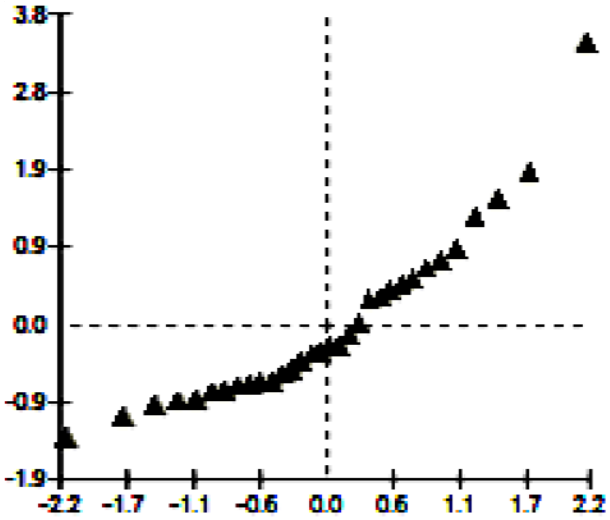


Figura 5. Gráfico de los errores a nivel entidad.

Se observa la existencia de datos atípicos en el extremo superior de la gráfica, la cual corresponde al estado de Jalisco, por lo que se ajusta el modelo (5), en el cual se ajusta el intercepto del estado Jalisco en forma separada, al evaluar su peso dentro de las estimaciones resulta significativa la estimación, es decir, los resultados de las estimaciones de los componentes de la varianza del modelo (5) cambian respecto al modelo (3).

Comparando este modelo (5), con los modelos anteriores, se observa que el modelo (5) es el que mejor se ajusta a los datos, pues se reduce la varianza entre las entidades considerablemente de 53.203 a 33.229, y la varianza entre los años disminuye de 0.694 a 0.649, e igualmente la reducción del valor de la deviance resultó significativa (670.994-665.936=5.058). Así, el *TIEMPO* y los ingresos de la *CFE* presentan influencia en el comportamiento de los ingresos del *SECPRIM*, siendo que por cada año transcurrido del periodo 2003 a 2008 los ingresos del *SECPRIM* se incrementaron en aproximadamente 0.172 miles de millones de pesos, y que por cada mil de millón de pesos que se incrementaron los ingresos de la *CFE*, los ingresos del *SECPRIM* se incrementaron en aproximadamente 0.117 miles de millones de pesos. Además se destaca que la entidad

federativa de Jalisco presenta un ingreso de 25.733 miles de millones de pesos en el *SECPRIM* más que las otras entidades federativas.

Del coeficiente obtenido para los ingresos de la *CFE* , el cual es de 0.117, se desprende que por cada mil millones de pesos que se incrementa el sector eléctrico, apenas el *SECPRIM* incrementa sus ingresos en 117 millones de pesos, es decir, los ingresos del *SECPRIM* son de aproximadamente del 18% de los ingresos de la *CFE* en el periodo bajo estudio.

Esto coincide con los estudios previos de Aburto y Hudlet (1989) en el sentido de que el sector eléctrico tiene influencia en la economía nacional, más esto se plasma de manera regional acentuándose las desigualdades económicas y productivas entre las entidades federativas.

Conclusiones

Como se ha apreciado, el campo mexicano tiene una deficiente productividad en comparación con el sector energético. Es de destacarse la intervención del Estado para mejorar esta situación, más no ha sido del todo acertada, puesto que a pesar de contar con un número considerable de programas agrícolas y una notable participación en el presupuesto federal, los resultados esperados no llegan a aterrizar en la realidad social del campo, lo que permite inferir vicios de funcionalidad de los organismos gubernamentales encargados de su aplicación, pudiendo pensar en la corrupción, el desvío de recursos o simplemente la deficiente función pública.

Lo que necesita el campo mexicano es la implementación de una política pública que logre coordinar todos estos aspectos para poder levantar este sector primario y volver a aprovechar sus recursos en beneficio de la sociedad mexicana; es cierto que la competencia es muy marcada y desigual, tanto en el exterior como en el interior, sin embargo el Estado Mexicano debe analizar la viabilidad de otros mercados diferentes al alimentario.

La variabilidad que presentan estos dos sectores llama mucho la atención, lo que denota un regionalismo que se traduce en una desigualdad económica y social en la población, con la necesidad del Estado de replantear sus políticas públicas al respecto para lograr una distribución más equitativa en todo el país tanto de los servicios energéticos como de los apoyos a la producción primaria. La empresa pública se establece como la

opción más viable, lo cual fue sustentado al verificar la eficiencia que han presentado *CFE* y *PEMEX* con la producción de electricidad y productos petroleros respectivamente; sin embargo, a pesar de que ambas cuentan con una infraestructura jurídica y organización similar, sus resultados varían dependiendo del mercado que cada una maneja, siendo el más favorecido el petrolero.

El comportamiento descrito por las paraestatales en estudio, y sobre todo sus resultados positivos, dan pauta a pensar en la viabilidad de la intervención activa del Estado en el campo mexicano a través de un ente paraestatal, mismo que sujeto a una infraestructura jurídica y orgánica, así como a condiciones exógenas a su producción, permitirá el aprovechamiento de los recursos que se pudieran obtener, y enfocarse en aquellos que tengan mayor demanda en el mercado internacional, dejando a su vez de lado la posibilidad de que los apoyos se regionalicen en determinadas entidades federativas, puesto que se trataría de una producción nacional con beneficios para todo el país.

Referencias

Aburto, A. J. y Hudlet, Y. R. (1989) "Algunas consideraciones sobre la estructura del balance energético, la participación del sector energético en la economía y la política general de ingresos" *Energía en México el arranque del siglo XXI*. Universidad Nacional Autónoma de México. México. pp 57-69.

Del Valle, M. y Lina, S.I. (1996) "Modernización y rezago tecnológico en el campo y las agroindustrias" *El cambio tecnológico en la agricultura y las agroindustrias en México* Siglo Veintiuno editores. México. pp. 51-94.

Goldstein, H. (1999) Multilevel Statistical Models. London. First Internet Edition.

Keynes, J. M. (1981) Teoría de la Ocupación, el Interés y el Dinero. Fondo de Cultura Económica, México.

López, D. V. (2004) "La agricultura veracruzana. Apertura y rezago regional" Estado, Economía y Hacienda Pública Número 6 Enero Junio 2004, pp. 115-133.

Raudenbush, S.W. and Bryk, A.S. (2002) Hierarchical Linear Models: Applications and Data Analysis Methods. Second edition, Newbury Park, CA: Sage.

Willars, A. J. (1989) "Consideraciones sobre la interrelación del sector petrolero y la economía en México" *Energía en México el arranque del siglo XXI*. Universidad Nacional Autónoma de México. México. pp 31-46.

www.conapo.gob.mx/prensa/2004/03boletin2004.htm, (2004) *Comunicado de Prensa*, México, 9 de enero 2004.

http://www.fao.org/docrep/x4400s/x4400s10.htm, (2011) *El estado mundial de la agricultura y la alimentación* "Los efectos sociales y económicos de la modernización de la agricultura". Depósito *de Documentos de la FAO. Departamento Económico y Social.*

Evaluación del Fondo de Aportaciones para la Infraestructura Social Municipal (FAISM) en el combate al rezago en infraestructura social de los municipios indígenas de Veracruz en el periodo 2000-2005

Arturo Abad Espíndola
Patricia Tapia Blásquez

Resumen

Uno de los fondos de aportaciones federales creados en 1998 por el Gobierno Federal que tiene como objetivo transferir recursos a los municipios para que éstos se encuentren en posibilidades de mejorar su infraestructura social para ir abatiendo las condiciones de pobreza extrema y de rezago social, es el Fondo de Aportaciones para la Infraestructura Social Municipal (FAISM). El objetivo de este trabajo es conocer y analizar si el FAISM ha contribuido en la mejoría de la infraestructura social de los municipios y regiones indígenas de Veracruz de 2000 a 2005, tomando como parámetro el Índice de Rezago en Infraestructura Social (IRIS) que elabora el Consejo Nacional de Evaluación (CONEVAL). Para tal fin se ajustó un modelo de regresión múltiple, y de esta manera determinar el efecto del FAISM y la región en la que se encuentra el municipio, sobre el IRIS. Se encontró que el FAISM influye significativamente en el IRIS; es decir, a medida que aumenta el fondo para el municipio, el índice de rezago se incrementa también. Asimismo, la región en la que se encuentra el municipio influye, por lo que se declararon diferencias entre regiones.

Palabras clave: Participación ciudadana, Regresión múltiple, Rezago social.

Abstract

One of the federal funds created in 1998 by the Government was the Contribution Fund for the Social Infrastructure of municipalities (FAISM), which has the purpose to transfer economical resources to the town councils in order to improve their social infrastructure, then struggle against poverty and decreasing the backwardness of rural areas. The aim of this work is to analyze if the FAISM has been contributing to develop the social infrastructure of indigenous municipalities in the state of Veracruz form 2000 to 2005 considering the Backwardness in Social Infrastructure Index (IRIS). With this purpose, we apply a multiple regression model to explain the effect of FAISM in IRIS. We conclude that the FAISM is statistically significant and the region in which the municipality is located does too. Also, the region the one that is the municipality influences, for what they declared differences between regions.

Keywords*: Civil participation, Multiple regression, Social backwardness.*

Introducción

En el marco del sistema nacional de coordinación fiscal durante el sexenio Zedillista se abrió el camino para la descentralización del dinero ejercido por la federación proveniente del Ramo 26, creando a partir de 1998, el Ramo 33, que tiene la finalidad de transferir recursos del presupuesto federal a los estados y municipios; destinadas estas transferencias a la atención de rubros tales como salud, educación y desarrollo social. Su importancia radica en que los fondos se ceden a autoridades locales por ley, y no por convenio, como sucedía hasta 1997. Uno de los fondos que constituye este Ramo 33 es el destinado para combatir la pobreza, el cual se denominó Fondo de Aportaciones para la Infraestructura Social. Este fondo a su vez, se integra con dos subfondos: El Fondo para la Infraestructura Estatal (FISE) y el FAISM. Para fines de esta investigación, el análisis se centrará en los recursos destinados de este último subfondo para el estado de Veracruz.

El objetivo del FAISM desde su creación ha sido el financiamiento de obras y acciones sociales básicas, a inversiones que beneficien directamente a sectores de su población que se encuentren en condiciones de rezago social y pobreza extrema. Dicho fondo cuenta con recursos etiquetados; es decir, no pueden usarse libremente por los municipios, deben ser designados exclusivamente en obras contempladas en la estructura programática que se determina con base en la Ley de Coordinación Fiscal y no al gasto corriente. Las principales áreas de aplicación son agua potable, alcantarillado, drenaje y letrinas, electrificación rural de colonias pobres, infraestructura básica de salud, infraestructura básica educativa, mejoramiento de vivienda, caminos rurales e infraestructura productiva rural.

La distribución del FAISM se realiza con base en una formula diseñada por la Secretaría de Desarrollo Social (SEDESOL) que toma en cuenta indicadores de pobreza extrema y marginación de los municipios. En la ley de Coordinación Fiscal sólo se pone a consideración del presidente municipal la posibilidad de someter a discusión de la ciudadanía la asignación de recursos de este fondo por medio del Comité de Planeación para el Desarrollo del Municipio (COPLADEMUN), sin embargo esta situación se agudiza al encontrarse que en el supuesto que se pongan a discusión, quien determina en última instancia es el presidente municipal.

Durante los últimos años, se han realizado trabajos de investigación cuyo objetivo ha sido analizar la descentralización financiera de los municipios Mexicanos, a través de la construcción de indicadores para explicar cómo se ha reflejado el gasto descentralizado del FAISM en indicadores como el índice de marginación. Al respecto, la Encuesta Nacional a Presidentes Municipales (2002) basándose en estadística social municipal del censo de 2000, establece que el FAISM contiene un criterio redistributivo débil, ya que en la asignación de los recursos que reciben los municipios por parte de los gobiernos estatales existe un espacio de discrecionalidad. También se han llevado a cabo investigaciones cuyo objetivo ha sido describir cómo se desarrolla la distribución de los recursos del FAISM de las entidades federativas a municipios rurales en un periodo de tiempo, como el que presenta Mathus (2008), quien describe la dinámica de asignación del FAISM de 1998 a 2005 en los municipios oaxaqueños.

En el presente trabajo, el objetivo es analizar y explicar la eficacia del FAISM en su interés por mejorar la infraestructura social de los municipios de Veracruz de 2000 a 2005, centrando la atención en las comunidades indígenas, por resultar los municipios más vulnerables a situaciones de pobreza extrema y los que requieren mayor inversión de recurso para mejorar su infraestructura. Para ello, se utilizará el IRIS indicador que construye el CONEVAL cada 5 años y que permite conocer el grado de rezago de infraestructura en que viven los municipios de las diversas entidades federativas y que refleja aspectos de carencia en infraestructura que por ley el FAISM debe de afrontar (Ley de Coordinación Fiscal, Art. 33).

El planteamiento del trabajo se expone a partir de que el proceso que se desarrolla por medio del COPLADEMUN para la planeación, vigilancia, asignación y supervisión al ejercicio del gasto descentralizado del FAISM (LCF, Art. 33), está influyendo negativamente en la eficacia para lograr los objetivos por los cuales fue puesto en marcha el FAISM, esto es, combatir el rezago social y la pobreza extrema mediante la inversión en infraestructura social, lo cual se refleja en el IRIS de los municipios indígenas de Veracruz de 2000 a 2005. Por lo cual es necesario conocer cómo ha contribuido el ejercicio del gasto a nivel municipal del FAISM en la infraestructura social en los municipios indígenas del Estado de Veracruz de 2000-2005.

Para cumplir con los objetivos, se definirá en primera instancia la estrategia metodológica a seguir y se explicará el proceso de obtención de datos. Posteriormente se muestran los resultados estadísticos, para finalmente exponer las conclusiones y principales discusiones.

Metodología

Los datos por municipio indígena de recursos económicos del FAISM del Estado hacia municipios indígenas se obtuvieron del Órgano de Regulación Financiera (ORFIS) de Veracruz expresados en millones de pesos. Con respecto a los datos del IRIS, estos se tomaron del indicador que construye el CONEVAL; en la composición del índice se consideran los siguientes indicadores: porcentaje de agua entubada, porcentaje de viviendas particulares sin agua entubada, porcentaje de vivienda sin drenaje, porcentaje de viviendas sin energía eléctrica y porcentaje de viviendas con piso de tierra. Para calcularlo el CONEVAL aplica la técnica de componentes principales, pues permite resumir en un indicador agregado las diferentes dimensiones del fenómeno en estudio. El índice resultante ordena las unidades de observación (localidad, municipio, estado) según sus carencias sociales. Es importante mencionar que, la información de estos indicadores es retomada del censo de población y vivienda elaborado por el Instituto Nacional de Estadística Geografía e Informática (INEGI). El IRIS se construye cada 5 años; es decir, se contó con este dato para 2000 y 2005, se recurrió a realizar una resta del IRIS 2005 menos el IRIS 2000 por cada municipio indígena de Veracruz, un IRIS negativo indica que el índice de rezago ha disminuido, mientras que sí es positivo ha aumentado; es decir, hay un mayor rezago en el 2005 que en el 2000. La diferencia resultante entre un año y otro, fue el dato que se utilizó para relacionarlo con el FAISM. Se analizaron los municipios veracruzanos catalogados como indígenas, que de acuerdo a la Comisión Nacional para el Desarrollo de los Pueblos Indígenas son 50 municipios, los cuales componen cuatro regiones indígenas de Veracruz, que son: Zongolica, Huasteca, Popoluca, y Totonaca. En la Tabla 1 se muestran cuáles son los municipios que integran cada región, así como el promedio del IRIS y del FAISM asignado para cada una.

Tabla 1. Estadísticas de IRIS y FAISM de las regiones y municipios indígenas del Estado de Veracruz.

Región	Municipios indígenas	IRIS (Promedio)	FAISM (Promedio)
ZONGOLICA (15)	Astacinga San Andrés Tenejapan Atlahuilco Soledad Atzompa Ixhuatlancillo Tehuipango Magdalena Tequila Mixtla de Altamirano Texhuacán Rafael Delgado Tlaquilpa Los Reyes Tilapan Zongolica	-0.22602	40.01
HUASTECA (14)	Benito Juárez Ixcatepec Citlaltépetl Ixhuatlán de Madero Chiconamel Platón Sánchez Chalma Tantoyuca Chicontepec Texcatepec Chontla Tlachichilco Ilamatlán Zontecomatlán de LyF Ixcatepec	-0.492207	84.60
POPOLUCA (9)	Cosoleacaque Soteapan Las Choapas Zaragoza Hueyapan de Ocampo Tatahuicapan de Juárez Mecayapan Uxpanapa Pajapan	-0.56705	83.32
TOTONACA (12)	Zozocolco de Hidalgo Chumatlán Tihuatlán Coyutla Papantla Coatzintla Mecatlán Coahuitlán Espinal Cazones Filomeno Mata Coxquihui	-0.76707	90.86
TOTAL	50	-0.49215	72.50

Primeramente se exploraron los datos para observar el comportamiento de las variables de interés y para analizar si el ejercicio del FAISM ha influido en el IRIS de los municipios indígenas del estado de Veracruz; se ajustó un modelo de regresión múltiple con una variable indicadora (Montgomery et al., 2004), el cual asume una relación lineal entre el IRIS, como variable respuesta y el FAISM como variable explicatoria, que permite estimar el valor de los parámetros y ver cómo se afecta el IRIS ante un cambio en el FAISM. También se incluye la variable cualitativa de región, con la finalidad de conocer si hay diferencia en la relación entre el IRIS y el FASIM dependiendo de la región a la que pertenezca el municipio. El modelo quedó especificado de la siguiente manera:

$$y_i = \beta_0 + \beta_1 FAISM_1 + \beta_2 POPOLUCA_2 + \beta_3 HUASTECA_3 + \beta_4 TOTONACA_4 + \varepsilon_i$$

$$i = 1,2,..50$$

$$\varepsilon_i \sim (N, \sigma^2)$$

donde y_i representa el IRIS para cada municipio *i*, β_0 representa el intercepto o el valor del IRIS para la región Zongolica cuando el FAISM es 0, β_1 es la pendiente y mide el cambio en el promedio del IRIS, cuando hay un cambio unitario en el FAISM. También se muestra en el modelo, las 3 variables dummy que se crearon, que indican si el municipio pertenece a determinada región, utilizando a la región Zongolica como la categoría de referencia. Así se tiene que en el modelo, se hizo la diferenciación a través de la siguiente categorización:

ZONGOLICA categoría de referencia.

HUASTECA = 1, si el municipio está en esta región, = 0 si pertenece a la Zongolica, Popoluca o Totonaca.

POPOLUCA = 1, si el municipio está en esta región, = 0 si pertenece a la Zongolica, Huasteca o Totonaca.

TOTONACA = 1, si el municipio está en esta región, = 0 si pertenece a la Zongolica, Huasteca o Popoluca.

Finalmente, ε_i representan al error el cual se postula que se distribuye como una normal con media cero y varianza constante; se supone también que los e_i son homocedasticos y no están correlacionados unos con otros. Para validar estos supuestos se realizó un análisis estándar de los residuos (Montgomery et al., 2004).

Resultados y discusión

Al explorar los datos se observó que la región que presentó mejores resultados en cuanto al IRIS en el año 2005 fue la Totonaca (Véase Figura 1). Es decir, en promedio disminuyó más su índice respecto al que había registrado en el año 2000 (-0.7670). Por el contrario, la región que mejoró muy poco fue la de Zongolica, pues solo disminuyó su índice en -0.22 unidades. Cabe mencionar que, en promedio, todas las regiones redujeron en el 2005 el

IRIS que presentaban en el año 2000. También se observa que, la región con menor variabilidad entre sus municipios respecto al IRIS, es la Popoluca.

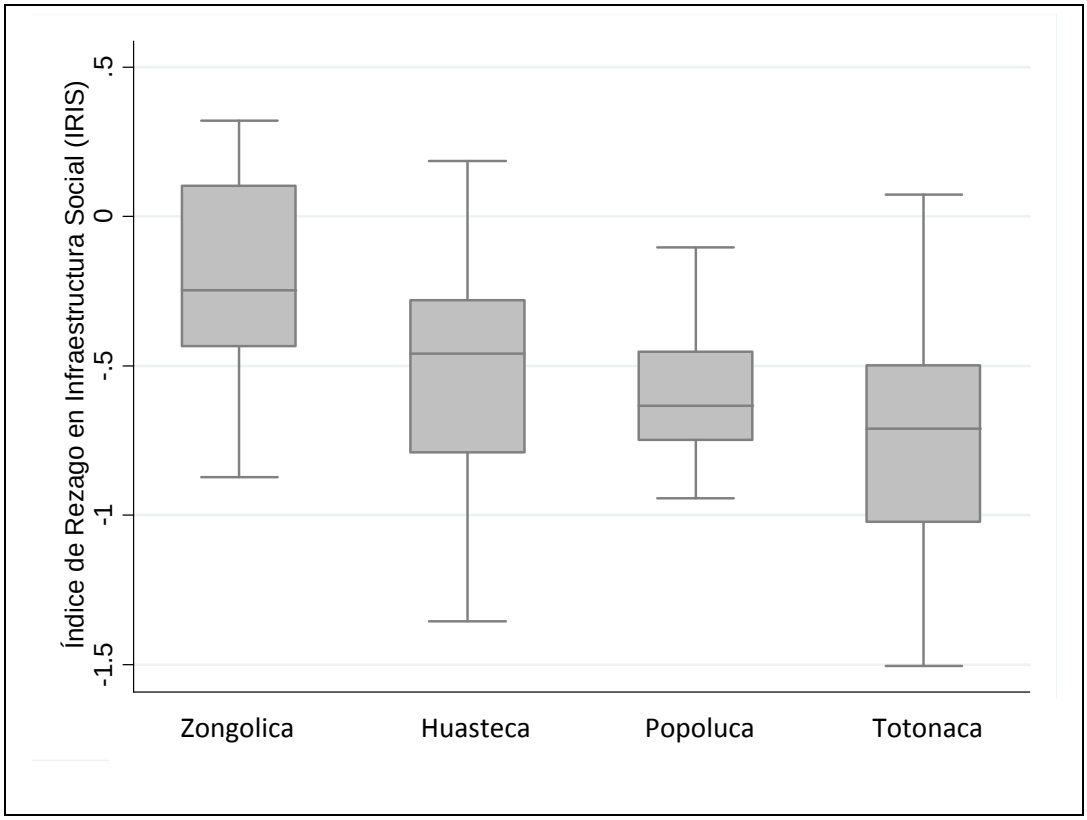


Figura 1. Estudio comparativo de rezago en infraestructura social por región

Haciendo el análisis a nivel municipal, se encontró que sólo cinco municipios (Atlahuilco, Ixhuatlancillo, Benito Juárez, Tlaquilpa y Soledad Atzompa) presentaron un comportamiento distinto al que se esperaba, lo que significa que en vez de disminuir su IRIS en el año 2005, éste aumentó. Los municipios que mejores resultados reportaron fueron Coahuitlán, Chumatlán y Mecatlán, de la región Totonaca, quienes redujeron su IRIS en -1.5, -1.4 y -1.05, respectivamente, y Texcatepec, de la Huasteca, registrando una diferencia de -1.35. Por otro lado, el monto promedio que recibieron del FAISM los municipios indígenas de Veracruz, fue de 72.50 millones de pesos, los cuales se distribuyeron en las cuatro regiones de la entidad de la siguiente manera: 33% fue para la región Huasteca, seguida con 30% para la Totonaca, 21% para la Popoluca y finalmente, la región que menos recurso económico recibió en este periodo fue la Zongolica, con 17%, siendo que ésta es la que más municipios concentra (15). El municipio que recibió mayor cantidad de recurso del FAISM fue Tantoyuca, de la región Huasteca y el que menos

recibió fue San Andrés Tenejapan de la región Zongolica. A pesar del monto recibido, Tantoyoca sólo disminuyó su IRIS en -0.07 y San Andrés Tenejapan en -0.29 unidades más.

Al visualizar la relación del IRIS respecto al FAISM (Véase Figura 2), se observa que en la región Popoluca dicha relación es lineal; es decir, a medida que aumenta el FAISM en los municipios, se incrementa la diferencia entre el IRIS 2005 y 2000, mientras que en las otras regiones no se visualiza una relación tan directa, particularmente en la región Zongolica, donde incluso se percibe una relación inversa.

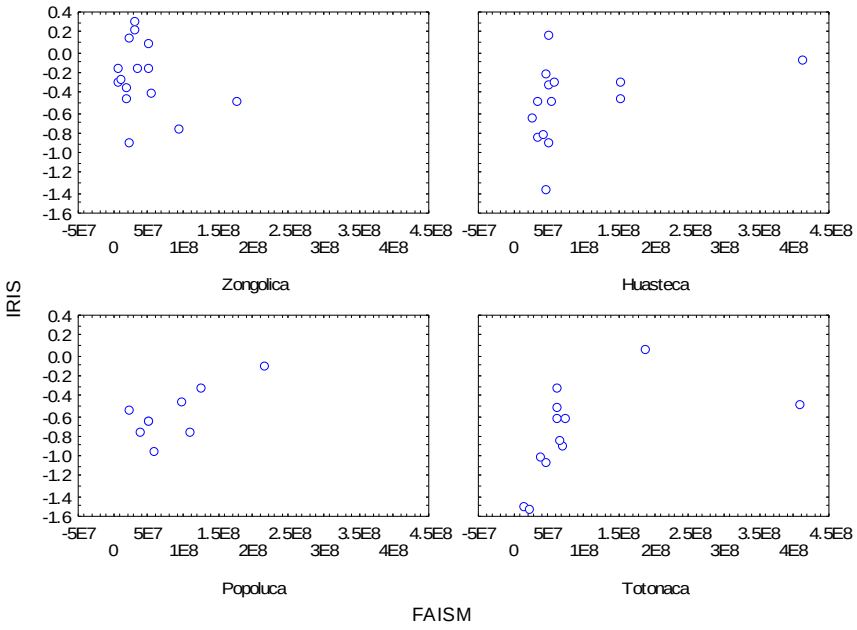


Figura 2. Relación entre el IRIS y el FAISM por región.

Los resultados de las estimaciones del modelo propuesto para estudiar la relación lineal entre e IRIS de los municipios indígenas del estado de Veracruz y el FAISM considerando la influencia de la región a la que pertenece el municipio, con una variable dummy, usando el método de Mínimos Cuadrados Ordinarios, MCO (Montgomery *et al*., 2004) se muestran en la Tabla 2.

Tabla 2. Resultados de las estimaciones de mínimos cuadrados ordinarios para el análisis de regresión, región Zongolica categoría de referencia.

	Coeficiente	**Error estándar**	**p-value**
Intercepto $\boldsymbol{\beta_0}$	-.28884	.094	0.004
FAISM $\boldsymbol{\beta_1}$	.0015699	.00062	0.016
Regiones			
Huasteca	-.3361	.1345	0.000
Popoluca	-.4110	.1517	0.010
Totonaca	-.6208	.1408	0.010
R^2	0.3325	F(4,45)=5.60	
R^2 ajustada	0.2731	Prob >F= 0.0010	

En la tabla anterior, se observa que la variable FAISM resulta significativa a un nivel de confianza del 5%. Esto quiere decir que el FAISM, sí influye en el IRIS; es decir, por cada millón de pesos que se aumente el FAISM, la diferencia entre el IRIS 2005 y 2000 de los municipios aumentará en .00156 unidades. Esta conclusión se aplica para todas las regiones.

Para el caso particular de cada región, que se crearon 3 variables dummy, se observa que las tres resultaron significativas, por lo que se interpreta que el promedio del IRIS para Zongolica (la categoría de referencia) es -.2888, mientras que para la región Huasteca es de -.6249 (-.3361+(-.2288)); para la Popoluca -0.6998 (-.4110+(-.2888) y finalmente para la Totonaca -.9096 (-.6208+(-.2884)). En este caso se obtiene un intercepto negativo porque indica que, en promedio, en cada región, la diferencia entre el IRIS del 2005 y 2000 ha sido negativa, es decir, el valor del IRIS ha disminuido, respecto al que se tenía en el año 2000. Asimismo, se observa en la Tabla 2, que el 33% de la varianza en el IRIS es explicada por el FAISM que se asigna a cada municipio indígena, así como por la región a la que pertenece. En algunas regiones la diferencia promedio es más grande, como en la Totonaca, y en otras es más pequeña, como en la región Zongolica. También se muestra, con el valor del estadístico F con 4 grados de libertad, que se rechaza la hipótesis de que alguno de los parámetros sea igual a 0, por lo que se concluye que las variables incluidas en el modelo contribuyen a explicar el comportamiento del ÍRIS en Infraestructura Social de los municipios indígenas del Estado de Veracruz.

Finalmente, dado que para validar las estimaciones del modelo, se corroboraron los supuestos de normalidad en los residuos, así como la homocedasticidad de la varianza, mencionamos que no se encontraron patologías significativas como se aprecia en el gráfico de la Figura 3 donde se percibe que no se viola el supuesto de normalidad.

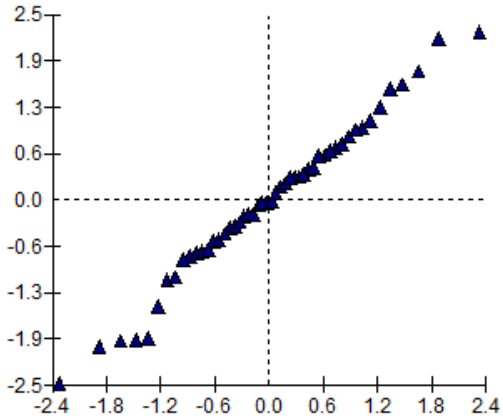


Figura 3. Gráfico de normalidad de los residuos del modelo ajustado.

Conclusiones

Los recursos económicos descentralizados que se otorgan a los municipios vía FAISM, tienen dentro de uno de sus objetivos ser invertidos en infraestructura social, como un medio para ir abriendo un proceso de lucha contra el rezago social en el que se encuentran los municipios indígenas de Veracruz. Sin embargo, resulta apremiante evaluar si dicho mecanismo está efectivamente cumpliendo con sus objetivos planteados. Para ello, en este trabajo se ajustó un modelo de regresión múltiple en el que se analizaron las variables FAISM y región, como explicativas de la variación de la diferencia del IRIS entre el año 2005 y 2000 de los municipios. Los resultados mostraron que estas variables fueron significativas; es decir, que a medida que se incrementan los fondos del FAISM destinados a cada municipio, la diferencia del IRIS aumenta. También se concluyó con las estimaciones que hay una diferenciación entre las regiones en las que se encuentran agrupados los municipios. Se estimó que la región Totonaca tiene en promedio una diferencia más alta del IRIS que la región Zongolica. Para futuras investigaciones, resultaría interesante estudiar analizar que características tiene cada región que explican esta diferencia.

Ante los resultados obtenidos, se muestra que el FAISM está cumpliendo sus objetivos para combatir los altos índices de rezago en infraestructura social que presentan los municipios indígenas de Veracruz, pues al aumentar el recurso la diferencia entre el IRIS 2005 y 2000 aumenta; es decir, que hubo mejoría.

Los resultados generan cuestionamientos al ejercicio del gasto del FAISM, porque la mejoría es muy poca dada la inversión que se realiza a través del fondo, ya que es de los gastos considerados por la política de descentralización financiera como determinantes para generar cambios a la infraestructura social de municipios y regiones, y si su influencia es pequeña, ello puede explicar la paulatina mejoría en el indicador del IRIS en el transcurso de 5 años en los municipios y regiones indígenas de Veracruz. Esto significa que el esfuerzo financiero para otorgar a municipios mayores recursos por medio del FAISM durante los últimos años está siendo poco eficaz para atender la problemática de los municipios indígenas con un alto índice de rezago en infraestructura social, lo cual está haciendo que dichos municipios cuente con una deficiente infraestructura social, que hará más complejo y largo el proceso de combate al rezago social de los municipios indígenas de Veracruz.

En consecuencia, es viable y pertinente que se construya una propuesta de coparticipación gobierno-ciudadanía para el ejercicio del FAISM en los marcos teóricos y polítologicos descritos, al considerar que el COPLADEMUN es un mecanismo que simula una coparticipación gobierno-ciudadanía, lo cual inhibe la posibilidad para que el ejercicio de los fondos mencionados tenga una influencia eficaz en términos de fortalecimiento de la infraestructura social de los municipios indígenas de Veracruz.

Ante la evidencia de los resultados en este análisis, se considera pertinente empezar a construir cambios al FAISM, y observar que está sucediendo particularmente en las regiones y por qué unas se encuentran más rezagadas que otras. Es evidente que la descentralización propia del enfoque financiero moderno, abre la posibilidad para que el ejercicio del FAISM se realice bajo un mecanismo de organización, a través del cual se desarrolla un proceso de coparticipación gobierno-ciudadanía para la asignación, vigilancia y supervisión del FAISM, de tal manera que se posibilite la eficacia en la transparencia y

control en las finanzas públicas de los municipios indígenas de Veracruz, sin embargo, es necesario vigilar los resultados obtenidos por dicho mecanismo.

Referencias

Cohen, J. (1997). "Procedure and Substance in Deliberative Democracy", Deliberative Democracy, Cambridge: MIT Press.

Díaz, A. y Silva, S. (2004). Descentralización a escala municipal en México: la inversión en infraestructura social. CEPAL - SERIE Estudios y perspectivas 15.

Duverger, M. (1981) "Sociología Política", Editorial Ariel, segunda reimpresión.

Gaudemet, P.M. (1996) Finances publiques, Tome 1, Paris, Edit. Montchrestien.

Habermas, J. (1995) Between Facts and Norms, Cambridge: mit Press.

Mathus, M.A.(2008). "Fondo de aportaciones para la infraestructura social municipal en los municipios de Oaxaca 1998-2002" en Observatorio de la Economía Latinoamericana, Nº 93.

Montgomery, D., Peck, E. y Geoffrey Yning (2004). Introducción al análisis de regresión lineal. CECSA. Primera reimpresión.México.

Ojeda, M. M., Díaz C., E., Apodaca, C. y Trujillo, I. (2004). Metodología de Diseño Estadístico. Universidad Veracruzana, Xalapa, Ver., México.

Lechner, N. (1988), Los patios interiores de la democracia, México: Fondo de Cultura.

Un análisis del impacto del Programa de Apoyos Directos al Campo (PROCAMPO) en la productividad del campo veracruzano, periodo 2002 – 2008

Mario Miguel Ojeda Ramírez
Arely González Hernández

Resumen

Las crisis económicas que ha sufrido México en los últimos años han provocado que el campo mexicano se encuentre ante una serie de influencias de procesos que han impactado en su desarrollo, produciendo múltiples contradicciones. Tal situación se ha presentado de manera diferenciada en los estados mexicanos y más aún a nivel municipal. Los programas públicos que se han diseñado e implementado para atender esta problemática han producido efectos diversos, tanto en las prácticas como en los resultados. Por todo esto, existe una necesidad de realizar estudios tanto a nivel local como regional para evaluar los impactos diferenciados de estos programas. En este trabajo se realiza el análisis de la superficie sembrada, la superficie cosechada y el monto del PROCAMPO, considerando como unidades de estudio a los 212 municipios del estado de Veracruz durante el periodo 2002 al 2008; se analiza la relación entre estas variables, identificando los patrones y tendencias a nivel municipal, usando la técnica de análisis cluster; también se lleva a cabo un estudio de modelación jerárquica para evaluar la relación entre estas variables analizando las series a nivel municipal. Se elaboraron grupos atípicos a un comportamiento de la mayoría de los municipios y se detecta que la superficie sembrada y el monto del PROCAMPO influyen en el comportamiento de la superficie cosechada.

***Palabras clave**:* Análisis estadístico, Análisis cluster, Evaluación de impactos de programas públicos, Modelos lineales jerárquicos, Modelación multinivel, Productividad agrícola.

Abstract

The economic crises that Mexico has suffered in the last years have caused that the Mexican countryside is facing a series of processes influences that have impacted their development producing multiple contradictions. Such situation has appeared of way differentiated the Mexican states and even more to municipal level. The Public programs that have been designed and implemented to attend this problematic situation have produced diverse effects that have occurred in the practice and in the results. For this, there is a need to realize studies both locally and regionally to evaluate the differential impacts of these programs. In this paper, it is realized an analysis of the planted surface, the harvested surface and the amount of PROCAMPO, considering as units of study to 212 municipalities in the state of Veracruz during the period 2002 2008, the relation between these variables is analyzed, identifying patterns and trends to the municipal level, using the technique of cluster analysis. Also there is carried out a study of hierarchical modeling to evaluate the relation between these variables analyzing the series to municipal level. Atypical groups

were elaborated to a behavior of the majority of the municipalities and there is detected that the planted surface and the amount of the PROCAMPO influence the behavior of the harvested surface.

Keywords: Statistical analysis, Cluster analysis, Evaluation of impact of public programs. Hierarchical linear models, Multilevel modeling, Agricultural productivity.

Introducción

México posee una gran riqueza y diversidad de recursos naturales inigualables, mismos que han sido perturbados debido a diversos factores, tanto naturales como por la intervención humana, afectando de manera directa al campo mexicano. Recientemente el Estado ha aplicado una serie de políticas en materia agropecuaria, debidas a la aplicación del Tratado de Libre Comercio de América del Norte (TLCAN), generando una transformación en la producción agropecuaria, en la liberalización comercial, y la eliminación de barreras arancelarias y no arancelarias. La justificación es la finalidad de vender los excedentes de producción debido a la baja capacidad de consumo interno. En la última década se ha sostenido un debate sobre el problema agrario, considerándose una transformación en la gestión de recursos a nivel federal, estatal y municipal, lo que ha conllevado a tener efectos positivos, principalmente en la actividad económica del campo y en las organizaciones o empresas integradoras de productores.

El sector agropecuario abarca tres amplios ramos: (1) la agricultura, donde se maneja el cultivo de la tierra para sembrar alimentos; (2) la ganadería, la cual se refiere a la crianza de los animales con fines de producción alimenticia; y (3), el control forestal, donde se considera la producción forestal maderable y la no maderable. Este sector en el estado de Veracruz es uno de los más importantes a nivel nacional, por el gran volumen con el que participa, por los empleos que genera y por la superficie que ocupa en su explotación. Además, se caracteriza por su ubicación, debido a que cuenta con canales de comercialización idóneos para transportar los principales cultivos y la producción en general.

El campo mexicano se ha adaptado a las políticas públicas que el Estado ha diseñado con el objeto de distribuir la riqueza de forma equitativa, lo que se ejerce a través

de la administración de recurso en programas específicos y sectoriales que atacan la problemática social. Actualmente, el sector agropecuario comparado con otros sectores económicos, tiene un menor ritmo de crecimiento pero una mayor volatilidad que ejerce un riesgo en la producción, cabe mencionar que es imprescindible el analizar la tendencia de los subsectores para la producción agrícola, ganadera, silvícola y forestal; siendo la producción del subsector agrícola, por una parte, la que aporta el mayor conjunto de actividades más importantes del sector; mientras que por otra parte, la producción del subsector forestal carece de incentivos y apoyos gubernamentales que recaen en bajos niveles de productividad. A pesar de que no existe gran discusión con respeto a las escasas inversiones que se destinan a la producción forestal, se debiese considerar como tema prioritario, tanto para el cuidado del medio ambiente como a la actualización del marco legal.

Las finanzas públicas sirven para impulsar cambios sociales y sólo a través de éstas se puede realizar un enfoque al desarrollo sustentable, que incluye la productividad, atención al medio ambiente pero también al desarrollo social. En este sentido el equilibrio social se puede lograr a través de la aplicación de las finanzas públicas redistributivas, por ello, el Presupuesto de Egresos de la Federación (PEF) contempla programas de política económica agropecuaria, para apoyar a los agricultores y garantizar un ingreso mínimo de acuerdo a su unidad o superficie de producción.

Uno de los programas es el de Ingreso Objetivo (IO), el cual tiene como propósito el garantizar a los agricultores un ingreso mínimo por unidad de producción. Este programa compensa el ingreso del productor cuando los precios del mercado domestico se encuentran por debajo de un nivel de precio fijado por el gobierno, conocido como "ingreso objetivo mínimo" y se designa para cada cultivo elegible. Este programa está diseñado para beneficiar a los productores de granos y oleaginosas, los pagos se realizan por tonelada. Otro programa es el Programa de Estímulos a la Productividad Ganadera (PROGRAN), su propósito es el promover prácticas de ganadería extensiva destinadas a crear incentivos para aumentar la producción de forraje en los pastizales y praderas del país. Se ejecuta a través de un sistema de identificación animal el cual es el Sistema Nacional de Identificación Individual del Ganado (SINIIGA), estos pagos están condicionados al registro de animales

a un padrón ganadero (SAGARPA, 2010). A principios de 1990, se anunció el Programa de Modernización del Campo, el cual estableció los lineamientos de la política agropecuaria hasta 1994. Estos lineamientos excluían del apoyo productivo a los campesinos pobres y delegan a la población rural los programas de tipo asistencial que intentan aminorar los efectos de la política de la población rural marginada.

A finales de 1993 surge el PROCAMPO, como un mecanismo de transferencia de recursos para compensar los productores nacionales por los subsidios que reciben sus competidores extranjeros, en sustitución del esquema de precios de garantía de granos y oleaginosas (SAGARPA, 2010); otorga un apoyo por hectárea o fracción de ésta a la superficie elegible, la cual debe estar inscrita en el directorio del PROCAMPO y que esté sembrada con cualquier cultivo lícito o que se encuentre bajo proyecto ecológico autorizado por la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). Dicho apoyo es entregado a los productores que acrediten ser propietarios o poseedores de buena fe o en posesión derivada (arrendamiento, usufructo, aparcería) de predios con superficies elegibles en explotación inscritos en el directorio del programa, donde se encuentran los productores del país que voluntariamente se inscribieron, independientemente del tamaño del predio, tipo de tenencia de la tierra, régimen hídrico, modo de producción o filiación política (SAGARPA, 2010).

El objetivo del PROCAMPO es transferir recursos en apoyo de la economía de los productores rurales, que siembren la superficie elegible registrada en el directorio del programa, cumplan con los requisitos que establezca la normatividad y acudan a solicitar por escrito el apoyo (ASERCA, 2010). Dada la importancia de este programa se han realizado diversos estudios los cuales analizan la situación de la implementación tanto de metodologías como de políticas públicas orientadas al desarrollo rural sustentable.

Uno de los trabajos que mayor impacto ha causado debido a su propuesta metodológica ha sido el de Ovando y Córdoba (2004), que propone clasificar la actividad agropecuaria y los productores agropecuarios del estado de Veracruz, con la finalidad de identificar los diferentes tipos de la actividad agropecuaria y estratos de las unidades productivas rurales del estado, para lo que propone la instrumentación del trato diferenciado tanto a las regiones como a los productores. Se utiliza la técnica de análisis

factorial donde analiza una serie de 38 variables de las que toma grupos de variables del criterio de especialización, de tecnología, de articulación al mercado, de marco institucional, de característica socioeconómicas (Ovando y Córdoba, 2004).

En el 2006, la OCDE publicó un informe temático *"*El Nuevo Paradigma Rural: Política y Gobernanza*"*, donde busca explicar cuál es el giro actual de las políticas de desarrollo rural, pero lo interesante es que lo resalta mediante un enfoque multidisciplinario, donde propone una coordinación sectorial que permita, a través de los órdenes de gobierno y los actores públicos y privados, enfrentar retos de cambios de paradigma. En el mismo año, la OCDE realizó el *"*Estudio de Política Rural*"* de México, el cual menciona que la política rural en México ha evolucionado de un sólo sector (agropecuario) hacia una política que involucra la integración de actores que a través de acciones conjuntas, genere un soporte productivo enfocado especialmente al sector agropecuario (OCDE, 2007). De la misma manera, demuestra que una buena política rural con agricultura intensiva, dará mejores resultados que una política rural pobre, enfatizando nuevamente el sector agrícola a nivel nacional.

A nivel nacional, la SAGARPA ha publicado en el 2009, el "Escenario Base 2009–2018; Proyecciones para el Sector Agropecuario de México", que genera un modelo de proyecciones macroeconómicas de largo plazo del sector agropecuario, basadas en el modelo escenario base 2009–2018. Abunda en el análisis del subsector agrícola y pecuario, mediante técnicas econométricas que permiten generar las posibles proyecciones a largo plazo, para evaluar y cuantificar los impactos de la política pública y cambios coyunturales en el sector. Consideran como escenario base un modelo econométrico que mide los siguientes aspectos: entorno y perspectiva macroeconómica, sector agropecuario internacional y política pública sectorial. Finalmente abarca submodelos agrícolas, pecuarios y agroindustriales (SAGARPA, 2009). Los resultados que se obtuvieron para el subsector agrícola fueron que para el 2009, el escenario base estima un incremento del 2% en la producción de granos y para el largo plazo se espera que esta tendencia positiva se mantenga. Además, se estima un crecimiento sostenido en los granos para uso humano y forrajero. El maíz predomina en la producción como en la demanda de la agricultura mexicana. Un dato importante es que se anticipa que la superficie sembrada de trigo, arroz

y algodón en México crecerá más lentamente que la del maíz, mientras que para la caña de azúcar se prevé un declive debido al precio internacional del azúcar. Asimismo, este estudio sugiere implementar estrategias que aumenten la producción ganadera para erradicar las enfermedades, conservar el suelo y tener un uso eficiente del agua en la producción de cultivos y forrajes.

Debido a los informes publicados por la SAGARPA, el análisis realizado aquí es precisamente porque se desconoce cuál ha sido el comportamiento de la superficie cosechada y sembrada en el estado de Veracruz, durante el periodo 2002-2008 y si es que existen diferencias entre los municipios. Se busca identificar aquellos que tienen comportamientos claramente diferentes, y así mismo estudiar la relación entre las variables productivas, mediadas por el tiempo.

En este artículo se aborda la temática referente a la agricultura en el estado de Veracruz; es decir, abarca aspectos de superficie cosechada y sembrada, superficie sembrada de riego y de temporal, productores beneficiados por el PROCAMPO y el monto pagado por dicho programa. Las interrogantes que surgen inmediatamente son conocer, durante el periodo 2002 al 2008, el comportamiento de la agricultura en el estado de Veracruz, saber si existe relación entre la superficie sembrada y la superficie cosechada. Para responder a estas inquietudes, como objetivo general se plantea que el conocer, estudiar y analizar el comportamiento de la economía de la producción agrícola, considerando básicamente tres variables señaladas. Así también se busca conocer la relación existente entre estas variables, considerando las series en el periodo de estudio.

Metodología

Los datos analizados se obtuvieron del Sistema Estatal y Municipal de Bases de Datos (SIMBAD) del Instituto Nacional de Estadística y Geografía (INEGI). La base de datos contiene cifras del sector agropecuario de los 212 municipios del estado de Veracruz durante los años 2002 al 2008. Las variables que se contemplan en la base de datos son: la superficie sembrada y la superficie cosechada medidas en hectáreas y el monto pagado por el PROCAMPO, considerado en miles de pesos. En primera instancia, se realizaron gráficas de dispersión para estudiar la relación existente entre la superficie sembrada y el

monto del PROCAMPO con la superficie cosechada. Así mismo, se construyeron series temporales de estos datos a nivel municipal para identificar las tendencias y comportamientos de las variables en estudio a nivel municipal. También se realizaron diferentes análisis de agrupación o análisis cluster (Legendre y Legendre, 1998) para identificar el grado de homogeneidad en los municipios respecto a los comportamientos de las series y de las asociaciones de las variables en estudio. Como análisis definitivo se realizó un análisis de modelación lineal jerárquica con el objetivo de ver la relación existente entre las variables de estudio considerando las series a nivel municipal. El software utilizado para el análisis estadístico fue el SPSS (SPSS, 2007) y el MLwiN versión 2.19. En la Tabla 1 se describen cada una de las variables estudiadas y el valor de la media nacional.

Tabla 1. Descripción de las variables de estudio y valor de la media nacional.

Variable	Descripción	Media nacional
SUPSEM	Superficie sembrada en cada municipio por año durante el periodo 2002-2008.	6930.89 hectáreas
MONPRO	Monto de los recursos asignados al programa PROCAMPO en cada municipio por año durante el periodo 2002-2008.	2875.05 miles de pesos
SUPCOS	Superficie cosechada en cada municipio por año durante el periodo 2002-2008.	8206.69 hectáreas

Dado que la información que se obtuvo presenta una estructura de anidamiento, y se desea estudiar la posible relación entre la superficie cosechada en cada municipio por años con respecto a las variables superficie sembrada y PROCAMPO se propone un modelo de dos niveles donde la variación de la cosecha a través del tiempo ocurre en el nivel-1 y la variación de la cosecha entre los municipios ocurre en el nivel-2 (Goldstein, 1999; Raudenbush y Bryk, 2002), Como unidades de nivel 1 se tomaron los 7 años que comprende este estudio y en un segundo nivel está formado por 209 municipios (Figura 1). La razón por la cual se tomaron únicamente 209 municipios de los 212 es que en los tres municipios no considerados no se tiene información completa sobre las variables en estudio.

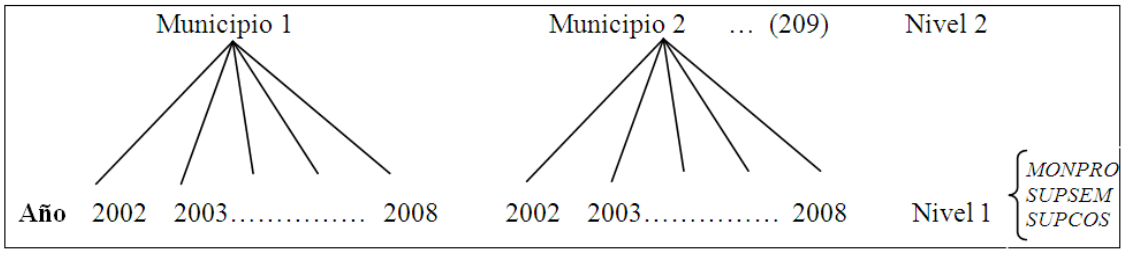


Figura 1. Diagrama de unidades y anidamiento para la estructura jerárquica de los datos en estudio.

A través de la modelación jerárquica, se pretende tener una mejor comprensión de la variabilidad de los datos, pues permite conocer la varianza entre los años y los municipios respecto a las hectáreas sembradas, teniendo en cuenta las variables explicativas consideradas en el estudio. Es decir, la variabilidad en los datos en cada nivel y entre niveles se analiza simultáneamente.

Se aplicó un modelo multinivel con intercepto aleatorio con variables explicatorias, entre el tiempo, la superficie sembrada, el monto de *PROCAMPO*, y la superficie cosechada para las 209 municipios, durante el periodo 2002-2008; quedando expresado el modelo en la siguiente ecuación:

$$
\begin{aligned}
&SUPCOS_{ij} = \beta_{0j} + \beta_1 TIEMPO + \beta_2 SUPSEM_{ij} + \beta_3 MONPRO_{ij} e_{ij} \\
&\beta_{0j} = \beta_0 + u_{0j} && i = 1,...,7 \\
&e_{ij} \sim N\left(0, \sigma_e^2\right) && j = 1,...,209 \\
&u_{0j} \sim N\left(0, \sigma_{u0}^2\right)
\end{aligned}
$$

donde β_0 denota el intercepto o la media global de la superficie cosechada en un municipio por año; β_1, β_2 y β_3 constituyen la pendiente o el cambio en la media de la superficie cosechada (*SUPCOS*) cuando hay un cambio unitario en cada variable explicatoria tiempo (*TIEMPO*), superficie sembrada (*SUPSEM*), y monto de PROCAMPO (*MONPRO*) respectivamente, manteniendo las otras variables constantes, e_{ij} denota el error aleatorio correspondiente a la i-ésima unidad de nivel 1 en la j-ésima unidad de nivel 2 y u_{0j} denota el error aleatorio a nivel 2. Con este modelo, lo que interesa es conocer si alguna variable como el tiempo (*TIEMPO*), la superficie sembrada (*SUPSEM*) o el monto de PROCAMPO (*MONPRO*) influyen en el comportamiento de la variable superficie

cosechada (*SUPCOS*). Para validar los resultados del modelo, se comprobó el cumplimiento de los supuestos de normalidad de los errores en los dos niveles,

Resultados y discusión

Al explorar los datos a nivel estatal, en la Figura 2 se observa que la superficie cosechada en el 2002 fue cerca de 1.5 millones de hectáreas, para el 2003 y 2004 se mantuvo en 1.3; sin embargo, durante el 2005 y 2007 bajó la superficie cosechada a 1.275; y para 2008, aumentó a 1.4. Es probable que el paso del Huracán Stan y Katrina hayan sido factores determinantes que afectaron de manera directa la superficie cosechada.

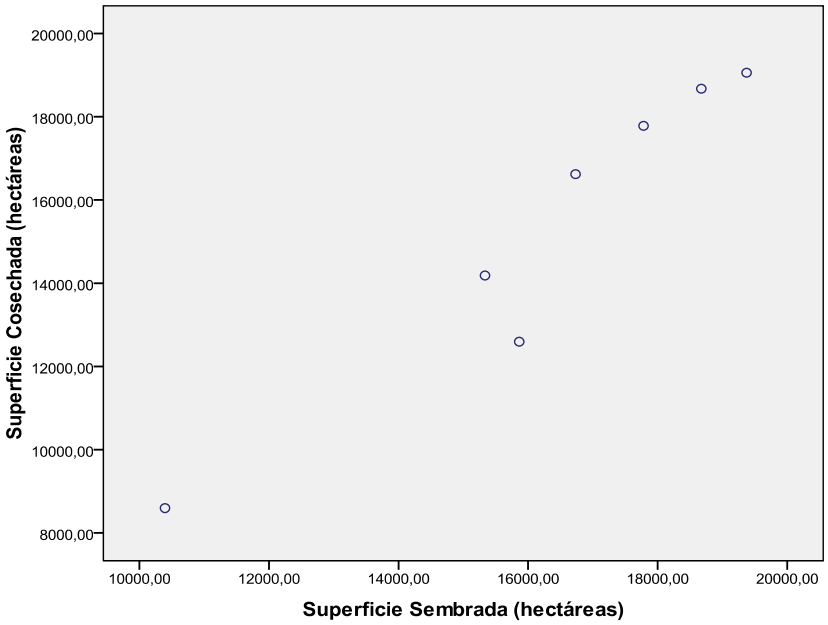


Figura 2. Superficie cosechada a nivel estatal del 2002 al 2008.

Al estudiar la relación entre la superficie sembrada y la superficie cosechada, se observa que existe una clara relación directa entre dichas variables. A pesar de que algunos valores sobresalen, es posible connotar que al menos por cada hectárea sembrada debería haber una hectárea cosechada; es decir, a medida que aumenta la superficie sembrada, aumenta la superficie cosechada, contemplado una relación directa positiva.

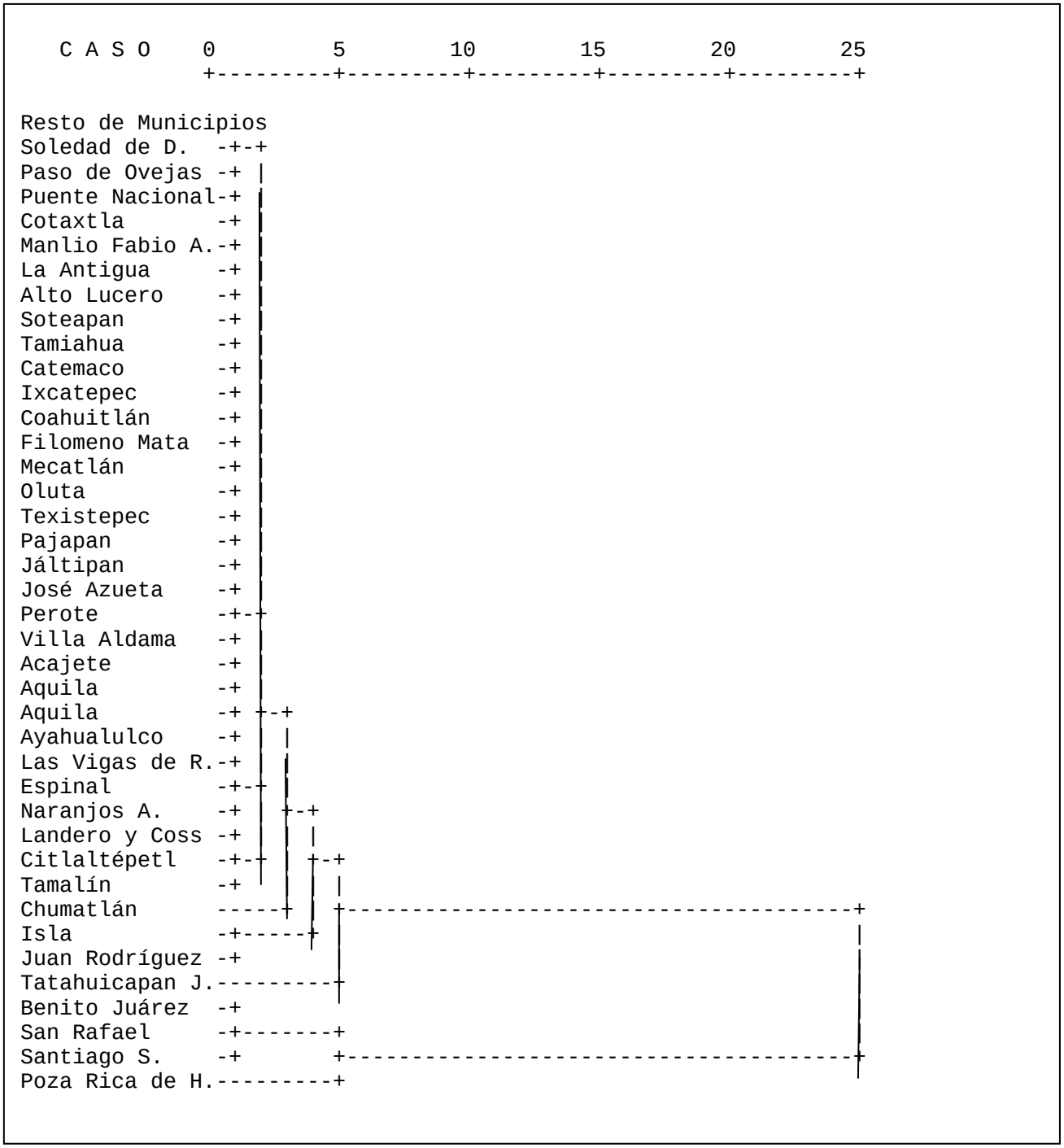


Figura 3. Dendrograma del porcentaje del monto de PROCAMPO.

En la Figura 3 se observa la formación de 2 grupos de municipios los cuales son semejantes respecto al porcentaje del monto asignado por PROCAMPO al municipio. Uno

de estos grupos se encuentran los municipios de Benito Juarez, San Rafael, Santiago Sochiapa y Poza Rica de Hidalgo. Mientras que en el otro los demás municipios.

En la figura 4 se observa la formación de 2 grupos de municipios los cuales son semejantes respecto al porcentaje de productividad. Uno de estos grupos está constituido por los municipios Poza Rica de Hidalgo, San Juan Rodríguez Clara, San Rafael y Santiago Sochiapa, mientras que el otro grupo está formado por todos los municipios restantes, es decir por 208 municipios.

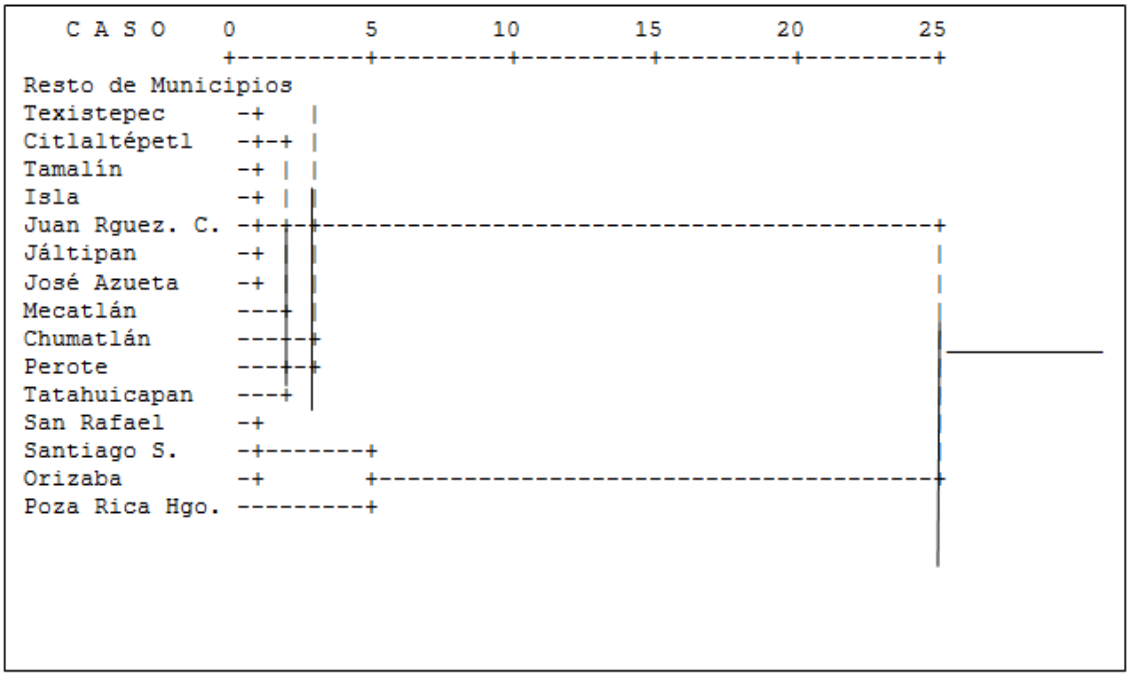


Figura 4. Dendrograma del porcentaje de productividad.

Para estudiar la posible relación entre la superficie cosechada con las variables explicatorias de tiempo, superficie sembrada y monto de PROCAMPO, se ajustaron 5 modelos multinivel, utilizando el método de Mínimos Cuadrados Generalizados Iterativos (Goldstein, 1999). Los resultados de las estimaciones se muestran en la Tabla 2. En el primer modelo ajustado, el de intercepto aleatorio, los resultados del ajuste muestran que se tiene una superficie cosechada en promedio de 6463.39 hectáreas en cada municipio por año; además de que existe variación tanto entre los años como entre los municipios. El porcentaje de la variabilidad de la superficie cosechada atribuida a los municipios es de

aproximadamente el 94% y sólo un 6% a los años. En el segundo modelo estimado se introdujo la variable tiempo como variable explicatoria, se mantuvo fija la pendiente y el intercepto aleatorio, los resultados del ajuste muestran que esta variable sí resulta significativa, aunque en forma negativa, es decir, que en cada año la superficie cosechada en los municipios se tuvo un decrecimiento de 79.69 hectáreas. Respecto a la variación entre los años y entre los municipios se mantienen en los valores anteriores, resultando que las variables explicatorias son significativas.

Tabla 2. Estimaciones de los parámetros y componentes de la varianza con sus errores estándar de 5 modelos multinivel.

	Modelo intercepto aleatorio, sin variables explicatorias (1)	Modelo intercepto aleatorio con la variable explicatoria *TIEMPO* (2)	Modelo intercepto aleatorio con variables explicatorias *TIEMPO* y *SUPSEM* (3)	Modelo intercepto aleatorio con la variables explicatorias *SUPSEM* y *MONPRO* (4)	Modelo inter. aleatorio con *SUPSEM*, *MONPRO* y municipios problema (5)
Parámetros fijos					
β_0 (Intercepto)	6463.39 (551.77)	6782.15 (561.87)	6550.13 (66.334)	6494.45 (39.8)	6474.039 (26.150)
β_1 (*TIEMPO*)		-79.69 (26.52)	**-21.68 (11.929)**		
β_2 (*SUPSEM*)			0.946 (0.005)	0.979 (0.0053)	0.977 (0.004)
β_3 (*MONPRO*)				-0.1029 (0.0092)	-0.067 (0.007)
β_4 (*Isla*)					-3900.76 (379.46)
β_5 (*Rod Clara*)					-3327.67 (383.128)
β_6 (*Azueta*)					-3251.54 (385.46)
β_7 (*Chicon*)					-2009.66 (364.22)
Componente de la varianza					
Nivel 2 σ^2_{u0}	63037720 (6224552)	63041504 (62244473)	324886 (43484)	214295 (32897)	29601 (13088)
Nivel 1 σ^2_e	4145173 (165542)	4115537 (164358)	832198 (33274)	806342 (32393)	685502 (27398)
Deviance					
-2*logVerosimilitud	27421.7	27412.7	24370.5	24156.9	23834

En el tercer modelo, se introdujo adicionalmente la superficie sembrada y se modela como fija (igual que *TIEMPO*). En los resultados mostrados en la Tabla 2, se observa que la variable *SUPSEM* es significativa, esto quiere decir, que ante un cambio unitario de la

superficie sembrada en cada municipio, la superficie cosechada se incrementa en promedio en 0.946 hectáreas, manteniendo la variable *TIEMPO* fija, cabe destacar que al introducir la variable *SUPSEM* al modelo, el *TIEMPO* deja de ser significativo, por lo consiguiente no se debe tomar en cuenta dicha variable en el modelo. Al comparar los modelos 2 y 3, se aprecia que la varianza a nivel municipio disminuyó de 63041504 a 324886, y la varianza a nivel año también presenta una reducción de 4115537 a 832198. También hay una disminución en el valor de la deviance de 27412.7 a 24370, es decir una reducción de 3042.2, que al compararlo con una distribución χ^2 con 1 grado de libertad, resulta significativa. Lo que indica que el tercer modelo es más adecuado para el ajuste de los datos.

Se ajustó un cuarto modelo, en éste se introdujo adicionalmente el monto de PROCAMPO y se modela como fija (igual que *SUPSEM*). En los resultados mostrados en la tabla 2, se observa que la variable *MONPRO* es significativa, esto quiere decir, que ante un cambio unitario del monto de PROCAMPO en cada municipio, la superficie cosechada tiene en promedio un decrecimiento de 0.103 hectáreas, manteniendo la variable *SUPSEM* fija. Al comparar los modelos 3 y 4, se aprecia que la varianza a nivel municipio disminuyó de 324886 a 214295, y la varianza a nivel año también presenta una reducción de 832198 a 806342. Lo que indica que el cuarto modelo es más adecuado para el ajuste de los datos. Sin embargo, en un análisis de los residuos, se observó la presencia de municipios atípicos, ver la Figura 5.

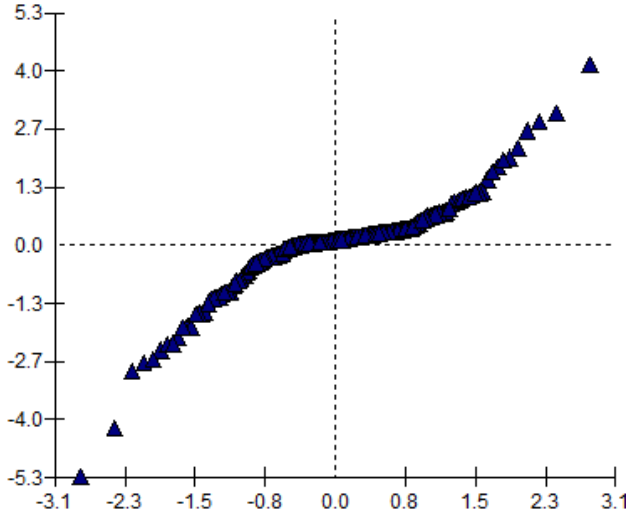


Figura 5. Gráfico de los residuos del modelo ajustado.

Para analizar si municipio atípico altera las estimaciones del modelo planteado, se incluyó este municipio como variable indicadora, lo cual llevó a otros municipios a aparecer como municipios atípicos, obteniéndose como resultado que se ajustará otro modelo que considerará estos municipios también como variables indicadoras. Comparando este quinto modelo con los modelos anteriores, se observa que éste es el que mejor se ajusta a los datos, pues se reduce la varianza entre los municipios considerablemente de 214295 a 29601, y la varianza entre los años disminuye de 806342 a 685502. Así un municipio en el que se siembra una superficie en promedio de 6930 hectáreas y el monto de PROCAMPO es de 2875 miles de pesos en promedio, tiene aproximadamente 6474 hectáreas cosechadas. Además tanto la superficie sembrada y el monto destinado por PROCAMPO presentan influencia en el comportamiento de la superficie cosechada, siendo que por hectárea sembrada del periodo 2002 a 2008 la superficie cosechada fue de aproximadamente de 0.977 hectáreas, y por cada mil pesos destinado por PROCAMPO la superficie cosechada disminuyó aproximadamente en 0.06 hectáreas.

Conclusiones

Del análisis cluster realizado, se encontró que hay 2 grupos de municipios de acuerdo a los porcentajes de productividad de la superficie sembrada y cosechada, y al porcentaje del monto del PROCAMPO, el primero de estos grupos formado por 4 municipios, mientras que el otro grupo por los municipios restantes. Del análisis de modelación jerárquica se concluye que existe relación directa entre la superficie cosechada y la superficie sembrada, donde el comportamiento resulta muy homogéneo en los municipios del estado de Veracruz del 2002 al 2008. Se destaca el hecho de que los montos del PROCAMPO influyan de manera negativa en la superficie cosechada en los municipios del estado de Veracruz, como se observa en los resultados. Sin embargo, pueden existir variables de otra índole que contravengan los resultados estadísticos, como pueden ser, por ejemplo el cambio climático, el que genere que no necesariamente se coseche la totalidad de una hectárea.

Finalmente, cabe mencionar que de acuerdo a los fundamentos teóricos de la sustentabilidad y la doctrina francesa, es necesaria la aplicación de políticas encaminadas a

incrementar el desarrollo rural sustentable a través de un enfoque multisectorial, que implique la creación de políticas integrales para la reactivación del campo mexicano.

Referencias

Legendre P. and Legendre L. (1998). *Numerical Ecology*. Second Edition. Elsevier, Amsterdam.

Gudynas, E. (2004). *Ecología, Economía y Ética del Desarrollo Sostenible.* CLAES.

Ovando, E. y Córdoba, L. (2004). *Política Agropecuaria Territorialmente Diferenciada: Propuesta Metodológica.* Estudios Agrarios.

Organización para la Cooperación y el Desarrollo Económico (2006). *The New Rural Paradigm: Policies and Governance*.

Organización para la Cooperación y el Desarrollo Económico (2007). *Estudios de política rural.*

SAGARPA, Agriculture and Food Policy Center, Food and Agricultural Policy Research Institute. (2009) *Escenario Base del Sector Agropecuario en México, Proyecciones 2009 – 2018*.

Sitios WEB:

http://www.spss.com

http://www.hks.harvard.edu/fs/pnorris/Classes/A%20SPSS%20Manuals/SPSS%20Statistics%20Brief%20Guide%2017.0.pdf

http://www.bristol.ac.uk/cmm/software/mlwin/download/mlwin-userman-09.pdf

Becas Pronabes: Una Mirada a su Evolución e Impacto en el Fortalecimiento del Desarrollo Humano 2002-2007

Ada Alicia Galván Herrera
Patricia Tapia Blásquez

Resumen

PRONABES es un instrumento de política pública que tiene por objetivo fomentar que los jóvenes en condiciones económicas adversas accedan a los servicios públicos de Educación Superior. El objetivo del presente trabajo es analizar si existe relación entre el porcentaje de becas que se asignan en las entidades federativas y el monto asignado a cada estado para este programa social, y con el Índice de Cobertura (COB) de cada entidad. Además, se estudiará si ha habido una variabilidad significativa entre las entidades en esta asignación a lo largo del periodo 2002-2007. Para ello se aplicó un modelo de dos niveles con intercepto y pendiente aleatoria, encontrándose que el monto asignado en cada entidad federativa influye positivamente en la proporción de becas asignadas del PRONABES, y que mientras aumente en promedio el índice de cobertura en educación superior de cada entidad, la proporción de becas disminuye. Se observó también que ha habido una variación significativa a lo largo del tiempo, entre los años del periodo de estudio, así como una diferencia entre las entidades federativas.

Palabras Clave: Becas PRONABES, Financiamiento a la Educación Superior, Desarrollo Humano, Gasto Nacional en Educación, Modelación Multinivel.

Abstract

PRONABES is a public policy tool that aims to encourage young people in difficult economic conditions to access public higher education. The objective of this work is to analyze if there is a relationship between PRONABES scholarships, the enrollment number, the coverage of higher education, the financial resources assigned and GDP, and also to find out if there is significant variability among the states of the country and the year of the study. Therefore, we apply a hierarchical lineal model, finding that GDP influences negatively the proportion of scholarships among estates. We also observed a significant variability among the years and the states in Mexico.

Introducción

El Estado presenta un gran desafío para mantener una política de financiamiento sostenido y creciente para la Educación Superior (ES), aunado a los evidentes problemas de equidad, y de proporcionar las mismas oportunidades de preparación para todos aquellos que lo deseen sin importar su condición socioeconómica. "Al hablar de igualdad de oportunidades debemos referirnos necesariamente, a las probabilidades de que los establecimientos educativos definan los términos de la competencia escolar en función de los dotes académicos de los individuos, con independencia de su origen social o cultural" (Rodríguez, 1996, p.50).

Por otro lado, el financiamiento a la Educación Superior Pública tiene que competir frente a las prioridades educativas de México, pues a pesar de que cada año se ha ido incrementado su presupuesto, aún representa una proporción muy pequeña del total del Gasto Nacional en Educación, colocándose muy por detrás del financiamiento federal a la educación básica. Esta problemática la agudizan las cifras presentadas por la UNESCO en su informe regional de "Educación para todos 2011", en la que señala, que mientras en educación básica cerca del 98 por ciento de los niños tienen acceso a la escuela, esta cifra cae a un 24 por ciento en la educación superior en nuestro país.

Uno de los principales medios utilizados para reducir las desigualdades sociales entre los alumnos del nivel superior y del resto de los niveles, es a través del programa de becas. La beca es definida como el apoyo económico en efectivo que se otorga de manera permanente y suficiente a aquellos estudiantes de bajos recursos cuya situación económica les dificulta solicitar su ingreso a los estudios o permanecer en ellos. (González, 2006, p. 280). Las becas, por tanto se convierten en uno de los principales mecanismos compensatorios para combatir las desigualdades sociales. En este sentido, el sexenio de Vicente Fox, dentro del Programa Nacional de Educación (PRONAE) 2001-2006, creó, en el 2001, el Programa Nacional de Becas para la Educación Superior (PRONABES) con el fin de fomentar que una mayor proporción de jóvenes en condiciones económicas adversas accedan a los servicios públicos de Educación Superior y terminen oportunamente sus

estudios. Se promovió que este programa de becas, no reembolsable, contara con el apoyo económico y la colaboración de los gobiernos de los estados (SESIC, 2003, p.2).

El programa está dirigido a alumnos que provienen de familias cuyos ingresos son menores a tres salarios mínimos, con la finalidad de propiciar que los estudiantes en situación económica adversa, pero con deseos de superación, puedan continuar su formación académica en el nivel superior, y lograr la equidad educativa mediante la ampliación de oportunidades de acceso y permanencia en programas educativos de reconocida calidad, ofrecidos por las instituciones públicas de Educación Superior del país. Por ello, resulta importante estudiar cómo ha sido la asignación del apoyo de PRONABES en las entidades federativas del país. Las becas PRONABES juegan un papel muy importante en el fortalecimiento del desarrollo humano; pues en la medida que estas logren cimentar su participación en la equidad educativa, podrán convertirse en un factor detonador del desarrollo nacional. Sin embargo, el éxito de éstas, se encuentra en gran parte condicionado por la ayuda financiera que el Estado brinde a este rubro, especialmente, a los sectores de la población más necesitados.

Se han llevado a cabo diversas investigaciones concernientes al financiamiento público de la Educación Superior, dentro de las cuales se destaca la postura de Guevara (2007) al considerar que a pesar de la necesidad de apoyar a la Educación Superior, su política de financiamiento se ha encontrado subordinada a la política económica neoliberal, restringiendo así el gasto público (disciplina fiscal). Según este estudio, las becas PRONABES durante el periodo 2001-2005, fueron el único tipo de subsidio de financiamiento extraordinario que mantuvo un crecimiento; pero a pesar de ello, sólo 1 de cada 10 alumnos en Educación Superior se han visto beneficiados.

Por otro lado, González (2006), a través de la óptica de las relaciones intergubernamentales, es decir, entre los vínculos que se establecen entre ámbitos de gobierno y entre éstos y otro tipo de actores e instituciones, se estudia al PRONABES; para que su operatividad se encuentre caracterizada por los principios de la no autoridad, la cooperación, la transferencia de políticas y la canalización de recursos. De esta manera, González plantea la posibilidad de que la legislación obligué a la Secretaría de Educación

Pública, a los estados y a las instituciones públicas de Educación Superior, a integrar todos los recursos destinados a becas en este rubro en un sólo fondo, administrado por un solo comité.

El principal objetivo de este artículo se centra en conocer la relación existente entre la proporción de becas PRONABES respecto a la matrícula de Educación Superior, frente al COB de cada entidad y al recurso financiero público que se asigna en este programa, así como conocer el comportamiento de la proporción de becas PRONABES a lo largo del tiempo durante el periodo 2002-2007. Asimismo, interesa conocer si existe variabilidad entre las entidades. Primeramente, se describirá la metodología utilizada para el desarrollo de la presente investigación, presentando las características de la base de datos bajo estudio y el tipo de análisis realizados. Posteriormente, se expondrán los resultados obtenidos con el ajuste del modelo propuesto, así como los puntos expuestos a discusión frente a los argumentos de los autores antes estudiados. Finalmente, se presentan las conclusiones alcanzadas después de haber realizado los análisis correspondientes.

Metodología

Para llevar a cabo el presente estudio, se obtuvieron datos del Anexo Estadístico del Cuarto Informe de Gobierno 2010 y se extrajo la siguiente información: proporción de becas PRONABES respecto al número de matrícula, los recursos públicos asignados a este programa en cada entidad federativa y el porcentaje de cobertura de cada una durante el periodo 2002-2007. En la tabla 1 se describen cada una de estas variables y el valor de la media nacional.

Tabla 1. Descripción y clasificación de las variables de estudio.

Variable	Descripción	Media nacional
BECAS	Porcentaje de becas PRONABES asignadas en el periodo 2002-2007 respecto al número de estudiantes matriculados en el nivel de educación superior en cada entidad federativa.	8.4%
REC	Monto de los recursos asignados al programa PRONABES en cada entidad durante el periodo 2002-2007.	20.806 millones de pesos
COB	Índice de cobertura. Porcentaje de alumnos en el nivel de educación superior, calculado respecto al número de personas en edad de estudiar este nivel (19 a 23 años). Es un indicador de la capacidad del sistema en Educación Superior de atender a la población en edad de estudiar.	22.4%

Se realizó un análisis exploratorio para conocer el comportamiento de las variables de estudio, utilizando un gráfico de perfil de tiempo para observar el comportamiento de la proporción de becas a lo largo del periodo 2002-2007 entre las entidades federativas, así también se emplearon gráficas de dispersión para conocer el tipo de relación existente entre las variables de cobertura y montos asignados para cada una de las entidades federativas.

Debido al interés de estudiar el patrón de crecimiento de la proporción de becas asignadas en cada entidad federativa a lo largo del periodo 2002-2007, se utilizó un modelo de dos niveles (Goldstein, 1999), en el que los años, considerados como unidades de nivel 1, están anidados dentro de las entidades federativas, representando las unidades de nivel 2, como se muestra en la figura 1. De esta manera, se tiene información que varía en cada uno de los años de estudio para cada entidad federativa.

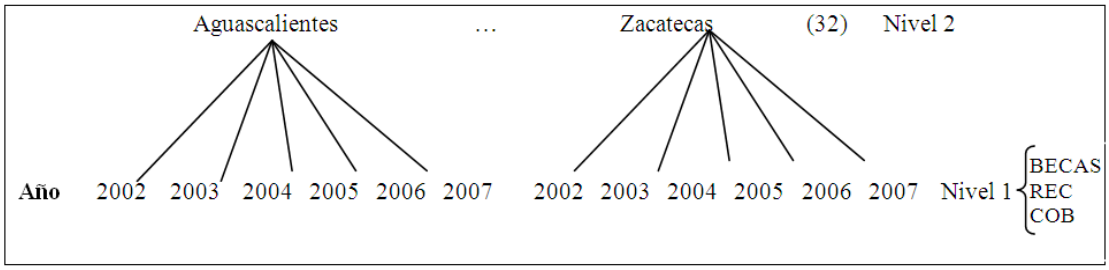


Figura 1. Diagramas de unidad para la estructura jerárquica de los datos bajo estudio.

A través de la modelación multinivel, se pretende tener una mejor comprensión de la variabilidad de los datos, pues esta metodología permite conocer la varianza entre los años y las entidades federativas respecto a la asignación de becas PRONABES, teniendo en cuenta las variables explicativas consideradas en el estudio. Es decir, la variabilidad en los datos en cada nivel y entre niveles se analiza simultáneamente.

La variable objeto de estudio es la proporción de becas asignadas de acuerdo al número de alumnos matriculados en el sistema de Educación Superior, por lo que se tienen para cada una de las entidades federativas, 6 mediciones correspondientes a la proporción de becas PRONABES que se registraron en los años del 2002-2007. Sea y_{ij} el valor de la proporción de becas**,** donde *i* representa el año (*i*=2002, 2003,…,2007) y *j* las entidades federativas (*j*=1, 2, …,32), con tres variables explicativas, AÑOS, REC y COB. El modelo en el primer y segundo nivel queda representado de la siguiente manera:

Nivel 1: $y_{ij}=\beta_{0j}+\beta_{1j}AÑOS_{ij}+\beta_{2j}REC_{ij}+\beta_{3j}COB_{ij}+e_{ij}$

Nivel 2: $\beta_{0j}=\gamma_{00}+u_{0j}$

$$\beta_{1j}=\gamma_{10}+u_{1j}$$

$$\beta_{2j}=\gamma_{20}$$

$$\beta_{3j}=\gamma_{30}$$

Modelo combinado: $y_{ij}=\gamma_{00}+\gamma_{10}AÑOS_{ij}+\gamma_{20}REC_{ij}+\gamma_{30}COB_{ij}+u_{0j}+u_{1j}+e_{ij}$

$$\begin{bmatrix}u_{0j}\\u_{1j}\end{bmatrix} \sim N(0,\Omega_u):\Omega_u=\begin{bmatrix}\sigma^2_{u0} & .\\ \sigma^2_{u01} & \sigma^2_{u1}\end{bmatrix}$$

donde γ_{00} representa el intercepto o la media global; mientras que γ_{10}, γ_{20} y γ_{30} constituyen la pendiente o el cambio en la media de la proporción de becas, cuando hay un cambio unitario en cada variable explicatoria (tiempo, (AÑOS), COB y recursos del programa (REC)) respectivamente, manteniendo la otra variable constante. Con este modelo, lo que interesa es conocer si alguna variable del primer nivel influye en el porcentaje de becas

respecto al número de matrícula y de esta manera conocer si estos factores explican la variabilidad existente.

Resultados y discusión

Un análisis preliminar de los datos, mostró que en el periodo 2002-2007, la entidad federativa que registró una mayor proporción de becas PRONABES fue el estado de Veracruz, con un 18.65%, y la que menos registró con un 2.93% fue Sonora. En cuanto al COB, se observó que el Distrito Federal seguido por el estado de Nuevo León, reportan los porcentajes más altos del país, 45 y 32% respectivamente, como se muestra en le Figura 1, lo que significa que casi la mitad de la población entre los 19 y 23 años que habita en el Distrito Federal, tiene acceso a la Educación Superior; mientras que el estado de Chiapas y Quintana Roo, registran el COB más bajos del país, con sólo un 13.1 y 12% respectivamente de su población accediendo a este nivel. Cabe destacar, como se indicó en la tabla 1, que la media nacional del COB fue de 22.4%, lo que significa que en el país sólo 2 de cada 10 jóvenes entre los 19 y 23 años tiene acceso a la Educación Superior. Asimismo, se puede apreciar en la Figura 1 que el COB varía considerablemente entre las entidades federativas.

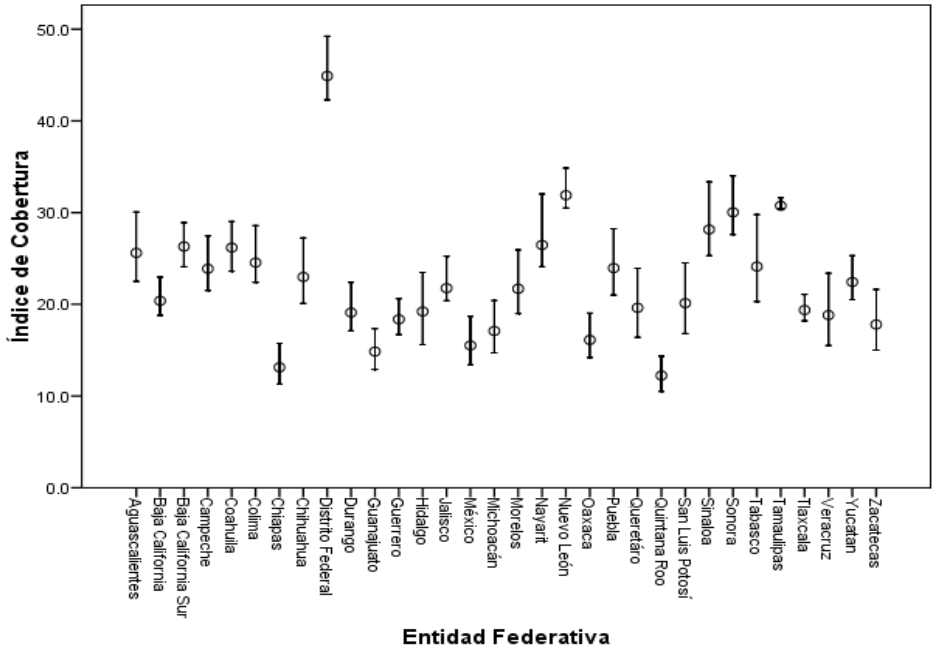


Figura 1. COB del periodo 2002-2007 por entidad federativa.

En cuanto a los recursos ejercidos por cada entidad federativa, en el rubro de PRONABES durante el periodo 2002-2007, se observó que Veracruz fue la entidad que ejerció una mayor cantidad de recursos con un total de 76.14 mil millones de pesos, después del Distrito Federal que encabeza la lista con un ejercicio de 76.35 mil millones de pesos, de lo que se interpreta que el estado de Veracruz tiene un peso importante a nivel nacional tanto en el porcentaje de becas PRONABES respecto a la matrícula, como en los recursos ejercidos bajo este rubro. Lo siguen el Estado de México y Puebla, con 63.9 y 29 mil millones de pesos respectivamente.

Por otra parte, se aprecia que el porcentaje de becas del PRONABES respecto a la matrícula ha ido en aumento a través del tiempo. En el año 2002, el promedio del porcentaje de becas asignadas a nivel nacional era de 4 % respecto a la matrícula de Educación Superior, cifra que se incrementó a un 11.5% en el año 2007. Además, se observa (Véase Figura 2) la presencia de datos atípicos en el 2002, 2004, 2005, y 2006, en donde Veracruz encabeza la lista de estados con mayor proporción de becas respecto al número de matrícula durante el periodo 2004-2006, confirmando que es una de las entidades que más ejerce este tipo de recursos. Sin embargo, en el año 2002, recién creado el programa, son las entidades de Yucatán, Quintana Roo, Chiapas y Guanajuato las que mayor porcentaje de becas registraron.

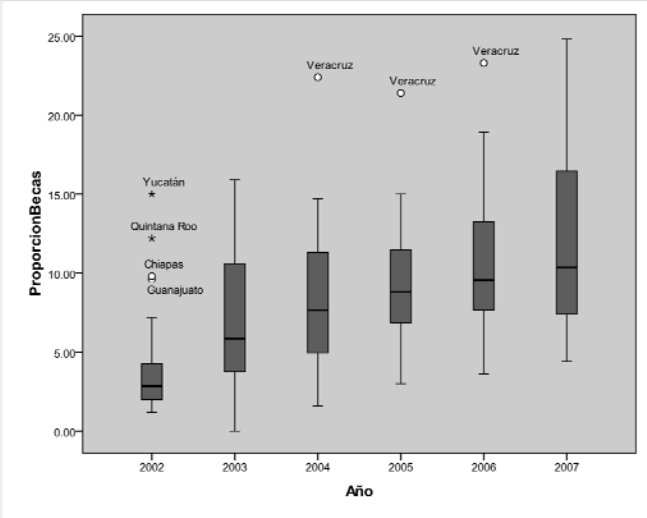


Figura 2. Tendencia y Variabilidad de la Proporción de becas asignadas a Educación Superior de acuerdo al número de matrícula (2002-2007).

Al relacionar la proporción de becas asignadas respecto a la matrícula (BECAS), con los años del periodo de estudio para cada entidad federativa, se aprecia una primera aproximación de la relación lineal existente entre BECAS y el tiempo. Es decir, la proporción de BECAS muestra una tendencia a aumentar a través de los años (Véase Figura 3). Asimismo, se observa que hay una variabilidad entre las entidades respecto a la proporción de BECAS.

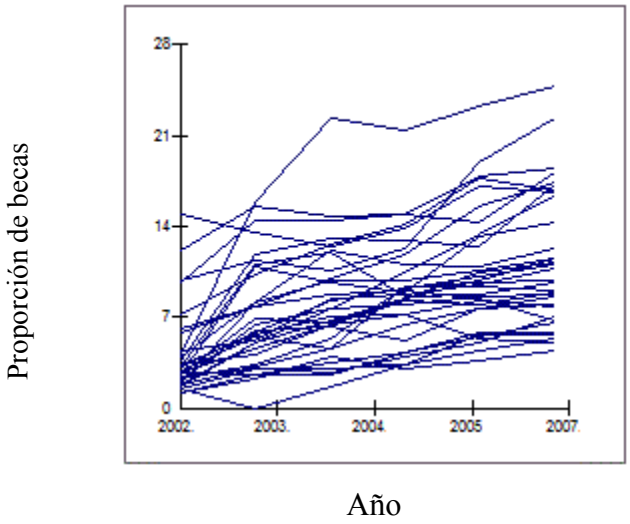


Figura 3. Relación entre la proporción de becas PRONABES asignadas respecto a la matrícula y los años del periodo de estudio.

Para corroborar los factores que contribuyen a explicar por qué hay variación en el tiempo y entre las entidades federativas respecto a las becas PRONABES, se ajustaron varios modelos multinivel, utilizando el método de Mínimos Cuadrados Generalizados Iterativos (Longford, 1999). Los resultados de las estimaciones se muestran en la tabla 2. En el primer modelo multinivel estimado, se mantuvo fija la pendiente y el intercepto aleatorio y se introdujo la variable años como variable explicativa, los resultados del ajuste muestran que la variable tiempo sí resulta significativa, es decir, que cada año, el porcentaje de becas PRONABES se incrementa en 1.379% en promedio. En los resultados del modelo, también se observó que hay variabilidad entre los años y entre las entidades.

En el segundo modelo, se introdujo adicionalmente el COB y se modeló como constante (igual que la variable AÑOS). En los resultados mostrados en la tabla 2, se observa que igualmente es significativa aunque con un efecto negativo, esto quiere decir,

que ante un cambio unitario en el COB de cada entidad, el porcentaje de becas disminuye en 0.263%, manteniendo la variable AÑOS constante. Al comparar los modelos 1 y 2, se aprecia que la varianza se reduce a 11.209 y que también hay una disminución en el valor de -2*log-likelihood de 910.73- 899.24= 11.49, que al compararlo con una distribución χ^2 con 2 grados de libertad, resulta significativa. Lo que indica que el segundo modelo está mejor ajustado a los datos.

Tabla 2. Resultados de las estimaciones del modelo de dos niveles

	Modelo intercepto aleatorio (1)	Modelo intercepto aleatorio 2 var ex. (2)	Modelo intercepto y pendiente aleatoria (3)
Parámetros fijos			
β_{0j} (Intercepto)	3.614 (0.759)	2.368 (0.759)	4.024 (0.560)
β_{1j} (Años)	1.379 (0.084)	1.734 (0.130)	1.261 (0.143)
β_{20} (Cob)		-0.263 (0.073)	- 0.242 (0.054)
β_{30} (Rec)			0.126 (0.013)
Parámetros aleatorios			
Nivel 2			
σ^2_{u0}	14.977 (3.910)	11.209 (2.944)	5.513 (1.711)
σ^2_{u1}			0.054 (0.294)
σ^2_{u0u1}			0.305 (0.098)
Nivel 1			
σ^2_{e}	3.968 (0.444)	3.903 (0.438)	1.500(0.187)
Deviance			
-2*loglikelihood	910.739	899.247	776.248

Sin embargo, falta incluir los recursos que destina cada entidad al programa PRONABES, lo que se realiza en el tercer modelo. Aunado a ello, como se había observado en el gráfico de perfiles de tiempo, las entidades varían en el tiempo tanto en su intercepto como en su pendiente, por lo que se optó por modelar la variable AÑOS como aleatoria, es decir que varíe de estado a estado. Para ello, se introduce el término de error u_{1j}, que representa la variación de la pendiente para cada entidad. Los resultados de las estimaciones de este tercer modelo muestran que las tres variables incluidas son significativas y que la variable recursos (REC) tiene un efecto positivo sobre el porcentaje de becas, es decir, por cada millón de pesos que los estados incrementen al programa PRONABES, la proporción de becas aumenta en un 0.126% manteniendo la variable AÑOS y COB constante.

Comparando este tercer modelo de coeficientes aleatorios, con los modelos anteriores, se observa que es el que mejor se ajusta a los datos, la varianza se reduce considerablemente entre las entidades, pasando de 14.977 a 5.513, la varianza entre los años también se reduce de 3.96 a 1.4, e igualmente la reducción del valor de -2*loglikelihood resultó significativa (910.73-776.248=134.48).

Una vez que se ha realizado el ajuste del modelo, es importante corroborar el cumplimiento de los supuestos y realizar un diagnóstico de los residuos para identificar datos atípicos. Como se había observado en la figura 2, el estado de Veracruz parece tener una mayor influencia en el modelo, lo que podría alterar el valor de las estimaciones. Con tal finalidad, se obtuvieron los gráficos de los residuos para los dos niveles de variación.

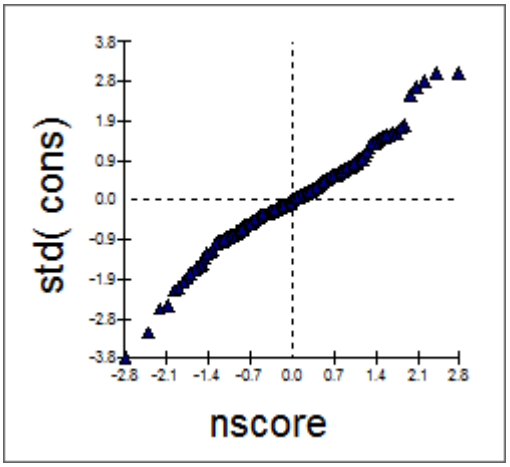


Figura 4. Gráfico de los errores a nivel año

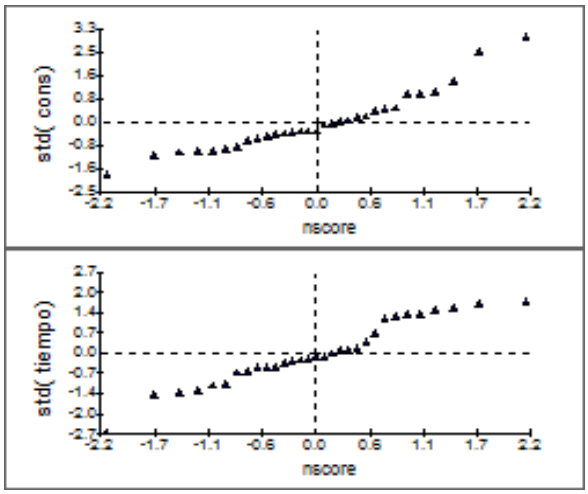


Figura 5. Gráfico de los errores nivel 2

Primeramente, se observa el cumplimiento de normalidad de los residuos. Por otro lado, a nivel año, sí se observa la existencia de datos atípicos en el extremo inferior de la gráfica, estos puntos representan a los estados de Chiapas y Puebla, así como un grupo de entidades en el extremo superior (Baja California Sur, Guanajuato y Yucatan). Sin embargo, ninguna de ellas es Veracruz y al evaluar su peso dentro de las estimaciones, no resultan significativas. Es decir, los resultados mostrados en la tabla 3, no se alteran considerablemente.

Los resultados obtenidos permiten corroborar las apreciaciones en el análisis exploratorio, sobre la variabilidad en el porcentaje de becas PRONABES de cada entidad; por lo que resulta razonable pensar que a través de los años se ha cumplido con el propósito de incrementar estos apoyos, permitiendo mejorar el COB en educación superior de cada entidad y lo que se destina en cuenta a recurso económico para este programa.

Es innegable el esfuerzo financiero que ha estado realizando el Estado mexicano en materia de subsidio a la Educación Superior Pública, sin embargo, las cifras deben mejorarse y aunque el programa de becas PRONABES ha resultado ser un buen apoyo, el COB aún no pasa del 24% en promedio a nivel nacional, lo que significa que cerca del 76% de la población en edad de recibir educación superior, no lo hacen.

Según Guevara (2007), en su estudio menciona que al 2005, sólo 1 de cada 10 alumnos contaba con la beca PRONABES; sin embargo, en el presente artículo se determinó que para el 2005 la proporción de asignación de becas oscilaba en 1 de cada 8, y es hasta el 2007 que se llega a la proporción de 1 de cada 10 alumnos, siendo Veracruz uno de los estados más beneficiados en este rubro.

Por último, y retomando la propuesta de González (2006), cabe señalar que si se desea que el PRONABES pueda contribuir aún más al apoyo de las becas, es importante crear los mecanismos intergubernamentales que permitan la operatividad de dicho programa, por lo que la creación de un fondo común (para canalizar todos los esfuerzos financieros tanto de federación, estados e instituciones) permitiría focalizar aún más la asignación de dicho beneficio económico, y de esta manera, ayudar realmente a quien más lo necesite, favoreciendo la reducción de las brechas de desarrollo.

Conclusión

A lo largo de este artículo se pudo conocer la evolución de la proporción de becas PRONABES entre las entidades federativas a través del tiempo, en el periodo 2002-2007. Este análisis permitió mostrar que el programa de becas PRONABES implementado desde el año 2001, ha resultado eficiente en el cumplimiento de sus objetivos, pues a lo largo del periodo analizado se han incrementado los recursos destinados para este programa que han reflejado un aumento en la proporción de becas que se asignan en cada entidad federativa. Esto ha permitido que cada vez más estudiantes en edad de cursar estudios en Educación

Superior accedan a este nivel, pues los resultados del modelo mostraron que al aumentar el COB de cada entidad federativa, la proporción de becas disminuye al haber menos estudiantes que se quedan si esta oportunidad.

Sin embargo, aunque se observó el incremento que ha habido en este apoyo, también se reflejaron las diferencias que aún existen entre las entidades federativas del país, pues hay algunos estados cuyo COB sigue estando muy por debajo de la media nacional, lo que significa que las brechas en cuanto al desarrollo en Educación Superior siguen ampliándose, pues las entidades con mayor riqueza son las que cuentan con un mayor acceso a la Educación Superior. La media nacional del COB, comparada con otros países, también muestra un rezago importante.

Se hace evidente la necesidad de incrementar los recursos públicos destinados a la Educación Superior, porque aunque se ha logrado el avance en los últimos años, los resultados muestra que aún hay mucho por hacer para mejorar los indicadores de acceso a la Educación Superior a nivel internacional.

Referencias

Cámara de diputados. (2010). *Constitución Política de los Estados Unidos Mexicanos*. Cámara de diputados, D.F.

Goldstein, H. (1999). *Multilevel Statistical Models*. London. First Internet edition.

González, J. (2006). El PRONABES: un enfoque de relaciones intergubernamentales. *Espacios públicos*, Febrero, 17 (9), 275-291.

Guevara, I. (2007). Reestructuración financiera a inicio del siglo XXI. El caso del gasto público en Educación Superior. *Economía Informal*. 349 (noviembre-diciembre), 58-68.

Rodriguez, R. (1996). *Educación Superior y Desigualdad Social. Un estudio sobre las determinaciones sociales y académicas de las trayectorias escolares en la UNAM*. El Colegio de México, D.F.

Subsecretaría de Educación Superior en Investigación Científica. (2003). *Reglas de operación e indicadores de evaluación y gestión del PRONABES.* Diario Oficial de la Federación, D.F.

Causalidad entre los ingresos y egresos de los gobiernos locales de México

Roberto Gallardo del Ángel[12]
Alberto Aguilar López[13]

Resumen

Este estudio examina las relaciones causales intertemporales entre los ingresos y egresos de los gobiernos locales de México, usando la técnica de vectores autorregresivos de Arellano-Bond, para analizar la causalidad entre las principales variables que determinan el presupuesto público. Se incluyeron datos de una muestra de 67 municipios medianos y grandes, para el periodo de 1995 hasta 2004. El mismo análisis se efectuó para los 31 estados de la república (excluyendo el Distrito Federal). Los resultados sugieren que tanto los ingresos como los egresos cambian simultáneamente, lo que implica doble causalidad en sentido de Granger. En otras palabras, los ingresos y los egresos municipales se determinan de manera simultánea, incluyendo rubros fuera del ámbito local como son las transferencias federales.

Palabras Clave: Ingresos y egresos locales, VAR, Causalidad Granger, municipios.

Abstract

This preliminary study explains the causal-intertemporal nexus between local revenues and expenditures in Mexico, using the VAR-Panel Arellano-Bond technique, to analyze causality among the main variables in the municipal budget. We included data from a sample of 67 medium-size and large municipalities for a ten years period of analysis (1995 to 2004). We also made the same analysis for 31 states (excluding Mexico city). We find out that both local revenue and expenditure change simultaneously therefore Granger-cause each other. In other words, local revenues and expenditures are determined at the same time, including receipts outside the local sphere like federal transfers.

Key words: Local revenues and expenditures, VAR, Granger Causality, Municipalities.

[12] rogallardo@uv.mx.
[13] alberto.aglo@gmail.com

Introducción

Este artículo se enmarca dentro de la literatura de las finanzas públicas locales, y contiene principalmente una reflexión sobre la secuencia en la determinación del presupuesto en los gobiernos locales. Es decir, sobre qué es lo que se determina primero, los ingresos o los egresos de los gobiernos municipales y los estados. Aunque parezca una hipótesis un poco ociosa, tiene implicaciones teóricas importantes, sobre todo cuando intentamos comprender el proceso político por el que atraviesa un presupuesto para su aprobación y ejecución.

En muchos países, la secuencia en la aprobación y ejecución de estas dos identidades financieras está determinada por las instituciones económicas y políticas, que incluyen diversas prácticas y reglamentaciones. Si para un país una determinada causalidad o secuencia es normal, para otro país podría no serlo, incluso las prácticas podrían ser diferentes entre niveles de gobierno. Estas diferencias en la determinación de los ingresos y gastos dadas las instituciones políticas locales, se debe reflejar en las estadísticas oficiales que presentan los municipios y los estados respecto a esos mismos ingresos y gastos.

El presente trabajo preliminar contiene un análisis econométrico que investiga la causalidad entre los ingresos y gastos de los gobiernos locales. El objetivo principal es describir el comportamiento fiscal de estas unidades de gobierno, así como incrementar los estudios acerca de las estructuras institucionales y los arreglos políticos que determinan las finanzas públicas de los gobiernos locales de México.

La literatura al respecto distingue los siguientes tres tipos de relaciones causales intertemporales entre los ingresos y egresos locales:

a) Ingresos y Egresos se deciden al mismo tiempo y cambian simultáneamente. En este caso, la toma de decisiones es llevada a cabo por votantes y burócratas al mismo tiempo. Queda por demostrar si la provisión de bienes públicos es decidida a través del voto o se refleja en el gasto público aprobado. Una ciudad o comunidad podría aprobar en un solo ejercicio político el ingreso fiscal y al mismo tiempo decidir sobre la cantidad de bienes públicos que requiere. Aunque qué

tanto del presupuesto lo deciden los individuos a través del voto también está influenciado por las instituciones políticas vigentes.

b) Los movimientos pasados en los ingresos locales explican el nivel actual de gastos. En este caso, los votantes deciden sobre los impuestos y demás ingresos locales y después deciden sobre el tipo de bienes públicos locales que necesitan. Es decir, primero se estructuran los recursos que fondearan los bienes públicos y luego se ejercerán en forma de gasto. Esta relación provee una explicación racional del proceso de toma de decisiones acerca de la provisión de bienes y servicios públicos. Holtz-Eakin, Newey y Rosen (1988), en un estudio de una muestra de 171 municipios de Estados Unidos durante 1971 hasta 1980, encontraron evidencia de que los ingresos pasados explican el nivel actual de gastos, pero el gasto pasado no altera el patrón futuro de ingresos.

c) Una tercera visión describe que los gastos locales cambian antes que los ingresos. Esta es probablemente la visión partidaria de la escuela Italiana de finanzas públicas, donde la demanda y la oferta de bienes públicos locales generan una creciente necesidad de gasto público. La evidencia sugiere que al nivel federal y ante algunas circunstancias históricas, ya sea en forma de guerras, crisis financieras, o recesiones económicas, el movimiento en el gasto público puede preceder a los ingresos. Un estudio de von Fustenberg, Green y Jeong (1986) demuestra que a nivel federal, en los Estados Unidos los cambios pasados en los niveles de impuestos no modifican el gasto actual, pero el nivel de gasto actual puede incrementar los ingresos futuros derivados de impuestos.

Una vez explicada brevemente las tres grandes visiones sobre la determinación de las identidades presupuestales, queda por identificar algunos aspectos adicionales acerca de la estructura fiscal de los gobiernos locales de México. Lo primero es que la competencia política en los municipios de México no se basa únicamente en la provisión de bienes públicos, debido a que éstos se financian en su mayoría por los niveles federal y estatal. El electorado no decide acerca del tipo y cantidad de bienes públicos locales como puede suceder en otros países. Los ingresos de los ayuntamientos dependen en gran medida de las transferencias desde niveles de gobiernos superiores, lo que implica una cantidad

relativamente fija de bienes públicos. En segundo lugar, el sistema electoral en México está ampliamente influenciado por arreglos institucionales con partidos políticos, confederaciones laborales, y otros grupos de poder que pueden incidir en el resultado de las elecciones. Estos arreglos pueden y tienden a modificar las propuestas programáticas, en el caso de que alguna propuesta se hubiese planteado en el proceso electoral y se hubiese planteado como ley. Por último, hay que subrayar que los Municipios dependen en gran medida de las transferencias o subvenciones, y tienen poca autonomía sobre las decisiones de ingresos y, en algunos casos, restricciones de gasto.

Otros estudios, como Moisio (2004), sobre una base panel de municipios de Finlandia, sugieren causalidad en el sentido de Granger de los gastos sobre los ingresos. Se halló unidireccionalidad durante el periodo en el que las transferencias eran condicionadas, pero durante el periodo donde se estableció una fórmula para su reparto se encontró fuerte evidencia a favor de la doble causalidad Dahlberg y Johansson (2000). En su estudio sobre municipios de Suecia estimaron una ecuación de gasto y encontraron causalidad Granger de los impuestos sobre las transferencias.

Estos estudios muestran significativas diferencias entre las estructuras fiscales de cada país, así como en los arreglos institucionales y políticos de cada uno de ellos. En México, la unidad política más pequeña de la administración pública es el Municipio, y varía en tamaño, población, recursos y actividad económica. Los acuerdos entre los gobiernos locales y la federación también varían.

En México hay poco más de 2400 municipios, distribuidos entre 31 Entidades Federativas, constituidos como entidades independientes y soberanas, con su propia estructura fiscal y administrativa. Sin embargo, en la práctica estas unidades políticas tienen menor autonomía que otros gobiernos locales en diferentes países del mundo, en términos de atribuciones fiscales y autonomía del gasto. La mayor limitante es la dependencia de las transferencias federales para sostener la administración pública, así como la falta de un servicio público de carrera, lo que dificulta el proceso de toma de decisiones. Para cada administración local, las políticas y los proyectos cambian de acuerdo a cómo se renueva la nómina de burócratas a cargo de las decisiones principales.

Una gran parte de los bienes públicos locales son suministrados por el gobierno federal y los estados, dejando al Ayuntamiento (Municipio) con pocas obligaciones. Esto fue ampliamente criticado durante la década de los ochentas, una época donde el gobierno federal centralizaba gran parte de la actividad económica y política, e incluso decidía sobre los aspectos locales. Para los años noventas, el gobierno federal redujo su injerencia sobre los asuntos de los gobiernos locales, y mayores fondos y autonomía fueron distribuidos hacia los Estados y Ayuntamientos. Un periodo de "devoluciones" inició con la reforma del sistema federal de transferencias, y su división en partidas de las cuales se tienen dos grupos principales: las transferencias no condicionadas llamadas "participaciones", y posteriormente los recursos federales condicionados llamados "aportaciones". Sin embargo, únicamente pequeñas funciones de asignación fueron transferidas a los Ayuntamientos. A pesar de ello, un periodo de mayor autonomía en la toma de decisiones locales se había iniciado.

Aunque este nuevo federalismo mexicano trató de crear una nueva relación con los gobiernos locales, la dependencia de los fondos federales se incrementó a nuevos niveles; pues los mecanismos legales no incentivaron la autonomía local. Lentamente, las reformas adheridas a las leyes buscan reactivar los gobiernos locales y dotarlos de mayores atribuciones para que sean ellos quienes obtengan sus recursos. Sin embargo, los gobiernos locales aun no pueden operar sus administraciones sin las transferencias federales.

Metodología

Este artículo estudia la causalidad entre ingresos y gastos locales usando una muestra de municipios medianos y grandes, así como de los estados de México. La muestra fue construida tomando en cuenta la población y la estructura económica. Al final, se integró una base de datos con 67 ayuntamientos. Se formó otra base integrada con datos de los 31 estados del país, excluyendo a la Ciudad de México, debido a su diferente situación política y administrativa. El punto principal es verificar cómo los movimientos de los ingresos y egresos se proceden uno a otro en el tiempo.

Los datos se recabaron de forma anual, y fueron obtenidos directamente del Instituto Nacional de Estadística y Geografía (INEGI); incluyen ingresos locales como impuestos,

transferencias y otros ingresos locales. Por el lado del gasto, éste se conforma por la inversión en capital físico, así como el gasto administrativo.

Se emplearon variables como impuestos locales, otros ingresos locales (multas y fianzas, ingresos por servicios municipales, etc.), y transferencias, la cuales incluyen los fondos condicionados y los no condicionados. Otras dos variables agregadas integraron las partidas de gastos, e incluyen los costos administrativos y operativos (gasto corriente), así como la inversión local en infraestructura y desarrollo social (gasto de inversión). Los datos fueron deflactados usando el IPC de 2002, y transformados en logaritmo natural.

Con el fin de probar causalidad entre ingresos y gastos locales, se empleará el concepto de causalidad en sentido de Granger, y se evaluarán las relaciones intertemporales entre estas cuentas. De acuerdo a Hamilton (1994), si un escalar *y* no ayuda a pronosticar a otro escalar *x*, decimos que *y* no es Granger-causal con *x*. Formalmente, *y* no es Granger-causal con *x* si para todo $s > 0$, el error cuadrático medio (MSE) de un pronóstico de x_{t+s} basado en $(x_t, x_{t-1},...)$, es el mismo que el error cuadrático medio de un pronóstico de x_{t+s} que usa los valores pasados de *y* $(y_t, y_{t-1},...)$, esto es,

$$MSE[E(x_{t+s} \mid x_t, x_{t-1},...)] = MSE[E(x_{t+s} \mid x_t, x_{t-1},..., y_t, y_{t-1},...)]$$

Por lo tanto, es necesario usar una proyección lineal de los ingresos sobre los gastos, y evaluar la significancia de los parámetros estimados usando un grupo de variables independientes. El principal problema de usar datos de panel es que las series, por lo regular, están disponibles solo para unos cuantos años. Holtz-Eakin, *et al*. (1988) describen la metodología para una estimación econométrica de este tipo de relaciones causales. Esta técnica emplea el análisis VAR y evalúa la significancia de la estimación. Desafortunadamente, esto suele complicarse, debido a que las variables están expresadas en términos de su evolución en el tiempo, y esto genera problemas entre los regresores y el término de error: la parte autorregresiva de la variable dependiente está correlacionada con los errores en el mismo periodo, así que en presencia de esta endogeneidad, los estimadores están sesgados y son ineficientes.

Dado lo anterior, se requiere usar el estimador de Variables Instrumentales, ampliamente detallado en Holtz-Eakin, *et al*. (1988) y Hsiao (2003). Por ello, este estudio

incorporará la técnica de Arellano-Bond que ocupa variables instrumentales de tiempo para evitar las complicaciones detalladas. También se construyeron ecuaciones de ingreso y de gasto para la estimación, para posteriormente analizar la significancia estadística de cada regresor. Además, Holtz-Eakin, *et al*. (1988) también describen cómo determinar el tamaño óptimo del rezago, con el fin de verificar la dependencia temporal del la causalidad-Granger. Por último, se estimó un test de causalidad-Granger por pares, el cual ayuda a establecer la causalidad en una base de datos panel, entre dos variables. Esta es una técnica simple que servirá para verificar los resultados de la regresión.

Usando la metodología de Arellano-Bond (1991) y Henr (1988), se construyó una ecuación de gasto de la siguiente forma:

$$GT_{i,t} = GT_{i,t-1} + GT_{i,t-2} + TAX_{i,t-1} + TAX_{i,t-2} + OR_{i,t-1} + OR_{i,t-2} + GR_{i,t-1} + GR_{i,t-2} + e_{i,t}$$

Esta ecuación indica que el gasto total *GT* es una función del nivel pasado de ingresos totales, impuestos *TAX*, otros ingresos *OR* (honorarios, multas, etc.) y transferencias *GR* con un rezago de tiempo de magnitud $m=2$. Esta regresión fue estimada usando una matriz de instrumentos, diferenciada y con la técnica GMM descrita por Arellano-Bond.

Un análisis diferente se efectuó para identificar la causalidad por el lado de los ingresos, usando los niveles anteriores de gasto corriente gasto de inversión. La ecuación para los ingresos totales *TR* queda como sigue:

$$TR_{i,t} = TR_{i,t-1} + TR_{i,t-2} + GC_{i,t-1} + GC_{i,t-2} + GI_{i,t-1} + GI_{i,t-2} + e_{i,t}$$

Aquí, los ingresos totales *TR* son una función de sus niveles pasados, más los rezagos del gasto corriente *GC* y de inversión *GI*.

Resultados

En promedio, 3/4 de los ingresos estatales viene de las transferencias federales. La Figura 1 muestra el panorama actual de los ingresos estatales mientras que la Figura 2 muestra la distribución de los ingresos municipales para los años 1995, 2000 y 2004. Las gráficas expresan promedios de las cuentas de ingresos. Coincidentemente con las entidades

federativas, las gráficas mostraron el incremento en la importancia de las transferencias sobre el periodo de 10 años, al tiempo que el gobierno reformó los estatutos legales e implementó el sistema de transferencias condicionadas. Los impuestos ocupan únicamente un 10 por ciento del total de los ingresos de los ayuntamientos, y 2/3 partes de dichos fondos vienen de transferencias federales y estatales.

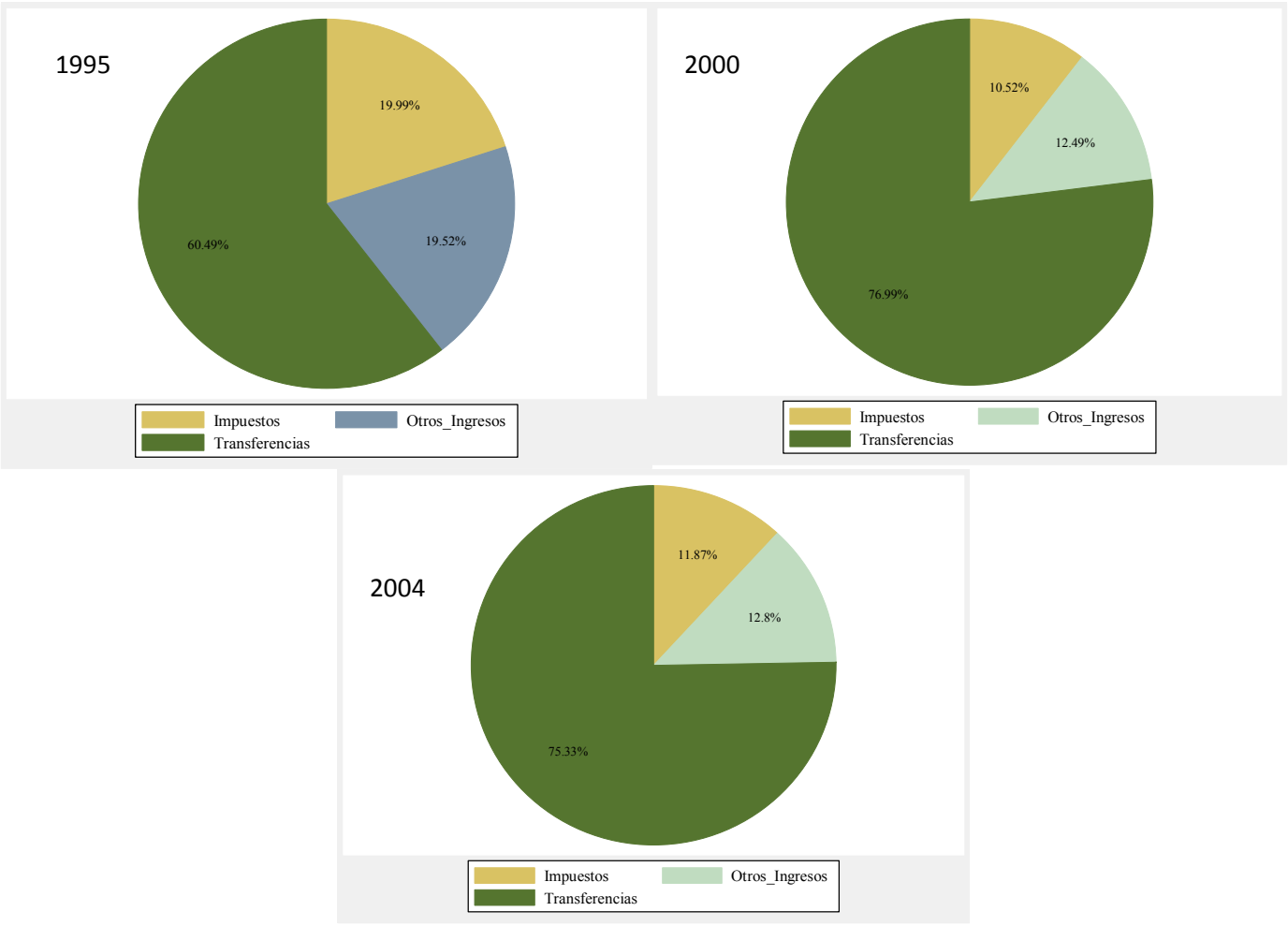


Figura 1. Ingresos Estatales promedio de 1995, 2000 y 2004.

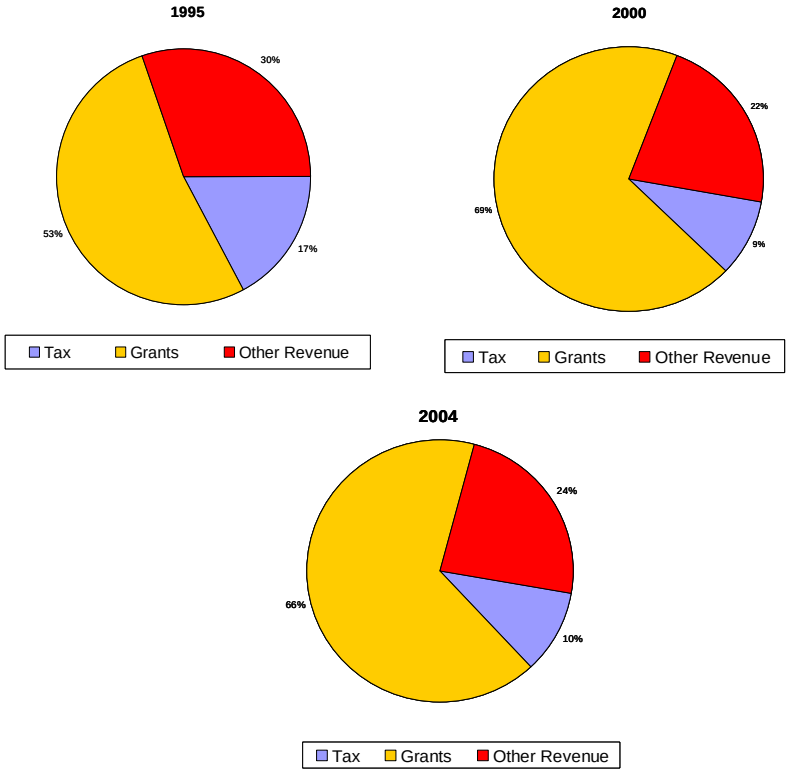


Figura 2. Ingresos Municipales promedio para 1995, 2000 y 2004.

Los resultados para la regresión de la ecuación de gasto se muestran en la Tabla 1 para el panel de 31 estados, que incluyó 248 observaciones con igual número de registros.

Tabla 1. Regresión de gasto para estados de México.

Variable	Coeficiente	Error estándar	Valor t	Prob.
GT(-1)	-0.064366	0.062831	-1.024426	0.3067
GT(-2)	-0.173977*	0.060872	-2.858102	0.0046
TAX(-1)	0.394933*	0.080073	4.932195	0
TAX(-2)	0.196022*	0.082577	2.373818	0.0184
OR(-1)	-0.249987*	0.098399	-2.540551	0.0117
OR(-2)	-0.008337	0.1018	-0.081898	0.9348
GR(-1)	-0.255712*	0.077983	-3.279091	0.0012
GR(-2)	0.152732**	0.082978	1.840636	0.0669

*Significativo al 5%. **Significant at 10%.

Los resultados muestran que la mayoría de las variables son significativas para el periodo de dos rezagos, con la única excepción de que otros ingresos únicamente lo es al segundo periodo rezagado. El mismo análisis se realizó para la muestra de 67 municipios, y

la tabla 2 da cuenta de los resultados. Los primeros rezagos para los impuestos y otros ingresos son significativos, al igual que las transferencias.

Tabla 2. Regresión de gasto para municipios de México.

Variable	Coeficiente	Error estándar	Valor t	Prob.
GT(-1)	-0.029679	0.051618	-0.574974	0.5656
GT(-2)	-0.2617*	0.04734	-5.528068	0
TAX(-1)	-0.102151*	0.047648	-2.14384	0.0326
TAX(-2)	0.013316	0.047127	0.282564	0.7776
OR (-1)	0.130731*	0.054302	2.407469	0.0165
OR (-2)	0.069162	0.054935	1.258976	0.2087
GR (-1)	0.312882*	0.080794	3.872604	0.0001
GR (-2)	-0.052873	0.07287	-0.725582	0.4685

*Significativo al 5%.

Los resultados para la regresión de la ecuación de ingresos se muestran en la tabla 3 para los Estados y para ayuntamientos se muestran en la tabla 4.

Tabla 3. Regresión para ingresos totales de los Estados.

Variable	Coeficiente	Error estándar	Valor t	Prob.
TR(-1)	-0.292816*	0.070734	-4.139659	0
TR(-2)	-0.223896*	0.071414	-3.135175	0.0019
GC(-1)	-0.105848	0.07154	-1.479554	0.1403
GC(-2)	0.019381	0.066436	0.291722	0.7707
GI(-1)	0.090334	0.055012	1.642075	0.1019
GI(-2)	0.115062*	0.053453	2.152581	0.0323

*Significativo al 5%.

Por la parte de la ecuación de ingresos, sólo en el segundo periodo el gasto en inversiones es significativo para estados. Por otro lado, la regresión para municipios muestra que solo el gasto corriente es significativo estadísticamente.

Tabla 4. Regresión para ingresos totales de municipios.

Variable	Coeficiente	Error estándar	Valor t	Prob.
TR(-1)	0.452498*	0.066315	6.82348	0
TR(-2)	0.009983	0.050286	0.198524	0.8427
GC(-1)	0.144224*	0.069687	2.069591	0.039
GC(-2)	-0.21046*	0.06292	-3.344877	0.0009
GI(-1)	-0.040433	0.048535	-0.833066	0.4052
GI(-2)	0.058091	0.044453	1.306797	0.1919

Una prueba de causalidad-Granger por pares (pair-wise) fue efectuada para ambas ecuaciones, con el fin de confirmar la doble direccionalidad encontrada en el análisis previo. El resultado, aunque no se incluyó en este artículo, muestra que la causalidad es significativa para los primeros cinco periodos tanto para gastos como para ingresos en ambas direcciones.

Los resultados preliminares mostraron una débil relación entre ingresos y gastos locales donde ambos están relacionados y se determinan simultáneamente. La causalidad es relativamente más fuerte por el lado de la ecuación de gasto, de tal modo que los niveles pasados de ingresos determinan los egresos actuales. Particularmente se ve que los ingresos de un año anterior influyen en el gasto total actual.

Por el lado de la ecuación de ingresos se tiene un resultado similar, que el nivel pasado de gastos causa los actuales ingresos. Aunque esta relación es débil para los estados, donde los ingresos parecen tener un criterio inercial donde el nivel de ingresos pasados determina el nivel de ingresos presente, a excepción del gasto de inversión en dos años anteriores. Por el lado de los municipios, el gasto corriente parece influir la decisión de ingresos totales actuales.

Aunque la provisión de bienes públicos no se decide totalmente por medio de las votaciones, y las elecciones locales no están necesariamente relacionadas con una cesta de bienes públicos, la explicación puede ser hallada en los arreglos institucionales del sistema fiscal mexicano. Los ingresos y egresos están vinculados a través de los candados administrativos impuestos por las leyes que regulan la relación entre los municipios y a la federación en términos de coordinación fiscal para la asignación de atribuciones y responsabilidades de tributación y ejercicio del gasto.

El presupuesto se determina en los congresos federal y locales, y es ahí donde se dictan las leyes de ingresos y egresos haciendo alusión a una estrategia de *planeación* en lugar de la revelación de preferencias a través del voto. En este sentido, tanto los ingresos como los egresos son decididos año con año por los representantes políticos de la sociedad que encuentran justificación en procesos de planeación institucional con poca participación de la sociedad civil.

Aunque los votantes determinan formalmente el resultado de los procesos políticos, las decisiones de provisión de los bienes públicos locales son tomadas en la esfera política. El presupuesto es decidido por instituciones y congresistas bajo un proceso preestablecido. Bajo esta perspectiva, se puede explicar la débil simultaneidad considerando que las preferencias a través del voto no son del todo tomadas en cuenta.

La débil simultaneidad hallada en este artículo requiere ser confirmada con más estudios. Posiblemente la determinación simultánea de los presupuestos locales se debe a rigideces institucionales. No ha habido grandes cambios en la forma de operar del sistema fiscal mexicano desde que la oposición ganó las elecciones federales del 2000. La estructura política y administrativa del sistema fiscal permaneció intacta y continúa funcionando bajo fórmulas específicas descritas en la Ley de Coordinación Fiscal y demás reglamentos federales y estatales, así como bajo procedimientos estables. Bajo estas condiciones, la suposición de simultaneidad parece ser razonable.

Referencias

Dahlberg, M. and Johansson, E. (2000). "*An examination of the dynamic behavior of local governments using GMM Bootstrapping methods"*, Journal of Applied Econometrics, No. 15.

Hamilton, J. D. (1994). *Time Series Analysis,* Princeton University Press, Princeton, pp. 302-306.

Hsiao, C. (2003). *Analysis of Panel Data.* 2nd Ed., Cambridge University Press, Cambridge, pp. 85-90.

Holtz-Eakin D., Newey W., and Rosen H. S. (1988). *Estimating vector autorregressions with panel data*, Econometrica, Vol. 56, No. 6, pp. 1371-1395.

Moisio, A. (2001). *Spend and tax, or tax and spend? Panel data evidence from Finnish municipalities during 1985-1999;* VATT; Government Institute for Economic Research, Helsinki.

Von Furstenberg G.M., Green J. and Jeong J. (1986). *Tax and Spend, or Spend and Tax?* The Review of Economics and Statistics, Vol. 68, No. 2, pp. 179-188.

Efecto de los contextos escolares en los resultados de la prueba ENLACE 2009: Un análisis multinivel.

Patricia Tapia Blásquez[14]
Mario Miguel Ojeda Ramírez[15]
Elizabeth Tapia Blásquez[16]

Resumen

En la última década, México ha experimentado un notable desarrollo en materia de evaluación educativa; se han aplicado sistemáticamente diversos instrumentos nacionales e internacionales para evaluar la calidad de la educación, sobre todo en la educación básica. Uno de tales instrumentos es la Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE). El objetivo del presente trabajo es analizar los resultados de esta prueba en al año 2009 considerando los contextos escolares, con el propósito de explicar los desempeños a nivel de escuela; para tal fin se utilizan estrategias de modelación multinivel, con lo que se valora la incidencia de los factores contextuales. Se demuestra que, características de las escuelas como el tipo de financiamiento que recibe (público o privado), el turno y el grado de marginación de la población en la que se ubican, influyen significativamente en el puntaje promedio obtenido en la prueba; así mismo, se concluye que los resultados enunciados se presentan a nivel nacional; es decir, las diferencias económicas y sociales de las entidades federativas no tienen efecto sobre este patrón en los resultados.

Palabras clave: Evaluación al desempeño, Educación primaria, Calidad de la Educación, Análisis estadístico (Modelos multinivel lineales), Investigación Educativa.

[14] Universidad Veracruzana. Av. Paseo de las Palmas No 15 esq. Los Mangos, Fracc. Jardines de las Ánimas C.P. 91190. Xalapa, Veracruz, México. Tel: +52 22 88 18 35 39: ptapia@uv.mx
[15] Profesor Investigador de la Universidad Veracruzana. Dirección General de Estudios de Posgrado: mojeda@uv.mx
[16] Docente frente a grupo de educación básica. Secretaría de Educación de Veracruz: ectb85@hotmail.com.

Introducción

A pesar de los esfuerzos que se venían realizando en educación, México presentaba un importante rezago hacia finales de los años noventa. La política educativa en ese momento, estaba enfocada a logros cuantitativos, los objetivos se traducían en el incremento de las tasas de matriculación y de cobertura, así como en la disminución de los índices de analfabetismo y deserción escolar. Sin embargo, se había dejado a un lado el tema de la calidad educativa, mismo que se hizo evidente ante los señalamientos de la OCDE (Organización para la Cooperación y el Desarrollo Económicos) con los resultados obtenidos por los alumnos mexicanos en la prueba que este organismo coordina, del Programa para la Evaluación Internacional de los Alumnos (PISA- Program for International Student Assesment). Los resultados mostraron que un porcentaje considerable de los estudiantes de quince años no cuenta con las habilidades mínimas para obtener la información a través de la lectura y no es capaz de resolver problemas matemáticos básicos[i].

Como señala Amador Hernández (2008), se considera que un indicador relevante para entender el problema de la calidad educativa es el desempeño de estudiantes de primaria y secundaria. Por dicha razón, el gobierno federal decidió adoptar estrategias diversas para elevar la calidad de la educación básica. El primer paso consistió en el diseño de un esquema que permitiera efectuar sistemáticamente la aplicación de pruebas del desempeño escolar para evaluar el aprendizaje de los alumnos a nivel estatal y nacional. De acuerdo con Rentería Castro (2010), para llevar a cabo las reformas de modernización educativa surgieron cuatro propuestas: el proyecto inicial del gobierno, autonombrado Modelo Pedagógico; el Nuevo Modelo Pedagógico elaborado por el equipo de trabajo del Secretario de Educación y analizado por el Consejo Nacional Técnico de la Educación (CONALTE); las siete propuestas para modernizar la primaria, formuladas por el Sindicato Nacional de Trabajadores de la Educación (SNTE) y el Acuerdo Nacional para la Modernización de la Educación Básica. Este último proyecto, que fue firmado en 1992, dio paso, diez años después, a la creación por decreto presidencial del Instituto Nacional para la Evaluación de la Educación (INEE). El acuerdo pretendió una redefinición de la relación entre el Estado y la sociedad, propiciando un acercamiento entre los actores que participan en el proceso educativo, revalorando el papel del maestro y de los padres de familia. De

esta manera, en el discurso, los gobiernos, federal y estatales, se comprometían a transformar el sistema de educación básica con el fin de asegurar a los niños y jóvenes una educación que los formara como ciudadanos de una comunidad democrática, proporcionándoles conocimientos para su ingreso a la vida productiva y social, y en general favorecería mejores niveles de vida (Pescador, 1992). Con esto se emprenden una serie de modificaciones, planes, programas, proyectos y sobre todo, evaluaciones aplicadas tanto a los profesores, como a los estudiantes.

El INEE inicia con la aplicación del Examen de Calidad y Logro Educativo (EXCALE) a los alumnos de sexto grado de primaria y tercer año de secundaria en las asignaturas de Español y Matemáticas, pero en mayo de 2006 el Secretario de Educación anunció que la Dirección General de Evaluación a través de la Dirección General de Evaluación de Políticas y Sistemas Educativos, iniciaría la aplicación de la prueba de Evaluación Nacional de Logros Académicos en los Centros Escolares (ENLACE), la cual ha ido adquiriendo, a lo largo de sus diferentes aplicaciones, una mayor fuerza, pues se ha convertido en un elemento determinante para catalogar el nivel educativo de las escuelas y los alumnos en México, posicionándose como un nuevo referente de evaluación académica. Tanto en el ámbito escolar como el de la opinión pública, el tema de la educación y su calidad están presentes en el análisis y la discusión (Amador Hernández, 2008: 21-22).

ENLACE consiste en la aplicación anual, al final del ciclo escolar, de un examen estandarizado de opción múltiple, a todos los estudiantes de escuelas públicas, privadas, indígenas y del sistema CONAFE (Consejo Nacional de Fomento Educativo)[ii] del país, que estén cursando tercer, cuarto, quinto y sexto grado de educación primaria, así como primer, segundo y tercer año de secundaria[iii], sobre las asignaturas de Español, Matemáticas y una tercera, que varía cada año de acuerdo al currículo vigente y que se elige en función de su cobertura en la carga horaria y de su perfil instrumental para abordar otros contenidos. Los resultados son expresando en un puntaje que va en un intervalo de 200 a 800 puntos y que se clasifica en cuatro niveles de logro: Insuficiente, elemental, bueno y excelente. La prueba se aplica con el propósito de "generar una escala de carácter nacional que proporcione información comparable de los conocimientos y habilidades que tienen los estudiantes en los temas evaluados". Su aplicación y resultados, como un determinante que

genera la creación de estrategias para la mejora del nivel de desempeño educativo, permite "sustentar procesos efectivos y pertinentes de planeación educativa y políticas públicas" (SEP, 2010).

Socialmente, los resultados de la prueba ENLACE se han convertido en el principal indicador para determinar la calidad de la educación que ofrecen los diferentes planteles educativos evaluados; se les cataloga y de acuerdo a su desempeño se valora su labor educativa, lo cual genera un ambiente de tensión entre los agentes relacionados con su aplicación (profesores, autoridades educativas, gobernantes, etc.). En muchas ocasiones su efectividad ha sido cuestionada, pues se confunde el objetivo de brindar una educación de calidad con el de solo mejorar los resultados recurriendo a prácticas que pueden cuestionarse (Alanís Herrera et al., 2009: 22-32).

Ante este complejo escenario, y considerando que México es un país con una enorme inequidad cultural, social y económica, se pregunta sobre la manera en que estos factores, contextuales a la escuela, pueden llegar a afectar el desempeño de los estudiantes en estas evaluaciones. Ante esta incertidumbre, "los resultados de ENLACE 2009 incorporan el grado de marginación por localidad conforme a los índices del Consejo Nacional de Población (CONAPO), de manera que una escuela pueda compararse de forma más equitativa y justa con aquellas ubicadas en comunidades con niveles socioeconómicos similares" (SEP, 2010). Tal cuestionamiento lleva a plantear la hipótesis de que las entidades cuyos factores socioeconómicos son más desfavorables siempre obtendrán los peores resultados. Así mismo, si este planteamiento es acertado, o si puede ocurrir que en un ambiente negativo puedan existir resultados positivos y de ser así, qué es lo que influye y cómo puede compararse este caso con sus pares en contextos equivalentes. Concretamente, la pregunta es de qué forma pueden influir en el puntaje obtenido por los estudiantes, el grado de marginación de la escuela a la que asisten, el tipo de financiamiento que recibe (público o privado), el turno en el que prestan sus servicios y, a nivel estatal, el PIB per cápita de la entidad y el presupuesto que destina a la educación a través del Fondo de Aportaciones para la Educación Básica (FAEB).

Recientemente, se han llevado a cabo diversas investigaciones sobre la prueba ENLACE, algunas de ellas de carácter meramente cualitativo y otros empíricos. Las primeras comprende trabajos en los que se analizan los reactivos desde un enfoque didáctico para situar los alcances y limitaciones de este tipo de examen (Padilla Magaña, R, 2009: 2-10), o se entrevistan a los actores involucrados en el proceso para desvelar las implicaciones que ha tenido la prueba ENLACE en la práctica docente (Alanís Herrera et al., 2009: 22-32). Respecto a las investigaciones empíricas, sus resultados muestran que las escuelas privadas reportan un mejor desempeño que las escuelas públicas (Rentería Castro, 2010: 10-13) y que el contexto socioeconómico de los estudiantes influye de manera importante en los resultados de ENLACE (Amador Hernández, 2008:22-24). Por otro lado, se ha hecho un análisis sobre el recurso destinado para la educación en México y la forma en que se debería emplear dicho gasto (López Suárez et al., 2005). Sin embargo, no se han integrado todos estos elementos en un modelo que permita explicar la variabilidad en los resultados de ENLACE y que muestre, qué tanto de esa variación se atribuye a los factores de los centros escolares y cuánto a las diferencias entre las entidades federativas.

Ante esta situación se plantean dos propósitos en este artículo: uno didáctico, para el que se describirá el uso y utilidad de la modelación lineal multinivel para encontrar las relaciones que existen entre los diversos factores; esto es, se explicará cómo un grupo de variables al nivel de la escuela y otras a nivel del contexto en que se encuentran, influyen en el resultado de la variable respuesta, que es el desempeño promedio obtenido por la escuela. Para el objetivo que busca la explicación de los resultados, se realizó un ejercicio de aplicación con esta metodología estadística, utilizando los datos de los puntajes obtenidos en el año 2009 de la prueba ENLACE aplicada a una muestra de las escuelas de toda la República Mexicana. Esto permitió tener una mirada sobre los factores que influyen a nivel escolar y estatal en el desempeño de los centros escolares y con ello dar una mejor visión de qué rumbos están tomando las estrategias para mejorar la calidad de la educación básica en nuestro país.

Metodología

Como ya se anticipó, los factores de los cuales interesa conocer el efecto que tienen en el resultado de la prueba ENLACE alcanzado por las escuelas (Véase Figura 1), se definen por centro escolar: con información relativa al TURNO al que asisten los estudiantes: matutino, vespertino o nocturno; el TIPO de financiamiento que recibe, si es Público o Privado, así como el GRADO DE MARGINACIÓN[iv] de la localidad en la que se encuentra ubicado. Los datos que se utilizaron para el análisis fueron tomados de la página oficial de la SEP para los resultados de la prueba ENLACE 2009, misma que fue aplicada a 9,604,980 estudiantes de educación primaria de 98,869 centros escolares evaluados, prescindiendo de la información relativa a las escuelas de tipo indígena, CONAFE y educación secundaria. A partir de estos datos, se aplicó un muestreo estratificado con asignación proporcional (Vivanco, 2005: 88-89) en función del tipo de escuelas en cada entidad federativa, para contar con información representativa de la población. De esta manera, se obtuvo una muestra de 14,997 centros escolares en proporción aproximada de 9:1 (Véase Anexo 1); es decir, en la mayor parte de los estados, 9 de cada 10 escuelas son públicas y 1 privada, salvo en el Distrito Federal, donde escuelas particulares representan el 35%, y en los estados de Morelos y Oaxaca, donde representan el 22 y 24%, respectivamente.

La variable objeto de estudio es la media del puntaje (PUNTAJE) obtenido por el centro escolar en los tres grados evaluados de las diferentes asignaturas; en el caso de la aplicación 2009 se refiere a Español, Matemáticas y Formación Cívica y Ética. Un análisis preliminar de los datos a nivel nacional muestra que la media fue de 503.46 puntos, y que en el promedio por entidad, el Distrito Federal encabeza la lista de los mejores resultados, seguido por los estados de Sonora, Nuevo León y Baja California Sur, y que las entidades que registraron los promedios más bajos son Tabasco, San Luís Potosí y Guerrero (Figura 1).

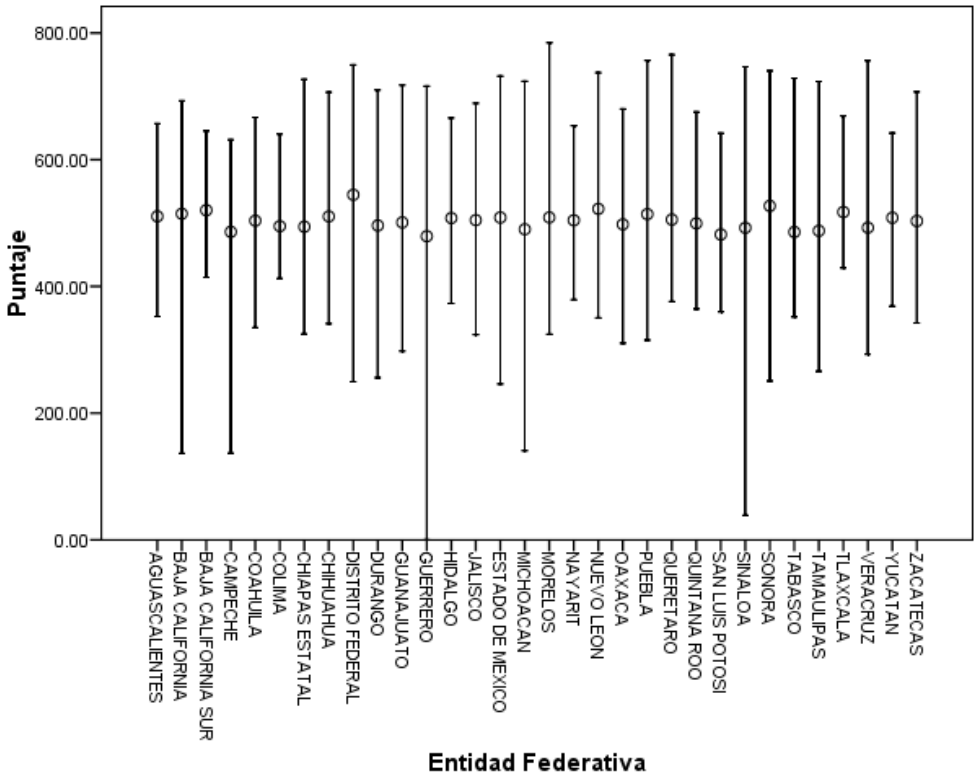


Figura 1. Resultados en la prueba de ENLACE 2009 por Entidad Federativa

Por otro lado, se observó que el 81% de las escuelas públicas prestan sus servicios en el turno matutino; tan solo un 17% lo hace en el turno vespertino y el resto en el turno de la noche. En el caso de las escuelas privadas, casi la totalidad (99%) lo hace en el turno matutino. Respecto al grado de marginación, se detectó que de los centros escolares evaluados en el país, el 40% están ubicados en localidades con bajo grado de marginación y el 27% en un alto grado, siendo los estados de Veracruz, Oaxaca y Guerrero los que cuentan con un mayor número de escuelas ubicadas en alta marginación, con el 20.8, 15.3 y 13%, respectivamente. En la figura 2 se aprecia que entre más alto es el grado de marginación de la población donde se encuentra ubicado el centro escolar, el promedio del puntaje obtenido en la prueba Enlace va disminuyendo y que este fenómeno se presentan tanto en escuelas públicas como privadas, observándose también que los resultados que obtienen las escuelas privadas son más altos que las escuelas públicas aunque se ubiquen en poblaciones con alto grado de marginación. Cabe mencionar, que Veracruz es el estado con mayor número de centros escolares de educación primaria en el país (7306), seguido por el Estado de México con 6136 escuelas.

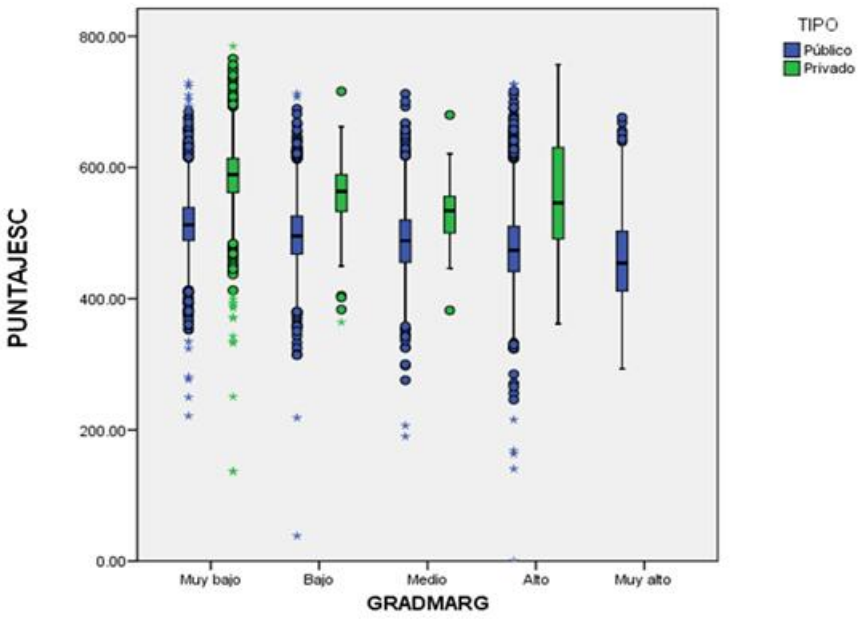


Figura 2. Resultados de la prueba Enlace por tipo de escuela y grado de marginación

Retomando el objetivo, se analizó si estos factores contribuyen a explicar las diferencias que existen entre las medias de los puntajes obtenidos por los centros escolares del país y si la variabilidad entre los estados es significativa, y si lo es, qué la provoca. Es decir, interesa estudiar si algunas variables a nivel estado, como el PIB Per Cápita de la entidad o el gasto que destina en educación a través del FAEB, influyen en los resultados de ENLACE. En la figura 3 y 4, se puede apreciar la distribución del PIB per capita y del FAEB por entidad federativa. Es importante resaltarse que Veracruz es el estado (después del Distrito Federal) que recibe una aportación más alta del Fondo para la Educación Básica, sin embargo su PIB Per cápita se encuentra dentro de los más bajos del país (48,240 pesos anuales) y el promedio del puntaje de la prueba Enlace de todos sus centros escolares, también lo colocan dentro de las últimas 8 entidades federativas. Por ello, resulta de suma importancia corroborar la influencia de estos factores socieconómicos en el desempeño de las escuelas del país.

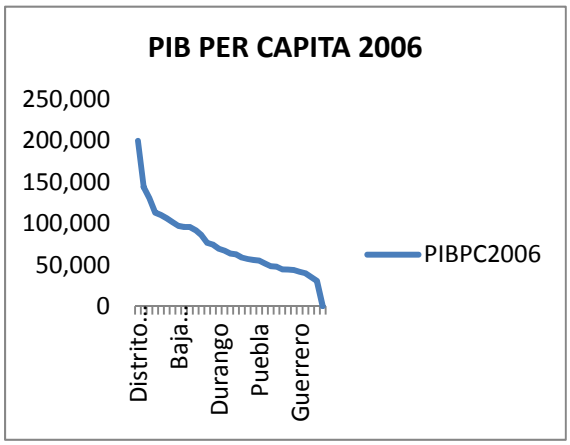


Figura 3. Distribución del PIB PER CAPITA por entidad federativa.

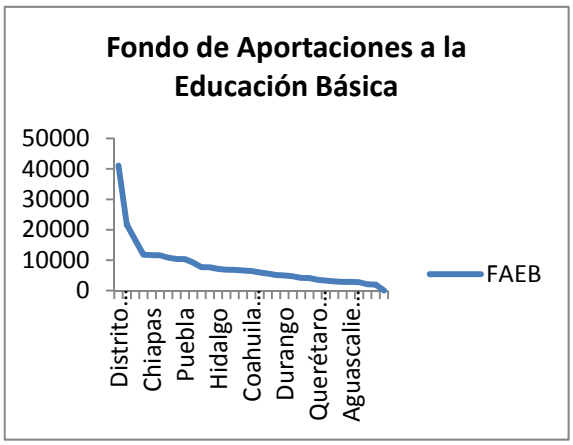


Figura 4. Distribución del FAEB por entidad federativa

Los datos con los que se contó para este estudio presentan una estructura jerárquica o de anidamiento; es decir, se tiene información del tipo, turno y grado de marginación de escuelas (nivel 1), que se encuentran agrupadas por entidad federativa (nivel 2). La metodología más indicada para analizar este tipo de datos es la modelación multinivel. Los modelos multinivel son diseñados para analizar variables de diferentes niveles simultáneamente, usando un modelo estadístico que apropiadamente incluye las diversas dependencias. La importancia de estos modelos radica en que se puede tener una mejor comprensión de la variabilidad de los datos, pues permite conocer la varianza entre las unidades de un mismo grupo o nivel y la covarianza entre los grupos o niveles, condición limitada en un análisis de regresión simple. Esta modelación de la varianza proporciona un marco más sólido que permite generar un amplio espectro de preguntas sobre el problema en cuestión, tales como los efectos contextuales. Por esta razón, ha sido aplicado y utilizado sobre todo en el área de Ciencias Sociales y particularmente en el tema de la Educación[v], pues el objetivo es investigar cómo un grupo de variables a nivel individual y grupal influyen en la variable respuesta, es decir, -en este caso- cómo el contexto afecta el rendimiento del centro escolar –que es aquí considerado en nivel individual-. En el área educativa se conoce como la "Teoría de la Rana en el Estanque" y hace referencia a la idea de que una rana en un estanque puede ser una pequeña rana en un estanque lleno de ranas grandes o una rana grande en un estanque de ranas pequeñas. Esta metáfora significa que los resultados individuales deben ser interpretados en relación a la media de su grupo.

El modelo lineal jerárquico o modelo multinivel busca estimar los parámetros desconocidos (intercepto y pendiente), pero además la varianza dentro de un grupo o nivel σ^2_e y la varianza entre los grupos o niveles σ^2_{u0}. El modelo se compone de una parte fija, en este caso con los parámetros que definen una línea promedio para todas las escuelas y de una parte aleatoria. La estimación de los coeficientes puede realizarse a través de diferentes enfoques como el de Máxima Verosimilitud o Estimación Bayesiana. Aplicado el modelo a este trabajo se pudieron conocer los efectos de las variables a nivel escuela y a nivel entidad federativa; en otras palabras, se contesta a: ¿qué tanto influyen las características de las entidades federativas en el logro de los resultados de ENLACE a nivel de centro escolar considerando el contexto de la escuela?

El modelo utilizado fue el modelo multinivel de intercepto aleatorio, lo que significa que el promedio del puntaje estimado en la prueba Enlace (PUNTAESC) de todas las escuelas, será el mismo para todas las entidades, es decir los parámetros estimados se consideran fijos para todas las escuelas de las entidades y comprendió dos niveles: el primero de ellos correspondió a las 14997 escuelas de la muestra y el segundo a las 32 entidades federativas, con sus respectivas variables de medición, como se muestra en la siguiente figura:

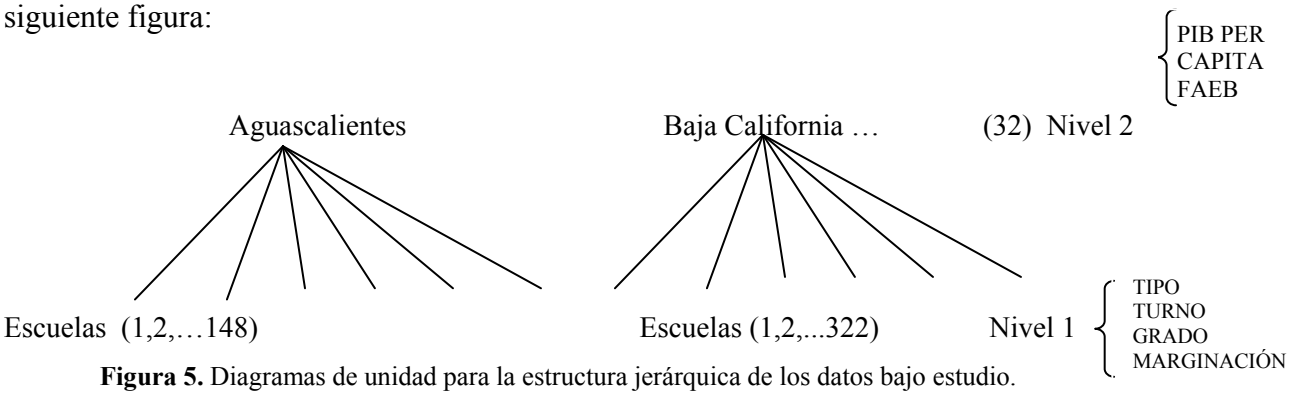


Figura 5. Diagramas de unidad para la estructura jerárquica de los datos bajo estudio.

La relación entre los parámetros a estimar quedó especificada de la siguiente manera:

$$PUNTAJE_{ij} = \beta_{0j} + \beta_1 PRIVADA_{ij} + \beta_2 TURNOVESP_{ij} + \beta_3 TURNONOC_{ij} + \beta_4 GMARG2_{ij} + \beta_5 GMARG3_{ij} + \beta_6 GMARG4_{ij} + \beta_7 GMARG5_{ij} + \beta_8 FAEB_j + \beta_9 PIBpc - 1_j + \beta_{10} PIBpc - 1_j + u_{0j} + e_{ij}$$

donde i= 1,2,…, 14997 j= 1,2,….,32

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$u_{0j} \sim N(0, \sigma^2_{u0})$$

$$e_{ij} \sim N(0, \sigma^2_e)$$

Resultados y discusión

Al estimar los parámetros del modelo[vi], se obtuvieron las estimaciones que se presentan en la Tabla 1. La media general del PUNTAJE obtenido en la prueba de ENLACE 2009 para los centros escolares públicos de todos los estados, que prestan sus servicios en el turno matutino y con un muy bajo grado de marginación fue de 527.08, con un error estándar de 8.136, resultando estadísticamente significativa.

Tabla 1. Resultado de las estimaciones para el modelo multinivel.

	Modelo intercepto aleatorio
Parámetros fijos	
β_{0j} (Intercepto)	527.08(8.13)
β_{1j} (Privada)	64.95 (1.48)
β_{2j} (Turnovesp)	-11.69 (1.118)
β_{3j} (Turnonoc)	-20.04 (1.44)
β_{5j} (GradMarg2)	-18.38 (1.301)
β_{6j} (GradMarg3)	-29.10 (1.44)
β_{7j} (GradMarg4)	-42.70 (1.250)
β_{8j} (GradMarg5)	-60.80 (2.231)
β_{9j} (FAEB)	-0.004 (0.632)
PIB (pc-1)	13.44 (7.619)
PIB (pc-2)	-7.549 (7.034)
Parámetros aleatorios	
Nivel 2 σ^2_{u0}	57.578
Nivel 1 σ^2_e	2511.130
-2*loglikelihood	160013.619

Asimismo, se observa que en promedio el resultado de las escuelas privadas fue más alto que las escuelas públicas, con una diferencia de media de 64.95 puntos. El puntaje disminuye en promedio, para las escuelas del turno vespertino y nocturno, por lo que se puede concluir que los centros escolares del turno matutino presentaron mejores resultados en la prueba de ENLACE. Respecto al grado de marginación, se deduce que entre más alto es el grado de marginación de la población en la que se encuentra ubicada la escuela, el promedio del puntaje obtenido irá disminuyendo hasta 60 puntos, como es el caso de las escuelas con muy alto grado de marginación.

Los resultados también muestran que las variables que se utilizaron a nivel estatal (nivel2) no resultaron significativas. Es decir, que los recursos destinados por cada entidad federativa, como Fondo de Aportación a la Educación Básica o el PIB per cápita de cada estado, no influyeron en los resultados de la prueba ENLACE a nivel escuela. Esto significa que la problemática de las escuelas con un alto grado de marginación es similar en cualquier estado de la república. Es decir, no influye si determinado estado invierte más en educación o económicamente se encuentra más desarrollado que otros, la realidad que aqueja a las escuelas en todo el país es muy parecida, y resulta declarada como homogénea en esta modelación. La explicación de las desigualdades entre las escuelas debe indagarse en las características de éstas, más que en factores relacionados a nivel estado.

En cuanto a la variabilidad explicada por el modelo, se observa en la Tabla 2 el análisis de los componentes de la varianza, donde se muestra que existe variabilidad entre los estados, es decir hay diferencias significativas en los resultados obtenidos en la prueba Enlace de una entidad a otra, así como también entre las escuelas de un mismo estado.

Con la información anterior, se obtuvo el Coeficiente de Correlación Intraclase, al dividir la variabilidad entre los estados con la variabilidad total. Este coeficiente mide el punto en el que el promedio que obtuvieron las escuelas en la prueba ENLACE dentro del mismo grupo (estado), se asemeja a otras escuelas de otros estados. El resultado obtenido fue de 0.0224, lo que significa que aproximadamente el 2% de la variabilidad de los datos es explicada por las diferencias entre los estados. Para validar los resultados del modelo, se comprobó el cumplimiento de los supuestos de normalidad de los residuos en todos los niveles, como se muestra en el gráfico siguiente:

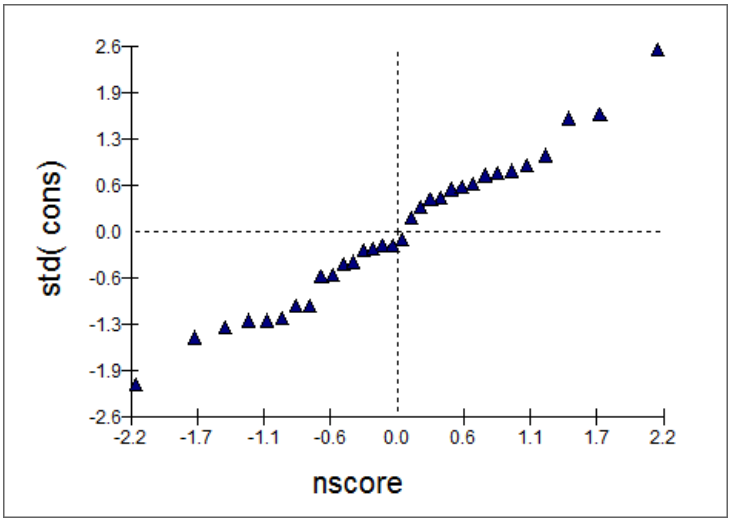


Figura 6. Gráfico de normalidad.

Conclusiones

La evaluación de la calidad educativa se ha constituido como un factor fundamental para el mejoramiento de la educación en México. Representa el motivo de debate entre docentes y grupo de docentes, pues es el principal instrumento que permite conocer la situación del país en materia educativa y determinar hacia donde deben dirigirse las acciones. Uno de los indicadores utilizados para su medición en nuestro país, ha sido a través de la aplicación de pruebas internacionales como PISA y nacionales como ENLACE, que son evaluaciones estandarizadas que aportan elementos que permiten obtener información contundente del panorama educativo en México. Al hacer un análisis particular de los resultados obtenidos por los centros escolares en la prueba ENLACE del año 2009, con la finalidad de detectar los factores que influyen en estos resultados, se encontró que el desempeño de las escuelas privadas es más alto que el de las escuelas públicas, con una diferencia de 65 puntos en promedio, lo que corrobora la hipótesis de Rentería Castro (2009) pues en este tipo de centros escolares se logra un mejor desempeño en el aprendizaje por la posibilidad que tienen los estudiantes de cursar materias extracurriculares y por tanto, recibir una formación más integral. Así mismo, se obtuvieron como variables significativas, el turno en el que la escuela presta sus servicios, siendo los centros escolares matutinos los que registraron mejores resultados y el grado de marginación, cuya incidencia es negativa, pues a mayor

grado de marginación de la población en la que se encuentra el centro escolar, los resultados obtenidos serán más bajos en promedio, coincidiendo con las conclusiones de Amador Hernández (2008). Comparando los resultados a nivel nacional, se observó que el Distrito Federal y las entidades de Sonora y Nuevo León encabezan la lista con los puntajes más altos y que los estados con una media de desempeño estadísticamente inferior son Tabasco, San Luis Potosí y Guerrero. Al buscar ubicar estos resultados en el contexto de las entidades federativas, utilizando variables como el PIB per cápita de los estados y el presupuesto que destinan a educación a través del FAEB, se aplicó un modelo multinivel, herramienta estadística que permite analizar las variables a nivel escuela y a nivel estado conjuntamente y medir qué porcentaje de la variabilidad de los resultados es explicada en este caso por los factores relacionados con la escuela y qué porción por las características de las entidades. El modelo arrojó como resultado que las variables a nivel estado no contribuyen a explicar las diferencias en los resultados de ENLACE de los centros escolares analizados. Es decir, que las escuelas catalogadas con muy alto grado de marginación presentan resultados homogéneos de bajo aprovechamiento en todas las entidades del país. La problemática que aqueja a las escuelas del país es muy similar, no teniendo una influencia estadísticamente significativa si una entidad tiene un ingreso per cápita por encima o por debajo de la media nacional o si el FAEB es más alto o menor.

Este hallazgo conduce a que los actores que intervienen en el sistema educativo, replanteen el enfoque encaminado al mejoramiento de la política educativa y a que se redoblen esfuerzos para atenuar la realidad que aqueja a las escuelas mexicanas.

Anexo 1

Proporción de la muestra de las escuelas por entidad federativa en función del tipo de centro escolar.

	TIPO			
	Pública		Privada	
	Recuento	%	Recuento	%
Aguascalientes	127	85.8%	21	14.2%
Baja California	269	83.5%	53	16.5%
Baja California Sur	68	87.2%	10	12.8%
Campeche	123	92.5%	10	7.5%
Chiapas Estatal	722	96.9%	23	3.1%
Chihuahua	413	90.2%	45	9.8%
Coahuila	335	89.8%	38	10.2%
Colima	83	91.2%	8	8.8%
Distrito Federal	461	65.1%	247	34.9%
Durango	385	96.3%	15	3.8%
Estado de México	1310	85.9%	215	14.1%
Guanajuato	844	92.1%	72	7.9%
Guerrero	429	94.9%	23	5.1%
Hidalgo	409	90.5%	43	9.5%
Jalisco	1028	90.1%	113	9.9%
Michoacán	366	85.9%	60	14.1%
Morelos	152	76.4%	47	23.6%
Nayarit	169	93.9%	11	6.1%
Nuevo León	487	89.5%	57	10.5%
Oaxaca	68	78.2%	19	21.8%
Puebla	535	84.3%	100	15.7%
Querétaro	208	84.9%	37	15.1%
Quintana Roo	120	83.9%	23	16.1%
San Luis Potosí	480	94.5%	28	5.5%
Sinaloa	460	94.1%	29	5.9%
Sonora	308	89.0%	38	11.0%
Tabasco	370	94.9%	20	5.1%
Tamaulipas	439	90.3%	47	9.7%
Tlaxcala	118	83.7%	23	16.3%
Veracruz	1560	95.6%	72	4.4%
Yucatán	198	87.6%	28	12.4%
Zacatecas	363	96.0%	15	4.0%

Fuente: Elaboración propia a partir de los datos de la muestra.

Anexo 2

Descripción de los factores a nivel escolar.

Nombre de la variable	Descripción	
TURNO	Horario al que asisten los estudiantes	1 Matutino 2 Vespertino 3 Nocturno
TIPO	Financiamiento que recibe la escuela	1 Público 2 Privado
GRADO DE MARGINACIÓN	Índice de clasificación de las localidades del país, según el impacto global de las privaciones que padece la población, las cuales limitan el pleno desarrollo de las personas.	1 Muy Bajo 2 Bajo 3 medio 4 Alto 5 Muy alto

Descripción de los factores a nivel estatal.

Nombre de la variable	Descripción	
PIB per cápita*	Es la relación entre el valor total del mercado de todos los bienes y servicios generados por la economía, en este caso, de una entidad federativa. Se utiliza para expresar su potencial económico.	**-2** Estados con un PIB per cápita muy bajo (2 desviaciones estándar de la media) **-1** Estados con un PIB per cápita con una desviación estándar por debajo de la media. **1** Estados con PIB per cápita a una desviación estándar por encima de la media. **2** Estados con un PIB per cápita muy alto. (2 desviaciones estándar arriba de la media).
FAEB	Fondo de Aportaciones para la Educación Básica	

Notas

[i] Para ahondar en los resultados de PISA, véase Información sobre México en PISA. Instituto Nacional para la Evaluación en la Educación.

[ii] El Consejo Nacional de Fomento Educativo (CONAFE) es un organismo descentralizado, de la Administración Pública Federal, con personalidad jurídica y patrimonio propios, creado por decreto presidencial del 11 de septiembre de 1971, con el objeto de allegarse recursos complementarios, económicos y técnicos, nacionales o extranjeros para aplicarlos al mejor desarrollo de la educación en el país, así como a la difusión de la cultura mexicana en el exterior.

[iii] En la aplicación de ENLACE 2009 se amplió la cobertura a alumnos de primero, segundo y tercer grado de secundaria. También se redefinió el enfoque de la evaluación a los alumnos del tercer grado de secundaria, al dejar de evaluar el contenido de los tres grados, para únicamente evaluar el currículum del grado.

[iv] Esta información la publica en su base de datos de la Secretaría de Educación Pública (SEP) conforme a los índices que elabora el Consejo Nacional de Población.

v Importantes desarrollos en modelación multinivel se iniciaron en el área de Educación. Véase Aitkin et. Al (1981) y Aitkin y Longford (1986).

[vi] Se utilizó como apoyo el software especializado en modelación multinivel para windows Mlwin.

Referencias

Aitkin, M. & Longford, N. (1986). Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society*, Series A: 149:1-43.

Alanís, J.P.; Ávila, F., Verónica y Elida Lerma Reyes (2009). "Las implicaciones de la prueba ENLACE en educación primaria y su relación con el contexto socioeconómico". *Revista Electrónica de la Red Durango de Investigadores Educativos A.C (México) Praxis Investigativa*, vol.1, año 1, pp. 22-33 (en línea). Disponible en: http://www.redie.org/librosyrevistas/redieinv01.pdf

Amador, J. C. (2008). "La evaluación y el diseño de políticas educativas en México", *Centro de Estudios Sociales y de Opinión Pública*, Documento de trabajo núm. 35, marzo de 2008, pp 1-44.

Bryk, A.S. & Raudenbush, S.W(1992). Hierarchichal Linear Models. Newbury Park, California: Sage.

Goldstein, H. (1999). *Multilevel Statistical Models*. London. First Internet edition.

Longford, N.T. (1993). *Random Coefficient models*. Oxford. Clarendon Press.

López, A.; Morales H., y Beltrán, Elvia E. (2005). El sostenimiento de la educación en México. *Papeles de población*. Red AlyC. (44) p.p.239-254 (en línea). Disponible en: http://redalyc.uaemex.mx/pdf/112/11204410.pdf.

Padilla, R. A. (2009). La prueba ENLACE desde un análisis didáctico. Más allá que una política de calidad para la educación básica. X Congreso Nacional de Investigación Educativa (en línea). Disponible en: http://www.comie.org.mx/congreso/memoria/v10/pdf/area_tematica_02/ponencias/1235-F.pdf

Pescador, J.A (1992). "Acuerdo nacional para la modernización de la educación básica. Una visión integral," en *El Cotidiano*, núm. 61. México, UAM.

Rentería, E. (2010). "La evaluación del desempeño escolar y la política educativa en México", *Revista Iberoamericana de Educación*, año 2, núm. 54, pp. 1-13 (en línea). Disponible en: http://www.rieoei.org/deloslectores/3791Renteria.pdf

SEP (2010). Secretaría de Educación Pública. www.enlace.sep.gob.mx.

Vivanco, M. (2005). "*Muestreo Estadístico: Diseño y Aplicaciones*". Editorial Universitaria (Chile).