UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLAN

ANALISIS DISCRIMINANTE
UNA APLICACIÓN AL NIVEL DE INGRESO DE LOS HOGARES

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A

CAROLE ODETTE SCHMITZ BASAÑEZ

ASESOR: FIS. MAT. JORGE LUIS SUAREZ MADARIAGA

AGOSTO 2006

« Les hommes se distinguent
par ce qu'ils montrent
et se ressemblent
par ce qu'ils cachent »

Paul Valéry

A mis Padres, Patrice y Amanda
A mi hermana Juliette
Gracias por todo lo que me han dado,
especialmente por todo su amor.
Los Adoro

Carole

Reconocimientos

Fis. Mat. Jorge Luis Suárez Madariaga, gracias por su tiempo e interés para la realización de este trabajo.

A mis sinodales: M. en C Víctor Ulloa Arellano, Act. Miguel Ángel Sánchez Barquín, Act. Alan Hernández Estrada y especialmente al Act. Mahil Herrera Maldonado, por el tiempo dedicado a la revisión de este trabajo y sus valiosas aportaciones.

A mi universidad por permitirme ser parte de ella y darme la formación que tengo, por las experiencias inolvidables y las extraordinarias personas que en ella conocí.

A mis maestros por compartir sus conocimientos y experiencias

Agradecimientos

Agradezco a todas las personas que me han brindado su amistad y su cariño, sobra decir que no me alcanzan las palabras para decir cuanto los quiero:

A Margot por su empuje para llegar a esta meta (¡Lo logramos!), por su compresión y cercanía, por las risas y las lágrimas compartidas, por ser una persona que vale oro (por poner un parámetro formal…), simplemente por ser mi gran amiga Maggi

A Carol, mi hermana por elección, por ser mi incansable compañera en esta travesía que es la vida, contagiándome su alegría y brindándome su amistad y cariño.

A Alain por ser una parte muy importante de mi vida, por su fortaleza y cariño que me sostuvieron en los momentos más difíciles. Por su compañía, su apoyo, sus risas, su inteligencia…. en resumen por ser Alain, pero sobre todo por haberme hecho tan feliz. "….Quizás", TAMHUB

A Oswaldo Palma, por ser un gran maestro, un buen jefe, pero sobre todo un excelente amigo. Gracias por todo lo enseñado y por todos los momentos compartidos.

A Isaac Sidhartha Salcedo Campos (Sr. Stich), por su solidaridad, camaradería y apoyo incondicional, pero sobre todo por su valiosísima amistad.

A Elsa Resano, por sus grandes enseñanzas, por las oportunidades y el cariño que nos ha brindado, por ser una gran mujer: "La Jefa"

A mis amigos Alan, Bere, Christian, Ivett, Mike, Pepe, Sonia, Wally…por los momentos compartidos, su cariño y su amistad.

A todos aquellos que el destino se ha llevado al viejo continente o a tierras insospechablemente más lejanas pero que siempre permanecerán en mi memoria y en mi corazón

Contenido

Introducción

Los problemas a los que se enfrentan actualmente los investigadores en todas las disciplinas resultan en su mayoría complejos y rara vez dependen de un factor único. La información recolectada para el análisis de estos problemas por lo general se compone de múltiples factores y se trata por medio de alguna técnica de análisis multivariado, que es la rama de la estadística dedicada al estudio de distribuciones multivariadas y sus muestras.

Independientemente del área de investigación, es frecuente enfrentarse a problemas de clasificación o diferenciación de grupos de individuos u objetos, cuya solución permitirá un mejor conocimiento de la realidad.

El análisis discriminante es una técnica estadística multivariada cuyo principal objetivo es la diferenciación de dos o más grupos de individuos, objetos etc.; y que hoy en día es aplicada en diferentes ámbitos como pueden ser la biología, la medicina, la psicología, las finanzas o la educación. Esta técnica multivariada tiene dos grandes vertientes: el análisis discriminante descriptivo, que permite distinguir las características que diferencian (discriminan) a los grupos; y el análisis discriminante predictivo, que, como su nombre lo indica permite obtener una función capaz de asignar individuos nuevos al grupo al que pertenecen con mayor probabilidad. El objetivo del presente trabajo es realizar una aplicación del análisis discriminante a un problema real, que en este caso se eligió fuera el nivel de ingreso en los hogares, presentando cada uno de los pasos necesarios para llevarla a cabo, así como la teoría que fundamenta este tipo de análisis.

El nivel de ingreso y su distribución dentro de la población están fuertemente asociados con problemas económicos y sociales, en diversas áreas como la educación, la salud etc., que actualmente son prioritarios no sólo en México sino a nivel mundial. Por ello, el estudio del nivel de ingreso resulta importante a la hora de querer afrontar dichos problemas, comenzando por su diagnostico, por medio del diseño y aplicación de instrumentos políticos como pueden ser programas de apoyo social o bien para realizar una evaluación de las políticas aplicadas.

Si bien el ingreso es un factor importante en muchos de los problemas socioeconómicos actuales, su medición resulta inexacta o errónea en la mayoría de los casos ya que es una variable difícil de captar debido a la renuencia de las personas por declarar su nivel de ingreso.

Por ello, se decidió aplicar el análisis discriminante al nivel de ingreso de los hogares en la República Mexicana de manera que, al definir grupos correspondientes a distintos niveles de ingreso, se puedan identificar los factores que diferencian estos grupos y que los hogares cuyo nivel de ingreso sea desconocido puedan ser adscritos a alguno de estos grupos De obtener resultados satisfactorios al aplicar la regla de clasificación, esto permitiría identificar el nivel de ingreso de los hogares y utilizar esta información como estimador en estudios en los que no se cuente con la información del nivel de ingreso o ésta sea inverosímil o incorrecta.

El presente trabajo busca entonces probar que por medio del análisis discriminante es posible identificar y estimar el impacto de los factores socioeconómicos que influyen en el nivel de ingreso de los hogares, así como generar una regla para la clasificación de nuevos hogares en grupos previamente definidos, a partir de características de ingreso fijo de los hogares y algunas características socioeconómicas de sus miembros.

Para la realización de la aplicación que se presenta en este trabajo, se utilizará información estadística a nivel nacional proveniente de la Encuesta Nacional de Ingreso y Gasto de los Hogares 2002 (ENIGH 2002), realizada por el Instituto Nacional de Estadística Geografía e Informática (INEGI). Se aplicarán las técnicas de análisis discriminante con la ayuda del paquete estadístico SPSS en su versión 13.

Dado que el análisis discriminante es una técnica multivariada, es necesario tener el conocimiento de varios conceptos de álgebra matricial para poder expresar y manejar información de muchas variables, por ello, en el primer capítulo de este trabajo se presentan brevemente los elementos de álgebra matricial necesarios para la comprensión y aplicación del análisis discriminante.

Del mismo modo, el segundo capítulo de este trabajo se avoca a la presentación de los conceptos básicos del análisis multivariado, necesarios para abordar la teoría del análisis discriminante. En este capítulo se presentarán los datos y algunas distribuciones multivariadas y las estadísticas de uso más frecuente en análisis multivariado, incluyendo brevemente conceptos básicos de análisis univariado cuyo entendimiento facilita la comprensión de los conceptos de análisis multivariado.

Una vez sentadas las bases del análisis multivariado, es posible abordar la teoría del análisis discriminante que se presenta en el capítulo 3. En éste capítulo, se revisarán los objetivos de está técnica multivariada, los supuestos que han de cumplirse para poder aplicarla, se definirán las funciones discriminantes y se presentará el modo en el que estas últimas se obtienen. También se presentarán diversas reglas de clasificación, los procedimientos para la selección de variables y finalmente la interpretación de los resultados.

En el capítulo 4, se presenta la aplicación del análisis discriminante al nivel de ingreso de los hogares, así como los resultados obtenidos por medio de dicha aplicación. En este capítulo se presentarán los detalles de la aplicación al nivel de ingreso de hogares, así como de la información que se empleará para la realización de ésta. También se detallarán cada uno de los pasos de la aplicación tales como la comprobación de los supuestos, la obtención de las funciones discriminantes, su interpretación y finalmente el resultado obtenido por medio de esta aplicación en cuanto a la clasificación de los individuos en los grupos predeterminados.

Aunado a esto, se incluyen dos apéndices. En el apéndice A se presenta brevemente la metodología de la ENIGH 2002; en el apéndice B se incluyen dos tablas con las principales distribuciones de probabilidad, sus funciones de densidad, sus parámetros, su media y su varianza.

Capítulo 1

Álgebra Matricial

El manejo de datos multivariados y el entendimiento de muchas de las técnicas de análisis multivariado, como el análisis discriminante, requiere del conocimiento de varios conceptos de álgebra matricial ya que estos permiten expresar y manejar de manera clara y sencilla la información de muchas variables.

En este capítulo se presentan brevemente los elementos de álgebra matricial necesarios para comprender y aplicar la teoría del análisis discriminante.

Las herramientas que nos brindan los paquetes estadísticos, como SPSS, facilitan en gran medida los cálculos necesarios para la aplicación de técnicas multivariadas, sin embargo, es necesario tener claros los conceptos de álgebra matricial para poder comprender e interpretar adecuadamente los resultados obtenidos.

1.1 Definiciones y terminología básica

1.1.1 Definiciones básicas

Definición 1 ***Vector*** *Un* ***vector*** *de n componentes es un conjunto ordenado de n números. De esta manera, definimos un* ***vector renglón*** *de n componentes como un conjunto ordenado de n números escritos de manera horizontal, análogamente, definimos un* ***vector columna*** *de n componentes como un conjunto ordenado de n números escritos de manera vertical.*

A lo largo del trabajo los vectores se denotarán por letras minúsculas negritas, los vectores columna se denotarán entonces por $\mathbf{u}, \mathbf{v}, \mathbf{a}, \mathbf{b}, \mathbf{c}$..., mientras que los vectores renglón se denotarán por $\mathbf{u}', \mathbf{v}', \mathbf{a}', \mathbf{b}', \mathbf{c}'$....

Definición 2 ***Matriz*** *Una* ***matriz*** $\mathbf{A}$ *de* $m \times n$ *es un arreglo rectangular de* mn *números ordenados en* m *renglones y* n *columnas.*

$$\mathbf{A}_{(m\times n)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

Las matrices se representarán en este trabajo con letras mayúsculas negritas $\mathbf{A}, \mathbf{B}, \mathbf{X}, ...$

El vector renglón $\mathbf{a'_{i(1\times n)}} = (a_{i1,}a_{i2,}..., a_{in})$ se conoce como el renglón i de la matriz $\mathbf{A}$ y el vector columna $\mathbf{a_{j(m\times 1)}} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$ se conoce como la j-ésima columna de la matriz $\mathbf{A}$.

Cualquier elemento a_{ij} de la matriz $\mathbf{A}$ puede identificarse gracias a su subíndice donde i denota el número de renglón y j el número de columna.

Una matriz $\mathbf{A}_{(m\times n)}$ puede verse también como un conjunto de m vectores renglón o un conjunto de n vectores columna. Si interpretamos un vector columna como m observaciones de una variable, la matriz $\mathbf{A}$ puede verse como m observaciones de n variables.

Definición 3 ***Igualdad de Matrices*** *Sean* $\mathbf{A}$ *y* $\mathbf{B}$ *dos matrices,* $\mathbf{A} = \mathbf{B}$ *si y sólo si* $\mathbf{A}$ *y* $\mathbf{B}$ *tienen el mismo tamaño y además cada elemento* a_{ij} *de* $\mathbf{A}$ *es igual al correspondiente* b_{ij} *de* $\mathbf{B}$*, esto es:*

$$\mathbf{A} = \mathbf{B} \text{ si y sólo si } a_{ij} = b_{ij} \ \forall i, j$$

1.1.2 Algunas Matrices Especiales

Definición 4 ***Matriz cuadrada*** *Sea* $\mathbf{A}$ *una matriz de* $m \times n$*, si* $m = n$ *entonces* A *se conoce como una* ***matriz cuadrada*** *de tamaño* n*.*

Definición 5 ***Transpuesta de una matriz*** *Sea* $\mathbf{A}$ *una matriz de* $m \times n$*, al intercambiar los renglones por las columnas de* $\mathbf{A}$*, se obtiene una matriz* $\mathbf{A}'_{(n\times m)}$ *llamada* ***matriz transpuesta****.*

$$\mathbf{A}_{(m\times n)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

$$\mathbf{A}'_{(n\times m)} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{i1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{i2} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{1j} & a_{2j} & \cdots & a_{ij} & \cdots & a_{mj} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{in} & \cdots & a_{mn} \end{bmatrix}$$

La transpuesta de la transpuesta de una matriz $\mathbf{A}$ *es igual a la misma matriz* $\mathbf{A}$*.*

$$(\mathbf{A}')' = \mathbf{A}$$

Definición 6 ***Matriz Simétrica*** *Sea* $\mathbf{A}$ *una matriz de* $n \times n$*, se dice que* $\mathbf{A}$ *es* ***simétrica*** *si y sólo si* $a_{ij} = a_{ji}\ \forall i, j$ *es decir,* $\mathbf{A}$ *es simétrica si y sólo si* $\mathbf{A} = \mathbf{A}'$*.*

Definición 7 ***Matriz Diagonal*** *Una matriz* $\mathbf{D}$ *cuadrada es llamada* ***diagonal*** *cuando los únicos elementos distintos de cero que tiene aparecen en la diagonal principal.*

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & d_{ii} & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & d_{nn} \end{bmatrix}$$

Definición 8 ***Matriz Identidad*** *Se conoce como* ***matriz identidad*** *a aquella matriz* $\mathbf{I}_{(n\times n)}$ *diagonal cuyos elementos de la diagonal principal son iguales a* 1*.*

$$\mathbf{I}_{(n\times n)} = (b_{ij}) donde b_{ij} = \begin{cases} 1 \ si\ i = j \\ 0 \ si\ i \neq j \end{cases}$$

1.2 Operaciones Matriciales

1.2.1 Suma de matrices

La suma de matrices está definida siempre y cuando las matrices sean del mismo tamaño. Sean $\mathbf{A}$ y $\mathbf{B}$ dos matrices de $m \times n$, entonces $\mathbf{C} = \mathbf{A} + \mathbf{B}$ es también una matriz de $m \times n$ y está dada por:

$$\mathbf{A}_{(m\times n)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \quad \mathbf{B}_{(m\times n)} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{i1} & b_{i2} & \cdots & b_{ij} & \cdots & b_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mj} & \cdots & b_{mn} \end{bmatrix}$$

$$\begin{aligned} \mathbf{C}_{(m\times n)} &= \mathbf{A} + \mathbf{B} \\ \mathbf{C}_{(m\times n)} &= \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1j}+b_{1j} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2j}+b_{2j} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1}+b_{i1} & a_{i2}+b_{i2} & \cdots & a_{ij}+b_{ij} & \cdots & a_{in}+a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mj}+b_{mj} & \cdots & a_{mn}+b_{mn} \end{bmatrix} \end{aligned}$$

Es decir $\mathbf{C}$ se obtiene al sumar las componentes correspondientes de A y $\mathbf{B}$.

$$\begin{aligned} \mathbf{C} &= \mathbf{A} + \mathbf{B} \\ \mathbf{C} &= (a_{ij} + b_{ij}) \forall i, j \end{aligned}$$

Esto se hace extensivo para la resta de matrices

$$\begin{aligned} \mathbf{D} &= \mathbf{A} - \mathbf{B} \\ \mathbf{D} &= (a_{ij} - b_{ij}) \ \forall i, j \end{aligned}$$

Algunas Propiedades de la suma de matrices:

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \mathbf{B} + \mathbf{A} \\ (\mathbf{A} + \mathbf{B}) + \mathbf{C} &= \mathbf{A} + (\mathbf{B} + \mathbf{C}) \\ (\mathbf{A} + \mathbf{B})' &= (\mathbf{A}' + \mathbf{B}') \end{aligned}$$

1.2.2 Producto de Matrices

Multiplicación de una matriz por un escalar

Sea $\mathbf{A} = (a_{ij})$ una matriz de $m \times n$ y α un escalar, definimos entonces $\alpha\mathbf{A}$ como una matriz de $m \times n$ cuyos componentes son el resultado de la multiplicación de cada elemento a_{ij} de $\mathbf{A}$ por α, es decir $\alpha\mathbf{A} = (\alpha a_{ij})$

$$\text{Sea} \quad \mathbf{A}_{(m\times n)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \quad \text{y} \quad \alpha \quad \text{un escalar, entonces}$$

$$\alpha A_{(m\times n)} = (\alpha a_{ij}) = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1j} & \cdots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2j} & \cdots & \alpha a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ \alpha a_{i1} & \alpha a_{i2} & \cdots & \alpha a_{ij} & \cdots & \alpha a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ \alpha a_{m1} & \alpha a_{m2} & \cdots & \alpha a_{mj} & \cdots & \alpha a_{mn} \end{bmatrix}$$

Algunas Propiedades del producto de una matriz por un escalar:

$$\begin{aligned} \alpha(A + B) &= \alpha A + \alpha B \\ (\alpha + \beta)A &= \alpha A + \beta A \ (\alpha, \beta \text{ escalares}) \end{aligned}$$

Producto escalar o producto interno

Sean $\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ y $\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ dos vectores de $n \times 1$, el *producto escalar o producto interno* de $\mathbf{a}$ y $\mathbf{b}$ esta entonces dado por:

$$\mathbf{a}'\cdot\mathbf{b} = a_1b_1 + ... + a_nb_n$$

El producto interno está definido solamente para vectores con el mismo número de componentes y el resultado de éste es un escalar.

Si en cambio realizamos el producto $\mathbf{a} \cdot \mathbf{b}'$ obtendremos una matriz de $n \times n$

$$\mathbf{a} \cdot \mathbf{b}' = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{bmatrix}$$

Algunas propiedades del producto interno:

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= \mathbf{b} \cdot \mathbf{a} \\ \mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} \\ (\alpha\mathbf{a}) \cdot \mathbf{b} &= \alpha(\mathbf{a} \cdot \mathbf{b}) \\ \mathbf{a}' \cdot \mathbf{a} &= a_1^2 + a_2^2 + ... + a_n^2 \end{aligned}$$

La raíz cuadrada de $\mathbf{a}' \cdot \mathbf{a}$ se conoce como *norma* de $\mathbf{a}$ y se denota por:

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}' \cdot \mathbf{a}}$$

Se dice que un vector está *normalizado* si $\mathbf{a}' \cdot \mathbf{a} = 1$, cabe señalar que un vector siempre puede normalizarse dividiendo cada una de sus componentes por la norma del vector.

Producto de dos matrices

Sean $\mathbf{A}_{(m \times n)}$ y $\mathbf{B}_{(n \times p)}$ dos matrices, el producto de $\mathbf{A}$ y $\mathbf{B}$ es una matriz $\mathbf{C}_{(m \times p)}$, donde cada elemento (c_{ij}) de $\mathbf{AB}$ es el resultado del producto interno del renglón i de $\mathbf{A}$ por la columna j de $\mathbf{B}$.

$$\begin{aligned} c_{ij} &= a_{i1}b_{i1} + a_{i2}b_{2j} + ... + a_{in}b_{nj} \\ c_{ij} &= \sum_k a_{ik}b_{kj} \end{aligned}$$

La multiplicación entre dos matrices está definida únicamente cuando el número de columnas de la primera sea igual al número de renglones de la segunda.

Por lo general el producto de matrices no es conmutativo es decir

$$\mathbf{AB} \neq \mathbf{BA}$$

de hecho, que el producto **AB** esté definido no implica que lo esté **BA**

Algunas propiedades del producto de matrices

$$\begin{aligned} \mathbf{AI} &= \mathbf{A} \\ \mathbf{A}(\mathbf{B}+\mathbf{C}) &= \mathbf{AB}+\mathbf{AC} \\ (\mathbf{A}+\mathbf{B})\mathbf{C} &= \mathbf{AC}+\mathbf{BC} \\ \mathbf{A}(\mathbf{BC}) &= (\mathbf{AB})\mathbf{C} \\ (\mathbf{AB})' &= \mathbf{B}'\mathbf{A}' \end{aligned}$$

Producto de una matriz por un vector

El producto de una matriz $\mathbf{A}_{(m\times n)}$ y un vector $\mathbf{b}_{(n\times 1)}$ está definido solamente cuando el número de columnas de **A** es igual al número de componentes de **b**, y está dado por:

$$\mathbf{c} = \mathbf{Ab}$$

que es equivalente a

$$\mathbf{c} = b_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{m1} \end{bmatrix} + b_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{m2} \end{bmatrix} + ... + b_i \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{mj} \end{bmatrix} + ... + b_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{in} \\ \vdots \\ a_{mn} \end{bmatrix}$$

El lado derecho de esta igualdad se conoce como *combinación lineal* de las columnas de la matriz **A**.

Vectores y Matrices ortogonales

Definición 9 ***Vectores ortogonales*** *Sean* **a** *y* **b** *dos vectores de* $n \times 1$ *si* $\mathbf{a}'\mathbf{b} = 0$**,** *entonces* **a** *y* **b** *son* ***ortogonales***

Definición 10 ***Matriz ortogonal*** *Sea* $\mathbf{A}_{(n\times n)}$*, si todas las columnas de* $\mathbf{A}$ *son mutuamente ortogonales y están normalizadas, entonces* $\mathbf{A}$ *es una* ***matriz ortogonal*** *y* $\mathbf{A}'\mathbf{A} = \mathbf{I}$.

Forma Cuadrática

Sea v un vector de $n \times 1$ y $A_{n\times n}$ una matriz simétrica, entonces el producto

$$\mathbf{v}'\mathbf{A}\mathbf{v}$$

conocido como *forma cuadrática* está dado por:

$$\begin{aligned}\mathbf{v}'\mathbf{A}\mathbf{v} &= \sum_i v_i^2 a_{ii} + \sum_{i\neq j} v_i v_j a_{ij} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j a_{ij}\end{aligned}$$

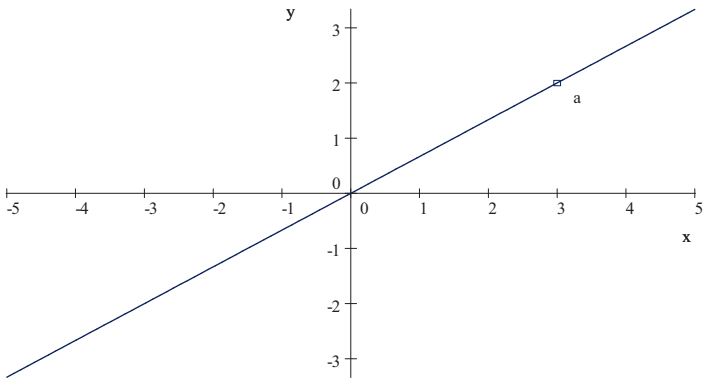
1.3 Geometría Matricial

El álgebra matricial es una herramienta muy útil para la formulación y desarrollo de conceptos matemáticos. La interpretación geométrica de los vectores y matrices, presentada en esta sección, suele ser de gran ayuda para la comprensión de los resultados obtenidos.

1.3.1 Espacio Vectorial

Sea $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$ un vector de $n\times 1$, los elementos de $\mathbf{v}$ pueden ser vistos como las coordenadas de un punto en un espacio n-dimensional o bien como el segmento que conecta al origen con este punto.

Por ejemplo en $\mathbb{R}^2$, si tenemos un vector $\mathbf{a} = \binom{3}{2}$, este puede ser visto como el punto $\mathbf{a}$ cuyas coordenadas son $x = 3$ y $y = 2$, o bien como el segmento que conecta el origen con el punto $\mathbf{a}$. Cualquier múltiplo escalar de $\mathbf{a}$ es un segmento de la línea que conecta el origen con el punto $\mathbf{a}$. La norma del vector $\mathbf{a}$, $\|\mathbf{a}\|$ es la distancia que hay entre el origen y el punto $\mathbf{a}$.



De esta manera llamaremos *espacio vectorial* a aquel conjunto de vectores V que cumplan con los siguientes axiomas:

1. Cerradura bajo la suma
$$\text{Sean } \mathbf{x}, \mathbf{y} \in V \Longrightarrow \mathbf{x} + \mathbf{y} \in V$$

2. Ley asociativa de la suma de vectores
$$\text{Sean } \mathbf{x}, \mathbf{y}, \mathbf{z} \in V \Longrightarrow (\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$$

3. Ley conmutativa de la suma de vectores
$$\text{Sean } \mathbf{x}, \mathbf{y} \in V \Longrightarrow \mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$$

4. Idéntico aditivo o vector cero
$$\begin{aligned} \exists \text{ un vector } \mathbf{0} &\in V \text{ tal que } \forall \mathbf{x} \in V \\ \mathbf{0} + \mathbf{x} &= \mathbf{x} + \mathbf{0} = \mathbf{x} \end{aligned}$$

5. Inverso aditivo
$$\text{Si } \mathbf{x} \in V, \text{ entonces existe } -\mathbf{x} \text{ tal que } \mathbf{x} + (-\mathbf{x}) = \mathbf{0}$$

6. Cerradura bajo la multiplicación por un escalar
$$\text{Sean } \mathbf{x} \in V \text{ y } \alpha \text{ un escalar} \Longrightarrow \alpha\mathbf{x} \in V$$

7. Primera ley distributiva de la multiplicación por escalares
$$\text{Sean } \mathbf{x}, \mathbf{y} \in V \text{ y } \alpha \text{ un escalar} \Longrightarrow \alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$$

8. Segunda ley distributiva de la multiplicación por escalares

$$\text{Sean } \mathbf{x} \in V \text{ y } \alpha, \beta \text{ escalares} \Longrightarrow (\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$$

9. Ley asociativa de la multiplicación por escalares

$$\text{Sean } \mathbf{x} \in V \text{ y } \alpha, \beta \text{ escalares} \Longrightarrow \alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$$

10. Idéntico multiplicativo

$$\forall \mathbf{x} \in V \quad \mathbf{1x} = \mathbf{x}$$

1.3.2 Base Vectorial

Una *base* de un espacio vectorial es un conjunto de vectores en ese espacio tales que cualquier vector de dicho espacio puede escribirse como una combinación lineal de los vectores de la base.

Dependencia Lineal

Un conjunto de vectores $\{\mathbf{v}_{1,}\mathbf{v}_2, ..., \mathbf{v}_n\}$ de un espacio vectorial V es *linealmente dependiente* si cualquiera de los vectores del conjunto puede escribirse como un combinación lineal de los otros, es decir si existen n escalares $\alpha_1, \alpha_2, ..., \alpha_n$ no todos cero tales que

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + ... + \alpha_n\mathbf{v}_n = 0$$

Independencia Lineal

Sea $\{\mathbf{v}_{1,}\mathbf{v}_2, ..., \mathbf{v}_n\}$ un conjunto de vectores en un espacio vectorial V, y sean $\alpha_1, \alpha_2, ..., \alpha_n$,n escalares, si la única solución a $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + ... + \alpha_n\mathbf{v}_n = 0$ es $\alpha_1 = \alpha_2 = ... = \alpha_n = 0$ entonces se dice que el conjunto de vectores $\{\mathbf{v}_{1,}\mathbf{v}_2, ..., \mathbf{v}_n\}$ es *linealmente independiente*. Esto es equivalente a decir que si el conjunto de vectores no cumple con la definición de dependencia lineal, entonces es linealmente independiente.

De esta manera, para un espacio n-dimensional, un conjunto de n vectores linealmente independientes forma una base para ese espacio vectorial.

1.3.3 Rango de una Matriz

El *rango* de una matriz es el número de columnas o renglones linealmente independientes.

$$\begin{aligned} Rango(\mathbf{A}) &= \text{columnas linealmente independientes} \\ &= \text{renglones linealmente independientes} \end{aligned}$$

$$Rango(\mathbf{A}) \leq \min(\text{columnas}, filas)$$

Matriz de rango completo

Se dice que una matriz $\mathbf{A}_{(n\times p)}$ es de *rango completo* si su rango es igual al más pequeño de entre n y p. Es decir si su rango es igual al número de filas (si $n \leq p$) o es igual al número de columnas (si $n \geq p$).

Por ejemplo sea $\mathbf{A}$ una matriz de $m \times n$ donde $m \leq n$ cuyo rango $Rango(\mathbf{A}) = m$ es equivalente a decir que la matriz tiene m renglones linealmente independientes y por lo tanto es de rango completo. Las n columnas de A son entonces linealmente dependientes

Algunas propiedades del rango de una matriz

$$Rango(\mathbf{AB}) \leq \min(Rango(\mathbf{A}), Rango(\mathbf{B}))$$

$$Rango(\mathbf{A}) = Rango(\mathbf{A}'\mathbf{A}) = Rango(\mathbf{A}\mathbf{A}')$$

1.4 Sistema de Ecuaciones Lineales

Sea $\mathbf{A}_{m\times n}$ una matriz cuyos componentes son conocidos, sea $\mathbf{b}_{m\times 1}$ un vector cuyos componentes también son conocidos y sea $\mathbf{x}_{n\times 1}$ un vector cuyos elementos son desconocidos.

Considerando el sistema de n ecuaciones lineales

$$\mathbf{Ax} = \mathbf{b}$$

se busca saber si existe una solución, de existir como encontrarla y si esta es única.

1.4.1 Sistema de ecuaciones homogéneas

Un sistema de ecuaciones lineales de la forma $\mathbf{Ax} = \mathbf{0}$ se llama sistema de *ecuaciones homogéneas*

1.4.2 Matriz Inversa

La solución de un sistema de ecuaciones lineales de la forma $\mathbf{Ax} = \mathbf{b}$ requiere de algo similar a lo que conocemos como inverso para los números reales, es decir $\forall a \neq 0 \quad \exists \quad b = \frac{1}{a}$.

Definición 11 ***Matriz Inversa*** *Sean* $\mathbf{A}$ *y* $\mathbf{B}$ *dos matrices de* $n \times n$ *si* $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$*, entonces* $\mathbf{B}$ *es la* ***inversa*** *de* $\mathbf{A}$ *y se denota por* $\mathbf{A}^{-1}$*. Tenemos entonces*

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Si una matriz $\mathbf{A}$ es *invertible* (existe su inversa), podemos resolver un sistema de ecuaciones lineales de la forma $\mathbf{Ax} = \mathbf{b}$ premultiplicando el sistema por $\mathbf{A}^{-1}$

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{Ix} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Decir que una matriz $\mathbf{A}$ es invertible es equivalente a decir que el sistema homogéneo $\mathbf{Ax} = \mathbf{0}$ tiene sólo la solución trivial $\mathbf{x} = \mathbf{0}$, es decir que todas las columnas de $\mathbf{A}$ son linealmente independientes.

Algunas propiedades de la matriz inversa

- Si $\mathbf{A}$ es invertible, entonces su inversa es única y $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- Sean $\mathbf{A}$ y $\mathbf{B}$ dos matrices invertibles de $n \times n$, entonces $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- Si $\mathbf{A}$ es invertible, entonces $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
- Si $\mathbf{A}$ es invertible y simétrica, entonces $\mathbf{A}^{-1}$ es también simétrica
- Si $\mathbf{A}$ es ortogonal, entonces $\mathbf{A}^{-1} = \mathbf{A}'$

Matrices Singulares y No Singulares

Sea $\mathbf{A}$ una matriz cuadrada de $n \times n$, si $\mathbf{A}$ es invertible, entonces decimos que $\mathbf{A}$ es una matriz *no singular*. Por el contrario si no existe la inversa de $\mathbf{A}$, esta última es una matriz *singular*.

Determinante de una matriz

La obtención de la matriz inversa de una matriz $\mathbf{A}_{(n\times n)}$ se hace con frecuencia a través del *determinante de la matriz* que es un número calculado como una función de todos los elementos de la matriz. El determinante, denotado por $\det(\mathbf{A})$ ó $|\mathbf{A}|$, se obtiene empleando una expansión por cofactores de la manera siguiente

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} |\mathbf{A}_{ij}| \quad i = 1, ..., n$$

donde $|\mathbf{A}_{ij}|$ es el determinante de la matriz que se obtiene al eliminar el renglón i y la columna j de la matriz $\mathbf{A}$. La expresión $(-1)^{i+j} |\mathbf{A}_{ij}|$ se conoce como cofactor ij de $\mathbf{A}$.

Por ejemplo, sea $A_{(3\times 3)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$

$$\begin{aligned} |\mathbf{A}| &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22}) \\ |\mathbf{A}| &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} \end{aligned}$$

Algunas propiedades del determinante de una matriz

$$\mathbf{A} \text{ es invertible si y sólo si } |\mathbf{A}| \neq 0$$

$$\begin{aligned} |\mathbf{A}| &= |\mathbf{A}'| \\ |\mathbf{AB}| &= |\mathbf{A}|\,|\mathbf{B}| \\ \left|\mathbf{A}^{-1}\right| &= \frac{1}{|\mathbf{A}|} \end{aligned}$$

Obtención de la matriz inversa por medio del determinante La inversa de una matriz $\mathbf{A}_{n\times n}$ puede obtenerse por medio del determinante de la siguiente manera

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} adj(\mathbf{A})$$

donde $adj(\mathbf{A})$ es la *matriz adjunta* de $\mathbf{A}$, es decir la transpuesta de la *matriz de cofactores*.

Por ejemplo para una matriz $\mathbf{A}_{(2\times 2)}$, se tiene:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad |\mathbf{A}| = (a_{11}a_{22} - a_{21}a_{12}) \neq 0 \qquad adj(\mathbf{A}) = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{(a_{11}a_{22} - a_{21}a_{12})} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

Para una matriz $\mathbf{A}_{(2\times 2)}$, la obtención de la matriz inversa resulta bastante sencilla, sin embargo para matrices de tamaño $n \geq 3$ este procedimiento resulta más tedioso, la obtención de las inversas de dichas matrices por otros métodos puede consultarse en cualquier libro de álgebra lineal.[1]

1.4.3 Matriz Regular

Si $\mathbf{A}$ es una matriz cuadrada de rango completo, entonces $\mathbf{A}$ es invertible y se conoce como *matriz regular*.

1.4.4 Matriz positiva (negativa) definida

Sea $\mathbf{A}_{n\times n}$ una matriz simétrica y tenemos la forma cuadrática

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij}$$

- Si $\mathbf{x}'\mathbf{A}\mathbf{x} > \mathbf{0} \quad \forall \quad \mathbf{x} \neq \mathbf{0}$, entonces $\mathbf{A}$ es una matriz *definida positiva*
- Si $\mathbf{x}'\mathbf{A}\mathbf{x} \geq \mathbf{0} \quad \forall \quad \mathbf{x} \neq \mathbf{0}$, entonces $\mathbf{A}$ es una matriz *definida no negativa o semidefinida positiva*
- Si $\mathbf{x}'\mathbf{A}\mathbf{x} < \mathbf{0} \quad \forall \quad \mathbf{x} \neq \mathbf{0}$, entonces $\mathbf{A}$ es una matriz *definida negativa*
- Si $\mathbf{x}'\mathbf{A}\mathbf{x} \leq \mathbf{0} \quad \forall \quad \mathbf{x} \neq \mathbf{0}$, entonces $\mathbf{A}$ es una matriz *definida no positiva o semidefinida negativa*

[1] Stanley I. Grossman (1996) "Álgebra Lineal" quinta edición, Ed. Mc Graw Hill
Lehmann, C.H. (1964) "Álgebra", Ed. Limusa

Algunas propiedades de las matrices definidas no negativas En la teoría del análisis multivariado, es muy común encontrar matrices definidas no negativas, por lo que las propiedades de éstas son un caso de particular interés.

- Si $\mathbf{A}$ es una matriz definida no negativa, entonces $|\mathbf{A}| \geq 0$
- La matriz identidad $\mathbf{I}$ es una matriz definida positiva
- Si $\mathbf{A}_{(n\times m)}$ es una matriz de rango completo con $n > m$, entonces $\mathbf{A}'\mathbf{A}$ es definida positiva y $\mathbf{A}\mathbf{A}'$ es una matriz definida no negativa
- Si una matriz $\mathbf{A}$ es definida positiva, su inversa $\mathbf{A}^{-1}$ es también una matriz definida positiva
- Si $\mathbf{A}$ es definida positiva y $\mathbf{B}$ es una matriz no singular, entonces $\mathbf{B}'\mathbf{A}\mathbf{B}$ es una matriz definida positiva

1.5 Valores y Vectores Propios

1.5.1 Ecuación característica

Muchas veces resulta útil encontrar un escalar λ y un vector $\mathbf{x}$ tales que para una matriz $\mathbf{A}_{n\times n}$ se cumpla

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

que es equivalente a

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

El sistema homogéneo de ecuaciones $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ tendrá una solución diferente a la trivial siempre y cuando la matriz $(\mathbf{A} - \lambda\mathbf{I})$ sea singular o cuando $|\mathbf{A} - \lambda\mathbf{I}| = 0$, esta ecuación se conoce como *ecuación característica de* $\mathbf{A}$.

1.5.2 Valores y vectores propios

Las soluciones de la ecuación $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ son los *valores propios* λ y los *vectores propios* $\mathbf{x}$. Estos también se conocen como raíces y vectores característicos o eigenvalores y eigenvectores, en este trabajo nos referiremos a ellos simplemente como valores y vectores propios.

Valores propios

Los *valores propios* se obtienen al resolver la ecuación $|\mathbf{A} - \lambda\mathbf{I}| = 0$, cabe señalar que las raíces de este polinomio no forzosamente son reales. Si la matriz $\mathbf{A}$ es de $n \times n$, entonces $\mathbf{A}$ tendrá n valores propios $(\lambda_1, \lambda_2, ..., \lambda_n)$

Vectores propios

Una vez hallados los valores propios, retomando la ecuación $(\mathbf{A}-\lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ es posible encontrar los *vectores propios* (diferentes de cero) correspondientes a λ.

Por ejemplo sea $\mathbf{A}_{2\times 2} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= \begin{vmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{vmatrix} \\ &= (4-\lambda)(4-\lambda) - 4 \\ &= \lambda^2 - 8\lambda + 12 \end{aligned}$$

resolviendo la ecuación obtenemos los dos valores propios de $\mathbf{A}$

$$\begin{aligned} \lambda_1 &= 6 \\ \lambda_2 &= 2 \end{aligned}$$

retomando la ecuación $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ y remplazando por los valores de λ, tenemos

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} &= \mathbf{0} \\ \begin{bmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= 0 \end{aligned}$$

Para $\lambda_1 = 6$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{cases} -2x_1 + 2x_2 = 0 \\ 2x_1 - 2x_2 = 0 \end{cases}$$

$$x_1 = x_2$$

Para $\lambda_2 = 2$

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{cases} 2x_1 + 2x_2 = 0 \\ 2x_1 + 2x_2 = 0 \end{cases}$$

$$x_1 = -x_2$$

De lo anterior tenemos que dos vectores propios de $\mathbf{A}$ son $\mathbf{x}_1 = \binom{1}{1}$ y $\mathbf{x}_2 = \binom{1}{-1}$, sin embargo es conveniente que los vectores propios estén normalizados, es decir que cumplan con la condición $x'x = 1$. Normalizando los vectores tenemos entonces

$$\mathbf{x}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} y \quad \mathbf{x}_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$$

Cabe señalar que los vectores propios normalizados de una matriz son ortogonales entre si, es decir $\mathbf{x}_i'\mathbf{x}_j = 0 \quad \forall \ i \neq j$

1.5.3 Descomposición espectral de una matriz

A partir de los n vectores propios normalizados de una matriz $\mathbf{A}_{(n\times n)}$ podemos formar una matriz $\mathbf{C}$, esta matriz cuyas columnas son ortogonales entre sí es una matriz ortogonal que por lo tanto cumple con la condición $\mathbf{C'C = I.}$ Del mismo modo podemos agrupar los n valores propios de una matriz $\mathbf{A}_{n\times n}$ en una matriz diagonal $\mathbf{\Lambda}_{n\times n}$. Tenemos entonces

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \quad \text{y} \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

La i-ésima columna del producto $\mathbf{AC}$ está formada por el producto de $\mathbf{A}$ por su i-ésimo valor propio es decir

$$\mathbf{A} \begin{pmatrix} c_{11} \\ \vdots \\ c_{n1} \end{pmatrix} = \mathbf{A}\mathbf{c}_i = \lambda_i \mathbf{c}_i$$

tenemos entonces

$$\mathbf{AC} = \begin{bmatrix} \lambda_1 c_{11} & \lambda_2 c_{12} & \cdots & \lambda_n c_{1n} \\ \lambda_1 c_{21} & \lambda_2 c_{22} & \cdots & \lambda_n c_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda_1 c_{n1} & \lambda_2 c_{n2} & \cdots & \lambda_n c_{nn} \end{bmatrix}$$

pero esto es equivalente a $\mathbf{C\Lambda}$**,** por lo que

$$\mathbf{AC} = \mathbf{C\Lambda}$$

premultiplicando por $\mathbf{C}'$

$$\mathbf{C}'\mathbf{AC} = \mathbf{C}'\mathbf{C\Lambda}$$

pero dado que $\mathbf{C}$ es una matriz ortogonal, entonces $\mathbf{C}'\mathbf{C} = \mathbf{I}$, y

$$\mathbf{\Lambda} = \mathbf{C}'\mathbf{AC}$$

Del mismo modo partiendo de que $\mathbf{AC} = \mathbf{C\Lambda}$ y postmultiplicando por $\mathbf{C}'$ obtenemos

$$\begin{aligned} \mathbf{ACC}' &= \mathbf{C\Lambda C}' \\ \mathbf{AI} &= \mathbf{C\Lambda C}' \\ \mathbf{A} &= \mathbf{C\Lambda C}' \end{aligned}$$

Definición 12 *La* ***diagonalización*** *de una matriz* $\mathbf{A}_{n\times n}$ *está dada por*

$$\mathbf{\Lambda} = \mathbf{C}'\mathbf{AC}$$

donde $\mathbf{\Lambda}$ *es la matriz diagonal cuyos elementos son los valores propios de* $\mathbf{A}$ *y* $\mathbf{C}$ *es la matriz cuyas columnas están formadas por los vectores propios normalizados de* $\mathbf{A}$*.*

Definición 13 *La* ***descomposición espectral*** *de una matriz* $\mathbf{A}_{n\times n}$ *está dada por*

$$\mathbf{A} = \mathbf{C\Lambda C}'$$

donde $\mathbf{\Lambda}$ *es la matriz diagonal cuyos elementos son los valores propios de* $\mathbf{A}$ *y* $\mathbf{C}$ *es la matriz cuyas columnas están formadas por los vectores propios normalizados de* $\mathbf{A}$*.*

Capítulo 2

Análisis Multivariado

El análisis multivariado (o análisis estadístico multivariado) es aquella rama de la estadística dedicada al estudio de distribuciones multivariadas (o multidimensionales) y sus muestras.

En un contexto aplicado, el análisis multivariado trata con un grupo (o varios) de individuos para los cuales se tienen valores de dos o más variables (características). Por medio del análisis multivariado se busca estudiar las interrelaciones entre las variables, las posibles diferencias entre grupos (en términos de estas variables), así como hacer inferencias sobre las poblaciones (grupos) de las que fueron extraídas las muestras.

En la práctica, pocas veces se utilizan modelos univariados, sin embargo, el entendimiento de estos está estrechamente ligado con el entendimiento de los modelos multivariados, por ello en este capítulo se presentarán tanto los conceptos univariados básicos como los multivariados.

En este capítulo se presentarán las distribuciones multivariadas, los datos multivariados y las estadísticas básicas, haciendo especial énfasis en la distribución normal multivariada y sus propiedades. Así mismo, se presentarán algunas pruebas de hipótesis. Los conceptos y definiciones básicos que forman el marco teórico del análisis multivariado se presentan brevemente en este capítulo, muchos de ellos a manera de recordatorio, ya que el tema principal de este trabajo es el análisis discriminante. Un estudio más profundo y detallado de estos temas puede consultarse en cualquier libro de probabilidad y estadística o análisis multivariado.[1]

[1] Mood, Graybill and Boes (1974) "Introduction to the theory of Statistics" 3^a Edición, Ed. Mc Graw Hill
C. Chatfield A.J. Collins (1980) "Introduction to Multivariate Analysis", Ed. Chapman and Hall

2.1 Definiciones básicas de probabilidad

Uno de los objetivos de la ciencia es describir y predecir eventos o fenómenos del mundo que nos rodea, para ello, se pueden construir modelos matemáticos que describan adecuadamente la realidad. Cualquier análisis estadístico debe construirse sobre la base de un modelo matemático. El modelo deberá parametrizarse de tal manera que cada parámetro sea interpretado con facilidad y corresponda a algún aspecto de la realidad. Por medio de la observación del fenómeno, que interpretamos como la realización de un fenómeno aleatorio, se obtendrán datos y de esta manera se podrá inferir sobre los parámetros del modelo.

En esta sección, se hace un breve resumen de las definiciones y conceptos de probabilidad necesarios para abordar la teoría de la distribución, indispensable a su vez para abordar la teoría del análisis multivariado.

Definición 14 ***Espacio Muestral*** *El* ***espacio muestral*** *, denotado por Ω, es la colección de todos los posibles resultados de un experimento conceptual.*

Definición 15 ***Evento y espacio evento*** *Un* ***evento*** *es un subconjunto del espacio muestral. La clase de todos los eventos asociados con un experimento dado se define como* ***espacio evento****. El espacio evento se denota generalmente por $\mathcal{A}$ y debe cumplir con las siguientes propiedades*

1. *$\Omega \in \mathcal{A}$.*
2. *Si $A \in \mathcal{A}$ entonces $A^c \in \mathcal{A}$.*
3. *Si A_1 y $A_2 \in \mathcal{A}$, entonces $A_1 \cup A_2 \in \mathcal{A}$.*

   Cualquier colección de eventos que cumpla con estas tres propiedades es conocida como ***álgebra*** *de eventos.*

Definición 16 ***Función*** *Una* ***función*** *$f(\cdot)$ es una regla que asocia cada punto en un conjunto A con uno y solamente uno de los puntos de otro conjunto B. El primer conjunto de puntos es llamado* ***dominio*** *de la función mientras que el segundo se conoce como* ***contradominio****.*

Definición 17 ***Función de Probabilidad*** *Sea Ω un espacio muestral y $\mathcal{A}$ una colección de eventos (se asume que $\mathcal{A}$ es un álgebra de eventos) de un experimento dado. La* ***función de probabilidad*** *$P[\cdot]$ es una función con dominio $\mathcal{A}$ y contradominio el intervalo $[0,1]$ que satisface los siguientes axiomas:*

1. *$P[A] \geq 0$ para todo $A \in \mathcal{A}$*

2. *$P[\Omega] = 1$*

3. *Si $A_1, A_2, ...$ es una secuencia de eventos mutuamente excluyentes en $\mathcal{A}$ (es decir $A_i \cap A_j = \phi$ para todo $i \neq j, i, j = 1, 2....$) y además $A_1 \cup A_2 \cup ... \cup A_n \cup ... = \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ entonces tenemos*

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i]$$

Definición 18 ***Espacio de Probabilidad*** *Un* ***espacio de probabilidad*** *es la tercia $(\Omega, \mathcal{A}, P[\cdot])$, donde Ω es un espacio muestral, $\mathcal{A}$ es una colección (álgebra) de eventos (subconjuntos de Ω) y $P[\cdot]$ es una función de probabilidad con dominio $\mathcal{A}$.*

Definición 19 ***Variable aleatoria univariada*** *Recordemos que, para un espacio de probabilidad dado $(\Omega, \mathcal{A}, P[\cdot])$,una* ***variable aleatoria univariada*** *(v.a.) denotada por X es una función con dominio Ω y contradominio la recta real. La función $X(\cdot)$ debe ser de tal forma que el conjunto A_r definido por $A_r = \{\omega : X(\omega) \leq r\}$ pertenezca a $\mathcal{A}$ para cada numero real r.*

Definición 20 ***Vector Aleatorio*** *Un* ***vector aleatorio o variable aleatoria $p-$dimensional*** **X** *se define de tal manera que*

$$\mathbf{X}' = [X_1, X_2, ...X_p] \text{ donde } X_1, X_2, ...X_p \text{ son v.a. univariadas}$$

2.2 Distribuciones Multivariadas

Las distribuciones de probabilidad sirven de modelos para fenómenos aleatorios cuya observación o medición genera los datos con los que se realizará el análisis.

Uno de los conceptos esenciales del análisis multivariado es la idea de una distribución de probabilidad multivariada, para poder entenderla es necesario tener claros los conceptos de distribución de probabilidad para el caso univariado. En esta sección se presentan las definiciones básicas de la teoría de la distribución para el caso univariado y se hace la extensión al caso multivariado.

2.2.1 Definiciones básicas

Definición 21 ***Función de distribución*** *La* ***función de distribución*** *de una v.a. X , denotada por $F_X(\cdot)$ es aquella función cuyo dominio es la recta real y contradominio el intervalo $[0,1]$, que satisface*

$$\begin{aligned} F_X(x) &= P[X \leq x] \\ &= P[\{\omega : X(\omega) \leq x\} \text{ para todo } x \in \mathbb{R} \end{aligned} \tag{2.1}$$

Definición 22 ***Variable aleatoria discreta*** *Una v.a. X se definirá como* ***discreta*** *si el rango de X es contable. Si una v.a. X es discreta, su función de distribución $F_X(\cdot)$ también será discreta.*

Definición 23 ***Variable aleatoria continua*** *Una v.a. X es llamada* ***continua*** *si existe una función $f_X(\cdot)$ tal que*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \quad \text{para cada número real } x.$$

Se dice que la función de distribución de una v.a. continua $F_X(x)$ es absolutamente continua.

2.2.2 Funciones de densidad para variables aleatorias univariadas

Definición 24 ***Función de masa de probabilidad*** *Si X es una v.a. discreta con valores distintos $x_{1,}x_2, ..., x_n, ...$, definimos la función de densidad o* ***función de masa*** *de probabilidad $f_X(\cdot)$ como:*

$$f_X(x) = \begin{cases} P[X = x] & si \;\; x = x_j, j = 1, 2, ..., n, ... \\ 0 & si\; x \neq x_j \end{cases} \tag{2.2}$$

Definición 25 ***Función de densidad para una v.a. continua*** *Si X es una v.a. continua, la función $f_X(x)$ en la expresión*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{2.3}$$

es la ***función de densidad*** *de X. De esta manera, es posible obtener $F_X(x)$ a partir de $f_X(x)$ y viceversa ($f_X(x) = \frac{dF_X(x)}{dx}$ para aquellos puntos x para los que $F_X(x)$ sea diferenciable).*

2.2.3 Distribución Conjunta

Definición 26 ***Función de distribución conjunta*** *Uno de los conceptos esenciales del análisis multivariado es, como ya se había mencionado, la idea de una distribución de probabilidad multivariada, para ello partimos de un vector aleatorio $\mathbf{X}' = [X_{1,}X_2, ...X_p]$ donde $X_{1,}X_2, ...X_p$ son v.a. univariadas definidas en el mismo espacio de probabilidad $(\Omega, \mathcal{A}, P[\cdot])$. La función de* ***distribución conjunta*** *de $X_{1,}X_2, ...X_p$, denotada por $F_{X_{1,}X_2,....X_p}(\cdot, \cdot, ..., \cdot)$ se define como*

$$F_{X_{1,}X_2,....X_p}(\cdot, \cdot, ..., \cdot) = P[X_1 \leq x_1, X_2 \leq x_2, ..., X_p \leq x_p] \;\; para\; todo\; (x_1, x_2, ..., x_p) \tag{2.4}$$

De esta manera, la función de distribución conjunta de $X_{1,}X_2, ...X_p$ es una función cuyo dominio es el espacio euclidiano p-dimensional y contradomino el intervalo $[0, 1]$.

Definición 27 ***Función de masa de probabilidad conjunta*** *Sea $\mathbf{X}' = [X_{1,}X_2, ...X_p]$ un vector aleatorio cuyas componentes son v.a. discretas. La función de densidad conjunta o* ***función de masa de probabilidad conjunta*** *$f_{X_{1,}X_2,....X_p}(\cdot, \cdot, ..., \cdot)$ se define de la siguiente manera:*

$$f_{X_{1,}X_2,....X_p}(\cdot, \cdot, ..., \cdot) = P[X_1 = x_1, X_2 = x_2, ..., X_p = x_p] \tag{2.5}$$

para $(x_1, x_2, ..., x_p)$ un valor de $[X_{1,}X_2, ...X_p]$, cero en otro caso.

Cabe señalar que $f_{X_1,X_2,....X_p}(\cdot,\cdot,...,\cdot) \geq 0 \; \forall \; (x_1, x_2, ..., x_p)$ y que $\sum f_{X_1,X_2,....X_p}(\cdot,\cdot,...,\cdot) = 1$ donde la suma se realiza sobre todos los posibles valores de $\mathbf{X}$.

Definición 28 ***Función de densidad conjunta***

Se dice que el vector aleatorio $\mathbf{X}' = [X_1, X_2, ..., X_p]$ *es una v.a. continua de dimensión p si y sólo si existe una función* $f_{X_1,X_2,....X_p}(\cdot,\cdot,...,\cdot) \geq 0$ *tal que*

$$F_{X_1,X_2,....X_p}(x_1, x_2, ..., x_p) = \int_{-\infty}^{x_p} ... \int_{-\infty}^{x_1} f_{X_1,X_2,....X_p}(u_1, u_2, ..., u_p)du_1...du_p \tag{2.6}$$

para todo $(x_1, x_2, ..., x_p)$.

Dicha función se conoce como ***función de densidad conjunta*** *y al igual que en el caso univariado,* $f_{X_1,X_2,....X_p}(\cdot,\cdot,...,\cdot)$ *cumple con las siguientes propiedades:*

1. $f_{X_1,X_2,....X_p}(x_1, x_2, ..., x_p) \geq 0$
2. $\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1,X_2,....X_p}(x_1, x_2, ..., x_p)dx_1...dx_p = 1$

La función de densidad conjunta puede obtenerse a partir de la función de distribución conjunta de la siguiente manera:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p F_X(x_1, x_2, ..., x_p)}{\partial x_1 \partial x_2 ... \partial x_p} \tag{2.7}$$

De hecho, a partir de la función de distribución conjunta es posible obtener la función de densidad conjunta y viceversa.

En el caso univariado, la función de densidad es utilizada para encontrar el área bajo la curva de $f_X(\cdot)$ en un intervalo (a, b), es decir

$$P[a < x < b] = \int_a^b f_X(x)dx$$

Para el caso bivariado por ejemplo, las probabilidades se encontrarán por medio del volumen bajo $f_{X_1,X_2}(\cdot,\cdot)$ en una cierta región R, es decir $P[(x_1, x_2) \in R] = \iint_R f_{X_1,X_2}(x_1, x_2)dx_1dx_2$. De esta manera, en el caso de un vector aleatorio p-dimensional, las probabilidades se pueden encontrar integrando sobre el subconjunto de interés en el espacio p-dimensional.

2.2.4 Distribución Marginal

Definición 29 ***Funciones de distribución Marginales***

Si $F_{X_{1,}X_2,....X_p}(\cdot,\cdot,...,\cdot)$ es la función de distribución conjunta de $X_{1,}X_2,....X_p$, entonces las funciones de distribución $F_{X_1}(\cdot)$, $F_{X_2}(\cdot)$,...,$F_{Xp}(\cdot)$ se conocen como funciones de ***distribución marginales.***

Definición 30 ***Función de masa de probabilidad marginal*** *A partir de la función de densidad conjunta de un vector aleatorio discreto $\mathbf{X}' = [X_{1,}X_2,....X_p]$ es posible obtener la función de* ***masa de probabilidad marginal*** *de cualquiera de los X_i de $\mathbf{X}'$ al sumar la función de masa conjunta sobre todas las demás variables, es decir*

$$\begin{aligned} f_{X_i}(x_i) &= P[X_i = x_i] \\ &= \sum f_{X_{1,}X_2,...,X_{i,},....X_p}(x_1, x_2, ..., x_{i,}..., x_p) \end{aligned} \tag{2.8}$$

donde la suma se realiza sobre todas las $\mathbf{x} = (x_1, x_2, ..., x_{i,}..., x_p)$ tales que el i-ésimo componente es fijo e igual a x_i, es decir sobre $x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_p$

Definición 31 ***Función de densidad marginal*** *Sea $\mathbf{X}' = [X_{1,}X_2, ..., X_p]$ una v.a. continua de dimensión p, la función de* ***densidad marginal*** *de cualquier componente de $\mathbf{X}'$, X_i puede encontrarse a partir de la función de densidad conjunta, al integrar esta última sobre todas las variables excepto X_i*

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_{1,}X_2,....X_p}(x_1, x_2, ..., x_p) dx_1, ..., dx_{i-1}, dx_{i+1}, ...dx_p \tag{2.9}$$

2.2.5 Distribución Condicional

Definición 32 ***Función de densidad condicional*** *Sea $\mathbf{X}' = [X_{1,}X_2, ...X_p]$ una v.a. (continua o discreta) de dimensión p y sean $X_{1,}X_2, ..., X_k$ y $X_{k+1,}..., X_p$ dos conjuntos disjuntos de variables de $\mathbf{X}$. La función de* ***densidad condicional*** *del vector aleatorio $k-$dimensional $[X_{1,}X_2, ..., X_k]$ dado el valor $(x_{k+1,}..., x_p)$ del vector $[X_{k+1,}..., X_p]$ se define como*

$$f_{X_{1,}X_2,...,X_k|X_{k+1,}...,X_p}(x_{1,}x_2, ..., x_k \mid x_{k+1,}..., x_p) = \frac{f_{X_{1,}X_2,...,X_k,X_{k+1,}...,X_p}(x_{1,}x_2, ..., x_k, x_{k+1,}..., x_p)}{f_{X_{k+1,}...,X_p}(x_{k+1}, ..., x_p)} \tag{2.10}$$

donde $f_{X_{k+1,}...,X_p}(x_{k+1,}..., x_p)$ es la función de densidad marginal de $[X_{k+1}, ..., X_p]$

2.2.6 Independencia

Definición 33 ***Independencia*** *Sea* $\mathbf{X}' = [X_1, X_2, ...X_p]$ *una v.a. p-dimensional, las variables* $X_1, X_2, ...X_p$ *son* ***independientes*** *si y sólo si*

$$F_{X_1, X_2, ...X_p}(x_1, x_2, ..., x_p) = \prod_{i=1}^{p} F_{X_i}(x_i) \text{ para todo } (x_1, x_2, ..., x_p) \tag{2.11}$$

o equivalentemente si

$$f_{X_1, X_2, ...X_p}(x_1, x_2, ..., x_p) = \prod_{i=1}^{p} f_{X_i}(x_i) \text{ para todo } (x_1, x_2, ..., x_p) \tag{2.12}$$

donde $F_{X_1, X_2, ...X_p}(\cdot, \cdot, ..., \cdot)$ *es la función de distribución conjunta de* $\mathbf{X}$ *,* $f_{X_1, X_2, ...X_p}(\cdot, \cdot, ..., \cdot)$ *es su función de densidad conjunta y* $F_{X_i}(x_i)$ *,* $f_{X_i}(x_i)$ *son las funciones marginales de distribución y densidad respectivamente.*

2.3 Datos Multivariados

Una variable es una característica o atributo de interés del individuo o sujeto de estudio, esta variable puede tomar diferentes valores, al ser observados para diferentes individuos constituyen los datos por medio de los cuales se hará el análisis estadístico.

Definición 34 ***Muestra aleatoria*** *Se dice que las variables aleatorias* $X_1, X_2, ..., X_n$ *forman una* ***muestra aleatoria*** *de una población con distribución* $F(\cdot)$ *si* $X_1, X_2, ..., X_n$ *son variables aleatorias independientes e idénticamente distribuidas con función de distribución común* $F(\cdot)$ *.*

Una parte importante de esta definición es el significado de las variables aleatorias $X_1, X_2, ..., X_n$, la variable X_i es la representación del valor numérico que tomará el i-ésimo elemento muestreado. Una vez realizadas las observaciones, los valores de $X_1, X_2, ..., X_n$ son conocidos y se denotarán por $x_1, x_2, ..., x_n$, estos valores son llamados algunas veces muestra aleatoria. La muestra aleatoria $X_1, X_2, ..., X_n$, no debe confundirse con el vector aleatorio $\mathbf{X}' = [X_1, X_2, ...X_p]$ presentado anteriormente, ya que en éste no se definía una distribución común para las p variables que lo conforman.

Definición 35 ***Matriz Multivariada*** *Una* ***matriz multivariada*** *está compuesta por una serie de medidas u observaciones de diferentes variables realizadas para un cierto número de individuos, sujetos u objetos. De esta manera obtenemos una* ***matriz de datos multivariados*** *de la forma:*

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{1p} \\ x_{21} & x_{22} & x_{2p} \\ \\ x_{n1} & x_{n2} & x_{np} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ individuos \\ \end{matrix}$$
$$variables$$

donde x_{ij} es el valor de la $j-$ésima variable para el $i-$ésimo individuo.

El número de individuos es n y el número de variables observadas para cada uno de estos n individuos es p, teniendo de esta forma un total de $n \times p$ observaciones (mediciones).

Definición 36 ***Muestra aleatoria multivariada*** *Se dice que un conjunto de datos constituye una* ***muestra aleatoria multivariada*** *si cada individuo ha sido extraído al azar de una población de individuos, y se ha realizado sobre él la medición u observación de ciertas características. El vector aleatorio $\mathbf{X}'$ no debe confundirse con la matriz de datos observados $\mathbf{X}$, por lo que será necesario especificar el significado según el contexto.*

Los datos obtenidos por medio de las observaciones o mediciones realizadas a los individuos suelen ser de diferentes tipos, dependiendo si se busca obtener información cuantitativa o cualitativa.

Los tipos de datos más comunes son:

- Nominal: Variables categóricas sin un orden específico (Por ejemplo el sexo, el alfabetismo o analfabetismo del individuo etc.)
- Ordinal: Variables categóricas en las que hay un orden sin que este implique una distancia entre los diferentes puntos de la escala (Por ejemplo el nivel de educación: Primaria, Secundaria, Preparatoria...)
- Escala: Por medio de este tipo de medición se puede cuantificar la "magnitud relativa" de las observaciones de las variables, así como las diferencias entre ellas y las diferencias respecto a la posición cero que es fija. (Por ejemplo la edad, la estatura y el peso)

La información cualitativa suele presentarse con códigos numéricos por ejemplo para el sexo de un individuo, podemos elegir arbitrariamente los códigos sexo=1 para los hombres y sexo=2 para las mujeres. Es importante notar que un mismo valor numérico puede representar información completamente diferente de acuerdo con la escala de medición de los datos.

2.4 Estadísticas Multivariadas Básicas

Al igual que en el caso de los datos univariados, es posible obtener las estadísticas básicas, tales como la media y la varianza, a partir de los datos multivariados. Para poder resumir un conjunto de datos multivariados, es necesario obtener las estadísticas básicas de cada una de las variables por separado, así como las relaciones entre las variables que generalmente se manejan por pares. En esta sección se presentan las estadísticas descriptivas de mayor interés, haciendo a la par un recordatorio de las definiciones de esperanza, varianza, covarianza y correlación.

2.4.1 Vector de Medias

Una matriz de datos multivariados $\mathbf{X}_{np} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$ puede verse como un conjunto de n observaciones para cada una de las p variables de un vector aleatorio $\mathbf{X}' = [X_1, X_2, ...X_p]$.

$$X_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} .... \quad X_p = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix}$$

La **media** de cada una de las variables de $\mathbf{X}$, que es una medida de localización central de la densidad está dada por

$$\mu_i = E[X_i] \tag{2.13}$$

Recordemos que la **esperanza** de una v.a. X está definida como:

$$E[X] = \sum_j x_j f_X(x_j) \tag{2.14}$$

si X es una v.a discreta con puntos de masa $x_1, x_2, ..., x_j, ...$ y función de masa de probabilidad $f_X(x)$

o

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) \text{ si } X \text{ es una v.a continua con función de densidad } f_X(x) \tag{2.15}$$

Algunas Propiedades de la Esperanza

Sean X una variable aleatoria; $c, c_{1,}c_2$ constantes y $g(\cdot), g_1(\cdot), g_2(\cdot)$ funciones con domino y contradomino la recta real

- $E[c] = c$
- $E[cg(X)] = cE[g(X)]$
- $E[c_1g_1(X) + c_2g_2(X)] = c_1E[g_1(X)] + c_2E[g_2(X)]$
- $E[g_1(X)] \leq E[g_2(X)]$ si $g_1(X) \leq g_2(X)$ para toda x

La **media muestral**, que es una estimación de la media a partir de las n observaciones de X, está dada por:

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{2.16}$$

De esta manera, para un vector aleatorio $p-$dimensional, tendremos un **vector de medias** cuyas componentes están conformadas por la media de cada una de las p variables en cuestión.

El vector de medias se representa entonces por

$$\boldsymbol{\mu}' = (\mu_1, \mu_2, ..., \mu_p) \text{ donde } \mu_i = E[X_i] \tag{2.17}$$

y para un conjunto de datos multivariado $\mathbf{X}_{np}$, el **vector de medias muestrales** estará dado por

$$\overline{\mathbf{X}} = (\overline{X}_1, \overline{X}_2, ..., \overline{X}_i, ..., \overline{X}_p) \text{ donde } \overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}. \tag{2.18}$$

2.4.2 Vector de Varianzas

Como ya habíamos mencionado, la media es una medida de localización central de la densidad de una variable, del mismo modo, es posible medir la dispersión de la densidad de una variable X a través de su **varianza**, denotada por σ_X^2 o $var[X]$.

En el caso univariado, para una variable aleatoria X con media μ_X definimos la **varianza** como:

$$var[X] = E[(X - \mu_X)^2] \tag{2.19}$$

es decir,

$$var[X] = \sum_j (x_j - \mu_X)^2 f_X(x_j) \tag{2.20}$$

si X es una v.a discreta con puntos de masa $x_1, x_2, ..., x_j, ...$ y función de masa de probabilidad $f_X(x)$

o bien

$$var[X] = \int_{-\infty}^{\infty} (x_j - \mu_X)^2 f_X(x_j) \text{ si } X \text{ es una v.a continua con función de densidad } f_X(x) \tag{2.21}$$

La **desviación estándar** de una variable X, denotada por σ_x está definida como $\sqrt{var[X]}$, la desviación estándar es también una medida de dispersión de los valores de la v.a., y en ocasiones es preferible utilizarla ya que a diferencia de la varianza, tiene las mismas unidades de medición que X.

La **varianza muestral**, que es una estimación de la varianza de una variable X a partir de las n observaciones de ésta, se denota por S^2 y se define como:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \text{ para } n > 1 \tag{2.22}$$

De esta manera, para el caso multivariado, podemos definir el **vector de varianzas**, denotado por $\boldsymbol{\sigma}'$, cuyas componentes son las varianzas de cada una de las variables del vector aleatorio.

$$\boldsymbol{\sigma}' = (\sigma_1^2, \sigma_2^2, ..., \sigma_i^2, ..., \sigma_p^2) \text{ donde } \sigma_i^2 = E[(X_i - \mu_i)^2] \tag{2.23}$$

Una estimación de $\boldsymbol{\sigma}'$, es el **vector de varianzas muestrales**

$$\mathbf{S}' = (S_1^2, S_2^2, ..., S_i^2, ..., S_p^2) \text{ donde } S_i^2 \text{ es la varianza muestral de la i-ésima variable} \tag{2.24}$$

2.4.3 Matriz de Varianza-Covarianza

La **covarianza** entre dos variables aleatorias X y Y denotada por $cov[X, Y,]$ o σ_{XY} está definida como:

$cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]$ donde μ_X y μ_Y son las medias de X y Y respectivamente (2.25)

alternativamente,

$$cov[X, Y] = E[XY] - \mu_Y \mu_X \quad \text{(por propiedades de la esperanza)} \tag{2.26}$$

La covarianza está definida siempre y cuando esta esperanza exista y es una medida de la relación lineal entre X y Y, en efecto, $cov[X, Y]$ será positiva cuando $(X - \mu_X)$ y $(Y - \mu_Y)$ tiendan a ser del mismo signo con alta probabilidad y será negativa cuando $(X-\mu_X)$ y $(Y-\mu_Y)$ tiendan a ser de signo contrario con alta probabilidad . Cabe señalar que la covarianza de una variable consigo misma es simplemente la varianza de la variable, es decir

$$\begin{aligned} cov[X, X] &= E[(X - \mu_X)(X - \mu_X)] \\ &= E[(X - \mu_X)^2] \\ &= var[X] \end{aligned} \tag{2.27}$$

Del mismo modo tenemos que $cov[X, Y] = cov[Y, X]$.

Por medio de las observaciones de dos variables X y Y podemos obtener la **covarianza muestral** S_{ij}, que se define de la manera siguiente

$$S_{ij} = \frac{\sum_{k=1}^{n}(X_{ki-}\overline{X}_i)(X_{kj-}\overline{X}_j)}{n-1} \tag{2.28}$$

Si tomamos $i = j$ en la expresión anterior, obtenemos la varianza muestral la i-ésima variable S_i^2, en el caso multivariado, se utilizará la notación S_{ii} en lugar de S_i^2, y para la desviación estándar se usará la notación S_i .

Para una variable aleatoria $p-$dimensional tenemos entonces p varianzas y $\frac{p(p-1)}{2}$ covarianzas distintas que pueden ordenarse en una matriz simétrica $\mathbf{\Sigma}_{p\times p}$ de tal forma que

$$\mathbf{\Sigma}_{p\times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \text{ donde } \sigma_{ij} = \sigma_{ji} = cov[X_i, X_j] \tag{2.29}$$

Esta matriz puede ser expresada en forma alternativa de la siguiente manera, utilizando las ecuaciones (2.25) y (2.26)

$$\begin{aligned} \mathbf{\Sigma}_{p\times p} &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] \\ &= E[\mathbf{X}\mathbf{X}'] - \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned} \tag{2.30}$$

Los elementos en la diagonal de esta matriz son las varianzas de las variables, mientras que aquellos fuera de esta son las covarianzas entre cada par de variables del conjunto y dado que $\sigma_{ij} = \sigma_{ji}$, la matriz $\mathbf{\Sigma}_{p\times p}$ es en efecto simétrica, además es positiva semidefinida. Esta matriz se conoce como **matriz de varianza covarianza** y análogamente para un conjunto de datos multivariados $\mathbf{X}_{np}$, podemos definir la **matriz de varianza covarianza muestral**, denotada por $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} \text{ donde } S_{ij} \text{ es la covarianza muestral entre la i-ésima y la j-ésima variable} \tag{2.31}$$

La matriz de varianza covarianza muestral es, al igual que la matriz de varianza covarianza, una matriz simétrica cuyos elementos en la diagonal son las varianzas muestrales de las variables y los elementos fuera de esta diagonal son las covarianzas muestrales entre cada par de variables.

El determinante de la matriz de varianza covarianza $V = |\mathbf{\Sigma}|$ es conocido como **varianza generalizada** y es una medida que sintetiza la dispersión de los datos, siendo una especie de varianza multivariada.

2.4.4 Matriz de Correlación

Dado que la covarianza entre dos variables depende de las unidades en las que se miden estas últimas, la interpretación de la covarianza suele ser difícil, de hecho la magnitud de la covarianza no tiene mucho significado ya que depende de la variabilidad de las dos variables. Por ello, es muchas veces preferible contar con una medida de la relación lineal entre las variables que no dependa de las unidades en las que estas se miden. Esto se logra estandarizando la covarianza al dividirla por el producto de las desviaciones estándar de las dos variables en cuestión.

A esta medida de relación lineal entre dos variables X y Y, que no tiene unidades de medición, se le conoce como **coeficiente de correlación**, generalmente se denota por ρ_{XY} y está dado por:

$$\rho_{XY} = \frac{cov[X,Y]}{\sigma_X \sigma_Y} \tag{2.32}$$

siempre y cuando $cov[X,Y], \sigma_X$ y σ_Y existan y además $\sigma_X > 0$ y $\sigma_Y > 0$.

El coeficiente de correlación mide el grado de relación lineal entre dos variables y toma valores ente -1 y 1, es decir $-1 \leq \rho_{XY} \leq 1$, un coeficiente de correlación cercano a 1 (ó-1) indicará una fuerte relación lineal entre las variables, dándose la igualdad cuando una de las variables es una función lineal de la otra. En cambio una correlación cercana a cero indicará que no existe relación lineal entre las variables, es importante señalar que una correlación igual a cero entre dos variables no implica que éstas sean independientes.

El **coeficiente de correlación muestral** entre dos variables X_i y X_j , denotado por r_{ij} está dado por

$$\begin{aligned} r_{ij} &= \frac{\sum_{K=1}^{n}(X_{ki}-\overline{X}_i)(X_{kj}-\overline{X}_j)}{(n-1)S_iS_j} \\ &= \frac{S_{ij}}{S_iS_j} \end{aligned} \tag{2.33}$$

Estos coeficientes pueden ser vistos como covarianzas muestrales normalizadas, cuyos valores están entre -1 y1.

Para un vector aleatorio con p variables, tenemos entonces $\frac{p(p-1)}{2}$ correlaciones distintas que al igual que en el caso de la matriz de varianza covarianza pueden ordenarse en una matriz simétrica $\mathbf{P}_{p\times p}$ llamada **matriz de correlación.**

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \text{ donde } \rho_{ij} = \rho_{ji} = \frac{cov[X_i, X_j]}{\sigma_{X_i}\sigma_{X_j}} \tag{2.34}$$

Los elementos de la diagonal de $\mathbf{P}$ son iguales uno y dado que $\rho_{ij} = \rho_{ji}$, la matriz es simétrica. Es posible relacionar esta matriz de correlaciones con la matriz de varianza covarianza al definir una matriz diagonal $\mathbf{D}$, cuyos elementos en la diagonal son las desviaciones estándar de las variables del conjunto, es decir

$$\mathbf{D} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ & & \\ 0 & 0 & \sigma_p \end{bmatrix}$$

de esta manera tenemos que:

$$\mathbf{\Sigma} = \mathbf{DPD} \tag{2.35}$$

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}$$

Al igual que $\mathbf{\Sigma}$, $\mathbf{P}$ es una matriz positiva semidefinida.

Del mismo modo es posible ordenar los coeficientes de correlación muestral en una matriz, simétrica y cuyos elementos en la diagonal son iguales a 1, $\mathbf{R}_{p\times p}$ conocida como **matriz de correlación muestral** y que está dada por

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{1p} \\ r_{21} & 1 & r_{2p} \\ & & \\ r_{p1} & r_{p2} & 1 \end{bmatrix} \tag{2.36}$$

donde r_{ij} es el coeficiente de correlación muestral entre la i-ésima y la j-ésima variable.

Análogamente con la ecuación (2.35), es posible calcular la matriz $\mathbf{R}$ a partir de la matriz de varianzas covarianzas muestrales $\mathbf{S}$, definiendo una matiz diagonal $\hat{\mathbf{D}}$ cuyos elementos en la diagonal son las desviaciones estándar muestrales de las variables.

$$\hat{\mathbf{D}} = \begin{bmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \\ \\ 0 & 0 & S_p \end{bmatrix}$$

entonces tenemos

$$\mathbf{R} = \hat{\mathbf{D}}^{-1}\mathbf{S}\hat{\mathbf{D}}^{-1} \tag{2.37}$$

2.4.5 Combinación Lineal de Variables

En muchos de los métodos del análisis multivariado, se utilizan combinaciones lineales de variables dado que son una alternativa para reducir la dimensión y de esta manera la manipulación de los datos es más simple. En particular, en análisis discriminante, las combinaciones lineales de variables son utilizadas para formar la función discriminante, como se verá en el siguiente capítulo. Por estos motivos es importante poder obtener los estadísticos relacionados con una combinación lineal de variables dada.

Definición 37 ***Combinación Lineal de variables*** *Sea* $\mathbf{a}' = (a_{1,}a_{2,\ldots,}a_n)$ *un vector de constantes y sea* $\mathbf{X}=\begin{bmatrix} X_1 \\ X_2 \\ \\ X_p \end{bmatrix}$ *un vector aleatorio de dimensión p , entonces* $Y = \mathbf{a}'\mathbf{X}$ *es una* ***combinación lineal*** *de* $\mathbf{X}$.

Y es una variable aleatoria univariada cuya media y varianza son:

$$E[Y] = \mathbf{a}'\boldsymbol{\mu} \text{ donde } \boldsymbol{\mu} \text{ es el vector de medias de } \mathbf{X} \tag{2.38}$$

$$var[Y] = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \text{ donde } \boldsymbol{\Sigma} \text{ es la matriz de varianza covarianza de } \mathbf{X} \tag{2.39}$$

La demostración de la ecuación(2.38) es inmediata por propiedades de la esperanza.

Para la varianza de Y, tenemos que:

$$\begin{aligned} var[Y] &= E[\mathbf{a}'(\mathbf{X}-\boldsymbol{\mu})^2] \\ &= E[\mathbf{a}'(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\mathbf{a}] \text{ puesto que } \mathbf{a}'(\mathbf{X}-\boldsymbol{\mu}) \text{ es un escalar y es igual a su transpuesta} \\ &= \mathbf{a}'E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})']\mathbf{a} \text{ por propiedades de la esperanza} \\ &= \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \text{ (ver 2.30)} \end{aligned}$$

Las ecuaciones (2.38) y (2.39) pueden generalizarse para el caso en que en lugar de un vector de constantes, se tenga una matriz de constantes $\mathbf{A}_{p\times m}$, entonces el vector aleatorio (de dimensión $(m \times 1)$) $\mathbf{A}'\mathbf{X}$ tendrá un vector de medias y una matriz de varianza covarianza dados por

$$E[\mathbf{A}'\mathbf{X}] = \mathbf{A}'\boldsymbol{\mu} \tag{2.40}$$

$$var[\mathbf{A}'\mathbf{X}] = \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} \tag{2.41}$$

2.4.6 Distancias

Distancia entre dos puntos

Una de las nociones importantes empleadas en la teoría del análisis discriminante es la de distancia. Consideremos en un principio la distancia entre dos puntos en un espacio bivariado (X, Y). Desde un punto de vista geométrico, la distancia entre los puntos $A(x_1, y_1)$ y $B(x_2, y_2)$ que denotaremos por d_{AB} es el segmento de recta que une los puntos A y B tal como se muestra en la figura 2.1 .

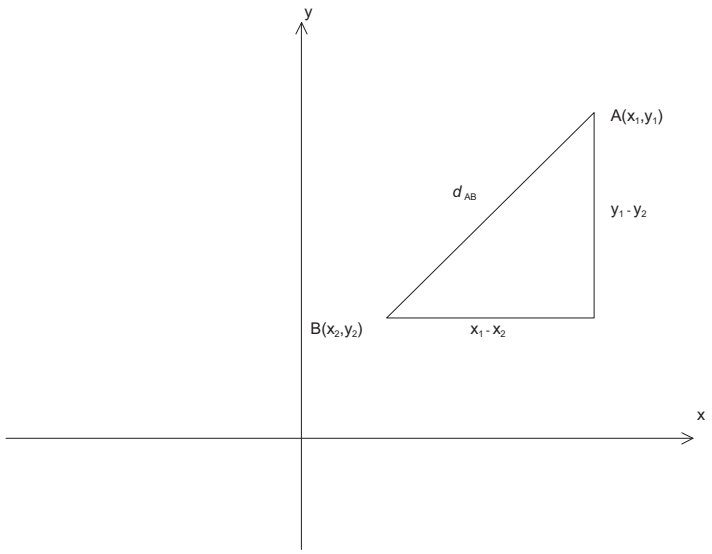


Figura 2.1

Por el teorema de Pitágoras, tenemos que

$$d_{AB}^2 = (x_2 - x_1)^2 + (y_1 - y_2)^2$$

que es el la medida euclidiana de la distancia con la que todos estamos familiarizados. Esta medida puede ser expresada en términos de vectores de la siguiente manera, considerando los vectores

$$\mathbf{a} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \text{ y } \mathbf{b} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$$

tenemos entonces

$$\begin{aligned} d_{AB}^2 &= [\mathbf{a} - \mathbf{b}]'[\mathbf{a} - \mathbf{b}] \\ &= [x_1 - x_2, y_1 - y_2] \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix} \\ &= (x_2 - x_1)^2 + (y_1 - y_2)^2 \end{aligned}$$

Esta medida es apropiada si se asume que no existe correlación entre las dos variables y que la varianza de cada una de ellas es igual a 1.

Si extendemos la noción de medida euclidiana a un espacio $p-$dimensional tenemos que:

$$d^2_{AB} = [\mathbf{X}_A - \mathbf{X}_B]'[\mathbf{X}_A - \mathbf{X}_B] \text{ donde } \mathbf{X}_A \text{ y } \mathbf{X}_B \text{ son vectores de dimensión } (p \times 1) \quad (2.42)$$

al igual que en el caso bivariado, este índice se basa en el supuesto que las variables no estén correlacionadas y que todas ellas tengan varianza igual a uno, es decir $\mathbf{\Sigma} = \mathbf{I}$.

Para comparar distancias entre dos o más variables, es necesario que se haya utilizado la misma métrica para medir las variables, si no es el caso, esto puede obtenerse al dividir las mediciones por las desviaciones estándar correspondientes. La intercorrelación entre las variables también debe tomarse en cuenta.

Si consideramos entonces que las variables no tienen varianzas iguales y su correlación es diferente de cero, podemos utilizar el índice de distancia cuadrada desarrollado por Mahalanobis:

$$\mathbf{\Delta}^2_{AB} = [\mathbf{X}_A - \mathbf{X}_B]'\mathbf{\Sigma}^{-1}[\mathbf{X}_A - \mathbf{X}_B] \text{ donde } \mathbf{\Sigma} \text{ es la matriz de varianza covarianza de } A \text{ y } B \quad (2.43)$$

Cabe señalar que si $\mathbf{\Sigma} = \mathbf{I}$, tenemos $\mathbf{\Delta}^2_{AB} = d^2_{AB}$

Éste es un índice de distancia (cuadrada) entre los puntos A y B definidos por los vectores $\mathbf{X}_A$ y $\mathbf{X}_B$ respectivamente y se conoce como índice de **distancia generalizada** de *Mahalanobis*.

Distancia entre centroides

Otro de los índices de distancia desarrollado por Mahalanobis que es importante en análisis discriminante, es aquel en el que los puntos entre los que se mide la distancia son los vectores de medias, conocidos como **centroides**. El centroide para una población k se denota como:

$$\boldsymbol{\mu}'_k = [\mu_{1k}, \mu_{2k}, ..., \mu_{pk}] \text{ donde } \mu_{ik} \text{ es la media de la i-ésima variable de la población } k \quad (2.44)$$

Si tenemos dos poblaciones con centroides $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$ la distancia entre estos dos centroides está dada por

$$\mathbf{\Delta}_{12} = [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^{1/2} \quad (2.45)$$

donde Σ es la matriz de varianza covarianza común a las dos poblaciones, es decir, se asume que las matrices de varianza covarianza de las dos poblaciones son iguales. Esta distancia es interpretada como la "distancia entre las dos poblaciones".

Distancia entre un punto y un centroide

El tercer índice de distancia de Mahalanobis es aquel en el que se mide la distancia entre un punto (un vector de p observaciones, que representa a un individuo) y el centroide de una población dada. Supongamos que tenemos g poblaciones de interés, la distancia entre $\mathbf{X}_i$, vector de observaciones para el i-ésimo individuo, y $\boldsymbol{\mu}_k$ centroide de la población k, está dada por:

$$\Delta_{ik} = [(\mathbf{X}_i - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_k)]^{1/2} \tag{2.46}$$

donde Σ_k es la matriz de varianza covarianza de la población k

Esta medida es de gran interés en el contexto de análisis discriminante como se verá en el siguiente capítulo, la idea es clasificar un individuo en una población a la que esté más cercano, usando una medida que estandarice las variables y que elimine el efecto de la correlación.

2.5 Distribución Normal Multivariada

Muchas de las distribuciones univariadas continuas tienen distribuciones análogas para el caso multivariado, una de ellas es la distribución **Normal**, cuyo equivalente para el caso multivariado es como su nombre lo indica la distribución **Normal Multivariada**. La importancia de la distribución normal multivariada viene en parte del teorema del límite central, pero también del hecho que algunos datos multivariados pueden aproximarse por medio de esta distribución. La distribución Normal multivariada es una de las más utilizadas en análisis multivariado ya que muchos de los procedimientos que éste emplea se hacen bajo el supuesto de normalidad de los datos. Para los fines de este trabajo, se supondrá que los datos se distribuyen normalmente.

En esta sección se define la distribución Normal Multivariada y se presentan algunas de sus propiedades más importantes, la demostración de estas últimas se puede consultar en cualquier libro de análisis multivariado.[1] Así mismo se presentan algunas de las pruebas de significancia que se utilizan en el análisis discriminante.

[1] Por ejemplo: C. Chatfield A.J. Collins (1980) "Introduction to Multivariate Analysis", Ed. Chapman and Hall

2.5.1 Definición de la distribución Normal Multivariada

Definición 38 ***Distribución Normal Univariada*** *La función de densidad de una v.a. univariada con* ***distribución normal*** *se define como:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{2.47}$$

donde μ es la media de X y σ^2 su varianza, utilizaremos entonces la notación $X \sim N(\mu, \sigma^2)$.

La función de distribución de $X \sim N(\mu, \sigma^2)$ está dada por

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2} dx \tag{2.48}$$

A continuación se muestran las gráficas de las funciones de densidad y distribución para una v.a. $X \sim N(0,1)$

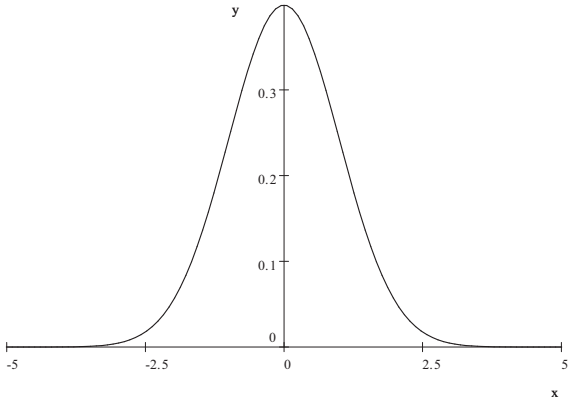


Figura 2.1: Función de densidad N(0,1)

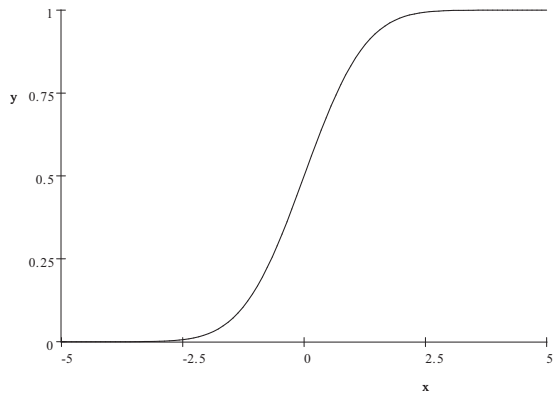


Figura 2.2: Función de distribución N(0,1)

La generalización de esta función de densidad para el caso multivariado da origen a la distribución Normal Multivariada

Definición 39 ***Distribución Normal Multivariada*** *Sea* $\mathbf{X}' = [X_1, X_2, ..., X_p]$ *una v.a. p-dimensional, se dice que* $\mathbf{X}$ *tiene una* ***distribución normal multivariada*** *si su función de densidad conjunta está dada por*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2.49}$$

donde $\boldsymbol{\mu}$ *es el vector de medias de* $\mathbf{X}$ *y* $\boldsymbol{\Sigma}$ *su matriz de varianzas covarianzas (que es una matriz simétrica positiva definida), utilizaremos entonces la notación* $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

De hecho, al igual que en el caso univariado, la función de densidad de una v.a. con distribución Normal multivariada está totalmente determinada por $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. Es interesante observar que la expresión $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ de (2.49) es el índice de distancia entre un punto y un centroide presentada en la sección anterior. Un valor pequeño de la varianza generalizada $|\boldsymbol{\Sigma}|$ indicará que las $\mathbf{x}$ están concentradas al rededor de $\boldsymbol{\mu}$ o que las variables están altamente correlacionadas. Si todas las correlaciones son iguales a cero, las p componentes de $\mathbf{X}$ son independientes.

Si las p variables aleatorias son independientes, tenemos $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)$ y $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ & & \\ 0 & 0 & \sigma_p^2 \end{bmatrix}$

por lo que la función de densidad conjunta de $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ puede reescribirse de la forma

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\prod_{i=1}^{p}\sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{p}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

2.5.2 Estimación de parámetros

Para una variable $p-$dimensional $\mathbf{X}$ con distribución Normal Multivariada, es posible demostrar que

$$E(\mathbf{X}) = \boldsymbol{\mu} \text{ y } \boldsymbol{\Sigma} \text{ es su matriz de varianza covarianza}$$

Los parámetros de una normal multivariada tienen entonces una interpretación inmediata, sin embargo, en la práctica estos parámetros son desconocidos, por lo que surge la necesidad de estimarlos.

Recordemos primero que un estimador de un parámetro θ es **insesgado** si la media de su distribución muestral es θ, es decir si $E[\hat{\theta}] = \theta$.

Recordemos también que la **función de verosimilitud** es la función de densidad conjunta definida como función del parámetro desconocido θ :

$$\begin{aligned} f(X_1, X_2, ...X_p, \theta) &= \prod_{i=1}^{n} f(X_i, \theta) \\ &= L(\theta \mid \mathbf{X}) \text{ donde } \mathbf{X} \text{ representa los datos muestrales} \end{aligned}$$

El valor que maximiza esta función es el **estimador máximo verosímil** de θ, denotado generalmente por $\hat{\theta}$.

La obtención de estos estimadores así como sus propiedades y otros estimadores pueden encontrarse en cualquier texto de probabilidad y estadística.[1]

Dada una muestra aleatoria de n observaciones independientes de un vector aleatorio $\mathbf{X}$, el vector $\overline{\mathbf{X}}$ y la matriz de varianza covarianza muestral $\mathbf{S}$ son estimadores insesgados de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. Cuando $\mathbf{X}$ tiene una distribución normal multivariada, puede demostrarse que $\overline{\mathbf{X}}$ es el estimador máximo verosímil de $\boldsymbol{\mu}$ y que $[(n-1)/n]\mathbf{S}$ es el estimador máximo verosímil de $\boldsymbol{\Sigma}$ y es sesgado. En este trabajo, utilizaremos como estimadores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ los estimadores insesgados $\overline{\mathbf{X}}$ y $\mathbf{S}$.

[1] Por ejemplo: Mood, Graybill and Boes (1974) "Introduction to the theory of Statistics" 3^a Edición, Ed. Mc Graw Hill

Además existe un estimador que es muy importante para el desarrollo del análisis discriminante, la **matriz de varianzas covarianzas ponderada** (**pooled**). Esta matriz es una medida de dispersión general entre los datos en la que se involucra la información de todas las poblaciones. Supongamos que tenemos dos poblaciones $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$ con un total de n_1 y n_2 observaciones respectivamente, las medias de las dos poblaciones son distintas $\mu_1 \neq \mu_2$ pero sus varianzas son iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Podemos obtener los estimadores de la varianza para cada una de las dos poblaciones S_1^2 y S_2^2. De aquí, podemos obtener otra medida que involucre la información de las dos poblaciones, conocida como **estimador combinado de la varianza**, lo denotaremos por S_{pl}^2 y está dado por:

$$S_{pl}^2 = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X_1}) + \sum_{i=1}^{n}(X_{2i} - \overline{X_2})}{n_1 + n_2 - 2} \tag{2.50}$$

Este estimador también puede calcularse como una ponderación de las varianzas de cada población, es decir

$$\begin{aligned} \mathbf{S}_{pl} &= \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \\ &= \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 - 1) + (n_2 - 1)} \end{aligned} \tag{2.51}$$

En necesario que $(n_1 + n_2 - 2)) > 0$, para que la matriz $\mathbf{S}_{pl}$ tenga inversa.

2.5.3 Algunas Propiedades de la distribución Normal Multivariada

Cerradura bajo transformaciones lineales

Una definición alternativa de la distribución normal multivariada es:

Definición 40 *Se dice que una variable aleatoria $p-$dimensional* $\mathbf{X}$ *tiene una distribución normal multivariada si y sólo si cada componente lineal de* $\mathbf{X}$ *tiene una distribución normal univariada.*

Una consecuencia inmediata de esta definición es el hecho que la normalidad multivariada se conserva bajo transformaciones lineales. Si $\mathbf{Y} = \mathbf{A}'\mathbf{X}$ donde $\mathbf{A}$ es una matriz de constantes de orden $(m \times p)$ y $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces

$$\mathbf{Y} \sim N_m(\mathbf{A}'\boldsymbol{\mu}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}) \tag{2.52}$$

dado que cualquier componente de $\mathbf{Y}$ es un componente lineal de $\mathbf{X}$ y por lo tanto tiene una distribución normal. La media y la varianza se obtienen de las ecuaciones $(2.40, 2.41)$.

Estandarización de Variables

En el caso univariado, es común transformar una variable aleatoria $X \sim N(\mu, \sigma^2)$ para obtener una variable con distribución $N(0, 1)$, para el caso multivariado, la transformación está dada por

$$\begin{aligned} \mathbf{U} &= \mathbf{B}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \\ \mathbf{U} &\sim N_m(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{2.53}$$

donde $\mathbf{B}$ es una matriz de dimensión $(p \times m)$ de rango completo, tal que $\mathbf{B}\mathbf{B}' = \boldsymbol{\Sigma}$, esto sólo es válido si $\boldsymbol{\Sigma}$ es de rango completo.

Distribución Ji-cuadrada

Se dice que una variable aleatoria X tiene distribución Ji-cuadrada con p grados de libertad $(X \sim \chi^2_{(p)})$, si está definida como la suma de cuadrados de p variables aleatorias con distribución normal estándar $N(0, 1)$.

Entonces, si $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma}$ de rango completo, se tiene que

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(p) \tag{2.54}$$

La función de densidad de una variable Ji-cuadrada se muestra en el Apéndice B.

Normalidad de las distribuciones marginales, independencia

Supongamos que el vector aleatorio $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ está particionado de la siguiente manera

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \text{ donde } \mathbf{X}_1 \text{ es un vector de } (q \times 1), q < p \tag{2.55}$$

con las correspondientes particiones

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2] \text{ y } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Entonces, tenemos las siguientes propiedades:

1. La distribución marginal de $\mathbf{X}_1$ es $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$

   En particular la distribución marginal de cualquier elemento de $\mathbf{X}$ tiene una distribución normal univariada, sin embargo el hecho que todos los componentes de vector aleatorio tengan una distribución normal no implica que el vector aleatorio tendrá una distribución normal multivariada.

2. $\mathbf{X}_1$ y $\mathbf{X}_2$ son independientes si y sólo si $\boldsymbol{\Sigma}_{12} = \mathbf{0}$

   ($\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$ y su dimensión es $(q \times (p - q))$).

Normalidad de la distribución condicional

Supongamos que el vector aleatorio $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ está particionado al igual que en (2.55), si $\mathbf{X}_1$ y $\mathbf{X}_2$ no son independientes es decir $\boldsymbol{\Sigma}_{12} \neq \mathbf{0}$, la distribución condicional de $\mathbf{X}_2$ dado $\mathbf{X}_1$, $f_{(\mathbf{X}_2|\mathbf{X}_1)}$ es una normal multivariada con $E[\mathbf{X}_2 \mid \mathbf{X}_1] = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1)$ y $cov[\mathbf{X}_2 \mid \mathbf{X}_1] = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$

Distribución de la media muestral

Sea $\mathbf{V} = \sum_{k=1}^{n} c_k \mathbf{X}_k$, donde $\mathbf{X}_k$ son variables aleatorias $p-$dimensionales y c_k son constantes, si las $\mathbf{X}_k$ están distribuidas normalmente, $\mathbf{V}$ también lo estará.

Supongamos que las $\mathbf{X}_k$ son independientes y tienen una distribución $N_p \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces la media muestral de una muestra aleatoria de n observaciones independientes está dada por

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k$$

y tenemos que

$$\overline{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \tag{2.56}$$

2.5.4 Pruebas de Hipótesis

En análisis multivariado, al igual que en el caso univariado, existen pruebas con respecto a las medias para una o más poblaciones y pruebas con respecto a la varianza. En esta sección se presentarán la prueba T^2 de Hotelling (que es el equivalente multivariado de la prueba t), la Λ de Wilks y la prueba M de Box, que son utilizadas en análisis discriminante.

Distribuciones F y t de Student

Las distribuciones F y t de Student juegan un papel esencial en el desarrollo de pruebas de hipótesis.

Recordemos que una v.a. X tiene **distribución** F con m y n grados de libertad, si está definida de la siguiente manera:

$$X = \frac{U/m}{V/n} \text{ donde } U \sim \chi^2(m) \text{ y } V \sim \chi^2(n) \tag{2.57}$$

$$X \sim F(m, n)$$

De mismo modo, una v.a. Y tiene **distribución** t **de Student** con k grados de libertad si está dada por

$$Y = \frac{Z}{\sqrt{U/k}} \text{ donde } Z \sim N(0,1) \text{ y } U \sim \chi^2(k) \tag{2.58}$$

$$Y \sim t(k)$$

Las funciones de densidad de estas variables se muestran en el Apéndice B.

Prueba respecto a la diferencia de medias (T^2 de Hotelling)

Primero consideremos el caso de la comparación de dos grupos. Supongamos que se realiza la medición de una v.a. univariada X para dos grupos diferentes X_1 y X_2 con tamaños de muestra n_1 y n_2 , obtenemos de esta manera dos medias $\overline{X_1}$ y $\overline{X_2}$ y podemos definir un estadístico t dado por

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(1/n_1 + 1/n_2)}S_{pl}} \tag{2.59}$$

donde S_{pl} es el estimador combinado de la varianza (2.50).

Este estimador puede compararse con una variable con distribución t de Student con $n_1 + n_2 - 2$ grados de libertad, es decir

$$t^2 = \frac{\left(\overline{X}_1 - \overline{X}_2\right)^2}{(1/n_1 + 1/n_2)S_{pl}^2}$$

Supongamos ahora que se obtienen dos muestras aleatorias de observaciones sobre un vector aleatorio p-dimensional $\mathbf{X}$, asumiremos que para las dos muestras se tiene una distribución normal multivariada con matriz de covarianza (desconocida) $\mathbf{\Sigma}$ de rango completo p. Deseamos realizar una prueba con hipótesis nula de igualdad en los vectores de medias, contra su alternativa de medias diferentes.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \; vs \; H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Si tenemos vectores $p-$dimensionales de medición con medias $\overline{\mathbf{X}}_1$ y $\overline{\mathbf{X}}_2$ respectivamente, el estadístico multivariado correspondiente a esta prueba está dado por:

$$T^2 = \frac{n_1 n_2}{n_{1+} n_2}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \mathbf{S}_{pl}^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \tag{2.60}$$

donde $\mathbf{S}_{pl}$ es el estimador ponderado de la varianza covarianza entre los grupos (2.51). Este estadístico se conoce como la T^2 **de Hotelling**. Es posible demostrar que

$$\frac{n_{1+} n_2 - p - 1}{(n_{1+} n_2 - 2)p} T^2 \tag{2.61}$$

tiene una distribución F con p y $(n_{1+}n_2 - p - 1)$ grados de libertad bajo la hipótesis de que no hay diferencia entre los grupos.

La regla de decisión para la prueba a un nivel de significancia α es de la forma:

Aceptar $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ si

$$T^2 \leq \frac{(n_{1+}n_2 - 2)p}{n_{1+}n_2 - p - 1} F^p_{n_{1+}n_2-p-1}(\alpha) \tag{2.62}$$

de lo contrario rechazar H_0

Lambda Wilks

En el caso univariado, para realizar el contraste de igualdad de medias para k poblaciones normales ($H_o : \mu_1 = \mu_2 = ... = \mu_k$), se utiliza el estadístico F (F ratio) dado por

$$F = \frac{\sum_{j=1}^{k} n_j(\bar{x}_{j.} - \bar{x})^2/(k-1)}{\sum_j \sum_i (x_{ji} - \bar{x}_{j.})^2/(n-k)}$$

Este estadístico tiene distribución F con $(k-1)$ y $(n-k)$ grados de libertad. La hipótesis de igualdad de medias se rechazará entonces si $F > c$, donde c es una constante tal que $P[F \geq c \mid H_o] = \alpha$. Este contraste se conoce generalmente como one-way-ANOVA.

El análisis multivariado de la varianza (MANOVA) se realiza en términos de las diferencias entre los vectores de medias o centroides. Para una v.a. $p-$dimensional, el centroide de la población k se denota por $\boldsymbol{\mu}_k$ y está definido como $\boldsymbol{\mu}'_k = [\mu_{1k}, \mu_{2k}, ..., \mu_{ik}, ..., \mu_{pk}]$. La hipótesis que queremos probar es:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g$$

Supondremos que las observaciones para las g poblaciones son independientes, tienen una distribución normal y sus matrices de varianza covarianza son idénticas.

El estadístico más común para realizar esta prueba es la **Lambda de Wilks** que se define como:

$$\boldsymbol{\Lambda} = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} \tag{2.63}$$

donde $\mathbf{W}$ es la matriz de sumas de cuadrados y productos cruzados al interior de los grupos, y $\mathbf{B}$ es la matriz de suma de cuadrados y productos cruzados de la hipótesis (o intergrupos),

tenemos entonces:

$$\begin{aligned} \mathbf{W} &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{X}_{ki}-\overline{\mathbf{X}}_k)(\mathbf{X}_{ki}-\overline{\mathbf{X}}_k)' \\ \mathbf{W} &= \mathbf{S}_1+\mathbf{S}_2+...+\mathbf{S}_g \end{aligned} \tag{2.64}$$

donde $\mathbf{S}_k$ es la matriz de sumas de cuadrados y productos cruzados totales de la $k-$ésima muestra, la matriz $\mathbf{W}$ es una generalización de la $\mathbf{S}_{pl}$ descrita anteriormente.

Por su parte la variación entre los grupos está dada por

$$\mathbf{B}=\sum_{k=1}^{g}(\mathbf{X}_k-\overline{\mathbf{X}})(\mathbf{X}_k-\overline{\mathbf{X}})' \tag{2.65}$$

También es posible expresar el estadístico Lambda en función de los eigenvalores de la matriz $\mathbf{W}^{-1}\mathbf{B}$ de la siguiente manera:

$$\mathbf{\Lambda}=\prod_{j=1}^{q}\frac{1}{1+\lambda_j} \text{ donde } \lambda_j, j=1,..,q \text{ son los eigenvalores de } \mathbf{W}^{-1}\mathbf{B} \tag{2.66}$$

La Lambda de Wilks para un conjunto de p variables mide las desviaciones dentro de cada grupo respecto a las desviaciones globales, sin diferenciar entre grupos. Los parámetros de este estadístico son p, $q-1$ y $n-q$, generalmente toma valores entre 0 y 1 y se rechaza la hipótesis para valores pequeños.

El valor de Lambda puede transformarse en un estadístico multivariante F, que permite contrastar la existencia de diferencias entre grupos.

Los puntos críticos exactos pueden encontrarse bajo ciertas condiciones (Schatzoff ,1966), sin embargo es más común utilizar una transformación de $\mathbf{\Lambda}$ que tiene una distribución aproximada a la Ji-cuadrada, esta transformación está dada por:

$$\mathbf{V}=-\left[n-1-\frac{(p+g)}{2}\right]\ln\mathbf{\Lambda} \tag{2.67}$$

y se aproxima a la distribución Ji-cuadrada con $p(g-1)$ grados de libertad (Bartlett, 1947).

Prueba M de Box

Algunas técnicas multivariadas como el análisis discriminante se realizan bajo el supuesto de homocedasticidad, es decir que las matrices de varianza covarianza son iguales entre los grupos. La hipótesis que se quiere probar es entonces:

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = ... = \mathbf{\Sigma}_g$$

El contraste de esta hipótesis se realiza a través del estadístico **M de Box** que está dado por:

$$M = (n - g) \log |\mathbf{S}| - \sum_{k=1}^{g} (n_k - 1) \log |\mathbf{S}_k|$$

donde S es la matriz de varianza covarianza combinada y $\mathbf{S}_k$ es la matriz de varianza covarianza del $k-$ésimo grupo. El estadístico M puede transformarse en un estadístico F para facilitar su interpretación, sin embargo, cabe señalar que este estadístico es muy sensible a pequeñas desviaciones de la normalidad multivariada y a tamaños de muestra grandes.[2]

[2]La transformación del estadístico M, en un estadístico con distribución F o Ji-cuadrada puede ser consultada en Harris " A primer of multivariate statistics" Ed Orlando Fl Academic Press 1985.

Capítulo 3

Análisis Discriminante

3.1 Introducción

El ser humano, por su naturaleza, siempre busca incorporar cierto tipo de información para conocer la realidad del mundo que lo rodea. Para poder conocer mejor la realidad, es necesario construir una representación limitada de ésta, reducida a elementos definidos y ordenados, que permita su manejo intelectual. Es decir podemos comprender mejor la realidad al dividirla en clases de elementos equivalentes entre sí y diferenciados de aquellos incluidos en otras clases. Uno de los procedimientos básicos que el hombre realiza para el conocimiento de la realidad es entonces la clasificación, que juega un papel de suma importancia en diversas áreas de la ciencia por ejemplo la química, la biología y la medicina.

La palabra clasificación puede tener varios significados de acuerdo con la naturaleza del problema que se está tratando. El primer tipo de problema asociado a la clasificación, conocido como agrupamiento (clustering), es aquel en el que a partir de una única población, se busca clasificar a los individuos u objetos de dicha población en grupos lo mejor diferenciados posible. De la misma manera, existen problemas en los que el objetivo es dividir en grupos una única población, sea homogénea o no, por medio de delimitaciones naturales o no naturales. Finalmente, existen problemas asociados a la clasificación conocidos como problemas de discriminación, en ellos se supone la existencia de grupos o poblaciones definidos a priori, de los que han sido extraídas las muestras. El problema consiste entonces en la adscripción de un nuevo individuo, cuyo grupo de origen es desconocido, a alguno de los grupos previamente definidos.

Para cierto tipo de problemas de clasificación, la percepción humana es suficiente y las técnicas matemáticas no son necesarias, sin embargo es necesario emplear estas técnicas cuando

las diferencias entre los grupos no son evidentes y se caracterizan por un gran número de variables. Además por medio de estas técnicas, se asegura que la clasificación no sea hecha por intuición y se elimina toda subjetividad. Existen técnicas estadísticas multivariantes para la clasificación de objetos o individuos, tales como el análisis de conglomerados (análisis cluster) o el análisis discriminante. La diferencia entre estos, es que en el análisis de conglomerados, se utiliza la información (variables) de los individuos para determinar clases lo mejor separadas entre sí, que incluyan a los individuos en cuestión; es decir que los grupos o clases surgen como resultado del análisis. En cambio en el análisis discriminante, los grupos o clases están definidos previamente y el objetivo es clasificar nuevos individuos en alguno de los grupos definidos a priori por medio de reglas basadas en la información de los individuos ya clasificados.

El principal objetivo del análisis discriminante es definir reglas de clasificación que permitan predecir el grupo o población de la que sea más probable que haya sido extraído un individuo, basándose en ciertas características conocidas de éste. Los problemas de discriminación suponen entonces la formulación de pronósticos o predicciones, a partir de las características del individuo u objeto, por lo que también hay que tomar en cuenta el error puede generarse al hacer un mal pronóstico y las consecuencias que esto puede tener.

Los orígenes de lo que hoy conocemos como análisis discriminante se remontan a los años veinte, cuando el estadista inglés Karl Pearson propuso un "índice de parecido racial" que era una especie de distancia entre grupos. Fue también en los años veintes que Mahalanobis empezó a estudiar en la India otro índice de distancia entre grupos, que quedaría formalizado en los años treinta. A partir de esta idea de distancia, R.A. Fisher desarrollaría la idea de una combinación lineal de variables para la discriminación entre grupos. En 1936 fue publicado en el *Annals of Eugenics*, el artículo pionero de Fisher, titulado "The use of multiple measurments in taxonomic problems", en él se presentaban tanto la idea de distancia entre grupos como la de combinación lineal de variables para la discriminación entre grupos y el artículo estaba enfocado a la resolución de problemas de antropología física y biología. Desde los años cuarenta, se han publicado múltiples extensiones y refinamientos de las ideas de Fisher, los primeros trabajos estaban básicamente enfocados a la predicción de la pertenencia entre grupos, pero a partir de los años sesenta surgió el interés por interpretar los efectos observados a través de un análisis multivariado de la varianza (MANOVA). Si bien las primeras aplicaciones del análisis discriminante se enfocaron a problemas relacionados con la medicina y la biología, hoy en día el análisis discriminante tiene muchas más áreas de aplicación tales como la educación, la psicología, el marketing, el análisis financiero, la agronomía, la arqueología etc. Cabe señalar también que gracias al desarrollo informático y de los paquetes de estadística, la aplicación del análisis discriminante se ha vuelto mucho sofisticada, permitiendo además manejar grandes cantidades de información de manera sencilla.

Si bien las palabras discriminación o discriminante pueden tener una connotación negativa o agresiva, la técnica del análisis discriminante trata con problemas de cuya solución suele obtenerse algún beneficio o progreso, sin buscar el perjuicio de alguien o algo. A continuación se presentan algunos ejemplos en los que ha sido aplicado el análisis discriminante.

- En educación, se han planteado problemas en los que se desea pronosticar si un alumno tendrá éxito académico o no, con el fin de establecer un plan de apoyo y refuerzo de estudio para aquellos alumnos a los que no se les pronostique éxito académico. En caso de que el pronóstico sea erróneo, se puede incurrir en diferentes tipos de errores con consecuencias distintas, como por ejemplo brindar apoyo a un alumno que en realidad no lo necesitaba o no brindarlo a un alumno que en realidad si lo requería, siendo ésta última una consecuencia quizás más grave.

- En el caso de la medicina, supongamos que un médico desea saber si un anestésico es seguro para un paciente o no lo es, basado en ciertas características conocidas de éste último. Si un anestésico es diagnosticado como inseguro para un paciente para el que en realidad es seguro, las consecuencias pueden no ser tan graves siempre y cuando se cuente con otro anestésico seguro para el paciente. Sin embargo, las consecuencias pueden ser nefastas si se diagnostica como seguro un anestésico para un paciente para él que en realidad no lo es.

- También existen diferentes aplicaciones del análisis discriminante en el ámbito financiero, por ejemplo estimar la viabilidad de un crédito. Es decir se busca clasificar a los clientes en el grupo de aquellos que son solventes y por lo tanto pagarán el importe de su deuda o en el grupo de aquellos que incurrirán en impagos y a los que por lo tanto no debería otorgarse el crédito. Las consecuencias de una mala clasificación de los clientes resultan entonces evidentes.

Una vez adquiridos los conocimientos básicos de álgebra lineal y análisis multivariado, es posible abordar la teoría del análisis discriminante, en sus dos grandes vertientes, el análisis discriminante descriptivo y el análisis discriminante predictivo. En este capítulo abordaremos diferentes aspectos de esta teoría tales como su objetivo y planteamiento del problema, las funciones discriminantes, las reglas de clasificación y las pruebas de significancia estadística empleadas en esta técnica. Finalmente, en el siguiente capítulo se realizará una aplicación de esta técnica multivariada al nivel de ingreso de los hogares.

3.2 Objetivo y planteamiento del problema

Como ya se había mencionado, el análisis discriminante es una técnica multivariada que busca obtener reglas para la clasificación de un nuevo individuo en la población o grupo a la que pertenezca más probablemente. Esto supone la existencia de g grupos o poblaciones previamente definidos que sean mutuamente excluyentes. Es decir cada individuo pertenece a uno y sólo uno de los grupos. La clasificación de un nuevo individuo se hará a través de ciertas características conocidas de éste. Estas características deberán ser las mismas que aquellas de los individuos ya clasificados, utilizadas para obtener las reglas de clasificación. Es decir el análisis discriminante pone en relación (por medio de una combinación lineal de variables) una variable categórica dependiente (la pertenencia a alguno de los g grupos) con una serie de variables continuas independientes (características del individuo) que serán llamadas variables discriminantes.

En este sentido, el análisis discriminante puede compararse con la regresión lineal, a excepción que la variable dependiente es categórica y no continua. En el caso de que la variable dependiente tenga únicamente dos categorías, el análisis discriminante podría compararse con la regresión logística y sería interesante comparar los resultados arrojados para un mismo problema por medio de estas dos técnicas.

En un problema típico de análisis discriminante, se cuenta con una muestra de n individuos, clasificados en g grupos mutuamente excluyentes que forman una partición de la muestra. Para cada uno de estos individuos, se tienen las observaciones de p características. Por medio de esta información, se tratará de obtener una función discriminante que permita la adscripción de nuevos individuos al grupo al que pertenezcan con mayor probabilidad. La variable dependiente categórica tendrá tantos valores discretos como grupos y a lo largo del trabajo será denotada por D_i, donde i indica que se trata del grupo del i-ésimo individuo. Las variables independientes o discriminantes por medio de las cuales suponemos se diferencian los grupos se notarán como $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_j, ..., \mathbf{X}_p$, de esta manera X_{ij} representará la observación de la j-ésima variable para el i-ésimo individuo . Si tenemos una muestra de n individuos, contaremos con n observaciones de cada una de las variables y tendremos g grupos $g_1, g_2, ..., g_k, ...g_g$ con tamaños iguales a $n_1, n_2, ..., n_k, ..., n_g$ donde $\sum_{k=1}^{g} n_k = n$. El problema principal radica entonces en encontrar los pesos de las variables discriminantes en una combinación lineal que permita asignar un grupo a un nuevo individuo con el menor error posible, esta combinación lineal de variables se conoce como función discriminante y se detallará en la siguiente sección.

La solución de un problema de este tipo puede tener dos propósitos diferentes, que son a la vez las dos vertientes del análisis discriminante, es decir el análisis discriminante descriptivo y el predictivo. Como su nombre lo indica, el análisis discriminante descriptivo está enfocado a la descripción de las diferencias entre los grupos, con el objetivo de determinar en que medida las

variables discriminantes contribuyen a la diferenciación entre grupos y cuáles son las variables que tienen mayor poder de discriminación. Por su parte, el análisis discriminante predictivo está enfocado a la obtención de funciones discriminantes que permitan la clasificación de nuevos individuos con el menor error posible.

Con el planteamiento de un problema tal, surgen diferentes cuestiones que hay que tomar en cuenta como cuáles son las variables que hay que incluir en el modelo y cómo seleccionarlas, cuáles son las más representativas, como clasificar a un individuo con el menor error posible y bajo que reglas realizar la clasificación. Estas cuestiones propias del análisis discriminante se tratarán en las siguientes secciones.

Como todas las técnicas de análisis multivariado, la teoría del análisis discriminante se basa en el cumplimiento de ciertos supuestos que teóricamente deberían ser comprobados antes de la aplicación de la técnica, pero cuya comprobación muchas veces resulta muy complicada y simplemente se dan por hecho. A continuación se presentan estos supuestos, así como algunas alternativas para su comprobación y algunas consecuencias de la violación de estos.

- Distribución Normal Multivariada

La aplicación del análisis discriminante supone que los datos con los que se trabaja representan una muestra aleatoria de una población con distribución Normal Multivariada. Es preciso que esto se cumpla para que puedan obtenerse correctamente las probabilidades de pertenencia a un grupo así como para que puedan ser empleadas las diferentes pruebas de significancia estadística empleadas en el análisis discriminante. Estas pruebas utilizan estadísticos que siguen un modelo teórico de distribución, y si el supuesto de normalidad multivariante se viola, el estadístico tendrá una distribución diferente a la del modelo.

Sin embargo la comprobación de este supuesto resulta bastante difícil, por lo que en la mayoría de los casos el supuesto queda sin comprobar. Si bien existen algunos métodos empíricos o gráficos para la comprobación de este supuesto (véase por ejemplo Stevens, 1986; Krzanowski, 1990) por lo general son difíciles de implementar. Cabe señalar que si una variable tiene distribución normal multivariada, todas las distribuciones marginales son normales univariadas, si bien el inverso de esta relación es falso, el hecho de que las variables por separado tengan una distribución normal univariada aumenta la probabilidad de que su distribución conjunta sea normal multivariada. Por ello, sería recomendable realizar pruebas de bondad de ajuste a cada variable por separado para comprobar su normalidad (por ejemplo la prueba Ji-cuadrada o la prueba de Kolmogorov-Smirnov).

Los resultados del análisis discriminante no se ven seriamente afectados si la distribución de las variables discriminantes no se aparta demasiado de la normal multivariada, aún menos si se trabaja con muestras muy grandes.

- Igualdad de las matrices de varianza covarianza

Otro supuesto del análisis discriminante es que las matrices de varianza covarianza para las poblaciones de las que fueron extraídos los grupos deben ser iguales. Este supuesto resulta bastante severo, y en la práctica raramente se cumple ya que los tamaños de las matrices son muy grandes y para que estas fueran iguales cada uno de sus elementos debería ser igual. Existen varias pruebas para comprobar la igualdad de matrices de varianza-covarianza como por ejemplo la prueba M de Box, descrita en el capítulo anterior. Al trabajar con muestras muy grandes, la diferencia entre las matrices de varianza covarianza suele resultar significativa aunque éstas no difieran en gran medida. Es así mismo importante tomar en cuenta que esta prueba es sensible a la violación del supuesto de normalidad multivariada.

- Linealidad

El modelo del análisis discriminante asume relaciones lineales entre las variables, lo que implica que existen relaciones lineales entre las variables de cada grupo, que pueden ser representadas por una recta, por lo que los diagramas de dispersión pueden ser útiles para comprobar este supuesto. Otro método para comprobar el cumplimiento de este supuesto está dado mediante el cálculo de los coeficientes de correlación lineal de Pearson.

- Ausencia de multicolinealidad y singularidad

La multicolinealidad ocurre cuando dos variables de la matriz de correlaciones muestran una correlación perfecta y tienen el mismo patrón de correlaciones con el resto de las variables, lo que indicaría que las variables tienen el mismo comportamiento y aportan información redundante. La singularidad se da cuando las puntuaciones alcanzadas en una variable son aproximadamente una combinación lineal del resto de las variables, es decir cuando existe una correlación múltiple entre una variable y las demás. De nueva cuenta una variable con estas características estaría aportando información redundante. De esta manera, si existe multicolinealidad o singularidad, no puede realizarse la inversión de la matriz de varianza covarianza ya que el determinante sería igual a cero. Esta inversión es necesaria para la obtención de los coeficientes de la función discriminante, como se verá más adelante, por lo que es importante que el supuesto de ausencia de multicolinealidad y singularidad se cumpla. En caso no cumplirse este supuesto, la solución más evidente es eliminar la o las variables en cuestión. Existen diferentes métodos para detectar multicolinealidad o singularidad en las matrices por ejemplo se puede obtener la matriz de correlaciones de Pearson y de arrojar valores cercanos a 0.99, esto nos indicaría la presencia de variables redundantes.

Además de estos supuestos, hay que tener en cuenta algunas consideraciones para la correcta aplicación del análisis discriminante.

Por ejemplo, el análisis discriminante supone que las variables independientes sean variables continuas, sin embargo en la práctica es muy común utilizar variables categóricas, que en caso de ser de más de dos categorías deberán recodificarse en variables dicotómicas que podrán ser incluidas en el modelo siempre y cuando la distribución muestral de la media sea normal. De acuerdo con el teorema del límite central, siempre que el tamaño de la muestra sea grande, la distribución muestral de la media es normal, por lo que para muestras grandes las variables dicotómicas pueden ser empleadas sin mayor problema.

Teóricamente, no existe límite para el número de variables independientes o discriminantes, pero este debe no debe ser mayor al número de casos que se tengan para el grupo más pequeño. En la bibliografía del análisis discriminante, es común encontrar la recomendación de que es preferible contar con al menos 20 casos por cada variable independiente incluida en el modelo, para así obtener conclusiones correctas.

En lo que respecta al tamaño de los grupos, no existe inconveniente en que estos difieran en su tamaño, de ser así únicamente habría que tomar en cuenta esto al momento de realizar la clasificación de los sujetos, al asignar probabilidades a priori diferentes a cada grupo de acuerdo con su tamaño.

Al igual que en otras técnicas estadísticas, deben tomarse en cuenta ciertas consideraciones sobre los datos tales como los datos desaparecidos o "missing data" y los caso aislados o "outliers". En caso de los missing data, si estos se presentan en la variable dependiente, los sujetos en cuestión pueden separarse y al finalizar el análisis podrán retomarse para así predecir su grupo de pertenencia. Si en cambio alguna de las variables discriminantes presenta datos desparecidos, podemos recurrir diferentes procedimientos (por ejemplo substitución por la media o regresión) para estimar estos valores, siempre y cuando los casos en los que se presenten no tengan un comportamiento diferente a los del resto de la muestra.

Una exploración previa de los datos puede tener como resultado la identificación de casos aislados, para el caso de datos multivariados, es común recurrir a la distancia de Mahalanobis entre un individuo y el centro de su grupo para la detección de estos casos. De encontrarse tales casos, habrá que tener cuidado en los efectos que pudieran tener en el análisis, de nueva cuenta al contar con muestras grandes el efecto de uno o dos casos aislados no tendrá gran impacto sobre los resultados.

3.3 Funciones Discriminantes

El objetivo básico del análisis discriminante es la obtención de funciones lineales, conocidas como funciones discriminantes que permitan tanto clasificar a los individuos en uno de los grupos definidos, como describir las diferencias entre esos grupos. Las funciones discriminantes tienen entonces un papel fundamental para el desarrollo de esta técnica. En esta sección se definirán estas funciones, así como la interpretación geométrica que se les puede dar. También se establecerá la manera en la que se obtienen los coeficientes de estas funciones.

3.3.1 Definición

Una **función discriminante** es una combinación lineal de variables discriminantes que permite la asignación de cada individuo a uno de los grupos previamente definidos, con el menor error posible. Supongamos que tenemos p variables independientes $(X_1, X_2, ..., X_p)$ medidas para n individuos divididos en g grupos preestablecidos. La función discriminante o función lineal discriminante para el i-ésimo individuo del grupo k está entonces dada por:

$$Y_{ki} = a_0 + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_p X_{ip} \tag{3.1}$$

donde:

Y_{ki} es el valor de la función discriminante para el *í*-ésimo individuo del grupo k. Este valor es también conocido como **puntuación discriminante.**

X_{ij} es el valor de la j-ésima variable para el i-ésimo individuo

a_j , $j = 1, ..., p$ es la **ponderación** o **coeficiente** de la j-ésima variable (a_0 es constante)

La asignación de los valores a_j, como se detallará más adelante, deberá de ser de tal manera que la variación dentro de los grupos (within) sea mínima y la variación entre los grupos (between) sea máxima. Los coeficientes de la primera función discriminante que se determinen, deben de ser de tal forma que las puntuaciones medias de cada grupo sean los más distintas posible. La segunda función seguirá el mismo criterio además de que sus valores no deben estar correlacionados con los obtenidos en para la primer función. Las funciones discriminantes sucesivas se obtendrán de manera similar.

Si se tienen p variables independientes y g grupos, el número máximo de funciones discriminantes que podremos obtener, denotado por q, estará dado por $q = \min(p, g-1)$, como

veremos más adelante. Si aplicamos todas las funciones discriminantes a las observaciones de un individuo, obtendremos todas sus puntuaciones discriminantes, denotadas por $z_{1i}, z_{2i}, ..., z_{qi}$ y gracias a ellas se podrá asignar un grupo de pertenencia al individuo en cuestión.

3.3.2 Interpretación geométrica de las funciones discriminantes

Las observaciones de las p variables para los n individuos pueden interpretarse como las coordenadas de n puntos en un espacio p-dimensional. Por ejemplo si contáramos con una muestra de observaciones de tres variables para n individuos, estas podrían verse como las coordenadas de n puntos en un espacio tridimensional, cuyos ejes están formados por las tres variables discriminantes o independientes. Ahora bien, si partimos de la idea que los n individuos de la muestra se dividen en tres grupos diferentes definidos previamente, es de esperar que los individuos de cada grupo se comporten de manera similar y por lo tanto se localicen en una región determinada del espacio. Cada grupo puede ser caracterizado por su centroide, el cual puede también ser interpretado como un punto en el espacio, que caracterizará la posición del grupo en este espacio, como se muestra en la siguiente figura.

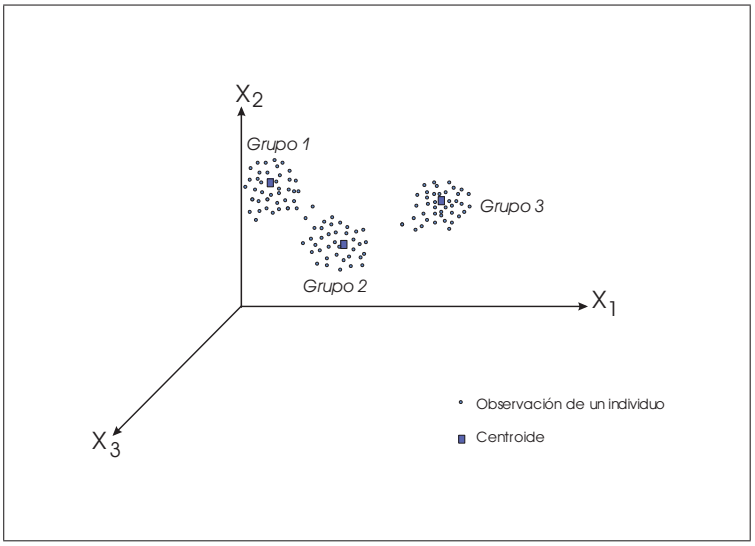


Figura 3.1 Observaciones y centroides de tres grupos, en un espacio definido por tres variables discriminantes

Para poder describir las diferencias de los grupos en función de las variables discriminantes, habrá que determinar, a través de la posición de sus centroides, si los grupos están lo suficientemente diferenciados en el espacio p dimensional. Las diferencias entre las posiciones

relativas de los centroides no necesariamente tienen que representarse en un espacio de dimensión p, de hecho los centroides definen un espacio de dimensión $(g-1)$ donde g es el número de centroides o grupos definidos previamente.

De esta manera si él número de grupos que se tiene es igual a tres, los centroides definirían un plano y dentro de este plano es también posible caracterizar el **centroide general** que es aquel en el que no se consideran las diferencias de grupos. Este centroide general será considerado como el origen en el espacio definido por los centroides de cada grupo. Una vez definido el origen, resta a definir la orientación de los ejes. El primer eje puede construirse de tal manera que los centroides de los grupos queden lo más separados posible. Si el número de grupos es mayor a dos, pueden generarse nuevos ejes hasta obtener $(g-1)$. Un segundo eje se obtiene con el mismo criterio de máxima separación entre los grupos con la condición adicional de que sea perpendicular al primer eje. En el caso de tener más de tres grupos, los ejes sucesivos se definen de la misma manera. Para el caso de tres variables discriminantes y tres grupos, tendríamos entonces un plano definido por los centroides dentro del cual es posible definir dos ejes como descrito anteriormente, tal y como se muestra en la siguiente figura.

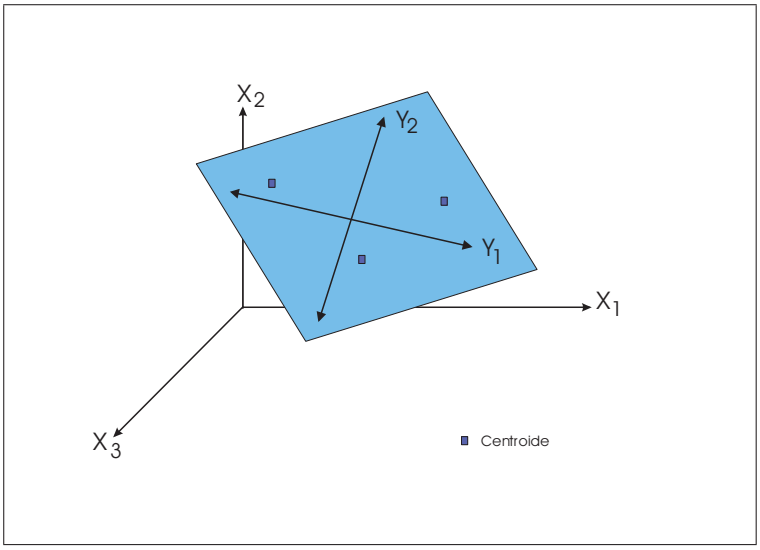


Figura 3.2: Funciones discriminantes para tres grupos y tres variables discriminantes

Estos ejes que maximizan la diferencia entre los grupos y además son ortogonales entre sí, corresponden a las funciones discriminantes. En otras palabras, las funciones discriminantes pueden ser vistas como **ejes** o **dimensiones** de un espacio en el que pueden ser examinadas las diferencias entre los grupos. Por medio de las ecuaciones de las funciones discriminantes,

se transforman las coordenadas de las observaciones, del espacio p dimensional al espacio definido por medio de los centroides. Es decir las **puntuaciones discriminantes** de las observaciones, obtenidas con cada ecuación discriminante, son las coordenadas de éstas sobre un eje discriminante.

Como hemos visto, el número de funciones discriminantes nunca a puede ser mayor al número de grupos menos uno. Sin embargo si el número de variables p es menor que número de grupos, tendremos p funciones discriminantes y en este caso las funciones discriminantes no permiten el cambio de coordenadas a un espacio de dimensión menor, sino simplemente generan un cambio de ejes dentro del mismo espacio p-dimensional.

3.3.3 Obtención de los coeficientes de las funciones discriminantes

Supongamos que contamos con p variables discriminantes, a partir de las cuales generamos una combinación lineal que permita la discriminación de n individuos en g grupos diferentes. Esta combinación sería entonces de la forma:

$$\mathbf{Y} = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \tag{3.2}$$

Por medio de esta combinación lineal, obtendremos para cada individuo una sola puntuación, y las puntuaciones de todos los individuos pueden agruparse en una única variable. A partir de esta variable podemos determinar las diferencias entre las medias de los grupos, al analizar el cociente entre las sumas de cuadrados intergrupos y las sumas de cuadrados intragrupos.

Recordemos primero las matrices $\mathbf{W}$ y $\mathbf{B}$, brevemente presentadas en el capítulo anterior (2.64 ,2.65). La matriz $\mathbf{W}$ es la matriz de sumas de cuadrados y productos cruzados intragrupos y está dada por:

$$\mathbf{W} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\mathbf{X}_{ki} - \overline{\mathbf{X}}_k)(\mathbf{X}_{ki} - \overline{\mathbf{X}}_k)' \tag{3.3}$$

Esta matriz recoge toda la información referente a la variabilidad al interior de los grupos.

La matriz $\mathbf{B}$ es la matriz de sumas de cuadrados y productos cruzados intergrupos, en ella se recaba toda la información referente a la variabilidad entre los grupos y está dada por:

$$\mathbf{B} = \sum_{k=1}^{g} (\mathbf{X}_k - \overline{\mathbf{X}})(\mathbf{X}_k - \overline{\mathbf{X}})' \tag{3.4}$$

Gracias a estas matrices, tenemos entonces toda la información referente a la variabilidad intergrupos e intragrupos, de hecho es posible demostrar que[1]:

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \tag{3.5}$$

donde $\mathbf{T}$ es la matriz total de sumas y productos cruzados dada por:

$$\mathbf{T} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\mathbf{X}_{ki} - \overline{\mathbf{X}})(\mathbf{X}_{ki} - \overline{\mathbf{X}})' \tag{3.6}$$

En esta matriz se recoge la información de la covarianza entre cada pareja de variables, de hecho bastaría dividir por $(n-1)$ cada elemento de la matriz para obtener la covarianza muestral entre cada par de variables. La ecuación (3.5) se conoce también como descomposición de la varianza.

Gracias a estas matrices, podemos expresar la variabilidad intergrupos e intergrupos para la variable $\mathbf{Y}$.

La variabilidad intragrupos en términos de $\mathbf{Y}$ está dada por:

$$\mathbf{a}'\mathbf{W}\mathbf{a} \tag{3.7}$$

y la variabilidad intergrupos en términos de $\mathbf{Y}$ es:

$$\mathbf{a}'\mathbf{B}\mathbf{a} \tag{3.8}$$

donde $\mathbf{a} = (a_1, a_2, ..., a_p)$

Recordemos que $\mathbf{W} = \mathbf{S}_1 + \mathbf{S}_2 + ... + \mathbf{S}_g$ (2.64) , donde $\mathbf{S}_i(\mathbf{X})$ es la matriz de sumas de cuadrados y productos cruzados totales del $i-$ésimo grupo, que puede expresarse de la forma $\mathbf{S}_i = \mathbf{X}'\mathbf{X} - \overline{\mathbf{X}'\mathbf{X}}$. Ahora bien para la matriz de sumas de cuadrados y productos cruzados totales de $\mathbf{Y}$, tenemos, para el $i-$ésimo grupo:

$$\mathbf{S}_i(\mathbf{Y}) = \mathbf{Y}'\mathbf{Y} - \overline{\mathbf{Y}'\mathbf{Y}}$$

donde $\mathbf{Y}$ puede expresarse como $\mathbf{X}\mathbf{a}$ y $\overline{\mathbf{Y}}$ como $\overline{\mathbf{X}}\mathbf{a}$, de esta manera tendríamos

$$\begin{aligned}
\mathbf{S}_i(\mathbf{Y}) &= (\mathbf{X}\mathbf{a})'(\mathbf{X}\mathbf{a}) - (\overline{\mathbf{X}}\mathbf{a})'(\overline{\mathbf{X}}\mathbf{a}) \\
&= (\mathbf{a}'\mathbf{X})(\mathbf{X}\mathbf{a}) - (\mathbf{a}'\overline{\mathbf{X}'})(\overline{\mathbf{X}}\mathbf{a}) \\
&= \mathbf{a}'(\mathbf{X}'\mathbf{X})\mathbf{a} - \mathbf{a}'(\overline{\mathbf{X}'\mathbf{X}})\mathbf{a} \\
&= \mathbf{a}'(\mathbf{X}'\mathbf{X} - \overline{\mathbf{X}'\mathbf{X}})\mathbf{a} \\
&= \mathbf{a}'\mathbf{S}_i(\mathbf{X})\mathbf{a}
\end{aligned}$$

[1] Ver por ejemplo Bardos Mireille "Analyse discriminante, Application au risque et scoring financier" Ed Dunod 2001

ahora bien respecto a la matriz de sumas de productos cruzados intragrupos en términos de $\mathbf{Y}$, que denotaremos como $\mathbf{SS_w(Y)}$, tenemos, aplicando lo anterior a cada uno de los grupos y sumando los resultados:

$$\mathbf{SS_w(Y)} = \mathbf{a'S}_1\mathbf{(X)a} + \mathbf{a'S}_2\mathbf{(X)a} + \cdots + \mathbf{a'S}_g\mathbf{(X)a}$$

o simplemente

$$\begin{aligned} \mathbf{SS_w(Y)} &= \mathbf{a'S}_1\mathbf{a} + \mathbf{a'S}_2\mathbf{a} + \cdots + \mathbf{a'S}_g\mathbf{a} \\ &= \mathbf{a'}(\mathbf{S}_1 + \mathbf{S}_2 + \cdots + \mathbf{S}_g)\mathbf{a} \end{aligned}$$

lo que es equivalente a

$$\mathbf{SS_w(Y)} = \mathbf{a'Wa}$$

La demostración para la variabilidad intergrupos se realiza de manera similar[2]

Utilizando las ecuaciones (3.7) y (3.8) podemos escribir el cociente entre la variabilidad intergrupos e intragrupos de $\mathbf{Y}$, como función del vector de coeficientes $\mathbf{a}$. De esta manera obtenemos:

$$\frac{\mathbf{a'Ba}}{\mathbf{a'Wa}} = \lambda \tag{3.9}$$

donde λ es un escalar que utilizaremos como criterio para medir la diferenciación de los grupos y que es conocido como **criterio de discriminación**. Cabe señalar que este cociente es similar a la $\mathbf{F}$ utilizada en el análisis de varianza ($F = \frac{SC_{entre}}{SC_{intra}} \frac{n-g}{g-1}$), con la diferencia que el segundo factor es constante para cualquier situación concreta, y por lo tanto sólo el primer factor es relevante para determinar en qué medida las medias de los grupos varían en relación con la variabilidad interna de los grupos.

El problema consiste entonces en encontrar los valores de $\mathbf{a}$, que maximicen el criterio de discriminación. Para ello, se calcula la derivada parcial del cociente respecto a cada elemento a_i de $\mathbf{a}$, y se iguala el resultado a cero.

Es posible demostrar[3] que $\frac{\partial}{\partial x}(\mathbf{x'Ax}) = (\mathbf{A} + \mathbf{A'})\mathbf{x}$, donde $\mathbf{x}$ es un vector de dimensión p y $\mathbf{A}$ una matriz de dimensión $p \times p$ (si $\mathbf{A}$ es simétrica esto se reduce a $\frac{\partial}{\partial x}(\mathbf{x'Ax}) = 2\mathbf{Ax}$) , y que B y W son matrices simétricas, es decir $\mathbf{B} = \mathbf{B'}$ y $\mathbf{W} = \mathbf{W'}$. De esta manera, la derivada parcial de λ respecto a $\mathbf{a}$, está dada por:

$$\begin{aligned} \frac{\partial \lambda}{\partial \mathbf{a}} &= \frac{\partial \left(\frac{\mathbf{a'Ba}}{\mathbf{a'Wa}}\right)}{\partial \mathbf{a}} \\ &= \frac{2\mathbf{Ba}(\mathbf{a'Wa}) - (\mathbf{a'Ba})2\mathbf{Wa}}{(\mathbf{a'Wa})^2} \end{aligned}$$

[2]Ver por ejemplo Tatsuoka Maurice, "Multivariate Analysis: Techniques for educational and psychological research", Ed John Wiley and Sons,1971

[3]Ver por ejemplo Tatsuoka Maurice, "Multivariate Analysis: Techniques for educational and psychological research", Ed John Wiley and Sons,1971

dividiendo el numerador y el denominador por $(\mathbf{a}'\mathbf{W}\mathbf{a})$ obtenemos:

$$\frac{2[\mathbf{B}\mathbf{a} - \lambda\mathbf{W}\mathbf{a}]}{(\mathbf{a}'\mathbf{W}\mathbf{a})} = 0$$

lo que es equivalente a

$$(\mathbf{B} - \lambda\mathbf{W})\,\mathbf{a} = 0$$

Asumiendo que $\mathbf{W}$ es invertible y postmultiplicando por $\mathbf{W}^{-1}$ tenemos

$$\left(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right)\mathbf{a} = 0 \tag{3.10}$$

que es una ecuación del tipo $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0$ y como sabemos las soluciones de λ son los eigenvalores de la matriz $\mathbf{A}$ y las soluciones de $\mathbf{v}$ corresponden a los eigenvectores asociados (ver sección 1.5). De esta manera, al obtener los eigenvalores y eigenvectores de $\mathbf{W}^{-1}\mathbf{B}$, se resuelve el problema de maximización del criterio discriminante.

El rango de una matriz producto no puede ser mayor al menor de los rangos de las matrices que conforman el producto. Suponiendo la no singularidad de $\mathbf{W}$ y que n-g>p, la matriz $\mathbf{W}$ será de rango p. Por otra parte, el rango de $\mathbf{B}$ es igual al mínimo entre $g-1$ y p. Por ello, el rango de la matriz $\mathbf{W}^{-1}\mathbf{B}$, que denotaremos como q será igual a $q = \min(g-1, p)$. Dado que el número de eigenvalores distintos de cero coincide con el rango de la matriz, para $\mathbf{W}^{-1}\mathbf{B}$ tendremos q eigenvalores no nulos y q eigenvectores asociados. Así, al resolver la ecuación (3.10) obtendremos q eigenvectores $\lambda_1, \lambda_2, ..., \lambda_q$ (ordenados de mayor a menor) y q eigenvectores asociados $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_q$ (ó $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_q$). Puesto que el eigenvalor λ_1 es el de mayor magnitud, la función

$$\mathbf{Y}_1 = (\mathbf{a}_{11}\mathbf{X}_1) + (\mathbf{a}_{12}\mathbf{X}_2) + \cdots + (\mathbf{a}_{1p}\mathbf{X}p) \tag{3.11}$$

es la combinación lineal que consigue una mayor discriminación.

La segunda función, asociada al segundo eigenvalor, estaría dada por

$$\mathbf{Y}_2 = (\mathbf{a}_{21}\mathbf{X}_1) + (\mathbf{a}_{22}\mathbf{X}_2) + \cdots + (\mathbf{a}_{2p}\mathbf{X}p) \tag{3.12}$$

y es la que presenta mayor poder de discriminación después de $\mathbf{Y}_1$ y además no está correlacionada con esta última.

Las funciones discriminantes sucesivas se definen de manera análoga, hasta completar q funciones discriminantes, que también se conocen como funciones lineales discriminantes.

Los coeficientes $\mathbf{a}_i$, conocidos como **coeficientes no estandarizados**, obtenidos de esta manera suelen utilizarse para la clasificación de los individuos, sin embargo no son interpretables ya que estos fueron calculados tomando en cuenta diferentes escalas. Para que estos

coeficientes fueran comparables bastaría multiplicar cada elemento del eigenvector $\mathbf{a}_i$ por la raíz cuadrada del elemento diagonal correspondiente de $\mathbf{W}$. Así tendríamos:

$$\mathbf{a}_i^* = \frac{\sqrt{w_{ii}}}{n-g}\mathbf{a}_i \quad i = 1, ..., p \tag{3.13}$$

Estos coeficientes se conocen como **coeficientes discriminantes estandarizados** y por medio de ellos es posible comparar las distintas contribuciones de las variables a las funciones discriminantes. Algunos algoritmos de análisis discriminante incluyen un ajuste a los coeficientes y una constante en el modelo, para hacer coincidir el origen de cada eje discriminante con el centroide global, esta pude calcularse de la manera siguiente:

$$a_i' = a_i\sqrt{n-g}$$

$$a_0 = -\sum_{i=1}^{p} a_i\overline{\mathbf{X}}_i$$

De esta manera, la puntuación discriminante asignada a un caso representa el número de desviaciones típicas que este se aleja del centroide global.

3.4 Reglas de Clasificación

Como ya lo habíamos mencionado, uno de los objetivos del análisis discriminante es la clasificación de los individuos en uno de los grupos previamente definidos, con el menor error posible. En esta sección, se describirán brevemente algunos de los métodos de clasificación, como por ejemplo aquellos en los que no es necesario el uso de funciones discriminantes, sino más bien de funciones de clasificación. Sin embargo, muchos autores señalan que la clasificación a través de las funciones discriminantes arroja mejores resultados en muchos de los casos y además es más fácil de implementar.

3.4.1 Funciones de Clasificación de Fisher

Como se mencionó anteriormente, Fisher fue el primero que sugirió una combinación lineal de variables para la clasificación de los individuos. Esta combinación debería construirse de tal manera que lograra la maximización de las diferencias entre los grupos al tiempo que minimizara la diferencia dentro de los grupos. Basada en esta idea, existe una función de clasificación, llamada **función de clasificación lineal** que está dada por:

$$h_k = b_{ko} + b_{k1}X_1 + b_{k2}X_2 + \cdots + b_{kp}X_p \tag{3.14}$$

donde h_k es la puntuación obtenida por un individuo para el grupo k $(k = 1...g)$.

En el caso de esta función de clasificación, conocida como **función de clasificación de Fisher**, los coeficientes se obtienen de la siguiente manera:

$$\begin{aligned} b_{k1} &= (n-g)\sum_{j=1}^{p} w_{ij}\overline{X}_j \\ b_{k0} &= -\frac{1}{2}\sum_{j=1}^{p} b_{kj}\overline{X}_j \end{aligned} \tag{3.15}$$

donde w_{ij} es un elemento de la matriz $\mathbf{W}$ definida anteriormente.

Estas puntuaciones serán más altas cuanto más cercano esté el caso del grupo, de esta manera se evalúan las puntuaciones de cada individuo para todos los grupos y el individuo será asignado al grupo para el cual haya obtenido la puntuación más elevada. Cabe señalar que este procedimiento es sensible a la violación del supuesto de igualdad de las matrices de varianza covarianza, de violarse este supuesto, los individuos tenderán a ser clasificados en el grupo de mayor dispersión. Algunos paquetes estadísticos tales como el SPSS permiten la obtención de los coeficientes de la función de clasificación de Fisher con facilidad, para así poder construir una función de clasificación para cada grupo que permitirá la obtención de las puntuaciones para cada individuo. El individuo quedará entonces clasificado en el grupo para el que haya obtenido una mayor puntuación. Si bien las funciones discriminantes ascienden a $q = \min(g-1, p)$, en el caso de las funciones de clasificación de Fisher, siempre tendremos tantas funciones como grupos.

3.4.2 Funciones de Distancia Generalizada

Otra manera de clasificar a un individuo dentro de un grupo se realiza por medio del cálculo de la distancia entre las observaciones de este individuo y los centroides de los grupos. En este caso, el individuo se asignaría al grupo de cuyo centroide se encuentre más próximo. La distancia de Mahalanobis entre un individuo con vector de observaciones $\mathbf{X}_i$ y el centroide de un grupo $\boldsymbol{\mu}_g$, como se menciono en el capítulo anterior (2.46) está dada por:

$\Delta_{ig} = [(\mathbf{X}_i - \boldsymbol{\mu}_g)'\Sigma_g^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_g)]^{1/2}$ donde Σ_g es la matriz de varianza covarianza de la población g

Para la clasificación de los individuos, es común utilizar una expresión equivalente a la

anterior dada por:

$$D^2(X \mid G_k) = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}(\overline{X}_i - \overline{X}_{ik})(\overline{X}_j - \overline{X}_{jk}) \tag{3.16}$$

donde G_k es el centroide del grupo k y w_{ij} es un elemento de la matriz $\mathbf{W}$ definida anteriormente.

De esta manera un individuo se asignará al grupo respecto al cual presenta la distancia más pequeña. Más delante veremos como este mismo concepto de distancia puede ser empleado para la clasificación de los individuos desde un punto de vista probabilístico.

3.4.3 Probabilidad de pertenencia a un grupo

Reglas de clasificación en general

Muchas de las reglas de clasificación utilizadas comúnmente se basan en el **principio de máxima verosimilitud**, es decir se pretende asignar un individuo al grupo para el cual su vector de observaciones tenga la mayor verosimilitud de ocurrencia. Esto puede verse en términos de las funciones de verosimilitud o de las funciones de densidad.

Por ejemplo consideremos el caso de una sola variable $\mathbf{X}$ y dos grupos distintos, denotados por G_1 y G_2. Supongamos además que los modelos (o funciones de distribución) para las dos poblaciones son como se representan en la siguiente figura:

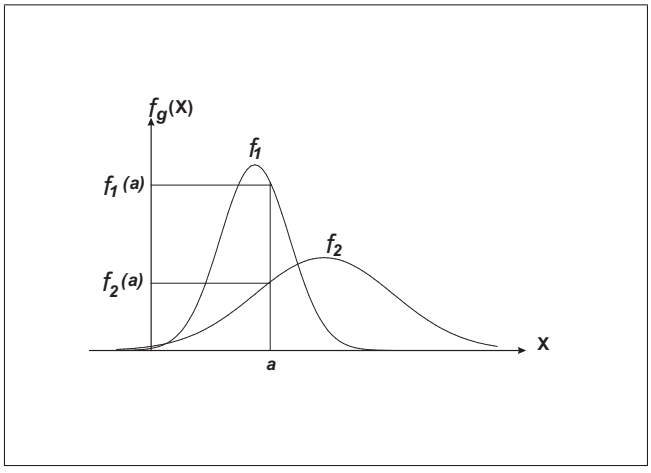


Figura 3.3: Representación gráfica de dos funciones de densidad.

Ahora bien, supongamos que tenemos una observación extraída del grupo 1, digamos $X = a$. Denotaremos la función de verosimilitud de para el grupo 1 como $f_1(a)$ y para el grupo 2 como $f_2(a)$. De acuerdo con el principio de máxima verosimilitud obtendríamos la siguiente regla de clasificación:

Asignar un individuo con $X = a$ al grupo 1 si $f_1(a) > f_2(a)$, de lo contrario asignarlo al grupo 2.

Para el caso de g grupos distintos, suponiendo que las distribuciones son iguales para los g grupos (por ejemplo todas son normales) y que f denota la función de densidad conjunta, la **regla de clasificación de máxima verosimilitud** sería:

$$\begin{gathered}\text{Asignar un individuo } i \text{ al grupo } k \text{ si} \\ f(\mathbf{X}_i \mid k) > f(\mathbf{X}_i \mid k') \\ \text{para todo } k \neq k'\end{gathered} \tag{3.17}$$

Es decir asignar un individuo i al grupo k si la verosimilitud de su vector de observaciones $\mathbf{X}_i$ es mayor para el grupo k que para cualquier otro grupo $k\prime$.

Esta regla también puede verse en términos de probabilidades más que en términos de funciones de verosimilitud. Sea $P(\mathbf{X} \mid k)$ la probabilidad de que un individuo tenga un perfil parecido a $\mathbf{X}$ dado que el individuo pertenece al grupo k, en otras palabras, es la probabilidad de que ocurra el vector de observaciones $\mathbf{X}$ dado que el individuo pertenece al grupo k. De esta manera, podemos obtener una segunda versión de la regla de clasificación de máxima verosimilitud, en términos de probabilidades:

$$\begin{gathered}\text{Asignar un individuo } i \text{ al grupo } k \text{ si} \\ P(\mathbf{X}_i \mid k) > P(\mathbf{X}_i \mid k') \\ \text{para todo } k \neq k'\end{gathered} \tag{3.18}$$

Esta regla puede ser vista desde otro punto de vista, al considerar la probabilidad de que un individuo i pertenezca a un grupo k dado que tiene un vector de observaciones $\mathbf{X}_i$, es decir $P(k \mid \mathbf{X}_i)$. Esta probabilidad se conoce como **probabilidad a posteriori de pertenencia a un grupo** y en este caso, se considera que todos los grupos tienen la misma probabilidad de ocurrencia, esto es $P(k) = P(k')$ para todo $k \neq k'$. De esta manera, los individuos serán clasificados en el grupo para el cual su probabilidad de pertenencia a posteriori sea más grande. La regla de clasificación estaría entonces dada por:

$$\begin{gathered}\text{Asignar un individuo } i \text{ al grupo } k \text{ si} \\ P(k \mid \mathbf{X}_i) > P(k' \mid \mathbf{X}_i) \\ \text{para todo } k \neq k'\end{gathered} \tag{3.19}$$

donde

$$P(k \mid \mathbf{X}_i) = \frac{P(\mathbf{X}_i \mid k)}{\sum_{k'=1}^{g} P(\mathbf{X}_i \mid k')} \tag{3.20}$$

Demostración.

Sea π_k la proporción de individuos de la muestra total que pertenecen al grupo k, π_k es conocida como la p**robabilidad a priori de pertenencia al grupo** k ya que representa la probabilidad de que un individuo seleccionado aleatoriamente en la población pertenezca al grupo k, en ocasiones π_k es denotada por $P(k)$.

Recordemos que por el teorema de Bayes tenemos:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

Tomando k y $\mathbf{X}_i$, como los eventos de interés tenemos:

$$P(k \cap \mathbf{X}_i) = P(\mathbf{X}_i)P(k \mid \mathbf{X}_i) = \pi_k P(\mathbf{X}_i \mid k)$$

despejando para $P(k \mid \mathbf{X}_i)$ tenemos:

$$P(k \mid \mathbf{X}_i) = \frac{\pi_k P(\mathbf{X}_i \mid k)}{P(\mathbf{X}_i)} \tag{3.21}$$

Sin embargo

$$\sum_{k'=1}^{g} P(k' \mid \mathbf{X}_i) = 1 = \sum_{k'=1}^{g} \frac{\pi_{k'} P(\mathbf{X}_i \mid k')}{P(\mathbf{X}_i)}$$

por lo que

$$P(\mathbf{X}_i) = \sum_{k'=1}^{g} \pi_{k'} P(\mathbf{X}_i \mid k')$$

Substituyendo esta expresión en (3.21) tenemos:

$$P(k \mid \mathbf{X}_i) = \frac{\pi_k P(\mathbf{X}_i \mid k)}{\sum_{k'=1}^{g} \pi_{k'} P(\mathbf{X}_i \mid k')} \tag{3.22}$$

si consideramos que todos los grupos tienen la misma probabilidad es decir que $\pi_k = \pi_{k'}$ para todo $k \neq k'$, la expresión anterior se simplifica a

$$P(k \mid \mathbf{X}_i) = \frac{P(\mathbf{X}_i \mid k)}{\sum_{k'=1}^{g} P(\mathbf{X}_i \mid k')} \tag{3.23}$$

■

Para todas estas reglas de verosimilitud es necesario calcular g valores distintos, para las dos ultimas reglas, la eficacia dependerá del cálculo adecuado de las probabilidades, por ello en caso de que los tamaños de los grupos sean diferentes, habrá que tomarlos en cuenta para el cálculo de estas probabilidades.

Al incorporar las probabilidades a priori de pertenencia a los grupos en la regla de decisión (3.19), como hemos visto en la demostración anterior, obtenemos la llamada **regla de máxima probabilidad Bayesiana**, dada por:

$$\begin{gathered}\text{Asignar un individuo } i \text{ al grupo } k \text{ si} \\ P(k \mid \mathbf{X}_i) > P(k' \mid \mathbf{X}_i) \\ \text{para todo } k \neq k'\end{gathered} \tag{3.24}$$

donde

$$P(k \mid \mathbf{X}_i) = \frac{\pi_k P(\mathbf{X}_i \mid k)}{\sum_{k'=1}^{g} \pi_{k'} P(\mathbf{X}_i \mid k')} \tag{3.25}$$

Dado que el denominador de la expresión anterior es constante para todos los grupos, la regla de clasificación podría basarse únicamente en los valores de $\pi_k P(\mathbf{X}_i \mid k)$ que además resultan proporcionales a los valores de $\pi_k f(\mathbf{X}_i \mid k)$. De hecho la expresión (3.25) podría verse de manera equivalente como:

$$P(k \mid \mathbf{X}_i) = \frac{\pi_k f(\mathbf{X}_i \mid k)}{\sum_{k'=1}^{g} \pi_{k'} f(\mathbf{X}_i \mid k')} \tag{3.26}$$

Reglas de clasificación basadas en el supuesto de normalidad

Hasta ahora, hemos presentado reglas de clasificación sin especificar alguna distribución para los grupos, sin embargo existen reglas de clasificación para cuando se supone distribución normal multivariada y estas son las que utilizaremos para el análisis discriminante ya que uno de los supuestos de este es la normalidad de las distribuciones.

Como se presentó en el capítulo anterior (2.49), la función de densidad de un vector aleatorio con distribución normal multivariada para un grupo k esta dada por:

$$f(\mathbf{X} \mid \mathbf{k}) = \frac{1}{(2\pi)^{p/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{X} - \boldsymbol{\mu}_k)\right] \tag{3.27}$$

donde $\boldsymbol{\mu}_k$ es el vector de medias de $\mathbf{X}$ y $\boldsymbol{\Sigma}_k$ su matriz de varianzas covarianzas.

Como también ya habíamos señalado el término $(\mathbf{X} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{X} - \boldsymbol{\mu}_k)$ es el índice de distancia Δ_{ig} entre un punto y el centroide del grupo. Usualmente, los parámetros son desconocidos, por lo que se remplazan por sus estimadores muestrales $\overline{\mathbf{X}}_k$ y $\mathbf{S}_k$,dando origen a la función de densidad muestral:

$$\hat{f}(\mathbf{X} \mid \mathbf{k}) = \frac{1}{(2\pi)^{p/2} \left|\mathbf{S}_k\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X} - \overline{\mathbf{X}}_k)' \mathbf{S}_k^{-1} (\mathbf{X} - \overline{\mathbf{X}}_k)\right] \tag{3.28}$$

La expresión anterior puede también escribirse para un individuo i como:

$$\hat{f}(\mathbf{X}_i \mid \mathbf{k}) = \frac{1}{(2\pi)^{p/2} \left|\mathbf{S}_k\right|^{1/2}} \exp\left[-\frac{1}{2} D_{ik}^2\right] \tag{3.29}$$

donde $D_{ik}^2 = (\mathbf{X}_i - \overline{\mathbf{X}}_k)' \mathbf{S}_k^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}_k)$ es el índice muestral de Mahalanobis de distancia entre un individuo y el centroide del grupo k.

Retomando la expresión presentada en (3.26), y substituyendo por la función de densidad muestral tendríamos:

$$\hat{P}(k \mid \mathbf{X}_i) = \frac{q_k \left|\mathbf{S}_k\right|^{-1/2} \exp\left[-\frac{1}{2} D_{ik}^2\right]}{\sum_{k'=1}^{g} q_{k'} \left|\mathbf{S}_{k'}\right|^{-1/2} \exp\left[-\frac{1}{2} D_{ik'}^2\right]} \tag{3.30}$$

donde $q_k = \hat{\pi}_k$ (dado que $(2\pi)^{-p/2}$ es un factor común del numerador y el denominador, éste queda simplificado)

De esta manera, la **regla de clasificación de máxima probabilidad para el caso de una distribución normal** $p-$**variada** estará dada por:

$$\begin{gathered} \text{Asignar un individuo } i \text{ al grupo } k \text{ si} \\ \hat{P}(k \mid \mathbf{X}_i) > \hat{P}(k' \mid \mathbf{X}_i) \\ \text{para todo } k \neq k' \end{gathered} \tag{3.31}$$

donde $\hat{P}(k \mid \mathbf{X}_i)$ se obtiene como en (3.30).

Consideremos ahora el caso de que las matrices de varianza covarianza de los g grupos sean iguales, esto es $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g$, en este caso un estimador de $\boldsymbol{\Sigma}$ es la matriz de

varianzas covarianzas ponderadas $\mathbf{S}_{pl}$ presentada en el capítulo anterior (2.51), de esta manera tendíamos:

$$\hat{f}(\mathbf{X}_i|\ \mathbf{k}) = \frac{1}{(2\pi)^{p/2}\left|\mathbf{S}_{pl}\right|^{1/2}} \exp\left[-\frac{1}{2}D^{*2}_{ik}\right] \tag{3.32}$$

donde $D^{*2}_{ik} = (\mathbf{X}_i - \overline{\mathbf{X}}_k)'\mathbf{S}^{-1}_{pl}(\mathbf{X}_i - \overline{\mathbf{X}}_k)$.

Dado que para este caso $(2\pi)^{-p/2}\left|\mathbf{S}_{pl}\right|^{-1/2}$, sería un factor común en la expresión (3.30), la **regla de máxima probabilidad para el caso de una distribución normal $p-$variada bajo el supuesto de igualdad en las matrices de varianzas covarianzas** estaría dada por:

$$\begin{gathered}\text{Asignar un individuo } i \text{ al grupo } k \text{ si} \\ \hat{P}(k \mid \mathbf{X}_i) > \hat{P}(k' \mid \mathbf{X}_i) \\ \text{para todo } k \neq k'\end{gathered} \tag{3.33}$$

donde

$$\hat{P}(k \mid \mathbf{X}_i) = \frac{q_k \exp\left[-\frac{1}{2}D^{*2}_{ik}\right]}{\sum_{k'=1}^{g} q_{k'} \exp\left[-\frac{1}{2}D^{*2}_{ik'}\right]} \tag{3.34}$$

Esta es la regla de clasificación que utilizaremos ya que para el modelo de análisis discriminante se supone distribución normal multivariada e igualdad de las matrices de varianza covarianza. Los estimados de las probabilidades de la expresión anterior son los que se calculan en el modelo discriminante y los paquetes estadísticos más comunes (como SPSS y SAS) las proporcionan para cada individuo y para todos los grupos, no sólo para el de mayor probabilidad.

Cabe señalar que a partir de estas reglas basadas en la normalidad pueden construirse otras reglas como por ejemplo las basadas en una función cuadrática derivada de (3.31) ,o en una función lineal basada en (3.33) ,o en la mínima distancia. Sin embargo, para los fines de este trabajo, la obtención de estas reglas se deja al lector.[4]

3.4.4 Clasificación por medio de la función discriminante.

Si bien la función discriminante es muy útil para analizar un problema de discriminación desde un punto de vista descriptivo, también resulta útil para la clasificación de los individuos. Las reglas de clasificación presentadas hasta ahora consideran el cálculo de g puntuaciones (obtenidas por combinaciones lineales, distancias o probabilidades) calculadas a partir de las p variables discriminantes. Para la clasificación de los individuos, el planteamiento sería el

[4] Ver por ejemplo Huberty "Applied Discriminant Analysis", 1994 Ed. John Wiley and Sons.

mismo que el presentado hasta ahora, con la diferencia que los cálculos se realizan a partir de las q funciones discriminantes y no de las p variables, lo que en la mayoría de los casos resulta más práctico para el cálculo de probabilidades y de distancias. Las clasificaciones basadas en las funciones discriminantes suelen ser las mismas que aquellas basadas en las variables independientes, a excepción de cuando se viola el supuesto de igualdad de matrices de varianzas covarianzas.

Para el caso concreto de la aplicación de este trabajo, la clasificación se hará entonces utilizando la regla de máxima probabilidad de Bayes aplicada a las puntuaciones discriminantes obtenidas por cada individuo por medio de las funciones de discriminación no estandarizadas. Supondremos que la distribución de estas últimas es normal multivariada y por medio del paquete estadístico SPSS obtendremos tanto las probabilidades de pertenencia a un grupo de los individuos como las puntuaciones discriminantes de cada sujeto.

3.5 Selección de Variables

Otro de los aspectos importantes que hay que tener en cuenta para una buena aplicación del análisis discriminante es la selección de variables. En realidad, al comienzo del análisis se cuenta con una serie de variables y el objetivo de la selección será guardar todas aquellas variables que tengan mayor capacidad de discriminación y eliminar aquellas que no lo tengan, de tal manera que el modelo resultante sea el más eficiente en cuanto a la clasificación de los individuos. Las variables cuyos valores medios sean iguales para todos los grupos no estarán aportando información por lo que serán eliminadas, y aquellas variables que aporten información redundante también se eliminarán. Si bien es posible realizar un análisis con todas las variables que se tienen disponibles (método de inclusión completa), la selección de variables permite obtener mejores resultados al determinar el subconjunto de variables que más discrimine los grupos, además que al eliminar variables se reduce la complejidad del problema. Una vez encontradas las variables para el modelo óptimo se realiza la obtención de las funciones discriminantes basada en estas variables y posteriormente se clasifican los individuos.

3.5.1 Condiciones mínimas para la selección de variables

Todas las variables, que se considere pueden ser incluidas en el modelo del análisis discriminante, deben cumplir ciertas condiciones antes de ser sometidas a los criterios de selección, de no cumplir con estas condiciones, no serán incluidas en el modelo. De igual manera, una vez realizada la selección, independientemente del criterio utilizado, las variables deberán ser

revisadas para comprobar que ninguna de las ya seleccionadas deba ser excluida del análisis. Las condiciones que las variables deben cumplir se basan en la tolerancia y en los estadísticos parciales $F_{\text{de entrada}}$ y $F_{\text{de salida}}$. En otras palabras, para que una variable sea incluida en el análisis deberá cumplir con las condiciones de tolerancia y de $F_{\text{de entrada}}$ para poder ser sometida a los criterios de selección previamente presentados. En cada paso de selección de variables, se comprobará que todas las variables incluidas hasta ese paso cumplan con las condiciones referentes al estadístico $F_{\text{de salida}}$, de no ser así, la o las variables que no satisfagan esta condición serán eliminadas.

Tolerancia

La **tolerancia** es una medida del grado de asociación lineal entre las variables. En otras palabras, la tolerancia de una variable es la proporción de varianza de esa variable que no depende del resto de las variables independientes. La tolerancia para una variable no seleccionada está dada por:

$$tolerancia = 1 - R^2 \tag{3.35}$$

donde R^2 es el cuadrado de la correlación múltiple entre la variable candidata a ser incluida en el modelo y todas las variables ya incluidas en el modelo.

Una variable con tolerancia cercana cero no deberá ser incluida en el modelo, pues esto indicaría una fuerte relación lineal entre esta variable y las variables previamente seleccionadas, es decir la variable no estaría aportando información adicional. Para poder ser incluida una variable, debe entonces tener una tolerancia mayor al valor fijado (para el paquete SPSS este valor es por defecto 0,001) y no debe provocar que el nivel de tolerancia de alguna de las variables previamente seleccionadas descienda por debajo del nivel fijado.

Estadístico F de entrada

El **estadístico parcial** $F_{\text{de entrada}}$ mide el incremento producido en la discriminación tras la incorporación de una variable respecto al nivel de discriminación obtenido por las variables ya seleccionadas. De esta manera si una variable arroja un valor pequeño en la $F_{\text{de entrada}}$ sería aconsejable no incluirla en el modelo. El estadístico $F_{\text{de entrada}}$ para una variable i esta dado por:

$$F_{\text{de entrada}} = \left(\frac{n-g-s}{g-1}\right)\left(\frac{1-\Lambda_i/\Lambda_{(i)}}{\Lambda_i/\Lambda_{(i)}}\right) \tag{3.36}$$

donde Λ_i es el valor de la lambda de Wilks (ver 2.63) tras la inclusión de la variable i, $\Lambda_{(i)}$ es el valor de lambda antes de incluir a la variable i y s es el número de variables ya incluidas en el modelo. Este estadístico tiene una distribución F con $(g-1)$ y $(n-g-s+1)$ grados de libertad.

En el paquete SPSS el valor fijado como mínimo para la $F_{\text{de entrada}}$ es de 3.84, si una variable obtiene un estadístico con un valor inferior a este será excluida del modelo. Sin embargo este paquete estadístico permite modificar este valor o bien modificar la probabilidad de entrada que es nivel crítico asociado al valor del estadístico $F_{\text{de entrada}}$ (por defecto este nivel se fija en 0.05, pero también es posible modificarlo).

Estadístico F de salida

El **estadístico parcial** $F_{\text{de salida}}$ mide el descenso en la discriminación suponiendo la eliminación de una variable del modelo. Un valor bajo del estadístico $F_{\text{de salida}}$ implicaría que la variable puede ser eliminada. El estadístico $F_{\text{de salida}}$ para una variable i esta dado por:

$$F_{\text{de salida}} = \left(\frac{n-g-s}{g-1}\right)\left(\frac{1-\Lambda_i/\Lambda_{(i)}}{\Lambda_i/\Lambda_{(i)}}\right) \tag{3.37}$$

donde Λ_i es el valor de la lambda de Wilks basado en todas las variables en el modelo (incluida la variable i), $\Lambda_{(i)}$ es el valor de lambda que se obtendría tras la eliminación de la variable i y s es el número de variables ya incluidas en el modelo. Este estadístico tiene una distribución F con $(g-1)$ y $(n-g-s+1)$ grados de libertad.

Un valor alto en el estadístico $F_{\text{de salida}}$ indicaría una perdida importante en la separación de los grupos si se eliminara la variable i, por lo que sería recomendable no eliminar esa variable. Del mismo modo que para el caso del estadístico $F_{\text{de entrada}}$, el paquete estadístico SPSS permite modificar el valor fijado para éste que es de 2.71, o bien modificar la probabilidad de salida que por defecto es de 0.10. De esta manera si una variable obtiene un estadístico $F_{\text{de salida}}$ menor que 2.71 será eliminada del modelo. Una vez terminado el proceso de selección de variables, el estadístico $F_{\text{de salida}}$ puede ser utilizado para ordenar las variables de acuerdo con su poder de discriminación, es decir aquellas que obtengan un valor de $F_{\text{de salida}}$ mayor serán las que más contribuyan a la discriminación entre grupos.

3.5.2 Métodos de selección de variables

Para resolver el problema de selección de variables existen básicamente tres tipos de algoritmos, el de **selección hacia adelante** (forward), el de **selección hacia atrás** (backward) y el de

selección por pasos (stepwise). El método más comúnmente utilizado es el de selección por pasos, que es el que se empleará para la aplicación en el siguiente capítulo.

Método de selección hacia adelante (Forward)

En el método de selección hacia delante, el primer paso consiste en determinar cual es la variable que logra mayor discriminación entre los grupos, es decir la que maximiza la separación de estos. Una vez seleccionada la primera variable, se procede a la elección de la segunda variable, se forman parejas entre la primer variable seleccionada y las restantes, para determinar la pareja que produce mayor discriminación. La variable que consiga junto con la primera variable la mayor discriminación posible, es seleccionada como segunda variable. Con estas dos variables seleccionadas y las restantes, se forman triadas para determinar cual será la tercera variable seleccionada, con los mismos criterios que para la segunda. Este proceso se repite hasta que sean incluidas en el modelo todas las variables necesarias, de tal manera que las restantes no contribuyan en forma significativa a la discriminación.

Método de selección hacia atrás (Backward)

El método de selección hacia atrás comienza considerando la inclusión de todas las variables, después de esto se considera la eliminación de la variable aparentemente menos útil para la discriminación. El proceso de selección hacia atrás termina cuando ninguna de las variables incluidas en el modelo pueda ser eliminada, es decir cuando la variable candidata a ser eliminada (esto es la que menos contribuye a la discriminación) resulta significativa.

Método de selección por pasos (Stepwise)

El método de selección por pasos o Stepwise, es el más comúnmente utilizado, en realidad este procedimiento es una combinación del método de selección hacia delante y hacia atrás. En este método las variables son seleccionadas para su inclusión en el modelo exactamente igual que en el método de selección hacia delante, la diferencia radica que en cada paso, es decir antes de incluir una nueva variable, se comprueba que todas variables previamente seleccionadas sean significativas. En realidad, las variables se seleccionan una a la vez, y en cada paso son reevaluadas para asegurar que ninguna de las incluidas anteriormente resulte redundante ante la presencia de otra variable. De esta manera, se obtiene un subconjunto óptimo de variables para la discriminación, pero cabe señalar que este método no resulta muy estable ante muestras pequeñas. Al emplear el método de selección de variables Stepwise en el paquete estadístico SPSS, se cumplen los supuestos de ausencia de multicolinealidad y singularidad requeridos para el análisis discriminante.

3.5.3 Criterios de selección

Para determinar que variables entran y cuales salen, independientemente del método utilizado, es preciso recurrir a algún criterio de discriminación. A continuación se presentan estos criterios, que generalmente suelen conducir a resultados muy similares, cabe también señalar que el paquete SPSS ofrece la opción de elegir el criterio que se desea utilizar.

Lambda de Wilks

Como se mencionó en el capítulo anterior (2.63), la **lambda de Wilks** mide las desviaciones dentro de cada grupo respecto a las desviaciones globales, sin diferenciar grupos. El estadístico Lambda de Wilks permite contrastar la hipótesis de igualdad de medias, este estadístico está dado por:

$$\mathbf{\Lambda} = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} \tag{3.38}$$

Este estadístico toma valores entre 0 y 1, y la hipótesis se rechaza para valores pequeños de lambda.

Siguiendo este criterio, en cada paso se selecciona la variable que produce el valor de lambda más pequeño para el conjunto de variables seleccionadas. En una selección hacia adelante, la primera variable seleccionada será la que presente el valor más pequeño para lambda. La segunda variable seleccionada será aquella que en conjunto con la primer variable seleccionada arroje el valor de lambda más pequeño. El proceso continuará hasta que todas las variables sean incluidas o bien ninguna de las restantes cumpla con ciertos criterios mínimos para la selección presentados en el apartado anterior.

Como también habíamos mencionado el estadístico lambda puede ser transformado en un estadístico F para poder contrastar la existencia de diferencias significativas entre los grupos. De igual manera, es posible calcular un estadístico F parcial como vimos en el apartado anterior. El uso de los estadísticos lambda, F exacta o F parcial suele conducir a los mismos resultados.

EL criterio de lambda Wilks es uno de los más comúnmente empleados, en realidad al utilizar este criterio se busca obtener un modelo que maximice las distancias entre los grupos al tiempo al tiempo que reduce la variabilidad al interior de los mismos. Este es el criterio que utilizaremos para realizar la aplicación del siguiente capítulo.

Distancia de Mahalanobis

A partir de la **distancia de Mahalanobis entre dos centroides** (2.45), se tiene una medida de la separación entre dos grupos. Para el caso de la selección de variables, el estadístico utilizado para medir la distancia multivariada entre dos grupos, digamos a y b, sería:

$$D^2_{ab} = (n-g)\sum_{i=1}^{p'}\sum_{j=1}^{p'} w^*_{ij}(\overline{\mathbf{X}}_i^{(a)} - \overline{\mathbf{X}}_i^{(b)})(\overline{\mathbf{X}}_j^{(a)} - \overline{\mathbf{X}}_j^{(b)}) \tag{3.39}$$

donde p' es el número de variables en el modelo y w^*_{ij} es el elemento ij de la inversa de la matriz $\mathbf{W}$.

El criterio de selección basado en la distancia de Mahalanobis supone el cálculo de este índice para todos los grupos respecto a las variables seleccionadas. De esta manera, se determina cual es la pareja de grupos más cercana, esto es la que tenga un valor de D^2 más pequeño. La siguiente variable que será incluida en el modelo, es aquella que maximice el valor de D^2 para los grupos que inicialmente estaban más cercanos. Al utilizar este criterio, se busca que la separación entre los grupos más cercanos sea máxima en cada paso, de tal manera que el modelo final logre una distancia máxima entre estos.

F intergrupos

El estadístico F, es un estadístico empleado para medir la diferencia entre dos grupos y contrastar la hipótesis de igualdad de medias para estos. Con base en la distancia de Mahalanobis (3.39) es posible construir un **estadístico F** que permita la selección de variables, éste quedaría dado por:

$$F = \frac{(n-1-p')n_a n_b}{p'(n-2)(n_a+n_b)} D^2_{ab} \tag{3.40}$$

donde p' es el número de variables en el modelo y n_a, n_b son los tamaños de los grupos a y b respectivamente.

En cada paso, la variable a ser seleccionada es aquella que produce un mayor valor de F para la pareja de grupos inicialmente más distantes. Este criterio difiere con el criterio basado en la distancia de Mahalanobis únicamente porque en él se consideran los tamaños de los grupos, es decir el estadístico F es la distancia de Mahalanobis ponderada por los tamaños de los grupos.

Varianza residual o Varianza no explicada

El criterio basado en la **varianza residual** utiliza como criterio de inclusión la suma de la variación entre todos los grupos no explicada por las variables ya incluidas en el modelo. Es decir en cada paso se incorpora al modelo la variable que minimiza la varianza residual. La varianza explicada por el modelo, denotada por R^2, resulta ser proporcional a la distancia de Mahalanobis en una constante c ($R^2 = D^2_{ab}$), para el cálculo de la varianza residual, generalmente se utiliza el estadístico R dado por:

$$R = \sum_{i=1}^{g-1} \sum_{j=i+1}^{g} \frac{4}{4 + D^2_{a_i b_j}} \tag{3.41}$$

V de Rao

Otro criterio de selección de variables se basa en una medida de distancia entre los grupos propuesta por Rao (1952), esta medida es conocida como la V **de Rao** o la traza de Lawley-Hotteling y está dada por:

$$V = (n-g) \sum_{i=1}^{p'} \sum_{j=1}^{p'} w^*_{ij} \sum_{k=1}^{g} n_k (\overline{\mathbf{X}}^{(k)}_i - \overline{\mathbf{X}}_i)(\overline{\mathbf{X}}^{(k)}_j - \overline{\mathbf{X}}_j) \tag{3.42}$$

donde p' es el número de variables en el modelo y w^*_{ij} es el elemento ij de la inversa de la matriz $\mathbf{W}$.

Cuanto mayor sea la distancia entre los grupos, mayor será el valor de V, por lo que la contribución de una variable a la discriminación puede medirse a través del incremento que esta produce en V. Es posible demostrar que V tiene una distribución aproximada a la Ji-cuadrada con $p\prime(g-1)$ grados de libertad. El cambio que se produce en V tiene una distribución análoga pero con $m(g-1)$ grados de libertad, donde m es el número de variables añadidas o eliminadas en cada paso. De este modo, es posible contrastar la significancia estadística del cambio en V tras la inclusión de una variable, si este no resultara significativo, la variable no será incluida en el modelo.

Al utilizar este criterio, se busca obtener un modelo que maximice la distancia entre los grupos, sin embargo, no se toma en cuenta la variabilidad intragrupos. Por medio del paquete SPSS, es posible especificar el incremento mínimo que se tiene que dar en el valor de V para que la variable sea incluida en el modelo.

3.6 Interpretación de los resultados

Una vez seleccionadas las variables se procede a la obtención de las funciones discriminantes y posteriormente a la clasificación de los individuos cuyo grupo de origen es conocido (muestra de individuos empleada para obtener la regla de clasificación). Con ello, surgen diversas cuestiones tales como el determinar si todas las funciones discriminantes son significativas, como contribuye cada función a la discriminación y como debe interpretarse. En esta sección trataremos de esclarecer estas cuestiones así como de determinar si una vez obtenidas las funciones discriminantes significativas, éstas conllevan a una buena clasificación de los individuos.

3.6.1 Funciones discriminantes significativas

Como vimos en la sección 3.3.3, las funciones discriminantes se obtienen a través de los eigenvalores y eigenvectores de la matriz $W^{-1}B$. Cada función discriminante tendrá entonces asociado un eigenvalor λ y un conjunto coeficientes, correspondientes a los eigenvectores asociados. Estos eigenvalores reflejan el grado en que una función discrimina entre los diferentes grupos, es decir mide las desviaciones de las puntuaciones discriminantes entre los grupos con respecto a las desviaciones totales dentro de los grupos, y su valor es siempre mayor o igual a cero. De esta manera, la función que tenga un valor de λ mayor será aquella que tenga mayor poder discriminante, por ello si una función discriminante tiene un eigenvalor cercano a cero, esto nos estará indicando que la función tiene un poder discriminante prácticamente irrelevante. El problema radica entonces en tener un parámetro de que tan pequeño debiera ser λ como para descartar la función discriminante en cuestión, que no resulta entonces útil para la discriminación.

Criterio a partir de porcentajes relativos

Es posible comparar las funciones discriminantes entre sí al sumar todos los eigenvalores y dividir cada eigenvalor por el total obtenido. De esta manera, el porcentaje relativo obtenido por cada función nos indicará que porcentaje tiene esta respecto poder discriminante total de las funciones. Estos porcentajes relativos, si bien dan información de que tan importante es una función respecto de las demás, no permiten determinar si una función debe ser excluida, debido a que no es posible fijar un porcentaje mínimo a partir del cuál se considere que la función no es útil para la discriminación.

Correlación canónica y coeficiente eta

Otra medida de las desviaciones de las puntuaciones discriminantes entre los grupos respecto a las desviaciones dentro de los grupos está dada por el **coeficiente de correlación canónica**, que a su vez permite determinar la importancia de las funciones discriminantes. Para la i-ésima función discriminante, el coeficiente de correlación canónica denotado por γ_i puede expresarse a partir del eigenvalor de la función i de la manera siguiente:

$$\gamma_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

El coeficiente de correlación canónica puede interpretarse como una medida de asociación entre el conjunto de las variables discriminantes y el conjunto que surge de representar los g grupos mediante $(g-1)$ variables dicotómicas (dummy). Así, un coeficiente de correlación alto, indicará que existe una relación entre la función discriminante y los grupos, es decir que la función tiene un buen poder de discriminación entre los grupos.

Desde el punto de vista del análisis de la varianza, el coeficiente de correlación canónica, que en esta caso recibe el nombre de **coeficiente eta**, mide el grado en que difieren las medias alcanzadas por la función discriminante en los grupos. En este caso se consideran como variables independientes los grupos y como variable dependiente la función discriminante. Desde este punto de vista tendríamos

$$\eta^2 = \frac{\text{Suma de cuadrados intergrupos}}{\text{Suma de cuadrados total}}$$

El coeficiente η^2 representaría entonces el porcentaje de varianza de la función discriminante explicada por la diferencia entre grupos.

Si bien el porcentaje relativo indica que tan importante es la función respecto a las demás, el coeficiente de correlación o coeficiente eta indica el grado de relevancia que esta tiene. Un coeficiente de correlación bajo, nos indicará que la función puede ser rechazada, aunque esta función tenga un porcentaje relativo importante, lo que sucedería en el caso de que los grupos no estén lo suficientemente diferenciados respecto a las variables tomadas en consideración.

Lambda de Wilks

El estadístico **lambda Wilks**, denotado por $\mathbf{\Lambda}$, es una medida de las diferencias entre los grupos debidas a varias funciones discriminantes, que considera las desviaciones de las puntuaciones discriminantes dentro de los grupos respecto a la desviación total sin considerar

grupos. Si el valor de lambda es cercano a 1, esto indicará que los grupos están poco separados. Este estadístico es generalmente usado para medir la discriminación que obtendríamos al eliminar una función discriminante del modelo, lo que nos permitiría en caso de no resultar significativa, eliminar una función discriminante. Como vimos en (2.66) el estadístico lambda puede escribirse como:

$$\mathbf{\Lambda} = \prod_{j=1}^{q} \frac{1}{1+\lambda_j}$$

donde $\lambda_j, j = 1, .., q$ son los eigenvalores de $\mathbf{W}^{-1}\mathbf{B}$

Si el valor de lambda es significativo, consideraremos que la primera función discriminante (que es la correspondiente al eigenvalor más grande) debe conservarse, después de esto se obtiene el valor de lambda sin incluir a la primer función es decir:

$$\mathbf{\Lambda} = \prod_{j=2}^{q} \frac{1}{1+\lambda_j}$$

En general, el valor de lambda después de la extracción de r funciones discriminantes está dado por

$$\mathbf{\Lambda} = \prod_{j=r+1}^{q} \frac{1}{1+\lambda_j}$$

Si después de extraer r funciones el residuo no resulta significativo, concluiremos que sólo las primeras r funciones deben de ser consideradas, en otras palabras, que bastará con r dimensiones para explicar las diferencias entre los grupos.

Como sabemos, el estadístico lambda toma valores entre cero y uno, y los valores cercanos a uno indican una escasa discriminación, sin embargo para determinar si este valor es significativo o no, nos basaremos en la transformación de lambda vista en el capitulo anterior(2.67), que tiene una distribución Ji-cuadrada con $p(g-1)$ grados de libertad. Esta transformación está dada por:

$$\mathbf{V} = -\left[n - 1 - \frac{(p+g)}{2}\right] \ln \mathbf{\Lambda}$$

Ahora bien dado que

$$\mathbf{\Lambda} = \prod_{j=1}^{q} \frac{1}{1+\lambda_j}$$

Tenemos que

$$\begin{aligned}\mathbf{V}= - & \left[n-1-\frac{(p+g)}{2}\right]\ln\left(\prod_{j=1}^{q}\frac{1}{1+\lambda_j}\right)\\ = & \left[n-1-\frac{(p+g)}{2}\right]\sum_{j=1}^{q}\ln(1+\lambda_j)\end{aligned}$$

Y dado que todas las funciones discriminantes no están correlacionadas, los términos $\ln(1+\lambda_j)$ de la expresión anterior son estadísticamente independientes y por lo tanto cada uno tiene una distribución aproximada Ji-cuadrada con $(p+g-2j)$ grados de libertad [5]. En general tendríamos, para el $j-$ésimo componente:

$$V_j = \left[n-1-\frac{(p+g)}{2}\right]\ln(1+\lambda_j)$$

Por lo tanto si extraemos V_1, V_2, etc. de V, el residuo tiene una distribución Ji-cuadrada y nos basaremos en estos estadísticos para determinar si la discriminación residual es significativa a un nivel α (generalmente fijado en 0.05) y de no ser así podremos eliminar la función o funciones en cuestión.

A continuación se presentan estos estadísticos residuales así como sus grados de libertad.

Residuo al eliminar	Estadístico aproximado χ^2	Grados de libertad
Primer función	$V - V_1$	$p(g-1)-(p+g-2)=(p-1)(g-2)$
Primeras dos funciones	$V - V_1 - V_2$	$(p-2)(g-3)$
Primeras tres funciones	$V - V_1 - V_2 - V_3$	$(p-3)(g-4)$
$\vdots$	$\vdots$	$\vdots$

Tabla 3.1 Estadísticos residuales χ^2 , con sus grados de libertad

Existen otros criterios para determinar la significancia en las diferencias de los grupos debidas a las funciones discriminantes, como por ejemplo el criterio de Bartlett-Pillai, el criterio de Roy o el criterio de Hotelling-Lawley, sin embargo para la aplicación del siguiente capítulo, nos basaremos en el criterio basado en la lambda de Wilks presentado en esta sección, y el estudio de los demás es dejado al lector.[6]

[5] Ver Apéndice B

[6] Ver por ejemplo Huberty "Applied Discriminant Analysis", 1994 Ed. John Wiley and Sons.

3.6.2 Interpretación de las funciones discriminantes

Una vez determinadas las funciones discriminantes significativas, se procede a la interpretación de las mismas, tomando en cuenta la posición que determinan para los casos así como para los centroides de cada grupo. Además, se analizará la relación entre las variables y las funciones discriminantes, para así poder medir la contribución de cada variable a la discriminación.

Puntuaciones discriminantes

Por medio del cálculo de las puntuaciones discriminantes para cada caso, podemos determinar las coordenadas de cada individuo sobre los ejes discriminantes. Sin embargo, para estudiar el comportamiento de los grupos, será conveniente calcular estas puntuaciones discriminantes para los centroides, más que para los casos aislados. Este cálculo se realiza por medio de los coeficientes no estandarizados $\mathbf{a}_i$ (3.2) y permite determinar la posición del centroide de cada grupo en el espacio discriminante. Si contamos con varios ejes discriminantes, resultará complicado determinar la lejanía o cercanía de los centroides, por lo que en este caso se recurrirá a procedimientos gráficos que ayuden a la visualización de las posiciones relativas tanto de los centroides como de los casos individuales. Entre estos procedimientos gráficos tenemos los histogramas totales y por grupo o bien los diagramas de dispersión.

Contribución de las variables

Los coeficientes no estandarizados pueden ser interpretados como la contribución absoluta de una variable a la determinación de la puntuación discriminante, sin embargo estos coeficientes no son comparables entre sí debido a la diferencia en las unidades de medida de las variables así como su variación. Sin embargo, es posible estandarizar estos coeficientes, como se mencionó en (3.13). Así obtendríamos:

$$\mathbf{a}_i^* = \frac{\sqrt{w_{ii}}}{n-g}\mathbf{a}_i \quad i = 1, ..., p \tag{3.43}$$

Por medio de estos coeficientes estandarizados, es posible comparar la contribución de cada variable a la discriminación a lo largo de las diferentes funciones discriminantes.

La contribución de una variable a las funciones discriminantes también puede medirse en términos de la correlación de Pearson entre las puntuaciones de las variables y las puntuaciones discriminantes.

Estas correlaciones se conocen también como **coeficientes de estructura** y pueden calcularse a partir de la siguiente fórmula:

$$s'_{ij} = \sum_{k=1}^{p} r'_{ik}\mathbf{a}^*_{kj} = \sum_{k=1}^{p} \frac{w_{ik}\mathbf{a}^*_{kj}}{\sqrt{w_{ii}w_{kk}}} \tag{3.44}$$

donde

s'_{ij} es el coeficiente de estructura intragrupos para la variable i y la función j

r'_{ik} es el coeficiente de estructura intragrupos entre la variable i y la variable k

$\mathbf{a}^*_{kj}$ es el coeficiente estandarizado correspondiente a la variable k en la función j

Estos coeficientes de estructura toman valores entre –1 y 1, un coeficiente cercano a estos valores indicará que la variable contribuye de manera importante a la función, mientras que un valor cercano a cero indicará que la variable no aporta información relevante a la función.

La contribución de una variable a la función discriminante puede medirse entonces a través de los coeficientes estandarizados, que representan la contribución de cada variable a las puntuaciones discriminantes, o a través de los coeficientes de estructura que miden la correlación entre las variables y las funciones discriminantes.

3.6.3 Clasificación de los individuos

Una vez realizada la parte descriptiva del análisis discriminante, ya se cuenta con la información de cómo están separados los grupos y de cuales son las variables que más contribuyen a la discriminación, sin embargo, si el problema se centra en la obtención de reglas para la clasificación de los individuos, esta información resultará poco interesante y el interés principal serán los resultados de clasificación. Como hemos visto en la sección 3.4, existen diferentes reglas de clasificación, la que utilizaremos para la clasificación de los individuos en el siguiente capítulo se basa en la regla de máxima probabilidad de Bayes y se realiza a través de las funciones discriminantes. Una vez realizada la clasificación, habrá que determinar que tan precisa es por medio de la cuantificación de los casos correctamente clasificados, pero también de los errores. De este modo podremos evaluar si esta regla es aplicable a nuevos individuos y saber que resultados pueden esperarse en este caso.

Al aplicar la regla de clasificación a los individuos en los que nos basamos para construir esta regla, podemos comparar los resultados obtenidos con los datos observados. Esta comparación se resume en la llamada **matriz de clasificación** o **matriz de confusión**, en ella se presentan los casos que resultaron clasificados en cada grupo, así como los esperados.

La estructura de esta matriz para el caso de tres grupos se muestra en la siguiente tabla:

		Grupo esperado			
		Grupo 1	Grupo 2	Grupo 3	Total
Grupo observado	Grupo1	n_{11}	n_{12}	n_{13}	n_1
	Grupo 2	n_{21}	n_{22}	n_{23}	n_2
	Grupo 3	n_{31}	n_{32}	n_{33}	n_3
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Tabla 3.2: Estructura de la matriz de clasificación

donde n_{ij} es el número de casos del grupo i, adscritos al grupo j por medio de la regla de clasificación.

Por medio de esta matriz es posible identificar los casos clasificados correctamente, que son los que se sitúan en la diagonal descendente, así como los errores que se están cometiendo. El paquete estadístico SPSS, arroja en su salida esta matriz, y en ella se presentan los casos en valores absolutos así como en porcentajes. Si el porcentaje de clasificación correcta total (para tres grupos estaría dado por $(n_{11} + n_{22} + n_{33})/n$) es alto, podríamos esperar que al aplicar esta regla a nuevos individuos se obtengan buenos resultados. El porcentaje de clasificación correcta es una medida de bondad de ajuste de la clasificación, pero también es un reflejo de las diferencias entre los grupos. Una buena clasificación indica que las variables permiten diferenciar correctamente los grupos, sin embargo una mala clasificación indicará que los grupos no están suficientemente diferenciados respecto a las variables discriminantes. Por ejemplo, en el caso de dos grupos, si estos están originalmente solapados o poco diferenciados, las probabilidades de error en la clasificación son grandes como se muestra en la siguiente figura:

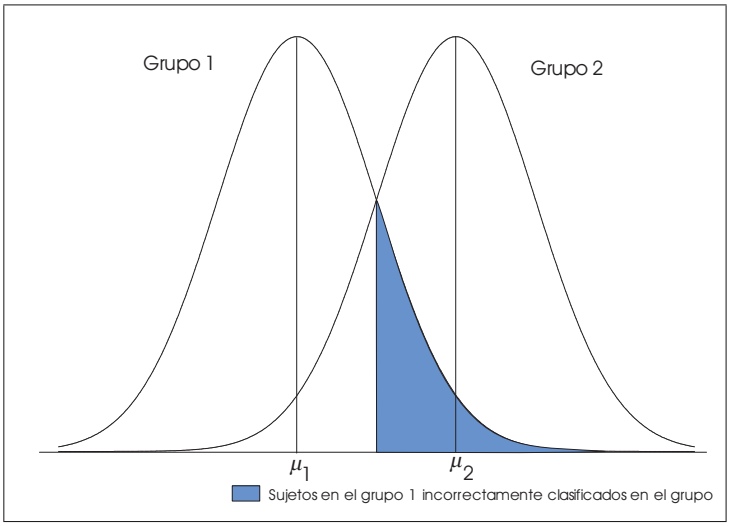


Figura 3.4: Errores de clasificación para dos grupos mal separados

El porcentaje de clasificación correcta nos da una idea de que tan buena es la regla de clasificación, sin embargo este porcentaje puede resultar optimista, ya que el modelo de discriminación se ajustará mejor a la muestra que sirvió para construirlo que a cualquier otra muestra extraída de la población. Por ello, existen algunos métodos para obtener una mejor estimación de los errores de clasificación.

- Uno de estos métodos, consiste en la división de la muestra en dos submuestras, una empleada para construir el modelo y la otra para la validación de la bondad de la clasificación. El inconveniente de este método es que buena parte de la información útil para la construcción de las funciones discriminantes queda inutilizada y además supone la utilización de muestras grandes.

- Otro método para reflejar con más realismo la efectividad de las funciones discriminantes es el método de "dejar uno fuera" o "leave-one-out" (Lachenbruch (1967)) que consiste en extraer un caso de la muestra y obtener las funciones discriminantes a partir de los $n-1$ casos restantes. El caso extraído es entonces clasificado con las funciones discriminantes obtenidas por los demás casos. Este procedimiento se realiza para los n casos, y los individuos no toman parte en la construcción de las funciones que permitirán su clasificación. Este método es empleado generalmente para muestras pequeñas o no muy grandes, ya que en estas últimas los resultados de la clasificación pueden diferir de los obtenidos originalmente.

Otro aspecto que podemos considerar para estimar la bondad de la clasificación, es que tanto mejora la clasificación por medio de las funciones discriminantes respecto a una clasificación realizada al azar. Por ejemplo en el caso de dos grupos, al clasificar a los individuos al azar (por ejemplo el lanzamiento de una moneda), esperaríamos obtener un 50% de los individuos correctamente clasificados, y si por medio de las funciones discriminantes obtenemos un porcentaje digamos de 60%, esto nos indicaría que nuestra regla de clasificación no es buena y esta muy cercana a la distribución al azar. Podemos medir este error de clasificación a través del estadístico

$$\tau = \frac{n_c - \sum_{i=1}^{g} \pi_i n_i}{n - \sum_{i=1}^{g} \pi_i n_i}$$

donde n_c es el número de casos correctamente clasificados, π_i es la probabilidad a priori de pertenencia al grupo i.

En efecto el valor de τ indicará la proporción en que se reduce el error de clasificación frente al error que se cometería por medio de una clasificación al azar. Un valor cercano a uno indicaría que no existen errores de clasificación.

En lo que respecta a los individuos mal clasificados, es decir aquellos fuera de la diagonal de la matriz de confusión, hay que tener en cuenta, como se mencionaba al inicio del capítulo,

que algunos de los errores cometidos, en determinadas circunstancias pueden tener serias repercusiones. Es el caso por ejemplo de pronosticar que un anestésico es seguro para un paciente para él que no lo es. Por ello, en muchas ocasiones, se incluyen en el análisis costos de clasificación [7]. El paquete estadístico SPSS no ofrece esta opción, y en el caso de la aplicación que realizaremos supondremos que todos los errores tienen el mismo costo.

[7] Ver por ejemplo Huberty "Applied Discriminant Analysis", 1994 Ed. John Wiley and Sons.

Capítulo 4

Aplicación

4.1 Introducción

Una vez adquiridos los conocimientos de la teoría del análisis discriminante, pueden realizarse distintas aplicaciones de éste en diferentes ámbitos, como se mencionó en el capítulo anterior. En lo que concierne a este trabajo, la aplicación se realizará sobre el ingreso per cápita en los hogares. Los objetivos de ésta aplicación son determinar los factores que influyen en el nivel de ingreso per cápita así como obtener una regla que permita la clasificación de los hogares en grupos correspondientes a distintos niveles de ingreso.

El ingreso es una variable muy importante que se utiliza con diversos propósitos, como puede ser el estudiar las diferencias de ingreso en una población determinada, es decir, evaluar la inequidad o equidad en la distribución del ingreso. Dado que dicha inequidad es actualmente uno de los problemas sociales más frecuentes en los países en vías de desarrollo, su evaluación resulta crítica para la implementación adecuada de programas prioritarios de desarrollo. De igual manera, el ingreso es un factor determinante en diversos problemas sociales tales como el analfabetismo, la deserción o rendimiento escolar, la desnutrición, la incidencia y prevalencia de ciertas enfermedades, entre otras. Resulta entonces deseable estimar con la mayor precisión posible el ingreso de las personas para una población dada.

Si bien el ingreso es una variable continua, es muchas veces necesario categorizarla para facilitar así su manejo. Existen diversas maneras de categorizar el ingreso, se pueden definir tantas categorías como se desee y los umbrales que las demarcan pueden ser establecidos bajo distintos criterios. La división del ingreso en rangos resultará entonces subjetiva puesto que el investigador podrá generar tantos grupos como desee y desplazar los umbrales entre un grupo y otro de manera tal que las categorías se adapten a sus necesidades.

Por ejemplo, es común en muchos estudios socio-demográficos dividir a la población en dos grupos, los "pobres" y los "no pobres", sin embargo, el concepto de pobreza es relativo al tiempo y a las distintas culturas o sociedades, además de que existen diferentes tipos de pobreza, tales como la pobreza en ingreso, la pobreza en salud, la pobreza en educación, la pobreza cultural, etc. Categorizar entonces el ingreso para obtener dos grupos de la población, los "pobres" y los "no pobres", depende del criterio utilizado para definir "pobreza". Podría basarse por ejemplo en el precio de la canasta básica y dividir así a aquellos hogares cuyo ingreso es insuficiente para adquirir la canasta básica de bienes y aquellos cuyo ingreso es suficiente para adquirirla. En 2002, la Secretaría de Desarrollo Social (SEDESOL), a través del Comité Técnico para la medición de la Pobreza,[1] propuso una metodología para la medición de la pobreza monetaria en México, en efecto se definen tres líneas de pobreza:

- Línea de pobreza 1: Considera a todos aquellos hogares cuyo ingreso es insuficiente como para cubrir las necesidades mínimas de alimentación.
- Línea de pobreza 2: Incluye a los hogares cuyo ingreso es insuficiente como para cubrir las necesidades de alimentación, así como para sufragar los gastos mínimos en educación y salud.
- Línea de pobreza 3: Se refiere a todos aquellos hogares cuyo ingreso es insuficiente como para cubrir las necesidades de alimentación, salud, educación, vestido, calzado, vivienda y transporte público.

Otra categorización frecuente del ingreso es aquella en la que se generan deciles, quintiles, cuartiles, terciles, etc. respecto a la población, es decir, se ordena la población en un orden creciente respecto al ingreso y se divide en n-grupos del mismo tamaño. Siendo la más frecuente la categorización por deciles, empleada por ejemplo por el Instituto Nacional de Estadística y Geografía e Informática (INEGI), para reportar la repartición del ingreso en la población. Del mismo modo, es posible categorizar el ingreso generando percentiles a partir de su distribución empírica.

Así mismo, es frecuente dividir el ingreso en rangos de salarios mínimos, tal como se hará en el presente trabajo. Al categorizar el ingreso, el investigador puede definir diversos umbrales por ejemplo podríamos encontrar categorización con base en salarios mínimos por zonas, en salarios mínimos de los hogares o per cápita, así como con diferentes umbrales, por ejemplo de 0 -2, de 2 - 4, de 4 - 8, de 8 -14 y más de 14 salarios mínimos, o bien, de 0-1, 1-3, 3-5, 5-7, 7-10 y más de 10 salarios mínimos.

[1] Ver: Comité Técnico para la Medición de la Pobreza. "Medición de la Pobreza: Variantes metodológicas y estimación preliminar". Serie de Documentos de Investigación, No. 1, SEDESOL, 2002.

Para realizar la aplicación propuesta en este trabajo, como se detallará más adelante, se optó por utilizar el ingreso mensual per cápita en el hogar y una categorización de éste en cinco grupos cuyos umbrales se definen de la siguiente manera:

- Grupo 1: Menos de 1 salario mínimo.
- Grupo 2: De 1 a 2 salarios mínimos.
- Grupo 3: De 2 a 4 salarios mínimos.
- Grupo 4: De 4 a 6 salarios mínimos.
- Grupo 5: Más de 6 salarios mínimos.

En un principio, se pensaba utilizar para la realización de esta aplicación el criterio propuesto por SEDESOL para distinguir entre la población "pobre" y la "no pobre", sin embargo, ésta clasificación no será empleada ya que si bien distingue tres categorías dentro de la población "pobre", en la población "no pobre" quedan mezclados aquellos hogares cuyo ingreso es bajo pese a haber cruzado el umbral propuesto de "pobreza", aquellos cuyo ingreso es relativamente alto y aquellos cuyo ingreso es muy elevado. En lo que respecta a la categorización por percentiles de población, ésta proporciona una aproximación de la posición que ocupan los hogares en la distribución del ingreso dentro de la población, es decir, el nivel de ingreso se aproxima mediante su posición dentro de la población y no de manera nominal. Por ello se optó por considerar una categorización con base en salarios mínimos, que proporciona una gama más amplia de los niveles de ingreso y refleja la mala distribución del ingreso en México como se detallará más adelante.

Cualquiera que sea la categorización empleada, es de esperarse que los datos reflejen la mala distribución del ingreso que hay en México, y dado que se trata de una variable continua que fue categorizada, los grupos tendrán tendencia a solaparse, es decir que los grupos no están completamente diferenciados entre sí, por ejemplo un hogar cuyo ingreso mensual per cápita es de 1 salario mínimo tendrá características similares de aquel cuyo ingreso es de 1.2 salarios mínimos mensuales per cápita. Además que el problema resulta bastante complejo ya que el comportamiento del ingreso suele ser en ocasiones bastante errático. Este análisis podría del mismo modo realizarse por medio de una regresión lineal que estimara el ingreso y después categorizar la variable resultante, sin embargo uno de los propósitos principales del presente trabajo es mostrar la aplicación del análisis discriminante, por lo que la comparación con la regresión lineal se deja al lector.

Uno de los aspectos difíciles de captar por medio de una encuesta es el ingreso en los hogares debido a que los informantes muchas veces no conocen con precisión el ingreso de

los demás miembros del hogar, lo que puede traer como consecuencia que se reporten valores inverosímiles o incorrectos. Muchas veces también se reportan valores perdidos ("missing") porque los informantes desconocen el nivel de ingreso de los miembros del hogar o sencillamente no desean responder por cuestiones personales que pueden ser de seguridad o privacidad por ejemplo. Así, en la mayoría de las encuestas, la variable del ingreso tiene valores perdidos, valores inverosímiles y/o incorrectos.

Por ello, en el presente trabajo, se propone realizar una aplicación para clasificar a los hogares en cinco grupos correspondientes a distintos niveles de ingreso, utilizando únicamente características de ingreso fijo del hogar y algunas de sus características socioeconómicas, en particular del jefe del hogar. De obtenerse una tasa de clasificación correcta elevada, esto permitiría aproximar el nivel de ingreso per cápita en encuestas que no cuenten con esta información o cuya información sea incompleta, inverosímil y/o incorrecta.

Para toda aplicación, es importante contar con una fuente de información veraz y confiable, por ello, la realización del presente trabajo se llevará a cabo utilizando la Encuesta Nacional de Ingresos y Gastos de los Hogares 2002 (ENIGH 2002), publicada por el INEGI, cuya metodología se presenta en el apéndice A. Se optó por utilizar esta fuente de información por ser una encuesta especializada, como su nombre lo indica, en el ingreso de los hogares, que tiene representatividad a nivel nacional y que es generalmente empleada para los estudios referentes al ingreso de los hogares en el país, por ejemplo, la metodología propuesta por SEDESOL utiliza como una de sus principales fuentes de información la ENIGH 2000. Cabe señalar también que se empleó la ENIGH 2002 por ser la versión más actualizada de las encuestas de ingreso y gasto de los hogares al momento de realizar el presente trabajo. Si bien como toda encuesta ésta puede contener información errónea o incompleta en lo referente al ingreso, la ENIGH 2002 , dada su metodología, resulta ser una fuente confiable en lo referente a la captación del ingreso en los hogares.

En el presente capítulo, se detallarán los aspectos técnicos de la realización de la aplicación, como son la construcción de la variable dependiente, las variables independientes consideradas para la construcción del modelo, así como otras consideraciones sobre los datos tales como los tamaños de muestra y el comportamiento de las variables. Posteriormente, se realizará la comprobación de los supuestos requeridos para la realización del análisis discriminante, para después llevar a cabo el análisis discriminante en sí mismo. Una vez realizado el análisis discriminante se procederá a la interpretación de los resultados, permitiendo así la descripción de los factores que influyen en los distintitos niveles de ingreso, la evaluación de las diferencias entre los grupos propuestos, así como la obtención de una regla de clasificación cuya eficacia deberá ser evaluada.

4.2 Consideraciones sobre los datos

La ENIGH 2002, es una encuesta cuyo objetivo general es proporcionar información sobre la distribución, monto y estructura del ingreso y el gasto de los hogares. Como se detallará en el apéndice A, el diseño muestral de la ENIGH se caracteriza por ser probabilístico, lo que permite generalizar los resultados obtenidos a toda la población de la República Mexicana. El marco muestral de la ENIGH se basa en la información demográfica y cartográfica del conteo de población y vivienda 1995 realizado por el INEGI. La base de datos de la ENIGH 2002 cuenta con 17,167 observaciones de hogares, que, una vez expandidos los datos, representan 24,650,169 hogares, con un promedio de 4.1 personas por hogar, tomando como definición de hogar aquel conjunto de personas, unidas o no por lazos de parentesco, que residen habitualmente en la vivienda y se sostienen de un gasto común para comer.

La encuesta contiene información de las 32 entidades federativas del país que se divide en dos estratos, urbano y rural; siendo una localidad rural aquella en la que haya menos de 2,500 habitantes y una urbana aquella en la que existen más de 2,500 habitantes. En la ENIGH 2002, tomando los datos expandidos, los hogares en áreas rurales representan el 36.5% del total, y aquellos en áreas urbanas el 63.5%.

La encuesta contiene también información sobre el ingreso corriente total en los hogares, que se compone del ingreso corriente monetario y del ingreso corriente no monetario. Para la realización de esta aplicación se consideró únicamente el ingreso corriente monetario de los hogares que es la cantidad de dinero que recibe un perceptor miembro del hogar por su trabajo, utilidades, rendimientos, indemnizaciones y transferencias corrientes, de acuerdo a sus diferentes fuentes [2]

[2] Ver Apéndice A

Las cifras publicadas por el INEGI, con base a la ENIGH 2002, muestran que el ingreso corriente monetario tiene una distribución desigual dentro de la población, como se puede ver en el cuadro siguiente:

Distribución del Ingreso Corriente Monetario Según Deciles de Hogares
(Miles de Pesos)

Deciles	Hogares	Ingreso Corriente Monetario	Porcentaje del ingreso total	
TOTAL	**24,650,169**	**493,997,718**	**100**	**100.00**
I	2,465,017	5,845,104	1.18	11.98
II	2,465,017	12,391,057	2.51	
III	2,465,017	17,919,047	3.63	
IV	2,465,017	23,044,450	4.66	
V	2,465,017	29,007,534	5.87	50.90
VI	2,465,017	36,097,309	7.31	
VII	2,465,017	45,183,146	9.15	
VIII	2,465,017	58,462,506	11.83	
IX	2,465,017	82,683,962	16.74	
X	2,465,016	183,363,603	37.12	37.12

[a] Los hogares a nivel nacional están ordenados en los deciles de acuerdo a su ingreso corriente monetario trimestral.
Los hogares que tuvieron cero ingreso corriente monetario, se clasifican en el primer decil.
FUENTE: **INEGI.** Encuesta Nacional de Ingresos y Gastos de los Hogares, Tercer Trimestre 2002.

Cuadro 4.1: Distribución del Ingreso Corriente Monetario

Como puede observarse, el 37.12% del ingreso corriente monetario total pertenece al último decil de hogares, mientras que los cuatro primeros deciles en conjunto sólo tienen el 11.98%.

La variable dependiente del modelo se construyó entonces a partir del ingreso per cápita mensual en el hogar (ingreso_percápita), que es a su vez el promedio de los ingresos corrientes monetarios totales del hogar de los últimos seis meses, dividido entre el número de personas que conforman el hogar. Para la categorización en rangos de salarios mínimos, se tomó el valor del salario mínimo legal en la República Mexicana (sin distinguir entre las tres zonas de salario) para 2002, que es de 1192.2 pesos mensuales.

Así, los cinco grupos propuestos en la sección anterior para la realización de la presente aplicación quedarían definidos de la siguiente manera:

- Grupo 1: ingreso_percápita <=$1192.2
- Grupo 2: $1192.2 < ingreso_percápita <= $2384.4
- Grupo 3: $2384.4 < ingreso_percápita <= $4768.8
- Grupo 4: $4768.8 < ingreso_percápita <= $7153.2
- Grupo 5: $7153.2 > ingreso_percápita

Al categorizar de esta manera la variable de ingreso obtenemos la siguiente distribución de la población:

Distribución de la población en cinco grupos, con base a salarios mínimos

		Frecuencia	Porcentaje
Grupo 1:	Menos de 1 salario mínimo	12,691,800	51.49
Grupo 2:	Entre 1 y 2 salarios mínimos	6,517,506	26.44
Grupo 3:	Entre 2 y 4 salarios mínimos	3,470,300	14.08
Grupo 4:	Entre 4 y 6 salarios mínimos	921,971	3.74
Grupo 5:	Más de 6 salarios mínimos	1,048,592	4.25
	Total	**24,650,169**	**100**

Cuadro 4.2: Distribución de la población en cinco grupos, con base a salarios mínimos

Como se puede observar en el cuadro anterior, el primer grupo es el que tiene una mayor concentración de hogares, esto debido a la mala distribución del ingreso mencionada anteriormente. En un principio, se consideró la categorización 0 - 2, de 2 - 4, de 4 - 8, de 8 - 14 y más de 14 salarios mínimos para realizar la presente aplicación del análisis discriminante, sin embargo, el primer grupo presentaba una concentración aún mayor de la población que el grupo propuesto de 0-1 salarios mínimos, además de que la barrera de 1 salario mínimo puede ser más representativa.

En la metodología propuesta por SEDESOL, se considera a una persona "no pobre" (en referencia a la Línea de pobreza 3) si su ingreso diario es equivalente a 41.8 pesos[3], tomando como referencia un mes de 30 días, esto sería equivalente a un ingreso mensual per cápita de 1,254 pesos, que es un valor no muy lejano al de un salario mínimo. Cabe señalar que estos umbrales son en términos del ingreso per cápita, por lo que bajo el criterio de SEDESOL, el ingreso mensual de un hogar de 4 personas debería ser superior a 5,016 pesos, si en ese hogar existe un solo perceptor, éste debería ganar aproximadamente 4.2 veces el salario mínimo para que dicho hogar dejará de ser considerado como pobre. En lo referente al último grupo, se optó por elegir el rango de los ingresos mayores a 6 salarios mínimos, ya que si este umbral se elevará, la proporción de la población en el último grupo quedaría aún más reducida.

Las variables que se consideraron para ser incluidas en el modelo como variables dependientes pueden dividirse básicamente en dos grupos, las características del hogar y sus miembros, en particular el jefe de familia; y las características de las viviendas y su equipamiento. Muchas de las variables contenidas en la ENIGH son categóricas, como suele ocurrir en la práctica, por lo que para poder ser empleadas para el análisis discriminante fueron recodificadas variables dicotómicas ("dummy"), algunas otras variables se generaron a partir de variables contenidas en la base de datos.

A continuación se presenta una lista de todas las variables dependientes consideradas para ser incluidas en el modelo, especificando la recodificación o cálculos realizados sobre ellas. Para las variables categóricas (más de dos categorías) se especifican los nombres de las variables dicotómicas creadas ("Dicotómicas"), tomando como códigos de estas variables, 1 si posee la categoría en cuestión, 0 de lo contrario.

[3] Pesos del 2000 en áreas urbanas

Características de los miembros del hogar, ENIGH 2002

Variable	Pregunta	Opciones	Códigos	Dicotómicas
Zona	Zona geográfica del país**	Norte	1	Centro
		Centro	2	Norte
		Sur	3	Sur
cla_hog	Clase de hogar	Hogar unipersonal	1	cla_hog1
		Hogar nuclear	2	cla_hog2
		Hogar ampliado	3	cla_hog3
		Hogar compuesto	4	cla_hog4
		Corresidentes	5	cla_hog5
urbano	Estrato	Urbano	1	
		Rural	0	
tam_hog	Número de personas en el hogar			
hacina	Indice de hacinamiento	personas/número de cuartos		
depend	Índice de dependencia económica	personas/ número de perceptores		
menores	Niños menores de 12 años			
nintraba	Niños entre 12 y 15 años que trabajan			
analfabe	Personas de 15 años y más analfabetas			
smedich	Al menos una persona en el hogar tiene servicio médico	Sí	1	
		No	0	

**Nota: Se consideran las zonas geográficas de acuerdo a la siguiente clasificación de las entidades federativas:

Zona Centro	Zona Sur	Zona Norte
01- Aguascalientes	04- Campeche	02- Baja California
06- Colima	07- Chiapas	03- Baja California Sur
09- Distrito Federal	12- Guerrero	05- Coahuila
11- Guanajuato	17- Morelos	08- Chihuahua
13- Hidalgo	20- Oaxaca	10- Durango
14- Jalisco	21- Puebla	18- Nayarit
15- México	23- Quintana Roo	19- Nuevo León
16- Michoacán	27- Tabasco	24- San Luis Potosí
22- Querétaro	30- Veracruz	25- Sinaloa
29- Tlaxcala	31- Yucatán	26- Sonora
		28- Tamaulipas
		32- Zacatecas

Cuadro 4.3.1: Características de los miembros del Hogar

Características del Jefe del hogar, ENIGH 2002

Variable	Pregunta	Opciones	Códigos	Dicotómicas
sexo_jef	Sexo del jefe de familia	Hombre	1	
		Mujer	0	
edad_jef	Edad del jefe de familia	Edad en años		
jef_alfa	¿El jefe de familia sabe leer y escribir?	Sí	1	
		No	0	
edociv	Estado civil del Jefe de familia	Unión libre	1	edo_civ1
		Casado	2	edo_civ2
		Separado	3	edo_civ3
		Divorciado	4	edo_civ4
		Viudo	5	edo_civ5
		Soltero	6	edo_civ6
educ	Educación del jefe de familia	Ninguna	0	edu_0
		Preprimaria	1	
		Primaria	2	
		Secundaria	3	edu_1
		Preparatoria, vocacional, normal o técnica	4	
		Universidad	5	edu_2
		Maestría, Doctorado	6	
jno_trab	El jefe de familia trabajó el mes pasado (por salario, por su cuenta, etc)	Sí	1	
		No	0	
servmed	El jefe de familia tiene derecho a servicios médicos del IMSS, ISSSTE,	Sí	1	
		No	0	
otraspre	El jefe de familia tiene derecho a otras prestaciones como aguinaldo, vacaciones	Sí	1	
		No	0	
prestac0	El jefe de familia no tiene derecho a prestaciones	Sí	1	
		No	0	

Cuadro 4.3.2 : Características del Jefe del Hogar

Características y equipamiento de la Vivienda, ENIGH 2002

Variable	Pregunta	Opciones	Códigos	Dicotómicas
tipo_viv	Tipo de vivienda	Casa sola que comparte muros	1	tviv1
		Casa sola que no comparte muros	2	
		Departamento en condominio horizontal	3	tviv3
		Departamento en edificio o condominio vertical	4	tviv4
		Departamento que comparte el servicio sanitario	5	tviv5
		Cuarto de azotea, local usado como vivienda, vivienda móvil, refugio	6	tviv6
tenencia	¿Esta vivienda es...	prestada?	1	viv_1
		recibida como prestación?	2	viv_2
		rentada o alquilada?	3	viv_3
		propia y la están pagando?	4	viv_4
		propia en terreno de asentamiento irregular?	5	viv_5
		propia en terreno ejidal o comunal?	6	viv_6
		propia y totalmente pagada en terreno propio?	7	viv_7
		propia y construida en terreno prestado	8	viv_8
		otro tipo de tenencia?	9	viv_9
cocina	¿Esta vivienda tiene cuarto para cocinar?	Sí	1	
		No	0	
cocinad	¿En el cuarto para cocinar también duermen?	Sí	1	
		No	0	
muros	¿De qué material es la mayor parte de las paredes o muros exteriores de esta vivienda?	Cartón, hule, tela, llantas, lámina de cartón, carrizo bambú, palma, tejamanil, embarro o bajareque	1	muros1
		Lámina de asbesto o metálica, fibra de vidrio, plástico o mica, vidrio o cristal	2	muros2
		Madera	3	muros3
		Adobe	4	muros4
		Panel de concreto, concreto monolítico, tabique, ladrillo, tabicón, block, piedra o cemento (incluye cantera)	5	muros5
		Otro material	6	muros6

Cuadro 4.3.3 : Características y equipamiento de la vivienda

Características y equipamiento de la Vivienda, ENIGH 2002

Variable	Pregunta	Opciones	Códigos	Dicotómicas
murosr	¿Los muros exteriores de la vivienda tienen algún recubrimiento por la parte externa?	Sí No	1 0	
techos	¿De qué material es la mayor parte de los techos de esta vivienda?	Cartón, hule, tela, llantas, lámina de cartón, palma. Tejamanil, Carrizo o bambú	1	techo1
		Madera	2	techo2
		Lámina metálica, fibra de vidrio, plástico o mica	3	techo3
		Terrado, teja, lámina de asbesto	4	techo4
		Panel de concreto, concreto monolítico, tabique, ladrillo, tabicón, block, losa de concreto	5	techo5
		Vigueta y poliuretano, vigueta y bovedilla, vigueta y cuña	6	techo6
		otros materiales	7	
techosr	¿Los techos de la vivienda tienen algún recubrimiento por la parte externa?	Sí	1	
		No	0	
pisos	¿De qué material es la mayor parte de los pisos de esta vivienda?	Tierra	1	piso1
		Cemento o firme	2	piso2
		Mosaico o terrazo	3	piso3
		Loseta de vinil o plástico, linóleum o congóleum	4	piso4
		Loseta de cemento (vitropiso), mármol	5	piso5
		Madera, duela o parquet	6	piso6
		Otros recubrimientos como alfombra etc.	7	piso7
agua	¿La vivienda tiene agua entubada...	dentro de la vivienda?	1	agua1
		fuera de la vivienda pero dentro del edificio o terreno?	2	agua2
		No dispone de agua entubada	3	agua3
drenaje	¿Esta vivienda cuenta con drenaje para el desalojo de las aguas jabonosas del fregadero, regadera, lavabo o lavadero ?	Sí No	1 0	
cbaño	¿Esta vivienda tiene cuarto de baño?	Sí No	1 0	
baño	¿Esta vivienda tiene...	hoyo negro o pozo ciego?	1	bano1
		letrina?	2	bano2
		excusado?	3	bano3
		no dispone del servicio sanitario?	4	bano4

Cuadro 4.3.3 : Características y equipamiento de la vivienda (Continuación 1)

Características y equipamiento de la Vivienda, ENIGH 2002

Variable	Pregunta	Opciones	Códigos	Dicotómicas
baño2	¿ A la letrina, excusado o sanitario...	Le echa agua con cubeta a la letrina, excusado o sanitario....	1	banoa
		La letrina excusado o sanitario tiene conexión de agua	2	banob
		No le echa agua a la letrina, excusado o sanitario...	3	banoc
basura	¿Habitualmente que hace con la basura?	Tiran la basura al río	1	basura1
		Queman la basura	2	basura2
		Tiran la basura en un terreno baldío o a la calle	3	basura3
		Entierran la basura	4	basura4
		Tiran la basura en el basurero público	5	basura5
		Utiliza el servicio de recolección público de la basura	6	basura6
		Utiliza el servicio de recolección particular	7	basura7
		Recicla la basura, la vende, la regala, elabora productos con los desechos etc.	8	basura8
combust	¿Habitualmente qué combustible utiliza para cocinar o calentar sus alimentos?	Leña	1	combust1
		Carbón	2	combust2
		Petróleo	3	combust3
		Electricidad	4	combust4
		Gas	5	combust5
		Otros	6	combust6
		No utiliza combustible	7	combust7
luz	¿Esta vivienda tiene luz eléctrica?	Sí	1	
		No	0	
luzf	¿De donde obtiene la luz eléctrica?	De un acumulador De una planta particular Del servicio publico De otra fuente	1 2 3 4	luz1 luz2 luz3 luz4
luz_ins	¿En todos los cuartos de la vivienda hay instalaciones fijas para focos?	Sí	1	
		No	0	
luz_con	¿Tiene contrato de luz?	Sí	1	
		No	0	
tinaco	¿Esta vivienda cuenta con tinaco(s)?	Sí	1	
		No	0	
cisterna	¿Esta vivienda cuenta con cisterna o aljibe?	Sí	1	
		No	0	
bomba	¿Esta vivienda cuenta con bomba de agua?	Sí	1	
		No	0	

Cuadro 4.3.3 : Características y equipamiento de la vivienda (Continuación 2)

Características y equipamiento de la Vivienda, ENIGH 2002

Variable	Pregunta	Opciones	Códigos
calentad	¿Esta vivienda cuenta con calentador de gas?	Sí	1
		No	0
s_aireac	¿Esta vivienda cuenta con sistema de aire acondicionado?	Sí	1
		No	0
s_calef	¿Esta vivienda cuenta con sistema de calefacción?	Sí	1
		No	0
lavabo	¿Esta vivienda cuenta con lavabo?	Sí	1
		No	0
regadera	¿Esta vivienda cuenta con regadera?	Sí	1
		No	0
lavadero	¿Esta vivienda cuenta con lavadero?	Sí	1
		No	0
fregader	¿Esta vivienda cuenta con fregadero?	Sí	1
		No	0
closet	¿Esta vivienda cuenta con closet?	Sí	1
		No	0
apoyoviv	En los últimos 6 meses le dieron un crédito para la compra, ampliación mejora de la	Sí	1
		No	0
teléfono	¿Esta vivienda tiene teléfono?	Sí	1
		No	0
celular	¿Cuenta con teléfono celular para el uso del hogar?	Sí	1
		No	0
tvcable	¿Cuenta con televisión por cable, Sky, Direct-tv o multivisión para el uso del hogar?	Sí	1
		No	0
inernet	¿Cuenta con Internet para el uso del hogar?	Sí	1
		No	0
automov	¿Cuentan con Automóvil, camioneta o camioneta de caja para el uso de este hogar?	Sí	1
		No	0
moto	¿Cuentan con motocicleta para el uso de este hogar?	Sí	1
		No	0
bici	¿Cuentan con bicicleta (que utilicen como medio de transporte) para el uso de este	Sí	1
		No	0
radio	¿Cuentan con Radio para el uso de este hogar?	Sí	1
		No	0
grabador	¿Cuenta con grabadora para el uso de este hogar?	Sí	1
		No	0
estereo	¿Cuenta con estéreo, modular o consola para el uso de este hogar?	Sí	1
		No	0
repro_cd	¿Cuenta con reproductor de discos compactos para el uso de este hogar?	Sí	1
		No	0
antena	¿Cuenta con antena parabólica para el uso de este hogar?	Sí	1
		No	0
tv_bn	¿Cuenta con televisión blanco y negro para el uso de este hogar?	Sí	1
		No	0

Cuadro 4.3.3 : Características y equipamiento de la vivienda (Continuación 3)

Características y equipamiento de la Vivienda, ENIGH 2002

Variable	Pregunta	Opciones	Códigos
tv_color	¿Cuenta con televisión a color para el uso de este hogar?	Sí	1
		No	0
video	¿Cuenta con videocasetera para el uso de este hogar?	Sí	1
		No	0
dvd	¿Cuenta con reproductor de video discos (dvd) para el uso de este hogar?	Sí	1
		No	0
compu	¿Cuenta con computadora para el uso de este hogar?	Sí	1
		No	0
impresor	¿Cuenta con impresora para el uso de este hogar?	Sí	1
		No	0
apcompu	¿Cuenta con escáner, quemador de cd, modem, lector de cd, y demás aparatos	Sí	1
		No	0
videoj	¿Cuenta con video juegos (Playstation, nintendo, sega u otros) para el uso de este	Sí	1
		No	0
licuador	¿Cuenta con licuadora para el uso de este hogar?	Sí	1
		No	0
batidor	¿Cuenta con batidora para el uso de este hogar?	Sí	1
		No	0
extracto	¿Cuenta con extractor de jugos para el uso de este hogar?	Sí	1
		No	0
tostador	¿Cuenta con tostador para el uso de este hogar?	Sí	1
		No	0
cafetera	¿Cuenta con cafetera para el uso de este hogar?	Sí	1
		No	0
sandwich	¿Cuenta con sandwichera para el uso de este hogar?	Sí	1
		No	0
jugos	¿Cuenta con exprimidor de jugos para el uso de este hogar?	Sí	1
		No	0
horno_el	¿Cuenta con horno eléctrico para el uso de este hogar?	Sí	1
		No	0
micro	¿Cuenta con horno de microondas para el uso de este hogar?	Sí	1
		No	0
refri	¿Cuenta con refrigerador para el uso de este hogar?	Sí	1
		No	0
estufa_g	¿Cuenta con estufa de gas para el uso de este hogar?	Sí	1
		No	0
molino	¿Cuenta con molino de mano para el uso de este hogar?	Sí	1
		No	0
lavadora	¿Cuenta con lavadora para el uso de este hogar?	Sí	1
		No	0
maq_cos	¿Cuenta con máquina de coser para el uso de este hogar?	Sí	1
		No	0
aire_ac	¿Cuenta con aparato de aire acondicionado (excluya sistema) para el uso de este hogar?	Sí	1
		No	0
calef	¿Cuenta con aparato calefactor (excluya sistema) para el uso de este hogar?	Sí	1
		No	0
aspira	¿Cuenta con aspiradora para el uso de este hogar?	Sí	1
		No	0

Cuadro 4.3.3 : Características y equipamiento de la vivienda (Continuación 4)

Para cada variable categórica de k categorías, se generaron k variables dicotómicas , para algunas de estas variables se eliminaron algunas categorías que no tuvieron incidencias en ningún caso (basura4, basura8, combust2, combust3, combust6, muros6, luz1, luz2, tviv6). Dado que el método de selección de variables que se utilizará es el método stepwise, para cada variable de k categorías, se dejará fuera al menos una categoría de manera tal que las variables no estén correlacionadas.

En la bibliografía del análisis discriminante, como ya se había mencionado, es común encontrar la recomendación de que al realizar una aplicación de éste, se cuenten al menos con 20 casos por cada variable discriminante; en el caso de la presente aplicación esta condición se cumple ampliamente, pese a que se contemplan inicialmente 145 variables discriminantes.

Al realizar una primera exploración de los datos se eliminaron 370 casos (datos no expandidos) que tenían valores perdidos en alguna de las variables discriminantes (por ejemplo hogares en los que el jefe del hogar esta ausente y no se cuenta con la información completa sobre este). Dado el tamaño de la muestra de la encuesta, se prefirió eliminar estos casos en lugar de estimarlos sea remplazándolos por las medias de las variables o realizando alguna regresión sobre ellos.

4.3 Comprobación de supuestos

4.3.1 Outliers

La mayoría de las variables consideradas para ser incluidas en el modelo son categóricas, sin embargo, también se incluyen algunas variables continuas. En un primer paso, se realiza la exploración de los datos para identificar los casos aislados (outliers univariados). Este procedimiento se realiza únicamente para las variables continuas y se consideran outliers aquellos casos que rebasen los puntos de corte $Z = \pm 3$ dentro de cada grupo.

A continuación se presentan los resultados obtenidos de esta primera exploración, presentando los estadísticos descriptivos de las variables en cuestión.

	GRUPO 1						
	N	Media	Desv. Típ	Min.	Max.	Zmin	Zmax
tam_hog	12,421,286	4.67	2.14	1	17	-1.71	**5.75**
hacina	12,421,286	2.33	1.58	0.13	15	-1.39	**8.00**
depend2	12,421,286	1.64	1.49	0	14	-1.10	**8.29**
menores	12,421,286	1.53	1.42	0	9	-1.08	**5.27**
nintraba	12,421,286	0.08	0.31	0	3	-0.25	**9.37**
analfabe	12,421,286	0.46	0.77	0	7	-0.60	**8.50**
edad_jef	12,421,286	47.51	15.91	15	97	-2.04	**3.11**
	GRUPO 2						
	N	Media	Desv. Típ	Min.	Max.	Zmin	Zmax
tam_hog	6,381,513	3.89	1.70	1	17	-1.70	**7.70**
hacina	6,381,513	1.44	0.88	0.1	9	-1.52	**8.58**
depend2	6,381,513	1.26	1.18	0	7	-1.07	**4.86**
menores	6,381,513	0.86	0.98	0	6	-0.87	**5.25**
nintraba	6,381,513	0.03	0.19	0	3	-0.15	**15.73**
analfabe	6,381,513	0.13	0.41	0	4	-0.32	**9.55**
edad_jef	6,381,513	46.96	14.79	14	97	-2.23	**3.38**
	GRUPO 3						
	N	Media	Desv. Típ	Min.	Max.	Zmin	Zmax
tam_hog	3,430,463	3.35	1.48	1	12	-1.58	**5.84**
hacina	3,430,463	1.01	0.56	0.08	6	-1.67	**8.90**
depend2	3,430,463	0.99	1.12	0	7	-0.89	**5.36**
menores	3,430,463	0.58	0.82	0	4	-0.71	**4.19**
nintraba	3,430,463	0.01	0.11	0	2	-0.10	**17.46**
analfabe	3,430,463	0.06	0.27	0	2	-0.21	**7.21**
edad_jef	3,430,463	47.03	14.64	17	94	-2.05	**3.21**
	GRUPO 4						
	N	Media	Desv. Típ	Min.	Max.	Zmin	Zmax
tam_hog	911,180	2.97	1.46	1	9	-1.34	**4.12**
hacina	911,180	0.82	0.46	0.14	3	-1.45	**4.70**
depend2	911,180	0.76	0.96	0	5	-0.79	**4.42**
menores	911,180	0.53	0.79	0	4	-0.67	**4.38**
nintraba	911,180	0.00	0.04	0	1	-0.04	**24.56**
analfabe	911,180	0.03	0.18	0	2	-0.15	**11.00**
edad_jef	911,180	45.34	14.45	18	89	-1.89	**3.02**
	GRUPO 5						
	N	Media	Desv. Típ	Min.	Max.	Zmin	Zmax
tam_hog	1,039,821	2.54	1.32	1	8	-1.17	**4.13**
hacina	1,039,821	0.66	0.51	0.13	5	-1.05	**8.52**
depend2	1,039,821	0.66	0.97	0	7	-0.68	**6.56**
menores	1,039,821	0.33	0.71	0	6	-0.47	**8.00**
nintraba	1,039,821	0.00	0.06	0	1	-0.06	**16.18**
analfabe	1,039,821	0.02	0.20	0	2	-0.11	**9.89**
edad_jef	1,039,821	46.70	14.08	20	86	-1.90	2.79

Cuadro 4.4: Estadísticos para la detección de outliers univariados dentro de cada grupo

Como podemos ver en el cuadro 4.4 existen outliers en todos los grupos y para prácticamente todas las variables. Dado que la muestra es lo suficientemente grande, se decidió eliminar estos casos del análisis. De esta manera, para los outliers univariados, fueron eliminados de análisis 1280 casos.

Una vez eliminados los casos aislados univariados, se procede a la identificación de los casos aislados multivariantes, esta detección puede llevarse a cabo por medio del cálculo de la distancia de Mahalanobis en cada grupo, que mide la distancia entre un caso y el centroide de su grupo. El siguiente cuadro presenta las 25 mayores distancias de Mahalanobis registradas en cada grupo.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
8652.00	1481.70	1936.00	490.00	470.00
1442.83	1442.57	1936.00	490.00	470.00
1431.85	1426.07	1936.00	490.00	470.00
1407.28	1356.79	1022.35	490.00	470.00
1362.41	1222.54	1022.35	490.00	470.00
1360.04	1214.19	885.21	490.00	470.00
1342.91	1158.54	805.74	490.00	470.00
1342.62	1156.85	797.58	490.00	470.00
1333.51	1139.69	759.66	490.00	399.30
1329.48	1103.88	733.53	490.00	399.30
1320.47	1092.07	726.87	490.00	399.30
1310.27	1073.92	721.04	490.00	365.65
1292.27	1072.37	719.25	490.00	365.65
1282.73	1061.81	713.69	357.04	333.84
1276.01	1061.41	689.04	357.04	333.84
1017.93	1055.86	642.06	287.62	333.38
998.89	1047.80	640.20	285.74	328.47
912.32	1046.64	636.45	285.10	323.69
895.97	1035.89	630.48	270.67	317.16
887.80	906.65	629.45	268.40	317.16
859.03	894.41	627.47	268.13	308.16
847.64	866.50	624.89	262.69	301.20
796.21	851.33	624.14	258.50	293.54
793.45	839.64	618.03	256.55	293.54
758.53	831.18	599.88	256.32	273.18

Cuadro 4.5 : 25 Mayores distancias de Mahalanobis dentro de cada grupo

La distancia de Mahalanobis sigue una distribución chi-cuadrada con 156 grados de libertad, el valor crítico para un nivel de significancia de α=0.01 es 200.006. El número de casos que se sitúan por encima de este valor, y por lo tanto son considerados como outliers multivariados, es de 2,232 casos, si bien es una cantidad importante de casos (13% de la muestra original), se decidió eliminar estos casos del análisis dado que la muestra restante sigue siendo de tamaño importante.

De esta manera, una vez eliminados los casos con valores perdidos en las variables discriminantes y los outliers tanto univariados como multivariados, obtenemos una muestra final de 13,285 casos, que una vez expandidos representan 19, 469,032 hogares de la República Mexicana. Si bien el número de casos eliminados es importante (22.6% de la muestra original), la muestra final sigue siendo suficientemente amplia para realizar la aplicación, y si sigue respetando el criterio de tener al menos 20 casos por variable discriminante.

4.3.2 Pruebas de Normalidad

Una vez eliminados los outliers y los casos con valores perdidos, se procede a la comprobación de los supuestos. Uno de los supuestos del análisis discriminante es, como ya habíamos mencionado, el de la distribución Normal Multivariada. Sin embargo, la comprobación de este supuesto es complicada, por lo que generalmente no se realiza, como es el caso del presente trabajo. La normalidad univariante no es una condición suficiente para la normalidad multivariante, sin embargo, si las variables siguen una distribución normal univariada aumenta la probabilidad de que en conjunto sigan una distribución Normal Multivariada. En el presente trabajo, la mayoría de las variables independientes son categóricas y fueron recodificadas a variables dicotómicas para poder ser incluidas en el modelo, sin embargo, también se incluyen algunas variables continuas sobre las cuales realizaremos la comprobación de la normalidad. Para comprobar la normalidad univariada, se utiliza la prueba Kolmogorov- Smirnov, aplicándola a cada variable por separado, en cada uno de los grupos definidos. A continuación se presentan los resultados obtenidos al aplicar esta prueba a las variables continuas incluidas en el modelo.

			tam_hog	hacina	depend2	menores	nintraba	analfabe	edad_jef
Grupo 1	Diferencias más extremas	Absoluta	0.1275	0.1799	0.1864	0.1744	0.5405	0.4153	0.0867
		Positiva	0.1275	0.1799	0.1864	0.1744	0.5405	0.4153	0.0867
		Negativa	-0.0931	-0.0885	-0.1260	-0.1296	-0.4061	-0.2638	-0.0495
	Kolmogorov-Smirnov Z		**10.9142**	**15.4029**	**15.9599**	**14.9318**	**46.2779**	**35.5606**	**7.4278**
	Sig. (bilateral)		**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Grupo 2	Diferencias más extremas	Absoluta	0.1411	0.1575	0.2280	0.2931	*/	0.5337	0.0648
		Positiva	0.1411	0.1575	0.2280	0.2931		0.5337	0.0648
		Negativa	-0.1263	-0.0824	-0.1342	-0.1897		-0.3774	-0.0316
	Kolmogorov-Smirnov Z		**8.2384**	**9.1984**	**13.3148**	**17.1187**		**31.1713**	**3.7829**
	Sig. (bilateral)		**0.0000**	**0.0000**	**0.0000**	**0.0000**		**0.0000**	**0.0000**
Grupo 3	Diferencias más extremas	Absoluta	0.1570	0.1967	0.2336	0.3448	*/	*/	0.0589
		Positiva	0.1379	0.1967	0.2336	0.3448			0.0589
		Negativa	-0.1570	-0.0842	-0.1738	-0.2222			-0.0279
	Kolmogorov-Smirnov Z		**6.4277**	**8.0548**	**9.5658**	**14.1214**			2.4104
	Sig. (bilateral)		**0.0000**	**0.0000**	**0.0000**	**0.0000**			0.0000
Grupo 4	Diferencias más extremas	Absoluta	0.1732	0.1450	0.2221	0.4278	*/	*/	0.0550
		Positiva	0.1732	0.1450	0.2221	0.4278			0.0550
		Negativa	-0.1410	-0.0891	-0.2097	-0.2744			-0.0288
	Kolmogorov-Smirnov Z		**3.6326**	**3.0413**	**4.6598**	**8.9741**			1.1536
	Sig. (bilateral)		**0.0000**	**0.0000**	**0.0000**	**0.0000**			0.1396
Grupo 5	Diferencias más extremas	Absoluta	0.2075	0.1529	0.2867	0.4768	*/	*/	0.0699
		Positiva	0.2075	0.1529	0.2867	0.4768			0.0699
		Negativa	-0.1164	-0.0917	-0.2427	-0.3185			-0.0404
	Kolmogorov-Smirnov Z		**4.2768**	**3.1526**	**5.9102**	**9.8302**			1.4406
	Sig. (bilateral)		**0.0000**	**0.0000**	**0.0000**	**0.0000**			0.0315

Cuadro 4.6: Estadísticos para prueba Kolmogorov-Smirnov

Como podemos ver en el cuadro anterior, los valores de z resultan altos en la mayoría de los casos y presentan probabilidades asociadas inferiores al nivel de significación $\alpha = 0.05$, por ello, se rechaza la hipótesis nula de que la distribución observada no se diferencia de la distribución normal (excepto para la edad del jefe del hogar (edad_jef) en los dos últimos grupos). Sin embargo, pese a que los resultados obtenidos para estas variables reflejan que la distribución univariada de los datos se diferencia de la distribución normal univariada, el resultado debe valorase tomando en cuenta que ligeras separaciones del supuesto de normalidad multivariante no afectan demasiado el resultado del análisis discriminante. Dado que la mayoría de las variables son dicotómicas, podríamos esperar que la distribución multivariada de los datos no se aleje excesivamente de la distribución normal multivariada.

4.3.3 Igualdad de Matrices de Varianza Covarianza

En lo referente a la comprobación del supuesto de igualdad de matrices de varianza covarianza en todos los grupos, esta puede llevarse a cabo por medio de la aplicación de la prueba M de Box, el estadístico M se calcula a partir de los determinantes de las matrices de varianza covarianza de cada uno de los grupos, sin embargo, dada la cantidad de variables incluidas en el modelo, el resultado rápidamente conduciría a rechazar la hipótesis de que las matrices de varianza covarianza son iguales. En el caso de esta aplicación las matrices de varianza covarianza resultan ser de entrada (es decir considerando todas las variables propuestas) singulares, por lo que el cálculo de del estadístico M de Box no puede llevarse a cabo. Aún eliminado variables de modo a obtener matrices de varianza-covarianza no singulares (como se realizará más adelante) lo más probable es que la hipótesis de que no existen diferencias entre las matrices de varianza covarianza sea rechazada, puesto que el estadístico M de Box, como se mencionó en la sección 2.5, es sensible a tamaños de muestra grandes y a pequeñas desviaciones de la distribución Normal. En el caso de la presente aplicación, los tamaños de muestra son grandes y desiguales respecto a los grupos, por lo que probablemente se concluirá que las matrices de varianza covarianza no son iguales, además de que los datos no se ajustan en su totalidad a la distribución Normal; la violación de estos dos supuestos puede llevar a un incremento en el porcentaje de casos mal clasificados por medio del análisis discriminante. Cabe señalar que en la práctica es difícil contar con datos que cumplan estos supuestos en su totalidad.

En lo referente a los supuestos de multicolinealidad y singularidad no serán revisados ya que, como hemos visto en la sección 3.5, al emplear el método "Stepwise" y el criterio de tolerancia para la inclusión de variables se asegura el cumplimiento de estos supuestos. De esta manera, la utilización del método Stepwise permitirá la eliminación de variables de modo tal que la matriz de varianza covarianza sea no singular y el estadístico M de Box podrá calcularse, teniendo en cuenta que muy probablemente se rechazará la hipótesis de igualdad de matrices de varianza covarianza por los motivos previamente expuestos.

4.4 Comportamiento de los datos

Antes de proceder a la obtención de las funciones discriminantes y a la clasificación, se llevará a cabo un análisis del comportamiento de las variables entre los grupos con la finalidad de comprobar si estas variables contribuirán o no a la diferenciación entre los grupos. En un primer acercamiento, se realiza un análisis descriptivo del comportamiento de algunas de las variables, los siguientes cuadros muestran los promedios (porcentajes si la variable es dicotómica) alcanzados por las variables para cada grupo y en total.

COMPORTAMIENTO DE LAS VARIABLES
CARACTERÍSTICAS DE LOS MIEMBROS DEL HOGAR

	Grupo					Total
	1	2	3	4	5	
Estrato Urbano/Rural	45.66%	84.25%	84.68%	88.31%	88.00%	64.96%
Al menos uno tiene servicio médico en el hogar	26.87%	58.86%	68.54%	68.08%	56.74%	44.18%
Número de personas en el hogar*	4.47	3.78	3.36	2.91	2.40	3.97
Índice de hacinamiento*	2.23	1.36	0.96	0.75	0.59	1.69
Índice de dependencia económica*	1.58	1.24	0.98	0.62	0.54	1.32
Niños menores de 12 anos*	1.43	0.80	0.59	0.41	0.24	1.05
Niños entre 12 y 15 años que trabajan*	0.05	0.00	0.00	0.00	0.00	0.02
Personas de 15 años y más analfabetas*	0.42	0.08	0.00	0.00	0.00	0.23
Sexo del jefe del hogar	80.76%	79.24%	77.43%	74.80%	77.04%	79.48%
Alfabetismo del jefe del hogar	79.90%	96.21%	100.00%	100.00%	100.00%	88.76%
Escolaridad del Jefe del hogar (NInguna, Preprimaria, Primaria)	74.56%	47.69%	26.33%	15.26%	4.07%	55.02%
Escolaridad del Jefe del hogar (Secundaria, Preparatoria, Normal)	23.91%	42.56%	39.77%	29.65%	22.11%	31.18%
Escolaridad del Jefe del hogar (Universidad, Maestría, Doctorado)	1.52%	9.75%	33.90%	55.09%	73.82%	13.80%
Estado civil del jefe del hogar (Unión libre)	14.47%	9.92%	5.43%	4.44%	5.14%	11.16%
Estado civil del jefe de familia (Casado(a))	64.22%	66.43%	64.26%	55.40%	56.60%	64.10%
Estado civil del jefe de familia (separado(a))	6.67%	6.93%	7.35%	9.90%	9.92%	7.11%
Estado civil del jefe de familia (divorciado(a))	0.85%	2.23%	4.63%	5.16%	8.16%	2.26%
Estado civil del jefe de familia (viudo(a))	10.44%	9.44%	9.84%	8.71%	6.27%	9.83%
Estado civil del jefe de familia (soltero(a))	3.36%	5.06%	8.48%	16.38%	13.90%	5.54%
El jefe de familia trabajo el mes pasado	83.69%	82.08%	83.63%	74.68%	84.12%	82.93%
No tiene prestaciones	79.94%	58.10%	47.79%	50.01%	48.55%	67.00%
Tiene derecho a servicios médicos del IMSS, ISSSTE etc.	18.20%	39.77%	50.99%	48.96%	49.68%	31.20%
Tiene derecho a otras prestaciones como aguinaldo, vacaciones etc.	19.03%	40.94%	52.06%	49.99%	51.45%	32.20%
Edad del jefe del Hogar*	47.46	47.00	46.60	46.45	46.62	47.14

*Promedio

Cuadro 4.7.1 : Comportamiento de las variables (Características de los miembros del hogar)

COMPORTAMIENTO DE LAS VARIABLES
CARACTERÍSTICAS DE LA VIVIENDA

	Grupo					Total
	1	2	3	4	5	
Vivienda prestada por algún pariente amigo u otra persona	12.96%	10.46%	10.98%	6.61%	4.50%	11.38%
Vivienda dada en su trabajo como prestación	0.15%	0.22%	0.00%	0.57%	0.75%	0.19%
Vivienda rentada o alquilada	8.91%	14.91%	18.54%	20.79%	28.85%	13.26%
Vivienda propia y aún la están pagando	1.77%	8.99%	11.08%	16.13%	11.33%	5.99%
Vivienda propia en terreno ejidal o comunal	15.92%	1.97%	0.97%	1.69%	0.08%	8.84%
Vivienda propia y totalmente pagada en terreno propio	58.35%	63.05%	58.07%	53.87%	54.49%	59.18%
Vivienda propia y construida en terreno prestado	1.68%	0.40%	0.37%	0.00%	0.00%	1.02%
¿Esta vivienda tiene cuarto para cocinar?	85.58%	92.57%	95.51%	95.52%	98.49%	89.82%
¿En el cuarto para cocinar también duermen?	3.80%	1.07%	0.23%	0.00%	0.22%	2.26%
Muros de cartón, hule, tela, llantas, lámina de cartón, carrizo, bambú, etc.	5.55%	0.00%	0.00%	0.00%	0.00%	2.83%
Muros de madera	7.63%	2.24%	1.17%	0.28%	0.43%	4.68%
Muros de asbesto	13.11%	4.87%	2.66%	1.52%	1.61%	8.47%
Muros de panel de concreto, concreto monolítico, tabique, ladrillo etc.	73.15%	92.89%	96.17%	98.21%	97.96%	83.74%
Techos de cartón, hule, tela, llantas, lámina de cartón, palma, etc.	8.39%	0.67%	0.00%	0.00%	0.00%	4.46%
Techos de Madera	1.44%	3.46%	4.31%	3.04%	1.83%	2.46%
Techos de lámina metálica, fibra de vidrio plástico o mica	19.98%	5.50%	2.84%	1.42%	1.22%	12.14%
Techos de terrado, lámina de asbesto o teja	20.50%	7.64%	3.23%	1.14%	0.00%	12.96%
Techos de panel de concreto, concreto monolítico, tabique, ladrillo, etc.	45.86%	79.09%	83.76%	88.74%	95.22%	63.95%
Techos de vigueta y poliuretano, vigueta y bovedilla, vigueta y cuña	3.76%	3.64%	5.86%	5.67%	1.73%	4.01%
Pisos de tierra	17.42%	0.81%	0.00%	0.00%	0.00%	9.10%
Pisos de cemento o firme	68.65%	54.12%	24.58%	13.38%	6.86%	53.49%
Pisos de mosaico o terrado	10.07%	26.56%	38.26%	45.31%	38.54%	21.11%
Pisos de loseta de vinil o plástico, linóleum o congóleum	0.61%	5.25%	10.11%	10.36%	5.04%	3.76%
Pisos de loseta de cemento (vitro piso),mármol	3.25%	13.26%	26.34%	26.41%	39.73%	11.79%
Pisos de madera, duela o parquet	0.00%	0.00%	0.00%	1.11%	1.64%	0.12%
Pisos de otros recubrimientos como alfombra etc.	0.00%	0.00%	0.71%	3.43%	8.20%	0.62%
Agua entubada dentro de la vivienda	38.85%	83.56%	96.00%	98.94%	100.00%	63.91%
Agua entubada fuera de la vivienda pero dentro del edificio o terreno	43.72%	14.49%	3.93%	1.06%	0.00%	26.68%
No dispone de agua entubada	17.43%	1.95%	0.07%	0.00%	0.00%	9.41%
¿Esta vivienda tiene cuarto de baño?	86.06%	99.16%	99.95%	100.00%	100.00%	92.67%
Vivienda con hoyo negro o pozo ciego	5.63%	0.04%	0.00%	0.00%	0.00%	2.88%
Vivienda con letrina	15.53%	2.73%	0.35%	0.00%	0.00%	8.68%
Vivienda con excusado o sanitario	67.32%	96.41%	99.61%	100.00%	100.00%	82.34%
Vivienda que no dispone del servicio sanitario	11.52%	0.81%	0.05%	0.00%	0.00%	6.09%
¿Esta vivienda tiene luz eléctrica?	96.12%	100.00%	100.00%	100.00%	100.00%	98.02%

Cuadro 4.7.2 : Comportamiento de las variables (Características de la vivienda)

COMPORTAMIENTO DE LAS VARIABLES
EQUIPAMIENTO DE LA VIVIENDA

	Grupo					Total
	1	2	3	4	5	
¿Esta vivienda cuenta con regadera?	36.22%	82.01%	94.91%	98.73%	100.00%	62.00%
¿Esta vivienda cuenta con closet?	10.86%	43.46%	71.92%	86.53%	91.88%	34.86%
¿Cuenta con teléfono para el uso del hogar?	19.32%	58.44%	74.08%	77.05%	83.79%	42.65%
¿Cuenta con teléfono celular para el uso del hogar?	8.28%	22.27%	34.11%	52.45%	56.38%	19.61%
¿Cuenta con televisión por cable para el uso del hogar?	3.04%	12.62%	29.68%	47.22%	62.32%	13.87%
¿Cuenta con Internet para el uso del hogar?	0.00%	3.33%	11.91%	32.08%	36.46%	5.54%
¿Cuenta con automóvil, camioneta o camioneta de caja	15.47%	42.64%	65.15%	83.59%	83.65%	35.53%
¿Cuenta con estéreo, modular o consola para el uso de este hogar?	31.99%	61.79%	73.69%	74.79%	74.65%	49.41%
¿Cuenta con reproductor de discos compactos para el uso de este hogar?	4.52%	15.99%	27.90%	38.50%	49.16%	14.28%
¿Cuenta con televisión a color para el uso de este hogar?	66.85%	94.07%	97.15%	99.82%	92.00%	80.75%
¿Cuenta con video casetera para el uso de este hogar?	18.02%	48.59%	64.81%	72.02%	84.46%	37.93%
¿Cuenta con reproductor de video discos (dvd) para el uso de este hogar?	0.46%	5.77%	10.44%	25.00%	34.07%	5.82%
¿Cuenta con computadora para el uso de este hogar?	0.07%	13.52%	35.41%	53.97%	57.62%	13.45%
¿Cuenta con impresora para el uso de este hogar?	0.07%	10.81%	29.27%	48.68%	47.42%	11.17%
¿Cuenta con (escáner, modem) y demás aparatos para la computadora ?	0.00%	4.31%	16.41%	24.60%	30.57%	5.87%
¿Cuenta con video juegos para el uso de este hogar?	2.27%	11.07%	15.69%	29.42%	23.11%	8.52%
¿Cuenta con licuadora para el uso de este hogar?	74.21%	94.17%	94.87%	93.74%	95.69%	84.15%
¿Cuenta con horno de microondas para el uso de este hogar?	10.39%	37.73%	58.75%	68.78%	81.18%	30.05%
¿Cuenta con refrigerador para el uso de este hogar?	61.04%	93.24%	96.89%	97.78%	98.73%	77.78%
¿Cuenta con estufa de gas para el uso de este hogar?	78.48%	98.07%	98.47%	96.59%	97.46%	88.05%
¿Cuenta con lavadora para el uso de este hogar?	38.09%	74.40%	81.46%	85.59%	81.56%	57.66%

Cuadro 4.7.3 : Comportamiento de las variables (Equipamiento de la vivienda)

Como se puede observar en los cuadros anteriores, dentro de las variables propuestas existen variables que tienen valores promedio (o porcentajes) claramente diferenciados entre los grupos de ingreso propuesto, principalmente entre el grupo de ingreso más elevado y el más bajo.

Por ejemplo, para las características de los miembros del hogar (cuadro 4.7.1), observamos que en 73.82% de los hogares del grupo 5 el jefe del hogar tiene un grado de escolaridad de Universidad, Maestría o Doctorado; mientras que esta misma característica se observa en tan sólo 1.52% de los hogares del Grupo 1.

Del mismo modo, en el caso de las características de la vivienda (cuadro 4.7.2), dentro de los grupos 3, 4 y 5 ninguno de los hogares observados cuenta con piso de tierra dentro de sus viviendas, mientras que para el grupo 1, el 17.42% de los hogares observados cuentan con

esta característica y para el grupo 2 este porcentaje es de 0.81%. Por el contrario, ninguno de los hogares observados en los grupos 1 y 2 cuenta con recubrimientos de piso tales como alfombras etc., mientas que para el grupo 5 este porcentaje es de 8.20%.

Sin embargo, también existen variables cuyos porcentajes no son demasiado distintos para todos los grupos, como es el caso por ejemplo de la variable (cocina) (¿Esta vivienda tiene cuarto para cocinar?).

En lo referente al equipamiento de las viviendas (cuadro 4.7.3), podemos ver, por ejemplo, que más de la mitad de los hogares observados en el grupo 5 (57.62%) cuentan con computadoras para su uso dentro del hogar, mientras que este porcentaje se reduce a sólo 0.07% para el grupo 1.

Además de permitir analizar el comportamiento de las variables dentro de los grupos, para ver si estas contribuirán a la diferenciación de estos, el realizar este análisis descriptivo aporta información sociodemográfica interesante, como por ejemplo el hecho de que en promedio el 98.02% de los hogares observados cuenta con luz eléctrica, el 9.41% no dispone de agua entubada, el 5.54% cuenta con Internet; o que la proporción de niños que trabajan, de acuerdo con la muestra observada, es de 0.02% etc.

Al realizar el análisis descriptivo de las variables, podemos observar que existen variables que resultan constantes dentro de los grupos y dado que la mayoría de las variables contempladas son originalmente categóricas, y por lo tanto recodificadas en variables dummy, dentro de un mismo grupo encontramos variables constantes idénticas, por lo que al obtener las matrices de varianza covarianza de cada grupo estas resultan singulares y de rangos diferentes.

Si bien el análisis discriminante puede llevarse a cabo omitiendo la singularidad de las matrices de varianza covarianza, puesto que las funciones discriminantes se obtendrán a partir de la matriz de varianza covarianza global cuya no singularidad es asegurada por la utilización del método Stepwise y el criterio de tolerancia; y a pesar de que probablemente la prueba de igualdad de matrices de varianza covarianza sea rechazada, las variables que resulten constantes dentro de un grupo determinado serán eliminadas, permitiendo de esta manera poder realizar la prueba M de Box y no alejarse tanto del supuesto de igualdad de matrices de varianza covarianza. La eliminación de estas variables se ve justificada también por el hecho de que el número de variables contempladas originalmente para el análisis es muy elevado, por lo que la eliminación de éstas no resulta demasiado importante. Para asegurar que las matrices de varianza covarianza de cada uno de los grupos sean no singulares, se eliminarán también las variables que presenten una correlación perfecta dentro de un grupo determinado.

Las 52 variables eliminadas en estos procesos se muestran el cuadro 4.8, las matrices de varianza covarianza de cada grupo, así como la matriz de varianza covarianza global se omiten debido a su gran tamaño.

GRUPO1	
Constante	Correlación Perfecta
cla_hog4	impresora
cla_hog5	combust1
viv_9	
pisos6	
pisos7	
combust4	
combust7	
s_calef	
apoyoviv	
internet	
antena	
apcompu	
calef	
aspira	

GRUPO2	
Constante	Correlación Perfecta
cla_hog4	combust1
cla_hog5	luz_ins
viv_5	luz_con
viv_9	
muros1	
muros2	
pisos6	
pisos7	
basura1	
basura3	
combust4	
combust7	
luz	
luz3	
luz4	
nintraba	

GRUPO3	
Constante	Correlación Perfecta
cla_hog5	bano2
viv_2	banoc
viv_5	luz_con
viv_9	bano4
muros1	cbano
muros2	luz_ins
techo1	
pisos1	
pisos6	
bano1	
basura1	
basura3	
combust1	
combust4	
luz	
luz3	
luz4	
nintraba	
analfabe	
jef_alfa	

GRUPO4	
Constante	Correlación Perfecta
cla_hog4	luz_con
viv_5	agua2
viv_8	agua1
cocinad	banoa
muros1	banob
muros2	otraspre
techo1	
pisos1	
agua3	
cbaño	
bano1	
bano2	
bano3	
bano4	
banoc	
basura1	
basura3	
combust1	
luz	
luz3	
luz4	
luz_ins	
nintraba	
analfabe	
jef_alfa	

GRUPO5	
Constante	Correlación Perfecta
cla_hog4	otraspre
tviv_5	
viv_5	
viv_8	
viv_9	
muros1	
muros2	
techo1	
techo4	
pisos1	
agua1	
agua2	
agua3	
cbaño	
bano1	
bano2	
bano3	
bano4	
banoa	
banob	
banoc	
drenaje	
basura1	
basura3	
combust1	
luz	
luz3	
luz4	
luz_ins	
luz_con	
regadera	
nintraba	
analfabe	
jef_alfa	

TODOS LOS GRUPOS
agua1
agua2
agua3
analfabe
antena
apcompu
apoyoviv
aspira
bano1
bano2
bano3
bano4
banoa
banob
banoc
basura1
basura3
calef
cbano
cla_hog4
cla_hog5
cocinad
combust1
combust4
combust7
drenaje
impresora
internet
jef_alfa
luz
luz_con
luz_ins
luz3
luz4
muros1
muros2
nintraba
otraspre
pisos1
pisos6
pisos7
regadera
s_calef
techo1
techo4
tviv_5
viv_2
viv_5
viv_8
viv_9

Cuadro 4.8: Variables eliminadas por ser constantes dentro de los grupos o por tener correlación perfecta.

Para determinar de manera más apropiada si existen diferencias estadísticamente significativas entre los valores de las medias, se realizará una prueba basada en el estadístico lambda de Wilks. Los resultados de esta prueba se presentan a continuación:

Pruebas de Igualdad de Medias en los Grupos
Características de los Miembros del Hogar

	Wilks' Lambda	F	df1	df2	Sig.
Estrato Urbano/Rural	0.829	1002941.789	4	19469027	0
Índice de hacinamiento	0.775	1410777.849	4	19469027	0
Indice de dependencia economica	0.938	319787.204	4	19469027	0
Número de personas en el hogar	0.89	603771.876	4	19469027	0
Niños menores de 12 años	0.877	683819.283	4	19469027	0
Niños entre 12 y 15 años que trabajan	0.976	121986.448	4	19469027	0
Personas de 15 años y más analfabetas	0.874	701602.975	4	19469027	0
Al menos uno tiene servivio medico en el hogar	0.869	734618.12	4	19469027	0
Sexo del jefe del hogar	0.998	7689.393	4	19469027	0
Edad del Jefe del Hogar	0.999	2753.887	4	19469027	0
Alfabetismo del jefe del hogar	0.916	444444.205	4	19469027	0
Escolaridad del Jefe del hogar (NInguna, Preprimaria, Primaria)	0.794	1265588.516	4	19469027	0
Escolaridad del Jefe del hogar (Secundaria, Prepartoria, Normal)	0.965	176927.131	4	19469027	0
Escolaridad del Jefe del hogar (Universidad, Maestria, Doctorado)	0.684	2247413.639	4	19469027	0
Estado civil del jefe del hogar (Union libre)	0.986	70436.034	4	19469027	0
Estado civil del jefe de familia (Casado(a))	0.997	14849.802	4	19469027	0
Estado civil del jefe de familia (separado(a))	0.999	5832.581	4	19469027	0
Estado civil del jefe de familia (divorciado(a))	0.983	85060.812	4	19469027	0
Estado civil del jefe de familia (viudo(a))	0.999	4821.878	4	19469027	0
Estado civil del jefe de familia (soletero(a))	0.978	110004.947	4	19469027	0
El jefe de familia trabajo el mes pasado (por salario, por su ccuenta, sin remuneración)	0.998	11117.783	4	19469027	0
No tiene prestaciones	0.916	447602.277	4	19469027	0
Tiene derecho a servicios médicos del IMSS, ISSSTE, PEMEX etc.	0.912	472033.824	4	19469027	0
Tiene derecho a otras prestaciones como aguinaldo, vacaciones etc.	0.911	476561.858	4	19469027	0

Cuadro 4.9.1 : Pruebas de igualdad de medias (Características de los miembros del hogar)

Pruebas de Igualdad de Medias en los Grupos
Características de la Vivienda

	Wilks' Lambda	F	df1	df2	Sig.
Vivienda prestada por algún pariente amigo u otra persona	0.995	22509.232	4	19469027	0
Vivienda dada en su trabajo como prestación	0.999	6750.551	4	19469027	0
Vivienda rentada o alquilada	0.976	121779.567	4	19469027	0
Vivienda propia y aún la estan pagando	0.964	183157.485	4	19469027	0
Vivienda propia en terreno ejidal o comunal	0.935	338229.1	4	19469027	0
Vivienda propia y totalmente pagada en terreno propio	0.997	13257.609	4	19469027	0
Vivienda propia y construida en terreno prestado	0.995	23172.206	4	19469027	0
¿Esta vivienda tiene cuarto para cocinar?	0.977	112139.639	4	19469027	0
¿En el cuarto para cocinar también duermen?	0.988	57064.713	4	19469027	0
Muros de cartón, hule, tela, llantas, lámina de cartón, carrizo, bambú etc	0.972	140146.358	4	19469027	0
Muros de Lámina de asbesto, lámina métalica, fibra de vidrio plástico etc.	0.997	13463.705	4	19469027	0
Muros de madera	0.979	104416.251	4	19469027	0
Muros de asbesto	0.97	150648.748	4	19469027	0
Muros de panel de concreto, concreto monolítico, tabique, ladrillo etc.	0.913	465208.069	4	19469027	0
Techos de cartón, hule, tela, llantas, lámina de cartón, palma, etc.	0.962	192439.506	4	19469027	0
Techos de Madera	0.995	26679.672	4	19469027	0
Techos de lámina metálica, fibra de vidrio plástico o mica	0.939	317495.489	4	19469027	0
Techos de terrado, lámina de asbesto o teja	0.944	289407.123	4	19469027	0
Techos de panel de concreto, concreto monolítico, tabique, ladrillo,etc.	0.847	879762.422	4	19469027	0
Techos de vigueta y poliuretano, vigueta y bovedilla, vigueta y cuña	0.998	11580.449	4	19469027	0
Pisos de tierra	0.913	465858.576	4	19469027	0
Pisos de cemento o frime	0.838	938297.772	4	19469027	0
Pisos de mosaico o terrado	0.911	477719.61	4	19469027	0
Pisos de loseta de vinil o plástico, linóleum o congóleum	0.964	183527.395	4	19469027	0
Pisos de loseta de cemento (vitro piso),mármol	0.891	595862.074	4	19469027	0
Pisos de madera, duela o parquet	0.987	65855.921	4	19469027	0
Pisos de otros recubrimientos como alfombra etc.	0.946	276773.562	4	19469027	0
Agua entubada dentro de la vivienda	0.706	2024057.089	4	19469027	0
Agua entubada fuera de la vivienda pero dentro del edificio o terreno	0.836	951838.15	4	19469027	0
No dispone de agua entubada	0.921	418016.064	4	19469027	0
¿Esta vivienda tiene cuarto de baño?	0.933	348921.691	4	19469027	0
Vivienda con hoyo negro o pozo ciego	0.972	140547.666	4	19469027	0
Vivienda con letrina	0.937	324598.828	4	19469027	0
Vivienda con excusado o sanitario	0.838	943832.228	4	19469027	0
Vivienda que no dispone del servicio sanitario	0.946	276062.316	4	19469027	0
¿Esta vivienda tiene luz electrica?	0.981	96301.953	4	19469027	0

Cuadro 4.9.2 : Pruebas de igualdad de medias (Características de la vivienda)

Pruebas de Igualdad de Medias en los Grupos
Equipamiento de la Vivienda

	Wilks' Lambda	F	df1	df2	Sig.
¿Esta vivienda cuenta con regadera?	0.695	2136971.378	4	19469027	0
¿Esta vivienda cuenta con closet?	0.662	2480513.442	4	19469027	0
¿Cuenta con telefono para el uso del hogar?	0.751	1616711.873	4	19469027	0
¿Cuenta con telefono celular para el uso del hogar?	0.871	719339.317	4	19469027	0
¿Cuenta con televisión por cable, Sky, Direct-tv o multivisión para el uso del hogar?	0.791	1289665.907	4	19469027	0
¿Cuenta con Internet para el uso del hogar?	0.818	1082404.319	4	19469027	0
¿Cuenta con automovil, camioneta o camioneta de caja	0.763	1510891.18	4	19469027	0
¿Cuenta con estéreo, modular o consola para el uso de este hogar?	0.866	751119.924	4	19469027	0
¿Cuenta con reproductor de discos compactos para el uso de este hogar?	0.872	712324.394	4	19469027	0
¿Cuenta con televisión a color para el uso de este hogar?	0.869	732699.819	4	19469027	0
¿Cuenta con videocasetera para el uso de este hogar?	0.795	1255371.848	4	19469027	0
¿Cuenta con reproductor de video discos (dvd) para el uso de este hogar?	0.873	710291.344	4	19469027	0
¿Cuenta con computadora para el uso de este hogar?	0.728	1814522.255	4	19469027	0
¿Cuenta con impresora para el uso de este hogar?	0.772	1436727.273	4	19469027	0
¿Cuenta con escaner, quemador de cd, modem, lector de cd, y demás aparátos inegrados a la computadora para el uso de este hogar?	0.861	782859.198	4	19469027	0
¿Cuenta con video juegos (Playstation, nintendom sega u otros) para el uso de este hogar?	0.928	375552.979	4	19469027	0
¿Cuenta con licuadora para el uso de este hogar?	0.923	406886.599	4	19469027	0
¿Cuenta con horno de microondas para el uso de este hogar?	0.756	1569091.106	4	19469027	0
¿Cuenta con refrigerador para el uso de este hogar?	0.83	997081.931	4	19469027	0
¿Cuenta con estufa de gas para el uso de este hogar?	0.909	485697.075	4	19469027	0
¿Cuenta con lavadora para el uso de este hogar?	0.833	972933.787	4	19469027	0

Cuadro 4.9.3 : Pruebas de igualdad de medias (Equipamiento de la vivienda).

Como puede observarse en los cuadros anteriores, ninguna de las variables presenta valores del estadístico lambda Wilks menores a 0.5 (recordemos que el estadístico lambda toma valores entre 0 y 1, donde valores cercanos a cero indican grandes diferencias entre los grupos y valores cercanos a 1 indican que no hay gran diferenciación entre los grupos), lo que indica que las diferencias entre las medias no son demasiado importantes. Los valores más bajos del estadístico lambda de Wilks en lo referente a las características de los miembros del hogar (cuadro 4.9.1) se presentan para las variables edu_2 (Escolaridad del jefe de familia (Universidad, Maestría, Doctorado)) y hacina (Índice de hacinamiento).

En cuanto a las características de la vivienda (cuadro 4.9.2), los valores más bajos obtenidos para el estadístico lambda corresponden a las variables agua1 y agua2 (agua entubada dentro de la vivienda y agua entubada fuera de la vivienda pero dentro del edificio o terreno).

Los valores más bajos del estadístico lambda para las variables del equipamiento de la vivienda (cuadro 4.9.3) se presentan en las variables closet y regadera (¿Esta vivienda cuenta con closet?, ¿Esta vivienda cuenta con regadera?).

Sin embargo, para todas las variables, estas diferencias resultan significativas al utilizar la transformación del estadístico lambda en un estadístico F, por lo que la prueba de igualdad de medias puede ser rechazada en todos los casos y todas las variables pueden ser consideradas como candidatas a ser incluidas en el análisis.

4.5 Selección de variables

La selección de variables que se incluyen en el modelo se realiza, como ya se había mencionado, a través del método Stepwise (método de inclusión por pasos), el criterio empleado para la selección de las variables es el criterio basado en el estadístico lambda Wilks descrito en el capítulo anterior, como condiciones mínimas para la inclusión de variables se fija un nivel de tolerancia de 0.001 y los estadísticos F de entrada y F de Salida se fijan en 3.84 y 2.71 respectivamente.

El proceso de selección de variables consta de 90 pasos en el caso de esta aplicación, al finalizar este proceso, únicamente 3 de las 93 variables contempladas quedan fuera del modelo al no cumplir con las condiciones de entrada, las demás variables no cumplen con las condiciones de salida por lo que se considerarán en el modelo. Los cuadros 4.10 a 4.13 corresponden a las salidas que ofrece el paquete SPSS en cuanto a la selección de variables y en ellos se muestra un resumen de los resultados obtenidos por medio de este proceso.

Variables introducidas/eliminadas(a,b,c,d)													
Paso	Introducidas	Lambda de Wilks											
		Estadístico	gl1	gl2	gl3	F Exacta				F aproximada			
						Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
1	closet	0.662	1	4	19469027	2480513.442	4	19469027	0				
2	edu_2	0.524	2	4	19469027	1857192.274	8	38938052	0				
3	fregader	0.461	3	4	19469027					1460147.331	12	51510198.71	0
4	tam_hog	0.415	4	4	19469027					1238618.045	16	59478851.43	0
5	compu	0.382	5	4	19469027					1086220.943	20	64571445.28	0
6	smedich	0.353	6	4	19469027					984416.899	24	67919304.34	0
7	tostador	0.332	7	4	19469027					897485.734	28	70196554.92	0
8	tvcable	0.316	8	4	19469027					824019.973	32	71798199.77	0
9	automov	0.303	9	4	19469027					761029.672	36	72959428.33	0
10	depend2	0.289	10	4	19469027					713958.017	40	73824244.58	0
11	urbano	0.279	11	4	19469027					671267.615	44	74483595.2	0
12	dvd	0.272	12	4	19469027					628380.826	48	74996687.7	0
13	cla_hog1	0.266	13	4	19469027					591334.855	52	75403172.97	0
14	telefono	0.261	14	4	19469027					558542.128	56	75730299.41	0
15	celular	0.255	15	4	19469027					530913.474	60	75997234.34	0
16	edu_1	0.251	16	4	19469027					504221.734	64	76217747.13	0
17	viv_3	0.248	17	4	19469027					479489.394	68	76401918.5	0
18	pisos5	0.245	18	4	19469027					457253.326	72	76557255.83	0
19	lavabo	0.242	19	4	19469027					436970.647	76	76689437.21	0
20	repro_cd	0.24	20	4	19469027					418354.57	80	76802817.77	0
21	sandwich	0.238	21	4	19469027					401443.792	84	76900780.83	0
22	pisos4	0.236	22	4	19469027					385802.336	88	76985985.54	0
23	tv_color	0.234	23	4	19469027					371446.369	92	77060544.32	0
24	pisos3	0.232	24	4	19469027					358134.81	96	77126151.75	0
25	edo_civ4	0.231	25	4	19469027					345685.089	100	77184179.58	0
26	cafetera	0.229	26	4	19469027					333919.955	104	77235747.5	0
27	videoj	0.228	27	4	19469027					322902.261	108	77281776.58	0
28	bomba	0.227	28	4	19469027					312625.119	112	77323029.91	0
29	pisos2	0.226	29	4	19469027					303056.845	116	77360143.98	0
30	video	0.225	30	4	19469027					294055.847	120	77393652.95	0
31	aire_ac	0.224	31	4	19469027					285603.009	124	77424007.78	0
32	jno_trab	0.223	32	4	19469027					277622.465	128	77451591.2	0
33	extracto	0.222	33	4	19469027					270124.43	132	77476729.77	0
34	techo5	0.221	34	4	19469027					263052.24	136	77499703.42	0
35	edo_civ6	0.22	35	4	19469027					256380.485	140	77520753.24	0
36	menores	0.219	36	4	19469027					250014.237	144	77540087.72	0
37	tviv_3	0.218	37	4	19469027					243987.208	148	77557887.9	0
38	edo_civ3	0.217	38	4	19469027					238273.923	152	77574311.62	0
39	lavadora	0.216	39	4	19469027					232808.674	156	77589496.95	0
40	horno_el	0.216	40	4	19469027					227617.225	160	77603565.1	0
41	viv_4	0.215	41	4	19469027					222650.775	164	77616622.83	0
42	techo2	0.214	42	4	19469027					217887.196	168	77628764.48	0
43	basura2	0.214	43	4	19469027					213290.305	172	77640073.62	0
44	s_aireac	0.213	44	4	19469027					208872.798	176	77650624.54	0
45	batidor	0.213	45	4	19469027					204624.746	180	77660483.41	0
46	estereo	0.212	46	4	19469027					200498.747	184	77669709.38	0
47	radio	0.212	47	4	19469027					196577.988	188	77678355.38	0
48	techosr	0.211	48	4	19469027					192790.986	192	77686468.97	0
49	sexo_jef	0.211	49	4	19469027					189141.956	196	77694092.89	0
50	edo_civ5	0.21	50	4	19469027					185613.798	200	77701265.72	0

Cuadro 4.10 : Variables que entran o salen en cada paso de la selección.

Variables introducidas/eliminadas(a,b,c,d)

Paso	Introducidas	Lambda de Wilks											
		Estadístico	gl1	gl2	gl3	F Exacta				F aproximada			
						Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
51	basura7	0.21	51	4	19469027					182213.363	204	77708022.31	0
52	refri	0.21	52	4	19469027					178926.369	208	77714394.19	0
53	tv_bn	0.209	53	4	19469027					175774.607	212	77720409.98	0
54	centro	0.209	54	4	19469027					172730.74	216	77726095.7	0
55	edad_jef	0.209	55	4	19469027					169784.405	220	77731475.01	0
56	viv_7	0.208	56	4	19469027					166944.78	224	77736569.52	0
57	moto	0.208	57	4	19469027					164189.604	228	77741398.94	0
58	servmed	0.208	58	4	19469027					161525.37	232	77745981.33	0
59	lavadero	0.208	59	4	19469027					158950.518	236	77750333.22	0
60	cla_hog3	0.207	60	4	19469027					156455.43	240	77754469.8	0
61	cla_hog2	0.206	61	4	19469027					154780.891	244	77758405.01	0
62	hacina	0.205	62	4	19469027					152422.676	248	77762151.7	0
63	sur	0.205	63	4	19469027					150134.403	252	77765721.69	0
64	basura6	0.205	64	4	19469027					147900.655	256	77769125.91	0
65	basura5	0.205	65	4	19469027					145762.171	260	77772374.44	0
66	techo6	0.205	66	4	19469027					143661.832	264	77775476.63	0
67	bici	0.204	67	4	19469027					141623.998	268	77778441.11	0
68	jugos	0.204	68	4	19469027					139643.067	272	77781275.9	0
69	graba	0.204	69	4	19469027					137717.292	276	77783988.45	0
70	edo_civ2	0.204	70	4	19469027					135846.132	280	77786585.66	0
71	viv_1	0.204	71	4	19469027					134029.941	284	77789073.98	0
72	calentad	0.203	72	4	19469027					132259.385	288	77791459.39	0
73	cisterna	0.203	73	4	19469027					130534.338	292	77793747.48	0
74	cocina	0.203	74	4	19469027					128850.415	296	77795943.46	0
75	muros5	0.203	75	4	19469027					127206.713	300	77798052.21	0
76	tviv_4	0.203	76	4	19469027					125603.293	304	77800078.26	0
77	combust5	0.203	77	4	19469027					124042.001	308	77802025.88	0
78	muros3	0.202	78	4	19469027					122511.592	312	77803899.05	0
79	murosr	0.202	79	4	19469027					121020.561	316	77805701.51	0
80	tviv1	0.202	80	4	19469027					119571.35	320	77807436.78	0
81	micro	0.202	81	4	19469027					118149.331	324	77809108.13	0
82	molino	0.202	82	4	19469027					116760.6	328	77810718.66	0
83	prestac0	0.202	83	4	19469027					115403.974	332	77812271.27	0
84	estufa_g	0.202	84	4	19469027					114070.378	336	77813768.72	0
85	maq_cos	0.202	85	4	19469027					112760.374	340	77815213.56	0
86	licuador	0.202	86	4	19469027					111480.014	344	77816608.23	0
87	tinaco	0.202	87	4	19469027					110223.493	348	77817955.02	0
88	muros4	0.201	88	4	19469027					108990.103	352	77819256.08	0
89	viv_6	0.201	89	4	19469027					107780.311	356	77820513.47	0
90	techo3	0.201	90	4	19469027					106591.138	360	77821729.09	0

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

a. El número máximo de pasos es 186.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para eliminar es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

Cuadro 4.10 : Variables que entran o salen en cada paso de la selección (Continuación).

Variables en el análisis

Paso		Tolerancia	F para eliminar	Lambda de Wilks
1	closet	1	2480513.442	
2	closet	0.983	1488373.65	0.684
	edu_2	0.983	1286748.382	0.662
3	closet	0.854	543100.924	0.512
	edu_2	0.98	1178523.464	0.572
	fregader	0.86	665800.891	0.524
⋮	⋮	⋮	⋮	⋮
90	closet	0.724	48126.347	0.203
	edu_2	0.653	343227.192	0.216
	fregader	0.545	15831.882	0.202
	tam_hog	0.335	83674.554	0.205
	compu	0.774	64604.601	0.204
	smedich	0.382	79825.546	0.205
	tostador	0.717	50513.448	0.204
	tvcable	0.862	137332.769	0.207
	automov	0.728	134983.758	0.207
	depend2	0.633	240130.808	0.211
	urbano	0.566	18786.859	0.202
	dvd	0.842	80346.529	0.205
	cla_hog1	0.027	22056.351	0.202
	telefono	0.638	53367.998	0.204
	celular	0.86	65829.493	0.204
	edu_1	0.619	15582.33	0.202
	viv_3	0.104	5870.431	0.202
	pisos5	0.324	28348.479	0.203
	lavabo	0.46	9917.113	0.202
	repro_cd	0.882	25033.309	0.202
	sandwich	0.778	26120.889	0.203
	pisos4	0.539	51119.742	0.204
	tv_color	0.51	10562.286	0.202
	pisos3	0.228	21359.024	0.202
	edo_civ4	0.676	34679.404	0.203
	cafetera	0.757	24472.285	0.202
	videoj	0.841	19250.058	0.202
	bomba	0.452	13705.271	0.202
	pisos2	0.237	12920.838	0.202
	video	0.737	12683.395	0.202
	aire_ac	0.776	27159.952	0.203
	jno_trab	0.668	15848.976	0.202

Cuadro 4.11 : Variables en el análisis en cada paso

Variables en el análisis

Paso		Tolerancia	F para eliminar	Lambda de Wilks
90	extracto	0.745	15427.222	0.202
	techo5	0.392	3335.03	0.202
	edo_civ6	0.478	21158.316	0.202
	menores	0.427	21420.392	0.202
	tviv_3	0.448	15702.574	0.202
	edo_civ3	0.422	19708.624	0.202
	lavadora	0.659	10868.198	0.202
	horno_el	0.83	21028.053	0.202
	viv_4	0.189	4858.428	0.202
	techo2	0.709	6462.208	0.202
	basura2	0.083	1965.22	0.202
	s_aireac	0.867	9867.831	0.202
	batidor	0.709	10129.336	0.202
	estereo	0.681	10298.591	0.202
	radio	0.921	11427.829	0.202
	techosr	0.707	9830.308	0.202
	sexo_jef	0.357	14490.832	0.202
	edo_civ5	0.317	7840.347	0.202
	basura7	0.375	12985.984	0.202
	refri	0.591	5175.891	0.202
	tv_bn	0.741	9600.823	0.202
	centro	0.386	3247.181	0.202
	edad_jef	0.44	6611.279	0.202
	viv_7	0.053	1555.625	0.202
	moto	0.978	6647.068	0.202
	servmed	0.075	6124.647	0.202
	lavadero	0.853	6654.139	0.202
	cla_hog3	0.01	39794.085	0.203
	cla_hog2	0.008	37350.601	0.203
	hacina	0.531	6951.382	0.202
	sur	0.439	4371.853	0.202
	basura6	0.061	6477.872	0.202
	basura5	0.227	5673.248	0.202
	techo6	0.701	4078.877	0.202
	bici	0.932	5107.689	0.202
	jugos	0.775	4713.863	0.202
	graba	0.864	4918.842	0.202
	edo_civ2	0.381	4305.829	0.202
	viv_1	0.117	2581.588	0.202
	calentad	0.526	4966.2	0.202
	cisterna	0.479	3841.139	0.202
	cocina	0.841	3915.807	0.202
	muros5	0.171	3398.945	0.202
	tviv_4	0.199	3468.546	0.202
	combust5	0.342	2360.527	0.202
	muros3	0.406	1840.128	0.202
	murosr	0.742	3645.244	0.202
	tviv1	0.167	3422.322	0.202
	micro	0.711	3165.988	0.202
	molino	0.772	2812.055	0.202
	prestac0	0.084	2783.8	0.202
	estufa_g	0.42	1884.573	0.202
	mag_cos	0.833	1767.464	0.202
	licuador	0.639	1745.731	0.202
	tinaco	0.513	1470.227	0.201
	muros4	0.245	1368.002	0.201
	viv_6	0.15	912.935	0.201
	techo3	0.65	527.336	0.201

Cuadro 4.11 : Variables en el análisis en cada paso (Continuación)

Variables no incluidas en el análisis

Paso		Tolerancia	Tolerancia min.	F para introducir	Lamdda de Wilks
0	tam_hog	1	1	603771.876	0.89
	centro	1	1	73335.875	0.985
	sur	1	1	370285.461	0.929
	norte	1	1	93159.856	0.981
	urbano	1	1	1002941.789	0.829
	cla_hog1	1	1	164711.086	0.967
	cla_hog2	1	1	5469.594	0.999
	cla_hog3	1	1	49274.226	0.99
	tviv1	1	1	166026.665	0.967
	tviv_3	1	1	63241.817	0.987
	tviv_4	1	1	199106.869	0.961
	viv_1	1	1	22509.232	0.995
	viv_3	1	1	121779.567	0.976
	viv_4	1	1	183157.485	0.964
	viv_6	1	1	338229.1	0.935
	viv_7	1	1	13257.609	0.997
	hacina	1	1	1410777.849	0.775
	depend2	1	1	319787.204	0.938
	cocina	1	1	112139.639	0.977
	muros3	1	1	104416.251	0.979
	muros4	1	1	150648.748	0.97
	muros5	1	1	465208.069	0.913
	murosr	1	1	667179.2	0.879
	techo2	1	1	26679.672	0.995
	techo3	1	1	317495.489	0.939
	techo5	1	1	879762.422	0.847
	techo6	1	1	11580.449	0.998
	techosr	1	1	1249537.313	0.796
	pisos2	1	1	938297.772	0.838
	pisos3	1	1	477719.61	0.911
	pisos4	1	1	183527.395	0.964
	pisos5	1	1	595862.074	0.891
	basura2	1	1	1081875.779	0.818
	basura5	1	1	26243.508	0.995
	basura6	1	1	765804.422	0.864
	basura7	1	1	29046.011	0.994
	combust5	1	1	817145.513	0.856
	tinaco	1	1	924826.184	0.84
	cisterna	1	1	631487.631	0.885
	bomba	1	1	785097.454	0.861
	calentad	1	1	2062918.884	0.702
	s_aireac	1	1	270822.129	0.947
	lavabo	1	1	2182690.131	0.69
	lavadero	1	1	75256.06	0.985
	fregader	1	1	2354426.724	0.674
	closet	1	1	2480513.442	0.662
	telefono	1	1	1616711.873	0.751
	celular	1	1	719339.317	0.871

Cuadro 4.12 : Variables que no están en el análisis en cada paso.

Variables no incluidas en el análisis

Paso		Tolerancia	Tolerancia min.	F para introducir	Lamdda de Wilks
0	tvcable	1	1	1289665.907	0.791
	automov	1	1	1510891.18	0.763
	moto	1	1	20934.461	0.996
	bici	1	1	140616.647	0.972
	radio	1	1	122432.345	0.975
	graba	1	1	31793.355	0.994
	estereo	1	1	751119.924	0.866
	repro_cd	1	1	712324.394	0.872
	tv_bn	1	1	113753.084	0.977
	tv_color	1	1	732699.819	0.869
	video	1	1	1255371.848	0.795
	dvd	1	1	710291.344	0.873
	compu	1	1	1814522.255	0.728
	videoj	1	1	375552.979	0.928
	licuador	1	1	406886.599	0.923
	batidor	1	1	1243497.296	0.797
	extracto	1	1	1229693.807	0.798
	tostador	1	1	1648777.581	0.747
	cafetera	1	1	1420386.054	0.774
	sandwich	1	1	1094299.187	0.816
	jugos	1	1	794556.65	0.86
	horno_el	1	1	849969.994	0.851
	micro	1	1	1569091.106	0.756
	refri	1	1	997081.931	0.83
	estufa_g	1	1	485697.075	0.909
	molino	1	1	267974.214	0.948
	lavadora	1	1	972933.787	0.833
	mag_cos	1	1	126426.239	0.975
	aire_ac	1	1	341671.556	0.934
	menores	1	1	683819.283	0.877
	smedich	1	1	734618.12	0.869
	sexo_jef	1	1	7689.393	0.998
	edad_jef	1	1	2753.887	0.999
	edu_0	1	1	1265588.516	0.794
	edu_1	1	1	176927.131	0.965
	edu_2	1	1	2247413.639	0.684
	edo_civ1	1	1	70436.034	0.986
	edo_civ2	1	1	14849.802	0.997
	edo_civ3	1	1	5832.581	0.999
	edo_civ4	1	1	85060.812	0.983
	edo_civ5	1	1	4821.878	0.999
	edo_civ6	1	1	110004.947	0.978
	ino_trab	1	1	11117.783	0.998
	prestac0	1	1	447602.277	0.916
	servmed	1	1	472033.824	0.912
⋮	⋮	⋮	⋮	⋮	⋮
88	norte	0	0	.	.
	viv_6	0.15	0.008	905.888	0.201
	techo3	0.65	0.008	520.29	0.201
	edu_0	0	0	76703.718	0.251
	edo_civ1	0	0	.	.
89	norte	0	0	.	.
	techo3	0.65	0.008	527.336	0.201
	edu_0	0	0	76703.718	0.251
	edo_civ1	0	0	.	.
90	norte	0	0	.	.
	edu_0	0	0	76703.718	0.251
	edo_civ1	0	0	.	.

Cuadro 4.12 : Variables que no están en el análisis en cada paso (Continuación).

Lambda de Wilks

Paso	Número de variables	Lambda	gl1	gl2	gl3	F exacta				F aproximada			
						Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
1	1	0.662	1	4	19469027	2480513.442	4	19469027	0				
2	2	0.524	2	4	19469027	1857192.274	8	38938052	0				
3	3	0.461	3	4	19469027					1460147.331	12	51510198.71	0
4	4	0.415	4	4	19469027					1238618.045	16	59478851.43	0
5	5	0.382	5	4	19469027					1086220.943	20	64571445.28	0
6	6	0.353	6	4	19469027					984416.899	24	67919304.34	0
7	7	0.332	7	4	19469027					897485.734	28	70196554.92	0
8	8	0.316	8	4	19469027					824019.973	32	71798199.77	0
9	9	0.303	9	4	19469027					761029.672	36	72959428.33	0
10	10	0.289	10	4	19469027					713958.017	40	73824244.58	0
11	11	0.279	11	4	19469027					671267.615	44	74483595.2	0
12	12	0.272	12	4	19469027					628380.826	48	74996687.7	0
13	13	0.266	13	4	19469027					591334.855	52	75403172.97	0
14	14	0.261	14	4	19469027					558542.128	56	75730299.41	0
15	15	0.255	15	4	19469027					530913.474	60	75997234.34	0
16	16	0.251	16	4	19469027					504221.734	64	76217747.13	0
17	17	0.248	17	4	19469027					479489.394	68	76401918.5	0
18	18	0.245	18	4	19469027					457253.326	72	76557255.83	0
19	19	0.242	19	4	19469027					436970.647	76	76689437.21	0
20	20	0.24	20	4	19469027					418354.57	80	76802817.77	0
21	21	0.238	21	4	19469027					401443.792	84	76900780.83	0
22	22	0.236	22	4	19469027					385802.336	88	76985985.54	0
23	23	0.234	23	4	19469027					371446.369	92	77060544.32	0
24	24	0.232	24	4	19469027					358134.81	96	77126151.75	0
25	25	0.231	25	4	19469027					345685.089	100	77184179.58	0
26	26	0.229	26	4	19469027					333919.955	104	77235747.5	0
27	27	0.228	27	4	19469027					322902.261	108	77281776.58	0
28	28	0.227	28	4	19469027					312625.119	112	77323029.91	0
29	29	0.226	29	4	19469027					303056.845	116	77360143.98	0
30	30	0.225	30	4	19469027					294055.847	120	77393652.95	0
31	31	0.224	31	4	19469027					285603.009	124	77424007.78	0
32	32	0.223	32	4	19469027					277622.465	128	77451591.2	0
33	33	0.222	33	4	19469027					270124.43	132	77476729.77	0
34	34	0.221	34	4	19469027					263052.24	136	77499703.42	0
35	35	0.22	35	4	19469027					256380.485	140	77520753.24	0
36	36	0.219	36	4	19469027					250014.237	144	77540087.72	0
37	37	0.218	37	4	19469027					243987.208	148	77557887.9	0
38	38	0.217	38	4	19469027					238273.923	152	77574311.62	0
39	39	0.216	39	4	19469027					232808.674	156	77589496.95	0
40	40	0.216	40	4	19469027					227617.225	160	77603565.1	0
41	41	0.215	41	4	19469027					222650.775	164	77616622.83	0
42	42	0.214	42	4	19469027					217887.196	168	77628764.48	0
43	43	0.214	43	4	19469027					213290.305	172	77640073.62	0
44	44	0.213	44	4	19469027					208872.798	176	77650624.54	0
45	45	0.213	45	4	19469027					204624.746	180	77660483.41	0
46	46	0.212	46	4	19469027					200498.747	184	77669709.38	0
47	47	0.212	47	4	19469027					196577.988	188	77678355.38	0
48	48	0.211	48	4	19469027					192790.986	192	77686468.97	0
49	49	0.211	49	4	19469027					189141.956	196	77694092.89	0
50	50	0.21	50	4	19469027					185613.798	200	77701265.72	0

Cuadro 4.13: Lambda Wilks en cada paso

Lambda de Wilks

Paso	Número de variables	Lambda	gl1	gl2	gl3	F exacta				F aproximada			
						Estadístico	gl1	gl2	Sig.	Estadístico	gl1	gl2	Sig.
51	51	0.21	51	4	19469027					182213.363	204	77708022.31	0
52	52	0.21	52	4	19469027					178926.369	208	77714394.19	0
53	53	0.209	53	4	19469027					175774.607	212	77720409.98	0
54	54	0.209	54	4	19469027					172730.74	216	77726095.7	0
55	55	0.209	55	4	19469027					169784.405	220	77731475.01	0
56	56	0.208	56	4	19469027					166944.78	224	77736569.52	0
57	57	0.208	57	4	19469027					164189.604	228	77741398.94	0
58	58	0.208	58	4	19469027					161525.37	232	77745981.33	0
59	59	0.208	59	4	19469027					158950.518	236	77750333.22	0
60	60	0.207	60	4	19469027					156455.43	240	77754469.8	0
61	61	0.206	61	4	19469027					154780.891	244	77758405.01	0
62	62	0.205	62	4	19469027					152422.676	248	77762151.7	0
63	63	0.205	63	4	19469027					150134.403	252	77765721.69	0
64	64	0.205	64	4	19469027					147900.655	256	77769125.91	0
65	65	0.205	65	4	19469027					145762.171	260	77772374.44	0
66	66	0.205	66	4	19469027					143661.832	264	77775476.63	0
67	67	0.204	67	4	19469027					141623.998	268	77778441.11	0
68	68	0.204	68	4	19469027					139643.067	272	77781275.9	0
69	69	0.204	69	4	19469027					137717.292	276	77783988.45	0
70	70	0.204	70	4	19469027					135846.132	280	77786585.66	0
71	71	0.204	71	4	19469027					134029.941	284	77789073.98	0
72	72	0.203	72	4	19469027					132259.385	288	77791459.39	0
73	73	0.203	73	4	19469027					130534.338	292	77793747.48	0
74	74	0.203	74	4	19469027					128850.415	296	77795943.46	0
75	75	0.203	75	4	19469027					127206.713	300	77798052.21	0
76	76	0.203	76	4	19469027					125603.293	304	77800078.26	0
77	77	0.203	77	4	19469027					124042.001	308	77802025.88	0
78	78	0.202	78	4	19469027					122511.592	312	77803899.05	0
79	79	0.202	79	4	19469027					121020.561	316	77805701.51	0
80	80	0.202	80	4	19469027					119571.35	320	77807436.78	0
81	81	0.202	81	4	19469027					118149.331	324	77809108.13	0
82	82	0.202	82	4	19469027					116760.6	328	77810718.66	0
83	83	0.202	83	4	19469027					115403.974	332	77812271.27	0
84	84	0.202	84	4	19469027					114070.378	336	77813768.72	0
85	85	0.202	85	4	19469027					112760.374	340	77815213.56	0
86	86	0.202	86	4	19469027					111480.014	344	77816608.23	0
87	87	0.202	87	4	19469027					110223.493	348	77817955.02	0
88	88	0.201	88	4	19469027					108990.103	352	77819256.08	0
89	89	0.201	89	4	19469027					107780.311	356	77820513.47	0
90	90	0.201	90	4	19469027					106591.138	360	77821729.09	0

Cuadro 4.13: Lambda Wilks en cada paso (Continuación)

En el cuadro 4.10 (variables introducidas/eliminadas) se muestran los estadísticos lambda Wilks y su correspondiente F asociados a las variables que entran o salen en cada paso del proceso, en cada paso, se selecciona la variable que en conjunto con las ya introducidas minimice el valor de lambda Wilks global. En esta tabla se muestra el estadístico correspondiente al método seleccionado, por lo que el encabezado de la tabla cambiará dependiendo del método seleccionado, que en el caso de la presente aplicación corresponde al criterio basado en el estadístico lambda de Wilks. Cabe señalar que el estadístico lambda Wilks que se presenta en este cuadro es el estadístico lambda Wilks global y este va disminuyendo conforme se introducen variables en el modelo.

El cuadro 4.11 (variables en el análisis) muestra los estadísticos de las variables incluidas en el modelo en cada paso, el nivel de tolerancia alcanzado, el estadístico F de Salida y el estadístico Lambda Wilks. Las variables con una tolerancia muy baja serán eliminadas del modelo ya que esto indica multicolinealidad. En este cuadro, se muestra únicamente un resumen de los pasos, correspondientes a los pasos 1, 2, 3 y 90 debido al gran espacio que ocuparía el mostrar cada uno de los 90 pasos del proceso, además de que esto resultaría poco ilustrativo.

El cuadro 4.12 (variables no incluidas en el análisis) presenta los estadísticos para las variables que no están incluidas en el modelo en cada paso (Tolerancia, Tolerancia mínima, F para introducir y lambda Wilks). De esta manera, en cada paso será introducida la variable que tenga asociado un valor de F para introducir más alto o equivalentemente un valor de lambda Wilks más bajo. De igual manera que para el cuadro 4.11. sólo se muestra en resumen de los pasos, que en este caso corresponden a los pasos 0, 88, 89 y 90.

El cuadro 4.13 (Lambda de Wilks en cada paso) muestra el estadístico lambda de Wilks global para cada paso de la selección, en el caso de la presente aplicación, dado que se emplea el criterio basado en el estadístico lambda de Wilks, la información contenida en esta tabla es la misma que la contenida en el cuadro 4.10.

Al iniciar el proceso de selección de variables (paso cero en el cuadro 4.12), la primera variable candidata a ser incluida es aquella que tiene en valor de lambda Wilks más pequeño, que en este caso es la variable closet (¿Esta vivienda cuenta con closet?) (lambda = 0.662), al revisar las condiciones mínimas para su inclusión, se comprueba que su estadístico F de Entrada (F Entrada closet = 2480513.442) es muy superior al valor fijado (F=3.84) , en este paso, la tolerancia es igual a 1 dado que no existen aún variables incluidas en el modelo. Puesto que cumple con las condiciones necesarias, la variable closet será introducida en el modelo en el paso 1 (ver cuadro 4.11 paso 1).

Después de incluir la variable closet, se seleccionará la variable que en conjunto con ésta logre la mayor discriminación entre los grupos, que es la variable edu_2 (Educación del jefe de familia (Universidad, Maestría, Doctorado)) con un valor de lambda asociado de 0.524 (cuadro

4.10), el valor del estadístico F asociado a esta variable es F=1286748.382 (correspondería al paso 1 en cuadro 4.12) que es muy superior al nivel fijado, el nivel de tolerancia (F=.983) sigue siendo superior al nivel fijado como mínimo, por lo que la variable edu_2 es incluida en el modelo.

Después del paso 1, el modelo cuenta ya con dos variables, por lo que es necesario revisar las condiciones de salida, en el cuadro 4.11 (paso 2) podemos ver que ninguna de las dos variables incluidas hasta este paso en el modelo tiene valores inferiores al nivel fijado para el estadístico F de salida (F=2.71) por lo que las dos variables seleccionadas permanecerán en el modelo.

El proceso de selección continúa de manera similar con el resto de las variables hasta el paso 89, en el que únicamente quedan fuera del modelo cuatro variables, la única candidata a ser seleccionada es la variable techo3 que presenta un valor de lambda de 0.201, y que cumple con las condiciones de entrada (F de entrada = 527.336 >3.84 , Tolerancia = 0.65, Tolerancia mínima =0.008) (ver cuadro 4.12 paso 89). Por lo que esta variable es incluida en el modelo. Al revisar las condiciones de salida tras la inclusión de la variable techo3, ninguna de las 90 variables incluidas en el modelo presenta valores F de salida inferiores a 2.71 (ver cuadro 4.11 paso 90), por lo que las 90 variables permanecerán en el modelo. En el paso 90, quedan tres variables fuera del modelo (norte (zona norte), edu_0 (Educación del jefe de familia Ninguna, Preprimaria, Primaria), edo_civ1 (Estado civil del jefe de familia: unión libre)), sin embargo ninguna de estas variables puede ser incluida en el modelo ya que el nivel de tolerancia que les está asociado es 0, lo que indica que estas variables tienen una relación lineal con alguna de las variables ya incluidas en el modelo, en efecto, por ejemplo la variable norte tiene una relación lineal con las variables centro y sur previamente incluidas en el modelo y como se mencionó, al utilizar el método Stepwise se garantiza que no haya multicolinealidad ni singularidad en la matriz de varianza covarianza global. Por ello, las tres variables mencionadas no son incluidas en el modelo por aportar información redundante.

De esta manera quedan seleccionadas 90 de las 93 variables contempladas para ser incluidas en el modelo y con ello se logra obtener un valor global de lambda de Wilks de 0.201, que resulta significativo (sig=0.000) (ver cuadro 4.10 paso 90 o cuadro 4.13 paso 90). Como se puede observar, en cada paso de la selección el valor del estadístico lambda de Wilks va disminuyendo conforme se introducen variables en el modelo, sin embargo, aún en el paso 90 este valor no es demasiado cercano a cero, lo que nos indica que los grupos están solapados.

Si bien la utilización del método Stepwise garantiza que no exista multicolinealidad ni singularidad en la matriz de varianza covarianza global, esto no es así para las matrices de varianza covarianza por grupos dado que la mayoría de las variables son variables dummy, es decir, el hecho de que no exista multicolinealidad en la matriz de varianza covarianza global no implica que esto se cumpla dentro de los grupos. De hecho, al realizar la prueba M de Box contemplando únicamente las 90 variables seleccionadas, las matrices de varianza covarianza

por grupos siguen siendo singulares y de rangos diferentes. Por ello, será necesario revisar las correlaciones parciales de las variables incluidas, especialmente aquellas que forman parte del conjunto de variables dummy generadas a partir de una variable de más de dos categorías. Al revisar dichas correlaciones parciales, encontramos que dentro de los grupos, existen variables con correlación parcial perfecta, y en cada caso, una de las variables deberá ser eliminada, las matrices de correlación se omiten de nueva cuenta por su gran tamaño. De esta manera quedan eliminadas 5 variables del modelo que son: basura7 (Utiliza el servicio de recolección particular), cla_hog2 (hogar nuclear), muros3 (muros de madera), techo2 (techos de madera), tviv1 (Casa sola que comparte muros o que no comparte muros). Una vez eliminadas estas 5 variables, se repetirá el proceso de selección de variables por el método Stepwise, de manera similar al descrito anteriormente.

Al realizar nuevamente este proceso con únicamente 85 variables, como era de esperarse, todas las variables son incluidas en el modelo y se obtiene un valor de lambda de 0.204 que resulta significativo (sig=0.000). Por medio de este proceso de selección, se contemplan entonces un total de 85 variables a ser incluidas en el modelo y que serán utilizadas para la construcción de las funciones discriminantes. Los cuadros correspondientes a este nuevo proceso de selección se omiten por ser similares a los del proceso de selección anterior.

4.5.1 Obtención M box

Una vez eliminadas las variables correlacionadas dentro de los grupos y sin tomar en cuenta la diferencia de grupos, es posible realizar la prueba M de Box, puesto que las matrices de varianza covarianza de todos los grupos son no singulares y son de rango completo, los resultados obtenidos por medio de la realización de esta prueba se muestran en los siguientes cuadros.

Logaritmo de los determinantes

Grupo	Rango	Logaritmo del determinante
1	85	-229.35
2	85	-204.619
3	85	-214.003
4	85	-237.537
5	85	-250.665
Intra-grupos combinada	85	-198.937

Los rangos y logaritmos naturales de los determinantes impresos son los de las matrices de covarianza de los grupos.

Cuadro 4.14 : Logaritmos de los determinantes de las matrices de varianza covarianza por grupo

Resultados de la prueba

M de Box		4E+008
F	Aprox.	30747.968
	gl1	14620
	gl2	3E+013
	Sig.	.000

Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.

Cuadro 4.15: Resultados de la prueba M de Box

El cuadro 4.14 muestra los log determinantes de cada una de las matrices de varianza covarianza (es decir el producto de sus eigenvalores), que sirven para distinguir los grupos cuyas matrices de varianza covarianza difieren más, como podemos observar, en el caso de los resultados obtenidos para esta aplicación, los grupos que difieren más son los grupos 1, 4 y 5.

El cuadro 4.15 muestra los resultados de la prueba M de Box realizada para la comprobación del supuesto de igualdad de matrices de varianza covarianza, el estadístico M se calcula a partir de los determinantes de las matrices, y para el caso de esta aplicación toma un valor de M= 449544989.4163. La significancia de este estadístico se basa en una transformación de éste a un estadístico F, que en este caso toma un valor aproximado de F=30747.9683654 con una probabilidad asociada de p=0.000, lo que permite rechazar la prueba de igualdad de matrices de varianza covarianza de los 5 grupos.

Como ya habíamos mencionado, pese a que la prueba arroja resultados que indican diferencias entre las matrices de varianza covarianza, estos deben ser matizados por los motivos previamente expuestos (ver 4.3.3).

4.6 Obtención de las funciones discriminantes

Una vez terminado el proceso de selección de variables, se procede a la obtención de las funciones discriminantes que en el caso de esta aplicación ascenderán a 4 (min(5-1, 85 variables), ver 3.3.1), y se llevará a cabo utilizando las 85 variables seleccionadas para el modelo. A continuación se presentan los coeficientes no estandarizados de estas funciones:

Coeficientes de las funciones canónicas discriminantes

	Función			
	1	2	3	4
tam_hog	-0.153	0.032	0.063	-0.017
centro	0.013	-0.034	0.135	0.413
sur	-0.047	-0.012	0.45	0.08
urbano	-0.028	-0.286	-0.644	0.109
cla_hog1	0.585	0.74	-0.586	-0.07
cla_hog3	-0.099	-0.097	-0.115	0.152
tviv_3	0.303	-0.865	-0.529	0.411
tviv_4	0.021	0.082	-0.561	-0.074
viv_1	-0.038	-0.333	0.505	-0.469
viv_3	0.363	-0.188	-0.03	-0.646
viv_4	0.28	-0.393	-0.158	0.522
viv_6	-0.086	-0.213	0.09	-0.027
viv_7	0.016	-0.328	-0.11	-0.258
hacina	0.025	0.08	-0.108	0.027
depend2	-0.259	0.056	0.024	-0.046
cocina	0.082	0.124	-0.145	-0.271
muros4	-0.039	0.33	-0.113	0.062
muros5	0.079	0.291	-0.125	0.066
murosr	0.001	-0.155	-0.05	0.069
techo3	0.017	0.042	-0.148	0.014
techo5	0.043	0.031	-0.397	0.048
techo6	-0.153	-0.005	0.542	0.323
techosr	0.109	-0.012	-0.261	-0.389
pisos2	-0.075	-0.469	0.296	0.012
pisos3	0.069	-0.49	1.24	0.248
pisos4	0.136	-0.985	2.761	-0.01
pisos5	0.337	-0.354	1.211	-0.959
basura2	-0.116	0.234	0.034	0.576
basura5	0.014	0.106	-0.17	1.356
basura6	0.054	0.158	0.208	0.569
combust5	-0.004	-0.217	-0.343	-0.148
tinaco	-0.065	0.008	-0.063	-0.022
cisterna	0.057	0.101	-0.226	0.341
bomba	0.167	0.044	0.477	-0.476
calentad	0.129	-0.024	0.015	0.169
s_aireac	0.26	0.14	0.334	-0.829
lavabo	0.06	-0.328	-0.164	-0.197
lavadero	-0.002	0.237	-0.161	0.423
fregader	0.093	-0.369	-0.059	-0.111
closet	0.338	-0.046	0.412	0.098
telefono	0.275	-0.387	-0.157	-0.439
celular	0.383	0.128	-0.097	0.257

Cuadro 4.16 : Coeficientes no estandarizados para las funciones discriminantes.

Coeficientes de las funciones canónicas discriminantes

	Función			
	1	2	3	4
tvcable	0.491	0.866	-0.246	-0.143
automov	0.529	0.051	0.355	0.559
moto	0.05	-0.742	0.103	-1.206
bici	-0.051	0.166	-0.285	0.017
radio	0.095	0.002	0.388	-0.036
graba	0.08	-0.012	0.011	-0.106
estereo	0.043	-0.207	0.265	-0.126
repro_cd	0.262	0.107	-0.053	-0.197
tv_bn	-0.133	-0.082	0.267	0.33
tv_color	0.056	-0.24	0.275	0.672
video	0.136	0.005	-0.295	-0.277
dvd	0.236	0.955	-1.807	0.795
compu	0.46	0.071	0.963	0.389
videoj	-0.062	-0.292	-0.119	1.257
licuador	-0.04	0.048	-0.231	-0.065
batidor	0.04	-0.253	0.128	0.09
extracto	0.205	0.014	-0.182	0.051
tostador	0.291	0.536	-0.069	0.07
cafetera	0.195	0.331	0.478	-0.017
sandwich	0.177	0.278	-0.219	-1.174
jugos	0.02	-0.004	0.195	0.494
horno_el	0.13	0.352	-0.879	0.209
micro	0.069	-0.006	0.069	-0.266
refri	0.078	-0.099	-0.385	-0.096
estufa_g	-0.002	0.126	-0.011	-0.428
molino	-0.049	0.116	-0.121	0.183
lavadora	-0.046	-0.267	-0.086	0.177
maq_cos	-0.05	0.001	0.064	-0.115
aire_ac	0.245	-0.025	-0.599	1.344
menores	-0.083	0.119	0.121	0.023
smedich	0.371	-0.635	0.286	0.459
sexo_jef	0.258	0.029	0.064	0.35
edad_jef	-0.002	0.006	0.001	-0.008
edu_1	0.173	-0.117	0.089	-0.134
edu_2	1.082	1.342	0.038	-0.155
edo_civ2	-0.083	-0.086	0.265	-0.122
edo_civ3	0.407	0.223	0.533	0.44
edo_civ4	0.731	0.649	0.429	-0.157
edo_civ5	0.173	-0.039	0.897	0.091
edo_civ6	0.381	0.304	0.758	1.114
jno_trab	0.126	-0.066	-0.044	-0.854
prestac0	-0.044	0.324	0.489	0.15
servmed	-0.06	0.55	0.599	-0.279
(Constante)	-1.063	-0.021	-0.574	0.044
Coeficientes no tipificados				

Cuadro 4.16 : Coeficientes no estandarizados para las funciones discriminantes.(Continuación)

4.6.1 Funciones discriminantes significativas.

Antes de proceder a interpretar los resultados, es necesario determinar si cada una de las funciones discriminantes obtenidas resulta significativa, para ello, en un primer paso, se obtienen los eigenvalores asociados a cada una de las funciones. El cuadro 4.17 presenta los eigenvalores asociados a cada una de las funciones discriminantes, recordaremos que los eigenvalores reflejan el grado en el que una función discrimina entre los diferentes grupos (ver 3.6.1).

Eigenvalores

Función	Eigenvalor	% de varianza	% acumulado	Correlación canónica
1	2.551(a)	87.7	87.7	0.848
2	.281(a)	9.7	97.4	0.468
3	.046(a)	1.6	99	0.211
4	.030(a)	1	100	0.17
a Se han empleado las 4 primeras funciones discriminantes canónicas en el análisis.				

Cuadro 4.17: Eigenvalores y porcentaje de varianza explicada por las funciones discriminantes.

Como podemos observar, la primera función discriminante es la que presenta un eigenvalor asociado más grande ($\boldsymbol{\lambda}_1 = 2.551$) y por lo tanto esta función es la que presenta un mayor poder discriminante. Las restantes tres funciones discriminantes presentan eigenvalores pequeños lo que nos indica que no tienen demasiado poder de discriminación, particularmente la cuarta función discriminante cuyo eigenvalor alcanza apenas un valor de $\boldsymbol{\lambda}_4 = 0.030$.

El examinar los porcentajes relativos, que permiten determinar que tan importante es una función discriminante respecto a las demás, podemos ver que la primer función es responsable del 87.7% de la varianza entre grupos, y que contando las tres primeras funciones discriminantes se explica el 99% de la dispersión entre los grupos. Las correlaciones que se presentan en el cuadro 4.17 muestran el grado de asociación entre las puntuaciones discriminantes y los grupos, podemos observar que esta medida es mucho más cercana a uno para la primera función (.848) que para la cuarta o incluso la tercera (.221 y .170 respectivamente).

Si bien estos resultados indican que la primera función discriminante es mucho más importante que las demás, e incluso que las últimas dos funciones podrían no ser muy relevantes, estos resultados no permiten valorar la significancia de cada una de las funciones obtenidas. Para determinar si las cuatro funciones discriminantes obtenidas son significativas, se realiza una prueba basada en el estadístico lambda Wilks (ver 3.6.1) , para contrastar la hipótesis nula de igualdad entre las puntuaciones alcanzadas por los grupos para las cuatro funciones discriminantes. Los resultados obtenidos por medio de la aplicación de esta prueba se presentan a continuación:

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 4	0.204	30947294.65	340	0
2 a la 4	0.724	6274802.651	252	0
3 a la 4	0.928	1453799.445	166	0
4	0.971	568956.48	82	0

Cuadro 4.18: Prueba de contraste para las funciones discriminantes

Para determinar la significancia de las funciones discriminantes, se utiliza una transformación de este estadístico en un estadístico Chi-cuadrado, al realizar esta transformación para los estadísticos lambda obtenidos, los valores de los estadísticos Chi cuadrados resultan significativos en todos los casos, incluyendo el caso en el que sólo se considera la cuarta función que es la que menos poder de discriminación tiene (χ^2= 568956.480, sig=0.000). Por lo tanto, se conservarán las cuatro funciones discriminantes, teniendo en cuenta que las dos primeras funciones son las que mayor poder de discriminación poseen.

4.7 Interpretación de los resultados.

Después de determinar que funciones discriminantes son significativas, y por lo tanto deberían de ser conservadas, en el caso de esta aplicación se conservaron las cuatro funciones discriminantes obtenidas, se procede a la interpretación de las mismas considerando la posición que determinan para cada uno de los casos y para los centroides de los grupos.

El siguiente cuadro presenta las puntuaciones discriminantes para los centroides de cada grupo que resultan de la evaluación de las funciones discriminantes no estandarizadas en los centroides de cada grupo.

Funciones en los centroides de los grupos

Grupo	Función 1	2	3	4
1	-1.332	.265	.028	.002
2	.397	-.708	-.213	.016
3	1.894	-.257	.405	-.155
4	3.088	.507	.137	.772
5	3.907	1.414	-.459	-.266

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Cuadro 4.19: Puntuaciones discriminantes para los centroides de cada grupo

El cálculo de las puntuaciones discriminantes para los centroides de los grupos (así como para cada caso), permite la determinación de las coordenadas de estos sobre los ejes discriminantes. Como podemos observar en el cuadro 4.19, las puntuaciones alcanzadas por cada grupo sobre el eje correspondiente a la primera función discriminante determinan una mayor separación entre los grupos, que las alcanzadas a lo largo de los ejes correspondientes a las otras funciones discriminantes, particularmente en el caso de la cuarta función discriminante.

Debido a que para esta aplicación se cuenta con cuatro ejes discriminantes, resulta complicado determinar la distancia entre los centroides, por lo que se recurrirá a diagramas de dispersión para determinar que tan lejanos o cercanos se encuentran los centroides respecto a las dos primeras funciones discriminantes, que son las que mayor poder de discriminación poseen. A continuación se presentan los diagramas de dispersión de los grupos respecto a las dos primeras funciones discriminantes de manera conjunta e individual.

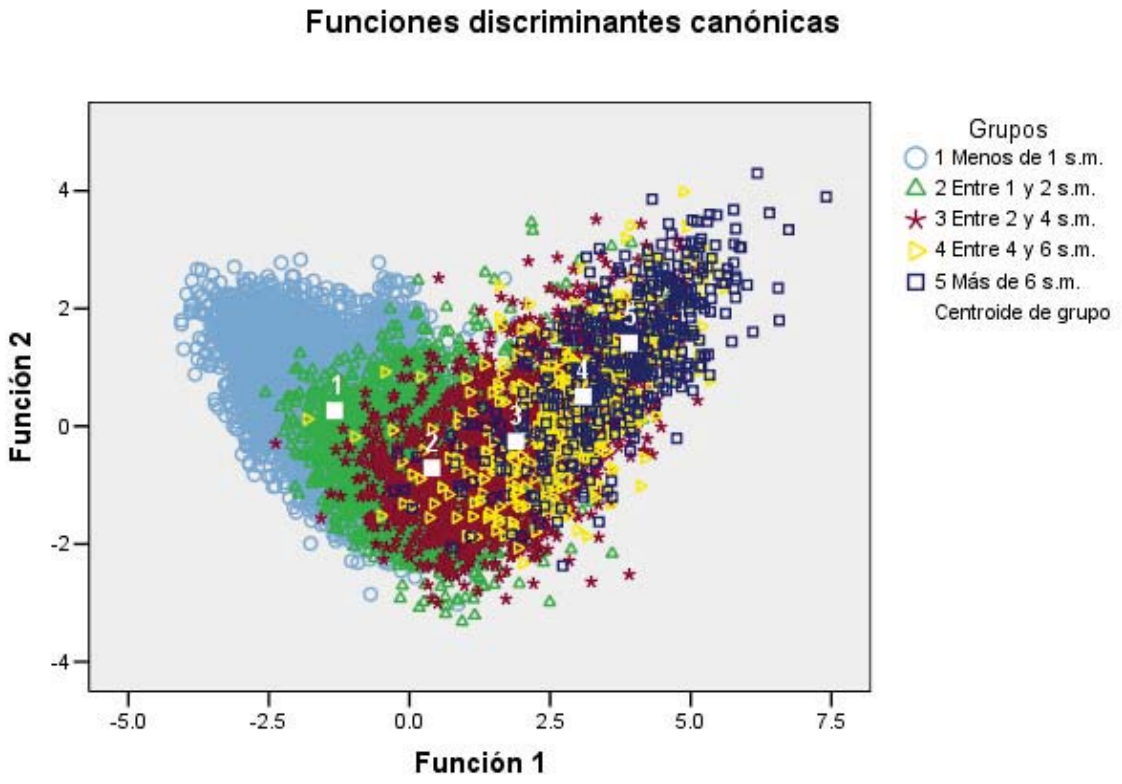


Figura 4.1: Diagrama de dispersión para todos los grupos

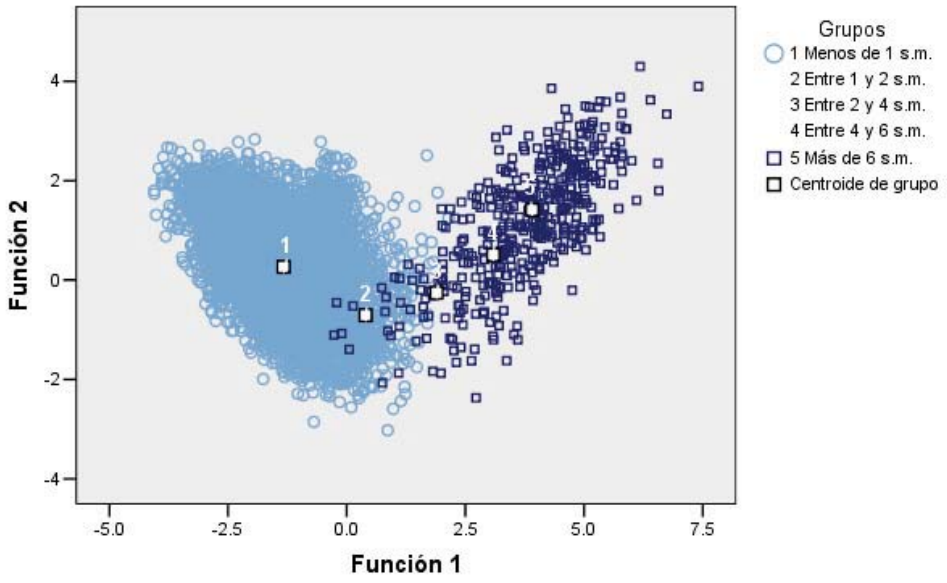


Figura 4.2: Diagrama de dispersión para los grupos 1 y 5

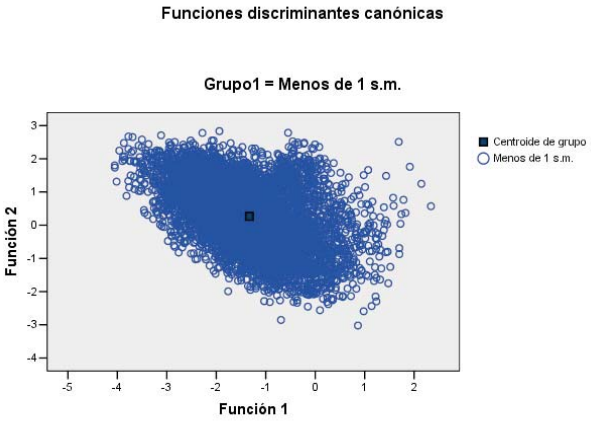


Figura 4.3: Diagrama de dispersión grupo 1

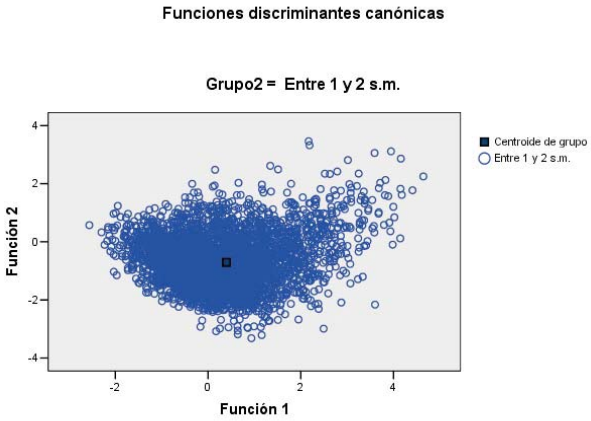


Figura 4.4: Diagrama de dispersión grupo 2

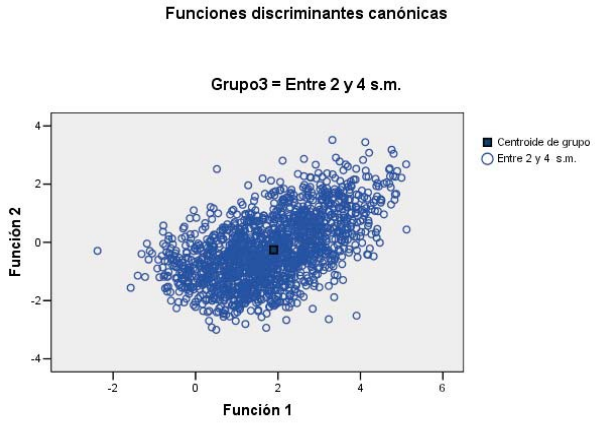


Figura 4.5: Diagrama de dispersión grupo 3

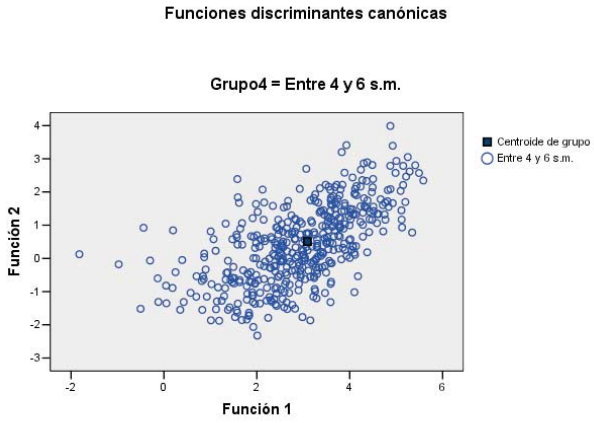


Figura 4.6: Diagrama de dispersión grupo 4

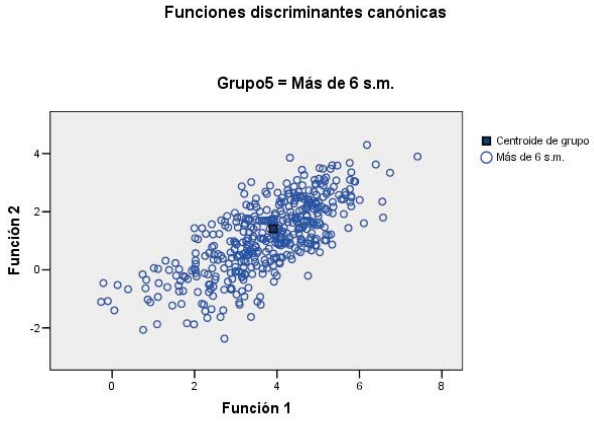


Figura 4.7: Diagrama de dispersión grupo 5

Al examinar el diagrama de dispersión de los grupos y sus centroides respecto a las dos primeras funciones discriminantes (Figura 4.1), lo primero que salta a la vista es el solapamiento de los grupos que, dadas las características de la aplicación que estamos realizando, era de esperarse como se mencionó anteriormente y que puede traer consigo una tasa de clasificación baja. Como podemos observar (Figura 4.2) incluso los grupos 1 y 5 presentan un ligero solapamiento.

Los diagramas de dispersión también permiten, como su nombre lo indica, ver que tan dispersos se encuentran los casos respecto a sus centroides, las figuras 4.3 a 4.7 reflejan que pese a la eliminación de outliers, los casos de los grupos en general se encuentran dispersos respecto a sus centroides, particularmente en los dos últimos grupos, lo que provoca a su vez el solapamiento de los grupos.

Al examinar más detalladamente el diagrama de dispersión de los grupos (figura 4.1) y las puntuaciones discriminantes de los centroides, podemos observar que pese al solapamiento de los grupos, estos se diferencian a lo largo del eje de las abscisas (Primera Función discriminante) y en menor medida a lo largo del eje de las ordenadas (Segunda función discriminante). La primera función discriminante permite diferenciar entonces en mayor medida los grupos, en particular el grupo 1 del grupo 5, mientras que la segunda función discriminante diferencía (en menor medida que la primera) los grupos 2 y 3 del resto de los grupos, en particular del grupo 5, y también permite una distinción, aunque en menor medida entre los grupos 1 y 5.

En lo referente a las dos últimas funciones discriminantes, no se presentan diagramas de dispersión correspondientes a estas funciones dado que su relevancia es menor que la de las dos primeras funciones como ya habíamos mencionado. Sin embargo, al analizar las puntuaciones de los centroides para estas dos funciones discriminantes, podemos ver que la tercera función discriminante permite diferenciar el grupo 3 de los demás, particularmente del grupo 5, sin embargo, las diferencias entre las puntuaciones de los centroides son mucho menores que para las funciones anteriores. Finalmente, como puede observarse en el cuadro 4.19, para la cuarta función discriminante, que resulta ser la de menor poder de discriminación, las puntuaciones de los centroides alcanzadas para esta función se encuentran relativamente próximas y únicamente se aprecia una ligera diferenciación entre el grupo 4 y los demás grupos.

4.7.1 Contribución de las variables

Otro aspecto esencial para la interpretación de las funciones discriminantes es la contribución de las variables discriminantes o variables independientes. Los coeficientes no estandarizados (cuadro 4.16) pueden interpretarse como la contribución absoluta de una variable a la puntuación discriminante obtenida, sin embargo, estos coeficientes proporcionan poca información acerca de la contribución de las variables al modelo ya que no son comparables como habíamos mencionado en el capítulo anterior. Por ello, es preferible utilizar los coeficientes estandarizados (que no dependen de la métrica de las variables) para analizar la contribución relativa de las variables independientes. El cuadro 4.20 presenta los coeficientes estandarizados de las funciones discriminantes.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función			
	1	2	3	4
tam_hog	-0.257	0.054	0.105	-0.029
centro	0.007	-0.017	0.067	0.203
sur	-0.021	-0.005	0.197	0.035
urbano	-0.012	-0.124	-0.28	0.047
cla_hog1	0.145	0.183	-0.145	-0.017
cla_hog3	-0.04	-0.039	-0.047	0.062
tviv_3	0.039	-0.111	-0.068	0.053
tviv_4	0.005	0.019	-0.128	-0.017
viv_1	-0.012	-0.106	0.16	-0.149
viv_3	0.121	-0.063	-0.01	-0.216
viv_4	0.065	-0.092	-0.037	0.122
viv_6	-0.024	-0.058	0.025	-0.007
viv_7	0.008	-0.161	-0.054	-0.126
hacina	0.028	0.087	-0.118	0.029
depend2	-0.32	0.069	0.03	-0.057
cocina	0.024	0.037	-0.043	-0.081
muros4	-0.011	0.091	-0.031	0.017
muros5	0.028	0.103	-0.044	0.023
murosr	0	-0.066	-0.021	0.03
techo3	0.005	0.013	-0.047	0.004
techo5	0.019	0.014	-0.175	0.021
techo6	-0.03	-0.001	0.106	0.063
techosr	0.048	-0.005	-0.114	-0.169
pisos2	-0.034	-0.214	0.135	0.006
pisos3	0.027	-0.191	0.483	0.097
pisos4	0.025	-0.184	0.516	-0.002
pisos5	0.103	-0.108	0.369	-0.292
basura2	-0.044	0.089	0.013	0.22
basura5	0.003	0.021	-0.033	0.266
basura6	0.023	0.068	0.089	0.243
combust5	-0.001	-0.071	-0.112	-0.048
tinaco	-0.03	0.003	-0.029	-0.01
cisterna	0.023	0.041	-0.092	0.139
bomba	0.067	0.018	0.192	-0.192
calentad	0.054	-0.01	0.006	0.07
s_aireac	0.043	0.023	0.056	-0.138
lavabo	0.025	-0.136	-0.068	-0.081
lavadero	0	0.07	-0.048	0.125
fregader	0.038	-0.151	-0.024	-0.046
closet	0.131	-0.018	0.16	0.038
telefono	0.118	-0.166	-0.067	-0.188
celular	0.142	0.048	-0.036	0.095

Cuadro 4.20: Coeficientes estandarizados de las funciones discriminantes.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función			
	1	2	3	4
tvcable	0.151	0.266	-0.076	-0.044
automov	0.221	0.021	0.149	0.234
moto	0.005	-0.067	0.009	-0.109
bici	-0.016	0.054	-0.093	0.006
radio	0.043	0.001	0.175	-0.016
graba	0.04	-0.006	0.005	-0.053
estereo	0.02	-0.096	0.123	-0.059
repro_cd	0.086	0.035	-0.017	-0.064
tv_bn	-0.049	-0.03	0.098	0.121
tv_color	0.021	-0.088	0.101	0.247
video	0.059	0.002	-0.128	-0.12
dvd	0.052	0.209	-0.395	0.174
compu	0.134	0.021	0.281	0.113
videoj	-0.017	-0.079	-0.032	0.338
licuador	-0.014	0.017	-0.081	-0.023
batidor	0.016	-0.103	0.052	0.037
extracto	0.074	0.005	-0.066	0.018
tostador	0.1	0.185	-0.024	0.024
cafetera	0.063	0.107	0.154	-0.006
sandwich	0.052	0.081	-0.064	-0.343
jugos	0.007	-0.001	0.07	0.177
horno_el	0.034	0.092	-0.23	0.055
micro	0.027	-0.002	0.027	-0.106
refri	0.03	-0.037	-0.146	-0.036
estufa_g	-0.001	0.039	-0.003	-0.132
molino	-0.017	0.04	-0.041	0.063
lavadora	-0.021	-0.121	-0.039	0.08
maq_cos	-0.022	0.001	0.028	-0.05
aire_ac	0.062	-0.006	-0.151	0.339
menores	-0.09	0.13	0.132	0.025
smedich	0.172	-0.294	0.133	0.212
sexo_jef	0.104	0.012	0.026	0.141
edad_jef	-0.03	0.093	0.014	-0.128
edu_1	0.079	-0.053	0.04	-0.061
edu_2	0.309	0.383	0.011	-0.044
edo_civ2	-0.04	-0.041	0.127	-0.059
edo_civ3	0.104	0.057	0.137	0.113
edo_civ4	0.108	0.096	0.063	-0.023
edo_civ5	0.051	-0.012	0.267	0.027
edo_civ6	0.086	0.069	0.171	0.252
jno_trab	0.047	-0.025	-0.017	-0.321
prestac0	-0.02	0.146	0.22	0.067
servmed	-0.027	0.243	0.265	-0.123

Cuadro 4.20: Coeficientes estandarizados de las funciones discriminantes. (Continuación)

Al examinar los coeficientes estandarizados podemos entonces determinar qué variables contribuyen más a las puntuaciones alcanzadas en esa función, de hecho, el valor absoluto de los coeficientes estandarizados nos indicará la importancia de la contribución de cada variable a las puntuaciones discriminantes.

El cuadro 4.21 muestra, en orden descendente, las veinte variables con coeficientes estandarizados de mayor magnitud para cada función discriminante (coeficientes sombreados en el cuadro 4.20).

Función 1		Función 2		Función 3		Función 4	
depend2	-0.32	edu_2	0.383	pisos4	0.516	sandwich	-0.343
edu_2	0.309	smedich	-0.294	pisos3	0.483	aire_ac	0.339
tam_hog	-0.257	tvcable	0.266	dvd	-0.395	videoj	0.338
automov	0.221	servmed	0.243	pisos5	0.369	jno_trab	-0.321
smedich	0.172	pisos2	-0.214	compu	0.281	pisos5	-0.292
tvcable	0.151	dvd	0.209	urbano	-0.28	basura5	0.266
cla_hog1	0.145	pisos3	-0.191	edo_civ5	0.267	edo_civ6	0.252
celular	0.142	tostador	0.185	servmed	0.265	tv_color	0.247
compu	0.134	pisos4	-0.184	horno_el	-0.23	basura6	0.243
closet	0.131	cla_hog1	0.183	prestac0	0.22	automov	0.234
viv_3	0.121	telefono	-0.166	sur	0.197	basura2	0.22
telefono	0.118	viv_7	-0.161	bomba	0.192	viv_3	-0.216
edo_civ4	0.108	fregader	-0.151	techo5	-0.175	smedich	0.212
sexo_jef	0.104	prestac0	0.146	radio	0.175	bomba	-0.192
edo_civ3	0.104	lavabo	-0.136	edo_civ6	0.171	telefono	-0.188
pisos5	0.103	menores	0.13	viv_1	0.16	jugos	0.177
tostador	0.1	urbano	-0.124	closet	0.16	dvd	0.174
menores	-0.09	lavadora	-0.121	cafetera	0.154	techosr	-0.169
repro_cd	0.086	tviv_3	-0.111	aire_ac	-0.151	centro	0.203
edo_civ6	0.086	pisos5	-0.108	automov	0.149	viv_1	-0.149

Cuadro 4.21: 20 variables con coeficientes estandarizados de mayor magnitud.

Como podemos observar en el cuadro anterior, la variable que contribuye de mayor manera a la primer función discriminante es la variable depend (índice de dependencia), seguida por las variables edu2 (educación del jefe del hogar: Universidad, Maestría, Doctorado), tam_hog (tamaño del hogar), automov (¿Cuenta con automóvil, camioneta o camioneta de caja?), smedich (Al menos uno tiene servicio médico en el hogar), etc. De manera análoga, las variables que más contribuyen al segundo eje discriminante son edu2, smedich, tvcable (¿Cuenta con televisión por cable, Sky, Direct-tv o multivisión para el uso del hogar?), servmed (El jefe del hogar tiene derecho a servicios médicos del IMSS, ISSSTE, etc.). La interpretación de la contribución de las variables se presentará brevemente para las dos primeras funciones discriminantes que son las que mayor importancia tienen y presentando sólo algunas de las variables ya que la interpretación para cada una de las 85 variables independientes resultaría poco ilustrativa.

Para determinar la contribución de las variables, también se recurre a la matriz de estructura que presenta las correlaciones entre las variables discriminantes y la función discriminante

canónica. Dentro de la matriz de estructura, las variables son ordenadas de manera descendente, de acuerdo con su grado de correlación con la función discriminante en cuestión. Es frecuente que en la matriz de estructura se incluyan todas las variables discriminantes, incluso aquellas que no serán incluidas en el modelo debido a que presentan colinealidad con otras variables. Si bien en algunos casos esto puede resultar útil para la interpretación de las funciones discriminantes, en el caso de la presente aplicación, sólo se mostrarán en la matriz de estructura las variables que serán incluidas en el modelo, esto debido principalmente a que el número de variables consideradas es grande y la interpretación de las funciones con un gran número de variables resulta complicada. A continuación se presenta la matriz de estructura.

Matriz de estructura

	Función			
	1	2	3	4
closet	.444(*)	-0.124	0.201	-0.006
fregader	.416(*)	-0.382	0.005	-0.065
edu_2	.404(*)	0.398	0.094	-0.02
lavabo	.399(*)	-0.388	-0.032	-0.102
calentad	.398(*)	-0.26	0.006	-0.032
compu	.378(*)	0.138	0.244	0.158
tostador	.361(*)	0.138	-0.009	-0.042
micro	.354(*)	-0.084	0.027	-0.108
automov	.347(*)	-0.08	0.125	0.131
telefono	.345(*)	-0.314	-0.05	-0.125
cafetera	.334(*)	0.15	0.108	-0.019
hacina	-.329(*)	0.224	0.053	0.043
extracto	.315(*)	0.012	-0.037	0.02
video	.313(*)	-0.152	-0.074	-0.12
batidor	.313(*)	-0.149	0.032	0.013
techosr	.312(*)	-0.158	-0.086	-0.128
tvcable	.312(*)	0.243	-0.034	0.004
sandwich	.291(*)	0.154	-0.049	-0.276
tinaco	.263(*)	-0.222	-0.005	-0.085
horno_el	.254(*)	0.172	-0.176	0.036
jugos	.252(*)	-0.021	0.092	0.112
bomba	.250(*)	-0.02	0.103	-0.148
celular	.239(*)	0.053	-0.056	0.165
repro_cd	.238(*)	0.074	-0.05	-0.033
menores	-.228(*)	0.154	0.129	0.028
cisterna	.225(*)	-0.016	0.012	-0.049
tam_hog	-.220(*)	-0.001	0.14	0.039
pisos3	.189(*)	-0.136	0.157	0.108
depend2	-.160(*)	-0.01	0.033	-0.097
s_aireac	.146(*)	0.064	0.033	-0.115
tviv_4	.125(*)	-0.025	-0.122	-0.055
bici	-.106(*)	0.03	-0.068	0.009
viv_3	.098(*)	0.012	-0.094	-0.09
cocina	.092(*)	-0.071	-0.03	-0.068
tv_bn	-.092(*)	0.081	0.046	0.049
graba	.049(*)	0.041	0.001	-0.015

Cuadro 4.22: Matriz de estructura

Matriz de estructura

	Función			
	1	2	3	4
combust5	0.217	-.402(*)	-0.228	-0.056
urbano	0.248	-.401(*)	-0.287	0.017
basura2	-0.266	.379(*)	0.157	0.032
refri	0.253	-.378(*)	-0.192	-0.045
tv_color	0.209	-.369(*)	-0.076	0.091
smedich	0.212	-.356(*)	0.095	0.037
lavadora	0.254	-.352(*)	-0.095	0.034
basura6	0.223	-.327(*)	-0.09	-0.058
edu_1	0.057	-.316(*)	-0.022	-0.059
estufa_g	0.166	-.314(*)	-0.172	-0.071
techo5	0.248	-.272(*)	-0.243	-0.044
licuador	0.156	-.266(*)	-0.169	-0.066
estereo	0.231	-.256(*)	0.027	-0.054
murosr	0.216	-.255(*)	-0.026	-0.04
muros5	0.176	-.236(*)	-0.104	-0.001
viv_6	-0.145	.232(*)	0.14	0.064
molino	-0.128	.216(*)	0.067	0.069
prestac0	-0.177	.200(*)	-0.059	0.1
servmed	0.183	-.199(*)	0.087	-0.096
techo3	-0.147	.190(*)	0.083	0.008
cla_hog1	0.098	.178(*)	-0.093	-0.008
tviv_3	0.057	-.123(*)	-0.072	0.046
muros4	-0.103	.118(*)	0.033	0
maq_cos	0.094	-.109(*)	0.023	-0.027
pisos2	-0.27	-0.072	-.326(*)	0.049
dvd	0.222	0.226	-.315(*)	0.154
pisos4	0.106	-0.137	.285(*)	0.074
sur	-0.156	0.203	.222(*)	-0.048
techo6	0.008	-0.028	.201(*)	0.06
radio	0.097	0.016	.141(*)	-0.004
viv_1	-0.039	-0.018	.119(*)	-0.05
centro	0.07	-0.081	-.110(*)	0.049
viv_7	-0.007	-0.087	-.095(*)	-0.039
edo_civ5	-0.016	-0.013	.074(*)	0.012
edad_jef	-0.014	0.009	-.017(*)	-0.005
videoj	0.17	-0.025	-0.018	.348(*)
pisos5	0.217	0.024	0.077	-.274(*)
aire_ac	0.162	-0.045	-0.074	.267(*)
jno_trab	-0.012	0.015	0.017	-.249(*)
edo_civ6	0.088	0.072	0.039	.216(*)
viv_4	0.114	-0.115	0.005	.184(*)
basura5	0.038	-0.039	-0.103	.153(*)
cla_hog3	-0.058	-0.056	0.054	.114(*)
edo_civ2	-0.018	-0.081	0	-.109(*)
moto	0.033	-0.065	0.018	-.105(*)
edo_civ4	0.081	0.044	0.009	-.099(*)
lavadero	0.075	-0.052	-0.059	.099(*)
edo_civ3	0.018	0.029	-0.024	.053(*)
sexo_jef	-0.024	0.003	-0.031	-.051(*)

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas

Variables ordenadas por el tamaño de la correlación con la función.

* Mayor correlación absoluta entre cada variable y cualquier función discriminante.

Cuadro 4.22: Matriz de estructura (Continuación)

Como podemos observar en el cuadro 4.22, la variable que tiene una mayor correlación con la primera función discriminante es la variable closet (¿Esta vivienda cuenta con closet?) que es a su vez una de las 10 primeras variables con coeficientes estándar de mayor magnitud. La segunda variable con mayor correlación con la primer función discriminante es la variable fregader (¿Esta vivienda cuenta con fregadero?) que no es una de las 20 primeras variables con coeficiente estándar de mayor magnitud, seguida por la variable depend2 (índice de dependencia) que es la variable con mayor coeficiente estándar.

Las variables que presentan una mayor correlación con la segunda función discriminante son combust5 (Utilizan gas para cocinar), urbano (Estrato Urbano/Rural) y basura2 (Queman la basura), de estas tres variables, únicamente la variable urbano figura en la lista de las 20 variables con mayores coeficientes estándar para la segunda función discriminante. Sin embargo, podemos observar que las variables edu_2, smedich y tvcable que son las tres primeras variables con mayores coeficientes estándar para la segunda función discriminante presentan valores de correlación no muy lejanos a los de las variables con mayor correlación para esta función.

En general, podemos observar que tanto los coeficientes de correlación como los coeficientes estándar de las funciones presentan valores relativamente pequeños para todas las variables, por lo que podemos anticipar errores de clasificación debido al solapamiento de los grupos que a su vez esta ligado tanto a la manera en la que se construyó la variable dependiente como a las características de esta última. Es decir, es de esperarse que un hogar con un ingreso mensual per cápita de por ejemplo 1 salario mínimo tenga características similares a las de un hogar con un ingreso mensual per cápita de 1.5 salarios mínimos.

Además de esto, el hecho de presentar alguna característica en particular no garantiza la pertenencia a un grupo en particular (por ejemplo que el jefe de familia tenga un nivel de educación de universidad, maestría o doctorado no implica que el hogar en cuestión no pueda pertenecer al grupo de hogares con menor ingreso), por lo que los coeficientes estándar son relativamente pequeños, sin embargo habrá que esperar que en conjunto las características del hogar logren discriminar el grupo de pertenencia de los hogares.

Los coeficientes estandarizados así como la matriz de estructura revelan información acerca de la importancia de las variables, para el análisis de la relación directa entre las variables y las funciones discriminantes se recurre a los coeficientes no estandarizados (cuadro 4.16) y se analiza la dirección en la que las puntuaciones se mueven respecto a los centroides de los grupos.

Por ejemplo, para las variables referentes a la educación del jefe de familia (edu_1 y edu_2) y respecto a la primera función discriminante, si el jefe de familia tiene un nivel de educación de "secundaria o preparatoria" (edu_1=1 y edu_2=0) tendríamos entonces 1*(0.173) y 0*(1.082) (ver cuadro 4.16), al tener estos valores la puntuación discriminante se movería un

poco más hacía los centroides de los grupos 3, 4 y 5 (ver cuadro 4.19), pero en menor medida que si el jefe de familia tuviera un nivel de educación de "universidad, maestría o doctorado" (edu_1=0 y edu_2=1 $\Longrightarrow$0*(0.173) y 1*(1.082)). Ahora bien, si el jefe de familia tiene un nivel de educación de "ninguna, preprimaria o primaria" (edu_1=0 y edu_2=0 $\Longrightarrow$0*(0.173) y 0*(1.082)) la puntuación discriminante no se mueve hacía los grupos de ingreso superior y permanece cercana a los centroides de los grupos 1 y 2. Esto resulta lógico si tomamos en cuenta que a mayor nivel de educación resulta más probable que se tenga un mejor ingreso, pero como habíamos mencionado, esto es relativo puesto que un hogar cuyo jefe familia no tenga educación puede tener un ingreso elevado.

Para las variables categóricas de dos categorías, como es el caso de la mayoría de las variables de equipamiento de la vivienda, la interpretación es muy similar por ejemplo, en el caso de la variable teléfono cuyo coeficiente no estandarizado para la primera función discriminante es 0.275, si el hogar cuenta con teléfono, tendríamos 1*0.275 lo que movería la puntuación hacía los centroides de los grupos 3, 4 y 5; si, por el contrario, el hogar no cuenta con teléfono tendríamos 0*.275 lo que no movería la puntuación. Como podemos observar en el cuadro 4.16, la mayoría de las variables correspondientes al equipamiento del hogar tienen un coeficiente no estandarizado positivo, lo que de nueva cuenta resulta lógico ya que supondríamos que un hogar que cuente con enseres tales como teléfono, celular, dvd etc. tendrá un mayor ingreso que aquellos hogares que carezcan de estos bienes, mencionado de nueva cuenta que este hecho no es válido en todos los casos.

En cuanto las variables continuas, tomemos por ejemplo la variable tam_hog (tamaño del hogar) cuyo coeficiente no estandarizado es -0.153, si un hogar cuenta únicamente con 2 miembros tendríamos (-0.153*2 = -.306) mientras que si el hogar cuenta con 8 miembros tendríamos (-1.53*8 =-1.224), es decir entre más personas haya en un hogar, más la puntuación se moverá hacía el centroide del grupo de menor ingreso (Grupo1). De nueva cuenta esto es concordante con el hecho de que muchos de los hogares con ingresos bajos tienden a ser hogares de familias numerosas, situación que también es discutible.

La interpretación para las demás variables en cuanto a su contribución a las puntuaciones discriminantes de cualquiera de las cuatro funciones obtenidas se realiza de manera similar, de acuerdo con el tipo de variable en cuestión (categórica o continua) y atendiendo siempre a la posición de los centroides de los grupos.

De esta manera hemos visto la importancia de las variables gracias a los coeficientes estandarizados y a la matriz de estructura, así como su contribución a las puntuaciones discriminantes gracias a los coeficientes no estandarizados. Respecto a los coeficientes no estandarizados resulta interesante ver de qué manera influyen diversos factores tales como la educación del jefe de familia, el tamaño del hogar etc.

4.7.2 Clasificación

Como habíamos mencionado, otro de los objetivos de esta aplicación es obtener una regla de clasificación que permita la clasificación de los hogares en los distintos niveles de ingreso, determinados por los grupos generados, a partir de las características de los hogares. El siguiente cuadro presenta los coeficientes de clasificación de Fisher para cada grupo (ver sec. 3.4.1). Al conocer los valores de las variables discriminantes de un hogar determinado, se realiza el cálculo de las puntuaciones de este hogar para cada una de las cinco funciones de clasificación y se asignará ese hogar al grupo en el que haya obtenido una puntuación más alta.

Coeficientes de la función de clasificación

	Grupo				
	1	2	3	4	5
tam_hog	0.65	0.34	0.167	-0.024	-0.139
centro	6.185	6.214	6.231	6.568	6.039
sur	8.164	7.987	8.175	8.063	7.662
urbano	2.315	2.701	2.114	2.136	2.125
cla_hog1	2.988	3.42	4.279	5.635	7.208
cla_hog3	-4.399	-4.447	-4.736	-4.757	-5.016
tviv_3	-1.917	-0.419	-0.754	-0.53	-1.178
tviv_4	-1.096	-1.006	-1.272	-1.102	-0.6
viv_1	76.041	76.171	76.356	75.486	75.339
viv_3	76.357	77.164	77.714	77.414	78.229
viv_4	76.937	77.848	77.903	78.463	77.887
viv_6	74.753	74.79	74.625	74.311	74.022
viv_7	72.962	73.331	73.183	72.743	72.791
hacina	4.377	4.37	4.372	4.518	4.648
depend2	0.309	-0.199	-0.538	-0.853	-0.981
cocina	8.524	8.576	8.711	8.691	9.238
muros4	12.116	11.757	11.767	12.061	12.332
muros5	9.934	9.818	9.979	10.39	10.725
murosr	0.422	0.587	0.475	0.434	0.252
techo3	7.143	7.167	7.117	7.222	7.347
techo5	2.737	2.877	2.702	2.928	3.179
techo6	-0.756	-1.143	-1.095	-1.127	-1.916
techosr	1.573	1.831	1.895	1.726	2.365
pisos2	4.401	4.657	4.513	3.997	3.321
pisos3	2.755	3.057	3.664	3.27	1.885
pisos4	3.519	4.047	5.515	4.176	1.758
pisos5	4.359	4.981	6.238	5.159	5.387
basura2	35.818	35.388	35.242	35.807	35.306
basura5	29.911	29.892	29.624	31.023	29.825
basura6	29.701	29.598	29.781	30.438	29.912
combust5	7.524	7.81	7.519	7.304	7.462
tinaco	-3.33	-3.436	-3.565	-3.641	-3.628
cisterna	-0.917	-0.858	-0.927	-0.405	-0.486
bomba	-0.625	-0.5	0.146	-0.19	0.196
calentad	1.651	1.897	2.06	2.349	2.249
s_aireac	0.713	0.934	1.736	1.296	2.298
lavabo	0.439	0.899	0.772	0.455	0.507
lavadero	7.238	7.049	6.982	7.596	7.468
fregader	0.28	0.813	0.769	0.511	0.403
closet	-0.835	-0.303	0.419	0.768	0.655
telefono	-1.329	-0.446	-0.232	-0.564	-0.142
celular	-0.328	0.237	0.765	1.584	1.805

Cuadro 4.23 : Coeficientes de clasificación

Coeficientes de la función de clasificación

	Grupo				
	1	2	3	4	5
tvcable	-0.562	-0.498	0.5	1.681	3.164
automov	0.256	1.044	1.983	3.078	2.766
moto	-2.146	-1.38	-1.369	-3.019	-2.461
bici	-0.038	-0.218	-0.399	-0.241	0.02
radio	0.732	0.801	1.19	1.169	1.055
graba	0.701	0.846	0.985	0.971	1.129
estereo	0.013	0.224	0.38	0.086	-0.094
repro_cd	-0.74	-0.381	0.061	0.287	0.835
tv_bn	2.022	1.813	1.685	1.699	1.014
tv_color	3.726	4.001	4.032	4.464	3.432
video	0.131	0.428	0.498	0.487	1.066
dvd	-0.625	-0.699	-1.169	1.063	2.373
compu	-0.037	0.464	1.714	2.42	1.883
videoj	-0.085	0.138	-0.375	0.523	-1.026
licuador	-0.04	-0.101	-0.27	-0.279	-0.063
batidor	-0.382	-0.096	-0.087	-0.183	-0.55
extracto	-0.602	-0.216	-0.024	0.328	0.565
tostador	1.444	1.443	2.065	2.906	3.599
cafetera	1.379	1.278	2.017	2.358	2.551
sandwich	-0.106	-0.034	0.423	-0.182	1.564
jugos	-0.285	-0.286	-0.221	0.205	-0.411
horno_el	-0.671	-0.573	-0.799	0.055	0.787
micro	-0.237	-0.133	0.055	-0.132	0.154
refri	0.957	1.28	1.131	1.163	1.467
estufa_g	2.199	2.068	2.188	1.888	2.451
molino	3.839	3.673	3.546	3.779	3.726
lavadora	-0.264	-0.06	-0.333	-0.406	-0.818
maq_cos	-1.56	-1.665	-1.681	-1.863	-1.822
aire_ac	2.09	2.701	2.457	4.137	3.278
menores	1.438	1.15	1.15	1.131	1.075
smedich	1.111	2.307	2.674	2.98	2.061
sexo_jef	14.285	14.692	15.07	15.707	15.544
edad_jef	0.489	0.479	0.481	0.475	0.488
edu_1	4.317	4.707	4.992	4.961	5.083
edu_2	4.905	5.459	7.734	9.897	12.138
edo_civ2	7.016	6.891	6.913	6.565	6.388
edo_civ3	21.328	21.693	22.656	23.577	23.337
edo_civ4	19.064	19.591	21.27	22.378	23.472
edo_civ5	20.513	20.635	21.415	21.436	20.913
edo_civ6	22.696	22.893	23.879	25.395	24.375
jno_trab	12.918	13.199	13.477	12.798	13.754
prestac0	58.811	58.304	58.661	58.863	58.674
servmed	55.295	54.507	55.083	55.012	55.394
(Constante)	-140.324	-141.433	-144.965	-149.34	-153.502
Funciones discriminantes lineales de Fisher					

Cuadro 4.23 : Coeficientes de clasificación (Continuación)

A continuación se presenta el mapa territorial que es una representación del espacio correspondiente a cada uno de los grupos, en el plano definido en este caso por las dos primeras funciones discriminantes (Función 1: eje de abscisas, Función 2: eje de ordenadas).

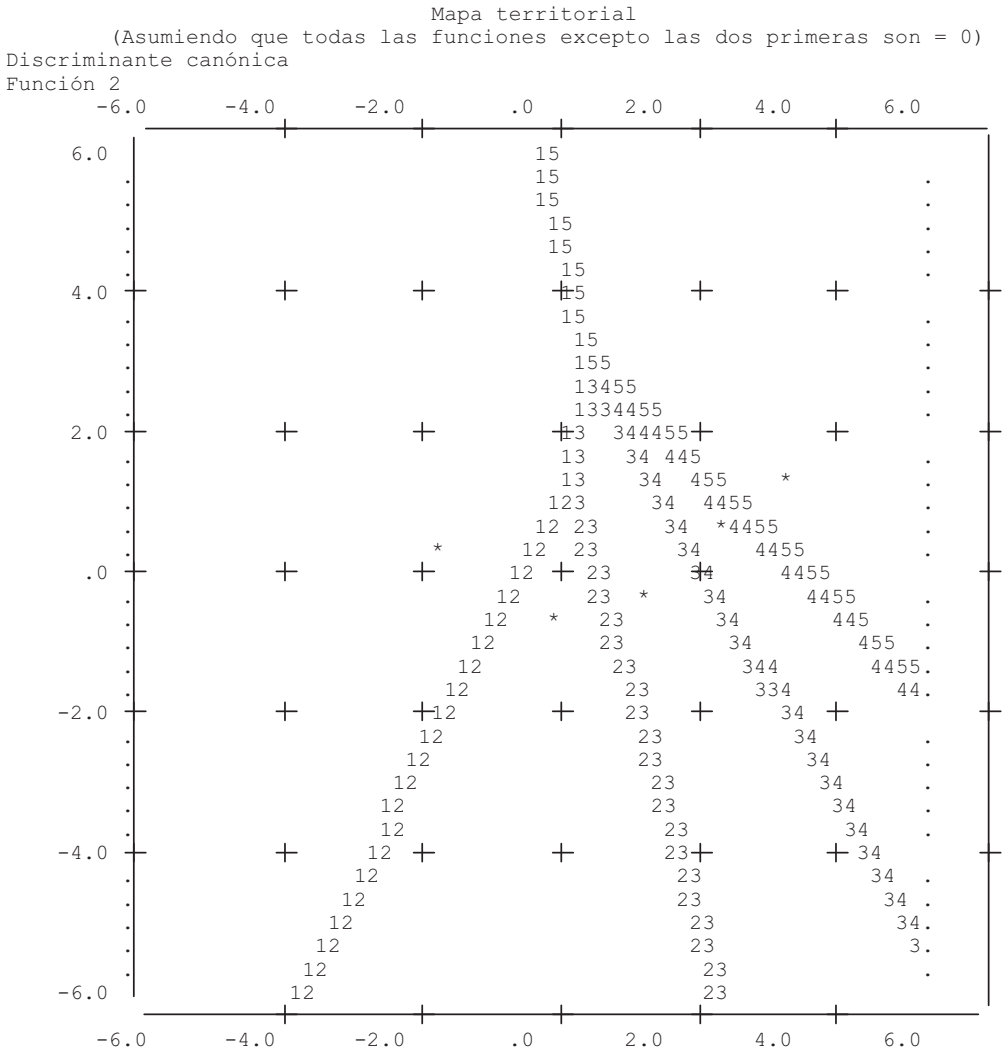


Figura 4.8: Mapa Territorial

Por medio de la figura 4.8 podemos observar también el solapamiento de los grupos, se observa como los centroides de los grupos se encuentran relativamente cercanos, además de que los territorios de los grupos 2, 3 y 4 quedan bastante reducidos, es decir que una ligera variación en las puntuaciones discriminantes puede traer consigo una mala clasificación de

los individuos, principalmente en estos grupos que presentan un mayor solapamiento. Como podemos observar, los grupos que se encuentran mejor diferenciados son el grupo 1 y el grupo 5, sin embargo, el mapa territorial muestra que existe una zona fronteriza entre estos dos grupos, lo que indica que no se encuentran completamente diferenciados.

Por ejemplo, al contar con los valores de todas las variables discriminantes para un hogar determinado, se podrían calcular las puntuaciones discriminantes de éste para las dos primeras funciones discriminantes (a partir de los coeficientes no estandarizados), tomando estas puntuaciones como coordenadas en el mapa territorial podríamos asignar este hogar al grupo correspondiente a la zona en la que se encuentre dicho hogar. Sin embargo no estaríamos tomando en cuenta las últimas dos funciones discriminantes que si bien son menos importantes resultaron significativas, además de que se presentarían problemas para los casos fronterizos.

La regla de clasificación que se empleará en la presente aplicación es la regla de máxima probabilidad de Bayes aplicada a las puntuaciones discriminantes, como se mencionó en el capítulo anterior (ver sec. 3.4.3 y 3.4.4). Al aplicar esta regla obtenemos los siguientes resultados:

Resultados de la clasificación(a)

		Grupo	Grupo de pertenencia pronosticado					Total
			1	2	3	4	5	
Original	Recuento	1	8027477	1797786	99869	3194	676	9929002
		2	774354	3194080	832787	198970	79048	5079239
		3	35750	744716	1219636	475684	313410	2789196
		4	607	39727	159016	309458	236257	745065
		5	0	21503	82382	189276	633369	926530
	%	1	80.8	18.1	1	0	0	100
		2	15.2	62.9	16.4	3.9	1.6	100
		3	1.3	26.7	43.7	17.1	11.2	100
		4	0.1	5.3	21.3	41.5	31.7	100
		5	0	2.3	8.9	20.4	68.4	100

a Clasificados correctamente el 68.7% de los casos agrupados originales.

Cuadro 4.24: Matriz de confusión

Como podemos observar en el cuadro 4.24, la tasa global de clasificación correcta asciende a 68.7% de los casos que no resulta una tasa muy elevada, sin embargo es interesante observar como en particular para el grupo1 se tiene una tasa de clasificación correcta considerablemente elevada (80.8%). De acuerdo con estos resultados, los grupos que presentan una mayor tasa de clasificación correcta son los grupos 1 y 5, resaltando el hecho de que para estos grupos la clasificación incorrecta se da en particular con el grupo más cercano y es casi nula para el grupo opuesto (0% de los originalmente pertenecientes al grupo1 son clasificados como

pertenecientes al grupo 5 y viceversa). Los grupos intermedios (2, 3 y 4) presentan una tasa de clasificación más baja, siendo el grupo 4 el que presenta un valor más bajo (41.5 %), esto debido al solapamiento de grupos principalmente.

A continuación se presentan los resultados obtenidos para dos casos particulares, ambos originalmente pertenecientes al grupo 1, estos resultados son los mismos que obtendríamos al tratar asignar un nuevo caso a alguno de los grupos preestablecidos.

Estadísticos por casos

Original	1	Grupo real			1
		Grupo mayor	Grupo pronosticado		1
			P(D>d | G=g)	p	0.156
				gl	4
			P(G=g | D=d)		0.889
			Distancia de Mahalanobis al cuadrado hasta el centroide		6.643
		Segundo grupo mayor	Grupo		2
			P(G=g | D=d)		0.102
			Distancia de Mahalanobis al cuadrado hasta el centroide		10.982
		Puntuaciones discriminantes	Función 1		-1.312
			Función 2		0.183
			Función 3		1.074
			Función 4		-2.352
	2	Grupo real			1
		Grupo mayor	Grupo pronosticado		2(**)
			P(D>d | G=g)	p	0.081
				gl	4
			P(G=g | D=d)		0.611
			Distancia de Mahalanobis al cuadrado hasta el centroide		8.3
		Segundo grupo mayor	Grupo		3
			P(G=g | D=d)		0.266
			Distancia de Mahalanobis al cuadrado hasta el centroide		9.961
		Puntuaciones discriminantes	Función 1		1.456
			Función 2		0.024
			Función 3		-2.608
			Función 4		-0.936

Para los datos originales, la distancia de Mahalanobis al cuadrado se basa en las funciones canónicas.
Para los datos validados mediante validación cruzada, la distancia de Mahalanobis al cuadrado se basa en las observaciones.
** Caso mal clasificado

Cuadro 4.25: Resultados de clasificación para dos casos particulares.

De acuerdo con los resultados mostrados en el cuadro anterior, el primer caso queda correctamente asignado al grupo 1, dado que al calcular la distancia de Mahalanobis al cuadrado respecto a los centroides de cada uno de los cinco grupos, el menor valor encontrado corresponde justamente al grupo 1 (6.643) seguido por el grupo 2 (10.982). Como podemos observar en el cuadro 4.25 para el primer caso, la probabilidad de encontrar casos más alejados del centroide dentro del grupo1 es de P(D>d | G=g) =1.56 y la probabilidad a posteriori de

pertenencia al grupo 1 para un caso con puntuaciones discriminantes como las observadas es de P(G=g | D=d) =.889, siendo la segunda mayor probabilidad la correspondiente al grupo 2, con un valor de 0.102. De esta manera, el primer caso queda correctamente clasificado en el grupo 1.

Podemos también observar el ejemplo del caso 2, que originalmente pertenece al grupo 1, pero queda incorrectamente clasificado en el grupo 2 debido que este se encuentra más cercano del centroide del grupo 2 e incluso del grupo 3 que del grupo 1 . La probabilidad de pertenencia al grupo 2 dadas las observaciones de este caso es de 0.611 y la segunda probabilidad mayor (correspondiente al grupo 3) es de 0.266, por lo que este caso queda mal clasificado en el grupo 2.

Para validar estos resultados, se realiza una segunda clasificación excluyendo de la obtención de las funciones discriminantes el caso que va a ser clasificado (método jacknife), los resultados obtenidos por medio de esta segunda clasificación se presentan a continuación:

Resultados de la clasificación(b,c)

		Grupo	Grupo de pertenencia pronosticado					Total
			1	2	3	4	5	
Validación cruzada(a)	Recuento	1	7968181	1857082	99869	3194	676	9929002
		2	805326	3106167	888172	200526	79048	5079239
		3	35750	822553	1103676	511180	316037	2789196
		4	607	62609	191509	202965	287375	745065
		5	0	21503	82883	314986	507158	926530
	%	1	80.3	18.7	1	0	0	100
		2	15.9	61.2	17.5	3.9	1.6	100
		3	1.3	29.5	39.6	18.3	11.3	100
		4	0.1	8.4	25.7	27.2	38.6	100
		5	0	2.3	8.9	34	54.7	100

a La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b Clasificados correctamente el 68.7% de los casos agrupados originales.

c Clasificados correctamente el 66.2% de los casos agrupados validados mediante validación cruzada.

Cuadro 4.26: Matriz de confusión para validación cruzada

Como podemos observar, la tasa de clasificación por medio de la validación cruzada disminuye un poco (pasa de 68.7% a 66.2%), podemos observar también que la tasa de clasificación se mantiene prácticamente igual para el grupo1, pero disminuye considerablemente para el resto de los grupos en particular para el grupo 4 cuya tasa de clasificación pasa de 41.5% a 27.2%. Con esto se comprueba que la bondad de los resultados no es demasiado buena debido al solapamiento de los grupos así como a la violación de ciertos supuestos, queda entonces por

comprobar que tanto aporta el análisis discriminante respecto a una clasificación realizada al azar. Como vimos en la sección 3.6.3, podemos medir la reducción del error de clasificación a través del estadístico

$$\tau = \frac{n_c - \sum_{i=1}^{g} \pi_i n_i}{n - \sum_{i=1}^{g} \pi_i n_i}$$

donde n_c es el número de casos correctamente clasificados, π_i es la probabilidad a priori de pertenencia al grupo i.

En efecto el valor de τ indicará la proporción en que se reduce el error de clasificación frente al error que se cometería por medio de una clasificación al azar, si consideramos que todos los grupos tienen la misma probabilidad ($\pi_i = 0.2$) y substituimos por el número casos correctamente clasificados (cuadro 4.24), obtenemos un valor de $\tau = 0.609$, lo que nos indica que pese a que la clasificación no es demasiado buena, se logra reducir los errores de clasificación en alrededor de 60.9%. Por ello, podríamos concluir que la regla de clasificación obtenida no es muy precisa (principalmente para los grupos intermedios), pero si aporta información importante para la clasificación de los individuos.

Si se incorporaran a la regla de clasificación las probabilidades a priori para cada grupo, basadas en los tamaños de muestra de éstos, obtendríamos los siguientes resultados:

Probabilidades a priori para los grupos

Grupo	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
1	0.51	7332	9929002
2	0.261	3411	5079239
3	0.143	1677	2789196
4	0.038	440	745065
5	0.048	425	926530
Total	**1**	**13285**	**19469032**

Resultados de la clasificación(a)

		Grupo	Grupo de pertenencia pronosticado					Total
			1	2	3	4	5	
Original	Recuento	1	8675906	1224973	27988	135	0	9929002
		2	1261063	3065667	631523	59650	61336	5079239
		3	96783	942362	1296733	211548	241770	2789196
		4	2009	100606	235166	177661	229623	745065
		5	0	29831	176978	105846	613875	926530
	%	1	87.4	12.3	0.3	0	0	100
		2	24.8	60.4	12.4	1.2	1.2	100
		3	3.5	33.8	46.5	7.6	8.7	100
		4	0.3	13.5	31.6	23.8	30.8	100
		5	0	3.2	19.1	11.4	66.3	100

a Clasificados correctamente el 71.0% de los casos agrupados originales.

Cuadro 4.27: Matriz de confusión con probabilidades a priori.

Como podemos observar en el cuadro 4.27, al incorporar las probabilidades a priori de los grupos la tasa de clasificación se incrementa ligeramente (pasa de 68.7% a 71%). Podemos ver que este aumento se debe al incremento en la tasa de clasificación correcta correspondiente al grupo 1 (pasa de 80.8% a 87.4%). En efecto la clasificación correcta en el grupo 1 se ve incrementada debido a que al ser el grupo con mayor tamaño de muestra, tiene una probabilidad a priori significativamente mayor a la de los demás grupos. Sin embargo, en lo referente a los cuatro grupos restantes podemos observar que las tasas de clasificación correcta disminuyen en todos los casos, siendo muy notorio el descenso en el grupo 4 (pasa de 41.5% a 23.8%). De esta manera, los tamaños desiguales en los tamaños de muestra, originados por la mala distribución del ingreso, provocan que al emplear las probabilidades a priori los casos tiendan a ser clasificados en el primer grupo, obteniendo una tasa de clasificación global relativamente elevada, pero con un incremento la clasificación incorrecta al interior de todos los grupos excepto el grupo1.

En general hemos visto que las diferentes reglas de clasificación tienden a clasificar de manera correcta a los individuos de los dos grupos extremos y no logran una muy buena clasificación en los grupos intermedios. Al querer clasificar nuevos individuos en alguno de los 5 grupos, el investigador deberá decidir si aplicar la regla de clasificación en la que todos los grupos tienen la misma probabilidad a priori, o bien incorporar al modelo probabilidades a priori basadas en los tamaños de muestra o algún otro criterio de acuerdo a sus necesidades e intereses.

Así mismo es posible, como se mencionó anteriormente, modificar el número de grupos y los umbrales que demarcan a estos últimos de acuerdo con los intereses que se tengan. Por ejemplo, podríamos reducir a tres el número de grupos de la siguiente manera:

- Grupo 1: Menos de 1 salario mínimo.
- Grupo 2: De 1 a 6 salarios mínimos.
- Grupo 3: Más de 6 salarios mínimos.

Al repetir el análisis discriminante para esta nueva categorización (considerando las mismas 85 variables), obtendríamos los siguientes resultados:

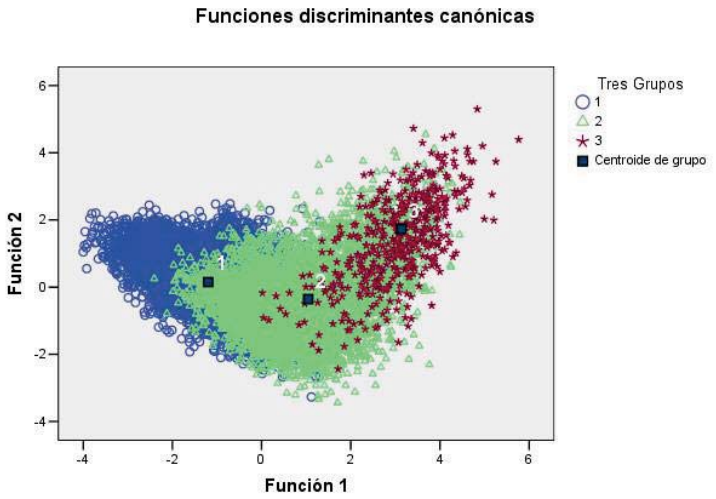


Figura 4.9: Diagrama de dispersión tres gurpos

Resultados de la clasificación(a)

		Grupo	Grupo de pertenencia pronosticado			Total
			1	2	3	
Original	Recuento	1	8626137	1301296	1569	9929002
		2	1260271	6140954	1212275	8613500
		3	0	125662	800868	926530
	%	1	86.9	13.1	0	100
		2	14.6	71.3	14.1	100
		3	0	13.6	86.4	100

a Clasificados correctamente el 80.0% de los casos agrupados originales.

Cuadro 4.28: Matriz de confusión para tres grupos

La Figura 4.9 muestra que al realizar el análisis discriminante con esta nueva categorización, los grupos aparecen mejor diferenciados que con la categorización original, sin embargo aún se observa un importante solapamiento de los grupos.Como podemos observar en el cuadro anterior, la tasa de clasificación global se incrementa considerablemente respecto al modelo con 5 grupos (pasa de 68.7% a 80%), y los tres grupos tienen tasas de clasificación correctas relativamente altas, siendo nuevamente el grupo intermedio el que presenta la menor tasa de clasificación correcta (71.3%).

Capítulo 5

Conclusiones

En el presente trabajo hemos presentado la teoría del análisis discriminante así como los resultados de la aplicación de éste sobre el nivel de ingreso de los hogares en la República Mexicana, con información estadística a nivel Nacional para el año 2002.

Como principales resultados del análisis discriminante descriptivo, podemos destacar que como era de esperarse, factores como la educación y el género del jefe de familia influyen de manera considerable en el nivel de ingreso de los hogares, así como la tenencia o carencia de ciertos bienes o servicios contribuyen a la diferenciación entre los distintos niveles de ingreso. Sin embargo, cabe destacar que no se cuenta con un factor que sea determinante a la hora de la diferenciación de los grupos correspondientes a los distintos niveles de ingreso, esto debido principalmente al comportamiento errático que presenta muchas veces el ingreso.

En cuanto a los resultados del análisis discriminante predictivo, hemos visto que la tasa de clasificación correcta alcanzada no es demasiado elevada, debido en parte al solapamiento de los grupos propuestos. Sin embargo, podemos concluir que el modelo aporta información significativa para la discriminación de los grupos. Como hemos visto en el capítulo anterior, el nivel de ingreso tiene una distribución desigual dentro de la población, lo que se ve reflejado en la mala diferenciación de los grupos.

En realidad existe una clara polarización entre el grupo con menor nivel de ingreso, que es a su vez el grupo que concentra la mayor parte de la muestra; y el grupo con mayor nivel de ingreso, grupo con una concentración de muestra significativamente menor. Los grupos con niveles de ingreso intermedios quedan entonces mal diferenciados. Esto es el claro reflejo de la inequidad que existe actualmente en México, pese a la implantación de programas sociales como un esfuerzo para combatir la pobreza y la desigualdad (por ejemplo: Oportunidades, Seguro Popular, Escuelas de Calidad etc.)

Al querer aplicar esta regla para determinar el nivel de ingreso en hogares en los que éste sea desconocido, habrá que considerar varios aspectos importantes como la temporalidad, ya que toda sociedad va sufriendo procesos evolutivos y las condiciones que prevalecían en el momento en el que fue captada la información por medio de la cual se generó la regla de clasificación presentada pueden cambiar con el paso del tiempo y perder su validez. Esperando que el cambio atenúe la polarización entre los grupos de mayor y menor ingreso y no la acentúe.

También habrá de tenerse en cuenta si el número de grupos propuestos así como sus umbrales son los más adecuados para los propósitos que se buscan. Además de que será necesario observar que el ingreso es una variable no fácil de tratar y que difícilmente se obtendrá una tasa de clasificación correcta muy elevada, por lo que será importante valorar los costos que traerían consigo los errores de clasificación cometidos.

Finalmente, podemos concluir que el análisis discriminante es una técnica estadística multivariada que puede ser de mucha utilidad en la búsqueda de la diferenciación de dos o más grupos de individuos u objetos, incluso cuando estos no estén perfectamente diferenciados y sean difíciles de tratar como es el caso de los grupos correspondientes a los distintos niveles de ingreso. Cabe señalar también que este tipo de problemas pueden ser tratados con otras técnicas estadísticas como son la regresión logística o los modelos multilogit.

Apéndice A

Metodología de la Encuesta

A continuación se presenta brevemente la metodología de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2002 (ENIGH 2002), empleada como principal fuente de información de este trabajo, con el fin de mostrar el marco conceptual de la encuesta.

A.1 Objetivos

Como se mencionó en el capítulo 4, el principal objetivo de la ENIGH 2002 es proporcionar información sobre la distribución, monto y estructura del ingreso y el gasto de los hogares, lo que permite, junto con las versiones anteriores de la encuesta evaluar en el tiempo los cambios en el nivel de vida de la población.

Algunos de los objetivos específicos de la ENIGH 2002 son la generación de información respecto a los siguientes rubros:

- La estructura del ingreso corriente de los hogares según la fuente de donde provenga.
- La estructura del gasto corriente de los hogares en la adquisición de bienes de consumo final (duraderos y no duraderos) así como las transferencias a otras unidades.
- Las características sociodemográficas de los miembros del hogar.
- La condición de actividad y las características ocupacionales de los miembros del hogar de 12 años y más.
- Las características de infraestructura de la vivienda y de equipamiento del hogar.

A.2 Metodología

A.2.1 Marco Conceptual

Dado que el ingreso determina la capacidad económica de los hogares para adquirir los bienes y servicios necesarios, éste resulta ser un condicionante principal en el nivel de bienestar de la población. Así, la ENIGH se basa en la consideración del que el tanto el monto del ingreso, su procedencia y su forma de distribución condicionan en gran medida el nivel de bienestar de la población. Para abordar el estudio de estos condicionantes, se seleccionaron como unidad de muestreo a la vivienda particular y como unidad de observación el hogar, considerando las siguientes definiciones de vivienda particular y de hogar:

- **Vivienda:**

Es el espacio delimitado por paredes, techos y pisos de cualquier material de construcción donde viven, duermen, preparan alimentos, comen y se protegen de las inclemencias del tiempo una o más personas. La entrada a la vivienda debe ser independiente, es decir, que sus ocupantes puedan entrar o salir de ella sin pasar por el interior de otra vivienda. Cabe aclarar que el (los) cuarto (s) de la vivienda cuyo acceso era independiente, si se rentaron o prestaron a otra (s) persona (s), que no eran miembros del hogar se consideró como otra vivienda.

Vivienda Particular

Es la vivienda regular de alojamiento de uno o más hogares.

Vivienda colectiva

Es la que se destina a servir de alojamiento habitual a personas sujetas a una subordinación de carácter administrativo u obligadas a cumplir normas de convivencia, en virtud de estar relacionadas por un objetivo público o interés personal común, como: razones de salud, disciplina, enseñanza, religión, trabajo, asistenciales, de alojamiento o militares. (Este tipo de viviendas no se consideran en la encuesta).

- **Hogar:**

Es el conjunto de personas unidas o no por lazos de parentesco que residen habitualmente en la vivienda y se sostienen de un gasto común para comer, una persona que vive sola o que no comparte gastos con otra(s) aunque viva en la misma vivienda también constituye un hogar. El criterio básico para la identificación de los miembros del hogar, es la existencia

de disposiciones básicas comunes para la vida doméstica, tales como compartir una provisión común de alimentos y como característica adicional la misma unidad de habitación.

Además de observar las variables relacionadas al ingreso (monto, procedencia, forma de distribución), será necesario observar las características sociodemográficas de los miembros del hogar, estrechamente relacionadas con los patrones de distribución del ingreso.

Será también importante precisar y conocer el marco conceptual referente a las transacciones económicas de ingresos y gastos realizadas por los miembros del hogar. Las transacciones económicas se clasifican en dos grandes rubros según su finalidad:

- **Transacciones económicas corrientes**

Las transacciones económicas corrientes son las que se realizan para cubrir las necesidades básicas y su resultado no es acumulable (como la compra de bienes de consumo final o el pago del alquiler de la vivienda). A su vez estas se dividen en:

•Ingreso Corriente Total

•Gasto Corriente Total

- **Transacciones económicas financieras y de Capital**

Se considera en este concepto de las transacciones económicas a las operaciones financieras y de capital ocurridas en el período de referencia, cuyo resultado es la modificación del acervo patrimonial de los hogares. Dentro de las transacciones financieras y de capital se incluyen también los actos cuyo propósito es el financiamiento para adquirir bienes y servicios de consumo final o bienes de capital (por ejemplo, el retiro de ahorros para comprar alimentos o los préstamos para comprar una casa). Las transacciones financieras y de capital, además, permiten explicar el déficit o superávit ente los ingresos y los gastos corrientes. A su vez se dividen en:

•Percepciones Financieras y de Capital Totales

•Erogaciones Financieras y de Capital Totales

Para los fines prácticos del presente trabajo, sólo se detallará brevemente el ingreso corriente total, el detalle del resto de las transacciones puede consultarse en el documento metodológico de la ENIGH 2002 [1]

[1] Encuesta Nacional de Ingreso y Gasto de los Hogares 2002, Documento Metodológico, INEGI 2002

Ingreso Corriente Total

Son las percepciones monetarias y no monetarias (en especie) que recibieron los miembros del hogar durante el período de referencia de la encuesta por su participación en el proceso productivo, por indemnizaciones y por transferencias corrientes sin contrapartida. Se registraron los ingresos netos que disponen los hogares para cubrir sus necesidades, es decir, después de descontar impuestos, cuotas a organizaciones laborales, a instituciones de seguridad social y deducciones similares. El ingreso corriente Total se clasifica en:

- **Ingreso Corriente Monetario**
- **Ingreso Corriente no Monetario**

A continuación se describen estos ingresos, principalmente el ingreso corriente monetario que es el ingreso considerado para la realización del presente trabajo.

Ingreso Corriente Monetario

Es la cantidad de dinero que recibe un perceptor miembro del hogar por su trabajo, por utilidades, por rendimientos e indemnizaciones y transferencias corrientes, de acuerdo a sus diferentes fuentes se clasifica en:

- **Remuneraciones al trabajo**

  Son los ingresos netos que las personas ocupadas miembros del hogar obtienen a cambio de la venta de su fuerza de trabajo a una empresa o patrón, con quien establecieron determinadas condiciones de trabajo mediante un contrato o acuerdo (verbal o escrito), incluye:

  - Sueldos, salarios y horas extras
  - Comisiones y propinas
  - Aguinaldos, gratificaciones y premios
  - Primas vacacionales
  - Reparto de utilidades

- **Renta Empresarial**

Son todas aquellas percepciones provenientes de un negocio propiedad de algún miembro del hogar o de una actividad productiva que se realiza en forma independiente o asociada.

Los ingresos provenientes de la renta empresarial, se clasifican de acuerdo a las características del negocio o a las actividades realizadas en:

- Ingresos por Negocios No Agropecuarios
- Ingresos por Negocios Agropecuarios
- Renta de la Propiedad
- Ingresos Netos de Cooperativas

- **Transferencias**

Percepciones monetarias que reciben los miembros de los hogares y que no constituyen el pago por trabajos realizados, incluye:

- Jubilaciones y pensiones
- Indemnizaciones de seguros contra riesgos a terceros
- Indemnización por despido y accidentes de trabajo
- Becas y donativos provenientes de instituciones
- Regalos y donativos originados dentro del país
- Regalos y donativos provenientes de otros países

- **Otros Ingresos**

Son todos aquellos ingresos monetarios no clasificados en los anteriores ingresos. Como por ejemplo: los ingresos provenientes de la venta de bienes muebles de segunda mano (automóviles, aparatos eléctricos, etc.). Siempre y cuando no sean de la actividad económica a la que se dedica alguno de los miembros del hogar, en cuyo caso se clasifican como ingreso provenientes de "Negocios Comerciales".

Ingreso Corriente No Monetario

Se refiere al valor estimado, a precios corrientes al consumidor, de los bienes y servicios para el consumo privado de los hogares.

Se clasifican en:

- **Autoconsumo**

  Valor estimado a precios corrientes al consumidor de los bienes y servicios producidos y consumidos por el propio hogar.

- **Pago en especie**

  Valor estimado a precios corrientes al consumidor de los bienes y servicios que proporcionan los patrones a sus obreros o empleados como pago a su trabajo.

- **Regalos**

  Valor estimado a precios corrientes al consumidor de bienes y servicios recibidos por los miembros de los hogares como regalo.

- **Estimación del alquiler de la vivienda**

  Es la estimación a precios corrientes, del alquiler de la casa habitación propia, la prestada por algún familiar o amigo y la vivienda recibida como prestación, a través de la empresa donde trabaja.

A.2.2 Captación de la Información

La ENIGH 2002, como ya se había mencionado, fue diseñada para poder presentar resultados a nivel nacional y para dos estratos urbano y rural, siendo una localidad urbana aquella que cuente con más de 2,500 habitantes y una localidad rural aquella con menos de 2,500 habitantes. El levantamiento de la encuesta se realizó en el periodo comprendido del 21 de Agosto al 15 de Noviembre de 2002, permitiendo por medio de visitas, la recolección de información en las viviendas seleccionadas. Para ello se utilizaron:

- Instrumentos de captación muy especializados que permitieron la operacionalización del marco de conceptos de la encuesta.

- Un equipo de entrevistadores, supervisores y jefes de área capacitados de manera especial sobre los procedimientos, lineamientos y criterios establecidos con base al marco de conceptos.

- Por otra parte se implementaron mecanismos de control para asegurar la calidad de la información.

A.3 Selección de la muestra

A.3.1 Diseño Muestral

Como habíamos mencionado en el capítulo 4 del presente trabajo, el diseño muestral de la ENIGH 2002 se caracteriza por ser probabilístico, lo que permite que los resultados arrojados por la encuesta puedan generalizarse a toda la población. El diseño de la muestra es también Polietápico, Estratificado y por conglomerados.

- **Probabilístico**.- Por que todas las unidades de muestreo tienen una probabilidad conocida y distinta de cero de ser seleccionadas.

- **Estratificado**.- Porque las unidades de muestreo con características similares de tipo geográficas y socioeconómicas se agrupan para formar estratos.

- **Polietápico**.- Porque la unidad última de selección (vivienda) es seleccionada después de varias etapas.

- **Por conglomerados**.- Porque previamente se conforman conjuntos de unidades muestrales de los cuales se obtiene la muestra.

A.3.2 Marco Muestral de la Encuesta

El marco muestral de propósitos múltiples del INEGI, constituido por la información demográfica y cartográfica obtenida por medio del Conteo de Población y Vivienda 1995, es el marco muestral que se utilizó para la realización de la ENIGH 2002.

Cada una de las 32 entidades federativas del país se divide en zonas que agrupan a las localidades de la siguiente manera:

Zona	Descripción
Urbano Alto	•Cuidades y áreas metropolitanas objeto de estudio de la Encuesta Nacional de Empleo Urbano (ENEU) •Resto de las ciudades de 100,000 y más habitantes y/o capitales de estado.
Complemento Urbano de Alta densidad	•Localidades de 20,000 a 99,999 habitantes •Localidades de 15,000 a 19,999 habitantes
Complemento Urbano de Baja densidad	•Localidades de 2,500 a 14,999 habitantes
Rural	•Localidades con menos de 2,500 habitantes

A.3.3 Formación de Unidades de Muestreo

1. Unidades Primarias de Muestreo (UPM)

Las UPM se constituyen por un Área Geoestadística Básica (AGEB), parte de un AGEB o varios AGEB colindantes, dependiendo de la zona de referencia, como se detalla a continuación.

UPM en urbano alto:

- Un AGEB con un mínimo de 480 viviendas.
- La unión de 2 o más AGEB contiguas del mismo estrato, con un mínimo de 480 viviendas en conjunto.

UPM en el resto de las zonas:

• Un AGEB o la unión de dos o más AGEB que contengan un mínimo de:

- 280 viviendas en localidades urbanas.
- 100 viviendas en localidades rurales

2. Unidad Secundaria de Muestreo (USM)

USM en urbano alto:

• La formación de la Unidad Secundaria de Muestreo se realiza únicamente en las ciudades ENEU. La USM o área de listado está conformada por la agrupación de viviendas bajo las siguientes condiciones:

- Puede estar formada por una manzana que tenga un mínimo de 40 viviendas habitadas.

- Puede estar formada por dos o más manzanas contiguas con al menos 40 viviendas habitadas.

USM en el resto de las zonas:

• En las zonas definidas como no ENEU, la unidad secundaria de muestreo está constituida por las viviendas particulares habitadas permanentemente o aptas para habitarse.

3. Unidad Terciaria de Muestreo (UTM)

Las unidades terciarias de muestreo se definen solamente en la zona denominada Ciudades ENEU y se conforman por las viviendas particulares, habitadas permanentemente o aptas para habitarse.

A.3.4 Tamaño de la Muestra

El tamaño de muestra está calculado para dar estimaciones a los siguientes niveles de desagregación:

- Nacional
- Localidades de 2500 y más habitantes
- Localidades de menos de 2500 habitantes

El tamaño de muestra para estos dominios se calcula con la proporción de perceptores de ingresos por vivienda, considerada como la variable principal de la encuesta y la que requiere los tamaños de muestra mayores, esto garantiza que las estimaciones del resto de las variables de interés queden cubiertas con este tamaño. Los tamaños de muestra obtenidos de esta manera son:

- Área Urbana: 14,539 viviendas en muestra
- Área Rural: 5,317 viviendas en muestra
- Total Nacional: 19,856 viviendas en muestra

Los detalles del cálculo de estos tamaños de muestra pueden ser consultados en el documento metodológico de la encuesta publicado por el INEGI.

A.3.5 Selección de la Muestra

La selección de la muestra en la ENIGH 2002, se realiza en forma independiente para cada entidad y estrato, el procedimiento varía dependiendo de la zona, los detalles de esta selección para cada una de las zonas puede consultarse en el documento metodológico de la encuesta.

Apéndice B

Distribuciones de Probabilidad

En este apéndice se resumen las distribuciones de probabilidad más usuales, se presentan sus funciones de densidad, sus parámetros, su media y su varianza .

B.1 Distribuciones Discretas

Nombre	**Función de Densidad**	**Parámetros**	**Media**	**Varianza**
Bernoulli	$p^x (1-p)^{1-x} I_{\{0,1\}}(x)$	$p \in]0,1[$	p	$p(1-p)$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x} I_{\{0,\ldots,n\}}(x)$	$p \in]0,1[$ $n = 1, 2, \ldots$	np	$np(1-p)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!} I_{\{0,1,\ldots\}}(x)$	$\lambda > 0$	λ	λ
Geométrica	$p (1-p)^x I_{\{0,1,\ldots\}}(x)$	$p \in]0,1[$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Binomial Negativa	$\binom{r+x-1}{x} p^r (1-p)^x I_{\{0,1,\ldots\}}(x)$	$p \in]0,1[$ $r > 0$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$

B.2 Distribuciones Continuas

Nombre	Función de densidad	Parámetros	Media	Varianza
Uniforme	$\frac{1}{b-a} I_{[a,b]}(x)$	$a, b \in \mathbb{R}$ $a < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial	$\lambda \exp(-\lambda x)\, I_{]0,\infty[}(x)$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}$ $\sigma > 0$	μ	σ^2
Normal Multivariada	$\frac{\exp\left(-\frac{1}{2}(X-\boldsymbol{\mu})' \sum^{-1}(X-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^p \lvert\sum\rvert}}$	$\sum_{p\times p}$ no singular, positiva definida	$\boldsymbol{\mu}$	$\sum$
Gamma	$\frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} I_{]0,\infty[}(x)$	$\lambda > 0$ $r > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Ji-Cuadrada	$\frac{1}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x} I_{]0,\infty[}(x)$	$k = 1, 2, ...$	k	$2k$
t	$\frac{\Gamma((k+1)/2)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{k\pi}} \frac{1}{\left(\frac{1+x^2}{k}\right)^{\frac{k+1}{2}}}$	$k > 0$	$\mu = 0$ para $k > 1$	$\frac{k}{k-2}$ para $k > 2$
F	$\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\,\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2}$ $\times \frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}} I_{(0,\infty)}(x)$	$m, n = 1, 2, ...$	$\frac{n}{n-2}$ para $n > 2$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ para $n > 4$

Referencias

[1984] Anderson T. "An Introduction to Multivariate Analysis", John Wiley and Sons.

[2001] Bardos Mireille, "Analyse Discriminante, Application au risque et scoring financier", Dunod

[1947] Bartlett, M.S., "Multivariate Analysis", Journal of the Royal Society, Series B, 1947, 9 , 176-197.

[1980] C. Chatefield A.J. Collins, "Introduction to Multivariate Analysis", Chapman and Hall.

[1984] Dillon, W.R. and Goldstein M., "Multivariate Analysis", John Wiley and Sons.

[2000] INEGI, "ENIGH 2000, Documento metodológico", INEGI.

[2001] Brian S. Everitt and Graham Dunn, "Applied Multivariate Data Analysis, Second Edition", Oxford University Press.

[1936] Fisher, R.A., "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 7, 178-188 .

[2001] Javier Gil Flores, Eduardo Garcia Jimenez, Gregorio Rodriguez Gomez, "Análisis Discriminante", La Muralla, Hespérides.

[1978] Dillon, W.R. and Goldstein M. "Discrte discriminant Analysis", John Wiley and Sons.

[1999] William H. Greene "Análisis Econométrico, 3^a Edición", Prentice Hall .

[1996] Santley I. Grossman, "Álgebra Linal 5° edición", Mc Graw Hill.

[1981] Hand D.J. "Discrimination and Clasification", John Wiley and Sons.

[1985] Harris, "A primer of multivariate statistics", Orlando Fl. Academic Press.

[1973] Kenneth Hoffman, Ray Kunze "Algebra Lineal", Prentice Hall.

[2000] Hosmer D.W., Lemeshow S. "Applied Logistic Regression", John Wiley and Sons.

[1994] Huberty, "Applied Discriminant Analysis" , John Wiley and Sons.

[1999] Johnson Dallas, "Metodos Multivariados aplicados al análisis de datos", Thomson international .

[1992] Johnson, R.A., Wichen, D.W. "Applied multivariate statistical analysis", Prentice Hall .

[1980] Kendall M "Multivariate Analysis, Second Edition", Charles Griffin & Company LTD London.

[1995] Krzanowski W.J. "Multivariate Analysis, Part 2", John Wiley and Sons.

[1990] Krzanowski W.J. "Pinciples of multivariate analysis. A user's perspective", Claredon Press, Oxford.

[1967] Lachenbruch, P.A. "An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis", Biometrics, 23, 639-645.

[1964] Lehman, C.H., "Álgebra", Limusa.

[1979] Maridia, K.V., Kent, J.T., Bibby, J.M. "Multivariate Analysis", Academic Press, London .

[1976] Donald F. Morrison "Multivariate Statistical Methods", McGraw Hill .

[1974] Mood, Graybill and Boes, "Introduction to the theory of statistics, third edition", Mc Graw Hill.

[2002] SEDESOL, Comite tecnico para la medicion de la pobreza, " Medición de la Pobreza: Variantes metodológicas y estimación preliminar", Serie de Documentos de Investigación N°1 SEDESOL.

[1966] Shatzoff, M., "Exact distributions of Wilk's likelihood ratio criterion and comparisons with competitive tests.", Harvard University.

[1986] Stevens, J, "Applied multivariate statistics for the social sciences", N.J., Hillsadale: Lawrence Eralbaum Associates, Pub.

[1971] Tatsuoka Maurice, " Multivariate Analysis: Techniques for educational and psychological research", John Wiley and Sons.