19

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

UNA APLICACIÓN DEL ANÁLISIS DISCRIMINANTE EN LAS CIENCIAS SOCIALES

T E S I S

Que para obtener el título de :

A C T U A R I A

P r e s e n t a :

MELBA BEATRIZ CASELLAS ARGÁEZ

DIRECTORA DE TESIS:
Mat. Margarita Elvira Chávez Cano

2001

VNIVERSIDAD NACIONAL
AVFNOMA DE
MEXICO

México D.F., a 10 de Septiembre de 2001

M. EN C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Una Aplicación del Análisis Discriminante en las Ciencias Sociales"

realizado por Melba Beatriz Casellas Argáez

con número de cuenta 8455149-0 , pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario Mat. Margarita Elvira Chávez Cano

Propietario Act. Jaime Vázquez Alamilla

Propietario M. en D. Alejandro Mina Valdés

Suplente Act. José Guadalupe Vázquez Vázquez

Suplente Mat. Hugo Villaseñor Hernández

Consejo Departamental de Matemáticas

M. en C. José Antonio Flores Díaz

Dedicatorias...

A Melba y Miguel.

Ese momento compartido algún mayo,
me permite andar por aquí.

A Miguel

A Luz

A Violeta, Queta y Alan

A Ileanita

A La Abuelita...

Agradecimientos

En todo libro, tesis e incluso documento, especialmente aquellos que se refieren a algún aspecto de investigación, no falta la trillada sección de los agradecimientos en la que aparecen un montón de nombres, por lo general, perfectos desconocidos para aquellos que en algún momento leerán el trabajo. Y lo que sucede es que es una sección ineludible, no porque el autor desee cumplir una formalidad, sino porque la gratitud siempre es importante, es un alimento para el que la ofrece y quien la recibe. Además, después de terminado el trabajo surge la necesidad de no dejar de lados a aquellas personas sin las cuales éste hubiera sido imposible realizar. La guía, comentario, consejo, paciencia, amor de mucha gente, se convierte en parte importante del trabajo.

Alguien tal vez leerá lo que se escribe, y el escrito está pensado para que llegue a ese alguien; pero la sección de los agradecimientos es un espacio muy personal de quien escribe, y de algún modo, también es un espacio que se desea compartir con el lector. Yo no deseo pasar por alto el agradecimiento que siento necesidad de expresar. Así, quiero agradecer

a Celina, por la voluntad puesta en el apoyo y la guía que me ha brindado, antes y durante este proceso de preparación, creación de la tesis y camino de trámites;

a Miros, quien tal vez sin saberlo, me dio las pistas para encontrar algo perdido hace mucho tiempo;

a la Mat. Margarita Chávez que pese a su agenda algo apretada aceptó ser la directora de este trabajo;

al Profesor Alejandro Mina por ese apoyo tan importante en el momento clave;

al Act. José Vázquez quién ha tenido la voluntad para un acercamiento muchos años después;

al Act. Jaime Vázquez por regalarme parte de su tiempo y permitir que las cosas fluyan;

al Mat. Hugo Villaseñor por su siempre amable atención;

a la UNAM por todas las oportunidades que me ha dado una y otra vez.

ÍNDICE

INTRODUCCIÓN

Este trabajo se presenta para obtener el título de la carrera de actuaría. Una carrera cuya formación se sustenta en las matemáticas, y como una parte relevante dentro de éstas, la estadística. Es por eso, que el objetivo de este trabajo de tesis es mostrar como el uso de las matemáticas puede ser de mucha utilidad en la solución, o al menos el entendimiento, de diversos problemas. En este caso se eligió el área de las ciencias sociales, específicamente el tema de la pobreza. El objetivo, entonces, es presentar el análisis discriminante, que es una técnica de lo que se conoce como análisis estadístico multivariado, y realizar una aplicación de éste sobre un problema concreto y real.

A nivel mundial el problema de la pobreza ocupa, en estos momentos, una posición de primer orden dentro de los retos por superar, que requiere atención inmediata, tanto por parte de los gobiernos, como de la sociedad. Basta darse una vuelta por las páginas web de los principales organismos internacionales, tales como el Banco Interamericano de Desarrollo (BID), el Banco Mundial (BM), la Consejo Económico para América Latina (CEPAL), para ver que gran parte de los trabajos actuales versan sobre la pobreza o por lo menos tratan algún aspecto particular de ella.

Esto refleja una preocupación real sobre el asunto, y es que, contrario a lo que en algún momento se pensó en cuanto a que los países desarrollados jalarían a los países en vías de desarrollo, para que estos últimos, pudieran seguir sus pasos y lograr su propio desarrollo, lo que ha ocurrido en las últimas décadas del siglo XX es que la brecha entre unos y otros se ha incrementado. Las causas son diversas, pero lo cierto es, que para enfrentar con éxito el problema de la pobreza, en estos momentos, es necesario que las causas que han propiciado estas desigualdades, sean entendidas; tanto por los países que directamente viven la pobreza, como por los países desarrollados.

La pobreza tiene muchas aristas, tal vez la más fácil de reconocer se relaciona con el aspecto de la percepción de un ingreso, pero ésta no es la única. Existe pobreza donde hay hacinamiento; donde hay niños, que en la infancia, tienen que salir en busca de dinero para vivir; existe pobreza donde no hay un techo que proteja y un hogar que cobije. La pobreza no sólo está circunscrita a lo material, también hay pobreza cuando un individuo carece de las habilidades necesarias para relacionarse provechosamente con su entorno, algo que ayuda a perpetuar esa pobreza material.

Un primer paso en la implementación de programas y políticas para el combate de la pobreza es contar con indicadores que apoyen en la identificación de las carencias dentro de una población. No existe una forma única de medir la pobreza, de hecho, la metodología que cada país decide emplear en gran medida está condicionada a la información con que cuenta o la que le sea factible generar. Tampoco existe una definición única exhaustiva y suficiente para definir la pobreza, el contexto mismo es relativo en el tiempo y el espacio de cada sociedad; sin embargo existen algunos conceptos generales para los que hay un consenso, como lo es la tesis de que la pobreza es un asunto de varias dimensiones, es decir que no está ligado a un solo factor sino a un número variado de factores que en momentos o permanentemente se relacionan entre sí en distintos planos.

Tal vez esta sea la razón por la cual no existe acuerdo en cuanto al porcentaje de población en pobreza en nuestro país, las cifras oficiales hablan de un 40% de población en pobreza y un 25% en pobreza extrema; algunos académicos dicen que estamos sobre el 75% de población en pobreza. Esta discusión se encuentra en boga en estos momentos. Lo cierto es que sea un 25 o un 75 por ciento, estamos hablando de al menos 25 millones de seres humanos que viven en condiciones de carencias y falta de oportunidades. No se puede negar que esto es un fuerte problema social y económico.

Pero no basta saber cuántas son las personas pobres del país o del mundo, para tratar de entender la pobreza, también se necesita conocer sus características, si éstas presentan o no diferencias de una región a otra. Seguramente la pobreza de las personas que viven en Chiapas es distinta a la de aquellas que están en Siberia o Sonora, pero qué sucede si hablamos de una región específica, ¿es igual la pobreza para todos los lugareños? Si no es así, ¿existen similitudes?, ¿cuáles son las diferencias? Contestar estas preguntas, y otras más, es tarea en la que se tienen que involucrar profesionales de distintas disciplinas que puedan abarcar las diferentes aristas que esto presenta.

Se puede comenzar buscando elementos que permitan asegura la existencia de diferencias entre las personas que viven en condiciones de pobreza. Con esto, las preguntas planteadas, y algunas más, no quedarán resueltas, pero sí nos aportarían una seguridad para movernos a mayores investigaciones con la esperanza, bien fundada, de no estar perdiendo el tiempo, sino, por el contrario, estar avanzando en nuestros conocimientos. Algo que puede resultar benéfico.

Este trabajo de tesis busca probar, con información concreta, que dentro de la población que vive en condiciones de pobreza, se pueden diferenciar dos grupos: el de la población en 'pobreza extrema' y el de la población en 'pobreza'; lo que daría sentido al hablar de estas dos poblaciones y

nos aseguraría que no es pura retórica los planteamientos donde se menciona esta distinción. Se estaría hablando de algo real.

En un estudio como este, se requiere contar con información sobre diversos aspectos de los individuos pobres. Información que contemple, sino todos, al menos algunos de los diferentes factores que inciden en la pobreza, su generación y su permanencia. Por otro lado, requerimos una herramienta adecuada para el manejo simultaneo de un conjunto amplio de datos, a cual sea confiable. Esto último, se solventa con el análisis estadístico multivariado, que permite incorporar tantas variables como consideremos adecuado a nuestro estudio.

De forma breve (ya más adelante se irán presentando cada punto con calma) se mencionarán los elementos utilizados para realizar este trabajo:

En cuanto a la información, se decidió trabajar con información oficial que el Programa de Educación, Salud y Alimentación (Progresa) genera para sus estudios sobre pobreza.

El ejercicio se lleva a cabo sobre localidades del estado de Yucatán.

Se utilizan técnicas de análisis estadístico multivariado para analizar la información. Específicamente el análisis discriminante que es una técnica adecuada para separa poblaciones y clasificar observaciones. Nuestra tarea versa en este sentido, ya que deseamos comprobar la existencia de dos poblaciones dentro de un cierto universo definido como hogares en condiciones de pobreza.

El paquete estadístico SPSS Ver.9 es la herramienta informática para el manejo de las bases de datos que contienen la información que sirve de insumo. Este paquete brinda una amplia gama de opciones para el análisis de datos, así como la aplicación de diversos métodos estadísticos.

La manera como está estructurada la tesis es en seis capítulos. En el capítulo uno se habla sobre la pobreza y se abordan algunas de las definiciones y enfoques que se tiene sobre el tema, además, se presenta algo de la historia de los programas de combate a la pobreza que se han sido desarrollados en nuestro país.

Para poder abordar la teoría de del análisis multivariado es necesario el conocimiento del álgebra de vectores y matrices, lo que permite expresar y manejar la información de manera tal que los conceptos del análisis multivariado también quedan plasmados y pueden ser aplicados. Personalmente, creo que la mayor dificultad para entender la teoría multivariada, radica en la notación, por lo que junto con el álgebra de vectores y matrices, en el capítulo dos, también se define la notación con la que serán expresados los datos y conceptos estadísticos.

El tercer capítulo presenta los resultados del análisis multivariado que se requieren para el análisis discriminante: la función normal multivariada así como algunas pruebas estadísticas multivariadas sobre las medias de dos y más poblaciones.

En el cuarto capítulo se presentará la teoría del análisis discriminante que será la metodología para abordar el tema que nos ocupa. Primero se dará una explicación general y posteriormente el desarrollo teórico donde usaremos lo ya visto en los capítulos anteriores

En el quinto capítulo se presentará la aplicación de esta técnica. Comenzando con la explicación de la construcción de las variables y algunas discusiones al respecto, para inmediatamente pasar a la aplicación y los resultados obtenidos.

Los resultados finales y las conclusiones se presentan en el sexto y último de los capítulos, así como algunos comentarios finales.

Capitulo I

EL PROBLEMA DE LA POBREZA

Pobreza... ese fantasma que arrastramos los ahora llamados "países emergentes", es hoy un lastre doloroso que como país nos dificulta la posibilidad de alcanzar un desarrollo sólido y equitativo para toda la población. Este pasivo impide capitalizar en toda su dimensión los importantes logros del crecimiento en nuestro producto interno bruto, mas aún cuando este crecimiento se concentra en un pequeño porcentaje de la población.

¿Qué sucede con esa población que se encuentra en situación de pobreza y que nunca llega a ver los beneficios de un crecimiento en el PIB? El gobierno se ve en la necesidad de implementar programas que permitan aliviar, en cierta medida, su situación de urgente atención; con lo que gran parte del gasto público tiene que enfocarse en programas sociales de apoyo, utilizando presupuesto que podría ser empleado en proyectos productivos con miras a crear las condiciones necesarias para que esta población genere su propia riqueza, y así, mejorar día a día el desarrollo del país.

La constitución establece que el gobierno tiene la obligación de velar por el bienestar de los ciudadanos. La pobreza obliga a éste a tomar medidas para su combate, pero si no se generan soluciones integrales, el dinero invertido en mitigar la pobreza será un dinero que aunque ciertamente logrará alivio temporal en esta población, no será capaz de producir, de construir y generar riqueza.

México tiene una larga tradición en programas sociales a lo largo de su historia reciente, que se comienzan a implementar después de nuestra revolución de principios del siglo pasado y como vía para lograr el objetivo de justicia social para un gran segmento de la población, que antes de ese momento histórico, no era considerado, cabalmente, dentro de nuestro sistema de gobierno. Aún así, los problemas de pobreza siguen afectando a una gran parte de la población. De hecho,

en los últimos años ésta se ha incrementado, en parte como consecuencia de las crisis económicas recurrentes de las últimas tres décadas.

Por el enfoque que tienen las instituciones para atender los grandes rezagos sociales, se puede hablar de etapas en la historia de los programas sociales de este país. La primera se da durante los primeros años posteriores al término de la revolución, a lo largo de los años 30, cuando se dieron reformas sociales de amplio alcance como la agraria y el reparto de la tierra en el campo o la organización obrera a través del sindicalismo, en las ciudades. En los años 40 se da otra etapa de las políticas sociales, con la creación de las grandes instituciones para el mejoramiento de la educación y la salud a nivel nacional. Es la época en que nace el IMSS, la secretaría de salud y la secretaría de educación.

Podemos hablar de una tercera etapa que se presenta a finales de los años 60 cuando, pese a los grades avances en materia de atención al gran universo de la población, no se ha logrado abatir la pobreza, y entonces surge el enfoque de atacar situaciones específicas de manera más dirigida. Se crean así, los primeros programas de acción específica, que son paralelos a los programas de gran alcance ya mencionados. El PIDER fue el primero de ellos.

Ya en la década de los 70's surge COPLAMAR; SAM y SOLIDARIDAD, en los 80 y durante la última década del siglo, Progresa (Rolando Cordera 1999[1])

Estos esfuerzos han logrado objetivos muy importantes, basta ver el avance alcanzado en materia de salud y de educación al pasar de un 77.7% de analfabetismo a principios del siglo XX a niveles de 10.6% al cierre del mismo. Sin embargo, no se puede negar que aún se tiene que trabajar para que además de los logros en cobertura de estos servicios, se alcance también, la calidad.

1.1 Algunas discusiones sobre pobreza

Existen varias definiciones sobre pobreza presentadas desde perspectivas y enfoques distintos, sin embargo, todas coinciden en un aspecto: la pobreza está relacionada con la satisfacción de necesidades. La diferencia radica en lo que se entiende por necesidades.

Un enfoque plantea el concepto de *necesidades* como 'lo como mínimo que requiere satisfacer un individuo para su subsistencia'. Entonces, se propone cierta canasta básica y con

[1] Alivio a la Pobreza. Análisis del Programa de Educación, Salud y Alimentación dentro de la política social.

base en su costo se define el nivel de ingreso mínimo que un individuo debe tener para cubrir dicha canasta. La variable utilizada para medir esto puede ser el gasto o el ingreso. Sin embargo, este enfoque no considera, por ejemplo, que la capacidad de las personas para transformar el ingreso en bienestar es variable, que es algo que está relacionado con la formación y educación de cada individuo. Por nuestra parte, nos inclinaremos más por un enfoque que considere de manera más amplia el concepto de necesidades.

Pensando que la pobreza se puede medir de acuerdo con la satisfacción de las necesidades básicas de un individuo o una familia, requerimos identificar cuáles son esas necesidades básicas. Al llegar a este punto, nos topamos con la discusión sobre qué es una necesidad básica. De forma general podemos decir que son dos las visiones que se tienen al respecto. Una plantea que no existen necesidades básicas generales, pues toda necesidad es cultural y la necesidad de ciertos bienes depende del significado social que la persona le dé. La otra dice que existen necesidades básicas para cualquier individuo independientemente del contexto cultural.

Los que plantean la primera, consideran que es posible hablar de necesidades básicas aplicables a cualquier ser humano independientemente de su historia y cultura; mientras que los defensores de la segunda plantean que las necesidades son relativas a las circunstancias históricas y culturales de los individuos.

Sin embargo es importante que más allá de disertaciones filosóficas, seamos capaces de aterrizar los elementos que tenemos en una aplicación práctica, evitando caer en discusiones interminables.

En este trabajo se quiere validar que la existencia de lo que se denomina pobreza extrema, y de ser posible, lo que implica la existencia de al menos dos grupos de individuos dentro de aquella población que se considera pobre. Se plantea entonces la necesidad de considerar en la medición, indicadores que puedan ser aplicados en cualquier contexto social y cultural. Algo que sea universal.

Por otro lado, se tiene un interés especial en llevar a cabo este ejercicio con base en datos concretos e ilustrar lo importante que es la estadística para entender los fenómenos sociales. Esto requiere el planteamiento de ciertos conceptos para que, con base en ellos, se identifiquen los factores que pueden ser incluidos en el ejercicio mismos que cumplan con esta idea de universalidad, planteada anteriormente.

Partiendo de esto, podemos considerar que en el caso de la pobreza, una base universal plantea: que el daño provocado por la ausencia de un bien determinado es igual para cualquier individuo.[2]

La satisfacción de necesidades básicas es lo que permite a los individuos la participación, tan activa como sea posible, en formas de vida que puedan elegir si tuvieran la oportunidad de hacerlo. Este planteamiento puede desglosase para identificar los distintos factores que hacen que el individuo esté en condición de realizar y llevar a cabo su elección. Aún de manera general, se puede decir que son necesidades básicas la salud y la autonomía personal. Esta última se puede entender basándonos en lo siguiente:

1. La comprensión que las personas tienen de sí mismas, de su cultura y de lo que esperan de ellas.
2. La capacidad psicológica del individuo para crear sus propias opciones.
3. Oportunidades objetivas que permiten que una persona actúe o deje de actuar.

Para alcanzar lo anterior es indispensable el factor educación, a través del cual se logra la libertad, y ésta da la posibilidad de tener mayores oportunidades de elección.

Así, salud y educación son dos de los factores que influyen en el desarrollo de las capacidades de los individuos.

Veamos lo que el Banco Mundial presenta en un documento llamado 'Datos y medición de la pobreza'[3], donde se presentan las recomendaciones de este organismo. En él se dan los resultados de un compendio de estudios realizados por 23 países que sufren condiciones de pobreza. Este documento aplica como definición de pobreza lo siguiente: "**la pobreza es carencia material, relaciones sociales malas, inseguridad y precariedad, poca confianza en uno mismo e impotencia**", y determina que existen cuatro dimensiones de la pobreza:

La pobreza de ingreso que se caracteriza por un grupo demográfico cuyos ingresos personales o consumo son inferiores a lo establecido por una línea de pobreza.

La pobreza en seguridad afecta a un grupo demográfico que se enfrenta a riesgos particularmente elevados como pueden ser, vivir en zonas de alto riesgo de desastres naturales debido al deterioro del medio ambiente o riesgos relacionados con la seguridad del individuo por situaciones de violencia o persecución política.

---

[2] Paulette Dieterlen, citando a Len Doyal en 'Alivio a la pobreza. Análisis del Programa de Educación Salud y Alimentación'

[3] Guía para la medición de la pobreza, las políticas a aplicar y métodos de evaluación para medir el impacto de las mismas

La pobreza en la educación para lo que se puede definir indicadores basados en el número de años cursados.

La pobreza en la salud que se caracteriza por un grupo demográfico con expectativas de salud por debajo de una línea establecida, misma que puede definirse de diversas maneras tales como hogares con niños desnutridos o la incidencia de una, varias enfermedades definidas para este fin, etcétera.

A esto podemos agregar la pobreza en la calidad de vida de las personas que se puede referir a situaciones de hacinamiento elevado de las familias en su vivienda, así como la posibilidad de tener espacios recreativos.

Como se puede ver el planteamiento del Banco Mundial lleva implícito mucho de los conceptos que se discutieron con anterioridad.

Se puede resumir el concepto de pobreza considerando todo lo planteado de la siguiente manera:

> "El concepto de pobreza es relativo al tiempo y al contexto de cada sociedad. Es una condición que limita las posibilidades de desarrollo de las capacidades básicas de las personas, debido a la falta de herramientas que permitan su desarrollo, lo que impide una participación plena en la sociedad. Así, el esfuerzo que se tiene que emplear para el logro de los satisfactores básicos no da los resultados que corresponderían al esfuerzo realizado".

Pero más allá de las definiciones, lo cierto es que este es un problema de primer orden en el mundo y principalmente porque afecta a las personas; les genera sufrimiento, sea este entendido como tal o no por ellos; les coarta sus posibilidades de satisfactores, además que afecta a aquellos que no la padecen directamente, pues invade toda la realidad de un país, se implanta dentro de la cultura, afecta la economía y puede llegar a perturbar la paz social.

No hay que perder de vista que la pobreza crea un círculo vicioso que ayuda a su reproducción generación tras generación. En muchas ocasiones es la única forma de vida que las personas conocen.

Partamos del concepto de que pobreza es carencia de oportunidades, de aquí se desprende una serie de carencias en la vida de los individuos que nacen en un entorno de pobreza, mismas que se van encadenando a lo largo de su vivir y que afectan su desarrollo personal.

Esto empieza desde el momento mismo de la gestación. Una madre mal alimentada, da a luz bebé con problemas de peso y talla, desde el momento mismo de nacer. Condición que no mejora

con el paso del tiempo. Ya en la edad escolar, pese a que en teoría, todos los mexicanos tenemos igualdad de oportunidades para accesar a la educación gratuita, en la práctica, no todos tenemos la misma oportunidad de hacerlo. La falta de oportunidades para una nutrición adecuada, para tener acceso a una escuela cercana con calidad educativa, para unos padres motivadores, son apenas algunos de los factores que van tejiendo las condiciones que perpetúan la condición de pobreza.

De acuerdo con la información oficial, en el país hay 26 millones de personas en pobreza extrema, sin embargo se dice que existe también aquella población que se encuentra en pobreza, aunque ésta no sea extrema. Podemos pensar entonces, que entre estas dos poblaciones existen factores que inciden para marcar una diferencia.

El objetivo de este trabajo versa sobre las familias que viven en las localidades marginadas de este país, las cuales por el simple hecho de habitar en estas localidades, suponemos que presentan condiciones de pobreza. Un primer objetivo es comprobar si hay fundamento suficiente para hacer la distinción entre 'pobreza' y 'pobreza extrema'. Es decir, se quiere valorar si es posible hacer una separación de estas familia en dos grupos, basándonos en información socioeconómica de éstas.

Si nos basamos únicamente en el nivel de los ingresos familiares y tomamos el costo de una canasta básica, encontraremos que efectivamente hay familias que no alcanzan a cubrir esta canasta, pero no llegaríamos mas que a identificar un solo grupo de familias que viven en esta condición. Por otra parte, el ingreso es un dato difícil de captar en las encuestas, ya que en muchos caso, la gente tiende a no dar información precisa. Es por eso que nuestro interés radica en probar si es posible hacer esta distinción con base en un conjunto de datos más amplio que tome en cuenta, también, otro tipo de características de las familias.

Así, esta falta de oportunidades deriva en la imposibilidad de satisfacer las necesidades; y la no satisfacción de éstas, evita que el individuo se pueda proveer de las herramientas que le ayuden a ser capaz de generar sus propias oportunidades.

Si esto es posible, el objetivo siguiente es identificar en dónde reside esta diferencia. Lo ideal sería identificar tanto las características manifiestas —lo que es posible ver—, así como aquellas que no son tan claras pero que inciden en favor de una diferenciación de las familias en pobreza. Sin embargo, esto último posiblemente resulte más complicado identificar, pues seguramente, estaríamos tocando terrenos que involucran factores cualitativos y la propia fuente de información tiene pocos datos al respecto, ya que, mas bien está enfocada a los factores cualitativos.

Pese a esto, el análisis de los resultados puede aportar alguna luz sobre aquellas características que sin ser manifiestas, pueden ser supuestas. Y tal vez podríamos llegar hasta el punto de sugerir el tipo de información que sería útil recolectar para probar su importancia.

Para verificar la existencia de estas dos poblaciones, se utilizará análisis discriminante, que es un método estadístico multivariado que aporta dos resultados, uno es la clasificación de las observaciones en un cierto número de grupos, para lo cual el método crea una función que separa las observaciones de manera óptima. Teniendo definida la función y con ella una regla de asignación, se obtiene el instrumento para clasificar nuevas observaciones en uno de los grupos.

El otro resultado es la identificación de los aspectos que mejor contribuyen a la clasificación, es decir, la función aporta información sobre las características que mejor separan (inciden más) a los grupos. Para este trabajo se están definieron dos grupos, el de las familias en pobreza y el de las familias en pobreza extrema. El análisis discriminante nos proporcionará los elementos para concluir si existe o no diferencias significativas entre ellos que vayan más allá del ingreso, además de indicar, en caso de que existan diferencian, cuáles son los aspectos que contribuyen más a crearlas.

Con base en lo discutido sobre la pobreza en este capítulo, y de acuerdo a los elementos que aportaron las fuentes de información, se identificaron los aspectos de interés para ser tomados en cuenta en el análisis. Cabe señalar, que fue necesario crear variables a partir de la información proporcionada por las fuentes, así como realizar algunas inspecciones sobre las variables para eliminar la existencia de información incongruente o extremadamente distinta a la demás información. Todo esto se explica con detalle en el capítulo V donde se presenta la aplicación.

Capítulo II

ÁLGEBRA DE MATRICES Y MANEJO DE DATOS MULTIVARIADOS

En las poblaciones univariadas (de una variable), en general, es posible caracterizar completamente la función de distribución a partir de dos parámetros, la media y la varianza. Sin embargo, en el caso multivariado (múltiples variables) donde se estudia una población para la cual se han medido p características (variables), se dispondrá de p medias, p varianzas y $\sum_{i=1}^{p-1} i$ covarianzas, que deben ser estimadas e interpretadas. El álgebra de vectores y matrices es la herramienta que permite el manejo algebraico de muchas variables, a la vez que puede expresar matemáticamente los conceptos del análisis multivariado.

Con las computadoras de la actualidad el cálculo de los estimadores no es problema, pero con la interpretación hay que ser cuidadoso, pues en ocasiones ante la perspectiva de contar con gran cantidad de información se puede caer en justificaciones que no son del todo válidas. Cada situación requiere una evaluación particular. En ocasiones no es necesario estimar todos los parámetros, ya que puede haber variables que no se incluyan en el modelo por alguna razón en particular. Un ejemplo es cuando existe redundancia en las variables.

Dentro del análisis multivariado muchos de los conceptos son extensión de los planteados en la estadística univariada, sólo que involucrando un mayor número de variables. Al igual que en el caso univariado, existen algunos estadísticos que permiten caracterizar a una población que son extensión del caso univariado, más no todos.

Los métodos estadísticos de análisis multivariado pueden agruparse en dos conjuntos: aquellos que permiten extraer información sobre la interdependencia de las variables que caracterizan a un individuo, y aquellos que permiten extraer información acerca de la dependencia entre una o varias variables, con otra u otras variables.

Para el desarrollo de la teoría multivariada, la herramienta básica para el manejo simultaneo de más de dos variables es el álgebra lineal. Es por eso que un adecuado manejo del álgebra de vectores y matrices es muy importante para poder comprender la teoría. De hecho, tal vez el aspecto más complicado dentro del análisis multivariado sea la notación, que en ocasiones se torna sobrecargada y confusa ante el manejo de las variables.

Considerando lo anterior, vale la pena hacer una escala y presentar algunos de los resultados del álgebra lineal que más útiles son en la estadística multivariada, así como algunos conceptos que son requisito manejar tanto conceptual como algebraicamente. En cuanto a la notación, es importante tener un claro entendimiento de ella para poder comprender sin equívocos los conceptos que expresan. Antes de abordar la teoría multivariada, que se aplica en el trabajo, vamos a establecer, con la mayor claridad posible, la manera como se expresan los vectores, las matrices y los escalares haciendo la liga de lo que representan dentro de la estadística multivariada. Así quedará establecida, de manera precisa, la notación que será utilizada a lo largo del trabajo.

2.1 Álgebra de vectores y matrices

Considerando, como ya se ha mencionado, la importancia que tiene el álgebra matricial para el desarrollo de la estadística multivariada, este apartado se dedica a dar un recorrido por aquellas características y resultados de interés para el desarrollo de la parte estadística.

Para ir ligando la notación matricial con los conceptos de la estadística multivariada, la manera de llamar a los vectores y las matrices será con una notación similar a la que se usará para expresar los vectores y matrices que constantemente se presentan en el análisis multivariado. Específicamente en el análisis discriminante tenemos información en tres niveles: poblaciones o grupos, observaciones al interior de las poblaciones o grupos y variables medidas para cada una de las observaciones. Empezaremos, entonces, por establecer la manera como se denotarán cada uno de estos conceptos:

1° las g poblaciones, que son equivalentes a decir a las g muestras o los g grupo.

2° las n observaciones o individuos que pertenecen a cada uno de los g grupos.

3° las p variables que se miden para cada uno de los n individuos dentro de los g grupos.

Los n individuos con sus p variables se representan en la matriz

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \qquad (2.1)$$

Donde el recorrido de las variables se indica con j=1,...,p; el de los individuos, con i=1,...,n.

Pero primero veamos cómo se llega a expresar las observaciones en una matriz de este tipo. Establezcamos, entonces, los elementos que nos posibiliten el uso claro de la notación vectorial para representar a las observaciones con sus variables.

Empecemos por lo más elemental, especificando la notación con que serán representados, a lo largo del trabajo, los vectores, las matrices y los escalares, así como las letras que servirán para indicar al número de grupos, el número de observaciones y el de variables, y las letras que servirán para representar los recorridos de cada uno.

Escalares se expresan en letras minúsculas	(a, x, z,...)
Vectores, con minúsculas negritas	(**a**, **x**, **z**,...)
Matrices en mayúsculas negritas	(**A**, **S**, **X**,...)
Número total de observaciones dentro de la muestra, población o grupo	n
Número total de variables se denotará con	p
Número total de grupos	g
Recorrido sobre los grupos	k
Recorrido sobre las variables	j
Recorrido sobre las observaciones	i

Un vector se expresa como columna. Así, el vector **x** de alguna observación con p variables (p×1) será:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}.$$

En caso de existir n observaciones, tendremos n vectores de observaciones

$$\mathbf{x}_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \cdots \quad \mathbf{x}_n = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \tag{2.2}$$

La transpuesta de un vector $\mathbf{x}$ (p×1) será un vector de (1×p) que se expresa $\mathbf{x}^t$. Lo mismo para una matriz, si se tiene $\mathbf{X}$ (n×p) su transpuesta es (p×n) y se expresa como $\mathbf{X}^t$. La transpuesta de la transpuesta de una matriz (o vector) es la misma matriz (o vector)

$(\mathbf{X}^t)^t = \mathbf{X}$ (matrix) $\qquad (\mathbf{x}^t)^t = \mathbf{x}$ (vector)

Una matriz cuadrada $\mathbf{S}$ (p×p) se llama ***matriz diagonal*** si todos sus elementos distintos a los de la diagonal principal, $s_{11}, s_{22}, \ldots, s_{pp}$, tienen valores 0.

$$\text{Diag } \mathbf{S} = \begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{pp} \end{pmatrix} \tag{2.3}$$

Si una matiz diagonal tiene todos sus elementos iguales a 1, se le da el nombre de ***matriz identidad*** $\mathbf{I}$, y es una matriz cuadrada.

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{2.4}$$

más adelante se verá que esta matriz tiene algunas de las propiedades del número entero 1.

Una matriz cuadrada $\mathbf{S}$ es ***simétrica*** si $\mathbf{S} = \mathbf{S}^t$, que equivale a una igualdad elemento a elemento $s_{ij} = s_{ji}$ para todo i, j.

2.1.1 Operaciones matriciales

Suma de matrices y vectores

La suma de matrices y vectores está definida cuando las matrices o vectores que se van a sumar son de igual tamaño, esto es, si **A** y **B** son matrices (n×p) entonces la suma **A+B** está definida y también es (n×p). Por ejemplo, para la matrices **A** y **B** ambas (2×1)

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{23} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{23} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{23}+b_{22} \\ a_{31}+b_{31} & a_{23}+b_{23} \end{pmatrix}$$

En caso de que las dimensiones de las matrices **A** y **B** sean diferente entre si, la suma no estará definida.

Producto de matrices

El producto de matrices **AB** está definido cuando el número de columnas de **A** es igual al número de renglones de **B**, entonces si **C** = **AB**, y el elemento c_{ij} de la matriz **C** está dado por

$$c_{ij} = \sum_k a_{ik} b_{kj}$$

Por ejemplo, si **A** es (2×3) y **B** es (3×2), **C** será (2×2)

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \bullet \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{23} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32} \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32} \end{pmatrix}$$

Que el producto de **AB** esté definido no significa que también lo esté el producto **BA**. Por ejemplo si **A** fuese (4×3) y **B** fuese de (3×2), **AB** estaría definido y arrojaría una matriz (4×2); pero **BA** no estaría definido pues el número de columnas de **B** es 2, que es distinto al número de renglones de **A** que es 4.

Cualquier matriz **X** de (n×p) puede ser multiplicada por su transpuesta, teniendo:

$\mathbf{X'X}$ es (p×p) se obtiene como el producto de sus columnas

$\mathbf{XX'}$ es (n×n) se obtiene como el producto de sus renglones (2.5)

Para cualquier matriz **A**, el producto con la matriz identidad es igual a **A**: $\mathbf{AI} = \mathbf{A}$

Algunas propiedades

Considerando la suma y el producto de matrices tenemos las siguientes propiedades. Sean **A**, **B** y **C** dos matrices cualesquiera y sean **x** y **z** dos vectores cualesquiera:

La suma e conmutativa.

$$\mathbf{A + B = B + A}$$

$$\mathbf{x + z = z + x}$$

$$\mathbf{(A + B)^t = B^t + A^t}$$

$$\mathbf{(x + z)^t = x^t + z^t} \qquad (2.6)$$

El producto es distributivo sobre la adición o sustracción:

$$\mathbf{A(B + C) = AB + AC}$$

$$\mathbf{(A + B)C = AC + BC}$$

$$\mathbf{A(B + C) \neq BA + CA}$$

$$\mathbf{(A - B)(C - D) = (A - B)C - (A - B)D = AC - BC - AD + BD} \qquad (2.7)$$

La transpuesta de un producto es el producto de las transpuestas en orden inverso:

$$\mathbf{(AB)^t = B^tA^t}$$

Si **A**, **B** y **C** son tres matrices tales que el producto **AB** existe y puede ser multiplicado por **C** se obtiene el ***triple producto*** **ABC** se puede factorizarse

$$\mathbf{ABC = A(BC) = (AB)C} \qquad (2.8)$$

$$\mathbf{ABC + ADC = A(B + D)C}$$

En algunos casos también es posible factorizar la suma del triple producto:

$$\mathbf{X^tX - X^tAX = X^t(X - AX) = X^t(I - A)X} \qquad (2.9)$$

Si **a** y **b** son vectores ambos de (n×1), $\mathbf{a^tb}$ será un escalar, ya que (1×n) (n×1) = (1×1)

$$\mathbf{a^tb} = (a_1 \quad \dots \quad a_n) \circ \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = a_1b_1 + \dots + a_nb_n$$

Mientras que $\mathbf{ab^t}$ es una matriz cuadrada (n×n) pues (n×1) (1×n) = (n×n).

$$\mathbf{ab^t} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \circ (b_1 \quad \dots \quad b_n) = \begin{pmatrix} a_1b_1 & \dots & a_1b_n \\ \vdots & & \vdots \\ a_nb_1 & \dots & a_nb_n \end{pmatrix}$$

De aquí se sigue que

$$\mathbf{a^ta} = a_1^2 + a_2^2 + \dots + a_n^2$$

que es una suma de cuadrados. Su raíz cuadrada es la distancia del origen al punto **a** que también se conoce como la ***norma*** de **a** y se denota $\|\mathbf{a}\| = \sqrt{\mathbf{a^ta}} = \sqrt{\sum_{i=1}^{n} a_i^2}$. Por su parte $\mathbf{aa^t}$ es una matriz cuadrada simétrica

$$\mathbf{aa}^t = \begin{pmatrix} a_1^2 & \cdots & a_1 a_n \\ \vdots & & \vdots \\ a_n a_1 & \cdots & a_n^2 \end{pmatrix}$$ y se le conoce como el ***producto matriz***

Cuando el producto $\mathbf{a}^t\mathbf{b}$ es un escalar

$$\mathbf{a}^t\mathbf{b} = (\mathbf{a}^t\mathbf{b})^t = \mathbf{b}^t(\mathbf{a}^t)^t = \mathbf{b}^t\mathbf{a} \qquad (2.10)$$

$$(\mathbf{a}^t\mathbf{b})^2 = (\mathbf{a}^t\mathbf{b})(\mathbf{a}^t\mathbf{b}) = (\mathbf{a}^t\mathbf{b})(\mathbf{b}^t\mathbf{a}) = \mathbf{a}^t(\mathbf{b}\mathbf{b}^t)\mathbf{a}$$

De (2.6), (2.7) y (2.10), se obtiene

$$(\mathbf{x}-\mathbf{z})^t(\mathbf{x}-\mathbf{z}) = \mathbf{x}^t\mathbf{x} - 2\mathbf{x}^t\mathbf{z} + \mathbf{z}^t\mathbf{z} \qquad (2.11)$$

Forma Cuadrática

Un caso especial de triple producto, que se presenta con mucha frecuencia en el análisis multivariado, se da cuando existe un vector $\mathbf{a}$ (n×1), su transpuesto el vector $\mathbf{a}^t$ (1×n) y una matriz simétrica $\mathbf{S}$ (n×n). Sustituyendo estos elementos en (2.8) $\mathbf{a}^t$ y $\mathbf{a}$ por $\mathbf{A}$ y $\mathbf{C}$ respectivamente, y $\mathbf{S}$ por $\mathbf{B}$, se obtiene lo que se conoce como la ***forma cuadrática***

$$\mathbf{a}^t\mathbf{S}\mathbf{a} = \sum_i a_i^2 s_{ii} + \sum_{i \neq j} a_i a_j s_{ij}$$

Si en lugar del vector $\mathbf{a}$, existe $\mathbf{b}$, entonces este producto se conoce como la ***forma bilineal***

$$\mathbf{a}^t\mathbf{S}\mathbf{b} = \sum_{i \neq j} a_i a_j s_{ij}$$

En el análisis multivariado, la forma cuadrática juega un papel central. La función de densidad de la normal multivariada expresada en términos de productos matriciales involucra la forma cuadrática. Las distancias multivariadas y la descomposición espectral, que se verán más adelante en este capítulo, también involucran esta expresión.

2.1.2 Independencia Lineal

Un conjunto de vectores $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$, se dice que son ***linealmente dependientes*** si existen constantes $c_1, c_2, \ldots, c_m$, no todas ceros, tal que $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \ldots + c_m\mathbf{a}_m = 0$. Si no existen tales constantes que satisfagan lo anterior, entonces se dice que el conjunto de vectores son ***linealmente independientes***. Cualquier vector $\mathbf{x}$ de orden n y el vector nulo $\mathbf{0}$ de orden n son linealmente dependientes.

Por ejemplo, si $\mathbf{a}_1 = (3\ 1\ -4)$ $\mathbf{a}_2 = (2\ 2\ -3)$ $\mathbf{a}_3 = (0\ -4\ 1)$ $\mathbf{a}_4 = (-4\ -4\ 6)$.

1. $2\mathbf{a}_2 + \mathbf{a}_4 = 2(2\ 2\ -3) + (-4\ -4\ 6) = (4\ 4\ -6) + (-4\ -4\ 6) = (0\ 0\ 0) = \mathbf{0}$ son linealmente dependientes
2. Si tomamos $\mathbf{a}_1$ y $\mathbf{a}_2$, $k_1\mathbf{a}_1 + k_2\mathbf{a}_2 = 0$

$$\Rightarrow\ k_1(3\ 1\ -4) + k_2(2\ 2\ -3) = 0$$

$$(3k_1 + 2k_2 \qquad k_1 + 2k_2 \qquad -4k_1 - 3k_2) = 0$$

$$3k_1 + 2k_2 = 0 \quad k_1 = 0 \ \Rightarrow\ k_2 = 0$$

$$k_1 + 2k_2 = 0$$

$$-4k_1 + 3k_2 = 0$$

Dependencia lineal implica que al menos un vector puede expresarse como combinación lineal de otros vectores en el conjunto. Así, dependencia lineal del conjunto, expresa redundancia en él.

2.1.3 Rango de una matriz

El ***Rango*** de una matriz **A** cuadrada o rectangular, se define como

Rango(**A**) = número de renglones linealmente independientes

Rango(**A**) = número de columnas linealmente independientes

Y es posible demostrar que el número de renglones linealmente independientes de una matriz es siempre igual al de columnas linealmente independientes.

Si **A** es (n×p) el rango máximo posible de **A** es el valor más pequeño entre n y p; si esto se da, se dice que **A** es de ***rango completo***.

Una consecuencia de la dependencia lineal es que cuando se tiene una matriz **A** (2x3) con los renglones linealmente independientes y rango=2 que equivale a que **A** de rango completo, entonces sólo dos de columnas son linealmente independientes, por lo que las columnas son linealmente dependientes, y existen la constantes c_1, c_2, c_3 como:

$$c_1\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} + c_2\begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} + c_3\begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

que se puede escribir como $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$\mathbf{Ac} = \mathbf{0} \qquad (2.12)$$

Esto nos lleva al resultado de que el producto de una matriz **A** y un vector **c** puede ser **0** aunque $\mathbf{A} \neq \mathbf{0}$ y $\mathbf{c} \neq \mathbf{0}$. Este resultado es de mucha utilidad al momento de calcular el vector propio de una matriz.

Otra consecuencia de la dependencia lineal de renglones y columnas es que existe la posibilidad de que **AB** = **CB** aunque **A** ≠ **C**, que es algo a tomar en cuenta, pues no es posible, en casos como este, cancelar matrices en ambo lados de la igualdad. Sin embargo existen dos excepciones en las que sí es posible realizar la cancelación en ambos lados de la igualdad, una es cuando **Ax** = **Bx**, donde es posible considerar que **A** = **B** ya que **x** toma todos los posibles valores.

El segundo caso de excepción se verá una vez que se defina lo que es una matriz regular y una singular.

2.1.4 Algunas matrices especiales

Matriz Inversa

En los número reales tenemos la existencia de a^{-1} como el inverso de a, tal que $aa^{-1} = a^{-1}a = 1$ si $a \neq 0$. Esto se extiende al álgebra de matrices de la siguiente manera: si existe **B** tal que **BA**=**AB**=1, entonces **B** es la llamada matriz inversa de **A** y se denota por $\mathbf{A}^{-1}$ con la propiedad de que $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

La condición técnica para que una matriz tenga inversa es que las k columnas $a_1, a_2, \ldots, a_k$ sean linealmente independientes. Así la existencia de $\mathbf{A}^{-1}$ es equivalente a que $c_1a_1 + c_2a_2 + \ldots + c_ka_k = 0$ sólo si $c_1 = c_2 = \ldots = c_k = 0$.[1]

Matriz regular y singular

Si **A** es una matriz cuadrada de rango completo, entonces podemos decir que **A** es ***regular*** (no singular) y tiene una inversa única que se denota por $\mathbf{A}^{-1}$ con la propiedad de que

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

Sea **A** una matriz cuadrada de (n×n). Si **B** es una matriz tal que **AB**=**I** y **BA**=**I** donde **I** es la matriz identidad (n×n), entonces se dice que $\mathbf{B} = \mathbf{A}^{-1}$ y ésta inversa en única.

Si **A** es una matriz cuadrada con rango incompleto, entonces **A** no tiene inversa y se llama ***singular***. Las matrices rectangulares no tienen inversa, aunque su rango sea completo.

[1] Para ver la forma como se calcula la matriz inversa se puede consultar cualquier libro de álgebra lineal, como por ejemplo: 'Álgebra lineal aplicada' de Ben Noble & James W. Daniel o 'Matrices' de la serie Schaum.

Si **A** y **B** son de igual tamaño y regulares, entonces el inverso de su producto es igual al producto de sus inversos en orden contrario:

$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Si **A** es rectangular de (n×p) con rango p<n, entonces $\mathbf{A}^t\mathbf{A}$ es una matriz cuadrada de (p×p) y tendrá inversa $(\mathbf{A}^t\mathbf{A})^{-1}$; pero para esta inversa no se cumple que $(\mathbf{A}^t\mathbf{A})^{-1} = (\mathbf{A}^t)^{-1}\mathbf{A}^{-1}$ ya que **A** es rectangular y por lo tanto no puede tener inversa, entonces $(\mathbf{A}^t)^{-1}\mathbf{A}^{-1}$ no existe.

Ahora si podemos abordar el segundo caso de excepción para el cual es posible cancelar matrices en ambos lados de una igualdad. Una matriz regular puede ser cancelada en ambos lados de la ecuación si aparece del mismo lado del producto, en sendos lados de la ecuación. Por ejemplo, si **B** es no singular, entonces $\mathbf{AB} = \mathbf{CB}$ implica que $\mathbf{A} = \mathbf{C}$.

Matriz positiva definida

Una matriz simétrica **A** se dice que es ***positiva definida*** si $\mathbf{x}^t\mathbf{Ax}>0$ para todo vector **x**, excepto $\mathbf{x}=0$. De manera similar, **A** es ***positiva semidefinida*** si $\mathbf{x}^t\mathbf{Ax}\geq 0$ para todo vector **x**, excepto $\mathbf{x}\neq 0$. Los elementos a_{ii} de la diagonal principal de una matriz positiva definida son positivos. De manera similar, cuando la matriz es positiva semidefinida los elementos de su diagonal principal $a_{ii}\geq 0$ para toda i.

Si una matriz es positiva definida o positiva semidefinida, entonces se define como una matriz no negativa.

Lo anterior es de mucha utilidad ya que las matrices de covarianzas y las de correlaciones siempre son matrices positivas, es decir, todos los valores propios de una matriz de covarianzas y de correlaciones serán números reales positivos. Esto también se cumple para las matrices de covarianzas y de correlaciones de una muestra.

2.1.5 Ortogonalidad

Dos vectores **a** y **b** de igual tamaño son ortogonales si: $\mathbf{a}^t\mathbf{b} = a_1b_1 + a_2b_2 + \ldots + a_mb_m = 0$.

Se dice que el vector **a** está ***normalizado*** cuando $\mathbf{a}^t\mathbf{a}=1$. El vector **a** siempre puede normalizarse al ser dividirlo por su norma $\|\mathbf{a}\| = \sqrt{\mathbf{a}^t\mathbf{a}}$, así el vector $\mathbf{c} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^t\mathbf{a}}}$ es normalizado, y por lo tanto $\mathbf{c}^t\mathbf{c}=1$.

De aquí, una matriz $\mathbf{C}=(\mathbf{c}_1, \mathbf{c}_2,...,\mathbf{c}_m)$ cuyas columnas están normalizadas y son mutuamente ortogonales, se llama ***matriz ortogonal***, y $\mathbf{C}'\mathbf{C} = \mathbf{I}$, ya que $\mathbf{c}_i'\mathbf{c}_i = 1$ y $\mathbf{c}_i'\mathbf{c}_j = 0$ para toda $i \neq j$.

$$\mathbf{C}'\mathbf{C} = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{n1} \\ c_{12} & c_{22} & \cdots & c_{n2} \\ \vdots & \vdots & & \vdots \\ c_{1n} & c_{2n} & \cdots & c_{nn} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{I} \qquad (2.13)$$

Cuando $\mathbf{C}'\mathbf{C} = \mathbf{I}$ se cumple, entonces las siguientes expresiones también se cumplen:

$$\mathbf{C}\mathbf{C}' = \mathbf{I}$$

$$\mathbf{C}' = \mathbf{C}^{-1}$$

Multiplicar por una matiz ortogonal produce el efecto de rotación de los ejes; de esta forma, si un punto $\mathbf{x}$ es transformado por $\mathbf{z} = \mathbf{C}\mathbf{x}$ donde $\mathbf{C}$ es ortogonal entonces

$$\mathbf{z}'\mathbf{z} = (\mathbf{C}\mathbf{x})'(\mathbf{C}\mathbf{x}) = \mathbf{C}'\mathbf{x}'\mathbf{C}\mathbf{x} = \mathbf{x}'\mathbf{I}\mathbf{x} = \mathbf{x}'\mathbf{x}$$

y la distancia a $\mathbf{z}$ es la misma que la distancia a $\mathbf{x}$.

2.1.6 Vectores y valores propios

Para la matriz cuadrada $\mathbf{A}$ se puede hallar un escalar λ y un vector $\mathbf{x}$ tal que $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. En esta ecuación λ es conocida como el valor propio y $\mathbf{x}$ como el vector propio de $\mathbf{A}$. Para que la ecuación $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ tenga una solución diferente a la trivial que permita hallar el vector $\mathbf{x}$, es condición necesaria y suficiente que $|(\mathbf{A} - \lambda\mathbf{I})| = 0$, es decir $(\mathbf{A} - \lambda\mathbf{I})$ sea una matriz regular, como en (2.12), para que $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x}=\mathbf{0}$ cuando $\mathbf{x}\neq\mathbf{0}$.

La ecuación $\mathbf{A} - \lambda\mathbf{I} = 0$ se conoce como ***ecuación característica***. Si $\mathbf{A}$ es una matriz de (n x n) tendrá n raíces, esto es, A tendrá n valores propios $\lambda_1,..., \lambda_n$ y por consiguiente n vectores propios $\mathbf{x}_1,...,\mathbf{x}_n$, que pueden ser hallados resolviendo la ecuación característica. Así, al valor propio λ_1 le corresponde el vector propio $\mathbf{x}_1$, al λ_2 el vector propio $\mathbf{x}_2$ y así sucesivamente. Estos vectores propios tienen la cualidad de ser linealmente independientes.

Dentro del análisis multivariado con frecuencia se requiere el uso de esta ecuación característica par realizar transformaciones dónde se obtengan vectores linealmente independientes.

Si $\mathbf{A}$ es una matriz simétrica positiva definida o semidefinida, lo que ocurrirá siempre que sea una matriz de covarianzas o de correlaciones, los valores propios de $\mathbf{A}$ siempre serán positivos, o positivos o cero según sea el caso. Es común ordenar los valores propios del más grande al más

pequeño, así los valores propios de **A** se denotan por $\lambda_1>\lambda_2>...>\lambda_p$ y los vectores propios se listan en el mismo orden como $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_p$ donde $\mathbf{x}_1$ corresponde a λ_1; $\mathbf{x}_2$, a λ_2 y así sucesivamente.

Los vectores propios de una matriz simétrica **A** (n×n) son mutuamente ortogonales, y si además, están normalizados e insertados como columnas de una matriz $\mathbf{C}=(c_1, c_2,...,c_n)$, entonces **C** es una matriz ortogonal que contiene los vectores propios de la matriz simétrica **A**.

2.1.7 Descomposición espectral de una matriz

Partiendo de la matriz ortogonal **C** que contiene los vectores propios de la matriz simétrica **A**, y dado que por ser **C** ortogonal cumple que $\mathbf{CC}^t = \mathbf{I}$, se puede multiplicar ambos lados de esta igualdad por **A**, entonces $\mathbf{ACC}^t = \mathbf{AI}$. Si tenemos que $\mathbf{C}=(\mathbf{x}_1, \mathbf{x}_2,...,\mathbf{x}_n)$, entonces podemos sustituir **C** por $(\mathbf{x}_1, \mathbf{x}_2,...,\mathbf{x}_n)$, por donde $\mathbf{x}^t$ son los vectores propios normalizados de **A**, tenemos

$$\mathbf{A} = \mathbf{A}(\mathbf{x}_1, \mathbf{x}_2,...,\mathbf{x}_n)\mathbf{C}^t$$

$$\mathbf{A} = (\mathbf{A}\mathbf{x}_1, \mathbf{A}\mathbf{x}_2,..., \mathbf{A}\mathbf{x}_n)\mathbf{C}^t$$

$$\mathbf{A} = \mathbf{A}(\lambda_1\mathbf{x}_1, \lambda_2\mathbf{x}_2,..., \lambda_n\mathbf{x}_n)\mathbf{C}^t \qquad (2.14)$$

Por lo tanto, la matriz diagonal $\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & ... & \lambda_n \end{pmatrix}$ se puede expresar (2.14) como

$$\mathbf{A} = \mathbf{CDC}^t$$

Esta relación entre una matriz simétrica **A** y sus valores y vectores propios se conoce como la ***descomposición espectral*** de la matriz **A**. Partiendo de esta descomposición espectral, **A** puede ser diagonalizada por una matriz ortogonal **C** que contiene los valores propios normalizados de **A** y así, la matriz diagonal que resulta contiene los valores propios de **A**.

$$\mathbf{AC} = \mathbf{CDC}^t\mathbf{C}$$

$$\mathbf{C}^t\mathbf{AC} = \mathbf{C}^t\mathbf{CDI}$$

$$\mathbf{C}^t\mathbf{AC} = \mathbf{D} \qquad (2.15)$$

2.2 Caracterización y manejo de datos multivariados

Con el manejo del álgebra de vectores y matrices podemos, ahora si, aprovechar al máximo los datos multivariados, manipularlos (en el buen sentido) para obtener la mayor cantidad de información posible.

El resultado obtenido en (2.2) muestra la manera como se expresa una observación con sus p variables $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$.

Si existen n observaciones (individuos) tendremos n vectores de observaciones

$$\mathbf{x}_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \cdots \quad \mathbf{x}_n = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

y podemos obtener la media de cada una de las p variable, detonándola

$$\bar{x}_1 = \frac{\sum_{i=1}^{n} x_1}{n}, \quad \bar{x}_2 = \frac{\sum_{i=1}^{n} x_2}{n}, \quad \cdots, \quad \bar{x}_p = \frac{\sum_{i=1}^{n} x_p}{n}$$

Para evitar confusión hay que hacer notar que en estas sumas cada $\bar{x}_j$ es la media de la variable j-ésima, además, no se está especificado el subíndice i que representa la observación i-ésima (i = 1,..,n), sino que al representar $\mathbf{x}_1$ en la suma, se está diciendo que estamos sumando la variable 1 de cada observación $\mathbf{x}_i$. Bajo esta notación, la manera general de expresar la media de una variable será:

$$\bar{x}_j = \sum_{i=1}^{n} x_1 \qquad \text{(llamaremos a ésta notación implícita)}$$

Estas p medias conforman el vector de medias $\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ x_p \end{pmatrix}$ (2.16)

Otra forma de representar lo anterior es expresando las n observaciones del vector de variables como:

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1j} \\ \vdots \\ x_{1p} \end{pmatrix} \qquad \mathbf{x}_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2j} \\ \vdots \\ x_{2p} \end{pmatrix} \qquad \cdots \qquad \mathbf{x}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nj} \\ \vdots \\ x_{np} \end{pmatrix}$$

de esta manera la media de cada una de las p variables se obtiene como

$$\bar{x}_1 = \sum_{i=1}^{n} x_{i1}, \quad \bar{x}_2 = \sum_{i=1}^{n} x_{i2} \quad \cdots \quad \bar{x}_p = \sum_{i=1}^{n} x_{ip}$$

cuya forma general se expresa como $\bar{x}_j = \sum_{i=1}^{n} x_{ij}$

Entonces, si queremos expresar en una matriz las observaciones con sus respectivas variables, podemos hacerlo a través de $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix}$ donde cada $\mathbf{x}_i^t$ es el vector de la observación i-ésima con sus p variables, que al transponerlo y hacerlos explícitos obtenemos la matriz mencionada al principio en (2.1). Podemos entonces dar la siguiente definición:

Definición 1. Se dice que un conjunto de datos constituye una muestra aleatoria multivariada si cada individuo a sido extraído al azar de una población de individuos, y en él se han medido u observado ciertas características. Sea x_{ij} la observación de la j-ésima variable en el i-ésimo individuo, $\mathbf{x}_i^t$ el vector fila que contiene las observaciones de todas las variables en el i-ésimo individuo, y $\mathbf{x}_j$ el vector columna que contiene las observaciones de la j-ésima variable. De esta manera se define la matriz de datos multivariados **X** como

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix} = \begin{array}{c} \text{Variables} \\ \begin{array}{cccccc} 1 & 2 & & j & & p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \begin{array}{c} 1 \\ 2 \\ \\ i \\ \\ n \end{array} \end{array} \quad \text{Observaciones} \qquad (2.17)$$

A partir de esta matriz que contiene toda la información estadística de la muestra, se puede calcular algunas funciones que permiten extraer conclusiones de los datos.

Para cada una de las p variables, es posible obtener la media, de la misma manera que en el caso univariado.

Definición 2. Dada una matriz de datos como la señalada en la definición anterior, se define la media de la variable j como $\bar{x}_j = \frac{\sum_{i=1}^{n} x_i}{n}$ y el vector promedio $\bar{x}$ de medias de las p variables como:

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{pmatrix} \qquad (2.18)$$

Este vector también se puede expresar en términos matriciales, par lo que se requiere el apoyo de un vector **j** (n×1) de valores uno.

$$\mathbf{j} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \qquad \text{Así} \quad \bar{\mathbf{x}} = \frac{\mathbf{X}^t\mathbf{j}}{n} \qquad (2.19)$$

Partiendo de la varianza de una variable $Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ y la covarianza entre dos variables $Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, tenemos la siguiente definición para el caso multivariado.

Definición 3. Dada una matriz de datos como la señalada en la definición 1, se define la varianza de la variable j por $s_{jj} = Var(x_j) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n}$; y la covarianza entre las variables j-ésima e h-ésima por $s_{ij} = Cov(x_j, x_h) = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{n}$ donde i, h se usan para enumerar las variables, mientras k se usan para enumerar los individuos. Así, la matriz de varianzas y covarianzas se expresa de la siguiente manera

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{i1} & s_{i2} & \cdots & s_{ij} & \cdots & s_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nj} & \cdots & s_{np} \end{pmatrix}$$

Retomando la matriz **X** de datos (2.17 y el vector de medias (2.16) y calculando la covarianza para cada variable obtenemos las entradas de la matriz **S**. De esta forma, la $\text{Cov}(x_1,x_1)=\text{Var}(x)=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}{n-1}$, la $\text{Cov}(x_1,x_2)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{y}_2)$, y así sucesivamente, lo que se expresa de manera general como

$$\text{Cov}(x_j,x_h) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ih}-\bar{y}_h)$$

Cada entrada de la matriz corresponde a la covarianza de dos de la p variables. Para verlo con claridad, a continuación se hace un desglose de algunas de las entradas de la matriz:

$\text{Cov}(x_1,x_1) = \text{Var}(x_1) = s_{11}$ → ocupará el renglón 1 columna 1
$\text{Cov}(x_1,x_2) = s_{12}$ → ocupará el renglón 1 columna 2
⋮
$\text{Cov}(x_1,x_j) = s_{1j}$ → ocupará el renglón 1 columna j
⋮
$\text{Cov}(x_1,x_p) = s_{1p}$ → ocupará el renglón 1 columna p

---

$\text{Cov}(x_2,x_1) = s_{21}$ → ocupará el renglón 2 columna 1
$\text{Cov}(x_2,x_2) = \text{Var}(x_2) = s_{22}$ → ocupará el renglón 2 columna 2
⋮
$\text{Cov}(x_2,x_j) = s_{2j}$ → ocupará el renglón 2 columna j
⋮
$\text{Cov}(x_1,x_p) = s_{2p}$ → ocupará el renglón 2 columna p
⋮
$\text{Cov}(x_p,x_p) = \text{Var}(x_p) = s_{pp}$ → ocupará el renglón p columna p

Así el total de las covarianzas para cada par de variables forman la matriz **S** de varianzas y covarianza de la muestra.

Puesto que la $\text{Cov}(x_j,x_{j+1})=\text{Cov}(x_{j+1},x_j)$, **S** es una matriz simétrica cuya diagonal principal tiene las varianzas de cada una de las p variables, mientras que las covarianzas de la i-ésima variable con las otras p−1 variables están fuera de la diagonal principal.

$$s_{jj} = s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1} \tag{2.20}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}^2 - n\bar{x}_j^2) \tag{2.21}$$

mientras que $s_{hj} = \sum_{i=1}^{n}(x_{ih} - \bar{x}_h)(x_{ij} - \bar{x}_j)$ (2.22)

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ih}x_{ij}\bar{x} - n\bar{x}_h\bar{x}_j) \tag{2.23}$$

Donde h, j son variables y cumplen que h, j=1,..., p; h≠j. Mientras que i=1,...,n representa las observaciones.

En el caso multivariado esta matriz de covarianzas se denota **S** (nunca $\mathbf{S}^2$) a diferencia del caso univariado donde la varianza se denota s^2.

En términos de los vectores de las observaciones (vectores observados) la matriz de covarianzas se expresa

$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t$ donde $\mathbf{x}_i$ es el vector (p×1) conformado por las p variables del individuo i-ésimo, mientras que $\mathbf{x}$ es el vector (p×1) de las medias de las p variables.

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ x_p \end{pmatrix} \quad \Rightarrow \quad (\mathbf{x}_i - \bar{\mathbf{x}}) = \begin{pmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_1 \\ \vdots \\ x_{ip} - \bar{x}_p \end{pmatrix}$$

$$\Rightarrow (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t = \begin{pmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_2 \\ \vdots \\ x_{ip} - \bar{x}_p \end{pmatrix} \begin{pmatrix} x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 & \dots & x_{ip} - \bar{x}_p \end{pmatrix}$$

Así, la matriz de covarianzas se ve de la siguiente manera:

$$\mathbf{S} = \begin{pmatrix} \sum(x_{i1}-\bar{x}_1)^2 & \sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \cdots & \sum(x_{i1}-\bar{x}_1)(x_{ip}-\bar{x}_p) \\ \sum(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & \sum(x_{i2}-\bar{x}_2)^2 & \cdots & \sum(x_{i2}-\bar{x}_2)(x_{ip}-\bar{x}_p) \\ \vdots & \vdots & & \vdots \\ \sum(x_{ip}-\bar{x}_p)(x_{i1}-\bar{x}_1) & \sum(x_{ip}-\bar{x}_p)(x_{i2}-\bar{x}_2) & \cdots & \sum(x_{ip}-\bar{x}_p)^2 \end{pmatrix}$$

Donde cada suma se aplica sobre i=1,...,n. y cada suma al interior de la matriz arroja un escalar. Esta matriz se expresa en términos de matriciales como:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}^t - n\bar{\mathbf{x}}_i\bar{\mathbf{x}}^t)$$

que se obtiene de manera análoga al caso univariado.

También es posible obtener **S** directamente de la matriz **X** planteada en (2.17) considerando:

1. $\sum_{i=1}^{n} x_{ih}x_{ij}$ es el producto de la h-ésima y j-ésima columna de **X**
2. $n\bar{x}_h\bar{x}_j$ es el elemento hj-ésimo de $n\bar{\mathbf{x}}\bar{\mathbf{x}}^t$ del producto de las columnas (ver 2.5)
3. Sabemos que $\bar{\mathbf{x}} = \frac{\mathbf{X}^t\mathbf{j}}{n}$ y de aquí $n\bar{\mathbf{x}} = \mathbf{X}^t\mathbf{j}$.
4. Usando la matriz de unos **J**, tenemos que $n\bar{\mathbf{x}}\bar{\mathbf{x}}^t = \mathbf{X}^t(\mathbf{J}/n)\mathbf{X}$

y finalmente sustituyendo estos resultados en (2.23), **S** se puede expresar como

$$\mathbf{S} = \frac{1}{n-1}\left[\mathbf{X}^t\mathbf{X} - \mathbf{X}^t\left(\frac{1}{n}\mathbf{J}\right)\mathbf{X}\right]$$

$$\mathbf{S} = \frac{1}{n-1}\left[\mathbf{X}^t\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{X}\right] \quad \text{por (2.9)} \qquad (2.24)$$

Como la covarianza depende de una escala de medición de las variables, resulta difícil comparar covarianzas entre diferentes pares de variables; por ejemplo si cambiamos millas por kilómetros, la covarianza cambia. Para resolver este problema es posible estandarizar la covarianza dividiéndola por la desviación estándar de las variables. A esta covarianza estandarizada se le llama ***correlación***.

Definición 4. A partir de la matriz **S** es posible calcular la matriz de correlación muestral **R** (pxp), de igual dimensión que **S** y cuyos elementos sean los coeficientes de correlación entre la j-ésima y la k-ésima variable (en lugar de las covarianzas como en el caso de **S**):

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}\cdot s_{kk}}} = \frac{s_{jk}}{s_j\cdot s_k} \qquad \text{Cuando } j = k,\ r=1. \qquad (2.25)$$

La diagonal principal estará formada por números uno y será simétrica como la matriz de covarianzas por ser $r_{jk} = r_{kj}$.

Para el caso bivariado, la correlación poblacional de dos variables x, y se expresa

$$\rho_{xy} = Cov(x,y) = \frac{\sigma xy}{\sigma x \, \sigma y}$$

y para la muestral se tiene $r_{xy} = \frac{Sxy}{SxSy}$

r y ρ nunca son iguales, en particular cuando $\rho_{xy} = 0$ no existe correlación ente las variables.

En términos de vectores, la correlación muestral es el coseno del ángulo θ. Sea θ el ángulo entre los vectores **a**, **b**; el vector que va del extremo de **a** hacia el de **b** lo llamaremos **c** y está dado por **c=b–a**.

Por la regla de los cosenos $\text{Cos } \theta = \mathbf{a}^T\mathbf{b} / [(\mathbf{a}^T\mathbf{a})(\mathbf{b}^T\mathbf{b})]^{-1/2}$

Cuando **a** y **b** son ortogonales, $\mathbf{a}^T\mathbf{b} = 0$, es decir $\text{Cos } \theta = 0 \;\therefore\; \theta=90°$. En un sentido geométrico, estos vectores son perpendiculares (figura 2.1).

Si el Cos θ es cercano a 1, r_{xy} será cercano a 1; si dos vectores son perpendiculares, el Cos θ

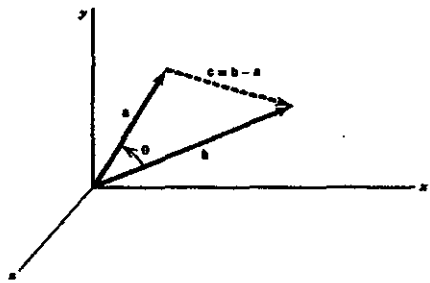

Figura 2.1 Vectores a y b en un espacio de 3 dimensiones

= 0 y $r_{xy} = 0$; cuando los vectores tienen direcciones opuestas r_{xy} estará cercano a –1.

Una medida que sintetiza la dispersión de los datos, es el determinante de la matriz **S** al que se le da el nombre de varianza generalizada $V = |S|$. Representa la varianza multivariada.

Existe un estimado, que es de vital importancia en el estudio del análisis multivariado y particularmente en el análisis de discriminante, la matiz de varianzas y covarianzas ponderada (pooled), que es una medida de las dispersión general de los datos en la que se involucra la información de todas las poblaciones.

El estimador de la varianza de una población esta dado por $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$, si tenemos dos poblaciones $X_1 \sim N(\mu_1,\sigma_1)$ y $X_2 \sim N(\mu_2,\sigma_2)$ con un total de n_1 y n_2 observaciones respectivamente, con medias distintas $\mu_1 \neq \mu_2$ y varianzas iguales $\sigma_1 = \sigma_2 = \sigma$, tendremos los estimadores s^2_1 y s^2_2. De aquí se puede obtener otra medida que es un estimador de la varianzas en el que se involucra la información de las dos poblaciones se denota por s_{pl}. Se conoce como ***el estimador combinado de la varianza***. Al dividir s^2_1 y s^2_2 entre la varianza conjunta, obtenemos:

$$\frac{\sum_{i=1}^{n}(x_{1i} - \bar{X}_1)^2}{\sigma^2} \sim {\chi_1}^2_{(n1-1)} \quad \text{y para la población 2} \quad \frac{\sum_{i=1}^{n}(x_{2i} - \bar{X}_2)^2}{\sigma^2} \sim {\chi_2}^2_{(n2-1)}$$

el estimador $s_{pl}^2 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n}(x_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2} \sim {\chi_2}^2_{(n1+n2-2)}$ es una medida conjunta de la varianza que considera las observaciones de las dos poblaciones. También puede calcularse como una ponderación de las varianzas de cada población.

$$[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}]\mathbf{S}_1 + [\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}]\mathbf{S}_2 = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 - 1) + (n_2 - 1)} \qquad (2.26)$$

Que es una expresión más clara de entender. Esta varianza ponderada es un estimador insesgado de la varianza poblacional $\boldsymbol{\Sigma}$. Es requisito que $n_1 + n_2 - 2 > p$, para que la matriz $\mathbf{S}_{pl}$ no sea singular y pueda tener inversa.

2.2.1 Combinación lineal de variables

En multivariado es muy común trabajar con combinaciones lineales de variables, pues es una manera de modificar la dimensión a una más simple y fácil de manipular. Como se verá más adelante, la función discriminante es una combinación lineal de variables que maximiza una función y existirán momentos en que se requiera conocer la media y demás estadísticos asociados a ella.

Sean $a_1, a_2, \ldots, a_n$ constantes y considérese la combinación lineal z de elementos del vector $\mathbf{x}$: $z = a_1x_1 + a_2x_2 + \ldots + a_px_p = \mathbf{a}'\mathbf{x}$, donde $\mathbf{a}' = (a_1,a_2,\ldots,a_p)$ es un vector de coeficientes. Si $\mathbf{a}$ se aplica a cada vector $\mathbf{x}_i$ de la muestra —donde i representa la observación i-ésima—, se puede expresar el valor z de la i-ésima observación como

$$z_i = a_1x_{i1} + a_2x_{i2} + \ldots + a_px_{ip} = \mathbf{a}'\mathbf{x}_i \qquad \text{para } i=1,\ldots,n.$$

La media muestral de z puede obtenerse como un promedio de los valores $z_1 = \mathbf{a}^t\mathbf{x}_1$, $z_2 = \mathbf{a}^t\mathbf{x}_2, \ldots, z_p = \mathbf{a}^t\mathbf{x}_p$ o como una combinación lineal del vector de medias de los x_i, $\bar{\mathbf{x}}$.

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \quad = \quad \mathbf{a}^t\bar{\mathbf{x}} \qquad (2.27)$$

Promedio de valore z_i — Combinación lineal de $\bar{\mathbf{x}}$

De forma similar, la varianza muestral de z puede obtenerse como la varianza de los $z_1, \ldots, z_n$ o directamente del vector de coeficientes **a** y la matriz de covarianzas **S** de los vectores de cada observación $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$

$$s_z^2 = \frac{\sum_{i=1}(z_i - \bar{z})^2}{n-1} = \mathbf{a}^t\mathbf{S}\mathbf{a} \qquad (2.28)$$

donde s_z^2 es el resultado multivariado análogo al de la varianza univariada de una muestra que está multiplicada por una constante *a*

$$s_z^2 = \frac{\sum_{i=1}^{n}(ax_i - a\bar{x})^2}{n-1} = a^2 s^2$$

Siempre la varianza es no negativa, entonces $s_z^2 \geq 0$ y por lo tanto $\mathbf{a}^t\mathbf{S}\mathbf{a} \geq 0 \ \forall \ \mathbf{a}$; tenemos entonces que **S** al menos es positiva semidefinida. Si además, las variables son continuas, no están linealmente relacionadas y $n-1 \geq p$, entonces **S** es de rango completo, por lo tanto, con toda seguridad **S** siempre es positiva definida.

Si definimos otra combinación lineal $w = \mathbf{b}^t\mathbf{x} = b_1x_1 + + b_2x_2 + \ldots + b_px_p$ con $\mathbf{b}^t \neq \mathbf{a}^t$, entonces $\mathrm{Cov}(z,w) = S_{zw} = \mathbf{b}^t\mathbf{S}\mathbf{a}$

y la correlación $r_{zw} = \dfrac{S_{zw}}{S_z^2 S_w^2} = \dfrac{b^tSa}{\sqrt{(a^tSa)(b^tSb)}}$

2.2.2 Las distancias

Consideremos la distancia entre los puntos $A=(x_1, y_1)$ y $B(x_1, y_2)$ en un espacio bivariado, como se representan en la figura 2.2. La distancia entre ambos está dada por la línea recta que va de uno a otro. Esta distancia es la hipotenusa del triángulo rectángulo APB, entonces, por el teorema de Pitágoras se tiene que $d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$.

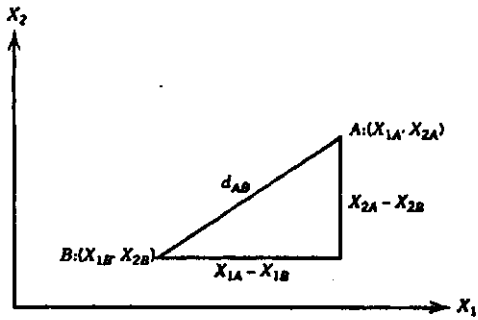

Figura 2.2 Distancia en el plano entre los puntos A y B

Esto también puede expresarse en término de vectores. Sean $\mathbf{x}_1$ y $\mathbf{x}_2$ vectores columna de dimensión (2x1), con las entradas $\mathbf{x}_1(x_{11}\ x_{12})$ y $\mathbf{x}_2(x_{21}\ x_{22})$. La distancia entre ellos será:

$$d^2 = \begin{pmatrix} x_{11} - x_{21} & x_{21} - x_{22} \end{pmatrix} \begin{pmatrix} x_{11} - x_{21} \\ x_{21} - x_{22} \end{pmatrix}$$

$$d^2 = (\mathbf{x}_1 - \mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2)$$

Extendiendo esta idea de distancia Euclideana a un espacio de p dimensiones, la distancia quedaría expresada:

$$d^2 = \sum_{j=1}^{p} (xj - xk)^2$$ y nuevamente en términos de vectores $d^2 = (\mathbf{x}_1 - \mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2)$

Siempre que no exista correlación entre las variables y que las varianzas sean igual a 1.

Para el caso multivariado, esta distancia no dice mucho; sin embargo, si la consideramos en términos de la desviación estándar, eso ya nos proporciona más información.

Para obtener una medida de distancia útil en el análisis multivariado, se deben considerar, no sólo las varianzas de las variables, sino también sus covarianzas o correlaciones. La distancia Euclideana de dos vectores no es útil, pues no está ajustada por las varianzas o las covarianzas. Para lograr una medida estadística de la distancia, se requiere estandarizar por medio de la inserción de la matriz inversa de la covarianza $d^2 = (\mathbf{x}_1 - \mathbf{x}_2)^t \mathbf{S}^{-1}(\mathbf{x}_1 - \mathbf{x}_2)$

2.2.2.1 La distancia D^2 de Mahalanobis

Considerando lo anterior, Mahalanobis desarrolló tres medidas sobre distancia (1936). La primera es la distancia entre dos puntos x_1 y x_2 en un espacio de dimensión p donde las variables

pueden estar correlacionadas y con varianzas distintas a 0. Esta medida involucra la matriz Σ de covarianzas de la población y se expresa como:

$$\Delta_{12} = (x_1 - x_2)^t \Sigma^{-1} (x_1 - x_2)$$

La segunda mide la distancia entre dos poblaciones que se obtiene a través del calculo de la distancia entre dos puntos donde cada punto representa un vector de medias de las p variables. A estos vectores se le llama **centroide** y es el valor promedio de un individuo dentro de la población. El centroide de la población k-ésima se denota por $\mu^t_k = \mu_{1k}\ \mu_{2k} \ldots \mu_{pk}$, donde μ_{ik} es la media de la variable i en la población k. Así la distancia entre dos poblaciones se expresa de la siguiente manera:

$$\Delta_{12} = [(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)]$$

Donde Σ es la matriz de covarianzas **común** a las dos poblaciones. Más adelante se verá que uno de los supuestos del análisis de discriminante es la igualdad en las varianzas y covarianzas, he aquí la razón de esta suposición.

La última de estas medidas es la distancia entre una población y un individuo. Así, se tiene un vector que representa el centroide de la población y otro vector con las p variables del individuo. Supóngase que se tienen g poblaciones. La distancia entre el centroide de la población g y el vector de p entradas (variables) para el individuo i, se representa como:

$$\Delta_{12} = [(x_i - \mu_1)^t \Sigma^{-1} (x_i - \mu_2)$$

Donde Σ es la matriz de covarianzas de la población k. Esta distancia es de interés particular en el análisis de discriminante, en el momento de la clasificación, pues para asignar cada individuo en un grupo, se comparan sus distancias al centroide de cada grupo y se asigna a aquel que le quede más cercano.

Es importante hacer notar que, si una variable tiene mayor varianza que otra, esta recibe menor peso en la distancia de Mahalanobis. De manera similar, dos variables con una fuerte correlación no contribuyen más que dos variables que tengan una menor correlación.

En esencia, el uso de la matriz de covarianzas inversa en la distancia de Mahalanobis tiene dos efectos: primero estandarizar todas las variables a la misma varianza, y segundo, eliminar la correlación.

2.3 Dos resultados importantes

1. Teniendo la matriz **A** (n×p), puede multiplicarse por su transpuesta y se obtiene $\mathbf{AA}^t$ que es el producto de (n×p) (p×n), por lo tanto obtenemos una matriz (n×n) que es el producto de sus renglones; mientras que $\mathbf{A}^t\mathbf{A}$ es el producto (p×n) (n×p) y se obtiene una matriz (p×p) que es el producto de sus columnas.

2. Un resultado de mucha importancia en el análisis discriminante es el que se presenta a continuación. Con su aplicación se obtiene la separación máxima entre grupos. Partiendo de la desigualdad de Cauchy-Schwartz se obtiene este resultado sobre maximación:

Si **B** es una matriz (p×p) definida positiva y **d** es el vector (p×1), entonces para un vector **x** (p×1) cualquiera, distinto de cero, el valor máximo del cociente $\frac{(\mathbf{x}^t\mathbf{d})^2}{\mathbf{x}^t\mathbf{B}\mathbf{x}}$ estará dado por $\mathbf{d}^t\mathbf{B}^{-1}\mathbf{d}$, cuando $\mathbf{x}=c\mathbf{B}^{-1}\mathbf{d}$ para cualquier $c\neq 0$. Lo que se expresa como:

$$\max \frac{(\mathbf{x}^t\mathbf{d})^2}{\mathbf{x}^t\mathbf{B}\mathbf{x}} = \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d} \qquad (2.29)$$

Capítulo III

EL ANÁLISIS MULTIVARIADO

El análisis estadístico multivariado o análisis multivariado es la parte de la estadística que estudia las distribuciones multivariada o multidimensionales —que tienen dos o más variables— y sus muestras. En las aplicaciones, el análisis multivariado trabaja con uno o varios grupos de individuos, cada uno de los cuales posee valores para una o más variables. Permite estudiar las interrelaciones entre las variables, las posibilidades de diferencias entre los grupos en términos de esas variables y delinear las inferencias relevantes de las variables concernientes a las poblaciones para las cuales se seleccionaron los grupos de muestra.

En este capítulo primero se presentará la distribución normal multivariada y algunas de sus propiedades más importantes y luego dos de las pruebas de significancia que son útiles para el análisis discriminate: la T^2 de Hotelling que es la contraparte multivariada de la prueba t en univariado, y la lambda de Wilks.

3.1 Distribución normal multivariada

Definición 3.1 Se dice que un vector $\mathbf{x}$ de dimensión (px1) tiene una distribución normal p-variada con vector media μ y matiz de covarianzas Σ si su función de densidad está dada por

$$f(\mathbf{x}) = (2\pi^p |\Sigma|)^{-1/2} \exp [-1/2 (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})]$$

Donde p es el número de variables. Cuando $\mathbf{x}$ tiene esta función de distribución, se dice que $\mathbf{x}$ se distribuye como un normal p con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, o simplemente $\mathbf{x} \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

El término $(\mathbf{x} - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ en el exponente de la normal multivariada es la distancia generalizada de $\mathbf{x}$ al cuadrado, que equivale a la distancia de Mahalanobis. Un valor pequeño de $|\boldsymbol{\Sigma}|$ indica que las $\mathbf{x}$ están concentradas alrededor de la media $\boldsymbol{\mu}$ en el espacio p o que existe una situación de alta correlación entre las variables (multicolinealidad). En este caso uno o más valores propios serán cero, con lo que se reduce la dimensión de la matriz.

Ya que es poco probable que se conozcan los valores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, es necesario estimarlos. De esta forma, la función estimada se expresa sustituyendo los valores poblacionales por los estimadores de los parámetros:

$$f(\mathbf{x}) = (2\pi^p |\mathbf{S}|)^{-1/2} \exp [-1/2 (\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})]$$

Donde $\mathbf{S}$ es la matriz (p×p) de covarianzas de la muestra y $\bar{\mathbf{x}}$ es el vector de (p×1) de medias de la muestra.

Podemos nombrar algunas de las características de la función normal multivariada que resultan útiles:

1. Para describir completamente la distribución se requiere estimar las medias, las varianzas y las covarianzas
2. Si las variables no están correlacionadas entonces son independientes
3. Las combinaciones lineales de la función, también son normales.
4. Una adecuada forma de la función de densidad proporciona un buen número de pruebas estadísticas.
5. Sirve como una buena aproximación a otras funciones que pueden aproximarse a ella con base en el teorema central del límite.

3.1.1 Propiedades de la normal multivariada

Listaremos algunas de las propiedades de un vector $\mathbf{x}$ aleatorio (p×1) de una función normal multivariada:

1. Si $\mathbf{a}$ es un vector de constantes, la función lineal $\mathbf{a}^t\mathbf{x} = a_1y_1 + a_2y_2 + \ldots + a_py_p$ se describe como $N(\mathbf{a}^t\boldsymbol{\mu},\mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a})$. La media $E(\mathbf{a}^t\mathbf{x}) = \mathbf{a}^t\boldsymbol{\mu}$ y la $Var(\mathbf{a}^t\mathbf{x}) = \mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a}$ para cualquier vector aleatorio $\mathbf{x}$. Adicionalmente, si $\mathbf{x} \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, entonces $\mathbf{a}^t\mathbf{x}$ tiene una distribución normal (univariada).

2. Si la es una matriz de constantes **A** (q×p) de rango q donde q ≤ p, entonces **Ax** consiste de q combinaciones lineales de las variables en **x**, con distribución $\mathbf{x} \sim N_p(\mathbf{A\mu}, \mathbf{A\Sigma A}^t)$. Nuevamente $E(\mathbf{Ax}) = \mathbf{A\mu}$ y $Cov(\mathbf{Ax}) = \mathbf{A\Sigma A}^t$; además, otra característica, es que las q variables en **Ax** tienen una distribución normal multivariada.

3. **Estandarización de variables.** Si $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se puede obtener un vector estandarizado $\mathbf{z} = (T^t)^{-1}(\mathbf{x} - \boldsymbol{\mu})$ donde $\boldsymbol{\Sigma} = T^tT$. Expresado de otra manera:

   $\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{x} - \boldsymbol{\mu})$ donde $\boldsymbol{\Sigma}^{1/2}$

   De la propiedad anterior, se sigue que **z** se distribuye como una $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{I})$; esto es, las **z**'s ~ N(0,1), lo que significa que en el vector de medias todas las entradas son 0, las varianzas son todas 1 y las correlaciones son 0.

4. **Distribución Ji-cuadrada.** Una variable aleatoria ji-cuadrada con p grados de libertad, está definida como la suma de cuadrados de las p variables aleatorias independientes estandarizadas normalmente. Si **z** es el vector estandarizado entonces $\sum_{i=1}^{p} z_i^2 = \mathbf{z}^T\mathbf{z}$ tiene una distribución χ^2 con p grados de libertad, y de la propiedad anterior se obtiene $\mathbf{z}^t\mathbf{z} = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$. De aquí, si $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces $(\mathbf{x} - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2$.

5. **Normalidad de las distribuciones marginales.** Cualquier subconjunto x's en **x** tiene una distribución normal multivariada con vector de medias dado por la el correspondiente subvector de μ y matriz de covarianza compuesta por su correspondiente submatriz **Σ**. Es decir, si $\mathbf{x}_1^t = (x_1, x_2, \ldots, x_r)$ denota un vector subconjunto de **x** con r elementos y $\mathbf{x}_2^t = (x_{r+1}, x_{r+2}, \ldots, x_p)$ otro vector subconjunto de **x**, entonces **x**, **μ** y **Σ** se pueden partir en:

$$\mathbf{x} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

   donde $\mathbf{x}_i$, **μ** son vectores (r x 1) y **Σ** es una matriz (r×r). Entonces $\mathbf{x}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ con $E(\mathbf{x}_1) = \boldsymbol{\mu}$ y $Cov(\mathbf{x}_1) = \boldsymbol{\Sigma}_{11}$.

6. **Independencia.**

   a) Los subvectores **x** y **y** son independientes si $\boldsymbol{\Sigma}_{xy} = 0$

   b) Dos variables individuales x_i y x_j son independientes si $\sigma_{xy} = 0$.

7. **Distribución Condicional.** Si $\mathbf{x}$ y $\mathbf{y}$ son variables no independientes, entonces $\Sigma_{xy} \neq \mathbf{0}$, y la distribución condicional de $\mathbf{y}$ dado $\mathbf{x}$, $f(\mathbf{y}/\mathbf{x})$ es una normal multivariada con

   $E(\mathbf{y}/\mathbf{x}) = \boldsymbol{\mu}_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$ y $Cov(\mathbf{y}/\mathbf{x}) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$

   La matriz $\Sigma_{yx}\Sigma_{xx}$ es llamada la ***matriz de coeficientes de regresión.***

8. **Distribución de la suma de dos vectores.** Si $\mathbf{x}$ y $\mathbf{y}$ son del mismo tamaño, por ejemplo ($p \times 1$), e independientes, entonces

   $\mathbf{x} + \mathbf{y}$ es $N_p(\boldsymbol{\mu}_y + \boldsymbol{\mu}_x,\ \Sigma_{yy} + \Sigma_{xx})$

   $\mathbf{x} - \mathbf{y}$ es $N_p(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x,\ \Sigma_{yy} + \Sigma_{xx})$

3.2 Pruebas de Hipótesis

En el análisis multivariado, al igual que para el caso univariado, existen pruebas con respecto a las mediad para una o más poblaciones. Lo mismo ocurre con las pruebas con respecto a la varianza. Sin embargo nuestro interés se centra en las pruebas que involucran dos muestras, por lo que se presentarán las estadísticas y distribuciones en las que se apoyan estas pruebas. Para hacer clara la presentación de cada caso, se partirá de la prueba univariada análoga correspondiente. Para la prueba con respecto a la diferencia de medias de dos muestras, comenzaremos con la prueba t para dos muestras del caso univariado para llegar a la T^2 de Hotelling; mientras que en lo referente a la Λ de Wilks, partiremos del análisis de varianza univariado.

3.2.1 Prueba con respecto a la diferencia entre las medias de dos muestras (T^2 de Hotelling)

Dentro de la estadística con frecuencia se requiere saber si dos poblaciones son diferentes entre sí, para lo que se comparan los valores de las medias de cada población y vemos si existe o no suficiente evidencia para asegurar que son diferentes —lo que nos indicaría que las poblaciones son diferentes— o en caso contrario, no poder asegurar que son distintas, con lo que aceptaríamos que no obtenemos evidencia suficiente para asegurar que hay deferencias.

Veamos primero el caso donde se estudia una variable y supongamos que tenemos poblaciones independientes que se distribuyen normalmente y tienen varianzas iguales y para cada población se toma una muestra ($x_1 \sim N_1(\mu_1,\sigma_1^2)$ y $x_2 \sim N_2(\mu_2,\sigma_2^2)$) cuyas medias muestrales son $\bar{x}_1$ y $\bar{x}_2$. Las inferencias se formulan con base en la diferencia que hay entre las medias muestrales. La distribución de $\bar{x}_1$ es normal con media μ_1 y varianza σ_1^2/n_1; mientras que la media de $\bar{x}_2$ es normal con media μ_2 y varianza σ_2^2/n_2. Dado que $\bar{x}_1$ y $\bar{x}_2$ son variables aleatorias independientes normalmente distribuidas, entonces la diferencia $\bar{x}_1 - \bar{x}_2$, también se distribuye como una normal con media $\mu_1 - \mu_2$ y varianza $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Por lo tanto, si se conoce σ^2, la distribución de

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1). \qquad (3.1)$$

Para cada una de las dos muestras aleatorias, se pueden definir las varianzas muestrales s_1^2 y s_2^2. Además, sabemos (de los resultados univariados) que $(n_1-1)s_1^2/n_1$ y $(n_2-1)s_2^2/n_2$ son dos variables independientes ji-cuadrada con (n_1-1) y (n_2-1) grados de libertad respectivamente. Sabemos que la suma de dos distribuciones ji-cuadrada se distribuye como una ji-cuadrada con grados de libertad igual a la suma de los grados de libertad de las dos distribuciones, así

$$W = \frac{(n_1-1)s_1^2}{\sigma^2} + \frac{(n_2-1)s_2^2}{\sigma^2} \sim X^2_{n1+n2-2} \qquad (3.2)$$

El cociente de z y la raíz cuadrada de w, da como resultado el cociente $\frac{N(0,1)}{X^2_{n1+n2-2}}$ que se distribuye como una t de Student con $n_1 + n_2 - 2$ grados de libertad ($t_{n1+n2-2}$), cuya expresión es

$$t = \frac{[\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]/\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{[(n_1-1)s_1^2 + (n_2-1)s_2^2]/\sigma^2}{n_1+n_2-2}}} = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Podemos tomar lo que se conoce como el estimador combinado (pooled) de la varianza común σ^2, definido como

$$s^2_{pl} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \quad \text{que es insesgado, ya que } E(s^2_{pl}) = \sigma^2.$$

Nótese que s^2_{pl} es la varianza ponderada de las muestras, siendo los factores de peso los grados de libertad. Así, obtenemos

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_{pl}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Esta es la expresión que sirve para formulara inferencias respecto a la diferencia de las medias de poblaciones.

Para probar $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

se usa $$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{pl}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3.3)$$

que se distribución como una t con n_1+n_2-2 grados de libertad cuando H_0 es cierta. Se rechazará H_0 si $| t | \geq t_{\alpha/2,\, n1+n2-2}$

Ahora veamos la prueba para el caso multivariado. Consideremos el caso donde p variables son medidas en cada observación de las muestras. Obtenemos los vectores de medias $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$, Queremos probar

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

De cada población obtenemos las muestras aleatorias; $\mathbf{x}_1=(x_{11}, x_{12},\ldots,x_{1n1}) \sim N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ de la población 1 y $\mathbf{x}_2= (x_{21}, x_{22},\ldots,x_{2n2}) \sim N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ de la 2. Las muestras son independientes y con varianzas iguales pero desconocidas $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$. Los vectores de medias estarán dados por:

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1} \quad \text{y} \quad \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

Definimos $\mathbf{W}_1$ y $\mathbf{W}_2$ como las matrices de suma de cuadrados y productos cruzados de las dos muestras:

$$\mathbf{W}_1 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)^t = (n_1 - 1)\mathbf{S}_1$$

$$\mathbf{W}_2 = \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)^t = (n_2 - 1)\mathbf{S}_2$$

Donde $(n_1 - 1)\mathbf{S}_1$ y $(n_2 - 1)\mathbf{S}_2$ son estimadores insesgados de $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$ respectivamente. Podemos combinarlos y obtener un estimador insesgado conjunto para la matriz de covarianzas poblacional común ($\mathbf{\Sigma}$): $\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 - 2}(\mathbf{W}_1 + \mathbf{W}_2)$

$$\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2]$$

Es insesgado, ya que $E(\mathbf{S}_{pl}) = \mathbf{\Sigma}$.

Partiendo (3.3) que presenta la expresión de t, podemos obtener t^2 univariada:

$$t^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{\left(s_{pl}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)^2} = \frac{(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)}{s_{pl}^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)} = \left(\frac{n_1 + n_2}{n_1 n_2}\right)\frac{(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)}{s_{pl}^2}$$

finalmente $t^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)(s_{pl}^2)^{-1}(\bar{x}_1 - \bar{x}_2)$

Que puede generalizarse al caso de p variables si se sustituye $(\bar{x}_1 - \bar{x}_2)$ por $(\bar{x}_1 - \bar{x}_2)$ y $\mathbf{S}_{pl}$ por s_{pl}^2, con lo que se obtiene

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^t(\mathbf{S}_{pl})^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (3.4)$$

Que se distribuye como $T^2_{p,\, n1+n2-2}$ cuando H_0 es verdadera. La prueba rechaza H_0 cuando T^2 excede el valor del punto crítico $T^2_{\alpha,p,n1+n2-2}$.

La estadística T^2 en (3.4) puede expresarse en su forma característica

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^t\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pl}\right]^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (3.5)$$

3.2.1.1 Algunas propiedades de la prueba T^2

1. Es necesario que $n_1+n_2-2>p$ para que $\mathbf{S}_{pl}$ sea regular (no singular). Al ser regular $\mathbf{S}_{pl}$ podemos obtener sus valores y vectores propios (ver sección 2.1.6).
2. La estadística T^2 también se puede expresar en términos de la distancia estandarizada entre $\bar{x}_1$ y $\bar{x}_2$, donde $D^2 = (\bar{x}_1 - \bar{x}_2)^t(\mathbf{S}_{pl})^{-1}(\bar{x}_1 - \bar{x}_2)$, entonces

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}D^2. \qquad (3.6)$$

Esta distancia en T^2 es un escalar, que cuando es grande, T^2 será grande y se rechazará la hipótesis nula de que $\mu_1 = \mu_2$.

3. Se espera que T^2 tenga una función de densidad sesgada ya que su límite inferior es 0 pero no tiene límite superior. Por su parte la estadística F también es una función sesgada. Estas dos funciones están directamente relacionadas y la estadística T^2 se transforma en una estadística F usando $\frac{n_1+n_2-p-1}{n_1+n_2-2}T^2 = F_{p,n1+n2-p-1}$ donde la dimensión p de T^2 en el primer parámetro de grados de libertad para F. Aunque únicamente se está mencionando la transformación sin presentar su deducción, es útil tenerla en cuenta pues al momento de ubicar el punto crítico se puede recurrir a las tablas de la función F.

3.2.2 Análisis de varianza multivariado (MANOVA)

En el análisis de varianza del caso univariado, tenemos un número g de grupos a los que se les aplican distintos tratamientos. Se quiere saber si la aplicación de estos tratamientos tiene efectos sobre los grupos, de ser así la variable respuesta x se verá afectada. Si los tratamientos producen un efecto, las medias de x para cada tratamiento serían distintas. Con esto, el valor de una observación dentro de un grupo se expresa en la siguiente ecuación

$$x_{ki} = \bar{x} + (\bar{x}_k - \bar{x}) + (x_{ki} - \bar{x}_k)$$

observación	media global	suma cuadrados de tratamientos	suma cuadrados de residuos

donde k es el recorrido de los tratamientos = 1,...,g
i es el recorrido de las observaciones = $1,...,n_k$

Sea $N = n_1 + n_2 + ... + n_g$, donde n_k es el número de observaciones del tratamiento k.

Para probar que las medias son distintas en cada tratamiento, se plantea la hipótesis nula sobre la no existencia de diferencias provocada por la aplicación de los tratamientos

$H_0 : \tau_1 = \tau_2 = ... = \tau_g = 0$

La estadística adecuada para probar la hipótesis nula es el cociente dado por

$$F = \frac{SCTR/(g-1)}{SCE/(N-g)} \qquad (3.7)$$

Donde SCTR es la suma de cuadrados de los tratamientos y SCE es la suma de cuadrados de los errores. La prueba F rechaza H_0 a un nivel de significancia α si $F > F_{g-1,N-g}$.

Para un F grande, son los tratamientos lo que ocasiona la variación que se presentan entre las poblaciones y se rechaza H_0; mientras que para un F pequeño, son los residuos los causantes de ésta, que es un efecto de la aleatoriedad, por tanto se acepta H_0.

La extensión del modelo de análisis de varianza al caso multivariado con g tratamientos, n_k observaciones en el tratamiento k y p variables en la medición de cada observación, queda de la siguiente manera:

$$\mathbf{x}_{kl} = \boldsymbol{\mu} + \tau_k + \boldsymbol{\varepsilon}_{kl} \quad \text{con } k=1,\ldots,g; \quad l=1,\ldots,n_k$$

donde $\boldsymbol{\varepsilon}_k$ son independientes y se distribuyen $N_p(0,\Sigma)$, el vector $\boldsymbol{\mu}$ es la media global de los tratamientos y τ_k representa el efecto del k-ésimo tratamiento de forma tal que $\sum_{k=1}^{g} n_k \tau_k = 0$. Los errores para los componentes de $\mathbf{x}_{kl}$ están correlacionados, pero la matriz de covarianzas Σ es la misma para toda la población. De acuerdo con el modelo, un vector de observaciones puede descomponerse

$\mathbf{x}_{kl}$	=	$\bar{\mathbf{x}}$	+	$(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$	+	$(\mathbf{x}_{kl} - \bar{\mathbf{x}}_k)$	(3.8)
observación		media global		estimación del efecto de los tratamientos		residual	

Se puede ver que esta descomposición es análoga las suma de cuadrados en el caso univariado. Partiendo de (3.8) se obtiene la siguiente expresión

$$\sum_{k=1}^{g}\sum_{l=1}^{n_k}(\mathbf{x}_{kl} - \bar{\mathbf{x}})(\mathbf{x}_{kl} - \bar{\mathbf{x}})^t = \sum_{k=1}^{g} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^t + \sum_{k=1}^{g}\sum_{l=1}^{n_k}(\mathbf{x}_{kl} - \bar{\mathbf{x}}_k)(\mathbf{x}_{kl} - \bar{\mathbf{x}}_k)^t$$

Suma de cuadrados y productos cruzados total (SCT)	Suma de cuadrados y productos cruzados de los tratamientos (SCTR)	Suma de cuadrados y productos cruzados de residuos (errores) (SCE)
	Variación **entre** Tratamientos	Variación **al interior** de cada tratamiento

La variación al interior (within) o suma de cuadrados de los errores (SCE), se denota con la matriz **W** y se expresa como:

$$\mathbf{W} = \sum_{k=1}^{g}\sum_{l=1}^{n_k}(\mathbf{x}_{kl} - \bar{\mathbf{x}}_k)(\mathbf{x}_{kl} - \bar{\mathbf{x}}_k)^t$$

$$\mathbf{W} = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \ldots + (n_g - 1)\mathbf{S}_g$$

Donde $\mathbf{S}_k$ es la matriz de covarianzas de la k-ésima muestra. Esta matriz es una generalización de la $\mathbf{S}_{pl}$ para dos muestras y juega un papel preponderante en las pruebas de los efectos de tratamientos.

Por su parte la variación entre los grupos (between) o suma de cuadrados de los tratamientos está dada por:

$$\mathbf{B} = \sum_{k=1}^{g} (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^t$$

La hipótesis nula $H_0 : \tau_1 = \tau_2 = ... = \tau_g = \mathbf{0}$ nos dice que no hay efecto de los tratamientos. La prueba se realiza considerando los tamaños relativos de las sumas de cuadrados y productos cruzados de los tratamientos (SCTR) y residuos (SCE). Equivalentemente hay que considerar los tamaños relativos de los residuos y la sumas de cuadrados y productos cruzados total (SCT).

La prueba donde $H_0 : \tau_1 = \tau_2 = ... = \tau_g = \mathbf{0}$ involucra varianzas generalizadas. Se rechaza H_0 Se rechazará H_0 si el cociente de las varianzas generalizadas es muy pequeño

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^t}{\sum_{k=1}^{g}\sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})^t} \qquad (3.9)$$

Es decir si $\Lambda \leq \Lambda_{\alpha,p,v_B,v_W}$. Existen tablas para hallar valor crítico exacto $\Lambda_{\alpha,p,v_B,v_W}$, pero también se puede realizar una transformación de la estadística F para el caso p=2.

La Λ compara la suma de cuadrados y productos cruzados al interior de las muestras con la suma de cuadrados y productos cruzados total que es similar a la prueba F en univariado (3.7) que compara la suma de cuadrados entre las muestras con la suma de cuadrados al interior de ellas. Al usar determinantes la prueba Λ se reduce a un escalar, así la información multidimensional que contienen **B** y **W** es transformada en un solo valor escalar que da la información suficiente para decidir si la separación de medias es significativas. Esto es algo típico en las pruebas multivariadas.

3.2.2.1 Algunas propiedades de Λ

1. Se necesita que $v_W \geq p$ para que el determinante de (3.9) sea positivo

2. La Λ de Wilks (3.9) se puede expresar en términos de los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_s$ de $\mathbf{W}^{-1}\mathbf{B}$ de la siguiente manera:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

El número de valores propios de $\mathbf{W}^{-1}\mathbf{B}$ distintos de cero es $s=\min(p,v_B)$, el rango de $\mathbf{B}$.

3. Λ es una prueba en sentido inverso a lo que estamos acostumbrados. Toma valores entre 0 y 1, y se rechaza para valores pequeños. Si los vectores de medias de las muestras son iguales, entonces $\mathbf{B}=0$ y por lo tanto $\Lambda=1$. En caso contrario, si la separación entre los vectores de medias es grande, $\mathbf{B}$ la variación entre las muestras, será más grande que $\mathbf{W}$ y Λ se aproxima a 0.

4. El valor crítico decrece conforme p crece. Es decir, la prueba pierde poder conforme aumenta el número de variables a menos que las variables que contribuyen a rechazar H_0 provoquen una reducción significativa en Λ.

5. Cuando $p=2$ o cuando $v_B=2$, Λ de Wilks se transforma exactamente en una estadística F. Para este caso especial, la trasformación de Λ a F se muestra en la tabla 3.1. La hipótesis H_0 se rechaza cuando el valor transformado de Λ excede el nivel superior de puntos porcentuales de la estadística F.

Cuadro 3.1

Transformación de la Λ de Wilks en una F exacta

Parámetros	Estadística F	Grados de libertad
Cualquier p, $v_B=1$	$\frac{1-\Lambda}{\Lambda}\frac{v_B-p+1}{p}$	p, v_B-p+1
Cualquier p, $v_B=2$	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{v_B-p+1}{p}$	$2p, 2v_W-9+1$
Cualquier v_B, p=1	$\frac{1-\Lambda}{\Lambda}\frac{v_B}{v_W}$	v_B, v_W
Cualquier v_B, p=2	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{v_W-1}{v_B}$	$2v_B, 2(v_W-1)$

6. Cuando el número de grupos es 2, el cociente de la Λ de Wilks se reduce a una función de la T^2 de Hotelling.

7. Aunque en el caso multivariado la prueba Λ rechaza la hipótesis nula, no implica, necesariamente, que la prueba F del caso univariado, si se aplica a cada una de las variables por separado, va a rechazar la hipótesis nula. Puede ser que lo haga para unas, pero no para todas. De igual manera puede ser que F rechace algunas variables mientras que Λ acepta H_0. Si se está realizando un estudio multivariado, es la prueba multivariada la que hay que tomar en cuenta, pero siempre es bueno plantear lo que puede ocurrir, puesto que la F univariada también tiene un uso dentro de multivariado cuando se está realizando la selección de las variables que conviene incluir en el análisis.

Aunque dentro de estas propiedades no está mencionado, para valores de p y v_B mayores a 2, existe una aproximación de F que viene a ser la extensión multivariada de la prueba F en el caso univariado. De hecho, se puede realizar una transformación a la Λ de Wilks y se obtiene una F.

Capítulo IV

ANÁLISIS DISCRIMINANTE

La metodología que se emplea para abordar el problema planteado, es una de las técnicas que hay en análisis multivariado para problemas de clasificación. En este caso el interés está en realizar una clasificación de las familias en dos grupos: familias en condiciones de pobreza extrema y familias en condiciones de pobreza. La pertenencia de cada familia a uno de los grupos debe ser explicada con la información que sobre ellas se tiene, misma que está comprendida en un conjunto de variables.

El análisis de discriminante es una técnica adecuada para abordar problemas de este tipo, ya que, su propósito básico es estimar la relación de una variable independiente no métrica (categórica) y un conjunto de variables independientes métricas, que se expresa a través de una combinación lineal de las variables métricas.

Para tener un panorama claro de lo que nos aporta esta técnica, primero haremos una explicación sobre ella para posteriormente, ya con más conocimiento abordar la parte teórica y el desarrollo de sus principales elementos.

4.1 Un primer acercamiento

Usaremos el término *grupo* para referirnos a una población o a una muestra de la población. Partiendo de que se tienen k grupos conocidos existen dos objetivos principales en la separación de estos; el primero, comprender las diferencias entre los grupos y el segundo, clasificar

correctamente individuos nuevos en grupos ya definidos, por lo tanto se le puede considerar tanto un tipo de análisis de perfil, como una técnica predictiva. Así, el análisis discriminante tiene dos vertientes. La primera es la descripción de los grupos, que se consigue encontrando una función que permita describir las diferencias entre ellos, la *función discriminante.* Las segunda se da cuando ya teniendo definida esta función, es posible evaluar y clasificar (predecir) cualquier otro individuo nuevo, con base en ella para determinar a qué grupo pertenece.

Esto permite su aplicación en situaciones donde se desea identificar el grupo al cual pertenece un individuo. Algunos ejemplos de casos en los que puede ser usado el análisis de discriminante son:

- Predicciones de éxito o fracaso de productos nuevos en el mercado.
- Clasificación de estudiantes de acuerdo a sus intereses vocacionales.
- Nivel de riesgo en el otorgamiento de créditos.

Sin embargo, no sólo puede servir para construir funciones, clasificar casos y probar la diferencia entre grupos, también puede aplicarse para explorar y describir

- Qué variables, entre muchas, son las más útiles para discriminar entre grupos.
- Si un grupo de variables funciona igual de bien que otro.
- Qué grupo es más parecido
- Cuáles son los casos extremos, es decir, dentro de su grupo, aquellos que son marcadamente diferentes a los demás.

Teóricamente no existen restricciones para el número de grupos de clasificación permitidos. En el caso de dos grupos, la técnica se conoce como análisis de discriminante de dos grupos. Cuando se tiene más de dos grupos, se llama análisis de discriminante múltiple.

Nosotros contamos con dos grupos definidos que serán los que a su vez definirán la variable dependiente no métrica : Grupo 1 : Familias en condiciones de pobreza extrema
Grupo 2 : Familias en condiciones de pobreza

La función discriminante es una combinación lineal de las variables independientes, que asigna cada individuo a uno de los grupos de clasificación definidos en un principio.

La discriminación tiene como objetivo que **la diferencia (varianza) al interior de los grupos sea la mínima, a la vez que la diferencia (varianza) entre los grupos sea la máxima posible.** Aquella combinación lineal de variables independientes que logre estas características en las varianzas, es la llamada **función discriminante,** que describe las diferencias entre los grupos e identifica la contribución relativa para separa los grupos con lo que se obtiene la

configuración óptima. De esta manera la función discriminante es la combinación lineal que separa los grupos. Para el individuo i, esta función sería la siguiente:

$$z_i = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n$$

donde a_0 = constante

a_j = Ponderación de la variable j

z_i = Puntuación discriminante del individuo i

Cada variable independiente es multiplicada por su correspondiente ponderación y los productos se suman. Como resultado se obtiene una única puntuación z discriminantes para cada individuo —Nótese que esta expresión es similar a una función de regresión—. Ya al interior de cada grupo, el promedio de estas puntuaciones es la medida conocida como **centroides**, que indican la situación promedio de un individuo dentro de su grupo. Mediante la comparación de los centroides –a través de sus distancias– es como se puede conocer que tan apartados están los grupos entre sí.

Una medida del éxito del análisis de discriminante es si la función discriminante ha dado lugar a grupos con centroides significativamente diferentes. La distancia D^2 de Mahalanobis, es una medida de la diferencia entre centroides y se cuenta con criterios que determinan si las diferencias son significativas.

Para analizar la **significación estadística** de la función discriminante, se considera una medida generalizada de la distancia entre los centroides. Esto se calcula comparando las distribuciones de las puntuaciones discriminantes de los grupos. La Lambda de Wilks, también llamada estadística U, es una prueba del análisis de varianza múltiple que toma valores entre 0 y 1, útil para probar la igualdad de los centroides. Valores pequeños indican diferencia entre las medias de los grupos.

La gráfica 4.1 muestra la distribución de las puntuaciones discriminantes de una función que separa bien los grupos:

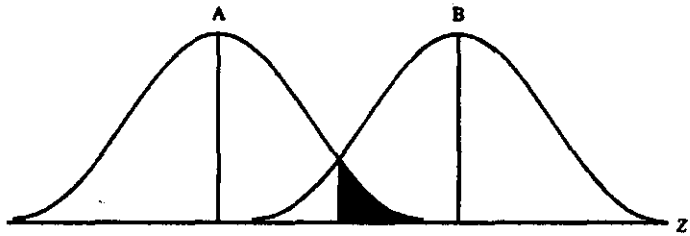

Figura 4.1 Función discrimiante para grupos bien separados

Mientras que la gráfica 4.2 muestra la distribución de las puntuaciones discriminantes de una función que separa mal los grupos. En ambos casos las áreas sombradas son la probabilidad de clasificar erróneamente observaciones del grupo A en el B.

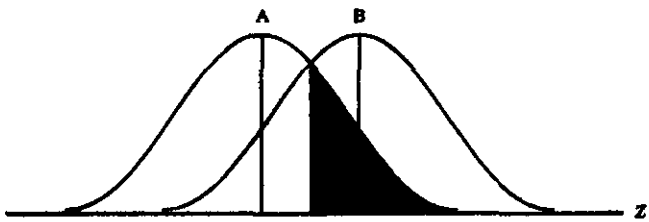

Figura 4.2 Función discrimiante para grupos mal separados

Otros criterios de que se dispone para medir la significancia estadística son:

- La traza de Hottelling
- El criterio de Pillai
- D^2 de Mahalanobis

Los supuestos claves para efectuar el análisis y obtener la función discriminante son:

1) Normalidad multivariada de las variables independientes.

2) Covarianzas y dispersión desconocidas pero iguales para los grupos.

4.1.1 Métodos para modelar la función discriminante

No siempre es necesario incluir todas las variables en el modelo discriminante. Puede suceder que algunas nos aporten la misma información, o que otras no aporten información. Si las variables están muy correlacionadas entre sí, posiblemente se tengan varias posibilidades de modelo óptimo.

Dentro de los métodos para el cálculo de la función discriminante, podemos mencionar dos que son los más conocidos:

1. **La estimación simultanea**. Calcular la función discriminante considerando todas las variables independientes a la vez, sin tomar en cuenta la capacidad discriminante de cada una.

2. **La estimación por etapas**. Calcula la función discriminante incluyendo las variables independientes una a una según su capacidad discriminante. Así la función se va formando, primero con la variable que más discrimina, luego, ya teniendo esta variable como base, se analizan las otras y se toma aquella a que en conjunto con la primera, discrimina más. El experimento se repite hasta ya no poder incluir variables que aporten algo a la discriminación o que su aportación sea tan baja que convenga más no incluirla en la función. Al final las variables excluidas no contribuyen significativamente a la discriminación.

En muchas aplicaciones es deseable descartar variables independientes que resulten redundantes ante la presencia de alguna otra variable. Así mismo, cuando no se tiene interés particular sobre la inclusión de una variable, sino lo que se quiere es identificar libremente aquellas variables que mejor separan los grupos, el método por etapas resulta adecuado.

Este método consta de dos acercamientos para la identificación de las variables más significativas. El primero es conocido como ***selección hacia delante*** (*forward selection*), donde la variable que entra en cada paso es aquella que maximiza la estadística parcial F basada en la lambda de Wilks (Λ de Wilks). Esto se consigue realizando el cálculo de la $\Lambda(x_i)$ para cada variable de manera individual y se elige aquella que sea menor (o la de mayor asociación F). En el segundo paso se calcula $\Lambda(x_i/x_1)$ para todas las variables restantes, donde x_1 es la primera variable incluida. Nuevamente se toma aquella con la menor $\Lambda(x_i/x_1)$. El proceso se repite ahora considerando $\Lambda(x_i/x_1,x_2)$ y continúa hasta que la F cae por debajo de algún valor crítico predeterminado.

El segundo acercamiento es el de ***selección hacia atrás*** (*backward selection*) que funciona de manera similar eliminando las variables que contribuyen menos en cada paso, de acuerdo a lo que arroja la estadística F parcial. La variable con la menor F parcial será removida del grupo de las seleccionadas.

La ***selección por pasos*** (*stepwise*) es una combinación de los acercamientos de selección hacia delante y selección hacia atrás. Las variables se seleccionan una a la vez en cada paso y son revaluadas para ver si alguna que haya entrado anteriormente se ha convertido en redundante ante la presencia de otra variable. De esta forma se obtiene la máxima separación adicional de grupos por arriba y por abajo de la separación ya alcanzada por otras variables.

El método por etapas no es recomendable en muestras pequeñas ya que se torna inestable y no es posible la generalización.

Hay que tener claro que aquí no se realiza es el cálculo de la función discriminante, sino el proceso de selección de variables paso a paso MANOVA (stepwise MANOVA). Una vez completada esta selección, entonces se calcula la función discriminante con las variables seleccionadas.

El análisis discriminante está relacionado con el análisis de varianza múltiple y con la regresión múltiple. Se comienza con dos o más grupos conocidos, como en el análisis de varianza, y entonces se aplica el procedimiento discriminante para obtener la combinación lineal de las variables dependientes que mejor caracterice la diferencia entre los grupos. Esta combinación lineal es equivalente a la ecuación de regresión múltiple. El procedimiento de discriminación estima los coeficientes y la función resultante puede ser empleada para clasifica casos nuevos, para lo que se define una regla de clasificación. Por otra parte la función discrimínate permite obtener para cada observación lo que se conoce como las **puntuaciones discriminantes**.

4.1.2 La regla de clasificación

Una regla de clasificación que es comúnmente usada es la que se basa en el principio de máxima verosimilitud que consiste en asignar un individuo a la población en la cual el vector de observaciones tiene la mayor verosimilitud de ocurrencia. Esto puede verse en términos de la función verosimilitud, o la función de densidad de probabilidad.

Veamos un ejemplo para el caso univariado: supóngase que se tienen dos poblaciones con sus respectivas funciones de densidad $f_1(x)$ y $f_2(x)$. Si una observación w es seleccionada aleatoriamente de f_1, la verosimilitud para la observación x=w es denotado por $f_1(w)$. Al aplicar el principio de máxima verosimilitud se obtiene la siguiente regla de asignación:

"asignar el individuo a con x=w a la población 1 si $f_1(x) > f_2(x)$".

Lo que significa, que si la verosimilitud de una observación x=w es mayor para la población 1 que para la población 2; si esto no se cumple, entonces asigna w a la población 2.

Trasladando lo anterior al caso multivariado, si $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ son las poblaciones con función de densidad multivariada, entonces la máxima verosimilitud será: asignar la observación i-ésima a la población 1 si la verosimilitud del vector $\mathbf{x}_i$ es mayor para el grupo 1 que para cualquier otro grupo.

Si $f_1(\mathbf{x}_i) > f_2(\mathbf{x}_i)$ asigne la i-ésima a la población 1
de otra forma asigne la i-ésima a la población 2.

Con la regla de clasificación, se determina un valor crítico, llamado también **punto de corte**, contra el cual se compara cada puntuación discriminante individual para determinar dentro de qué grupo debe clasificarse.

Una regla adecuada de clasificación va a depender de la bondad de los estimadores de probabilidad, así como del tamaño de la muestra en que se basa la estimación. Considerando esto, puede resultar adecuado contar con información adicional como los tamaños relativos de las poblaciones en consideración. Esto es, obtener la probabilidad p_g de que al tomar aleatoriamente una observación del universo ésta provenga de la población g. A este valor se le da el nombre de probabilidad *a priori* de la población.

Otro aspecto importe a considerar es el error de asignación que pueda presentarse. Es casi imposible definir una regla de asignación perfecta. Siempre va a existir una probabilidad de cometer algún error al momento de la asignación. Lo importante es asegurar que este error sea lo más pequeño posible.

Como se verá más adelante, si se cuenta con información que permitan conocer estas probabilidades pueden ser incorporadas a la regla de clasificación. No necesariamente se tienen que conocer ambas, puede ser que sólo una de ellas sea conocida.

4.2 Desarrollo Teórico

La función discriminante es aquella combinación lineal de las variables medidas para cada observación que mejor separa los grupos.

Supongamos que se tienen dos poblaciones Π_1 y Π_2 que están compuestas por n_1 y n_2 observaciones respectivamente y p variables que se miden para cada observación y que se asocian a cada Π_i; hay una función de densidad de probabilidad (fdp) $f_k(\mathbf{x})$ en R^p, tal que, para cualquier individuo que provenga de la población Π_1 tiene una fdp $f_1(\mathbf{x})$. Ambas poblaciones tienen la misma covarianza, pero medias diferentes; esto es $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$ y $\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$. El vector de las observaciones de la muestra de Π_1 es $\mathbf{x}_{11}, \mathbf{x}_{12},\ldots,\mathbf{x}_{1n1}$, y el de Π_2 $\mathbf{x}_{21}, \mathbf{x}_{22},\ldots,\mathbf{x}_{2n2}$, donde cada vector $\mathbf{x}_{ij}$ contiene los valores de las p variables. El objetivo del análisis discriminante es asignar cada individuo a una de las poblaciones con base en las medidas de $\mathbf{x}$, donde $\mathbf{x}$ es el vector de características de un individuo. Lo deseable es cometer la menor cantidad de errores posibles en la asignación.

Una buena regla de clasificación es aquella que presenta pocos errores al asignar cada individuo. En otras palabras, aquella donde la probabilidad de clasificar incorrectamente una observación sea pequeña. Existen otros dos aspectos que vale la pena tomar en cuenta. El primero es que puede darse el caso de contar con información sobre la probabilidad de que una observación provenga de una u otra población. De esta manera podemos decir que p_1 es la probabilidad de que la observación pertenezca a Π_1 y p_2, de que pertenezca a Π_2.

El segundo aspecto a considerar es el costo de un error de clasificación, por ejemplo, se desea clasificar a las personas en dos grupos, aquellas que son enfermos potenciales de cierto mal y aquellas que no lo son. Es más costoso clasificar incorrectamente a una persona que es un enfermo potencial como sano, que a una persona sana como enfermo potencial.

Una buena regla de clasificación es aquella que involucra estos dos aspectos; sin embargo, no siempre se puede contar con información al respecto.

Considerando nuevamente las poblaciones Π_1 y Π_2, sea Ω el espacio muestral que es una colección de todas las posibles observaciones $\mathbf{x}$ que pueden clasificarse en Π_1 o Π_2, y sean R_1 el conjunto de las observaciones de Ω que se clasifican en Π_1, mientras que R_2 el conjunto de las restantes $\mathbf{x}$ que se clasifican en R_2. $R_1+R_2= \Omega$, es decir, R_1 y R_2 son mutuamente excluyentes y exhaustivas.

La probabilidad condicional de clasificar una observación en Π_2 cuando realmente pertenece a Π_1 se expresa como:

$$P(2/1) = P(\mathbf{x} \in R_2/\Pi_1) = \int_{R_2} f(x_1)dx$$

que representa el volumen formado por la función de densidad de $f(\mathbf{x}_1)$ sobre R_2. Igualmente, la probabilidad condicional de clasificar una observación en Π_1 cuando realmente pertenece a Π_2 se expresa

$$P(2/1) = P(\mathbf{x} \in R_1/\Pi_2) = \int_{R_1} f(x_2)dx$$

La figura 4.3 presenta una gráfica donde lo anterior se ve con claridad.

Sea p_1 la probabilidad inicial de Π_1 y p_2, la de Π_2 donde $p_1+ p_2=1$. La probabilidad global de clasificar correcta o incorrectamente una observación $\mathbf{x}$ es el producto de la probabilidad inicial y la condicional.

P(clasificar correctamente en Π_1) = P(observación provenga de Π_1 y se clasifica correctamente)
$= P(\mathbf{x} \in R_1/\Pi_1)\, P(\Pi_1) = P(1/1)\, p_1$

P(error al clasificar en Π_1) = P(observación provenga de Π_2 y se clasifica incorrectamente en Π_1)
$= P(\mathbf{x} \in R_1/\Pi_2)\, P(\Pi_2) = P(1/2)\, p_2$

P(clasificar correctamente en Π_2) = P(observación provenga de Π_2 y se clasifica correctamente)
$= P(\mathbf{x} \in R_2/\Pi_2)\, P(\Pi_2) = P(2/2)\, p_2$

P(error al clasificar en Π_2) = P(observación provenga de Π_1 y se clasifica incorrectamente en Π_2)
$= P(\mathbf{x} \in R_2/\Pi_1)\, P(\Pi_1) = P(2/1)\, p_1$

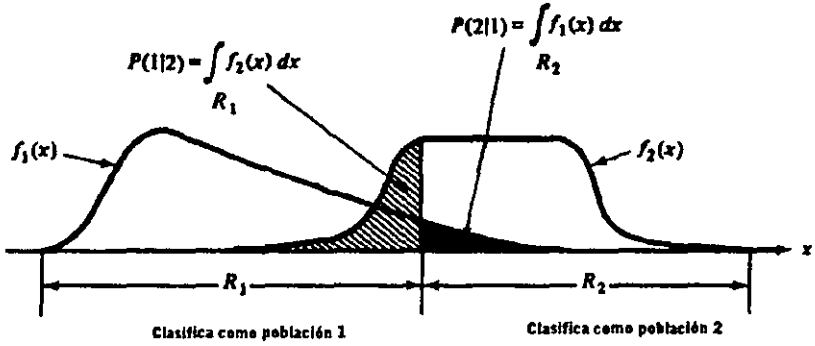

Figura 4.3 Error de clasificación para dos regiones cuando p =1

Estas clasificaciones están considerando las probabilidades de los errores de clasificación pero ignoran los costos que un error de este tipo genera. Los costos pueden afectar una regla de clasificación ya que aunque se tengan probabilidades de errores pequeñas, por ejemplo P(1/2)=0.06, si el costo es muy alto, la probabilidad puede hacerse muy grande y por consiguiente la regla no es eficiente.

El costo de una mala clasificación puede plantearse a través de una matriz de comparación

		Clasificado como	
		Π_1	Π_2
Población	Π_1	0	c(2/1)
Verdadera	Π_2	c(1/2)	0

Donde c(2/1) es el costo de clasificar la observación en Π_2 siendo de Π_1, c(1/2) es el costo de clasificar la observación en Π_1 cuando realmente es de Π_2 y cero es el costo de no cometer error.

Así para cualquier regla, el costo esperado de una mala clasificación (EMC por sus siglas en ingles 'expected misclassification error') se obtiene al multiplicar la diagonal inversa por sus probabilidades de ocurrencia

$$ECM = c(2/1)\,P(2/1)\,p_1 + c(1/2)\,P(1/2)\,p_2$$

Una buena regla de clasificación es cuando EMC es lo más pequeño posible.

Resultado 1.

Las regiones R_1, R_2, que minimizan el error de clasificación esperado, están definidas por los valores **x** para los cuales se cumple la siguiente desigualdad:

$$R_1 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_1}{p_2}\right)$$

Proporción entre fdp $\geq$ Proporción de costos — Proporción de las probabilidades iniciales

$$R_2 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

Proporción entre fdp $\geq$ Proporción de costos — Proporción de las probabilidades iniciales

Cuando las probabilidades iniciales no se conoce, entonces se supone iguales: $p_1/p_2=1$, Así

$$EMC=c(2/1)P(2/1)+c(1/2)P(1/2)$$

Y las regiones se representan

$$R_1 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1/2)}{c(2/1)}\right) \quad \text{y} \quad R_2 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1/2)}{c(2/1)}\right)$$

En caso de que sea el costo de clasificar incorrectamente lo que no se conoce, se suponen iguales: $c(1/2)/c(2/1)=1$. Las regiones de clasificación óptima estarán dadas por la comparación de las funciones de densidad y las probabilidades originales.

$$R_1 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_1}{p_2} \quad \text{y} \quad R_2 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$

Pero también puede darse que ni las probabilidades iniciales ni el costo de clasificar erróneamente una observación se conozcan, en este caso ambos serán considerados iguales, es decir, $p_1/p_2=1$ y $c(1/2)/c(2/1)=1$. Las regiones de clasificación óptima estarán determinadas por la comparación de las funciones de densidad

$$R_1 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \quad \text{y} \quad R_2 = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

En la práctica es común usar esta regla de clasificación.

Hasta ahora no se ha especificado alguna distribución para las poblaciones; nos hemos limitado a considerar cualquier función de probabilidad. Ahora vamos a hablar de poblaciones que tienen una distribución normal multivariada.

Si $f_1(x)$ y $f_2(x)$ son normales multivariadas, la función de densidad conjunta de Π_1 y Π_2 estará dada por

$$f_i(x)=\frac{1}{2\pi\,|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \qquad \text{para } i = 1, 2$$

supóngase que los parámetros poblacionales $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}$ son conocidos.

Para obtener la región en la que se minimiza el EMC, después de cancelar 2π y $\boldsymbol{\Sigma}$ se tiene

$$R_1 = \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \geq \left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_1 = \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \geq \left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

Dadas estas regiones, podemos construir la siguiente regla de clasificación:

Sean las poblaciones Π_1 y Π_2 con función de densidad de probabilidad normal multivariada. La regla de asignación que minimiza el costo esperado de mala clasificación está dada por:

Asignar x_0 a Π_1 si

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_0-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (4.1)$$

Asignar x_0 a Π_2 para cualquier otro caso.

Hay que considerar que en la gran mayoría de las situaciones los valores poblacionales $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}$ son desconocidos por lo que es necesario remplazarlos por sus respectivas estimaciones $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ y $\mathbf{S}$. Consideremos la variable aleatoria multivariada $x^t=(x_1, x_2,...,x_p)$ y supongamos que para la población Π_1 se tienen n_1 observaciones, mientras que para la población Π_2, se tienen n_2 observaciones y que $n_1+ n_2-2\geq p$. Las respectivas matrices de datos serán:

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{12} \\ \vdots \\ \mathbf{x}_{1n_1} \end{pmatrix} \quad \text{y} \quad \mathbf{x}_2 = \begin{pmatrix} \mathbf{x}_{21} \\ \mathbf{x}_{22} \\ \vdots \\ \mathbf{x}_{2n_2} \end{pmatrix}$$

Cuyas medias muestrales y matriz de covarianzas estarán dados por

$$\text{Medias} \quad \bar{x}_1 = \frac{\sum_{j=1}^{n} x_{1j}}{n_1}, \quad \bar{x}_2 = \frac{\sum_{j=1}^{n} x_{2j}}{n_2}$$

$$\text{Varianzas} \quad \mathbf{S}_1 = \frac{\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_j)(x_{1j} - \bar{x}_j)^t}{n_1 - 1}, \quad \mathbf{S}_2 = \frac{\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_j)(x_{2j} - \bar{x}_j)^t}{n_2 - 1}$$

$$\text{Varianza ponderada} \quad \mathbf{S}_{pl} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 - 1) + (n_2 - 1)}$$

Sustituyendo los valores poblacionales de (4.1) por sus respectivos valores muestrales, obtenemos la siguiente regla de clasificación:

Asignar x_0 a Π_1 si

$$(\bar{x}_1 - \bar{x}_2)^t \mathbf{S}_{pl}^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2) + \mathbf{S}_{pl}^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln\left[\left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (4.2)$$

Asignar x_0 a Π_2 para cualquier otro caso.

Cuando los costos de mala clasificación y las probabilidades iniciales son iguales para cada población $\left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)=1$, pero ln(1)=0. En este caso la regla de clasificación tendrá como punto crítico el punto medio (m) entre los centroides de las poblaciones. El resultado 4.6, que se presenta más adelante en este capítulo, ilustra la manera como se encuentra este punto. Por ahora simplemente mencionaremos la regla de asignación que corresponde cuando los costos del error de clasificación y las probabilidades iniciales son iguales.

Asignar x_0 a Π_1 si

$$\text{si } z_0 = (\bar{x}_1 - \bar{x}_2)^t \mathbf{S}_{pl}^{-1} x_0 \geq m \qquad (4.3)$$

Asignar x_0 a Π_2 para cualquier otro caso.

4.2.1 La función discriminante

Con la misma idea de lograr una buena clasificación, analizaremos ahora la función discriminante. Volvamos a las poblaciones Π_1 y Π_2, con las características que ya hemos definido. Trabajaremos con las muestras de las poblaciones $(x_{11}, x_{12},..., x_{1n_1})$ y $(x_{11}, x_{12},..., x_{1n_2})$ y el objetivo es identificar la población a la que pertenece una observación $\mathbf{x}$. La idea original de Fisher fue transformar las observaciones $\mathbf{x}$ multivariadas en univariada aplicando una combinación lineal z de las observaciones, así obtuvo $\mathbf{z} = \mathbf{a}'\mathbf{x}$, que explícitamente y de manera general se expresa como:

$z_i = a_1x_{i1} + a_2x_{i2} + ... + a_px_{ip}$ donde $i=1,...,n_1$.

Por su parte, los coeficientes $a_1, a_2,...,a_p$ pueden ser visto como indicadores de cuáles variables contribuyen más a delinear la separación.

Para verlo más claro, consideremos las poblaciones expresadas en una matriz, como la que se presenta abajo, donde las observaciones de Π_1 son los primeros n_1 renglones, mientras que los restantes n_2 renglones corresponden a Π_2. Podemos obtener las combinaciones lineales de cada población y expresarlas de la siguiente manera

$z_{1j} = a_1x_{11} + a_2x_{12} + ... + a_px_{1p}$ donde $i=1,...,n_1$ } combinación lineal para Π_1

$z_{2j} = a_2x_{21} + a_2x_{22} + ... + a_px_{1p}$ donde $i=1,...,n_2$ } combinación lineal para Π_2

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 p} \\ \hline x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n_2 1} & x_{n_2 2} & \cdots & x_{n_2 p} \end{pmatrix}$$

Los vectores de las $n_1 + n_2$ observaciones de las poblaciones $\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n_1} \end{pmatrix}$, $\mathbf{x}_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n_2} \end{pmatrix}$ se

transforman en los escalares $\mathbf{z}_1 = \begin{pmatrix} z_{11} \\ z_{12} \\ \vdots \\ z_{1n_1} \end{pmatrix}$, $\mathbf{z}_2 = \begin{pmatrix} z_{21} \\ z_{22} \\ \vdots \\ z_{2n_2} \end{pmatrix}$.

Las medias de estos escalares estarán dada por:

$$\bar{z}_1 = \frac{\sum_{i=1}^{n_1} z_{1i}}{n_1} = \mathbf{a}^t\bar{\mathbf{x}}_1 = \text{media de los valores z obtenidos de las observaciones provenientes de } \Pi_1$$

$$\bar{z}_2 = \frac{\sum_{i=1}^{n_2} z_{2i}}{n_2} = \mathbf{a}^t\bar{\mathbf{x}}_2 = \text{media de los valores z obtenidos de las observaciones provenientes } \Pi_2 \qquad (4.4)$$

Al hacer esto se logra reducir la dimensión a través de una combinación lineal que separe los grupos, sin embargo, no sólo se quiere hacer la separación de los grupos sino que al mismo tiempo se quiere que ésta sea la mayor separación posible, así, deseamos aquella combinación lineal que también proporcione la mejor separación. Es entonces, que se involucra el concepto de distancia entre grupos. Esta distancia se obtiene de la diferencia entre los centroides de cada grupo $(\bar{z}_1 - \bar{z}_2)$. Estandarizando esta distancia y tomando el cuadrado para evitar problemas con valores negativos, obtenemos la expresión $\frac{(\bar{z}_1 - \bar{z}_2)^2}{s^2}$. La mejor combinación lineal será aquella que maximice esta distancia. Así,

$$\frac{\text{Distancia al cuadrado de las medias}}{\text{La varianza de las z}} = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{(\mathbf{a}^t\bar{\mathbf{x}}_1 - \mathbf{a}^t\bar{\mathbf{x}}_2)^2}{\mathbf{a}^t\mathbf{S}_{pl}\mathbf{a}}$$

que se obtiene de los resultados (2.27) para el numerador y (2.28) para el denominador, y finalmente se puede expresar

$$\frac{\text{Distancia al cuadrado de las medias}}{\text{La varianza de las z}} = \frac{[\mathbf{a}^t(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{a}^t\mathbf{S}_{pl}^{-1}\mathbf{a}} \qquad (4.5)$$

Tomando (2.29) que es un resultado importante relacionados con la desigualdad de Cauchy-Schwartz y sustituyendo el vector **d** por nuestra diferencia de medias $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, el vector **x** por nuestro vector de coeficientes **a** y la matriz **B** por nuestra la $\mathbf{S}_{pl}$, obtenemos que el valor máximo del cociente (4.5) es igual a $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t\,\mathbf{S}_{pl}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ cuando $\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

$$\max \frac{[\mathbf{a}^t(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{a}^t\mathbf{S}_{pl}^{-1}\mathbf{a}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t\,\mathbf{S}_{pl}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

Con este resultado, la combinación lineal $\mathbf{z} = \mathbf{a}'\mathbf{x}$ nos dará la distancia máxima entre las muestras (entre sus centroides) que será cuando $\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. Así, obtenemos los que se conoce como **la función discriminante lineal de Fisher**

$$\mathbf{z} = \mathbf{a}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pl}^{-1}\mathbf{x} \qquad (4.6)$$

Este resultado, que es la solución dada por Fisher para el problema de la separación para dos poblaciones, también se puede ver como una forma para clasificar una nueva observación. La figura 4.4 presenta una gráfica que ilustra esta solución.

Con base en (4.6) se puede definir $z_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pl}^{-1}\mathbf{x}_0$ como el valor de la función discriminante para cualquier observación nueva $\mathbf{x}_0$. Por ejemplo, tomemos el punto medio entre las medias de los dos grupos:

$m = (\bar{z}_1 + \bar{z}_2)/2$ por el resultado (4.4) $m = (\mathbf{a}'\bar{\mathbf{x}}_1 + \mathbf{a}'\bar{\mathbf{x}}_2)/2$

como $\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \Rightarrow \mathbf{a}' = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pl}^{-1}$ y sustituyéndolo

$$m = \frac{1}{2}\left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_1 + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_2\right]$$

$$m = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \qquad (4.7)$$

m es el punto medio entre las medias $\bar{z}_1$ y $\bar{z}_2$ de los grupos, entonces para cualquier observación nueva $\mathbf{x}_0$ que provenga de Π_1 se espera que $z_0 \geq m$, mientras que si $\mathbf{x}_0$ proviene de Π_2 se espera que $z_0 < m$.

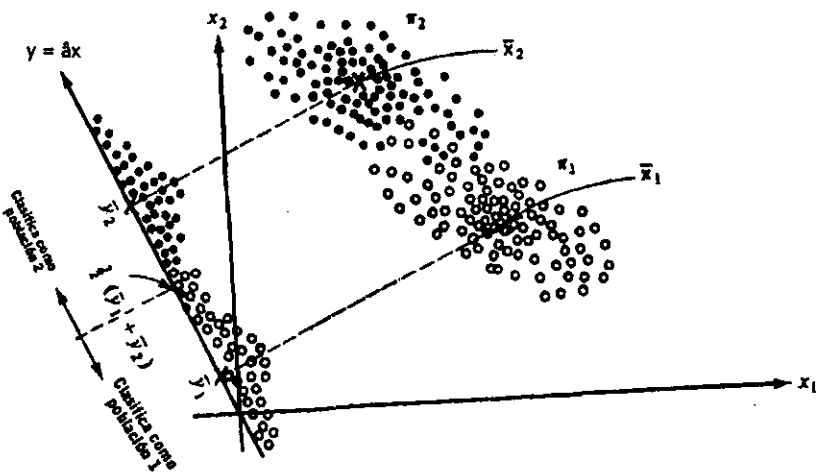

Figura 4.4 Representación pictórica del procedimiento de Fisher cuando p=2

Así, podemos dar una regla para la asignación de las observaciones:

$$\text{Asignar } \mathbf{x}_0 \text{ a } \Pi_1 \text{ si } z_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t \mathbf{S}_{pl}^{-1} \mathbf{x}_0 \geq m$$
$$\text{Asignar } \mathbf{x}_0 \text{ a } \Pi_2 \text{ si } z_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t \mathbf{S}_{pl}^{-1} \mathbf{x}_0 < m \qquad (4.8)$$

Cuando z se aplica a los valores de x ésta los proyecta sobre un nuevo eje (eje z) que separa de manera óptima los dos grupos. En la figura 4.5, el punto en que se corta el eje z con la línea que intersecta las dos elipses, que es el de máxima separación de los puntos que expresan las medias de los grupos. Este punto es el m que se presenta en el resultado (4.7)

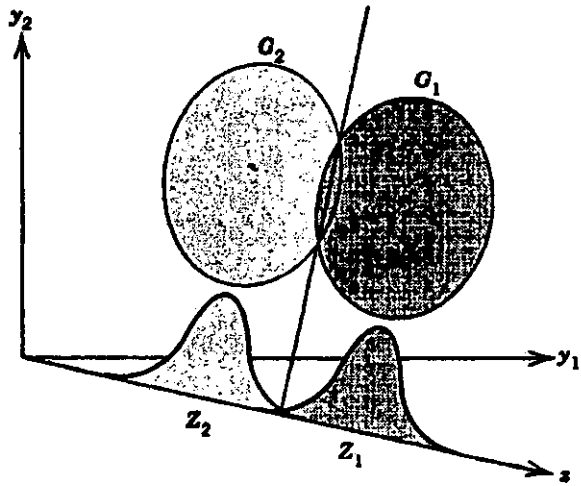

Figura 4.5 Ilustración gráfica del analisis discrimiante de dos grupos

4.2.1.1 La escala

Los coeficientes $a_1, a_2,...,a_p$ de la función discriminante proporcionan la contribución relativa que cada variable aporta al modelo para separa los grupos. Sin embargo, para que esta información sea interpretable se requiere que las observaciones x_i estén expresadas en las mismas escala y con varianzas comparables. Si las x's no cumplen lo anterior, necesitamos coeficientes $\mathbf{a}^*$ que al aplicarse estandaricen las variables. Para la i-ésima observación x_{ki} del k-ésimo grupo, podemos expresar la función discriminante en términos de sus variables estandarizadas.

$$z_i = a_1^* \frac{(x_{ki1} - \bar{x}_{k1})^2}{s_1^2} + a_2^* \frac{(x_{ki2} - \bar{x}_{k2})^2}{s_2^2} + ... + a_p^* \frac{(x_{kip} - \bar{x}_{kp})^2}{s_p^2} \qquad k=1,2 \quad i=1,...,n_k$$

Donde $\bar{\mathbf{x}}_i^t = (\bar{x}_{k1}, \bar{x}_{k2}, ..., \bar{x}_{kp})$ es el vector de medias del grupo k y s_j es la desviación estándar al interior de la muestra, de la j-ésima variable, que se obtiene con la raíz cuadrada de la j-ésima diagonal de la matriz $\mathbf{S}_{pl}$. Así los coeficientes estandarizados deben ser de la forma

$$a_j^* = s_j a_j \qquad j=1,...,p$$

En la forma vectorial, se expresa $\mathbf{a}^* = (\text{diag } \mathbf{S}_{pl})^{1/2}\mathbf{a}$

4.2.2 Pruebas de significancia

No es suficiente haber encontrado la expresión que permita separar de la mejor manera los grupos de interés, también necesitamos asegurar que la separación obtenida sea significativamente buena. Para medir la distancia entre los grupos, se toman los centroides de cada uno de ellos y se mide la distancia de éstos. Podemos, entonces, utilizar la prueba T^2 que es una prueba sobre las diferencias entre las medias de poblaciones. La aplicación de esta prueba requiere suponer poblaciones con función normal multivariada. Se pude ver que para el desarrollo de la función discriminante esta suposición no se requirió.

Algo que resulta muy conveniente, pues se usa para probar si las medias son significativamente diferentes, es que la máxima distancia relativa entre las medias de dos grupos es $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t \mathbf{S}_{pl}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. Por otro lado, hemos visto en (3.6) que la estadística T^2 se puede expresar en términos de distancia estandarizada entre los centroides de dos grupos $\bar{\mathbf{x}}_1$ y $\bar{\mathbf{x}}_2$ y que cuando la distancia es grande, T^2 también es grande. Como consecuencia, el vector de coeficientes de la función discriminante es significativamente distinto de 0.

De las propiedades mencionadas en la sección 3.2.1.1, de la prueba T^2 para dos medias,

sabemos que T^2 puede transformarse en una $F_{p,n1+n2-p-1}$.

Hay que tener en cuenta que una separación significativa no necesariamente implica buena clasificación. Independientemente de cualquier prueba de separación, se puede evaluar la eficiencia del procedimiento de clasificación. Por otro lado, si la separación no es significativa puede resultar infructuosa la búsqueda de una clasificación útil.

Capítulo V

LA APLICACIÓN

Ya con el conocimiento de lo que es el análisis discrimínate en el nivel intuitivo y teórico, estamos en posibilidad de abordar nuestro problema de clasificación y ver cómo esta técnica se puede aplicar en un problema concreto.

Más allá de lo que el sentido común nos indica, queremos verificar con base en información concreta, que hablar de pobreza extrema no es simple retórica, sino algo real. Además de crear un modelo que clasifique a las familias que viven en las localidades marginadas de zonas rurales, resulta interesante ir recogiendo las experiencias que nos aporta el recorrido para llegar a nuestro modelo.

En este capítulo se presenta el ejercicio práctico del análisis discriminante y el modelo que mejor discrimina los dos grupos que estamos considerando. Pero antes de llegar a ese modelo, fue necesario ir analizando, con cuidado, lo que nos iba diciendo la información. Nunca hay que perder de vista que los números, son número, y sin la interpretación del ser humano, no tienen vida, nada dicen. Es por esto, que más allá de ver una variable en función, únicamente, del modelo, es muy importante verla en todo su contexto, donde el modelo es una parte del todo, mas no el todo. Hay sutilizas que la estadística no nos dice si nosotros no somos lo suficiente hábiles como para preguntarle. Ella nos dará datos, pistas; nosotros tendremos que irlas juntando, buscando la lógica que rija lo que obtenemos. Hay ocasiones en que tal vez se le tenga que poner algunas trampas y observar su respuesta. Ella no se equivoca, pues sólo aporta datos, somos nosotros quienes le damos información poco confiable o interpretamos mal lo que nos muestra.

La estadística sirva para entender relaciones, pero requiere de datos para procesar. Si la alimentamos con información errónea, obtendremos relaciones, pero tal vez éstos no tengan sentido en el mundo real. Las fuentes de información son el alimento de la estadística, si éstas son

buenas, los resultados serán buenos, ya sea que encontremos o no lo que buscamos. Siempre habrán resultados que nos enseñan más sobre lo que investigamos, nos dan conocimiento y nos muestran un camino para mejorar o replantear nuestra información.

Veamos, entonces, las fuentes de información tomadas para este trabajo para luego pasar a la preparación de la información, y finalmente, su incorporación en la metodología.

5.1 Las fuentes de información

Dos de los factores indispensables para obtener resultados útiles y reales son: un planteamiento adecuado del problema y la correcta identificación de los factores que nos permitan caracterizar el fenómeno de la pobreza. Por otra parte, también se requieren fuentes de información confiables y veraces. Carece de información confiable dará resultados que no expresan una situación real. De esta forma ni el resultado final ni los resultados parciales serán de utilidad real.

Tomando en cuenta lo anterior y en el ánimo de presentar un trabajo apegado a la realidad, las fuentes de información que se emplean son principalmente dos: El índice de marginación construido por el Consejo Nacional de Población (Conapo) en 1995 y la Encuesta de Características Socioeconómicas de los Hogares 1998 (Encaseh), levantada entre abril y julio de ese año.

El índice de marginación está construido para todas las localidades reportadas en el censo de 1990 y el conteo rápido de 1995. Indica el nivel de marginación de las localidades y las clasifica definiendo 5 grados de marginación:

Grado 1 Localidades de muy alta marginación
Grado 2 Localidades de alta marginación
Grado 3 Localidades de media marginación
Grado 4 Localidades de baja marginación
Grado 5 Localidades de muy baja marginación

Por su parte, la Encaseh fue levantada por el Programa de Educación Salud y Alimentación (Progresa). El levantamiento de la encuesta se realizó a todos los hogares de las localidades donde se aplicó, entendiendo por hogar: "El conjunto de personas que viven dentro de la vivienda, unidos o no por parentesco, comparten los gastos de manutención y preparan los alimentos en la misma cocina".

Así, el análisis se enfocará sobre los hogares que viven en localidades rurales de alta, muy alta y media marginación, grado 3, 4 y 5 respectivamente. A nivel mundial se reconoce una localidad rural como aquella que cuenta con menos de 2,500 habitantes.

5.2 Consideraciones sobre los datos

El modelo supone variables continuas, independientes entre sí que se distribuyen normalmente; sin embargo, es difícil que en la realidad esto se cumpla al pie de la letra; además es común contar con variables categóricas que aportan información importante y que puede resultar conveniente tomarlas en cuenta para la discriminación.

El uso de la función discriminante de Fisher requiere que las características sean medidas cuantitativamente por lo que el uso de variables binarias o nominales debe realizarse con cuidado. Hay que considerar los aspectos que pueden incidir negativamente en la precisión de las clasificación para evitar caer en ellas.

En el caso de las variables nominales, los problemas pueden surgir cuando al asignar un valor numérico a una categoría se olvida que ésta no es métrica y se le dé un manejo como si lo fuera. No hay que olvidar que el asignar números sirve para dar una referencia a la categoría a la que pertenece un individuo y en ningún momento es una medida métrica. Si las medidas de distancias propuestas se generan únicamente a partir de medidas nominales, las propiedades del resultado métrico no estarán bien definidas.

Para cuantificar datos cualitativos generalmente se emplean técnicas analíticas, como el análisis de factores que transforman los datos en índices o puntuaciones, antes de construir la función; pero el tratamiento de estas variable depende del número de categorías que presentan. Para el análisis discriminante de dos grupos las variables dicotómicas pueden ser tratadas asignando valores de 0 y 1 a las variables nominales.

Para el caso de variables dicotómicas que expresan la presencia o ausencia de una característica, el criterio para la asignación de los valores se define con base en la existencia o ausencia de la característica. Usualmente se crea una variable numérica que toma el valor 1 si la observación posee la característica y 0 si no la posee. De esta manera, la variable es tratada como las demás en el proceso de clasificación y discriminación.

Al aplicar este tratamiento hay que ser cuidadoso en cuanto al sentido en que se relacionan las variables independientes con la dependiente. Si el planteamiento de la variable dependiente es tomar el valor 1 en caso de presencia de la característica y 0 para la ausencia de ésta; el planteamiento de las variables independientes debe ser en el mismo sentido.

Por ejemplo, para el ejercicio que aquí se plantea, el interés radica en clasificar cada una de las observaciones en uno y sólo uno de los dos grupos definidos:

Grupo 1 Familia en condiciones de pobreza extrema (poseen característica de interés)

Grupo 2 Familia en condición de pobreza (no poseen característica de interés)

Se creó la variable ***lin_pob*** para identificar a qué grupo pertenece cada observación evaluada, y esta variable tomará los valores 1 para identificar que la observación pertenece al grupo 1; y 0, si pertenecen al grupo 2. Así, a las variables dicotómicas se les asigna el valor de 1 si la presencia de la característica que se está midiendo acerca más la observación al grupo 1 y cero si la acerca más al grupo 2. De esta forma al asignar valor a la variable que refleja la posesión de un automóvil, se asignará 1, a aquellas familias que no lo tengan, y 0 a aquellas que si lo tenga.

Con base en experimentos computacionales se ha podido observar que la función discriminante de Fisher puede funcionar o no, dependiendo de las correlaciones entre las variables categóricas y las continuas. Una correlación baja en una población y alta en otra puede indicar que no es muy conveniente el uso de esta función; lo mismo ocurre cuando se presenta un cambio en el signo de la correlación entre las dos poblaciones.

Si se va a realizar una investigación donde las variables de interés son dicotómicas (0 – 1) puede usarse algún otro método de clasificación, como la regresión logísitica.

5.3 Inspección de datos y construcción de variables

Con el objetivo de determinar aquellos factores a considerar para realizar una medición de la pobreza, retomaremos la discusión que se presentó en el primer capítulo y a partir de esto iremos ubicando cada una de las variables útiles para el análisis. Algunas tendrán que construirse a partir de la información original y otras simplemente tendrán que inspeccionarse buscando detectar cualquier problema que pudiera afectar los resultados del análisis. En cualquier caso, es claro que un primer paso para tener las variables independientes, es conocer cómo vienen los datos que

proveen las fuentes de información. Pueden presentarse tres situaciones: que exista información incongruente, presencia de valores extremos o variables con distribuciones extrañas.

Aquellas observaciones donde se encontró incongruencia o falta de respuesta, para alguna variable de interés, fueron eliminadas del análisis. Un caso típico es la variable ingreso, que es difícil de captar en las entrevistas y por consecuencia presenta respuestas incongruentes o falta de ellas porque, simplemente, los entrevistados no quieren contestar.

El criterio para tratar los valores extremos, tanto de las variables a incluir en el análisis, como aquellas que se usan para generarlas, es eliminar las observaciones que caen más allá del percentil 95. No se usó el rango intercuartil, pues muchas de las variables presentan una gran cantidad de casos con valor cero y es así como se requiere la información. Por ejemplo, para la proporción de niños que trabajan por un ingreso (*pcp_trab*) que es una variable cuyos valores están contenidos en el intervalo [0,1], un alto porcentaje de los hogares tendrán valor 0, lo que es importante para la discriminación, por tanto, no es posible prescindir de estas observaciones. Sin embargo, la variable del total de niños que trabajan por un ingreso —que se usa para crear *pcp_trab*— sí fue tratada previa la generación de pcp_trab. Lo que se hizo fue excluir las observaciones que están más allá del percentil 95.

Para el caso de variables con distribuciones asimétricas aún después de la eliminación de los valores extremos, se optó por alguna transformación. Como ejemplo, volvamos al caso de pcp_trab. Se decidió no trabajar con la variable del total de niños que trabajan, y en su lugar se generó pcp_trab que tiene una distribución más estable y contiene la información que nos interesa.

Considerando estos criterios fueron construidas las variables de interés, mismas que a su vez se definieron con base en la discusión sobre la pobreza y los factores que inciden en ellas, especialmente aquellos que propician su existencia.

La encuesta proporciona información sobre los servicios con que cuentan las viviendas, el nivel educativo de cada uno de sus integrantes, el trabajo que desempeñan, si hay o no presencia de discapacitados, ingresos de cada uno de los integrantes del hogar, quién habla lengua indígena, equipamiento de la vivienda, animales que posee el hogar y número de hectáreas de tierra.

El análisis discriminante parte de la existencia de un número de poblaciones y su objetivo básico es producir una regla de asignación que permita predecir la población a la que pertenece una nueva observación, clasificándola en una y sólo una de las poblaciones definidas de antemano. En nuestro caso, tenemos dos poblaciones:

Población 1 Familia en condiciones de pobreza extrema

Población 2 Familia en condición de pobreza

Basándonos en la información del ingreso per cápita de cada familia y considerando el nivel de ingreso mínimo que se requiere para cubrir las necesidades básicas de un individuo podemos definir las dos poblaciones de interés. Este nivel mínimo se establece de acuerdo al costo de la canasta básica que a mediados de 1998 era de $339.70 mensuales.

El ingreso per cápita de la familia no toma en cuenta los ingresos de los menores de 15 años, pues se esperaría que a esa edad los niños estén dedicados a estudiar. Si se tomará en cuenta este ingreso, se pensaría que la familia tiene mejores condiciones de vida cuando por el contrario, ante la necesidad en que viven, deciden incorporar al trabajo productivo a los menores impidiéndoles recibir una educación que podría ayudarlos, a que en el futuro, mejoren sus condiciones de vida.

Se construye entonces la variable ***lin_pob*** que tomará los siguientes valores:

lin_pob = 0, si el ingreso mensual per cápita > $339.7 Familias en pobreza extrema

lin_pob = 1, si el ingreso mensual per cápita ≤ $339.7 Familias en pobreza

A partir de esto, la función discrimínate y con ello la regla de clasificación se obtiene incorporando al análisis las características de cada familia.

Como ya se ha mencionado, el ingreso es una variable difícil de captar en las encuestas, es por eso que para evitar ambivalencias, únicamente se consideró para el análisis aquellos hogares que cuentan con información completa sobre ingresos y que no presenten incongruencias en lo declarado.

Por otra parte, se definió un conjunto de variables independientes que, en principio, todas fueron consideradas en el modelo. A través del procedimiento de discriminación se fue encontrando el conjunto idóneo de estas variables que conformaron modelo final. Para definir las variables que interesa considerar en el modelo se identificaron cuáles son aquellos aspectos podían cubrirse con la información disponible y con base en lo discutido en el primer capítulo.

Comencemos con lo relativo a la autonomía personal. Un factor que influye en la condición de pobreza de un hogar es la presencia de personas con alguna discapacidad. De la información sobre personas ciegas, mudas, sordas, con problemas mentales, necesita ayuda, se crea la variable ***dicapacl*** que identificar a los hogares que tienen a algún miembro del hogar discapacitado. Esta variable toma valores de 1 cuando hay discapacitados y 0 cuando no los hay.

Dentro de las variables sobre las características del jefe del hogar, se crea la variable sobre la edad de jefe: ***edad_jef***.

En principio, la influencia que los padres ejercen sobre los hijos para que asistan a la escuela, está relacionada con su propio nivel de escolaridad. A mayor escolaridad más es la importancia que los padres dan a la escuela y por tanto la motivación que les dan a los hijos para asistir. Para captar esta información, se crea la variable ***esc_jefe*** que son los años de escuela cursados por el jefe del hogar considerando hasta el último grado aprobado.

También relacionado con la escolaridad, aquellos hogares que tienen niños entre 5 y 15 años que no asisten a la escuela se consideran en desventaja, ya que estos niños tienen mayor posibilidad de reproducir la pobreza de sus padres al no tener acceso a la educación formal. Así tenemos la variable ***noasi1*** que es el número de niños que no asisten a la escuela. A partir de ésta se obtiene la proporción de niños entre 5 y 15 años que no asisten a la escuela (variable ***prp_nasi***), que es el cociente dado por:

$$\textbf{prp\_nasi} = \frac{\text{Total de niños entre 5 y 15 años que no asisten}}{\text{Total de niños entre 5 y 15 años}}$$

Se toman los niños a partir de los 5 años pues la educación preescolar es importante en el desarrollo de sus habilidades y ayuda a un mejor desempeño del niño en los años posteriores. En cuanto al límite de los 15 años, es la edad a que un individuo termina la educación media.

Caso similar ocurre para aquellos hogares que cuentan con niños entre 8 y 15 años que trabajan por un ingreso. En este caso se considera la edad de 8 años, no por otra razón, sino que, la encuesta sólo cuenta con información de actividades económicas de los individuos a partir de esa edad; el límite de los 15 años es debido a que se espera que un individuo, como mínimo, concluya educación media. Dentro este rango de edad, se esperaría que los niños asistan a la escuela. Si trabajan percibiendo ingresos, es lógico pensar que su rendimiento y aprovechamiento escolar se verá afectado de manera negativa —esto en el caso de que asistan a la escuela— ya que muchas veces se da preferencia al trabajo sobre la asistencia a la escuela, terminando por abandonándola. Con esto se mantienen en el mismo esquema de posibilidades mínimas para acceder a una mejor perspectiva de vida y una mejor remuneración por el trabajo que desempeñan. La variable ***trab15*** es el número de niños que trabajan por un ingreso y de aquí se desprende la variable de la proporción de niños entre 8 y 15 años que no asisten a la escuela ***prp_trab***

$$\textbf{prp\_trab} = \frac{\text{Total de niños entre 8 y 15 años que trabajan por un ingreso}}{\text{Total de niños entre 8 y 15 años}}$$

Cuando no hay niños en el hogar estas dos variables toman el valor de 0.

De la información sobre el de total de personas en el hogar y el número de cuartos de la vivienda se obtiene un índice que representa el nivel de hacinamiento en que vive la familia, y se crea la variable ***i_hacina*** = $\frac{\text{Total de niños entre cuartos}}{\text{Total de personas}}$, que es el número de personas por habitación.

Otro índice, es el que se obtiene para representar el número de personas que dependen de cada preceptor: índice de dependencia, que está dado en la variable ***i_depen*** y se calcula como el cociente ***i_depen*** = $\frac{\text{Total de personas del hogar}}{\text{Total de perceptores}}$.

Mucho se ha hablado sobre que los grupos indígenas son los más marginados de los marginados. Para identificar las familias indígenas nos basamos en el hecho de que algún miembro del hogar hable lengua indígena. La variable par representar esto es ***lengua_i*** que toma valor de 0 si nadie habla lengua indígena y vale 1 si existe algún miembro en el hogar que hable lengua indígena.

Otro grupo de variables son las que se crean para representan las características de la vivienda en lo referente a los servicios con que ésta cuenta y los materiales de que está construida.

Aspecto importante es el material del piso de la vivienda, que tiene una estrecha relación con las condiciones de higiene y salud de las familias. Contar con piso de tierra representa una condición más desventajosa que tener un piso de material firme, ya sea cemento o losetas. Para caracterizar esta condición se crea la variable ***m_piso*** que toma valor de 0 cuando se tiene algún material firme y 1 cuando el piso es de tierra.

Los techos y paredes también son clasificados en dos categorías, asignando valor de 0 a los materiales firmes y 1 a los que brindan menor resguardo. Para el caso de los techos la variable es ***m_techos*** y los materiales que se consideran firmes son: teja, losa de concreto, tabique, ladrillo y block. La categoría de blandos comprende: cartón, tela, hule, lámina de asbesto, cartón o metálica; palma, carrizo, bambú, madera, fibra de vidrio y plástico.

Para las paredes se genera la variable ***m_pared*** y los materiales que considera firmes son: la madera, concreto, vidrio, adobe, tabique, ladrillo, block, cemento y cantera; mientras que como blandos están el cartón, tela, hule, lámina de asbesto, de cartón o metálica; palma, carrizo, bambú, madera, fibra de vidrio, plástico y tabla roca.

El servicio de luz se representa con la variable ***serv_luz*** que será 0 en caso de contar con el servicio y 1 cuando no sea así.

La información sobre los servicios con que cuenta el hogar, tales como, existencia de agua entubada en el terreno o en el interior de la vivienda, además de lo referente a la existencia de baño o letrina identificando si estos cuentan o no con agua corriente, se expresan en una variable: ***aguabaio*** que se construye con información de una serie de preguntas sobre servicios de agua entubada y baño, y consta de 9 categorías:

1 = Sin agua en el terreno y sin baño
2 = Sin agua en el terreno y baño sin agua
3 = Con agua en la vivienda y sin baño
4 = Con agua en el terreno y baño sin agua
5 = Con agua en terreno, sin agua en vivienda y baño con agua
6 = Agua en el terreno, sin agua en la vivienda y sin baño
7 = Con agua en la vivienda y baño sin agua
8 = Con agua en la vivienda y baño con agua
9 = Sin agua en el terreno o la vivienda y baño con agua

Para considerar que una familia cuenta con los servicios básicos para una adecuada higiene debe contar con servicio de agua potable en su vivienda, ya sea en el interior o en el terreno en que está construida; y tener baño al que llegue el agua. Con base en este criterio, se generó la variable ***s_aguaba*** que toma valores de 0 en caso de contar con los servicios —categorías 5, 7 y 8 de *aguabaio*— y 1 cuando no es así —categorías 1, 2, 3, 4 y 6—. Los casos que caen en la categoría 9 se eliminan del análisis por ser incongruentes.

Otra parte de la información se refiere a las pertenencias con que cuenta el hogar, esto abarca tanto los enseres domésticos, como la posesión de animales, tierras y vehículos.

De animales se tiene información sobre número de caballos; mulas o burros; bueyes; reses o vacas; chivos, borregos o cabras; gallinas o guajolotes; cerdos y conejos. La forma de adecuar esta información es homogenizando las unidades en que se contabilizan los animales. La fuente de información ya presenta una variable en unidades homogéneas.

De la referente a la posesión de tierras para trabajar, se cuenta con información sobre el número de hectáreas totales que se cultivaron en los últimos 12 meses ***tot_hect***.

Para representar la presencia o no de enseres domésticos se crea una variable por cada enser, que toma valor de 0 si el hogar cuenta con el enser, y en caso de no ser así, toma valor de 1. Se generan once variables (una por cada una): ***licuador, refri, estufgas, cal_agua, radio, modular, tele, vhs, lavaropa, ventila*** y ***vehículo***, cuyos nombres hablan por sí solos.

Habiendo revisado las variables, generado aquellas que son necesarias para el ejercicio y tratado las que lo requerían, obtenemos una lista de 26 variable independientes más la variable *lin_pob* que identifica la población a que pertenece originalmente cada observación. El cuadro 5.1 presenta el listado de estas variables. En un principio, todas las variables fueron consideradas para definir la función discriminante.

Cuadro 5.1 Listado de variables que se consideran en el modelo

	Variable	Descripción	Tipo
	lin_pob	Línea de pobreza definida con base en el ingreso per-cápita de los hogares	Real
1	Edad_jef	Edad del jefe del hogar	Real
2	esc_jefe	Años de escolaridad del jefe del hogar	Real
3	pcp_nasi	Proporción de niños entre 5 y 15 años que no asisten a la escuela	Real
4	pcp_trab	Proporción de niños entre 8 y 15 años que trabajan	Real
5	i_hacina	Índice de hacinamiento	Real
6	i_depen	Índice de dependencia	Real
7	discapac	Presencia de discapacitados en el hogar	Dicotómica
8	lengua_i	Algún miembro habla lengua indígena	Dicotómica
9	tot_anim	Número total de animales en unidades homogéneas	Real
10	tot_hect	Número total de hectáreas para trabajar	Real
11	m_piso	Material del piso de la vivienda	Dicotómica
12	m_techo	Material del techo de la vivienda	Dicotómica
13	m_pared	Material de las paredes de la vivienda	Dicotómica
14	s_aguaba	Servicio de agua y baño	Dicotómica
15	serv_luz	Servicio de luz eléctrica	Dicotómica
16	licuador	Licuadora	Dicotómica
17	refri	Refrigerador	Dicotómica
18	estufgas	Estufa de gas	Dicotómica
19	cal_agua	Calentado de agua	Dicotómica
20	radio	Radio	Dicotómica
21	modular	Tocadiscos o modular	Dicotómica
22	tele	Televisor	Dicotómica
23	vhs	Videocasetera	Dicotómica
24	lavaropa	Lavadora de ropa	Dicotómica
25	ventila	Ventilador eléctrico	Dicotómica
26	vehiculo	Vehículo	Dicotómica

Cada una de las observaciones pertenece a uno de los dos grupos que se identifican con la variable *lin_pob* cuyos valores están dados por:

lin_pob = 0 Familia en pobreza extrema

lin_pob = 1 Familia en pobreza

Se puede observar que contamos con un importante número de variables dicotómicas a las que se asignó valores 0 o 1 considerando lo ya comentado sobre la congruencia que debe existir entre la construcción de la variable dependiente (también dicotómica) y las variables independientes dicotómicas. Es la pobreza extrema, la característica de interés a la que se le ha asignado el valor de 1. Las variables independientes se han construido manteniendo esta relación.

De acuerdo a lo mencionado anteriormente, debido a la presencia de estas variables, es conveniente analizar las correlaciones entre las variables reales y las dicotómicas, dentro de cada grupo definido para asegurar que no se presenten correlaciones que difieran de manera importante entre uno y otro grupo, o que tengan cambio en el sentido de la correlación.

5.3.1 Correlaciones

En el análisis de las correlaciones de cada grupo (*lin_pob* = 0 y *lin_pob* = 1) no se observan grandes diferencias entre uno y otro. Haciendo un recorrido por toda la matriz de correlaciones, los niveles de estos fluctúan en el intervalo (-.250,.250), y en general los niveles de correlación son bajos.

La correlación más alta se da entre las variables techos y paredes; el grupo de hogares pobres tiene un 0.502 de correlación, y el grupo de hogares pobres extremos .498. Poco por debajo, está la correlación entre los años de escolaridad del jefe (*escolar*) y su edad (*edad_jef*) con -0.435 y –0.495 para hogares pobres y hogares no pobres, respectivamente. Por debajo de éstas se encuentran las correlaciones de la edad del jefe, con los índices de hacinamiento y dependencia; así como, las que se presentan en el caso de los enseres, que son las más numerosas. El cuadro 5.2 muestra las 25 correlaciones más altas de un total de 325 que se tuvieron con las 26 variables.

Finalmente tenemos los casos donde se aprecia un cambio de signo en la correlación (cuadro 5.3). Sin embargo, en algunos casos la prueba de significancia para probar que correlaciones son distintas a 0, resulta significativa. Aún así, los niveles de las correlaciones son muy bajos, muy cercanos a cero. En estos casos, se optó por no excluir, de momento, estas variables y esperar par ver su comportamiento y qué tanto afectan al momento de realizar el análisis para definir la función discriminante.

Cuadro 5.2 Correlaciones más altas

Variables correlacionadas			Grupo de hogares pobres	Grupo de hogares solventes
Edad_jef	v.s.	esc_jefe	-0.435	-0.465
Edad_jef	v.s.	l_depend	-0.346	-0.347
l_depend	v.s.	l_hacina	0.344	0.285
pcp_trab	v.s.	pcp_nasi	0.298	0.343
tot_hect	v.s.	t_animal	0.287	0.290
m_techo	v.s.	m_pared	0.498	0.502
m_techo	v.s.	estufgas	0.292	0.342
m_techo	v.s.	refri	0.276	0.332
m_techo	v.s.	ventila	0.296	0.324
serv_luz	v.s.	tele	0.315	0.310
licuador	v.s.	refri	0.375	0.449
licuador	v.s.	estufgas	0.418	0.513
licuador	v.s.	ventila	0.388	0.445
licuador	v.s.	lavaropa	0.338	0.390
licuador	v.s.	tele	0.311	0.377
refri	v.s.	estufgas	0.397	0.482
refri	v.s.	ventila	0.330	0.384
refri	v.s.	tele	0.267	0.324
refri	v.s.	lavaropa	0.290	0.368
estufgas	v.s.	tele	0.267	0.333
estufgas	v.s.	lavaropa	0.366	0.419
estufgas	v.s.	ventila	0.380	0.451
tele	v.s.	radio	0.353	0.382
tele	v.s.	ventila	0.369	0.412
Lavaropa	v.s.	ventila	0.307	0.347

Para todos estos casos es significativa la prueba de la correlación distinta a 0.

SPSS brinda la opción de un procedimiento para análisis discriminante y éste ya comprende una serie de pruebas de interés, resultados estadísticos que puede ser de ayuda y la opción de realizar la selección de variables a través procedimiento de selección por pasos (stepwaise). Ya se irán explicando los resultados que se obtienen con el uso de este procedimiento.

Cuadro 5.3 Correlaciones con signo distinto para cada población

Variables correlacionadas			Grupo de hogares pobres	Grupo de hogares solventes
Edad_jef	v.s.	m_techo	-0.034	0.002
Edad_jef	v.s.	ventila	-0.014	0.031
i_depend	v.s.	serv_luz	0.006	-0.038
i_hacina	v.s.	serv_luz	0.042	-0.028
i_hacina	v.s.	tele	0.051	-0.076
i_hacina	v.s.	radio	0.019	-0.057
i_hacina	v.s.	discapaci	-0.007	0.033
t_animal	v.s.	lavaropa	-0.021	0.009

5.4 Selección de las variables para un modelo óptimo

En el análisis discriminante lo importante es encontrar la combinación de variables adecuada que nos lleve a obtener la función que mejor separe los grupos que se estudian. Puesto que estamos suponiendo la existencia de dos poblaciones distintas, nos interesa seleccionar aquellas variables que más discriminen entre una y otra población. Lo que equivale a tomar las variables que son significativamente diferentes. Si no existe diferencia entre las variables, no tendría caso intentar realizar una separación de los grupos con base en la información que éstas aportan, pues sería un trabajo infructuoso.

Así, un primer paso en la selección de las variables es probar si, efectivamente, al tomar las variable de uno y otro grupo, éstas son diferentes de manera significativa. Se realiza, entonces, una prueba sobre la diferencia de medias de las variables.

En este primer paso se considera a todas las variables, pues nos interesa saber cuáles de ellas conviene tener en cuenta para el proceso de discriminación. El cuadro 5.4 muestra el resultado que se obtiene para cada una de las variables al efectuar la prueba.

Cuadro 5.4 **Prueba de igualdad de medias de los grupos**

	Variables	Lambda de Wilks	F	df1	df2	Sig.
1	EDAD_JEF	0.98431	298.4874	1	18731	0.00000
2	ESC_JEFE	0.99864	25.4589	1	18731	0.00000
3	PCP_NASI	0.99375	117.8271	1	18731	0.00000
4	PCP_TRAB	0.99680	60.0678	1	18731	0.00000
5	I_DEPEN	0.85403	3,201.4166	1	18731	0.00000
6	I_HACINA	0.93687	1,262.0832	1	18731	0.00000
7	DISCAPAC	0.99976	4.5837	1	18731	0.03229
8	LENGUA_I	0.99952	8.9491	1	18731	0.00278
9	TOT_ANIM	0.99858	26.6299	1	18731	0.00000
10	TOT_HECT	0.99720	52.5751	1	18731	0.00000
11	M_PISO	0.99996	0.6730	1	18731	0.41203
12	M_TECHO	0.99164	157.8425	1	18731	0.00000
13	M_PARED	0.99287	134.4678	1	18731	0.00000
14	S_AGUABA	0.99361	120.4030	1	18731	0.00000
15	SERV_LUZ	0.99901	18.5235	1	18731	0.00002
16	LICUADOR	0.98123	358.2576	1	18731	0.00000
17	REFRI	0.98557	274.3103	1	18731	0.00000
18	ESTUFGAS	0.97711	438.8160	1	18731	0.00000
19	CAL_AGUA	0.99921	14.8970	1	18731	0.00011
20	RADIO	0.99649	65.9180	1	18731	0.00000
21	MODULAR	0.99284	135.0435	1	18731	0.00000
22	TELE	0.99298	132.3605	1	18731	0.00000
23	VHS	0.99234	144.4959	1	18731	0.00000
24	LAVAROPA	0.98730	240.8960	1	18731	0.00000
25	VENTILA	0.98761	234.9272	1	18731	0.00000
26	VEHICULO	0.99525	89.4197	1	18731	0.00000
27	VEHICULO	0.99525	89.4197	1	18731	0.00000

La variable F y el valor de significancia (columnas 3 y 6 respectivamente) son resultado de la prueba de *un criterio de clasificación*[1] (de una vía), calculado para cada variable de manera individual. En este caso en que se manejan dos grupos, la estadística F es equivalente a la estadística t de la prueba t para *varianza combinada de dos medias*[2].

[1] En inglés se llama *one-way-anova*

[2] En inglés se le nombra *two-sample-pooled variance*

La Λ de Wilks también aporta información respecto a la diferencia entre los grupos. Mientras la F es el cociente de la variabilidad al interior de los grupos, la Λ de Wilks es el cociente de la suma de cuadrados entre los grupos y la suma de cuadrados total. El rango de valores de Λ va de 0 a 1, así, valores pequeños indican diferencias entre grupos, mientras que valores cercanos a 1 indican que no hay diferencia.

Para nuestro ejercicio, la única variable que no es significativa ni al 1% ni al 5% es *m_piso*. Esto indica que la variable *m_piso* no va a aportar información para la discriminación, por lo que no la incluiremos para el análisis. Los casos de *discapac* y *lengua_i*, pese a no ser significativas al 1%, si lo son al 5%, por lo que permanecen en el ejercicio.

Antes de proseguir, vale la pena hacer una comparación de las características promedio de las familias en cada grupo (ver cuadro 5.5).

Este cuadro muestra los niveles que tienen, en promedio, cada una de las características medidas. En la columna 'Total' se presentan los niveles promedio del total de individuos; en la columna 'No extremos', los promedios del grupo de familias en condición de pobreza; en la 'Extremos', los promedios de las familias clasificadas en condiciones de pobreza extrema.

Otro punto importante que se debe tomar en cuenta es el de las variables que en la función discrimínate presenta cambio de signo. Si se diera el caso, hay que descartar del análisis la variable.

Tomando en consideración todas las posibles variables, salvo *m_piso* que no presenta diferencias significarivas entre uno y otro grupo, se realizó una primera corrida para definir la función discrimínate. El fin era observar el comportamiento de cada variable ante la presencia de todas las demás. En ese momento se utilizó el método de estimación simultanea que se presentó en el capítulo IV.

Se obtuvo la función discriminante (primer modelo) que incluye todas las variables. El cuadro 5.6, que se presenta más adelante, muestra la función discriminante obtenida: los coeficientes estandarizados de cada variable. Se observa que las variables *tot_anim*, *tot_hect*, *m_pared*, *serv_luz* y *lengua_i* presentan el problema del signo. Cabe señalar que en los coeficientes no estandarizados la situación es la misma.[3]

[3] Más adelante, cuando se presente el valor de los centroides, se dará una explicación detallada del cambio de signo.

Cuadro 5.5 Características de los hogares

Clasificación original (canasta básica)

	Total	No extremos	Extremos
Hogar			
Personas en el hogar [2]	4.76	3.84	5.04
Edad del jefe [2]	43.75	48.63	42.65
Años de escolaridad del jefe [2]	3.16	3.25	3.10
Habla lengua indígena el jefe del hogar [1]	93.20	92.30	93.80
Niños de 0 a 11 años [2]	1.42	0.63	1.68
Niños que no asisten a la escuela [2]	18.46%	17.63%	18.56%
Proporción de niños que trabajan por un ingreso [2]	9.40%	5.36%	10.84%
Con discapacitados [1]	3.70	3.90	4.00
Índice de dependencia [2]	2.07	0.93	2.42
Índice de hacinamiento [2]	3.62	2.58	3.91
Servicios			
Agua entubada dentro de la vivienda o en el terreno [1]	86.40	87.90	85.80
Baño con agua corriente [1]	9.40	14.40	7.50
Agua dentro de la vivienda y baño con agua corriente [1]	6.80	11.00	5.20
Luz eléctrica [1]	89.80	91.20	89.30
Piso de tierra [1]	11.40	10.10	11.20
Techo firme [1]	30.80	38.50	28.00
Cuartos por vivienda [1]	1.62	1.72	1.58
Pertenencias y propiedades			
Ucuadora [1]	27.00	37.10	23.30
Refri [1]	19.60	28.00	16.40
Estufa de gas [1]	17.00	26.60	13.50
Calentador de agua	3.80	4.90	3.40
Radio [1]	53.20	57.60	51.50
Tocadiscos [1]	8.40	12.70	6.80
Televisión [1]	58.40	65.10	55.90
Videocasetera [1]	4.20	7.20	3.10
Lavadora [1]	12.50	18.50	10.30
Ventilador [1]	33.10	42.70	29.60
Vehículo [1]	3.50	6.10	1.30
Poseen tierras 1	45.30	38.20	47.90
Número total de animales de trabajo [2]	3.60	2.90	3.90
Total de animales (unidades homogéneas) [2]	164.65	145.54	171.70
Total de hectáreas (unidades: predios de temporal) [2]	1.08	0.94	1.36
4. Ingresos			
Ingreso mensual per cápita [1]	294.25	622.50	173.09
Porcentaje de observaciones		27.0%	73.0%

[1] Porcentaje

[2] Promedio

Cuadro 5.6

Coeficiente canónicos estandarizados

	Variables	Coeficientes		Variables	Coeficientes
1	EDAD_JEF	0.0047	14	SERV_LUZ	-0.0472
2	ESC_JEFE	-0.1296	15	LICUADOR	0.0952
3	PCP_NASI	0.0101	16	REFRI	0.0238
4	PCP_TRAB	0.1718	17	ESTUFGAS	0.1061
5	I_DEPEN	0.8707	18	CAL_AGUA	0.0185
6	I_HACINA	0.1374	19	RADIO	0.0553
7	DISCAP AC	0.0553	20	MODULAR	0.0796
8	LENGUA_I	-0.0020	21	TELE	0.0672
9	TOT_ANIM	0.0631	22	VHS	0.0708
10	TOT_HECT	0.0438	23	LAVAROPA	0.0784
11	M_TECHO	0.0338	24	VENTILA	0.0223
12	M_PARED	-0.0178	25	VEHICULO	0.0863
13	S_AGUABA	0.0580			

Así, estas cinco variables se eliminaron del análisis. Las restantes 20 aún serían analizadas para determinar cuáles conformarían el modelo óptimo. De este análisis se obtuvo un segundo modelo para el cual se presentan y explican los cuadros de resultados que arroja el SPSS. En este punto pasamos a una siguiente etapa en la selección de las variables, ya que de aquí en adelante, esta selección se basa en los resultados obtenidos de las pruebas estadísticas la Λ de Wilks (3.9) y F que se presentan en la sección 3.2 del capítulo III.

Para seleccionar las variables se puede aplicar el método de selección para adelante, selección para atrás o el de selección por pasos que es la combinación de los dos anteriores. Éste último es el que se utilizó por ser el más completo.[4]

Primero se aplica la selección hacía adelante. Partiendo del hecho de que ninguna variable ha sido seleccionada y con base en los resultados de las estadísticas, se van seleccionando las variables. El cuadro con el título "Variables Not in the Análisis"[5], que se despliega abajo es una de las salidas que proporciona SPSS. En ella se despliegan las estadísticas F y Λ de cada una de las variables que aún no han sido incluidas en el análisis. Las estadísticas se calculan para cada paso

[4] El análisis discriminante de SPSS tiene la opción de aplicar este método.
[5] Sólo se despliega el resultado de los primero dos pasos y del último. Se seleccionaron 17 variables. Contando desde el paso 0, el total de pasos fue 18.

del procedimiento. El cuadro despliega los valores de las estadísticas obtenidas para cada variable un cada uno de los pasos.

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	EDAD_JEF	1.000	1.000	591.753	0.985
	ESC_JEFE	1.000	1.000	74.357	0.998
	PCP_NASI	1.000	1.000	285.843	0.993
	PCP_TRAB	1.000	1.000	115.909	0.997
	I_DEPEN	1.000	1.000	6466.828	0.858
	I_HACINA	1.000	1.000	2743.109	0.934
	DISCAPAC	1.000	1.000	16.760	1.000
	M_TECHO	1.000	1.000	402.435	0.990
	S_AGUABA	1.000	1.000	285.026	0.993
	LICUADOR	1.000	1.000	749.862	0.981
	REFRI	1.000	1.000	664.745	0.983
	ESTUFGAS	1.000	1.000	965.311	0.976
	CAL_AGUA	1.000	1.000	42.277	0.999
	RADIO	1.000	1.000	118.407	0.997
	MODULAR	1.000	1.000	354.065	0.991
	TELE	1.000	1.000	268.065	0.993
	VHS	1.000	1.000	330.964	0.992
	LAVAROPA	1.000	1.000	476.301	0.988
	VENTILA	1.000	1.000	601.766	0.985
	VEHICULO	1.000	1.000	314.129	0.992
1	EDAD_JEF	0.883	0.883	10.043	0.858
	ESC_JEFE	0.967	0.967	474.188	0.848
	PCP_NASI	0.997	0.997	140.032	0.855
	PCP_TRAB	0.987	0.987	342.508	0.850
	I_HACINA	0.890	0.890	634.463	0.844
	DISCAP	0.999	0.999	39.805	0.857
	M_TECHO	0.998	0.998	484.623	0.847
	S_AGUABA	0.999	0.999	331.354	0.851
	LICUADOR	0.999	0.999	765.216	0.841
	REFRI	1.000	1.000	629.422	0.844
	ESTUFGAS	0.998	0.998	1006.841	0.836
	CAL_AGUA	1.000	1.000	39.732	0.857
	RADIO	1.000	1.000	136.312	0.855
	MODULAR	1.000	1.000	337.047	0.851
	TELE	0.997	0.997	359.012	0.850
	VHS	0.999	0.999	363.362	0.850
	LAVAROPA	0.998	0.998	537.724	0.846
	VENTILA	0.999	0.999	642.935	0.844
	VEHICULO	0.997	0.997	424.734	0.849
17	EDAD_JEF	0.691	0.627	0.027	0.808
	PCP_NASI	0.876	0.627	1.442	0.808
	CAL_AGUA	0.949	0.627	1.199	0.808

El procedimiento arranca incluyendo la variable que presenta las medias más diferentes, lo que se decide con base en los resultados del paso 0. La variable que tiene la diferencia mayor para la prueba de diferencia de medias (la que arroja la estadística F más grande), es la que se selecciona. En este paso *i_depend* es la variable seleccionada.

En el paso 1, nuevamente se calculan las estadísticas para todas las *variables que no están en el análisis*, es decir, las que no han sido seleccionadas. Aquí *estufgas* es la variable seleccionada con la F mayor y es la que se selecciona.

El procedimiento se repite hasta que ninguna de las variables, que no han sido seleccionadas (que no están en el análisis) alcanza a tener una F significativa. Para nuestro ejercicio, esto sucede en el paso 17. Se puede ver que en este paso ya ninguna de la variable tiene una F que le permita entrar al modelo. El valor de F mínimo para seleccionar una variable es 3.4.

Con esto concluye la parte de la selección para adelante. Falta el procedimiento de selección hacia atrás que es la misma idea, pero en sentido inverso. En la selección hacía atrás se parte de que todas las variables están seleccionadas y paso a paso se van eliminando aquellas que tienen una estadística F mayor que cierto valor dado. Sin embargo, el procedimiento hacia atrás sufre algunos ajustes cuando forma parte del método de selección por paso (stepwaise).

Las variables se analizan conforme van siendo seleccionas, esto es, en lugar de comenzar con todas e ir eliminando, aquí lo que se hace es que se van analizando conforme van siendo seleccionadas. En el cuadro "Variables in the Análisis"[6] se muestra la estadística calculada una vez que la nueva variable ha sido incorporada.

Variables in the Analysis

Step		Tolerance	Min. F to Remove	Wilks' Lambda
1	I_DEPEN	1.000	6466.828	
2	I_DEPEN	0.998	6513.930	0.976
	ESTUFGAS	0.998	1006.841	0.858
3	I_DEPEN	0.879	4473.462	0.924
	ESTUFGAS	0.954	719.618	0.844
	I_HACINA	0.850	349.819	0.836
4	I_DEPEN	0.847	4701.427	0.924
	ESTUFGAS	0.928	571.315	0.837
	I_HACINA	0.845	304.791	0.831
	ESC_JEFE	0.928	209.302	0.829

[6] Sólo se despliegan los primeros cuatro pasos de un total de 17.

En el primer paso sólo se tiene la variable que fue seleccionada en el paso 0 de la selección hacía adelante, *i_depend*. Así, con cada nueva variable que se integra al grupo de las seleccionada, se aplica la prueba estadística F.

La selección continúa incluyendo en cada paso la variable con la mayor diferencia entre sus medias (selección hacia delante), a la vez que verifica si existe alguna variable, que ante la entrada de la nueva variable, tenga que ser eliminada del grupo seleccionado. En la sección 4.1.1 del capítulo 4 se explica este método.

El resultado que arrojó el ejercicio fue: 17 variables seleccionadas de las 20 originales. Las tres restantes no alcanzaron los niveles de significancia necesarios para ser seleccionadas. Ninguna de las variables seleccionas en algún momento tuvo que ser eliminadas ante la presencia de alguna nueva variable.

El cuadro "Variables Entered/Removed" que se despliega a continuación muestra el resumen del procedimiento de selección por pasos.

Variables Entered/Removed [a,b,c,d]

		Wilks' Lambda							
						Exact F			
Step	Entered	Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	I_DEPEN	.858	1	1	39050.000	6466.828	1	39050.000	.000
2	ESTUFGAS	.836	2	1	39050.000	3820.120	2	39049.000	.000
3	I_HACINA	.829	3	1	39050.000	2686.142	3	39048.000	.000
4	ESC_JEFE	.825	4	1	39050.000	2077.679	4	39047.000	.000
5	PCP_TRAB	.821	5	1	39050.000	1708.076	5	39046.000	.000
6	LICUADOR	.817	6	1	39050.000	1459.454	6	39045.000	.000
7	VEHICULO	.814	7	1	39050.000	1276.563	7	39044.000	.000
8	MODULAR	.812	8	1	39050.000	1129.849	8	39043.000	.000
9	TELE	.811	9	1	39050.000	1009.984	9	39042.000	.000
10	VHS	.810	10	1	39050.000	914.047	10	39041.000	.000
11	DISCAPAC	.810	11	1	39050.000	834.373	11	39040.000	.000
12	S_AGUABA	.809	12	1	39050.000	767.634	12	39039.000	.000
13	LAVAROPA	.809	13	1	39050.000	710.615	13	39038.000	.000
14	VENTILA	.808	14	1	39050.000	661.040	14	39037.000	.000
15	M_TECHO	.808	15	1	39050.000	617.802	15	39036.000	.000
16	RADIO	.808	16	1	39050.000	579.841	16	39035.000	.000
17	REFRI	.808	17	1	39050.000	546.203	17	39034.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 40.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Contando de izquierda a derecha, la tercera columna da el resultado de la selección por pasos. Son los valores de Λ con los que las variables fueron seleccionadas. Estos valores son los mismos que los del cuadro "Variable not in the análisis". La séptima columna es la F calculada para desechar las variables que ya no funcionan para el modelo. Es el resultado de la selección hacía atrás, de hecho, estos valores son los mismos que presenta el cuadro "Variables in the analisys".

Haciendo la analogía con la función discriminante $z = \mathbf{a}'\mathbf{x}$, las 17 variables seleccionadas, conforman el vector $\mathbf{x}$, el cual existe para cada observación (podríamos llamarle $\mathbf{x}_i$). El vector $\mathbf{a}'$ es el vector de coeficientes.

SPSS proporciona el cuadro "Canonical Discriminant Funtion Coefficients" —que se muestra abajo— donde se expresan los valores de los coeficientes para las variables que integran la función.

Canonical Discriminant Function Coefficients

	Function
	1
ESC_JEFE	-.052
PCP_TRAB	.933
I_DEPEN	.586
I_HACINA	.088
DISCAP DISCAPACITADOS EN EL HOGAR	.344
M_TECHO	.082
S_AGUABA	.125
LICUADOR	.195
REFRI	.091
ESTUFGAS	.300
RADIO	.073
MODULAR	.312
TELE	.087
VHS	.345
LAVAROPA	.154
VENTILA	.083
VEHICULO	.572
(Constant)	-3.469

Unstandardized coefficients

Pero como bien sabemos, para conocer qué tanto contribuye cada variable al modelo, se requiere estandarizar estos coeficientes, con lo que se obtiene el cuadro de coeficientes estandarizados, que es equivalente a los que se presentaron en el cuadro 5.6, sólo que ahora la función involucra 17 variables, ya que 3 no fueron seleccionadas en el procesos paso a paso (cuadro 5.7).

Cuadro 5.7

Coeficiente canónicos estandarizados (17 variables)

	Variables	Coeficientes		Variables	Coeficientes
1	ESC_JEFE	-0.1347	10	ESTUFGAS	0.1109
2	PCP_TRAB	0.1667	11	RADIO	0.0364
3	I_DEPEN	0.8528	12	MODULAR	0.0856
4	I_HACINA	0.1600	13	TELE	0.0426
5	DISCAPAC	0.0649	14	VHS	0.0682
6	M_TECHO	0.0376	15	LAVAROPA	0.0505
7	S_AGUABA	0.0523	16	VENTILA	0.0386
8	LICUADOR	0.0858	17	VEHICULO	0.1046
9	REFRI	0.0356			

Esta información en combinación con la correlación de cada variable con la variable canónica z (cuadro 5.8), ayuda a entender la importancia que cada variable tiene para el modelo. En el cuadro de coeficientes estandarizados, SPSS sólo despliega la información de las 17 variables con que se estima z. El cuadro de las correlaciones, presenta todas las variables, incluso aquellas que no resultaron seleccionadas.

Cuadro 5.8

Correlaciones con la variable canónica

	Variables	Coeficientes		Variables	Coeficientes
1	I_DEPEN	0.8344	11	VHS	0.1888
2	I_HACINA	0.5434	12	VEHICULO	0.1839
3	ESTUFGAS	0.3224	13	S_AGUABA	0.1752
4	LICUADOR	0.2841	14	TELE	0.1699
5	REFRI	0.2675	15	PCP_NASI °	0.1624
6	VENTILA	0.2545	16	RADIO	0.1129
7	EDAD_JEF °	-0.2540	17	PCP_TRAB	0.1117
8	LAVAROPA	0.2264	18	ESC_JEFE	-0.0895
9	M_TECHO	0.2081	19	CAL_AGUA °	0.0551
10	MODULAR	0.1952	20	DISCAPAC	0.0425

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

° This variable not used in the analysis.

Un aspecto importante del análisis discriminante, que proporciona SPSS, es el valor de los centroides de cada grupo. Lo que se muestra en el cuadro "Funcrions at Group Centroids". Los centroides son los valores z promedio de cada grupo. Justamente la función discriminante se obtiene maximizando la distancia entre los centroide, lo que ocurre cuando $\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Functions at Group Centroids

	Function
LIN_POB	1
.00	-.804
1.00	.296

Unstandardized canonical discriminant functions evaluated at group means

Se puede observar que para nuestro ejercicio, el grupo de las familias en pobreza tiene su centroide en un valor menor al del grupo de familias en pobreza extrema. Esto es resultado de la manera como se construyó la variable que identifica cada grupo. Coeficientes negativos acercan a la observación hacia el grupo de las familias solventes.

En este ejercicio tenemos tres casos en cuanto a la relación de los coeficientes de la función discriminante y la dirección en que la mueven. Esta relación es lo que nos indica si alguna variable tiene algún problema de cambio de signo. Veamos algunos ejemplos de cómo deben analizarse las variables de acuerdo a lo que representan.

Proporción. Usemos como ejemplo la variable *pcp_trab* cuyo coeficiente no estandarizado es 0.933, y supongamos un hogar con 5 niños entre 8 y 15 años, entonces:

Si 2 trabajan, pcp_trab = 0.4 ⇒ 0.933 * 0.4 = 0.373
Si 4 trabajan, pcp_trab = 0.8 ⇒ 0.933 * 0.8 = 0.746

A mayor número de niños que trabajan, la puntuación de la variable se mueva hacia el centroide del grupo de pobres extremos. Si el signo del coeficiente fuera negativo, esta variable se movería en sentido contrario pues la puntuación nos acercaría al centroide de los pobres.

Dicotómicas. Supongamos que el hogar cuenta con licuadora, entonces 0.195 * 0 = 0, no aporta nada a la combinación lineal; pero si no hay licuadora, la puntuación sería 0.195*1=.0195, con lo que contribuye a incrementar el valor de la combinación lineal. Con esta contribución, se mueve hacia el grupo de los pobres extremos lo cual es lógico ya que suponemos que las familias que no cuentan con este enser están en condiciones más difíciles de aquellas que si lo posean. Esto pasa para todas las variables 0–1 ya que **se construyeron en concordancia con la definición y construcción de los dos grupos.**

Reales. Como ejemplo tomemos *i_hacina* cuyo coeficiente es 0.088, entonces:

Si *i_hacina* = 2 ⇒ 0.088 * 2 = 0.172

Si *i_hacina* = 4 ⇒ 0.088 * 4 = 0.352

A mayor índice de hacinamiento, mayor puntuación, por lo tanto contribuye a llevar la combinación lineal a valores más grandes, es decir, al grupo de pobres extremos. Algo que resulta lógico si pensamos que mientras más hacinadas están las familias peores condiciones de vida tienen.

El valor z es aquel que maximiza la distancia entre z_1 y z_2. Pues bien, SPSS también proporciona los valores de z_1 y z_2. El cuadro "Classification Function Coefficients" muestran estos valores. Al igual que los coeficientes estandarizados y los no estandarizados, para el ejercicio SPSS presenta, únicamente, información de las 17 variables seleccionadas.

Classification Function Coefficients

	LIN_POB	
	.00	1.00
ESC_JEFE	.784	.727
PCP_TRAB	1.191	2.217
I_DEPEN	.366	1.011
I_HACINA	.462	.559
DISCAP DISCAPACITADOS EN EL HOGAR	1.438	1.816
M_TECHO	3.846E-02	.129
S_AGUABA	1.330	1.468
LICUADOR	-.877	-.662
REFRI	-.630	-.530
ESTUFGAS	.395	.724
RADIO	1.401	1.481
MODULAR	6.746	7.089
TELE	-.140	-4.419E-02
VHS	16.857	17.236
LAVAROPA	3.058	3.228
VENTILA	-.929	-.838
VEHICULO	22.758	23.388
(Constant)	-25.728	-29.264

Fisher's linear discriminant functions

Otro de los resultados que brinda la salida del SPSS es el cálculo de la Λ de Wilks que sirve para probar la hipótesis nula de que las medias de todas las variables a través de los grupos definidos, con el análisis discriminante, son iguales. Se presenta lambda con una aproximación a la distribución ji-cuadrada.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.808	8331.496	17	.000

Para nuestro ejercicio, la ji-cuadrada muestra que la diferencia entre las medias de los grupos definidos por la función discriminante es altamente significativa.

Finalmente tenemos un cuadro cruzado entre la clasificación original y la obtenida con la función discriminante. También se presenta el porcentaje de observaciones que en ambos casos, con la función discriminate y con la clasificación por *lin_pob*, se clasificaron en el mismo grupo.

El inciso c. expresa el porcentaje total de observaciones clasificadas en el mismo grupo con el procedimiento de validación que resulta del método cross-validate

Classification Results [b,c]

			Predicted Group Membership		
		LIN_POB	.00	1.00	Total
Original	Count	.00	8381	2161	10542
		1.00	9657	18996	28653
	%	.00	79.5	20.5	100.0
		1.00	33.7	66.3	100.0
Cross-validated [a]	Count	.00	8377	2165	10542
		1.00	9668	18985	28653
	%	.00	79.5	20.5	100.0
		1.00	33.7	66.3	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 69.8% of original grouped cases correctly classified.

c. 69.8% of cross-validated grouped cases correctly classified.

El segundo modelo, aún contiene un número grande de variables en la función discriminante, por lo que en un intento por reducir este número, se realizaron varias pruebas en las que se excluyeron algunas variables que menos contribuían al modelo, que menos correlación mostraban o que, con una combinación de estos dos criterios, se pensaba que su aportación no era tan importante. Sin embargo el criterio que mejor identificó a estas variables, llegó por otro camino.

Se mantuvieron en el análisis las variables identificadas como claves para bosquejar la situación socioeconómica de una familia —aquellas que son menos temporales, por tanto permanentes y más arraigadas— que son las relacionadas con los indicadores demográficos, así como aquellas que expresan las condiciones de los servicios básicos. Aunque éstas últimas en ocasiones pueden ser engañosas, cuando por ejemplo, en alguna localidad se ha llevado a cabo

algún programa para dotar de ciertos servicios a la población, tal es el caso del servicio de luz eléctrica o drenaje. En Yucatán, por ejemplo, el gobierno estatal implementó un programa para poner piso firme en todas las viviendas de las familias que no tuvieran; seguramente esta es la razón por la cual la variable *m_piso* no funcionó para la discriminación.

Sin embargo, el modelo incluye un importante número de variables que hacen referencia a la existencia de enseres domésticos (casi todos). Bajo el supuesto de que no se requiere incluir tantos de ellos, se planteó la opción de reducir este número para incluir únicamente aquellos, que de entrada, muestran una fuerte diferencia entre los dos grupos, no desde el punto de vista de la significación estadística —que con la prueba de medias ya vimos qué expresa—, sino de una manera más intuitiva.

Así, con base en los promedios de posesión de cada enser para los dos grupos de familias, se identificaron los enseres con promedios menos diferentes. Y estos fueron los que se excluyeron del análisis. El indicador que su utilizó para esto fue:

$$\text{indicador} = \frac{\text{diferencia en puntos porcentuales, entre los hogares solventes y los pobres, que cuentan con el enser}}{\text{porcentaje del total de hogares que cuentan con el enser}}$$

Cuadro 5.9

Enser	Porcentaje de hogares con enser: Todos los Hogares	Hogares Solventes	Hogares Pobres	Indicador
Licuadora	27.00	37.10	23.30	51.1%
Refri	19.60	28.00	16.40	59.2%
Estufa de gas	17.00	26.60	13.50	77.1%
Calentador de agua	3.80	4.90	3.40	39.5%
Radio	53.20	57.60	51.50	11.5%
Tocadiscos	8.40	12.70	6.80	70.2%
Televisión	58.40	65.10	55.90	15.8%
Videocasetera	4.20	7.20	3.10	97.6%
Lavadora	12.50	18.50	10.30	65.6%
Ventilador	33.10	42.70	29.60	39.6%
Vehículo	3.50	6.10	1.30	137.1%

Indicador =
Diferencia en puntos porcentuales pobres y pobres extremos que cuentan con el enser, dividido entre el porcentaje de la población que cuenta con el enser

El cuadro de arriba muestra el resultado obtenido para cada enser. Los casos en que el indicador es mayor a 50% fueron incluidos en el modelo. Cuatro fueron las variables que se excluyeron: calentador de agua, radio, televisión y ventilador.

Obtuvimos, entonces, el tercer modelo que proporcionó un buen resultado, pues no se presentaron cambios importantes en ninguno de los indicadores ya presentados para el segundo modelo (de las 17 variables). La función discriminante quedó conformada por 14 de las 16 variables consideradas. Por otro lado se observa que no se incorporó ninguna de las dos variables que en el ejercicio previo no hubieran sido consideradas en la función anterior.

En resumen, el modelo no sufre cambios. La ganancia está en que, con menos variables, alcanzamos los mismos resultados en los indicadores del modelo a los obtenidos para el segundo. Los resultados de este tercer y definitivo modelo, su regla de asignación, se presentan en el siguiente capítulo.

Capítulo VI

RESULTADOS Y CONCLUSIONES

En el capítulo anterior se presentaron los resultados del segundo de los modelos obtenidos, el cual contiene 17 variables, y se mencionó un tercer modelo que es el definitivo para nuestro ejercicio. Hay que mencionar que antes de llegar a este tercer modelo fue necesario efectuar varios ejercicios con el propósito de responder distintas interrogantes que se fueron derivando de los resultados que se iban obteniendo. En este capítulo se presentan los resultados de este tercer modelo, así como algunos comentarios que permitan ilustrar la razón por lo que se decidió adoptar este modelo como definitivo.

6.1 El modelo definitivo

Regresemos al punto en que tenemos el segundo modelo conformado por 17 variables. Con la idea de reducir el número de variables y viendo que el número de aquéllas que representan la presencia de enseres en el hogar es grande, se decidió eliminar algunas de ellas, y verificar si las que permanecían ayudaban a explicar a las que se dejaron fuera.

Después de efectuar el análisis sobre la diferencia de la presencia de enseres entre los grupos clasificados con línea de pobreza (cuadro 5.9), nos quedamos con un total de 16 variables elegibles para la definición de la función discriminante. El cuadro 6.1 presenta la lista de estas variables.

Nuevamente se aplicó el método de selección por pasos para estas 16 variables y en este caso, son dos las que no alcanzan el valor de F mínimo para ser seleccionadas. Nuestra función discriminante queda conformada con las 14 variables seleccionas, obteniendo así el tercer modelo que finalmente es el que se consideró como definitivo.

Cuadro 6.1 Listado de variables que se consideran en el modelo definitivo

1	Edad_jef	Edad del jefe del hogar	Real
2	esc_jefe	Años de escolaridad del jefe del hogar	Real
3	pcp_nasi	Proporción de niños entre 5 y 15 años que no asisten a la escuela	Real
4	pcp_trab	Proporción de niños entre 8 y 15 años que trabajan	Real
5	i_hacina	Índice de hacinamiento	Real
6	i_depen	Índice de dependencia	Real
7	Discapac	Presencia de discapacitados en el hogar	Dicotómica
8	m_techo	Material del techo de la vivienda	Dicotómica
9	s_aguaba	Servicio de agua y baño	Dicotómica
10	licuador	Licuadora	Dicotómica
11	refri	Refrigerador	Dicotómica
12	estufgas	Estufa de gas	Dicotómica
13	modular	Tocadiscos o modular	Dicotómica
14	vhs	Videocasetera	Dicotómica
15	lavaropa	Lavadora de ropa	Dicotómica
16	vehiculo	Vehículo	Dicotómica

El hecho de que únicamente dos —un número muy pequeño— son las variables excluidas, puede explicarse por los niveles de correlación que presentan entre ellas, los cuales en general son bajos, por lo que es necesario incluir un número importante de variables para poder alcanzar un modelo significativo. Si tuviéramos variable mejor correlacionadas, es posible que el número de las que conformarían el modelo, fuera menor.

Al compara la diferencia entre grupos obtenida en el segundo modelo, con la obtenida para este último, se puede apreciar que el resultado de la discriminación es bastante satisfactorio. Sólo en el caso de la televisión se pierden algunos puntos porcentuales en la diferencia; en el resto de las variables, en general, se mantiene lo alcanzado con el segundo modelo (ver cuadro 6.2). Realizando una inspección comparativa de los grupos definidos originalmente con la línea de pobreza (cuadro 5.5), y las características de los grupos obtenidos con la función discriminante del

último modelo, podemos apreciar que se obtiene una mejor separación con este último, ya que las diferencias entre las características de cada grupo se acentúa.

Cuadro 6.2 **Características de los hogares**
Clasificación con función discriminante

	Total	No extremos	Extremos
Hogar			
Personas en el hogar [2]	4.76	3.83	5.55
Edad del jefe [2]	43.75	48.50	39.73
Años de escolaridad del jefe [2]	3.16	3.26	3.08
Habla lengua indígena el jefe del hogar [1]	93.20	92.20	94.00
Niños de 0 a 11 años [2]	1.42	0.63	2.08
Niños que no asisten a la escuela [2]	18.46%	14.42%	19.70%
Proporción de niños que trabajan por un ingreso [2]	9.40%	5.24%	10.85%
Con discapacitados [1]	3.70	2.50	4.70
Índice de dependencia [2]	2.07	0.93	3.03
Índice de hacinamiento [2]	3.62	2.58	4.50
Servicios			
Agua dentro de la vivienda o en el terreno [1]	86.40	88.30	84.60
Baño con agua corriente [1]	9.40	14.20	5.20
Agua dentro de la vivienda y baño con agua corriente [1]	6.80	10.70	3.30
Luz eléctrica [1]	89.80	91.40	88.40
Piso de tierra [1]	10.60	9.20	11.90
Techo firme [1]	30.80	39.70	23.20
Cuartos por vivienda [1]	1.62	1.74	1.51
Pertenencias y propiedades			
Licuadora [1]	27.00	38.70	17.10
Refri [1]	19.60	29.00	11.40
Estufa de gas [1]	17.00	27.60	7.90
Calentador de agua	3.80	4.60	3.10
Radio [1]	53.20	55.50	51.00
Tocadiscos [1]	8.40	13.00	4.40
Televisión [1]	58.40	62.90	54.10
Videocasetera [1]	4.20	7.10	1.70
Lavadora [1]	12.50	18.80	7.10
Ventilador [1]	33.10	42.00	25.30
Vehículo [1]	3.50	6.10	1.30
Poseen tierras [1]	45.30	41.20	48.90
Número total de animales de trabajo [2]	3.60	3.40	3.90
Total de animales (unidades homogéneas) [2]	164.65	155.77	172.03
Total de hectáreas (unidades: predios de temporal) [2]	1.08	1.00	1.16
4. Ingresos			
Ingreso mensual per cápita [2]	294.25	410.94	195.43
Porcentaje de observaciones		33.5%	66.5%

[1] Porcentaje
[2] Promedio

Veamos ahora los coeficientes a_j de la función discriminante obtenida y las correlaciones de cada una de las variables consideradas. El cuadro 6.3 presentan la función discriminante de coeficientes estandarizados, mientras que el cuadro 6.4, las correlaciones de cada una de las variables con la función canónica. El análisis de estos cuadros nos permite identificar las variables que tienen más importancia dentro del modelo.

Cuadro 6.3

Función Discriminante Definitiva

Coeficiente estandarizados

	Variables	Coeficientes
1	ESC_JEFE	-0.1402
2	PCP_TRAB	0.1675
3	I_DEPEN	0.8549
4	I_HACINA	0.1571
5	DISCAPAC	0.0654
6	M_TECHO	0.0496
7	S_AGUABA	0.0555
8	LICUADOR	0.1086
9	REFRI	0.0475
10	ESTUFGAS	0.1225
11	MODULAR	0.0810
12	VHS	0.0674
13	LAVAROPA	0.0587
14	VEHICULO	0.1069

Cuadro 6.4

Correlaciones

Coeficiente vs Función

	Variables	Coeficientes
1	I_DEPEN	0.8363
2	I_HACINA	0.5447
3	ESTUFGAS	0.3229
4	LICUADOR	0.2855
5	REFRI	0.2691
6	EDAD_JEF °	-0.2516
7	LAVAROPA	0.2261
8	M_TECHO	0.2092
9	MODULAR	0.1951
10	VHS	0.1889
11	VEHICULO	0.1854
12	S_AGUABA	0.1753
13	PCP_NASI °	0.1635
14	PCP_TRAB	0.1124
15	ESC_JEFE	-0.0912
16	DISCAP	0.0429

° No incluida en la función discriminante

La importancia que cada variable tiene dentro de la función discriminante está dada por el valor absoluto de su coeficiente. Aquí vemos que el índice de dependencia aparece como la variable con el coeficiente mayor: 0.8459, así como la de mayor correlación con la función canónica 0.8368. Por lo tanto, es la variable de mayor importancia en este caso. Le siguen a una distancia considerable la proporción de niños que trabajan, el índice de hacinamiento y la escolaridad del jefe. Finalmente, vienen los enseres con la estufa de gas a la cabeza y a cierta distancia, licuadora y vehículo. Los demás presentan coeficientes relativamente pequeños.

Resalta el hecho de que las variables que comprenden las características menos temporales (más estables en el tiempo) son las que mejor discriminan los grupos. Este tipo de variables reflejan aspectos estructurales de la familia. Son aspectos en los que hay menos influencia de terceros Es decir, en el caso de servicios de agua, luz, etcétera, las familias están sujetas a que estén disponibles en la comunidad; pero, aspectos como la escolaridad del jefe, el hacinamiento de la familia, están más relacionados a las condiciones internas de ésta.

También llama la atención que dentro de los considerados como enseres, el que mejor discrimina es la estufa de gas, además de ser una de las variables que mejor correlación presenta con la función discriminante. La estufa de gas es un enser que requiere un manejo más complejo, pues hay que contar con una instalación especial, así como tener acceso el servicio del gas; mientras que para otros enseres lo que importa es tener, únicamente, el servicio de electricidad.

Se habrá observado que en ningún momento hemos considerado la probabilidad *a priori* p, ni la del error de clasificación y tampoco los costos de éste. Y es que en ningún momento se supone que la clasificación original es la mejor. De hecho, estamos suponiendo que es perfectible, pues la variable ingreso presenta algunas dificultades desde su origen. Aunque hay que subrayar que con la depuración realizada quedaron fuera del análisis todos los hogares para los cuales se tenía la mínima duda de la veracidad de su información.

Tampoco se está asegurando que la clasificación original esté mal. Lo que sí suponemos es que es un punto de partida, y que al tomar en cuenta otro tipo de variables, la clasificación que se obtenga —si realmente se logra— puede ser más robusta.

El siguiente cuadro presenta los centroides de cada grupo en el tercer modelo, mismos que son similares a los obtenidos para el segundo, y que fueron presentados en el capítulo anterior.

Functions at Group Centroids

	Function
LIN_POB	1
.00	-.601
1.00	.295

Unstandardized canonical discriminant functions evaluated at group means

El análisis discriminante también nos proporciona una regla de asignación que permite clasificar nuevas observaciones en cualquiera de los grupos definidos. En el capítulo IV se plantea la regla de asignación (4.8) para el caso donde el error de mala clasificación y la probabilidad original son iguales. Esta regla establece el punto medio entre los centroides de las poblaciones como el punto crítico para la asignación.

La función discriminante del modelo final con 14 variables, que se expresa cuadro 6.3, nos proporciona una regla de asignación que se construye considerando lo siguiente:

Sean

$\mathbf{z}_1$ el vector de valores z del grupo de familias en condiciones de pobreza extrema,

entonces el valor z más pequeño del grupo 1 esta dado por $\min\{\mathbf{z}_1\} = z_{1\min}$

$\mathbf{z}_2$ el vector de valores z del grupo de familias en condiciones de pobreza,

entonces el valor z más grande del grupo 2 esta dado por $\max\{\mathbf{z}_2\} = z_{2\max}$

z_{pc} el valor en el que se ubica el punto de corte.

Para el grupo 1 que contiene a los hogares pobres extremos, los valores z, que son las combinaciones lineales de cada hogar clasificado en este grupo, se encuentran comprendidos en el intervalo (–0.25288,3.95063). Por su parte los valores z del grupo de hogares pobres (grupo 2) se distribuyen en el intervalo (–3.71837,–0.25309). Del grupo de hogares pobres extremos, tomamos el valor z_1 más pequeño: $z_{1\min}$=–0.25288. Para el grupo de los hogares pobre, el valor z_2 más grande: $z_{2\max}$=–0.25309. Con estos dos valores obtenemos nuestro punto de corte: z_{pc}= –0.25309.

La regla de asignación queda definida de la siguiente manera:

Si $z > z_{2\max} = -0.25309$, asigna **x** al Grupo 1: Familias en pobreza extrema

de otra forma, asigna **x** al Grupo 2: Familias en pobreza.

Donde x es el factor con la información del individuo.

Ante la evidencia de que las variables que más aportan a la discriminación son aquellas con la característica de ser más estables en el tiempo, y dado que la mayoría de ellas son de tipo real, surge la inquietud de observar lo que sucedería con la clasificación si consideramos únicamente las variables menos volátiles, a la vez, nos permitiría aislar en cierta medida los efectos de las variables de tipo real. Probablemente se lograría construir un modelo adecuado que tuviera un menor número de variables, algo que en términos técnicos sería deseable. Faltaría ver si ya puesto en la realidad resulta igual de deseable.

No presentaremos aquí los resultados obtenidos para este nuevo ejercicio, bastará mencionarlos y contrastarlos con lo obtenido en el modelo definitivo y el modelo original (con línea de pobreza).

Este ejercicio logra identificar de manera clara, dos grupos de familias. De hecho se alcanza una mejor separación entre los grupos al obtenido en la clasificación original; sin embargo, esta separación es menos eficiente a la alcanzada con el tercer modelo de 14 variables.

Vale la pena resaltar que las variables que presentaron problemas de signo (total de animales y número de hectáreas), ante la ausencia de las variables de enseres, mantuvieron la misma problemática. Esto habla de que las variables enseres no están contribuyendo a modificar el comportamiento de las otras. Por el lado de las correlaciones, la variable con la más alta de ellas fue, nuevamente, el índice de dependencia *i_depend*. Esta variable también fue la que más peso presentó en el modelo. Se puede apreciar que en términos generales el comportamiento de las variables fue muy se,ejante al del modelo definitivo.

Resumiendo: la separación lograda fue buena, las diferencias alcanzadas entre variables, con este cuarto modelo, fueron muy similares a las del modelo definitivo, salvo en el caso de los enseres, donde las diferencias en puntos porcentuales claramente diminuye. Sin embargo, a pesar de esto, aún en las variables de enseres se alcanzó una mejor diferenciación entre los dos grupos de hogares, que es obtenido por el modelo original.

6.2 CONCLUSIONES Y COMENTARIOS

Podemos mencionar 3 conclusiones de este ejercicio:

1. Si hace sentido hablar de una diferenciación entre las familias pobres

Con base en los indicadores obtenidos por el propio análisis discriminante, se puede asegurar que sí se logró hacer una distinción entre las familias pobres de las localidades marginadas. Al analizar de manera comparativa las características de los grupos obtenidos con el modelo original (cuadro 5.5) y con el modelo definitivo (cuadro 6.2) se logra identificar la existencia de dos grupos bien definidos entre las familias pobres. Con este resultado, a la par que se verifica que, efectivamente, existe una diferencia entre las familias pobres, también se obtuvo una nueva clasificación que separa mejor los grupos y muestra con mayor claridad las diferencias que existen entre uno y otro.

Podemos decir, entonces, que hablar de pobreza extrema tiene sentido, faltaría ver cuáles pueden ser las directrices de las acciones para apoyar a esas familias. Esfuerzos, sin duda se están realizando.

Aún teniendo identificados los dos grupos, podemos ver que aspectos tan importantes, como el nivel de escolaridad de los jefes, es en general muy bajo en ambas poblaciones. Igual sucede en los ingresos familiares y los servicios de baño con agua corriente.

2. Las variables estructurales son las que acentúan las diferencias entre los grupos

Dentro del modelo estadístico, las características que están marcando la diferencia son esas que representan las condiciones estructurales de las familias que ayudan a la repetición de la pobreza generación tras generación. Estas son las características más difíciles de modificar, pues requiere un esfuerzo sostenido que debe trascender los tiempos sexenales y cualquier lógica que no sea la de combatir una situación estructural.

Pese a que estas variables muestran su importancia dentro de la discriminación, no es conveniente dejar de lado lo que nos están aportando la información de los enseres, ya que al incorporarla, las diferencias entre uno y otro grupo se hacen más evidentes, a la vez que las diferencias ya obtenidas para las otras características se mantiene. Podemos decir que las variables estructurales son sobre las que recae el peso de la discriminación.

3. Es adecuado aplicar una técnica que incorpore variables reales.

Lo anterior habla de un aspecto interesante: el hecho de que en un ejercicio de este tipo, donde las variables más estables son de tipo real, resulta correcto emplear una metodología que permita trabajar con variables de tipo real, ya que son éstas donde recae la discriminación. Son las que más aportan a la definición del modelo.

Este comentario surge en contraposición a la idea de que el uso de una metodología diferente, como pudiera ser la regresión logística, sería más adecuada por el número de variables dicotómicas que se tiene. Sin embargo, por lo que representan las variables reales, es importante incluirlas, como reales, y así aprovechar toda la información que contiene. En el caso de un modelo de regresión logística se perdería información al transformarlas a dicotómicas.

La enseñanza que podemos obtener de lo anterior, es que, aunque la teoría indica que ante la presencia de un número importante de variables dicotómicas es preferible usar un modelo de regresión logística, esto no debe tomarse como receta; mas bien, es importante incluir la experiencia propia para discernir sobre el tipo de información que aporta cada variable y elegir el método a emplear. En todo caso, sería conveniente realizar algunas pruebas para ver la importancia que las variables tienen en el modelo.

Referencias

1. Anderson, T. (1984) 'An introduction to multivariate analysis'. Ed. John Wiley & Sons.
2. Ayres, F. Jr (1970) 'Matrices. Serie Schaum'. Ed. McGraw Hill.
3. Canavos, G. C. (1988) 'Probabilidad y estadística. Aplicaciones y métodos'. Ed. McGraw Hill.
4. Ciesas, Progresa (1999) 'Alivio a la pobreza: análisis del programa de educación salud y alimentación dentro de la política social'.
5. Ciesas, Progresa (1999) 'Memorias del seminario 'Alivio a la pobreza'.
6. Pieck, E. & Aguado, E. (1995) 'Educación y pobreza (de la desigualdad social a la equidad'. El colegio Mexiquense, UNICEF.
7. Goldstein, M. y Dillon, W. R. (1978) 'Discrete discriminant analysis'. Ed. John Wiley & Sons.
8. Hair J. F. jr., Anderson, R. E., Tatham, R. L. y Black, W. C. (1999) 'Análisis multivariante'. Ed. Prentice Hall.
9. Hand, D. J. (1981) 'Discrimination and classification' . Ed. John Wiley & Sons.
10. Huberty, C.J. (1994) 'Applied dicriminant analysis'. Ed John Wiley & Sons.
11. INEGI 'Estadísticas Históricas de México – Tomo I'
12. Johnson, R. A. y Wichern, D. W. (1992) 'Applied multivariate statistical analysis'. Ed. Prentice Hall
13. Johnson, D. E. (2000) 'Métodos multivariados aplicados al análisis de datos'. Ed. International Thomson editores, S.A. de C.V.
14. Lang, S (1976) 'Álgebra lineal'. Ed. Fondo educativo interamericano, S.A.
15. Lehmann, C. H. (1964) 'Álgeba'. Ed. Limusa.
16. Lipschutz, S. (1970) 'Álgebra lineal. Serie Schaum'. Ed. McGraw Hill.
17. McLachlan, G. J. (1992) 'Discriminant analysis and statistical pattern recognition'. Ed. John Wiley & Sons.
18. Meyer, P. L. (1973) 'Probabilidad y aplicaciones estadísticas'. Ed. Fondo educativo interamericano, S.A.
19. Rencher, A. (1995) 'Methods of multivariate analysis'. Ed. John Wiley & Sons.
20. SPSS 'Bases 7.5 Aplication guide'
21. Swokowski, E. W. (1982) 'Cálculo con geometría analítica'. Ed. Wadsworth international iberoamérica.
22. Tutsuoka, M. (1971) 'Multivariate analysis. Techniques for educational and psychological research'. Ed. John Wiley & Sons