UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA.

"COMPARACIÓN ENTRE EL ANÁLISIS DISCRIMINANTE Y LA REGRESIÓN LOGÍSTICA EN LA CLASIFICACIÓN DE UNA COLONIA DE CANGREJOS HERRADURA (*Limulus polyphemus*)"

TRABAJO DE GRADUACIÓN PRESENTADO POR:

RENÉ ARMANDO PEÑA AGUILAR

PARA OPTAR AL GRADO DE:

MAESTRO EN ESTADÍSTICA

CIUDAD UNIVERSITARIA, JUNIO DE 2011

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA.

TRABAJO DE GRADUACIÓN

"COMPARACIÓN ENTRE EL ANÁLISIS DISCRIMINANTE Y LA REGRESIÓN LOGÍSTICA EN LA CLASIFICACIÓN DE UNA COLONIA DE CANGREJOS HERRADURA (*Limulus polyphemus*)"

PRESENTADO POR:

RENÉ ARMANDO PEÑA AGUILAR

ASESORES:

DE LA UNIVERSIDAD DE EL SALVADOR:

DR. JOSÉ NERYS FUNES TORRES

DE LA UNIVERSIDAD COMPLUTENSE DE MADRID:

DR. JOSÉ MIGUEL GARCÍA-SANTESMASES MARTÍN-TESORERO

CIUDAD UNIVERSITARIA, JUNIO DE 2011

UNIVERSIDAD DE EL SALVADOR

RECTOR: MSC. RUFINO ANTONIO QUEZADA SÁNCHEZ

SECRETARIO GENERAL: LIC. DOUGLAS VLADIMIR ALFARO CHÁVEZ

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO : DR. RAFAEL ANTONIO GÓMEZ ESCOTO

SECRETARIA: LICDA. MARÍA TRINIDAD TRIGUEROS DE CASTRO

ESCUELA DE MATEMÁTICA

DIRECTOR : ING. CARLOS MAURICIO CANJURA LINARES.

SECRETARIO : LIC. ERNESTO AMÉRICO HIDALGO

ÍNDICE GENERAL

Introducción

En el desarrollo de una investigación, frecuentemente obtenemos características o variables que pueden ser entre categóricas y cuantitativas. Con estas variables se pretende dar solución a un determinado problema que se plantea previamente. La elección de los métodos estadísticos apropiados para los análisis de los datos son importantes para obtener buenos resultados que orienten a la toma de decisiones correctas.

El problema planteado en este trabajo consiste en estudiar la efectividad de clasificación entre los métodos "Análisis Discriminante", que consiste en ubicar a los individuos en una determinada población con base a características medibles y observadas en los individuos; y la "Regresión Logística", que se utiliza para estimar las probabilidades de que un individuo pertenezca a un determinado grupo dado que las características del individuo han tomado ciertos valores concretos.

Los datos para buscar resolver este problema fueron tomados de libro de *Agresti (2002)*, que consisten en una muestra de 173 individuos marinos cuyo nombre científico es *Limulus polyphemus*, comúnmente conocidos como "cangrejo herradura". Para estos individuos, las hembras tienen su propio macho en su nido, pero, atraen a otros machos a los que se les llama individuos satélites.

El objetivo fundamental de este trabajo es ver la efectividad de los métodos "Análisis Discriminante" y la "Regresión Logística" clasificando el cangrejo herradura hembra con o sin individuos satélites, partiendo de características como el peso, el ancho de caparazón, el color y el estado de las espinas del cangrejo herradura hembra.

Es de señalar que en el libro de *Agresti (2002)* , se han realizado otros tipos de análisis estadísticos como la predicción del número de satélites con la regresión de Poisson, en ningún momento se han empleado los métodos que en este trabajo son nuestro objetivo para la clasificación que se pretende hacer con la muestra de los 173 cangrejos de herradura hembra.

Para lograr nuestro objetivo fundamental, nuestro trabajo está organizado de la siguiente forma:

En un primer capítulo trataré de describir los fundamentos teóricos del Análisis Discriminante y de la Regresión Logística, este último es un modelo lineal generalizado y por tanto, se describirán los fundamentos teóricos de los modelos lineales generalizados.

En un segundo capítulo, con ayuda del software SPSS, se desarrolla la aplicación de los métodos de clasificación “Análisis discriminante” y “Regresión Logística” y poder obtener los resultados necesarios para la comparativa de ambos métodos.

No hay duda que el presente trabajo ayudará en los procesos de enseñanza aprendizaje de los estudiantes en Estadística, específicamente cuando se estudian los métodos de clasificación.

1. Clasificación usando el análisis discriminante y la regresión logística.

1.1.1. Introducción al análisis discriminante

¿Qué es el análisis discriminante?

Al intentar elegir una técnica apropiada en el análisis estadístico de datos, algunas veces encontramos un problema que incluye una variable dependiente categórica y varias variables independientes métricas. Por ejemplo, podemos querer distinguir entre riesgo de crédito alto y bajo. Si tuviéramos una medida métrica del riesgo de crédito, podríamos utilizar la regresión multivariante. Pero puede ocurrir que sólo queramos conocer si alguien se encuentra en una categoría de riesgo bueno o malo. Esta no es la medida de tipo métrico requerida para el análisis de regresión múltiple.

El análisis discriminante es las técnica estadística apropiada cuando la variable dependiente es categórica (nominal o no métrica) y las variables independiente son métricas. En muchos casos, la variable dependiente consta de dos grupos o clasificaciones, por ejemplo, masculino frente a femenino o alto frente a bajo o bueno frente a malo. En otras situaciones, se incluyen más de dos grupos o clasificaciones, como en una clasificación de tres grupos que comprenda clasificaciones bajas, medias y altas. El análisis discriminante tiene la capacidad de tratar tanto dos grupos como grupos múltiples.

El análisis discriminante implica obtener un valor teórico, es decir, una combinación lineal de dos o más variables independientes que discrimine mejor entre los grupos definidos a priori. La discriminación se lleva a cabo estableciendo las ponderaciones del valor teórico para cada variable de tal forma que maximicen la varianza entre-grupos frente a la varianza intra-grupos. La combinación lineal para el análisis discriminante, también conocida como función discriminante, se deriva de una ecuación que adopta la siguiente forma[1]

donde:

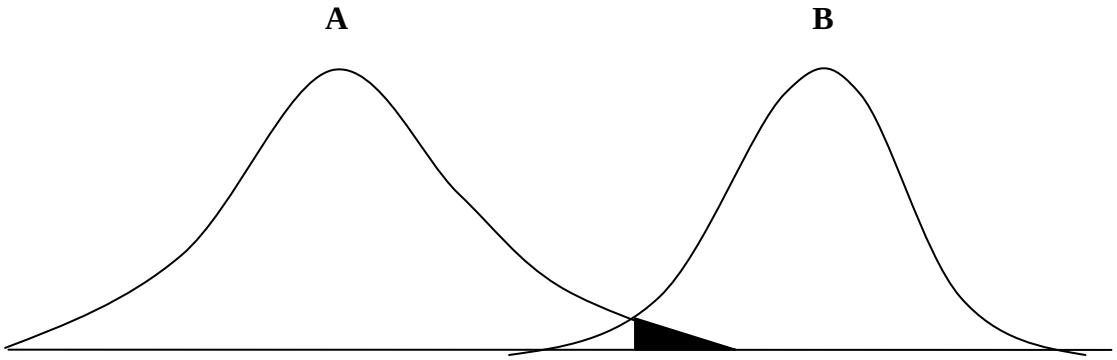
$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \cdots + W_p X_{pk}$$

$Z_{jk} = \text{Puntuacion z de la funcion discriminante j para el objeto k}$

[1] **Información tomada de Anderson y otros (1999)**

$a = \text{Constante}$
$W_i = \text{Ponderacion discriminante para la variable independiente i}$
$X_{ik} = \text{Variable independiente i para el objeto k}$

Promediando las puntuaciones discriminante Z compuesta para cada individuo en el análisis y en particular para todos los individuos dentro de un grupo, obtenemos la media del grupo. Esta media del grupo es conocida como centroide. Los centroides indican la situación más común de cualquier individuo de un determinado grupo, y una comparación de los centroides de los grupos muestra lo apartados que se encuentran los grupos a lo largo de la dimensión que se está trabajando.
El contraste para la significación estadística de la función discriminante es una medida generalizada de la distancia entre los centroides de los grupos.
Se calcula comparando las distribuciones de las puntuaciones discriminantes para los grupos. Si el solapamiento en la Distribución es pequeño, la función discriminante separa bien los Grupos. En caso contrario la función es un mal discriminador entre los grupos. La siguiente ilustración representa la distribución de puntuaciones discriminantes para una buena función discriminante que separa bien los grupos A y B.

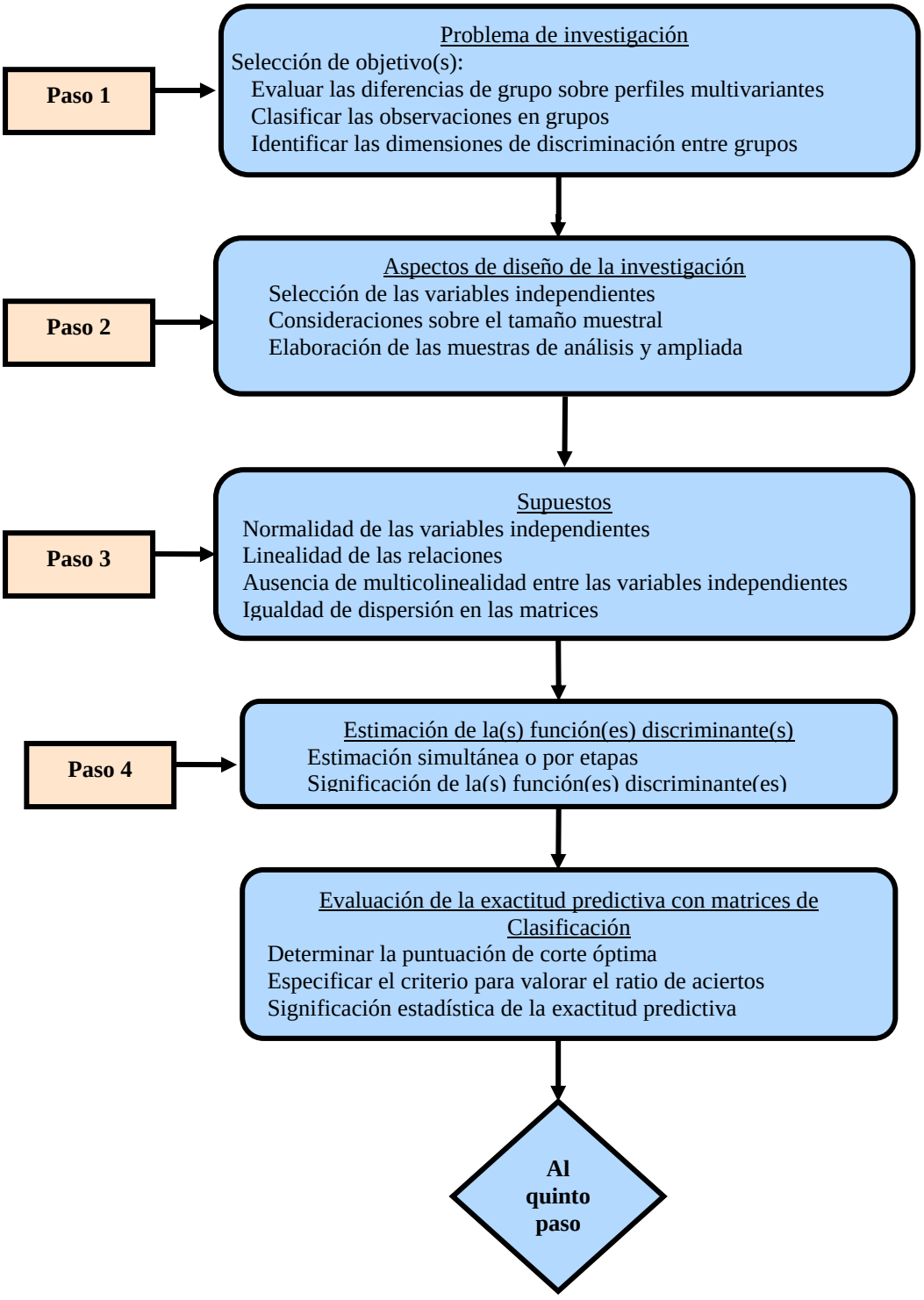


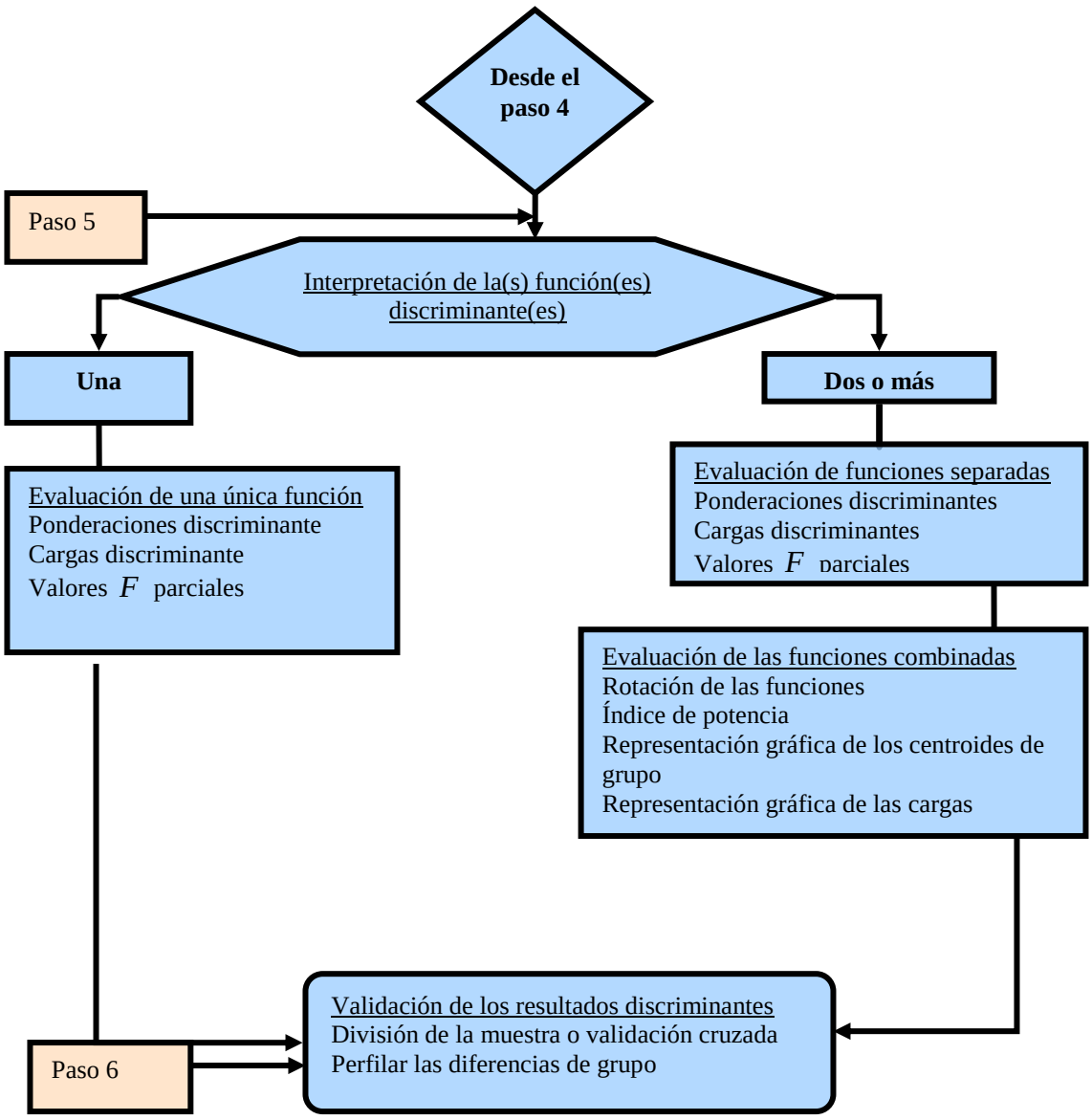
Función discriminante z

1.1.2. El proceso de decisión en el análisis discriminante

Se puede considerar la aplicación del análisis discriminante desde el punto de vista de la construcción de un modelo de clasificación en seis pasos que se ilustra en el siguiente diagrama[2]:

[2] **Información tomada de *Anderson y otros (1999)***


Paso 1
Problema de investigación
Selección de objetivo(s):
Evaluar las diferencias de grupo sobre perfiles multivariantes
Clasificar las observaciones en grupos
Identificar las dimensiones de discriminación entre grupos
Paso 2
Aspectos de diseño de la investigación
Selección de las variables independientes
Consideraciones sobre el tamaño muestral
Elaboración de las muestras de análisis y ampliada
Paso 3
Supuestos
Normalidad de las variables independientes
Linealidad de las relaciones
Ausencia de multicolinealidad entre las variables independientes
Igualdad de dispersión en las matrices
Paso 4
Estimación de la(s) función(es) discriminante(s)
Estimación simultánea o por etapas
Significación de la(s) función(es) discriminante(es)
Evaluación de la exactitud predictiva con matrices de Clasificación
Determinar la puntuación de corte óptima
Especificar el criterio para valorar el ratio de aciertos
Significación estadística de la exactitud predictiva
Al quinto paso



1.1.3. Objetivos del análisis discriminante

El análisis discriminante puede tratar cualquiera de los siguientes objetivos de investigación:

1. Determinar si existen diferencias significativas entre los perfiles de las puntuaciones medias sobre un conjunto de variables de dos (o más) grupos definidos a priori.
2. Determinar cuál de las variables independientes cuantifica mejor de las diferencias en los perfiles de las puntuaciones medias de dos o más grupos.

3. Establecer los procedimientos para clasificar objetos(individuos, empresas, productos, etc.), dentro de los grupos, en base a sus puntuaciones sobre un conjunto de variables independientes.
4. Establecer el número y la composición de las dimensiones de la discriminación entre los grupos formados a partir del conjunto de variables independientes.

Como se puede observar a partir de estos objetivos, el análisis discriminante es útil cuando el investigador está interesado en comprender las diferencias de los grupos o en clasificar correctamente objetos en grupos o clases. Por tanto, se puede considerar el análisis discriminante tanto un tipo de análisis de perfil como una técnica predictiva analítica. En cualquier caso, la técnica es la más apropiada cuando existe una única variable dependiente categóricas y varias variables independientes escaladas métricamente. Como en el análisis de perfil, el análisis discriminante proporciona una valoración objetiva de las diferencias entre grupos sobre un conjunto de variables independientes. Para comprender las diferencias del grupo, el análisis discriminante tiene en cuenta tanto el papel de la variables independientes como las combinaciones que se construyen con estas variables que representan dimensiones de discriminación entre los grupos. Estas dimensiones son los efectos conjuntos de varias variables que trabajan unidas para diferenciar entre grupos. El uso de los métodos de estimación secuencial permite también identificar subconjuntos de variables con la mayor capacidad discriminante. Finalmente, para fines de clasificación, el análisis discriminante proporciona una base, no sólo para clasificar la muestra utilizada para estimar la función discriminante, sino también cualesquiera otras observaciones que puedan tener valores para todas las variables independientes.

1.1.4. Diseño de la investigación mediante el análisis discriminante

El éxito en la aplicación del análisis discriminante requiere tener en cuenta varias cuestiones. Estas cuestiones incluyen la selección tanto de la variable dependiente como de las independientes, el tamaño muestral necesario para la estimación de las funciones discriminantes y la división de la muestra con fines de validación.

1.1.4.1. Selección de las variables dependiente e independientes

Para aplicar el análisis discriminante, el investigador primero debe especificar qué variables van a ser independientes y qué variable va a ser dependiente. Recordar que la variable dependiente es categórica y las variables independientes son métricas.

El investigador debe centrarse primero en la variable dependiente. El número de grupos de la variable dependiente (categorías) puede ser de dos o más, pero estos grupos deben ser mutuamente excluyentes y exhaustivos. Con esto se quiere decir que cada observación debe estar colocada dentro de un grupo solamente.
Hay algunas situaciones donde el análisis discriminante es apropiado incluso aunque la variable dependiente no sea una variable categórica. Podemos tener una variable dependiente que sea ordinal o medida a intervalos que queremos utilizar como variable dependiente categórica. En tales casos, tendremos que crear una variable categórica.

Cuando se crean tres o más categorías, se presenta la posibilidad de examinar solamente los grupos extremos en un análisis discriminante de dos grupos. A este proceso se le denomina *enfoque de los extremos polares*.
Después de tomarse una decisión sobre la variable dependiente, el investigador debe decidir qué variables independientes incluye en el análisis. Las variables independientes generalmente se seleccionan de dos formas. La primera implica identificar las variables tanto en la investigación previa como desde el modelo teórico que sirve de fundamento a la pregunta de la investigación. La segunda forma es intuitiva, utilizando el conocimiento del investigador y seleccionando intuitivamente las variables para las cuales no existe investigación previa o teoría, pero que lógicamente podrían relacionarse para predecir los grupos de la variable dependiente.

1.1.4.2. Tamaño muestral

El análisis discriminante es bastante sensible a la razón entre el tamaño muestral y el número de variables predictoras. Muchos estudios sugieren una razón de 20 observaciones por cada variable

predictora[3]. Aunque esta razón puede ser difícil de conseguir en la práctica, el investigador debe tener en cuenta que los resultados podrían llegar a ser inestables a medida que el tamaño de la muestra disminuye en relación con el número de variables independientes. El tamaño mínimo recomendado es de cinco observaciones por variable independiente. Nótese que esta razón se aplica a todas las variables consideradas en el análisis, incluso si todas las variables consideradas no entran en la función discriminante (como en la estimación por etapas).

Además del tamaño muestral total, el investigador debe también considerar el tamaño muestral de cada grupo. Como mínimo, el tamaño del grupo más pequeño debe ser mayor que el número de variables independientes. Como una regla práctica, cada grupo debe tener al menos 20 observaciones, el investigador debe también considerar los tamaños relativos de los grupos. Si los grupos varían ampliamente en tamaño, esto puede afectar a la estimación de la función discriminante y a la clasificación de las observaciones. En la etapa de clasificación, los grupos más grandes tienen una posibilidad desproporcionadamente más grande de clasificación. Si los tamaños de los grupos varían de forma importante, puede que el investigador quiera muestrear aleatoriamente desde el grupo más grande, y con ello reducir su tamaño a un nivel comparable con el grupo más pequeño.

1.1.4.3. División de la muestra

La muestra original se divide en dos submuestras; una, utilizada para la estimación de la función discriminante, y otra con fines de validación. Es esencial que cada submuestra tenga un tamaño adecuado para apoyar las conclusiones de los resultados.

Se han sugerido un conjunto de procedimientos para dividir la muestra, pero el más utilizado implica desarrollar la función discriminante con un grupo y luego probarla con un segundo grupo. El procedimiento habitual consiste en dividir aleatoriamente la muestra la muestra total de encuestados en dos grupos. Uno de estos grupos, la muestra de análisis, se usa para construir la función discriminante. El segundo grupo, la ampliación de la muestra, se usa para validar la función discriminante. Este método de validación de la función se denomina ***división de la muestra o enfoque de validación cruzada.***

[3] **Información tomada de** ***Anderson y otros (1999)***

No se ha establecido una manera definitiva para dividir la muestra en los grupos de análisis y ampliación (o validación). El procedimiento más común es dividir el total del grupo, de tal forma que la mitad de los encuestados pertenezca a la muestra de análisis y la otra mitad a la ampliación de la muestra. No obstante, no se ha establecido ninguna regla fiable, y algunos investigadores prefieren una división 60-40 ó 75-25 entre los grupos de análisis y ampliación.

Cuando se seleccionan los individuos para los grupos de análisis y validación, generalmente se sigue un proceso de muestreo estratificado proporcional. Si los grupos categóricos del análisis discriminante están igualmente representados en el total de la muestra, se selecciona un número igual de individuos. Si los grupos categóricos son desiguales, los tamaños de los grupos seleccionados para la ampliación de la muestra deben ser proporcionales a la distribución total de la muestra. Por ejemplo, si una muestra consiste en 50 hombres y 50 mujeres, la ampliación de la muestra tendría 25 hombres y 25 mujeres. Si la muestra contiene 70 mujeres y 30 hombres, entonces la ampliación de la muestra consistiría en 35 mujeres y 15 hombres.

1.1.5. Supuestos del análisis discriminante

Es deseable encontrar ciertas condiciones para la correcta aplicación del análisis discriminante. Los supuestos claves para obtener la función discriminante son el de normalidad multivariante de las variables independientes y el de estructuras (matrices) de covarianza y dispersión desconocidas pero iguales para los grupos. Si los supuestos no se cumplen, el investigador debería identificar los métodos alternativos disponibles y la influencia que cabría esperar sobre los resultados. Los datos que no cumplan el supuesto de normalidad multivariante pueden causar problemas en la estimación de la función discriminante. Por ello, se sugiere que se use la regresión logística como una técnica alternativa, si es posible[4].

Las matrices de covarianzas distintas pueden afectar desfavorablemente al proceso de clasificación. Si los tamaños muéstrales son pequeños y las matrices de covarianzas son distintas, la significación estadística del proceso de estimación se ve afectada desfavorablemente. Finalmente, en muchos de los programas estadísticos están disponibles técnicas de clasificación cuadráticas, si existen grandes diferencias entre las matrices de covarianzas de los grupos y otras soluciones no minimizan el efecto.

[4] **Información tomada de *Anderson y otros (1999)***

Otra característica de los datos que puede afectar a los resultados es la multicolinealidad entre las variables independientes. La multicolinealidad consiste en que dos o más variables independientes están altamente correlacionadas, por lo que una variable puede venir muy bien explicada o predicha por otras variables y, por ello, añadir poca capacidad explicativa al conjunto completo. El investigador, al interpretar la función discriminante, debe conocer el nivel de multicolinealidad y su influencia al determinar que variables entran en la solución por etapas.

Un supuesto implícito, al utilizar funciones discriminantes lineales, es que todas las relaciones son lineales. Las relaciones no lineales no están reflejadas en la función discriminante a menos que se realicen transformaciones específicas de la variable para representar los efectos no lineales. Finalmente, los casos atípicos pueden tener una influencia sustancial en la precisión clasificadora de cualquier resultado del análisis discriminante. Se aconseja al investigador examinar todos los resultados por la presencia de casos atípicos y eliminarlos si fuera necesario.

1.1.6. Estimación del modelo discriminante y valoración del ajuste global

Para obtener la función discriminante, el investigador debe decidir el método de estimación y determinar después el número de observaciones que se van a mantener. Una vez que se han estimado las funciones, puede valorarse el ajuste global del modelo de varias formas. Primero pueden calcularse las **puntuaciones discriminantes Z.** La comparación de las medias de los grupos sobre las puntuaciones Z ofrece una medida de la discriminación entre grupos. La capacidad de la predicción se valora por el número de observaciones clasificadas dentro de los grupos adecuados. Se dispone de vario criterios para valorar si el proceso de clasificación alcanza significación estadística y/o práctica. Finalmente la validación por casos puede identificar la precisión en la clasificación de cada caso y su influencia relativa sobre la estimación global del modelo.

1.1.6.1. Método de cálculo

Se pueden utilizar dos métodos de cálculo para derivar una función discriminante: el método simultaneo (directo) y el método por etapas. La **estimación simultanea** implica el cálculo de la

función discriminante donde todas las variables independientes son consideradas simultáneamente, sin considerar la capacidad discriminante de cada variable independiente.

La **estimación por etapas** [5] es una alternativa al enfoque simultáneo. Incluye las variables independientes dentro de la función discriminante de una en una, según su capacidad discriminatoria. El enfoque por etapas comienza eligiendo la variable que mejor discrimina. La variable inicial se empareja entonces con cada una de las variables independientes(de una en una), y se elige la variable que más consigue incrementar la capacidad discriminante de la función en combinación con la primera variable. La tercera y posteriores variables se seleccionan de una manera similar. Mientras se incluyen variables adicionales, algunas variables seleccionadas previamente pueden ser eliminadas si la información que contienen sobre las diferencias del grupo está contenida en alguna combinación de otras variables incluidas en posteriores etapas. Al final, o bien todas las variables habrán sido incluidas en la función, o se habrá considerado que las variables excluidas no contribuyen significativamente a una mejor discriminación.

El método por etapas es útil cuando el investigador quiere considerar un número relativamente grande de variables independientes para incluir en la función.

El conjunto reducido es generalmente tan bueno como, y algunas veces mejor que, el conjunto completo de variables.

1.1.6.2. Significación estadística

Después de calcularse la función discriminante, el investigador debe valorar el nivel de significación. Se dispone de varios criterios estadísticos. Las medidas del lambda de Wilks, la traza de Hotelling y el criterio de Pillai evalúan la significación estadística de la capacidad discriminatoria de la función(es) discriminante(s). El criterio convencional de 0.05 o superior se utiliza a menudo. El análisis discriminante estima una función discriminante menos que grupos existentes. Todos los programas de computador proporcionan al investigador la información necesaria para averiguar el número de funciones necesarias para obtener significación estadística, sin incluir funciones discriminantes que no incrementen la capacidad discriminatoria significativamente. Si se consideran una o más funciones que no son estadísticamente significativas, el modelo discriminante debería

[5] **Información tomada de *Anderson y otros (1999)***

reestimarse con el número de funciones que se hayan obtenido, limitado por el número de funciones significativas. De esta manera, la valoración de la precisión en la predicción y la interpretación de las funciones discriminantes estarán basadas solamente en funciones significativas.

1.1.7. Valoración del ajuste global

Una vez que se han identificado las funciones discriminantes significativas, la atención se desplaza a averiguar el ajuste global de la(s) función(es) discriminante(s) considerada(s). Esta valoración conlleva a tres tareas: calcular la puntuación Z discriminante para cada observación, evaluar diferencias de grupo sobre las puntuaciones Z discriminantes y valorar la precisión en la predicción de pertenencia al grupo.

1.1.7.1. Cálculo de las puntuaciones Z discriminantes

Como ya hemos dicho, las puntuaciones Z vienen dadas por

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \cdots + W_p X_{pk}$$

donde

$Z_{jk} =$ Puntuación z discriminante de la función discriminante j para el objeto k

$a =$ Constante

$W_i =$ Ponderación discriminante para la variable independiente i

X_{ik} =Variable independiente i para el objeto k

Estas puntuaciones Z pueden emplear valores y ponderaciones estandarizados o no estandarizados. La versión estandarizada es más útil en la interpretación, pero la versión no estandarizada es más fácil de utilizar en el cálculo de la puntuación Z discriminante.

Debemos darnos cuenta de que la ***función discriminante*** difiere de la ***función de clasificación***, también conocida como la ***función discriminante lineal de Fisher***. Las funciones de clasificación, una para cada grupo, pueden utilizarse al clasificar observaciones. En este método de clasificación,

unos valores de la observación para las variables independientes se incluyen en las funciones de clasificación y se calcula una puntuación de clasificación para cada grupo para esa observación. La observación se clasifica entonces en el grupo con la mayor puntuación de clasificación. Utilizamos la función discriminante como el medio de clasificar porque ofrece una representación resumida y simple de cada función discriminante, simplificando el proceso de interpretación y la valoración de la contribución de las variables independientes.

1.1.7.2. Evaluación de diferencias entre grupos

Una forma de valorar el ajuste global del modelo es determinar la magnitud de las diferencias entre los miembros de cada grupo en términos de las puntuaciones Z discriminantes. Una medida resumen de las diferencias entre grupos es una comparación de los ***centroides*** de grupo, la puntuación Z discriminante media para todos los miembros del grupo. Una medida de éxito del análisis discriminante es su capacidad para definir funciones discriminantes que den lugar a centroides de grupo significativamente diferentes. Las diferencias entre centroides se miden en términos de la medida D^2 de Mahalanobis, para la cual se dispone de contrastes que determinan si las diferencias son significativamente distintas.

1.1.7.3. Valorando la exactitud en la predicción de pertenencia al grupo

Aquí se valora si cada observación es correctamente clasificada. Para hacer esto, deben realizarse una serie de consideraciones: la razón de ser práctica y estadística para elaborar matrices de clasificación, la determinación de la puntuación de corte, la construcción de matrices de clasificación y los estándares para valorar la exactitud clasificatoria.

Por qué se elaboran matrices de clasificación Los contrastes estadísticos para valorar la significación de las funciones discriminantes no informan sobre lo correctamente que predice la función. Para determinar la capacidad predictiva de una función discriminante, el investigador debe construir matrices de clasificación donde se revele la razón de aciertos o porcentaje correctamente clasificados.

Determinación de la puntuación de corte Antes de construir la matriz de clasificación el investigador debe determinar la puntuación de corte. La **puntuación de corte** es el criterio (puntuación) frente al cual cada puntuación discriminante individual es comparada para determinar dentro de qué grupo debe ser clasificado cada objeto.

Si los tamaños de grupo son iguales, la puntuación óptima es

$$Z_{CE} = \frac{Z_A + Z_B}{2}$$

donde

Z_{CE} = Valor de la puntuación de corte crítica para grupos de igual tamaño
Z_A = Centroide del grupo A
Z_B = Centroide del grupo B

Cuando la muestra es tomada aleatoriamente de la población, la mejor estimación de las probabilidades son las proporciones muéstrales.

Si los tamaños de grupo son distintos, la puntuación óptima es

$$Z_{CU} = \frac{N_B Z_A + N_A Z_B}{N_A + N_B}$$

donde

Z_{CU} = Valor de la puntuación de corte crítica para grupos de distinto tamaño
N_A = Número del grupo A
N_B = Número del grupo B
Z_A = Centroide del grupo A
Z_B = Centroide del grupo B

Construcción de las matrices de clasificación Partimos de que tenemos la muestra dividida en dos grupos, la muestra de análisis y la muestra de validación. Con la muestra de validación elaboramos la matriz de clasificación. El proceso consiste en multiplicar las ponderaciones generadas por la muestra de análisis por las medidas de la variable primaria de la muestra de validación. Después, las puntuaciones discriminantes individuales para la muestra de validación se comparan con el valor de la puntuación de corte crítica y se clasifica de la siguiente forma:

Clasificar un individuo dentro del grupo A si $Z_n < Z_{ct}$.

o

Clasificar un individuo dentro del grupo B si $Z_n > Z_{ct}$.

donde

Z_n = Puntuación Z discriminante para el individuo n-ésimo

Z_{ct} = Valor de la puntuación de corte crítica.

Medición de la capacidad predictiva mediante la aleatoriedad La capacidad predictiva de la función discriminante se mide con la razón de aciertos, el cual se obtiene en la matriz de clasificación. El investigador podría preguntarse sobre qué se considera un nivel aceptable de capacidad predictiva para una función discriminante. Por ejemplo, ¿es el 60% un nivel aceptable o debería esperarse un 80% o un 90% de la capacidad predictiva? Para responder a esta pregunta, el investigador debe determinar primero el porcentaje que podría ser clasificado correctamente de forma aleatoria (sin la ayuda de la función discriminante).

Determinación del criterio basado en la aleatoriedad[6]

Cuando los tamaños muéstrales son iguales, la determinación de la clasificación aleatoria es bastante simple; la probabilidad se obtiene dividiendo 1 por el número de grupos. Por ejemplo, en una función de dos grupos la probabilidad sería de 0.5; para una función de tres grupos la probabilidad sería de 0.33, y así sucesivamente.

Si los tamaños de los grupos son distintos, un criterio de clasificación aleatoria es basarse en el tamaño muestral del grupo más grande, este criterio es conocido como **el criterio de máxima aleatoriedad.** Se determina calculando el porcentaje de la muestra completa representado por el más grande de los dos (o más) grupos. Por ejemplo, si los tamaños de los grupos son 65 y 35, el criterio de máxima aleatoriedad es el 65 por ciento de clasificaciones correctas. Por tanto, si la razón de aciertos por la función discriminante no excede el 65% , entonces no nos ayudaría a predecir según este criterio. Este criterio debería utilizarse cuando el único objetivo del análisis discriminante es maximizar el porcentaje clasificado correctamente.

Otro criterio para tamaños de grupos desiguales es el **criterio de aleatoriedad proporcional**, la fórmula para este criterio es

$$C_{PRO} = p^2 + (1-p)^2$$

[6] **Información tomada de *Anderson y otro (1999).***

donde

$p =$ Proporción de individuos del grupo 1

$1 - p =$ Proporción de individuos del grupo 2

Al aplicar algún criterio de aleatoriedad, también es sugerible seguir el siguiente criterio: la precisión clasificatoria debería ser por lo menos un cuarto mayor que aquella obtenida por la aleatoriedad. Por ejemplo, si la precisión aleatoria es del 50%, la precisión clasificatoria debería ser del 62.5%. El criterio es fácil de aplicar con grupos de igual tamaño. Con grupos de tamaño diferente, se alcanza una cota superior cuando se utiliza el modelo de aleatoriedad máxima para determinar la precisión aleatoria.

Medidas de precisión clasificatoria fundamentadas estadísticamente relacionadas con la aleatoriedad

Un contraste estadístico para contrastar la capacidad discriminatoria de la matriz de clasificación cuando se compara con un modelo de aleatoriedad es el **estadístico** Q **de Press**[7]. Esta medida sencilla compara el número de clasificaciones correctas con el tamaño muestral total y el número de grupos. Se compara el valor hallado con un valor crítico (el valor de la chi-cuadrado para un grado de libertad al nivel de confianza deseado). Si éste excede el valor crítico, la matriz de clasificación puede considerarse estadísticamente mejor que la aleatoriedad. El estadístico Q se calcula mediante la siguiente fórmula:

$$Q \; de \;\; \text{Pr}ess = \frac{[N - (nK)]^2}{N(K-1)}$$

donde

$N =$ tamaño muestral total
$n =$ número de observaciones correctamente clasificadas

$K =$ número de grupos

[7] **Información de** ***Anderson y otros (1999)***

En el análisis discriminante múltiple, si tenemos G grupos o G clasificaciones, sólo necesitamos determinar

$$r = min(G-1, p)$$

vectores W de ponderaciones y consecuentemente r funciones discriminante Z[8] . En primer lugar, observamos que, necesitamos construir

$$2\binom{G}{2} = G(G-1)$$

vectores W para los G grupos, pero si el mínimo es G-1, sólo determinamos los primeros G-1 de los W y el resto quedan determinados por los G-1 primeros. Esto es como sigue:

Suponemos que los grupos tienen distribuciones normales p-dimensionales con la matriz de varianzas covarianzas iguales y si μ_i y μ_j son las medias de los grupos i y j entonces, el vector de ponderaciones para los grupos i y j es

$$W_{i,j} = V^{-1}(\mu_i - \mu_j)$$

así, podemos determinar los vectores $W_{i,i+1}$, para $i = 1,2,...,G-1$ y obtener cualquier otro a partir de estas G-1 direcciones. Por ejemplo:

$$W_{i,i+2} = V^{-1}(\mu_i - \mu_{i+2}) = V^{-1}(\mu_i - \mu_{i+1}) + V^{-1}(\mu_{i+1} - \mu_{i+2}) = W_{i,i+1} + W_{i+1,i+2}$$

Cuando $p \le G-1$, como estos vectores pertenecen a $\mathbb{R}^p$, el número máximo de vectores linealmente independientes es p y el resto quedan determinados por los primeros p vectores W.

La frontera de separación entre los grupos i y j es el conjunto

$$\left\{ X \in \mathbb{R}^p / W_{ij}^t X = \frac{1}{2} W_{ij}^t (\mu_i + \mu_j) \right\}$$

La función discriminante Z para los grupos i y j es

$$Z_{ij} = W_{ij}^t X - \frac{1}{2} W_{ij}^t (\mu_i + \mu_j) \quad con \quad X = (x_{1k}, x_{2k}, ..., x_{pk})^t$$

Donde k denota el número del individuo en los grupos i y j .

[8] **Información de *Peña (2002)***

Si $Z_{ij} > 0$ para todo $j \neq i$ clasificamos a X en el grupo i.

Propiedades de Z_{ij}[9]:

1. $Z_{ij}(X) = -Z_{ji}(X)$
2. $Z_{rs}(X) = Z_{is}(x) - Z_{ir}(X)$

Si hay G grupos, para clasificar un individuo X se tienen que determinar un total de

$$2\binom{G}{2} = G(G-1)$$

funciones discriminantes Z_{ij} para todo $j \neq i$.

Por las propiedades de Z_{ij} sólo necesitamos encontrar una cantidad de

$$r = min(G-1, p)$$

de ellas y las demás son función de las primeras r.

1.1.8. Contrastes en el análisis discriminante

En el análisis discriminante con G grupos, previamente, se deben realizar los siguientes contrastes:

a) Hipótesis de homocedasticidad.

b) Hipótesis de normalidad.

c) Hipótesis de diferencia entre las medias poblacionales de los G grupos.

La hipótesis de homocedasticidad asume que la matriz de covarianzas de los G grupos es constante igual a $\sum$.

La hipótesis de normalidad asume que cada uno de los grupos tienen distribución normal multivariante, es decir, $x_g \rightarrow N(\mu_g, \Sigma)$ para $g = 1, 2, ..., G$.

La respuesta que se de a la hipótesis c) es crucial para la justificación de la realización del análisis discriminante. En el caso de que la respuesta sea negativa carecería de interés continuar con el análisis, ya que significaría que las variables introducidas como variables clasificadoras no tienen capacidad discriminante significativa**.**

[9] **Información de *Cuadras (2008)***

1.1.9. Contraste de igualdad de varias medias multivariadas

Supongamos que hemos observado una muestra de tamaño *n* de una variable p dimensional que puede estratificarse en G clases o grupos con n_g observaciones cada uno para $g = 1, 2, ..., G$. Un problema importante es contrastar que las medias de las G clases o grupos son iguales. La hipótesis a contrastar es:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_G = \mu$$

donde, además, $\sum$ es la matriz de varianza covarianza, es definida positiva, e idéntica en los grupos. La hipótesis alternativa es:

$$H_1 : no\ todas\ las\ \mu_i\ son\ iguales$$

con las mismas condiciones para $\sum$.

El test de la razón de verosimilitudes[10] es

$$\lambda = n \ln\left(\frac{|S|}{|S_w|}\right)$$

donde la función $|.|$ es el determinante y

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t ;\ \ donde\ \ x_i = (x_{i1}, x_{i2}, ..., x_{ip})^t\ \ y\ \ \bar{x} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)^t$$

$$S_w = \frac{1}{n}W\ \ ;\ \ \ donde\ \ \ W = \sum_{g=1}^{G}\sum_{h=1}^{n_g}(x_{hg} - \bar{x}_g)(x_{hg} - \bar{x}_g)^t$$

La matriz W es conocida como la *suma de cuadrados dentro de los grupos*.

El estadístico λ tiene asintóticamente una chi-cuadrada con g grados de libertad, donde $g = p(G-1)$.

Rechazamos H_0 a un nivel α si $\lambda > \chi^2_{\alpha,g}$.

[10] **Información de *Peña (2002)***

1.1.10. Contraste de igualdad de matrices de varianzas covarianzas

Consideremos las poblaciones normales p-dimensionales $N(\mu_i, \sum_i)$ *para* $i = 1, 2, ..., G$

Estamos interesados en realizar el siguiente contraste:

$$H_0 : \sum_1 = \sum_2 = \cdots = \sum_G$$

este contraste se resuelve mediante la prueba de la razón de verosimilitud[11]

$$\lambda_R = \frac{|S_1|^{n_1/2} \times \cdots \times |S_G|^{n_G/2}}{|S|^{n/2}}$$

donde S_i es la matriz de varianzas covarianzas de los datos de la población i, estimación máximo verosímil de $\sum_i$ y

$$n = n_1 + \cdots + n_G$$
$$S = \frac{1}{n}(n_1 S_1 + \cdots + n_G S_G) = \frac{W}{n}$$

es la estimación máximo verosímil de $\sum$, matriz de covarianzas común bajo H_0. Rechazamos H_0 si el estadístico

$$-2\ln(\lambda_R) = n\ln|S| - (n_1 \ln|S_1| + \cdots + n_G \ln|S_G|) \sim \chi^2_q$$

es significativo, donde $q = Gp(p+1)/2 - p(p+1)/2 = (G-1)p(p+1)/2$ son los grados de libertad de la ji-cuadrado. Si rechazamos H_0, entonces resulta que no disponemos de unos ejes comunes para representar todas las poblaciones (la orientación de los ejes viene dada por la matriz de covarianzas).

Debido a que la prueba anterior puede ser sesgada, conviene aplicar la corrección de Box,

$$c(n-G)\ln|S| - ((n_1 - 1)\ln|\hat{S}_1| + \cdots + (n_G - 1)\ln|\hat{S}_G|)$$

donde $\hat{S}_i = \frac{n_i}{n_i - 1} S_i$, y la constante c es

[11] **Información de *Peña (2002)***

$$c = \left[1 - \left(\frac{2p^2 + 3p - 1}{6(p+1)(G-1)} \right) \left(\sum_{k=1}^{G} \frac{1}{n_G - 1} - \frac{1}{n-G} \right) \right]$$

1.1.11. Contraste de normalidad multivariante

Partimos de que se tienen n vectores $p-$dimencionales , digamos, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^t \quad para \quad i = 1, 2, ..., n$ y el objetivo es estudiar si esta muestra viene de una normal $p-$dimencional .

Sabemos que,

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^t ; \quad donde \quad \bar{x} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)^t$$

y que la distancia de Mahalanobis d_{ij} entre x_i y x_j es[12]

$$d_{ij} = (x_i - \bar{x})^t S^{-1} (x_j - \bar{x})$$

Definimos por A_p y K_p como el coeficiente de asimetría y curtosis multivariantes respectivamente y,

$$A_p = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^3$$

$$K_p = \frac{1}{n} \sum_{i=1}^{n} d_{ii}^2$$

Si los datos vienen de una normal multivariada, asintóticamente se verifica:

$$\frac{nA_p}{6} \sim \chi_f^2 \quad con \quad f = \frac{1}{6} p(p+1)(p+2)$$

$$K_p \sim N\left(p(p+2); \frac{8p(p+2)}{n} \right)$$

La potencia de esta prueba no es muy alta a no ser que se tenga una muestra relativamente grande.

[12] **Información de** ***Peña (2002)***

1.1.12. Análisis discriminante en poblaciones desconocidas

Partimos de que tenemos G poblaciones posibles. Como caso particular, la clasificación clásica es para $G = 2$. La matriz general de datos X, de dimensión nxp, (n individuos y p variables), puede considerarse ahora en G matrices, correspondientes a las subpoblaciones.

Vamos a llamar x_{ijg} a los elementos de estas submatrices, donde i representa el individuo, j la variable y g el grupo o submatriz. Llamamos n_g al número de elementos en el grupo g y el número total de observaciones es:

$$n = \sum_{g=1}^{G} n_g$$

Vamos a llamar $x'_{ig} = (x_{i1g}, ..., x_{ipg})$ al vector fila 1xp que contiene los p valores de las variables para el individuo i en el grupo g. El vector de medias dentro de cada subpoblación será:

$$\overline{x}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{ig}$$

La matriz de varianzas covarianzas para los elementos de la subpoblación g será:

$$\hat{S}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (x_{ig} - \overline{x}_g)(x_{ig} - \overline{x}_g)'$$

donde se ha dividido por $n_g - 1$ para tener estimaciones centradas de las varianzas y covarianzas.

Si suponemos que las G poblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de las matrices de covarianza en cada población, con peso proporcional a su precisión. Por tanto:

$$\hat{S}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \hat{S}_g$$

Para obtener la funciones discriminantes utilizamos $\overline{x}_g$ como una estimación de μ_g, y $\hat{S}_w$ como la estimación de la matriz de varianzas covarianzas poblacional $\sum$. Clasificamos un nuevo elemento x_0 en aquella población g donde

$$\min_g (x_0 - \bar{x}_g)' \hat{S}_w^{-1} (x_0 - \bar{x}_g) = \min_g \hat{T}_g' (\bar{x}_g - x_0)$$

es decir en el grupo g, en el que la distancia de Mahalanobis entre x_0 y $\bar{x}_g$ sea la más pequeña.

1.2.1. Introducción al análisis de regresión logística

La regresión logística son las técnicas estadísticas apropiadas cuando la variable dependiente es categórica (nominal o no métrica) y las variables independiente son métricas.

La regresión logística también conocida como análisis logit, está restringida en su forma básica a dos grupos, aunque en formulaciones alternativas puede considerar más de dos grupos considerando la distribución multinomial.

1.2.2. Modelos lineales generalizados

Consideremos particularmente una variable aleatoria univariante Y cuya distribución de probabilidades depende solamente del parámetro θ y ϕ es el parámetro de dispersión que se supone conocido o se estima. La distribución pertenece a la familia exponencial si puede escribirse de la forma[13]

$$f(y;\theta,\phi) = \exp\left(\frac{t(y)q(\theta) + r(\theta)}{p(\phi)} + s(y,\phi) \right) \qquad \text{(A)}$$

Donde $p(.), t(.), q(.), r(.)$ y $s(.)$ son funciones conocidas para cada variable aleatoria Y.

Si t(y) = y, se dice que la distribución está escrita en la forma canónica o estándar y $q(\theta)$ es llamada algunas veces el parámetro natural de la distribución.

Un modelo lineal generalizado tiene tres componentes:

1. *La componente aleatoria* que consiste en variables de respuesta, $Y_1, ..., Y_n$, independientes con funciones de densidad de la familia exponencial que son de la forma estándar.

[13] **Para modelos lineales generalizados se siguieron los libros:** ***Lindsey (1997), Dobson (2002) , Hardin (2007), McCullagh y Nelder (1989)***

2. *La componente sistemática* consiste en el conjunto de predictores lineales $(\eta_1,...,\eta_n)$ tales que

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, \ldots, n$$

Para p variables explicativas recogidas en el vector

$$x_i' = (x_{i1}, \ldots, x_{ip})$$

3. *La tercera componente* consiste en la función de enlace o nexo $g(.)$ que conecta las componentes aleatoria y sistemática . Sea $\mu_i = E[Y_i]$, $i = 1,2,3,...,n$

El modelo enlaza μ_i a η_i por $\eta_i = g(\mu_i)$, es decir, $g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}$, $i = 1,2,...,n$.

El nexo $g(.)$ es una función monótona y diferenciable.

La función de enlace o nexo $g(.)$ que transforma la media a el parámetro natural es llamado el enlace canónico o nexo canónico.

Esto es, $g(\mu_i) = q(\theta_i)$ y

$$q(\theta_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, \ldots, n$$

1.2.2.1. El modelo de regresión logística

El modelo de regresión logística es un modelo lineal generalizado que describimos a continuación:

Consideremos $Y_i = Bernoulli(\pi_i)$ para $i = 1,2,...,n$ y que son independientes, entonces

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left\{ y_i\left[\ln(\pi_i) - \ln(1-\pi_i)\right] + \ln(1-\pi_i) + \ln\binom{1}{y_i} \right\}$$

$$= \exp\left\{ \frac{y_i q(\pi_i) - b(\pi_i)}{a_i(\phi)} + c(y_i,\phi) \right\}$$

donde

$$E[Y_i] = \pi_i, \quad q(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \log it(\pi_i) = g(\pi_i) \quad \textit{que es el enlace canonico}$$

$$b(\pi_i) = -\ln(1-\pi_i),\ a(\phi) = 1,\ c(y_i, \phi) = \ln\binom{1}{y_i}$$

El predictor lineal es $\eta_i = g(\pi_i) = x_i'\beta$ es decir

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = q(\pi_i) = \sum_{j=1}^{p} x_{ij}\beta_j$$

y despejando π_i se tiene:

$$\pi_i = \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\sum_{j=1}^{p} x_{ij}\beta_j}} = \frac{1}{1 + e^{-\sum_{j=1}^{p} x_{ij}\beta_j}}$$

Es decir,

$$\hat{\pi}_i = \frac{1}{1 + e^{-\sum_{j=1}^{p} x_{ij}\hat{\beta}_j}}$$

El modelo

$$\hat{\pi}_i = \frac{1}{1 + e^{-\sum_{j=1}^{p} x_{ij}\hat{\beta}_j}}$$

Es llamado *el modelo de regresión logística* con enlace *logit* y representa la probabilidad estimada de tener éxito en el i-ésimo ensayo de n experimentos de Bernoulli, dado el vector i de las variables explicativas.

1.2.2.2. El modelo de regresión de Poisson

Consideremos $Y_i = Poisson(\mu_i)$ para $i = 1, 2, ..., n$ y que son independientes, entonces

$$f(y_i; \mu_i) = \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} = \exp\{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\}$$

$$= \exp\left\{\frac{y_i \ln(\mu_i) - b(\mu_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

donde

$$E[Y_i] = \mu_i \ ,$$
$$q(\mu_i) = \ln(\mu_i) = g(\mu_i) \ \ que\ es\ \ el\ \ \ enlace\ canonico.$$
$$b(\mu_i) = \mu_i, \ a(\phi) = 1, \ c(y_i, \phi) = -\ln(y_i!)$$

El predictor lineal es $\eta_i = g(\mu_i) = x_i'\beta$ y $x_i'\beta = \sum_{j=1}^{p} x_{ij}\beta_j$, es decir,

$$\mu_i = \exp\left[\sum_{j=1}^{p} x_{ij}\beta_j\right]$$

que muestralmente es

$$\hat{\mu}_i = \exp\left[\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right]$$

Este modelo es conocido como *el modelo de regresión de Poisson* con enlace canónico logaritmo natural y representa una estimación para el promedio del individuo *i* dado el vector *i* de las variables explicativas.

1.2.3. Estimación por máxima verosimilitud en los modelos lineales generalizados

Dadas las observaciones independientes $Y_1 = y_1, ..., Y_n = y_n$, con funciones de densidad de probabilidad que pertenecen a la familia exponencial canónica o estándar, es decir:

$$f(y_i;\theta_i,\phi)=\exp\left\{\frac{y_i q(\theta_i)-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right\},\quad para\ \ i=1,2,...,n.$$

Si consideramos la transformación $\omega_i = q(\theta_i)$ y si esta es uno a uno, la densidad anterior puede escribirse de la siguiente forma:

$$f(y_i;\omega_i,\phi)=\exp\left\{\frac{y_i\omega_i-d(\omega_i)}{a_i(\phi)}+c(y_i,\phi)\right\},\quad para\ \ i=1,2,...,n$$

La función de verosimilitud es

$$L(\omega,\phi,y)=\prod_{i=1}^{n} f(y_i;\omega_i,\phi)=\exp\left\{\sum_{i=1}^{n}\frac{y_i\omega_i-d(\omega_i)}{a_i(\phi)}+\sum_{i=1}^{n}c(y_i,\phi)\right\}$$

donde $y=(y_1,...,y_n)$ y $\omega=(\omega_i,...,\omega_n)$ es el parámetro natural.

La función de log-verosimilitud es

$$l(\omega,\phi,y)=\ln[L(\omega,\phi,y)]=\sum_{i=1}^{n}\frac{y_i\omega_i-d(\omega_i)}{a_i(\phi)}+\sum_{i=1}^{n}c(y_i,\phi)$$

Supongamos que el parámetro de escala, ϕ, es conocido. Estamos interesados en estimar $\beta=(\beta_1,...,\beta_p)$ por máxima verosimilitud. El estimador máximo verosímil,

$$\hat{\beta}=\underset{\beta}{argmax}\, l(\omega,\phi,y)$$

se obtiene resolviendo el sistema de ecuaciones de verosimilitud

$$U_j=\frac{\partial l(\omega,\phi,y)}{\partial\beta_j}=0,\quad j=1,2,...,p$$

ó equivalentemente $U=(U_1,...,U_p)^t=0_{p\times 1}$, en notación vectorial.

Proposición 2.1. Las ecuaciones de verosimilitud son

$$0=\sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ij}}{V[Y_i]g'(\mu_i)},\quad j=1,2,...,p$$

y en forma matricial $X^t M^{-1}(y-\mu)=0$, donde

$$X=(x_{ij})_{n\times p}\ ,\ y=(y_1,...,y_p)^t\ ,\ \mu=(\mu_1,...,\mu_p)^t\ ,\ M=diag(m_1,...,m_n),\ con\ \ m_i=V[Y_i]g'(\mu_i)$$

Demostración. En la función de log-verosimilitud,

$$l(\omega,\phi,y)=\ln[L(\omega,\phi,y)]=\sum_{i=1}^{n}\frac{y_i\omega_i-d(\omega_i)}{a_i(\phi)}+\sum_{i=1}^{n}c(y_i,\phi)=\sum_{i=1}^{n}l_i(\omega_i,\phi,y_i)$$

se verifica la relación de dependencia

$$l_i\leftarrow\omega_i\leftarrow\mu_i\leftarrow\eta_i\leftarrow\beta$$

donde $\omega_i \leftarrow \mu_i$ significa que ω_i es función de μ_i. Primero vemos que $E[Y_i]=\mu_i=d'(\omega_i)$,

Sea $U=\dfrac{\partial \ln f(y_i;\omega_i,\phi)}{\partial\omega_i}=\dfrac{\partial \ln f_{\omega_i}}{\partial\omega_i}$, entonces

$$E_{\omega_i}[U]=E\left[\frac{\partial \ln f(y_i;\omega_i,\phi)}{\partial\omega_i}\right]=\int_{Y_i}\frac{\partial \ln f_{\omega_i}}{\partial\omega_i}f_{\omega_i}dy_i=\int_{Y_i}\frac{\frac{\partial f_{\omega_i}}{\partial\omega_i}}{f_{\omega_i}}f_{\omega_i}dy_i=\int_{Y_i}\frac{\partial f_{\omega_i}}{\partial\omega_i}dy_i=\frac{\partial}{\partial\omega_i}\int_{Y_i}f_{\omega_i}dy_i=\frac{\partial}{\partial\omega_i}(1)=0$$

Entonces

$$0=E\left[\frac{\partial \ln f(y_i;\omega_i,\phi)}{\partial\omega_i}\right]=E\left[\frac{\partial}{\partial\omega_i}\left(\frac{y_i\omega_i-d(\omega_i)}{a_i(\phi)}+c(y_i,\phi)\right)\right]=E\left[\frac{y_i-d'(\omega_i)}{a_i(\phi)}\right]=E[Y_i]-d'(\omega_i)$$

Por tanto

$$\mu_i=E[Y_i]=d'(\omega)$$

$$0=\frac{\partial}{\partial\omega_i}\int_{Y_i}\frac{\partial \ln f_{\omega_i}}{\partial\omega_i}f_{\omega_i}dy_i=\int_{Y_i}\frac{\partial}{\partial\omega_i}\left[\frac{\partial \ln f_{\omega_i}}{\partial\omega_i}f_{\omega_i}\right]dy_i=\int_{Y_i}\left[\frac{\partial^2 \ln f_{\omega_i}}{\partial\omega_i^{\,2}}f_{\omega_i}+\frac{\partial \ln f_{\omega_i}}{\partial\omega_i}\frac{\partial f_{\omega_i}}{\partial\omega_i}\right]dy_i$$

$$=\int_{Y_i}\left[\frac{\partial^2 \ln f_{\omega_i}}{\partial\omega_i^{\,2}}f_{\omega_i}+\left(\frac{\partial \ln f_{\omega_i}}{\partial\omega_i}\right)^2 f_{\omega_i}\right]dy_i=E_{\omega_i}\left[\frac{\partial}{\partial\omega_i}U\right]+E_{\omega_i}\left[U^2\right]$$

Por tanto,

$$E_{\omega_i}\left[U^2\right] = -E_{\omega_i}\left[\frac{\partial}{\partial\omega_i}U\right]$$

A sí,

$$V_{\omega_i}\left[U\right] = E_{\omega_i}\left[U^2\right] = -E_{\omega_i}\left[\frac{\partial}{\partial\omega_i}\left(\frac{\partial}{\partial\omega_i}\left(\frac{y_i\omega_i - d(\omega_i)}{a_i(\phi)} + c(y_i,\phi)\right)\right)\right] = -E_{\omega_i}\left[\frac{\partial}{\partial\omega_i}\left(\frac{y_i - d'(\omega_i)}{a_i(\phi)}\right)\right] = \frac{d''(\omega_i)}{a_i(\phi)}$$

También

$$E_{\omega_i}\left[U^2\right] = E_{\omega_i}\left[\left(\frac{y_i - d'(\omega_i)}{a_i(\phi)}\right)^2\right] = \frac{E_{\omega_i}\left[Y_i^2\right] - (d'(\omega_i))^2}{\left(a_i(\phi)\right)^2} = \frac{V_{\omega_i}\left[Y_i\right]}{\left(a_i(\phi)\right)^2}$$

Por tanto

$$V_{\omega_i}\left[Y_i\right] = a_i(\phi)d''(\omega_i)$$

Ahora vemos las siguientes relaciones:

$$(a)\quad l_i = l_i(\omega_i) = \frac{y_i\omega_i - d(\omega_i)}{a_i(\phi)} + c(y_i,\phi) \Rightarrow \frac{\partial l_i}{\partial\omega_i} = \frac{y_i - d'(\omega)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}.$$

$$(b)\quad \mu_i = \mu_i(\omega_i) = d'(\omega_i) \Rightarrow \omega_i = \omega_i(\mu_i) = (d')^{-1}(\mu_i) \Rightarrow \frac{\partial\omega_i}{\partial\mu_i} = \frac{1}{\frac{\partial\mu_i}{\partial\omega_i}} = \frac{1}{d''(\omega_i)}.$$

$$(c)\quad \eta_i = \eta_i(\mu_i) = g(\mu_i) \Rightarrow \mu_i = \mu_i(\eta_i) = g^{-1}(\eta_i) \Rightarrow \frac{\partial\mu_i}{\partial\eta_i} = \frac{1}{\frac{\partial\eta_i}{\partial\mu_i}} = \frac{1}{g'(\mu_i)}.$$

$$(d)\quad \eta_i = \eta_i(\beta) = x_i^t\beta \Rightarrow \frac{\partial\eta_i}{\partial\beta_j} = x_{ij}.$$

De aplicar la regla de la cadena y la igualdad $V_{\omega_i}\left[Y_i\right] = a_i(\phi)d''(\omega_i)$, se deduce que

$$\frac{\partial l_i}{\partial\beta_j} = \frac{\partial l_i}{\partial\omega_i}\frac{\partial\omega_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} = \frac{\left(y_i - \mu_i\right)x_{ij}}{a_i(\phi)d''(\omega_i)g'(\mu_i)} = \frac{\left(y_i - \mu_i\right)x_{ij}}{V\left[Y_i\right]g'(\mu_i)}$$

Finalmente, sustituyendo en las ecuaciones de verosimilitud

$$\sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} l_i(\omega_i, \phi, y_i) = 0, \qquad j = 1, ..., p$$

Se obtiene el resultado enunciado; es decir

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{a_i(\phi) d''(\omega_i) g'(\mu_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{V[Y_i] g'(\mu_i)} = 0, \qquad j = 1, ..., p.$$

Ejemplo 2.1. Aplicación de la proposición para deducir las ecuaciones de verosimilitud para el modelo de Poisson con dos parámetros.

Consideremos el modelo $Y_i \overset{d}{=} Poiss(\mu_i), \quad con \quad \ln(\mu_i) = \beta_0 + \beta_i x_i, \ i = 1, ..., n.$ En este caso $\omega_i = \ln(\mu_i), \ g(x) = \ln x, \ d(\omega_i) = e^{\omega_i} \ y \ a_i(\phi) = 1$. Así pues, $g'(\mu_i) = \frac{1}{\mu_i} \quad y \quad d''(\omega_i) = e^{\omega_i} = \mu_i$. Por último, la matriz de diseño o de datos es

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Aplicando la proposición, se obtiene el siguiente sistema de ecuaciones de verosimilitud.

$$0 = \frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \exp\{\beta_0 + \beta_1 x_1\}] \Rightarrow n\bar{y} = e^{\beta_0} \sum_{i=1}^{n} \exp\{\beta_1 x_i\}$$

$$0 = \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} [x_i y_i - x_i \exp\{\beta_0 + \beta_1 x_1\}] \Rightarrow \sum_{i=1}^{n} x_i y_i = e^{\beta_0} \sum_{i=1}^{n} x_i \exp\{\beta_1 x_i\}$$

Que es un sistema de dos ecuaciones no lineal en las variables $\beta_0 \ y \ \beta_1$ y que no se puede resolver explícitamente. Es necesario acudir a métodos numéricos para obtener aproximaciones de $\beta_0 \ y \ \beta_1$.

1.2.4. Métodos numéricos para la obtención de estimadores máximo-verosímiles

Las ecuaciones de verosimilitud no siempre proporcionan soluciones explícitas para β_j, $j = 1,...., p$. De ahí la necesidad de disponer de métodos numéricos para obtener estimaciones máximo verosímiles, $\hat{\beta}_j$, $j = 1,...p$, y en consecuencia obtener el ajuste

$$\hat{\mu}_i = g^{-1}(x_i{}^t \hat{\beta}).$$

1.2.4.1. El método de Newton-Raphson

El algoritmo de Newton-Raphson se apoya en el desarrollo en serie de Taylor. Si β^* es solución de las ecuaciones de verosimilitud; es decir;

$$U(\beta^*) = 0,$$

y $\beta^{(0)}$ es un valor arbitrario de β, entonces el desarrollo de Taylor de primer orden garantiza la siguiente aproximación

$$0 = U(\beta^*) \cong U(\beta^{(0)}) + H(\beta^{(0)})(\beta^* - \beta^{(0)})$$

donde $H = \frac{\partial U}{\partial \beta} = \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right)$ es la matriz **Hessiana**. Despejando β^* de la aproximación, se obtiene

$$\beta^* \cong \beta^{(0)} - H^{-1}(\beta^{(0)})U(\beta^{(0)})$$

que sirve de base para plantear la ecuación recurrente

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} - H^{-1}(\hat{\beta}^{(r-1)})U(\hat{\beta}^{(r-1)}),$$

donde

$U = \left(\frac{\partial l}{\partial \beta_1}, ..., \frac{\partial l}{\partial \beta_p} \right)^t$, $\hat{\beta}^{(r)} = (\hat{\beta}_1^{(r)}, ..., \hat{\beta}_p^{(r)})^t$, $H = \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right)_{j,k=1,...,p}$, H es conocida como la matriz **Hessiana** de orden $p \times p$ y esta formada por las segundas derivadas parciales respecto a los parámetros betas.

Además, $\hat{\beta}^{(r)}$ es el valor estimado de $\hat{\beta}$ en la r -ésima iteración del algoritmo y $H^{-1}(\hat{\beta}^{(r-1)}), U(\hat{\beta}^{(r-1)})$ son H^{-1} y U evaluadas en $\hat{\beta}^{(r-1)}$.

1.2.4.2. El método de puntuaciones de Fisher

El método de puntuaciones de Fisher utiliza el mismo algoritmo de Newton-Raphson, pero sustituye la matriz Hessiana H, por su esperanza; es decir, por la matriz información de Fisher cambiada de signo $I = -E[H]$. La ecuación recurrente sería

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} + I^{-1}(\hat{\beta}^{(r-1)})U(\hat{\beta}^{(r-1)})$$

donde

$$I = -E[H] = \left(-E\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] \right)_{j,k=1,...,p}$$

Observación. La estimación inicial de β para los métodos de Newton-Raphson y de puntuaciones de Fisher se pude obtener tomando $\hat{\mu}^{(0)} = y$. En tal caso

$g(y) = X\hat{\beta}^{(0)} \Rightarrow X^t g(y) = X^t X \hat{\beta}^{(0)} \Rightarrow \hat{\beta}^{(0)} = \left(X^t X \right)^{-1} X^t g(y)$.

1.2.4.3. Mínimos cuadrados ponderados iterados

Consideremos el modelo lineal normal $Y = X\beta + e$, con $e = N_n(0, \sigma^2 V)$. Supongamos que $V_{n \times n}$ es una matriz conocida, definida positiva y simétrica, entonces existe una matriz invertible $K_{n \times n}$ tal que

$V = KK^t$. Para obtener el estimador de máxima verosimilitud de β es suficiente con transformar el modelo de la siguiente forma:

$$\left.\begin{cases} \xi = K^{-1}Y \\ M = K^{-1}X \\ \varepsilon = K^{-1}e \end{cases}\right\} \Rightarrow \xi = M\beta + \varepsilon$$

donde $\varepsilon \overset{d}{=} N_n(0, \sigma^2 I)$.

Es fácil comprobar que el estimador de máxima verosimilitud, $\hat{\beta} = \left(M^t M\right)^{-1} M^t \xi$, coincide con el estimador por mínimos cuadrados ponderados; es decir

$$\hat{\beta} = \underset{\beta}{argmin} \left(\xi - M\beta\right)^t \left(\xi - M\beta\right).$$

Además

$$\hat{\beta} = \left(X^t \underbrace{(K^{-1})^t K^{-1}}_{(KK^t)^{-1}=V^{-1}} X \right)^{-1} X^t \underbrace{(K^{-1})^t K^{-1}}_{(KK^t)^{-1}=V^{-1}} Y = \left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Y. \qquad \text{(B)}$$

Proposición 2.2. El método de puntuaciones de Fisher se puede expresar de la siguiente forma

$$\hat{\beta}^{(r)} = \left(X^t W(\hat{\beta}^{(r-1)})^{-1} X\right)^{-1} X^t W(\hat{\beta}^{(r-1)})^{-1} Z(\hat{\beta}^{(r-1)}) \qquad (C)$$

donde

$$W = diag\left(V[Y_i] g'(\mu_i)^2;\ i = 1,...,n\right)$$
$$Z = X\beta + diag\left(g'(\mu_1),...,g'(\mu_n)\right)\left(y - \mu\right)$$

Dada la similitud de (B) con (C), el método de puntuaciones de Fisher también recibe el nombre de "*algoritmo de mínimos cuadrados ponderados iterados".*

Demostración. Para llegar a la ecuación (C), comenzamos multiplicando la ecuación recurrente de Fisher por $I(\hat{\beta}^{(r-1)})$. Entonces

$$I(\hat{\beta}^{(r-1)})\hat{\beta}^{(r)} = I(\hat{\beta}^{(r-1)})\hat{\beta}^{(r-1)} + U(\hat{\beta}^{(r-1)}).$$

Ahora calculamos las expresiones de los términos $I(\beta)$. Se obtiene

$$I(\beta)_{jk} = -E\left[\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right] = -\sum_{i=1}^{n} E\left[\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right] = \sum_{i=1}^{n} E\left[\frac{\partial l_i}{\partial\beta_j}\frac{\partial l_i}{\partial\beta_k}\right]$$
$$= \sum_{i=1}^{n} E\left[\frac{(y_i-\mu_i)^2 x_{ij}x_{ik}}{V[Y_i]^2 g'(\mu_i)^2}\right] = \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{V[Y_i]g'(\mu_i)^2}$$

después de aplicar la proposición **2.1** en la penúltima igualdad.

En definitiva, se ha comprobado que

$$I(\beta) = X^t W(\beta)^{-1} X\ ,$$

donde $W = diag\left(V[Y_i]g'(\mu_i)^2;\ i=1,...,n\right)$.

Las ecuaciones de verosimilitud escritas en notación matricial son

$$U(\beta) = 0_{p\times 1}$$
$$X^t W(\beta)^{-1} Z^*(\beta) = 0_{p\times 1}$$

donde $Z^*(\beta)_{n\times 1} = \left((Y_i-\mu_i)g'(\mu_i);\ i=1,...,n\right)^t = diag\left(g'(\mu_i),...,g'(\mu_n)\right)(Y-\mu)$

Así:

$$I(\hat\beta^{(r-1)})\hat\beta^{(r)} = X^t W(\hat\beta^{(r-1)})^{-1} X\hat\beta^{(r)}$$

y

$$I(\hat\beta^{(r-1)})\hat\beta^{(r-1)} + U(\hat\beta^{(r-1)}) = X^t W(\hat\beta^{(r-1)})^{-1} X\hat\beta^{(r-1)} + X^t W(\hat\beta^{(r-1)})^{-1} Z^*(\hat\beta^{(r-1)})$$
$$= X^t W(\hat\beta^{(r-1)})^{-1}\left[X\hat\beta^{(r-1)} + diag\left(g'(\mu_1),...,g'(\mu_n)\right)(y-\mu)\right]$$
$$= X^t W(\hat\beta^{(r-1)})^{-1} Z(\hat\beta^{(r-1)})$$

Igualando nuevamente ambos lados, y operando, se obtiene

$$\hat\beta^{(r)} = \left(X^t W(\hat\beta^{(r-1)})^{-1} X\right)^{-1} X^t W(\hat\beta^{(r-1)})^{-1} Z(\hat\beta^{(r-1)})$$

que es lo que se quería demostrar.

1.2.5. Inferencia en los parámetros de un modelo lineal generalizado

Proposición 2.3. Bajo condiciones de regularidad se verifica:

$$a)\ U(\beta) \sim N_p(0, I(\beta))$$
$$b)\ U^t(\beta) I^{-1}(\beta) U(\beta) \sim \chi^2_p$$

donde " $\sim$ " se usa para denotar la igualdad en distribución asintótica.

Demostración. La demostración puede verse en "Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximun likelihood estimator in generalized linear models. Annals of Statistics 13,342-368".

Proposición 2.4. Se verifica que

$$a)\ \hat{\beta} \sim N_p(\beta, I^{-1}(\beta))$$
$$b)\ (\hat{\beta}-\beta)^t I(\beta)(\hat{\beta}-\beta) \sim \chi^2_p \quad (\textit{estadístico de Wald})$$

Demostración. a) Desarrollando en serie de Taylor el vector de puntuaciones $U(\beta)$ se tiene

$$U(\beta) \sim \underset{=0}{U(\hat{\beta})} + \underset{\xrightarrow{n\to\infty} -I(\beta)}{H(\hat{\beta})}(\beta-\hat{\beta}) \sim I(\beta)(\hat{\beta}-\beta) \Rightarrow (\hat{\beta}-\beta) \sim I^{-1}(\beta)U(\beta)$$

Por la proposición **2.3**,

$U(\beta) \sim N_p(0, I(\beta)) \Rightarrow I^{-1}(\beta)U(\beta)$, *tiene asintóticamente también una normal con parámetros*:

$$E\left[\hat{\beta}-\beta\right] \sim E\left[I^{-1}(\beta)U(\beta)\right] = I^{-1}(\beta)E\left[U(\beta)\right] = 0_{p\times 1} \quad \textit{por la proposición } 12.$$

Por tanto,

$$E\left[\hat{\beta}\right] = \beta$$

También

$$V\left[\hat{\beta}-\beta\right]\sim V\left[I^{-1}(\beta)U(\beta)\right]=I^{-1}(\beta)V\left[U(\beta)\right]\left(I^{-1}(\beta)\right)^{t}$$
$$=I^{-1}(\beta)I(\beta)I^{-1}(\beta)=I^{-1}(\beta)$$

ya que por ser I una matriz simétrica se tiene $(I^{-1})^{t}=(I^{t})^{-1}=I^{-1}$

Por tanto,

$$V\left[\hat{\beta}\right]\sim V\left[I^{-1}(\beta)U(\beta)\right]=I^{-1}(\beta)$$

Como $(\hat{\beta}-\beta)\sim I^{-1}(\beta)U(\beta)\Rightarrow\hat{\beta}\sim\beta+I^{-1}(\beta)U(\beta)\sim N_{p}(\beta,I^{-1}(\beta))$, queda por tanto a) demostrado.

La parte b) es inmediata, pues, es simplemente una aplicación de la parte b) de la proposición **2.3**.

1.2.6. Medición del ajuste

El proceso de ajustar un modelo a unos datos puede contemplarse como una forma de reemplazar un conjunto de datos $y=(y_1,...,y_n)$ por un conjunto de valores ajustados $\hat{\mu}=(\hat{\mu}_1,...,\hat{\mu}_n)$. El vector $\hat{\mu}$ es una función de un número reducido de parámetros estimados $\hat{\beta}$ (los parámetros del modelo).

En general $\hat{\mu}_i\neq y_i$ y la cuestión a responder es si la discrepancia es alta o baja. Normalmente, una discrepancia alta no es tolerable, mientras que una discrepancia baja sí lo es. Cuando ajustamos un modelo lineal generalizado, juzgamos su adecuación comparando la verosimilitud del modelo ajustado con la verosimilitud del modelo saturado.

El modelo saturado es un modelo de forma similar al propuesto que describe de modo "perfecto" los datos (es decir, tal que $\hat{\mu}_i=y_i\ \ \forall i$). Por tanto tiene muy poca utilidad desde el punto de vista de la explicación del comportamiento medio del fenómeno aleatorio modelado. El modelo saturado asigna toda la variabilidad a la componente sistemática, no dejando variabilidad para la componente aleatoria. El modelo saturado se utiliza para medir cómo un ajuste concreto se parece al ajuste "perfecto" que describe el 100% de la variabilidad de los datos.

En el otro extremo se encuentra el modelo nulo. Se trata de un modelo con un único parámetro (es decir, tal que $\hat{\mu}_i = \hat{\mu} \quad \forall i$) y que asigna toda la variabilidad a la componente aleatoria.

En la práctica el modelo nulo es demasiado simple y el modelo saturado no aporta información alguna, pues se limita a repetir los datos uno a uno. El arte de modelar consiste en seleccionar un modelo con pocos parámetros y con poca discrepancia respecto del modelo saturado.

El modelo saturado asociado a un modelo propuesto viene caracterizado por la utilización de:

- La misma distribución para la respuesta (no necesariamente con los mismos parámetros),
- El mismo número de parámetros que datos ($p = n$). Ello implica que no quedan grados de libertad para los residuos,
- El mismo nexo.

El modelo saturado se define especificando sus componentes; es decir,

1. Respuestas $y_1, ..., y_n$ independientes con función de densidad de la forma (A),
2. Parámetros $\beta = (\beta_1, ..., \beta_n) = \eta$ y matriz de diseño $X = I_{n \times n}$ es la identidad de $n \times n$,
3. Nexo $g(\cdot)$ (función monótona y diferenciable) tal que $\eta_i = \beta_i = g(\mu_i)$, con $\mu_i = E[Y_i]$.

En este tipo de modelos, al ser $X = I_{n \times n}$, el parámetro β carece de interés y en consecuencia conviene plantear la función de log-verosimilitud como función de μ ; es decir

$$l(\mu, \phi; y) = \sum_{i=1}^{n} l_i(\mu_i, \phi; y_i) \quad con \quad \mu = (\mu_1, ..., \mu_n) \quad e \quad y = (y_1, ..., y_n).$$

Para obtener el estimador máximo verosímil de $\mu_i = E[Y_i]$ en el modelo saturado,

$$\tilde{\mu}_i = \underset{\mu_i}{argmax}\, l_i(\mu_i, \phi; y_i),$$

basta con plantear las ecuaciones de verosimilitud

$$0 = \frac{\partial l_i}{\partial \mu_i} = \frac{\partial l_i}{\partial \omega_i} \frac{\partial \omega_i}{\partial \mu_i} = \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{d''(\omega_i)} \Leftrightarrow \tilde{\mu}_i = y_i$$

Sea $\hat{\beta}$ el **EMV** (Estimador de Máxima Verosimilitud) de β, $\hat{\mu} = \mu(\hat{\beta})$ el **EMV** de μ y $\hat{\omega} = \omega(\hat{\mu}) = \omega(\mu(\hat{\beta}))$ el **EMV** de ω en un modelo lineal generalizado arbitrario. Podemos escribir

(en función de μ) el máximo en β de la función de log-verosimilitud del modelo de la siguiente forma

$$l(\hat{\mu},\phi;y) = \max_{\beta} l(\mu(\beta),\phi;y)$$

Para su modelo saturado, sean $\tilde{\mu} = y$ y $\tilde{\omega} = \omega(y)$ los **EMV** de μ y ω respectivamente. En consecuencia $l(\tilde{\mu},\phi;y)$ es el máximo de su función de log-verosimilitud.

Definición 2.1. El "estadístico desviación" (del inglés "deviance") es

$$S(y,\hat{\mu},\phi) = 2\left[l(\tilde{\mu},\phi;y) - l(\hat{\mu},\phi;y)\right] = 2\sum_{i=1}^{n} \frac{y_i(\tilde{\omega}_i - \hat{\omega}_i) - \left(d(\tilde{\omega}_i) - d(\hat{\omega}_i)\right)}{a_i(\phi)}$$

Para deducir la distribución asintótica de $S(y,\hat{\mu},\phi)$ nos basamos en el cociente de la razón de verosimilitudes RV.

$$RV(y) = \frac{\sup_{w=w(\beta),\,\beta\in R^p \text{ bajo } H_0} L(\omega,\phi,y)}{\sup_{w=w(\beta),\,\beta\in R^n \text{ bajo } H_1 (m.\,saturado)} L(\omega,\phi,y)}$$

donde

$$H_0 : g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \ , \qquad i = 1,\ldots,n$$

Si tomamos los estimados de máxima verosimilitud $\hat{\omega} = \omega(\hat{\beta})$ para el modelo propuesto bajo H_0 y $\tilde{\omega} = \omega(\tilde{\beta})$ para el modelo saturado bajo H_1, se tiene:

$$RV(y) = \frac{L(\hat{\omega},\phi,y)}{L(\tilde{\omega},\phi,y)}$$

Si el tamaño de la muestra es grande y H_0 es cierta, se sabe que, $-2lnRV(y)$ se distribuye asintóticamente como una χ^2 con un número de grados de libertad igual a la diferencia de dimensión entre los espacios $\mathbb{R}^n$ y $\mathbb{R}^p$ para este caso de modelos lineales generalizados. Esto es,

$$-2lnRV(y) \sim \chi^2_{n-p}$$

Ahora, sólo es de observar que:

$$\begin{aligned} -2lnRV(y) &= -2\left(\ln L(\hat{\omega},\phi,y) - \ln L(\tilde{\omega},\phi,y)\right) \\ &= 2\left(\ln L(\tilde{\omega},\phi,y) - \ln L(\hat{\omega},\phi,y)\right) \\ &= 2\left(l(\tilde{\omega},\phi,y) - l(\hat{\omega},\phi,y)\right) \\ &= S(y,\hat{\omega},\phi) \end{aligned}$$

Se concluye así que $S(y,\hat{\mu},\phi) \sim \chi^2_{n-p}$ y sirve para contrastar la hipótesis nula H_0.

Definición 2.2. El "estadístico de las puntuaciones de Rao" para contrastar la idoneidad del modelo lineal generalizado, y más concretamente para contrastar

$$H_0 : g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \ , \qquad i = 1,...,n$$

ó en notación matricial

$$H_0 : g(\mu) = X\beta$$

es

$$R_n(\hat{\mu},\phi) = U^t(\hat{\mu})I^{-1}(\hat{\mu})U(\hat{\mu}) \sim \chi^2_{n-p}$$

donde

$\hat{\mu} = \mu(\hat{\beta}) = (\hat{\mu}_1,...,\hat{\mu}_n)^t$ es el EMV de μ restringido a H_0.

$U(\mu) = \left(U_1(\mu),...,U_n(\mu)\right)^t$, *con* $U_i(\mu) = \dfrac{\partial}{\partial\mu_i} l(\mu_1,...,\mu_n;y_1,...,y_n)$ es el vector de puntuaciones del modelo saturado.

$I(\mu) = -\left(E\left[\dfrac{\partial}{\partial\mu_j}U_i(\mu)\right]\right)_{n\times n}$ es la matriz de información de Fisher del modelo saturado.

Definición 2.3. El "estadístico de Mahalanobis" para contrastar

$$H_0 : g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \ , \qquad i = 1,...,n$$

es

$$M_n(y,\hat{\mu},\phi) = (y-\hat{\mu})^t I(\hat{\mu})(y-\hat{\mu}) \sim \chi^2_{n-p}$$

Donde $\hat{\mu} = \mu(\hat{\beta})$ e $I(\hat{\mu})$ es la matriz de información de Fisher del modelo saturado.

1.2.7. El análisis de residuos

En este apartado introducimos distintos tipos de residuos (y estadísticos en general) que se utilizan en modelos lineales generalizados para analizar la idoneidad de las hipótesis efectuadas.

En el caso de modelos lineales generalizados se consideran los siguientes residuos:

1. Los residuales de Pearson.
2. Los residuales de Ascombe.
3. Los residuales desviación.
4. Los residuales verosimilitud

El *residual de Pearson* se define

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}[Y_i]}}.$$

El residual de Pearson estandarizado es

$$r_i^{PS} = \frac{y_i - \hat{\mu}_i}{\sqrt{1-h_{ii}}}$$

donde h_{ii} es el elemento diagonal de la fila i-ésima y columna i-ésima de la matriz sombrero

$H = W^{-1/2} X (X^t W^{-1} X)^{-1} X^t W^{-1/2}$ en la que

$$W = diag(V[Y_1] g'(\mu_1)^2, ..., V[Y_n] g'(\mu_n)^2)$$

que depende de los parámetros. Por tanto se trabaja con su estimación máximo-verosímil $\hat{W}$ y X es matriz de diseño de $n \times p$ cuyas columnas son los valores de las variables explicativas $X_1, ..., X_p$.

Una desventaja de los residuales de Pearson es que su distribución, para respuestas no normales, es a menudo marcadamente asimétrica. Por tanto, sus propiedades pueden diferir de las de los residuos del modelo lineal normal. Ascombe propuso definir otro tipo de residuos cuya distribución fuera más próxima a la normal en modelos en los que la varianza de la respuesta sea función de la media; es decir, donde

$$V[Y_i] = v(\mu_i), \quad i = 1, \ldots, n.$$

Para ello propuso modificar los residuos de Pearson Cambiando y_i y $\hat{\mu}_i$ por $A(y_i)$ y $A(\hat{\mu}_i)$ respectivamente. Los *residuos de Ascombe* son

$$r_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{\sqrt{\hat{V}[A(y_i)]}}, \ i = 1, \ldots, n$$

Wedderburn demostró que la función

$$A(y) = \int_{-\infty}^{y} \frac{d\mu}{v^{1/3}(\mu)}$$

es una buena candidata para conseguir que los r_i^A sean aproximadamente normales.

Por Taylor, observemos que $A(y_i) = A(\mu_i) + A'(\mu_i)(y_i - \mu_i) + o(1)$. Entonces,

$$V[A(y_i)] \approx [A'(\mu_i)]^2 V[Y_i] = [A'(\mu_i)]^2 v(\mu_i)$$
$$\hat{V}[A(y_i)] = [A'(\hat{\mu}_i)]^2 v(\hat{\mu}_i)$$

Por tanto, los residuos de Ascombe se suelen escribir de la siguiente forma

$$r_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{v(\hat{\mu}_i)}}, \ i = 1, \ldots, n$$

Además en el caso $A(y) = \int_{-\infty}^{y} \frac{d\mu}{v^{1/3}(\mu)}$, se obtiene

$$A'(y) = v^{-1/3}(y) \quad \text{y} \quad A'(\hat{\mu}_i)\sqrt{v(\hat{\mu}_i)} = v^{-1/3}(\hat{\mu}_i)v^{1/2}(\hat{\mu}_i) = v^{1/6}(\hat{\mu}_i).$$

Por tanto,

$$r_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{v^{1/6}(\hat{\mu}_i)}, \ i = 1, \ldots, n$$

Ejemplo 2.2. Distribución de Poisson. Se tiene que

$$V[Y_i] = \mu_i = v(\mu_i)$$

$$A(y) = \int_0^y v^{-1/3}(\mu)d\mu = \int_0^y \mu^{-1/3}d\mu = \left[\frac{3}{2}\mu^{2/3}\right]_0^y = \frac{3}{2}y^{2/3}$$

Por tanto

$$r_i^A = \frac{\frac{3}{2}\left(y_i^{2/3} - \hat{\mu}_i^{2/3}\right)}{\hat{\mu}_i^{1/6}},\ i = 1,\ldots,n$$

Cuando $a_i(\phi) = a_i\phi$, el estadístico desviación $S(y,\hat{\mu},\phi)$ se puede escribir de la siguiente forma

$$S(y,\hat{\mu},\phi) = \frac{D(y,\hat{\mu})}{\phi}$$

donde

$$D(y,\hat{\mu}) = 2\sum_{i=1}^{n}\frac{y_i(\tilde{\omega}_i - \hat{\omega}_i) - \left(d(\tilde{\omega}_i) - d(\hat{\omega}_i)\right)}{a_i}$$

Es conocida como la *desviación no escalada*.

Si el estadístico desviación no escalada se utiliza como una medida de discrepancia en un modelo lineal generalizado, entonces cada unidad contribuye una cantidad D_i a dicha medida, de forma que:

$$D = \sum_{i=1}^{n} D_i = \sum_{i=1}^{n}\frac{2}{a_i}\left[y_i(\tilde{\omega}_i - \hat{\omega}_i) - \left(d(\tilde{\omega}_i) - d(\hat{\omega}_i)\right)\right]$$

El *residuo desviación* es

$$r_i^D = signo(y_i - \hat{\mu}_i)\sqrt{D_i}\ ,\ i = 1,\ldots,n.$$

y verifica que $D = \sum_{i=1}^{n}(r_i^D)^2$.

El *residuo desviación estandarizado* es

$$r_i^{DS} = \frac{r_i^D}{\sqrt{1-h_{ii}}},\ i = 1,\ldots,n$$

y hace resaltar cualquier observación que "contribuye demasiado", a la desviación del modelo.

El *residuo verosimilitud* es

$$r_i^L = signo(y_i - \hat{\mu}_i)\sqrt{h_{ii}(r_i^{PS})^2 + (1-h_{ii})(r_i^{PS})^2}$$

La utilidad de este residuo reside en el hecho de que $(r_i^L)^2$ es aproximadamente igual al cambio en la desviación escalada, S, que resulta cuando eliminamos la i-ésima observación en el ajuste. Así la sensibilidad del ajuste de un modelo a la eliminación de la i-ésima observación se puede describir considerando el tamaño de r_i^L.

2. Aplicación del análisis discriminante y la regresión logística en clasificar el cangrejo herradura con o sin individuos satélites. Comparación de los resultados

2.1. Introducción de la aplicación con el análisis discriminante

EL presente capítulo tendrá los objetivos siguientes:

Objetivo general

Clasificar si el cangrejo herradura hembra tiene o no tiene individuos satélites utilizando el Análisis Discriminante y la Regresión Logística con la ayuda del SPSS.

Objetivo específico

Determinar la efectividad de ambos métodos Análisis Discriminante y la Regresión Logística, comparando los resultados generados por el SPSS, el número de casos en que coinciden ambos métodos y el número de clasificaciones correctas que hacen ambos métodos.

Comenzamos mostrando una imagen del cangrejo herradura cuyo nombre científico es *Limulus polyphemu* el cual es nuestro individuo de estudio:

El Cangrejo hembra de esta especie tiene su propio macho en su hábitat, adicionalmente puede o no tener otros machos al rededor a los que llamamos “individuos satélites”.

El objetivo de esta aplicación es estudiar el poder que tienen ciertas características del cangrejo herradura hembra para clasificarlo con presencia o sin presencia de individuos satélites, las características observadas en dicho cangrejo son:

1. La anchura en centímetros.
2. El peso en kilogramos
3. El color: 1= medio claro; 2 = claro; 3 = medio oscuro; 4 = oscuro.
4. Condición de las espinas: 1 = buenas; 2 = unas rotas; 3 = todas rotas

Estas características fueron tomadas de 173 cangrejos herradura hembra, información disponible en el libro *Categorical Data Analysis*.

El factor discriminante tiene dos atributos “Hay individuos satélites” y “No hay individuos satélites”, el cual genera los dos grupos siguientes:

GRUPOS	Frecuencia	Porcentaje
G1: No hay individuos satélites	62	35.8
G2: Hay individuos satélites	111	64.2
Total	173	100

2.1.1. Clasificación del cangrejo herradura con o sin individuos satélites usando el análisis discriminante

Descriptivos del grupo 1 (G1) : No hay individuos satélites

La siguiente tabla presenta estadísticos del número de cangrejos según el tipo de espina y el tipo de color que posee el cangrejo de herradura hembra para G1.

Espina * color Crosstabulation

Count

		color				
		1= Medio claro	2= Claro	3= Medio oscuro	4= Oscuro	Total
Espina	1 = Espinas buenas	2	8	0	1	11
	2 = Unas espinas rotas	0	5	2	1	8
	3 = Todas las espinas rotas	1	13	16	13	43
Total		3	26	18	15	62

Como se aprecia, el cangrejo hembra con más frecuencia en el G1, es el que tiene todas las espinas rotas y colores claro, medio oscuro y oscuro con frecuencias de 13, 16 y 13 respectivamente.

La siguiente tabla presenta descriptivos según las variables cuantitativas peso y anchura para el grupo G1.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewnes	Kurtosis
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
anchura	62	21.0	28.7	25.169	1.6741	.000	-.319
peso	62	1.20	3.20	2.1391	.44839	.244	-.055
Valid N (listwise)	62						

Hay medias diferentes y la simetría y curtosis es aceptable para la normalidad en las variables anchura y peso.

Descriptivos del grupo 2 (G2) : Hay individuos satélites

La siguiente tabla presenta estadísticos según el tipo de espina y el tipo de color que posee el cangrejo de herradura hembra para G2.

Espina * color Crosstabulation

Count

		color				
		1= Medio claro	2= Claro	3= Medio oscuro	4= Oscuro	Total
Espina	1 = Espinas buenas	7	16	3	0	26
	2 = Unas espinas rotas	2	3	2	0	7
	3 = Todas las espinas rotas	0	50	21	7	78
Total		9	69	26	7	111

Como se aprecia, el cangrejo hembra con más frecuencia en el G2, es el que tiene todas las espinas rotas y colores claro y medio oscuro con una frecuencias de 50 y 21 cangrejos respectivamente.

La siguiente tabla presenta descriptivos según las variables cuantitativas peso y anchura para el grupo G2.

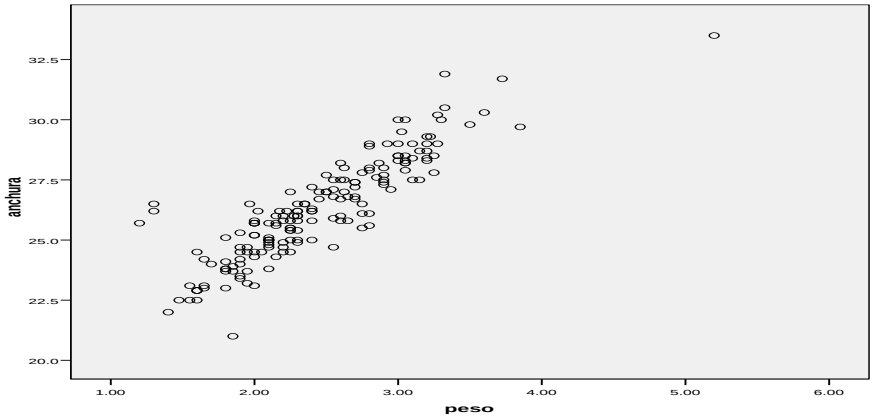
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewnes	Kurtosis
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
anchura	111	22.5	33.5	26.930	2.0689	.282	.186
peso	111	1.48	5.20	2.6037	.57539	.759	2.491
Valid N (listwise)	111						

La simetría y curtosis de normalidad es aceptable para la variable anchura, no así para el peso.

Aunque tenemos identificadas cuatro variables como independientes, que son la **anchura**, **el peso**, **el color** y **condición de las espinas**, vamos a evaluar la importancia de estas en el modelo:

Primero, no debe de haber multicolinealidad entre las variables incluidas en el modelo. Entre las variables cuantitativas **anchura** y **peso** presentan multicolinealidad, lo que se puede apreciar en el siguiente diagrama de dispersión acompañado de su coeficiente de correlación:



Correlations

		anchura	peso
anchura	Pearson Correlation	1	.887**
	Sig. (2-tailed)		.000
	N	173	173
peso	Pearson Correlation	.887**	1
	Sig. (2-tailed)	.000	
	N	173	173

**. Correlation is significant at the 0.01 level

Como se ve en el diagrama de dispersión y en el coeficiente de correlación 0.887 y una significancia de 0.000, están altamente correlacionadas linealmente, indicando que el peso y la anchura no son independientes. Para evitar la multicolinealidad una de ellas debe salir del modelo, la que tenga menor capacidad de clasificación correctamente. Dejaremos que el SPSS con los métodos de selección por etapas y hacia atrás eliminen la variable adecuada.

Estudiaremos también la influencia que pueden tener las variables categóricas sobre la variable factor discriminante. Las hipótesis que quereos probar son:

1. H_0 : El factor discriminante es independiente de la condición de las espinas.
2. H_0 : El factor discriminante es independiente del color.

Los estadísticos proporcionados por el SPSS son:
Para la primera hipótesis son:

factor * Espina Crosstabulation

			Espina			
			1 = Espinas buenas	2 = Unas espinas rotas	3 = Todas las espinas rotas	Total
factor	1 = No hay individuos satélites	Count	11	8	43	62
		Expected Count	13.3	5.4	43.4	62.0
	2 = Hay individuos satélites	Count	26	7	78	111
		Expected Count	23.7	9.6	77.6	111.0
Total		Count	37	15	121	173
		Expected Count	37.0	15.0	121.0	173.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.602	2	.272
Likelihood Ratio	2.526	2	.283
Linear-by-Linear Association	.133	1	.716
N of Valid Cases	173		

En estos resultados se observa un valor de la chi-cuadrado igual a 2.602 con una significancia alta de 0.272, llevando al no rechazo de la hipótesis 1, es decir, es razonable una independencia entre el factor discriminante y las condiciones de las espinas. Por esta razón no son importantes en las funciones discriminantes y no se utilizaran para el modelo de clasificación.

Para la segunda hipótesis son:

factor * color Crosstabulation

			color				
			1=Claro	2=Medio claro	3= Medio oscuro	4= Oscuro	Total
factor	1 = No hay individuos satélites	Count	3	26	18	15	62
		Expected Count	4.3	34.0	15.8	7.9	62.0
	2 = Hay individuos satélites	Count	9	69	26	7	111
		Expected Count	7.7	61.0	28.2	14.1	111.0
Total		Count	12	95	44	22	173
		Expected Count	12.0	95.0	44.0	22.0	173.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	14.078[a]	3	.003
Likelihood Ratio	13.698	3	.003
Linear-by-Linear Association	12.334	1	.000
N of Valid Cases	173		

a. 1 cells (12.5%) have expected count less than 5. The minimum expected count is 4.30.

En este caso, se observa un valor de la chi-cuadrado igual a 14.078 con una significancia baja de 0.003, llevando al rechazo de la hipótesis 2, es decir, que el color incide en el factor de discriminación. Por tal razón, utilizaremos la variable color en el modelo y dejaremos que el SPSS seleccione el color(es) más importantes en el modelo.

El análisis anterior sugiere las siguientes variables independientes en el modelo: la anchura, el peso y la variable cualitativa color dumitizada de la siguiente manera:

cj= 1 si se observa el color j (j=1,2,3,4) en el cangrejo hembra i-ésimo
(i=1,2,3,…,n=173) de la muestra.
= 0 si no es el color j.

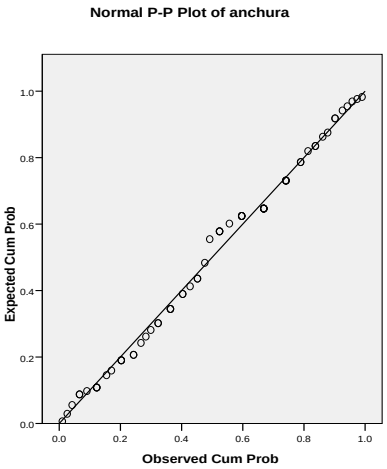
Consideraciones sobre el tamaño muestral total

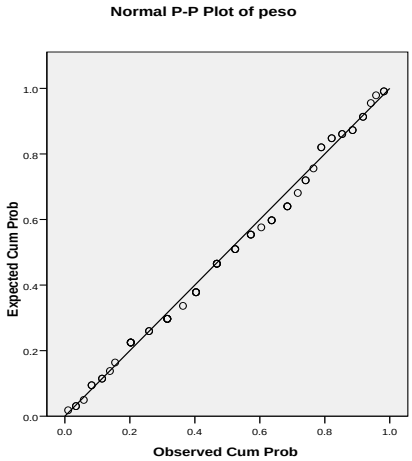
Los estudios sugieren como mínimo 20 observaciones por cada variable independiente. En nuestro investigación hay un total de 6 variables independientes 2 cuantitativas y 4 variables dumy que representan el color. Así, el tamaño muestral sugerido debería ser de 20x6 = 120 y nuestra muestra es de 173 elementos, indicando un buen tamaño relativamente a lo sugerido.

Chequeo de los supuestos del análisis discriminante

Normalidad: Debemos ver que las variables anchura y peso son normales separadamente en el G1 y en el G2.

Para el grupo G1 se tiene:





One-Sample Kolmogorov-Smirnov Test

		anchura
N		62
Normal Parameters[a,b]	Mean	25.169
	Std. Deviation	1.6741
Most Extreme Differences	Absolute	.078
	Positive	.063
	Negative	-.078
Kolmogorov-Smirnov Z		.616
Asymp. Sig. (2-tailed)		.842

a. Test distribution is Normal.

b. Calculated from data.

One-Sample Kolmogorov-Smirnov Test

		peso
N		62
Normal Parameters[a,b]	Mean	2.1391
	Std. Deviation	.44839
Most Extreme Differences	Absolute	.070
	Positive	.070
	Negative	-.063
Kolmogorov-Smirnov Z		.548
Asymp. Sig. (2-tailed)		.925

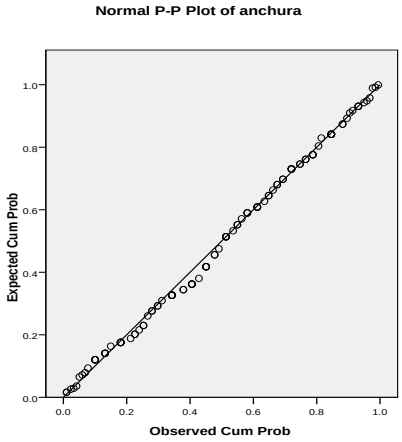
a. Test distribution is Normal.

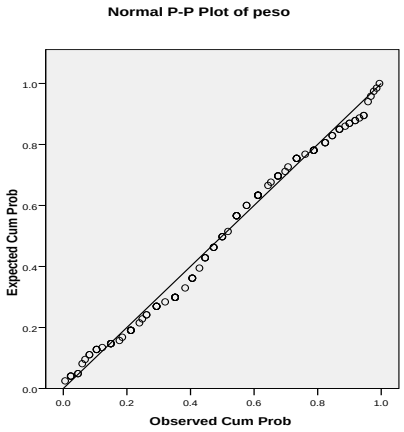
b. Calculated from data.

Los gráficos y las pruebas no paramétrica de normalidad, señalan que no pueden rechazarse las hipótesis de que la anchura y el peso tenga normalidad para G1 y por tanto, la normalidad es razonable.

La prueba de normalidad conjunta aquí no es necesario ya que solo hay dos variables cuantitativas y por la dependencia entre ellas una debe salir del análisis, así, solo es justificable la normalidad individual de las dos variables peso y anchura.

Para el grupo G2 se tiene:





One-Sample Kolmogorov-Smirnov Test

		anchura
N		111
Normal Parameters[a,b]	Mean	26.930
	Std. Deviation	2.0689
Most Extreme Differences	Absolute	.061
	Positive	.061
	Negative	-.039
Kolmogorov-Smirnov Z		.646
Asymp. Sig. (2-tailed)		.799

a. Test distribution is Normal.

b. Calculated from data.

One-Sample Kolmogorov-Smirnov Test

		peso
N		111
Normal Parameters[a,b]	Mean	2.6037
	Std. Deviation	.57539
Most Extreme Differences	Absolute	.080
	Positive	.080
	Negative	-.048
Kolmogorov-Smirnov Z		.838
Asymp. Sig. (2-tailed)		.483

a. Test distribution is Normal.

b. Calculated from data.

Los gráficos y las pruebas no paramétrica de normalidad, señalan que no pueden rechazarse las hipótesis de que la anchura y el peso tenga normalidad para G2 y por tanto, la normalidad es razonable en el G2.

Homocedasticidad y discriminación del factor: Aquí revisaremos la igualdad de las varianzas de la anchura y del peso separadamente ya que una de las dos deberá eliminarse en los G1 y G2. También, revisaremos que haya poder de discriminación en el factor, es decir, que las medias poblacionales en las variables anchura y peso en los grupos G1 y G2 sean distintas.

Estadísticos por grupo para la variable anchura:

Group Statistics

	factor	N	Mean	Std. Deviation	Std. Error Mean
anchura	1 = No hay individuos satélites	62	25.169	1.6741	.2126
	2 = Hay individuos satélites	111	26.930	2.0689	.1964

Independent Samples Test

		Levene's Test for Equality of Variances					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
anchura	Equal variances assumed	2.820	.095	-5.731	171	.000	-1.7604
	Equal variances not assumed			-6.082	149.2	.000	-1.7604

En los resultados mostrados en estas dos tablas para la variable anchura, la prueba de Levene´s señala para la igualdad de varianzas una significancia de 0.095 que al 5% no puede rechazarse que las varianzas en los grupos G1 y G2 sean iguales y al mismo tiempo señala que para la igualdad de las medias en los grupos G1 y G2 tienen una significancia de 0.000, la cual es muy baja llevando al rechazo de que las medias sean iguales. Los resultados de esta variable son muy positivos ya que podemos asumir medias diferentes y varianzas iguales en la variable anchura en los grupos G1 y G2.

Estadísticos por grupo para la variable peso:

Group Statistics

	factor	N	Mean	Std. Deviation	Std. Error Mean
peso	1 = No hay individuos satélites	62	2.1391	.44839	.05695
	2 = Hay individuos satélites	111	2.6037	.57539	.05461

Independent Samples Test

		Levene's Test for Equality of Variances					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
peso	Equal variances assumed	4.656	.032	-5.492	171	.000	-.46457
	Equal variances not assumed			-5.888	153.03	.000	-.46457

En los resultados mostrados en estas dos últimas tablas para la variable peso, la prueba de Levene´s señala para la igualdad de varianzas una significancia de 0.032 que al 5% se rechaza que las varianzas en los grupos G1 y G2 sean iguales y al mismo tiempo señala que para la igualdad de las medias en los grupos G1 y G2, asumiendo varianzas no iguales, tienen una significancia de 0.000, la cual es muy baja llevando al rechazo de que las medias sean iguales. Los resultados de esta variable no son muy positivos ya que se viola el supuesto de que las varianzas sean iguales en la variable peso en los grupos G1 y G2.

Recordemos que hay una de las variables anchura y peso que debe de salir del análisis discriminante por haber una dependencia lineal entre ellas y esperemos que sea la variable peso la que tenga que salir por violar uno de los supuestos del análisis discriminante.

Resultados del análisis discriminante usando SPSS

Usando la estimación por etapas[14] se tiene:

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	peso	1.000	1.000	30.157	.850
	anchura	1.000	1.000	32.845	.839
	Medio claro	1.000	1.000	.654	.996
	Claro	1.000	1.000	6.754	.962
	Medio oscuro	1.000	1.000	.655	.996
	Oscuro	1.000	1.000	12.139	.934
1	peso	.250	.250	.929	.834
	Medio claro	.995	.995	.148	.838
	Claro	.978	.978	2.585	.826
	Medio oscuro	.981	.981	.000	.839
	Oscuro	.992	.992	7.359	.804
2	peso	.250	.249	.839	.800
	Medio claro	.988	.984	.022	.804
	Claro	.834	.834	.359	.802
	Medio oscuro	.912	.912	.538	.802

En el paso cero no hay variable en el modelo discriminante. En el paso uno entra la variable anchura. En el paso dos entra el color "oscuro" o sea la variable dumy "c4". Esto se ve en la siguiente tabla

[14] ***El modelo comienza sin ninguna variable y se van introdociedo sólo si son importantes en la discriminación y pueden salir en caso de perder iportancia cuando entra una nueva. Puede revisarse el libro: Pérez López (2005)***

Variables Entered/Removed[a,b,c]

Step	Entered	Wilks' Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	anchura	.839	1	1	171	32.845	1	171	.000
2	Oscuro	.804	2	1	171	20.713	2	170	.000

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

a. Maximum number of steps is 12.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

Las variables importantes en el modelo sólo son la anchura y c4 el color oscuro. Afortunadamente la variable peso que violaba uno de los supuestos del análisis discriminante y tenía dependencia lineal con la anchura, esta, ha sido sacada del análisis de clasificación tal como se esperaba que saliera una de las dos.

Las funciones de clasificación lineales de Fisher´s se muestran en la siguiente tabla:

Classification Function Coefficients

	factor	
	1 = No hay individuos satélites	2 = Hay individuos satélites
anchura	6.799	7.245
Oscuro	6.042	4.581
(Constant)	-87.317	-98.146

Fisher's linear discriminant functions

Para G1, se tiene:

Z_1= -87.317+6.799(anchura)+6.042c4

Para el grupo G2, se tiene:

Z_2= -98.146+7.245(anchura)+4.581c4

Clasificamos al individuo "cangrejo herradura hembra" con cierta anchura y c4=1 si es color oscuro y si otro color c4=0, en el grupo G_i en el que Z_i sea el más grande para i= 1,2.

Ejemplo: Para individuo 1, anchura=28.3 y c4=0 y Z_1=105.0947 y Z_2=106.8875. Por tanto, el primer caso va al grupo G2.

La clasificación de los primeros 15 individuos se muestra en la siguiente tabla:

Casewise Statistics

	Case Number	Actual Group	Predicted Group
Original	1	2	2
	2	1	1
	3	2	2
	4	1	2
	5	2	2
	6	1	1
	7	1	2
	8	1	2
	9	1	1
	10	1	2
	11	1	2
	12	1	2
	13	2	2
	14	1	1
	15	2	2

En estos primeros 15 se han clasificado mal los casos 4, 7, 8, 10, 11 y 12.

Un resumen global de la clasificación así obtenida resulta en la siguiente:

Classification Results[b,c]

			Predicted Group Membership		
		factor	1 = No hay individuos satélites	2 = Hay individuos satélites	Total
Original	Count	1 = No hay individuos satélites	29	33	62
		2 = Hay individuos satélites	14	97	111
	%	1 = No hay individuos satélites	46.8	53.2	100.0
		2 = Hay individuos satélites	12.6	87.4	100.0
Cross-validated[a]	Count	1 = No hay individuos satélites	27	35	62
		2 = Hay individuos satélites	14	97	111
	%	1 = No hay individuos satélites	43.5	56.5	100.0
		2 = Hay individuos satélites	12.6	87.4	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. 72.8% of original grouped cases correctly classified.

c. 71.7% of cross-validated grouped cases correctly classified.

Como vemos este modelo discriminante ha clasificado correctamente el 72.8% a nivel global de los 173 individuos, habiendo clasificado el 46.8% correctamente de 62 individuos en el grupo G1 y el 87.4% correctamente de 111 individuos en el grupo G2.

La clasificación con la validación cruzada(clasificación del individuo cuando no se toma en cuenta en la estimación del modelo) es idéntica en el grupo G2 con el 87.4% en ambas clasificaciones; no así en el grupo G1, la clasificación cruzada es un poco menos mejor que la clasificación global en un 3.3%. del porcentaje bien clasificado por la clasificación global.

Características del modelo:

La probabilidad de clasificar un cangrejo herradura hembra correctamente es:

$$P = P(ser\ del\ G1\ \ y\ \ Clasificado\ en\ G1) + P(ser\ del\ G2\ \ y\ \ Clasificado\ en\ G2)$$
$$= \frac{29}{173} + \frac{97}{173} = \frac{126}{173} = 0.7283$$

La probabilidad de error o probabilidad de clasificar un cangrejo herradura hembra incorrectamente es:

$$P = P(ser\ del\ G1\ \ y\ \ Clasificado\ en\ G2) + P(ser\ del\ G2\ \ y\ \ Clasificado\ en\ G1)$$
$$= \frac{33}{173} + \frac{14}{173} = \frac{47}{173} = 0.2717$$

La probabilidad de clasificar un cangrejo herradura hembra en el grupo G1 dado que es del grupo G1, es:

$$P = P\left(Clasificado\ \ en\ G1 \Big/ es\ \ de\ \ G1\right) = \frac{P(ser\ del\ G1\ \ y\ \ Clasificado\ en\ G1)}{P(ser\ \ del\ \ G1)} = \frac{\frac{29}{173}}{\frac{62}{173}} = \frac{29}{62} = 0.4677$$

La probabilidad de clasificar un cangrejo herradura hembra en el grupo G2 dado que es del grupo G2, es:

$$P = P\left(Clasificado\ \ en\ G2 \Big/ es\ \ de\ \ G2\right) = \frac{P(ser\ del\ G2\ \ y\ \ Clasificado\ en\ G2)}{P(ser\ \ del\ \ G2)} = \frac{\frac{97}{173}}{\frac{111}{173}} = \frac{97}{111} = 0.8739$$

Estas dos últimas característica, señalan que el modelo clasifica mejor en el grupo G2.

Otra medida de la capacidad predictiva por el modelo discriminante propuesto

El estadístico Q de Press es

$$Q\ \ de\ \ \ Press = \frac{\left[N - (nK)\right]^2}{N(K-1)}$$, tiene asintóticamente una χ_1^2.

Donde

N= Tamaño muestral total

n = número de observaciones correctamente clasificadas

K= Número de grupos

$$Q\ \ de\ \ \ Press = \frac{\left[173 - (126(2))\right]^2}{173(1)} = 36.08$$

Al 5% , el cuantil $1-\alpha$ de la chi-cuadrado con un grado de libertad es

$$\chi^2_{1,1-\alpha} = \chi^2_{1,1-0.05} = \chi^2_{1,0.95} = 3.841$$

Como el *Q de Press* es mayor que dicho cuantil, esta clasificación es altamente significativa buena, es decir no es pobre.

2.2. Introducción de la aplicación con el análisis de regresión logística.

A través del SPSS y usando las variables peso, anchura, c1, c2, c3 y c4 como variables independientes en la regresión logística y 1 para identificar a los del grupo G2 y, 0 para identificar los del grupo G1; se tiene:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	anchura	,497	,102	23,887	1	,000	1,644
	Constant	-12,351	2,629	22,075	1	,000	,000
Step 2[b]	anchura	,478	,104	21,084	1	,000	1,613
	c4	-1,301	,526	6,116	1	,013	,272
	Constant	-11,679	2,693	18,814	1	,000	,000

a. Variable(s) entered on step 1: anchura.

b. Variable(s) entered on step 2: c4.

Usando un método de selección de variables hacia adelante, se termina en el segundo paso, quedando las variables anchura y c4. Los parámetros son significativamente distintos de cero al 5%.

El modelo es:

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -11.679 + 0.478(anchura(i)) - 1.301(c4)$$

despejando $\hat{\pi}_i$ se tiene:

$$\hat{\pi}_i = \frac{1}{1+e^{11.679-0.478(anchura(i))+1.301(c4)}}$$

Si $\hat{\pi}_i \geq 0.5$ para el individuo i, entonces este se clasifica en el grupo G2, en caso contrario en el grupo G1.

La validación de este modelo la analizamos con la siguiente tabla:

Goodness of Fit[b]

	Value	df	Value/df
Deviance	90.841	76	1.195
Scaled Deviance	90.841	76	
Pearson Chi-Square	77.431	76	1.019
Scaled Pearson Chi-Square	77.431	76	
Log Likelihood[a]	-93.979		
Akaike's Information Criterion (AIC)	193.958		
Finite Sample Corrected AIC (AICC)	194.278		
Bayesian Information Criterion (BIC)	201.066		
Consistent AIC (CAIC)	204.066		

Dependent Variable: factordumy
Model: (Intercept), anchura, c4

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

La hipótesis que contrastamos con esta tabla es:

$$H_0: \ \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(anchura(i)) + \beta_2(c4)$$

El estadístico de prueba es el estadístico desviación:

$$S(y,\hat{\mu},\phi) = 2\left[l(\tilde{\mu},\phi;y) - l(\hat{\mu},\phi;y)\right] = 2\sum_{i=1}^{n}\frac{y_i(\tilde{\omega}_i - \hat{\omega}_i) - \left(d(\tilde{\omega}_i) - d(\hat{\omega}_i)\right)}{a_i(\phi)}$$
$$= Deviance = 90.841$$

Que tiene una chi-cuadrada con 76 grados de libertad χ^2_{76} (De los 173 casos sólo hay 79 clases covariantes distintas y tres parámetros estimados) y $\chi^2_{76,0.95} = 97.3$. Por tanto no puede rechazarse H_0 y el modelo estimado es aceptable para estimar la probabilidad de pertenencia al grupo G2.

Lo mismo pasa con la suma de los cuadrados de los errores de Pearson que es 77.431< $\chi^2_{76,0.95} = 97.35$, llevando al no rechazo de H_0.

Análisis de la sobredispersión

Queremos que no haya sobredispersión $(\phi = 1)$ para lo cual debemos probar $H_0: \phi = 1$, sabemos que $\hat{\phi} = \frac{S(y,\hat{\mu},\phi)}{n-p} = \frac{Deviance}{n-p} = \frac{90.841}{76} = 1.195$ (dado en la tabla salida de SPSS) y un punto de corte puede ser $\frac{\chi^2_{76,0.95}}{76} = 1.281$ y no puede rechazarse H_0, la sobre dispersión que hay es tolerable.

2.2.1. Clasificación del cangrejo herradura con o sin individuos satélites usando el análisis de regresión logística.

La siguiente matriz proporciona un resumen de clasificación para las etapas de convergencia 1 y la etapa final 2.

Classification Table[a]

Observed			Predicted		
			factor		
			1 = No hay individuos satélites	2 = Hay individuos satélites	Percentage Correct
Step 1	factor	1 = No hay individuos satélites	27	35	43,5
		2 = Hay individuos satélites	16	95	85,6
	Overall Percentage				70,5
Step 2	factor	1 = No hay individuos satélites	29	33	46,8
		2 = Hay individuos satélites	14	97	87,4
	Overall Percentage				72,8

a. The cut value is ,500

Se clasifican 126 individuos correctamente o un 72.83% de los casos

2.3. Comparación de los resultados producidos de la clasificación del cangrejo herradura por el análisis discriminante y regresión logística.

Ambos métodos clasifican correctamente la misma cantidad de 126 del total 173 casos que es el 72.83%.

Para ver realmente si hay diferencia en los casos clasificados por ambos métodos hacemos un cruce de la clasificación discriminante y de la clasificación con la regresión logística, esta es:

Predicted Group for Analysis 1 * Predicted group Crosstabulation

Count

		Clasificación logística		
		1 = No hay individuos satélites	2 = Hay individuos satélites	Total
Clasificación discriminante	1 = No hay individuos satélites	41	2	43
	2 = Hay individuos satélites	2	128	130
Total		43	130	173

En esta tabla se ve que ambos métodos coinciden en clasificar 41 casos en G1 de los cuales, sólo 29 coincidieron en clasificar correctamente . También coinciden en clasificar 128 en G2 de los cuales, sólo 97 coincidieron en clasificar correctamente.

Los 4 casos en que no coinciden ambos métodos son:

Caso	Grupo original	Clasif. Discriminante	Clasif. logística
11	1	2	1
29	1	1	2
82	1	1	2
144	1	2	1

Esta diferencia es sólo el 2.31% que es mínima y en el 97.69% clasifican de forma igual. Por lo tanto, ambos métodos son prácticamente iguales en capacidad para clasificar.

Conclusiones

1. Según los resultados, los métodos "Análisis Discriminante" y la "Regresión Logística" han clasificado correctamente 72.83% cada uno de ellos y en este sentido no hay ninguna diferencia.

2. Hay una diferencia, pero, mínima en el número de clasificaciones que no coinciden, es decir, coinciden en el 97.69% de los 173 cangrejos y difieren en un 2.31%. Debo aclarar que no todas las clasificaciones de coincidencia son correctas y en este sentido esta conclusión es diferente a la primera, pero, no menos importante.

3. Ambos métodos clasifican correctamente el 87.4% en el grupo G2 y el 46.8% en el grupo G1, y este sentido tampoco hay ninguna diferencia.

4. De manera general, según los resultados, no hay mejor efectividad de un método respecto del otro, tanto el Análisis Discriminante como la Regresión Logística son iguales en efectividad para clasificar. Soló que, cuando intervienen variables independientes categóricas en el modelo, puede ser más adecuado el modelo de regresión logística ya que este si las permite; mientras que en el análisis discriminante se requieren que sean cuantitativas.

5. A pesar de la indiferencia de ambos métodos, para este tipo de problemas es más conveniente la Regresión Logística porque en este método si se pueden usar variables cualitativas como variables independientes, mientras que en el Análisis Discriminante se requiere que todas las variables independientes sean cuantitativas.

Bibliografía

Agresti, Alan. (2002), *Categorical Data Analysis*, segunda edición, John Wiley & Sons, United States of America

Christensen, Ronald. (1997), *Log-Linear Models and Logistic Regression*, second edition, Springer Verlag, United States of America.

Cuadras, Carles M. (2008), *Nuevos Métodos de Análisis Multivariante*, Universidad de Barcelona, España.

Dobson, Annette J. (2002), An introduction to generalized linear models, second edition, Chapman & Hall/crc, United States of America.

Fahrmeir, L. And Kaufman, H. (1985). Consistency and asymptotic normality of the maximun likelihood estimator in generalized linear models

Hardin, James W.; Hilbe, Joseph M. (2007), Generalized Linear Models and Extensions, second edition, Stata Press. United States of America.

J. F. Hair, Jr., R. E. Anderson, R. L . Tatham, W. C. Black, 1999, Análisis Multivariante, quinta edición, Pearson Prentice Hall.

Lindsey, James K. (1997), Applying Generalized Linear Models, Springer Verlag.

McCullagh, P and Nelder, J. A. (1989), Generalized Linear Models, second edition, Chapman and Hall.

Peña, Daniel. (2002), Análisis de Datos Multivariante, primera edición, McGraw-Hill \ Interamericana de España.

Pérez López, César. (2005), Técnicas Estadística con SPSS 12, primera edición, Pearson Prentice Hall.