UNIVERSIDAD
DE GRANADA

Facultad de Ciencias

Grado en Matemáticas

TRABAJO DE FIN DE GRADO

Estudio del Análisis Discriminante. Aplicación a datos reales

Presentado por:
María Polonio Sánchez

Curso académico 2021-2022

Estudio del Análisis Discriminante. Aplicación a datos reales

María Polonio Sánchez

María Polonio Sánchez *Estudio del Análisis Discriminante. Aplicación a datos reales*.
Trabajo de Fin de Grado. Curso académico 2021-2022.

Responsable de tutorización

Desiré Romero Molina
Departamento de Estadística e Investigación Operativa

Nuria Rico Castro
Departamento de Estadística e Investigación Operativa

Grado en Matemáticas

Facultad de Ciencias

Universidad de Granada

Declaración de originalidad

D./Dña. María Polonio Sánchez

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2021-2022, es original, entendida esta, en el sentido de que no ha utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 11 de septiembre de 2022

Fdo: María Polonio Sánchez

Índice general

Summary

The main aim of the study is to analyze Fisher's Linear Discriminant Analysis and to later apply real data using the R statistics software.

The LDA came about from a study made by Fisher in 1936 where, apart from the morphology of flowers, he studied and evaluated a lineal function to establish the differences between varieties of Iris (Setosa, Versicolor and Virgnica). Fisher did not verify all the hypotheses that are currently considered when applying said technique, but established their foundations. It consists in using a variable category that is a lineal combination of discriminating variables, measured at intervals or through use of reason, to find existing differences between the groups.

The LDA has two main aims. The first being to build discriminating functions, that allow us to explain the belonging of an individual to a group, as well as establish the weight of each variable in the discrimination. The second objective is to predict to which group it is most probable the individual belongs, knowing only certain variables.

This classifying technique, included in multivariable dependency techniques (those where variables are divided into two groups: dependent variables and independent variables), is applicable to many areas of knowledge. For example, in education, one tries to estimate students' academic performance based on educational and social factors. In medicine, it's used to diagnose illnesses and prescribe the most adequate treatment based on the characteristics of the patient. Finally, one of the most remarkable uses is in the economic scope for estimating cost effectiveness of a business based on variables such as income, debts and the patrimony of said business. Furthermore, it deduces whether it would be beneficial for a financial entity to approve a mortgage to its customers.

To conclude, while carrying out the study, said technique has been mathematically developed. In Chapter 1, we will investigate Fisher's Linear Discriminant Analysis in the case of 2 groups and following that, we will generalize in the case of K groups. This analysis is the most widely used and is implemented in the majority of software statistics. The main goal behind this technique is to minimize the variability in some groups and maximize the variability between others, as well as procuring the inverse of the first matrix for the second. From said matrix, we can obtain the highest own value and its associated vector. The elements of said vector will be the coefficients of the discriminant function we are looking for. Afterwards, we will explain the fundamental geometry of our discriminant functions. Furthermore, we will carry out an in depth study of the necessary assumptions for correct usage of said technique such as the absence of multicollinerity, the equality of variance-covariance matrices of each group and the normality of the discriminant variables in others.

In Chapter 2 we will carry out a study on Bayes' function of discrimination for the group 2 case and further groups, initially, without discriminant variables that follow a normal dis-

tribution and latterly, under normal circumstances, we will find said discriminant functions. Furthermore, we will study the cost of classification due to committing an error.

In Chapter 3, we will apply the theory from Chapter 1 to a set of real data. Specifically, the data used has been extracted from GENEIDA investigation project from the Poniente hospital in Almeria, whose aim is to study the effect on growth of a pregnant woman's fetus when exposed to atmospheric pollutants during pregnancy. In this study, only those women who where participants for the pregnancy control program in said hospital and between 12 and 14 weeks pregnant, were considered. Furthermore, they must be over 15, to have conceived naturally and be pregnant with one child. Moreover, women who have difficulty understanding the Spanish language or those who have been clinically diagnosed as having a chronic illness and are receiving treatment prior to pregnancy, will not be considered for the study. From the data received, we will look to establish discriminant functions that allow us to predict the type of birth. Firstly, we will classify as natural birth or assisted birth, assisted meaning cesarean or any surgical procedure from a set of 16 variables of which we will have previously selected based on a step by step method that I previously implemented, to be able to find the most significant. Secondly, we will analyze the model for classifying 3 groups, the objective being to distinguish between a natural birth, a vaginal birth and a cesarean using a lineal combination of the most significant variables.

Relating to the objectives of the study we can affirm that all of them have been fulfilled. For the elaboration of the study, a bibliographic recompilation has been made about all the subjects treated, gathering the most relevant information about each of them. Subsequently, the mathematical study of the technique has been developed along Chapters 1 and 2 and after, it has been applied and implemented in Chapter 3 to a set of real data.

Introducción

Con la finalidad de encontrar las diferencias existentes entre varios grupos a partir de las observaciones o mediciones de ciertas características y asignar un nuevo individuo o elemento basándonos en dichas características a uno de los grupos surge el análisis discriminante.

Esta técnica multivariante parte de un conjunto amplio de observaciones o individuos bajo estudio de las distintas variables de interés que proceden de distintos grupos para poder así construir unas funciones discriminantes que nos permitan explicar la pertenencia de cada observación a su grupo, el peso de cada variable de las observaciones en la discriminación y la clasificación de los nuevos individuos en uno de los grupos.

Una vez realizado el análisis obtenemos una regla de clasificación, es decir, un criterio que nos permita cuantificar la probabilidad de pertenencia de nuevos individuos a cada grupo observando las mismas variables de interés.

El análisis discriminante apareció por primera vez en un artículo en 1936 de Fisher, biólogo y estadístico inglés, que desarrolló un modelo lineal que distinguía tres especies de Iris entre sí (setosa, versicolor y virginica) en función de la forma y tamaño de los sépalos y pétalos de un conjunto de 50 muestras de cada una de las especies recolectados por el botánico Anderson en 1935. Sin embargo, no cumplía exactamente todas las hipótesis necesarias para el uso correcto del análisis. Aun así, se sigue utilizando como recurso para aprender el mecanismo de algunas técnicas discriminantes.

Esta técnica de clasificación, incluida dentro de las técnicas multivariantes de dependencia (aquellas en las que las variables están divididas en dos grupos: las variables dependientes y las independientes), es aplicable a muchas áreas del conocimiento. Por ejemplo, con respecto a la educación, se busca estimar el rendimiento académico de los alumnos, basándose no solo en resultados académicos sino en variables personales como pueden ser el entorno familiar, la autoestima personal, la seguridad en sí mismo o algún tipo de enfermedad. Otras aplicaciones en este ámbito son establecer si exite discriminación de raza o sexo en los colegios o establecer qué tipo de estudios sería más recomendable llevar a cabo según las capacidades del alumno.

En el ámbito de la medicina, esta técnica es utilizada para diagnosticar enfermedades, recetar los medicamentos más adecuados según las características del paciente, reconocer si un tumor es benigno o maligno, entre otras muchas.

Otros ejemplos de aplicación son estimar el nivel de rentabilidad de una empresa a través del estudio de unas ciertas variables como pueden ser ingresos, deudas y patrimonio de dicha empresa y estudiar si a una entidad financiera le saldría beneficiario conceder una hipoteca a sus clientes basándose en sus edades, rentas y antigüedades en el trabajo.

En este trabajo vamos a desarrollar la teoría del análisis discriminante centrándonos en el lineal. En el capítulo 1, desarrollaremos el análisis discriminante lineal de Fisher para el caso de 2 grupos y posteriormente, generalizaremos para el caso de k grupos siguiendo la idea de Rencher [7], Gil [3] y Cuadras [1]. Además, llevaremos a cabo un estudio exhaustivo de los supuestos necesarios para el uso correcto de dicha técnica.

En el capítulo 2, llevaremos a cabo el estudio de la función de discriminación de Bayes basándonos fundamentalmente en los desarrollos de Peña [5] y Cuadras [1]. Además, estudiaremos los costes de clasificación al cometer un error. Por último, en el capítulo 3, aplicaremos la teoría del capítulo 1 a un conjunto de datos reales sobre la distinción entre diferentes tipos de partos a partir de ciertas variables relacionadas con el desarrollo del embarazo de distintas mujeres del proyecto GENEIDA usando el software estadístico R [6]. También se ha implementado una función que nos permita seleccionar las variables que mayor discriminación provoquen entre los distintos grupos.

Con respecto a los objetivos del trabajo se puede afirmar que se han cumplido todos ellos. Para la elaboración del trabajo, se ha realizado una recopilación bibliográfica sobre todos los temas tratados, extrayendo la información más relevante de cada uno de ellos. Por consiguiente, se ha desarrollado el estudio matemático de la técnica en los capítulos 1 y 2 y posteriormente, ha sido aplicada e implementada en el capítulo 3 a un conjunto de datos reales.

1 Discriminante lineal de Fisher

En este capítulo estudiaremos los supuestos en los que se basa el análisis discriminante lineal de Fisher y desarollaremos el modelo para realizar una clasificación en dos o más grupos.

1.1. Clasificación en 2 grupos

Consideremos dos grupos $(k = 2)$ que denotaremos por G_1 y G_2. Nuestro objetivo es construir funciones discriminantes que nos permitan aplicar una regla de clasificación para asignar nuevos individuos en uno de los dos grupos conociendo las diferencias existentes de un conjunto de individuos en los distintos grupos, partiendo de una serie de supuestos a tener en cuenta:

- Se tiene una variable categórica y un conjunto de variables cuantitativas discriminantes (propiedades conocidas de los individuos bajo estudio) medidas en una escala de intervalo o de razón.
- Cada uno de los grupos debe tener al menos dos individuos.
- Los grupos han de ser mutuamente excluyentes, es decir, un individuo no puede pertenecer a dos grupos distintos.
- El número de variables discriminantes ha de ser menor que el número de individuos bajo estudio menos 2, es decir, sean $(X_1, X_2, ..., X_p)$, p variables discriminantes y n el número de individuos bajo estudio, entonces $p < (n - 2)$.
- Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes, es decir, debe haber ausencia de multicolinealidad entre ellas.
- Las matrices de varianzas-covarianzas de cada grupo han de ser aproximadamente iguales (lo que se conoce como propiedad de homogeneidad de matrices de varianzas-covarianzas).
- Las variables discriminantes tienen que regirse por una distribución normal multivariante.

1.1.1. Una sola variable clasificadora

En el caso en el que solo tengamos una variable clasificadora, X, nuestro objetivo será buscar una función lineal de dicha variable que permita la clasificación de un nuevo elemento en uno de los dos grupos según el valor que tome dicha función minimizando los errores de clasificación, siempre que ambos grupos tengan la misma forma y varianza, por lo que la

diferencia entre los grupos estará en la localización. Por tanto, se puede tomar como función lineal

$$C = \frac{\overline{X}^1 + \overline{X}^2}{2}$$

donde la media muestral del primer grupo, $\overline{X}^1$, es menor que la media muestral del segundo grupo, $\overline{X}^2$. Así pues, clasificaremos el nuevo elemento x según:

$$\begin{cases} \text{Si } x < C \Rightarrow x \in G_1. \\ \text{Si } x > C \Rightarrow x \in G_2. \end{cases}$$

En la Figura 1.1 se puede apreciar la zona problemática entre ambos grupos como el intervalo en la recta X donde se confunden los valores mayores del grupo 1 con los inferiores del grupo 2. Gracias al eje discriminante y a una buena elección de la regla de clasificación minimizamos el error de discriminación en dicha zona.

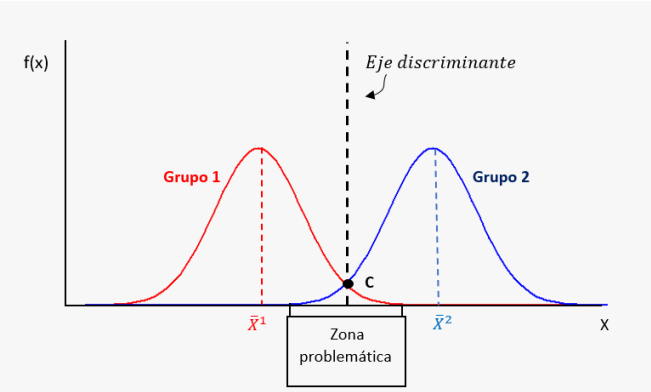

Figura 1.1: Análisis del problema en dos grupos con una sola variable discriminante.

1.1.2. Dos variables clasificadoras

Supongamos que disponemos de dos variables clasificadoras (X_1, X_2). Podemos pensar que solo es necesario utilizar una de las dos variables ya que nuestro objetivo es minimizar el error de clasificación. Por lo que se elegirá aquella variable que al proyectar el conjunto de datos observados en el eje correspondiente utilizando la misma función lineal que en el caso de una sola variable aparezca una zona problemática menor. Dicha zona se origina por la superposición de las distribuciones normales de los grupos. Por ejemplo, en la Figura 1.2 se puede observar que al proyectar los datos sobre el eje X_2, la zona de problema es menor que si proyectamos sobre el eje X_1. En consecuencia, la variable X_2 discrimina mejor que la variable X_1.

Sin embargo, como nuestro objetivo es minimizar la región problemática, Fisher buscó una función que minimice la dispersión dentro de cada grupo y maximice la separación entre ellos. Dicha función representada en la Figura 1.2 tiene la siguiente forma:

$$D = w_1 X_1 + w_2 X_2,$$

donde w_j son los coeficientes de ponderación de cada variable clasificadora para $j = 1, 2$ cuya obtención se explicará para el caso general de p variables clasificadoras que veremos a continuación.

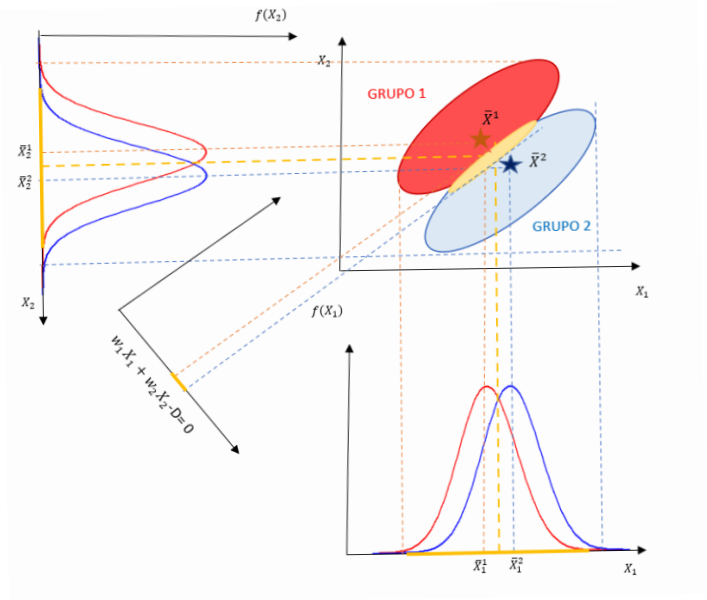

Figura 1.2: Análisis del problema en dos grupos con dos variables.

1.1.3. Generalización a p variables clasificadoras

Generalizamos los casos anteriores a $(X_1, \ldots, X_p)$ variables discriminantes basándonos en Rencher [7]. Fisher resolvió el problema minimizando la variabilidad dentro de los grupos y maximizando la variabilidad entre grupos mediante una combinación lineal de estas variables. Dicha combinación se conoce como función discriminante lineal de Fisher:

$$D = w_1X_1 + w_2X_2 + \ldots + w_pX_p \tag{1.1}$$

donde w_j son los coeficientes de ponderación de cada variable clasificadora para $j = 1, 2, \ldots, p$. Por tanto, nuestro objetivo es calcular dichos coeficientes y para ello, partimos de n observaciones.

Puesto que la puntuación discriminante correspondiente a la observación i-ésima viene dada por $D_i = w_1X_{i1} + w_2X_{i2} + \ldots + w_pX_{ip}, \quad i = 1, \ldots, n$, se puede expresar de manera matricial la función (1.1) como:

$$\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} \tag{1.2}$$

Reescribiendo el sistema (1.2) respecto a las desviaciones de las medias quedaría:

$$\begin{pmatrix} D_1 - \overline{D} \\ D_2 - \overline{D} \\ \vdots \\ D_n - \overline{D} \end{pmatrix} = \begin{pmatrix} X_{11} - \overline{X}_1 & X_{12} - \overline{X}_2 & \cdots & X_{1p} - \overline{X}_p \\ X_{21} - \overline{X}_1 & X_{22} - \overline{X}_2 & \cdots & X_{2n} - \overline{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \overline{X}_1 & X_{n2} - \overline{X}_2 & \cdots & X_{np} - \overline{X}_p \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} \tag{1.3}$$

donde $\overline{D} = w_1\overline{X}_1 + w_2\overline{X}_2 + \ldots + w_p\overline{X}_p$ siendo $\overline{X}_l = \frac{\sum_{i=1}^{n} X_{il}}{n}$, la media muestral de la variable l, para $l = 1, \ldots, p$.

Por tanto, se puede resumir (1.3) como la función discriminante en diferencias:

$$d = Xw.$$

A partir de esto, podemos obtener la variabilidad de la función discriminante, es decir, la suma de los cuadrados de las desviaciones de las variables con respecto a su media:

$$d'd = w'X'Xw \tag{1.4}$$

donde $X'X$ es la matriz simétrica de dimensión pxp que expresa las desviaciones cuadráticas con respecto a la media de las variables.

Como vimos en la asignatura de Inferencia Estadística:

Teorema 1.1. *La matriz de desviaciones cuadráticas $X'X$ se puede descomponer como la suma de la variabilidad entre grupos (E) y la variabilidad dentro de los grupos o intragrupos (I).*

En primer lugar, definimos E e I como

Definición 1.1. Variabilidad entre grupos (E):

$$E = \sum_{j=1}^{2} n_j(\overline{X}^j - \overline{X})(\overline{X}^j - \overline{X})'$$

$$\text{donde} \quad \overline{X} = \frac{\sum_{j=1}^{2} n_j \cdot \overline{X}^j}{n} \quad \text{con } n = n_1 + n_2$$

siendo n_i el número de observaciones bajo estudio en cada uno de los 2 grupos y $\overline{X}^j$ el centroide del j-ésimo grupo definido como

$$\overline{X}^j = \left(\overline{X}^j_1, \ldots, \overline{X}^j_l, \ldots, \overline{X}^j_p\right)' \quad j = 1, 2$$

$$\text{con} \quad \overline{X}^j_l = \frac{\sum_{i=1}^{n_j} X^j_{il}}{n_j}, \quad l = 1, \ldots, p.$$

Definición 1.2. Variabilidad dentro de los grupos o intragrupos:

$$I = \sum_{i=1}^{n_1}(X_i - \overline{X}^1)(X_i - \overline{X}^1)' + \sum_{i=1}^{n_2}(X_i - \overline{X}^2)(X_i - \overline{X}^2)'$$

donde $X_i = (X_{i1}, \ldots, X_{il}, \ldots, X_{ip})'$, $i = 1, \ldots, n_j$ para cada $j = 1, 2$.

A continuación, demostramos el teorema para el elemento l-ésimo de la diagonal de la matriz $X'X$ ya que para los elementos cruzados el procedimiento es análogo.

$$\sum_{i=1}^{n}(X_{il} - \overline{X}_l)^2 \underbrace{=}_{n=n_1+n_2} \sum_{i=1}^{n_1}(X_{il} - \overline{X}_l)^2 + \sum_{i=1}^{n_2}(X_{il} - \overline{X}_l)^2 =$$
$$\sum_{i=1}^{n_1}(X_{il} - \overline{X}_l^1 + \overline{X}_l^1 - \overline{X}_l)^2 + \sum_{i=1}^{n_2}(X_{il} - \overline{X}_l^2 + \overline{X}_l^2 - \overline{X}_l)^2 =$$
$$\sum_{i=1}^{n_1}(X_{il} - \overline{X}_l^1)^2 + \sum_{i=1}^{n_1}(\overline{X}_l^1 - \overline{X}_l)^2 + \underbrace{2\sum_{i=1}^{n_1}(X_{il} - \overline{X}_l^1)(\overline{X}_l^1 - \overline{X}_l)}_{=0,\text{suma de las desviaciones respecto a la media}} +$$

$$\sum_{i=1}^{n_2}(X_{il} - \overline{X}_l^2)^2 + \sum_{i=1}^{n_2}(\overline{X}_l^2 - \overline{X}_l)^2 + \underbrace{2\sum_{i=1}^{n_2}(X_{il} - \overline{X}_l^2)(\overline{X}_l^2 - \overline{X}_l)}_{=0,\text{suma de las desviaciones respecto a la media}} =$$
$$n_1(\overline{X}_l^1 - \overline{X}_l)^2 + n_2(\overline{X}_l^2 - \overline{X}_l)^2 + \sum_{i=1}^{n_1}(X_{il} - \overline{X}_l^1)^2 + \sum_{i=1}^{n_2}(X_{il} - \overline{X}_l^2)^2 =$$
$$\sum_{j=1}^{2} n_j(\overline{X}_l^j - \overline{X}_l)^2 + \sum_{j=1}^{2}\sum_{i=1}^{n_j}(X_{il} - \overline{X}_l^j)^2.$$

Por consecuente, $X'X = E + I$.

Por tanto la expresión (1.4) quedaría:

$$d'd = w'X'Xw = w'(E + I)w = w'Ew + w'Iw.$$

Luego, para buscar los D_i que mayor discriminación provoquen entre los grupos, maximizamos la variabilidad entre grupos y minimizamos la variabilidad dentro de los grupos, es decir, se debe encontrar

$$\text{máx}\left[\frac{w'Ew}{w'Iw}\right] \tag{1.5}$$

Equivalentemente, podemos calcular el máximo de $w'Ew$ tal que $w'Iw = 1$ ya que (1.5) es invariante frente a cambios de escala.

Ahora, aplicamos transformadores de Lagrange:

$$L = w'Ew + \lambda(w'Iw - 1)$$

y buscamos el máximo de dicha función igualando a 0 su derivada respecto a ω, es decir, $0 = \frac{\partial L}{\partial w}$.

$$\frac{\partial L}{\partial w} = 2Ew + 2I\lambda w = 0 \;\rightarrow\quad Ew = \lambda Iw \quad\rightarrow\quad (I^{-1}E)w = \lambda w.$$

Y como $Ew = \lambda Iw$ tenemos que $w'Ew = \lambda w'Iw = \lambda$.

Por lo que tomando el vector propio asociado al mayor valor propio se obtendrá los coeficientes de la función que mayor discriminación provoque.

1.1.4. Clasificación

Ahora queremos clasificar un nuevo elemento $x \in \mathbb{R}^p$ en alguno de los dos grupos G_1 o G_2. Para ello, calculamos su puntación discriminante, a la que denotaremos por d, introduciendo los valores de cada variable en la función discriminante.

Y por otro lado, calculamos el punto de corte o la frontera discriminante que viene dado por:

$$C = \frac{\overline{D}_1 + \overline{D}_2}{2}$$

donde

$$\begin{cases} \overline{D}_1 = w_1\overline{X}_1^1 + w_2\overline{X}_2^1 + \ldots + w_p\overline{X}_p^1 \\ \overline{D}_2 = w_1\overline{X}_1^2 + w_2\overline{X}_2^2 + \ldots + w_p\overline{X}_p^2 \end{cases} \tag{1.6}$$

con $w_1, \ldots, w_p$ los elementos del vector propio asociado al mayor valor propio de la matriz $I^{-1}E$.

Así pues, aplicaremos el siguiente criterio para clasificar el nuevo elemento:

$$\begin{cases} \text{Si } d < C\ (\Leftrightarrow d - C < 0), \text{entonces } x \text{ pertenece al grupo } 1. \\ \text{Si } d > C\ (\Leftrightarrow d - C > 0),\ \text{entonces } x \text{ pertenece al grupo } 2. \end{cases}$$

Hay que tener en cuenta que para la obtención de dicho criterio de clasificación no se ha considerado el tamaño de los individuos observados en cada grupo. Esto provocaría que la proporción de individuos mal clasificados en el grupo de menor tamaño será mayor que en el otro grupo. Para solucionarlo, podemos desplazar dicho punto al centroide del grupo con menor tamaño para igualar los errores de clasificación, es decir,

$$C = \frac{n_1 \times \overline{D}_1 + n_2 \times \overline{D}_2}{n_1 + n_2} \tag{1.7}$$

1.1.5. Coeficiente de correlación canónica

Para ver hasta qué punto discrimina nuestra función, es útil estudiar el coeficiente de correlación canónica que viene dada por

$$\eta = \sqrt{\frac{\lambda}{1+\lambda}} \tag{1.8}$$

donde λ es el valor propio asociado a la función discriminante.

Este coeficiente toma valores entre 0 y 1, entendiendo que para valores cercanos a 1, la discriminación que provoca la función será mayor.

1.1.6. Ejemplo

Con el siguiente ejemplo extraído de Cuadras [1] ponemos en práctica la técnica utilizada en el epígrafe 1.1. Para ello, consideramos dos especies de moscas distintos: Amerohelea pseudofascinata (G_1) y Amerohelea fascinata (G_2) y 2 variables: X_1 y X_2, donde X_1 = longitud de la antena (mm)y X_2 = longitud del ala (mm). Recolectamos del primer grupo 9 observaciones y del segundo 6, es decir, $n_1 = 6$ y $n_2 = 9$, luego $n = 15$.

En las siguientes tablas observamos los datos recolectados:

Grupo 1

Grupo	X_1	X_2
1	1.14	1.78
1	1.20	1.86
1	1.18	1.96
1	1.30	1.96
1	1.26	2.00
1	1.28	2.00

Grupo 2

Grupo	X_1	X_2
2	1.38	1.64
2	1.40	1.70
2	1.24	1.72
2	1.36	1.74
2	1.38	1.82
2	1.48	1.82
2	1.54	1.82
2	1.38	1.90
2	1.56	2.08

En primer lugar, calcularemos la función discriminante lineal de Fisher.
Para ello, como hemos visto en 1.1.3 tenemos que hallar el valor propio de la matriz $I^{-1}E$ y el vector propio asociado a dicho valor propio serán los coeficientes de nuestra función discriminante.

Calculamos E e I según (1.1) y (1.2), para lo cual necesitamos previamente obtener los valores de las medias de cada variable en cada grupo y en total:

$$\begin{cases} \overline{X}^1 = (1.227, 1.927)' \\ \overline{X}^2 = (1.413, 1.804)' \end{cases}$$

$$\overline{X} = \frac{n_1 \overline{X}^1 + n_2 \overline{X}^2}{n} = (1.3386, 1.8532)'.$$

Así pues, las matrices E e I vienen dadas:

$$E = \begin{pmatrix} 0.0544 & -0.00823 \\ -0.0823 & 0.12420 \end{pmatrix} \quad I = \begin{pmatrix} 0.1739 & 0.0864 \\ 0.0864 & 0.0981 \end{pmatrix}.$$

Luego,

$$I^{-1}E = \begin{pmatrix} 1.297326 & -1.202573 \\ -1.981539 & 2.325202 \end{pmatrix}. \tag{1.9}$$

Calculamos el valor propio dominante de la matriz (1.9) y es $\lambda = 3.43825$ y su vector propio asociado es

$$w = \begin{pmatrix} 0.4897366 \\ -0.8718705 \end{pmatrix}.$$

Por tanto, la función discriminante lineal de Fisher es:

$$D = 0.4897366 X_1 - 0.8718705 X_2.$$

Una vez obtenida dicha función vamos a clasificar las siguientes tres observaciones en una de las dos especies. Dichas observaciones son:

- Obs 1 : $(1.37, 1.79)$.
- Obs 2 : $(1.17, 1.95)$.
- Obs 3 : $(1.4, 1.83)$.

Para ello, calculamos el valor en cada caso de la función discriminante:

$$\begin{cases} D_{obs1} = 0.4897366 \times 1.37 - 0.8718705 \times 1.79 = -0.889709. \\ D_{obs2} = 0.4897366 \times 1.17 - 0.8718705 \times 1.95 = -1.12715. \\ D_{obs3} = 0.4897366 \times 1.4 - 0.8718705 \times 1.83 = -0.90989. \end{cases}$$

Calculemos ahora el punto discriminante (1.7) pero previamente necesitamos hallar (1.6)

$$\begin{cases} \overline{D}_1 = 0.4897366 \times 1.227 - 0.8718705 \times 1.927 = -1.079187. \\ \overline{D}_2 = 0.4897366 \times 1.413 - 0.8718705 \times 1.804 = -0.880856. \end{cases}$$

Luego,

$$C = \frac{6 \times -1.079187 + 9 \times -0.880856}{15} = -0.96018.$$

Entonces, aplicando el criterio de clasificación del epígrafe 1.1.4 , es decir, si $D_{obs} < C \Rightarrow obs \in G_1$.

$$\begin{cases} D_{obs_1} = -0.889709 > C = -0.96018 \Rightarrow obs_1 \in G_2. \\ D_{obs_2} = -1.12715 < C = -0.96018 \Rightarrow obs_2 \in G_1. \\ D_{obs_3} = -0.90989 > C = -0.96018 \Rightarrow obs_3 \in G_2. \end{cases}$$

Las observaciones 1 y 3 serán del tipo Amerohelea fascinata mientras que la observación 2 será una mosca de especie Amerohelea pseudofascinata.

Por último, para ver el grado de discriminación de nuestra función calculemos el coeficiente de correlación canónica dado por (1.8)

$$\eta = \sqrt{\frac{3.43825}{1 + 3.43825}} = 0.8801.$$

Luego, como se trata de un valor muy cercano a 1, la función lineal de Fisher provoca una gran discriminación entre los grupos.

1.2. Clasificación en más de 2 grupos

El objetivo ahora es generalizar la idea del epígrafe 1.1 para el caso de k grupos, con la finalidad de obtener funciones discriminantes que nos permitan clasificar a cada individuo al grupo de pertenencia más probable.

Para poder aplicar de forma correcta el análisis han de cumplirse los supuestos 1.1. Además, el número máximo de funciones discriminantes que se podrán construir bajo estos supuestos viene dado por $m = min(p, k-1)$.

Denotaremos por $(D_1, D_2, \ldots, D_m)'$ a estas funciones discriminantes, que serán combinaciones lineales de nuestras variables con la misma idea que la utilizada para dos grupos, es decir, minimizando la varianza entre grupos (E) y maximizando la varianza dentro de los grupos (I).

El modelo vendrá dado por:

$$\begin{cases} D_1 = w_{11}X_1 + w_{12}X_2 + \ldots + w_{1p}X_p \\ D_2 = w_{21}X_1 + w_{22}X_2 + \ldots + w_{2p}X_p \\ \vdots \\ D_m = w_{m1}X_1 + w_{m2}X_2 + \ldots + w_{mp}X_p \end{cases}$$

donde cada $D_i = w_{i1}X_1 + w_{i2}X_2 + \ldots + w_{ip}X_p$ verifica que $Corr(D_i, D_j) = 0, \forall i \neq j$.

1.2.0.1. Obtención de las funciones discriminantes

Con la idea de obtener dichas funciones discriminantes bajo las condiciones mencionadas anteriormente, realizaremos el siguiente procedimiento:

- D_1 será combinación lineal de las p variables discriminantes que mayor distinción ocasionen entre los grupos.
- D_2 será combinación lineal de las p variables discriminantes que mayor discriminación provoque entre los grupos, después de D_1, con la condición de que $Corr(D_1, D_2) = 0$.
- En general, calcularemos los D_i como combinación lineal de las variables discriminantes que mayor discriminación produzcan, después de D_{i-1} tal que verfique que $Corr(D_i, D_j) = 0, \forall j = 1, \ldots, i-1$.

Definición 1.3. Llamaremos **funciones discriminantes canónicas** a aquellas funciones $D_i = w_{i1}X_1 + w_{i2}X_2 + \ldots + w_{ip}X_p$ tales que son combinación lineal de $(X_1, \ldots, X_p)'$ p variables tipificadas.

Expresando matricialmente obtenemos $D_i = w_i'X$, buscando así D_i como la función que provoque mayor discriminación entre grupos siempre bajo la condición de que la correlación entre dichas funciones sea nula.

Generalizando para el caso de k grupos, las expresiones de la variabilidad entre grupos E y la variabilidad dentro de los grupos I se reescriben como:

Definición 1.4.

$$E = \sum_{j=1}^{k} n_j(\overline{X}^j - \overline{X})(\overline{X}^j - \overline{X})' \tag{1.10}$$

Definición 1.5.

$$I = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_i^j - \overline{X}^j)(X_i^j - \overline{X}^j)' \tag{1.11}$$

siendo

$$X_i^j = \left(X_{i1}^j, \ldots, X_{il}^j, \ldots, X_{ip}^j\right)' \quad j = 1, \ldots, k$$

donde X_{il}^j corresponde al individuo i observado en la variable l en el grupo j.

Aplicando el teorema 1.1 para k grupos, obtenemos que

$$Var(D_i) = Var(w_i'X) = w_i'Var(X)w_i = w_i'(E + I)w_i = w_i'Ew_i + w_i'Iw_i.$$

Puesto que buscamos minimizar la varianza entre grupos (E) (1.10) y maximizar la varianza dentro de los grupos (I) (1.11), llevaremos a cabo la misma idea que en el epígrafe 1.1.3, luego maximizamos:

$$\text{máx}\left[\frac{w_i'Ew_i}{w_i'Iw_i}\right]$$

Y como es invariante bajo cambio de escala, bastará con calcular el máx$[w_i'Ew_i]$ tal que $w_i'Iw_i = 1$.

Aplicando los multiplicadores de Lagrange, se define:

$$L = w_i' E w_i - \lambda(w_i' I w_i - 1) \;\rightarrow\; \tfrac{\partial L}{\partial w_i} = 2Ew_i + 2I\lambda w_i = 0$$
$$\rightarrow\; Ew_i = \lambda I w_i \quad \rightarrow \quad (I^{-1}E)w_i = \lambda w_i.$$

Y como $Ew_i = \lambda I w_i$ tenemos que $w_i' E w_i = \lambda w_i' I w_i = \lambda$.

Por consiguiente, para obtener las funciones discriminantes se sacan los autovectores de la matriz $(I^{-1}E)$ asociados a los valores propios elegidos de manera decreciente de la siguiente forma:

- λ_1, es el mayor valor propio de la matriz $(I^{-1}E)$ y w_1 el vector propio asociado a ese valor.
- λ_2, es el segundo mayor valor propio de la matriz $(I^{-1}E)$ y w_2 el vector propio asociado a ese valor.

$$\vdots$$

- λ_i, es el valor propio que ocupa la posición i ordenados de manera decreciente de la matriz $(I^{-1}E)$ y w_i el vector propio asociado a ese valor.

$$\vdots$$

- λ_m, es el último mayor valor propio de la matriz $(I^{-1}E)$ y w_m el vector propio asociado a ese valor.

Todos estos vectores son linealmente independientes, siempre que provengan de valores propios distintos, y dan lugar a funciones incorreladas siempre y cuando la matriz $(I^{-1}E)$ sea simétrica.

Para conocer la variabilidad total explicada por esta técnica se suman todos los valores propios, es decir, $\sum_{i=1}^{m} \lambda_i$. Por lo que, el porcentaje explicado por la función discriminante D_i del total de las funciones discriminantes viene dada por:

$$\frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} 100\,\%$$

1.2.0.2. Clasificación

Una vez obtenidas las funciones discriminantes podemos utilizar las propias observaciones utilizadas en la elaboración de dichas funciones para ver el grado de eficacia de las mismas desde el punto de vista de la clasificación.

Si dichos resultados son favorables, podemos usar estas funciones discriminantes para clasificar nuevos individuos en los distintos grupos conociendo los valores de las respectivas

variables discriminantes $(X_1, X_2, \ldots, X_p)'$ construyendo dichas funciones para cada grupo. Por lo que se clasificaría el nuevo individuo en el grupo cuya puntuación discriminante sea mayor.

1.2.1. Significado de los valores propios

La idea es determinar si cada uno de los valores propios λ_i son estadísticamente significativos, es decir, si aportan o no información distintiva entre los grupos.

Para ello, se realiza un test de contraste denominado test V de Barlett, en el que se contrasta:

$$\begin{cases} H_0 : \lambda_1 = \lambda_2 = ... = \lambda_m = 0 \\ H_1 : \lambda_i \neq 0 \quad \text{para al menos un } i \end{cases}$$

cuyo estadístico viene dado por:

$$V_0 = (n - 1 - \frac{p+k}{2}) \sum_{i=1}^{m} \log(1 + \lambda_i) \overset{n\to\infty}{\underset{H_0}{\rightsquigarrow}} \chi^2_{p(k-1)}$$

En el caso de que no se rechace la hipótesis nula (H_0), es decir, si el valor del estadístico en nuestra muestra no supera a $\chi^2_{p(k-1),\alpha}$ para un nivel de significación α fijado, no tendría sentido proseguir con el análisis, puesto que esto indicaría que las variables clasificadoras implicadas en el estudio no tendrían ningún tipo de poder discriminante significativo.

En caso contrario, si se rechaza la hipótesis nula significa que al menos el primer valor propio es estadísticamente significativo, ya que los valores propios estaban ordenados de manera descendente en cuanto a su poder de discriminación. En consecuencia, se pasaría a contrastar la significación conjunta del resto de valores propios, es decir:

$$\begin{cases} H_0 : \lambda_2 = ... = \lambda_m = 0 \\ H_1 : \lambda_i \neq 0 \quad \text{para al menos un } i \end{cases}$$

cuyo valor estadístico viene dado por:

$$V_1 = (n - 1 - \frac{p+k}{2}) \sum_{i=2}^{m} \log(1 + \lambda_i) \overset{n\to\infty}{\underset{H_0}{\rightsquigarrow}} \chi^2_{(p-1)(k-2)}$$

Por lo que si no se rechaza la hipótesis nula significaría que solo el primer valor propio proporciona un poder de discriminación significativo, en caso contrario, tanto el primero como el segundo valor propio, serían significativos.

Este procedimiento se realiza de manera reiterada hasta que ya no se rechace la hipótesis nula, es decir, si $V_i > \chi^2_{(p-i)(k-i-1),\alpha}$ para $i = 0, 1, \ldots, j-1$, se vuelve a hacer un test de

contraste para los $m - j$ siguientes valores propios, es decir:

$$\begin{cases} H_0 : \lambda_{j+1} = ... = \lambda_m = 0 \\ H_1 : \lambda_i \neq 0 \quad \text{para al menos un } i \end{cases}$$

cuyo valor estadístico es:

$$V_j = (n - 1 - \frac{p+k}{2}) \sum_{i=j+1}^{m} \log(1 + \lambda_i) \overset{n\to\infty}{\underset{H_0}{\rightsquigarrow}} \chi^2_{(p-j)(k-j-1)}$$

Por último, para determinar el poder discriminatorio de una función discriminante se utiliza el coeficiente de correlación canónica que mide la proporción de la variabilidad total debido a las diferencias entre grupos para cada función discriminante. Esto viene dado por:

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

donde λ_i es el valor propio asociado a la función discriminante i-ésima.
Por tanto, para valores cercanos a 1, la función discriminante i-ésima provocará una buena distincción entre grupos.

1.2.2. Significado geométrico de las funciones discriminantes

Siguiendo a Gil [3], veamos el significado geométrico de nuestras funciones discriminantes obtenidas en el apartado 1.2.

Para ello, dado que consideramos n el número de individuos bajo estudio y p el número de variables independientes podríamos representar los datos en la matriz de dimensión $n \times p$ siguiente:

Variable / Individuo	X_1	X_2	...	X_p
Individuo 1	X_{11}	X_{12}	...	X_{1p}
Individuo 2	X_{21}	X_{22}	...	X_{2p}
⋮	⋮	⋮	⋱	⋮
Individuo n	X_{n1}	X_{n2}	...	X_{np}

Consideramos las p variables discriminantes como los ejes del espacio que generan y cada individuo (fila) como un punto de dicho espacio. Por tanto, los valores obtenidos de un nuevo individuo en las p variables representan sus coordenadas en los ejes del espacio p dimensional que definen.
Como caso particular, consideramos 3 variables y apreciamos en la Figura 1.3 la posición que ocupa el punto m (un individuo) respecto a las variables X_1, X_2 y X_3.

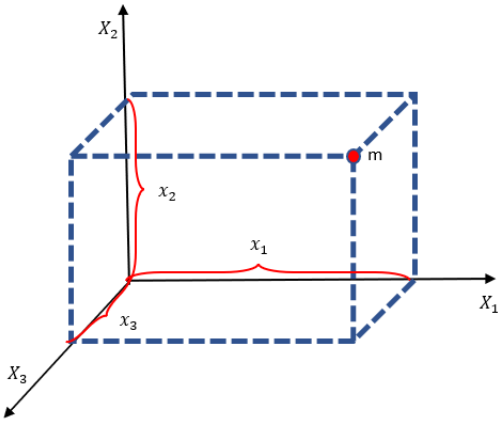

Figura 1.3: Posición del punto $m = (x_1, x_2, x_3)$ respecto a las variables X_1, X_2 y X_3.

Por consiguiente, un inviduo pertenecerá a un grupo si sus coordenadas espaciales son similares a las del resto de individuos que pertenecen a dicho grupo quedando localizados en la misma región del espacio como se puede apreciar en la Figura 1.4 :

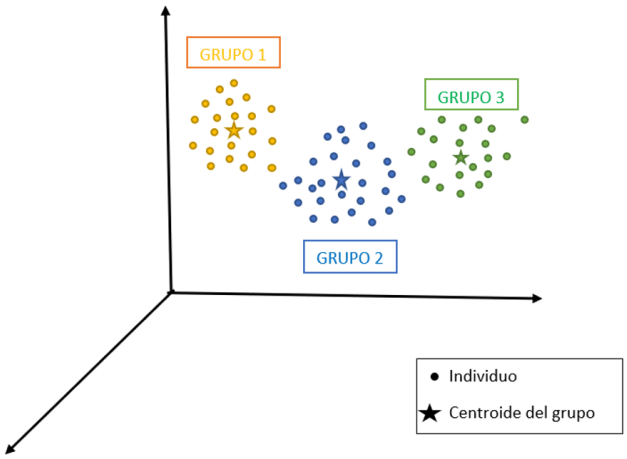

Figura 1.4: Posición de los individuos bajo estudio y el centroide de cada grupo.

Como el objetivo es maximar las diferencias entre grupos, se elegirán los ejes de manera que cumplan eso y a la vez, sean ortogonales entre sí, luego nuestro primer eje podemos situarlo en la dirección en la que los centroides aparezcan más separados, es decir, en la que el vector propio produzca una mayor discriminación. El segundo eje se elegirá de la misma manera pero a su vez tiene que ser ortogonal al primero y así sucesivamente, definimos los ejes que mayor discriminación provoquen siempre ortogonales entre sí.

Hay que tener en cuenta que para representar las posiciones de los centroides de cada grupo no es necesario que permanezcan al espacio p dimensional, es decir, k centroides definen un espacio de dimensión $k - 1$.

Como caso particular, en la Figura 1.5 apreciamos 3 centroides de 3 grupos distintos que definirán un plano y dos ejes ortogonales entre sí, siendo el primer eje (Y_1) el que corresponde a la función discriminante que mayor discriminación entre grupos produzca.

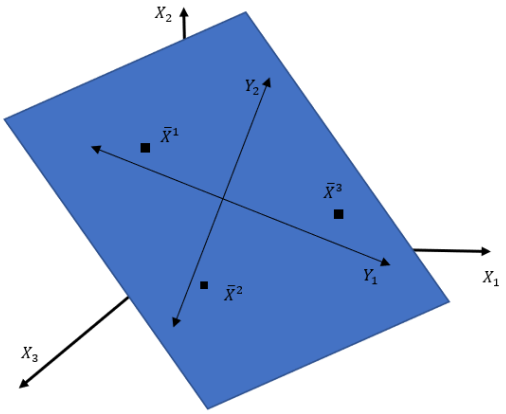

Figura 1.5: Ejes ortogonales de mayor dispersión para los centroides de cada grupo ocasinados por las variables X_1, X_2, X_3.

1.3. Supuestos y consecuencias por incumplimiento

Resumiendo todas las hipótesis indicadas en 1.1, el análisis discriminante se fundamenta sobre los siguientes requisitos:

a. Normalidad Multivariante.

b. Homogeneidad de matrices de varianzas-covarianzas.

c. Ausencia de multicolinealidad

A continuación, vamos a estudiar algunas técnicas estadísticas que nos permitirán comprobar estas hipótesis así como la robustez frente a violaciones de dichos supuestos.

1.3.1. Normalidad Multivariante

En primer lugar, que la distribución conjunta de las variables siga una normal multivariante implica que cada variable aleatoria también siga una distribución normal, sin embargo, el recíproco no es cierto en general. En nuestro caso, puesto que las variables son incorreladas, bajo normalidad, equivale a que sean independientes entre sí. Por tanto, comprobar que se verifica el supuesto de normalidad multivariante es equivalente a comprobar la normalidad univariante de cada variable.

Este supuesto será fundamental para los test de significación ya que, en el caso en el que no se verifique, la distribución del estadístico será diferente de la distribución teórica de la muestra.

La robustez frente a violaciones de la normalidad es mayor cuando esta proviene de la asimetría, siendo más preocupante para pequeñas muestras de tamaño desigual. Esto provocaría que el cálculo de probabilidades de pertenecia a un grupo sea menos preciso por lo que habría que estudiar con cuidado los casos que se encuentran en el límite entre dos grupos, es decir, aquellos en los que el valor de su probabilidad esté en torno a 0.5 ya que un pequeño error conduciría a una incorrecta clasificación del individuo.

A continuación, comprobamos la normalidad univariante de cada variable de manera gráfica (cuantil-cuantil) o a través de test de significación como pueden ser: el test de Kolmogorov-Smirnov o el test de Shapiro-Wilk.

1.3.1.1. Gráficos cuantil-cuantil

Los gráficos cuantil- cuantil, conocidos como Q-Q Plot, son gráficos de dispersión donde se enfrentan los cuantiles de la distribución teórica (eje X) frente a los cuantiles de la distribución de un conjunto de datos (eje Y). Por tanto, en nuestro caso, compararemos la distribución teórica normal con la distribución de un conjunto de datos. Si ambos cuantiles proceden de la misma distribución, se podrá observar que los puntos del gráfico se aproximan a una línea recta.

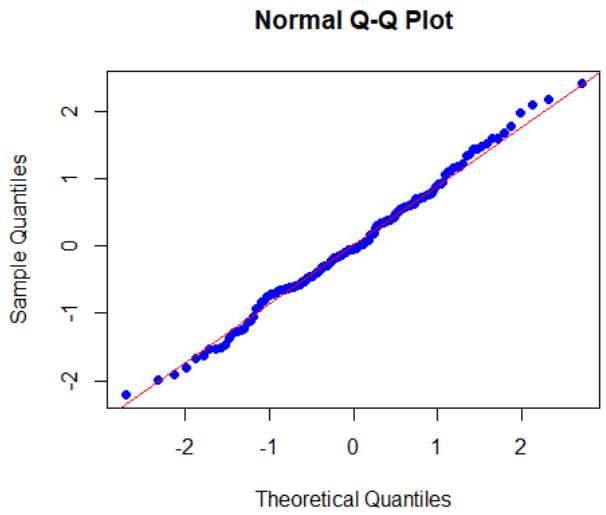

Figura 1.6: Ejemplo de Q-Q Plot Normal

En la Figura 1.6 podemos ver un gráfico Q-Q normal de un conjunto de datos que proceden de una variable con distribución normal. Aunque este gráfico nos permite ver de forma rápida si se cumple el supuesto de normalidad univariante a veces su interpretación puede ser algo subjetivo. Es por ello por lo que sería recomiendable confirmar la validación de la hipótesis si el gráfico origina dudas mediante un test de significación como los que presentan a continuación:

1.3.1.2. Test de Kolmogorov-Smirnov

Consideremos $(X_1, \ldots, X_p)$ una muestra aleatoria simple de una variable aleatoria continua X que se distribuye según una función de distribución F cualquiera y $F^*_{X_1,\ldots,X_p}$ la función de distribución muestral o empírica.

El contraste de hipótesis a estudiar sería:

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases} \tag{1.12}$$

Puesto que el test trata de comparar F y F_0, siendo esta última la función teórica de la variable X, en el caso de que sea F_0 la función distribución normal pero esta no esté totalmente determinada, es decir, que se desconozca el valor de su media y de su varianza, será necesario estimar estos parámetros mediante máxima verosimilitud en función de las observaciones de la muestra. En este caso el test pasa a denominarse test de Lilliefors.
Supongamos que en nuestro caso F_0 está totalmente determinado.

Para resolver (1.12) utilizamos el estadístico de Kolmogorov-Smirnov:

$$D(X_1, ..., X_p) = \underset{x\in\mathbb{R}}{sup}|F^*_{X_1,\ldots,X_p} - F_0(x)|$$

Por tanto, el contraste se resolvería con el siguiente test:

$$\varphi(X_1, \ldots, X_p) = \begin{cases} 1 & D(X_1, ..., X_p) \geq d_\alpha \\ 0 & D(X_1, ..., X_p) < d_\alpha \end{cases}$$

donde d_α verifica

$$P_{H_0}(D(X_1, ..., X_p) \geq d_\alpha) = \alpha$$

o mediante la obtención del

$$p_{valor} = P_{H_0}(D(X_1, ..., X_p) \geq D_t),$$

siendo

- D_t el valor estadístico teórico en la muestra observada.
- $D(X_1, ..., X_p)$ bajo la hipótesis nula (H_0) sigue una distribución de Kolmogorov, ya que F_0 es continua.

En conclusión, se rechazaría la hipótesis nula si la diferencia entre $F^*_{X_1,\ldots,X_p}$ y $F_0(x)$ es muy grande. Hay que tener en cuenta que este test es recomendable cuando el número de observaciones es mayor que 50 puesto que no requiere que los datos se agrupen u ordenen de una manera concreta.

1.3.1.3. Shapiro-Wilk

Consideremos $(X_1, \ldots, X_p)$ una muestra aleatoria simple de una variable aleatoria continua X.

Mientras que el test de Kolmogorov- Smirnov sirve para cualquier distribución continua, el de Shapiro-Wilk solo sirve para resolver el contraste de ajuste de una normal pero tiene la ventaja de que no hace falta que la distribución esté completamente determinada.

El contraste de hipótesis a estudiar sería:

$$\begin{cases} H_0 : X \backsim N(\mu, \sigma^2) \\ H_1 : X \nsim N(\mu, \sigma^2) \end{cases}$$

Para resolverlo utilizaremos el estadístico de Shapiro-Wilk:

$$W = \frac{\left(\sum_{i=1}^{n} a_i (X_{p-i+1} - X_i) \right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

donde a_i es el valor de un coeficiente que se encuentra tabulado en las tablas de Shapiro-Wilk que se pueden encontrar, por ejemplo, en de la Garza [2] para la posición i de cada observación de la muestra, $\overline{X}$ es la media muestral y los términos $X_{p-i+1} - X_i$ corresponden a las diferencias de restar el primer valor al último valor, el segundo al penúltimo y así sucesivamente, habiendo ordenado previamente los valores de manera ascendente.

Por lo tanto, el test de Shapiro-Wilk quedaría:

$$\varphi(X_1, \ldots, X_p) = \begin{cases} 1 & W \geq w_\alpha \\ 0 & W < w_\alpha \end{cases}$$

donde w_α verifica

$$P_{H_0}(W \geq w_\alpha) = \alpha$$

o mediante la obtención del

$$p_{valor} = P_{H_0}(W \geq W_{exp}),$$

siendo

- w_α el valor estadístico proporcionado por la tabla de Shapiro-Wilk para ese tamaño muestral y con nivel de significación α dado.
- W_{exp} el valor del estadístico en la muestra observada.
- W bajo la hipótesis nula (H_0) sigue una distribución de Shapiro-Wilk.

En conclusión, no se rechazaría la hipótesis nula de normalidad si el estadístico W_{exp} es mayor que el valor estadístico tabulado.

Sin embargo, también hay tests en los que directamente se comprueba la normalidad multivariante, como es el caso de los test de Mardia y Shapiro-Wilk generalizado.

1.3.1.4. Test de Mardia

El test de Mardia está basado en el estudio multivariante de la asimetría y la curtosis de la distribución, que vienen dadas según Mardia [4] por:

- Asimetría:
$$\gamma = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}^3$$
- Curtosis:
$$\delta = \frac{1}{n}\sum_{i=1}^{n} D_{ii}^2$$

donde $D_{ij} = (X_i - \overline{X})'\Sigma^{-1}(X_j - \overline{X})$, es decir, la distancia de Mahalanobis, siendo Σ la matriz global de varianzas-covarianzas.

En nuestro caso, X es una matriz de dimensión $n \times p$ y el contraste que se desea resolver con este test sería:

$$\begin{cases} H_0 : X \backsim N_{n\times p}(\mu, \Sigma) \\ H_1 : X \nsim N_{n\times p}(\mu, \Sigma) \end{cases} \tag{1.13}$$

Puesto que la distribución normal es simétrica ($\gamma = 0$) y mesocúrtica ($\delta = 0$), resolver el contraste antrior (1.13) es equivalente a resolver dos test distintos que solventar de manera independiente, el de asimetría y el de curtosis:

Asimetría	**Curtosis**
$\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma \neq 0 \end{cases}$	$\begin{cases} H_0 : \delta = 0 \\ H_1 : \delta \neq 0 \end{cases}$

En primer lugar, el estadístico para el test de asimetría $EAS = (\frac{n}{6})\gamma$ se aproxima por una distribución χ^2 con $f = \frac{p(p+1)(p+2)}{6}$ grados de libertad. De manera que el p-valor de la prueba es $P[\chi_f^2 > EAS]$. Se rechazará la hipótesis nula cuando el p-valor sea menor que el nivel de significación α dado.

Y en segundo lugar, el estadístico δ para el test de curtosis se aproxima por una normal de media $p(p+2)$ y varianza $\frac{8p(p+2)}{n}$. De manera que el p-valor de la prueba es $P\left[\frac{\delta - p(p+2)}{\sqrt{8p(p+2)/n}} > \delta_{exp}\right]$, donde δ_{exp} es el valor del estadístico en la muestra observada. Se rechazará la hipótesis nula cuando el p-valor sea menor que el nivel de significación α dado.

1.3.1.5. Test de Shapiro-Wilk generalizado

A partir del concepto de invarianza, un vector aleatorio es normal multivariante ($X \backsim N(\mu, \Sigma)$) si y solo si su estandarización $Z = \Sigma^{-\frac{1}{2}}(X - \mu) \backsim N(0, I_p)$ también lo es.

El estadístico de Shapiro-Wilk multivariante a considerar viene dado por:

$$W_{VG} = \frac{1}{p}\sum_{j=1}^{p} W_j$$

donde W_j, $\forall j = 1, \ldots, p$ son los estadísticos de Shapiro-Wilk, mencionados anteriormente 1.3.1.3, univariantes para cada Z_j como se puede ver en Villasenor [9] .

Al igual que en el caso univariante, se rechazaría la normalidad cuando el estadístico W_{VG} es menor que el valor estadístico tabulado por la tabla de Shapiro-Wilk.

1.3.2. Homogeneidad de matrices de varianzas-covarianzas

Las matrices de varianzas-covarianzas que han sido extraídas de cada grupo han de ser iguales. Este supuesto es conocido como homocedasticidad.

Para verificar dicho supuesto podemos recurrir a la prueba M de Box. Según Timm [8], esta prueba consiste en un test que resuelve el contraste:

$$\begin{cases} H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_k \\ H_1 : \Sigma_i \neq \Sigma_j, \text{para algún } i \neq j \end{cases}$$

donde Σ_i es la matriz de varianzas-covarianzas para el grupo i.

Entonces, el estadístico M de Box vendrá dado por:

$$M = v_e \log(|\Sigma|) \sum_{n=1}^{k} v_i \log(|\Sigma_i|)$$

donde $v_e = n - k$ y $v_i = n_i - 1$.
Multiplicando M por $1 - C$ donde

$$C = \frac{2p^2 + 3p - 1}{6(p-1)(k-1)} \left[\sum_{n=1}^{k} \left(\frac{1}{v_i} - \frac{1}{v_e}\right)\right],$$

se tiene que $\chi^2 = (1 - C)M$ sigue asintóticamente y aproximadamente bajo la hipótesis nula una distribución χ^2 con $f = \frac{p(p+1)(k-1)}{2}$ grados de libertad.
En consecuencia, la hipótesis nula del test se rechaza si $\chi^2 > \chi^2_{f,\alpha}$ para un nivel de significación α. Hay que tener en cuenta que esta aproximación es razonable si $n_i \geq 20$. Si no fuera así, se usaría la distribución F de Snedecor en lugar de la χ^2 cuyos grados de libertad dependen de una serie de consideraciones que se puede encontrar en de la Garza [2].

Cuando los tamaños de los grupos son grandes o iguales, la robustez del análisis discriminante ante la violación de este supuesto es considerable. Sin embargo, cuando estos son pequeños o desiguales, los procedimientos de clasificación son especialmente sensibles a la violación de este supuesto. Si se incumpliera, los individuos bajo estudio deberían ser clasificados en los grupos que cuentan con mayor dispersión.

1.3.3. Ausencia de multicolinealidad

La multicolinealidad se produce cuando una variable discriminante es combinación lineal de las demás, es decir, cuando existe una correlación múltiple entre dicha variable y las restantes. En ese caso, la variable que es combinación lineal de otras no aporta información nueva.

La aparición de dicha multicolinealidad entre las variables hace que las matrices de varianzas-covarianzas presenten singularidades, es decir, que su rango no sea máximo, por consiguiente, su determinante sea 0. Debido a esto, no se podrá realizar la inversión de estas matrices, las cuales son necesarias para calcular los coeficientes de la función discriminante lineal (1.1).

No obstante, no hace falta que la multicolinealidad sea perfecta para provocar singularidades sino que basta con que el determinante de la matriz de varianzas-covarianzas sea un valor próximo a 0 para que su inversa tenga valores inestables.

Para detectar multicolinealidad en estas matrices existen diferentes procedimientos. Uno de ellos, es utilizar la matriz de correlación de Pearson. Si al observar dicha matriz encontramos valores cercanos a 1 nos indicaría la presencia de variables redundantes.

Sin embargo, hay que tener en cuenta que esta comprobación solo detecta correlación entre parejas de dos variables. De ahí que sea conveniente de acuerdo con Gil [3] utilizar la regresión múltiple, tomando de forma reiterada cada vez una variable como dependiente y las restantes como independientes, por lo que, si la correlación múltiple al cuadrado (R^2) entre una combinación lineal de las variables independientes y la variable dependiente es alta, estas darán lugar a situaciones cercanas a la aparición de singularidades. En el caso de que esto ocurra la solución más rápida es eliminar dicha variable ya que no aporta suficiente información adicional sobre el resto de variables.

El problema de multicolinealidad se puede intentar evitar aplicando un método de selección de variables.

1.3.3.1. Selección de variables

Según Gil [3], no todas las variables discriminan de igual forma entre los grupos. Es por ello, que no es necesario incluir todas las variables de partida, sino que seleccionaremos las óptimas. Para ello, utilizaremos el método conocido como método paso a paso que se puede desarollar en tres direcciones diferentes:

- **Hacia adelante o *forward*.**
  En primer lugar, partimos de la variable que mayor discrimación produzca. Después, dicha variable formará pareja con las restantes, eligiéndose como segunda variable

aquella que pertenezca a la pareja que resulte más discriminante. A continuación, esta segunda variable formará pareja con las restantes y de manera reiterada se irán eligiendo variables hasta que no quede ninguna o se considere que las restantes no proporcionan más discriminación.

- **Hacia atrás o *backward*.**
  Partimos inicialmente de todas las variables y se elimina la que produzca menor discriminación. A continuación, eliminamos la siguiente que produzca menor discriminación y así sucesivamente se van excluyendo variables una a una hasta que entre las variables no eliminadas aporten una significante discriminación entre los grupos.

- ***Stepwise*.**
  Es una combinación de los métodos anteriores. Se parte de una selección hacia adelante de variables, aunque revisando en cada paso el conjunto de variables resultantes por el método hacia atrás, por si pudieramos excluir alguna de ellas.

Para determinar cuáles variables se incluyen, continúan o se excluyen del método paso a paso, nos basamos en el cálculo de algunos de los siguientes estadísticos:

- Λ de Wilks: mide las desviaciones dentro de cada grupo respecto a las desviaciones globales, sin diferenciar entre grupos. El valor Λ se calcula a partir de las diferencias entre los grupos y la homogeneidad de los mismos. Es decir,

  $$\Lambda = \frac{|I|}{|E+I|}$$

  donde E representa la variabilidad entre grupos (1.10) y I, la variabilidad dentro de los grupos (1.11).

  Por tanto, la variable que produce mayor discriminación es aquella cuyo valor de Λ sea el más pequeño.

- F de Snedecor: Se compara para una variable X_i, $\forall i = 1, \ldots, p$, las desviaciones de las medias de cada uno de los grupos con respecto a la media total, entre las desviaciones a la media dentro de cada grupo, es decir,

  $$F = \frac{|E|}{|I|} \tag{1.14}$$

  Por lo que,
  - Si F toma un valor elevado para una variable es indicio de que las medias de cada grupo están muy separadas y la variable discrimina bien.
  - Si F toma un valor pequeño para una variable, esto indica que dicha variable discrimina poco puesto que habrá homogeneidad en los grupos.

  Por tanto, la variable con mayor discriminación será la que produzca mayor F.

Tendremos también en cuenta que estas variables seleccionadas deben cumplir unas condiciones necesarias y en el caso de que no se verifiquen, han de ser descartadas. Estas condiciones vienen evaluadas por el nivel de tolerancia de dichas variables, dado por:

$$T = 1 - R^2$$

siendo R^2 la correlación múltiple entre esta variable y todas las variables ya incluidas, cuando estas han sido obtenidas a partir de la matriz de variabilidad entre grupos (1.2).

Luego, según Gil [3] para que una variable sea incluida es recomendable que $R^2 < 0.999$, es decir, un nivel de tolerancia mayor que 0.001.

2 Discriminador de Bayes

En este capítulo, desarollaremos otra función discriminante, la conocida como discriminador de Bayes, la cual aplicaremos para el estudio del análisis de dos o más grupos basándonos en los desarrollos de Peña [5] y Cuadras [1].

2.1. Clasificación en dos grupos

Con la finalidad de clasificar un nuevo individuo o elemento $x = (x_1, \ldots, x_p)'$ en uno de los dos grupos G_1 o G_2, denotaremos por f_1 y f_2 a las funciones de densidad, no necesariamente normales, de cada población, respectivamente. Además, suponemos conocidas las probabilidades a priori (π_1 y π_2, respectivamente) de que un individuo bajo estudio proceda de un grupo u otro.

Una vez observado el elemento x, pasaremos a calcular las distribuciones a posteriori gracias a la regla de Bayes de que este individuo provenga de un grupo u otro.

Teorema 2.1. ***Regla de Bayes:*** *Sean $G_1, \ldots, G_k$, k grupos mutuamente excluyentes y exhaustivos con probabilidades distintas de 0, osea, $\pi_i \neq 0 \ \forall i = 1, \ldots, k$. Sea x un individuo cualquiera del que se conocen las distribuciones condicionadas $f_i(x)$, entonces*

$$P(G_i|x) = \frac{f_i(x)\pi_i}{\sum_{j=1}^{n} f_j(x)\pi_j}$$

Por tanto, las probabilidades a posteriori $P(G_i|x)$, $i = 1, 2$ para cada grupo, respectivamente serán:

$$P(G_1/x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)} \tag{2.1}$$

$$P(G_2/x) = \frac{\pi_2 f_2(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)} \tag{2.2}$$

Para clasificar x en uno de los dos grupos comparamos únicamente los numeradores ya que los denominadores son iguales, es decir,

$$\text{si} \quad \pi_1 f_1(x) > \pi_2 f_2(x) \Rightarrow x \in G_1.$$

Por lo que la clasificación de x se hace en el grupo que tenga mayor probabilidad.

2.1.1. Coste de clasificación

Como cualquier otro método de clasificación, se pueden llegar a cometer errores que producirán unas consecuencias a la hora de clasificar un individuo en un grupo. Esto afectará en gran medida a la hora de tomar decisiones. Es por ello que será necesario encontrar una regla discriminante.

Supongamos que las posibles decisiones que tenemos que tomar en el problema son únicamente dos: asignar en G_1 o en G_2. Entonces, la regla de decisión es:

$$\begin{cases} \text{Si } \ D(x_1 \ldots, x_p) \geq 0 \Rightarrow x \in G_1 \\ \text{Si } \ D(x_1 \ldots, x_p) < 0 \Rightarrow x \in G_2 \end{cases}$$

Por lo tanto, esta regla divide en dos el espacio muestral E_x (Consideremos $\mathcal{R}^p$).

- $e_1 = \{\mathbf{x} \in \mathcal{R}^p \ / D(\mathbf{x}) > 0 \ \}$
- $e_2 = \{\mathbf{x} \in \mathcal{R}^p \ / D(\mathbf{x}) < 0 \ \}$

Luego, en el caso en que podamos cuantificar las consecuencias de los errores, intentamos incluirlos en la solución del problema formulándolo como un problema de decisión de Bayes.

Para ello,

a. Denotamos por $c(G_i|G_j)$ al coste de clasificar x en el grupo G_i cuando debería haber sido clasificado en G_j. Estos costes serán conocidos.

b. Minimizamos el coste esperado.

Luego, si clasificamos x en G_1 puede ocurrir:

a. Acertamos con probabilidad $P(G_1|x)$, por lo que, el coste de penalización sería 0.

b. No acertamos con probablidad $P(G_2|x)$, por lo que, el coste de penalización sería $c(G_1|G_2)$.

En la Figura 2.1 podemos ver cómo un problema de clasificación entre dos grupos se puede ver como un problema de toma de decisiones bajo las ideas anteriores:

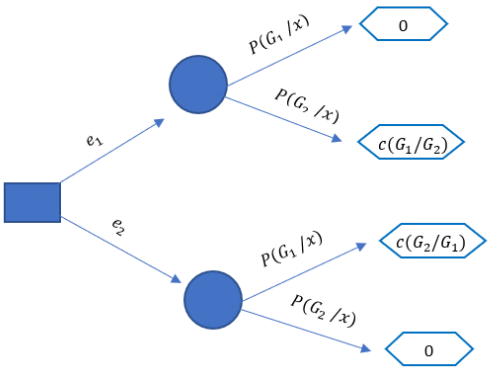

Figura 2.1: Problema de toma de decisiones

Por tanto, el coste promedio o valor esperado de tomar la decisión e_1, es decir, clasificar en el grupo 1, será:

$$E(e_1) = 0P(G_1|x) + c(G_1|G_2)P(G_2|x) = c(G_1|G_2)P(G_2|x).$$

Análogamente, el coste promedio de tomar la decisión e_2 será:

$$E(e_2) = c(G_2|G_1)P(G_1|x) + 0P(G_2|x) = c(G_2|G_1)P(G_1|x).$$

En consecuencia, asignaremos x al grupo 1 si su coste esperado es menor, es decir, utilizando (2.1) y (2.2) si:

$$\frac{f_1(x)\pi_1}{c(G_1|G_2)} > \frac{f_2(x)\pi_2}{c(G_2|G_1)}.$$

Por lo que bajo esta condición, clasificaríamos el elemento x en G_1 si:

- Su probabilidad a priori es más alta.
- La función de densidad evaluada en el elemento es más alta.
- El coste de clasificación es menor si nos equivocamos.

2.1.2. Función discriminante de Bayes bajo normalidad

En este apartado, analizaremos el problema planteado anteriormente 2.1.1 con el supuesto de que f_1 y f_2 siguen distribuciones normales con distinto vector de medias (denotemos por μ_1 al vector de medias de G_1 y μ_2 al de G_2) pero idéntica matriz de varianzas-covarianzas (Σ).

La función de densidad de una distribución normal p variante de media $\mu_i \in \mathbb{R}^p$ y matriz de varianzas-covarianzas $\Sigma_p \times p$ viene dada por:

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp - \frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i) \tag{2.3}$$

De acuerdo a la conclusión del epígrafe anterior 2.1.1, clasificaremos el individuo x en la población G_1 si

$$\frac{f_1(x)\pi_1}{c(G_1|G_2)} > \frac{f_2(x)\pi_2}{c(G_2|G_1)}$$

Puesto que la función de densidad para ambas poblaciones solo toma valores positivos, podemos aplicar logaritmos a ambas lados de la desigualdad de modo que queda:

$$\ln \frac{f_1(x)\pi_1}{c(G_1|G_2)} > \ln \frac{f_2(x)\pi_2}{c(G_2|G_1)}$$

Ahora, sustituyendo la expresión de la función de densidad 2.3, aplicando propiedades de los logaritmos y puesto que $\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}$ es común en ambas lados de la desigualdad, obtenemos que:

$$-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \ln \frac{\pi_1}{c(G_1|G_2)} > -\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) + \ln \frac{\pi_2}{c(G_2|G_1)} \tag{2.4}$$

Llamando D_i^2 a la distancia de Mahalanobis entre el individuo a clasificar y la media de la población i, para $i = 1, 2$, es decir:

$$D_i^2 = (x - \mu_i)' \Sigma^{-1} (x - \mu_i). \tag{2.5}$$

La ecuación (2.4) se puede escribir como:

$$\frac{1}{2} D_2^2 - \ln \frac{\pi_2}{c(G_2|G_1)} > \frac{1}{2} D_1^2 - \ln \frac{\pi_1}{c(G_1|G_2)} \tag{2.6}$$

Caso particular: Si las probabilidades a priori de pertenecer a un grupo son iguales, osea, $\pi_1 = \pi_2$, y los costes de penalización también son iguales, es decir, $c(G_1|G_2) = c(G_2|G_1)$, entonces la ecuación (2.6) se quedaría:

$$D_2^2 > D_1^2.$$

Por consecuente, se clasificaría el elemento en G_1 si $D_2^2 > D_1^2$.

2.1.2.1. Cálculo de probabilidades de error

La utilidad de la regla de clasificación depende de los errores esperados, por lo que buscamos calcular la probabilidad de error de clasificar cada una de las observaciones en una de las dos poblaciones para establecer la probabilidad de tomar una decisión errónea. Para ello, consideremos la variable $z = w'X$ donde $w = \Sigma^{-1}(\mu_2 - \mu_1)$. Puesto que z transforma la variable multivariante X que sigue una distribución normal en una combinación lineal de los valores con coeficientes dados por el vector w, entonces z es una variable unidimensional con distribución normal de media $m_i = w'\mu_i$ y varianza $\sigma^2 = (m_2 - m_1)^2$. Luego, la probabilidad de una decisión errónea cuando $x \in G_2$ es:

$$P(G_1/G_2) = P\{z \geq \frac{m_1 + m_2}{2} \ / \ z \sim N(m_2, \sigma)\}$$

Con la idea de buscar una distribución normal con media 0 y varianza 1, tipificamos la variable z tomando $y = (z - m_2)/\sigma$ y quedaría:

$$P(G_1/G_2) = P\{y \geq \frac{\frac{m_1+m_2}{2} - m_2}{\sigma}\} = 1 - \Phi(\frac{\sigma}{2}) \tag{2.7}$$

donde Φ es la función de distribución de una $N(0, 1)$.

De la misma manera, se calcula la probabilidad de una decisión errónea cuando $x \in G_1$ es:

$$P(G_2/G_1) = P\{z \leq \frac{m_1 + m_2}{2} \ / \ z \sim N(m_1, \sigma)\} = P\{y \leq \frac{\frac{m_1+m_2}{2} - m_1}{\sigma}\} = \Phi(\frac{-\sigma}{2})$$

2.1.2.2. Probabilidades a posteriori

Una vez realizado el análisis, calculamos las probabilidades a posteriori para establecer el grado de confianza al clasificar un nuevo individuo.
La probabilidad a posteriori de que un nuevo individuo pertenezca al grupo 1 viene dada por:

$$P(G_1/x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}$$

Sustituimos el valor de f_i, $\forall i = 1, 2$, dado por (2.3) y quedaría:

$$P(G_1/x) = \frac{\pi_1\, exp(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)}{\pi_1\, exp(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \pi_2\, exp(-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)}$$

De nuevo, llamando a D_1^2 y D_2^2 a la distancia de Mahalanobis (2.5), obtenemos que:

$$P(G_1/x) = \frac{\pi_1 D_1^2}{\pi_1 D_1^2 + \pi_2 D_2^2} = \frac{1}{1 + \frac{\pi_2}{\pi_1} exp(-\frac{1}{2}(D_2^2 - D_1^2))}. \tag{2.8}$$

Dicha probabilidad solo depende de las probabilidades a priori y de las distancias del punto a la media de cada población.

Caso particular: Si las probabilidades a priori de pertenecer a un grupo son iguales, osea, $\pi_1 = \pi_2$, entonces cuanto más alejado esté el nuevo individuo de la primera población, es decir, cuanta más diferencia haya entre entre D_1^2 y D_2^2, menor será la probabilidad de que el nuevo individuo x pertenezca a G_1, ya que el denominador será mayor.

2.1.2.3. Ejemplo

En el siguiente ejemplo extraído de Peña [5] vamos a clasificar un retrato entre dos posibles pintores. Para ello, medimos dos variables, la primera representa la profundidad del trazo de la pintura y la segunda, la proporción que ocupa el retrato sobre la superficie del lienzo.

El pintor A presenta como medias de estas variables 2 y 0.8, respectivamente y el pintor B, sus medias son 2.3 y 0.7. Las desviaciones típicas de estas variables en ambos pintores son: 0.5 y 0.1, respectivamente y la correlación entre ellas es 0.5.

Nuestro objetivo es indicar a qué pintor pertenecería la obra que tiene como medidas 2.1 y 0.75, respectivamente.

En primer lugar, calculamos las distancias de Mahalanobis de cada grupo. Para ello, calculamos la matriz Σ, donde la covarianza es el producto de la correlación por las desviaciones típicas, es decir:

$$\Sigma = \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}$$

La distancia de Mahalanobis para el grupo A es:

$$D_A^2 = \begin{pmatrix} 2.1-2 & 0.75-0.8 \end{pmatrix} \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.1-2 \\ 0.75-0.8 \end{pmatrix} = 0.52.$$

Y la del pintor B es:

$$D_B^2 = \begin{pmatrix} 2.1-2.3 & 0.75-0.7 \end{pmatrix} \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.1-2.3 \\ 0.75-0.7 \end{pmatrix} = 0.8133.$$

Por tanto, puesto que $D_B^2 > D_A^2$, la obra se la atribuimos al pintor A.

Ahora, obtengamos el error esperado de clasificación si clasificamos la nueva obra en el pintor A. Para ello, calculemos la distancia entre las medias:

$$\sigma^2 = \begin{pmatrix} 2-2.3 & 0.8-0.7 \end{pmatrix} \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2-2.3 \\ 0.8-0.7 \end{pmatrix} = 2.61333.$$

Luego, $\sigma = 1.6166$ y la probabilidad de error (2.7) es:

$$P(A/B) = 1 - \Phi(\frac{1.6166}{2}) = 1 - 0.7905 = 0.2095.$$

Luego, podemos tener un 20.95 % de error.

Por último, calculamos la probabilidad a posteriori (2.8) de que la obra pertenezca al pintor A en el caso particular de que las probabilidades a priori son iguales, es decir, $\frac{\pi_2}{\pi_1} = 1$.

$$P(A/x) = \frac{1}{1 + exp(-\frac{1}{2}(0.8133 - 0.52))} = 0.5365.$$

Esta probabilidad indica que el grado de confianza de que la obra pertenezca al pintor A es muy baja puesto que la probabilidad de que pertenezca al grupo B es de $1 - 0.5365 = 0.4635$, es decir, ambas son muy semejantes.

2.2. Clasificación en más de dos grupos

2.2.1. Función discriminante de Bayes bajo normalidad

Partiendo de la idea del epígrafe 2.1 generalizamos para el caso de k grupos: $G_1, G_2, \ldots, G_i, \ldots, G_k$. Dividimos el espacio muestral E_x, que en la mayoría de los casos será $\mathbb{R}^p$ en k regiones: $e_1, e_2, \ldots, e_k$ de manera que si el nuevo indviduo x pertenece a la región e_i, entonces será clasificado en el grupo G_i. Además, supondremos que los costes son constantes e independientes de la población en la que se clasifique.

Entonces, la región e_i vendrá dada por:

$$e_i = \{x \in \mathbb{R}^p / \pi_i f_i(x) > \pi_j f_j(x); \forall i \neq j\}. \tag{2.9}$$

Considerando el caso en el que las probabilidades a priori son iguales, π_j, $\forall j = 1, \ldots, k$, y las distribuciones $f_i(x)$ son normales con la misma matriz de varianzas-covarianzas, la condición de pertenecer a la región (2.9) equivale a calcular la distancia de Mahalanobis. Por tanto, clasificaremos el nuevo individuo en el grupo en el que dicha distancia sea mínima.

Además, eliminamos de (2.5) el término $x'\Sigma^{-1}x$ que es común en todos los grupos, $\mu_i'\Sigma x = x'\Sigma\mu_i$ ya que Σ es simétrica y nos queda que la ecuación lineal a minimizar es:

$$L_i = -2\mu_i'\Sigma^{-1}x + \mu_i'\Sigma^{-1}\mu_i. \tag{2.10}$$

Llamando $w_i = \Sigma^{-1}\mu_i$, la regla es $\min_i(w_i'\mu_i - 2w_i'x)$.

Una propiedad a destacar de la regla de decisión es que cumple la propiedad transitiva.

Por otro lado, la frontera de separación entre dos poblaciones G_i y G_j viene dada:

$$C_{ij}(x) = L_i(x) - L_j(x) = 0.$$

Sustituyendo los valores de la ecuación (2.10) y agrupando términos obtenemos:

$$C_{ij}(x) = -2(\mu_i - \mu_j)'\Sigma^{-1}x + (\mu_i - \mu_j)'\Sigma(\mu_i + \mu_j) = 0.$$

Llamando $w_{ij} = \Sigma^{-1}(\mu_i - \mu_j)$, establecemos como frontera de separación

$$C_{ij}(x) = 2w_{ij}'x + w_{ij}'(\mu_i + \mu_j) = 0 \Rightarrow w_{ij}'x = \frac{1}{2}w_{ij}'(\mu_i + \mu_j)$$

A partir de ello, se obtiene que la variable indicadora entre dos grupos G_i y G_j es

$$z = (\mu_i - \mu_j)'\Sigma^{-1}x. \tag{2.11}$$

2.2.2. Clasificación

Por tanto, para llevar a cabo un análisis discriminante de Bayes con $k > 2$ grupos, se puede realizar de dos maneras distintas: la primera, es calcular las distancias de Mahalanobis para cada grupo y clasificar el nuevo elemento en el grupo en el que la distancia sea mínima y la segunda forma, hacer el análisis comparando los grupos de dos en dos.

Para la segunda forma, primero, se calcula el valor medio de la variable indicadora en cada grupo. A continuación, se establece la frontera de separación entre ambos grupos. Por último, se clasificará el nuevo elemento en el grupo cuyo valor indicador esté más próximo.

Veamos un caso específico suponiendo 5 grupos con $p > 4$ variables indepedientes, por lo que habrá 4 reglas de clasificación independientes y las demás se deducen de ellas.

Comparamos la pertenencia del nuevo individuo de dos en dos grupos, calculando la variable indicadora (2.11) entre ambos grupos. De esta expresión obtendríamos la clasificación del nuevo individuo en el grupo cuyo valor indicador esté más próximo. De esta manera, se clasificaría el nuevo individuo en el grupo i antes que en el grupo j, es decir, $G_i > G_j$.

Repitiendo el proceso, obtenemos, por ejemplo, que:

$$G_1 > G_2$$
$$G_2 > G_3$$
$$G_4 > G_3$$
$$G_5 > G_4$$

Por tanto, puesto que se clasificaría antes en los grupos 1 y 5 que en los grupos $2, 3, 4$, el problema se reduce a clasificar el elemento entre ambos grupos aplicándose lo estudiado en 2.1.

2.2.3. Ejemplo

En el siguiente ejemplo extraído de Peña [5], se considera una máquina que admite monedas. Esta lleva a cabo 3 tipos de mediciones: X_1 (peso), X_2 (espesor) y X_3 (densidad de estrías en su canto) para establecer su valor comparándolas con 3 clases diferentes de monedas: M_1, M_2 y M_3.

La distribución del peso, el espesor y la densidad de estas monedas es de tipo normal, con vectores de medias :

$$\begin{cases} \mu_1 = (20, 8, 8)' \\ \mu_2 = (19.5, 7.8, 10)' \\ \mu_3 = (20.5, 8.3, 5)' \end{cases}$$

y matriz de varianzas-covarianzas:

$$\Sigma = \begin{pmatrix} 4 & 0.8 & -5 \\ 0.8 & 0.25 & -0.9 \\ -5 & -0.9 & 9 \end{pmatrix}$$

En primer lugar, se quiere indicar qué tipo de menoda es el más adecuado para una de medidas $(x_1, x_2, x_3) = (22, 8.5, 7)'$.

Para ello, se puede realizar de dos formas distintas. La primera es calcular las distancias de Mahalanobis (2.5) y asignarla al grupo que menor distancia produzca. Así pues, calculamos dichas distancias:

$$D_1^2 = (x - \mu_1)'\Sigma^{-1}(x - \mu_1) = \begin{pmatrix} 22-20 & 8.5-8 & 7-8 \end{pmatrix} \begin{pmatrix} 4 & 0.8 & -5 \\ 0.8 & 0.25 & -0.9 \\ -5 & -0.9 & 9 \end{pmatrix}^{-1} \begin{pmatrix} 22-20 \\ 8.5-8 \\ 7-8 \end{pmatrix} = 1.8421.$$

De igual forma, obtendríamos que $D_2^2 = 2.01$ y $D_3^2 = 6.69$, por lo que deberíamos de asignar dicha moneda al primer grupo.

Para la segunda forma, calculamos las variables indicadoras (2.11) entre los grupos.

En primer lugar, calculamos la variable indicadora para clasificar entre M_1 y M_2 que es:

$$z_1 = (\mu_1 - \mu_2)'\Sigma^{-1}x = -0.93x_1 + 1.74x_2 - 0.56x_3.$$

La media de esta variable para la primera moneda es $-0.93 \times 20 + 1.74 \times 8 - 0.56 \times 8 = -9.16$ y para la segunda es $-0.93 \times 19.5 + 1.74 \times 7.8 - 0.56 \times 10 = -10.163$, luego el punto de corte es -9.6615. Y como para la moneda que queremos clasificar se tiene que $-0.93 \times 22 + 1.74 \times 8.5 - 0.56 \times 7 = -9.59$, tendríamos que $G_1 > G_2$.

De la misma forma, relizaremos la variable indicadora entre clasificar M_2 y M_3.

$$z_2 = (\mu_2 - \mu_3)'\Sigma^{-1}x = 2.6947x_1 - 5.0536x_2 + 1.5473x_3.$$

La media de esta variable para la segunda moneda es $2.6947 \times 19.5 - 5.0536 \times 7.8 + 1.5473 \times 10 = 28.6016$ y para la tercera es $2.6947 \times 20.5 - 5.0536 \times 8.3 + 1.5473 \times 5 = 21.033$, luego el punto de corte es 24.8173. Y como para la moneda a clasificar se tiene que $2.6947 \times 22 - 5.0536 \times 8.5 + 1.5473 \times 5 = 17.8751$, tendríamos que $G_3 > G_2$.

Por lo que ahora, tenemos que estudiar la variable indicadora entre M_1 y M_3:

$$z_3 = (\mu_1 - \mu_3)'\Sigma^{-1}x = 1.77x_1 - 3.31x_2 + 0.98x_3.$$

La media de esta variable para la primera moneda es $1.77 \times 20 - 3.31 \times 8 + 0.98 \times 8 = 16.71$ y para la tercera es $1.77 \times 20.5 - 3.31 \times 8.3 + 0.98 \times 5 = 13.65$, luego el punto de corte es 15.17. Y como para la moneda a clasificar se tiene que $1.77 \times 22 - 3.31 \times 8.5 + 0.98 \times 7 = 17.61$, tendríamos que $G_1 > G_3$.

Por tanto, se clasificaría en M_1, luego en M_3 y finalmente en M_2.

En segundo lugar, calculamos las probabilidades a posteriori considerando las probabilidades a priori son iguales, de igual forma que para el caso de dos grupos 2.1.2.2, es decir:

$$P(M_i/x) = \frac{exp(-\frac{1}{2}D_i^2)}{exp(-\frac{1}{2}D_1^2) + exp(-\frac{1}{2}D_2^2) + exp(-\frac{1}{2}D_3^2)}.$$

Luego, en nuestro caso, sustituyendo las respectivas distancias de Mahalanobis calculadas anteriormente, las probabilidades a posteriori quedarían

$$\begin{cases} P(M_1/x) = 0.5 \\ P(M_2/x) = 0.46 \\ P(M_3/x) = 0.4 \end{cases}$$

3 Aplicación con datos reales en R

> En este capítulo, realizaremos una aplicación con datos reales de la técnica del análisis discriminante lineal de Fisher con el software estadístico R. Todas las funciones utilizadas en dicho programa se pueden encontrar en la página web de R [6].

Los datos proceden el proyecto GENEIDA (*Genetics, Early Life Environmental Exposures and Infant Development in Andalucía*), el cual tiene como objetivo estudiar el efecto que produce la exposición de la mujer durante el embarazo a contaminantes ambientales en el crecimiento del feto.

El proyecto GENEIDA se desarrolla en el Hospital de Poniente de la provincia de Almería. Se incluyen en el estudio aquellas mujeres embarazadas que acuden a dicho hospital para el programa de control del embarazo entre la semana 12 y 14 de gestación. Además, tienen que ser mayores de 16, con embarazo único y sin uso de técnicas de reproducción asistida. Tampoco se tendrán en cuenta aquellas mujeres que tengan dificultad en la compresión del idioma castellano o que hayan sido diagnosticadas clínicamente de una enfermedad crónica con carácter previo al embarazo y que estén bajo tratamiento médico.

La información se obtuvo mediante cuestionarios sobre las exposiciones ambientales y ocupacionales, estilo de vida, dieta, etc, así como ecografías en las visitas de seguimiento del embarazo de la mujer en el hospital.

3.1. Análisis de 2 grupos

Nuestro objetivo es encontrar una función discriminante que nos permita distinguir, a partir de los datos de ciertas variables de la base bajo estudio, entre los embarazos que terminaron en un parto natural o en un parto con ayuda, entiendo estos últimos como los que acaban con cesárea o con algún procedimiento quirúrgico como puede ser el uso de ventosa, fórceps o espátula.

En primer lugar hemos considerado la variable categórica *grupos*2. Esta variable indica si el parto ha sido con ayuda (grupo 1) o natural (grupo 2). Por otro lado, para discriminar ambos grupos, se dispone de 15 variables en la base bajo estudio. Dichas variables son:

- *Edad*: Edad de la mujer.
- *Peso_mujer*: Peso de la mujer antes del embarazo en kg.
- *talla_madre_mt*: Altura de la madre en metros.
- *Té_T*3: Cantidad de té ingerida al día por la mujer en el tercer trimestre. Una taza de té son 50 cc.

- *Alcohol_T3*: Cantidad de alcohol en g al día en el tercer trimestre.
- *Café_T3*: Cantidad de café en g ingerido al día en el tercer trimestre.
- *VitB12_T3*: Ingesta total de vitamina B12 medida en μg/día en el tercer trimestre.
- *Calcio_T3*: Cantidad de calcio en mg en el tercer trimestre.
- *BMI_preembarazo*: Índice de masa corporal de la mujer antes del embarazo.
- *BMI_postembarazo*: Índice de masa corporal de la mujer después del embarazo.
- Medidas ecográficas del embarazo más comunes:
  - *BDP_39*: Diámetro biparietal del feto a las 39 semanas en mm.
  - *CA_39*: Circunferencia abdominal del feto a las 39 semanas en mm.
  - *CC_39*: Circunferencia cefálica del feto a las 39 semanas en mm.
  - *PFE_39*: Peso fetal estimado del feto a las 39 semanas en mm.
- *Edad_gest_numerico*: Edad gestacional del feto en semanas al nacimiento.

Ahora vamos a estudiar estas variables con idea de comprobar si están incorreladas entre sí para poder aplicar correctamente la técnica.

Empezamos cargando los datos en R y, para ello, leemos los datos almacenados en una hoja de texto a partir de la función *read.table* y le indicamos que la primera fila corresponde al nombre de las variables con el argumento *header=TRUE*. Después lo guardamos en la variable *datos*.

```
datos <- read.table( file="datos2grupos.txt", header=TRUE)
```

A continuación analizamos la incorrelación entre dichas variables. Para ello utilizamos la función *cor* que sirve para establecer la relación existente entre cada par de variables calculando el coeficiente de correlación de Pearson. Esta función tiene los siguientes argumentos:

```
cor(x,y,method="pearson")
   x es un vector numérico, matriz o data.frame
   y es un vector numérico, matriz o data.frame de la misma dimesión que x
   method="pearson" calcula el coeficiente de correlación de Pearson
```

Aplicamos dicha función a nuestras variables.

```
correlaciones=cor(datos[,1:ncol(datos)],method="pearson")
```

Y representamos gráficamente su resultado usando la función *corrplot* que se puede encontrar en la librería *corrplot* de R.

```
library(corrplot)
col4 <- colorRampPalette(c("#7F0000", "red", "#FF7F00", "yellow", "#7FFF7F",
                           "cyan", "#007FFF", "blue", "#00007F"))
corrplot(correlaciones,method = 'color',type = "upper",col = col4(50), diag=FALSE)
```

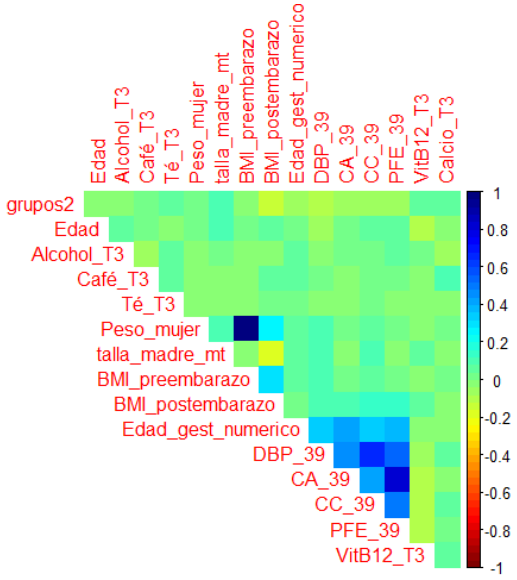

Figura 3.1: Gráfico de correlación entre las variables.

En la Figura 3.1 se puede apreciar que existe una fuerte correlación entre las variables: *Peso_mujer* y *BMI_preembarazo*. Además, también muestran una fuerte asociación positiva todas las medidas ecográficas entre sí, es decir, *PFE_39*, *CA_39*, *CC_39* y *DBP_39*. Por último, la variable *Edad_gest_numerico* también está fuertemente correlacionada con las medidas ecográficas.

3.1.1. Implementación propia: método de selección de variables

A continuación llevamos a cabo un método de selección de variables (método paso a paso) para obtener aquellas que mayor discriminación aporten a la hora de aplicar la técnica. Para ello hemos implementado una función en R.

Para comenzar con el método de selección de variables se determina el valor del estadístico F (1.14), proporcionado por la salida *svd* al cuadrado de la función *lda*, para cada una de las 15 variables bajo estudio, una a una, con respecto a la variable categórica *grupos2* y se elige aquella variable cuyo valor de F sea mayor. A continuación, se calcula el estadístico F de dicha variable con todas las demás y se elige la pareja cuyo valor sea mayor. Por consiguiente, se realiza este proceso de manera iterativa, añadiendo en cada paso aquella variable que junto con las variables anteriores seleccionadas produzca un mayor valor del estadístico hasta que la diferencia entre el valor máximo de F del modelo actual y el valor máximo de F del modelo anterior sea menor que 3, cota establecida en la mayoría de los softwares estadísticos.

Previo a la implementación de nuestra función veamos la sintáxis de la función *lda* en R:

```
lda(fórmula,data,prior,tol)
  fórmula: variable categórica~X1+X2+... siendo Xi las variables discriminantes.
  data: data.frame a partir del cuál se extraen las variables discriminantes usadas
    en la fórmula.
```

```
  prior: son las probabilidades de pertenencia a cada grupo. Si se indican se harán
    en el mismo orden de los grupos. En caso contrario, serán proporcionales al tama
    ño de los grupos.
  tol: valor de tolerancia para decidir si una variable discriminate forma parte del
    modelo.
```

El código de nuestra función es:

```
  library(MASS)

  nvariables<-ncol(datos)-1
  Fest<-numeric()
  velegida<-numeric()
  valorelegida<-numeric()

  for(i in 1:nvariables)
  {
    Fest[i]<-(lda(datos[,1]~datos[,i+1],tol=0.001)$svd)^2
  }

  velegida[1]<-which.max(Fest)
  valorelegida[1]<-Fest[which.max(Fest)]

    for(j in 2:n)
  {
    ncolelegidas<-(velegida[1:(j-1)]+1)
    x<-paste0("datos[,",ncolelegidas[1:(j-1)],"]")

    for(i in 1:n)
    {
      formula<-formula<-as.formula(paste("datos[,1]~ datos[,i+1]+", paste(x,collapse=
    "+")))
      ifelse(any(i==velegida[1:(j-1)]),Fest[i]<-NA,Fest[i]<-(lda(formula,tol=0.001)$
    svd)^2)
    }

    velegida[j]<-which.max(Fest)
    valorelegida[j]<-Fest[which.max(Fest)]


    if(abs((valorelegida[j]-valorelegida[j-1]))<3)
    {
      velegida<-velegida[-j]
      break
    }
  }

datos<-datos[,c(1,ncolelegidas)]
```

Como resultado obtenemos que las variables seleccionadas por nuestro método son *talla_madre_mt, BMI_postembarazo* y *DBP_39*.

Una vez seleccionadas las variables aplicamos la técnica del análisis discriminante. Pero, para un uso correcto de la técnica han de cumplirse los supuestos 1.3.

3.1.2. Comprobación de los supuestos

En primer lugar, la variable *grupos*2 es categórica y las variables discriminantes seleccionadas, *talla_madre_mt*, *BMI_postembarazo* y *DBP_*39, están medidas en una escala de intervalo o razón. Además, en cada grupo hay más de un individuo ya que tenemos que $n_1 = 206$ y $n_2 = 244$, y cada individuo solo pertenece a uno de los dos grupos.

En la figura 3.1 se ve evidente que la correlación entre las variables *talla_madre_mt*, *BMI_postembarazo* y *DBP_*39 es casi nula. Por lo que no habría problemas de multicolinealidad entre ellas.

Continuamos comprobando la normalidad en cada grupo.

Almacenamos en *datos*1 los invididuos que corresponde al grupo 1 (parto con ayuda) y en *datos*2 aquellos que corresponde al grupo 2 (parto natural) con la función *subset* cuyos argumentos principales son:

```
subset(x,condition)
  x: objeto del que queremos extraer un subconjunto.
  condition: expresión lógica que indica aquellos elementos que se quieren conservar
    eliminado el resto.
```

En nuestro caso,

```
datos1<-subset(datos, datos$grupos2==1)
datos2<-subset(datos, datos$grupos2==2)
```

Veamos gráficamente la normalidad de cada variable mediante los gráficos Q-Q plot estudiados en 1.3.1.1 con las funciones *qqnorm* y *qqline*.

```
par(mfcol = c(2,3))
for (k in 2:4) {
  j0 <- names(datos)[k]
  x0 <- seq(min(datos[, k]), max(datos[, k]), le = 50)
  for (i in 1:2) {
    i0 <- levels(datos$grupos2)[i]
    x <- datos[datos$grupos2 == i0, j0]
    if(i==1) col<-"blue" else col<-"magenta"
    qqnorm(x, main = paste("Grupo",i0,":", j0), pch = 19, col =col)
    qqline(x)
  }
}
```

En base a los gráficos de la Figura 3.2 las variables *talla_madre_mt* y *DBP_*39 siguen una distribución normal. Sin embargo, la variable *BMI_postembarazo* presenta algunas irregularidades que descartan la normalidad. Para corregir dicho problema podemos aplicar transformaciones logarítmicas como el logarítmo neperiano, el logarítmo en base 10 o en base 2. Aun así, la variable *BMI_postembarazo* sigue sin regirse por una distribución normal.

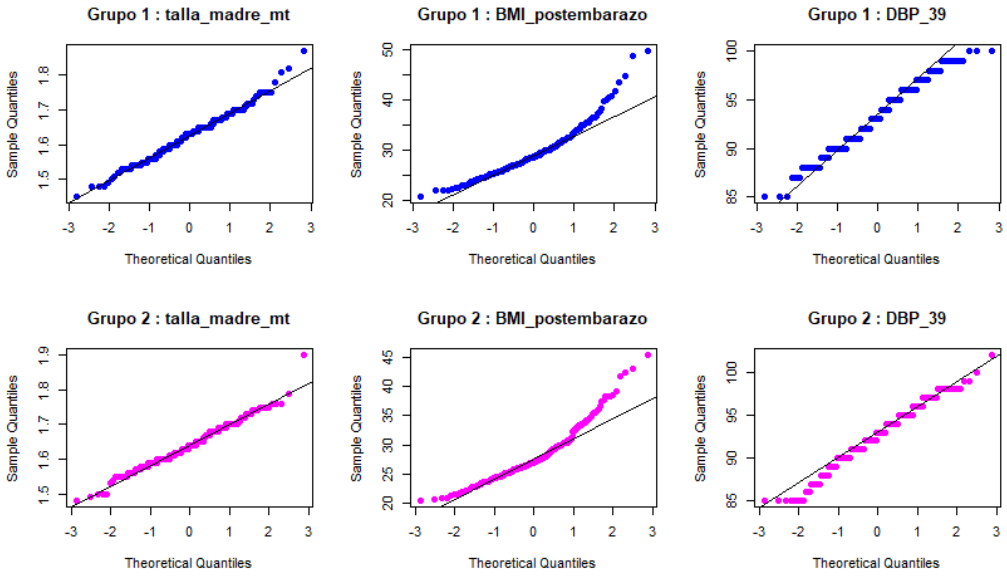

Figura 3.2: Gráficos Q-Q plot

Continuamos con el análisis aunque teniendo en cuenta que esa variable no cumple la normalidad. Esto se puede corroborar usando el test de Kolmogorov-Smirnov 1.3.1.2. Para ello, utilizamos la función *ks.test* que tiene los siguientes argumentos.

```
ks.test(x,y,...)
 x: vector numérico de valores de datos
 y: función de distribución acumulativa real. En este caso, "pnorm".
 ...: parámetros de la distribución especificada por y.
```

Los resultados de los test son

- Grupo 1.

```
ks.test( x=datos1$talla_madre_mt,"pnorm", mean(datos1$talla_madre_mt),sd(datos1
    $talla_madre_mt))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos1$talla_madre_mt
## D = 0.064581, p-value = 0.3567
## alternative hypothesis: two-

ks.test(x=datos1$BMI_postembarazo,"pnorm",mean(datos1$BMI_postembarazo),sd(
    datos1$BMI_postembarazo))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos1$BMI_postembarazo
## D = 0.097824, p-value = 0.03879
## alternative hypothesis: two-sided

ks.test(x=datos1$DBP_39,"pnorm",mean(datos1$DBP_39),sd(datos1$DBP_39))

##  One-sample Kolmogorov-Smirnov test
```

```
##
## data:  datos1$DBP_39
## D = 0.094377, p-value = 0.05097
## alternative hypothesis: two-sided
```

- Grupo 2.

```
ks.test( x=datos2$talla_madre_mt,"pnorm",mean(datos2$talla_madre_mt), sd(datos2
    $talla_madre_mt))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$talla_madre_mt
## D = 0.068201, p-value = 0.2064
## alternative hypothesis: two-sided

ks.test(x=datos2$BMI_postembarazo,"pnorm",mean(datos2$BMI_postembarazo),sd(
    datos2$BMI_postembarazo))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$BMI_postembarazo
## D = 0.11547, p-value = 0.002986
## alternative hypothesis: two-sided

ks.test(x=datos2$DBP_39,"pnorm",mean(datos2$DBP_39), sd(datos2$DBP_39))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$DBP_39
## D = 0.08355, p-value = 0.06631
## alternative hypothesis: two-sided
```

Por último validamos que se cumpla la igualdad de las matrices de varianzas-covarianzas. Para ello aplicamos la prueba M de Box (1.3.2) usando la función *boxM* del paquete de R *biotools*.

Dicha función tiene la siguiente sintáxis:

```
boxM(data,grouping)
  data: data.frame o matriz que contiene los n individuos de las p variables.
  grouping: vector de longitud n que contiene la pertenencia de cada individuo a su
   grupo correspondiente.
```

Luego, aplicándola a nuestros datos obtenemos:

```
library(biotools)
boxM(data=datos[2:4],grouping = datos$grupos2)

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  datos[2:4]
## Chi-Sq (approx.) = 9.1235, df = 6, p-value = 0.1668
```

Como se puede apreciar, el p-valor del test es superior a 0.15, por lo que no hay evidencias de rechazar la hipótesis de igualdad de matrices de varianzas-covarianzas.

3.1.3. Aplicación de la técnica

Una vez comprobados los supuestos, nuestro objetivo es aplicar la técnica del análisis discriminante lineal. Pero, previamente, veamos como discriminan nuestras variables por separado con un diagrama de dispersión utilizando la función *pairs*. El color azul representa a los individuos del grupo 1 y el color magenta a los del grupo 2.

```
pairs(x=datos[,2:4],col=c("blue","magenta")[datos$grupos2],pch=16)
```

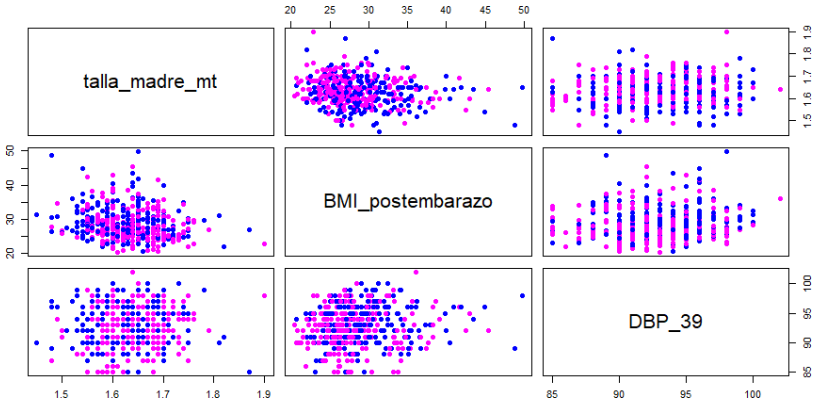

Figura 3.3: Diagrama de dispersión de las tres variables discriminantes.

En los gráficos de la Figura 3.3 se aprecia que ninguna de las variables discrimina totalmente entre los 2 grupos. Es por ello que usaremos la técnica para intentar encontrar una función de las variables seleccionadas que discrimine entre ambos grupos. Para ello, usamos la función *lda* del paquete *MASS* de R.

Esta función devuelve

```
Call: fórmula que se ha considerado.
Prior probabilities of groups: probabilidad a priori de pertenencia a cada grupo.
Group means: medias de cada grupo en cada variable.
Coefficients of linear discriminants: coeficientes de la función lineal
   discriminante de Fisher.
```

En nuestro caso, tomamos como variable categórica *grupos*2 y como variables independientes: *talla_madre_mt*,*BMI_postembarazo* y *DBP*_39. Guardamos el resultado en la variable *lda* y obtenemos:

```
lda<-lda(grupos2~talla_madre_mt+BMI_postembarazo+DBP_39,data=datos)
lda

##
## Call:
## lda(grupos2 ~ talla_madre_mt + BMI_postembarazo + DBP_39, data = datosutiles,
## tol = 1e-04)
##
## Prior probabilities of groups:
##         1         2
## 0.4577778 0.5422222
```

```
##
## Group means:
##  talla_madre_mt BMI_postembarazo  DBP_39
## 1      1.627184      29.44191    93.26699
## 2      1.640697      27.99150    92.63115
##
## Coefficients of linear discriminants:
##                         LD1
## talla_madre_mt    7.4720163
## BMI_postembarazo -0.1526478
## DBP_39           -0.1415996
```

Luego se tiene que la función lineal discriminante de Fisher viene dada por:

$$D = 7.4720163 \cdot talla\_madre\_mt - 0.1526478 \cdot BMI\_postembarazo - 0.1415996 \cdot DBP\_39.$$

Una vez construída nuestra función, podemos hacer predicciones a partir de los resultados obtenidos por la función *lda*. Para ello usamos la función *predict*, cuya sintáxis es:

```
predict(x,new.data)
 x: objeto para el que se desea la predición.
 new.data: data.frame con los individuos a clasificar.
```

La función *predict* devuelve:

```
class: vector de dimensión n que devuelve el resultado de la clasificación.
posterior: devuelve las probabilidades a posteriori.
```

La aplicamos a nuestra variable *lda* y la guardamos en *clasificación*. Previamente tenemos que cargar y leer el paquete *stat*.

```
library(stat)

clasificacion<-predict(lda,datos[2:4])
```

En la Figura 3.4 se pueden apreciar las puntuaciones discriminantes de cada grupo.

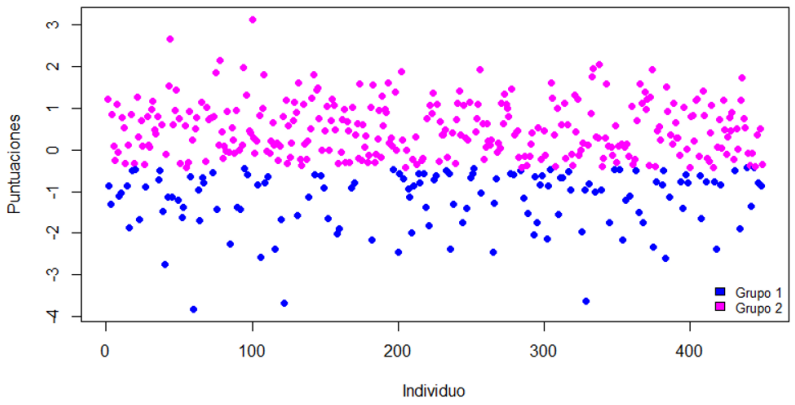

Figura 3.4: Puntuaciones discriminantes de cada grupo

A continuación, podemos ver con la función *confusionmatrix* del paquete *biotools* el porcentaje de acierto de la clasificación.

La sintáxis de dicha función es:

```
confusionmatrix(data,reference)
  data: un factor de clases predichas.
  reference: un factor de clases para ser utilizado como los verdaderos resultados.
```

Luego, adaptándonos a nuestros datos tenemos

```
aciertoclasificacion<-confusionmatrix(datos$grupos2,clasificacion$class)
aciertoclasificacion

##
##       1   2
##   1  87 119
##   2  47 197
```

En la tabla se puede apreciar que de los 206 individuos que hay en el grupo 1 han sido clasificados correctamente 87 y han sido mal clasificados 119. En cambio, en el grupo 2, de los 244 individuos han sido clasificados correctamente 197 y erróneamente 47.

A partir de ella, obtenemos que la probabilidad de acertar de cada grupo viene dado por:

```
diag(prop.table(aciertoclasificacion,1))

##
##         1         2
## 0.4223301 0.8073770
```

Luego, la probabilidad de aciertos al clasificar un individuo en el grupo 1 es de 42.23 % y en el grupo 2 es de 80.73 %. Así pues, la probabilidad de acertar del modelo es 63.11 %.

En conclusión, partiendo de la base de que cada variable por separado mostraba una zona de confusión entre ambos grupos muy grande, la función discriminante lineal de Fisher discrimina en un 63.11 % de los casos, aunque tiene más tendencia a discriminar mejor a los individuos del grupo 2 que a los del grupo 1.

3.2. Análisis de 3 grupos

Nuestro objetivo ahora es encontrar funciones discriminantes que nos permitan distinguir entre los embarazos que terminaron en un parto con cesárea, en un parto vaginal o en un parto eutócico o natural. Al igual que en el caso de 2 grupos partimos de 450 individuos bajo estudio y 15 variables en la base.

En primer lugar hemos considerado la variable categórica *grupos*3. Esta variable indica el tipo de parto que se ha producido, donde el grupo 1 corresponde a un parto por cesárea, el grupo 2 a uno vaginal y el grupo 3 a un parto eutócico.

Leemos los datos y los guardamos en la variable *datos*.

```
datos <- read.table( file="datos3grupos.txt", header=TRUE)
```

3.2.1. Implementación propia: método de selección de variables

A continuación llevaremos a cabo de nuevo el método implementado en R para la selección de las variables más significativas. En este caso, puesto que tenemos menos datos de cada grupo aumentamos la restricción del criterio de parada a 1 ya que en caso contrario solo entraría la variable *BMI_postembarazo* y buscamos una función discriminante más completa.

```
  nvariables<-ncol(datos)-1
  Fest<-numeric()
  velegida<-numeric()
  valorelegida<-numeric()

  for(i in 1:nvariables)
  {
    Fest[i]<-(lda(datos[,1]~datos[,i+1])$svd)^2
  }

  velegida[1]<-which.max(Fest)
  valorelegida[1]<-Fest[which.max(Fest)]

    for(j in 2:n)
    {
    ncolelegidas<-(velegida[1:(j-1)]+1)
    x<-paste0("datos[,",ncolelegidas[1:(j-1)],"]")

    for(i in 1:n)
    {
      formula<-formula<-as.formula(paste("datos[,1]~ datos[,i+1]+", paste(x,collapse=
    "+")))
      ifelse(any(i==velegida[1:(j-1)]),Fest[i]<-NA,Fest[i]<-(lda(formula)$svd)^2)
    }

    velegida[j]<-which.max(Fest)
    valorelegida[j]<-Fest[which.max(Fest)]


    if(abs((valorelegida[j]-valorelegida[j-1]))<1)
    {
      velegida<-velegida[-j]
      break
    }
  }


datos<-datos[,c(1,ncolelegidas[1:length(ncolelegidas)]]
```

Al aplicar la técnica obtenemos que debemos seleccionar las variables *talla_madre_mt*, *BMI_postembarazo*, *DBP_3*9 y *Calcio_T*3.

Una vez seleccionadas las variables bajo estudio aplicamos la técnica del análisis discriminante. Pero, para un uso correcto de la técnica han de cumplirse los supuestos 1.3.

3.2.2. Comprobación de los supuestos

En primer lugar, la variable *grupos*3 es categórica y las variables discriminantes *talla_madre_mt*, *BMI_postembarazo*, *DBP_*39 y *Calcio_T*3 están medidas en una escala de intervalo o razón.

En este caso, trabajamos con 3 grupos y en cada uno de ellos hay más de un individuo ya que tenemos que $n_1 = 142$, $n_2 = 64$ y $n_3 = 244$.

Veamos que no hay correlación entre las variables *talla_madre_mt*, *BMI_postembarazo*, *DBP_*39 y *Calcio_T*3. Para ello utilizamos la función *cor*.

```
correlaciones=cor(datos[,2:ncol(datos)],method="pearson")

correlaciones


##                   talla_madre_mt  BMI_postembarazo  DBP_39    Calcio_T3
## talla_madre_mt       1.00000000       -0.16951364    0.08548206 -0.02832612
## BMI_postembarazo    -0.16951364        1.00000000    0.09131056 -0.01638763
## DBP_39               0.08548206        0.09131056    1.00000000  0.05512000
## Calcio_T3           -0.02832612       -0.01638763    0.05512000  1.00000000
```

En la matriz de correlaciones se puede observar que el coeficiente de correlación lineal de Pearson entre ellas es muy bajo por lo que no exite correlación entre las variables.

A continuación comprobemos que las variables siguen una distribución normal en cada grupo. Almacenamos en *datos*1 los invididuos que corresponde al grupo 1 (parto por cesárea), en *datos*2 aquellos que corresponde al grupo 2 (parto vaginal) y en *datos*3 aquellos que corresponde al grupo 3 (parto natural) con la función *subset*.

```
datos1<-subset(datos, datos$grupos3==1)
datos2<-subset(datos, datos$grupos3==2)
datos2<-subset(datos, datos$grupos3==3)
```

Veamos gráficamente si se cumple la normalidad univariante.

```
par(mfcol = c(3,4))
for (k in 2:5) {
  j0 <- names(datos)[k]
  x0 <- seq(min(datos[, k]), max(datos[, k]), le = 50)
  for (i in 1:2) {
    i0 <- levels(datos$grupos3)[i]
    x <- datos[datos$grupos3 == i0, j0]
    if(i==1) col<-"blue" else if(i==2) col<-"green" else col<-"magenta"
    qqnorm(x, main = paste("Grupo",i0,":", j0), pch = 19, col =col)
    qqline(x)
  }
}
```

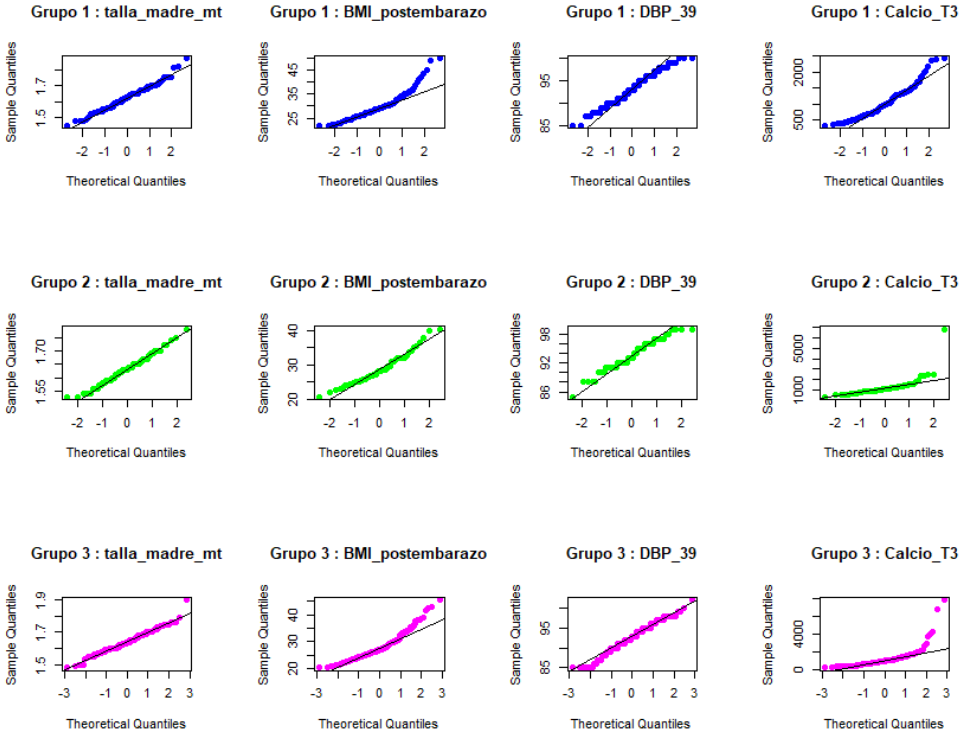

Figura 3.5: Gráficos Q-Q plot

En base a los gráficos de la Figura 3.5 las variables *talla_madre_mt* y *DBP_39* siguen una distribución normal. En cambio, las variables *BMI_postembarazo* y *Calcio_T3* presentan algunas irregularidades que descartan la normalidad. Para corregirlo podemos aplicar transformaciones logarítimas. En este caso, aplicamos la función *log*10 a la variable *BMI_postembarazo* y la función *log* a la variable *Calcio_T3*.

```
datos$log10_BMI_postembarazo<-log10(datos$BMI_postembarazo)
datos$log_Calcio_T3<-log(datos$Calcio_T3)
```

Con dichas transformaciones disminuimos la asimetría que se origina a la derecha provocando que las variables sigan una distribución normal.

Comprobemos que siguen una distribución normal con el test de Kolmogorov-Smirnov.

- Grupo 1.

```
ks.test(x=datos1$talla_madre_mt,"pnorm",mean(datos1$talla_madre_mt),
sd(datos1$talla_madre_mt))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos1$talla_madre_mt
## D = 0.062171, p-value = 0.6426
## alternative hypothesis: two-sided
```

```
ks.test(x=datos1$log_BMI_postembarazo,"pnorm",
mean(datos1$log10_BMI_postembarazo),sd(datos1$log10_BMI_postembarazo))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos1$log10_BMI_postembarazo
## D = 0.094312, p-value = 0.1599
## alternative hypothesis: two-sided

ks.test(x=datos1$DBP_39,"pnorm",mean(datos1$DBP_39),sd(datos1$DBP_39))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos1$DBP_39
## D = 0.10526, p-value = 0.08601
## alternative hypothesis: two-sided

ks.test(x=datos1$log_Calcio_T3,"pnorm",mean(datos1$log_Calcio_T3),
sd(datos1$log_Calcio_T3))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos1$log_Calcio_T3
## D = 0.07402, p-value = 0.418
## alternative hypothesis: two-sided
```

- Grupo 2.

```
 ks.test(x=datos2$talla_madre_mt,"pnorm",mean(datos2$talla_madre_mt),
 sd(datos2$talla_madre_mt))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$talla_madre_mt
## D = 0.080741, p-value = 0.7983
## alternative hypothesis: two-sided

ks.test(x=datos2$log10_BMI_postembarazo,"pnorm",
mean(datos2$log10_BMI_postembarazo),sd(datos2$log10_BMI_postembarazo))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$log10_BMI_postembarazo
## D = 0.080707, p-value = 0.7987
## alternative hypothesis: two-sided

ks.test(x=datos2$DBP_39,"pnorm",mean(datos2$DBP_39),sd(datos2$DBP_39))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$DBP_39
## D = 0.10943, p-value = 0.4275
## alternative hypothesis: two-sided

ks.test(x=datos2$log_Calcio_T3,"pnorm",mean(datos2$log_Calcio_T3),
sd(datos2$log_Calcio_T3))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos2$log_Calcio_T3
```

```
## D = 0.10481, p-value = 0.483
## alternative hypothesis: two-sided
```

- Grupo 3.

```
ks.test(x=datos3$talla_madre_mt,"pnorm",mean(datos3$talla_madre_mt),
sd(datos3$talla_madre_mt))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos3$talla_madre_mt
## D = 0.068201, p-value = 0.2064
## alternative hypothesis: two-sided

ks.test(x=datos3$log10_BMI_postembarazo,"pnorm",
mean(datos3$log10_BMI_postembarazo),sd(datos3$log10_BMI_postembarazo))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos3$log10_BMI_postembarazo
## D = 0.085785, p-value = 0.05513
## alternative hypothesis: two-sided

ks.test(x=datos3$DBP_39,"pnorm",mean(datos3$DBP_39),sd(datos3$DBP_39))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos3$DBP_39
## D = 0.08355, p-value = 0.06631
## alternative hypothesis: two-sided

ks.test(x=datos3$log_Calcio_T3,"pnorm",mean(datos3$log_Calcio_T3),
sd(datos3$log_Calcio_T3))

##  One-sample Kolmogorov-Smirnov test
##
## data:  datos3$log_Calcio_T3
## D = 0.05734, p-value = 0.3987
## alternative hypothesis: two-sided
```

Como el p-valor es mayor que 0.05 en todos los test no se rechaza la hipótesis nula que las distribuciones de las variables sean normales en cada grupo.

Por último validamos que se cumpla la igualdad de las matrices de varianzas-covarianzas con la función *boxM*.

```
library(biotools)
boxM(data=datos[2:5],grouping = datos$grupos3)

##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  datos[2:5]
## Chi-Sq (approx.) = 18.799, df = 20,  p-value = 0.5349
```

Puesto que el p-valor es muy grande no se rechaza la hipótesis nula de la igualdad de las matrices de varianzas-covarianzas.

3.2.3. Aplicación de la técnica

Previamente a la aplicación de la técnica veamos como discriminan nuestras variables por separado con la función *pairs*. En este caso, representaremos a los individuos del grupo 1 por el color azul, a los del grupo 2 por verde y a los del grupo 3 por magenta.

```
pairs(x=datos[,2:5],col=c("blue","green","magenta")[datos$grupos3],pch=16)
```

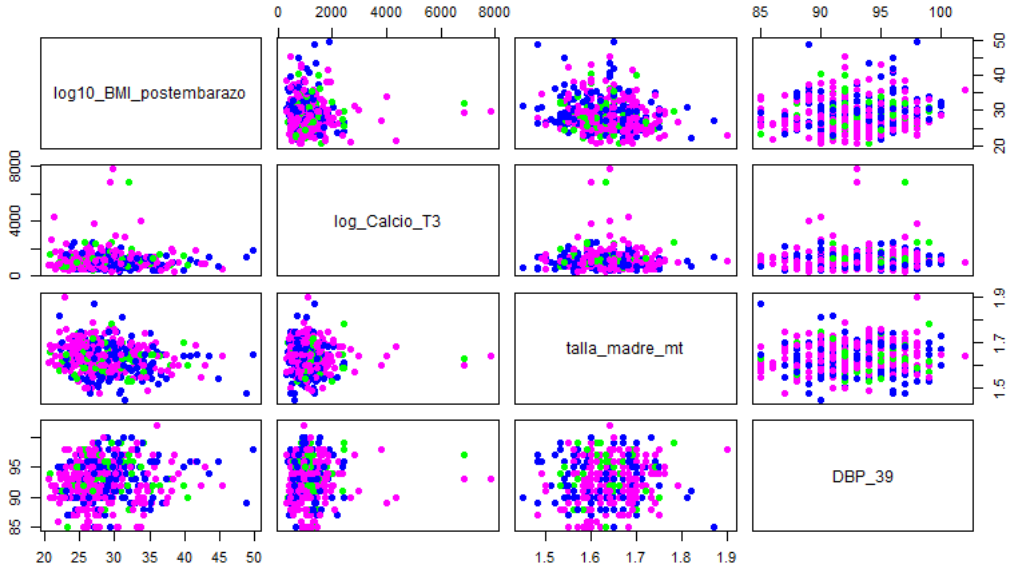

Figura 3.6: Diagrama de dispersión de las cuatro variables discriminantes

En la figura 3.6 se puede apreciar que las variables por separado no provocan una clara discriminación entre los distintos grupos. Por ello aplicamos la técnica usando la función *lda*. Guardamos el resultado en la variable *lda*.

```
lda<-lda(grupos3~talla_madre_mt+log10_BMI_postembarazo+DBP_39+log_Calcio_T3,
data=datos, tol=1.0e-4)

lda

## Call:
## lda(grupos3 ~ talla_madre_mt +  log10_BMI_postembarazo + DBP_39+
## log_Calcio_T3, data = datos, tol = 1e-04)
##
## Prior probabilities of groups:
##         1         2         3
## 0.3155556 0.1422222 0.5422222
##
## Group means:
##    talla_madre_mt log_10_BMI_postembarazo   DBP_39     log_Calcio_T3
## 1       1.624648          1.468689          93.14789     6.850981
## 2       1.632812          1.453399          93.53125     7.010245
## 3       1.640697          1.442366          92.63115     6.935399
```

```
##
## Coefficients of linear discriminants:
##                                LD1         LD2
## talla_madre_mt           -7.06051040  1.7915909
## log10_BMI_postembarazo   10.86357959  0.1042222
## DBP_39                    0.09803895 -0.1968533
## log_Calcio_T3            -0.81249800 -1.5657640
##
## Proportion of trace:
##  LD1    LD2
## 0.8216 0.1784
```

Como se puede ver obtenemos dos funciones discriminantes cumpliendose así la hipótesis de que el número máximo de funciones discriminantes es $min(k-1, p)$, siendo $k = 3$ y $p = 4$. La función *lda* devuelve ahora un nueva salida *Proportion of trace* que nos indica el porcentaje de discriminación de cada función. La primera discrimina un 82.16 % mientras que la segunda 17.84 %.

A partir de los valores de *Coefficients of linear discriminants* obtenemos las funciones discriminantes:

$$D_1 = -7.06051040 \cdot talla\_madre\_mt + 10.86357959 \cdot log10\_BMI\_postembarazo + 0.09803895 \cdot DBP\_39 - 0.81249800 \cdot log\_Calcio\_T3.$$

$$D_2 = 1.7915909 \cdot talla\_madre\_mt + 0.1042222 \cdot log10\_BMI\_postembarazo - 0.1968533 \cdot DBP\_39 - 1.5657640 \cdot log\_Calcio\_T3.$$

A continuación buscamos obtener el porcentanje de acierto de la clasificación.
Empezamos haciendo las prediciones de nuestros resultados a partir de los datos con la función *predict*. Lo guardamos en la variable *clasificacion3g*.

```
clasificacion3g<-predict(lda,datos[2:5])
```

A partir de ellas obtengamos la clasificación del número de individuos que pertenecen a cada grupo usando la matriz de confusión.

```
aciertoclasificacion3g<-confusionmatrix(datos$grupos3,clasificacion3g$class)
aciertoclasificacion3g

##    new 1 new 2 new 3
## 1    31     0   111
## 2     6     0    58
## 3    25     0   219
```

En la tabla se puede apreciar que de los 142 individuos del grupo 1 se han clasificado correctamente 31 y erróneamente 111 al grupo 3. De los 64 invidividuos del grupo 2 no se han clasificado correctamente ningún individuo. De hecho, 8 han sido clasificados al grupo 1 y 58 al grupo 3. Por último, de los 244 individuos del grupo 3 han sido correctamente clasificados 219 y errónamente 25 al grupo 1.

La probabilidad de acertar de cada grupo es

```
diag(prop.table(aciertoclasificacion3g,1))

## 0.2183099 0.0000000 0.8975410
```

Y la probabilidad de acierto total es 55.55 %.

```
sum(diag(prop.table(aciertoclasificacion3g)))

## 0.5555556
```

En conclusión, el modelo tiene tendencia a clasificar los individuos en el grupo 3 puesto que es el más numeroso. De hecho, con dichas variables seleccionadas y con los datos proporcionados en el estudio las funciones discriminantes obtenidas no son capaces de clasificar a ningún individuo en el grupo 2. Esto se debe a que cuando el número de individuos de cada grupo está muy desequilibrado el modelo tiende a clasificarlos en el grupo mayoritario. Luego para hacer un uso correcto de la técnica los grupos deberían de estar compensados o usar otro tipo de función discriminante, como la de Bayes, que tenga en cuenta dicha descompensación.

Bibliografía

[1] Cuadras, C. M. *Nuevos Métodos del análisis multivariante*. CMC Editions, Barcelona, 2007. [Citado en págs. XII, 7, and 25]

[2] Garza García, J. d. l. *Análisis estadístico multivariante : un enfoque teórico y práctico*. McGraw Hill, México, D.F, 2013. [Citado en págs. 18 and 20]

[3] Gil Flores, J. *Análisis discriminante*. Cuadernos de Estadística 12. La Muralla, Madrid, 2001. [Citado en págs. XII, 13, 21, and 23]

[4] Mardia, K. V. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 3 (12 1970), 519–530. [Citado en pág. 19]

[5] Peña, D. *Análisis de datos multivariantes*. McGraw-Hill, Madrid, 2010. [Citado en págs. XII, 25, 29, and 32]

[6] Proyect, R. https://www.r-project.org/. [Citado en págs. XII and 35]

[7] Rencher, A. C. *Methods of multivariate analysis*, 3rd ed. ed. Wiley series in probability and statistics. Wiley, Hoboken, N.J, 2012. [Citado en págs. XII and 3]

[8] Timm, N. H. *Applied multivariate analysis*. Springer texts in statistics. Springer, New York, 2002. [Citado en pág. 20]

[9] Villasenor Alva, J. A., and Estrada, E. G. Una generalización de la prueba de normalidad multivariada de Shapiro-Wilk. *Comunicaciones en Estadística—Teoría y Métodos*. [Citado en pág. 20]