UNIVERSIDAD DE GRANADA

Escuela Internacional de Posgrado

Máster en Estadísica Aplicada

TRABAJO DE FIN DE MÁSTER

Estudio de la técnica multivariante de Análisis Discriminante. Aplicación a datos reales.

Presentado por:

Ana García Burgos

Curso académico 2020-2021

Estudio de la técnica multivariante de Análisis Discriminante. Aplicación a datos reales.

Ana García Burgos

Ana García Burgos ***Estudio de la técnica multivariante de Análisis Discriminante. Aplicación a datos reales.***

Trabajo de fin de Máster. Curso académico 2020-2021.

Responsable de tutorización

Desirée Romero Molina
Departamento de Estadística e Investigación Operativa

Nuria Rico Castro
Departamento de Estadística e Investigación Operativa

Máster en Estadísica Aplicada

Escuela Internacional de Posgrado

Universidad de Granada

Declaración de originalidad

Dña. Ana García Burgos

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Máster (TFM), correspondiente al curso académico 2020-2021, es original, entendida esta, en el sentido de que no ha utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 16 de septiembre de 2021

Fdo: Ana García Burgos

Agradecimientos

Quiero expresar, en primer lugar, un profundo agradecimiento a mi familia por haber sido partícipes en este periodo de formación que finaliza, ayudándome a entender algunas cuestiones médicas y apoyándome en todo momento durante el desarrollo de este trabajo. A mi padre, por ser una fuente de optimismo desde el principio de mis estudios. A mi madre, por las innumerables charlas de apoyo en los momentos más difíciles. A mi hermana, por haber convivido conmigo y haberme ayudado en todos los momentos de mi trayectoria como estudiante. Y, por último, a mi pareja, por haber creído en mí desde el principio y por haberme aportado la seguridad que me faltaba en los momentos más complicados de mi formación.

También deseo dejar constancia en estas líneas de mi gratitud con el doctor D. Antonio Jesús Láinez Ramos-Bossini y su equipo médico, quienes amablemente me han cedido los datos que recopilaron durante las semanas más duras de la pandemia causada por la COVID-19.

Por último, quisiera agradecer y reconocer el esfuerzo de mis responsables de tutorización de este Trabajo Fin de Máster, Desirée y Nuria, por el trabajo que han realizado ayudándome y el tiempo que han dedicado a leer, corregir, aportarme conocimiento y servirme de guía en el transcurso de los últimos años, pues mi trayectoria profesional se ha visto beneficiada gracias a la ayuda que me han prestado incondicionalmente. Siempre serán un referente como docentes y como personas.

Índice general

Introducción

El análisis discriminante es una técnica estadística que sirve para clasificar individuos u objetos según el grupo al que es más probable que pertenezcan. Esta probabilidad se establece a partir de la observación de diferentes variables, las cuales, a diferencia del grupo de pertenencia, deben ser directamente observables. Por tanto, el análisis discriminante trata de dar solución al problema de clasificación cuando existe información de las variables observables y del grupo de pertenencia para un conjunto y se pretende determinar, para un nuevo individuo, cuál es el grupo al que pertenece a partir de la observación o medición de sus características observables. Por ejemplo, en los bancos existen unos sistemas automáticos que, a partir de variables medibles como pueden ser ingresos, patrimonio o tiempo en el trabajo, preveen un comportamiento futuro e informan a la entidad financiera si debe conceder un crédito o no.

Otros ejemplos de aplicación de la técnica del análisis discriminante son clasificar una declaración de impuestos como defraudadora o no, una empresa como en riesgo de quiebra o no, reconocer si un tumor es benigno o maligno, etc.

En el ámbito de la ingeniería al uso de esta técnica se le llama reconocimiento de patrones. La finalidad es diseñar máquinas que sean capaces de clasificar automáticamente, como puede ser el caso de las máquinas que clasifican monedas o billetes, o reconocen sonidos y voces.

Siguiendo a Cea (2016), la primera aplicación del análisis discriminante se mostró en un artículo de Fisher (1936). En una excavación, encontraron restos de cráneo. Se quería saber si éstos pertenecían a humanos o a antropoides. Para ello, se usaron distribuciones de medidas físicas para cráneos de antropoides y de humanos. En el artículo se usó el análisis discriminante lineal y tuvo tal impacto que actualmente es comúnmente llamado análisis discriminante lineal de Fisher. En el artículo, no se cumplían estrictamente

todas las hipótesis necesarias para el correcto uso de la técnica del análisis discriminante lineal, que se verán en la sección 1.2.1, pero sí que dio forma a la idea de usar una variable categórica que fuera combinación lineal de varias variables independientes para la diferenciación entre grupos.

Para el desarrollo de este artículo influyeron propuestas anteriores de medidas de distancias entre grupos, como el coeficiente de semejanza racial, medida de semejanza entre dos razas propuesta por Pearson (1926) y que desarrollarían posteriormente su discípulo Morant (1936) y Mahalanobis (1936).

Welch (1939) adaptó el análisis discriminante lineal de Fisher bajo hipótesis de normalidad multivariante. A partir de estos resultados, Smith (1947) creó una variación del análisis para cuando las matrices de varianzas-covarianzas sean significativamente diferentes en los grupos que se estén considerando: el análisis discriminante cuadrático. Fix y Hodges (1951) trabajaron un método no paramétrico para el análisis discriminante: el análisis discriminante del k-ésimo vecino más próximo que se basa en la idea de clasificar a un nuevo individuo en el mismo grupo que su vecino más próximo. Para medir esta proximidad entre grupos utilizaron la distancia de Mahalanobis. Se clasificará al individuo en el grupo más próximo. En este método no es necesaria la hipótesis de normalidad.

Siguiendo a Cea (2016), los objetivos del análisis discriminante pueden resumirse en:

(i) Describir las características que distinguen a los individuos de un grupo.

(ii) Clasificar a nuevos individuos en los grupos que ya están diferenciados.

El análisis discriminante analiza la relación que existe entre una única variable dependiente y un conjunto de variables independientes. La variable dependiente será categórica y las variables independientes serán de intervalo o de razón. Las categorías de la variable dependiente serán los grupos en los que se quiere discriminar y las variables independientes serán las que se utilicen para poder realizar tal discriminación entre los grupos.

El análisis discriminante se incluye dentro de las técnicas multivariantes de dependencia, ya que trabaja con un conjunto de múltiples variables explicativas para obtener un modelo que permita determinar la pertenencia a un grupo u otro. En la Figura 1 se puede ver un esquema donde se recogen algunas de las técnicas multivariabtes más usuales según se trate de modelizar o no el comportamiento de una o varias variables dependientes.

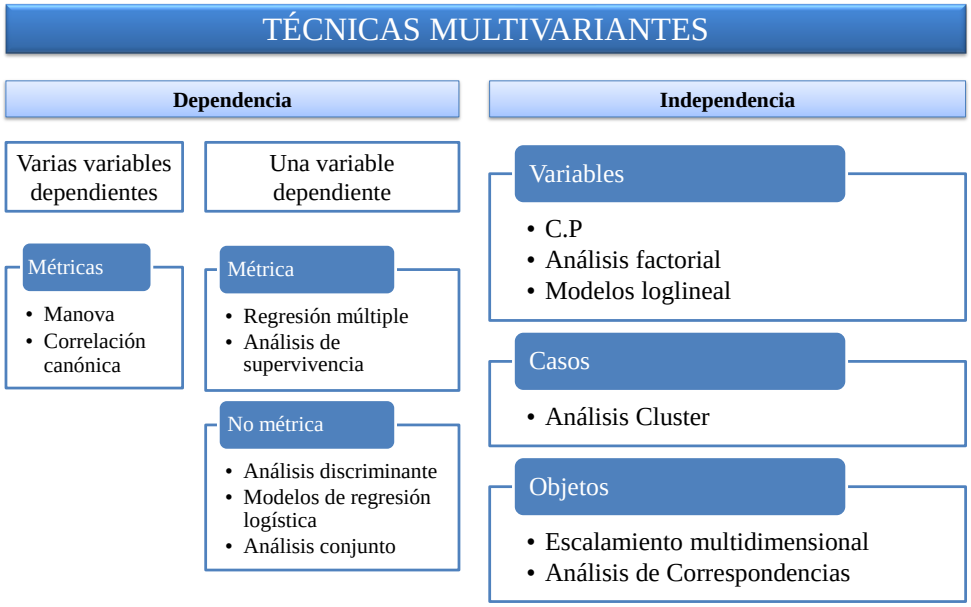

Figura 1: Clasificación de las técnicas multivariantes.

En este trabajo nos vamos a centrar en el estudio del análisis discriminante lineal y en el análisis discriminante cuadrático, ya que son los tipos de análisis discriminante más usados y suelen estar implementados en la mayoría del software estadístico. En el capítulo 1 se expondrán las bases tanto del análisis discriminante lineal de Fisher como del análisis discriminante cuadrático en dos grupos. En el capítulo 2 generalizaremos la idea al caso de g grupos. En el capítulo 3, se explicarán los principales métodos de selección de variables discriminantes y cómo se validan las hipótesis necesarias para el desarrollo de la técnica. Por último, en el capítulo 4 haremos una aplicación con datos reales con el software estadístico R. Las funciones utilizadas en el programa estarán explicadas en el Anexo.

Capítulo 1

Funciones discriminantes en dos grupos

Para desarrollar este capítulo partimos del supuesto de que conocemos la discriminanción de un conjunto de individuos en dos grupos diferenciados, es decir, sabemos a qué grupo pertenecen ciertos individuos y además tenemos información observable de ellos.

Nuestro objetivo será encontrar un modelo capaz de relacionar la clasificación de los individuos con la información que se tiene de ellos para poder, en un futuro, clasificar de forma correcta nuevos casos a partir solamente de la información observable en ellos.

1.1. Planteamiento del problema

Según Cuadras (2018), podemos plantear el problema de clasificación de la siguiente forma. Consideremos G_I y G_{II} dos grupos en los cuales tenemos definido un vector aleatorio $\mathbf{X} = (X_1, \ldots, X_p)'$ continuo.

Definición 1. *Se denominan variables discriminantes o variables independientes cada una de las variables que componen el vector aleatorio continuo* $\boldsymbol{X} = (X_1, \ldots, X_p)'$.

Nuestro problema consistirá en clasificar un individuo nuevo en uno de los dos grupos, conociendo su valor en las p variables, $\mathbf{x} = (x_1, \ldots, x_p)'$. Denotaremos al nuevo individuo como $\mathbf{p}$.

Definición 2. *Se define la regla discriminante como el criterio que clasifica un individuo,* $\mathbf{p}$*, en un grupo, conociendo* $\mathbf{x}$ *las observaciones del vector* $\boldsymbol{X}$ *sobre* $\mathbf{p}$.

Definición 3. *Se dice que $S(x_1, \ldots, x_p)$ es una función discriminante cuando permite aplicar una regla discriminante. Diremos que $\mathbf{p} \in G_I$ cuando $S(x_1, \ldots, x_p) > 0$ y en otro caso diremos que $\mathbf{p} \in G_{II}$.*

En este capítulo veremos dos funciones discriminantes, la función discriminante lineal y la función discriminante cuadrática.

1.2. Análisis discriminante lineal

En esta sección estudiaremos en detalle el discriminante lineal, buscando una función discriminante que será una combinación lineal de las variables discriminantes que minimice los errores de clasificación.

1.2.1. Hipótesis

Para efectuar de forma correcta el análisis debemos considerar una serie de supuestos:

(1) Disponemos de una matriz que contiene una variable categórica, donde se recoge el grupo de pertenencia, y el resto de variables son de intervalo o de razón.

(2) Debe haber al menos dos grupos y cada uno de los grupos debe contener al menos dos individuos.

(3) El número de variables discriminantes debe ser menor que el número de individuos menos 2. Esto es, si $(X_1, \ldots, X_p)'$ es el vector de variables, tiene que verificarse que $p < (N - 2)$, siendo N el número de objetos.

(4) **Ausencia de multicolinealidad**. Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.

(5) **Igualdad de matrices de varianzas-covarianzas**. Las matrices de varianzas-covarianzas dentro de cada grupo deben ser aproximadamente iguales.

(6) **Normalidad multivariante**. Las variables deben seguir una distribución normal multivariante.

1.2.2. Caso de una variable discriminante

Supongamos que solo disponemos de una variable discriminante, X. El objetivo es encontrar una función lineal de la variable discriminante X que permita clasificar cada observación en G_I o en G_{II}, minimizando el error de clasificación y teniendo en cuenta que las distribuciones en G_I y en G_{II} solo se diferencian en su localización pero tienen la misma forma y la misma varianza.

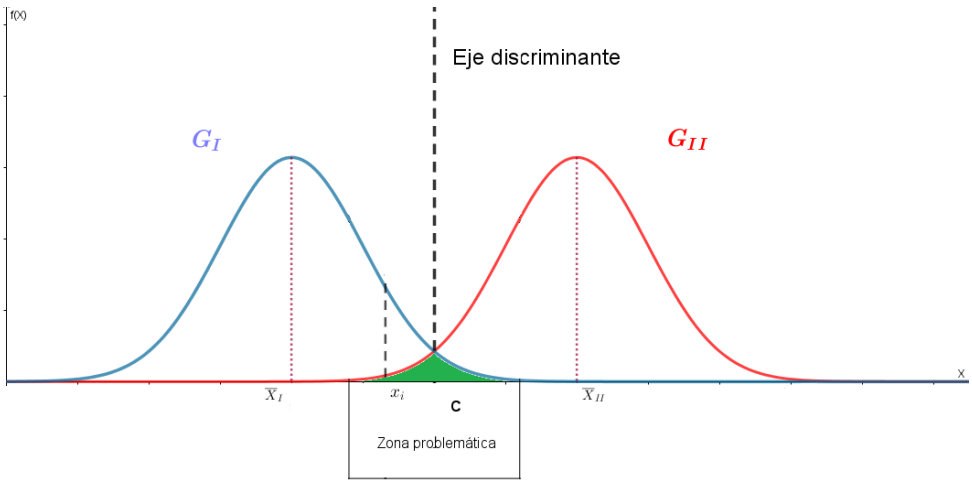

Figura 1.1: Análisis del problema en dos grupos con una sola variable discriminante.

Si nos fijamos en la Figura 1.1, podemos tomar como función o eje discriminante la recta $C = \dfrac{\overline{X}_I + \overline{X}_{II}}{2}$ siendo $\overline{X}_I$ la media muestral de la variable X en el grupo G_I y siendo $\overline{X}_{II}$ la media muestral de la variable X en el grupo G_{II}. De esta manera, la zona problemática indicada en la Figura 1.1 quedará minimizada. Por tanto, usando esta función discriminante se cometerá un error mínimo para el problema de clasificación.

1.2.3. Caso de dos variables discriminantes

Supongamos ahora que disponemos de dos variables discriminantes (X_1, X_2). Queremos encontrar una función lineal de las variables discriminantes X_1 y X_2 que permita clasificar cada observación en G_I o en G_{II}, minimizando el error de clasificación.

Para ello, consideramos X_1 y proyectamos sobre el eje correspondiente los datos observados y usamos la misma solución que hemos obtenido en el caso de una variable discriminante. Al hacer esto, se crea una zona problemática debido a la superposición de las distribuciones normales de los dos grupos. A continuación, hacemos lo mismo con X_2 y se crea otra nueva zona problemática. Por tanto, tenemos dos zonas problemáticas distintas.

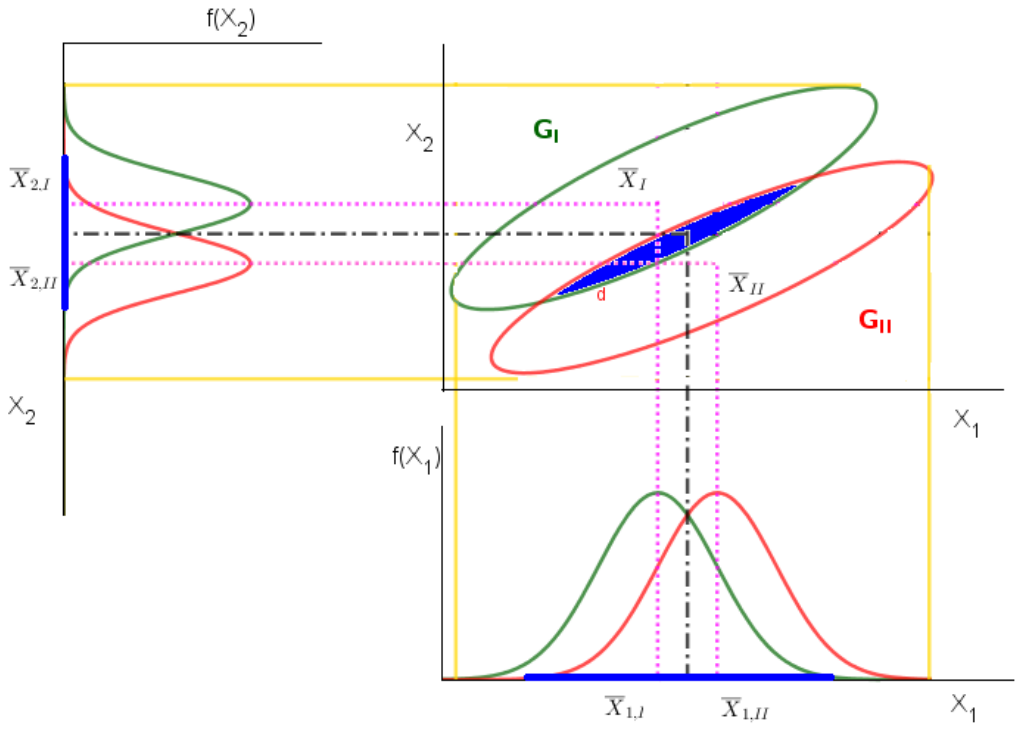

Figura 1.2: Análisis del problema en dos grupos con dos variables discriminantes.

Si consideramos el ejemplo de la Figura 1.2, se tiene que la zona problemática creada por la variable discriminante X_2 es menor que la generada por X_1. En consecuencia, X_2 discrimina mejor que X_1. Pero como nuestro objetivo es minimizar la región problemática, buscaremos una función lineal de X_1 y X_2

$$D = \omega_1 X_1 + \omega_2 X_2$$

que minimice dicha región, como se muestra en la Figura 1.3.

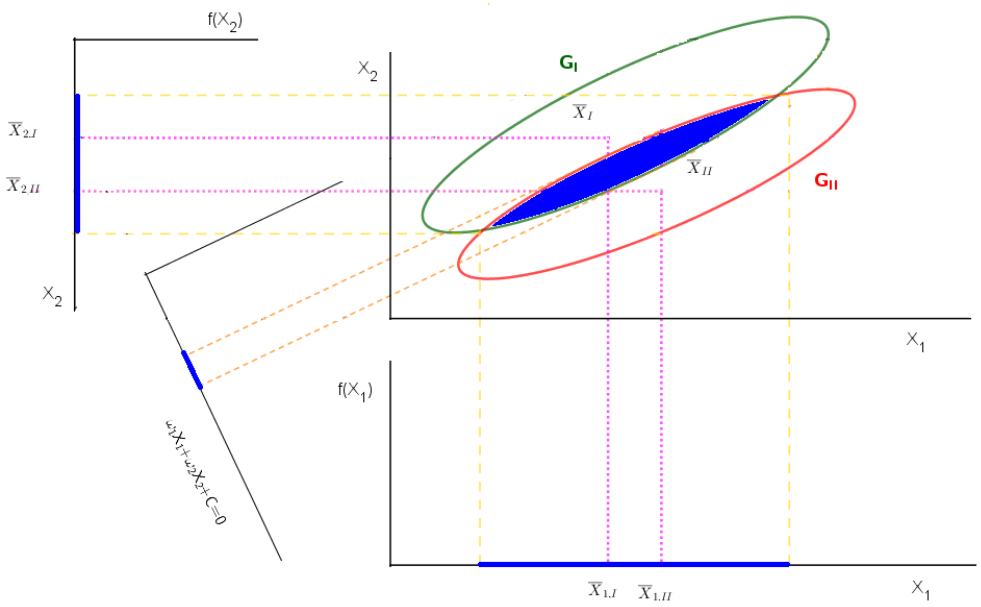

Figura 1.3: Función lineal que minimiza la zona problemática con dos variables discriminantes.

Fisher resolvió este problema creando una función que maximizaba la separación entre los grupos, maximizando la distancia entre sus medias, y minimizaba la dispersión dentro de los grupos. La función lineal que se obtiene al aplicar estas condiciones se denomina discriminante lineal de Fisher. A continuación veremos cómo se obtiene dicha función para el caso general de p variables discriminantes.

1.2.4. Generalización a p variables discriminantes

Como hemos indicado, siguiendo la idea de Fisher (1936), queremos encontrar una función discriminante capaz de minimizar la variabilidad dentro de los grupos y maximizar la variabilidad entre los grupos que sea combinación lineal de las p variables de las que se dispone:

$$D = \omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_p X_p.$$

El objetivo es hallar los coeficientes ω_j. Para ello consideremos N observaciones.

Para cada observación $i = 1, \ldots, N$, obtenemos la función discriminante:

$$D_i = \omega_1 X_{1i} + \omega_2 X_{2i} + \ldots + \omega_p X_{pi}.$$

Si expresamos las funciones discriminantes anteriores matricialmente, obtenemos:

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{p1} \\ X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} & X_{2N} & \dots & X_{pN} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_p \end{bmatrix} \tag{1.1}$$

Podemos reescribir la expresión (1.1) en función de las desviaciones de la media, obteniendo:

$$\begin{bmatrix} D_1 - \overline{D} \\ D_2 - \overline{D} \\ \vdots \\ D_N - \overline{D} \end{bmatrix} = \begin{bmatrix} X_{11} - \overline{X}_1 & X_{21} - \overline{X}_2 & \dots & X_{p1} - \overline{X}_p \\ X_{12} - \overline{X}_1 & X_{22} - \overline{X}_2 & \dots & X_{p2} - \overline{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} - \overline{X}_1 & X_{2N} - \overline{X}_2 & \dots & X_{pN} - \overline{X}_p \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_p \end{bmatrix} \tag{1.2}$$

siendo

$$\overline{D} = \omega_1 \overline{X}_1 + \omega_2 \overline{X}_2 + \ldots + \omega_p \overline{X}_p.$$

La expresión (1.2) se resume en la función discriminante en diferencias:

$$d = X\omega.$$

Obtengamos a partir de d la variabilidad de la función discriminante, es decir, la suma de cuadrados de las desviaciones de las variables discriminantes con respecto de su media:

$$d'd = \omega' X' X \omega,$$

siendo $X'X$ la matriz simétrica que expresa las desviaciones cuadráticas respecto de la media de las variables.

A continuación, veamos un teorema que nos resultará útil para seguir con los cálculos.

Teorema. *La matriz de desviaciones cuadráticas $X'X$ se puede descomponer como la suma entre la matriz que incluye las variabilidades entre los grupos para las variables, E, y la matriz que incluye las variabilidades dentro de cada grupo o intragrupos I. Es decir*

$$X'X = E + I,$$

con:

$$E = n_I(\overline{X}_I - \overline{X})(\overline{X}_I - \overline{X})' + n_{II}(\overline{X}_{II} - \overline{X})(\overline{X}_{II} - \overline{X})'$$

donde:

$$\overline{X} = \frac{n_I \overline{X}_I + n_{II} \overline{X}_{II}}{N} \quad con \quad N = n_I + n_{II}.$$

$$\overline{X}_I = (\overline{X}_1^I, \ldots, \overline{X}_p^I)' \quad con \quad \overline{X}_j^I = \frac{\sum_{i=1}^{n_I} X_{ij}}{n_I}$$

$$\overline{X}_{II} = (\overline{X}_1^{II}, \ldots, \overline{X}_p^{II})' \quad con \quad \overline{X}_j^{II} = \frac{\sum_{i=1}^{n_{II}} X_{ij}}{n_{II}}$$

$$I = \sum_{i=1}^{n_I} (X_{iI} - \overline{X}_I)(X_{iI} - \overline{X}_I)' + \sum_{i=1}^{n_{II}} (X_{iII} - \overline{X}_{II})(X_{iII} - \overline{X}_{II})',$$

donde:

$$X_{iI} = (X_{i1}^I, \ldots, X_{ip}^I)'$$

$$X_{iII} = (X_{i1}^{II}, \ldots, X_{ip}^{II})'$$

siendo X_{ij}^I el individuo i observado en la variable j en el grupo I y siendo X_{ij}^{II} el individuo i observado en la variable j en el grupo II.

Nota. *Denotaremos como $\overline{X}_I$ y $\overline{X}_{II}$ los centroides de los grupos I y II, respectivamente.*

Demostración. *Para verlo, tomemos el elemento k-ésimo de la diagonal principal de la matriz $X'X$:*

$$\sum_{i=1}^{N} (X_{ik} - \overline{X}_k)^2 = \sum_{i=1}^{n_I} (X_{ik} - \overline{X}_k)^2 + \sum_{i=1}^{n_{II}} (X_{ik} - \overline{X}_k)^2 =$$

$$= \sum_{i=1}^{n_I} (X_{ik} - \overline{X}_k^I + \overline{X}_k^I - \overline{X}_k)^2 + \sum_{i=1}^{n_{II}} (X_{ik} - \overline{X}_k^{II} + \overline{X}_k^{II} - \overline{X}_k)^2 =$$

$$= \sum_{i=1}^{n_I} (X_{ik} - \overline{X}_k^I)^2 + \sum_{i=1}^{n_I} (\overline{X}_k^I - \overline{X}_k)^2 + \sum_{i=1}^{n_{II}} (X_{ik} - \overline{X}_k^{II})^2 + \sum_{i=1}^{n_{II}} (\overline{X}_k^{II} - \overline{X}_k)^2 =$$

$$= \sum_{i=1}^{n_I} (X_{ik} - \overline{X}_k^I)^2 + n_I (\overline{X}_k^I - \overline{X}_k)^2 + \sum_{i=1}^{n_{II}} (X_{ki} - \overline{X}_k^{II})^2 + n_{II} (\overline{X}_k^{II} - \overline{X}_k)^2 \quad k = 1, \ldots, p.$$

Análogamente, se podría demostrar para los elementos cruzados y se tiene:

$$X'X = E + I.$$

Por tanto, usando el teorema obtenemos que:

$$d'd = \omega' X' X \omega = \omega'(E + I)\omega = \omega' E \omega + \omega' I \omega.$$

Si seguimos la idea de Fisher, para encontrar las funciones D_i que consigan discriminar de la mejor manera posible, tenemos que maximizar la varianza entre grupos y minimizar la varianza dentro de los grupos.

Esto se resume en hallar

$$\text{máx} \left(\frac{\omega' E \omega}{\omega' I \omega} \right).$$

Calcular este máximo es equivalente a calcular máx$(\omega' E \omega)$ con $\omega' I \omega = 1$, pues la función $\dfrac{\omega' E \omega}{\omega' I \omega}$ es invariante frente a cambios de escala.

Para calcular el máximo aplicaremos multiplicadores de Lagrange:

$$L = \omega' E \omega - \lambda(\omega' I \omega - 1) \Rightarrow \frac{\partial \mathrm{L}}{\partial \omega} = 2E\omega - 2\lambda I\omega = 0$$
$$\Rightarrow E\omega = \lambda I\omega \Rightarrow (I^{-1}E)\omega = \lambda\omega.$$

Como $E\omega = \lambda I\omega$, obtenemos que $\omega' E \omega = \omega' \lambda I \omega = \lambda$.

Luego si tomamos el vector propio asociado al máximo valor propio obtendremos la función que mejor discrimina.

1.2.5. Clasificación

Supongamos ahora que queremos clasificar un individuo **p** cuya observación viene dada por $\mathbf{x_0} = (x_1, \ldots, x_p)'$. Entonces calculamos la función discriminante d_0, sustituyendo los valores correspondientes de las p variables en ella.

Podemos calcular la frontera discriminante:

$$C = \frac{\overline{D}_I + \overline{D}_{II}}{2},$$

siendo:

$$\overline{D}_I = \omega_1 \overline{X}_{1I} + \ldots + \omega_p \overline{X}_{pI}$$
$$\overline{D}_{II} = \omega_1 \overline{X}_{1II} + \ldots + \omega_p \overline{X}_{pII}$$

con $\omega_1, \ldots, \omega_p$ los elementos del vector propio asociado al mayor valor propio de la matriz $I^{-1}E$. Entonces, clasificaremos la observación en G_I si

$$d_0 < C \quad \Leftrightarrow \quad d_0 - C < 0,$$

y clasificaremos la observación en G_{II} si:

$$d_0 > C \quad \Leftrightarrow \quad d_0 - C > 0.$$

Esta regla de clasificación tiene un problema y es que no tiene en cuenta que los grupos puedan tener distinto tamaño. Si esto ocurre, al usar C como frontera de clasificación, la proporción de casos mal clasificados en el grupo de menor tamaño será mucho mayor que en el grupo de mayor tamaño.

Entonces, cuando los tamaños son desiguales se puede usar una regla de clasificación que desplaza el punto de corte acercándolo al centroide del grupo de menor tamaño para igualar los errores de clasificación, como por ejemplo la distancia ponderada:

$$C = \frac{n_I \overline{D}_I + n_{II} \overline{D}_{II}}{n_I + n_{II}}.$$

Otra opción es calcular las funciones discriminantes para el grupo G_I y para el grupo G_{II} y clasificar la observación en el grupo en el cual la función tenga mayor valor.

1.3. Análisis discriminante cuadrático

En esta sección estudiaremos en detalle el discriminante cuadrático, buscando una función discriminante que, en este caso, será cuadrática.

1.3.1. Hipótesis

Para realizar de forma correcta el análisis discriminante cuadrático se deben satisfacer las mismas hipótesis que el discriminante lineal, vistas en el apartado 1.2.1, a excepción de la hipótesis de igualdad de matrices de varianzas-covarianzas, pues en esta técnica se tiene en cuenta la matriz de varianzas-covarianzas en cada grupo. Este hecho da lugar a una función discriminante cuadrática, que estudiaremos a continuación.

1.3.2. Desarrollo de la técnica

Sean $\mathbf{X} = (X_1, \ldots, X_p)'$ el vector aleatorio continuo que contiene las p variables discriminantes, $\mathbf{x} = (x_1, \ldots, x_p)'$ las observaciones de dichas variables y G_I y G_{II} los grupos considerados. Vamos a suponer que el vector $\mathbf{x}$ sigue una $\mathcal{N}_p(\boldsymbol{\mu_1}, \Sigma_1)$ en G_I y una $\mathcal{N}_p(\boldsymbol{\mu_2}, \Sigma_2)$ en G_{II}.

Sean $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ las funciones de densidad de $\mathbf{x}$ en los grupos G_I y G_{II} respectivamente. Entonces clasificaremos $\mathbf{p}$ en G_I cuando:

$$f_1(\mathbf{x}) > f_2(\mathbf{x})$$

y clasificaremos $\mathbf{p}$ en G_{II} en otro caso. Operando, obtenemos:

$$f_1(\mathbf{x}) > f_2(\mathbf{x}) \Leftrightarrow log\left(f_1(\mathbf{x})\right) - log\left(f_2(\mathbf{x})\right) > 0.$$

Luego, según lo especificado en la definición 3, el discriminante cuadrático puede definirse como:

$$Q(\mathbf{x}) = log\left(f_1(\mathbf{x})\right) - log\left(f_2(\mathbf{x})\right).$$

Sabemos que la función de densidad de una Normal viene dada por:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right\}. \tag{1.3}$$

Luego, operando, se obtiene:

$$log\left(f_i(\mathbf{x})\right) = -\frac{p}{2} log(2\pi) - \frac{1}{2} log(|\Sigma_i|) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \quad \forall i = 1, 2.$$

Por tanto el discriminante cuadrático queda de la forma:

$$Q(\mathbf{x}) = \frac{1}{2}\mathbf{x}'(\Sigma_2^{-1} - \Sigma_1^{-1})\mathbf{x} + \mathbf{x}'(\Sigma_1^{-1}\mu_{\mathbf{1}} - \Sigma_2^{-1}\mu_{\mathbf{2}}) +$$

$$+\frac{1}{2}\mu_{\mathbf{2}}{}'\Sigma_2^{-1}\mu_{\mathbf{2}} - \frac{1}{2}\mu_{\mathbf{1}}{}'\Sigma_1^{-1}\mu_{\mathbf{1}} + \frac{1}{2} log(|\Sigma_2|) - \frac{1}{2} log(|\Sigma_1|).$$

1.3.3. Clasificación

Usando el discriminante cuadrático clasificaremos $\mathbf{p}$ en G_I cuando $Q(\mathbf{x}) > 0$ y clasificaremos $\mathbf{p}$ en G_{II} cuando $Q(\mathbf{x}) < 0$.

Capítulo 2

Funciones discriminantes en más de dos grupos

En este capítulo trataremos de generalizar la idea expuesta en el capítulo 1 al caso en el que estemos trabajando con g grupos.

Consideremos $G_1, \ldots, G_g$ los grupos en los cuales tenemos definido un vector aleatorio continuo $\mathbf{X} = (X_1, \ldots, X_p)'$.

Nuestro problema será clasificar un nuevo individuo en uno de los g grupos conociendo su valor en las p variables. En este caso, no será posible encontrar una única función discriminante. Serán necesarias $g - 1$ funciones discriminantes.

2.1. Análisis discriminante lineal

2.1.1. Hipótesis

Para efectuar de forma correcta el análisis se tienen que verificar las hipótesis enunciadas en la sección 1.2.1. Además, se tiene que el número máximo de funciones que se pueden calcular viene dado por:

$$\text{mín}\,(p, g - 1).$$

2.1.2. Motivación geométrica

A continuación, siguiendo a Gil y cols. (2001), realizamos una motivación geométrica que nos facilitará comprender el problema. Podemos construir

una matriz de datos de N filas y p columnas, siendo N el número de individuos y p el número de variables discriminantes, como la que se muestra en el Cuadro 2.1:

Variable / **Individuo**	X_1	X_2	.....	X_p
Individuo 1	x_{11}	x_{21}	.....	x_{p1}
Individuo 2	x_{12}	x_{22}	.....	x_{p2}
.....	.....	.....	.....	.....
Individuo N	x_{1N}	x_{2N}	.....	x_{pN}

Cuadro 2.1: Motivación geométrica.

Podemos considerar las variables discriminantes como ejes en el espacio p-dimensional que generan y cada individuo, es decir, cada fila de la matriz anterior, como un punto en dicho espacio. Por ejemplo, supongamos que tenemos únicamente tres variables discriminantes. Representando la idea anterior, obtenemos lo que se muestra en la Figura 2.1.

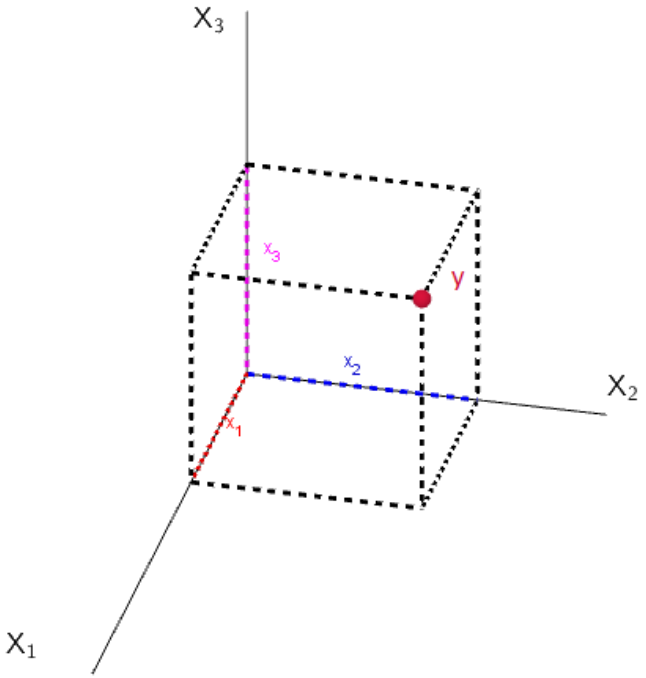

Figura 2.1: Posición del punto y en el espacio generado por las variables discriminantes X_1, X_2 y X_3.

Por tanto, cuando los individuos pertenecen a un mismo grupo, tendrán una situación en el espacio similar, como se muestra en la Figura 2.2.

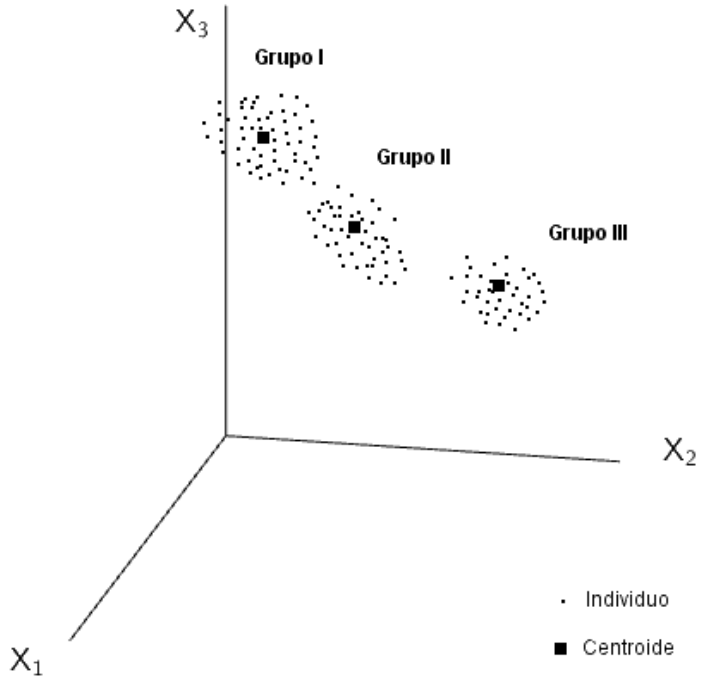

Figura 2.2: Posición de las observaciones y de los centroides.

Entonces, si el problema que queremos resolver consiste en ver las diferencias existentes entre los grupos a partir de las variables discriminantes, bastará con analizar la posición de los centroides para ver si los grupos están bien diferenciados en el espacio generado por las p variables.

Si observamos la Figura 2.2, podemos construir un plano que pase por los tres centroides. Nuestro objetivo será determinar los ejes del plano que maximicen la distancia entre los centroides.

Intuitivamente, podemos construir un eje apuntando al punto en el que los centroides estén más separados, consiguiendo así la máxima dispersión posible. Para construir el segundo eje, seguiremos el mismo criterio pero además este nuevo eje debe ser perpendicular al primero que hemos construido. De esta manera, conseguimos construir los dos ejes que proporcionan la mayor dispersión entre los tres grupos en el espacio, como se observa en la Figura 2.3.

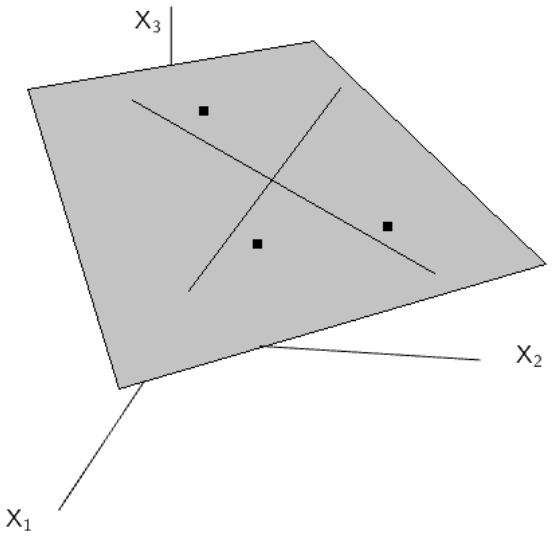

Figura 2.3: Ejes que consiguen la máxima dispersión para los tres centroides en el espacio generado por las variables discriminantes X_1, X_2 y X_3.

Cada uno de estos ejes es una función discriminante.

2.1.3. Caso de p variables discriminantes

Podemos generalizar la idea expuesta en la sección 1.2.4. Supongamos que tenemos g grupos donde se asignan una serie de individuos y p variables medidas sobre ellos, $(X_1, \ldots, X_p)'$.

El objetivo es encontar funciones discriminantes que nos permitan clasificar a cada individuo en su correspondiente grupo. Buscamos funciones discriminantes $(D_1, \ldots, D_m)'$ que sean funciones lineales de $(X_1, \ldots, X_p)'$, es decir:

$$D_1 = \omega_{11}X_1 + \omega_{12}X_2 + \ldots + \omega_{1p}X_p$$

$$D_2 = \omega_{21}X_1 + \omega_{22}X_2 + \ldots + \omega_{2p}X_p$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$D_m = \omega_{m1}X_1 + \omega_{m2}X_2 + \ldots + \omega_{mp}X_p,$$

siendo $m = \text{mín}(g-1, p)$.

Queremos que estas funciones discriminen lo máximo posible a los g grupos. Luego, basándonos en la idea de Fisher, las combinaciones lineales de las p variables tienen que maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos para las N observaciones de las que se dispone.

2.1.4. Obtención de las funciones discriminantes

El objetivo es obtener m funciones $(D_1, \ldots, D_m)'$ a partir de $(X_1, \ldots, X_p)'$ variables observadas en g grupos tal que:

$$D_j = \omega_{j1}X_1 + \omega_{j2}X_2 + \ldots + \omega_{jp}X_p, \quad j = 1, \ldots, m,$$

siendo $m = \text{mín}(g-1, p)$ y verificando $Corr(D_j, D_k) = 0 \quad \forall j \neq k$.

Entonces, calcularemos $(D_1, \ldots, D_m)'$ de forma que:

- D_1 será la combinación lineal de $(X_1, \ldots, X_p)'$ que proporcione la mayor discriminación entre los centroides de los grupos.
- D_2 será la combinación lineal de $(X_1, \ldots, X_p)'$ que proporcione la mayor discriminación entre los centroides de los grupos, después de D_1, y que verifique $Corr(D_1, D_2) = 0$.
- En general, D_j será la combinación lineal de $(X_1, \ldots, X_p)'$ que proporcione la mayor discriminación entre los centroides de los grupos, después de D_{j-1}, y que verifique $Corr(D_j, D_k) = 0$, con $\forall k = 1, \ldots, (j-1)$.

De forma análoga al razonamiento seguido en la sección 1.2, se obtienen las funciones discriminantes lineales asociadas al modelo.

2.1.5. Problema de clasificación

Una vez halladas las funciones discriminantes podemos clasificar los individuos usados para crear dichas funciones para ver el grado de eficacia en el ámbito de clasificar.

Otra opción más precisa, según comenta Berrar (2018), sería dividir la muestra en dos partes, la primera sería una muestra de entrenamiento con el que se construirán las funciones discriminantes y, la segunda, una muestra para probar el grado de eficacia de la clasificación.

Si los resultados son buenos, podemos usar las funciones discriminantes para clasificar nuevos individuos de los que se desconozca su procedencia

conociendo sus valores en las variables discriminantes $(X_1, \ldots, X_p)'$ construyendo las funciones discriminantes para cada grupo y clasificando en el grupo en el que la puntuación discriminante sea mayor.

2.2. Análisis discriminante cuadrático

2.2.1. Hipótesis

Para efectuar de forma correcta el análisis se tienen que verificar las hipótesis enunciadas en la sección 1.3.1. Además, se tiene que el número máximo de funciones discriminantes viene dado por:

$$\text{mín}\,(p, g-1).$$

2.2.2. Desarrollo de la técnica

Consideremos $f_i(\mathbf{x})$, con $i = 1, \ldots, g$ las funciones de densidad de $\mathbf{x}$ en los grupos G_i, con $i = 1, \ldots, g$. Clasificaremos el individuo $\mathbf{p}$ en el grupo que tenga mayor función de densidad. Es decir, clasificaremos $\mathbf{p}$ en G_i cuando:

$$f_i(\mathbf{x}) = \text{máx}\left\{f_1(\mathbf{x}), \ldots, f_g(\mathbf{x})\right\}.$$

Se puede definir el discriminante cuadrático de la forma:

$$Q_{ij}(\mathbf{x}) = log\left(f_i(\mathbf{x})\right) - log\left(f_j(\mathbf{x})\right) \quad \forall i \neq j, \quad i, j = 1, \ldots, g.$$

Desarrollando la expresión se obtiene:

$$Q(\mathbf{x}) = \frac{1}{2}\mathbf{x}'(\Sigma_j^{-1} - \Sigma_i^{-1})\mathbf{x} + \mathbf{x}'(\Sigma_i^{-1}\mu_{\mathbf{1}} - \Sigma_j^{-1}\mu_{\mathbf{j}}) +$$

$$+\frac{1}{2}\mu_{\mathbf{j}}{}'\Sigma_j^{-1}\mu_{\mathbf{j}} - \frac{1}{2}\mu_{\mathbf{i}}{}'\Sigma_i^{-1}\mu_{\mathbf{i}} + \frac{1}{2}log(|\Sigma_j|) - \frac{1}{2}log(|\Sigma_i|) \quad \forall i \neq j, \quad i, j = 1, \ldots, g.$$

Capítulo 3

Consideraciones sobre los datos

3.1. Selección de variables discriminantes

Como hemos visto, la idea fundamental del análisis discriminante lineal de Fisher es construir funciones lineales a partir de las variables discriminantes que discriminen entre los distintos grupos.

No todas las variables discriminan de igual forma. Debido a este hecho, no debemos incluir todas las variables discriminantes iniciales consideradas en el estudio en la función lineal que construyamos.

Se suelen utilizar tres métodos para seleccionar las variables: el método hacia delante, el método hacia atrás y el método paso a paso.

Para la utilización de dichos métodos es necesario el cálculo de los siguientes estadísticos:

- **F de Snedecor**. Para una variable X_i, con $i = 1, \ldots, p$, se comparan las desviaciones de las medias de cada grupo a la media total, entre las desviaciones a la media dentro de cada grupo.

  - Cuando F es grande para X_i, las medias de cada grupo están muy separadas y, en consecuencia, X_i discrimina bien.
  - Cuando F es pequeña para X_i, hay poca homogeneidad en los grupos, luego los grupos están muy próximos. En consecuencia, la variable no discrimina bien.

- **Λ de Wilks**. Este estadístico mide el poder discriminante del conjunto

de variables discriminantes. Viene dado por:

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

es decir, por las desviaciones de la media dentro de cada grupo, entre las desviaciones a la media total sin distinguir grupos. Toma valores entre 0 y 1, entonces:

- Cuando Λ se aproxima a 0, la variable discrimina bien.
- Cuando Λ se aproxima a 1, la variable no discrimina bien.

Describamos brevemente los métodos citados anteriormente.

- **Método hacia delante**. En primer lugar, se considera la variable con mayor F o con menor Λ. Después, consideramos la segunda variable con mayor F o con menor Λ y seguimos haciendo esto hasta que ninguna variable que quede por elegir discrimine significativamente.
- **Método hacia atrás**. Suponemos que todas las variables son necesarias. Eliminamos la variable que menos discrimina entre los grupos y repetimos este procedimiento hasta que entre las variables que hayamos eliminado no quede ninguna que discrimine de manera significativa. Este método suele seleccionar pocas variables.
- **Método paso a paso**. Este método es una combinación del método hacia delante y del método hacia atrás. Se aplica el método hacia delante y a continuación, se aplica el método hacia atrás, es decir, cuanto se introduce una variable se reevalúan todas para ver si alguna de ellas debe salir. Para determinar qué variables entran y salen en cada paso se hace lo siguiente:
  - Proporcionar un p-valor de entrada y otro de salida.
  - Si el p-valor obtenido al introducir una variable no es inferior al p-valor de entrada, la variable que hemos considerado no entra.
  - Si el p-valor obtenido al eliminar una variable no es superior al de salida, la variable no sale del conjunto de discriminación

3.2. Validación de las hipótesis

En esta sección explicaremos la importancia de las hipótesis:

(i) Normalidad multivariante.

(ii) Igualdad de matrices de varianzas-covarianzas.

(iii) Ausencia de multicolinealidad.

Estudiaremos técnicas que permitan comprobar si se verifican estos supuestos.

3.2.1. Ausencia de multicolinealidad y singularidad

La multicolinealidad existe cuando hay una fuerte correlación entre dos variables y cuando estas muestran el mismo patrón de correlaciones que las restantes variables.

La singularidad ocurre cuando las observaciones de una variable son aproximadamente una combinación lineal del resto de variables.

Por tanto, al añadir esta hipótesis se tiene que una variable discriminante no puede ser combinación lineal del resto de variables discriminantes consideradas.

Esta hipótesis ha de ser considerada para asegurar que tengan sentido los cálculos matemáticos que se hacen al poner en marcha la técnica, pues la ausencia de multicolinealidad y singularidad hace que el rango de la matriz de varianzas-covarianzas disminuya y en consecuencia su determinante será nulo, luego no podremos obtener las expresiones que requieran la inversa de la matriz de varianzas-covarianzas.

En el caso de casi multicolinealidad o singularidad, el determinante de la matriz de varianzas-covarianzas no será nulo, pero su valor sí que será próximo a cero, haciendo que la inversa de la matriz de varianzas-covarianzas tenga valores muy inestables.

Además de asegurar la estabilidad de la inversa de la matriz, cumplir esta hipótesis asegura que todas las variables discriminantes aporten información diferente sobre los individuos y no sean redundantes.

Existen diferentes métodos para detectar matrices con multicolinealidad o sigularidad. Una opción sería utilizar la matriz de correlaciones de Pearson. Si en ella encontramos valores próximos a -1 y 1 no estaríamos cumpliendo la hipótesis. El problema es que este método únicamente detecta la colinealidad

bivariada.

Para solucionar esto, podemos usar regresión múltiple, tomando una variable como dependiente y el resto como independientes una y otra vez, y cuando la correlación múltiple al cuadrado, R^2, entre la combinación lineal de las variables independientes y la variable dependiente sea alta diremos que los valores de la variable se pueden predecir por una combinación lineal de las variables restantes, es decir, las variables serán casi singulares.

Si detectamos que se incumple la hipótesis, podemos prescindir de la variable afectada, pues la pérdida de información no nos preocupa, ya que el único problema sería que la información de esta variable resultará redundante.

El problema de la multicoliealidad y singularidad se puede evitar aplicando los métodos de selección de variables indicados en la sección 3.1.

3.2.2. Normalidad multivariante

Gracias a este supuesto, obtendremos con precisión las probabilidades de que un individuo peretenezca a un grupo o a otro, que son las probabilidades a posteriori.

Sabemos que si la distribución conjunta de las variables sigue una normal multivariante, entonces cada variable se distribuirá como una normal también. El recíproco no es cierto, aunque si cada variable se distribuye normalmente, aumentará la probabilidad de que la distribución siga una normal multivariante.

En la práctica, el incumplimiento de esta hipótesis no tendrá graves consecuencias siempre que la distribución de las variables discriminantes no se aleje demasiado de una normal.

El incumplimiento de la normalidad es más grave cuando hay asimetrías. Cuando trabajamos con muestras grandes el problema no será tan grave como cuando trabajamos con muestras pequeñas y de diferente tamaño. En el caso de incumplimiento de la hipótesis de normalidad, debemos prestar especial atención a las clasificaciones obtenidas, pues será menos preciso y hay que tener cuidado en los casos límite de los dos grupos, es decir, en los casos en los que la probabilidad tome valores en torno a 0,50, pues podríamos clasificar

al individuo en el grupo incorrecto debido al incumplimiento de esta hipótesis.

Cabe destacar que si la probabilidad a posteriori de los individuos clasificados correctamente es alta, el incumplimiento de la hipótesis de normalidad no debe haber sido grave mientras que si la probabilidad a posteriori de los individuos clasificados erróneamente es alta debemos pensar en la gravedad del incumplimiento de las hipótesis o en si hemos elegido buenas variables discriminantes.

Test de Kolmogorov-Smirnov

Gil y cols. (2001) aconsejan la comprobación de la normalidad multivariante viendo la normalidad en cada variable. Existen diferentes alternativas para estudiar la normalidad. Cuando se trabaja con tamaños muestrales pequeños es recomendable utilizar el test de Shapiro-Wilks. Otra de las diferentes alternativas es aplicar el test de Kolmogorov-Smirnov. Expliquemos en qué consiste este test.

Consideremos $(X_1, \ldots, X_n)$ una m.a.s. de una variable aleatoria X continua que se distribuye siguiendo una función de distribución F cualquiera. El contraste que se plantea es el siguiente:

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

Este problema se resuelve usando el estadístico de Kolmogorov-Smirnov, que viene dado por:

$$D(X_1 \ldots, X_n) = \sup_{x \in \mathbb{R}} |F^*_{X_1,\ldots,X_n}(x) - F_0(x)|$$

con $F^*_{X_1,\ldots,X_n}$ la función de distribución muestral. El estadístico de Kolmogorov-Smirnov propociona una medida de la discrepancia entre $F^*_{X_1,\ldots,X_n}$ y F_0. El test de Kolmogorov-Smirnov viene dado por:

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

donde d_α verifica:

$$\mathrm{P}_{H_0}(d(X_1, \ldots, X_n) \geq d_\alpha) = \alpha$$

y

$$p - valor = \mathrm{P}_{H_0}[D(X_1, \ldots, X_n) \geq D_{exp}]$$

siendo D_{exp} el valor del estadístico en la muestra observada.

Podemos encontrar la información detallada de este test en el libro de Canavos (2003).

El problema que presenta este test es que debe usarse cuando se conocen los parámetros de la distribución normal que se quiere contrastar, ya que, de lo contrario, los resultados serán muy conservadores. Esto hace que el test no resulte muy útil en la aplicación, ya que la media y la varianza poblacionales no suelen ser conocidas y, con el test de Kolmogorov-Smirnov se deben estimar, lo que conlleva a que no se rechace la hipótesis nula en múltiples ocasiones.

Para solventar este problema se estudiará el test de Lilliefors.

Test de Lilliefors

Lilliefors (1967) creó otra opción menos conservativa basada en el test de Kolmogorov-Smirnov, el test de Lilliefors. El autor realiza una corrección al test desarrollado anteriormente estimando los parámetros de la distribución normal de la muestra a partir de los datos muestrales. Este test puede aplicarse cuando no se conoce ni la media ni la varianza de la distribución, hecho que es especialmente interesante cuando trabajamos con una base con datos observables.

Se puede encontrar la información detallada de este test en la publicación de Lilliefors (1967).

Además de realizar contrastes de hipótesis para probar la normalidad es aconsejable utilizar métodos gráficos para observar visualmente si los datos de la muestra se aproximan a una normal. Se estudiará a continuación.

Procedimientos gráficos

Siguiendo a Gnanadesikan (1977), podemos comprobar la normalidad a partir de procedimientos gráficos, usando por ejemplo los gráficos cuantil-cuantil (gráficos Q-Q). Este procedimiento permite observar la cercanía entre la distribución de un conjunto de datos y la distribución teórica.

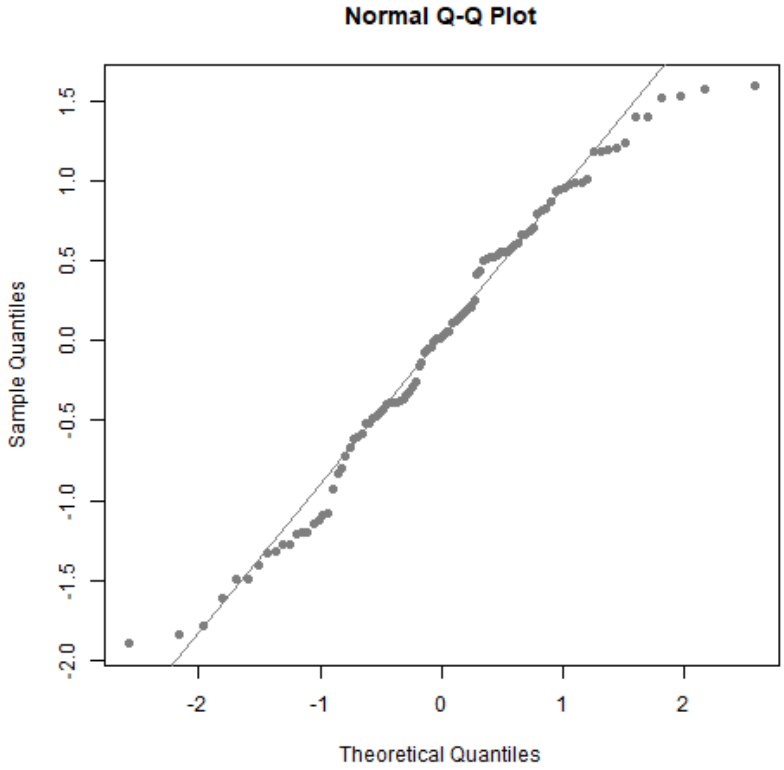

Figura 3.1: Ejemplo del gráfico Q-Q.

En el eje X se grafican los cuantiles de la distribución teórica y en el eje Y, los cuantiles de la distribución del conjunto de datos. Si la distribución del conjunto de datos y la distribución teórica tienen la misma distribución el gráfico cuantil-cuantil será lineal. Podemos ver un ejemplo en la Figura 3.1

3.2.3. Igualdad de matrices de varianzas-covarianzas

El cumplimiento de que las matrices de varianzas-covarianzas extraídas de los grupos sean iguales es necesario para reducir las fórmulas usadas en la obtención de la función discriminante.

Según Gil y cols. (2001), esta hipótesis es muy severa y es cierto que raramente se cumple estrictamente.

Podemos usar diferentes pruebas para comprobar la igualdad de las matrices de varianzas-covarianzas. Destacaremos la prueba M de Box.

Siguiendo la notación usada en el trabajo, es decir:

$$\begin{aligned} N &\equiv \text{número total de individuos,} \\ g &\equiv \text{número de grupos,} \\ n_i &\equiv \text{número de individuos en el grupo } i, \\ I_i &\equiv \text{matriz de varianzas-covarianzas en el grupo } i, \\ E &\equiv \text{matriz entre los grupos,} \end{aligned}$$

el estadístico M de Box viene dado por:

$$M = (N-g)\log|E| - \sum_{i=1}^{g}(n_i - 1)\log|I_i|.$$

El estadístico M no tiene una distribución muestral conocida pero si consideramos el estadístico

$$C = \frac{2p^2+3p-1}{6(p+1)(g-1)}\left(\sum_{i=1}^{g}\frac{1}{n_i-1} - \frac{1}{N-g}\right),$$

entonces se tiene que $M \cdot (1-C)$, bajo la hipótesis de que el vector de variables es normal multivariante en cada grupo, se distribuye aproximadamente como una χ^2 con $\dfrac{p(p+1)(k-1)}{2}$ grados de libertad, siempre que $n_i > 20$.

El contraste de hipótesis viene dado por:

$$\begin{cases} H_0: & \Sigma_1 = \Sigma_2 = \ldots = \Sigma_g \\ H_1: & \text{otro caso} \end{cases}$$

con Σ_i la matriz de varianzas-covarianzas muestral en el grupo i-ésimo.

La prueba M de Box es muy sensible cuando no se cumple la hipótesis de normalidad multivariante. De hecho, hay casos en los que las matrices de varianzas-covarianzas son prácticamente iguales y se han considerado distintas por no verificarse la hipótesis de normalidad multivariante.

La técnica es sensible al incumplimiento de la igualdad de matrices de varianzas-covarianzas, pues en el caso de incumplimiento, los individuos tenderían a ser clasificados en las poblaciones que tengan mayor dispersión. Debemos prestar especial atención cuando las muestras son pequeñas o desiguales.

Capítulo 4

Aplicación a datos reales

En este capítulo se realizará una aplicación del análisis discriminante a datos reales utilizando datos procedentes de 832 pacientes con sospecha de COVID-19 que fueron admitidos en el servicio de urgencias del Hospital Virgen de las Nieves de la ciudad de Granada. Para llevar a cabo el desarrollo de la técnica, se utilizará fundamentalmente el software estadístico R.

La base de datos ha sido cedida por el Doctor D. Antonio Jesús Láinez Ramos-Bossini del Hospital Virgen de las Nieves. Consta de 832 observaciones y contiene los valores de 222 variables de diferente naturaleza, recogidos por el equipo médico del hospital a partir de la anamnesis, de analíticas de sangre y de distintas pruebas médicas.

Se pretende encontrar una función discriminante que permita separar entre pacientes que reciben el alta y pacientes que ingresan en el hospital tras visitar urgencias a partir de los datos de los que se dispone de las variables incluidas en la base de datos. Para ello, se ha considerado como variable categórica la variable `Des_Urg`, que indica si el paciente ha sido ingresado en el hospital o ha recibido el alta.

Se diferenciarán cuatro apartados dentro del capítulo. En primer lugar, se depurarán los datos. Después, se comprobará qué hipótesis verifican las observaciones de la base de datos. A continuación, se aplicará la técnica y, por último se mostrarán las conclusiones. Además, existe un Anexo en el que se muestran las funciones utilizadas con el software R junto con los argumentos necesarios para su uso adecuado.

4.1. Depuración de datos

Antes de comenzar a aplicar la técnica, como suele ser usual en cualquier análisis estadístico, es necesario comenzar depurando los datos. La base de datos se ha reducido de 832 observaciones inciales a 394, que son los casos donde se conoce el desenlace en urgencias de los pacientes tras las pruebas, es decir, donde la variable `Des_Urg` no tiene un dato perdido. Después, se han seleccionado las 89 variables continuas que tiene la base de datos, que son aquellas que se pueden usar en la técnica del análisis discriminante. Dentro de estas variables seleccionadas, se han eliminado las que tienen un porcentaje de datos perdidos superior al 10 %, por lo que la base de datos se reduce a 38 variables continuas. Finalmente, se ha calculado el coeficiente de correlación de Spearman para saber qué variables estaban asociadas con la variable `Des_Urg`, obteniendo un total de 26 variables. Estas variables han sido revisadas para estudiar sus datos anómalos y se han eliminado aquellos que claramente habían sido introducidos de forma errónea en la base de datos. Por último, se ha realizado una imputación de valores perdidos mediante el método de la media.

Como se comentaba anteriormente, el objetivo es encontrar una función discriminante que permita hacer una clasificación de los pacientes que acuden a urgencias, diferenciando los que serán ingresados de los que no, en base a sus resultados en las variables indicadas previamente. Por lo tanto, se aplicará la técnica considerando como variables independientes las 26 elegidas anteriormente y, como variable dependiente o categórica, `Des_Urg`. Si el paciente recibe el alta tras su visita a urgencias diremos que pertenece al grupo `Alta`. Si, por lo contrario, el paciente es ingresado en el hospital diremos que pertenece al grupo `Ingreso`.

Trabajaremos con un `data.frame` de 27 columnas. Las primeras 26 columnas estarán constituidas por las observaciones de las variables comentadas anteriormente y la última columna indicará el grupo al que pertenece cada una de las observaciones.

A continuación, leeremos los datos que tenemos almacenados en un archivo con formato `.csv` a partir de la función `read.csv`. Después, los guardaremos en `datos`.

```
read.csv("datosTFMdef.csv")->datos
```

4.2. Validación de las hipótesis

Para realizar correctamente la técnica del análisis discriminante, tanto el lineal de Fisher como el cuadrático, se tienen que verificar una serie de hipótesis que veremos a continuación.

La primera hipótesis que tenemos que validar es que la variable dependiente sea categórica y las independientes sean de intervalo o de razón. Según lo indicado anteriormente, las variables consideradas la cumplen.

En primer lugar, ya que vamos a trabajar con dos grupos como se ha comentado anteriormente, veamos por cuántos individuos está compuesto cada uno.

```
data.frame(table(datos$Des_Urg))
```

```
##   Var1 Freq
## 1    0  117
## 2    1  277
```

Se puede observar que de los 394 pacientes, 117 reciben el alta y 277 ingresan en Urgencias. Por tanto, como $n_1 = 177$ y $n_2 = 277$, cada uno de los grupos tiene al menos dos individuos, luego la segunda hipótesis también se cumple.

El número de variables discriminantes viene dado por $p = 26$ y el número total de individuos es $N = 394$, luego se verifica que $p < (N - 2)$.

4.2.1. Ausencia de multicolinealidad

Las 26 variables bajo estudio deben ser linealmente independientes entre sí. Este supuesto es inviolable, pues, si no se verifica, la técnica carece de sentido. Crearemos dos variables, `data` y `grupos`. La primera de ellas recogerá las observaciones de cada una de las variables y, la segunda, el grupo al que pertenece cada individuo.

```
data<-datos[,1:26]
grupos<-datos[,27]
```

Para saber si existen problemas de multicolinealidad calcularemos la matriz de correlaciones de las 26 variables bajo estudio, a partir de la función

`cor`, descrita en el Anexo. Al trabajar con un número elevado de variables es complicado visualizar qué variables presentan una fuerte dependencia lineal a partir de la matriz de correlaciones. Resulta útil apoyarse en métodos gráficos. Para ello, utilizaremos la función `corrplot`, descrita en el Anexo, que procede de la librería `corrplot`. Esta función devuelve un correlaciograma en el que las variables que presenten relación lineal directa estarán señaladas en azul y, las que presenten una relación lineal inversa estarán señaladas en naraja. Cuanto más potente sea la relación lineal, mayor será el círculo que relaciona ambas variables.

```
cor(data)->m.cor
corrplot(m.cor, type="upper", tl.col="black", tl.srt=45)
```

En principio, como puede verse en el gráfico 4.1, la ausencia de multicolinealidad no se verifica, pues algunas de las variables presentan una correlación elevada.

Ahora, hay que plantearse qué criterio utilizar para resolver el problema de multicolinealidad. Se detallarán brevemente las opciones seguidas para resolver el problema.

En primer lugar, se utilizó la correlación de Spearman con el software R para saber qué variables tenían una asociación potente con la variable categórica `Des_Urg`. Se seleccionó la variable que presentaba mayor asociación y se eliminaron aquellas variables que eran linealmente dependientes con ella, seleccionado progresivamente las variables con mayores correlaciones con la variable de agrupación, hasta que se obtuvo un conjunto de variables linealmente independientes entre sí y con una alta correlación con el grupo. Este criterio sin embargo no resultó ser preciso, puesto que daba lugar a una función discriminante con poco poder clasificatorio.

Finalmente, se optó por utilizar un método de selección de variables discriminantes. El problema es que R no tiene ninguno implementado para la técnica del análisis discriminante. Por tanto, se decidió utilizar SPSS, que sí que lo tiene implementado para el análisis discriminante lineal. Sin embargo, surge otra dificultad, pues SPSS no tiene implementado ningún método de selección de variables discriminantes para el análisis discriminante cuadrático que, como se verá posteriormente, será el que se utilice para el desarrollo de la aplicación. En definitiva, no existe un método específico para solventar el problema de multicolinealidad en el análisis discriminante cuadrático. Se ha optado por utilizar el método de inclusión por pasos de SPSS para el análisis

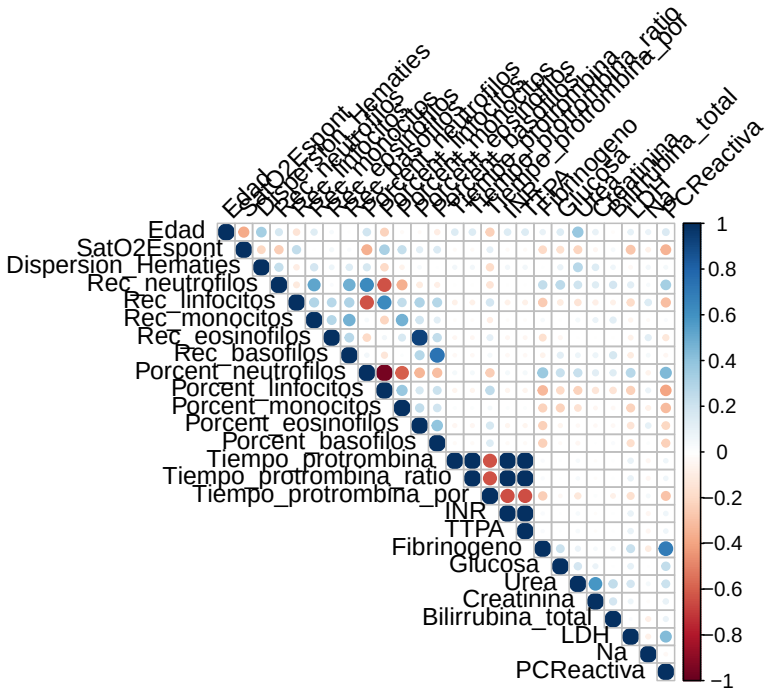

Figura 4.1: Correlaciograma 26 variables.

discriminante lineal, aunque luego se use dicha selección en el cuadrático, ya que se llega a una función discriminante con un poder discriminatorio aceptable. Para ello, una vez cargados los datos se utilizará la siguiente sintaxis en SPSS:

```
DATASET ACTIVATE ConjuntoDatos1.
DISCRIMINANT
/GROUPS=Des_Urg(0 1)
/VARIABLES=Edad SatO2Espont Dispersion_Hematies
Rec_neutrófilos Rec_linfocitos Rec_monocitos
Rec_eosinófilos Rec_basófilos Porcent_neutrófilos
Porcent_linfocitos Porcent_monocitos Porcent_eosinófilos
Porcent_basófilos Tiempo_protrombina Tiempo_protrombina_ratio
Tiempo_protrombina_por INR TTPA Fibrinogeno
Glucosa Urea Creatinina Bilirrubina_total LDH Na PCReactiva
/ANALYSIS ALL
/METHOD=WILKS
/FIN=3.84
/FOUT=2.71
/PRIORS SIZE
/HISTORY
/CLASSIFY=NONMISSING POOLED.
```

El programa selecciona 9 variables, que son las siguientes:

- **Fibrinogeno:** Indica el fibrinógeno en sangre. Se trata de una proteína que se encarga de la formación de los coágulos sanguíneos para frenar el sangrado.
- **Porcent_basofilos:** Indica el porcentaje de basófilos que hay en sangre. Los basófilos son un tipo de glóbulo blanco.
- **LDH:** Indica el LDH en sangre. Es una proteína que se encarga de producir energía necesaria para nuestro organismo.
- **Edad:** Indica la edad del paciente.
- **Porcent_eosinofilos:** Indica el porcentaje de eosinófilos que hay en sangre. Los eosinófilos son un tipo de glóbulo blanco.
- **Rec_eosinofilos:** Indica el recuento total de eosinófilos en sangre.

- **SatO2Espont:** Indica la saturación de Oxígeno en sangre del paciente, es decir, indica la cantidad de Oxígeno disponible en sangre.
- **Na:** Indica la cantidad de sodio en sangre. Es un mineral con carga eléctrica que ayuda a que los músculos y los nervios funcionen de forma correcta.
- **Rec_monocitos:** Indica el recuento total de monocitos en sangre del paciente, que son un tipo de glóbulo blanco.

Tras eliminar las 17 variables que no selecciona el método de inclusión por pasos, veamos qué correlaciograma se obtiene.

```
cor(data)->mat_cor
corrplot(mat_cor, type="upper",
tl.col="black", tl.srt=45,method="number")
```

Eliminamos `Rec_eosinófilos` por tener dependencia lineal con Porcent_eosinófilos, como puede verse en la Figura 4.2. Por tanto, de aquí en adelante trabajaremos con 8 variables independientes.

4.2.2. Normalidad multivariante

El siguiente supuesto que debemos validar es si las variables siguen una normal multivariante en cada grupo. Cabe destacar que, aunque esta hipótesis no se verifique, la técnica puede aplicarse, pero existe la posibilidad de que los resultados pierdan precisión, aunque en múltiples ocasiones se obtienen resultados apropiados.

Se han transformado algunas de las variables utilizando funciones matemáticas como el logaritmo, la raíz cuadrada o la función racional $\frac{1}{x}$, con el fin de poder validar la hipótesis de normalidad. Las variables transformadas son las siguientes:

- **Log_Mono:** Logaritmo de `Rec_monocitos`.
- **Raiz_PEos:** Raíz cuadrada de `Porcent_eosinofilos`.
- **RLog_Baso:** Raíz del logaritmo de `Porcent_basofilos`.
- **Log_Fibri:** Logaritmo de `Fibrinogeno`.
- **Div_LDH:** Cociente de uno y `LDH`.

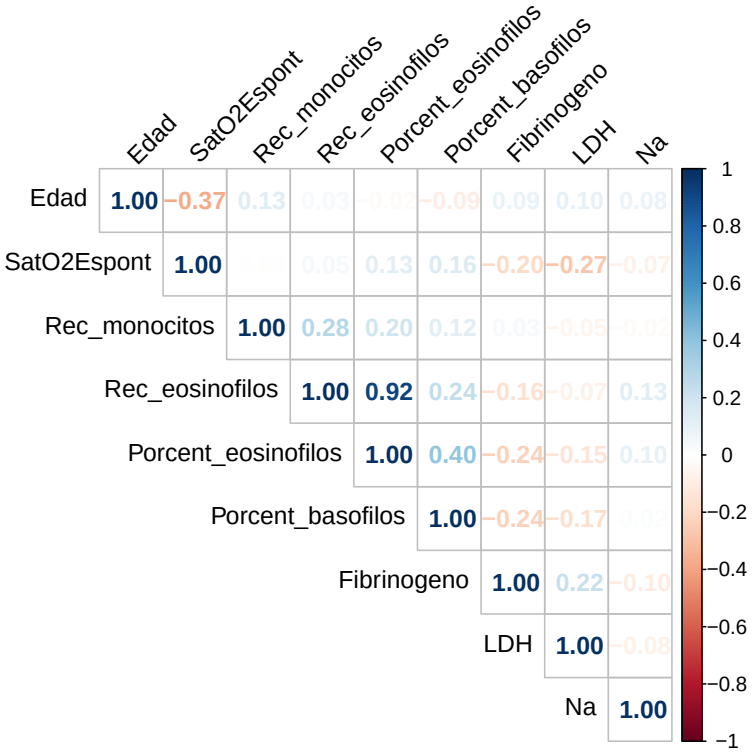

Figura 4.2: Correlaciograma 9 variables.

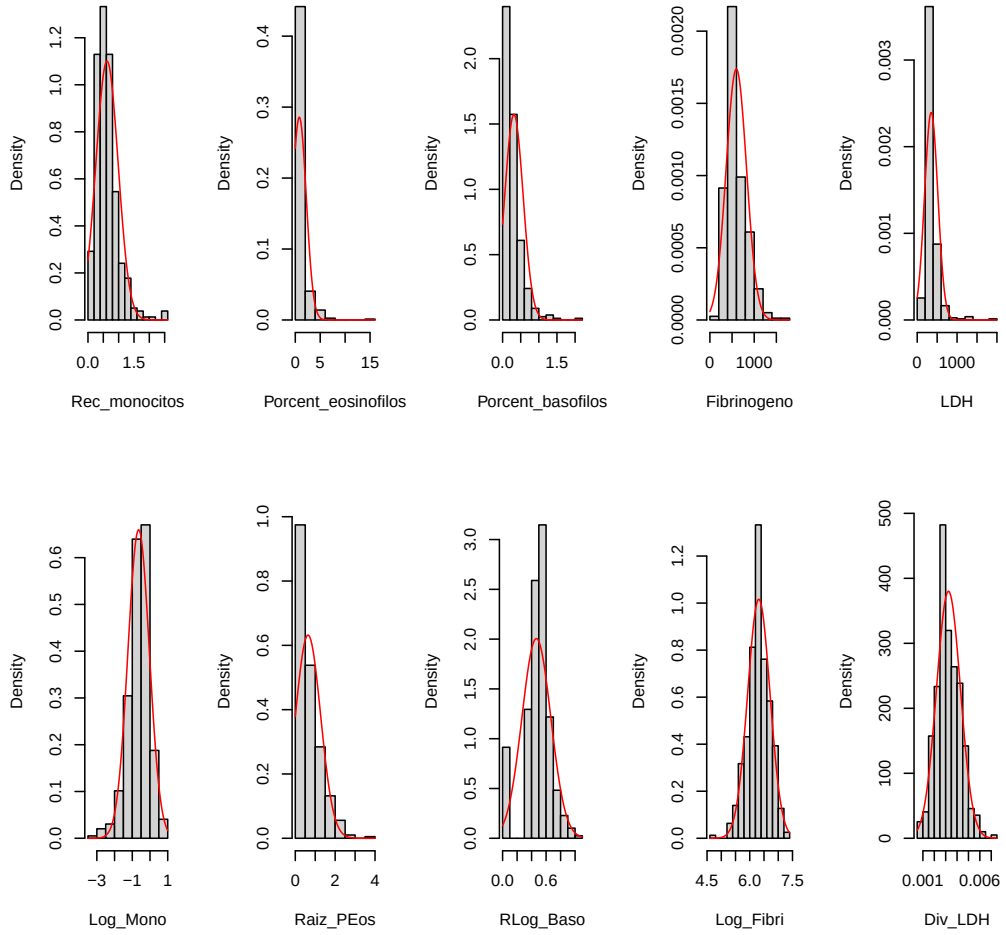

Figura 4.3: Transformaciones de las variables para validar la normalidad.

En la Figura 4.3 pueden verse los histogramas de las variables originales y los histogramas de las variables transformadas.

A continuación, vamos a almacenar los individuos que pertenecen al grupo `Alta` en la variable `datos.alta` y los del grupo `Ingreso` en la variable `datos.ingreso`, utizando la función `subset`, recogida en el Anexo.

```
datos.alta<-subset(datos,datos$Des_Urg==0)
datos.ingreso<-subset(datos,datos$Des_Urg==1)
```

Se probará esta hipótesis observando si cada una de las variables sigue una normal unidimensional en cada uno de los dos grupos. Para ello usaremos

el test de Lilliefors (Kolmogorov-Smirnov) mediante la función `lillie.test` de la librería `nortest`, descrita en el Anexo. El código necesario para la ejecución del test se muestra a continuación para la variable `Edad`. Para el resto de variables el código es análogo.

```
lillie.test(datos.alta$Edad)
lillie.test(datos.ingreso$Edad)
```

Veamos el resultado del test para las 8 variables bajo estudio:

```
lillie.test(datos.alta$Edad)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$Edad
## D = 0.082027, p-value = 0.05113

lillie.test(datos.ingreso$Edad)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$Edad
## D = 0.080105, p-value = 0.0001931
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$SatO2Espont
## D = 0.2033, p-value = 6.369e-13
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$SatO2Espont
## D = 0.161, p-value < 2.2e-16
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$Log_Mono
## D = 0.077839, p-value = 0.07824
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$Log_Mono
## D = 0.052623, p-value = 0.0613
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$Raiz_PEos
## D = 0.061735, p-value = 0.3356
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$Raiz_PEos
## D = 0.19755, p-value < 2.2e-16
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$RLog_Baso
## D = 0.13039, p-value = 4.439e-05
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$RLog_Baso
## D = 0.23276, p-value < 2.2e-16
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$Log_Fibri
```

```
## D = 0.055134, p-value = 0.5167
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$Log_Fibri
## D = 0.066163, p-value = 0.005191
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$Div_LDH
## D = 0.044372, p-value = 0.8277
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$Div_LDH
## D = 0.08017, p-value = 0.0001898
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.alta$Na
## D = 0.15928, p-value = 1.008e-07
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos.ingreso$Na
## D = 0.13776, p-value = 7.354e-14
```

En base a los p-valores resultantes, se obtiene que las variables `Edad` en el grupo `Alta`, `Log_Mono` en ambos grupos, `Raiz_PEos` en el grupo `Alta`, `Log_Fibri` en el grupo `Alta` y `Div_LDH` en el grupo `Alta` pueden considerarse normales. Como se comentaba al principio de este apartado, la ausencia de normalidad es un supuesto que puede no cumplirse debido a la robustez de la técnica, pero debe tenerse en cuenta que se puede llegar a resultados menos precisos.

Estudiemos ahora la normalidad gráficamente a partir de las funciones `qqnorm` y `qqline`. A continuación, se muestra el código para la variable `Edad`. Para el resto de variables el código es análogo.

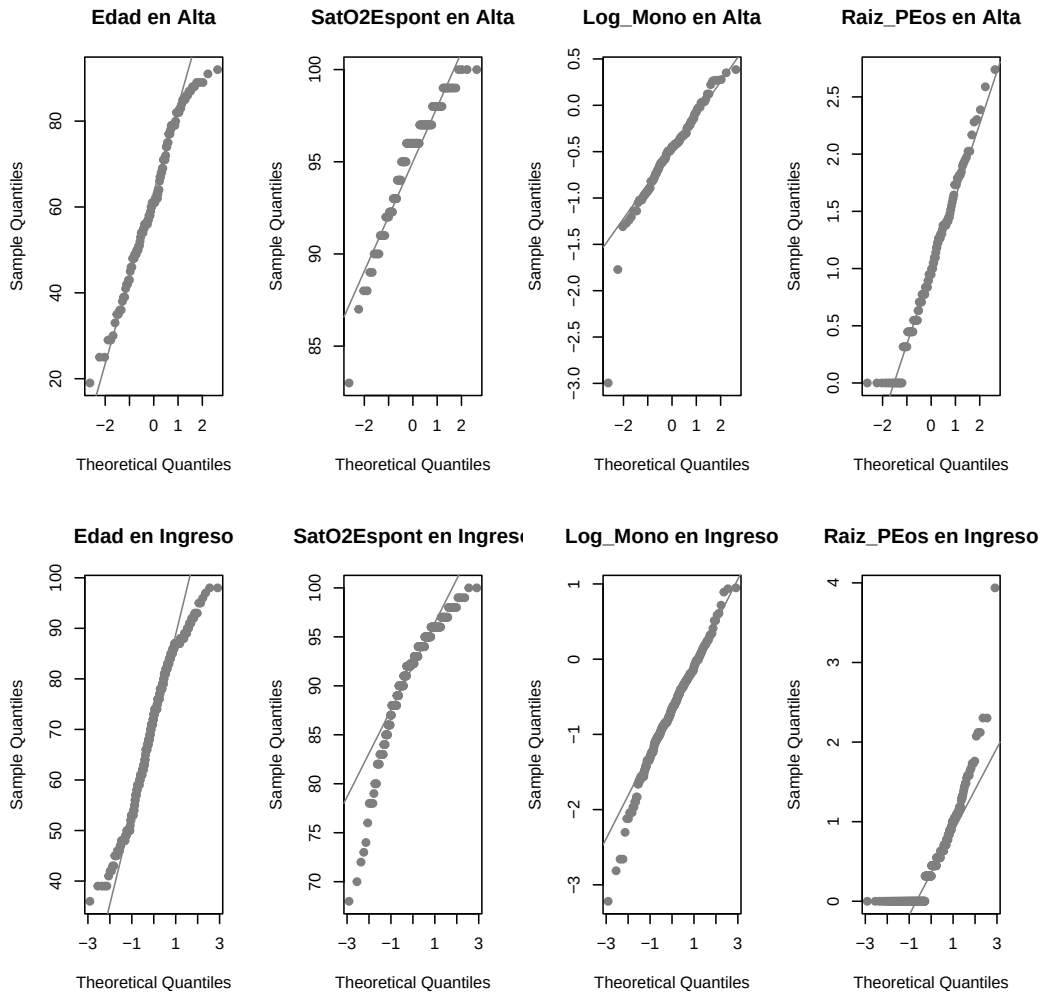

Figura 4.4: Gráficos Q-Q.

```
qqnorm(datos.alta$Edad,pch=19,
col="gray 50",main="Edad en Alta")
qqline(datos.alta$Edad,pch=19,col="gray 50")

qqnorm(datos.ingreso$Edad,pch=19,
col="gray 50",main="Edad en Ingreso")
qqline(datos.ingreso$Edad,pch=19,col="gray 50")
```

Como puede verse en la Figura 4.4, las variables `Edad` y `Log_Mono` se asemejan a una distribución Normal. También es el caso de la variable `Raiz_PEos` en el grupo `Alta`. Sin embargo, la variable `SatO2Espont` presenta problemas

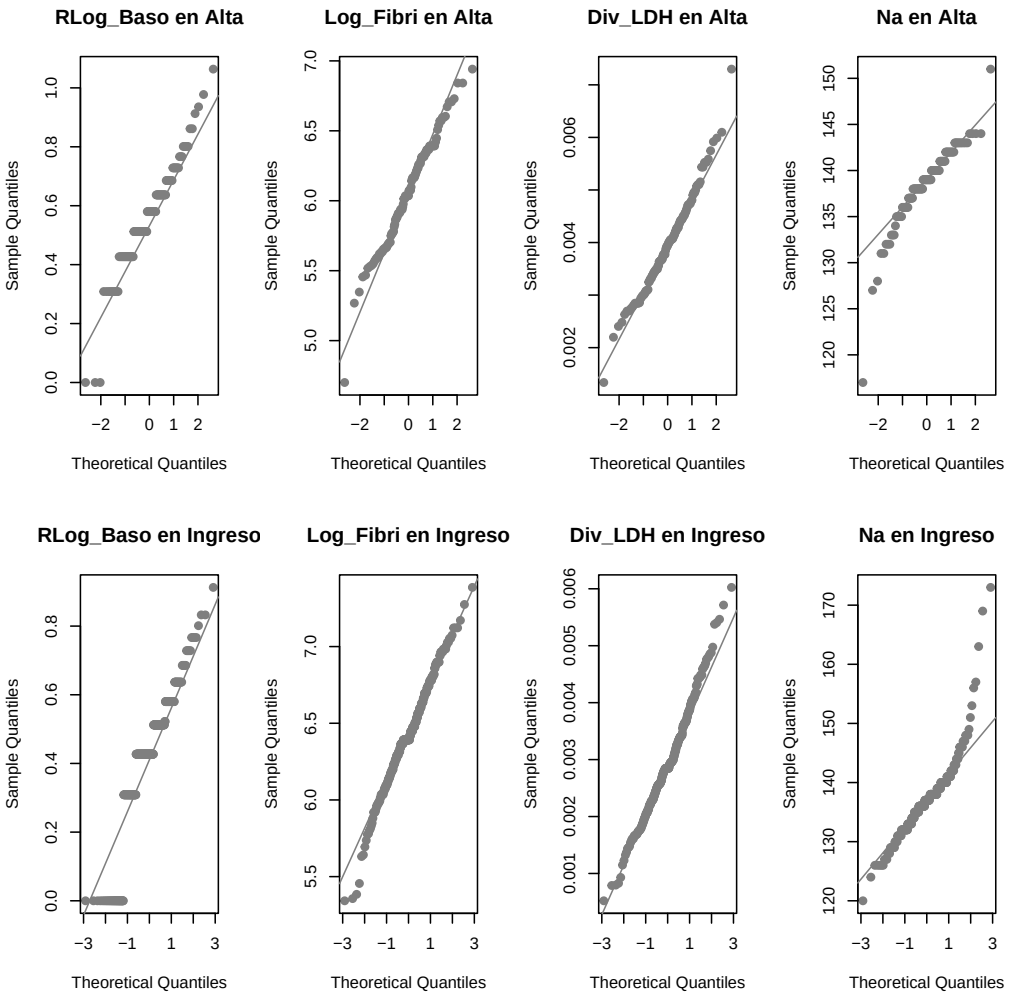

Figura 4.5: Gráficos Q-Q.

de normalidad en ambos grupos. Aunque los resultados reales de la prueba de medición de la saturación de Oxígeno son continuos, la variable que se utiliza en la práctica no es continua puesto que es usual que el equipo médico trabaje con el número entero resultante. Por otra parte, la normalidad en el grupo `Ingreso` de la variable `Raiz_PEos` falla debido a que exite una gran cantidad de individuos de los que son ingresados que tienen un porcentaje de eosinófilos nulo.

En la Figura 4.5 puede verse que las variables `Log_Fibri` y `Div_LDH` se aproximan a una Normal en ambos grupos. La variable `Na` presenta un problema similar a la variable `SatO2Espont`, descrito anteriormente. La variable `RLog_Baso` no se asemeja a una Normal.

4.2.3. Igualdad de matrices de varianzas-covarianzas

El último supuesto que debemos validar es la igualdad de matrices de varianzas-covarianzas. Para ello, usaremos la función `boxM`, descrita en el Anexo, que sirve para aplicar la prueba M de Box que resuelve el contraste de hipótesis de igualdad entre las matrices de varianzas-covarianzas de los grupos bajo estudio. Se debe cargar la librería `biotools`.

```
boxM(data=data[1:8],grouping=grupos)

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data[1:8]
## Chi-Sq (approx.) = 177.06, df = 36, p-value < 2.2e-16
```

Como podemos ver, la función nos devuelve el p-valor del test que, al ser un valor muy cercano a 0, nos indica que, en base a los datos, se rechaza la hipótesis de igualdad de las matrices, por lo tanto no podemos validar esta hipótesis. Asumimos que uno de los grupos es más variable que el otro.

Cabe destacar que la prueba M de Box se ve afectada cuando no existe normalidad multivariante, por lo que es común que matrices iguales aparezcan significativamente distintas cuando no se cumple la hipótesis de normalidad. Además, cuando el tamaño de la muestra es elevado, es muy sencillo rechazar la hipótesis nula. No obstante, en estos casos el análisis discriminante cuadrático es muy útil, pues no necesita que se satisfaga esta hipótesis.

4.3. Aplicación de la técnica

Antes de aplicar la técnica, vamos a dividir la muestra aleatoriamente en dos bases de datos. La primera se llamará `entrenamiento` y estarán recogidas el 80 % de las observaciones. La segunda recibirá el nombre de `prueba` y contendrá el 20 % de las observaciones. Realizaremos el método utilizando el conjunto de datos `entrenamiento` y, posteriormente, se clasificará a los individuos del conjunto `prueba` utilizando el discriminante cuadrático estimado. Para la creación de una muestra aleatoria, se utilizará la función `sample`, descrita en el Anexo.

```
muestra<-sample(1:nrow(datos),315)
entreno<-datos[muestra,]
prueba<-datos[-muestra,]
```

A continuación, realizaremos la aplicación utilizando el discriminante cuadrático, pues, al no verificarse la hipótesis de igualdad de matrices de varianzas-covarianzas, el análisis dicriminante lineal de Fisher pierde eficacia. Se utilizará la función `qda` de la librería `MASS`, descrita en el Anexo.

Previo a su uso podemos hacer un diagrama de dispersión de las nueve variables usando la función `pairs`, que nos permite visualizar los dos grupos en cada par de variables.

```
pairs(x = entreno[, 2:9], col=ifelse(entreno$Des_Urg==1,
"firebrick", "green3"), pch = 19, oma=c(3,3,3,16))
par(xpd=TRUE)
legend(x="bottomright",title = "Grupo",legend=c("Alta",
"Ingreso"),     fill = c("green3","firebrick" ))
```

En base a la Figura 4.6, es evidente que ninguna de las variables discrimina claramente entre los dos grupos bajo estudio. Veamos si el análisis discriminante cuadrático proporciona una función de todas las variables que cumpla con el objetivo. Para ello, como ya hemos indicado previamente, hacemos el análisis usando la variable categórica `Des_Urg` y las variables discriminantes `Edad`, `SatO2Espont`, `Log_Mono`, `Raiz_PEos`, `RLog_Baso`, `Log_Fibri`, `Div_LDH` y `Na` y lo guardamos en la variable `qda.entreno`, obteniendo:

```
qda.entreno <- qda(Des_Urg ~ Edad+SatO2Espont+Log_Mono+
Raiz_PEos+RLog_Baso+Log_Fibri+Div_LDH+Na,
data = entreno)
```

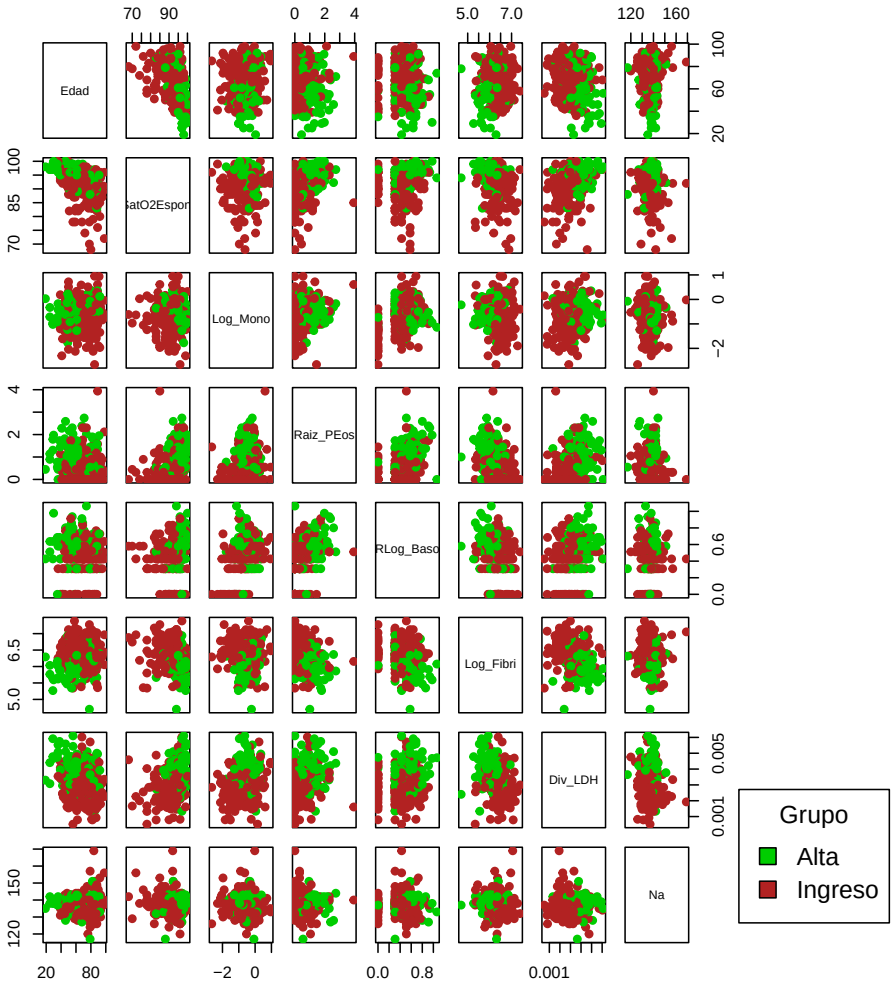

Figura 4.6: Diagrama de dispersión.

```
qda.entreno

## Call:
## qda(Des_Urg ~ Edad + SatO2Espont + Log_Mono + Raiz_PEos+
## RLog_Baso + Log_Fibri + Div_LDH + Na, data = entreno)
##
## Prior probabilities of groups:
##         0         1
## 0.3047619 0.6952381
##
## Group means:
##       Edad SatO2Espont   Log_Mono Raiz_PEos RLog_Baso
## 0 61.05208    95.36275 -0.4714798 1.0323744 0.5591219
## 1 71.12785    91.15795 -0.6706510 0.4756547 0.4214347
##  Log_Fibri      Div_LDH       Na
## 0 6.045849  0.003897398 138.5104
## 1 6.433894  0.002907135 137.1330
```

Usaremos también la función `predict`, descrita en el Anexo, que es una función capaz de hacer predicciones a partir de los resultados obtenidos con la función `qda`. A partir de ella determinaremos el porcentaje de acierto en cada grupo que se obtiene con la función discriminante obtenida y el porcentaje de acierto total de ella. La usamos aplicándola a la variable `qda.entreno`, donde anteriormente guardamos el resultado de la aplicación de la función `qda`, y la guardamos en la variable `p.qda.entreno`:

```
p.qda.entreno<-predict(qda.entreno,entreno[2:9])
```

Podemos ver el porcentaje de acierto con los datos de entrenamiento en la clasificación haciendo lo siguiente:

```
table(entreno$Des_Urg,
p.qda.entreno$class)->m.qda.ent
m.qda.ent

##
##       0   1
##   0  70  26
##   1  23 196
```

Con la función cuadrática, de los 96 pacientes que reciben el alta en urgencias de la muestra de entrenamiento, se han clasificado correctamente 70 y se

han clasificado incorrectamente 26. De los 219 pacientes que son ingresados tras su llegada a urgencias en la muestra de entrenamiento se han clasificado correctamente 196 y se han clasificado incorrectamente 23.

Se tiene por tanto que el porcentaje de acierto por grupo viene dada por:

```
diag(prop.table(m.qda.ent,1))

##         0         1
## 0.7291667 0.8949772
```

Por lo tanto, se puede afirmar que el porcentaje de acierto al clasificar individuos en `Alta` viene dada por 0.7291667 y en `Ingreso`, por 0.8949772, con un porcentaje de acierto total en la muestra de entrenamiento de:

```
sum(diag(prop.table(m.qda.ent)))

## [1] 0.8444444
```

Teniendo en cuenta la capacidad de discriminación que mostraron las variables en el estudio previo, donde se observaba que los datos tenían una gran zona de confusión, la función obtenida nos permitiría clasificar nuevos casos a partir de las variables consideradas con un porcentaje aceptable de acierto.

A continuación, trabajaremos con las observaciones que no han sido utilizadas para la construcción de la función discriminante. Veamos si las clasifica en el grupo correcto o no. Para ello, usaremos de nuevo la función `predict` aplicándola a `qda.entrenamiento`, variable donde está guardado el análisis discriminante cuadrático de la base `entrenamiento`, y a la base de datos `prueba`.

```
p.qda.prueba<-predict(qda.entreno,prueba)
table(prueba$Des_Urg,
p.qda.prueba$class)->m.qda.prueba
m.qda.prueba

##
##      0  1
##   0 17  4
##   1  7 51
```

Se tiene que de los 21 invidivuos que recibieron el alta en la muestra, 17 han sido bien clasficados y 4 no. Por otra parte, de los 58 invidivuos que fueron ingresados en la muestra, 51 han sido bien clasficados y 7 no, con un porcentaje de acierto por grupo de:

```
diag(prop.table(m.qda.prueba,1))

##         0         1
## 0.8095238 0.8793103
```

En vista al resultado, el porcentaje de acierto en el grupo `Alta` viene dado por 0.8095238 y, en el grupo `Ingreso`, por 0.8793103, con un porcentaje de acierto total en la muesta de:

```
sum(diag(prop.table(m.qda.prueba)))

## [1] 0.8607595
```

Por lo tanto, se puede decir que la función discriminante cuadrática clasifica de manera satisfactoria a los nuevos individuos.

4.4. Conclusión

En resumen, el objetivo era encontrar una función discriminante que permitiera separar entre pacientes que reciben el alta y pacientes que son ingresados en el hospital tras visitar urgencias. Se consideró la variable categórica `Des_Urg`, que indica si el paciente recibe el alta o es ingresado en el hospital.

Partíamos de una base de datos con 832 observaciones de pacientes con sospecha de COVID-19 distribuidas en 222 variables de diferente naturaleza. Tras hacer una depuración de datos y seleccionar las variables continuas que se tenían en la base de datos se tienen 394 observaciones y 89 variables continuas. Después, se han seleccionado las variables cuyo porcentaje de datos perdidos era inferior al 10 % para no perder demasiada información, reduciendo las variables a 38. A continuación se ha aplicado la correlación de Spearman de las variables continuas con la variable categórica para saber cuáles de ellas tenían asociación con la categórica, reduciendo el número de variables a 26.

Tras lo anterior se ha procecido a validar las hipótesis de ausencia de multicolinealidad, normalidad multivariante e igualdad de matrices de varianzas-covarianzas. Las 26 variables mencionadas anteriormente presentaban problemas de multicolinealidad, por lo que se ha usado el método de inclusión por pasos de SPSS para solucionarlo, quedándonos con 8 variables que no presentaban este inconveniente. Después, se han transformado cinco de las ocho variables seleccionadas utilizando funciones matemáticas para conseguir normalidad. Por último, se ha intentado validar la hipótesis de igualdad de matrices de varianzas-covarianzas, sin éxito. Este hecho ha sido el detonante para descartar la aplicación del análisis discriminante lineal, pues la igualdad de matrices de varianzas-covarianzas es una hipótesis que se debe satisfacer en dicha técnica. Por tanto, la aplicación se ha llevado a cabo utilizando el análisis discriminante cuadrático, que no necesita que se satisfaga la hipótesis descrita.

Antes de aplicar la técnica, se ha dividido aleatoriamente la base de datos en dos partes, `entrenamiento` y `prueba`, destinando el 80 % y el 20 % de los datos a cada parte, respectivamente. De esta forma, se ha pretendido obtener un porcentaje de acierto en la clasificación realista del modelo creado y se ha obtenido que los nuevos individuos han sido correctamente clasificados en un 86.07595 % de los casos, resultado bastante satisfactorio.

Conclusiones

Como ha podido comprobarse en este Trabajo Fin de Máster, el objetivo consistía en encontrar una función discriminante que permitiera clasificar a nuevos individuos basándonos en las observaciones de una serie de individuos. Antes de comenzar con la aplicación, desarrollada en el capítulo 4, se han formulado detalladamente las funciones discriminantes más usuales en la actualidad, así como las hipótesis necesarias para su correcto desarrollo. En el capítulo 1 se han explicado ambas funciones desde un punto de vista matemático para el problema de dos grupos y, en el capítulo 2 se ha generalizado la idea al caso en el que se trabajen con más de dos grupos.

Además, en el capítulo 3 se han dado nociones sobre cómo seleccionar variables cuando tenemos una gran cantidad de variables discriminantes y cómo validar las hipótesis necesarias para que la técnica del análisis discriminante funcione correctamente.

La aplicación a datos reales de este TFM únicamente se ha desarrollado utilizando dos grupos, pues, a partir de una base de datos procedente del Hospital Virgen de las Nieves de la ciudad de Granada, se ha pretendido encontrar una función discriminante que fuera capaz de diferenciar entre pacientes que reciben el alta y pacientes que son ingresados en el hospital tras su visita a urgencias. Por tanto, se tiene en cuenta una variable categórica que indica si el individuo pertenece a un grupo o a otro. Se ha utilizado la función discriminante cuadrática, debido a que la hipótesis de igualdad de matrices de varianzas-covarianzas no se satisfacía, por lo que el análisis discriminante lineal no era válido. Tras aplicar el análisis discriminante cuadrático en R se ha obtenido un porcentaje de acierto total del 84.44444 % y del 86.07595 % en las muestras de entrenamiento y de prueba, respectivamente, lo que quiere decir que la función discriminante estimada tiene un porcentaje de acierto elevado y puede utilizarse para clasificar a futuros nuevos individuos.

Las perspectivas futuras de este Trabajo Fin de Máster deberían ir di-

rigidas a implementar un método de selección de variables con el software estadístico R, pues carece de alguno de los estudiados para la técnica del análisis discriminante. Sería muy apropiado para mejorar la selección de variables para evitar los problemas de multicolinealidad y poder perfeccionar la función discriminante conseguida.

Anexo

En este Anexo se describen los parámetros que admiten las funciones utilizadas en R para hacer la aplicación a datos reales.

cor

cor(x,y,method)

`x`: Vector numérico o matriz de datos.

`y`: Vector o matriz con dimensiones compatibles con `x`.

`method`: Indica qué coeficiente de correlación se va a calcular, como *pearson*, *kendall* o *spearman*.

corrplot

corrplot(corr,type,tl.col,tl.srt,method)

`corr`: Matriz de correlaciones.

`type`: Se utiliza para mostrar la matriz completa, 'full', triangular superior, 'upper', o triangular inferior, 'lower'.

`tl.col`: Color de la etiqueta del texto.

`tl.srt`: Indica los grados de rotación del texto.

`method`: Método de visualización de la matriz de correlaciones. Entre otros se encuentra la representación a partir de círculos, 'circle', o la que indica el coeficiente de correlación, 'number'.

subset

subset(x,subset)

`x`: Objeto del que queremos obtener el subconjunto.

`subset`: Expresión lógica que indica elementos o filas que se quieren conservar. Los valores restantes se eliminarán.

lillie.test

lillie.test(x)

`x`: Vector numérico de datos.

qqnorm y qqline

qqnorm(x,pch,col,main)

Esta función genera un gráfico Q-Q de los valores de la muestra.

`x`: Muestra de datos.

`pch`: Diferentes símbolos gráficos a elegir que se usan para representar los puntos de la muestra.

`col`: Color de los símbolos que representan la muestra.

`main`: Nombre del gráfico generado.

qqline(x,pch,col,main)

Esta función genera una recta al gráfico anterior que pasa por el primer y el tercer cuantil normal teórico.

`x`: Muestra de datos.

`pch`: Permite alterar el formato de la recta.

`col`: Color de los puntos de la muestra.

boxM

boxM(data,grouping)

`data`: `Data.frame` o matriz que contiene N observaciones de p variables.

`grouping`: Vector de longitud N que indica a qué grupo pertenece cada observación.

sample

sample(x,n)

`x`: Vector de uno o más elementos del que se elige la muestra.

`n`: Número de elementos que se desean elegir.

pairs

pairs(x,col,pch,oma)

`x`: Matriz o `data.frame` que indica las coordenadas.

`col`: Color de los puntos de la muestra.

`pch`: Diferentes símbolos gráficos a elegir que se usan para representar los puntos de la muestra.

`oma`: Sirve para desplazar la Figura. En este caso se ha utilizado para incluir la legenda al gráfico.

qda

qda(formula,data,prior)

`formula`: Variable categórica$\sim X_1 + X_2 \ldots$, siendo X_i las variables discriminantes.

`data`: `Data.frame` a partir del cual se toman las variables especificadas en `formula`.

`prior`: Probabilidades a priori de pertenencia a cada grupo. Si no se especifica se calcularán automáticamente las proporciones en función del número de individuos.

Valores que devuelve `qda`

`Call`: Función que se ha considerado.

`Prior probabilities of groups`: Probabilidades a priori de pertenencia a cada grupo.

`Means groups`: Medias de cada grupo.

`Coefficients of linear discriminants`: Coeficientes del discriminante cuadrático.

predict

predict(x,newdata)

`x`: Objeto para el que se desea la predicción.

`newdata`: `Data.frame` con los nuevos individuos que se desean clasificar.

Valores que devuelve predict

`class`: Vector que devuelve la categoría a la que pertenece cada individuo.

`posterior`: Vector que devuelve las probabilidades a posteriori de clasificación.

`x`: Puntuaciones de las funciones discriminantes.

Referencias

Berrar, D. (2018). *Cross-Validation*. Reference Module in Life Sciences.

Canavos, G. C. (2003). *Probabilidad y Estadística. Aplicaciones y Métodos*. McGraw-Hill.

Cea, M. A. (2016). *Análisis Discriminante*. Centro de Investigaciones Sociológicas. CIS.

Cuadras, C. M. (2018). *Nuevos Métodos del Análisis Multivariante*. CMC Editions.

Fisher, R. (1936). *The use of multiple measurements in taxonomic problems*. Annals of Eugenics.

Fix, E., y Hodges, J. (1951). *An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation*. International Statistical Review.

Gil, J., García, E., y Rodríguez, G. (2001). *Análisis Discriminante*. La Muralla.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. John Wiley and Sons Inc.

Lilliefors, H. (1967). *On the Kolmogorov Smirnov Test for Normality with Mean and Variance Unknown*. Journal of the American Statistical Association.

Mahalanobis, P. C. (1936). *On the Generalised Distance in Statistics*. Science and Culture.

Morant, G. M. (1936). *A Biometric Study of the Human Mandible*. Biometrika.

Pearson, K. (1926). *On the Skull and Portraits of George Buchanan.* Oliver and Boyd.

Smith, C. (1947). *Some examples of discrimination.* Annals of Eugenics.

Welch, B. L. (1939). *Note on Discriminant Functions.* Biometrika.